Assignment 4

Predictive Analysis - NYC Real Estate

Moiz Deshmukh

BU ID: U63698143

Metropolitan College, Boston University

AD517: Applied Business Analytics (Summer 2)

Professor Prasad Kothapalli

July 30th, 2024

Table of Contents

Executive Summary 3
Time Series Analysis 4
Regression Analysis 5
Multiple Regression Analysis 6
Residual Regression Analysis 6
References 8
Appendix 9
Figure 1-A 9
Figure 1-B 9
Figure 2 10
Figure 3 10
Figure 4 11
Figure 5 11
Figure 6 12
Figure 7 12
Figure 8 13
Figure 9 13
Figure 10 14
Figure 11 14

Executive Summary

This report provides a predictive analysis of property prices in the Bayside neighborhood, Queens Borough, New York City. A comprehensive time series analysis in R Studio evaluated total residential sales since 2009 in Bayside. The analysis included properties from 2009 onward, excluding those sold for less than $75,000 and those smaller than 350 square feet. The 'MAM' smoothing model, selected based on Akaike Information Criterion (AIC) scores, fit the historical data well. Despite no clear seasonal trend, the forecast predicts a positive upward trend in sales prices over the next eight quarters, with a wide 95% confidence interval indicating variability.

The regression analysis on the same dataset showed time and quarter as significant predictors of property sale prices, with values estimated to increase over time. However, these factors account for only 59% of the variation in sale prices, with time alone explaining 48%. A multiple regression analysis found building type and gross square feet to be the most useful predictors, while sale date and year built had a smaller impact. The analysis highlighted the need to address multicollinearity by removing variables with high p-values and low t-values.

Lastly, a residual regression analysis identified the best and worst deals for buyers. The most overpriced property in Bayside was 36-21 213TH STREET, sold in 2013 for $6 million despite a market value of $2.6 million. Similarly, 39-30 214TH PLACE was sold for $4.8 million, while its market value was $2.8 million. Conversely, properties like 209-20 18TH AVE, sold in 2010 for $2.1 million despite a market value of $2.8 million, represented good deals. These findings highlight discrepancies between sale prices and market values, underscoring both overpriced and underpriced properties in Bayside.

Time Series Analysis

A time series analysis was conducted in R Studio to examine the total residential sales for each year since 2009, focusing on the Bayside neighborhood in the Queens Borough of New York City. This analysis utilized residential data filtered for properties in Bayside from 2009 onward, excluding earlier data as 2008 marked the end of a real estate price bubble caused by the Subprime Mortgage Crisis. To ensure accuracy, properties sold for less than $75,000 and those smaller than 350 square feet were excluded from the dataset.

For the time series prediction, as seen (in Figure 1-A), a smoothing model was selected based on the characteristics observed in the historical data. The selection process involved comparing the Akaike Information Criterion (AIC) scores for various models, and the model with the lowest AIC score was chosen for further analysis (Figure 2). Also, on running the ZZZ model, the model suggested to use the ‘MAM’ model, which was determined to best fit the data (See figure 3).

The 'MAM' model's parameters (Figure 4) include alpha, beta, and gamma values, all set at 0.0001, as seen in the summary. These parameters influence the smoothing process for level, trend, and seasonal components, respectively. The initial states and error measures, such as the Mean Error (ME), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE), indicate the model's accuracy and reliability. The AIC score of 1943.042, along with other metrics like AICc and BIC, supports the model's robustness in fitting the data.

The analysis graph, derived from the 'MAM' model, displays peaks and valleys within the data. However, these fluctuations do not occur at regular intervals or during the same quarter year-over-year, confirming the absence of a seasonal trend. The forecast generated by the 'MAM' model, as shown in the chart, predicts a positive upward trend in sales prices over the next eight quarters. Nevertheless, the forecast includes a wide 95% confidence interval, highlighting the potential range in which future sale prices could fall.

Due to the significant deviations in sale prices (Figure 1-B) and the lack of a clear seasonal trend, the forecast's wide confidence interval indicates a high level of uncertainty. While the general trend suggests that prices will continue to rise, this variability implies that a property purchased now may not guarantee a profit or loss within the next two years. Consequently, this high level of risk must be considered by investors aiming to buy properties for short-term resale. The 'MAM' model's forecast shows the importance of cautious investment strategies, taking into account the potential for significant price fluctuations in the Bayside neighborhood's real estate market.

Regression Analysis

The regression analysis used the same filtering criteria as the time series analysis and focused on the Bayside neighborhood. In this analysis, time and quarter were used as predictors for sales price. The results as seen in figure 5, show that both time and quarter together are statistically significant predictors of property sale prices. The analysis suggests that property values are estimated to increase by approximately $1 million over time. However, the adjusted R-squared values for both the combined and simple regression models indicate that time and quarter together account for only 59% of the variation in sale prices.

Moreover, the p-values for sales occurring in all four quarters exceed 0.05, indicating no statistically significant seasonal effects in the data. Given this lack of significance, an additional regression analysis was performed using only time as a predictor of sales price (Figure 6). The findings show that time alone explains 48% of the variation in sale prices, implying that including the quarter as a predictor adds an extra 11% to the explanation of historical price changes. This highlights the significance of time as a key factor in predicting property values in Bayside, while seasonal factors appear to have a minimal impact.

Multiple Regression Analysis

A multiple regression analysis was conducted in R Studio to determine whether factors such as sale date, year built, building type, gross square feet, and number of units were statistically significant predictors of sale price. As indicated in the figure 7, these predictors together explain 43% of the variation in sale price. Certain building types significantly predict sale price, alongside sale date, year built, and gross square feet. The findings suggest that the type of building being sold is a more crucial indicator for determining sale price than the sale date or year built, although these factors do have a statistically significant, but smaller, impact on sale price. Hence, we can say that the building type and the gross square feet are the most useful predictors of amount of sale and the sale date or year-built count as the least predictors.

Property types such as Four Families Apartment, Large Suburban Residence and Two Stories Detached SM are statistically significant and have a substantial positive impact on sale price. The Year-Built and Gross Square Footage have a positive impact on the price as well. However, the number of residential units have a negative impact, which means that as the number of units increases, the actual sale price decreases. This could be possibly due to the investors competing with each other. Using twenty-four variables as predictors of sale price carries the risk of redundancy and multicollinearity among the independent variables. Therefore, removing variables with high p-values and low t-values is essential in reducing redundant variables and ensuring a more robust analysis.

Residual Regression Analysis

The final analysis aimed to identify the properties that represented the best and worst deals for buyers. This determination was made by examining the residual values, which indicate the difference between the actual sale price and the predicted value at a given time.

In Figure 8, it is evident that the most overpriced property in our Bayside neighborhood was 36-21 213TH STREET, sold in 2013 for $6 million, despite its actual market value being $2.6 million. Similarly, the second most overpriced property, also illustrated in Figure 8, was sold in 2015 for $4 million, although its real market value was $3.3 million. The third most overpriced property, 39-30 214TH PLACE in Bayside, was sold for $4.8 million, while its market value at that time was $2.8 million, as seen in Figure 9.

When examining properties sold at the best deals, we focus on those with the largest negative residual values, as depicted in Figure 11. The first notable property is 209-20 18TH AVE, sold in 2010 for $2.1 million, although its market value was $2.8 million. The second is 38-12 213TH STREET, sold in 2017 for $262,000, while its market worth was $2.5 million at that time, representing an exceptionally good deal.

Overall, these figures highlight significant discrepancies between sale prices and market values, emphasizing both overpriced and underpriced properties in the Bayside neighborhood.

References

38-12 213th St #1, Bayside, NY 11361 | Trulia. (n.d.). Trulia Real Estate Search. https://www.trulia.com/home/38-12-213th-st-1-bayside-ny-11361-2070407010

Redfin. (n.d.). *18-40 211th St Unit 2J, Bayside, NY 11360 - 1 bed/1 bath*. Redfin. https://www.redfin.com/NY/Bayside/1840-211th-St-11360/unit-2J/home/56720905

Redfin. (n.d.-c). *36-21 215th Pl, Bayside, NY 11361 - 6 beds/3 baths*. Redfin. https://www.redfin.com/NY/Bayside/3621-215th-Pl-11361/home/20836166

Appendix

Figure 1-A

Forecast using smoothing model 'MAM'

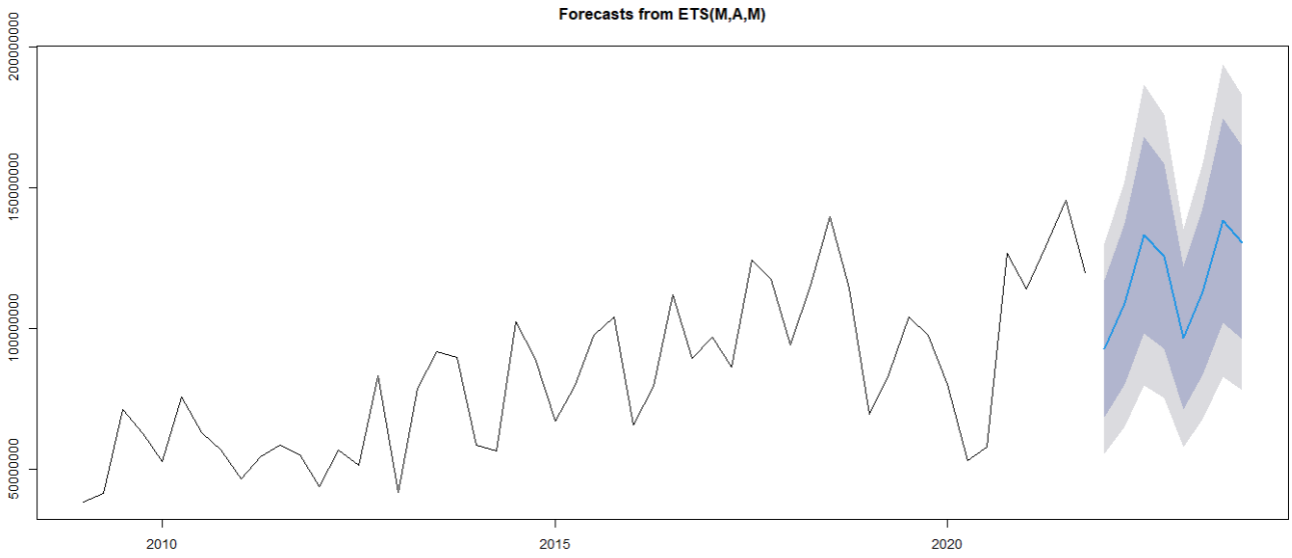


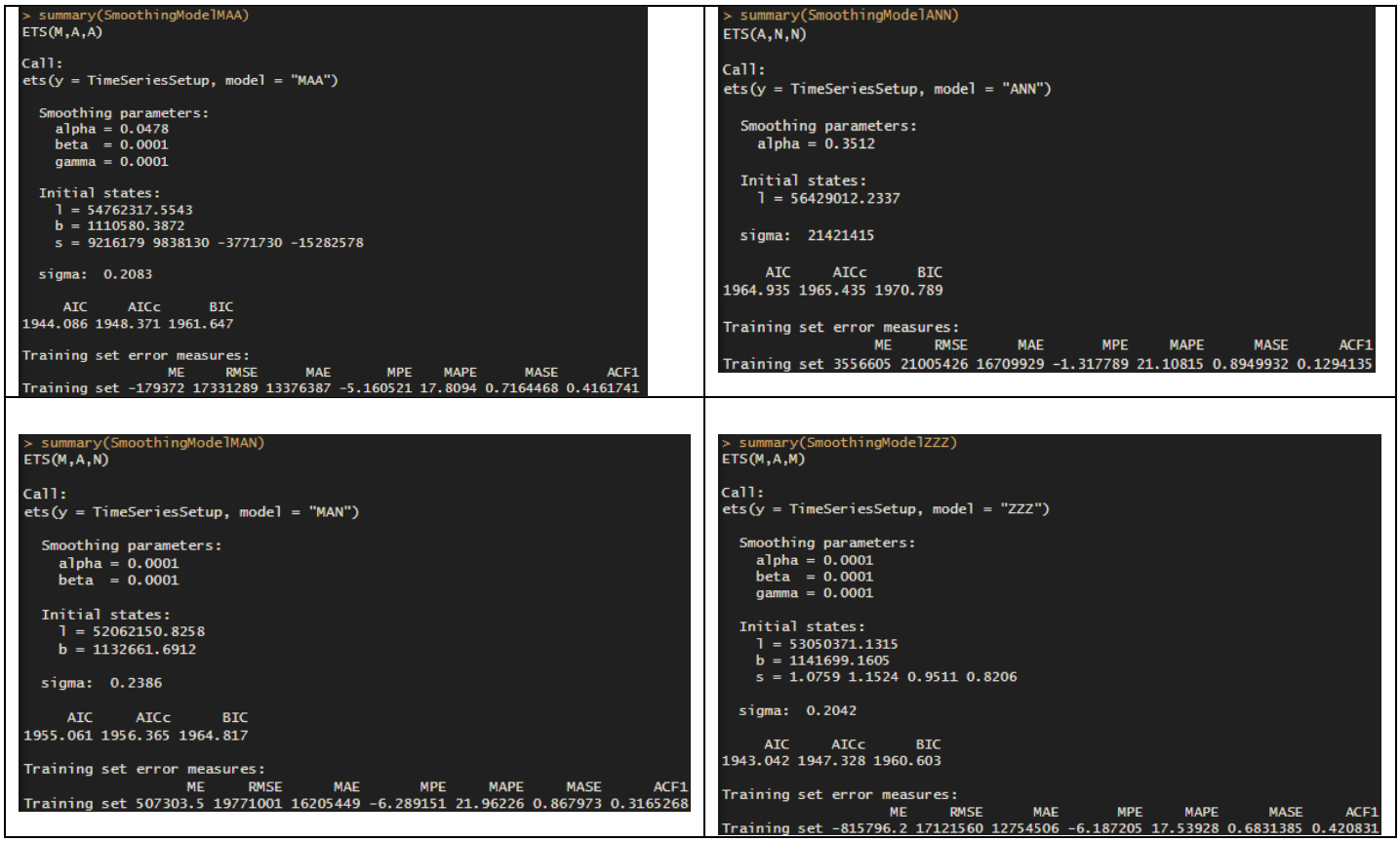
Figure 1-B

Sales Forecast of 8 Quarters

```
> ForecastSales
        Point Forecast     Lo 80     Hi 80    Lo 95     Hi 95
2022 Q1       92960909  68628853 117292965 55748241 130173577
2022 Q2      108828110  80342893 137313328 65263730 152392491
2022 Q3      133179914  98320731 168039096 79867395 186492433
2022 Q4      125570321  92702909 158437732 75303952 175836689
2023 Q1       96694931  71385506 122004357 57987505 135402357
2023 Q2      113156020  83537980 142774061 67859137 158452904
2023 Q3      138424115 102192265 174655965 83012283 193835948
2023 Q4      130466680  96317641 164615719 78240231 182693129
```

Figure 2

Testing Different Models



```
> summary(SmoothingModelMAA)
ETS(M,A,A)

Call:
ets(y = TimeSeriesSetup, model = "MAA")

  Smoothing parameters:
    alpha = 0.0478
    beta  = 0.0001
    gamma = 0.0001

  Initial states:
    l = 54762317.5543
    b = 1110580.3872
    s = 9216179 9838130 -3771730 -15282578

  sigma:  0.2083

     AIC     AICc      BIC
1944.086 1948.371 1961.647

Training set error measures:
                    ME     RMSE      MAE       MPE    MAPE      MASE      ACF1
Training set -179372 17331289 13376387 -5.160521 17.8094 0.7164468 0.4161741
```

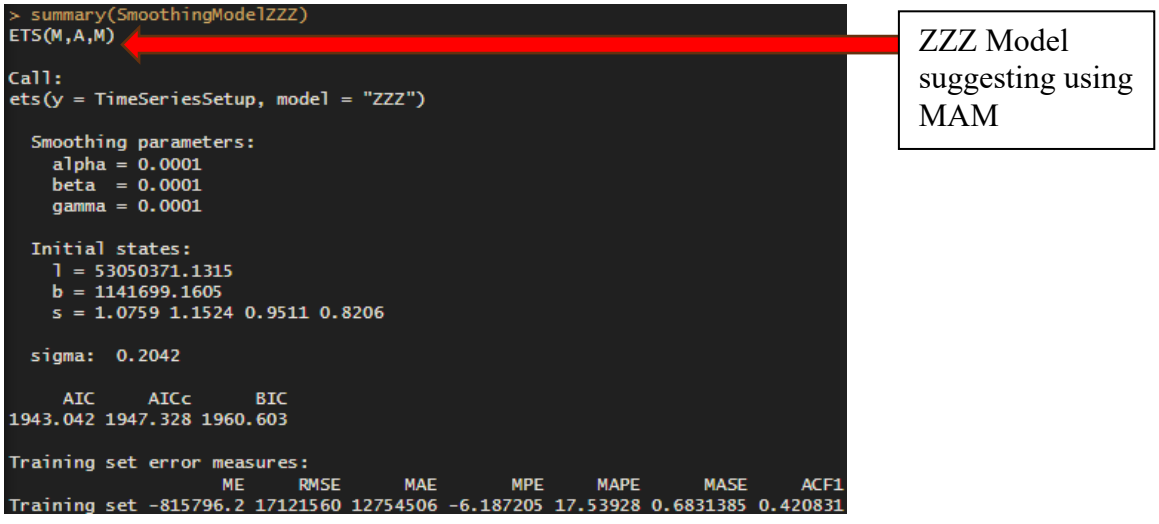
```
> summary(SmoothingModelANN)
ETS(A,N,N)

Call:
ets(y = TimeSeriesSetup, model = "ANN")

  Smoothing parameters:
    alpha = 0.3512

  Initial states:
    l = 56429012.2337

  sigma:  21421415

     AIC     AICc      BIC
1964.935 1965.435 1970.789

Training set error measures:
                   ME     RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set 3556605 21005426 16709929 -1.317789 21.10815 0.8949932 0.1294135
```

```
> summary(SmoothingModelMAN)
ETS(M,A,N)

Call:
ets(y = TimeSeriesSetup, model = "MAN")

  Smoothing parameters:
    alpha = 0.0001
    beta  = 0.0001

  Initial states:
    l = 52062150.8258
    b = 1132661.6912

  sigma:  0.2386

     AIC     AICc      BIC
1955.061 1956.365 1964.817

Training set error measures:
                   ME     RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set 507303.5 19771001 16205449 -6.289151 21.96226 0.867973 0.3165268
```

```
> summary(SmoothingModelZZZ)
ETS(M,A,M)

Call:
ets(y = TimeSeriesSetup, model = "ZZZ")

  Smoothing parameters:
    alpha = 0.0001
    beta  = 0.0001
    gamma = 0.0001

  Initial states:
    l = 53050371.1315
    b = 1141699.1605
    s = 1.0759 1.1524 0.9511 0.8206

  sigma:  0.2042

     AIC     AICc      BIC
1943.042 1947.328 1960.603

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set -815796.2 17121560 12754506 -6.187205 17.53928 0.6831385 0.420831
```

Figure 3

Forecast using smoothing model ‘MAM’



```
> summary(SmoothingModelZZZ)
ETS(M,A,M)

Call:
ets(y = TimeSeriesSetup, model = "ZZZ")

  Smoothing parameters:
    alpha = 0.0001
    beta  = 0.0001
    gamma = 0.0001

  Initial states:
    l = 53050371.1315
    b = 1141699.1605
    s = 1.0759 1.1524 0.9511 0.8206

  sigma:  0.2042

     AIC     AICc      BIC
1943.042 1947.328 1960.603

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set -815796.2 17121560 12754506 -6.187205 17.53928 0.6831385 0.420831
```

ZZZ Model suggesting using MAM

Figure 4

Selecting model 'MAM' as per the recommendation, with least AIC

```
> summary(SmoothingModelMAM)
ETS(M,A,M)

Call:
ets(y = TimeSeriesSetup, model = "MAM")

  Smoothing parameters:
    alpha = 0.0001
    beta  = 0.0001
    gamma = 0.0001

  Initial states:
    l = 53050371.1315
    b = 1141699.1605
    s = 1.0759 1.1524 0.9511 0.8206

  sigma:  0.2042

     AIC     AICc      BIC
1943.042 1947.328 1960.603

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE     ACF1
Training set -815796.2 17121560 12754506 -6.187205 17.53928 0.6831385 0.420831
```

Figure 5

Regression Analysis Summary

```
> summary(Regression)

Call:
lm(formula = TOTALSALES ~ TIME + QUARTER, data = Reg1)

Residuals:
      Min        1Q    Median        3Q       Max
-60851228  -8712536    223814  11508771  30842026

Coefficients:
            Estimate Std. Error t value       Pr(>|t|)
(Intercept) 35762431    6402456   5.586 0.000001133702 ***
TIME         1251007     164166   7.620 0.000000000944 ***
QUARTERQ2    7969301    6951397   1.146          0.25742
QUARTERQ3   24271407    6957210   3.489          0.00107 **
QUARTERQ4   22082877    6966888   3.170          0.00268 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17720000 on 47 degrees of freedom
Multiple R-squared:  0.628,     Adjusted R-squared:  0.5963
F-statistic: 19.83 on 4 and 47 DF,  p-value: 0.000000001275
```

Figure 6

Simple Regression Analysis Summary

```
> summary(SimpleRegression)

Call:
lm(formula = TOTALSALES ~ TIME, data = Reg1)

Residuals:
      Min        1Q    Median        3Q       Max
-54452374 -11360597   -781971  15877253  40959903

Coefficients:
            Estimate Std. Error t value        Pr(>|t|)
(Intercept) 48129346    5622643   8.560 0.0000000000229 ***
TIME         1296817     184622   7.024 0.0000000055005 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19980000 on 50 degrees of freedom
Multiple R-squared:  0.4967,    Adjusted R-squared:  0.4866
F-statistic: 49.34 on 1 and 50 DF,  p-value: 0.0000000055
```

Figure 7

Multiple Regression Analysis Summary

```
Call:
lm(formula = SALE_PRICE ~ ., data = MultiRegPrep)

Residuals:
     Min       1Q   Median       3Q      Max
-3088091  -103003    -7099    89585  3413457

Coefficients:
                                 Estimate       Std. Error t value             Pr(>|t|)
(Intercept)                -8060456.91384578   404271.58142673 -19.938 < 0.0000000000000002 ***
BUILDING_CLASS_FINAL_ROLLA1   127389.38376143    13980.00336378   9.112 < 0.0000000000000002 ***
BUILDING_CLASS_FINAL_ROLLA2    31975.48789313    15789.07157171   2.025               0.0429 *
BUILDING_CLASS_FINAL_ROLLA3   264855.22183062    26049.16880306  10.168 < 0.0000000000000002 ***
BUILDING_CLASS_FINAL_ROLLA5  -126143.03144289    14477.84054664  -8.713 < 0.0000000000000002 ***
BUILDING_CLASS_FINAL_ROLLA9    -3959.14772337    79547.88176570  -0.050               0.9603
BUILDING_CLASS_FINAL_ROLLB1   156088.63213877    18730.99137810   8.333 < 0.0000000000000002 ***
BUILDING_CLASS_FINAL_ROLLB2   194137.02801301    17497.18380166  11.095 < 0.0000000000000002 ***
BUILDING_CLASS_FINAL_ROLLB3   204503.79210796    18900.19138870  10.820 < 0.0000000000000002 ***
BUILDING_CLASS_FINAL_ROLLB9   187512.63543790    33554.52284635   5.588  0.00000002411913036 ***
BUILDING_CLASS_FINAL_ROLLC0   253505.99037504    22710.89162954  11.162 < 0.0000000000000002 ***
BUILDING_CLASS_FINAL_ROLLC1  1308880.58898993   102246.83332175  12.801 < 0.0000000000000002 ***
BUILDING_CLASS_FINAL_ROLLC2   226405.82687716   107256.66771077   2.111               0.0348 *
BUILDING_CLASS_FINAL_ROLLC3   402939.09005478    57180.04267308   7.047  0.00000000000207313 ***
BUILDING_CLASS_FINAL_ROLLC6   432047.00065940   200117.25030467   2.159               0.0309 *
BUILDING_CLASS_FINAL_ROLLC7  2084378.19512172   261288.91738946   7.977  0.00000000000000183 ***
BUILDING_CLASS_FINAL_ROLLC9  2511331.01859420   136574.04202323  18.388 < 0.0000000000000002 ***
BUILDING_CLASS_FINAL_ROLLD4  2363758.18722465   309053.21201273   7.648  0.00000000000002417 ***
BUILDING_CLASS_FINAL_ROLLR2  -511785.26317738   107534.41743772  -4.759  0.00000199652685823 ***
BUILDING_CLASS_FINAL_ROLLR3  -315467.49985229    43574.91674873  -7.240  0.00000000000051666 ***
BUILDING_CLASS_FINAL_ROLLR4  -295448.53238475    43303.27970548  -6.823  0.00000000000996687 ***
SALE_DATE                         0.00116789        0.00003147  37.110 < 0.0000000000000002 ***
YEAR_BUILT                     3660.55601010      205.93074933  17.776 < 0.0000000000000002 ***
GROSS_SQUARE_FEET                23.67594800        1.55699797  15.206 < 0.0000000000000002 ***
RESIDENTIAL_UNITS             -35981.47032403     2383.33137626 -15.097 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 260700 on 5110 degrees of freedom
Multiple R-squared:  0.4383,    Adjusted R-squared:  0.4356
F-statistic: 166.1 on 24 and 5110 DF,  p-value: < 0.00000000000000022
```

Figure 8

Top Overpriced Sales of the Properties

BUILDING_CLASS_FINAL_ROLL	SALE_DATE	YEAR_BUILT	GROSS_SQUARE_FEET	RESIDENTIAL_UNITS	SALE_PRICE	RESIDUAL	ADDRESS
D4	2017-09-13	1961	373772	240	6100000	2646509.060444769915193	18-40 211TH STREET
C9	2013-11-22	1957	27440	36	6000000	3413457.291315825190395	36-21 213TH STREET
D4	2017-07-31	1961	363000	248	5400000	2493838.002548291813582	18-35 CORP KENNEDY ST
C1	2018-04-30	1931	10000	12	4800000	2896967.568693246226758	39-30 214TH PLACE
C9	2012-12-21	1968	28648	30	4600000	1762606.251838794909418	39-06 CORPORAL STONE ST
A5	2015-07-02	1959	2450	1	4030000	3346688.838051640428603	42-39 208TH ST
D4	2017-11-08	1960	99500	120	3800000	2520392.047295061871409	209-15 18TH AVENUE
B3	2019-05-24	1910	3758	2	3000000	2026934.139830267988145	220-16 43RD AVENUE
C1	2012-12-26	1968	8577	14	2862500	1126548.578189596300945	213-03 41ST AVENUE
C7	2010-10-13	1988	7761	8	2700000	-0.000000001552053	211-63 46TH AVENUE
A1	2021-10-25	1910	2992	1	2460000	1456900.970892727375031	218-15 40TH AVENUE
A0	2017-03-09	1945	1329	1	2450000	1656176.181558731710538	58-33 220TH STREET
A3	2017-07-18	1991	2700	1	2400000	1127256.961728466209024	215-15 27TH AVENUE
A1	2016-09-22	1960	4627	1	2300000	1262747.403018101816997	24-25 LITTLE NECK BLVD
A2	2020-06-26	1965	2608	1	2300000	1249116.206397710135207	39-33 223RD STREET
D4	2017-09-29	1961	489440	310	2300000	-1374952.067668046103790	211-35 23RD AVENUE
A1	2021-04-28	1940	3332	1	2250000	1147197.563143039355054	214-41 29TH AVENUE

Figure 9

Property Details of 39-30 214TH PLACE, Bayside, NYC

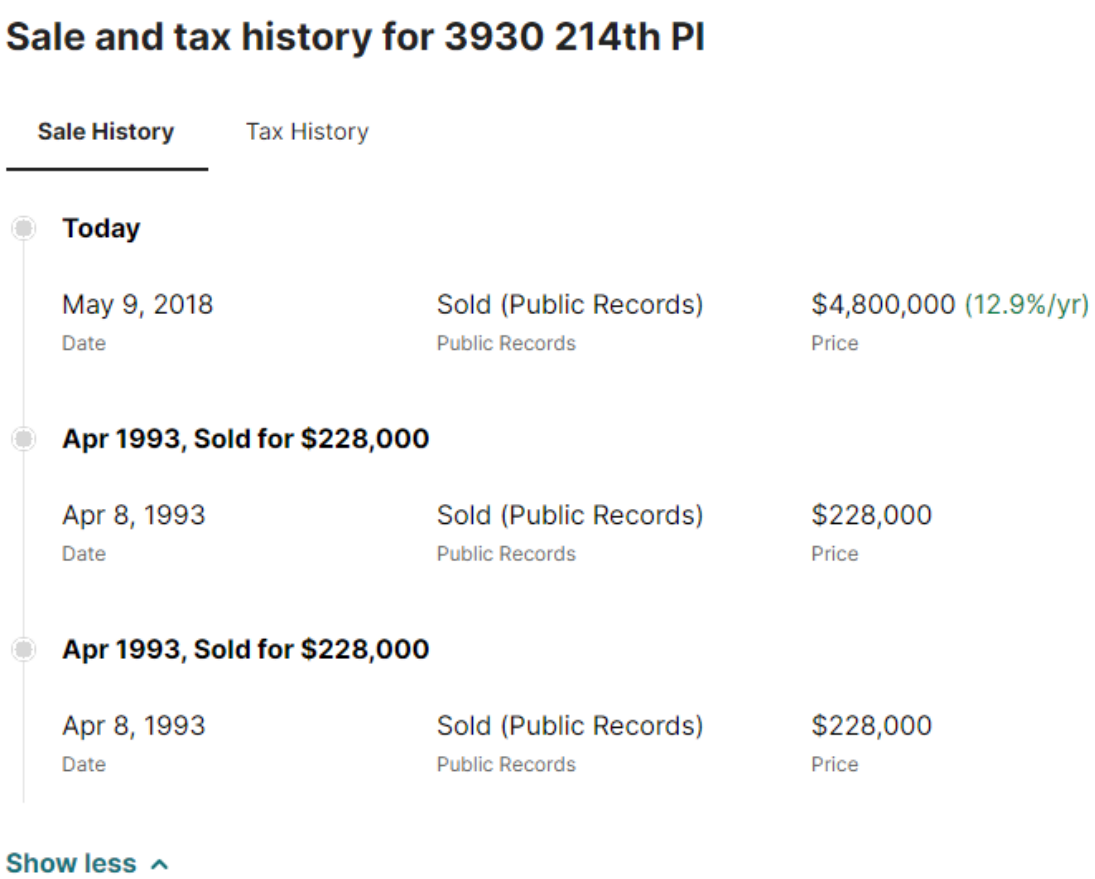


Figure 10

Top Best Bargains of the Properties

	BUILDING CLASS FINAL ROLL	SALE DATE	YEAR BUILT	GROSS SQUARE FEET	RESIDENTIAL UNITS	SALE PRICE	RESIDUAL	ADDRESS
1	D4	2010-10-07	1960	196000	120	215000	-3088091.1	209-20 18TH AVE
2	C9	2017-09-27	1931	8640	12	262667	-2788921.8	38-12 213TH STREET
3	D4	2015-03-19	1961	363000	248	199000	-2619878.2	18-75 CORPORAL KENNEDY STR
4	C9	2017-10-05	1925	25000	32	311000	-2387141.8	215-37 43RD AVENUE
5	C1	2017-09-27	1931	6720	8	95000	-1852606.4	39-41 213TH STREET
6	C1	2021-08-27	1922	14080	19	400000	-1437415.9	202-18 43RD AVENUE
7	D4	2017-09-29	1961	489440	310	2300000	-1374952.1	211-35 23RD AVENUE
8	B2	2019-12-04	2005	2853	2	161780	-1146821.0	209-54 45TH ROAD
9	C0	2021-06-29	1950	2680	3	200000	-984381.1	47-25 BELL BOULEVARD
10	B1	2021-05-24	1965	2656	2	200000	-973652.7	35-02 203RD STREET
11	B2	2017-07-06	2008	2090	2	280000	-932619.6	215-07 36TH AVENUE
12	B2	2016-05-04	1975	3498	2	200000	-881969.2	15-54 208TH PLACE
13	A1	2014-11-17	2005	3200	1	220000	-880080.4	50-40 214 STREET
14	C0	2020-12-31	1950	2680	3	300000	-866218.0	47-40 215TH STREET
15	C0	2013-11-11	1960	3076	3	103000	-846137.1	212-06 43RD AVENUE
16	B3	2018-05-21	1955	2210	2	223100	-840907.1	58-34 OCEANIA STREET

Figure 11

Top 3 Properties with Highest Negative Residual Values.