Advanced Data Analytics (ADA) W24

Assignment 1

Due on Jan 26, 11:59pm, 2024

1. Submission Requirements

This assignment is an individual assignment. Your task is to analyze house sales data released by the New York government and build a regression model that could predict house sales price.

You need to package your submission in one zipped file, named as "ADA-1234-Assn1.zip", where 123 stands for the last 4 digits of your student ID. The zipped folder contains:

1) Only one Python Jupyter Notebook file (.ipynb) containing the reproduceable analysis process and results. The notebook should be self-explained, i.e., it is clear for Tas to judge what you did and why you did so. The notebook should answer all the raised questions in section 3 and 4.
2) Supporting documents to ensure your notebook file can be executed on TA's laptop, such as an preprocessed csv file that your .ipynb file needs to take as input.
3) A reference file containing all conversations between you and ChatGPT as a record to avoid academic integrity issue (if we detect highly similar code between you and another student). You can copy paste the links to conversation.

2. Background

The target dataset is published at: https://www.nyc.gov/site/finance/property/property-rolling-sales-data.page
We are interested in analyzing 3 years' New York City Sales Data (2021,2022, 2023) by borough. You can find past years' dataset at: https://www.nyc.gov/site/finance/property/property-annualized-sales-update.page.
We also provide a data.zip file containing all required raw dataset.

3. Data Exploration and Hypothesis Test (50 points)

Q1 (10 points) Create one csv file that contains all house sales records from five boroughs over the target years (this may involve some manual or automated reformatting work). Perform statistical data exploration to report the 1) statistics of house price for each borough), 2) types of houses involved in the given dataset, 3) missing values in the provided dataset.

Q2 (10 points) Suppose you want to pick two boroughs and analyze the recovery of house price in NYC real estate market post-COVID 19. Propose two questions under this topic and answer them by performing hypothesis tests. Briefly describe why you believe your questions are well-motivated and how you make decisions on which test to be performed. Ensure your answer to the above question is presented in the text cell in your notebook file before your code and results.

Q3 (10 points) Suppose you want to compare the house prices and house sales for three types of houses in terms of BUILDING CLASS CATEGORY, choose proper data visualization methods to draw plots and report your findings in the text cell in your notebook file after your code and plots.

4. Feature Engineering and Regression Analysis (50 points)

   Q4 (10 points) Based on your findings and exploration from previous questions, raise a regression modeling task targeting house sale price prediction. You are free to choose the portion of data you want to use for the prediction task, but you should explain why you make such decision. Ensure your answer to the above question is presented in the text cell in your notebook file before your code.

   Q5 (20 points) Analyze the raw features and decide the lists of features to be included for the proposed regression task in Q4. Detect if multicollinearity exists in selected features. Explain how you perform feature engineering in the text cell before your code.

   Q6 (20 points) Build one regression model based on your previous analysis, let us call it model A. Compared the prediction power of model A (80% training, 20% test) with another model B, which is made of raw features (without applying feature engineering raised in Q5). Note that, you should use the same set of training data/testing data, and report the performance on both train and test set.