



# **IBM MACHIN LEARNING CAPSTONE PROJECT**

## **A CASE STUDY TO PREDICT THE SEVERITY OF AN ACCIDENT**

Atena Vahedian

# Introduction

- A public data set provided by Coursera is used to demonstrate Data Analysis and Machine Learning skills obtained in the IBM Data Science Program
- The collision data set provided by Coursera includes all collision reports in Seattle since 2004.
- The data set has 194673 records and 37 attributes, and each record is labelled by the accident's severity level.
- The objective of this project is to: understand the data, define relevant attributes that cause the accident and build a predictive model.
- This analysis will help in potential driving hazard identification and can help drivers choose the best route based on weather and road condition.

# Data Acquisition and Cleaning

- Before the modeling process, it is important to investigate the data.
- 10 features were finalized for modeling purposes.

	SEVERITYCODE	ADDRTYPE	COLLISIONTYPE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	WEATHER	ROADCOND	LIGHTCOND	SPEEDING
0	2	Intersection	Angles	2	0	0	2	Overcast	Wet	Daylight	NaN
1	1	Block	Sideswipe	2	0	0	2	Raining	Wet	Dark - Street Lights On	NaN
2	1	Block	Parked Car	4	0	0	3	Overcast	Dry	Daylight	NaN
3	1	Block	Other	3	0	0	3	Clear	Dry	Daylight	NaN
4	2	Intersection	Angles	2	0	0	2	Raining	Wet	Daylight	NaN

# Statistical description of the data

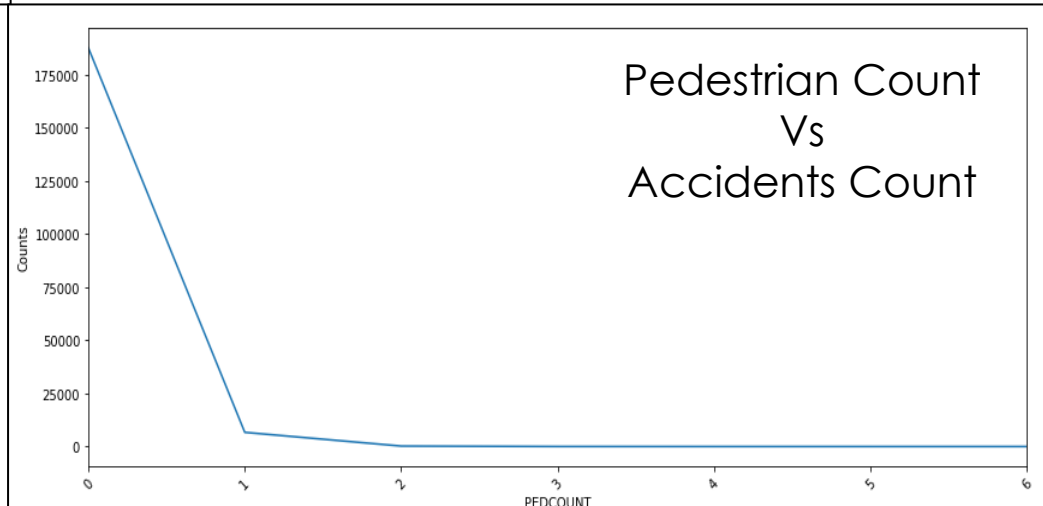
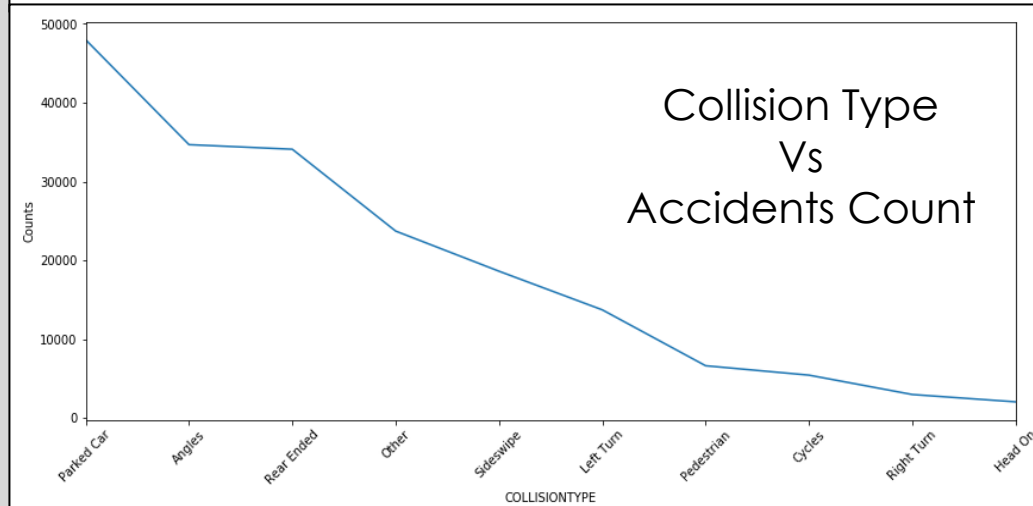
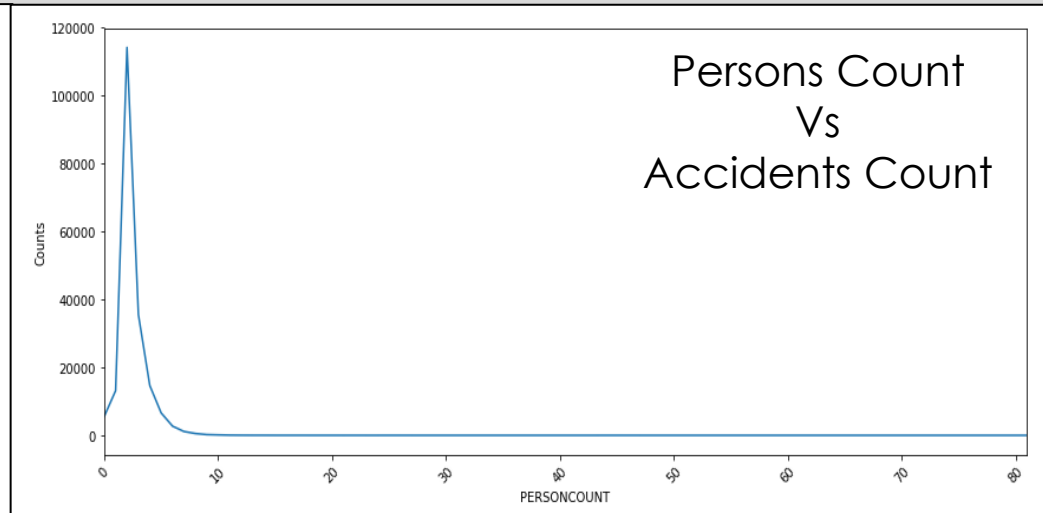
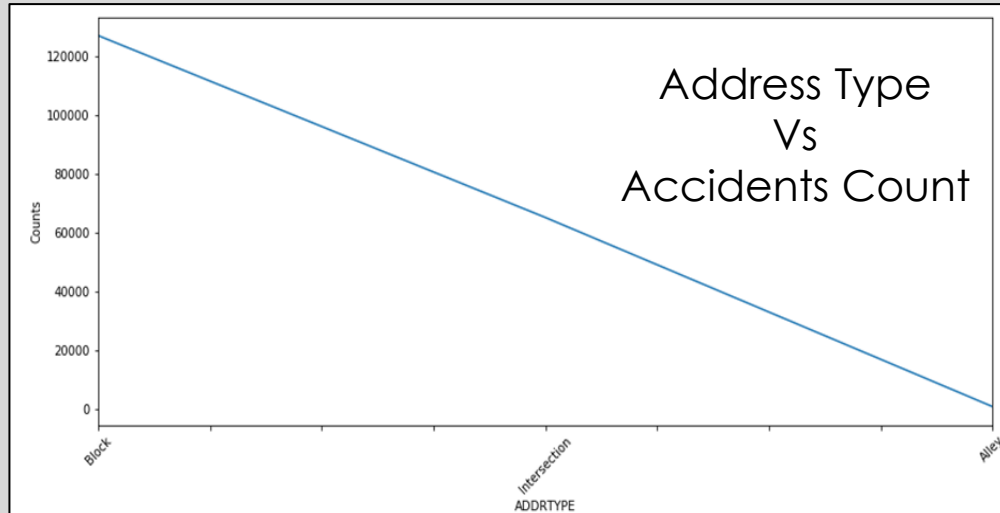
	SEVERITYCODE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT
count	194673.000000	194673.000000	194673.000000	194673.000000	194673.000000
mean	1.298901	2.444427	0.037139	0.028391	1.920780
std	0.457778	1.345929	0.198150	0.167413	0.631047
min	1.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	2.000000	0.000000	0.000000	2.000000
50%	1.000000	2.000000	0.000000	0.000000	2.000000
75%	2.000000	3.000000	0.000000	0.000000	2.000000
max	2.000000	81.000000	6.000000	2.000000	12.000000

	ADDRTYPE	COLLISIONTYPE	WEATHER	ROADCOND	LIGHTCOND	SPEEDING
count	192747	189769	189592	189661	189503	9333
unique	3	10	11	9	9	1
top	Block	Parked Car	Clear	Dry	Daylight	Y
freq	126926	47987	111135	124510	116137	9333

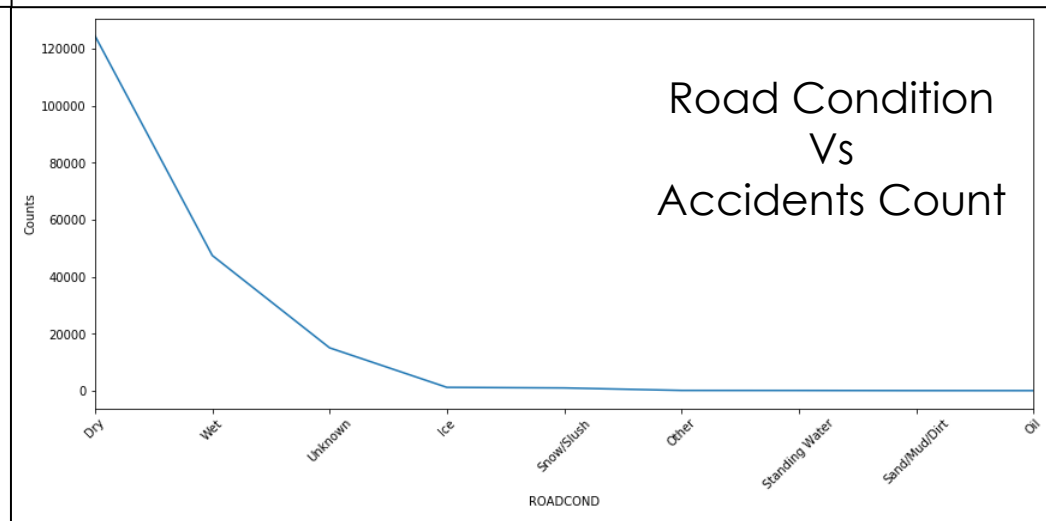
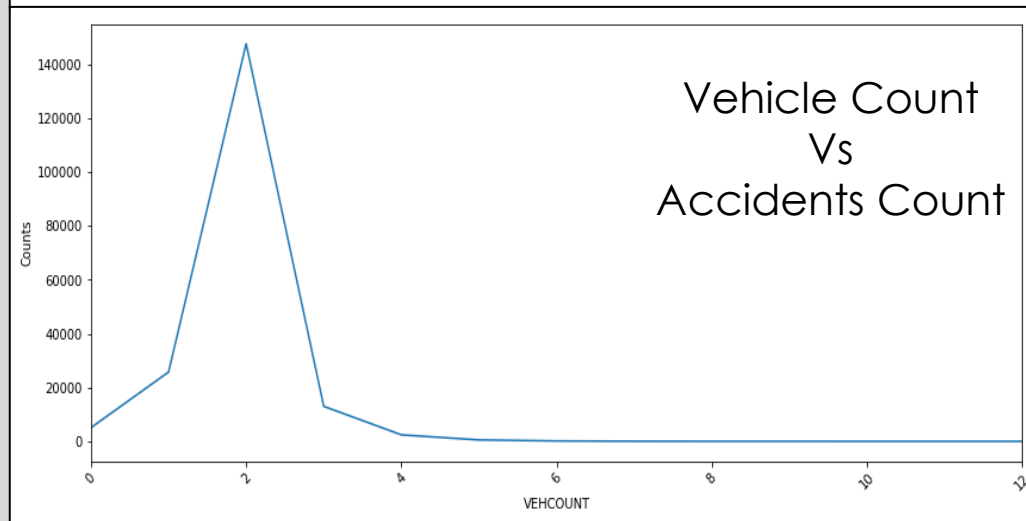
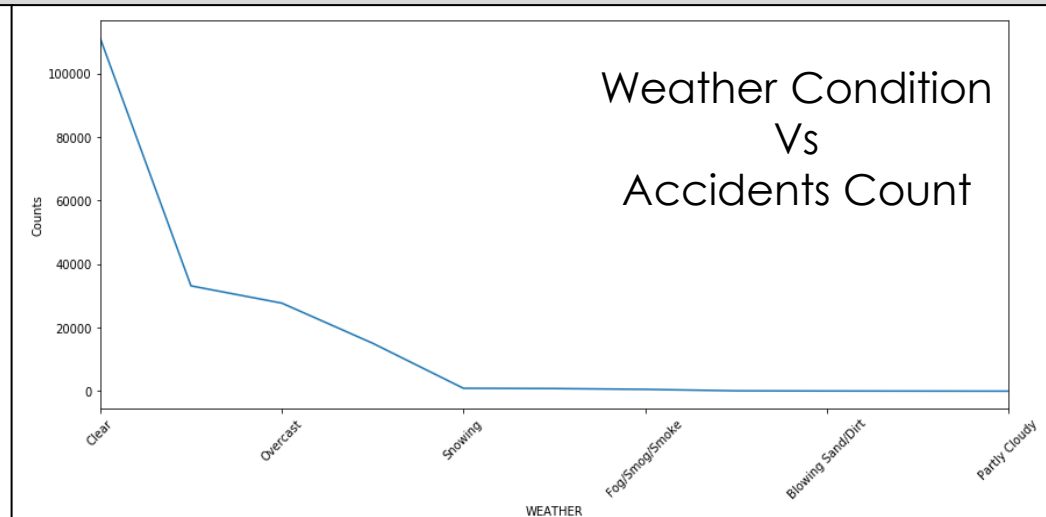
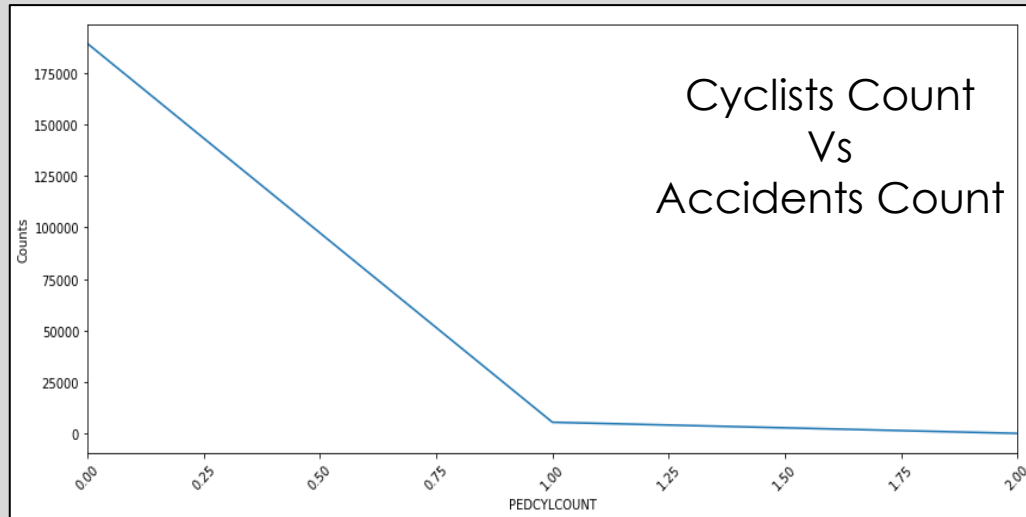
# Distribution of accidents vs attributes

SEVERITYCODE				SEVERITYCODE				SEVERITYCODE				SEVERITYCODE			
ADDRTYPE		SEVERITYCODE		LIGHTCOND		SEVERITYCODE		ROADCOND		SEVERITYCODE		WEATHER		SEVERITYCODE	
Alley	1	0.890812		Dark - No Street Lights	1	0.782694		Dry	1	0.678227		Blowing Sand/Dirt	1	0.732143	
	2	0.109188			2	0.217306			2	0.321773			2	0.267857	
Block	1	0.762885		Dark - Street Lights Off	1	0.736447		Ice	1	0.774194		Clear	1	0.677509	
	2	0.237115			2	0.263553			2	0.225806			2	0.322491	
Intersection	1	0.572476		Dark - Street Lights On	1	0.701589		Oil	1	0.625000		Fog/Smog/Smoke	1	0.671353	
	2	0.427524			2	0.298411			2	0.375000			2	0.328847	
				Dark - Unknown Lighting	1	0.636364		Other	1	0.674242		Other	1	0.880577	
					2	0.363636			2	0.325758			2	0.139423	
				Dawn	1	0.670663		Sand/Mud/Dirt	1	0.693333		Overcast	1	0.684456	
					2	0.329337			2	0.306667			2	0.315544	
				Daylight	1	0.668116		Snow/Slush	1	0.833665		Partly Cloudy	1	0.400000	
					2	0.331884			2	0.166335			2	0.337185	
				Dusk	1	0.670620		Standing Water	1	0.739130		Raining	1	0.662815	
					2	0.329380			2	0.260870			2	0.337185	
				Other	1	0.778723		Unknown	1	0.950325		Severe Crosswind	1	0.720000	
					2	0.221277			2	0.049675			2	0.280000	
				Unknown	1	0.955095		Wet	1	0.668134		Sleet/Hail/Freezing Rain	1	0.752212	
					2	0.044905			2	0.331866			2	0.247788	
												Snowing	1	0.811466	
													2	0.188534	
												Unknown	1	0.945928	
													2	0.054072	
SPEEDING		SEVERITYCODE													
N	1	0.705099													
	2	0.294901													
Y	1	0.621665													
	2	0.378335													

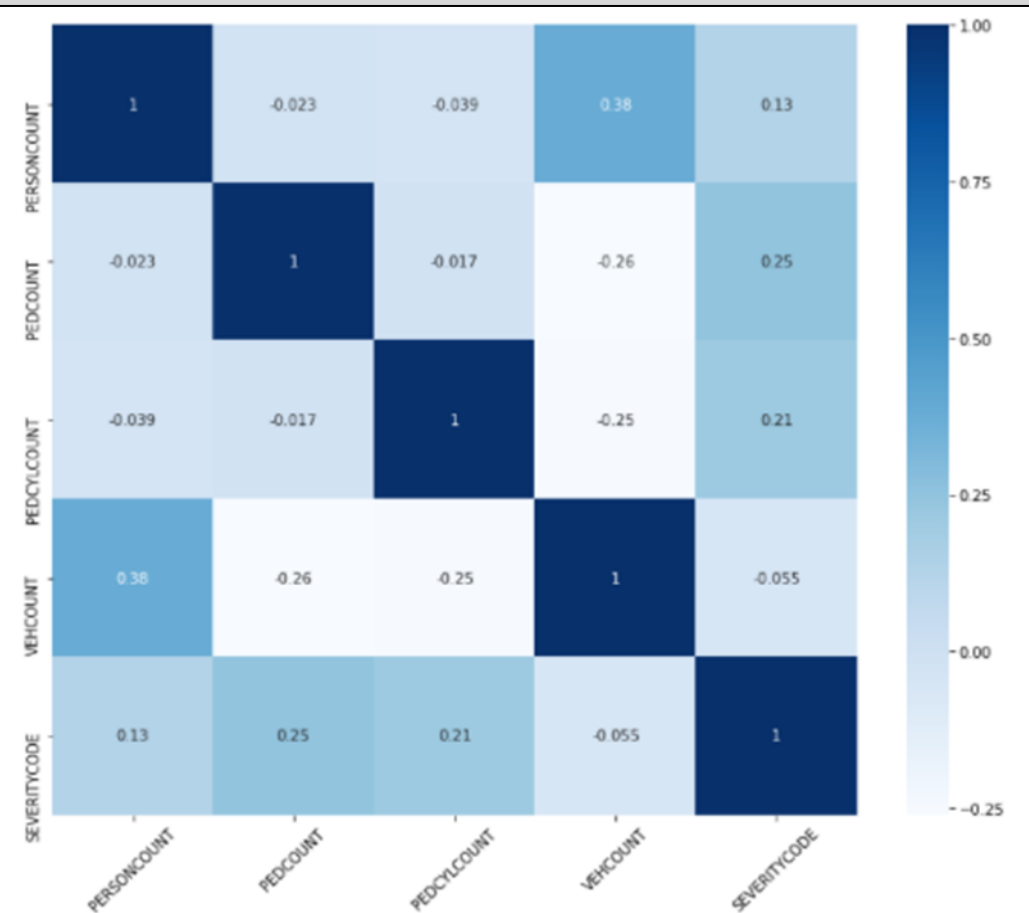
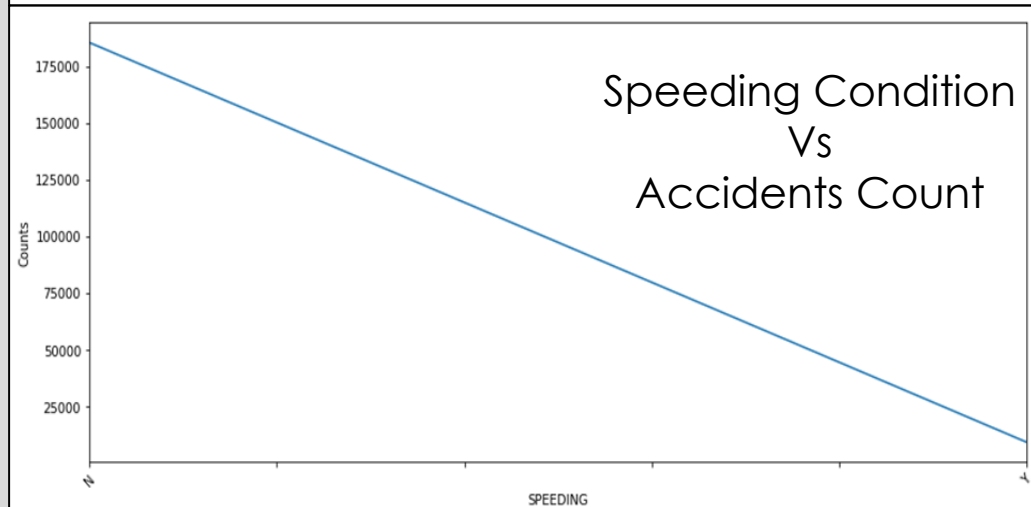
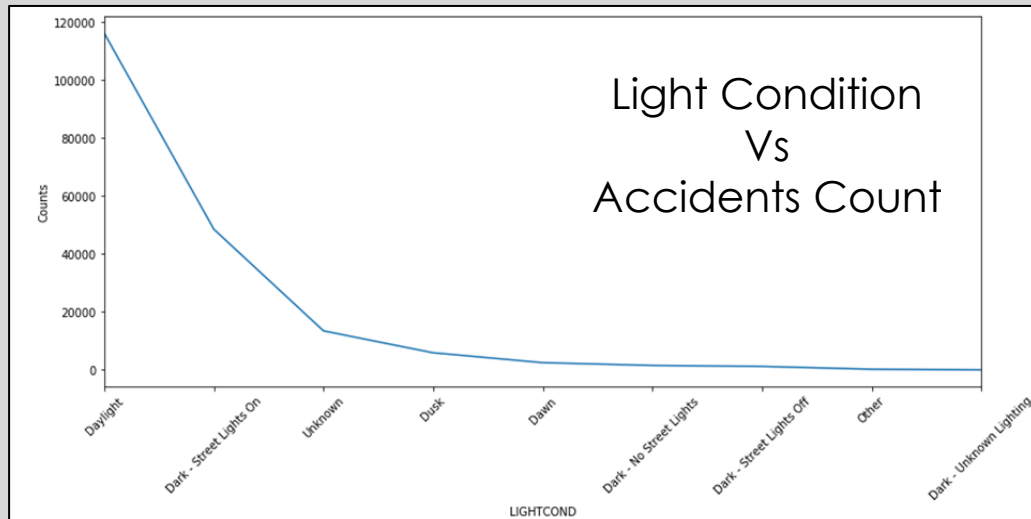
# Number of accidents vs attributes (cont'd)



# Number of accidents vs attributes (cont'd)



# Number of accidents vs attributes

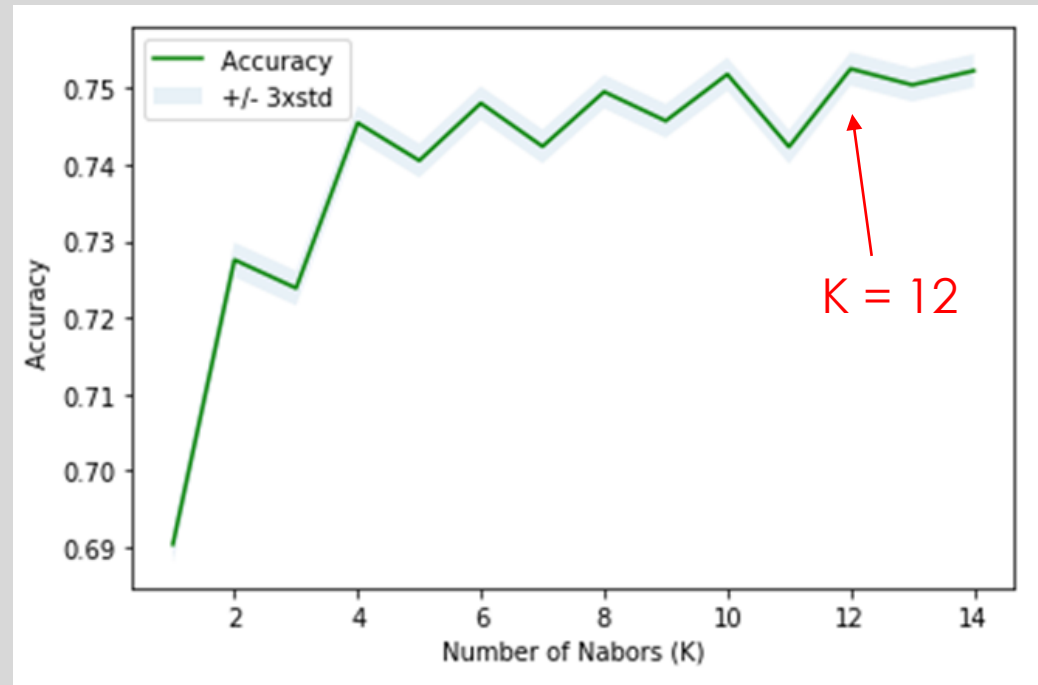


Heatmap showing the correlation between numerical values in the dataframe.



# Modeling

- Four different model would be applied on the data set and will be evaluated using various metrics to find the best option:
  1. K Nearest Neighbor (KNN)
  2. Decision Tree
  3. Support Vector Machine
  4. Logistic Regression
- The train\_test\_split parameters:  
X, y, test\_size=0.2, and random\_state=4
- K value for KNN algorithm is 12.



# Model Evaluation

Algorithm	Jaccard	F1-Score	Log Loss
KNN	0.75	0.72	NA
Decision Tree	0.75	0.69	NA
SVM	0.76	0.71	NA
Logistic Regression	0.76	0.72	0.48

# Results

- Most accidents are labeled as category “1” or less severe accidents.
- Most accidents happen at the block type address, however, they are less severe than the accidents happening at intersections.
- Most accidents reported, did not involve speed violation.
- Number of pedestrian, cyclist, and persons involved in the accident have positive relationship with the severity level of the accident and all accidents have less than 10 people involved and mostly no pedestrian.
- Most collisions are labeled as parked car type and lastly, most accidents happen at clear weather and dry road condition during daylight hours.
- All four models predicted the severity code for the test set with relatively high accuracy score.
- KNN and SVM algorithms required longer computation time compared to the other two methods.

# Discussion

- Data set was split into training and testing sets (80%, 20%) to compare the predicted data labels (severity code) with the recorded values.
- All models have similar Jaccard accuracy values, but Decision Tree and Logistic Regression have comparatively less computation times and would be the optimum models for future predictions.
- It is recommended to collect additional data with less missing information as this could help improve the model accuracy.

# Conclusion

- There is no particular relationship between bad weather, light and road conditions that affect collisions.
- There were a lot more collisions that happened on dry roads and clear weather conditions during day time compared to when conditions were not ideal.
- Drivers tend to be more careful in driving in adverse weather, road and light conditions.
- Author suggests using this model to predict a possibility of severe car accidents based on the time and location of driving.
- This information could help drivers choose the safest route for their travel and reduce the number of dangerous car accidents.

Thank  
you

