Problem Set 1

Applied Stats/Quant Methods 1

Name: Darragh McGee (18319331)

Question 1: Education

A school counselor was curious about the average of IQ of the students in her school and took a random sample of 25 students' IQ scores. The following is the data set:

```
Student_IQ <- c(105, 69, 86, 100, 82, 111, 104, 110, 87, 108, 87, 90, 94, 113,
    112, 98, 80, 97, 95, 111, 114, 89, 95, 126, 98)
```

1. **Find a 90% confidence interval for the average student IQ in the school.**

   90 Percent Confidence Interval = Point Estimate i.e. Mean +/- Margin of Error (Critical Value*Standard Error)

   Step 1: Input Data Set of Student IQs and Create Vector

   ```
   Student_IQ <- c(105, 69, 86, 100, 82, 111, 104, 110, 87, 108, 87, 90,
   94, 113, 112, 98, 80, 97, 95, 111, 114, 89, 95, 126, 98)
   ```

   Step 2: Calculating Mean Student Height

   ```
    mean_IQ <- mean(Student_IQ)
    print(mean_IQ)
   ```

   Mean Student IQ = 98.44

   Step 3: Calculating Standard Error

   ```
    standard_error_IQ <- sd(Student_IQ)/sqrt(length(Student_IQ))
    print(standard_error_IQ)
   ```

Standard Error = 2.618575

Step 4: Calculate Test Statistic for 90 Percent Confidence Level As the sample size is 25 (Less than 30) a T-Distribution should be used for Critical Value.

T-Statistic Formula requires Degrees of Freedom Degrees of freedom (df = n - 1)

```
 df <- 25 - 1
 print(df)
```

Degrees of Freedom = 24

T-Statistics for 90 Percent Confidence Interval is interested in the Critical Value for the First 5 Percent and Last 5 Percent of the Distribution

```
t_score_lower <- qt(0.05, df) # Critical Value for First 5% (Lower Bound)
print(t_score_lower)
t_score_upper <- qt(0.95, df) # Critical Value for Last 5% (Upper Bound)
print(t_score_upper)
```

t-score = +/- 1.710882

Step 5: Construct 90 Percent Confidence Interval

```
 lower_bound <- (mean_IQ+(t_score_lower)*(standard_error_IQ))
 print(lower_bound)
 upper_bound <- (mean_IQ+(t_score_upper)*(standard_error_IQ))
 print(upper_bound)
```

90 Percent Confidence Interval = 93.95933 to 102.9201

Using in-built t.test() formula in R to check results

```
 t.test(Student_IQ, conf.level = 0.9, alternative = "two.sided")
```

```
 90 percent confidence interval:
 93.95993 102.92007
 sample estimates:
 mean of x
 98.44
```

The results from the in-built t.test function are equal to the step-by-step calculations (allowing for marginal rounding differences).

2. Next, the school counselor was curious whether the average student IQ in her school is higher than the average IQ score (100) among all the schools in the country.

   Using the same sample, conduct the appropriate hypothesis test with $\alpha = 0.05$.

Step 1: Assumptions about Data

- Sample Size is below 30. Therefore, T-Statistic should be used.
- Population IQ = 100
- IQ represents continuous data
- Random Sampling Conducted
- The data is approximately normally distributed.
- Observations are independent of one another.

Step 2: Setting Up Hypothesis

- **Null Hypothesis:** The average Student IQ in the sample school is equal to the average IQ score (100) among all the schools in the country
- **Alternative Hypothesis:** The average student IQ in the sample school is greater than the average IQ score (100) among all the schools in the country

Steps 3 and 4: Calculating Test Statistic and P-Value using T-Test formula:

```
t.test(Student_IQ, mu = 100, alternative = c("greater"), conf.level=0.95)
```

T-test Output

```
t = -0.59574, df = 24, p-value = 0.7215
alternative hypothesis: true mean is greater than 100
95 percent confidence interval:
93.95993      Inf
sample estimates:
mean of x
98.44
```

Step 5: Conclusion

- Fail to reject the Null Hypothesis as the p-value is greater than 0.05 (exceeding the 5 percent significance level).
- There is insufficient evidence to conclude that the sample mean is greater than 100.

Question 2: Political Economy

Researchers are curious about what affects the amount of money communities spend on addressing homelessness. The following variables constitute our data set about social welfare expenditures in the USA.

Figure 1.1.png

`State`	*50 states in US*
`Y`	*per capita expenditure on shelters/housing assistance in state*
`X1`	*per capita personal income in state*
`X2`	*Number of residents per 100,000 that are ”financially insecure” in state*
`X3`	*Number of people per thousand residing in urban areas in state*
`Region`	*1=Northeast, 2= North Central, 3= South, 4=West*

Explore the `expenditure` data set and import data into `R`.

```r
df <- 25 - 1
```

- Please plot the relationships among *Y*, *X1*, *X2*, and *X3*? What are the correlations among them (you just need to describe the graph and the relationships among them)?

- Please plot the relationship between *Y* and *Region*? On average, which region has the highest per capita expenditure on housing assistance?

- Please plot the relationship between *Y* and *X1*? Describe this graph and the relationship. Reproduce the above graph including one more variable *Region* and display different regions with different types of symbols and colors.