1. Business Problem

1.1. Background

Every year, car accidents cause more injuries and deaths than any other type of personal injury. There are many factors that go into whether a person gets into a car accident that the severity of the event. With the progress of technology included in cars and their new capabilities, it would be important to have the tools and means available to provide drivers with a warning, given the weather and road conditions about the possibility of getting into a car accident and how severe it would be. Therefore, the driver would drive more carefully or even change his/her driving route if he/she is able to do so.

1.2. Problem

Data that might contribute to determining the severity of a car accident may include weather, road conditions, lighting conditions, and other metrics that would help describe the surrounding environment of the driver. This project aims to predict how severe an accident would be based on these data

1.3. Interest

Obviously, it would be of great interest for drivers to have to avoid an accident from knowing how severe it may be by driving more carefully. This will also enable transport, security, and emergency agencies all around the world to have different predictive models for quickly analyzing crashes when they happen and dispatch an appropriate response swiftly.

2. Data

2.1 Data Sources

The data utilized for this report came from collision data provided by SPD and recorded by Traffic Records. This dataset includes all types of collisions from 2004 to present, where collisions will display at the intersection or mid-block of a segment.

2.2 Data Cleaning

The data was downloaded from the course organizers at IBM. We first filled in missing values for the columns of interest. PEDROWNOTGRNT, SPEEDING, INATTENTIONIND, and UNDERINFL had only Y entered and so it was assumed all other values were N. These values were then converted to quantitative values. UNDERINFL has both quantitative and qualitative values on top of missing

values. We assumed the missing values to be N and then converted the qualitative values to quantitative ones.

We also had variables that had ambiguous values, particularly “Other” and “Unknown” for WEATHER, ROADCOND, LIGHTCOND. These values were converted to NaN and then rows containing these values were dropped from our dataset.

2.3 Feature Selection

The columns that were decided to be kept from the original dataset to predict our severity score, SEVERITYCODE, is as follows:

- SEVERITYCODE: A code that corresponds to the severity of the damage
    - 0 – Unknown
    - 1 – Property Damage
- ADDRTYPE: Collision address type
- PERSONCOUNT: The total number of people involved in the collision
- PEDCOUNT: The number of pedestrians involved in the collision
- PEDCYCLCOUNT: The number of bicycles involved in the collision
- VEHCOUNT: The number of vehicles involved in the collision
- SDOT_COLCODE: A code given to the collision by SDOT
- INATTENTIONIND: Whether or not the collision was due to inattention
- UNDERINFL: Whether or not the driver was under the influence of drugs or alcohol
- WEATHER: A description of weather conditions during the time of the collision
- ROADCOND: The condition of the road during the collision
- LIGHTCOND: The light conditions during the collision
- PEDROWNOTGRNT: Whether or not the pedestrian right of way was granted
- SPEEDING: Whether or not speeding was a factor in the collision
- ST_COLCODE: A code provided by the state that describes the collision

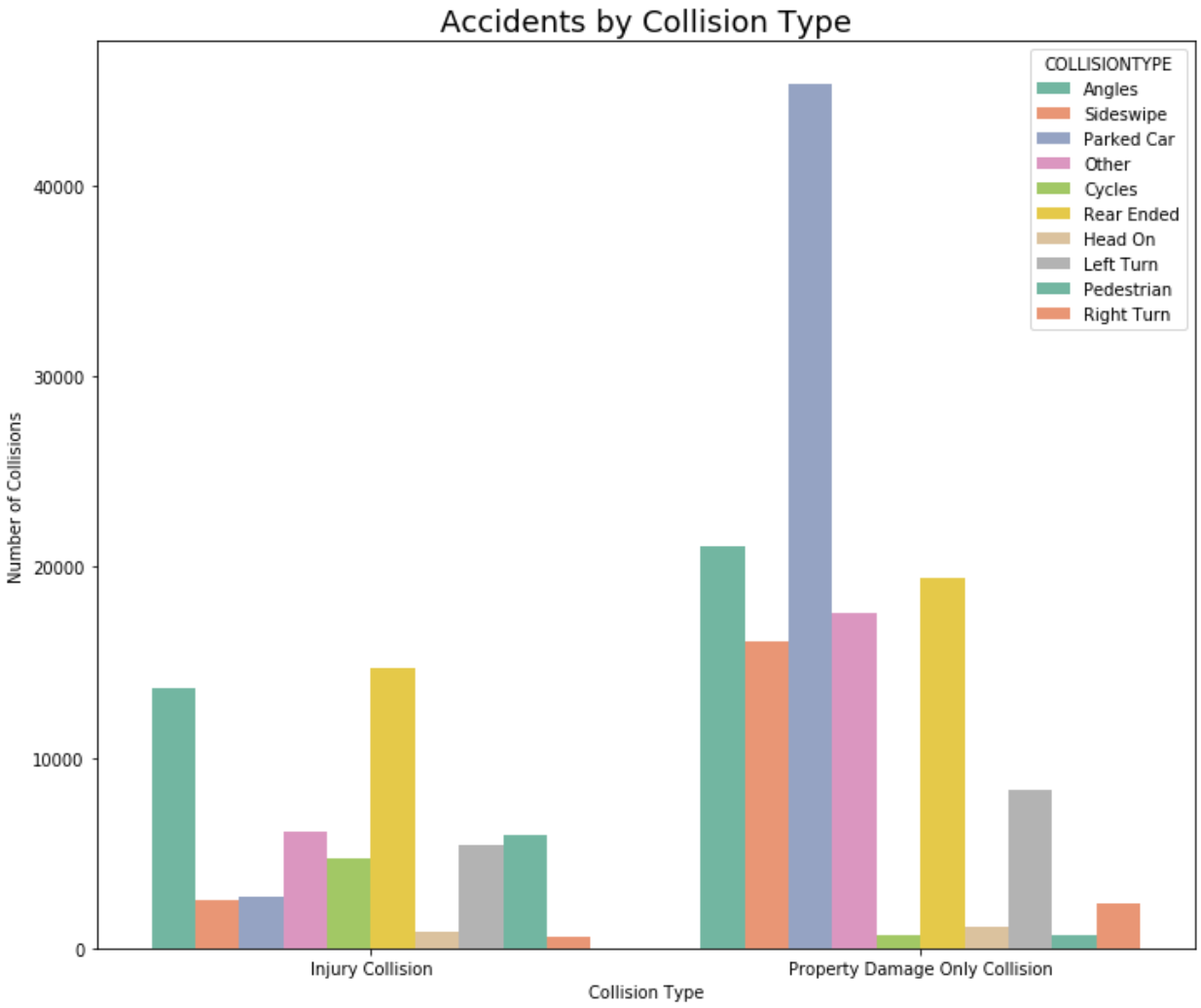
The other columns in the dataset are either redundant to the features chosen above or do not have any significance in determining accident severity.

3. Methodology

3.1 Exploratory Data Analysis

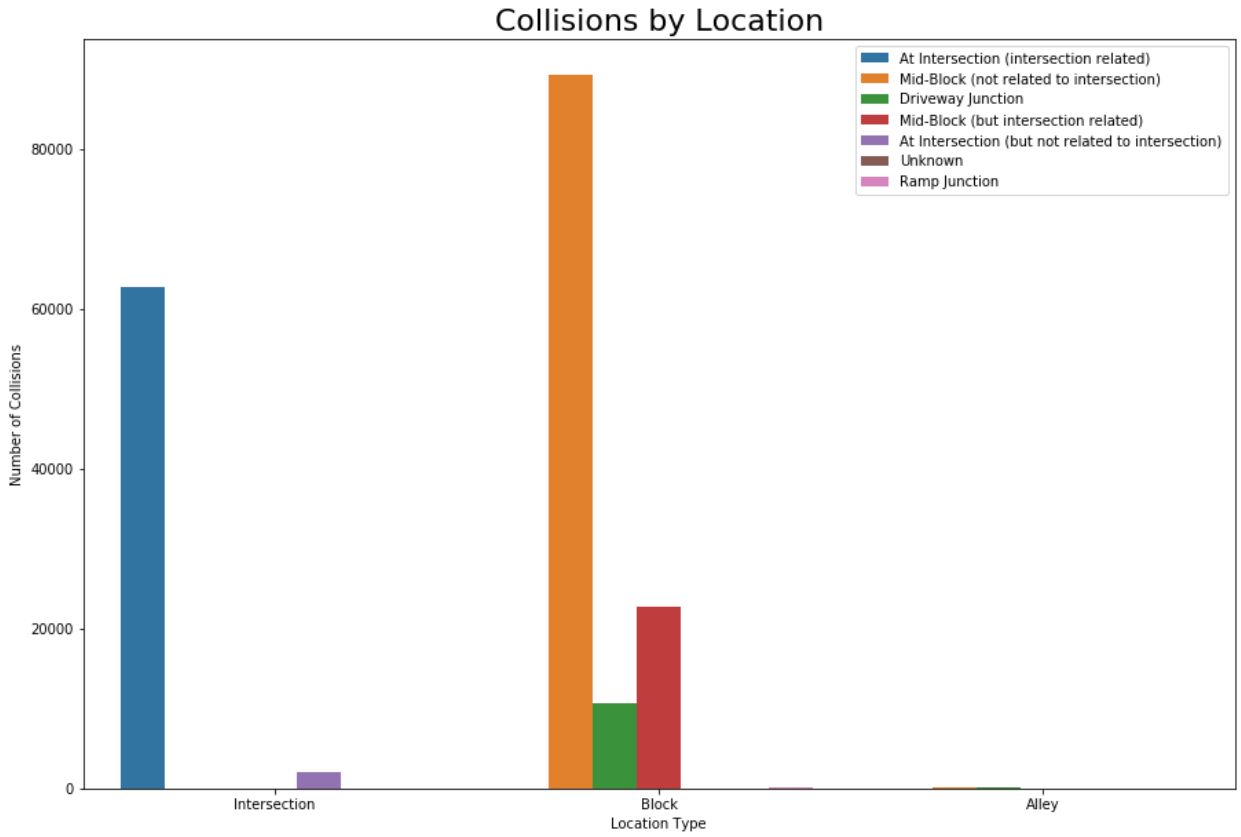
Collision Type and Number of Collisions

For injury collisions, we see that most car accidents occur due to cars getting hit on the rear end or at an angle. For property damage, most collisions happen from hitting a parked car.



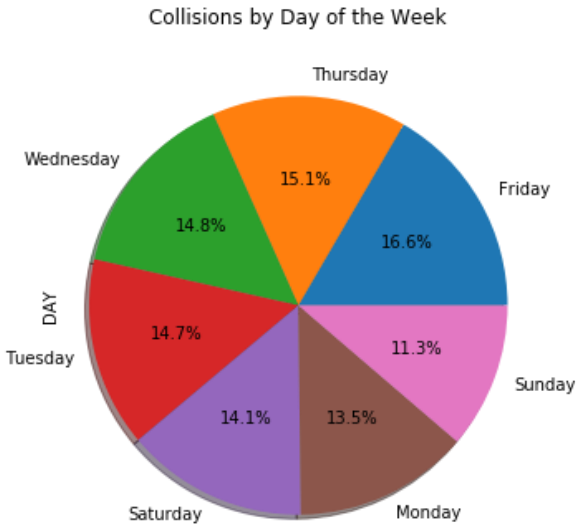
Location Type and Number of Collisions

Most block-related collisions seem to happen mid-block, away from the intersection. Also, there seems to be very little collisions occurring in an alley.



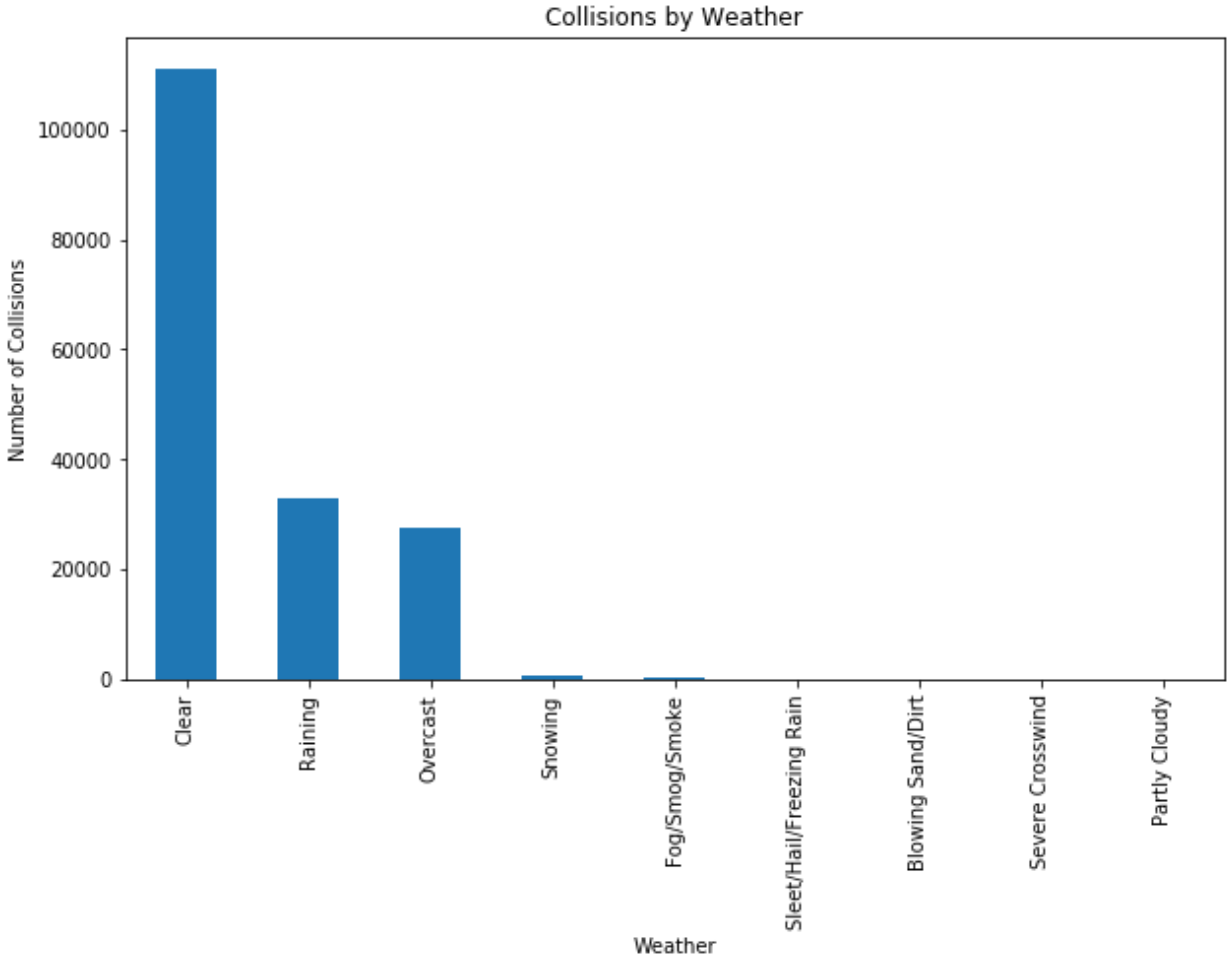
Days of the Week and Number of Collisions

Although the percentage of getting into an accident is roughly the same across days, it seems as though most accidents occur on a weekday as opposed to the weekend.



Weather and Number of Collisions

Based on our visualization analysis, it seems that most collisions occurred on a clear day. Therefore, it does not seem as though weather plays an important factor when predicting collisions and collision severity.



3.2 Inferential Statistical Testing

Upon correlation analysis between collision severity and all of our variables of interest, it was found again that weather did not seem to have any strong correlation. Similarly, location of the accident did not seem to play a role in its severity as well.

	SEVERITYCODE	Intersection	Alley	Block
SEVERITYCODE	1.000	0.199	-0.026	-0.185
Intersection	0.199	1.000	-0.044	-0.970
Alley	-0.026	-0.044	1.000	-0.085
Block	-0.185	-0.970	-0.085	1.000

When analyzing whether the people involved in an accident played a role in accident severity, it was found that the most severe injuries involved pedestrians and cyclists.

	SEVERITYCODE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT
SEVERITYCODE	1.000	0.131	0.246	0.214	-0.055
PERSONCOUNT	0.131	1.000	-0.023	-0.039	0.381
PEDCOUNT	0.246	-0.023	1.000	-0.017	-0.261
PEDCYLCOUNT	0.214	-0.039	-0.017	1.000	-0.254
VEHCOUNT	-0.055	0.381	-0.261	-0.254	1.000

3.3 Selection of Machine Learning Model

In this project we will direct our efforts in predicting **accident severity**.

In the first step we have collected the required data, cleaned the data, chose variables that were related to the problem at hand, and converted all of our qualitative variables into quantitative ones

The second step involved exploratory analysis where we examined various factors that played a role in collisions from a data visualization perspective. Moreover, we examined the correlation between severity code and all of our variables.

In the next step, we will utilize **classification** as our machine learning algorithm method since we are categorizing unknown items into a discrete set of categories. In this case, we need to determine which severity code, 1 or 2, a case falls into given the weather and road conditions.

Note: K-Nearest Neighbors was attempted but the system kept crashing when trying to determine the best value for k. Support Vector Machine was not utilized because it does not work well with very large datasets.

4. Analysis

We first had to normalize our data since the values of our dataset were not in the same range. We then split our data set into a train set and test set, with the test size being 20% and random state being 3.

The machine learning models we used for classification were **Decision Tree** and **Logistic Regression.** The results of our model evaluation are as follows:

Evaluation	Decision Tree	Logistic Regression
Jaccard Index	0.73	0.73
Precision	0.71	0.72
Recall	0.99	0.97
F1 Score	0.66	0.68

5. Discussion

Based on our analysis of the data, we have some key observations:

1. Most accidents occur during weekdays at intersections
2. Weather conditions do not play a significant role in accidents
3. Road and lighting conditions have a weak correlation with accidents
4. Between blocks, maximum accidents occur at mid-blocks
5. In collision accidents, maximum damage is done to parked cars

In this project we have identified the relation between accidents and accident severity with several human, environmental and location attributes. Maximum accidents occur at intersections related to pedestrians or cyclists. We analyzed different machine learning classification methods to classify accidents as injury or collision accidents. The Logistic Regression model offered maximum accuracy. It correctly predicted 72% as injury collisions. This data could be used by enable transport, security, and emergency agencies to better understand car accidents and take preventative measures to prevent them. It is also aimed at us, whether we are pedestrians, cyclists or vehicle owners to be more careful at intersections to prevent an accident.

6. Conclusion

We were able to achieve an accuracy of 72% using the Logistic Regression Classifier. There are a lot of variances which have not been accounted for. However, using this project we could really narrow down to the location where maximum accidents occur and the most affected. We also understood that there is very less importance of human and weather factors in causing an accident. The prediction could be improved by capturing real-time data during traffic incidents.