NYC Real Estate Analysis During COVID

Presented by DataTreasure Consultants

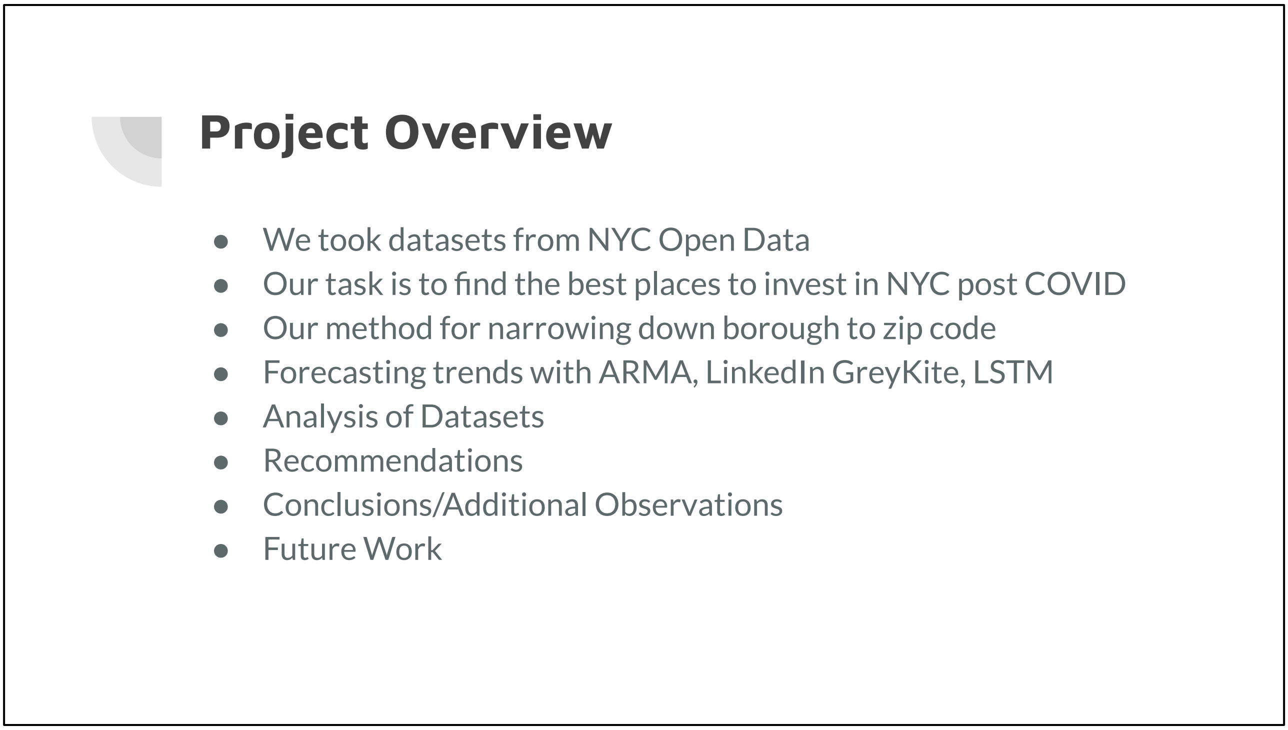
Source: http://www.destination360.com/north-america/us/new-york/images/s/free-things-to-do-in-new-york-city.jpg

Agenda

- Project Overview
- Datasets Utilized
- Methodology
- Time Series Models and Analysis
- How We Chose the Zip Codes
- Recommendations
- Future Work
- Questions

Agenda

- Project Overview
- Datasets Utilized
- Methodology
- Time Series Models and Analysis
- How We Chose the Zip Codes
- Recommendations
- Future Work
- Questions

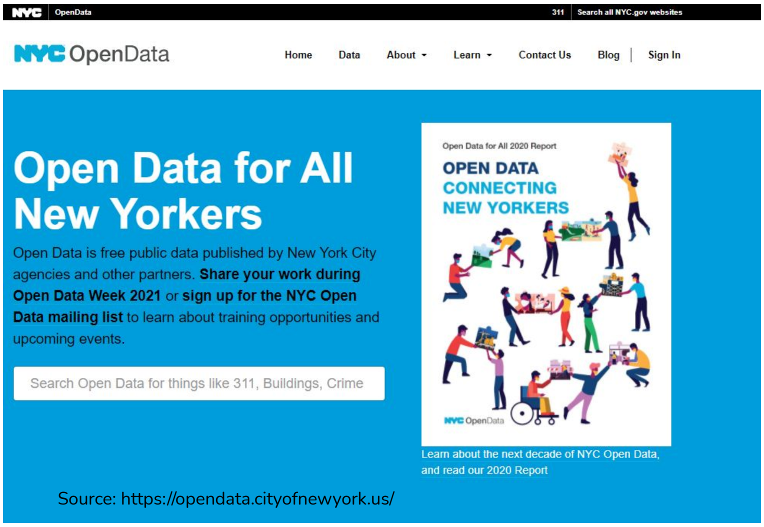


Project Overview

- We took datasets from NYC Open Data
- Our task is to find the best places to invest in NYC post COVID
- Our method for narrowing down borough to zip code
- Forecasting trends with ARMA, LinkedIn GreyKite, LSTM
- Analysis of Datasets
- Recommendations
- Conclusions/Additional Observations
- Future Work

- We took datasets from NYC Open Data
    - Who knew they tracked everything!
- Our task is to find the best places to invest in NYC post COVID
    - Initially I wanted to do pre-COVID as well but that was shot down due to not being able to do anything with the 2GB dataset
- Our method for narrowing down borough to zip code
- Forecasting trends with ARMA, LinkedIn GreyKite, LSTM
- Analysis of Datasets
- Recommendations
- Conclusions/Additional Observations
- Future Work

The Data We Used


NYC OpenData
311
Search all NYC.gov websites
NYC OpenData
Home
Data
About
Learn
Contact Us
Blog
Sign In
Search Open Data for things like 311, Buildings, Crime

Open Data for All New Yorkers

Open Data is free public data published by New York City agencies and other partners. **Share your work during Open Data Week 2021** or **sign up for the NYC Open Data mailing list** to learn about training opportunities and upcoming events.

Open Data for All 2020 Report

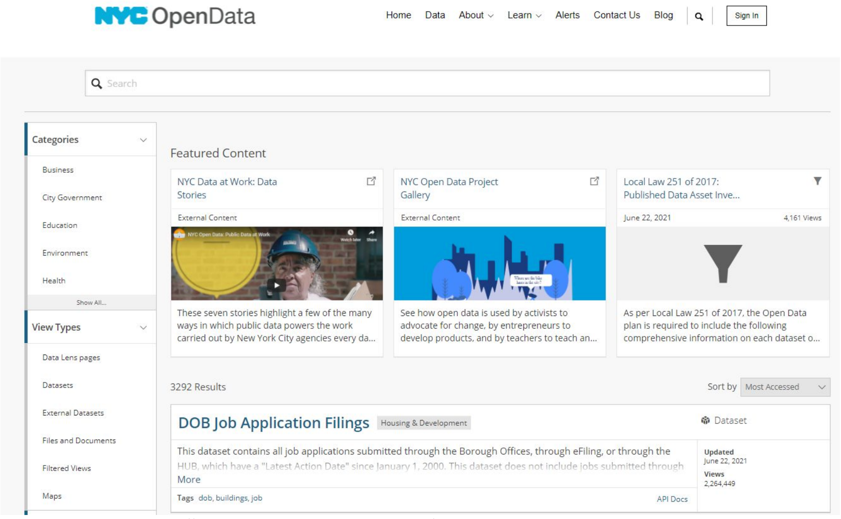
OPEN DATA CONNECTING NEW YORKERS

Learn about the next decade of NYC Open Data, and read our 2020 Report

Source: https://opendata.cityofnewyork.us/

We got our data from NYC Open data. This is their website.

The Data We Used



Featured Content

NYC Data at Work: Data Stories
External Content
These seven stories highlight a few of the many ways in which public data powers the work carried out by New York City agencies every da...

NYC Open Data Project Gallery
External Content
See how open data is used by activists to advocate for change, by entrepreneurs to develop products, and by teachers to teach an...

Local Law 251 of 2017: Published Data Asset Inve...
June 22, 2021 4,161 Views
As per Local Law 251 of 2017, the Open Data plan is required to include the following comprehensive information on each dataset o...

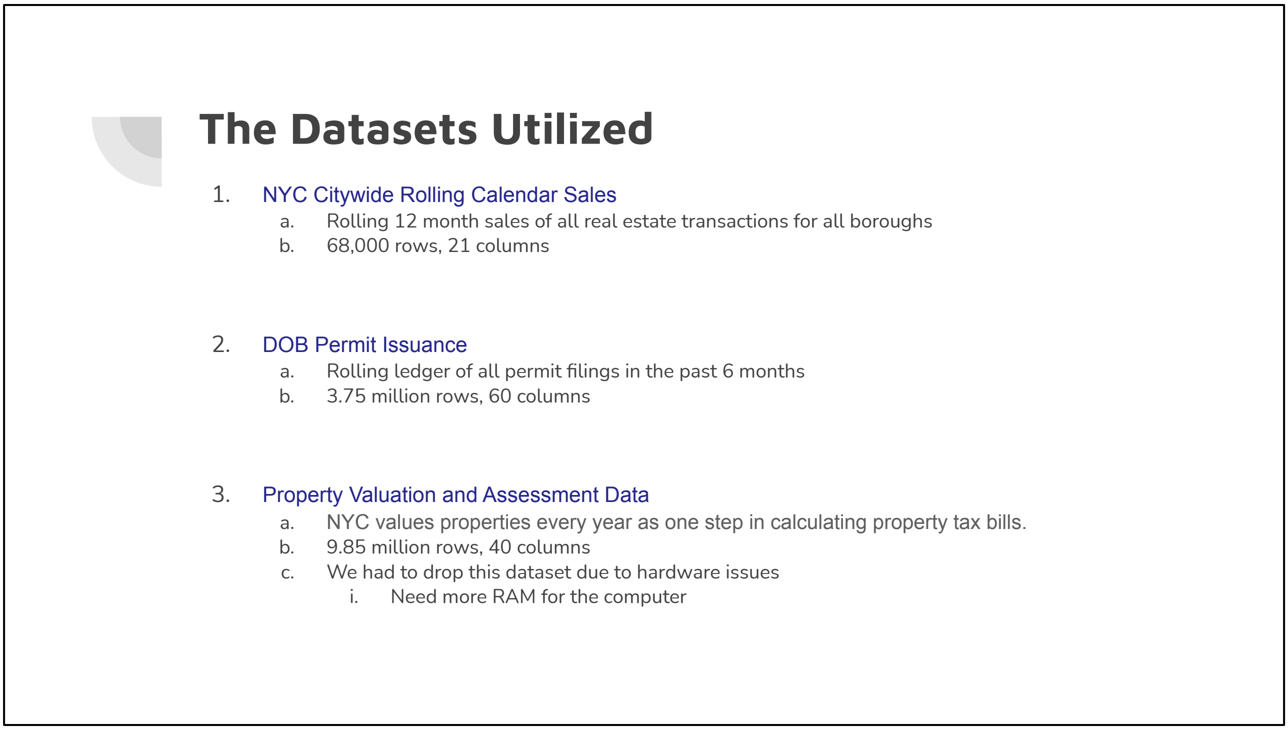
3292 Results
Sort by Most Accessed

DOB Job Application Filings Housing & Development
Dataset
This dataset contains all job applications submitted through the Borough Offices, through eFiling, or through the HUB, which have a "Latest Action Date" since January 1, 2000. This dataset does not include jobs submitted through
More
Tags dob, buildings, job
API Docs
Updated June 22, 2021
Views 2,264,449

Source: https://opendata.cityofnewyork.us/

We got our data from NYC Open data. This is their website.

Lo and behold, they had tracked many different categories. This is a massive undertaking on their part and very useful for us to data mine.



The Datasets Utilized

1. NYC Citywide Rolling Calendar Sales
   a. Rolling 12 month sales of all real estate transactions for all boroughs
   b. 68,000 rows, 21 columns

2. DOB Permit Issuance
   a. Rolling ledger of all permit filings in the past 6 months
   b. 3.75 million rows, 60 columns

3. Property Valuation and Assessment Data
   a. NYC values properties every year as one step in calculating property tax bills.
   b. 9.85 million rows, 40 columns
   c. We had to drop this dataset due to hardware issues
      i. Need more RAM for the computer

So, from this website, we obtained the following datasets:

1. NYC Citywide Rolling Calendar Sales
   a. Rolling 12 month sales of all real estate transactions for all boroughs
   b. 68,000 rows, 21 columns

We used this one as the main dataset for analysis. During the time of the project, they updated and merged all datasets into one. BEfore it was 5 separate datasets with less rows but same columns per borough. I used the new dataset for comparison of the model to new fresh data

2. DOB Permit Issuance
   a. Rolling ledger of all permit filings in the past 6 months
   b. 3.75 million rows, 60 columns

We used this for the finding of the top zip codes after we checked to see how the boroughs were doing from the rolling sales data. There were some interesting findings with the names of the top people who were submitting permit requests. I believe we found some useful information here on which zip codes to focus on .

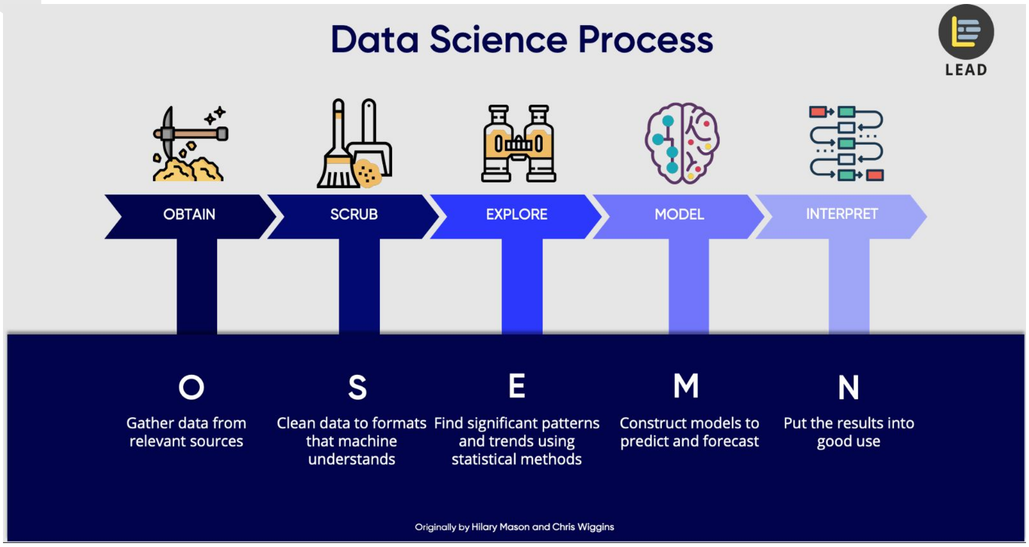
Property Valuation and Assessment Data

c. NYC values properties every year as one step in calculating property tax bills.
d. 9.85 million rows, 40 columns

This dataset was too big. We had to drop it during the project because our software would keep crashing trying to use it due to memory problems

Methodolgy

Source : https://towardsdatascience.com/tagged/osemn



Data Science Process

O Gather data from relevant sources

S Clean data to formats that machine understands

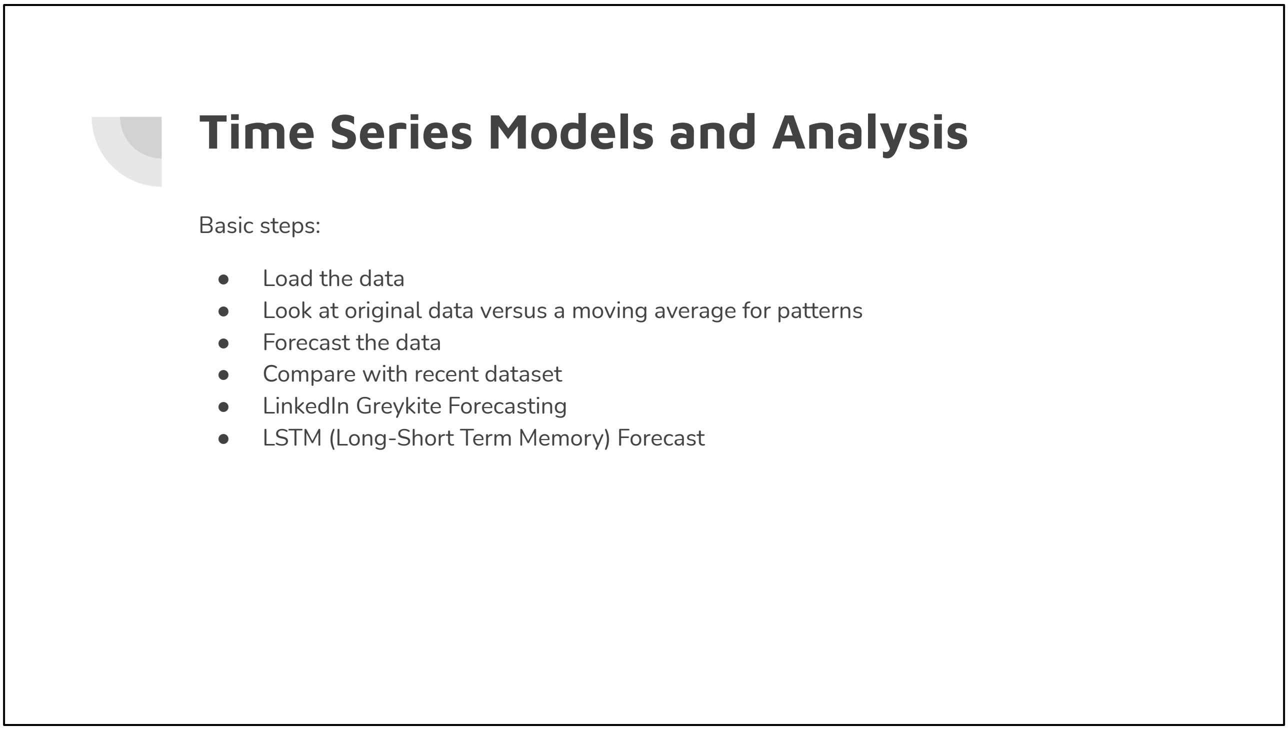
E Find significant patterns and trends using statistical methods

M Construct models to predict and forecast

N Put the results into good use

Originally by Hilary Mason and Chris Wiggins

We used the OSEMN methodology!



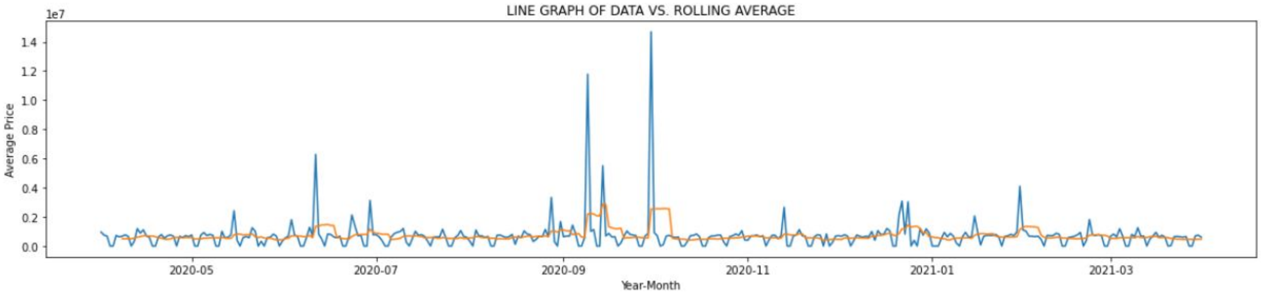
Time Series Models and Analysis

Basic steps:

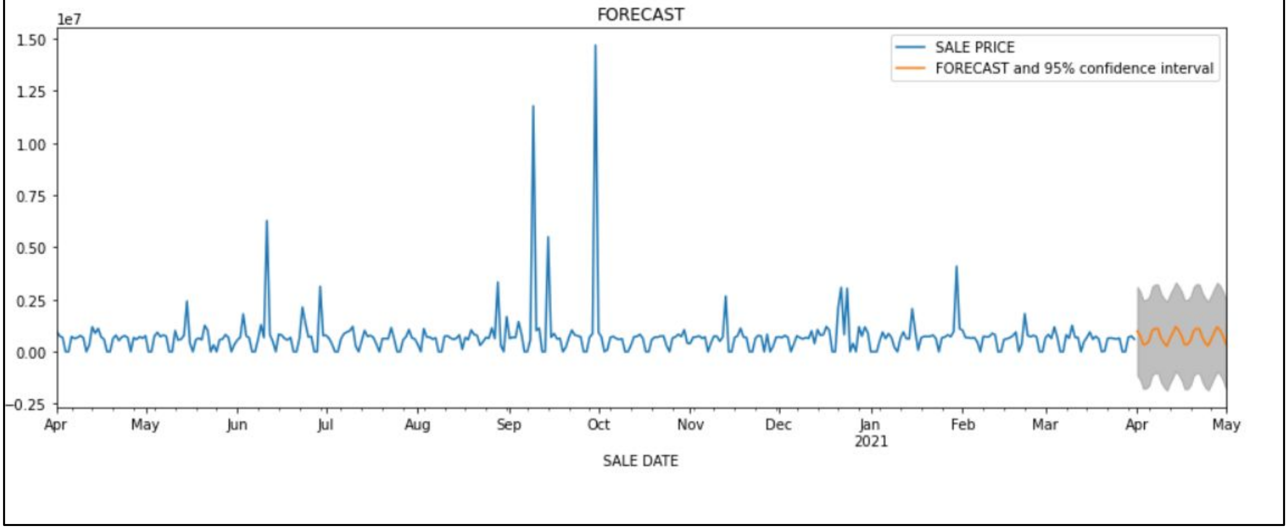
- Load the data
- Look at original data versus a moving average for patterns
- Forecast the data
- Compare with recent dataset
- LinkedIn Greykite Forecasting
- LSTM (Long-Short Term Memory) Forecast

Bear with me, these will be alot of slides. Here is a general outline of what is to come:

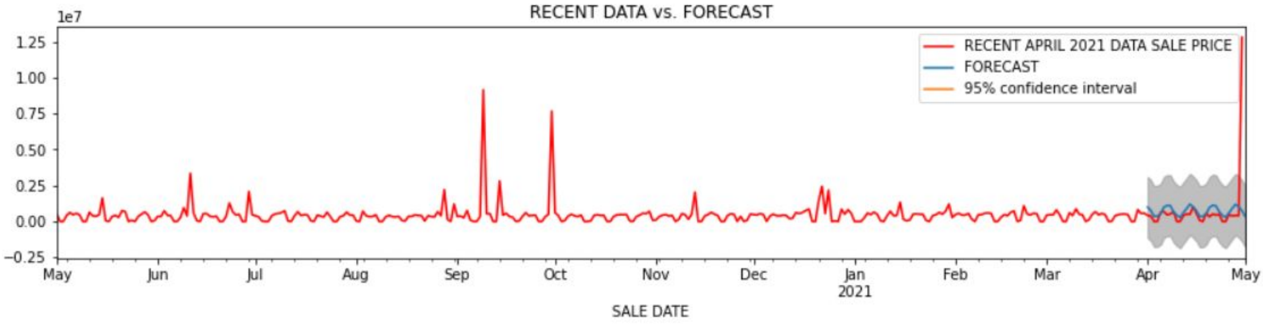
QUEENS ARMA


LINE GRAPH OF DATA VS. ROLLING AVERAGE
1e7
Average Price
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
Year-Month

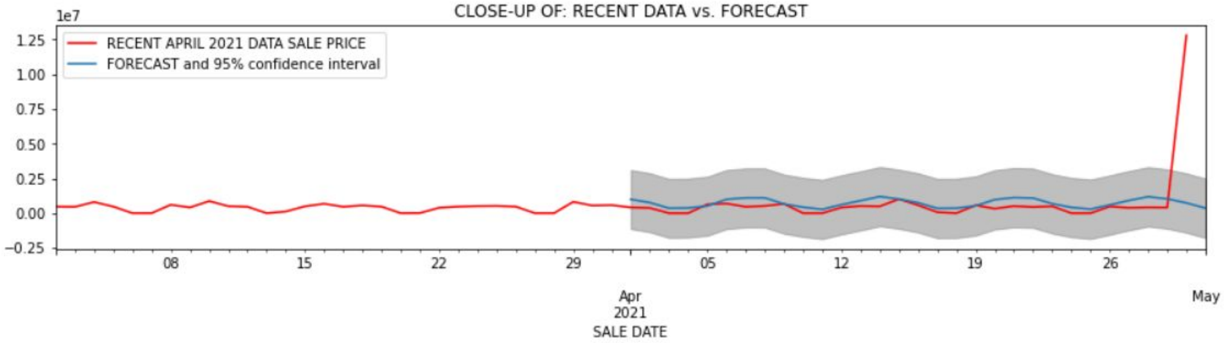
QUEENS ARMA


FORECAST
1e7
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
SALE PRICE
FORECAST and 95% confidence interval
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
2021
Feb
Mar
Apr
May
SALE DATE

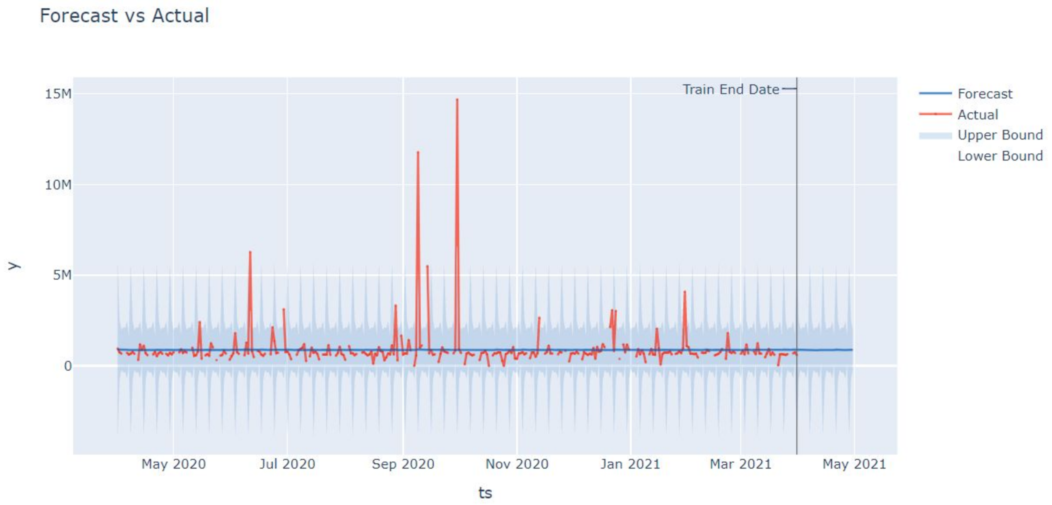
QUEENS ARMA


RECENT DATA vs. FORECAST
1e7
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
RECENT APRIL 2021 DATA SALE PRICE
FORECAST
95% confidence interval
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
2021
Feb
Mar
Apr
May
SALE DATE

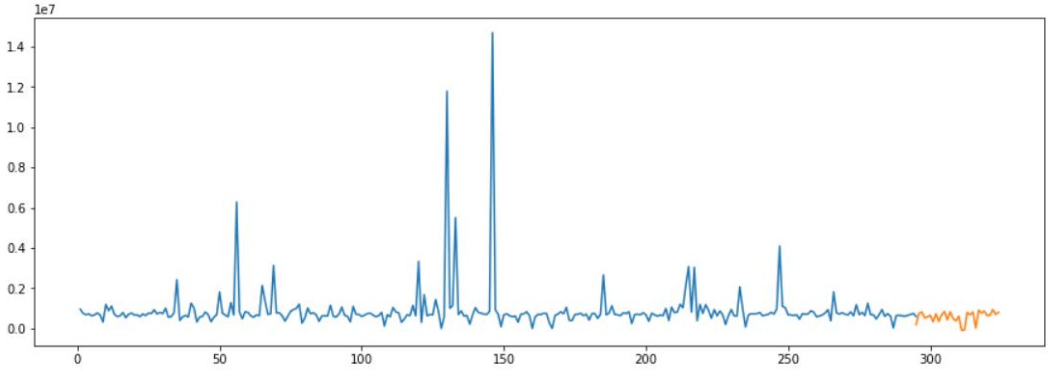
QUEENS ARMA


CLOSE-UP OF: RECENT DATA vs. FORECAST
1e7
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
RECENT APRIL 2021 DATA SALE PRICE
FORECAST and 95% confidence interval
08
15
22
29
05
12
19
26
Apr
2021
May
SALE DATE

QUEENS - LinkedIn GreyKite Forecast



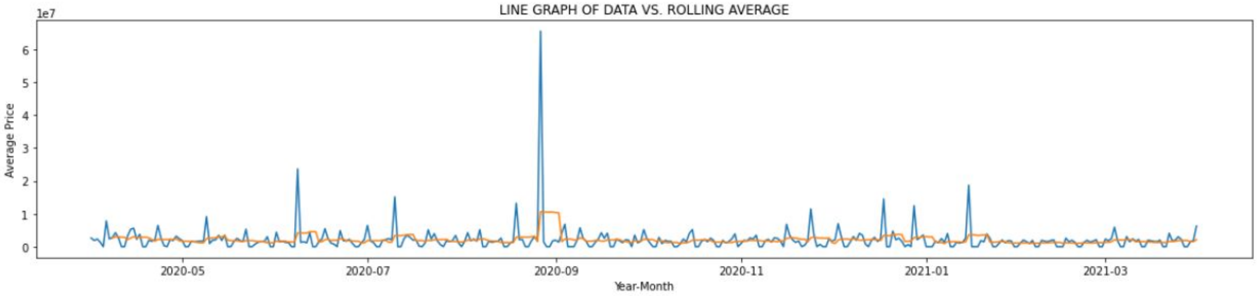
QUEENS - LSTM Model Forecast



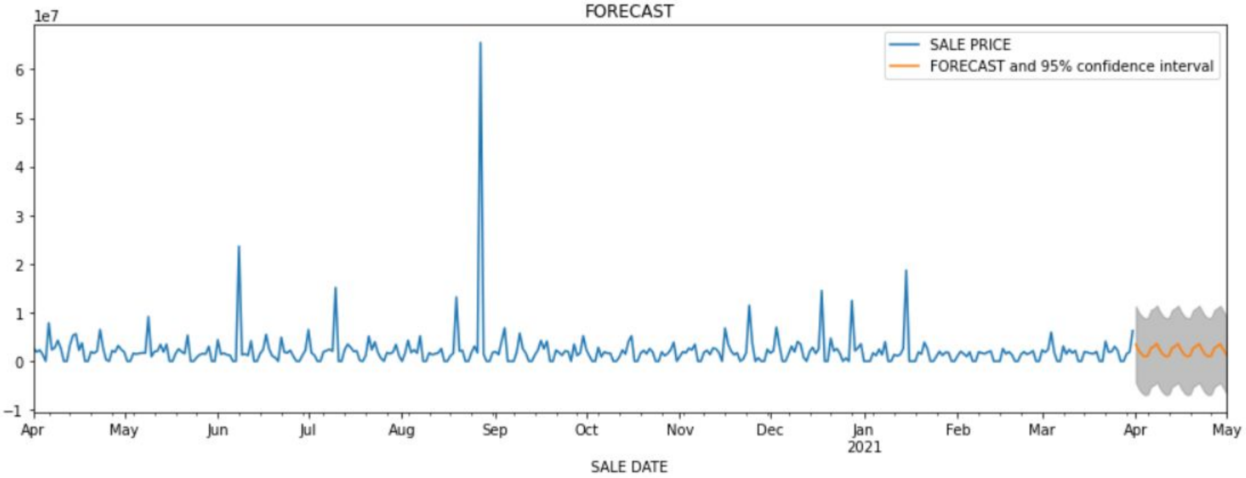
Observation

Queens prices per model show stable with some dips

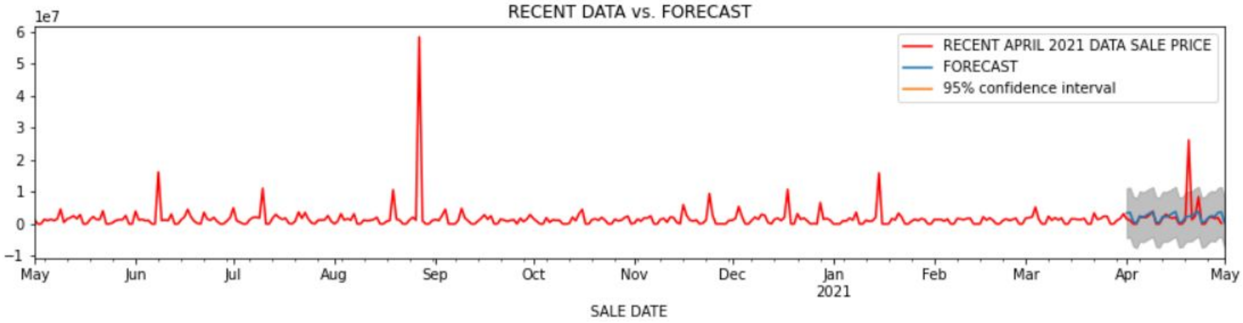
MANHATTAN ARMA



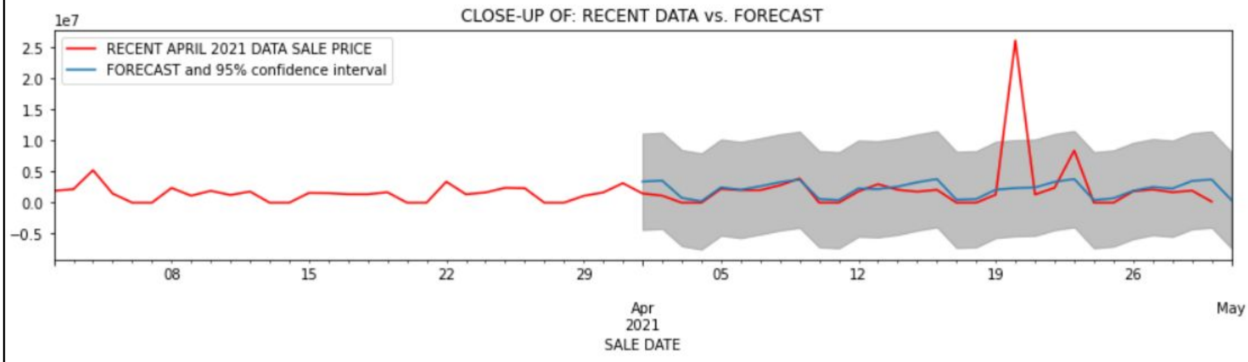
MANHATTAN ARMA



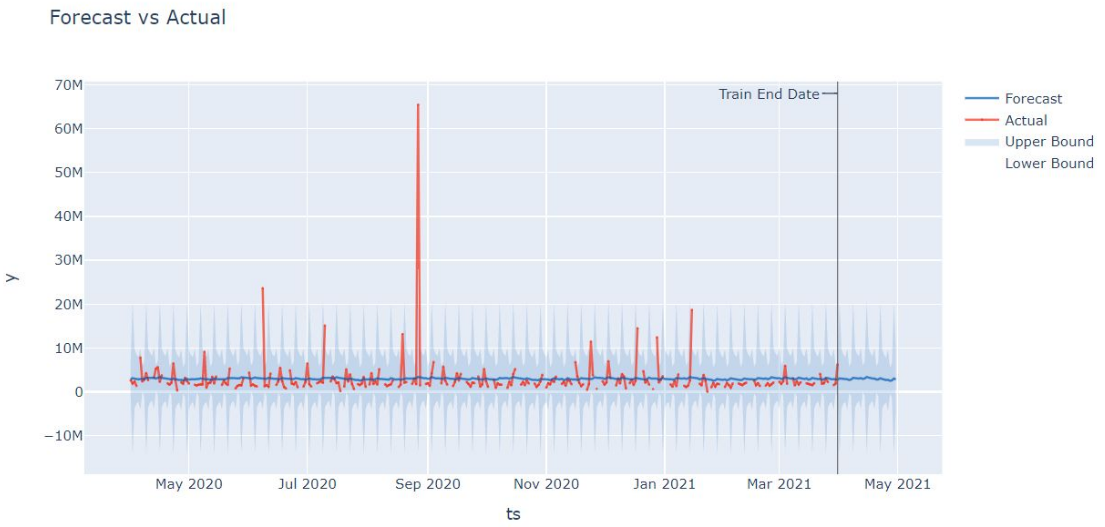
MANHATTAN ARMA



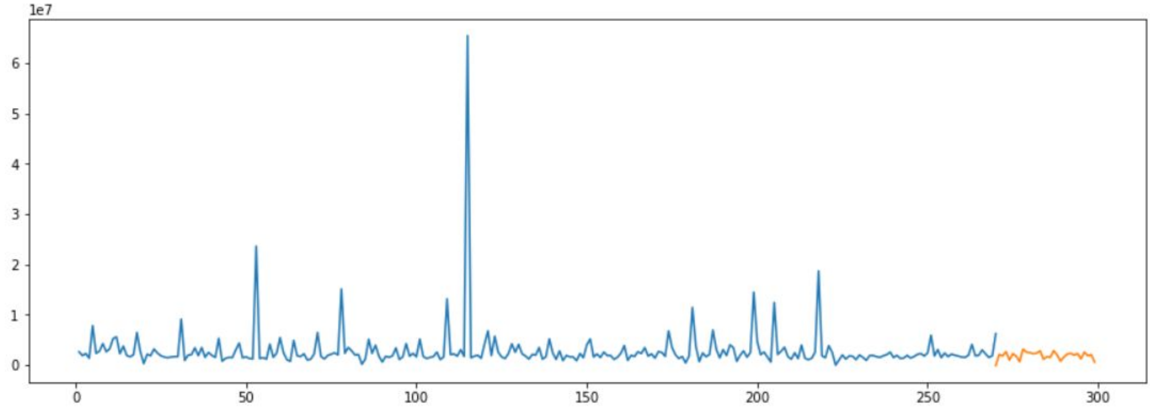
MANHATTAN ARMA



MANHATTAN - LinkedIn GreyKite Forecast



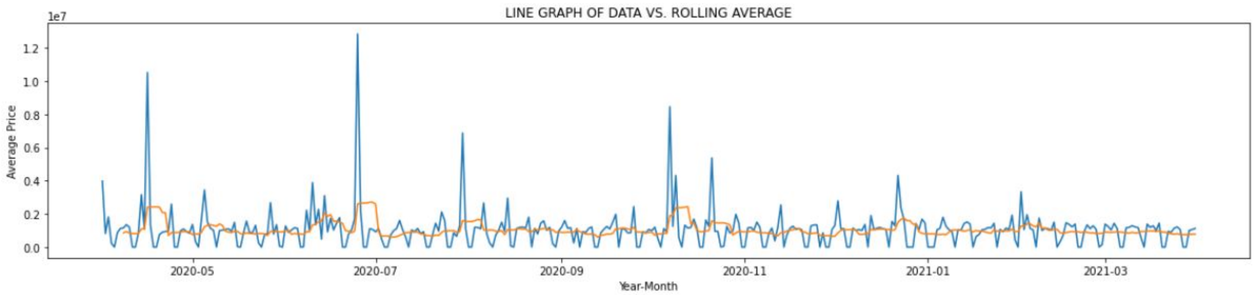
MANHATTAN - LSTM Model Forecast



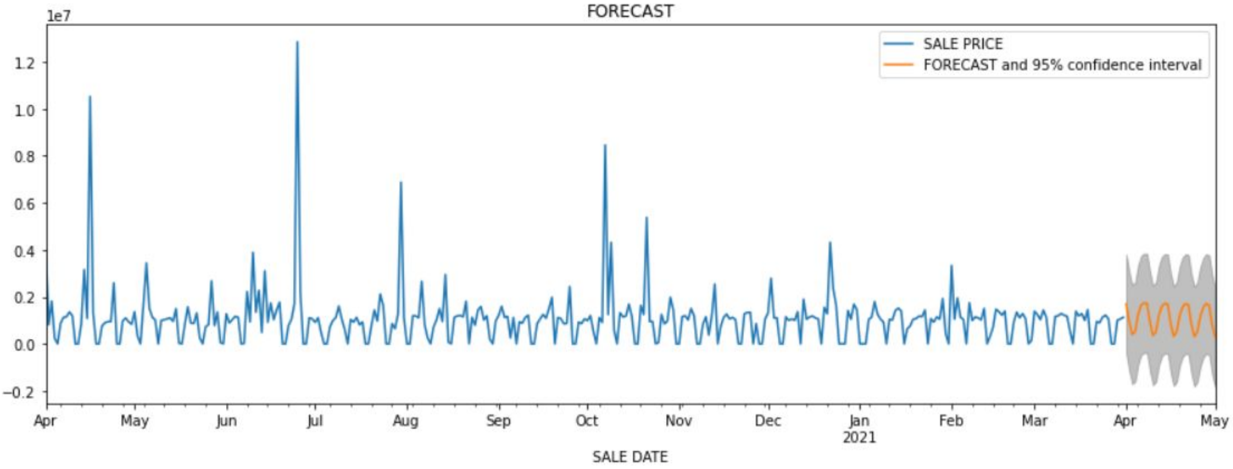
Observations:

1. Predictions seem lower than the origional data for Manhattan

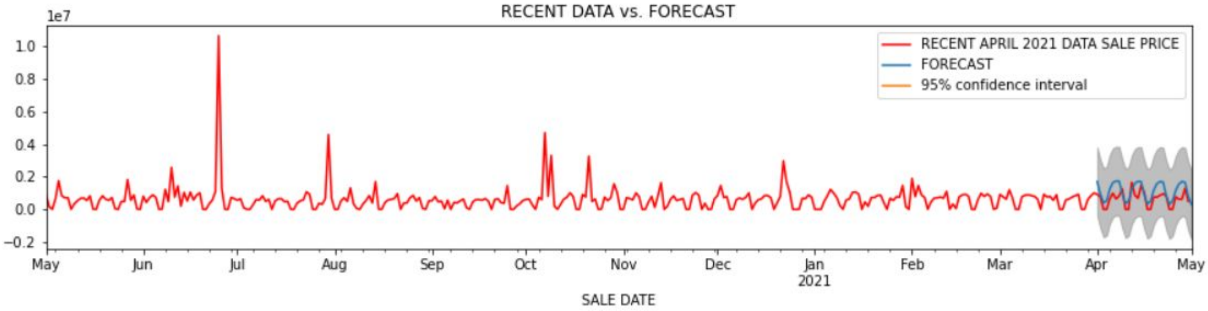
BROOKLYN ARMA



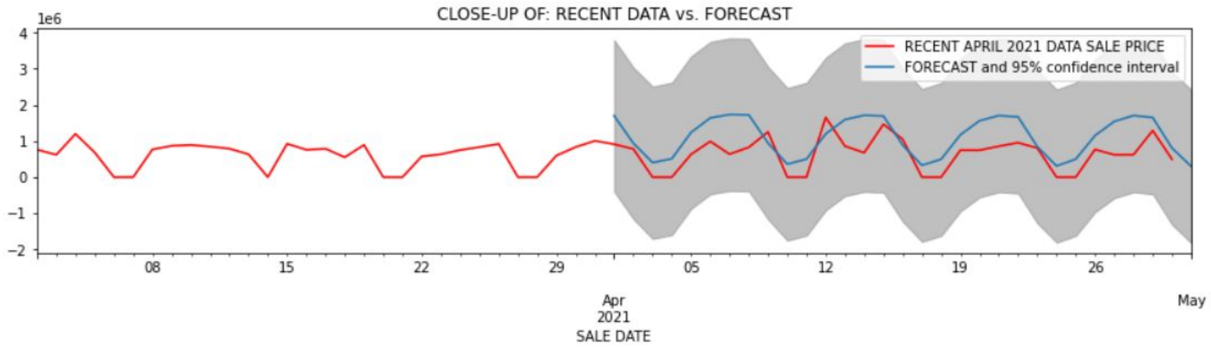
BROOKLYN ARMA



BROOKLYN ARMA



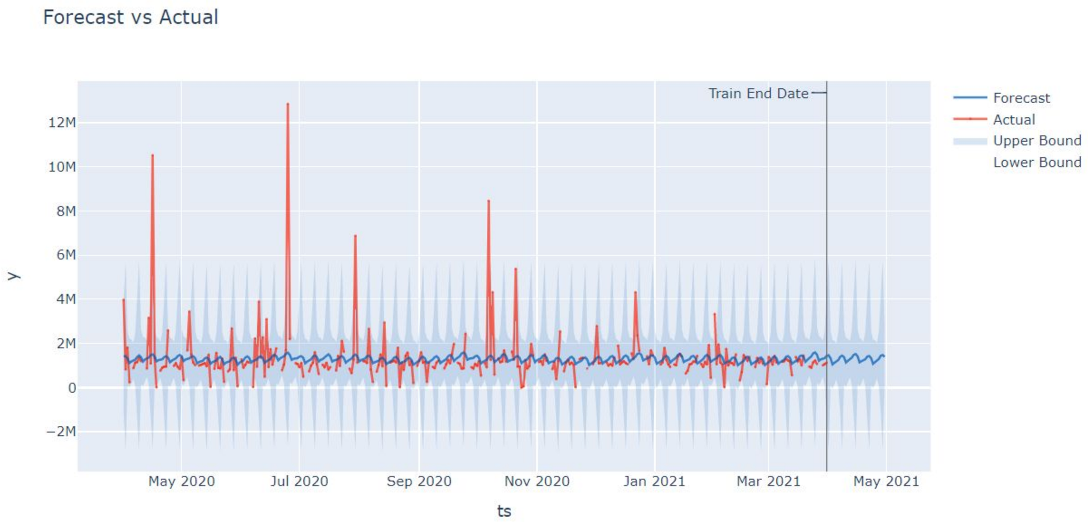
BROOKLYN ARMA



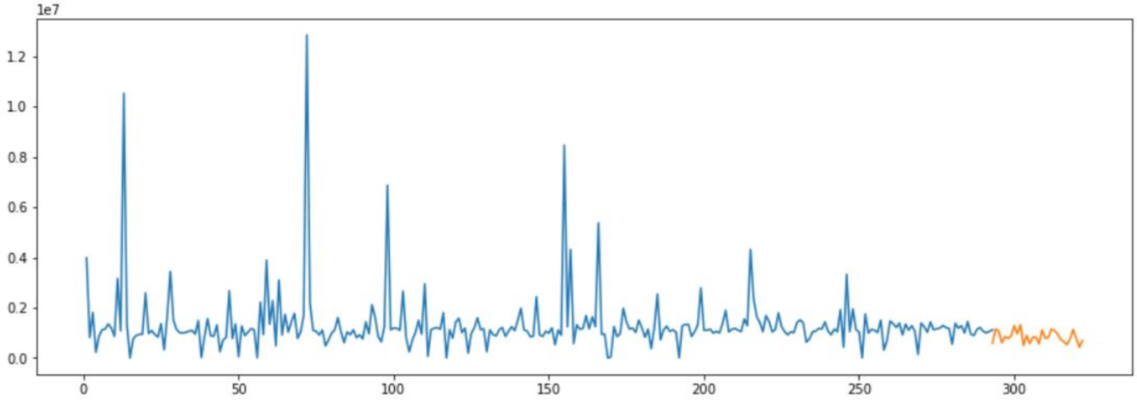
Observation

- We see that the model looks like it fits well versus the test data of 4/1/2021 until 4/31/2021

BROOKLYN - LinkedIn GreyKite Forecast



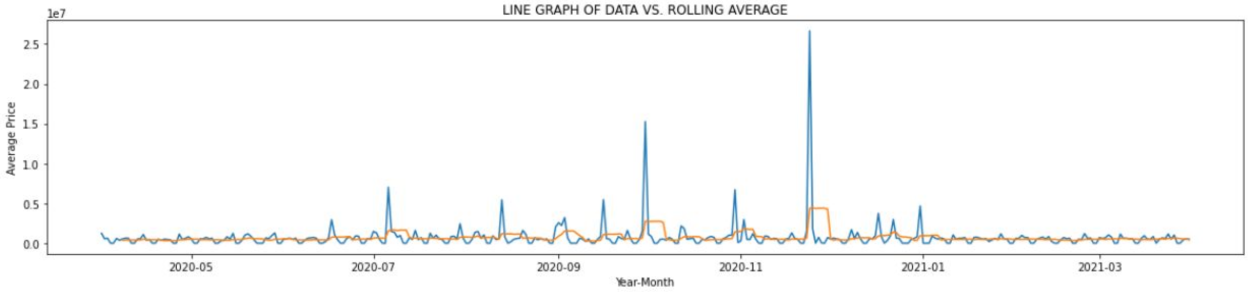
BROOKLYN - LSTM Model Forecast



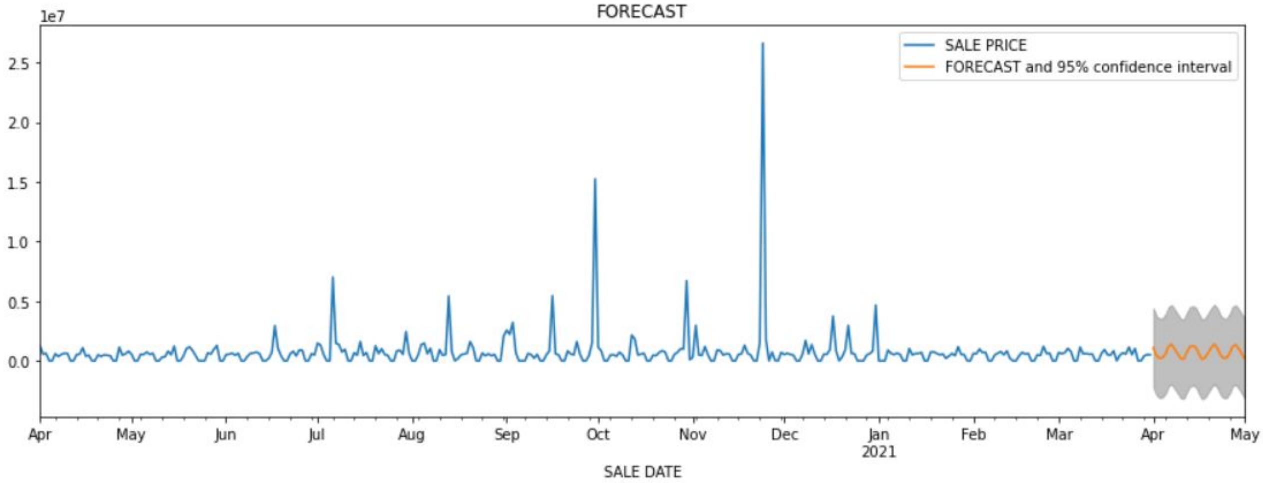
Observation:

Brooklyn prices are also predicted to be lower per this model.

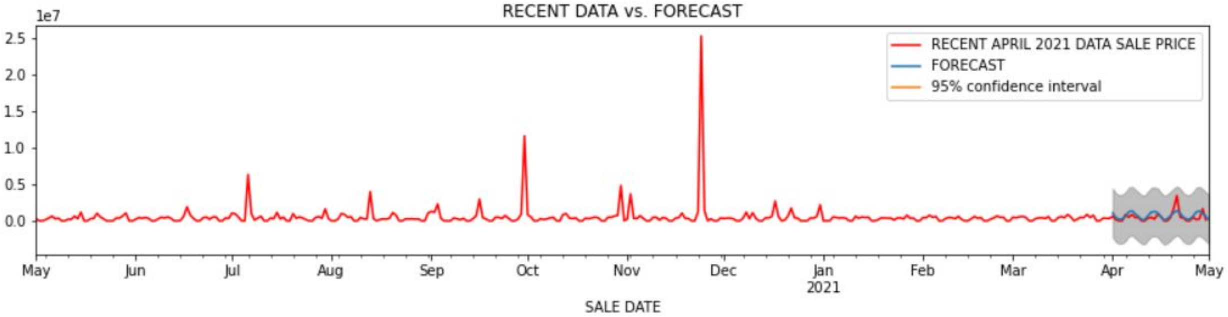
BRONX ARMA


LINE GRAPH OF DATA VS. ROLLING AVERAGE
1e7
2.5
2.0
1.5
1.0
0.5
0.0
Average Price
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
Year-Month

BRONX ARMA

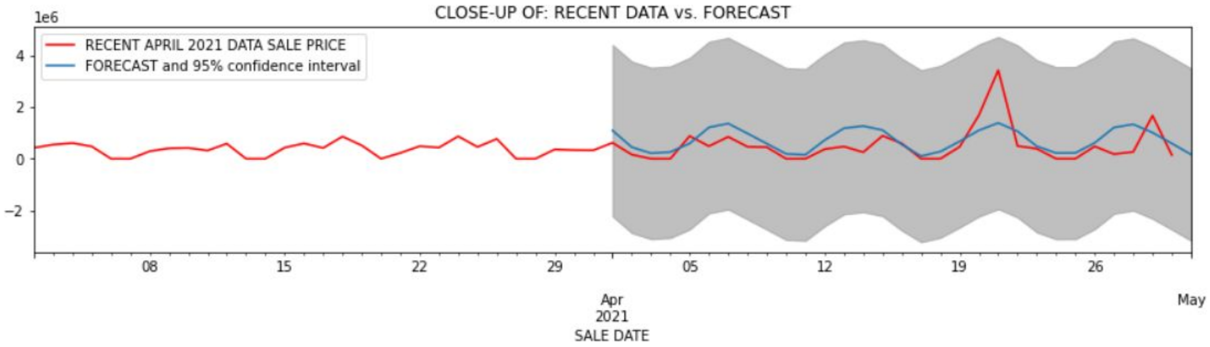

FORECAST
1e7
2.5
2.0
1.5
1.0
0.5
0.0
SALE PRICE
FORECAST and 95% confidence interval
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
2021
Feb
Mar
Apr
May
SALE DATE

BRONX ARMA

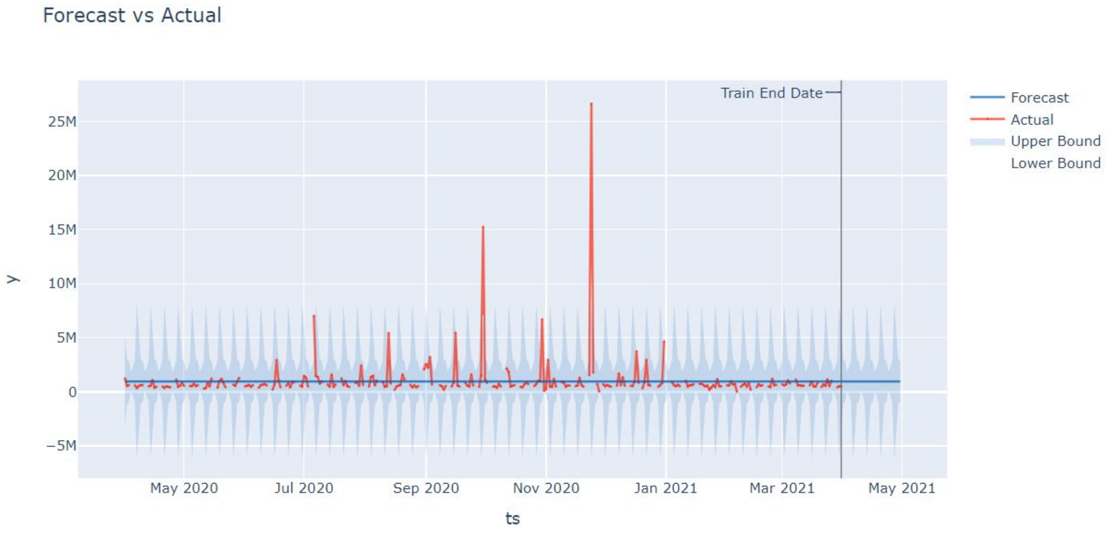


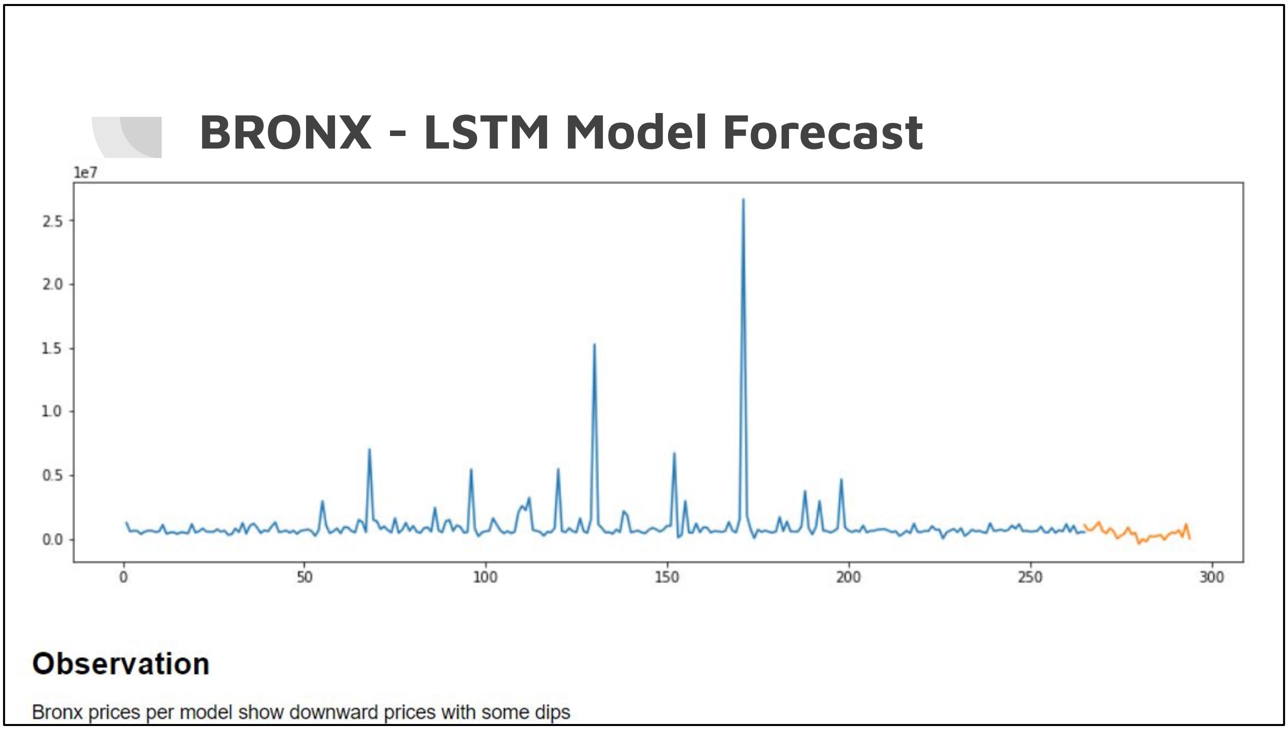
RECENT DATA vs. FORECAST
1e7
2.5
2.0
1.5
1.0
0.5
0.0
RECENT APRIL 2021 DATA SALE PRICE
FORECAST
95% confidence interval
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
2021
Feb
Mar
Apr
May
SALE DATE

BRONX ARMA


CLOSE-UP OF: RECENT DATA vs. FORECAST
1e6
RECENT APRIL 2021 DATA SALE PRICE
FORECAST and 95% confidence interval
4
2
0
−2
08
15
22
29
05
12
19
26
Apr
2021
May
SALE DATE

BRONX - LinkedIn GreyKite Forecast



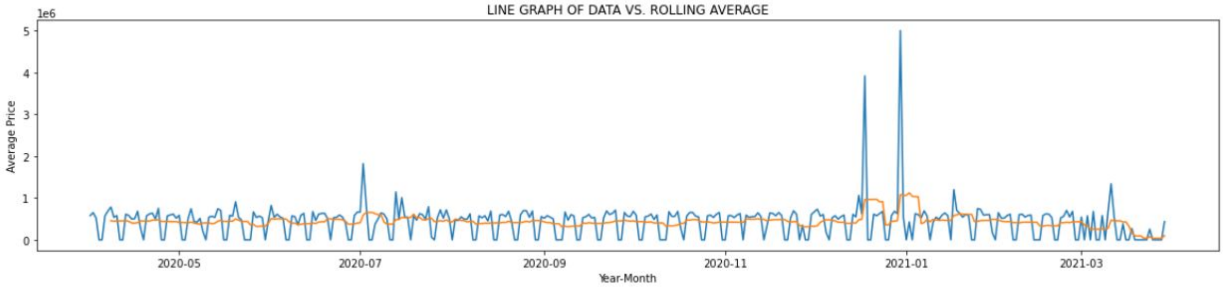


BRONX - LSTM Model Forecast

Observation

Bronx prices per model show downward prices with some dips

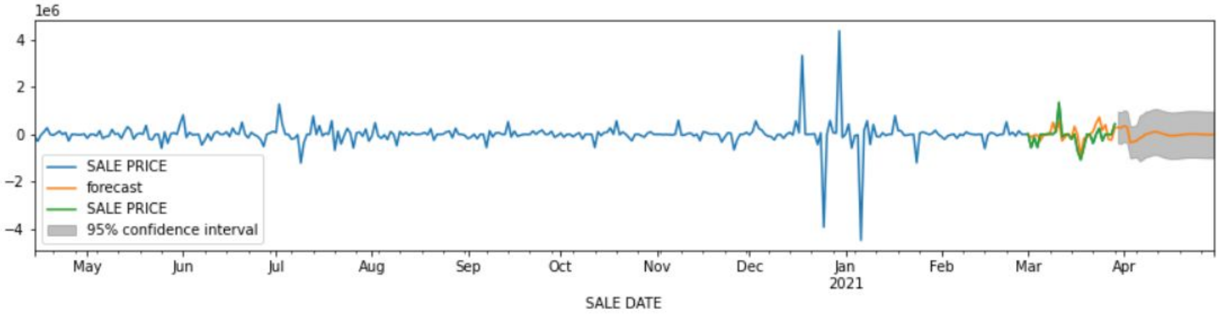
STATEN ISLAND ARMA



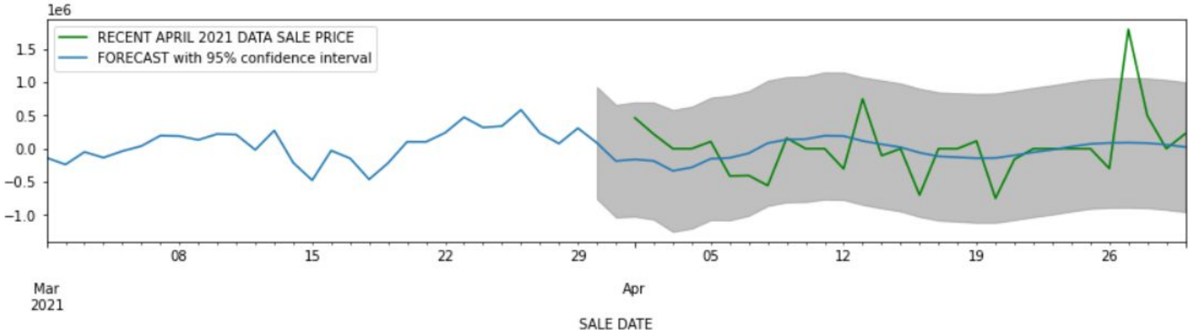
Observation

```
- The spikes in the data where the price goes to the millions or tens of millions is due to buildings being bought.
- Other than that, the rest are residential properties well under a million in price
- Near the end of the last month, there is a drop in sales
```

STATEN ISLAND ARMA

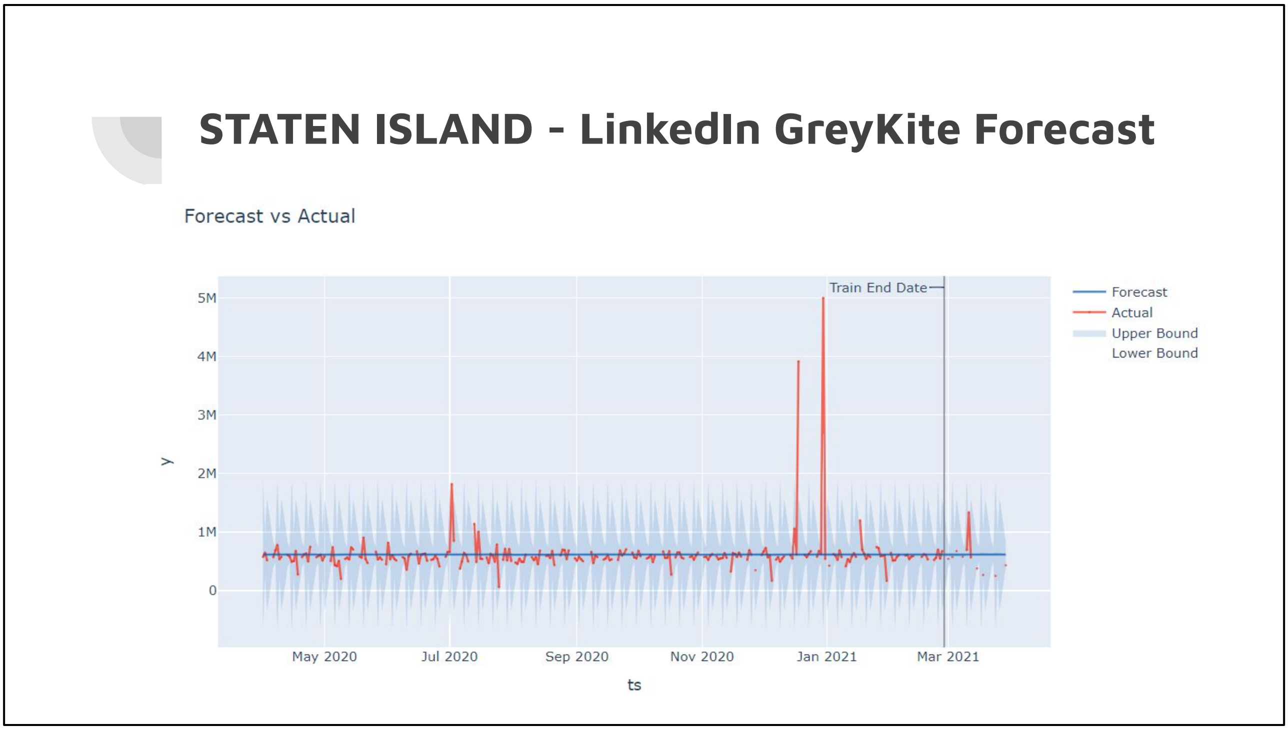


STATEN ISLAND ARMA



Observation

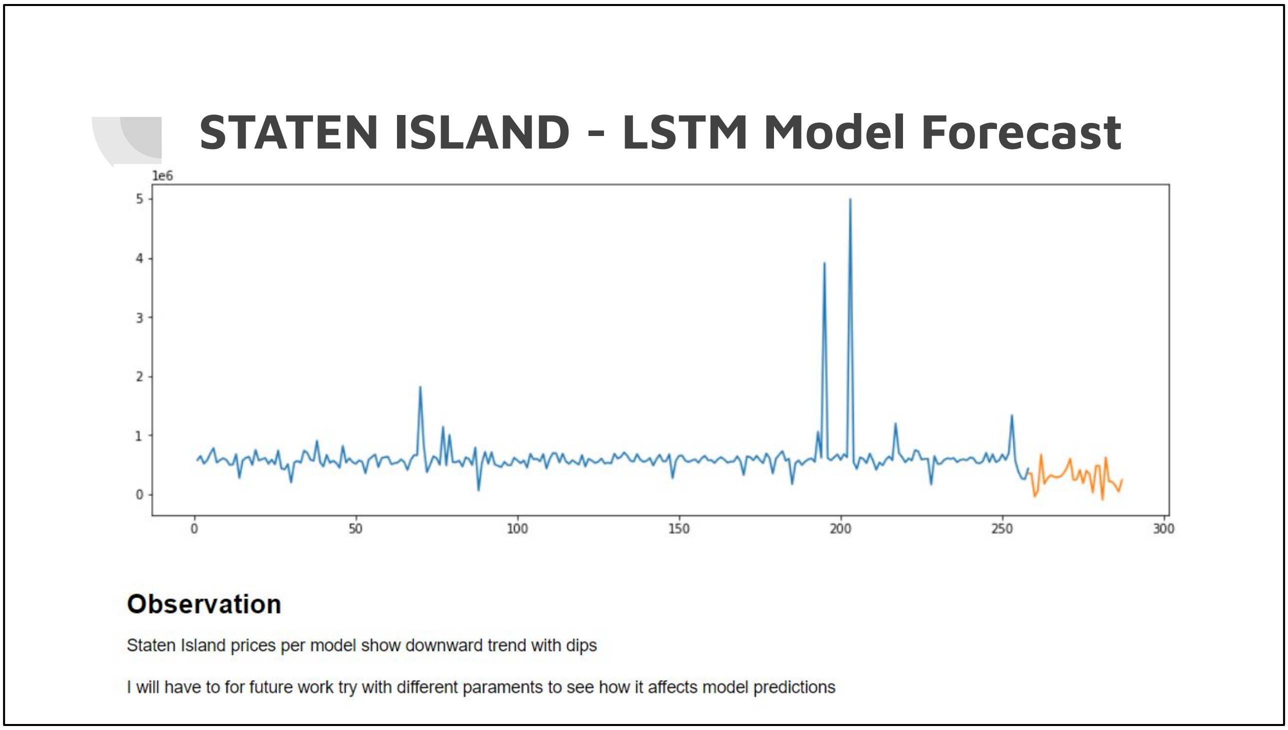
- The model does not look like it fits well



STATEN ISLAND - LinkedIn GreyKite Forecast

We see here that since Staten Island had less sales per day than the other boroughs which reflects in the gaps in the data where there were days with no sales.

For future work, we will have to revisit this dataset with different methods such as “intermittent time period forecasting.”



STATEN ISLAND - LSTM Model Forecast

Observation

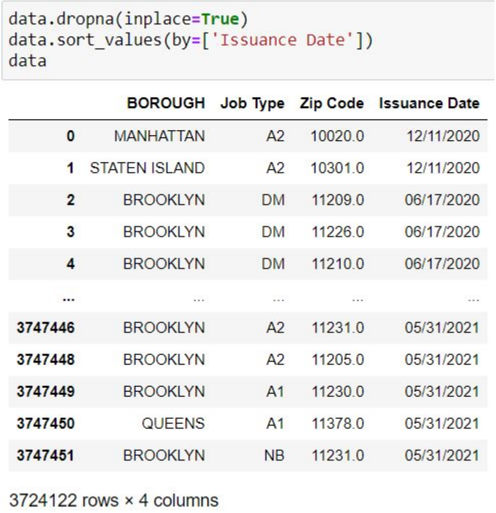
Staten Island prices per model show downward trend with dips

I will have to for future work try with different paraments to see how it affects model predictions

We see here that since Staten Island had less sales per day than the other boroughs which reflects in the gaps in the data where there were days with no sales.

For future work, we will have to revisit this dataset with different methods such as “intermittent time period forecasting.”

Permit Data



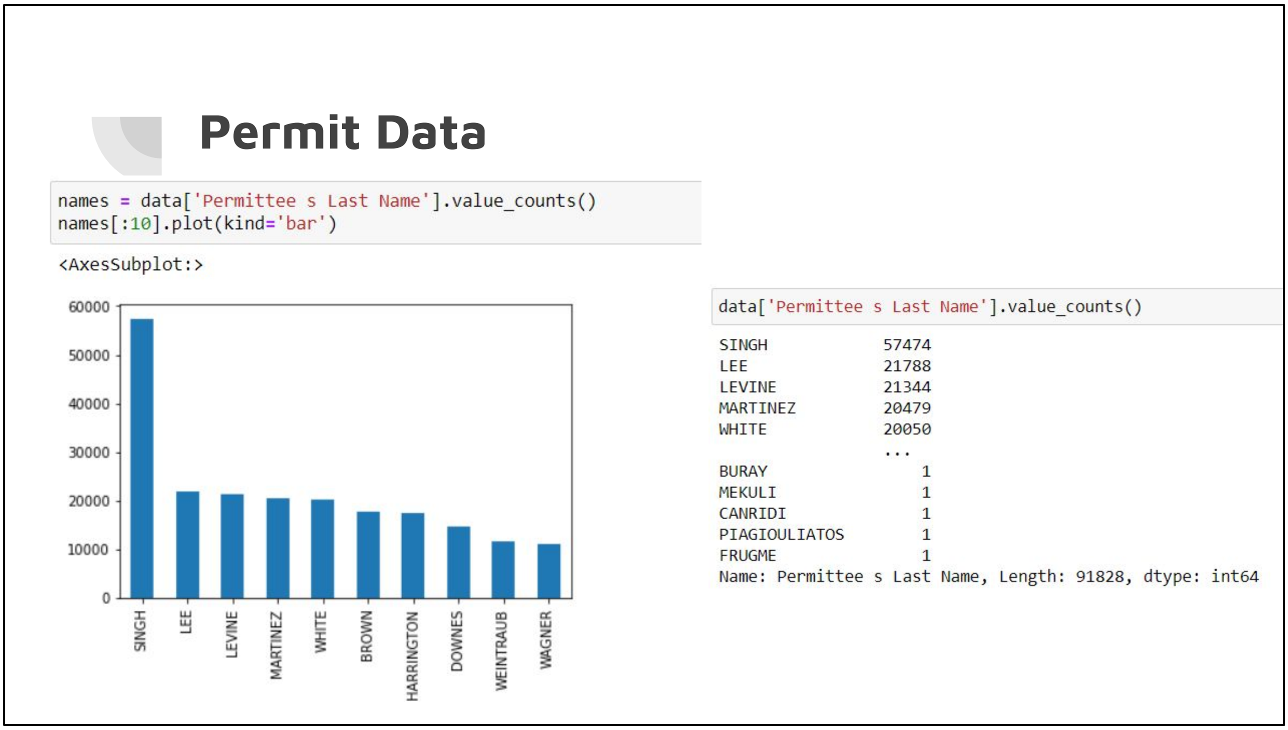
```
data.dropna(inplace=True)
data.sort_values(by=['Issuance Date'])
data
```

	BOROUGH	Job Type	Zip Code	Issuance Date
0	MANHATTAN	A2	10020.0	12/11/2020
1	STATEN ISLAND	A2	10301.0	12/11/2020
2	BROOKLYN	DM	11209.0	06/17/2020
3	BROOKLYN	DM	11226.0	06/17/2020
4	BROOKLYN	DM	11210.0	06/17/2020
...	...	...	...	...
3747446	BROOKLYN	A2	11231.0	05/31/2021
3747448	BROOKLYN	A2	11205.0	05/31/2021
3747449	BROOKLYN	A1	11230.0	05/31/2021
3747450	QUEENS	A1	11378.0	05/31/2021
3747451	BROOKLYN	NB	11231.0	05/31/2021

3724122 rows × 4 columns

This dataset was huge and reflected all filings of permits in the past 6 months

A quick analysis of names shows how the permits are broken down by the top people in this field.



Permit Data

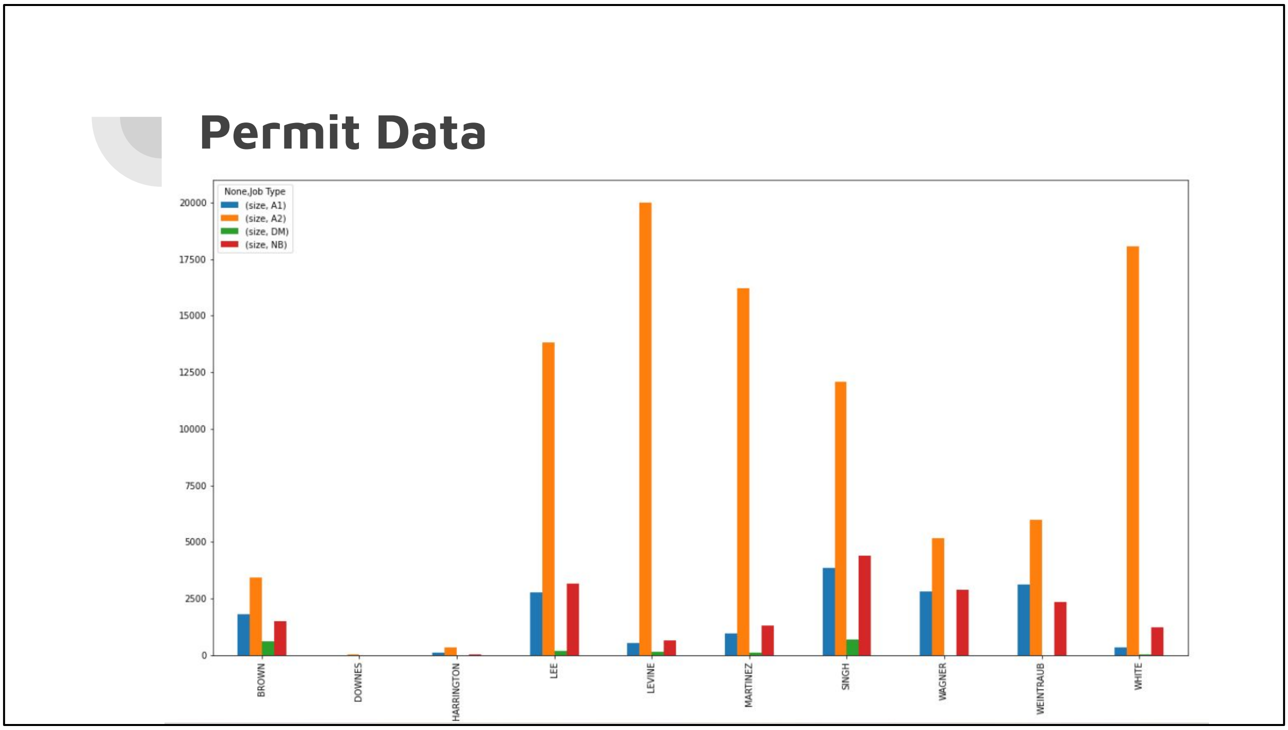
```
names = data['Permittee s Last Name'].value_counts()
names[:10].plot(kind='bar')
```

```
<AxesSubplot:>
```

```
data['Permittee s Last Name'].value_counts()
```

```
SINGH            57474
LEE              21788
LEVINE           21344
MARTINEZ         20479
WHITE            20050
                 ...  
BURAY                1
MEKULI               1
CANRIDI              1
PIAGIOULIATOS        1
FRUGME               1
Name: Permittee s Last Name, Length: 91828, dtype: int64
```

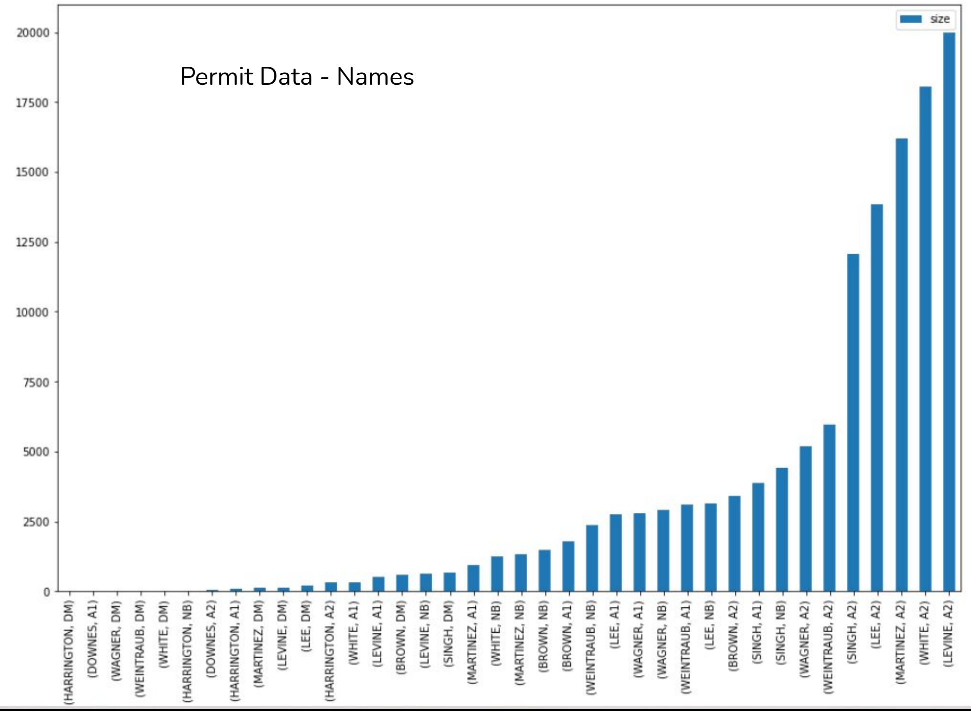
This dataset was huge and reflected all filings of permits in the past 6 months

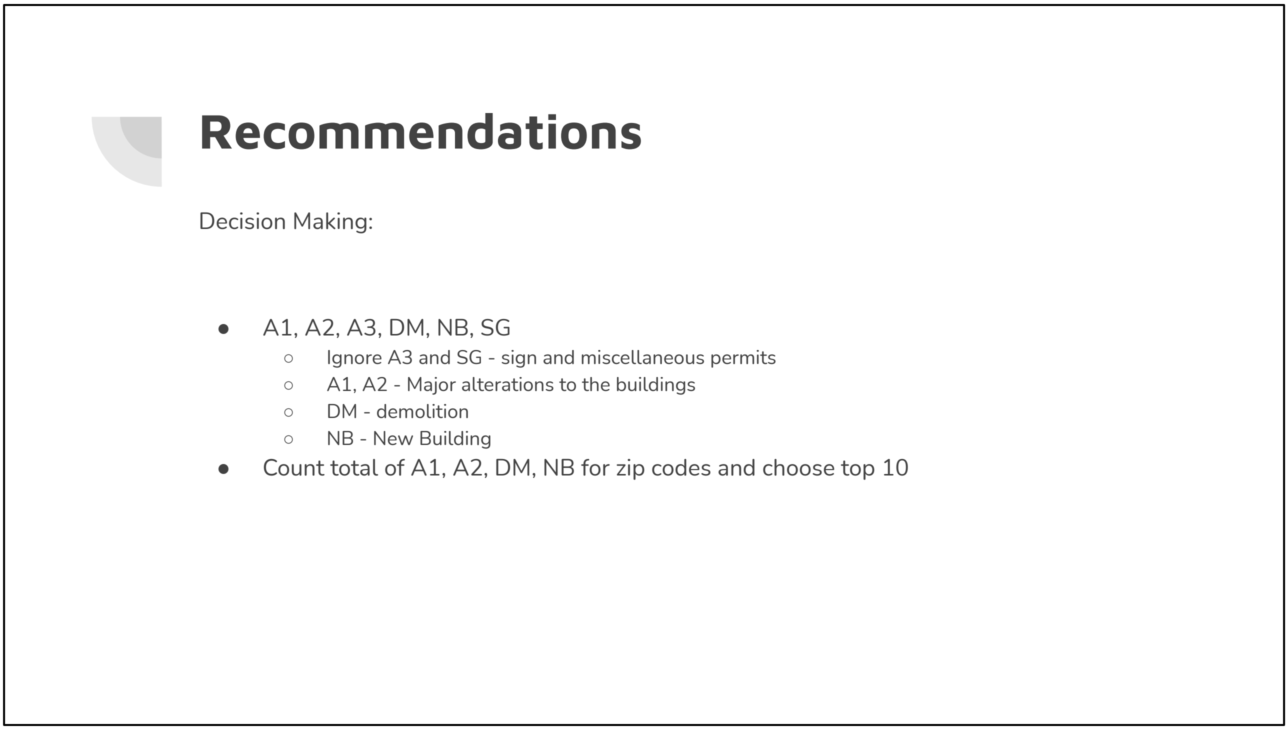


Permit Data

We saw that Singh was leading in permits but upon a closer look they are leading in new building permits. Levine is the winner here with A2 permit type.

Permit Data - Names





Recommendations

Decision Making:

- A1, A2, A3, DM, NB, SG
  - Ignore A3 and SG - sign and miscellaneous permits
  - A1, A2 - Major alterations to the buildings
  - DM - demolition
  - NB - New Building
- Count total of A1, A2, DM, NB for zip codes and choose top 10

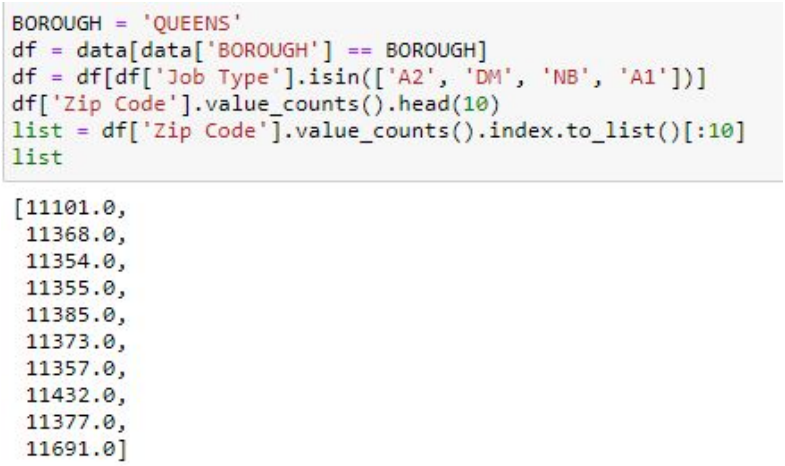
Decision Making:

We chose based on what we saw for permit data and what websites have told about different alteration types.

A2 is the major value add and fixes to buildings.

- A1, A2, A3, DM, NB, SG
  - Ignore A3 and SG - sign and miscellaneous permits
  - A1, A2 - Major alterations to the buildings
  - DM - demolition
  - NB - New Building
- Count total of A1, A2, DM, NB for zip codes and choose top 10

Recommendations - QUEENS Zip Codes

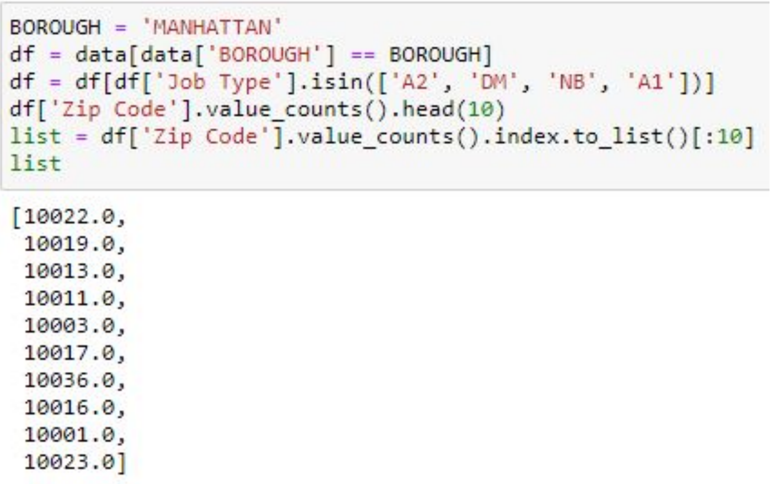


```
BOROUGH = 'QUEENS'
df = data[data['BOROUGH'] == BOROUGH]
df = df[df['Job Type'].isin(['A2', 'DM', 'NB', 'A1'])]
df['Zip Code'].value_counts().head(10)
list = df['Zip Code'].value_counts().index.to_list()[:10]
list
```

```
[11101.0,
 11368.0,
 11354.0,
 11355.0,
 11385.0,
 11373.0,
 11357.0,
 11432.0,
 11377.0,
 11691.0]
```

Based on the permit data, these are our recommendations per zip code. You can view these at your time later but these zip codes are where the most alterations and building is being done.

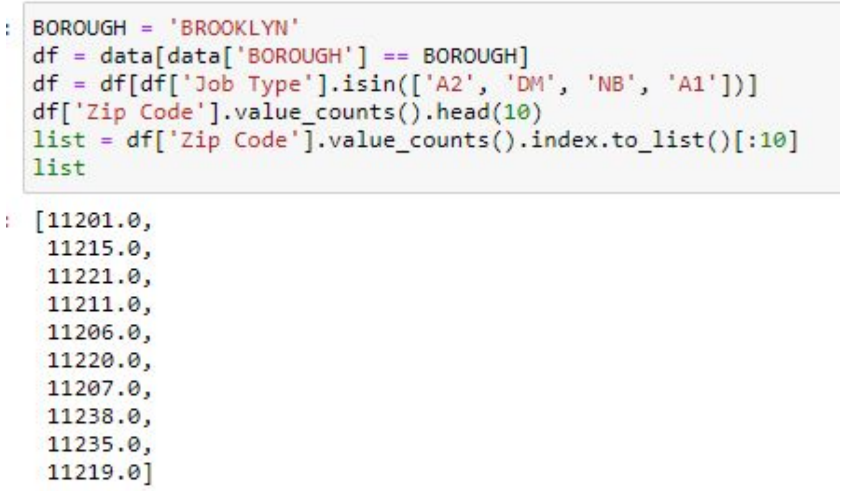
Recommendations - MANHATTAN Zip Codes



```
BOROUGH = 'MANHATTAN'
df = data[data['BOROUGH'] == BOROUGH]
df = df[df['Job Type'].isin(['A2', 'DM', 'NB', 'A1'])]
df['Zip Code'].value_counts().head(10)
list = df['Zip Code'].value_counts().index.to_list()[:10]
list
```

```
[10022.0,
 10019.0,
 10013.0,
 10011.0,
 10003.0,
 10017.0,
 10036.0,
 10016.0,
 10001.0,
 10023.0]
```

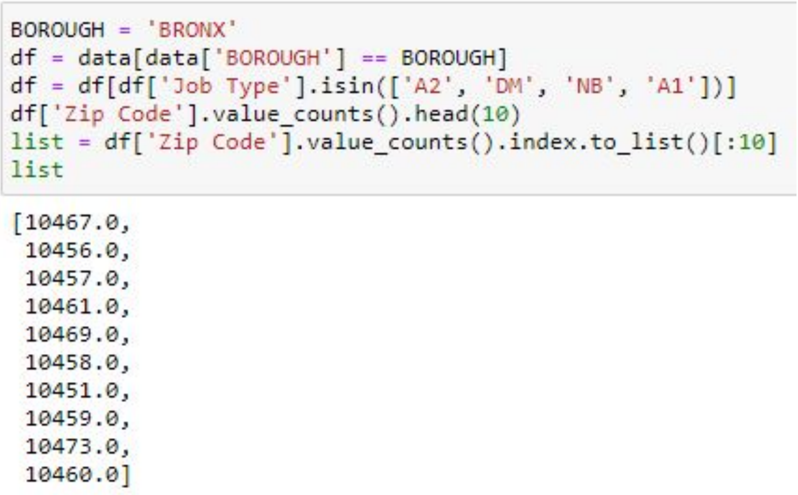
Recommendations - BROOKLYN Zip Codes



```
BOROUGH = 'BROOKLYN'
df = data[data['BOROUGH'] == BOROUGH]
df = df[df['Job Type'].isin(['A2', 'DM', 'NB', 'A1'])]
df['Zip Code'].value_counts().head(10)
list = df['Zip Code'].value_counts().index.to_list()[:10]
list
```

```
[11201.0,
 11215.0,
 11221.0,
 11211.0,
 11206.0,
 11220.0,
 11207.0,
 11238.0,
 11235.0,
 11219.0]
```

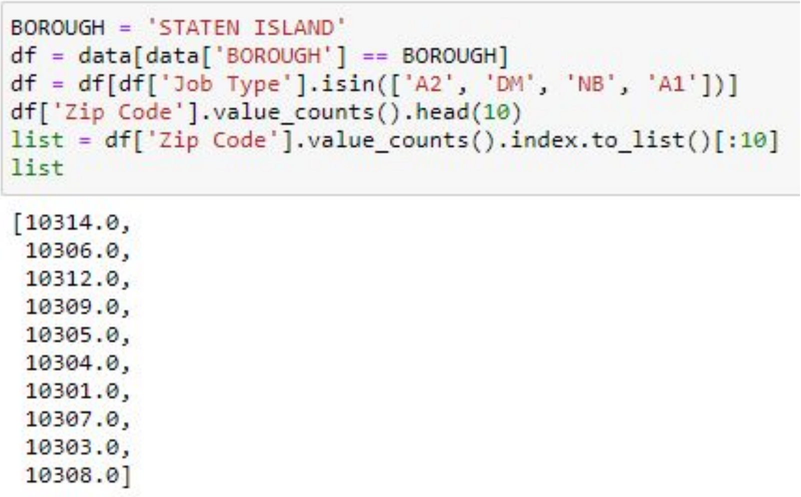
Recommendations - BRONX Zip Codes



```
BOROUGH = 'BRONX'
df = data[data['BOROUGH'] == BOROUGH]
df = df[df['Job Type'].isin(['A2', 'DM', 'NB', 'A1'])]
df['Zip Code'].value_counts().head(10)
list = df['Zip Code'].value_counts().index.to_list()[:10]
list
```

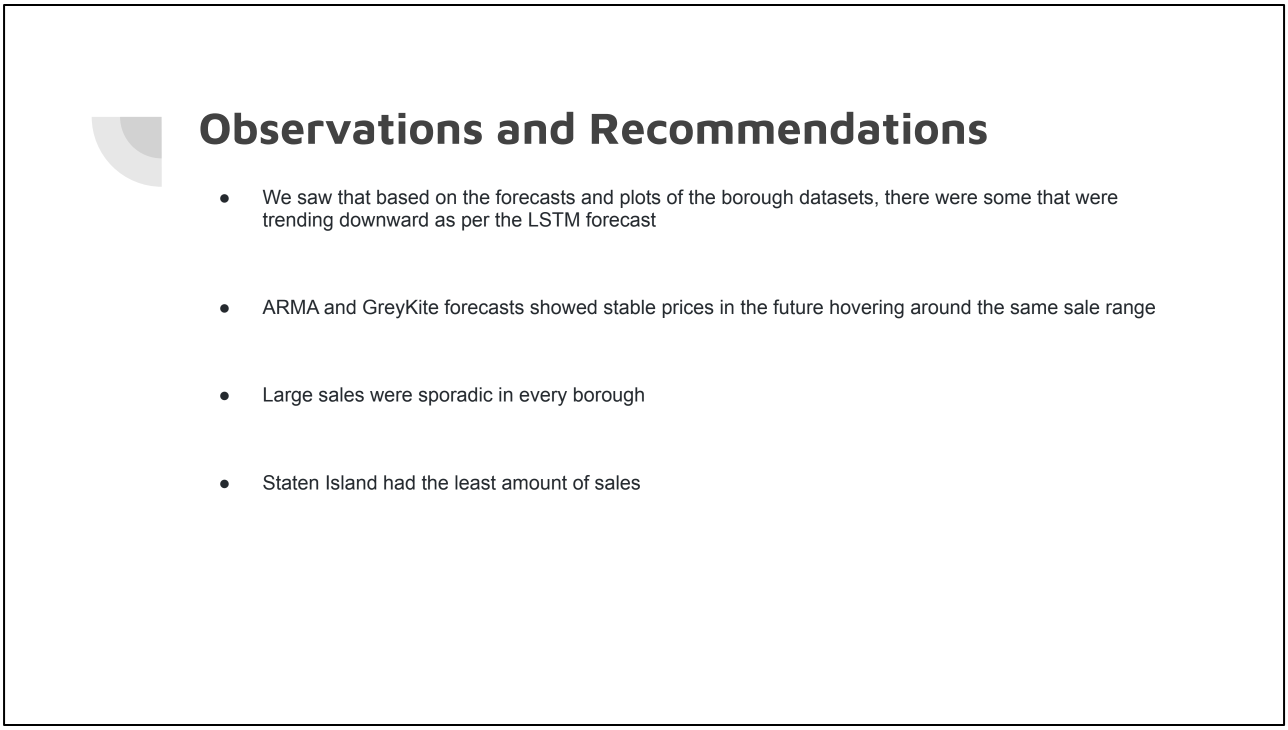
```
[10467.0,
 10456.0,
 10457.0,
 10461.0,
 10469.0,
 10458.0,
 10451.0,
 10459.0,
 10473.0,
 10460.0]
```

Recommendations - STATEN ISLAND



```
BOROUGH = 'STATEN ISLAND'
df = data[data['BOROUGH'] == BOROUGH]
df = df[df['Job Type'].isin(['A2', 'DM', 'NB', 'A1'])]
df['Zip Code'].value_counts().head(10)
list = df['Zip Code'].value_counts().index.to_list()[:10]
list
```

```
[10314.0,
 10306.0,
 10312.0,
 10309.0,
 10305.0,
 10304.0,
 10301.0,
 10307.0,
 10303.0,
 10308.0]
```

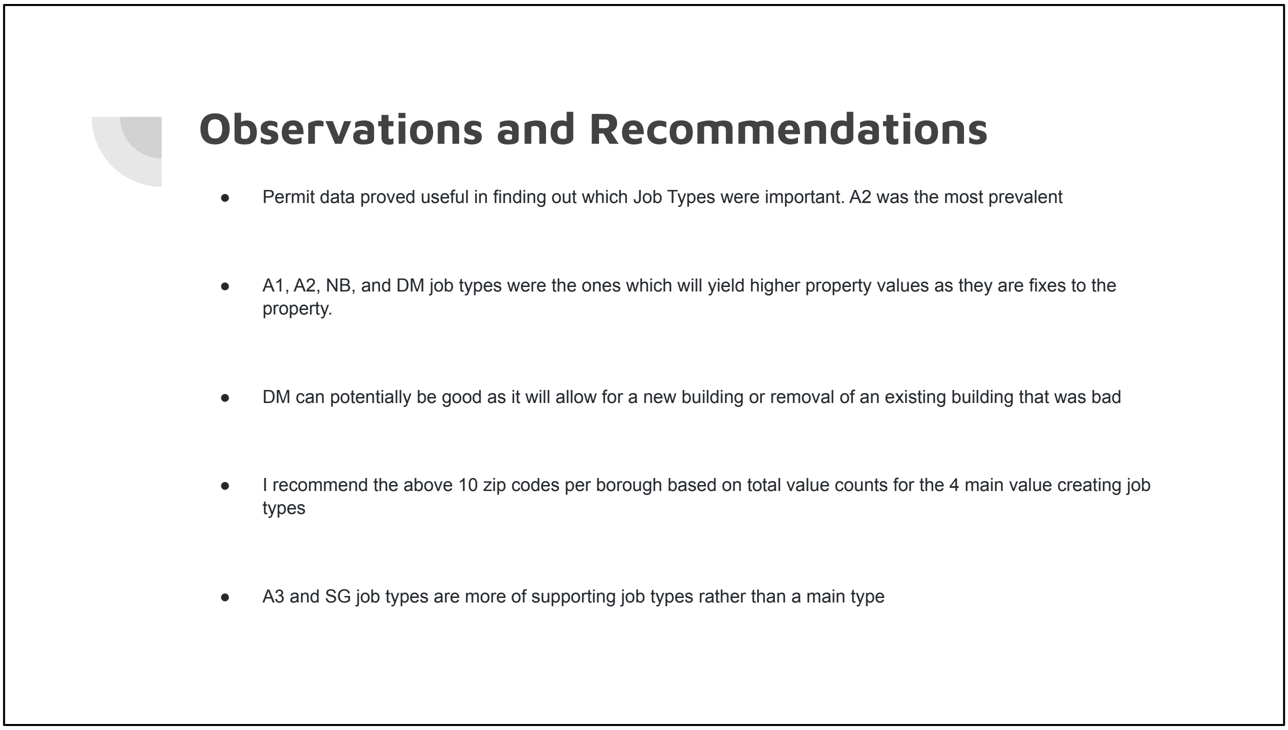


Observations and Recommendations

- We saw that based on the forecasts and plots of the borough datasets, there were some that were trending downward as per the LSTM forecast
- ARMA and GreyKite forecasts showed stable prices in the future hovering around the same sale range
- Large sales were sporadic in every borough
- Staten Island had the least amount of sales

Observations/Recommendations

1. We saw that based on the forecasts and plots of the borough datasets, there were some that were trending downward as per the LSTM forecast
2. ARMA and GreyKite forecasts showed stable prices in the future hovering around the same sale range
3. Building sales were sporadic in every borough
4. Staten Island had the least amount of sales
5. Permit data proved useful in finding out which Job Types were important. A2 was the most prevalent
6. A1, A2, NB, and DM job types were the ones which will yield higher property values as they are fixes to the property.
7. DM can potentially be good as it will allow for a new building or removal of an existing building that was bad
8. I recommend the above 10 zip codes per borough based on total value counts for the 4 main value creating job types
9. A3 and SG job types are more of supporting job types rather than a main type

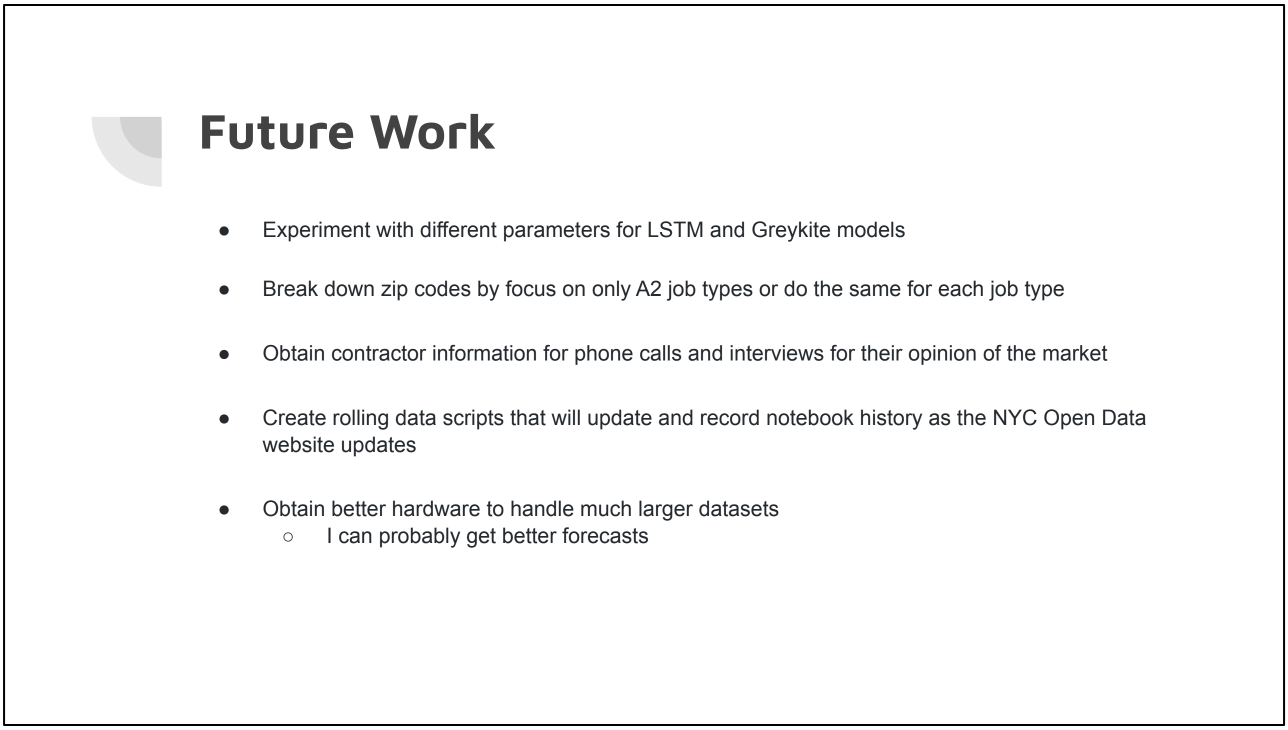


Observations and Recommendations

- Permit data proved useful in finding out which Job Types were important. A2 was the most prevalent
- A1, A2, NB, and DM job types were the ones which will yield higher property values as they are fixes to the property.
- DM can potentially be good as it will allow for a new building or removal of an existing building that was bad
- I recommend the above 10 zip codes per borough based on total value counts for the 4 main value creating job types
- A3 and SG job types are more of supporting job types rather than a main type

Observations/Recommendations

1. We saw that based on the forecasts and plots of the borough datasets, there were some that were trending downward as per the LSTM forecast
2. ARMA and GreyKite forecasts showed stable prices in the future hovering around the same sale range
3. Building sales were sporadic in every borough
4. Staten Island had the least amount of sales
5. Permit data proved useful in finding out which Job Types were important. A2 was the most prevalent
6. A1, A2, NB, and DM job types were the ones which will yield higher property values as they are fixes to the property.
7. DM can potentially be good as it will allow for a new building or removal of an existing building that was bad
8. I recommend the above 10 zip codes per borough based on total value counts for the 4 main value creating job types
9. A3 and SG job types are more of supporting job types rather than a main type

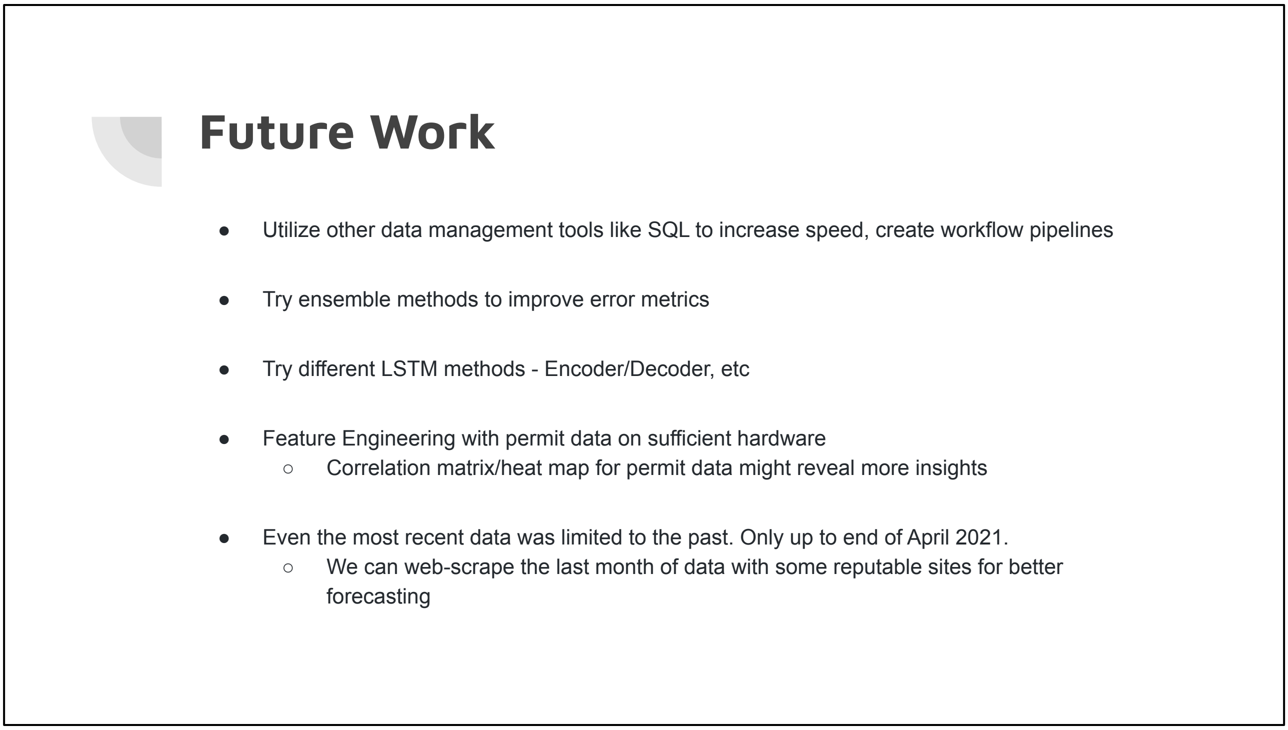


Future Work

- Experiment with different parameters for LSTM and Greykite models
- Break down zip codes by focus on only A2 job types or do the same for each job type
- Obtain contractor information for phone calls and interviews for their opinion of the market
- Create rolling data scripts that will update and record notebook history as the NYC Open Data website updates
- Obtain better hardware to handle much larger datasets
  - I can probably get better forecasts

Future Work

1. Experiment with different parameters for LSTM and Greykite models
2. Break down zip codes by focus on only A2 job types or do the same for each job type
3. Obtain contractor information for phone calls and interviews for their opinion of the market
4. Create rolling data scripts that will update and record notebook history as the NYC Open Data website updates
5. Obtain better hardware to handle much larger datasets
   i. I can probably get better forecasts
6. Utilize other data management tools like SQL to increase speed, create workflow pipelines
7. Try ensemble methods to improve error metrics
8. Try different LSTM methods - Encoder/Decoder, etc
9. Feature Engineering with permit data on sufficient hardware
   i. Correlation matrix/heat map for permit data might reveal more insights
10. Even the most recent data was limited to the past. Only up to end of April.
    i. We can webscrape the last month of data with some reputable sites to better have the latest predictions



Future Work

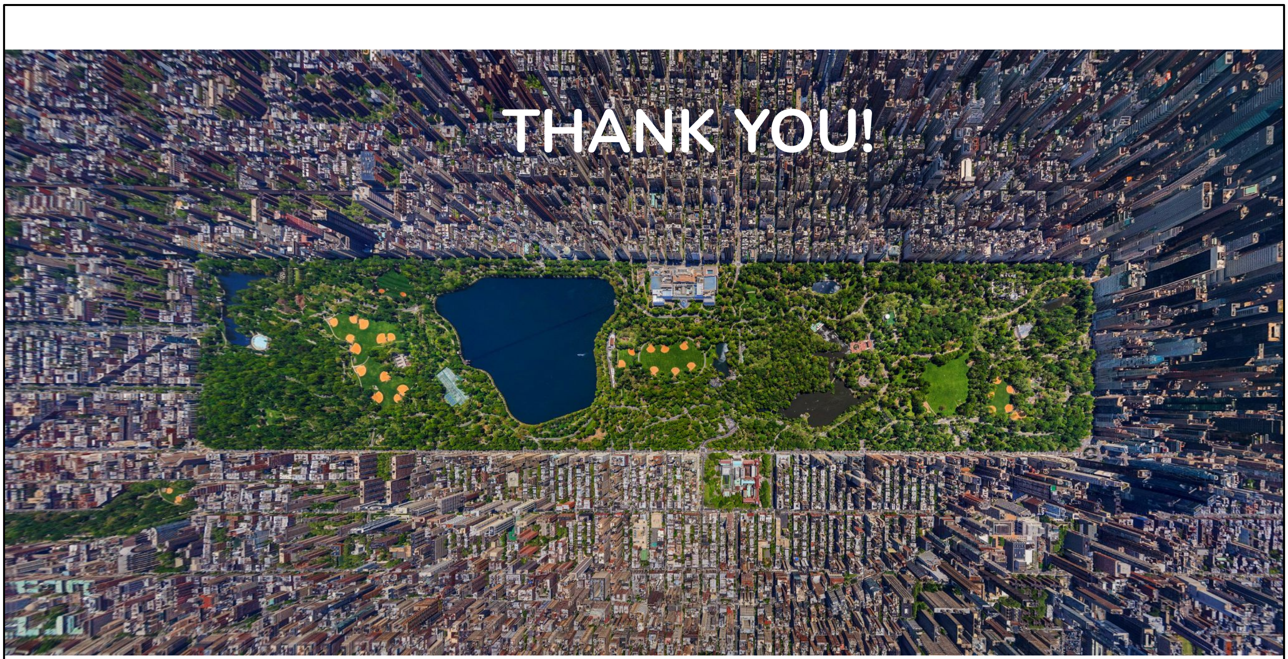
- Utilize other data management tools like SQL to increase speed, create workflow pipelines
- Try ensemble methods to improve error metrics
- Try different LSTM methods - Encoder/Decoder, etc
- Feature Engineering with permit data on sufficient hardware
  - Correlation matrix/heat map for permit data might reveal more insights
- Even the most recent data was limited to the past. Only up to end of April 2021.
  - We can web-scrape the last month of data with some reputable sites for better forecasting

Reputable sites like Zillow, Trulia, and other real estate websites local to NYC

Future Work

1. Experiment with different parameters for LSTM and Greykite models
2. Break down zip codes by focus on only A2 job types or do the same for each job type
3. Obtain contractor information for phone calls and interviews for their opinion of the market
4. Create rolling data scripts that will update and record notebook history as the NYC Open Data website updates
5. Obtain better hardware to handle much larger datasets
   i. I can probably get better forecasts
6. Utilize other data management tools like SQL to increase speed, create workflow pipelines
7. Try ensemble methods to improve error metrics
8. Try different LSTM methods - Encoder/Decoder, etc
9. Feature Engineering with permit data on sufficient hardware
   i. Correlation matrix/heat map for permit data might reveal more insights
10. Even the most recent data was limited to the past. Only up to end of April.
    i. We can webscrape the last month of data with some reputable sites to better have the latest predictions

THANK YOU!



Source: https://assets.sbnation.com/assets/1994067/nyc-panorama-serget-semenov.jpeg

QUESTIONS?

Source: https://s3-us-west-2.amazonaws.com/supermoney-blog/wp-content/uploads/2016/01/house-question.jpg