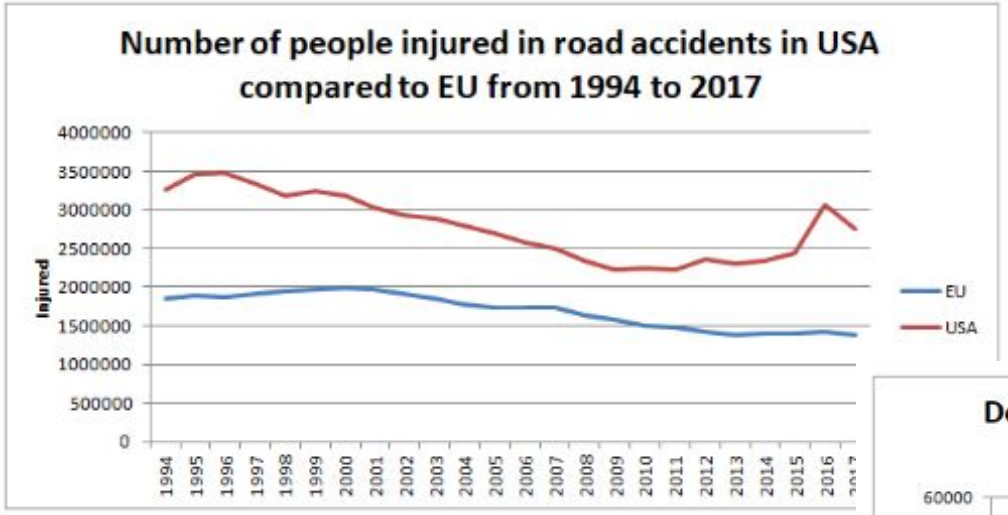
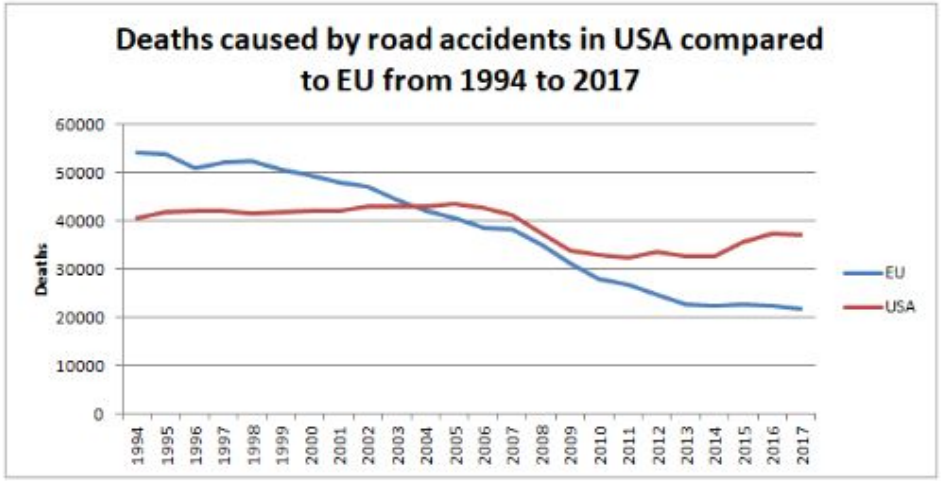
ANALYSIS OF COLLISIONS IN SEATTLE

TRENDS OF INJURIES AND DEATHS IN ACCIDENTS





GOAL OF ANALYSIS AND ITS RECIPIENTS

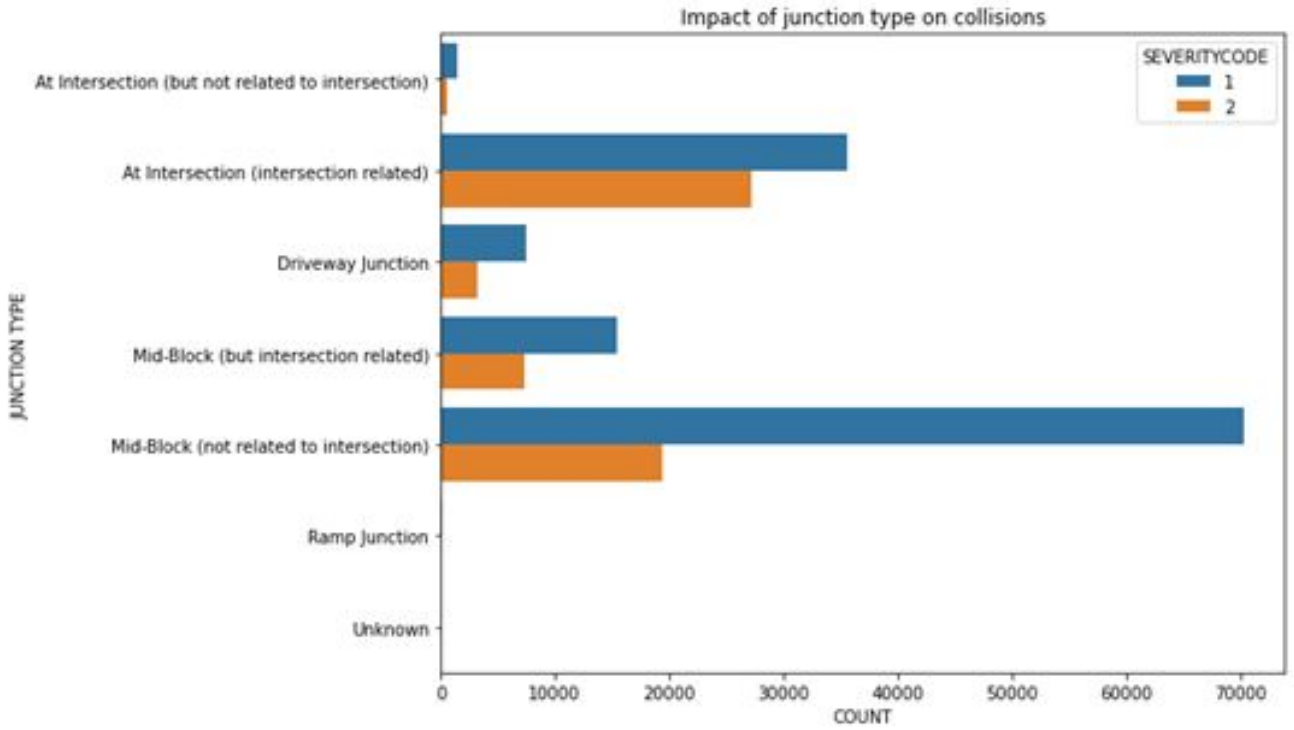
- The goal is to detect factors that contribute the most to the number of collisions and their severity which could help to reduce their frequency and negative consequences.
- Results of analysis could be helpful for authorities and traffic participants and enable them to avoid unnecessary costs and stay healthy.

DATASET AND DATA PREPROCESSING

- Seattle collisions dataset containing of 35 features and 1 target variable was used
- Rows with missing values were dropped
- Incorrect data types were handled
- Following features were taken into further consideration:
  - X - longitude
  - Y - latitude
  - INCDATE
  - INCDTTM
  - JUNCTIONTYPE
  - UNDERINFL
  - WEATHER
  - ROADCOND
  - LIGHTCOND
  - HITPARKEDCAR

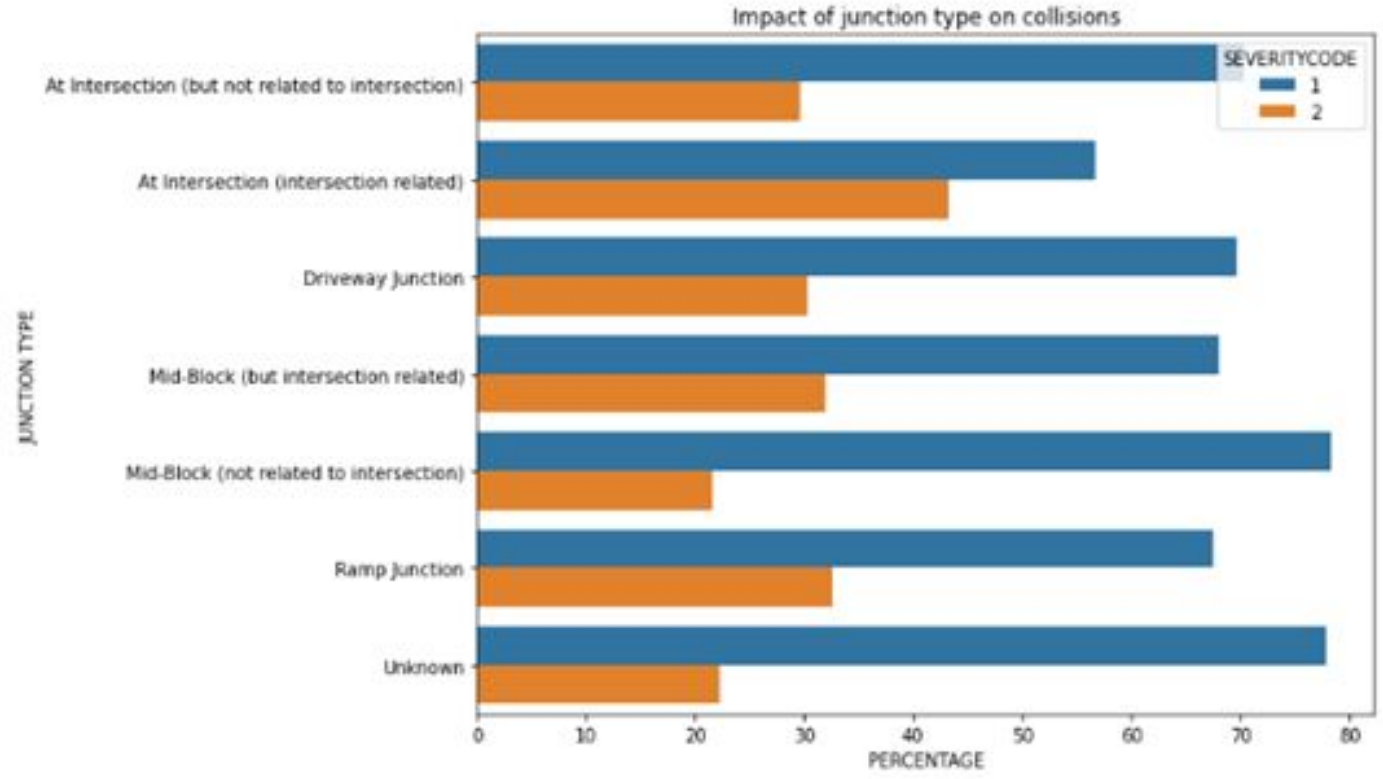
EXPLORATORY ANALYSIS

OVERWHELMING MAJORITY OF COLLISIONS HAPPEN AT MIDBLOCK AND AT INTERSECTION

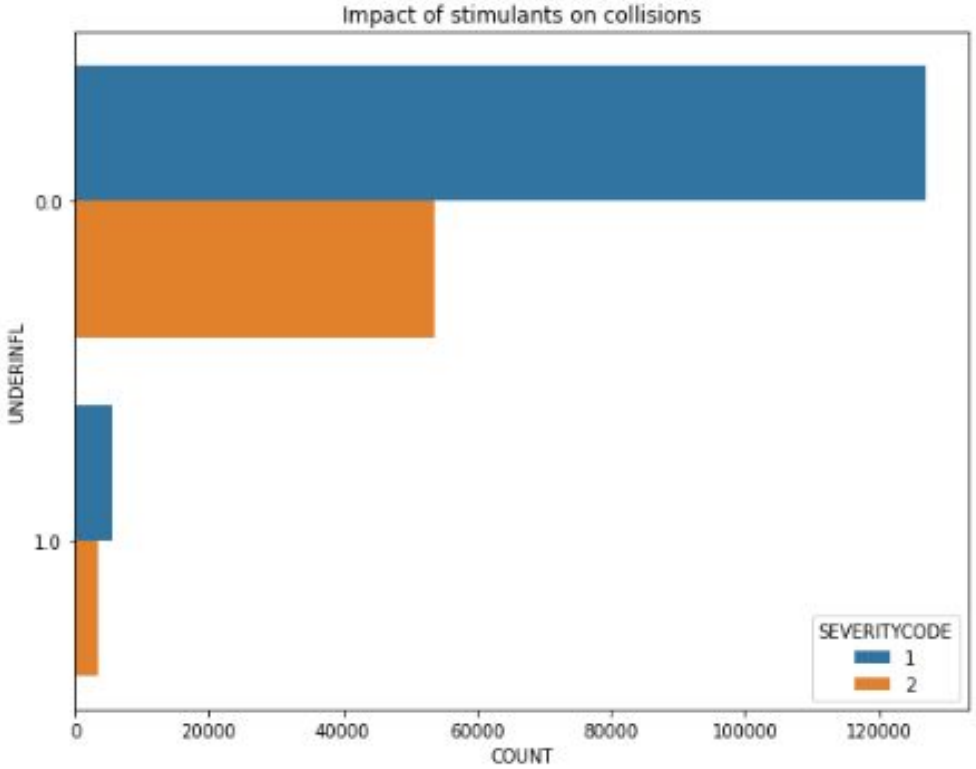


MOST OF PEOPLE GET INJURED AT INTERSECTION

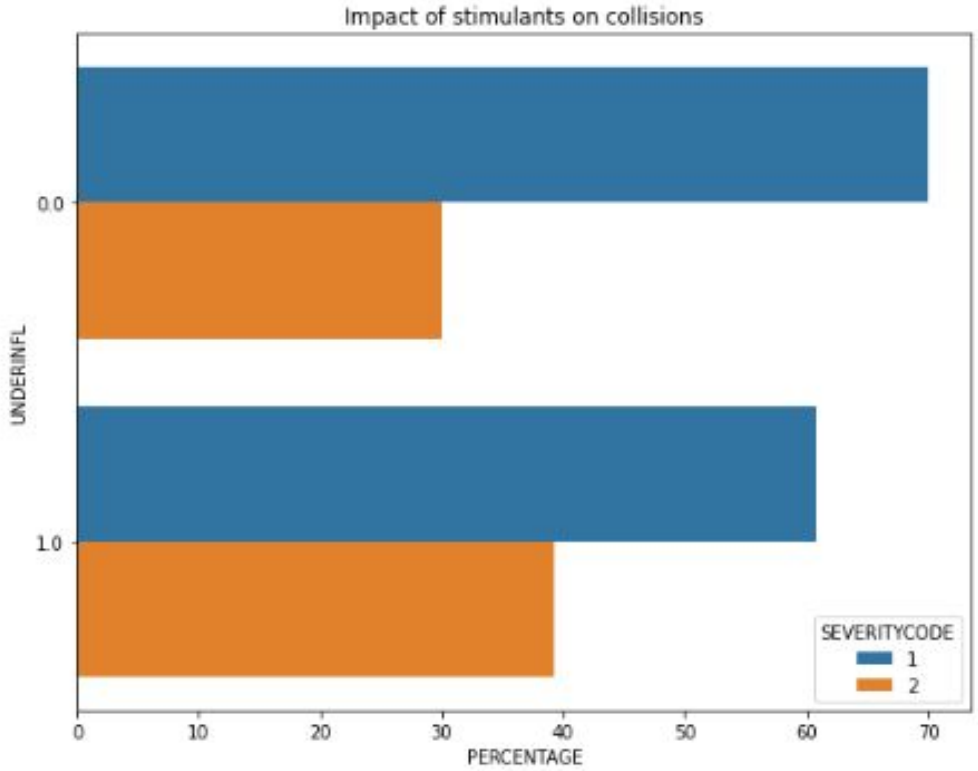
ONLY 20% OF COLLISIONS AT MIDBLOCK END UP WITH INJURIES WHEREAS AT INTERSECTION - ABOUT 40%



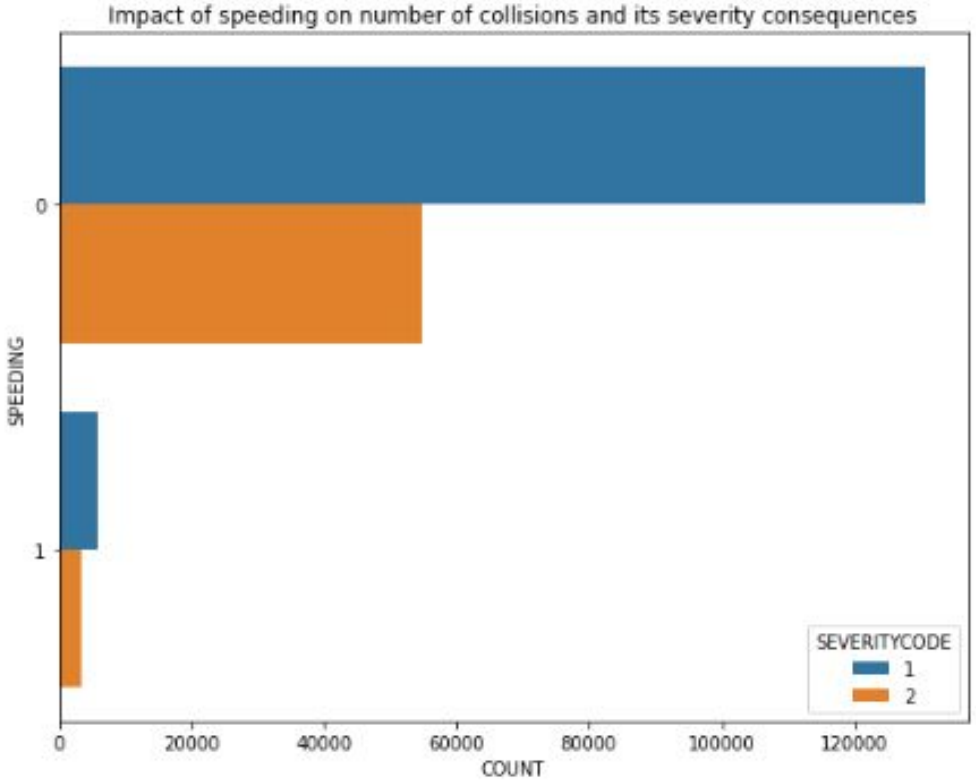
COLLISIONS ARE RARELY CAUSED BY DRIVERS UNDER INFLUENCE OF STIMULANTS



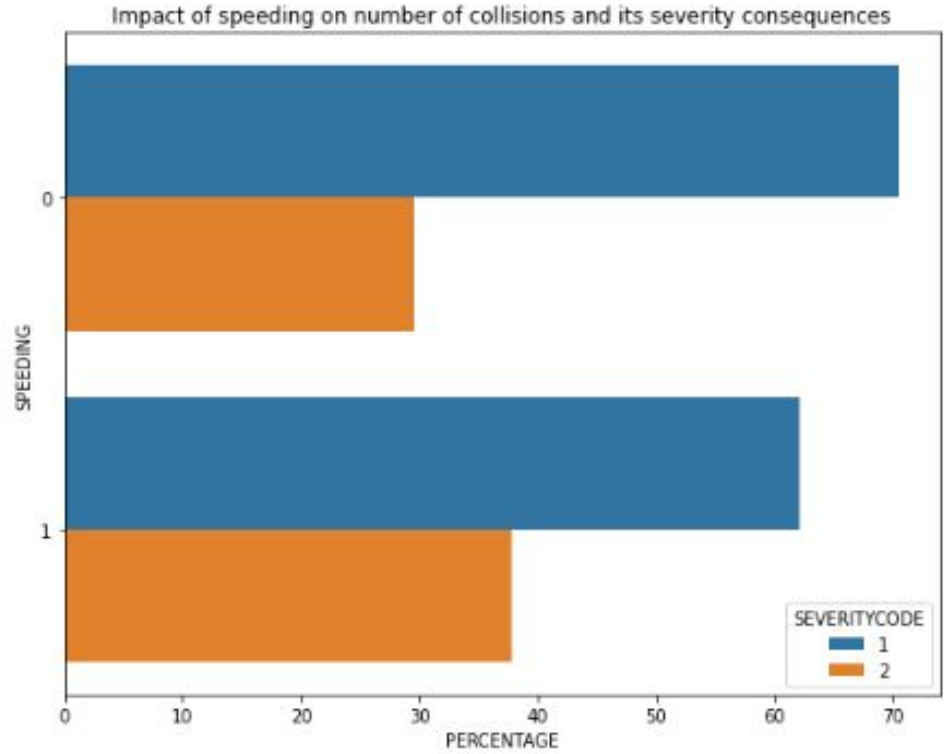
MORE PEOPLE GET INJURED IN COLLISIONS CAUSED BY DRIVERS UNDER INFLUENCE OF STIMULANTS COMPARED TO SOBER DRIVERS



ONLY 5% OF COLLISIONS WERE CAUSED BY SPEEDING DRIVERS



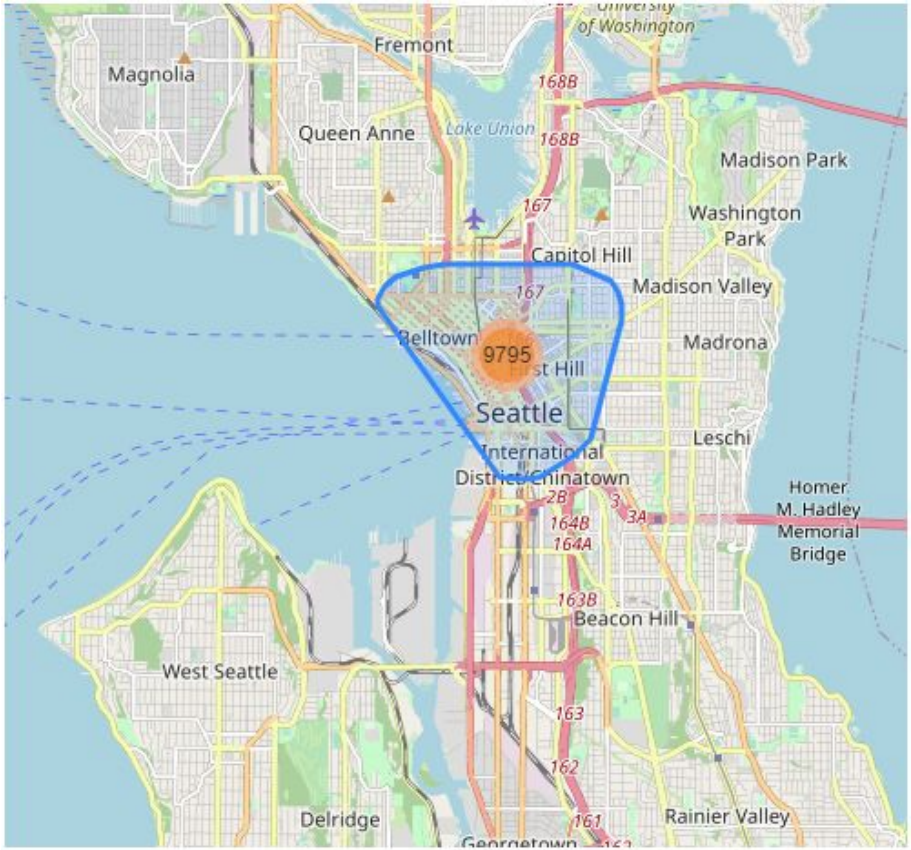
COLLISIONS CAUSED BY SPEEDING ENDED UP WITH INJURIES MORE FREQUENTLY(35% OF THEM)



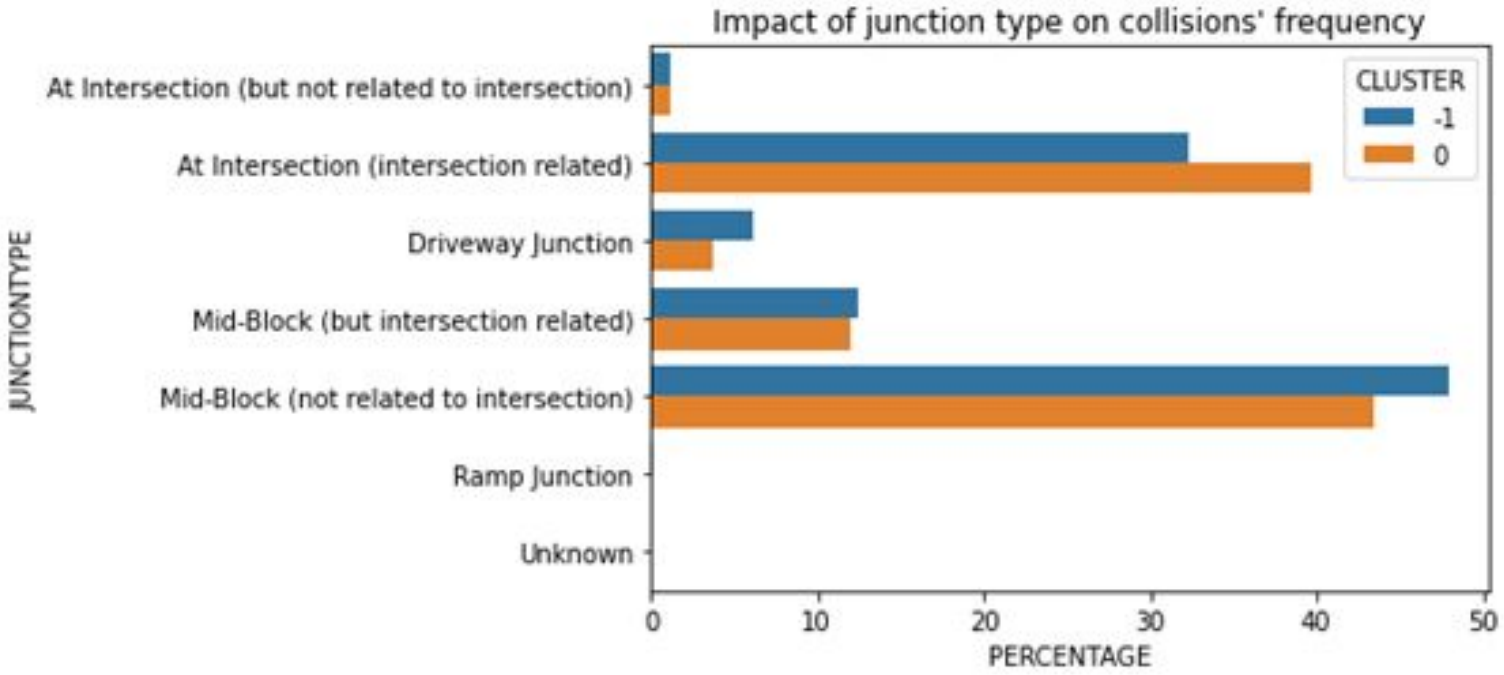
CLUSTERING LOCATIONS OF HIGH COLLISIONS DENSITY

- DBSCAN algorithm with haversine metric was used
- Original dataset was downsampled to 50000 collisions with fraction maintenance as DBSCAN is very memory expensive
- 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

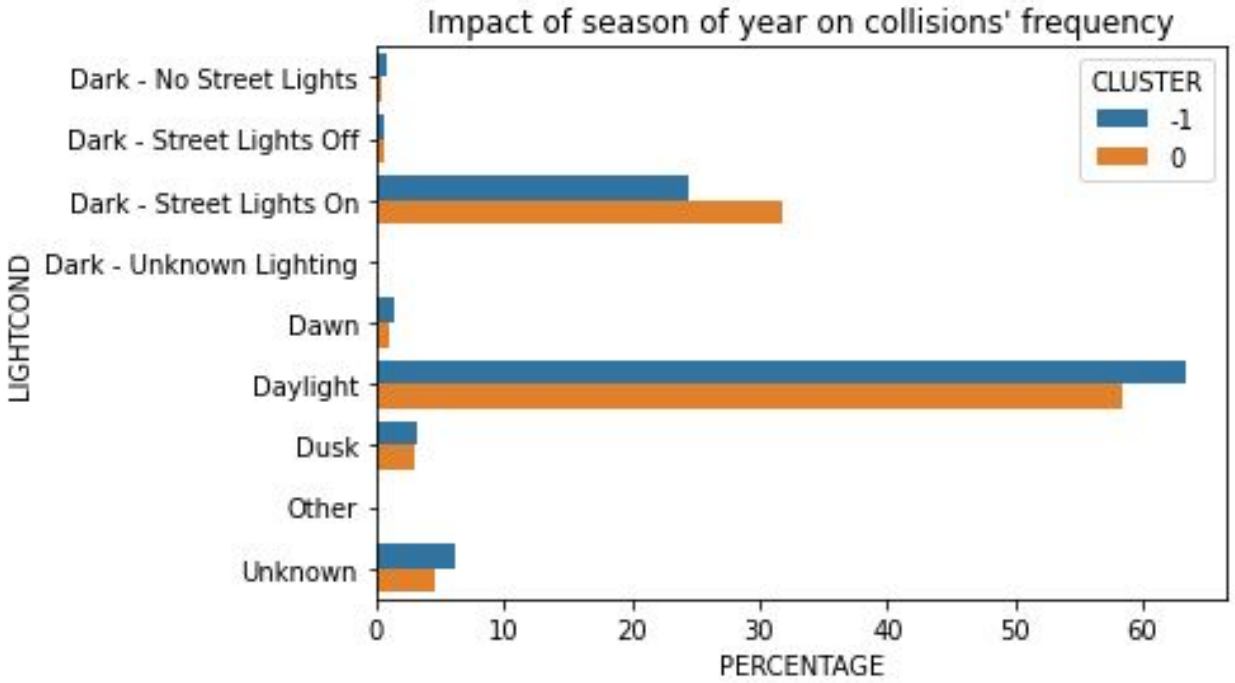
ALL INLIERS WERE VISUALISED USING FOLIUM LIBRARY



IN CITY CENTER THERE WERE MORE COLLISIONS AT INTERSECTION AND LESS AT MID-BLOCK(CLUSTER 0 IS THE MAIN GROUPING AND CLUSTER -1 CONTAINS OUTLIERS)



IN CITY CENTER THERE WERE MORE COLLISIONS IN THE DARK AND LESS IN THE DAYLIGHT



COLLISIONS' CLASSIFICATION AND FEATURES IMPORTANCE

LOGISTIC REGRESSION AND RANDOM FOREST

BOTH ALGORITHMS MANAGED TO PREDICT COLLISIONS' SEVERITY 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

RANDOM FOREST FOR GUESSING FEATURES IMPORTANCE

	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,

SUMMARY

- Findings can help traffic participants be more aware of potential dangers and pay more attention in the most hazard places and conditions
- High frequency of collisions in city center and higher fraction of them caused at intersections and in the dark(mainly with street lights on) than in other locations should draw attention of authorities
- Also, the whole Seattle should observe collisions at mid-block and intersections and try to detect their causes
- Number of injuries and deaths caused by road accidents in US compared to EU is unsatisfactory and can be certainly reduced
- This analysis may help to understand what are the causes 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.