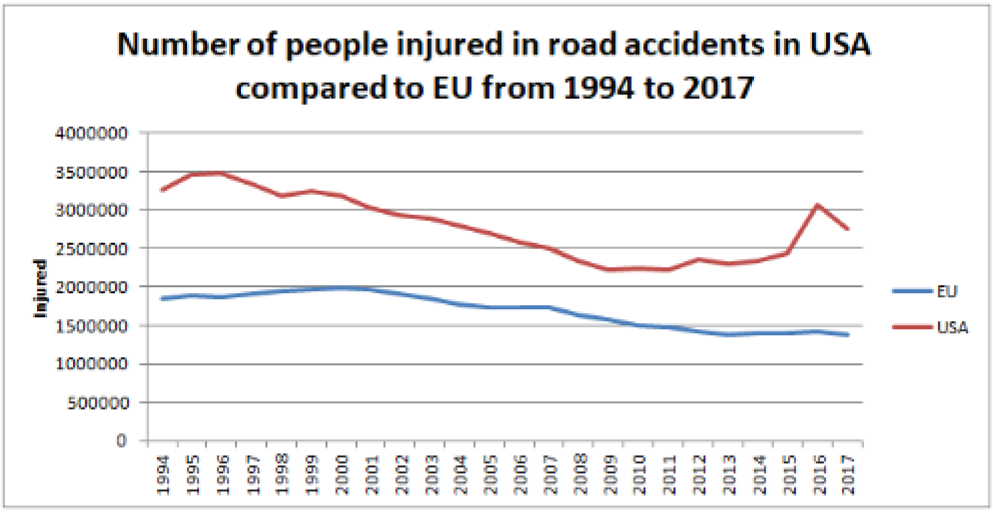
ANALYSIS OF COLLISIONS IN SEATTLE

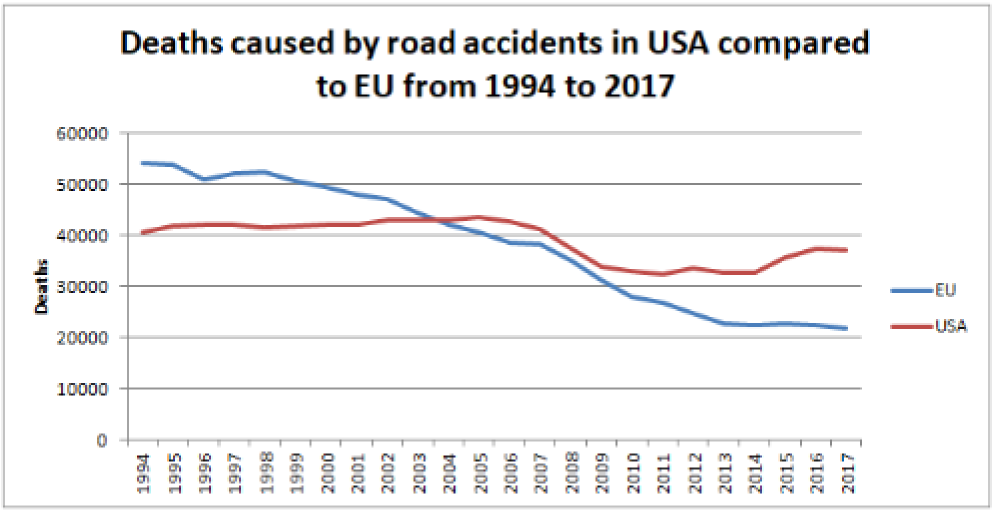
INTRODUCTION/PROBLEM PRESENTATION

As data from OECD suggest, there is significantly **more injuries caused by road accidents in US than in all European Unions' countries**. Although until 2011 number of accidents related injuries regularly decreased, since 2012 rising trend has come back, while EU succeeded in reduction of this negative phenomenon.



Source: Own study based on data available at: https://data.oecd.org/transport/road-accidents.htm

Moreover, whereas in EU number of deaths in road accidents has been decreasing over time, the contrary trend can be noticed in US where significantly declining trend is visible only in years 2006-2011. **Lately, more and more deaths in this country are caused by road accidents.**



Source: Own study based on data available at: https://data.oecd.org/transport/road-accidents.htm

DATA

As a result, it may be useful to take into consideration different conditions and dependencies, e.g. weather or road conditions, influence of illicit stimulants, drivers' inattention, speeding, pedestrians' behaviour or light conditions to determine which of them **contributes the most to the number of accidents' injuries or fatalities and try to reduce number of accidents and their negative consequences.**

It should be of special interest of different stakeholders, e.g. **traffic participants** themselves or **authorities**. People being aware of dangers would be available to avoid them, stay healthy and save money spent on repairs, and country could save money spent on people's treatment and experience their gratitude and recognition.

It could also help in projecting some social campaigns aimed at making people awared of dangers or even deploying of an early warning system that could alert people if some hazardous weather or other negative conditions are expected.

To achieve the goal of explaining potential contributors to accidents and their consequent injuries or deaths, **Seattle's collision dataset** will be used. It contains following attributes:

OBJECTID	ObjectID	ESRI unique identifier
SHAPE	Geometry	ESRI geometry field
INCKEY	Long	A unique key for the incident
COLDETKEY	Long	Secondary key for the incident
ADDRTYPE	Text, 12	Collision address type: • **Alley** • **Block** • **Intersection**
INTKEY	Double	Key that corresponds to the intersection associated with a collision

LOCATION	Text, 255	Description of the general location of the collision
EXCEPTRSNCODE	Text, 10	
EXCEPTRSNDESC	Text, 300	
SEVERITYCODE	Text, 100	A code that corresponds to the severity of the collision: • **3**—fatality • **2b**—serious injury • **2**—injury • **1**—prop damage • **0**—unknown
SEVERITYDESC	Text	A detailed description of the severity of the collision
COLLISIONTYPE	Text, 300	Collision type
PERSONCOUNT	Double	The total number of people involved in the collision
PEDCOUNT	Double	The number of pedestrians involved in the collision. This is entered by the state.
PEDCYLCOUNT	Double	The number of bicycles involved in the collision. This is entered by the state.
VEHCOUNT	Double	The number of vehicles involved in the collision. This is entered by the state.
INJURIES	Double	The number of total injuries in the collision. This is entered by the state.
SERIOUSINJURIES	Double	The number of serious injuries in the collision. This is entered by the state.
FATALITIES	Double	The number of fatalities in the collision. This is entered by the state.
INCDATE	Date	The date of the incident.
INCDTTM	Text, 30	The date and time of the incident.
JUNCTIONTYPE	Text, 300	Category of junction at which collision took place
SDOT_COLCODE	Text, 10	A code given to the collision by SDOT.
SDOT_COLDESC	Text, 300	A description of the collision corresponding to the collision code.
INATTENTIONIND	Text, 1	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Text, 10	Whether or not a driver involved was under the influence of drugs or alcohol.

WEATHER	Text, 300	A description of the weather conditions during the time of the collision.
ROADCOND	Text, 300	The condition of the road during the collision.
LIGHTCOND	Text, 300	The light conditions during the collision.
PEDROWNOTGRNT	Text, 1	Whether or not the pedestrian right of way was not granted. (Y/N)
SDOTCOLNUM	Text, 10	A number given to the collision by SDOT.
SPEEDING	Text, 1	Whether or not speeding was a factor in the collision. (Y/N)
ST_COLCODE	Text, 10	A code provided by the state that describes the collision. For more information about these codes, please see the State Collision Code Dictionary.
ST_COLDESC	Text, 300	A description that corresponds to the state's coding designation.
SEGLANEKEY	Long	A key for the lane segment in which the collision occurred.
CROSSWALKKEY	Long	A key for the crosswalk at which the collision occurred.
HITPARKEDCAR	Text, 1	Whether or not the collision involved hitting a parked car. (Y/N)

DATA PREPROCESSING

Following features has been dropped at the beginning because they do not carry any useful information:

- LOCATION,
- SEVERITYDESC,
- COLLISIONTYPE,
- SDOT_COLCODE,
- SDOT_COLDESC,
- SDOTCOLNUM,
- ST_COLCODE,
- ST_COLDESC,
- SEGLANEKEY,
- CROSSWALKKEY,
- EXCEPTRSNCODE,
- EXCEPTRSNDESC,
- OBJECTID,
- INCKEY,
- COLDETKEY,
- REPORTNO,
- STATUS,
- ADDRYPE,
- INTKEY,
- SEVERITYCODE.1.

As a result of initial analysis following features were also dropped because of overwhelming lacks in values:

- PEDROWNOTGRNT,
- INATTENTIONIND.

Incorrect data types in dataset were taken care of and rows with missing values in any columns were removed. For smaller datasets we should rather impute them with mean and most frequent values but now that our dataset is so large, we can easily drop it as they were only 2-2,5% of our dataset(our ML algorithms will be computational expensive anyway).

FEATURES SELECTION

Goal of this analysis was to explain which conditions contribute to collisions and their severity and to discover their impact on severity, so it was determined to use only features that are known right after collision and are direct causes of collision or its severity. As a result, following features have been removed from further analysis:

- PERSONCOUNT,
- PEDCOUNT,
- PEDCYLCOUNT,
- VEHCOUNT.

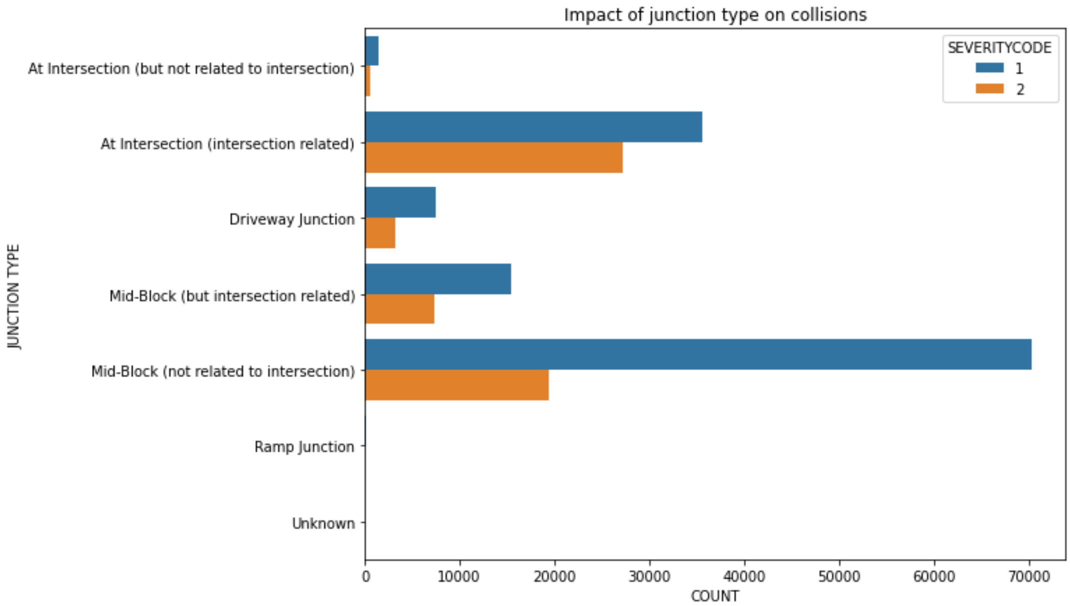
To sum up, only following features have been took into consideration in analysis:

- SEVERITYCODE – target variable,
- X – longitude,
- Y- latitude,
- INCDATE,
- INCDTTM,
- JUNCTIONTYPE,
- UNDERINFL,
- WEATHER,
- ROADCOND,
- LIGHTCOND,
- HITPARKEDCAR.

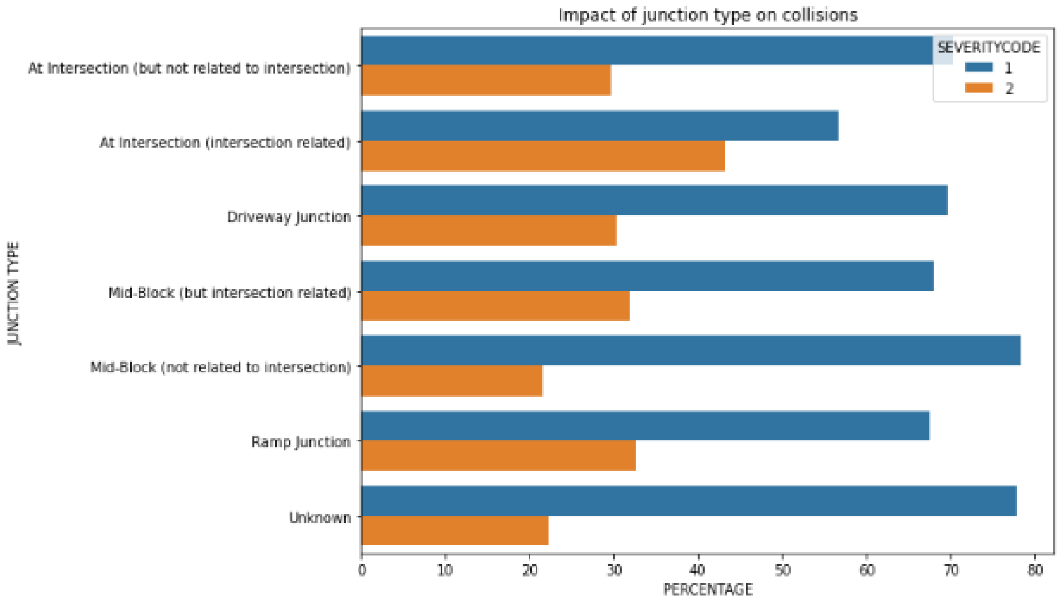
EXPLORATORY ANALYSIS

In this section data were grouped by particular independent variables and dependencies between different values of them and target variable were sought. Following conclusions were drawn:

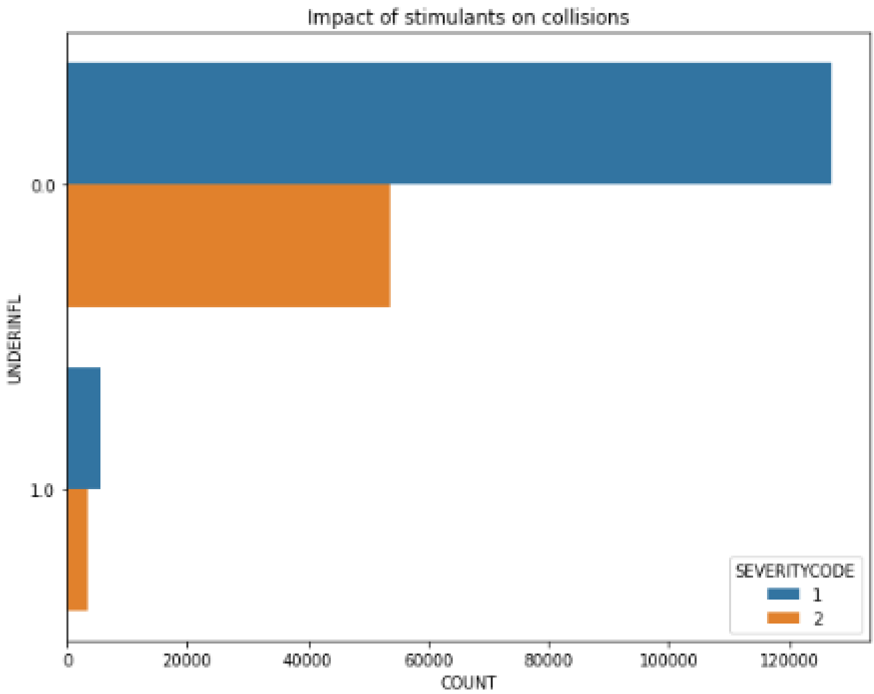
1. Overwhelming majority of collisions happen at midblock and at intersection.
2. Most of people get injured at intersection.



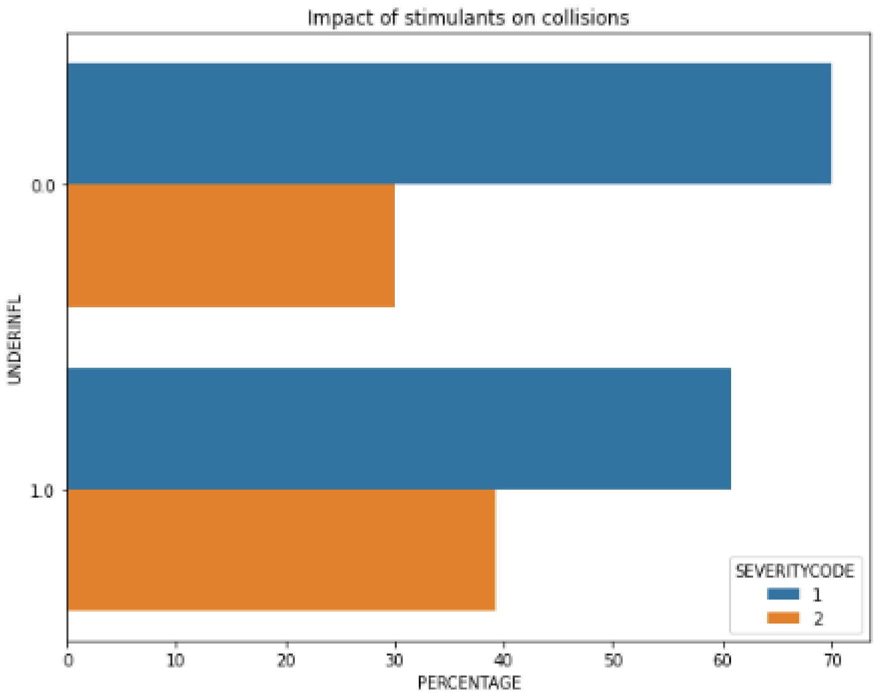
3. Only 20% of collisions at midblock end up with injuries whereas at intersection – about 40%.



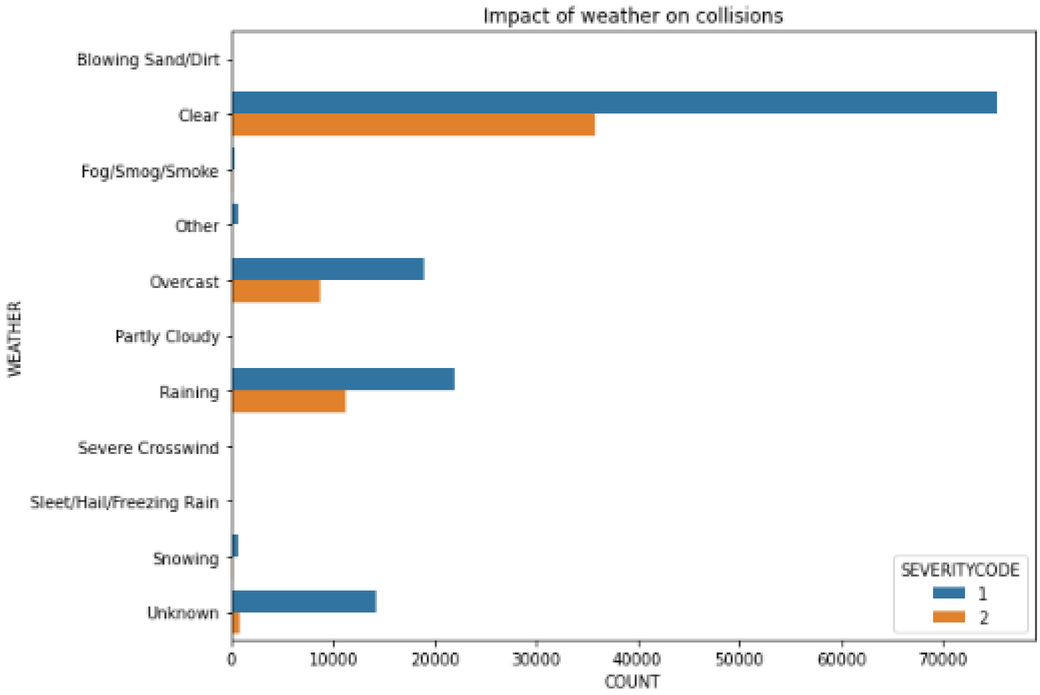
4. Collisions are rarely caused by drivers under influence of stimulants.



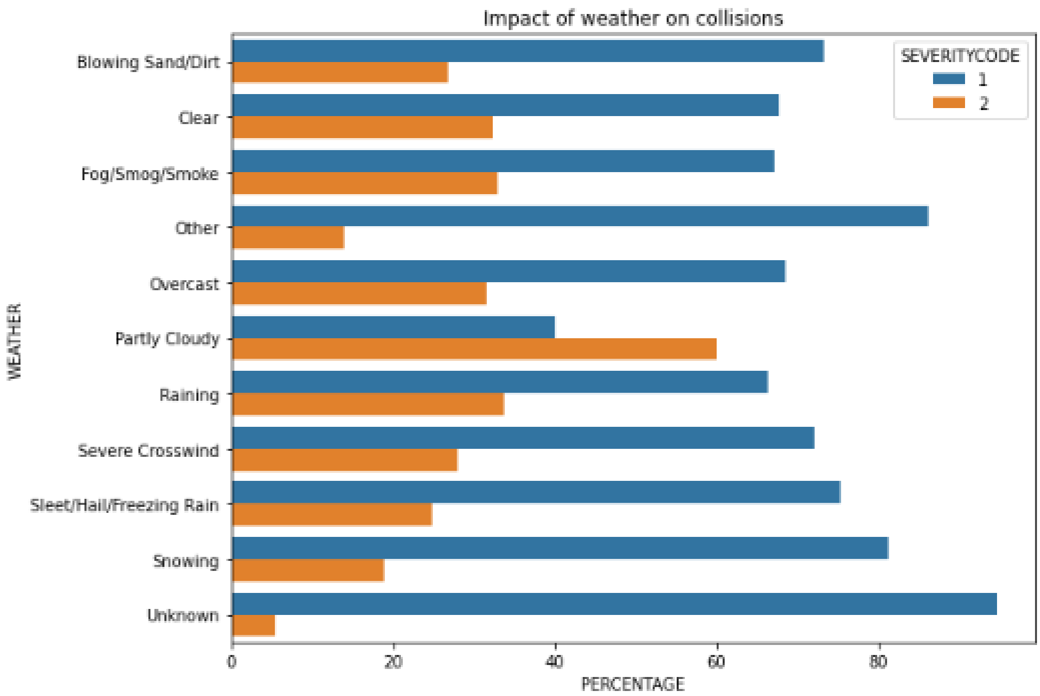
5. More people get injured in collisions caused by drivers under influence of stimulants compared to sober drivers.



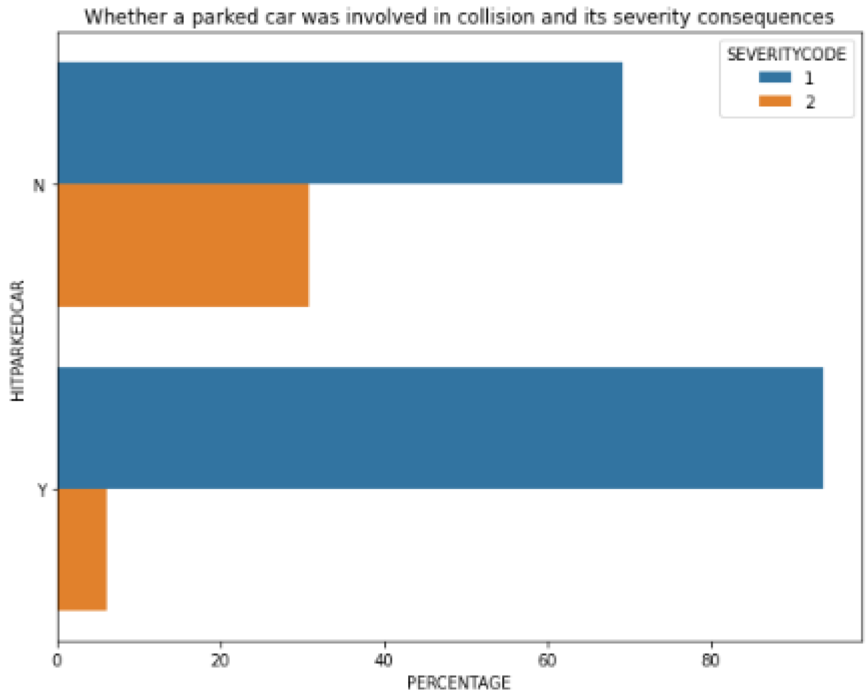
6. Most of collisions happen when the weather is clear, but it can be caused by the fact that days with such weather are the most frequent in a year. Taking into account that rainy days are probably rare, number of collisions during them seems to be quite significant.



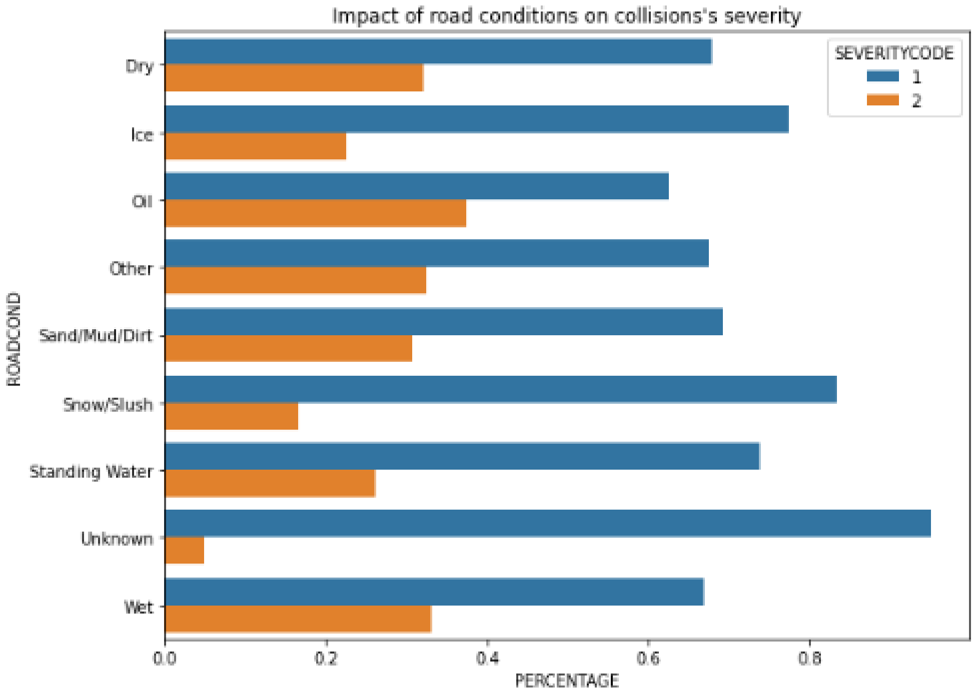
7. It can be noticed that weather conditions impact severity of collisions. Injuries are caused most frequently when it is partly cloudy and rainy, whereas the least of collisions end up with injuries when it is snowing.



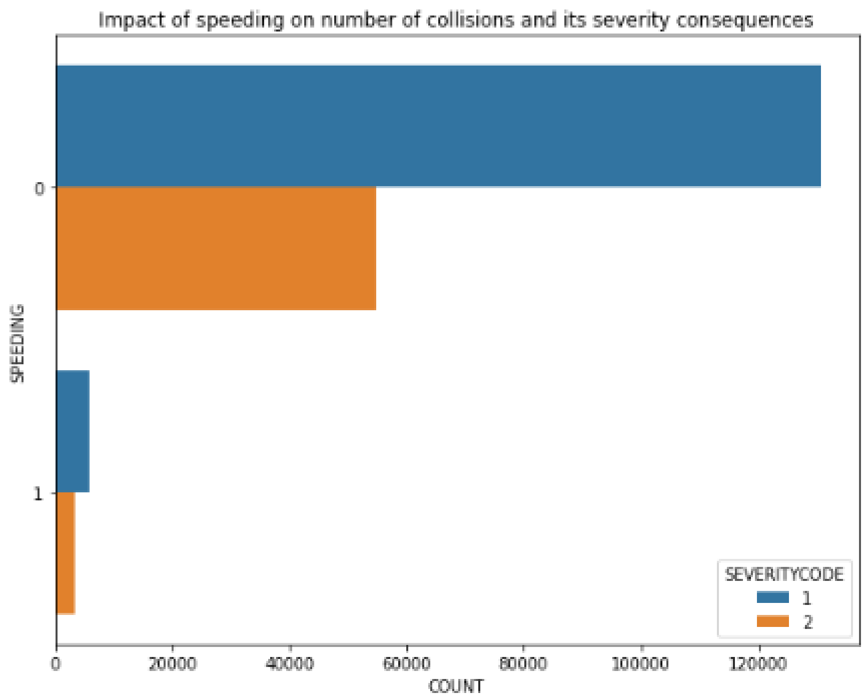
8. Collisions relatively rarely involve hitting a parked car. Moreover, in collisions involving that people get injured only in 6% of them.



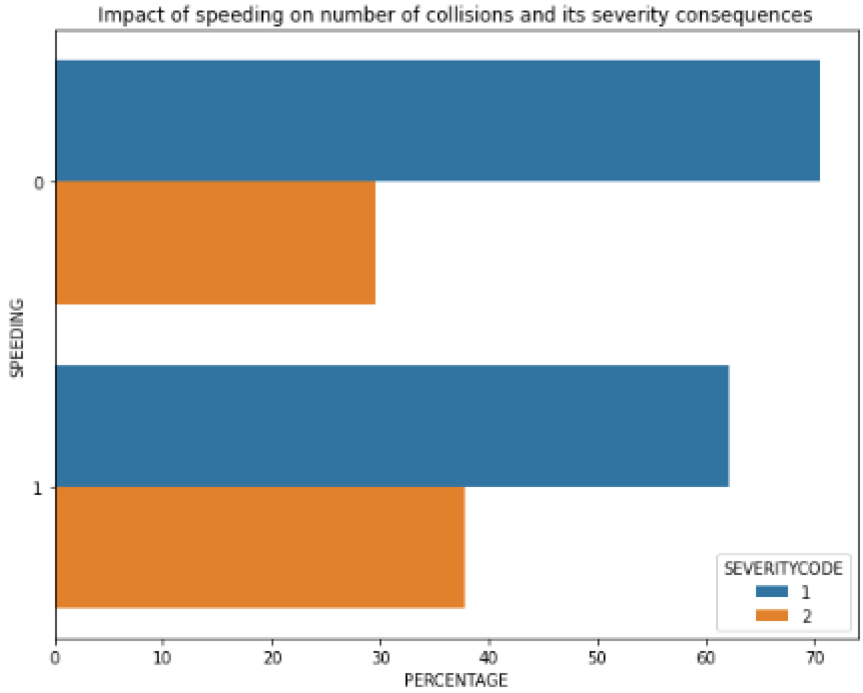
9. The least of collisions end up with injuries when roads are covered with snow. It can be caused by extensive attention paid by drivers and limited speed.



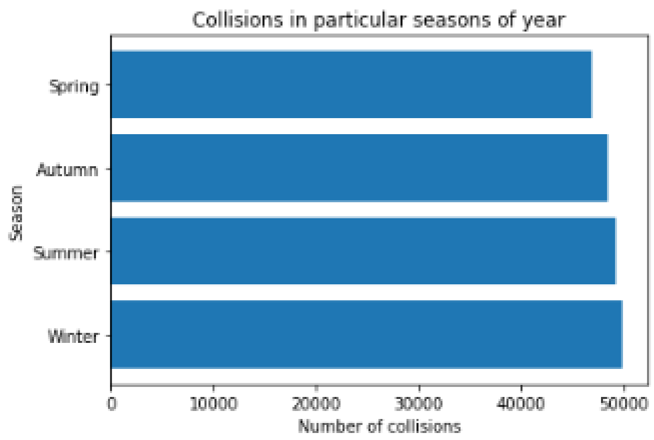
10. Only about 5% of collisions were caused by speeding drivers.

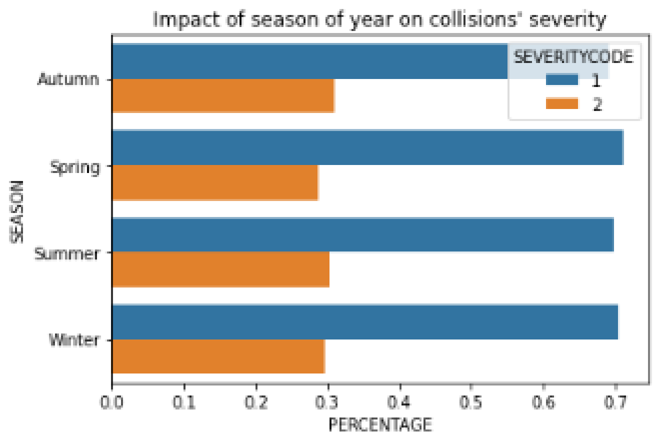


11. Collisions caused by speeding ended up with injuries most frequently (35% of them).



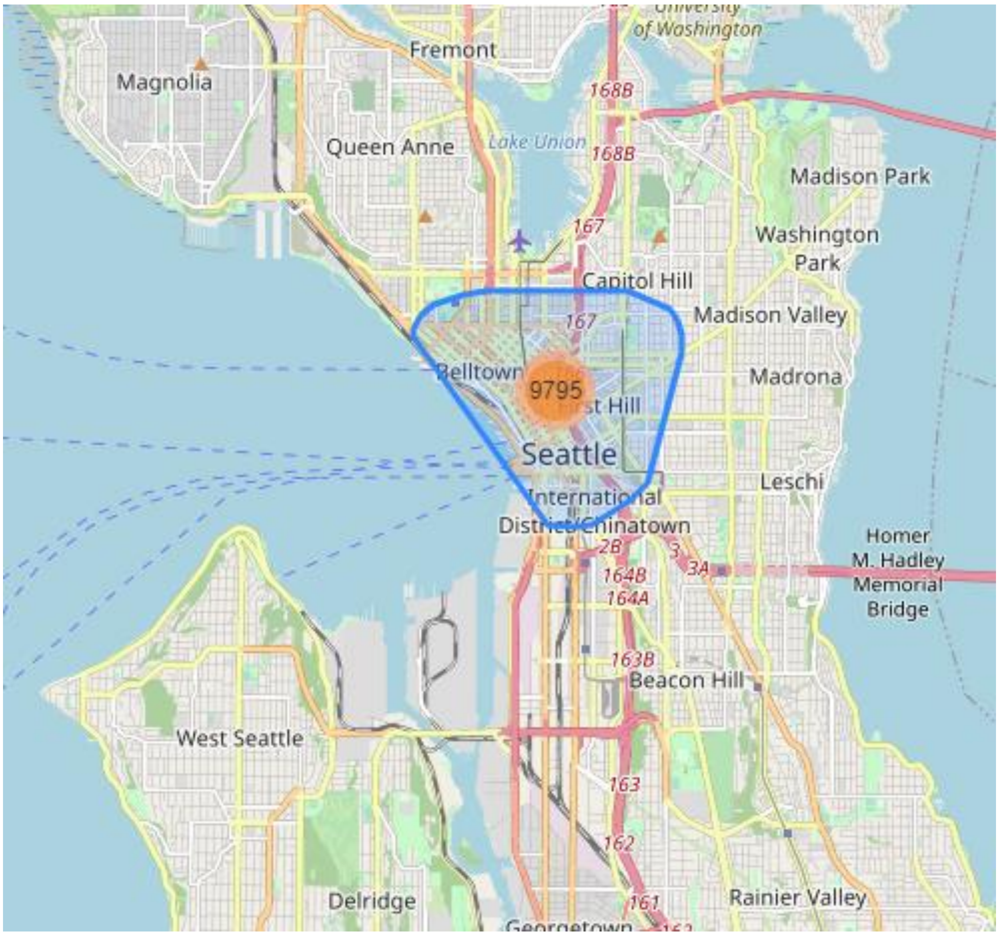
12. Seasons of year have insignificant impact on number of collisions and their severity.





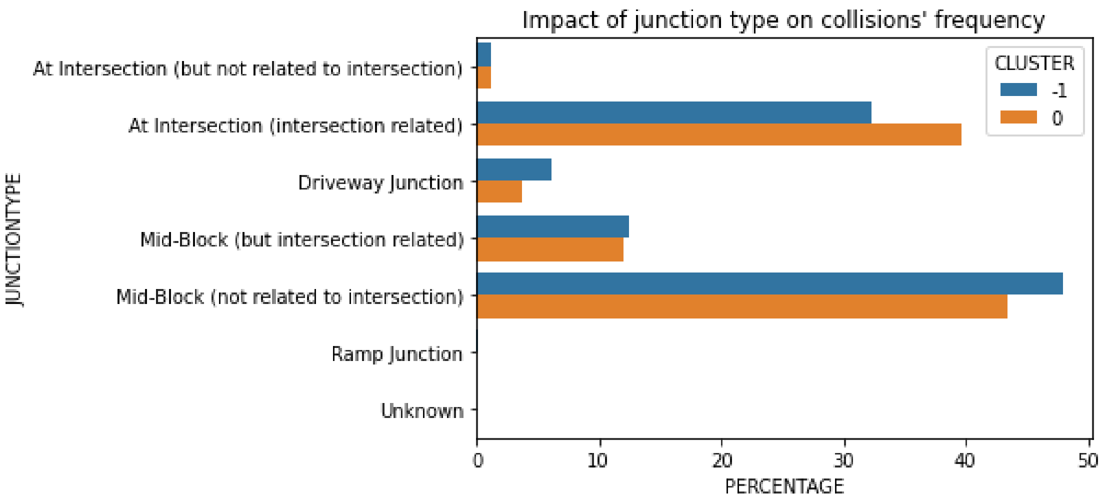
CLUSTERING LOCATIONS OF HIGH COLLISIONS DENSITY

DBSCAN algorithm was used to discover places in Seattle where the highest density of collisions occurs. Due to the fact that this algorithm is very memory expensive, it was decided to downsample original dataset to 50000 collisions with respect to their original fraction of severity. This algorithm was chosen from clustering methods because it enables to detect outliers and discover one main grouping. Moreover, haversine metric was used as it is recommended for spatial data analysis. 1000 collisions were supposed to be in a radius of 500m from a specific location in order to consider it a core point. 5944 such locations were discovered and all of them were clustered in the city center. All inliers were visualised using Folium library and below map represents an area on which they were located.

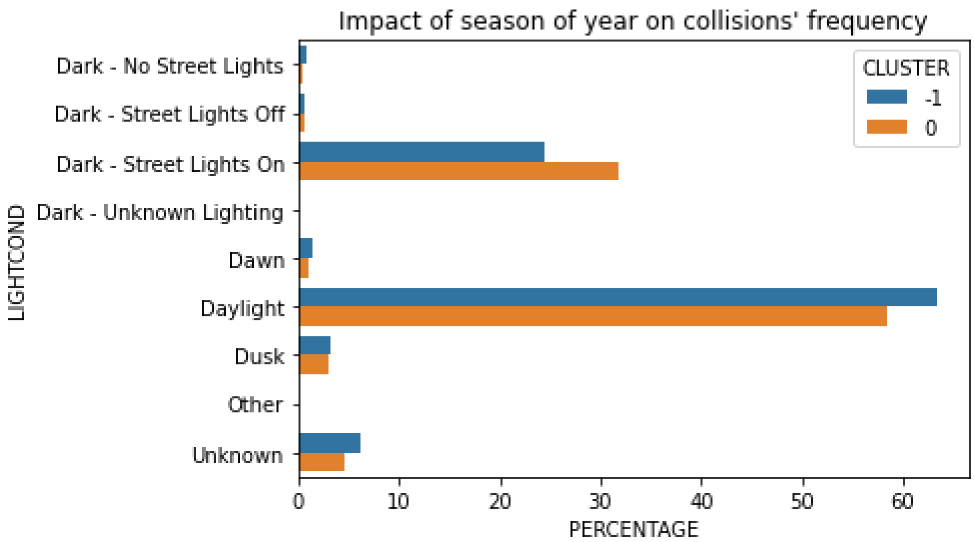


Analysis was conducted to determine if there are any differences between created cluster and other locations. Cluster 0 is the main grouping and cluster -1 contains outliers. Two conclusions were drawn:

- In city center there were more collisions at intersection and less at mid-block.



- In city center there were more collisions in the dark and less in the daylight.



COLLISIONS' CLASSIFICATION AND FEATURES' IMPORTANCE

In this section two algorithms were trained to be able to predict collisions' severity:

- Logistic Regression,
- Random Forest.

Both of them managed to do that quite well and achieved similar accuracy and F1 scores on both training and test set:

- training set

	LogisticRegression	RandomForest
Accuracy	69.026432	69.393620
F1 score	81.442420	81.653033

- test set

	LogisticRegression	RandomForest
Accuracy	68.760064	68.807263
F1 score	81.298408	81.308955

Moreover, random forest was used in order to determine features importance. The following table presents them:

	Feature	Importance
1	JUNCTIONTYPE	0.474961
3	ROADCOND	0.134914
4	LIGHTCOND	0.133125
2	WEATHER	0.117241
5	HITPARKEDCAR	0.061424
6	SPEEDING	0.041196
0	UNDERINFL	0.037139

Indisputably, junction type has the highest impact on collisions' severity. It carries about 47% of discriminative information. Secondly, there goes road conditions, light conditions and weather.

SUMMARY

The goal of this study was too identify which factors contribute to number of collisions and their severity. It was discovered that junction type is the main of them and carries 47% of discriminative information. Such findings can help traffic participants be more aware of potential dangers and pay more attention in the most hazard places and conditions.

Two machine learning algorithms were conducted to try to predict severity of collisions. Both of them performed similarly and satisfactorily with accuracy of about 69% and F1 score of about 81%.

During the analysis, it was found out that city center's neighbourhood is a region of the highest density of collisions, but only two significant differences were spotted between collisions in city center and other locations. Namely, in city center more collisions occurred at intersections and in the dark(mainly with street lights on) than in other locations. It should draw attention of authorities in order to make intersections safer and check out whether street lights perform well. The whole Seattle should also observe collisions at mid-block and try to detect their causes.

Number of injuries and deaths caused by road accidents in US compared for example to EU is unsatisfactory and can be certainly reduced. This analysis may help to understand what are the causes of such a situation and draw authorities' attention to take action. Intervention will cut costs for both traffic participants and authorities who are bound to get people's gratitude and recognition for saving their money, health and lives
.