Final Project

Final Project Policies

Grading Policy

Regardless of your grading basis, in order to pass the course, you must achieve **a score of at least 50% (at least 13/25)** on the final project.

Independent Work Policy

You **may not** copy someone else's code or write-up. You **may not** enable someone to copy your work by sharing your code or write-up. Any submitted project that is deemed by the instructor to be in violation of the independent work policy will **receive a score of 0**. Since a passing score on the project is necessary for passing the class, anyone deemed to be in violation of the policy will automatically fail the class.

Submission Policy

Due data can be found in syllabus's course schedule. **No late submission will be accepted**.

Project Description

Introduction

This project will give you an opportunity to apply many of the data analytical techniques covered in class. It will also allow you to indulge in working with a comprehensive data set collected from a real business context. In this project, you will have a chance to showcase your analytical techniques as well as decision-making skills on a new business problem. As the outcome of this project, you will make business decisions that are informed by data analyses.

Major Objective

You are an investment consultant. An important client of yours wants to invest in real estate in a U.S. city that has been experiencing an uptick in growth. The client wants to better understand the sales prices of the real estate properties. Your main job is to help her explain and eventually predict the sales prices of some properties as close to reality as possible.

The Data

The broker of the properties has given you the relevant information of 150 houses sold in the city over the past 2 years. These data are contained in the "house.csv" data set, which is posted on the course website. The data collected on the houses include:

1. PID – the property ID, which is the unique IDs of the real estate properties
2. House Size – the size of the house in square feet
3. Lot Size– the size of the lot in acres
4. Rooms – the total number of rooms in the house

5. Bathrooms –the number of bathrooms in the house
6. Utilities – this describes the average monthly utility cost (in $)
7. Year Built – this describes the year when the house was initially constructed
8. Overall Condition – this is a rating of the overall condition of the house. The numeric scale of this rating is as follows:

   1 – Very Poor 2 – Poor 3 – Fair 4 – Average 5 – Good 6 – Excellent 7 – Very Excellent

9. Brick – whether or not the house is made primarily of brick.
10. Price (in $) – this is the sales price of the houses

Main Task Overview and Data Preparation

First, randomly select a subset of the 150 houses. This random sample should contain 130 houses. Create a data frame of this sample and call it "training.houses" (i.e., the training set). Create another data frame of the rest of the houses (n = 20) and call it "testing.houses" (i.e., the testing set). Make sure you save and export your training and testing sets at the very beginning of your work. Both sets need to be submitted along with your final report. You can read "*A conceptual note on the project procedure*" in the *appendix* for more information.

Second, use multiple regression to explain house sales price. Your goal is to find the most fit regression model based on the data in the training set. Interpret the results associated with your model.

Lastly, you need to make prediction using the data in the testing set. Based on your prediction, you need to make the decision on "which 10 properties ranked highest in sales price". Then validate your prediction using the real price values in the testing set. Is your prediction correct (and/or to what extent)?

Specific Requirements in the Project Report

Your end-product for the project will be an R Markdown report. An R Markdown file with a brief outline of the report is provided to you. You need to input code chunks (and R codes), descriptions, and interpretations/discussion in the R Markdown file.

Insert code chunks in your report where needed (you can add as many code chunks as you need to address the project problems). You can also add titles and/or subtitles, where appropriate, to make your report more readable.

In your report, you need to address the following problems:

[Using the training set]:

Data Exploration

1. Is there a significant difference in sales prices between brick and non-brick houses? Also use a side-by-side boxplot to present the prices of the two types of houses.
2. Calculate the *age of the houses* in 2018. In addition, present a histogram of the house age variable.

Regression Modeling and Interpretations

3. Fit a regression model to explain house sales price.
4. Interpret the regression results. You should follow the 5 steps covered in class to analyze the results. The 5th step – prediction – is conducted in the next question.

[Using the testing set]:

Prediction and Validation

5. Based on your regression model, predict which 10 properties in the testing set ranked highest in sales price. Clearly state your answer in the report.
6. Validate your prediction using the real price values in the testing set. Is your prediction correct (and/or to what extent)? Provide a discussion/conclusion of your analysis.

Note 1: Figures, tables, test results that are presented without accompanying description/discussion will receive at most half credit. To earn full credit, you must describe what each table/figure/result is showing and discuss any key takeaways. In other words, it is not sufficient to simply display R output. You must also provide thoughtful discussion of the output in the project context.

Note 2: In your discussion of the results, you must use in-line code chunks to assess statistical significance, that is, to report p-values and regression coefficients. Points will be deducted if you do not use in-line code chunks.

Submitting Files

Files to be submitted:

- The **Rmd** file that generates your analysis
- The resulting **html** file produced by knitting
- The **training set** (a csv file) that you generated
- The **testing set** (a csv file) that you generated

It is impossible for me to evaluate your report if I don’t have your training and testing sets. **Therefore, project will be not graded unless all the files above are submitted.**

Appendix

I. A conceptual note on the project procedure (also see Unit 9)

A fancy name of what we do in this project is called "***data mining***", which is essentially a statistical method that involves validating analytical models against real data. The procedure of data mining starts with split the original data set into two subsets: one subset for estimation (called the ***training set***) and one subset for validation (called the ***testing set***). A regression equation is estimated from the first subset. Then the values of explanatory variables from the second subset are substituted into this equation to obtain predicted values for the dependent variable. Finally, these predicted values are compared to the *known* values of the dependent variable in the second subset. If the agreement is good, there is reason to believe that the regression equation will predict well for new data.

II. Grading Rubric

Components	Criteria		
	Unsatisfactory	Marginal	Satisfactory
Data Preparation (15%)	The submission did not include explanations on the tests/procedures performed to address the project problems. R codes were incorrectly written. Results were not produced by the codes. The submission did not explain what results were obtained. Interpretation and discussion were unsatisfactorily provided or were entirely ignored. The overall report was difficult to read and understand. No HTML file was submitted.	The submission included some but not complete explanations on the tests/procedures performed to address the project problems to the extent that they left readers wonder what exactly were done. Or the explanations were not stated in a clear fashion. Either or both R codes and result output involve issues. The submission provided interpretations of the results, but focused on the incorrect R output. The discussion and conclusion were provided but was not satisfactory. In-line code chunks were shown incorrectly in the HTML file or were entirely missed. The overall report was acceptable but not satisfactory.	The submission clearly explained the tests/procedures performed to address the project problems. R codes were correctly written to run analyses. And the analyses were done according to the requirements covered in this course (e.g., 5 steps of interpreting regression results). The results were correctly interpreted and clearly reported. Assessment of statistical significance (e.g., p-values and coefficients) were provided in discussion using in-line code chunks. The overall report was well-formatted and clearly-presented and easy to read.
Data Exploration (20%)			
Regression Model (15%)			
Model Results Interpretations (20%)			
Prediction (15%)			
Validation and Conclusion (10%)			
Overall R Markdown format & report readability (5%)			