Lecture 8: Grouped and Bivariate Analysis

Problem Set for Lecture 8

These questions will assist in bolstering your understanding of the material in Lecture 8. 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 9 and Lecture 10.

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

1. Using the `College_data` dataset within the `MSMU` library, answer the following questions. When you are working these steps out, only display the first 10 observations to show your results using the `head()` function at the end of your pipe chain.

   (a) There is a lot of data present, so only display the columns which are: Name, Region, Acceptance, Enrollment, and the Out of State Tuition. Filter it so that only rows which have more than 500 accepted students are shown. Use this dataset for the rest of the parts.

   (b) Some of the Colleges have the word "Private" at the end of it, implying it is a private college. Create a single new column which indicates the private colleges and the public colleges using `grepl()` and then remove the "Private" indication within the `Name` column. What is the mean Out of State Tuition cost based on each region and private status?

   (c) Determine the `Conversion` rate by calculating the percentage of students which enroll in a college after they are accepted. Then create a new column which breaks the `Conversion` rate into 4 quartiles (where roughly 25% of the data is in each category)

   (d) Determine how many schools are in each `Conversion` category based on their Private/Public status.

   (e) What proportion of schools have higher than a 90% "Conversion" rate? What about a 75% "Conversion" rate? A 50% "Conversion" rate?

   (f) Create a scatter-plot for Accepted and Enrolled based on their "Conversion" Category. Include a line of best fit (linear model with no standard error bars). What does this tell us?

2. Using the `txhousing` dataset in the `ggplot2` library, answer the following questions about the Texas housing market.

   (a) Focus on the variables `city`, `year`, `month`, `sales`, `listings`, `median`, and `inventory`. Restrict the data to observations to only looking at complete cases. Use this dataset for the rest of the parts.

(b) Determine the average number of home sales for each combination of `city` and `year`. Arrange the results from largest to smallest average sales.

(c) Determine the average inventory for each `month`. Based on your summary, during which times of year does inventory tend to be largest?

(d) For each year, identify the city with the highest percentage of listed homes that were sold. (For 2000, Garland is the highest city at 2836/5569 = 0.509)

(e) Create a summary table for each `city` that includes the total number of sales and the mean median home price. Then create a new column showing how far each city's total sales is from the overall mean of the summarized sales values. Arrange from largest to smallest based on this difference.

(f) Create a visualization showing how the average median home price changes across years for the 5 cities with the largest average number of sales (Houston, Dallas, Austin, San Antonio, Collin County). To get the `geom_line()` layer to work properly you need to specify the grouping variable. Write 2–3 sentences describing what your graph suggests.

3. Look at the `mlb_players_18` dataset in the `openintro` library to answer the following questions:

(a) Calculate the total number of home-runs (HR) each team's outfield had (LF, CF, RF). Who were the top 5 teams? (Dodgers at 282)

(b) Show the top 3 players in home-run (HR) total for each position

(c) Create a variable which determines if a players last name starts with a letter in the first half of the alphabet or the second half of the alphabet. Do the same for their first name. Then determine the average number of hits (H) each combination has along with the number of people in each grouping. (Hint: use `grepl()` with regular expressions and conditional statements). You should get the 1st-half first name and 1st-half last name with 561 people

(d) Determine the average batting average for each position type (note that batting average is Hits/AB). Notice that it matters if we find the batting average for each person and then average the group together or if we sum the group statistics together and then find the average. (1B is 0.253, not 0.243)

(e) Look at just players on St. Louis' team (STL). Order the team based on the number of games they played (if there is a tie break it using alphabetical order for their name) and just display the name, position, games played, and hits. I then want you to go down the list of players and total up the number of hits for that player and everyone above them on the list

(f) Order the player in terms of at bats (AB) with the most at the top. Only show the name and AB column. Assign the minimum rank to the players (Turner, T is ranked #1). How many distinct ranks are there? (389)

(g) With the same code as above, I want you to determine what the largest At Bat "gap" in values is between ranks (so if rank 14 has 500 AB and rank 15 has 490 AB the gap would be 10).

Visualizations to Match for Problem 1 and 2: