ADS4293 LAB2

Chapter 2, 3, 4 - Data Cleaning, Wrangling, and Visualization

TA: Malak Al-Hanafi

18 September, 2025

LAB SETUP

- Excel file will be emailed at the beginning of the lab
- We'll go over one question at a time
- 10–15 minutes for you to solve in pairs or individually
- I'll put a slide with formulas to help you
- After that, I'll go over it
- Submit your **Excel file** on *Brightspace (Owl)* at the end of the lab
- Today's lab has 4 questions — goal is to finish them all within 2 hours

QUESTION 1

33. FILE ***Longitudinal_Partial.*** The accompanying data file contains the data from the National Longitudinal Survey (NLS), which follows over 12,000 individuals in the United States over time. Variables in this analysis include the following information on individuals: Urban (1 if lives in urban area, 0 otherwise), Siblings (number of siblings), White (1 if white, 0 otherwise), Christian (1 if Christian, 0 otherwise), FamilySize, Height, Weight (in pounds), and Income (in $). Remove Height and Weight from the data set.

    a. How many individuals live in an urban area and have more than three members in their households?

    b. Subset the data to create two data sets, one that includes observations who live in an urban area and have incomes over $40,000 and one that includes the rest of the observations. How many observations are in each subset?

QUESTION 1-A

a. How many individuals live in an urban area and have more than three members in their households?

- Use the Urban column (1 = lives in urban area, 0 = otherwise)
- Use the FamilySize column (> 3 means more than 3 people in the household)
- You can try:
  - An *indicator/helper column* with **IF()** to flag rows that meet both conditions, then use **SUM()**

OR

  - use **COUNTIFS()** directly with both conditions

QUESTION 1-B

b. Subset the data to create two data sets, one that includes observations who live in an urban area and have incomes over $40,000 and one that includes the rest of the observations. How many observations are in each subset?

- Use the Income column and the Urban column
- Think about creating an indicator/helper column that flags:
  - 1 = if Urban = 1 and Income > 40,000
  - 0 = otherwise
- You can then count how many observations fall in each group
- Use the **filter** to create a subset of the data that falls into each criteria

QUESTION 2

59. FILE ***Patients.*** Refer to Exercise 2.48 for a description of the problem and data set.

a. The variable Race contains five unordered categories: American Indian, Asian/Pacific Islander, Hispanic, Non-Hispanic Black, and Non-Hispanic White. Reduce the number of categories to four by combining the two least frequent categories. Name the new category “Other.” How many observations are in the “Other” category?

b. Transform the Race variable with the new categories into dummy variables. Use the most frequent race category in the data as the reference category. Which category is the reference category? How many dummy variables should be created?

c. The variable Education contains four ordered categories: HS (lowest educational attainment level), Some College, College, and Graduate (highest educational attainment level). Replace the category names with category scores ranging from 1 (lowest) to 4 (highest). What is the average category score for Education?

QUESTION 2-A

a. The variable Race contains five unordered categories: American Indian, Asian/Pacific Islander, Hispanic, Non-Hispanic Black, and Non-Hispanic White. Reduce the number of categories to four by combining the two least frequent categories. Name the new category “Other.” How many observations are in the “Other” category?

- Create a **PivotTable** with the Race variable
  - Place **Race** in the ***rows*** and ***values*** to get a count
    - This shows how many individuals fall into each race category
- Create a new column called ***new_race***
  - Identify the two smallest categories and group them into “Other”
  - Keep all other categories as they are
  - Count how many in “other”

QUESTION 2-B

b. Transform the Race variable with the new categories into dummy variables. Use the most frequent race category in the data as the reference category. Which category is the reference category? How many dummy variables should be created?

- After the first question, you should have:
  - *Hispanic, Non-Hispanic Black, Non-Hispanic White, and Other*
- For dummy variable creation
  - You always create k − 1 dummies if there are k categories
  - The most frequent group is the reference category (so no dummy for it)
  - Create a new column for each of the other categories

- Dummy variables are numeric variables (0/1) that represent categories of a qualitative (non-numeric) variable so they can be used in models and calculations. Because Excel, regression, and most statistical models need numbers, not text.

QUESTION 2-C

c. The variable Education contains four ordered categories: HS (lowest educational attainment level), Some College, College, and Graduate (highest educational attainment level). Replace the category names with category scores ranging from 1 (lowest) to 4 (highest). What is the average category score for Education?

- Create a new column called *Education_Score* and assign numeric scores to these categories:
  - HS = 1, Some College = 2, College = 3, Graduate = 4
- Create a new column called *Average_Education*
  - Find the average score

QUESTION 3

33. FILE ***Gas_Prices.*** The accompanying data file shows the price of gas (Price in $ per gallon) at 51 gas stations across the United States.

a. Construct a boxplot for the Price variable. Does the boxplot suggest that outliers exist?

b. Use *z*-scores to determine if there are any outliers for the Price variable. Are your results consistent with part a? Explain why or why not.

QUESTION 3-A

a. Construct a boxplot for the Price variable. Does the boxplot suggest that outliers exist?

- Select the Price column.
- Go to Insert → Charts → Box & Whisker
- Outliers will show as dots beyond the whiskers

QUESTION 3-B

b. Use *z*-scores to determine if there are any outliers for the Price variable. Are your results consistent with part a? Explain why or why not.

- Calculate the **mean** of the Price variable
  - Using =AVERAGE(range)
- Calculate the **standard deviation** of the Price variable
  - Use the =STDEV.S(range) function
- Identify the **minimum** and **maximum** values of the Price variable
  - using =MIN(range) and =MAX(range)
- Calculate **z-scores** for the min and max values using:
  - z = (x − mean) / stdev
    - If |z| > 3 → it's an outlier
    - If |z| ≤ 3 → it's not an outlier

- A z-score tells us how far a value is from the mean in terms of standard deviations.

QUESTION 4

35. FILE ***Healthy_Living.*** Healthy living has always been an important goal for any society. Most would agree that a diet that is rich in fruits and vegetables (FV) and regular exercise have a positive effect on health, while smoking has a negative effect on health. The accompanying data file shows the percentage of these variables observed in various states in the United States.

    a. Construct a scatterplot of Health against Exercise. Describe the relationship.

    b. Construct a scatterplot of Health against Smoking. Describe the relationship.

Thank you!