Hye Lim Lee – Group Assignment

1. Objective

   Goal: "Reducing the ratio of bike collisions among bike sharing riders in Toronto"

   The following procedure was used to analyze data to achieve the goal.

   - Measure feature importance for each column according to the bike collisions
   - Analyze 3 major relevance for the collisions
     - ✓ Find causes
     - ✓ Suggest solutions
   - Find the data that appears in both data set (Bike sharing and Bike collision)
   - Suggest solutions

2. Data Preparation

   Open data was used which are "Bikeshare Ridership (2017)" and "Toronto Police Cyclists Data"

   Data quality is relatively reliable since the data was gained from the official city of Toronto website and Toronto police website.

   To clean and organize the data, I had to score the data columns, combine the data and find sharing topics among data sets.

3. Analysis

To clean the data, I dropped 23 out of 57 columns which has more than 50 percentage of data missing and dropped another 6 columns that don't have unique values more than 5. Since X and Y have the same data as LATITUDE and LONGITUDE, those additional 2 columns were dropped. FATAL_NO, ACCLOC, TIME and INJURY were not used as well since similar information exists. ACC, Index_, ACCNUM, HOOD_ID and ObjectId columns were also deleted because of noninformation. Therefore, the total features that are useful in this analysis are 22 columns from "Toronto Police Cyclists Data" file. However, only 5 columns out of 22 meaningful columns were used to get the final result with "Bikeshare Ridership (2017)" file.

Among the meaningful data features, numerical values and classified values are separately analyzed. First, the "YEAR" column from the "Toronto Police Cyclists Data" file, 2017 year's information takes only 9% which is the same year of data from "Bikeshare Ridership (2017)". Therefore, it is reasonable to consider other years' information to analyze the data and get the result.

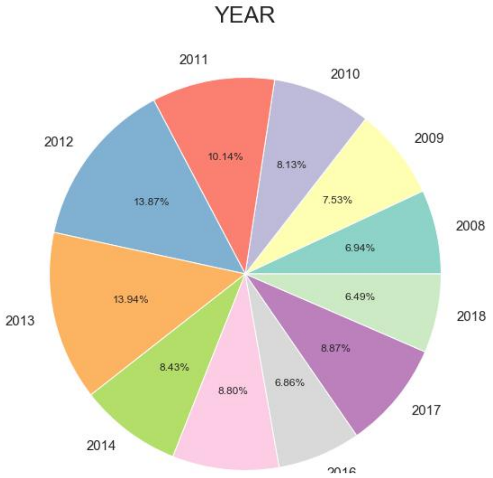

Figure 1

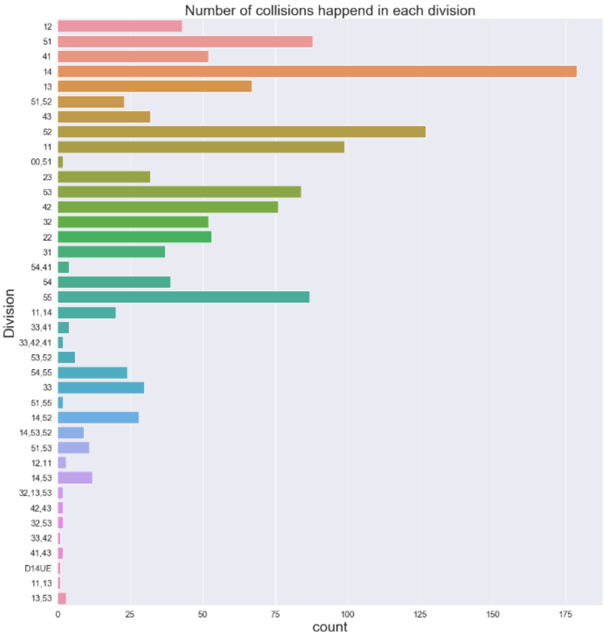

Figure 2

As figure 2 shows, Division 14 has the highest number of collisions and Division 52 and 11 followed.

An interesting note is that the Divisions start with the number 5 which are 51, 52, 53, and 55 have relatively higher collisions than other Divisions and the district for these 4 Divisions is Toronto and East York.

Since Division 14 is in the Toronto and East York district as well, we can reasonably assume that the stations in this district for the bike shares need to be extra careful. The stations in this district are Broadview, Chester, Pape, Donlands, Greenwood, Coxwell, Woodbine, Main Street, and Victoria Park Station.

Figure 3 represents the age range that was involved in the bike collisions. As shown, except unknown, age between 50 to 54 were involved in accidents the most. Furthermore, the age in 20's and 30's has high number of collision as well. As Figure 4 shows, the collisions happened the most often during Daylight. From the Hour column, the portion of number of collision during day time are also higher since the day time is longer than night time.

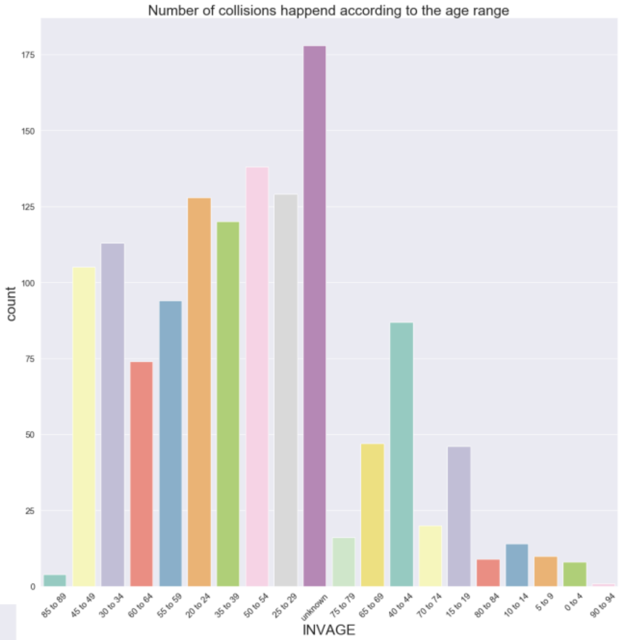

Figure 3

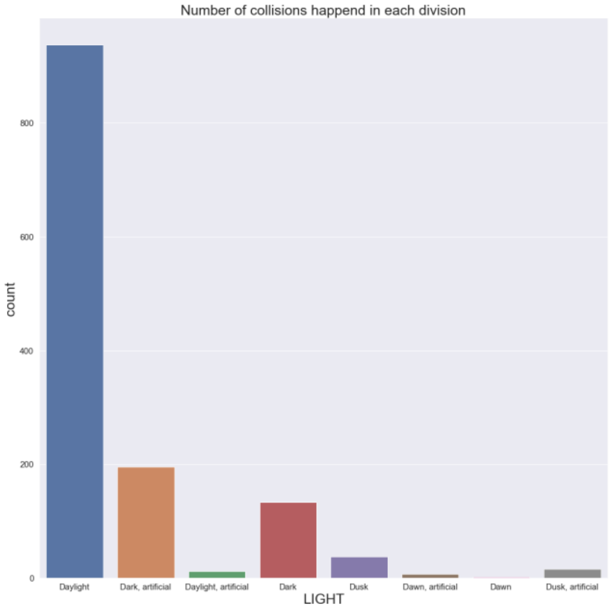

Figure 4

Figure 5 represents the number of collisions that happened according to the type of Manoeuver. From this data, the collisions happened the most often when the cyclist go straight. From this data, cyclists need to be cautious the object around them even when they just go straight.

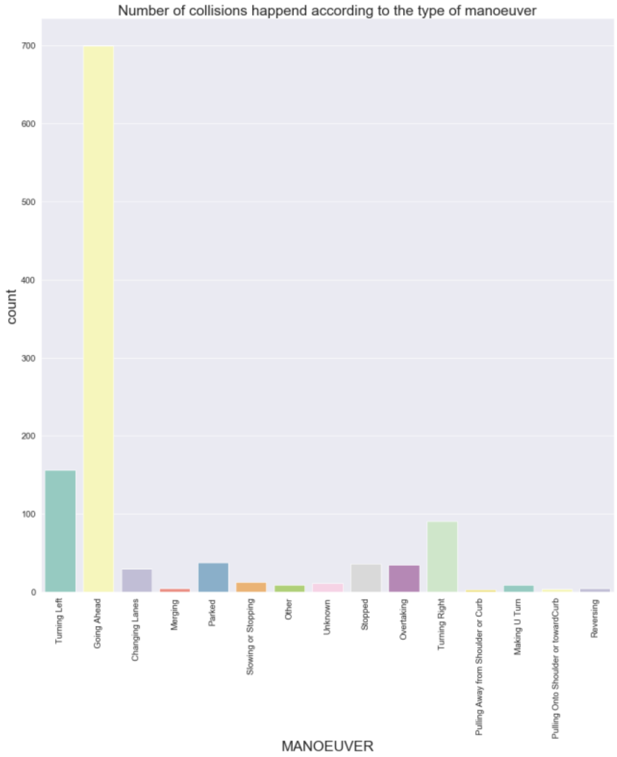

Figure 5

4. Conclusion

Since there is not enough data for bike sharing, only station data was able to be used with district data from "Toronto Police Cyclists Data". It was analyzed that the bike sharing riders need to be careful when they ride in the Toronto and East York district. Furthermore, people who are in the age range of 50 to 54 need to be extra careful. Not only riders need to be cautious when they turn left, but also they need to be more cautious when they just go straight during the daytime.