GEORGETOWN UNIVERSITY / McDONOUGH SCHOOL of BUSINESS

Final Project – Fall 2024

This is a group project for the following two courses:

- **OPAN 6601 (Statistics for Business Analytics)**
- **OPAN 6606 (Programming I – Introduction to Data)**

Project Description and Guidelines

Airbnb is a company that provides an online marketplace for short-term rentals of homes and apartments. Much of the data from Airbnb's website has been compiled and made publicly available on the website Inside Airbnb. For this assignment, you will analyze a sample of the Airbnb listings from some of the neighborhoods in Washington, DC. Each row in this dataset represents a single Airbnb listing.

The main goal of the assignment is to use your R, statistics, and data analysis skills in analyzing this dataset. In the real world, you would, of course, have very substantial latitude in how to undertake the analysis and present the results. For the sake of setting expectations for this assignment, though, you will need to do the following:

1. Report your exploratory analysis of the data. This can include data visualization, summary tables, changes made to the data, or any other insightful findings about the data.

2. Which combination of neighborhood and room type has the highest average price? Which one has the lowest? Which combination has the highest variability? Which combination has the lowest?

3. Write an R function that takes three arguments: a confidence level, the name of a numerical variable, and a data frame. The function should return the confidence interval for the average value of the specified variable at the given confidence level. Apply this function to compute a 95% confidence interval for the average price of the listings in the dataset. Provide an interpretation of the computed confidence interval in the context of the Airbnb listings.

4. Test whether the average price of all listings in the population is more than $200 (at 95% level of confidence). What is the statistical conclusion based on this result? Is this conclusion in line with the estimated confidence interval reported in question 3?

5. Visualize price to test for normality and comment on the results (diagnostics plots generated for this question will be counted as only 1 plot in your report - see below for details).

6. What's the best simple linear regression model for "price" of the listings based on R-squared and residual standard error? Compare/present the results of the tested models as a table in your report. (Note: You should not manually create the comparison table in your submitted report. Your code should generate the table.)

7. Implement a multiple linear regression model for the "price" of the listings. Report the regression coefficients and measures of fit, and write an interpretation of the regression coefficients in the context of this model. Are there any violations of model assumptions?

8. Are there any multicollinearity concerns among the independent variables selected for the multiple regression model of question 7? Explain.

9. **Summary/Recommendation Section**: Write a summary/recommendation section based on your findings. In 1-2 paragraphs, provide an investment recommendation or market analysis summary for Airbnb in Washington, DC, based on the data you analyzed. Use insights from your analysis to justify your recommendation. (Note: You should start your submitted report with this summary.)

When writing your R code for addressing the questions and analyzing the dataset, consider the following:

- Combine the two data sets ("Listings.csv" and "Reviews.csv") and use the complete dataset for analysis. The complete dataset used for the analysis should not have any missing values.
- In your final report, include at least three but no more than five figures. Each figure may have multiple panels. Regression output plots are not counted toward the total number of plots).
- Have at least two chunks of code written in the tidyverse package's piping form. This excludes using the piping form for ggplot.
- Include question numbers and relevant (but not excessive) annotations in your R code.

Deliverables:

You should submit your **well-commented** and **fully functional** R code and a PDF report answering project questions and displaying and explaining the result of your analysis. You should submit these two files to both your Statistics and Programming class Canvas pages. Each group should make only one submission for each class.

- **A single R code/file**: This is a file where one can see how you've done the work you explained in the report. The R file should contain every step, from loading the data to producing the report. (no need to submit the Google Colab / notebook version).
- **A written report (in PDF)**: This is the written report where you share your analysis and answer the questions. Include group member names on the first page, start the report with the summary of your findings (as noted Question 9), and report any data pre-processing steps and assumptions you have made in analyzing the data. Include question numbers in your report. No R code should be shown in this report. If you plan to use R output in your report, it should be clean and formatted properly. Describe each table, plot, or output in your report briefly but clearly. You can choose your own overall format, but do not use a font size larger than 12 and use at least 1.5 or double line spacing. You can choose your own overall format, but you should use a font size between 10 and 12 and use at least 1.5 or double line spacing. Limit the total pages of your PDF document to **8 page,** including the title page.

Evaluation:

Each student's grade will be made up of three different components: **the code (40%), the written report (40%), and peer evaluation (20%)** based on the following framework.

- **Coding part (40%):** The functionality, cleanness, quality and efficiency of your code (including annotations), clarity/visual appeal of the figures and generated outputs, whether you followed the instructions, and whether you included all the coding tasks mentioned above account for 40% of your grade. Your Programming I instructor determines this grade. Each student in the team is expected to get the same grade for this part.
- **Statistics and data analysis (40%):** The quality and depth of your analysis, clarity of the explanation, format of the report, and whether you answered all the questions asked in the analysis section. Your Statistics for Business Analytics instructor determines this grade. Each student in the team is expected to get the same grade for this part.

- **Peer evaluation (20%):** The feedback of your group members about your performance on the team accounts for 20% of your grade. If you fail to submit the peer evaluation form, you will automatically lose the feedback score.

The instructors reserve the right to adjust individual grades and the grading distribution based on feedback from team members. In a circumstance that a student does not actively participate in team's activities and does not contribute to the team deliverables, the instructors reserves the right to assign 0 for the individual's total project grade.