README.md

NYC Real Estate Analysis during COVID and Recommendations

from: https://lp-cms-production.imgix.net/2021-05/GettyRF_494607942.jpg?auto=format&fit=crop&sharp=10&vib=20&ixlib=react-8.6.4&w=850

README Table of Contents

6. LinkedIn GreyKite forecasting of data - all NYC boroughs and observations
7. LSTM Forecasting data - all NYC boroughs and observations
8. Permit Data Analysis and decisions
9. Observations/Recommendations
10. Future Work

1. Project Goals and Abstract

The project started off initially as a means to understand for myself the real estate market before and after the peaks of the COVID pandemic within the city. Seeing no people on the streets was an eerie sight. The data I took was from the open data repository maintained by the city. NYC tracks all real estate transactions and maintained rolling datasets per borough about the sales that occured in each borough. At the time of me doing this project, I had data from 04/01/2020 up until 3/31/2021. The datasets were huge. The largest that I tried to use was 9.85 million rows and was making short work of my computer. The goals then solidified into the following:

- Analyze the boroughs to see which ones are trending and in which direction
- Forecast the data to see if the prices of the properties will go up or down
- Create 10 recommended zip codes per borough that a potential investor can look into as a lucrative area to buy property

  The last bullet point is important. Real estate transactions are still going on in the city and I want to know where there are potential hotspots to purchase property in. Upon analysis of the datasets of the five boroughs and permit data. I made recommendations on which zip codes an investor should take a look at in depth if they are looking to purchase property.

2. Project Summary

First, I went to the NYC Open Data Website and perused their volumes of records that they keep on every moving part in the city. Real Estate is no exception. They had a dataset that is continuously growing of all transactions that went back into the late 2000s. Sadly, I could not use this dataset because my hardware did not have the RAM to do so. I then downloaded the rolling data for all 5 boroughs and forecasted the data for the next 30 days with ARMA models, LinkedIN Greykite Forecasting software, and LSTM (Long-Short Term Memory) Neural Network using Keras. Upon analysis, the markets looked stable with the exception of one or two trending slightly downward. Within each of the datasets, there were massive spikes in sale price due to buildings being sold but they were being sold unpredictably. After analyzing the boroughs and seeing their trends and forecast, I looked at the permit data for each borough. Basically, contractors/architects/owners have to file for different job types based on what kind of renovation they want to do to their property. The job types I focused on were Alteration Type 1, Alteration Type 2, New Building, and Demolition. Respectively, the job type acronyms are: A1, A2, NB, and DM. I based my zip code recommendations on the value_counts of the sum of these job types in each zip code. I was able to get recommendations for each borough this way.

3. The Data Used and Challenges

The data used are as follows and are all taken from https://data.cityofnewyork.us/

NYC Citywide Rolling Calendar Sales

- Please note upon writing this project, they updated and merged all the datasets into one giant one called "NYC Citywide Rolling Calendar Sales"

  ### DOB Permit Issuance

  ### Property Valuation and Assessment Data

Challenges with the data:

1. "Property Valuation and Assessment Data" is 2GB in size. I could open it through jupyter notebook at the expense of 100% RAM utilization. If I tried to do anything with the data, the notebook crashed. I could not export any data out of this dataset. For future work, I would require more RAM in a computer or try running the notebook through a cloud VM with sufficient compute power.
2. The Rolling Calendar sales datasets had nulls and Nans that had to be dealt with
3. The column names required to be changed to string type
4. I had to make qualitative decisions on which columns to keep and what data to ignore
5. Permit Issuance Data, although smaller, was still a large file and contained 3.75 million rows. I had clean and sort through the data for my recommendations.
6. The rolling datasets came in .xls format

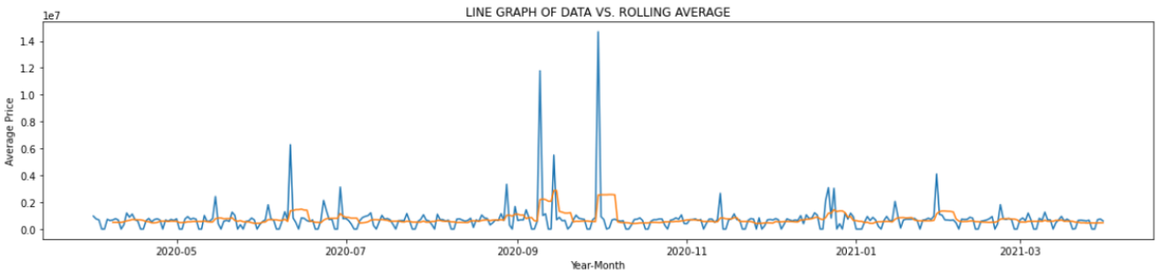
4. Cleaning and Prepping the data:

I had to make the following changes to the rolling datasets:

Change	Reason
Skip rows when loading .xls file	This allowed for proper loading of the data. There were some rows at the top which pandas was not reading properly.
Take only a couple columns	I chose: 'TAX CLASS AT PRESENT','ZIP CODE', 'RESIDENTIAL UNITS', 'TOTAL UNITS', 'LAND SQUARE FEET', 'GROSS SQUARE FEET', 'YEAR BUILT', 'TAX CLASS AT TIME OF SALE', 'SALE PRICE', 'SALE DATE', because I thought they were the most significant information I can analyze
dropna()	There were nulls and nans in the dataset which I dropped. It was not much
reset_index()	Index had to be reset after dropping rows
Manhattan specific	I chose the bare minimum relevant data because there were ALOT of missing data in the other columns, espescially the square footage column. I chose: 'TAX CLASS AT PRESENT','ZIP CODE', 'YEAR BUILT','SALE PRICE', 'SALE DATE'

5. ARMA Forecasting data - all NYC boroughs and observations

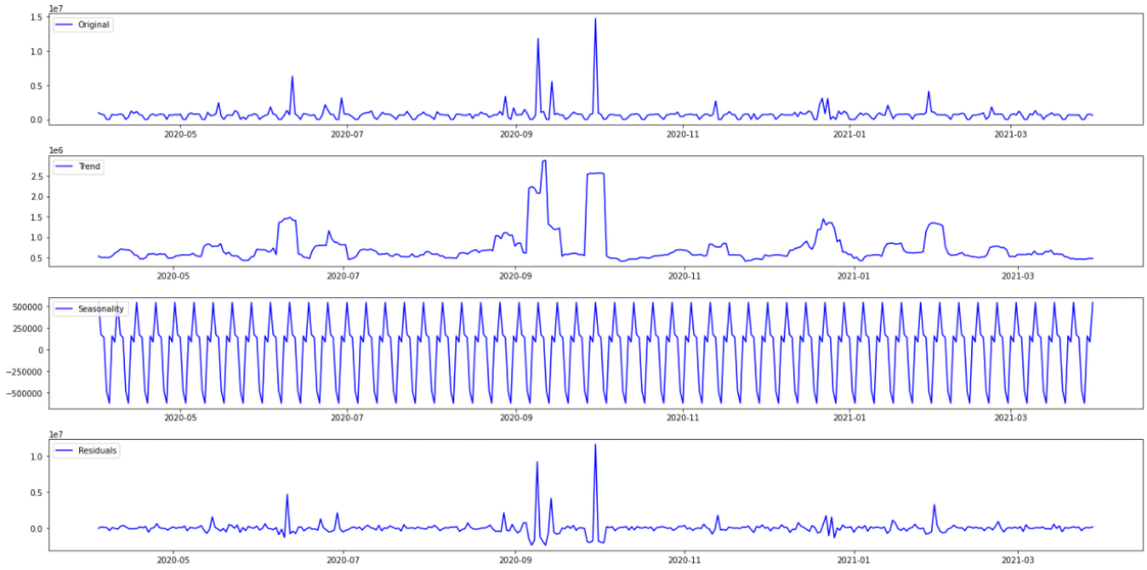
QUEENS Line Graph of Data vs 7-day rolling average



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
```

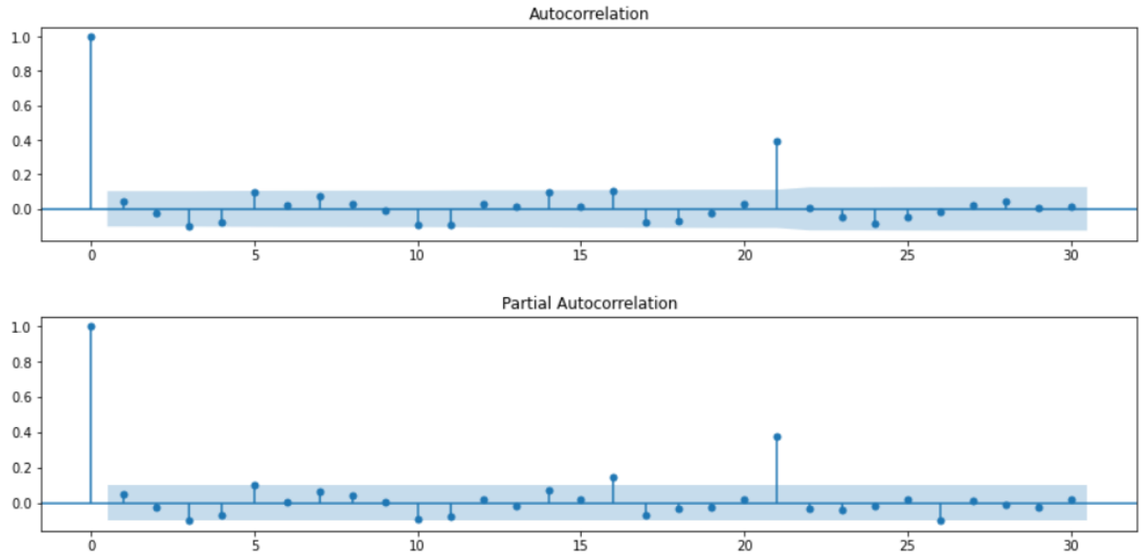
QUEENS Statsmodels decomposition



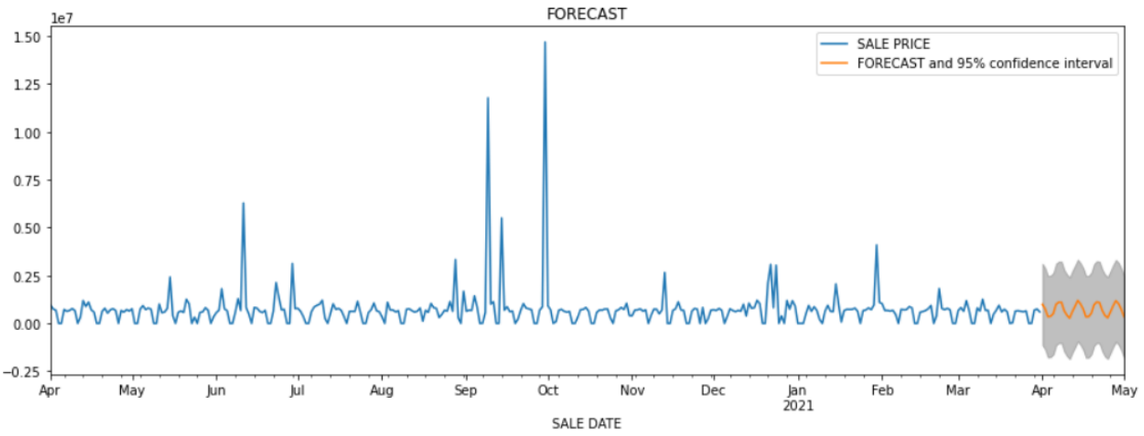
Observations:

- A large amount of sales happened between August 2020 and November 2020.
- Looks like there may be some seasonality every month

QUEENS Auto-Correlation and Partial Auto-Correlation

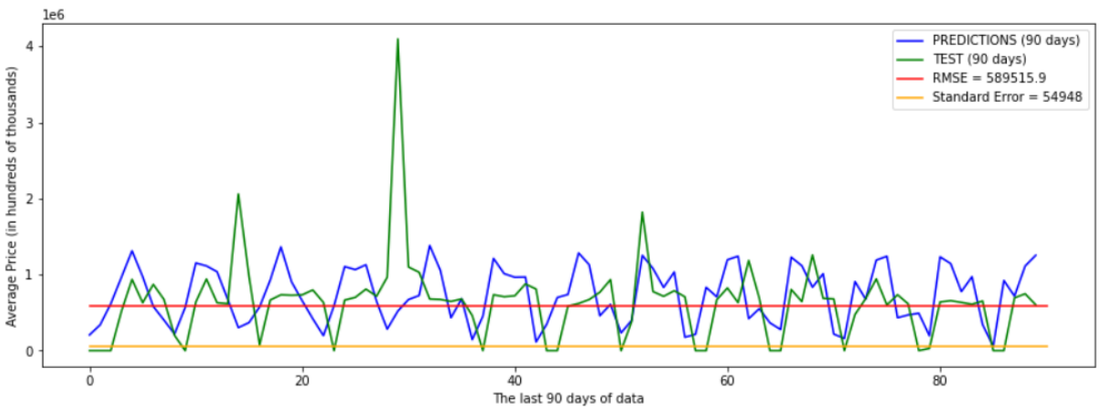


QUEENS ARMA forecast



Error Analysis

1. Length of Predictions : 90
2. Length of Test data : 90
3. RMSE : 584288.5171829152
4. Standard Error : 54947.9592801162

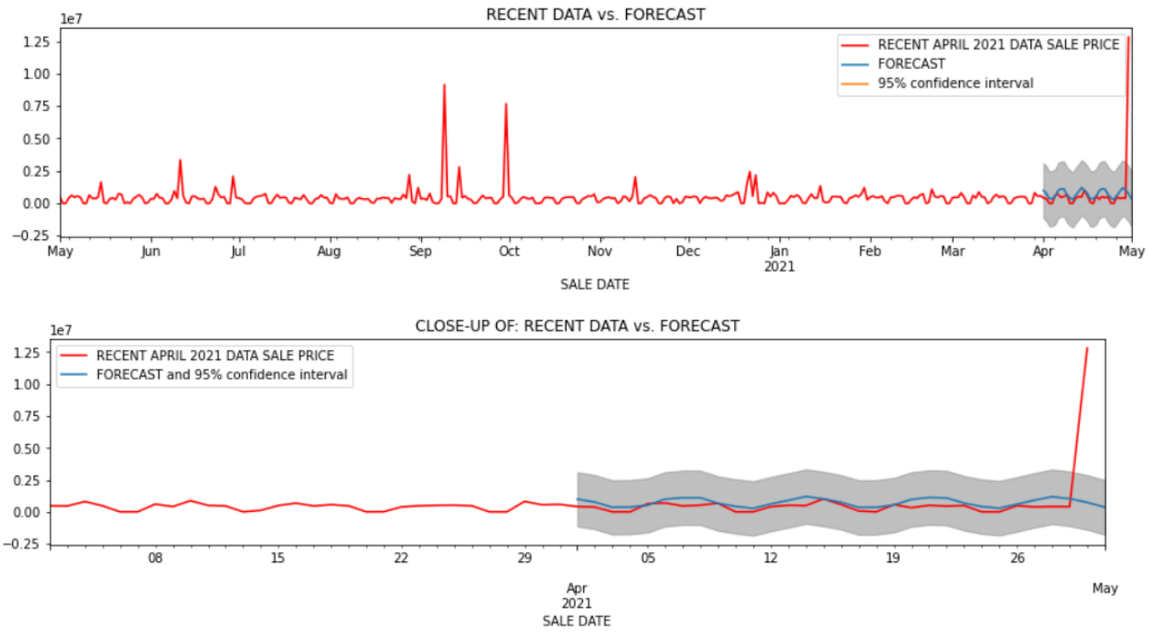


Observation:

RMSE is not too high or low. Lower values of RMSE indicate better fit. I believe this is a good range and model fits well.

- RMSE is 584288.5171829152
- Standard error is 54947.9592801162

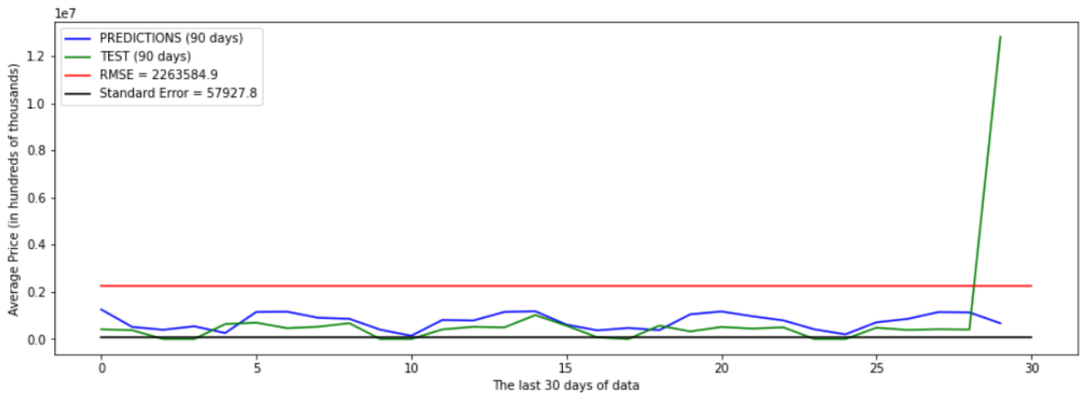
QUEENS Comparing predictions with fresh data from June 2021 dataset (4/1/2021 - 4/31/2021)



Observation

We see that the model has a decent fit with the test data of the last 30 days. However, the outlier will affect the error

1. Length of Predictions : 30
2. Length of Test data : 30
3. RMSE : 2263584.943411371
4. Standard Error : 54947.9592801162

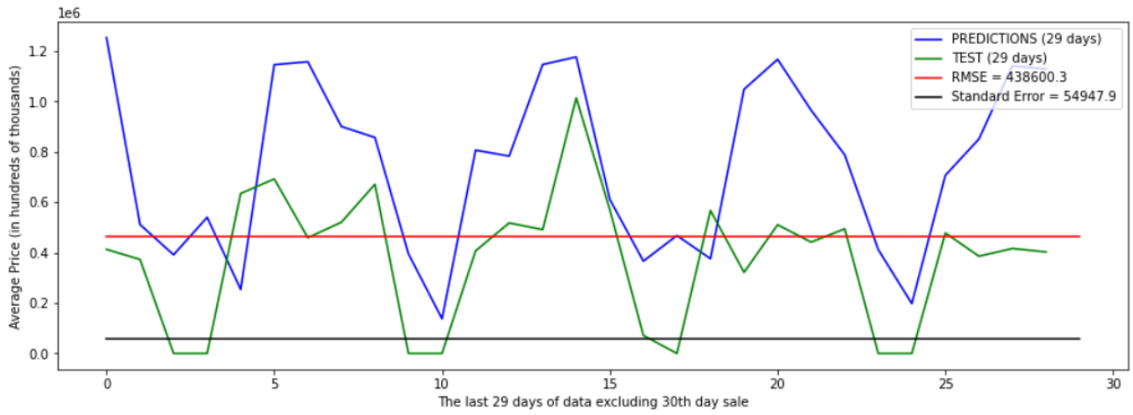


Observation:

- We see that RMSE is much higher due to a large sale that occured near the end of the month.
- Below I remove that outlier and re-check RMSE and Standard error

QUEENS Comparing predictions with fresh data - removed outlier

```
Length of Predictions :  29
Length of Test data :  29
RMSE :  463989.1499686349
Standard Error :  57927.81077050332

Text(0, 0.5, 'Average Price (in hundreds of thousands)')
```



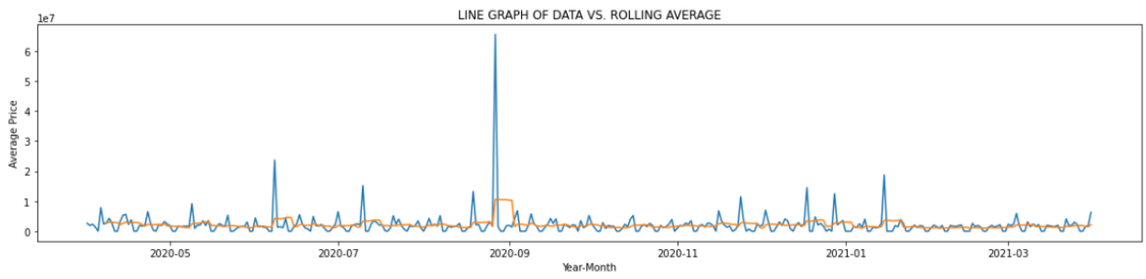
Observation

- RMSE is lower and so is the standard error than before. This can indicate a good fit. RMSE is not too high.
- With the outlier removed, RMSE is lower, indicates good fit

QUEENS Observations/Conclusions/Recommendations

```
1. The point of this analysis was to see if the borough was good to invest in
2. Based on the model:
    - We can enter to buy or exit to sell based on when the market will do well
3. The borough sales look predictable
4. There are unpredictable building sales which are very large amounts in the
millions to tens of millions
5. We can look at the top 10 building permit heavy locations further
```

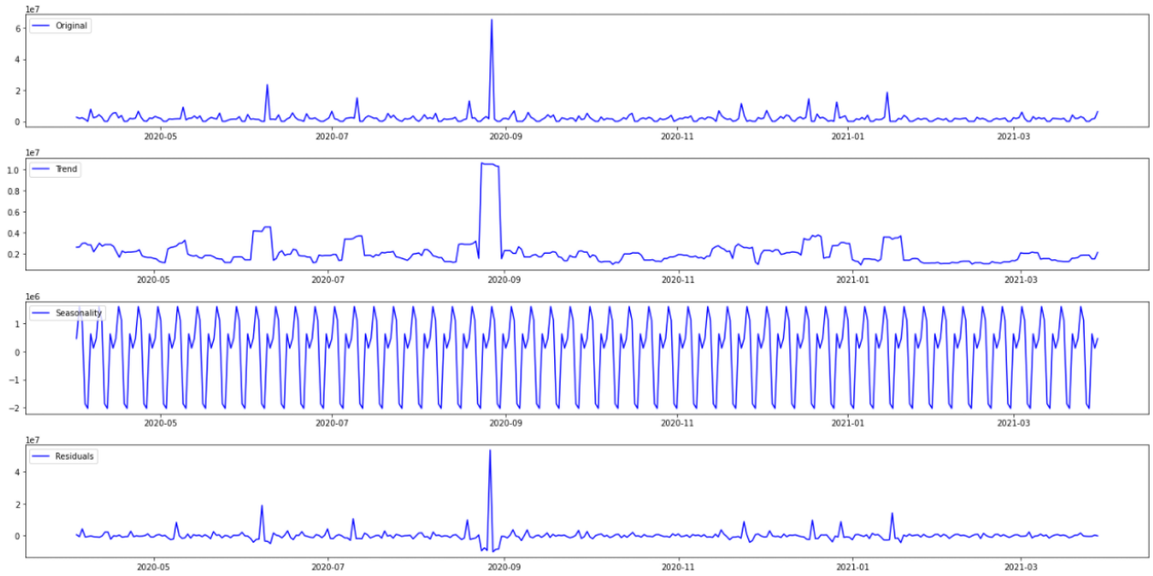
MANHATTAN Line Graph of Data vs 7-day rolling average



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
```

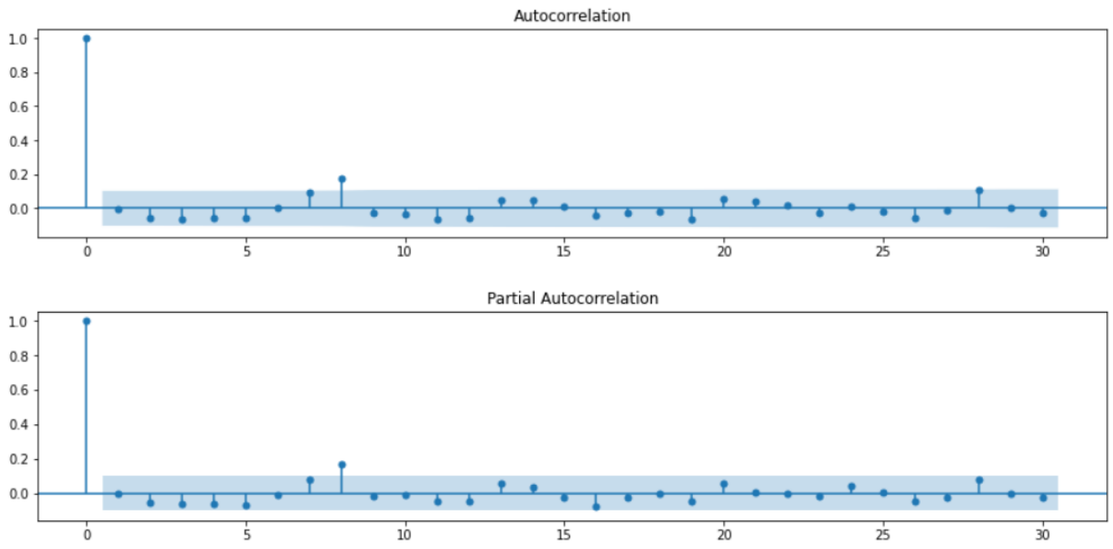
MANHATTAN Statsmodels decomposition



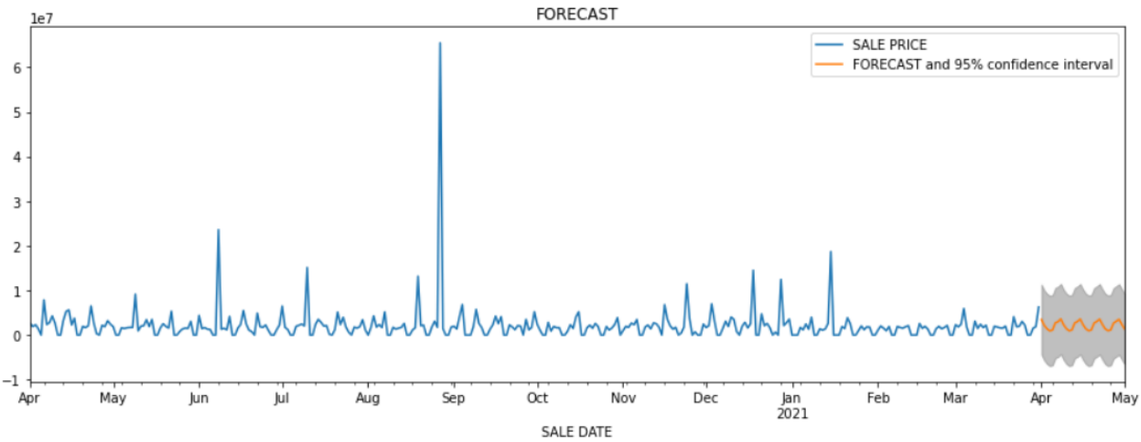
Observations:

- A large amount of sales happened between August 2020 and November 2020.
- Looks like there may be some seasonality every month

MANHATTAN Auto-Correlation and Partial Auto-Correlation

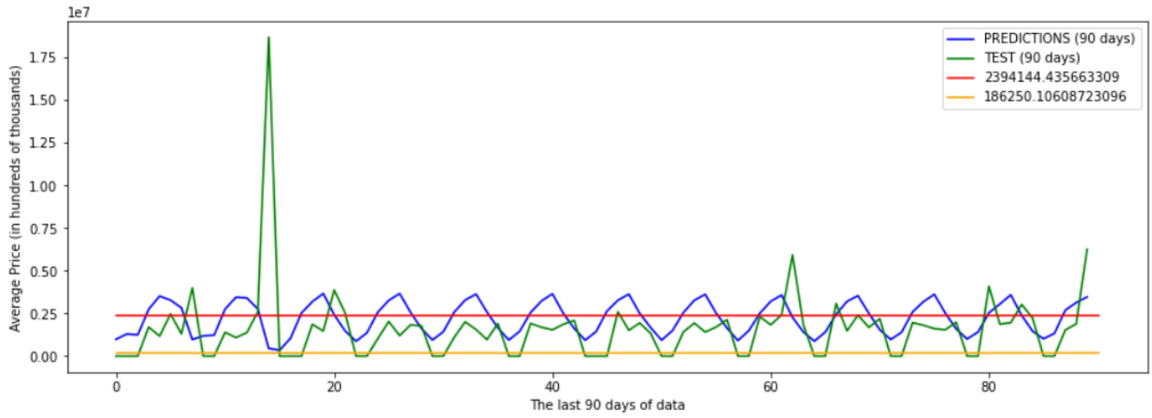


MANHATTAN ARMA forecast



Error Analysis

```
Length of Predictions :  90
Length of Test data :  90
RMSE :  2394144.435663309
Standard Error :  186250.10608723096

Text(0, 0.5, 'Average Price (in hundreds of thousands)')
```

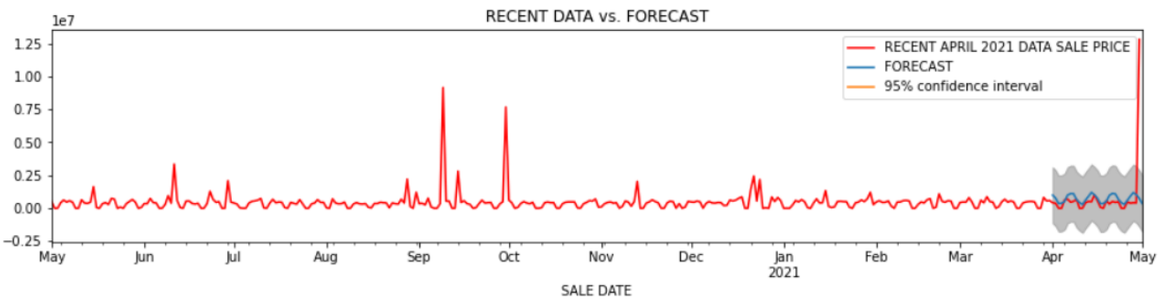


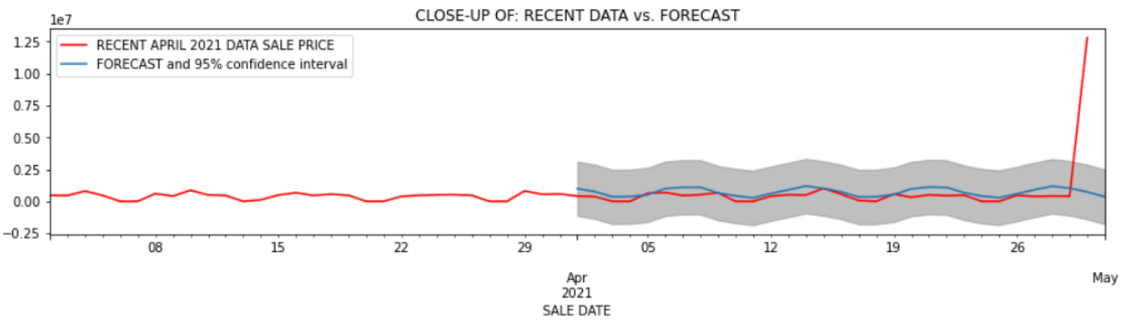
Observation:

RMSE is on the higher side of the data point values. This might indicate the higher price volatility in manhattan versus other boroughs.

- RMSE is 239414.4
- Standard error is 186250.1

MANHATTAN Comparing predictions with fresh data from June 2021 dataset (4/1/2021 - 4/31/2021)





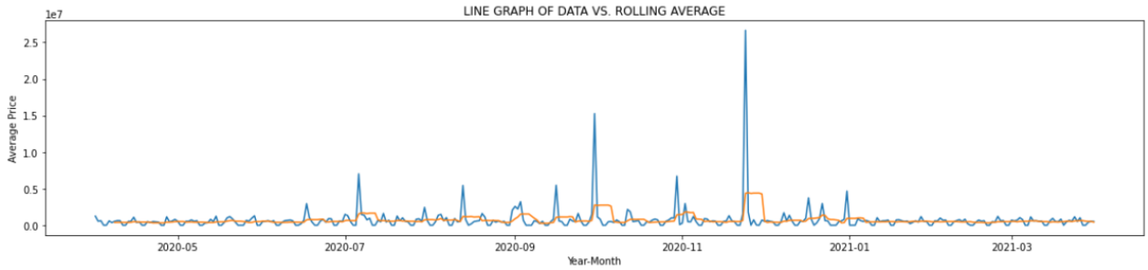
Observation

- We see that the model looks like it fits well versus the test data of 4/1/2021 until 4/31/2021
- There is a spike in April 2021, probably a building got sold for millions
    - This will affect our RMSE

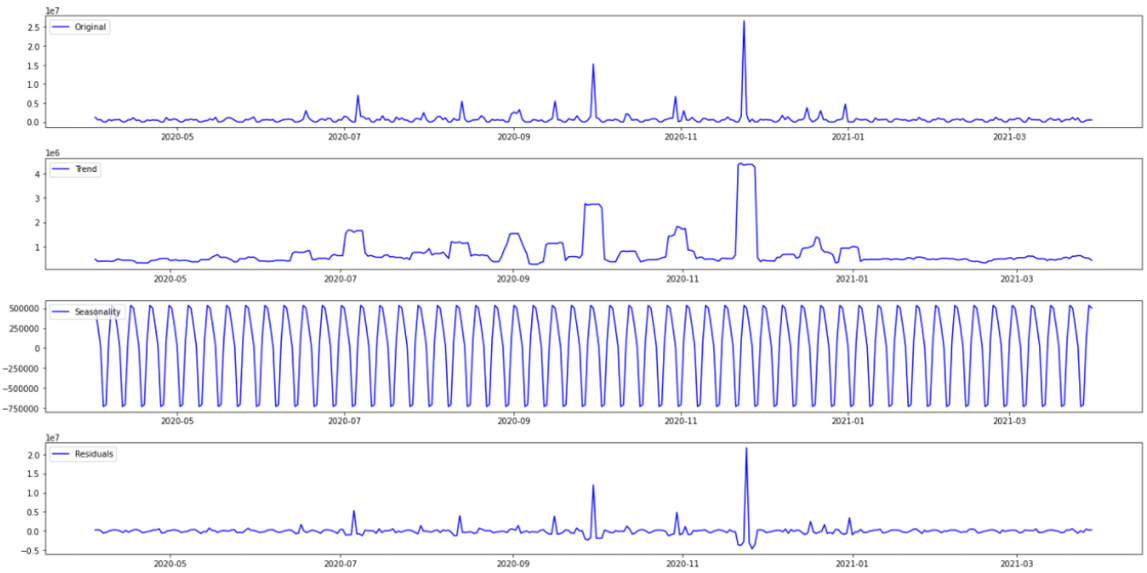
MANHATTAN Observations/Conclusions/Recommendations

```
1. The point of this analysis was to see if the borough was good to invest in
2. Based on the model:
    - We can enter to buy or exit to sell based on when the market will do well
3. The borough sales look predictable
    - Espescially for Manhattan, there is a predictable fluctuation
4. There are unpredictable building sales which are very large amounts in the
millions to tens of millions
5. We can look at the top 10 building permit heavy locations further
```

BRONX Line Graph of Data vs 7-day rolling average



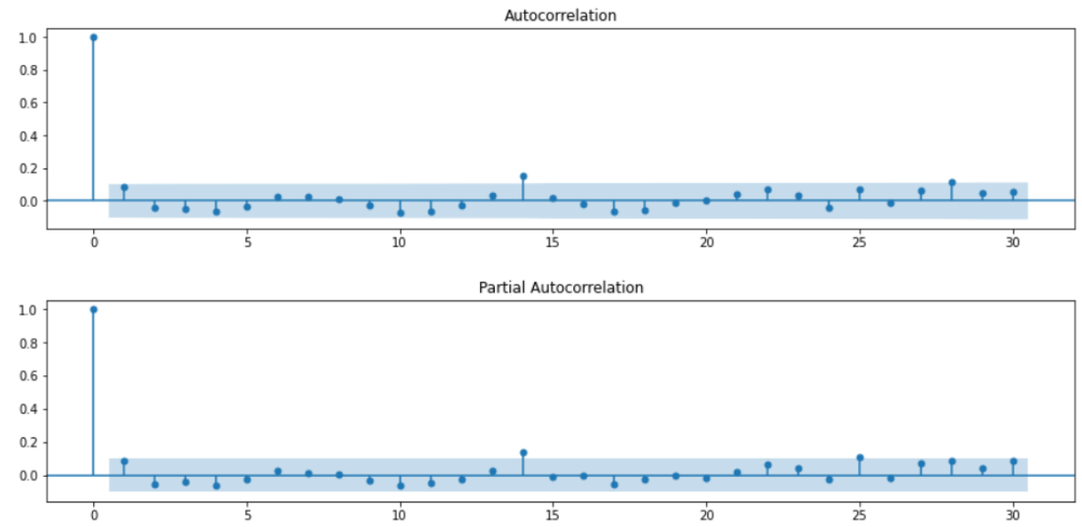
BRONX Statsmodels decomposition



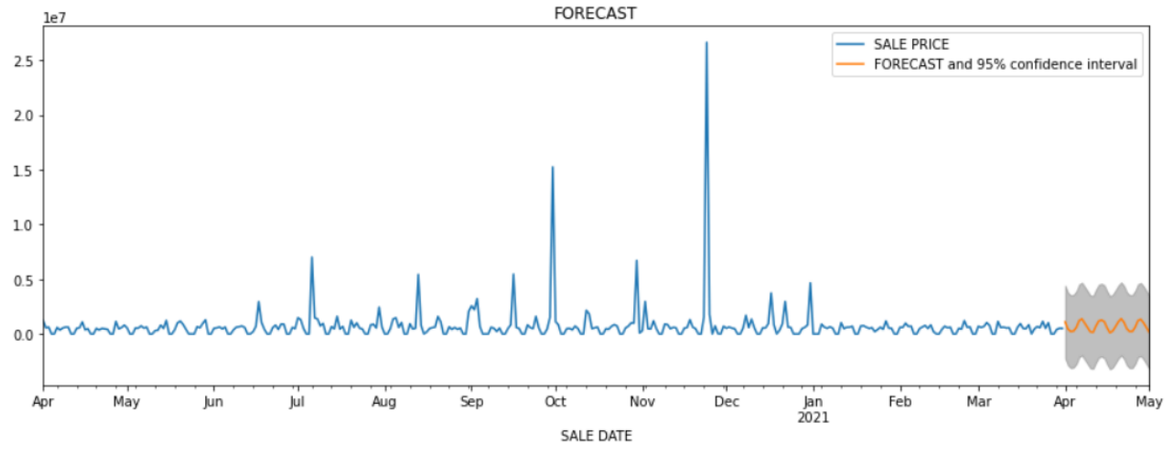
Observations:

- Looks like there may be some seasonality every month

BRONX Auto-Correlation and Partial Auto-Correlation



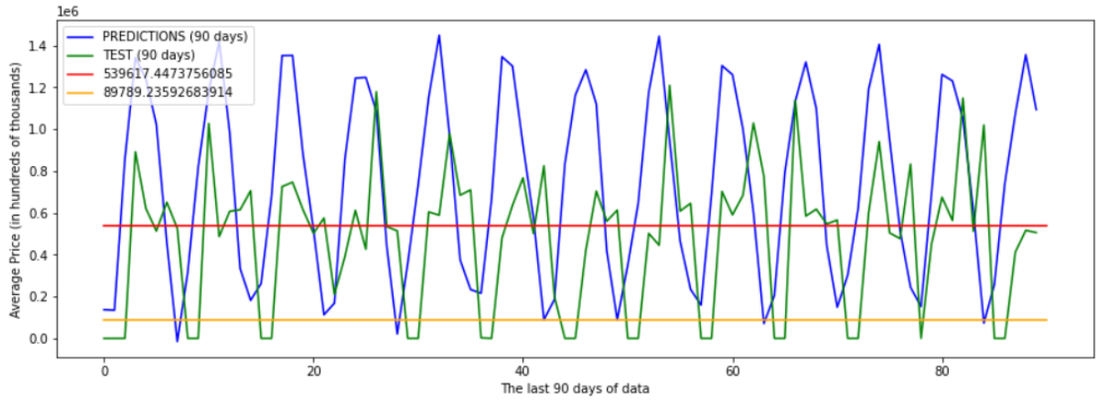
BRONX ARMA forecast



Error Analysis

Length of Predictions : 90
Length of Test data : 90
RMSE : 539617.4473756085
Standard Error : 89789.23592683914

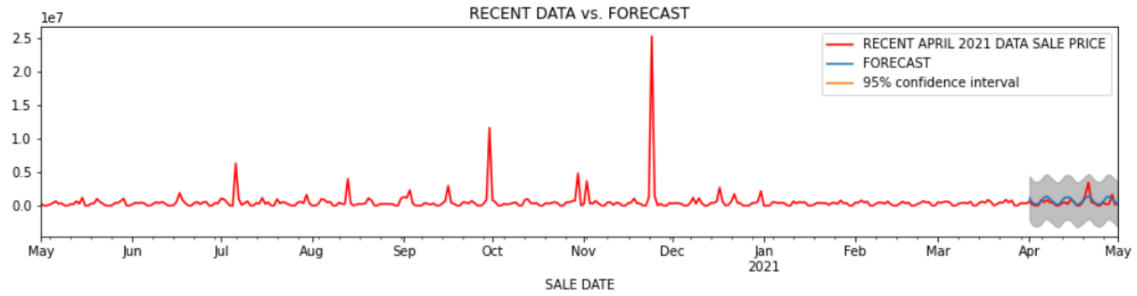
```
Text(0, 0.5, 'Average Price (in hundreds of thousands)')
```

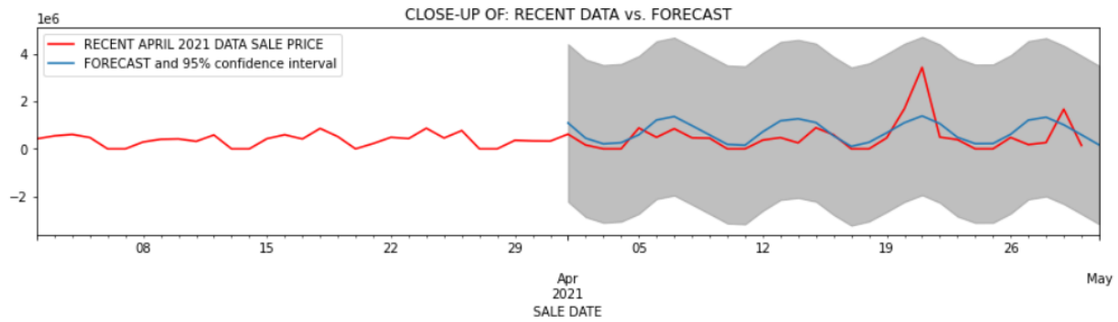


Observation:

RMSE is not too high and not too low compared to the data. Does not indicate a bad fit nor a good fit

BRONX Comparing predictions with fresh data from June 2021 dataset (4/1/2021 - 4/31/2021)



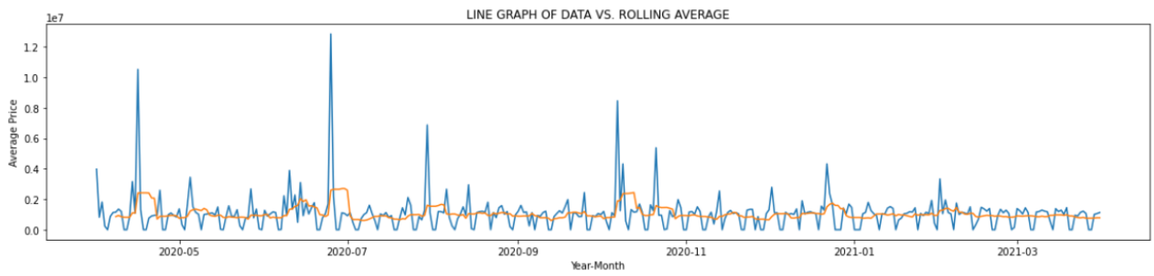


BRONX Observations/Conclusions/Recommendations

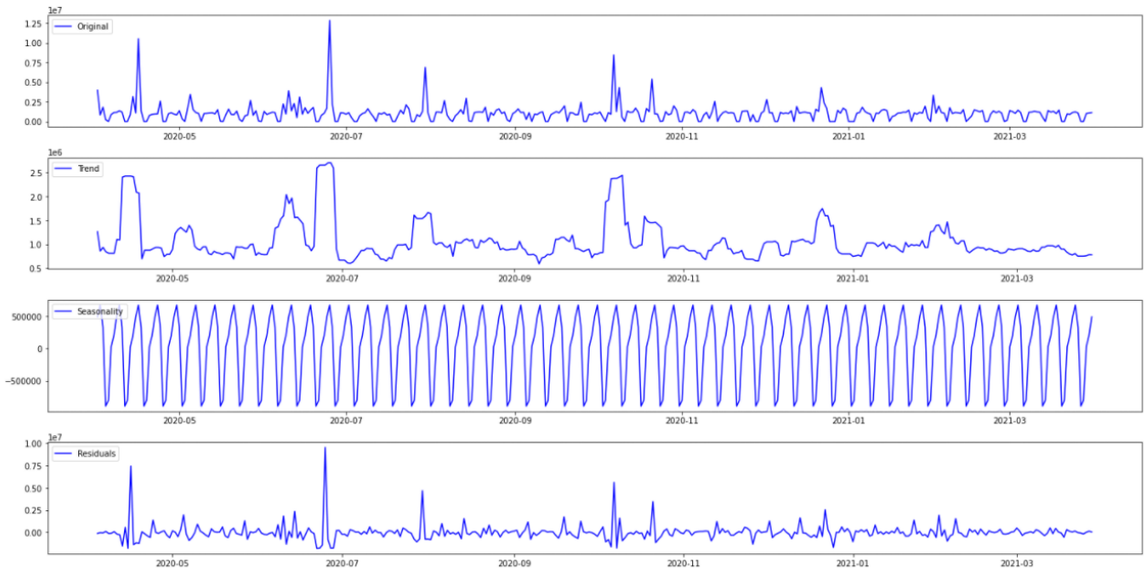
1. The point of this analysis was to see if the borough was good to invest in
2. Based on the model:

```
    - We can enter to buy or exit to sell based on when the market will do well
3. The borough sales look predictable
    - There is predicable fluctuation in Bronx
4. There are unpredictable building sales which are very large amounts in the
millions to tens of millions
5. We can look at the top 10 building permit heavy locations further
```

BROOKLYN Line Graph of Data vs 7-day rolling average



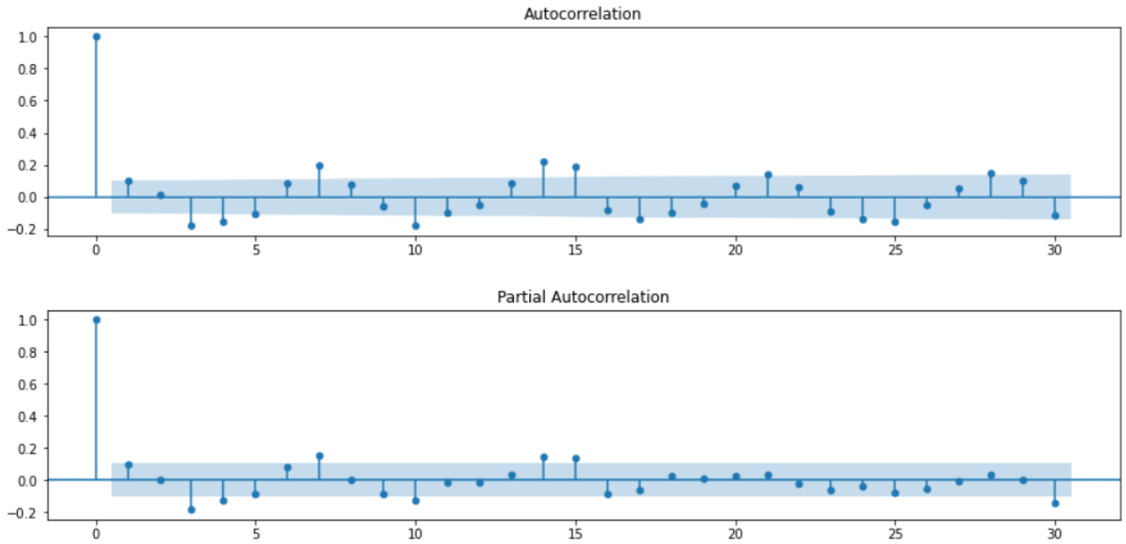
BROOKLYN Statsmodels decomposition



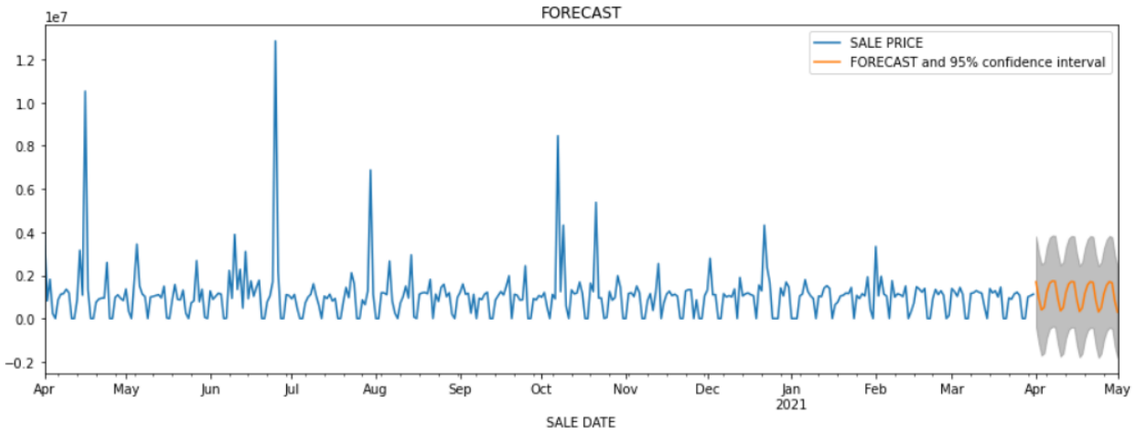
Observations:

- Looks like there may be some seasonality every month

BROOKLYN Auto-Correlation and Partial Auto-Correlation

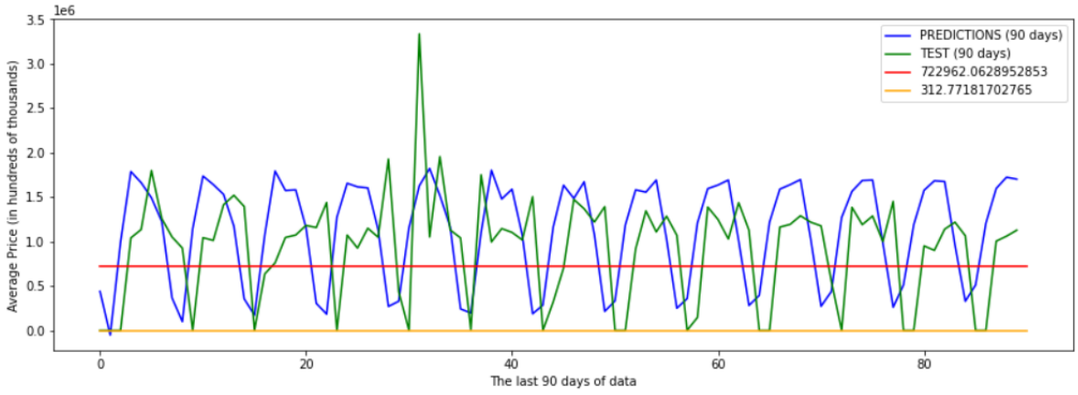


BROOKLYN ARMA forecast



Error Analysis

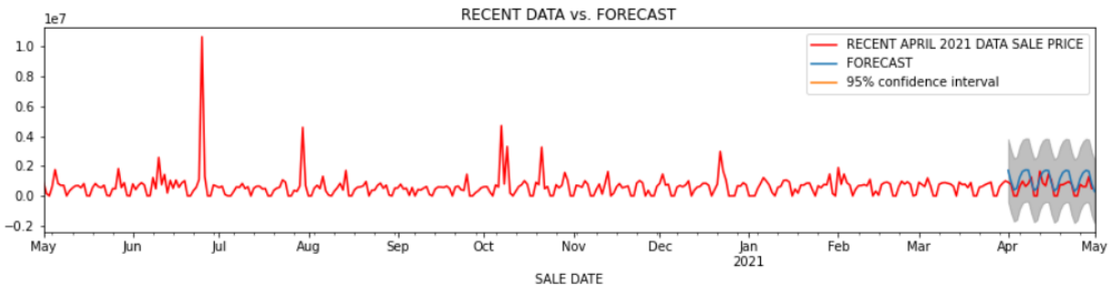
```
Length of Predictions :  90
Length of Test data :  90
RMSE :  722962.0628952853
Standard Error :  312.77181702765

Text(0, 0.5, 'Average Price (in hundreds of thousands)')
```

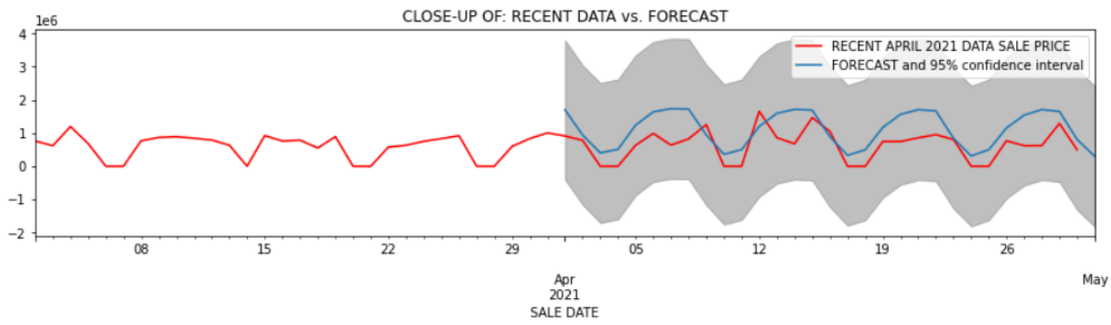


Observation:

- Here RMSE is lower than original model. We will stick with new model.

BROOKLYN Comparing predictions with fresh data from June 2021 dataset (4/1/2021 - 4/31/2021)

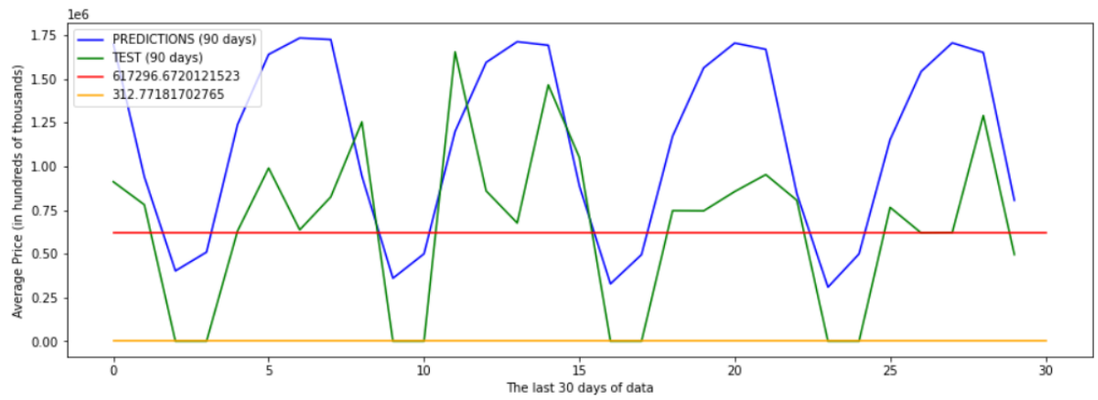




Observation

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

```
Length of Predictions :  30
Length of Test data :  30
RMSE :  617296.6720121523
Standard Error :  312.77181702765

Text(0, 0.5, 'Average Price (in hundreds of thousands)')
```



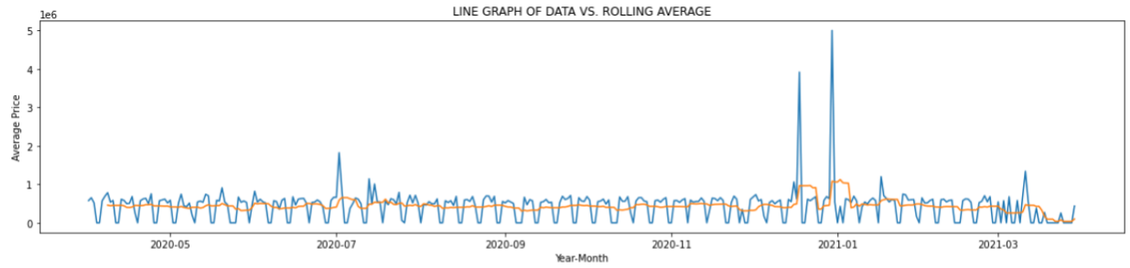
Observation

1. RMSE is lower here when comparing the new month data with predicted values

BROOKLYN Observations/Conclusions/Recommendations

1. The point of this analysis was to see if the borough was good to invest in
2. Based on the model:
    - We can enter to buy or exit to sell based on when the market will do well
3. The borough sales look predictable
    - There is predicable fluctuation in Brooklyn
4. We can look at the top 10 building permit heavy locations further

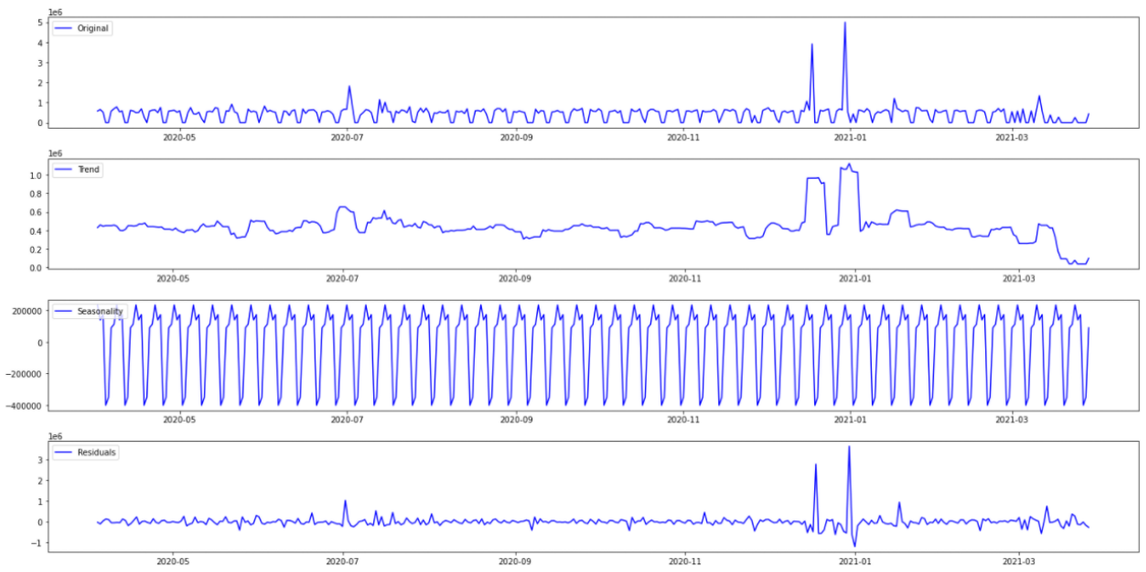
STATEN ISLAND Line Graph of Data vs 7-day rolling average



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

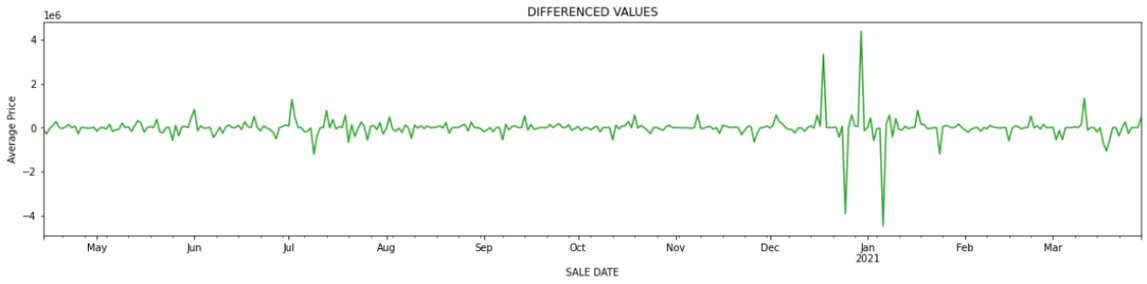
STATEN ISLAND Statsmodels decomposition

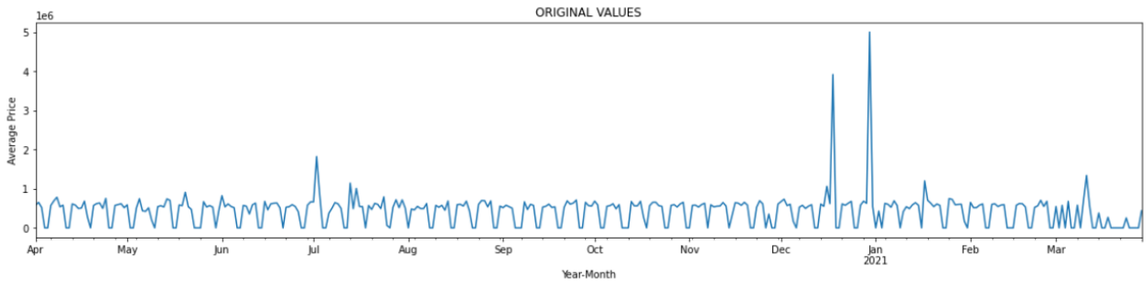


Observations:

- Looks like there may be some seasonality every month

STATEN ISLAND Inducing Stationarity






```
Test Statistic                  -5.182300
p-value                          0.000010
#Lags Used                      16.000000
Number of Observations Used    333.000000
Critical Value (1%)             -3.450141
Critical Value (5%)             -2.870259
Critical Value (10%)            -2.571415
dtype: float64
```

Results of ADF Test

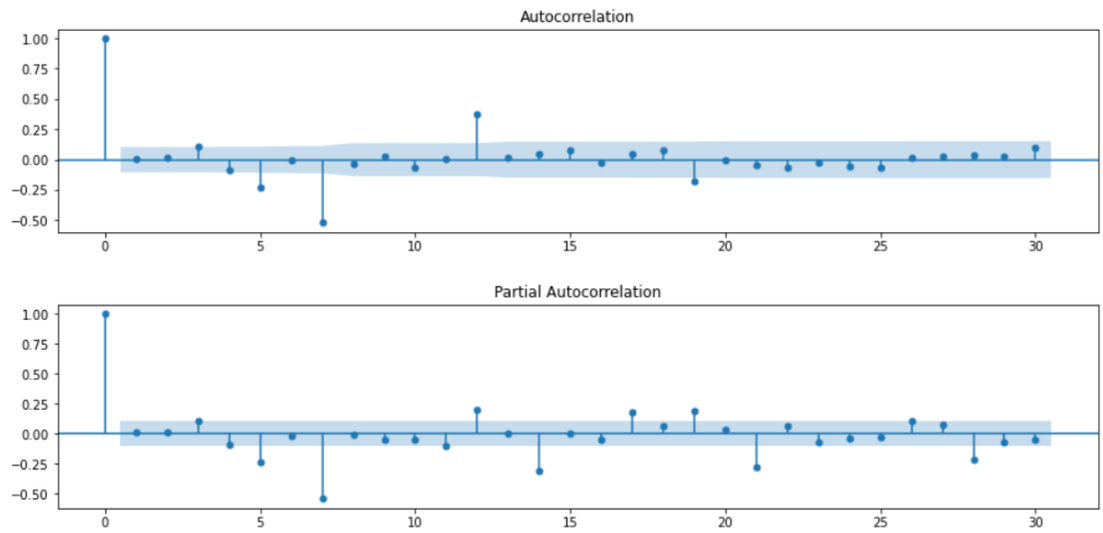
Test Statistic vs. Critical Values

- Initial test shows Test Statistic of **-5.182300**, this is greater than the critical values for 1% and 5%.
  - *We **REJECT** the null hypothesis! The data does not have a unit root and is stationary*

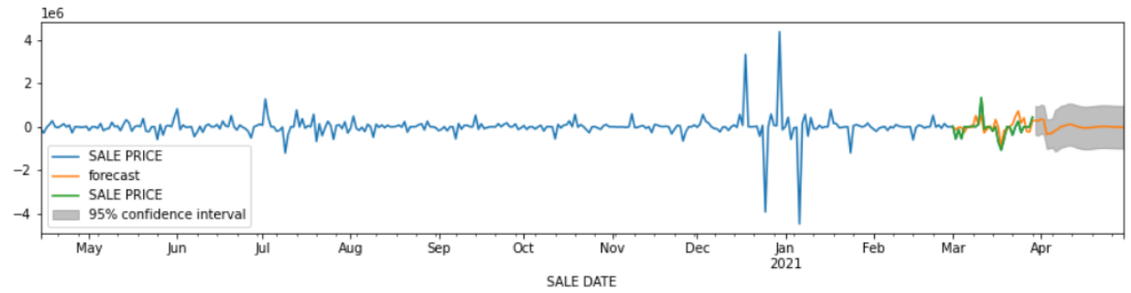
P-Value Analysis

- Our current p-value is **0.000010**
  - *This means: p-value <= 0.05:*
  - *We **REJECT** the null hypothesis! The data does not have a unit root and is stationary*

STATEN ISLAND Auto-Correlation and Partial Auto-Correlation



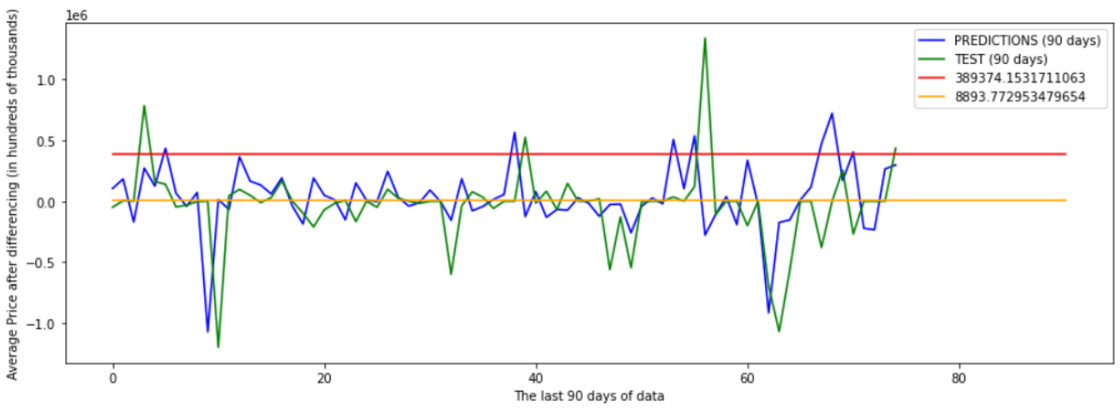
STATEN ISLAND ARMA forecast



Error Analysis

Length of Predictions : 75
Length of Test data : 75
RMSE : 389374.1531711063
Standard Error : 8893.772953479654

```
Text(0, 0.5, 'Average Price after differencing (in hundreds of thousands)')
```

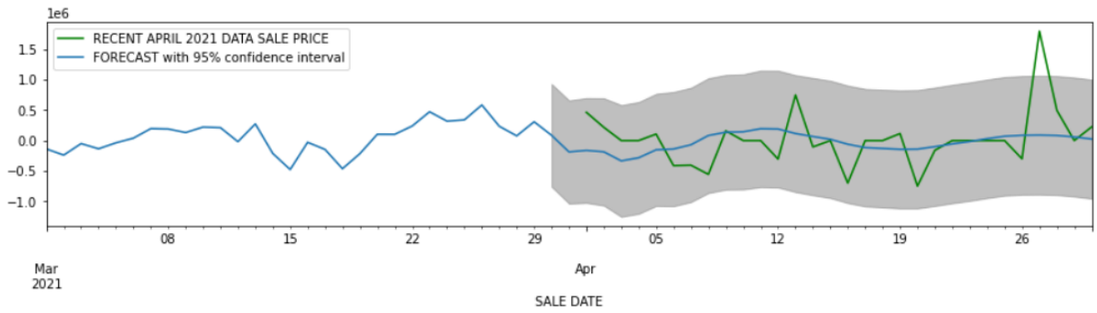


Observation:

RMSE is on the higher side of the data but this is after differencing

- RMSE is 389374.15
- Standard error is 8893.7

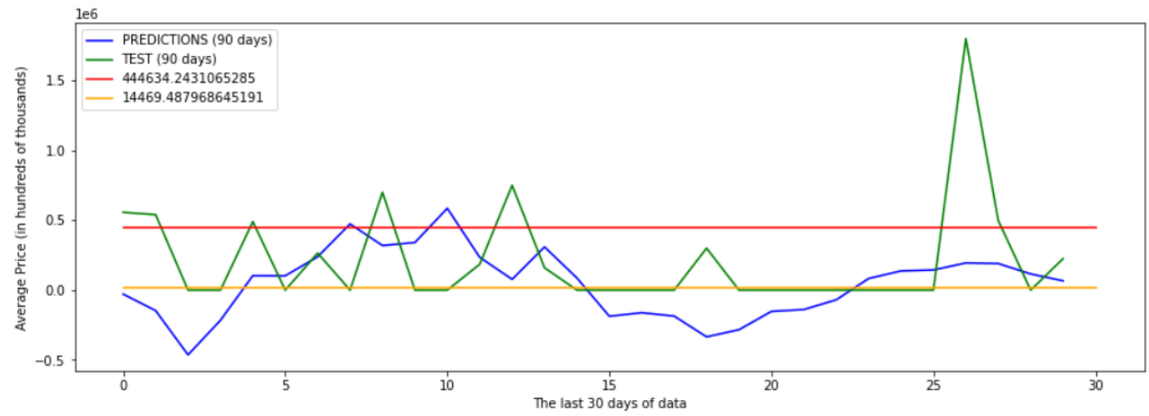
STATEN ISLAND Comparing predictions with fresh data from June 2021 dataset (4/1/2021 - 4/31/2021)



Observation

- The model does not look like it fits well

```
Length of Predictions :  30
Length of Test data :  30
RMSE :  444634.2431065285
Standard Error :  14469.487968645191

Text(0, 0.5, 'Average Price (in hundreds of thousands)')
```



Observation

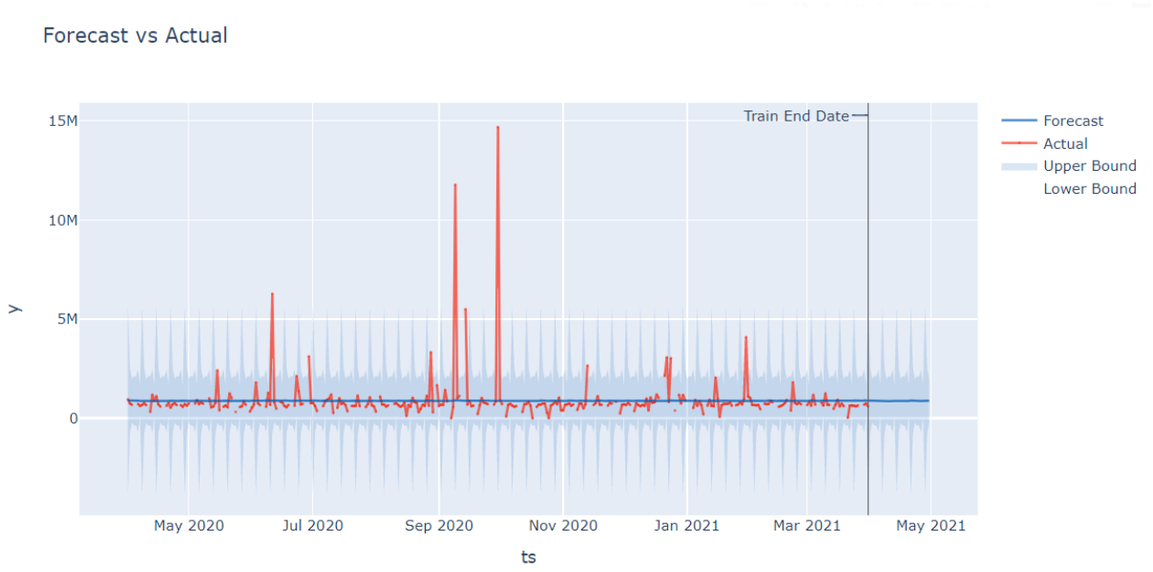
1. RMSE is alot higher here when comparing the new month data with predicted values
   - this is probably due to the large sales that occured during the last month

STATEN ISLAND Observations/Conclusions/Recommendations

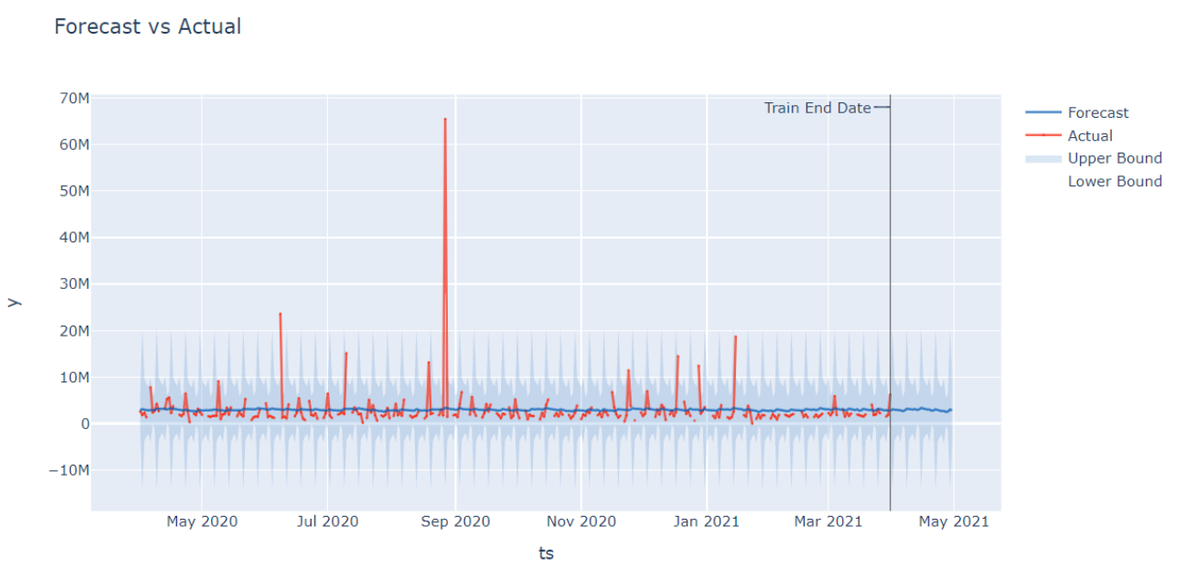
```
1. The point of this analysis was to see if the borough was good to invest in
2. Based on the model:
    - We can enter to buy or exit to sell based on when the market will do well
3. The borough sales look predictable
    - There is predicable fluctuation in Staten Island
4. We can look at the top 10 building permit heavy locations further
```

6. LinkedIn GreyKite forecasting of data - all NYC boroughs and observations

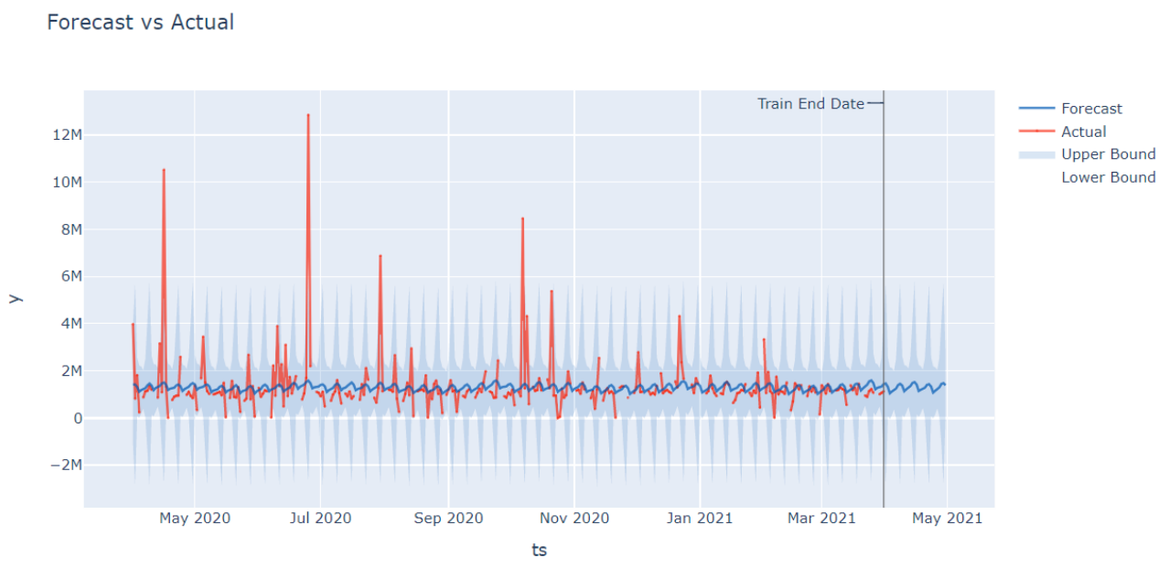
LinkedIn GreyKite - QUEENS



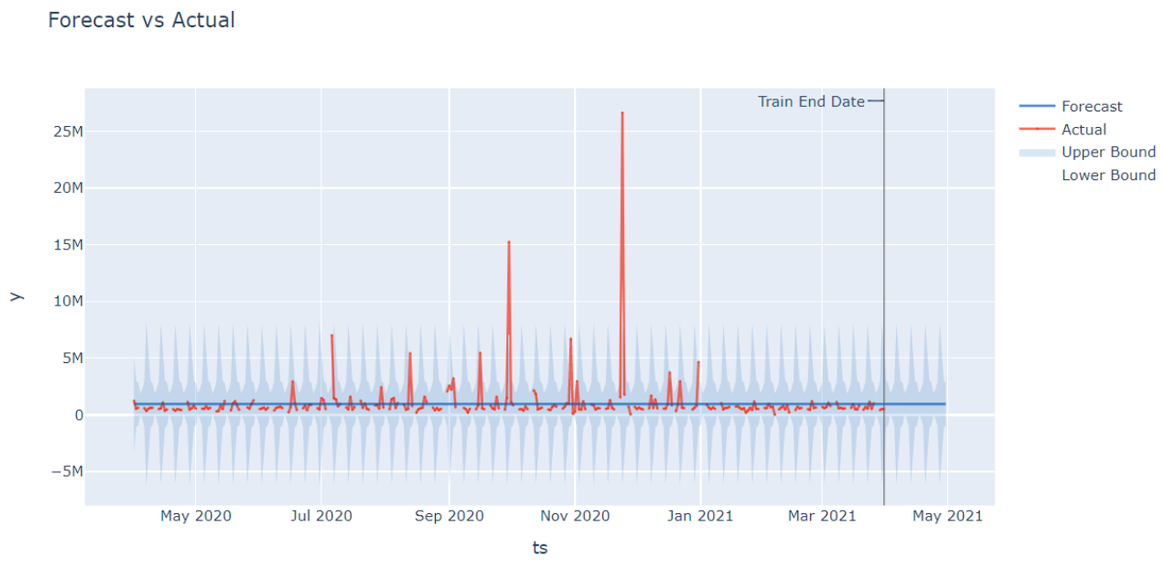
LinkedIn GreyKite - MANHATTAN



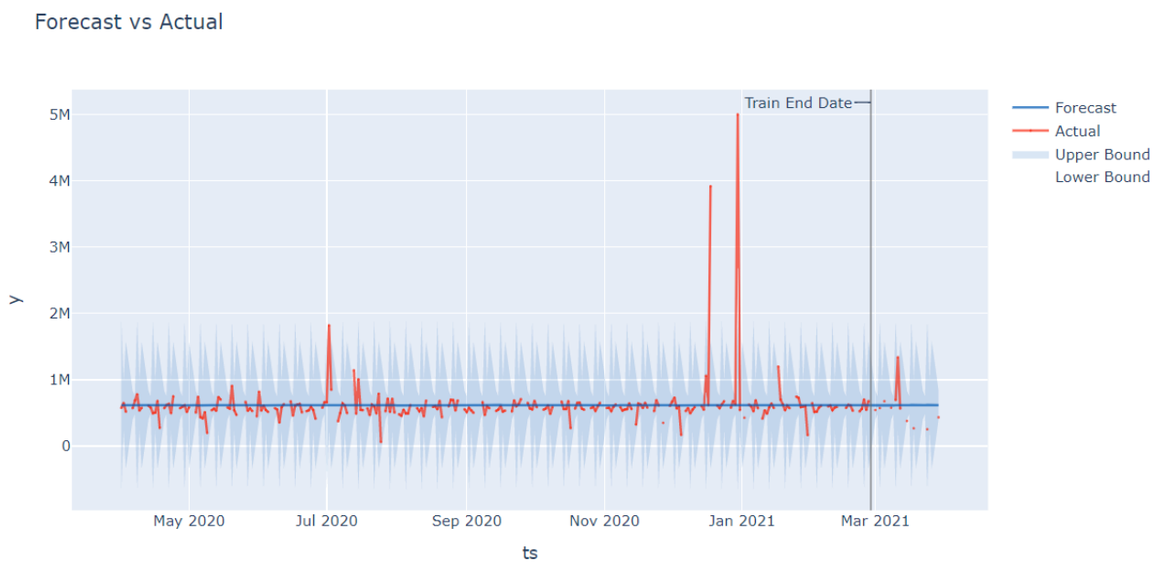
LinkedIn GreyKite - BROOKLYN



LinkedIn GreyKite - BRONX

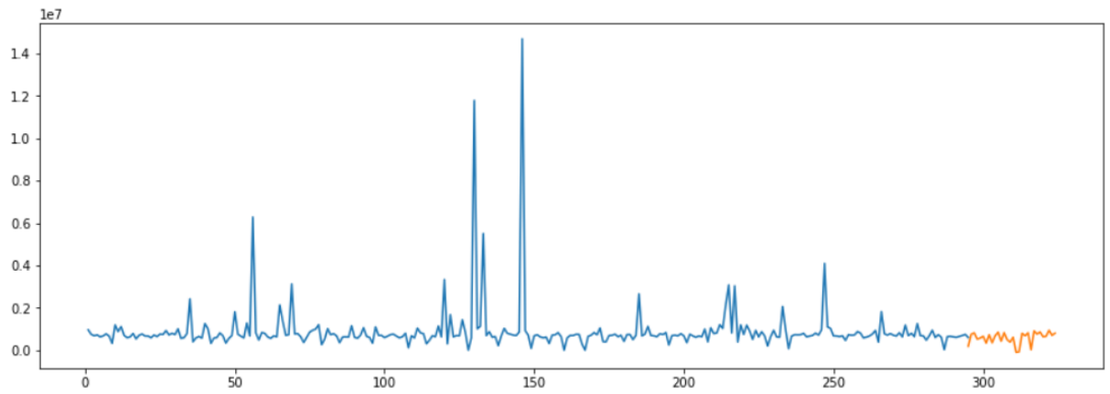


LinkedIn GreyKite - STATEN ISLAND



7. LSTM Forecasting data - all NYC boroughs and observations

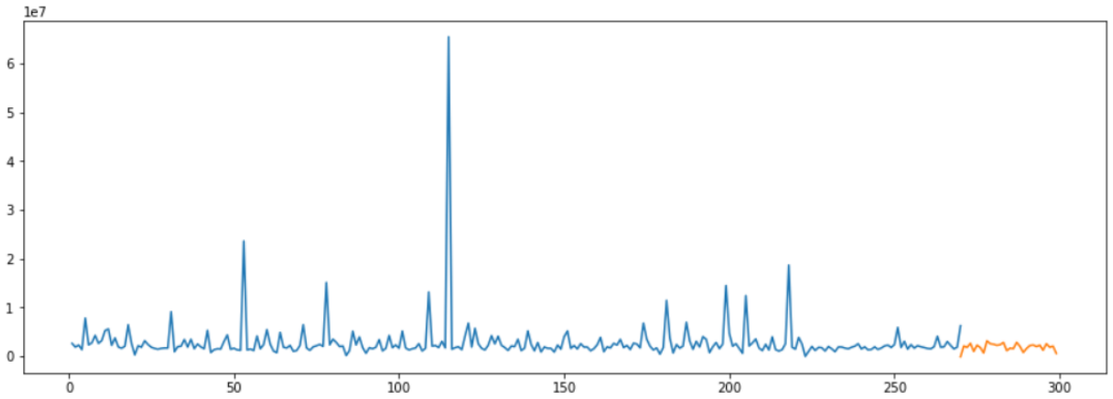
LSTM QUEENS FORECAST



Observation

Queens prices per model show stable with some dips

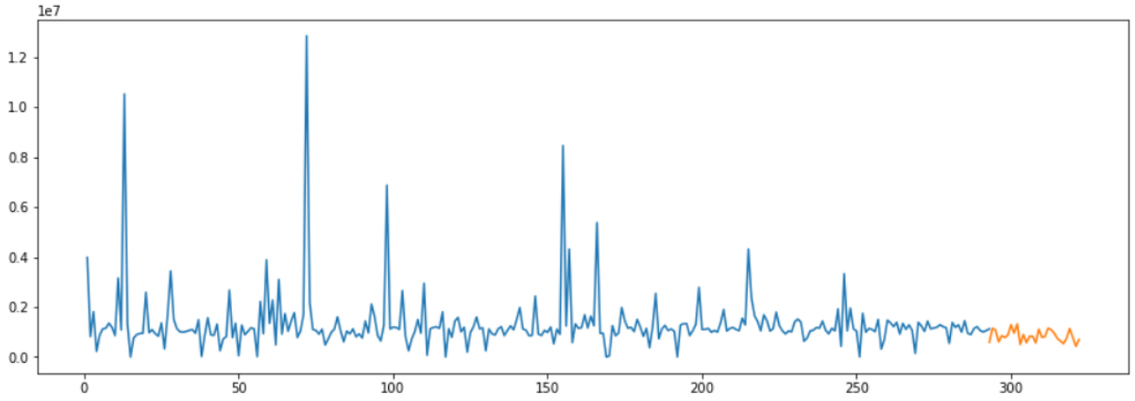
LSTM MANHATTAN FORECAST



Observations:

1. Predictions seem lower than the origional data for Manhattan

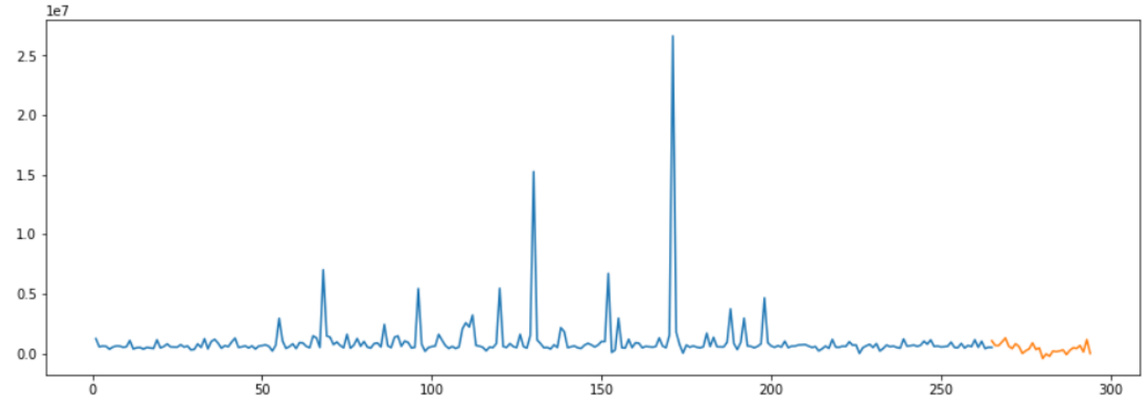
LSTM BROOKLYN FORECAST



Observation:

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

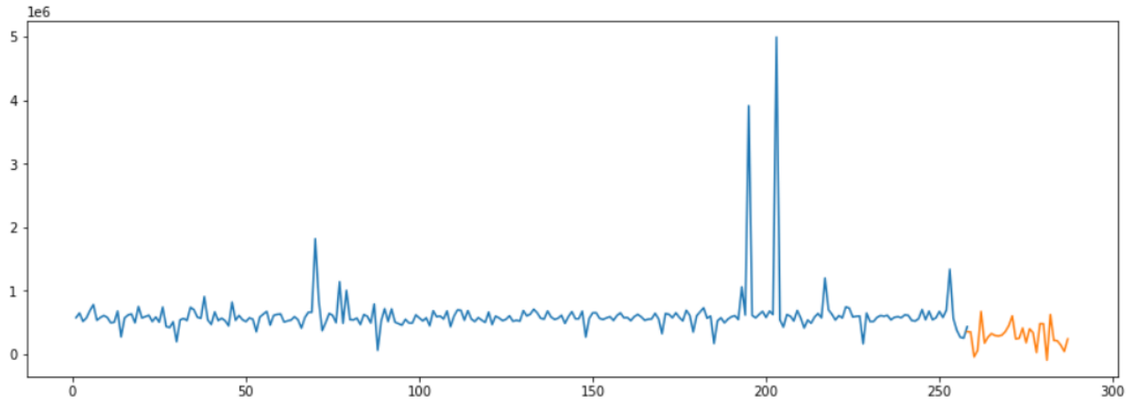
LSTM BRONX FORECAST



Observation

Bronx prices per model show downward prices with some dips

LSTM STATEN ISLAND 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

8. Permit Data Analysis and decisions

Top Names Analysis

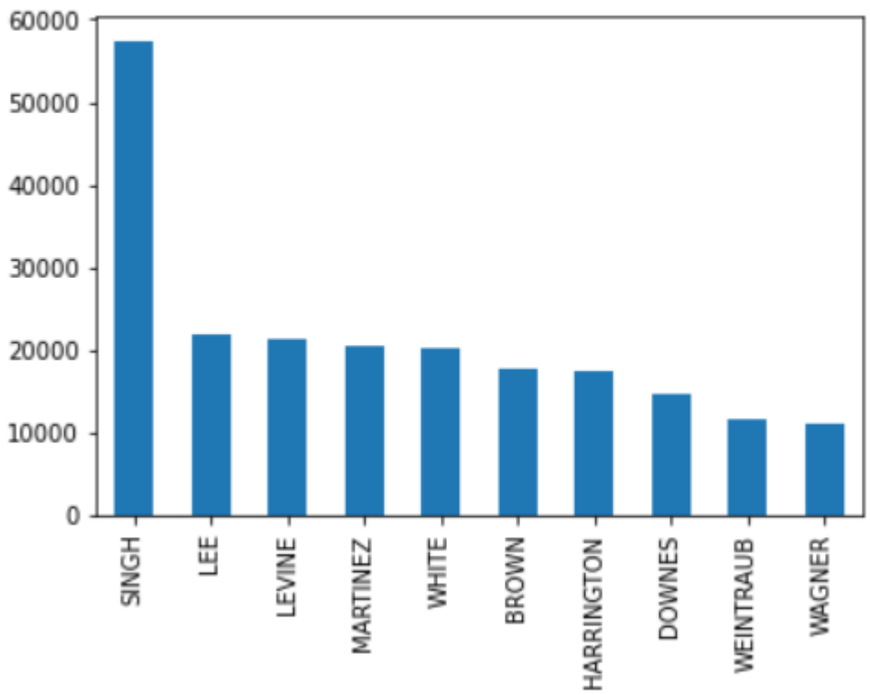
```
data.dropna(inplace=True)
data.sort_values(by=['Issuance Date'])
data
```

	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

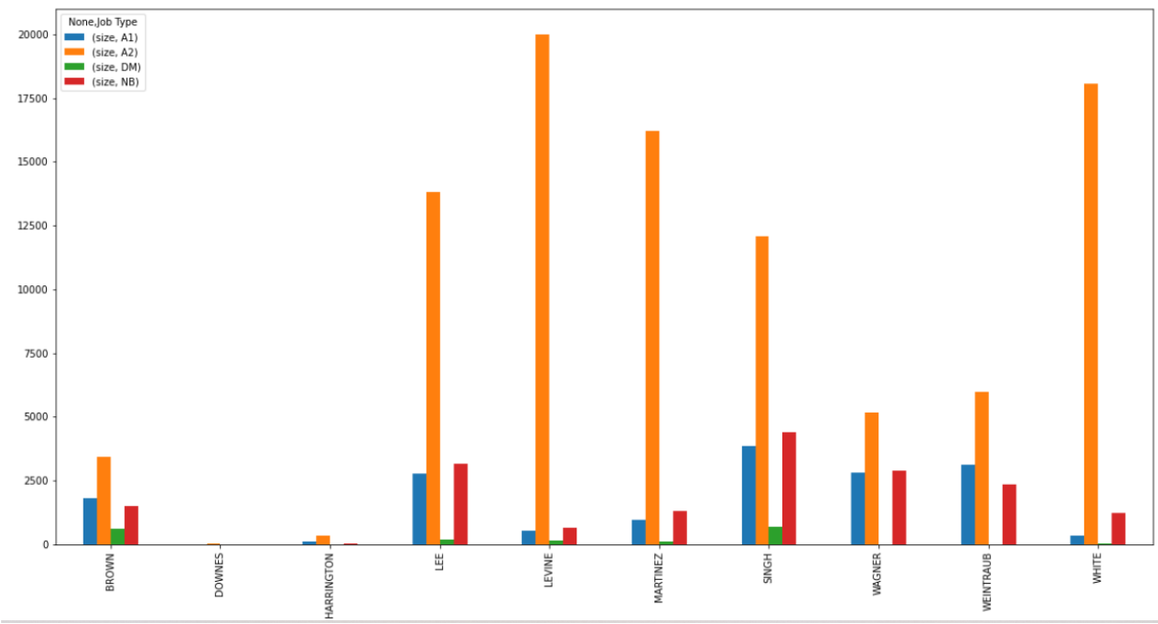
```
names = data['Permittee s Last Name'].value_counts()
names[:10].plot(kind='bar')
```

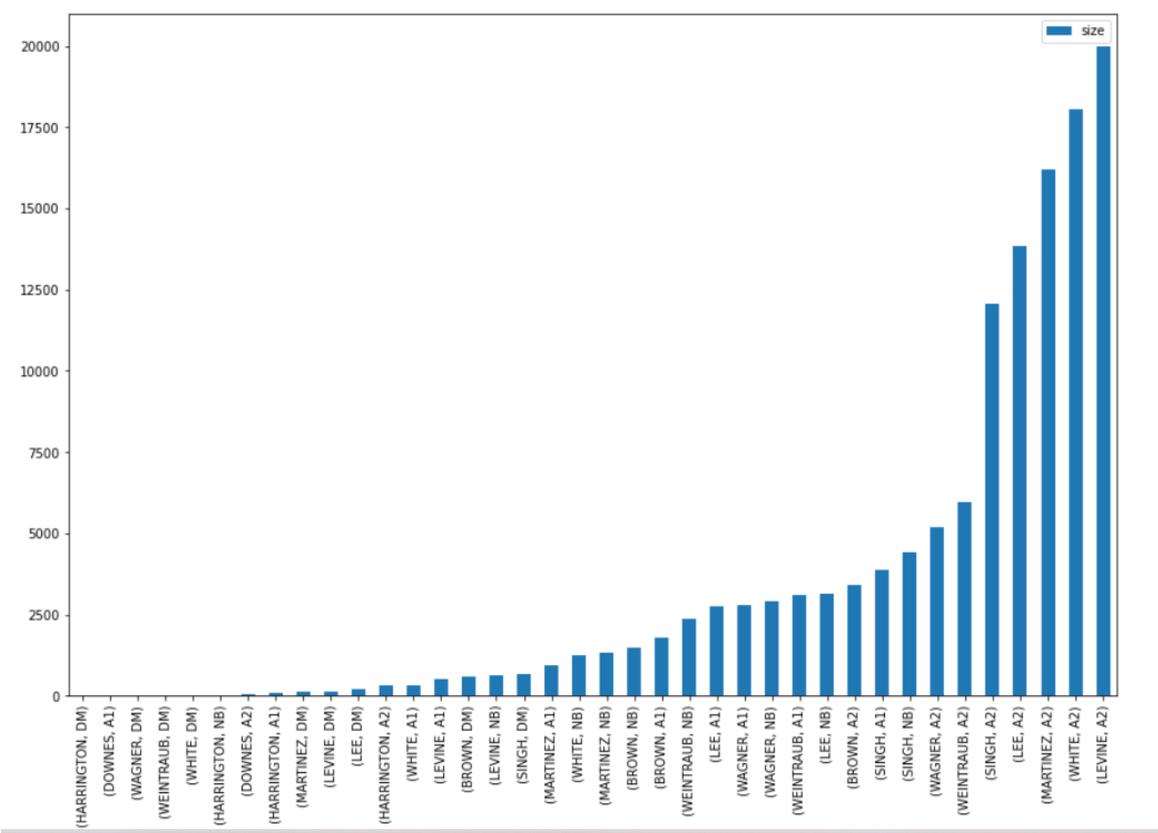
```
<AxesSubplot:>
```



```
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
```





Top 10 Zip Codes Per Borough

QUEENS

```
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]
```

MANHATTAN

```
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]
```

BRONX

```
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]
```

BROOKLYN

```
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]
```

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]
```

9. 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

10. 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

Releases

No releases published

Packages

No packages published

Languages

- **Jupyter Notebook** 100.0%