Analysis Road Accident Severity in Seattle

Data Science Professional Certificate capstone project by IBM/Coursera

Table of contents

1. Introduction: Business Problem

Approximately 1.35 million people die each year as a result of road traffic crashes. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury. It is one of the major causes of premature deaths. They are many factors that influence the severity of a road accident whenever it occurs, including weather conditions, speed of the vehicle, condition of the road, etc.

The seaport city of Seattle is the largest city in the state of Washington, as well as the largest in the Pacific Northwest. As of the latest census, there were 713,700 people living in Seattle. Seattle residents get around by car, trolley, streetcar, public bus, bicycle, on foot, and by rail. With such bustling streets, it's no surprise that Seattle sees car accidents every day.

In order to reduce the severity of a road accident or prevent it altogether, there is a need to understand how the different factors affect the severity of a road accident. By understanding the relations between these factors, and developing a model that can predict the severity of a road accident with high accuracy, road users can use these predictions to adjust their driving or change their travels if possible, so that they can reduce the severity of the accidents or avoid it. This would greatly reduce the severity of accidents and some accidents will be avoided.

The audience of this project are the policy makers of Seattle and road users in general. By having these data, policy makers can make policies that reduces the number of road accidents and their fatality.

2. Data

The data that is used in this project is provided by SPD and recorded by Traffic Records. The data is from the year 2004. The data set provided for this work allows the analysis of a record of 200,000 accidents in the state of Seattle. The target of the data is the severity of the accident. The data has different features that relate to the severity of the data like speed, number of vehicles involved, number of passengers, etc.
A data exploratory will be done on the data to understand the data and identify relationships. These relationships will be used in the creation of a machine learning model that will predict the severity of data using supervised classification algorithms.
This data will be split into training and testing data to be used to train the model and test for out of case accuracy.

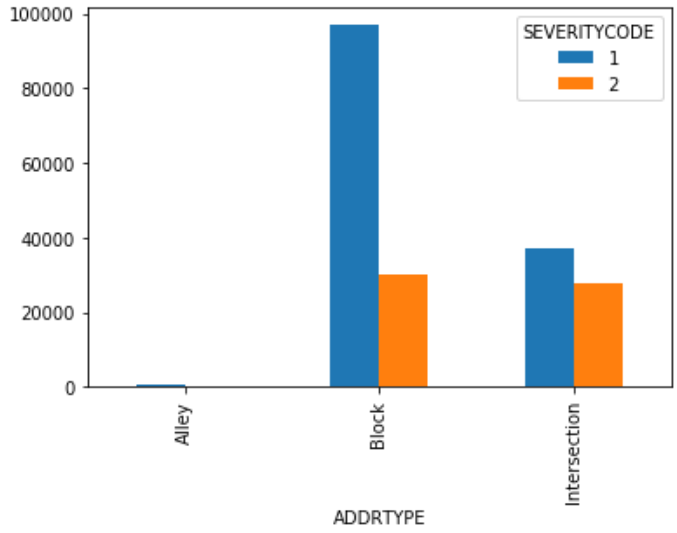
3. Methodology

3.1 Exploratory Data Analysis¶

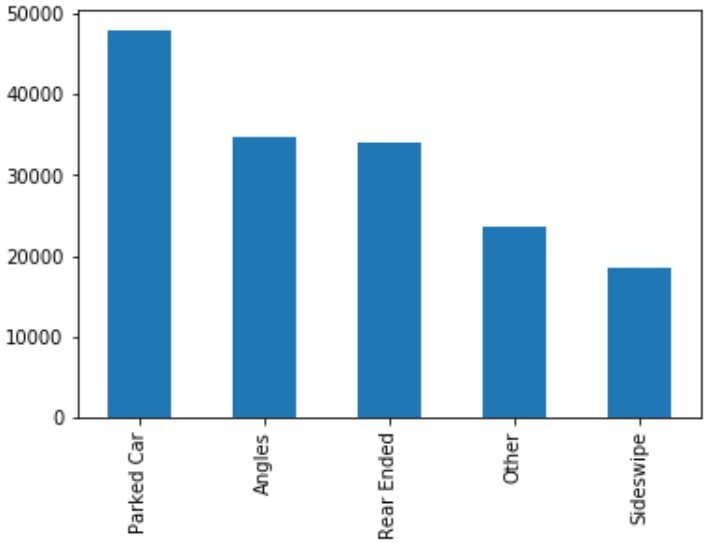
In order to gain a deeper understanding of the data I did an exploratory data analysis.

3.1.2 Variables and their Relationships¶

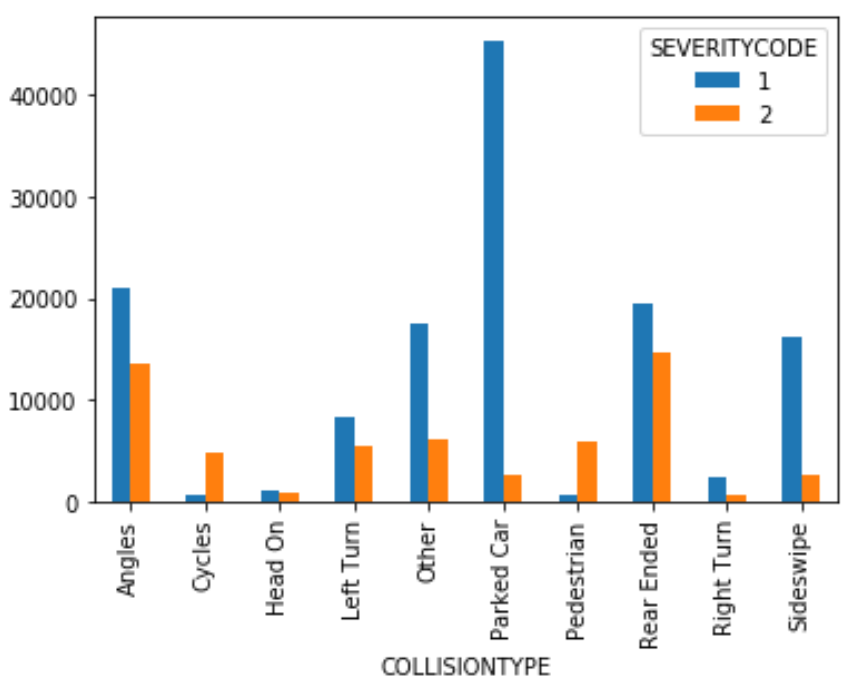
a) Relationship between ADDRTYPE and Severity of accidents¶



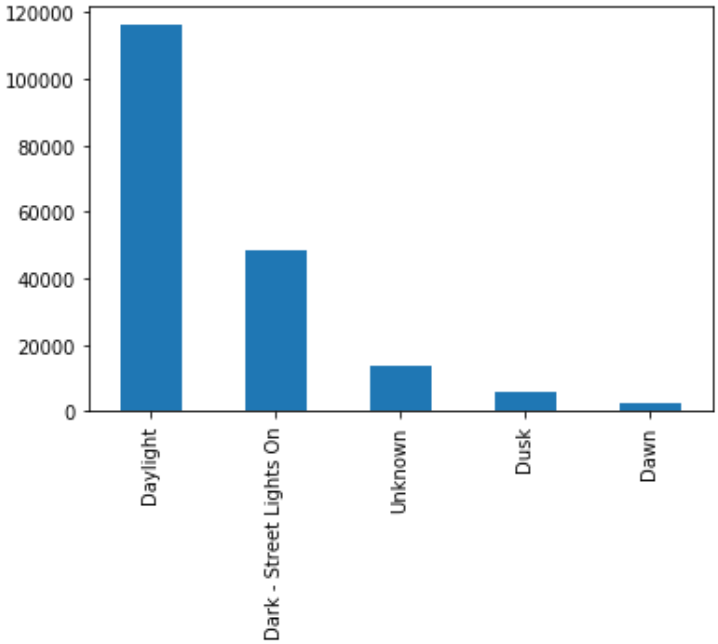
b) COLLISIONTYPE¶



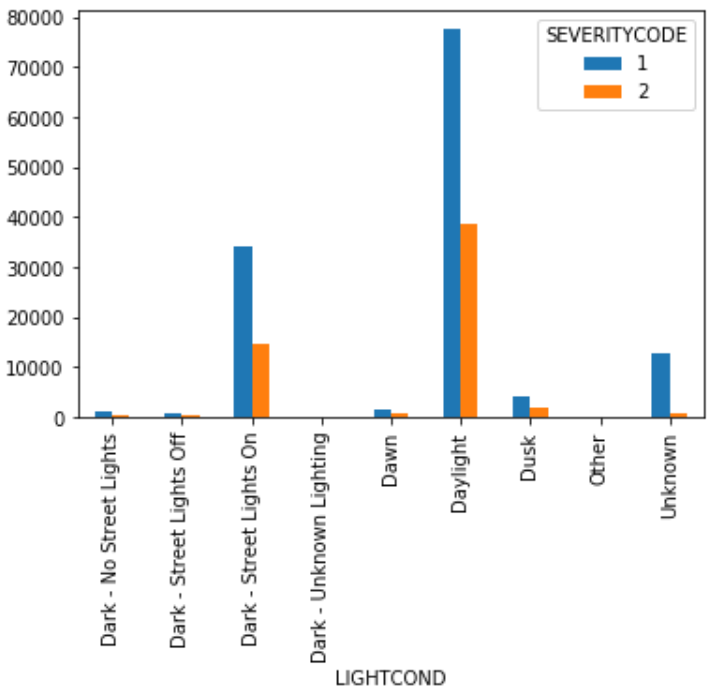
Relationship¶



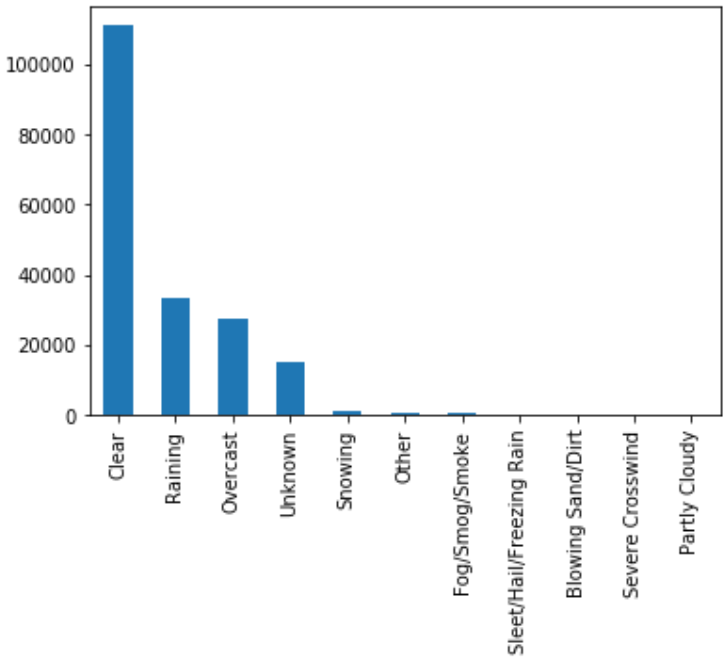
c) Lighting conditions



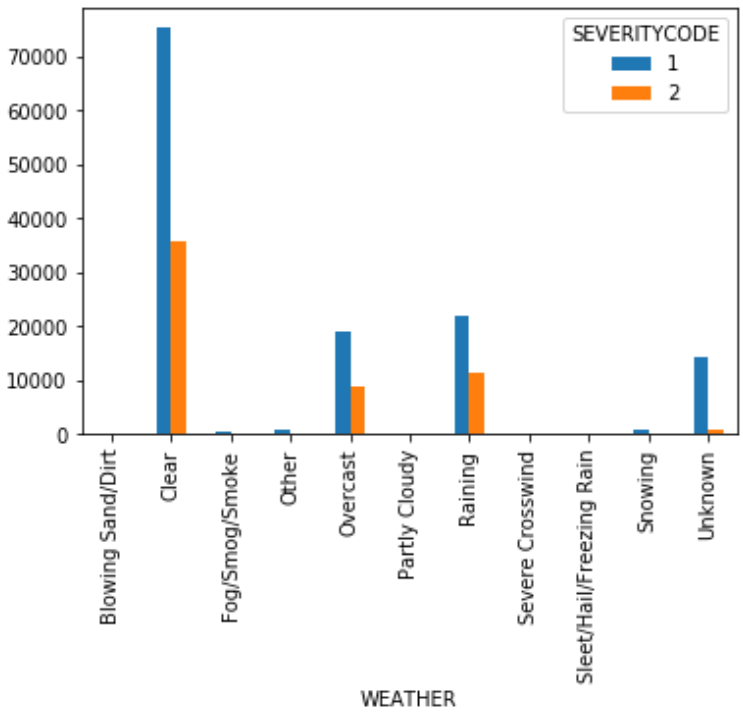
Relationship¶



d) Weather Conditions



Relationship



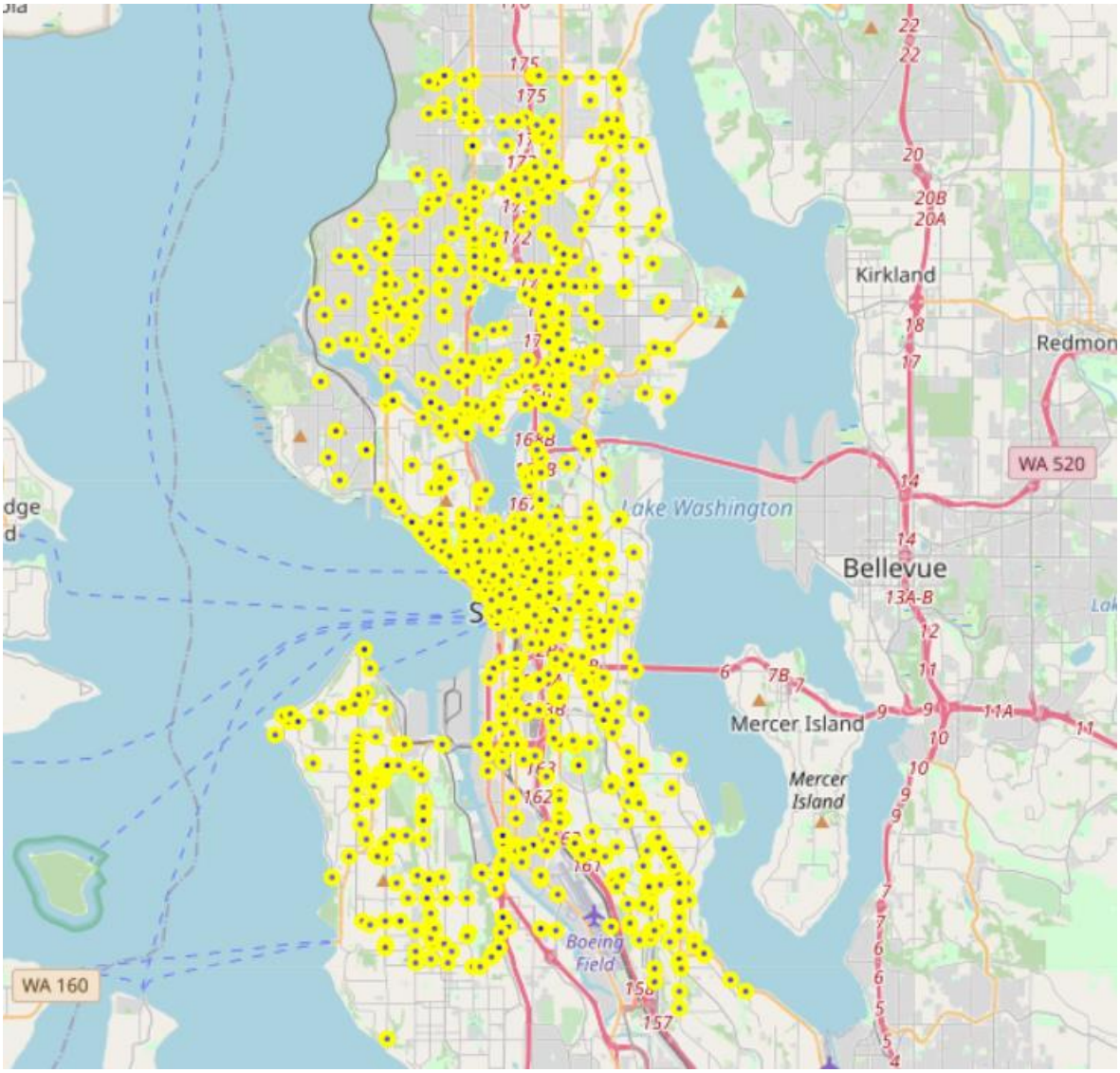
e) Other Exploratory Data Analysis¶

Accident as a Result of Pedestrian Right of Way not Granted

		PEDROWNOTGRNT
SEVERITYCODE	**JUNCTIONTYPE**	
1	**At Intersection (but not related to intersection)**	3
	At Intersection (intersection related)	353
	Driveway Junction	38
	Mid-Block (but intersection related)	9
	Mid-Block (not related to intersection)	53
	Ramp Junction	0
	Unknown	0
2	**At Intersection (but not related to intersection)**	14
	At Intersection (intersection related)	3385
	Driveway Junction	324
	Mid-Block (but intersection related)	74
	Mid-Block (not related to intersection)	400
	Ramp Junction	0
	Unknown	0

Visualization of the Accident Locations

I used Folium library to visualize the accidents in a map, as shown below:



3.2 Data Cleaning

After getting a deeper understanding of the data, the next step was to clean up the data. This step involved dropping of some columns that were not necessary, including: 'REPORTNO','OBJECTID', 'INCDATE', 'INCDTTM', 'COLDETKEY', 'INCKEY', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC', 'SDOT_COLCODE', 'SDOTCOLNUM', 'STATUS', 'SEVERITYCODE.1', 'ST_COLCODE', 'ST_COLDESC',, 'HITPARKEDCAR', 'CROSSWALKKEY'.

I also needed to convert some string columns into integer categorical variables, including: 'SPEEDING','PEDROWNOTGRNT', 'INATTENTIONIND', 'UNDERINFL'.

3.3 Predictive Modelling

I build two predictive machine learning learning classification models. These models uses supervised learning.
The classification models used are:
Support Vector Machine (SVM) and Decision Tree
The models performance was evaluated using Jaccard and f1-score as shown below.

Model Performance Comparison

Algorithm	Jaccard	F1-score
Decision Tree	0.72	0.74
SVM	0.80	0.76

4. Results and Discussion ¶

Our analysis has shown that most accidents occur within blocks, during the daytime. The most common type of collision from the data is hitting a parked car. These accidents mostly happen in dry road conditions and therefore we cannot blame rain.
The locations with the most accidents are: BATTERY ST TUNNEL and NORTHGATE WAY BETWEEN MERIDIAN AVE N AND CORLISS AVE. Some of the possible explanations for the accidents happening in blocks is that the drivers are distracted. There is not data to show the types of distractions.
From the folium map, we can see that accidents are distributed throughout Seattle, though not evenly.

5. Conclusion

In conclusion, we have established that most accidents occur in the blocks and mostly involves drivers getting distracted and hitting parked cars. Some of the policies that can me made to mitigate this is by adding guard rails to the roads along the blocks, and creating parking lots that are not too close to the road.
We have also created two machine learning prediction models that can predict the fatality of an accident with an average certainity.