NOVA IMS Information Management School

Lisbon accident evaluation:

Which factors lead to accidents and wounded drivers and passengers

Fábio Oliveira

m20201057@novaims.unl.pt

Gonçalo Duarte

m20201329@novaims.unl.pt

Válter Frade

m20201052@novaims.unl.pt

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

CONTEXT

- Portugal 2017-2020 - more than 30k accidents, 40k wounded and 400 deaths per year

Many factors can influence these incidents

(Road profile, atmospheric conditions, drivers, ...)

Need for DS approach to manage, transform, analyse and present results

LxDataLab challenge:

Identify high risk zones and influence of external factors in Lisbon and

Identify causes for wounded drivers/passengers

STRATEGY

Total of 7 datasets regarding Lisbon in 2019:

2 - accidents' info (ANSR and RSB);

4 – Road profiles (ANSR);

1 – traffic jams (Waze).

Cleaning and agglomeration of the different data sources in order to allow **4 distinct analysis**:

- group accidents through a grid system to identify highest incident zones (geopandas, folium and H3: Uber's system);
- road profile features and traffic by zone and correlation to the previous analysis (spearman test);
- identification of traffic jam peaks;
- accident characteristics leading to wounded drivers and passengers (lasso regression).

STRATEGY

Expected outcomes:

- Identification of high frequency accident zones (city center);
- Crossroads, slopes and traffic impacting accident frequency positively;
- Peak traffic during rush hours;
- Driving at night and non cautious behavior can increase the likelihood of wounded people.

CHALLENGES

Data:

- Some accidents didn't have geolocation → creation of specific sub-dataset for geographic based analysis;
- Crossroad and traffic light data were very similar → analysis of both datasets.

Grid system:

- Grid resolution selection → balance between detail and generalization.

Feature engineering:

- Associate accidents to traffic delays→ find the nearest traffic location close to 10 minutes before the incident.

GEOGRAPHIC ANALYSIS – ZONES OF HIGHEST NUMBER OF ACCIDENTS

- TOP 3 zones with the highest number of accidents:

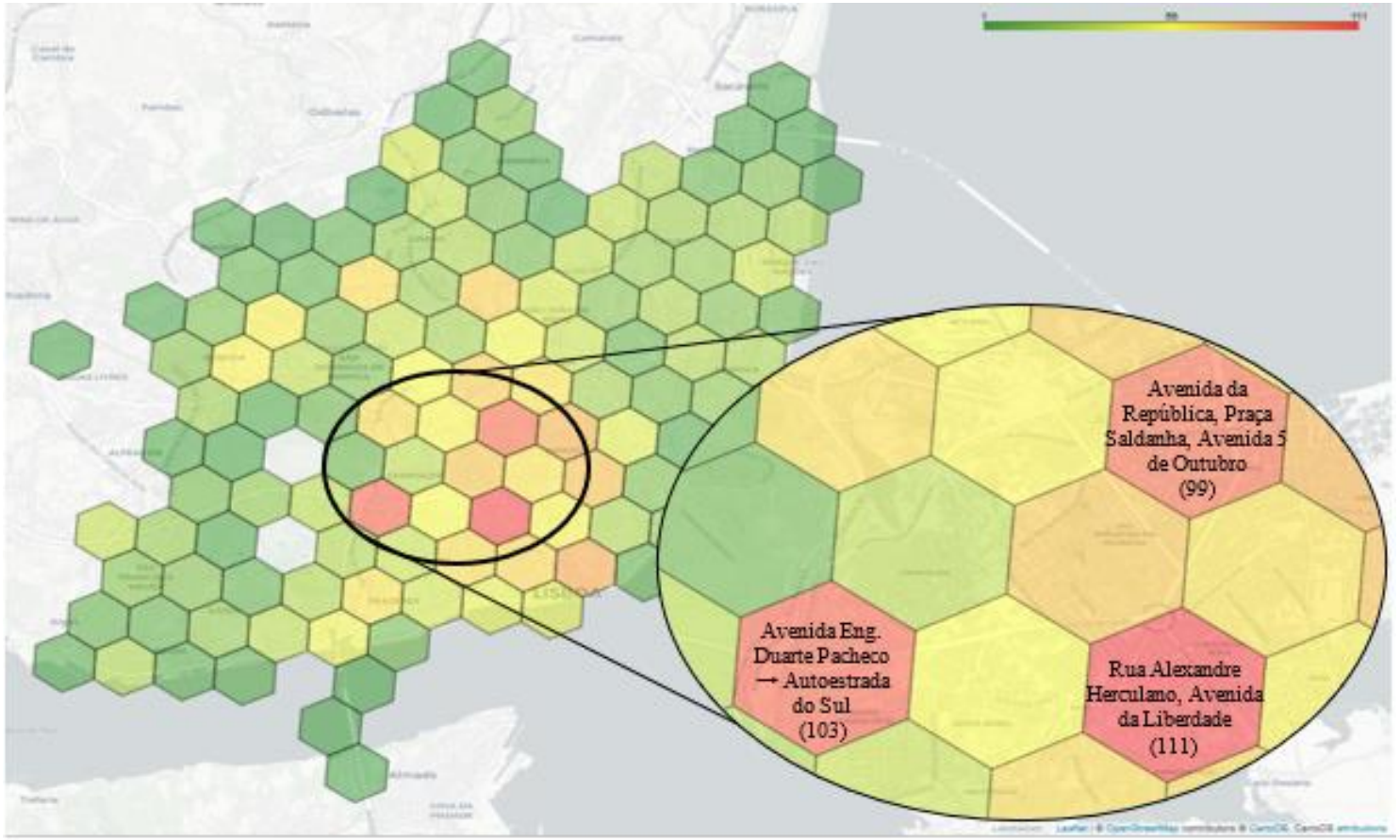


Fig. 1: Geographic representation of the H3 hex cells, with resolution=8, colored by the number of accident occurrences that ranges from 1 (green) to 111 (red).

GEOGRAPHIC ANALYSIS - CROSSROADS

Moderate positive correlation

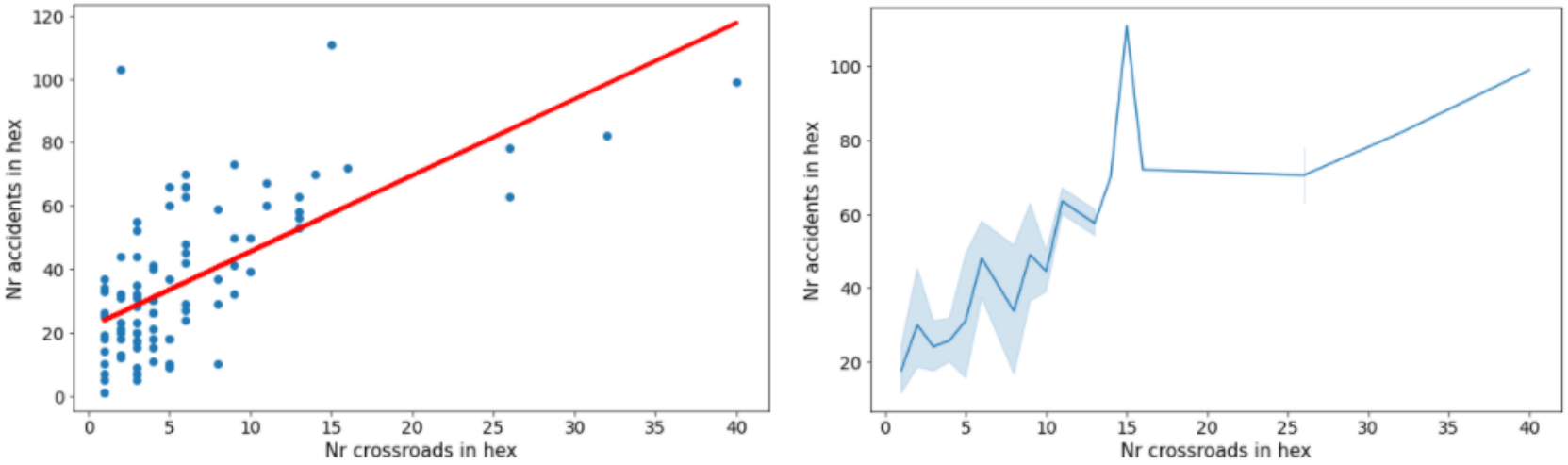


Fig. 2: Scatter plot of the number of crossroads and accidents per hex with the best linear fit (spearmen coefficient= 0.63, p-value=1.67×10^{-11}) and respective line plot.

- Street elevation and slope analyses didn't show a correlation with the number of accidents.

TRAFFIC CHARACTERIZATION

Highest traffic levels – January to April and during 8 and 18 o'clock

Lowest traffic levels – August and during 4 and 5 o'clock

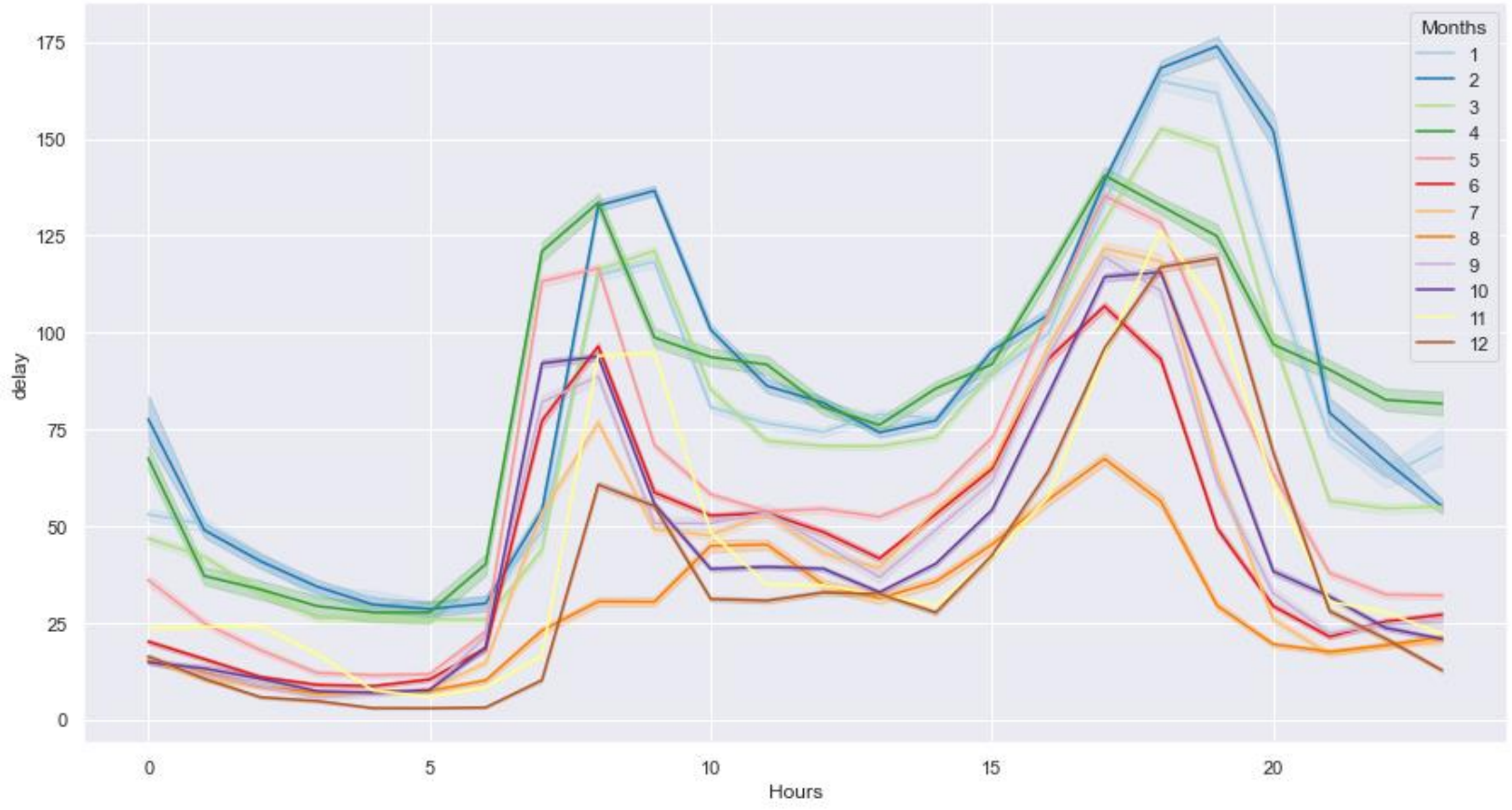


Fig. 3: Hourly profile of Lisbon traffic grouped by month.

CHARACTERISTICS LEADING TO WOUNDED DRIVERS AND PASSENGERS

Positive influence – two-wheeled vehicles, age, unsafe behavior, among others

Negative influence – gender “man” and type of accident

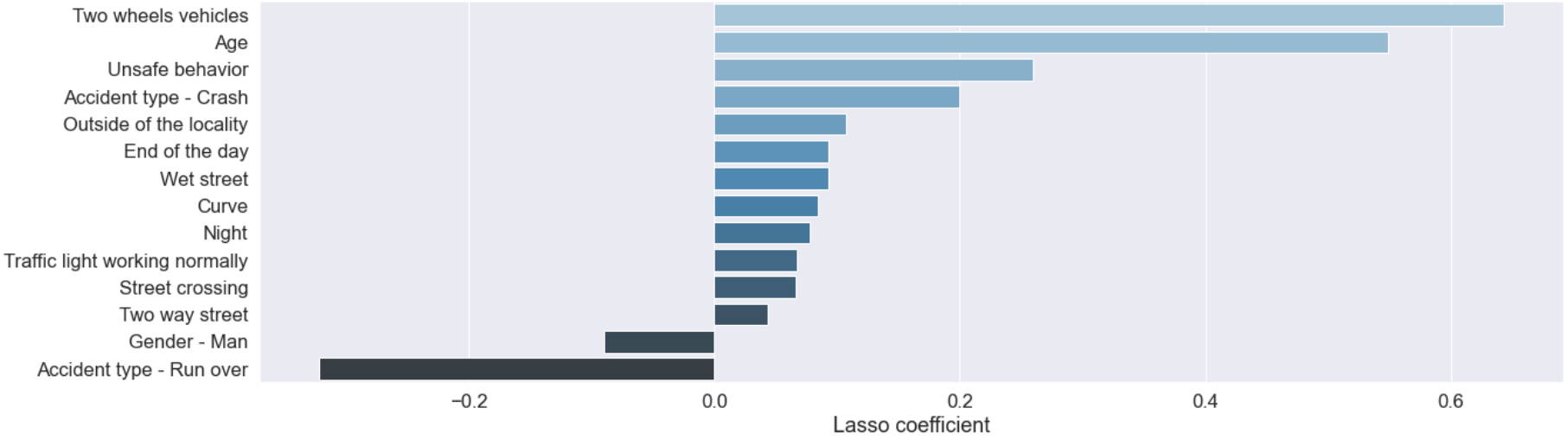


Fig. 4: Lasso coefficients for different factors.

CRITIQUE

Typical **Data Science** study case:

- data management (pandas) ;
- data cleaning;
- statistical analysis (SciPy);
- data visualization (matplotlib, seaborn);

+ handling of geospatial data (geopandas, H3 and folium).

Challenge allowed **exploratory data analysis** and **critical thinking**

KEY FINDINGS

1. Classification of 3 main areas with a higher number of accidents:
   - Rua Alexandre Herculano + Avenida da Liberdade;
   - Avenida Eng. Duarte Pacheco → Autoestrada do Sul;
   - Avenida da República + Avenida 5 de Outubro;

2. Traffic characterization of Lisbon over 2019:
   - Highest traffic levels – January to April and around 8 and 18 o'clock
   - Lowest traffic levels – August and around 4 and 5 o'clock

3. Factors that can increase the likelihood of accidents/ wounds:
   - Crossroads;
   - Two-wheeled vehicles;
   - Age;
   - Unsafe behavior.

NOVA IMS Information Management School

Q&A

CONVERSATION STARTERS

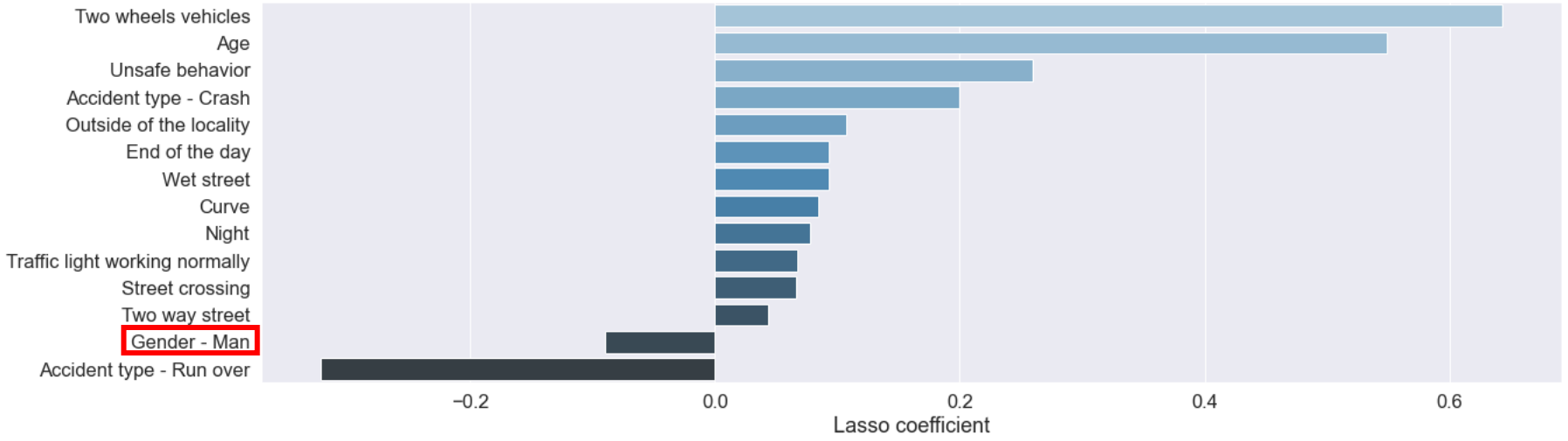
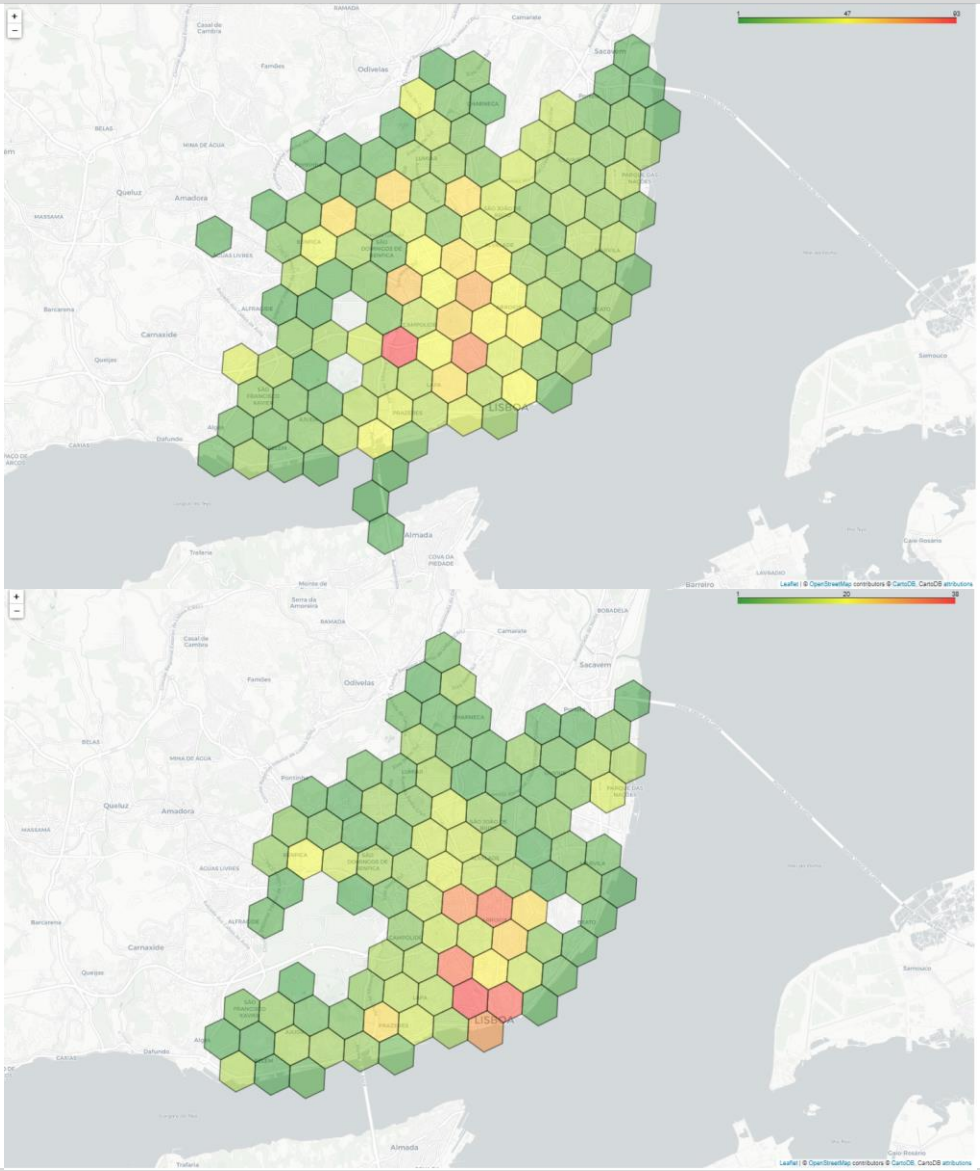


Fig. 4: Lasso coefficients for different factors.

CONVERSATION STARTERS



CONVERSATION STARTERS

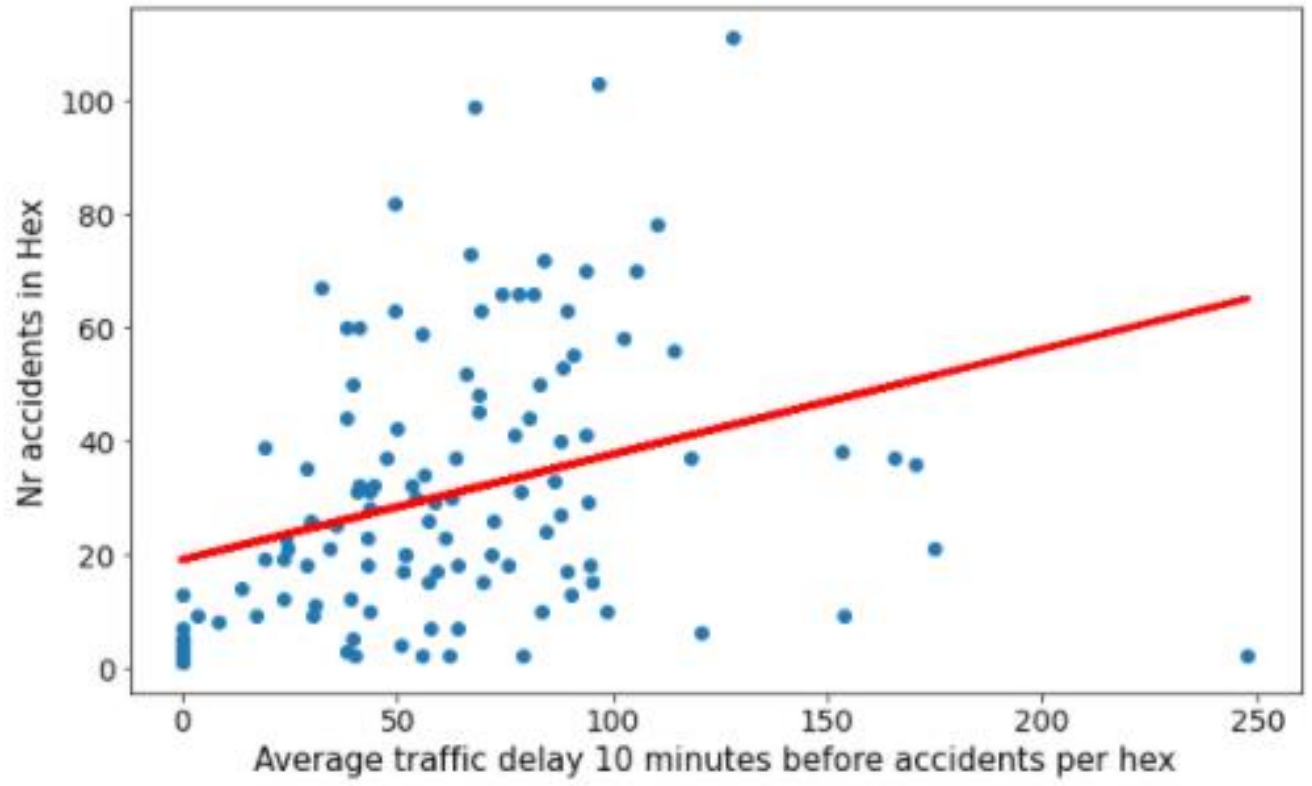


Fig. 6: Scatterplot of the average traffic delay 10 minutes before an accident and the number of accidents per hex with the best linear fit (spearmen coefficient= 0.43, p-value=$8.48x10^{-7}$)

CONVERSATION STARTERS

CONVERSATION STARTERS