**Assignment 4**: Data Analytics (Spring 2024) (15% written)

**Due: 04/02/2024** (by 11:59pm ET) Submission method: written document via LMS

Please use the following file naming for electronic submission: **DataAnalytics\_A4\_YOURFIRSTNAME\_YOURLASTNAME.xxx**

Late submission policy: first time with valid reason – no penalty, otherwise 20% of score deducted each late day

Note: Your report for this assignment should be the result of your own individual work. Take care to avoid plagiarism (“copying”), and include references to all web resources, texts, and class presentations. You may discuss the problems with other students, but do not take written notes during these discussions, and do not share your written solutions.

**General Assignment**: Pattern, trend, relations: model development and evaluation of housing (Brooklyn, Manhattan, Queens) **NYC Citywide Annualized Calendar Sales Update** datasets available:

[**https://data.cityofnewyork.us/City-Government/NYC-Citywide-Annualized-Calendar-Sales-Update/w2pb-icbu**](https://data.cityofnewyork.us/City-Government/NYC-Citywide-Annualized-Calendar-Sales-Update/w2pb-icbu)

The weighting score for each question is included below.

Please use the question numbering below for your written responses for this assignment.

Please include code (fragments and/or scripts) and the plots you generate for the questions below.

1. For any **one** of the Brooklyn, Manhattan, Queens sales datasets, perform the following:

a). Describe the type of patterns or trends you might look for and how you plan to

model them. Describe any exploratory data analysis you performed. Include

plots and other descriptions. Min. 5 sentences (1%)

**I would expect the number of units to correlate with the sales price to a very high extent. I would also expect sales prices to be very different depending on zipcodes. Additionally, I think sales prices correlate with the land area but not to an extremely high degree since real estate is very expensive in NYC even on small land areas. I did some EDA just looking at the distribution of sales prices. There were an extraordinary amount of zeros and very long tails. I also observed a very large number of values below 1,000, which I think is quite unrealistic, so I limited the range from 10,000 to 100,000,000 to get a look at the shape of the distribution. A screenshot of this histogram is below.**

**A graph of a sales graph

Description automatically generated with medium confidence**

b). Identify the outlier values in the data for Sales Price or on a variable you choose, explain

why you consider those data points are outliers? Use the Cook’s Distance and IQR (Inter

Quartile Range) to identify the outlier points (1%)

**Cook’s Distance identified outliers mainly as values towards the low end, such as prices in the 1,000-2,000 range. The IQR is from around 650,000 to 2,500,000, so the outliers are numbers below and above that range.**

c). Conduct Multivariate Regression on the chosen dataset to predict the Sales Price using

Gross Square feet, Land Square feet. When you conduct the multivariate regression,

make sure to draw at least 3 samples from the data and compare the different results you

obtained. Explain the results Min. 5 sentences (1%)

**The regression results varied heavily based on the samples. I took three samples of 10,000 rows and performed the same regression fit to all of them, however, one of the samples ended up with a multiple R-squared value of 0.377, and the other two ended up with multiple R-squared values of over 0.85. The model has lots of data points, which is why I think the R-squared values are so high for the second two samples. However, the gross square feet and land square feet are clearly good indications of sales price since they predicted sales price with a high accuracy 2/3 of the time.**

d). Pick one or more models (these need not be restricted to the models you’ve learned so far

[Decision Trees, KNN, K-Means, RandomForest…]) to explore the chosen data. Interpret

the model fits and indicate significance. Describe any cleaning you had to do and why.

Min. 5 sentences (2%)

2. For your chosen dataset:

a). Apply the model(s) to predict quantities of interest (that you choose).

Describe (contingency table) or plot the predictions. Min. 2-3 sentences (4000-level 5%,

6000-level 3%)

b). Examine the fit(s). Perform a significance test that is suitable for the variables you are

investigating and describe the results. Min. 2-3 sentences (4000-level 4%, 6000-level

3%)

c). Discuss any observations you had about the datasets/ variables, other data in the dataset

and/or your confidence in the result. Min 1-2 sentences (1%)

3. 6000-level question (3%). Draw conclusions from this study – about the model type and

suitability/ deficiencies. Describe what worked and why/ why not. Min. 4-5 sentences