Lecture 8: Exploratory Data Analysis

For each question below, write the R code you would use and record the result when appropriate. Answer in complete sentences when asked to explain.

Part 1: EDA Goal One with `football`

1. Load the `MSMU` package and access the `football` dataset.

   (a) Install the package and load it into R. You only need to put the library call in the first R chunk of the RMarkdown document.

   (b) Display the first 6 observations of the dataset.

2. Before beginning EDA, narrow the dataset to a smaller group of variables by creating a new dataframe called `football_small` containing only the following columns:

   `Player, Tm, Age, Wins, Yds, TD, Int, Rate, QBR`

   (a) Write the code to create `football_small`.

   (b) Display the first 6 observations of `football_small`.

3. **Step 1: What is the Data?**

   (a) According to the documentation, what does this dataset represent?

   (b) According to the documentation, what does one row represent?

4. **Step 2: Logical vs Physical**

   (a) Display the structure of `football_small` using `str()`.

   (b) How many observations and how many variables are present?

   (c) What data types appear in the dataframe?

   (d) Does the documentation appear to match the actual dataset? Explain briefly.

5. **Step 3: Data Conditions**

   (a) Determine how many missing values are present in each column of `football_small`.

   (b) Determine how many missing values are present in the `Wins` column.

   (c) Determine how many missing values are present in the `QBR` column.

   (d) Use `which()` to determine where the missing values occur in the `QBR` column.

   (e) Display all observations where `Wins` is missing.

6. Continue **Step 3: Data Conditions**

   (a) Display the unique team abbreviations in sorted order.

   (b) Determine the range of `QBR`. You should get a value, not NA.

   (c) Determine the mean of `QBR`. You should get a value, not NA.

7. Still in **Step 3: Data Conditions**, use `complete.cases()`.

   (a) Determine how many complete observations are in `football_small`.

   (b) Determine how many observations contain at least one missing value.

   (c) Determine the proportion of complete observations.

   (d) Display the first 6 complete observations.

8. **Step 4: Data Preparation** Create a copy of the data called `football2`.

   (a) Write the code to create `football2`.

(b) Use the `cut()` function to create a new column called `Age_group` based on the `Age` variable using the breaks

`20, 25, 30, 35, 50`

and the labels

`"21-25", "26-30", "31-35", "36+"`

(c) Display a table of `Age_group`.

(d) In one sentence, explain why creating `Age_group` might be useful.

9. Continue **Step 4: Data Preparation**

(a) Create a new column called `Tm_f` by converting `Tm` to a factor.

(b) Display the structure of `football2`.

(c) In one sentence, explain why treating `Tm` as categorical may be helpful.

10. **Step 5: Initial Summary**

(a) Display a summary of `football2`.

(b) What are two things you learn from the summary?

11. **Step 6: Basic Visualizations**

(a) Create a histogram of the `Age` variable.

(b) Create a histogram of the `QBR` variable, making sure to handle missing values appropriately.

(c) Create a barplot of the `Age_group` variable.

(d) Create a barplot of the `Tm_f` variable.

(e) In one or two sentences, describe one thing you notice from one histogram and one thing you notice from one barplot.

Part 2: Reflection

12. In one or two sentences, explain why it is important to carry out EDA before jumping into deeper analysis of a dataset.