Homework 2 – BBUS 301 100 points

Material covered: 4. Statistical Graphs
5. Base R and Data Cleaning

Follow the directions below to complete the assignment. Once you have completed the assignment, submit your answer on Canvas before the due date.

Submission Guidelines:

Due date: July 15, 2025 at 11:15 am

Submission means: You will submit an .R script containing all code and explanations (remember to use # at the beginning of your comment) for each question in order to receive credit for your work.

Submission details: Students must make sure that their answers are readable, complete, and submitted before the deadline. Late submission for ANY reason, whether in part or in whole, or unreadable files will trigger the late penalties. The time when I receive the last part of your answers will count when computing late penalties, not the time when you sent the answers.

Late Penalties: If an assignment is turned immediately after the due date, your score will be reduced by 10% of the maximum assignment grade. Additional 10% will be deducted from your score for every 24 hours after that.

Academic Integrity: This is an individual assignment. Students need to submit their own answers and understand those answers. I reserve the right to ask you to walk me through the code for your assignment. If you fail to explain the code in detail and the choices you made, you will receive 0 points for the part of the assignment. Students who refuse to provide the explanation for their work will receive 0 points for the entire assignment.

Using code not discussed in class: Students who are submitting code not discussed in the class will temporary receive 0 points for that part of the assignment until they meet with me and they walk me through their code. Students need to meet with me within one week of returning the graded assignment, no later than August 19. If you fail to meet with me or if you cannot explain some of the code & the choices you made, you will keep the 0 points for that part of the assignment.

Homework Questions to Answer:

The data we will use in the homework (Seattle_wages.csv) was provided by Seattle Department of Human Resources.

Variables:

department: The department in which the City of Seattle employee works.
last_name: The last name (also known as a surname) of the City of Seattle employee.
first_name: The first name of the City of Seattle employee.
job_title: The job title of the City of Seattle employee.
Hourly Rate: The amount, listed in U.S. dollars, the City of Seattle employee makes per hour.

.

Part 1: Organizing your work (0 pts)

1. (0 pts) Create an R script called "hmw_2_name.R" (use your first name instead of name; don't use the quotation marks). This R script will contain all code and explanations (remember to use # at the beginning of your comment) for each question in this homework.
2. (0 pts) Use the RStudio keyboard shortcut Ctrl + Shift + R (or the Mac equivalent) to create a section header in your script. An "Insert Section" dialog box should appear, asking you to input the name of the section header—call it "Homework 2" and press OK.
3. (0 pts) Set your working directory to the folder where you downloaded the data for homework 2.
4. (0 pts) Import the *Seattle_wages.csv* file from your downloaded folder. Remember to use the assign operator in order to create a new data frame called *wages_df* that will contain the imported data.

Part 2: Base R and Data Cleaning – use Base R (55 pts)

5. (5 pts) Have a look at the handout for Module 1. Introduction to R Programming. The last few slides discuss the appropriate style for naming objects. Look at the names of the variables (columns) in the data frame you just created. Identify the variable name(s) that need changed/edited to comply with these style rules, and please do the change.
6. Look at the variable that records the department where employees work. For this variable:
   a. (10 pts) Create a count / frequency table for this variable. Which is the department with most employees?
   b. (5 pts) When recording these values a typo has been made. Is there more than one value in this column that represents the same department? Which ones?
   c. (10 pts) Combine them: choose one value and replace the other one(s) with the value you chose.
7. (10 pts) What are the names of the Seattle City Light department employees making an hourly wage higher than 150? Make sure that you filter the date using both conditions.
8. (5 pts) How many missing values are present per column?
9. (10 pts) Look at your numerical variable. Decide if you would like to either impute missing values as the value of the mean, replace the missing values with 0 or use a different value. Do it. Briefly explain your rationale for inputting the value of your choosing.

Part 3: Statistical Graphs – use Base R, not ggplot2 (45 pts)

10. (15 pts) Create a boxplot for the hourly rate variable. Draw the boxplot vertically. Mention two insights that we learn from this visual representation of our data.
11. (0 pts) Run the following command to increase the plot margins to accommodate longer labels on the x-axis for the next two plots:

    ```
    par(mar = c(7, 4, 4, 2))
    ```
12. (15 pts) Create a boxplot for the hourly wage by department. Draw the boxplot vertically. Rotate the x-axis labels to be perpendicular to the axis. Choose an appropriate title and label your y axis. Mention two insights that we learn from this visual representation of our data.
13. (15 pts) Create a bar chart of the departments. Choose an appropriate title and label your y axis. Rotate the x-axis labels to be perpendicular to the axis. Mention two insights that we learn from this visual representation of our data.