UNIVERSITAT DE BARCELONA

FUNDAMENTALS OF DATA SCIENCE MASTER'S THESIS

A study of escoltes catalans census

Author:
David SOLANS

Supervisors:
Dra. Mireia RIBERA and Dr.
Eloi PUERTAS

A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamentals of Data Science

in the

Facultat de Matemàtiques i Informàtica

September 2, 2018

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

A study of escoltes catalans census

by David SOLANS

This thesis concludes the Masters studies in Data Science of the University of Barcelona during the course 2017/2018. It contains a study of the escoltas catalans census, an non-profit organization, part of the scouts foundation, with the aim of obtaining insights that might be useful for enhacing their internal organization or assisting future decisions by adding information.

The work done during this thesis includes different stages of the datascience pipeline, from data gathering and data enriching to the use of state of the art techniques for predicting future behaviours. The analysis performed on the obtained data resulted in two main outcomes:

- **Infographics**. A visualization explaining data characteristics and some interesting facts found while performing the data analysis was created. In order to create this visualization, techniques learned during the course of the Presentation and Visualization subject were applied.
- **Analysis of implication**. Aiming to have a better understanding of which factors determine the implication of the association members, a study of the length of stay and its correlated circumstances has been performed.

Apart of those outcomes, the job done around data cleaning, formatting and enriching might be considered another interesting outcome of this project.

To perform the analysis, open data has been leveraged to add the necessary information about the context and domain. To perform all done work regarding data acquisition, processing and analysis, different open source tools have been used.

All the information and details about project outcomes and development are explained in the present document.

Generated notebooks not containing personal of members of the association were pushed to: https://github.com/dsolanno/TFM-escoltas

Acknowledgements

I would like to thank my supervisors, Dra. Mireia Ribera and Dr. Eloi Puertas for their valuable and constructive suggestions during the planning and development of this project.

I am also grateful to all the staff-members of Fundació Escolta Josep Carol for their help and assistance into designing the objectives of this work.

Contents

Chapter 1

Introduction and motivation

1.1 Introduction

"If you can't measure it, you can't improve it.", this well known quote, written by Peter Drucker, an Austrian-born American management consultant, educator, and author, gives a complete explanation of the problem and objectives of this thesis.

With a growing interests on data analysis in multiple domains, all kinds of organizations are giving more and more importance to the information they have been storing for years. Profiles related to data analysis are being incorporated in multidisciplinary teams with the purpose of analyzing information to obtain insights from data and inform decisions and actions. The process of incorporating knowledge extracted from the stored information to assist taking informed decisions is called data-driven transformation and has became one of the big key topics in the domain of information technologies.

Although organizations are starting to recognize the value of this information could provide, even when it was traditionally ignored in the past, there are multiple technical challenges that make its use an arduous task. Between the main technical obstacles, the growing size of the information and the technical difficulty of the data cleaning procedures are key barriers to overcome before being in a position where any conclusion can extracted from data.

In this context, the work done during this thesis corresponds to the analysis of the census of an educative foundation, part of the Scout Movement. The Scout Movement was created at the beginning of the twentieth century and is a global initiative that aims to support young people in their physical, mental and spiritual development, so they may play constructive roles in society, with a strong focus on the outdoors and survival skills.

The mentioned census contains information obtained from three different associations: Escoltes Catalans, Acció Escolta and Minyons i Escoltes Catalans, all of them part of the Federatió Catalana d'Escoltime(Catalan Scout Federation).

The dataset analyzed during the development of this thesis was provided by another organization called Fundació Escolta Josep Carol, which can be undestood as the umbrella under which the other three associations are placed in the Scout hierarchy. The obtained data set corresponds of the historical census of the members of those three associations from around 1940 up to the end of year 2016.

Those three associations are all of them sharing a similar structure, composed by sub-units, called *"agrupaments"* that can be translated to *groups*. The groups have specific headquarters where they organize different activities for the members. In this sense, we can say that the information analyzed in the present document can be easily divided in two levels of granularity:

- **Groups**. Corresponds to the headquartes of the associations, where individuals or members assist for celebrating and paritipating in the different activities. In this documents, the terms *groups* and *agrupaments* are used to reference this centers.

- **Individuals**.Correspond to the members of the association. They are also categorized into a hierarchy depending on their age and the total amount of time they have been part of the organization. Members are part of only one grouping, although the grouping each member is part of might change across years. Along the present document, the terms *association members*, *members*, *escoltas* and *scouts* are used to name the individuals forming the Catalan Scouts Federation.

With the analysis described in the following pages, the objective of extracting useful information from data that might help decision-makers into better understanding their organization and, therefore, adapt their decisions accordingly was accomplished.

Chapter 2

Objectives and analysis

2.1 Obtained data

The census, as it was received, was composed by 7 files, in *.dta* format, each of them being binary files containing the census of a up to a different year between 2007, 2008, 2010, 2011, 2012, 2014 and 2016. For each individual in the census, attributes such as name, surname, date of birth, home address, gender, city and postal code was gathered. In total, there were 35.020 rows.

One important particularity of this data set refers to the way it was created. Traditionally, each grouping had their own census, typically stored in a paper. After a decision taken by board members across the three associations, it was decided to unify the census in a single file for each year, so all the papers were sent and one individual was in charge of the transcription from the physical to the digital formats. Taking into account that the information was manually created in its origin and then the information was digitalized manually, there are clearly two sources of errors. Also, as will be explained in the following sections, the attributes stored for the members were not consistent across groups, which made this data to contain a high percentage of null values.

The main challenge imposed was the fact of having individuals represented in more than one of the files, as for example, one individual member of a given group between 2009 and 2012 would appear in the census of 2010, 2011 and 2012. The way this challenge was solved is explained in section 3, in the Data Cleaning subsection.

2.2 Objectives and definition methodology

Apart of the scientific and technical challenges imposed by this project, there was also a component of project management. At the very beginning of the work, the key objectives and questions to be answered were not predefined, so the task of understanding what where the requirements of the organization and which of those requirements could be answered by analyzing the data, was part of the initial phases of project itself.

To do this task, monthly meetings where celebrated with board members from the Fundació Escolta Josep Carol, who where the contact point for this project. After some iterations, the objectives of the project were clarified and both, the necessities of the organization and the requirements of this capstone project were aligned.

In this case, the organization recognized the value of analyzing their census to understand how members behave. As an association that depends on the assistance of young members, their main focus is about understanding what influences or determines the total amount of time that a certain member keeps being part of the association, so they can run especial actions to achieve their objective of have people engaged with the association for longer periods of time.

During the different meetings celebrated with members of the association, the following list of questions was collected, containing all those queries that were mentioned as interesting for the board-members.

- **Q1.-** Ratio men-women in the association
- **Q2.-** Where do members live, what are the socio economic factors of those areas?
- **Q3.-** Where are groups located? What are the socio economic aspects of those places?
- **Q4.-** What are the differences between the rural and urban environments?
- **Q5.-** How are distances affecting the group members select to enrol?
- **Q6.-** Are there any interesting facts to be remarked about the length of the stay of association members?
- **Q7.-** Do having family members in the association affect the length of the stay of members?
- **Q8.-** Is there any relation between the location of the group and other formal educative places (schools/highscools)?
- **Q9.-** Is there any trend in the number of members of the different groups?
- **Q10.-** Is the length of the stay significantly different in urban environment compared to rural areas?
- **Q11.-** Which are the areas with higher density of escoltas?
- **Q12.-** Which are the areas with higher density of groups?
- **Q13.-** Is there any difference between the places where escoltas are poorly represented compared to places with higher density of escoltas?

The objectives of the work were about finding answers to those questions by using data. As can be observed, the purpose of understanding those answers was about obtaining insights that might help board members to have a better understanding of the organization and its members.

As hypothesis, socio-economic and transport-related factors such as distance from home to the groups headquarters where proposed as possible factors affecting the time that a given member remains part of the association.

Having said that, after some conversations with representatives from the foundation and the advisors of this thesis, it was decided to work in two different directions:

- **An infographics**. Creating a visualization explaining data characteristics and some interesting and relevant facts found while performing the data analysis. For this task, a subset of the questions [Q1-Q6] was selected with the idea of create a visualization that could answer them. Results found for the rest of the questions are gathered in A The explanation of how the infographic was designed and created can be found on the Chapter 4 of this document.

  The reasion for what this special type of visualization was selected were becouse infographics are known to simplify information in a visually engaging way in order to attract and inform a large audience, that makes infographics a good choice for engaging an heterogeneous community as the escoltas is. In other hand, by creating an static visualization rather than a dashboard, the complex task of setting up an infrastructure hosting the dashboard and other maintenance endevours were avoided.

- **A study of implication in the association**. By using the total amount of time that a scout remains part of the association, as the way of measuring implication, the initial purpose was about predicting how much time will a new member be part of the association given her characteristics.

  Although having a predictive model to estimate the total length of stay of a given member was interesting for them, when asking representatives about how were them expecting to consume and use the model, it was said that the model won't be directly consumed but an interpretation of which data features do the model use to perform the prediction are interesting for the association, as it might help into understanding which factors are related to the length of stay of a given individual.

  As happens with the created infographic, chapter 4 contains information about the process and results of the work done on studying the length of the stay of association members.

Given that the members of the association were all coming from the region of Catalonia, it was also decided that for certain types of analysis, it would make more sense to focus on the city of Barcelona, as it was found to be the city with higher density of members and groups per area unit.

Chapter 3

Development

3.1 Data cleaning

As has been explained in the first chapter of this document, the census information received was stored in 7 different binary files.

The work done in this section was required in order to facilitate future analysis. As the quality of the data was too low as it was received.

As Figure**??** depicts, tasks related to formats unification, data grouping, data enriching and geocoding were required. Those tasks are explained in the following subsections.

3.1.1 Unifying formats

To load the content of those files in Python, the Pandas module was utilized.

The loaded files were containing tabular data, where each row was containing the information relative to one member. Each member had , between others, the following fields (the corresponding field names written in Catalan language in the file header).

- **Name**. Called Nom, contains the name of the member.
- **Surname**. Called CogNom, contains the surnaname of the member.
- **MemberId**. NumSoci, created as unique identifier for each escolta in the census.
- **EntryDate**. DataAlta, data on which the member joint the association.
- **LeavingDate**.DataBaixa, data on which the member left the association.
- **Gender**. Sexe, gender of the member.
- **Address**. Adresa, member's home address.

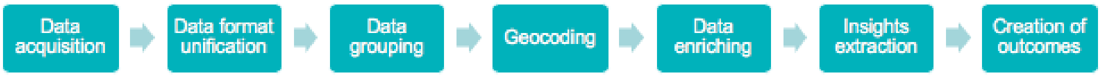


FIGURE 3.1: Data pipeline implemented during the development of this thesis

- **City**. Ciutat, member's city of residence address.
- **PostalCode**. PostalCode, member's home zipcode.
- **CurrentGroup**. AgrupamentActual, name of the last group the member was part of.
- **Email**. Email of the individual
- **SocialSecurityNumber**. NumSegSocial, social security number of the association member.
- **Telephone**.Phone number of the person
- **SexualOffencesCertificate**. certificatdelictessexuals people registered from 2014 as educators was required by Spanish legislation to present this certificate, confirming that they had never been charged as defendants in sexual offences.

Although it might not be the case for an English reader (attribute names where written in Catalan) the names were enough self-explanatory, which was the first good new. Apart of this, the information was more or less consistent among years, the field names were different per file, it was decided to unify information so all the records were having an unified list of attribute names. It was decided to keep Catalan field names to facilitate the possible future re-usage of the resulting after the aggregation of records by members of the association. The final list of attribute names was:
['Adresa', 'AgrupamentActual', 'CertificatDelictesSexuals', 'CodiPostal', 'Nom', 'Nom2', 'Comarca', 'Correu', 'DataAlta', 'DataBaixa', 'DataCertificatDelictesSexuals', 'DataNaixament', 'Edat', 'Fax', 'IdSoci', 'Mobil', 'NIF', 'CogNom', 'CogNom2', 'NomAgrupaUltim', 'NumSS', 'NumSoci', 'Poblacio', 'Provincia', 'Sexe', 'Sexe-def', 'Telefon', 'Telefon2', 'UnitatActual', 'Year', 'codi', 'notUsed', 'v26', 'v27']

Apart of the work done on unifiying field names, there were also language encoding problems with accents and special characters that were solved by substituting them by other characters.

Once the field names were unified across the files, they were merged into a single file. This action partially solved the initial problem of heterogeneity in the data source. The created file was load in memory in a single data frame containing 35.020 records that had a common set of fields.

3.1.2 Data grouping

To solve the problem of having individuals represented in more than one record in the data source, it was required to devise an strategy that could help into grouping those repeated rows into a single one, so conclusions were not biased by the fact of having repeated records.

In this sense, the first stragegy was about using the field *NumSoci* that, as was explained, was suposed to be a unique identifier for each member. When analizing the values of this field, there where 21.544 unique values with 2685 of them being not valid data (NaN in *Python*).

Nom	**Nom2**	**Cognom**	**Cognom2**
David	Solans	Noguero	
Solans		Noguero	David

TABLE 3.1: Example of two rows containing name and surnames of a given user with their values swapped across columns.

When analizing more in detail the fields with equal *NumSoci*, it was found that the identifier is unique per each year, but members did not have the same *NumSoci* across years. As an example, the member with name Pepito Sánchez, could have *NumSoci* 19991 in the file of 2009 but *NumSoci* equal to 00001 in the file of 2010, so it could not be used for the data grouping as it might be definitively adding a source of errors when performing the data aggregation.

The next option was about using the *NIF* field, which was suposed to contain the identity number of each member. Regarding the *NIF*, there where 28783 different values in total, meaning that almost each row had a different value for the *NIF* field. The reason behind this fact was rapidly diagnosed: *NIF* formats were not consistent accross years, as members were writting their *NIF* sometimes with the letter in first place and other times, when the letter was written in the last position, it was written using a dash or a hyphens to separate the digits from the letters. Again, non perfect aggregation was possible with this technique.

Fields containing names and surnames such as *Nom* or *Cognom* could also be usable for the aggregation task. Although they were also challenging, as sometimes their values are misspelled or swapped (e.g.: surnames placed in the name field and vice versa). The solution for this field was about concatenating all the names and surnames for each row into a single string containing the values of the fileds *Nom*, *Nom2*, *Cognom*, *Cognom2* lowered, sorted alphabetically and concatenated.

Following this approximation, a user having in two rows in the dataset where whose names and surnames are wrongly placed might be aggregated into a single one, as if applying the stated technique, the names and surnames will be aggregated in new field having the same value for both records.

Following the example of Table 3.2, after this procedure, both records will contain a new field with value *david,noguero,solans*, which makes it possible to automatically identify both records as representing one single user.

Although those three approximations could work for a subset of the records, it was clear that none of them was giving quality results at 100%. For this reason, it was required to devise a methodology to measure the quality of each aggregation or grouping.

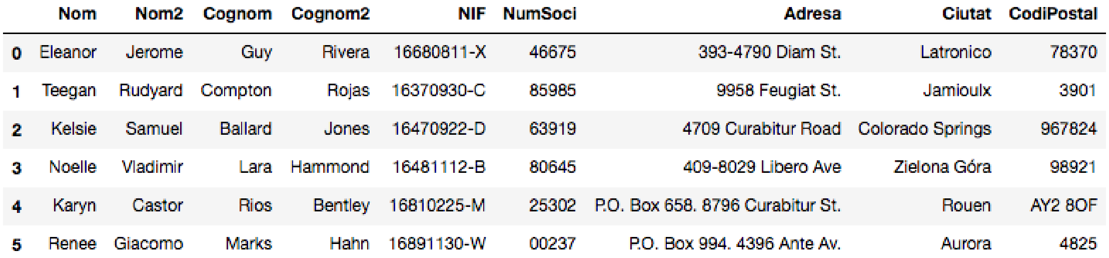
Measuring groups quality

To solve this challenge, a sub set of golden samples where used. A golden sample is typically hand-crafted with a lot of attention to the fine details, meaning that for this subset, the relations of which records were containing information of a given user were known.

To manually create those golden samples, an automatic data generator was used to craft records for each individual, see Figure3.2 for example of generated information. The values of each record was slightly modified to simulate those stated

Technique	Correctly aggregated rows
Using numSoci	51.03%
Using NIF	68.72%
Using Name	73.12%
Using NIF and Name with edit distance	89.64%

TABLE 3.2: Percentage of correctly aggregated rows for each of the stated data grouping techniques.



	Nom	Nom2	Cognom	Cognom2	NIF	NumSoci	Adresa	Ciutat	CodiPostal
0	Eleanor	Jerome	Guy	Rivera	16680811-X	46675	393-4790 Diam St.	Latronico	78370
1	Teegan	Rudyard	Compton	Rojas	16370930-C	85985	9958 Feugiat St.	Jamioulx	3901
2	Kelsie	Samuel	Ballard	Jones	16470922-D	63919	4709 Curabitur Road	Colorado Springs	967824
3	Noelle	Vladimir	Lara	Hammond	16481112-B	80645	409-8029 Libero Ave	Zielona Góra	98921
4	Karyn	Castor	Rios	Bentley	16810225-M	25302	P.O. Box 658. 8796 Curabitur St.	Rouen	AY2 8OF
5	Renee	Giacomo	Marks	Hahn	16891130-W	00237	P.O. Box 994. 4396 Ante Av.	Aurora	4825

FIGURE 3.2: Example of generated information for the golden samples

behaviours observed in the census. The created samples where representing a total of 25 members. For each of them, a randomly (uniform distribution) selected number of samples (between 2 to 7) was selected. By using this crafted records, that we knew how where they related (in the sense that we knew wich records where representing the same individual), the three techniques stated above where tested, obtaining poor results with all of them.

The reason making the three techniques failing in more cases than expected was the presence of null values and typos, making it impossible in most of the cases to perform an exact-matching between fields to make the aggregation.

The final solution was about using a combination of two ideas. To develop this solution, the *NIF* field and the created field containing the name strings concatenated were used. Comparison of both fields where done in a one by one basis. Accepting as valid those matches between a real value and a null one (to avoid discarding aggregations by the fact of one record not having the field informed) and those having an edit (Levenshtein) distance below 10% the length of the compared strings (to allow matched strings to be slightly difference, probably because of transcription errors and typos).

This solution worked reasonably well for the golden samples and after applying it to the census data set, zero errors where detected in a visual inspection. See Table**??** to know more about the accuracy of each method.

As a result, the information was aggregated in a new data frame containing a total number of 15.092 samples. This data frame was used in all the calculations explained in the following sections of this document.

3.1.3 Data enriching

During this phase, a search for public and open data sets was conducted. The selected datasets where:

- **Postal codes polygons of Catalonia**: Not found in the internet, as one must pay to the spanish public entity of shipments to obtain it. Luckily, I kept a copy from and past project.
- **Postal codes polygons of Barcelona**: Obtained by querying file above for polygons with identifiers of Barcelona's postal codes.
- **Neighborhood polygons of Barcelona**: Obtained from https://github.com/martgnz/bcn-geodata. Whose fantastic work I would like to thank.
- **Barcelona level of studies per area**: Obtained from the Barcelona city council web: http://www.bcn.cat/estadistica/catala/index.htm.
- **Barcelona average income per area**: Obtained from the Barcelona city council web: http://www.bcn.cat/estadistica/catala/index.htm.
- **Catalan municipalities socio economic factors**: Publicly available at https://www.idescat.cat/

As happens too often, usually open data ends not being so open in reality, given the difficulties that one might find when trying to consume "public" data to include it in certain automatic reasoning. Although the problem with finding the postal code polygons was already expected, as I had to overcome this difficulty in past projects, there was a non-expected problem with obtaining data from Idescat for each Catalan city.

Although it is possible and quite easy to access a table-based visualization of all the socio-economic factors of a certain municipality from Catalonia (e.g.: https://www.idescat.cat/emex/?
a procedure to download all this information was not found. While considering the option of manually opening the urls for the 947 municipalities that Catalonia has and click the button with the message *In Excel* to download an *.xsls* file containing the displayed information, but at the time of trying to download the information from Idescat (and it remains the same at the time of writting this document), the action launched by the button to download the file was not working properly, so it was not possible to download the file per each Catalan city or town in a manual manner.

Given this situation, I decided to create a scrapper by using two well known Python packages to first, download the information displayed in Idescat for each municipality into a file and then, parse the content of each downloaded file to obtain those attributes which where identified as interesting for the project.

In concrete, the list of attributes found interesting for the project where:

1. **Population**
2. **Population. Men**
3. **Population. Women**
4. **Population. From 0 to 14 years**
5. **Population. From 15 to 64 years**

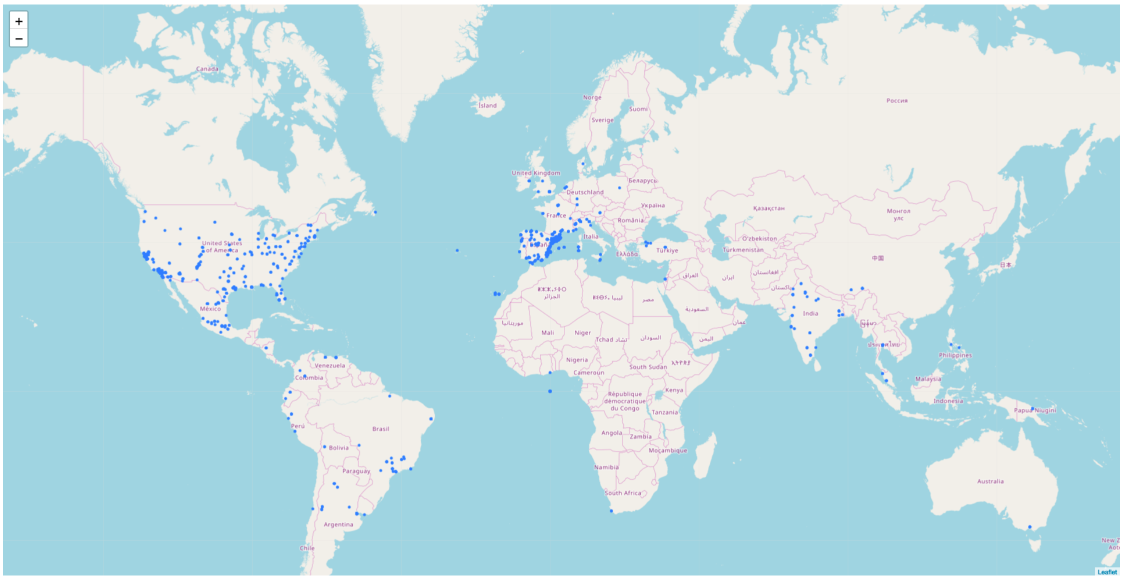


FIGURE 3.3: (Blue)Points where geocoding disambiguated addresses from the dataset. A significant fraction of them was wrongly found located outside Catalonia

6. **Total weath by declarer (euros)**
7. **Total amount of sports facilities**
8. **Affiliations to Social Security by labour sectors**

After this work, the result was a dictionary per each municipality containing the corresponding values, so this information could be easily merged with the census data during the next developments for this thesis.

At the time of writting this document, an offical HTTP API of the Idescat web was found as accessible from: https://www.idescat.cat/dev/api/emex/

3.1.4 Geocoding

According to the Oxford Dictionary, geocoding refers to the fact of providing the geographical coordinates corresponding to a location or physical address. In the context of this project, geocoding has been used for transforming addresses to geospatial points (latitude/longitude coordinates) so calculations such as distances between individuals residences and group locations can be done.

During the first attempt of geocoding the addresses of the data set, each of the addresses was concatenated with the city and the postal code to create a new column for the data that received the name of *ComposedAddres*.

By using the official Google Maps client for geocoding, after obtaining an ApiKey, as it is required by the provider, the values of the *ComposedAddress* field where geocoded one by one. Obtaining the points (lat/long) depicted in the Figure 2.2.

As it can be easily see in Figure3.3, a significant number of points where situated out of Catalonia. Analyzing the results, only 83% of the points where placed inside the polygon of the catalan region, even without chequing the quality of the information of those points placed inside Catalonia, the obtained results were already poorer than expected.

Before starting working in solving issues, a way of validating the obtained geocodings was required. To do so, a hierarchical validator based on geospatial polygons was used. Given that for each record, the address, postal code and city were known, the validator was checking quality of the obtained points by checking if they were located inside the corresponding polygon. To avoid performance issues, the checking of the point in Catalonia was the first check so if this condition was not accomplished the rest of calculations was skipped, saving this way a significant amount of CPU computations for each incorrect location.

The flow of this validation is depicted in the pseudocode contained in *Algorthim1*:

Algorithm 1 Geocoding validator

```
Require: $records, point_{obtained}$ while N $\leq records_{length}$ do
    $point_N = point_{obtained}[N]$
    if $point_N \in Catalonia_{polygon}$ then
      if $record \ni postalCode_{value}$ then
        if $\exists postalCode_{polygon} \wedge point_N \notin postalCode_{polygon}$ then
          $returnFALSE$
        end if
      end if
      if $record \ni city_{value}$ then
        if $\exists city_{polygon} \wedge point_N \notin city_{polygon}$ then
        $returnFALSE$
        end if
      end if
    else
      $returnFALSE$
    end if
  end while
```

Once the geoconding validator was prepared, a geocoder supporting multiple providers was used.In concrete, the list of used geocoding providers was:

1. Open Street Maps
2. Google Maps
3. ArcGIS

After some experiments, it was found that for specially complex addresses, the variance of the results returned by each provider was significant, up to the point that, for addresses located in Barcelona, points with differences of up to 3km distance between results where found. See Figure3.4 for an example.

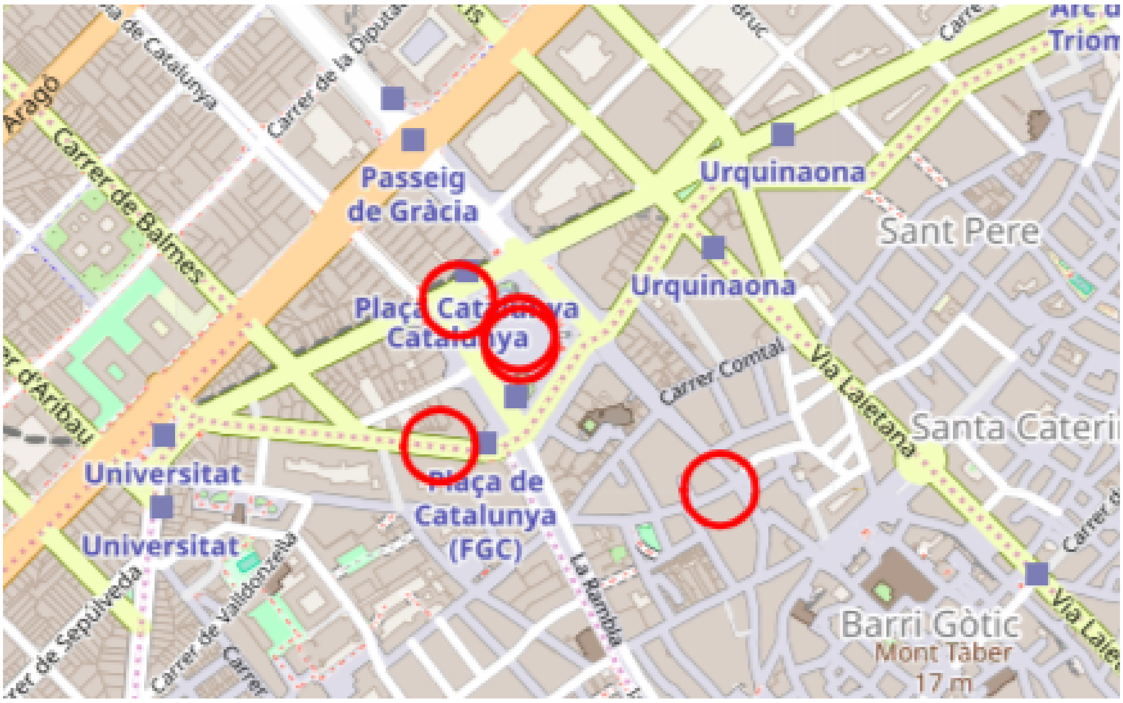


FIGURE 3.4: Example of variance in geocoding returned by different providers (in this case, looking for Plaza Catalunya, main square of Barcelona).

With that, main geocoding service used was Open Street Maps as it allowed more calculations without reaching the API limit. For each point not correctly located, other providers are checked looking if any of the resting geocoding providers was locating the addresses inside the corresponding polygon.

As result of all this work, a final number of 87% of the addresses where correctly located inside their geospatial polygon. Reducing significantly the original error on geocoding addresses.

Even when more work could have been doe in this topic, for example, by incorporating natural language processing techniques to structure the addresses, it was decided to not follow this direction, as it was not the main focus of this project.

3.2 Tools used during development

To perform the data analysis of this thesis, *Python* was used as programming language. The version used during the development was *v3.6*. Together with *Python*, the following modules where used for certain tasks:

- **Pandas**: Used for managing, filtering and transforming the data obtained as source of information.
- **Geopandas**: Given the geospatial nature of the information, geopandas was also used for performing geospatial-based transformations as well as for creating map-based plots.
- **Matplotlib**: Used to create charts, to help obtaining insights from the data.

- **Selenium** and **BeautifulSoup**: Used to scrap information from the web, key task for the data enriching prodecure.
- **Scikit-learn**: Used for data mining and machine learning models training and evaluation.
- **Pymining**Used to extract association rules of the data.
- **Geocoder**: Used to compare results among different geocoding services
- **Microsoft PowerPoint** Used to create the infographic layout as a composition of images created with other packages.
- **Draw.io** Drawing tool used to create the initial mock of the infography.
- **generatedata.com** For generating fake personal information to be used as golden truth in the data aggregation task.

.

Chapter 4

Results

4.1 Infographic

An infographic is a clipped compound of information and graphics. Infographics are graphic visual representations of information, data or knowledge intended to present information quickly and clearly. They can improve cognition by utilizing graphics to enhance the human visual system's ability to see patterns and trends.

As any other visualization effort, the infographic was designed as a communication mechanism between the end user and the designer. For this reason, the first step was about working on an user analysis to obtain information about the characteristics of the potential users to consume the visualization.

4.1.1 User analysis

Although in principle the infographics was designed to be consumed by board members of the organization, who could be considered as expert users, it might be also used as dissemination asset, to it should be also understandable and interpretable by novice users.

Unless it will be typically consumed by individual users (audience size equal to one) and it was assumed that users will consume it by using the browser of a laptop with an standard screen, the infographic design must be also prepared to be printed to create, for example, a poster to be shown in each group venue.

The main objective of the user interpreting the visualization will be about performing an exploratory analysis.

4.1.2 Tasks and functional analysis

By interpreting the infographics, the user should be able to understand different characteristics of the escoltas and groups forming the association.

Given that the association is mainly composed by young people, the visualization shoudl be clear and prepared for all types of audiences.

4.1.3 Initial design

Initial design of the infograpics can be seen in the appendix B of this document.

The entry-point, as object to obtain the attention of the user, a big number with the historical total amount of members of the association will be displayed.

From here, the information is organized in different sections, each of them grouping information around a certain topic.

Ratio men/women in the association

In this first section, the total percentage of men is compared with the percentage of women and the total percentage of members with unknown gender (as the sum of women and men percentages is not 100% and it could cause confusion).

Analysis of member's home locations and association's groups locations

The following two sections have an especial layout in two columns, with the biggest column containing a big map of Catalonia and the smallest, containing a equivalent map for the city of Barcelona, as it is the one where most of the groups and members are based.

To relate the map of Catalonia with the map of Barcelona, a *zoom* has been used to create illusion of a real magnifier lens placed over the biggest map.

After different compositions, it was decided to place the map of Barcelona in different sides in each of the sections, as it helps the user to understand that both plots are not containing the same information.

Distance to the closest group and real membership

This section compares two situations.

The first situation corresponds to an imaginary environment where all the members were enlisted to the group with was closer to their home locations.

The second one corresponds to the reality, where not all the members are using distance as the only criteria to select their group.

Length of stay

This last section was reserved for some interesting fact around the topic of the length of stay, as it was clearly pointed as one of the most interesting analysis by the *customers* of the project.

4.1.4 Obtained result

The final result of the infographics can be seen in the AppendixB of this document. As it explained the Appendix, the layout of the final implementation of the infographics was finally modified to landscape to facilitate the readability of the visualization in computer screens.

In addition to this, the landscape-based implementation was also modified to use the historical number of escoltes as element to catch the attention of the user.

To select an accessible combination of colors, a combination that is accessible to people with disabilities, http://www.colorsafe.co was used to select a safe color palette.

Although this version of the infographics was created in English language, copies of this design but written in Spanish and Catalan were also created and will be provided to the association.

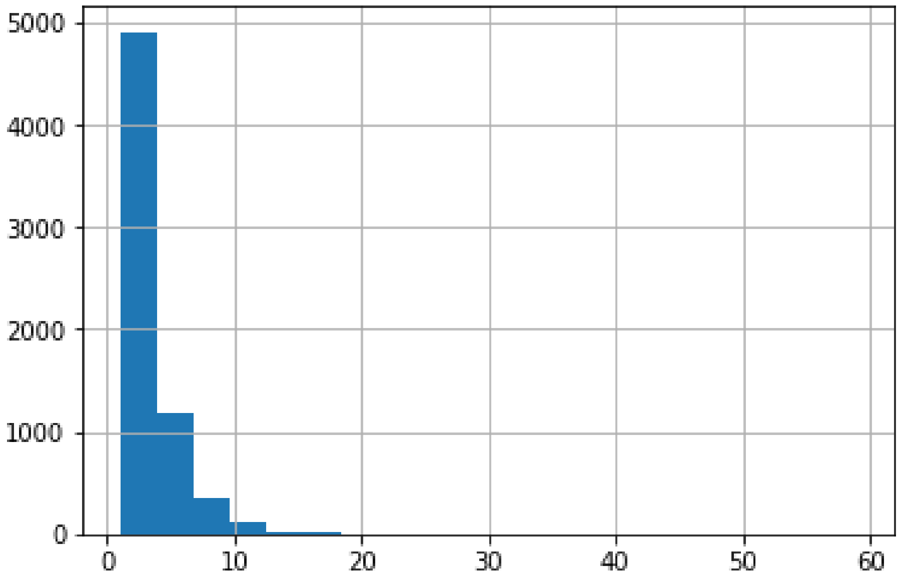


FIGURE 4.1: Distribution of number of years of stay across the association

4.2 Length of stay analysis

As has been already mentioned above in this document, even when the initial objective of this work was to create a model to predict the length of the stay for a given member characteristics, it was clear after some discussions that the most interesting part of the model would be which where the relevant features, rather than the accuracy of the model itself.

This means that the association board members where not interesting in using the model for predicting the length of the stay for a given individual, but understanding what are the found features that are affecting more dramatically the length of this stay. The following sections depict the work done around this topic.

Main obstacle in the analysis performed in this section corresponds to the fact of having a percentage of 25.83% of the rows with a valid value for entry and leaving dates. Making this analysis to be biased from the beginning.

4.2.1 Distribution of stay duration across groups

To calculate the length of the stay, those rows having valid values in both the entryDate and leavingDate columns where used. The elapsed time was calculated both in years (see Figure4.1) and months (see Figure**??**fig:stay$_{m}onths).Havingameanof2.830834yearsand30.409684m$
The following work done for understanding the length of the stay in the organization was about comparing this length between different groups.

The distribution of length of stays per gender can be seen at Figure**??**

One of the most interesting analysis was about calculating the stay duration per group unit. Groups units are assigned to members in relation to their experience as escoltas. The distributions obtained per each group where clearly reflecting this. See Figurefig:stay$_{u}nit$

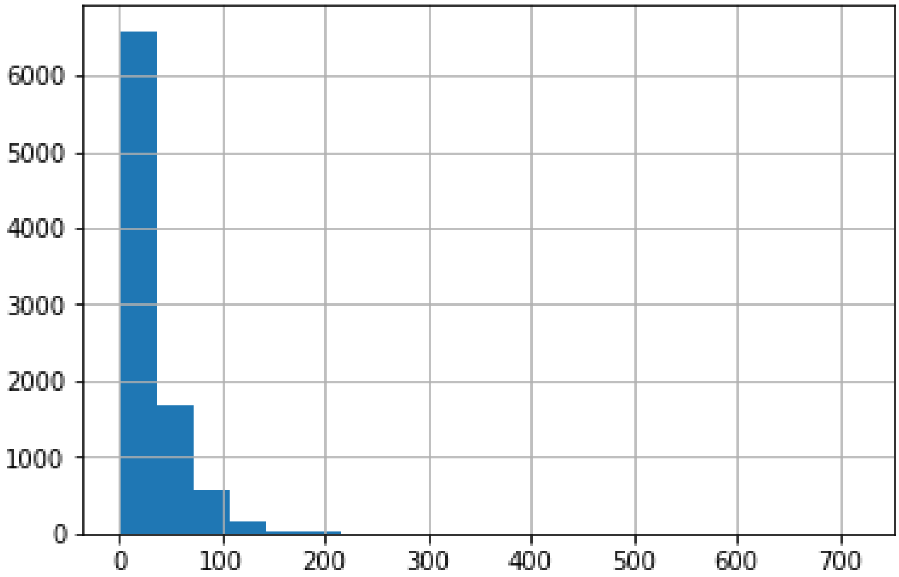


FIGURE 4.2: Distribution of number of months of stay across the association

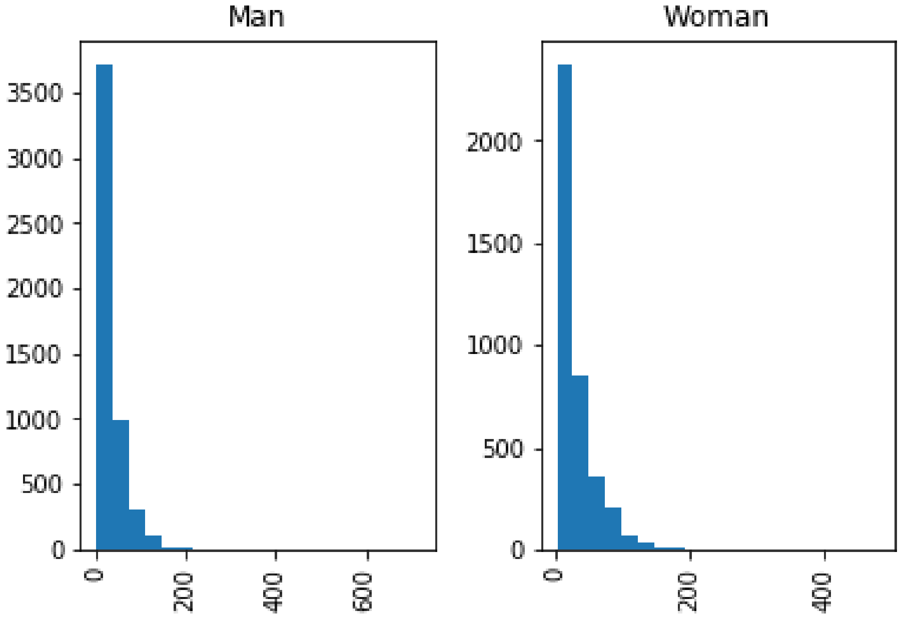


FIGURE 4.3: Distribution of stays per gender

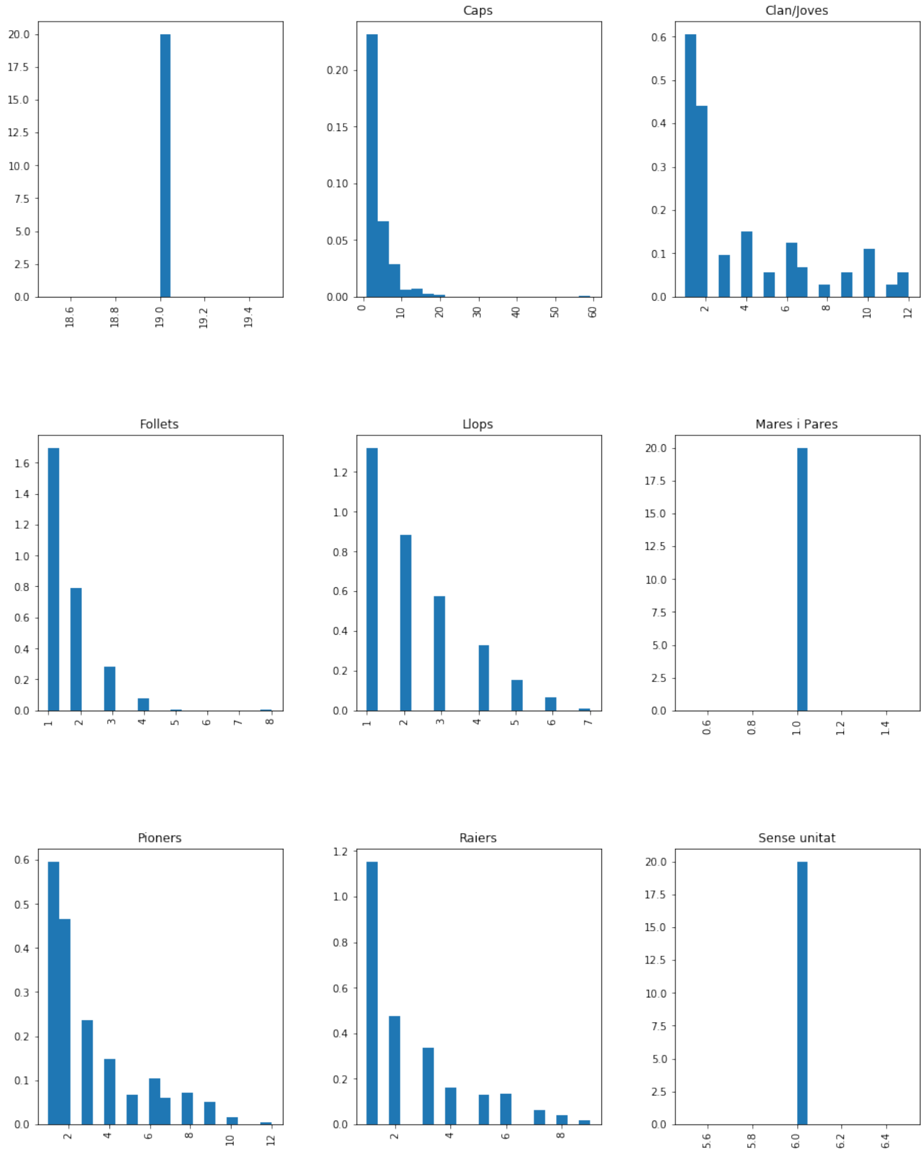


FIGURE 4.4: Distribution of stays per unit

4.2.2 Predictive model

The predictive model was intended to predict the length of the stay for a given member. To do so, the historical data set was used.

Converting the forecasting problem into a classification problem

The forecasting problem was converted in a classification problem as the extact value of the prediction is not so important, but it was more relevant to know if a given member would stay in the association during a long, long-medium, short-medium or short period of time.

To do so, lengths of stay values were converted into categories, following two strategies:

- **Percentiles. Balanced data** Creating categories that were grouping data between the 4 percentiles (25%,50%,75%, 100%). This technique was optimal for learning, as it was creating balanced data (all the classes containing 25% of the samples).

  item **Manual categorization. Unbalanced data** As it was discussed, understanding if the new member would stay during less than one year, up to two years, up to five years or more than that, would make classification more user-friendly, as those would be grouping data per groups of interest for the board members although it had some problems for learning, as the number of elements per each class was not balanced.

Both categorizations were compared, training and testing for them a set of 10 different models.

Selected models

Multiple models where compared for the first attempt of creating the predictive model. In concrete, the following list contains the selected collections of models, their parameters and the id used to identify them during the computations:

- **'l2logreg'** Logistic Regression CV(penalty='l2'cv=5fitIntercept=True)
- **'lasso'** LassoCV(cv=5,fitIntercept=True)
- **'xgb'** XGBoost Classifier(maxDepth=3 nEstimators=300 learningRate=0.05)
- **'logreg'** Logistic Regression(fitIntercept=True)
- **'SVM'** Support Vector Machine(C=1)
- **'No-params-logReg'** Logistic Regression
- **'RandomForest'** Random Forest(maxDepth=2)
- **'RandomForest2'** Random Forest(maxDepth=4)
- **'DecissionTree'** Decision Tree Classifier(minSamplesSplit=25)
- **'DecissionTree2'** Decision Tree Classifier(minSamplesSplit=50)

Model	Accuracy
l2logreg	0.675167
lasso	0.207800
xgb	**0.680477**
logreg	0.675167
SVM	0.675167
No-params-logReg	0.675167
RandomForest	0.675167
RandomForest2	0.678445
DecissionTree	0.675429
DecissionTree2	0.673463

TABLE 4.1: Obtained classification accuracies for unbalanced data

Results

To calculate the results, cross validation was used. Cross-validation is used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

In concrete, K-fold validation was the methodology used to test the quality of the learning of each classifier for each portion of training data. The general procedure of the K-fold is as follows:

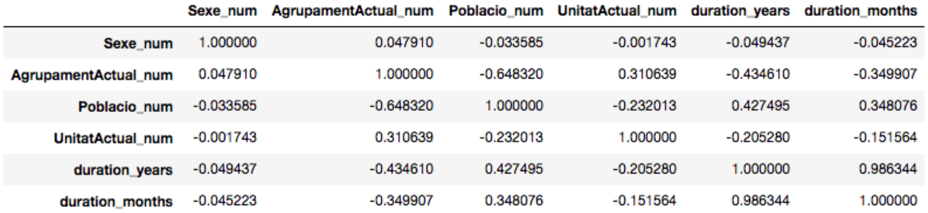
1. Shuffle the dataset randomly.
2. Split the dataset into k groups
3. For each unique group:
4. Take the group as a hold out or test data set
5. Take the remaining groups as a training data set
6. Fit a model on the training set and evaluate it on the test set
7. Retain the evaluation score and discard the model
8. Summarize the skill of the model using the sample of model evaluation scores

After this process, the obtained results where the following:

With the classifier *XGBoost* obtaining slightly better accuracies in both the balanced and the unbalanced use cases. A summary of the obtained accuracies for the different models can be seen in tables: Table4.1 and Table4.2.

Model	Accuracy
l2logreg	0.7316
lasso	0.6216
xgb	**0.7475**
logreg	0.7349
SVM	0.7346
No-params-logReg	0.7349
RandomForest	0.7278
RandomForest2	0.7474
DecissionTree	0.7459
DecissionTree2	0.7470

TABLE 4.2: Obtained classification accuracies for balanced data



	Sexe_num	AgrupamentActual_num	Poblacio_num	UnitatActual_num	duration_years	duration_months
Sexe_num	1.000000	0.047910	-0.033585	-0.001743	-0.049437	-0.045223
AgrupamentActual_num	0.047910	1.000000	-0.648320	0.310639	-0.434610	-0.349907
Poblacio_num	-0.033585	-0.648320	1.000000	-0.232013	0.427495	0.348076
UnitatActual_num	-0.001743	0.310639	-0.232013	1.000000	-0.205280	-0.151564
duration_years	-0.049437	-0.434610	0.427495	-0.205280	1.000000	0.986344
duration_months	-0.045223	-0.349907	0.348076	-0.151564	0.986344	1.000000

FIGURE 4.5: Correlations between values of the dataset

4.2.3 Relevant features

To perform this analysis, two basic techniques were used:

In one hand, correlation between variables, once their values converted into numeric format, was performed. Although without being able to extract any relevant result due to the amount of missing values found in the data, as the source was the result of selecting those rows having enough information to calculate to length of the stay and adding the socio-economic information for the municipality where the row was located. In addition, to this, not all the fields in information are available for all the municipalities, making correlations with socioeconomic data not possible to be calculated. For correlations between variables of the dataset, see Figure4.5

In the other hand, a technique called *Association rules*, commonly used in data mining was used. *Association rules* are if/then statements that help uncover relationships between seemingly unrelated data in a information repository. An example of an association rule would be *"If a customer buys a dozen eggs, he is 80% likely to also purchase milk."*, where the first condition being *"If a customer buys a dozen eggs"*, the *80%* corresponds to finding the second condition, *"purchase milk"*, given the first.

This rules are helpful to detect dependencies between groups of variables. The algorithm used for extracting the rules allows the user to select two parameters:

1. **Confidence Level**. Being a percentage, allows the user to filter the probability of finding the second condition to be true, given the first.

2. **Support**. Support is used to reference the minimum number of appearances of a given rule to be extracted.

Although many different values of both parameters where used, a large number of rules were always extracted, not being able to extract any final conclusions from this work.

Chapter 5

Conclusions

5.1 Conclusions

Apart of the obtained results, one of the lessons learned from this work is the importance of the quality of the data. As approximation, 80% of the effort done has been around data cleaning and formatting. And even with all this effort, there was space for more work, for example in the field of geocoding.

5.1.1 Causality

Although it was expected from the beginning of the project, that causes of any event observed in the data might be difficult to find, it is clear that understanding causality in such complex environments its and arduous task. In this attempt of understanding relations between social factors and some attributes and trends of the association, it has been nos possible to prove causality, as given the nature of the scenario, only correlations were observable.

This means that there is still a lot of space to improve this work, specially from the point of view of extra information, as adding all existing information and domain knowledge might be useful to understand which factors are causing certain observed events.

5.1.2 Acquired experience

During the development of this project, I have had the opportunity to work in a real world problem, with real customers and raw data.

Although data quality has been one of the main challenges, it gave the opportunity to devise original cleaning strategies such the techniques used for aggregating those records containing information from the same individual.

I really see all the work done as an interesting step for growing my expertise in the field of data science, as it is a nice example of the complete pipeline: from the data gathering to the final conclusions.

Appendix A

Interesting facts

A.1 Analyzing proposed questions (not included in the infographics)

A.1.1 Q7.- Do having family members in the association affect the length of the stay of members?

Answering this question has not been possible due to the fact of not having names ordered in the data set. This makes detecting family members almost impossible, as surnames can not be extracted and therefore, compared.

A.1.2 Q8.- Is there any relation between the location of the group and other formal educative centres (schools/highscools)?

After locating the educative centres in Catalonia, see FigureA.1, it was obvious that their locations are related to the locations where groups are located, as all of them are always located in city centres.

The found average distance between a group and it's closer educative centre for Catalonia was 0.5003 meters.

A.1.3 Q9.- Is there any trend in the number of members of the different groups?

In general, the trend for all the groups is stable among the time. With a clear drop in the number of years that a member remains part of the association after being part of the *Caps* unit. One hypothesis for the reason causing this fact is stated in the following answer.

A.1.4 Q10.- Is the length of the stay significantly different in urban environment compared to rural areas?

Understanding rural areas, to those municipalities having a population below 20k citizens, the obtained result concluded that even when it's a common trend, people from rural areas are more likely to leave at the stage of being Caps (18 years old). This might be caused by the necessity of change their residences for starting their new stage at the university.

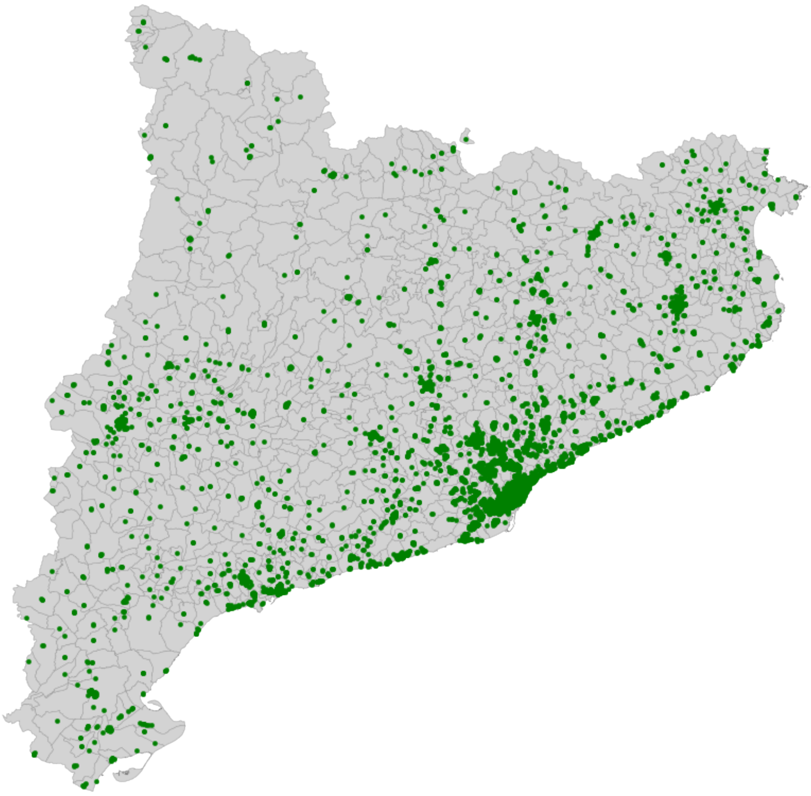


FIGURE A.1: Location of educative centers in Catalonia

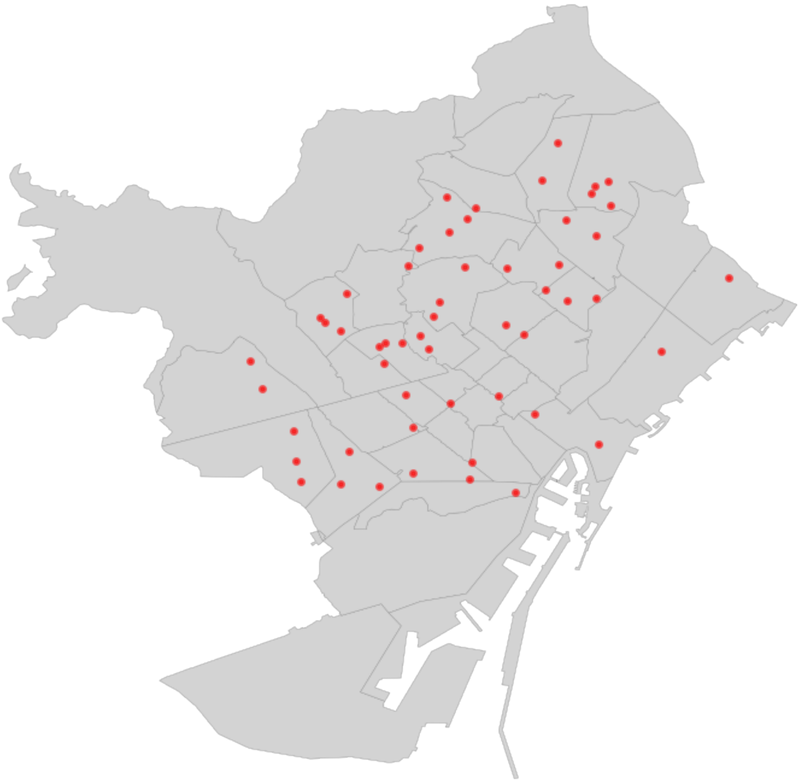


FIGURE A.2: Group headquarter locations in Barcelona

A.1.5 Q11.- Which are the areas with higher density of escoltas?

Whereas the city with a bigger historical number of members is Barcelona, with 2.899 members, the city with a higher percentage is Alpicat, with a ratio $number_of_escoltes/population$ of 0.0382

A.1.6 Q12.- Which are the areas with higher density of groups?

The area with higher density of groups is the city of Barcelona with a density of 0.5325 groups per km^2. See FigureA.2 for understanding how are groups located in the city.

A.1.7 Q13.- Is there any difference between the places where escoltas are poorly represented compared to places with higher density of escoltas?

Doing some calculations, this questions is equivalent to the comparison of the rural and urban environments, so no further calculations were done.

Appendix B

Infographics design and final results

This appendix contains the initial design of the infographic layout and the final implementation after creating the charts and composing them into the layout.

B.1 Mockup of the initial design

See FigureB.1

B.2 Implementation of the design

The result created while implementing the created design can be seen in FigureB.3

B.3 Final version

To facilitate the visualization in a screen, a horizontal layout was finally chosen. Also, the total number of historical members was moved to the centre as element to catch the user attention. See FigureB.3

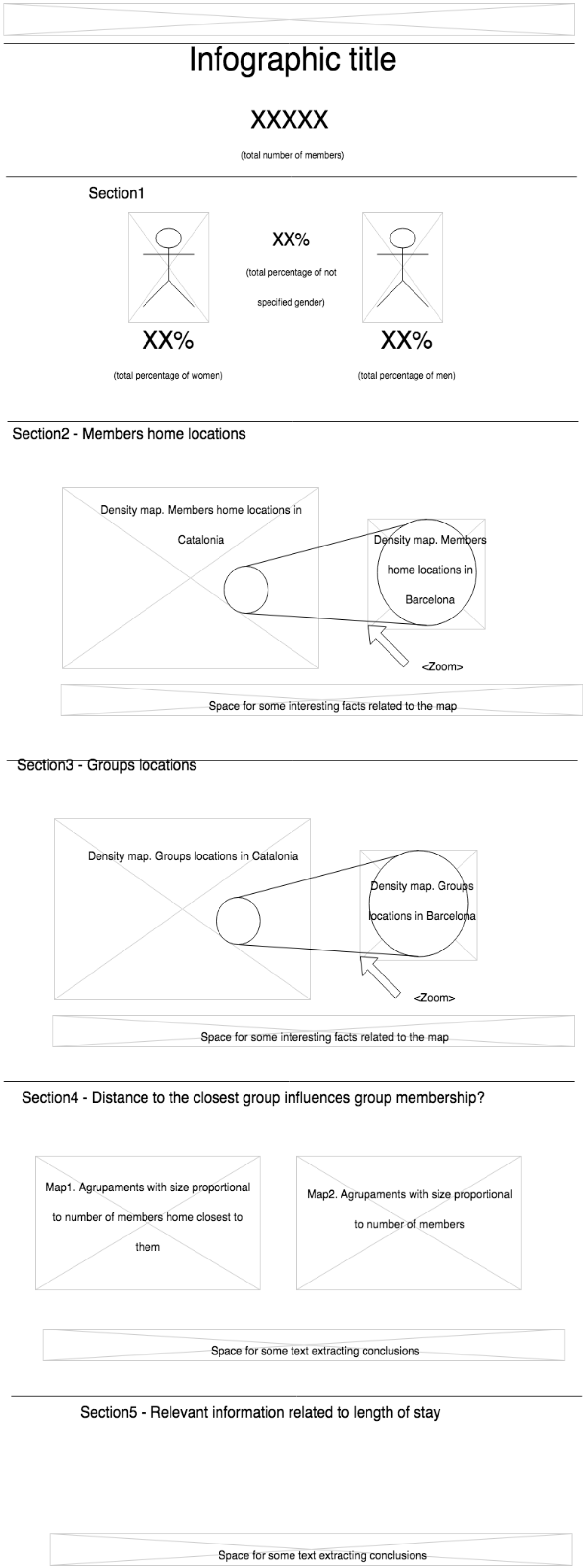
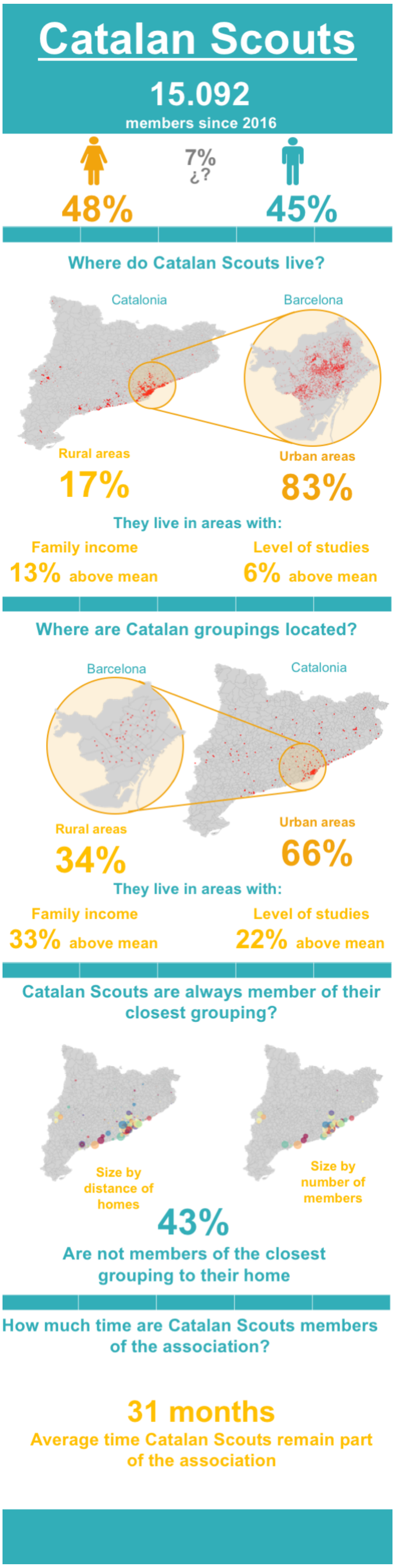

Infographic title
XXXXX
(total number of members)
Section1
XX%
(total percentage of not specified gender)
XX%
(total percentage of women)
XX%
(total percentage of men)
Section2 - Members home locations
Density map. Members home locations in Catalonia
Density map. Members home locations in Barcelona
<Zoom>
Space for some interesting facts related to the map
Section3 - Groups locations
Density map. Groups locations in Catalonia
Density map. Groups locations in Barcelona
<Zoom>
Space for some interesting facts related to the map
Section4 - Distance to the closest group influences group membership?
Map1. Agrupaments with size proportional to number of members home closest to them
Map2. Agrupaments with size proportional to number of members
Space for some text extracting conclusions
Section5 - Relevant information related to length of stay
Space for some text extracting conclusions

FIGURE B.1: Initial mock of the infographics design



Catalan Scouts

15.092

members since 2016

7%
¿?

48%

45%

Where do Catalan Scouts live?

Catalonia

Barcelona

Rural areas

17%

Urban areas

83%

They live in areas with:

Family income

13% above mean

Level of studies

6% above mean

Where are Catalan groupings located?

Barcelona

Catalonia

Rural areas

34%

Urban areas

66%

They live in areas with:

Family income

33% above mean

Level of studies

22% above mean

Catalan Scouts are always member of their closest grouping?

Size by distance of homes

Size by number of members

43%

Are not members of the closest grouping to their home

How much time are Catalan Scouts members of the association?

31 months

Average time Catalan Scouts remain part of the association

FIGURE B.2: Final version of the infographics

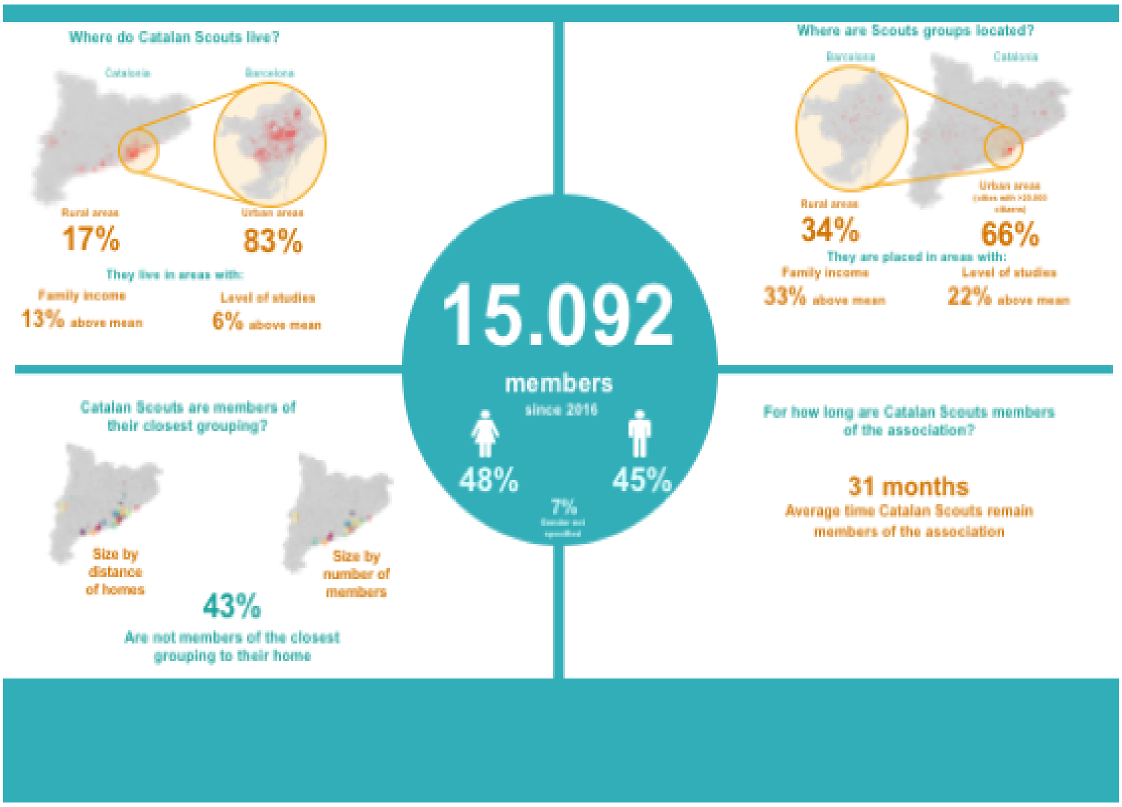


FIGURE B.3: Final version of the infographics