Predicting car accident severity

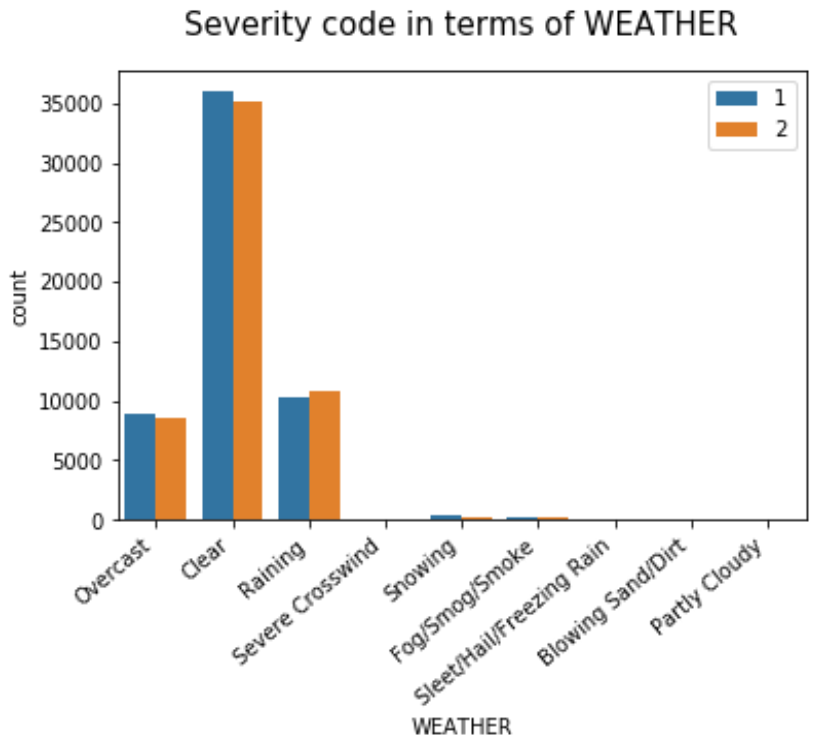
Predicting car accident severity would have several applications

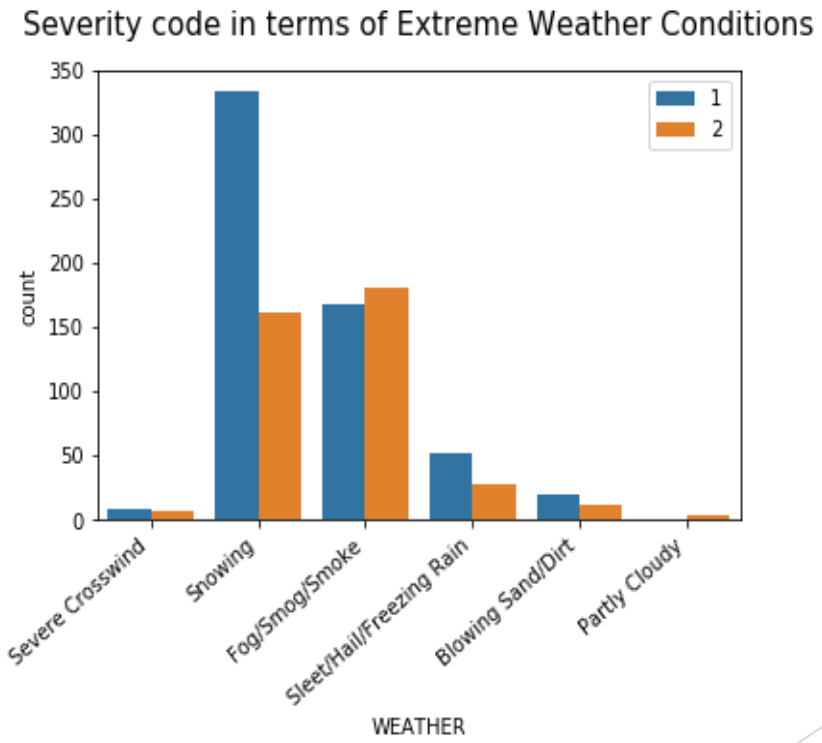
- Authorities could use this knowledge to take measures in order to reduce the number of severe accidents
- Navigation applications would improve their recommendations based on the probability that a severe accident might occur in any potential journey

Data acquisition and cleaning

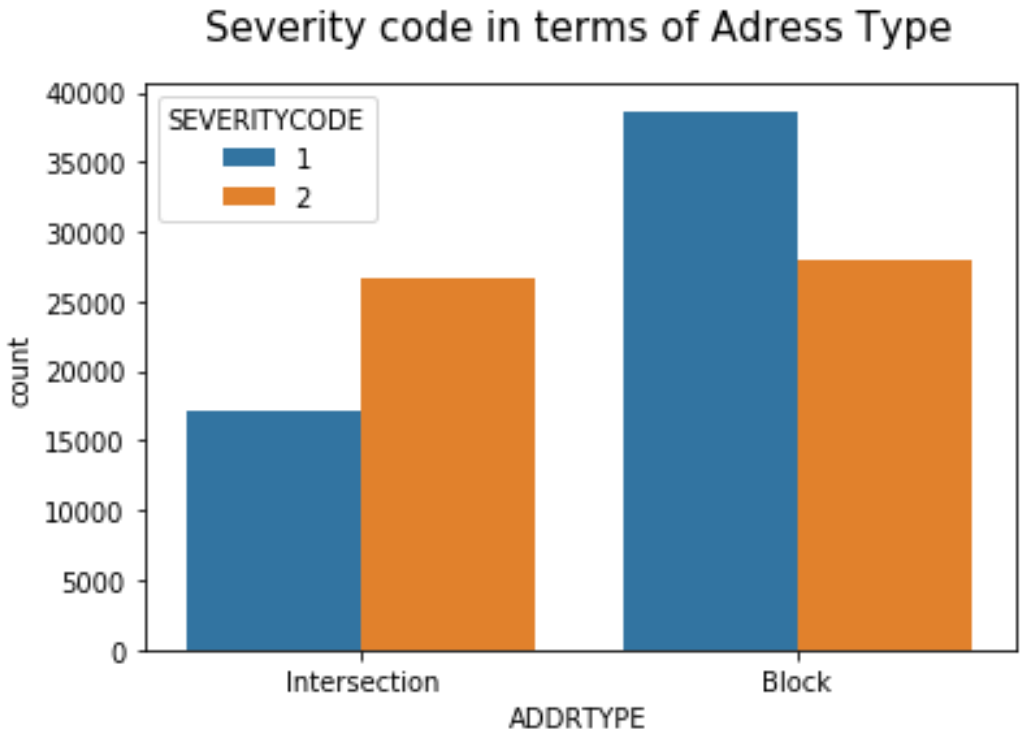
- Rows with empty values are deleted from the dataset
- The dataset is not balanced: severity 2 accidents (which caused injuries) are undersampled in the dataset. In order to solve this severity 1 rows are undersampled
- Features which are codes of identification or describe the accident are eliminated from the dataset

Weather and road conditions have little impact on severity, when conditions are not extreme

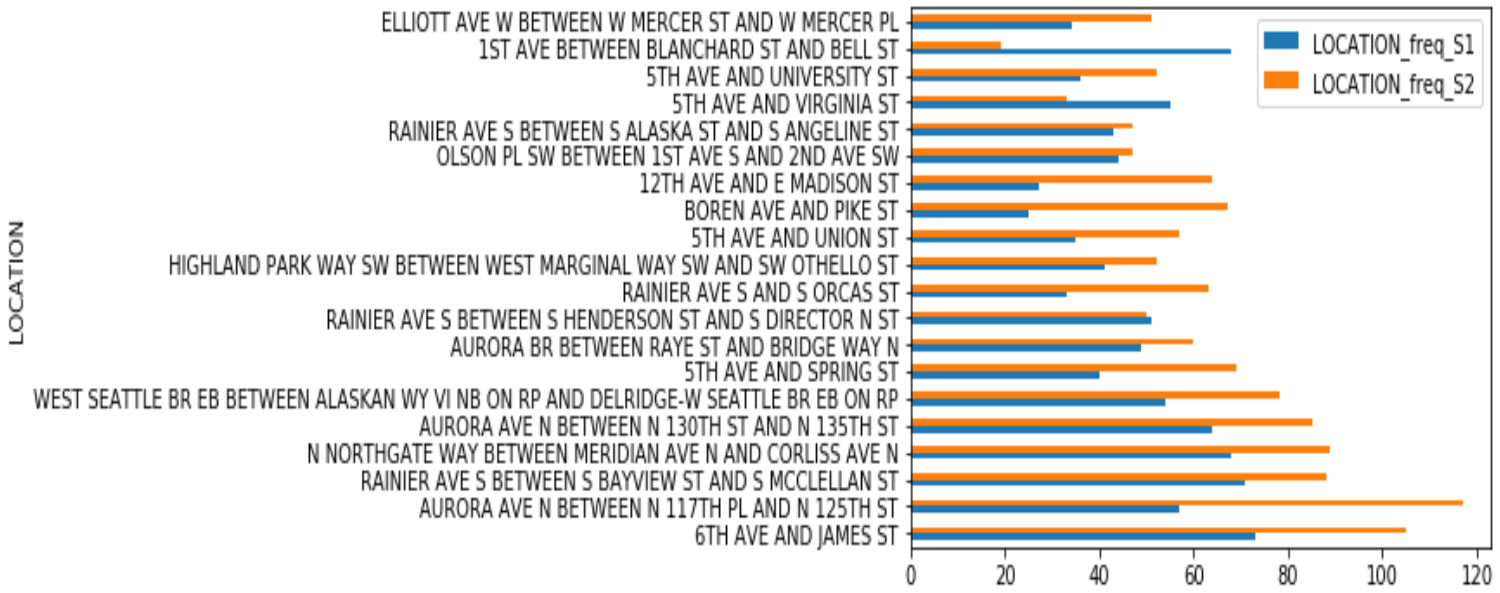




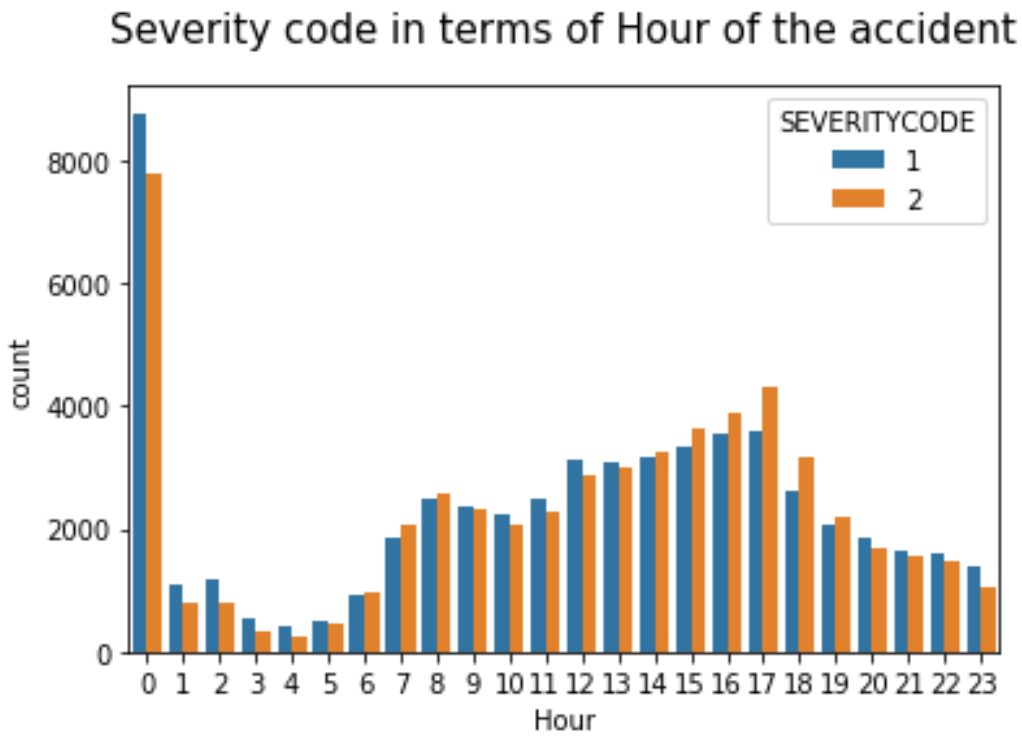
Accidents in intersections tend to be more severe



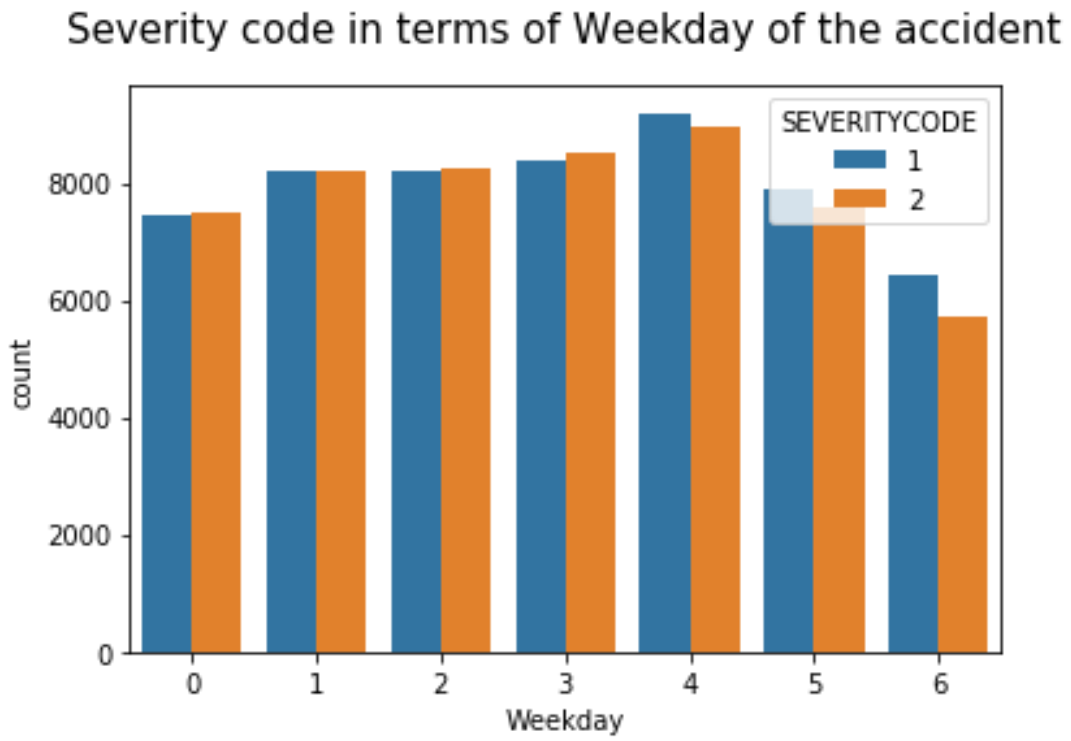
There are black spots where accidents are more frequent



Accidents at night are less frequent



Accidents in the weekend are less frequent



Classification models

	KNN	Logistic Regression	Decision Tree
Accuracy	0.68	0.66	0.69
Confusion matrix	[[11203, 5572][5095, 11270]]	[[11671, 5104][6308, 10057]]	[[10477, 6298] [4451, 11914]]

- Performance of the three models is similar

-Feature importance shows that *LOCATION_FREQ_S1 and LOCATION_FREQ_S2* were the best predictors by far

Conclusion and discussion

- Information about the location is the most valuable in order to predict severity
- Accidents in intersections tend to be more severe
- More features which describe the location of the accident would be useful
- For instance: speed limit in each location