Wrangling Data for Capstone Project #1

Intro

Choosing a topic for the project seemed the most daunting part until you realize which data you want to work with after choosing your topic. Indeed, I was lucky enough to find the data I was looking for both from Kaggle and from the FCC site, where the former originated from. However, the problem I faced was choosing between the two: the former is much cleaner and thus the data table contains less missing values, while the latter provides much more data but has more missing values. After some deciding, I have decided to go with FCC's data since it not only contains all of Kaggle's, but also the date range is much wider and it has proportionally about as much missing values as does Kaggle.

Cleaning the Data

For this step I extracted the data I see necessary for this project. The dataset contained the following columns (columns without a missing value are **bold**):

- **Ticket ID**
- **Ticket Created (date and time)**
- Date of Issue (date only, may contain null values)
- Time of Issue (time only, may contain null values)
- **Form (Phone, TV, Internet, etc.)**
- Method
- Issue
- Caller ID Number
- Type of Call or Message
- Advertiser Business Number
- City, State, Zip columns
- Location (Center Point of the Zip Code)

Since the project focuses on robocalls, we only need to focus on data regarding phone issues. Next, I filtered out the rows that contained no location whatsoever (missing values for all of Caller ID Number, City, State, Zip). As there were only a small fraction of them, doing this would negligible effect to the overall data.

Next, I noticed many of the dates and times of issue had missing values, so I filled those with the date and time from the column “Ticket Created”. The challenging part comes from the “Time of Issue” column, which contains data that do not follow a particular string format. Using the filled-in data, I combined the time and date columns into a new column, “Issue Occurred.”

Finally, since the values in the data are mainly text-based and time-based, it would be impractical to determine if there are outliers; thus, it is safe to say there will be no outliers to be handled.

Wrangling the data is only part of the challenge of this project. The next part’s challenge comes from using your creativity in order to establish data visualizations.