Capstone Project #1

Sangyeol Baek

Overview

- Problem
    - The number of robocalls in America has increased dramatically over the past few years.
    - According to USA Today, there were 58.5 million robocalls in 2019 alone, 22% increase from 2018. Because of this, the government passed a law attempting to deter them, such as fining robocallers up to $10,000 per call.[1]

Data Acquisition

I acquired the dataset from the FCC site:

https://opendata.fcc.gov/Consumer/CGB-Consumer-Complaints-Data/3xyp-aqkj

NOTE: the database is constantly updated daily, but the dataset used is dated up to April 9, 2020.

Data Wrangling

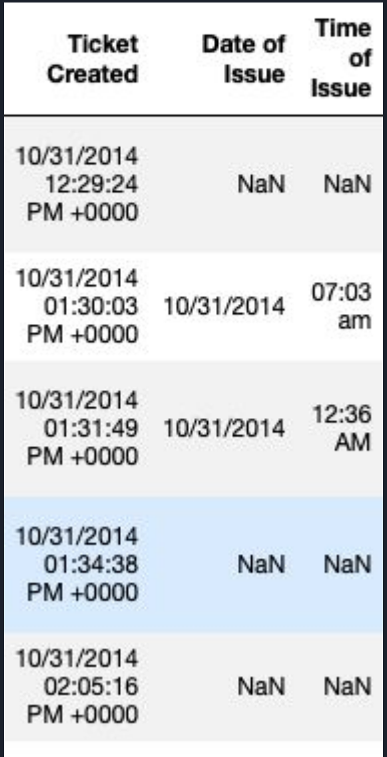
- Since this concerns robocalls, filter rows in which “Form” matches “Phone”
- Three datetime-based columns: “Ticket Created,” “Date of Issue,” and “Time of Issue”
  - For null values in the latter two columns, fill with the date/time component from “Ticket Created”
  - Combine the three into one datetime column: “Issue Occurred”
- Filter out all rows that have null values for ALL of columns “City”, “State”, and “Zip”

- **Null Values**: except for the above mentioned, filled the null values with “Unknown”
- **Outliers**: technically, there are no purely numerical columns, thus no outliers.
  - However, for the column “Date of Issue”, filter rows to those with years ranging from 2014-2020, since a few entries contain out of range dates such as year 2916, 0001, etc.

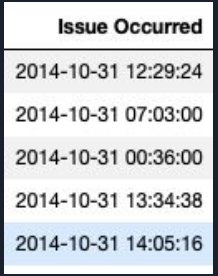
Challenges with Data Wrangling

The Main Challenge

- Combine the three time-formatted columns into one column. (“Ticket Created” has **NO Null Values**)
  - If “Date of Issue” or “Time of Issue” is null, use data from “Ticket Created”
  - Otherwise, use the data from the former two columns



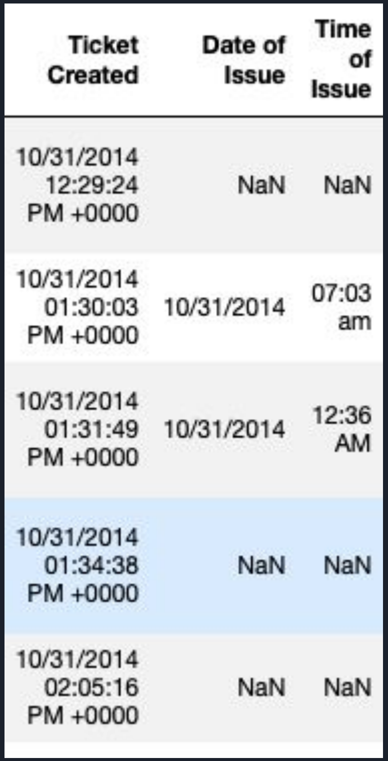
Ticket Created	Date of Issue	Time of Issue
10/31/2014 12:29:24 PM +0000	NaN	NaN
10/31/2014 01:30:03 PM +0000	10/31/2014	07:03 am
10/31/2014 01:31:49 PM +0000	10/31/2014	12:36 AM
10/31/2014 01:34:38 PM +0000	NaN	NaN
10/31/2014 02:05:16 PM +0000	NaN	NaN



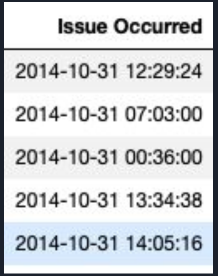
Issue Occurred
2014-10-31 12:29:24
2014-10-31 07:03:00
2014-10-31 00:36:00
2014-10-31 13:34:38
2014-10-31 14:05:16

Challenges with Data Wrangling (cont.)

- CHALLENGES
  - Values in **Time of Issue** had inconsistent if not peculiar formatting
    - E.g. “07:03 am”, “12:36 AM”, “4:30 P.M.”, “2:30PM”, “9:6 pm”, etc.



Ticket Created	Date of Issue	Time of Issue
10/31/2014 12:29:24 PM +0000	NaN	NaN
10/31/2014 01:30:03 PM +0000	10/31/2014	07:03 am
10/31/2014 01:31:49 PM +0000	10/31/2014	12:36 AM
10/31/2014 01:34:38 PM +0000	NaN	NaN
10/31/2014 02:05:16 PM +0000	NaN	NaN



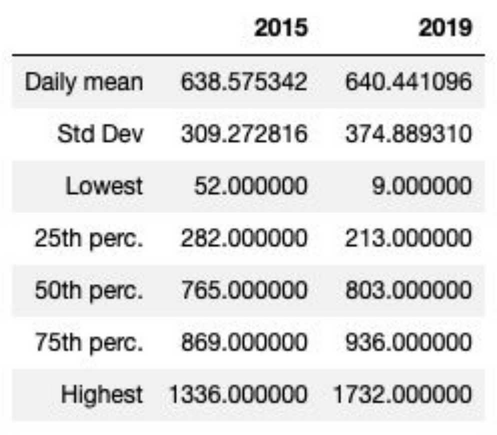
Issue Occurred
2014-10-31 12:29:24
2014-10-31 07:03:00
2014-10-31 00:36:00
2014-10-31 13:34:38
2014-10-31 14:05:16

Initial Hypothesis

Determine whether there is an increase in daily total number of robocalls in America from 2015 to 2019. In other words, were there more in 2019 than in 2015?

<u>**Null Hypothesis**</u>: There is no increase from 2015 to 2019.

For this, use the t-test for the difference of means using the bootstrap method, with α = 0.05 as the statistical significance. Also I ensured each “population” contained 365 unique entries.



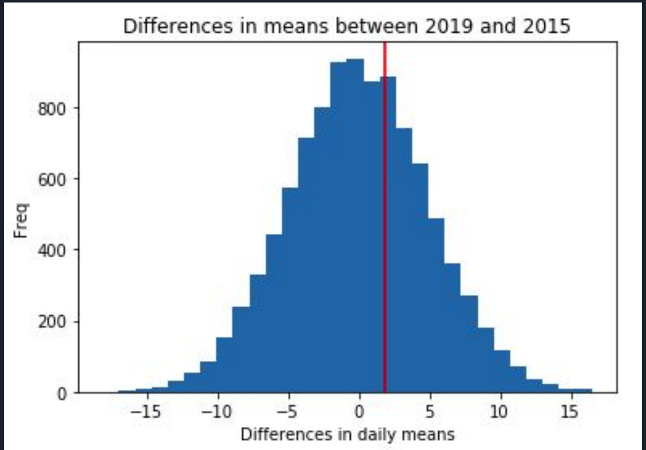
	2015	**2019**
Daily mean	638.575342	640.441096
Std Dev	309.272816	374.889310
Lowest	52.000000	9.000000
25th perc.	282.000000	213.000000
50th perc.	765.000000	803.000000
75th perc.	869.000000	936.000000
Highest	1336.000000	1732.000000

Initial Hypothesis (cont.)

Results

The p = 0.352 > α, so we cannot reject the null hypothesis.

Although we cannot reject the null hypothesis, the graph shows that the hypothesis being true is more likely.



Further Analysis

Determine whether there is an increase in proportions for robocalls targeting wireless users from 2015 to 2019.

Null Hypothesis: There is no increase in robocalls for wireless users.

To do this, extract the “Method” column with “Issue Occurred” matching years 2015 and 2019 to use as the populations. For this, use the z-test for difference in proportions. With $\alpha = 0.05$, $z > 1.64$ to reject the null hypothesis.

Results:

Surprisingly, the z-value resulted as -1.92, so clearly $z < 1.64$, so we cannot reject the null hypothesis. In fact, it implies there were actually less wireless users targetted in 2019 than in 2015.

References

1. https://www.usatoday.com/story/tech/2020/01/15/robocalls-americans-got-58-5-billion-2019/4476018002/