Homework for Lecture 9: Loading Data in R & EDA

The following homework assignment should be completed as an RMarkdown document. Only show your working code in the code chunks, and place comments/explanations in the "white space" to help me evaluate your work. Please submit both the knitted RMarkdown document (`.doc`) and the `.Rmd` file. You may discuss the assignment with your classmates, but your write-up and work should be done independently of your classmates. The use of AI is considered a violation of academic policy for this assignment. When in doubt, ask me what is and is not allowed.

Answer the following questions using the `housing_data` dataset in the `MSMU` library. Make sure to include headers in your RMarkdown document to separate each section. When you run code, explain what you are doing and what the output is telling us using complete sentences in the written "white space" of the document.

1. **Load and Inspect the Data**
   (a) Load the `MSMU` library and display the first few rows of the `housing_data` dataset.
   (b) Use `str()` to look at the structure of the dataset.
   (c) Which variables are quantitative? Which variables are categorical or could reasonably be treated as categorical?

2. **Summarizing Bedrooms and Bathrooms**
   (a) What is the distribution of bedrooms? Create a table and visualize it with a barplot.
   (b) Do the same for the distribution of bathrooms.
   (c) Write 2–3 sentences explaining what these summaries tell us about the dataset.

3. **Visualizing Price and Area**
   (a) Create a histogram of house prices.
   (b) Create a histogram of house area.
   (c) Write 2–3 sentences explaining what the histograms help us see about the distribution of price and area.

4. **Using Tables and Logical Vectors**
   (a) How many houses have exactly 3 bedrooms and 4 bathrooms?
   (b) Create a two-way table showing all combinations of bedrooms and bathrooms.
   (c) What does the two-way table tell us that separate bedroom and bathroom tables do not?

**** MORE PROBLEMS ON PAGE 2 ****

5. **Using `aggregate()` for Price and Area**

   (a) What is the average price of a house?

   (b) What is the average price of a house by the number of bedrooms it has?

   (c) What is the average area of a house?

   (d) What is the average area of a house by the number of bedrooms it has?

   (e) Write 2–3 sentences explaining what you learned from these averages.

6. **Creating a New Categorical Variable**

   (a) Create a new variable called `area_cat` that classifies each house as:

   - `small` if the area is from 1,000 to less than 2,000 square feet,
   - `regular` if the area is from 2,000 to less than 3,500 square feet,
   - `large` if the area is 3,500 square feet or above.

   (b) Use `table()` to determine how many houses fall into each area category.

   (c) What is the average price of a house grouped by both number of bedrooms and `area_cat`?

   (d) Write 3–5 sentences explaining how area category and number of bedrooms seem to relate to house price.

7. **Final Data Story**

   Write a short paragraph summarizing what you learned from this dataset. Your paragraph should include:

   - one result from a table,
   - one result from a visualization,
   - one result from `aggregate()`,
   - and one sentence explaining why grouped summaries are useful in exploratory data analysis.