

Grau en Estadística

Títol: Anàlisi de dades al núvol amb BigQuery

Autor: Anna Salazar Belver

Director: Montserrat Guillén i Estany

Departament: Departament d'Econometria, Estadística i Economia Aplicada

Convocatòria: Gener 2023



Resum

Aquest treball està centrat en el funcionament de BigQuery, un programa per a l'emmagatzement i l'anàlisi de dades al núvol. Es pretén crear una guia pràctica per l'ús d'aquesta eina en la seva versió gratuïta. L'escrit aborda les limitacions a respectar per la utilització de la plataforma sense cap cost, així com la connexió de BigQuery amb altres programes per facilitar l'anàlisi i visualització de les dades. Finalment, es presenten dos mètodes d'anàlisi de dades per discutir l'abast de l'eina en aquest moment.

Paraules clau: Big Data, Magatzem de dades, Visualització de dades, API, Anàlisi de dades, Google Cloud, Plataforma al núvol.

Abstract

This work focuses on the performance of BigQuery, a programme for storing and analysing data in the cloud. It aims to create a practical guide for the use of this tool in its free version. The paper addresses the limitations to be respected for the use of the platform at no cost, as well as the connection of BigQuery with other programmes to facilitate the analysis and visualization of data. Finally, two data analysis methods are presented to discuss the availability of the tool at this time.

Keywords: *Big Data, Data Warehouse, Data Visualization, API, Data Analysis, Google Cloud, Cloud Platform.*

Classificació AMS: 47R50 Data bases, information systems
62-07 Data analysis
68T09 Computacional aspects of data analytics and big data
94A16 Informational aspects of data analytics and big data

Índex

1	Introducció	1
1.1	Motivació	1
2	Metodologia	2
3	Què és BigQuery?	3
3.1	Per què hauríem d'utilitzar BigQuery en lloc d'altres eines?	3
4	Creació i treball amb conjunts de dades i taules	5
4.1	Configuració de la Plataforma de Google Cloud (GCP)	5
4.1.1	Limitacions	5
4.2	Creació d'un conjunt de dades	6
4.3	Definició d'una taula de BigQuery des de la interfície d'usuari	7
4.3.1	Afegir dades a una taula de BigQuery senzilla	9
4.4	Càrrega de dades per crear una taula a BigQuery	11
4.5	Consulta de dades i visualització d'estadístiques de consultes	12
4.6	Creació d'una taula a partir d'un resultat de consulta	13
5	Dades públiques i dades externes	15
5.1	Conjunts de dades públiques a BigQuery	15
5.2	Taules externes de BigQuery	17
6	Integració de BigQuery amb Looker Studio	19
6.1	Ús de Looker Studio des de BigQuery	19
6.2	Ús de Looker Studio des de la pròpia plataforma	21
7	Connexió d'R a BigQuery	23
7.1	Connectem R a BigQuery	23
7.2	Consultes amb el paquet "bigrquery"	24

8 Implementació de BigQuery a l'aula	25
8.1 Consultes amb 'bigquery'	25
8.2 Comentaris dels estudiants sobre la pràctica	29
9 Anàlisi de dades mitjançant la connexió d'R i BigQuery	31
9.1 Descripció de la base de dades	31
9.1.1 Lectura de les dades	32
9.1.2 Objectius del projecte	33
9.2 Preprocessament	34
9.2.1 Dades mancants	34
9.2.2 Dades atípiques	35
9.2.3 Categoritzar	37
9.2.4 Variable resposta	39
9.3 Anàlisi descriptiva	40
9.3.1 Anàlisi univariant	40
9.3.2 Anàlisi bivariant	45
9.4 Anàlisi per components principals	47
9.4.1 Projectió de variables numèriques	49
9.4.2 Projectió de variables categòriques	51
10 Aprenentatge automàtic amb BigQuery ML	57
10.1 Regressió logística	57
10.1.1 Entrenament del model	57
10.1.2 Avaluació del model	59
10.1.3 Interpretació del model	62
11 Recopilació del material a Github	65
12 Conclusions	66
Bibliografia	69

1 Introducció

L'objectiu d'aquest treball és conèixer un dels recursos de la plataforma Google Cloud, anomenat BigQuery. A partir de l'enteniment de com funciona la computació al núvol i l'emmagatzemament de dades en el mateix entorn, es vol crear un material pedagògic per a incloure en el programari de l'assignatura de Fitxers i Bases de Dades del grau d'Estadística.

Amb aquest propòsit en ment, el treball es divideix en tres apartats principals. En primer lloc, s'aprofundirà en la descripció de la plataforma al llarg de les primeres cinc seccions. Primer es descriurà BigQuery a alt nivell per, més endavant, poder entendre l'estructura de la interfície d'usuari i les maneres en què es pot interactuar amb el sistema. En aquest apartat també s'inclouran les connexions amb altres plataformes i *softwares* per a poder treure el màxim profit de les dades d'estudi. Seguidament, el segon apartat està conformat per una única secció, que contempla l'aplicació de BigQuery a l'entorn de classe i inclou la valoració dels mateixos estudiants sobre la seva experiència i visió. En últim lloc, es presenta l'anàlisi d'una base de dades mitjançant la connexió amb el programari R. En aquest apartat, es realitza part de l'exploració de les dades a R i part a BigQuery, tot tenint en compte les limitacions del llenguatge que s'utilitza en cada entorn i els recursos visuals que ofereixen.

1.1 Motivació

Al llarg de la meua trajectòria acadèmica he anat identificant, cada cop de manera més precisa, els meus interessos i ambicions, així com les tècniques d'aprenentatge que em permeten aprendre i formar-me de manera més efectiva. En el meu cas, l'aprenentatge basat en la pràctica em permet aplicar els coneixements teòrics que he adquirit a problemes reals, la qual cosa fa que entengui millor els conceptes i vegi la seva utilitat. Tot això és el que m'ha donat motivació per continuar aprenent i intentar anar sempre una mica més enllà, i el motiu principal que em va portar a interessar-me per crear un recurs pedagògic pràctic per compartir-lo amb altres estudiants amb inquietuds similars a les meves.

La meua tutora, la professora Montserrat Guillén, em va donar l'oportunitat d'ajudar-la en aquest projecte, que va captar la meua atenció des del primer moment. Arran de cursar l'assignatura de Fitxers i Bases de Dades, el meu interès en les macrodades o *Big Data* va créixer i, cada cop més, penso que la solució del futur per treballar amb conjunts de dades massius serà fer-ho al núvol o d'alguna manera que no impliqui la descàrrega de dades a l'equip local. Fent aquest treball he conegut una de les alternatives per treballar amb aquest tipus d'informació, però encara queda camí al davant i moltes ganes d'aprendre i conceptes en les quals aprofundir.

2 Metodologia

Per poder fer un seguiment del treball i exemplificar tota la informació teòrica que s'anirà veient en el transcurs d'aquest, s'ha fet ús de dos conjunts de dades.

Per una banda, s'ha tractat una base de dades pública a l'entorn de BigQuery, anomenada *Catalonia Cell Coverage*. Aquesta conté informació sobre la cobertura de telefonia mòbil de la població catalana que va ser recopilada des de l'any 2015 fins al 2017, ambdós inclosos. Algunes de les variables que es van tenir en compte en aquest estudi van ser el senyal mitjà del dispositiu, el nom de la xarxa i de l'operador, la velocitat estimada de la font i el codi postal del lloc on es va adquirir la telemetria.

Per altra part, s'han utilitzat unes dades que provenen de l'agència estatal de trànsit de Washington, Estats Units. L'Administració Nacional de Seguretat del Trànsit a les Carreteres, *National Highway Traffic Safety Administration* (NHTSA) en anglès, va fer públiques tres taules que feien referència als accidents ocorreguts al llarg de l'any 2015, les persones involucrades en aquests (siguin conductors, passatgers o vianants) i un inventari de tots els vehicles que van ser afectats. Les dades tractades són una mostra que conté tots els accidents produïts el mes de desembre d'aquell any. Aquesta informació es pot trobar al web de la NHTSA: <https://www.transportation.gov/briefing-room/traffic-fatalities-sharply-2015>.

Per poder dur a terme aquest treball s'ha emprat el *software* R, mitjançant la interfície R-Studio, i dos llenguatges de programació. Principalment, s'ha fet ús del llenguatge SQL, tant a la pàgina de BigQuery com a l'entorn R, fent ús del paquet "sqldf", que es pot trobar al repositori oficial CRAN. Així i tot, també s'ha usat el mateix llenguatge R per acomplir alguns altres aspectes de l'anàlisi que inclouen el preprocessament de les dades i la creació de gràfics descriptius.

3 Què és BigQuery?

BigQuery és un motor d'anàlisi de macrodades (*Big Data*) que permet executar consultes amb llenguatge SQL (*Structured Query Language*) al núvol sobre les dades emmagatzemades en aquest, sense importar el volum de les dades ni el tipus de consultes que es volen fer. El motor de consulta és capaç de treballar sobre terabytes de dades en qüestió de segons, i sobre petabytes en pocs minuts. Avui en dia, les empreses estan adoptant cada cop més la presa de decisions basades en dades, i fomentant una cultura oberta en la qual les dades no estan aïllades dins dels departaments. BigQuery, en proporcionar els mitjans tecnològics per a promoure un canvi cultural cap a l'agilitat i l'obertura, realitza un paper molt important en l'augment del ritme de la innovació.

Treballar amb dades a BigQuery implica tres aspectes principals: l'emmagatzemament, la incorporació de les dades i la consulta d'aquestes, Google s'encarrega de tota la resta. Com BigQuery és un servei totalment gestionat, no és necessari configurar ni instal·lar res en el nostre ordinador i, pel mateix motiu, no necessitem un administrador de la base de dades. Simplement, podem entrar al nostre projecte de Google Cloud des del mateix navegador i començar a analitzar.

Pel que fa a l'emmagatzemament, les dades es guarden en una taula estructurada, la qual cosa significa que es pot utilitzar SQL estàndard per a facilitar la consulta i l'anàlisi de dades. BigQuery és perfecta pel *Big Data* perquè gestiona tot aquest emmagatzemament i està proveïda d'operacions d'escalabilitat que funcionen de forma automàtica sense que l'usuari tingui la necessitat d'involucrar-se, per la qual cosa mai haurem de preocupar-nos per la grandària de les dades amb les quals treballem. Part de la consideració de disseny darrere de BigQuery és animar als usuaris a centrar-se en els coneixements en lloc de la infraestructura.

Per a més informació sobre BigQuery, es pot consultar la pàgina de [Google Cloud](#).

3.1 Per què hauríem d'utilitzar BigQuery en lloc d'altres eines?

Una de les característiques més rellevants que presenta BigQuery és que es tracta d'una plataforma sense servidor local, és a dir, que els servidors s'executen en segon pla, sense la intervenció de l'usuari. A més, presenta una alta disponibilitat, la qual cosa es tradueix en que no cal preocupar-se per la caiguda dels servidors, ja que la plataforma s'encarrega del seu manteniment. També té propietats d'escalabilitat automàtica que fan possible gestionar fins a petabytes de dades. Aquestes característiques no estan disponibles a la majoria de plataformes d'emmagatzemament de dades tradicionals, i fan destacar BigQuery entre moltes.

Com en molts altres magatzems de dades, BigQuery és capaç de treballar amb moltes fonts de

dades diferents. Es poden pujar les dades des del propi sistema d'arxius, des de Google Cloud Storage o des de Google Drive, entre moltes més fonts. Després de fer-ho, es poden consultar aquestes dades utilitzant SQL estàndard o SQL heretat. Els resultats de les consultes solen emmagatzemar-se en la memòria cau durant 24 hores, de manera que les següents execucions d'aquesta consulta només hauran d'obtenir les dades de la memòria cau en lloc de fer-ho del disc. És a dir, estarà lliure de cost fer la mateixa consulta en un plaç de 24 hores, ja que no s'haurà de tornar a examinar la base de dades per obtenir el resultat.

4 Creació i treball amb conjunts de dades i taules

4.1 Configuració de la Plataforma de Google Cloud (GCP)

Per utilitzar aquesta eina d'anàlisi només caldrà crear-se un compte a Google Cloud i treballar a la zona de proves (*Sandbox*) que ofereix Google per operar de forma gratuïta. Per fer servir la zona de proves seguirem els passos següents:

1. En primer lloc, ens dirigim a la interfície d'usuari (UI) de [BigQuery](#). Des d'aquesta interfície es poden realitzar la majoria de les operacions.
2. Accedim al nostre compte de Google o creem un nou compte si encara no en tenim cap. Si és el primer cop que iniciem sessió a Google Cloud, haurem de marcar el país on som i acceptar les condicions de servei.
3. Un cop dins, podem veure com és l'espai de treball SQL. Hi ha una secció de l'Explorador a l'esquerra que ens permet navegar entre projectes, conjunts de dades i taules. Per tal de fer servir la zona de proves, és necessària la creació d'un projecte.

S'introdueix un nom al projecte i es fa clic a *Create*. En el nostre cas, l'hem anomenat `el_meu_projecte` (Figura 4.1), i treballarem sobre aquest per il·lustrar el funcionament de la plataforma.

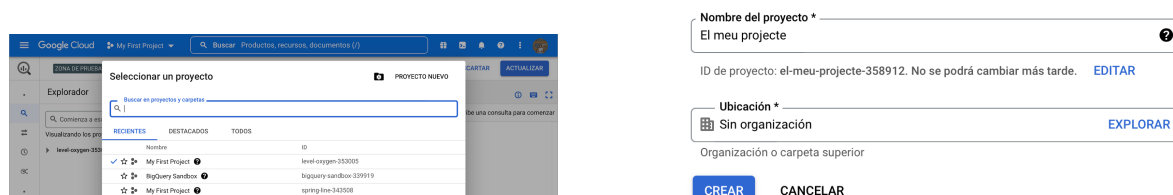


Figura 4.1: Creació d'un projecte

4. Un cop creat el projecte, el navegador ens redirigeix a la interfície web de BigQuery.
5. Ara ja podrem carregar o consultar dades en el nostre projecte sense cap compte de facturació adjunta.

4.1.1 Limitacions

Això no obstant, per a l'ús de la zona de proves gratuïta que ofereix Google, haurem de tenir en compte un seguit de limitacions.

En primer lloc, ens trobem amb un màxim de 10 Gb d'emmagatzemament i 10 Tb de consulta al mes. Al llarg d'aquest exemple no utilitzarem un volum de dades més gran ni sobrepassarem el límit d'espai de consulta, però s'han de tenir en compte aquestes limitacions si l'objectiu és

treballar amb el format gratuït.

A més, ens trobem que tots els conjunts de dades tenen el temps de caducitat de la taula per defecte establerta en seixanta dies. Per tant, totes les taules, vistes o particions de les taules caducaran automàticament passat aquest temps.

Una altra característica destacable és que els projectes de la zona de proves no són compatibles amb:

- La transmissió de dades.
- Sentències de llenguatge de manipulació de dades (DML).
- El servei de transferència de dades de BigQuery.

4.2 Creació d'un conjunt de dades

Ara que ja disposem d'un projecte en el qual treballar, ha arribat el moment de crear un nou conjunt de dades dins d'aquest. Es pot pensar en un conjunt de dades a BigQuery com una agrupació lògica de taules. Alhora, diferents conjunts de dades s'integren en un mateix projecte.

Per a crear-ne un, només s'ha de desplegar el menú i triar l'opció de crear un nou conjunt de dades. Tot seguit, hi ha diversos detalls per al conjunt de dades que es poden establir. En primer lloc, hi ha l'opció de canviar el projecte que l'encabirà. Això farà que aparegui un navegador on es podrà especificar el projecte. Una altra possibilitat serà escollir la ubicació de les dades. Això determina on s'aprovisionaran els recursos subjacents, com la computació i l'emmagatzematge, per al servei BigQuery. Les consideracions a l'hora de triar una ubicació inclouran el rendiment per als usuaris finals, l'alta disponibilitat i també qualsevol restricció d'auditoria o compliment. I, per últim, es pot establir un temps d'expiració per defecte per a les taules dins d'un conjunt de dades.

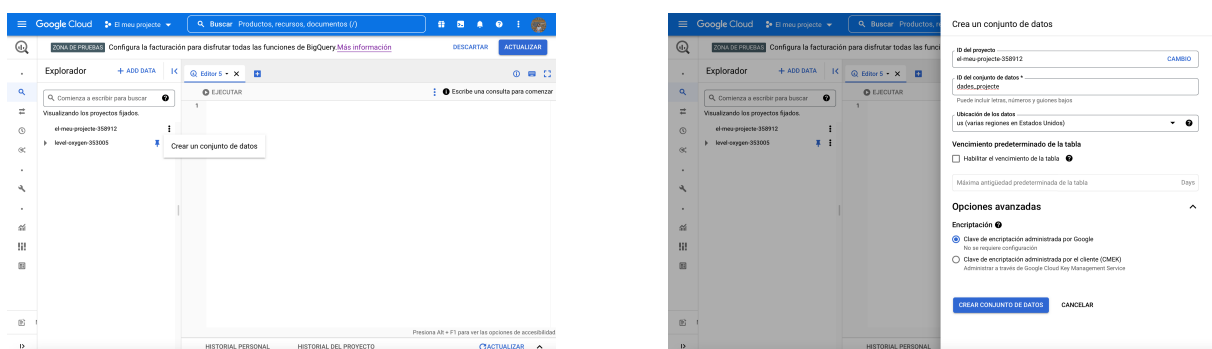


Figura 4.2: Creació d'un conjunt de dades

Tal com es pot veure a la figura 4.2, hem creat un nou conjunt de dades anomenat `dades_projecte` que estarà ubicat en el projecte `el_meu_projecte`, la ubicació de les dades s'ha establert a

diverses regions dels Estats Units i, per últim, no s'ha habilitat un temps de venciment de la taula, de manera que, per defecte, BigQuery l'emmagatzemarà per 60 dies.

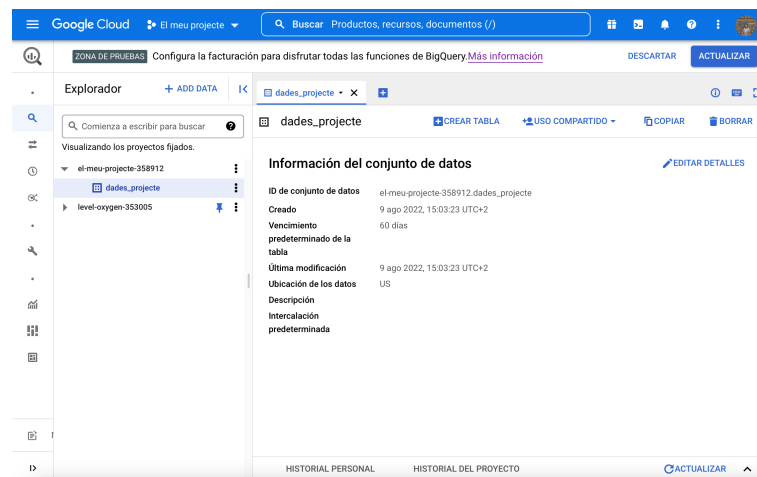


Figura 4.3: Informació del conjunt de dades

Un cop creat `dades_projecte`, es pot comprovar que ara apareix dins de `el_meu_projecte` a la interfície d'usuari (UI) de BigQuery, i s'observa que no hi ha taules dins d'aquest. També es pot donar un cop d'ull als detalls associats a aquest conjunt de dades (Figura 4.3). Des d'aquest menú, podem triar obrir-lo, i es desplegarà a la dreta tota la informació del conjunt de dades. En aquesta secció es troba l'identificador del conjunt de dades, que també assenjala el projecte en el qual s'ha creat, i després altres detalls que inclouen les hores de creació i modificació, així com la ubicació d'aquest.

A més, des d'aquesta finestra es pot compartir el conjunt de dades amb altres usuaris. Hi ha opcions per a copiar i eliminar aquest conjunt de dades. I després, a l'opció *editar detalls*, és possible reconfigurar el temps de caducitat de les taules, establir una descripció o afegir etiquetes. Per exemple, si volem marcar aquest conjunt de dades com a pertanyent a un equip, podem establir una etiqueta amb la clau d'equip i el valor corresponent. En acabant, quan el guardem, les etiquetes apareixen a l'apartat d'informació.

4.3 Definició d'una taula de BigQuery des de la interfície d'usuari

Després d'haver creat un conjunt o base de dades en un projecte, ja es pot crear una taula dins d'ell. Si es té la informació de la base, s'hauria de veure aquesta opció per a crear una nova taula des d'aquí. Alternativament, hom es pot dirigir al projecte, després al conjunt de dades i, finalment, triar l'opció de crear una taula. Apareixerà un formulari i donarà l'opció d'especificar una font per a la taula en qüestió. Això permetrà importar dades de fonts ja existents, com l'emmagatzematge en el núvol de Google, o bé un arxiu dels propis sistemes de l'ordinador.

La primera taula que crearem serà bastant simple, i ens servirà per explorar una mica la plataforma. De fet, serà una taula buida anomenada accidents (Figura 4.4).

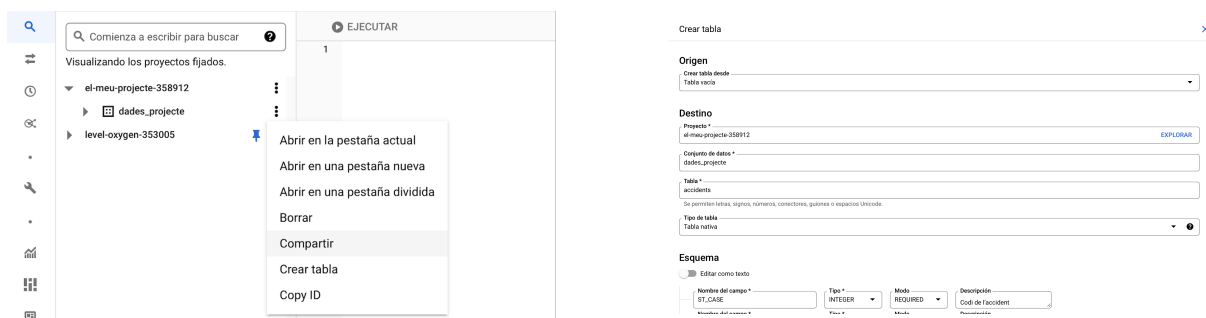


Figura 4.4: Creació d'una taula

A continuació, passem a la secció d'Esquema. Podem fer ús d'aquesta interfície per a establir les columnes de la nostra taula, incloent-hi els tipus i altres configuracions. La primera columna que definirem és l'identificador de l'accident, que s'anomenarà ST_CASE. Per al tipus de variable, podem triar d'entre un menú d'opcions, que inclou tots els tipus amb els quals ja estem familiaritzats. Quant a la manera (columna *modo* a la figura 4.5), aquesta determinarà si els valors que pren la variable poden ser nuls o si es requereix un valor (com és el cas de l'identificador), i també podem establir que els valors siguin d'un tipus que es pugui repetir, marcant *indistint*. Finalment, es pot escriure una descripció per a la variable, que és opcional.



Figura 4.5: Esquema de la nostra taula

La taula 4.1 que acabem de crear està formada per 8 variables, 6 de les quals són numèriques i 2 categòriques, i es descriuen tal com es pot veure a continuació.

En el transcurs del treball, farem ús d'aquesta taula, juntament amb dues més, que prenen de nom de persones i vehicles, per analitzar les dades que es van prendre d'un conjunt d'accidents de trànsit que es van donar a Washington, Estats Units, durant el mes de desembre de l'any 2015. L'agència estatal de trànsit va fer públiques aquestes dades al seu portal web (<https://www.transportation.gov/briefing-room/traffic-fatalities-sharply-2015>).

Variable	Tipus	Descripció
ST_CASE	Numèrica	Codi de l'accident
DAY	Numèrica	Dia de l'accident (de l'1 al 31)
HOURL	Numèrica	Hora de l'accident (99 = desconeguda)
MINUTE	Numèrica	Minut de l'accident (99 = desconegut)
RUR_URB	Categòrica	Informació sobre la localització (1 = Rural, 2 = Urbà, 6 = Via no classificada, 8 = No registrat, 9 = Desconegut)
DAY_WEEK	Categòrica	Dia de la setmana (1 = Diumenge, 2 = Dilluns, ..., 7 = Dissabte)
FATALS	Numèrica	Nombre de víctimes a l'accident (poden der ferits o morts)
DRUNK_DR	Numèrica	Nombre de conductors beguts involucrats a l'accident

Taula 4.1: Especificacions de la taula Accidents

4.3.1 Afegir dades a una taula de BigQuery senzilla

Ara que ja s'ha creat una taula de consulta, podem centrar-nos a treballar amb ella. Per a això, ens desplaçarem cap avall i donarem un cop d'ull al primer esquema de la taula (a la figura 4.6), on es troba a alguna informació interessant. Més enllà de la identificació de la taula, a l'esquerra de la figura, també es pot comprovar la grandària de la taula a la dreta, que ens donarà una indicació de la quantitat de dades que es processarien en cas d'executar consultes sobre aquesta. La grandària d'emmagatzematge a llarg termini assenyala les dades a les quals no s'ha accedit en els últims 90 dies, i després, per descomptat, tenim les hores de creació i modificació, juntament amb la ubicació de les dades de la taula. Des d'aquesta interfície també és possible editar els detalls existents d'aquesta taula. Aquí es permet l'establiment d'un temps de caducitat en cas que es vulgui anul·lar el que s'ha establert en el nivell del conjunt de dades. També hi ha l'opció d'establir una descripció o afegir etiquetes.

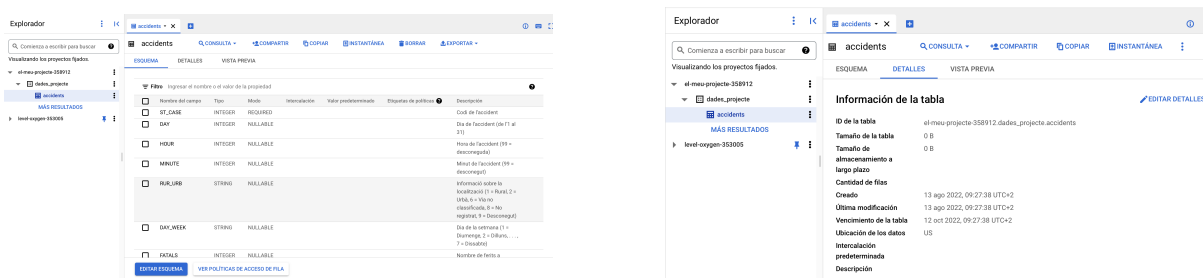


Figura 4.6: Detalls de la taula

Una altra característica la qual es pot consultar és la vista prèvia de la taula i, com és lògic, veurem que aquesta que acabem de crear encara no conté dades, ja que simplement hem creat l'esquema de la taula, sense inserir cap dada en aquesta. Si féssim ús de SQL, en qualsevol altre context es podrien afegir dades a partir d'una simple consulta a la taula, que tindria l'estructura següent:

```
INSERT INTO `el-meu-projecte-358912.dades_projecte.accidents`  
(ST_CASE, DAY, HOUR, MINUTE, RUR_URB, DAY_WEEK, FATALS, DRUNK_DR)  
VALUES (20055, 1, 20, 55, "1", "3", 3, 0);
```

A partir d'aquesta consulta afegiríem a la taula el cas d'un accident amb identificador 20055, que es va produir el dia 1 del mes a les 20:55 a una zona rural (RUR_URB = 1) un dimarts (DAY_WEEK = 3), i en el que hi va haver 3 víctimes i cap conductor begut involucrat en l'accident.

Això no obstant, quan intentem executar la consulta, BigQuery ens informa d'un error (Figura 4.7). Si recordem, prèviament s'han definit algunes de les limitacions per a l'ús de la zona de proves de BigQuery. Entre aquestes s'hi troba que no podem utilitzar el llenguatge de manipulació de dades (DML), és a dir, que no podem modificar la taula amb sentències com INSERT INTO, UPDATE o DELETE, per exemple. Per aquest motiu, l'error ens avisa que no tenim el nostre projecte vinculat a un compte i, per tant, no ens avaluarà la nostra consulta.

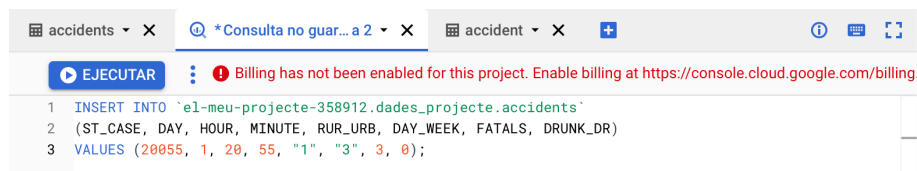


Figura 4.7: Inserció de dades a la taula

4.4 Càrrega de dades per crear una taula a BigQuery

Ara explorarem un cas d'ús més comú per als usuaris de BigQuery, que tracta de crear una taula a partir de dades existents. Per a això, ens dirigirem al nostre conjunt de dades, `dades_projecte`, i triarem crear una nova taula. Aquest cop, la font no serà una taula buida, sinó que carregarem un arxiu CSV del nostre propi sistema d'arxius. Un cop seleccionem importar les dades, es pot seleccionar diferents tipus d'arxiu, com ara CSV, JSON, Avro o Parquet, principalment. Per a respectar les limitacions de la zona de proves, hi ha algunes restriccions quant a la grandària de l'arxiu que es pot pujar, recordem que aquestes han de ser menors a 10 Gb. Procedim llavors a navegar pels nostres sistemes d'arxius per a buscar l'arxiu a pujar. Una vegada que l'arxiu ha estat seleccionat, el format ha sigut identificat i s'ha establert automàticament en CSV (Figura 4.8).

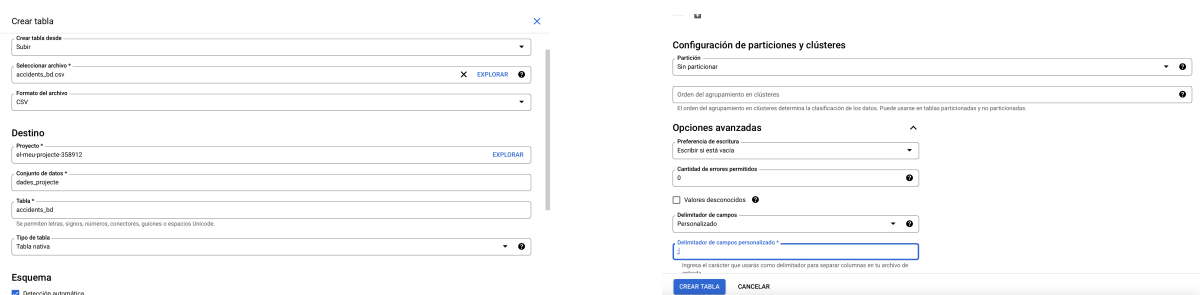


Figura 4.8: Lectura d'un arxiu extern

Quant al projecte i al conjunt de dades, els deixarem com estan, i escollirem el nom de la taula, `accidents_bd`, el qual farà saber que inclou informació sobre diversos accidents de trànsit. A continuació, es troba l'opció de definir explícitament l'esquema. Això no obstant, atès que es tracta d'un arxiu CSV amb múltiples columnes, es pot triar l'opció de detectar automàticament. D'aquesta manera, BigQuery donarà un cop d'ull al contingut de cada columna i determinarà quin ha de ser l'esquema.

Més enllà d'això, al final de la finestra de creació de la taula ens apareixeran unes opcions avançades. Aquestes opcions permeten la lectura de diferents tipus de CSV, entre altres coses. Sabem que el delimitador de camps d'un CSV pot ser una tabulació o una coma entre altres possibilitats. En el nostre cas, cal especificar que el nostre tabulador és el punt i coma “;”.

Si hem establert totes aquestes especificacions, ja podem començar amb l'anàlisi.

Ara es pot comprovar que `accidents_bd` apareix dins la nostra base de dades, `dades_projecte`. A continuació, podem accedir a la informació de la taula i al seu contingut desplegant el menú i triant Obrir. En l'esquema de la taula, sortirà que s'ha detectat automàticament el tipus dels diferents camps. Donem un cop d'ull als detalls de la taula. Aquí es pot veure que la grandària total és de poc més de 280 kB. El nombre de files és d'unes 2.780. I després, quan ens dirigim a la vista prèvia, tenim un primer contacte amb el contingut (Figura 4.9).

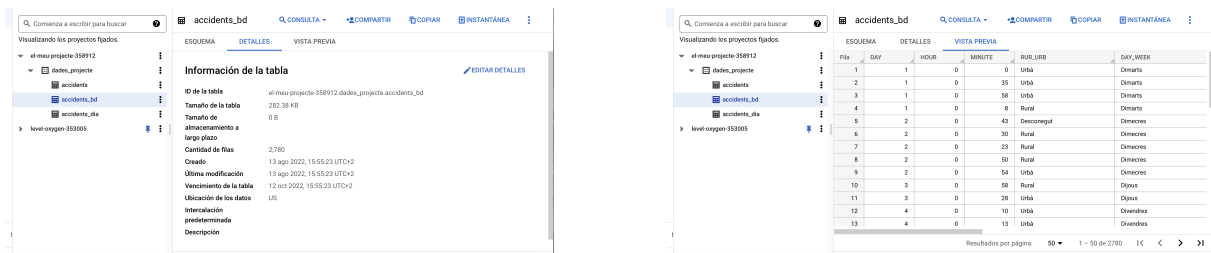


Figura 4.9: Informació sobre la taula

4.5 Consulta de dades i visualització d'estadístiques de consultes

Per a executar consultes en aquesta taula, ens dirigim al botó de consulta i s'obrirà en una nova pestanya. Aquesta pot ser una pestanya completament nova, és a dir, ocultant aquesta vista de detalls, o bé pot ser una pestanya dividida, que ens permetrà veure alhora aquesta vista de detalls de la taula mentre construïm una consulta. Mitjançant aquest procés, ha aparegut una nova pestanya cap a la dreta, i la consulta que apareix per defecte inclou una clàusula SELECT, però no inclou cap camp (Figura 4.10). Precisament per això hi ha un error de sintaxi, com es mostra a la dreta de la figura.



Figura 4.10: Elaboració d'una consulta

Ara, per a completar la clàusula SELECT, podríem escriure els noms dels atributs que volem consultar, a més de condicions, per exemple, mitjançant la clàusula WHERE. Concretament, l'esquema que s'haurà de seguir per a consultar la base de dades té la forma següent:

```
SELECT atribut1, atribut2,...
FROM '[nom_projecte].[nom_base_de_dades].[nom_taula]'
(WHERE condició)
```

Si escrivim a l'editor la nostra consulta, apareixerà un validador d'aquesta a la part superior dreta de la finestra. Aquest validador l'hem vist anteriorment quan ens indicava un error en voler executar una consulta fent servir llenguatge de manipulació de dades (DML), i pot agafar dues formes:

- Si la consulta és vàlida, apareixerà una icona de verificació en color verd.
- Si la consulta no és vàlida, apareixerà una icona d'exclamació en color vermell.

A més, el validador també mostra la quantitat de dades que la consulta processarà quan s'executi. Per exemple, si demanem en una consulta que ens retorni la columna sencera DAY, el validador

INFORMACIÓN DEL TRABAJO	RESULTADOS	JSON	DETALLES DE LA EJECUCIÓN
ID de trabajo	el-meu-projecte-358912:US.bqxjob_784c0bbb_1829bc02edc		
Usuario	salazar.correu@gmail.com		
Ubicación	US		
Hora de creación	14 ago 2022, 11:47:20 UTC+2		
Hora de inicio	14 ago 2022, 11:47:20 UTC+2		
Hora de finalización	14 ago 2022, 11:47:20 UTC+2		
Duración	0 s		
Bytes procesados	21.72 KB		
Bytes facturados	10 MB		
Prioridad del trabajo	INTERACTIVE		
Usar SQL heredado	false		
Tabla de destino	Temporary table		

Figura 4.11: Primera consulta

de la dreta ens marca que es processaran una quantitat de gairebé 22 kB, tal com es pot veure a la figura 4.11.

Efecturem aquesta consulta prement *Ejecutar*. Els resultats apareixen sota la finestra d'editor i mostra certs detalls com, per exemple, que la consulta s'ha executat en uns 0 segons. Per descomptat, podem desplaçar-nos i donar un cop d'ull a tots els resultats (a la dreta de la Figura 4.11). Entre els detalls que es mostren, per a cada consulta s'hi troben la informació del treball, els resultats en forma de taula, els resultats en format JSON i certs detalls sobre l'execució de la consulta.

Si posem el focus en la informació del treball, es troba l'identificador d'aquest, l'usuari que l'ha executat, la ubicació on s'emmagatzemen les dades, l'hora de la creació i l'execució d'aquesta, el temps d'execució i els bytes processats i facturats. Veiem que el nombre de bytes facturats és de 10 MB, aquesta és la quantitat mínima que surt per defecte per a cada consulta a la plataforma de Google Cloud, i té en compte les despeses generals. Per a consultes a bases de dades més extenses, aquesta facturació serà major i impedirà l'ús de la zona de proves. L'última característica que crida l'atenció de la informació del treball és que els resultats s'emmagatzemen en una taula temporal. Això vol dir que aquesta taula resultant no es guardarà com una més en el nostre conjunt de dades i, per tant, no la podrem consultar.

4.6 Creació d'una taula a partir d'un resultat de consulta

Una altra alternativa a les taules temporals serà crear una taula en el nostre conjunt de dades a partir d'una taula temporal o vista. En aquest cas, farem una sèrie temporal de 31 observacions que ens compti el nombre d'accidents ocorreguts cada dia del mes a partir de la consulta següent:

```
SELECT DAY, COUNT(*) AS FREQ
FROM 'el-meu-projecte-358912.dades_projecte.accidents_bd'
GROUP BY DAY
```

Un cop realitzada la consulta, s'han d'exportar totes aquestes dades a una nova taula i, per a fer-ho, revisarem algunes de les opcions d'exportació que es troben al menú *Guardar resultados*. En aquest, apareixen diverses opcions per a la manera de guardar els resultats (Figura 4.12). Es poden guardar com un arxiu CSV a Google Drive o en una carpeta local. En el nostre cas, exportarem el contingut a una nova taula de BigQuery. Una vegada feta aquesta selecció, podem decidir el nom del projecte i el conjunt de dades on s'aprovisionarà la taula i, a més, establir un nom de taula.

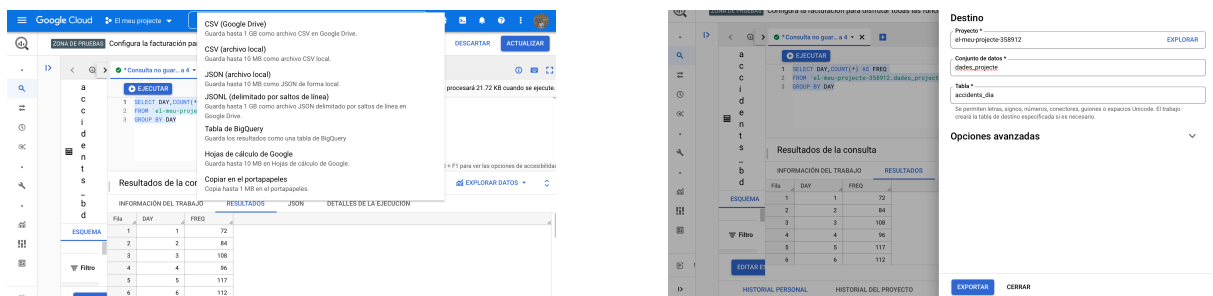


Figura 4.12: Creació d'una taula a partir d'una consulta

Llavors, quan guardem les coses, això haurà iniciat un treball per a guardar la nova taula i carregar-la amb dades. Un cop tanquem aquesta notificació, podem treure el pin de l'Explorador i, avall, hi apareix com una taula. En obrir-la confirmem que l'esquema apunta a les mateixes columnes que havíem referenciat en la clàusula SELECT de la consulta que va crear aquesta taula. Des dels detalls es pot confirmar que el nombre de files coincideix amb el dels resultats de la consulta, concretament 31. I, a més, la vista prèvia mostra quines són les dades exactament.

Hom es pot preguntar: *Quina és la finalitat d'aquesta taula?* Bé, atès que només conté un subconjunt de la taula original, les consultes cap a aquesta taula tindran potencialment menys dades per a processar que les consultes que s'executen directament a la taula original. Potser, en aquest cas no tornarem a necessitar aquesta consulta, però imaginem que només ens interessa estudiar els accidents que s'hagin produït a zones rurals i creem una consulta que filtri aquest tipus de casos. En aquesta situació, serà millor executar la resta de consultes sobre la taula petita, que inclou tota la informació que necessitem i, probablement, tindrà un cost de consulta menor.

5 Dades públiques i dades externes

5.1 Conjunts de dades públiques a BigQuery

Mentre continuem familiaritzant-nos amb la plataforma, podem explorar una altra opció que ofereix BigQuery, que són dades que estan disponibles públicament per a que qualsevol usuari pugui executar les seves consultes. Aquests conjunts de dades públiques es troben en un projecte anomenat *BigQuery Public Data*. Per a accedir a elles, farem servir al botó d'agregació de dades (ADD DATA) que es troba al costat de l'explorador, i escollirem l'opció *Explorar conjuntos de datos públicos* (Figura 5.1).

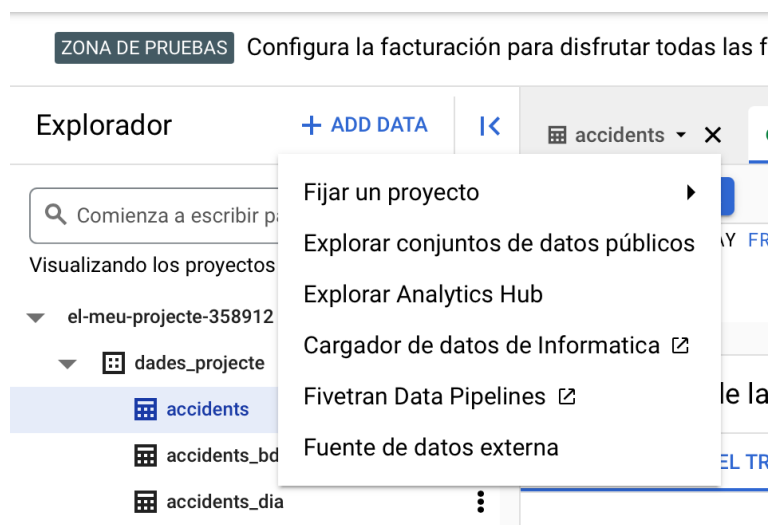


Figura 5.1: Accés als conjunts de dades públiques

Podem donar un cop d'ull als diferents conjunts de dades i taules d'aquest projecte i navegar entre les diferents opcions per a poder triar aquells que cridin la nostra atenció i, en definitiva, amb els que treballarem. Per exemple, entre totes aquestes taules s'hi troba una que conté informació sobre la població catalana recopilada per l'aplicació *GenCat Mobile Coverage* (Figura 5.2). Dins d'aquesta taula s'hi troben dades recollides mitjançant crowdsourcing¹, que contenen informació sobre l'estat de la cobertura de la telefonia mòbil a Catalunya. La plataforma utilitza una aplicació Android per a registrar les dades dels ciutadans a través dels seus dispositius mòbils sobre el nivell de cobertura per operador, la xarxa utilitzada (2G, 3G i 4G) i la ubicació del dispositiu. Aquestes dades en concret van ser recopilades durant els anys 2015-2017.

Si volguéssim descarregar la taula i situar-la dins del nostre projecte, BigQuery ens permet aquesta operació prement la opció *Ver conjunto de datos*, seguida de *crear tabla*. Això no

¹La pràctica d'obtenir informació o aportacions a una tasca o projecte recorrent als serveis d'un gran nombre de persones, remunerades o no, normalment a través d'Internet.

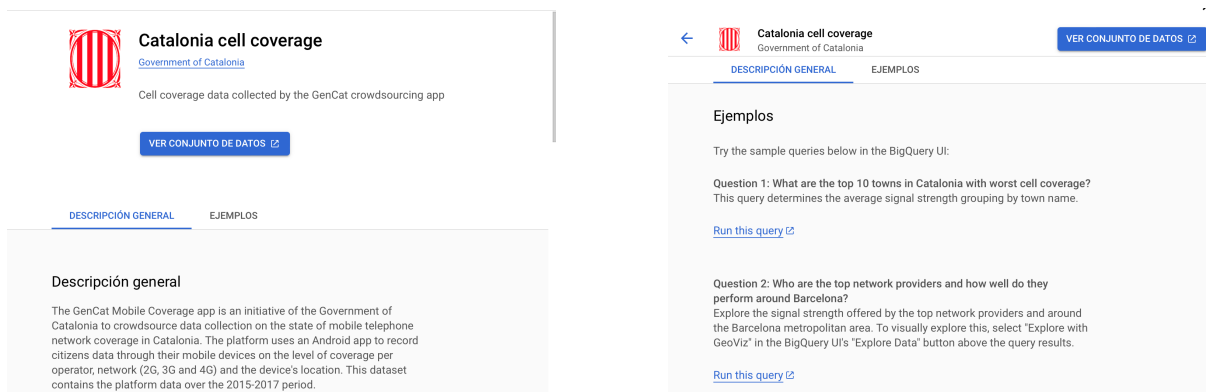


Figura 5.2: Creació d'una taula a partir d'una consulta

obstant, també és possible consultar la taula sense necessitat de descarregar-la dins del nostre projecte, fent servir el projecte `bigquery-public-data`, que és el que farem en aquest cas.

Dins la informació de la taula que ens proporciona la plataforma, apareix una secció amb consultes suggerides (les podem trobar a la dreta de la Figura 5.2). D'entre aquestes, la primera fa referència als 10 pobles catalans amb pitjor cobertura de mòbil. Si premem el botó *Run this query*, el navegador ens redirigeix a l'editor amb la consulta preparada, i nosaltres la podem executar sense cap cost, ja que tant la taula com el resultat de la consulta estan emmagatzemats a BigQuery i les dades no s'han de processar (Figura 5.3).

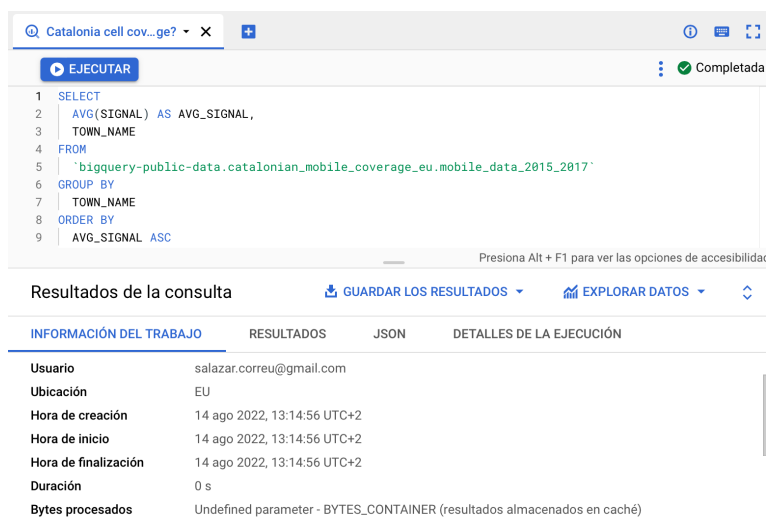


Figura 5.3: Consulta a un conjunt de dades públic

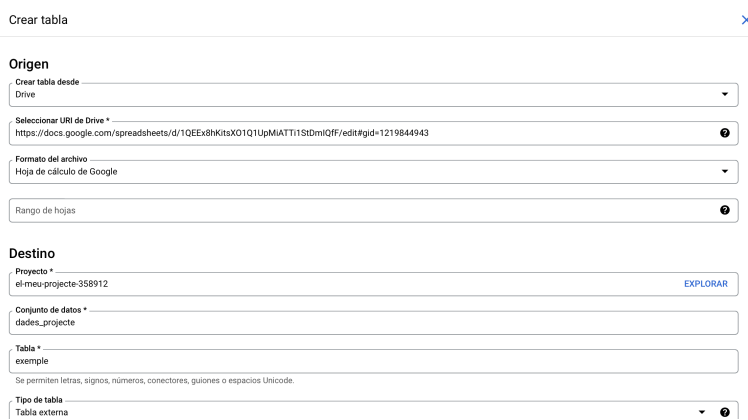
Els resultats d'aquesta consulta assenyalen que els pobles catalans amb menys cobertura mòbil són Canejan, Boadella i les Escaules, Cabó, la Vajol, Gaià, Farrera, Gisclareny, Viver i Serrateix, Savallà del Comtat i Torroja del Priorat. A més, els bytes processats en aquesta consulta apareixen com un paràmetre indefinit, ja que aquests resultats estan emmagatzemats a la memòria cau de BigQuery.

Tal com hem vist, BigQuery emmagatzema a la seva memòria cau els resultats de la consulta perquè les dades puguin ser recuperades més ràpidament la pròxima vegada que s'executi la mateixa consulta. Això no obstant, cal tenir en compte que només s'accedeix a les dades de la memòria cau quan s'executa exactament la mateixa consulta després de la seva creació. Per exemple, si es modifiqués una mica aquesta consulta i demanés tan sols el nom del poble, en comptes d'aquest i la mitjana del senyal mòbil, podríem arribar a pensar que els resultats de la consulta d'aquesta execució haurien de retornar-nos un subconjunt de les dades que ja són presents en la memòria cau (ja que és un subconjunt de la nostra consulta anterior). Però, per la forma en què funciona la memòria cau de BigQuery, quan executem això, observarem que la informació no s'ha recuperat d'aquesta. En canvi, quan aquesta consulta es torna a executar, és quan la memòria cau s'activa, i és d'on es recuperen les dades.

Per tant, la memòria cau només funciona si és la mateixa consulta la que es torna a executar. Aquestes característiques són de gran interès, ja que l'emmagatzematge en memòria cau és una gran manera de reduir el cost d'execució de les consultes, i també millora el rendiment d'aquestes.

5.2 Taules externes de BigQuery

Una altra funció que presenta la plataforma és la lectura d'arxius externs que s'actualitzen de forma periòdica. Això és d'especial utilitat en els casos en què la informació de la qual es disposa és a temps real, que és una característica prou habitual quan es treballa amb volums de dades molt elevats. Per a il·lustrar el funcionament de BigQuery en aquests casos, crearem una nova taula que estarà vinculada, en aquest cas, a Google Drive, concretament als fulls de càlcul de la plataforma (Figura 5.4). És molt important que el propietari de la taula sigui el mateix compte que està vinculat a BigQuery, perquè en cas contrari sorgeix un missatge d'error i no és capaç de vincular la taula externa.



Crear tabla

Origen

Crear tabla desde
Drive

Seleccionar URI de Drive *
<https://docs.google.com/spreadsheets/d/1QEEEx8HktsX01Q1UpMATT11StDmIQf/edit#gid=1219844943>

Formato del archivo
Hoja de cálculo de Google

Rango de hojas

Destino

Proyecto *
el-meo-projecte-358912 [EXPLORAR](#)

Conjunto de datos *
dades_projecte

Tabla *
exemple

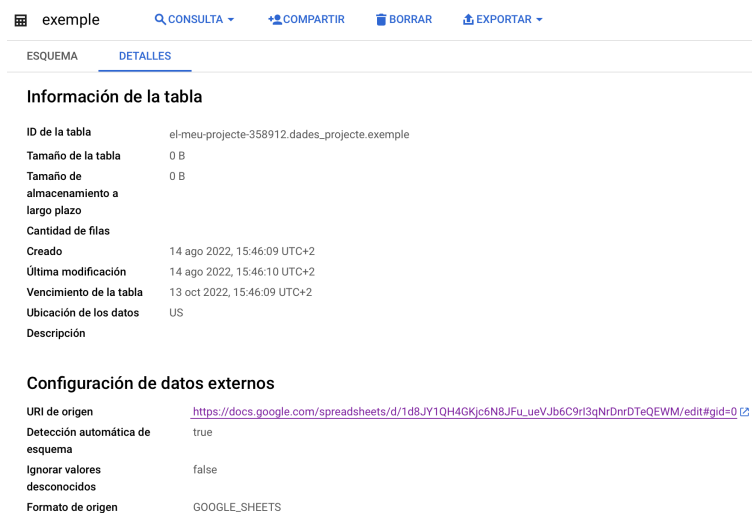
Se permiten letras, signos, números, conectores, guiones o espacios Unicode.

Tipo de tabla
Tabla externa

Figura 5.4: Connexió a una taula externa

Si ens dirigim als detalls, aquí és on veiem quelcom interessant. La grandària de la taula és de zero bytes, atès que les dades són externes a BigQuery (Figura 5.5). Si ens desplaçem, es poden veure els detalls de les dades externes. Això significa que quan s'actualitza el full de càlcul, qualsevol consulta cap a aquesta taula recollirà automàticament les dades més recents, ja que la nostra consulta de la taula gran no és només una còpia del full de càlcul, sinó que és de fet una referència a ella.

Quan una consulta s'executa, totes les dades són retornades a nosaltres, i podem accedir a elles com ho faríem amb qualsevol dada que resideixi en una taula nativa de BigQuery.



The screenshot shows the BigQuery interface for a table named 'exemple'. At the top, there are navigation options: 'CONSULTA', 'COMPARTIR', 'BORRAR', and 'EXPORTAR'. Below this, there are two tabs: 'ESQUEMA' and 'DETALLES', with 'DETALLES' being the active tab. The main content is divided into two sections: 'Información de la tabla' and 'Configuración de datos externos'.

Información de la tabla	
ID de la tabla	el-meu-projecte-358912.dades_projecte.exemple
Tamaño de la tabla	0 B
Tamaño de almacenamiento a largo plazo	0 B
Cantidad de filas	
Creado	14 ago 2022, 15:46:09 UTC+2
Última modificación	14 ago 2022, 15:46:10 UTC+2
Vencimiento de la tabla	13 oct 2022, 15:46:09 UTC+2
Ubicación de los datos	US
Descripción	

Configuración de datos externos	
URI de origen	https://docs.google.com/spreadsheets/d/1d8JY1QH4GKjc6N8JFu_ueVJb6C9rI3qNrDnrDTeQEWM/edit#gid=0
Detección automática de esquema	true
Ignorar valores desconocidos	false
Formato de origen	GOOGLE_SHEETS

Figura 5.5: Característiques d'una taula externa

6 Integració de BigQuery amb Looker Studio

Ara que hem cobert els diferents tipus de taules de BigQuery, ens centrarem en com es poden visualitzar les consultes.

Looker Studio és una eina gratuïta que permet convertir les dades en panells o informes complets, fàcils de llegir i de compartir, i totalment personalitzables. Algunes de les seves funcionalitats són:

- Descriure les dades amb gràfiques, que inclouen gràfics de línies, de barres i circulars (o *pie plots*), mapes geogràfics, gràfics d'àrea i de bombolles, taules de dades dinàmiques i molt més.
- Permet que els informes siguin interactius amb filtres de visualització.
- Inclou enllaços i imatges en les quals es pot clicar per crear catàlegs de productes, biblioteques de vídeo i altres continguts amb URL.
- Facilita l'anotació i descripció dels informes amb text i imatges.

A més de presentar totes aquestes característiques, amb Looker Studio es poden elaborar fàcilment informes sobre dades procedents d'una gran varietat de fonts, sense necessitat de programar. En tan sols uns instants, es pot connectar a conjunts de dades com BigQuery.

6.1 Ús de Looker Studio des de BigQuery

Imaginem que volem tornar a consultar el nombre d'accidents de trànsit que es van donar cada dia durant aquell mes als Estats Units. Per a fer aquesta consulta, farem servir la *query* anterior:

```
SELECT DAY, COUNT(*) AS FREQ
FROM 'el-meu-projecte-358912.dades_projecte.accidents_bd'
GROUP BY DAY
```

Per a visualitzar la taula resultant, podem ampliar el menú *Explorar datos*, tot seguit d'*Explorar con Looker Studio* (Figura 6.1).

Una vegada feta aquesta selecció, sorgeix una nova interfície que ja té una taula que conté alguna informació i un histograma amb les dades d'aquesta (Figura 6.2).

Looker Studio, per defecte, ha entès que una taula de recompte es visualitza normalment a partir d'un gràfic de barres vertical o histograma, i, per això, l'ha creat sense que nosaltres ho hàgim especificat. Així i tot, nosaltres podem seleccionar una visualització abans de configurar-la per a presentar la informació que necessitem. Pose'm-nos en el cas que preferim un gràfic de barres horitzontal per a la visualització d'aquestes dades. Premem l'opció *Añadir un gráfico* i

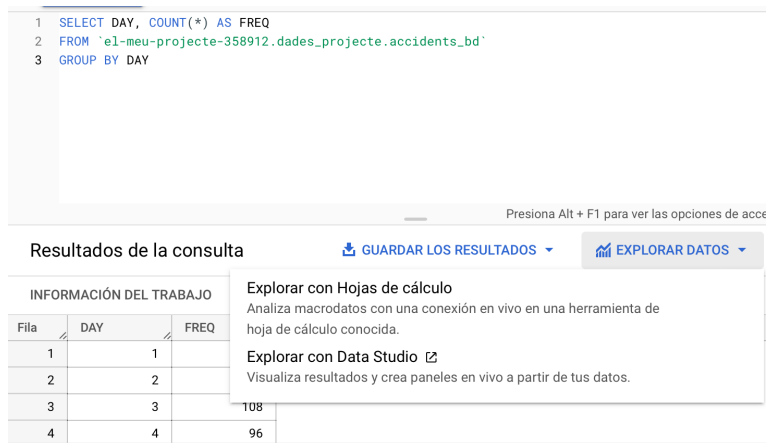


Figura 6.1: Visualització d'una consulta a Looker Studio

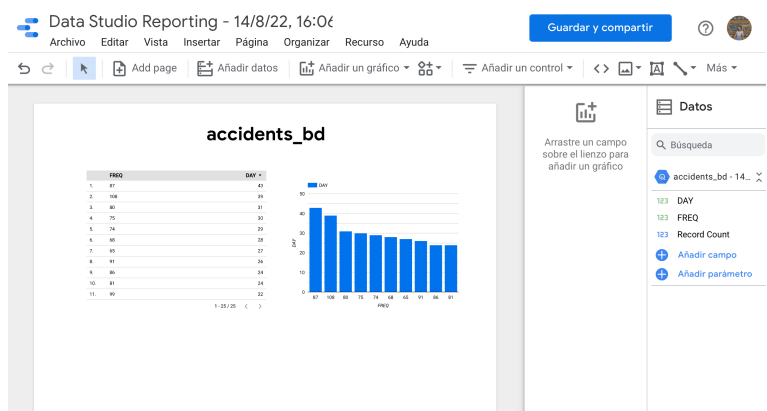


Figura 6.2: Consulta des de Looker Studio

ens assegurem que la dimensió, que en aquest cas són els dies, i la mètrica, la freqüència dels accidents, estiguin seleccionades segons el que vulguem representar (Figura 6.3).

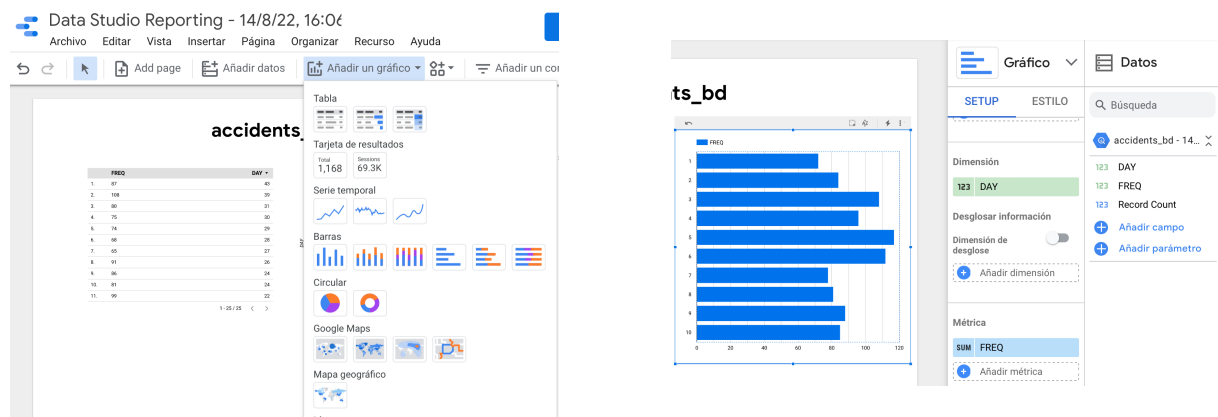


Figura 6.3: Creació d'un gràfic a partir de Looker Studio

6.2 Ús de Looker Studio des de la pròpia plataforma

Per a utilitzar aquesta eina és necessari disposar d'un compte a Google. Per accedir a la pàgina, naveguem a [Looker Studio](#) i iniciem sessió amb el nostre compte de Google. Un cop dins, es troba una pantalla d'inici amb les característiques següents:

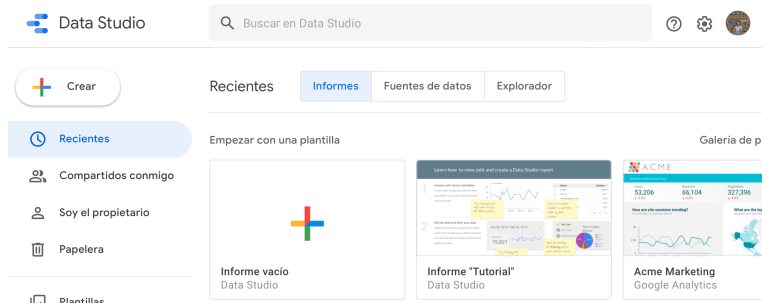


Figura 6.4: Inici de Looker Studio

A la pàgina d'inici hi ha una sèrie de plantilles, que són una forma entretinguda d'explorar les capacitats de Looker Studio. En el nostre cas, com el que ens interessa és crear un informe des de zero, clicarem a *Informe vacío*. Un cop dins l'informe en blanc, s'hauran d'afegir les dades que volem representar en aquest (Figura 6.5). Tenim moltes opcions a l'hora d'escollir la font de les dades, però per a fer-ho més senzill vincularem l'informe a BigQuery, específicament al nostre conjunt de dades `accidents_bd`.

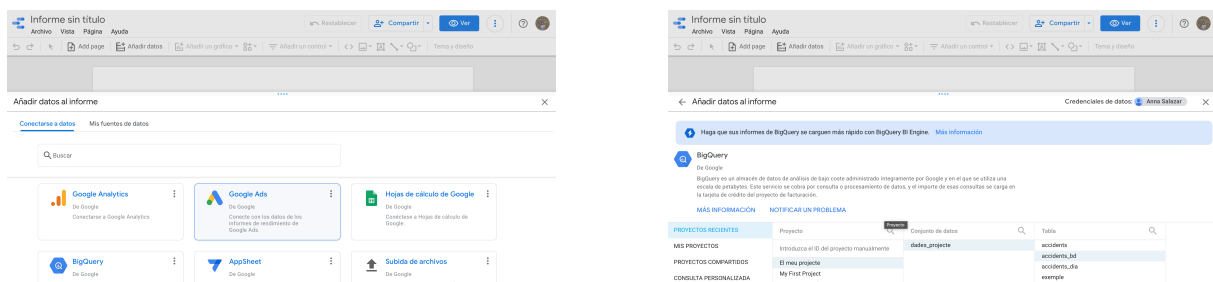


Figura 6.5: Afegir dades a Looker Studio

Per a crear un gràfic, podem clicar a *Añadir un gráfico* i, d'aquesta manera, escollir el que ens interessa representar en el panell de la dreta, ajustant la dimensió i la mètrica del nostre interès. Un cop creat el gràfic, es pot editar la seva mida fent-lo més gran o més petit, segons la nostra preferència, i es pot moure de lloc dins la pàgina.

A més de gràfiques, es poden decorar les pàgines per fer-les més boniques, o per afegir informació. També és possible assignar un nom a la pàgina, i crear-ne tantes com siguin necessàries per al nostre informe. Així mateix, podem establir un nom per a tot l'informe, de manera que a l'hora de guardar-lo i compartir-lo sigui més fàcilment interpretable.

Per exemple, un informe de dues pàgines pot tenir un aspecte semblant a la Figura 6.6.

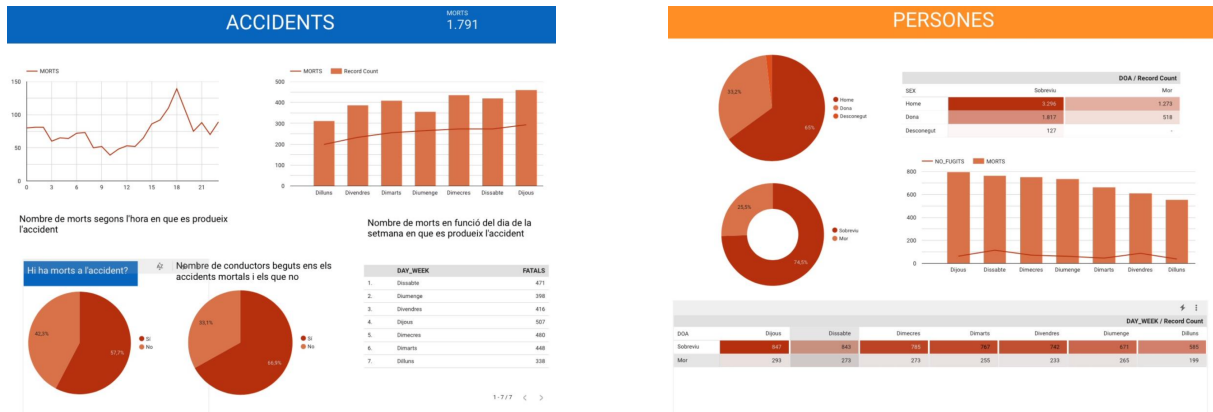


Figura 6.6: Informe a Looker Studio

Per acabar, podem compartir el nostre informe perquè altres usuaris puguin tenir accés. Això es pot fer des de la pantalla d'inici de Looker Studio, seleccionant l'opció de compartir l'informe (Figura 6.7) i afegint les adreces de correu dels qui vulguem fer lectors de l'informe (aquells qui poden veure l'informe, però no tenen permisos d'edició) o bé editors (poden veure i editar l'informe).

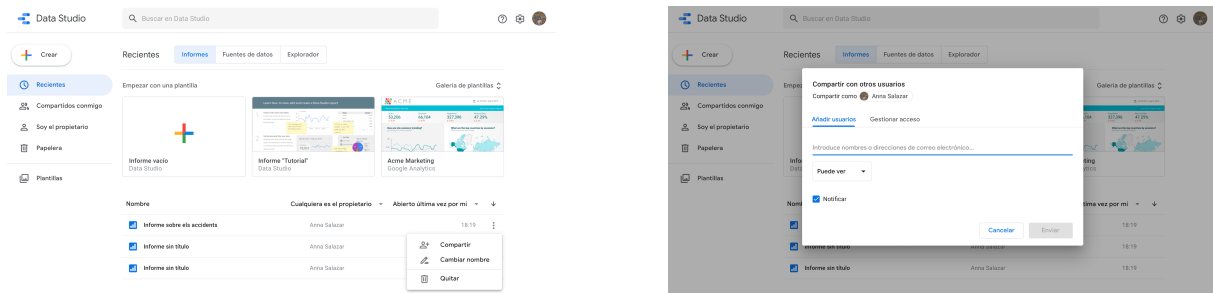


Figura 6.7: Compartir un informe

7 Connexió d'R a BigQuery

7.1 Connectem R a BigQuery

Per a entendre com funciona la connexió entre BigQuery i R, primer s'ha de definir el concepte d'API (*Application Programming Interface*) o interfície de programació d'aplicacions.

Una API és un conjunt de processos, funcions i mètodes que ofereix una determinada biblioteca de programació i que serveix com a capa d'abstracció perquè sigui emprada per un altre programa informàtic. Dit d'una altra manera, podem entendre les APIs com un codi que indica a les aplicacions com comunicar-se entre elles, de manera que puguin interactuar.

En el nostre cas, volem que R sigui capaç d'interactuar amb BigQuery, perquè així ens permeti realitzar les nostres consultes a la base de dades sense haver d'importar les dades a R. Haurem d'utilitzar les llibreries `bigrquery`¹ i `DBI`² per aconseguir-ho.

```
install.packages("bigrquery")
install.packages("DBI")
library(bigrquery)
library(DBI)
```

Asignem a l'objecte `projecte` el nom del nostre projecte, que a l'exemple correspon a *el-meu-projecte-358912*.

```
projecte <- "el-meu-projecte-358912"
```

I per últim, mitjançant la funció `dbConnect()` obrirem la connexió amb Bigquery indicant el nom del nostre projecte i la base de dades a la qual volem accedir dins d'aquest, `dades_projecte`. A l'apartat *billing* s'haurà d'introduir l'identificador del projecte amb la font de facturació. Nosaltres indiquem el nom del nostre projecte, que està sotmès a les limitacions de la *Sandbox* o zona de proves gratuïta.

```
dades <- dbConnect(
  bigrquery::bigquery(),
  project = projecte,
  dataset = "dades_projecte",
```

¹Una interfície per a la API de 'BigQuery' de Google

²Una definició d'interfície de base de dades per a la comunicació entre R i els sistemes de gestió de bases de dades relacionals (SGBD). Totes les classes d'aquest paquet són virtuals i han de ser esteses per les diferents implementacions de R/SGBD.

```
    billing = projecte
  )
```

Executant aquest codi no es produeix gran cosa, tret que es crea una variable de connexió. Però la primera vegada que intentem fer ús d'aquesta connexió (per exemple, fent una consulta a una de les taules de la base de dades), se'ns demana que ens autentifiquem a través del nostre compte de Google en una finestra del navegador. Un cop fet això, ja podrem començar a consultar les nostres dades de BigQuery, així com els conjunts de dades públics.

7.2 Consultes amb el paquet “bigrquery”

Per a mostrar com seria una consulta a la nostra taula des de R buscarem el llistat d'accidents que van ocórrer un cap de setmana a les 21:00.

```
query1 <- "SELECT ST_CASE
          FROM 'el-meu-projecte-358912.dades_projecte.accidents_bd'
          WHERE (DAY = 1 or DAY = 7) and HOUR = 21 and MINUTE = 0"
```

Un cop hem formulat la nostra *query*, a través de la funció `dbGetQuery()`, ens comunicarem amb l'API de Google. A l'esquerra s'hi troba l'esquema de la comunicació, mentre que a la dreta està la funció utilitzada.

```
dbGetQuery(connexió, consulta, n)      dbGetQuery(dades, query1, n = 10)
```

Essent n el nombre de casos a mostrar (si calen restriccions).

En executar aquesta consulta, R retorna la següent taula:

ST_CASE
62353

Aquesta correspon a que únicament va ocórrer un accident un cap de setmana a les 21:00, i aquest té l'identificador 62353.

8 Implementació de BigQuery a l'aula

BigQuery és una eina d'anàlisi de dades d'alt rendiment, capaç de processar un volum molt gran d'informació en segons, és senzilla d'utilitzar i ofereix una zona de proves des de la que es pot treballar de forma gratuïta. És per aquest motiu que s'ha presentat aquest programa als estudiants del grau d'estadística a l'assignatura de Fitxers i Bases de Dades, perquè ells mateixos puguin avaluar la seva experiència amb la plataforma de Google Cloud, així com el grau de dificultat que troben en el seu ús.

Per portar a terme aquesta dinàmica es van necessitar dues sessions de dues hores cadascuna. A la primera, es va presentar tota la informació teòrica sobre la plataforma. S'explicava les propietats de BigQuery, així com les limitacions que presenta la seva zona de proves. La professora Montserrat Guillén va fer unes diapositives per presentar-les a l'aula on s'incorporava tota la informació necessària per crear-se un compte a BigQuery i poder dur a terme l'activitat proposta. Aquestes diapositives van ser creades a partir de la recerca contemplada en aquest treball, i en elles es van fer algunes modificacions a l'hora de penjar-les al campus virtual, per evitar publicar un material que encara no s'havia presentat.

8.1 Consultes amb “bigrquery”

Per portar a terme la sessió pràctica es va fer ús del conjunt de dades públic prèviament esmentat: *Catalonia Cell Coverage*, que conté informació sobre l'estat de cobertura de la telefonia mòbil que es va recollir en el període dels anys des del 2015 fins al 2017.

En primer lloc, es presentava la taula de dades amb la qual s'havia de treballar, que conté les variables següents:

Variable	Descripció
date	Data de telemetria en format AAAA-MM-DD
hour	Hora de telemetria en format HH24:MM:SS
lat	Latitud
long	Longitud
signal	Senyal mitjana
network	Nom de la xarxa
operator	Nom de l'operador
status	Codi de l'estat = {0, 1, 2, 3}
description	Descripció de l'estat - En servei (0) - Fora de servei (1) - Estat d'emergència (2)

net	- Apagat Tipus de xarxa = 2G, 3G, 4G
speed	Velocitat estimada de la font
satellites	Nombre de satèl·lits GPS
precision	Constant que decriu la precisió del proveïdor
provider	Nom del proveïdor de la posició
activity	Activitat de l'usuari: - En un vehicle - Parat - A peu - Inclinat - Amb bicicleta - Desconegut
downloadSpeed	Velocitat de descàrrega actual
uploadSpeed	Velocitat de càrrega actual
postal_code	Codi postal
town_name	Nom de la ciutat on es va adquirir la telemetria
position_geom	Representació gràfica de la posició

Taula 8.1: Variables de les dades Catalonia Cell Coverage

Un cop els estudiants es trobaven a l'entorn BigQuery, les consultes a crear havien de resoldre les demandes següents:

1. Feu una taula de freqüències de l'activitat que fa cada usuari segons les dades recollides.

```
SELECT activity, count(date) AS Freq
FROM 'bigquery-public-data.catalonian_mobile_coverage_eu.
      mobile_data_2015_2017'
GROUP BY activity
```

2. Retorneu un llistat de les 5 ciutats amb millor connexió mitjana de Catalunya.

```
SELECT AVG(SIGNAL) AS AVG_SIGNAL, TOWN_NAME
FROM 'bigquery-public-data.catalonian_mobile_coverage_eu.
      mobile_data_2015_2017'
```

```
GROUP BY TOWN_NAME
ORDER BY AVG_SIGNAL DESC
LIMIT 5
```

3. Feu un resum numèric (mínim, mitjana, desviació típica i màxim) de la precisió de la xarxa per aquelles 10 que tenen la mitjana més elevada.

```
SELECT network, min(percission) AS Minim, avg(percission) AS
Mitjana, stddev(percission) AS Desv_Tipica, max(percission) AS
Maxim
FROM 'bigquery-public-data.catalonian_mobile_coverage_eu.
mobile_data_2015_2017'
GROUP BY network
ORDER BY Mitjana DESC
LIMIT 10
```

4. Creeu una taula que reculli les ciutats i la xarxa dels 10 primers usuaris que es van quedar sense servei mòbil, obviant els casos en què el nom de la ciutat o de la xarxa sigui *NULL*.

```
SELECT town_name, network
FROM 'bigquery-public-data.catalonian_mobile_coverage_eu.
mobile_data_2015_2017'
WHERE description="STATE_OUT_OF_SERVICE" AND
town_name IS NOT NULL AND network IS NOT NULL
ORDER BY date ASC
LIMIT 10
```

5. Construiu una taula que reculli la xarxa, el seu tipus i la precisió mitjana. Llisteu els 10 primers casos, en ordre descendent, segons la precisió mitjana.

```
SELECT network, net, avg(percission) AS Precisio_mitj
FROM 'bigquery-public-data.catalonian_mobile_coverage_eu.
mobile_data_2015_2017'
```

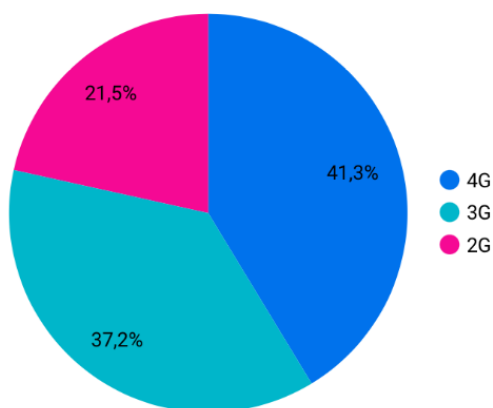


```

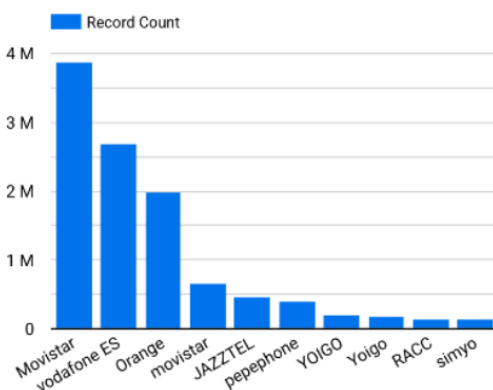
WHERE precision IS NOT NULL AND
      net IS NOT NULL AND network IS NOT NULL
GROUP BY network, net
ORDER BY Precisio_mitj DESC
LIMIT 10

```

6. Feu un gràfic de sectors sobre l'aparició dels diferents tipus de xarxa, excloent els casos en què aquesta variable pren el valor *NULL*.



7. Creeu un histograma que mostri la freqüència en què apareix cada operador a la base de dades. Utilitzeu Looker Studio per a fer-ho.



8.2 Comentaris dels estudiants sobre la pràctica

D'entre els estudiants matriculats a l'assignatura de Fitxers i Bases de Dades, van ser vint-i-un els qui van realitzar la pràctica i van valorar la seva interacció amb BigQuery. Les preguntes que van respondre aquests estudiants eren:

Responen a les següents preguntes sobre l'ús de BigQuery. Valoreu (0: molt fàcil - 10: molt difícil)

1. Entorn BigQuery (connexió)
 2. Entorn BigQuery (consultes SQL)
 3. Connexió amb R
 4. Valoració global
- Suggeriu alguna milora en la introducció a BigQuery?

A continuació es farà una recopilació dels resultats obtinguts, d'on es pot veure un resum gràfic a la figura 8.1.

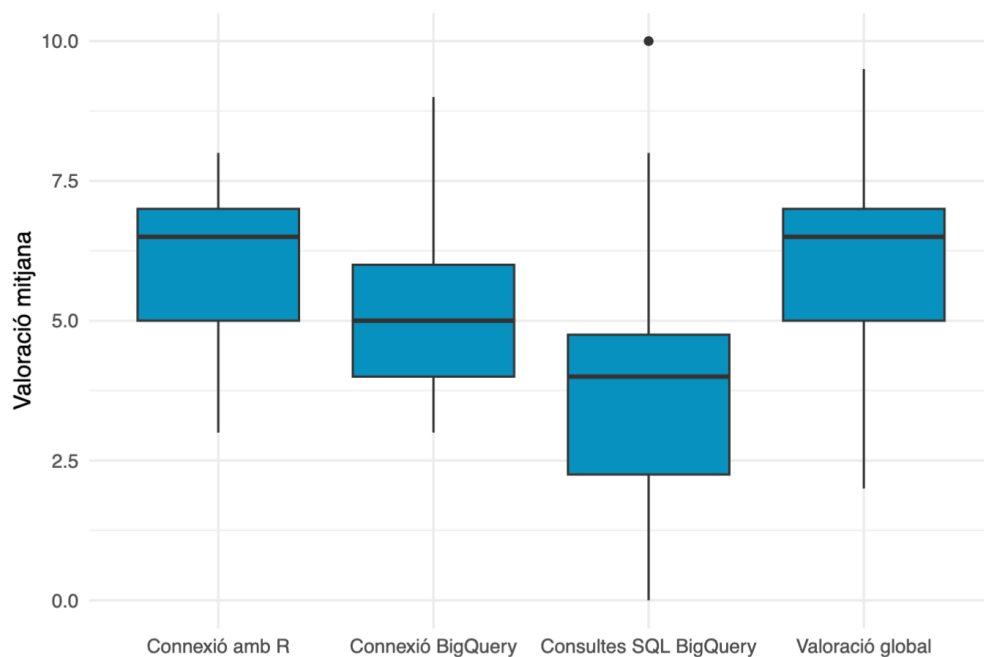


Figura 8.1: Valoració mitjana dels estudiants

Pel que fa a la connexió a BigQuery, la puntuació mitjana va ser de 5,44. Els estudiants no ho consideren fàcil, però tampoc li troben un grau de dificultat molt elevat a connectar-se a la

plataforma per primera vegada. Alguns comenten que, en ser un entorn nou, els hi costa perquè no hi estan familiaritzats i els fa sortir de la seva zona de confort.

En relació a les consultes SQL (*Structured Query Language*) a l'entorn BigQuery, la puntuació mitjana baixa a 4,06. És a dir, als estudiants, en general, els sembla més senzill l'ús de BigQuery que l'entrada a la plataforma. El llenguatge que s'utilitza a BigQuery és molt similar a d'altres vistos als programaris ACCES, R o SAS, i així ho fan saber els estudiants.

En quant a la connexió amb R, la puntuació mitjana ha sigut de 6, una mica més elevada que les anteriors. S'ha trobat una mica més de complicació en aquest punt, ja que el funcionament de les APIs no s'havia tractat prèviament i era un concepte nou.

Per últim, la valoració global ha estat de 6,09 en quant a la dificultat del programa.

També s'ha preguntat si els estudiants volien deixar suggerències per la millora en la introducció a BigQuery, aquelles coses que canviarien i/o els hi facilitaria la introducció a aquesta eina. Molts dels comentaris que s'han repetit en les valoracions és que els hi hagués ajudat el fet de dedicar-hi més pràctiques, perquè suposava un volum de informació molt gran que se'ls hi ha explicat en poc temps. A més, demanden que l'explicació de la teoria vingui acompanyada d'un suport pràctic des de la primera sessió, per entendre millor des del principi el funcionament de BigQuery.

Tot i que aquesta mostra d'estudiants difícilment és representativa, perquè només han participat vint-i-un estudiants a les sessions pràctiques, els resultats han sigut força positius i els estudiants han mostrat a les valoracions el seu interès en seguir coneixent aquest tipus d'eines, perquè els hi semblen interessants de cara al futur.

9 Anàlisi de dades mitjançant la connexió d'R i BigQuery

9.1 Descripció de la base de dades

La base de dades que serà utilitzada al llarg de l'estudi prové de l'agència estatal de trànsit dels Estats Units i conté tres taules, entre les quals es troba un llistat d'accidents de trànsit ocorreguts el desembre de 2015 als Estats Units, juntament amb un recompte de totes les persones (conductors, passatgers o vianants) involucrades als accidents i, finalment, un inventari de tots els vehicles també involucrats.

L'enllaç a la base esmentada és el següent: <https://www.transportation.gov/briefing-room/traffic-fatalities-sharply-2015>

Més concretament, en cada taula es poden trobar les variables següents:

Accident és un llistat d'accidents de trànsit ocorreguts al desembre de 2015 als Estats Units.

Variable	Tipus	Descripció
ST_CASE	Categòrica	Codi de l'accident
DAY	Categòrica	Dia de l'accident (de l'1 al 31)
HOUR	Numèrica	Hora de l'accident (99 = desconeguda)
MINUTE	Numèrica	Minut de l'accident (99 = desconegut)
RUR_URB	Categòrica	Informació sobre la localització (1 = Rural, 2 = Urbà, 6 = Via no classificada, 8 = No registrat, 9 = Desconegut)
DAY_WEEK	Categòrica	Dia de la setmana (1 = Diumenge, 2 = Dilluns, ..., 7 = Dissabte)
FATALS	Numèrica	Nombre de víctimes a l'accident (poden ser ferits o morts)
DRUNK_DR	Numèrica	Nombre de conductors beguts involucrats a l'accident

Taula 9.1: Llistat de variables de la taula Accident

Person és un llistat de totes les persones (conductors, passatgers o vianants) involucrades als accidents.

Variable	Tipus	Descripció
ST_CASE	Categòrica	Codi de l'accident al qual està involucrada la persona
PER_NO	Categòrica	Nombre de la persona dins de cada accident
AGE	Numèrica	Edat de la persona (998 = No registrada, 999 = Desconeguda)
SEX	Categòrica	Sexe de la persona (1 = home, 2 = dona, 8 = No registrat, 9 = Desconegut)
PER_TYP	Categòrica	Tipus de persona (1 = conductor, 2 = ocupant, resta de codis = altres)
DOA	Categòrica	Tipus de víctima (0 = sobreviu, 7 = mort a l'accident, 8 = mort al trasllat, 9 = Desconegut)

Taula 9.2: Llistat de variables de la taula Person

Vehicle és un llistat de tots els vehicles involucrats als accidents.

Variable	Tipus	Descripció
ST_CASE	Categòrica	Codi de l'accident al qual està involucrat el vehicle
NO_VEH	Numèrica	Nombre de vehicles implicats en l'accident
HIT_RUN	Categòrica	Identificador de vehicle fugit (0 = No, 1 = Sí, 9 = Desconegut)
TRAV_SP	Numèrica	Velocitat estimada (mph) del vehicle quan va tenir l'accident (997,998 i 999 = Desconegut)
PREV_SP	Categòrica	Indicador d'existència de límit de velocitat permesa just abans de l'accident (997,998 i 999 = Desconegut)

Taula 9.3: Llistat de variables de la taula Vehicle

9.1.1 Lectura de les dades

A partir d'aquestes tres taules, es definiran dues taules noves amb les quals es treballarà al llarg del projecte. La lectura i transformació de les taules originals s'ha fet mitjançant la connexió del programa R-Studio amb BigQuery, que és el lloc on s'hi trobaven guardades. Aquesta tècnica ha permès que l'ordinador no hagi de treballar directament amb el volum de dades complet, sinó amb un subconjunt d'aquestes.

En primer lloc, es tindrà en compte la informació dels accidents. D'aquesta manera es podrà estudiar les característiques dels diferents accidents registrats, així com es podrà fer prediccions sobre els nous accidents en funció de les seves característiques. S'ha anomenat aquesta base "accident", i està conformada per les variables següents: DAY, HOUR, MINUTE, RUR_URB, DAY_WEEK, FATALS, DRUNK_DR, NO_PER, MORTS, NO_VEHICLE, HIHAMORTS.

Les variables MORTS, NO_PER, NO_VEHICLE i HIHAMORTS han sigut creades a posteriori

a partir de les taules de les que es disposava mitjançant consultes SQL, i es defineixen a continuació:

Variable	Tipus	Descripció
MORTS	Numèrica	Nombre de morts en l'accident
NO_PER	Numèrica	Nombre de persones implicades en l'accident
NO_VEHICLE	Numèrica	Nombre de vehicles implicats en l'accident
HIHAMORTS	Categòrica	Variable identificadora dels accidents mortals (0: no hi ha morts en l'accident, 1: hi ha morts en l'accident)

Taula 9.4: Llistat de variables definides a posteriori per a la taula Accidents

D'altra banda, s'estudiarà la informació sobre les persones implicades en aquests accidents. D'aquesta manera es podrà perfilar el tipus de conductors en els casos en que hi hagi morts en l'accident, així com en els que no hi hagi. Aquesta informació també ens facilitarà l'elaboració de possibles models per predir el tipus de víctima que serà una persona involucrada en un accident de trànsit en base a les seves característiques. En aquest cas, s'ha anomenat aquesta taula "persones", i està conformada per les variables següents: DAY, HOUR, MINUTE, RUR_URB, DAY_WEEK, FATALS, DRUNK_DR, NO_PER, MORTS, NO_VEHICLE, NO_FUGITS, AGE, SEX, PER_TYP i DOA.

Las variable NO_FUGITS ha sigut creada a posteriori a partir de les taules de les que es disposava mitjançant consultes SQL, i es defineix a continuació:

Variable	Tipus	Descripció
NO_FUGITS	Numèrica	Nombre de vehicles fugits implicats en l'accident

Taula 9.5: Llistat de variables definides a posteriori per a la taula Persones

9.1.2 Objectius del projecte

Estudiant aquestes taules sobre persones que s'han vist implicades, de forma directa o indirecta, en accidents de trànsit es pretén:

- Descriure les característiques dels accidents registrats
- Tractar de forma adequada les dades mancants i els valors atípics de les taules
- Estudiar les relacions de dependència entre variables
- Desenvolupar un model de predicció que ens permeti establir si hi haurà víctimes mortals o no a un accident de trànsit en funció de les característiques que presenti aquest.

9.2 Preprocessament

La taula d'accidents està formada per 2078 casos (accidents) i 11 variables. En canvi, la taula de persones la conformen 7087 individus (files) i 15 variables (columnes).

Les variables que tenim són DAY, HOUR, MINUTE, RUR_URB, DAY_WEEK, FATALS, DRUNK_DR, NO_PER, MORTS, NO_VEHICLE, HIHAMORTS, NO_FUGITS, AGE, SEX, PER_TYP i DOA, definides prèviament.

9.2.1 Dades mancants

Per a poder tractar les dades mancants (o *missings*) d'ambues taules, en primer lloc haurem de transformar-les, ja que hi ha variables que presenten dades mancants, però les tenen codificades.

En el cas de les variables numèriques amb *missings*, que són l'edat (AGE), l'hora (HOUR) i el minut (MINUTE) en el qual es va tenir l'accident, les codificacions per aquestes dades són 99, 998 o 999, depenent de cada cas.

	NA	Percentatge
DAY	0	0.00
HOUR	33	1.19
MINUTE	34	1.22
RUR_URB	0	0.00
DAY_WEEK	0	0.00
FATALS	0	0.00
DRUNK_DR	0	0.00
NO_PER	0	0.00
MORTS	0	0.00
NO_VEHICLE	0	0.00
HIHAMORTS	0	0.00

Taula 9.6: Taula accident

	NA	Percentatge
DAY	0	0.00
HOUR	55	0.78
MINUTE	58	0.82
RUR_URB	0	0.00
DAY_WEEK	0	0.00
FATALS	0	0.00
DRUNK_DR	0	0.00
NO_PER	0	0.00
MORTS	0	0.00
NO_VEHICLE	0	0.00
NO_FUGITS	0	0.00
AGE	222	3.13
SEX	0	0.00
PER_TYP	0	0.00
DOA	0	0.00

Taula 9.7: Taula person

Un cop transformades aquestes dades, es poden visualitzar els *missings* per cada variable, tant en terme absolut com relatiu. A les taules 9.6 i 9.7 s'hi poden trobar les variables de les taules d'accidents i de persones, respectivament, juntament amb el nombre de dades mancants que presenten, i el tant per cent que aquestes suposen al total de la informació de la variable.

Tal i com es pot observar, a la taula d'accidents s'hi troben *missings* per a les variables HOUR i MINUTES, mentre que per la taula de persones, hi ha *missings* per a les variables HOUR, MINUTES i AGE. En ambdós casos, totes les variables són numèriques i, per aquest motiu, es pot usar l'algoritme KNN per a la imputació de valors a les dades mancants.

K-nearest neighbors (KNN) és un tipus d'algoritme d'aprenentatge supervisat que s'utilitza tant per a la regressió com per a la classificació. La seva funció és intentar predir la classe correcta per a unes dades de prova (que, en el nostre cas, seran les variables que presenten dades mancants) en base a la seva similitud amb altres mostres de dades conegudes (en el nostre cas, les variables completes). Tot això es fa assumint que les dades amb trets similars es troben juntes, i utilitza mesures de distància en el seu nucli.

Un cop s'ha aplicat l'algoritme per a les variables corresponents, es pot veure, a continuació, com cap de les dues taules presenta cap *missing* a les variables conflictives.

	NA	Percentatge
HOUR	0	0.00
MINUTE	0	0.00

Taula 9.8: Taula Accident

	NA	Percentatge
HOUR	0	0.00
MINUTE	0	0.00
AGE	0	0.00

Taula 9.9: Taula Person

Recordem que les taules 9.8 i 9.9 i mostren les bases de dades d'accidents i de les persones implicades en els accidents, respectivament.

9.2.2 Dades atípiques

Pel que fa a les dades atípiques (*outliers*), en destaca el nombre de persones implicades a l'accident. Més específicament, hi ha un cas en que 53 persones estan involucrades en un accident. A priori, res ens fa pensar que aquesta dada, tot i ser atípica, no pugui ser certa. Per exemple, un accident pot tenir un nombre elevat de persones involucrades si involucra un autobús com a vehicle. Això no obstant, a l'hora de la segmentació les dades es podrien veure afectades per aquest valor, ja que alguns algoritmes són molt sensibles als *outliers*.

Al següent gràfic (figura 9.1) es representa la variable nombre de persones (NO_PER), on es poden identificar de forma clara aquests valors atípics. També es pot trobar un resum numèric de la variable a la taula 9.10, que indica el valor màxim que pren la variable.

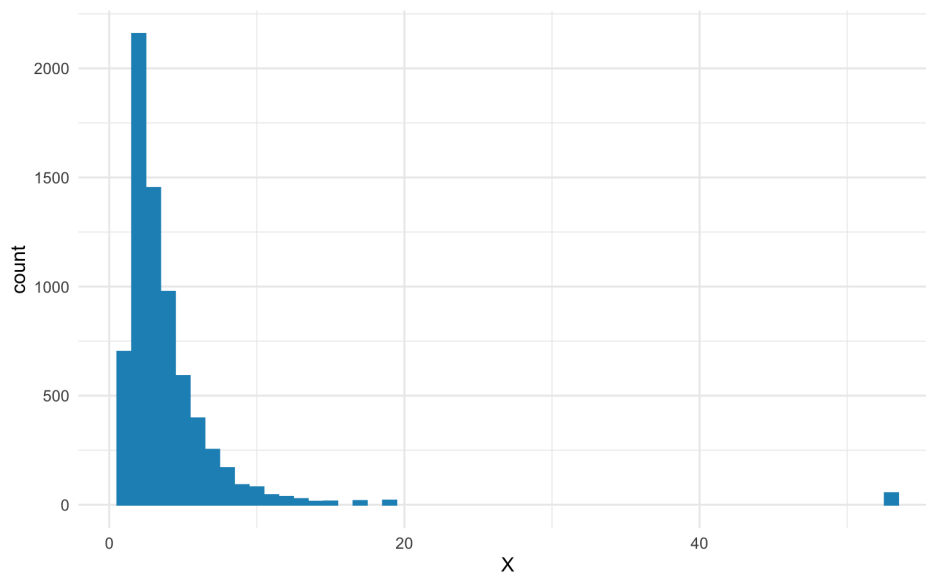


Figura 9.1: Histograma de la variable Nombre de persones

N.Valid	Min	Q1	Mediana	Mitjana	Desv. estàndar	Q3	Max	IQR
7087	1	2	3	4.015098	4.938707	5	53	3

Taula 9.10: Resum numèric de la variable Nombre de persones

Per tal d'assegurar-nos que aquesta dada no afecta al nostre anàlisi, i tenint en compte que disposem d'una base de dades molt gran, treurem aquests casos d'ambdues taules.

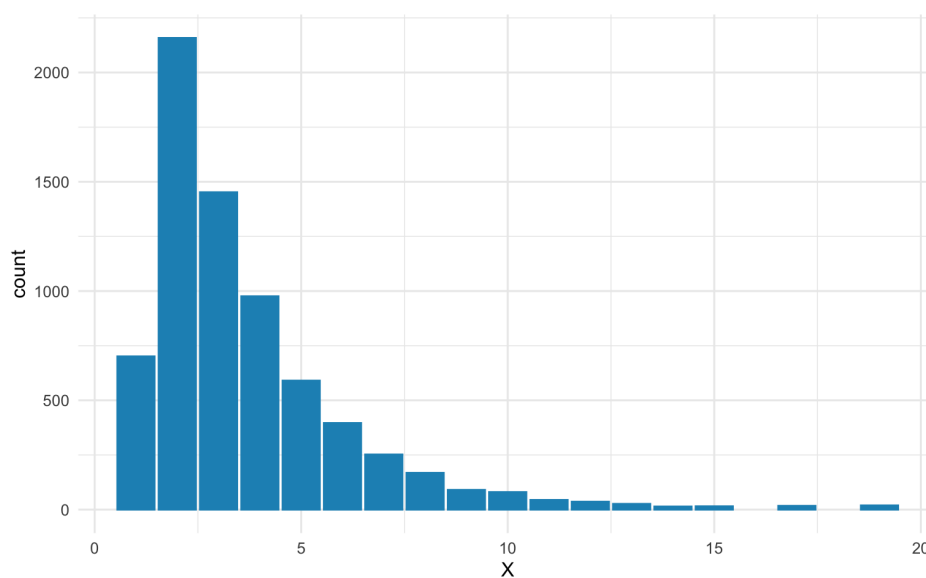


Figura 9.2: Histograma de la variable Nombre de persones després d'eliminar els outliers

N.Valid	Min	Q1	Mediana	Mitjana	Desv. estàndar	Q3	Max	IQR
7034	1	2	3	3.64005	2.521077	4	19	2

Taula 9.11: Resum numèric de la variable Nombre de persones

9.2.3 Categoritzar

En el cas de les dades mancants que es troben a variables categòriques, el que es farà serà factoritzar-les i, seguidament, definir els nivells que presenta el factor. Així, per exemple, la variable PER_TYP presenta 8 nivells que s’han d’agrupar en 3 (Conductor, Ocupant i Altres).

A continuació es mostren els canvis realitzats a algunes de les variables categòriques de la base de dades:

PER_TYP: Tipus de persona (1 = conductor, 2 = ocupant, resta de codis = altres).

- Abans: 1, 2, 3, 4, 5, 6, 8, 9
- Després: Conductor, Ocupant, Altres

DAY_WEEK: Dia de la setmana (1 = Diumenge, 2 = Dilluns, . . . , 7 = Dissabte).

- Abans: 1, 2, 3, 4, 5, 6, 7
- Després: Diumenge, Dilluns, Dimarts, Dimecres, Dijous, Divendres, Dissabte

SEX: Sexe de la persona (1 = home, 2 = dona, 8 = No registrat, 9 = Desconegut).

- Abans: 1, 2, 8, 9
- Després: Home, Dona

Per la variable variable SEX hi ha les categories anomenades “No registrat” i “Desconegut”, que representen aquelles persones de les quals no tenim informació sobre el seu sexe. Com aquestes categories no ens aporten informació d’utilitat a l’hora de realitzar l’estudi ni per a realitzar models predictius, prescindirem dels individus que corresponguin a aquestes per a realitzar el nostre anàlisi.

RUR_URB: Informació sobre la localització (1 = Rural, 2 = Urbà, 6 = Via no classificada, 8 = No registrat, 9 = Desconegut).

- Abans: 1, 2, 6, 8, 9

- Després: Rural, Urbà, Desconegut

HI HA MORTS: Variable identificadora dels accidents mortals (0: no hi ha morts en l'accident, 1: hi ha morts en l'accident).

- Abans: 0, 1
- Després: No, Sí

DOA: Tipus de víctima (0 = sobreviu, 7 = mort a l'accident, 8 = mort al trasllat, 9 = Desconegut)

- Abans: 0, 7, 8, 9
- Després: Sobreviu, Mor

Per aquesta última variable, DOA, hi ha una categoria anomenada "Desconegut", que representa aquelles persones que no se sap si sobrevien a l'accident o no. Ja que en aquest estudi el fet de sobreviure o no a l'accident és de gran interès, i aquesta categoria no ens aporta informació útil, prescindirem dels individus enmarcats en aquesta categoria per a realitzar el nostre anàlisi. A més a més, agruparem les categories 7 i 8 (mort a l'accident i mort al trasllat, respectivament) en una única.

Per últim, es crearan dues variables noves a partir de HOURS i DAY, que ja són presents a ambdós taules. Aquest pas es realitza perquè les variables HOURS i DAY tenen un rang de valors molt elevat que aporta poca informació.

HOURS_agrupat

En el cas de la variable HOURS, es tindrà en compte que, comunament, es considera que el dia està format per 5 intervals de temps segons la posició del sol. Aquest són la matinada (de les 0 a les 5 h incloses), el matí (de les 6 a les 11 h incloses), el migdia (de les 12 a les 14 h incloses), la tarda (de les 15 a les 19 h incloses) i la nit (de les 20 a les 23 h incloses). S'han fet servir aquests intervals per definir la nova variable HOURS_agrupat.

- Abans: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23
- Després: Matinada, Matí, Migdia, Tarda, Nit

SETMANA

Pel que fa a la variable SETMANA, s'han definit les setmanes del mes en que es va realitzar el seguiment que presenten les dades. S'ha considerat el primer dia de la setmana el dilluns i l'últim el diumenge, tenint en compte que el dia 1 del mes era un dimarts. Per aquest motiu les setmanes 1 i 5 són les més curtes, especialment la cinquena, ja que el dia 31 va caure en dijous.

- Abans: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31
- Després: Setmana 1, Setmana 2, Setmana 3, Setmana 4, Setmana 5

9.2.4 Variable resposta

Per últim, definirem les variables resposta per a cada taula, és a dir, aquelles característiques que ens interessa poder predir tant en els futurs accidents com en les pròximes persones que es vegin involucrades en aquests.

Per una banda, és d'interès classificar els accidents segons si aquests han ocasionat morts o bé no ha sigut el cas. D'aquesta manera, es podria crear un model de predicció que permeti establir si un accident serà mortal o no en el futur en funció de les característiques que presenti.

Per tant, la variable d'interès és HIHAMORTS, que es mostra a la figura 9.3.

Variable	Categories	Freqüències	Missings
Hi ha morts	1. No	1176 (42.3%)	0
[factor]	2. Sí	1604 (57.7%)	(0.00%)

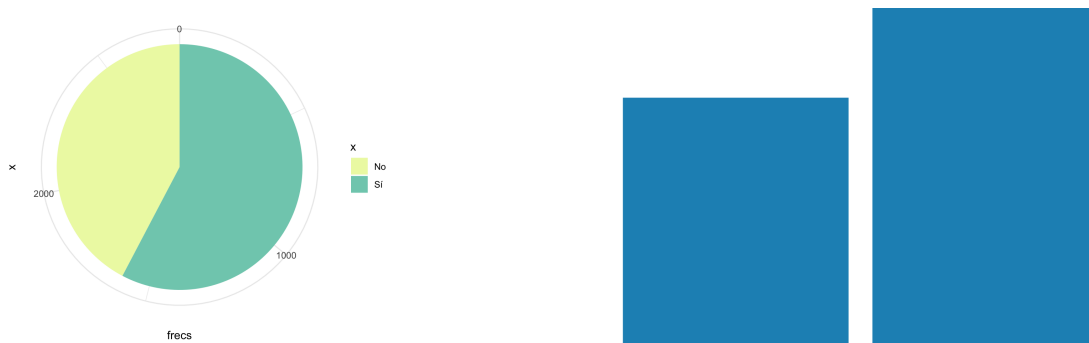


Figura 9.3: Anàlisi descriptiu de la variable Hi ha morts

El tant per cent d'accidents amb morts és molt elevat perquè tots són accidents amb víctimes que es van enregistrar per l'autoritat i, per tant, devia anar-hi un agent de trànsit al lloc on es va produir. Els accidents més lleus o superficials no van ser enregistrats per aquest estudi.

Seguint aquesta línia, serà també de gran importància el tipus de víctima que esdevindran cadascuna de les persones implicades en un accident. En aquest cas, la variable d'interès serà DOA, de la qual es pot trobar un breu anàlisi descriptiu a la figura 9.4.

Variable	Categories	Freqüències	Missings
Tipus de víctima	1. Sobreviu	5113 (74.1%)	0
[factor]	2. Mor	1791 (25.9%)	(0.00%)



Figura 9.4: Anàlisi descriptiu de la variable Tipus de víctima

9.3 Anàlisi descriptiva

9.3.1 Anàlisi univariant

Variables numèriques

Variables vinculades als accidents

Variable	Tipus	Descripció
HOUR	Numèrica	Hora de l'accident (99 = desconeguda)
MINUTE	Numèrica	Minut de l'accident (99 = desconegut)
FATALS	Numèrica	Nombre de víctimes a l'accident (poden ser ferits o morts)
DRUNK_DR	Numèrica	Nombre de conductors beguts involucrats a l'accident
MORTS	Numèrica	Nombre de morts en l'accident
NO_PER	Numèrica	Nombre de persones implicades en l'accident
NO_VEHICLE	Numèrica	Nombre de vehicles implicats en l'accident

Taula 9.12: Variables numèriques vinculades als accidents

Per a comprovar que les dades són correctes i que s'han tractat bé en termes de *missings* i *outliers*, farem ús del resum numèric de cada variable que es mostra a la taula 9.13.

Les variables HOUR i MINUTE prenen valors entre 0 i 23 i entre 0 i 59 respectivament, els quals són rangs esperats i no donen peu a cap dada atípica en la seva distribució. A la tercera fila de la taula es troba el nombre de víctimes a l'accident, FATALS, que assenyala que no hi ha cap accident en què no hi hagi, com a mínim, una víctima. També veiem com el nombre màxim de víctimes és 5, però en termes generals i d'acord amb els quartils 1, 2 i 3 i la mitjana s'observa que la gran majoria dels accidents (el 75%) només en presenten una. Així mateix, pel que fa al nombre de conductors beguts involucrats a l'accident, DRUNK_DR, aquesta variable pren el valor 0 en la majoria dels casos (el 75% com a mínim, com indica el tercer quartil). Això no obstant, hi ha casos en què hi ha fins a 2 conductors beguts involucrats en un mateix accident.

Variable	N.Valid	Min	Q1	Q2	Mitjana	Desv.	Q3	Max	IQR
HOUR	2780	0	7	15	13.0122	6.7949	18	23	11
MINUTE	2780	0	13	28	28.0417	17.2255	43	59	30
FATALS	2780	1	1	1	1.1000	0.3833	1	5	0
DRUNK_DR	2780	0	0	0	0.2457	0.4429	0	2	0
MORTS	2780	0	0	1	0,6442	0.6346	1	5	1
NO_PER	2780	1	1	2	2.5302	1.6805	3	19	2
NO_VEHICLE	2780	1	1	1	1.5079	0.6998	2	6	1

Taula 9.13: Resum de les variables numèriques vinculades als accidents

Per la variable NO_PER, s'observa que el mínim de persones que es trobaven als vehicles en el moment de l'accident era un, cosa que té sentit, ja que no podria haver-hi cap accident de trànsit si no hi hagués, com a mínim, un conductor. La mitjana de gent implicada en cada accident va de 2 a 3 persones, i l'accident en el qual va haver-hi més participants van arribar a ser fins a 19 persones. Aquest tipus d'accident és creïble, ja que quatre cotxes poden arribar a col·lidir en una carretera si hi ha molt poc temps de reacció i, si aquests cotxes van plens, es podria arribar a parlar de vint persones en un accident, comptant que no hi hagués cap vianant.

Pel que fa a MORTS, es pot veure com en el 25% dels casos, com a mínim, no hi ha cap víctima mortal en aquests accidents. La mitjana no arriba a ser un, cosa que indica que la majoria dels accidents no presenten cap mort, i en el 75% dels casos el màxim de víctimes és d'un. Així i tot, hi ha un cas en el qual hi va haver 5 morts, que destaca per sobre de la resta.

Per acabar, en tots els accidents de trànsit hi ha un mínim d'un vehicle, i així ho mostra la variable NO_VEHICLE. L'accident amb més vehicles implicats en té 6.

Quant als *missings*, la primera columna de la taula és un recompte dels casos vàlids per a cada variable, és a dir, aquells que no presenten valors mancants per a aquella variable concreta. Es pot comprovar com el mètode d'imputació s'ha realitzat correctament, perquè totes les variables presenten el mateix recompte de casos vàlids, que coincideix amb el nombre d'accidents total a la taula.

Variables vinculades a les persones

Sobre les variables vinculades a les persones (Taula 9.14), només varien les variables NO_FUGITS i EDAT respecte a la taula anterior. De fet, la distribució de les variables HOUR, MINUTE, FATALS, DRUNK_DR, NO_PER, MORTS i NO_VEHICLE no presenta canvis rellevants respecte a la distribució que es troba a la taula anterior. No es troba cap cas de dada atípica ni cap dada mancanted per aquestes variables. Per aquest motiu, ens centrarem ara en els resums de les variables NO_FUGITS i EDAT, que es troben a la taula 9.15.

Variable	Tipus	Descripció
HOUR	Numèrica	Hora de l'accident (99 = desconeguda)
MINUTE	Numèrica	Minut de l'accident (99 = desconegut)
FATALS	Numèrica	Nombre de víctimes a l'accident (poden ser ferits o morts)
DRUNK_DR	Numèrica	Nombre de conductors beguts involucrats a l'accident
MORTS	Numèrica	Nombre de morts en l'accident
NO_PER	Numèrica	Nombre de persones implicades en l'accident
NO_VEHICLE	Numèrica	Nombre de vehicles implicats en l'accident
NO_FUGITS	Numèrica	Nombre de vehicles fugits implicats en l'accident
AGE	Numèrica	Edat de la persona

Taula 9.14: Variables numèriques vinculades a les persones

Variable	N.Valid	Min	Q1	Q2	Mitjana	Desv.	Q3	Max	IQR
HOUR	6904	0	8.0	15	13.4422	6.5239	18	23	10.00
MINUTE	6904	0	13.0	28	27.8872	17.2487	43	59	30.00
FATALS	6904	1	1.0	1	1.1705	0.5296	1	5	0.00
DRUNK_DR	6904	0	0.0	0	0.2219	0.4346	0	2	0.00
MORTS	6904	0	0.0	1	0.6966	0.7340	1	5	1.00
NO_PER	6904	1	2.0	3	3.6548	2.5291	4	19	2.00
NO_VEHICLE	6904	1	1.0	2	1.7784	0.8766	2	6	1.00
NO_FUGITS	6904	0	0.0	0	0.0529	0.2478	0	3	0.00
AGE	6904	0	23.5	37	39.9005	20.4028	55	98	31.25

Taula 9.15: Resum de les variables numèriques vinculades a les persones

Pel que fa al nombre de vehicles fugits a l'accident, el tercer quartil indica que al 75% dels casos no hi ha cap vehicle fugit, i si en fixem en el valor màxim que pren aquesta variable, el cas amb més vehicles fugits en presenta 3. Finalment, sobre l'edat de les persones, el rang de valors va des de 0 fins a 98 anys, és a dir, en alguns accidents hi ha implicats nadons i, en altres, persones de mitjana edat.

Novament, ens trobem que la primera columna de la taula és un recompte dels casos vàlids per a cada variable, és a dir, aquells que no presenten valors mancants per a aquella variable concreta. Es pot comprovar com el mètode d'imputació s'ha realitzat correctament, perquè totes les variables presenten el mateix recompte de casos vàlids, que coincideix amb el nombre de persones total.

Variables categòriques

Variables vinculades als accidents

DAY: Dia de l'accident (de l'1 al 31). Tipus de variable: Factor

Variable	Categoria	Freqüència	Missings
DAY [factor]	1. 1	71 (2.6%)	0 (0.0%)
	2. 2	84 (3.0%)	
	3. 3	108 (3.9%)	
	4. 4	96 (3.5%)	
	5. 5	117 (4.2%)	
	6. 6	112 (4.0%)	
	7. 7	78 (2.8%)	
	8. 8	81 (2.9%)	
	9. 9	88 (3.2%)	
	10. 10	85 (3.1%)	
	[21 altres]	1859 (66.9%)	

La freqüència d'accidents segons el dia del mes sembla prou estable, i no es troba cap patró significatiu.

RUR_URB: Informació sobre la localització (1 = Rural, 2 = Urbà, 6 = Via no classificada, 8 = No registrat, 9 = Desconegut). Tipus de variable: Factor

Variable	Categoria	Freqüència	Missings
RUR_URB [factor]	1. Rural	1174 (42.2%)	0
	2. Urbà	1288 (46.3%)	(0.0%)
	3. Desconegut	318 (11.4%)	

En 318 casos no hi ha informació sobre la localització de l'accident, però per la resta dels casos, el 88,6%, es pot identificar si va ser en una zona rural o urbana, i els casos estan prou balancejats en aquestes dues categories.

DAY_WEEK: Dia de la setmana (1 = Diumenge, 2 = Dilluns, . . . , 7 = Dissabte). Tipus de variable: Factor

Variable	Categoria	Freqüència	Missings
DAY_WEEK [factor]	1. Diumenge	356 (12.8%)	0
	2. Dilluns	311 (11.2%)	(0.0%)
	3. Dimarts	410 (14.7%)	
	4. Dimecres	436 (15.7%)	
	5. Dijous	460 (16.5%)	
	6. Divendres	387 (13.9%)	
	7. Dissabte	420 (15.1%)	

El mes de desembre de 2015 es van donar la majoria d'accidents de trànsit els dijous, mentre que el dia de la setmana on es van donar menys va ser el dilluns.

HIHAMORTS: Tipus de variable: Factor

Variable	Categoria	Freqüència	Missings
HIHAMORTS	1. No	1176 (42.3%)	0
[factor]	2. Sí	1604 (57.7%)	(0.0%)

Aquesta és la variable resposta de la taula, és a dir, aquella que ens interessa predir a partir de la relació amb la resta de variables de la taula. Pren dues categories (sí i no) i, pel que es pot veure, està una mica desbalancejada. Això vol dir que trobem més accidents a la taula en què hi ha hagut víctimes mortals que en què no s'hagin donat. La característica del balanceig, si no es compleix, pot ser un inconvenient a l'hora de crear el model, ja que pot donar més pes a uns casos per sobre dels altres. Tanmateix, com la diferència és petita, s'ha decidit treballar amb les dades que tenim, sense aplicar cap tècnica de balanceig que podria afectar a aquestes.

Variables vinculades a les persones

SEX: Sexe de la persona (1 = home, 2 = dona). Tipus de variable: Factor

Variable	Categoria	Freqüència	Missings
SEX	1. Home	4569 (66.2%)	0
[factor]	2. Dona	2335 (33.8%)	(0.0%)

Per aquesta variable s'han eliminat els individus amb el sexe no registrat o desconegut, perquè són categories de la variable que no aporten informació extra sobre l'individu. Cal destacar, a més a més, que la majoria de les persones d'aquesta taula són homes (66,2%).

PER_TYP: Tipus de persona (1 = conductor, 2 = ocupant, resta de codis = altres). Tipus de variable: Factor

Variable	Categoria	Freqüència	Missings
PER_TYP	1. Conductor	4057 (58.8%)	0
[factor]	2. Ocupant	2096 (30.4%)	(0.0%)
	3. Altres	751 (10.9%)	

Quant al tipus de persona, se'n desconeix la informació d'aproximadament l'11% dels individus, mentre que gairebé el 60% de les persones involucrades als accidents eren conductors. Per tant, d'aquí es pot extreure que en la majoria de vehicles que van patir un accident en aquest període de temps només viatjava una persona, el conductor o conductora del vehicle.

DOA: Tipus de víctima (0 = sobre viu, resta de codis = mor). Tipus de variable: Factor

Variable	Categoria	Freqüència	Missings
DOA	1. Sobre viu	5113 (74.1%)	0
[factor]	2. Mor	1791 (25.9%)	(0.0%)

Per acabar, pel tipus de víctima, que és la variable resposta en aquest cas, no hi ha balanceig entre les categories “Sobre viu” i “Mor”. Si es vol crear un model per intentar predir la classe d’aquesta variable, molt probablement s’hauria d’emprar alguna tècnica de balanceig sobre aquestes dades.

9.3.2 Anàlisi bivariant

Per acabar l’anàlisi descriptiva de les dades, s’estudiarà la relació que existeix entre diferents parells de variables. Aquest tipus d’anàlisi ajudarà a esbrinar si existeix una associació entre les variables i, en cas afirmatiu, quina és la força d’aquesta.

Per a la visualització d’aquestes relacions entre variables s’ha fet ús de la plataforma de Google, [Looker Studio](#). Aquesta permet convertir les dades en panells o informes complets, fàcils de llegir i de compartir, així com totalment personalitzables.

Amb Looker Studio es poden elaborar fàcilment informes sobre dades procedents d’una gran varietat de fonts, sense necessitat de programar. En tan sols uns instants, permet la connexió a grans conjunts de dades com els que es troben a BigQuery. Amb aquesta finalitat, s’han pujat les taules preprocessades altra vegada a BigQuery, i s’ha establert la connexió amb l’eina de visualització.

Variables vinculades als accidents

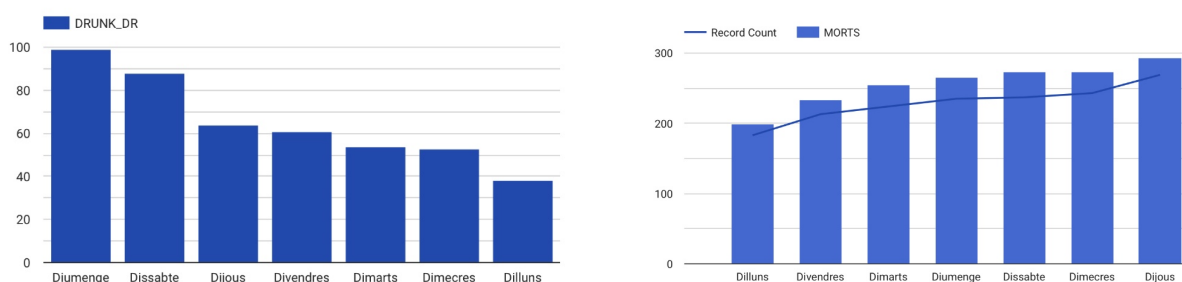


Figura 9.5: Nombre de conductors beguts i quantitat d’accidents, respectivament

Pel que fa al nombre de conductors beguts segons el dia de la setmana en el qual succeeix l’accident, s’observa de forma clara a la gràfica esquerra de la figura 9.5 com la majoria dels conductors beguts es concentren al cap de setmana. Sembla que podria ser un patró perquè

hi ha diferències notables entre la quantitat de conductors beguts a finals de la setmana, en comparació als dilluns, dimarts i dimecres. D'altra banda, si ens fixem, en aquest cas, en el nombre de morts segons el dia de la setmana, el dilluns es troba altra vegada en l'última posició, ja que és el dia en què es donen menys morts en accidents de trànsit. Paral·lelament, els últims dies de la setmana presenten un major nombre d'accidents mortals. Aquesta és la informació que presenta la gràfica dreta de la figura 9.5.

Si es té en compte la informació d'aquestes últimes gràfiques, es podria pensar que existeix una relació entre el nombre de conductors beguts i el nombre de ferits mortals als accidents de trànsit. S'haurà de tenir en compte aquesta hipòtesi per anàlisis posteriors de les dades.

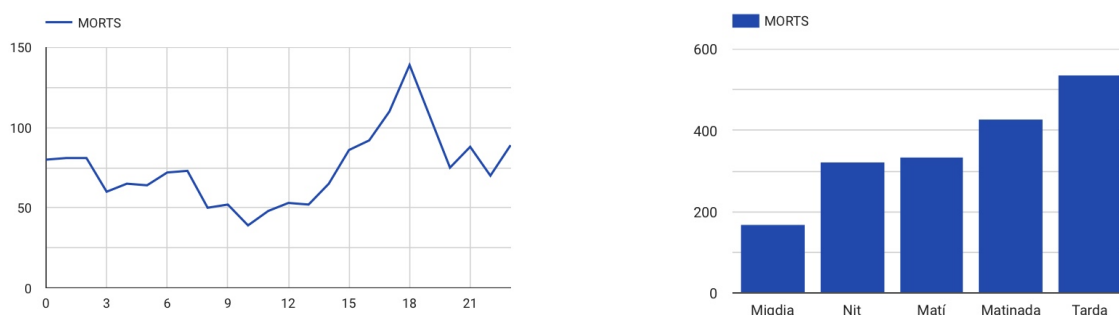


Figura 9.6: Hores sense agrupar i hores agrupades, respectivament

Si centrem l'atenció en les hores del dia, a la figura 9.6 es poden veure les freqüències absolutes quant a la quantitat de morts en els diferents moments del dia. A l'esquerra, es divideix el dia en les seves hores, i s'observa com l'hora en què es produeixen més accidents és a les 6 de la tarda. Perquè la gràfica sigui més informativa, s'han agrupat les hores segons els moments del dia per crear la gràfica de la dreta, amb la informació de la variables HOURS_agrupat.

En definitiva, de les dues gràfiques s'extreu que la majoria de les morts es produeixen a la tarda i a la matinada, mentre que el moment del dia on hi ha menys morts és el migdia.

Variables vinculades a les persones

Pel que fa a la informació que tenim registrada sobre les persones involucrades en els accidents, i en la línia de les anàlisis anteriors, la figura 9.7 presenta una gràfica que mostra com el nombre de vehicles fugits en els accidents de trànsit varia en funció del moment del dia en que es produeix l'accident, així com en funció de la setmana del mes.

A la figura 9.7 es troba, a l'esquerra, la quantitat de vehicles fugits en els diferents moments del dia, i aquests es concentren sobretot a la tarda (de 15 a 19 hores, incloses) i la matinada (de les 0 a les 5 hores, incloses). En canvi, pel que fa a les diferents setmanes del mes es veu com la majoria de cotxes fugits es troben a la primera i la quarta setmana, i la distribució no és uniforme durant totes les setmanes del mes, com s'esperaria si no hi hagués cap relació entre

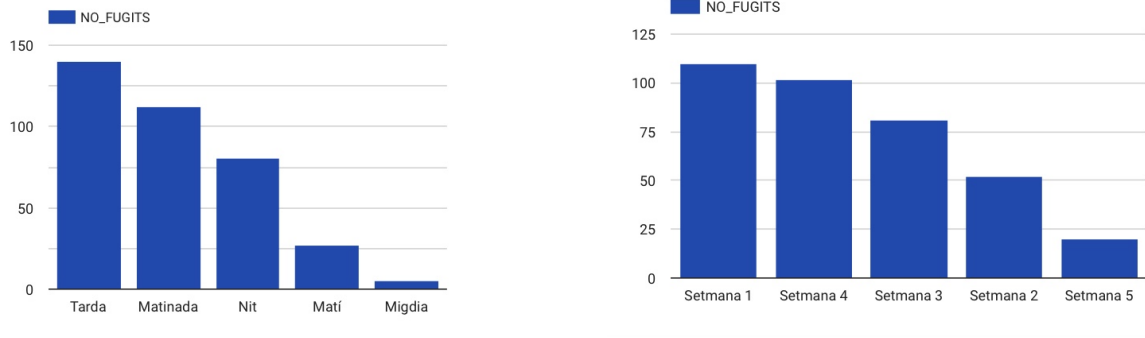


Figura 9.7: Nombre de vehicles fugits segons el dia de la setmana i segons el dia del mes, respectivament

ambdues variables.

En últim lloc, la figura 9.8 mostra una taula de contingència entre les variables SEX i DOA (és a dir, el sexe de la persona i el tipus de víctima, si sobreviu o no a l'accident). Aquesta taula marca com a categoria més abundant a la base de dades els homes que sobreviuen a l'accident, i, com a menys abundant, les dones que no sobreviuen al mateix.

SEX	DOA / Record Count	
	Sobreviu	Mor
Home	3.296	1.273
Dona	1.817	518

Figura 9.8: Quantitat de supervivents segons el sexe

Si es considera aquesta informació de forma relativa, aproximadament el 39% dels homes implicats en els accidents han resultat ferits mortalment, mentre que d'entre les dones ho ha sigut aproximadament el 29%. Tots els accidents enregistrats a la base de dades, ho són perquè en aquests hi ha hagut ferits o alguna complicació que ha portat als agents de trànsit a intervenir en aquest. Per tant, sembla haver-hi una relació en la qual els accidents que tenen algun ferit i després una víctima, aleshores és molt més freqüent que sigui un home qui mori.

9.4 Anàlisi per components principals

Per fer aquest anàlisi s'usarà la taula d'accidents depurada, que té un total de 7 variables numèriques. Per tant, l'anàlisi de components principals tindrà un total de 7 components.

Després de realitzar els càlculs corresponents, obtenim l'ACP de les variables numèriques a la figura 9.9.

```
## Standard deviations (1, ..., p=7):
## [1] 1.3664595 1.1889767 1.0075529 0.9935906 0.9379098 0.6780110 0.6142997
##
## Rotation (n x k) = (7 x 7):
##           PC1      PC2      PC3      PC4      PC5
## HOUR      0.019581482  0.36162096 -0.45475857  0.5291117  0.60251155
## MINUTE    -0.009586389 -0.03536516  0.85203068  0.4472944  0.26826732
## FATALS    0.526538509 -0.35531998 -0.12336459  0.2791675 -0.03396552
## DRUNK_DR  0.007444661 -0.41701761  0.02401409 -0.5214861  0.74096957
## NO_PER    0.539765106  0.38662738  0.08167484 -0.1537734  0.04925564
## MORTS     0.449980848 -0.49708983 -0.10881041  0.1996512 -0.10126310
## NO_VEHICLE 0.477909311  0.41381709  0.18147281 -0.3264650  0.04628810
##           PC6      PC7
## HOUR      -0.13779312  0.008670414
## MINUTE    -0.01677126 -0.019499405
## FATALS    0.50426055  0.497914912
## DRUNK_DR  0.06356915 -0.020963683
## NO_PER    0.40899951 -0.599290534
## MORTS     -0.58198865 -0.386980152
## NO_VEHICLE -0.46521033  0.492222591
```

Figura 9.9: Coeficients de les variables per les 7 components principals

Sabem que cada component representa una inèrcia concreta. Ho podem veure gràficament en el següent gràfic de barres (Figura 9.10).

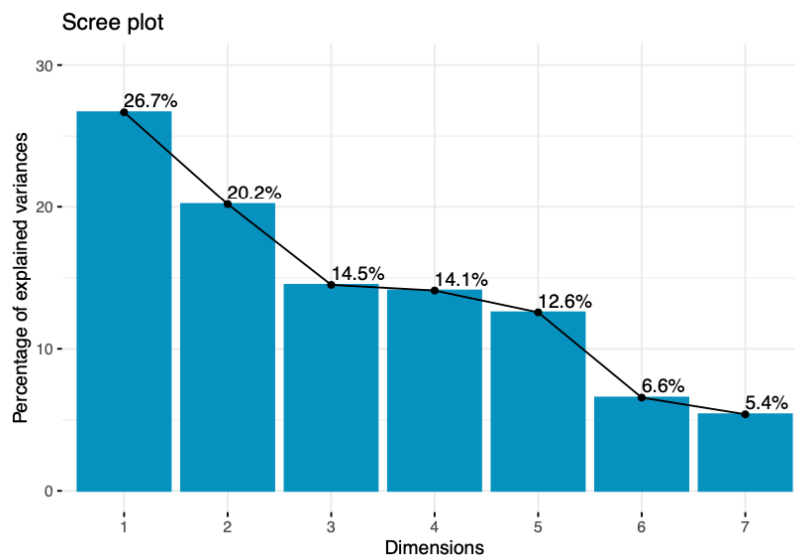


Figura 9.10: Barplot de la inèrcia de cada component

Tenint en compte que la inèrcia equival a la proporció de la variabilitat de les dades, sabem que amb un 80% d'inèrcia, podem obtenir gairebé tota la informació. Donant un cop d'ull a

la gràfica de la inèrcia, es pot veure que amb les quatre primeres components ja aconseguim gairebé el 80% de la inèrcia, així que ens podem servir d'aquestes per la nostra anàlisi.

A continuació, realitzem un gràfic de dispersió per a totes les combinacions possibles (Figura 9.11). Diferenciarem els individus, que en el nostre cas són accidents, depenent de si hi ha hagut víctimes mortals o no en aquest.

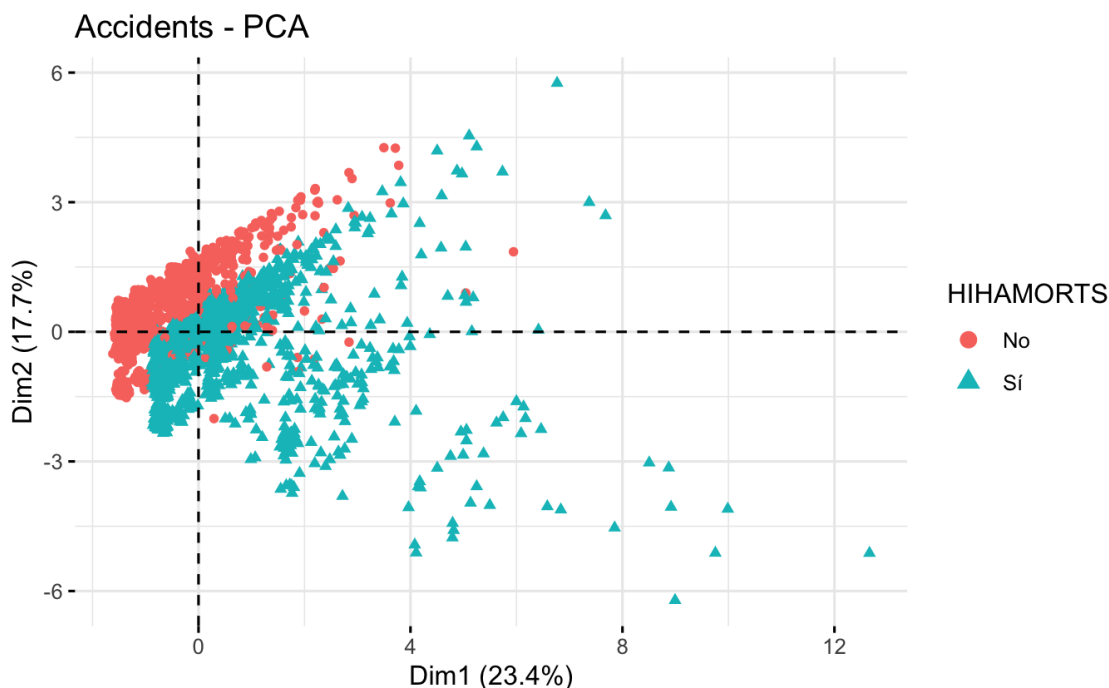


Figura 9.11: Gràfica de la projecció dels accidents

Es pot observar que les dues primeres components no aconseguen diferenciar de forma clara ambdós grups, i s'esdevenen solapaments. Això no obstant, es veu una lleugera tendència dels accidents sense víctimes mortals a situar-se a valors més elevats de la segona component, i a més baixos de la primera component.

9.4.1 Projecció de variables numèriques

A la gràfica corresponent a la figura 9.12 es pot veure totes les variables numèriques representades en la primera i segona component.

Veiem que la majoria de variables estan representades sobre l'eix horitzontal, que correspon a la primera component. Aquestes variables són NO_VEHICLE, MORTS, FATALS i NO_PERSONES. A més, aquestes dues últimes prenen un valor de gairebé 0,8 i, per tant, són les que expliquen amb més precisió la primera component. Ens fixem, també, que totes aquestes variables tenen relació amb el nombre de persones, de manera que a la primera component se li pot assignar

l'etiqueta de “Nombre de persones involucrades”.

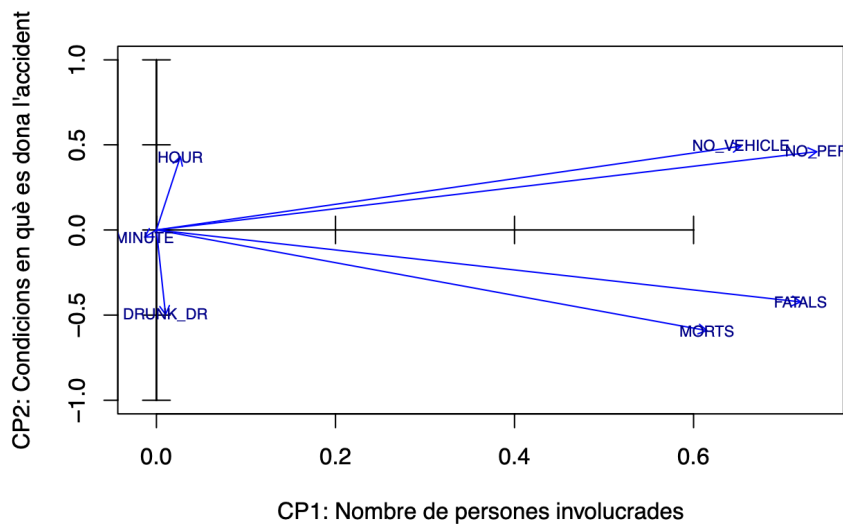


Figura 9.12: Gràfica de la projecció de variables numèriques

Pel que fa a l'eix vertical, només hi ha dues variables que estiguin una mica relacionades amb la segona component. Aquestes són HOUR i DRUNK_DR, que prenen un valor prop del 0,5. Com que a priori, aquestes dues variables no tenen gaire relació l'una amb l'altra, assignarem a la segona component l'etiqueta de “Condicions en què es dona l'accident”, ja que a simple vista cap de les dues destaca sobre l'altre en la seva aportació a la segona component.

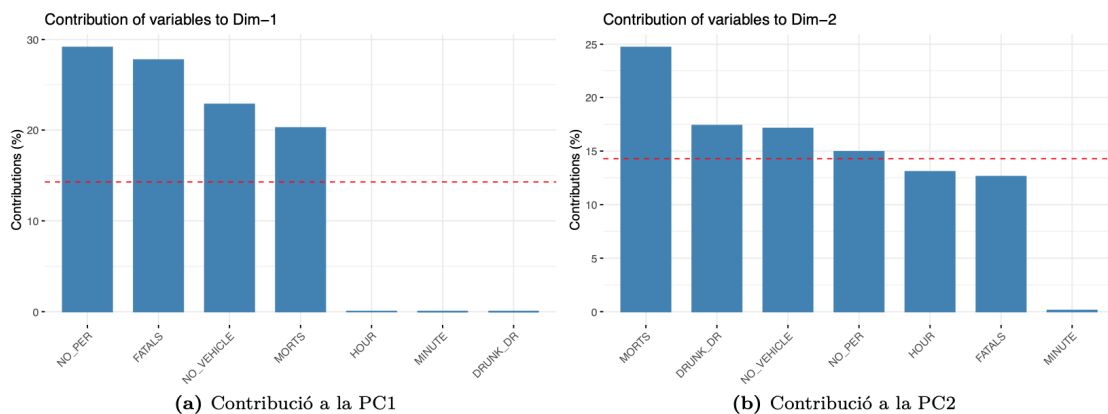


Figura 9.13: Gràfiques de contribució de les variables a les components principals

Als gràfics de la figura 9.13, la línia vermella discontinua indica el valor mitjà de contribució. Per una determinada component, una variable amb una contribució major a aquest límit pot considerar-se important a l'hora de contribuir a aquesta component. A l'esquerra, es pot veure que les variables NO_PER i FATALS són les que més contribueixen a la PC1, tal com s'havia

destacat prèviament. D'altra banda, a la gràfica de la dreta destaca la variable MORTS com la més contribuent a la PC2.

Pel que fa als vectors (variables) de la Figura 9.12, ens podem fixar, també en la seva longitud i en l'angle respecte als eixos de les components principals i entre ells mateixos.

L'angle que formen les variables HOUR i DRUNK_DR respecte l'eix vertical és petit, la qual cosa ens indica que, tot i no ser les variables més contribuents, estan estretament relacionades amb la creació d'aquesta component. Paral·lelament, es pot dir el mateix de les variables NO_PER i FATALS respecte a la PC1.

Una altra lectura que se li pot fer als angles entre vectors són les correlacions que mostren entre variables. D'una banda, el nombre de persones i el nombre de vehicles presenten una correlació positiva, ja que formen un angle proper a zero entre elles, així com les variables FATALS (nombre de víctimes) i MORTS. D'altra banda, però, també s'entreveu que les variables MORTS i HOUR estan incorrelacionades, és a dir, són independents l'una de l'altra, perquè formen un angle de noranta graus, així com les variables NO_PER i NO_VEHICLE respecte DRUNK_DR.

Quant a la longitud, com major sigui la llargària d'un vector relacionat amb una variable (en un rang normalitzat del 0 a l'1), major variabilitat d'aquesta variable està continguda en la representació de les dues components representades, és a dir, millor està representada la seva informació a la gràfica. En aquest cas, NO_VEHICLE, NO_PER, FATALS i MORTS són les variables millor representades per les dues primeres components.

9.4.2 Projecció de variables categòriques

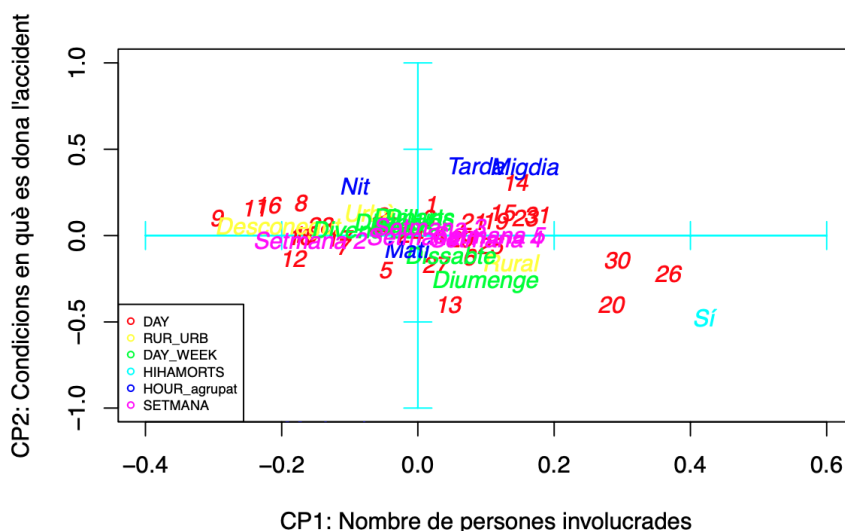


Figura 9.14: Projecció de les variables categòriques

També es pot fer una projecció de les variables categòriques sobre aquestes components, si en el mateix gràfic s'afegeixen totes les categories de les variables categòriques. El resultat es pot consultar a la figura 9.14.

Clarament aquest gràfic no es pot interpretar, ja que hi ha tantes categories que no es poden distingir. Per tant, el que es farà serà crear un gràfic per a cada variable categòrica, a veure si d'aquesta manera és més senzilla la seva interpretació.

Variable Dia

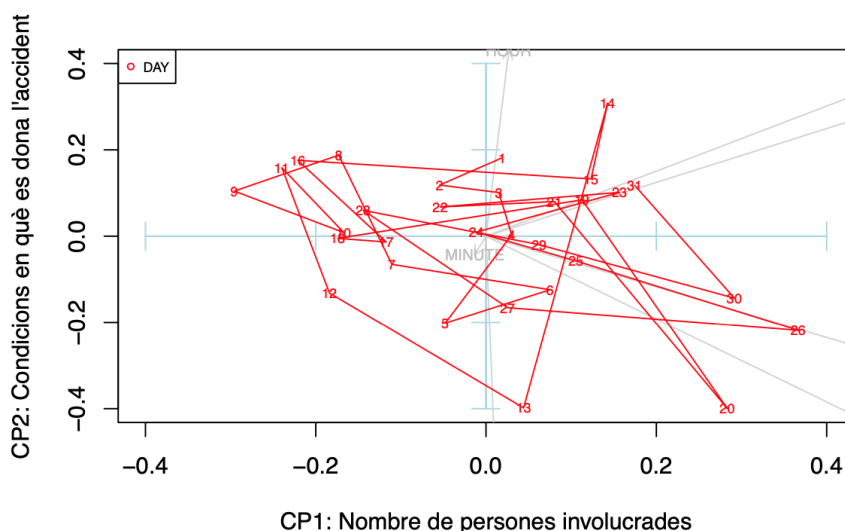


Figura 9.15: Gràfica de la projecció de la variable Dia

Amb la variable Dia veiem que les categories no formen cap patró rellevant respecte a les components. N'hi ha moltes, cosa que dificulta veure quines són les que expliquen millor una component o altra.

Variable Localització

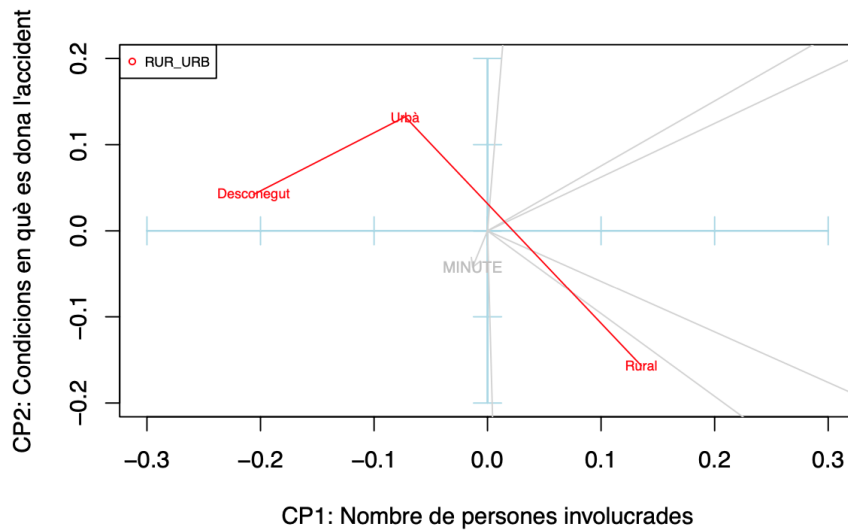


Figura 9.16: Gràfica de la projecció de la variable Localització

Per la variable Localització (RUR_URB) s’observa clarament que les categories “Urbà” i “Rural” expliquen la segona component, ja que són més properes a aquesta, mentre que la categoria “Desconegut” està més lligada a la primera component.

Variable Dia de la setmana

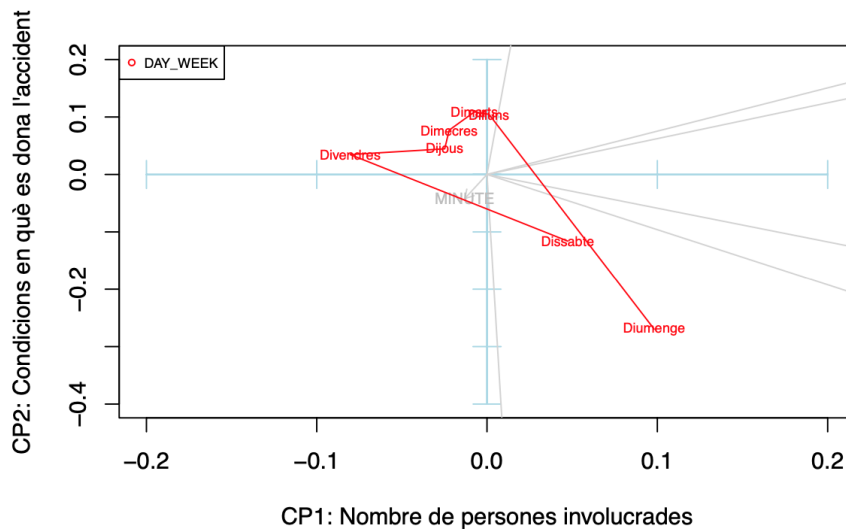


Figura 9.17: Gràfica de la projecció de la variable Dia de la setmana

Per la variable Dia de la setmana (DAY_WEEK) ens trobem amb que les categories que estan millor representades per les primeres components són divendres, dissabte i diumenge, ja que presenten els vectors amb major longitud. D’entre aquestes, divendres sembla més lligada

a la primera component, mentre que dissabte i diumenge contribueixen de forma semblant a ambdues components.

Variable HIHAMORTS

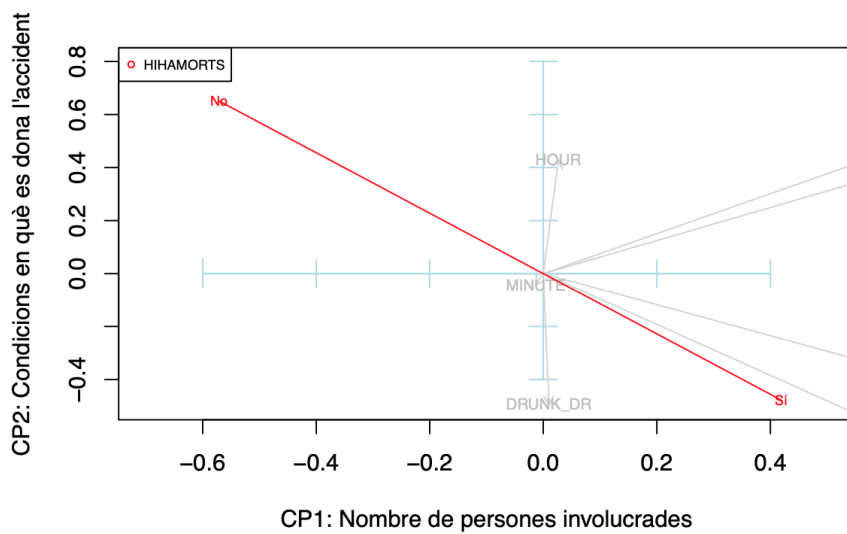


Figura 9.18: Gràfica de la projecció de la variable HIHAMORTS

Per la variable HIHAMORTS tenim dues categories que presenten una correlació negativa total, és a dir, si no pertany a una categoria, pertany a l'altra, evidentment. D'entre aquestes, la categoria "NO" queda més explicada per les primeres components, i ambdues categories contribueixen de forma igualitària en les dues components, ja que presenten aproximadament quaranta-cinc graus respecte els dos eixos.

Variable Hora grupada

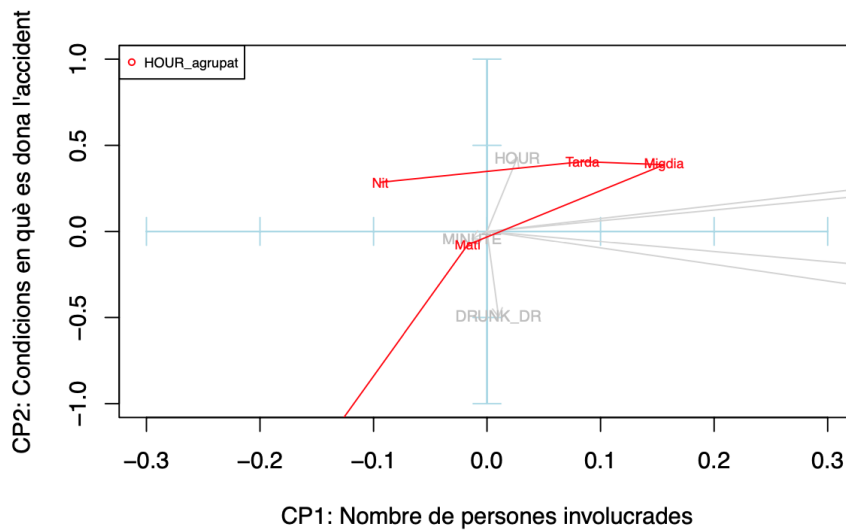


Figura 9.19: Gràfica de la projecció de la variable Hora agrupada

Per la variable hora agrupada (HOUR_agrupat) hi ha cinc categories. D'entre aquestes, els accidents que queden millor explicats per les primeres components són els que es donen a la matinada. Sembla que les categories “migdia” i “nit” estan més lligades a la primera component, mentre que “matinada” ho està a la segona. A més, la categoria que queda menys explicada és el matí.

Variable Setmana

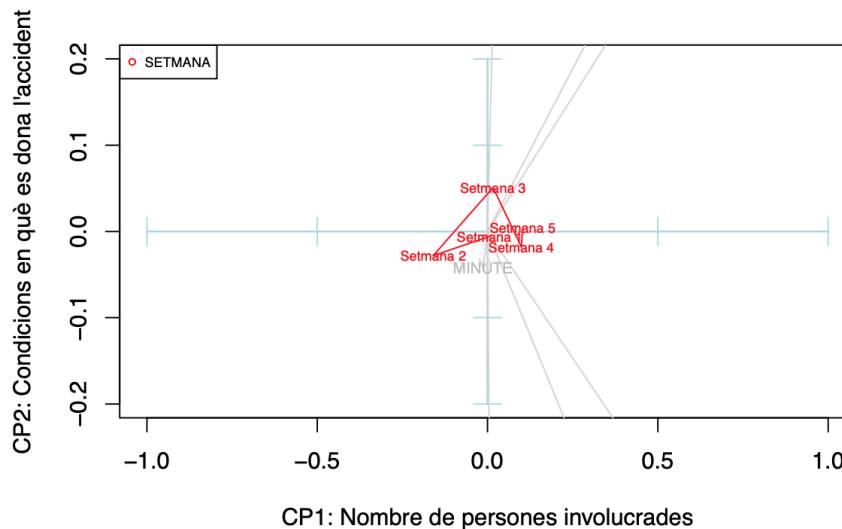


Figura 9.20: Gràfica de la projecció de la variable Setmana

Per la variable Setmana s'aprecia que cap de les categories està gaire explicada per les primeres

components, ja que els vectors són de molt poca longitud (tenen un mòdul proper a zero), i tampoc s'entreveu cap patró destacable.

10 Aprenentatge automàtic amb BigQuery ML

10.1 Regressió logística

BigQuery ML és una eina d'anàlisi que pertany a BigQuery i permet crear i executar models d'aprenentatge automàtic mitjançant consultes SQL estàndard. Com forma part de la plataforma de Google Cloud, permet continuar treballant amb les dades des del núvol, sense haver d'exportar-les localment. Aquest és un gran avantatge, ja que en funció del volum de dades que s'estiguin tractant, el fet de moure-les pot ser molt costós a escala computacional, així com restrictiu per temes d'espai. Alguns dels models compatibles amb BigQuery ML són la regressió lineal, la regressió logística (binària i multinomial), sèries temporals, arbres de decisió i xarxes neuronals profundes, entre d'altres.

Les dades que s'estan analitzant tenen com a variable resposta `HIHAMORTS`, que és una variable binària que pren el valor 1 en el cas d'haver-hi alguna víctima mortal a l'accident, i 0 altrament. En ser la variable d'interès binària, s'haurà de crear un model de classificació que ens permeti determinar si un accident tindrà víctimes mortals o no.

La regressió logística és un enfocament estadístic que s'utilitza pels problemes de classificació, ja que permet modelar la probabilitat de què es doni una de les categories de la variable resposta (classe) en funció d'un conjunt de variables predictores.

10.1.1 Entrenament del model

Abans de crear el model a la consola de BigQuery, dividirem la nostra base de dades en dues, les dades d'entrenament, o *train*, que representaran aproximadament el 80% del total, i les dades de prova, o *test*, que en seran el 20%. El conjunt *train* s'utilitzarà per entrenar el model, mentre que *test* s'utilitzarà un cop creat el model, i servirà per comprovar el bon funcionament d'aquest. Aquesta divisió de dades sovint és emprada per evitar el sobre ajust del model, que és un problema que es dona quan un model s'ajusta massa bé a les dades d'entrenament i no aconsegueix ajustar-se de forma fiable a dades addicionals.

	Hi ha morts	
	0 (No)	1 (Sí)
Dades totals	1604	1176
Train	1342	967
Test	262	209

Tal com s'observa a la taula, les dades estan aproximadament balancejades en els tres subconjunts. En cadascun, els accidents on hi ha víctimes mortals representen entre el 56 i el 58% dels

casos. Aquesta divisió s'ha pogut efectuar directament a la consola, a partir de les dades preprocessades, creant una variables fictícia perquè actui com a clau primària i identifiqui cada accident de forma única.

Per crear el model logístic a BigQuery ML, s'utilitzarà la següent sintaxi:

```
CREATE OR REPLACE MODEL [nom_projecte].[nom_base_de_dades].[nom_model]
OPTIONS(input_label_cols=['[variable_resposta]'], model_type='logistic_reg')
AS
SELECT * EXCEPT ([variables_no_utilitzades])
FROM '[nom_projecte].[nom_base_de_dades].[nom_taula]';
```

També cal destacar que s'ha usat totes les variables explicatives excepte les variables MORTS, HOUR i SETMANA. S'ha pres aquesta decisió per l'alta correlació que tenen aquestes variables amb d'altres de la mateixa taula.

Les variables predictores del model creat són les següents:

<u>Variables numèriques</u>	<u>Variables categòriques</u>
MINUTE	RUR_URB
FATALS	DAY_WEEK
DRUNK_DR	HOUR_AGRUPAT
NO_PER	
NO_VEHICLE	

En el cas de les variables categòriques, s'haurà d'especificar a les opcions del model que la seva codificació es faci mitjançant el mètode *dummy*. Això crearà columnes noves a la matriu de dades, tantes com possibles valors prengui la variable categòrica menys un, que serà la categoria base del model. Aquesta especificació s'indica amb la comanda següent: `CATEGORY_ENCODING_METHOD = 'DUMMY_ENCODING'`.

Seguidament, ajustem el model a les dades d'entrenament amb una nova consulta.

```
SELECT *
FROM ML.EVALUATE(MODEL '[nom_projecte].[nom_base_de_dades].[nom_model]',
(
  SELECT *
  FROM '[nom_projecte].[nom_base_de_dades].train'));
```

10.1.2 Avaluació del model

Aquest model, un cop entrenat, retorna un seguit de mètriques que fan referència a la funció de pèrdua, la duració de cada iteració i la taxa d'aprenentatge que presenta el model a cada iteració.



Figura 10.1: Entrenament del model

Per entendre com ha sigut l'entrenament d'aquest model, donarem un cop d'ull als resultats obtinguts (Figura 10.1).

La funció de pèrdua és un mètode per avaluar que tan bé un algoritme modela o s'ajusta a les dades d'entrenament. Si les prediccions es desvien massa dels resultats reals, la funció de pèrdua pren valors elevats. En canvi, si té l'ajuda d'una funció d'optimització, la funció de pèrdua aprèn a reduir l'error de predicció a cada iteració. En aquest sentit, en el nostre model no es redueix prou l'error de predicció, ja que es comença amb un valor de 0,684 abans de començar les iteracions, i s'acaba amb 0,636 a la cinquena iteració.

La següent mètrica que es troba és la duració de les iteracions. Aquest és un procés àgil en el qual la iteració que ha ocupat més temps ha sigut de 1,78 segons, aproximadament. Això, a escala d'usuari és molt atractiu, ja que permet el modelatge ràpid del model en conjunts de dades grans, cosa que normalment pot ser una limitació.

En últim lloc, es troba la taxa d'aprenentatge o *learning rate*. Aquest valor es calcula mitjançant l'algoritme de descens del gradient, i afecta la velocitat en què aquest algoritme convergeix en les ponderacions òptimes pels paràmetres. Una taxa d'aprenentatge massa elevada podria impedir que les ponderacions arribin a ser les òptimes, mentre que una taxa petita podria requerir moltes iteracions per arribar a calcular-les.

Seguidament, es pot avaluar el rendiment del model a partir de les dades de prova (*test*). Es fa servir una consulta com l'anterior, però aquest cop dirigida a aquesta altra taula de dades. En aquest cas el model retorna un seguit de mètriques i uns gràfics que permeten l'avaluació del model amb les noves dades.

Una de les primeres mètriques que trobem és la precisió del model, que és el tant per cent de les dades que s'han classificat correctament. Aquesta mètrica pren valors entre 0 i 1, i en el nostre cas és de 0,6336, és a dir, el model ha predit pel 63,36% de les dades la classe correcta de la variable resposta.

La mesura de recuperació fa referència al tant per cent de les dades classificades correctament d'entre les que la resposta correcta era positiva. Aquesta mètrica pren el valor de 0,8178, i significa que d'entre els casos positius (accidents amb víctimes mortals), el 81,78% s'han classificat efectivament com a positius. Al contrari, l'exactitud (o *accuracy*) quantifica el nombre d'elements classificats correctament d'entre aquells que s'han classificat com a positius i, en aquest cas, el valor baixa a 0,6228.

D'altra banda, la puntuació F1 es calcula a través de la mitjana harmònica entre les mètriques de precisió i recuperació. S'utilitza la mitjana harmònica en comptes de l'aritmètica perquè aquesta té la propietat de comportar-se com la mitjana aritmètica si els valors són molt semblants, però s'apropa més cap al valor petit quan hi ha diferències entre les dues mesures.

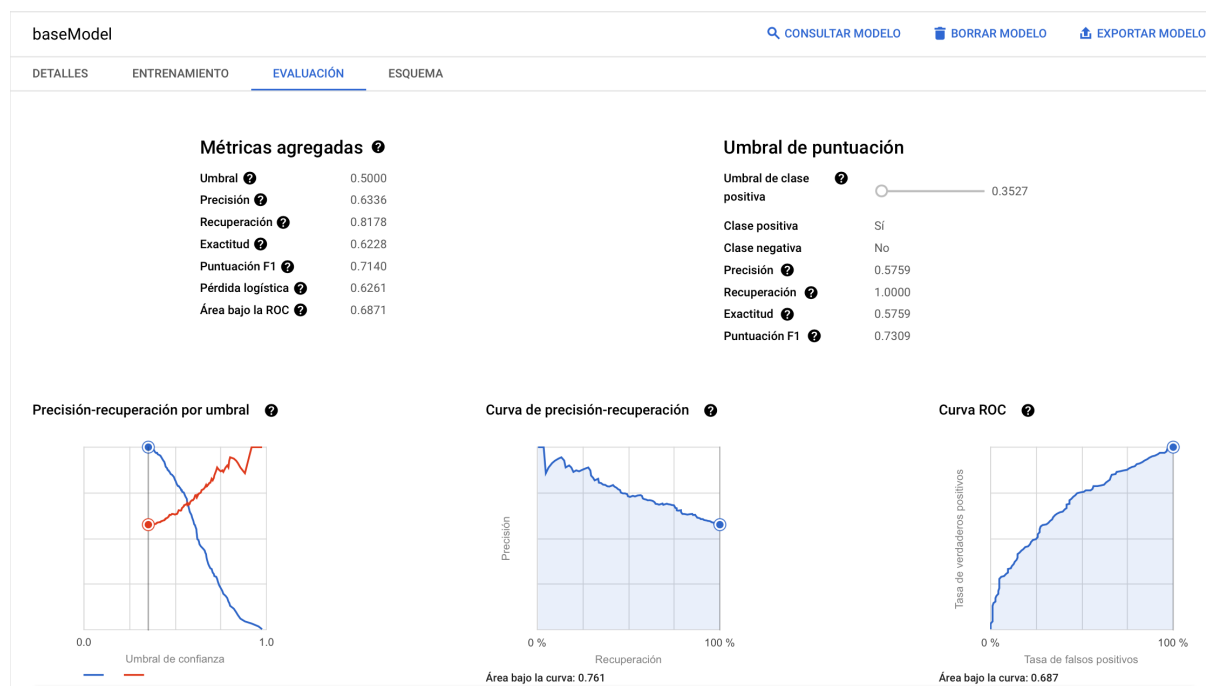


Figura 10.2: Avaluació del model

Per acabar, es troben les mètriques de pèrdua logística (*log loss*) i àrea sota la corba ROC (AUC). La corba ROC és un gràfic que mostra el rendiment d'un model de classificació en tots els llindars de classificació, presentant la taxa de falsos positius a l'eix d'abscisses i la de veritables positius a l'eix d'ordenades. L'AUC mesura l'àrea sota la corba d'aquest model i, segons el valor que prengui es pot saber com de bo o dolent és aquest:

- 0,5 - 0,6: El model és molt dolent
- 0,6 - 0,7: El model és dolent
- 0,7 - 0,8: El model és bo
- 0,8 - 0,9: El model és molt bo
- 0,9 - 1: El model és excel·lent

El valor de l'AUC es pot interpretar de la següent manera: donat un parell d'individus (accidents), un amb resposta positiva ($Y = 1$, hi ha víctimes mortals) i un altre amb resposta negativa ($Y = 0$, no hi ha víctimes mortals), es té un 69% de probabilitats que l'accident on sí que hi ha hagut víctimes tingui una probabilitat predita pel model superior a la probabilitat predita per a l'accident amb resposta negativa.

Pel model creat, el valor de l'àrea sota la corba és de 0,6871. Per tant, aquest model no és bo en termes d'AUC.

La pèrdua logística, per la seva banda, és un indicador de què tan a prop està la probabilitat de predicció del valor real corresponent i, per tant, pren valors del 0 a l'1. Com més s'apropi la probabilitat predita al valor real, major serà el valor de la pèrdua logística. En aquest model pren el valor de 0,6261.

Llindar de puntuació

A la dreta de la figura 10.2 s'hi troba una sèrie de mètriques. Per una banda, s'identifica que la classe positiva és la categoria "Sí", de la variable HIHAMORTS, mentre que la classe negativa és "No", i també s'hi troben altra vegada les mesures de precisió, recuperació, exactitud i puntuació F1. L'única cosa que canvia és el llindar de classe positiva, que es pot modificar a conveniència per l'usuari.

Per decidir a quina classe pertany un individu es pot establir un llindar. En funció d'aquest llindar, la probabilitat estimada obtinguda classifica les observacions. El valor per defecte que apareix a la figura és de 0,3527, i implica que tots els individus que tinguin una probabilitat de resposta positiva d'aquest valor o més alt es classificaran com a positius. Aquest és el valor mínim en el nostre cas, perquè no hi ha cap individu pel qual s'hagi calculat una probabilitat menor. Si ens fixem en la corba ROC, es veu que amb aquest valor de llindar, els falsos positius estan presents en el 100% dels casos, perquè cap individu s'ha classificat a la classe negativa. Si movem el valor del llindar, veurem com canvien les mètriques d'avaluació del model i en quin punt de la corba ens situem, ja que es reajusta la manera en què el model assigna la classe de la variable resposta a les observacions.

En el cas de la classificació logística binària, totes les mètriques reflexen els valors calculats

quan el llindar es fixa en 0,5, que és un criteri molt comú quan es treballa amb una variable resposta balancejada.

10.1.3 Interpretació del model

La interpretació dels pesos o coeficients beta calculats a la regressió logística és diferent de la interpretació dels pesos de la regressió lineal, ja que el resultat de la regressió logística és una probabilitat i, per tant, pren valors entre el 0 i l'1. La suma ponderada dels paràmetres es transforma, a través de la funció logística, en una probabilitat.

$$\log \left(\frac{P(y=Si)}{P(y=No)} \right) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

On p és el nombre de paràmetres que conformen el model.

Al terme esquerre de la igualtat se l'anomena *log-odds*. Els *odds*, que és l'equació que es troba dins del logaritme, són la probabilitat de resposta positiva entre la probabilitat de resposta negativa. Podem entendre el model logístic com un model lineal pels *log-odds*, ja que a través la combinació de termes es pot quantificar com canvia la classificació o la predicció en modificar alguna de les variables. Per fer-ho, s'haurà de treballar amb els exponents de la funció anterior.

$$\frac{P(y=Si)}{P(y=No)} = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}$$

Si canviem el valor d'una d'aquestes variables, per exemple augmentant en una unitat el valor de la variable x_1 , es pot calcular l'odds ratio de les dues prediccions:

$$\frac{odds_{x_1+1}}{odds} = \frac{e^{\beta_0 + \beta_1 (x_1+1) + \dots + \beta_p x_p}}{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}$$

$$\frac{odds_{x_1+1}}{odds} = e^{\beta_1}$$

Un canvi en una variable d'una unitat augmenta el *log-odds* tant com el valor del seu pes assignat.

Per entendre l'efecte del pes de les variables del model en la probabilitat de resposta positiva (accidents on hi ha víctimes mortals), executem la següent consulta, que retornarà els coeficients beta del model:

```
SELECT *
FROM ML.WEIGHTS(MODEL [nom_projecte] . [nom_base_de_dades] . [nom_model]);
```

Row	processed_input	weight	category_weights.category	categ... weight
1	MINUTE	-0.0024866...	null	null
2	RUR_URB	null	Urbà	0.0
			Rural	0.41865228...
			Desconegut	-0.0342466...
3	DAY_WEEK	null	Dissabte	0.0
			Dimecres	-0.0067840...
			Dimarts	-0.1741146...
			Diumenge	0.28300118...
			Dijous	-0.0604582...
			Dilluns	0.23710177...
			Divendres	0.01060096...
4	FATALS	0.79631994...	null	null
5	DRUNK_DR	0.27705565...	null	null
6	NO_PER	-0.0081475...	null	null
7	NO_VEHICLE	0.13381452...	null	null
8	HOUR_agrupat	null	Tarda	0.0
			Matí	0.11085031...
			Migdia	-0.1428410...
			Matinada	0.60195278...
			Nit	0.04825263...
9	__INTERCEPT__	-1.0228353...	null	null

Figura 10.3: Coeficients del model

El primer que farem serà mirar l'intercept (β_0). Quan totes les variables numèriques prenen el valor zero i les variables categòriques estan en la categoria de referència o categoria base, les *odds* estimades de resposta positiva són $e^{\beta_0} = e^{-1,0228353} = 0,3595739934$.

Els *odds* es descriuen com $odds = \frac{P(y=Si)}{P(y=No)}$ i, per tant, per calcular la probabilitat de resposta positiva haurem de fer el càlcul següent:

$$odds = \frac{P(y=Si)}{P(y=No)}$$

$$odds = \frac{P(y=Si)}{1-P(y=Si)}$$

$$P(y = Si) = \frac{odds}{1+odds}$$

$$P(y = Si) = \frac{e^{-1,0228353}}{1+e^{-1,0228353}}$$

$$P(y = Si) = 0,2644755$$

Les categories de referència de les variables categòriques són "urbà", pel que fa a la informació

de la localització; “Dissabte”, en el cas de dies de la setmana; i “Tarda”, per la variable d’hores agrupades. Per tant, la probabilitat de què hi hagi víctimes mortals en un accident que es dona en aquestes condicions és de 0,2644755.

Pel que fa a les variables numèriques, que són MINUTE, FATALS, DRUNK_DR, NO_PER i NO_VEHICLE, tenim que l’augment del valor de cada variable en una unitat afecta les probabilitats estimades en un factor de e^{β_i} , on i és el subíndex de cada variable numèrica.

En el cas de FATALS, per exemple, es troba una relació positiva entre el nombre de víctimes a l’accident i el fet d’haver-hi hagut alguna mort en el mateix. Per cada augment unitari en el nombre de víctimes, la probabilitat de classe positiva canvia en un factor de $e^{0,79631994} = 2,217365856$.

Altrament, les variables categòriques, RUR_URB, DAY_WEEK i HOUR_agrupat s’interpreten de forma que canviar la variable x_i de la categoria de referència a una de les altres afecta les probabilitats estimades en un factor de β_i .

Si ens fixem en la variable RUR_URB, es pot calcular que si, en comptes d’ubicar-se l’accident en una zona urbana, s’ha donat en una localització rural, la probabilitat canvia en un factor de $e^{0,4165228} = 1,519911759$, mentre que si la zona és desconeguda, el canvi és de $e^{-0,0342466} = 0,9663331775$.

11 Recopilació del material a Github

El codi en R i SQL, així com les dades utilitzades per la realització d'aquest treball es troben a un repositori de Github al que es pot arribar a partir de l'enllaç següent: <https://github.com/AnnaSalazar/TFG>.

A partir d'aquest material, es poden reproduir els resultats als quals s'ha arribat, i es poden visualitzar les comandes íntegres fetes servir en cada secció del treball.

El repositori està organitzat de manera que s'hi troben totes les dades, les originals i les preprocessades, així com el codi pel tractament i transformació d'aquestes. D'altra banda, hi ha una carpeta on es troben les valoracions dels estudiants, anonimitzades, per la consulta d'aquests i, per últim, el codi de l'anàlisi per components principals i de la creació del model logístic.

Tota aquesta informació està exposada en un arxiu README al mateix repositori, per facilitar la navegació dins d'aquest.

12 Conclusions

Per acabar, un cop explorada la plataforma BigQuery, tractant conceptes com la computació al núvol i l'emmagatzemament de dades en el mateix entorn, es farà un recull de conclusions i idees que han sorgit al llarg de la realització d'aquest treball.

L'objectiu principal d'aquest treball era la creació d'un material pedagògic per a incloure en el programari de l'assignatura de Fitxers i Bases de Dades del grau d'Estadística. Per fer-ho, es va crear un material enfocat al coneixement de BigQuery com a motor d'anàlisi de macrodades i la seva diferenciació envers altres eines d'anàlisi de dades. Aquest material també incloïa exemples pràctics a l'hora d'interactuar amb la plataforma, com pot ser la creació de conjunts de dades, la consulta a taules incloses en aquests i, fins i tot, la visualització dels resultats de les consultes. Un cop recopilada tota aquesta informació, es van dissenyar unes diapositives que encabien tot el ventall teòric, així com una pràctica guiada perquè els estudiants tinguessin l'oportunitat de posar-se a prova i experimentar ells mateixos el funcionament d'aquesta eina d'anàlisi.

La primera dificultat amb la qual ens vam trobar va ser que, en ser BigQuery una eina en desenvolupament que s'adapta de forma contínua a les necessitats del mercat, encara hi ha aspectes estètics o de presentació cap als usuaris que poden variar, com es va donar el cas amb Looker Studio. Durant el procés de recerca d'informació i la creació del material pedagògic, Looker Studio prenia el nom de Data Studio. Això no va comportar canvis pel que fa al funcionament, però va portar els estudiants a la confusió, perquè no trobaven el que exactament estava descrit al material. Així i tot, aquest imprevist es va poder resoldre a classe i no va suposar cap impediment per a la realització de la pràctica.

Quant a la implementació de BigQuery a l'aula, tot i que les sessions es van donar sense gaires complicacions pels estudiants i els comentaris van ser positius, no s'ha analitzat la valoració que van fer més enllà de la mitjana i la variància entre les puntuacions per falta de representativitat. Si es pogués haver realitzat la mateixa pràctica a un grup més gran, es podrien extreure conclusions que ajudessin a millorar el material de cara a l'any següent, o la manera de treballar amb aquest.

En últim lloc, pel que fa a l'anàlisi d'un conjunt de dades pròpiament, s'ha trobat que BigQuery permet alguns tipus d'anàlisi, però d'altres no. Això pot ser degut al fet que s'hagi volgut enfocar cap a l'aprenentatge automàtic o *machine learning*, i per això alguns altres tipus d'anàlisi com el de components principals no estan al seu abast, o són molt més complicats de realitzar en comparació a totes les alternatives que presenta R amb la seva gran varietat de paquets. Potser en un futur aquesta problemàtica es resolgui i BigQuery arribi a ser una eina completa per l'anàlisi estadística.

La realització d'aquest treball ha sigut una experiència molt positiva, personalment, ja que ha

suposat portar a les aules una cosa que no s'havia fet fins ara. Evidentment, no és factible estar actualitzat sobre tots els programes que existeixen actualment per l'anàlisi de dades, però com a estudiant trobo molt atractiu que s'explorin les possibilitats que existeixen i s'entengui a quines necessitats responen aquest tipus de programes emergents.

Referències

- [1] Comtois, D. (2021). Package 'summarytools'. *Tools to Quickly and Neatly Summarize Data*. URL <https://CRAN.R-project.org/package=summarytools>. R package version 1.0.0.
- [2] Cooksey, B. (2014). An Introduction to APIs. *Zapier, Inc.* Cvetojevic, S. Juhasz, L., & Hochmair, H. (2016). *Positional Accuracy of Twitter and Instagram Images in Urban Environment*. URL https://Doi.Org/10.1553/Giscience2016_01_s191.
- [3] Skoglund, K. (2019). *Git Essential Training: The Basics*. LinkedIn Learning. URL <https://www.linkedin.com/learning/git-essential-training-the-basics>.
- [4] Google. n.d. Quickstarts | BigQuery | Google Cloud. *Google*. URL <https://cloud.google.com/bigquery/docs/quickstarts>.
- [5] Grothendieck, G., & Grothendieck, M. G. (2017). Package 'sqldf'. *Manipulate r Data Frames Using SQL*. URL <https://CRAN.R-project.org/package=sqldf>. R package version 0.4-11.
- [6] Kassambara, A., & Mundt, F. (2020). Package 'factoextra'. *Extract and Visualize the Results of Multivariate Data Analyses*. URL <https://CRAN.R-project.org/package=factoextra>. R package version 1.0.7.
- [7] Kowarik, A., & Templ, M. (2016). Imputation with the R Package VIM. *Journal of Statistical Software* 74, 1–16. doi:10.18637/jss.v074.i07.
- [8] Lakshmanan, V., & Tigani, J. (2019). *Google BigQuery: The Definitive Guide: Data Warehousing, Analytics, and Machine Learning at Scale*. O'Reilly Media.
- [9] Ohri, A. (2014). *R for Cloud Computing: An Approach for Data Scientists*. Springer.
- [10] Rischin, J. (2022). *Looker Studio for Beginners*. LinkedIn Learning. URL <https://www.linkedin.com/learning/google-data-studio-for-beginners-2022>.
- [11] Wickham, H., & Müller, K. (2022). Package 'DBI'. *RDatabase Interface*. URL <https://CRAN.R-project.org/package=DBI>. R package version 1.1.3.
- [12] Wickham, H., & Bryan, J. (2022). Package 'bigrquery'. *An Interface to Google's 'BigQuery' 'API'*. URL <https://CRAN.R-project.org/package=bigrquery>. R package version 1.4.1.
- [13] Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., ... & Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. doi: 10.21105/joss.01686 (URL: <https://doi.org/10.21105/joss.01686>).

- [14] Xie, Y. (2022). Package 'knitr'. *A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.41.
- [15] Zhu, H. (2021). Package 'kableExtra'. *Construct Complex Table with 'kable' and Pipe Syntax*. URL <https://CRAN.R-project.org/package=kableExtra>. R package version 1.3.4.