Data Scientists: Bingqi Zhou, Yuexiang Zhang, Yuxiao Ran

Dataset: Traffic and Collisions

20 Apr 2019

Rankings of Streets on Collision Probability

Our team mainly focus on which street has the highest probability of collision. We calculate the collision probability by dividing the total number of collisions happened on the street by the total traffic counts on the street.

When cleaning the data, we noticed that street names in the collision table doesn't match those in the traffic counts table. We have to extract the street type out of the street names and convert the abbreviations to full names. For example, we need to convert "DR" to "DRIVE." Therefore, Bingqi created a dictionary with abbreviations as keys and full names as values. Then we can apply the dictionary to the abbreviations to get the full names of the street types. Finally we match rows in both tables according to their street names and their street types.

Another obstacle we meet when cleaning data is how to group the rows on the street names. Noticing that the "groupby" method of pandas doesn't create a column counting the total number of times collisions occur on the street, Yuexiang decided to add another column with 1 as the value of each row. Then, when we group the street names, we got the total counts of collisions.

After we finished the data cleansing, we divided the number of collisions occurred on each street with its corresponding traffic counts to get the ratio of collisions occurred. To see which street has the highest probability of collisions. The result appears to be the "VIKING

WY.” Though it only has 2 collisions, the relatively small number of total traffic counts makes it stand out.