Report - Predicting Car Accident Severity in Seattle

Overview

There were 194,674 reported collisions between 2004 and 2020 in the Seattle Metropolitan Area. There is an opportunity to review common characteristics, including the time of day, weather, road and lighting conditions, geographic location, types of vehicles involved, presence of impairment by drugs or alcohol of individuals involved, among other factors to determine the severity of bodily harm of the individuals involved in the associated collisions via a predictive model.

Audience

Individuals interested in understanding the factors that are associated to accidents and how data science could be leveraged in the future to mitigate the severity of injuries.

Aim of the Analysis

Explore 39 variables provided the City of Seattle that associate to roadway accidents and specifically the severity of the accidents. The ability to see outcomes will allow for us to review relevant variables to develop predictive models to determine the likely outcome of an accident.

Data Used and Overview of Features

2004 and 2020 accident data provided by the City of Seattle. The feature variables selected are based on environment conditional factors:

1. Weather Condition – illustrates the weather conditions at the site of the accident e.g. - rainy
2. Road Condition – illustrates the roadway conditions at the site of accident e.g. – wet roads
3. Light Condition – illustrates the lightening condition of the roadway – e.g. daylight
4. Road Junction type – illustrates the section of a roadway an accident occurred – e.g. intersection

Data Cleansing

The data was reviewed and null or NaN values to ensure models can act accordingly. Rows with invalid values were dropped to improve the general accuracy of the ML model outputs.

```
# Check for invalid values, such as Null or NaNs.
accidents[col].isnull().sum()
```

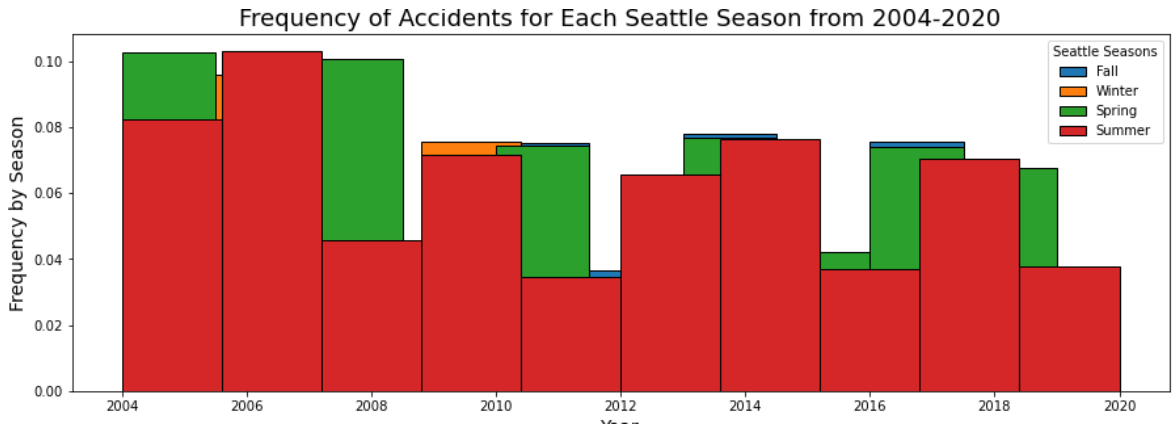
```
Out[80]: SEVERITYCODE       0
         INCDTTM            0
         PERSONCOUNT        0
         PEDCOUNT           0
         PEDCYLCOUNT        0
         VEHCOUNT           0
         WEATHER         5081
         ROADCOND        5012
         LIGHTCOND       5170
         JUNCTIONTYPE    6329
         dtype: int64
```

A review of the selected variables was done to give a sense of what conditional factors are most often present. Data transformation was also conducted to convert all object values into integers.

Report - Predicting Car Accident Severity in Seattle

Observations and Results

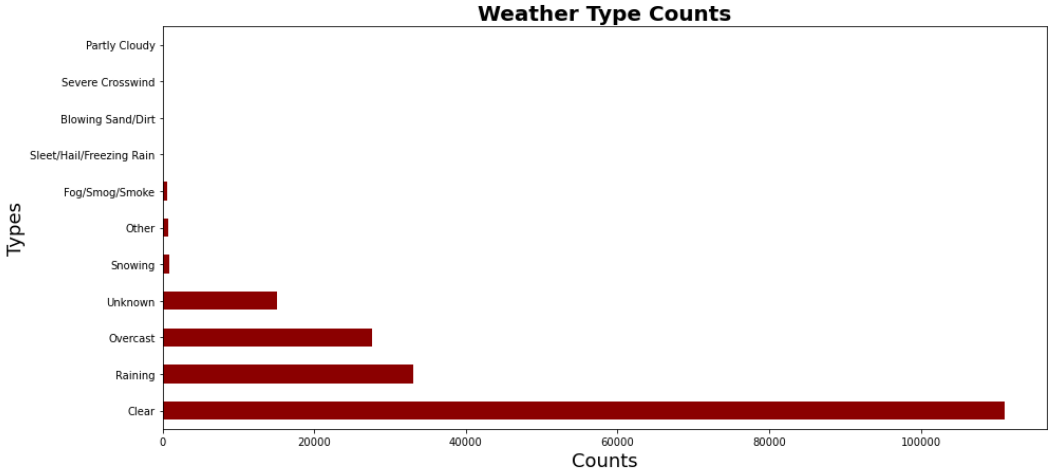
No direct finding was found between seasons and severe accidents outside of the fact summer months have a higher volume of accidents. This could be in part from greater volume of roadways based on seasonal travel that is most likely take place during warmer months such as tourism and agriculture.



Clear conditions are present for most accidents

```
weather = accidents.WEATHER.value_counts()
print(weather)
create_barh_plot(weather, 'Weather Type Counts', 'weather_counts', 'darkred')
```

```
Clear                     111008
Raining                    33117
Overcast                   27681
Unknown                    15039
Snowing                      901
Other                        824
Fog/Smog/Smoke               569
Sleet/Hail/Freezing Rain     113
Blowing Sand/Dirt             55
Severe Crosswind              25
Partly Cloudy                  5
Name: WEATHER, dtype: int64
```

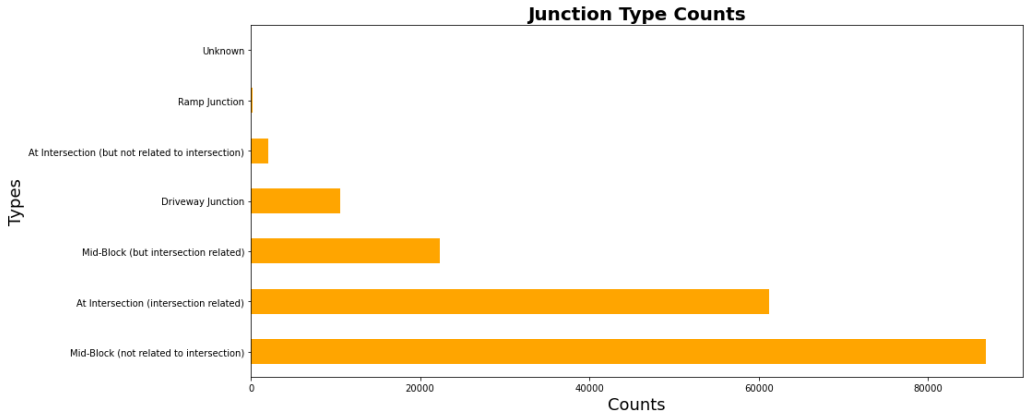


Mid-block accidents are common areas of accidents

Report - Predicting Car Accident Severity in Seattle

```
In [51]: JUNCTIONTYPE = accidents.JUNCTIONTYPE.value_counts()
         print(JUNCTIONTYPE)
         create_barh_plot(JUNCTIONTYPE, 'Junction Type Counts', 'Junction_counts', 'orange')
```

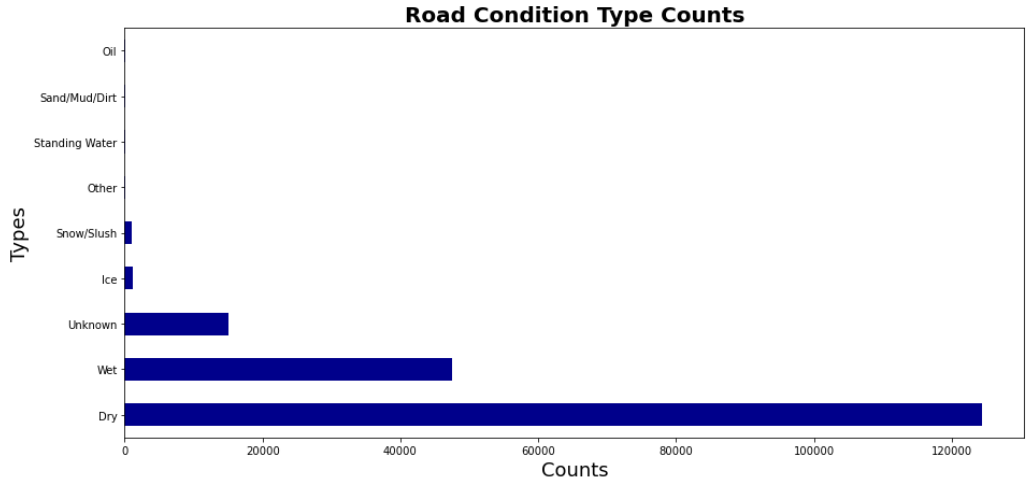
```
Mid-Block (not related to intersection)              86856
At Intersection (intersection related)               61241
Mid-Block (but intersection related)                 22353
Driveway Junction                                    10520
At Intersection (but not related to intersection)     2057
Ramp Junction                                          162
Unknown                                                  7
Name: JUNCTIONTYPE, dtype: int64
```



Dry conditions are present for most accidents

```
In [14]: roadcond = accidents.ROADCOND.value_counts()
         print(roadcond)
         create_barh_plot(roadcond, 'Road Condition Type Counts', 'roadcond_counts', 'darkblue')
```

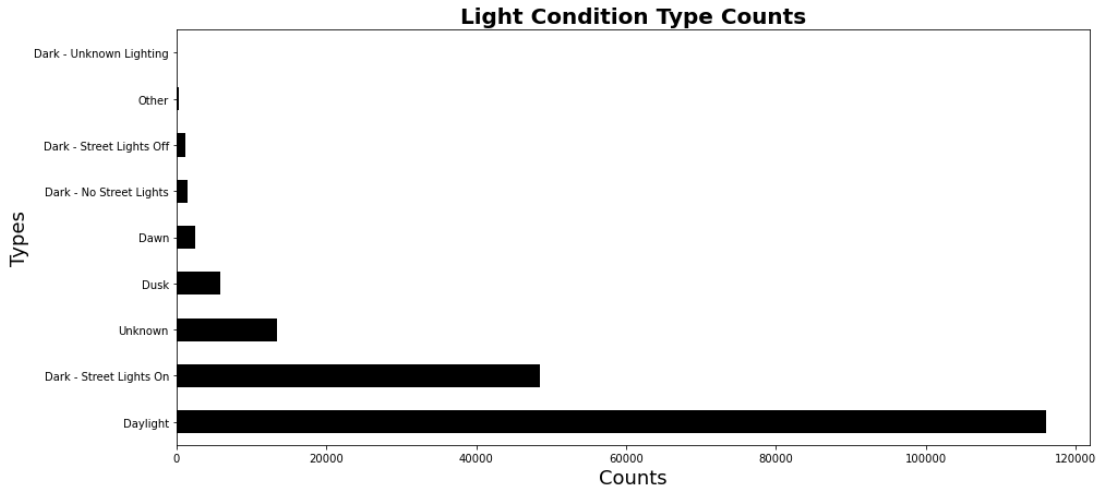
```
Dry              124300
Wet               47417
Unknown           15031
Ice                1206
Snow/Slush          999
Other               131
Standing Water      115
Sand/Mud/Dirt        74
Oil                  64
Name: ROADCOND, dtype: int64
```



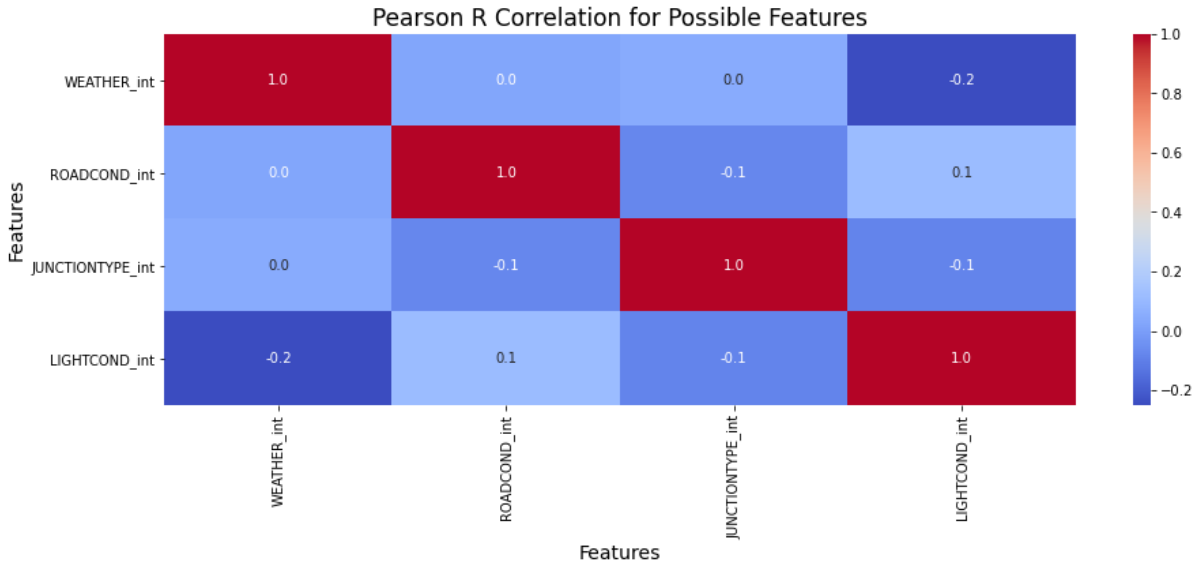
Roadways are generally well lit when accidents occur

```
lightcond = accidents.LIGHTCOND.value_counts()
print(lightcond)
create_barh_plot(lightcond, 'Light Condition Type Counts', 'lightcond_counts', 'black')
```

```
Daylight                    116077
Dark - Street Lights On      48440
Unknown                      13456
Dusk                          5889
Dawn                          2502
Dark - No Street Lights       1535
Dark - Street Lights Off      1192
Other                          235
Dark - Unknown Lighting         11
Name: LIGHTCOND, dtype: int64
```



After classifying variable outcomes by rank order into integers. Running a Pearson R Correction of the variables highlights a small to no strength of association amongst variables



Report - Predicting Car Accident Severity in Seattle

Model Results

- **Decision Tree Classifier** - A Decision Tree is a simple representation for classifying examples. It is a Supervised Machine Learning where the data is continuously split according to a certain parameter. The three parameters are nodes, edges, leaf of nodes. A regression tree was selected as target variable were continuous values.
- **K-Nearest Neighbor** - The k-nearest neighbors algorithm is a supervised classification algorithm. It takes labeled points and uses them to learn how to label other points. To label a new point, it looks at the labeled points closest to that new point which are its nearest neighbors, and has those neighbors vote. So whichever label, the most of the neighbors have is the label for the new point. Here “k” in K-Nearest Neighbors is the number of neighbors it checks. It is supervised because you are trying to classify a point based on the known classification of other points.
- **Logistic Regression-** A statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression(or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

Decision Tree

	precision	recall	f1-score	support
1	0.69	1.00	0.82	25330
2	0.35	0.00	0.01	11310
accuracy			0.69	36640
macro avg	0.52	0.50	0.41	36640
weighted avg	0.58	0.69	0.57	36640

K- Nearest Neighbor

	precision	recall	f1-score	support
1	0.70	0.83	0.76	25330
2	0.34	0.20	0.25	11310
accuracy			0.64	36640
macro avg	0.52	0.52	0.51	36640
weighted avg	0.59	0.64	0.60	36640

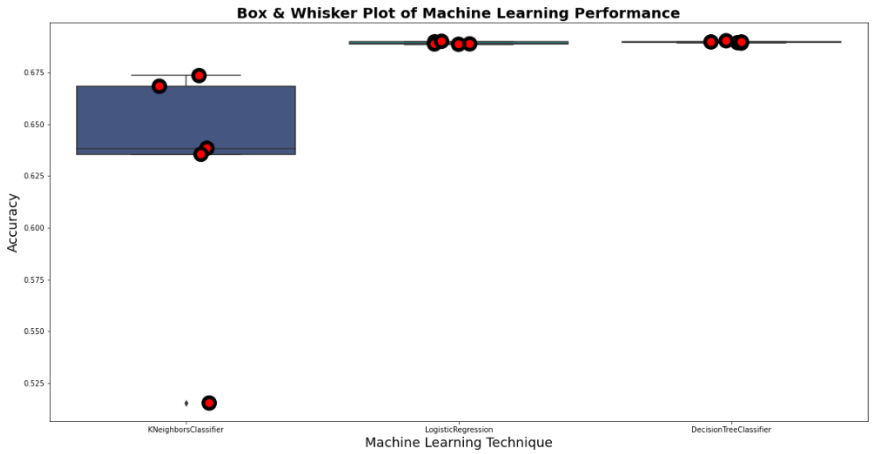
Logistic Regression

	precision	recall	f1-score	support
1	0.69	1.00	0.82	25330
2	0.33	0.00	0.01	11310
accuracy			0.69	36640
macro avg	0.51	0.50	0.41	36640
weighted avg	0.58	0.69	0.57	36640

- **Precision** quantifies the number of positive class predictions that actually belong to the positive class.
- **Recall** quantifies the number of positive class predictions made out of all positive examples in the dataset.
- **F-Measure** provides a single score that balances both the concerns of precision and recall in one number.

The Decision Tree and Logistic Regression score similarly, but the Decision Tree has a slight edge at with a F-1 Score of.567 vs .566

Report - Predicting Car Accident Severity in Seattle



The variability of K-nearest nearest is notable when compared with the Logistic and Decision Tree Models when reviewing via Box and Whisker Plot.

Average Accuracy by Model

Decision Tree Classifier 0.698437

Logistic Regression 0.697983

K neighbors Classifier 0.685582

Conclusion

Opportunities exist to improve the accuracy of the models. A few other areas of research could be:

1. Geographic location for “clusters” of accidents
2. Appending the congestion or traffic of a roadway
3. Review of the days of the week that correlate with accidents

The existing model however, could be a useful addition to real-time GPS direction applications on phones and within cars to notify a driver when the presence of adverse conditions are present to increase their caution and alertness while driving to mitigate being a victim of a severe accident.