NYC Property Sales

Our Original Dataset

- Comprised of the details of all property sales made in 2016-2017 in the 5 boroughs of NYC.
- Made up of 22 columns and 84,549 rows of property sales.
  - Columns in our data include many attributes such as:
    - Borough
    - Year built
    - Land square feet
- Dataset found on kaggle
  - Original dataset can be found in the nyc.gov website
  - Corresponding metadata in a glossary file

Goals of the Project

- Create visualizations about where sales are occurring the most and which boroughs have the highest revenue using matplotlib.
- Use pandas to find correlation between fields in our dataset and sales.

Data Cleaning

Looking at the Data

- Before making any changes, we looked at data to check for missing values.
  - Instead of NaNs there were strings comprised of ‘ - ‘.
  - There were values with zeros in the sale price column but were not considered NaN.
- Although most of the columns had descriptions, only a few seemed worth focusing on.
  - Some of the data types associated with certain columns were not appropriate.

Dealing with NaN and Zero Values

- The first thing we did was identify the NaN values while importing the file.
  - Then, we filled in the NaNs with a more standard value (-1).
- The zero values in the data set were real sales
  - Transfer of ownership without cash consideration
  - Does not accurately represent house price.
- Decided to filter out both NaN and zero values

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt

        data = pd.read_csv("nyc-rolling-sales.csv", na_values = ' -  ')
```

```
In [2]: data['SALE PRICE'] = data['SALE PRICE'].fillna(-1)
        print(data.shape)

        (84548, 22)
```

```
In [3]: data['BOROUGH'][data['SALE PRICE'] == -1].value_counts()

Out[3]: 4    8295
        1    3867
        5    2399
        Name: BOROUGH, dtype: int64
```

Dropping columns

- Too many blank rows
  - Easement, gross square feet, land square feet, and apartment number
- Redundancy
  - ZIP code, building class at present, building class category, index
- Irrelevant
  - Block, lot, year built and tax class at present
- Vague
  - Residential units, commercial units and total units

Looking at dtypes

```
BOROUGH                          int64
NEIGHBORHOOD                    object
ADDRESS                         object
BUILDING CLASS AT TIME OF SALE  object
SALE PRICE                     float64
SALE DATE                       object
```

Converting columns to appropriate types

- Converted “Sale Price” from float to integer
- Converted “Borough” from int to string. Instead of #1-5, Manhattan, Queens, etc.
- Converted “Sale Date” from object to dateandtime; dropped the time-stamp at the end.

```
In [58]: data.dtypes

Out[58]: BOROUGH                          object
         NEIGHBORHOOD                     object
         ADDRESS                          object
         BUILDING CLASS AT TIME OF SALE   object
         SALE PRICE                        int64
         SALE DATE                datetime64[ns]
         dtype: object
```

Cleaning up data: Results

- Remaining columns
  - Borough
  - Neighborhood
  - Address
  - Building class at time of sale
  - Sale price
  - Sale date
- Remaining rows:
  - 59759

Data Analysis

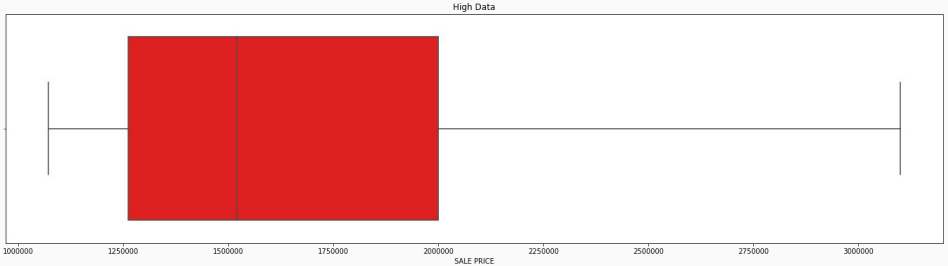
Dividing the dataset into subsets

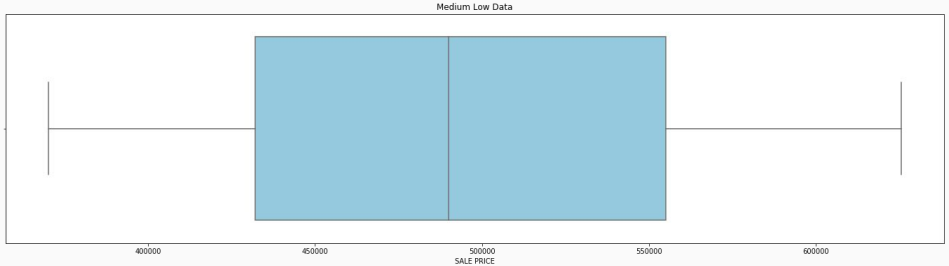
- Dropped features unnecessary to our analysis and narrowed dataset to 6 columns.
- Divided original dataset into 4 subsets by property sales price due to large number of total property sales and to see possible correlations between variables and respective price ranges.
- By specifying lower and upper values of sale price for a subset, we divided the original dataset into low, medium low, medium high, and high.

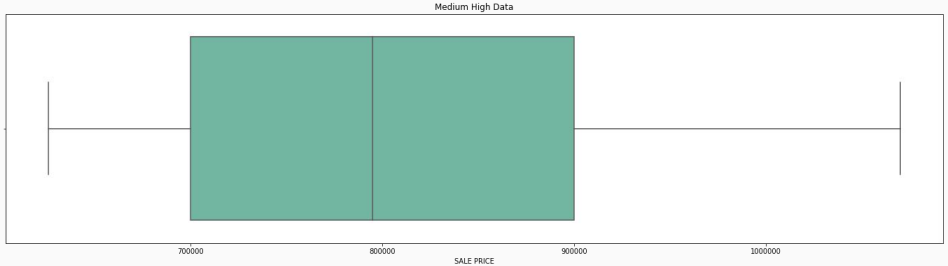
Issues with Data

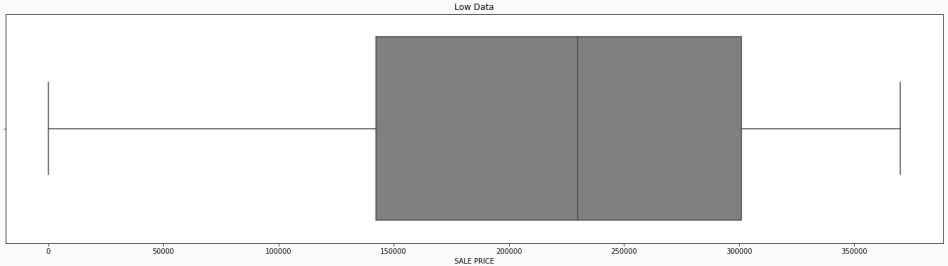
- Low category has deceptive sales prices.
- Trying to describe a subjective market.
- Placed a limit on High category to remove outliers.

Distributions of each set





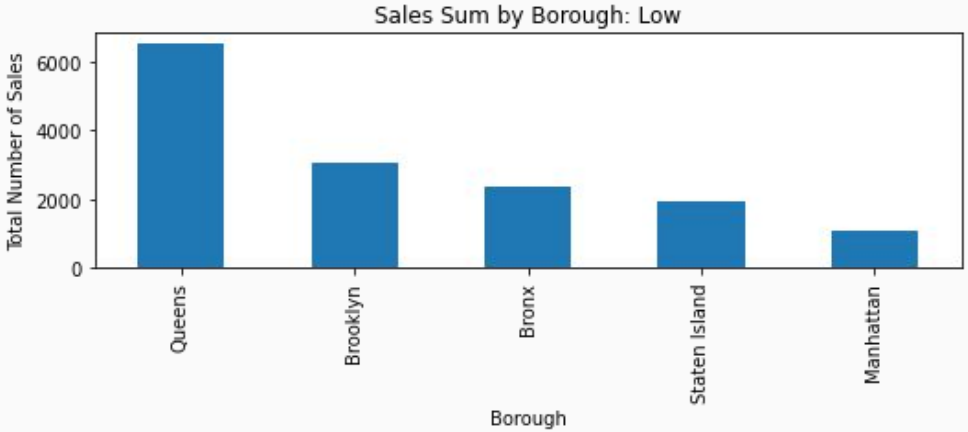


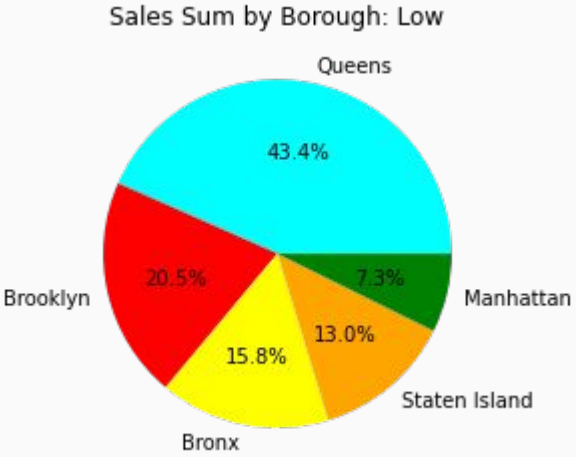


Non-Normal Distributions

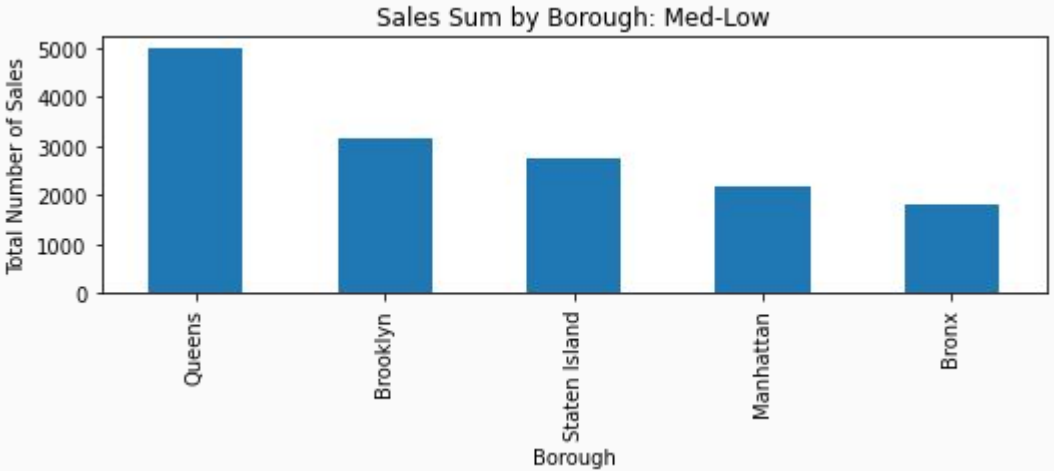
- The resulting distributions indicate heavy skews in the data.
- Mean should not be used for data analysis.

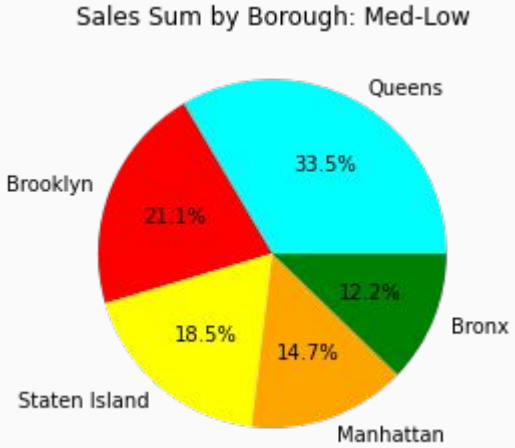
Total sales per borough: Low end



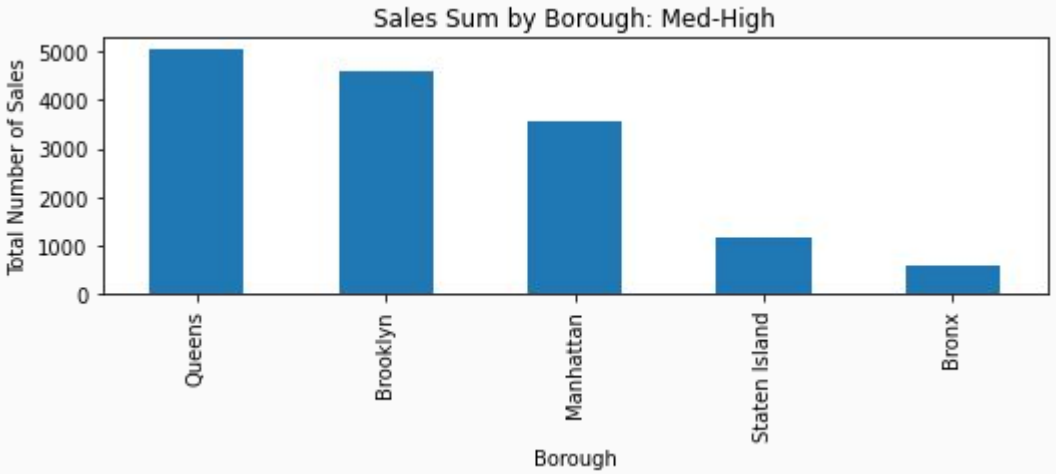


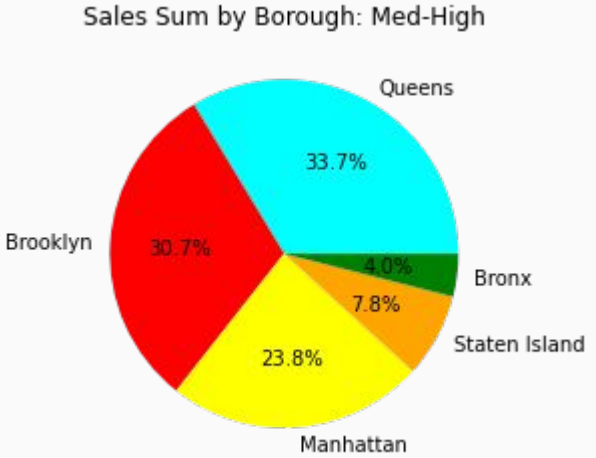
Total sales per borough: Medium-low end



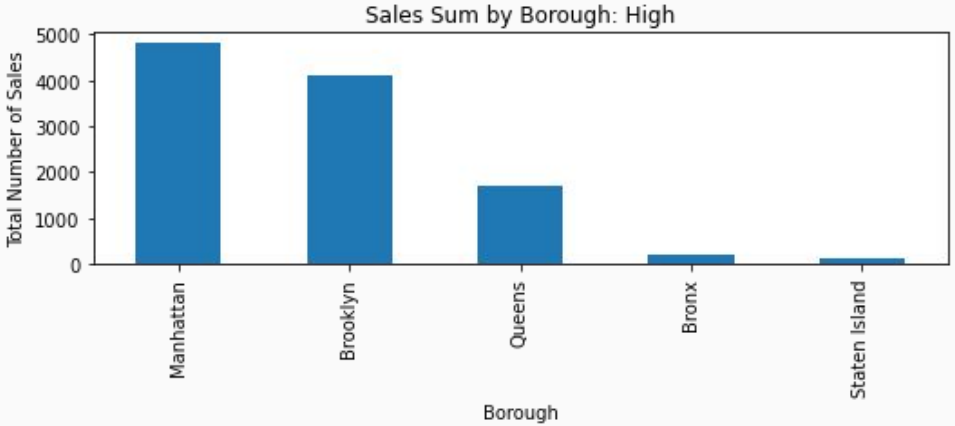


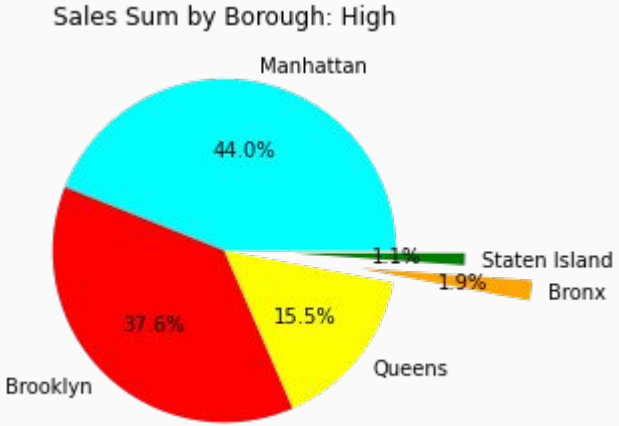
Total sales per borough: Medium-high end



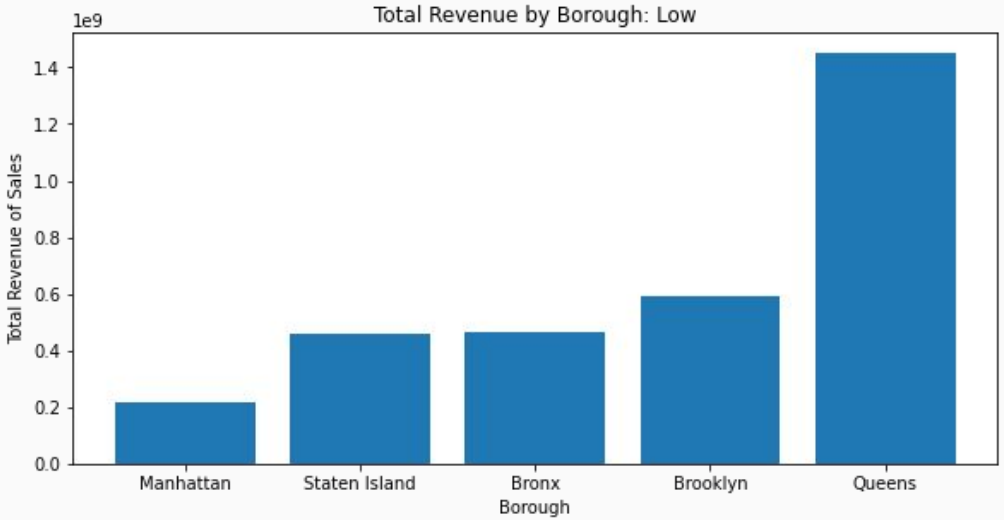


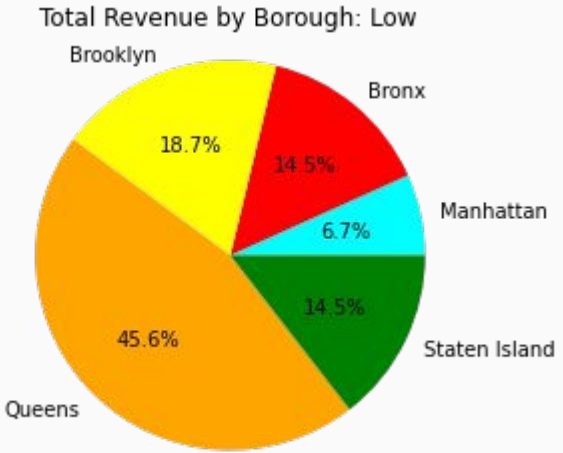
Total sales per borough: High end



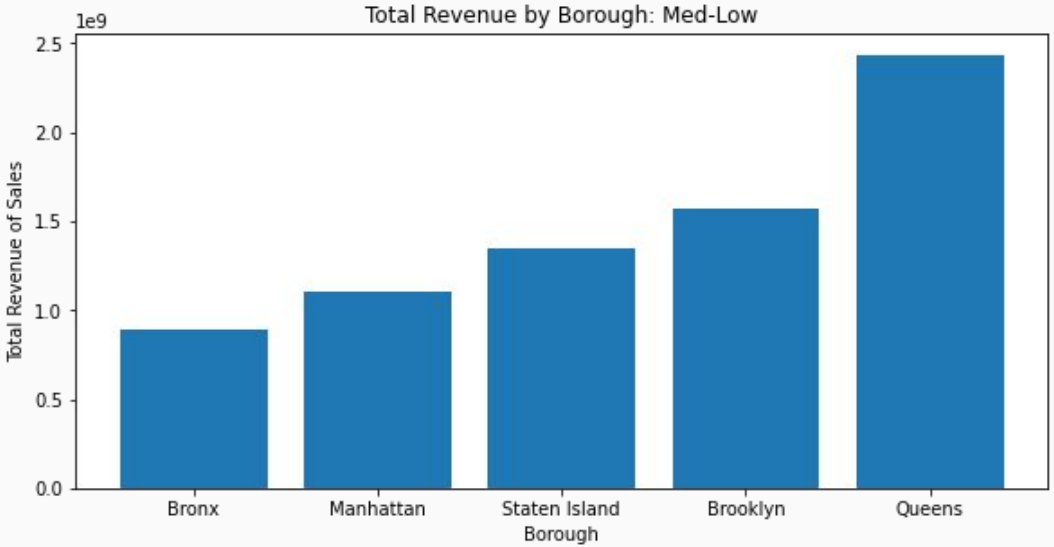


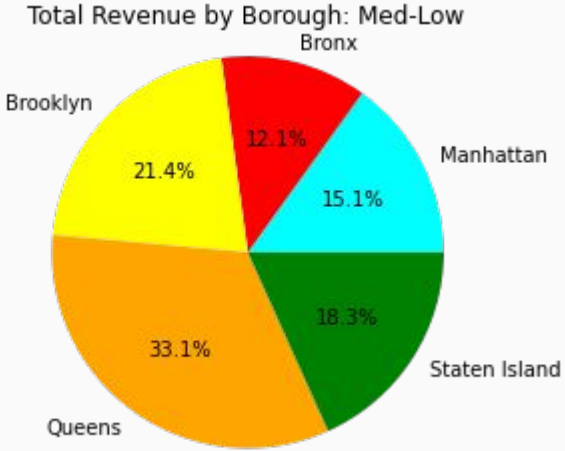
Proportion of Sales in Total Revenue per Borough: Low End



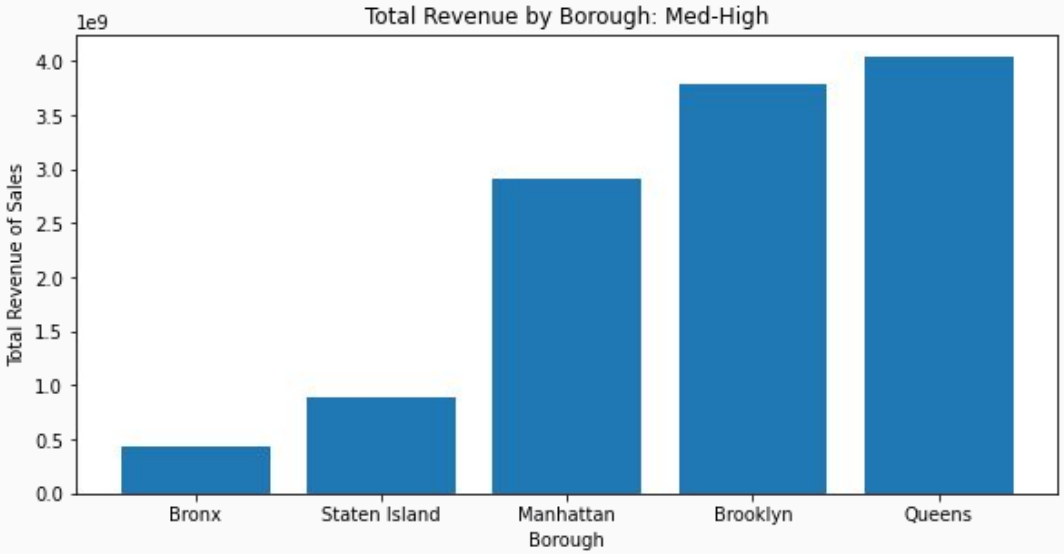


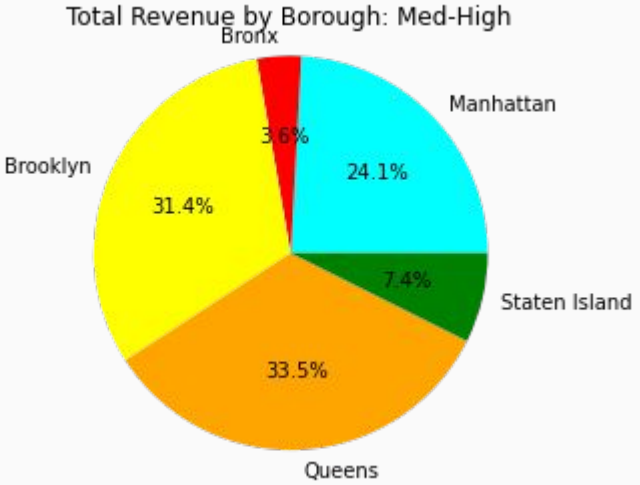
Proportion of Sales in Total Revenue per Borough: Medium-low end



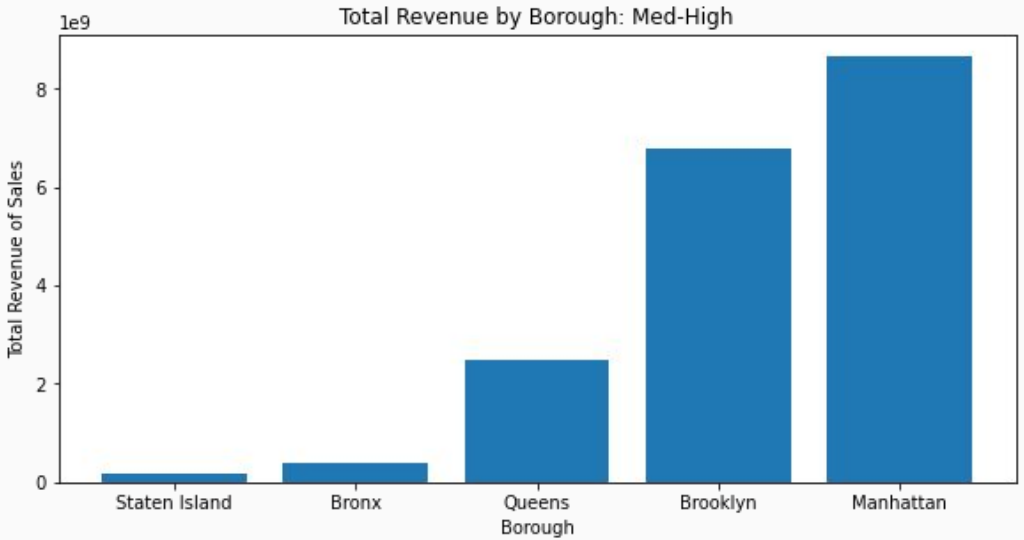


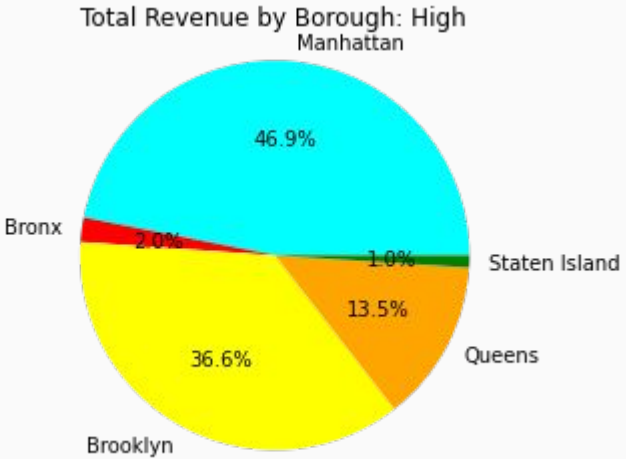
Proportion of Sales in Total Revenue per Borough: Medium-high end



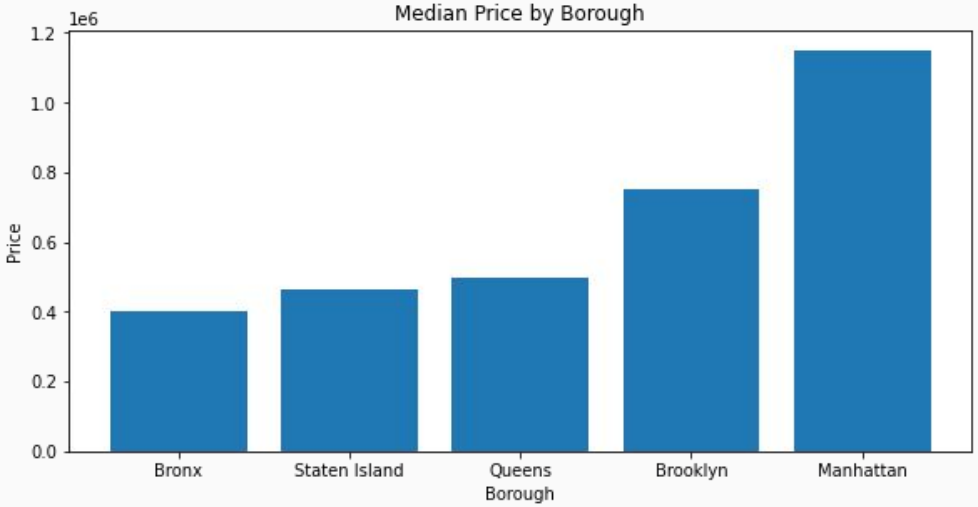


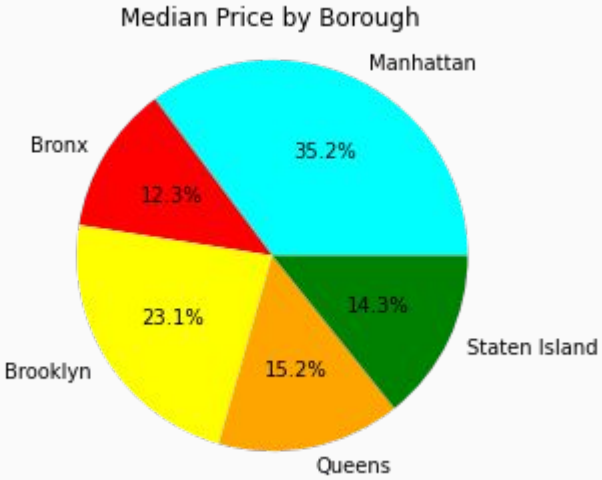
Proportion of Sales in Total Revenue per Borough: High end





Median Sale Price per Borough

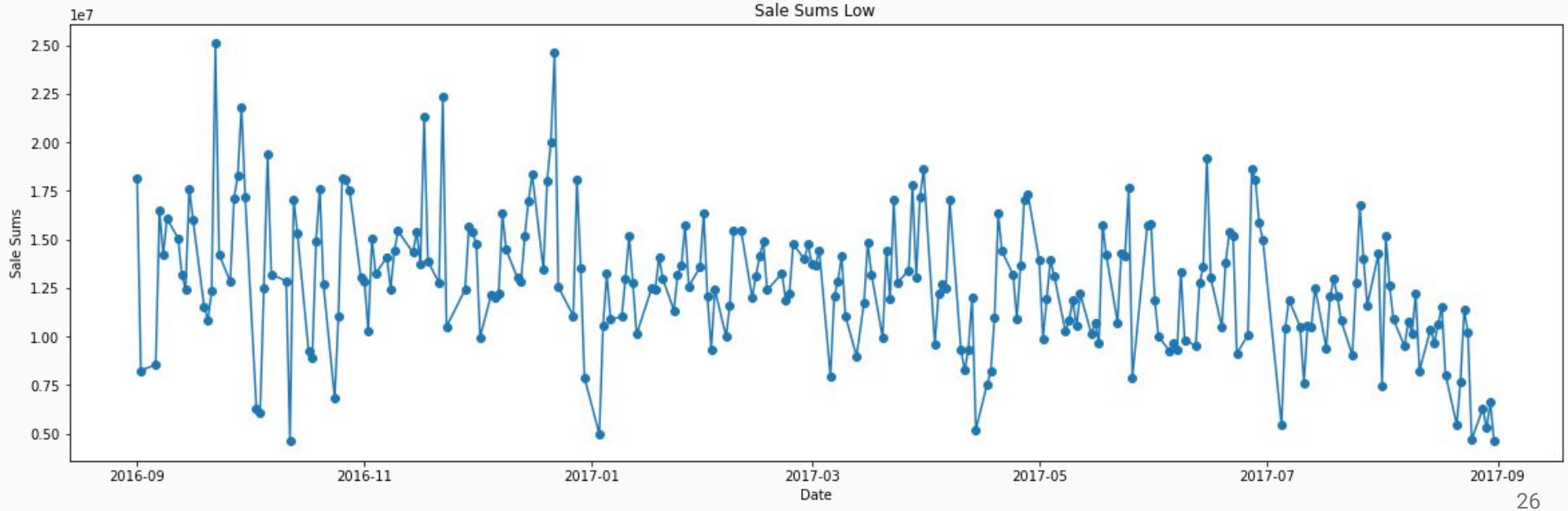




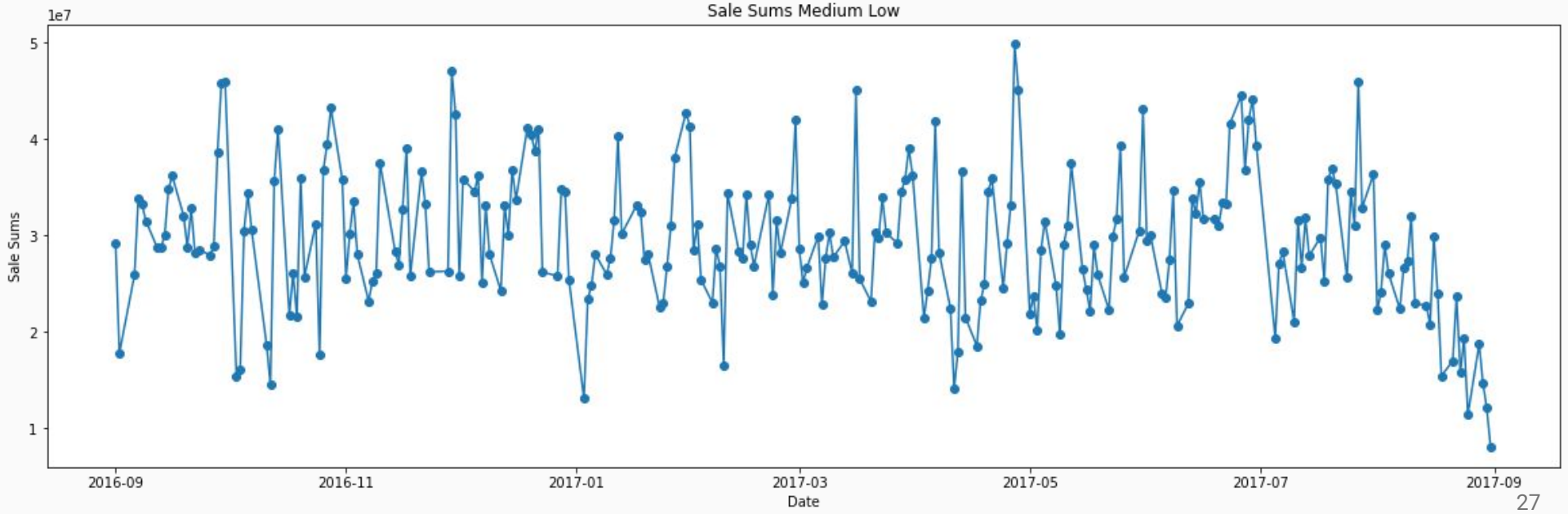
Mapping out sales over time

- One of the challenges we looked at was plotting the sales of buildings by the sale date. Using the four separate price ranges (low, medium low, medium high, and high) four time series were created.
- Note: Days with no sales were removed.

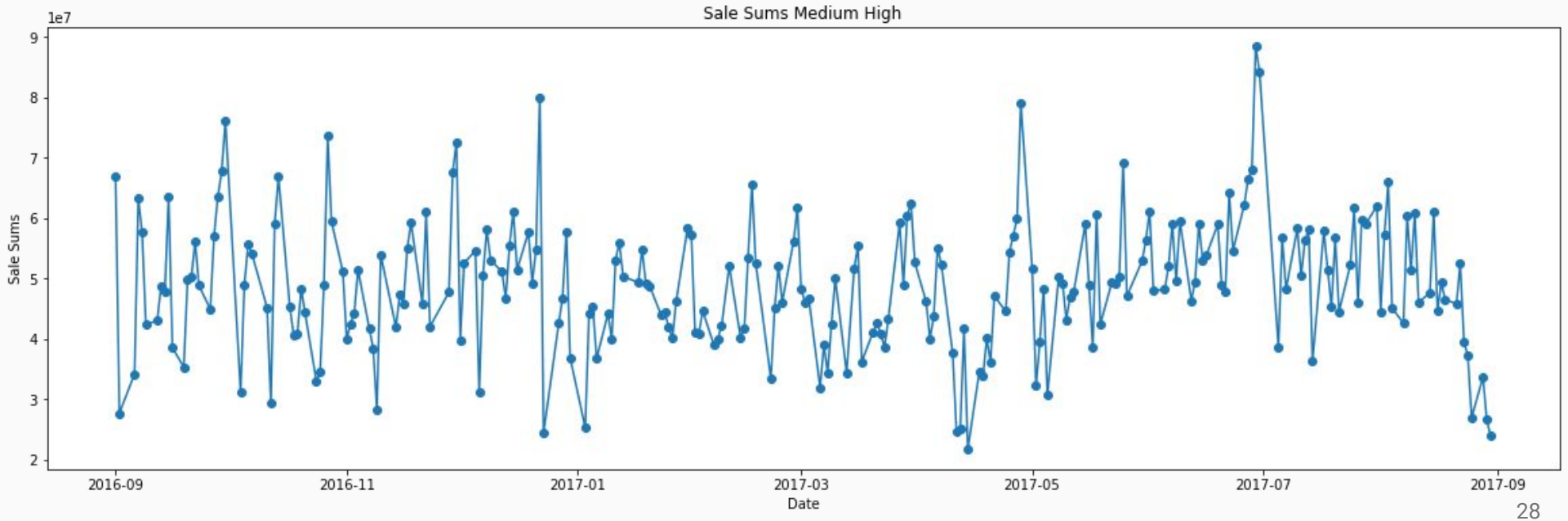
Low


Sale Sums Low
1e7
2.50
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
Sale Sums
2016-09
2016-11
2017-01
2017-03
2017-05
2017-07
2017-09
Date

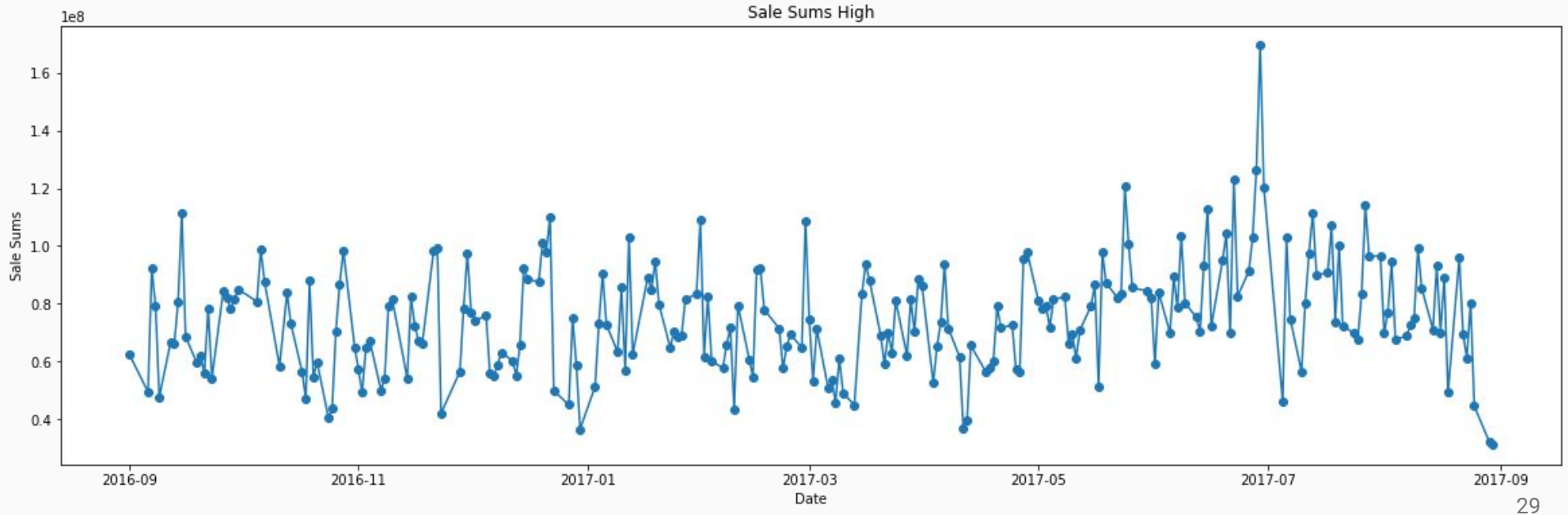
Medium Low


Sale Sums Medium Low
1e7
5
4
3
2
1
Sale Sums
2016-09
2016-11
2017-01
2017-03
2017-05
2017-07
2017-09
Date

Medium High


Sale Sums Medium High
1e7
9
8
7
6
5
4
3
2
Sale Sums
2016-09
2016-11
2017-01
2017-03
2017-05
2017-07
2017-09
Date

High


Sale Sums High
1e8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
Sale Sums
2016-09
2016-11
2017-01
2017-03
2017-05
2017-07
2017-09
Date

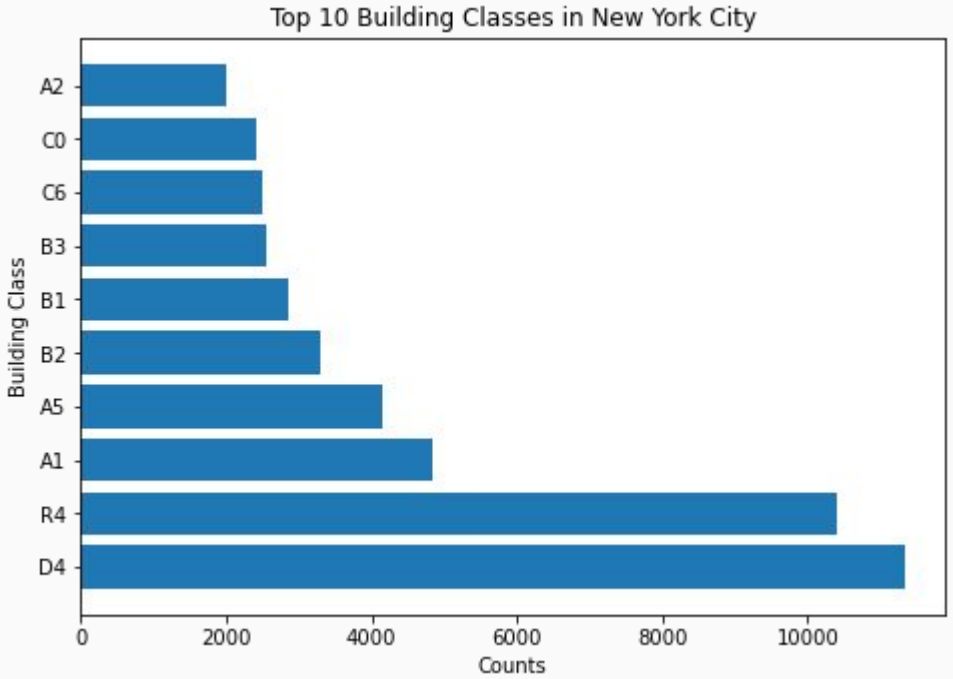
Explanation for the dip?

- November elections for Mayor.
- Not properly understanding data.

Building Classes

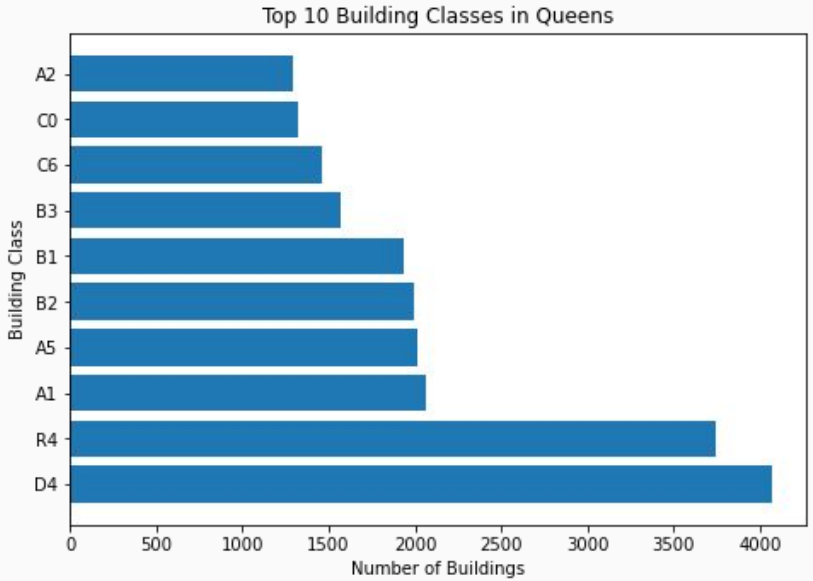
- To separate properties based on their qualities, the data included a building classification column
- They are labeled as 2 character ID’s
- The building class ID’s are defined in the nyc.gov website

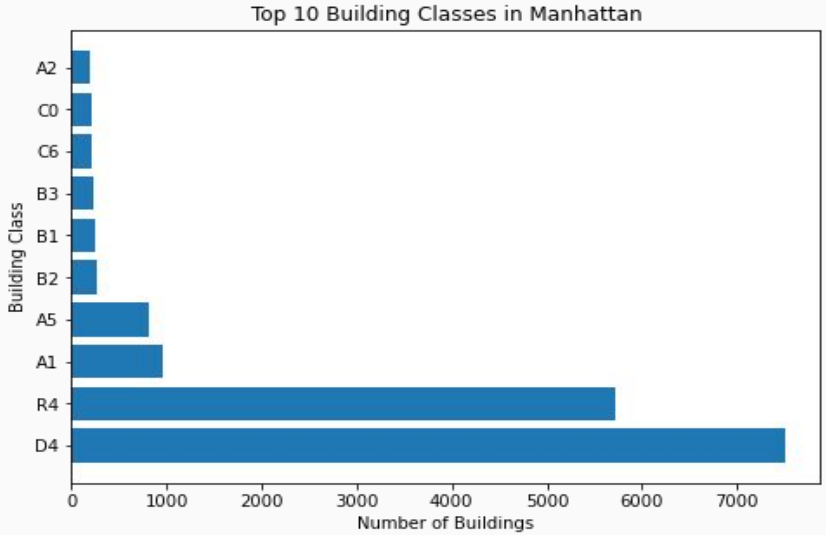
Top 10 Building Classes



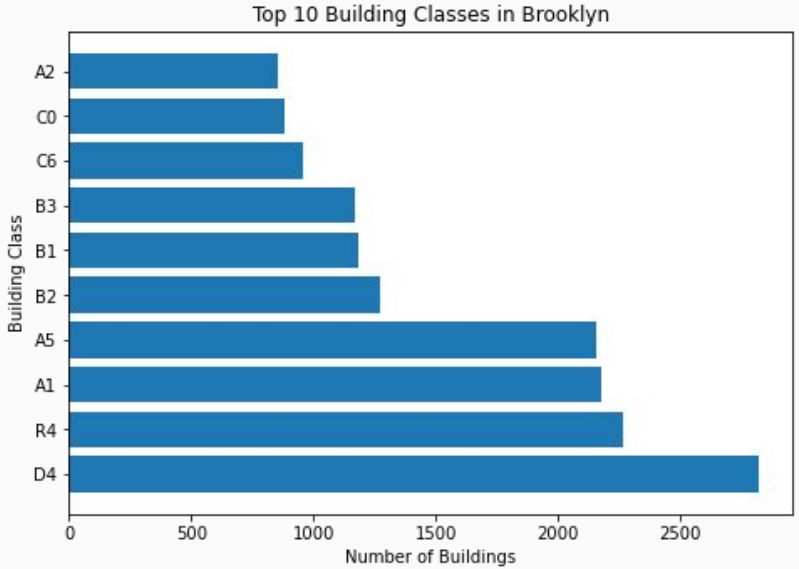
- A2: One story permanent living quarter
- C0: Three families
- C6: Walk-up cooperative
- B3: Two family converted from one family
- B1: Two family brick
- B2: Two family frame
- A5: One family attached or semi-attached
- A1: Two stories: detached sm or mid
- R4: Condo; Residential unit in elevator bldg
- D4: Elevator Cooperative

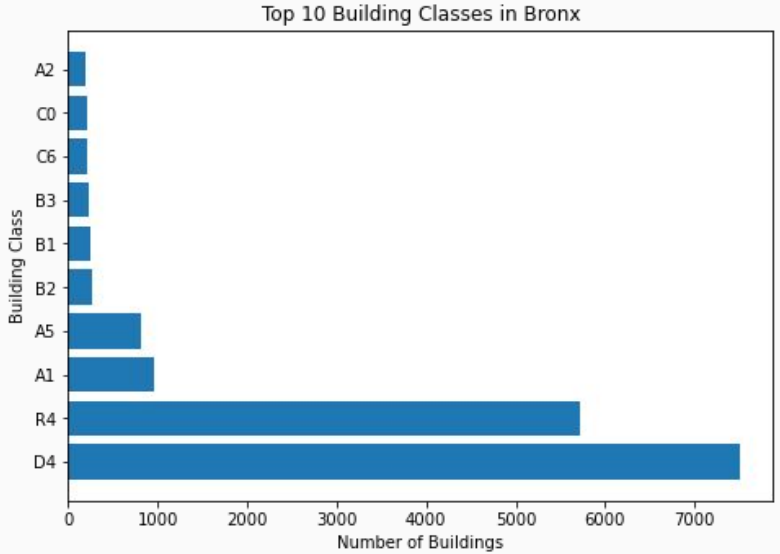
Queens vs Manhattan



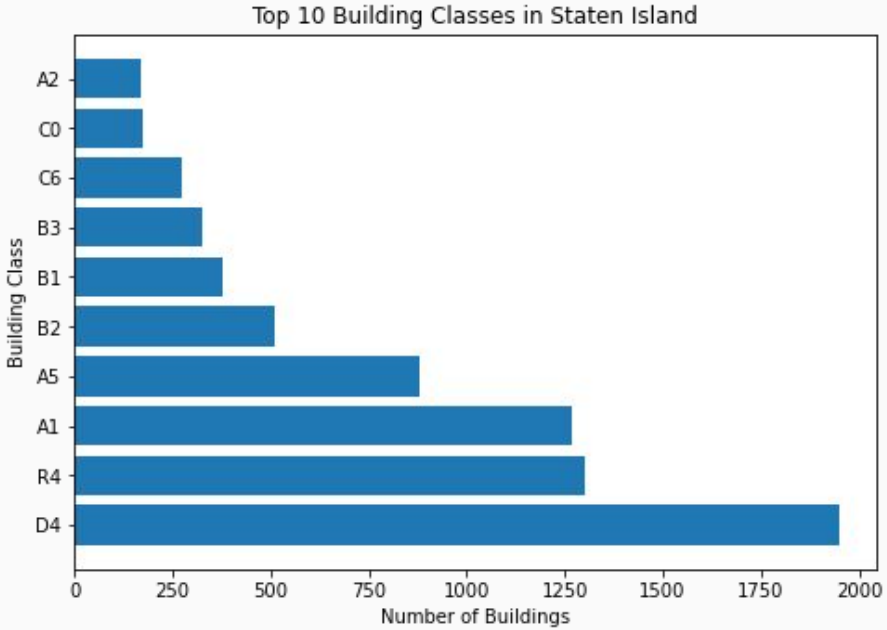


Brooklyn vs Bronx





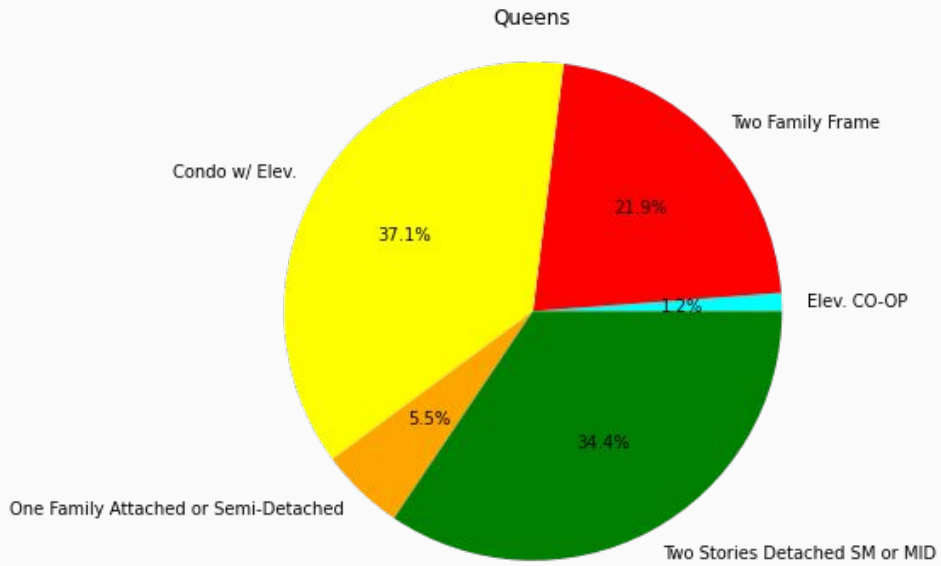
Staten Island

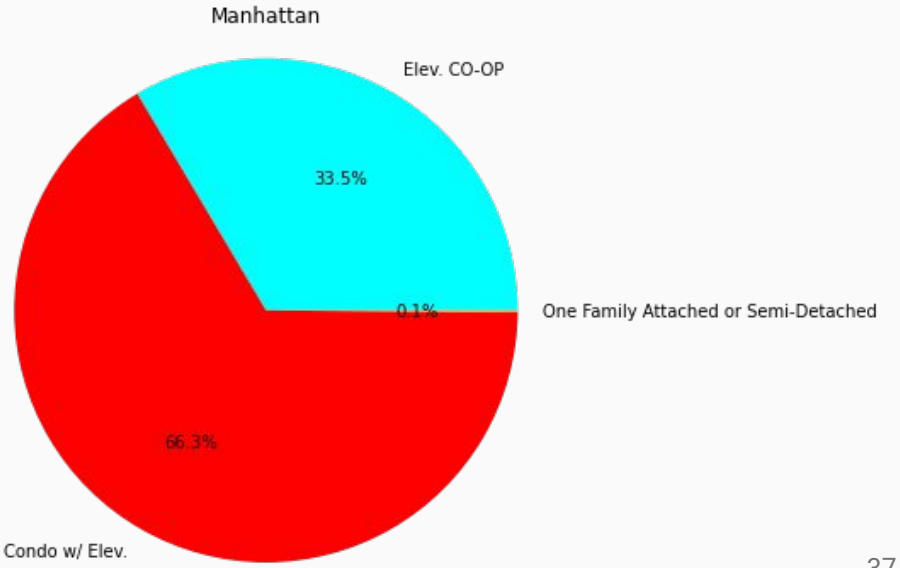


Combining Building and Sales Range

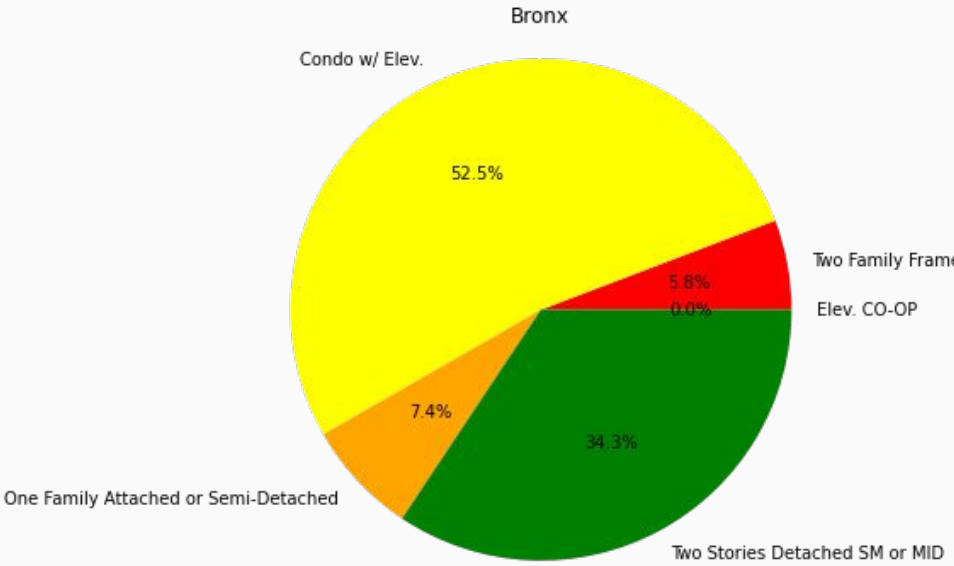
- Using building class? Which ones?
- How much analysis?

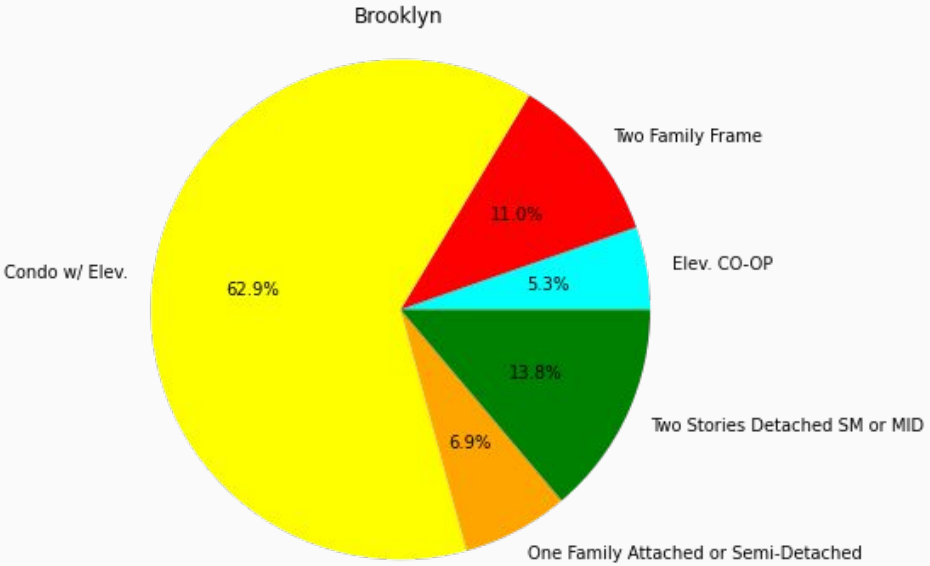
Where are the high end sales coming from?



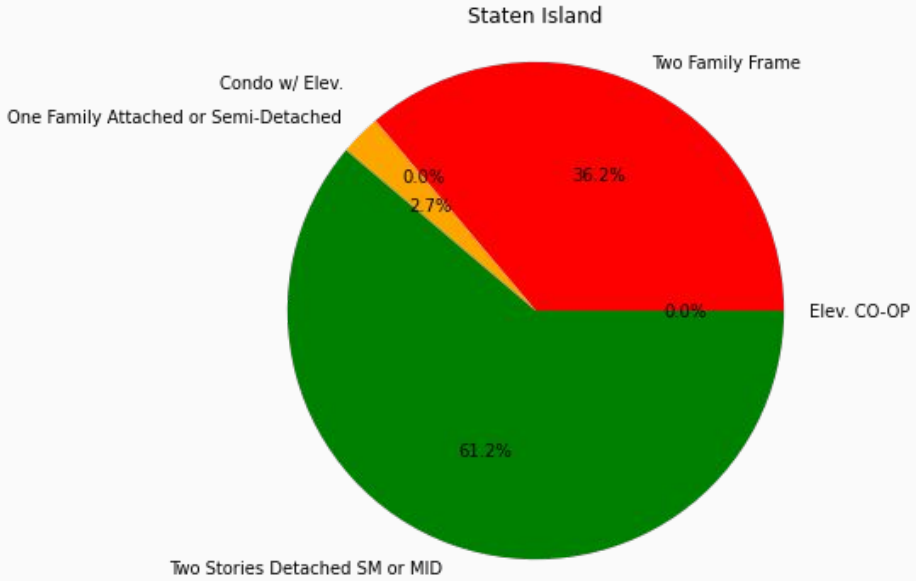


High-end sales from the Bronx and Brooklyn

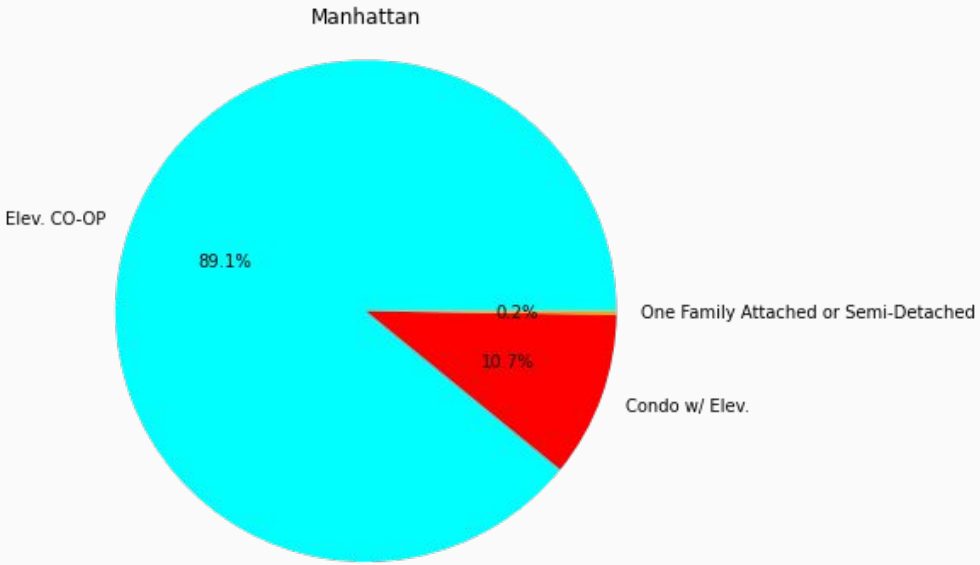


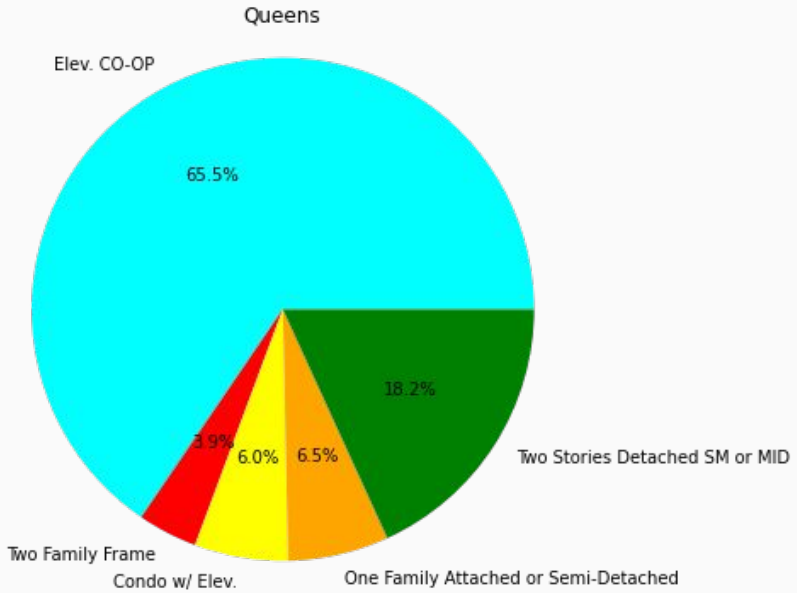


High end Sales from Staten Island

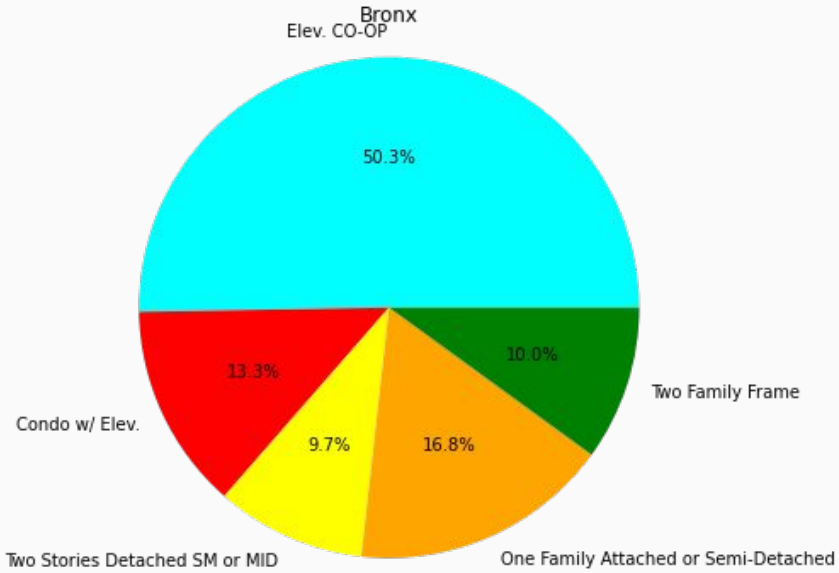


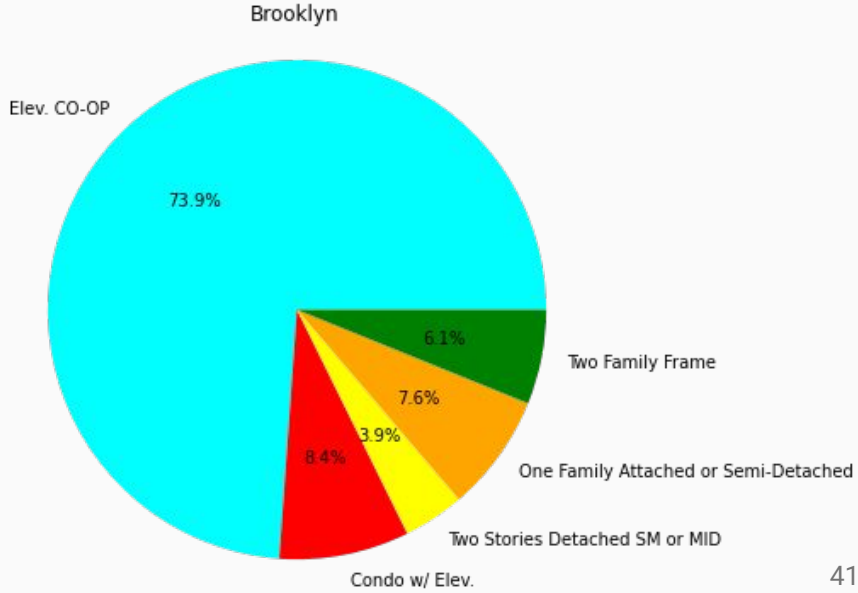
Where are the low end sales coming from?



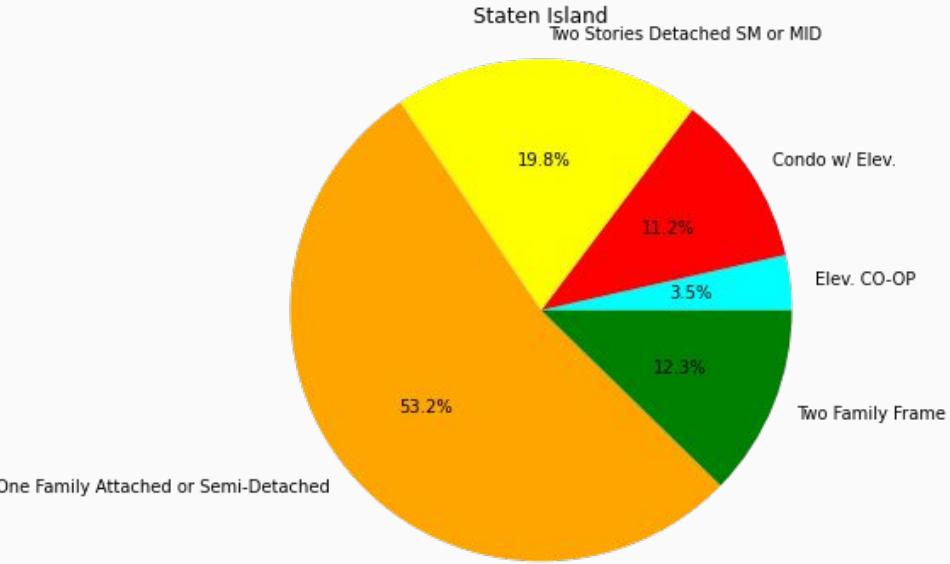


Low end Sales from the Bronx and Brooklyn



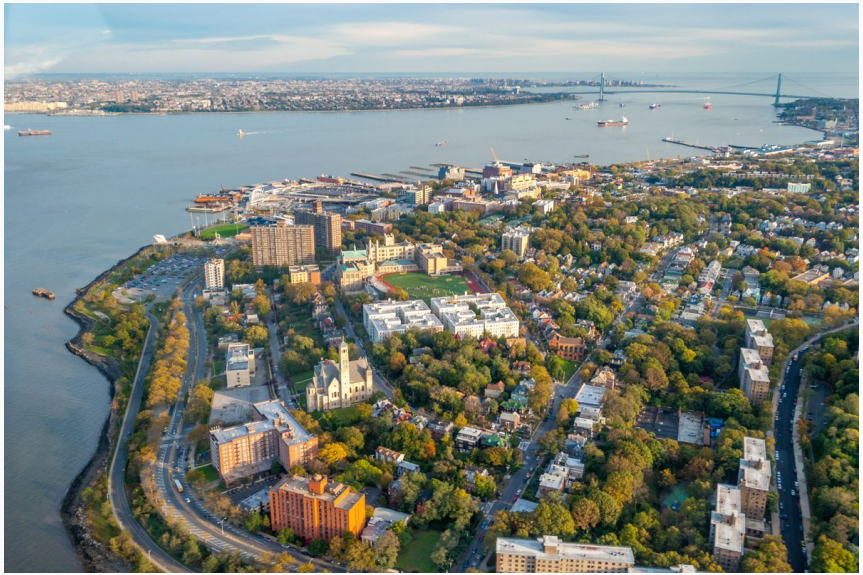


Low end sales from Staten Island



Aerial View of Staten Island and Queens

Staten Island



Queens

Aerial View of Brooklyn and the Bronx

Brooklyn

Bronx

Aerial View of Manhattan

Manhattan

Findings from our Analysis

- Queens had the highest number of property sales except where the price of the property was higher than than $1,000,000. Manhattan had the highest sales in this category.
- Brooklyn had the 2nd highest number of property sales throughout the price ranges.
- Bronx and Staten Island had the lowest number of property sales.
- Queens also had the highest amount of revenue from property sales for the subset with the low price range.
- Manhattan had the highest amount of revenue from property sales for the subset with the highest price range.

What could be analyzed further?

- Find trends in transfer sales.
- Spatial data for heat map.
- Reusing the original data set, remove all missing values and see if remaining data can be studied.
- Median prices of building classes within the different boroughs (Creating sets with better distributions).
- Use neighborhood column and compare prices of certain neighborhoods within a borough
- Using Address column