BACKGROUND

In today's dynamic and competitive retail environment, understanding consumer shopping behaviour is crucial for businesses to thrive and effectively meet customer needs. This Group Assignment will allow you to investigate the complexities of consumer behaviour.

The dataset that you will be working with captures a wide range of consumer purchases, offering a multifaceted view of shopping patterns and preferences. This includes:

1. Demographic information (Age, Gender)
2. Purchase details (Item Purchased, Category, Purchase Amount)
3. Indicator of customer behaviour (Frequency of Purchases)
4. Contextual factors (Season)
5. Transaction details (Payment Method, Shipping Type)

This collection of variables allows for a nuanced exploration of how various factors interact to shape consumer choices and shopping patterns. Your task is to apply a range of statistical and data visualization techniques to uncover meaningful patterns and relationships within this data.

ASSIGNMENT TASK

Part A: Descriptive Analysis – Age Group and Frequency of Purchases (25 marks)

Where necessary, values should be presented in 2 decimal places.

1. Create a new column, “Age Group”, using an appropriate Excel function that contains the following categories:
   - 18 – 24
   - 25 – 34
   - 35 – 44
   - 45 – 54
   - 55 – 64
   - 65 or older

   Ans:

   VLOOKUP function was used to categorize the age accordingly into age groups of 18 – 24, 25 – 34, 35 – 44, 45 – 54, 55 – 64, and 65 or older. Function was written as =VLOOKUP(B2,$N$4:$O$9,2). The same function but with different cell referencing was used to group the age for the rest of the observation. The table used to reference the VLOOKUP function is given below:

Start Age	Age Group
18	18-24
25	25-34
35	35-44
45	45-54
55	55-64
65	65 or older

2. Using the appropriate Excel function, obtain a frequency distribution for Age Group by filling in Table 1.

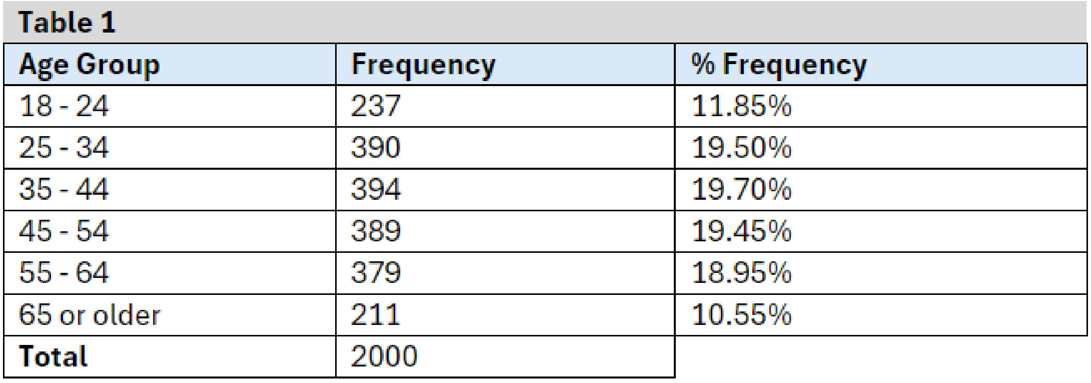


Table 1

Age Group	Frequency	% Frequency
18 - 24	237	11.85%
25 - 34	390	19.50%
35 - 44	394	19.70%
45 - 54	389	19.45%
55 - 64	379	18.95%
65 or older	211	10.55%
Total	2000	

Function used to find the frequency is COUNTIF by refencing the 'n=2000' sheet. The total of frequency was calculated using the SUM function. Then the frequency was calculated by diving the specific frequency from each age group and divided by the sum of frequency of all age groups.

Before attempting the subsequent questions, sort the data by Age Group in ascending order. Filter the data for those aged below 35 years and copy and paste these data into the sheet named "Below 35". Do the same for observations for those aged 35 and above into the sheet named "35 and above".

3. Obtain two frequency distributions for the Frequency of Purchases by using the appropriate Excel function – one for the "Below 35" category and another for the "35 and above" category. Fill in your answers in Table 2. In your own words, explain why using percentage frequency (i.e. proportion) is better than using frequency when describing categories.

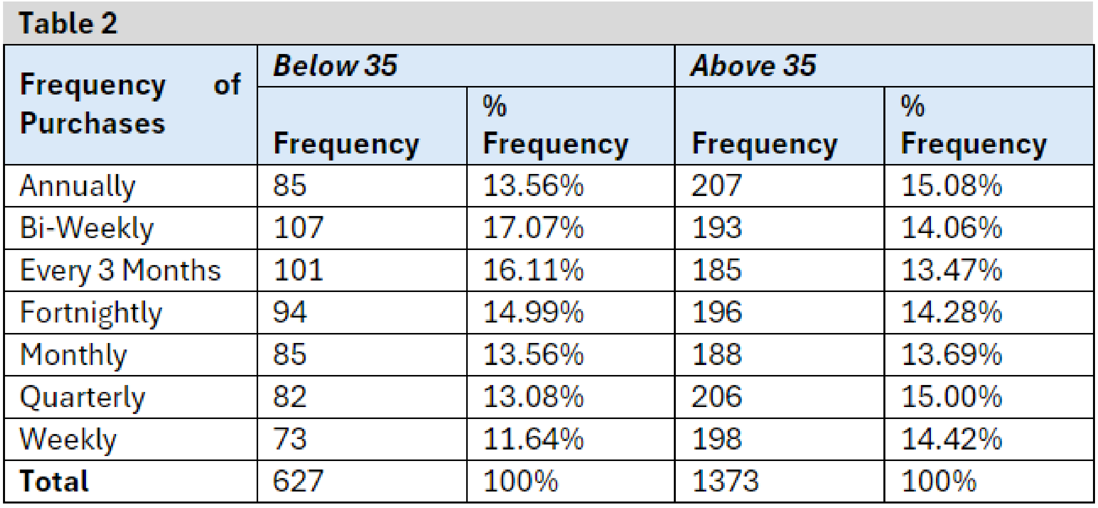


Table 2

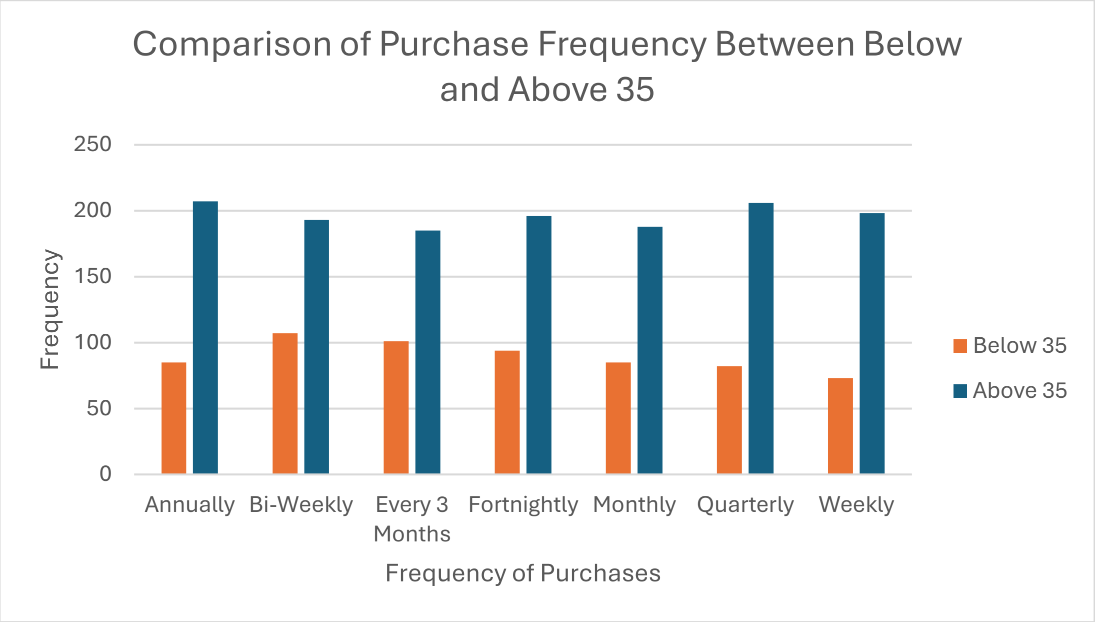
Frequency of Purchases	*Below 35*		*Above 35*	
	Frequency	% Frequency	Frequency	% Frequency
Annually	85	13.56%	207	15.08%
Bi-Weekly	107	17.07%	193	14.06%
Every 3 Months	101	16.11%	185	13.47%
Fortnightly	94	14.99%	196	14.28%
Monthly	85	13.56%	188	13.69%
Quarterly	82	13.08%	206	15.00%
Weekly	73	11.64%	198	14.42%
Total	627	100%	1373	100%

Function used to find the frequency of purchases for both age group of below 35 and above 35 is COUNTIF by refencing the ‘Below 35’ and ’35 and Above’ sheet respectively. The total of frequency for both age group was calculated using the SUM function. Lastly, the percentage frequency was calculated by dividing the frequency of frequency of purchases with the total of the frequency using the respective age group.

Using percentage frequency allows easier comparison among categorical data no matter the sample size while providing a clear view on the relative size of each category within the whole. For example, from Table 2, it shows the frequency and percentage frequency of purchasing habits into different time frames, such as annually, bi-weekly, every 3 months, fortnightly, monthly, quarterly, and weekly, made by individuals below 35 and 35 and above the age groups. The percentage frequency allows direct comparison between the two age groups. For example, it is easier to compare between age groups where 13.56% of customers below the age of 35 make purchases annually whereas 15.08% of customers above the age of 35 also do the same which indicates a slightly higher purchasing frequency among the age group of 35 and above. Furthermore, the purchase frequency for bi-weekly shows a significant difference, where 17.07% is from below 35 age group compared to 14.06% from above 35 age group and so on. This highlights the trends in the behaviour of the consumer by representing how habits of purchasing can vary by age. Not only it is easier to compare among the age groups, but percentage frequency also helps to identify the highest and lowest frequency. For instance, the most frequently bought by customers is Bi-Weekly

(17.07%) and the lowest is Weekly (11.64%) for the age group of below 35. For the age group of 35 and above, the highest frequency is Annually (15.08%) and the lowest is Every 3 Months (13.47%). To conclude, the use of percentage frequency increases the understanding and interpretation of the data, which makes it easier to identify which age group are more likely or not likely to purchase frequently across different intervals.

4. Use an appropriate chart to visualize the information in Table 2.



5. Based on Part A: Questions 3 and 4, discuss the relationship between Age Group and Frequency of Purchases.

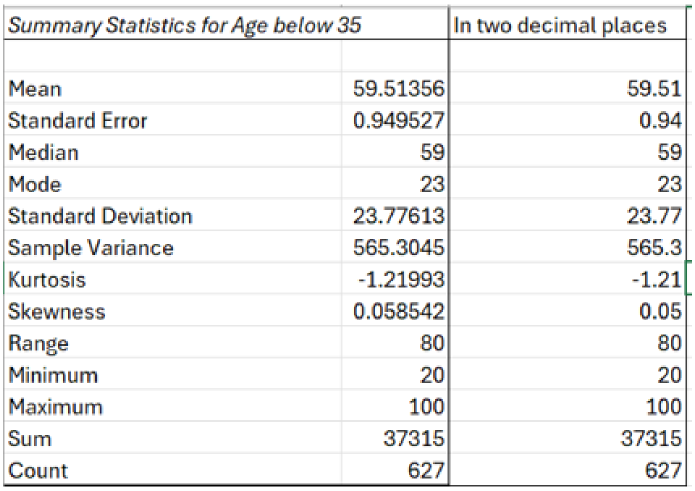
Based on the Table 2 and the chart generated, more purchases have been made by the age group of above 35 years old compared to the age group of below 35 years old across all the categories. Consumers from the age group of below 35 has made the highest purchases Bi-Weely with a frequency percentage of 17.07% whereas the lowest purchases made was Weekly with a frequency percentage of 11.64%. Meanwhile, consumers from the above 35 age group have made highest frequency of purchases Annually with a percentage of 15.08% whereas the lowest frequency purchases were made Every 3 Months with a percentage of 13.47%.

Part B: Descriptive Analysis – Age Group and Purchase Amount

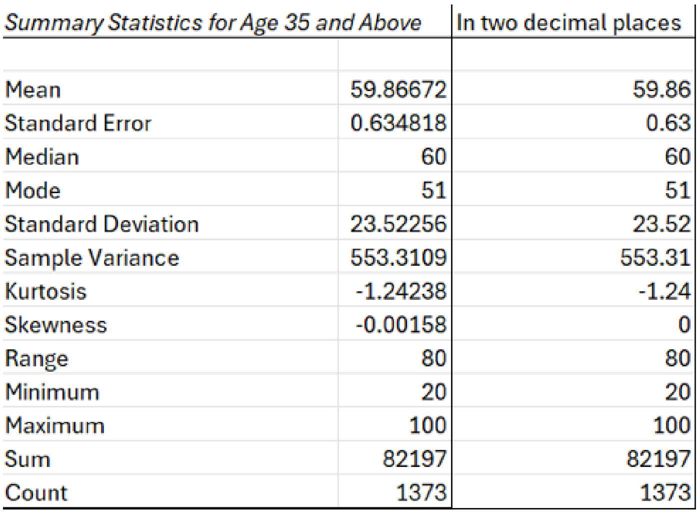
Where necessary, values should be presented in 2 decimal places.

1) Obtain the summary statistics for Purchase Amount for the "Below 35" and "35 and above" categories. This analysis can be done in the respective sheets containing the data for the two categories.

   Data analysis tool in Excel has been used to find out the summary statistics for both age groups.



Summary Statistics for Age below 35		In two decimal places
Mean	59.51356	59.51
Standard Error	0.949527	0.94
Median	59	59
Mode	23	23
Standard Deviation	23.77613	23.77
Sample Variance	565.3045	565.3
Kurtosis	-1.21993	-1.21
Skewness	0.058542	0.05
Range	80	80
Minimum	20	20
Maximum	100	100
Sum	37315	37315
Count	627	627



Summary Statistics for Age 35 and Above		In two decimal places
Mean	59.86672	59.86
Standard Error	0.634818	0.63
Median	60	60
Mode	51	51
Standard Deviation	23.52256	23.52
Sample Variance	553.3109	553.31
Kurtosis	-1.24238	-1.24
Skewness	-0.00158	0
Range	80	80
Minimum	20	20
Maximum	100	100
Sum	82197	82197
Count	1373	1373

2) Transfer the information from Part B: Question 1 into Table 3. Fill in the remaining information in Table 3 using the relevant Excel functions and formulas.

Information from the summary statistics of both age groups has been transferred to Table 3. The remaining information (First Quartile, Third Quartile, Interquartile range (IQR) and Coefficient of variation) has been calculated using Excel function.

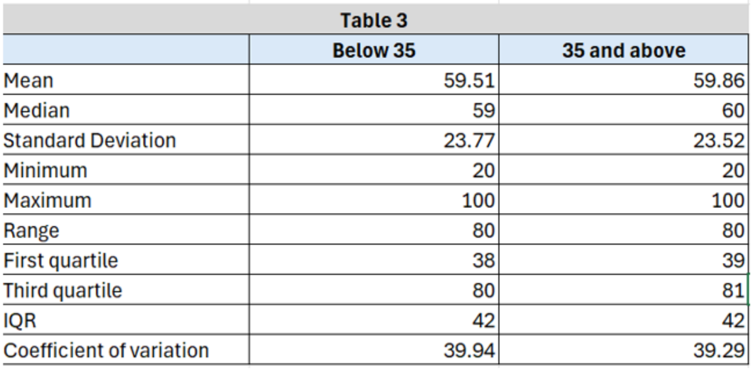


	Table 3	
	Below 35	35 and above
Mean	59.51	59.86
Median	59	60
Standard Deviation	23.77	23.52
Minimum	20	20
Maximum	100	100
Range	80	80
First quartile	38	39
Third quartile	80	81
IQR	42	42
Coefficient of variation	39.94	39.29

3) Discuss and compare the measures of location for the two age categories.

The median purchase amount for the age group below 35 is 59, whereas for the age group 35 and above is higher than the age group below 35 which is 60.

The mean purchase amount for the below 35 age category is 59.51, whereas for the 35 and above age category, it is higher than the below 35 age category which is 59.87. This shows that customers in the older age group spend more than those in the younger age group.

4) Discuss and compare the measures of dispersion for the two age categories.

The range of purchase amounts for customers in both age groups (age below 35, age 35 and above) is the same, which is $20 to $100. This shows that the minimum amount spent by the customers is $20 while the maximum amount is $100.

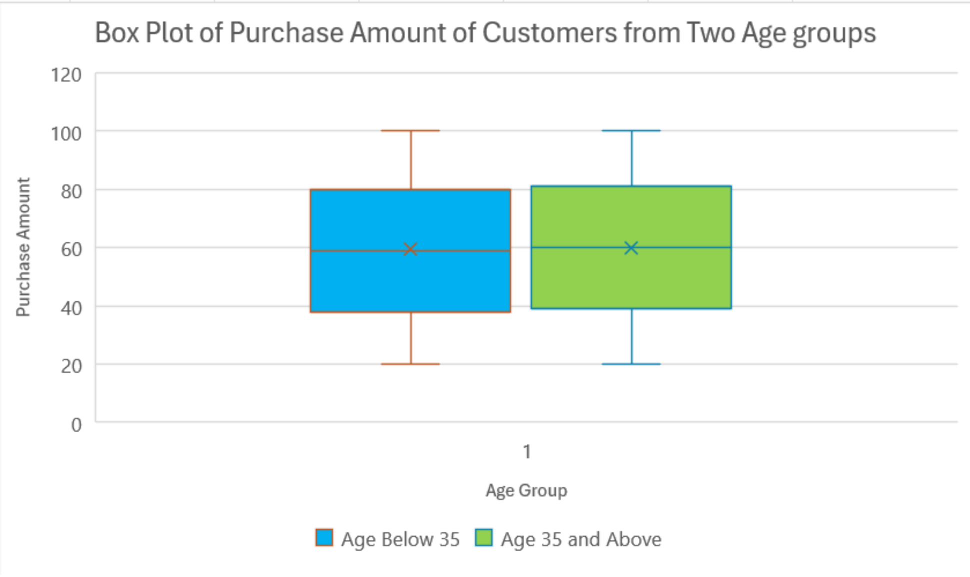
In the below 35 age group, its first quartile ($38) and third quartile ($80) shows that 50% of the customer made purchases between $38 and $80, while in the 35 and above 35 age group, its first quartile ($39) and third quartile ($81) shows that 50% of the customer made purchases between $39 and $81. The interquartile range for both age groups is the same, which is $42. This shows that 50% of data spread out over a range of $42.

The standard deviation for the customers aged below 35 is $23.776, while for age 35 and above is $23.523. This shows that there is a similar spread in purchase amounts, but it is more spread out in the younger age group compared to the older age group. Also, the variability in purchasing behaviour of customers in both age group is almost the same, but it is higher in the age group of below 35.

The coefficient of variance for customers in below 35 age group is 39.95%. It means that the standard deviation is approximately 39.95% of the mean. This shows that it has a high variability in purchase amounts compared to the average purchase amount. While the coefficient of variance for customers in 35 and above age group is 39.29%. It means that the standard deviation is approximately 39.29% of the mean. This also shows that it also has a high variability in purchase amounts compared to the average. To conclude, the purchase amount or spending behaviour of the younger age group is slightly more inconsistent than the older age group.

5) Use an appropriate chart to visualize the distribution of Purchase Amount for the two age categories. Ensure that you provide an interpretation of the chart.

Box Plot has been created to visualize the distribution of Purchase Amount for the two age categories.



The box plot of purchase amount of customers from age group of below 35 shows minimum (20), maximum (100), first quartile (38), median (59) and third quartile (80). It does not have any outliers outside the box. Since the box is wider from the median to the third quartile than from the first quartile to the median, the box plot is somewhat right-skewed (positive distribution). This means that most of the data which is the purchase amount appears below the median.

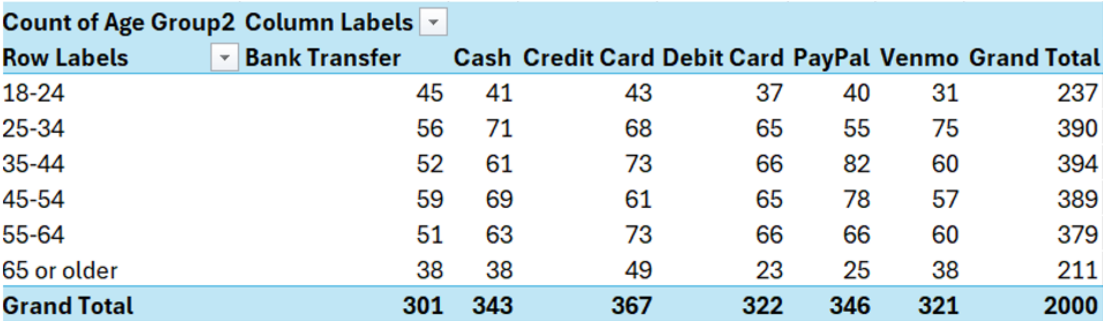
The box plot of purchase amount of customers from age group of 35 and above shows minimum (20), maximum (100), first quartile (39), median (60) and third quartile (81). It does not have any outliers outside the box. Since the median is appear in the middle of the box, the box plot is symmetry. This means that the data (purchase amount) spread evenly on both sides of the median.

Part C: Probability – Age Group and Payment Method

Answers for this part should be presented in 3 decimal places. Ensure that you show all calculations.

1. Construct a PivotTable for Age Group and Payment Method in a new sheet and name this sheet “Part C (PivotTable)”.

   Pivot table for age group and payment method:



Count of Age Group2	Column Labels						
Row Labels	Bank Transfer	Cash	Credit Card	Debit Card	PayPal	Venmo	Grand Total
18-24	45	41	43	37	40	31	237
25-34	56	71	68	65	55	75	390
35-44	52	61	73	66	82	60	394
45-54	59	69	61	65	78	57	389
55-64	51	63	73	66	66	60	379
65 or older	38	38	49	23	25	38	211
Grand Total	**301**	**343**	**367**	**322**	**346**	**321**	**2000**

2. What is the probability that a randomly selected customer is between 45 – 54 years of age?

   P(45 – 54 years of age) = 389/2000 = 0.1945

3. What is the probability that a randomly selected customer uses a credit card as their payment method?

   P(credit card) = 367/2000 = 0.1835

4. What is the probability that a randomly selected customer is between 18 – 24 years old and uses Venmo as their payment method?

   P(18 – 24 years old and uses Venmo) = 31/2000 = 0.0155

5. What is the probability that a randomly selected customer uses a credit card as their payment method, given that the customer is between 18 – 24 years old? Are these events independent?

P(credit card | 18 – 24) = P(credit card)
P(credit card | 18 – 24) = 43/237 = 0.1814
P(credit card) = 0.1835

P(credit card | 18 – 24) ≠ P(credit card)

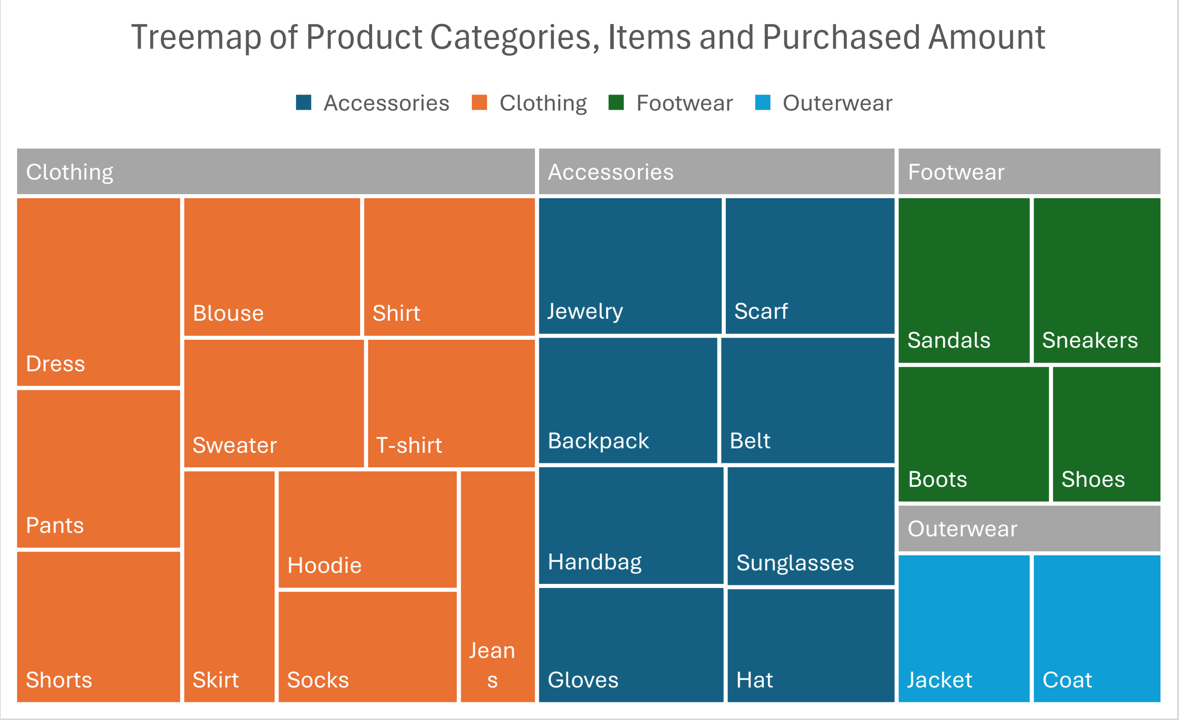
∴ Hence, these events are not independent.

Part D: Visualization – Category, Item Purchased, and Purchase Amount

This part requires you to do some additional research. Use a new Excel sheet to answer this part and name the sheet "Part D".

1. There are four main categories in the dataset: Accessories, Clothing, Footwear, and Outerwear. Each of these categories contains a variety of specific items that customers have purchased. Create a comprehensive visualization that presents a hierarchical view of the product categories, and their specific items along with the purchase amount for each item. You should avoid using bar and column charts in your visualization. Note: This should be answered using a single chart.
2. Provide some insights into the visualization produced.

Graph analysing categories with their specific items.



From the tree map, it is represented that the category of Clothing has the highest purchase amount among other categories, which shows that customers are likely to buy clothing compared to accessories, footwear and outerwear. Meanwhile, Outwear category has the lowest sale. Furthermore, Dress from Clothing category is the most popular items which has the highest purchase amount compared to the other items. Whereas the least popular and lowest frequently purchased item is Shoes from Footwear category.