**Collisions in Seattle**

INFO W18 Summer 2019

Project 2 Analysis Report

August 16th, 2019

Daniel Martin, Ewan Seo, Julie Maeng

**Introduction**

Traffic collisions are frequent in urban areas. Do road conditions, weather, or popular events affect the occurrence of such accidents? Our project seeks to gain insights on what factors may be relevant to traffic collisions. For this purpose, our project will focus on analyzing a large data source consisting of information about historic accidents occurred in Seattle.

**Questions**

Our analysis will focus on the following questions:

1. What road conditions result in the greatest amount of car collisions?
2. What weather conditions result in the greatest amount of car collisions?
3. What weather conditions result in the greatest amount of car collisions as a result of drivers speeding?
4. What road conditions result in the greatest amount of car collisions as a result of drivers speeding?
5. Is there a correlation between drivers under the influence and federal holidays?
6. What percentage of collisions are caused by drivers under the influence on non-holidays?
7. Is there a correlation between collisions and weather conditions, road conditions, and time of day?

**Data Files**

* [collisions.csv](https://catalog.data.gov/dataset/collisions)

This dataset is the main source used in our analysis. It covers all traffic collisions in Seattle from 2004 to 2017, with approximately 210,000 data entries. The source includes extensive documentation of factors in terms of the collisions, such as collision type, incident date, weather condition, street location, number of people injured, and more.

* usaholidays.csv

This data provides information about historical dates of federal holidays. The dataset covers the dates of holidays from 1966 to 2019, summing up to a total of 485 entries.

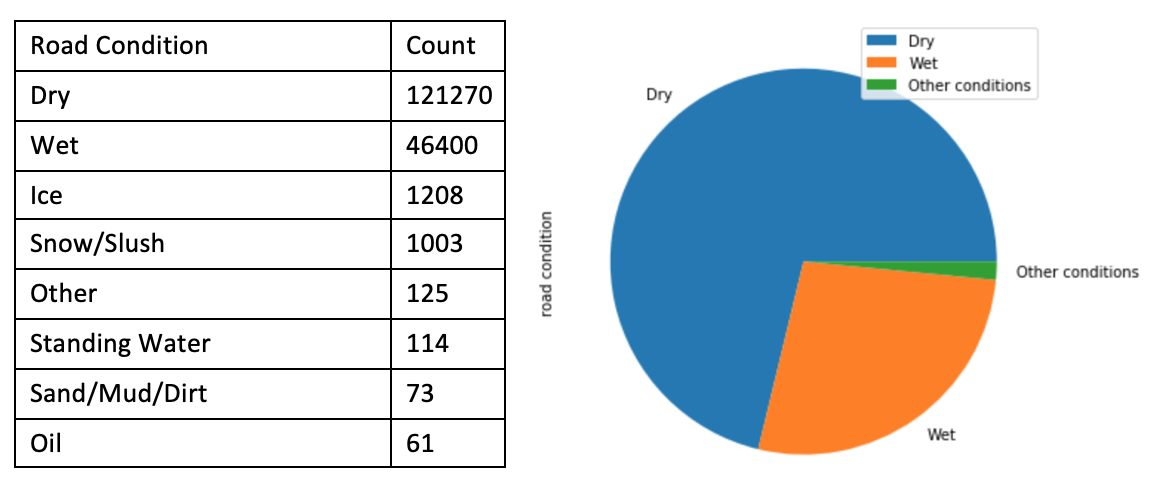
**Data Optimization**

As with any other dataset, we took several steps to optimize our access to the information. First, since our analysis involves comparing dates to collision types, the dates in collisions.csv had to be converted into usable format. The dates in collisions.csv are recorded in the format of YYYY-MM-DD with additional values. For instance, the entry for March 22nd in 2012 is recorded as 2012-03-22T00:00:00.000Z. In order to match the year-month-date format used in usaholidays.csv, we utilized a function to keep the first 10 characters that encompass the date as the same format.

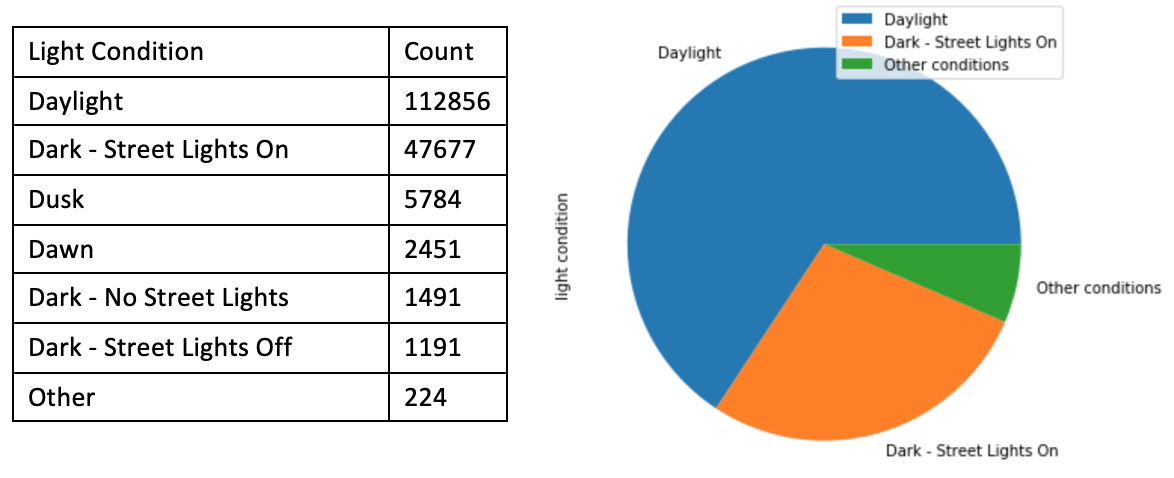
In addition, we organized individual variables in the dataset into data frames in order to understand the frequencies relating to collisions. We noticed that collisions.csv contains numerous data entries with missing values. By using a filtering function for each variable, we filtered out data that were unknown or “empty”. We utilized various functions to add flexibility and modularity to our analysis, as the functions could be used to selectively filter out rows for analysis of specific variables, keeping as many valid data points in the dataset per individual trial. With this method we proceeded to count the frequency for the variables of our choice named “under influence”, “road conditions”, “light conditions”, “weather”, “in attention”, and “speeding”.

Next, we applied the recorded federal holiday dates from usholidays.csv to investigate whether or not an incident occurred on a federal holiday or not. This opened a new facet of analysis for this dataset, to compare the collision variables to whether or not it was a holiday at the time.

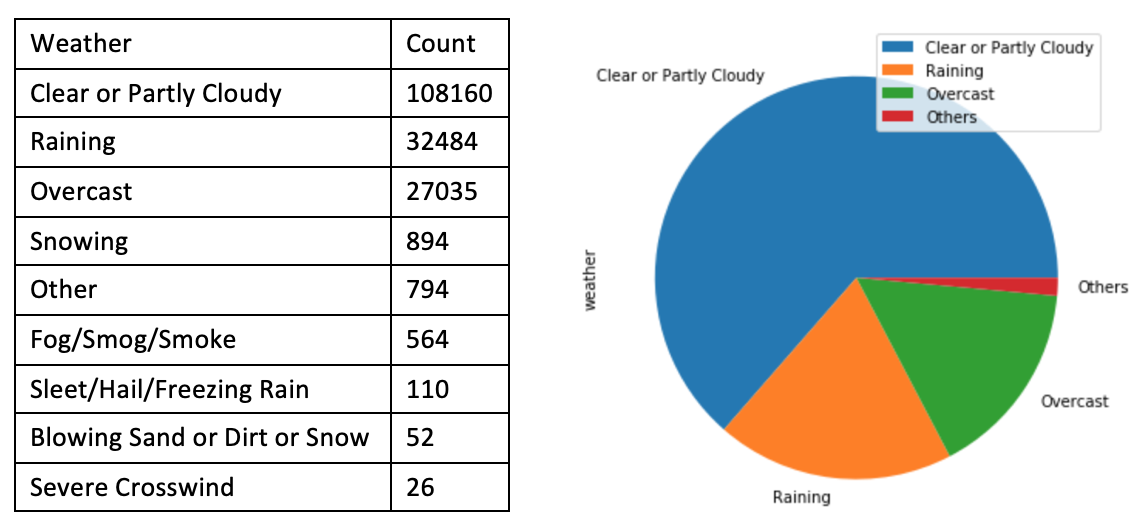
**Plots**

With the optimized data, we proceeded to perform analysis that investigates the questions asked above. First, we will take a look at certain conditions that were present during traffic collisions. In order to investigate the relationship between various environmental conditions and traffic collisions, we take a look at how many times any type of collisions in Seattle occurred under the respective conditions.

*Table and Pie Chart Representation of Road Conditions during Collisions*

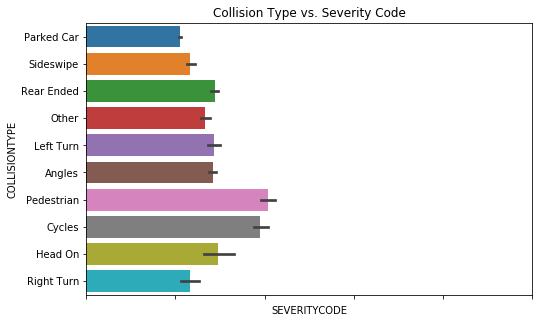


*Table and Pie Chart Representation of Light Conditions during Collisions*

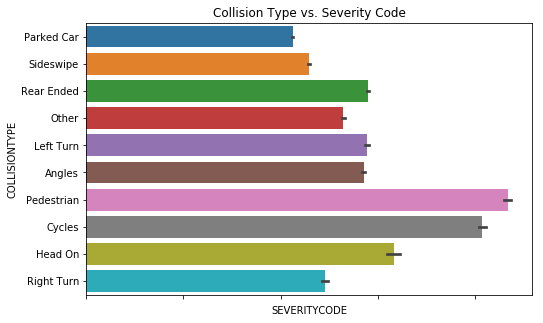


*Table and Pie Chart Representation of Weather during Collisions*

From the tables and graphs above, we can observe that collisions occurred most on dry roads than those in any other conditions, during daytime than at other times of day, and on clear or partly cloudy days compared to other weather conditions. We can conclude from the dataset that traffic collisions in general occurred most under these conditions. Next, we look at how these conditions are related to collisions caused by speeding.

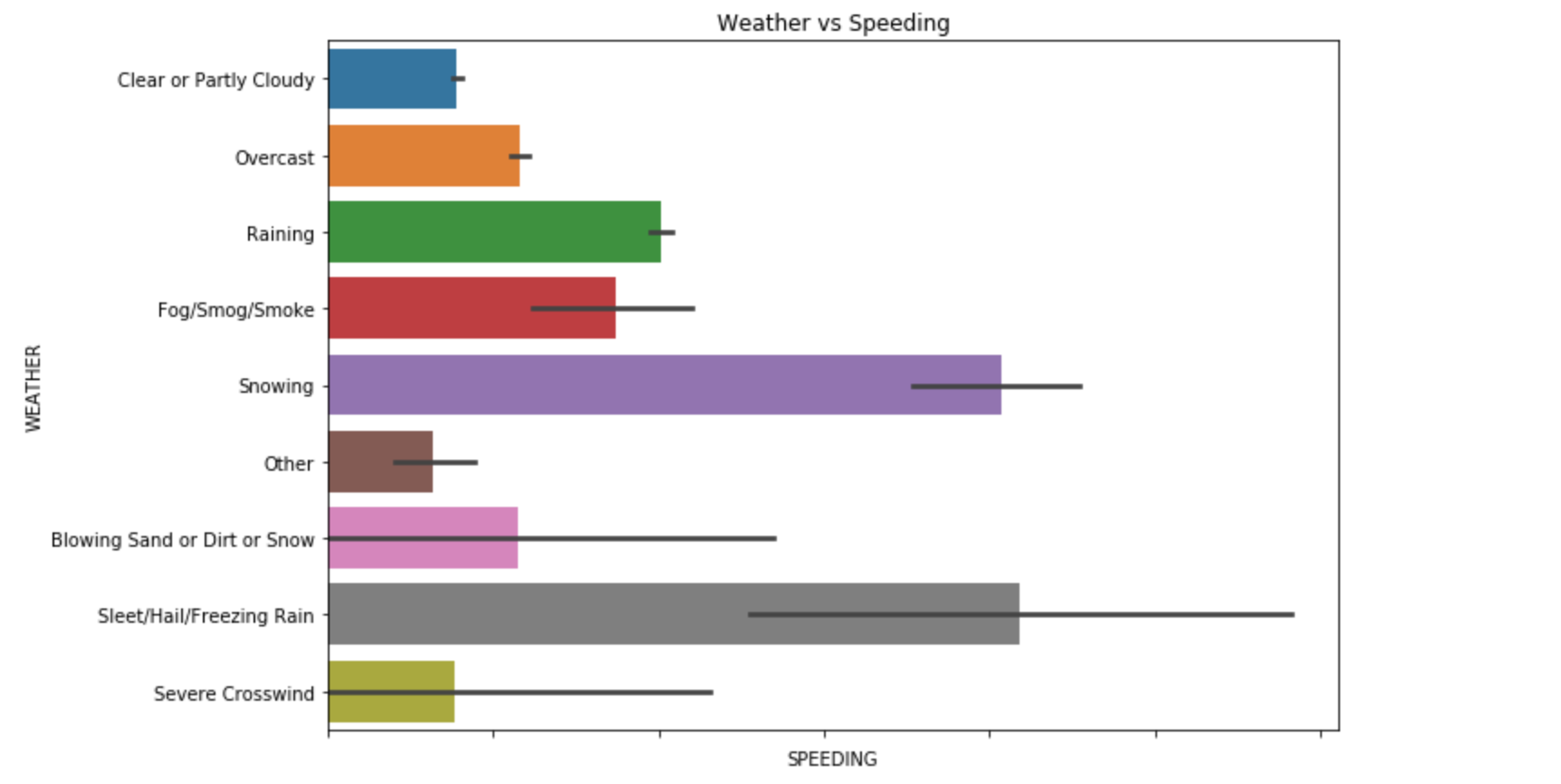


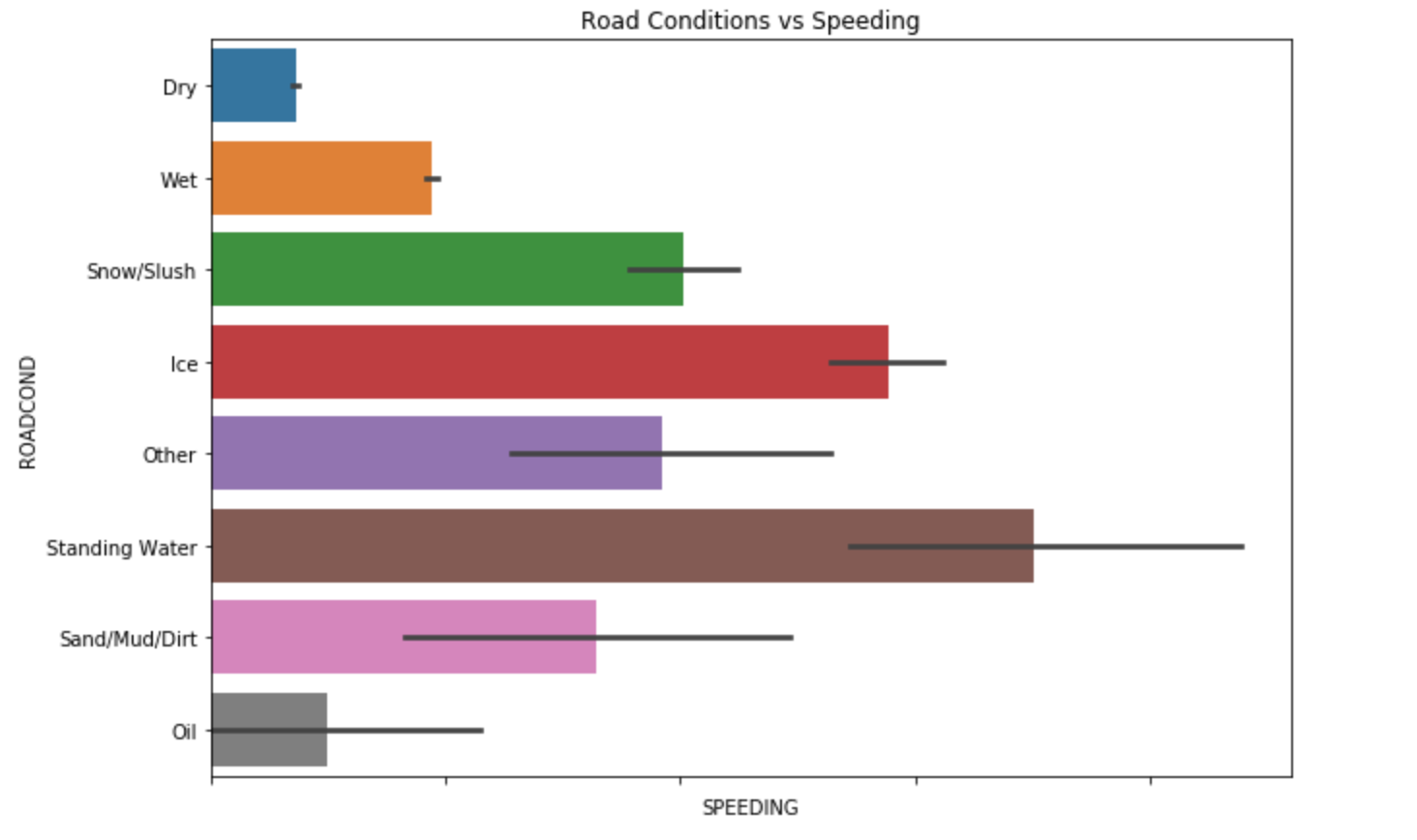
**Barplot Representation of Severity Code vs. Collision Type**



**Barplot Representation of Severity Code vs. Collision Type**

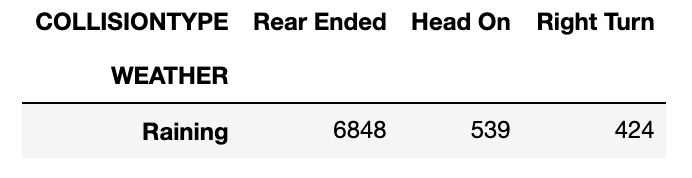
Utilizing our function to delineate between holiday dates and non-holiday dates, we created some plots to compare severity levels of accidents between holidays and all days, including holidays. We did this by summing severity codes by accident type, 0 being negligible, 1 being property damage, 2 being injuries, and 3 being fatalities. We found that the severity of these situations were pretty consistent whether it was a holiday or not, in terms of scale. This data gave insight into the heightened dangers involved in pedestrian and cycle-involved accidents, which intuitively made sense.

***Barplot Representation of Weather Collisions result of Speeding***

**

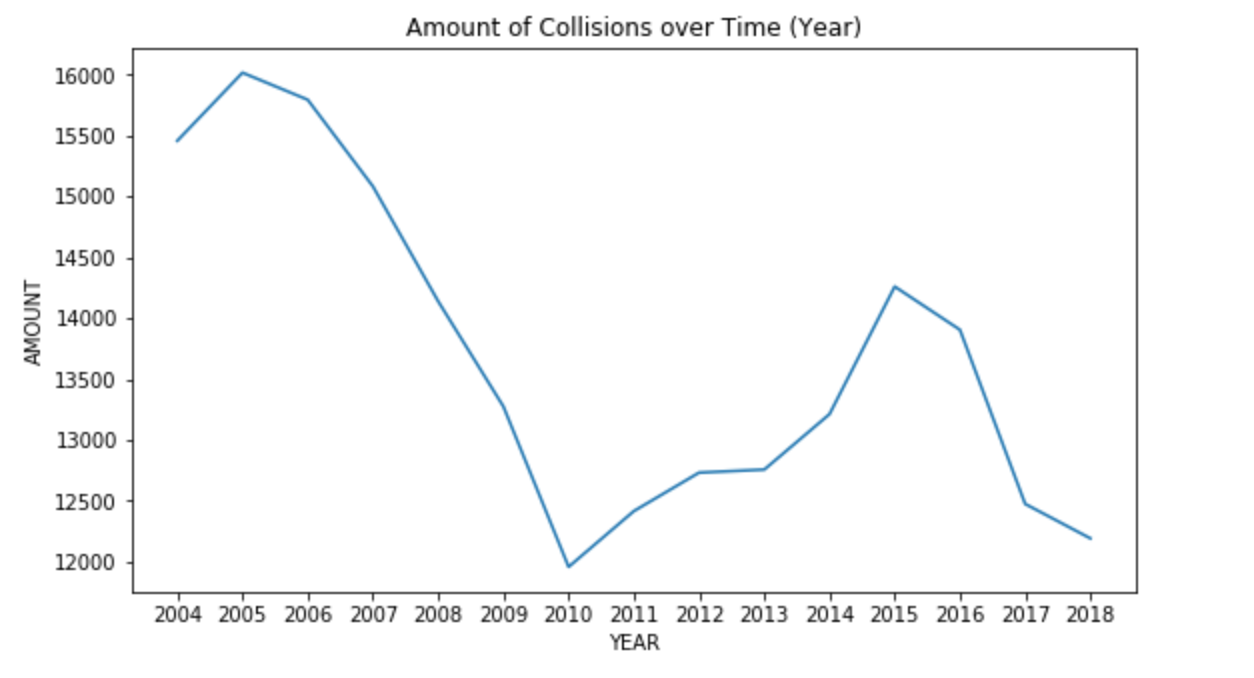
***Barplot Representation of Road Condition Collisions result of Speeding***

The information displayed from the pie charts above shows that collisions occur most in *Clear or Partly Cloudy* weather conditions. From the barplots above, we see that speeding results in the greatest number of collisions when weather and road conditions are either *Raining or Snowing.* This tells us that the collisions that occured during *Clear and Partly Cloudy* weather is not do to drivers speeding but do to a different cause. The bar plots also act as a sanity check. One would assume that most collisions that occur in *Raining or Snowing* conditions is a result of speeding because the road conditions do not offer drivers enough time to break.



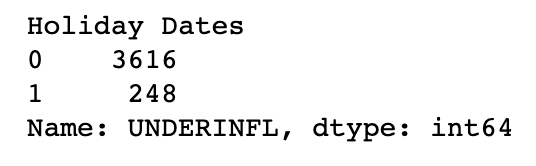
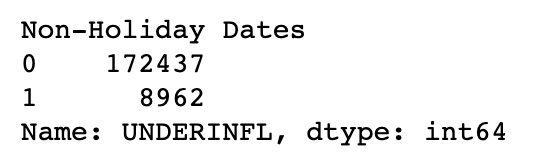
*DataFrame Representation of Collision Type when Raining Weather*

Out of the major types of car collisions (*Rear Ended, Head On, Right Turn) Rear Ended* car collisions make up approximately **87%** of these collisions. This reiterates the assumption that when weather conditions are *Raining* most collisions are a result of the drivers speeding and not being able to stop their car in time.



***Lineplot Representation of Car Collisions over Time***

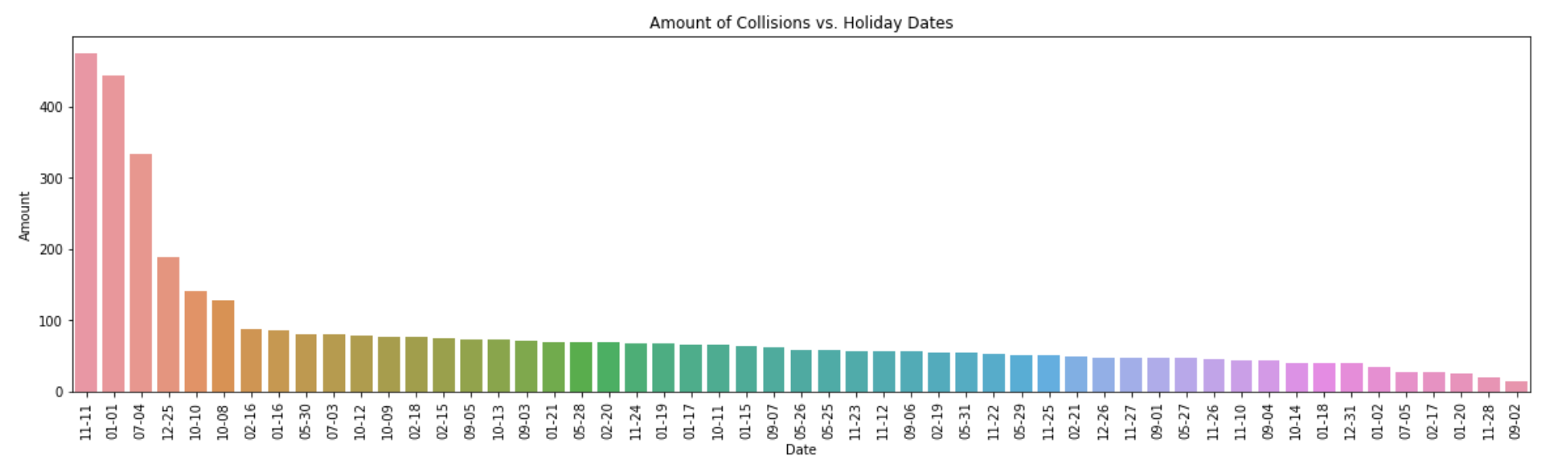
The line plot above shows the amount of car collisions that have occurred from the year 2004 to 2018. I excluded years 2003 and 2019 as the data in those years was not complete and did not represent an entire years worth of collisions. As we can see the highest point occurs in 2005 with approximately 16,000 collisions. The minimum occurs in 2010 with roughly 12,000 collisions. However, it is interesting how there is a gradual decline in collisions starting in 2007 to 2010. This could be a result of better infrastructure for public transportation or a result of increased unemployment due to the Great Recession. These scenarios would limit the amount of individuals driving on Seattle roadways.



***Collisions do to Drivers Under the Influence***

The above representation shows the amounts of collisions that were do to drivers under the influence of either drugs or alcohol. The table to the left represents collisions that occurred on non-holidays, while the table to the right represents collisions that occurred on holidays.

Drivers under the influence made up approximately **4%** of all total collisions on non-holidays spanning from 2003 to 2019. On the other hand, drivers under the influence made up approximately **6%** of all total collisions on holidays. This re-enforces the common thinking that there are more drunk drivers on the roads when the day is a holiday.



***Barchart Representation of Collisions on Holiday Dates***

The barchart above displays the total amount of collisions that occurred on holiday dates. As we can see **Veteran’s Day** (11-11) contains the highest amount of collisions at roughly 450 followed by New Years (01-01) and Fourth of July. The least amount of collisions occurs on Labor Day (09-02).