BUSINESS ANALYTICS IN PRACTICE

PORTFOLIO TASKS

Bhavik Sachin Raut

Table of Contents

PORTFOLIO TASK 1

Fresco's marketing management team

Aim of the project

Our main Aim in this report is to examine data from Fresco Supermarket, in order to identify trends and patterns in a sample of weekly data collected for a number of their loyalty cardholders over a 26-week period, and to make any necessary adjustments in order to predict the value of a single customer's shopping basket.

Summary

The total sample size is 75.

Initially, the dependent and independent variables were identified, and all independent factors that have an influence on our dependent variable were examined.

VARIABLES	VARIABLE DESCRIPTION
Spending's (Dependent Variable)	Low spender=0, Medium Spender=1, High Spender= 2 (Categorical Variable)
Store Type (Independent Variable)	0=Convenient Stores, 1=Superstore, 2= Online (Categorical)
Gender (Independent Variable)	0=Male, 1= Female (Categorical Variable)
Age (Independent Variable)	Continuous variable representing the age of customers
Value Products (Independent Variable)	Continuous variable
Brand Products (Independent Variable)	Continuous variable
Top Fresco Products (Independent Variable)	Continuous variable

Three groups were assigned to the shopping basket: Low, Medium, and High spenders. By using this as our dependent variable we Identified the spending's of the potential customer.

Following that, we estimated our model, and then we used the parsimony model for our further analysis.

To test the adequacy of our final mode, we ran a multicollinearity test, did residual analysis, observed standardised residuals, and computed the cook distance and DFBeta. Finally, to establish the goodness of fit, we looked at the pseudo r square, Hosmer and Lemeshow's tests, and classification accuracy, finding that each category's accuracy was similar.

We identified a parsimonious model that can be used for prediction and classification after passing the adequacy tests.

Justification of the methods considered

We have used Multinomial logistic regression in this analysis as we have a categorical dependent variable with three levels in this task. This method helped us in our analysis to understand whether factors such as customers gender, age, shopping frequency per week and shopping basket price affect our spending of the potential customer.

overview of the results.

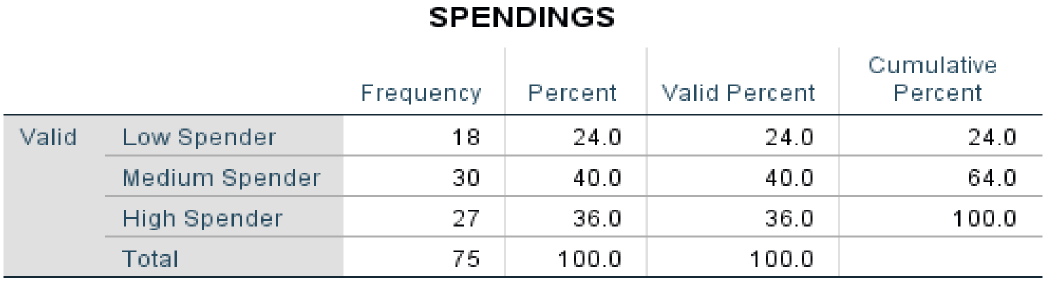
After all the analysis we found accurate model which correctly identified the spendings of the potential customer according to the following three groups:

Low Spender	shopping basket value of £25 or less	88.9% correctly predicted
Medium Spender	shopping basket value between £25.01 and £70	76.7% correctly predicted
High Spender	shopping basket greater than £70	85.2% correctly predicted

Part B:

In this report we will be using multinomial logistic regression as our dependent variable has more than two outcomes such as Low spender, Medium Spender, High Spender. And our independent variables are customers gender, age , shopping frequency per week and shopping basket price

We look at the highest frequency to determine our reference category, from the table below we discovered that the medium spender had the highest frequency, therefore we have used that as our reference category in our analysis. Using a combination of binary logistics models such as Low vs Medium and High vs Medium, it compares different groups.



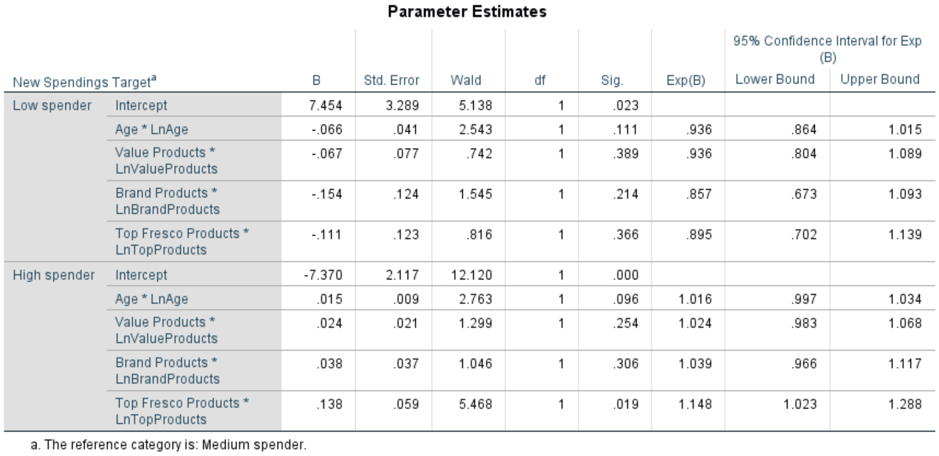
SPENDINGS

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Low Spender	18	24.0	24.0	24.0
	Medium Spender	30	40.0	40.0	64.0
	High Spender	27	36.0	36.0	100.0
	Total	75	100.0	100.0	

Assumption- Linearity

We created new variable that is the natural logs for each continuous independent variable to ensure the linear relationship between continuous variables and the logit of the dependent variable. After that, we performed a Logistic regression where we included all the independent variable as well as interactions between each of them and their logarithmic variables. The table below shows that most of the values are more than 0.05, indicating that linearity is not an issue.

Now that the 1st Assumption has been met, we may go on to logistic regression.

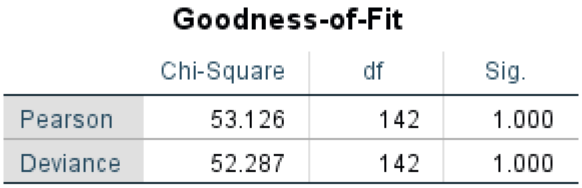


Parameter Estimates

							95% Confidence Interval for Exp (B)	
New Spendings Target[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	Lower Bound	Upper Bound
Low spender	Intercept	7.454	3.289	5.138	1	.023			
	Age * LnAge	-.066	.041	2.543	1	.111	.936	.864	1.015
	Value Products * LnValueProducts	-.067	.077	.742	1	.389	.936	.804	1.089
	Brand Products * LnBrandProducts	-.154	.124	1.545	1	.214	.857	.673	1.093
	Top Fresco Products * LnTopProducts	-.111	.123	.816	1	.366	.895	.702	1.139
High spender	Intercept	-7.370	2.117	12.120	1	.000			
	Age * LnAge	.015	.009	2.763	1	.096	1.016	.997	1.034
	Value Products * LnValueProducts	.024	.021	1.299	1	.254	1.024	.983	1.068
	Brand Products * LnBrandProducts	.038	.037	1.046	1	.306	1.039	.966	1.117
	Top Fresco Products * LnTopProducts	.138	.059	5.468	1	.019	1.148	1.023	1.288

a. The reference category is: Medium spender.

Independence of errors – Looking at the chi square goodness of fit and the DF column below we can observe that this assumption passes as the Chi square / Df = 53.126/142 **<u>= 0.374 which is less than threshold of 2.</u>**



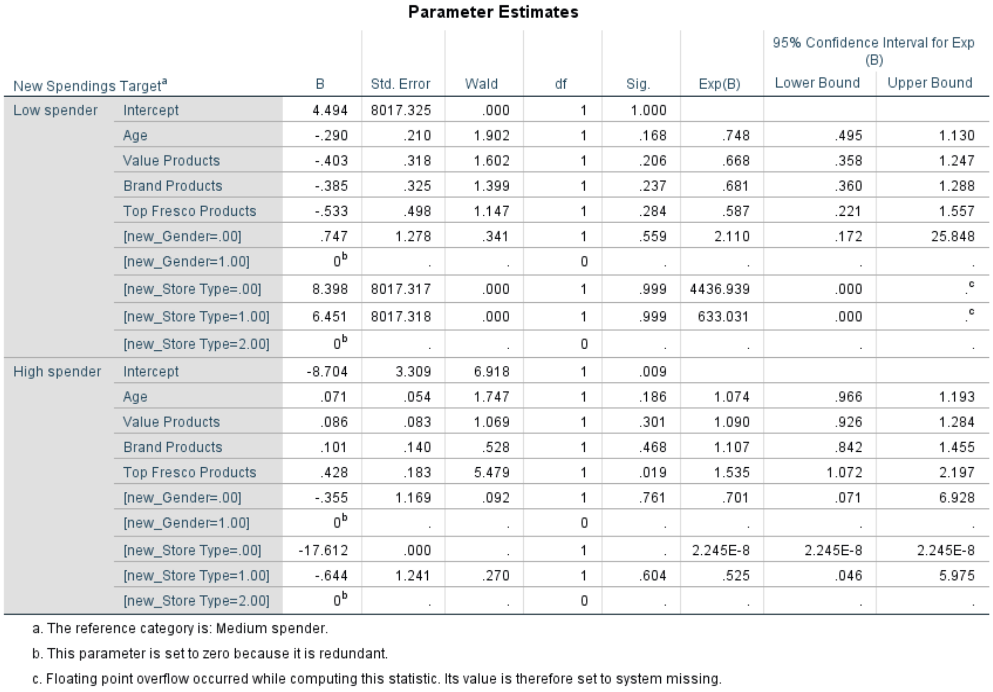
Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	53.126	142	1.000
Deviance	52.287	142	1.000

ESTIMATION OF MODEL

All the independent variables are insignificant since the significance of wald statistics is <0.05 except the interaction of Top fresco Product.

Looking at the table, we can see that store type is the least important variable, with a value of 0.999 which is >0.05, therefore we eliminated it first. To obtain a parsimonious model, all insignificant variables were gradually deleted from our model.



Parameter Estimates

								95% Confidence Interval for Exp (B)	
New Spendings Target[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	Lower Bound	Upper Bound
Low spender	Intercept	4.494	8017.325	.000	1	1.000			
	Age	-.290	.210	1.902	1	.168	.748	.495	1.130
	Value Products	-.403	.318	1.602	1	.206	.668	.358	1.247
	Brand Products	-.385	.325	1.399	1	.237	.681	.360	1.288
	Top Fresco Products	-.533	.498	1.147	1	.284	.587	.221	1.557
	[new_Gender=.00]	.747	1.278	.341	1	.559	2.110	.172	25.848
	[new_Gender=1.00]	0[b]	.	.	0	.	.	.	.
	[new_Store Type=.00]	8.398	8017.317	.000	1	.999	4436.939	.000	.[c]
	[new_Store Type=1.00]	6.451	8017.318	.000	1	.999	633.031	.000	.[c]
	[new_Store Type=2.00]	0[b]	.	.	0	.	.	.	.
High spender	Intercept	-8.704	3.309	6.918	1	.009			
	Age	.071	.054	1.747	1	.186	1.074	.966	1.193
	Value Products	.086	.083	1.069	1	.301	1.090	.926	1.284
	Brand Products	.101	.140	.528	1	.468	1.107	.842	1.455
	Top Fresco Products	.428	.183	5.479	1	.019	1.535	1.072	2.197
	[new_Gender=.00]	-.355	1.169	.092	1	.761	.701	.071	6.928
	[new_Gender=1.00]	0[b]	.	.	0	.	.	.	.
	[new_Store Type=.00]	-17.612	.000	.	1	.	2.245E-8	2.245E-8	2.245E-8
	[new_Store Type=1.00]	-.644	1.241	.270	1	.604	.525	.046	5.975
	[new_Store Type=2.00]	0[b]	.	.	0	.	.	.	.

a. The reference category is: Medium spender.

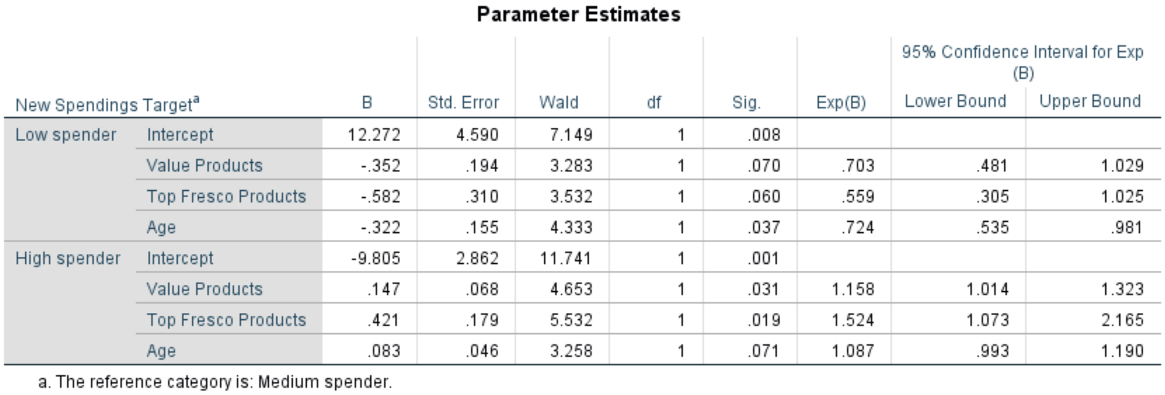
b. This parameter is set to zero because it is redundant.

c. Floating point overflow occurred while computing this statistic. Its value is therefore set to system missing.

These results provide information comparing Low Spender and High Spender against the reference category (Medium Spender). Specifically, the regression coefficients indicate which predictors significantly discriminate between people with low spending and medium spending as well as High spenders and medium spenders.

Final model:

After removing all the insignificant variables we found our best model, where we have all significant variables.



Parameter Estimates

New Spendings Target[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
Low spender	Intercept	12.272	4.590	7.149	1	.008			
	Value Products	-.352	.194	3.283	1	.070	.703	.481	1.029
	Top Fresco Products	-.582	.310	3.532	1	.060	.559	.305	1.025
	Age	-.322	.155	4.333	1	.037	.724	.535	.981
High spender	Intercept	-9.805	2.862	11.741	1	.001			
	Value Products	.147	.068	4.653	1	.031	1.158	1.014	1.323
	Top Fresco Products	.421	.179	5.532	1	.019	1.524	1.073	2.165
	Age	.083	.046	3.258	1	.071	1.087	.993	1.190

a. The reference category is: Medium spender.

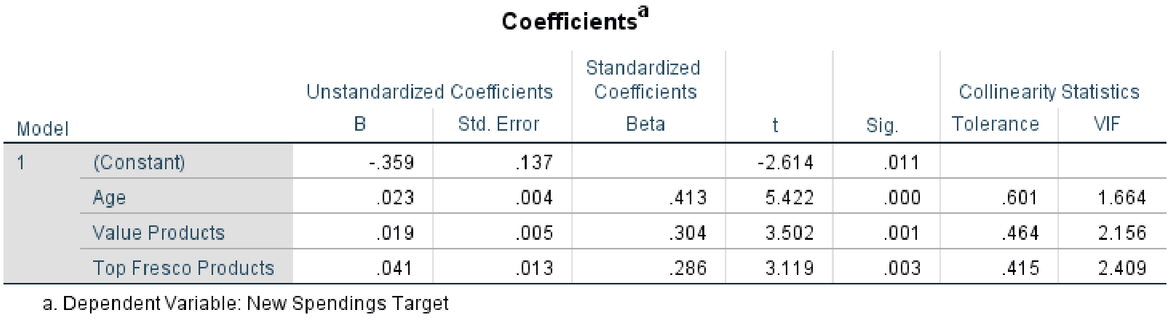
The first set of coefficients represents comparison between Low spenders and Medium spenders. Here we can see that age is the most significant predictor in the model as p value of age is 0.37 which is <0.05.

The second set of coefficients represents comparison between High Spender and Medium Spender. Here the most significant predictor is top fresco Products in the model as p value is 0.019 which is <0.5

In simple words, for a unit change in age of the customers, the odds of the customer being a low spender is decreased by 0.724 times the age of customer, similarly , the odds of the customers being a high spenders are increased by 1.524 times the Top Fresco Products purchased also the odds for being a high spenders are increased by 1.158 times the value products

MODEL’S ADEQUACY

Checking multicollinearity



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-.359	.137		-2.614	.011		
	Age	.023	.004	.413	5.422	.000	.601	1.664
	Value Products	.019	.005	.304	3.502	.001	.464	2.156
	Top Fresco Products	.041	.013	.286	3.119	.003	.415	2.409

a. Dependent Variable: New Spendings Target

Looking at the collinearity statistics we can observe that VIF is <10 and Tolerance is >0.1 which shows that there is no multicollinearity among the independent variables.

It is also crucial to examine the model's adequacy, apart from its goodness of fit. Diagnostic tests were used to do this.

Standardised residuals

We observed the cases which are above 2 (only 1 residual are above 2), and only 1% are above 2.5 (only 1 residual is around 2.5), which is important to check the model's standardised residuals

Cook's distance

Cook's distance is another diagnostic metric of model adequacy, and we observed that no residuals had a Cook's distance greater than 1, indicating that our model is adequate.

DFBetas

DFBetas for each independent variable are less than 1 which passes this test as well. (all are below 1)

Goodness of fit

Pseudo R-Square

Cox and Snell	.767
Nagelkerke	.868
McFadden	.676

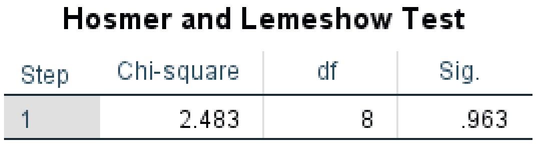
These are pseudo-R square values that are treated as similar to the R square value in OLS regression.

Here the **Cox and Snell value is 0.767** and **Nagelkerke = 0.868**, both the values are close to 1 which is a good sign for our model. Looking at these values we can say this is a good performing model.

Hosmer and Lemeshow's test.

The Hosmer and Lemeshow's test of overall fit tests whether the model's predicted output is different to the actual observations. The null hypothesis is that they are the same while the alternative is that they are different. As a result, we want the test to accept the null hypothesis, which is the case here since the reported Sig. of 0.963 is higher than 0.05 and therefore we need to accept the null hypothesis. Hence, the set of independent variables has some relevance to the dependent variable.

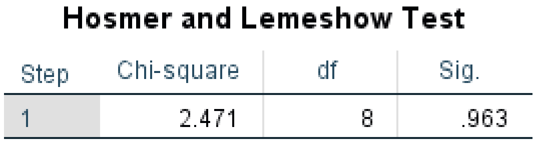
Looking at the sig value we can conclude that this is a good model as the values in both the binary models are above 0.05.



Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	2.483	8	.963

-> Medium Spender & High Spender



Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	2.471	8	.963

-> Medium Spender & Low Spender

Classification Accuracy

Observed	Predicted			
	Low spender	Medium spender	High spender	Percent Correct
Low spender	16	2	0	88.9%
Medium spender	4	23	3	76.7%
High spender	0	4	23	85.2%

Overall Percentage	26.7%	38.7%	34.7%	82.7%

These are classification statistics used to determine which group were best predicted by the model.

Low spenders were correctly predicted by the model 88.9% of the Spenders (as 16 of the 18 people who are low spenders)

Medium Spenders were correctly predicted by the model 76.7%

High spenders were correctly predicted by the model 85.2%

Looking at the overall percentage 82.7% we can conclude that this is a good model.

In terms of model accuracy, we can state that our model accurately predicted the spending of potential customers in these three categories.

PORTFOLIO TASK 2

CONJOINT ANALYSIS

Introduction

The aim of this report is to understand consumer preferences in mobile phone buying for the launch of a new mobile phone, as well as the characteristics they believe are most significant during the purchase. This analysis provides us with information into what a consumer expects and supports us with the launch of a new mobile phone.

Method

In this analysis, we used conjoint analysis, as it is a multivariate technique created to analyse preferences for any sort of product. Conjoint analysis is also frequently utilised in the design of new products or product expansions. In our case, we used this method to learn about customer preferences for the launch of a new mobile phone.

Selection of features

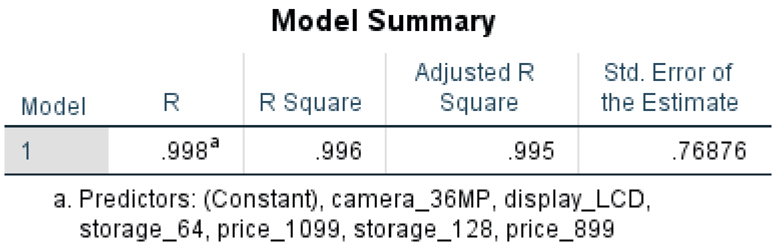
The four features listed below in table 1 have levels of 3,3,2,2 respectively. These features were deemed relevant because buyers examine them before making a purchase. Then we used to create 36 different (3x3x2x2) combinations of the product. These product combinations were ranked by ten friends based on their preferences, and the results were then aggregated. The rankings were given from 1 to 36, with 1 being the worst and 36 being the best.

PRICE (GBP)	**STORAGE (GB)**	**DISPLAY TYPE**	**CAMERA (MP)**
£799 £899 £1099	256GB 128 GB 64 GB	OLED LCD	36MP 16MP

Table 1: Features of Mobile phone

Later, in order to do regression analysis, we created Dummy variables (combinations of 0 and 1) and employed the collected information, ranking of the various product opinions which were ranked from 1 to 36. This information was subsequently used in a linear regression for conjoint analysis.

Data Analysis- Regression Output

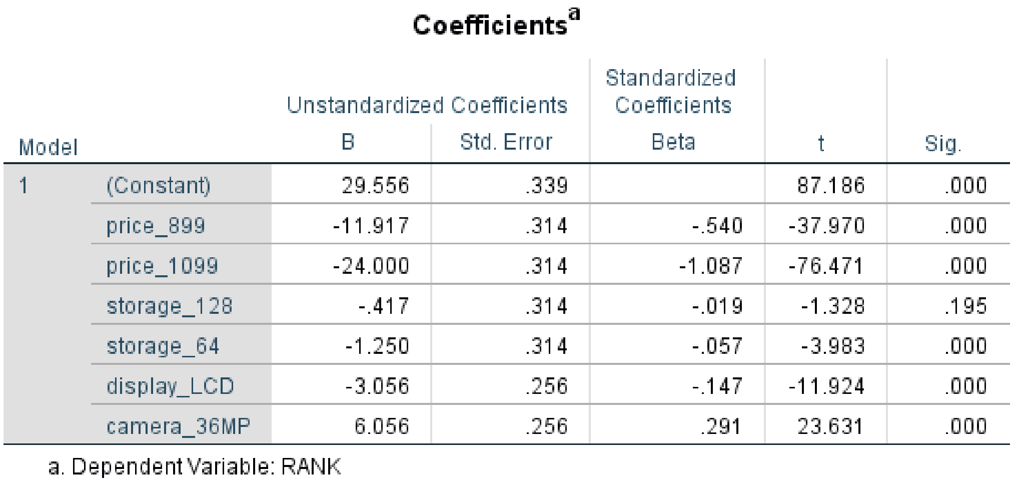


Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.998[a]	.996	.995	.76876

a. Predictors: (Constant), camera_36MP, display_LCD, storage_64, price_1099, storage_128, price_899

In the model summary we can observation that our R-Square value is .996, indicates that our dummy variables (product attributes) are capable of accurately predicting the rank order.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	29.556	.339		87.186	.000
	price_899	-11.917	.314	-.540	-37.970	.000
	price_1099	-24.000	.314	-1.087	-76.471	.000
	storage_128	-.417	.314	-.019	-1.328	.195
	storage_64	-1.250	.314	-.057	-3.983	.000
	display_LCD	-3.056	.256	-.147	-11.924	.000
	camera_36MP	6.056	.256	.291	23.631	.000

a. Dependent Variable: RANK

In the Coefficients table above, we first look at the Standardized Coefficients Beta, which is a measure of the impact of a certain attribute's level on response preference (RANK), which is the dependent variable in our linear model. The relative importance of contribution and impact of the independent variables on the Rank is identified by looking at the standardized coefficients. So, looking at the numbers we can see that there are negative values and one positive value.

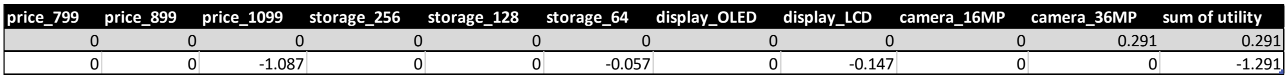
The 1st level of price (£799) is not appeared in the above table because we created dummy variables (n-1). 799 GBP is not producing dummy variable so we can see the standardized coefficient of PRICE section £899 is -0.540 also £1099 have a negative value (-1.087) because consumers won't be happy to pay more so they prefer buying mobile phone with less price rather than spending more.

As we can see that 128 GB and 64GB has negative values -0.019, -0.057 respectively which indicates that most of the consumers prefer buying high storage mobile phone with 256GB.

Consumers prefer OLED displays for their phone because, as we can see, LCD displays have a negative value (-0.147), which is not preferred by customers.

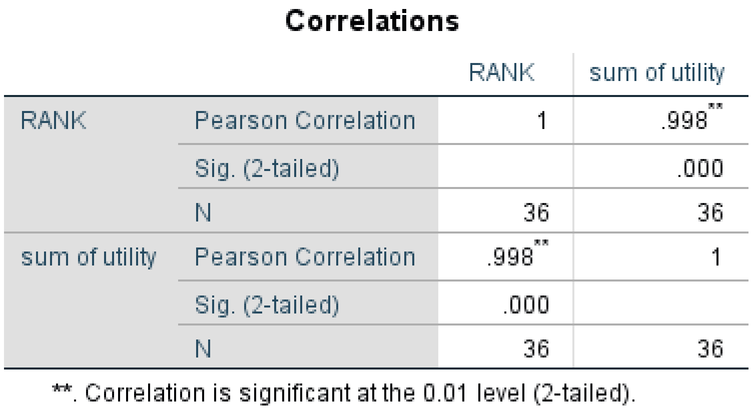
The camera with 36 Megapixel has a positive value (0.291), indicating that customers appreciate high quality cameras and are unwilling to use poor quality cameras such as 16 Megapixel having value of 0, which has been excluded from the table.

Later for each combination of product we added the utility figures that we got from the regression analysis. Some examples are shown below.



price_799	price_899	price_1099	storage_256	storage_128	storage_64	display_OLED	display_LCD	camera_16MP	camera_36MP	sum of utility
0	0	0	0	0	0	0	0	0	0.291	0.291
0	0	-1.087	0	0	-0.057	0	-0.147	0	0	-1.291

This gives us an indication of that the features of a mobile phone provides us with higher utility (e.g.0.291) or lower utility (e.g. -1.291).



Correlations

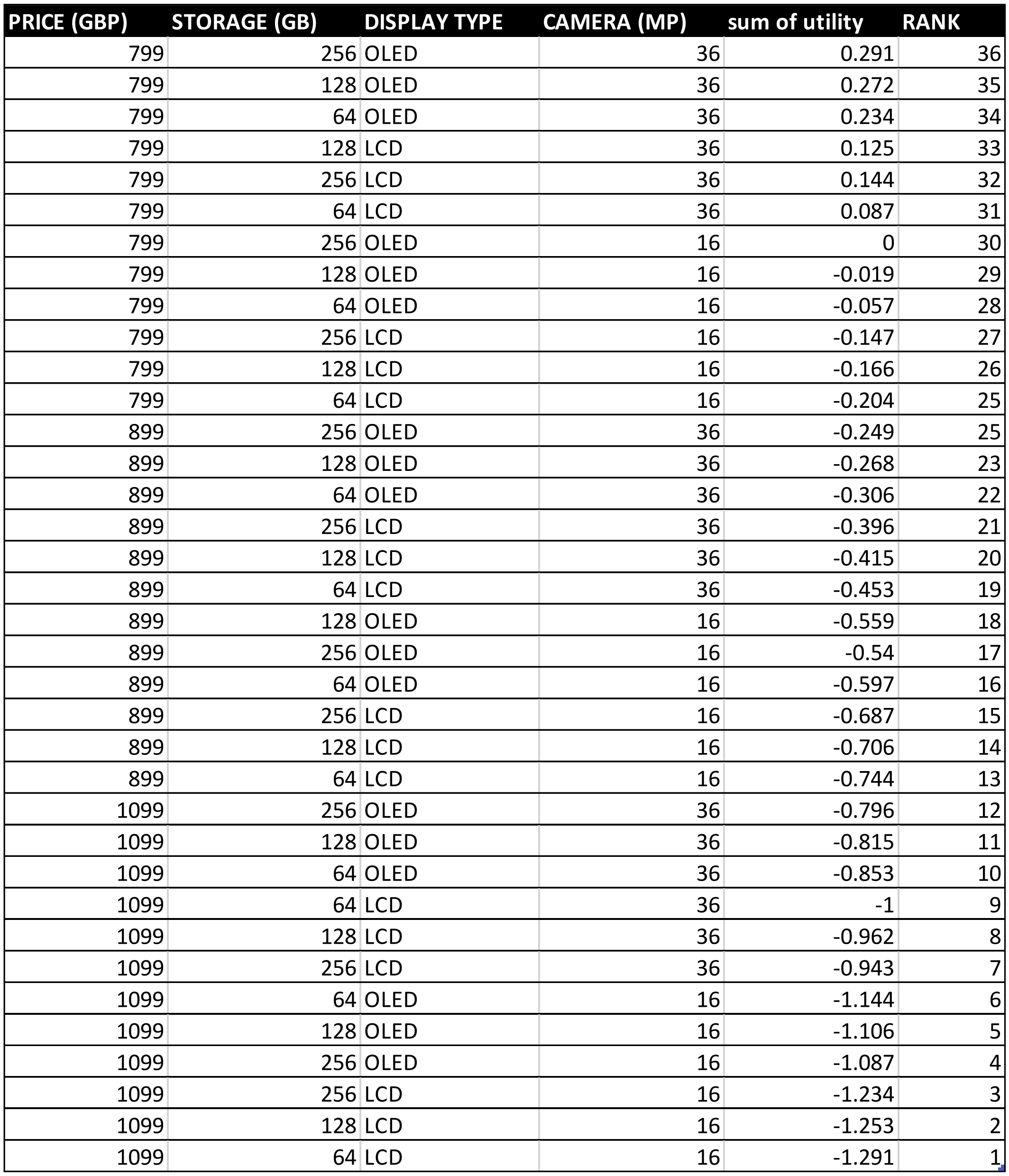
		RANK	sum of utility
RANK	Pearson Correlation	1	.998**
	Sig. (2-tailed)		.000
	N	36	36
sum of utility	Pearson Correlation	.998**	1
	Sig. (2-tailed)	.000	
	N	36	36

**. Correlation is significant at the 0.01 level (2-tailed).

We can correlate our utility estimates with our earlier rank orders, and the correlation coefficient is .998, indicating that our utility estimates are correct.

For more details we can observe the table 2 below.

We have shown the overall utilities for each product in the table below. We can also see that the majority of consumers choose to buy a mobile phone that is less expensive yet has great features. For example, we can see that the highest rank (36) is given to the most preferred product having the highest sum of utility.

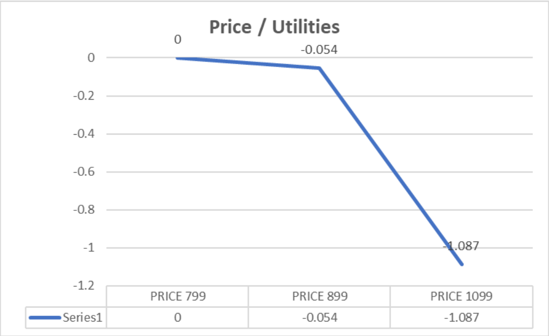


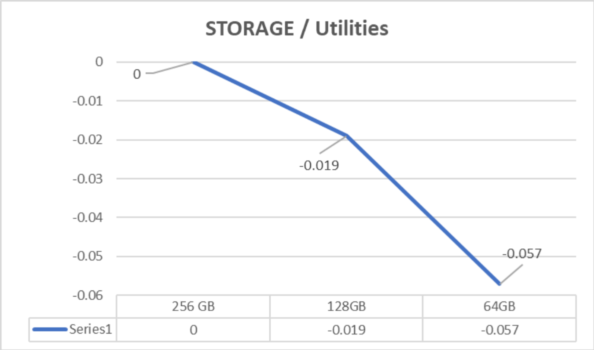
PRICE (GBP)	STORAGE (GB)	DISPLAY TYPE	CAMERA (MP)	sum of utility	RANK
799	256	OLED	36	0.291	36
799	128	OLED	36	0.272	35
799	64	OLED	36	0.234	34
799	128	LCD	36	0.125	33
799	256	LCD	36	0.144	32
799	64	LCD	36	0.087	31
799	256	OLED	16	0	30
799	128	OLED	16	-0.019	29
799	64	OLED	16	-0.057	28
799	256	LCD	16	-0.147	27
799	128	LCD	16	-0.166	26
799	64	LCD	16	-0.204	25
899	256	OLED	36	-0.249	25
899	128	OLED	36	-0.268	23
899	64	OLED	36	-0.306	22
899	256	LCD	36	-0.396	21
899	128	LCD	36	-0.415	20
899	64	LCD	36	-0.453	19
899	128	OLED	16	-0.559	18
899	256	OLED	16	-0.54	17
899	64	OLED	16	-0.597	16
899	256	LCD	16	-0.687	15
899	128	LCD	16	-0.706	14
899	64	LCD	16	-0.744	13
1099	256	OLED	36	-0.796	12
1099	128	OLED	36	-0.815	11
1099	64	OLED	36	-0.853	10
1099	64	LCD	36	-1	9
1099	128	LCD	36	-0.962	8
1099	256	LCD	36	-0.943	7
1099	64	OLED	16	-1.144	6
1099	128	OLED	16	-1.106	5
1099	256	OLED	16	-1.087	4
1099	256	LCD	16	-1.234	3
1099	128	LCD	16	-1.253	2
1099	64	LCD	16	-1.291	1

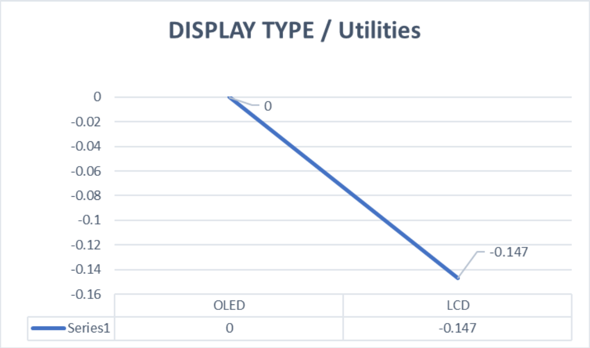
Table 2: overall utilities for each product

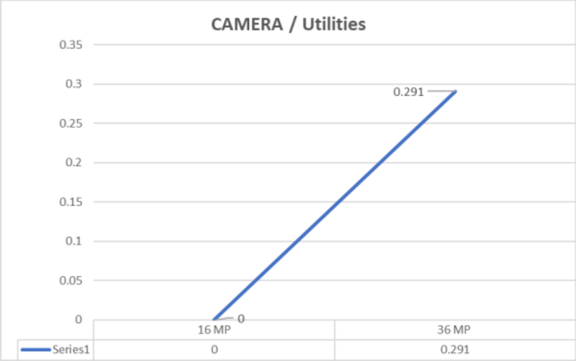
Utility Graphs

We generated Utility graphs for each aspect and discovered that there is a utility spike observed in price starting at £899, and in storage (128GB) indicating that this is where the utility drastically changes, and the rest are constant or slightly decreasing.









Conclusion

From this analysis we can conclude that conjoint analysis is a powerful tool for companies for deciding the market strategies before launching a new mobile phone also useful to decide the prices according to consumer preferences.

For example, the most desired combination of product had the maximum utility of 0.291 when the mobile cost was £799 and included 256GB of storage, an OLED display, and a camera with 36 megapixels.

The least desired combination had the lowest utility of -1.291 when the mobile price was £1099 and included 64GB of storage, an LCD display, and a 16-megapixel camera.

PORTFOLIO TASK 3

CLUSTER ANALYSIS

Introduction

Our goal in this report is to find trends and patterns in a sample of records obtained from a few of their customers by undertaking segmentation analysis for a UK Bank. The goal of Segmentation analysis is customer grouping with high similarities and also the different groups being unique and not combined.

Customers' gender, age, savings and current account balances, marital status, homeownership, and employability, as well as how long they've been customers with the bank, how long they've been in employment, and their credit risk, were all considered in this analysis. These factors are classified as Demographic variables.

Method used to apply for the market segmentation

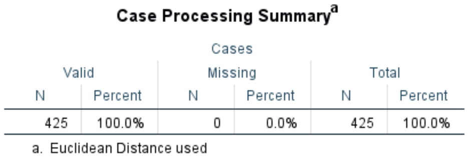
In this study, we chose Hierarchical cluster analysis to do the market segmentation because the goals of market segmentation and cluster analysis are identical. As the objective is to achieve data for group of customers in a bank where consumers belonging to that group have attributes that are similar to each other but differ from customers belonging to other groups. Based on distance and similarity between the individuals, this can be accomplished using the clustering approach. Finally, every customer of the bank will be assigned to a cluster. A dendrogram can be used to represent this.

Characteristics used for cluster analysis

To begin with the analysis, we took into account all of the continuous and categorical variables, including current account, savings account, months customer, months employed, age, gender, marital status, housing, job, and credit risk. Among these variables we converted all categorical variables into numeric so that we could utilise it in our initial clustering.

For our better understanding, we labelled the cases with unique customer IDs, which aided us in evaluating the proximity matrix and dendrogram.

We didn't mention the number of clusters in the initial stage because we couldn't decide how many clusters we'd have in our solution. We considered the dendrogram before deciding on the number of clusters to make. In this case, we used the Within Linkage and used the Euclidean distance to measure the distance. We also normalised the data between the ranges of 0 and 1, which is always preferable because we don't want the measurement unit to affect our results. Below is the result of the initial clustering method:



Case Processing Summary[a]

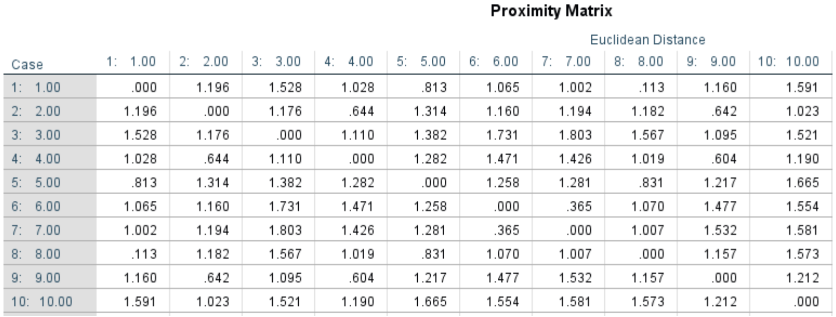
Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
425	100.0%	0	0.0%	425	100.0%

a. Euclidean Distance used

Looking at the case processing summary we can see that in total we have 425 cases with 0 missing cases.

Proximity matrix

The proximity matrix which calculates the distance between cases. We can see that the distance between case 1 and 2 is 1.196, and the distance between case 1 and 3 is 1.528. Because the distance is relatively large, we expect that they will not be in the same cluster. On the other hand, the similarity between cases 1 and 8 is 0.113, i.e., the Euclidean distance between them, therefore these cases may end up being the same cluster.



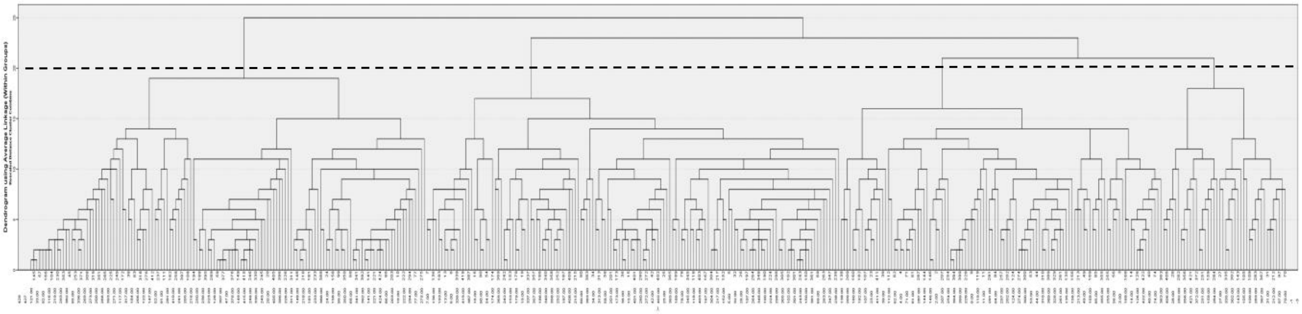
Proximity Matrix

	Euclidean Distance									
Case	1: 1.00	2: 2.00	3: 3.00	4: 4.00	5: 5.00	6: 6.00	7: 7.00	8: 8.00	9: 9.00	10: 10.00
1: 1.00	.000	1.196	1.528	1.028	.813	1.065	1.002	.113	1.160	1.591
2: 2.00	1.196	.000	1.176	.644	1.314	1.160	1.194	1.182	.642	1.023
3: 3.00	1.528	1.176	.000	1.110	1.382	1.731	1.803	1.567	1.095	1.521
4: 4.00	1.028	.644	1.110	.000	1.282	1.471	1.426	1.019	.604	1.190
5: 5.00	.813	1.314	1.382	1.282	.000	1.258	1.281	.831	1.217	1.665
6: 6.00	1.065	1.160	1.731	1.471	1.258	.000	.365	1.070	1.477	1.554
7: 7.00	1.002	1.194	1.803	1.426	1.281	.365	.000	1.007	1.532	1.581
8: 8.00	.113	1.182	1.567	1.019	.831	1.070	1.007	.000	1.157	1.573
9: 9.00	1.160	.642	1.095	.604	1.217	1.477	1.532	1.157	.000	1.212
10: 10.00	1.591	1.023	1.521	1.190	1.665	1.554	1.581	1.573	1.212	.000

Dendrogram Analysis

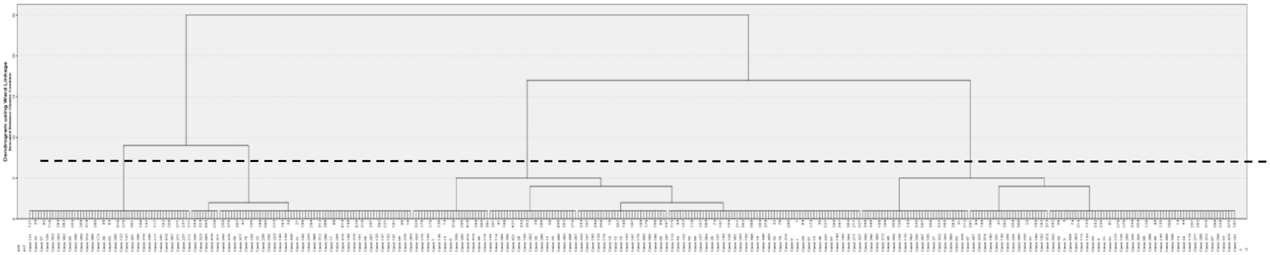
Dendrogram (Within Group Linkage) :

We explored several clusters by looking at the dendrogram below, thus based on our analysis, we considered our imaginary line and determined that 4 clusters would be an adequate solution based on the distribution.



Dendrogram (Ward Method):

We also used the Ward Method in our analysis; looking at the distribution, we don't think it's good distribution, but after considering our imaginary line it does show that 4 clusters would be the best choice for our further analysis.

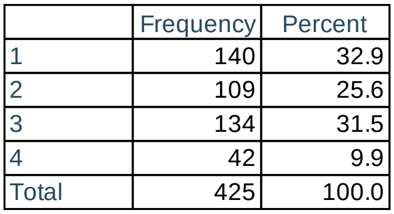


Creating Different Clusters

We iteratively built clusters from the above output (dendrogram) to discover the suitable solution by comparing both the methods using their frequencies.

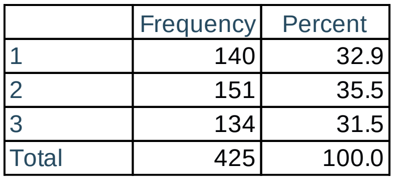
Frequency tables for Within Linkage method

Table 1



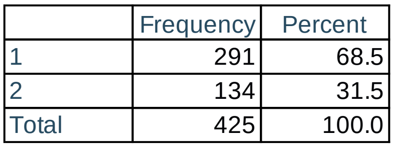
	Frequency	Percent
1	140	32.9
2	109	25.6
3	134	31.5
4	42	9.9
Total	425	100.0

Table 2



	Frequency	Percent
1	140	32.9
2	151	35.5
3	134	31.5
Total	425	100.0

Table 3

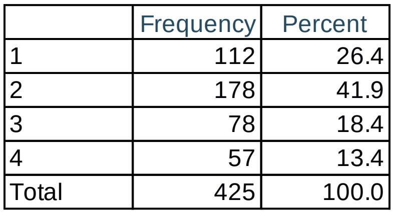


	Frequency	Percent
1	291	68.5
2	134	31.5
Total	425	100.0

We Based on the frequency table above for Within Linkage Method. We can observe that the frequency of table 1 and table 2 are not evenly distributed, as the 4th cluster in table 1 has only 42 customers compared to the other groups, which is less. And the values in table 3 are approximately 50% less than cluster 1 in that table. In comparison, an equal distribution can be seen in table 2

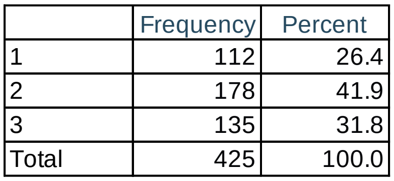
Frequency tables for Ward Method

Table 1



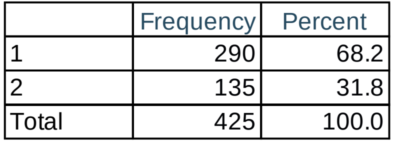
	Frequency	Percent
1	112	26.4
2	178	41.9
3	78	18.4
4	57	13.4
Total	425	100.0

Table 2



	Frequency	Percent
1	112	26.4
2	178	41.9
3	135	31.8
Total	425	100.0

Table 3



	Frequency	Percent
1	290	68.2
2	135	31.8
Total	425	100.0

According to the Ward Method frequency tables above, starting with 4 clusters and ending with 2 clusters, the groups are distributed equally in table 2 compared to table 1 and table 3.

CONCLUSION

After examining the frequency tables for both methods, we decided that the total clusters should be 3 (table 2) of the Within Linkage Method because the percentage of values in all three clusters are evenly distributed, which will support the bank's product development team in developing financial products and promotions for their customers with respect to their segments.

PORTFOLIO TASK 4

Time series forecasting with the decomposition technique using Excel

Introduction

This report summarises the results of a time series forecasting performed on a sample of monthly data on behalf of a business analytics consulting firm using the decomposition technique. The aim of this report is to compute future values for each individual component and then combine them all up to create a prediction. In this analysis we are considering data of 11 years and will be forecasting the number of passengers for the 12th year.

Does the data have a trend? Does it have a seasonal component?

Considering 12 months for the period of 12 years

The data for airline passengers is displayed with this plot showing an upward trend. The blue line in figure 1 estimates the trend line of the time series. Looking at the figure 1 we can see the number of airline passengers have increased from 1949 till the end of 1959. This trend gives us sort of a sense of the overall growth in airline travel.

We can observe a 12-month recurring sequence that repeats throughout the time and this recurring sequence indicates that seasonality is present in this data. In this dataset and we can see that the increasing and decreasing pattern is repeated year after year. The gap between the data and the trend line at each year is referred to seasonality.

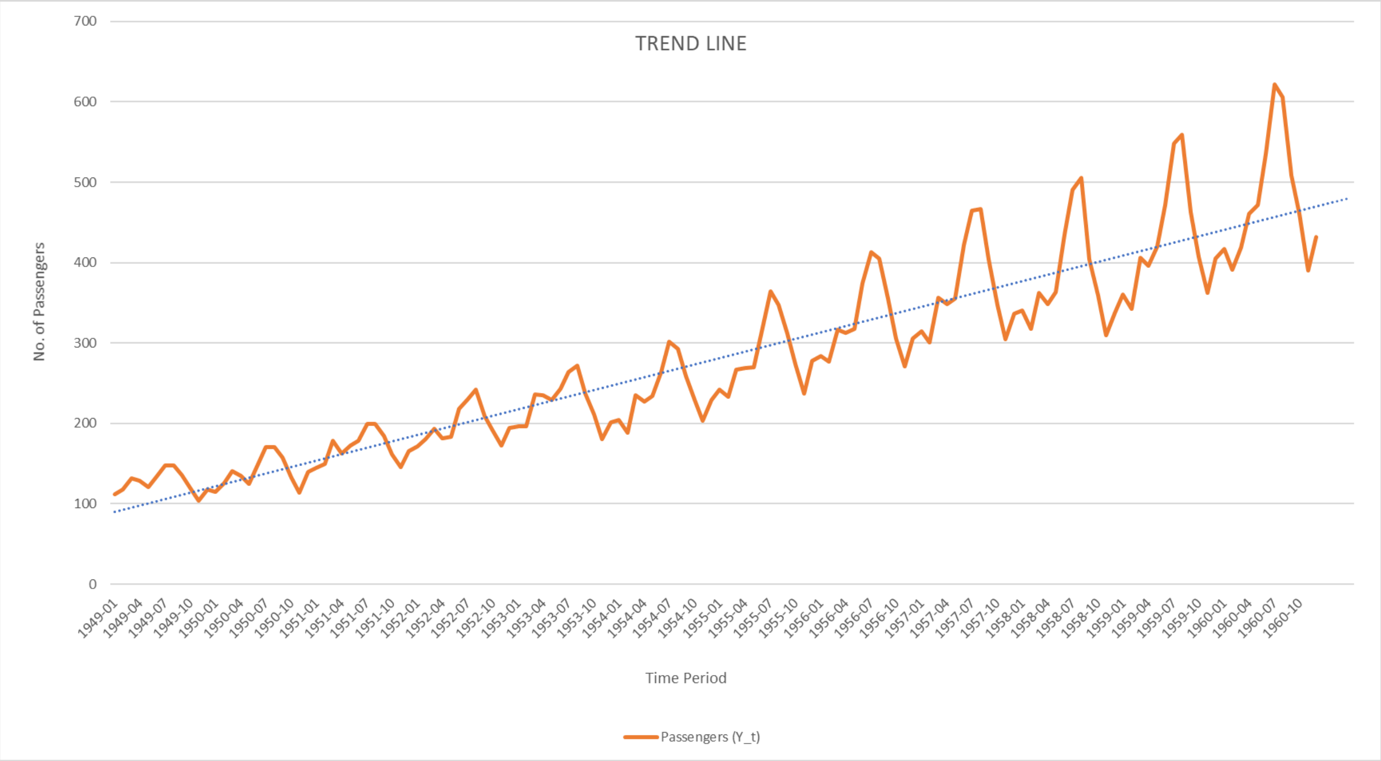


Figure 1: Trend line & Seasonality

How many seasons can be recognised in this data set?

12 Seasons are recognised in this data set.

There is an upward trend, with a similar sequence commencing every year. In each year there is a big increase in flights around the summer time (May-Aug) also we can see little increase during Christmas time(Nov-Dec).

Every year, there appears to be a seasonal variation. A noticeable peak appears in the middle of the year, indicating seasonality with a **12-months.** Perhaps due to increasing demand for flight travel of airlines in that time period

Calculate appropriate moving averages for this data set to smooth out the trend. Then calculate the seasonal components values. Provide an interpretation for the seasonal factor values.

We will use 12MA as an appropriate moving average for this dataset to smooth out the trend. We calculated the centred moving average because we had an even number of points in the average. In order to calculate seasonal component values, we fitted the multiplicative model by looking at the plot of the data.

By looking at the seasonal factor values, we can see that June, July, and August are the busiest months, with more people travelling possibly due to summer vacations. Also, if we see in the autumn there is a dip, but a minor increase in December, i.e. Christmas time.

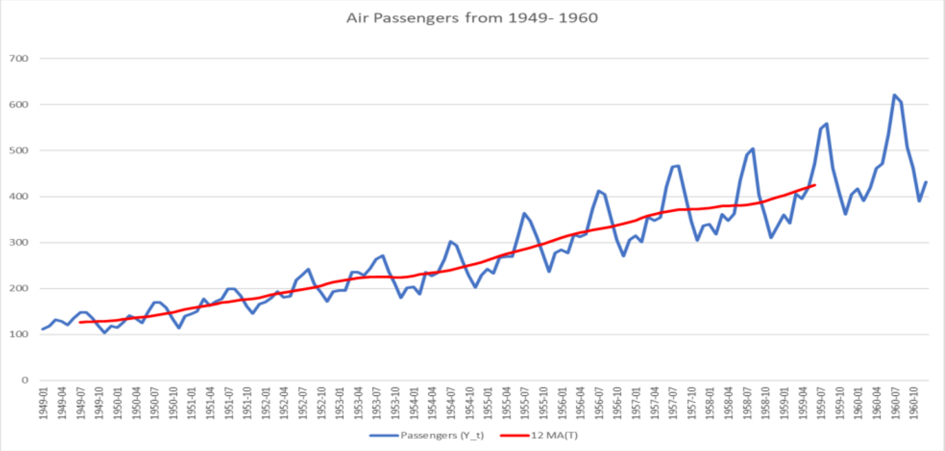
After looking at the seasonal factors we found that the values corresponding to each other in different year are not equal, they are very close to each other but not exactly the same.

For example, 1950-July = 1.206386753 but 1951-July = 1.162043796. We used the correction factor, which is equal to 12 divided by all the seasonal factors, to get the same value for each month corresponding to the year, which is termed the typical seasonal factor.

Interpretation- For the first month (January), we got a typical SF of 0.910, which suggests that the data in the first month is nearly 91 % greater than it would be if there was no seasonal effect. We obtained it in the same way for the rest of the months, refer table 1.

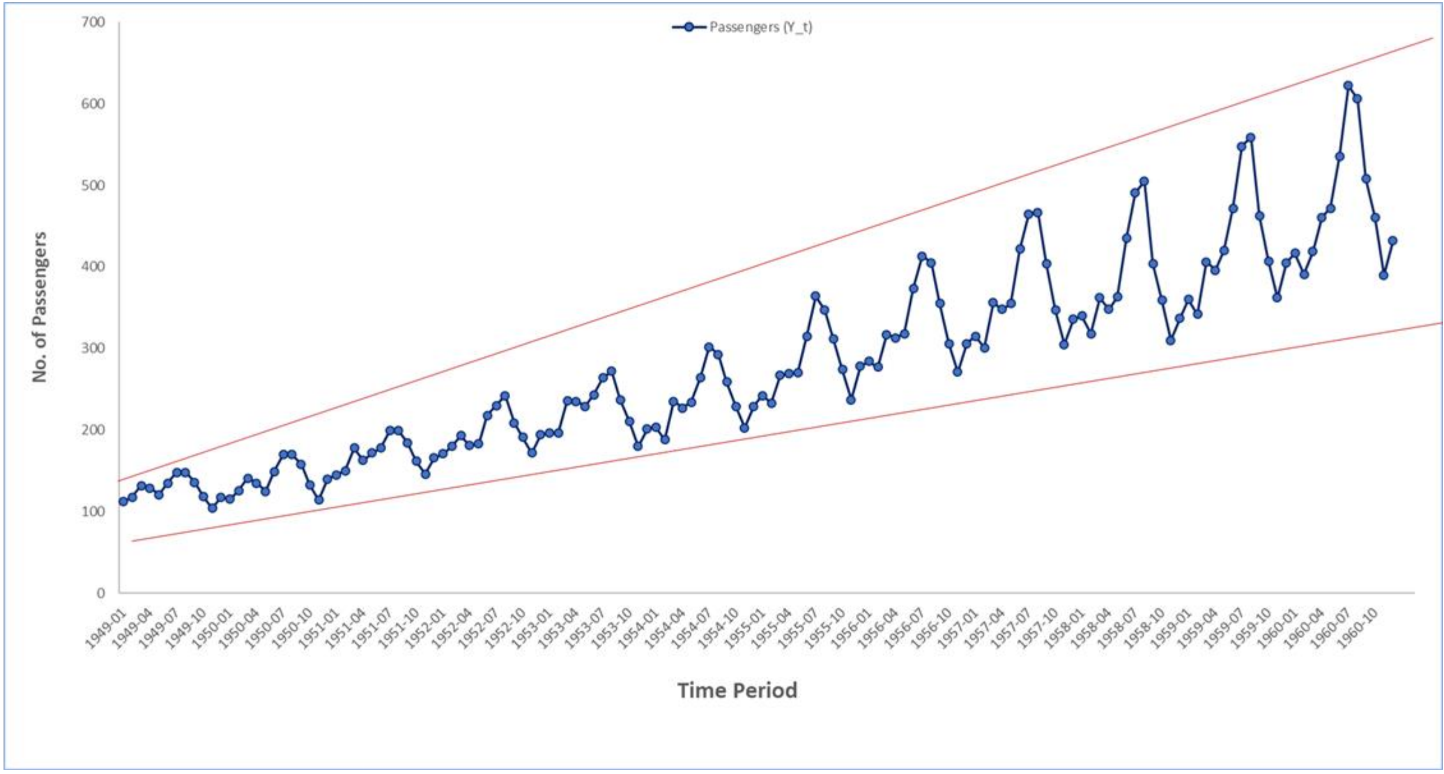
Table 3: Typical Seasonal factors

Months	Typical SF
Month 1	0.910003709
Month 2	0.887376502
Month 3	1.018203704
Month 4	0.975411976
Month 5	0.979812827
Month 6	1.111589812
Month 7	1.222146626
Month 8	1.213596104
Month 9	1.060916842
Month 10	0.921767026
Month 11	0.800213228
Month 12	0.898961644



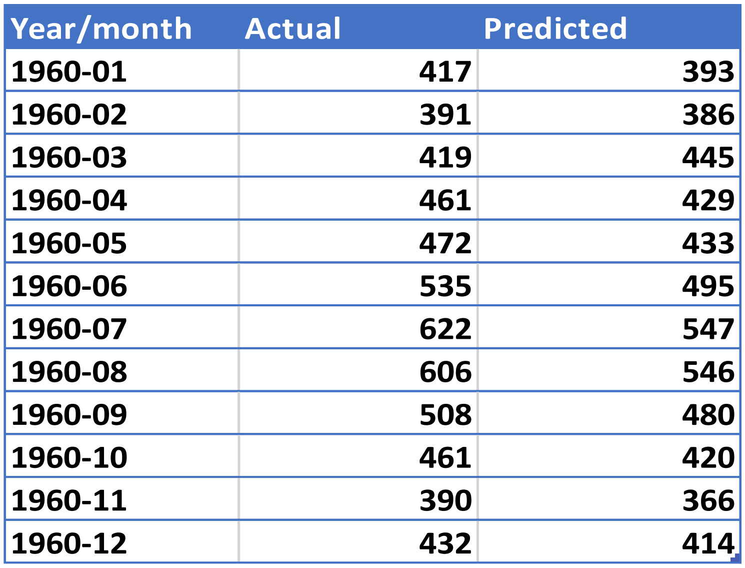
Which model describes this data set the best – additive or multiplicative? Why?

Figure 1 shows a proportional swift around the trend, indicating that multiplicative model can describe the behaviour of time series much better. The height of the cycles appears to be increasing, showing that it is multiplicative model. And the major reason for this is when seasonal variation increases over time, a multiplicative model is appropriate to use.



Next forecast the number of airline passengers for the last year according to the data of previous years.

We will forecast the values for 1960 based on the pervious data (1949-1959)
We estimated the typical seasonal factors, De-seasonalised the data, fitted a regression line between the De-seasonalised data, and forecasted future De-seasonalised values of the time series.
To forecast the number of airlines passengers we calculated the intercept and the slope and found the line of best fit. We later calculated the De-seasonalised forecast using slope and intercept y=a + b * t for predicting.
Then, using the seasonal factors and the De-seasonalized forecast, we calculated the actual prediction.



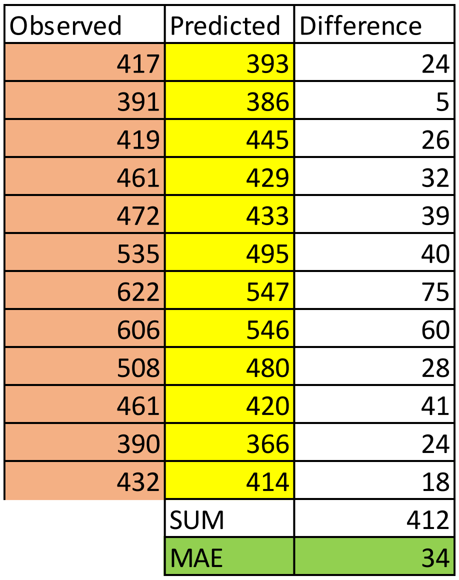
Year/month	Actual	Predicted
1960-01	417	393
1960-02	391	386
1960-03	419	445
1960-04	461	429
1960-05	472	433
1960-06	535	495
1960-07	622	547
1960-08	606	546
1960-09	508	480
1960-10	461	420
1960-11	390	366
1960-12	432	414

Finally, calculate the mean absolute error and mean square error for your forecasts.

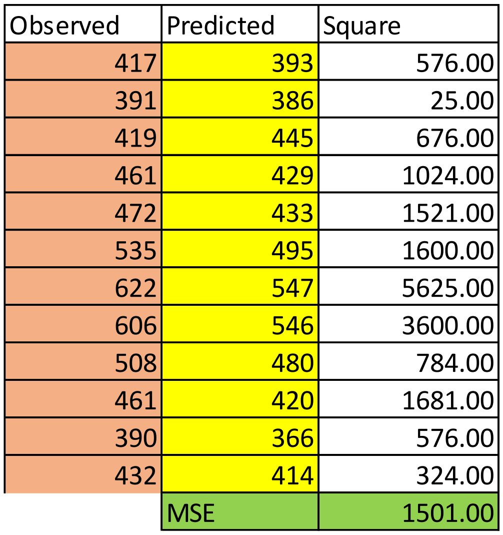
Mean Absolute Error and Mean Squared Error were used to evaluate our model.

MAE =34 which means on average the forecast distance from the actual value is 34

MSE= 1501



Observed	Predicted	Difference
417	393	24
391	386	5
419	445	26
461	429	32
472	433	39
535	495	40
622	547	75
606	546	60
508	480	28
461	420	41
390	366	24
432	414	18
	SUM	412
	MAE	34



Observed	Predicted	Square
417	393	576.00
391	386	25.00
419	445	676.00
461	429	1024.00
472	433	1521.00
535	495	1600.00
622	547	5625.00
606	546	3600.00
508	480	784.00
461	420	1681.00
390	366	576.00
432	414	324.00
	MSE	1501.00

PORTFOLIO TASK 5

Forecasting UK Covid-19 cases using ARIMA Model

INTRODUCTION

The Aim of this report is to use the ARIMA model to forecast UK covid 19 cases for the next seven days (15th – 21st June 2020).In this Analysis we have used data from: EU Open Data Portal: COVID-19 Coronavirus data from the European Centre for Disease Prevention and Control. This dataset contains the UK's Covid-19 number of cases and deaths, each day, from the 1st of January to the 14th of June 2020.

Method

We have used the ARIMA model (Autoregressive Integrated Moving Average Model) for time series data analysis and forecasting. The ARIMA model is used to analyse historical data and forecast future data in a series.

Summary of our Analysis

After inspecting the dataset, we first determined whether the data is stationary or non-stationary. Also, there was an negative value for 21st may which was` handled by using linear interpolation method. Examining the plot 1, we discovered that this is a non-stationary time series. Then we used differencing as it can change a non-stationary time series to stationary one.

After looking at the plot we can say that it's a non-stationary time series.

In this plot of time series (Figure 1), we can identify some noticeable trends and changing levels in series, such as the fact that there were no cases observed until February, and then the cases began to rise in March, and then gradually decreased thereafter. As a result, we eliminated the first part of the data (January to the end of February) and focused our analysis on the second half of the data (after March 1, 2020). After the data has been removed, we can assume from (figure 2) that the mean of the time series data is increasing over time. By this we can claim that this is a non-stationary time series, but that is insufficient to determine. Hence, we used the autocorrelation method to be more exact in determining this, which helped us claim it as a non-stationary time series.

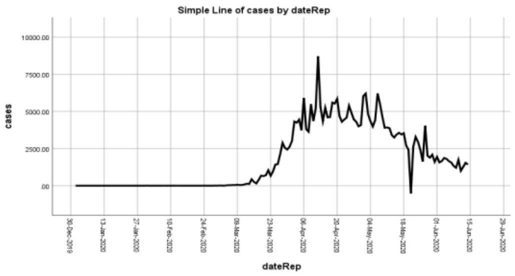


Figure 2

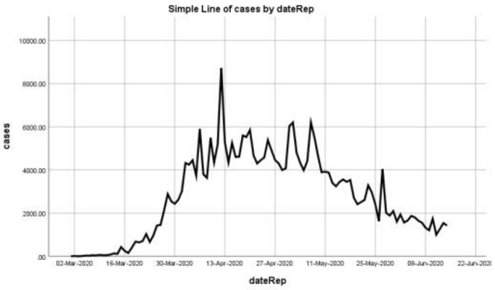


Figure 3

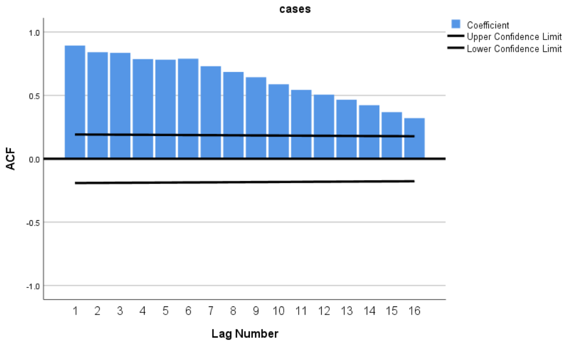
Autocorrelation method

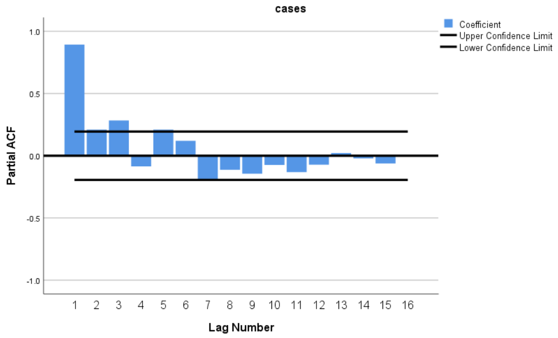
After conducting autocorrelation method, we found the following results:

In the ACF plot, we can observe that for increasing lags, all spikes are significant and slowly tend to degrade to zero, indicating that this is a non-stationary time series, since the values quickly degrade to zero for the stationary.

Also, in the PACF figure the first spike is very significant, the coefficient is almost equal to 1, and the following lags are almost zero.

From these two plots we can conclude that this is a non-stationary time series.





So, in order to make this a stationary time series, we first differenced the data before proceeding with our analysis. We have selected a difference of 1 since it is sufficient to transform non-stationary to stationary time series. Differencing has helped in the stabilisation of the mean by lowering the time series fluctuation.

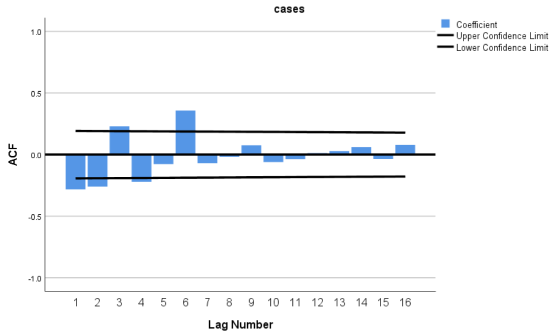
INTIAL MODEL

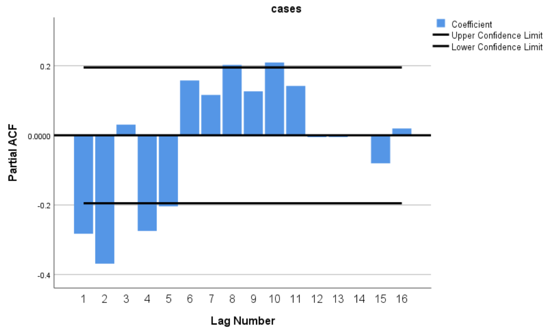
The new ACF and PACF graphs for the differenced timeseries data are shown below.

We can see from the ACF plot (Q) that the lag 1,2,3,4,6 is significant. In the PACF (P)plot, lags 1, 2, 4 and 5 are significant, while the others are nearly zero.

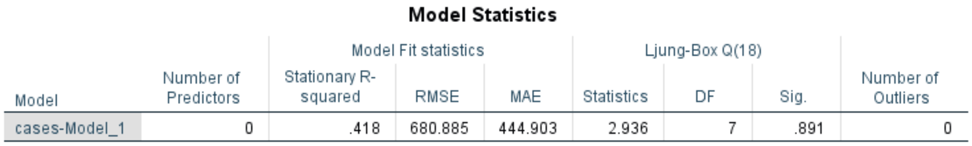
In the initial model we have considered P= 5 based on PACF, Q =6 based on ACF and D= 1 based on number of differencing.

This identifies an ARIMA (5,1,6) as a candidate to estimate our time series.





We can say that the model is acceptable based on the Ljung-Box test findings because it is more than 5% the assumption is satisfied, and the model may fit the data, but we must also examine the model parameters to claim that this is an adequate model.



Model Statistics

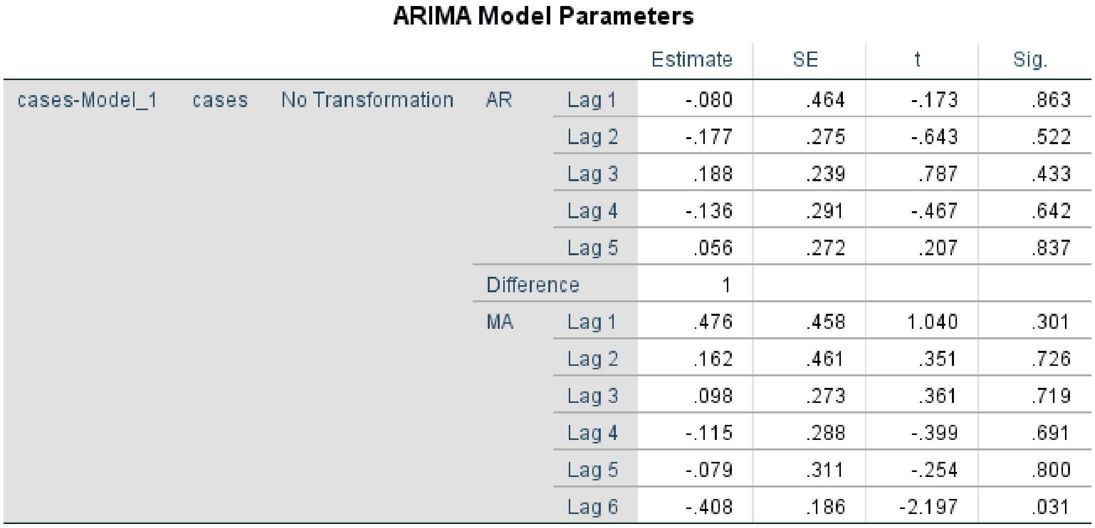
		Model Fit statistics			Ljung-Box Q(18)			
Model	Number of Predictors	Stationary R-squared	RMSE	MAE	Statistics	DF	Sig.	Number of Outliers
cases-Model_1	0	.418	680.885	444.903	2.936	7	.891	0

We discovered the following result when we looked at the model parameters.

In AR section, none of the lags are significant because they are more than 5%.

In the MA section, we can see that lag 1 to 5 are insignificant, whereas lag 6 is significant.

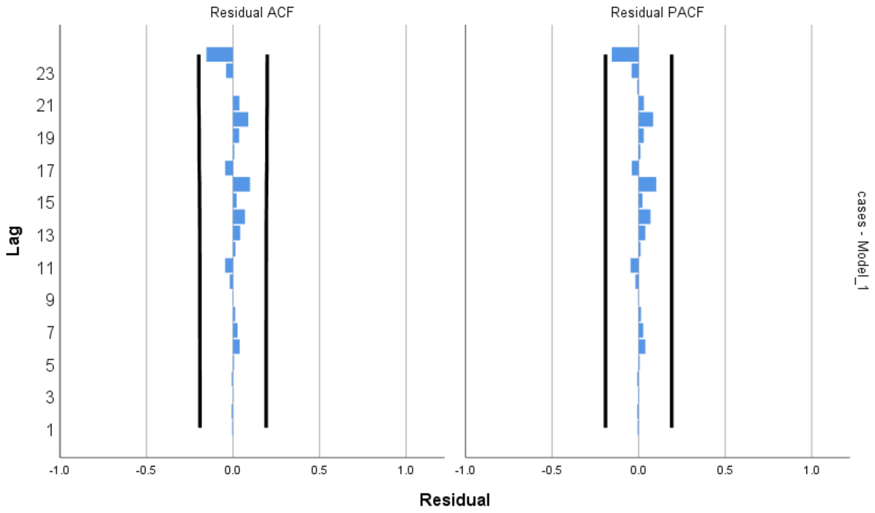
To achieve a more parsimonious model, we need to remove the insignificant lags from our model.



ARIMA Model Parameters

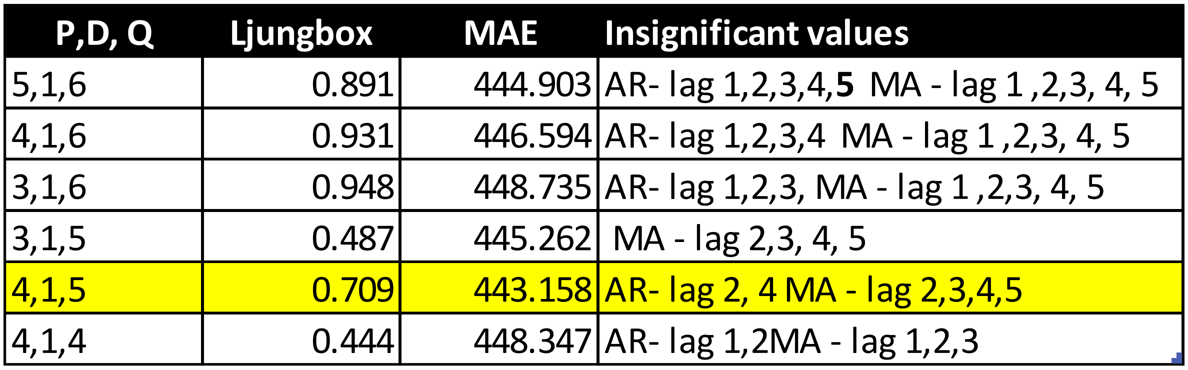
					Estimate	SE	t	Sig.
cases-Model_1	cases	No Transformation	AR	Lag 1	-.080	.464	-.173	.863
				Lag 2	-.177	.275	-.643	.522
				Lag 3	.188	.239	.787	.433
				Lag 4	-.136	.291	-.467	.642
				Lag 5	.056	.272	.207	.837
			Difference		1			
			MA	Lag 1	.476	.458	1.040	.301
				Lag 2	.162	.461	.351	.726
				Lag 3	.098	.273	.361	.719
				Lag 4	-.115	.288	-.399	.691
				Lag 5	-.079	.311	-.254	.800
				Lag 6	-.408	.186	-2.197	.031

Looking at the residual plot of ACF and PACF we can observe that all the lags are within their significance interval which assumes that this model is predicting the behaviour of time series well. However, for a better model, we had to rerun the model, removing the insignificant lags one by one.



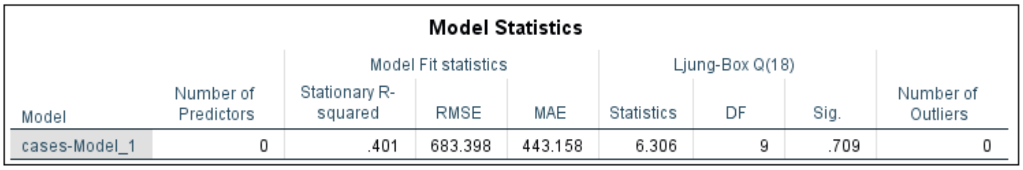
FINAL MODEL

After removing all the insignificant lags, we found the ARIMA (4,1,5) model as our parsimonious model.



P,D, Q	Ljungbox	MAE	Insignificant values
5,1,6	0.891	444.903	AR- lag 1,2,3,4,**5** MA - lag 1 ,2,3, 4, 5
4,1,6	0.931	446.594	AR- lag 1,2,3,4 MA - lag 1 ,2,3, 4, 5
3,1,6	0.948	448.735	AR- lag 1,2,3, MA - lag 1 ,2,3, 4, 5
3,1,5	0.487	445.262	MA - lag 2,3, 4, 5
4,1,5	0.709	443.158	AR- lag 2, 4 MA - lag 2,3,4,5
4,1,4	0.444	448.347	AR- lag 1,2MA - lag 1,2,3

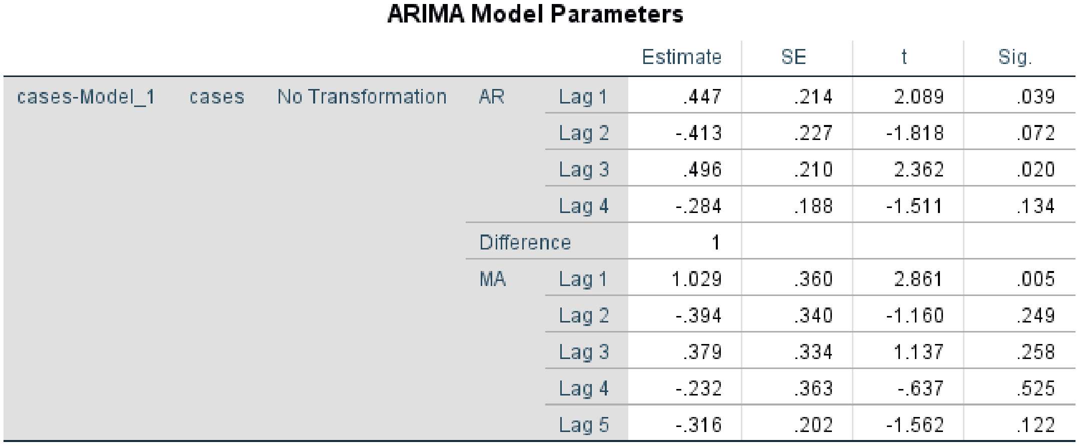
The final model is ARIMA 4,1,5 because the MAE value has been reduced, and the Ljungbox test suggests that it is an adequate model because it is more than 5%, and we conclude that this is the final model after comparing it to other models.



Model Statistics

Model	Number of Predictors	Model Fit statistics			Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	RMSE	MAE	Statistics	DF	Sig.	
cases-Model_1	0	.401	683.398	443.158	6.306	9	.709	0

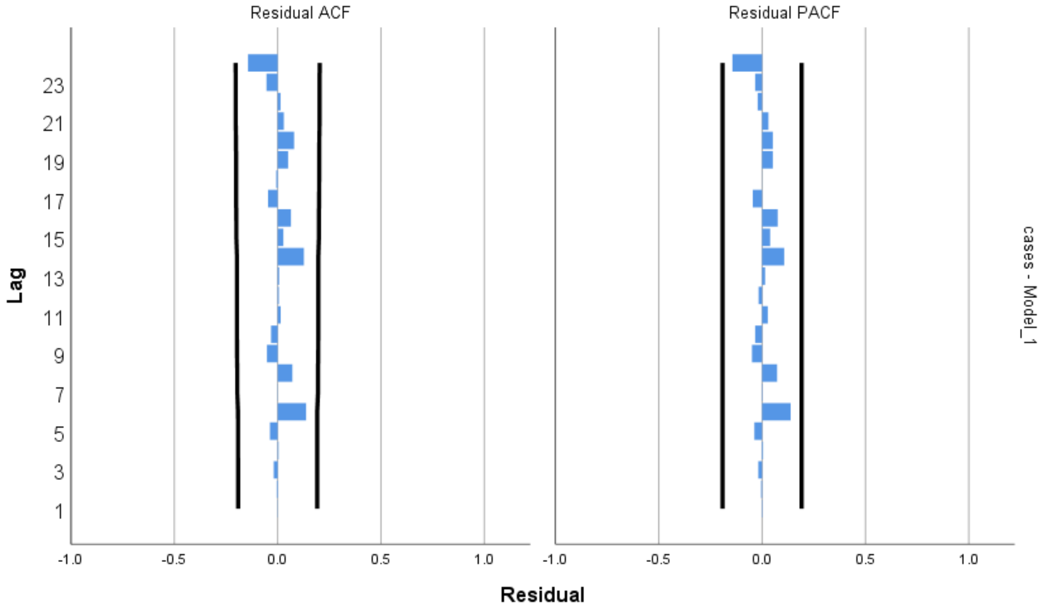
Looking at the model parameters, in AR section we can see that lag 1 and 3 is significant, but lag 4 is not. However, there are insignificant (lag 4 in AR & lag5 in MA) which could be eliminated but after removing those the MAE is increasing, hence we considered ARIMA (4,1,5) to be our final one.



ARIMA Model Parameters

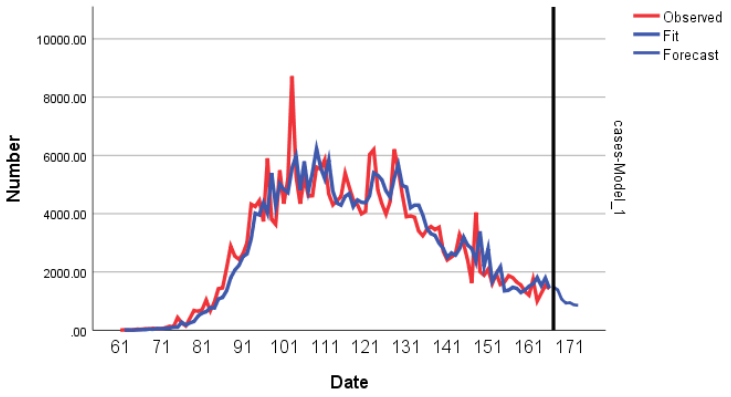
					Estimate	SE	t	Sig.
cases-Model_1	cases	No Transformation	AR	Lag 1	.447	.214	2.089	.039
				Lag 2	-.413	.227	-1.818	.072
				Lag 3	.496	.210	2.362	.020
				Lag 4	-.284	.188	-1.511	.134
			Difference		1			
			MA	Lag 1	1.029	.360	2.861	.005
				Lag 2	-.394	.340	-1.160	.249
				Lag 3	.379	.334	1.137	.258
				Lag 4	-.232	.363	-.637	.525
				Lag 5	-.316	.202	-1.562	.122

Finally, we can see that all of the lags in the residual plot of ACF and PACF are inside their significance intervals, indicating that this model correctly predicts the behaviour of time series and that it is the most parsimonious model.



We predicted the number of Covid-19 cases for the next seven days using ARIMA (4,1,5). We can also observe the line graph below which shows the observed cases (Red line) and the forecasted value (Blue line). The predicted cases are listed in the table below.

Date	PREDICTED CASES
15 JUN 2020	1480.38
16 JUN 2020	1394.52
17 JUN 2020	1064.97
18 JUN 2020	932.08
19 JUN 2020	946.28
20 JUN 2020	868.46
21 JUN 2020	855.52



If we could have eliminated lags that were less important but were in the centre of two significant lags, we could have developed a more parsimonious model. Unfortunately, this was not achievable with SPSS.

PORTFOLIO TASK 6

Forecasting the exchange rate using an artificial neural network in IBM SPSS.

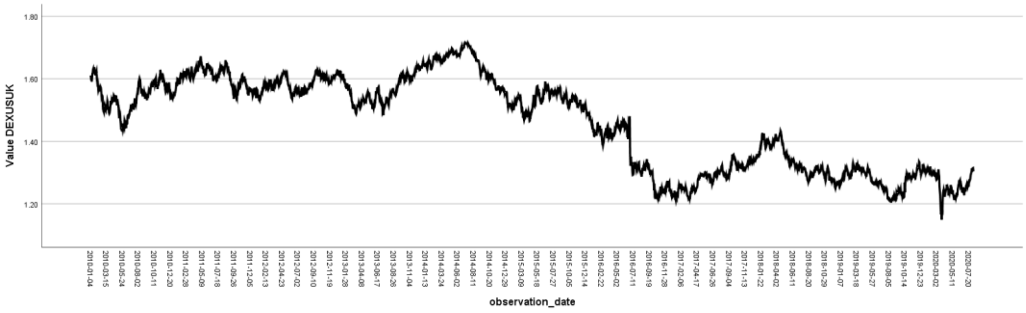
INTRODUCTION

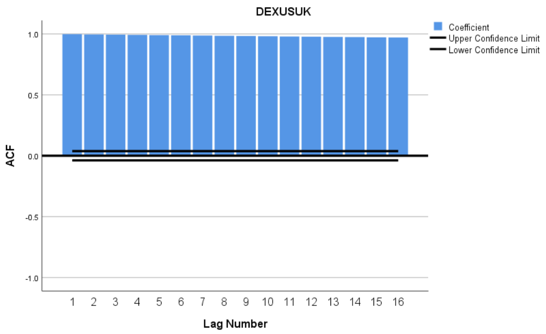
In this report we have forecasted the US dollar against the UK pound (GBP) using an Artificial Neural Network with help of the features of exchange rates time series, by developing the best neural network to know the accuracy of forecasted values. The dataset we have used in this analysis includes the daily exchange rate from January 4, 2010 until August 7, 2020. As a result our goal is to forecast the exchange rate for 8th August 2020.

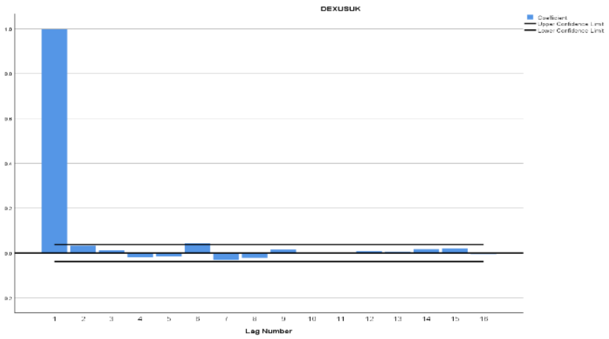
Selection of inputs and outputs

There were many NA values in the initial data, which were treated by taking the average of the previous four days because we can't consider values below the NA for this averaging because it's not appropriate to consider future exchange rate values.

Using the auto correlation method, we chose the predictors (inputs) and forecasts (output). After examining the Autocorrelation Function (ACF) and partial auto-correlation function (PACF) plots, we discovered that there are 6 lags, which assisted us in selecting our inputs and output. As seen in the PACF plot, there are 6 significant lags, while in the ACF plot, all spikes are significant. These two plots assisted us in fitting the neural network into this timeseries dataset, as well as selecting 6 inputs as the lag values of the time series and 1 output as the current value of time series. To put it another way, the exchange rate at y(t) is an output, whereas the exchange rates at y(t-1), y(t-2), y(t-3), y(t-4), y(t-4), y(t-5), and y(t-6) are inputs.





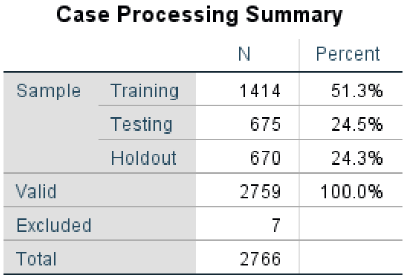


Output from the SPSS and interpretation of the results.

In this analysis we used a multilayer perceptron (MLP) neural network model to estimate the exchange rate because it is a good neural network model for financial forecasting.

After selecting the inputs and output, We first chose the output as dependent variable and the inputs as covariates. To improve network training, these covariates were rescaled by using the Standardized method (subtract the mean and divide by the standard deviation).

Later The current dataset was partitioned into training, testing, and holdout samples by the Partition Dataset method. Looking at the case processing summary, we can see the assigned percentage that is 50% of the sample of whole dataset for training, 25% for testing the neural network, and 25% for holdout.



Case Processing Summary

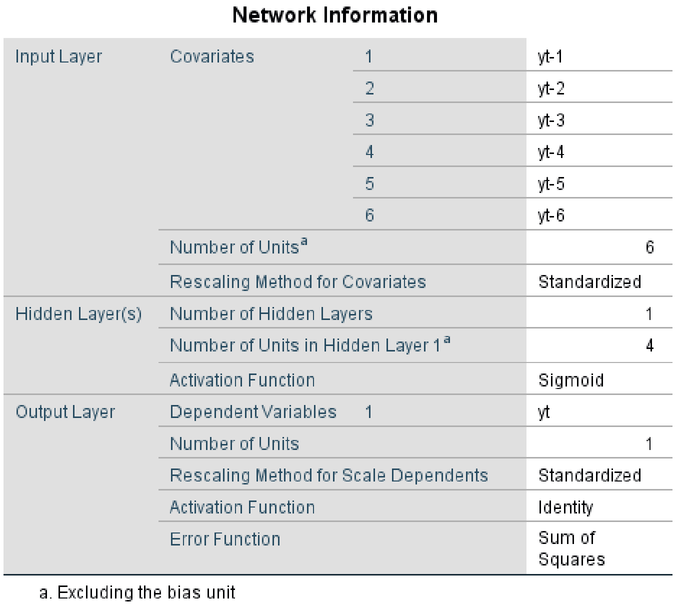
		N	Percent
Sample	Training	1414	51.3%
	Testing	675	24.5%
	Holdout	670	24.3%
Valid		2759	100.0%
Excluded		7	
Total		2766	

ARCHITECTURE OF NEURAL NETWORK

As multilayer perceptron consists of input layer, output later and one or more hidden layer with sigmoid activation function in the hidden neurons. We used 1 hidden layer perceptron for this time series analysis and the software helped to select the best hidden neurons. We used the Sigmoid for the input layer and the identity activation function for the output layer because it connects the weighted sums of units in one layer to the values of units in the next layer.

Batch training was chosen because it uses all the data from training dataset to determine the network that helps to handle the records.

We described the neural network model of this timeseries using the Network Information table. Looking at the network information and network diagram, we observe that there are 6 units in our Input layer as expected and one hidden layer with 4 units (selected by SPSS) and output layer with 1 unit.

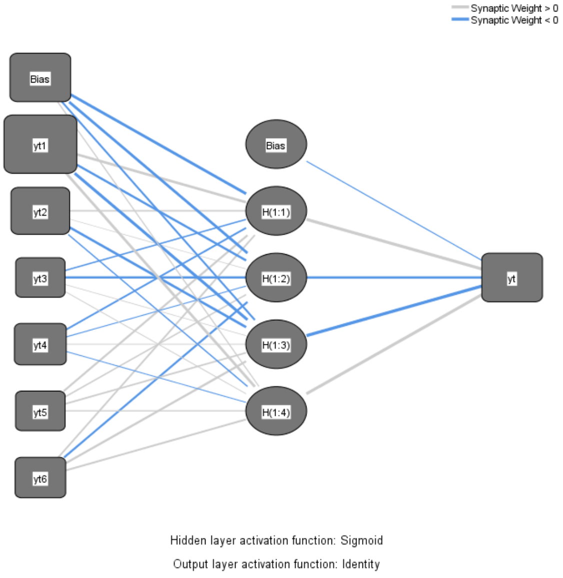


Network Information

Input Layer	Covariates	1	yt-1
		2	yt-2
		3	yt-3
		4	yt-4
		5	yt-5
		6	yt-6
	Number of Units[a]		6
	Rescaling Method for Covariates		Standardized
Hidden Layer(s)	Number of Hidden Layers		1
	Number of Units in Hidden Layer 1[a]		4
	Activation Function		Sigmoid
Output Layer	Dependent Variables	1	yt
	Number of Units		1
	Rescaling Method for Scale Dependents		Standardized
	Activation Function		Identity
	Error Function		Sum of Squares

a. Excluding the bias unit

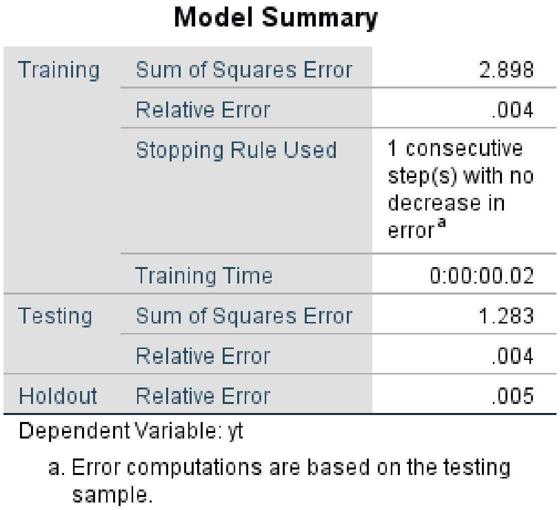
Positive synaptic weight (>0) is represented by a grey line, whereas negative synaptic weight (0) is represented by a blue line in this network diagram. Also, there is one bias unit at each input and hidden layer which helps in increasing the model accuracy .



NNAR (6,4) is the neural network defining our time series, according to this information from our model.

Performance analysis

The model summary explains the details of our neural network. The relative error is nearly 0.4% in each training, testing, and 0.5% in holdout phase, which is a good sign for our model because smaller error values indicate a better model. Specifically, the holdout phase shows better fitting and forecasting, so NNAR (6,4) works well to forecast the exchange rate.



Model Summary

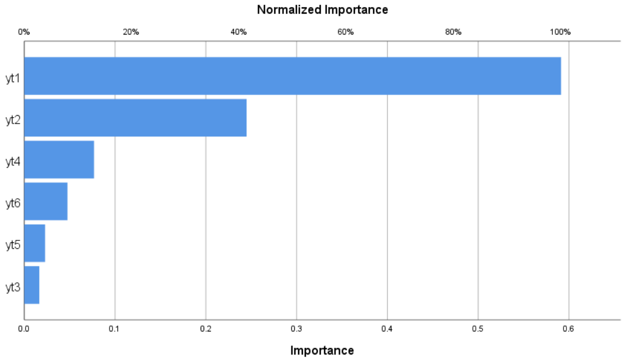
Training	Sum of Squares Error	2.898
	Relative Error	.004
	Stopping Rule Used	1 consecutive step(s) with no decrease in error[a]
	Training Time	0:00:00.02
Testing	Sum of Squares Error	1.283
	Relative Error	.004
Holdout	Relative Error	.005

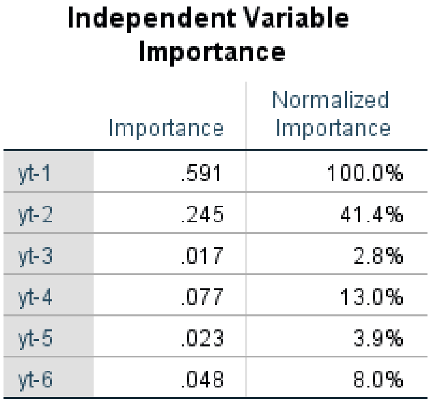
Dependent Variable: yt

a. Error computations are based on the testing sample.

Independent variable importance analysis.

The table and chart below is the result of sensitivity analysis which was conducted to determine the importance of each predictor in determining the neural network. The analysis is based on a sample set that includes both training and testing. This table and chart shows that input 1 (yt-1) is the most important predictor also yt-2 contributes well in this forecasting and the yt-3 is the least important input.

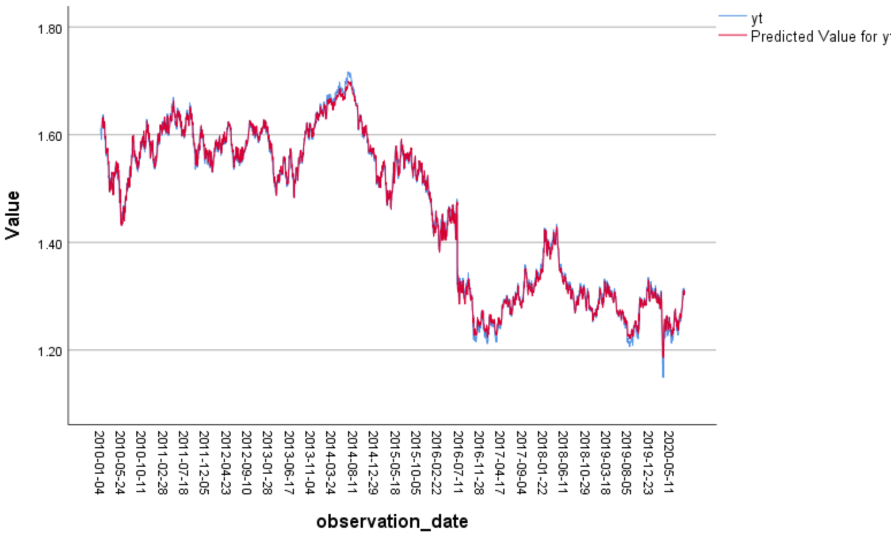




Independent Variable Importance

	Importance	Normalized Importance
yt-1	.591	100.0%
yt-2	.245	41.4%
yt-3	.017	2.8%
yt-4	.077	13.0%
yt-5	.023	3.9%
yt-6	.048	8.0%

Graph representing the performance of the forecasting



In the above graph, the blue line indicates the actual rate of exchange, while the red line reflects the predicted values of exchange rate.

We can see from the graph that the two lines closely overlap, which shows a good prediction.

RESULT

Forecasted exchange rate for 8th August 2020 is 1.3037 $ which is the one step ahead prediction for time series variable