Final Project

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 and The Data

You are a business consultant. An important client of yours has invested in a couple of fine dining restaurants in a U.S. city. The client wants to better understand the performance of the fine diners. She has gathered data on 200 restaurants in the city. The data are contained in the "restaurants.csv" data file, which is posted on the course website. The variables include:

1. ID – the restaurant ID, which is the unique IDs of the restaurants
2. Advertising – This is the yearly advertising expenditure (in thousands of $ dollars)
3. Adsplay – is the number of commercials play per week on webpages (e.g., Yelp, YouTube)
4. Year – this describes the year when the restaurant started its business
5. Days – the number of open days per week
6. Price _ this is the average spend per person (in $)
7. Parking _ whether the restaurant has valet parking (coded as 1) or not (coded as 0).
8. Rating – this is a rating of the overall service, a scale from 1 (very poor) to 10 (very good).
9. Cuisine – whether the food is marketed as traditional (coded as 0) or creative (coded as 1)
10. Table.Turns – this measures the number of times tables are turned per week. This indicates the volume of customers and is a key measure of success/revenues in dining business.

Your client would like you to help her interpret and eventually predict the table turns in the city as close to reality as possible. She also wants to know whether the data can help her determine what can be done to improve table turns.

Task Overview and Data Preparation

First, randomly select a subset of the 200 restaurants. This random sample should contain 170 restaurants. Create a data frame of this sample and call it "rt.training" (i.e., the training set). Create another data frame of the rest of the restaurants (n = 30) and call it "rt.testing" (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 table turns. 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 restaurants ranked highest in table turns". Then validate your prediction using the real table turns values in the testing set. Discuss your predictions (i.e., answer the specific questions below; see "Specific Requirements in the Project Report").

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 questions:

[Using the training set]:

Data Exploration

1. In the sample, how many restaurants market their food as traditional? And how many market their food as creative? Is there a significant difference in table turns between restaurants whose food is marketed as traditional versus creative? Also use a side-by-side boxplot to present the table turns of the two types of restaurants.
2. Create a new variable "age" in the restaurants data set, which is the *age* of the restaurants in 2019. Report the mean, standard deviation, maximum, and minimum of the age variable. In addition, present a histogram of the age variable.

Regression Modeling and Interpretations

3. Fit a regression model to explain restaurant table turns.
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. When interpreting the results, make sure you address your client's concern regarding what can be done to improve table turns. For example, your client is interested in knowing if advertising is worth it. What would you recommend your client to do?

[Using the testing set]:

Prediction and Validation

5. Based on your regression results, predict which 10 restaurants in the testing set ranked highest in table turns. Clearly state your answer in the report (i.e., the 10 restaurant IDs).
6. Validate your prediction using the real table turns values in the testing set: Is your prediction correct? That is, are the 10 predicted restaurants really ranked highest? How many predicted top 10s are actually in the top 10s?

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 (25%)			
Regression Model (15%)			
Model Results Interpretations (20%)			
Prediction (10%)			
Validation and Conclusion (10%)			
Overall R Markdown format & report readability (5%)			