



# SAN FRANCISCO POLICE DEPARTMENT CALLS FOR SERVICE A DATA VISUALIZATION REPORT

CA682 ASSIGNMENT

DIVYA AREN | 20210762 | MCM-DA | DIVYA.AREN2@MAIL.DCU.IE



# ABSTRACT

The project aims to analyze the incident calls received by the San Francisco Police Department (SFPD) regarding criminal activities (unverified) from 2016-2020. Data is cleaned and transformed to visualize the service calls for top 10 criminal activities reported to San Francisco Police using a racing bar chart and analyze the increase in reports of these commonly reported crimes from 2016-2020. In another visualization we analyze the number of service calls received during daytime and at night by SFPD (2016-present) using an area chart and compare it with a stacked bar chart, showing the comparison between the number of calls to report criminal activities during daytime and at night and the total number of service calls over the span of these four years and discuss how area chart is a more justified choice contrary to popular belief.

With the help of the racing bar chart, we visualize the number of incidents which are most commonly reported from 2016 to present with an animated timeline. We see that Passing Calls is the highest reported crime and with the help of the stacked area chart we can conclude that the number of criminal incident reports are more in the daytime than during the night hours. We are also able to visualize the total crimes reported to SFPD.

# INTRODUCTION

## DATASET

Dataset - Police Department Calls for Service - is created by the City and County of San Francisco and is openly available for use at <https://datasf.org/>. It caters to the records of service calls received by San Francisco Police Department every day from 2016 to present regarding reports of crimes (unverified) in the city.

## ATTRIBUTES OF DATA

The dataset is 543 MB and has 3.75 million records. It has 14 columns like Crime Name, Report Date, Call Date etc. with different data types like Text, Number, Datetime.

## BIG DATA ASPECTS

The dataset is changed and published daily and the same is updated on the website with a lag of 14 days. The daily frequency of change in the dataset indicates the aspect of velocity of big data. As the data is created and updated by the Police Department of San Francisco, it also caters to another aspect of big data - Veracity. Value and Variability are other important aspects we witness during our analysis of this dataset.

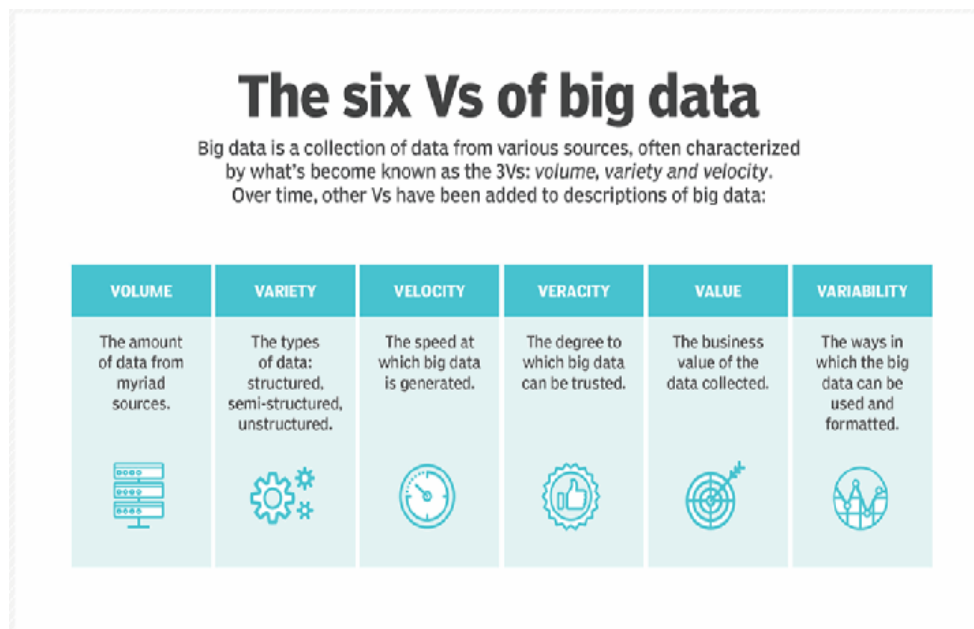


Figure 1: Aspects of Big Data (<https://searchdatamanagement.techtarget.com/definition/big-data>)



## INCIDENT CALLS

- More calls for service during daytime than at night
- 50% calls for top 10 crimes
- Maximum calls received in 2017 (from 2016 to 2020)



## TOP 10 CRIMES REPORTED

- Passing Call
- Traffic Stop
- Suspicious Person
- Homeless Complaint
- Traffic Violation Cite
- Improper parking stopping or standing
- Audible Alarm
- Muni Inspection
- Well-Being Check
- Trespasser

# DATA HANDLING

## DATA EXPLORATION

Data exploration was carried on different aspects of the dataset and to better understand the relationship between different features. For example, finding top 10 crimes based on the number of service calls received from 2016 till date and visualizing the total number of calls in each year through a line chart. It is found from the exploratory analysis that in 2017 SFPD received the highest number of service calls among these 4 years. Interestingly, out of 3.75 million records, 1.89 million records are for the calls received for reporting the top 10 most common crimes.

## DATA CLEANING

Data was cleaned and some of the law codes mentioned in the Original Crime Type were renamed to their description for creating more understandable visualizations. For example, Law code “22500e” was replaced with “Improper Parking or Standing”.

## DATA TRANSFORMATION

While exploring the data, some peculiar features were identified. They were insightful and required to be presented in a more readable and easily comprehensible manner. Visualization of these attributes like the time of the day for maximum incident calls, calls received during day and night-time and understanding the increase in the crimes reported to SFPD were particularly important to the analysis and for understanding the service calls received by SFPD.

Data is transformed by creating several different data frames in Pandas. Year wise data for the top 10 crimes was extracted from the cleaned data frame by transposing the data frame created by using function *pivot\_table* for creating the racing bar chart.

Top reported crimes were identified using *value\_counts* and were then sorted according to the time of calls into two groups – daytime and night. This data was then used for second visualization.

# VISUALIZATIONS

## THE RACING BAR CHART

The racing bar chart is a visualization phenomenon and is engaging for users as they can play the timeline of events interactively. Presenting the most commonly reported crimes through the bar chart races tells a compelling story and plays the data in a unique, interesting, and understandable format, with different points of interest like change in crime rank, change in number of incidents in a year and their relative positions for different years. This helps users visualize the development of crimes during the span of these five years.

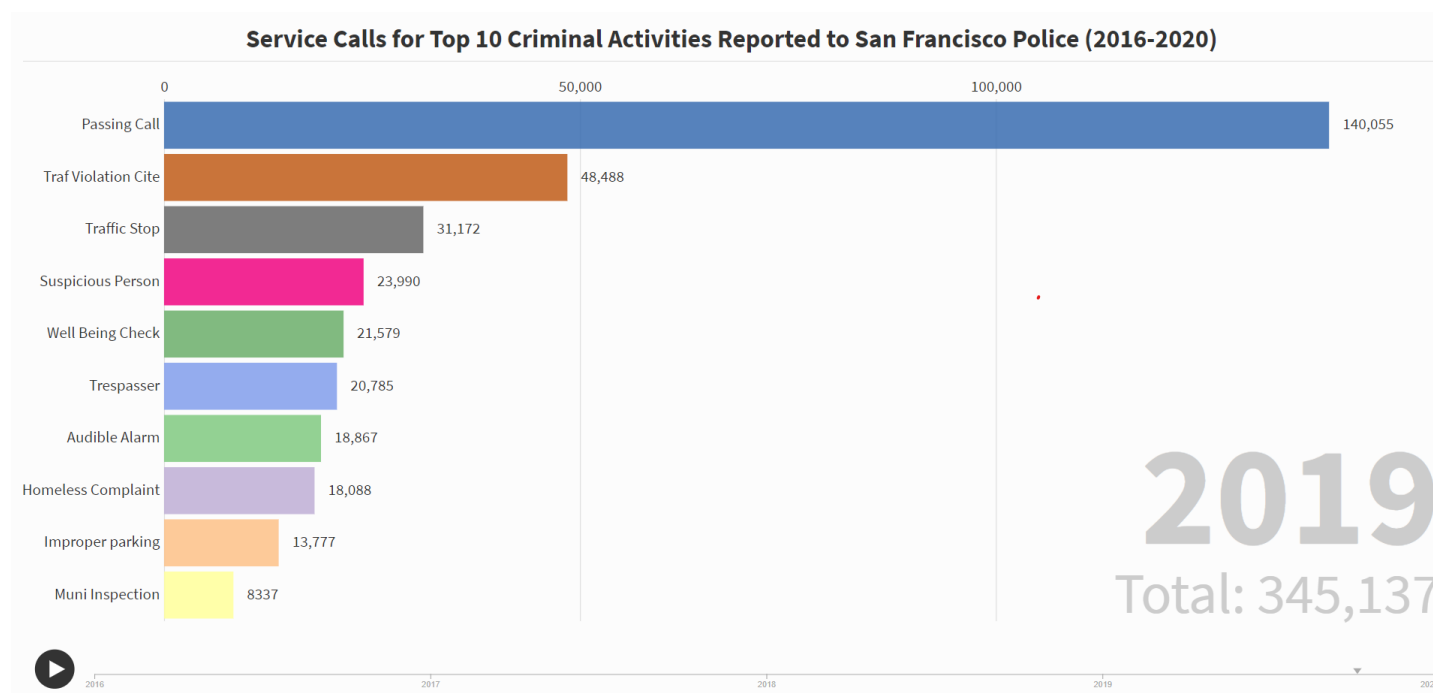


Figure 2: Racing Bar Chart – Service Calls for Top 10 Criminal Activities Reported to San Francisco Police (2016-2020)

[https://preview.flourish.studio/4514482/7L\\_aar4pzEFE4ox7nFFxJtBIB6uI0tqCoFCrN1cK5UxfC8rjSri\\_mp4I\\_mckKpsAo/](https://preview.flourish.studio/4514482/7L_aar4pzEFE4ox7nFFxJtBIB6uI0tqCoFCrN1cK5UxfC8rjSri_mp4I_mckKpsAo/)

## STACKED AREA CHART VS STACKED BAR CHART

Below data is extracted from the main dataset and further we discuss the apt visual representation for this data.

| YEAR | Incident Calls at Night | Incident Calls during Daytime | Total Incident Calls |
|------|-------------------------|-------------------------------|----------------------|
| 2016 | 221091                  | 410373                        | 631464               |
| 2017 | 279999                  | 561948                        | 841947               |
| 2018 | 257869                  | 546116                        | 803985               |
| 2019 | 261350                  | 573614                        | 834964               |
| 2020 | 192246                  | 417570                        | 609816               |



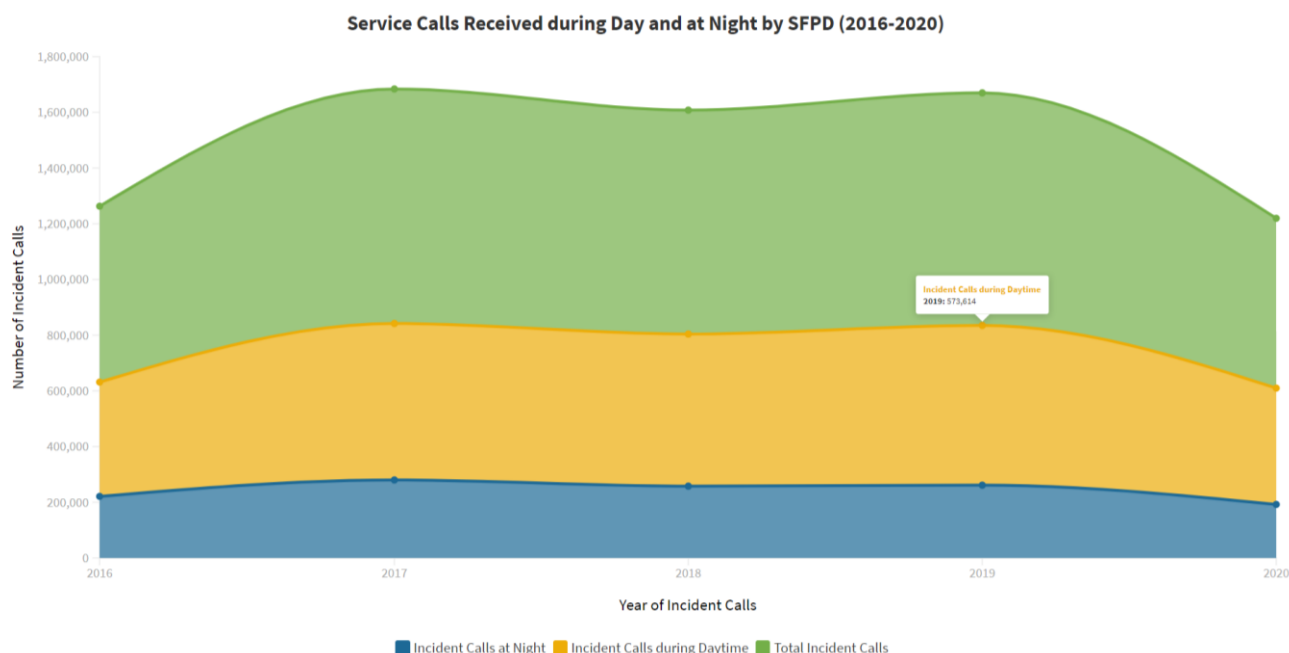


Figure 3: Stacked Area Chart – Service Calls Received during Day and at Night by SFPD (2016-2020)  
[https://preview.flourish.studio/4680629/BfHyBz7nMQqy4MC5RkKI93TaxVq\\_ms5alfG7aEsNckhvjilbfnf\\_q7Fd4zpuauyB/](https://preview.flourish.studio/4680629/BfHyBz7nMQqy4MC5RkKI93TaxVq_ms5alfG7aEsNckhvjilbfnf_q7Fd4zpuauyB/)

Above data can be visualized using stacked area chart or stacked bar graph as one of the main features of this visualization is showing comparison among the number of incident calls during day and at night. However, we justify using stacked area chart over bar chart contrary to the data viz community which finds stacked bar graphs handier as per the research (please see references)

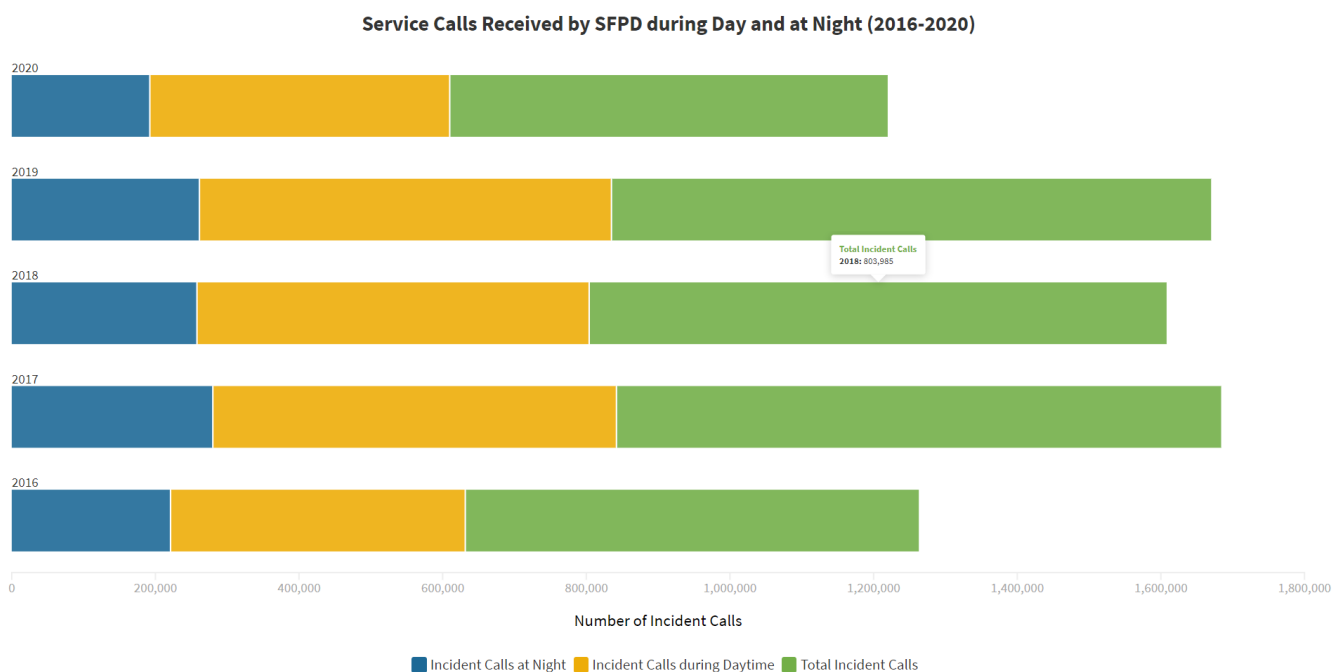


Figure 4: Stacked Bar Chart – Service Calls Received by SFPD during Day and at Night (2016-2020)  
<https://preview.flourish.studio/4505393/uhtpRlFRRuG4hrIYITkEdMrXJwk4K6Lo07GDBKq6D77Hj0uFLAogxB-TIFVdi1V/>

Both the graphs very well show the comparison of values but the reason we prefer stacked area chart is that it gives the visual impression of how largely the values differ not only yearly, but we may compare the areas visually across the graph. For example, when we look at data for 2019, without using hover and checking the exact value plotted we can assume that the daytime calls are at least double than calls at night. Now the bar chart also provides that insight if we compare the length of stacked bars but when we must look at how values developed in between 2018 and 2019, only the stacked area chart can do that, making the comparison easier in terms of the entire timeline and not just points in time. Moreover, area chart makes it easier to compare different proportions i.e., if we must compare total calls and calls at night, it seems easier to do so from this visualization over comparison of length of bars (first bar vs third bar).

## DESIGN CHOICES AND INTERACTIVITY

The racing bar chart comes with set colour palettes in Flourish, having selected the 'accented hues' which in my opinion are the right mix as it fills dark colours for higher valued bars and lighter colours for lower valued bars. The graph is sorted from highest to lowest in terms of incident calls received by SFPD. This helps to know which crimes are the most reported in which year. The colours used for the area chart – yellow and a darker shade of blue – represent the time of the incident calls day and night respectively and the green colour represents the total of the calls as it is a combination of yellows and blues thus the total oof calls received during daytime and at night.

The labels are kept horizontal over diagonal due to visual appeal and simplicity of the graphs. Header is centered and legends, if added, are on the footer of the graph to manage the space well and for the overall visual to look neat and not over-cluttered. Hover text can be seen instead of direct labelling for reading exact values. For the racing bar chart labelling is done on the left-hand side as the length of labels is more than the size of the bars.

The interactivity of these visualizations makes them shareable, readable, user-friendly, and very easily understandable. It highlights various insights being derived from the same visualization at different points in time as compared to static bar graphs which show statistics at just one instance and do not cover how the values developed over time.

## TOOLS

Python was used for cleaning the dataset. Pandas library was useful in doing transformations in data and for creating interactive and animated visualization Flourish was used as the visualization tool. Data is imported from Python to Flourish for creating visualizations. For creating a word cloud in a mask, libraries Numpy, Matplotlib, WordCloud were used.

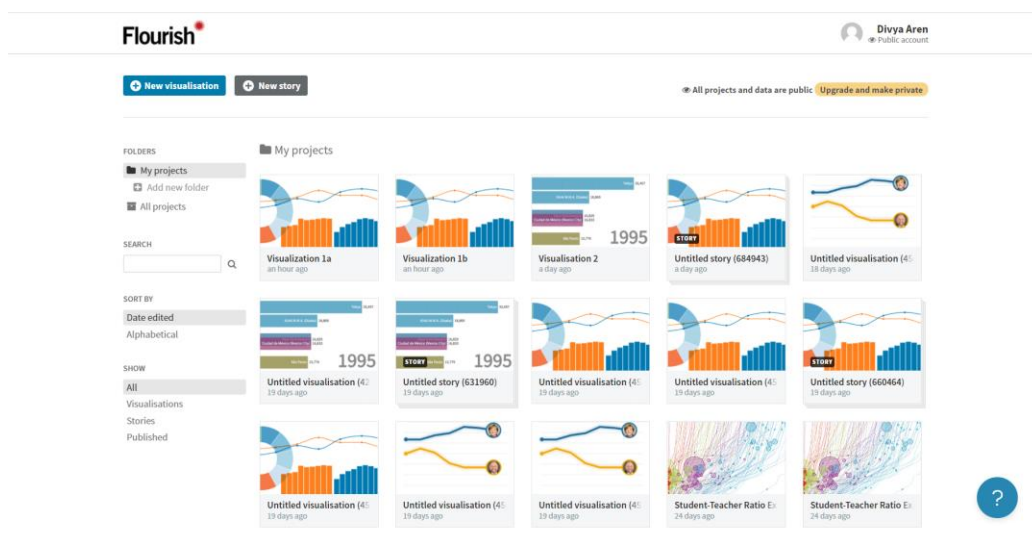


Figure 5: Flourish Projects Snapshot



The area chart could be improved by placing labelling inside the chart instead of labelling through legends that makes the viewers read the chart faster. However, with Flourish we could just find the option of labelling on the right side of the chart or with legends, and I used legends and representative colour scheme as an alternative to labelling inside the chart area.

**“In good information visualization, there are no rules, no guidelines, no templates, no standard technologies, no stylebooks.. You must simply do whatever it takes.”**

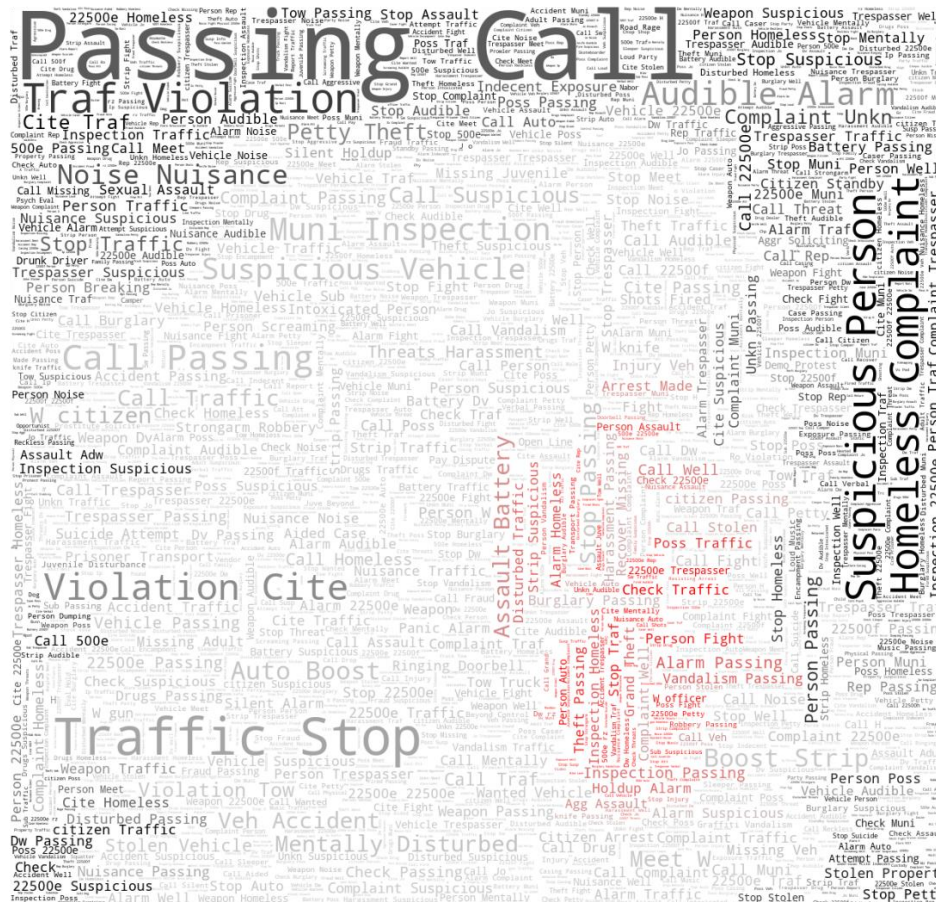


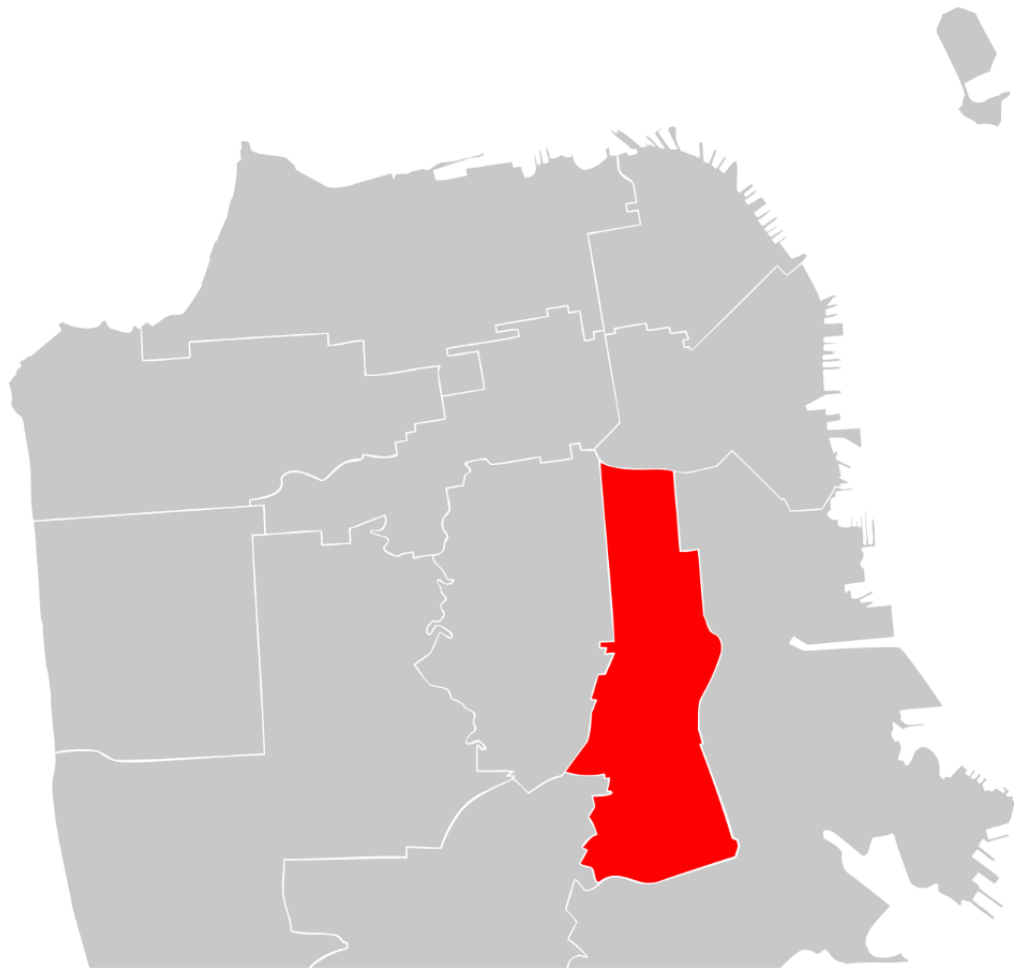
Figure 6: Word cloud on San Francisco map mask





# REFERENCES

1. Dataset: <https://sf.gov/>
2. Flourish Tutorial: <https://www.youtube.com/watch?v=E85arQdPNfs>
3. Resources supporting design choices:
  - <https://academy.datawrapper.de/article/128-what-to-consider-when-creating-area-charts#:~:text=Use%20area%20charts%20only%20if,as%20important%20as%20its%20shares>
  - <https://chartio.com/learn/charts/area-chart-complete-guide/>
  - <https://pandable.co/resources/bar-chart-races#:~:text=Finally%2C%20the%20bar%20chart%20races,of%20interest%20with%20the%20race>
4. Image used as word cloud mask: <https://www.cleanpng.com/png-san-francisco-district-4-san-francisco-district-10-3190232/>



*Figure 7: San Francisco map used as mask for word cloud*

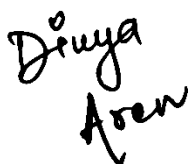
# DECLARATION ON PLAGIARISM

*This form must be filled in and completed by the student submitting an assignment*

|                            |                    |
|----------------------------|--------------------|
| <b>Name:</b>               | Divya Aren         |
| <b>Student Number:</b>     | 20210762           |
| <b>Programme:</b>          | MCM                |
| <b>Module Code:</b>        | CA682              |
| <b>Assignment Title:</b>   | Data Visualization |
| <b>Submission Date:</b>    | 18 Dec 2020        |
| <b>Module Coordinator:</b> | Dr Suzanne Little  |

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I have read and understood the Assignment Regulations. I have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged, and the sources cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

I have read and understood the referencing guidelines found at <http://www.dcu.ie/info/regulations/plagiarism.shtml>, <https://www4.dcu.ie/students/az/plagiarism> and/or recommended in the assignment guidelines



Name: Divya Aren  
Date: 18<sup>th</sup> December 2020