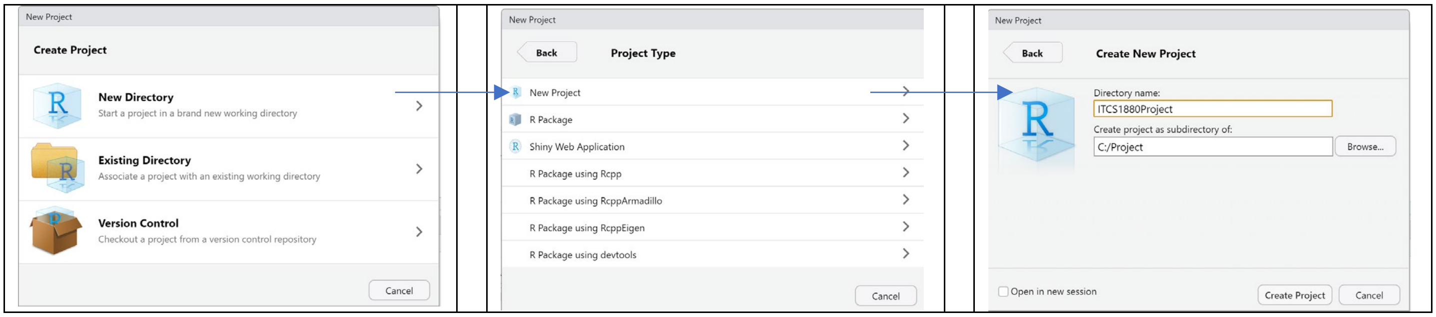
ITCS1880 – Final Project

Create a new project



Directory name: ITCS1880Project

Create project as subdirectory of: C:/Project

This creates a new project space for your final project. You will include all scripts in this project. You will have access to history and session data related specifically to your final project.

Download and examine the Mall_Customers.csv data file.

Read the Mall_Customers.csv data into a data frame called **customer_data**.

You will be using this data to conduct the following analysis steps:

1. Data Analysis
2. Plotting
3. Probability and Statistical Analysis
4. Summary Analysis

Step 1. – Data Analysis

In R Studio, create a new script and using your customer_data variable, perform the following data and statistical analysis:

- Use R built-in functions to find:
    - The number of rows and the number of columns in the data set.
    - The names of the columns in the data set
    - The top 10 rows of the data set
    - The class of the dataset and each column
        - Determine if there are any factor variables and what the levels are
- Summarize the data in the Mall Customers data set:
    - By each variable and as a whole
    - Be sure to include: Min, Max, Mean, Median, Quartiles and Standard Deviation
- For the factor variable (gender) in your dataset group by average: Age, Annual Income
- Aggregate the average Age and Annual Income by Gender in one analysis (hint: the must be co**m**bi**ne**d)

In a MS Word document, summarize the data based on your activities in Step 1 and include screen shots.

Step 2 – Plotting

You will now plot the data in your Mall Customers dataset to get a visual representation of the data. Create a new script in your project to create the following plots:

- Boxplots for each numeric variable in the dataset
- A histogram for each numeric variable – separated by Gender
- An appropriate scatterplot

In your Word document, include screen shots and an analysis based on the visual representation of data in your plots.

Step 3 – Probability and Statistical Analysis

- Use the dnorm function to verify whether or not the Age, Annual Income, and Spending Score variables follow the normal distribution
  - Plot each of these variables using geom_line()
  - Hint: You will need means and standard deviations from above
- Use the Shapiro-Wilk normality test to test for normality of each of the variables, and then each of the variables based on Gender.
  - What are the results? It they do not follow the normal distribution – which are closest?
- Conduct the Ansari-Bradley test to see if you can conduct a two-sample t-test.
  - If results are favorable – perform a t-test on Gender-Age, Gender-Income, and Gender-Spending Score.
- Use pnorm to find the probability of a person's age being above 40, and below 40. Do the same with age 50.
- Create a new script that will contain a function called – **corMallData()** - that will have 2 arguments and loop through the customer_data data frame. The function details are as follows:
  - The arguments you will pass to the function are: **data** and **gender**. So your function definition will look like this: **corMallData(data, gender) {}**.
    - You will pass your customer_data frame to the function for the **data** argument, and either the string "Female" or "Male" for the **gender** argument when you call the function
  - Inside the function you will:
    - Create an empty data frame called gData (**gDF <- data.frame()**)
    - Create a **for** loop – use **i** for loop variable (first part) and **1:200** for the vector (third part)
    - Inside the loop create an IF statement that will do the following:
      - Check if the data being passed in for gender is equal to the current value in the current row of the data passed into the function (customer_data passed to data argument)
        - Your IF statement might look something like this: **if (data[i,2]==gender){}**
          - The i variable represents the current row in the **for** loop, and 2 represents the gender column variable.
      - Use cbind to bind the Age (3rd), Annual Income (4th) and Spending Score (5th) data values to a row into a data frame called **gData**
        - gData <- cbind(data[i,3],data[i,4], data[i,5])
      - Use rbind to add this row to the gData data frame
        - gDF <- rbind(gDF, gData)
      - Once the loop is finished, return the gDF variable
    - Run the code for the function and call it using the function call: corMallData(customer_data, gender="Female"). Store this function call in a **fData** variable.
    - Change the names of the columns in the **fData** data frame to "Age", "Income" and "Spending"
  - Use the correlation function in R to see if there is a strong positive or negative correlation between age and income and Age and Spending
    - Do this for gender = Female and gender = Male

In your Word document, summarize and evaluate the results of your application of statistical functions on the Mall Customers data. Include appropriate screen shots.

Step 4 – Summary

Include an overall summary of your analysis of the Mall Customers data. Your analysis should be a minimum of 300 words. In your analysis you may include what was your favorite thing you learned in this class, as well as the most challenging concept.

Project Submission

When you are ready to submit, you will zip up your entire project folder including your Word document with your analysis of each step and all screen shots.