Capstone project report:
Car Collisions Data Analysis

1 Problem Statement

Many car accidents occur due to bad road or weather conditions in certain places. Some of the accidents could have been avoided if preventive measures would have been taken like setting speed limits, warnings for drivers or improving the traffic flows.

In order to take preventive actions, need to analyze the accident data for the previous years and determine correlation of the number and severity of the accidents with weather conditions, road conditions, time of the day and location.

This will help the government to improve the situation on roads.

2 Data Description

The data about collisions is provided by SPD and recorded by Traffic Records. The dataset includes information about 194,673 collisions of all types in Seattle from beginning of 2014 to May of 2020.

For each collision the dataset contains coordinates of the incident, its severity, number of people and cars involved, incident date and time, weather and road condition, whether a driver involved was under the influence of drugs or alcohol.

The data is available via the following link: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

2.1 Attributes not considered in the study

There are 38 attributes in total in the collected data set. Some of them are not useful for statistical analysis, so they are not considered:

1) OBJECTID - ESRI unique identifier;
2) COLDETKEY - Secondary key for the incident;
3) REPORTNO – number of the report;
4) STATUS – no description is given;
5) ADDRTYPE - Collision address type: Alley, Block, Intersection;
6) INTKEY - Key that corresponds to the intersection associated with a collision;
7) EXCEPTRSNCODE, EXCEPTRSNDESC – inaccurate location info. The data with these values will be removed from the dataset.
8) INCDTTM - The date and time of the incident. The data contains inaccurate data: “am” or “pm” or timestamp itself are often missed, so not useful for statistical analysis.
9) INATTENTIONIND – Whether or not collision was due to inattention. Subjective details, not useful for statistical analysis.
10) SDOTCOLNUM - A number given to the collision by SDOT.
11) ST_COLCODE - A code provided by the state that describes the collision.
12) ST_COLDESC – A description that corresponds to the state’s coding designation.
13) SEGLANEKEY - A key for the lane segment in which the collision occurred.
14) CROSSWALKKEY - A key for the crosswalk at which the collision occurred.
15) HITPARKEDCAR - Whether or not the collision involved hitting a parked car.

2.2 Statistical overview of the dataset

Statistical overview including the count of a certain attribute in the set, its frequency, number of unique values and the top frequent value can be found in the table below.

Attribute name	Number of values	Number of unique values	Top frequent value	Frequency
LOCATION	189339	23890	N NORTHGATE WAY BETWEEN MERIDIAN AVE N AND COR...	265
SEVERITYDESC	189339	2	Property Damage Only Collision	132221
COLLISIONTYPE	184582	10	Parked Car	46381
INCDATE	189339	5985	2006/11/02 00:00:00+00	88
JUNCTIONTYPE	185146	7	Mid-Block (not related to intersection)	87390
SDOT_COLDESC	189339	39	MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END ...	84538
UNDERINFL	184602	4	N	97639
WEATHER	184414	11	Clear	108959
ROADCOND	184481	9	Dry	122076
LIGHTCOND	184327	9	Daylight	113582
PEDROWNOTGRNT	4645	1	Y	4645
SPEEDING	8720	1	Y	8720
ST_COLCODE	189321	115	32	26650
ST_COLDESC	189339	63	One parked--one moving	42869
HITPARKEDCAR	189339	2	N	182334

2.3 Attributes that will be used for the analysis in this study

From the remaining attributes the most useful seem to be the following:

1) Location (including latitude, longitude and location description),
2) Incident Date,
3) Weather conditions,
4) Road conditions and
5) Light conditions.

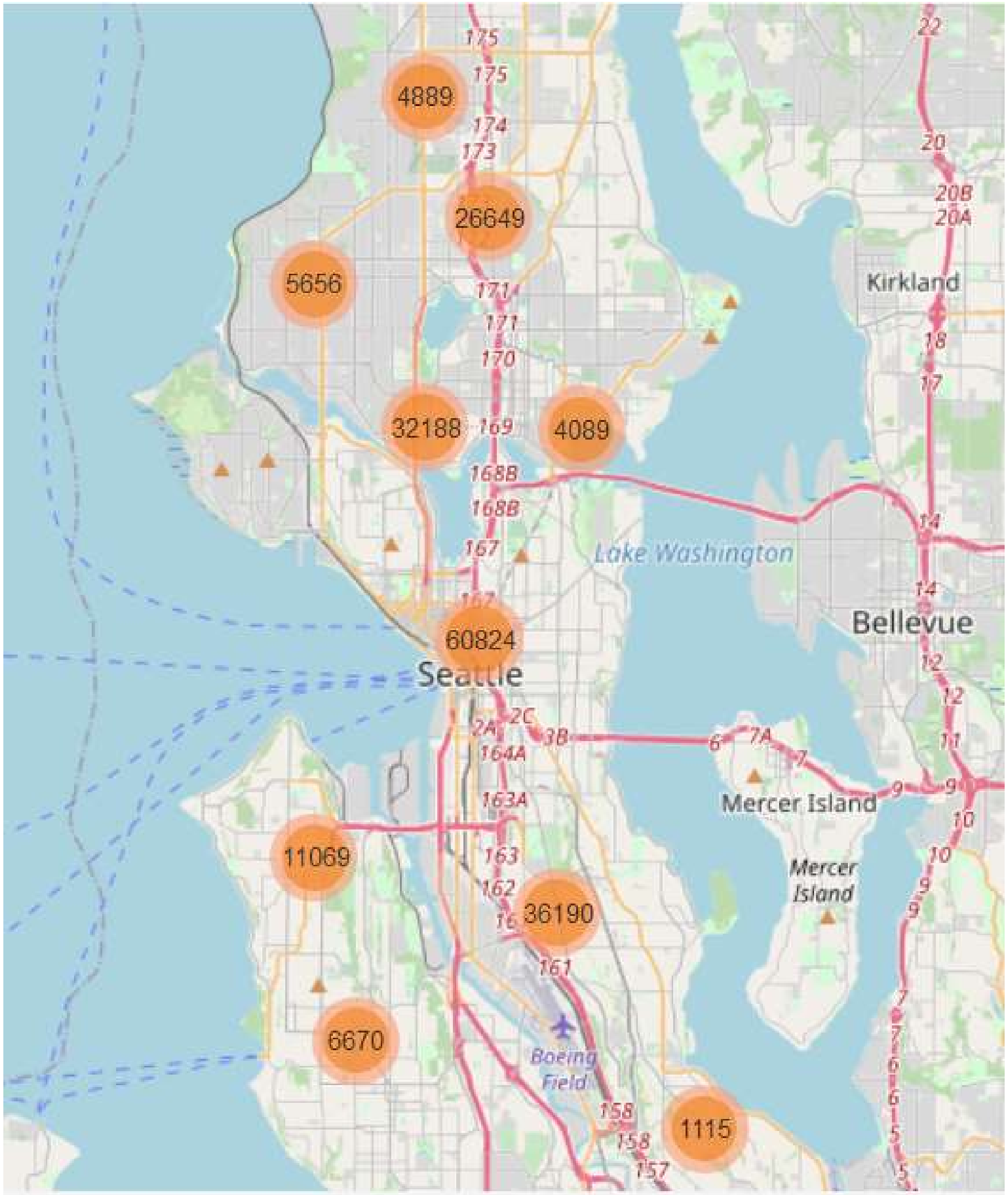
Weather, road and light conditions may impact the number of accidents. Poor traffic conditions in certain locations can be also the reason of higher number of accidents.

The study will cover the correlation between different conditions and the number of accidents per day. It will also show the established trend of the accident frequency in time.

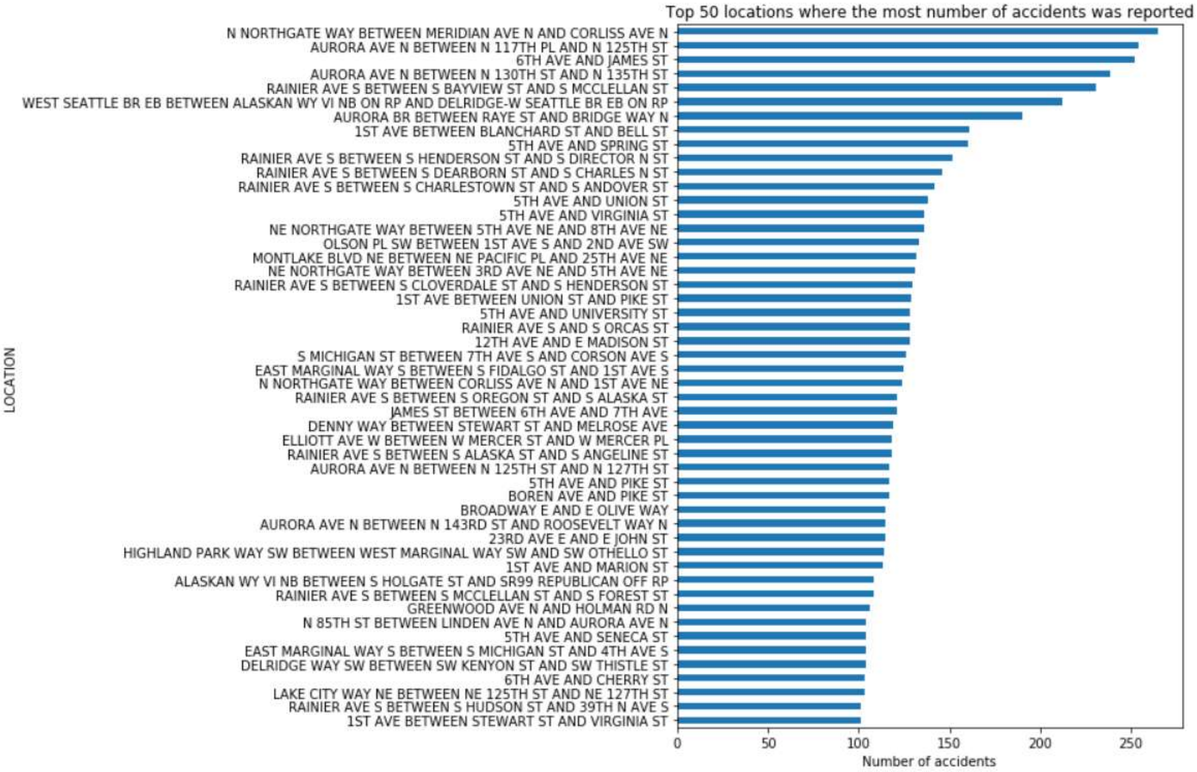
3 Exploratory data analysis

3.1 Frequency of the accidents in certain locations

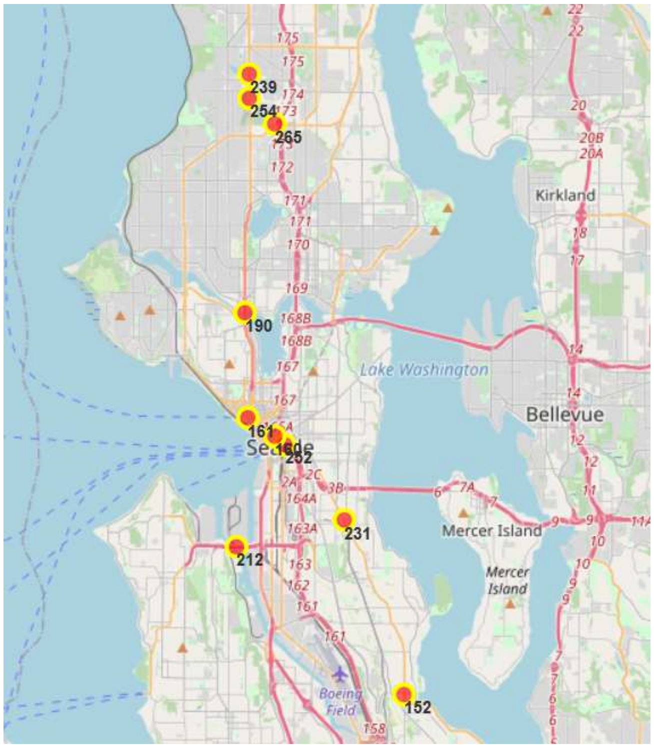
There are 23,890 unique locations in the city where the accidents were reported from. in certain places however the number of accidents is significantly higher than in others. The following map shows the numbers of accidents reported from various spots.



Top 50 places where the number of accidents for the last 5 years is more that 100 require special attention of the authorities and are shown on the diagram below.

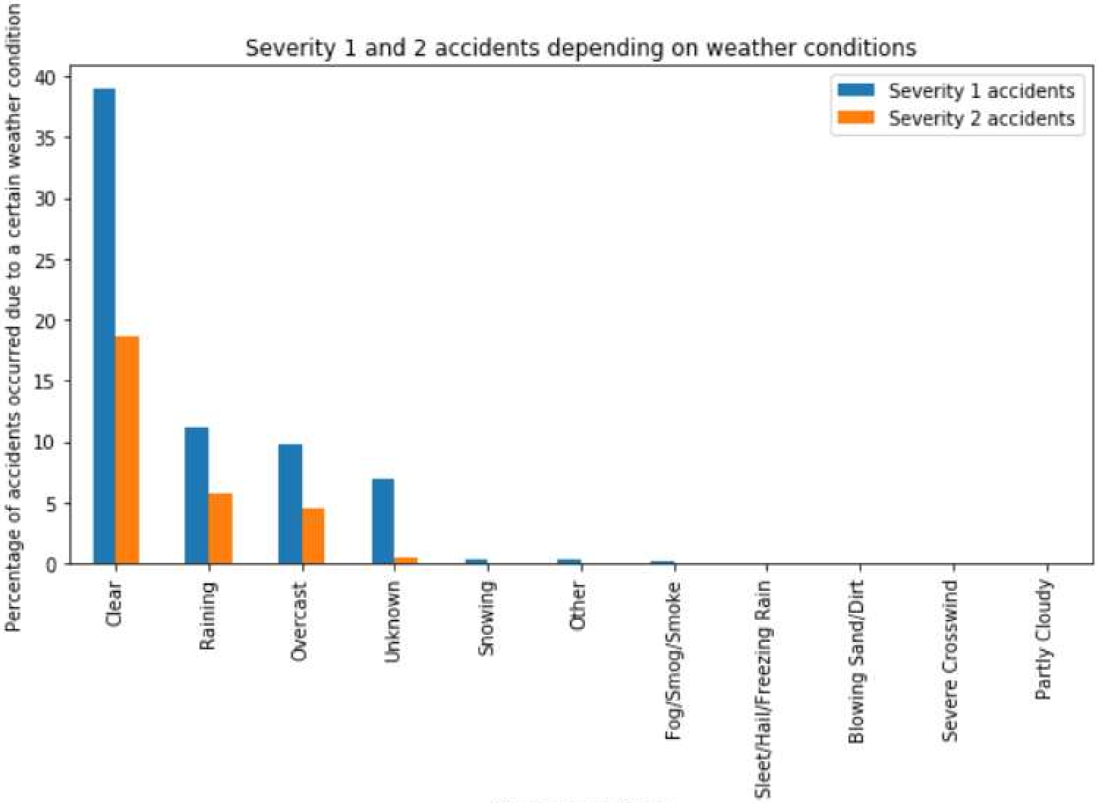


Top 10 locations with the highest number of accidents on a map:



3.2 Influence of weather conditions on the number of accidents

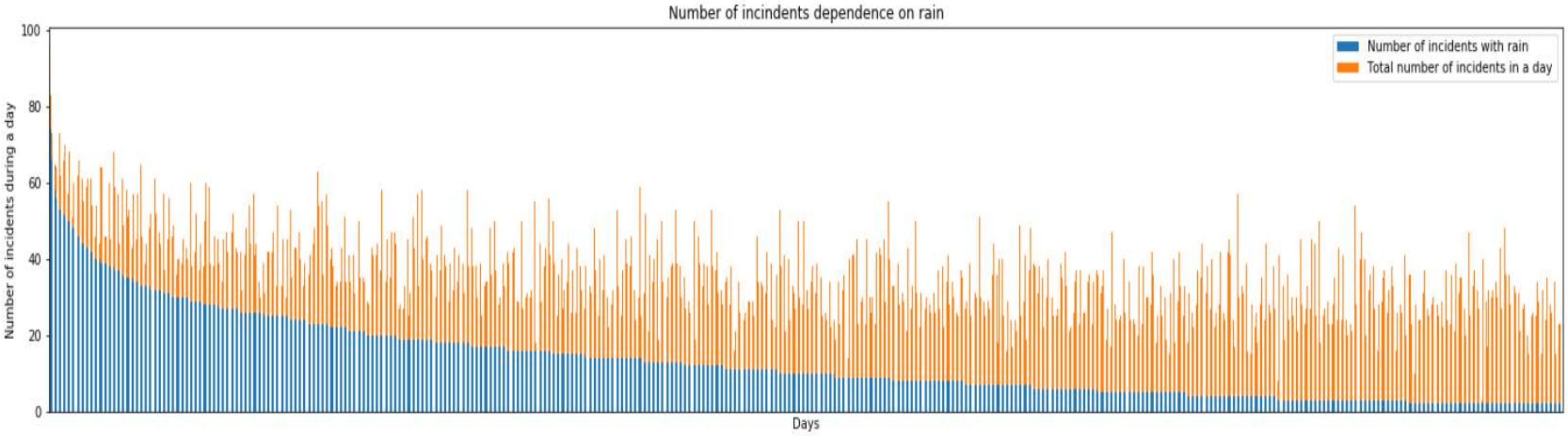
Let’s determine how weather conditions affect the number of incidents per day. The chart below shows that most of the accidents occurred when the weather was clear followed by rain.



Find out the dates, where the number of incidents with certain weather conditions were at maximum levels.

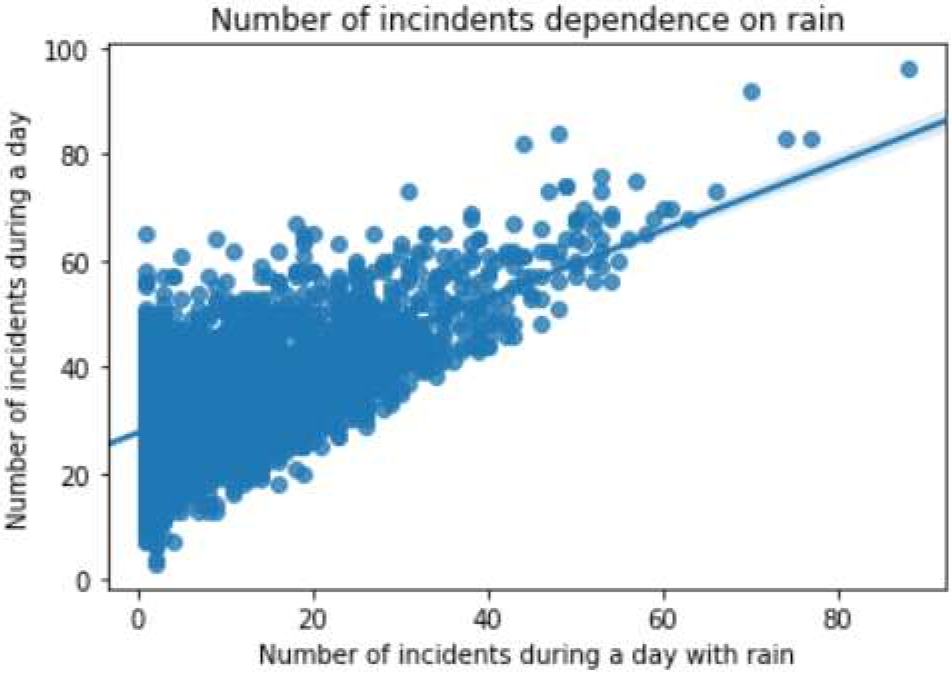
3.2.1 Number of accident dependency on rain

On the following diagram the orange bars show the number of accidents at a certain day, and the blue part of it – the ones where weather conditions were marked as "Rainy". It is clear that during rainy days the amount of accidents is higher than during sunny days.



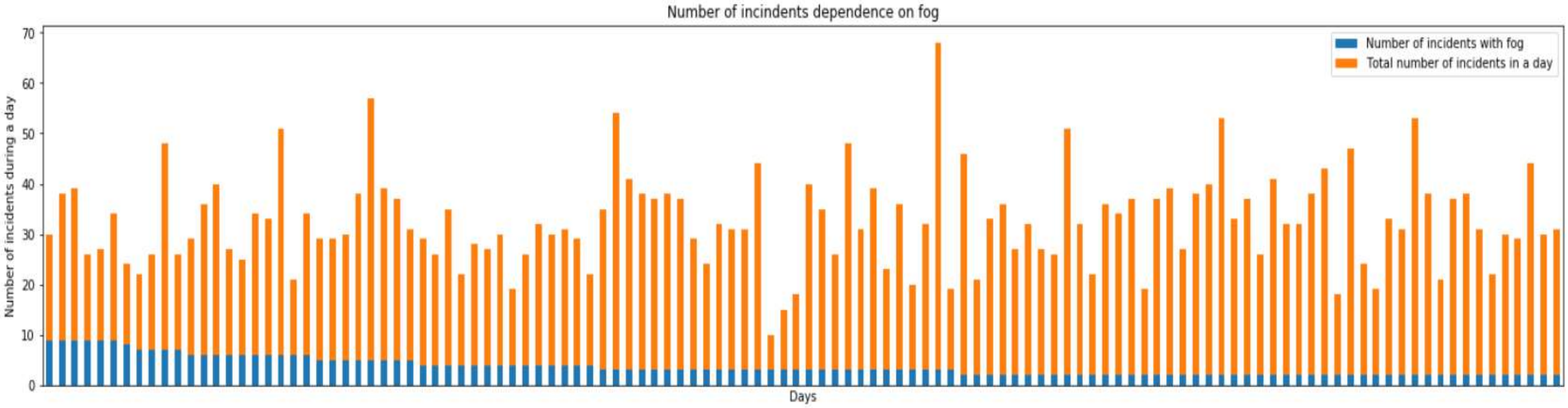
The correlation between the number of accidents reported during rainy conditions and the total number of accidents reported within the day is 0.648, which proves higher probability of an accident during a rainy day.

Regression plot below also shows high dependency of the number of accidents on the rain.



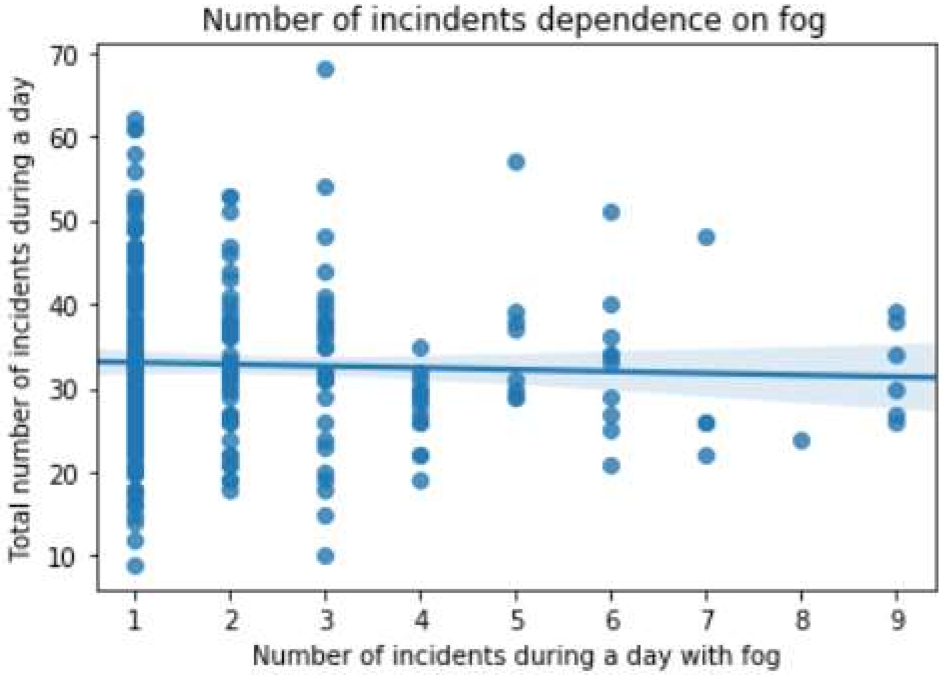
3.2.2 Number of accident dependency on fog

As on the previous chart, on the following diagram the orange bars show the number of accidents at a certain day, and the blue part of it – the ones where weather conditions were marked as "Fog/Smog/Smoke". The diagram shows that the number of accidents does not depend on fog.



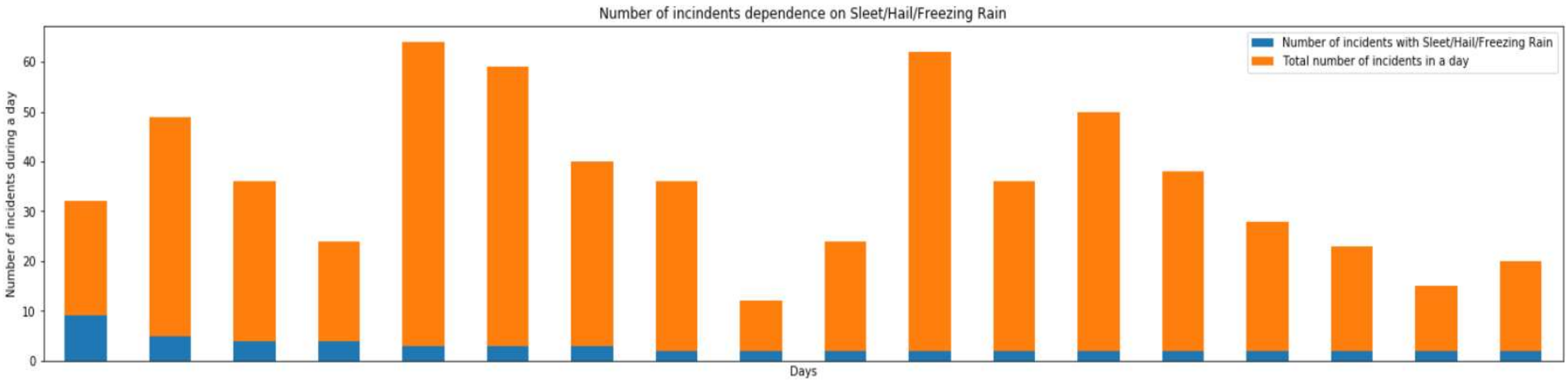
The correlation between the number of accidents reported during fog conditions and the total number of accidents reported within the day is 0.0397, which proves that the probability of an accident during fog is not higher than during a sunny day.

Regression plot below also shows no dependency of the number of accidents on fog.



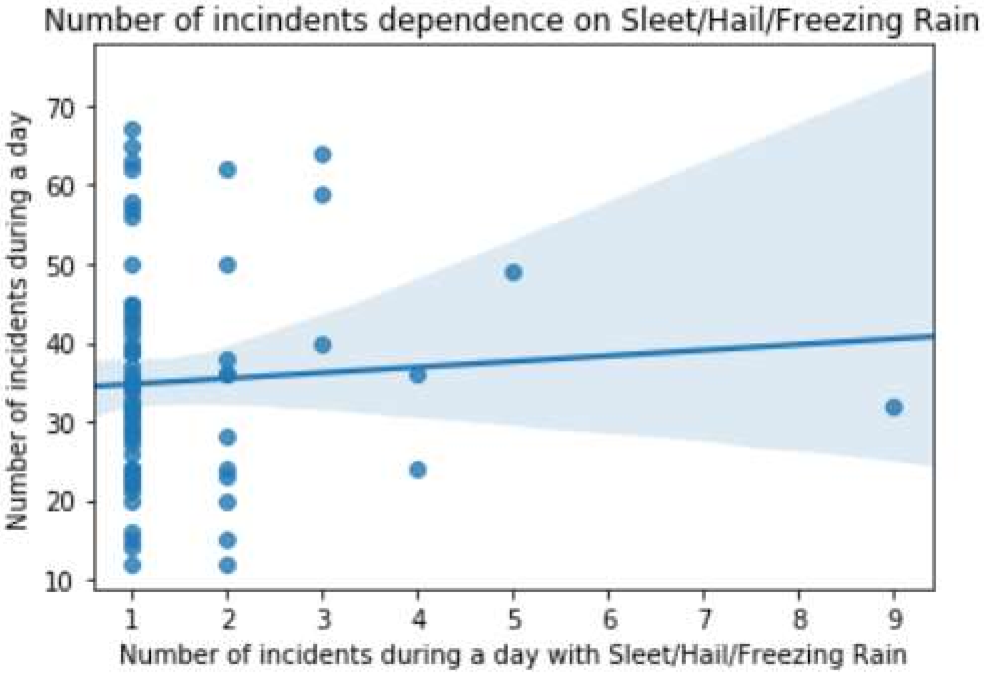
3.2.3 Number of accident dependency on Sleet/Hail/Freezing Rain

As on the previous chart, on the following diagram the orange bars show the number of accidents at a certain day, and the blue part of it – the ones where weather conditions were marked as "Sleet/Hail/Freezing Rain". The diagram shows that the number of accidents does not depend on sleet.



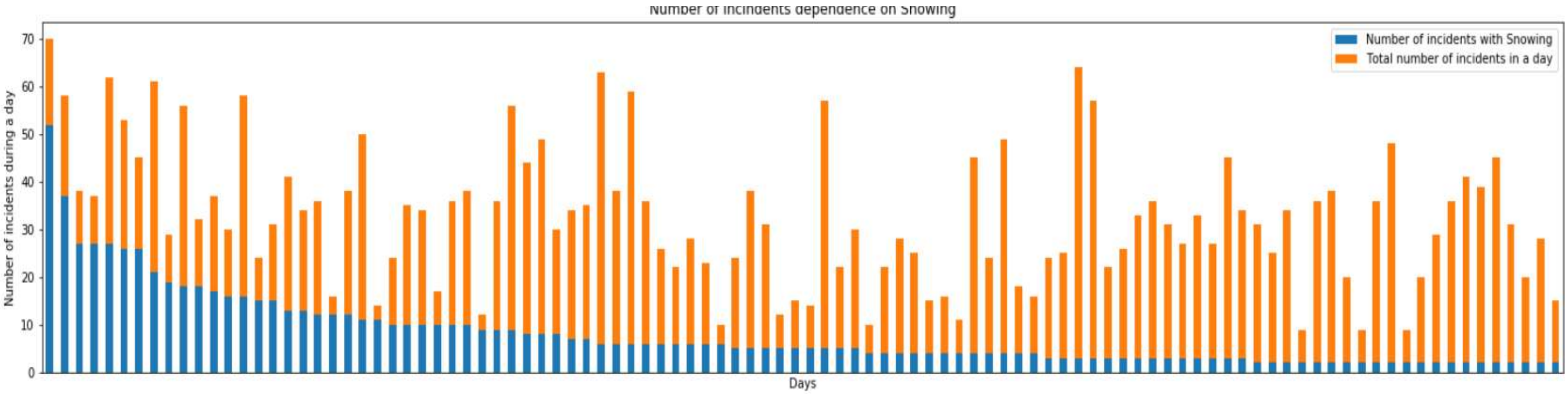
The correlation between the number of accidents reported during fog conditions and the total number of accidents reported within the day is 0.0626, which proves that the probability of an accident during sleet is not higher than during a sunny day.

Regression plot below also shows no dependency of the number of accidents on fog.



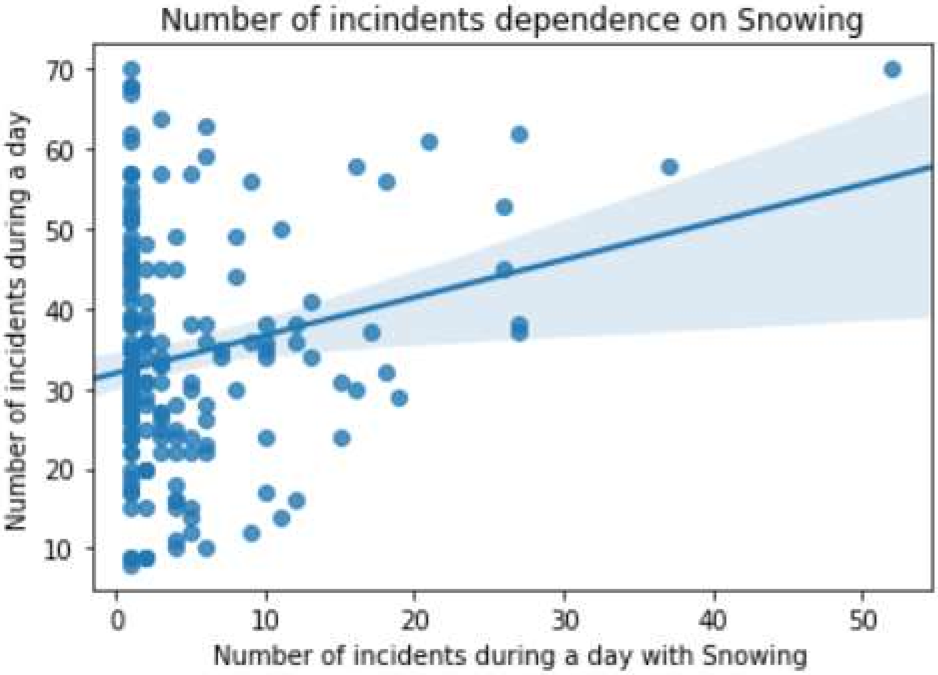
3.2.4 Number of accident dependency on snow

As on the previous chart, on the following diagram the orange bars show the number of accidents at a certain day, and the blue part of it – the ones where weather conditions were marked as "Snowing". The diagram shows that the number of accidents is slightly higher when it is snowing than during a sunny day.



The correlation between the number of accidents reported during fog conditions and the total number of accidents reported within the day is 0.2326, which proves that slightly higher probability of an accident when it is snowing.

Regression plot below also shows low dependency of the number of accidents on snow.



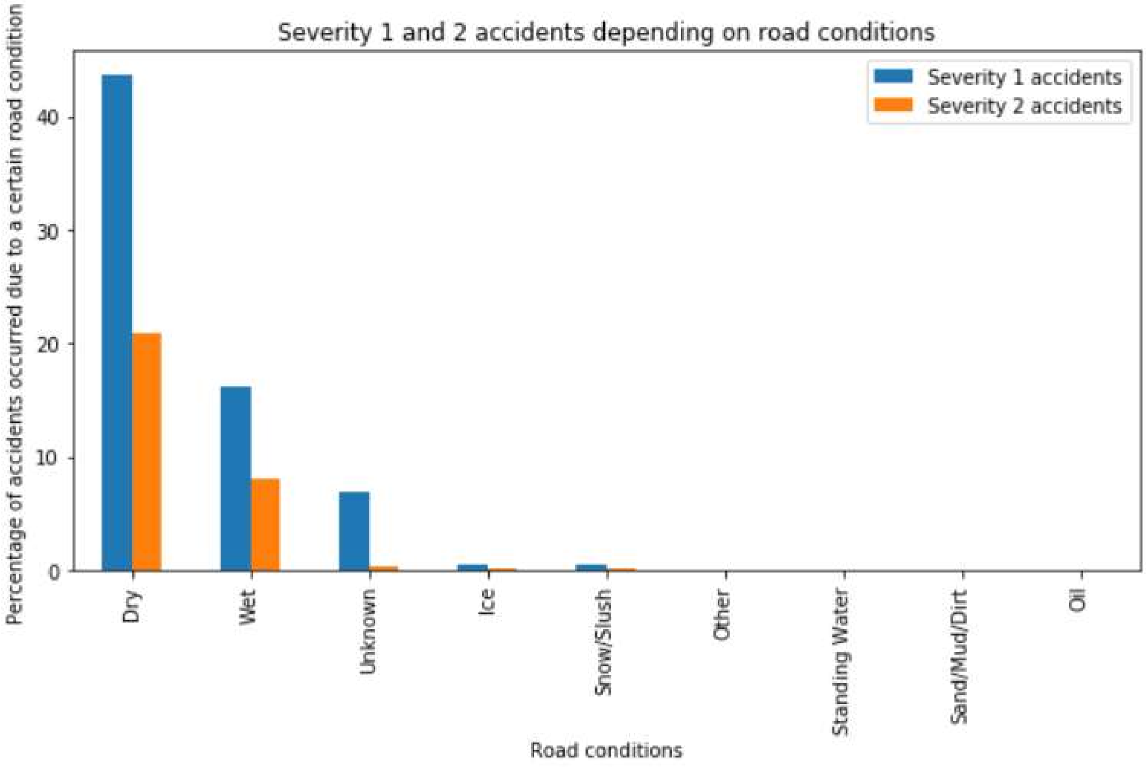
3.2.5 Conclusions about weather conditions correlation with the number of accidents

As could be seen from the previous paragraphs, the largest number of accidents occurred during normal weather conditions. During a rainy day the probability of getting into an accident is significantly higher. Snow also increases the probability slightly, while sleet and fog did not affect the number of accidents, therefore there is no influence on the probability.

Therefore, drivers should be extremely cautious when it is raining or snowing. Authorities could also use warning signs to stress drivers' attention on the weather conditions.

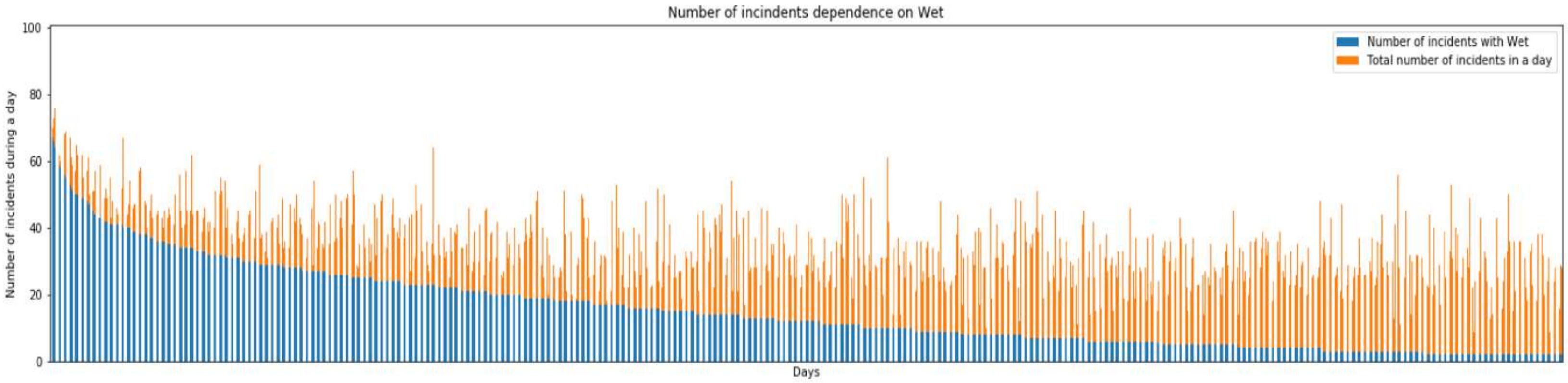
3.3 Influence of road conditions on the number of accidents

More than 40% of the accidents were severity 1 and reported when the road was dry, around 20% of the accidents were reported with severity 2 from dry roads. Wet roads take the second place followed by ice. Let's study in more details how road conditions influence the amount of car accidents during the day.



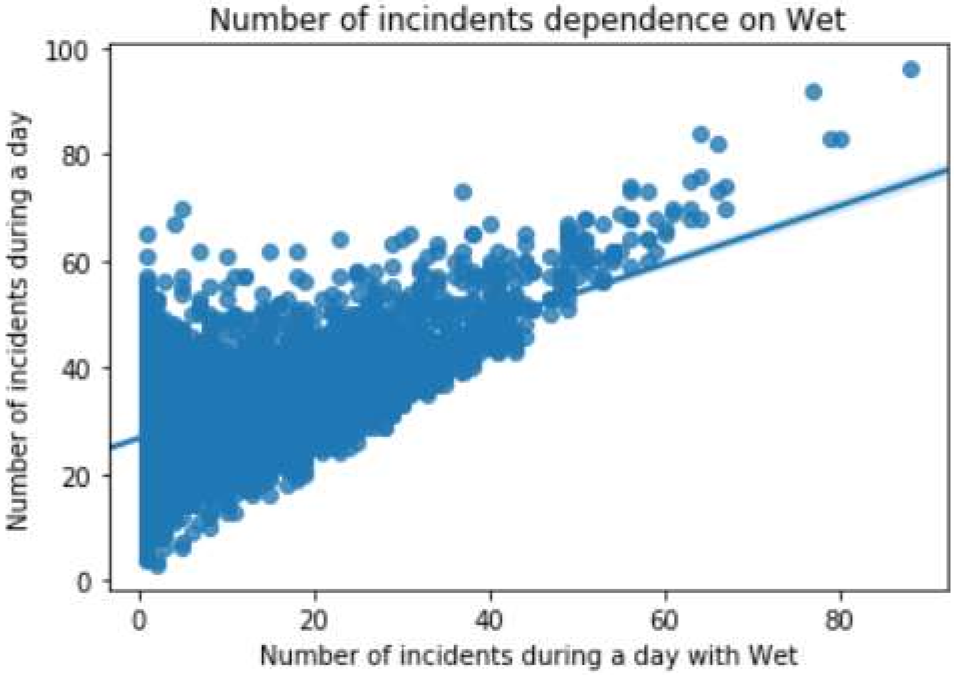
3.3.1 Number of accident dependency on wet road

On the following diagram the orange bars show the number of accidents at a certain day, and the blue part of it – the ones where road conditions were marked as “Wet”. The chart illustrates that when the road is wet, the number of accidents increases.



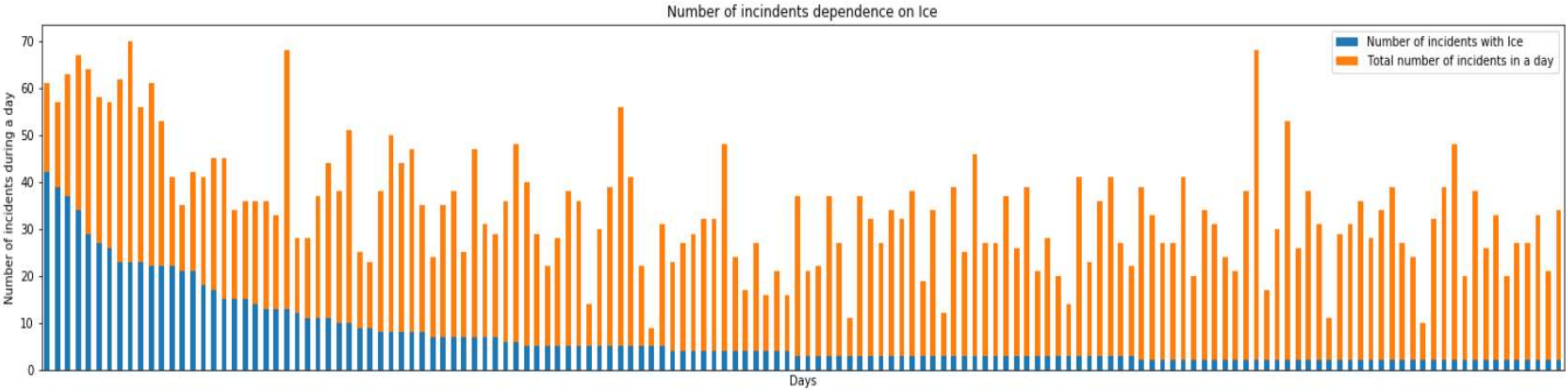
The correlation between the number of accidents reported during rainy conditions and the total number of accidents reported within the day is 0.6327, which proves higher probability of an accident on a wet road.

Regression plot below also shows high dependency of the number of accidents on the road wetness.



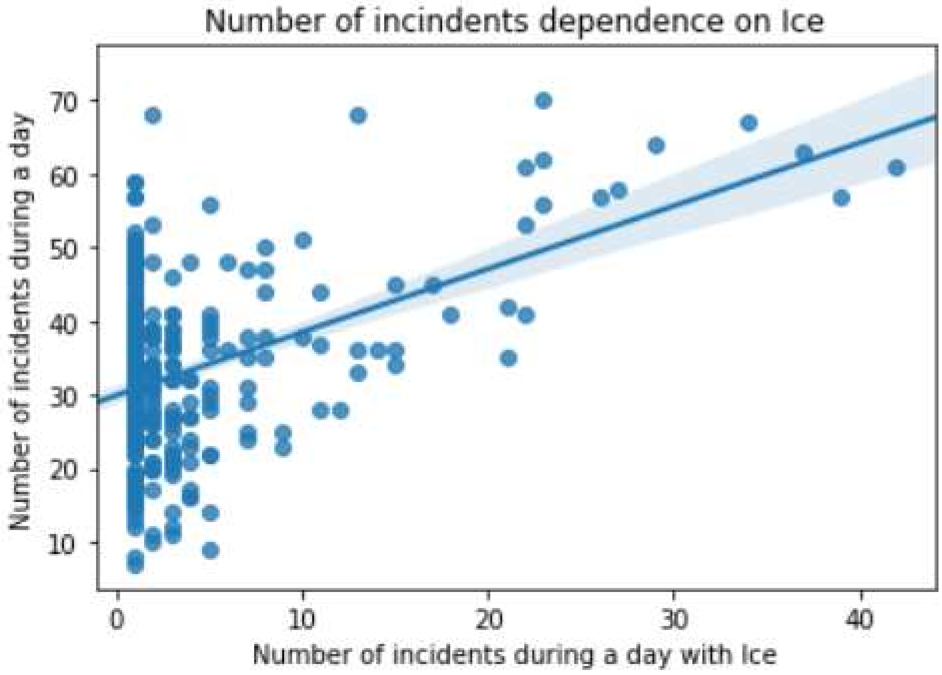
3.3.2 Number of accident dependency on ice road

On the following diagram the orange bars show the number of accidents at a certain day, and the blue part of it – the ones where road conditions were marked as "Ice". The chart illustrates that when the road is icy, the total number of accidents increases.



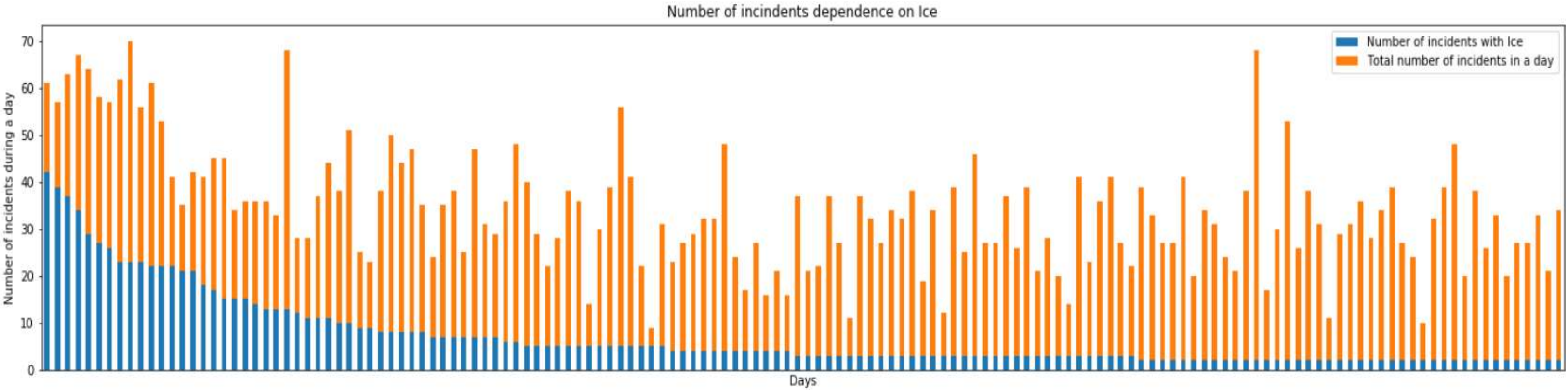
The correlation between the number of accidents reported during rainy conditions and the total number of accidents reported within the day is 0.4630, which proves higher probability of an accident on an ice road.

Regression plot below also shows high dependency of the number of accidents on the road with ice.



3.3.3 Number of accident dependency on snow and slush on the road

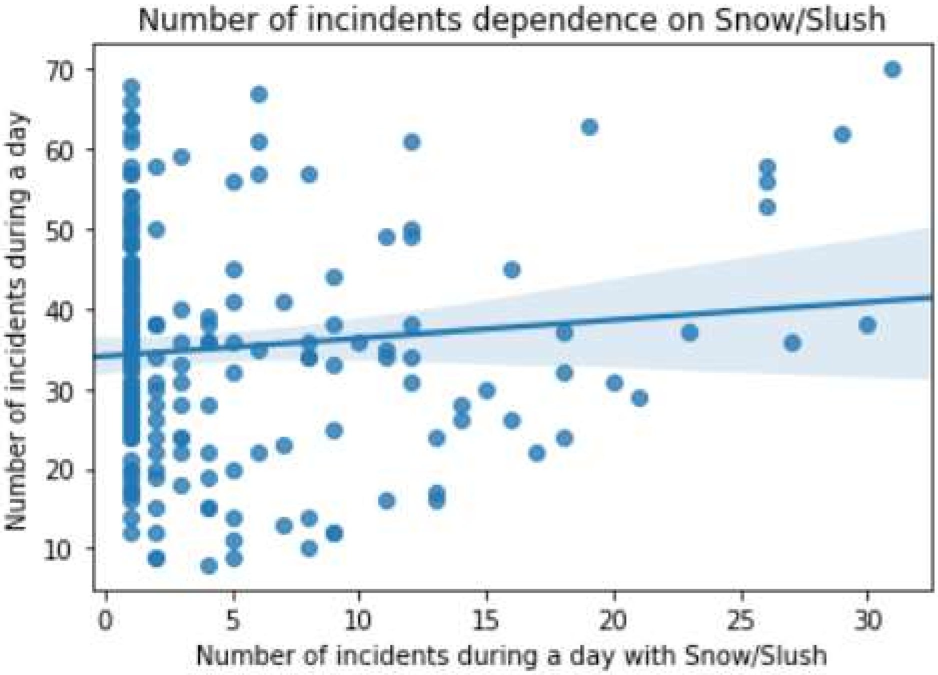
On the following diagram the orange bars show the number of accidents at a certain day, and the blue part of it – the ones where road conditions were marked as “Snow/Slush”. The chart illustrates slight increase of the total number of accidents when there is snow on the road.



The correlation between the number of accidents reported during snow conditions and the total number of accidents reported within the day is 0.1034, which proves low dependency of the probability of an accident on the road with snow.

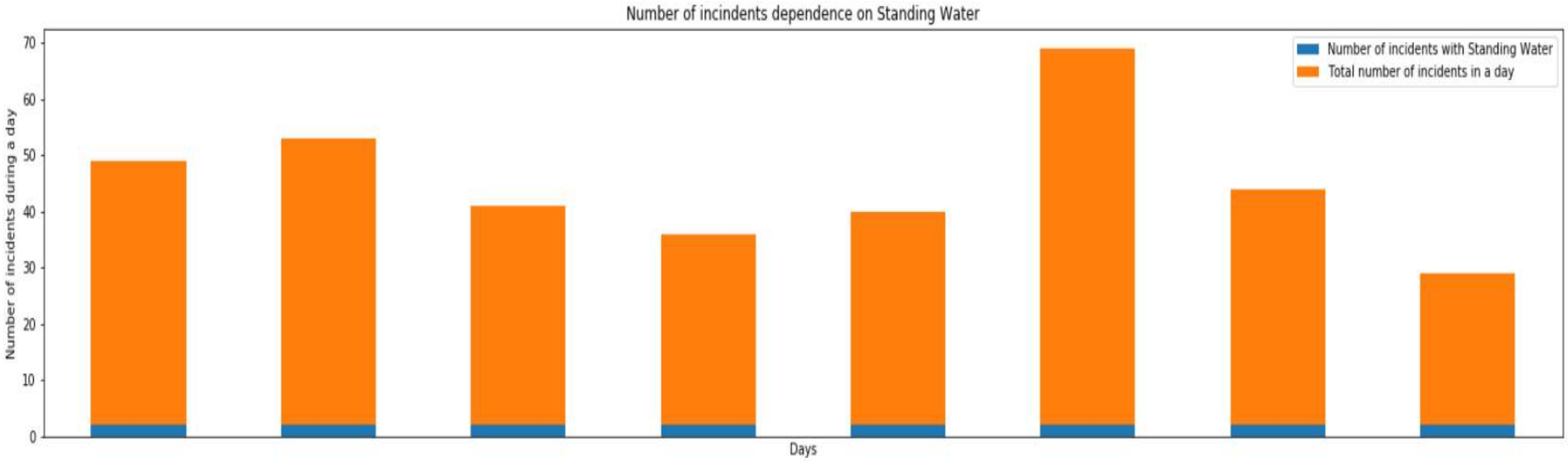
However, when high number of cases were reported with snow on the roads (on the left-hand side of the chart), the total number of incidents is higher than the average amount of accidents in a given month. Cutting the data set with low amount of the number of accidents reported with ice on the road, would probably improve the model.

Regression plot below also shows low dependency of the number of accidents on the road with snow.



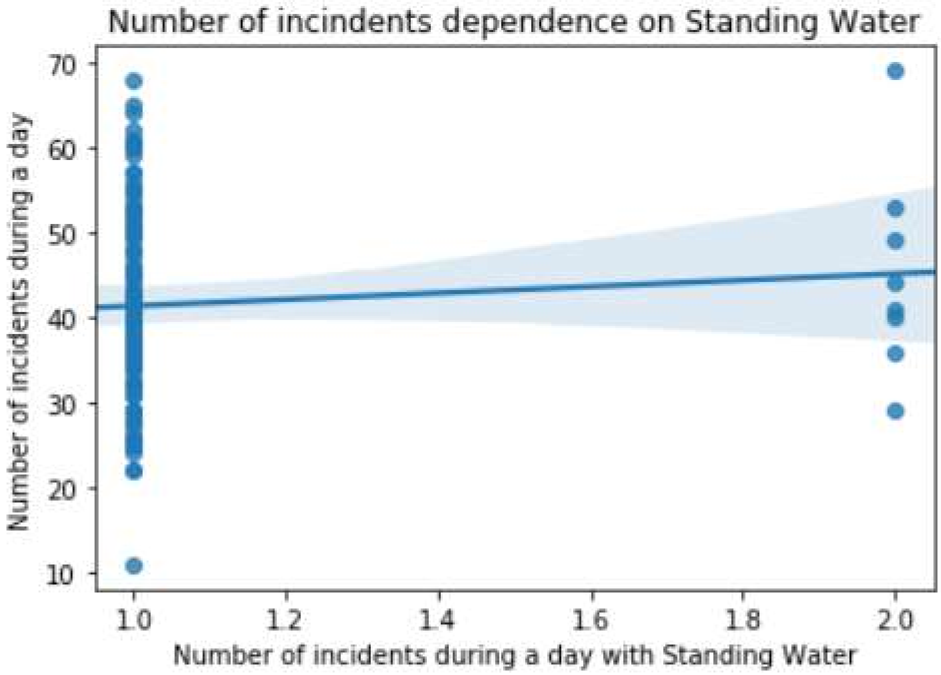
3.3.4 Number of accident dependency on standing water on the road

On the following diagram the orange bars show the number of accidents at a certain day, and the blue part of it – the ones where road conditions were marked as "Standing Water". The chart does not illustrate dependency of the number of accidents from the standing water on the road.



The correlation between the number of accidents reported during rainy conditions and the total number of accidents reported within the day is 0. 0866, which proves that there is no dependency of the probability of an accident on the road with standing water.

Regression plot below also shows low dependency of the number of accidents on the road with standing water.



3.3.5 Conclusions about road conditions correlation with the number of accidents

As could be seen from the previous paragraphs, the largest number of accidents from dry roads. When a road is wet or icy, the probability of getting into an accident is significantly higher. Snow also increases the probability slightly, while standing water on the road did not affect the number of accidents, therefore there is no influence on the probability.

Dirt and oil on the roads were not reported often, so the available data does not allow to build any conclusion on their influence on the number of accidents.

Therefore, drivers should be extremely cautious when the road is wet or ice. Authorities could also use warning signs to stress drivers' attention on the road conditions.

3.4 Influence of light conditions on the number of accidents

In order to check the maximum correlation of the light condition with the total number of collisions, a pivot table with all conditions and the total number of accidents was created:

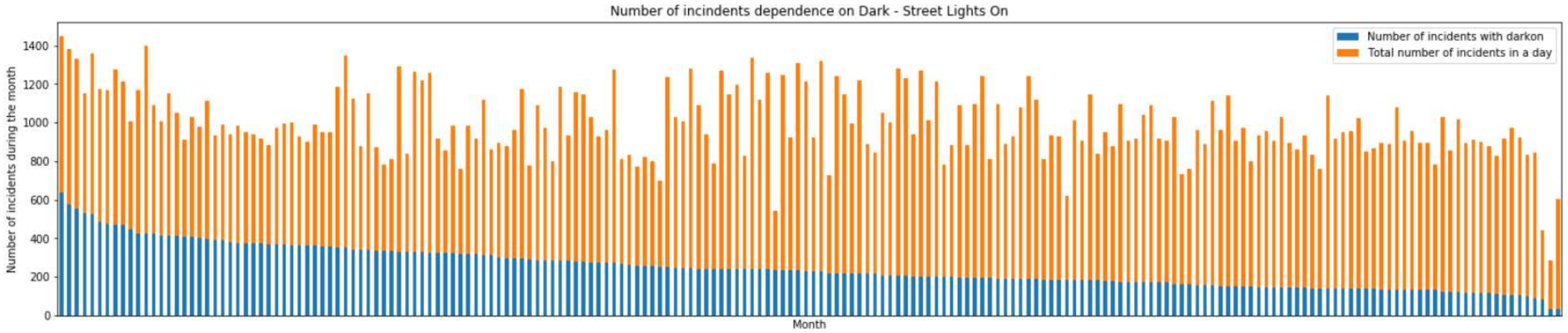
strmonth	Dark - No Street Lights	Dark - Street Lights Off	Dark - Street Lights On	Dawn	Daylight	Dusk	Total number of collisions
2004-01	8	5	400	24	453	31	976
2004-02	7	7	303	3	504	23	893
2004-03	5	5	273	15	642	32	1030
2004-04	8	6	207	6	753	30	1053
2004-05	9	2	217	15	832	35	1148
2004-06	7	2	175	11	837	24	1098
2004-07	7	5	189	9	786	32	1079
2004-08	4	8	195	5	798	34	1092
2004-09	11	5	246	4	683	37	1028
2004-10	8	5	319	25	559	24	984

Based on this, the maximum correlation of the total number of collisions is with the Dusk condition:

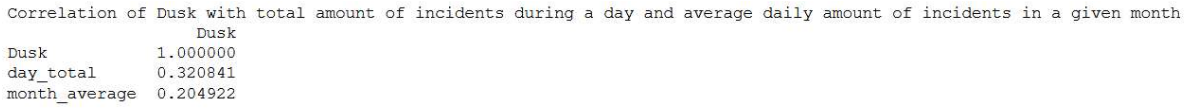
	Total number of collisions	**Dusk**	**Daylight**	**Dark - Street Lights On**	**Dawn**	**Dark - No Street Lights**	**Dark - Street Lights Off**
Total number of collisions	1	0.726853	0.717462	0.417451	0.402899	0.393712	0.307687

3.4.1 Light conditions: Dark - Street Lights on

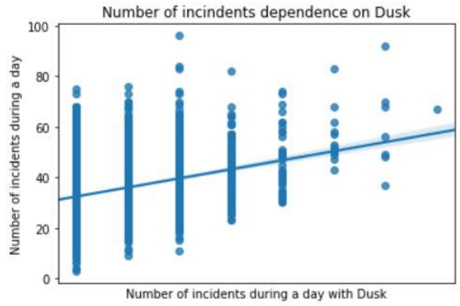
The maximum number of collisions appear when it is dark the street lights are on.



However, the correlation with the total number of collisions is not very high:

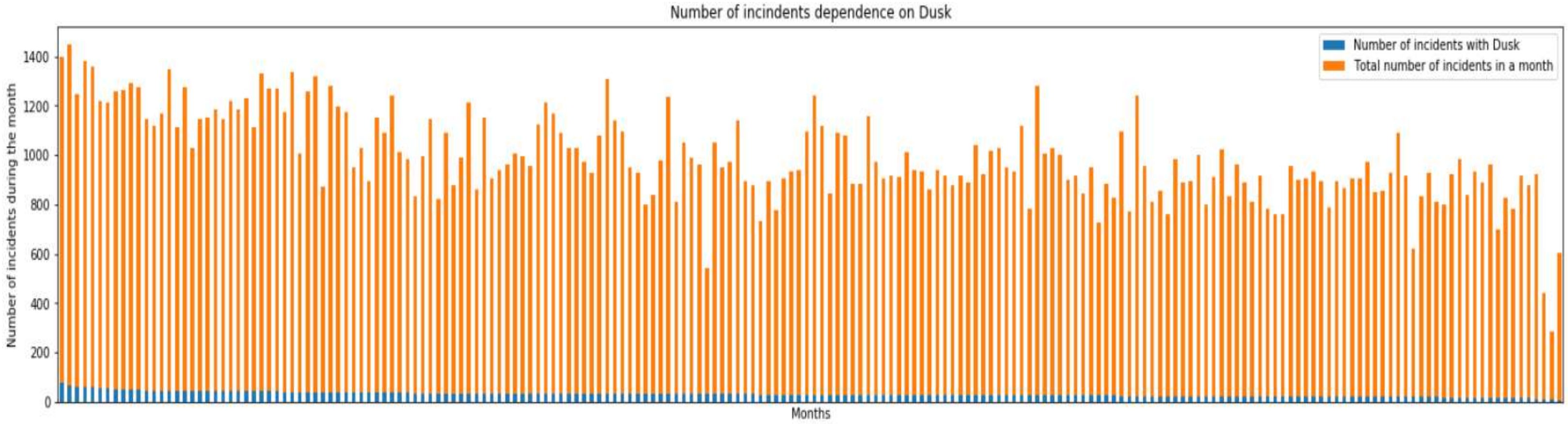


```
Correlation of Dusk with total amount of incidents during a day and average daily amount of incidents in a given month
                    Dusk
Dusk           1.000000
day_total      0.320841
month_average  0.204922
```

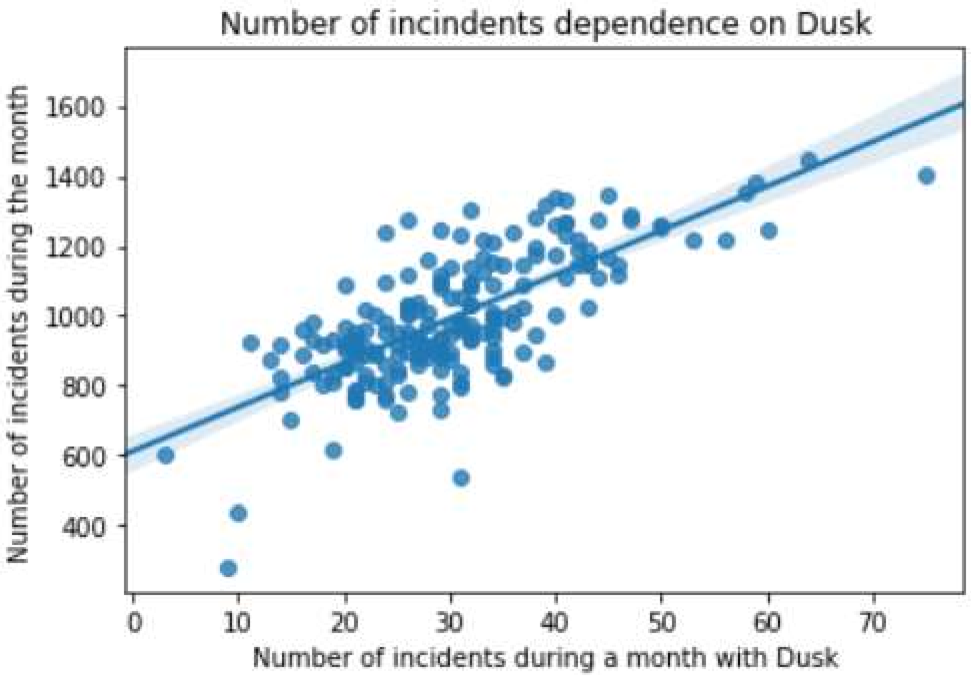


3.4.2 Light conditions: Dusk

The highest correlation of the total number of accidents is with Dusk light condition, the value is 0.726853. The contribution in the total amount of collision every month is low as can be seen on the diagram below:



Regression plot shows positive correlation:

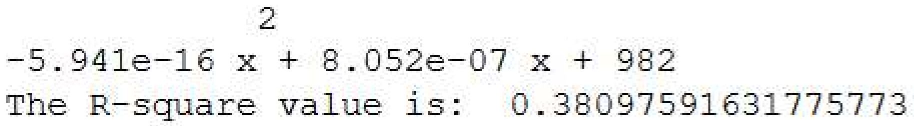


But this should not be considered due to low occurrences.

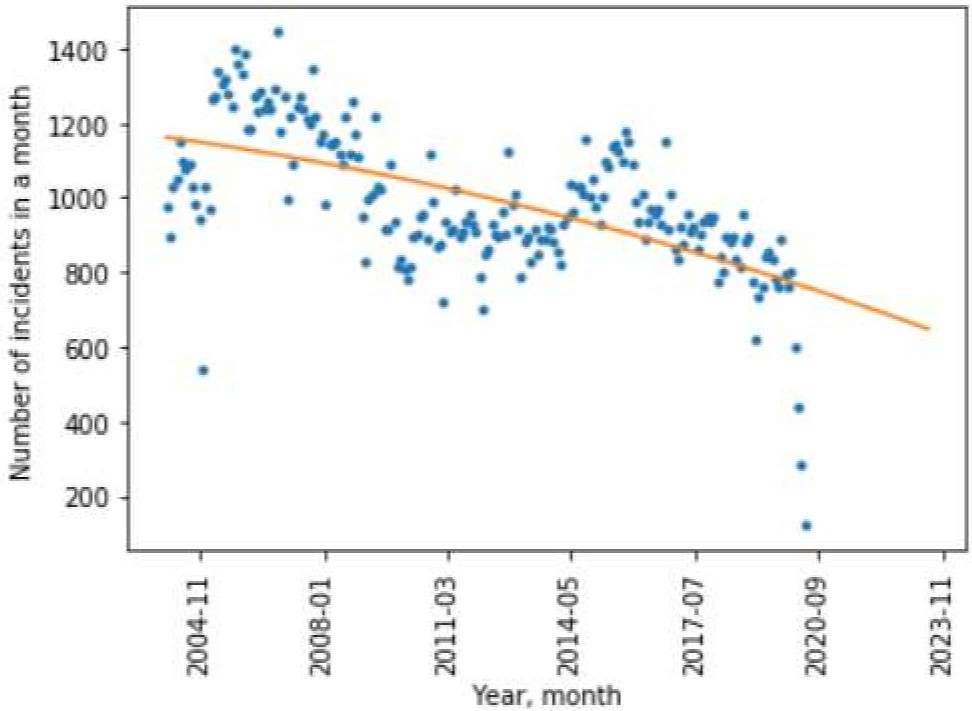
3.5 Regression analysis of the accidents frequency

It is hard to predict the frequency of accidents in the coming years with the regression models as the trend is decreasing and high degree polynomial functions give negative estimations, while low degree or linear regression give low R-square value.

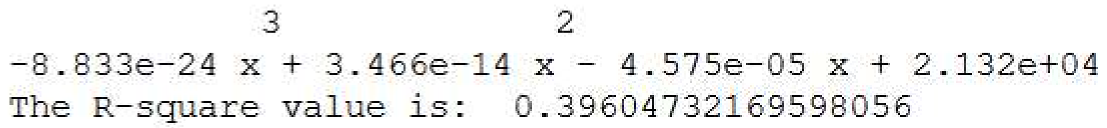
2nd degree polynomial regression:



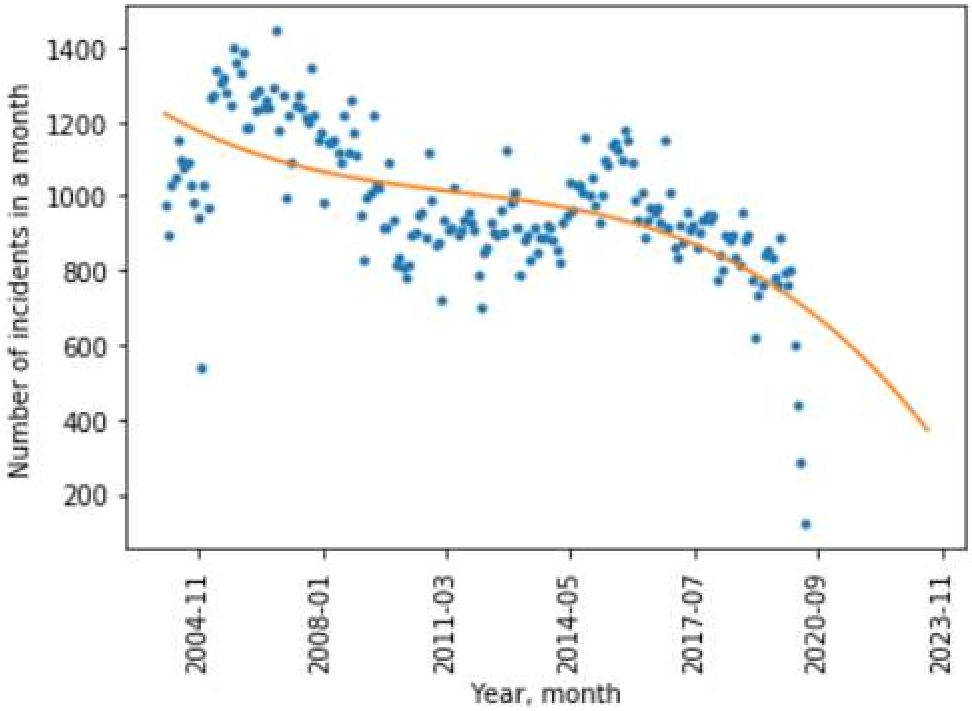
```
            2
-5.941e-16 x + 8.052e-07 x + 982
The R-square value is:  0.38097591631775773
```



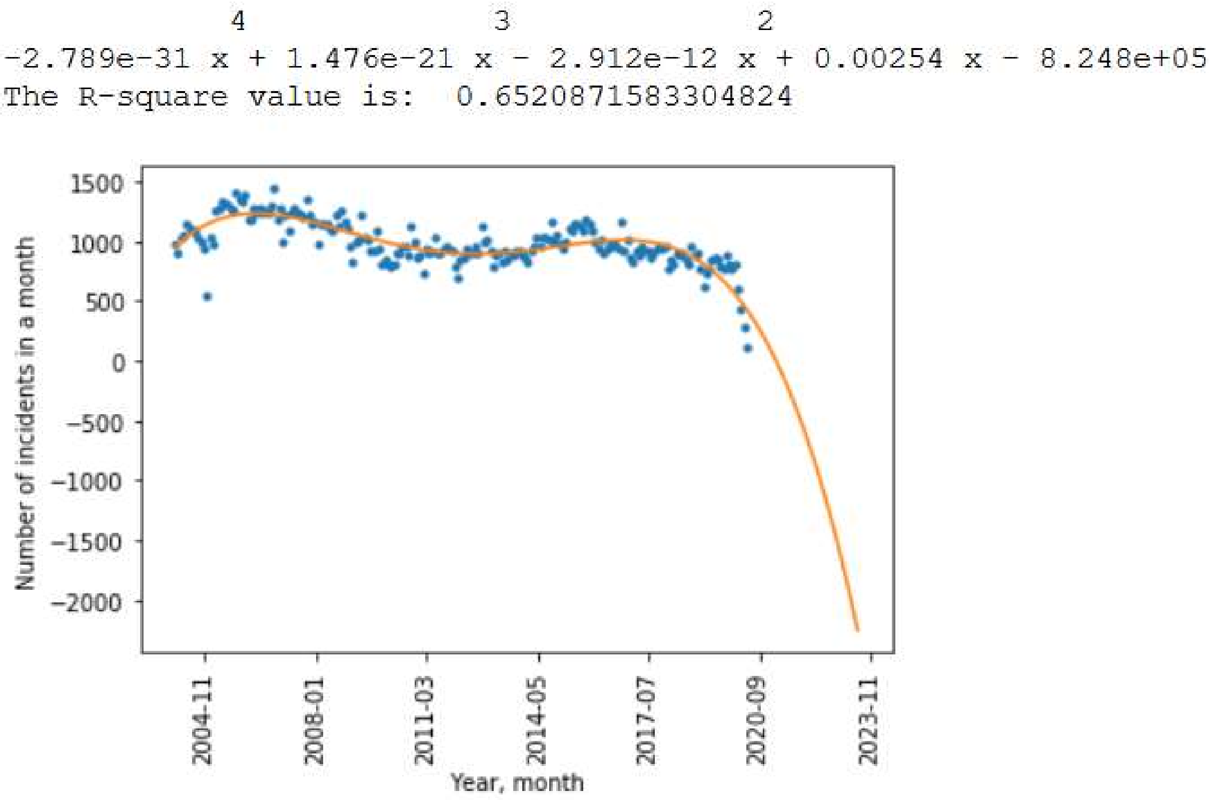
3rd degree polynomial regression:



```
            3              2
-8.833e-24 x + 3.466e-14 x - 4.575e-05 x + 2.132e+04
The R-square value is:  0.39604732169598056
```



4th degree polynomial regression:



```
            4              3              2
-2.789e-31 x + 1.476e-21 x - 2.912e-12 x + 0.00254 x - 8.248e+05
The R-square value is:  0.6520871583304824
```

4 Results

The data of car accidents in Seattle from 2014 till May 2020 was analyzed, 4 main attributes affecting the probability of an accident were discovered:

1) Location;
2) Weather conditions;
3) Road conditions;
4) Light conditions.

The study identified places in the city where most of the accidents occur. Authorities should check the traffic flows in those locations and determine what is causing the higher number of accidents there.

Probability of getting into an accident is higher when it is raining or snowing, and the road is wet or icy. Drivers must be more careful when driving in such environment. Other weather or road conditions do not have high influence on the probability.

Number of accidents tends to decrease in the last years. Linear and polynomial regression models is not suitable to predict the number of accidents in the next years. Low degree of polynomial function gives low R-square value, and high degree produces unrealistic estimations.

5 Discussion

The given dataset provides a lot of information that can be used to study statistics of the accidents and develop preventive measures to reduce the number. In particular, statistical analysis shows that some places require special attention, as the probability of getting into an accident is high. Perhaps this is related to the poor organization of the traffic flows.

Further observation and analysis of statistical data from problematic places can give more understanding of the reasons of higher amount of collisions and help reduce it. Every place has to be studied individually and is out of scope of the current work.

Polynomial and linear regression cannot be used to predict probability of an accident with categorical attributes on extended time scale. Neural networks algorithms are more suitable for this task.

6 Conclusion

Statistical data for car collisions for Seattle city was analyzed with statistical methods: correlation, regression plots, linear and polynomial regression analysis. Problematic places were shown on maps with identification of the number of incidents in each place. Recommendations were given with regards to improving the situation and further studies were suggested.