# Part 1 : Business Analytics : Predicting Potential Return

The business questions are solved using the Cross-Industry Standard Process Model for Data Mining (CRISP-DM). This process model applies to predictive analytics and will be of great assistance for a structured approach.

## 1. Business Understanding

The objectives of this business problem are:

- To determine the factors that cause customers to spend more or less on the website's online platform.
- To predict customer spending on the website by building a model and evaluating the model with a new dataset of additional customers.

The data used is the data provided by the PhysicalSound Company with sample orders and additional customers to predict the spending.

To measure the business objectives:

- Descriptive statistics, visualisation, correlation analysis and regression analysis were used to determine the factors.
- R-Squared($R^2$) and Additional R-Squared.

The categorical variables have been encoded as numerical variables. All the variables are quantified and can be used directly.

The key goal is to evaluate different groups of customers by building a predictive model in order to increase customer spending and estimate future revenue.

## 2. Data Understanding

### 2.1 Data Collection

The company provided two data files.

The order_july24.csv file has 2000 orders with 6 variables.

The new_customer24.csv file has 20 additional orders with 6 variables.

Instead of 'spend', it has 'order'.

The unit of analysis is the individual customer order while the target variable is the spending('spend').

There are 5 dependent variables: ‘past_spend’, ‘age’, ‘ad_channel’, ‘time_web’, and ‘voucher’.

There are 2 categorical variables: ‘ad_channel’ and ‘voucher’. The rest of the 4 variables are quantitative.

### 2.2 Data Exploration

For the quantitative variables, data is summarised by the measures of central tendency and dispersion.

| Variables | Mean | Median | Mode | Range | Variance | Standard Deviation |
|---|---|---|---|---|---|---|
| **spend** | 40.14648 | 40 | 43 | 66 | 96.42022 | 9.819380 |
| **past_spend** | 13.01975 | 12 | 0 | 51 | 110.0832 | 10.49205 |
| **age** | 34.13998 | 34 | 34 | 28 | 15.77687 | 3.972011 |
| **time_web** | 81.21324 | 83 | 92 | 124 | 275.3044 | 16.59230 |

spend:

The mean and median are quite close, and the distribution is positively skewed. Most customers spent 43 GBP. The difference between the highest and lowest spending amount is 66 GBP. The variance and standard deviation are high which shows greater variability, or presence of an outlier.

past_spend:

The mean and median are quite close, and the distribution is positively skewed. Most customers have not spent anything as the mode is 0 GBP. The difference between the highest and lowest spending amount is 51 GBP. Hence, the variance and standard deviation are also quite high.

age:

The mean and median are quite close, and the distribution is positively skewed. Most customers are close to 34 years old. The difference between the age of highest and lowest spending amount is 28 years.

time_web:

The median is greater than the median and the distribution is negatively skewed. Most customers spent 92 seconds on the website before making a purchase. The difference between the highest and lowest amount of time spent is 124 seconds.  The variance and standard deviation are quite high which suggest that some customers spent more time on the website than the others.

Frequency and relative frequency distribution are used for categorical variables.

ad_channel:

| Measures | Frequency | Relative Frequency |
|---|---|---|
| **Leaflet (1)** | 514 | 0.2570 |
| **Social media (2)** | 543 | 0.2715 |
| **Search engine (3)** | 519 | 0.2595 |
| **Influencer (4)** | 408 | 0.2040 |

Most customers visit the website through social media. The next popular advertisement channels are search engines and leaflets. Influencers seem to be the least popular or effective among customers.

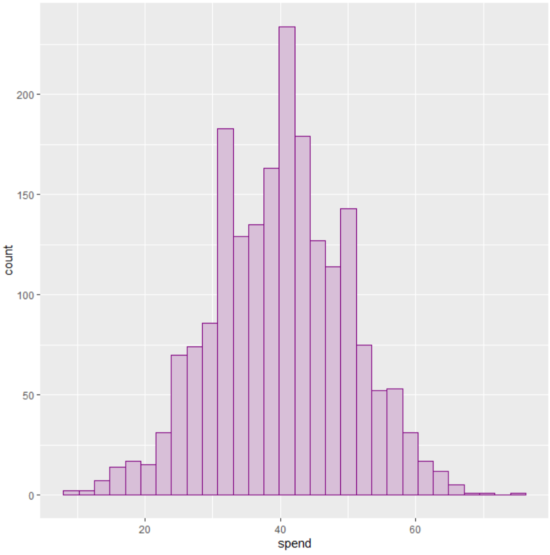
voucher:

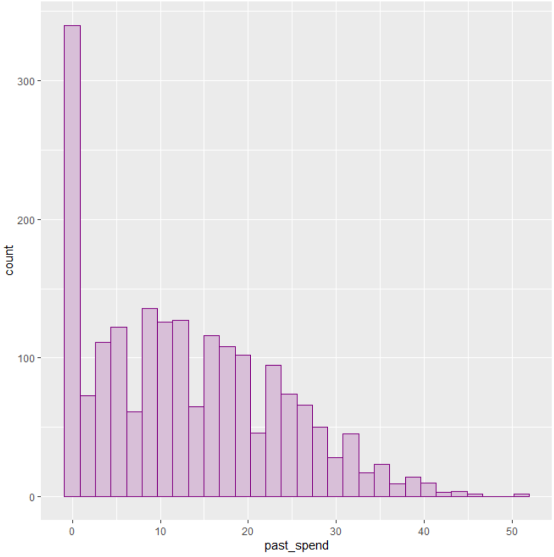
| Measures | Frequency | Relative Frequency |
|---|---|---|
| **Used (1)** | 459 | 0.2295 |
| **Not used (2)** | 1517 | 0.7585 |

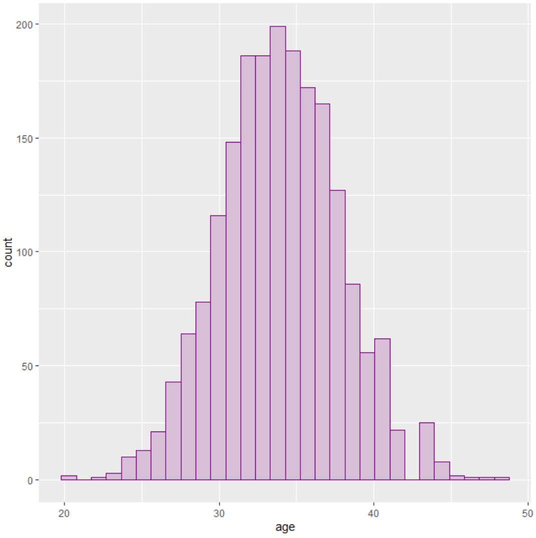
Around 75% of the customers have not used the 5% discount voucher. This indicates that either they are not popular, or the customers do not want to take advantage of the vouchers due to some reasons.

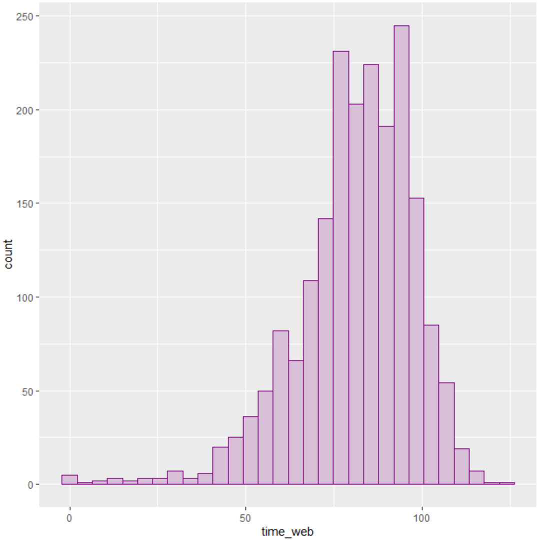
### 2.3 Data Visualisation

Data visualisation also helps in understanding and summarising the variables and provides some key insights.

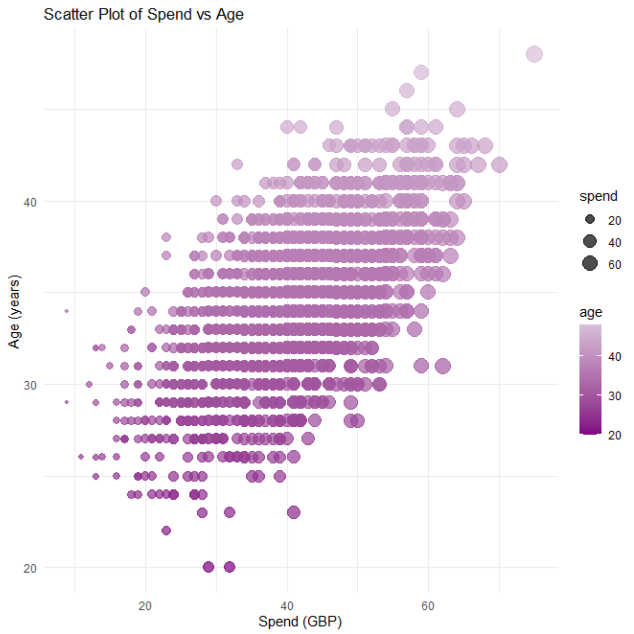


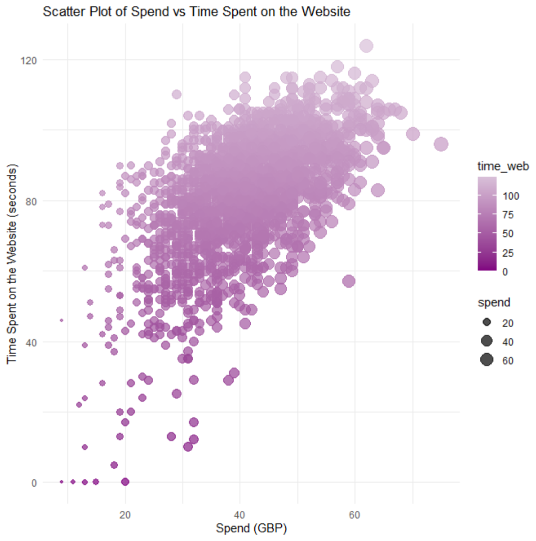


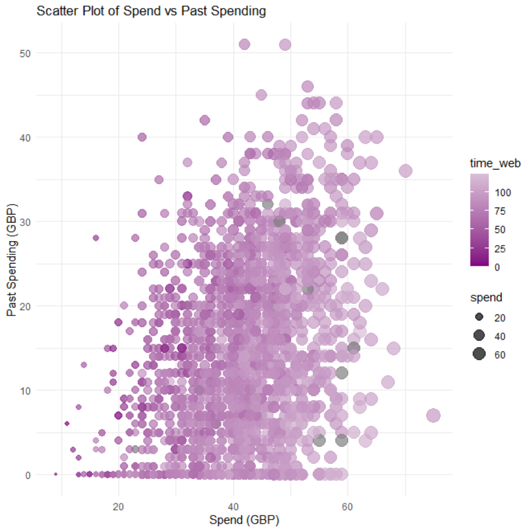




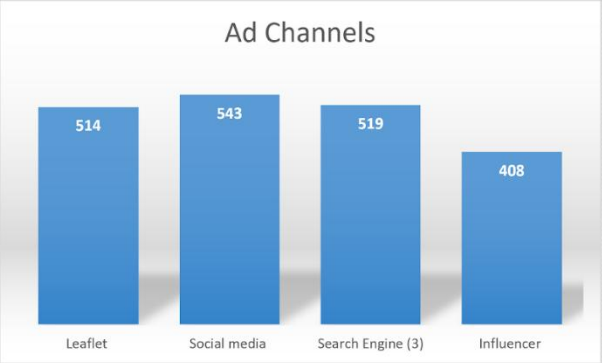
Since the target variable is spend, it will be helpful to construct a scatterplot against other variables to describe their relationship.

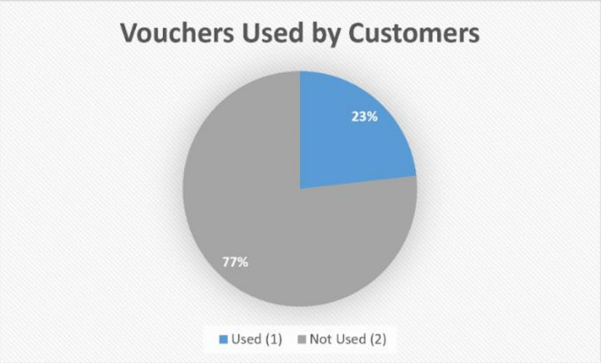

Scatter Plot of Spend vs Age


Scatter Plot of Spend vs Time Spent on the Website


Scatter Plot of Spend vs Past Spending

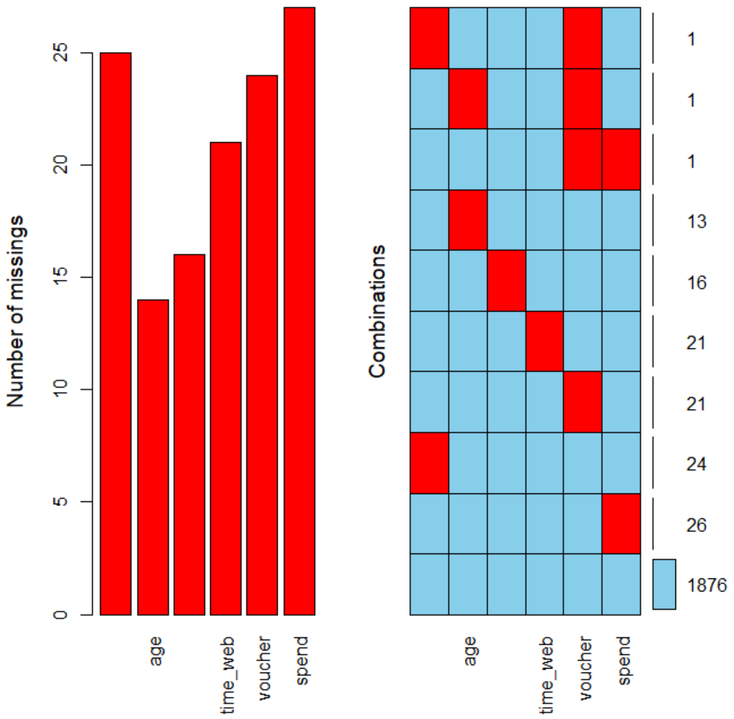
The scatter plots show a positive correlation between customer spending and the age of the customers, time spent by them on the website, and past spending.



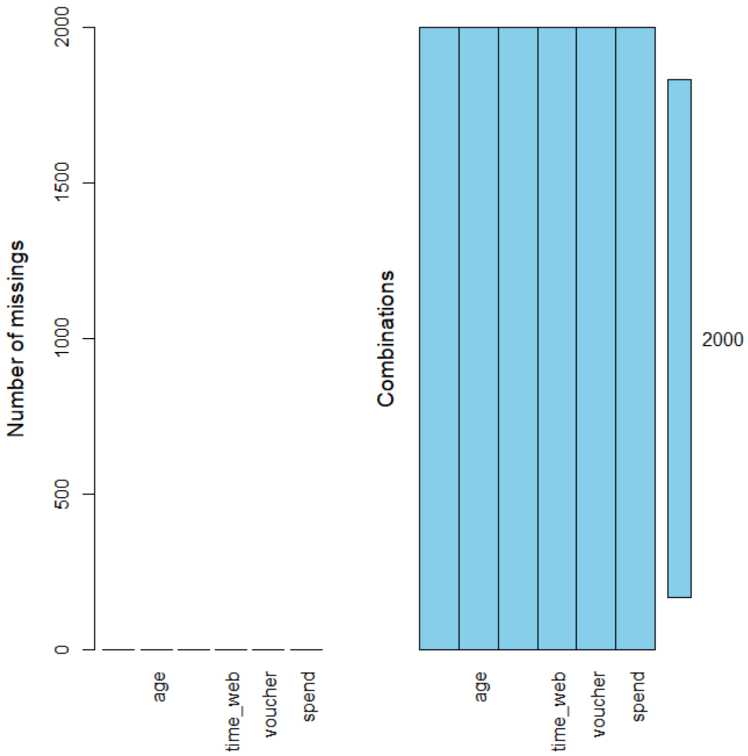


## 2.4 Handling Missing Values

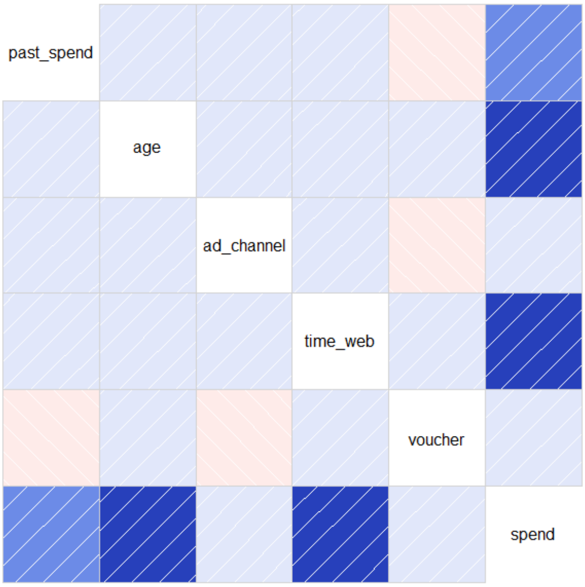
There are 124 rows that are missing data.

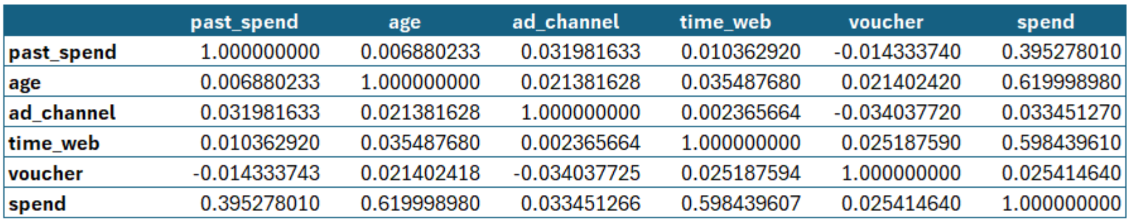


All the variables have high bars suggesting a significant number of missing values. Hence multiple imputations would be of great assistance in handling the missing data while maintaining the data integrity and reducing bias. The approach is also robust and provides accurate results.



It will be helpful to check correlation between the variables.





| | past_spend | age | ad_channel | time_web | voucher | spend |
|---|---|---|---|---|---|---|
| **past_spend** | 1.000000000 | 0.006880233 | 0.031981633 | 0.010362920 | -0.014333740 | 0.395278010 |
| **age** | 0.006880233 | 1.000000000 | 0.021381628 | 0.035487680 | 0.021402420 | 0.619998980 |
| **ad_channel** | 0.031981633 | 0.021381628 | 1.000000000 | 0.002365664 | -0.034037720 | 0.033451270 |
| **time_web** | 0.010362920 | 0.035487680 | 0.002365664 | 1.000000000 | 0.025187590 | 0.598439610 |
| **voucher** | -0.014333743 | 0.021402418 | -0.034037725 | 0.025187594 | 1.000000000 | 0.025414640 |
| **spend** | 0.395278010 | 0.619998980 | 0.033451266 | 0.598439607 | 0.025414640 | 1.000000000 |

- There is a strong correlation between age and spend.
- There is a strong correlation between time_web and spend.
- There is a moderate correlation between past_spend and spend.
- There is a negative correlation between past_spend and voucher.
- There is a negative correlation between ad_channel and voucher.

### 2.5 Assumptions

**Linearity -** The expected value of the dependent variable has a linear relation with the explanatory variable(s).

Figure 1

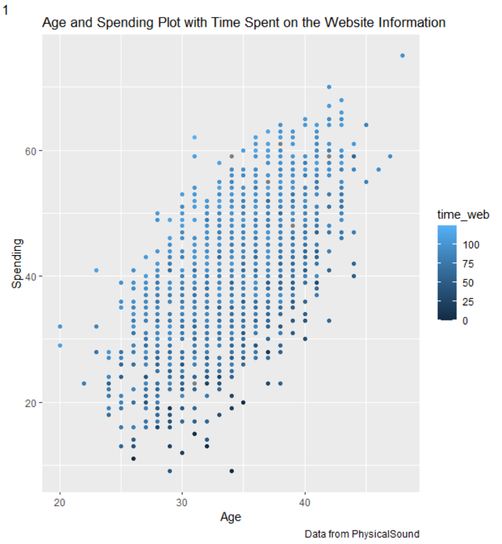
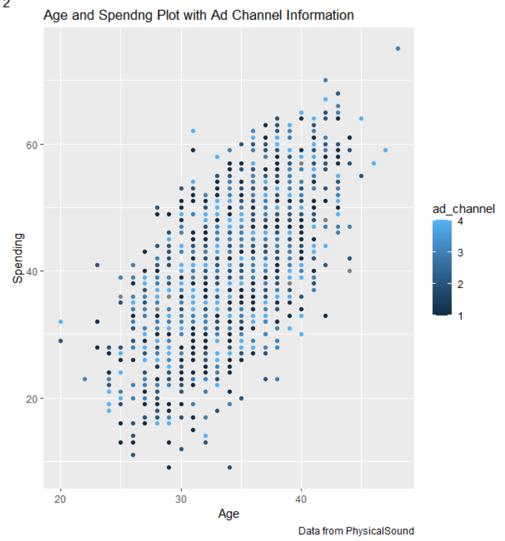
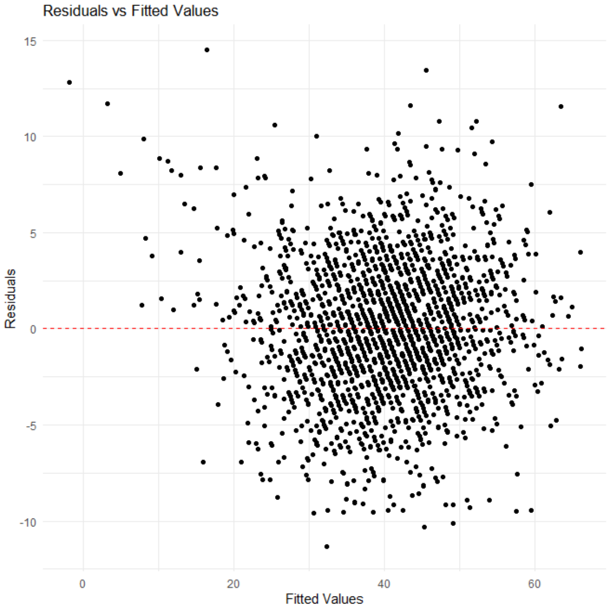


Figure 2



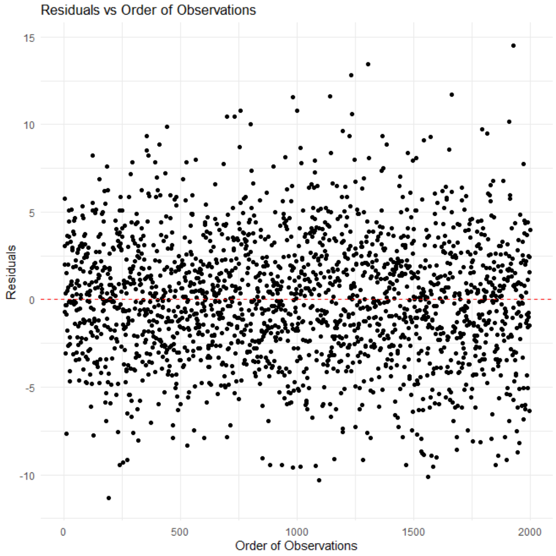
The points form a straight line demonstrating linearity, and the assumption holds good.

**Homoscedasticity -** The variance of the dependent variable should be the same for all values of the explanatory variable(s).



As the residuals are randomly scattered along the horizontal line, the assumption holds good.

**Independence -** The error (distance from the line) of each point should be independent from the error of other points.

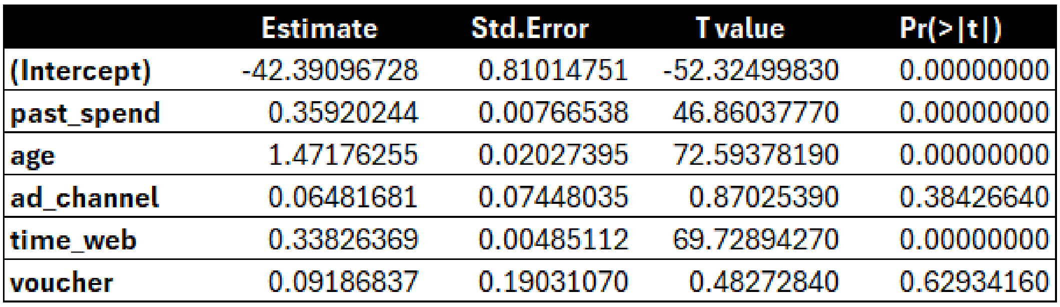


The residuals are randomly scattered with no clear pattern, the assumption holds good.

## 3. Modelling

The target variable is a scalar value of customer spending. Linear regression model is highly suitable for predicting a continuous variable. It is simple, and effective, and serves as a great baseline model. It helps to predict customer spending and provides valuable insights into the factors.

## 4. Evaluation



| | Estimate | Std.Error | T value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -42.39096728 | 0.81014751 | -52.32499830 | 0.00000000 |
| past_spend | 0.35920244 | 0.00766538 | 46.86037770 | 0.00000000 |
| age | 1.47176255 | 0.02027395 | 72.59378190 | 0.00000000 |
| ad_channel | 0.06481681 | 0.07448035 | 0.87025390 | 0.38426640 |
| time_web | 0.33826369 | 0.00485112 | 69.72894270 | 0.00000000 |
| voucher | 0.09186837 | 0.19031070 | 0.48272840 | 0.62934160 |

- The R-squared value is 0.8655424 and the adjusted R-squared value is 0.8652052.
- Thus, the model can explain 86% variation within spending information.
- Since both the values are high and close to 1, the model is fit, and the predictions are meaningful.

## 5. Deployment

- Age is highly significant in relation to spending by 1.4718 years with some deviation of 0.02096.
- Past spending and time spent on the website is significant with 0.3592 GBP and 0.3382 seconds with deviations of 0.00787 and 0.00496.
- The ad channel and voucher used are not statistically significantly related to spending.

## 6. Conclusion

Based on the regression model, the factors affecting customer spending are age, past spending, and time spent on the website.

- Older customers spend more which suggests to develops strategies marketing tactics that targets them. The needs of older customers should be recognised to increase overall spending.

- The customers who spend more time on the website tend to spend more. Hence, website design needs to be optimised, and user experience needs to be enhanced to encourage more spending.

- Customers who spend in the past are likely to spend again, suggesting a loyal customer base. Loyalty and referral programs along with customer service would increase the spending.

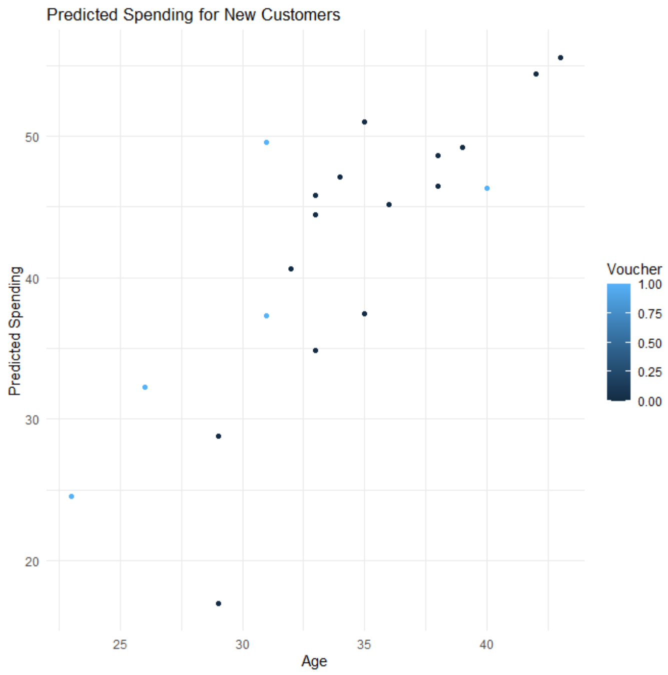
- The advertisement channel that attract0ed customers to the website has no impact on their spending. This suggests that their decisions are more influenced by other factors.

- Using voucher has no impact on customer spending, which suggests that other promotions need to be developed.

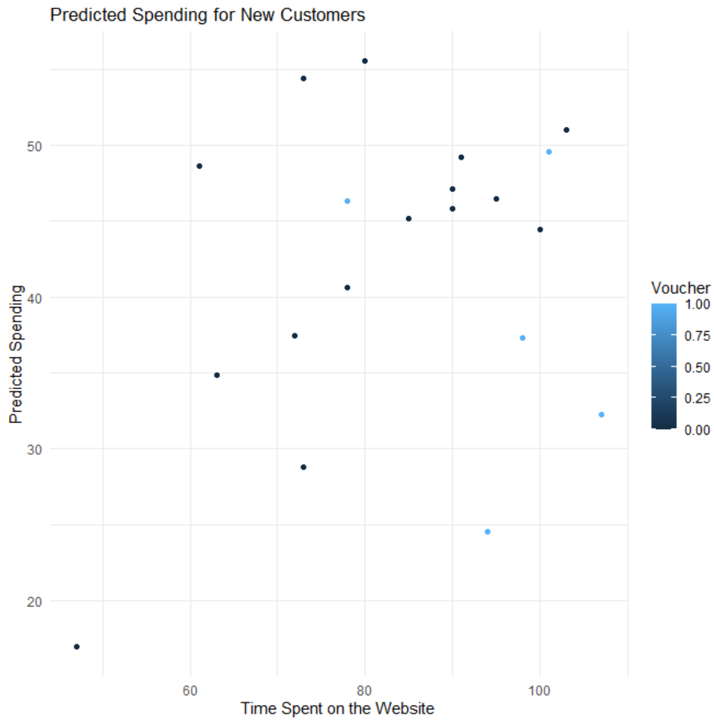
## Training Set

The following table shows the predictions of additional new customers:

| order | prediction |
|---|---|
| 1 | 16.97176 |
| 2 | 45.15695 |
| 3 | 46.31843 |
| 4 | 46.45428 |
| 5 | 34.86631 |
| 6 | 44.44363 |
| 7 | 47.13757 |
| 8 | 54.36843 |
| 9 | 51.04600 |
| 10 | 28.83469 |
| 11 | 49.60312 |
| 12 | 24.55547 |
| 13 | 37.49175 |
| 14 | 40.62372 |
| 15 | 37.32342 |
| 16 | 49.21704 |
| 17 | 45.79544 |
| 18 | 48.60301 |
| 19 | 32.29058 |
| 20 | 55.56399 |



Predicted Spending for New Customers
Predicted Spending
50
40
30
20
25
30
35
40
Age
Voucher
1.00
0.75
0.50
0.25
0.00



Predicted Spending for New Customers
Predicted Spending
50
40
30
20
60
80
100
Time Spent on the Website
Voucher
1.00
0.75
0.50
0.25
0.00

# Part 2 : Business Decision-Making : Trial of Delivery Robot

## Business Question 1

The business question is solved using TOPSIS due to the following reasons:

- The question requires choosing the best option from multiple alternatives, not finding a compromise solution.
- The data requires ranking multiple alternatives based on their proximity to an ideal solution.
- The dataset is relatively small with pre-defined weights based on management priorities – one alternative has a degree of dominance over the other and needs ranking.
- The majority of the criteria are quantitative in nature, with only mobility and aesthetic as subjective criteria.
- Avoid the complexity of pairwise comparisons.

Among all the techniques, TOPSIS is the simpler and more straightforward approach and aligns with the management's clearly stated priorities.

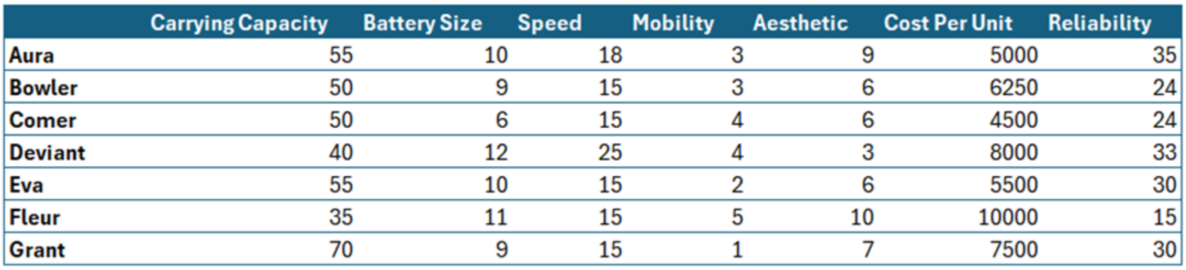
### Data Collection

The 'Robot_Info.csv' provided by the management has 7 rows and 8 columns.

There are 8 criteria: carrying capacity, battery size, speed, mobility, aesthetic, cost per unit, and reliability.

Also, the management has provided an Excel file 'Management_Priority,xlsx' that is used for assigning weights to criteria for business plan 1.

1. **Performance Table**

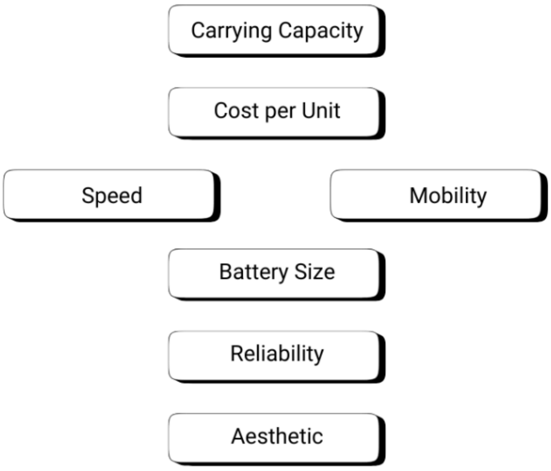


| | Carrying Capacity | Battery Size | Speed | Mobility | Aesthetic | Cost Per Unit | Reliability |
|---|---|---|---|---|---|---|---|
| Aura | 55 | 10 | 18 | 3 | 9 | 5000 | 35 |
| Bowler | 50 | 9 | 15 | 3 | 6 | 6250 | 24 |
| Comer | 50 | 6 | 15 | 4 | 6 | 4500 | 24 |
| Deviant | 40 | 12 | 25 | 4 | 3 | 8000 | 33 |
| Eva | 55 | 10 | 15 | 2 | 6 | 5500 | 30 |
| Fleur | 35 | 11 | 15 | 5 | 10 | 10000 | 15 |
| Grant | 70 | 9 | 15 | 1 | 7 | 7500 | 30 |

2. **Assigning Weights**

**For Business Plan 1:**

According to the management team's primary business model, the importance of factors that impact robot selection is as below:



- Carrying capacity is the most important criterion and aesthetic is the least important criterion.

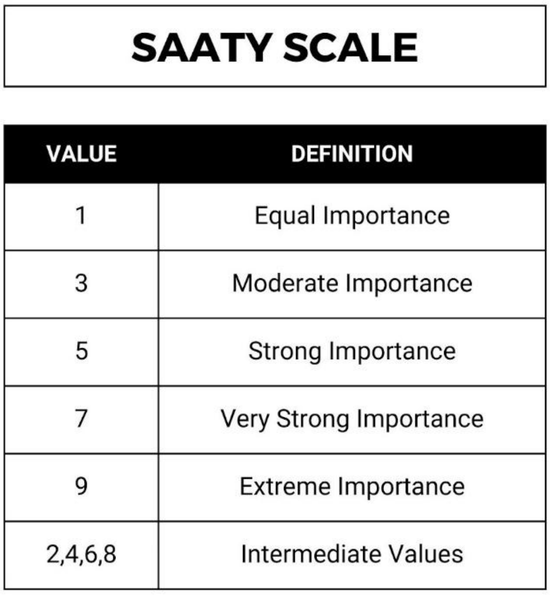
- Carrying capacity is of strong importance over cost per unit, very strong importance over speed, mobility, and battery size, and extreme importance over reliability and aesthetic.

- Cost per unit has strong importance over speed and mobility, very strong importance over battery size and reliability, and extreme importance over aesthetic.

- Speed and mobility are of equal importance. Both have strong importance over battery size and very strong importance over reliability and aesthetic.

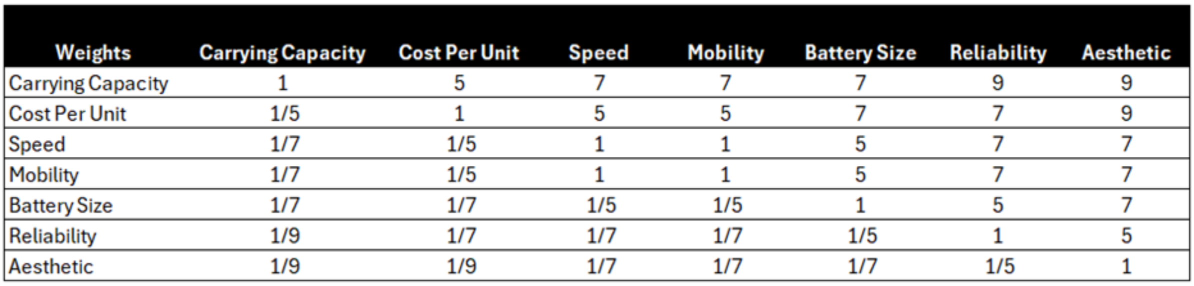
- Reliability has strong importance over aesthetics.

Values are assigned according to relative importance using the Saaty scale.

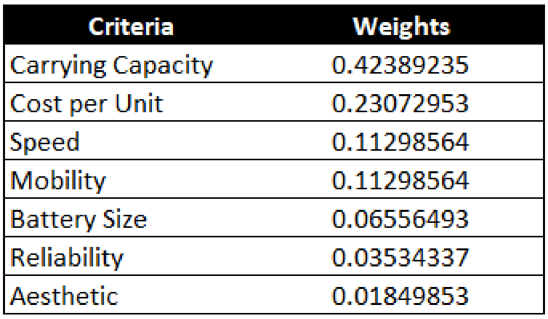


SAATY SCALE

| VALUE | DEFINITION |
|---|---|
| 1 | Equal Importance |
| 3 | Moderate Importance |
| 5 | Strong Importance |
| 7 | Very Strong Importance |
| 9 | Extreme Importance |
| 2,4,6,8 | Intermediate Values |



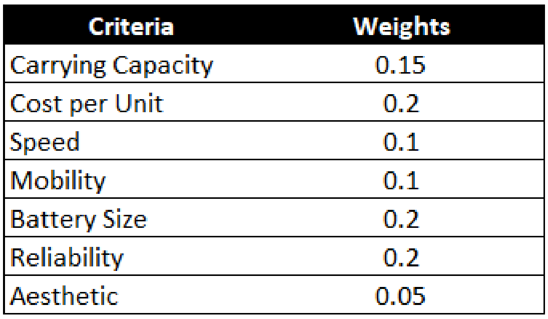
| Weights | Carrying Capacity | Cost Per Unit | Speed | Mobility | Battery Size | Reliability | Aesthetic |
|---|---|---|---|---|---|---|---|
| Carrying Capacity | 1 | 5 | 7 | 7 | 7 | 9 | 9 |
| Cost Per Unit | 1/5 | 1 | 5 | 5 | 7 | 7 | 9 |
| Speed | 1/7 | 1/5 | 1 | 1 | 5 | 7 | 7 |
| Mobility | 1/7 | 1/5 | 1 | 1 | 5 | 7 | 7 |
| Battery Size | 1/7 | 1/7 | 1/5 | 1/5 | 1 | 5 | 7 |
| Reliability | 1/9 | 1/7 | 1/7 | 1/7 | 1/5 | 1 | 5 |
| Aesthetic | 1/9 | 1/9 | 1/7 | 1/7 | 1/7 | 1/5 | 1 |



| Criteria | Weights |
|---|---|
| Carrying Capacity | 0.42389235 |
| Cost per Unit | 0.23072953 |
| Speed | 0.11298564 |
| Mobility | 0.11298564 |
| Battery Size | 0.06556493 |
| Reliability | 0.03534337 |
| Aesthetic | 0.01849853 |

**For Business Plan 2:**

- Battery capacity, cost, and reliability are considered to be the most important criteria and assigned equal weightage.
- The next important criterion is assumed to be carrying capacity.
- Speed and mobility are given equal weightage.
- Aesthetics is assumed to be the least important criterion.



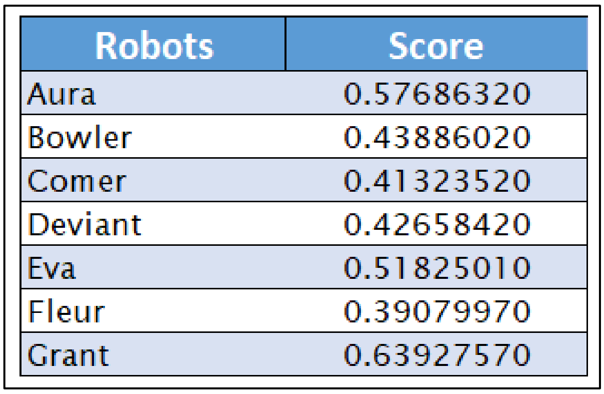
| Criteria | Weights |
|---|---|
| Carrying Capacity | 0.15 |
| Cost per Unit | 0.2 |
| Speed | 0.1 |
| Mobility | 0.1 |
| Battery Size | 0.2 |
| Reliability | 0.2 |
| Aesthetic | 0.05 |

## 3. Criteria

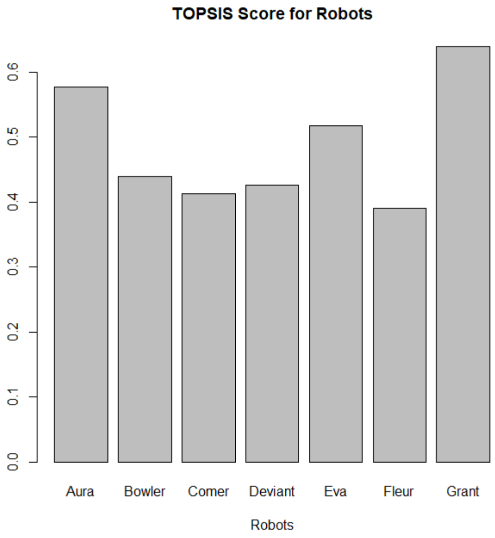
- Cost per unit needs to be minimised.
- The rest of the factors need to be maximised.

## 4. Result

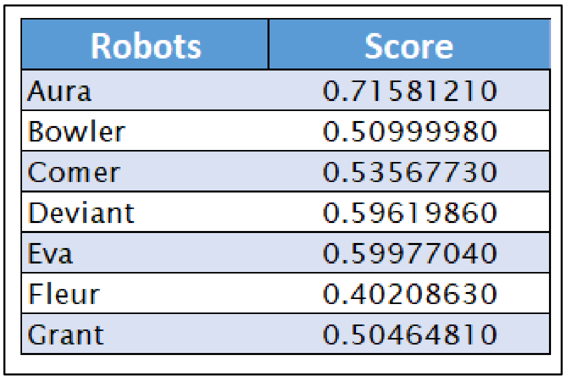
**For Business Plan 1:**



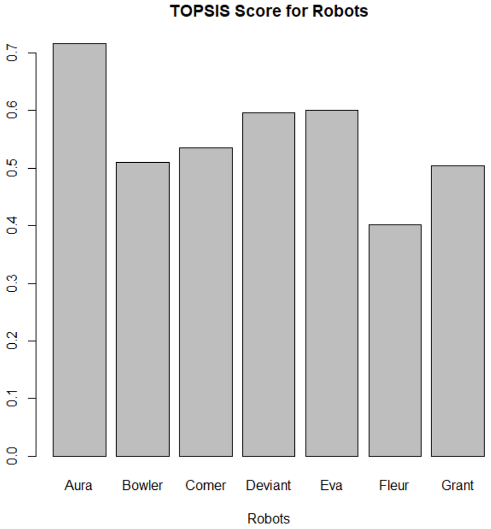
| Robots | Score |
|---|---|
| Aura | 0.57686320 |
| Bowler | 0.43886020 |
| Comer | 0.41323520 |
| Deviant | 0.42658420 |
| Eva | 0.51825010 |
| Fleur | 0.39079970 |
| Grant | 0.63927570 |



**For Business Plan 2:**



| Robots | Score |
|---|---|
| Aura | 0.71581210 |
| Bowler | 0.50999980 |
| Comer | 0.53567730 |
| Deviant | 0.59619860 |
| Eva | 0.59977040 |
| Fleur | 0.40208630 |
| Grant | 0.50464810 |



## 5. Conclusion:

**For Business Plan 1:**

Grant should be chosen. Grant has the highest score followed by Aura and Eva.

Grant is mostly favourable due to its high carrying capacity and reasonable cost option. Though it has low mobility and battery size, it is still acceptable as capacity is the most important criterion.

**For Business Plan 2:**

Aura should be chosen. Aura has the highest score, followed by Eva and Deviant.

Aura has 10 hours of battery life and is balanced with all the three critical factors, making it suitable for selling the robot technology.

## REFERENCES

Dean Abbott 2014. *Applied Predictive Analytics : Principles and Techniques for the Professional Data Analyst.* Wiley.

Hodgett, R.E., Siraj, S. and Hogg, E.L. 2024. *Smart decisions : a structured approach to decision analysis using MCDA*. Hoboken: John Wiley & Sons, Inc.