XIAMEN UNIVERSITY MALAYSIA

Couse Code	:	MAT203
Course Name	:	Statistics
Academic Session	:	September 2024
Assessment	:	Project

Prepared by : 1. WANG FANGYU ID: MAT2309449

Mark: /100

CY Hapiness Report 2024

1.Introduction

CY owns a chain of stores called CY Hapiness, offering a variety of products across eight categories: Cosmetics, Clothing, Books, Shoes, Food and Beverages, Toys, Souvenirs, and Technology. He operates one store in each of the states of Kuala Lumpur, Selangor, Perak, and Johor. CY is an elderly man, he has not yet implemented a comprehensive system to gather data from all his stores, but he plans to install one soon. As the end of 2024 approaches, CY is preparing for the upcoming year. To aid his planning, he has collected sales data from 20 days in 2024 as a sample, which he intends to analyse to optimize his inventory and explore potential business activities in the coming year.

2.Sampling Method

The data consists of information from 20 randomly selected days. Each branch has data for these 20 days. We believe that the sampling method is inappropriate. From the data, we can observe that samples were taken on five days in January and April, four days in February, three days in March, and one day each in June, August, and September. It is known that seasons can significantly impact a store's sales performance, and the current sampling heavily concentrates on the period from January to April. Therefore, we consider this sampling method inappropriate. We recommend using a stratified sampling method instead. Specifically, the twelve months of the year can be divided into four quarters: January to March, April to June, and so on. Then, five days (five samples) should be randomly selected from each quarter.

3.Distribution of Payment Methods and Gender Proportions

The proportions of payment methods used by customers

The available payment methods are Cash, Credit Card, and Debit Card. Below is the proportion of customers using each payment method.

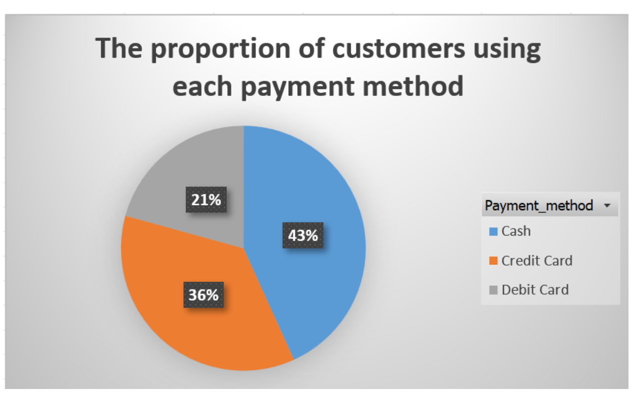


Chart 1: The proportion of customers using each payment method.

From the chart above, we can clearly see the proportions of payment methods used by customers during transactions. According to our statistical data, 43% of customers paid with cash, over one-third used credit cards, and approximately one-fifth opted for debit cards.

The proportions of female and male customers

The proportions of female and male customers are as follows:

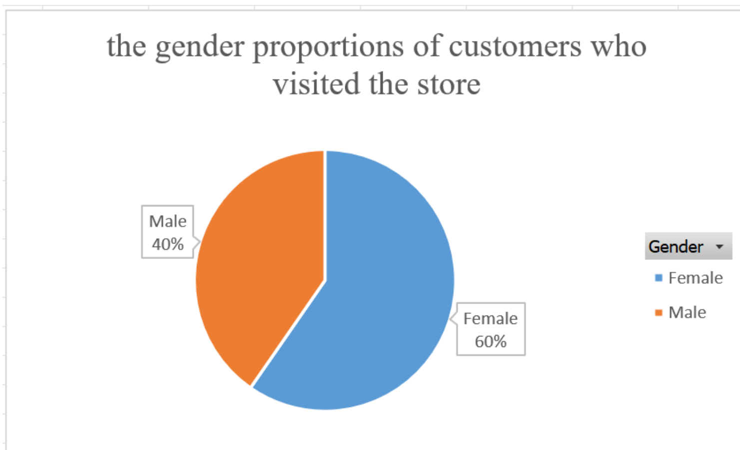


Chart 2: The proportions of female and male customers.

According to our statistical data, we can determine the gender distribution of in-store customers, with females accounting for 60% and males accounting for 40%.

4.Average Total Sales per Day

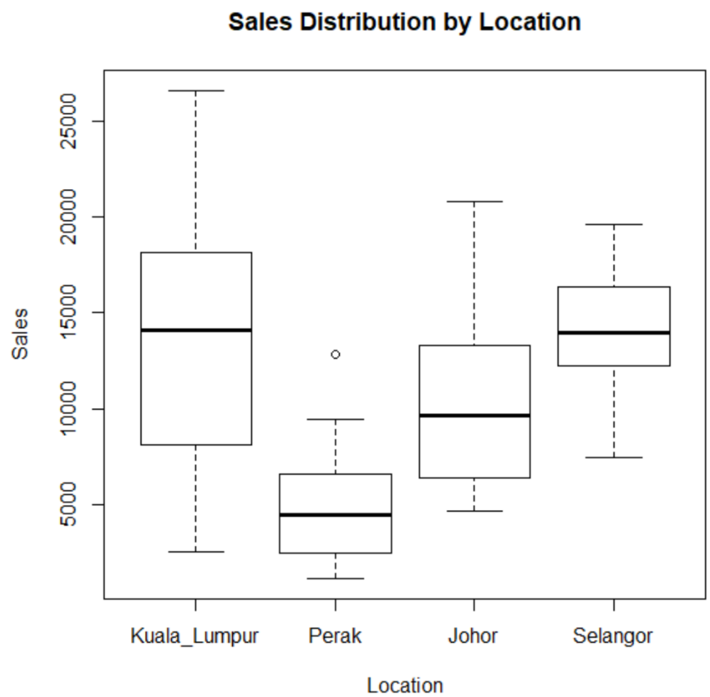
Analysis of the average total sales per day

Based on the provided data, the estimated average total sales per day for each location are as follows:

Location	Average Total Sales per Day (RM)
Johor	10356.651
Kuala Lumpur	13501.72
Perak	4913.1685
Selangor	14139.967

Table 1: The estimated average total sales per day for each location.

The boxplots below display the average total sales per day for each location:



Boxplot 1: The boxplots of the average total sales per day for each location.

From the boxplot, we can infer that Kuala Lumpur and Selangor have higher sales compared to the other two regions, as their medians are similar and higher than the others. Perak has the lowest sales, as its median is the lowest among the four, while Johor's sales are moderate. Additionally, we can deduce that Kuala Lumpur and Johor exhibit higher variability in sales, as their interquartile ranges are larger than those of the other two regions, with Kuala Lumpur showing the greatest variability. Furthermore, we also notice an outlier in the data for Perak.

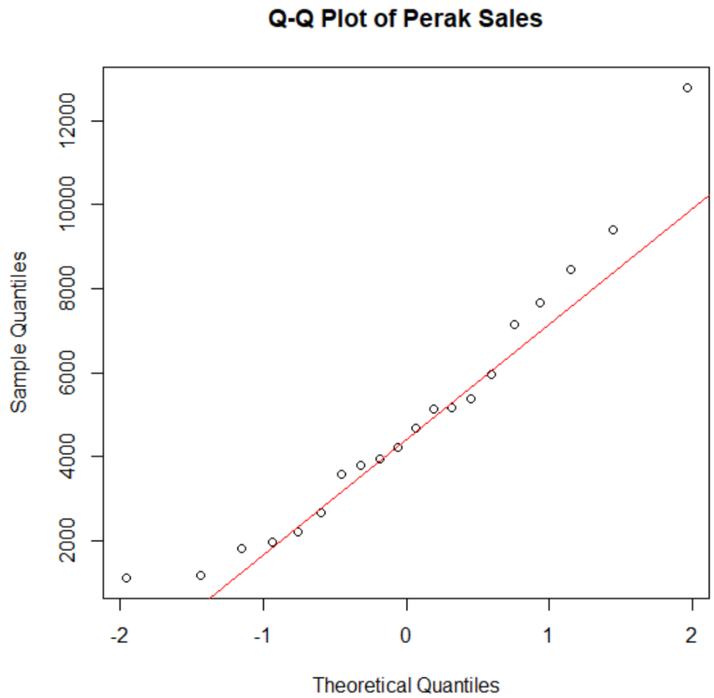
Hypothesis test assisting in making decision

CY is considering closing the branch in Perak if the average total sales per day fall below RM5,000. To assist in this decision, the following hypothesis test has been conducted.

We organized the sales data for 20 days in Perak, resulting in the following 20 data points, each corresponding to the sales for those days.

```
3570.85 12804.84  5389.48  7660.39  4665.49  8452.57  5165.16  1817.76
7152.15  2202.97  3804.23  4217.23  1097.39  1180.69  2654.97  9417.62
5967.30  3952.02  1955.60  5134.66
```

Using R, we created a Q-Q plot based on this data.



From the Q-Q plot above, we can observe that the data points are almost all distributed near the red line. Therefore, we can conclude that our data follows a normal distribution.

Here we gives a hypothesis test of a significance level of 0.05: let μ be the real average total sales per day, then the two hypotheses:

$$H_0\colon \mu = 5000$$
$$H_a\colon \mu < 5000.$$

Using R, we can get the sample standard deviation as follows:

```
> data <- c(3570.85, 12804.84, 5389.48, 7660.39, 4665.49, 8452.57,
+           5165.16, 1817.76, 7152.15, 2202.97, 3804.23, 4217.23,
+           1097.39, 1180.69, 2654.97, 9417.62, 5967.3, 3952.02,
+           1955.6, 5134.66)
> sample_sd <- sd(data)
> print(sample_sd)
[1] 3016.384
```

So the standard deviation:

$$\sigma \approx s = 3016.384.$$

From the above data, the sample mean of total sales per day for Perak is:

$$\bar{X} = 4913.1685.$$

And the number of sample: $n = 20$.

Assume H_0 is true, then

$$Z = \frac{\bar{X} - \mu}{\sigma \Big/ \sqrt{n}} = \frac{4913.1685 - 5000}{3016.384 \Big/ \sqrt{20}} \approx -0.13.$$

The P-value:

$$p\left(Z < -0.13\right) = 0.4483 > 0.05$$

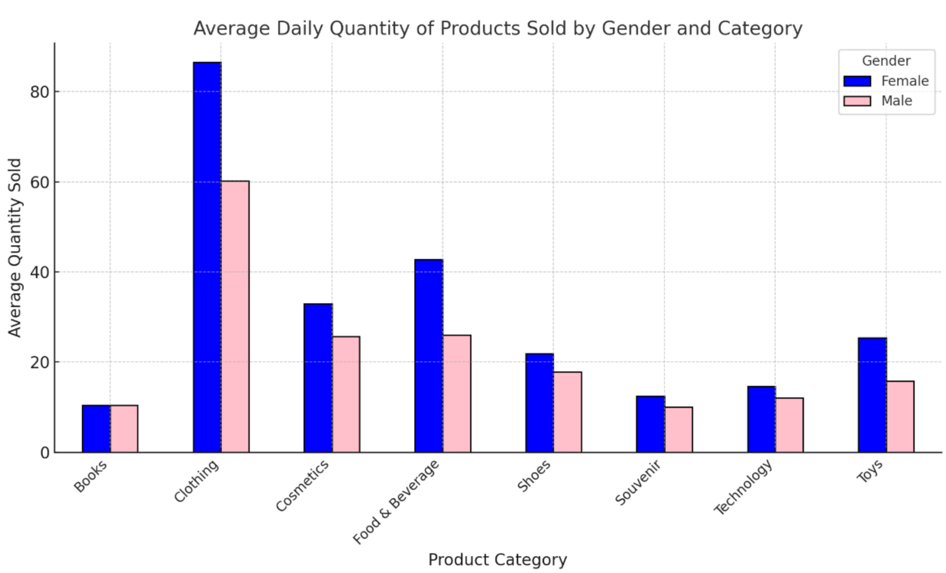
Decision:

Fail to reject H_0.

So we do not have compelling evidence for concluding that the true average total sales per day for Perak is less than 5000, thus the decision is not closing the branch in Perak.

5.Clothing Inventory Replenishment

The chart below illustrates the average daily quantity of products sold to male and female customers across different product categories.



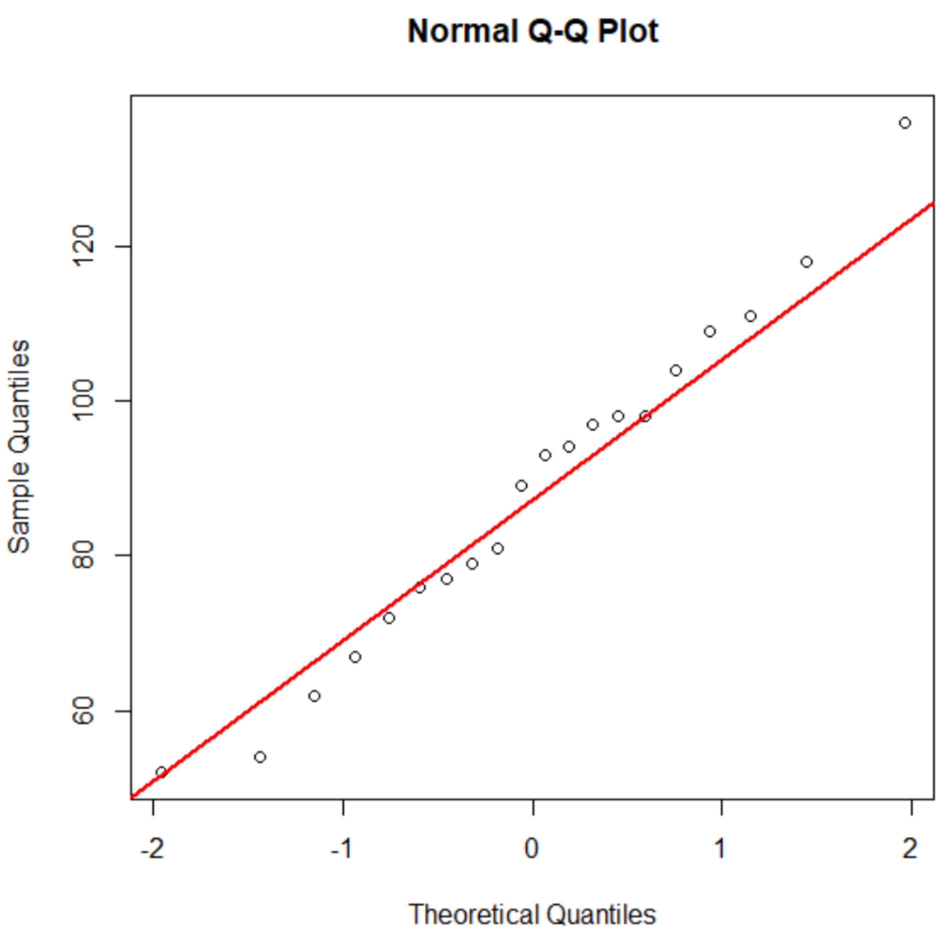
From the chart, we can observe that, apart from books where the average daily consumption is equal between males and females, in all other categories, females' average daily consumption exceeds that of males. Notably, in the categories of clothing, food & beverage, and toys, females' average daily consumption is much higher than that of males.

CY plans to manage the clothing inventory for next year, as it has the highest daily sales. However, he is uncertain whether he should order more clothing for females than for males.

By consolidating the statistical data, we obtain the sales data of clothing for twenty days as follows.

67	118	62	81	93	52	79	109	77	98
98	111	94	97	89	136	104	76	54	72

Using R, we created a Q-Q plot based on this data.



From the Q-Q plot above, we can observe that the data points are almost all distributed near the red line. Therefore, we can conclude that our data follows a normal distribution.

By organizing the statistical data, we obtain the clothing consumption data for women and men over twenty days as follows.

Female									
37	74	32	54	52	27	41	65	49	67
60	63	70	62	63	86	59	52	32	59
Male									
30	44	30	27	41	25	38	44	28	31
38	48	24	35	26	50	45	24	22	13

Using R, we can get the sample variances as follows:

```
> female <- c(37, 74, 32, 54, 52, 27, 41, 65, 49, 67,
+             60, 63, 70, 62, 63, 86, 59, 52, 32, 59)
> male <- c(30, 44, 30, 27, 41, 25, 38, 44, 28, 31,
+           38, 48, 24, 35, 26, 50, 45, 24, 22, 13)
>
> female_variance <- var(female)
> male_variance <- var(male)
>
> cat("Female sample variance:", female_variance, "\n")
Female sample variance: 233.7474
> cat("Male sample variance:", male_variance, "\n")
Male sample variance: 99.81842
```

Here we gives a hypothesis test of a significance level of 0.05 to determine if it is plausible that the variances of the average daily quantities of clothing sold are equal for male and female customers: let ${\sigma_1}^2, {\sigma_2}^2$ be the real variances of the average daily quantities of clothing sold for male and female, respectively, then the two hypotheses:

$$H_0\colon {\sigma_1}^2 = {\sigma_2}^2$$
$$H_a\colon {\sigma_1}^2 \neq {\sigma_2}^2.$$

Let ${S_1}^2, {S_2}^2$ be the sample variances of the average daily quantities of clothing sold for male and female, respectively, n be the number of samples, so

$$\begin{aligned} {S_1}^2 &= 99.8184 \\ {S_2}^2 &= 233.7474 \\ n &= 10. \end{aligned}$$

Assume H_0 is true, then the Test Statistic:

$$f = \frac{{S_1}^2}{{S_2}^2} = \frac{99.8184}{233.7474} = 0.4270.$$

And

$$
\begin{aligned}
P - value &= 2p\left(F_{9,9} < 0.4270\right) \\
&= 2p\left(F_{9,9} > 2.3419\right) \\
&> 2p\left(F_{9,9} > 2.44\right) \\
&= 2 \times 0.1 \\
&= 0.2 > 0.05.
\end{aligned}
$$

Decision:

$$\text{Fail to reject } H_0.$$

So we do not have compelling evidence for concluding that the real variances of the average daily quantities of clothing sold for male and female are different. So it is plausible that the variances of the average daily quantities of clothing sold are equal for male and female customers.

Base the statistical data we organized, and using R, we can get the sample means as follows:

```
> female <- c(37, 74, 32, 54, 52, 27, 41, 65, 49, 67,
+             60, 63, 70, 62, 63, 86, 59, 52, 32, 59)
> male <- c(30, 44, 30, 27, 41, 25, 38, 44, 28, 31,
+           38, 48, 24, 35, 26, 50, 45, 24, 22, 13)
> female_mean <- mean(female)
> male_mean <- mean(male)
>
> cat("Female sample mean:", female_mean, "\n")
Female sample mean: 55.2
> cat("Male sample mean:", male_mean, "\n")
Male sample mean: 33.15
```

Now we gives a hypothesis test of a significance level of 0.01 to determine if the average daily quantity of female clothing sold is greater than that of male clothing: let μ_1, μ_2 be the real means of the average daily quantities of clothing sold for male and female, respectively, then the two hypotheses:

$$
\begin{aligned}
H_0\colon \mu_1 - \mu_2 &= 0 \\
H_a\colon \mu_1 - \mu_2 &< 0.
\end{aligned}
$$

Let $\bar{X}_1, \bar{X}_2$ be the sample means of the average daily quantities of clothing sold for male and female, respectively, n be the number of samples, so

$$
\begin{aligned}
\bar{X}_1 - \bar{X}_2 &= -22.05 \\
n &= 10.
\end{aligned}
$$

Assume H_0 is true, then the Test Statistic:

$$
\begin{aligned}
Z &= \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{{\sigma_1}^2}{n} + \frac{{\sigma_2}^2}{n}}} \\
&= \frac{-22.05 - 0}{\sqrt{\frac{99.8184}{10} + \frac{233.7474}{10}}} \\
&= -3.8178
\end{aligned}
$$

And

$$P - value = 2p\left(Z < -3.8178\right) \approx 0 < 0.01.$$

Decision:

$$\text{Reject } H_0.$$

So we have compelling evidence for concluding that the average daily quantity of female clothing sold is greater than that of male clothing. So the decision that CY should make is ordering more female clothing.

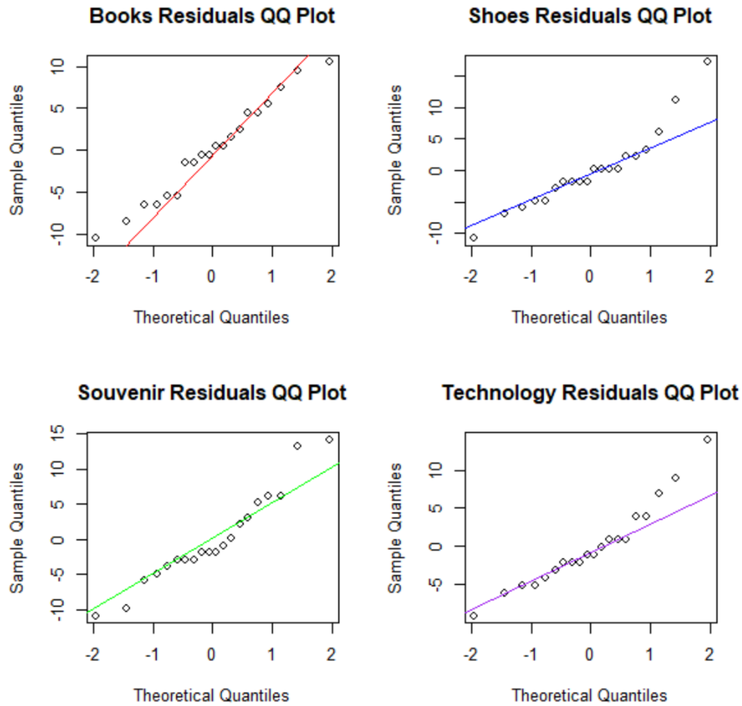
6. Promotional Campaign

CY is planning to launch a promotional campaign for Books, Shoes, Souvenirs, and Technology. Before finalizing his promotional idea, he wants to determine if the average quantity sold per day is the same across these four categories.

By organizing the statistical data, we can obtain twenty residuals for each of the four categories as follows

Books	Shoes	Souvenir	Technology
-5. 5	-1. 8	6. 15	-2. 05
-0. 5	0. 2	-10. 85	-6. 05
0. 5	-2. 8	-0. 85	3. 95
-1. 5	-4. 8	6. 15	-2. 05
4. 5	2. 2	-1. 85	3. 95
9. 5	-1. 8	-4. 85	0. 95
-10. 5	0. 2	-9. 85	-4. 05
-8. 5	0. 2	-2. 85	-1. 05
10. 5	-1. 8	3. 15	0. 95
1. 5	11. 2	-2. 85	-1. 05
-6. 5	6. 2	-1. 85	0. 95
4. 5	17. 2	5. 15	-3. 05
5. 5	-1. 8	-3. 85	13. 95
2. 5	-6. 8	0. 15	-0. 05
7. 5	2. 2	-1. 85	-2. 05
-6. 5	3. 2	13. 15	8. 95
-5. 5	-10. 8	2. 15	-5. 05
-0. 5	0. 2	14. 15	-9. 05
-1. 5	-4. 8	-5. 85	-5. 05
0. 5	-5. 8	-2. 85	6. 95

Using R, we can get the Q-Q plot of these four categories' residuals.



From the plots, we can observe that the residuals for books are almost entirely distributed near the line, while the residuals for the other three categories, except for a few outliers, are also mostly distributed near the line. This indicates that the daily quantity sold of all four categories are generally stable at the daily quantity sold, with few instances of deviation from the daily quantity sold.

Now we gives a single-factor ANOVA test of a significance level of 0.05 to determine whether the average daily quantity sold is the same across all four categories: let $\mu_1, \mu_2, \mu_3, \mu_4$ be the real the average daily quantity sold of four categories, respectively, then the two hypothesises are

$$H_0\colon \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$H_a\colon \text{at least one } \mu_i \neq \mu_j, i \neq j.$$

After organizing the statistical data, we obtain the table of sales for 20 days across the four product categories as follows:

Books	Shoes	Souvenir	Technology
7	22	18	12
12	24	1	8
13	21	11	18
11	19	18	12
17	26	10	18
22	22	7	15
2	24	2	10
4	24	9	13
23	22	15	15
14	35	9	13
6	30	10	15
17	41	17	11
18	22	8	28
15	17	12	14
20	26	10	12
6	27	25	23
7	13	14	9
12	24	26	5
11	19	6	9
13	18	9	21

Using R, we could get the SST, SSTR and SSE of our data as follows:

```
> books <- c(7, 12, 13, 11, 17, 22, 2, 4, 23, 14, 6, 17, 18, 15, 20, 6, 7, 12, 11, 13)
> shoes <- c(22, 24, 21, 19, 26, 22, 24, 24, 22, 35, 30, 41, 22, 17, 26, 27, 13, 24, 19, 18)
> souvenir <- c(18, 1, 11, 18, 10, 7, 2, 9, 15, 9, 10, 17, 8, 12, 10, 25, 14, 26, 6, 9)
> technology <- c(12, 8, 18, 12, 18, 15, 10, 13, 15, 13, 15, 11, 28, 14, 12, 23, 9, 5, 9, 21)
>
> all_data <- c(books, shoes, souvenir, technology)
> group_means <- c(mean(books), mean(shoes), mean(souvenir), mean(technology))
> overall_mean <- mean(all_data)
>
> n_books <- length(books)
> n_shoes <- length(shoes)
> n_souvenir <- length(souvenir)
> n_technology <- length(technology)
>
> sstr <- n_books * (group_means[1] - overall_mean)^2 +
+   n_shoes * (group_means[2] - overall_mean)^2 +
+   n_souvenir * (group_means[3] - overall_mean)^2 +
+   n_technology * (group_means[4] - overall_mean)^2
>
> sse <- sum((books - group_means[1])^2) +
+   sum((shoes - group_means[2])^2) +
+   sum((souvenir - group_means[3])^2) +
+   sum((technology - group_means[4])^2)
>
> sst <- sum((all_data - overall_mean)^2)
>
> cat("SSTR :", sstr, "\n")
SSTR : 1866.1
> cat("SSE :", sse, "\n")
SSE : 2791.7
> cat("SST :", sst, "\n")
SST : 4657.8
```

So we have

$$
\begin{aligned}
SSTr &= 1866.1 \\
SSE &= 2791.7 \\
SST &= 4657.8 \\
MSTr = \frac{SSTr}{I-1} &= \frac{1866.1}{4-1} = 622.033 \\
MSE = \frac{SSE}{n-I} &= \frac{2791.7}{80-4} = 36.7329 \\
f = \frac{MSTr}{MSE} &= \frac{622.033}{36.7329} = 16.9339
\end{aligned}
$$

and

$$P - value = p\left(F_{3,76} > 16.9339\right) < 0.05.$$

Decision:

$$\text{Reject } H_0.$$

So we do not have compelling evidence for concluding that the average daily quantity sold is the same across all four categories. So at least exist one categories' average daily quantity sold is different from the other categories.

Now we will use Tukey's Procedure to check which category or categories differ from the others.
Base on the organized data, using R, we can get all the pairwise confidence intervals for $\mu_i - \mu_j$, and the result are as follows:

```
                   Pair      Lower     Upper
1        Books vs Shoes -15.202930 -7.397070
2     Books vs Souvenir  -3.348484  4.648484
3   Books vs Technology  -5.201749  2.101749
4     Shoes vs Souvenir   7.853608 16.046392
5   Shoes vs Technology   5.988839 13.511161
6 Souvenir vs Technology -6.061148  1.661148
```

Fome the result above we obtain that '0' is not included in the pairwise confidence intervals for the combinations of Books vs Shoes, Shoes vs Souvenir, and Shoes vs Technology, where Shoes appear in all three combinations. In contrast, '0' is included in the confidence intervals for the other three combinations, indicating that the average daily quantity sold of Shoes is different from the other three.

Student ID	**Member**	**Contribution**
MAT2309449	WANG FANGYU	• All the work.