Lecture 9: More EDA

For each question below, write the R code you would use and record the result when appropriate. I encourage you to complete this in an RMarkdown document so that you will have the code/explanations for your personal records. Answer in complete sentences when asked to explain.

Part 1: EDA with `mlb_df.csv`

1. Download the `mlb_df.csv` file from Canvas and load it into R.

   (a) Write the code using `read.csv()` to load the dataset and save it as `mlb_df`.

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

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

   (a) Based on the variable names, what does this dataset appear to represent?
   (It is ok if you are unsure, the `mlb_eda` dataset in the `MSMU` library contains the documentation for this example)

   (b) What does one row of the dataset appear to represent?

   (c) Which variables seem to describe player identity, and which seem to describe player performance?

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

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

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

   (c) Which variables appear to be quantitative?

   (d) Which variables appear to be categorical?

4. **Step 3: Data Conditions**

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

   (b) Display all observations where `games` is missing.

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

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

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

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

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

   (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.

6. **Step 4: Data Preparation** Create a copy of the data called `mlb_df2`.

   (a) Write the code to create `mlb_df2`. This way if we make a mistake we do not overwrite our original dataset.

   (b) Create a new column called `team_f` by converting `team` to a factor.

   (c) Create a new column called `position_f` by converting `position` to a factor.

   (d) Display the structure of `mlb_df2`. What do you notice?

7. Continue **Step 4: Data Preparation** using `cut()`.

   (a) Use the `cut()` function to create a new column called `HR_group` based on the `HR`:

   Using the breaks: `0, 10, 20, 30, 40, 60`
   and the labels: `"0-9", "10-19", "20-29", "30-39", "40+"`

   (b) Display a table of `HR_group`.

   (c) In one sentence, explain what the `cut()` function gave you here.

8. **Step 5: Initial Summary**

   (a) Display a summary of `mlb_df2`. What can we learn from the summary?

9. **Step 6: Basic Visualizations**

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

   (b) Create a barplot of the `position_f` variable.

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

10. **Step 7: Using `table()`**

    (a) Create a table showing the number of players in each position.

    (b) Create a table showing the number of players in each home run group.

    (c) Create a two-way table showing the number of players in each combination of `position_f` and `HR_group`. How can we interpret this result?

11. **Step 8: Logical Vectors**

    (a) Use a logical operator/vector to determine how many players are pitchers.

    (b) Use a logical operator/vector to determine how many players are catchers on the `KC` team.

    (c) Use logical indexing to find the mean number of home runs for players on the `BOS` team.

    (d) Use logical indexing to find the mean batting average for players whose position is `RF`.

12. **Step 9: Using `aggregate()`**

    (a) Use `aggregate()` to find the mean number of home runs by position.

    (b) Use `aggregate()` to find the mean batting average by team.

    (c) Use `aggregate()` to find the mean OPS by position and team.

    (d) Find the sum of `AB` for players on the `KC` team.

    (e) Find the sum of `H` for players on the `KC` team.

    (f) Compute `H/AB` for the `KC` team using your two totals above.

    (g) Compare this value to the mean `AVG` for `KC`. Why are these two values different?

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