### 

### Assignment 3: Predictive & Prescriptive Analytics

**Part 1:** Predictive analytics - Multiple linear regression

* Use a multiple regression model to determine the sale of a given residential property in your neighborhood since 2009 (year of transaction). Include:
  + Year built
  + BUILDING\_CODE\_ID (or BUILDING\_CODE\_FINAL\_ROLL depending on how you joined tables; categorical variable)
  + Gross square feet
  + Number of units

For this, I used the ‘lm’ function in R. I used the year built, building code, gross square feet, and the number of units as predictor variables. I used the sale price as the outcome variable.

* After building the model, answer the question "What are the most and least useful predictors of the amount of a sale?"

To figure out the most and least useful predictors of the amount of a sale from the model, we need to look at the coefficients and their corresponding p-values.

Most Useful Predictors:

1. Gross square footage: The model shows a coefficient of 5.053e+02 for the variable gross square foot. This means that the amount of the sale is anticipated to rise by about $505.3 for every additional unit of gross square footage. The extremely significant p-value for this prediction is 2e-16 or 0.00000000000000022. Which is a very small number. This suggests a strong relationship with the sale amount.
2. Residential Units: The coefficient is 6.884e+04 for this variable. Thus the sale amount should rise by about $68,840 for every new residential unit. The significant p-value for this prediction is 0.00451.
3. Building class D4: This predictor represents a specific category of building class. The coefficient of -3.384e+07 means that being classified as "D4" is related to a significant decrease in the sale amount. It has a highly significant p-value of < 2e-16 as well.

Least Useful Predictors:

1. Year Built: According to the coefficient of -1.160e+01, the sale amount should drop by about $11 for each year as the building gets older. The non-significant p-value of 0.93316 for this predictor, however, suggests that it might not be an accurate predictor of the sale amount.
2. The remaining building classes: These predictors reflect distinct building class subcategories. They may not be accurate predictors, as they have non-significant p-values.

**Part 2:** Predictive analytics - Time series forecasting

* Use the sales of residential properties in your neighborhood since 2009 to build a forecast model for the next 8 quarters of sales.

To create a time series forecasting model, I used the forecast package in R. I created a time series object with the filtered data with the frequency set to 4 to have a quarterly time unit.

Then I used the time series object to model and forecast the next 8 quarters of sales.

* Present your findings to include all the below:
  + Define the model type that you used (Only use "N" and/or "A") and why you used it.

For this forecast, I have used the Holt-Winters or triple exponential smoothing model (AAA). I used this model as the forecast is a time series forecast and the data has trends and seasonality. Even though the plot (Figure 1) is quite flat and does not appear to have a very distinct upward or downward trend, I am assuming that there might be a slight upward trend. By considering both trend and seasonality, the model provides a more accurate and robust forecast for data with seasonal variations.

If I am to disregard the existence of a trend in this time series data, the forecast might not be as accurate.

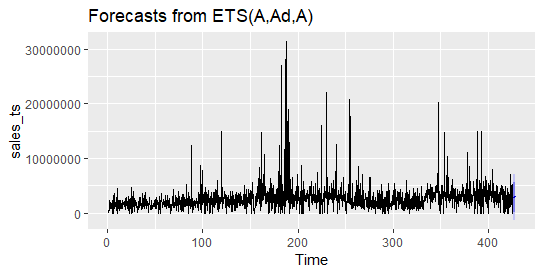


Figure 1

* + A table that shows the forecast numbers, confidence bands for the next 8 quarters.

The below table (Table 1) shows the forecast numbers and confidence bands for the next 8 quarters.

Quarter Forecasted\_Sales Lower\_CI Upper\_CI

1 426.75 2841100 -1408688 7090889

2 427.00 2888436 -1364842 7141714

3 427.25 3051049 -1205732 7307831

4 427.50 2948751 -1311547 7209048

5 427.75 2839586 -1424258 7103430

6 428.00 2886963 -1380422 7154348

7 428.25 3049617 -1221320 7320555

8 428.50 2947358 -1327143 7221860

Table 1

The lower and upper confidence intervals provide a range of values around the forecasted sales that capture the uncertainty or variability associated with the forecast. The wider the confidence interval, the greater the uncertainty in the forecasted sales. Decision-makers can use these intervals to assess the range of possible outcomes and make informed decisions based on the level of risk they are willing to accept.

Part 3: Prescriptive analytics

* Develop an optimization model for your real estate business by following the instructions in the attached file.
* Write your findings with a focus on the output, interpretation of the output, and what the insights mean for our decision-making process.

By focusing on the maximization of NPV, I have created an Excel optimization model using the tool Solver. I took three decision variables. They are - the number of employees, the number of offices, and the commission rate. Then I calculated the total cost using the relationship between the decision variables and the constraints. I used Solver to find the optimal solution where the NPV is the highest over the next eight quarters.

The results show that in order to get the maximum NPV, the real estate company needs to open one office and hire one employee. The commission rate needs to be 5%.

I took the budget of 20,000 USD as the cash flow and the difference between the cost and the budget as the profit based on which I calculated the NPV. In order to stay under budget and get positive results, the real estate company needs to operate with less manpower and within a small office space. The calculated average cost per square foot in the Park Slope area for commercial properties is quite high. It is 1296.937 USD per square foot.

Thus the decision makers need to take all of the constraints and the profitability into consideration before deciding to open a new real estate office in the Park Slope area.