2. Data Acquistion and cleaning

The dataset is available as comma-separated values (CSV) files, KML files, and ESRI shapefiles that can be downloaded from the Seattle Open GeoData Portal. The data is also available from RESTful API services in formats such as GeoJSON.

2.1 Loading the dataset

We download the dataset to our project directory and take a look at the data types and the dimensionality of the data. We can see that the dataset contains 194,673 records and 40 fields.

The metadata of the dataset can be found from the website of the Seattle Department of Transportation. On reading the dataset summary, we can determine the description of each of the fields and their possible values.

The data contains several categorical fields and corresponding descriptions which could help us in further analysis. We make an attempt at understanding the data in terms of the fields that we shall take into account for later stages of model building.

The X and Y fields denote the longitude and latitude of the collisions. We can visualize the first few non-null collisions on a map.

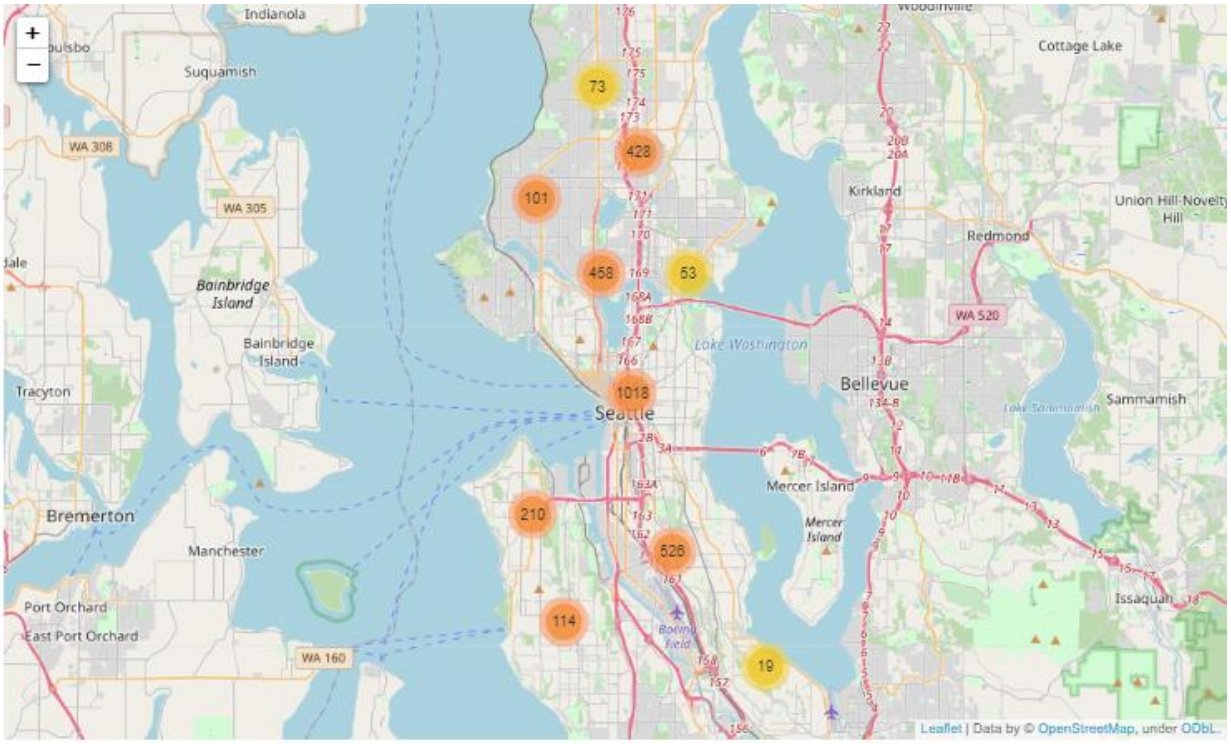

Fig 1. Map of Seattle city with accident data across the city

2.2 Data Cleaning

The dataset used for this project is based on car accidents which have taken place within the city of *Seattle, Washington* from the year *2004* to *2020*. This data is regarding the *severity of each car accidents* along with the time and conditions under which each accident occurred. The data set used for this project can be found on Seattle website. The model aims to predict the severity of an accident, considering that, the variable of Severity Code was in the form of *1* (Property Damage Only) and *2* (Physical Injury). The dataset has 194673 rows and 38 features.

In order to deal with the issue of columns having a variation in frequency, arrays were made for each column which were encoded according to the original column and had equal proportion of elements as the original column. Then the arrays were imposed on the original columns in the positions which had *Other* and *Unknown* in them. This entire process of cleaning data led to a loss of almost 5000 rows which had redundant data, whereas other rows with unknown values were filled earlier.

2.3 Feature Selection

After data cleaning, there were 169957 rows and 38 features (columns) in the dataset. Upon examining it was clear that there are some redundancies in the features. For e.g. 'INCDTTM' and 'INCDATE' both features contain the data about the data and time of the accident. The difference between the two is that 'INCDATE' feature only contains the date of the accident and 'INCDTTM' contains the date as well as time data of the accident.

There were other features that had no effect on the severity of the accident. Those features are

- 'OBJECTID'
- 'INCKEY'
- 'COLDETKEY'
- 'REPORTNO'
- 'STATUS'
- 'ADDRTYPE'
- 'INTKEY'
- 'LOCATION'
- 'EXCEPTRSNCODE'
- 'EXCEPTRSNDESC'
- 'SEVERITYDESC'
- 'COLLISIONTYPE'
- 'SDOT_COLCODE'
- 'SDOT_COLDESC'

- 'SDOTCOLNUM'

- 'ST_COLCODE'
- 'ST_COLDESC'
- 'SEGLANEKEY'
- 'CROSSWALKKEY'
- 'HITPARKEDCAR'

These features are either unique identifier for the dataset set such as INTKEY, OBJECTID, etc. or are the description of the codes used for severity of an accident example SDOT_COLCODE, SDOT_COLDESC, etc.

The INCDTTM feature was used to analyze the number of accidents over the years but later dropped. On analyzing the data using this feature we found out that the number of accidents decreased over the years. Correlation also showed that there is some relation between the features (Pearson correlation coefficient > 0.3). For example, PERSONCOUNT was related to VEHCOUNT as this seems logical too.

The features that were selected for the prediction are:

- X
- Y
- PERSONCOUNT
- PEDCOUNT
- PEDCYLCOUNT
- VEHCOUNT
- UNDERINFL
- ROADCOND
- LIGHTCOND
- WEATHER
- SPEEDING