Introduction

The problem is Seattle has car accidents that can be prevented. Accidents happen at all times, but if the main causes of accidents are determined, advance warning or mitigating methods can be performed. For example, certain intersections may be more susceptible to accidents due to heavy usage or the way they are constructed. As a result, better street lights can be added (only protected left and right turns) or traffic personnel can be used to direct the cars. If it is determined that accidents occur the majority of a time a driver is speeding, has a high blood alcohol level, or was not paying attention, the data can be used as evidence for enacting harsher laws and regulations. In addition, the data can be advertised to the public to show them the consequences of driving under these conditions. This will hopefully dissuade people in the future. Finally, there are also uncontrollable factors such as weather and road conditions. If certain patterns are discovered to cause many accidents, local government can know when to send alerts to the public to drive more cautiously or even avoid the roads entirely.

The target audience of this analysis is the Seattle government and transportation department. It should identify key causes of accidents and allow them to identify trends for when accidents can be prevented. This will reduce the number of accidents and injuries for the city

Data

The data comes from collision and accident reports in Seattle during the years 2004-present. It was collected by the Seattle Police Department and Traffic Records department. The data will be used to identify the key variables that cause accidents. For example, the “WEATHER” column can be used to show the types and number of accidents that occur for different categories. In addition, the “INTKEY” column can be grouped and the sum of the accidents in that intersection can be calculated. This list can be sorted descending to identify the more dangerous intersections that need improvements or closer monitoring. Finally, a supervised learning model will be used to come up with a formula that can predict the severity of an accident based on the inputs. The data has 37 independent variables and 194,673 records. The dependent variable, “SEVERITYCODE”, has numbers that correspond to different levels of severity caused by the accident. Many of the columns are object types. In addition, other columns that appear to be integer types are also actually objects, because the numbers correspond to different categories. Finally, some columns and rows have null values, which will be dealt with during the data pre-processing phase.

Methodology

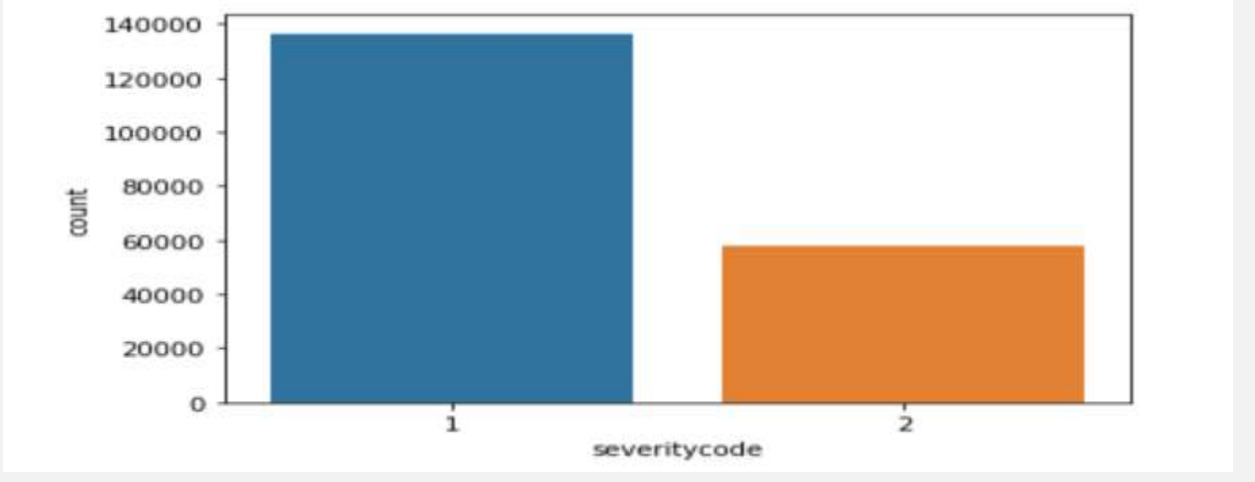
Exploratory data analysis was performed on the relevant categorical variables: address type, collision type, person count, pedestrian count, cyclist count, vehicle count, junction type, SDOT type, under the influence, weather, road conditions, light conditions, lane key, crosswalk key, and if a parked car was hit. The amount of categories in each variable ranged from a few to over a dozen. As a result, categories with less than a dozen variables were turned into countplots to better visualize the data. Statistical testing was not performed because the data revolved around categorical variables, not numerical ones.

Unfortunately, key variables such as pedestrian right of way, inattentive drivers, and whether the car was speeding had a majority of null values. Therefore, they were dropped and not part of the analysis. However, it is likely that these variables play a key factor in vehicle accidents.

Results

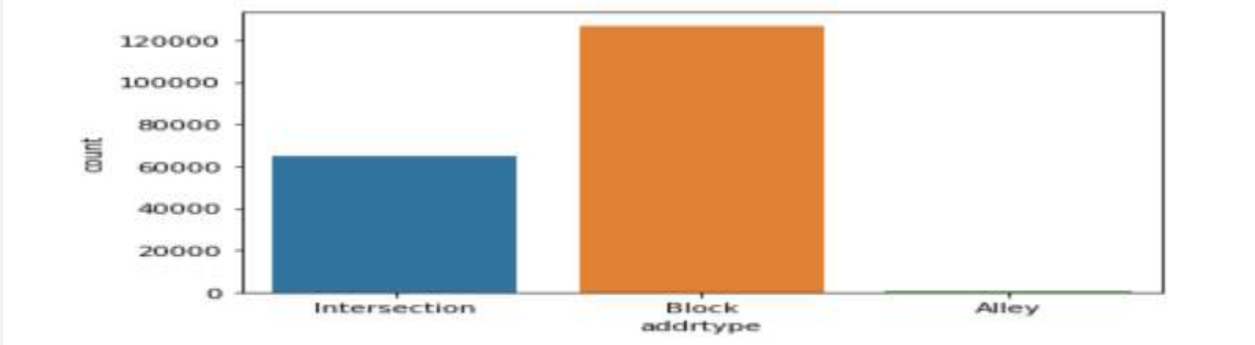
Interestingly, the data shows that most accidents occur during the day with normal drivers and conditions.

To start with, here is the count plot for the dependent variable, severity code:

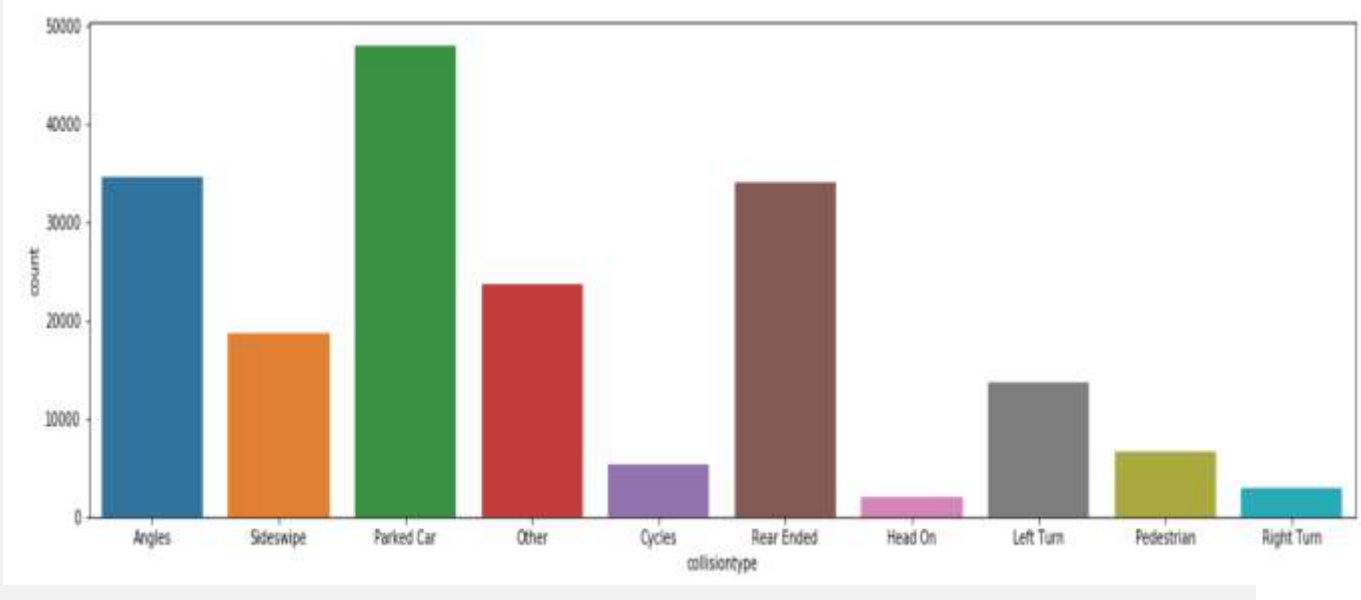


Most of the cases involved property damage.

The first variable involves where accidents took place:

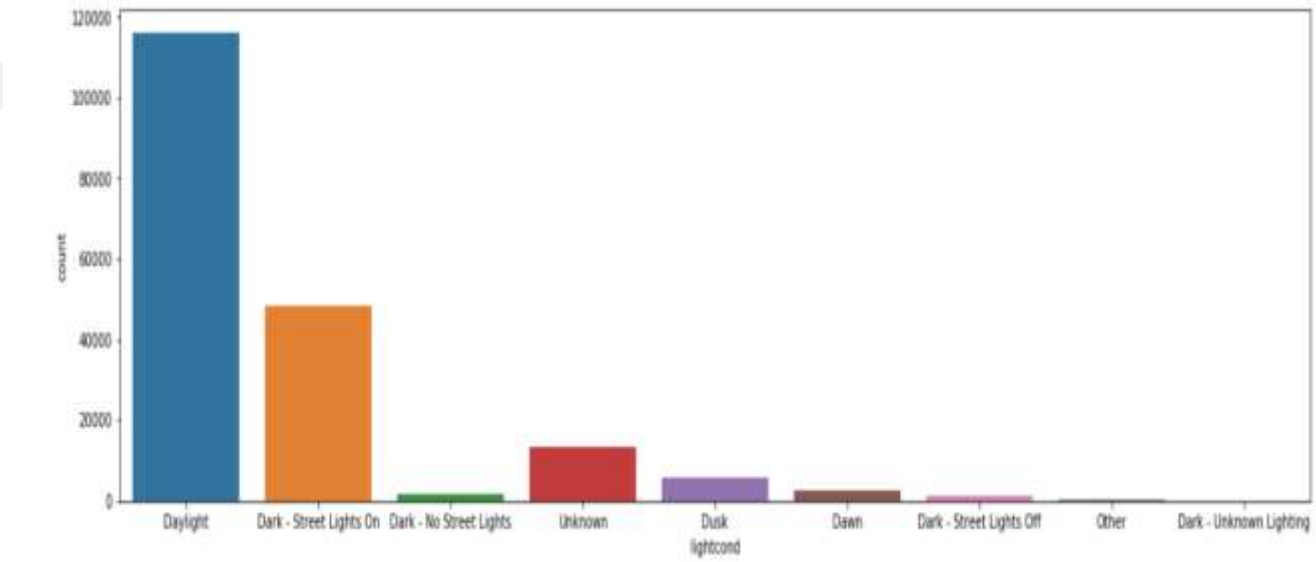


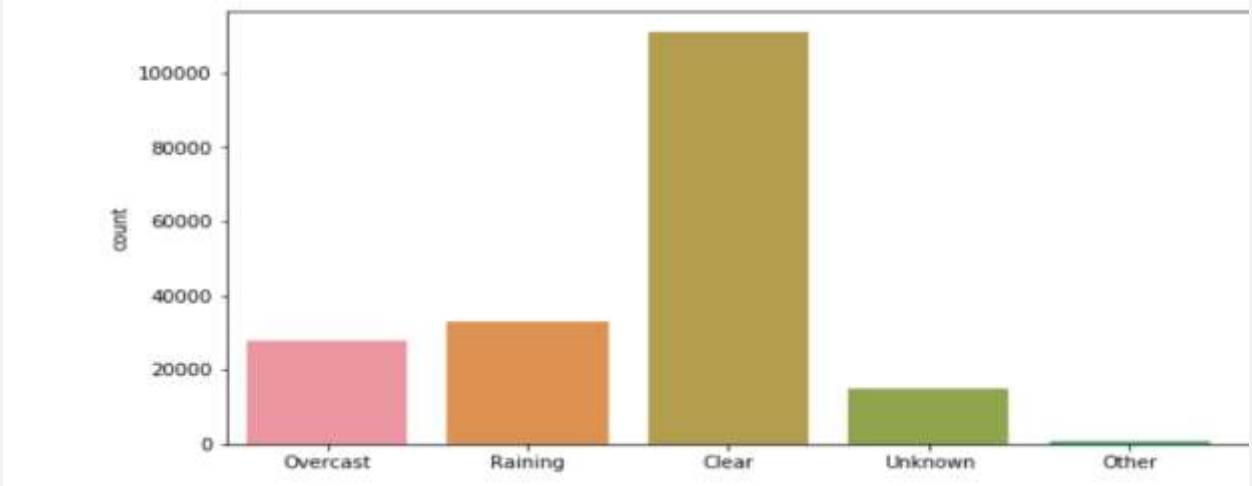
As expected, most occur at a block or intersection.

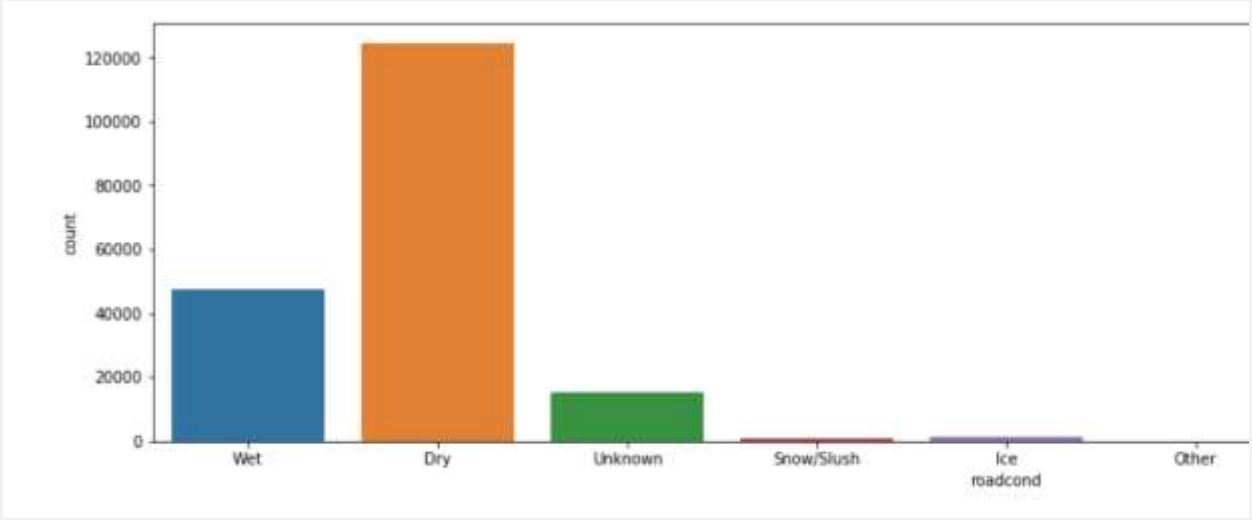


The above chart shows the types of collisions. The highest category involves park cars, which aligns with what we already know. The majority of cases involves property damage, and this explains why. Hitting a parked car usually does not involve an injury.

Next, analysis was done on certain uncontrollable factors such as weather and road conditions:







Most vehicle accidents occur during the best driving times when it is clear, during the day, and the roads are dry.

Discussion

The first striking observation has to do with the dependent variable. It seems highly unlikely that over the date range of the data no serious injuries or fatalities occurred. This may be a warning sign that the severity codes were somehow altered when

the data set was being created, or that the sample data is incomplete and missing those reports.

The main recommendation has to do with key variables such as pedestrian right of way, inattentive drivers, and if the car was speeding. In many of the records, these values were null. However, this data should be collected in order to draw new insights or create better prediction models.

It was determined that most accidents occur during normal weather and road conditions. However, further data is needed to analyze this trend. It may be that these types of days constitute the highest number of days in the year. Therefore, further data on the weather needs to be analyzed. For example, it may be sunny for 100 days and then snow on 1 day. If you look at data for accidents, there may be 1000 accidents occur during sunny days and only 20 on snowy days. Really, the average number of accidents per weather type is much higher on snowy days. Though, because there are few days like that during the year the total number of accidents appears low.

Conclusion

The data showed that most vehicle accidents occur during good conditions with normal drivers. This means it will be harder for the Seattle transportation department to mitigate accidents. However, as most accidents only involve property damage or minor injuries, there is not a serious problem that needs to be dealt with right away. This shows that infrastructures are being designed and operating properly. Therefore, the focus should be an emphasis on drivers being more careful.