New York City (NYC) 311 Dataset

Group Project for the course IE6600 – Computation and Visualization

Group Number: 3

GROUP MEMBERS

Saravanan Arumugam

Satyajit Lanka

Shriram Vijaykumar

Varun Kumar Kumaravel

Chapter 1 – Introduction

311 NYC is a non-emergency service number that people in New York City can call to report incidents or request city services. The service is available 24 hours a day, 7 days a week, and is meant to be an easy and convenient way for people to get in touch with the city government. Examples of the types of things that people can report or request through 311 include noise complaints, graffiti removal, and requests for tree trimming. In addition to calling the 311 number, people can also use the 311 website or mobile app to submit service requests. Each record in this dataset refers to a 311 call. Date and Time, as the names suggest, refer to the data and time of the 311 call. Department Abbreviation and Department Name refer to the department that the call was handled by. Service Name describes the nature of the call content (e.g. Bulky Item Pick-up, Trash Container Services, or Animal Service Centers). The Call Resolution categories how the call was handled (e.g. call was transferred, information was given to the caller, or a service request was processed).

The 311 system is designed to make it easier for people to get in touch with the city government and request services or report incidents. Some examples of the types of things that people can request or report through 311 include:

- Noise complaints
- Street light outages
- Graffiti removal
- Potholes
- Illegal dumping
- Request for tree trimming

In addition to providing a convenient way for people to access city services, the 311 system also helps the city to track and manage requests and incidents more effectively. For example, the 311 website and mobile app provide real-time updates on the status of requests, and the 311 system also allows the city to analyze data on the types and locations of requests to identify trends and areas that may require additional resources.

Chapter 2 - Summary of Results

Our main objective of the project is to create a comprehensive interactive dashboard to visualize the 311 dataset of New York City and to evaluate the impact of Covid-19 on various municipal complaints. Finally, to compare multiple variables affecting those complaints in different periods of timeline.

We have used Tableau to bring out the insights of the 311 dataset like the different boroughs where there are more than 470 complaints. So, every complaint was explained in different parts of each borough.

Our dataset consists of more than 30 million data. Since it is impossible to clean the data using Google Co-lab and Jupyter Notebook, we have used Big Query to clean and process the data. It was a complex step, because it is not easy to clean 31 million data within a minute. It took hours to completely clean the data. Once, the data is cleaned, we have connected the data to Tableau.

Our project consists of two dashboards and there are six research questions that utilize the same dataset to bring out different visualizations. The first dashboard gives the NYC's area covered in our study which is the average response time that is based on the complaint types and the responsiveness of the Agencies before and during COVID pandemic.

The 2nd dashboard answers the questions - "Did the COVID variants impact the volume of 311 service requests" and "Did New Yorkers sleep at night".

Chapter 3 - Data Sources

We analyzed a subset of more than 31 million service requests logged in the NYC 311 Service requests from 2010-Present dataset, which is made publicly available via NYC Open Data's Socrata API.

NYC311 is the non-emergency New York City call center providing the public 24/7 access to city services and government information. The calls cover a broad range of topics with request encompassing more than 470 complaint types, including everything from rodent sightings, missed trash collection to hot water issues and street/sidewalk repairs.

We queried the NYC 311 dataset with Big Query, to filter the data down to the years 2017-2021, and to reduce the original 41 columns to only the most relevant features. The resulting dataset has 12 million and 25 columns.

Below I have attached the public dataset web URL of New York City 311 Service requests:

https://data.cityofnewyork.us/Social-Services/311-Website/b7y6-82dk

Census data for NYC Population:

https://www.census.gov/library/stories/state-by-state/new-york-population-change-between-census-decade.html

Cleaning tools that we have used in this project are:

1. Google Co-lab:

   In Google Co-lab, we have made some visualization such as bar plots, pie charts, removing null values, and dropping the duplicate values.

2. Google Big Query:

   With the help of Big query, we have resolved the issue of Timestamp, which is one of the biggest issue that we faced while querying the data.

   One of the biggest use of Google Big Query is it acts as a data warehouse. We have stored 18GB size of data.

3. Tableau Prep:

   With the use of Tableau prep, we came up with different visualizations and mainly used for analyzing the data in multiple ways.

Chapter 4 – Results and Methods

Code Snippet that we have used in Google BigQuery to store the data:

```
## to create final_dataset table from public dataset with the relevant columns
SELECT
  unique_key, created_date, closed_date, agency, agency_name, complaint_type, descriptor, location_type,
incident_zip, city, resolution_action_updated_date, open_data_channel_type,location
FROM
  `bigquery-public-data.new_york_311.311_service_requests`;

## adding a column in the final dataset
ALTER TABLE `ie-6600-comp-and-viz-project.311dataset.final_dataset`
ADD COLUMN  covid_status STRING;

## DEleting rows to filter the data for 32+32 months of pre and during covid
DELETE `ie-6600-comp-and-viz-project.311dataset.final_dataset`
WHERE created_date < "2017-04-01 00:00:00";

## Adding value to covid_status column pivoting on 1-1-2020
UPDATE `ie-6600-comp-and-viz-project.311dataset.final_dataset`
SET covid_status = "Before"
WHERE created_date < "2020-01-01 00:00:00";

UPDATE `ie-6600-comp-and-viz-project.311dataset.final_dataset`
SET covid_status = "During"
WHERE created_date >= "2020-01-01 00:00:00";

## agency wise complaint count
SELECT agency, agency_name, count(*) FROM `ie-6600-comp-and-viz-project.311dataset.final_dataset`
group by agency, agency_name
order by count(*) DESC;

## Remove Null
DELETE `ie-6600-comp-and-viz-project.311dataset.final_dataset`
WHERE complaint_type IS NULL
OR agency IS NULL
OR resolution_action_updated_date IS NULL
OR open_data_channel_type IS NULL;


## to create the cumulative_complaints_list table
SELECT DISTINCT(complaint_type) AS Complaints, count(*) AS occurance,
round(count(*)/(SUM(count(*))over())*100,3) AS percent,
FROM `ie-6600-comp-and-viz-project.311dataset.final_data`
GROUP BY
  Complaints
ORDER BY
 occurance DESC
```

Code Snippet of Python for cleaning the data below:

```
Import pandas as pd
Import numpy as np
Import matplotlib.pyplot as plt
Import seaborn as sns
Import datetime
Import scipy.stats as stats
Import statsmodels.api as sm
Import warnings

Df=pd.read_csv("data.csv")  #reading the data file

Display(df.head()) #printing the first 5 rows from df

df.shape #displaying the dimensions of the dataset

#converting columns created_date and closed_date to datetime format
df.created_date = pd.to_datetime(df.created_date)
df.closed_date=pd.to_datetime(df.closed_date)

#calculating the response time in days
df['response_time'] = df.closed_date – df.created_date
df['response_time'] = df['response_time'] / np.timedelta64(1.'D')

#remove the records with invalid response times( less than zero)

df=df[df['response_time'] > 0]

# this is the required file that we have used in tableau for creation of a visualization in our
dashboard.
df.to_csv('311_clean_data.csv', index=False)
```

The entire code can be found in the link below:

https://colab.research.google.com/drive/1GbmTisqi0llo9KNnVRuGWr9ln5QA2TsP?usp=share_link

For the tableau visualization, we have come up 6 research questions. The first question: “Does the infamous Pareto Principle apply to the 311 use case?” The graph below will provide us with the answer.

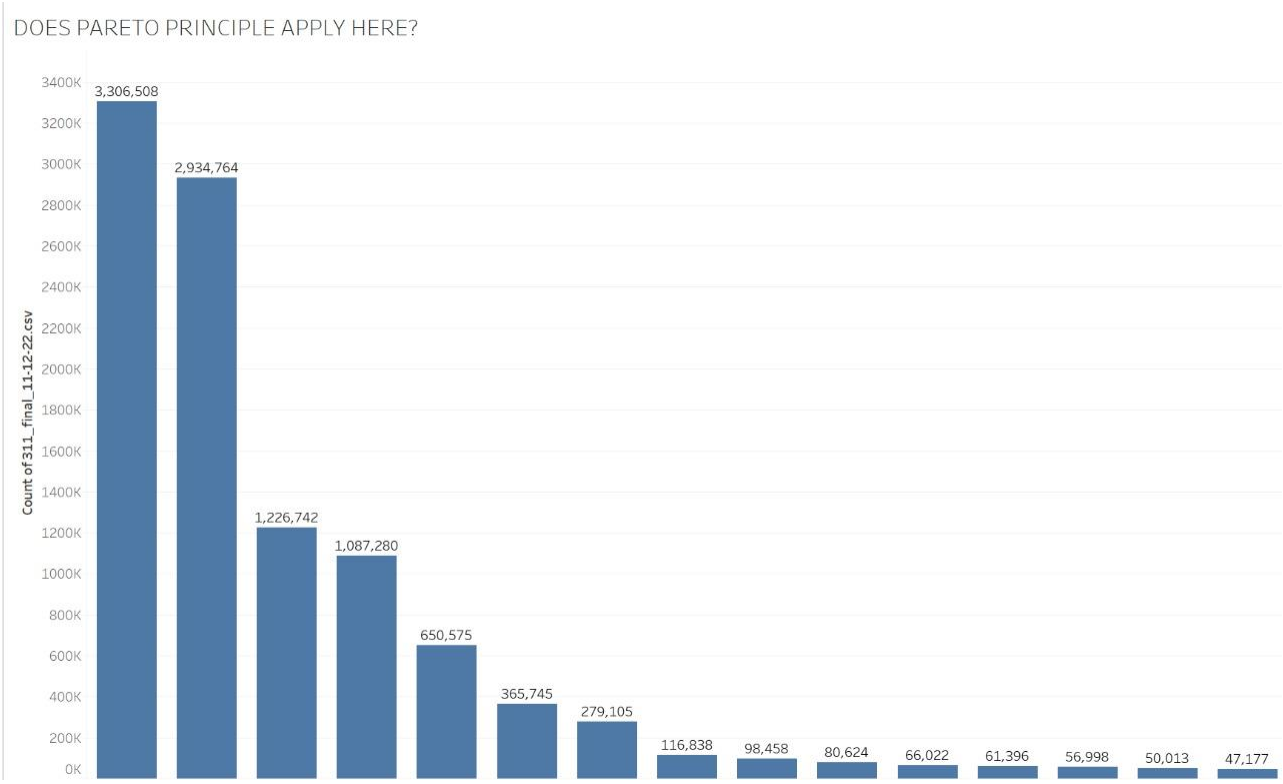

Fig. 1 - Depicts the visualization of number of complaints for different agencies.

In the context of 311 dataset, Pareto Principle can be defined as “80% of number of complaints raised falls under 20% of types of complaints”.
Here you can see, the first 4 complaints type accounts for approximately 8 million complaints which accounts for little more than 80% of the total complaints which is 11 million. Thereby, with the help of the visualization, it is clear that the Pareto Principle is applicable here.

The second research question that we are focusing on: "Is population directly correlated with the volume of 311 complaints?" The graph below will provide us with the answer.

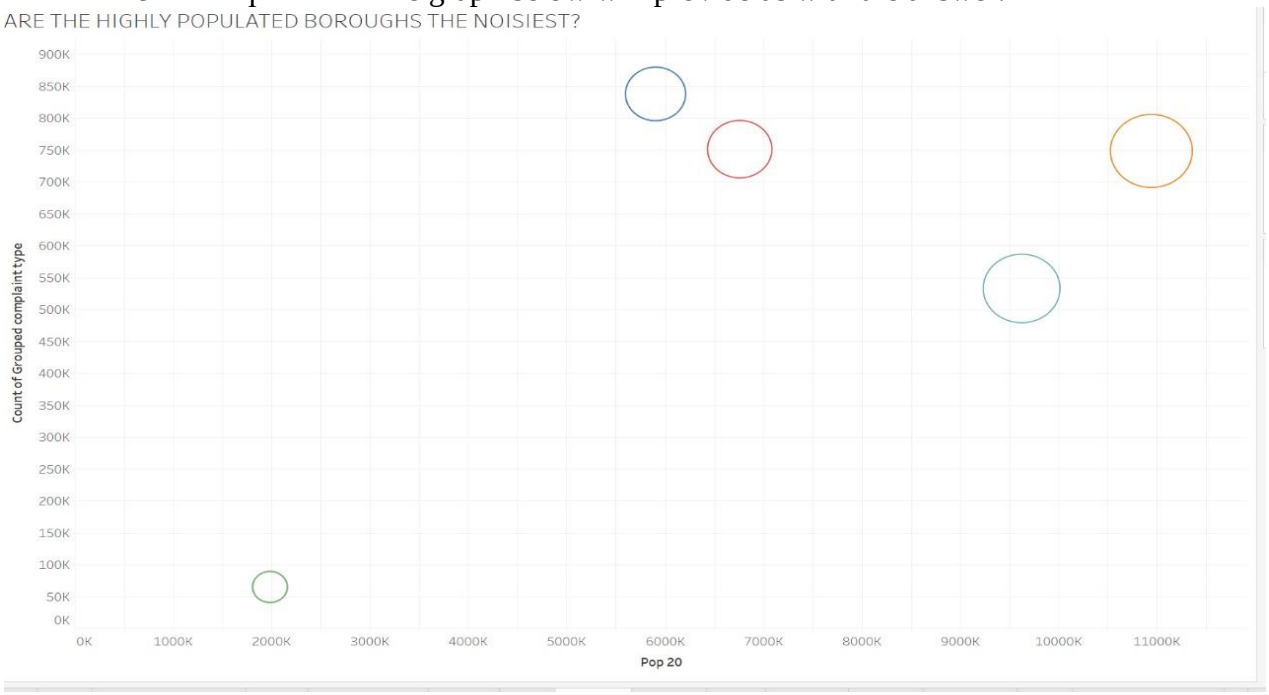

Fig 2. Scatter plot between population and number of complaints

The above graph is a scatter-plot between population and number of complaints. It is easier to assume that higher population would correlate with more complaints but through this visualization, we can infer that it's not the case. The highest number of noise complaint were recorded in Bronx – which is the second lowest populated Borough.

The 3rd research question that we are focusing on "Do people become more tolerant during holiday season?".

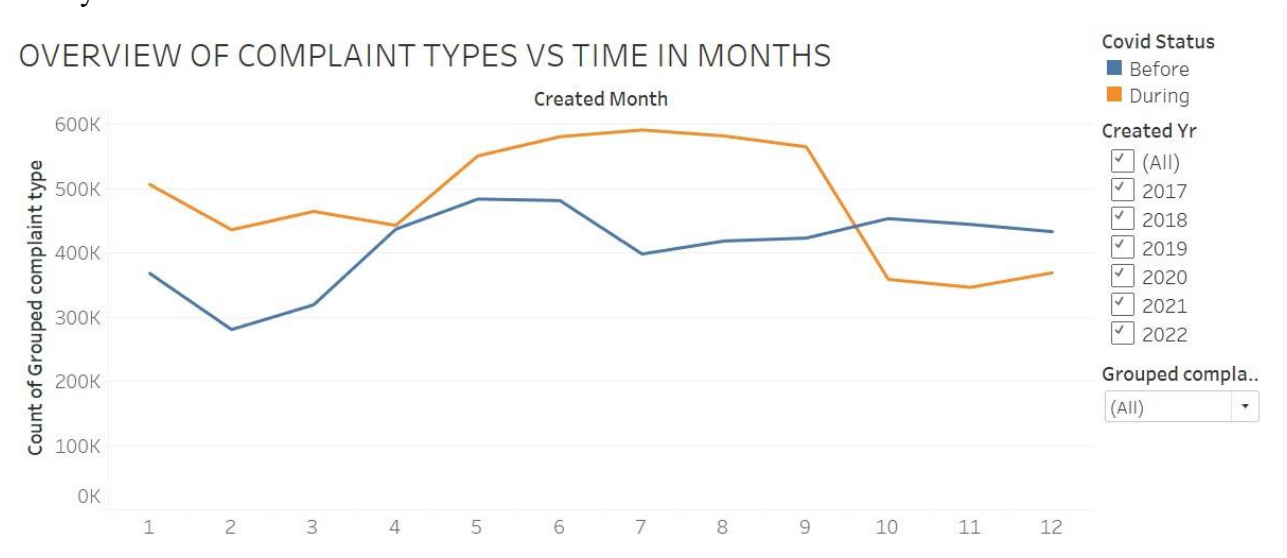

Fig 3. Line-plot between the complaint types of each month during and before COVID

This graph talks about the trend of complaints against months. We can clearly see that there is a rise in the number of complaints during COVID-19 than before except for the months of October,

November and December. Through this, we may infer that people become more tolerant during holiday season but correlation is different from causation. Given the dataset and the context, we cannot confirm that people are not tolerant during holiday season, since there can be a multitude of reasons for the drop of complaints during those months.

The 4th question that we are focusing on: "Did COVID actually impact the efficiency in NYC agencies service? ".

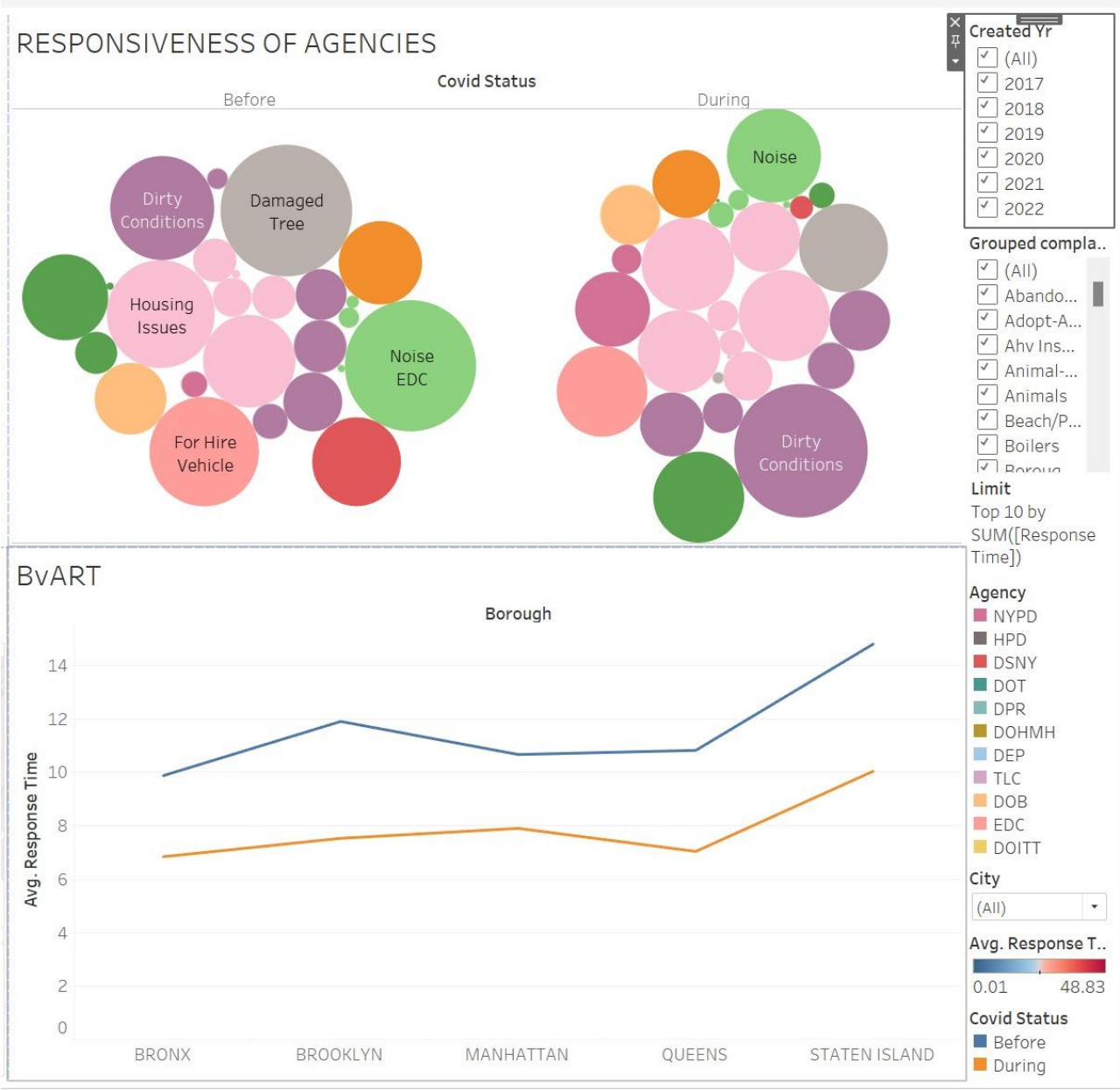

Fig.4 – Bubble plot and line plot of complaints from each borough

From the above graph, we can infer that there is a clear drop in the time taken to respond to the complaints. So, we can conclude that the agencies like NYPD (New York Police Department), DCA (Department of Consumer Affairs) have reacted much faster during COVID than pre-COVID.

The 5th research question that we are focusing on: “Did COVID impact the volume of 311 requests?”.

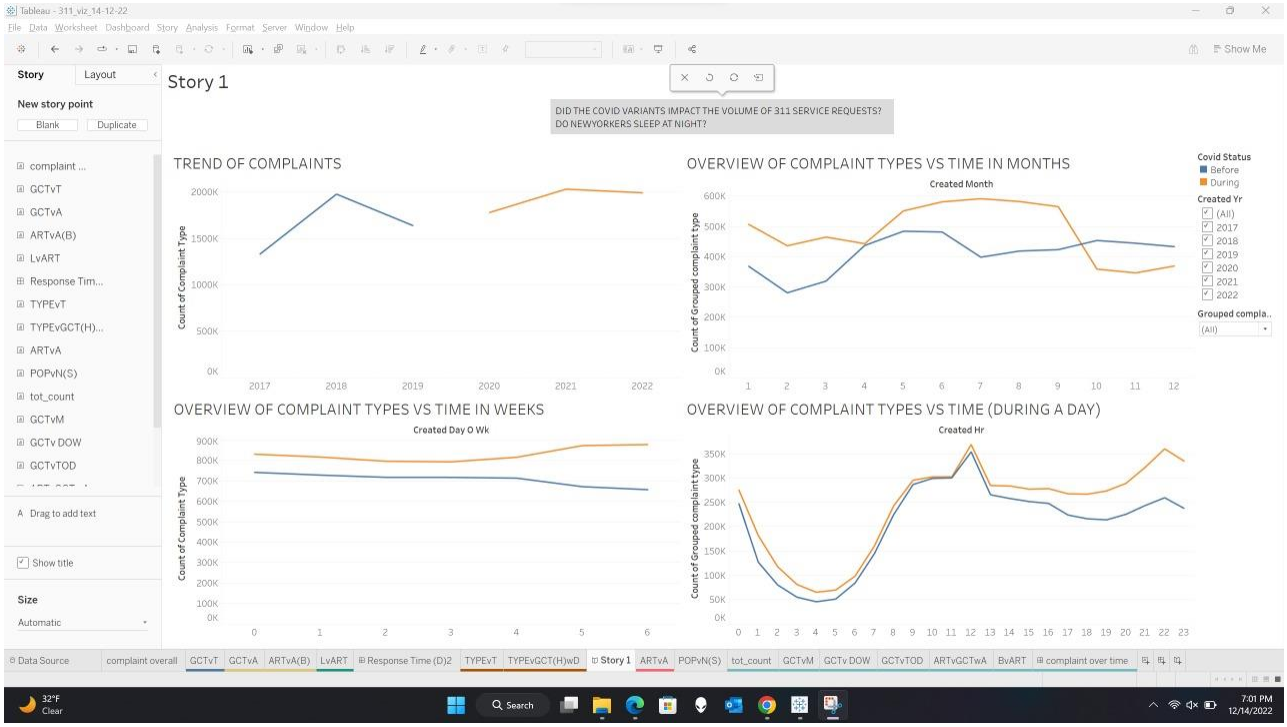

Fig.5 – Line plots of number of service requests before COVID and during COVID

It is clear from the first graph that there is a rise in the number of service requests during COVID than pre-COVID. Yes, it does make a moderate impact on the volume of 311 requests.

The 6th research question that we are focusing on: “Do New Yorkers sleep at night?”.

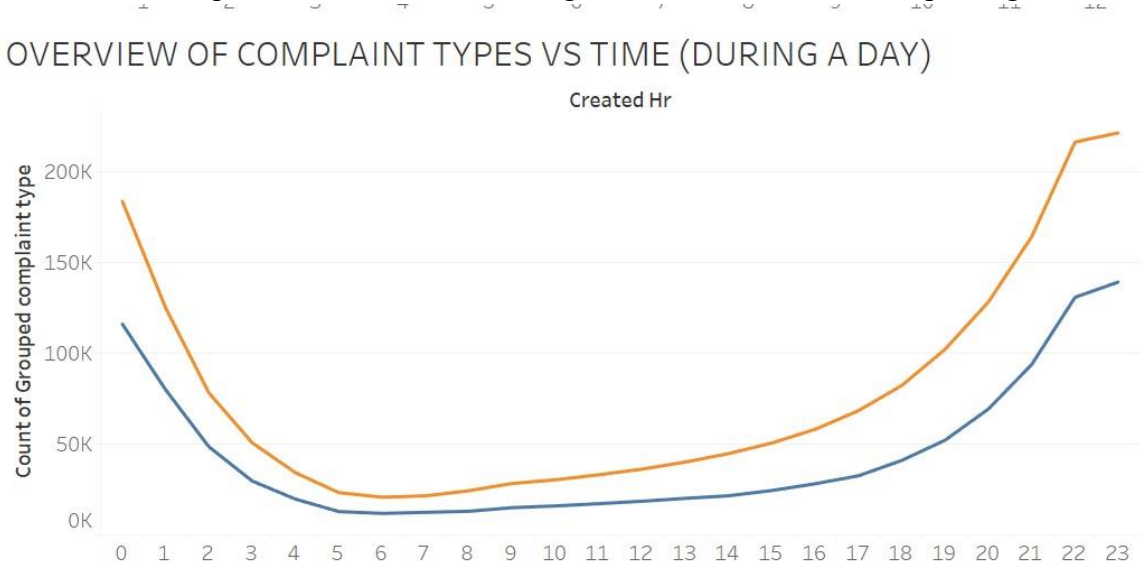

Fig. 6 – Line plot of number of complaints in each hour during and before COVID

From the graph, we can infer that after 8 p.m., the noise complaints increased which indicates

that New Yorkers stay up late night complaining.

Please find below one sample dashboard:

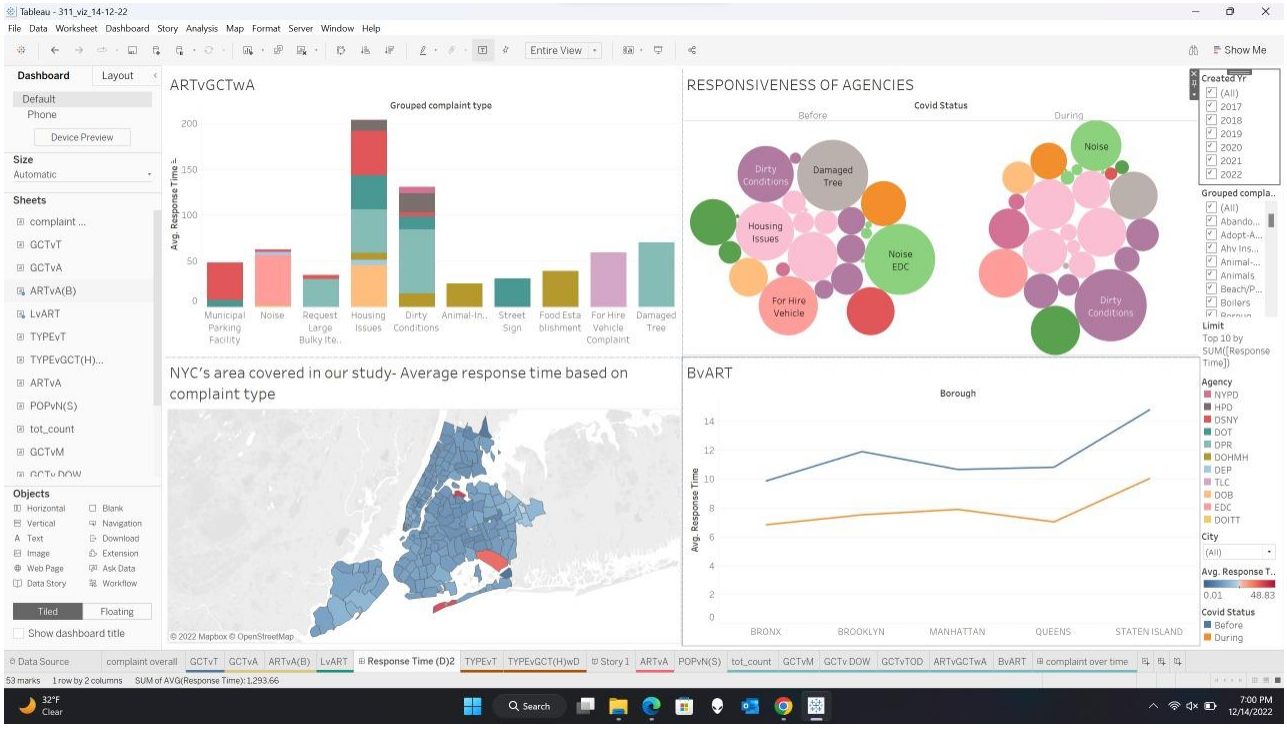

Fig. 7 - This dashboard talks about responsiveness of agencies against complaint types, boroughs and zip codes for the years 2017-2022. We can play around with filters based on the business requirements to find insights.

Chapter 5 – Limitations and Future Scope

- Our project gives an overview into the NYC 311 from 2010 – 2022 and also explains the 2017-2022 in-depth analysis of the complaints that are received to the NYC 311.

- However, we plan to extend it to all the years i.e. from 2010-2022, but since the data that we have is more than 30 million which is around 18 GB, it is complicated to process such a massive dataset. So, we did the data pre-processing in Python (Pandas) and reduced the dataset to 2.5GB which is around 12 million data.

- Implement a machine learning based prediction model to predict the number of complaints generated by NYC to aid in allocation of resources more efficiently.

- Compute more data points and parameters with much higher computation power which will result in better in-depth analysis of the 311 dataset.