Mini Project 1 Assignment

Due date: Tuesday, September 3 at 11:59pm

Objectives

In this project, you will demonstrate your ability to:

- Load a data set into R.
- Calculate descriptive statistics and plots for numeric and categorical/binary variables.
- Draw appropriate conclusions based on statistics and plots.
- Organize an exploratory analysis into a readable report.
- Create a report using R Markdown.

Data

The data for this project is the Pierce County House Sales data set, found in Canvas in the Mini Project 1 module. This data set represents home sales in 2020 in Pierce County, Washington. Please visit this site for variable names and definition:

https://www.openintro.us/data/index.php?data=pierce_county_house_sales

Assignment

Form a group of 3-4 with your classmates. Before beginning the project, make sure that you discuss expectations about frequency of communication, timeline for completing the work, and expectations of group members. Write a short report that describes a typical home sold in this area in 2020.

First, subset the data to include only listings whose sale price is less than $1,000,000. Choose 4 variables to analyze in your project. For each variable you choose, please do the following:

- Create an appropriate plot or table to display the distribution of the variable.
- Calculate appropriate summary statistics for each variable
- Write 2-3 sentences to summarize the shape, center, and variability of the distribution (for numeric variables), the most frequent values (for binary/categorical variables), and any usual features of the data such as outliers, missing values, categories with low counts, or any apparent data errors.

Next, look at the houses whose sale price is greater than $1,000,000. Repeat the previous steps for this subset of the data. Write a few sentences describing the similarities and differences between the houses in the two groups.

Your analysis should be presented as a short report, with tables and plots embedded in the report and formatted to be appropriately sized and readable. Make sure to write clearly. It is fine to write in a conversational tone. Create the report in an R Markdown file (.Rmd) and knit to an html file to submit it. Please use the echo=FALSE option when creating your code chunks. Please do not include unnecessary computer output (e.g. printing the data).

Other policies

- If you use any outside resources, make sure to include an appropriate citation. For websites or code resources, include as many of the following pieces of information that are available: url, name of author, date accessed, name of webpage.
- Use of generative AI such as ChatGPT is strongly discouraged for this project. Under no circumstances may you upload the data set into an AI tool.
- Under no circumstances may you paste output from ChatGPT and present it as your work.
- After project grades are posted, students may request a chance to resubmit a corrected project. Re-submissions will be re-graded using the original rubric, subject to a 20% deduction.

Rubric

Your project will be graded according to this rubric. Make sure to follow the specific guidelines about the required plots, statistics, and discussion elements.

Description	Points possible
Presentation style: the results are presented clearly in a readable, narrative form. The writing is clear and concise. All quantitative results, including plots, tables, and summary statistics are referenced and interpreted in the text. Captions, graph titles, and axis labels are included where appropriate.	2
Report is submitted in .html form and created from R Markdown.	1
The data are subsetted correctly.	1
Appropriate tables or plots are created for 4 variables. At least 1 variable is numeric and at least 1 variable is categorical.	2
Appropriate summary statistics are calculated for 4 variables. At least 1 variable is numeric and at least 1 variable is categorical. Robust measures of center and spread are used when appropriate.	2

Accurate conclusions are drawn and the conclusions are communicated clearly. The text includes an exploration of outliers and missing values and, if applicable, a discussion of their potential impact on overall data quality.	2