MSC 550 Intro Analytics & Programming

Midterm Exam

Due by 10/8/2022

The primary objective of this housing.csv dataset is to predict the housing price based on certain factors like house area, bedrooms, furnished, nearness to main road, etc. Variable description is provided below:

- **area** = area of a house
- **bedrooms** = number of house bedrooms
- **bathrooms** = number of house bathrooms
- **stories** = number of house stories
- ***mainroad*** = whether connected to the main road
- ***guestroom*** = whether has a guestroom
- ***basement*** = whether has a basement
- ***hotwaterheating*** = whether has hot water heater
- ***airconditioning*** = whether has an airconditioning
- ***parking*** = the number of parking
- ***price*** = price of the house

You are expected to build the "best" multiple linear regression model using housing_train.csv by 1. addressing the potential violations of the error term assumptions, 2. removing the unnecessary predictors, and 3. dealing with the potential outliers. Using the "best" model that you built, answer the following questions:

1. Do guestroom, basement, hot water heater, and air conditioning significantly influence the housing price (significance), and how (magnitude)?
2. Make prediction on the test set *housing_test.csv*, and report the mean square error (MSE).

Please submit your midterm on canvas by the end of next Sunday, Oct. 8. You should submit both an analysis report (in .doc or .pdf file) and your code (in .py or .ipynb file). **Please present all the results that you get from your analysis in the analysis report (including diagnostic plots, fitted model, etc.), and explicitly discuss what you are doing for each step.** Please let me know if you have any questions.