# IBM Capstone Project – Allegheny County / City of Pittsburgh 2017 Crash Data

## Introduction

Allegheny County in the City of Pittsburgh government official wants to analyze the data collected for car accidents. They want to know by using data science methodologies and algorithms of the circumstances recorded, whether there is a way to predict and decrease the frequency of accidents. The decrease of accidents can help city officials better plan their personnel hours and resources where needed most. The goal is to plan resources more efficiently and effectively based on the predicted severity type of an accident. If an accident severity is level 1 (killed) or 2 (major injury), then more resources should be directed for accident assistance or to educate the drivers.

## Data

The historical data of the crash data in 2017 contains locations and information about every crash incident reported to the police in Allegheny Country in 2017. This will be used to train and test the developed model.

The link of the data can be found in: https://data.wprdc.org/datastore/dump/bf8b3c7e-8d60-40df-9134-21606a451c1a

This data will be used as an input to build a machine learning mechanism to predict the severity of the types of the accident based on key factors that might caused the accident or contribute to the accidents. The key factors can include “time of the day” “month” “day of the week,” “road conditions”, “illumination”, etc. The prediction is the “Max_Severity_Level”.

The results will be evaluated using the logistic regression methodology if possible.

There are 12537 rows with 190 columns in this data. Therefore, first we want to clean up the data.

There are different things to clean including:

1. There are columns that are not needed in this study
2. There are missing data or mislabeled data in cell

## Methodology

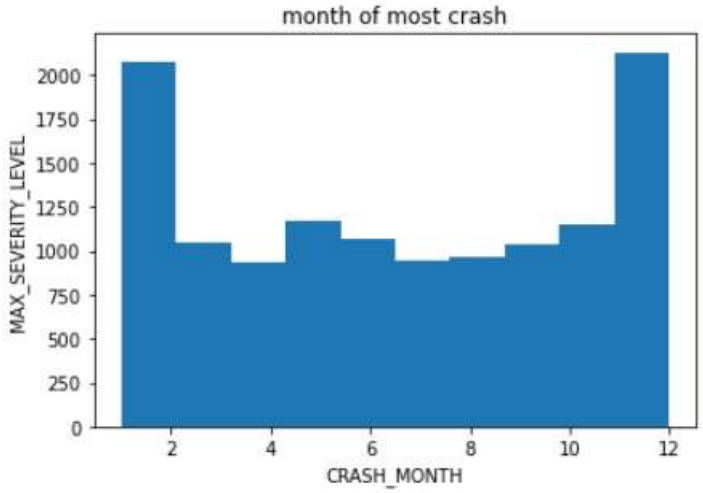
After cleaning up the data. First we get a broad view on the correlation between all the remaining data to see whether we can spot any obviously correlation between each other. Please find the table below.

| | CRASH_MONTH | DAY_OF_WEEK | TIME_OF_DAY | HOUR_OF_DAY | ILLUMINATION | WEATHER | ROAD_CONDITION | MAX_SEVERITY_LEVEL | SPEED_LIMIT |
|---|---|---|---|---|---|---|---|---|---|
| CRASH_MONTH | 1.000000 | 0.007969 | 0.010516 | 0.010500 | 0.013293 | -0.032476 | -0.056818 | 0.004807 | 0.028470 |
| DAY_OF_WEEK | 0.007969 | 1.000000 | 0.032011 | 0.032008 | -0.009154 | 0.034740 | 0.046551 | -0.032301 | 0.011434 |
| TIME_OF_DAY | 0.010516 | 0.032011 | 1.000000 | 0.999807 | -0.004916 | -0.006869 | -0.009614 | 0.020512 | -0.015587 |
| HOUR_OF_DAY | 0.010500 | 0.032008 | 0.999807 | 1.000000 | -0.004827 | -0.007216 | -0.009862 | 0.020512 | -0.015874 |
| ILLUMINATION | 0.013293 | -0.009154 | -0.004916 | -0.004827 | 1.000000 | 0.114472 | 0.094815 | 0.004508 | 0.012053 |
| WEATHER | -0.032476 | 0.034740 | -0.006869 | -0.007216 | 0.114472 | 1.000000 | 0.578450 | -0.018432 | 0.012105 |
| ROAD_CONDITION | -0.056818 | 0.046551 | -0.009614 | -0.009862 | 0.094815 | 0.578450 | 1.000000 | -0.036391 | 0.036046 |
| MAX_SEVERITY_LEVEL | 0.004807 | -0.032301 | 0.020512 | 0.020512 | 0.004508 | -0.018432 | -0.036391 | 1.000000 | -0.034709 |
| SPEED_LIMIT | 0.028470 | 0.011434 | -0.015587 | -0.015874 | 0.012053 | 0.012105 | 0.036046 | -0.034709 | 1.000000 |

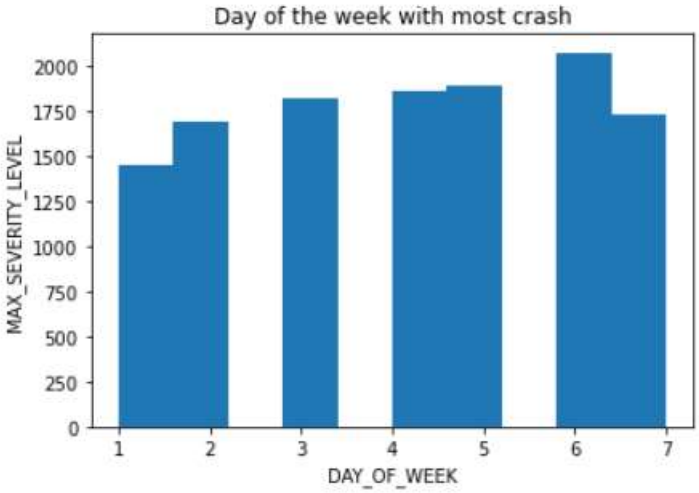
However, there are very few correlation between the different data. One is the time of the day verses the hour of the day. It is expected because they are truly related to one another. Second is the correlation between the weather and the road condition. This is probably due to the fact that a rainy day means it is going to be a wet road. A snowy day will mean either sleet road or snowy road, etc.

## Possible linear correlation based on when drivers are on the road

The historical data of the crash data in 2017 contains times, hour and month or days of the week for each and every crash. Therefore we plot them to get a look at whether there are obviously correlation or they can be explained by logic.

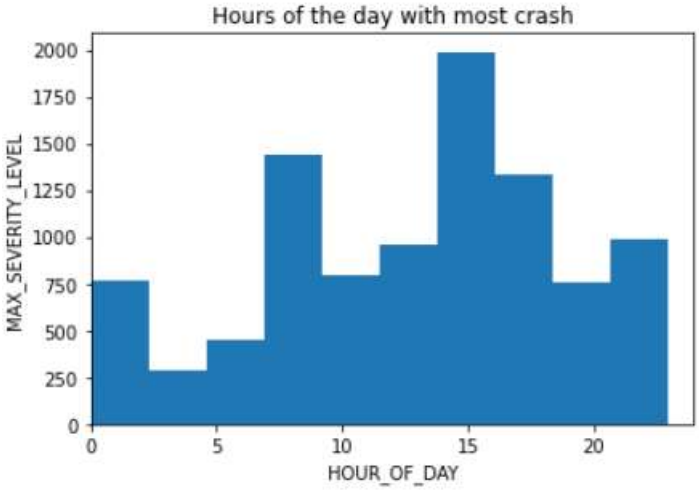



The crash data based on the month you can see that January and December the number are higher compare to rest of the year. This can be either based on those are the months that people are traveling and not at work, or the weather (snow and sleet road) can possibly contribute to higher accident rates.




However, looking at the days of the week, they are pretty evenly spread out. The crash numbers are not affected by whether it is work week or weekend.

Next, we look at the hours of the day when most accidents can happen. The graphic below showed the most accidents happen around rush hour after work. The accident numbers are low at nights are expected because of people are sleeping and less vehicles are on the road.




After looking at the time correlation to the accident rate, now we will build and compare using the following Machine Learning techniques:

- Logistic Regression
- K-Nearest Neighbor (KNN)
- Decision Tree

The data is also split between the Training set and the testing set. They are randomly choose on whether they will be training or testing set. During the Decision Tree prediction tool testing, we use 6837 as training set, and 2931 as testing set. During the KNN and logistic regression prediction, we use 5860 as training set and 3908 as testing set.

## Final Result

The result for the prediction tool are shown below. The Decision tree provided the best prediction result with highest accuracy.

| Algorithm | Jaccard | F1-score | Accuracy |
|---|---|---|---|
| Decision Tree | 0.4387 | 0.5300 | 0.6571 |
| KNN | 0.4311 | 0.5287 | 0.6407 |
| Logistic Regression | 0.4392 | 0.5283 | 0.6627 |