ALY_6000 Project 2 Report

Junchen Yi

College Professional Studies in Analytics, Northeastern University

ALY_6000: Introduction to Analytics

Zhi He

Submission Date: 02/10/2023

Introduction

For this project, I practiced using two different datasets. In the first part, I analyzed the country's happiness and freedom index. In the second part, I analyzed various data on players in Major League Baseball. In both of these parts of the assignment, I was asked to learn about filtering, sorting, expanding, and presenting data, which gave me a deeper understanding of data analysis.

Analysis and Key Findings

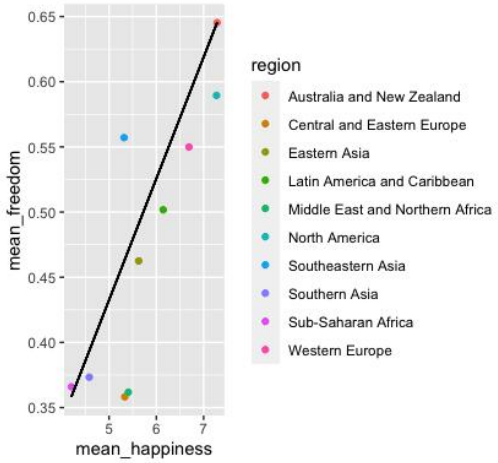
Assignment One:

In this part, I learned some useful function:

1. Use the function ‘names’ to produce the column names for your data set.
2. Use the ‘view’ function to view the data set in a separate tab.
3. Use the ‘glimpse’ function to view data set.
4. Use ‘clean_names’ to normalize the column names (variable names) of data frames to make them easier to use in R.
5. Use ‘mutate’ to create new columns or modify existing columns, or to perform calculations or conversions based on variables in an existing data set.
6. Sorting rows in a dataset with the 'arrange' function.
7. Use ‘summarize’ function to summarize or aggregate data.

In question 13, I was asked to compare the average GDP of two different regions. Firstly, I filter two kinks of countries from dataset data_2015, check that each value of "region" on the left of the symbol appears in the set of values on the right. Then, filter rows with "Region" equal to "Western Europe", and rows with "Region" equal to "Sub-Saharan Africa". Secondly, Ordered data and selected 10 countries for each region. Finally, calculate average GDP and printed the result into a data frame.

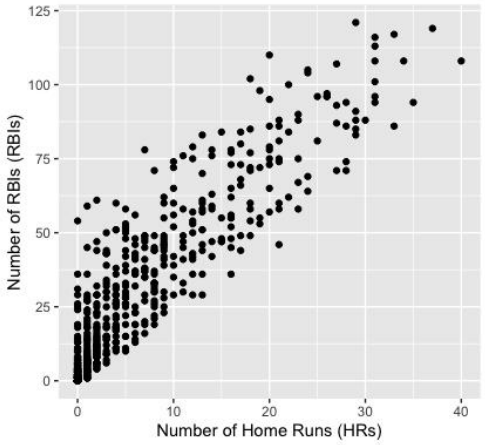
In question 14, I created a scatter plot using the ggplot2 package. The first parameter of the ‘ggplot’ function is the dataset. The 'aes' function specifies the data for the x-axis (mean happiness index) and y-axis (mean degrees of freedom), as well as the color of the points (color = region). In addition, the 'geom_point' function is used to add points to the scatter plot, and the 'geom_segment' function is used to add lines.



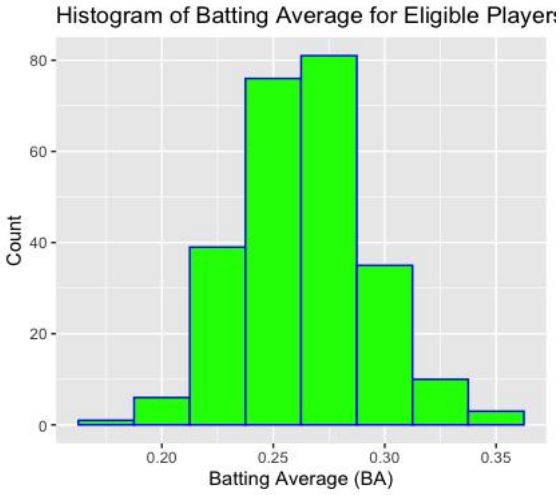
Assignment 2:

This assignment is more about using functions to modify the original dataset by adding columns in order to better analyze the player data and find the most valuable players.

In question 11, the scatter plot of home runs versus RBIs gives an idea of the relationship between a player's home runs and RBI. The visualisation results show that the two are positively correlated, meaning that as HR increases, so does RBI. Also, I've learned that the more HR a player has, the more threatening they are in attack. A player's RBI number usually indicates their ability to drive runners to score. Looking at the images, we can see that most of the players are mediocre, but there are also many very good attackers and team contributors.



In question 13, the result of the code run is the histogram, which was successfully created once in last week's job. Histograms allow us to visually see and analyze the batting average distribution of eligible players, which is valuable for evaluating player performance and making informed decisions related to end-of-season awards. We can see that the batting average of the players is somewhat like a normal distribution, indicating that most of the players' batting average is concentrated between 70 and 80.



In question 19, I think the two players who can compete for MVP are Mattingly and Schmidt. both of them have the same overall ranking. Although Schmidt's HR and RBI rankings are very high, his OBP rankings are very low compared to Mattingly, so Mattingly is more consistent in terms of rankings. Combined with their ages, Mattingly is younger and I think he has more potential. To

summarize, I think Mattingly should be elected MVP.

Reference:

1986 Major League Baseball Standard Batting. Baseball. (n.d.).
https://www.baseball-reference.com/leagues/majors/1986-standard-batting.shtml