Lecture 9: Handling Missing Data

Problem Set for Lecture 9

These questions will assist in bolstering your understanding of the material in Lecture 9. Emphasis should be placed on having the correct code/output as well as communication. The deliverable should be a knitted RMarkdown document to pdf without any code running off the page. You will be able to present these in office hours as an oral assessment interview along with the problem set for Lecture 8 and Lecture 10.

***** First, make sure you have read the Lecture Material's writeup *****

1. "Clean" the `baseball_teams` dataset within the `MSMU` library to deal with missing values. If you remove observations or alter values, you must explain what you did and why. Include a short write-up as you answer the following questions:

   (a) At the top of the document, include bullet points summarizing what you did to address the missing values. For instance: "I found unusual negative values that appeared to represent missing data, recoded them as `NA`, and then imputed them using the mean because they appeared to be missing at random."

   (b) How many missing values are there in total? What proportion of each column is missing?

   (c) How many complete cases do we have? Can we simply omit all observations with missing values? Explain why or why not.

   (d) Are there any missing values that are not coded as `NA`? For example, check for unusual negative values or impossible outliers. What should be done with them?

   (e) Are there any columns with too many missing values to be useful? If so, what should be done with those columns?

   (f) Is there any noticeable pattern in the missing data? For example, do the same variables tend to be missing together for certain observations?

   (g) Can any missing values be logically determined from other variables already in the dataset? If so, fill them in and explain your reasoning.

   (h) Create a scatterplot with year on the x-axis and win total on the y-axis. Color the teams based on whether or not they made the World Series. Add separate lines of best fit for teams that made the World Series and teams that did not.

   (i) Write the cleaned dataset to a CSV file and verify that row names are not included. Upload this CSV file to Canvas along with your problem set solutions.

2. Using the "Lecture-11-Assessment.csv" dataset on Canvas, carry out the same steps that you did above, except for the visualization.