Assignment 5: Data Analytics (Fall 2025) (15% written)

Due: March 28th, 2025 (by 10:00pm ET) Submission method: email (eleisa2@rpi.edu) or LMS

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 dataset - available:

https://rpi.box.com/s/cwgjifo8he6iacqdp1t5yodmp3oarei4

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:

NOTE: There are rows where the BOROUGH column contains the name (e.g. "Manhattan") and others with the borough number (e.g. "1"). Just choose one of those, no need to combine them.

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 regression analysis on the **1 borough dataset** to predict the Sale Price using other variables. After you find a well-performing model test it on a subset of the **1 borough dataset** based on any criterion of your choice (e.g. neighborhood). You may have to try

multiple models and drop variables with very low significance. Explain the results. Describe any cleaning you had to do and why. Min. 5 sentences (2%)

d). Using the same **1 borough dataset**, train and evaluate 3 supervised learning models e.g., Naïve Bayes, k-NN, Random Forest to predict the neighborhood based on the quantitative variables (price, area). Evaluate the results using contingency tables & precision/recall metrics. Describe any cleaning you had to do and why. Min. 5 sentences (2%)

2. Create another derived dataset for a different borough from the one in question 1.

a). Apply the best performing regression model(s) from **1.c** to the new dataset 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 new 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 neighborhood in the new dataset. Evaluate the results (contingency tables & metrics). Explain how well (or not) the models generalize to the new 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. 5 sentences.