**311 Calls in San Francisco**

**April18, 2020**

**Group #2 Team Members:**

**Paul Pineda**

**James Ye**

**Tanvir Khan**

**Travis Le**

**Project Description:**

The 311 San Francisco is a free public service and platform that serves as a tool to accept, process, and address various non-emergency requests, complaints, and inquiries from the community, such as requests for city street cleaning, graffiti, road conditions, etc.

Analysis of 311 calls can be of great use for a wide variety of purposes,such as understanding the status of a city, effectiveness of the government services in addressing such requests.

In this project, we use the 311 requests data from the City of San Francisco to determine if there are any relationships between the number of requests and day of the week, income, population, and shelter-in-place lockdown.

This project spans 311 complaints from January 1, 2017 through April 9, 2020.

**Hypothesis (Team Member in charge):**

* Middle of the week would be more complaint (Paul Pineda)
* With higher income would be more complaint (James Ye)
* Number of service requests/complaints increase by population (Tanvir Khan)
* Number of service requests/complaints increase by shelter-in-place (Tavis Le)

**Data Sources Used:**

* 311 call data from City of San Francisco from 2017 - Present:

<https://sf311.org/information/reports>

* Income and Population per zip code data from US Census:

<https://www.census.gov/data.html>

* Google Map API

**Data Cleaning:**

* Pandas to remove unnecessary data from initial .csv file
* Slice Method to separate the date data and put it the different columns
* Use of the US Census of Bureau and Google fu to retrieve zip code for each neighborhood, latitude and longitude and removing invalid zip codes that are not from SanFrancisco County.
* Clean the dates using Datetime method to obtain values in different formats because the exact time is not required for this analysis. We also convert the relevant columns to keep them as numbers.

**Challenges:**

* 311 Raw Data was 1.7 GB file, should have used API key
* Finding relevant data from US Census Bureau
* Even though raw 311 cases data file has Zipcode column, however there was no data on Zip code.
* Raw data was based on neighborhood rather than Zipcode.
* Retrieving zip codes with neighborhood for each service requests
* Finding the valid zip code related to that neighborhood reference for City of San Francisco was time consuming.
* Combining and merging data from different sources in a correct way

**Tools for our analysis:**

* Bar Chart
* Pie Chart
* Scatter Plot
* Heatmap
* Linear Regression

**Cases:**

For this project, each case/request/complaint is associated with their respective zip code. Each case count is a group of all 311 complaints for that zip code for the selected time period.

## Transformation from Raw to Final DataSet:

## 311 Raw Data

* + This project spans 311 complaints from January 1, 2017 through April 7, 2020 .
  + 311 receives millions of complaints each year across all city agencies: Health, Environmental Protection, Sanitation, Police, and others.
  + For this project, the data was filtered as follows:
    - Selected only data that were Opened cases from 2017 - Present.
    - Only complaints which had a neighborhood as part of the complaint, to make data analysis possible using Zip code file.

## Zip Code by Borough

* + A list of zip codes and their associated neighborhood was extracted from the google fu and Census data
* Income and Population
  + Income and population information by zip code was retrieved from the US Census using API

## Combined Dataset

## The datasets were merged and transformed to display the required data. The combined datasets were uploaded to GitHub for reproducibility.

* + Final DataFrame
    - Data Collected - 2017 - 2020
    - Case Count 1,873,843
    - From 47 Column to 18 Column



**Analysis strategy**

* Bar Chart
* Pie Chart
* Scatter Plot
* Heatmap
* Linear Regression
* Statistical Analysis

**Analysis:**

**Hypothesis 1:**

HA = There will be more 311 cases during the middle of the week

H0 = There will not be more 311 cases during the middle of the week

**Hypothesis 2:**

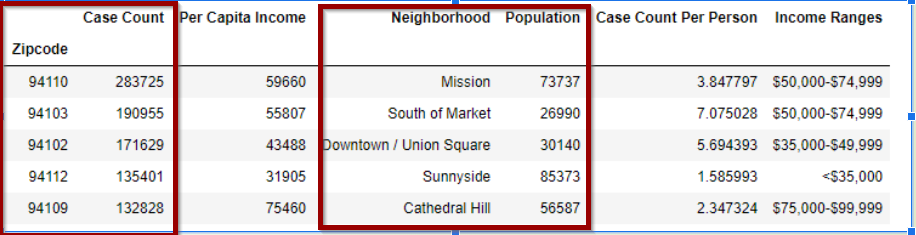
HA = With higher income there would be more complaint

H0 = With higher income there would not be more complaint

**Hypothesis 3:**

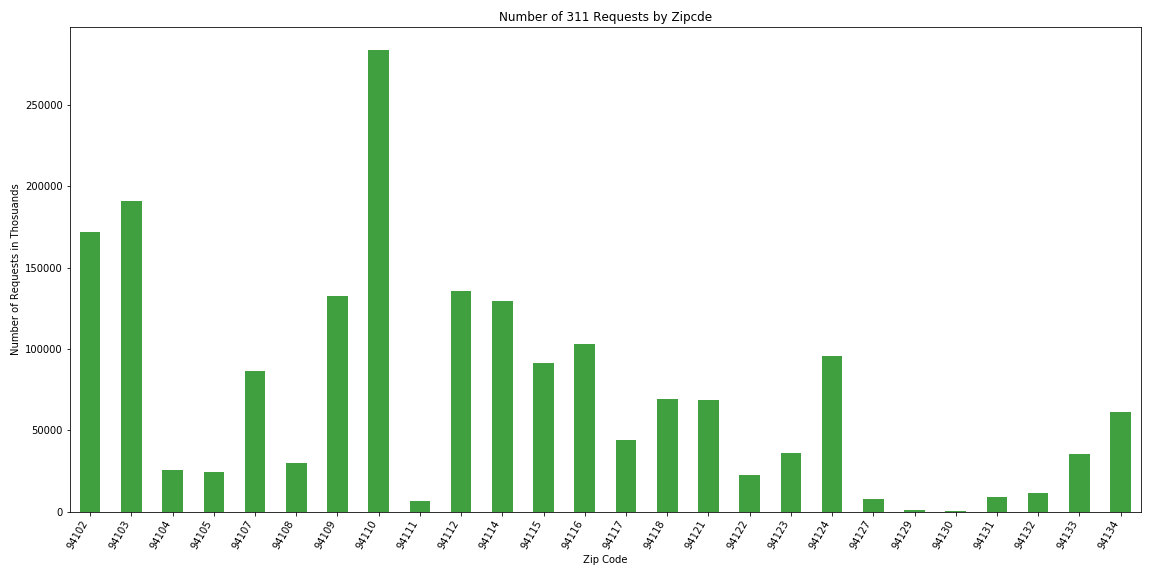
HA = If the population number is higher, then the number of service requests/complaints is higher in that Zip code

H0 = If the population number is higher, then the number of service requests/complaints is not higher in that Zip code



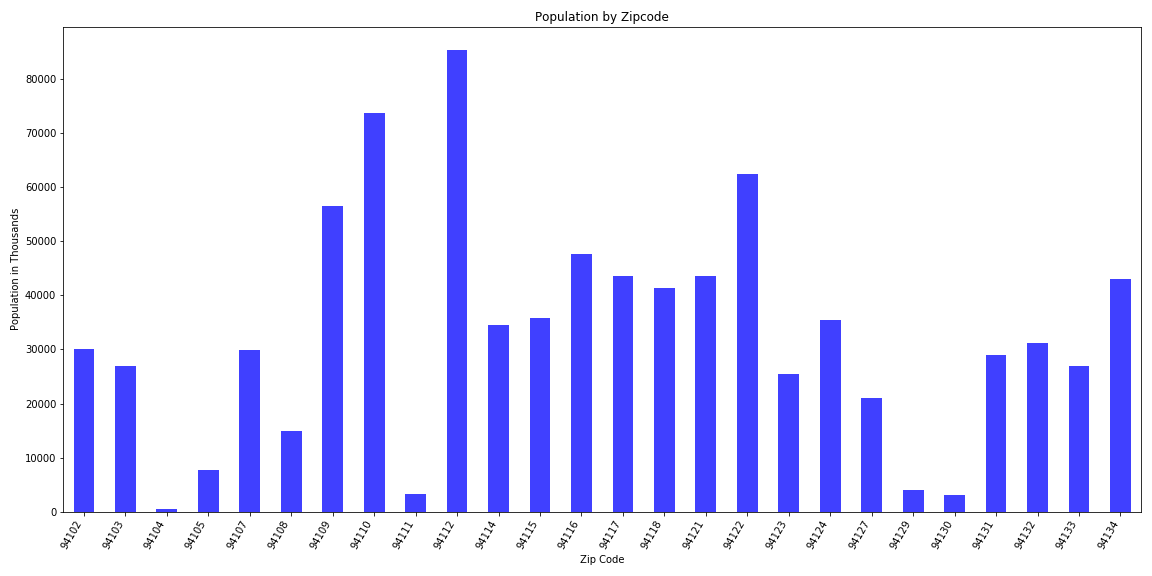
To analyze if there is any relationship between number complaints and population number for each zipcode, series of bar graph, google heat map, regression analysis and hypothesis testing were carried out.

Here’s the bar graph that shows how the number of requests varies in each zip code.



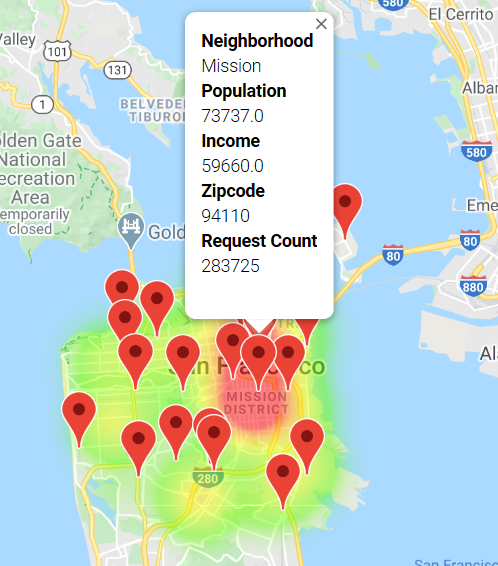
We found that zipcode 94110 Mission district had the most requests followed about 283725 by SOMA 190955 and then Tenderloin UnionSquare area about 171629 requests were made

This bar graph shows population size in that zipcode



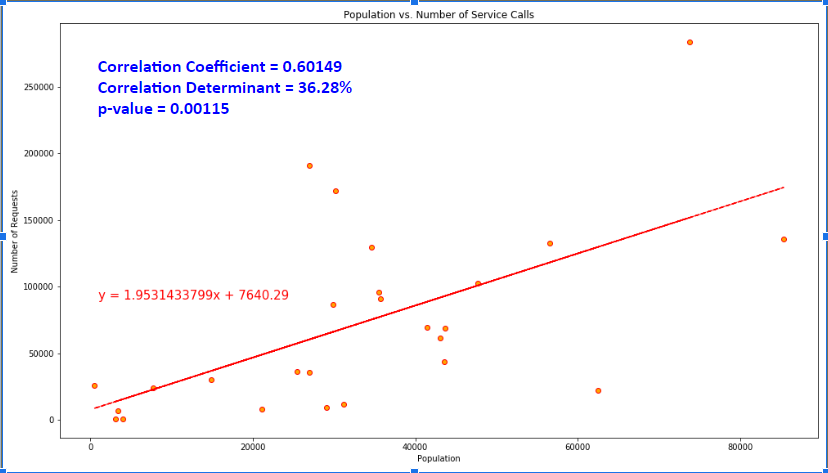
The highest populated neighborhood associated with zipcode is 94112 is Sunny Side OuterMIssion has population size of about 85237 then followed by 94110 Mission district has population size of about 73737 and then followed by 94122 Inner sunset has population size of about 62516

To further show and project the relationship between these variables, google heat map was created with markers that show the neighborhood name with the number of requests, income & population size. As from the Bar graphs, Mission district has the most requests & second highest population which is reflected as the most heated area (red) in the Heat Map below.



Some statistical tests on the variables were carried out to verify the findings and assumption is correct or not.

Regression modeling & Pearson Correlation test:



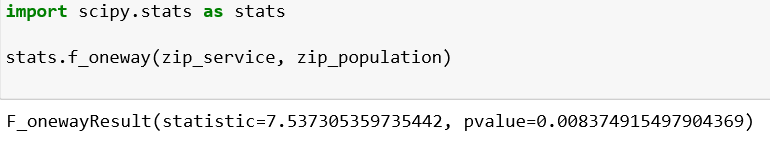
Here the Regression line shows a positive increasing relationship between the number of requests/complaints and populations. The regression line slope shows that no matter what the population size is, each zip code commits about 2 requests per day!

The correlation coefficient is 0.6 shows there is a moderately strong relationship between the two observed variables number of requests and population numbers for each zip code. And 36% change/variation in number of requests is caused by population number.

F-Test Statistical Testing:

The correlation coefficient and determinant shows that there is a relationship between the number of requests/complaints and population number and how well a regression model fits the data set, however it doesn’t tell us the entire story hence we did further testing to test the validity of our data.

The hypothesis testing was conducted on our variables using F- statistic Test and F-statistic test was used to check if the data conforms/fits to a regression model.



Result: Since p-value is very small 0.00837 ( ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis. We found our hypothesis is statistically significant at 5% significance level and were able to reject the null.

Conclusion:

From the bar plots, google heat maps, linear modeling, and statistical analysis shows the validity of our model and hence our hypothesis assumption is correct, if the population number is higher, then the number of service requests/complaints is higher in that Zip code.

**Hypothesis 3:**

HA = The number of service requests/complaints is higher during Shelter in Place

H0 = The number of service requests/complaints is not higher during Shelter in Place

# **Conclusion:**

## Summary:

The initial question was to determine if a correlation exists between:

* Middle of the week would be more complaint (Paul Pineda)
* With higher income would be more complaint (James Ye)
* Number of service requests/complaints increase by population (Tanvir Khan)
* Number of service requests/complaints increase by shelter-in-place (Tavis Le)

Using plots, linear modeling, and statistical analysis, on each hypothesis, middle of the week, income, shelter-in-place did not appear to be correlated with the number of requests/complaints, hence initial assumptions were rejected. However only population size and number of complaints and with their respective zip codes were correlated. The validity of the data was indicated by summary statistics for the chosen variables (number of complaints & population number, which showed low p-values less than 0.05.

Further data exploration beyond the original scope of the hypothesis resulted in the discovery of significant correlations in complaints by neighborhood. Staten Island has a greater percentage of complaints per capita. Early spring months have the greatest volume of complaints, which is logical as they are after snow plow contact, salt treatment, and snow melting, and when individuals increase outdoor activities.

## Insights

* There may be a self-selection of 311 complainants. Certain people may not want to call 311, for example if they are not comfortable speaking on the phone, or do not have time in their workday, or do not speak English. 311 does provide service in multiple languages, but callers may not know this.
* 311 complaints also come from web forms, which is limited to individuals with computers and internet access. While 311 does have a mobile app, it is limited. Some of these complaint categories are not available on the mobile app and require the web interface or a phone call to report.
* Median income might have been a better indicator than mean since high incomes in large cities skew the average higher. In reality, close to 12% of San Francisco is living below or near the poverty line, a very different story from a mean income of $64,900.
* Data set was very large and it was difficult to process, so filtered empty columns on many levels to reduce the data size.

## Future Research

* These two datasets are extremely rich and much more research could be done.
* Incorporating more Census demographic data like age, ethnicity, and race, of the residents. This could have developed predictive models of complaints by population characteristics.
* Monthly complaints or particular complaints could be broken down into neighborhoods of some neighborhood showing unresolved or increasing conditions at a different rate than the other neighborhood.
* Incorporating weather data and comparing average response time for complaints during different weather conditions and seasonality.
* Predict time required in terms of range of days to resolve specific complaints in the specific zip code.

.