Predicting Car Accident Severity in Seattle

Prepared by : **Suganthi Sathiyamoorthi**
For : **IBM DataScience Capstone Project**
Submitted On : **Oct 4th 2020**

Predicting Car Accident Severity in Seattle

2. Data Preprocessing

2.1 Data Sources

Data provided with this project request is used for this analysis. A more raw data can from Seattle GeoData where most of the data and attributes can be studied. Even though raw dataset has records from 2004 and has data for all severitycodes and studied for possible utilization for this project, course provided dataset with just 2 severity codes was utilized due to very unbalanced nature of the data. Models were evaluated for utilizing unbalanced data but could not use it because of heavily unbalanced data that requires further analysis and utilization of advanced techniques. The imported data has 194673 observations of various attributes such as severity code, severity code description, Address type, Junction type, collision type, Road Condition, Light Condition, Weather Condition, Speeding, Driving Under Influence indicator, Longitude, Latitude, Injury count, Fatality count, SDOT code, description, Data and Time of Accident etc., The attribute Types and descriptions can be found at Seattle DOT

2.2 Data Cleaning

Data downloaded from data source with nulls and Nan values were replaced with Unknown/Other values. Data formats were converted to other formats as required. A few columns such as Severity Code, Description & SDOT Code, Description & ST Code and Description were combined to new columns for analysis purposes. X and Y co-ordinates missing Latitude and Longitude were substituted with Seattle's Latitude and Longitude. Upon cleaning, it can be observed that the observations are available from April 1st 2004 to May 20th 2020.

2.3 Exploratory Data Analysis

As the source data has many attributes, all the attributes were studied for possible selection as feature for this project. Individual Attribute was checked for percentage valid data that might be a possible reason for the accident. Such attributes were further analyzed with respect to the target attribute of severity. The following shows the results of this analysis:

Total Data Observations : 194673

Attribute	Selected(Y)/Not Selected (X)	Reason
OBJECTID	X	Doesn't provide relevant information for Prediction
SHAPE	X	Doesn't provide relevant information for Prediction
INCKEY	X	Doesn't provide relevant information for Prediction

Predicting Car Accident Severity in Seattle

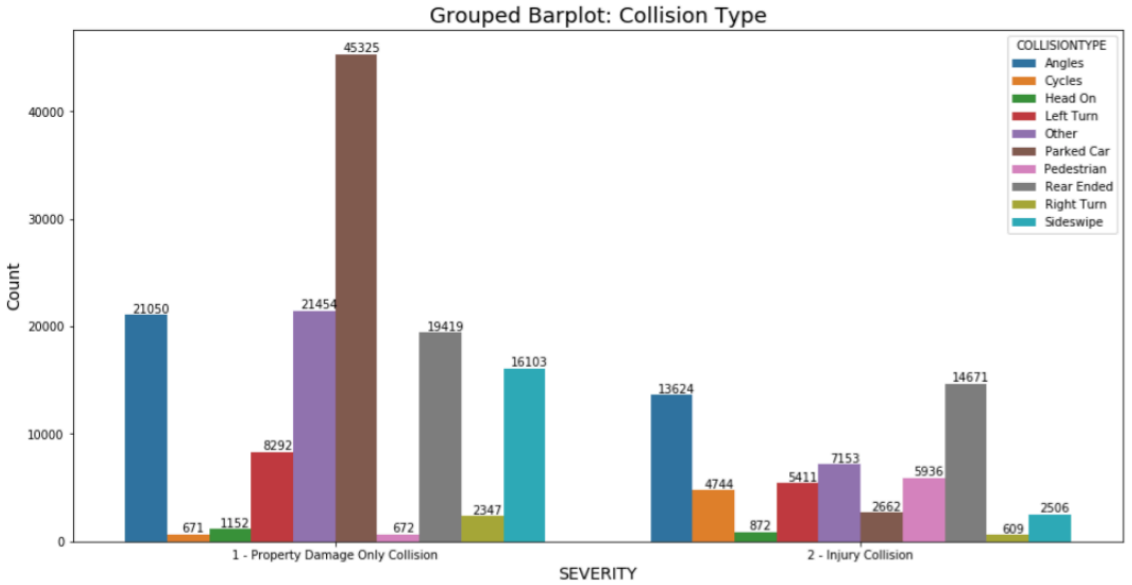
COLDETKEY	X	Doesn't provide relevant information for Prediction
ADDRTYPE	X	Collision address type: • Alley • Block • Intersection **Junction type provides more information than Address Type **
INTKEY	X	Doesn't provide relevant information for Prediction
LOCATION	X	Description of the general location of the collision , ** **X and Y co-ordinate provides more info than Location** **
EXCEPTRSNCODE	X	Doesn't provide relevant information for Prediction
EXCEPTRSNDESC	X	Doesn't provide relevant information for Prediction
SEVERITYCODE	Y	A code that corresponds to the severity of the collision: • 3—fatality • 2b—serious injury • 2—injury - Count - **58188** • 1—prop damage – Count - **136485** • 0—unknown
SEVERITYDESC	Y	A detailed description of the severity of the collision
COLLISIONTYPE	Y	Collision type Provides more information about the collision
PERSONCOUNT	X	The total number of people involved in the collision ** Not selected as this information pertains to particular accident **
PEDCOUNT	X	The number of pedestrians involved in the collision. This is entered by the state. ** Not selected as this information pertains to particular accident **
PEDCYLCOUNT	X	The number of bicycles involved in the collision. This is entered by the state. ** Not selected as this information pertains to particular accident **
VEHCOUNT	X	The number of vehicles involved in the collision. This is

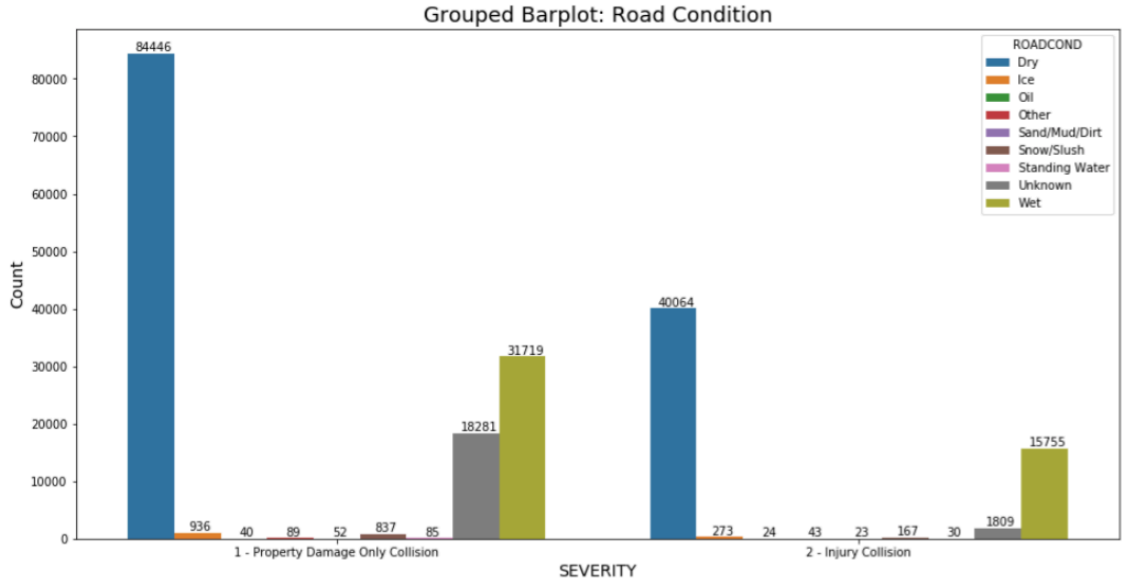
		entered by the state. ** Not selected as this information pertains to particular accident **
INJURIES	X	The number of total injuries in the collision. This is entered by the state. ** Not selected as this information pertains to particular accident **
SERIOUSINJURIES	X	The number of serious injuries in the collision. This is entered by the state. ** Not selected as this information pertains to particular accident and dataset doesnot contain this severity**
FATALITIES	X	The number of fatalities in the collision. This is entered by the state. ** Not selected as this information pertains to particular accident and dataset doesnot contain this severity**
INCDATE	X	The date of the incident.
INCDTTM	X	The date and time of the incident. **Incident Date and Time was used in evaluation but was not selected as time of the day was not available for more that 25000 of the cases. Day of the week was also evaluated but not significant difference was found in the day of the week**
JUNCTIONTYPE	Y	Category of junction at which collision took place
SDOT_COLCODE	Y	A code given to the collision by SDOT.
SDOT_COLDESC	Y	A description of the collision corresponding to the collision code.
INATTENTIONIND	X	Whether or not collision was due to inattention. (Y/N) ** Y datacount was very less **
UNDERINFL	X	Whether or not a driver involved was under the influence of drugs or alcohol. ** Y datacount was very less **
WEATHER	Y	A description of the weather conditions during the time of the

		collision.
ROADCOND	Y	The condition of the road during the collision.
LIGHTCOND	Y	The light conditions during the collision.
PEDROWNOTGRNT	X	Whether or not the pedestrian right of way was not granted. (Y/N)
SDOTCOLNUM	X	A number given to the collision by SDOT.
SPEEDING	X	Whether or not speeding was a factor in the collision. (Y/N) ** N count was very high and Y count very less . Y – `9333 count` **
ST_COLCODE	X	A code provided by the state that describes the collision. For more information about these codes, please see the State Collision Code Dictionary. ** Similar to SDOT **
ST_COLDESC	X	A description that corresponds to the state's coding designation. ** Similar to SDOT **
SEGLANEKEY	X	A key for the lane segment in which the collision occurred.
CROSSWALKKEY	X	A key for the crosswalk at which the collision occurred.
HITPARKEDCAR	X	Whether or not the collision involved hitting a parked car. (Y/N) ** Very Unbalanced datacontent **

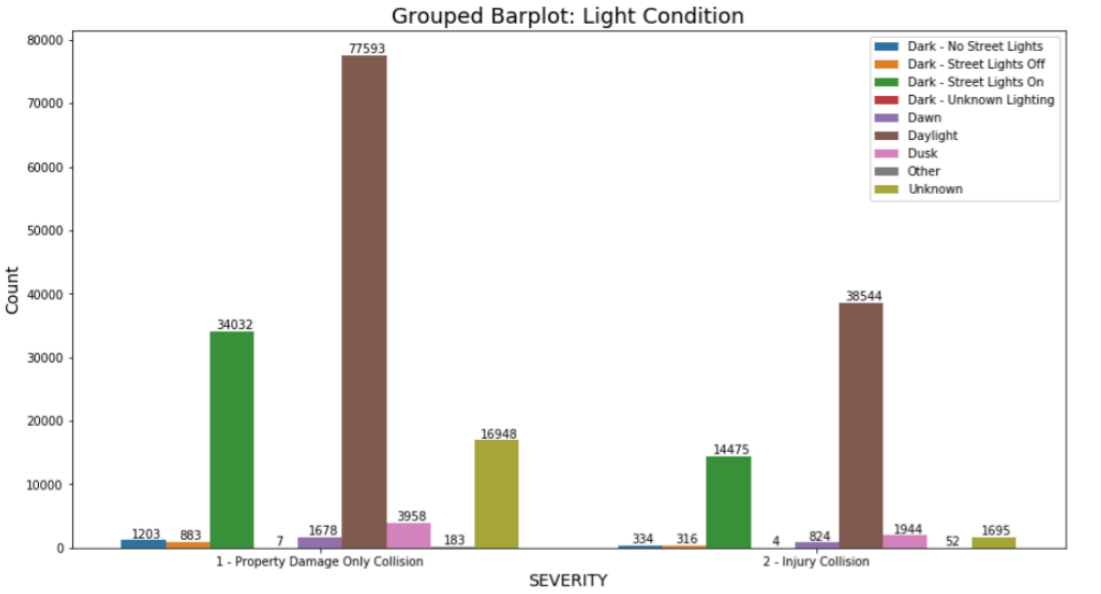
Predicting Car Accident Severity in Seattle

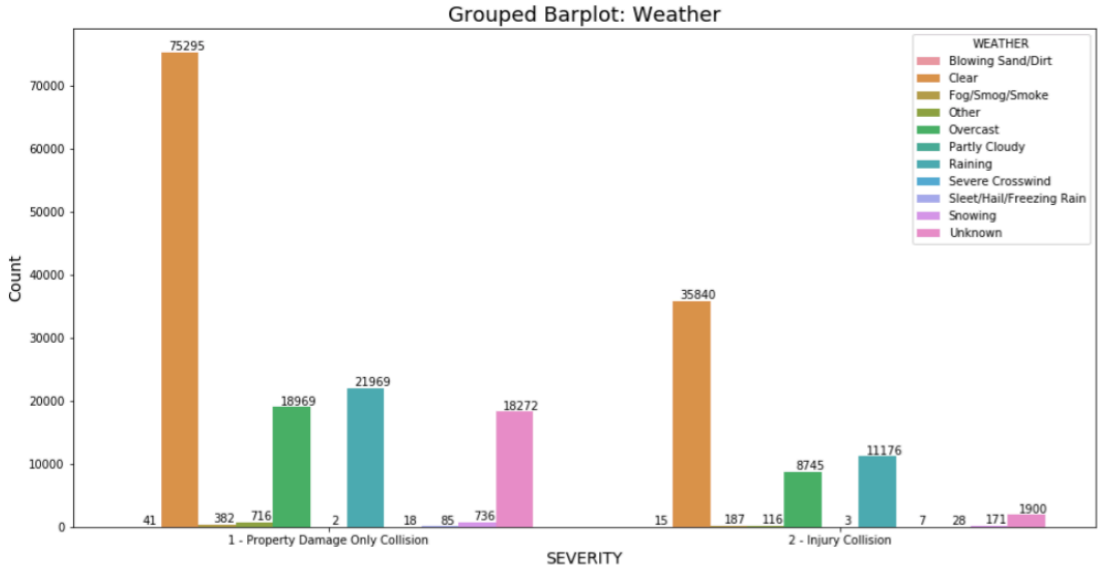
Selected Attributes were analyzed against severity as follows:



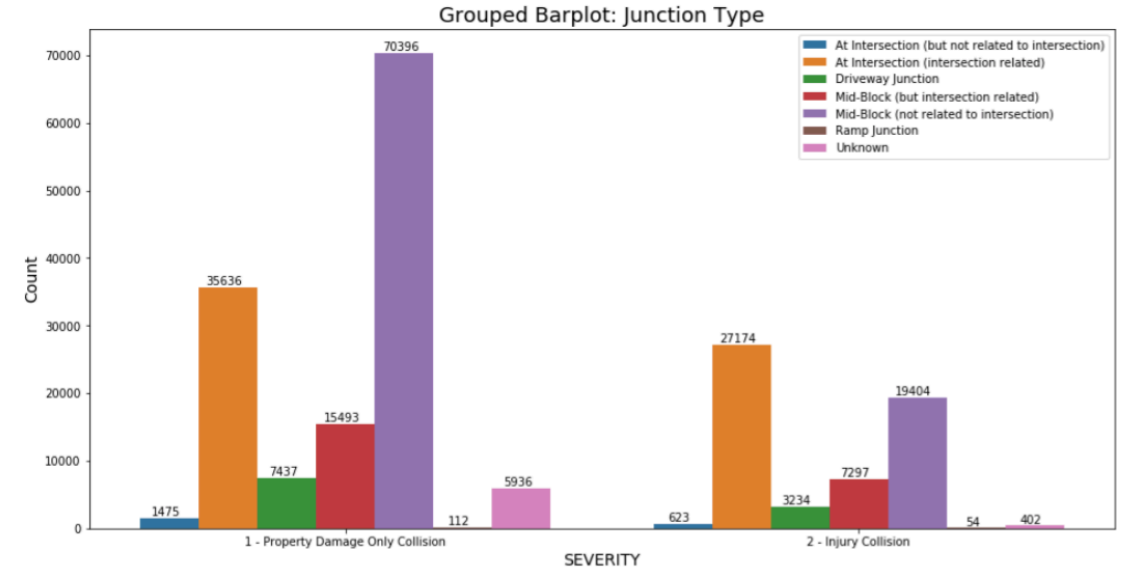


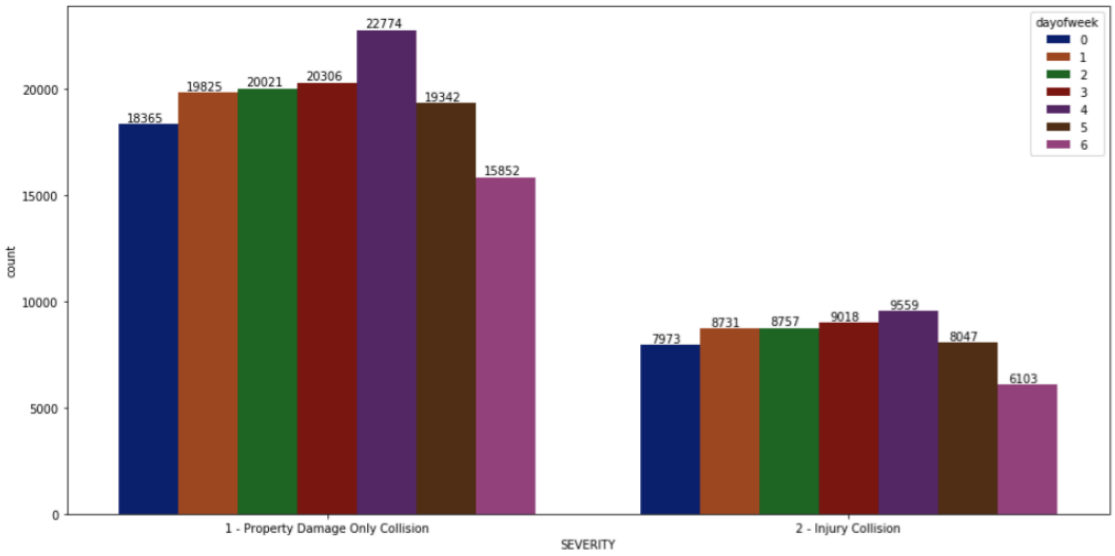
Predicting Car Accident Severity in Seattle



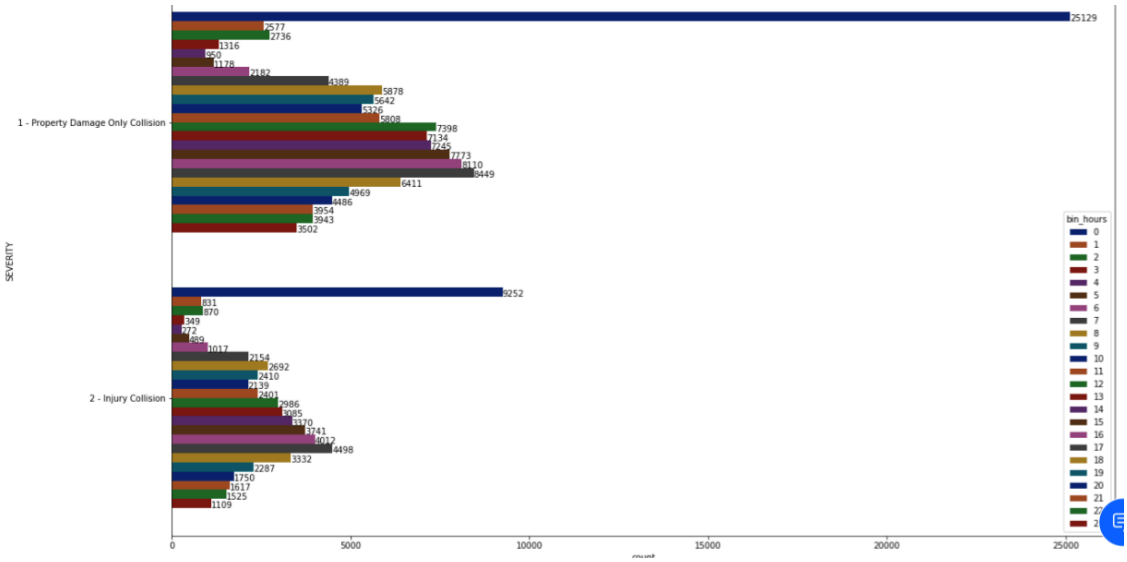


Predicting Car Accident Severity in Seattle

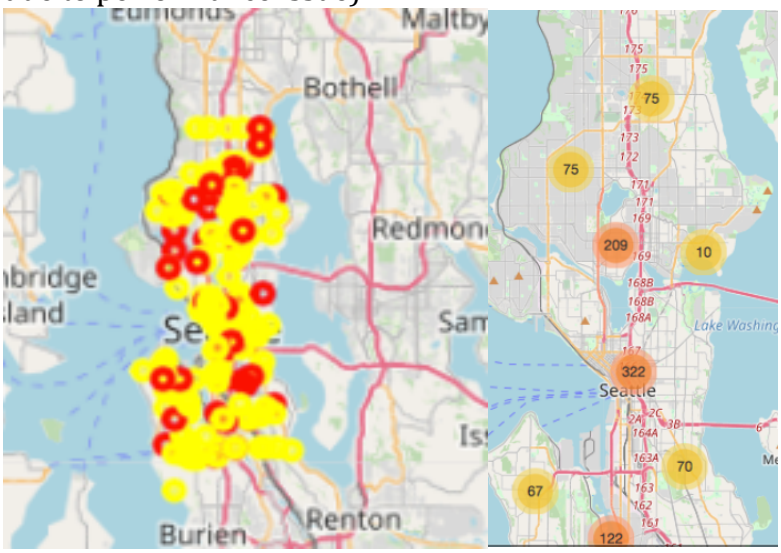




Predicting Car Accident Severity in Seattle



X and Y Co-ordinates were plotted using folium (1000 samples only were plotted due to performance issue)



Predicting Car Accident Severity in Seattle

Based upon the above analysis, the following attributes are selected for feature engineering:

- SEVERITY CODE, SEVERITY CODE DESCRIPTION

```
-       1 - Property Damage Only Collision    136485
        2 - Injury Collision                   58188
```

- X, Y,
- JUNCTION TYPE, COLLISION TYPE
- ROAD CONDITION, LIGHT CONDITION, WEATHER CONDITION
- SDOT CODE, SDOT DESCRIPTION
- INCIDENT DATE and TIME.

The following attributes were not selected based upon 194673 observations:

- SPEEDING – Only 9333 are with Y values
- DRIVING UNDER INFLUENCE – 9121 with Y or 1 values, 185552 with N or 0 values
- ADDRESS TYPE is similar to JUNCTION TYPE. JUNCTION TYPE has more information than ADDRESS TYPE.
- PERSON Counts were not selected as it is related to specific accident.
- INATTENTIONIND was not selected due to very less Y count – 29285
- PEDROWNOTGRNT – very less Y count – 4667
- HITPARKEDCAR - Y count – 7216
- Keys and Ids were not selected for analysis