Assignment 5: Data Analytics (Fall 2024) (15% written)

Due: November 29th, 2024 (by 08:00pm ET) Submission method: email (eleisa2@rpi.edu)

Please use the following file naming for electronic submission:
DataAnalytics_A5_YOURFIRSTNAME_YOURLASTNAME.xxx

Late submission policy: first time – 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: Patterns, trends, relations: model development and evaluation of housing - NYC Citywide Annualized Calendar Sale Update datasets - available:

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 the code scripts and plots you generate for the questions below.

1. Create a derived dataset containing data points from only **one** of the five boroughs of NYC:

a). Describe the type of patterns or trends you might look for and how you plan to explore and model them. Min. 3 sentences (0%) ;-)

b). Perform exploratory data analysis (variable distributions, etc.) and describe what you did including plots and other descriptions. Identify the outlier values in the data for Sale Price and generate suitable plots to demonstrate the outliers relative to the other data points. Min. 5 sentences (2%)

c). Conduct Multivariate Regression on the **1 borough dataset** to predict the Sale Price using other variables that may have a meaningful connection to price. After you identify a well-performing model test it on **2** subsets of **1 borough dataset** (based on meaningful criteria of your choice, such as building class or sq footage value) and compare the results you obtained. You may have to try multiple models and drop variables with very low significance. Explain the results. Min. 5 sentences (2%)

d). Pick more than one supervised learning model (these need not be restricted to the models you've learned so far), e.g., Naïve Bayes, k-NN, Random Forest, SVM to explore a classification problem using the data. You may choose which categorical variable (e.g. neighborhood, building class) to use as class label. Evaluate the results (contingency tables & metrics). Describe any cleaning you had to do and why. Min. 5 sentences (2%)

2. For the entire (**5 boroughs**) dataset:

a). Apply the best performing regression model(s) from **1.c** to predict Sale Price based on the variables you chose. Plot the predictions and residuals. Explain how well (or not) the models generalize to the whole dataset and speculate as to the reason. Min. 3-4 sentences (4000-level 5%, 6000-level 3%)

b). Apply the classification model(s) from **1.d** to predict the categorical variable of your choice. Evaluate the results (contingency tables & metrics). Explain how well (or not) the models generalize to the whole dataset and speculate as to the reason. Min. 3-4 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 (0%) ;-)

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. 6-7 sentences.