Format instructions for the assignment report

Perform descriptive statistics on your dataset and report the results.

Format

1. First, tell about the dataset and its source.

   Example: The dataset is an employee's salary data and fetched from .

2. Provide information on the number of instances and attributes.

   Example: The dataset contains X number of records and Y number of attributes.

3. Select some attributes from the datasets and perform the following

   1. Plot histogram and add an explanation of your interpretation of the histogram. Add proper labels to your x and y-axis [You can use xlab and ylab parameters].
   2. Compute the central measure (Mean, Median, Mode) and explain which one you would prefer and why.
   3. Compute dispersion (variance or standard deviation) and explain it (your interpretation).

4. Compute the correlation between your attributes (report it in a table) and see if there is any relationship between attributes.

   If you find a correlation higher than .6, report it using a graph and explain it.

5. Formulate 4-5 queries to extract data from your dataset which involve the use of `select` `filter` `summarize` `group-by` `mutate` functions from `dplyr` package in R.

   Example:

   How many employees have salary less high than 15500 in the population.csv dataset (available here)

   Following is the R code (you can take a snapshot from R-Studio or simply put your code)

   ```
   data %>% filter(salary > 15500) %>% summarize(n=n())
   ```

   The above command gave us the following result

   ```
     n

   1 1567
   ```

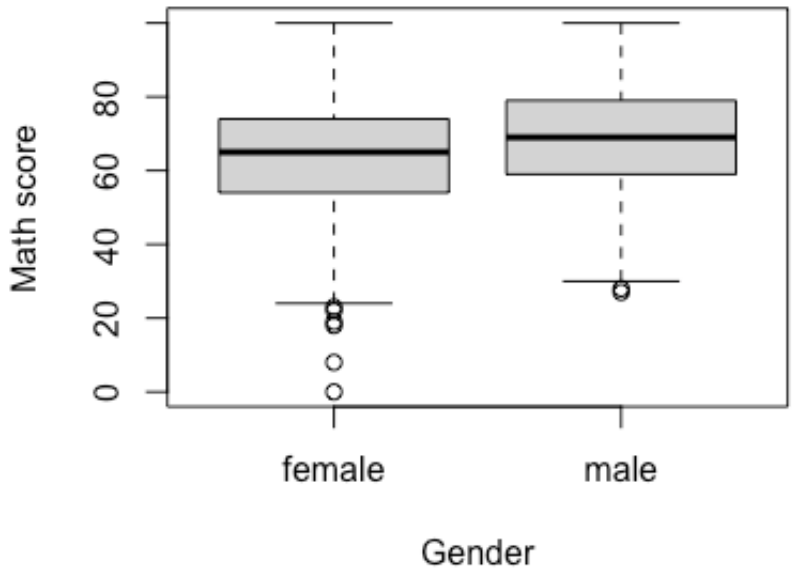
   We can say that there are 1567 employees in the population.csv dataset having salaries higher than 15500.

   1. Similarly, you need to demonstrate the use of other functions using the same format (Your question, R code, Output, Explanation)

Statistical tests

Format

You either have questions to investigate or formulate your questions. Now, to formulate questions you first need to have a look at the given data. So you can do descriptive statistics. Here, I am using a simple plot. So let's say we plotted the data for Male and Female students and got the following



From the graph, we can see that the mean math score for male students is higher than for female students. So we want to investigate the question on differences in scores in male and female students.

- First, write your question you want to investigate and your hypothesis (null and alternative)

  Example:

> H_0: Female students have scored the same as male students.
>
> H_1: Female students have not scored the same as male students.

Check for assumptions: In this case, normal distribution for scores in both male and female groups and variance in both male and female groups.

- State which tests you would perform and why

  Example: Here, we will apply an independent sample t-test. As the number of gender groups is two and samples are not from the same persons. Plus, the math score attribute is numeric (interval) type.

- Specify your level of significance (α) (e.g., 5%, 1%, .1%)

  Example: We are specifying 5% of the level of significance.

- Perform the test and report your R code (snapshot or code)

Example:

```
> t.test(data$`math score`~data$gender,var=T)

  Two Sample t-test

data:  data$`math score` by data$gender
t = -5.3832, df = 998, p-value = 9.12e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.952285 -3.237737
sample estimates:
mean in group female   mean in group male
            63.63320             68.72822
```

- Draw the conclusion
- *Example:* From the result, we can see that the p-value is less than .05 (5% level of significance) therefore we reject the null hypothesis and the alternative hypothesis is accepted. Thus, a significant difference is found in the math score of male and female students.