# PART – 1

# Introduction to the Business Problem :

A new start-up based in the city of Leeds, “AutonomousShipment”, have developed 4 Autonomous robot drones. The company aims to conduct a one month testing phase in order to check the usefulness of these drones in the application of doorstep delivery. As their Data Analyst, I have been tasked to –

* Determine which robot of the four (Archer, Bowler, Corner and Deviant) robots would be the best to use in the tests based on a number of factors (Carrying Capacity, Battery Size, Average Speed, Cost per unit and Reliability).

The data provided has information about every factor of each robot and also the preferences of which factors have higher priority over others as per the management teams’ perspective.

* Establish the number of robots (of the same type) to assign to stores of various types (Grocery, Clothing and Sport). This has to be done in accordance with the given constraints and goal in mind. The constraints in this case are that –

a) The budget stays within 250,000 GBP

b) The number of technician staff hours does not exceed a 1000 for the entire month, and

c) Each store should be assigned atleast 5 robots.

The Goal is to have the robots complete as many orders as possible in a day.

# TASK 1 :

The data provided by the management team consists of values corresponding to each factor of each type of robot in the file named “Robot\_Info”. There are no missing values in the dataset. The values are all of integer datatype. Out of the five factors, all are maximizing in nature except for the factor, “Cost per unit” which is minimizing in nature.

The file “Management\_Priority” consists of the management teams’ consideration of importance of each factor over another. We shall make use of this file to assign weights to each factor. The management feels like the weights of each factor are of the below mentioned order.

*Carrying capacity < Average Speed < Battery Size < Cost per Unit < Reliability*

Out of these factors, it is mentioned that Battery Size is considered 3 out of 5 stars important. Considering that Reliability is the most important factor out of the five, we shall give it 5 out of 5 stars, that is, a weight of 1. Consequently, Battery Size will have a weight of 0.6. Assuming that all of the factors’ weights’ are equidistant, we can assign a weight of 0.2 to Carrying capacity, 0.4 to Average Speed and 0.8 to Cost per Unit. Upon normalizing these weights, we end up with the table below :

|  |  |  |
| --- | --- | --- |
| **S.No.** | **Factor** | **Normalized Weight** |
| 1 | Carrying capacity | 0.066666667 |
| 2 | Average Speed | 0.2 |
| 3 | Battery Size | 0.133333333 |
| 4 | Cost per Unit | 0.266666667 |
| 5 | Reliability | 0.333333333 |

Coincidentally, the weight of Cost per Unit ended up to comprise more than 25 per cent of the total weight, which was the request of one of the members of the management team.

**Selection of MCDA method :**

Since we do not have data in the form of pairwise comparisons, we can reject AHP and make use of WSM, TOPSIS or VIKOR to solve this business problem. While WSM is a viable method, an ideal point method such as TOPSIS or VIKOR offers more clarity as it’s outputs are distances which are relative to the Positive Ideal Solution (PIS) and Negative Ideal Solution (NIS).

However, out of the two, we shall make use of **VIKOR** as it takes into consideration conflicting criteria which may impact our decision. It also offers us more flexibility to apply Sensitivity analysis on the parameter “v” as well, through which we can gain more insights.

**Implementation of the MCDA method :**

With a value of v = 0.5, the VIKOR method determined the **Robot DSXX – Deviant** to be the most ideal to be deployed as per the given criteria. With a large difference between the Q value of the Deviant and other robots, we can say that it is most likely the best option.



However, in order to be absolutely certain, we shall employ Sensitivity anlysis onto two variables, namely, the weights and the value v.

**Sensitivity Analysis on weights :**

The weights assigned to get the above mentioned values, are equidistant in nature, we shall look at another 4 sets of weights which are not equidistant between factors. 2 of the weight sets are skewed such that the higher weights gain even more magnitude at the cost of the lower weights going lower (nos.1 and 2 in the table below) and the other 2 sets consist of weights in which the lower weights increase at the cost of the higher weights decreasing (nos. 4 and 5 in the table below) (3 is the original set of weights) . The other weight sets are still in accordance with the criteria set by the management team.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Weight sets** | | | | |
|  | **1** | **2** | **3 (Original set)** | **4** | **5** |
| Carrying Capacity | 0.014881 | 0.035714 | 0.066667 | 0.130719 | 0.12381 |
| Average Speed | 0.059524 | 0.071429 | 0.133333 | 0.163399 | 0.152381 |
| Battery Size | 0.241071 | 0.214286 | 0.2 | 0.166667 | 0.171429 |
| Cost Per Unit | 0.282738 | 0.321429 | 0.266667 | 0.261438 | 0.266667 |
| Reliability | 0.401786 | 0.357143 | 0.333333 | 0.277778 | 0.285714 |

Now, when these sets of weights are fed into VIKOR, the below attached graph is obtained.

As we can see from the graph above, the Deviant is generally the best choice. However, in the cases where the weights are very close to one another, the Corner is the better option. In truth, one can make an argument for both cases, where the weights are distributed evenly to ensure that words such as “less” and “more” carry equal weight on both sides of the spectrum and also for close knit weights because why would a start-up design a robot in which its’ carrying capacity matters for only 1 to 6 percent? (Weight sets 1, 2 and 3). The best thing to do in this scenario would be to reconsult the management team.

However, since I cannot do so, I will go ahead with the **Deviant** for further analysis as it seems improbable that the difference in importance between Cost per unit and Reliability, and Average Speed and Battery Size, are only around 1 to 2 percent (Weight sets 4 and 5).

**Sensitivity Analysis on v :**

Even with changes in v, the Deviant is the most preferred in every case. This means that the Deviant is not only closest to the most ideal solution, but also the safest solution.

# TASK 2 :

Assuming the number of Deviant robots dispatched to grocery stores is G, clothing stores is C and sports equipment stores is S, we have the following set of equations from the given constraints –



Therfore, in the course of a month, the constraint equations are -

* 8700G + 8100C + 7800S <= 250000 (Budget) (throughout the month)
* 40G + 28C + 20S <=1000 (Technician working hours) (throughout the month)
* G >= 5
* C >= 5
* S >= 5

We need to maximize the number of orders done by the robots which is given by the equation :

*270G + 180C + 120S*

**Selection of MCDA method :**

In order to solve this task, we shall make use of an Optimization method, in this case **Linear Optimization.** Goal programming methods would prove useless in this current scenario as we have a goal which needs to be maximized rather than it be another target to be met. Moreover, there is no information provided as to which constraints take priority over another nor the amounts by which one constraint takes priority over another (weights).

**Implementation of the MCDA method :**

We shall conduct Linear Optimization in order of ranking of constraints, so that incase AutonomousShipment decide to remove any of the constraints in the future, they can do so. I have ranked the constraints based on my personal preference and recorded them accordingly in the table below -



Now, we find the values of G, C and S which give the highest orders per day considering only the constraints of equal or lower rank.



Hence, we see that by assigning **19 Deviants to the grocery stores, 5 to the clothing stores, and 5 to the sports equipment stores**, the robots complete **221** orders a day.

# Response to the Management of AutonomousShipment :

In conclusion, it is recommended that the management deploy the **Robot DSXX – Deviant** for the one month trial. The greatest number of orders can be achieved by assigning **19 Deviants to the grocery stores, 5 to the clothing stores, and 5 to the sports equipment stores.** The robots are expected to complete **221** orders a day. The above table can be referred to incase the management decides to removes a certain constraint in order to achieve even more orders per day.

# PART - 2

# Business Understanding :

‘Drinks@home.uk’ is an ecommerce website which sells both alcoholic and non-alcoholic beverages from all over the world to the people of Great Britain. My manager has provided me with data of 400 customers on the website over the course of the past week and expects me to derive key insights in order to help him answer two business questions :

* Determine which factors out of the many factors (Estimated\_Age, Time\_On\_Site, Seen\_Voucher, Estimated\_Income and Advertisement\_Channel) monitored by the website impacts the Revenue obtained from the average customer in a positive or negative way.
* Given the insights obtained from the first question, establish which of the three marketing strategies will be the most profitable for ‘Drinks@home.uk’. The three strategies are -

a) Running advertisements targeting customers older than 45 years old.

b) Provide a voucher for 20 GBP off the customers’ next orders, and

c) Spending money on advertising with an influencer.

# Data Understanding :

The data provided by the manager consists of values corresponding to each factor of 400 unique customers in the file named “Transactions\_Customer”. There are no missing values in the dataset. The values are all numerical in value. We hence do not need to impute any missing values or clean up the dataset and can directly move on to analyzing the data.

The summary of the data reveals to us that –

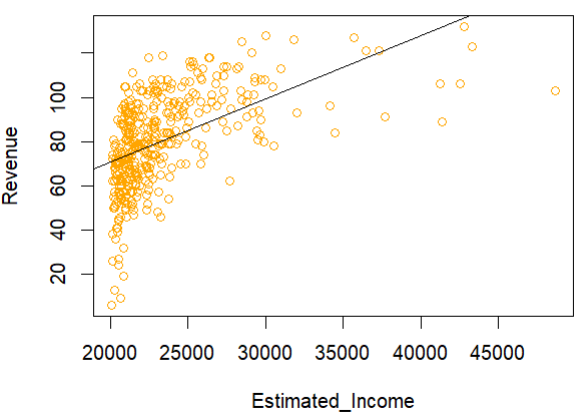
* “Estimated\_Age”, “Time\_On\_Site”, “Estimated\_Income” and “Revenue” are numerical in nature.
* “Seen\_Voucher” and “Advertisement\_Channel” are categorical in nature.
* Most of the data is evenly spread out, with the exception of a few outliers. This is good as we have equal representation of all types of customers which will give us reliable results in the modelling stage.
* There are exactly 100 people who have seen each advertisement channel.
* There are 193 people who have not seen a voucher and 207 people who have.
* The mean “Estimated\_Age” is 32.09.
* The mean “Time\_On\_Site” is 164.
* The mean “Revenue” is 80.09.

The correlation matrix of the dataset provides us with more insights -

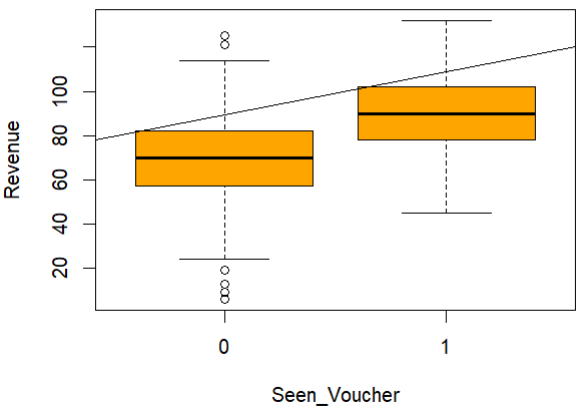


We represent the correlations of Revenue in graphical format as depicted below -

**From the above graph, we observe that Seen\_Voucher, Estimated\_Income and Advertisement\_Channel have significant positive correlation with Revenue.** We can gain a better insight by plotting each of the variables against Revenue separately. Regression lines are also plotted on the graph to better understand the general trend of the data.

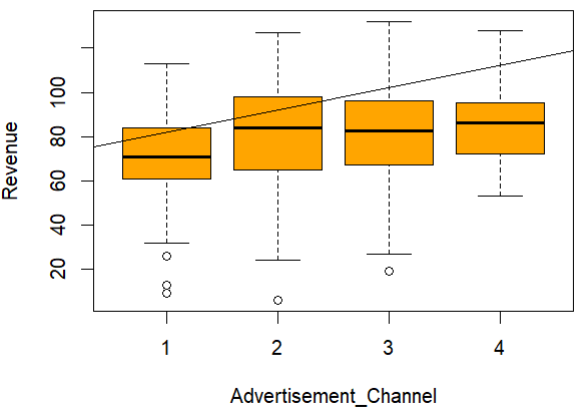


We can observe that there is a steep increase in revenue when the Estimated\_Income increases from 20000 to 25000 after which although the slope reduces it still maintains a trend upward.



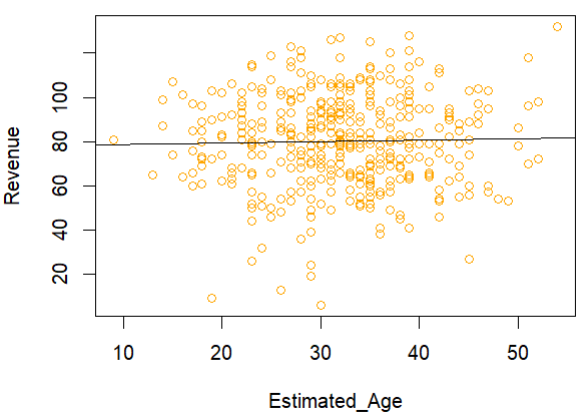
Customers who have seen a voucher are likely to spend about 90 GBP on the website while Customers who haven’t seen a voucher generally spend about 70 GBP.

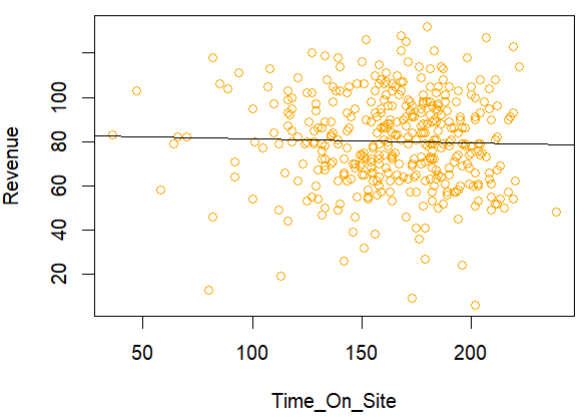
.



Influencer based ads (number 4) are significantly better than other types of ads in bringing revenue as it has the highest median out of the 4, but it also has the smallest average inter-quartile range which makes it reliable.

**Estimated\_Age and Time\_on\_Site have almost zero correlation with Revenue.** This can be visually verified in graphical form as shown below -





# Data Preparation :

In order to solve this business problem, we shall make use of **Linear Regression modelling**.

Linear Regression is a supervised model which learns the characteristics of a set of training data, using which it predicts the outputs of a different set of test data.

The steps we take in order to get the data ready for modelling are -

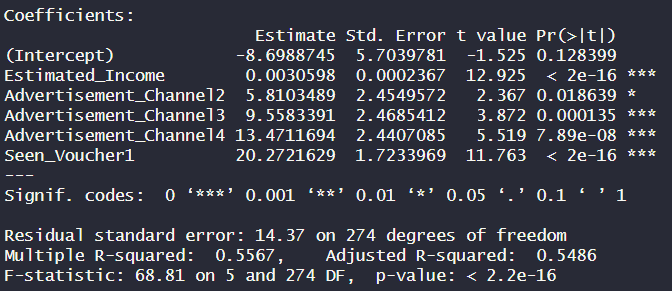
* We firstly make a new column, “Greaterthan45” which is derived from whether the “Estimated\_Age” is greater than 45. 1 if yes and 0 if no.
* Now we have 3 categorical variables, namely “Seen\_Voucher”, “Advertisement\_Channel” and “Greaterthan45”. Using the as.factor() command, we make these columns into multiple columns of just zeroes and ones.
* The last and final step is to split the dataset into training and test data. This is done in order to check the efficiency of our model.

We have made use of a 70:30 split between training and test data. (that is, 280 values in training data and 120 values in test data)

* Our data is now ready for the modelling phase.

# Modelling :

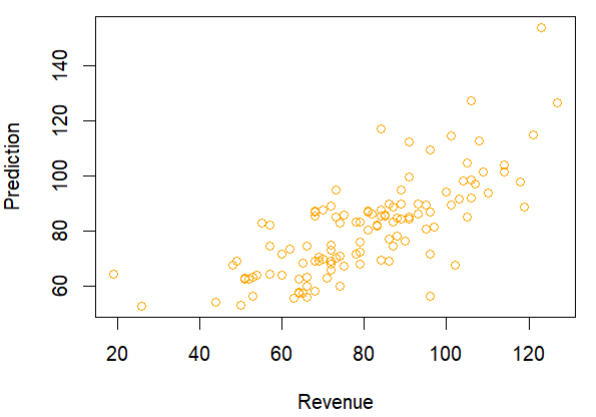
Upon allowing the Linear Regression Model to train on our training set, the model now showcases the following features -



With these values of t > 1.96, p < 0.05 and an Adjusted R-squared value of 0.5486, we can say that the **model is well-trained** and **valid**.

The model attempts to predict the test data and does so with an RMSE value of **13.53242**.

The graph below showcases the trend of prediction done by the model –

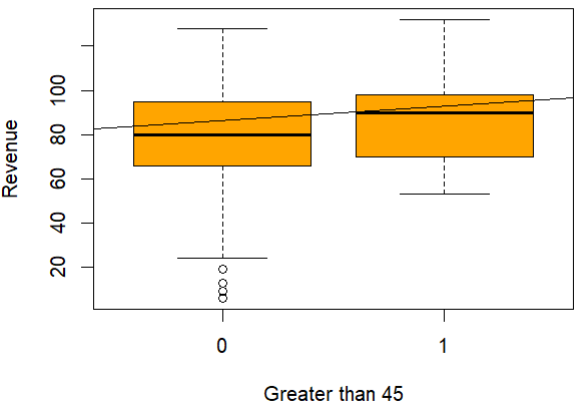


# Evaluation :

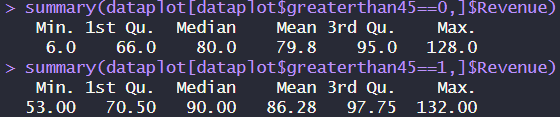
We shall now use the previously generated models of data which satisfies the conditions of the three marketing strategies in order to decide which marketing strategy will be the most profitable.

1. **Running advertisements targeting customers older than 45 years old –**

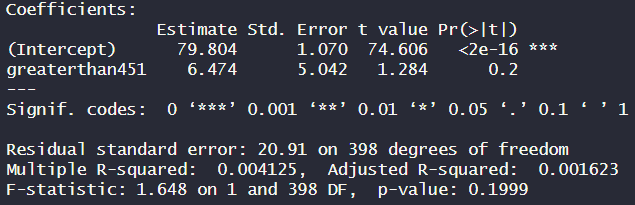
At first glance this marketing strategy seems viable as customers older than 45 old do indeed provide more revenue than customers who are younger than 45. This can be verified in the graph showcased below along with the regression line showcasing the trend of the boxplots -



On average, customers aged greater than 45 generate 6.48 GBP more revenue per purchase than a customer aged less than or equal to 45.



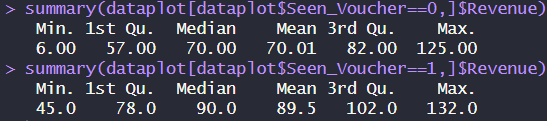
However, it should be noted that only 18 out of the 400 customers in the dataset are aged greater than 45. This makes the result acquired highly unreliable. This unreliability is also showcased in the model summary in which p > 0.05, t <1.96 and a low adjusted R squared.

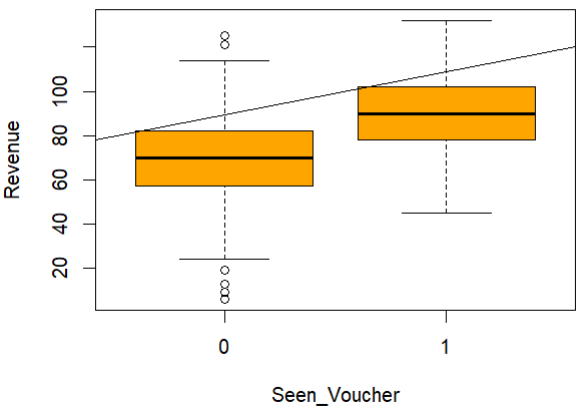


Hence, this strategy is considered unacceptable for deployment.

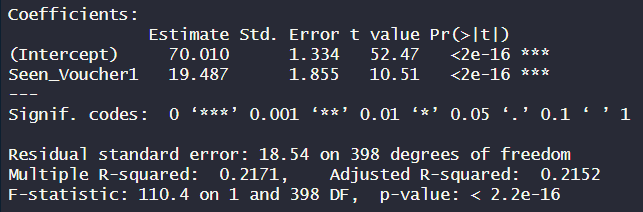
1. **Provide a voucher for 20 GBP off the customers’ next orders –**

We are already aware that customers who have seen a voucher generate more revenue than customers who haven’t. The difference in this case is found to be 19.49 GBP on average.





The model summary showcases promising results with t > 1.96, p < 0.05 and an adjusted R squared value of 0.2152.



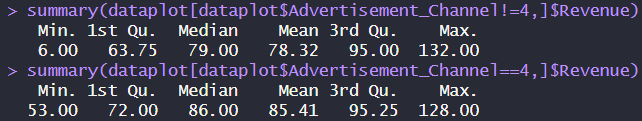
However, we need to consider that although the customers will purchase for an extra of 19.49 GBP, they will make use of the voucher provided which will ultimately result in a loss of

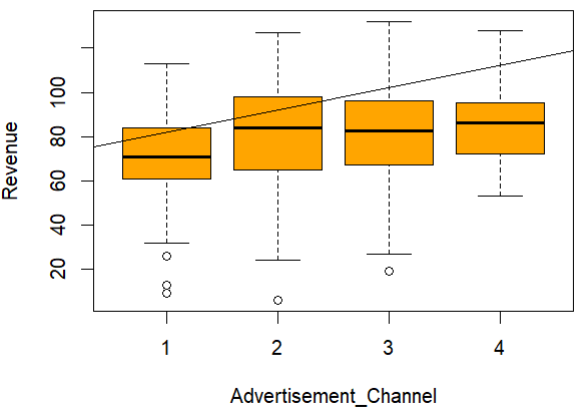
0.51 GBP per customer purchase for ‘drinks@home.uk’.

Hence, this strategy is considered unacceptable for deployment.

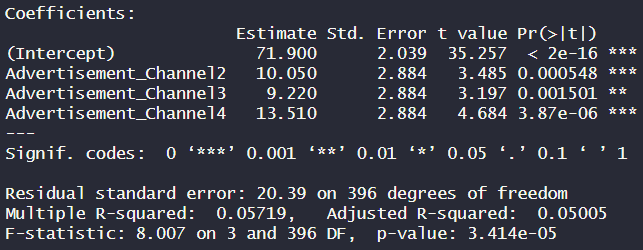
1. **Spending money on advertising with an influencer –**

This marketing strategy, at face value, also shows good promise as this mode of advertisement generates the most revenue; 7.09 GBP more than other types of advertisements.





The model summary also shows acceptable values of t, p and R square.



Hence, we have determined that ‘drinks@home.uk’ will be profitable if they **spend money on advertising with an influencer.**

# Conclusion :

In conclusion, the impact of each factor monitored by ‘drinks@home.uk’ on the per customer Revenue is as follows –

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **Factor** | **Impact on Revenue** | **Correlation** |
| 1 | Estimated\_Age | **No impact** | 0.02628325 |
| 2 | Time\_On\_Site | **No impact** | -0.0282623 |
| 3 | Seen\_Voucher | **Positive** | 0.46597667 |
| 4 | Estimated\_Income | **Positive** | 0.53165184 |
| 5 | Advertisement\_Channel | **Advertisement 1 - 4th Rank (Least Revenue), Advertisement 2 - 2nd Rank,**  **Advertisement 3 - 3rd Rank,**  **Advertisement 4 - 1st Rank (Most Revenue)** | 0.21240069 |

In order for the website to be most profitable, it is recommended that ‘drinks@home.uk’ **spend more money on advertising with an influencer.**