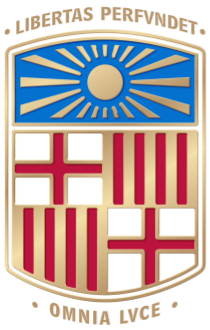



UNIVERSITAT DE BARCELONA

Trabajo Final de Grado

# ESTUDIO ESTADÍSTICO DEL GASTO TURÍSTICO EN ESPAÑA

**Joan Parera Ferreira**

Grado en Administración y Dirección de Empresas

Tutor: Sergi Ramírez Mitjans

Facultad de Economía y Empresa

Curso académico 2022-2023

# RESUMEN Y PALABRAS CLAVE

El turismo desempeña un papel crucial en la economía española, siendo uno de los sectores más importantes. Con el fin de comprender mejor este fenómeno, se lleva a cabo un estudio estadístico con el programa RStudio utilizando datos recopilados del Instituto Nacional de Estadística (INE). Este estudio se centra en analizar las variables que influyen en el gasto turístico en España, tanto a nivel individual como grupal, con el objetivo de comprender su comportamiento.

Para llevar a cabo esta investigación, se va a examinar minuciosamente diversas variables que pueden influir en el gasto turístico, como la procedencia de los turistas, el lugar de destino, el número de pernoctaciones y otros factores relevantes. Mediante el análisis de estas variables, se busca identificar patrones y tendencias que puedan ayudar a comprender y predecir el comportamiento del gasto turístico. Para este análisis se elaborarán diferentes gráficas y valores extraídos del tratamiento de la base de datos en RStudio.

Para finalizar, se realizan diferentes métodos de previsión para lograr la mejor aproximación posible al comportamiento del gasto turístico a nivel individual. Estos métodos incluyen técnicas estadísticas y modelos de pronóstico avanzados realizado en RStudio que permiten desarrollar predicciones futuras con cierto grado de precisión.

**Palabras clave**: turismo, gasto, métodos de previsión, variable, modelo, función y base de datos.

## ABSTRACT AND KEYWORDS

Tourism plays a crucial role in the Spanish economy, being one of the most important sectors. In order to better understand this phenomenon, a statistical study is carried out using RStudio program and data collected from the National Institute of Statistics (INE). This study focuses on analyzing the variables that influence tourist expenditure in Spain, both as an individual and as a group level, with the aim of understanding its behavior.

To carry out this research, various variables that can influence tourist spending will be thoroughly examined, such as tourists' origin, destination, number of overnight stays, and other relevant factors. By analyzing these variables, the goal is to identify patterns and trends that can help understand and predict tourist spending behavior. For this analysis, different graphs and values extracted from the treatment of the database in RStudio will be elaborated.

In conclusion, various forecasting methods are employed to achieve the best possible approximation of individual-level tourist spending behavior. These methods include statistical techniques and advanced forecasting models performed in RStudio, which enable the development of future predictions with a certain degree of accuracy.

**Keywords:** tourism, spent, forecasting methods, variable, model, function, and database.

# ÍNDICE

# 1. Introducción

Este trabajo de final de grado (TFG) consiste en detectar patrones que influyen en el gasto de los turistas en España, esto se va a realizar según la encuesta EGATUR que proporciona el Instituto Nacional de Estadística (INE). Este estudio se va a realizar con el soporte del programa informático RStudio, en el cual se llevará a cabo toda la parte numérica del trabajo. De esta manera se intentará encontrar aquellos datos más determinantes para el gasto del turismo en España y la obtención de modelos de predicción fiables en base algunos patrones.

## 1.1 Propósitos y contextualización del trabajo

Este estudio viene motivado por encontrar modelos de predicción hábiles y datos relevantes en uno de los sectores más importantes del país como es el turismo.

Mediante la encuesta realizada por el Instituto Nacional de Estadística el cual recoge muchas variables influyentes del turismo; destino, motivo, época del año, situación familiar… Se busca encontrar un modelo predictivo eficiente en el que poder encontrar el comportamiento de determinadas personas.

Cabe destacar que el contexto actual de la situación del turismo está cambiando respecto a los años anteriores. En 2020 se produce una pandemia mundial derivada del SarsCOV2 la cual comporto innumerables restricciones a nival nacional como internacional. Este hecho afecto directamente al turismo, por lo tanto, los datos escogidos para esta muestra se encuentran condicionados por este hecho que se produjo hace tres años atrás.

## 1.2 Justificación

Por lo que hace referencia al tema en cuestión; que es un estudio estadístico del gasto turístico en España con RStudio quería dividirlo en dos partes.

La primera es el tema del turismo en España, actualmente el turismo es una de las principales fuentes de ingresos para el país, generando empleos y contribuyendo al crecimiento económico del país. Además, el turismo impulsa otros sectores como la hostelería, la restauración, el transporte, el comercio y la cultura.

Esto es así gracias a que España es uno de los destinos turísticos más visitados del mundo con una gran variedad de atracciones turísticas como playas, ciudades históricas, festivales y eventos culturales, entre otros.

Por estas razones considero que es algo primordial y esencial entender este sector, para ello se necesita hacer un estudio estadístico para conocerlo bien y poder potenciarlo más. Con este estudio pretendo aprender de qué manera se gasta el dinero aquí en España y de donde proviene; los países que más gastan en nuestro país, las comunidades autónomas que perciben el mayor gasto por los turistas y en que se servicios se gasta, entre otras cosas.

Una vez aprendido el comportamiento llega la parte de predecirlo, si conseguimos pronosticar la conducta de los turistas se puede llegar ayudar a las empresas con sus decisiones estratégicas.

La segunda parte en la cual defino la motivación por este trabajo es la parte estadística, en mi ámbito personal y profesional me gustan mucho los números y la tecnología. En este trabajo pretendo aprender mezclar la estadística con un programa informático diseñado para ello, RStudio.

Con el trabajo final de grado lograre aprender un poco de lenguaje de programación y a programar con RStudio, estas competencias me parecen super interesantes para mi vida profesional. Lograré extraer datos, situarlos o nombrarlos como considere mejores y extraer predicciones fiables de este estudio.

El conocimiento que pretendo adquirir en este proyecto me puede ayudar a desarrollar estudios similares de diferentes temáticas. La estadística se puede aplicar a muchos ámbitos de la vida personal y profesional.

Antes de acabar con este punto una de las razones para escoger esta temática de trabajo final de grado fue mi buen grado de satisfacción la asignatura métodos de previsión, una asignatura en la que se aplicaba de forma práctica nuestros conocimientos de estadística. El profesor que tuve explicó la asignatura con mucha dedicación y eso incentivo aún más mi pasión por la estadística.

## 1.3 Objetivos

El objetivo principal del proyecto es lograr detectar patrones de gastos según la encuesta EGATUR que proporciona el Instituto Nacional de Estadística para conseguir un sistema de predicción del gasto del turismo en España en base a las características principales encontradas en el estudio, es decir lograr un método de predicción óptimo.

Otros objetivos generales del trabajo son los siguientes:

Aprender a utilizar RStudio y su lenguaje de programación: En la universidad nos han enseñado fundamentos muy básicos de este programa, por tanto, uno de los objetivos es adquirir el conocimiento necesario para desarrollar este estudio estadístico de forma autónoma.

Describir y caracterizar las variables más relevantes del estudio: Para poder desarrollar un buen estudio estadístico se han de entender las variables sujetas a estudio y cuáles son sus principales características, después de eso encontrar las más relevantes.

Organización y planificación: Para poder desarrollar correctamente un trabajo final de grado se necesita una buena planificación inicial con el tiempo de trabajo para poder cumplir con las entregas previamente estructuradas. Todo esto ha de ir acompañado de una buena organización y ejecución de las diferentes tareas que hay en el trabajo.

Demostrar el dominio de los conocimientos adquiridos durante el grado de Administración y Dirección de Empresas: Durante la carrera universitaria he adquirido muchos conocimientos necesarios para este trabajo, es por lo que uno de los objetivos es tener la capacidad de aplicarlos de manera efectiva. Estos conocimientos pueden venir obtenidos por asignaturas como son estadística y métodos de previsión entre otras.

## 1.4 Metodología

Para la elaboración de la parte práctica del TFG se va a utilizar como información principal los datos extraídos del Instituto Nacional de Estadística de la encuesta de gasto turístico (EGATUR) cogiendo el mes de agosto de 2022 como muestra, ya que consideramos que es el mes más representativo para el turismo. Es preciso comentar que en un inicio se intento realizar este trabajo con una muestra de doce meses, pero debido a la gran extensión de esta base de datos el ordenador utilizado no conseguía ejecutarlos en el programa escogido, por esa razón se reduce la base de datos al mes de agosto.

En el instituto nacional de estadística encontramos datos desde febrero de 2015 hasta septiembre de 2022, con aproximadamente 16.400 personas encuestadas cada mes, esto hace tener una muestra muy extensa durante un largo periodo de tiempo.

En estas encuestas encontramos que nos crea una base de datos con la información del país de procedencia, el motivo del viaje, tipo de alojamiento, actividades que realiza y gastos, entre otros datos interesantes.

Otro soporte para la realización de la parte práctica es sobre la utilización del programa RStudio, con el que se realizara el trabajo. Para este caso usaremos un material de información de la propia página de RStudio y de la página web Stake Overflow donde hay lecciones para aprender a utilizar la herramienta estadística del trabajo.

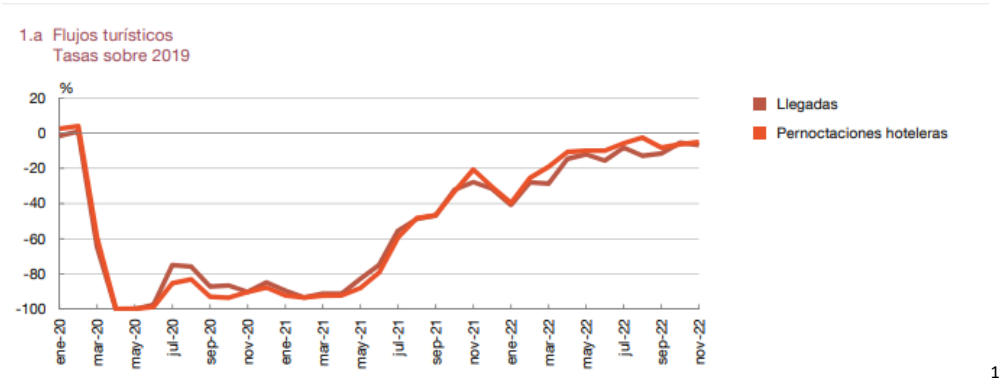
# 2. Metodología

## 2.1 Contexto del turismo en España

El turismo es uno de los sectores más relevantes en España. Según EXCELTUR, una asociación sin ánimo de lucro, formada por las 33 empresas más relevantes de la cadena de valor turística ha informado que en el año 2022 el PIB turístico alcanza los 159 mil millones de euros nominales, superando por un 1,4% la actividad en 2019, época antes de la pandemia. A más también explica que el 61% del crecimiento económico de España en 2022 es gracias al turismo. Con estos datos se puede ver la gran importancia del sector en la economía española.

También hay que destacar que en el instituto nacional de estadística se elaboró un informe en 2022, en el que mostraba los datos del 2021. En este año el turismo representaba el 8% del PIB total de la economía española, también es uno de los sectores que aporta más al empleo español con 2,27 millones de puestos de trabajo, un 11,4% del total.

Por otro lado, también cabe destacar un documento elaborado por analistas de la situación económica del banco de España en la que destacan que el sector se está recuperando y que los ingresos turísticos alcanzan ya niveles precrisis. El nivel de afluencia aún no está recuperado en su totalidad debido a la debilidad que mantiene el turismo de larga distancia, haciéndose especial énfasis en el asiático.




[1]

Hay que destacar también que las empresas turísticas necesitan saber sobre sus clientes para poder personalizar más sus productos. En otros sectores se hace mucha inversión en conocer los datos de los consumidores a través de las redes sociales, en este caso es igual, saber los datos de sus clientes los turistas.

[1] Gráfico 1a de García Esteban, Coral, Ana Gómez Loscos y César Martín Machuca. (2023). "La recuperación del turismo internacional en España tras la pandemia". Boletín Económico - Banco de España, 2023/T1, 01. https://doi.org/10.53479/25114

## 2.2 Modelos *Machine Learning* Supervisados

El *machine learning*, o aprendizaje automático, es una rama de de la inteligencia artificial (IA) que se enfoca en desarrollar algoritmos y técnicas que permiten a las "maquinas" aprender y mejorar automáticamente a través de la experiencia y los datos, en lugar de ser programadas explícitamente para llevar a cabo tareas específicas.

En lugar de seguir instrucciones detalladas, los algoritmos de *machine learning* se entrenan con conjuntos de datos para aprender patrones y relaciones inherentes a esos datos. Utilizan métodos estadísticos y técnicas de optimización para ajustar sus modelos y hacer predicciones o tomar decisiones basadas en nuevos datos de entrada.

El objetivo principal del *machine learning* es desarrollar modelos y algoritmos que puedan generalizar a partir de los datos de entrenamiento para realizar predicciones precisas o tomar decisiones informadas sobre nuevos datos sin etiquetar.

Dentro del *machine learning* encontramos el supervisado y el no supervisado. En los supervisado durante el proceso de entrenamiento se les proporciona un conjunto de datos de entrada, junto con las etiquetas o respuestas correctas correspondientes. Es decir, una persona le proporciona los datos de entrenamiento y los de respuesta.

El objetivo de estos modelos es aprender una función o un patrón en los datos de entrenamiento, de modo que puedan generalizar y realizar predicciones precisas sobre nuevos datos no etiquetados.

En los supervisados encontramos que existen dos tipos: los de clasificación, que se encargan de clasificar un objeto dentro de diversas clases, y de regresión, que predice un valor numérico.

Dentro de los supervisados encontramos modelos como la regresión lineal, los árboles de decisión y las máquinas de vectores de soporte (SVM), entre otros.

El *machine learning* no supervisado utiliza técnicas para descubrir patrones, estructuras ocultas o relaciones en conjuntos de datos. A diferencia del aprendizaje supervisado, no se proporcionan etiquetas o respuestas previas a la base de datos durante el proceso de entrenamiento.

Estos algoritmos exploran las características y la distribución de los datos para generar agrupamientos o segmentaciones automáticas.

Algunos algoritmos del *machine learning* no supervisado es el clustering, clasifica en grupos los datos de salida, y asociación, descubre reglas dentro del conjunto de datos.

## 2.3 Métodos de previsión

Para realizar este trabajo se necesitan diferentes métodos de previsión para encontrar los mejores y así obtener las mejores predicciones. Por esa razón se van a realizas: *XGBoost,* árbol de decisión y *Random Forest*. A más necesitamos comprender correctamente las variables que se van a estudiar, por esa razón, aparte de realizar análisis univariantes y bivariantes con gráficas, también se realizara un *Clustering.*

***Clustering:***

El clustering, también conocido como análisis de agrupamiento, es una técnica de aprendizaje que se utiliza para identificar patrones y estructuras ocultas en un conjunto de datos. El objetivo del clustering es agrupar variables en grupos homogéneos y con la mayor diferencia entre los grupos.

Estos grupos o *clusters* de observaciones que comparten características similares o tienen una estructura similar pueden ser atributos numéricos, categóricos o una combinación de ambos. El algoritmo de esta herramienta cada observación a un grupo de acuerdo con su similitud con otros objetos de esa misma base de datos.

Dentro del *Clustering* encontramos dos tipos diferentes; divisivo y aglomerativo.

El divisivo consiste en que desde un primer grupo donde se encuentran todas las variables este se va dividiendo en subgrupos, esto se repite recursivamente en cada grupo hasta que haya un grupo para cada observación

El aglomerativo se realiza a la inversa del divisivo, este parte que cada variable tiene su propio clúster y estas se van agrupando por similitud entre ellas hasta acabar fusionándose en uno.

***Arboles de decisión***

Los árboles de decisión son un método usado en distintas disciplinas como modelo de predicción, estos se basan en tomar decisiones en base a una regla e ir creando ramas.

Hay diferentes maneras de obtener árboles de decisión, entre ellas está el CART: *Classification And Regression Trees*. Esta técnica lo que hace es buscar la variable objetivo mediante una función a partir de variables predictoras, y el objetivo de este método es encontrar esta función.

La técnica CART puede obtener arboles de clasificación para variable objetivo de carácter discreto y de regresión para cuando es continua.

En este trabajo utilizaremos el *Recurive Partitioning and Regression Trees* (RPART), lo que hace este algoritmo es encontrar la variable independiente que mejor separa nuestros datos en grupos, que corresponden con las categorías de la variable objetivo. Esta separación se expresa con una regla que, a su vez, esta representa un nodo.

Una vez hecha la separación en grupos a partir de la regla, se repite el proceso. Esto se hace porque buscamos la variable que mejor separa los datos en grupo, por tanto, se realiza esto de manera recursiva hasta que nos es imposible obtener una mejor partición, en ese momento el algoritmo se detiene. Aquellos grupos que no pueden ser mejor separados se les llama nodo terminal u hoja.

Una característica de este algoritmo es que cuando una variable ha sido elegida para separar los datos ya no se vuelve a usar en los grupos que ha creado. Se buscan variables distintas que mejoren la separación de datos.

El resultado de todo esto es una serie de bifurcaciones que tiene la apariencia de un árbol que va creciendo ramas.

Las ventajas de este método es su interpretabilidad ya que nos da un conjunto de reglas a partir de las cuales podemos tomar decisiones y a la hora visualizarlo es más sencillo de entender.

Por otro lado, encontramos sus principales desventajas y es que es un tipo de clasificación débil, sus resultados pueden variar mucho dependiendo de la muestra escogida para entrenar un modelo. Y generalmente se sobre ajusta el modelo.

***Random forest***

Para entender en que consiste un *Random Forest*, primero tenemos que entender las desventajas de un árbol de decisión. Los árboles de decisión tienden a sobre ajustar, es decir, tienden a predecir bastante bien pero su generalización no es tan buena. Una forma de mejorar la generalización de los árboles es combinar varios árboles y en eso consiste el *Random Forest*.

Este modelo es un conjunto de árboles de decisión combinados con *bagging. Bagging* lo que hace es que cada modelo se entrena con subconjuntos del conjunto de entrenamiento. Por tanto, podemos decir que este método predictivo crea distintos árboles que ven distintas porciones de los datos y estos se entrenan con distintas muestras para un mismo problema. De esta manera, al combinar los resultados, unos errores se compensan con otros y tenemos una predicción que generaliza mejor.

***XGBoost:***

*XGBoost* es un algoritmo de aprendizaje automático basado en un árbol de decisiones y utiliza un marco de potenciación de gradientes. Su nombre es la abreviatura de las palabras extreme *gradient boosting* (refuerzo de gradientes extremo).

Este método de previsión lo que hace en primer lugar es realizar una predicción inicial. Los residuales se calculan en función del valor predicho y de los valores observados. Se crea un árbol de decisión con los residuales. Se calcula la similitud de los datos de una hoja, así como la ganancia de similitud de la división posterior. Se comparan las ganancias para determinar una entidad y un umbral para un nodo. El valor de salida de cada hoja también se calcula mediante los residuales. Para la clasificación, los valores se calculan generalmente utilizando

el registro de momios y probabilidades. La salida del árbol se convierte en el nuevo residual para el *dataset*, que se utiliza para construir otro árbol. Este proceso se repite hasta que los residuales dejan de reducirse, o bien el número de veces especificado. Cada árbol subsiguiente aprende a partir de los árboles anteriores y no tiene asignado el mismo peso.

Con todos los árboles creados y dando más peso a los últimos creados ya que tienen una tasa de aprendizaje más alto se suma a la predicción inicial para llegar a un valor final.

*XGBoost* utiliza una serie de parámetros y métodos para proporcionar mejores resultados; estos son la regularización, se utiliza un parámetro para reducir la sensibilidad a los datos individuales y evitar el exceso del ajuste. El corte, se selecciona un parámetro para comparar y así evitar el exceso de ajuste al cortar las ramas innecesarias y reducir la profundidad de los árboles. Boceto de cuantil ponderado, para poder llevar a cabo una mejor predicción se evalúa cada valor con cuantiles ponderados en vez de como umbral para dividir los datos. Aprendizaje paralelo, este método divide los datos en bloques para crear árboles. Acceso sensible al caché, este método usa la memoria cache para calcular las puntuaciones de similitud y los valores de salida.

En resumen, *XGBoost* genera múltiples modelos de predicción "débiles" secuencialmente, y que cada uno de estos tome los resultados del modelo anterior, para generar un modelo más "fuerte", con mejor poder predictivo y mayor estabilidad en sus resultados. Durante el entrenamiento, los parámetros de cada modelo débil son ajustados iterativamente tratando de encontrar el mínimo de una función objetivo, que puede ser la proporción de error en la clasificación, el área bajo la curva (AUC), la raíz del error cuadrático medio (RMSE) o alguna otra.

# 3. Desarrollo y resultados

## 3.1 Inicio proyecto y descarga de datos

El estudio se desarrollará por la plataforma RStudio, para poder llevar un correcto control de todo el material utilizado y necesario procedemos a crear una carpeta en el escritorio. Posteriormente generaremos las subcarpetas donde almacenaremos la siguiente información; bases de datos del estudio, inputs descargados para su utilización, outputs creados por nosotros y los scripts con el código correspondiente del trabajo.

Al abrir el script de RStudio en el que vamos a trabajar el primer paso que vamos a realizar es descargar los paquetes necesarios para elaborar el análisis estadístico correctamente.

El siguiente paso que hemos de realizar es la elaboración de la base de datos inicial conjunta, en la página web del Instituto Nacional de Estadística (INE), podemos encontrar los datos de las encuestas de turismo realizadas por la entidad. Los resultados de estas encuestas están divididos en periodo mensual desde 2015 hasta 2022.

Descargamos el periodo muestral a analizar que comprende desde agosto de 2021 hasta setiembre de 2022. Una vez tenemos estos documentos, hemos de escoger los que realmente nos interesan, son aquellos que están en formato excel y acabados en .csv para que lo pueda leer R.

Una vez tenemos los datos que vamos a utilizar los juntamos todos en una misma base de datos para trabajar con ellos y los guardamos como nuestra base de datos inicial.

## 3.2 Transformación Variables Categóricas.

Para poder desarrollar el análisis de todas las variables hemos de cambiar los números de las variables categóricas en texto. En total hay 103 variables categóricas que hemos de cambiar por su explicación correspondiente.

Primero descargamos el documento de diseño de registros y valores válidos de las variables del INE y lo ponemos en la carpeta de input. En este documento nos aparece el glosario de cada una de las variables categóricas que tenemos en la base de datos y gracias a este podemos relacionarlo con nuestra base de datos para poder realizar las transformaciones, a este glosario lo llamaremos “diccionario”.

El diccionario ofrecía el código de cada variable con su respectiva explicación, en algunos casos esta era demasiado extensa y para poder facilitar la interpretación de datos en un futuro se abreviaron algunas de las descripciones.

Mediante RStudio se hace la transformación de cada una de las variables con los diferentes casos que nos encontramos. Para poder hacer este proceso se hace lo siguiente:

Por ejemplo, con la variable categórica CCAADEST (Comunidad Autónoma de Destino principal):

1. Cargamos el excel que contiene la información sobre las comunidades autónomas.
2. Seleccionamos las filas exclusivas de la tabla del diccionario y eliminamos aquellas vacías.
3. Como en el "diccionario" los números aparecen con dos dígitos (00,01,02…) convertimos los datos de CCAADEST en la en números de dos dígitos para así poder hacer la transformación, si no lo hiciéramos RStudio no consideraría que sean los mismos números.
4. Hacemos la transformación de datos realizando un match entre el diccionario y la base de datos en la variable escogida.

Cuando hemos acabado este proceso vamos a ver en la base de datos que se han cambiado los números por los nombres de las comunidades autónomas correspondientes.

Otro caso particular en este proceso han sido algunas variables categóricas en las que tenían un nombre similar y no tenían "diccionario" pero se tenía que cambiar los números uno y dos por sí y no.

Un ejemplo de este caso son las variables VIAJA_XXX, en estas encontrábamos diferentes combinaciones de con que personas viaja y con quienes no. Para realizar esta transformación hacemos lo siguiente:

1. Hacemos que RStudio busque todas las columnas que empiecen con VIAJA_
2. Realizamos un bucle para que intercambie todos los valores uno por SI y todos los valores dos por NO de aquellas columnas que ha encontrado con la condición de que le hemos puesto previamente.

Con estos dos procesos se transformaron todas las variables categóricas en texto incluso aquellas que no tenían un diccionario propio.

### 3.3 Imputar los datos faltantes

Para poder desarrollar este punto primero hay que instalar y cargar dos paquetes de Rstudio; *mice* y *VIM*. Después cargamos la base de datos extraída del apartado anterior.

Para poder imputar los datos faltantes primero hay que dividir la base de datos entre las variables categóricas y las numéricas ya que para la imputación se tratan de maneras distintas. Es por esa razón que creo dos vectores en RStudio que agrupen estos dos grupos.

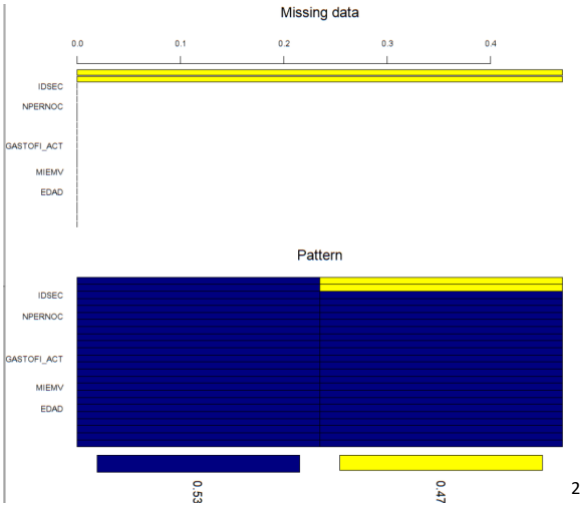
Empezamos con las variables categóricas y lo que hacemos es crear un bucle en el que Rstudio nos busque los valores con "NA" y nos los sustituya por "No responde".

Una vez acabado con las categóricas, creamos una copia de los datos sin estas variables la cual quedara formada únicamente por las numéricas.

Con este conjunto de datos buscaremos con la función "*md.pattern()"* el patrón de datos faltantes en el conjunto. Con esto encontramos que tenemos dos variables con NA que son grado de satisfacción y fidelidad al destino.

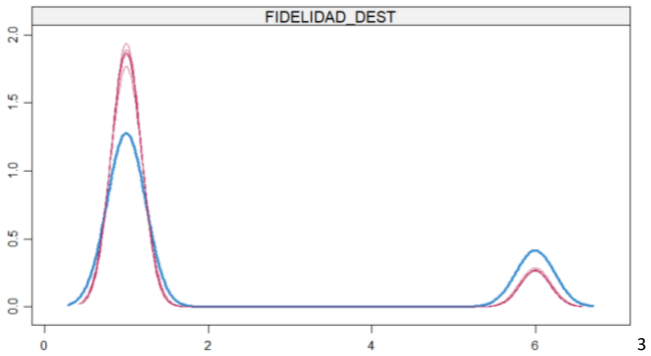
Para comprobar esto que hemos visto anteriormente lo vamos a graficar, entonces con la función "*aggr()"* creamos un gráfico de barras que muestra el porcentaje de datos faltantes para las variables de mi conjunto, así como sus patrones. Los patrones aparecen en el eje y, mientras que la variable correspondiente se muestra en el eje x. Junto al grafico que nos aparece también nos da el porcentaje de datos faltantes por variables, dando como resultado que tanto grado de satisfacción como fidelidad al destino tienen un 47% de datos faltantes.

En el grafico que extraemos podemos corroborar los datos que nos ha dado la consola de RStudio.




[2] Gráfico extraído de RStudio

Para solucionar este problema se utiliza la función "*mice()"* para imputar los valores perdidos en los datos numéricos. Esta se realiza utilizando el método de imputación múltiple que se ejecuta cinco veces con un máximo de 50 interacciones. Una vez acabado guardamos el resultado y creamos un gráfico de densidad que muestra la distribución de los valores imputados en cada variable. En rojo se pueden ver los datos imputados y en azul los datos previamente existentes




Para acabar reemplazamos los valores faltantes con los valores imputados con la función "*complete()"* y juntamos las variables categóricas con las numéricas imputadas.

De esta manera se crea la base de datos imputada con la que podemos empezar a trabajar en el análisis de datos y obtener las predicciones.

### 3.4 Visualización de los datos

Para poder visualizar los datos utilizaremos RStudio y los graficaremos de diversas maneras para obtener más información de nuestros datos.

Antes de empezar a graficar nuestras variables cargaremos los paquetes *ggplot2* y *psych.* La primera de estas es una librería para la visualización de datos que ofrece la opción de crear diversos gráficos personalizados de manera fácil y rápida. La segunda librería es para el análisis psicométrico y estadístico ofrece una amplia variedad de funciones para el análisis de datos.

Una vez tenemos las librerías, cargamos los datos ya imputados y los parametrizamos en categóricos y numéricos.

***Univariante***

Con estos datos empezamos el estudio descriptivo univariante de los mismos. Un estudio descriptivo univariante es una técnica de análisis estadístico que se utiliza para resumir y describir las características de una sola variable.

[3] Gráfico extraído de RStudio.

Para empezar, realizaremos este estudio con variables categóricas y las graficaremos en un diagrama de barras (*Barplot*). El *barplot* representa el número de efectivos por cada categoría, en el eje horizontal se encuentra la categoría de la variable y en el eje vertical las frecuencias.

En RStudio realizaremos un bucle para todas nuestras variables categóricas en la que se calcula la frecuencia relativa de cada categoría de la variable. Posteriormente con esas frecuencias las convertimos en porcentaje multiplicándolo por cien. Una vez tenemos estos datos generamos un gráfico de barras para cada variable categórica. Con estos datos los guardamos en nuestra carpeta de output en formato png para analizarla en el siguiente punto.

Una vez acabado con las categóricas procedemos a crear los gráficos descriptivos univariantes de las variables numéricas. Para este caso extraeremos dos graficas distintas, *boxplot* e histograma. Un diagrama de caja o *boxplot* representa los indicadores robustos y los *outliers* de la variable, está constituido por una caja delimitada por el cuantil 1 y 3, y con una línea interior que representa la mediana. Tiene unos bigotes y por encima o por debajo de estos bigotes encontramos los *outliers* que son esos puntos que podemos considerar anómalos. Un histograma muestra la distribución de la variable en barras, la superficie de cada barra es proporcional a la frecuencia de los valores representados. En el eje horizontal están las variables y en el eje vertical las proporciones.

Para realizar esto en RStudio creamos un bucle con las variables numéricas de nuestra base de datos y creamos un resumen estadístico de cada una. Después se utiliza *ggplot2* para crear un gráfico de caja de la variable numérica con la función "*geom_boxplot()"* y guardamos el grafico en png en la carpeta correspondiente. Posteriormente y dentro del mismo bucle generamos un histograma con la función "*geom_histogram()"* y lo guardamos en la misma carpeta que el *boxplot* creado anteriormente.

***Bivariante***

Un estudio descriptivo bivariante permite analizar la relación entre dos variables y puede proporcionar información importante para entender el comportamiento y las relaciones de campos.

Nuestro estudio bivariante lo realizaremos todo el rato contra la variable importe gasto total ya que es la variable a estudiar en este trabajo. Con esto se busca ver cómo afecta y que relación tienen el resto de las variables tanto categóricas como numéricas al gasto total.

Primero realizaremos las variables categóricas contra nuestra variable respuesta, gasto total, la cual es una variable numérica. Lo graficaremos mediamente un diagrama de barras bivariante. Este presenta las mismas características que el univariante con la diferencia de que en el diagrama de barras en se encuentran las dos variables superpuestas.

La ejecución en RStudio consiste en realizar un bucle en el que se hace lo siguiente, se genera un archivo que contiene el resumen estadístico de la variable respuesta desglosado por cada nivel de la variable categórica. Posteriormente a la creación de este resumen se crea un diagrama

de barras que le incorporamos los colores azul oscuro y claro para visualizar la distribución del gasto total en cada nivel de las categóricas, finalmente guardamos dicho grafico para analizarlo.

El siguiente estudio bivariante que vamos a ejecutar es uno entre nuestras variables numéricas contra el gasto total, exceptuando la propia variable respuesta. El grafico que realizaremos para analizar es un diagrama de dispersión (*scatterplot*) este es una representación de puntos los cuales muestran una relación entre ambas variables, en el eje horizontal encontramos la variable explicativa y en el eje vertical de la variable respuesta.

Para realizar el diagrama de dispersión en RStudio primero se crea un bucle en el que el código comprueba si la variable actual del vector de numéricas es gasto total, si no es así continua con el análisis. Se genera un resumen que contiene el coeficiente de correlación entre la variable actual y la respuesta y posteriormente crea un gráfico de dispersión con estos datos. Este *scatterplot* se ha creado con las funciones "*geom_point()"* y "*geom_smooth()"* las cuales agregan puntos en el gráfico y una línea de tendencia con el método de regresión lineal. Finalmente se guarda el grafico para analizarlo.

Para saber aquellos datos que son más relevantes también extraemos la correlación completa de las variables numéricas entre ellas para luego fijarnos en la correlación con el gasto total para así saber aquellas que influyen más con esta.

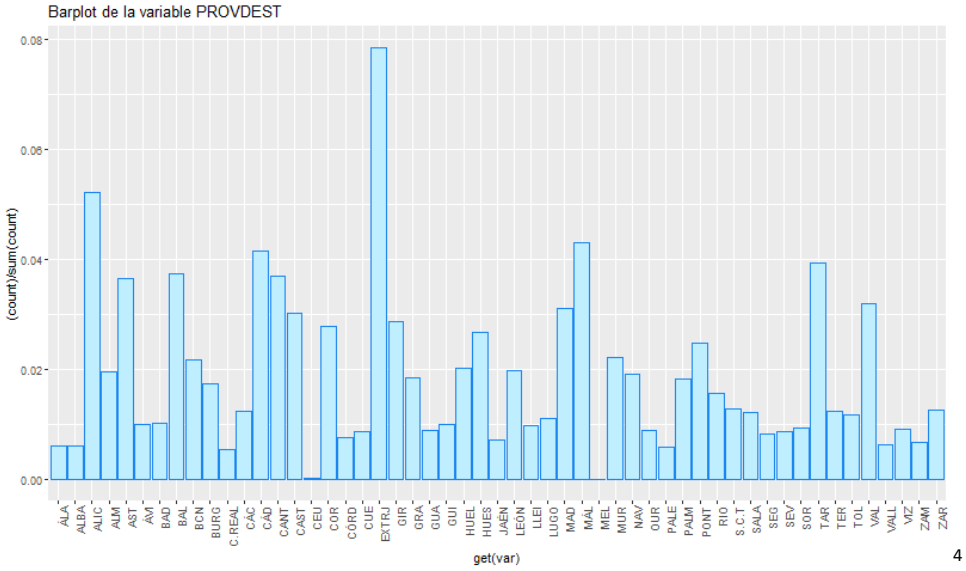
### 3.5 Análisis de gráficos

En este apartado se van a comentar los gráficos más relevantes en base a los datos obtenidos de la correlación entre las variables con el gasto total. El resto de graficas se encuentran en el anexo del trabajo.

***<u>Gráficos Univariantes</u>***

**Variables categóricas:**

- Provincia de destino



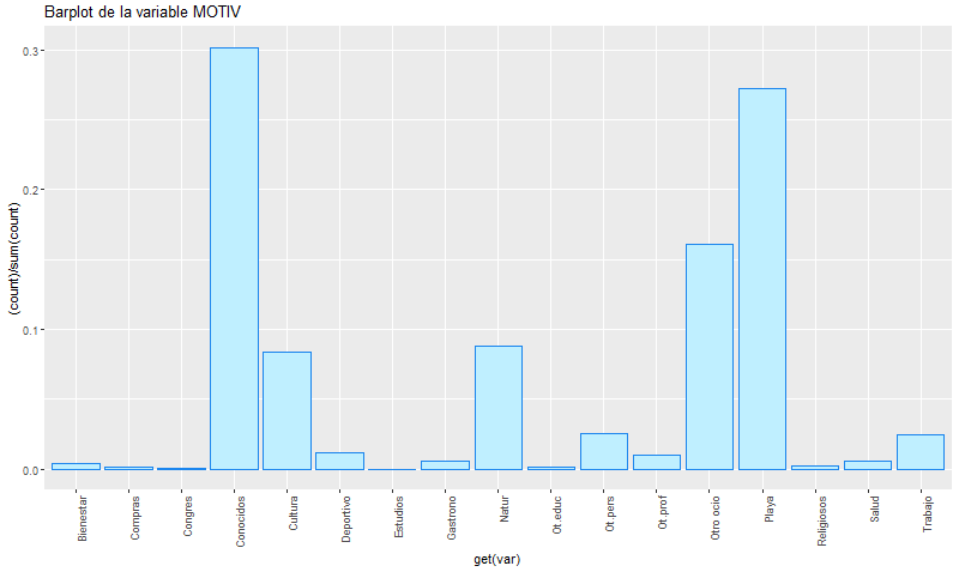
[4]

Una de las variables más importantes en cuanto al turismo es el lugar de destino al que se va. Se puede apreciar que la localización más visitada, con diferencia, es el extranjero, acercándose al 8% del total. Hay que matizar que en el extranjero se contemplan muchos lugares distintos. La provincia de España más visitada es Alicante, con un 5,2%. Seguida de esta provincia, podemos observar que el resto también son provincias costeras con playa: Málaga, con un 4,3%; Cádiz, con un 4,2%; y Tarragona, con un 4%. Con estos datos, podríamos afirmar que el turismo que predomina es el de las provincias costeras y, por ende, el de playa.

Por el otro lado las provincias que reciben menos turismo son Ceuta y Melilla que no llegan ni al 0,1%.

[4] Gráfico extraído de RStudio

- Motivo principal del viaje



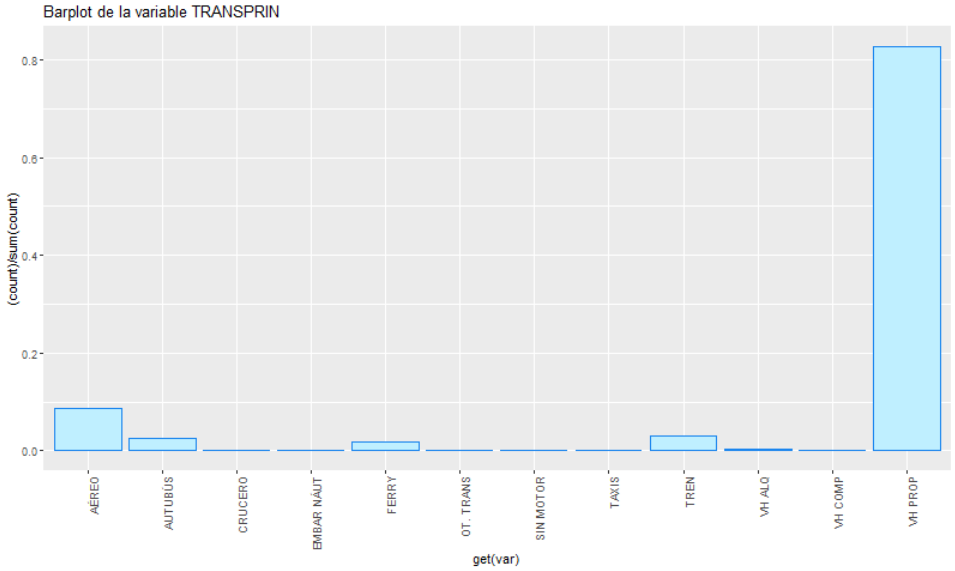
[5]

Para conocer bien los datos del turismo, es primordial conocer las causas de dicho movimiento. Como podemos observar en la gráfica, la principal razón de desplazamiento es para visitar a amigos y familiares, con un 30%. Por detrás de las visitas a conocidos, podemos ver que le sigue de cerca el turismo de sol y playa, con un 27,2%. Luego encontramos otras actividades de ocio más allá de las que aparecen en la encuesta, con un 16%. Estos tres turismos diferentes acumulan más del 70% del total de motivos.

Cabe destacar también el turismo de naturaleza y cultural, que son de los más importantes y destacan más en este estudio, con un 8,8% y un 8,4% respectivamente.

[5] Gráfico extraído de RStudio

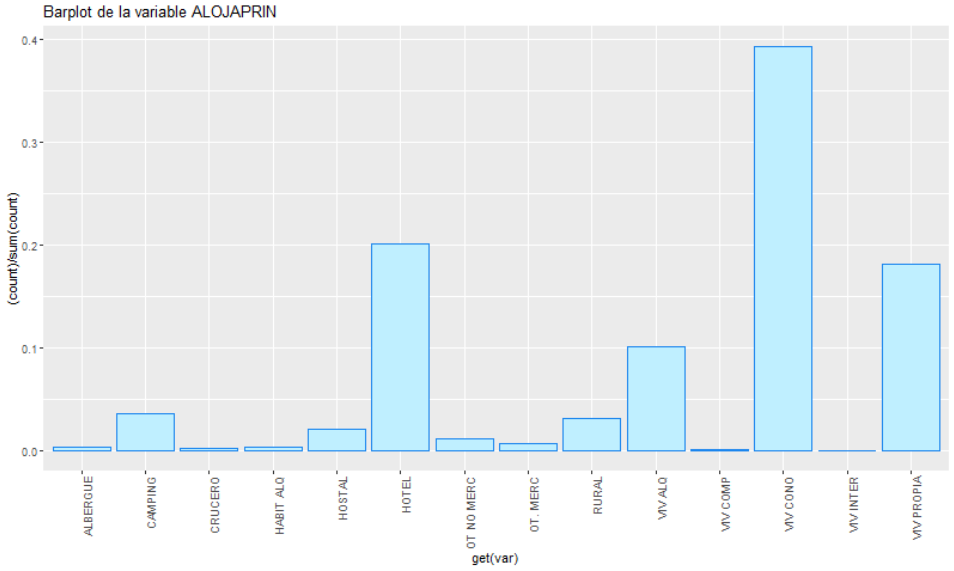
- Transporte principal




Un aspecto fundamental a tener en cuenta con el gasto total es el transporte, por esa razón es importante saber cuál es el medio de transporte principal del turista. Como podemos observar, más del 80% de los encuestados utilizan su vehículo propio como medio de transporte para viajar, estando muy por delante de los demás. En segundo lugar, encontramos el transporte aéreo con un 8,6%, el cual conlleva unos gastos considerablemente superiores al anterior. El resto de los medios de transporte, a excepción del tren y el autobús, tienen un uso bastante reducido en comparación con los mencionados anteriormente.

[6] Gráfico extraído de RStudio

- Alojamiento principal:



[7]

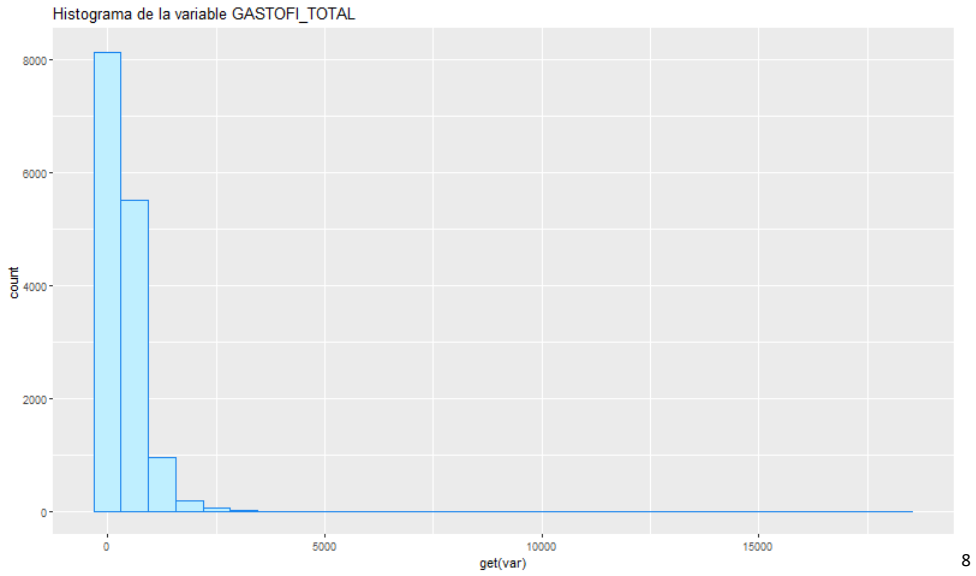
El alojamiento principal está condicionado por el motivo del viaje, por lo tanto, resulta interesante analizar este gráfico teniendo en cuenta los datos anteriores. Podemos observar que el 39,3% de los turistas se hospedan en la vivienda de algún conocido, lo cual está estrechamente relacionado con el hecho de que el principal motivo del turismo sea visitar a familiares y amigos.

En segundo lugar, después de las viviendas de familiares o amigos, se encuentra el hotel con un 20% de elección como alojamiento principal. Le sigue en tercer lugar la opción de viviendas propias o segundas residencias, con un 18%. En este gráfico, podemos notar que estas tres variables mencionadas representan más del 75% del total de alojamientos principales elegidos por los turistas.

[7] Gráfico extraído de RStudio

**Variables numéricas:**

- Gasto total

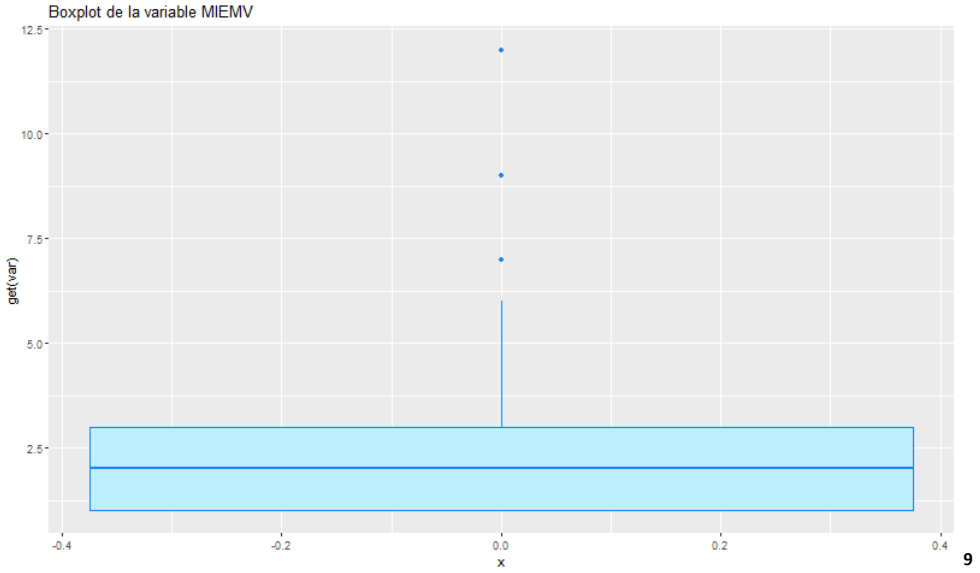



[8]

Nuestra variable a estudiar es el gasto total, como se puede observar la media del gasto total del turismo es unos 400€. La mayoría de encuestado se encuentran gastando menos de 500€ por viaje.

[8] Gráfico extraído de RStudio

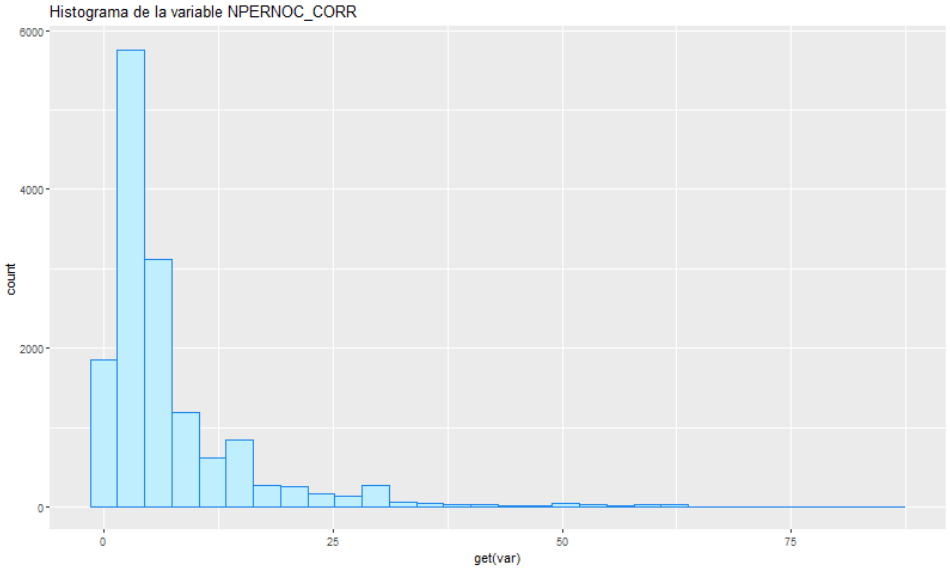
- Número de miembros que participan en el viaje.




Es importante saber cuántos miembros hay en cada viaje ya que en función del número de miembros los gastos variaran mucho, como podemos observar en este *boxplot* el número medio de personas que viajan es dos y hasta seis personas no es extraño. A partir de tres personas ya son valores que consideramos atípicos.

[9] Gráfico extraído de RStudio

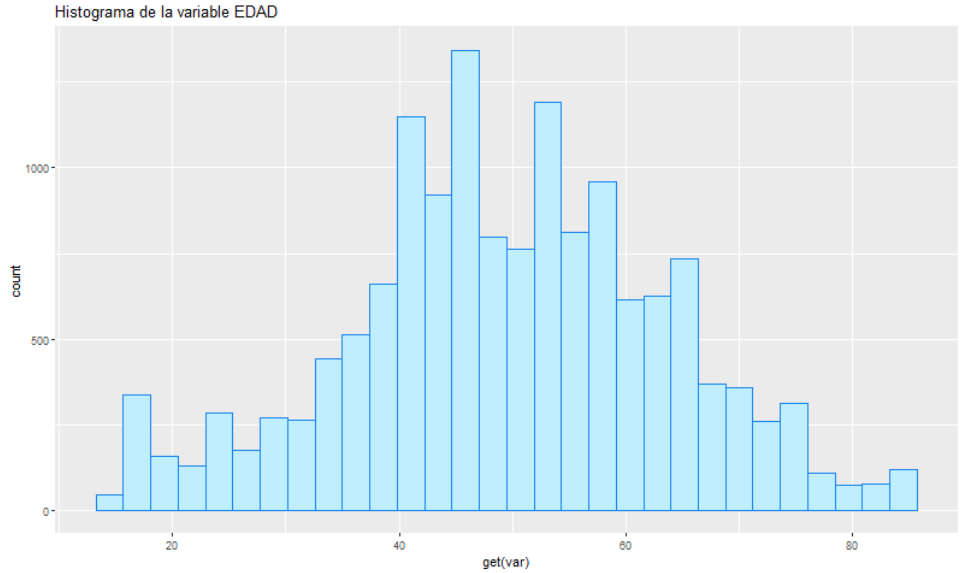
- Número de pernoctaciones corregido




Uno de los factores más determinantes para el gasto total es el coste en el alojamiento, este coste está muy relacionado con el número de pernoctaciones, como podemos observar en el histograma la mayoría de los turistas no pasan más de cinco noches fuera de su hogar.

[10] Gráfico extraído de RStudio

- Edad




[11]

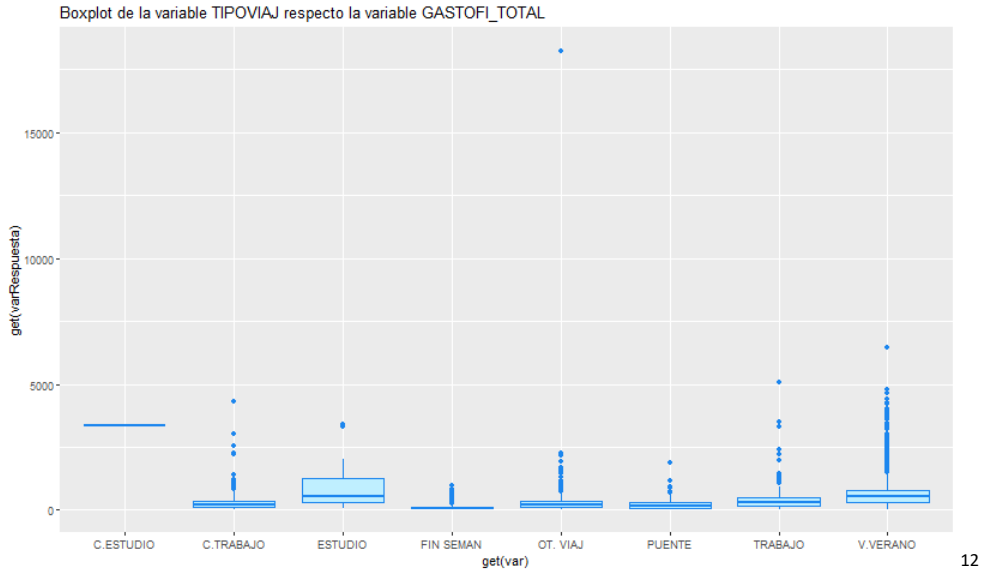
En cuanto a la edad de las personas que viajan nos encontramos que los rangos de edad más frecuentes son entre los cuarenta y cuarenta y ocho años con el pico más alto seguido de los cincuenta y dos y los cincuenta y ocho. Estas son personas con un poder adquisitivo mayor al de los jóvenes y eso hace que les sea más sencillo viajar.

[11] Gráfico extraído de RStudio

## ***Gráficos bivariantes***

### Variables categóricas contra el gasto total

- Tipo de viaje contra gasto total



[12]

En cuanto al tipo de viaje, se puede observar que el centro de estudios es aquel que supone mayores gastos, sin embargo, dado que solo contamos con un valor para esta variable, no lo consideramos representativo. De manera similar, la variable de estudio, con solo veintiún respuestas, tampoco la consideramos significativa.

Excluyendo las variables mencionadas anteriormente, se destaca que las vacaciones de verano tienen la mayor media de gasto, lo cual es lógico debido a que este tipo de viaje suele ser de mayor duración. También se encuentran los viajes por motivos de trabajo, que ocupan el segundo lugar en cuanto a gastos. Además, es importante mencionar que las vacaciones de verano presentan algunos valores atípicos más altos que el resto de las categorías.

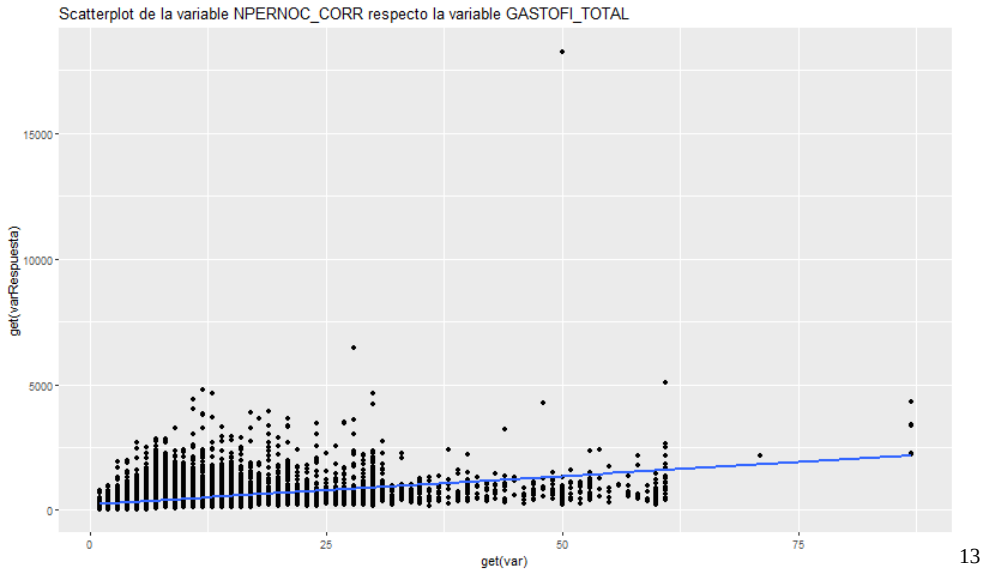
[12] Gráfico extraído de RStudio

### Variable numérica contra el gasto total

Para analizar estos gráficos primero vamos a analizar la correlación de las variables numéricas fijándonos en nuestra variable respuesta, de esto podemos extraer que la variable numérica más influyentes en el gasto total son el número de pernoctaciones, con una correlación positiva del 45,92%. El resto de las variables muestra una relación muy baja con la variable respuesta.

Estas correlaciones que hemos mencionado tienen bastante sentido, ya que a medida que aumenta el número de noches, el gasto tiende a aumentar.

- Número de pernoctaciones contra gasto total.



[13]

Como podemos observar la correlación entre las dos variables es positiva, lo que significa que a medida que aumenta el número de pernoctaciones, también aumenta el gasto total. Esta correlación positiva está representada por una línea de tendencia, la cual indica que el 46% de la variación en el gasto total puede explicarse por el número de noches.

La mayoría de los puntos se concentran en la parte inferior izquierda del gráfico, lo que indica que hay una gran cantidad de observaciones donde el gasto total es bajo y las pernoctaciones lo es también. Sin embargo, a medida que se avanza en el eje horizontal, también aumenta la dispersión de los datos en el eje vertical, es decir, el gasto total.

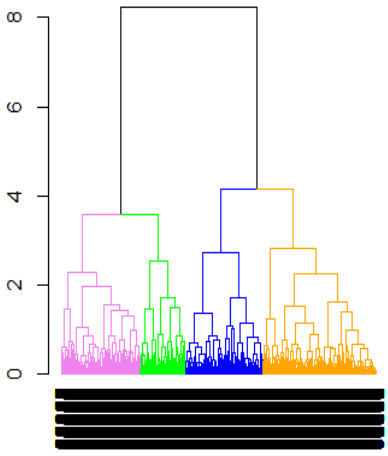
[13] Gráfica extraída de RStudio

### 3.6 *Clustering*

Para llevar a cabo este análisis necesitamos instalar las siguientes librerías; "*cluster"* esta nos proporciona funciones para el análisis de clustering, incluyendo cálculos de distancia y clustering jerárquico. "*NbClust",* este se utiliza para determinar el número óptimo de *clusters* en un conjunto de datos utilizando varios índices de clustering. "*Dendextend"* con esta librería ampliamos la funcionalidad, manipulación y visualización de los dendograma. Por último, también es necesario *"ggplot2"* para que nos ayude a crear gráficos.

Para poder crear el clustering de datos primero se convierte las variables categóricas en factores y se calcula una matriz de disimilitud utilizando la distancia de Grower. Luego se realiza el clustering jerárquico mediante la función *"hclust()"* utilizando el método de enlace completo. Cuando se grafica este dendograma se visualiza que la mejor forma de separar los *clusters* es en cuatro partes ya que en dos partes se estaría siendo poco preciso.

En base al dendograma extraído previamente se corta este en cuatro *clusters* y se le asignan colores para poder diferenciarlo.




[14]

Una vez dividido procedemos a separar las variables categóricas de las numéricas para poder desarrollar un mejor análisis de los grupos elaborados por el *clustering*.

Para las variables categóricas se realiza un análisis descriptivo bivariado agrupado por *clusters* para cada variable categórica. Se generan tablas que muestran la distribución de las categorías en cada *cluster*. Luego, se realiza una prueba de chi-cuadrado para determinar si la variable

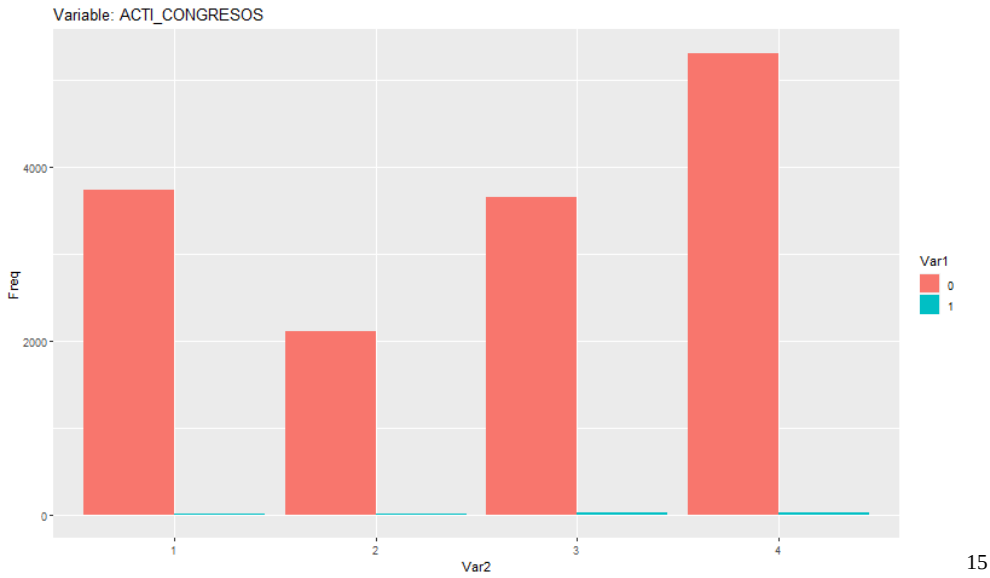
[14] Gráfico extraído de RStudio

categórica varía significativamente entre los *clusters*. También se generan gráficos de barras bivariados que muestran la relación entre las variables categóricas y los *clusters*.

Para las variables numéricas Se realiza un análisis descriptivo bivariado agrupado por *clusters* para cada variable numérica. Se utiliza la función *"describeBy()"* del paquete *"psych"* para obtener estadísticas descriptivas para cada *cluster*. Luego, se realiza una prueba de ANOVA para determinar si la variable numérica varía significativamente entre los *clusters*. También se generan gráficos de que visualizan la distribución de las variables numéricas para cada *cluster*.

En este análisis descartaremos algunas gráficas ya que consideramos que no nos dan información suficiente para sacar conclusiones del *clustering* realizado, un ejemplo de estas es la siguiente.

- Gráfico *clustering* Actividad Congresos

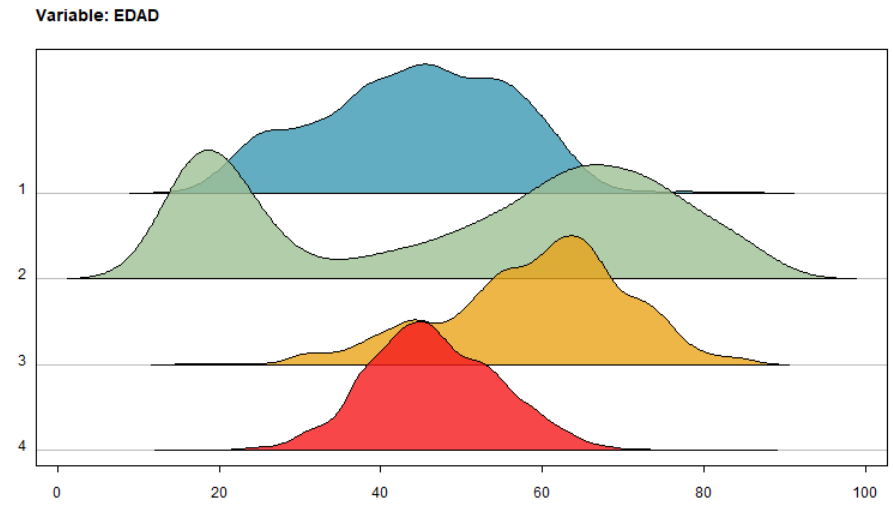



15

Por otro lado, sí que encontramos diversas graficas que podemos relacionar entre ellas para desarrollar información útil para el estudio. A continuación, comentaremos alguna de estas para desarrollar perfiles de clientes. El resto de las gráficas se encuentran en el anexo.

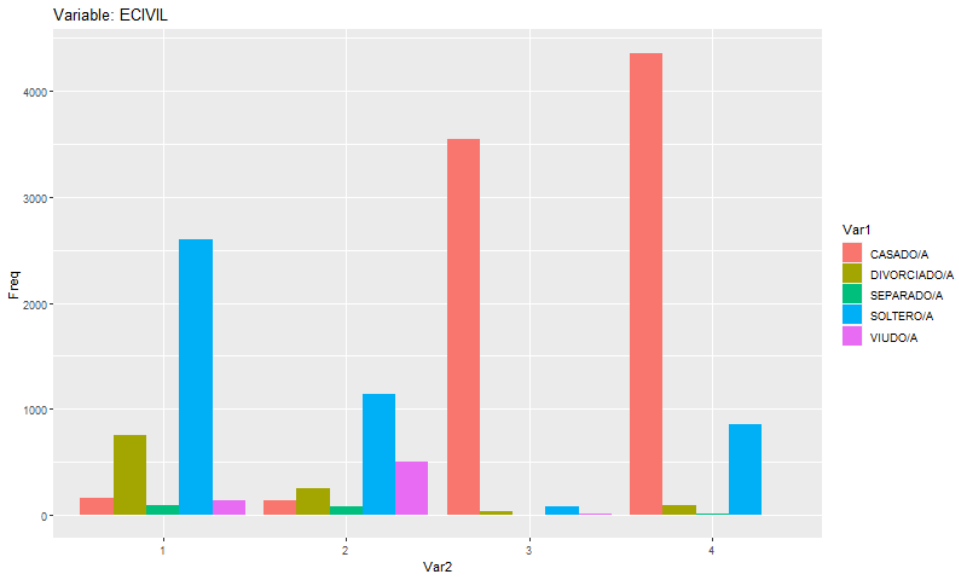
[15] Gráfico extraído de RStudio

- Gráfica *clustering* Edad




[16]

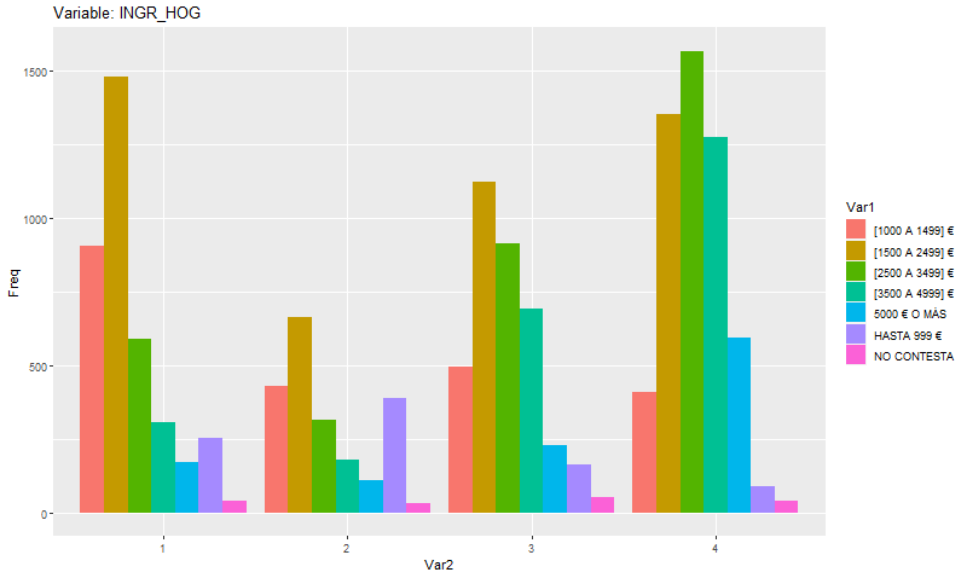
- Gráfica *clustering* Estado Civil




[17]

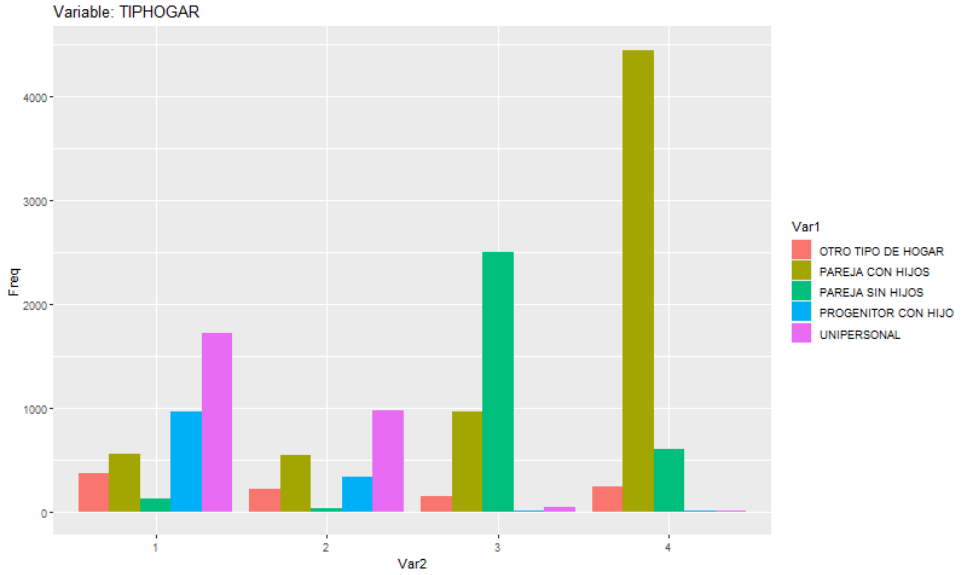
[16] Gráfica extraída de RStudio
[17] Gráfica extraída de RStudio

- Gráfico *clustering* ingresos en el hogar




[18]

- Gráfico *clustering* Tipo de Hogar




[19]

Se va a realizar el análisis por cada parte del *clustering*:

[18] Gráfico extraído de RStudio
[19] Gráfico extraído de RStudio

Primer *cluster:* Se puede apreciar que son personas entre 20 y 65 años, un rango de edad muy elevado, donde la mayoría son solteros o divorciados. Los ingresos por hogar que predominan en estas personas son los más bajos, entre 1.000€ a 2.499€, hecho que tiene bastante concordancia con el hecho que el tipo de hogar es unipersonal, por tanto, solo hay un sueldo.

Segundo *cluster:* Aquí podemos encontrar personas jóvenes de entre 15 y 25 años y también personas mayores de 50 años. Este grupo de personas la mayoría son solteros o viudos, hecho que también tiene sentido por el rango de edades y el tipo de hogar que tienen que es unipersonal en su mayoría. Los ingresos son bastante bajos, entre 0 y 2.499€

Tercer *cluster:* En este grupo encontramos las edades comprendidas entre 45 y 75 años, donde la mayoría están casados, pero tienen un hogar sin hijos, esto tiene bastante sentido ya que puede venir derivado de que en esas franjas de edad los hijos ya se hayan independizado. Por otro lado, los ingresos del hogar están entre 1.500€ a 4.999€.

Cuarto *cluster:* Aquí encontramos a personas adultas de entre 30 y 60 años las cuales están casadas y con hijos en el hogar, cabe destacar que es el grupo con mayores ingresos por hogar estando desde los 1.500€ hasta más de 5.000€.

Como se puede observar tenemos cuatro perfiles bien marcados que buscan llenar diferentes necesidades durante su proceso turístico, hecho que han de saber aprovechar bien las empresas para conocer cuáles son los perfiles de sus clientes.

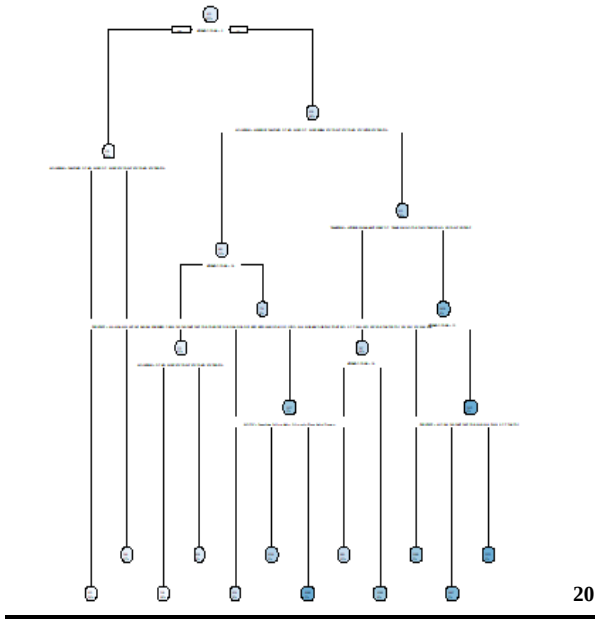
### 3.7 Modelo de árbol de decisión

Una vez se han realizado los análisis individuales y grupales de las variables podemos realizar diferentes modelos predictivos con distintos algoritmos.

Para realizar el modelo de árbol de decisión hemos de cargar las librerías; "*tidyverse"* esta es un conjunto de paquetes de r donde están incluidos *"tidyverse", "ggplot2"," dplyr", "tidyr" y "readr"* estos se utilizan para la transformación, limpieza y visualización de datos. Otra librería es "*rpart",* esta es necesaria ya que es la que proporciona funciones para el ajuste y visualización de modelos de árboles de regresión y clasificación utilizando el algoritmo CART. Para poder unificar y simplificar el proceso de entrenamiento y evaluación se instala también *"caret".*

Después de instalar las librerías se carga la base de datos y estos se dividen en un conjunto de prueba y otro de entrenamiento, en este caso la fracción de entrenamiento representara el ochenta por ciento de los datos, esto se lleva a cabo con la función "*sample_frac"* y "*setdiff",* este último simplemente es para encontrar el veinte por ciento de prueba haciendo la diferencia del total con el de entrenamiento.

Una vez dividida la base de datos con la función "*rpart()"* se entrena un modelo de árbol de regresión con la variable objetivo gasto total. Para poder mostrar este árbol se ejecuta *"rpart.plot()"*. Una vez está el árbol se realiza la función "*predict()"* con el set de prueba para generar un vector con los valores predichos por el modelo entrenado.




Como se puede observar el primer criterio de división se basa en la variable "*Npernoc_Corr*", si el valor de esta variable es menor que 4,5, el árbol se divide en el nodo 2, de lo contrario se divide en el 3.

Para aquellos valores menores a 4,5 pernoctaciones, tenemos 6.133 datos los cuales nos muestran que tienen un gasto total de 179,32€. La división de este nodo podemos ver que se divide principalmente en alojamiento principal de bajo coste, nodo 4, aquí se encuentra la vivienda de un conocido o un camping, entre otras, el coste es de 99,78€. Por el otro lado encontramos el nodo 5, en este están los alojamientos algo más costosos como podría ser un hotel o una vivienda de alquiler, esto se refleja en su gasto que asciende a 344,46€.

Por el lado del nodo 3 encontramos que para estancias de más de 4,5 noches el gasto es de 636,15€, dentro de este nodo podemos destacar varios puntos, el primero es que hay lugares de destinos en los que el coste es mayor que en otros. Otro punto a destacar es el tipo de transporte principal, en el nodo 15 donde se refiere al aéreo/crucero podemos ver como en esta ocasión el gasto es de 1.408,62€. Colgando de este nodo también se aprecia un dato relevante y es la diferencia entre las dos hojas terminales donde el destino es nacional, nodo 62, o el extranjero, nodo 63, en el segundo caso el coste es mucho mayor con 2.502,08€

[20] Gráfico extraído de RStudio

Antes de acabar se realizan varias predicciones y mediante la función *"confusionMatrix()"* y *"rmse()"* que compara los datos reales con los predichos podemos saber la precisión de nuestro modelo.

Nuestras predicciones han sido para diversos casos: 680,76 en el primero de ellos y 318,97 en el segundo. Una vez se tienen las predicciones de nuestro modelo se encuentra la desviación de este, esta es de 257,16€. Esto es bastante y nos hace ver que este método de previsión nos explica muy bien el modelo, pero es muy poco preciso.

Esta sobrestimación se puede producir porque como hemos visto en la parte teórica el árbol de decisión varía mucho dependiendo de la muestra escogida y suele sobre ajustar el modelo. Esto no nos interesa, realizar predicciones erróneas sobrestimando puede producir innumerables perdidas en previsión de ventas. Un empresario puede contratar a más personas pensando que el gasto del turista será mayor y eso le comportaría incurrir en unos gastos que no cubrirá con el dinero que dejen sus clientes.

### 3.8 Modelo de *Random Forest*

El *Random Forest* es un modelo predictivo que mejora los árboles de decisión, para poder realizar este modelo predictivo se necesita la librería especifica *"randomForest",* esta implementa un algoritmo *Random Forest*, es un método de aprendizaje automático utilizado para la clasificación y regresión, con este se puede obtener predicciones. Aparte de esta librería también será necesaria *ggplot2* para poder realizar unos gráficos de árbol óptimos y *tidyverse* por todas sus funcionalidades.

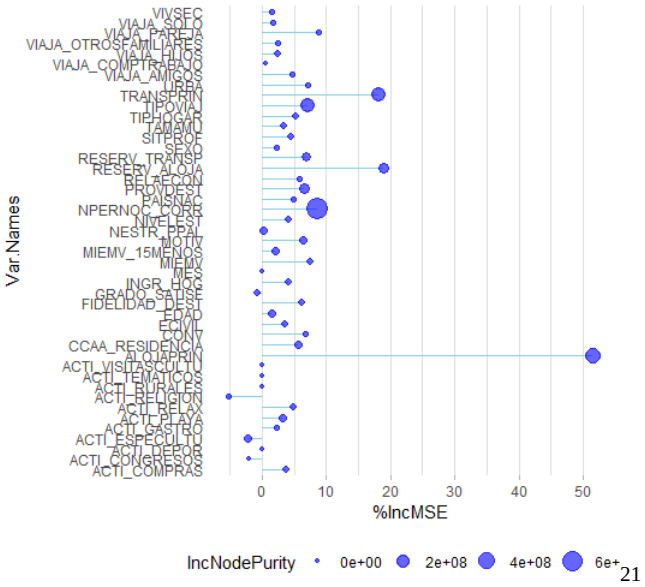
Una vez cargadas las librerías y los datos, se procede a realizar la separación de la base entre *Train* y *Test.* Estos estarán compuestos por un ochenta por ciento en la parte de entrenamiento y un veinte por ciento en la de prueba. Para hacer esto utilizamos la función "*sample_frac()"* y posteriormente por diferencia encontramos la parte de test.

Cuando ya se tiene la base de datos dividida se ejecuta la función *"randomForest()"* con la parte de entrenamiento, la variable objetivo que pretendemos buscar con este modelo es el gasto total y le solicitamos al RStudio que se creen quinientos árboles guardando la información importante de las variables.

Una vez realizado el *Random Forest* podemos ver que con este método conseguimos explicar el 69,8% de nuestro modelo, este porcentaje es bastante significativo.

Como se han guardado dicha información con la función "*importance()"* extraemos estas variables importantes y con la función *"ggplot()"* graficamos dichos parámetros. De esta manera podemos ver que variables explican mejor nuestro objetivo principal que es el gasto total.

- *Random Forest* contra gasto total



[21]

Como podemos apreciar en el gráfico la variable que más afecta positivamente al gasto total del turismo es el alojamiento principal, esta parece ser la variable más relevante seguida de la propia reserva del alojamiento, del transporte principal y del número de pernoctaciones. Esto tiene mucho sentido ya que en el análisis anterior de los árboles de decisión ya pudimos apreciar el número de noches afecta directamente en el gasto y que el coste en transporte es muy relevante, un viaje en avión hace que el gasto sea mucho mayor. Por el otro lado se puede ver que negativamente las actividades religiosas y culturales.

Una vez tenemos ya nuestro modelo podemos realizar predicciones y encontrar sus errores. Para el grupo de predicciones de entrenamiento vemos una desviación de 121,35€, esta es elevada, pero mejora al modelo anterior. Por lo que se refiere a las variables para el grupo de prueba nos encontramos una desviación mayor, 202,89€. Con esto podemos concluir que nuestro modelo de predicción sobre estima los resultados, la diferencia entre el grupo de entrenamiento y el de prueba es de 81,54€, hecho que nos lleva afirmar esta sobrestimación.

Con este modelo hemos podido clasificar aquellas variables más relevantes de nuestro modelo y pese a mejorar la precisión del modelo anterior sigue siendo deficiente.

[21] Gráfico extraído de RStudio

### 3.9 Modelo predictivo *XGBoost*

Para ejecutar el modelo *XGBoost* primero tenemos que cargar las librerías necesarias que están relacionadas con este modelo predictivo, estas son: "*xgboost", "tidyverse", "caret" y "fastDummies".* La primera librería nos proporciona un conjunto de funciones y métodos para entrenar y ajustar modelos de *Gradient Boosting.* La librería "*caret"* ofrece una serie de algoritmos para ayudar en la selección de variables y su validación. La última librería es utilizada para la creación de variables *dummy* a partir de variables categóricas en un conjunto de datos. Las variables *dummy* son variables binarias que se utilizan para representar las categorías de una variable categórica en un modelo predictivo.

Después de cargar las librerías y nuestra base de datos, eliminaremos la variable meses porque es la misma para todo el modelo. Para poder llevar a cabo el método de previsión necesitamos transformar las variables categóricas en variables *dummy*, esto se realiza mediante la función *"dummy_cols()".*

Para continuar con el análisis se realiza una división de los datos en conjuntos de entrenamiento, ochenta por ciento, y prueba veinte por ciento.

Una vez tenemos la división hecha se convierte ambos conjuntos en objetos de *DMatrix* con la función *"xgb.DMatrix()"*.

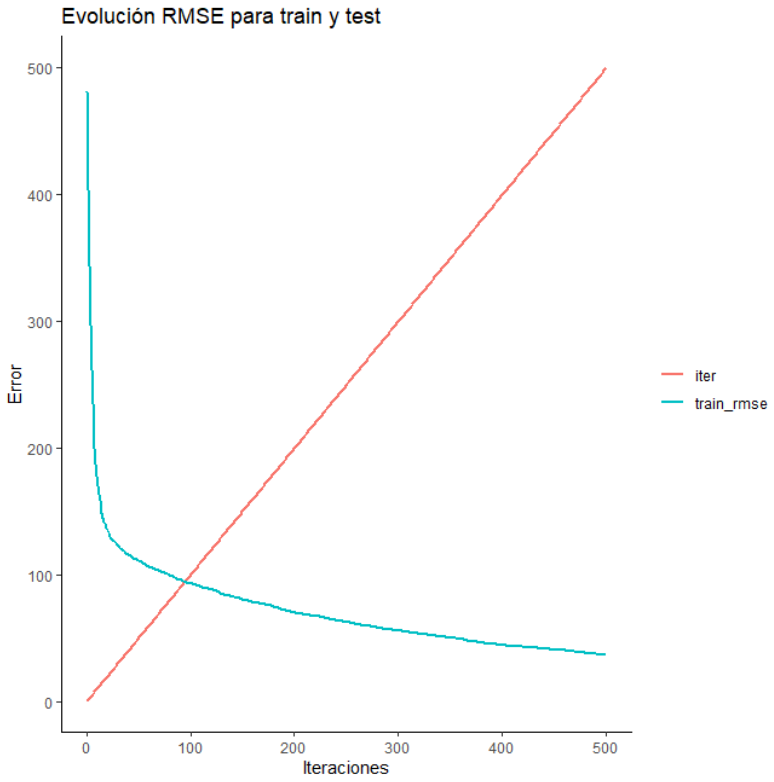
Luego mediante las funciones del paquete *"xgboost"* se escogen los mejores parámetros para el modelo, estos son escogidos por el propio algoritmo del modelo ya que se encarga de buscar los parámetros que encuentra mejores para predecir el modelo, esto también de le conoce como validaciones cruzadas. Una vez tenemos los parámetros óptimos entrenamos nuestro modelo y elaboramos predicciones.

Una vez hemos generado nuestras predicciones para los grupos de entrenamiento y test procedemos a ver que errores nos han generado. Para el caso de entrenamiento tenemos una desviación de 36,74€, el error más bajo que hemos obtenido de cualquier otro modelo, por otro lado, encontramos que para el grupo de prueba se obtiene un error cuadrático medio de 365€, el más alto de todos los modelos.

Esto nos lleva a la conclusión que el modelo esta sobre ajustado, nuestro método de previsión funciona correctamente para encontrar datos ya conocidos, pero no es capaz de predecir correctamente aquellos datos no vistos, no acaba de encontrar los parámetros correctos.

Para ver mejor esta sobre estimación se realiza una curva *logloss* con *train* y *test.*

- Curva *Logloss train y test*



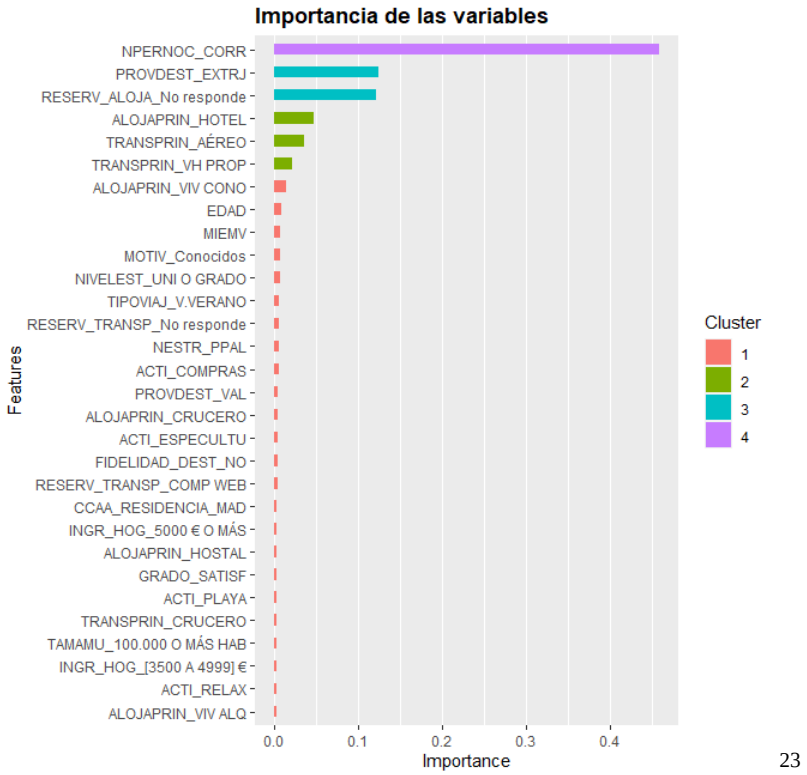

[22]

Como se puede apreciar perfectamente en la imagen el modelo esta sobre ajustado, a medida que aumentan las interacciones el error cuadrático medio del grupo de prueba aumenta, por otro lado, en el de entrenamiento podemos apreciar como va disminuyendo. Es decir, siguen pendientes opuestas y eso hace que nuestro modelo no pueda ser valido.

En el método *XGBoost* hemos podido extraer y averiguar que variables ha considerado más importantes para el análisis:

[22] Gráfico extraído de RStudio

- Importancia de las variables *XGBoost:*



23

Tal y como muestra el gráfico la variable más importante es el número de pernoctaciones, luego encontramos en el segundo clúster el viajar al extranjero y en el tercer bloque se encuentran el tipo de transporte, tanto propio como aéreo, y el alojamiento en hotel.

[23] Gráfico extraído de RStudio

# 4. Conclusiones

Para concluir se va a recoger toda la información importante elaborada en los diferentes puntos de estudios y análisis.

Para poder entender el turismo, es relevante saber aquellos factores clave que lo caracterizan y que contribuyen a su desarrollo. A través de los análisis univariantes y bivariantes hemos podido desarrollar diferentes conclusiones, los turistas prefieren destinos costeros para realizar turismo de playa, lo cual prevalece como una de las tendencias dominantes en el sector.

Por otro lado, también hemos visto que el alojamiento y el transporte principales aquellos hechos que más influencian en el gasto total. En estos análisis la mayoría de las personas utilizada su propio vehículo y se hospeda en casa de conocidos. Esta elección es lógica teniendo en cuenta que es el principal motivo de los viajes.

Además, en este estudio se ha podido elaborar perfiles de turistas bastante diferenciados gracias al proceso de *clustering*, esto es muy relevante ya que las empresas que viven del turismo necesitan saber hacia qué tipo de cliente quieren orientar sus servicios.

Estos perfiles los hemos dividido en cuatro tipos de clientes que son parecidos entre ellos pero diferentes al resto.

El primer perfil de cliente que nos podemos encontrar es gente que vive sola por diferentes motivos, como estar soltero o divorciado. Estos disponen de unos ingresos bajos, entre 1.000 y 2.500€. Este grupo se mueve especialmente por ver a conocidos y por ir a la playa.

El segundo perfil de cliente es gente con un nivel de ingresos bastante bajo y viajan menos que el resto de los perfiles. Estas son personas muy jóvenes que están estudiando y aún no tienen un trabajo, y también encontramos a las personas ya jubiladas, que también presentan unos ingresos bajos y ya no están trabajando.

El tercer perfil: aquí encontramos personas sin hijos con pareja en una edad más adulta. Esto hace que tengan unos ingresos por hogar más altos que los dos perfiles anteriores. El motivo del turismo es principalmente para ver a conocidos, ir a la playa y otras actividades de ocio.

El cuarto y último perfil de cliente son parejas con hijos con un ingreso superior al resto de perfiles. En este caso, el mayor motivo del turismo es ir a la playa. Hay que destacar que, al ser grupos de familias con hijos, sus viajes están enfocados a lugares con niños y donde puedan estar todos.

Conociendo estos perfiles, cada empresa, en función de sus servicios, ha de entender a qué público se está enfocando o se quiere enfocar. Un lugar de ocio nocturno no puede pretender enfocar su producto a perfil de clientes de familias con hijos ya que debido a sus características no van a buscar ese producto ni realizaran mucho gasto. Por otro lado, si es un hotel con

actividades para niños si que puede centrarse en ese perfil y no tanto en el de personas que viven solas.

En España el turismo es uno de los sectores más importantes y relevantes de su economía, por lo que es imprescindible saber que variables son las que proporcionan mayor ingreso en el país, es decir mayor gasto de los turistas, esto es importante para así poder explotarlas más o bien desarrollar otras y no depender tanto de unos pocos factores.

Gracias a nuestros métodos de predicción hemos podido extraer aquellas variables más relevantes en el momento del gasto, estas son: el número de pernoctaciones, el alojamiento y el transporte principales.

Como se pudo ver en el árbol de decisión, el primer nodo era el número de pernoctaciones. Esta es una variable muy importante y es sencillo de entender: cuanto más tiempo estás haciendo turismo, más gasto se realiza. En ese mismo método, también pudimos apreciar que el tipo de alojamiento era algo que condicionaba bastante en los nodos. En función de la categoría de este, ocurría lo mismo con el tipo de transporte. El viajar en avión o crucero era determinante.

Siguiendo con los otros métodos de previsión realizados, *random forest* y *XGBoost*, también nos especifican que las variables que han detectado como más relevantes son el alojamiento principal, sobre todo en el método de *random forest*, el transporte principal y el número de noches, en especial en *XGBoost*.

Con estos datos, podemos concluir que las empresas encargadas del alojamiento y del transporte, principalmente aéreo y marítimo, pueden obtener más beneficios procedentes del turismo. Esta información también es interesante en términos de hacer paquetes. Un hotel es consciente de que el mayor gasto del cliente será el hospedaje en ese lugar. Por tanto, puede estar dispuesto a asumir algo más de precio por pequeños servicios, como por ejemplo el transporte al hotel. De esta forma, reduciría una parte de su gasto.

Para finalizar, tenemos las predicciones de nuestros modelos, en vista a los resultados de este hemos concluido que ninguno de los tres modelos escogidos ha sido capaz de predecir con suficiente precisión. Todos los modelos presentan errores muy elevados y acaban sobreestimando el modelo, hecho que perjudica gravemente el análisis. Lo que podemos concluir de estas sobre estimaciones es que nuestra base de datos era, seguramente, demasiado corta en cuanto a su temporalidad y eso nos ha provocado estos errores.

### 4.1 Líneas futuras

En este trabajo nos hemos encontrado con algunas incidencias, la más destacada ha sido que no hemos podido realizar el trabajo con una base de datos más amplia. En un principio, se realizó con una muestra de un año. Al llevar a cabo la ejecución de los métodos de predicción en los scripts de RStudio, el ordenador utilizado se bloqueaba debido a que no tiene la potencia suficiente para soportar bases de datos tan amplias. Por esa razón, se decidió reducir la base de datos para poder continuar con el desarrollo del trabajo. Sin embargo, como hemos podido observar, es necesario ampliar esta muestra para obtener conclusiones más sólidas.

Por esta razón podemos indicar que una línea futura del trabajo debería ser realizar el mismo análisis, pero con una base de datos más extensa en términos de tiempo. En lugar de utilizar el mes de agosto como base de datos, sería conveniente utilizar uno o dos años completos para que así los métodos de predicción sean más fiables.

Además de ampliar la muestra, también se podrían realizar más métodos de predicción para intentar encontrar mejores predicciones. Uno de ellos podría ser la regresión lineal múltiple, la cual permite generar un modelo lineal en el que el valor de la variable dependiente se determina a partir de un conjunto de variables independientes. Con este modelo, también podríamos evaluar la influencia que tienen los predictores sobre la variable objetivo.

Otro método sería el *Support Vector Machine* (SVM), el cual consiste en que mediante los algoritmos del método kernel se puede mapear no linealmente los datos desde el espacio original a un espacio dimensional diferente, que suele ser superior.

Aparte de estas líneas futuras también se podría realizar un análisis del gasto de forma más detallada y segmentada, separando el gasto total por los diferentes gastos que lo componen, como es gasto en alojamiento, transporte, ocio y alimentación entre otros.

## 5. Bibliografía y webgrafía

Amazon Web Services, Inc. «¿Qué es la regresión lineal? - Explicación del modelo de regresión lineal - AWS». Accedido 1 de junio de 2023. https://aws.amazon.com/es/what-is/linear-regression/.

BRAINTRUST-EKM, Comunicacion. «¿Por qué la empresa turística debe conocer a sus clientes?» *BrainTrust CS* (blog), 19 de septiembre de 2019. https://www.braintrust-cs.com/empresa-turistica-conocer-clientes/.

«ChatGPT». Accedido 4 de mayo de 2023. https://chat.openai.com.

«Cómo funciona el algoritmo XGBoost—ArcGIS Pro | Documentación». Accedido 30 de mayo de 2023. https://pro.arcgis.com/es/pro-app/latest/tool-reference/geoai/how-xgboost-works.htm.

«Exceltur | PIB Turístico Español». Accedido 2 de marzo de 2023. https://www.exceltur.org/pib-turistico-espanol/.

García Esteban, Coral, Ana Gómez Loscos, y César Martín Machuca. «La recuperación del turismo internacional en España tras la pandemia». *Boletín Económico*, n.º 2023/T1 (12 de enero de 2023): 08. https://doi.org/10.53479/25114.

INE. «INEbase / Servicios /Hostelería y turismo /Cuenta satélite del turismo de España / Últimos datos». Accedido 7 de junio de 2023. https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=estadistica_C&cid=1254736169169&menu=ultiDatos&idp=1254735576863.

INE. «INEbase / Servicios /Hostelería y turismo /Encuesta de turismo de residentes / Resultados». Accedido 24 de enero de 2023. https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176990&menu=resultados&idp=1254735576863#!tabs-1254736195369.

«Random Forest (Bosque Aleatorio): combinando árboles - IArtificial.net», 10 de junio de 2019. https://www.iartificial.net/random-forest-bosque-aleatorio/.

«Random Forest Regression in R: Code and Interpretation | HackerNoon». Accedido 30 de mayo de 2023. https://hackernoon.com/random-forest-regression-in-r-code-and-interpretation.

«RPubs - Arboles de decisión con R - Clasificación». Accedido 30 de mayo de 2023. https://rpubs.com/jboscomendoza/arboles_decision_clasificacion.

«RPubs - Clustering Jerárquico en R». Accedido 29 de mayo de 2023. https://rpubs.com/mjimcua/clustering-jerarquico-en-r.

«RPubs - Clustering Jerárquico en R». Accedido 1 de junio de 2023. https://rpubs.com/mjimcua/clustering-jerarquico-en-r.

«RPubs - Ejemplos aplicados de árboles de regresión y clasificación». Accedido 30 de mayo de 2023. https://rpubs.com/kevortiz10/arboles-decision.

«RPubs - How to Tune SVM Parameters ?» Accedido 29 de mayo de 2023. https://www.rpubs.com/CHENW05/520528.

«RPubs - Regresión Lineal Múltiple en R». Accedido 29 de mayo de 2023. https://rpubs.com/Joaquin_AR/226291.

«RPubs - XGBoost en R». Accedido 29 de mayo de 2023. https://rpubs.com/jboscomendoza/xgboost_en_r.

Statologos: El sitio web para que aprendas estadística en Stata, R y Phyton. «XGBoost en R: un ejemplo paso a paso», 4 de mayo de 2021. https://statologos.com/xgboost-en-r/.

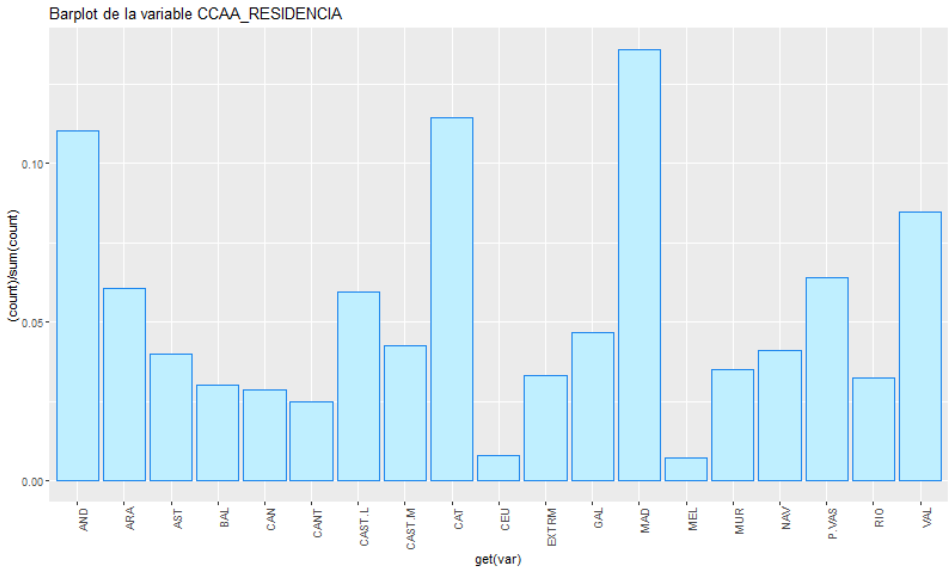
«Support Vector Machine (SVM)». Accedido 1 de junio de 2023. https://es.mathworks.com/discovery/support-vector-machine.html.

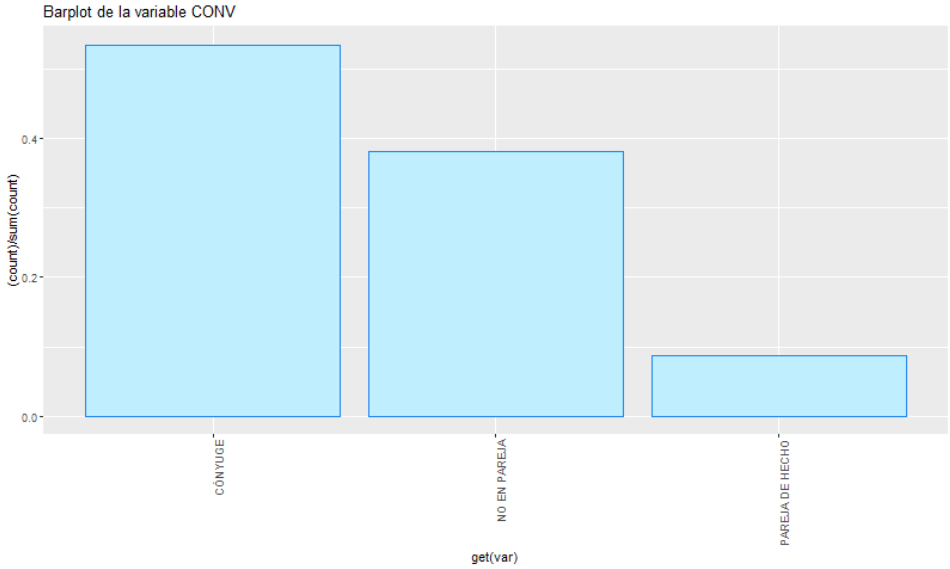
«XGBoost — Programación — DATA SCIENCE». Accedido 1 de junio de 2023. https://datascience.eu/es/programacion/xgboost-4/.

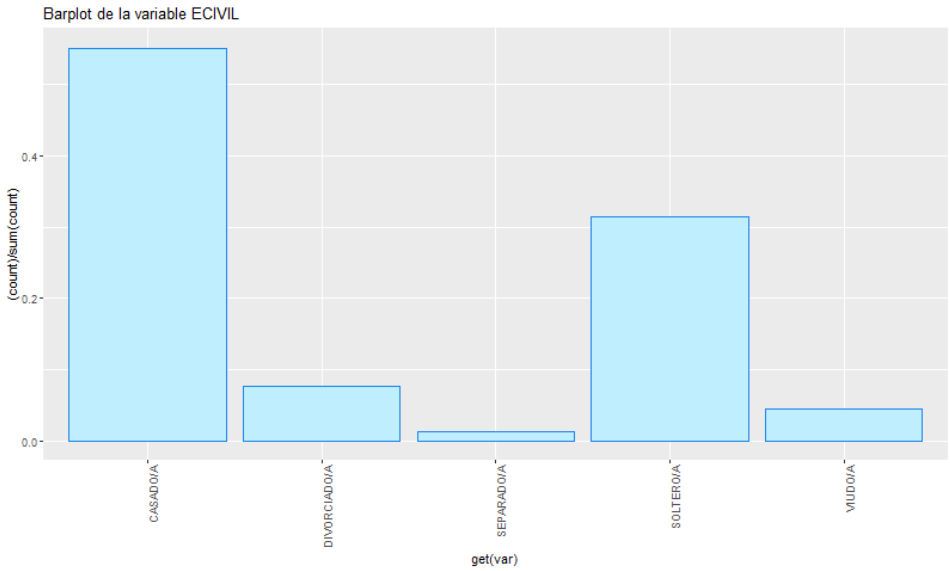
# 6. Anexo

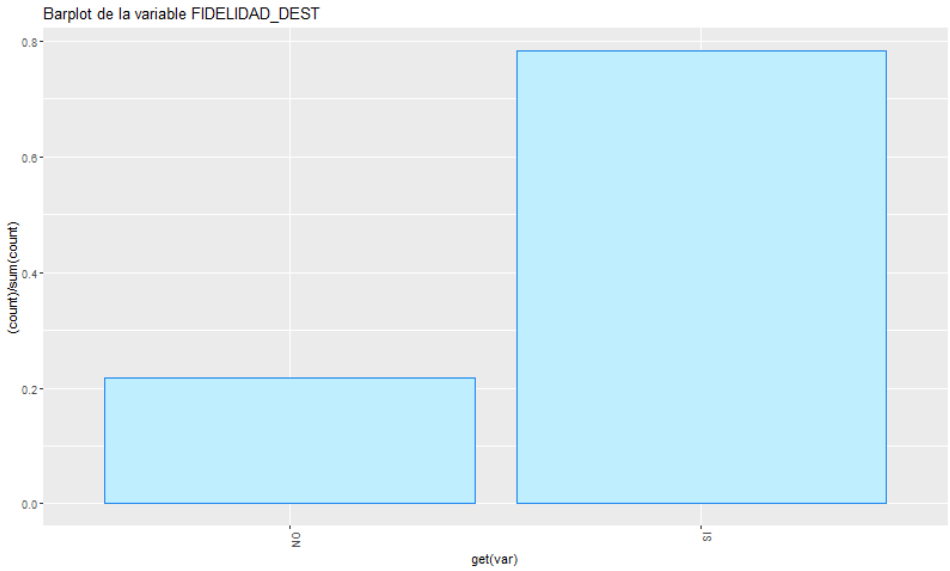
Todos los gráficos que aparecen en el anexo han sido elaborados en este trabajo por RStudio

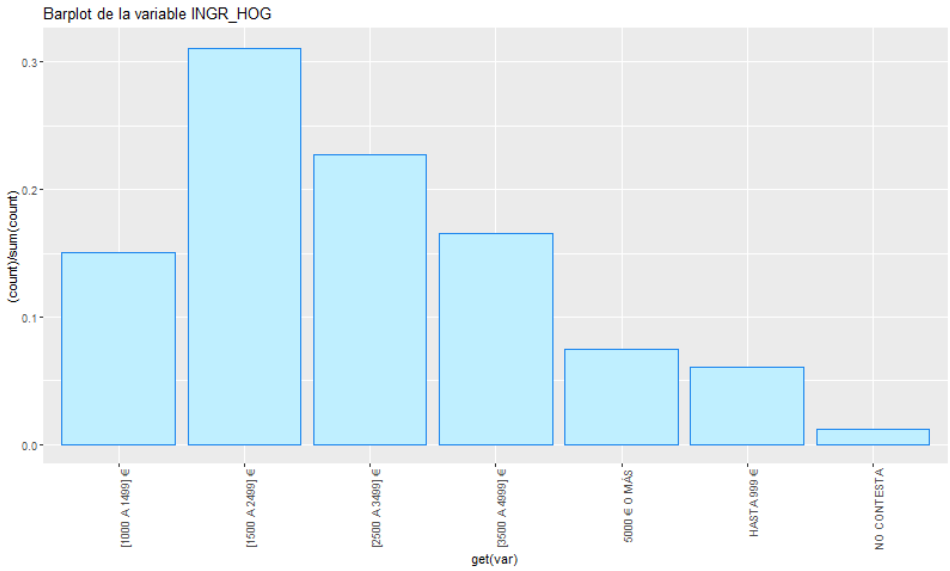
Gráficos Univariantes, variables categóricas:

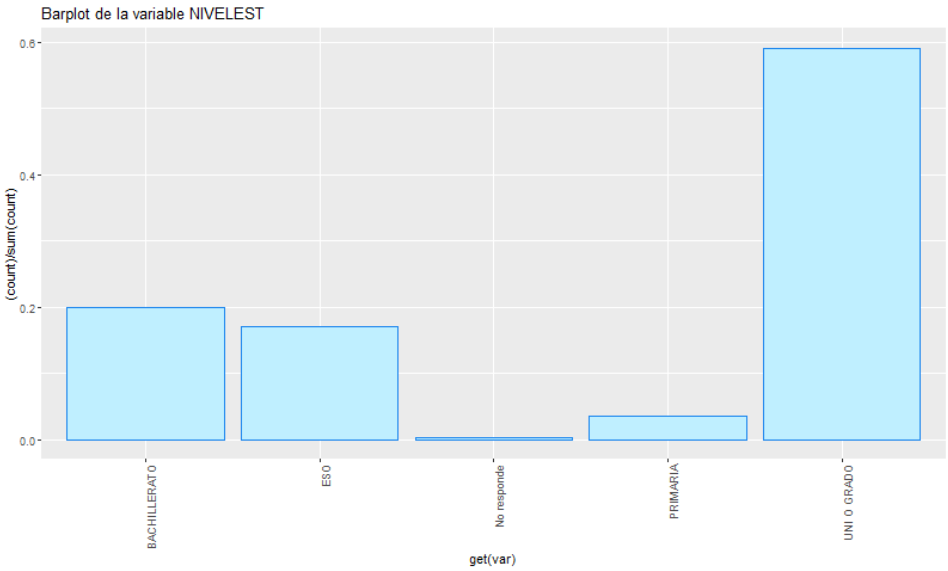


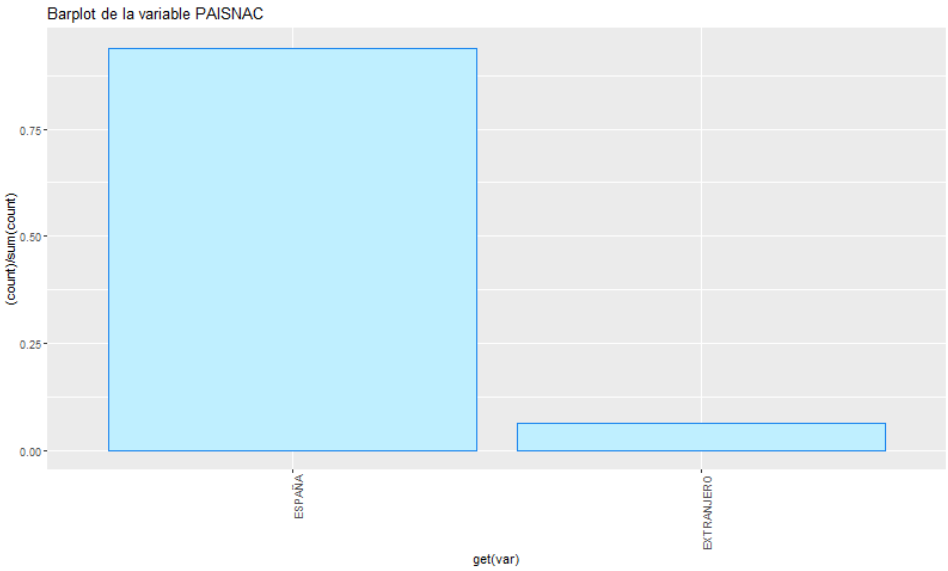



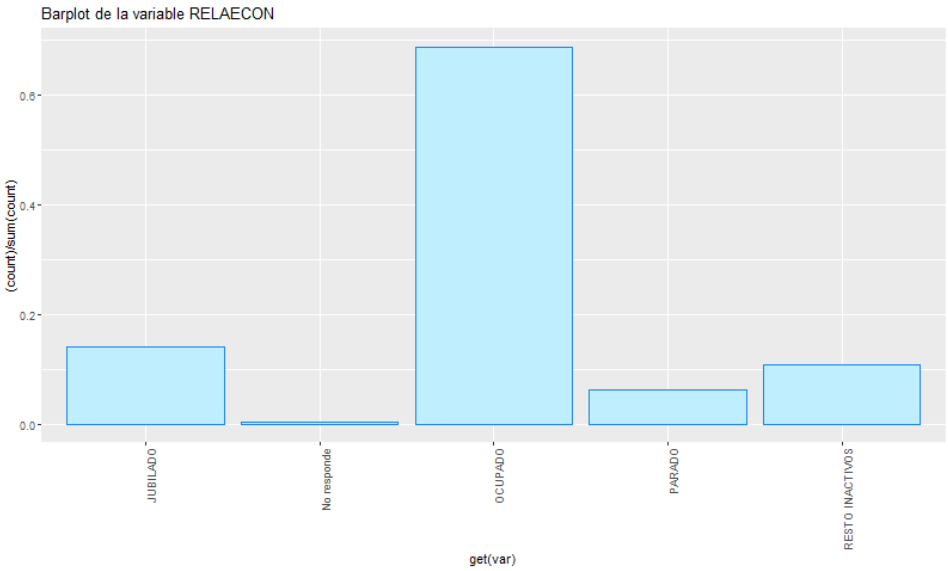

Barplot de la variable ECIVIL
(count)/sum(count)
0.4
0.2
0.0
CASADO/A
DIVORCIADO/A
SEPARADO/A
SOLTERO/A
VIUDO/A
get(var)

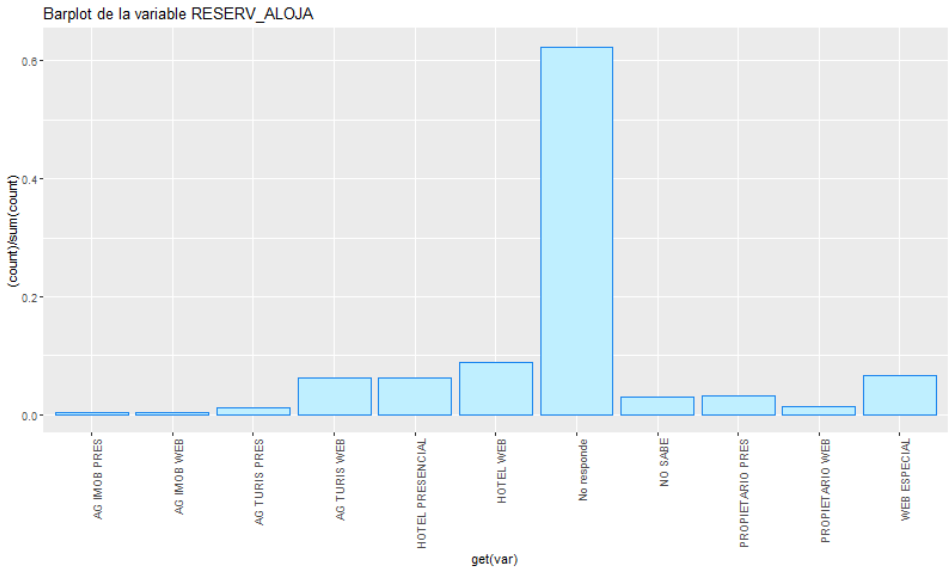

Barplot de la variable FIDELIDAD_DEST
(count)/sum(count)
0.8
0.6
0.4
0.2
0.0
NO
SI
get(var)

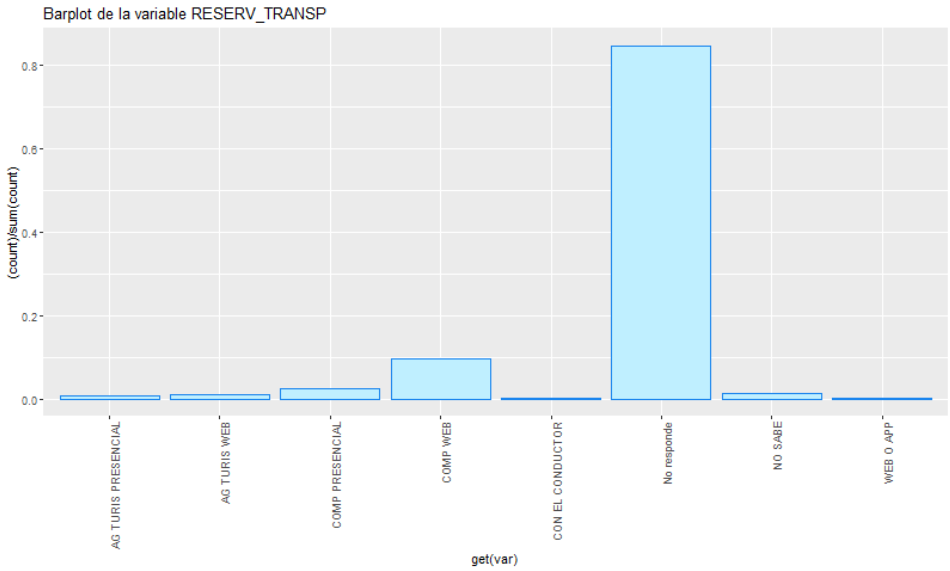

Barplot de la variable INGR_HOG
(count)/sum(count)
0.3
0.2
0.1
0.0
[1000 A 1499] €
[1500 A 2499] €
[2500 A 3499] €
[3500 A 4999] €
5000 € O MAS
HASTA 999 €
NO CONTESTA
get(var)

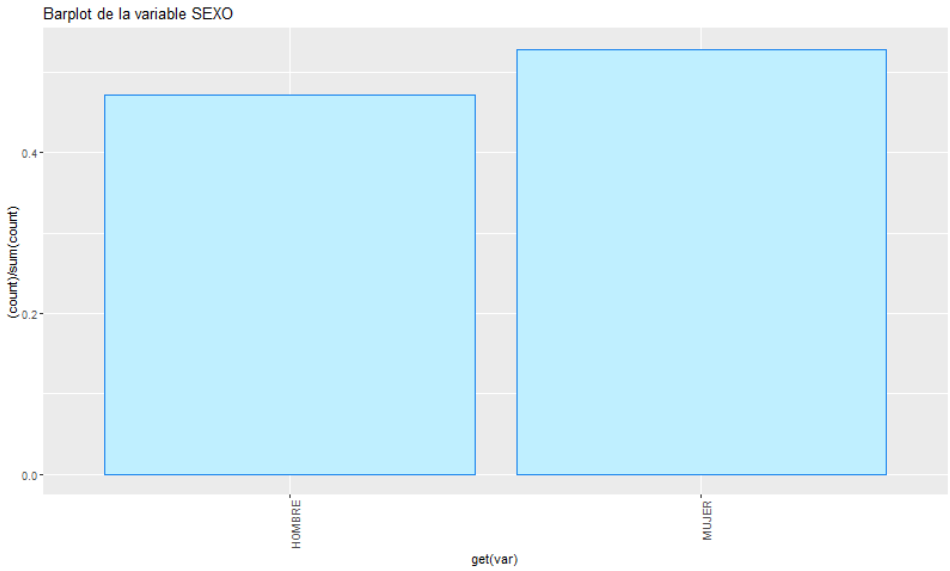

Barplot de la variable NIVELEST
(count)/sum(count)
0.0
0.2
0.4
0.6
BACHILLERATO
ESO
No responde
PRIMARIA
UNI O GRADO
get(var)

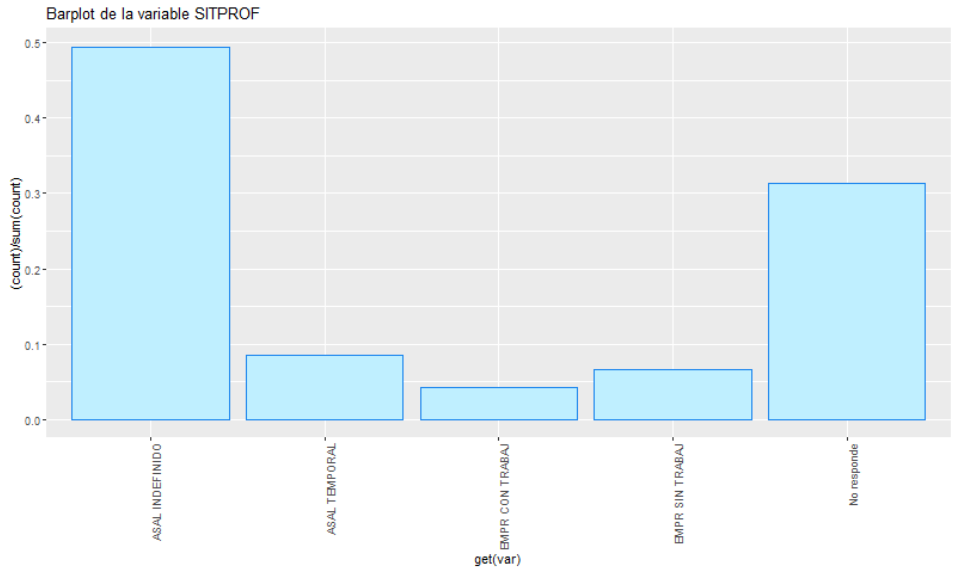

Barplot de la variable PAISNAC
(count)/sum(count)
0.00
0.25
0.50
0.75
ESPAÑA
EXTRANJERO
get(var)

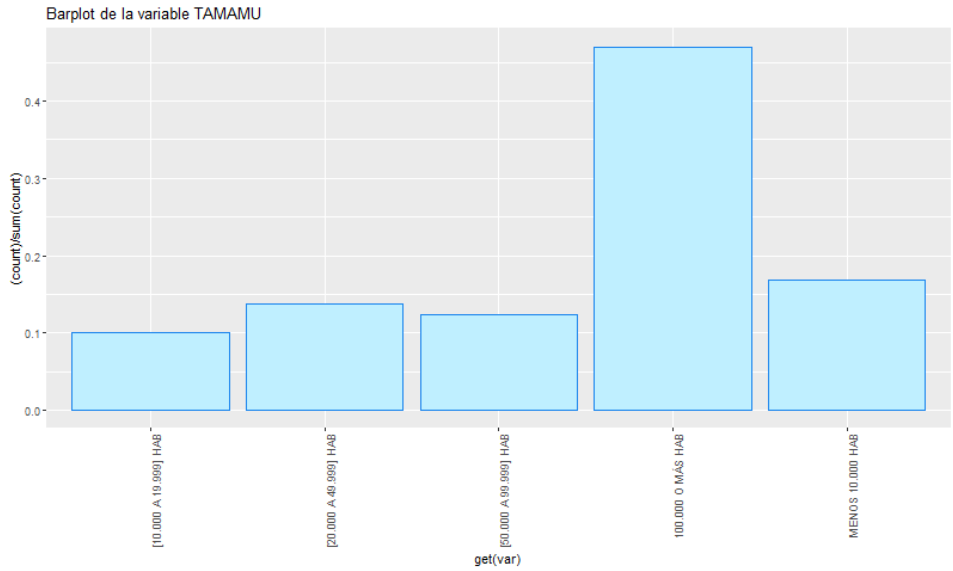

Barplot de la variable RELAECON
(count)/sum(count)
0.0
0.2
0.4
0.6
JUBILADO
No responde
OCUPADO
PARADO
RESTO INACTIVOS
get(var)

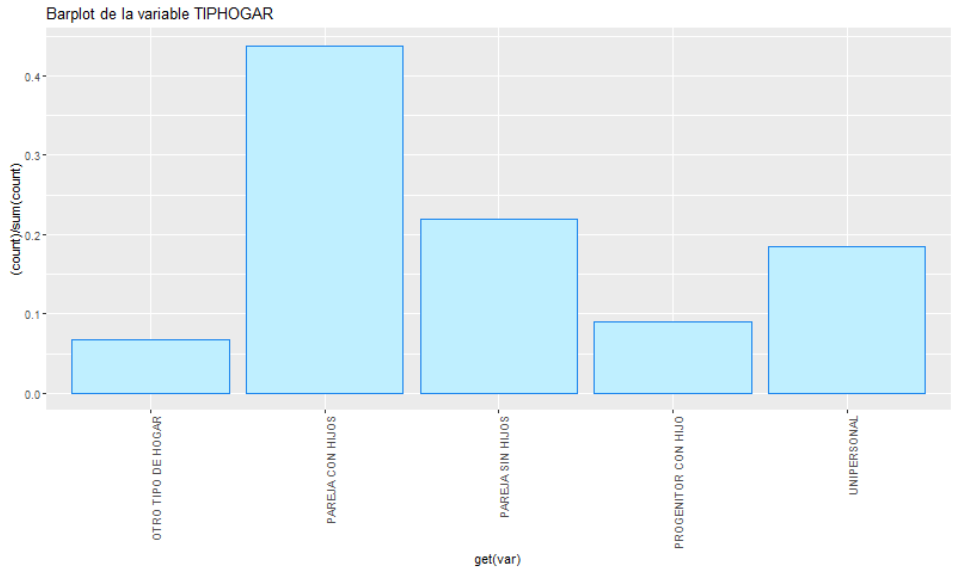

Barplot de la variable RESERV_ALOJA
(count)/sum(count)
0.0
0.2
0.4
0.6
AG IMOB PRES
AG IMOB WEB
AG TURIS PRES
AG TURIS WEB
HOTEL PRESENCIAL
HOTEL WEB
No responde
NO SABE
PROPIETARIO PRES
PROPIETARIO WEB
WEB ESPECIAL
get(var)

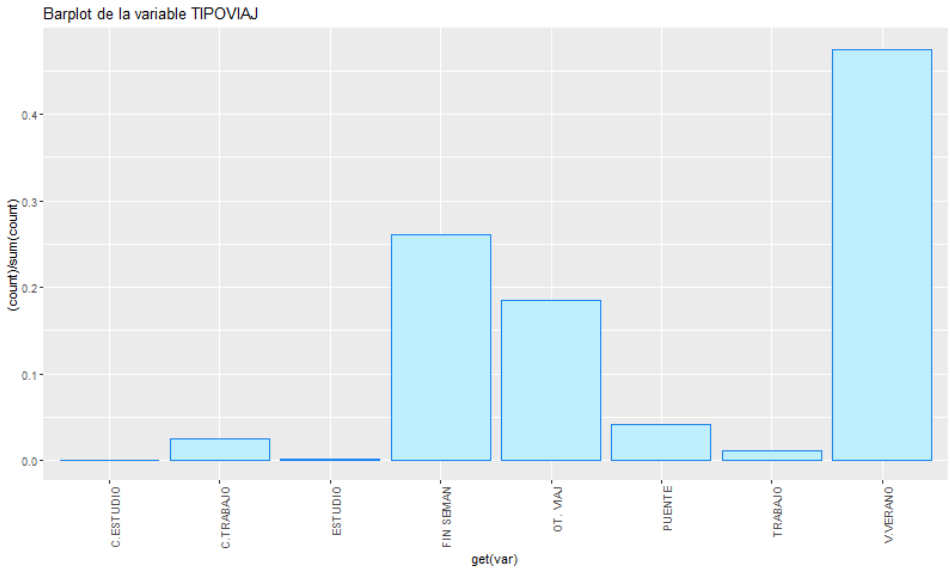

Barplot de la variable RESERV_TRANSP
(count)/sum(count)
0.0
0.2
0.4
0.6
0.8
AG TURIS PRESENCIAL
AG TURIS WEB
COMP PRESENCIAL
COMP WEB
CON EL CONDUCTOR
No responde
NO SABE
WEB O APP
get(var)

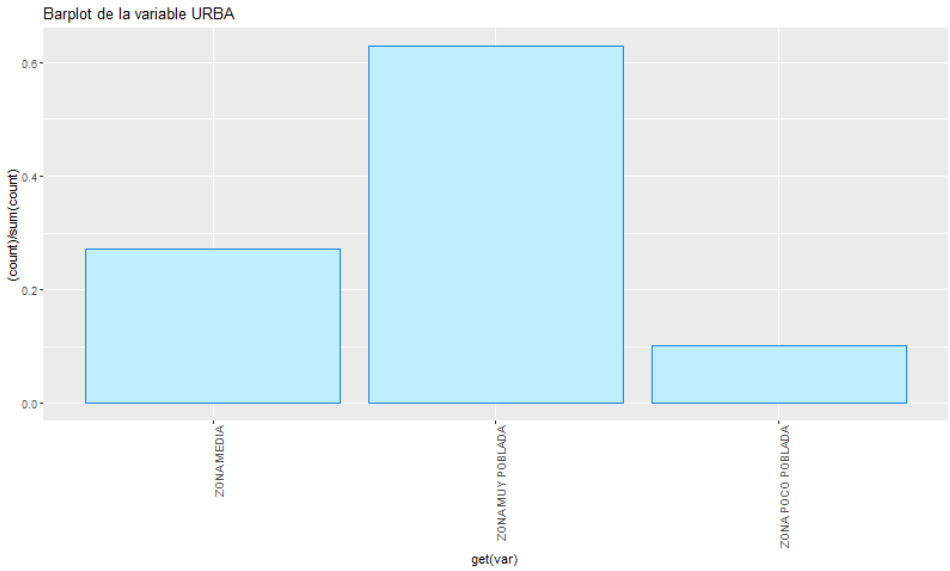

Barplot de la variable SEXO
(count)/sum(count)
0.0
0.2
0.4
HOMBRE
MUJER
get(var)

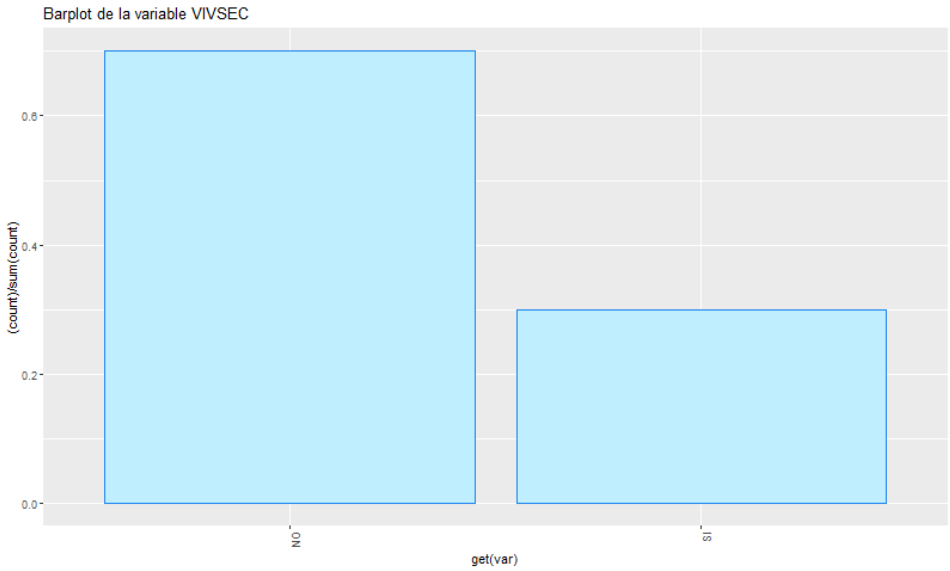

Barplot de la variable SITPROF
(count)/sum(count)
0.0
0.1
0.2
0.3
0.4
0.5
ASAL INDEFINIDO
ASAL TEMPORAL
EMPR CON TRABAJ
EMPR SIN TRABAJ
No responde
get(var)


Barplot de la variable TAMAMU
(count)/sum(count)
0.0
0.1
0.2
0.3
0.4
[10.000 A 19.999] HAB
[20.000 A 49.999] HAB
[50.000 A 99.999] HAB
100.000 O MÁS HAB
MENOS 10.000 HAB
get(var)

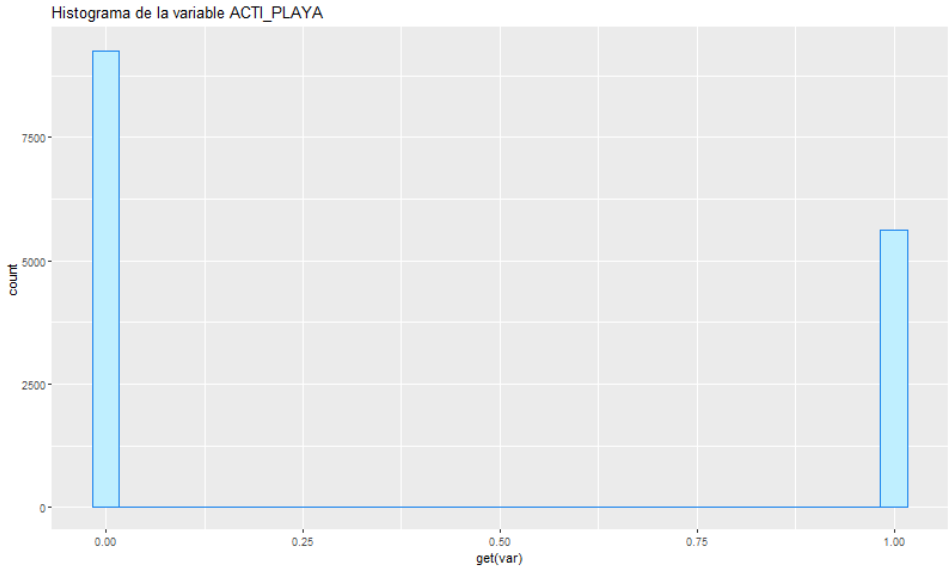

Barplot de la variable TIPHOGAR
(count)/sum(count)
0.0
0.1
0.2
0.3
0.4
OTRO TIPO DE HOGAR
PAREJA CON HIJOS
PAREJA SIN HIJOS
PROGENITOR CON HIJO
UNIPERSONAL
get(var)

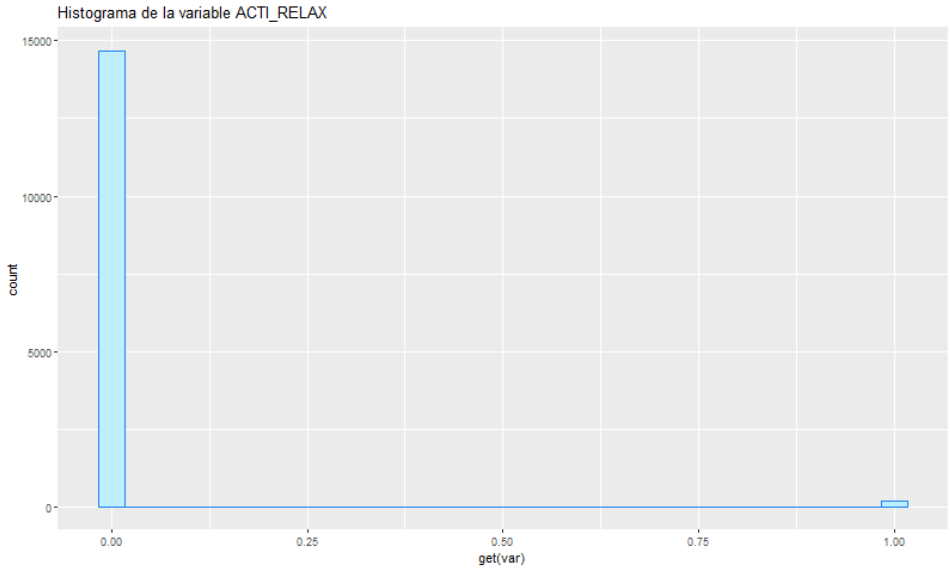

Barplot de la variable TIPOVIAJ
(count)/sum(count)
0.0
0.1
0.2
0.3
0.4
C.ESTUDIO
C.TRABAJO
ESTUDIO
FIN SEMAN
OT. VIAJ
PUENTE
TRABAJO
VVERANO
get(var)

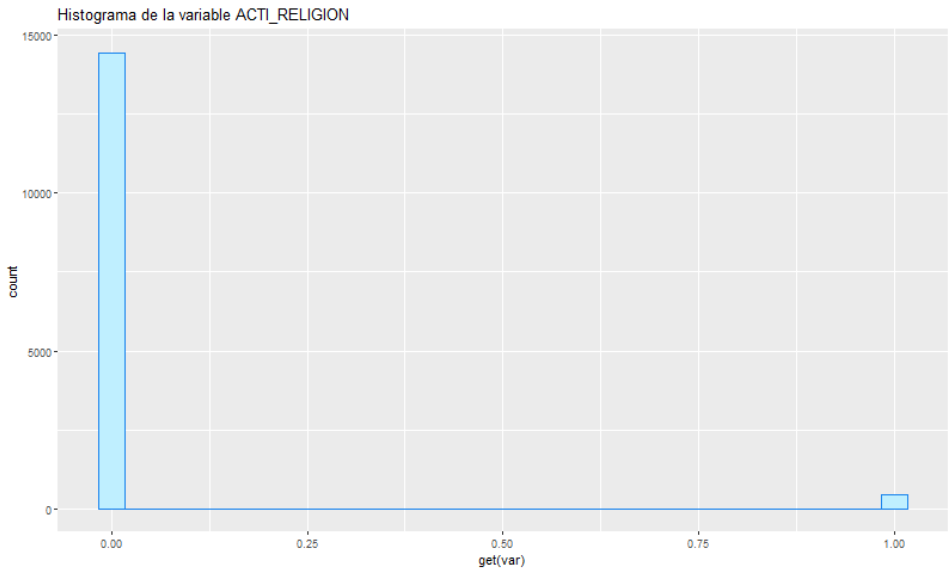

Barplot de la variable URBA
(count)/sum(count)
0.0
0.2
0.4
0.6
ZONA MEDIA
ZONA MUY POBLADA
ZONA POCO POBLADA
get(var)

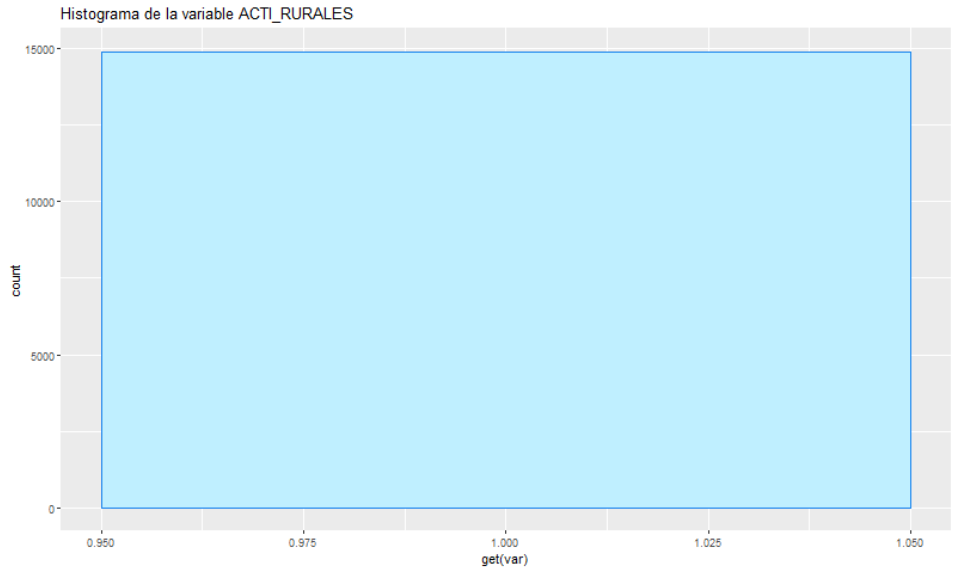

Barplot de la variable VIVSEC
(count)/sum(count)
0.0
0.2
0.4
0.6
NO
SI
get(var)

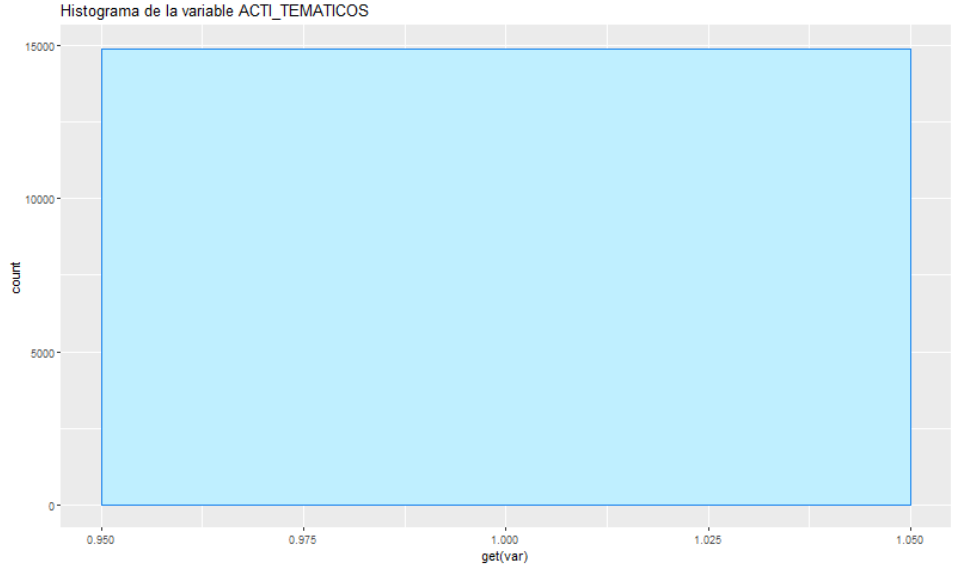
Gráficos univariantes, variables numéricas:

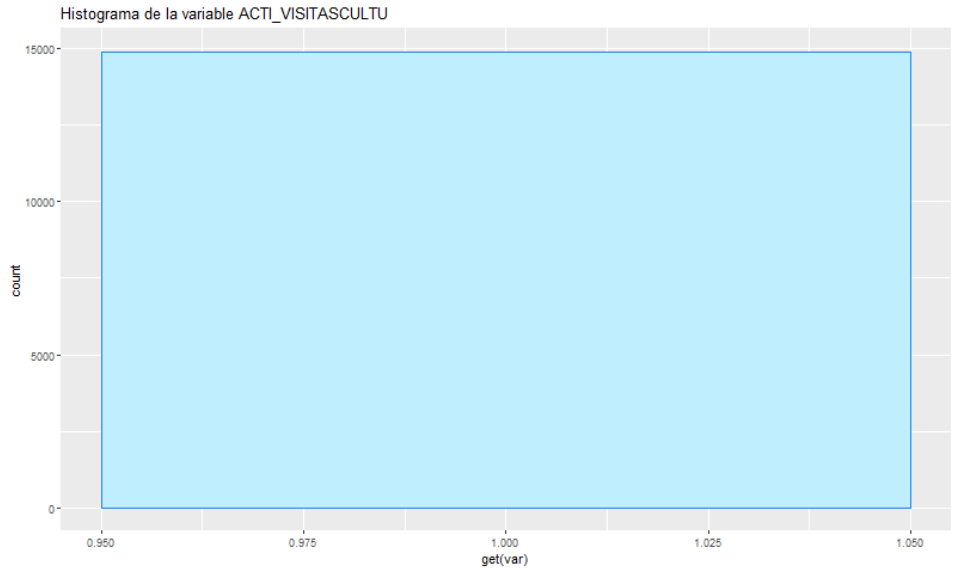


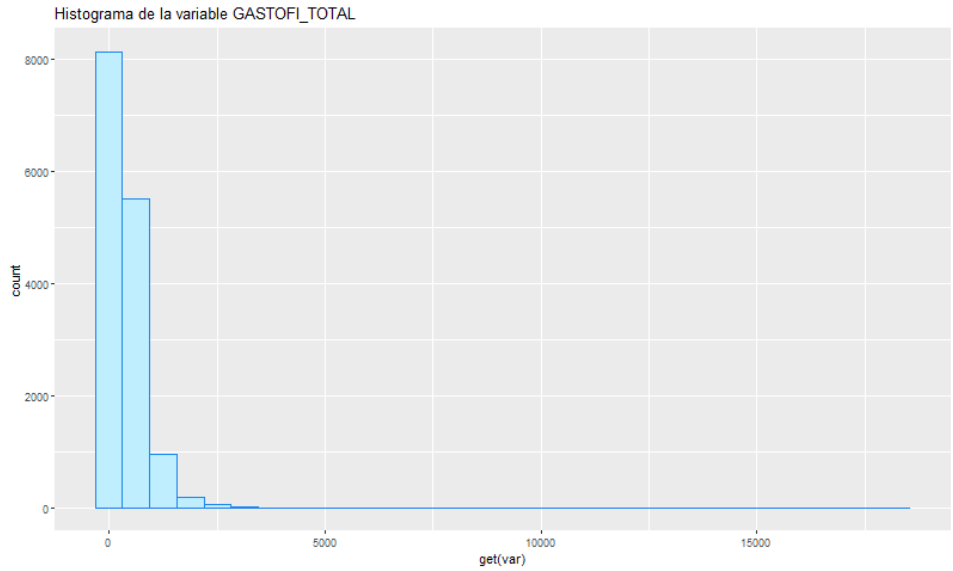



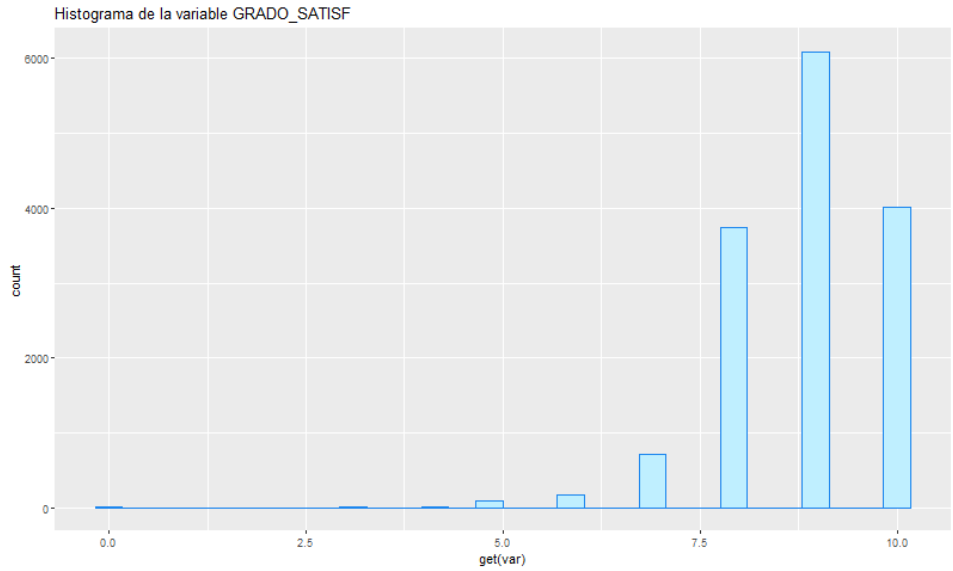



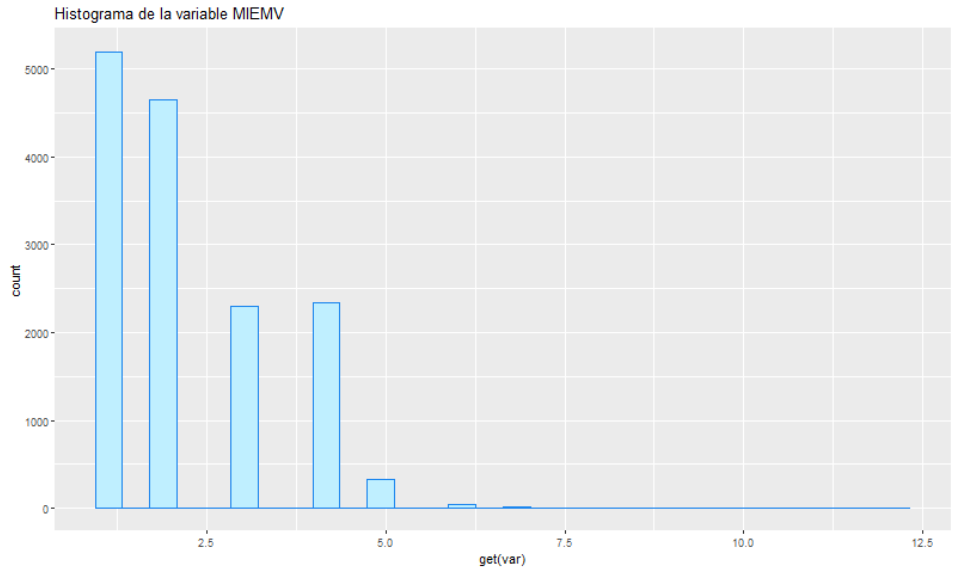

Histograma de la variable ACTI_RURALES
15000
10000
5000
0
count
0.950
0.975
1.000
1.025
1.050
get(var)

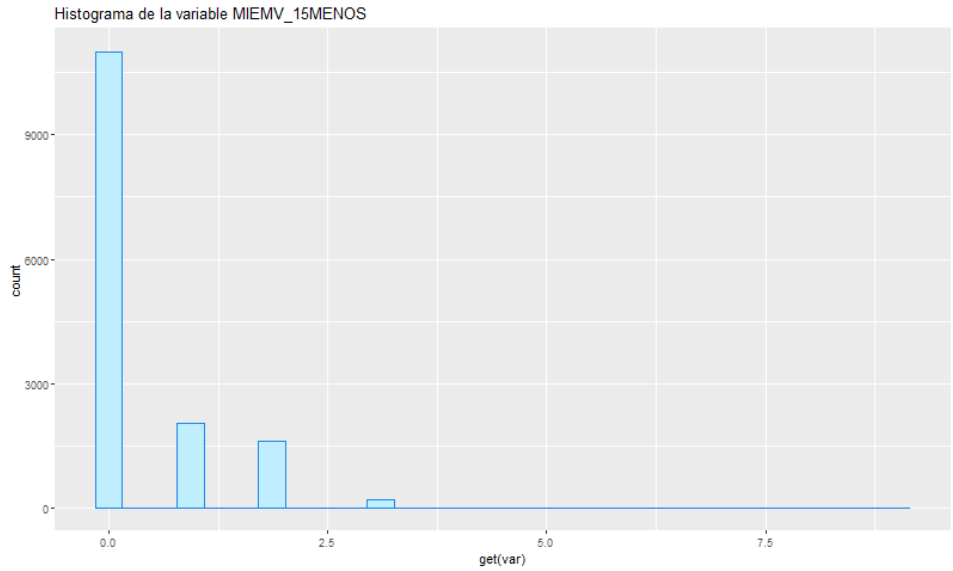

Histograma de la variable ACTI_TEMATICOS
15000
10000
5000
0
count
0.950
0.975
1.000
1.025
1.050
get(var)

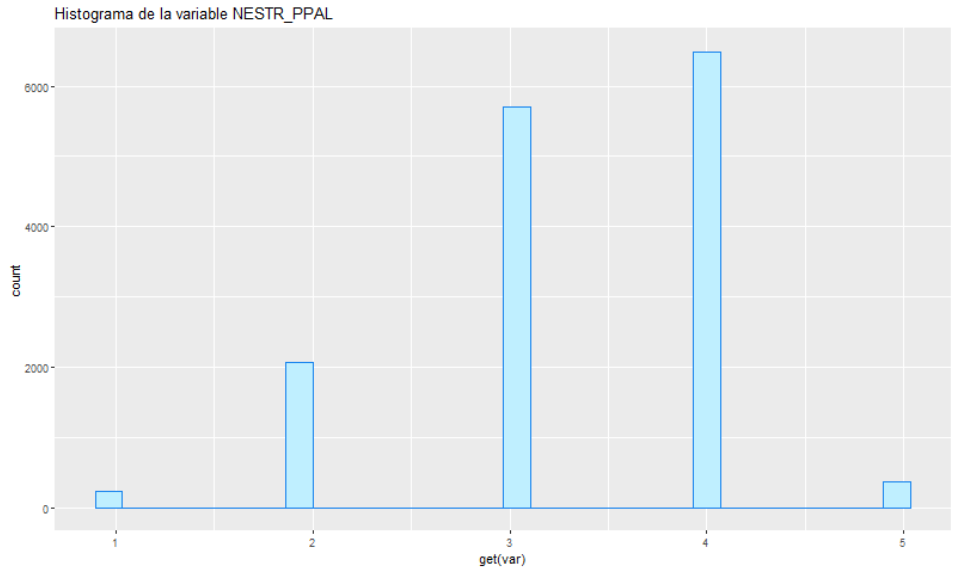

Histograma de la variable ACTI_VISITASCULTU
15000
10000
5000
0
count
0.950
0.975
1.000
1.025
1.050
get(var)

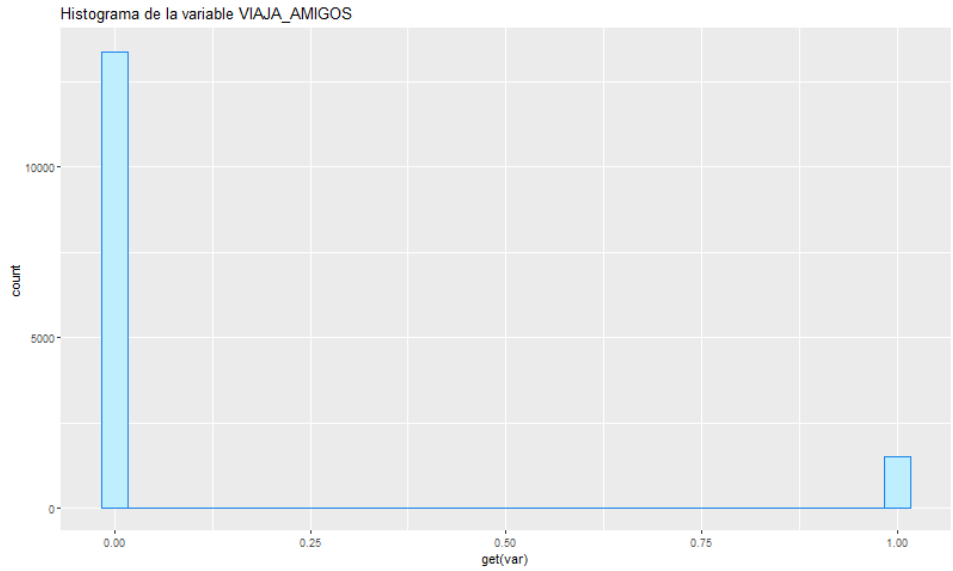

Histograma de la variable GASTOFI_TOTAL
count
get(var)
0
2000
4000
6000
8000
0
5000
10000
15000

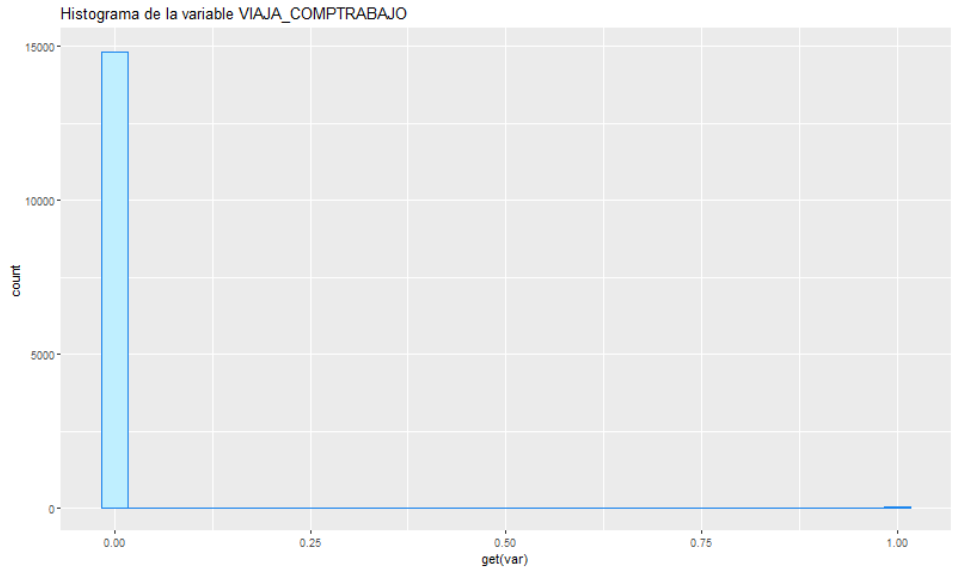

Histograma de la variable GRADO_SATISF
count
get(var)
0
2000
4000
6000
0.0
2.5
5.0
7.5
10.0

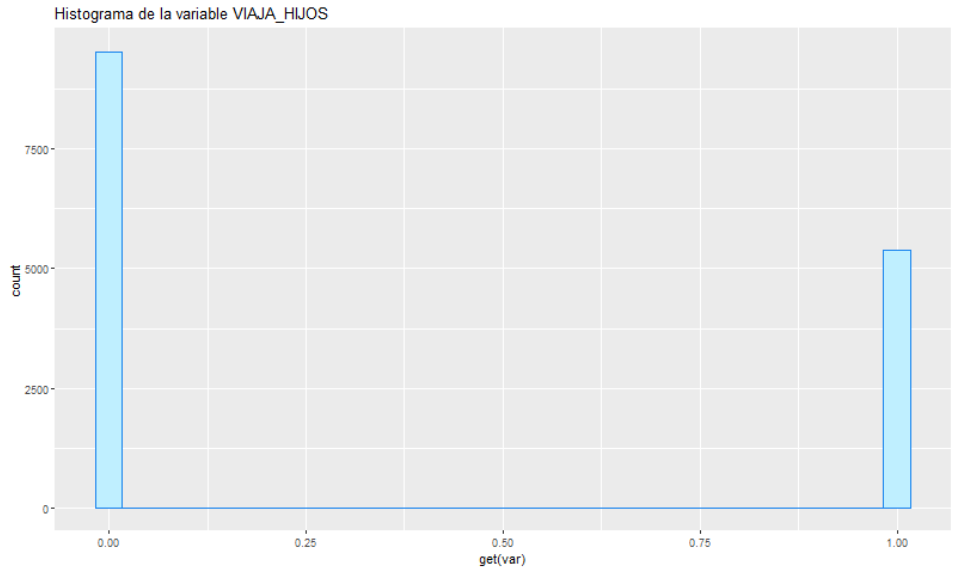

Histograma de la variable MIEMV
count
get(var)
0
1000
2000
3000
4000
5000
2.5
5.0
7.5
10.0
12.5

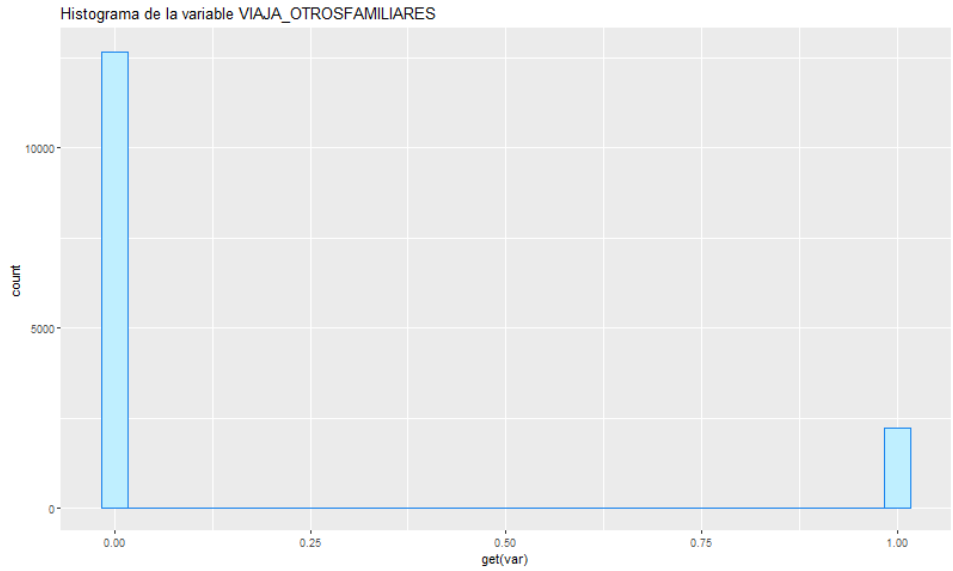

Histograma de la variable MIEMV_15MENOS
count
9000
6000
3000
0
0.0
2.5
5.0
7.5
get(var)

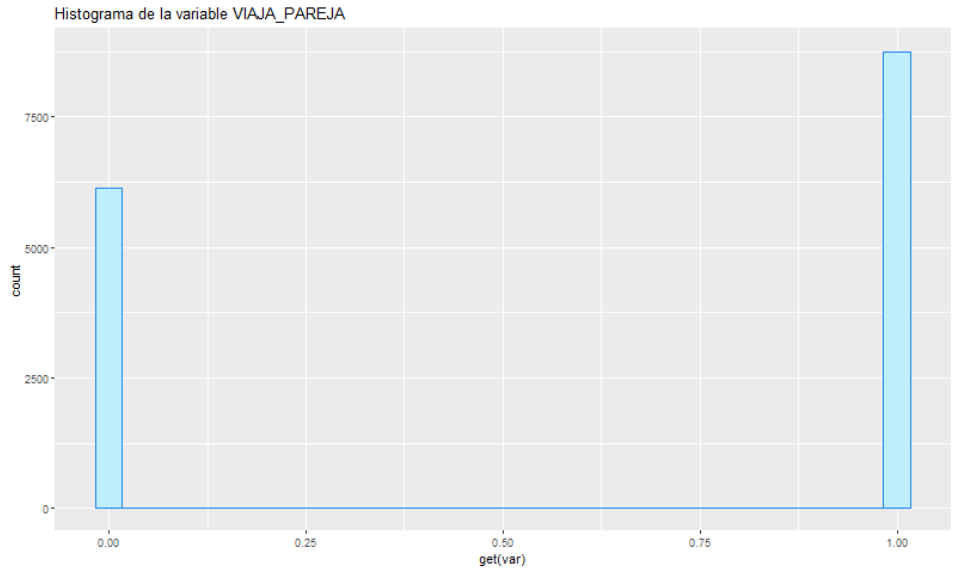

Histograma de la variable NESTR_PPAL
count
6000
4000
2000
0
1
2
3
4
5
get(var)

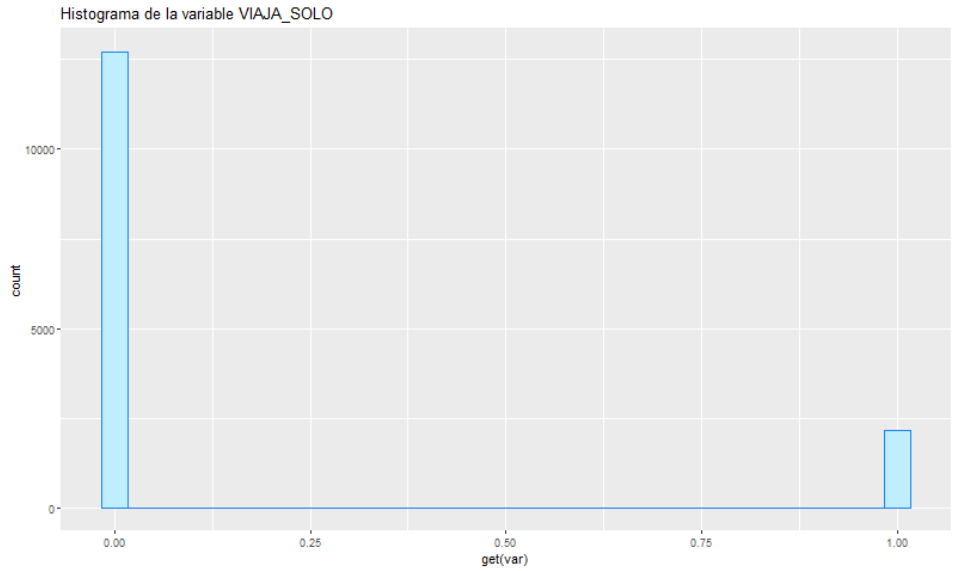

Histograma de la variable VIAJA_AMIGOS
count
10000
5000
0
0.00
0.25
0.50
0.75
1.00
get(var)

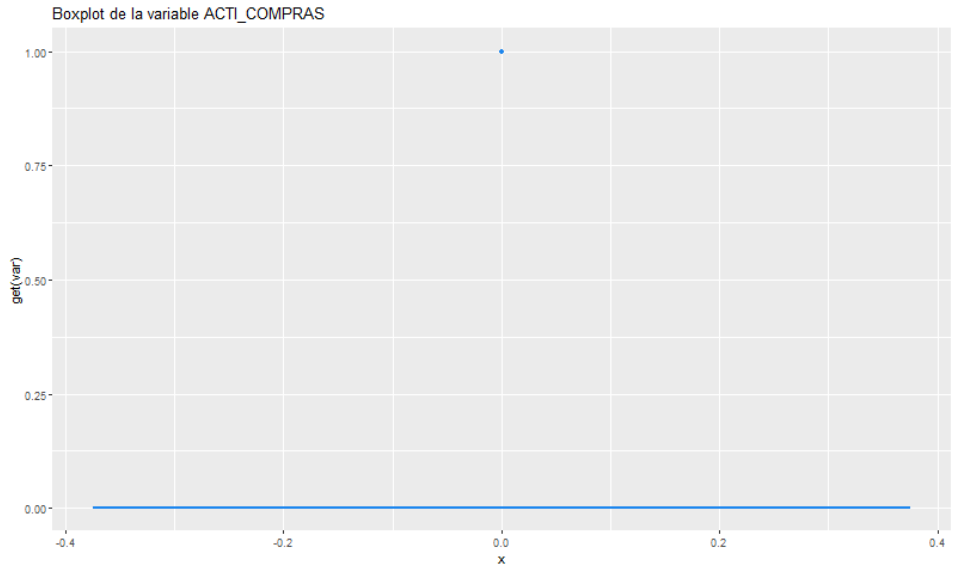

Histograma de la variable VIAJA_COMPTRABAJO
15000
10000
5000
0
count
0.00
0.25
0.50
0.75
1.00
get(var)

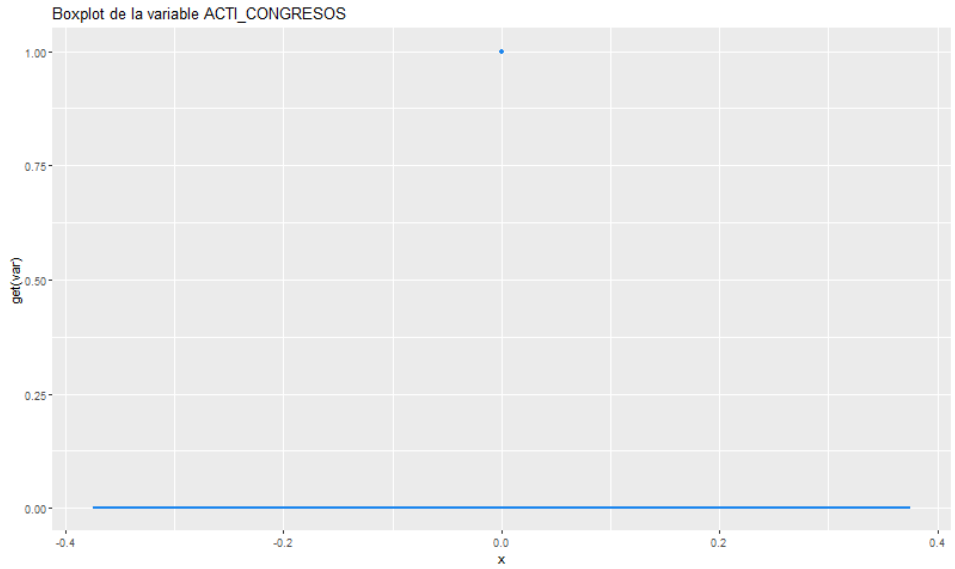

Histograma de la variable VIAJA_HIJOS
7500
5000
2500
0
count
0.00
0.25
0.50
0.75
1.00
get(var)

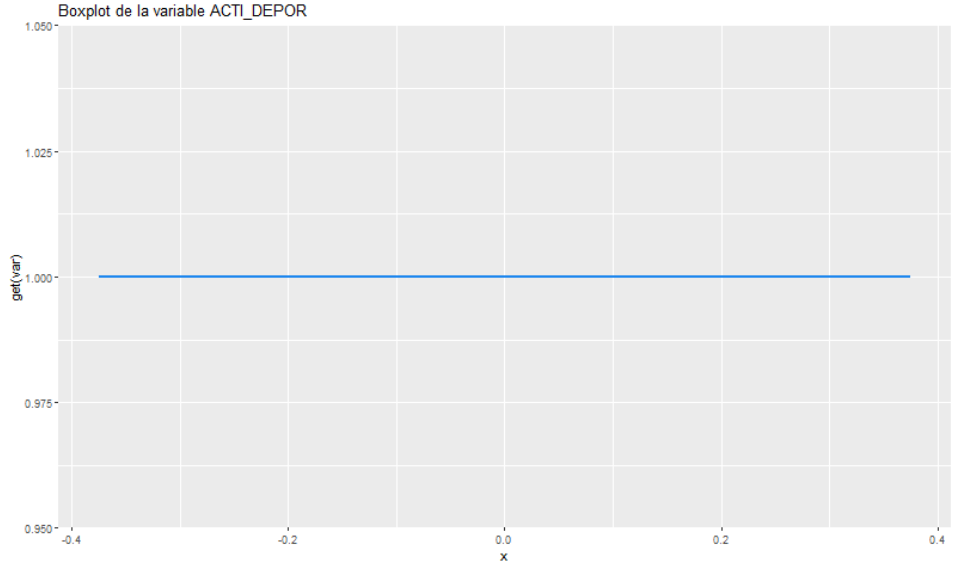

Histograma de la variable VIAJA_OTROSFAMILIARES
10000
5000
0
count
0.00
0.25
0.50
0.75
1.00
get(var)

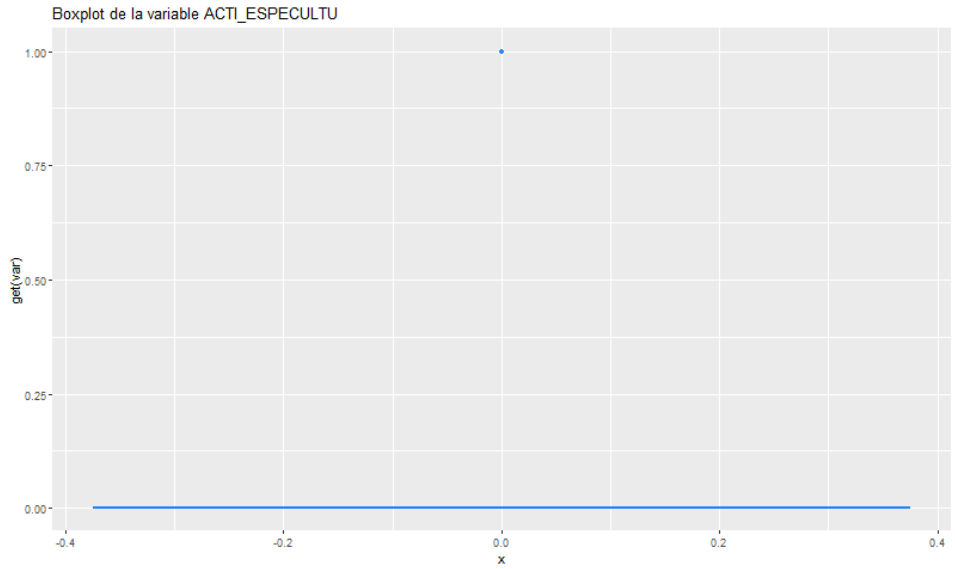

Histograma de la variable VIAJA_PAREJA
count
7500
5000
2500
0
0.00
0.25
0.50
0.75
1.00
get(var)

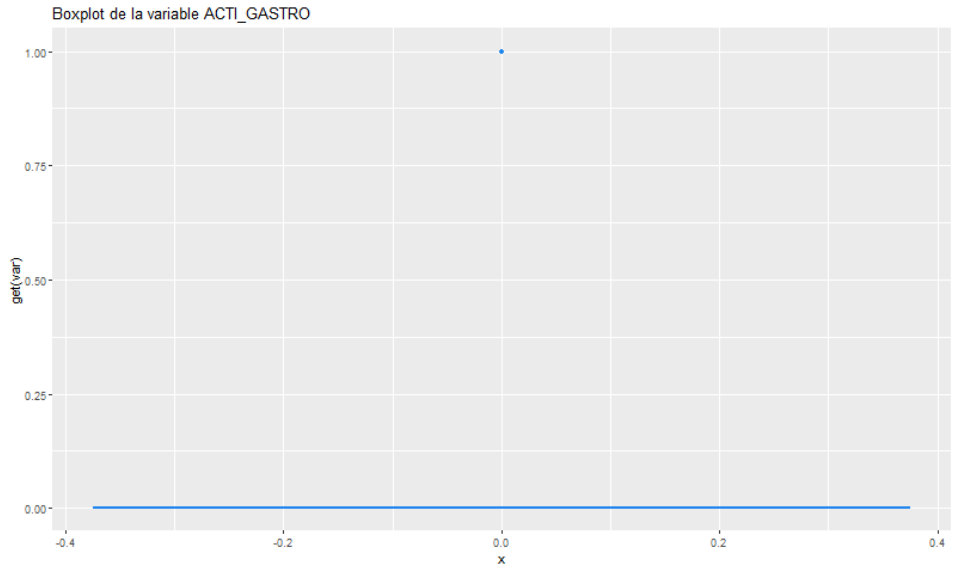

Histograma de la variable VIAJA_SOLO
count
10000
5000
0
0.00
0.25
0.50
0.75
1.00
get(var)

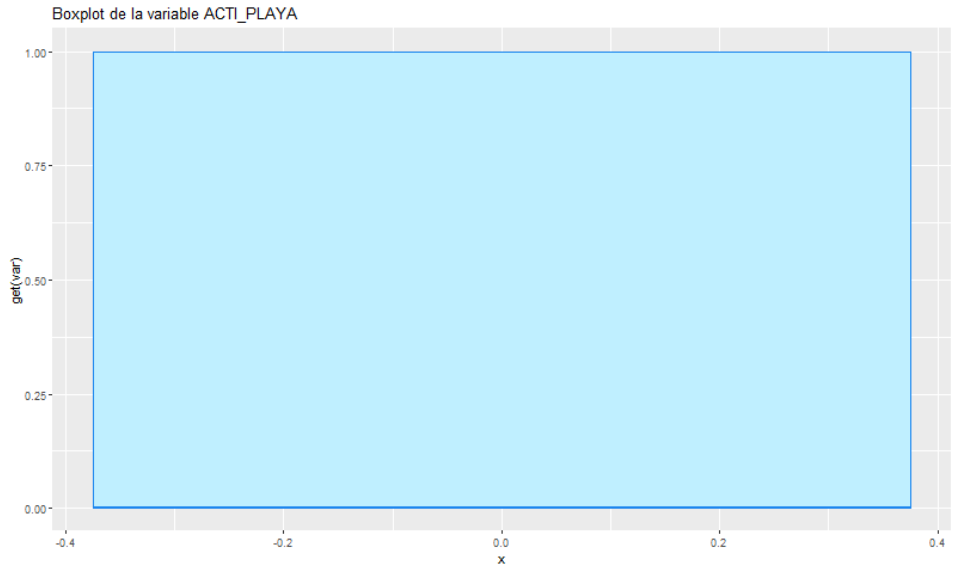

Boxplot de la variable ACTI_COMPRAS
get(var)
1.00
0.75
0.50
0.25
0.00
-0.4
-0.2
0.0
0.2
0.4
x

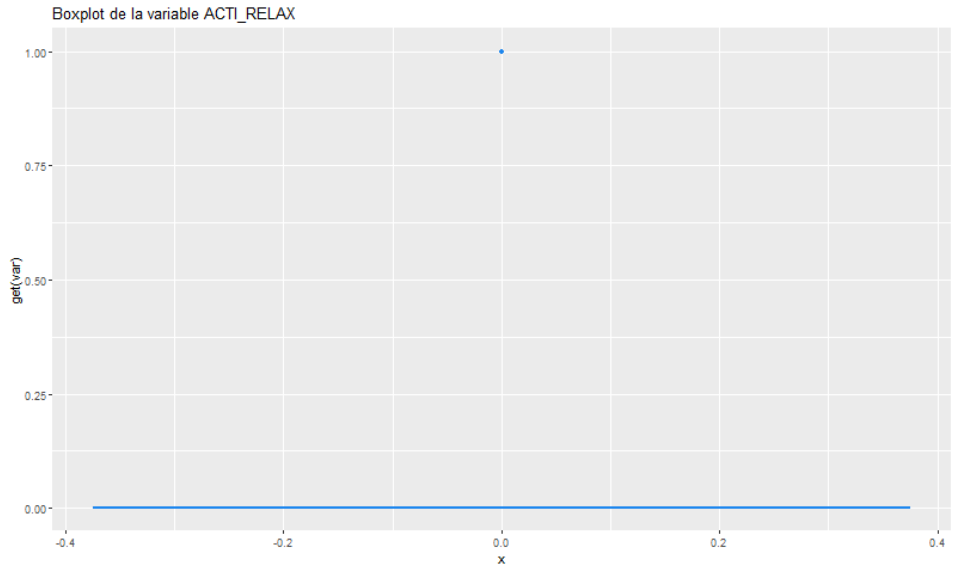

Boxplot de la variable ACTI_CONGRESOS
get(var)
1.00
0.75
0.50
0.25
0.00
-0.4
-0.2
0.0
0.2
0.4
x

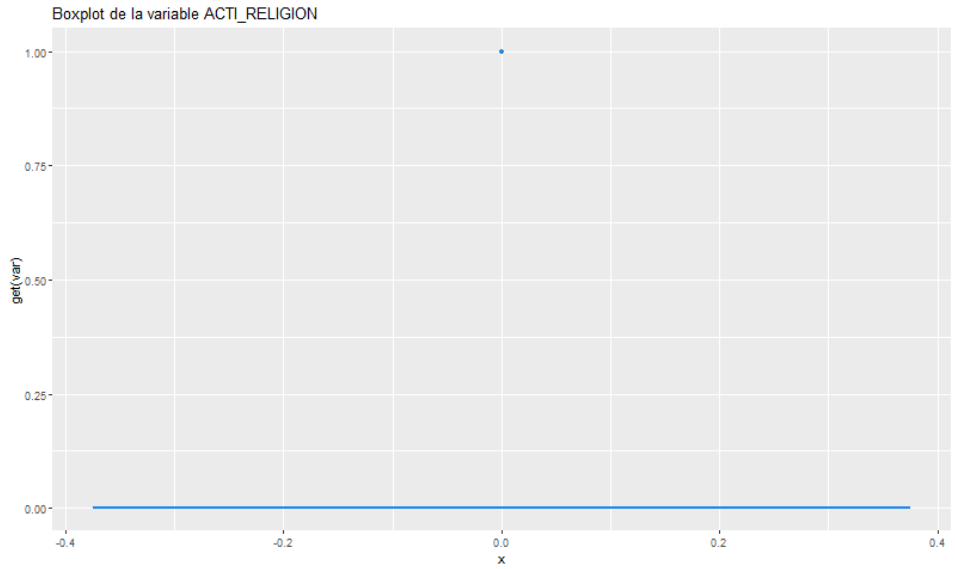

Boxplot de la variable ACTI_DEPOR
get(var)
1.050
1.025
1.000
0.975
0.950
-0.4
-0.2
0.0
0.2
0.4
x

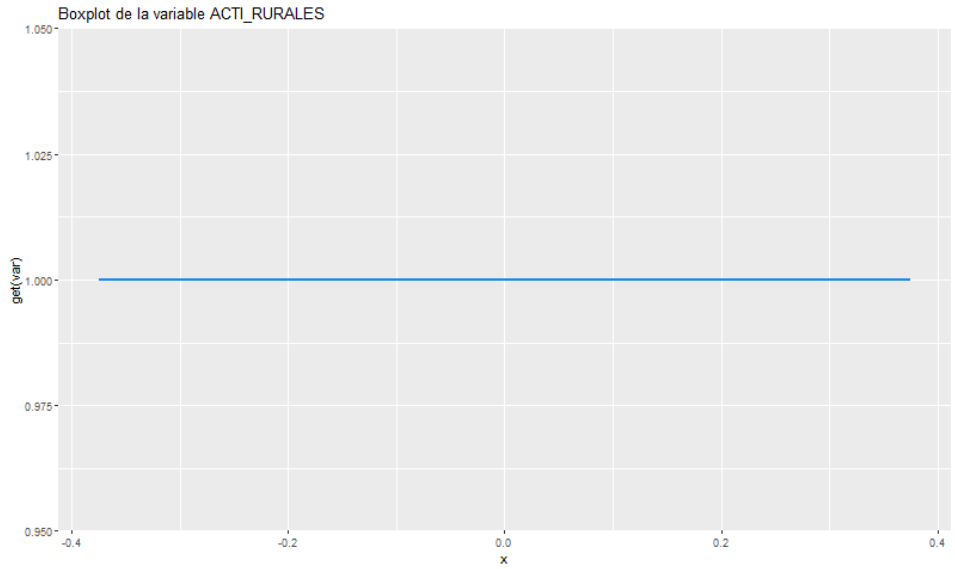

Boxplot de la variable ACTI_ESPECULTU
get(var)
1.00
0.75
0.50
0.25
0.00
-0.4
-0.2
0.0
0.2
0.4
x

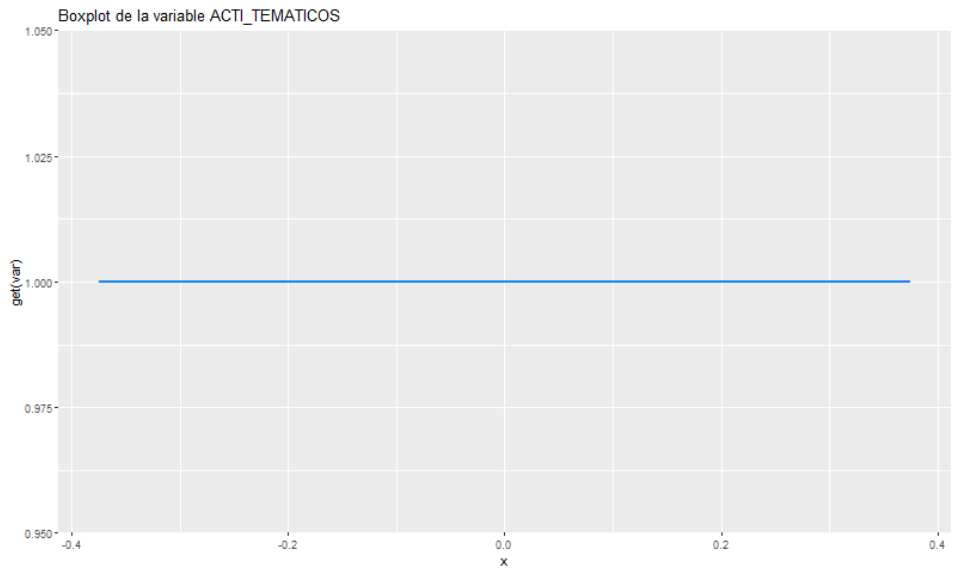

Boxplot de la variable ACTI_GASTRO
get(var)
1.00
0.75
0.50
0.25
0.00
-0.4
-0.2
0.0
0.2
0.4
x

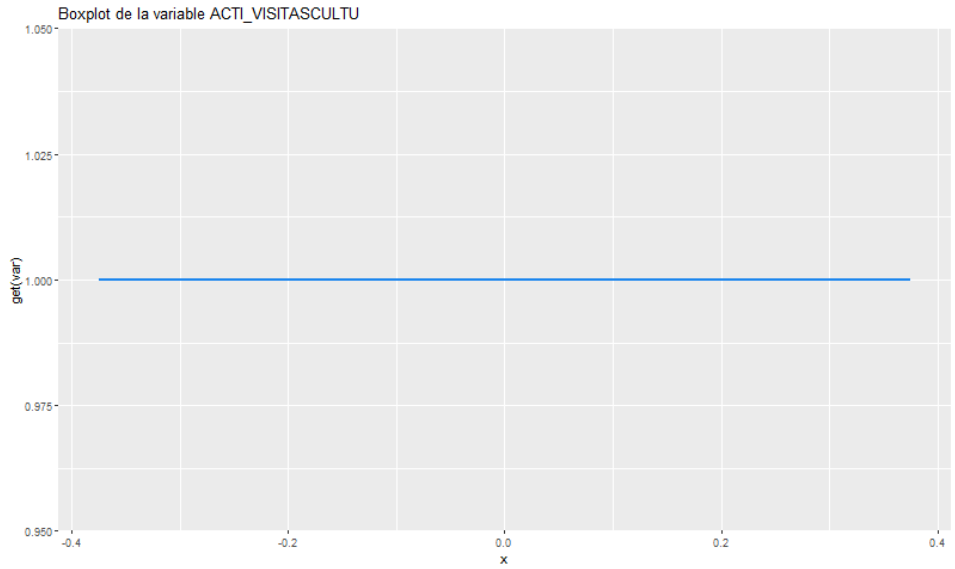

Boxplot de la variable ACTI_PLAYA
get(var)
1.00
0.75
0.50
0.25
0.00
-0.4
-0.2
0.0
0.2
0.4
x

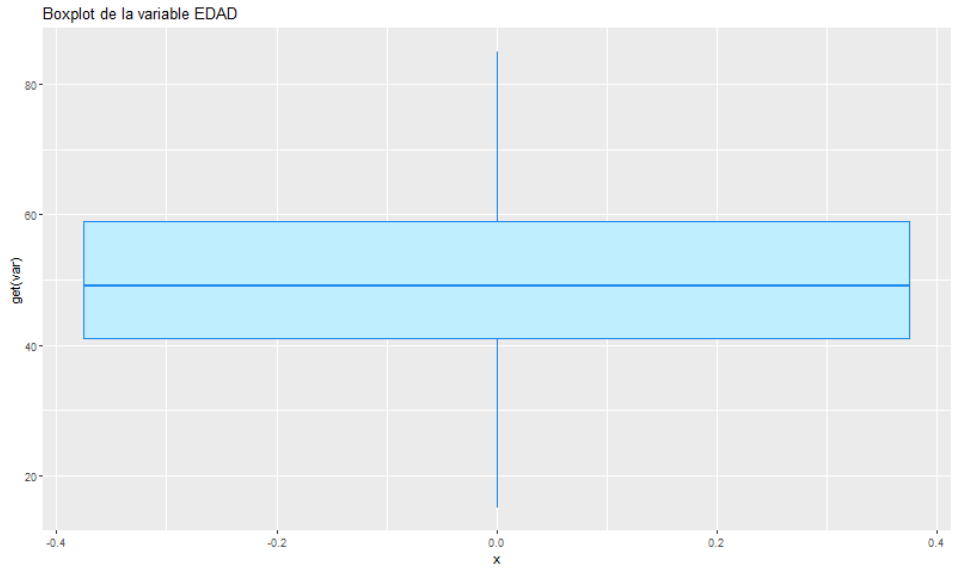

Boxplot de la variable ACTI_RELAX
get(var)
1.00
0.75
0.50
0.25
0.00
-0.4
-0.2
0.0
0.2
0.4
x

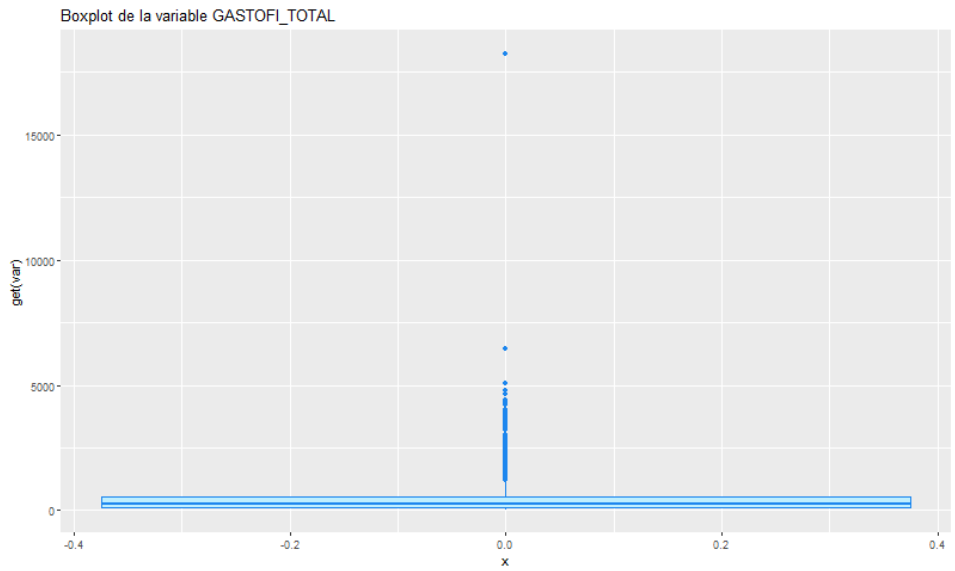

Boxplot de la variable ACTI_RELIGION
get(var)
1.00
0.75
0.50
0.25
0.00
-0.4
-0.2
0.0
0.2
0.4
x

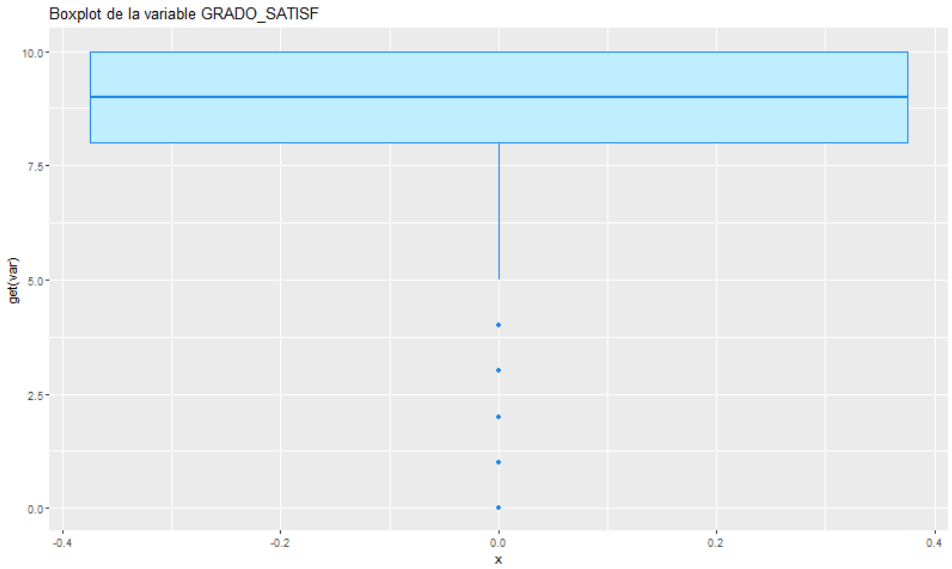

Boxplot de la variable ACTI_RURALES
get(var)
1.050
1.025
1.000
0.975
0.950
-0.4
-0.2
0.0
0.2
0.4
x

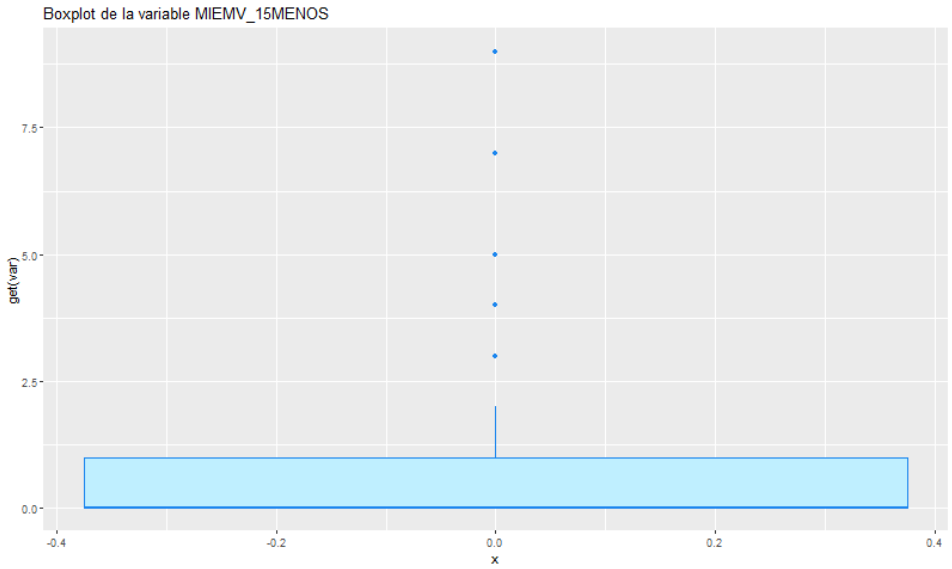

Boxplot de la variable ACTI_TEMATICOS
get(var)
1.050
1.025
1.000
0.975
0.950
-0.4
-0.2
0.0
0.2
0.4
x

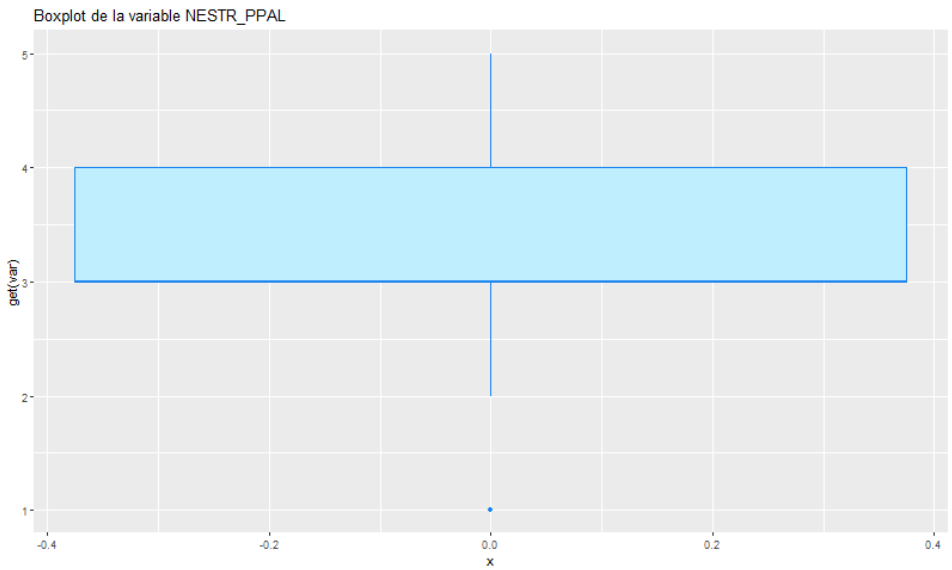

Boxplot de la variable ACTI_VISITASCULTU
1.050
1.025
1.000
0.975
0.950
get(var)
-0.4
-0.2
0.0
0.2
0.4
x

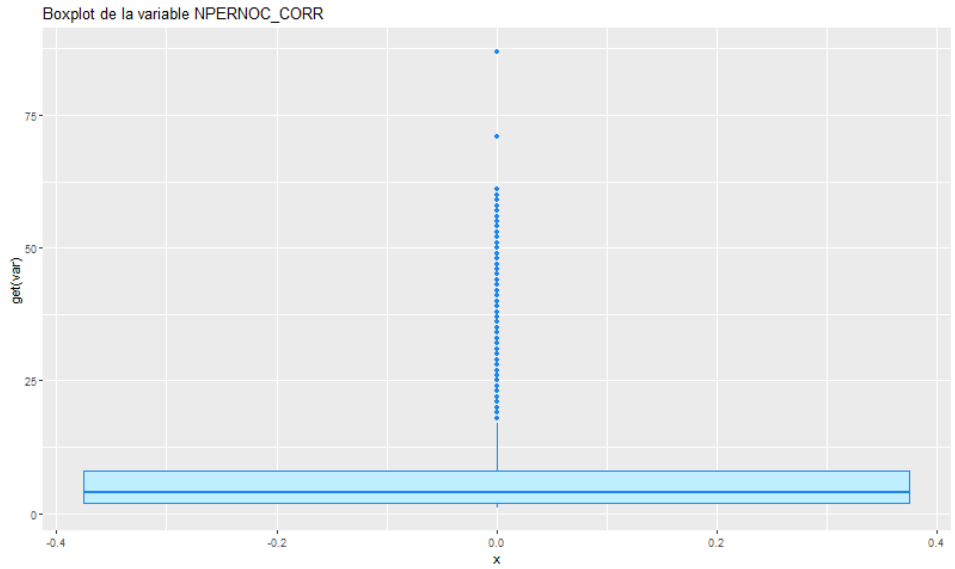

Boxplot de la variable EDAD
80
60
40
20
get(var)
-0.4
-0.2
0.0
0.2
0.4
x

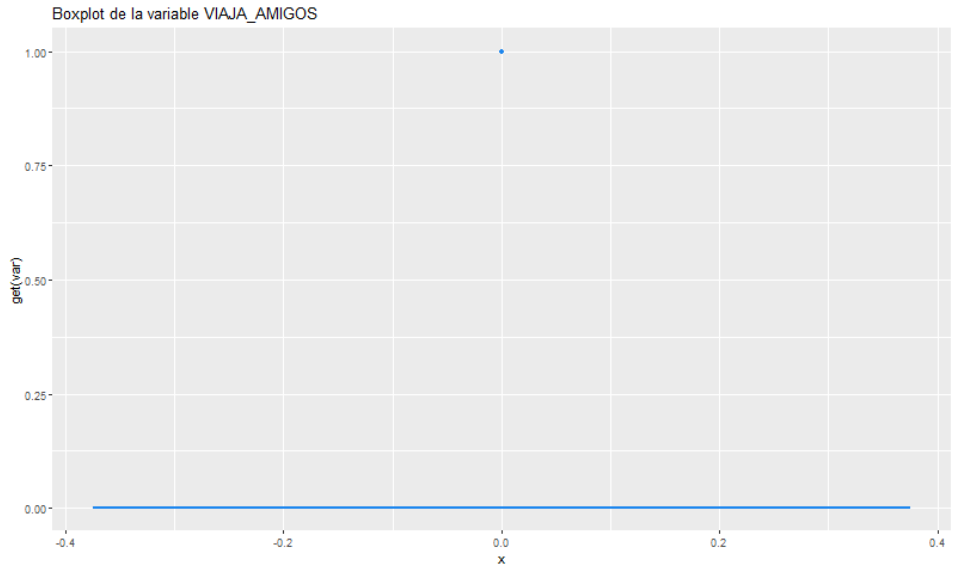

Boxplot de la variable GASTOFI_TOTAL
15000
10000
5000
0
get(var)
-0.4
-0.2
0.0
0.2
0.4
x

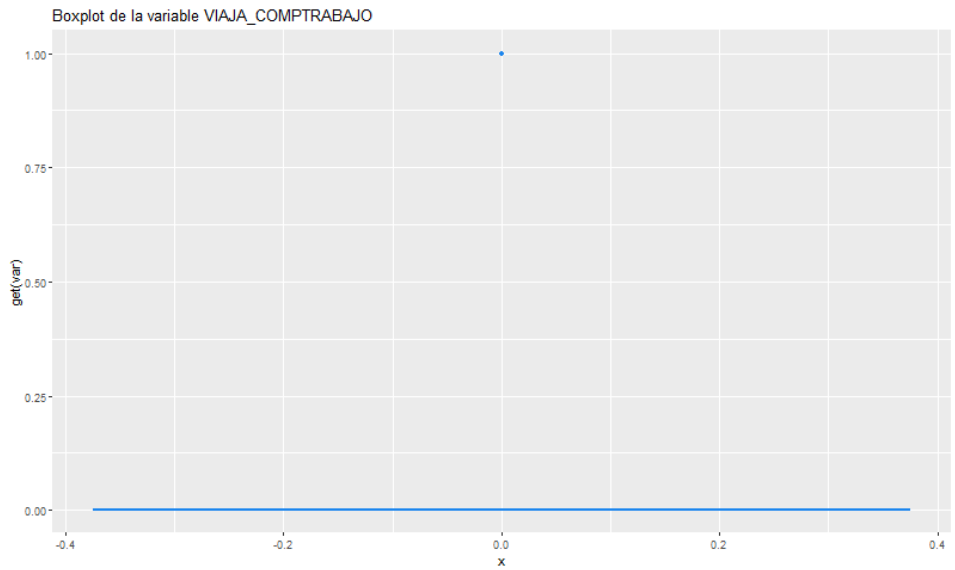

Boxplot de la variable GRADO_SATISF
10.0
7.5
5.0
2.5
0.0
get(var)
-0.4
-0.2
0.0
0.2
0.4
x

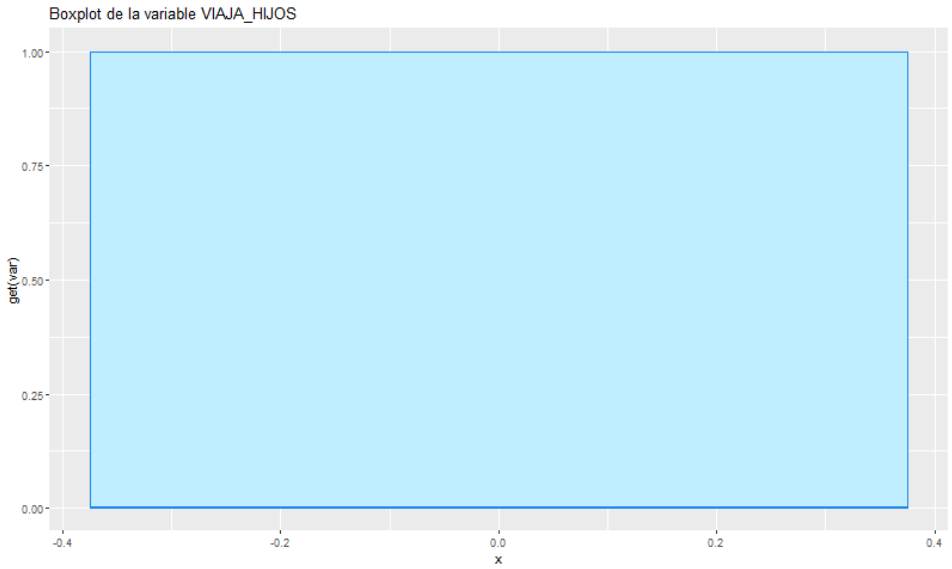

Boxplot de la variable MIEMV_15MENOS
7.5
5.0
2.5
0.0
get(var)
-0.4
-0.2
0.0
0.2
0.4
x

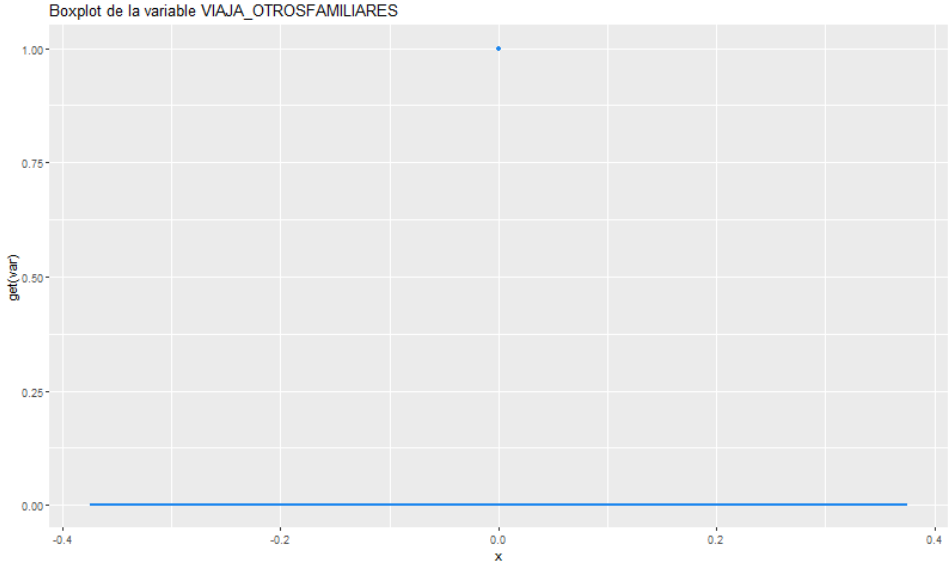

Boxplot de la variable NESTR_PPAL
5
4
3
2
1
get(var)
-0.4
-0.2
0.0
0.2
0.4
x

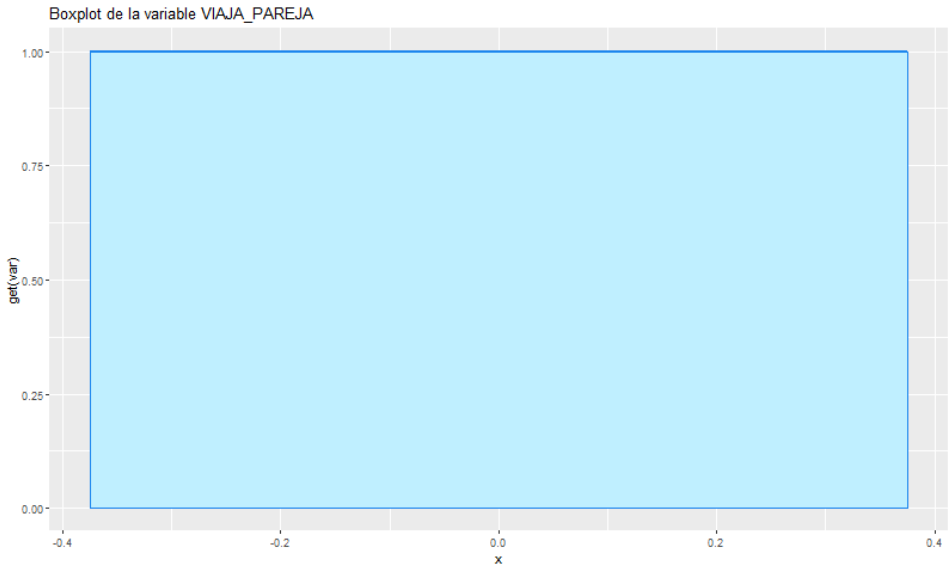

Boxplot de la variable NPERNOC_CORR
get(var)
75
50
25
0
-0.4
-0.2
0.0
0.2
0.4
x

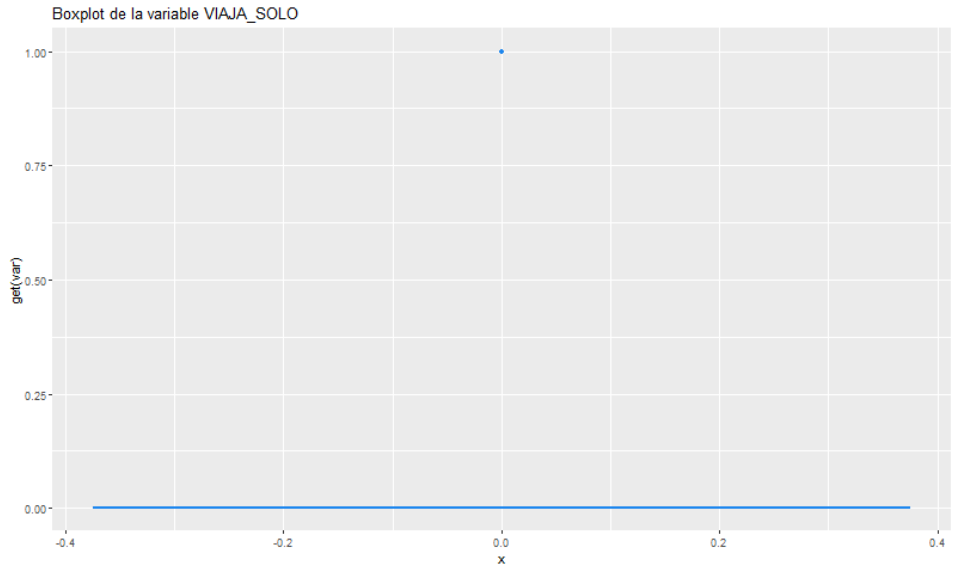

Boxplot de la variable VIAJA_AMIGOS
get(var)
1.00
0.75
0.50
0.25
0.00
-0.4
-0.2
0.0
0.2
0.4
x

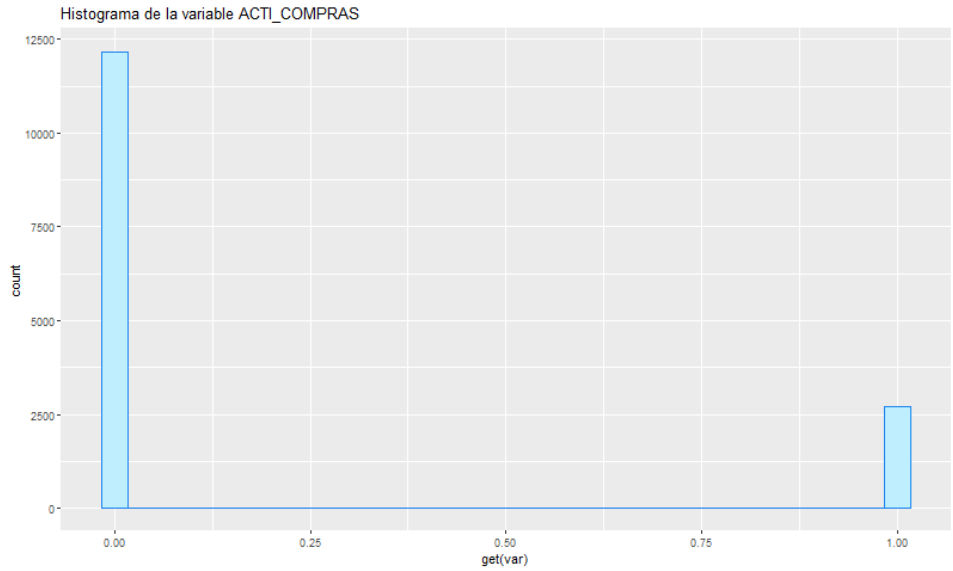

Boxplot de la variable VIAJA_COMPTRABAJO
get(var)
1.00
0.75
0.50
0.25
0.00
-0.4
-0.2
0.0
0.2
0.4
x

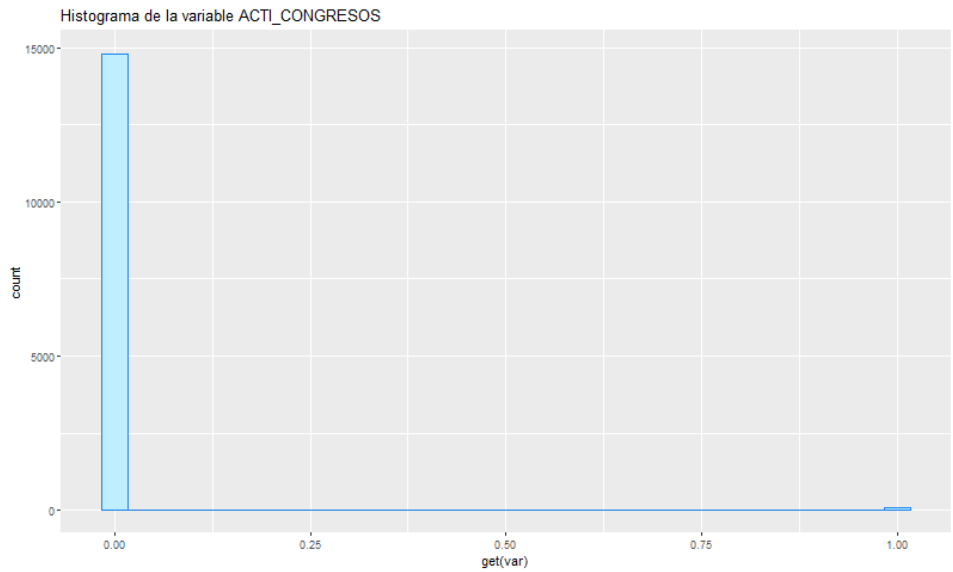

Boxplot de la variable VIAJA_HIJOS
1.00
0.75
0.50
0.25
0.00
get(var)
-0.4
-0.2
0.0
0.2
0.4
x

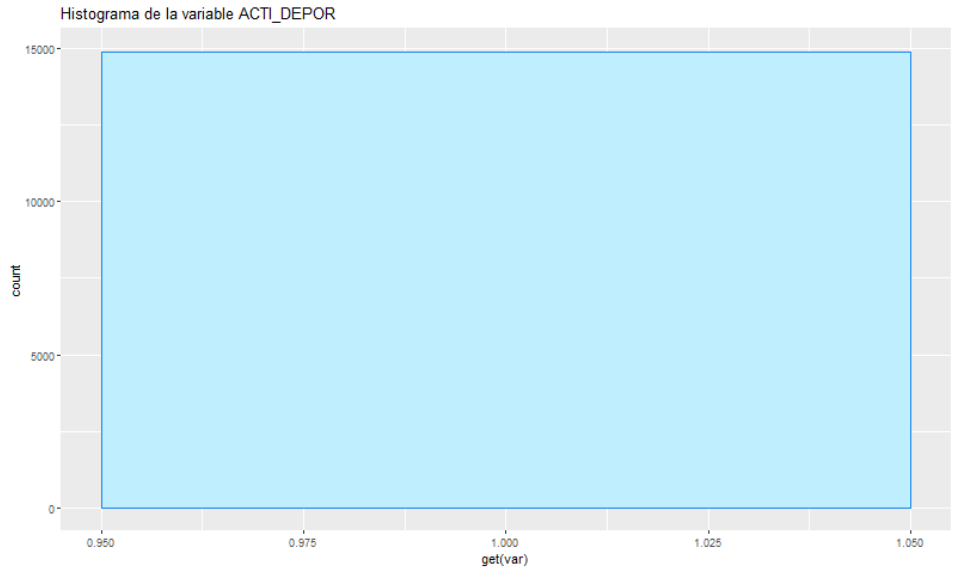

Boxplot de la variable VIAJA_OTROSFAMILIARES
1.00
0.75
0.50
0.25
0.00
get(var)
-0.4
-0.2
0.0
0.2
0.4
x

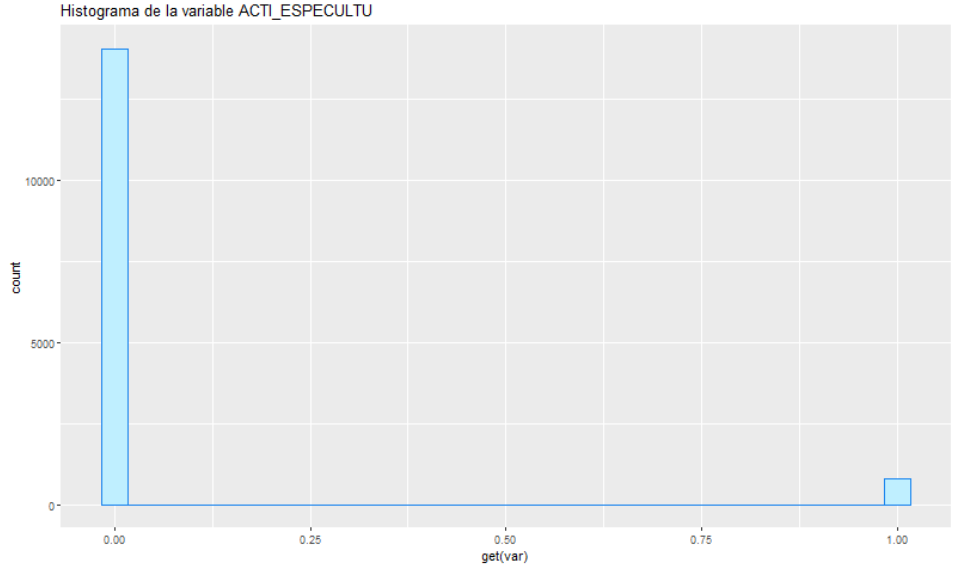

Boxplot de la variable VIAJA_PAREJA
1.00
0.75
0.50
0.25
0.00
get(var)
-0.4
-0.2
0.0
0.2
0.4
x

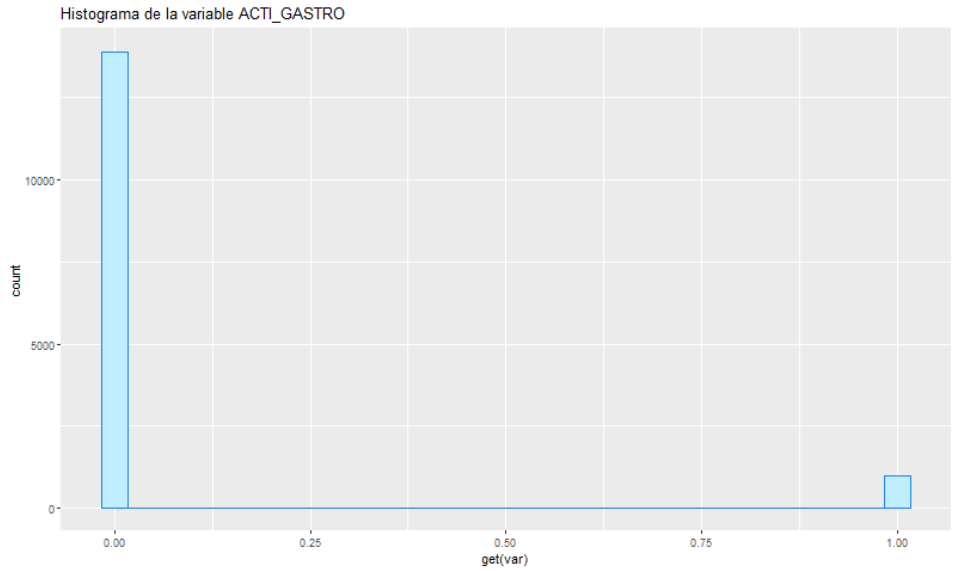

Boxplot de la variable VIAJA_SOLO
get(var)
1.00
0.75
0.50
0.25
0.00
-0.4
-0.2
0.0
0.2
0.4
x


Histograma de la variable ACTI_COMPRAS
count
12500
10000
7500
5000
2500
0
0.00
0.25
0.50
0.75
1.00
get(var)

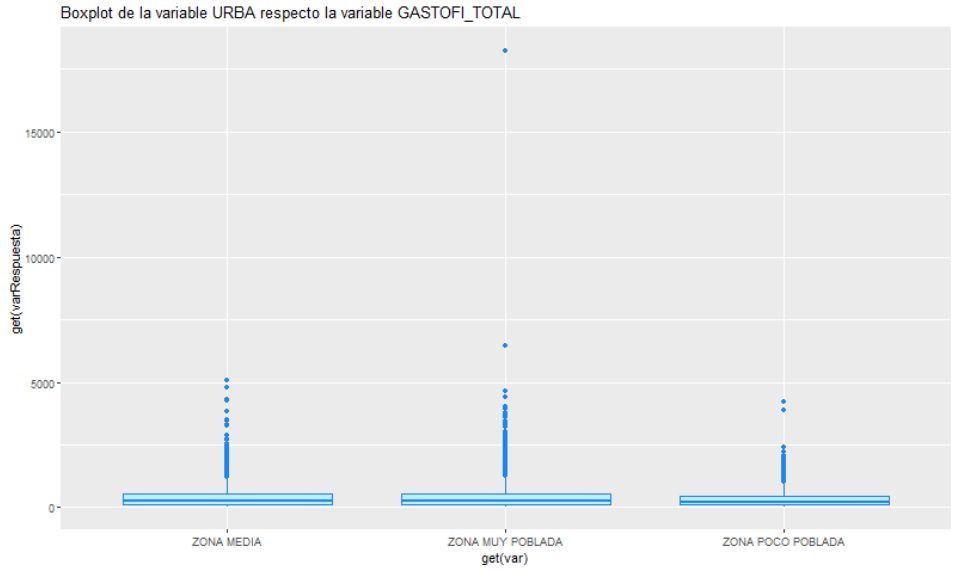

Histograma de la variable ACTI_CONGRESOS
count
15000
10000
5000
0
0.00
0.25
0.50
0.75
1.00
get(var)

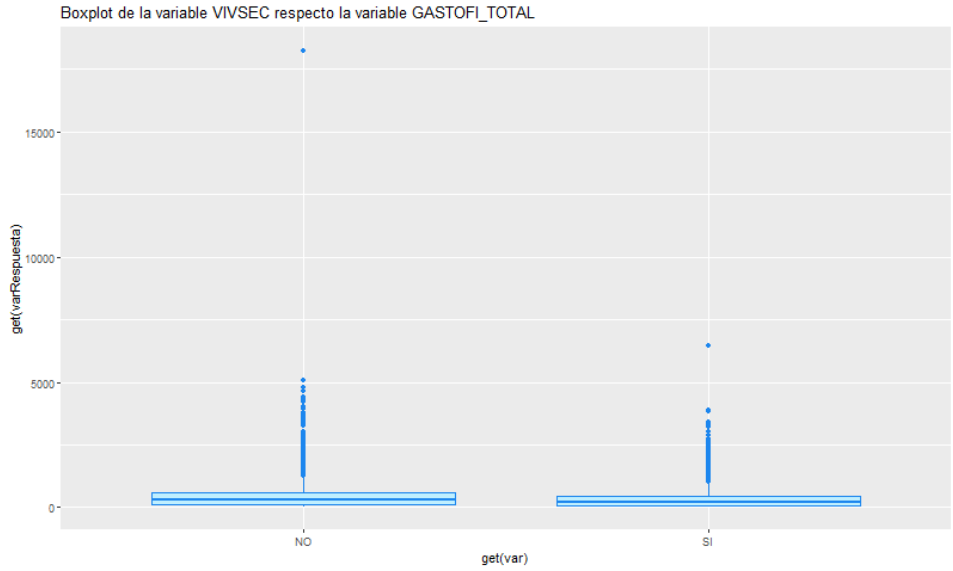

Histograma de la variable ACTI_DEPOR
15000
10000
5000
0
count
0.950
0.975
1.000
1.025
1.050
get(var)

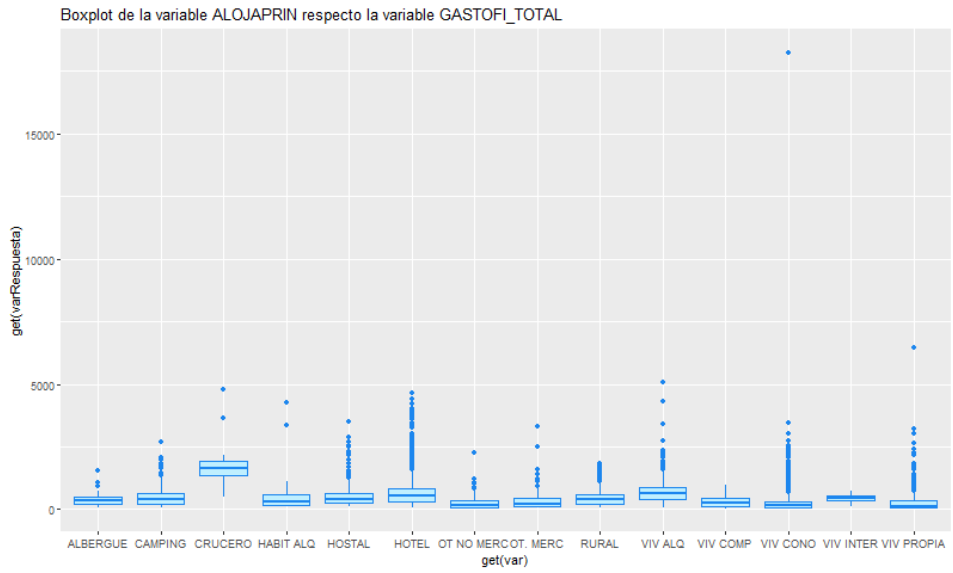

Histograma de la variable ACTI_ESPECULTU
10000
5000
0
count
0.00
0.25
0.50
0.75
1.00
get(var)

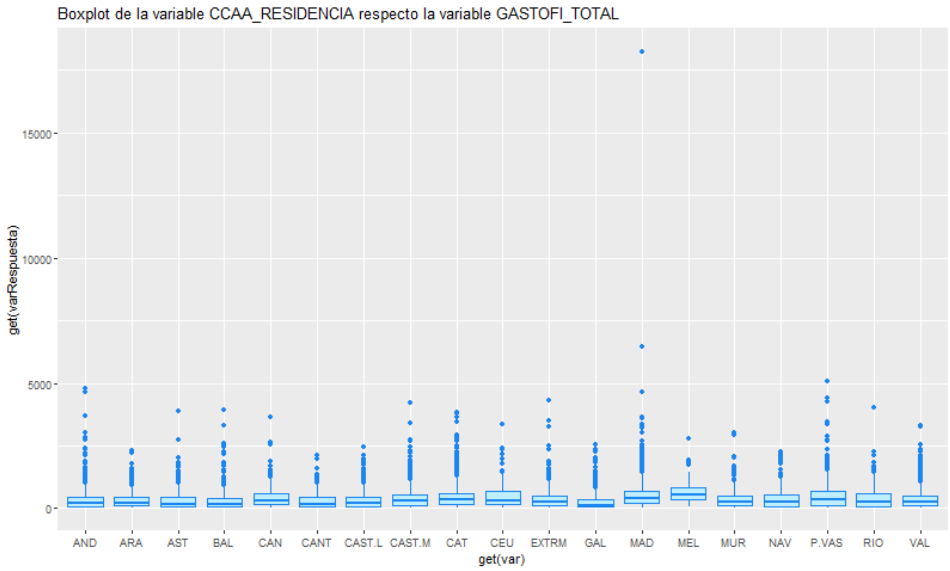

Histograma de la variable ACTI_GASTRO
10000
5000
0
count
0.00
0.25
0.50
0.75
1.00
get(var)

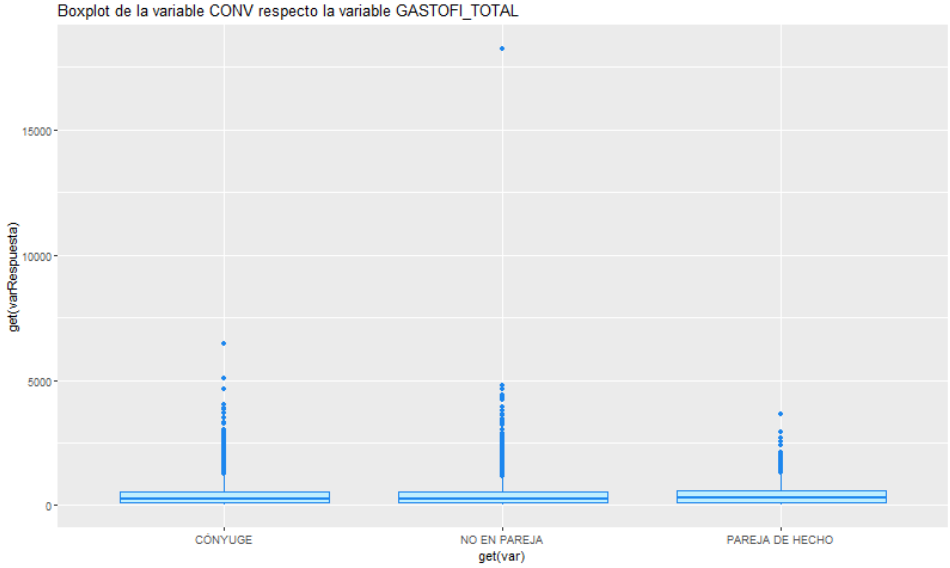
Gráficos bivariantes, variables categóricas contra gasto total:

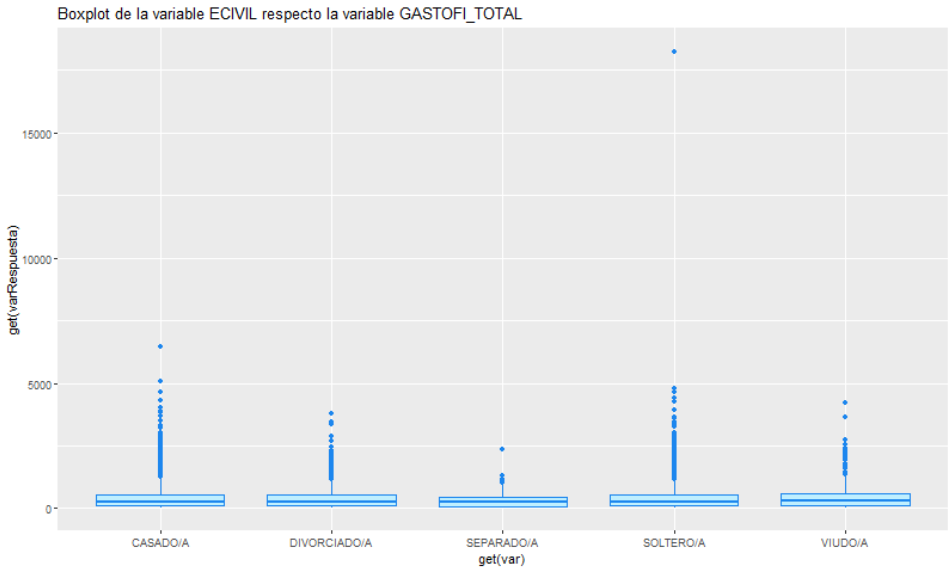


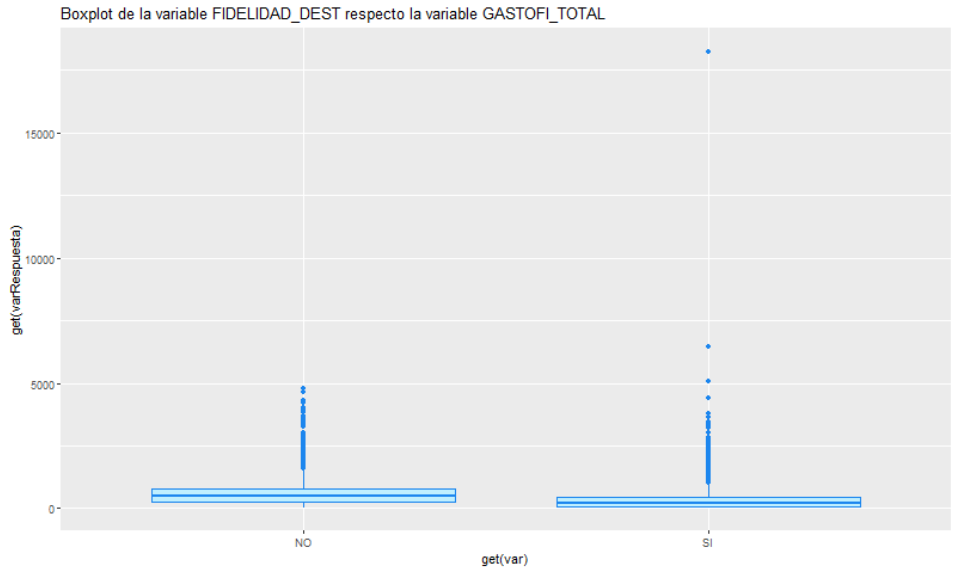



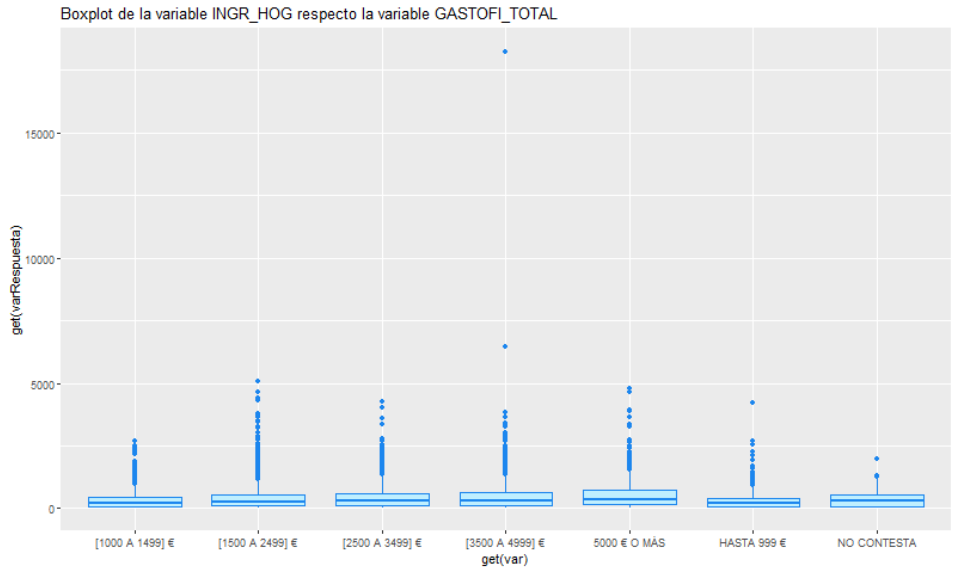



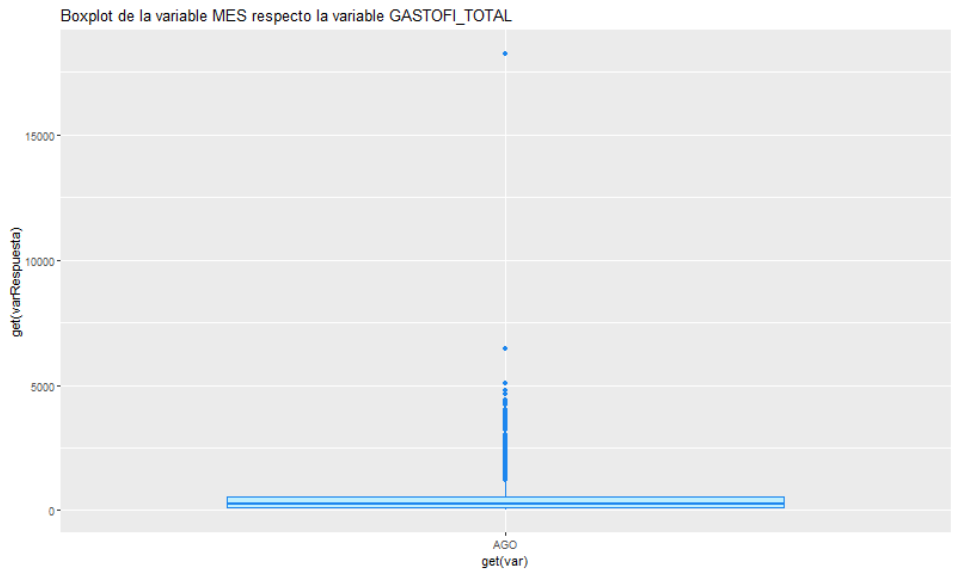

Boxplot de la variable CCAA_RESIDENCIA respecto la variable GASTOFI_TOTAL
get(varRespuesta)
15000
10000
5000
0
AND
ARA
AST
BAL
CAN
CANT
CAST.L
CAST.M
CAT
CEU
EXTRM
GAL
MAD
MEL
MUR
NAV
P.VAS
RIO
VAL
get(var)

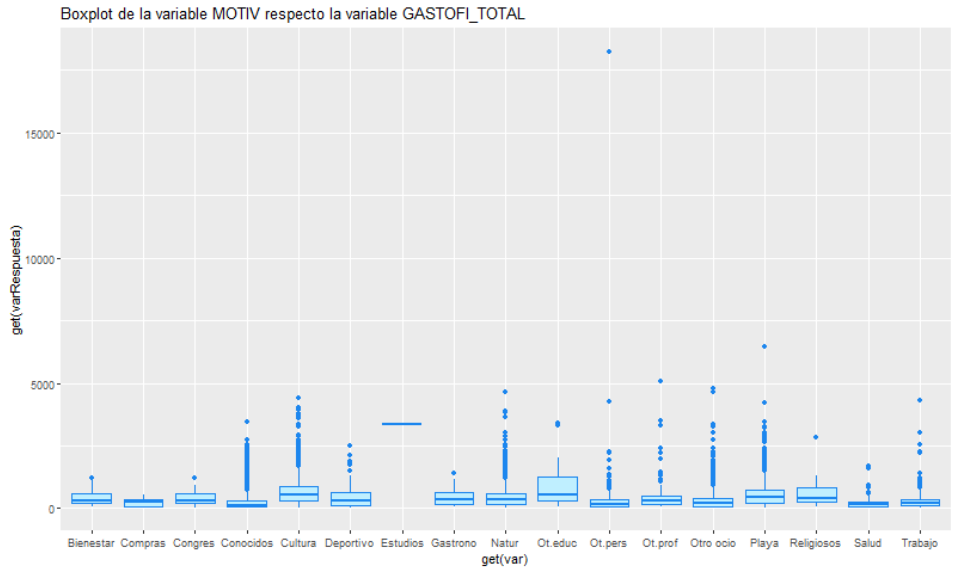

Boxplot de la variable CONV respecto la variable GASTOFI_TOTAL
get(varRespuesta)
15000
10000
5000
0
CÓNYUGE
NO EN PAREJA
PAREJA DE HECHO
get(var)

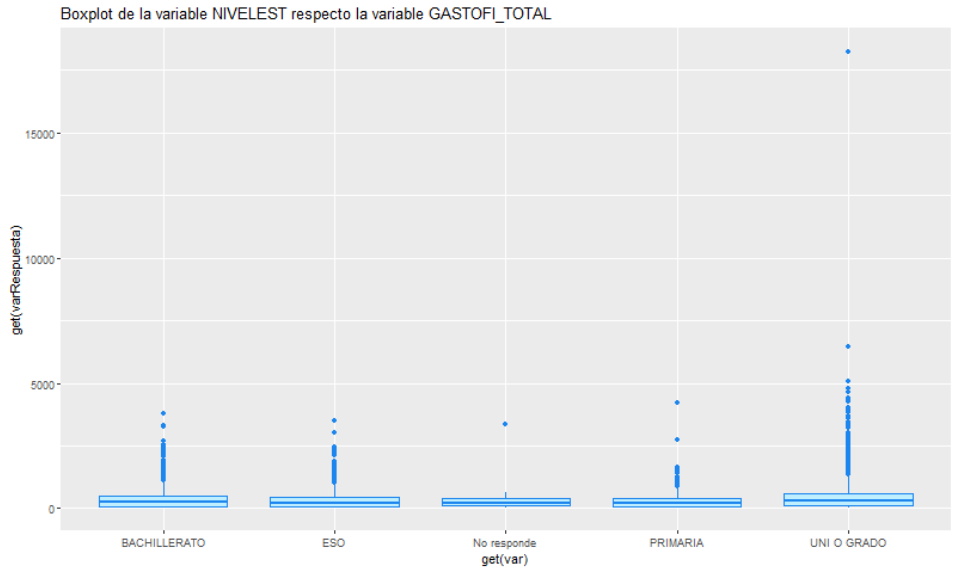

Boxplot de la variable ECIVIL respecto la variable GASTOFI_TOTAL
get(varRespuesta)
15000
10000
5000
0
CASADO/A
DIVORCIADO/A
SEPARADO/A
SOLTERO/A
VIUDO/A
get(var)

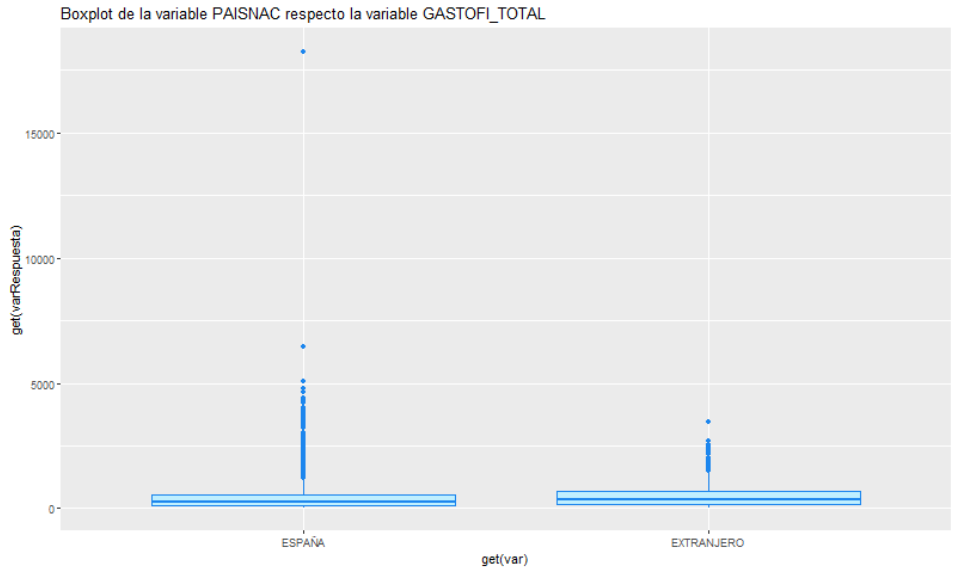

Boxplot de la variable FIDELIDAD_DEST respecto la variable GASTOFI_TOTAL
15000
10000
5000
0
get(varRespuesta)
NO
SI
get(var)

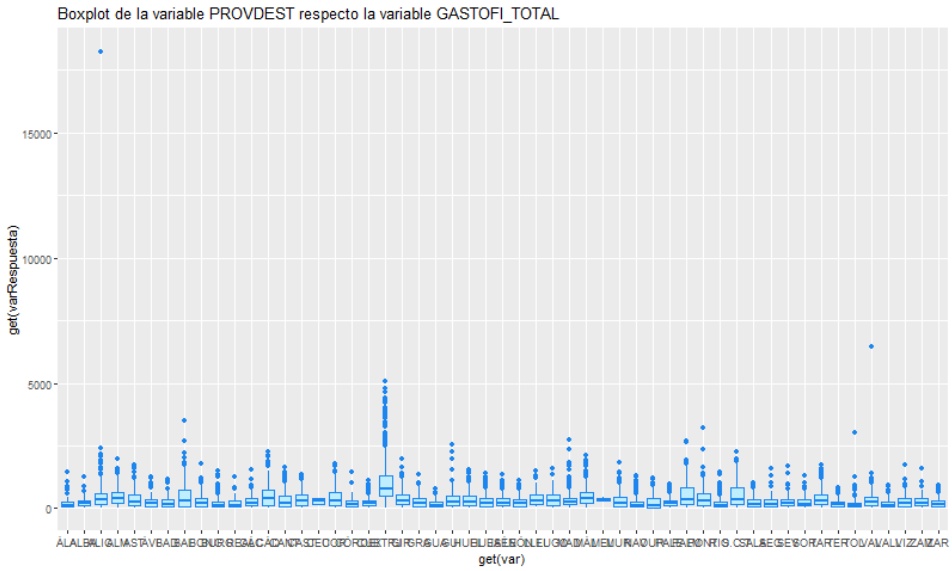

Boxplot de la variable INGR_HOG respecto la variable GASTOFI_TOTAL
15000
10000
5000
0
get(varRespuesta)
[1000 A 1499] €
[1500 A 2499] €
[2500 A 3499] €
[3500 A 4999] €
5000 € O MÁS
HASTA 999 €
NO CONTESTA
get(var)

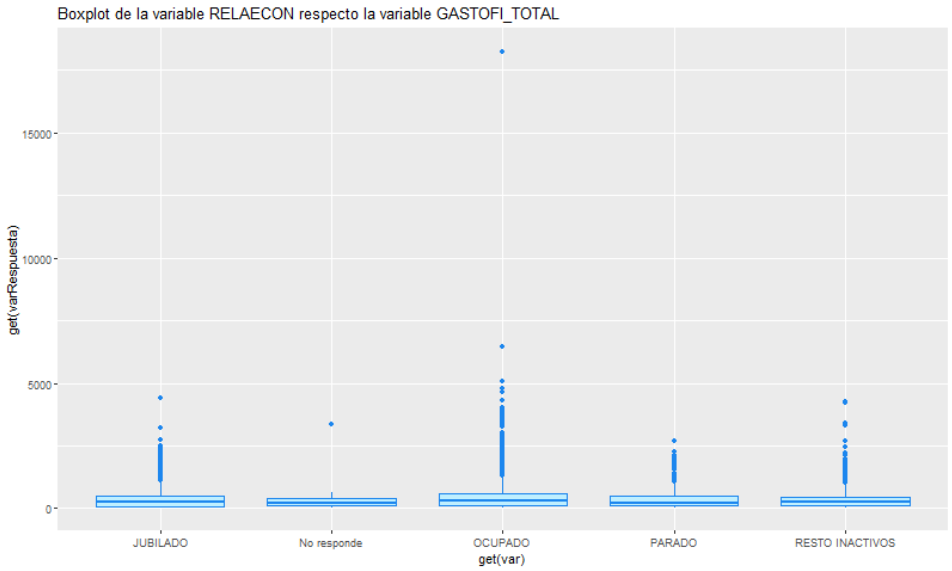

Boxplot de la variable MES respecto la variable GASTOFI_TOTAL
15000
10000
5000
0
get(varRespuesta)
AGO
get(var)

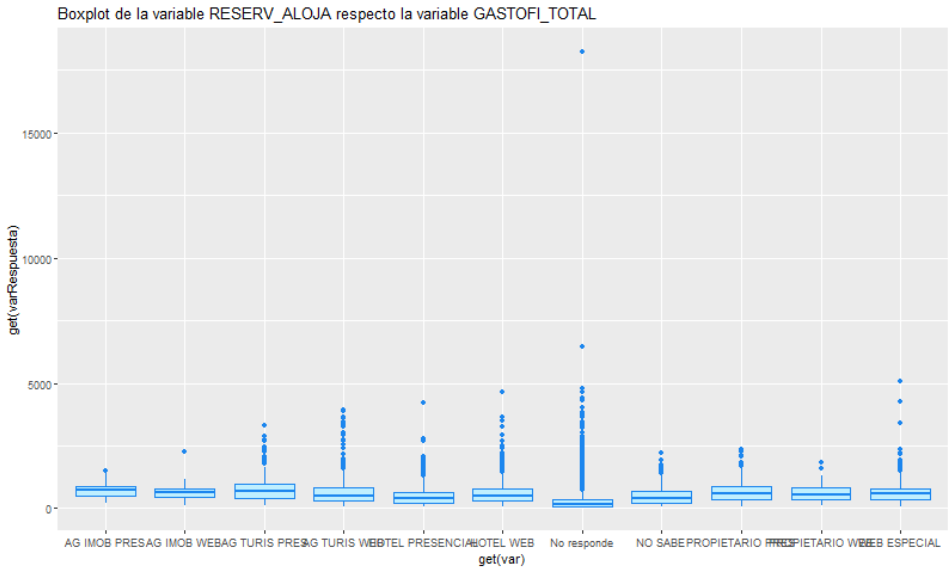

Boxplot de la variable MOTIV respecto la variable GASTOFI_TOTAL
get(varRespuesta)
15000
10000
5000
0
Bienestar
Compras
Congres
Conocidos
Cultura
Deportivo
Estudios
Gastrono
Natur
Ot.educ
Ot.pers
Ot.prof
Otro ocio
Playa
Religiosos
Salud
Trabajo
get(var)

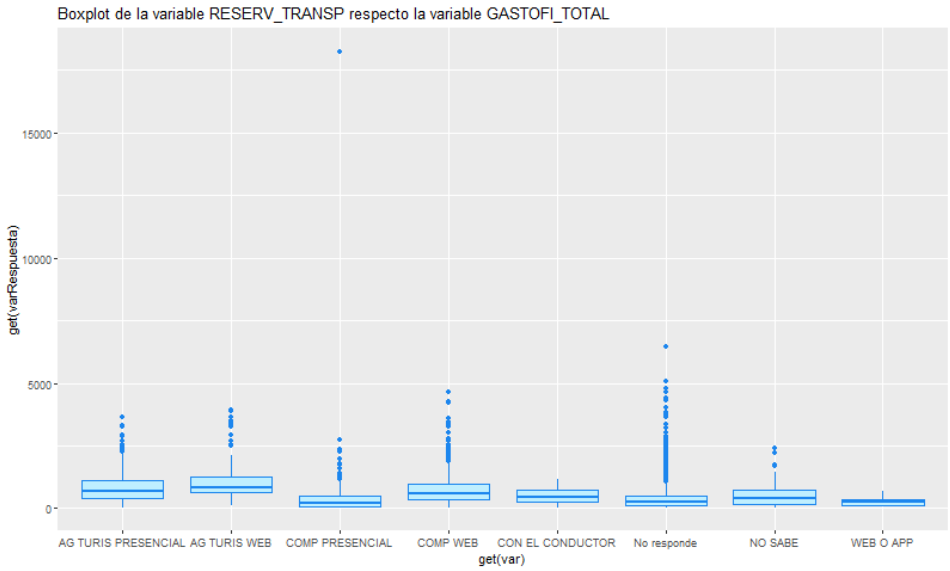

Boxplot de la variable NIVELEST respecto la variable GASTOFI_TOTAL
get(varRespuesta)
15000
10000
5000
0
BACHILLERATO
ESO
No responde
PRIMARIA
UNI O GRADO
get(var)

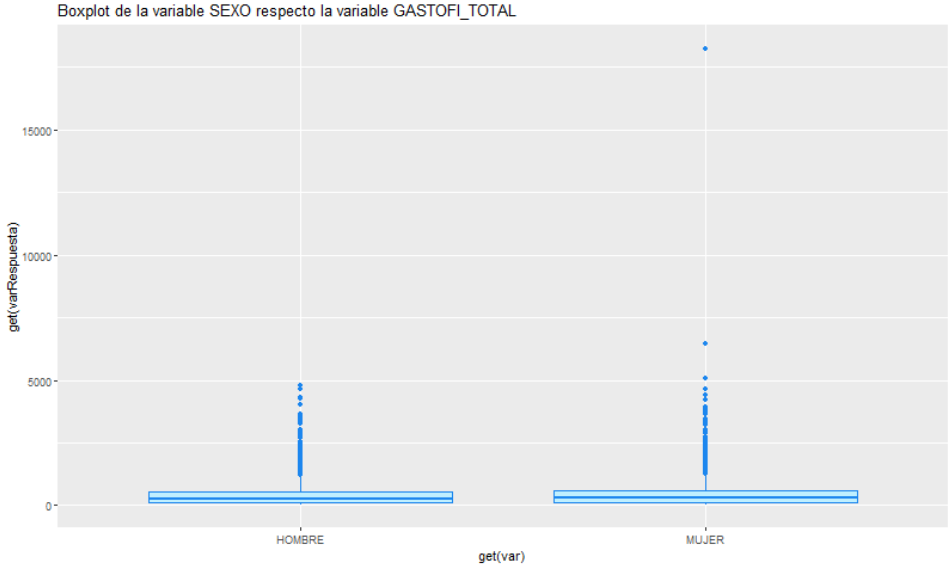

Boxplot de la variable PAISNAC respecto la variable GASTOFI_TOTAL
get(varRespuesta)
15000
10000
5000
0
ESPAÑA
EXTRANJERO
get(var)

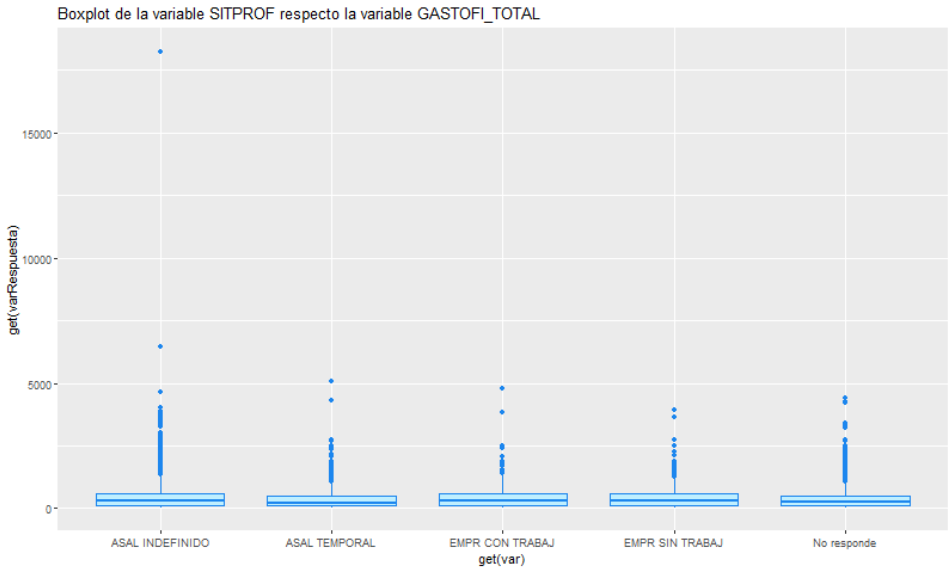

Boxplot de la variable PROVDEST respecto la variable GASTOFI_TOTAL
get(varRespuesta)
15000
10000
5000
0
get(var)

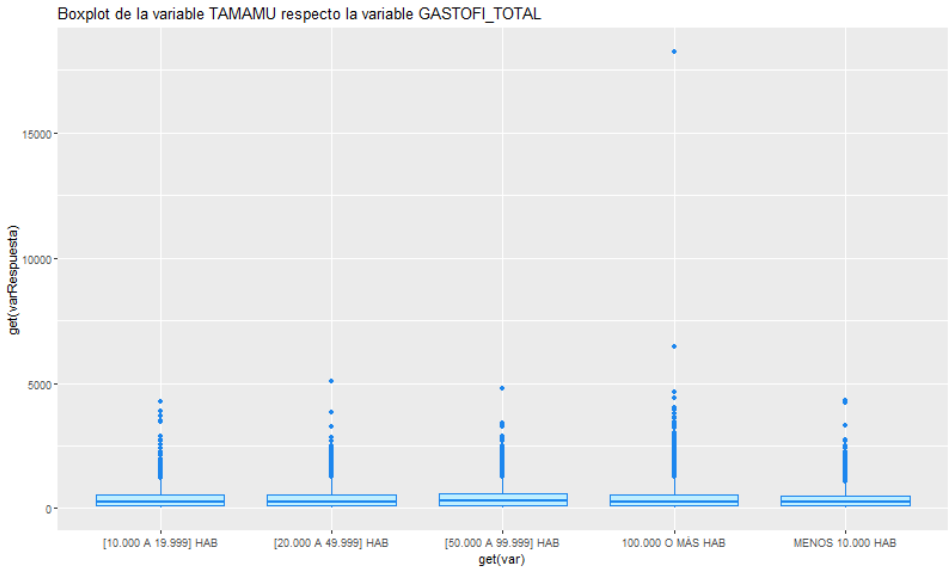

Boxplot de la variable RELAECON respecto la variable GASTOFI_TOTAL
get(varRespuesta)
15000
10000
5000
0
JUBILADO
No responde
OCUPADO
PARADO
RESTO INACTIVOS
get(var)

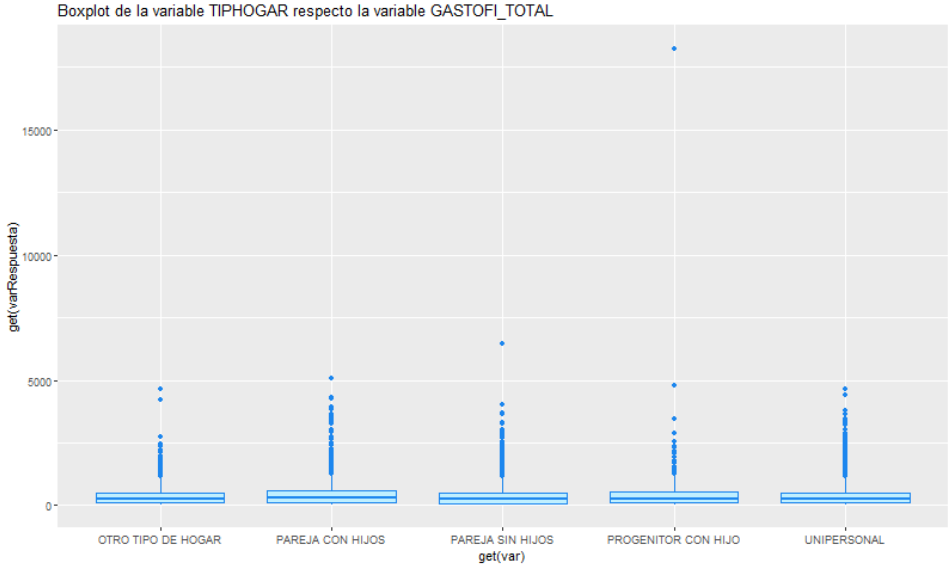

Boxplot de la variable RESERV_ALOJA respecto la variable GASTOFI_TOTAL
get(varRespuesta)
15000
10000
5000
0
No responde
get(var)

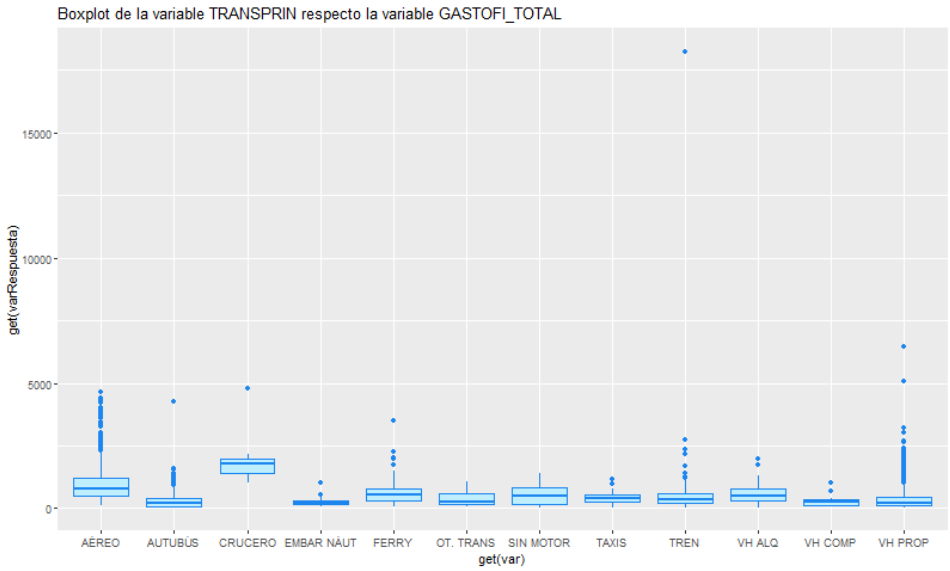

Boxplot de la variable RESERV_TRANSP respecto la variable GASTOFI_TOTAL
15000
10000
5000
0
get(varRespuesta)
AG TURIS PRESENCIAL
AG TURIS WEB
COMP PRESENCIAL
COMP WEB
CON EL CONDUCTOR
No responde
NO SABE
WEB O APP
get(var)


Boxplot de la variable SEXO respecto la variable GASTOFI_TOTAL
15000
10000
5000
0
get(varRespuesta)
HOMBRE
MUJER
get(var)

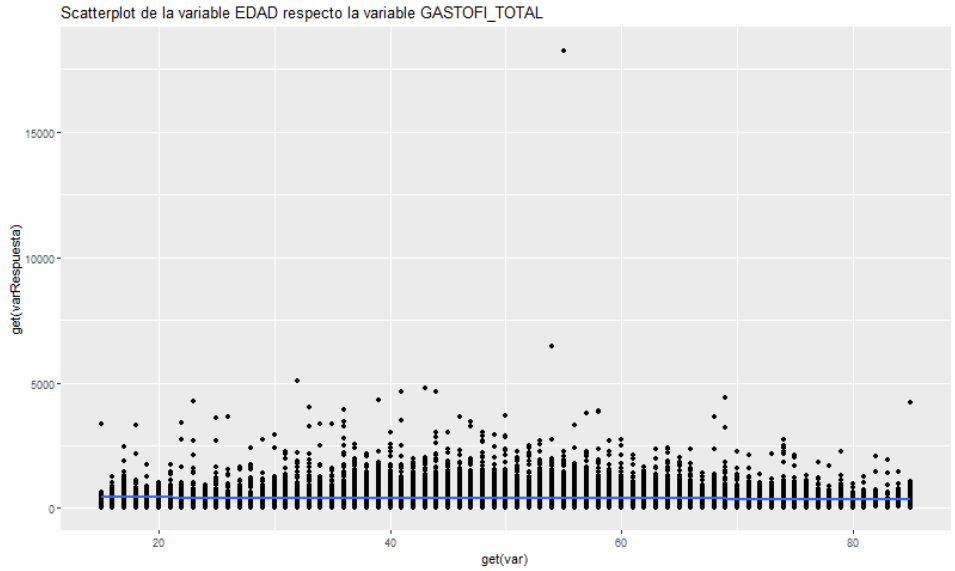

Boxplot de la variable SITPROF respecto la variable GASTOFI_TOTAL
15000
10000
5000
0
get(varRespuesta)
ASAL INDEFINIDO
ASAL TEMPORAL
EMPR CON TRABAJ
EMPR SIN TRABAJ
No responde
get(var)

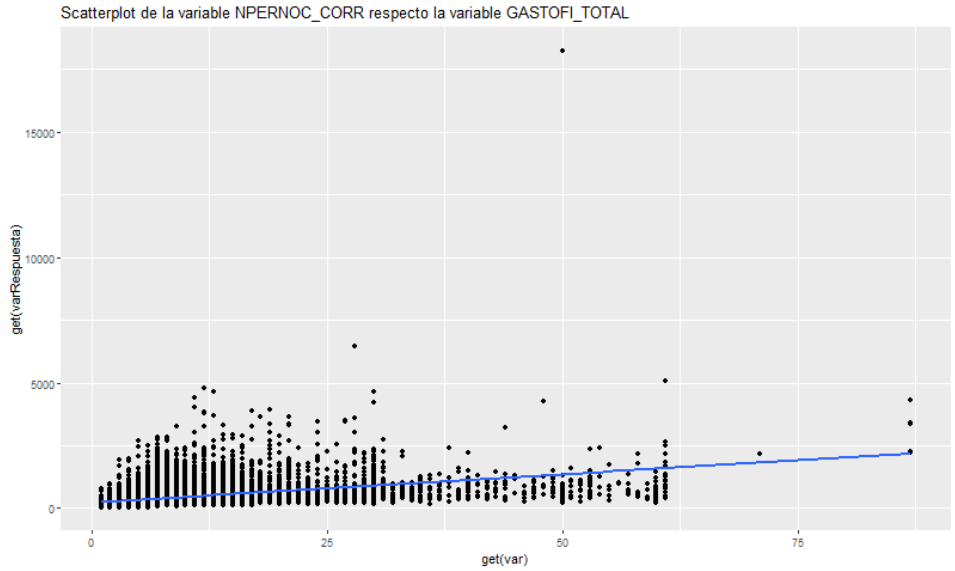

Boxplot de la variable TAMAMU respecto la variable GASTOFI_TOTAL
get(varRespuesta)
15000
10000
5000
0
[10.000 A 19.999] HAB
[20.000 A 49.999] HAB
[50.000 A 99.999] HAB
100.000 O MÁS HAB
MENOS 10.000 HAB
get(var)

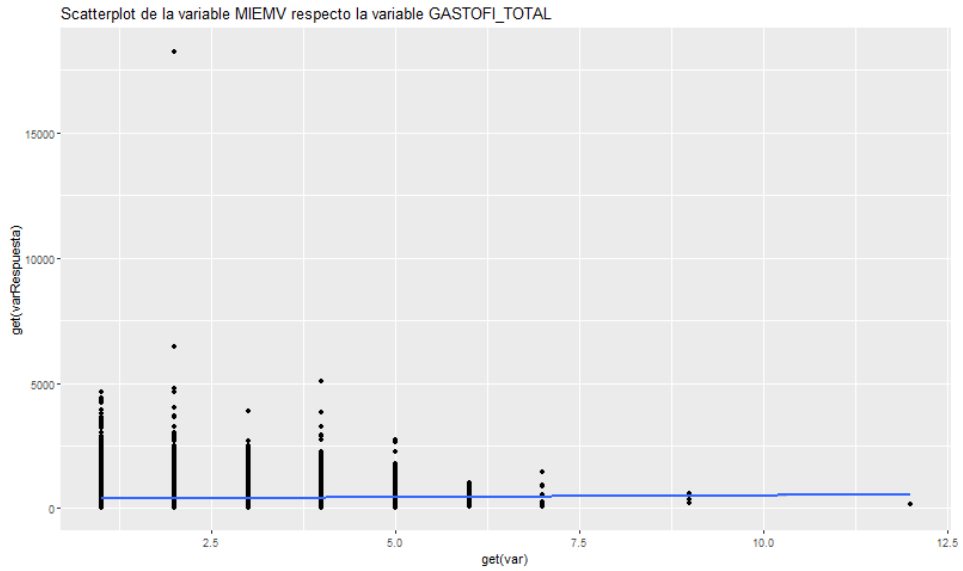

Boxplot de la variable TIPHOGAR respecto la variable GASTOFI_TOTAL
get(varRespuesta)
15000
10000
5000
0
OTRO TIPO DE HOGAR
PAREJA CON HIJOS
PAREJA SIN HIJOS
PROGENITOR CON HIJO
UNIPERSONAL
get(var)


Boxplot de la variable TRANSPRIN respecto la variable GASTOFI_TOTAL
get(varRespuesta)
15000
10000
5000
0
AÉREO
AUTUBÚS
CRUCERO
EMBAR NÁUT
FERRY
OT. TRANS
SIN MOTOR
TAXIS
TREN
VH ALQ
VH COMP
VH PROP
get(var)

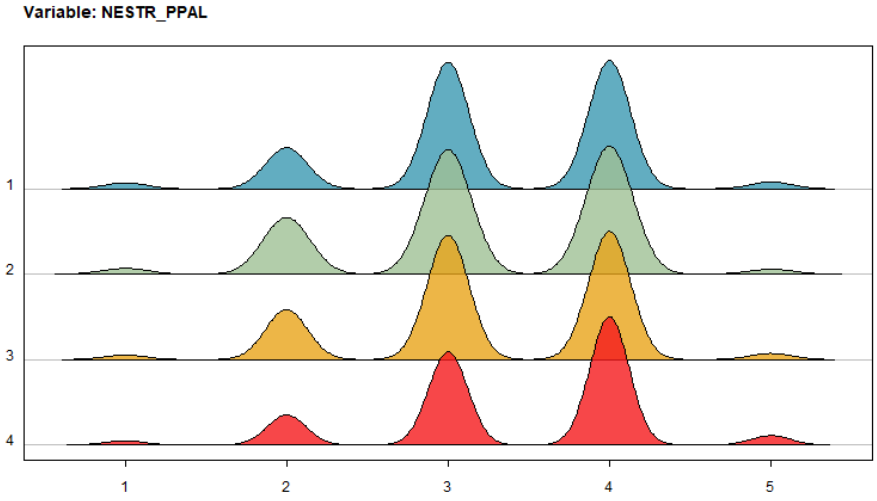
Gráficos bivariantes, variables numéricas contra gasto total:

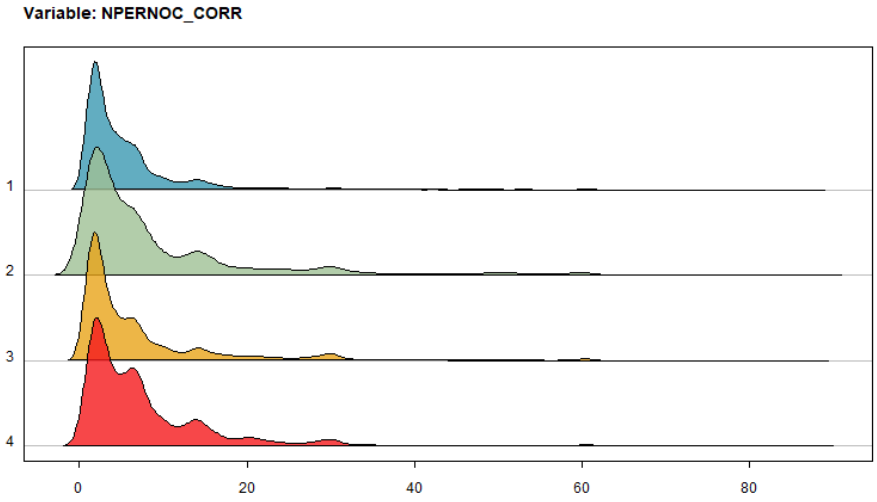


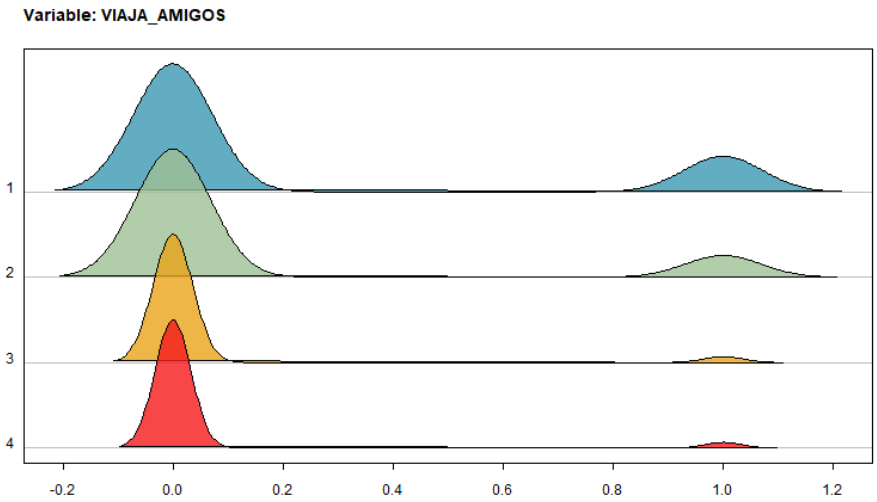



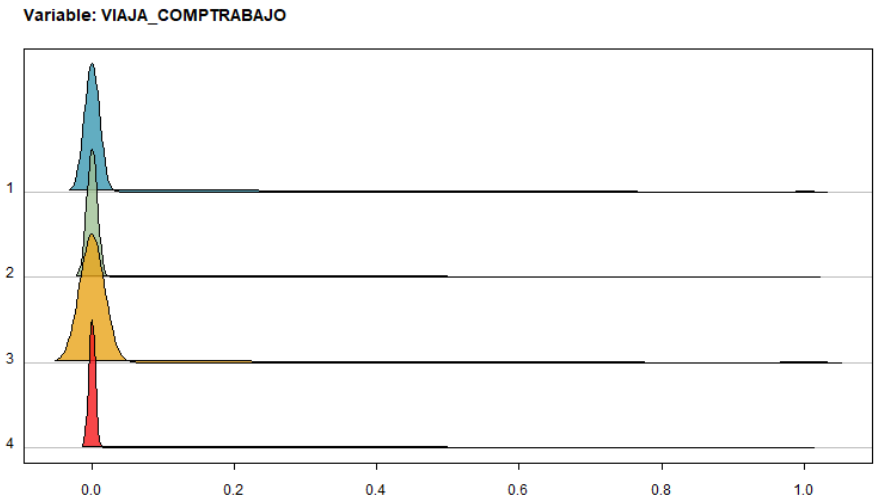



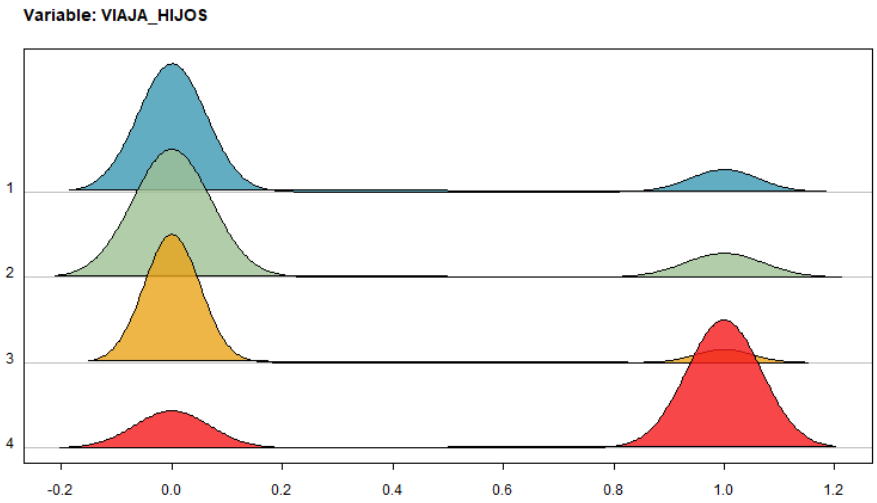
Clustering-Profiling:

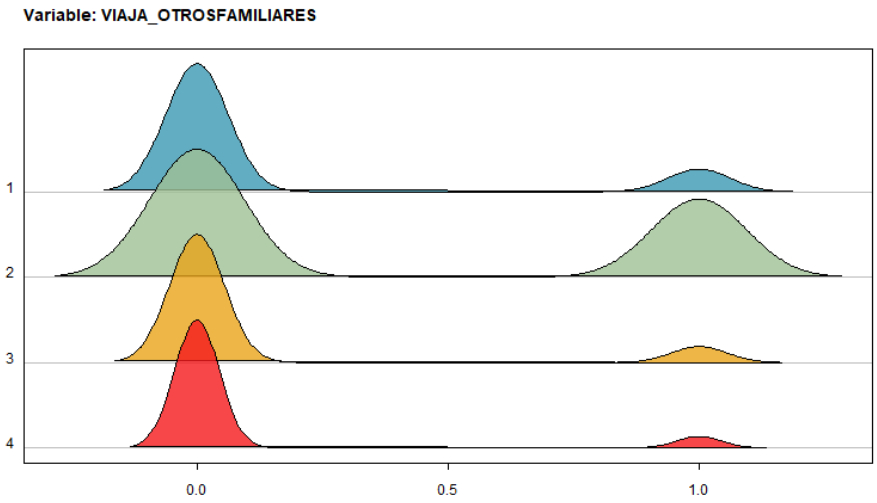


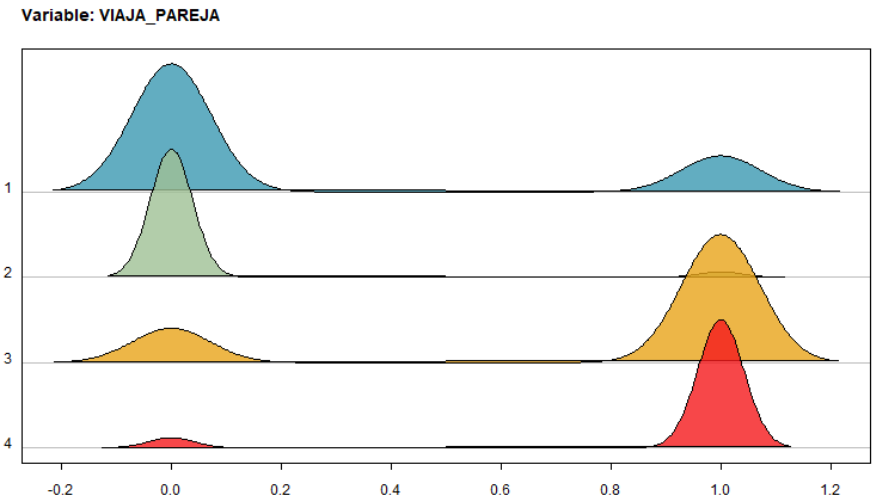



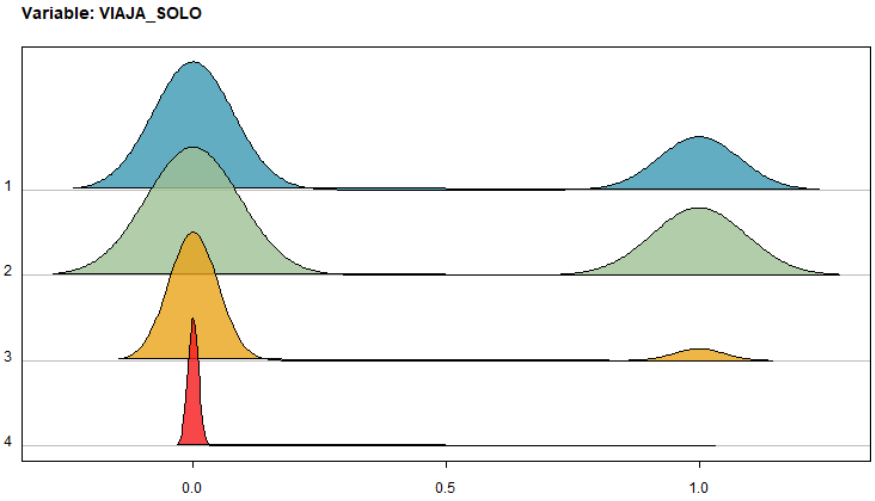



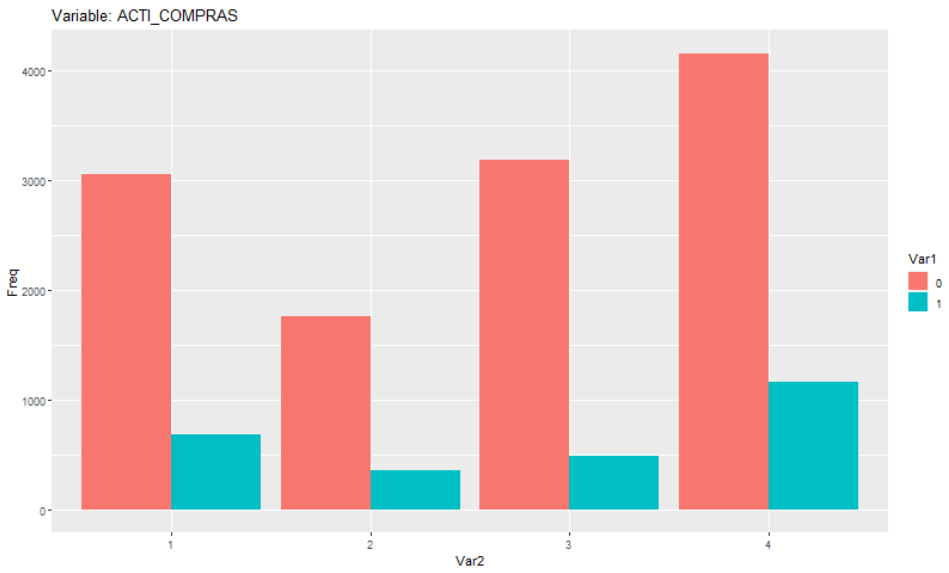

Variable: VIAJA_COMPTRABAJO
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0

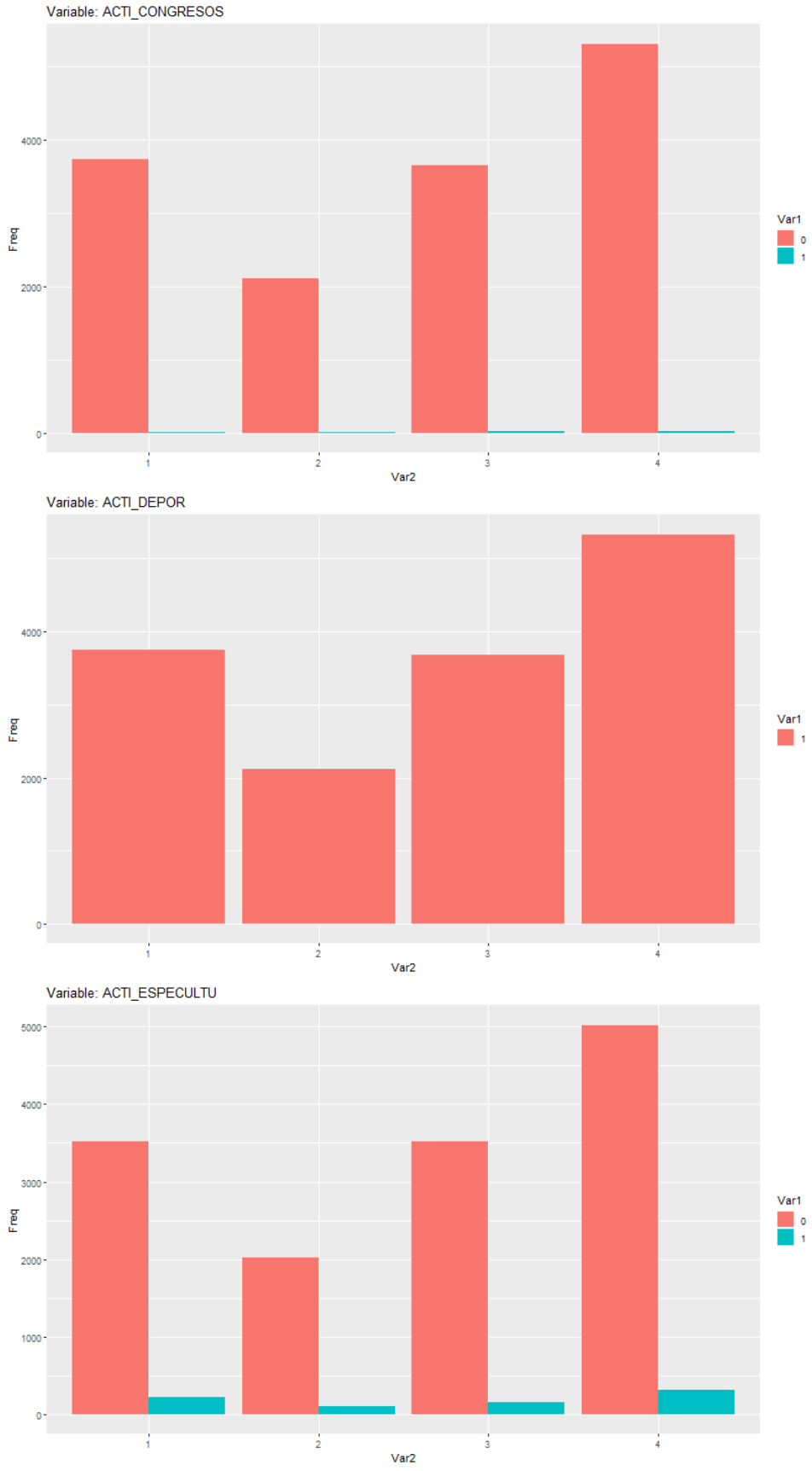

Variable: VIAJA_HIJOS
1
2
3
4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2

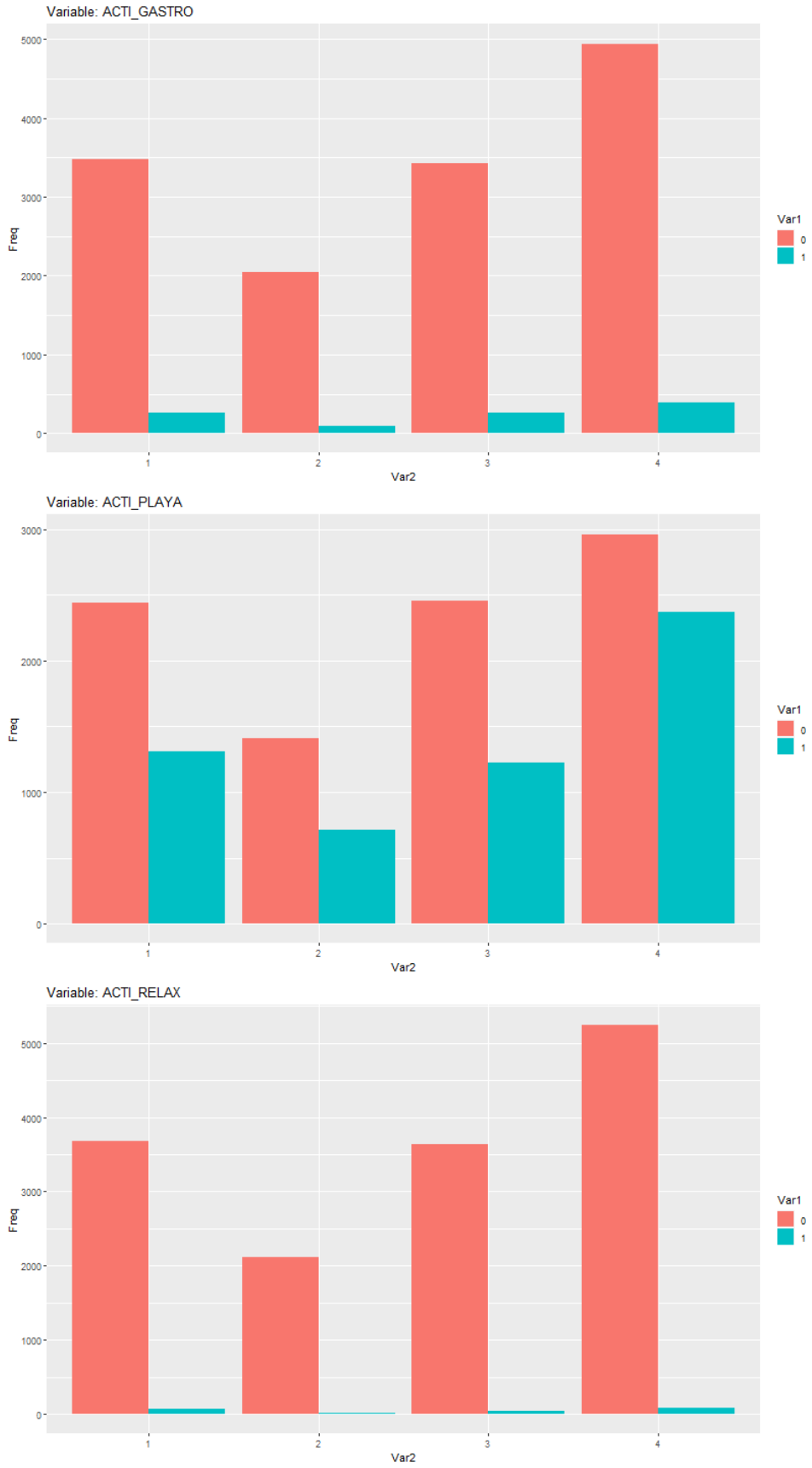

Variable: VIAJA_OTROSFAMILIARES
1
2
3
4
0.0
0.5
1.0


Variable: VIAJA_PAREJA
1
2
3
4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2

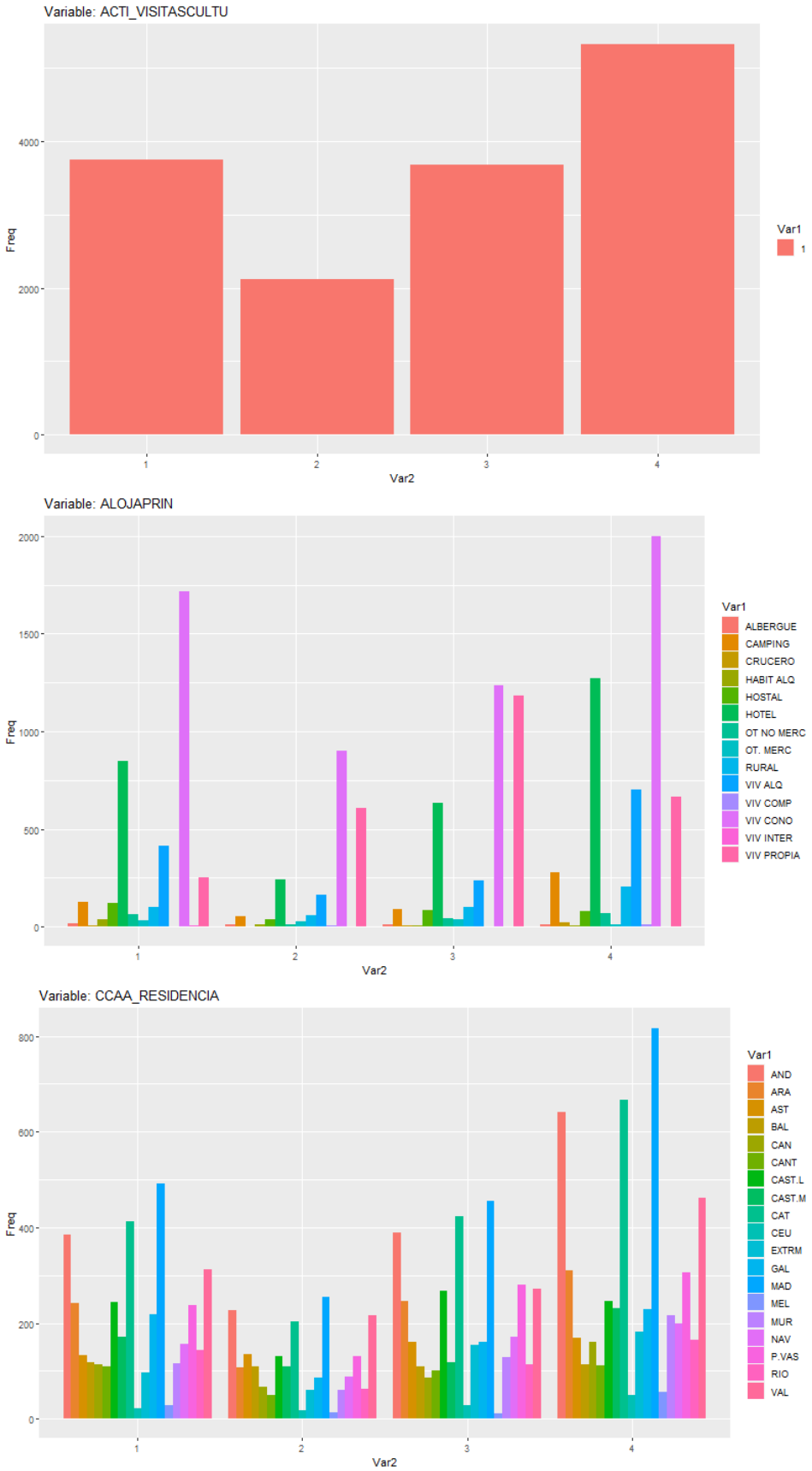

Variable: VIAJA_SOLO
1
2
3
4
0.0
0.5
1.0

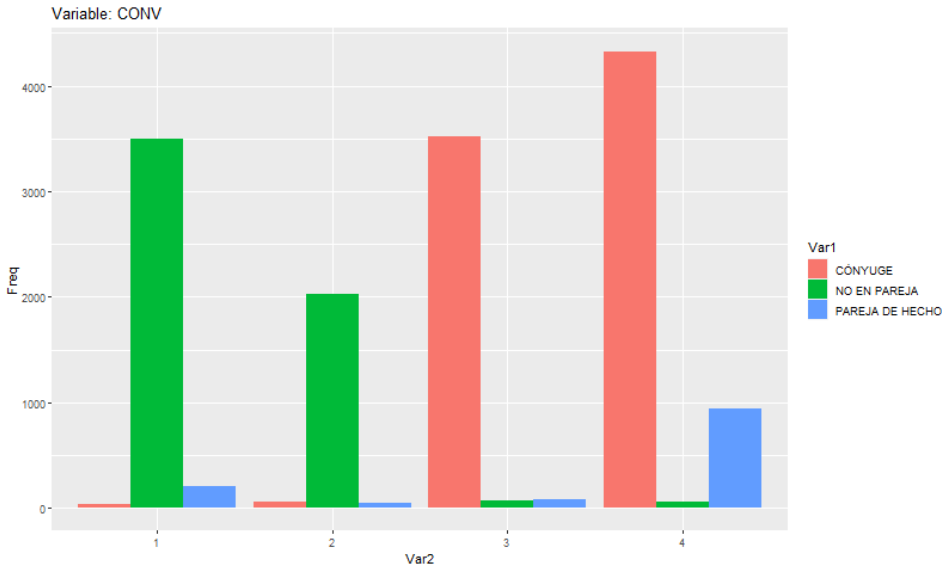

Variable: ACTI_COMPRAS
Freq
4000
3000
2000
1000
0
1
2
3
4
Var2
Var1
0
1

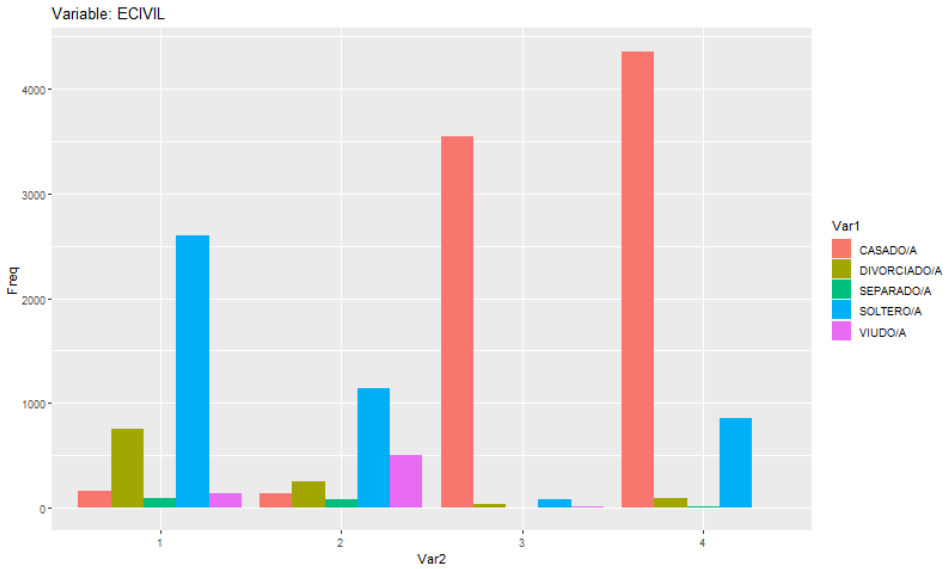

Variable: ACTI_CONGRESOS
Freq
4000
2000
0
1
2
3
4
Var2
Var1
0
1
Variable: ACTI_DEPOR
Freq
4000
2000
0
1
2
3
4
Var2
Var1
1
Variable: ACTI_ESPECULTU
Freq
5000
4000
3000
2000
1000
0
1
2
3
4
Var2
Var1
0
1

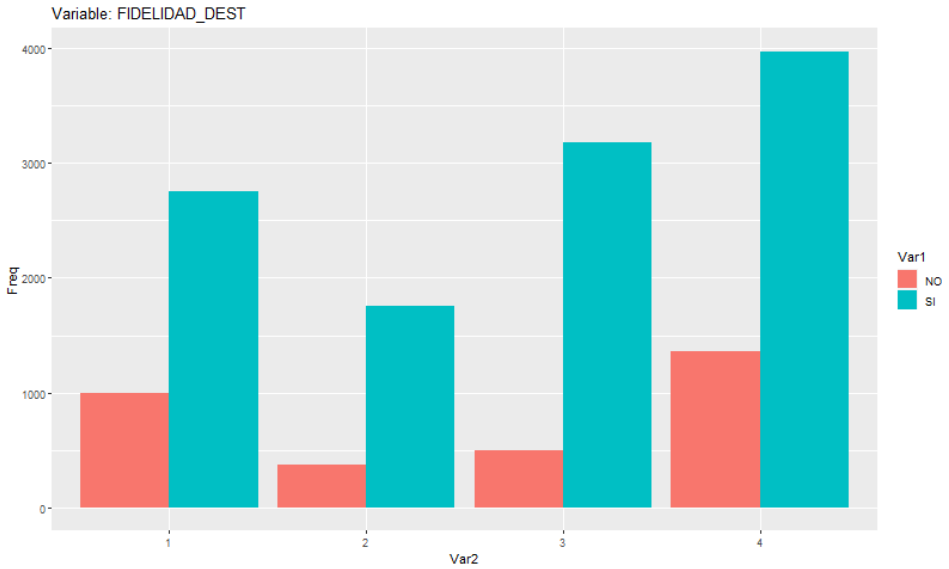

Variable: ACTI_GASTRO
Freq
Var2
Var1
0
1
Variable: ACTI_PLAYA
Freq
Var2
Var1
0
1
Variable: ACTI_RELAX
Freq
Var2
Var1
0
1

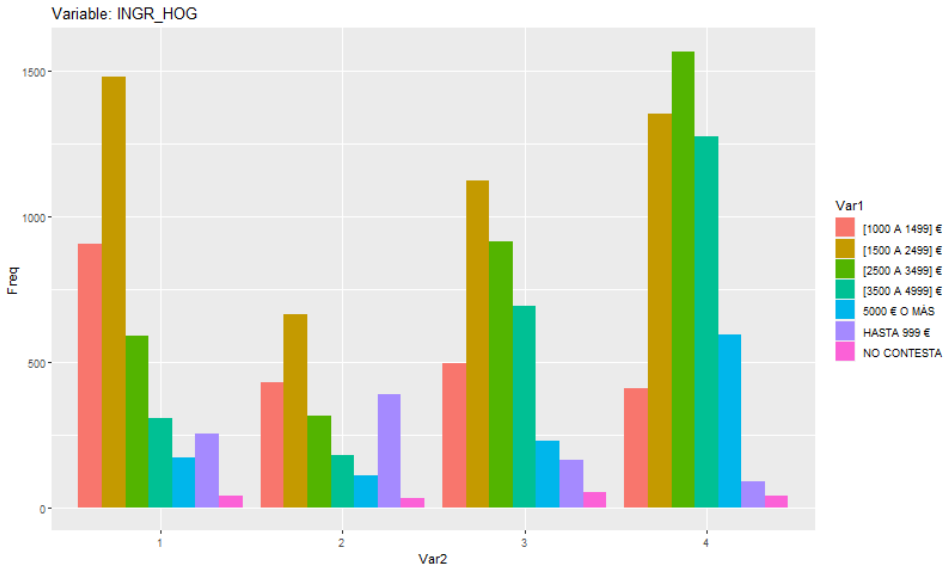

Variable: ACTI_RELIGION
Freq
Var2
Var1
0
1
Variable: ACTI_RURALES
Freq
Var2
Var1
1
Variable: ACTI_TEMATICOS
Freq
Var2
Var1
1

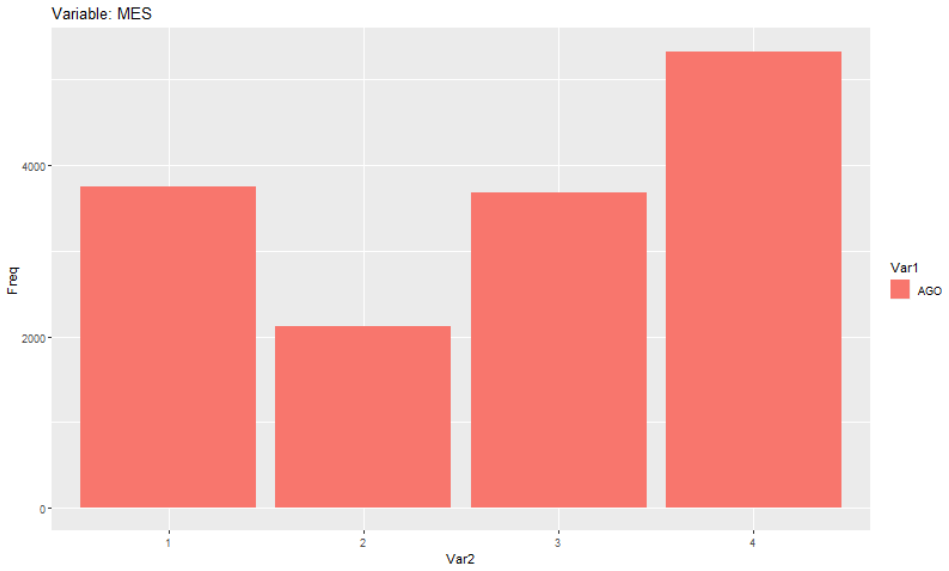

Variable: ACTI_VISITASCULTU
Freq
Var2
Var1
1
Variable: ALOJAPRIN
Freq
Var2
Var1
ALBERGUE
CAMPING
CRUCERO
HABIT ALQ
HOSTAL
HOTEL
OT NO MERC
OT. MERC
RURAL
VIV ALQ
VIV COMP
VIV CONO
VIV INTER
VIV PROPIA
Variable: CCAA_RESIDENCIA
Freq
Var2
Var1
AND
ARA
AST
BAL
CAN
CANT
CAST.L
CAST.M
CAT
CEU
EXTRM
GAL
MAD
MEL
MUR
NAV
P.VAS
RIO
VAL

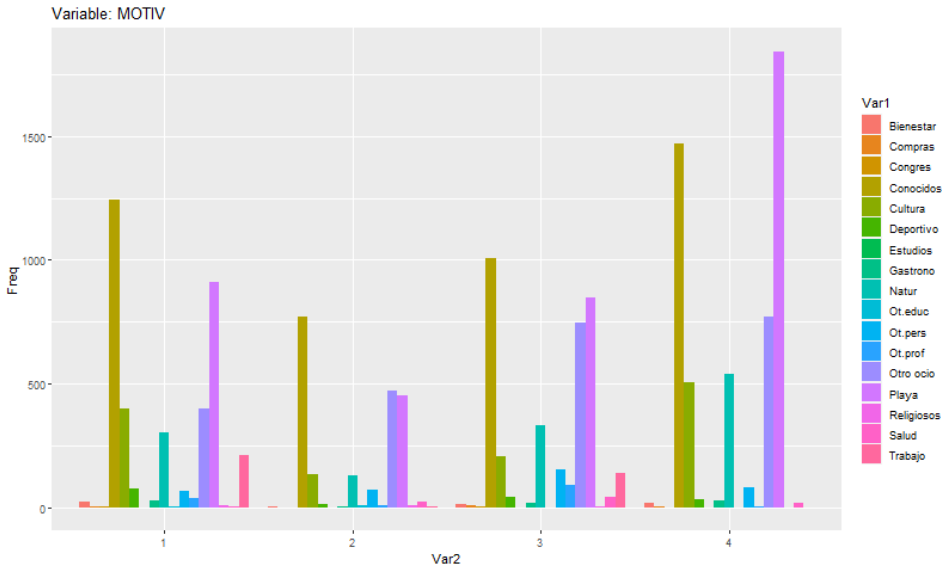

Variable: CONV
Freq
Var2
Var1
CÓNYUGE
NO EN PAREJA
PAREJA DE HECHO

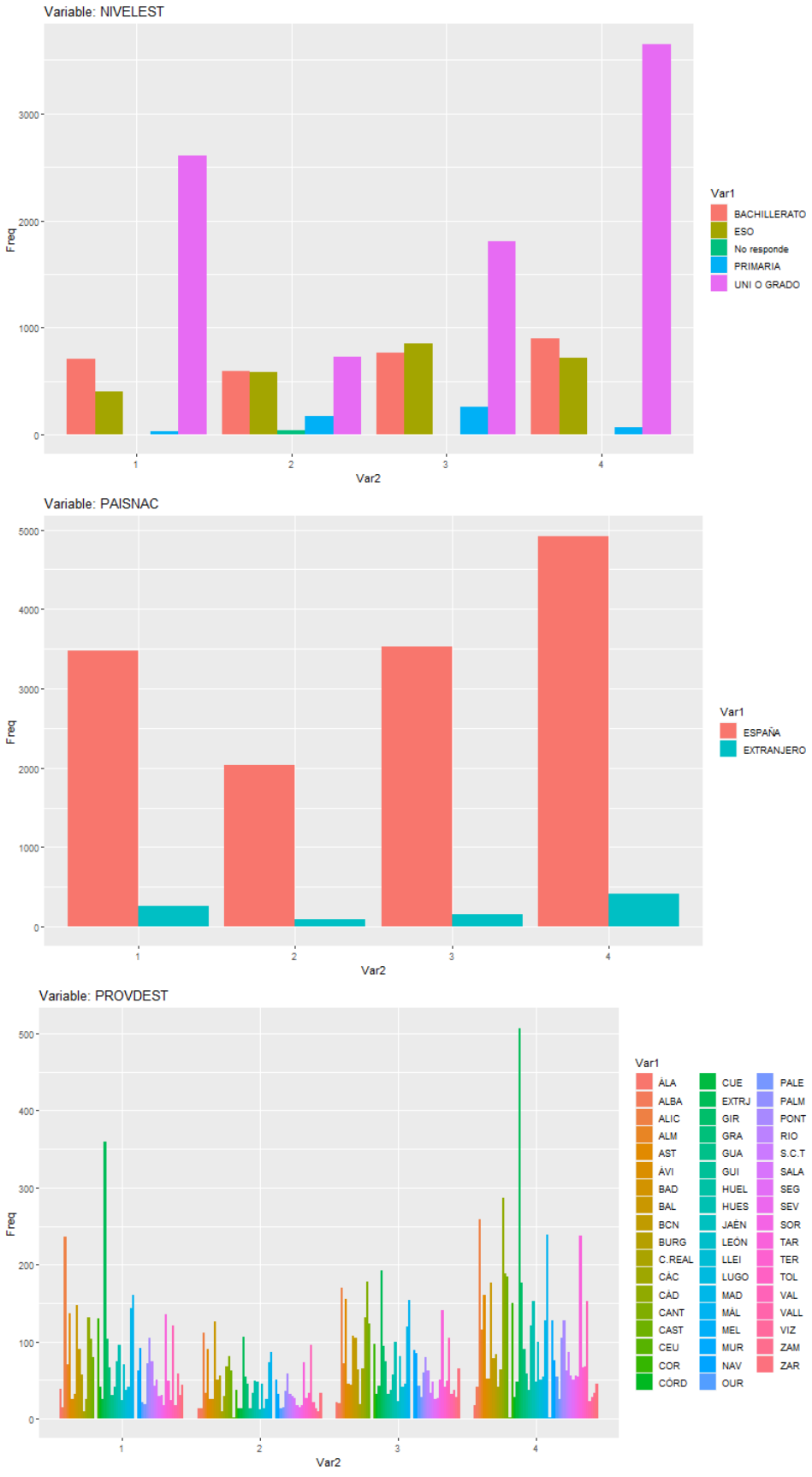

Variable: ECIVIL
Freq
Var2
Var1
CASADO/A
DIVORCIADO/A
SEPARADO/A
SOLTERO/A
VIUDO/A

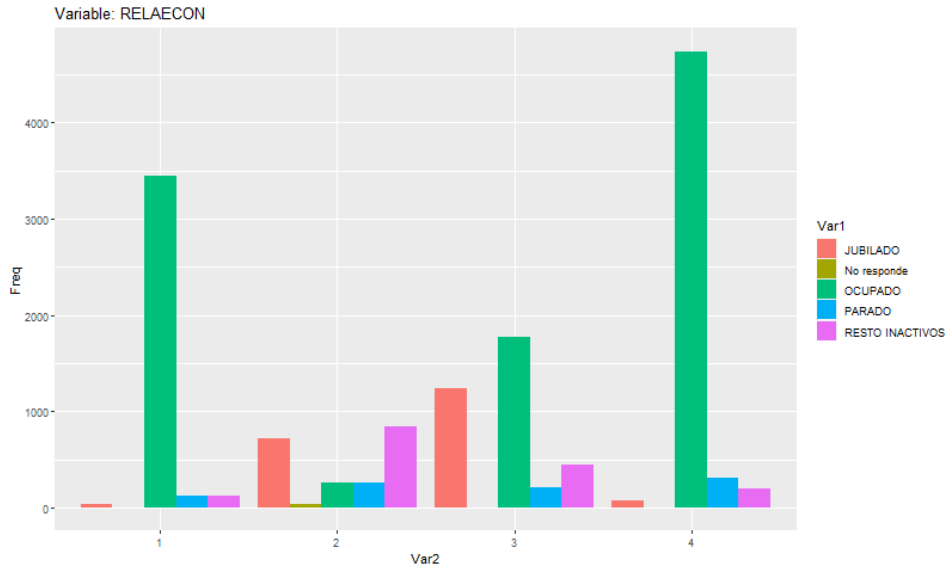

Variable: FIDELIDAD_DEST
Freq
Var2
Var1
NO
SI

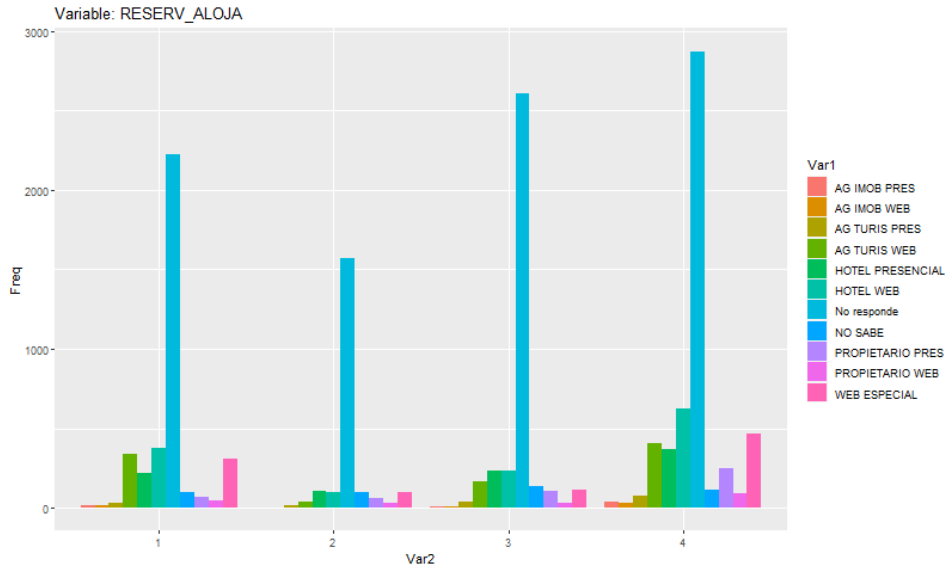

Variable: INGR_HOG
Freq
1500
1000
500
0
1
2
3
4
Var2
Var1
[1000 A 1499] €
[1500 A 2499] €
[2500 A 3499] €
[3500 A 4999] €
5000 € O MÁS
HASTA 999 €
NO CONTESTA

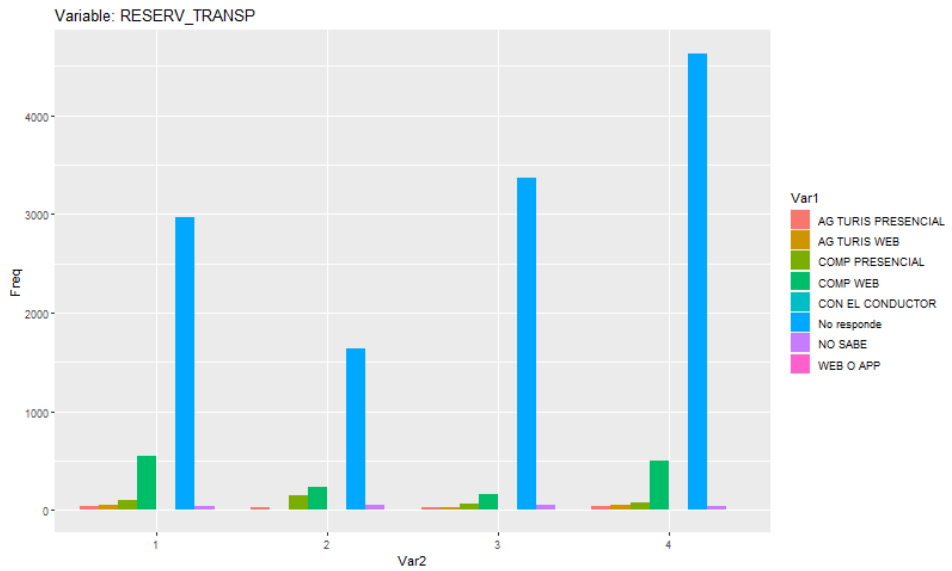

Variable: MES
Freq
4000
2000
0
1
2
3
4
Var2
Var1
AGO

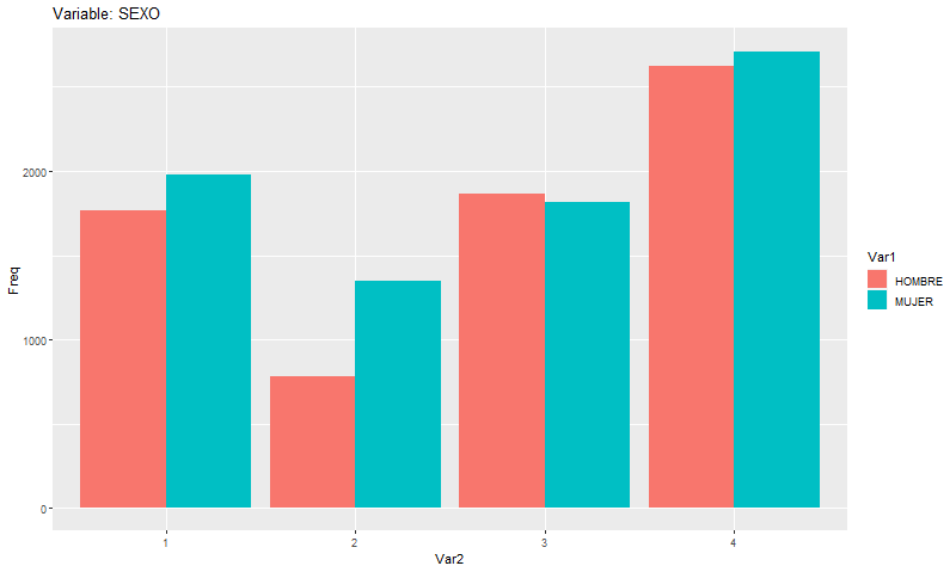

Variable: MOTIV
Freq
1500
1000
500
0
1
2
3
4
Var2
Var1
Bienestar
Compras
Congres
Conocidos
Cultura
Deportivo
Estudios
Gastrono
Natur
Ot.educ
Ot.pers
Ot.prof
Otro ocio
Playa
Religiosos
Salud
Trabajo

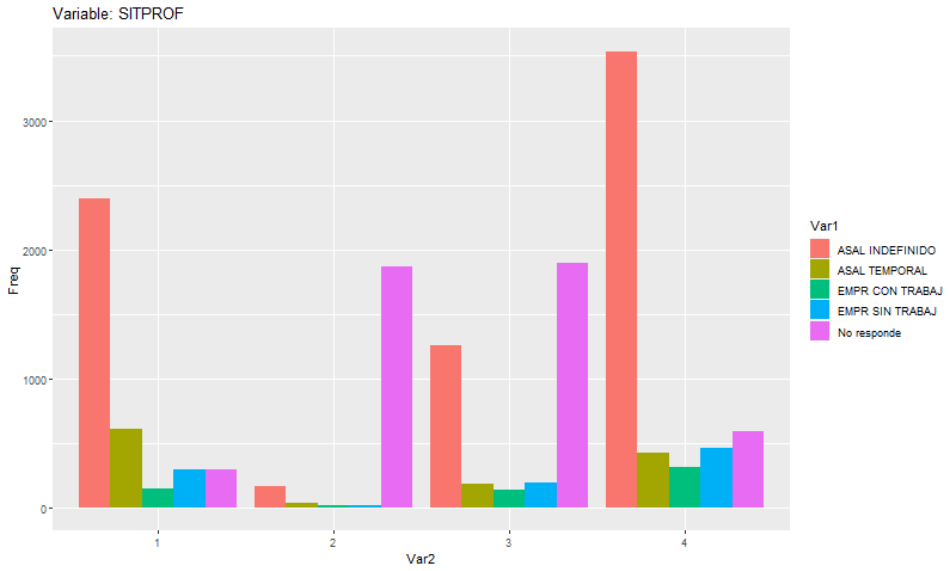

Variable: NIVELEST
Freq
Var2
Var1
BACHILLERATO
ESO
No responde
PRIMARIA
UNI O GRADO
Variable: PAISNAC
Freq
Var2
Var1
ESPAÑA
EXTRANJERO
Variable: PROVDEST
Freq
Var2
Var1
ÁLA
ALBA
ALIC
ALM
AST
ÁVI
BAD
BAL
BCN
BURG
C.REAL
CÁC
CÁD
CANT
CAST
CEU
COR
CÓRD
CUE
EXTRJ
GIR
GRA
GUA
GUI
HUEL
HUES
JAÉN
LEÓN
LLEI
LUGO
MAD
MÁL
MEL
MUR
NAV
OUR
PALE
PALM
PONT
RIO
S.C.T
SALA
SEG
SEV
SOR
TAR
TER
TOL
VAL
VALL
VIZ
ZAM
ZAR

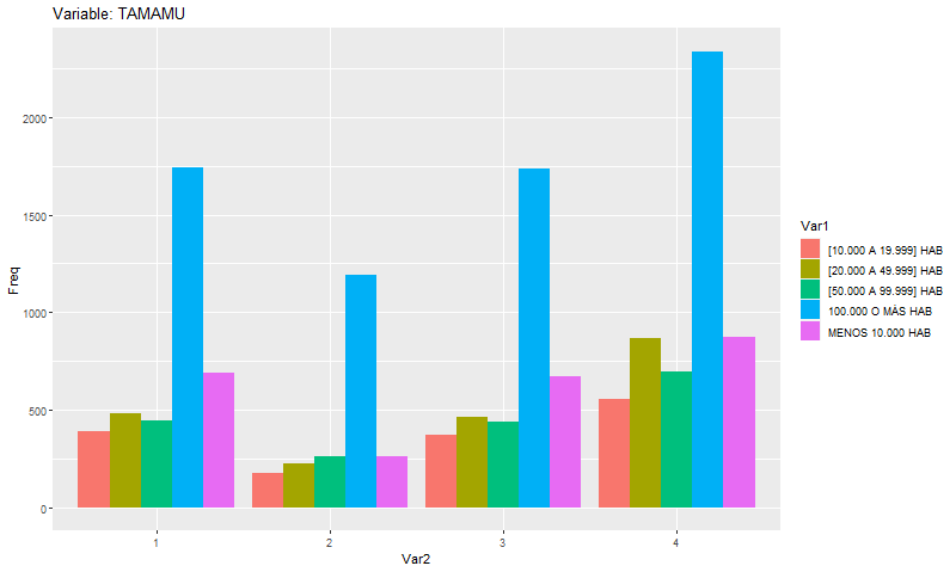

Variable: RELAECON
Freq
Var2
Var1
JUBILADO
No responde
OCUPADO
PARADO
RESTO INACTIVOS

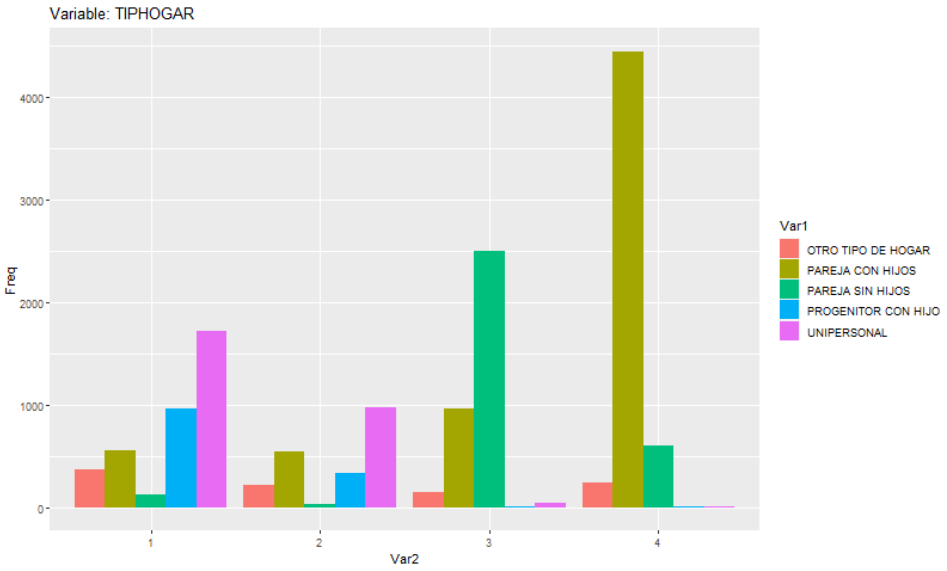

Variable: RESERV_ALOJA
Freq
Var2
Var1
AG IMOB PRES
AG IMOB WEB
AG TURIS PRES
AG TURIS WEB
HOTEL PRESENCIAL
HOTEL WEB
No responde
NO SABE
PROPIETARIO PRES
PROPIETARIO WEB
WEB ESPECIAL

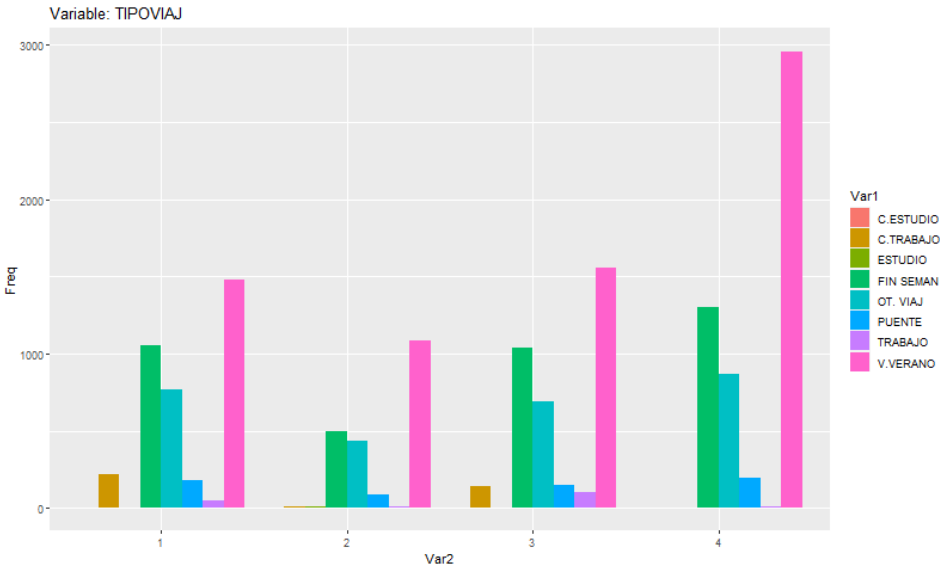

Variable: RESERV_TRANSP
Freq
Var2
Var1
AG TURIS PRESENCIAL
AG TURIS WEB
COMP PRESENCIAL
COMP WEB
CON EL CONDUCTOR
No responde
NO SABE
WEB O APP

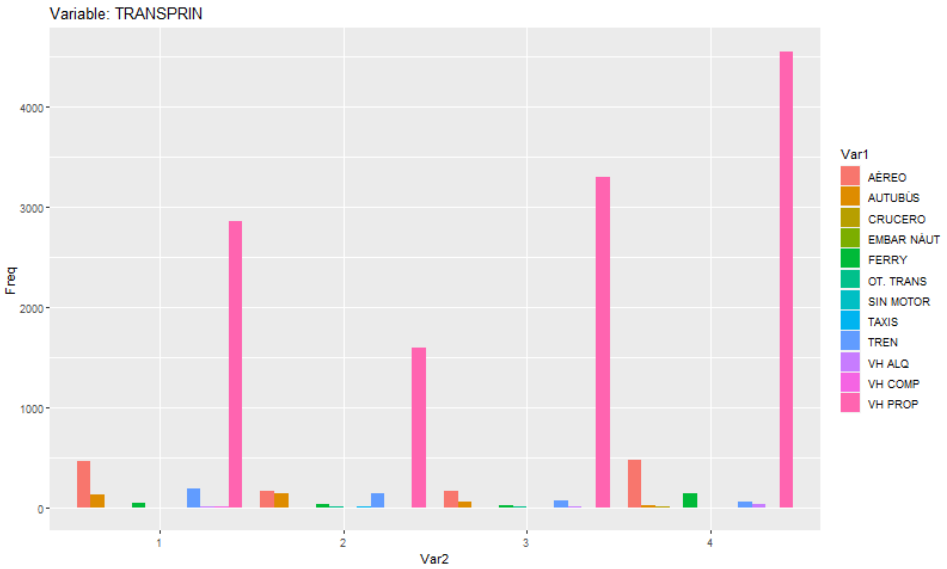

Variable: SEXO
Freq
2000
1000
0
1
2
3
4
Var2
Var1
HOMBRE
MUJER

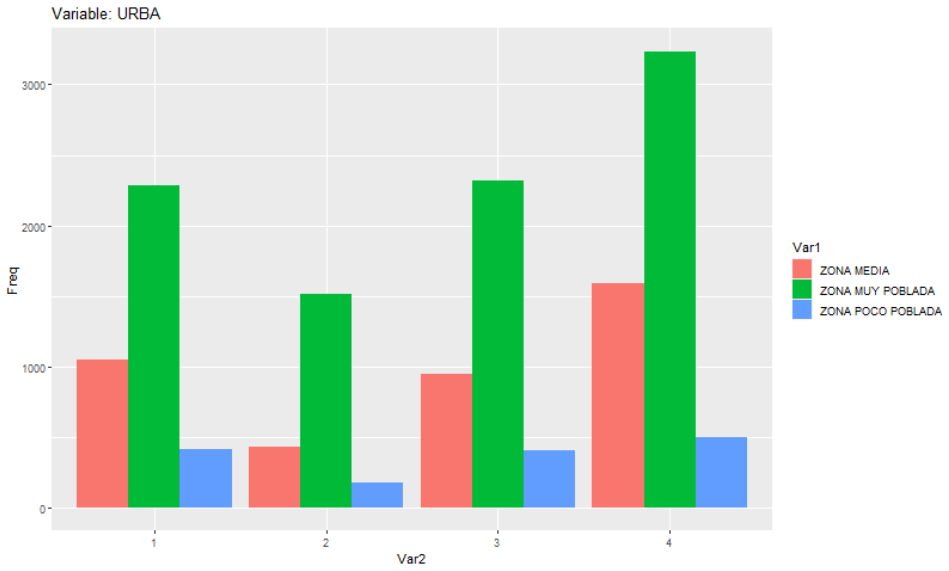

Variable: SITPROF
Freq
3000
2000
1000
0
1
2
3
4
Var2
Var1
ASAL INDEFINIDO
ASAL TEMPORAL
EMPR CON TRABAJ
EMPR SIN TRABAJ
No responde

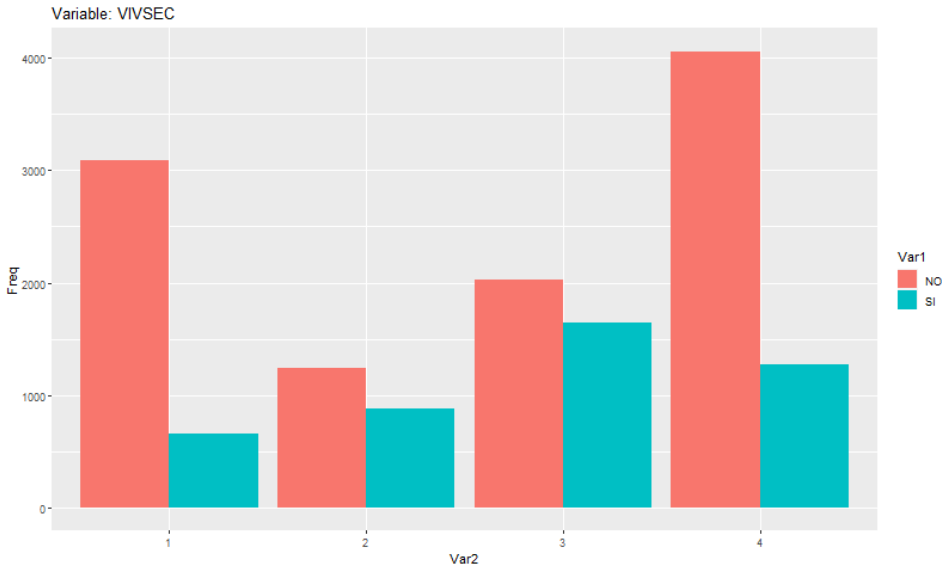

Variable: TAMAMU
Freq
2000
1500
1000
500
0
1
2
3
4
Var2
Var1
[10.000 A 19.999] HAB
[20.000 A 49.999] HAB
[50.000 A 99.999] HAB
100.000 O MÁS HAB
MENOS 10.000 HAB

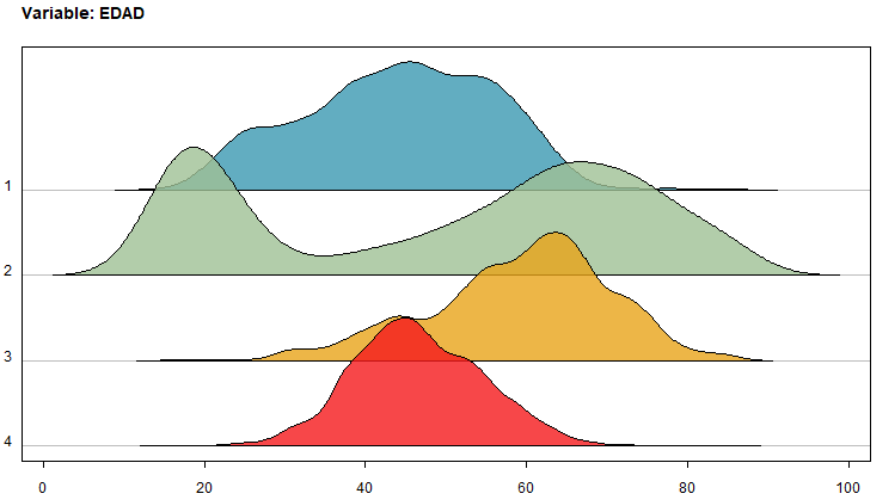

Variable: TIPHOGAR
Freq
Var2
Var1
OTRO TIPO DE HOGAR
PAREJA CON HIJOS
PAREJA SIN HIJOS
PROGENITOR CON HIJO
UNIPERSONAL

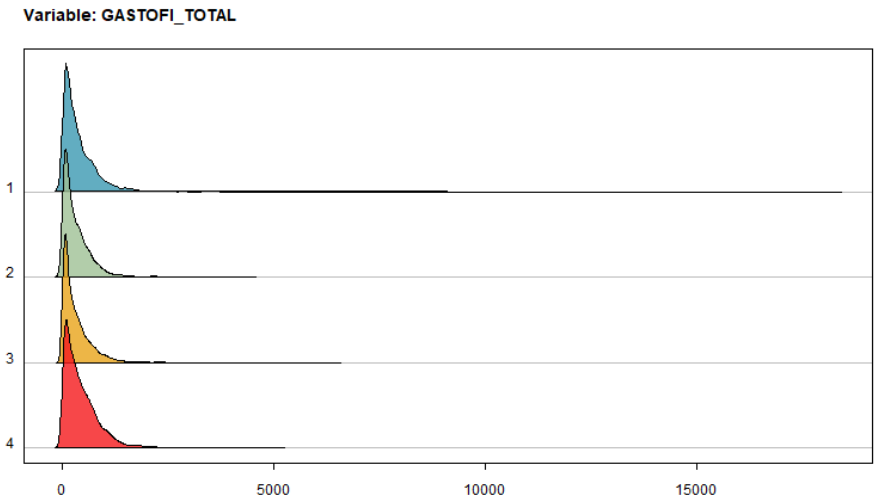

Variable: TIPOVIAJ
Freq
Var2
Var1
C.ESTUDIO
C.TRABAJO
ESTUDIO
FIN SEMAN
OT. VIAJ
PUENTE
TRABAJO
V.VERANO

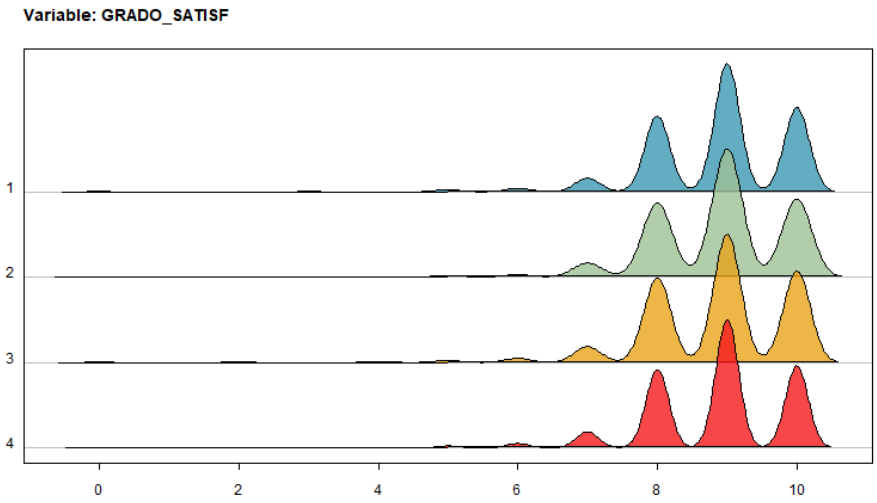

Variable: TRANSPRIN
Freq
Var2
Var1
AÉREO
AUTUBÚS
CRUCERO
EMBAR NÁUT
FERRY
OT. TRANS
SIN MOTOR
TAXIS
TREN
VH ALQ
VH COMP
VH PROP

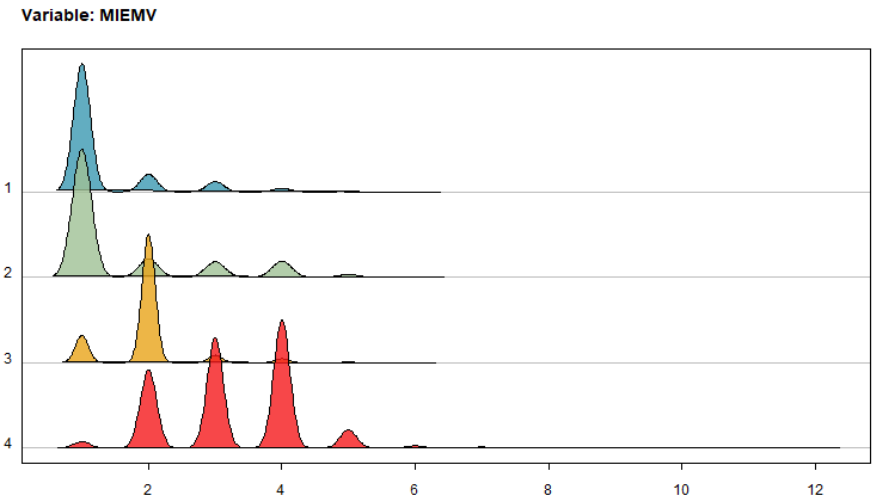

Variable: URBA
Freq
3000
2000
1000
0
1
2
3
4
Var2
Var1
ZONA MEDIA
ZONA MUY POBLADA
ZONA POCO POBLADA

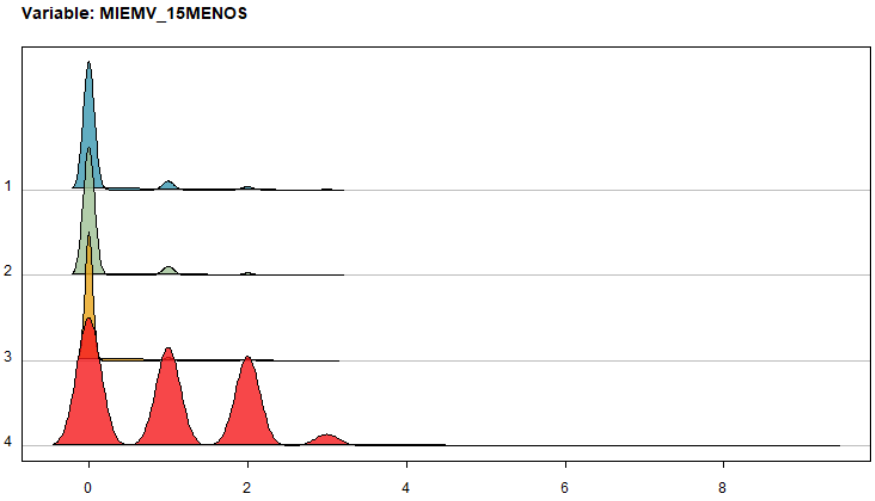

Variable: VIVSEC
Freq
4000
3000
2000
1000
0
1
2
3
4
Var2
Var1
NO
SI


Variable: EDAD
1
2
3
4
0
20
40
60
80
100

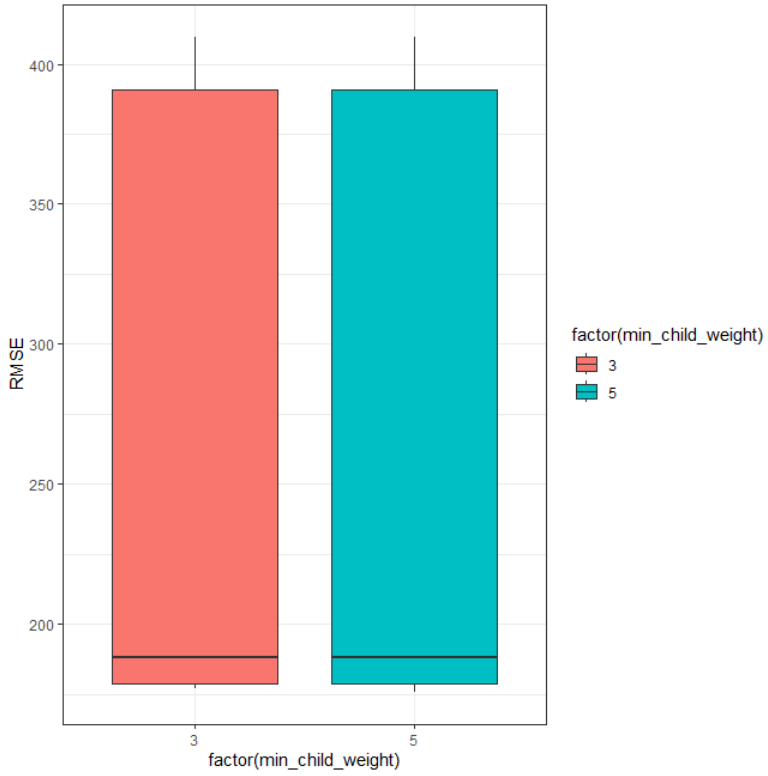

Variable: GASTOFI_TOTAL
1
2
3
4
0
5000
10000
15000

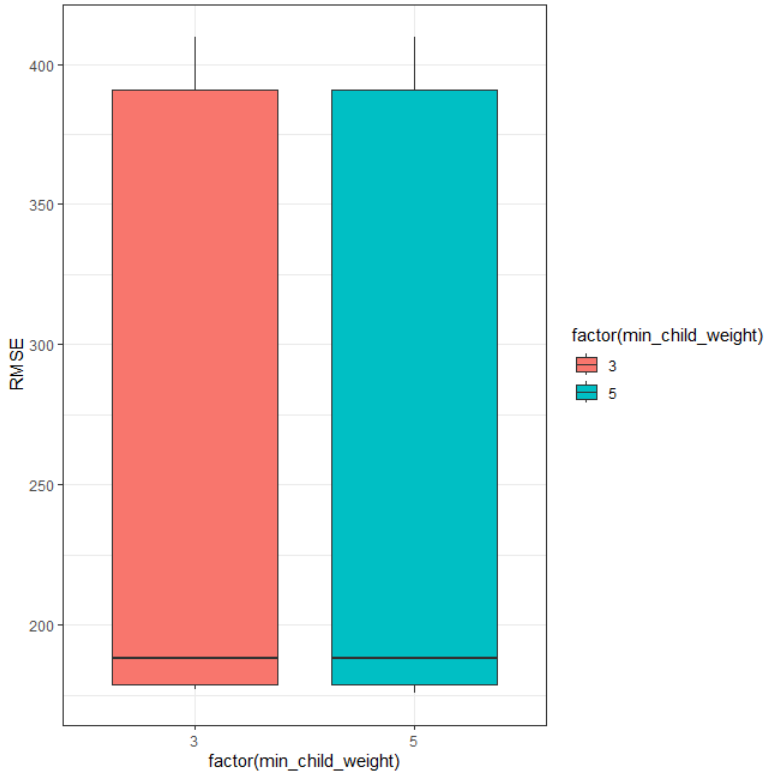

Variable: GRADO_SATISF
1
2
3
4
0
2
4
6
8
10

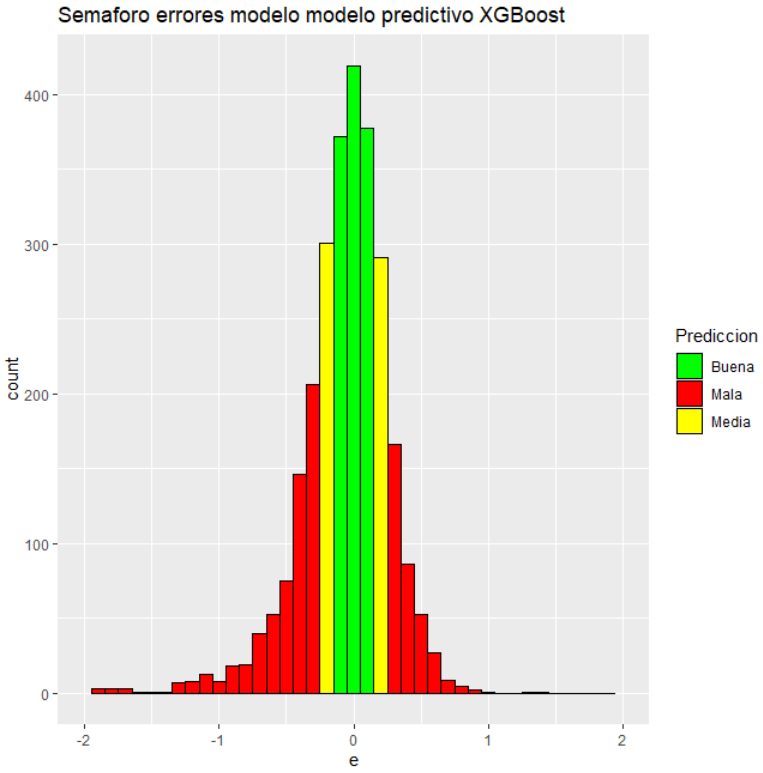

Variable: MIEMV
1
2
3
4
2
4
6
8
10
12

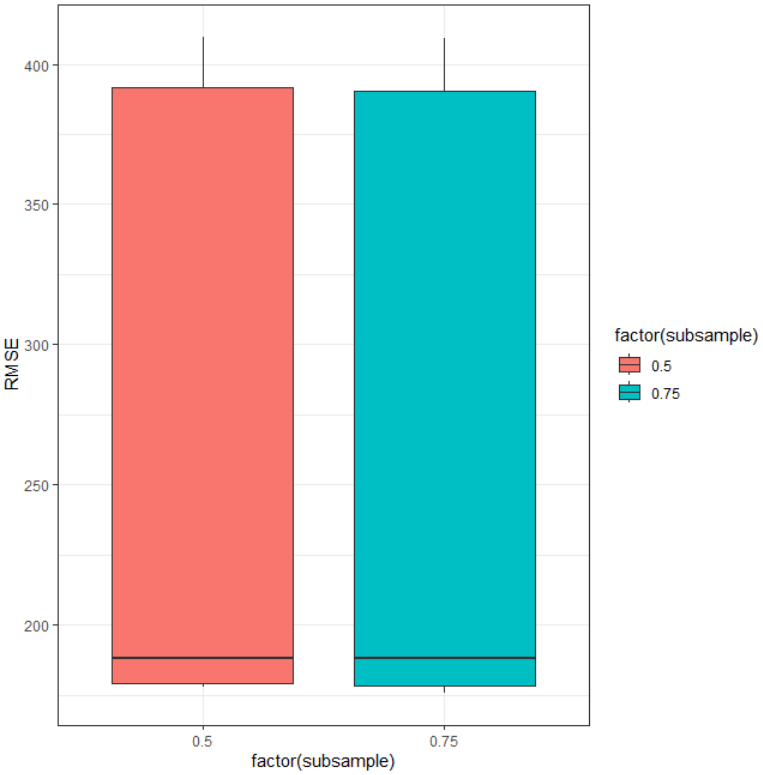

Variable: MIEMV_15MENOS
1
2
3
4
0
2
4
6
8

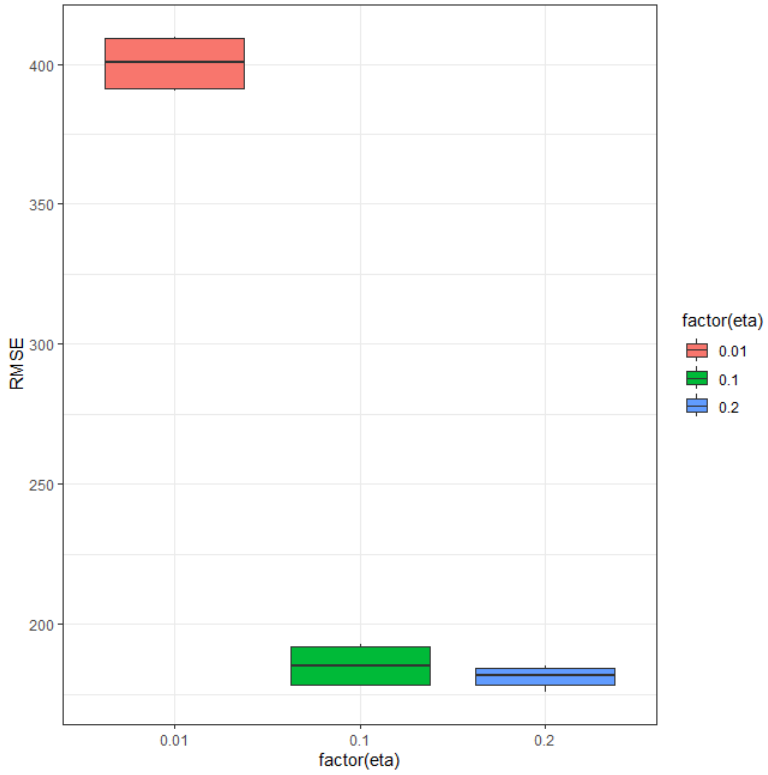
*XGBoost:*

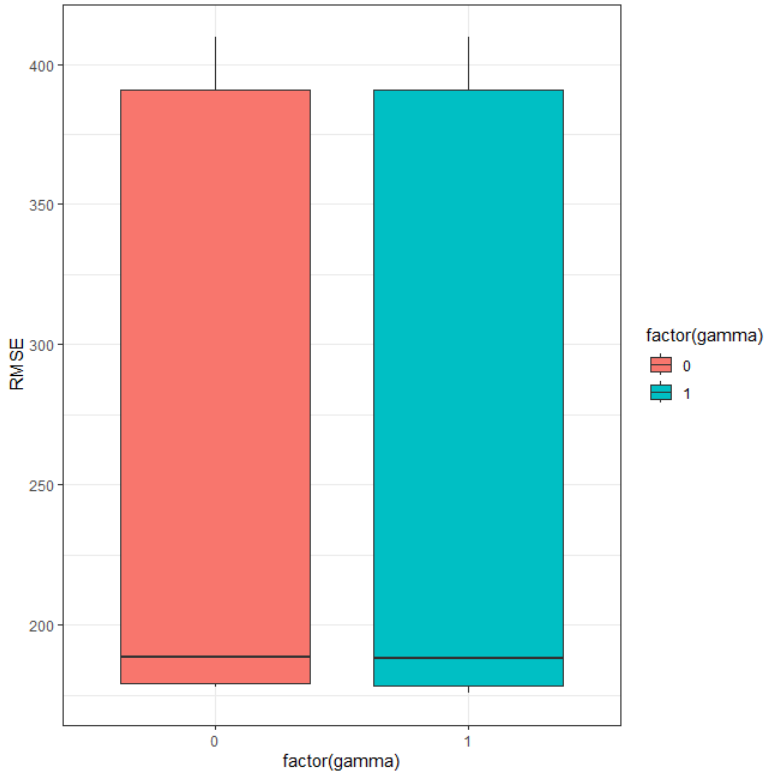

400
350
300
250
200
RMSE
3
5
factor(min_child_weight)
factor(min_child_weight)
3
5

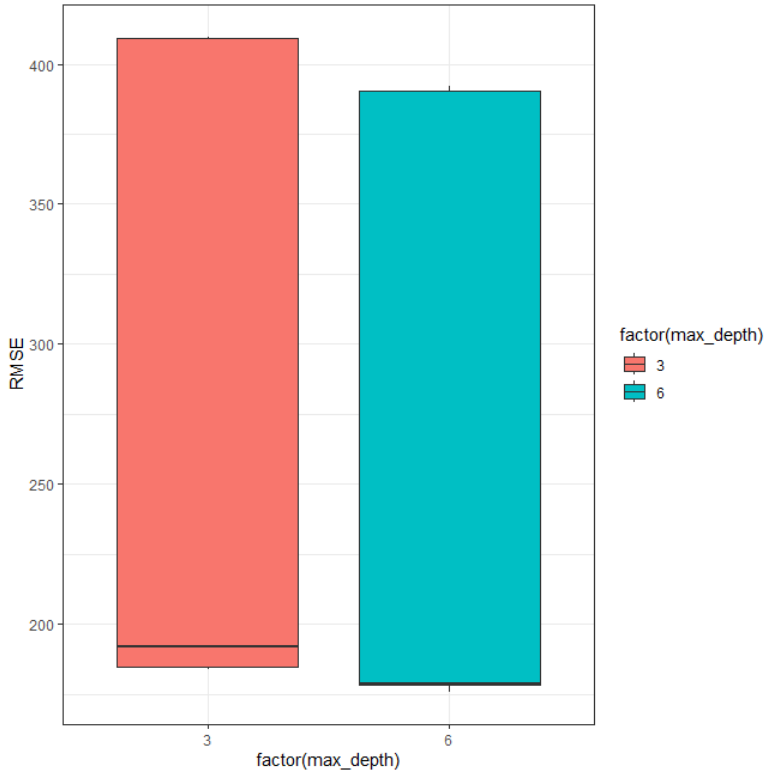

400
350
300
250
200
RMSE
3
5
factor(min_child_weight)
factor(min_child_weight)
3
5


Semaforo errores modelo modelo predictivo XGBoost
count
e
Prediccion
Buena
Mala
Media


RMSE
factor(subsample)
factor(subsample)
0.5
0.75


400
350
300
250
200
RMSE
0.01
0.1
0.2
factor(eta)
factor(eta)
0.01
0.1
0.2


400
350
300
250
200
RMSE
0
1
factor(gamma)
factor(gamma)
0
1


400
350
300
250
200
RMSE
3
6
factor(max_depth)
factor(max_depth)
3
6