Determining Factors which play a role in Accident Severity for the city of Seattle.

Introduction/Business Problem

The Seattle Police Department (SPD) has recorded all car collision accident from 2004 to present. Basing on those historical data (194,673 records), we can understand the high-risk areas, understand car injury factors to avoid accident, and plan a safe trip to Seattle in future. The reduction in severity of accidents can be beneficial to the whole society, including the Public Development Authority of Seattle which works towards improving those road factors, and the car drivers themselves who may take precaution to reduce the severity of accidents.

Data Understanding

Collisions data from 2004 to present. Those data are provided by the Traffic Records Group in the SDOT Traffic Management Division from Seattle, WA. It includes all collisions (194,673 records) provided by the Seattle Police Department and recorded by the Traffic Record, displayed at the intersection or mid-block of a segment from 2004 to the present. Each record has 38 variables/attributes which contains information such as severity code, address type, location, collision type, weather, road condition, speeding, among others. Of the 194,673 records, only 58,188 records are injury collision, so this is an unbalanced dataset for our project.

Data Cleaning and Preparation

Feature selection

As we want to analyze what factors will probably cause a car collision and the severity of the accident, we would drop those unrelated features and useless information, only keep 15 features/attributes of the original dataset. Those 23 features we dropped/deleted are: 'OBJECTID', 'INCKEY', 'COLDETKEY', ‘REPORTNO', 'STATUS', 'INTKEY', 'LOCATION', ‘EXCEPTRSNCODE', 'EXCEPTRSNDESC','PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', ‘VEHCOUNT', 'SDOT_COLCODE', ‘SDOT_COLDESC', 'PEDROWNOTGRNT', 'ST_COLCODE', ‘ST_COLDESC', 'SEGLANEKEY', 'CROSSWALKKEY', ‘HITPARKEDCAR’.

Handling missing values

- Convert "?" to NaN.

We replace "?" with NaN (Not a Number), which is Python's default missing value marker, for reasons of computational speed and convenience.

- Replace missing value by frequency

Whole columns should be dropped only if most entries in the column are empty. In our dataset, none of the columns are empty enough to drop entirely. Basing on the character of the data features we choose, we mainly replace the missing value with the most frequent values.

However, features like "INATTENTIONIND" and "SPEEDING" only have "Y" value, thus we replace the missing value in those two columns with "Y".

Correct data format

Two features ("INCDTTM" and "INCDATE") should NOT be object type, thus we change those two columns with "datetime64" type. And then use those data to calculate which hour ("hourofday") and which weekday ("dayofweek") those accidents happened.

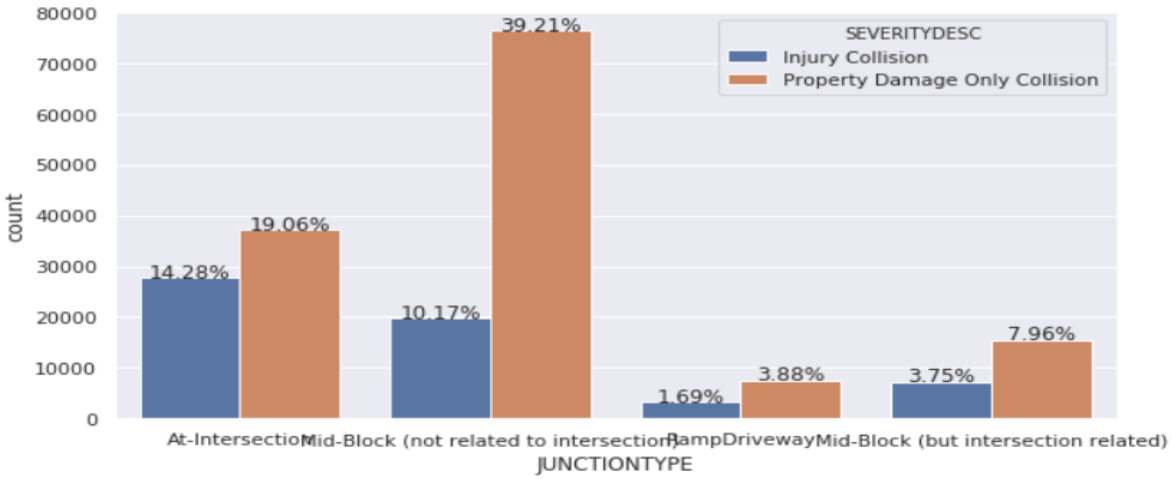
Delete/Drop some rows

We want to use location data to map the accident, so have to drop more than 5000 records who don't have X and Y data.

Data Visualization and Evaluation

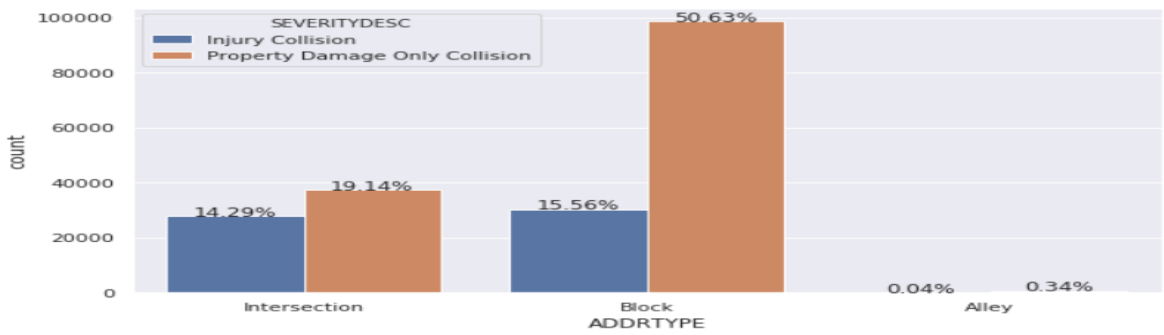
After binning the X and Y data, we are able to plot the following graphs and understand the implication of several features such as address type, road condition, light condition, etc. on the accident severity

1. Junction Type



From the plot above, we can see "At Intersection" is an important factor, where the accidents are more likely to involve injury, only a little less than the chances of property damage.

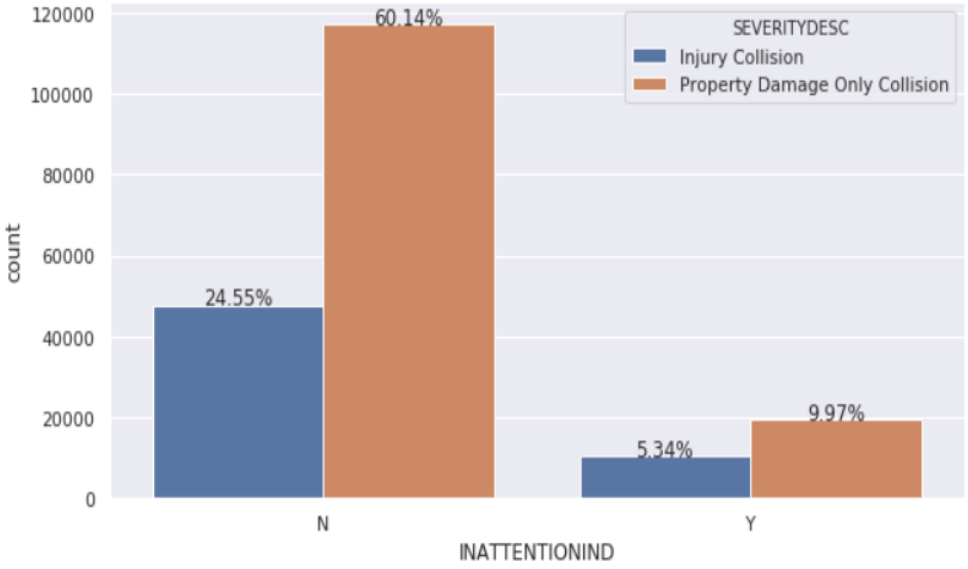
2. Address Type



Address Type data also show that Intersection and Block greatly influences the severity

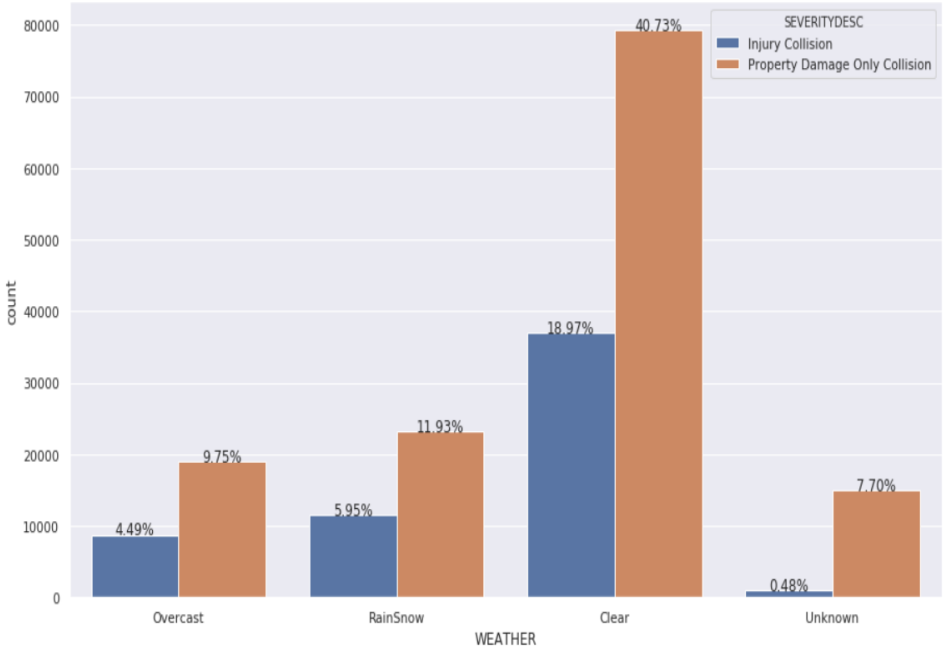
We can see that an accident which happened at Intersection or Block will be more likely to have people injured. Car collision occurred at Alley are also different than other two types, but data are not big enough.

3. Driver Attention



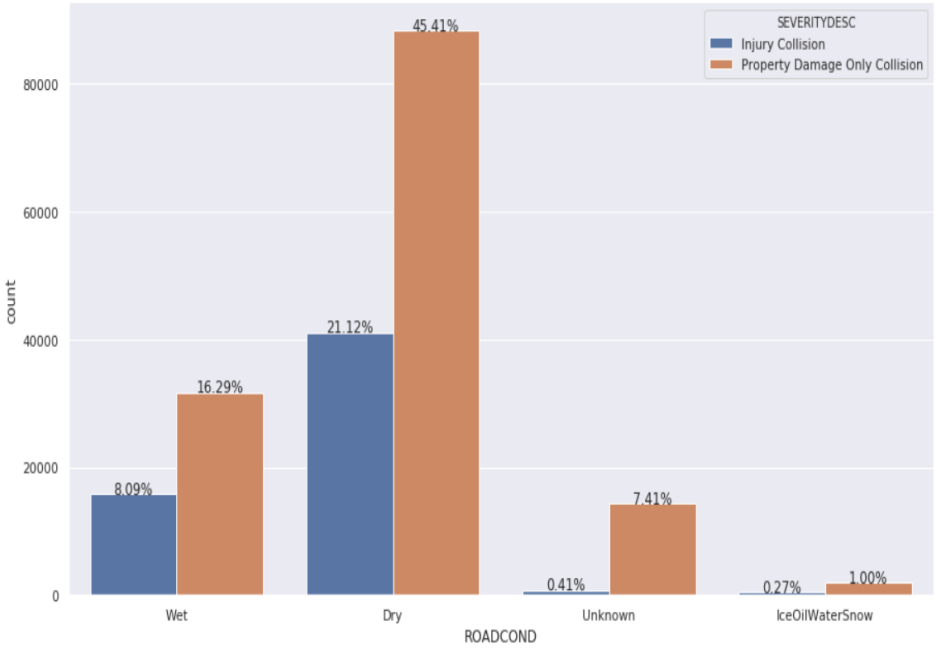
We see that drivers not paying attention results in a higher accident ratio where people are also injured.

4. Weather



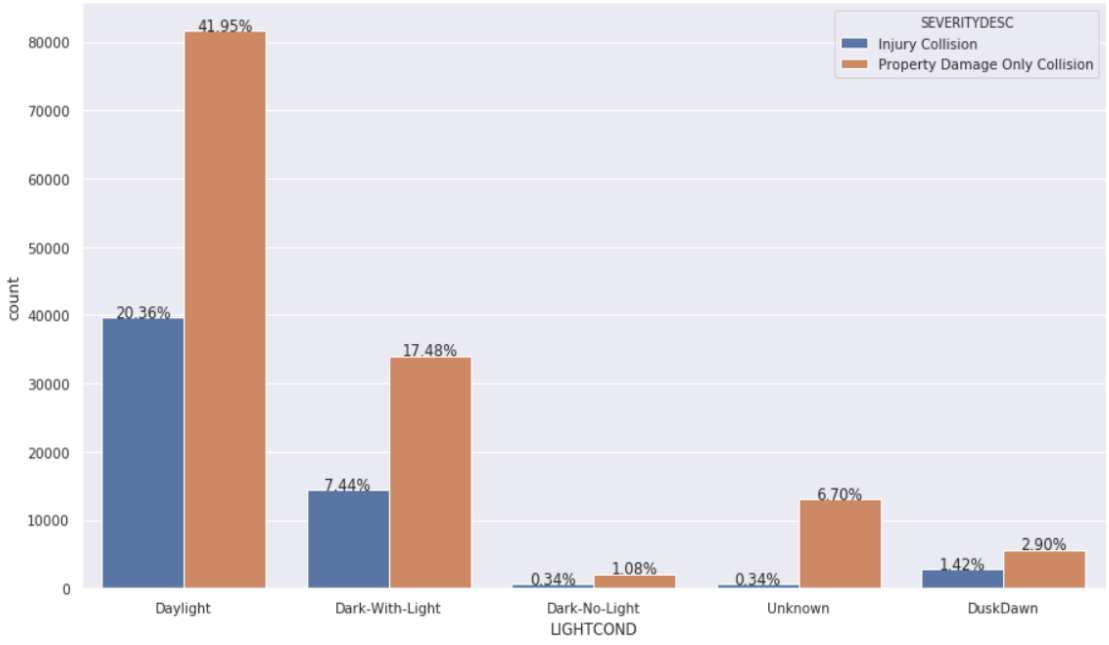
The graph shows us most accidents happen in clear weather.

5. Road Condition



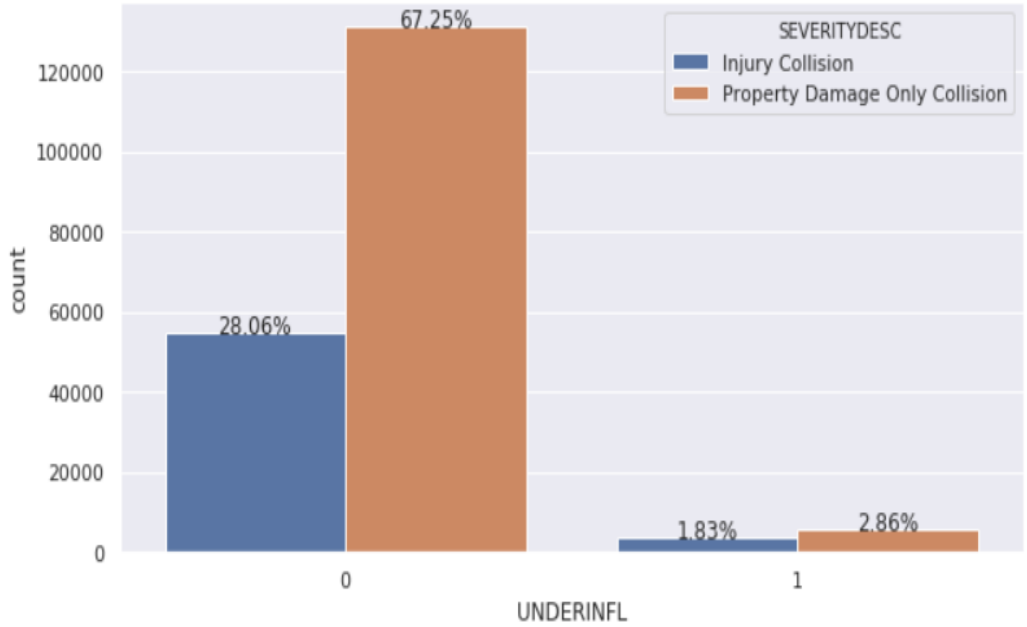
The plot shows us that a majority of accidents happen when the road is dry or wet as compared to the effect of ice, oil or snow.

6. Light Condition



The data shows us that majority of accidents happen in daylight.

7. Under Influence



The plot shows that driving under the influence of drugs or alcohol will greatly increase the probability of an accident.

Conclusion

We have analysed certain predominant factors such as Junction Type, Address Type, road, light and weather conditions, etc and have inferred the following.

1. Driving under the influence of intoxicants can greatly increase the possibility of an accident.
2. Driver inattentiveness is a major cause of accidents involving people and property.
3. Intersections are where the maximum number of accidents take place.