Assignment 5: Data Analytics (Spring 2026) (15% written)

Due: March 31st, 2026 (by 10:00pm ET) Submission method: 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

Assignment Topic: Regression & classification analysis with housing data.

NYC Citywide Annualized Calendar Sale Update dataset:

https://rpi.box.com/s/om9gvmtfpssleoih40uo4cf3q0g7k6z2

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. Regression

- Create a subset 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). 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. Examine at least four variables, preferably the ones you will use as inputs to models in the next sections. **(4000-level: 3%, 6000-level: 2%)**

b). Conduct regression analysis on the **1 borough dataset** to predict the *Sale Price* using other variables (e.g. land square feet, number of units). You should try multiple models with different combinations of features and report on the best performing model. You may evaluate your models with a simple one round cross validation. Explain the results and describe any cleaning you had to do and why. **Min. 3 sentences (4000-level: 5%, 6000-level: 4%)**

2. Classification

- Take a subset of the **1 borough dataset** keeping only 3-4 **neighborhoods**.

a) Train and evaluate 3 supervised learning models (e.g. kNN, Random Forest) to predict the **neighborhood** based on quantitative variables (e.g. price, square feet, number of units). Evaluate the results using contingency tables & precision/recall/F1 metrics. You may evaluate your models with a simple one round cross validation. Explain the results and describe any cleaning you had to do and why. **Min. 3 sentences (4000-level: 5%, 6000-level: 4%)**

3. Conclusions

- Draw conclusions from this study about the quality of the dataset and the suitability/deficiencies of the models you tested. Describe what worked, what did not, and why. **(4000-level: 2%, 6000-level: 5%)**

EOF