Running head: ASSIGNMENT 4

Assignment 4

Yuqi Zheng

AD 571 Business Analytics Foundations

Boston University

2019 Fall

10/28/2019

Table of Contents

Executive Summary

This report uses a predictive analysis model and uses the R language to perform a deep prediction and analysis of my neighbor, Sunnyside. We will also analyze the residential properties of Sunnyside.

The sales of Sunnyside for the following eight quarters can be forecasted through R by applying time series analysis as well as the multiple regression model. Even though at the beginning, the first Quarter of 2018 always lower than the last Quarter of 2017, both of the two forecasts exhibit the same cycle that the total sale of the last Quarter is most significant than the previous seven quarters in 2018. Consequently, the strategy of the firm can be adjusted by putting more resources into the last Quarter. During the determination of sale by multiple regression model, the YearBuilt is redundant independent variables, meanwhile, the Building A2, A3, A6, A9, B2, B3, B9, C5, C6, D9, E1, M1, o8, o9, and S1 also have no influence on sale price in the Building Class Final Roll. Meanwhile, Gross square feet and SaleDate are the most useful predictors compared with the least useful predictor, "Building Class Final C5," which located at 48 Grand Ave. Through careful analysis, the building D1 is the most prominent bargain property. When sold in 2010, it was the biggest bargain with a residual value of -11055168.5. Building A5, which is sold in 2004, is the most overpriced property with a residual value of 23076717.2. The company needs to focus more on the building in One family attached or semi-detached type because it usually has a higher average GrossSqFt per unit because it will impact the sale price a lot.

Time Series Analysis

Before making predictions, we need to filter out my target neighbor, Sunnyside, like the previous project and use the deletion time and state to narrow the data. Time series prediction is a model to predict future values based on previously observed values. Using this method to predict Sunnyside's sales for the next eight quarters, we need to use the formula mutate(t=Year*4+SalesQrt-2008*4-4) to let R sort the sales quarter from 1 to 32 from 2009 to 2016. Then build the model with ts () and ets () and make predictions.

By using this method, we can see that the sales situation in the next eight quarters is quite unstable. As shown in Appendix 1, sales of Sunnyside in 2017 are not as optimistic as in 2018. In particular, there were two substantial fluctuations from 2013 to 2017, with the highest peak reaching around 2.0e+08. At the same time, prices in the First Quarter of 2018 are always lower than in the last Quarter of 2017, and in the last Quarter of 2018, sales are higher than at the end of 2017 (except for exceptional circumstances). However, for the second and third quarters, their performance was relatively flat no matter what year. According to Appendix1 and 2, Sunnyside's sales changes in the next eight quarters will not change very sharply. Compared with the sharp rise and fall in previous years, the next two years will rise with slight fluctuations.

Multiple Regression Model to Forecast

Unlike the previous section, this section will analyze the use of multiple regression models to predict Sunnyside's sales for the next eight quarters. Using the cbind() add a new column named Quarter in the database, then through lm() with TotalSale, Quarter, and t to build the model. We could find the result in Appendix 3.

As the result shows, the p-value is equal to 0.0281, which is less than 0.05. So, we can show that time and quarter changes do affect the sales of Sunnyside. Also, we can find that the r-squared value is equal to 0.2217, which is a relatively large value, meaning that 22.17% of sales are related to seasonal changes.

Through the predict.lm(), R outputs the forecast for the next eight quarters, which could be found in Appendix 4. Similar to Appendix1, the First Quarter of 2018, which as the fifth Quarter of the table, is always smaller than the last Quarter of the previous year. But the amount of eighth Quarter is more significant than the fourth Quarter's, indicating that sales in 2018 are still higher than in 2017. In terms of gains, the increase in 2018 is also higher than in 2017.

Multiple Regression Model to Sale

In this section, we will use a multiple regression model to determine the sale of specific residential properties in Sunnyside, which include: Sale Date, Year built, Building type (categorical), Gross Square Feet, and Number of Units. We use lm() with SalePrice, SaleDate, YearBuilt, BuildingClassFinalRoll, GrossSqFt, and ResidentialUnits to build the model (see results in Appendix 5).

A p-value higher than 0.05 (> 0.05) is not statistically significant and indicates weak evidence against the null hypothesis. And it is opposite when p-value less than 0.5. From the table, SaleDate and GrossSqFt are the most useful predictors for Sunnyside sales because their p-value is equal to 2e-16, which is that the p-value is extremely close to zero, but it does not equal to zero. It also means it has an excellent impaction on Sunnyside's sales. Thus, SaleDate and GrossSqFt are key factors we need to consider in future forecasts.

Appendix 5 also shows that BuildingClassFinalRollC5 is the least useful predictor because his p-value is equal to 0.467380, very close to 0.5. Explain that C5 will not be sold repeatedly, and his presence will not affect our forecast results.

According to p-value, we can see that YearBuilt's p-value is 0.214516 and Building A2, A3, A6, A9, B2, B3, B9, C5, C6, D9, E1, M1, o8, o9, S1 also have p-values greater than 0.5, which means that they will less impact the sales of Sunnyside. When the p-value is greater than 0.5, we generally say that the connection between this variable and our hypothesis is smaller. Thus, Building A2, A3, A6, A9, B2, B3, B9, C5, C6, D9, E1, M1, o8, o9, S1, and YearBuilt are redundant variables for Sunnyside' sale price.

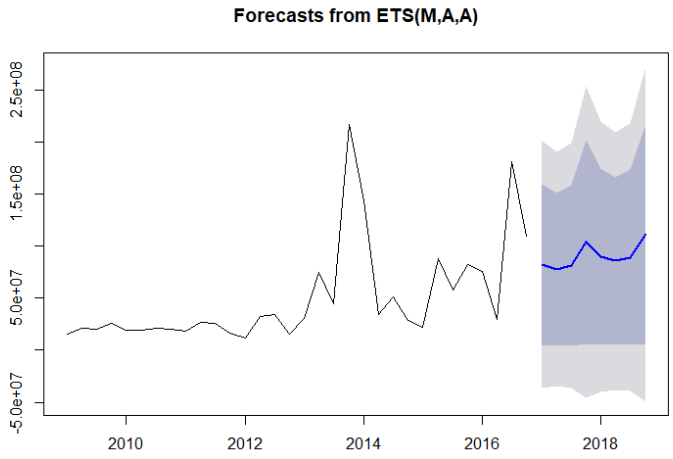
To find the biggest bargains and the most overpriced properties, though mutate() add the new column, which is the residual for the model. From the table, we could find the biggest bargain property is the Buildingclassfinalroll D1, which is located at 41-22 42ND St. It has the lowest residential, which is -11055168.5. The residential type is Elevator Apt; Semi-fireproof without stores. It's was built in 1937 and was sold at $57000 on 5/20/2010. Its gross square feet are 61770, with 60 residential units. The most overpriced property is the Buildingclassfinalroll A5, which is located at 4518 Greenpoint Ave. The residential type is One family attached or semi-detached. It's was built in 1925 and was sold at $23514799 on 03/31/2004. Its gross square feet are 1680, with only one residential unit and the highest residential, 23076717.2.

According to our previous analysis, these differences exist mainly because of the factors of SaleDate and GrossSqFt. The time difference between the biggest bargain property and the most overpriced property is about six years. Also, the average GrossSqFt per unit of D1 is 61770/60=1029.5, and the GrossSqFt of A5 is 1680 per unit, which also confirms our analysis above.

Appendix

```
> forecast(ts.model,8)
         Point Forecast   Lo 80     Hi 80      Lo 95     Hi 95
2017 Q1       82454736 4718848 160190623 -36432045 201341516
2017 Q2       78137596 4471778 151803413 -34524551 190799742
2017 Q3       81607008 4670326 158543691 -36057494 199271511
2017 Q4      104107016 5957985 202256047 -45998973 254213006
2018 Q1       90187551 5161357 175213746 -39848792 220223894
2018 Q2       85870411 4914265 166826558 -37941330 209682153
2018 Q3       89339824 5112790 173566858 -39474309 218153957
2018 Q4      111839832 6400439 217279226 -49415805 273095469
```

Appendix 1



Appendix 2

```
> summary(reg)

Call:
lm(formula = TotalSale ~ Quarter + t, data = df.Historical2)

Residuals:
      Min        1Q    Median        3Q       Max
-53230963 -20994158  -6627060   5943702 146819009

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -553795   20125556  -0.028  0.97825
QuarterQ2   -4187000   22043167  -0.190  0.85077
QuarterQ3    6839071   22092242   0.310  0.75927
QuarterQ4   13706595   22173794   0.618  0.54166
t            2854521     849702   3.359  0.00234 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 44050000 on 27 degrees of freedom
Multiple R-squared:  0.3221,    Adjusted R-squared:  0.2217
F-statistic: 3.208 on 4 and 27 DF,  p-value: 0.0281
```

Appendix 3

```
> predict.lm(reg,x,interval = "confidence")
        fit       lwr       upr
1  93645410 48855529 138435290
2  92312931 47523051 137102812
3 106193523 61403643 150983403
4 115915569 71125689 160705449
5 105063495 55138864 154988126
6 103731017 53806386 153655648
7 117611608 67686978 167536239
8 127333654 77409024 177258285
```

Appendix 4

```
> summary(model)

Call:
lm(formula = SalePrice ~ SaleDate + YearBuilt + BuildingClassFinalRoll +
    GrossSqFt + ResidentialUnits, data = df.Historical3)

Residuals:
      Min        1Q    Median        3Q       Max
-11055169   -176210      4349    159430  23076717

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.249e+06  2.191e+06   0.570 0.568575
SaleDate                 1.394e-03  1.676e-04   8.315  < 2e-16 ***
YearBuilt               -1.386e+03  1.116e+03  -1.242 0.214516
BuildingClassFinalRollA2 -5.497e+03  1.138e+05  -0.048 0.961480
BuildingClassFinalRollA3 -1.209e+05  5.129e+05  -0.236 0.813697
BuildingClassFinalRollA5  2.499e+05  9.208e+04   2.714 0.006679 **
BuildingClassFinalRollA6  1.377e+05  6.175e+05   0.223 0.823525
BuildingClassFinalRollA9  4.386e+04  1.590e+05   0.276 0.782654
BuildingClassFinalRollB1  2.723e+05  8.904e+04   3.058 0.002247 **
BuildingClassFinalRollB2  1.817e+05  1.151e+05   1.579 0.114494
BuildingClassFinalRollB3  2.210e+05  1.176e+05   1.880 0.060265 .
BuildingClassFinalRollB9  1.374e+05  1.659e+05   0.828 0.407790
BuildingClassFinalRollC0  3.759e+05  1.098e+05   3.424 0.000624 ***
BuildingClassFinalRollC1  2.603e+06  1.528e+05  17.035  < 2e-16 ***
BuildingClassFinalRollC2  5.206e+05  1.084e+05   4.803 1.64e-06 ***
BuildingClassFinalRollC3  6.195e+05  2.843e+05   2.179 0.029400 *
BuildingClassFinalRollC5  6.321e+05  8.696e+05   0.727 0.467380
BuildingClassFinalRollC6 -7.827e+03  8.699e+05  -0.009 0.992822
BuildingClassFinalRollC7  1.283e+07  4.384e+05  29.261  < 2e-16 ***
BuildingClassFinalRollD1  9.245e+06  2.129e+05  43.415  < 2e-16 ***
BuildingClassFinalRollD4 -2.802e+06  5.709e+05  -4.909 9.66e-07 ***
BuildingClassFinalRollD7  1.346e+07  3.308e+05  40.682  < 2e-16 ***
BuildingClassFinalRollD9  8.570e+05  8.729e+05   0.982 0.326281
BuildingClassFinalRollE1 -3.944e+05  8.771e+05  -0.450 0.652974
BuildingClassFinalRollM1  5.060e+05  8.695e+05   0.582 0.560614
BuildingClassFinalRollO8  1.859e+05  6.184e+05   0.301 0.763779
BuildingClassFinalRollO9  1.063e+05  8.698e+05   0.122 0.902772
BuildingClassFinalRollS1  8.208e+04  6.178e+05   0.133 0.894311
GrossSqFt                 1.030e+02  1.181e+01   8.720  < 2e-16 ***
ResidentialUnits         -7.204e+04  9.970e+03  -7.225 6.33e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1225000 on 2964 degrees of freedom
Multiple R-squared:  0.7171,    Adjusted R-squared:  0.7144
F-statistic: 259.1 on 29 and 2964 DF,  p-value: < 2.2e-16
```

Appendix 5