
Homework for Lecture 8: Exploratory Data Analysis

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 `babies` dataset in the `openintro` library. If this is your first time using the `openintro` library, you may need to install the package before loading it. Make sure to include headers in your RMarkdown document to separate each step. Explain your code and output in complete sentences.

1. Step 1: What is the Data?

Use the documentation to learn about the `babies` dataset.

- (a) What code can we use to load the `openintro` library?
- (b) What code can we use to view the documentation for the `babies` dataset?
- (c) In your own words, what does this dataset represent?
- (d) What are two variables in the dataset that you think may be useful to investigate?

2. Step 2: Logical vs. Physical

Compare the documentation to the actual dataset in R.

- (a) Use `str()` to display the physical structure of the dataset.
- (b) Does the documentation seem to match the dataset in R? Explain your answer.
- (c) How many observations are in the dataset?
- (d) How many variables are in the dataset?

3. Step 3: Data Conditions

Investigate whether the data have missing, special, or unusual values.

- (a) Are there any actual missing values coded as `NA` in the dataset?
- (b) How many complete cases are there?
- (c) What proportion of cases contain at least one missing value?
- (d) Use `sort()` and `unique()` on at least two variables to look for unusual values. Do you notice anything that may need to be investigated?

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

4. Step 4: Data Preparation

Prepare the data so that it is easier to summarize and interpret.

- (a) Identify at least one column that could be better represented as a factor. Create a new factor version of that column.
- (b) Find the mean height of the mothers. If there are missing or unusual values, make sure your calculation handles them correctly.
- (c) Create a new variable that breaks birth weight into 4 categories. You may choose your own break points, but explain why you chose them.
- (d) Use `table()` to check how many observations fall into each birth weight category.

5. Step 5: Initial Summary

Use summary tools to learn more about the dataset.

- (a) Display a summary of the full dataset.
- (b) Display a summary of your new birth weight category variable.
- (c) Write 3–5 sentences explaining what you learned from the summaries. Mention at least one thing about the center, spread, categories, missing values, or unusual values.

6. Step 6: Data Visualization

Create visualizations to better understand the data.

- (a) Create a visualization for one quantitative variable.
- (b) Create a visualization for one qualitative variable.
- (c) Create one additional visualization of your choice.
- (d) For each visualization, write 1–2 sentences explaining what the graph helps us see.

7. Final EDA Reflection

Write a short paragraph answering the following questions:

- What is one thing you learned about the `babies` dataset?
- What is one possible issue, limitation, or unusual feature of the dataset?
- Why is it important to complete exploratory data analysis before making conclusions from a dataset?