# Prediction and Analysis of Traffic Accident in Seattle, Washington

# Introduction

Dataset that we have is about collision mainly on transportation that record from by SPD and will display at the intersection or mid block a segment. So many collisions in this dataset cause that record from 2004 until Present

So, the goals of this project is reduce the accident on transportation by understanding about characteristic each data by severity condition. Finding the insight for each level on severity code, how we can give a guidance when we using means of transportation, how we can more protective when we driving and many more the insight would be suggest in this project.

**More info about this dataset.**

SDOT Traffic Management Division, Traffic Records Group | SDOT GIS Analyst | DOT_IT_GIS@seattle.gov

# HELLO!

**I am Anugrah Nurhamid**

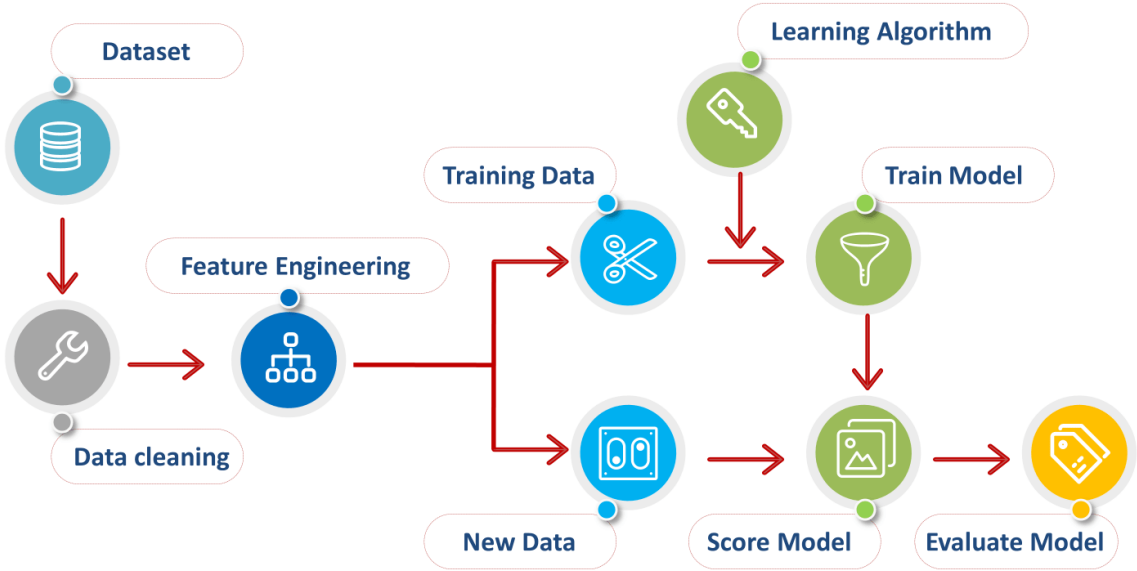
I am here for present final capstone project in IBM Data Science.

You can find me at medium blog @anugrahn

# 1.

# Data Acquisition & Cleaning

# Methodology Project



# Methodology Project

**Data cleaning :**

In this cleaning data, we can try with detection outlier for each columns (if any). Pre-processing the data like inconsistent data or missing value that we have in this dataset.

**Data Visualization :**

Creating an interactive visualization that lets we dive down into each data point. From data exploration we can start from hypotheses about our data and the our problem need to be tackling. Finally we can take many insight for many solution or many characteristic (detection pattern or something) for each level of severity.

**Feature Engineering :**

Feature engineering attempts to increase the predictive power of learning algorithms by creating features from raw data that help facilitate the learning process. Beside of future engineering, we can selects the key subset of original data features, its call feature selection

**Training and Test Model :**

Determine optimal data features & create an informative machine learning model that predicts the target most accurately. In this case we can use some alogrithm such as on classification method (Logistic regression, Random Forest Classifier, and many more). The final result would be give a best performance for each modeling that we have done.

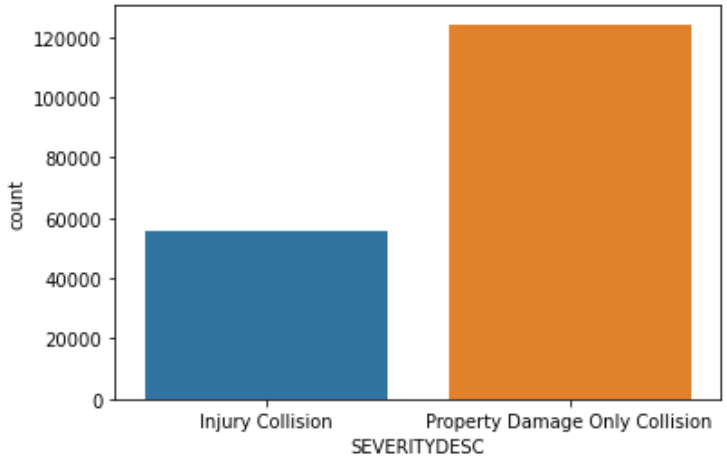
**Evaluation Model :**

The various ways to check the performance of our machine learning especially in classification method, such as recall & precission, f1-score, accuracy score and many more.

# 2.

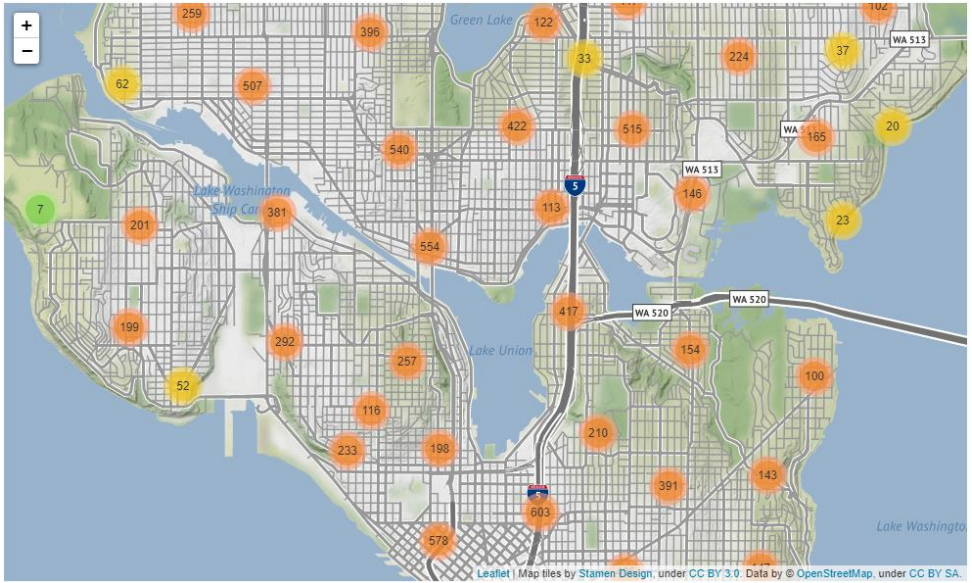
# Exploratory Data Analysis

# Exploratory Data Analysis



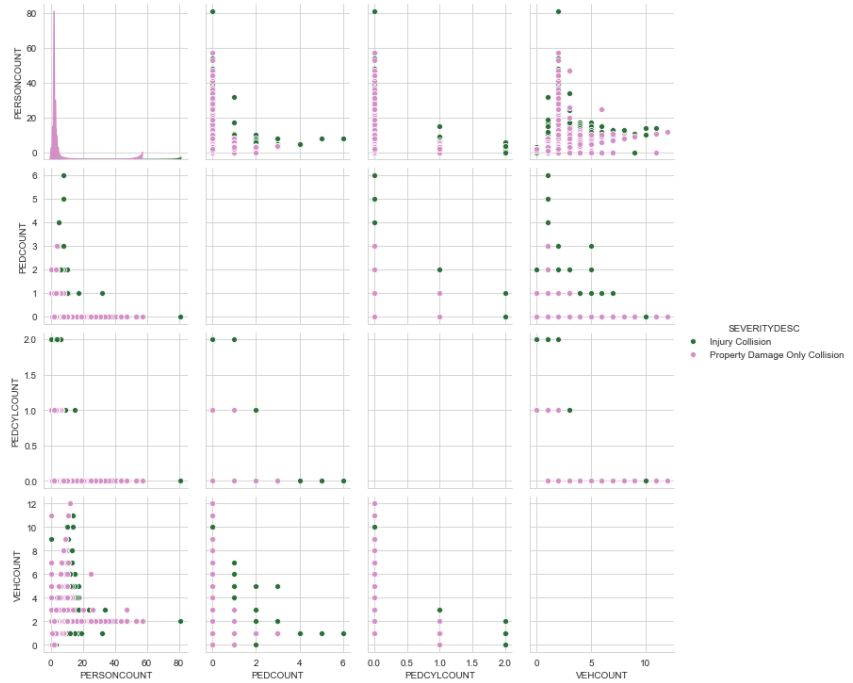

This is the our target we can use *'severity description'* or *'severity code'* to explore the our data. Some visualization in this case we can compare with our target. Because of that, we can take some insight for more deep to learn and understanding some features, what features have high correlation with our target.

# Distribution of collisions by grouping location



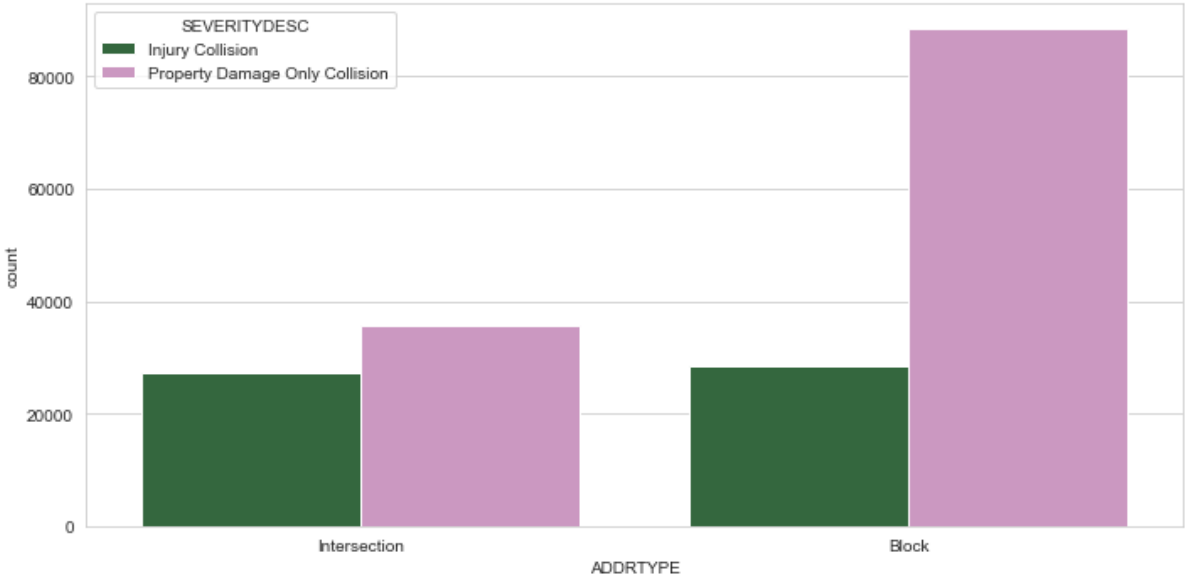

As we see on this folium graph, we grouping for each location on an accident happens. This is a distribution and the total of accident in Seattle.

# Distribution of collisions by grouping location



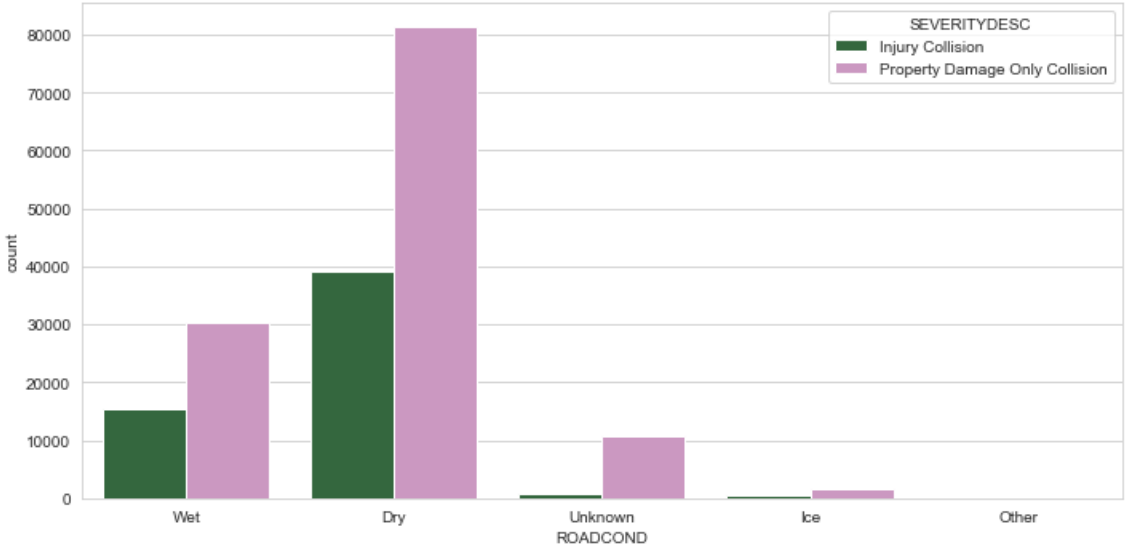

From this graph we conclude, more vehicle count or person count not affected on our target, are the type of accident is injury collision or just property damage?

## Distribution of collisions by grouping location



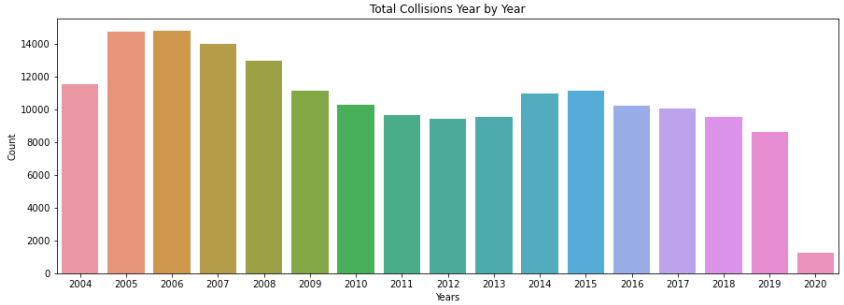

From percentage we know about many people have an accident on address type in block, I assume that block more than dangerous if we know aware about safety riding and the percentage of injury collisions on intersection **44%** more than on block with **24%**.

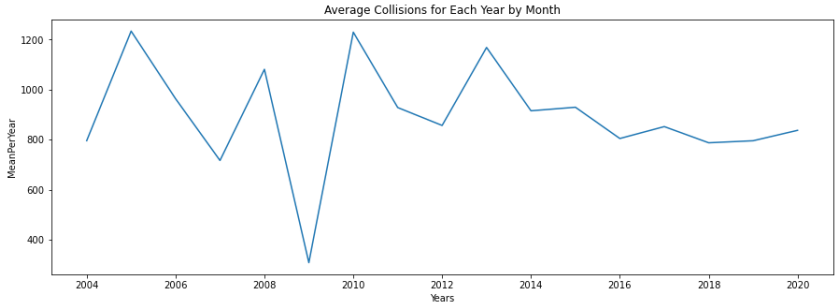
## Distribution of collisions by grouping location




The condition of road when accident happens dominated by the road condition when wet and dry with around **33 %** of the data injury collisions. ***The conclusion about this categorical features analysis, maybe around 40% percent weather, light condition, condition of road is affected to the accident collisions.***

# Distribution of collisions by grouping location




From this 2 graph above we can conclude that.

- Average collisions per month in one year the lowest year is 2009 in range 2004 until 2020 with 300 collisions accident.
- Pattern in total collisions for each year is decreased year by year until 2020.
- The highest total collisions in the data from 2004 increasing to 2005 and hold until 2006 and decrease after that.
- I assume in 2004 until 2006, the Regulations not yet completed and many people many people are not aware of the importance of safe driving. As time goes by until now, the collisions accident is better than the previous year.
- The dataset so many lacks data on 2020, because 2020 not yet completed.

3.

# Predictive Modeling

In this modeling we wanna try Logistic Classifier, Random Forest Classifier & Light GBM Classifier

# Hyperparameter tuning

### Logistic Regression

Final hyperparameter tuning in Logistic Regression we have **'{'class_weight': None, 'penalty': 'l1', 'solver' : 'liblinear'}'**. After that predict with data test and the final result for each metrics in classification report we can describe after this.

### Random Forest Classifier

Final hyperparameter tuning in Random Forest Classifier we have **{ 'max_depth': 4, 'max_features' : 2, 'min_samples_leaf' : 2, 'min_samples_split' : 10}**. After that predict with data test and the final result for each metrics in classification report we can describe after this.

### Light GBM Classifier

Final hyperparameter tuning in Light GBM we have **{'boosting_type': 'dart', 'class_weight': {0: 1, 1: 8}, 'num_leaves': 50}**. After that predict with data test and the final result for each metrics in classification report we can describe after this.

# Result

From 3 model that we build, we can conclude that the best model and we can take from our conslusion is modeling with Random Forest Classifier. Because, the model have a score **recall 83%** and **the accuracy is 65%**, like we discuss before. In this case we can optimize recall after that we evaluation the best accuracy score. If we compare with **LGBM Model that have recall 96% but the accuracy score is 55 %**. Beside we optimze recall, we compare the accuracy score too, for balancing prediction of the model.


CONCLUSION

## Confusion Matrix

**Precision**

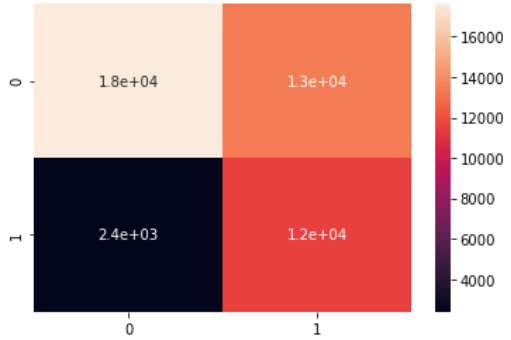
**Precision is the accuracy model predict a positive or negative class**

**Recall**

**Recall is the accuracy of positive or negative predictions from the actual data.**

**Accuracy**

**Accuracy is the accuracy of the overall prediction**




# 83 %

Score Recall (Great > 80%)

# 65 %

Accuracy Score (acceptable > 60%)

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.88 | 0.57 | 0.69 | 31065 |
| 1 | 0.46 | 0.83 | 0.59 | 13952 |
| accuracy | | | 0.65 | 45017 |
| macro avg | 0.67 | 0.70 | 0.64 | 45017 |
| weighted avg | 0.75 | 0.65 | 0.66 | 45017 |

# THANKS!

Finally, we have the best model for our case. But, we can more evaluation this model with f1-score and ROC Curve, how the performance our model when we evaluation with that model, is the best model or we have another model to give more best and effective model.