CDS62124

Statistical Data Analysis

Assignment 1: Descriptive Statistics & Data Visualization Analysis

Tutorial: TT5L

Student ID	Name
1211108075	Hiew Wei Cheng

Table of Contents

Dataset **3**

Part A: Exploratory Descriptive Analysis **3**

Part B: Distributional Assessment **7**

Part C: Bivariate Relationship Analysis **9**

Part D: Application & Critical Reflection **10**

Dataset - https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python

This dataset is about mall customer segmentation and has gender, age, annual income and spending score variables. The purpose of choosing this dataset is to find the relationship between age and spending score to understand how customers' age affects their spending habits.

Part A: Exploratory Descriptive Analysis

Summary of Quantitative Variables

This dataset has 2 quantitative variables, which are age and spending score from 1 to 100. Below is the R code and summary of these two variables.

Age

```
> age <- data$Age
> spending_score <-data$Spending.Score..1.100.
> mean(age)
[1] 38.85
> median(age)
[1] 36
> as.numeric(names(sort(table(age), decreasing = TRUE)[1]))
[1] 32
> sd(age)
[1] 13.96901
> var(age)
[1] 195.1332
> range(age)
[1] 18 70
> quantile(age)
   0%   25%   50%   75%  100%
18.00 28.75 36.00 49.00 70.00
> IQR(age)
[1] 20.25
> sd(age)/mean(age)
[1] 0.3595626
```

Above is the code and detailed summary for age variable. The mean for age is 38.85, with the median of 36 and a mode of 32. The standard deviation for age 13.97 and have the variance of 195.13. The range for this variable starts from 18 to 70. The quantiles are 28.75 (Q1), 36.0 (Q2) and 49.0 (Q3). The interquartile range (IQR) is 20.25 and the coefficient variation (CV) is 0.3596.

Spending Score

```
> mean(spending_score)
[1] 50.2
> median(spending_score)
[1] 50
> as.numeric(names(sort(table(spending_score), decreasing = TRUE)[1]))
[1] 42
> sd(spending_score)
[1] 25.82352
> var(spending_score)
[1] 666.8543
> range(spending_score)
[1]  1 99
> quantile(spending_score)
   0%   25%   50%   75%  100%
 1.00 34.75 50.00 73.00 99.00
> IQR(spending_score)
[1] 38.25
> sd(spending_score) / mean(spending_score)
[1] 0.5144128
```

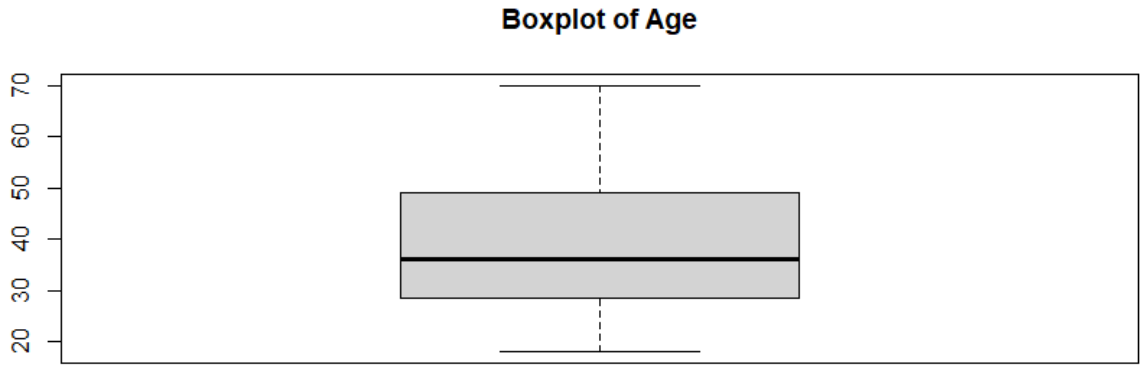
Above is the code and detailed summary for spending score variable. The mean for spending score is 50.2, with a median of 50 and a mode of 42. The standard deviation for spending score is 25.82 and has the variance of 666.85. The range for this variable starts from 1 to 99. The quantiles are 34.75 (Q1), 50.00 (Q2) and 73.0 (Q3). The interquartile range (IQR) is 38.25 and the coefficient variation (CV) is 0.5144.

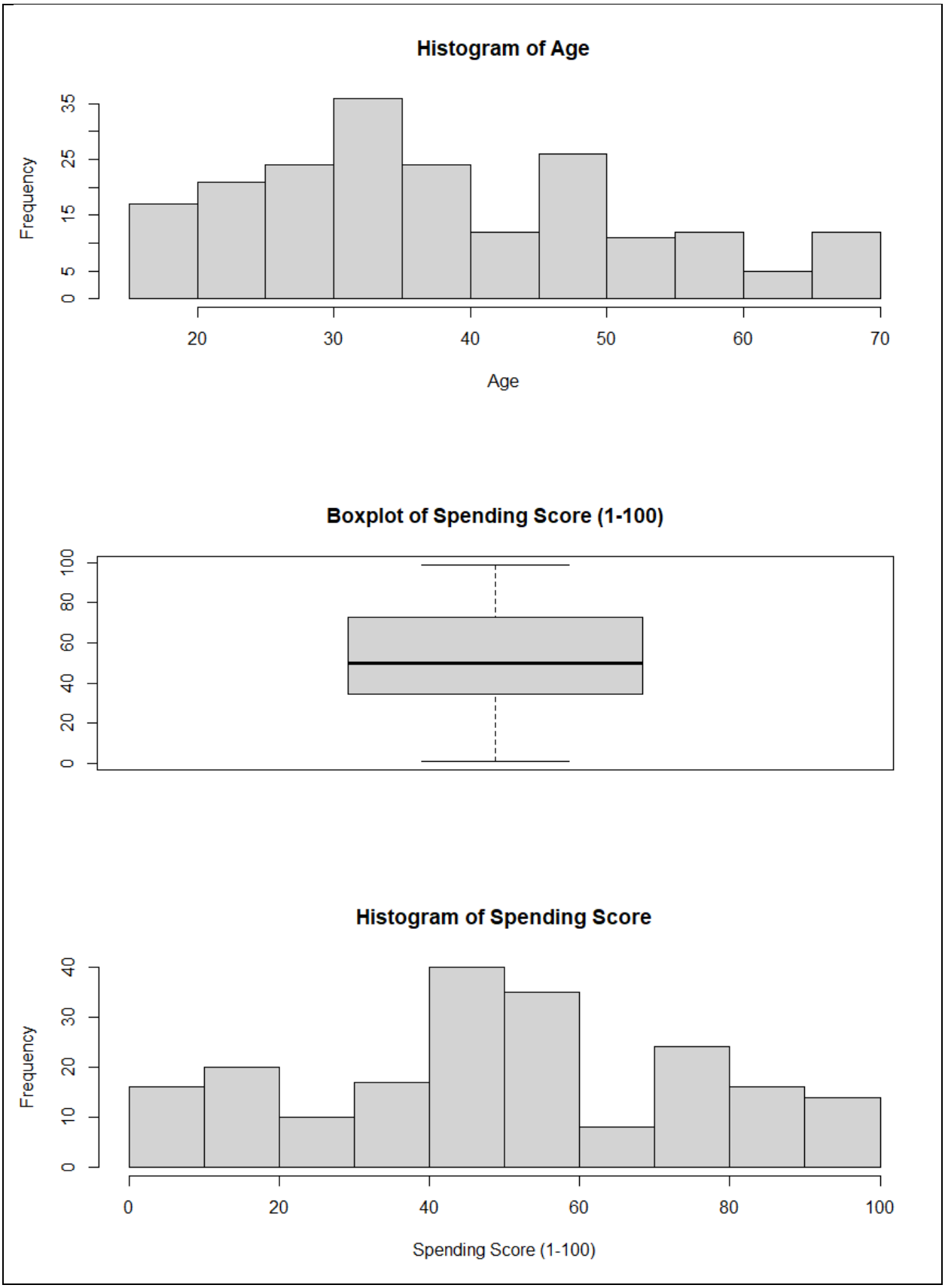
Boxplot and Histogram for Annual Income and Spending score

Code:

```
> boxplot(age, main= "Boxplot of Age")
> boxplot(spending_score, main= "Boxplot of Spending Score (1-100)")
> hist(age, main="Histogram of Age", xlab="Age")
> hist(spending_score, main="Histogram of Spending Score", xlab="Spending Score (1-100)")
```

Boxplot and Histogram:




Histogram of Age
Frequency
0
5
15
25
35
20
30
40
50
60
70
Age
Boxplot of Spending Score (1-100)
0
20
40
60
80
100
Histogram of Spending Score
Frequency
0
10
20
30
40
0
20
40
60
80
100
Spending Score (1-100)

Comparison between Age and Spending Score

Above is the code and output for boxplot and histogram based on age and spending score. According to the boxplot and histogram for age, it shows that the age distribution is right skewed and does not contain any outlier. Since the mean (38.85) is greater than the median (36) for age variable, it proofs that the age distribution is right skewed. This suggests that this variable is asymmetric.

For the spending score, the histogram seems to be not normally distributed. But according to boxplot of spending score, it shows that the distribution is right skewed and does not have any outlier. By comparing the mean (50.2) and median (50), the mean is greater than median which shows that its distribution is slightly right skewed. This suggests that this variable is asymmetric.

The standard deviation and coefficient variation (CV) for age is 13.97 and 0.3596 while the standard deviation and coefficient variation for spending score is 25.82 and 0.5144. From these data, we can see that spending score has higher standard deviation and coefficients variation. This means that spending score has a wider spread and shows more variability which makes it less consistent. In the other hand, standard deviation and coefficient variation for age is smaller, showing that the data is more consistent and clustered around the mean. Therefore, age is a more consistent variable than spending score as its value less spread out compared to spending score.

Part B: Distributional Assessment

Empirical Rule R Code:

```
> range_first <- c(mean_age  - sd_age , mean_age  + sd_age )
> range_second <- c(mean_age  - 2*sd_age , mean_age  + 2*sd_age )
> range_third <- c(mean_age  - 3*sd_age , mean_age  + 3*sd_age )
```

Output:

range_first	num [1:2] 24.9 52.8
range_second	num [1:2] 10.9 66.8
range_third	num [1:2] -3.06 80.76

Proportion:

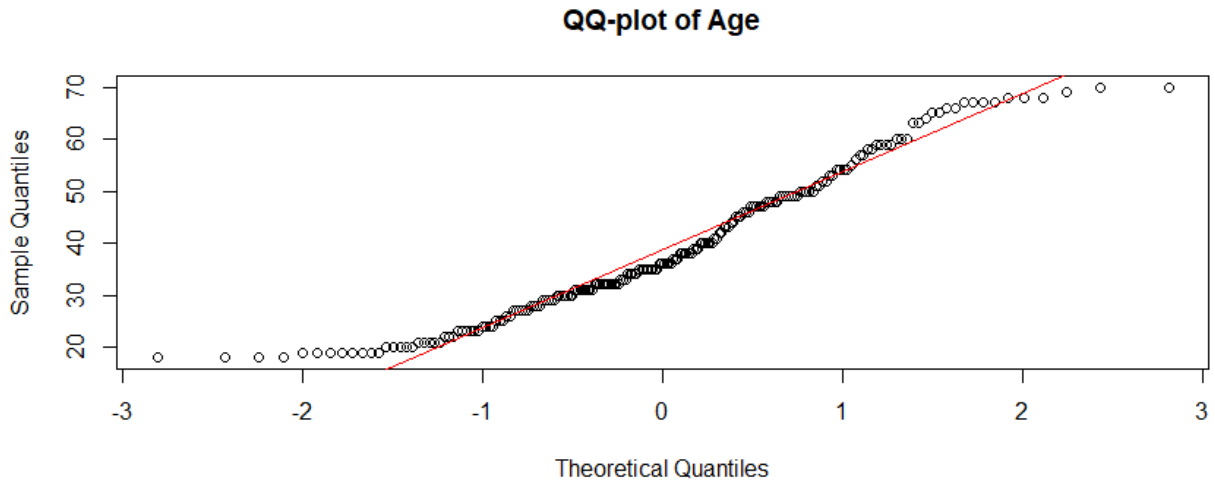
```
> prop_first <- mean(age >= range_first[1] & age <= range_first[2])
> prop_second <- mean(age >= range_second[1] & age <= range_second[2])
> prop_third <- mean(age >= range_third [1] & age <= range_third [2])
```

Output:

prop_first	0.645
prop_second	0.95
prop_third	1

QQ-Plot:

```
> qqnorm(age, main="QQ-plot of Age")
> qqline(age, col="red")
```



Shapiro-Wilk Test:

```
> shapiro.test(age)

        Shapiro-Wilk normality test

data:  age
W = 0.95162, p-value = 2.711e-06
```

For this part we will be using age variable. Empirical Rule is applied to examine whether the age variable follows a normal distribution. Based on the calculation, we can see that the range of ±1 standard deviation is between 24.9 to 52.8, covering 64.5% of data which is slightly below the expected 68%. For ±2 standard deviations, the range is between 10.9 to 66.8, covering 95% of data which is the same as expected value. The range of ±3 standard deviation is between -3.06 to 80.76, covering 100% of the data which is slightly higher than the expected 99.7%. These results show that the distribution for age is close to normal.

A QQ-plot is also generated to see whether the age variable appears to be normally distributed. Based on the QQ-plot for age, the points mostly follow the red reference line but there is visible deviation at the tails, showing potential skewness. This suggests that the age distributions have minor deviations from perfect normal.

The Shapiro-Wilk test is also applied to examine the normality of age variable. The result we get is p-value = 2.711e-06 which is equal to 0.000002711. Since the p-value is less than 0.05, we can reject the null hypothesis that the data is normally distributed. This confirms that age variable does not follow a perfect normal distribution.

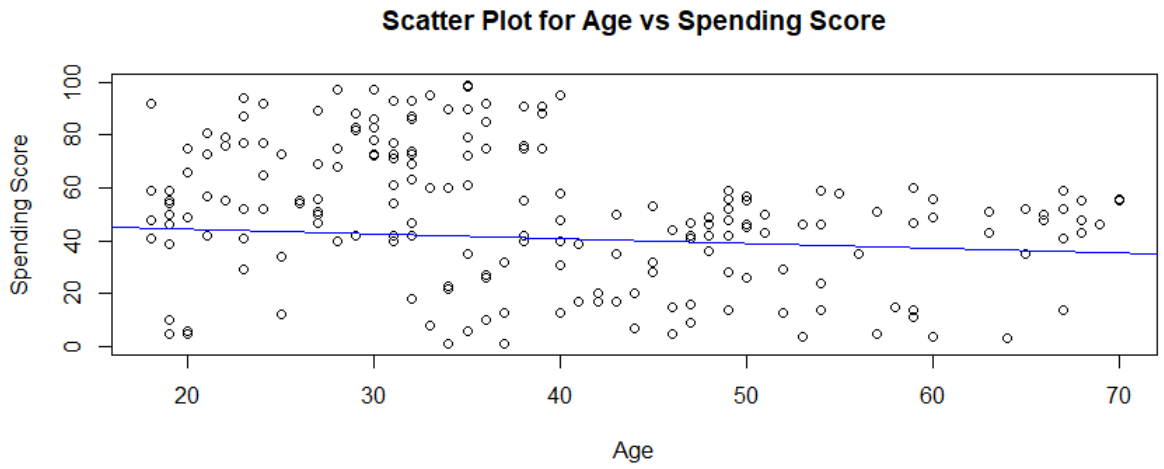
Which method is most reliable and practical

To determine whether the age variable follows a normal distribution, the empirical rules, QQ-plot and Shapiro-Wilk test are applied. The empirical rules only provide a quick numerical check while the QQ-plot only provides visual interpretation. For the Shapiro-Wilk test, it provides strong statistical evidence against normality. Therefore, the Shapiro-Wilk test is the most reliable and practical method because it is based on a formal statical test.

Part C: Bivariate Relationship Analysis

Scatter Plot:

```
> plot(age, spending_score, main="Scatter Plot for Age vs Spending Score", xlab="Age", ylab="Spending Score")
> abline(lm(age ~ spending_score), col="blue")
```



Relationship between Age and Spending Score variables

Based on the scatter plot, we can see the scatter plot shows an unclear linear trend but with slightly downward trend. This suggests that the relationship between the two variables is weak negative relationship, meaning that when the age increase, the spending scores tend to decrease slightly.

Pearson Correlation Coefficient Formula in R:

```
> S_xx <- sum((age - mean_age)^2)
> S_yy <- sum((spending_score - mean_spending_score)^2)
> S_xy <- sum((age - mean_age) * (spending_score - mean_spending_score))
> r <- S_xy / sqrt(S_xx * S_yy)
```

Output:

S_xx	38831.5
S_xy	-23490
S_yy	132704
r	-0.32722684603909

Pearson correlation coefficient

The result of Pearson correlation coefficient is r = -0.327. Since the r value is between 0 to -1, it shows that it is weak negative linear relationship between the variables. This means that when age increase, the spending score tends to decrease but the relationship is not strong.

Evaluate whether the relationship is causal or associative

Although the Pearson correlation coefficient shows a week negative relationship between age and spending score. It only shows a trend but not prove that shows the age causes a decrease in the spending score. Therefore, the relationship is associative and not causal.

Part D: Application & Critical Reflection

1. **Summarize insights for non-technical audience**
   In this dataset, we looked at customer's age and spending score to understand the patterns and relationships. Based on the calculations, the average age is 39, and most customers are between 18 to 70 years old. For the average spending score, we get 50 and the range between 1 to 99. While comparing these two variables we see that the age data is more consistent while spending scores vary more from person to person.

   When we checked if the age variable follows a normal pattern, we found that it is close but not perfect match. A detailed check using Shapiro-Wilk test confirmed that age variable does not perfectly follow a normal distribution.

   We also explored the relationship between age and spending score. It showed us that the relationship between them is a weak negative relationship which means when people get older, they tend to spend slightly less. However, it does not mean that it is the causes of lower spending, it's just an observed pattern.

   In summary, the analysis shows a pattern that customer tends to spend less when they get older but there may be various factors that cause the spending score decrease.

2. **One limitation of descriptive statistics and how inferential stats could help**
   One limitation of descriptive statistics is that its results are only applicable to study sample and cannot be speculated to the whole population. For example, while we observed a weak relationship between age and spending score in this dataset, it does not mean that it will be the same for the whole population.

   On the other hand, inferential stats allow us to use data from sample to make inferences or predictions for a larger population. It helps us determine whether the observed patterns are statistically significant or occurred by chance. This help makes inferential stats more reliable.

3. **Reflect on how visualizations enhance understanding beyond numbers.**
   Visualizations enhance understanding beyond numbers by turning numbers into visuals that are easier to understand and interpret. Charts such as boxplot, histogram, QQ-plot, scatter plot help us see the patterns, trends, outliers, skewness immediately. For example, boxplot can help us see the skewness of the variable and outliers which is difficult to determine from numbers only. Visualization also helps non-technical audiences have better understanding instead of throwing all the numbers for them. In summary, visualization helps us understand a bunch of data faster.