Group 3 Project Proposal:

"Group 3 Asset Management"

By: Kevin Camacho, Polly Chu

CUNY: Bernard M. Baruch College

CIS 4400: Data Warehousing for Analytics

Professor: Richard Holowczak

December 17, 2021

Group Email Addresses:

Keve.cam@gmail.com

pChu98@gmail.com

Type of 311 Complaint: Residential Noise Complaint

Business Problem: Are residential rent prices influenced by the volume of 311 residential noise complaints?

Narrative Description:
A brand new alternative investment management company called Group 3 is looking to determine rent prices on their properties. During the height of the COVID-19 pandemic, the newly formed company acquired new properties because they inferred that the decline of the residential real estate market in New York was transitory. As a result, Group 3 began their micro and macro research to determine their rent prices, looking at external factors influencing rent prices within NYC.

One of the external factors Group 3 wants to focus on is the recent noise complaints 311 NYC Services have received from the period of 2018 - 2021and whether those residential noise complaints affect rent prices within the area. Therefore, during the research, Group 3 decided to focus on critical key predictors or key performance indicators that will help see if residential noise complaints directly influence rent prices within a neighborhood. Some of the KPIs Group 3 will be focusing on are median rent prices by Zip code and/or neighborhood for the month, 311 complaints type volume in a month. With these KPIs, Group 3 will be one step closer to identifying the possible relationship between median rent prices and noise complaints.

Identifying whether residential noise complaints influence median rent prices within a neighborhood will help Group 3 find a baseline on setting the rent and attract new residents to their properties, promoting growth within their business.

Potential KPIs:
- Median neighborhood rental price to complaint ratio
- Noise Complaint Volume on monthly basis

Grain: Periodic, we will be using a monthly period for this project. Ranging from January 1, 2018 through November 2021.

Citations:

Asher. "311 Noise Complaints: NYC Open Data." *311 Noise Complaints | NYC Open Data*, 28 July 2015, https://data.cityofnewyork.us/Social-Services/311-Noise-Complaints/p5f6-bkga. Accessed 10 Sep. 2021.

"Streeteasy Data Dashboard: StreetEasy." *StreetEasy Blog*, 10 Dec. 2021, https://streeteasy.com/blog/data-dashboard/. Accessed 10 Sep. 2021.

"U.S. ZIP Codes: Free Zip Code Map and ZIP Code Lookup." *UnitedStatesZipCodes*, 2012, https://www.unitedstateszipcodes.org/. Accessed 6 Dec. 2021.

The dimensional model diagram

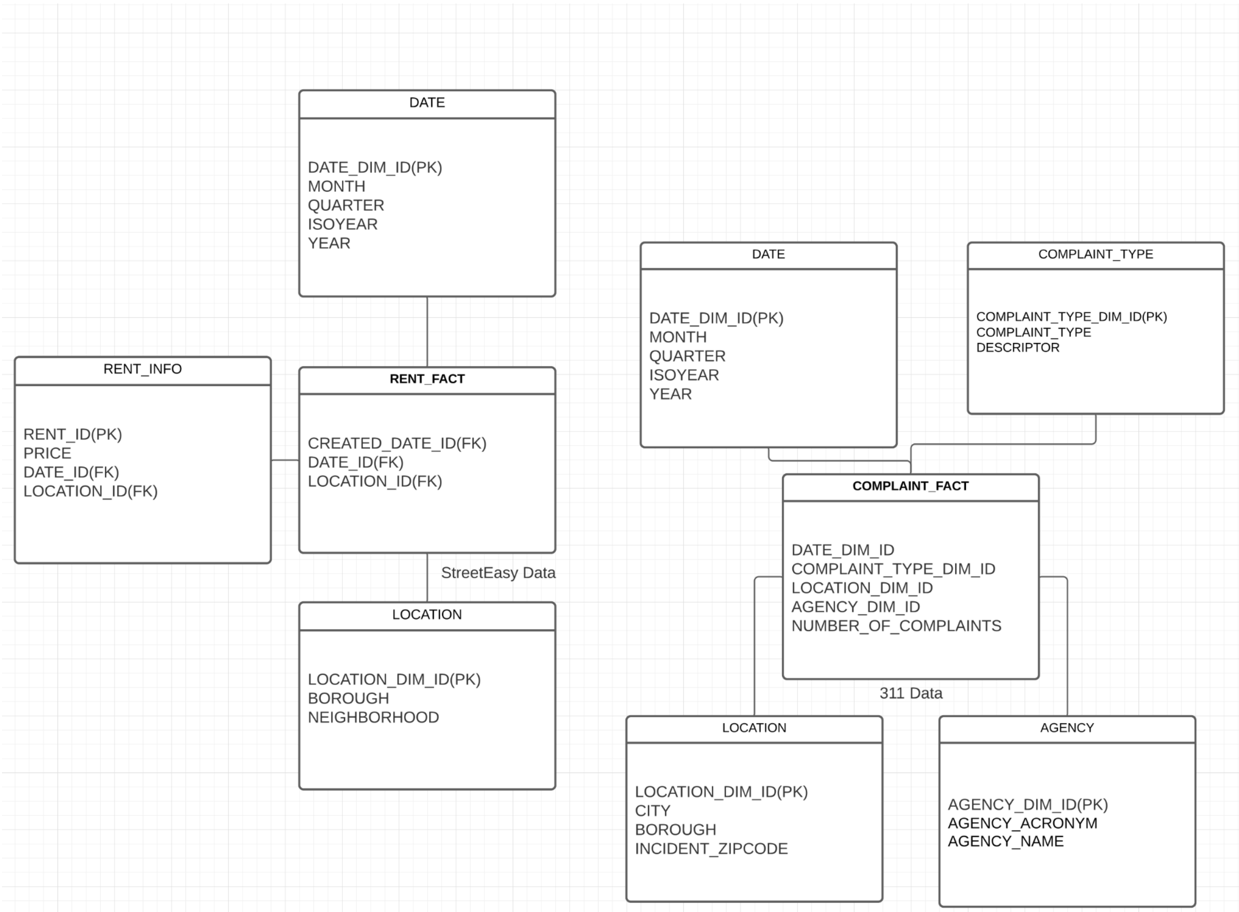
The first two dimensional models on the following page:

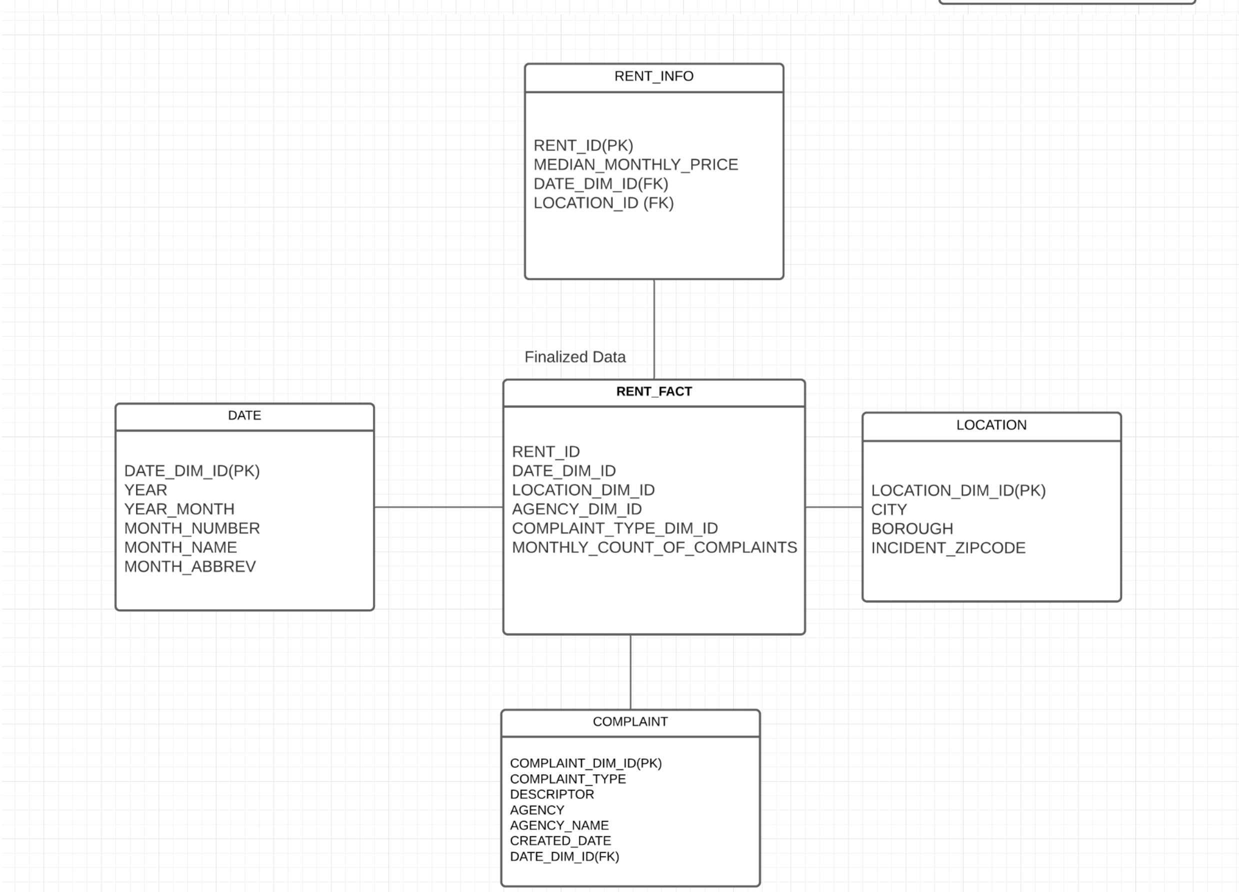
One is labeled for Street Easy, which is the rental data we are going to be extracting which showcases neighborhoods and their median monthly rent over a period of time of 2010-2021.

One is labeled 311 Data, which is the opendata we are going to be extracting which contains the noise complaints and some location information.

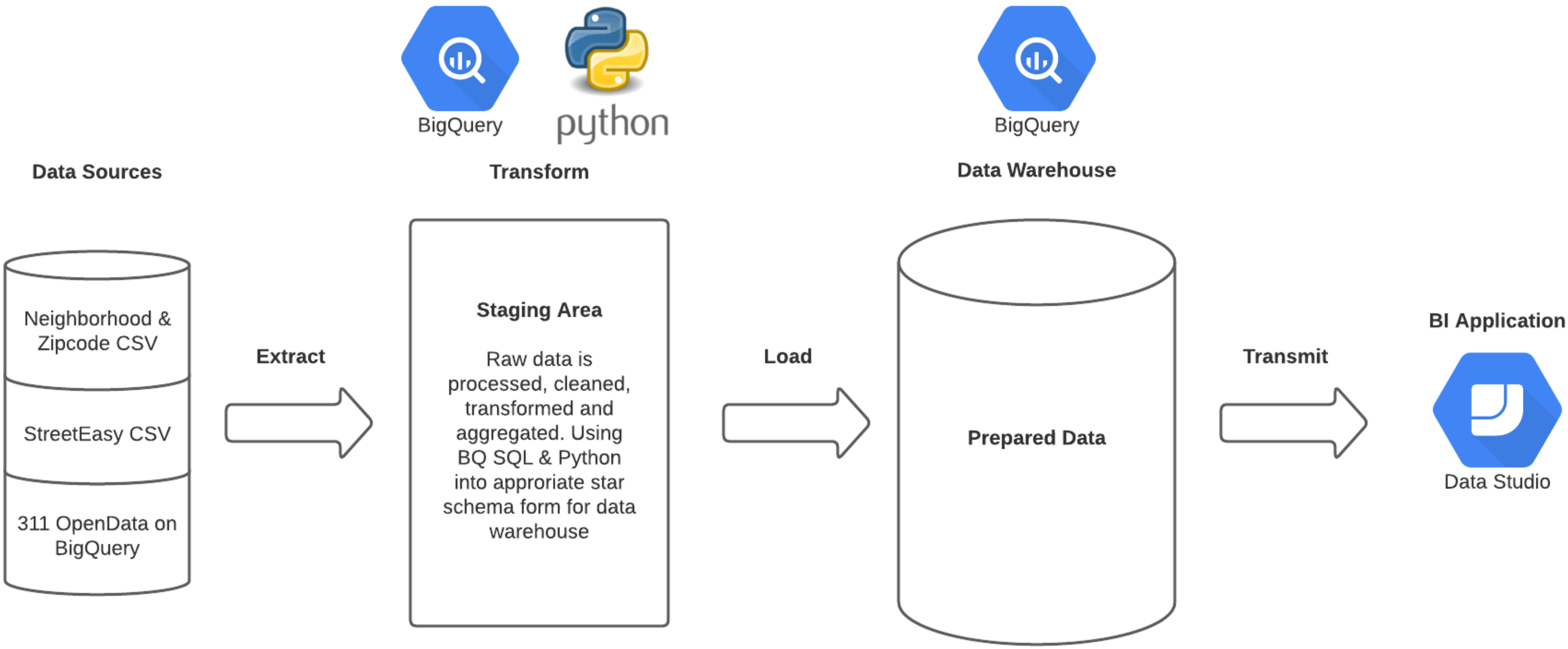
One is labeled finalized data, which is the completed dimensional model diagram for the project.

(viewed on next page)


DATE
DATE_DIM_ID(PK)
MONTH
QUARTER
ISOYEAR
YEAR
RENT_INFO
RENT_ID(PK)
PRICE
DATE_ID(FK)
LOCATION_ID(FK)
RENT_FACT
CREATED_DATE_ID(FK)
DATE_ID(FK)
LOCATION_ID(FK)
StreetEasy Data
LOCATION
LOCATION_DIM_ID(PK)
BOROUGH
NEIGHBORHOOD
DATE
DATE_DIM_ID(PK)
MONTH
QUARTER
ISOYEAR
YEAR
COMPLAINT_TYPE
COMPLAINT_TYPE_DIM_ID(PK)
COMPLAINT_TYPE
DESCRIPTOR
COMPLAINT_FACT
DATE_DIM_ID
COMPLAINT_TYPE_DIM_ID
LOCATION_DIM_ID
AGENCY_DIM_ID
NUMBER_OF_COMPLAINTS
311 Data
LOCATION
LOCATION_DIM_ID(PK)
CITY
BOROUGH
INCIDENT_ZIPCODE
AGENCY
AGENCY_DIM_ID(PK)
AGENCY_ACRONYM
AGENCY_NAME


RENT_INFO
RENT_ID(PK)
MEDIAN_MONTHLY_PRICE
DATE_DIM_ID(FK)
LOCATION_ID (FK)
Finalized Data
DATE
DATE_DIM_ID(PK)
YEAR
YEAR_MONTH
MONTH_NUMBER
MONTH_NAME
MONTH_ABBREV
RENT_FACT
RENT_ID
DATE_DIM_ID
LOCATION_DIM_ID
AGENCY_DIM_ID
COMPLAINT_TYPE_DIM_ID
MONTHLY_COUNT_OF_COMPLAINTS
LOCATION
LOCATION_DIM_ID(PK)
CITY
BOROUGH
INCIDENT_ZIPCODE
COMPLAINT
COMPLAINT_DIM_ID(PK)
COMPLAINT_TYPE
DESCRIPTOR
AGENCY
AGENCY_NAME
CREATED_DATE
DATE_DIM_ID(FK)

A description and screen pictures of each of the ETL processes and any custom code used to process the source data.



ETL/ELT methodology

We will be using BigQuery:Public Data to extract data from NYC OpenData.
We will extract StreetEasy Rent Data from their website as a CSV.
We will create a CSV to extract with neighborhood and zip codes to match up with StreetEasy, as the Street Easy data only has neighborhoods as identifiers.
We will then use Python & BigQuery SQL to format the data we extract appropriately, we will also be using a csv/excel file from StreetEasy to get median rental prices and transform that data as well.
We will then set up our dimensions and fact tables using Big Query.
We will be using BigQuery to store our data.

- **Generating UUID as Surrogate Keys - this is an example repeated for each dimension**

```
CREATE OR REPLACE TABLE `group3-311.dimensions.Date_Dim`
AS SELECT GENERATE_UUID() Date_Dim_ID, * FROM `group3-311.dimensions.Date_Dim`;
```

- **Inverting StreetEasy rent data**

```
import pandas as pd

# Turn this: areaName,Borough,areaType,_2010_01,_2010_02,_2010_03,_2010_04,etc.
# Into this: areaName,Borough,areaType,month,rent

#
areaName,Borough,areaType,_2010_01,_2010_02,_2010_03,_2010_04,_2010_05,_2010_06,_2
010_07,
#
_2010_08,_2010_09,_2010_10,_2010_11,_2010_12,_2011_01,_2011_02,_2011_03,_2011_04,
_2011_05,
#
_2011_06,_2011_07,_2011_08,_2011_09,_2011_10,_2011_11,_2011_12,_2012_01,_2012_02,
_2012_03,
```

```
#
_2012_04,_2012_05,_2012_06,_2012_07,_2012_08,_2012_09,_2012_10,_2012_11,_2012_12,
_2013_01,
#
_2013_02,_2013_03,_2013_04,_2013_05,_2013_06,_2013_07,_2013_08,_2013_09,_2013_10,
_2013_11,
#
_2013_12,_2014_01,_2014_02,_2014_03,_2014_04,_2014_05,_2014_06,_2014_07,_2014_08,
_2014_09,
#
_2014_10,_2014_11,_2014_12,_2015_01,_2015_02,_2015_03,_2015_04,_2015_05,_2015_06,
_2015_07,
#
_2015_08,_2015_09,_2015_10,_2015_11,_2015_12,_2016_01,_2016_02,_2016_03,_2016_04,
_2016_05,
#
_2016_06,_2016_07,_2016_08,_2016_09,_2016_10,_2016_11,_2016_12,_2017_01,_2017_02,
_2017_03,
#
_2017_04,_2017_05,_2017_06,_2017_07,_2017_08,_2017_09,_2017_10,_2017_11,_2017_12,
_2018_01,
#
_2018_02,_2018_03,_2018_04,_2018_05,_2018_06,_2018_07,_2018_08,_2018_09,_2018_10,
_2018_11,
#
_2018_12,_2019_01,_2019_02,_2019_03,_2019_04,_2019_05,_2019_06,_2019_07,_2019_08,
_2019_09,
#
_2019_10,_2019_11,_2019_12,_2020_01,_2020_02,_2020_03,_2020_04,_2020_05,_2020_06,
_2020_07,
#
_2020_08,_2020_09,_2020_10,_2020_11,_2020_12,_2021_01,_2021_02,_2021_03,_2021_04,
_2021_05,
# _2021_06,_2021_07,_2021_08,_2021_09,_2021_10

# Read in the file  (change this file name to your source data file)
df = pd.read_csv('bquxjob_18073950_17db504c0f8.csv')
```

```
# Make a list of the column names of just the year-months
column_list=df.columns.values.tolist()
column_list.remove('areaName')
column_list.remove('Borough')
column_list.remove('areaType')

# Set up the output data frame with the desired columns
output_column_list = ['areaName','Borough','areaType','yearmonth','rent']
output_df = pd.DataFrame(columns = output_column_list)


# Iterate over the rows
for index, row in df.iterrows():
  # Iterate over the names of the year-months
  for yearmonthcol in column_list:
     # Turn the _2021_10 into an integer 202110
     yearmonthint = int(yearmonthcol.replace('_',''))
     # Build a new record
     new_record = dict()
     new_record['areaName'] = row['areaName']
     new_record['Borough'] = row['Borough']
     new_record['areaType'] = row['areaType']
     new_record['yearmonth'] = yearmonthint
     new_record['rent'] = row[yearmonthcol]
     # Append the new record to the output dataframe
     output_df=output_df.append(new_record, ignore_index=True)

# Save the output file
output_df.to_csv('inverted_rent_data.csv', ignore_index=True)
```

- **A CSV with neighborhoods and zip codes within that area was created to use as a geo-location to drill down.**

- **Joining StreetEasy Neighborhood Data with Neighborhood & Zipcode CSV, where a match on Neighborhood is found, we did this because the StreetEasy data does not come with zip code, just neighborhood names, so we used our own CSV with**

neighborhoods and zip codes paired together, to join these tables when it found a match.

```
SELECT a.*, b.*
FROM `group3-311.se_data.zipcode_neighborhood_to_join` AS a
LEFT JOIN `group3-311.se_data.se_rent` AS b
ON a.Neighborhood = b.areaName;
```

*Note**

We did use other SQL code which is essentially Select statements to create the dimensions themselves, this was early on in the project and overlooked that we had to record those queries, so we don't have those queries saved. There is also one query we used to aggregate the noise complaints into a monthly basis instead of the daily format that the source comes in.

Final Dimensional Schema

Complaint_Dim

SCHEMA DETAILS PREVIEW

Table schema

Filter Enter property name or value

Field name	Type	Mode
complaint_dim_id	STRING	NULLABLE
complaint_type	STRING	NULLABLE
descriptor	STRING	NULLABLE
Location_Dim_ID	STRING	NULLABLE
agency	STRING	NULLABLE
agency_name	STRING	NULLABLE
date_dim_id	STRING	NULLABLE
createdDate	DATE	NULLABLE

Date_Dim_CSV

SCHEMA DETAILS PREVIEW

Table schema

Filter Enter property name or value

Field name	Type	Mode
date_dim_id	STRING	NULLABLE
year	STRING	NULLABLE
year_month	DATE	NULLABLE
month_number	STRING	NULLABLE
month_name	STRING	NULLABLE
month_abbrev	STRING	NULLABLE

Location_Dim

SCHEMA DETAILS PREVIEW

Table schema

Filter Enter property name or value

Field name	Type	Mode
Location_Dim_ID	STRING	NULLABLE
Borough	STRING	NULLABLE
Neighborhood	STRING	NULLABLE
Zipcode	STRING	NULLABLE

Rent_Dim QUERY SH

SCHEMA DETAILS PREVIEW

Table schema

Filter Enter property name or value

Field name	Type	Mode
Rent_Dim_ID	STRING	NULLABLE
median_rent	FLOAT	NULLABLE
date_dim_id	STRING	NULLABLE
location_dim_id	STRING	NULLABLE

Rent_fact

SCHEMA DETAILS PREVIEW

Table schema

Filter Enter property name or value

Field name	Type	Mode
date_dim_id	STRING	NULLABLE
location_dim_id	STRING	NULLABLE
complaint_dim_id	STRING	NULLABLE
rent_dim_id	STRING	NULLABLE
Monthly_Complaint_Count	STRING	NULLABLE

KPI Visualizations

The following two maps are showing the overlay of median rent in a neighborhood, and the number of complaints in a neighborhood.
This specific heat map was a live interaction, and you can press play and it will go over time from Jan 2018 - 2021 Nov showing you the changes.

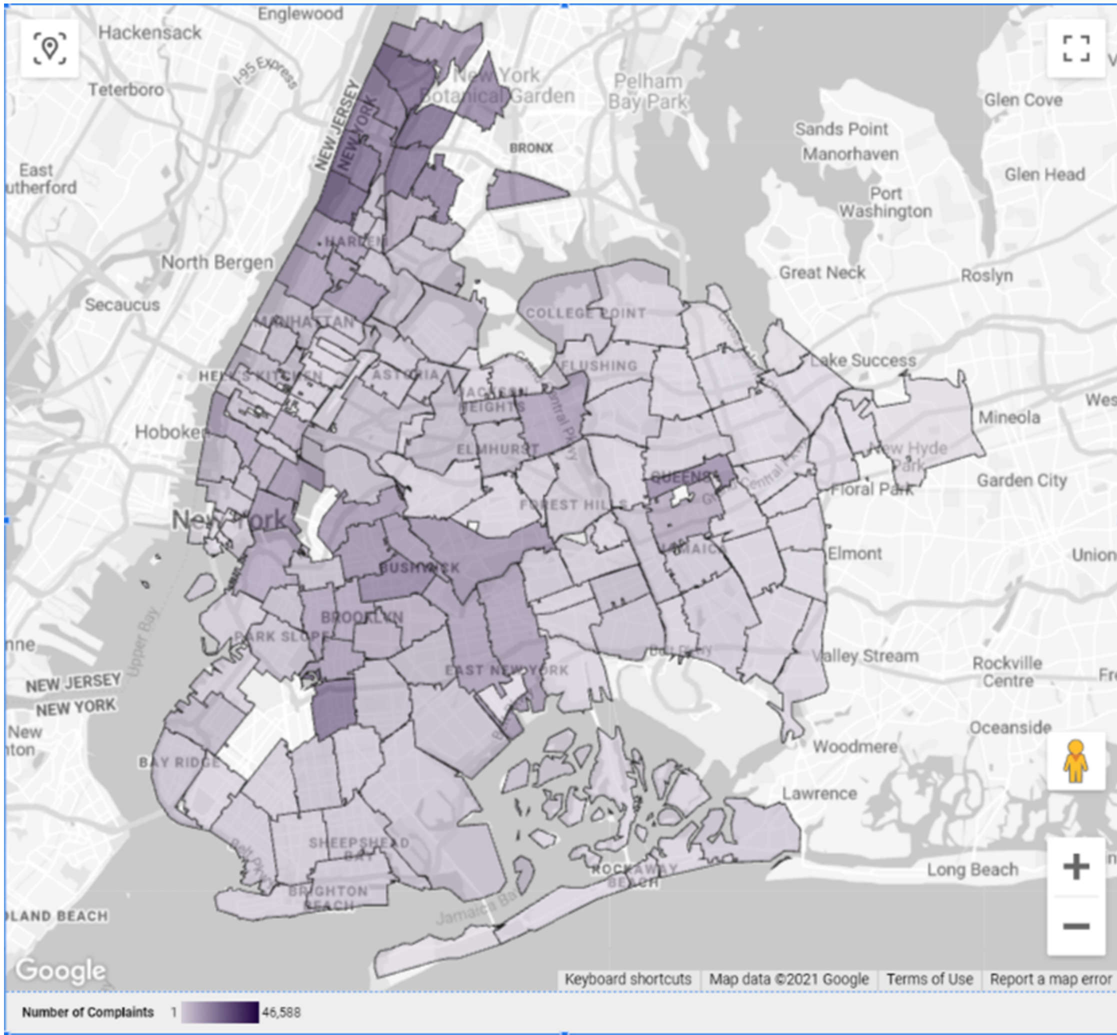
These heat maps are showing the amount of complaints, to the median rent price of a specific neighborhood in a period of time. The legend is on the bottom left of the two maps.

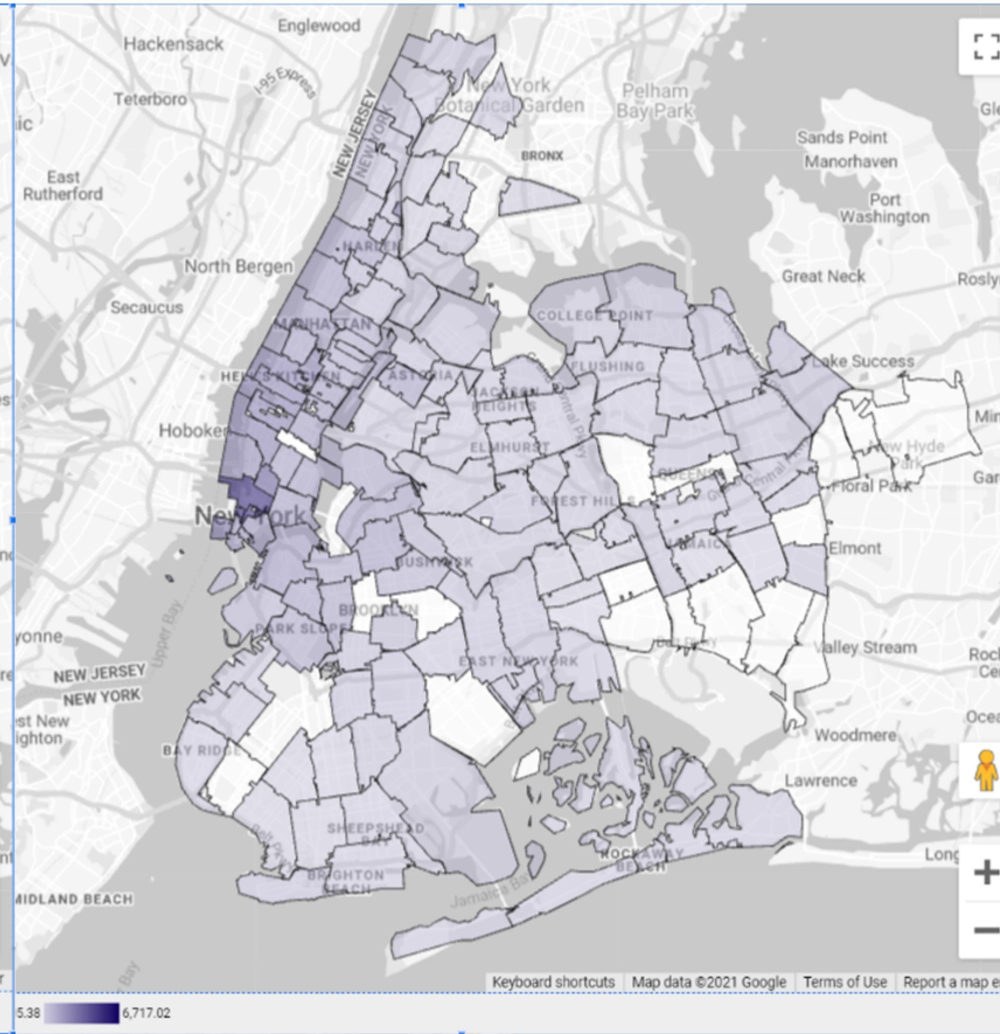
For the purposes of this screenshot, this period of time is Jan 2018.

(viewed on next page)

It is going to be apparent but not as easy to see here, that there is no correlation for the amount of noise complaints and the median rent price of a neighborhood.

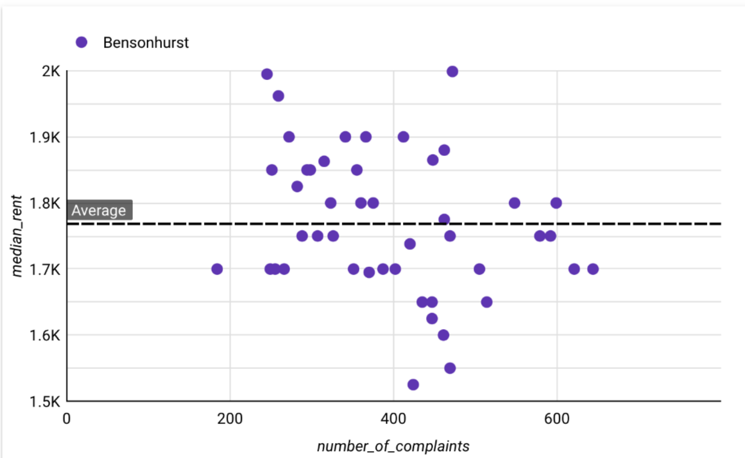
KPI Visualization



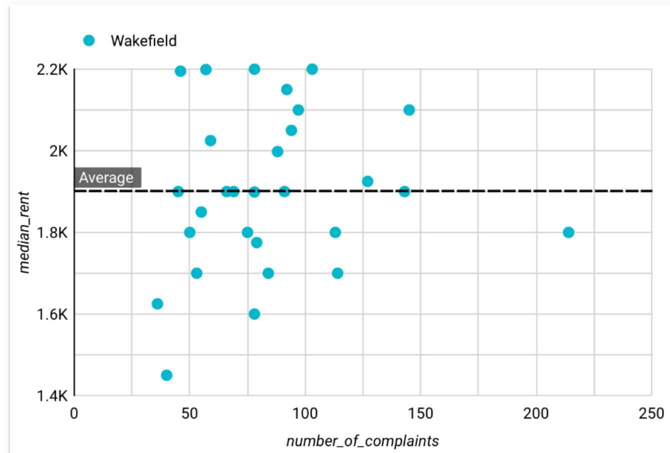


Here we selected 1 neighborhood from each borough that had a potentially interesting distribution. Each point represents a period in time, from Jan 2018 to Nov 2021 for their respective neighborhood. In Bensonhurst, the more reports of noise complaints, the lower the median rent was. For Wakefield, it landed between 40-100 complaints for the month, not many people seem to use 311 there. Rent to Noise complaint was not correlated. Corona was interesting as the more complaints in this neighborhood, the lower the median rent was. People love complaining in the upper east side, and it might be because the median rent is so high, but we still see no correlation here.

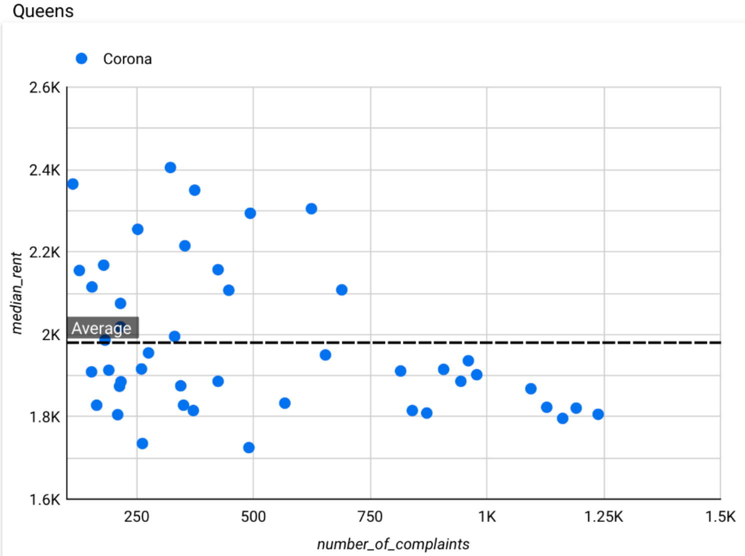
Brooklyn



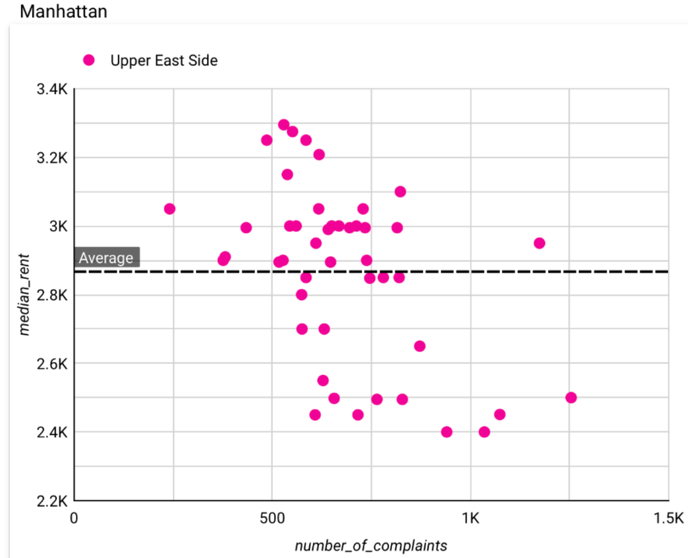
Bronx



Queens

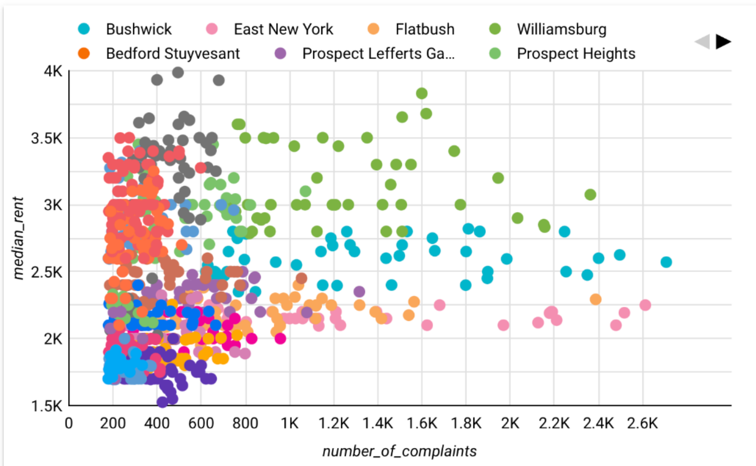


Manhattan

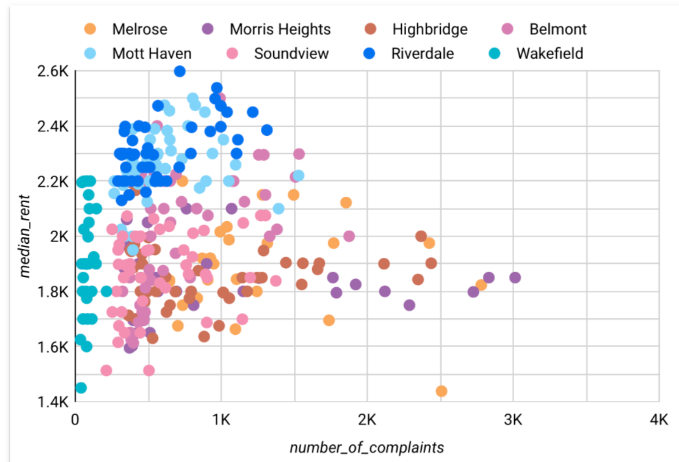


At this level we are looking through the boroughs, and their respective neighborhoods. Each point is a period of time between Jan 2018 - Nov 2021. If we look at the orange bubbles in Manhattan which represent Tribeca, the median rent is high but relatively low complaints here. Williamsburg in brooklyn, which is the dark green point, shows that as the number of complaints increase, the median rent stays relatively the same. Again in Manhattan, the blue bubbles increasing in number of complaints is Washington Heights, the complaints increase by a large margin, and the median rent stays relatively the same there.

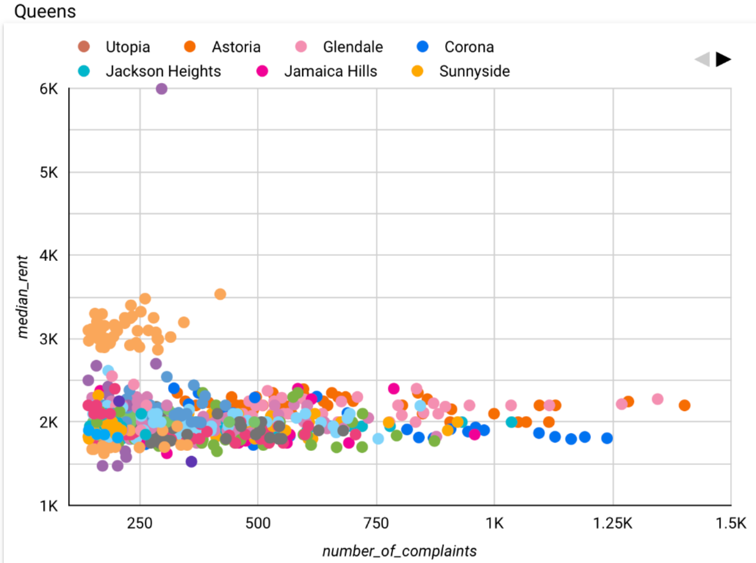
Brooklyn



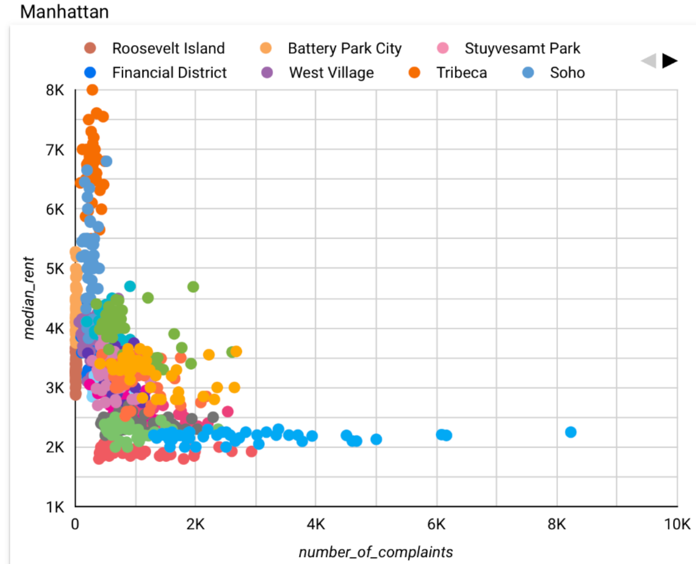
Bronx



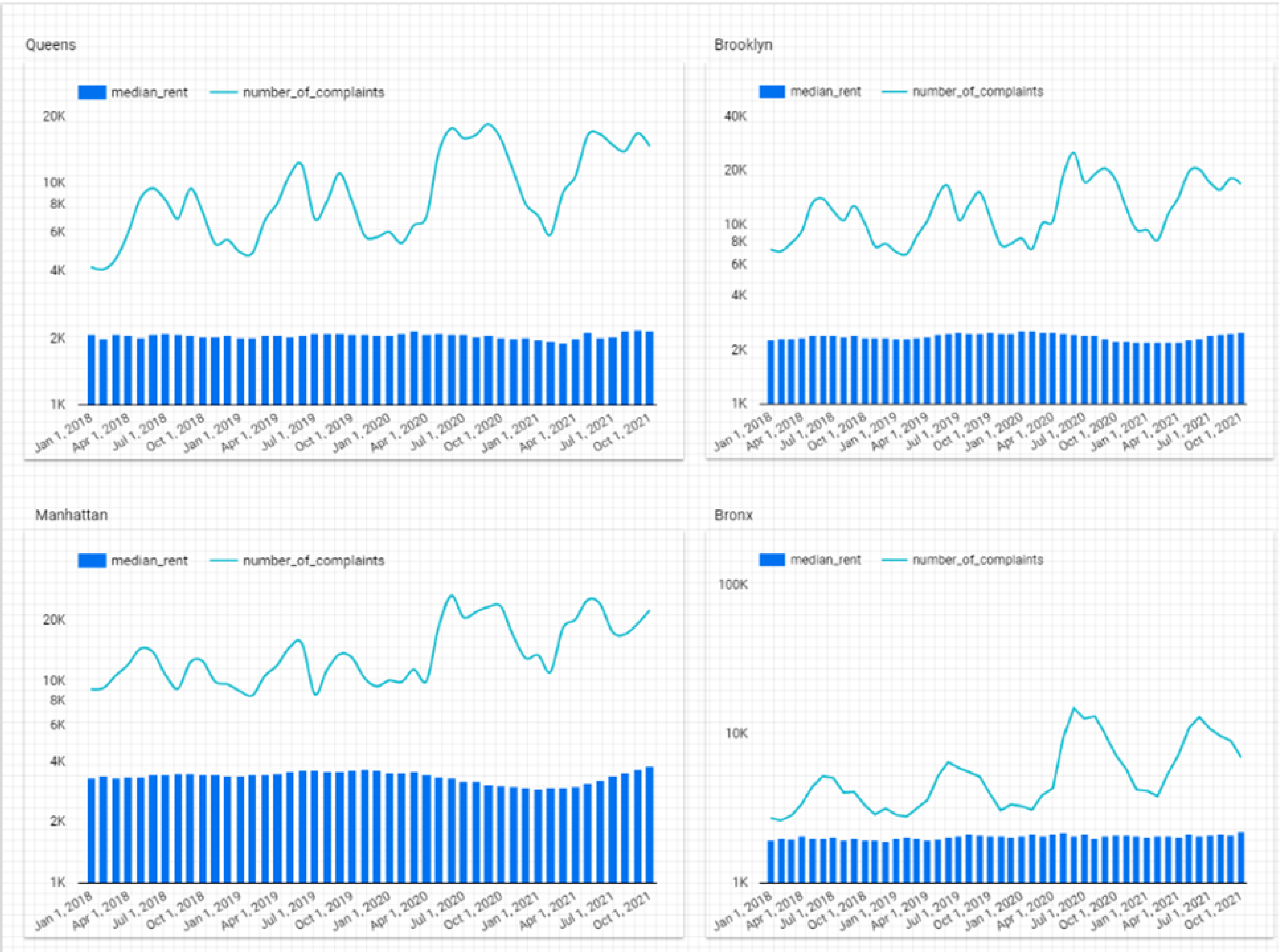
Queens



Manhattan



Here we have a time series graph showing median rent and number of complaints for each month through the period noted on the x-axis. The Y axis for median_rent is in logscale for the rents to show changes more apparently. Number of complaints is at a normal scale. At this level we are seeing if the borough itself has median rent changes based on the number of complaints. From our findings, there are no real changes based on the number of complaints. This trend was apparent at the neighborhood scale as well.



Tools

<u>LucidChart</u>: It is a platform that allows users to draw and share charts and diagrams with others. Users are also able to collaborate with each other simultaneously and revise each other's work. The ETL/ELT diagram will be created here, along with the Dimension/Fact Table model.

<u>Google Big Query</u>: It is a data warehouse that uses SQL. Users are able to share the dataset by creating service keys for other users. We will also be using it as the staging area for our data. This will also extract the 311 Opendata for us.

<u>Google Colab</u>:It is a development environment that allows users to write and execute python code. Users are able to collaborate with other users on the platform as well.

<u>Google Data Studio</u>: It is an online tool that uses our prepared data into customizable visualizations, such as graphs and charts.

Narrative Conclusion

A)

- **Whatsapp**: We used this as our main form of communication to coordinate what to do and when to do it.
- **Zoom**: We met up on Zoom to further discuss the tasks that we had to complete, and it allowed us to work collectively without being together.
- **Google Doc**: This is where we collaborated and merged all our work together.
- **StreetEasy Dataset - Median Asking Rent**: This dataset lists all the median asking rent prices for different neighborhoods and boroughs each month from January 2010 to October 2021.
- **311 Residential Noise Complaint Dataset from NYC Open Data**: This dataset shows all the noise complaints filed in New York City from 2010 until now, and it includes many descriptors of the complaints, such as incident ZIP code, date created, and complaint type.

B)
Overall, it was a difficult project but a big learning experience due to several reasons.
First, many of our group members dropped out of the class, so we only had two members left to do the work.
Next, the datasets were hard to connect together since 311 dataset was according to ZIP codes and the StreetEasy dataset was categorized by neighborhoods, so we had to go find another source to help us connect them and manually create another document to link them together.
Lastly, the most difficult part was coding Python (thanks professor) during the E**T**L to modify our rent data because it was in a different format than what we needed, but after that step it was a lot easier creating the visualizations of the KPIs. If we were to do the project over again, we would like to examine even more external factors, such as income level for each area.

C)
The overall data did not show any correlation between the number of noise complaints in an area versus the median asking rent price. In most neighborhoods, even as the number of complaints increased, the rent price remained the same; however, there was some data that showed a negative correlation between the two variables. For instance in the Manhattan, Bensonhursts, Corona and Upper East Side scatter plots, as the number of noise complaints increased (after a certain threshold), the median rent prices decreased. In addition, the heat maps showed that the places with the most noise complaints coincide with areas with medium to low median rent and vice versa in some situations.

D)
This was a difficult task, and it felt as if it was a real job. You have employees who leave an organization and the tasks no longer done by those people fall onto you. Tasks that could have been spread out, and the amount of knowledge on the team was lowered. So the remaining team members had to really dig deep and learn to do new things to see this through completion and contacting a more knowledgeable person to assist when we would hit deadends we could not overcome (thanks professor)

Reference List

https://cloud.google.com/bigquery/docs/reference/standard-sql/timestamp_functions
Documentation to adjust the daily 311 data to monthly by formatting the timestamp data type.

https://cloud.google.com/bigquery/docs/reference/standard-sql/uuid_functions
Documentation to create UUID for each dimension. This helped create unique keys.

https://cloud.google.com/bigquery/docs/reference/standard-sql/conversion_functions
Documentation to change data types in Big Query by casting

https://cloud.google.com/bigquery/docs/reference/standard-sql/aggregate_functions
Documentation on how aggregation works in Big Query.

http://holowczak.com/category/datawarehouse/
Various resources, and how dimensional models should be, Star Schema, Grain among others

https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/
Fact table techniques and Dimension Table Techniques and Fundamental Concepts

Group Meeting Log Sheet

Date of Meeting: varied **Time of Meeting**: Last meeting

Group : 3 **Recorder**: none

Attending:

Non-held - done via whatsapp

Absent	**Excused**
N/A	**N/A**

Topics Discussed:

- Dimensional modeling finialization

Tasks Assigned	Team Member	Delivery Date
ETL Visualization	All	12/15/2021

Meeting Ending Time: n/a

(continued below)

Performance Appraisal & Sign-off

Team Member Name(print)	Signature	Weekly Contribution
Polly Chu		50.00%
Kevin Camacho		50.00%