**Accident Prediction**

**Daniil Litvinov**

**October 1,2020**

**1. Introduction**

**1.1 Background**

Road traffic accidents are extremely common. Because of their frequency, road traffic accidents are the leading cause of death worldwide, claiming millions of lives per year. Thus, a system that can predict the occurrence of road traffic accidents or crash-prone areas could potentially save lives.

**1.2 Problem**

Although difficult, predicting an accident is not impossible. Accidents do not occur in a purely stochastic manner; their occurrence is influenced by many factors, such as the physical condition of the driver, the type of vehicle, driving speed, traffic conditions, road structure and weather. Examining historical accident records would help us understand the (potentially causal) relationship between these factors and road traffic accidents, which in turn would enable us to construct an incident predictor.

**1.3 Interest**

Obviously, City goverment and residents would be very interested in accurate prediction of accidents. It will safe a lot of lifes, and create less traffic on the roads.

**2. Data acquistion and cleaning**

**2.1 Data sources**

[Seatle Accidents data](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv) shows Collision situations from 2004 to Present.

It consists of 37 columns, which record information such as location, severity of accidents, as well as various meteorological data and traffic backgrounds.

**2.2 feature selection**

Firstly, we decided to remove the columns that duplicate the information and leave only the most convenient columns for use. For example, we will use Severeritycode and remove the severitycode.1 and Severitydesc columns. We also leave X and Y and remove Location, Leave st\_colcode and remove st\_coldesc, remove junctiontype and leave addrtype.

Secondly, we will remove columns that do not carry any useful information for us, such as assigned id numbers ('objectid, coldetkey, inckey, reporno, intkey, sdot\_colcode, sdot coldesc , sdotcolnum, Seglanekey and crosswalkkey

Thirdly, we will delete the columns that do not carry any information at all, according to the metadata, the exceptrsncode and exceptrsndesc

**2.3 data cleaning**

To begin with, we will delete the lines where there is no coordinate data, because in the future we will not be able to break this data into clusters and see in what places certain types of accidents most often occur.

Then we have a bunch of columns with binary data in them, we will replace the missing data and the False value like N by 0, also we will replace the True value like Y by 1. So we will bring everything to one form and it will be easier to work with such data.

We also have columns in which the ‘unknown’ value already exists, so let's replace the missing data by ‘unknown’ value.

There are only 18 lines left with the missing data from the St\_colcode column and we will simply delete these lines, since we cannot predict what is there and this will not greatly affect our result.

**3. Methodology section**

In this project we will try to figure out where,why and when most accidents occur.

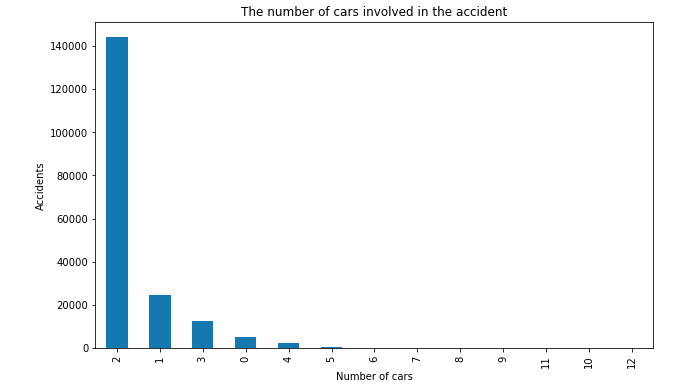
In the first step we prepare our data to be ready to work with

Second step in our analysis will be calculation and exploration of reasons of accidents in Seatle and the dependence of accidents on time and day of week.

In third and final step we will focus on location of accidents and within those create clusters of locations which are showing typical accident areas. Also we will create a machine learning model that will predict Severity code.

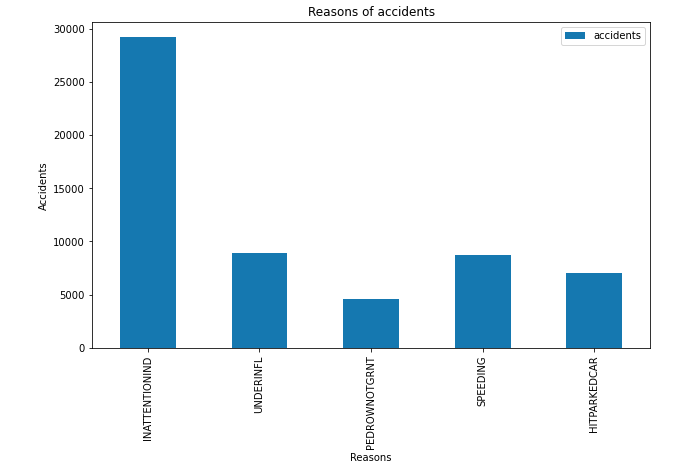
**3.1 Data Analysis**

**3.1.1 how many vehicles are involved in the accident?**

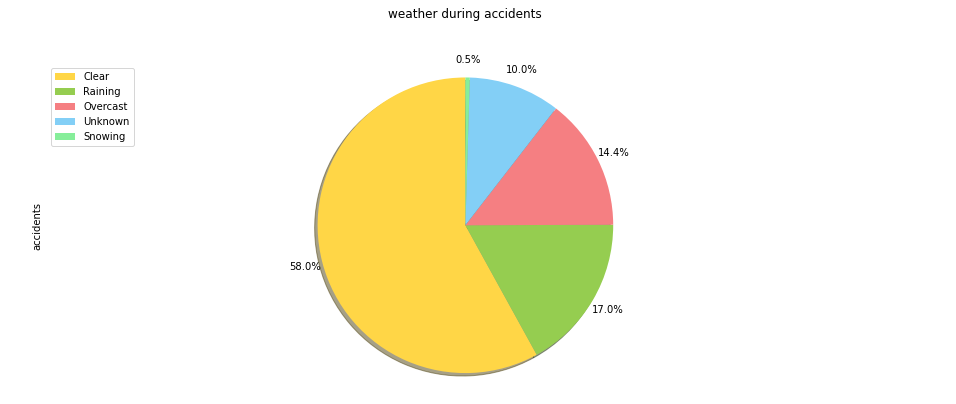


The histogram above shows the distribution of accidents depending on the number of vehicles. This indicates that most accidents involve one or two vehicles. This makes sense intuitively; most accidents occur as a result of a collision with the road and / or a collision between two vehicles.

**3.1.2 Why do people get into accidents?**

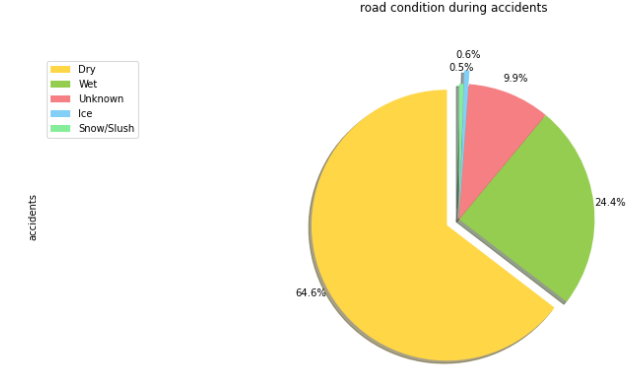


The histogram above shows that the most important reason for people to get into an accident is inattention, but we also see that a decent number of accidents also occur due to other reasons.

**3.1.3 Weather during accidents**

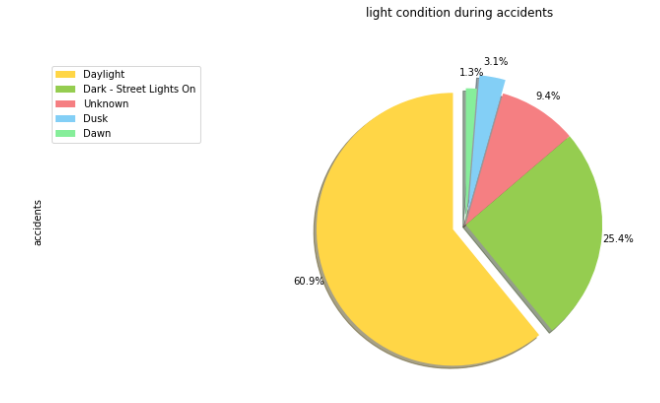
How we can see most of the accidents happened in dry weather, so we can conclude that the weather does not greatly affect the accidents.

**3.1.4 Road Conditions during accidents**

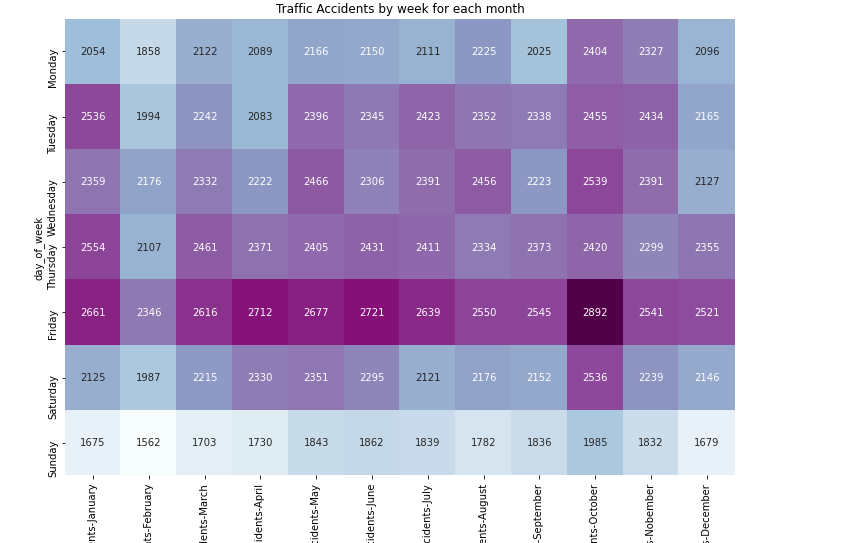


Consider how road conditions affect accidents. As we can see that most of the accidents occurred on a dry road, from which we can conclude that road conditions do not greatly affect the number of accidents.

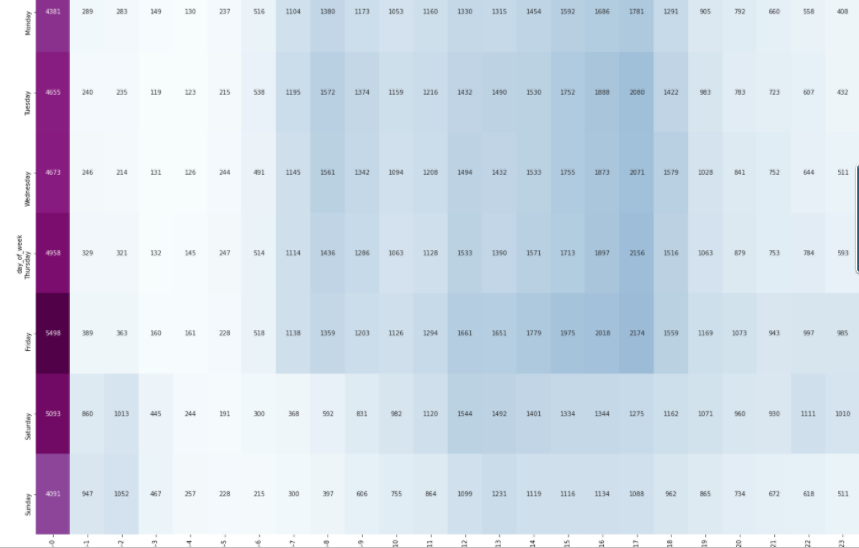
**3.1.5 Light Conditions during accidents**

Now let's consider what kind of road light conditions people most often get into accidents.  
As we can see, people were in accidents most often in normal daylight and we can conclude that road light conditions does not greatly affect the number of accidents

**3.1.6 What days, months and what time most accidents happen?**



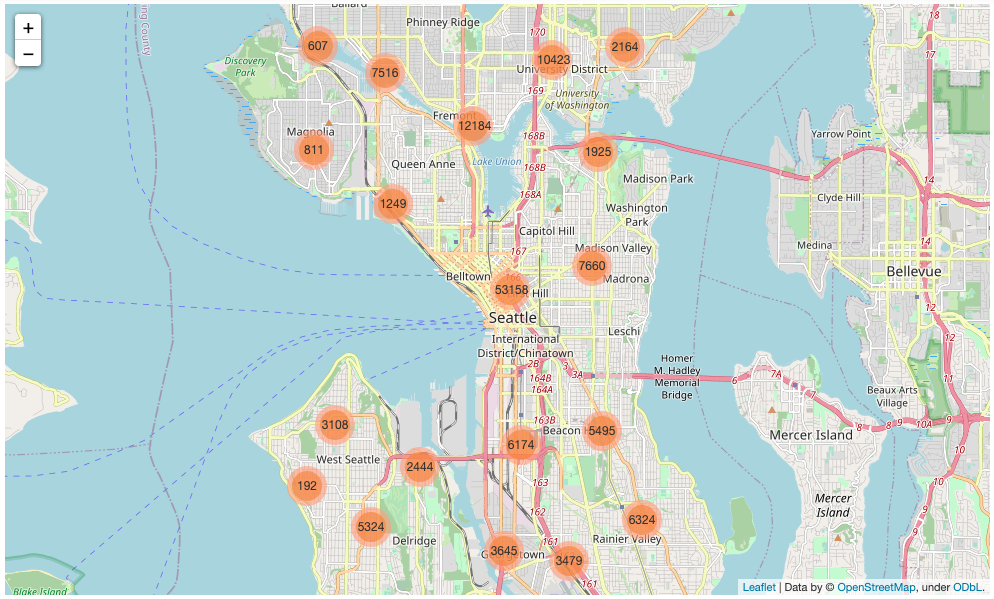
The above heat map shows the distribution of accidents by months and by days of the week. as we can see the distribution of accidents at different times of the year is approximately the same, but on the days of the week we can see that the most accidents occur on Friday.



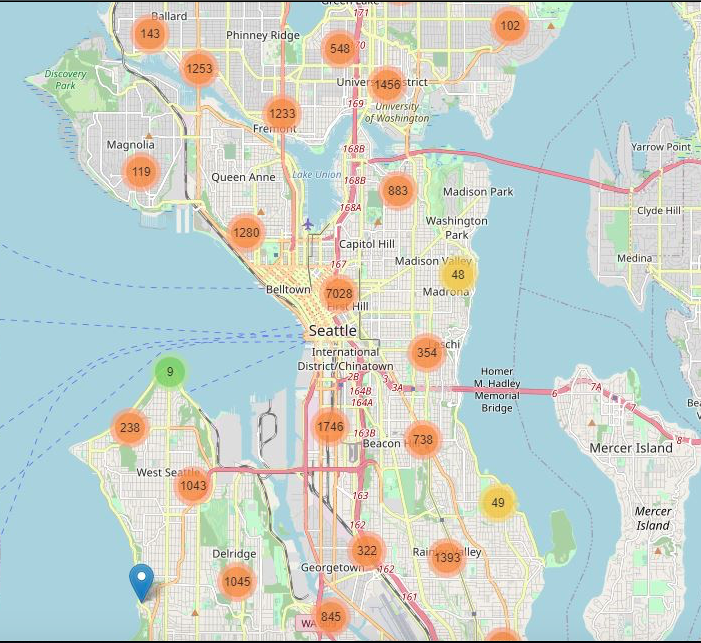
Now let's consider the distribution of accidents over time  
As we can see on the heat map, most accidents happen at 12 am and between 7 am and 5 pm.

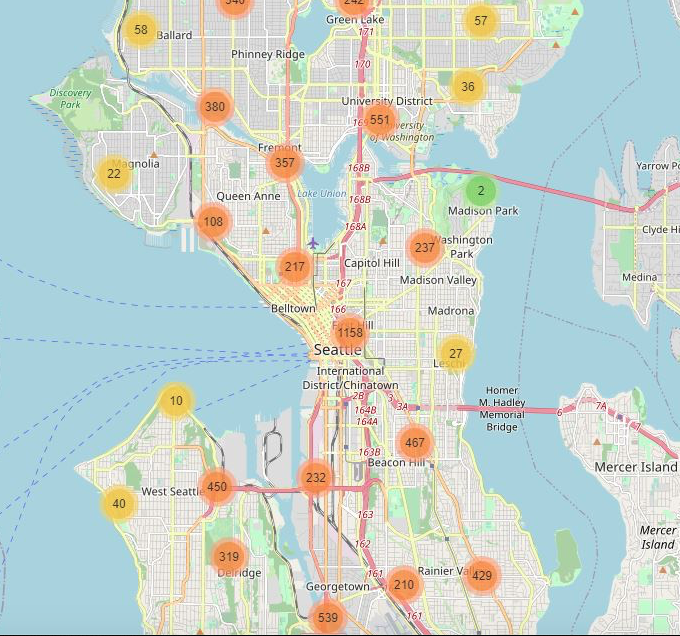
**3.2 Accidents visualization on Map.**

Let's break down all the crashes into clusters and see where the most crashes happened.

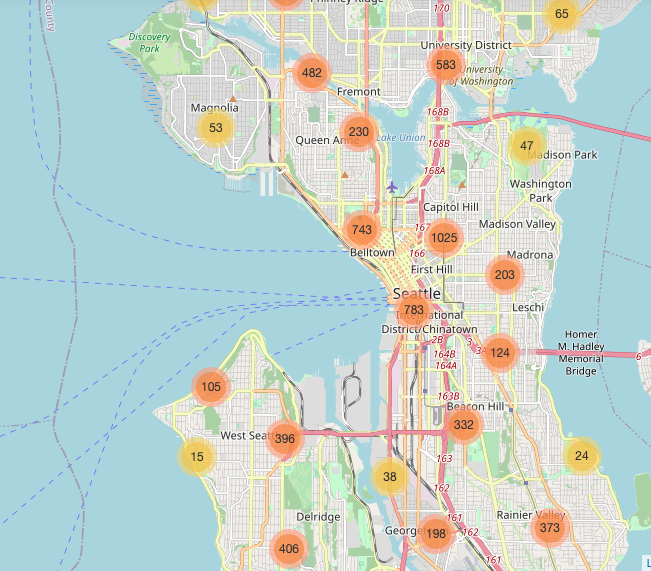


As we can see most of the accidents happened in the city center, at the university and in the Fremont area. Also we can see St\_colcode for every accident.

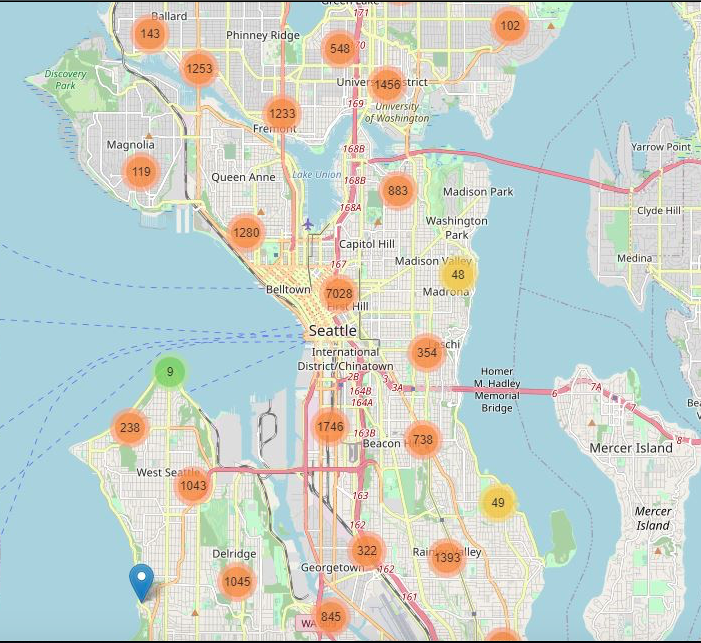


Now let's look at the map above where people got into accidents due to inattention.

On this Map we can see areas where accidents were due to speed limit.



And on the last Map we can see that most of accidents Underinfluence happen near center of the city and at University areas.



On the Map above we can see areas where we have parking problems. Also we can see Address type for each accident.

**3.3 Machine learning for Severity code prediction**

We will use classification machine learning.At the begining we will delete some columns not to make our data overfit. Ans than will look what method is better to use

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Methods | Knn | DT | SVM | LR |
| Jaccard index | 0.57 | 0.74 | 0.74 | 0.74 |
| F1-score | 0.58 | 0.68 | 0.69 | 0.70 |
| Logloss | 0.97 | 0.56 | - | 0.54 |

Based on the below table, Logistic Regression is the best model to predict car accident severity. Now we can predict was it property damage (class 1) or injury (class 2).

**4. Results section**

Firstly, it became clear for what reasons most accidents occur, at what time and on what days of the week they most often occur.

Secondly, we realized that the seasons, temperature, road conditions and light conditions do not affect the number of accidents at all.

We also analyzed and looked at which places most accidents occur and in which places they occur due to the fact that the driver was under influence, speeding, hit the parked car and was inattententive.

Thirdly, thanks to machine learning, we can predict the severity code.

**5. Discussion section**

Now we can use the information on the maps and take some action. For example, to place police officers in those places where the most accidents occurred, prohibit parking where most accidents with parked cars occur. And also use the received data in navigation to help people bypass these hot spots.

**6. Conclussion**

In conclussion I would like to saythat if we will use such information, we can make less traffic problems and save some lifes. Thank for reading my report.