# Introduction

Recklessly driving and not following the traffic rules are the major reasons why there occurs a lot of collisions and accidents leading to deaths. The pedestrians are in the area of severe loss, though there also exists a great loss to the public and private property.

Some of the major reasons of accidents includes:

- Not following the traffics rules
- Talking on the phone
- Drink and Drive
- Driving above the speed limits
- Parking in the Non-Parking areas
- Entering the No-Entry areas
- Not slowing down in the areas of schools and heavy population areas

# Business Problem

Road Accidents are the major cause of death all around the globe. These accidents account for nearly 70% to 80%(Approx.) of deaths.

In this model we wish to target the most common reason why road accidents occur. We look at the most important aspects and attributes for the road accident, targeting the most common and the most fatal roads/areas in terms of most accidents/fatality rates.

The model basically segments the data points in terms of their severity codes using unsupervised learning algorithms. The data basically consists of several severity codes but the dataset used contains only two severity code viz. 1 and 2.

Hence this classification algorithm falls under the binary class classification problem.
The classification approach uses K – Nearest Neighbors, Support Vector Machine as well as Logistic Regression for the determination of the best accuracy and best classification.

# Target Audience

This model wishes to answer who is most affected by these accidents.

- Pedestrians
- Motorists
- Cyclists

The model seeks to answer whether or not the future accidents are likely to occurs, given the condition of the day, roads and the conditions of lights.

Given the data points the model wish to determine the severity in the collision as well as the number of accidents that might happen.

# Data

The Data used for the analysis and prediction contains a lot of rows and columns regarding the collision's statistics.
The dataset is rich in attribute and contains a lot of information needed for analysis.
It includes severity, fatal rates, timezones, no of accidents, types of injuries and many more attributes.

## Data Set Summary

The dataset contains all collisions statistics provided by SPD and recorded by Traffic Records. This includes all types of collisions.
***Timeframe:*** 2004 to Present.

The dataset has a lot of missing values and outliers. Upon examination of the data set we came down to 9 features which were to be used by the model for the training and evaluation.
These data features are listed below.

### *The various attributes of the dataset are:*

- **OBJECTID**: unique identifier
- **SHAPE**: Geometry field

- **INCKEY**: Long unique key for the incident
- **ADDRTYPE**: Collision address type:
    - Alley
    - Block
    - Intersection
- **INTKEY**: Key that corresponds to the intersection associated with a collision.
- **LOCATION**: Location where the collision occurred
- **SEVERITYCODE**: A code that corresponds to the severity of the collision:
    - 3—fatality
    - 2b—serious injury
    - 2—injury
    - 1—prop damage
    - 0—unknown
- **WEATHER**: 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 not granted. (Y/N)
- **SPEEDING**: Whether or not speeding was a factor in the collision. (Y/N)
- **UNDERINFL**: Whether or not a driver involved was under the influence of drugs or alcohol.
- **INATTENTIONIND**: Whether or not collision was due to inattention. (Y/N)
- **JUNCTIONTYPE**: Category of junction at which collision took place

and many more.

*Some of the essential attribute for the training and building the model are as follows:*

- LOCATION
- WEATHER
- ROADCOND
- LIGHTCOND
- PEDROWNOTGRNT
- SPEEDING
- UNDERINFL
- INATTENTIONIND
- JUNCTIONTYPE

### *State Collision Code:*

| Code | Description |
|---|---|
| 0 | Vehicle Going Straight Hits Pedestrians |
| 1 | Vehicle Turing Right Hits Pedestrians |
| 2 | Vehicle Turning Left Hits Pedestrians |
| 3 | Vehicle Backing Hits Pedestrians |
| 4 | Vehicle Hits Pedestrians – All Other Actions |
| 5 | Vehicle Hits Pedestrians –Actions Not Started |
| 10 | Entering At Angle |
| 11 | From Same Direction – Both Going Straight – Both Moving Sideswipe |
| 12 | From Same Direction – Both Going Straight – One Stopped- Sideswipe |
| 13 | From Same Direction – Both Going Straight – Both Moving – Rear End |

# *Exploratory Data Analysis*

## *Selecting a Target Variable*

As the accidents are the major cause of the deaths all around the world, the model seeks to answer the severity of these accidents given the appropriate data points causing the accidents. The severity code is to be grouped in to its label using unsupervised learning approach.

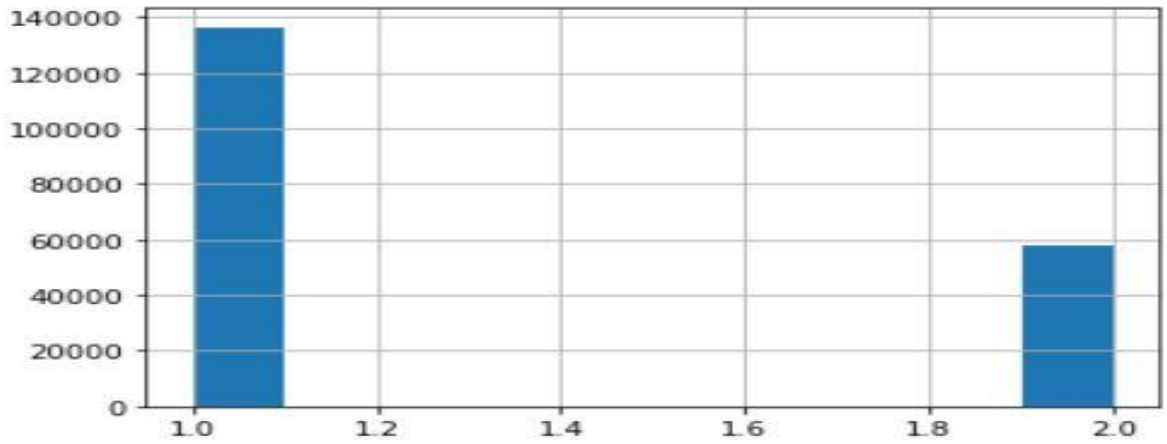
So, our target variable will be severity code. The dataset contains only two different kinds of severity code, namely 1 and 2.

### *Severity Code – Number of Accidents*

Analyzing the data gives us the insight of the severity code and the number of accidents related to that severity code.

- There was a total of ***136485*** numbers of accidents, which were of ***Severity type: 1***
- There was a total of ***58188*** numbers of accidents, which were of ***Severity type: 2***

The plot clearly shows the relation between the Severity code on the X-axis and the number of accidents on the Y-axis.

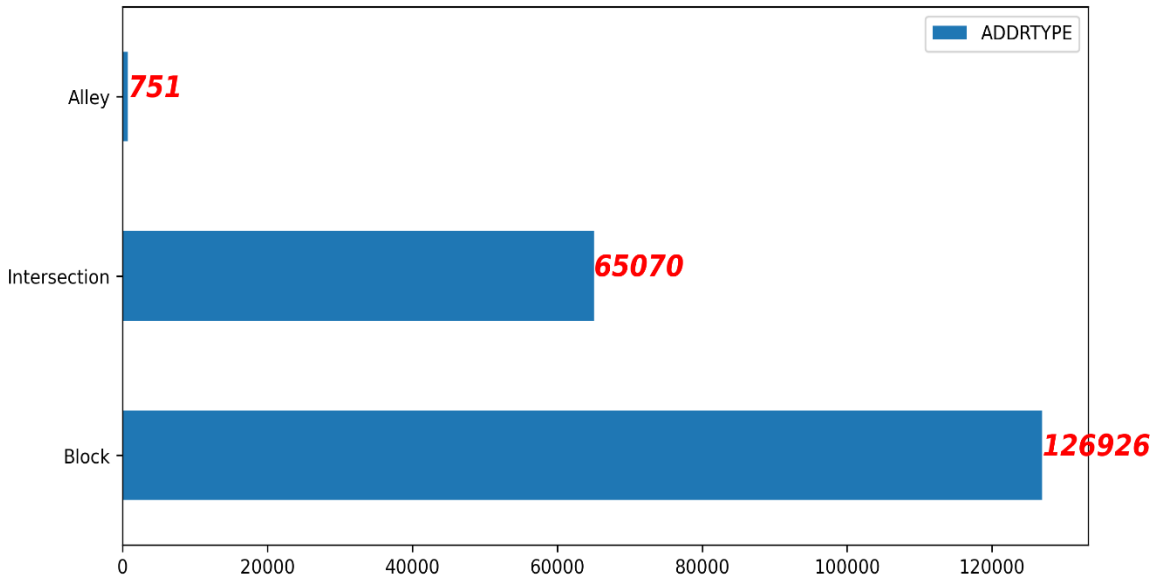



## *Address Type and Number of Accidents*

A total of ***192747*** accidents have taken place on the different address type.

The total number of accidents on the address type

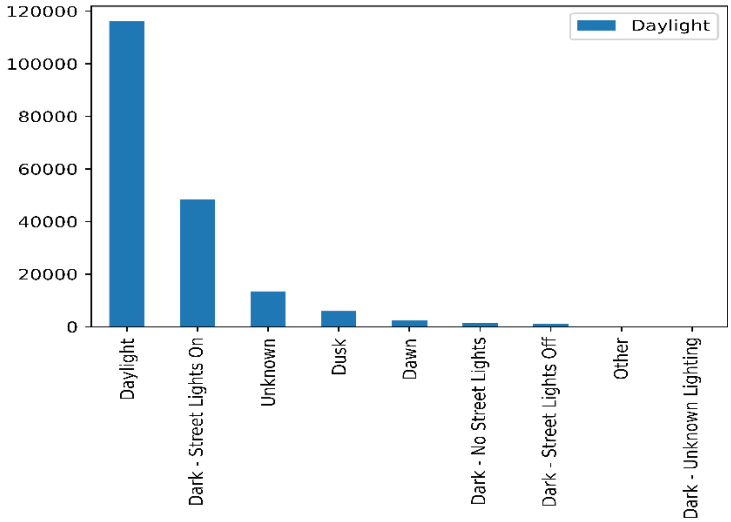
- ***Alley*** accounts for a total of ***751*** accidents
- ***Intersection*** accounts for a total of ***65070*** accidents, and
- ***Block*** accounts for a total of ***126926*** accidents.
- The rest of ***1926*** accidents were not reported in any of the address types



## *Light Condition and Number of Accidents*

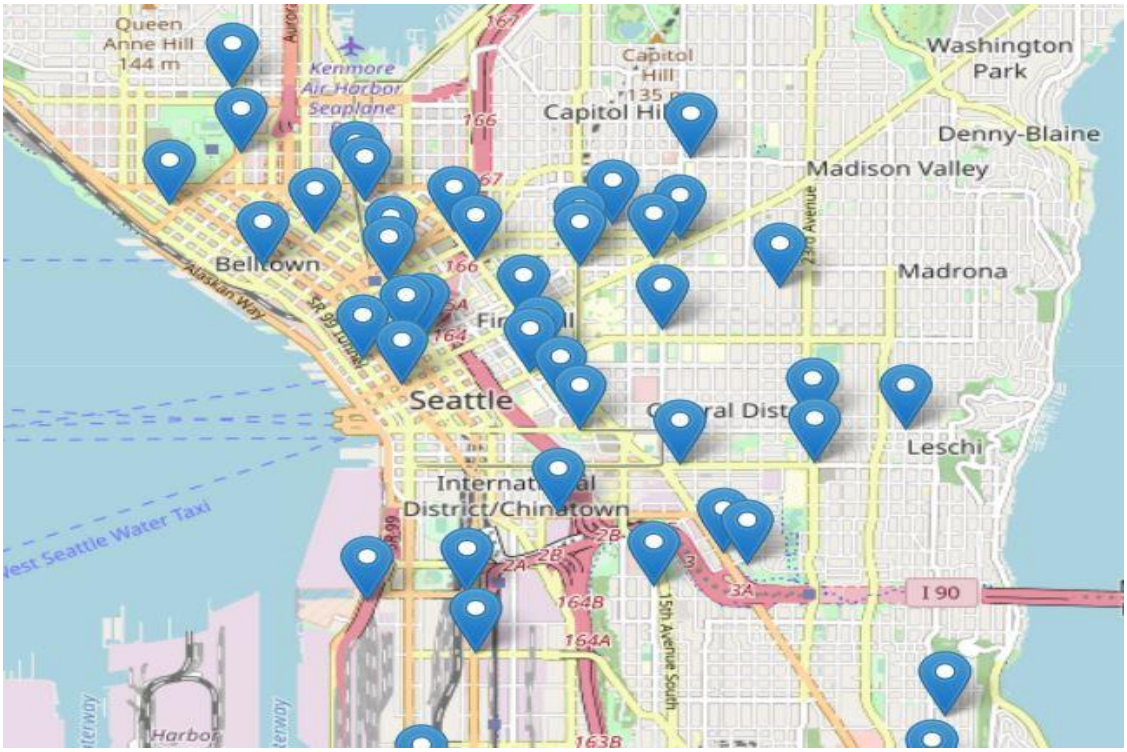
The data frame consisting of number of accidents in different light condition is:

| | LIGHTCOND |
|---|---|
| **Daylight** | 116137 |
| **Dark - Street Lights On** | 48507 |
| **Unknown** | 13473 |
| **Dusk** | 5902 |
| **Dawn** | 2502 |
| **Dark - No Street Lights** | 1537 |
| **Dark - Street Lights Off** | 1199 |
| **Other** | 235 |
| **Dark - Unknown Lighting** | 11 |

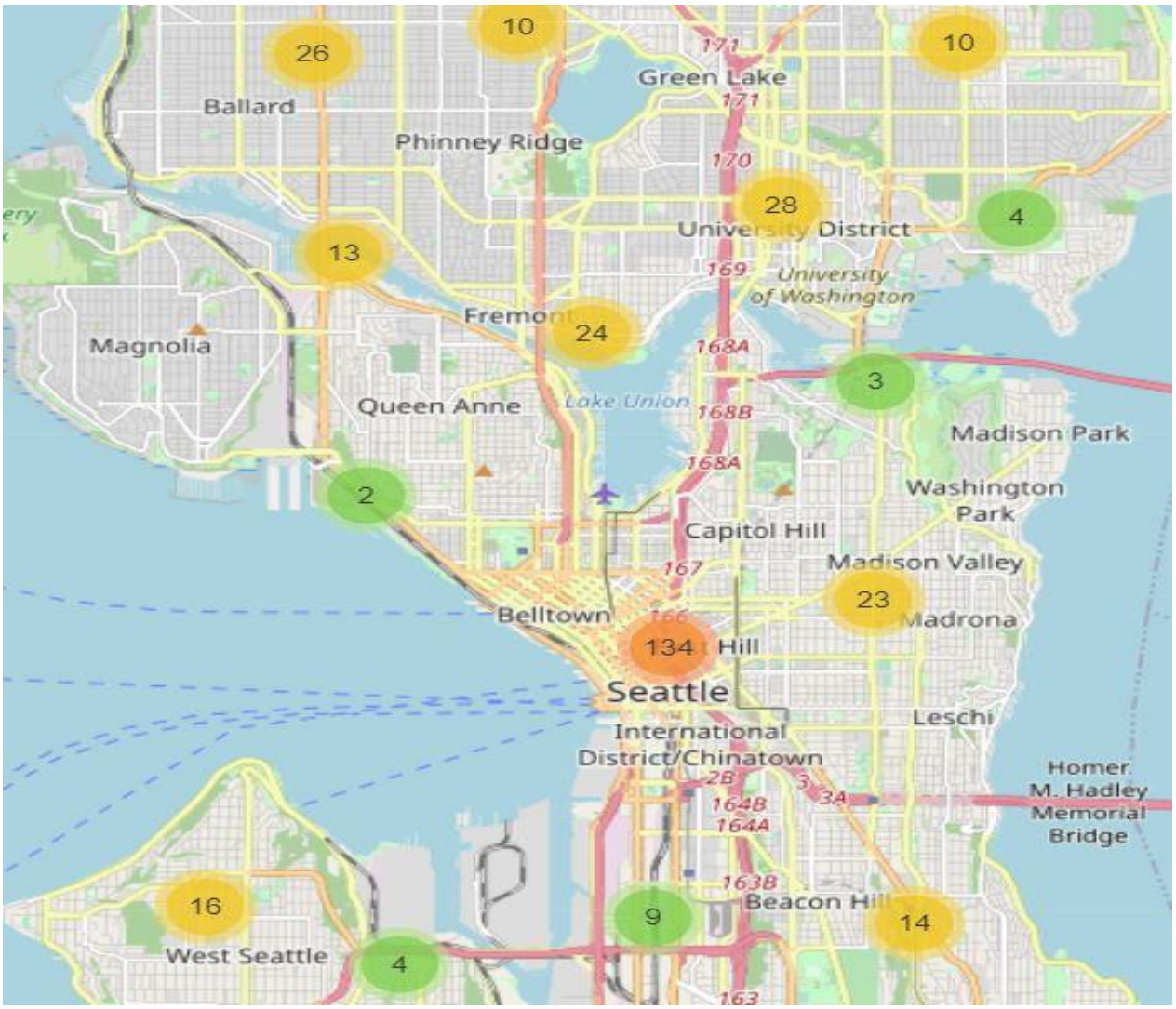



## *Number of Collisions in different Areas*

***Plotting 100 accidents on the map show***



***Accidents grouped into clusters form***




# Modeling

There are various different algorithms that can be used to model the collisions data. The model utilizes classification approach to group the data in to two categories.

As the SEVERITYCODE has only two labels hence, the model uses binary class classification approach to classify the data points.

For the different types of categorical variable, the values were substituted as corresponding to its numerical for.

For the values of ROADCOND, they were changed to:

- 'Clear' - 0
- 'Raining' - 1
- 'Overcast' - 2
- 'Unknown' - 3
- 'Snowing' – 4
- 'Other' - 5
- 'Fog/Smog/Smoke' - 6
- 'Sleet/Hail/Freezing Rain' - 7
- 'Blowing Sand/Dirt' - 8
- 'Severe Crosswind' - 9
- 'Partly Cloudy' – 10

Similarly, the variable for the LIGHTCOND were changed to:

- 'Daylight' - 0
- 'Dark - Street Lights On' - 1
- 'Unknown' - 2
- 'Dusk' - 3
- 'Dawn' - 4
- 'Dark - No Street Lights' - 5
- 'Dark - Street Lights Off' - 6
- 'Other' - 7
- 'Dark - Unknown Lighting' – 8

For variable with only two discrete values, dummy variables were used in place:

- For UNDERINFL:
  - N: 0
  - Y: 1
- For HITPARKEDCAR:
  - Y: 1
  - N: 0

The data is split into train and test data and the test size was 0.25, i.e. 25% of the data was utilized for testing

# Model Evaluation

Using the Support Vector Machine (SVM) algorithm, the following results were obtained:

| | Precision | Recall | F1 – Score | Support |
|---|---|---|---|---|
| 1 | 0.74 | 0.97 | 0.84 | 32659 |
| 2 | 0.75 | 0.21 | 0.33 | 14217 |