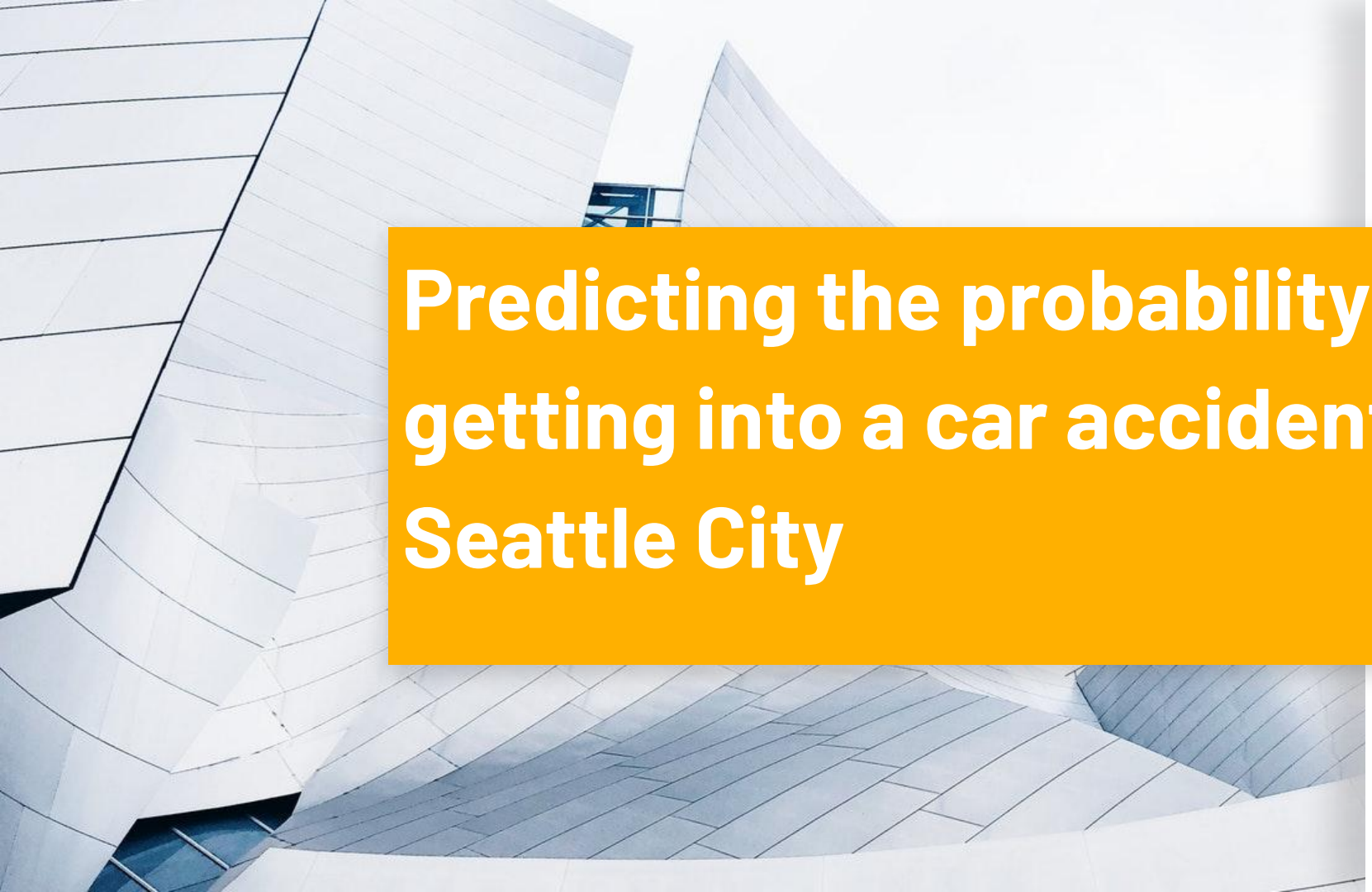


Predicting the probability of getting into a car accident in Seattle City

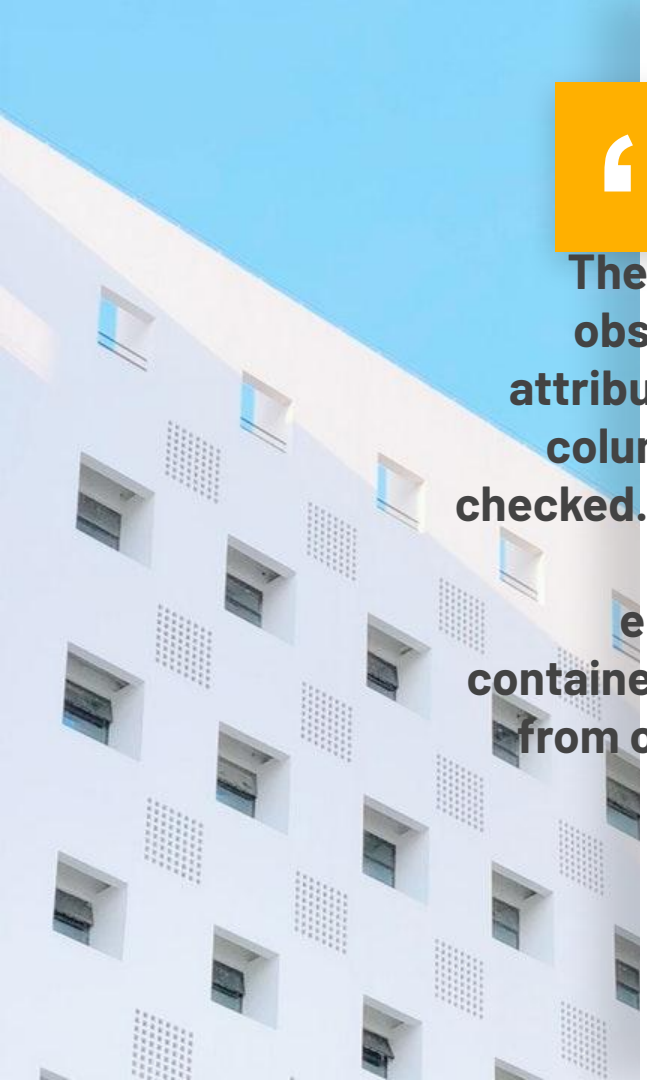
Background

Seattle city has really troubles to deal with the accidents data. When people are driving to another city for work or to visit some friends. It is rainy and windy, and on the way, they could come across a terrible traffic jam on the other side of the highway. Long lines of cars barely moving. As the people keep driving, police car start appearing from afar shutting down the highway. It is an accident and there's a helicopter transporting the ones involved in the crash to the nearest hospital. They must be in critical condition for all of this to be happening.

Problem

The objective of the project is describing fatality of accidents. Given that the number of accidents on busy roads is creating problems in the city's transportation. It has the necessity to determine under what circumstances a serious traffic accident is most likely to occur.

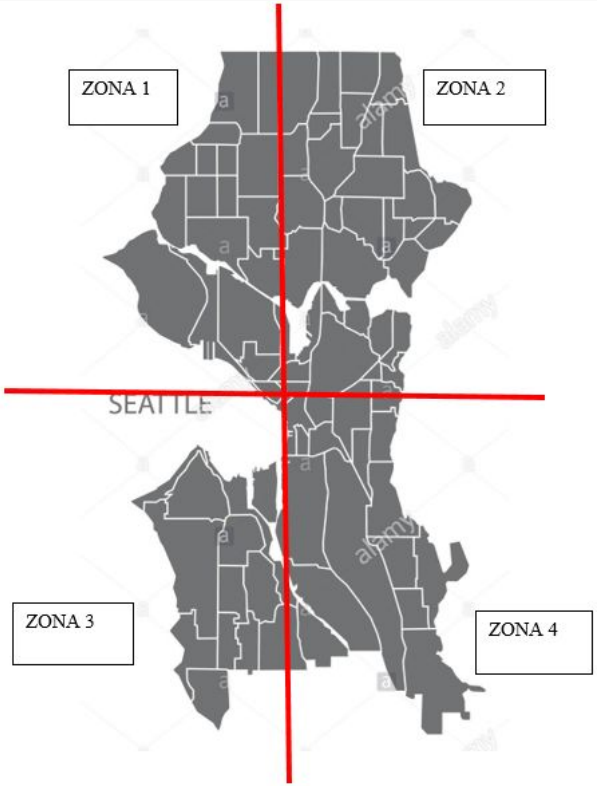
Data cleaning



The dataset contain 194673 observations (rows) and 38 attributes (columns). First the columns that have NaN were checked. Columns that had more than half like this were eliminated. Columns that contained duplicate information from other columns were also removed.

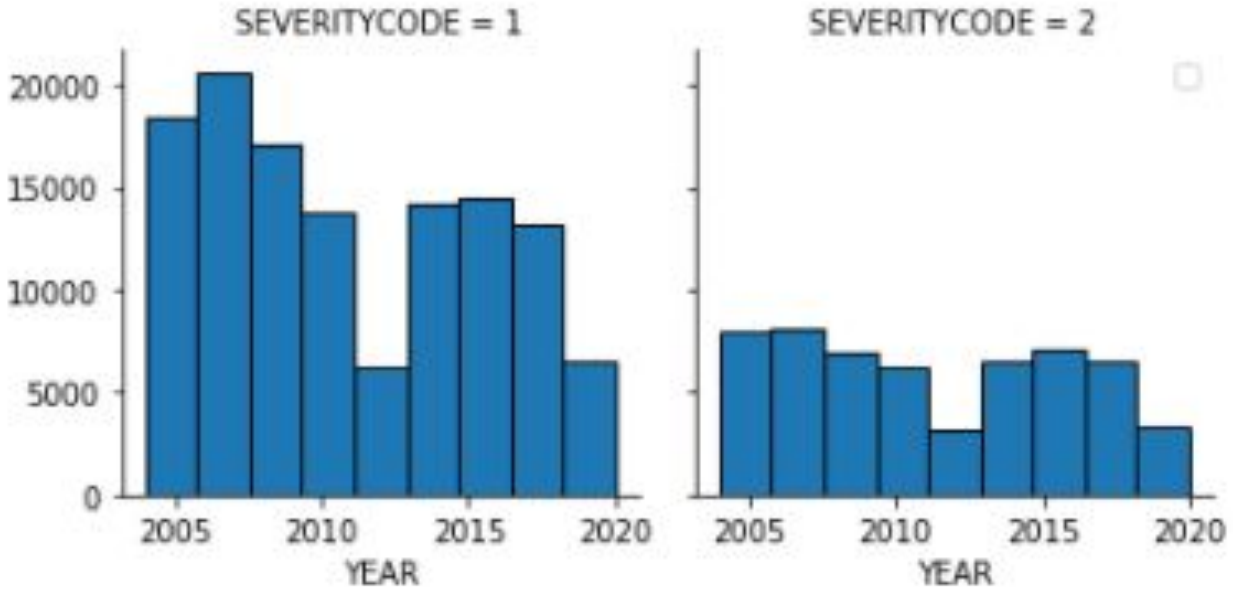
```
df.drop(['INCKEY', #irrelevant
         'OBJECTID', #irrelevant
         'REPORTNO', #irrelevant
         'STATUS', #irrelevant
         'LOCATION', #Duplicate Information,
         'SDOTCOLNUM', #Duplicate information
         'SDOT_COLCODE',
         'SDOT_COLCODE',
         'INTKEY', #irrelevant
         'HITPARKEDCAR',
        'COLDETKEY',
        'EXCEPTRSNCODE',
        'EXCEPTRSNDESC',
        'SEVERITYCODE.1',
        'INATTENTIONIND',
        'SDOTCOLNUM',
        'SPEEDING', #many missing data
        'ST_COLCODE',
        'ST_COLDESC',
        'SEGLANEKEY',
        'ST_COLCODE',
        'CROSSWALKKEY',
        'PEDROWNOTGRNT',
         'UNDERINFL',
         'PERSONCOUNT',
         ], axis = 1, inplace = True)
```

In such a way that each data will be grouped in one of these Zones and know which of these is more prone to accidents.

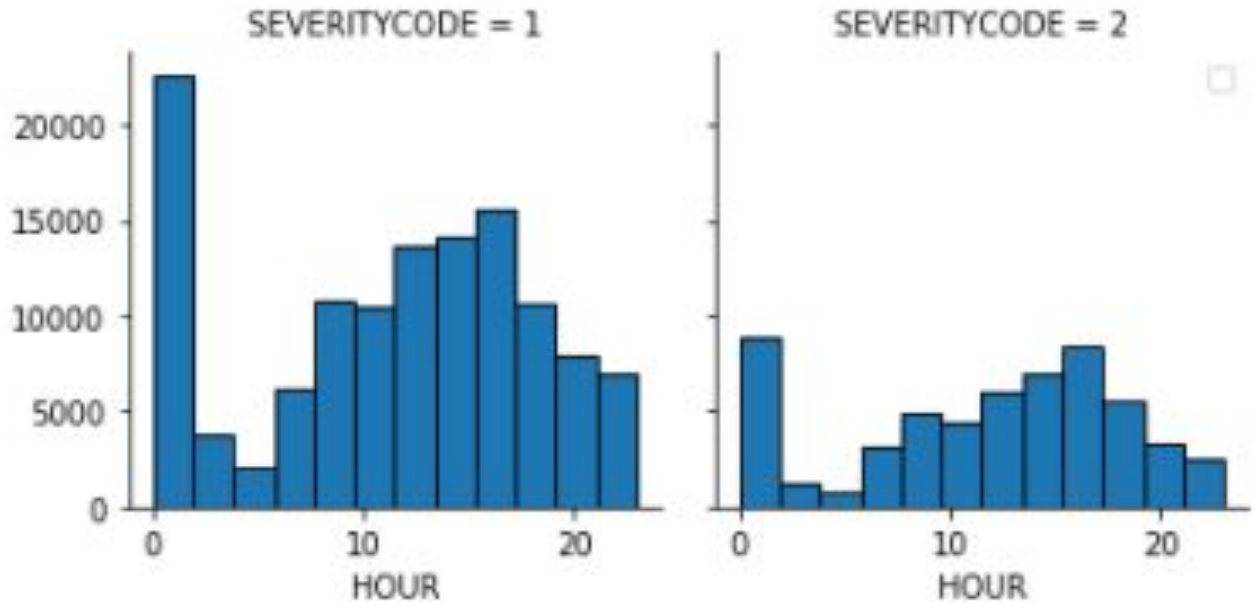


Exploratory Data Analysis

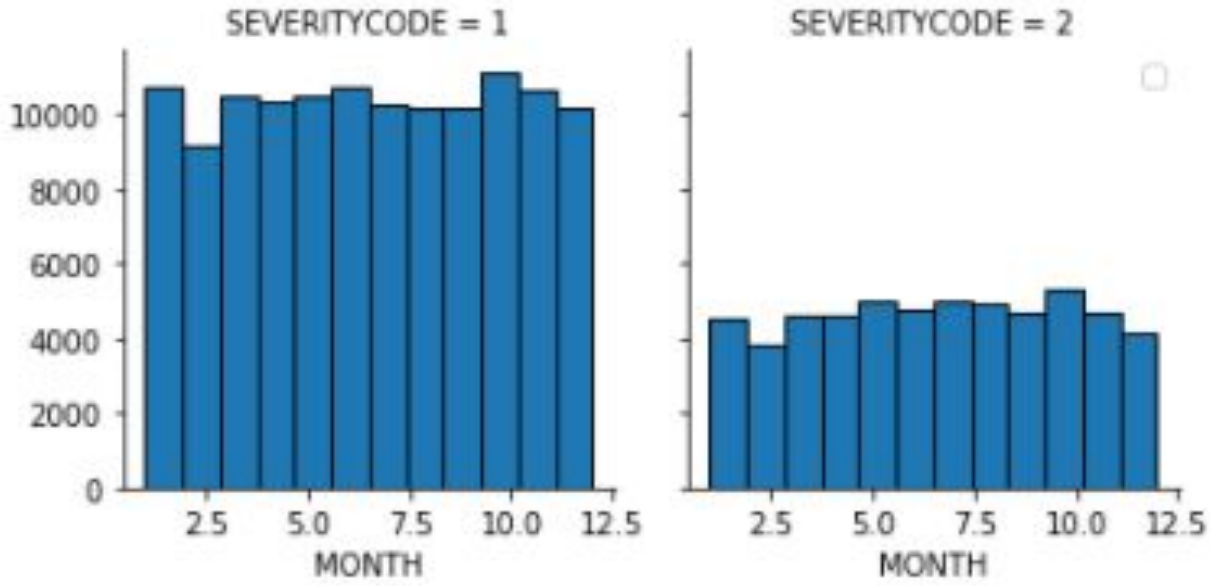
'YEAR' and "SEVERITY"



'HOUR' and "SEVERITY"



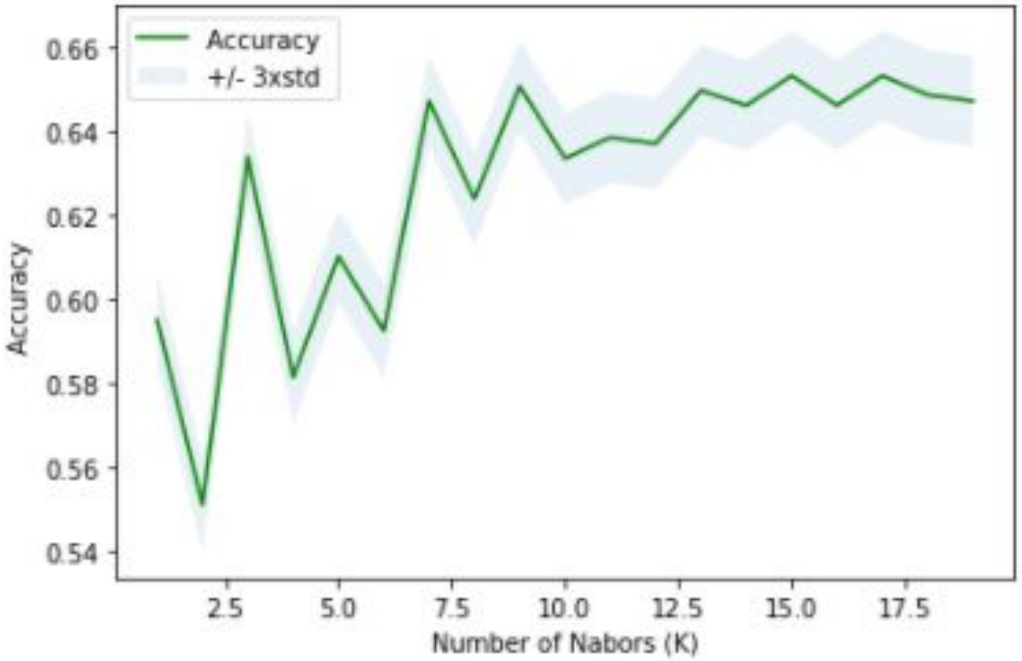
'MONTH' and "SEVERITY"



Predictive Modeling

In this case, given that the variables are not continuous, that is, they are discrete. A Classification method had to be used. In this case, K nearest Neighbors was chosen.

4.1.1 K Nearest Neighbor



The best accuracy was with 0.6530303030303031 with k= 15

jaccard KNN: 0.6530303030303031

F1-score KNN: 0.6259979935509098

THANKS!