Budapest University of Technology and Economics
Faculty of Electrical Engineering and Informatics
Department of Computer Science and Information Theory

Comprehensive analysis of road traffic accidents in the city of Barcelona

Scientific Students' Association Report

Author:

Ákos Schneider

Advisor:

László Kabódi

2022

Contents

Kivonat **i**

Abstract **ii**

1 Introduction **1**

1.1 Project structure 1

1.2 Used tools 2

1.2.1 JetBrains DataSpell 2

1.2.2 Anaconda 2

1.2.3 Python libraries 2

1.2.3.1 pandas 3

1.2.3.2 googletrans 3

1.2.3.3 numpy 3

1.2.3.4 utm 3

1.2.3.5 scipy 3

1.2.4 Tableau 3

2 Data preparation **5**

2.1 Data collection 5

2.2 Data cleaning 8

2.2.1 Column switching 8

2.2.2 Incorrect hour value range 8

2.2.3 Missing latitude and longitude values 10

2.2.4 Type of day 10

2.2.5 Datasets in Catalan 10

2.2.6 Incorrect character coding 10

2.2.7 Data Imputation 10

2.2.8 Categorical column value set changing 11

2.2.9 *Unharmed* extent of injury values in 2020 11

2.3 Cleaned dataset dimensions . . . 11

3 Data exploration **12**

4 Data visualization with Tableau **18**

4.1 Connecting to data . . . 18
4.2 First worksheet . . . 21
4.3 First dashboard . . . 22
4.4 Interactive dashboard . . . 23
4.5 Using spatial files . . . 23
4.6 Comments . . . 24

5 Data analysis **25**

5.1 Demography . . . 25
5.1.1 Age composition in 2021 . . . 25
5.1.2 Education in 2021 . . . 27
5.1.3 Trends . . . 30
5.2 Accidents . . . 32
5.2.1 Time distributions of accidents in different districts . . . 33
5.2.2 Which district is the most dangerous? . . . 35
5.2.2.1 *Damage score* of accidents . . . 35
5.2.2.2 Normalizer, Normalized index . . . 36
5.2.3 In-depth analysis of different aspects and their combinations . . . 38
5.2.4 Difference in accident participation age . . . 44
5.2.4.1 Tableau calculated fields, bins, and parameters . . . 45
5.2.5 Biggest accidents . . . 46
5.2.6 Trends . . . 47
5.2.6.1 Tableau forecasting . . . 47

6 Conclusion **50**

6.1 Key findings . . . 50
6.2 Limitations and challenges . . . 51
6.3 Future research . . . 52

Bibliography **53**

Kivonat

Ma, az informatika korában, az adatok felhasználása a rendszerek fejlesztésére egyre fontosabbá válik. A technológiai fejlődés lehetővé tette, hogy az adatgyűjtés, -feldolgozás és -elemzés rendkívül hatékony legyen, és így egyre több és több felhasználási területet nyitott meg. Akár üzleti döntésekhez, akár tudományos kutatáshoz, az adatok valamilyen formában történő felhasználása elkerülhetetlen a mai világban. A közúti közlekedési balesetek elemzése betekintést ad egy város közlekedési infrastruktúrájának hibáiba, előrejelzése pedig megmutathat negatív trendeket. Ezen aspektusok figyelembevételével történő beavatkozás egy város úthálózatába csökkentheti az emberi áldozatok számát, biztonságosabbá téve a városban zajló közlekedést.

A projekt témája a közúti közlekedési balesetek elemzése Barcelona városában. A dolgozatban az OpenData BCN nevű nyílt adatportálról használok fel adathalmazokat, köztük demográfiai és közúti baleseti adatokat, amelyeket 2010 és 2021 között gyűjtöttek össze Barcelonában. A projekt első fázisa az adatok előkészítéséből és tisztításából áll, hogy az adathalmazok konzisztens állapotba kerüljenek. Ezt követi egy feltáró adatelemzési szakasz, ahol a közúti közlekedési balesetek kezdeti jellemzőit vizsgálom a Python Pandas könyvtár segítségével. Ezután a Tableau üzleti intelligencia szoftverrel interaktív grafikonokat és egyéb vizualizációkat készítek, amelyek segítségével részletes ismereteket tudok szerezni a balesetekről és Barcelona demográfiájáról. Ezeket a rálátásokat az utolsó fázisban átfogó adatelemzéssel nyerem ki, ahol az érdekes trendeket, korrelációkat és kiugró értékeket vizsgálom meg, objektív és szubjektív érvelést nyújtva. Konkrétabban részletekbe menően megvizsgálom Barcelona korösszetételét és iskolázottságát, valamint a közúti balesetek időbeli, térbeli és demográfiai eloszlását. Kitekintésként bemutatom a fent említett üzleti intelligencia szoftver előrejelzési képességeit is.

Azért választottam ezt a témát a projektemhez, mert ezt a területet tartom a legérdekesebbnek az összes informatikai terület közül, amellyel eddig találkoztam. Az összetett rendszerek elemzése, összefüggések és ok-okozati kapcsolatok keresése, lényegében a környezetünk működésének megválaszolása olyan téma, amelyben el tudok mélyedni. Az adatvizualizáció kiaknázása nem csak szemet gyönyörködtetőbbé teszi a munkámat, hanem segíti kreativitás és a művészet iránti rajongásom kibontakozását is.

Abstract

In today's age of information technology, the usage of data for the improvement of systems has become increasingly important. The technological advancement enabled data gathering, processing, and analysis to be highly efficient and therefore opened up more and more powerful use cases. Whether for business decisions or scientific research, the utilization of data in some form is inevitable in today's world. Analysis of road traffic accidents gives insight into the failures of a city's transport infrastructure and its prediction can forecast negative trends. Interventions in a city's road network, taking these aspects into account, can reduce the number of human casualties, making transport in the city safer.

The topic of the project is the analysis of road traffic accidents in the city of Barcelona. The starting point of the paper is the gathering of datasets from an open data portal called OpenData BCN, including demographical and road traffic accident data, collected between the years 2010 and 2021. The first phase of the project comprises data preparation and cleaning to reach a consistent state of datasets. An exploratory data analysis phase follows, where I look at the initial characteristics of road traffic accidents using the Pandas library in Python. Next, I use a business intelligence tool called Tableau to make interactive dashboards and visualizations which allows me to get detailed insights into accidents and the demography of Barcelona. These insights are extracted in the last phase with comprehensive data analysis being carried out, where I look into interesting trends, correlations, and outliers, providing objective and subjective reasoning. More specifically, I look into the age composition and education of Barcelona as well as the time-, spatial-, and demographical distribution of road traffic accidents. As an outlook, I also demonstrate the forecasting capabilities of the business intelligence tool mentioned above.

I chose this topic for my project because I find this field the most intriguing of all the fields of information technology that I have encountered. Analyzing complex systems and looking for correlations and causations, essentially answering how our surroundings work is a topic that I can delve into. Mixing it with the aspect of data visualization is not only going to make my work more appealing to the eye but also let me indulge in creativity and allow my admiration for art to flourish.

Chapter 1

Introduction

"You can have data without information, but you cannot have information without data."

This quote by Daniel Keys Moran, an American computer programmer, and science fiction writer is a fundamental aspect of processes that work with data. Information can only be gained through data, it is essentially processed data. However, to gain valuable, relevant insights for our use case, we need to ask a question, about what it is that we want to answer with the information. We need to process the data in a way, that our questions are answered quickly if not immediately. It is important to assess what could be valuable in a dataset, and what is irrelevant for the use case. Sometimes the data might be inadequate, but it is still better to use than using none at all. The representation of information or data can help people understand the underlying aspects better, and makes the analysis quicker. Data visualization is a form of representation, that makes presentations of analyses more intuitive as well.

I spent one year abroad as an Erasmus+ exchange student in Barcelona, which inspired me to choose the city as the subject of analysis. The council of the city of Barcelona has an OpenData service which they describe as the following: "Open Data or Public Sector Information Openness is a movement driven by public administrations with the main objective of maximizing available public resources, exposing the information generated or guarded by public bodies, allowing its access and use for the common good and for the benefit of anyone and any entity interested."[1] The project started in 2010, the year from which most of the datasets collection gathering was initiated. I used this service to browse through the 548 categorized datasets and I will be using the ones about the demography and road traffic accidents. From now on, I will refer to road traffic accidents either as RTA or simply accidents.

1.1 Project structure

From the two categories mentioned above, I will be using three and two datasets respectively. The two datasets from demography contain data on the population by age and the population by academic level from 2010 to 2021. When it comes to accidents, the three datasets contain data on accidents by people involved, accidents by cause from 2010 to 2021, and accidents by severity from 2016 to 2021. On top of this, I will be using datasets containing data on areas and registered vehicles of different districts. All of the datasets

[1] `https://opendata-ajuntament.barcelona.cat/en/open-data-bcn`

are stored in comma-separated text files, one file for every year in each category. In most of the datasets, there are coding errors, differences in column names, -counts, value inconsistencies, and so on. For this reason, the datasets need to be prepared and cleaned for later use. This will be the first phase of the project. After the data is in a consistent format it can be used as an input for a data visualization and business intelligence software called Tableau. With this software, I will create visualizations of the demography and accidents of Barcelona, more specifically interactive dashboards. This will be the second phase of the project. When the dashboards are created, I will carry out an initial exploratory data analysis and a thorough analysis of interesting trends, correlations, and outliers, providing objective and subjective reasoning. The final part of my project will be focusing on the predictive aspects of Tableau.

1.2 Used tools

There were several different tools and software that helped me complete my project. In the sections below, I provide the complete list of tools used during the completion of the project.

1.2.1 JetBrains DataSpell

DataSpell is an IDE developed by JetBrains and released in 2021. It is specifically designed for those involved in exploratory data analysis, and it combines the interactivity of Jupyter notebooks with the intelligent Python and R coding assistance of PyCharm in one ergonomic environment [9]. I will be using DataSpell 2022.1.2 for the data preparation and cleaning phase.

1.2.2 Anaconda

Anaconda is a distribution of the Python (and R) programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. Anaconda distribution comes with over 250 packages automatically installed, and over 7,500 additional open-source packages can be installed from PyPI as well as the conda package and virtual environment manager [21]. To execute notebooks or Python code in DataSpell, it is needed to configure at least one virtual environment based on a Python interpreter. When attaching a directory to the DataSpell workspace, a virtual environment is created by the IDE. The environment I will be using is Conda, based on the Anaconda installation, however, any other Python environment can be chosen. More specifically conda 4.13.0, and Python 3.9.12 will be used.

1.2.3 Python libraries

Throughout the data preparation phase, I will use the previously mentioned environment, writing code in Python notebooks. To provide a solution, some third-party packages will be needed to manage the datasets. I will introduce the used packages in the following sections.

1.2.3.1 pandas

Pandas is a software library written for the Python and R programming languages for data manipulation and analysis [16]. More specifically, it provides data structures and operations for manipulating numerical tables and time series. Perhaps the DataFrame object is the most notable feature of the library, this is the object used to represent tabular data and is used for data manipulation. Pandas allows importing data from various file formats such as comma-separated values (csv), JSON, Parquet, SQL database tables or queries, and Microsoft Excel, as well as exporting data to said file formats. The most notable operations supported on DataFrame objects are merging, reshaping, selecting, as well as data cleaning features. This is the library that I will use the most features from, as I will be manipulating, and changing the datasets of my use case.

1.2.3.2 googletrans

The datasets I will be using as I said before are from the council of Barcelona, and the language of all of them is Catalan, which I want to translate into English. Googletrans is a python library that implements the Google Translate API and can be used for this purpose. It uses the same servers as `translate.google.com`, has built-in language detection, and can provide translations in bulk [8].

1.2.3.3 numpy

NumPy is an essential package for scientific computing in Python [12]. It is a library that provides several objects such as multidimensional array objects, masked arrays, or matrices; with fast operations on said objects including mathematical and logical operations. To name a few: sorting, selecting, discrete Fourier transforms, basic linear algebra, basic statistical operations, and random simulations. I will only be using a few features of NumPy during the completion of the project.

1.2.3.4 utm

Utm is a python library to convert bidirectionally between UTM and WGS84 standards [3]. I will need this library as before 2016, the location of accidents was only given in the UTM standard, however, I will need Longitude and Latitude values for locations for Tableau.

1.2.3.5 scipy

SciPy is a free and open-source Python library used for scientific computing and technical computing [14]. It contains modules for optimization, linear algebra, special functions, FFT, signal and image processing, and other tasks common in science and engineering. I will be using this package in section 5.2.2, calculating the Pearson correlation coefficient of certain metrics.

1.2.4 Tableau

Tableau Software is a US-based company focused on providing business intelligence tools such as interactive data visualization software. The company has multiple versions of its

platform, both licensed, and free versions. I will be using the free version of their platform called *Tableau Public* [17]. The software can be injected with data from multiple data sources such as relational databases, text files, spreadsheets, and cloud databases. The injected data is used to generate rich graph-type data visualizations. It has a mapping functionality, it can plot latitude and longitude coordinates and also connect to spatial files. It has built-in geo-coding which allows administrative places (states, districts, postal codes) to be mapped automatically. I will dive deeper into the functionalities of the software in the *Data visualization* chapter, where I will show interactive Tableau dashboards.

Chapter 2

Data preparation

2.1 Data collection

The data was collected from the sources shown in table 2.7, all from the OpenData website of Barcelona. In each dataset category I chose, a separate csv file contained data records for each year. The starting point was the *People involved in accidents managed by the Police in the city of Barcelona* dataset, which I will refer to as *accidents by people* from now on. I decided to include the *Description of the accidents handled by the police in the city of Barcelona causality* dataset into the collection, as I thought it would be interesting to investigate further into accident causes. I will refer to this dataset as *accidents by cause* from now on. I also thought it would be a great idea to compare the age distribution in accidents with the actual population age distribution, as well as, look into academic levels of certain districts, and their possible correlation to accidents. Hence came two population datasets, *Population of the city of Barcelona according to sex and year of registration in the Register* and *Population of the city of Barcelona by sex and academic level*, which I will refer to as *population by age* and *population by academic level*. Lastly, I thought it would be good not only to group accidents by people involved but also individually, showing the number of casualties, their severity, and the number of vehicles involved. The *Accidents managed by the local police in the city of Barcelona* dataset contains all this information, which I will refer to as *accidents by severity*. During the analysis phase, a lot of different ideas came to my mind. I figured ranking districts based on their risk of accidents would be a great idea, as that could be a potential sign of infrastructure flaws. Using only the number of accidents would not be enough, as there are several different factors in play that can determine and influence this number. Some factors are listed in section 5.2.2. After browsing through the dataset portal, I have found two categories that are good to normalize accident numbers with. These are *Register of vehicles, evolution of the vehicle fleet in the city of Barcelona* (from now on *registered vehicles*) and *Neighborhoods' area size of the city of Barcelona* (from now on *area*). The population of districts can also be used to normalize the accident numbers. I have developed a formula for the normalization that will be described in the analysis phase later on. In the following, I will describe the cleaned datasets in detail. After the description, I will provide insights into data cleaning and how I got to the cleaned datasets.

Column	Type of field	Description
File number	Numerical	File number of the accident that the person of the data record was involved in
District code	Numerical	Code of district the accident occurred in
District name	Categorical	Name of district the accident occurred in
Neighborhood code	Numerical	Code of neighborhood the accident occurred in
Neighborhood name	Categorical	Name of neighborhood the accident occurred in
Day of week	Categorical	Name of the day the accident occurred in
Type of day	Categorical	Workday or weekend
Date	Date	Date of the accident the person was involved in
Part of the day	Categorical	Morning/Afternoon/Night
Hour	Numerical	Hour the accident occurred in
Type of vehicle involved	Categorical	Vehicle the person was using (if not pedestrian)
Gender	Categorical	Gender of the person involved in the accident
Pedestrian's fault	Categorical	Indicates if the pedestrian was at fault, and what the fault was
Type of person	Categorical	Driver/Passanger/Pedestrian
Age	Numerical	Age of the person involved in the accident
Extent of injury	Categorical	Injury severity (serious, slight, fatal)
UTM-X	Numerical	UTM-X location of the accident
UTM-Y	Numerical	UTM-Y location of the accident
Longitude	Numerical	Longitude location of the accident
Latitude	Numerical	Latitude location of the accident

Table 2.1: Accidents by people column descriptions

Column	Type of field	Description
File number	Numerical	File number of the accident
Description of cause	Categorical	Description of cause of the accident

Table 2.2: Accidents by cause column descriptions

Column	Type of field	Description
File number	Numerical	File number of the accident that the person of the data record was involved in
Number of deaths	Numerical	Number of fatal casualties in the accident
Number of people slightly injured	Numerical	Self-explanatory
Number of people seriously injured	Numerical	Self-explanatory
Number of victims	Numerical	Number of fatal-, slight-, serious casualties in the accident
Number of vehicles involved	Numerical	Self-explanatory

Table 2.3: Accidents by severity column descriptions

Column	Type of field	Description
Year	Numerical	Year of data recording
District code	Numerical	Code of district of population
District name	Categorical	Name of district of population
Neighborhood code	Numerical	Code of neighborhood of population
Neighborhood name	Categorical	Name of neighborhood of population
Gender	Categorical	Gender of population
Age	Numerical	Age of population
Number of inhabitants	Numerical	Number of inhabitants with the criteria's listed above

Table 2.4: Population by age column description

Column	Type of field	Description
Year	Numerical	Year of population
District code	Numerical	Code of district of population
District name	Categorical	Name of district of population
Neighborhood code	Numerical	Code of neighborhood of population
Neighborhood name	Categorical	Name of neighborhood of population
Gender	Categorical	Gender of population
Academic level	Numerical	Academic level of population
Number of inhabitants	Numerical	Number of inhabitants with the criteria's listed above

Table 2.5: Population by academic level column description

Column	Type of field	Description
Year	Numerical	Year of metrics.
District name	Categorical	Name of district of metrics.
District code	Numerical	Code of district of metrics.
Population	Numerical	Population scaled between 1 and 2.
Area(ha)	Numerical	Area scaled between 1 and 2.
Nr_vehicles	Numerical	Number of vehicles scaled between 1 and 2.
Nr_accidents	Numerical	Number of accidents scaled between 1 and 2.
Total_severity	Numerical	Total severity aggregated taking into account the extent of injury of accident participants.
Pop_density	Numerical	Population density calculated from population and area and then scaled between 1 and 2.
Normalized_index	Numerical	The metric used to rank districts.
Normalizer	Numerical	The metric used to denominate the total severity.

Table 2.6: Normalization dataset column description

2.2 Data cleaning

There are six different target datasets, *Accidents by people, Accidents by cause, Accidents by severity, Population by age, Population by academic level, and Normalization parameters*. I will use these datasets for visualizations, analysis, and predictions. From table 2.1 up until 2.6 the column descriptions of said datasets can be seen. These are the final schemas after the cleaning process. The dataset in table 2.6 requires some additional explanation, which I will provide in section 5.2.2.

The first and most important dataset is the one containing the *Accidents by people*, seen in table 2.1. An important thing to know about the datasets containing accidents is that through the *File number* attribute they can be joined as *File number* is a unique identifier of a singular accident. It goes without saying that in *Accidents by people* File number is not unique, as multiple people can be involved in a single accident. In *Accidents by cause* and *Accidents by severity*, however, *File number* is unique, as each row in those datasets equal one accident.

As I said previously, each dataset category had separate csv files for each year, which I needed to merge to be able to handle them as one dataset. To merge the category years, I needed to have the same relational schema for each year. This was the source of most inconsistencies, as the number, name, and order of columns differed greatly throughout the years. There were problems with values too, instead of empty ones, the values were *Unkown*, which required some digging to figure out. Another problem with values was the inconsistent type of attributes. The reason for said inconsistencies is purely based on the data collection. I summed all the inconsistencies into table 2.7, where I provide a possible solution for each inconsistency. The provided solutions are written in short, and might not be intuitive right off the bat, for this reason, I will show the least trivial ones in detail in the following subsections. The data cleaning of each data category was carried out in separate Python notebooks using the tools mentioned in subsection 1.2.3.

2.2.1 Column switching

I defined a function for this issue, named *df_column_switch*, accepting a DataFrame object in question, and two column names to be switched as parameters. In the *accidents by people* dataset collection, from 2016 to 2021 the *Age* and *Extent of injury* columns were switched up compared to the previous years as well as the UTM-X and UTM-Y columns from 2010 to 2015.

2.2.2 Incorrect hour value range

In the *accidents by people* collection, between 2014 and 2015, the value range of the hour column is from 1 to 12, whereas in all the other years, the value range is from 0 to 23. Not only this, but the *Part of the day* column is missing, which means it is impossible to deduct the correct hour value. Unfortunately, there is no way to fix this, so I will preemptively avoid including these years in analyses where either the part of the day or hour is present.

Data	Source	Inconsistency/Challange	Solution
Accidents by people 2010-2021	OpenData BCN	Different column names across all the years.	Renaming columns.
		Not needed columns across all the years.	Deletion of columns.
		Column order switched up from 2016 to 2021.	Switching columns using a function.
		Hour range 1-12 instead of 0-23. Part of the day not present (2014, 2015).	Unable to fix.
		Longitude, latitude columns not present from 2010 to 2015.	Generate from UTM-X and UTM-Y values.
		Street name column not present in 2019.	Street code mapping from other years, column insertion.
		Type of day column not present in 2021.	Labour, weekend day mapping, column insertion.
		UTM-X, UTM-Y columns order switched up from 2010 to 2015.	Usage of function to switch columns.
		Dataset in Catalan language.	Translation using googletrans python library.
		Character coding incorrect in values for seven distinct columns.	Replacing with correct ones.
		Unknown values instead of empty ones.	Replace to empty, then clear rows with empty values.
		Inconsistent data types in numeric columns.	Cast to consistent numeric values.
		Categorical columns' set of values too broad.	Merging values.
Accidents by cause 2010-2021	OpenData BCN	Only two columns needed.	Drop all of the others.
		Dataset in Catalan language.	Translation using googletrans python library.
Population by age 2010-2021	OpenData BCN	Age column value not numerical.	Convert to numerical.
		Character coding incorrect in some columns.	Replacing with correct ones.
		Dataset in Catalan language.	Translation using googletrans python library.
Population by academic level 2010-2021	OpenData BCN	Character coding incorrect in some columns.	Replacing with correct ones.
		Dataset in Catalan language.	Translation using googletrans python library.
Accidents by severity 2016-2021	OpenData BCN	Only 6 relevant columns needed.	Drop all of the others.
Registered vehicles 2016-2021	OpenData BCN	Aggregation by district of registered vehicle numbers.	Saved relevant data to dictionary, for later use in normalization.
Area	OpenData BCN	Aggregation by district of areas in hectar.	Saved relevant data to dictionary, for later use in normalization.

Table 2.7: Datasets and challanges

2.2.3 Missing latitude and longitude values

In the *accidents by people* collection, between 2010 and 2015, only UTM-X and UTM-Y values are given, not latitude and longitude values. I decided to use the *utm* python library to transform between the two metrics. I added the two columns to the dataframe and the correct values to the years between 2010 and 2015.

2.2.4 Type of day

In the *accidents by people* collection, in 2021 the *Type of day* column was not present. I created the column and mapped *Day of week* values to 'Labour' or 'Weekend' values in the designated dataframes.

2.2.5 Datasets in Catalan

There are categorical fields that need to be translated from Catalan to English. Some examples of this are *Pedestrian's fault* and *Type of vehicle involved* columns. I used the *googletrans* python library to get translations for values, put the translations in a map, and replaced the Catalan values with English ones. There is a designated function on a dataframe object called replace, in which it is possible to pass a dictionary object and set the *inplace* parameter to true. This will replace all values in the whole dataframe that are keys of the passed dictionary with the values of the keys. Some translations were incorrect due to unknown reasons, those I had to map manually. The full list of columns that needed to be translated is the following from table 2.1: *Day of the week, Part of the day, Type of vehicle involved, Gender, Pedestrian's fault, Type of person, Extent of injury.* In the population data collections, only the *Gender* column needed to be translated, while in the *accidents by cause* collection the *Description of cause* column.

2.2.6 Incorrect character coding

The character codings in seven columns were incorrect in the *accidents by people* data collection. The most notable ones were in *District name* and *Neighborhood name* columns. The recommended IANA encoding of the Catalan language is windows-1252, I used that to find the right characters and replace values with the right encoding format.

2.2.7 Data Imputation

The collection of the data by the council of Barcelona was conducted in a way that they did not put empty values, instead referred to empty values with the *Unknown* text. In order to be able to use built-in function *isnull()* of the dataframe column I needed to replace it with *NAN* NumPy constant . I used the built-in *replace()* function of the dataframe to do this task. After that, I could observe how many empty values were in each column, and delete rows that contained empty values in sensitive cases.

I defined a function called *print_uniquevalues_and_emptyrate* that accepts two parameters, one is a dataframe object, and the other is a column name from said dataframe object. The purpose was to see the range of values in each column, check what percentage of total rows have empty values in the column passed in the parameters, and then delete rows containing empty values. For example from the *accidents by people* collection a total of $(0.18\% + 0.22\% + 0.034\% + 0.82\% + 1.15\% + 0.12\%) = 2.52\%$ of records were deleted.

With the deletion, the total number of records is 134945 across 11 years. The deletion of around 2.5% is a price that I was willing to pay for the consistency of the dataset and for now being able to do computations on the remaining valid values.

2.2.8 Categorical column value set changing

In the *accidents by people* collection, the *Extent of innjury* column contained 10 unique values, however, I only wanted 3: *Slight injury, Serious injury,* and *Death*. The previous, more detailed range of values contained information about the death time occurrence (right after the accident, or more than 24 hours after the accident), as well as hospitalization and so on. The reason for this was that after thorough research on papers on road traffic accidents, the range of accident severity was comprised of the above-mentioned three values in said papers [6].

2.2.9 *Unharmed* extent of injury values in 2020

In the year 2020, the data recording was slightly different compared to the rest of the observed years. Instead of the three major categories of *Death*, *Slight injury*, and *Serious injury*, there was an additional one, called *Unharmed*. It resulted in there being twice as many accident participants as the previous year, which clearly would have been a problem had I not deleted the corresponding rows. To maintain consistency throughout the years I got rid of rows containing *Unharmed* extent of injury values.

2.3 Cleaned dataset dimensions

Dataset	**Number of records**	**Number of data points**
Accident by people	∼135k	20
Accidents by cause	∼111k	2
Accidents by severity	∼54k	6
Population by age	∼174k	8
Population by academic level	∼10.5k	8
Normalization parameters	$\rightarrow 60$	11

Table 2.8: Cleaned dataset dimensions

In table 2.8 above, the dimensions of the used datasets are detailed. The reason for only 54 thousand data records in *Accidents by severity* is that I purposely selected data between 2016 and 2021.

Chapter 3

Data exploration

Exploratory data analysis (or EDA) is part of the data analysis process to summarize the main characteristics of datasets often using data visualization after data cleaning [22]. I will perform an exploratory data analysis, I will go through the most important attributes of each dataset related to accidents.

First I will observe time-related attributes such as year, month, day, and hour. After that, I will look at spatial aspects, namely districts and neighborhoods, and lastly the rest of the categorical attributes such as cause, the extent of injury, types of vehicles involved, and so on.

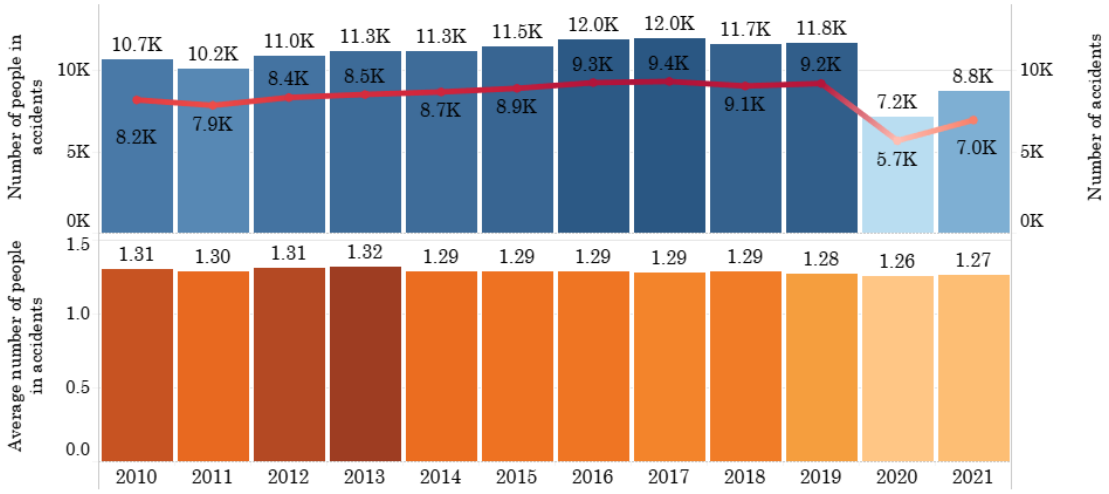

Figure 3.1: People in accidents by year

In the upper part of figure 3.1, the number of people involved in accidents can be seen with blue bars, while the number of accidents can be seen with the red line. On the bottom part, with orange, the average number of people involved in a single accident can be seen for each year. Right off the bat, we can see that from 2010 up until 2019, both the number of accidents and the number of people involved in accidents increased slightly. Then came 2020 with a huge drop-off in numbers and a slight "recovery" in 2021. The reason for this drop-off is the thing that has been a part of our lives for the last 2 years as of the time of writing this document, the COVID-19 pandemic. As for the average number of people involved in an accident, it has decreased slightly over the years. The lowest value was in 2020, understandably, more and more people traveled alone as a form of isolation during the pandemic.

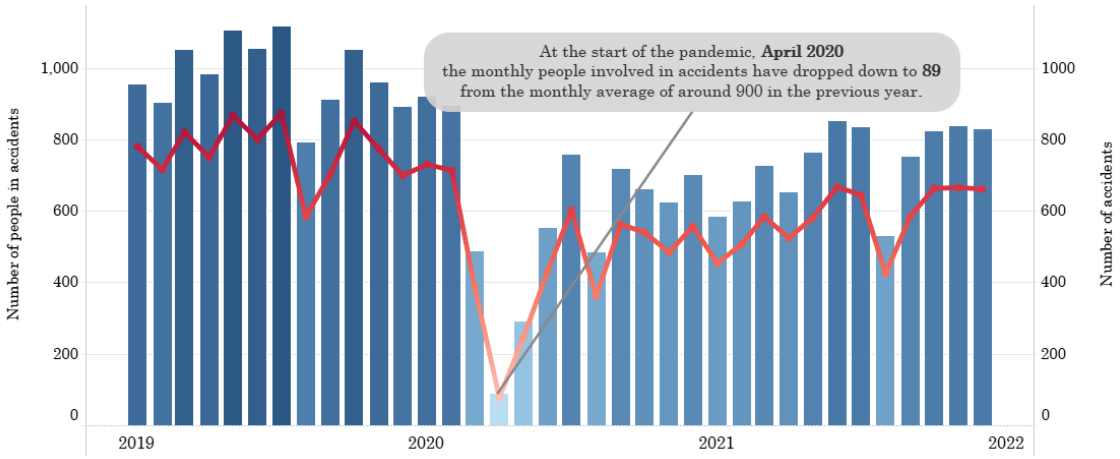

Figure 3.2: COVID-19 effect on accident numbers

In figure 3.2, one of the most important aspects of the pandemic can be observed. The government introduced mobility restrictions after the first outburst which resulted in less mobility, not just personal but vehicular. The fewer vehicles on the road, the fewer accidents happen.

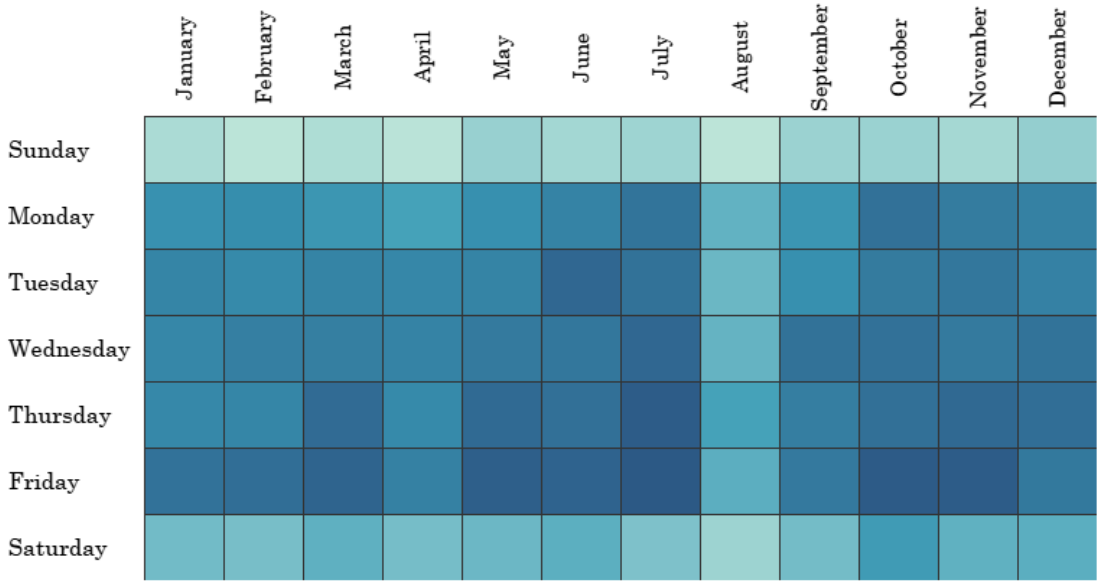

Figure 3.3: Month by day heatmap

The figure above shows the number of people involved in accidents on weekdays by month. From the heatmap in figure 3.3 two clear statements can be made: first, on the weekends, the number of people involved in accidents is much lower than compared on weekdays, second, in August, the number of people involved in accidents is lower than in other months. Friday is the day when most people are involved in accidents, and the strongest month in this regard is July. A possible reason for Fridays might be that there is more commuting than compared on other days: people visiting Barcelona for a weekend, or going home to the countryside can affect the commuting on Fridays. The three most busy months are June, July, and October, a possible reason is an increase in tourism during those periods.

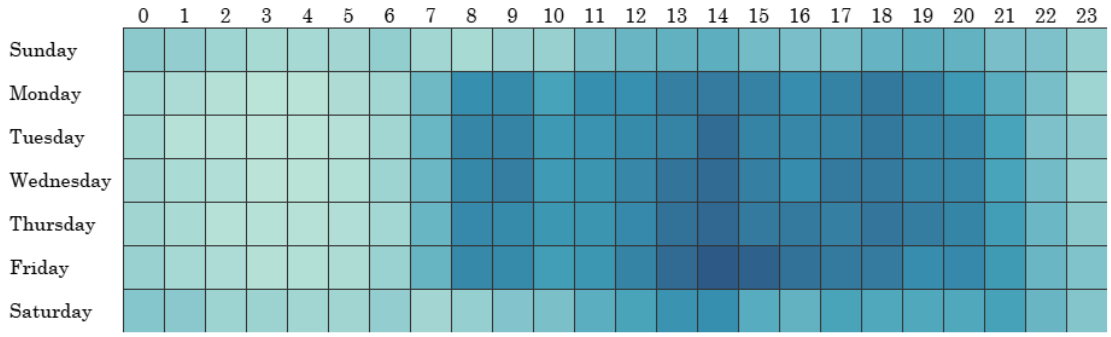

Figure 3.4: Hour

The figure above (figure 3.4[1]) tells the hour distribution of accidents throughout the seven days of a week. A darker rectangle is formed on the heatmap with period intersections of Monday to Friday, and 8 AM to 8 PM. This intuitively means that most of the accidents happen during that period. The absolute peak time is Friday 2 PM. Throughout the 11 years (2010-2021) 1509 people were involved on Fridays between 2 and 3 PM, while Tuesday at 3 AM produced the lowest number, 36 people involved in accidents.

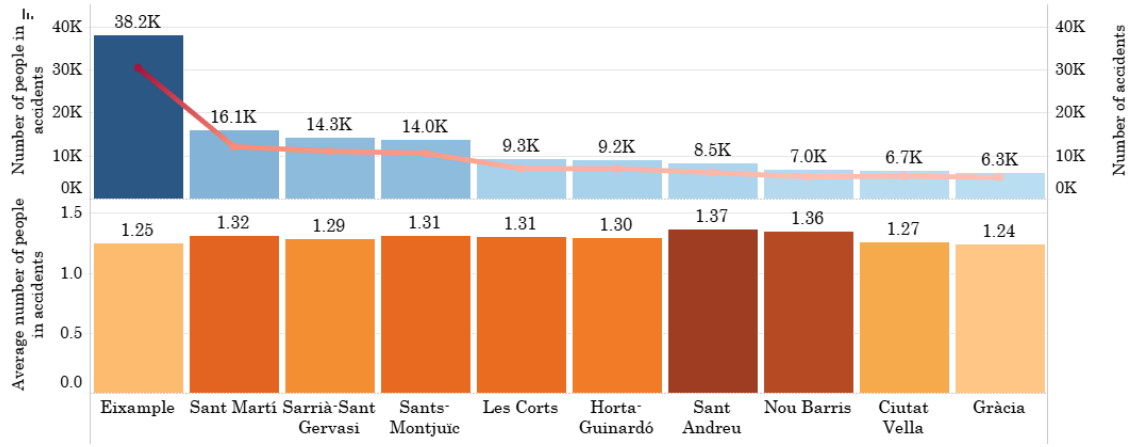

Figure 3.5: Districts

In figure 3.5 the Pareto distribution of accidents in districts can be seen. Eixample is the district with the most accidents, more than doubling the amount of the second district. Later on, in the project, I will provide a thorough analysis of these districts including demographic data. I will dive deeper into why in some districts the average number of people in accidents is higher than in others, but as of now, this graph is only sufficient to grasp initial information on districts.

When it comes to neighborhoods, I only plotted the top 10 neighborhoods with the highest accident numbers. Number one is la Dreta de l'Eixample, which is part of the Eixample district. Out of around 130 thousand people involved in accidents throughout the 2010-2021 time period, the top 10 neighborhoods account for 54 thousand. There are 73 neighborhoods and 10 districts in Barcelona.

In the following, I will show figures for the rest of the categorical features of datasets regarding accidents.

[1]2014 and 2015 data has been excluded for the reason explained in section 2.2.2.

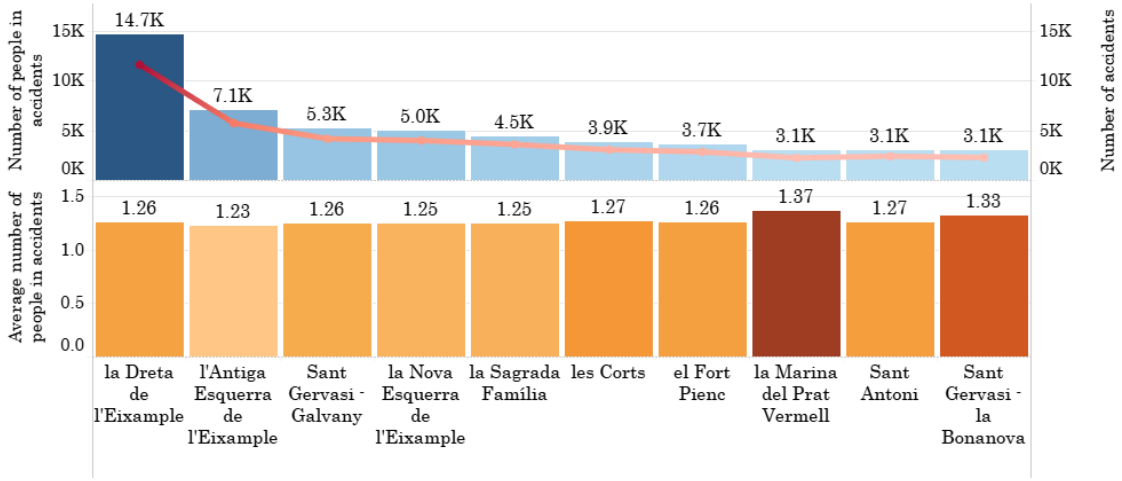

Figure 3.6: Neighborhoods

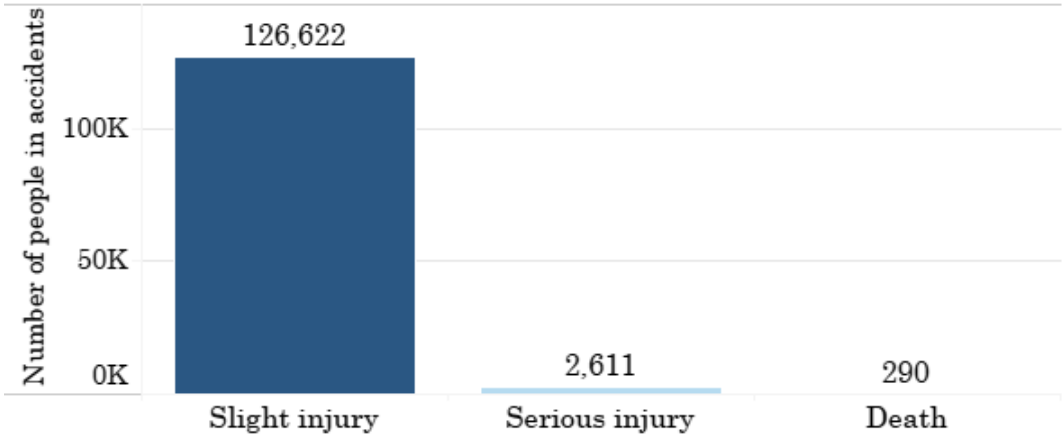

Figure 3.7: Extent of injury

In figure 3.7 it can be seen that the vast majority of people involved in an accident suffered only a slight injury. Throughout the 12 years, there were 290 fatal casualties of accidents.

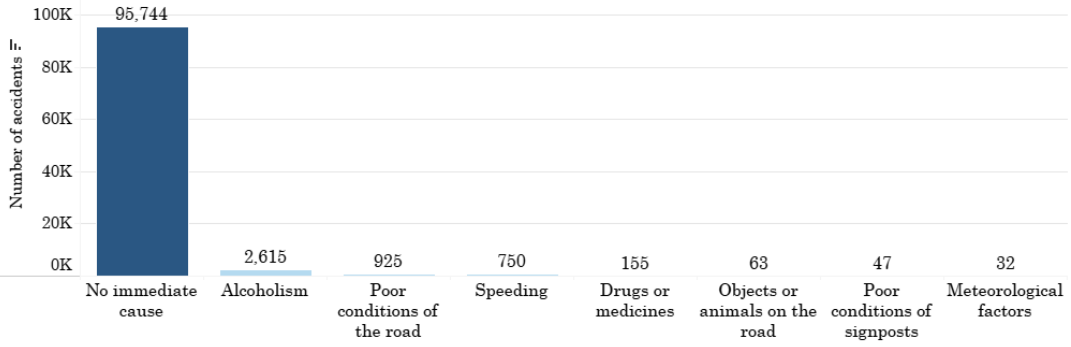

Figure 3.8: Cause of accidents

At the time of data recording in most cases, the authorities were unable to determine the direct cause of the accident. For this reason out of 100 thousand accidents, approximately 96 thousand were labeled with *No immediate cause* as can be seen in figure 3.8. When

the actual cause could be determined immediately, most accidents were a result of alcohol consumption. After that comes poor road conditions and speeding.

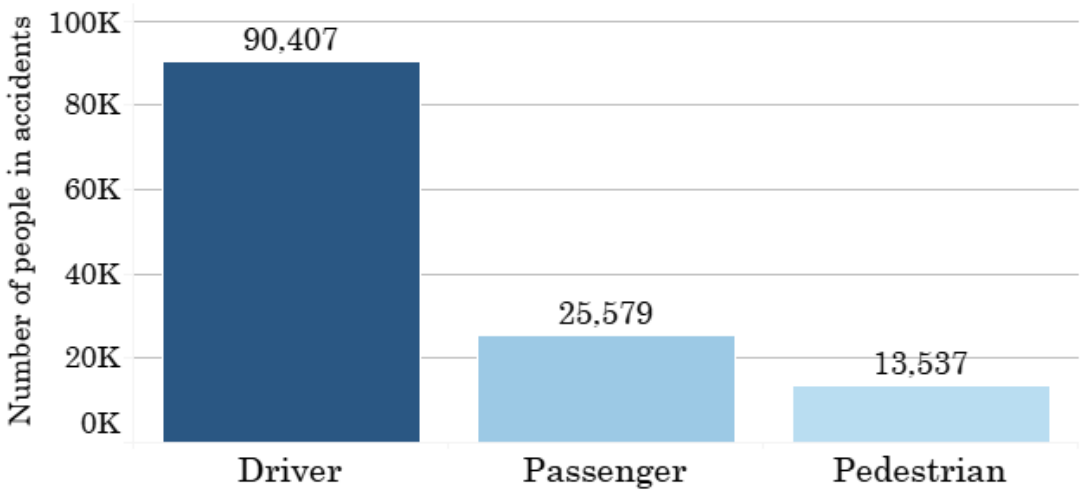

Figure 3.9: Type of person

In figure 3.9, the role of each person in an accident can be seen. Most of the people who participated in one, was the driver, while the smallest group are the pedestrians. In the following, I will show the categorical column that signals whether the pedestrian was at fault in the accident.

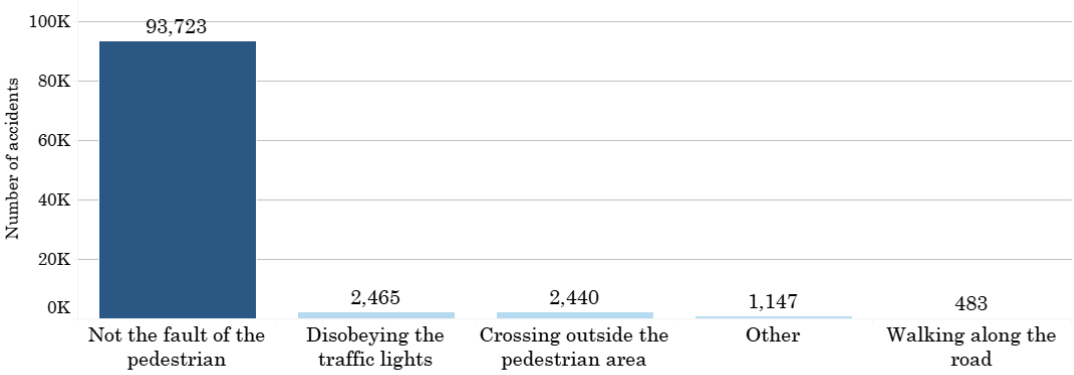

Figure 3.10: Fault of the pedestrian

As seen in figure 3.10, out of around 100 thousand accidents throughout the 12 years, 94 thousand were not a result of pedestrian negligence. As for the other amount of 6 thousand, disobedience of traffic lights as well as crossing the road outside pedestrian area were the most prominent causes.

There were several different vehicles listed, in a total of 15 types. I included the 10 vehicles that occurred most frequently in accidents. Most of the people that were involved in accidents were using motorcycles. This is not a surprise as it is the most used type of vehicle in the city of Barcelona. There are multiple reasons for its popularity; there are regulations in place for older cars, some of them are banned from entering the city because of the higher amounts of emission [13]; as a very flexible urban vehicle for use in the average short and medium distance journeys, there are more motorcycles in Barcelona

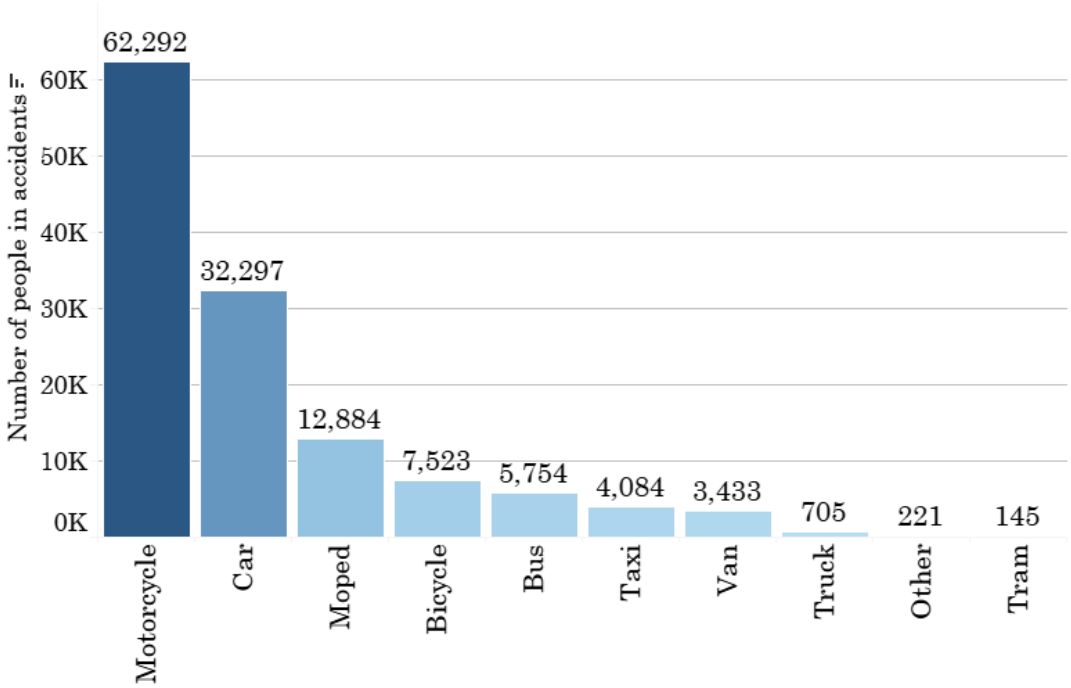

Figure 3.11: Type of vehicle

than cars [7]. Still, the second type of vehicle is the car, accounting for around one-fourth of all means of transport in this case.

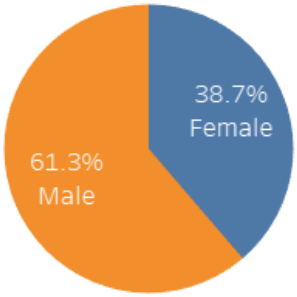

Figure 3.12: Gender

When it comes to gender distribution, around 61% of people involved in accidents are male and 39% are female.

Chapter 4

Data visualization with Tableau

As I mentioned in section 1.2.4, I will be using Tableau Public for data visualization, providing simple worksheets as well as highly interactive dashboards. I have already included many graphs in chapter 3 as part of the Exploratory data analysis. I will go through a quick example from scratch to showcase the usage of the software. I will introduce the most used features.

4.1 Connecting to data

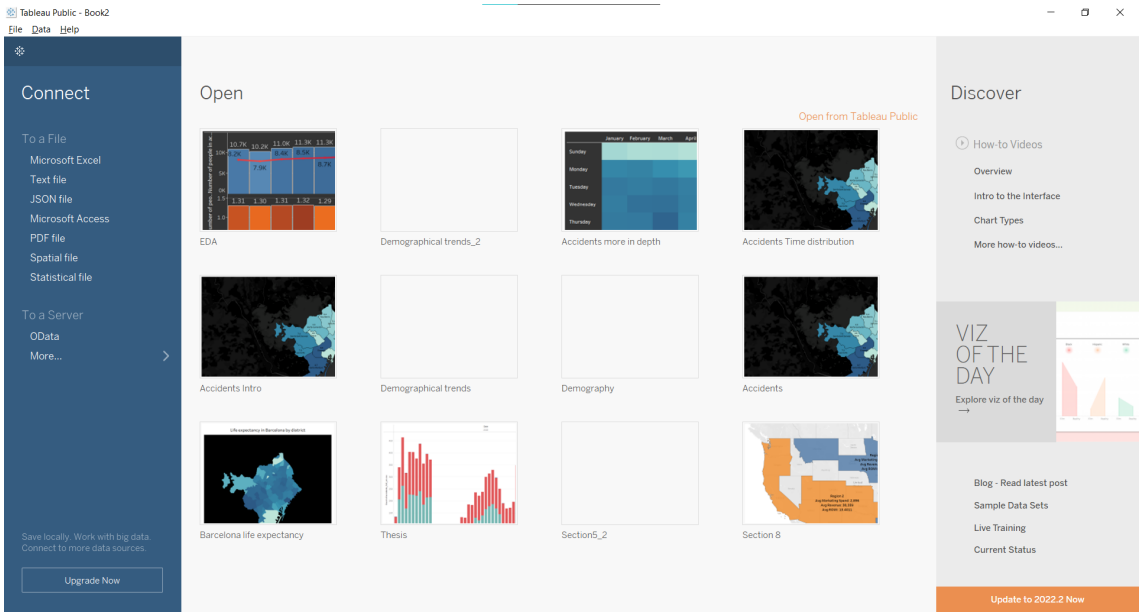

Figure 4.1: Tableau opening page

In figure 4.1, the opening page of Tableau Public can be seen. This is the page that pops up upon the first start of the software. Right off the bat, we can browse through previously created Tableau workbooks[1]. Looking at the left part of the picture there is a *Connect* pane. This can be used to connect a workbook to a data source. There are several different options to connect to a data source (file) including Microsoft Excel, Text

[1] "Tableau uses a workbook and sheet file structure, much like Microsoft Excel. A workbook contains sheets. A sheet can be a worksheet, a dashboard, or a story."[20]

files (csv), JSON files, Microsoft Access, Spatial files, and so on. In my projects, I only used csv and spatial files, which means Text and Spatial file data sources were the ones I needed to use.

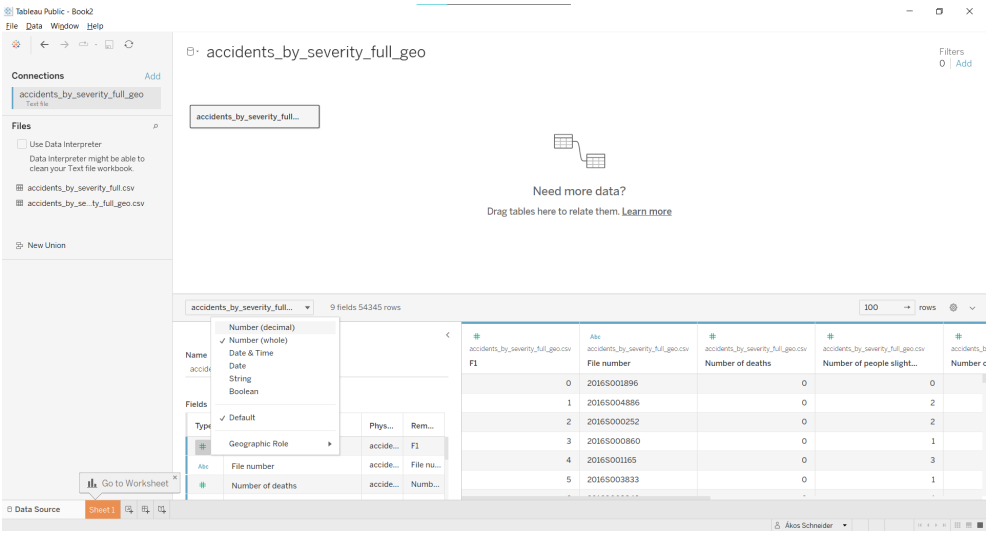

Figure 4.2: Tableau data source page

Right after connecting to a data source, the page in figure 4.2 pops up. We have the option to add different files to the already existing connection with the *Add* button. On the bottom part, we can see a preview of the data that was loaded in. A detailed description of fields is also present. Tableau loads the data and automatically assigns a data type to each field. We can modify that type to our liking, but the default behavior is often correct. We can assign geographical roles for more complex behavior as Tableau has built-in location detection for regions, area codes, countries, and so on. In the middle part of the page, the data can be seen as a tabular entity. Tableau calls this representation the *Logical layer*, where if we connect more tabular data, we can connect the two on a logical level based on a field. This is not a join, but a logical association, so that Tableau knows in the future that the two tables are closely tied.

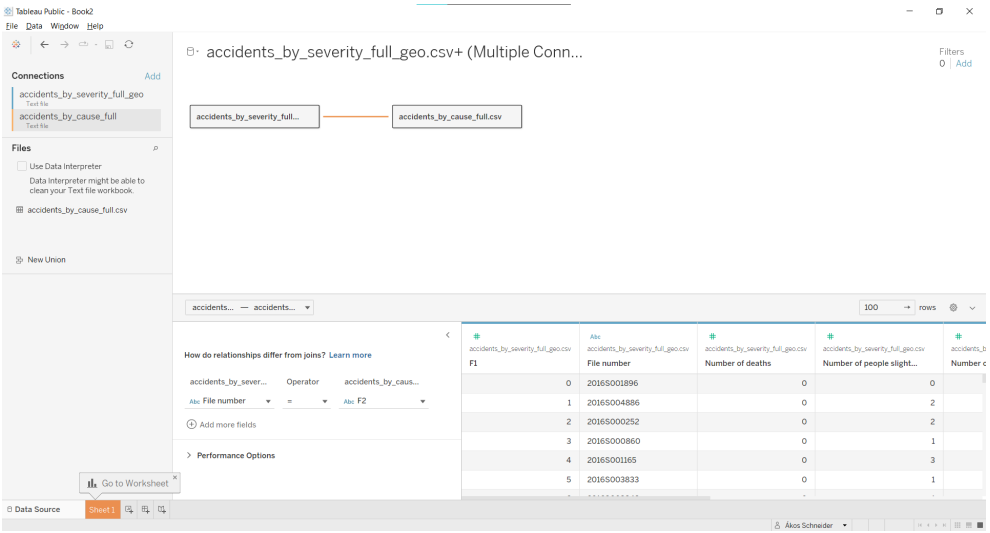

Figure 4.3: Tableau logical association

In figure 4.3, I selected another file as an additional connection and drag-and-dropped it to the middle which prompted a connection. I had to specify which fields I want to connect the two tables on, which ended up being the *File number* (it is called F2 in the second table). After this connection is specified, I can make visualizations using both tables at the same time as Tableau now knows how to connect them. Think of this relationship as a contract between two tables. When creating visualization with fields from these two tables, Tableau brings in data from these tables using that contract to build a query with the appropriate joins [19]. When double-clicking one of the two tables in the canvas, the so-called *Physical layer* appears.

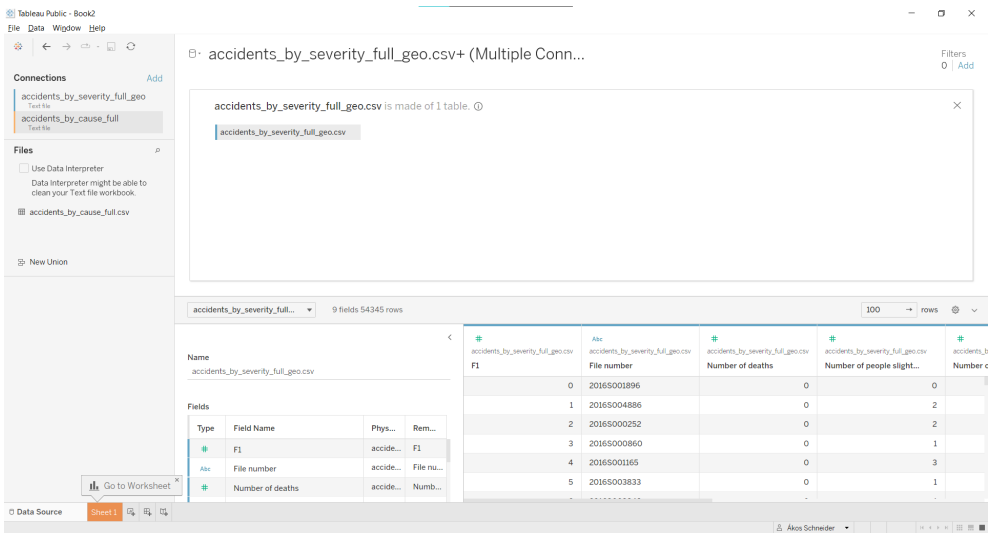

Figure 4.4: Tableau physical layer

This is the layer that can be used to perform joins. The same drag-and-drop method can be used to prompt a join on the two tables, where the user has to specify the fields the two tables should be joined on, and also the type of join to perform (left, right, inner, outer). Tableau recommends avoiding using joins and instead advises the use of logical relationships for better performance.

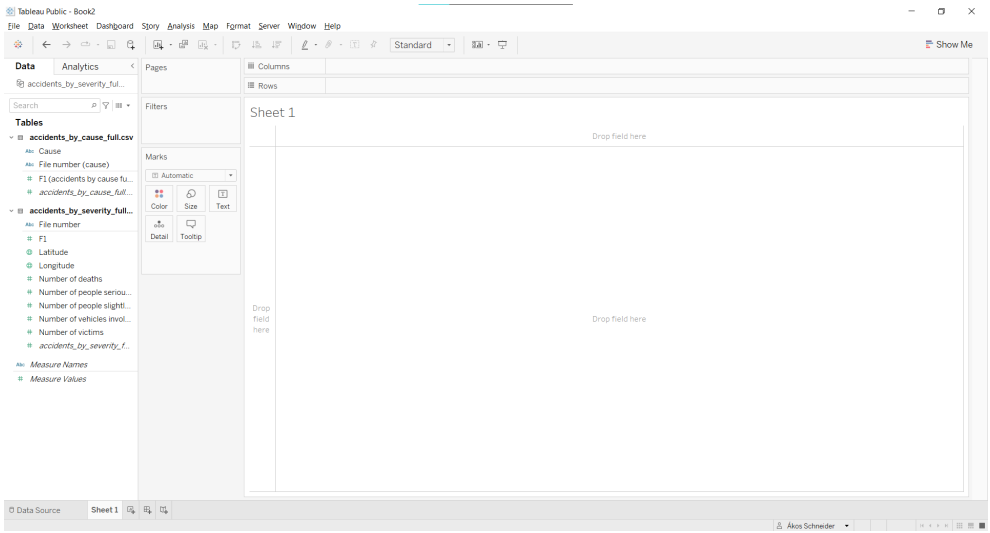

Figure 4.5: Tableau worksheet view

4.2 First worksheet

After the data connections have been specified and we are satisfied with the preview of data, it is possible to move on to creating visualizations by clicking on the *Go to worksheet* button or the sheet named *Sheet1*, which we can rename later on. A view as in figure 4.5 can be seen where the creation of worksheets can start. On the left, we can find all the fields of our data grouped into two categories: dimensions and measures. Dimensions are best described as categorical data, while measures are numerical. This type of field determines the visualizations, later on, however, we can always create a dimension out of a measure.

The visualizations can be done in a drag-and-drop fashion. Grabbing one of the fields and throwing it onto either the columns or rows will start the visualizations. After some practice, it becomes very intuitive and easy to use. In the *Marks* pane we can drag and drop fields as well. For example, if we drop the field *Cause* to color, it will intuitively create distinct colors for each Cause on the visualization. The color palette can be changed by clicking on the *Color* option. The same goes with *Size, Label, Detail,* and *Tooltip*. Also in the *Marks* pane, we can select the type of visualizations we want, however, there are some limitations depending on the number of columns and rows. These limitations are described under the *Show Me* dropdown.

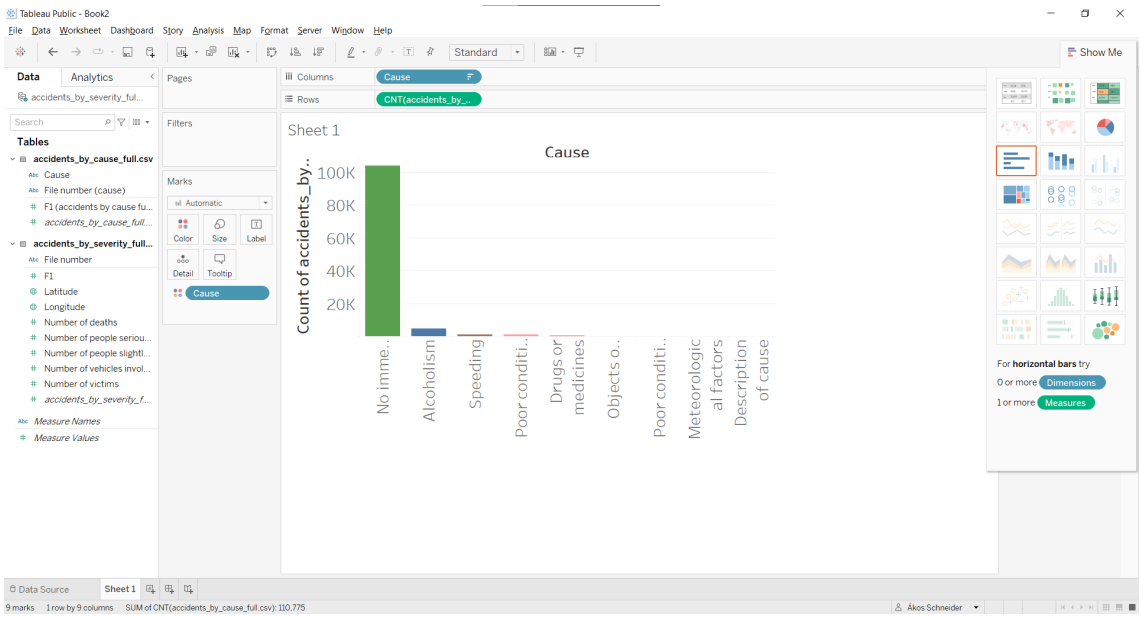

Figure 4.6: First worksheet

There is a wide variety of formatting options including worksheet-level font and size options; row, column alignment/shading/borders/lines. The view of graphs is highly customizable.

In the *Filters* pane, it is possible to filter the data presented on the visualization by any field. The filtering can be done by specifying values, by a condition, or by top (highest amount based on a field). These filters can be shown and formatted as per the need of the user. In figure 4.7 on the left of the graph can be seen a simple default filtering on causes of the accidents, where the option *No immediate cause* is filtered out.

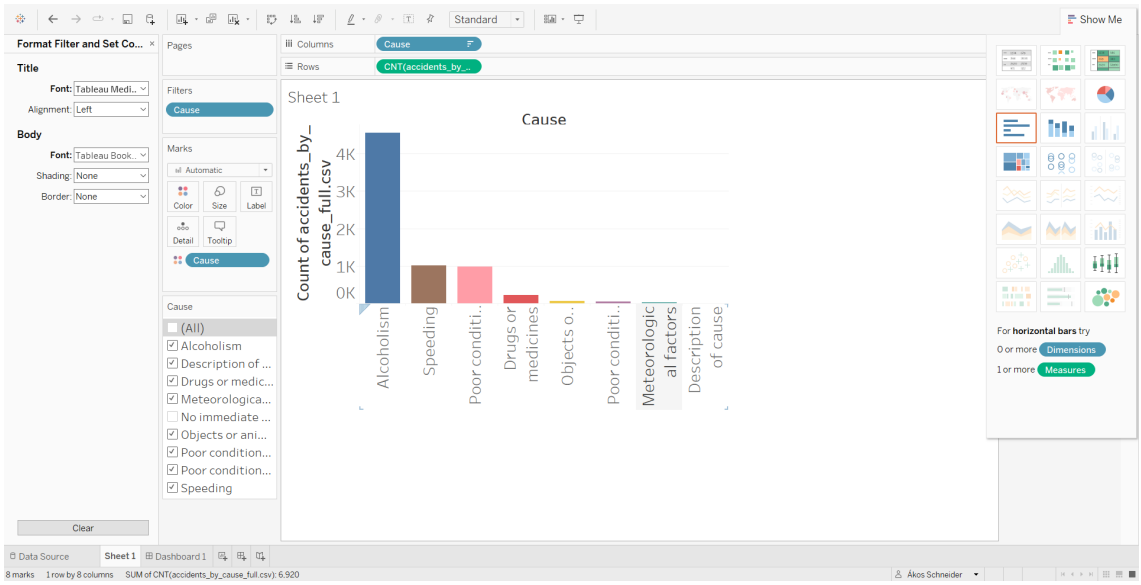

Figure 4.7: First worksheet with filter

4.3 First dashboard

A dashboard can be created by clicking the designated button on the bottom right next to the sheets listed. The layout and size can be set, there are various standard sizes for tablets, phones, and computers. The fixed pixel size can be set as well, allowing great flexibility.

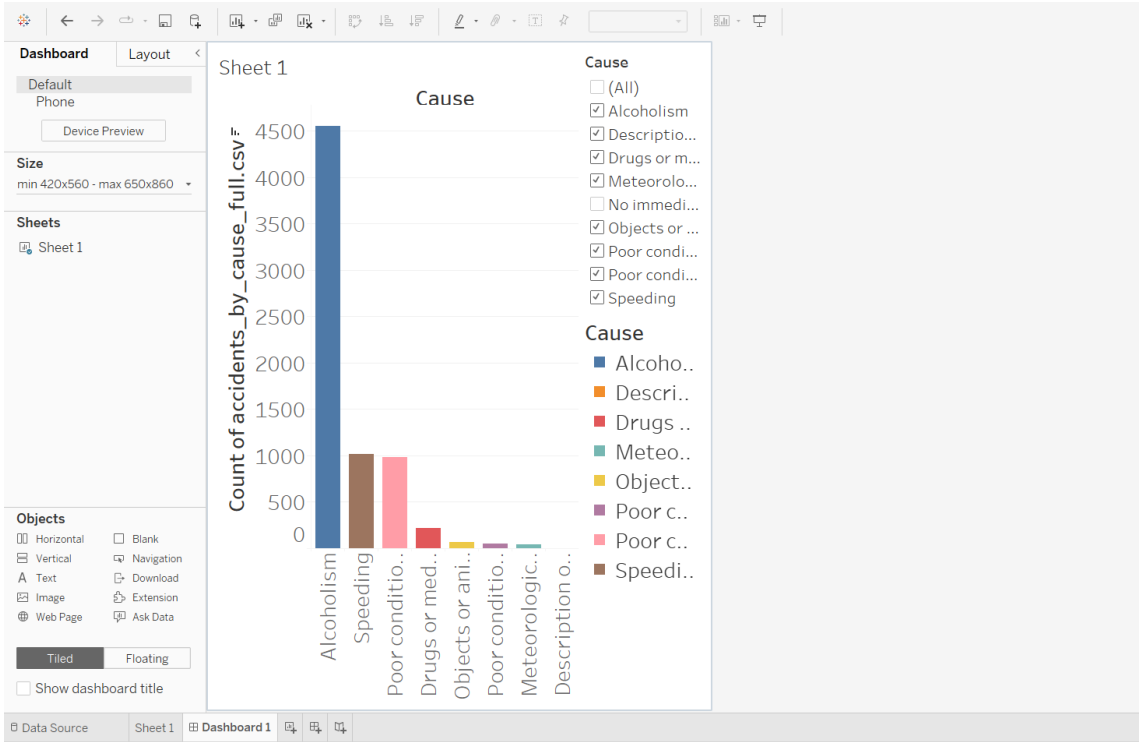

Figure 4.8: First dashboard

The created sheets will be listed on the left and can be dragged onto the dashboard either as Tiled or Floating objects. Tiled objects will fill the dashboard, while with floating

objects there is flexibility to choose the size and direct position of the object. Various other objects can be put onto the dashboard such as images, texts, extensions, and so on.

4.4 Interactive dashboard

The filter that was created in the worksheet is included when dropping the worksheet on the dashboard. When we have multiple worksheets on a dashboard, these filters can be selected to be applied to all worksheets on the dashboard. Right-clicking each worksheet on the dashboard, we can ask for it to be used as a filter. The default behavior is that clicking a part of the worksheet will filter it to the clicked value range. The above-mentioned quick filter creation is in essence a creation of *Actions*. Actions can be defined or modified in the *Dashboard/Actions* path where it is possible to specify all the details of the action. *Actions* have source and target sheets, where an action can be run by selection or hovering on sheets. When the action is fired, the filter is applied to the target sheets. With the usage of *Actions*, highly interactive dashboards can be created.

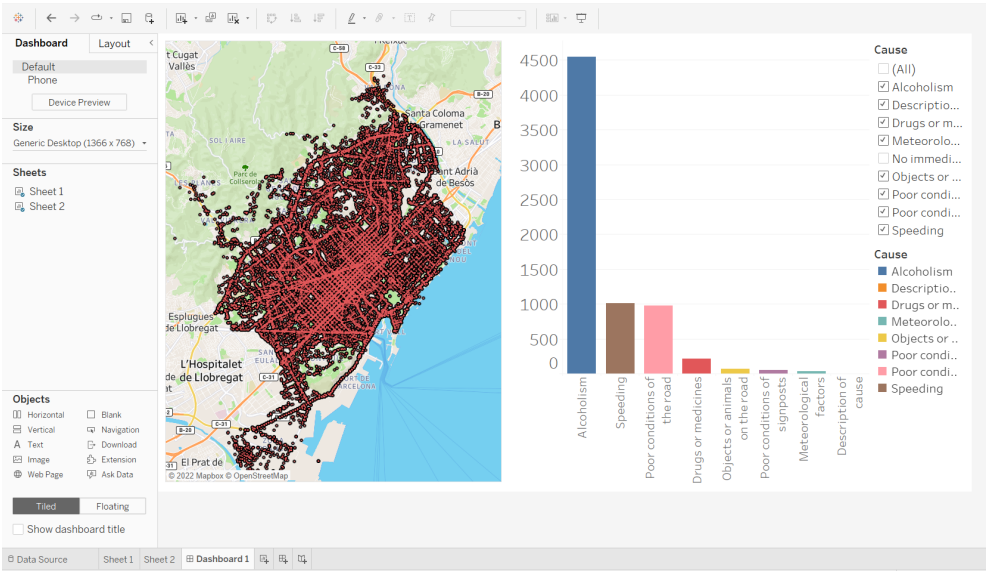

Figure 4.9: Formatted dashboard with two worksheets

In figure 4.9, I included two sheets, the first on the right is already introduced, the second shows the locations on a map of the accidents between 2016 and 2021. After a little formatting, the result is in the figure. This is just an example to showcase how quickly a complex interactive dashboard can be created in Tableau.

4.5 Using spatial files

Tableau has built-in visualizations for certain areas. Based on state names it can visualize the USA states for example with options for labeling, colorizing, and so on. There are some areas, however, that are not included. The districts of Barcelona are not built in, which meant I had to look for a spatial file showing the districts. After a while, I found a website[2] where a lot of different shapefiles of cities can be found. I chose the districts of Barcelona and downloaded them from the website in the shapefile format.

[2] `https://data.metabolismofcities.org/dashboards/barcelona/maps/567058/view/`

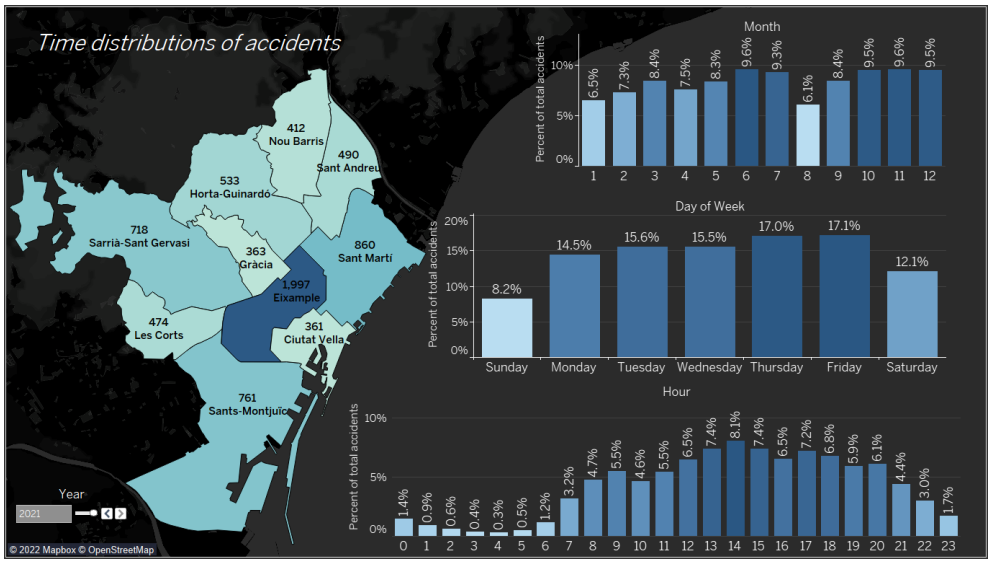

Figure 4.10: A dashboard containing a map using shapefiles

The dashboard above contains a map, which the previously mentioned shapefile made happen. In the data source setting page I needed to include the file just like any other source file, and connect the accident data on the district code field, which tells Tableau that the districts mentioned in the shapefile and the accident datasets are equivalent, making it possible to colorize them based on the number of accidents in each district or any other metric included in the dataset.

4.6 Comments

As I said previously, I will use this software for the visualizations and have already had the previous section include some histograms and heatmaps. This chapter was an introduction to how Tableau works, there are loads more features that I have not included as it is not the sole goal of the paper. In the future I will not get into detail about the creation of dashboards and worksheets, however, if I encounter difficulties and use important special features I will not hesitate to mention them.

Chapter 5

Data analysis

In this chapter, I will provide a comprehensive analysis of the demographics and accidents of the city of Barcelona using the visualizations created with Tableau. The analysis of both of these aspects will be structured similarly. First I will analyze the data of the latest year and its characteristics, then I will look at trends throughout the period between 2010 and 2021. However, a more detailed analysis will be carried out on accidents, as that is the main focus of the paper. Throughout the analysis, I will be looking for interesting outliers and connections and will provide subjective and objective explanations for them.

5.1 Demography

In the following subsection, I will show the dashboards and worksheets created in Tableau for the demographic analysis of Barcelona. It is important to know the demography because normalizing accident data with demographic data allows for more accurate analysis.

5.1.1 Age composition in 2021

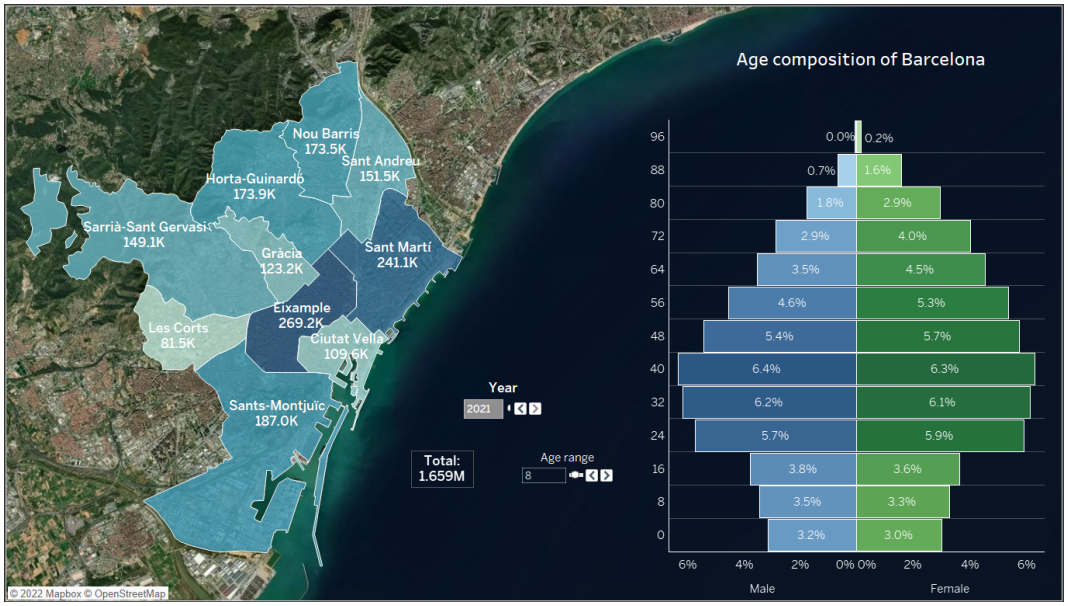

Figure 5.1: Age composition dashboard

The dashboard in figure 5.1 contains two main elements, on the left, a map showing the number of inhabitants in each district, and on the right, a population pyramid[1] showing the age distribution by gender. At first glance, I can tell that the population pyramid of the whole of Barcelona city is constrictive [4], meaning elderly and shrinking. This type of pyramid is distinctive in countries with higher levels of socioeconomic development, where access to quality education and health care is available to a large portion of the population. I will talk about the education of Barcelona in the following parts of this chapter.

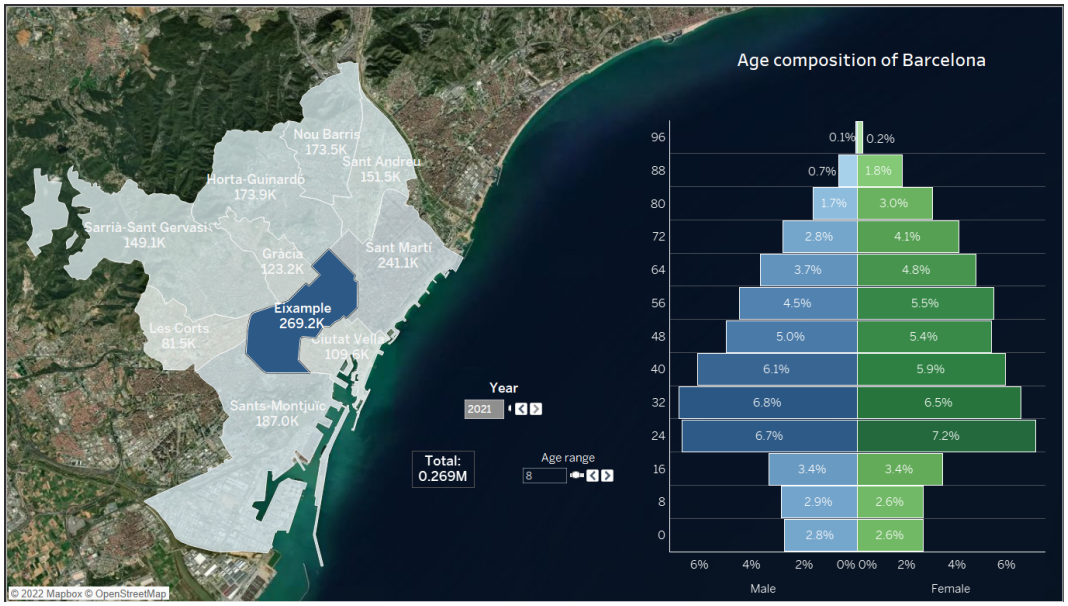

Figure 5.2: Age composition in Eixample

Looking at the picture above, the age composition of the Eixample district can be seen. It is clear, that the population pyramid of Eixample falls into the same category as the whole of Barcelona. I went on and looked at every district, trying to find something that would break the pattern, however, all of the districts fell into the constrictive category, with very few changes in the shape of the graph.

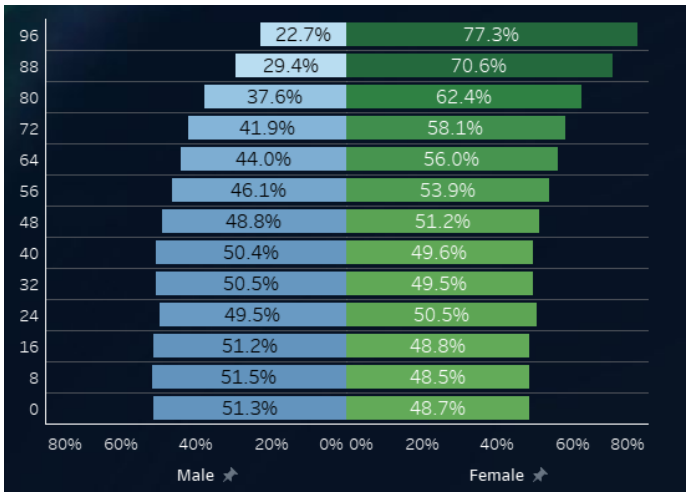

Figure 5.3: Age composition

[1] A simple graph that conveys the complex social narrative of a population through its shape.[4]

An interesting thing to mention here is that looking at the older age groups (64 and after) a significant difference in gender representation is shown. Women aged 64 or more are almost double the amount compared to men aged 64 or more. In figure 5.3, the gender distribution is shown in each age group. As we look at the older age groups, a significant tilt is showing, signaling that men die earlier than women, or in other words, the life expectancy of men is lower than that of women. More than 70% of people aged 88 or older are women. There is a recent study that claims to explain the reason behind it: men lose the male sex chromosome as they get older resulting in scarring in the heart muscle that can increase the chance of a fatal heart failure [11].

5.1.2 Education in 2021

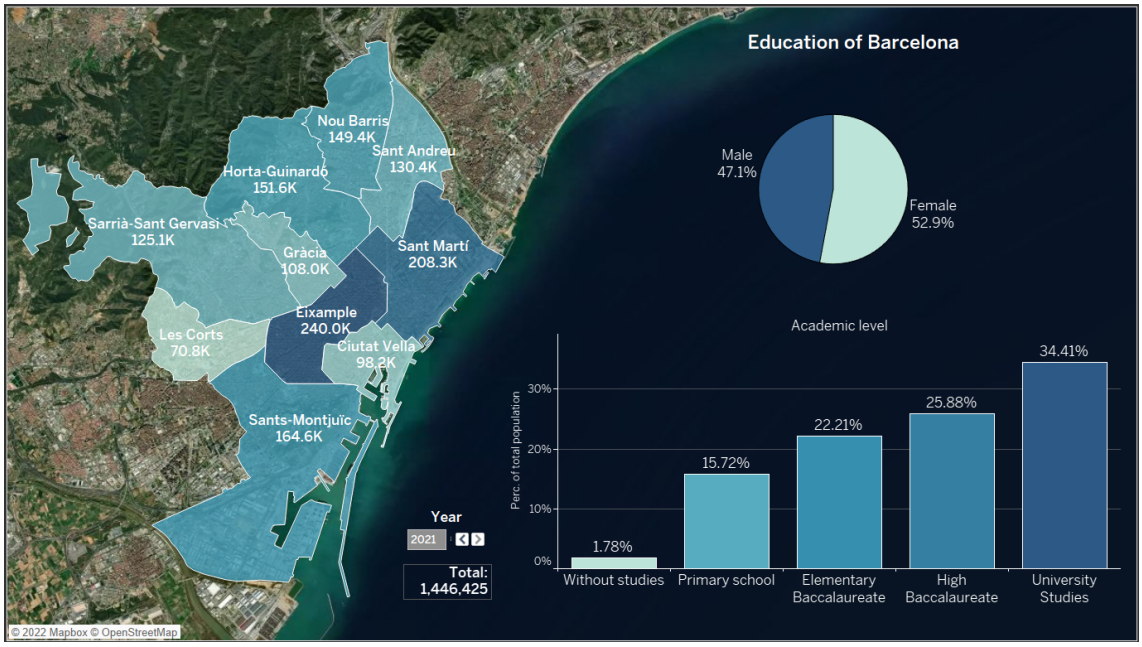

Figure 5.4: Academic distribution

The data for education applies to people aged 16 or older. One-third of the residents have a university degree, whilst less than 2% of residents have had no education at all in their lives as per figure 5.4.

Looking at different districts' education, there are noticeable differences. Sarrià-Sant Gervasi is the district with the best education numbers, with around 54% of people having a university degree whilst 0.5% of people have had no education at all. Eixample, Gràcia and Les Corts are the following districts with great numbers: around 45% of people with a university degree and 1% of people with no education at all. These top four districts can be clustered into better-performing districts. The worst performing district by far is Nou Barris, where only 16% of people have a university degree and 3.5% have no education at all. The majority (around one-third) of people only completed the Elementary Baccalaureate, which can be finished by the age of 18. Sant Andreu is the second worst district and can be clustered with Nou Barris. All the rest of the districts are following the patterns of the whole of Barcelona in terms of education and can be put into the moderate educational cluster. These three clusters can be seen in figure 5.5.

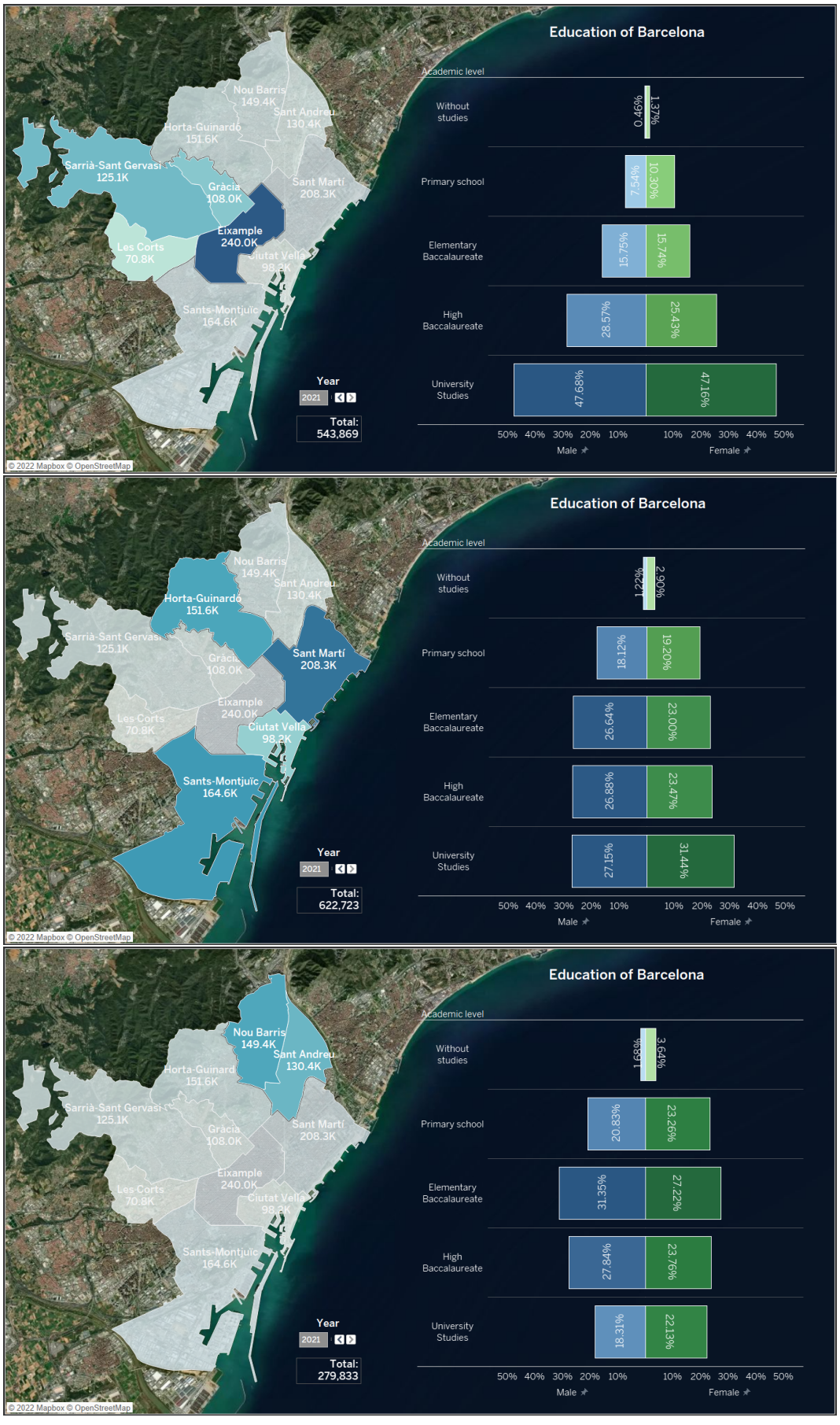

Figure 5.5: Academic clustering of districts

Taking a closer look at figure 5.4, it is clear that 52.9% of people aged 16 or older are female. Using the capabilities of the highly interactive dashboard, I looked at each academic category and saw that out of the people in *Primary school* and *University studies*, the ratio of females was 55.3%, which means compared to the total age distribution, there are proportionally more women in these academic categories (figure 5.6).

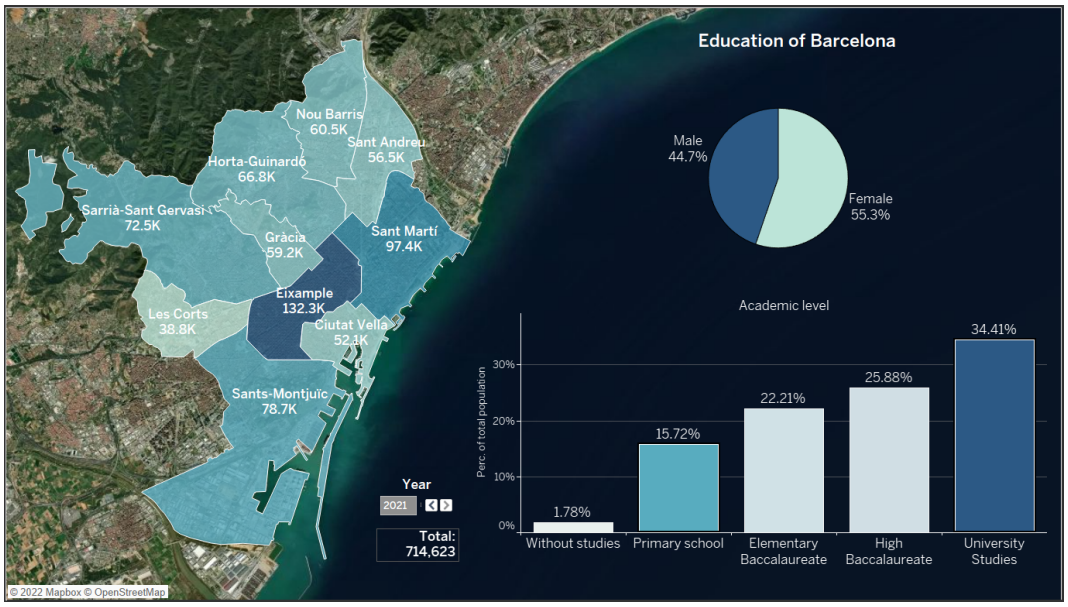

Figure 5.6: University and primary school gender distribution

The same can be said for men in the Baccalaureate categories from figure 5.7: the ratio of males in those two academic categories is 50.1%, which is more than the total average for males by 3%.

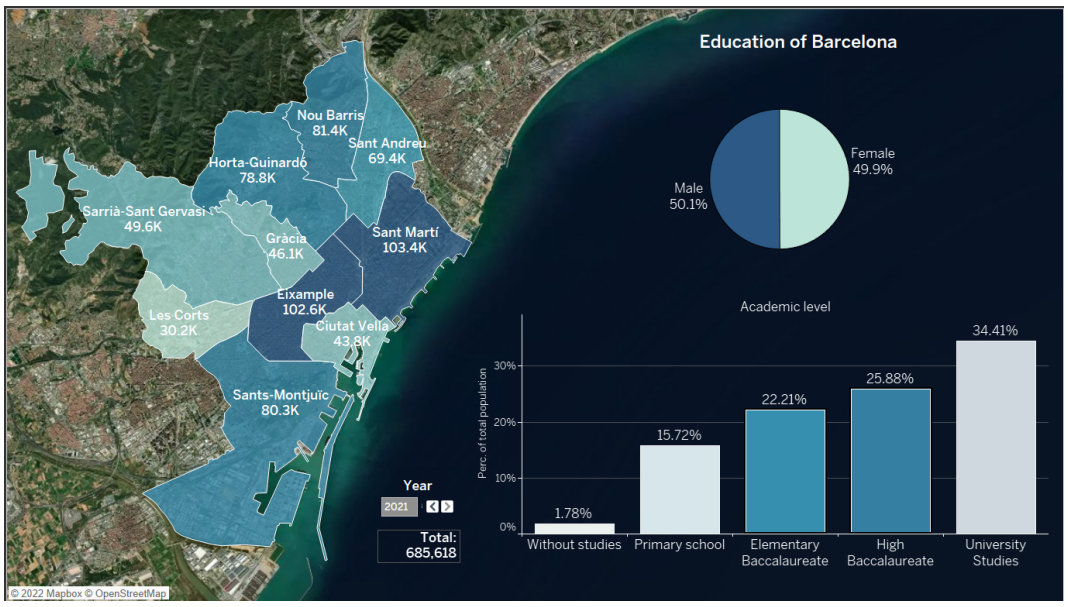

Figure 5.7: Baccalaurate gender distribution

As for people without studies, out of the 25428 people, 72.9% are female, which is a significant difference from the total average of 52.9% (figure 5.8).

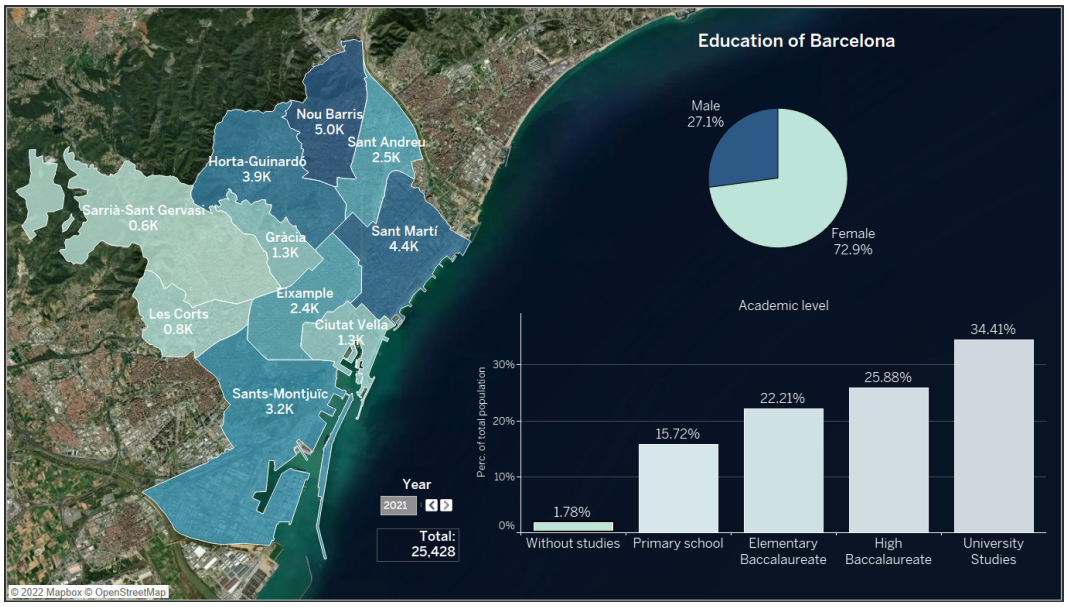

Figure 5.8: Without studies gender distribution

From figure 5.5 a more detailed analysis can be carried out, looking at how in each district cluster the above-mentioned phenomena are present. The moderate and bad educational district clusters behave the same way as the total of Barcelona. However, the academically better performing cluster differs from the other two in the university studies category: the ratio of men and women with a university degree are the same in their respective group as compared to the total average which means women are proportionally less represented in said category.

5.1.3 Trends

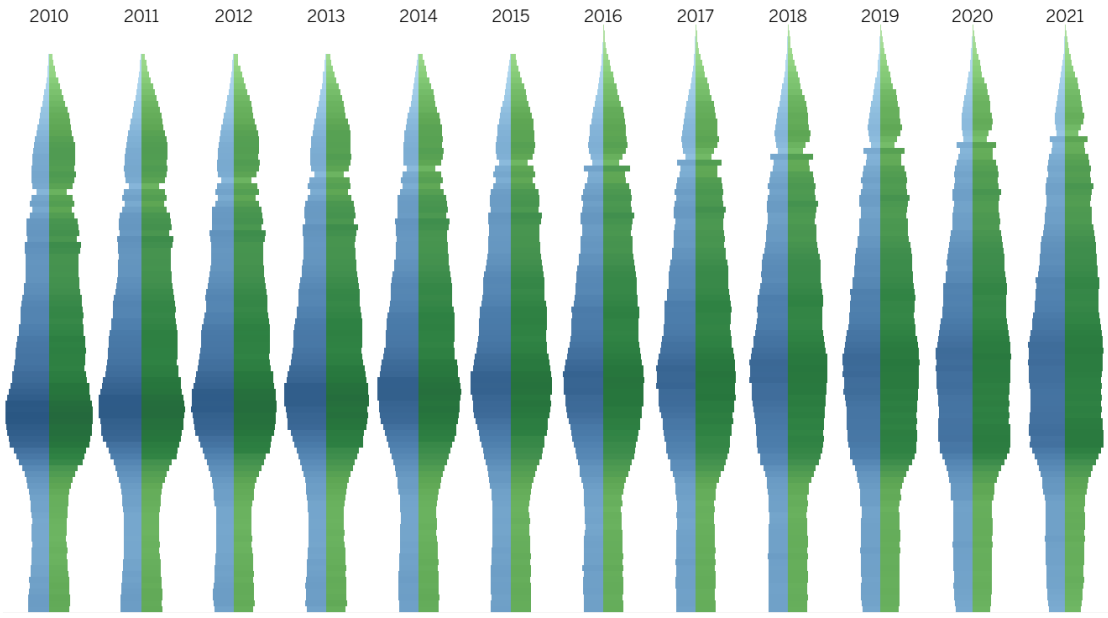

Figure 5.9: Population pyramid over the years

Looking at figure 5.9, it is clear that Barcelona has an aging demographic. The bottom part of the population pyramid has shrunk, meaning there are fewer and fewer children born every year. However, the top of the pyramid seems to grow signaling a higher presence in older age groups. As per the World Health Organization, this is a global phenomenon and is a result of the continued decline of fertility rates and increased life expectancy [15].

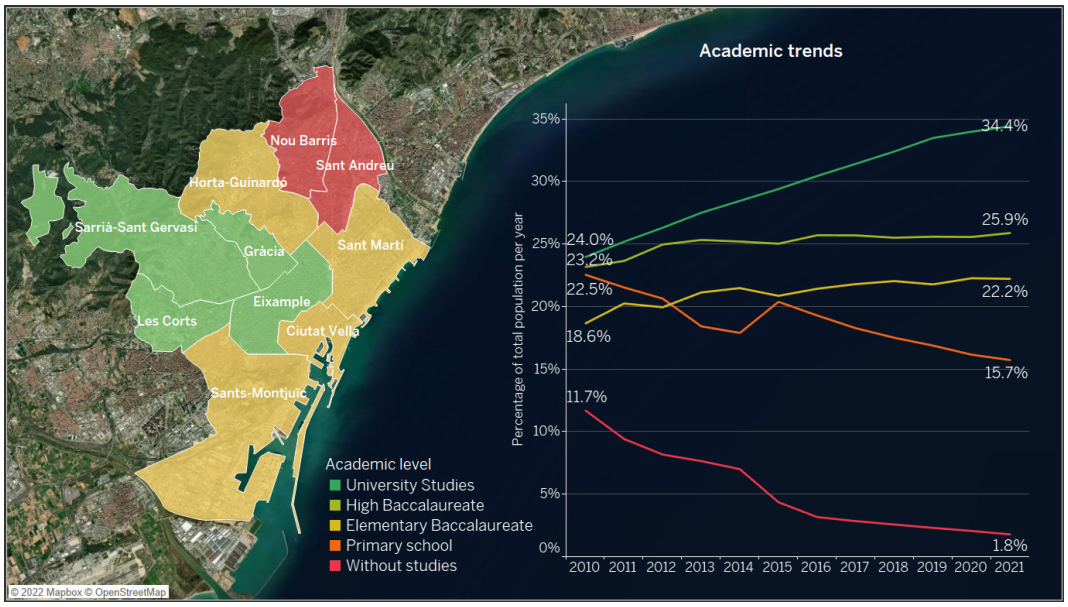

Figure 5.10: Academic trends

Moving on to academic trends, a positive outline can be seen. The percentage of people with no education whatsoever has shrunk from 11.7% down to 1.8%. The ratio of people with a university degree has grown by the same amount as the previous metric shrunk, by around 10%. Overall, around 72% of people have been studying at least until the age of 18, as Baccalaureate studies are meant from the age of 16 to 18.

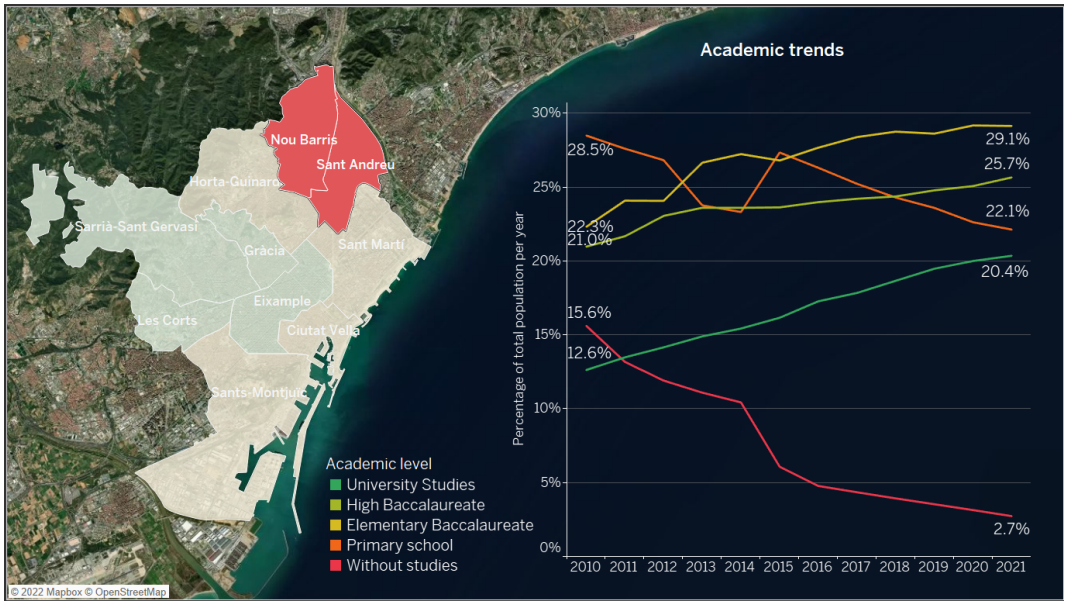

Figure 5.11: Academic trends in underperforming districts

Even when looking at underperforming districts overall from the 2021 clustering, a positive trend is visualized. From these two districts in figure 5.11, the share of people with a university degree was the lowest in 2010, however, as of 2021, it is close to the top 3, while the non-educated population ratio has shrunk with around 13% points. When observing the other two clusters, a similar trend is showing.

5.2 Accidents

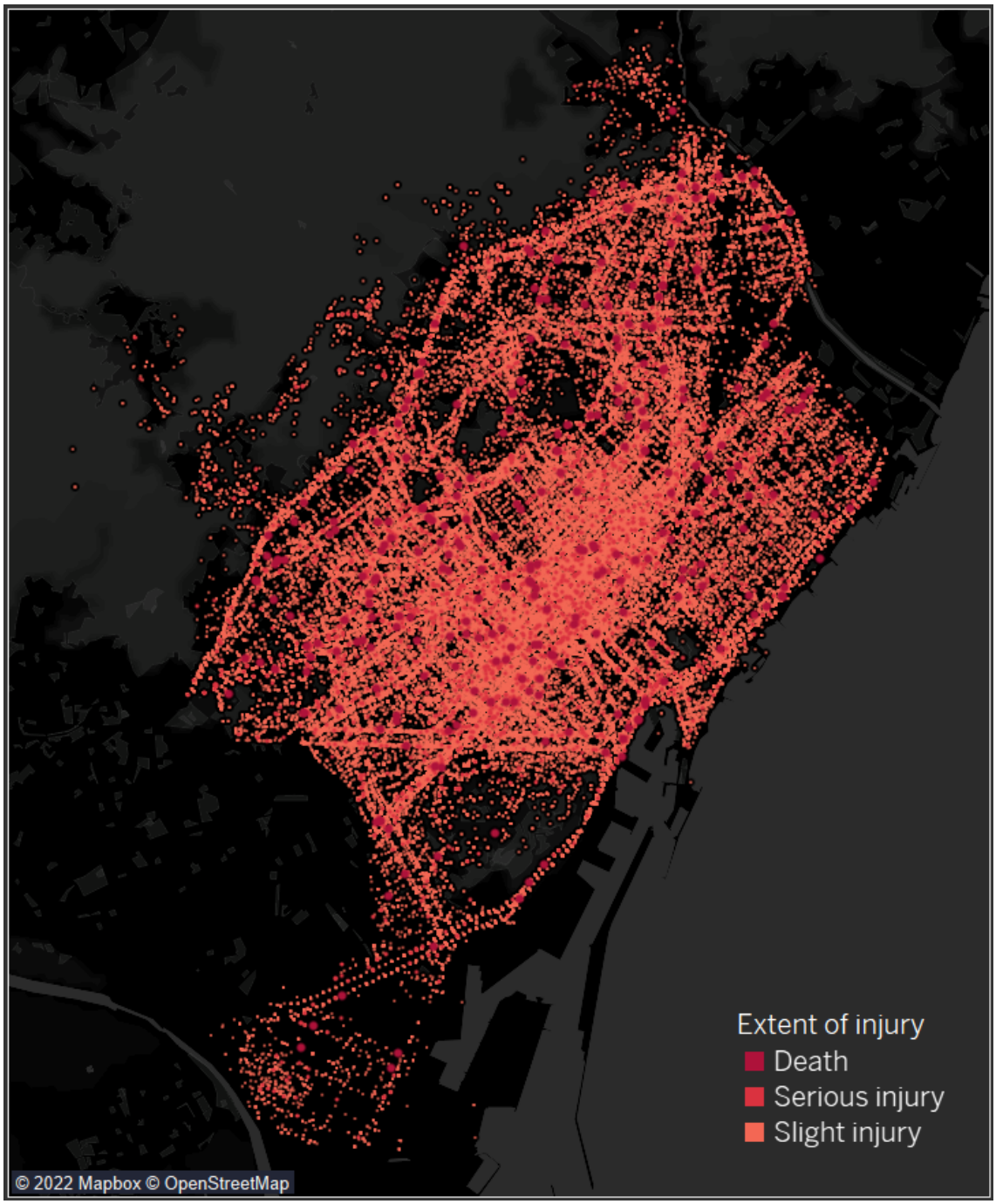

Figure 5.12: People involved in accidents

In figure 5.12 the spatial occurences of people being involved in accidents can be seen. At first glance it is clear that the closer we move to the city centre, the more accidents will happen. There are three color codes: the more serious the injury, the darker the red in the point of interest.

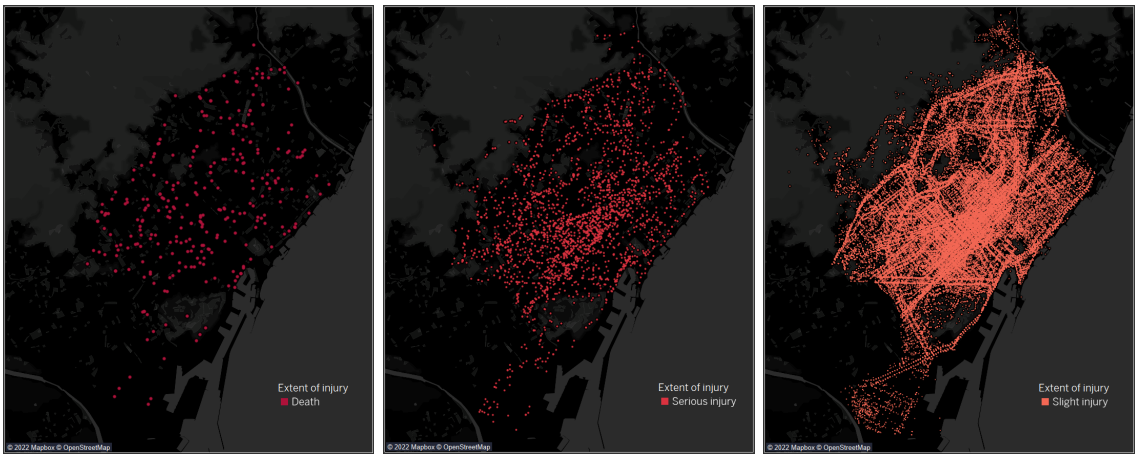

Figure 5.13: People involved in accidents (death, serious, slight)

In the figure above, the seperate injury levels and their spatial occurences can be seen. Most people who are involved in an accident, leave with a slight injury. This information has been deducted from the exploratory data analysis, however, it is interesting to see it from a different angle.

5.2.1 Time distributions of accidents in different districts

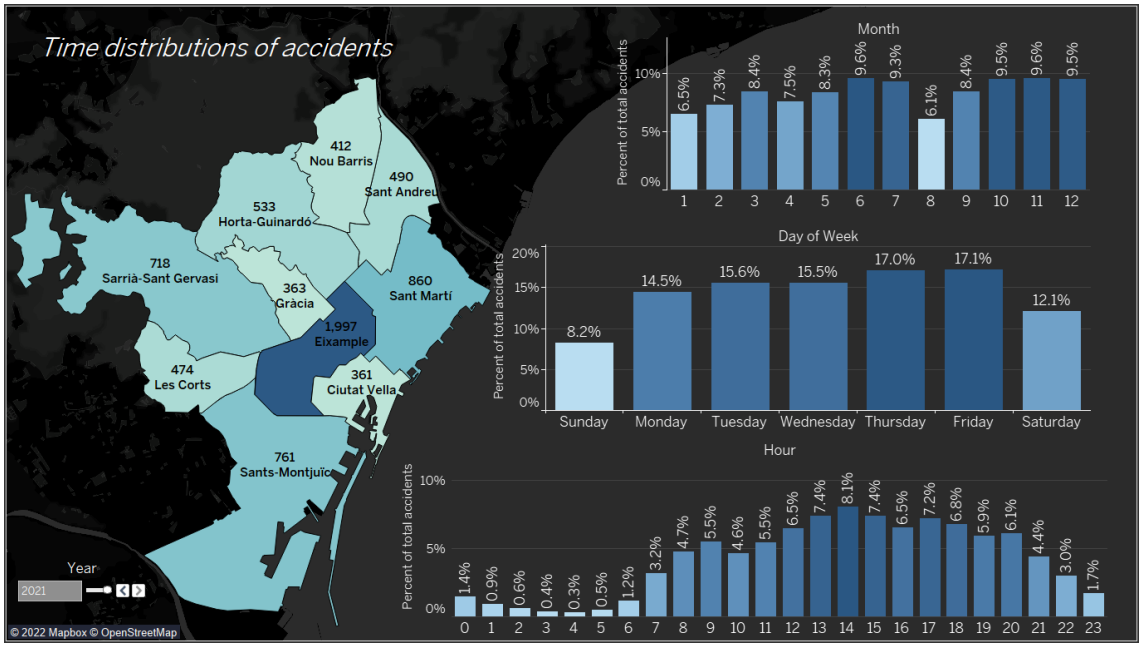

Figure 5.14: Time distributions of accidents in 2021

In figure 5.14, a highly interactive dashboard can be seen, currently showing the time distribution of accidents in 2021 with accident numbers for each district. The map, the

year selector, and each time graph can be used to filter the data. With this dashboard it is possible to answer questions such as *"How many accidents happened on Fridays in July 2021 between 14 and 16 in each district?"*.

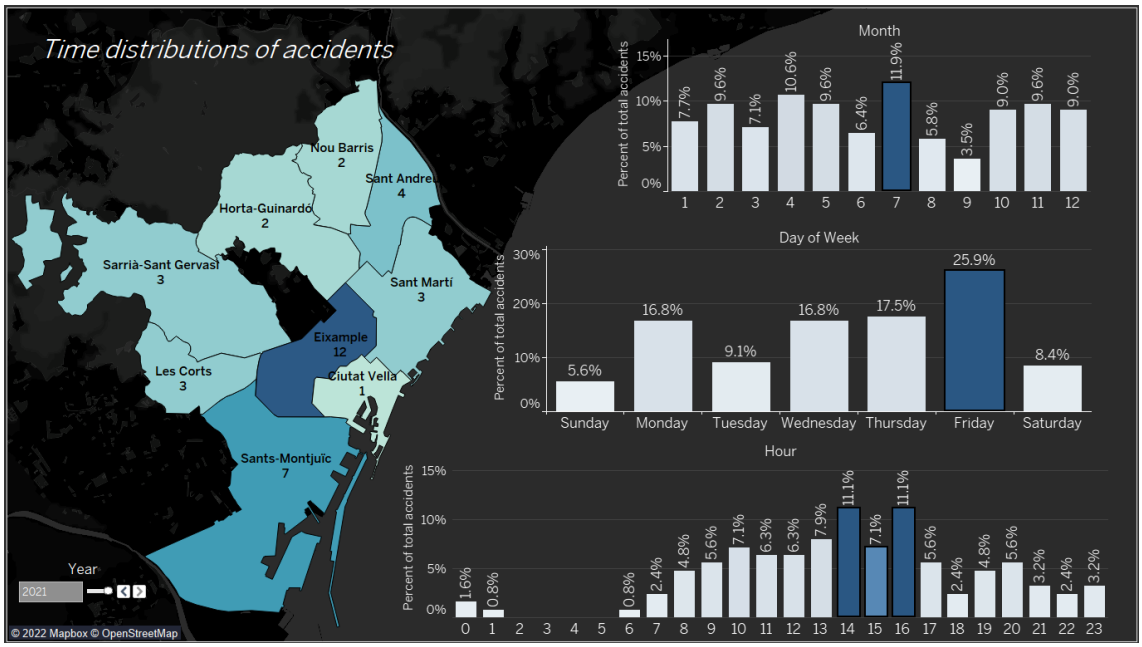

Figure 5.15: Accidents in each district between 14 and 16, Fridays, July 2021

The figure above answers the question immediately, showcasing the power of highly interactive Tableau dashboards. With it, very specific questions can be answered in a few seconds. We may ask question from another angle. Let the question be *"What was the hour the most accidents happened on Fridays in the summer of 2019 in Eixample district?"*.

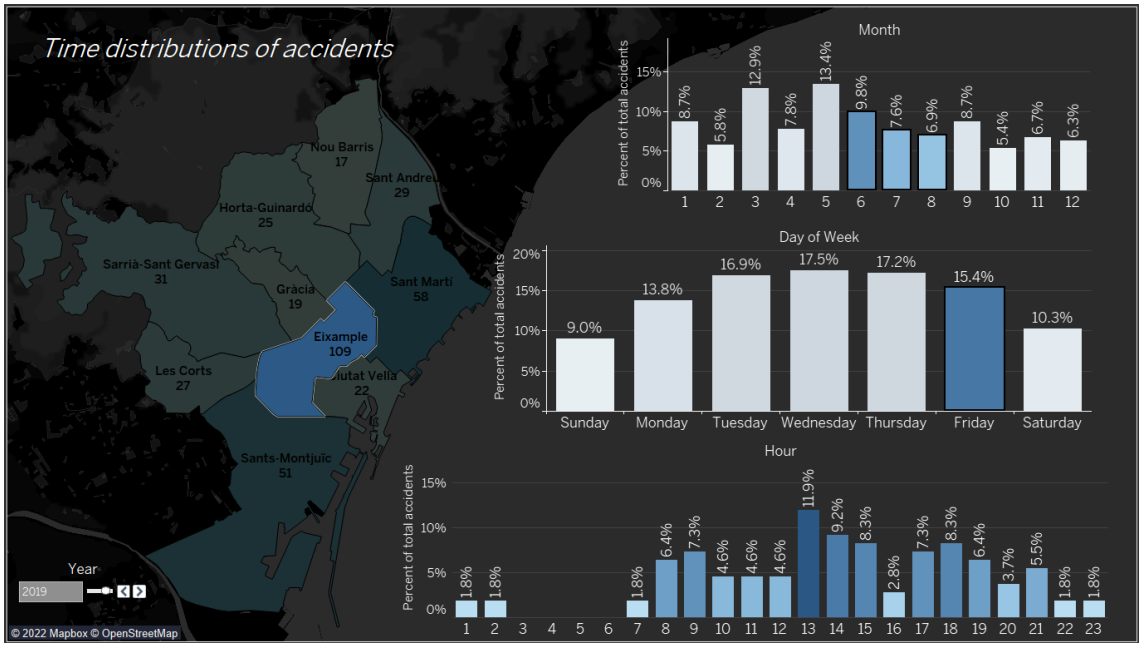

Figure 5.16: Hourly distribution of accidents on Fridays in the summer of 2019 in Eixample

The short answer is between 13 and 14. The longer answer is: between 13 and 14, 11.9% of accidents happened on Fridays in the summer of 2019 in Eixample, which out of the 109 accidents happenening those Fridays, 13 of them happened between 13 and 14.

This kind of dashboard is good if we want to answer very specific question and look at time distributions, however, it does not say too much about dangerousness of districts.

5.2.2 Which district is the most dangerous?

The goal is to locate districts that have proportionally more accidents than others. Ranking each district by the number of accidents is not enough as there are various different aspects that can determine this number. A few of these aspects can be the accident severity, traffic, population, number of pedestrians, number of cars registered and size (in terms of area, road intersections, road distance). These aspects need to be used to normalize the number of accidents in each district. The normalized value can be used to rank districts in terms of dangerousness. The optimal solution would be to use all of the influencing aspects, however, the availability of the data is a challange. I will be able to use accident severity, population, area and number of vehicles registered to normalize the number of accidents. I have to mention here that the datasets used for this calculation are between years 2016 and 2021, covering 6 years. The resulting dataset has 60 rows, because the granularity is by year and by district. With 10 districts and 6 years, 60 different records are present.

5.2.2.1 *Damage score* of accidents

A question that appeared was the choosing of the damage value of accidents in terms of the extent of injury of people being involved in said accident. In the *Accidents by severity* dataset I have a breakdown of accident victims in terms of extent of injury which I used to calculate the damage value of each accident. The equation is as follows:

$$\mathbf{D(a_i) = a_i(slight) * SLIGHT + a_i(serious) * SERIOUS + a_i(death) * DEATH}, \tag{5.1}$$

where $\mathbf{D(a_i)}$ is the damage score of accident $\mathbf{a_i}$, $\mathbf{a_i(slight)}$ is the number of slightly injured participants in accident $\mathbf{a_i}$, and **SLIGHT** is a constant weight. The **SLIGHT**, **SERIOUS**, and **DEATH** constants were calculated as follows:

$$\mathbf{SLIGHT = 1} \tag{5.2}$$

$$\mathbf{SERIOUS = 3} \tag{5.3}$$

$$\mathbf{DEATH = 5} \tag{5.4}$$

The choosing of the weights was based on a report on *Identifying and ranking dangerous accident locations: Overview sensitivity analysis* [10]. The cited paper thoroughly details the consequences of choosing different weights and comes to the conclusion that the choice of weights will have an important effect on the type of accident locations in the ranking system (e.g. locations with high traffic volumes resulting in many small accidents compared to regions with less traffic but more serious accidents). The paper says the governments should therefore carefully decide which priorities should be stressed in the traffic safety policy. As for my case, I went with weights of 1, 3 and 5, as the paper considers it a moderate choice. With equation 5.1, each accident now has a *damage score*.

5.2.2.2 Normalizer, Normalized index

Another big question with this approach is the choice of the significance of the metrics (cars registered, population, area). One may be more important than the other, I have to find a way to determine the significance of metrics. My thought process was, that the more a certain metric correlates with the number of accidents, the more significance it should carry. I used the scipy python package to calculate the Pearson correlation coefficient between the number of accidents and all of the metrics listed above.

ρ	**Vehicles registered**	**Population**	**Area**
Number of accidents	0.81087	0.73796	0.11416

Table 5.1: Correlation of metrics

The registered number of vehicles correlates the most with the number of accidents, which means it will carry the most significance. The value set of these metrics can differ greatly, which is why I decided to scale each of them between values of 1 and 2 and then multiply each metric with its significance. A normalizer was formed for each district and year:

$$\mathbf{N}(d,y) = \rho(pop) * \mathbf{pop}(d,y)_2^1 + \rho(veh) * \mathbf{veh}(d,y)_2^1 + \rho(area) * \mathbf{area}(d,y)_2^1, \tag{5.5}$$

where *d, y* are district and year, $\rho(pop)$ is the correlation between the number of accidents and population, and $\mathbf{pop}(d,y)_2^1$ is the scaled value ([1;2]) of the population in the parameterized year and district. For each year and district, a weighted, scaled value is created called the *Normalizer*. This *Normalizer* is then used to correct for the effects of the listed metrics on the accident numbers. Another metric is created, called the *Normalized index* which is the metric that will be used to rank districts each year. This index is calculated as follows:

$$\mathbf{NI}(d,y) = \frac{\sum_{i}^{d,y} \mathbf{D}(\mathbf{a_i})_2^1}{\mathbf{N}(d,y)} \tag{5.6}$$

where $\mathbf{NI}(d,y)$ is the index for a certain year and district, $\sum_{i}^{d,y} \mathbf{D}(\mathbf{a_i})_2^1$ is the scaled ([1;2]) summation of *Damage score* values of individual accidents for every year in each district.

This ranking by *Normalized index* will give a better view of the dangerousness of districts rather than using the plain metric, the number of accidents.

In figure 5.17, the ranking of districts can be seen based on the created *Normalized index*. The bottom bar chart shows the number of accidents in comparison, to get an idea of how much the ranking has changed using the *Normalizer*. After the adjustment, Eixample is only fourth in terms of dangerousness, and the most dangerous district with this metric is Les Corts. Once again, there were four aspects taken into account when normalizing: accident severity, registered vehicles, population, and area. The selection of aspects could have been much broader, however, out of the listed, potentially affecting metrics these were the ones I found useful and suitable data for. It is important to mention that the selection is purely subjective based on intuition, which implies that the new ranking is subjective as well.

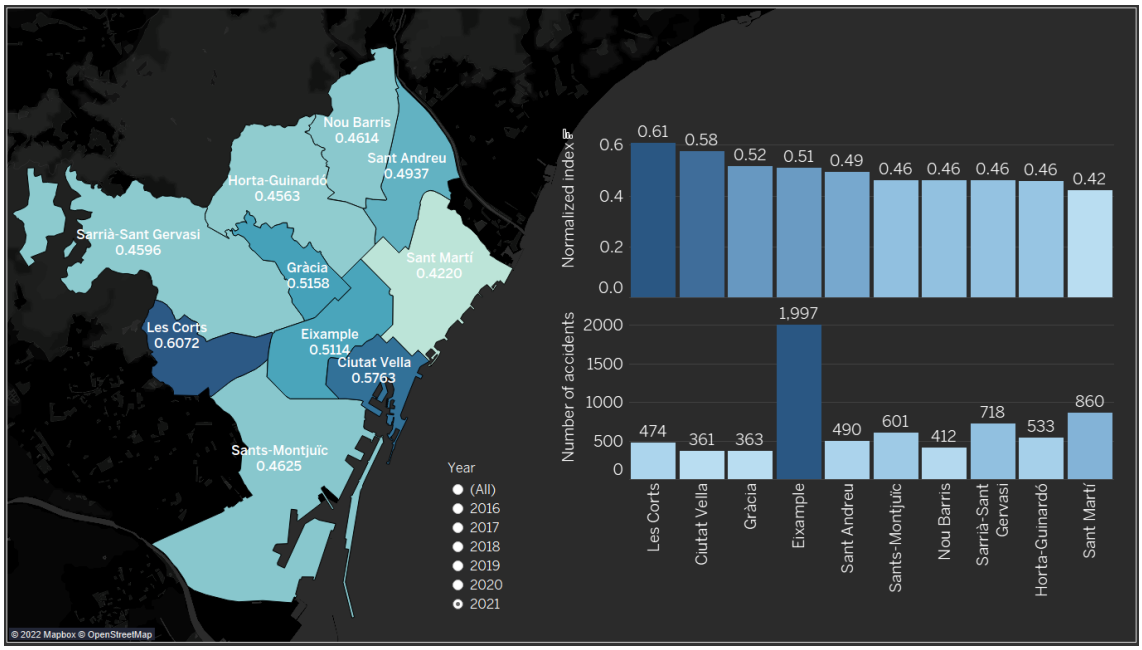

Figure 5.17: Ranking of districts based on the *Normalized index*

Intuitively, the higher the denominator (or the *Normalizer*) the lower the *Normalized index*. The *Normalizer is made up of a weighted sum of three aspects*. Ultimately, this is only a correction for the skews resulting from the three aspects and not anything else. It is not a surprise that the top three districts had the lowest value of *Normalizer*s. The *Normalizer* shows how much a combination of certain aspects affects the number of accidents metric. In the following, I will continue the analysis based on numerous combinations of fields and filters in the datasets.

5.2.3 In-depth analysis of different aspects and their combinations

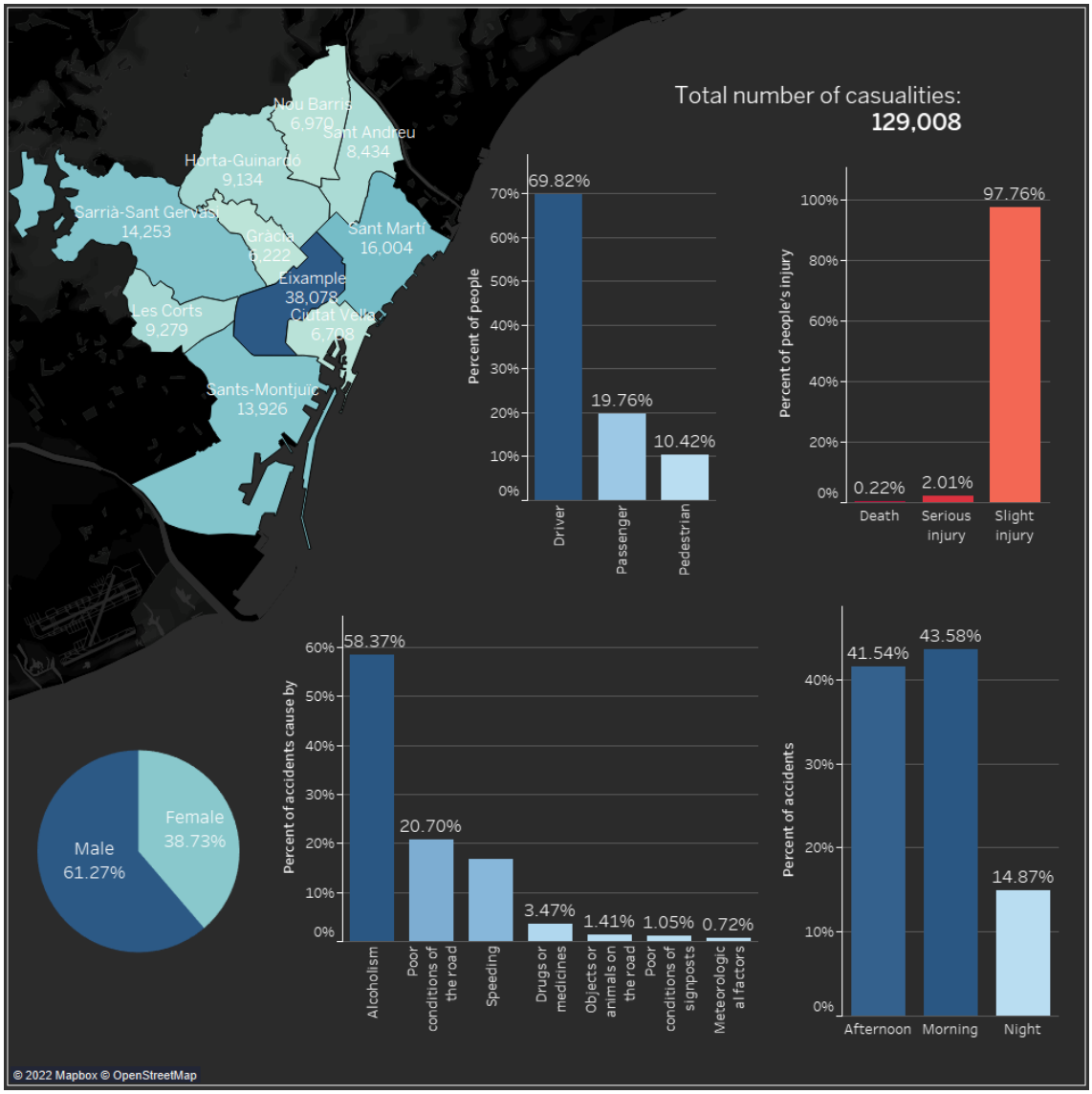

Figure 5.18: Dashboard for in-depth analysis of accidents

Besides others, primarily I will be using the dashboard above to further analyze accident participation between 2010 and 2021. During the data exploration, all of these aspects have been shown individually, however, using them as a filter on a dashboard of this complexity shows great insights into underlying aspects. There are five different metrics represented with distribution: gender, cause of accident, type of person, the extent of injury, and part of the day. By clicking on either graph a filter is applied to the rest of the dashboard. The total number of casualties label also changes based on the filter applied.

Looking at the data on deaths is important as finding either group of people highly affected, or areas with proportionally more deaths can show where the problem is, and allow certain authorities to make the necessary precautions and adjustments in the future of the road network. By clicking on deaths in the histogram of the extent of injury, I filtered the rest of the dashboard.

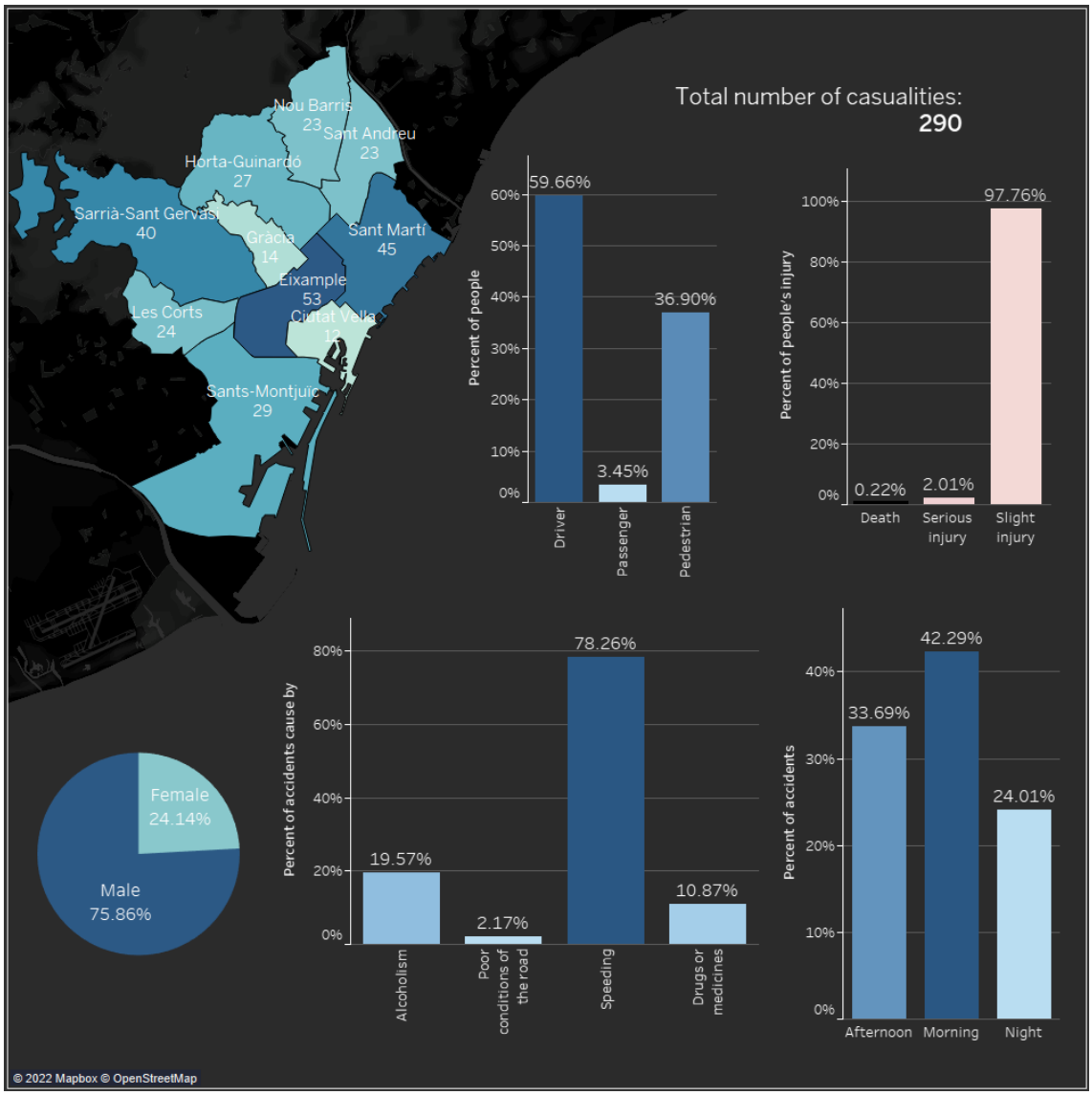

Figure 5.19: Deaths analysis

There are big deviations in metrics compared to the original, unfiltered dashboard. Around 76% of people that died in accidents were male, compared to the overall accident participation representation of around 61%. This shows that men are more in danger compared to women, however, there is an important aspect that needs to be taken into account. The number of accidents should be normalized with the distance driven in each group, and unfortunately, I did not find any relevant data on this for my case. However, there is a study from the United States, in 1990, where the data suggests men in a younger age group are more involved in fatal accidents than women [6]. On the other hand, in non-fatal, but injury-causing accidents, women are more represented overall. An important aspect this study analyzes is age. I will analyze accident participation by age and gender in a later section.

Moving onto causes of accidents resulting in fatal casualties, around 78% of fatal accident participants were a result of speeding if the cause could be determined immediately. This is a huge increase compared to the 17% in all of the accidents.

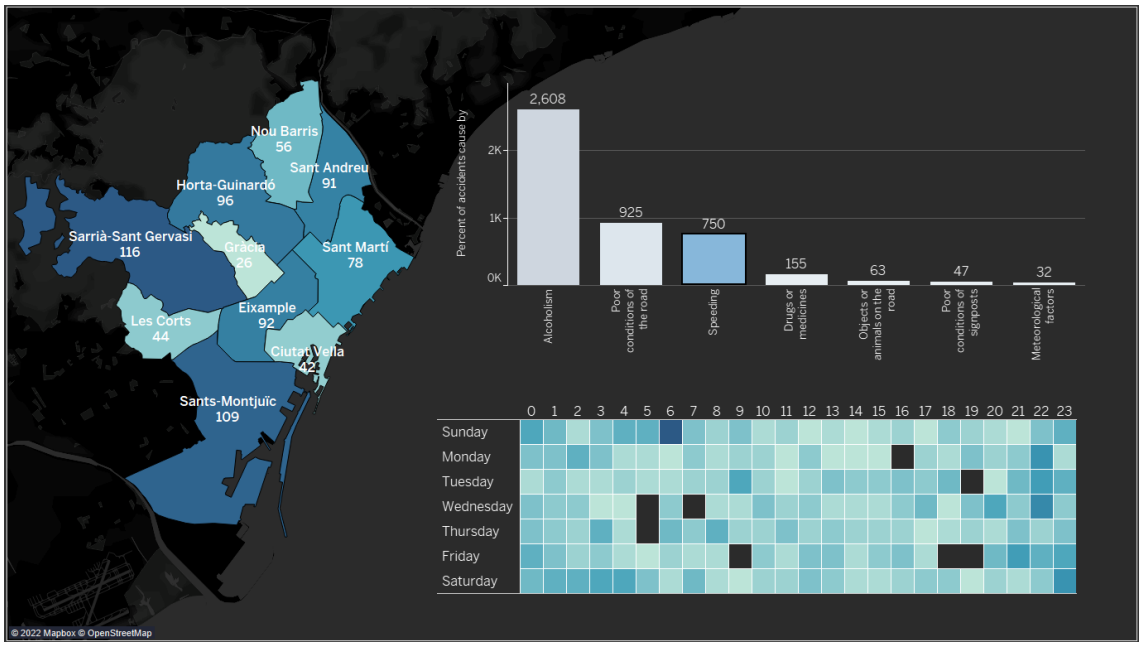

Figure 5.20: Speeding as cause occurrences

An interesting aspect when looking at *Speeding* as the cause of the accidents, is that as opposed to most accidents happening in the city center in general, accidents caused by Speeding happen more in the outskirts. The reason for that is the maximum speed limit being lower in the center of the city. This is perfectly visible in figure 5.20. The dashboard in the figure also contains a heatmap of accidents by day of week and hour. Accidents caused by speeding do not follow the regular accident pattern (high traffic hours). Proportionally more accidents happen caused by speeding at night (between 20 and 6). Potential reasons for this might be that traffic is not as high during the night which gives people more room to go faster as well as poor visual conditions with the lack of light during the night.

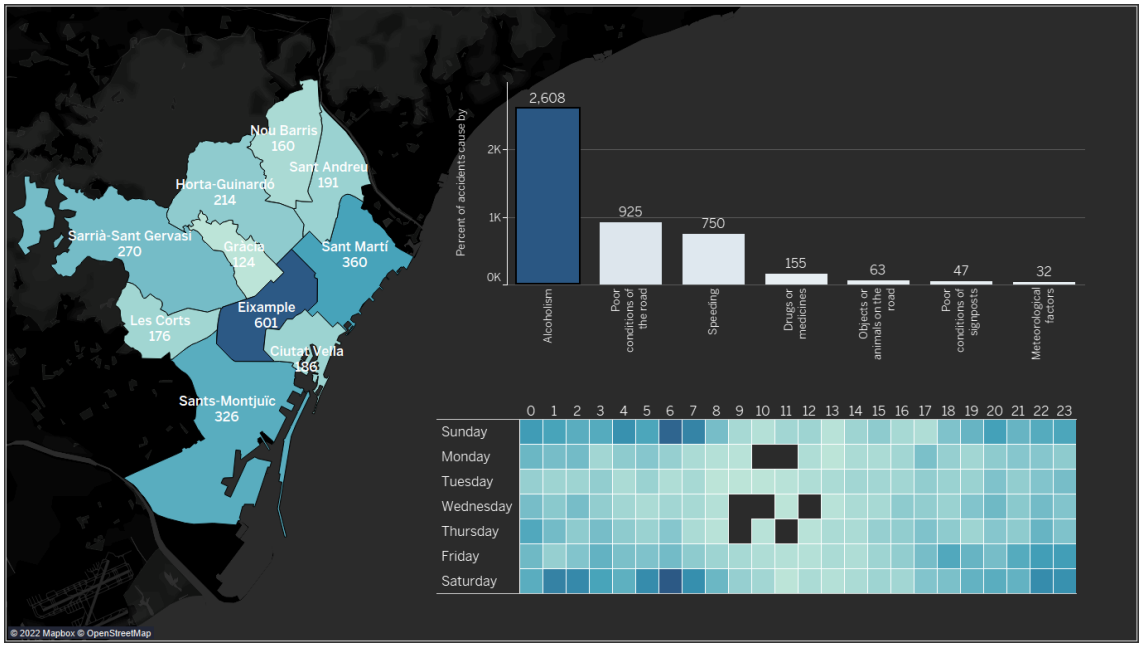

Figure 5.21: Alcohol as cause occurrences

When looking at the heatmap in figure 5.21, accidents caused by alcohol can be seen. A similar pattern is visible as was in Speeding: at night, there are proportionally more accidents caused by alcohol consumption than during the day. On top of this, looking at weekends and weekdays, during the weekend much more accidents happen that are caused by alcohol consumption. In general, more people drink at night as well as during the weekends which is a potential reason for this pattern.

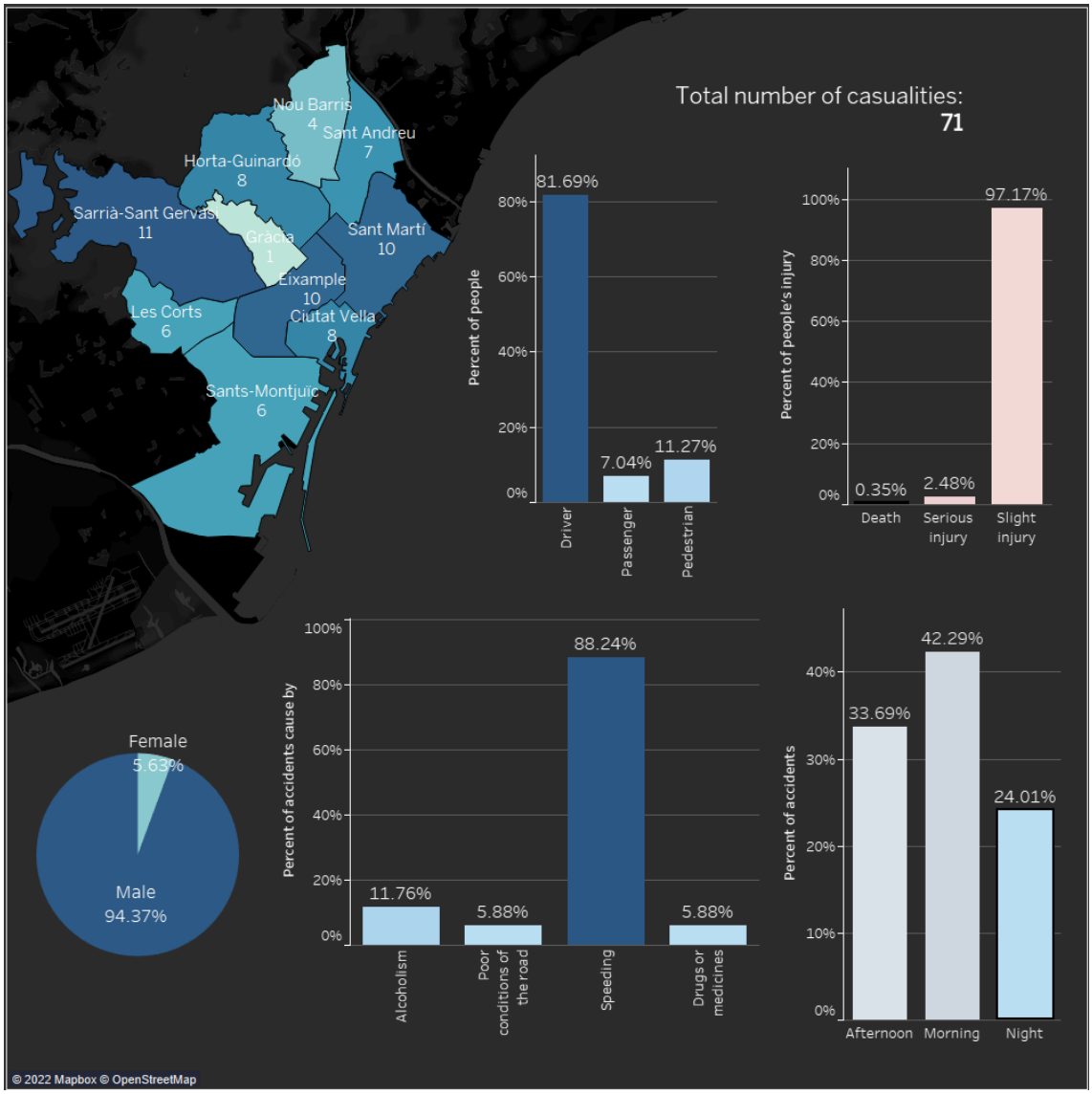

Figure 5.22: Deaths at night

Looking back at the dashboard in figure 5.19, another interesting feature is obvious when filtering to death. The person involved in an accident with this classification can either be a driver, passenger, or pedestrian. In general, around 10.5% of accident participants are pedestrians, however, this number jumps up to 37% in fatal participants. At the same time, passenger representation decreases from around 20% to 3.5%, essentially indicating that the chances of a fatal casualty being a passenger are quite low (out of the 290 deaths, 10 were pedestrians).

Looking at the time of day distribution, there is a 10% increase in accidents at night when comparing overall accident participation and fatal accident participation. This shows that there are proportionally more fatal accidents at night than overall accidents at night.

When observing deaths in accidents at night, it is clear that speeding as the cause has increased even more, to around 88%. At the same time, the share of men in this specific group increased to around 95%, which is an overwhelming majority. The takeaway here is that in fatal accidents at night, the casualty is most likely going to be a male driver, with the cause of the accident being speeding. This indicates that men at night are very careless and fast drivers. The exact numbers are present in figure 5.22.

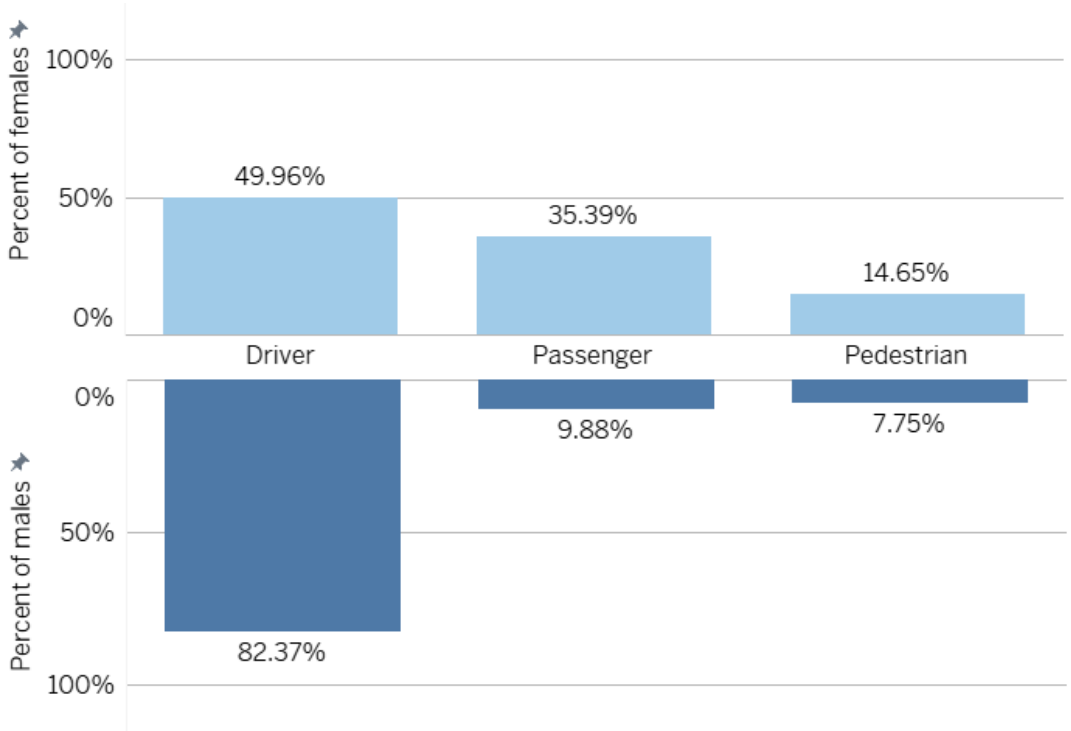

Figure 5.23: Gender distribution of drivers, passengers and pedestrians

The *type of person* distribution of males and females can be read from the dashboard used previously, however, I wanted to put both side by side, hence figure 5.23. It shows that the vast majority (∼82%) of males involved in accidents are drivers, while only half of the females are drivers. The share of females both in the groups of passengers and pedestrians is greater than males.

Moving onto accident representations in different age groups for the two genders, we can clearly see, that up until the age of 70, there are almost twice as many men involved in accidents than women. If we look back at the age composition in figure 5.3, for people aged 70 or more, it is clear why the representation changes. There are simply much more women after the age of 70 than men, increasing the chance of women being involved in more accidents.

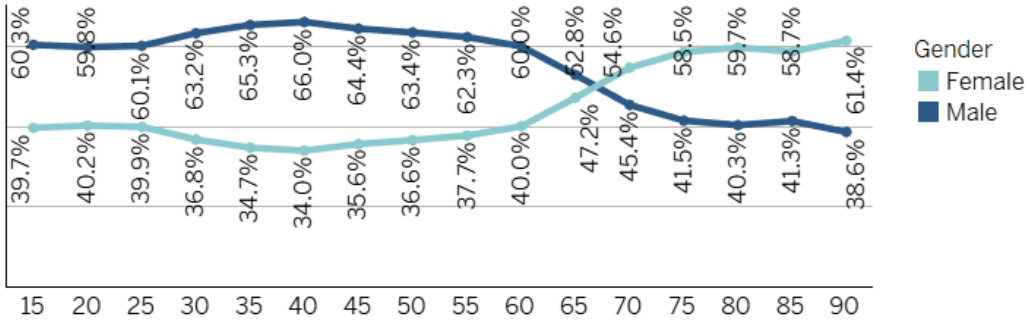

Figure 5.24: Gender representation in accidents for different age groups

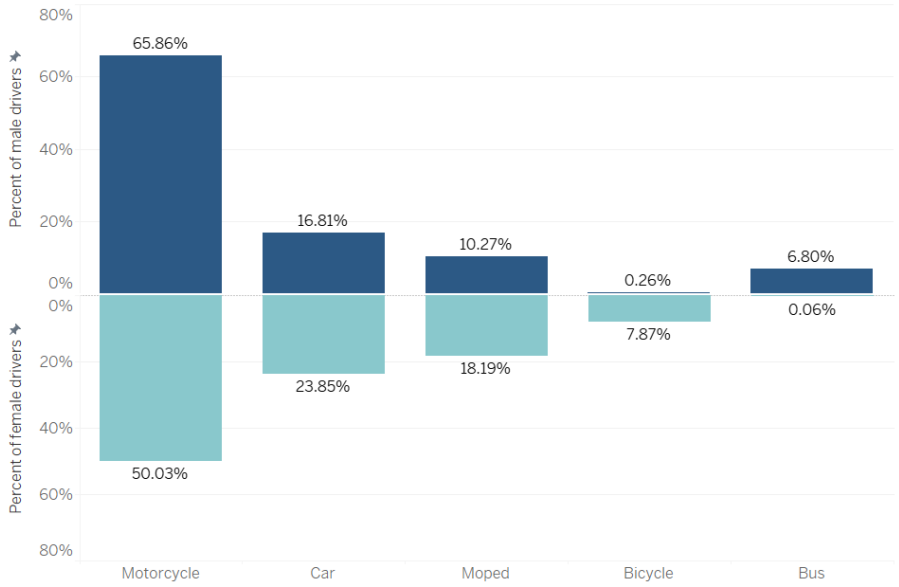

Figure 5.25: Gender distribution of types of vehicles used by driver

The figure above shows the distribution of vehicles used by male and female drivers. There are three main takeaways. First, the vast majority of male drivers involved in accidents rode motorcycles, while that is true only for half of the female drivers. Secondly, when it comes to riding a bicycle, almost no male bicycle riders were involved in accidents as opposed to around 8% of female drivers involved in accidents riding bicycles. Lastly, almost no female bus drivers were involved in accidents as opposed to around 7% of male drivers involved in accidents driving a bus. The reason for this last point might be, that there are very few female bus drivers, to begin with. I could not find data for Barcelona to back this up. However, I have found it for London, and since culturally and socioeconomically both London and Barcelona belong to Western Europe, I figured the data from London can be of guidance. According to the website of the government in London, UK [2], the share of women bus drivers rose to 7% by 2014, which is a good indicator, that the number of female bus drivers is very low compared to male.

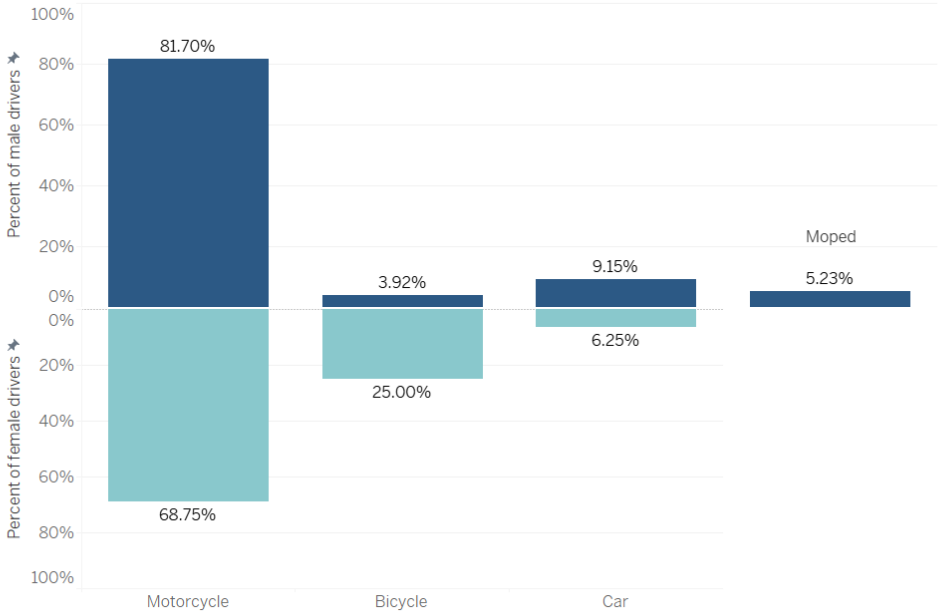

Figure 5.26: Gender distribution of types of vehicles used by a driver with fatal injury

Comparing the histograms shown in figure 5.25 and 5.26, it is clear that the representation of motorcycles in fatal accidents is higher than in accidents with all sorts of severity: especially in the male group, if a driver dies in an accident, he most likely was driving a motorcycle (81.7% of fatal male casualties driving).

5.2.4 Difference in accident participation age

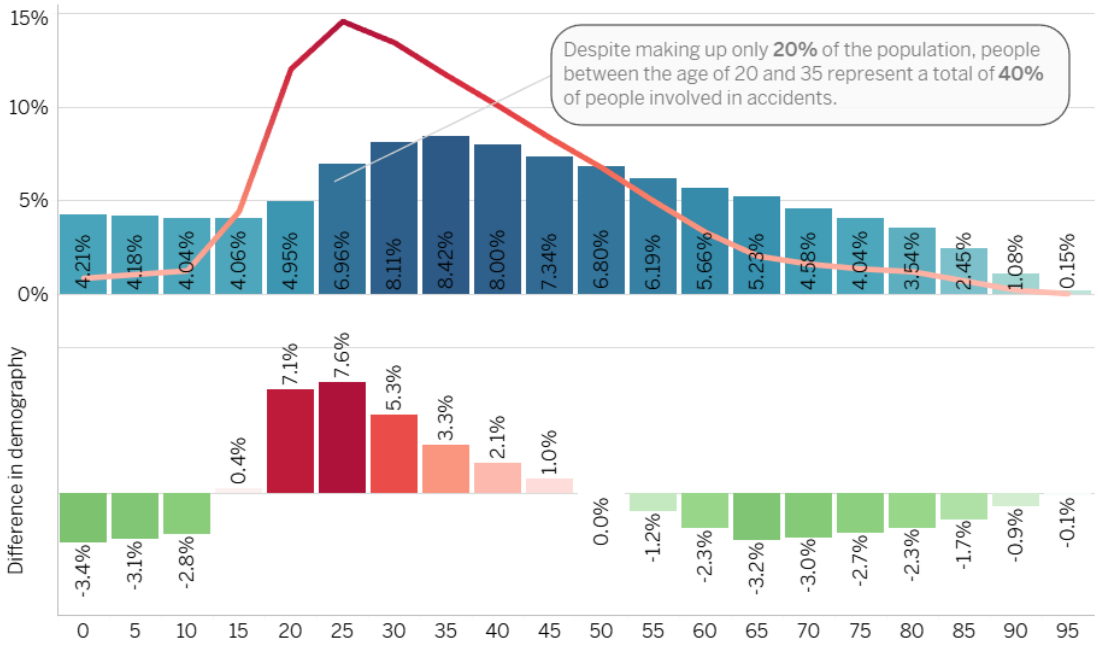

Figure 5.27: Difference in accident participation by age group

In figure 5.27, the age group participation in accidents can be compared to the population tree. On the bottom part, the percentage difference between accident participation and age group representation can be seen. The figure clearly shows that the younger age groups tend to participate in more accidents relative to other age groups. A possible reason might be driver inexperience. The earliest a Spanish citizen can acquire a B-1 driver's license is the age of 16 [5]. This means that people between the age of 20 and 35 can have relatively little experience on the roads, resulting in an increased risk of traffic accidents. Not only the lack of experience, but the reckless driving of youngsters result in that age group being involved in more accidents relative to others. The paper I have previously cited also backs up this fact [6], where the authors dissected accident data collected from surveys, and the accident rates per distance driven clearly show that youngsters were proportionally more represented in accidents than others. The paper shows distributions by age group for all sorts of scenarios, including fatal accidents, crashes without injury and crashes with injury. The signal is clear, supporting my statement above. Unlike the cited paper, I have not had access to data regarding distance driven by age group, which means I can not be sure this is true here, however, given the cultural indifference in countries from better socioeconomic backgrounds, the assumption is straightforward.

5.2.4.1 Tableau calculated fields, bins, and parameters

As I mentioned in the section where I introduced Tableau, if it comes to using a more advanced feature of the software that helps me create dashboards and worksheets, I will detail the implementation. This is a good example in figure 5.27 as I used both *bins* and *calculated fields*[2].

Calculated fields in Tableau allow users to create additional data from the data that already exists, essentially creating a new dimension or measure (in tabular data terms we would call it a column). There are a lot of options with calculated fields including ratio calculations, data aggregation, data segmentation, and result filtering. To define a calculated field you have to click on either a measure or a dimension which will result in a pop-up window where the calculated field can be defined. Looking at figure 5.27 as an example, I needed to define three calculated fields to create the Worksheet. They were *SumOfPopulationPercentOfTotal*, *CountOfPeopleInAccidentsPercentOfTotal*, and *Diff. in demography%*. The calculations were as follows:

$$\textbf{SumOfPopulationPercentOfTotal} = \frac{\textbf{SUM}(\textbf{nr\_inhab})}{\textbf{TOTAL}(\textbf{SUM}(\textbf{nr\_inhab}))} \tag{5.7}$$

$$\textbf{CountOfPeopleInAccidentsPercentOfTotal} = \frac{\textbf{CNT}(\textbf{people\_in\_accidents})}{\textbf{TOTAL}(\textbf{CNT}(\textbf{people\_in\_accidents}))} \tag{5.8}$$

And finally *Diff. in demography%* equals the difference of *SumOfPopulationPercentOfTotal* and *CountOfPeopleInAccidentsPercentOfTotal*. There are numerous other imperatives that can be used to define a calculated field such as conditionals, operations, and so on.

Bins in Tableau are an abstraction for discrete groups. For example, any discrete series of numbers can be considered bins. Furthermore, it is possible to define bins from discrete numbers, with the option of setting the size of the bin. In my example, there were people in accidents from the age of 1 to 100. To be able to handle them in groups of integer units, I defined a bin with a size of 5 at first. From now on if I use that bin instead of the age,

[2] I have used bins previously, in numerous other dashboards.

the data will be aggregated to the group size. This means that with the bin size of 5, I will have twenty separate age groups: 0 to 5, 5 to 10, 10 to 15, and so on.

Another great feature of Tableau to combine bins with is *Parameters*. Instead of giving a constant value of 5 for the bin, I can pass a parameter that can also be visible on the Tableau dashboard, allowing manual setting of the size of the bin during the analysis of the dashboard. This gives a lot of options when observing the age characteristics of a population or people involved in accidents.

5.2.5 Biggest accidents

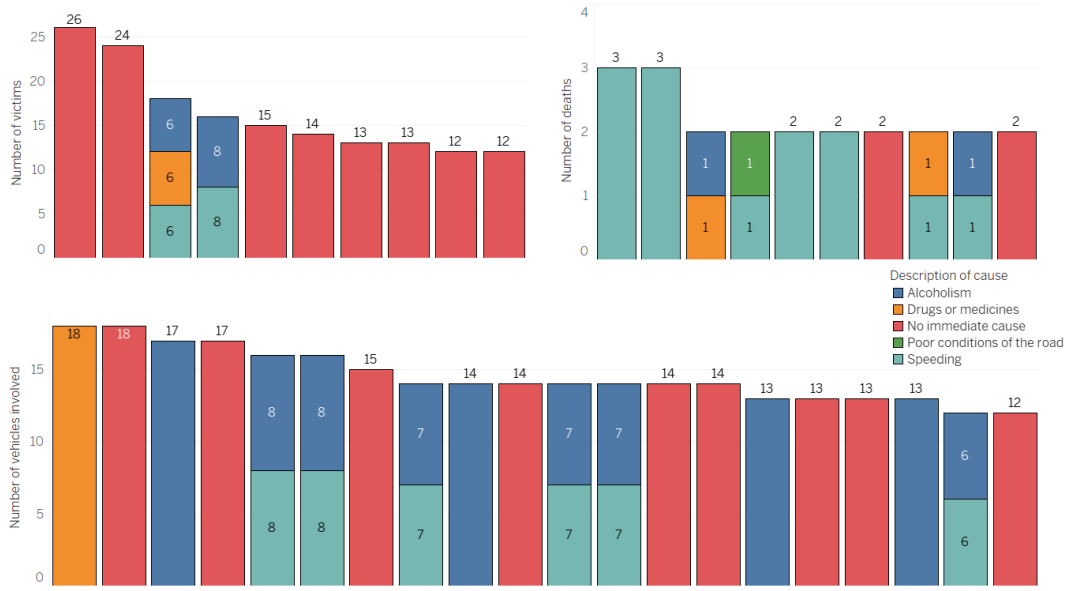

Figure 5.28: Biggest accidents in terms of victims, deaths, and vehicles involved

Based on the accidents by severity dataset, I created a dashboard showcasing the accidents with the highest number of personal and vehicular involvements as well as accidents with the most deaths. The different colors show the cause (or causes) of the accidents.

Looking at figure 5.28, in the top left corner, accidents with the highest number of victims can be seen. In terms of this metric, the most severe accident included 26 people, with 24 being second on this list. Out of the top 10, none included fatal casualties, and out of the 24 and 18 victims of second and third accidents, 3 people were seriously injured in each accident. When looking at causes, a key takeaway is that in most cases (8 out of 10), authorities could not immediately determine the cause of the accident. It is understandable as with more casualties come more factors to take into account, making it harder to point to a limited number of causes which results in uncertainty.

Looking at the top right section of the figure, the accidents with the highest number of deaths can be seen. The most severe accidents in terms of this metric had 3 fatal casualties and there were two accidents like this. In the top 10 list, the light blue color is overwhelming compared to others, this signals that accidents with the most fatal casualties resulted from speeding. It is interesting how the immediate cause here is only present in 2 out of 10 accidents. A possible explanation could be that these accidents being caused by speeding were so severe that there were no questions about the cause. In order for 3 people to die inside a city in a road traffic accident, the car in question must have traveled way

over the speed limit making it easy to determine the cause of the accident. Other causes here include two casualties regarding alcohol consumption and poor road conditions each.

Inspecting the bottom half of the figure, the top twenty accidents are listed with the highest number of vehicles involved in accidents. The highest number of vehicles is 18 followed by 17 and 16. Interestingly, the joint first accident on this list was caused by drug or medicine consumption. 9 out of the first 20 accidents had no immediate cause, and when the cause could be determined, it was mostly alcohol consumption and speeding.

5.2.6 Trends

Tableau provides an option for predictive modeling: forecasting. It uses a technique called exponential smoothing. Exponential smoothing models iteratively predict future values of a time series of values from weighted averages of past values [18]. It is called exponential because every value is influenced by every preceding value with exponentially decreasing degrees, more recent values have a greater weight in the calculation of the forecast. Exponential smoothing models with trend and seasonality are effective when the data exhibits trends (growing or decreasing over time) or seasonality (repeating variation in value over a certain period of time). Road traffic accidents both exhibit trends (in the case of Barcelona a slight increase through the 2010s) and seasonality (every August the number of accidents are significantly lower than in other months).

5.2.6.1 Tableau forecasting

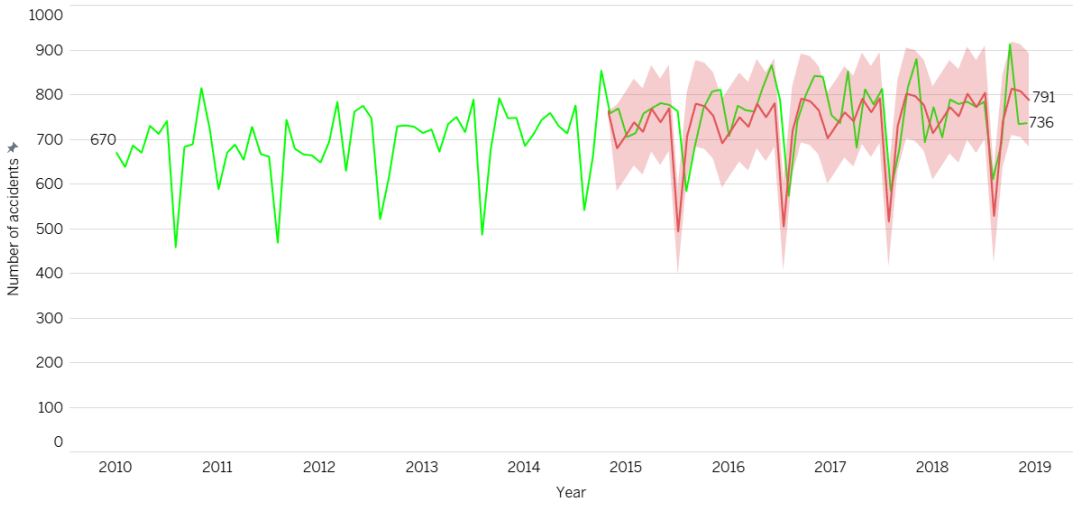

Figure 5.29: Comparison of Tableau forecast with actual numbers

There are different model types one can use in Tableau. The two options for trend and seasonality are additive and multiplicative. An additive is one in which the contributions of the model are summed, and a multiplicative is one in which at least some component contributions are multiplied.

The level of granularity in the data in this scenario is in months. I used the period from 2010 to 2014 as input for the model, on which the prediction is based. I then asked Tableau to predict the time series for the next four years, until the end of 2018. I chose this time period because in 2020 a huge anomaly was present: the coronavirus pandemic.

The numbers decreased significantly, falling completely out of the pattern, and potentially ruining the prediction. I used an additive model for both trend and seasonality. The forecasting (red) compared to the actual values (green) can be seen in figure 5.29. The pale red area shows the prediction intervals, the model having a 95% confidence that the values will fall into that area in the prediction. We can see that the model clearly picked up on the seasonality present in Augusts. A slight increase over time is predicted as well. Now it is time to numerically evaluate the model.

Options Used to Create Forecasts

Time series:	Month of Date
Measures:	Number of accidents
Forecast forward:	48 months (January 2015 – December 2018)
Forecast based on:	January 2010 – December 2014
Ignore last:	No periods ignored
Seasonal pattern:	12 month cycle

Number of accidents

Initial			Change From Initial	Seasonal Effect				Contribution		
January 2015			January 2015 – December 2018	High		Low		Trend	Season	Quality
695	±	86	96	October 2018	47	August 2018	-206	3.1%	96.9%	Ok

All forecasts were computed using exponential smoothing.

Number of accidents

Model			Quality Metrics					Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Additive	Additive	Additive	44	35	0.44	5.0%	488	0.010	0.000	0.225

Figure 5.30: Tableau forecast options and model

In the figure above, the forecast options and model tabs can be seen from Tableau. There is various information that can be seen about the forecast, some more straightforward, others not so clear at first glance. The upper part of the figure describes the general characteristics of the forecast model such as what granularity of the data is at, the length of the forecast, the length the forecast is based on, and the length of the seasonal pattern. The middle part of the figure shows how the prediction numbers have changed, how they fluctuated, what seasonal effect they had, and how important the trend and seasonality were in the model. This last metric is quite interesting, 96.9% was the importance of seasonality in the prediction. What this says is that trend was much less of a factor when taking into account the two aspects. In the bottom part, numerical quality metrics, and other information can be seen such as the type of the model and smoothing coefficient. In the quality metrics, RMSE is short for root mean squared error, MAE for mean absolute error, MASE for mean absolute scaled error, MAPE for mean absolute percentage error, and AIC for Akaike information criterion. Scale-dependent metrics are not suitable for comparing different time series meaning RMSE and MAE are not good indicators of model quality in this example. For this reason, I will be looking at MASE and MAPE in the first place.

MASE is calculated by first computing the MAE and then dividing it by the MAE of an in-sample naive benchmark. MASE practically tells how the model performs compared to an in-sample one-step forecast from the naive model. A MASE value of 1 means that the model is doing neither better nor worse than the naive model, which is objectively not a

good thing. Why bother making a model if it performs the same as a naive one? A MASE value below 1 says that the model performs better than a naive model and a value above 1 means the model is even worse than a naive one. With that said, a MASE value of 0.44 tells us that the model more than doubled the accuracy compared to the in-sample naive one. This result is acceptable as stated by Tableau (Quality is Ok).

MAPE is calculated by taking the average (mean) of the absolute difference between actuals and predicted values divided by the actuals. MAPE can be interpreted as the inverse of model accuracy. More specifically it is the average percentage difference between actual and predicted values. Having a MAPE of 5%, in this case, tells us that on average, the predictions are 5% off from the actual values that the model aimed for [1]. As per the cited website, a MAPE below 10% tells us that the model is *very good*.

Chapter 6

Conclusion

In this final chapter, I look back at the key findings of the paper, some difficulties, and limitations I encountered as well as opportunities for further development and future research.

6.1 Key findings

Throughout the paper, I went through all the phases of data analysis: data collection, preparation, exploratory data analysis, and using data visualization to thoroughly analyze the demographics and road traffic accidents in Barcelona.

To begin with, I demonstrated the capabilities of the Pandas Python library as well as several different third-party libraries during the data preparation/cleaning phase. I managed to go from an inconsistent, incorrect, inaccurate set of datasets in Catalan language to a consistent set of logically merged, cleansed, imputed, and translated sets of datasets that contained little to no irrelevant data for my subject of analysis.

Next, I put the capabilities of the business intelligence tool called Tableau to display whilst creating interactive graphs and maps that helped me during the exploratory data analysis as well as the in-depth analysis.

During the last phase, I conducted a thorough analysis of the demographics and accidents in Barcelona. Looking at the demographics included establishing that the city has an aging population with a constrictive age composition, as well as seeing in detail the difference in education spatially, and also in terms of gender. The districts of Barcelona could be clustered in terms of educational performance. I then looked at the trends where it was clear, that education was getting better throughout the 2010s in each educational cluster. Moving to road traffic accidents, I first looked into the time distributions of occurrences. I showcased the power of interactive dashboards by being able to answer complex questions in just seconds.

Following this, I developed a subjective method of ranking districts in terms of danger, taking into account underlying aspects that could influence the number of accidents such as area, registered vehicles, and population of a district. I also gave importance to accident severity, as I used different weights for accidents with a different number of fatal, seriously injured, or lightly injured casualties. I called the final metric *Normalized index*, which resulted in a different ranking compared to just the ranking by the number of accidents. Although there are a lot of different factors that need to be taken into account, this metric can give a better ranking in terms of danger for districts.

In the next sections, I looked at certain aspects and their combinations. A good example of a key finding was the analysis of deaths, and seeing how different the spatial, time, and gender distribution is in that category compared to the overall distributions. There were numerous interesting other findings in section 5.2.3, including the gender distribution of types of vehicles used by the driver as well as a driver with a fatal injury. A key following aspect was the comparison of the age composition with accident participation age distribution. A key aspect was also derived in the next section, in people between the age of 20 and 35: despite making up only 20% of the population, they represent 40% of people involved in accidents.

The analysis phase was finished by first looking at the biggest accidents in terms of overall victims, fatal casualties, and vehicles involved and then experimenting with the forecasting capabilities of Tableau.

6.2 Limitations and challenges

During the project, I encountered numerous difficulties and limitations. At the start, data cleaning was the most challenging, as there were multiple steps I had to make to reach a consistent state. I overcame most problems, however, there was something that I was not able to do. As I mentioned in the designated data preparation chapter, in the accidents by people collection, between 2014 and 2015, the value range of the hour column is from 1 to 12, whereas in all the other years, the value range is from 0 to 23. On top of this, the part of the day column only contained one value throughout all the records, which made it impossible to deduct the correct hour the accident happened between 2014 and 2015. This resulted in me having to avoid using (filtering the two years out) in visualizations where I presented time-related information, more specifically about hour distributions. A challenge was also to validate the quality of the data, as I did not account for the seemingly unrecognizable defects of the miscollection of said data. I eliminated the most defects, however, I am sure that others are hard to see, even when diving deep into the qualities, and underlying aspects of the data.

Another limitation was that when looking at accident causes, around 96% of accidents were labeled with *No immediate cause*. This resulted in me having to exclude 96 thousand accidents when looking at direct causes, and only being able to analyze 4 thousand between the years 2010 and 2021. Since the available data sample became less, the analysis is not as accurate.

When looking into accident participation by age and gender, I can not directly make comparisons between age groups and genders, as there is a key factor that needs to be taken into account: distance driven. It is a very important normalizing factor since, without this metric, the amount of participation in accidents does not say enough. In my scenario, around 61% of people involved in an accident were male. I can not directly say that this means men are more careless when it comes to traffic, because the distance driven needs to be taken into account. I, unfortunately, did not have access to such data, which means my analysis is not complete, as it remains a big challenge to explore datasets that are giving additional value to the analysis. I can envision a lot of other underlying aspects, which would be beneficial to use, but it is unrealistic to include every one of them, even if they are manifested in datasets.

6.3 Future research

I have a lot of ideas moving forward with the project, with the cleaned, consistent data at hand, it is very easy to find domains that can further utilize it. One of these domains can be prediction or forecasting. By creating negative samples, and using them to create classifiers or machine learning models, we can answer such questions or problems about whether a certain situation, at a time and place, with certain characteristics, can be an accident or not. This can lead to implementing systems such that they predict the probability of an accident when choosing a destination to travel to.

Another interesting thing to explore could be the usage of time series prediction with machine learning models (such as *Random Forest Classifier* and *Logistic Regression*) and comparing the performance of said models with the forecasting functionalities of Tableau.

There is an algorithm called DBSCAN, which stands for density-based spatial clustering that can be used to utilize the spatial aspect of the data. It essentially clusters accident locations and also selects some (that belong to none of the clusters) as random noise. The clusters can be the accident hotspots, where the algorithm tells us that there is more to the location than just a simple chance that increased the numbers in that area. Determining accident hot spots in a city, essentially selecting the most dangerous intersections can tell the governing officials that there is something wrong with the infrastructure of the city. This has the potential to save lives, as it is direct feedback on the road infrastructure of a city, and with the right amount of change or influence, the intersection can be modified to be safer.

Bibliography

[1] Stephen Allwright. How to interpret mape. https://stephenallwright.com/interpret-mape/.

[2] London Assembly. Bus driver numbers. https://www.london.gov.uk/questions/2014/1209.

[3] Tobias Bieniek. utm project description. https://pypi.org/project/utm/.

[4] Lauren Boucher. What are the different types of population pyramids? https://populationeducation.org/what-are-different-types-population-pyramids/, 2016.

[5] Right Casa. Young adults in spain can now drive from the age of sixteen. https://rightcasa.com/young-adults-in-spain-can-now-drive-from-the-age-of-sixteen/.

[6] Kenneth L. Campbell Dawn L. Massie. Analysis of accident rates by age, gender, and time of day based on the 1990 nationwide personal transportation survey, February, 1993. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/1007/83596.0001.001.pdf?sequence=2.

[7] Ajuntament de Barcelona. Transport by motorcycle. https://www.barcelona.cat/mobilitat/en/means-of-transport/motorcycle.

[8] SuHun Han. Googletrans project description. https://pypi.org/project/googletrans/.

[9] JetBrains. Dataspell - the ide for professional data scientists. https://www.jetbrains.com/dataspell/, 2021.

[10] Tom Brijs Koen Vanhoof Karolien Geurts, Geert Wets. Identifying and ranking dangerous accident locations: Overview sensitivity analysis. https://www.academia.edu/20602268/Identifying_and_ranking_dangerous_accident_locations_Overview_sensitivity_analysis.

[11] Neurosciencenews.com. Loss of male sex chromosome leads to earlier death for men. https://neurosciencenews.com/y-chromosome-loss-death-21049/.

[12] NumPy. Numpy documentation. https://numpy.org/doc/stable/.

[13] The Urban Mobility Observatory. Barcelona is to ban older cars and vans on days of high pollution. https://www.eltis.org/in-brief/news/barcelona-ban-older-cars-and-vans-days-high-pollution.

[14] Travis Oliphant. Scipy documentation. https://docs.scipy.org/doc/scipy/.

[15] World Health Organization. Ageing: Global population. https://www.who.int/news-room/questions-and-answers/item/population-ageing.

[16] Pandas. Pandas documentation. https://pandas.pydata.org/docs/.

[17] Tableau Software. Tableau public community. https://www.tableau.com/community/public.

[18] Tableau. How forecasting works in tableau. https://help.tableau.com/current/pro/desktop/en-us/forecast_how_it_works.htm, .

[19] Tableau. Use relationships for multi-table data analysis. https://help.tableau.com/current/server/en-us/datasource_multitable_normalized.htm, .

[20] Tableau. Workbooks and sheets. https://help.tableau.com/current/pro/desktop/en-us/environ_workbooksandsheets.htm, .

[21] Wikipedia. Anaconda (python distribution). https://en.wikipedia.org/wiki/Anaconda_(Python_distribution), .

[22] Wikipedia. Exploratory data analysis. https://en.wikipedia.org/wiki/Exploratory_data_analysis, .