Project Title

First Author1, Second Author2, Third Author3 & Fourth Author4

1first author student number and e-mail contact

2second author student number and e-mail contact

3third author student number and e-mail contact

4fourth author student number and e-mail contact

Abstract**:** maximum of 250 words, font Times New Roman, size 10, line spacing 1.0

Keywords: maximum three keywords separated by semicolon  
Statement of Contribution: clearly state the contributions of each group member to the project in particular who performed the following: Data collection, Data Cleaning, Data Analysis, Wrote the report, Design the study, Discussion of the Results.

# Introduction

Motor vehicle accidents can have a high impact on our everyday life. Not only regarding costs for the public infrastructure but also through traffic and most importantly when we consider the persons involved, which can get wounded. Solely in Portugal, in the last years, there has been an average of more than 30 thousand accidents per year that cause more than 40 thousand wounded and 400 deaths (ref 1). For these reasons, many cities around the world have an increased interest in understanding not only the causes that lead to accidents but also the main factors that cause the occurrence of wounded passengers and pedestrians. One of those cities is the Portuguese capital Lisbon, which launched a competition to not only pinpoint and evaluate the locations where accidents occur but also to infer which characteristics lead to wounded persons.

However, accident occurrence is a complex problem, where many factors play a part. Normally, these factors are divided into two groups. One of the groups focuses on driver characteristics, such as skill level, experience, risk-taking behaviors (cellphones usage for example), and age, while the other group is focused on external conditions such as quality of the road, traffic, atmospheric conditions number of persons in the vehicle, and vehicle type.

Several studies have been made where different factors of both these groups are evaluated (ref2), however, this type of evaluations normally presents some limitations.

For example, in Portugal, you have the option of reaching a mutual agreement when an accident occurs. This type of accident is only reported to the insurance companies and for this reason, never reach the police and fireman reports. If one just has access to the police and fireman accident reports this can lead to a non-representative sample of the accidents that occurred in the city and may lead to incorrect conclusions.

Another limitation is regarding the driver characteristics and behavior. Although most of the external factors can be obtained and evaluated easily, some of the driver's characteristics can be difficult to obtain. How to evaluate if a driver is inexperienced or not? And how to get information regarding the behavior before the accident. Was the driver talking on the cellphone, for example?

Nonetheless, and even with these limitations one can evaluate the police and fireman accident reports and reach some conclusions, and that's exactly what we try to do in this project.

# Data

## Description and Extraction

For this study, we used several datasets supplied by LxDataLab focused on the accidents that occur in Lisbon, several city characteristics, and traffic jams. The data sets where:

* Accidents in Lisbon registered in 2019 by the “Autoridade Nacional de Segurança Rodoviária” (ANSR)

This data set has multiple information regarding the accident characteristics, vehicles and persons involved.

* Lisbon accidents registered in 2019 by the “Regime de Sapadores Bombeiros” (RSB)

This data set had only information regarding the time, type, and location of the accidents.

* Lisbon street height, slopes, crossings, and traffic lights

This data set had multiple information regarding street characteristics.

* Lisbon traffic jams in 2019

This data set had information regarding the city traffic jams measured by waze in the year 2019.

## Transformation

We started by evaluating the accident data consistency for errors and misclassifications[[1]](#footnote-1). Has previously stated the RSB data had little information, however, the ANSR accident presented a high number of descriptive variables regarding the accidents that had to be evaluated.

To evaluate the accident data, we needed to create a single data set, however, we had some limitations since some of the accidents in the ANSR data set didn’t have geolocation.

1. For this reason, and to allow different evaluations we decided to divide the accidents data into three data sets: Accident data without geolocation and considering only the general accident data (date and type)
2. Accident data with geolocation and considering general accident data (date, type and location)
3. Accident data with only the ANSR descriptive data;

Next, we evaluated the remaining data sets namely the supplied shapefiles (Lisbon street information) and traffic jams data performed the required cleaning.

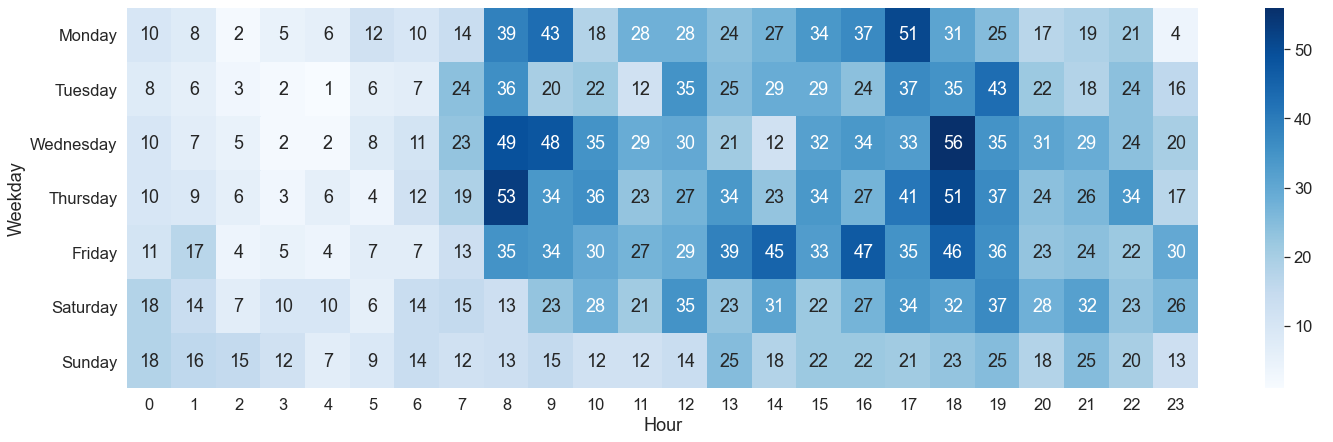
Finally, when performing geolocation evaluation, we needed to aggregate the data by some means. For this purpose, we used the python package “h3” developed by Uber which convert each geolocation to an unique hex, allowing the aggregation of our accident, traffic jam and street characteristics.

VALTER E GONÇALO CONFIRMEM ESTA PARTE

# Results and Discussion

1. **Accident occurrence over time**

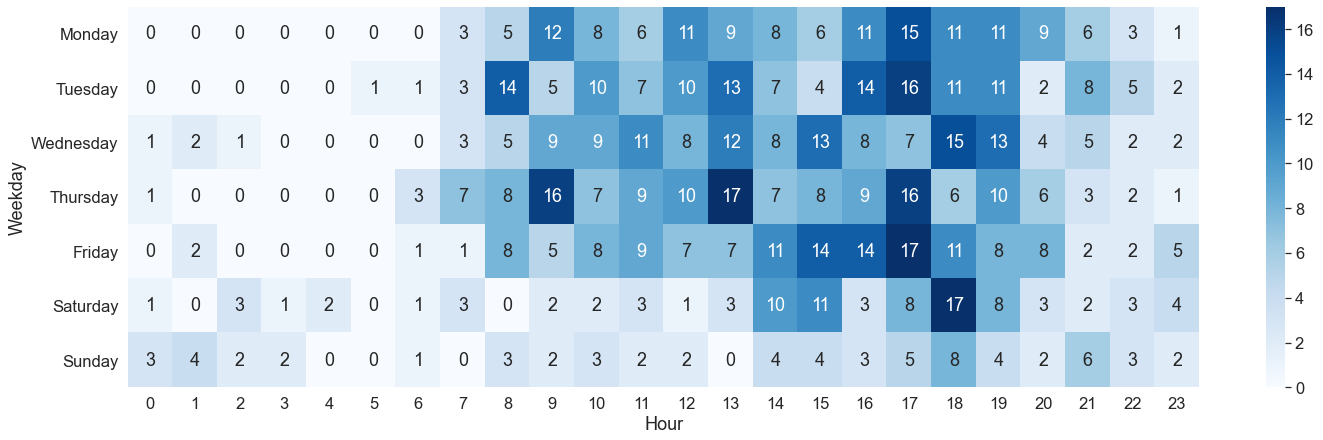
We started by asking the question: “How does the accident rate vary over time?”. The premise was that there were times during the day/week that had a higher rate of accidents. To evaluate this, we aggregated all the 2019 collisions/crashes and runovers, by hour and weekday.



**Fig. 1:** 2019 number of collisions and crashes per hour and weekday

When we evaluate the heatmap with the collisions/crashes over a week we can see several patterns:

* There is a higher concentration of accidents during the working days hours, especially around the rush hours. This intuitively makes sense since a higher concentration of cars may lead to the occurrence of more accidents;
* The occurrence of accidents during the weekends presents a different behavior. On Saturday the concentration of accidents occurs more during the morning and end of the afternoon, probably due to people going for a walk in the city. On Sundays, the number of accidents is much lower and is highly concentrated in the afternoon;
* Another interesting fact is that we can see that the occurrence of accidents during Friday and Saturday nights is much higher than during the week. One possible explanation for this is that during these days a lot of persons go out at night in Lisbon;



**Fig. 2:** 2019 number of run overs per hour and weekday

When evaluating the run over heatmap we can see that:

* Like the collision/crashes data most of the runovers happen during the working day hours, especially at 17 o'clock;
* During the weekends most of the run overs happen after lunch which is normally when people are going for a walk;

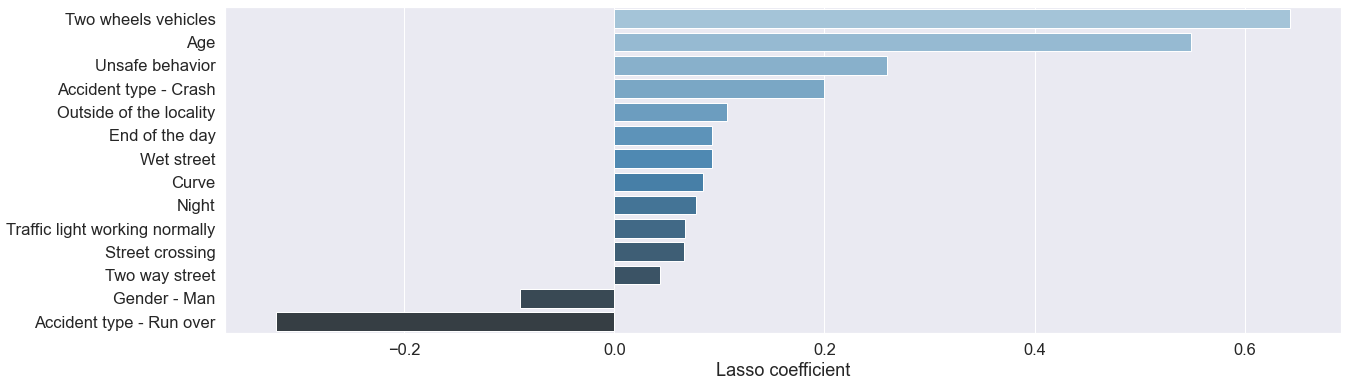
An important point is that although we make several assumptions using intuition, namely that more traffic causes more accidents, we need to dive deeper into the subject and determine the root causes leading to accidents.

1. **Accident characteristics that lead to wounded drivers and passengers**

The main objective of this point was to evaluate the ANSR police reports and determine which accident characteristics lead to wounded drivers and passengers. We mainly focused on driver, atmospheric, and road conditions to perform this evaluation[[2]](#footnote-2).

To perform this analysis, we used the cleaned ANSR data frame and took several steps:

* Converted the column “Lesões a 30 dias” to a binary variable ad use it as our objective label. If a person was wounded it would get a value of 1, if not of 0 meaning we don’t differentiate between lightly or badly wounded and deaths;
* Converted all the accident characteristics to dummy variables;
* Applied a lasso regression using our dependent and independent variables. Lasso regression was selected due to its simplicity and because this regularization method tends to convert the coefficient of the variables that are not important to 0;
* Evaluate the obtained lasso regression coefficients to determine which factors have a higher influence on the occurrence of wounded drivers and passengers.



**Fig. 3:** Lasso coefficients for different factors

As can be seen from the graph most of the factors increase the occurrence of wounded drivers and passengers, and only two reduce it.

* The factor that most increases the occurrence of wounded is if you travel in a two wheels vehicle. This makes sense since accidents with two wheels vehicles normally lead to a fall that can be at high velocities;
* Regarding the driver and passenger characteristics, we have several factors that increase the occurrence of wounded, such as a higher age or low-security behavior (such as not using a helmet or safety belt). On the other hand, we can also see if the driver or passenger is a man the occurrence of wounded tends to be lower;
* When we evaluate the atmospheric conditions, we see that being late afternoon or night also influences positively the occurrence of wounded;
* The type of accident also influences the occurrence of the wounded. Crashes tend to increase the occurrence t of wounded, while runovers lower the occurrence of driver and passengers wounded;
* Lastly, some of the street characteristics also increase the occurrence of wounded. Namely, if the accident occurs in a curve or a crossing, if it is outside a locality, and if the streets have two ways.

Although, many conclusions can be taken from this study further evaluations need to be performed to estimate the impact of these characteristics. However, they are outside of the scope of this work.

# Conclusions

# Acknowledgements

Use this section to acknowledge the contribution of a third party for your project, for instance additional mentoring or data acquisition, or if you are using data from your company for instance.

# References

1 pordata: <https://www.pordata.pt/Portugal/Acidentes+de+via%C3%A7%C3%A3o+com+v%C3%ADtimas++feridos+e+mortos+++Continente-326>

2 ipma site para os dados de precipitação

Footnote 1: see cleaning data notebook

The report should have no more than 3500 words, plus Figures, Tables, and max of 30 References.

Figures and tables should be correctly enumerated and have a self-explanatory caption.

The font should be Times New Roman with size 10 and line spacing of 1.5. Text must be justified.

Follow the Physical Review Style for references, <https://cdn.journals.aps.org/files/styleguide-pr.pdf>.

Consider using a reference managing tool such as Mendeley, Zotero, or Read Cube.

References should be done using the APA style and must reference meaningful sources (don’t cite blogs for instance, consider instead citing research articles). Maximum of 15 references allowed.

References made to particular content in the Jupyter Notebook should be done as a footnote[[3]](#footnote-3).

Don’t forget to correctly specify the original data sources, and correctly credit their authors.

Consider sharing your Python materials and all the resources you used for this project in a public GitHub repository.

1. Check Jupyter notebook: 0.0-data-preparation.ipynb [↑](#footnote-ref-1)
2. Check Jupyter notebook: 0.2-accident-police-evaluation.ipynb [↑](#footnote-ref-2)
3. Check Jupyter Notebook XXX [↑](#footnote-ref-3)