Analysis of Road Accidents in Catalonia

Iana Pokrovskaia

October XX, 2020

1. Introduction

According to ACEA (Europe Automobile Manufacturers Association), there are 312.7 million motor vehicles in circulation on the EU's roads, which is more than one to every two Europeans. For a long time, cars have been an important part of our daily life, and it is not going to change anytime soon. Many types of public transport use the same roads as cars. As a result, the number of people driving the roads every day is enormous. This leads to the high importance of road safety since road injuries are one of the most common causes of accidental death. However, improving road conditions also requires a lot of resources, which makes it necessary to take a systematic approach to determine which roads are more dangerous and need immediate attention.

An effective way to define the actual riskiness of the specific road is to analyze data about road accidents. This project aims to find features that affect road safety negatively, weigh their importance, conduct a statistical analysis of the current state of the roads, as well as suggest options for improvements.

It is obviously interesting for car drivers and citizens in general. Besides, this analysis can be useful for people responsible for future road improvements. The traffic police also may be interested in this analysis, since it identifies the conditions that require additional attention.

2. Data acquisition and cleaning

2.1 Data sources

For this analysis, I have used Catalonia Road Traffic Injuries & Deaths dataset from Kaggle (link), which consists of data from 2010 to 2020. It provides a lot of data about car collision circumstances, including the types of the road intersection, weather/light conditions, allowed speed limit and whether it was exceeded, date and time of the accident, and more.

2.2 Data cleansing

Firstly, I translated the dataset from Catalan to English for better understanding. After basic analysis, I captured a lot of missing values in some additional attributes like allowed speed, regulation of priority, type of section and so on. I decided to replace these values with “Not

specified", since there is a lot of 'Not specified' values in the dataset and it holds the same meaning. Furthermore, I changed the datatype of allowed speed column from float to string, because it presented not a continuous value, but a list of options in numeric format. Also, in the 'ALLOWED_SPEED' column I found the value 999, which could not be correct, so it was replaced with 'Not specified'.

In all other respects, the dataset was perfect for further processing.

2.3 Feature selection

This dataset has a huge number of features, so an important task is to sort essential features from irrelevant ones.

To highlight important features, I analyzed the number of unique values and found out that for columns 'ROAD', 'CITY', 'TIME', 'KM', their number is 679, 854, 1298 and 2150 respectively. Since the total number of entries in this dataset is 16774, there is no benefit in using these features data during analysis. These columns have been deleted. There were also a large number of unique values in the 'DATE' column. The exact date of the accident is not suitable for further analysis, but information about the month may be useful. Therefore, based on the 'DATE' column, the 'MONTH' column was formed, and the 'DATE' column was deleted. The 'YEAR' column was deleted too.

I decided to remove the features containing data that can be obtained only after an accident. Although the data from the 'N_DEATHS' and 'N_MAJ_INJURIES' columns will be used for the subsequent analysis to determine the hazard level of the road, the rest of such features have no practical use for this project. This list includes columns 'N_VICTIMS', 'N_VEHICLES', 'N_PEDESTRIANS' and so on. 'FOG_INFLUENCE', 'TRAFFIC_INFLUENCE' and similar columns were also removed, and the same for 'ACC_TYPE', 'ACC_CLASSIFICATION' and 'SPEED_LIMIT_DISPLAY'. The 'WEEKDAY' column contains information from the 'WORKING DAY' column, so 'WORKING DAY' can be deleted too.

To determine the strength of correlation between values, it is necessary to convert them to a numerical format. A temporary dataset was created using LabelEncoder from sklearn library. The obtained values of the Pearson coefficient are shown in Figure 1.

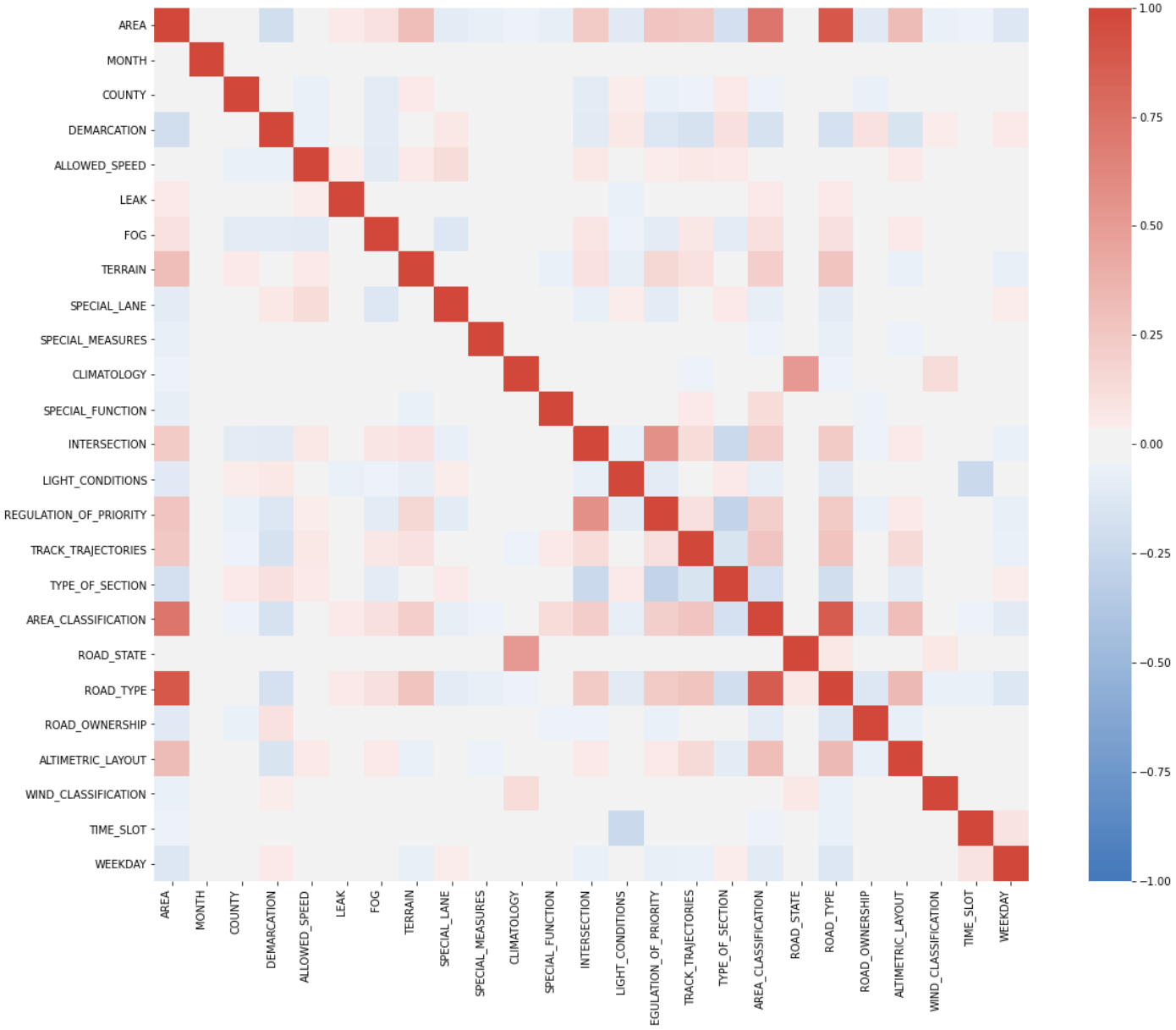

Figure 1. Pearson coefficient heatmap

As can be seen on the heatmap, there is practically no correlation between the values (the red diagonal line corresponds to identical features). The found relationships with the absolute value of the Pearson coefficient exceeding 0.7 are given in Table 1. P-values in all the cases are exactly 0.0.

Table 1. Features with the high value of Pearson coefficient

First feature	Second feature	Pearson coefficient
'AREA'	'AREA_CLASSIFICATION'	0.723
'AREA'	'ROAD_TYPE'	0.884
'AREA_CLASSIFICATION'	'ROAD_TYPE'	0.873

Thus, all three features are related. Values of ‘AREA’ and ‘AREA_CLASSIFICATION’ columns are close: ‘Urban area’ and ‘Road’ for the first and ‘Urban area’, ‘Road’ and ’Crossing’ for the second. Thus, the ‘AREA’ column can be dropped. The ‘ROAD_TYPE’ column can be dropped, too, since most of the values divide into two groups: ‘Urban road’ and ‘Conventional road’, which is directly related to the area type.

After all, 23 features were selected.