UTM

UNIVERSITI TEKNOLOGI MALAYSIA

FACULTY OF COMPUTING

Group Project

COURSE NAME : **STATISTIC FOR DATA SCIENCE (MCSD1113)**

GROUP NO. : **2**

PREPARED BY :

NO.	NAME	STUDENT ID
1	GRACE LING KIAN HWAI	MCS231026
2	MIR TAMZID HASAN	A20EC4037
3	YAP QI YUAN	MCS231025

LECTURER'S NAME : **DR. MOHAMAD SHUKOR TALIB**

DATE OF SUBMISSION : **4th JULY 2024**

TABLE OF CONTENT

INTRODUCTION

The purpose of this study is to analyse customer shopping trends using a dataset containing demographic and purchasing information. This analysis aims to uncover patterns and insights that can help businesses tailor their marketing strategies, optimize product offerings, and improve customer experience. By understanding these trends, we expect to identify key factors that influence customer spending and behaviour.

DATASET

2.1 Dataset Description

The "Customer Shopping Trends Dataset" dataset which consists of 3900 rows x 18 columns was collected from Kaggle (BANERJEE, 2023) and the all the variables are categorised by level of measurement, type of data and type of variables as shown in the Table 2.1. The variables which have been used for the project's analysis are highlighted in blue.

Variables	Description	Level of Measurement	Type of Data	Type of Variable
Customer ID	Unique identifier for each customer	Nominal	Categorical	Qualitative
Age	Age of the customer in years	Ratio	Continuous	Quantitative
Gender	Gender of the customer (e.g., Male, Female)	Nominal	Categorical	Qualitative
Item Purchased	Specific item purchased by the customer	Nominal	Categorical	Qualitative
Category	Category to which the purchased item belongs (e.g., Clothing)	Nominal	Categorical	Qualitative
Purchase Amount (USD)	Amount spent by the customer in USD	Ratio	Continuous	Quantitative
Location	State or region where the customer is located	Nominal	Categorical	Qualitative
Size	Size of the purchased item (e.g., S, M, L)	Ordinal	Categorical	Qualitative
Color	Color of the purchased item	Nominal	Categorical	Qualitative
Season	Season during which the purchase was made (e.g., Winter)	Nominal	Categorical	Qualitative
Review Rating	Customer's rating of the purchased item on a scale	Interval	Continuous	Quantitative
Subscription Status	Whether the customer is a subscriber (Yes/No)	Nominal	Categorical	Qualitative
Shipping Type	Type of shipping selected (e.g., Express)	Nominal	Categorical	Qualitative
Discount Applied	Whether a discount was applied to the purchase (Yes/No)	Nominal	Categorical	Qualitative
Promo Code Used	Whether a promo code was used for the purchase (Yes/No)	Nominal	Categorical	Qualitative
Previous Purchases	Number of previous purchases made by the customer	Ratio	Discrete	Quantitative
Payment Method	Method of payment used (e.g., Venmo, Cash)	Nominal	Categorical	Qualitative
Frequency of Purchases	How frequently the customer makes purchases (e.g., Fortnightly)	Ordinal	Categorical	Qualitative

Table 2.1 : Variables Description and Categorisation.

2.2 Data Preprocessing

```
> sum(is.na(shopping_data))
[1] 0
> sum(duplicated(shopping_data))
[1] 0
```

```
> str(shopping_data)
'data.frame':   3900 obs. of  18 variables:
 $ Customer.ID           : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Age                   : int  55 19 50 21 45 46 63 27 26 57 ...
 $ Gender                : chr  "Male" "Male" "Male" "Male" ...
 $ Item.Purchased        : chr  "Blouse" "Sweater" "Jeans" "Sandals" ...
 $ Category              : chr  "Clothing" "Clothing" "Clothing" "Footwear" ...
 $ Purchase.Amount..USD. : int  53 64 73 90 49 20 85 34 97 31 ...
 $ Location              : chr  "Kentucky" "Maine" "Massachusetts" "Rhode Island"
 $ Size                  : chr  "L" "L" "S" "M" ...
 $ Color                 : chr  "Gray" "Maroon" "Maroon" "Maroon" ...
 $ Season                : chr  "Winter" "Winter" "Spring" "Spring" ...
 $ Review.Rating         : num  3.1 3.1 3.1 3.5 2.7 2.9 3.2 3.2 2.6 4.8 ...
 $ Subscription.Status   : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ Shipping.Type         : chr  "Express" "Express" "Free Shipping" "Next Day Air
 $ Discount.Applied      : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ Promo.Code.Used       : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ Previous.Purchases    : int  14 2 23 49 31 14 49 19 8 4 ...
 $ Payment.Method        : chr  "Venmo" "Cash" "Credit Card" "PayPal" ...
 $ Frequency.of.Purchases: chr  "Fortnightly" "Fortnightly" "Weekly" "Weekly" ...
```

Figure 2.2 : Check for Null, Duplicates and Data Type.

Before the data is used to do the analysis, we ensured that there are no missing and duplicate values, and the data types are correct for each variable as shown in Figure 2.2.

DATA ANALYSIS

3.1 Descriptive Analysis

3.1.1 Bar Chart

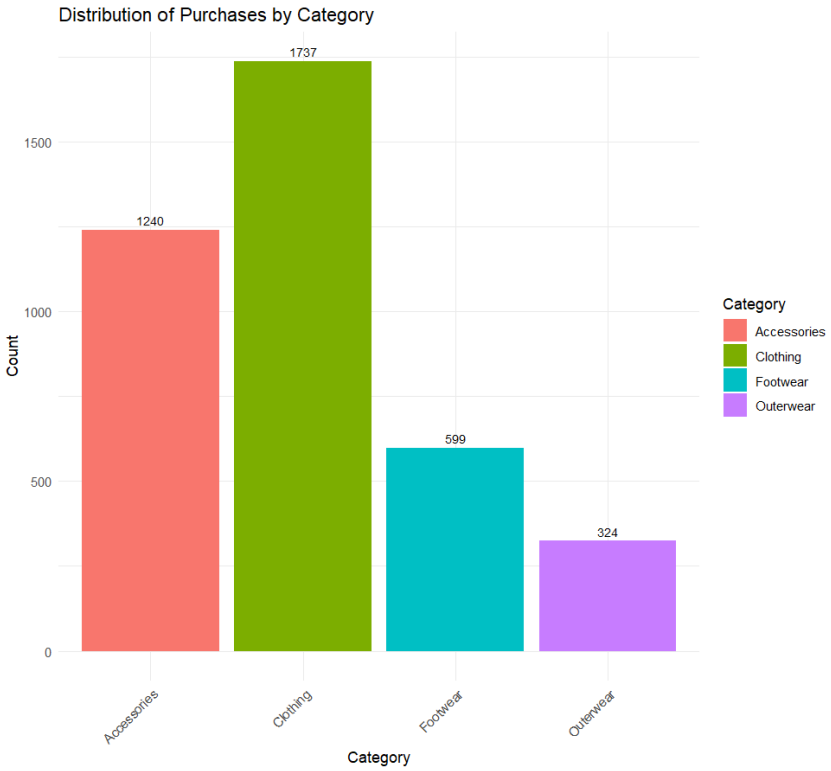


Figure 3.1.1 Distribution of Purchases by Category

From the bar chart as shown in Figure 3.1.1, most of the customers purchases clothing (1737 pax) followed by Accessories (1240 pax), Footwear (599 pax) and Outerwear (324 pax).

3.1.2 Pie Chart

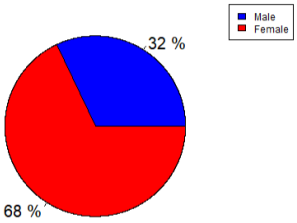


Figure 3.1.2 Gender Percentage

According to the pie chart as shown in Figure 3.1.2, female is more preferring to do shopping than male as female buyer is more than male buyer by 36%.

3.1.3 Stem & Leaf

```
The decimal point is 1 digit(s) to the left of the |

25 | 0000000000000000000000000000000000000000000000000000000000000000000000
26 | 0000000000000000000000000000000000000000000000000000000000000000000000+79
27 | 0000000000000000000000000000000000000000000000000000000000000000000000+74
28 | 0000000000000000000000000000000000000000000000000000000000000000000000+56
29 | 0000000000000000000000000000000000000000000000000000000000000000000000+90
30 | 0000000000000000000000000000000000000000000000000000000000000000000000+82
31 | 0000000000000000000000000000000000000000000000000000000000000000000000+77
32 | 0000000000000000000000000000000000000000000000000000000000000000000000+72
33 | 0000000000000000000000000000000000000000000000000000000000000000000000+72
34 | 0000000000000000000000000000000000000000000000000000000000000000000000+102
35 | 0000000000000000000000000000000000000000000000000000000000000000000000+76
36 | 0000000000000000000000000000000000000000000000000000000000000000000000+69
37 | 0000000000000000000000000000000000000000000000000000000000000000000000+76
38 | 0000000000000000000000000000000000000000000000000000000000000000000000+62
39 | 0000000000000000000000000000000000000000000000000000000000000000000000+83
40 | 0000000000000000000000000000000000000000000000000000000000000000000000+101
41 | 0000000000000000000000000000000000000000000000000000000000000000000000+68
42 | 0000000000000000000000000000000000000000000000000000000000000000000000+91
43 | 0000000000000000000000000000000000000000000000000000000000000000000000+67
44 | 0000000000000000000000000000000000000000000000000000000000000000000000+78
45 | 0000000000000000000000000000000000000000000000000000000000000000000000+59
46 | 0000000000000000000000000000000000000000000000000000000000000000000000+94
47 | 0000000000000000000000000000000000000000000000000000000000000000000000+68
48 | 0000000000000000000000000000000000000000000000000000000000000000000000+64
49 | 0000000000000000000000000000000000000000000000000000000000000000000000+86
50 | 0000000000000000000000000000000000000000000000000000000000000000000000
```

Figure 3.1.3 Review Rating

Based on the Review Rating Stem and Leaf Plot as shown in Figure 3.1.3, most of the customers were satisfied with the product they purchased as the overall rating started from mid score and above. The lowest rating is 2.5 out of 5.0 and the highest rating is 3.4 out of 5.0.

3.1.4 Histogram

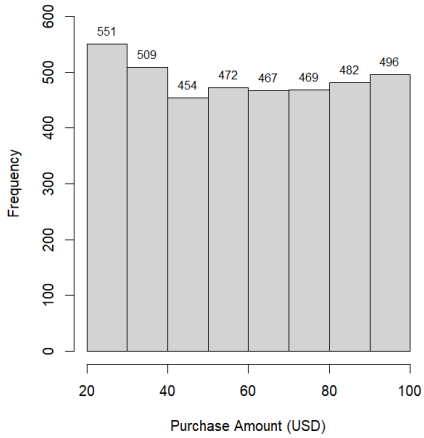


Figure 3.1.4 Purchase Amount Frequency

The histogram in Figure 3.1.4 shows the distribution of the Frequency of Purchase Amount (USD). Most customers spend around 20-40 USD for the products.

3.1.5 Boxplot

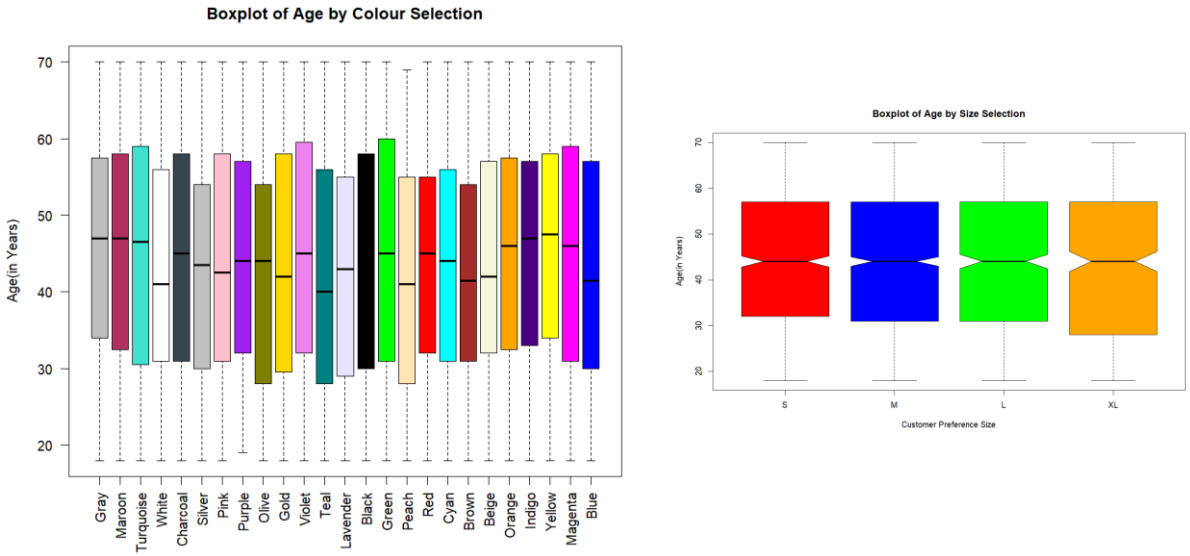


Figure 3.1.5 Colour (left) and Size (right) Selection

The boxplots in Figure 3.1.5 show the customer's preference of colour and size of the product purchased. From the left boxplot, there is no single colour that stands out as preferred across all ages, but certain colours like turquoise, charcoal, gold, black, green, magenta and blue show consistent popularity across a broader age range while from the right boxplot, the size preferences are more consistent compared to colour preferences, indicate a more uniform distribution of age across different sizes.

3.1.6 Descriptive Analysis (mean, median, mode, etc.)

```
> summary(Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   31.00   44.00   44.07   57.00   70.00
> summary(Review.Rating)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.50    3.10    3.70    3.75    4.40    5.00
> summary(Purchase.Amount..USD.)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.00   39.00   60.00   59.76   81.00  100.00
```

Figure 3.1.6 Descriptive Analysis for Age, Review Rating and Purchase Amount (USD)

From the analysis as shown in Figure 3.1.6, the central 50% of the ages lie between 31 and 57 years, showing a wide age range of customers. The statistic shows most of the customer is middle-aged individuals where the median age is 44 years. The central 50% of the purchase amounts lie between $39 and $81, suggesting a broad range of spending among customers. It also indicates the Purchase amounts vary significantly, with a balanced distribution around $60.

The central 50% of the ratings lie between 3.1 and 4.4, indicating generally positive reviews obtained from the customer.

3.2 Inferential Analysis

3.2.1 Hypothesis Testing 1-sample or 2-sample

```
> t.test(Age, mu=44.07, alternative="less", conf.level = 0.95)

        One Sample t-test

data:  Age
t = -0.0063177, df = 3899, p-value = 0.4975
alternative hypothesis: true mean is less than 44.07
95 percent confidence interval:
     -Inf 44.46911
sample estimates:
mean of x
 44.06846
```

Figure 3.2.1.1 Hypothesis Testing 1-sample (Age)

From previous session 3.1.6, we obtained the mean age is 44.07. We claimed that the buyer age should less than 44.07. From the hypothesis testing as in Figure 3.2.1.1, we can conclude that p-value (0.4975) is greater than the significance level of 0.05 (assuming a 95% confidence level), we fail to reject the null hypothesis. Therefore, there is not enough evidence to conclude that the true mean age is less than 44.07 years based on the sample data.

```
        Welch Two Sample t-test

data:  Purchase.Amount..USD. by Shipping.Type
t = 1.5108, df = 1297.9, p-value = 0.1311
alternative hypothesis: true difference in means between
group Express and group Standard is not equal to 0
95 percent confidence interval:
 -0.6014504  4.6314255
sample estimates:
 mean in group Express mean in group Standard
              60.47523               58.46024
```

Figure 3.2.1.2 Hypothesis Testing 2-sample (Purchase Amount by Shipping Type)

From the result as in Figure 3.2.1.2, customers appear to spend similarly regardless of the shipping type chosen as p-value (0.1311) > significant level (0.05).

3.2.2 Goodness of Fit Test

```
        Chi-squared test for given probabilities

data:  observed_frequencies
X-squared = 1247.2, df = 3, p-value < 2.2e-16
```

Figure 3.2.2 Goodness of Fit Test (Category)

According to the result shown in Figure 3.2.2, there is extremely strong evidence where p-value (2.2e^-16) < significant level (0.05) to reject the null hypothesis. The observed data does not align with the expected frequencies based on the given probabilities.

3.2.3 Chi Square Test of Independence

```
        Pearson's Chi-squared test

data:  contingency_table
X-squared = 4.1573, df = 5, p-value = 0.527
```

Figure 3.2.3 Chi Square Test of Independence (Payment Method & Discount Applied)

Based on result shown in Figure 3.2.3, we can conclude that there is no significant evidence where p-value (0.527) > significant level (0.05) to suggest that the variables represented in the contingency table are dependent.

3.2.4 Correlation

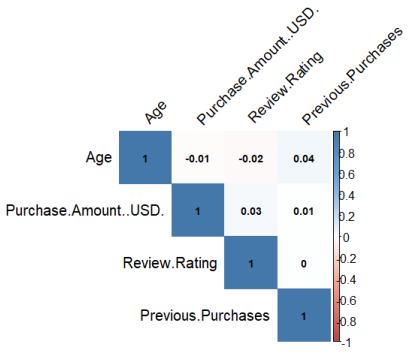
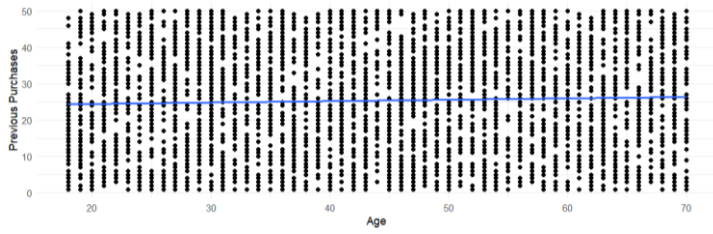


Figure 3.2.4 Correlogram

From the correlation plot as shown in Figure 3.2.4, there is no correlation between previous purchase and review rating, while have weak positive correlation between age and previous purchases, purchase amount and review rating, and purchase amount and previous purchase. Weak negative correlation happened between age and purchase amount and age and review rating.

3.2.5 Regression




```
Residuals:
     Min       1Q   Median       3Q      Max
-25.3479 -12.4745  -0.0415  12.4580  25.6116

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.65834    0.70876  33.380   <2e-16 ***
Age          0.03842    0.01520   2.527   0.0115 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.44 on 3898 degrees of freedom
Multiple R-squared:  0.001636,  Adjusted R-squared:  0.00138
F-statistic: 6.387 on 1 and 3898 DF,  p-value: 0.01154
```

Figure 3.2.5 Regression Plot between Age and Previous Purchases (left) and Results (right)

Since the correlation between age and previous purchases is the highest among the datasets, therefore, a linear regression line is plotted to further investigate their relationship as shown in Figure 3.2.5. From the result, there is a statistically significant relationship between age and previous purchases (p-value = 0.0115), but the effect size is very small. The coefficient for age is positive (0.03842), suggesting that previous purchases slightly increase as age increases.

However, the R-squared value is very low (0.001636), indicating that age explains very little of the variation in previous purchases. This means that while age does have a statistically significant effect, it is not practically significant in explaining previous purchases.

3.2.6 ANOVA

```
Analysis of Variance Table

Response: Previous.Purchases
            Df Sum Sq Mean Sq F value  Pr(>F)
Age          1   1331 1331.18  6.3866 0.01154 *
Residuals 3898 812466  208.43
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3.2.6 ANOVA Test Result between Age and Previous Purchases

From the ANOVA Test Result as shown in Figure 3.2.6, the p-value for Age (0.01154) is less than 0.05, indicating that the relationship between Age and Previous Purchases is statistically significant at the 5% significance level. The sum of squares for Age (1331) is relatively small compared to the sum of squares for Residuals (812466), indicating that while Age is a statistically significant predictor of Previous Purchases, it explains only a small portion of the total variation.

CONCLUSION

From this project, we learnt the method to choose appropriate dataset by determine the level of measurement (nominal, ordinal, interval and ratio) for the analysis. Moreover, we managed to use what we had learnt in class to do data pre-processing and data analysis using RStudio. From the analysis, we found that:

1. Customer Demographics: The gender distribution in the dataset is not uniform, with more females than males and customers span a broad age range, with a median age of 44 years, indicating a diverse customer base.

2. Product Preferences: Clothing is the most purchased category, followed by Accessories, Footwear, and Outerwear. No significant preference for the colour and the size of the products.

3. Review Ratings: Generally positive review ratings (equal or above 2.5) suggest satisfaction with purchased products.

4. Statistical Relationships: Age shows a statistically significant but weak relationship with previous purchases.

From the analysis, we can recommend tailoring marketing efforts based on demographic insights, suggest product optimization and enhance customer experience based on product preferences and review ratings.

REFERENCES

BANERJEE, S. (2023, 10). *Customer Shopping Trends Dataset*. Retrieved from kaggle: https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trends-dataset?select=shopping_trends_updated.csv