KINGSTON UNIVERSITY LONDON

DATA ANALYTICS AND VISUALISATION (CI7330)

COURSEWORK - I

NAME	Sujay Grama Suresh Kumar
KU ID	K2201621
EMAIL ID	K2201621@kingston.ac.uk
MODULE LEADER	Dr. Rosie McNiece

[1] Descriptive Statistics

i. Quantitative Data

1. Spend (Amount spent on weekly supermarket shop):

- On average, families spend around **£104.88** per week. This gives us a sense of the typical expenditure.
- Most families spend vary within **£16.39** of the average. This suggests that spending habits are relatively consistent.
- The spending range varies from **£50.49 to £146.71**. So, while most are close to the average, there is a notable spread in spending extremes.
- A significant number of families, about half of them (middle 50%), fall within a spending **range of £21.63**. This provides us with an understanding of what is considered a common spending range for the majority.

2. Qtime (Time spent queuing at checkout/collection point):

- On average, families spend about **6.98 minutes** waiting. This provides an understanding of the typical waiting time.
- Most wait times vary within **2.68 minutes** of the average. Indicating that waiting times are generally consistent.
- Wait times range from **1 to 17 minutes**. There's a considerable spread, with some experiencing longer waits.
- For the majority of families, about half of them (middle 50%), wait times fall within a **range of 4 minutes**, providing insight into the common waiting experience.

3. Income (Estimated household income):

- The average household income is **£38,525.57** per year. This gives us a snapshot of the typical income.
- Incomes vary around **£4,667.13** from the average. There's some diversity in income levels.
- Incomes range from **£23,164.00 to £52,369.89**. There's a substantial range in income levels within the surveyed population.
- For the majority of households, about half of them (middle 50%) , incomes fall within a **range of £6,511.82,** providing insight into the common income distribution.

4. Famsize (Number of household members in addition to the buyer):

- The average family size is **4 members**. This gives us an idea of the typical family composition.
- Family sizes vary, with most falling within **2 members of the average**. Indicating that family sizes are generally close to the average.
- Family sizes range from **0 to 10 members**. There's a wide range of family sizes within the surveyed population.
- For the majority of families, about half of them (middle 50%) , family sizes fall within a **range of 3 members**, providing a sense of the common family size.

ii. Categorical Data

5. Store (Store type):

- **Most families shop in Urban stores**. This indicates a preference for urban shopping environments (186 Observations).

6. Shoptype (Shopping type):

- **The majority of families prefer in-store shopping**. Highlighting a prevalent inclination towards traditional in-store shopping (363 Observations).

Variables	Mean	Median	Standard Deviation (SD)	Range (Min-Max)	IQR Range (50%)
Spend	104.8811 £	104.8154 £	16.3906 £	50.491 £ – 146.710 £	21.6302 £
Q-Time	6.976 min	7 min	2.682 min	1 min – 17 min	4 min
Income	38525.57 £	38449.73 £	4667.129 £	23164.00 £ – 52369.89 £	6511.815 £
Famsize	4.004	4	1.987	0 – 10	3

Table 1 Descriptive Statistics for Quantitative Variables

[2] Visualisation: 1

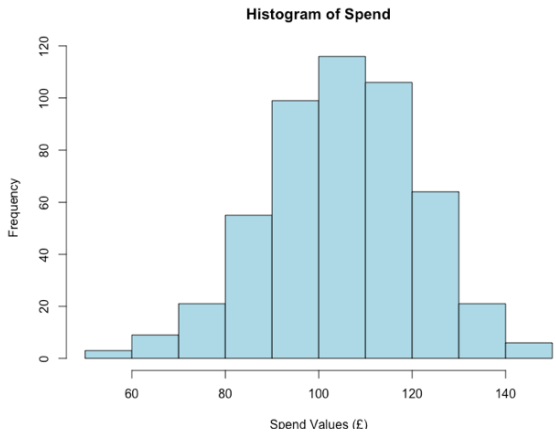

Figure 1 Weekly Spending Distribution (Univariate)

For this assignment, the chosen variable to visualize is the **"spend"** variable, representing the amount people spend on their weekly supermarket shopping.

Summary:

- The Spend Range: The x-axis shows how much money people spend, ranging from 0 to 150 pounds.
- Frequency: The y-axis indicates how many people fall into each spending range.

Observation

- Common Spending Range: The tallest bars is in the centre has spending values between 90 to 110 pounds. This means that a lot of people spend around this amount each week.
- Less Common Spending (Outliers): Towards the edges (left and right sides), very few people spend either less than 80 pounds or more than 120 pounds weekly.

Interpretation:

- Most Common Spending: The majority of individuals tend to spend around 90 to 110 pounds on their weekly supermarket shopping.
- Variation in Spending: Although many people fall within this typical spending range, there are some who spend less or more than the average.

[3] Visualisation: 2

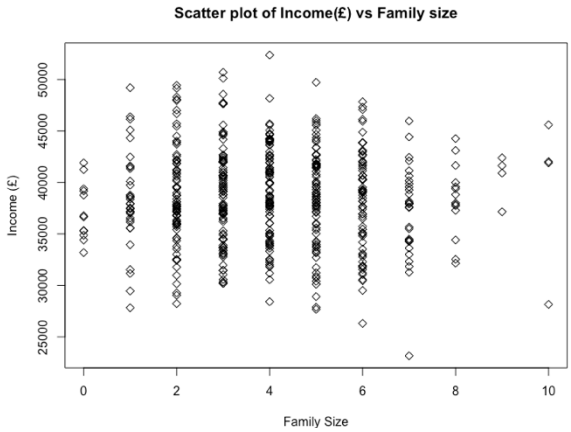

Figure 2 Income (£) vs Family Size (Bivariate)

Summary:

The scatter plot of **Income vs. Family Size** provides a visual overview of how family size relates to income. It is observed that some interesting patterns such as, where most families fall in terms of size and income, as well as some exceptional cases that stand out.

Observation:

1. Where Most Families Are:

- Families with 2 to 5 members are the most common, with incomes typically ranging between £35,000 and £46,000.
- In this range, a bunch of circles overlaps, indicating that many families within the dataset share similar income levels

2. Income Spread for Different Family Sizes:

- Smaller families (size 1) and larger families (size 7) can have varied incomes, ranging from £30,000 to £43,000.

- The scatter plot reveals that income is not solely determined by family size; there's a mix of incomes within each size category.

3. Outliers:
   - An outlier was observed in the dataset, specifically a family with four members earning an unusually high income, potentially above £50,000
   - This family is exceptional and stands out from the usual income patterns in the plot.

4. Larger Families Have Moderate Incomes:
   - Families with 8, 9, and 10 members are less common, but they still have moderate incomes, typically between £35,000 and £44,000.
   - Even though there are fewer of these families, they show a similar income range as smaller families.

5. Exceptional Case with Lower Income:
   - There's a unique case where a family with 7 members has a lower income, falling below £25,000.
   - This family size is an exception, as most families with similar sizes have higher incomes.

Interpretation:
The scatter plot simplifies the relationship between family size and income. Families with 2 to 5 members typically have incomes in a specific range, but there are outliers. For instance, a family of 4 stands out for its exceptionally high income, while a family of 7 has a lower income compared to others of similar size. This visual representation aids in understanding these patterns without delving into complex statistical details.

[4] The solution below thoroughly examines the association between the continuous/quantitative variable "spend" and the categorical variable "store".

Selected Method: ANOVA
Hypotheses:
- Null Hypothesis (H0): There is no significant difference in the mean spend across different levels of store.
- Alternative Hypothesis (H1): There is a significant difference in the mean spend across different levels of store.

Rationale for Chosen Method (ANOVA):
- Nature of Variables: ANOVA is suitable when examining the relationship between a continuous variable (spend) and a categorical variable with more than two levels (store), which matches the scenario.
- Number of Groups: ANOVA allows for the comparison of means across multiple groups, making it appropriate for analysing the association between spend and the categorical variable store.

ANOVA Results:

```
> #Annova
> group = factor(weeklyshop$spend)
> out = aov(weeklyshop$store~weeklyshop$spend)
> summary1 <- summary(out)
> p_value <- summary1[[1]]$"Pr(>F)"[1]
> summary(out)
                  Df Sum Sq Mean Sq F value Pr(>F)
weeklyshop$spend   1    2.7  2.6685   3.937 0.0478 *
Residuals        498  337.5  0.6778
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> cat("P-value:", p_value, "\n")
P-value: 0.04777853
```

Figure 3 ANOVA test results

- P-value: The p-value from the ANOVA test is 0.0478, which is less than the significance level of 0.05 ($P<0.05$).
- Conclusion: We reject the null hypothesis (H0). There is evidence to suggest a significant difference in the mean spend across different levels of the store.

T-Test Consideration:
- The t-test was not appropriate in this case because it requires exactly two levels in the grouping variable. Since "store" has more than two levels, a t-test cannot be performed directly on this variable.

Chi-squared Test Consideration:
- The chi-squared test was not appropriate in this case because it is designed for analysing the association between two categorical variables. Here, "store" is a categorical variable, but "spend" is continuous.

Overall Conclusion:
Based on the results of the **ANOVA test**, there is a significant association between spend and store. The choice of ANOVA was appropriate due to the nature of the variables involved, and the rejection of the null hypothesis suggests that the mean spend varies significantly across different levels of the store.

[5] Correlation

- To investigate the association using correlation, the choice of a pair of variables is based on selecting those likely to have a linear relationship. The correlation results are examined, and the pair with a stronger correlation is identified.

Correlation Analysis:

Correlation Pairs	Correlation Coefficient
Spend and Qtime	**-0.06504031**
Spend and Income	**-0.0342992**
Qtime and Famsize	**-0.04921866**
Famsize and Income	**-0.01088618**

Choice of Variables:
While none of the correlations are very strong, Correlation between "Spend" and "Qtime" has the strongest negative correlation among the options.

Interpretation and Explanation:

Chosen Variables: Spend and Qtime

Correlation Coefficient: -0.06504031

Interpretation:

- The negative correlation suggests a weak linear relationship between the amount spent and the shopping time.
- As shopping time increases, the amount spent tends to decrease slightly. However, the correlation is weak, indicating that other factors likely contribute to the variability in spending.

Consideration for Other Variables:

- Correlations 2, 3, and 4 show very weak or close to zero correlations, indicating a limited linear relationship between the respective pairs of variables.

Overall Conclusion:
Chosen Pair: Spend and Qtime
Explanation: The negative correlation suggests a weak inverse relationship between the amount spent and shopping time, with spending tending to decrease slightly as shopping time increases.
Choosing variables that exhibit a more substantial correlation helps in identifying potential patterns or relationships between variables. In this case, the correlation between Spend and Qtime, while weak, is the strongest among the options provided.