**Final Project: San Francisco Human Waste Report**

**Joe Hernandez**

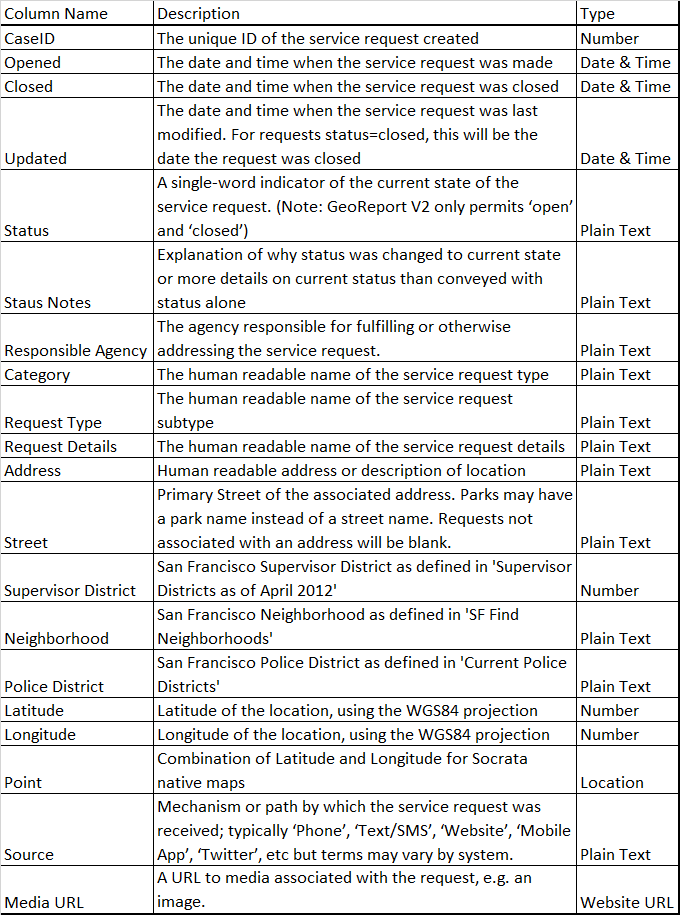
**IST 652 - Scripting for Data Analysis**

**Introduction**

San Francisco has been long considered amongst the biggest cities in the west coast. The San Francisco Bay area is considered to have engineering marvels such as the Golden Gate Bridge, pristine vintage and operational cable cars, and a technical hub with Silicon Valley in the immediate vicinity. With 25 million tourists and the population approaching one million the city faces challenges with over 7,500 in the homeless population leading to health and sanitation residuals such as drug use, litter (such as discarded needles), and human feces in the streets and sidewalks. As a response, San Francisco’s Department of Public Works instituted the “Poop Patrol” as a new position to address the human waste issue

**Data and Source:**

The data used for this analysis is a comma separated file provided by DataSF, who belonging to the Office of the Chief Data Officer for San Francisco, seek to transform city operations and services through data resulting in improved quality of life and work for their residents, employers, employees, and visitors. This 1.14 GB file is a collection of San Francisco’s municipal 311 calls for non-emergency services (graffiti, sidewalk cleanup, etc), which is accessed through the link provided following the data dictionary, from July 2008 to March 2019 and continues to be updated over time. The dataset consists of over 3.5 million records and 20 columns with the following dictionary:



<https://data.sfgov.org/City-Infrastructure/311-Cases/vw6y-z8j6>

Preprocessing the data required a reduction in the size of the dataset. To answer the analysis questions the number of columns were reduced to 11 consisting of: Opened, Closed, Updated, Status Notes, Request Details, Street, Neighborhood, Latitude, Longitude, Source, and Media URL. Additionally, to focus on the human waste issue only the Request Details of that category will be used. Finally, for the most complete data reports without Opened nor Closed datetime values are removed. These actions reduced are then compiled into a dataframe of about 135,000 rows and 11 columns and when saved is less than 10 MB.

**Methods of Analysis**

The methods of analysis will focus on the following questions:

* *Question 1: Which times of the day have the highest frequency of reporting, least frequency?*
  + This question will observe the time of day for residents and visitors where the most human waste is seen and its location. Residents exposed to persisting human waste are vulnerable to health and sanitation hazards. Visitors also can see this information to see what percentage of the human waste occurs during the day.
  + Units of analysis: Opened, Longitude, and Latitude
  + Comparison: Morning (0500 – 1159), Afternoon (1200 – 1659), Evening (1700 – 2359), and Night (0000 – 0459)
* *Question 2: How much waste is each neighborhood and street location?*
  + Given the geographical representation of the previous question, this question looks to identify specific neighborhoods and streets and the frequency on how much human waste is reported there. This can be useful information for the city to prioritize their resources to cleaning up that specific area.
  + Units of analysis: Neighborhood and Street
  + Comparison: Sum of human waste reports per neighborhood and per street
* *Question 3: Are there any longer-term trends in the reporting and resolution of human waste? Is reporting and resolution declining?*
  + This question is reshaped from the original question to address the severity of human waste reports over time and adds descriptive statistics to how responsive the city is at cleaning up and if there exists an inverse relationship between the two.
  + Units of analysis: Opened and Closed
  + Comparison: Sum of human waste reports per year and descriptive statistics in the difference between the time Opened to Closed
* *Question 4: Has there been a transition in the source of reporting human waste covered by the data?*
  + The emergence of technologies has allowed for telecommunications and the internet to be used as convenient resources to reporting and responding to human waste reports. This analysis will look at the trend to see if there are any changing trends over time.
  + Units of analysis: Opened and Source
  + Comparison: Cumulated and daily sum of human waste reports by source

**Description of Program**

The aforementioned questions provided several challenges in changing attributes to the appropriate data types. In order to do so several libraries are loaded into the program such as pandas, numpy, statistics, plotly, datetime, time, matplotlib, and plotly.io. The program begins with logging into plotly with the appropriate credentials and api key. Afterwards the preprocessing loads the file and extracts the appropriate columns and rows relevant to the analysis then puts it into a dataframe where it is saved as an excel file.

As a visual, the code first generates a plot using plotly of all the human wastes points in San Francisco using longitude and latitude coordinates as a data object with a layout object centering the map at the median latitude and longitude of only unique coordinates. A figure object is comprising of the data object and layout object with the human waste report title and is then plotted using plotly and is then called for display.

In the first question, a new dataframe is made where columns Opened and Closed are converted to datetime objects. The Updated column is used to populate the hour of the Opened column. Then four subsetted dataframes are created to represent the time of day for morning, afternoon, evening, and night. A printout of the number of reports and the percentage is produced to show the results. Additionally, a similar geographic map plot to the San Francisco human waste report is produced for each time of the day dataframe.

For the second question, two pivot tables were created. The first pivot table indexes the Neighborhood column and takes the aggregate length of each report in that neighborhood, whereas, the second pivot table indexes the Neighborhood column to the first level and the Street column to the second level and takes the aggregate length for each report in that street. The pivot table is then sorted in descending order, where dataframes of the top 5 neighborhoods and top 25 streets are produced and saved to an excel file. A bar chart is produced for the Neighborhood with coloring for each neighborhood and for street using the same coloring based on neighborhood listing streets with high human waste frequency.

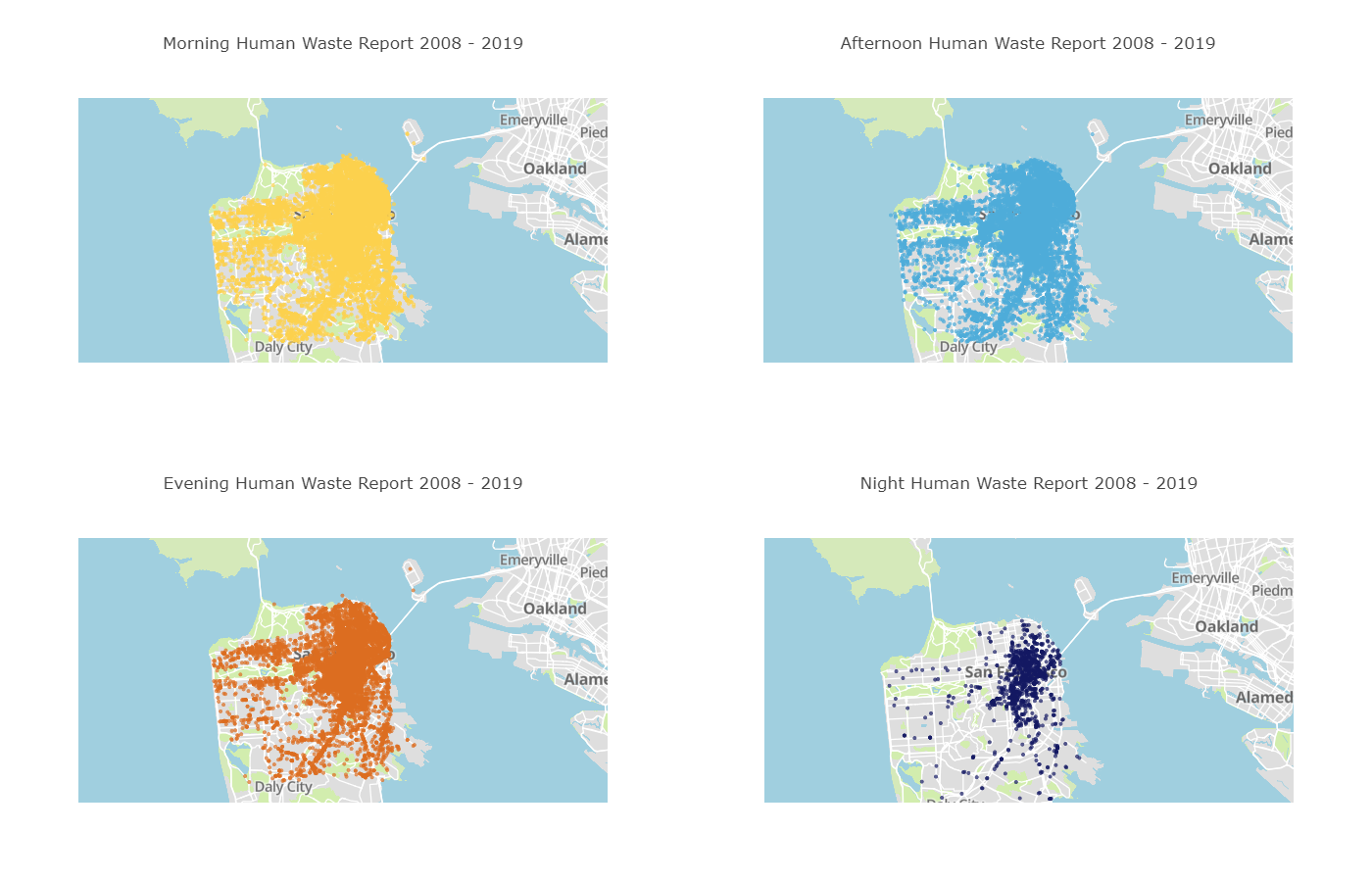
The third question generates a new dataframe where columns Opened and Closed are converted to datetime objects again; however, the Updated column takes the difference of the Opened and Closed columns and converts the objects into minutes elapsed. It is then put into a dataframe where each year is subset to run descriptive statistics. To prevent the calculations from producing exponential values, the pandas options were set to display float format. When the descriptive statistics run for each year it is then merged into one dataframe and saved to an excel sheet. Graphically a boxplot for the first, middle, and last year is produced by month.

In the final question a new dataframe is created where Opened is converted to a date object. It is then converted to a pivot table where Source is the first level index and Opened is the second level index. The Updated column is used to aggregate the sum and cumulative sum of which method the human waste was reported. The pivot table is then saved to an excel file. In the graphics, two time series plots are produced showing the number and cumulative of reports for that day.

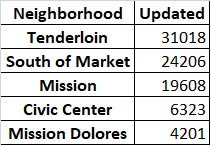
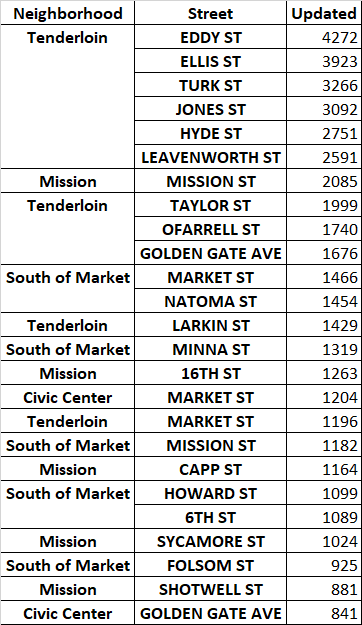
**Output File Description**

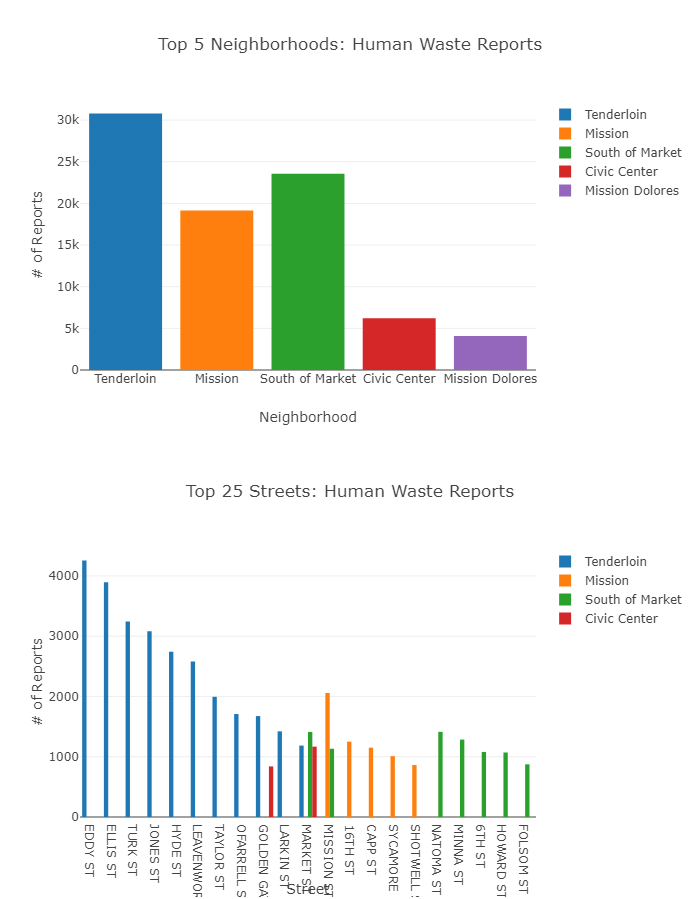
The following output, produces the printout, excel document, and graphing to answer each question:

* *Question 1: Which times of the day have the highest frequency of reporting, least frequency?*
  + Output: Printout identifying the number of reports per time of day and percentage: Graphical maps plotting the location of the reports per time of day  
    *65368 Total Morning Reports: 48.26 %  
    49992 Total Afternoon Reports: 36.91 %  
    18770 Total Evening Reports: 13.86 %  
    1320 Total Night Reports: 0.97 %*

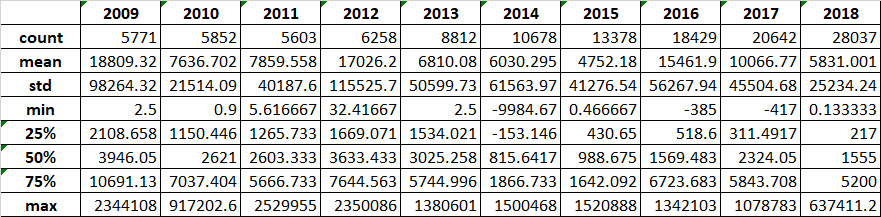
**

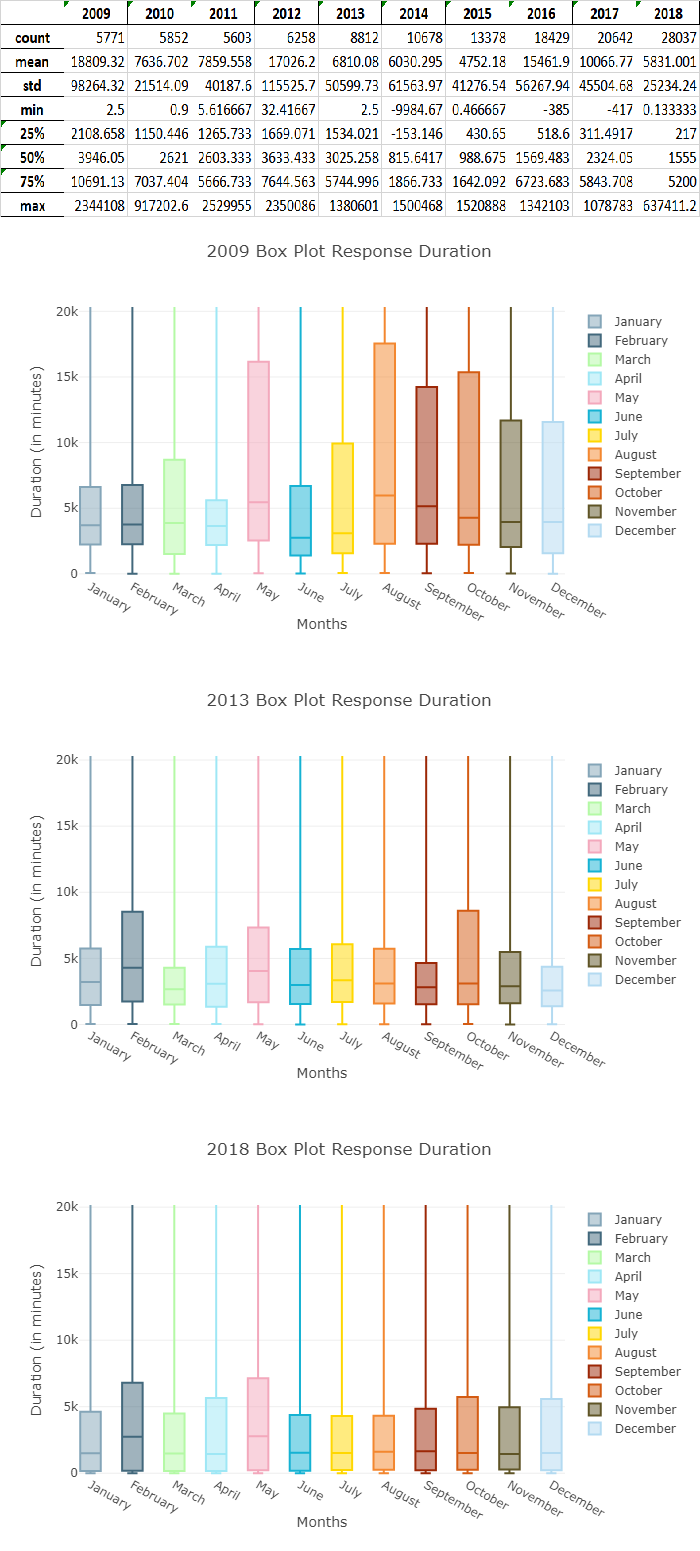
* + Upon observation, the analysis identifies the morning to have the most reporting of human waste; however, given the Night has the smallest number of reporting, it is also the time of day with the least activity. It can be speculated that some of the Morning case reports are residuals from the night human waste activities. Additionally, for the most active time of day, the afternoon shows a high percentage of reports.
* *Question 2: How much waste is each neighborhood and street location?*
  + Output: Excel file of top five neighborhoods with highest reports (2 columns: Neighborhood(category), and Updated(count)), Excel file of top 25 streets with highest reports (3 columns: Neighborhood(category level 1), Street(category level 2), and Updated(count)), Graphical bar chart of Top 5 Neighborhoods and Top 25 Streets

*  
*

**

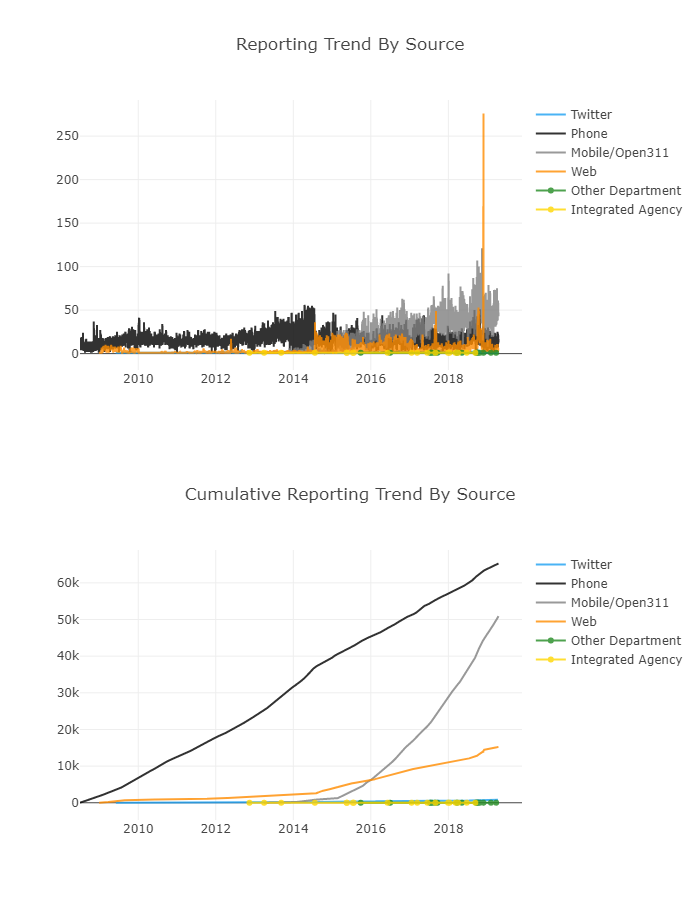
* + Among the top five neighborhoods, the Tenderloin, South of Market, and Mission show significantly higher reports of human waste in the city, such that, 11 streets of the top 25 street reports belong to the Tenderloin, 7 belongs to South of Market, and 5 belong to Mission. This useful information can be used by the city to allocate more cleanup resources in those areas.
* *Question 3: Are there any longer-term trends in the reporting and resolution of human waste? Is reporting and resolution declining?*
  + Output: Excel file of descriptive statistics for human waste reports and cleanup per year, Graphical boxplot for 2009, 2013, and 2019

**

**

* + *Although there is an increasing trend in the number of human waste reports since 2011, there is also a decreasing resolution time for these cases. The best way to examine the responsiveness of these cases is to look at the measures of central tendencies of the mean and median. While they show an ebb and flow trend, this would need further investigation over time with the newly instated “Poop Patrol” in mid 2018.*
* *Question 4: Has there been a transition in the source of reporting human waste covered by the data?*
  + Output: Excel file of daily source report, Excel file of cumulated daily source report, Time series plot for daily and cumulative daily source report

*Excel files excluded due to record length of 9340.*

**

* + The phone has been dominant in human waste reporting since this data has been recorded; however, mobile phone is emerging as the most convenient method of reporting. This is reflected in 2014 where there was nearly a swap in the source reporting of human waste. Additionally, the web offers technology where the community can take pictures of their cases to ensure it is cleaned up. Lastly, Twitter can pose a new upward trending reporting method depending on how DPW can take advantage of geotags and hashtags.

**Conclusion**

San Francisco faces many challenges in addressing the human waste issues. In such a finite area with a high population density there is a social disparity between the homeless and the residents and visitors. Throughout the day, the city is plagued with human waste issues where the morning and afternoon needs significant cleaning to remove it. The top three neighborhoods where human waste is reported compose of streets in the top 25 and serve as an indicator to where to allocate cleanup resources. Although there is an increasing trend in human waste reporting there is also an improvement in the cleanup responsiveness. Lastly, the reporting process has caught up technologically where it offers convenient alternatives to report human waste aiding in the rapid responsiveness.

The human waste problem is one of many issues for the city. The information provided by DataSF is useful and can help prioritize issues by region and time. Such abundance of information has allowed for improvement to processes such as responsiveness and convenience to reporting. Such innovations in recording data and compiling it for data scientists to process brings to the attention for the city to be accountable for handling the issues of their community.