Capstone Project #1 Milestone Report

Sangyeol Baek

Problem Statement[1]

In America, telephones have become a necessity in our lives. Not too recently, telephones have gone mobile so that we can carry them around with us on the go so that we do not miss a call from people we know from our colleagues or family members. Need to make a last-minute update to your significant other? No problem. Just call him or her right away. Stuck at home due to a pandemic from all of your friends and/or relatives? No problem. Our phones are here to keep us in touch. Need a day off due to sudden illness? No problem. Just contact your employer. Because of this, long-distance communication is basically a non-issue in America.

However, since we can call practically anyone, chances are you have encountered a call from a phone number or someone you do not recognize. Rarely, even the caller ID may be unknown. Suspiciously, when you pick up, most often you hear a peculiar robotic voice with various purposes ranging from the non-malicious telemarketing to the malicious scams that can drive unsuspecting people to exposing their personal or financial information. In much rarer but more severe cases, the scammer you might encounter is an actual person from another country. The federal government actively combats against such calls (loosely) called “robocalls.” Thus, by finding trends of robocalls in America this project hopes to make combating them easier.

The Data[2]

The data I obtained was from FCC and Kaggle, where the latter had only part of the data from the FCC database. To decide which one to use, I explored through each dataset in a temporary Jupyter notebook. Ultimately, I chose the FCC’s version since it’s constantly updated and has a wider date range than Kaggle’s. Also, surprisingly, the amount of missing data from FCC’s and Kaggle’s datasets were almost proportionally identical.

Ticket ID	Ticket Created	Date of Issue	Time of Issue	Form	Method	Issue	Caller ID Number	Type of Call or Messge	Advertiser Business Number	City	State	Zip	Location (Center point of the Zip Code)
534	10/31/2014 12:29:24 PM +0000	NaN	NaN	Phone	Wireless (cell phone/other mobile device)	Cramming (unauthorized charges on your phone b...	None	NaN	None	Minnetonka	MN	55345	MN 55345\n(44.91531, -93.484053)
535	10/31/2014 01:30:03 PM +0000	10/31/2014	07:03 am	Phone	Internet (VOIP)	Telemarketing (including do not call and spoof...	978-957-4464	Live Voice	NaN	Berwick	PA	18603	PA 18603\n(41.073476, -76.248784)

The majority of the dataset columns consisted of text-based data. Otherwise, the most relevant columns of any numerical sort were the timestamps from the “Ticket Created,” “Date of Issue,” and “Time of Issue” columns. As depicted above, while the “Ticket Created” column has NO

missing values, the other two do, and additionally the dates and/or times could be different from that from the “Ticket Created” column. To combine the three columns, I created a new column called “Issue Occurred,” which will contain the date and time an issue occurred. For the missing dates and/or times, I used the dates and/or times from the “Ticket Created” columns to fill them out. For example:

Ticket Created	**Date of Issue**	**Time of Issue**
10/31/2014 12:29:24 PM +0000	NaN	NaN

=> **Issue Occurred**: 2014-10-31 12:29:24 PM

10/31/2014 01:30:03 PM +0000	10/31/2014	07:03 am

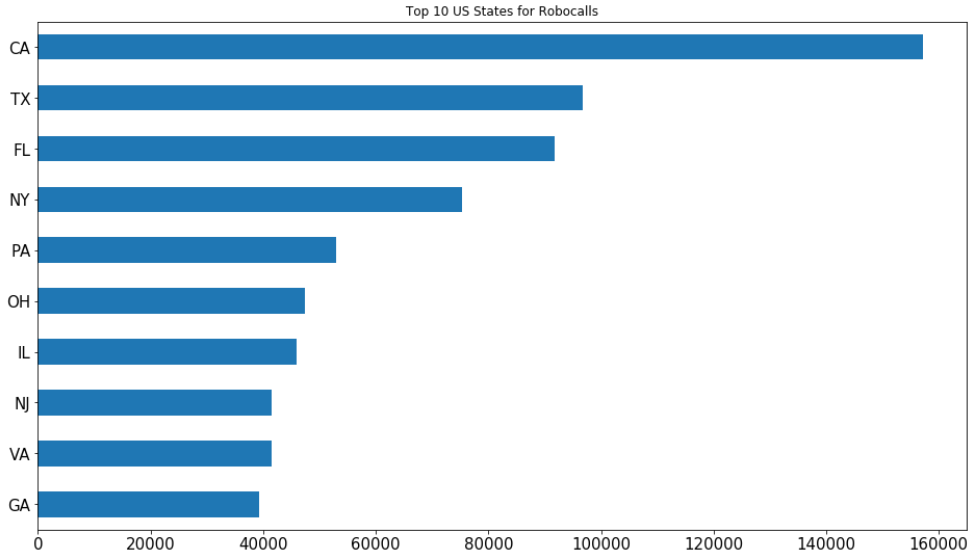
=> **Issue Occurred**: 2014-10-31 07:03:00 AM

After that, since the focus of this project is on robocalls, I selected out the rows in which the entries of the column “Form” matched “Phone.” Next, I would filter out columns that I deemed irrelevant or had too many missing values (specifically “Caller ID Number”, “Advertiser Business Number”, and “Location”). Lastly, I filtered out rows in which all of “City”, “State”, and “Zip” had missing values. After the data wrangling, it resulted in the table below that will be used for later:

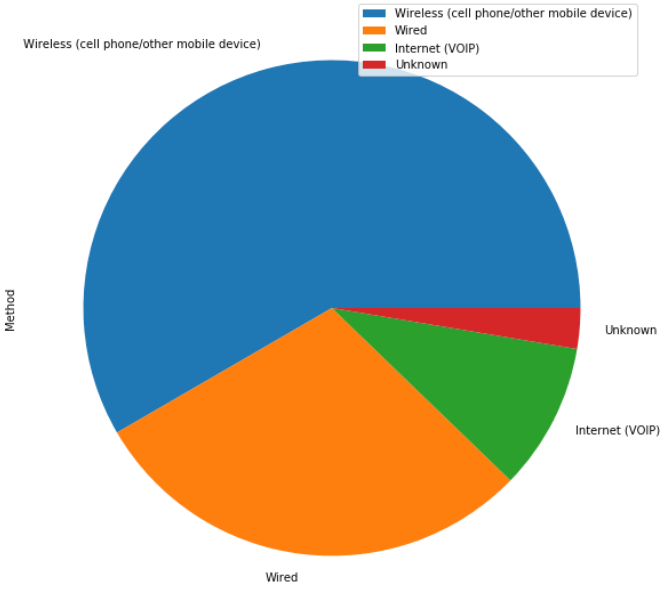
Issue Occurred	Form	Method	Issue	City	State	Zip
2014-10-31 12:29:24	Phone	Wireless (cell phone/other mobile device)	Cramming (unauthorized charges on your phone b...	Minnetonka	MN	55345
2014-10-31 07:03:00	Phone	Internet (VOIP)	Telemarketing (including do not call and spoof...	Berwick	PA	18603
2014-10-31 00:36:00	Phone	Wireless (cell phone/other mobile device)	Telemarketing (including do not call and spoof...	Johnstown	PA	15902

Initial Findings[(3)]

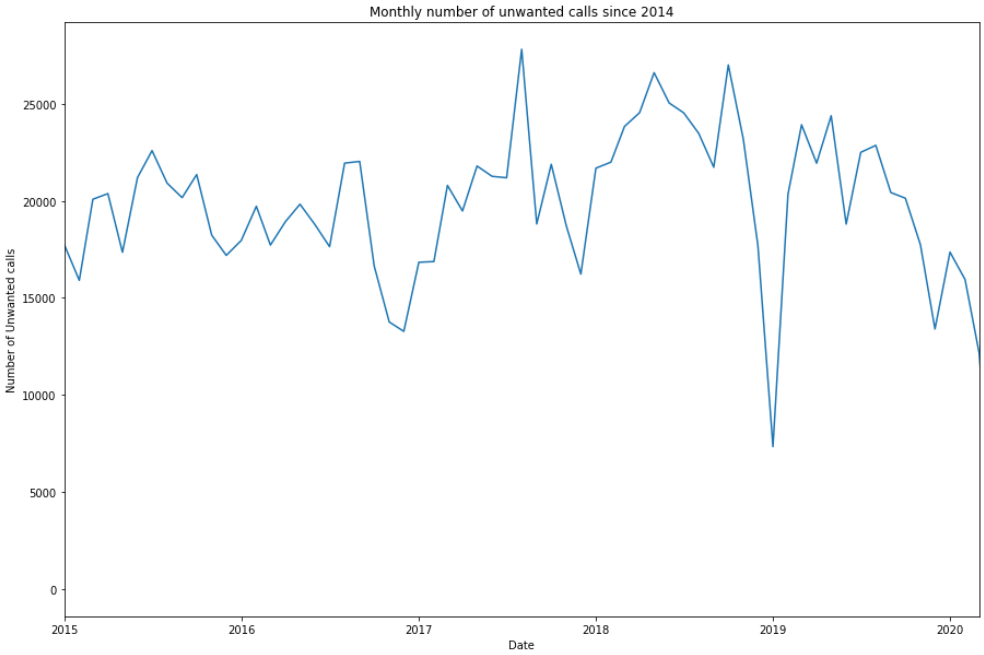
Below is a brief summary of the findings from the wrangled dataset.



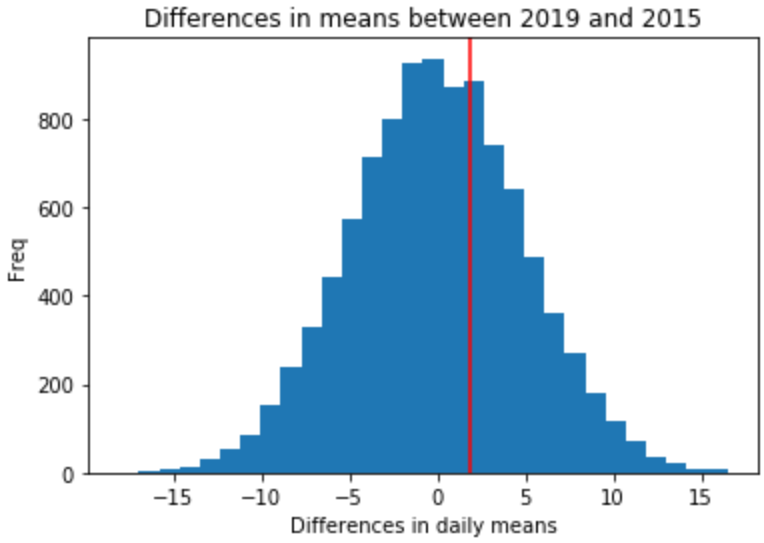
Unsurprisingly, the robocalls tend to target the regions with greater population, which are also subjectively the most well known regions in America. Therefore, there is no doubt that California becomes the most prominent target (by a long run), with Texas at a distant second.



As the number of wireless users increases overtime, it is also no surprise that most robocalls target wireless users.



The most surprising findings I have encountered, however, regards the amount of targeted wireless users over time. The graph seems to indicate no significant increase or decrease. The hypothesis test on whether the daily mean amount of robocalls increases further supports this, as the test fails to reject the null hypothesis that the daily mean has increased from 2015 to 2019. However, with the observed mean difference (indicated by the red line) slightly over 0, it can still imply there was an increase, albeit not to a significant degree.



Challenges Encountered

Needless to say, data will start as a cluttered mess. Ideally, it would be clean enough to grant us a head start. However, very often this is not practical. The dataset for this project is no

exception. In fact, one of the biggest challenges was finding a way to combine the three time-based columns into one new column. Luckily, the columns “Ticket Created” and “Date of Issue” all had entries with a consistent format; “Time of Issue” did not. Looking closer to the latter column, the entries came in various formats:
(these are arbitrarily chosen to demonstrate format inconsistencies)

```
07:03 am, 12:36 PM, 12:30 A.M., 4:13 P.M., 05:02pm, 9:6 P.M., etc.
```

Because of this, trying to make them into datetime objects was cumbersome. Luckily, all had the colon separator and the AM/PM labels. To tackle this, I removed all whitespaces and formatted the AM/PM parts to have consistent formats:

```
07:03AM, 12:36PM, 12:30AM, 4:13PM, 05:02PM, 9:6PM, etc.
```

After that, I am able to create a column of times with consistent format.

However, another challenge I encountered came from trying to fill out the missing values for the date and time of issue columns. The idea was to fill them with the time and/or date from the “Ticket Created” column; however, I ran into trouble trying to use the “fillna” function provided by the Pandas library as all of the date and time columns were filled with the values from the latter regardless of whether there were missing values. Eventually, I decided to go safe and create three numpy arrays from each of the columns and manually fill in the values for each missing value. After that, I would create a new column array that would combine the date and time columns into one datetime column, “Issue Occurred.”

The challenges I encountered after data wrangling was finding use of the newly cleaned dataset. As aforementioned, the dataset I used hardly had any numerical values in the data, which limited my options especially from those designed to test correlations of two different variables. Luckily, I was able to use the timestamps in the dataset as populations, although it required me to wrangle the data a bit more.

Final Thoughts

Doing the project was not only a good use of the skills instilled in this course, but also a learning experience. Of course, there are also ways I could have wrangled the dataset differently and thus yield different results, but after all, data analysis can potentially have unlimited possibilities that it would be impractical to put it all at once given the time constraints and possibly even one’s skillset. Nevertheless, I can only hope that doing this project will be beneficial to FCC’s combating robocalls to ensure safety in long-distance communication.

References:

(1) Capstone Project 1 Proposal
(2) Data Wrangling for Capstone 1
(3) Data Inferencing for Capstone 1