BUSINESS ANALYTICS IN PRACTICE

PORTFOLIO TASKS

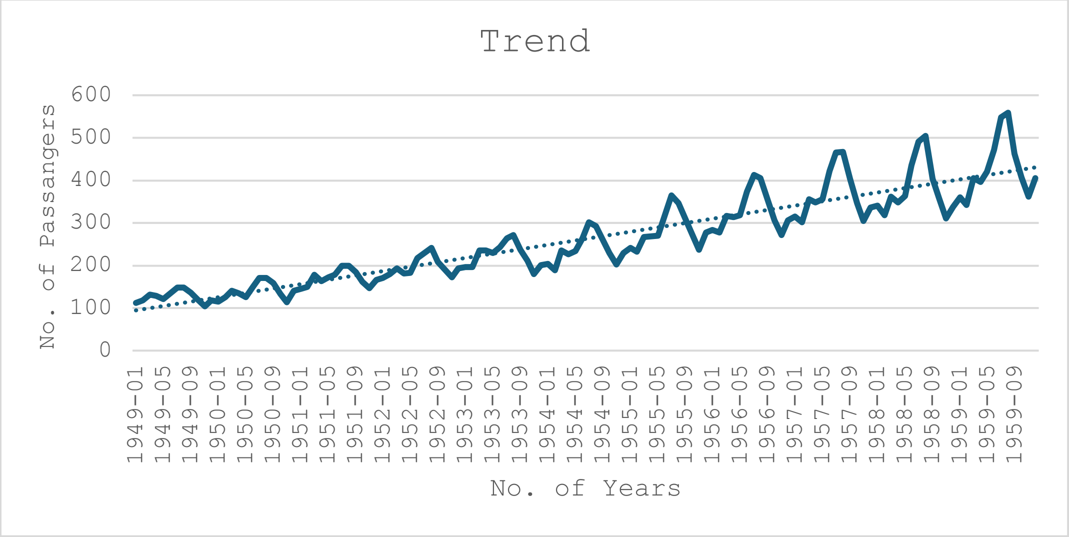
TABLE OF CONTENTS

Portfolio: Task 1

Time series forecasting with the decomposition technique.

Predicting future results and making well-informed decisions depend heavily on a grasp of prior data trends in the quickly developing subject of business analytics. We conducted a study of the monthly passenger counts using a widely referenced time series dataset. Our findings shed light on consumer behavior, seasonal fluctuations, and the industry's general growth over the past ten years. To guarantee a thorough analysis, we exclude the data from the dataset's last year. This lets us concentrate on the identified patterns without being influenced by missing or anomalous data from the last year.

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



The Graph includes the data of Airline passengers from 1949 to 1959 and shows a clear upward trend, indicating that the number of passengers increases over time. We can interpret that the data has both trend and a seasonal component as there's a visible seasonal pattern within each year, with peaks and troughs recurring at regular intervals.

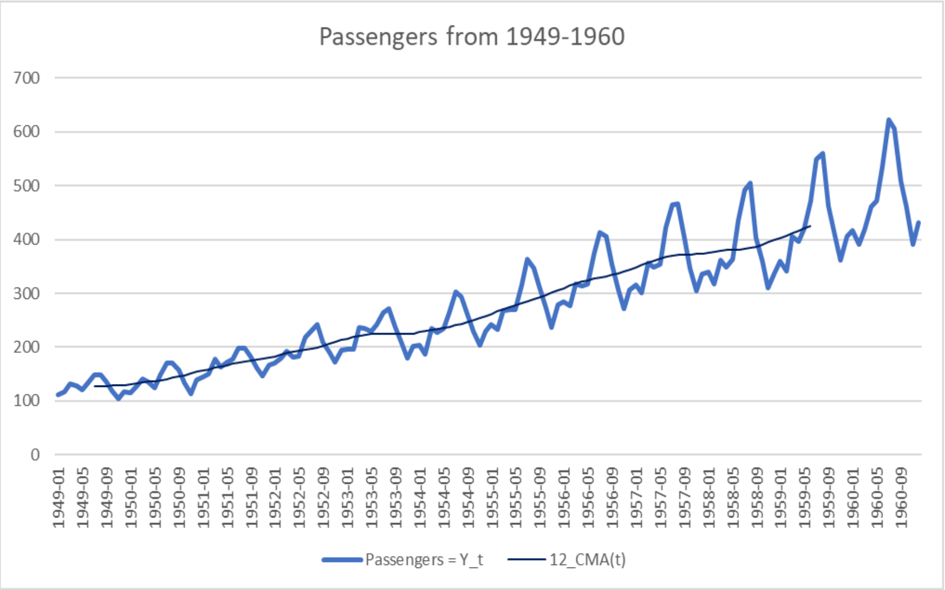
How many seasons can be recognized in this data set?

There are probably 12 seasons (one for each month) because the seasonal patterns in the plot match the months in a year. Regular increases and falls in passenger numbers characterize these patterns, which correspond to seasonal travel trends.

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

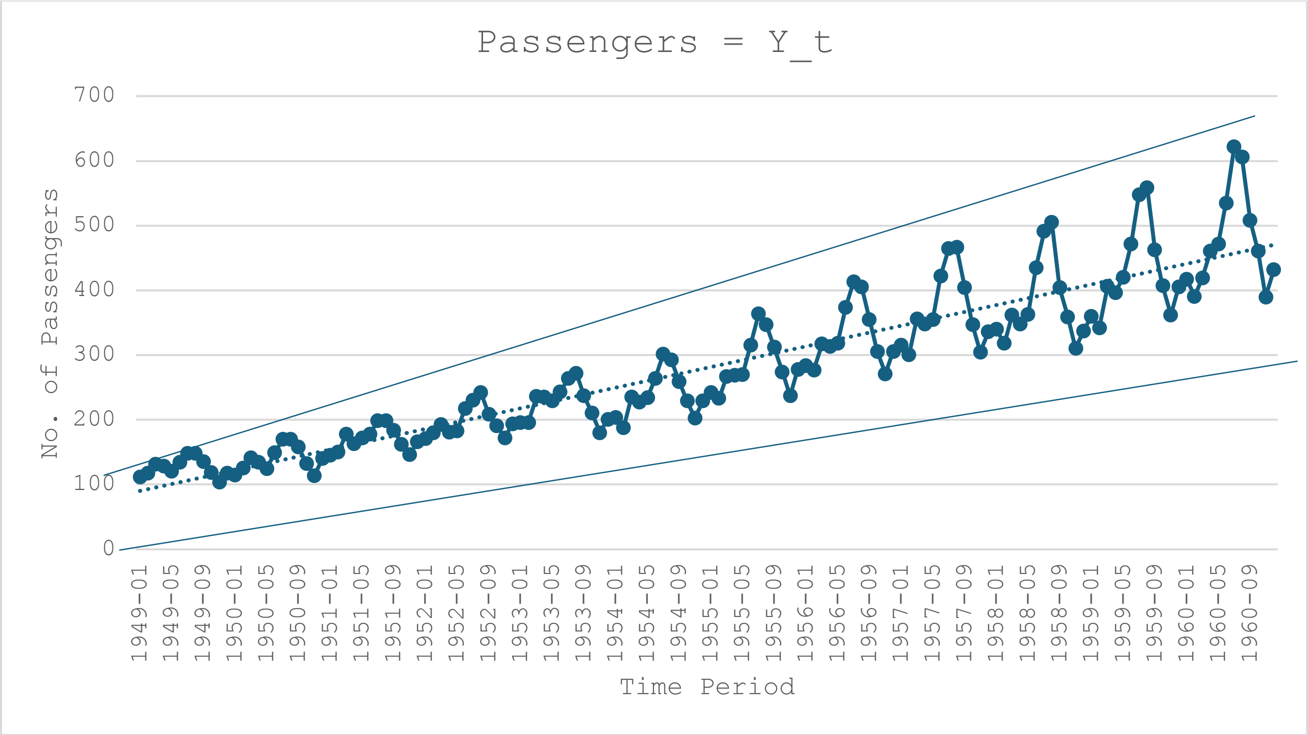
The 12-month moving average has been calculated and plotted alongside the original number of passengers. The moving average smooths out the data, making it easier to identify the underlying trend without seasonal fluctuations. As expected, the first 11 months don't have moving average values due to the lack of preceding data to calculate the average. We can observe that June, July & August are the busiest months according to the seasonal factor values. We can also observe that during Autumn there is slight decrease and in December the number of passengers increases. We can interpret those values corresponding to each other in year are not the same according to the seasonal. For example, 1951- June has the value of 1.052735338 but 1952-06 is 1.113191489. To acquire the same number for each month that corresponds to the year, we utilized the corrective factor, which is equal to 12 divided by all the seasonal factors. This is known as the typical seasonal factor.

Months	St
1	0.91000371
2	0.8873765
3	1.0182037
4	0.9754120
5	0.9798128
6	1.1115898
7	1.2221466
8	1.2135961
9	1.0609168
10	0.9217670
11	0.8002132
12	0.89896164

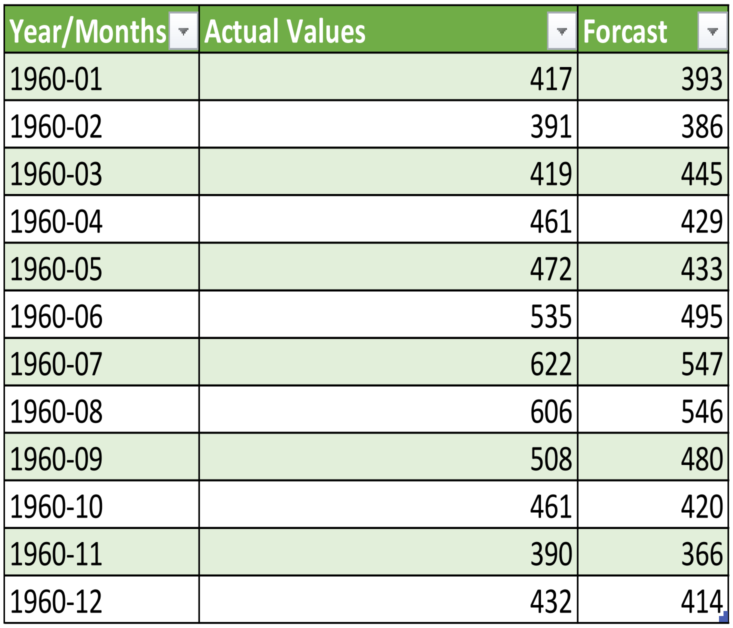


Which model describes this data set the best – additive or multiplicative? Why?

By observing the visible trend and seasonality, we have initially proceeded with a multiplicative model since the seasonal variation seems to increase with the trend, a characteristic more aligned with multiplicative models.



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

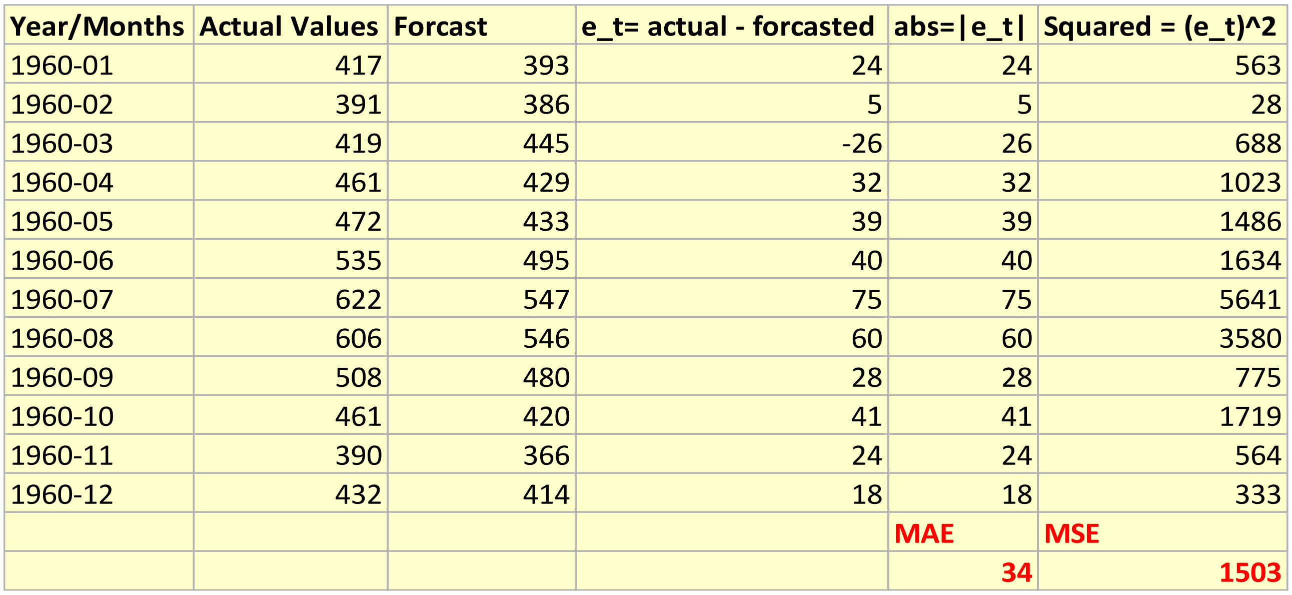


Year/Months	Actual Values	Forcast
1960-01	417	393
1960-02	391	386
1960-03	419	445
1960-04	461	429
1960-05	472	433
1960-06	535	495
1960-07	622	547
1960-08	606	546
1960-09	508	480
1960-10	461	420
1960-11	390	366
1960-12	432	414

With reference to the historical data (1949–1959), we shall Forecast the values for 1960. We de-seasonalized the data, evaluated the typical seasonal components, fitted a regression line between the de-seasonalized data, and forecast the time series' future de-seasonalized values. We determined the intercept, slope, and line of best fit to forecast the number of airline passengers. Later, we used the slope and intercept formula, y=a + b * t, to create the De-seasonalized forecast for prediction. We then computed the actual prediction using the De-seasonalized forecast and the seasonal parameters.

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

To Evaluate the model, we calculated the Mean Absolute Error and Mean Squared Error. Our MAE = 34.37 which means on average the forecast distance from the actual value is 34.37 MSE=1502.89



Year/Months	Actual Values	Forcast	e_t= actual - forcasted	abs=|e_t|	Squared = (e_t)^2
1960-01	417	393	24	24	563
1960-02	391	386	5	5	28
1960-03	419	445	-26	26	688
1960-04	461	429	32	32	1023
1960-05	472	433	39	39	1486
1960-06	535	495	40	40	1634
1960-07	622	547	75	75	5641
1960-08	606	546	60	60	3580
1960-09	508	480	28	28	775
1960-10	461	420	41	41	1719
1960-11	390	366	24	24	564
1960-12	432	414	18	18	333
				MAE	MSE
				34	1503

Portfolio: Task 2

Forecasting UK Covid-19 cases

INTRODUCTION

This study examines the COVID-19 status in the UK from 1 January to 14 June 2020, with a particular emphasis on daily confirmed cases and fatalities. We seek to determine patterns and peak times of the early phases of the pandemic in the UK using data from the European Centre for Disease Prevention and Control via the EU Open Data Portal. To fight the global health epidemic, this analysis aims to shed light on the transmission and effects of COVID-19. It does this by helping future research and policy development.

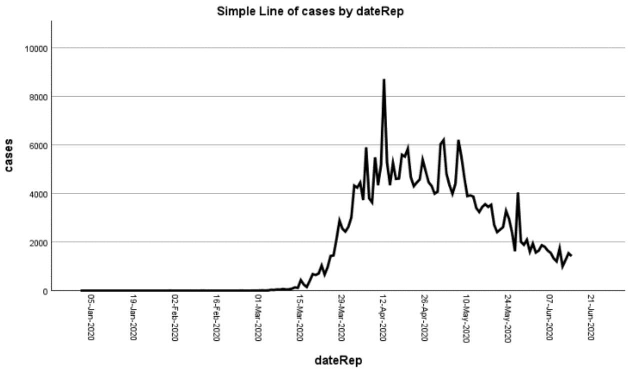
Method

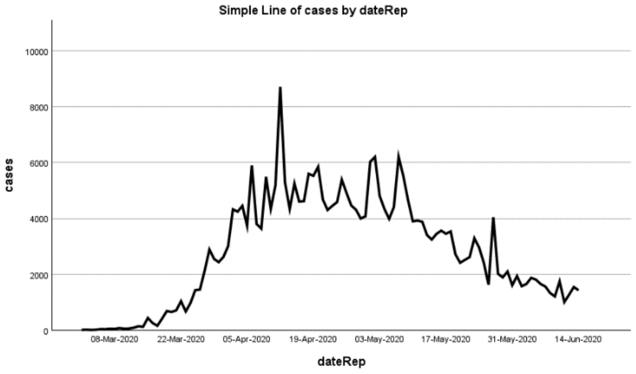
The ARIMA model (Autoregressive Integrated Moving Average Model) has been used in this study for time series data analysis and forecasting, historical data analysis and future data in a series.

Analysis

We initially ascertained whether the data is stationary or non-stationary by looking over the dataset. Additionally, the 21st of May had a negative value, which was handled by taking the average of the upper and lower value from the dataset. By observing the plot we found out that our data was Non-stationary hence, we carried out differencing which can convert a non-stationary time series into a stationary one.

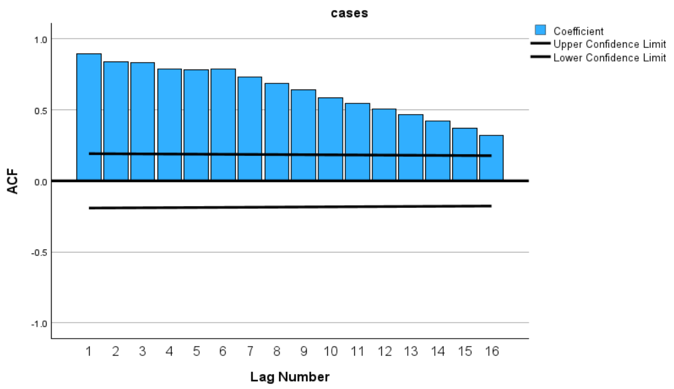
We can see some clear trends and varying series levels in the time series plots, such as the fact that no cases were seen until February, after which they started to grow in March and then progressively declined. To concentrate our research on the second half of the data (after March 1, 2020), we removed the initial portion of the data, which included the months of January through February 28. We can infer that the time series data mean is rising with time once the data has been eliminated. This allows us to assert that the time series is non-stationary, but it is not enough to decide. Thus, we employed the autocorrelation technique to be more precise which resulted in determining it as a non-stationary time series.

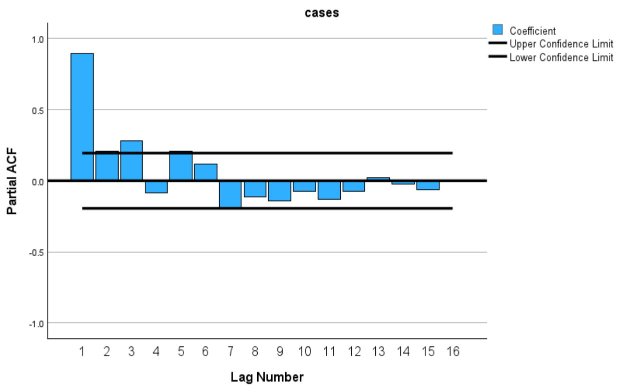




Autocorrelation

As the data is non-stationary, the ACF plot indicates a large initial positive correlation that progressively drops, showing a significant influence of past values on future ones. The PACF plot suggests that the influence of an autoregressive component is greatest at the first lag, with a substantial correlation at the first lag and sporadic significance at higher lags. These trends point to the potential value of autoregressive models for forecasting and hint that the data may require differencing to attain stationarity.

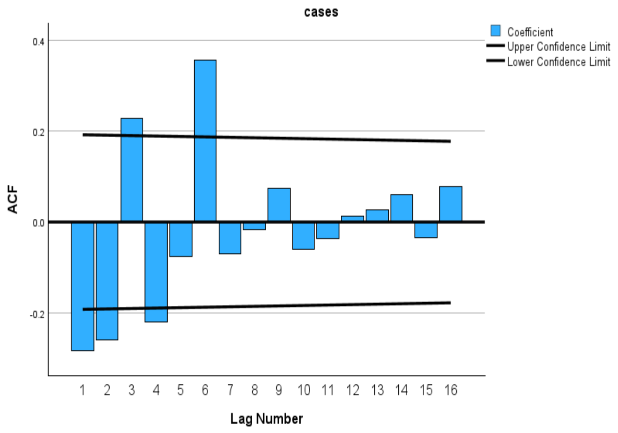


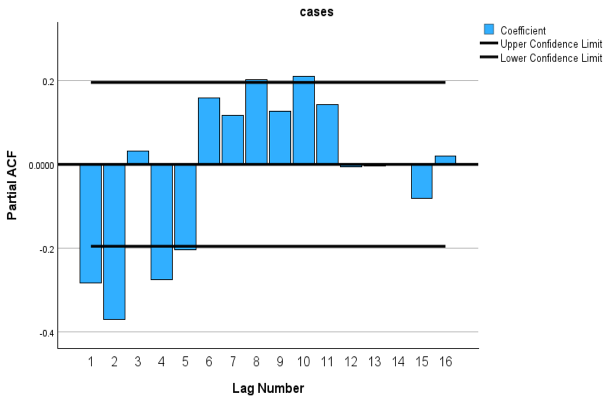


Therefore, before continuing with our study, we first did the differencing of data to create a stationary time series. Because it is adequate to convert non-stationary time series into stationary ones, we have chosen a difference of 1. By reducing the time series fluctuation, differencing has contributed to the mean's stability.

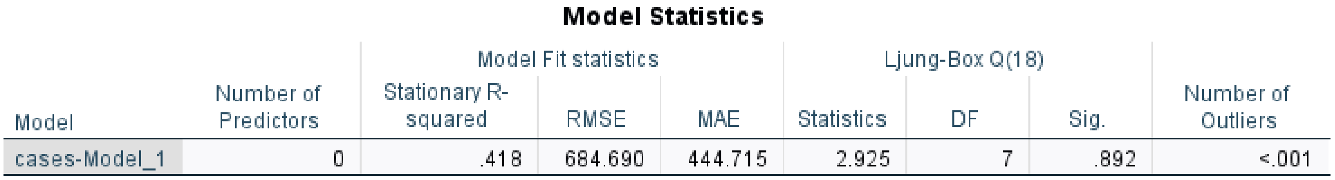
First Model

The autocorrelation is considerable for a few lags (lags 2 and 3 in particular), as the ACF plot shows, with the Lags going over the upper confidence limit. After taking into consideration the values at all intervening lags, the PACF plot indicates strong negative partial autocorrelation at lags 1, 2, and 4. This signifies that the value at those lags is inversely connected to the current value. A mixed autoregressive moving average (ARMA) process or an over-differencing may be appropriate for modelling this time series, as indicated by the alternating signals and the rapid drop-off in the PACF plot. To Estimate our time series, we identified an ARIMA Model of (5,1,6) to check the adequacy of the model.





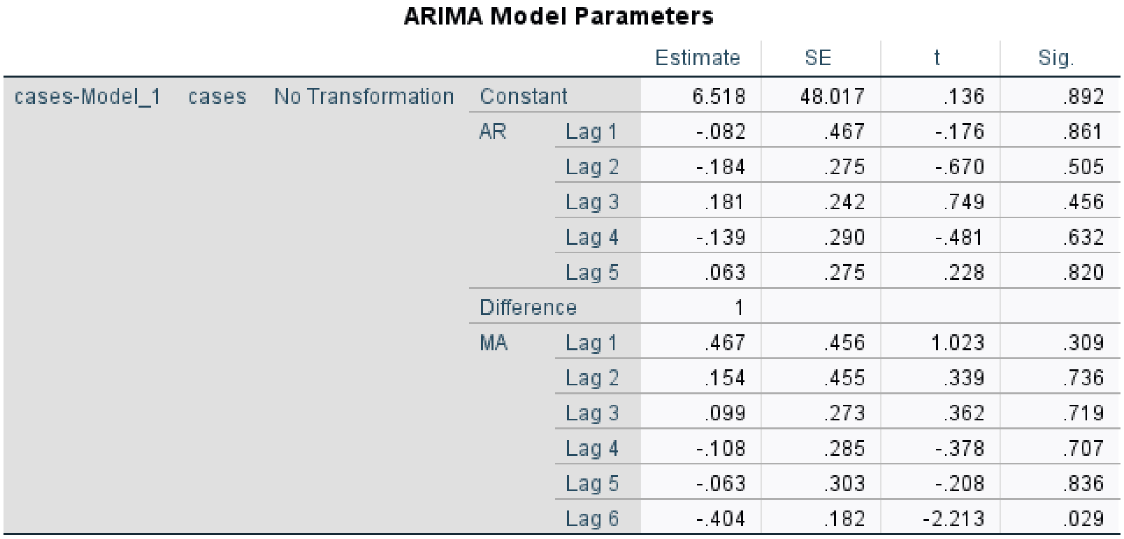
Based on the results of the Ljung-Box test, we can conclude that the model is acceptable because the assumption is satisfied in more cases than 5% of the cases. The model may also match the data, but we need to look more closely at the model's parameters before we can conclude that it is suitable.



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	684.690	444.715	2.925	7	.892	<.001

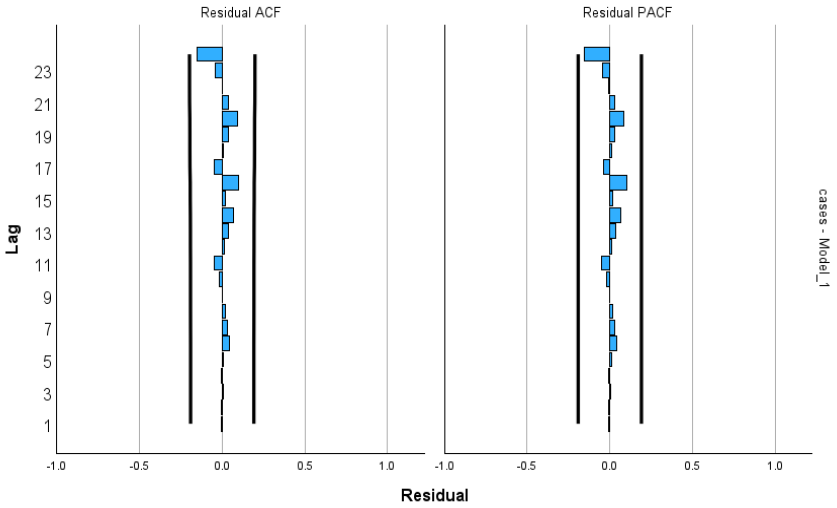
Examining the model parameters revealed the following outcome. Since all the lags in the AR portion are greater than 5%, none of them are significant. Lags 1 to 5 are shown to be insignificant in the MA section except lag 6 which is significant. The insignificant lags in our model must be eliminated to produce a more parsimonious model.



ARIMA Model Parameters

					Estimate	SE	t	Sig.
cases-Model_1	cases	No Transformation	Constant		6.518	48.017	.136	.892
			AR	Lag 1	-.082	.467	-.176	.861
				Lag 2	-.184	.275	-.670	.505
				Lag 3	.181	.242	.749	.456
				Lag 4	-.139	.290	-.481	.632
				Lag 5	.063	.275	.228	.820
			Difference		1			
			MA	Lag 1	.467	.456	1.023	.309
				Lag 2	.154	.455	.339	.736
				Lag 3	.099	.273	.362	.719
				Lag 4	-.108	.285	-.378	.707
				Lag 5	-.063	.303	-.208	.836
				Lag 6	-.404	.182	-2.213	.029

All the lags in the residual plots of the ACF and PACF are within their significance intervals, indicating that the model is presumably doing a good job of forecasting the behavior of time series. But to improve the model, we had to run it again and eliminate each lags one at a time.

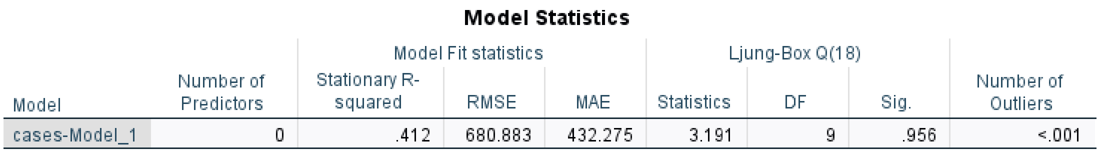


Final Model

ARIMA (5,1,4) model is our most parsimonious mode after removing all the insignificant lags.

P, D, Q	MAE	Ljung box Q	Insignificant Lags
5,1,6	444.715	0.892	AR- lag 1- 5 MA - lag 1- 5
5,1,4	**432.275**	**0.956**	**None**
5,1,3	479.796	0.096	MA - lag 2,3
10,1,6	439.733	0.479	AR- lag 1- 10 MA - lag 1,2,4,5 6

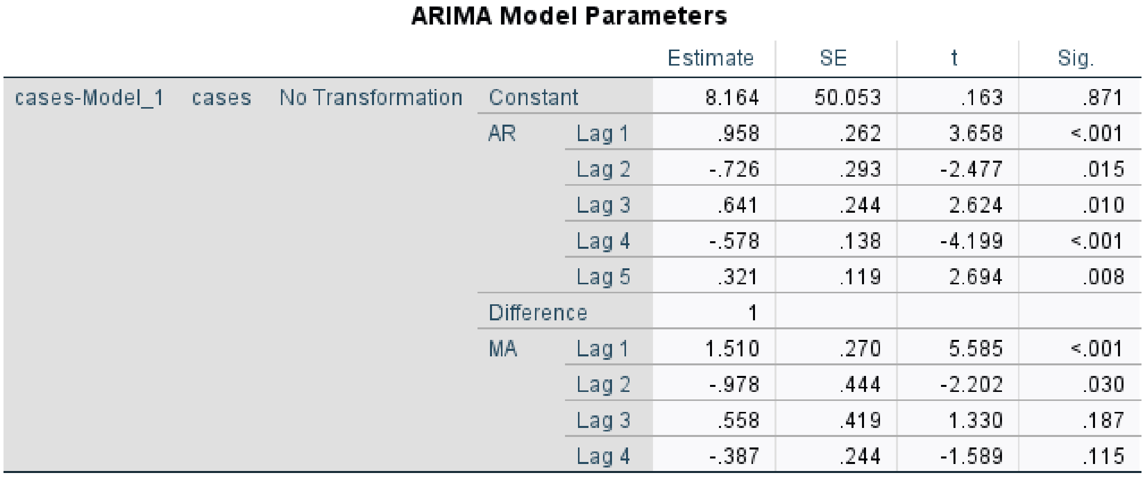
We can conclude that ARIMA 5,1,4 is our final model after comparing it to other models as it has the lowest MAE (Mean Absolute Error) and the Ljungbox test suggests that it is an adequate model because it is more than 5% Significance.



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	.412	680.883	432.275	3.191	9	.956	<.001

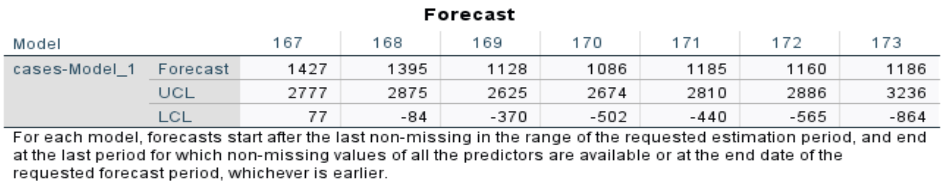
In our final Model parameters, we can see that Lags 2, 3, and 5 are positive in the AR section, suggesting a direct relationship with the current value, while lags 1 and 4 are negative, indicating an inverse relationship. Lags 2, 3, and 5 are statistically significant (p-values less than .05), which means they are likely to provide meaningful information in predicting the future values of the series. Also, lags 1 & 2 are significant and lags 3 & 4 are insignificant as the values is less than .05. We tried eliminating the lags from MA section to get a more accurate model. However, it resulted in an increase of MAE & Significance level more than 0.05.



ARIMA Model Parameters

					Estimate	SE	t	Sig.
cases-Model_1	cases	No Transformation	Constant		8.164	50.053	.163	.871
			AR	Lag 1	.958	.262	3.658	<.001
				Lag 2	-.726	.293	-2.477	.015
				Lag 3	.641	.244	2.624	.010
				Lag 4	-.578	.138	-4.199	<.001
				Lag 5	.321	.119	2.694	.008
			Difference		1			
			MA	Lag 1	1.510	.270	5.585	<.001
				Lag 2	-.978	.444	-2.202	.030
				Lag 3	.558	.419	1.330	.187
				Lag 4	-.387	.244	-1.589	.115

Below are the forecasted values which we had to predict for the next seven days, 15–21 June 2020. We can observe that the cases gradually decrease after every passing day and a sudden increase on 21 June.

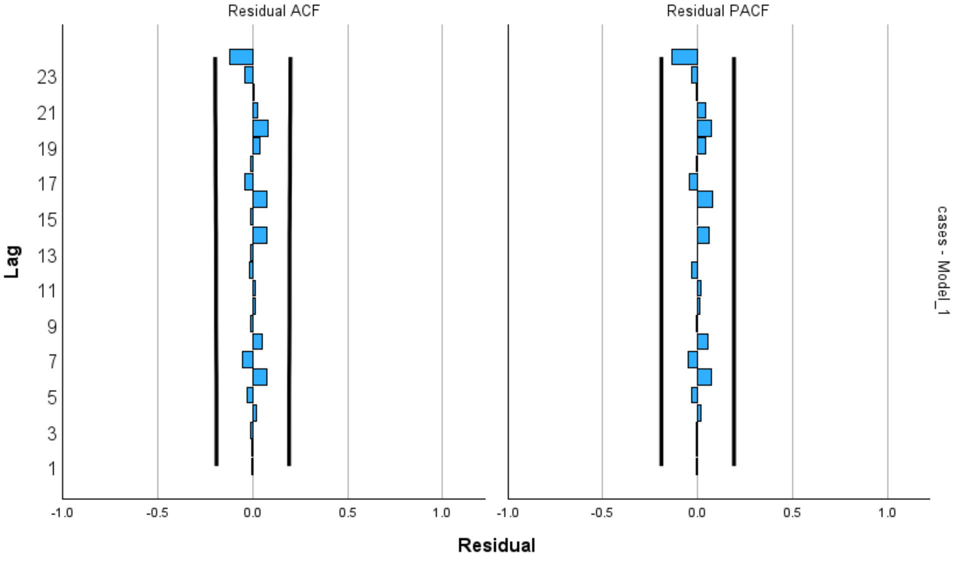


Forecast

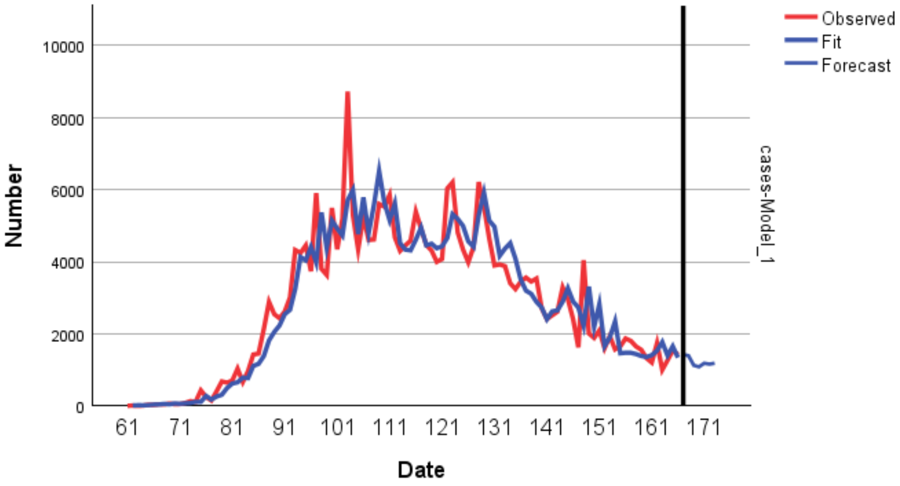
Model		167	168	169	170	171	172	173
cases-Model_1	Forecast	1427	1395	1128	1086	1185	1160	1186
	UCL	2777	2875	2625	2674	2810	2886	3236
	LCL	77	-84	-370	-502	-440	-565	-864

For each model, forecasts start after the last non-missing in the range of the requested estimation period, and end at the last period for which non-missing values of all the predictors are available or at the end date of the requested forecast period, whichever is earlier.

Lastly, we can observe that every lag in the ACF and PACF residual plot is within its significance intervals, showing that this model is the most parsimonious and accurately predicts the behavior of time series.



The actual observed values over time are represented by the red line. The model's fit to the historical data is depicted by the blue line, which demonstrates how well the seasonality and underlying trend of the observed values were represented by the ARIMA model. The boundary between the historical data and the prediction period is most likely indicated by the dark vertical line. The light blue line, which extends into the future, displays the predicted values beyond this line. The prediction seems to be maintaining the declining trend that the later portion of the historical data showed.



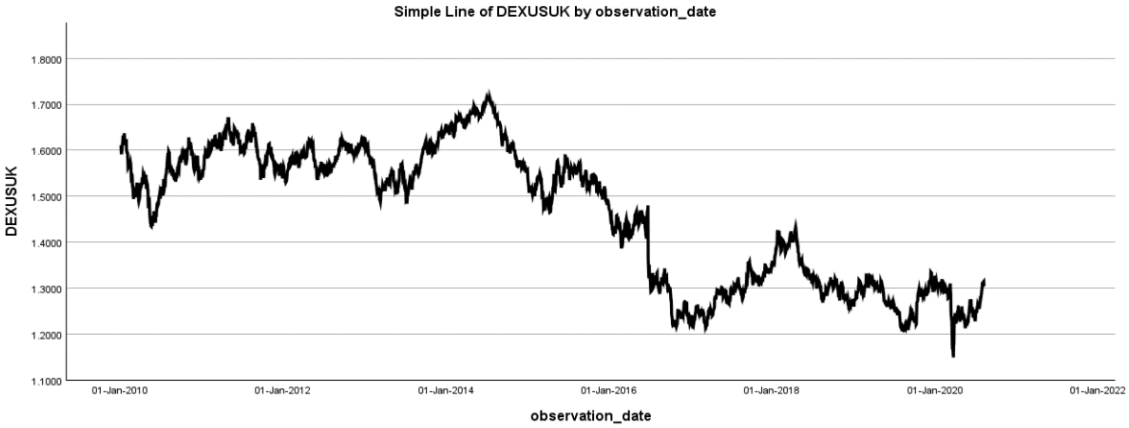
Portfolio task 3

Forecasting time series using ANNs

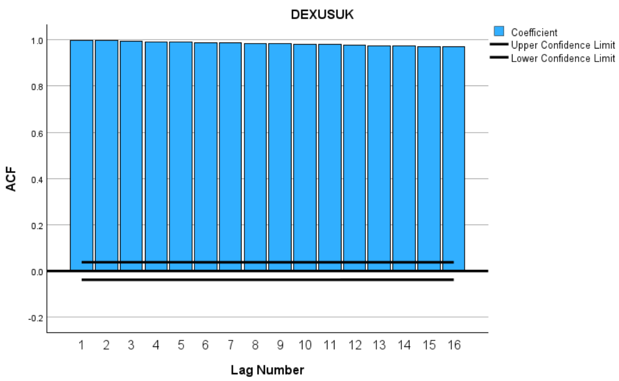
INTRODUCTION

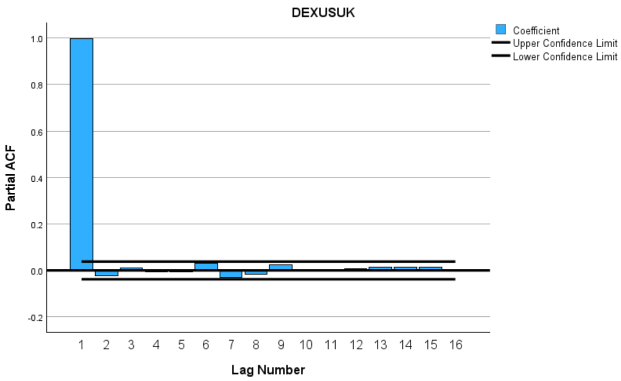
This analysis delves into the exchange rate variations between the US dollar and the British pound sterling, using data from January 4, 2010, to August 7, 2020, obtained from the Federal Reserve Economic Data (FRED) website, curated by the US Federal Reserve System's Board of Governors. Our focus is on a period marked by a notable downtrend in the pound's value against the dollar, starting from July 2014, despite periodic fluctuations. Through our study, we seek to identify the factors contributing to these exchange rate movements, understand the role of global economic and political events in these dynamics, and discuss their wider implications on international economics and trade relations.

There were many NA values in the Dataset which were cleaned using the average of the upper value and lower value. Then we plotted a line graph shown below using the Chart Builder function in SPSS which showed us that there is a trend in Data from Jan 2010 to Jan 2020.

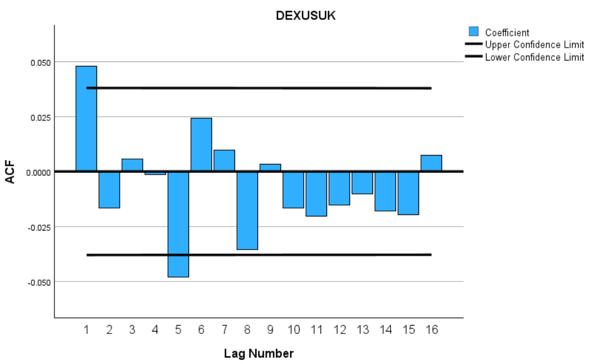


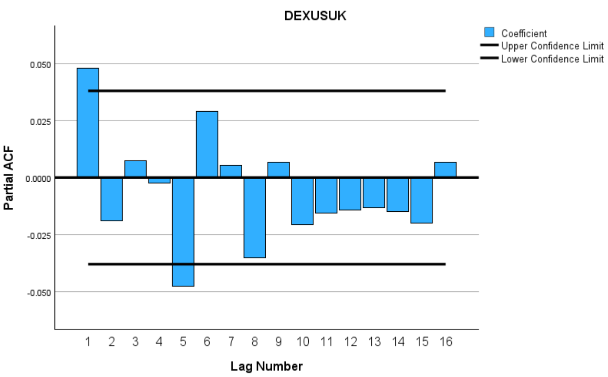
We selected the forecasts (output) and predictors (inputs) using the auto correlation method. After running the Autocorrelation, we found that the lags are insignificant and the ACF & PACF plots depicted that the data is non-stationary. Hence, we had to do the differencing and run the Autocorrelation plots again.





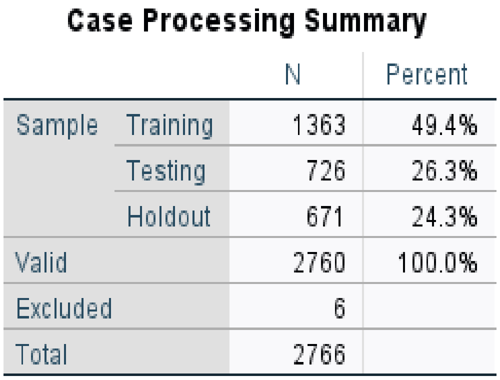
After Differencing





We found that there are 5 lags after differencing looking at the plots of the partial auto-correlation function (PACF) and autocorrelation function (ACF), which helped us choose our inputs and output. These two plots helped us choose 5 inputs to represent the time series' lag values and 1 output to represent the time series' current value. Put otherwise, the exchange rates at y(t) and y(t-1), y(t-2), y(t-3), y(t-4), y(t-5) are inputs and the exchange rate at y(t) is an output.

Analysis



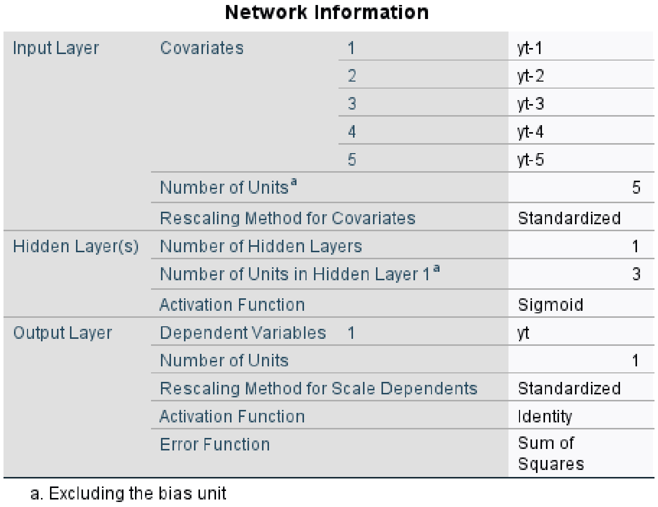
Case Processing Summary

		N	Percent
Sample	Training	1363	49.4%
	Testing	726	26.3%
	Holdout	671	24.3%
Valid		2760	100.0%
Excluded		6	
Total		2766	

Since the multilayer perceptron (MLP) neural network model is an effective tool for financial forecasting, we employed it in our investigation to estimate the exchange rate. We first designated the output as the dependent variable (Yt) and the inputs as covariates y(t-1), y(t-2), y(t-3), y(t-4), y(t-5) after choosing the inputs and output. Later, the Partition Dataset method was used to divide the current dataset into training, testing, and holdout samples. The allocated percentage, which is 50% of the sample of the entire dataset for training, 25% for testing the neural network, and 25% for holdout, is visible when we look at the case processing summary.

Architecture

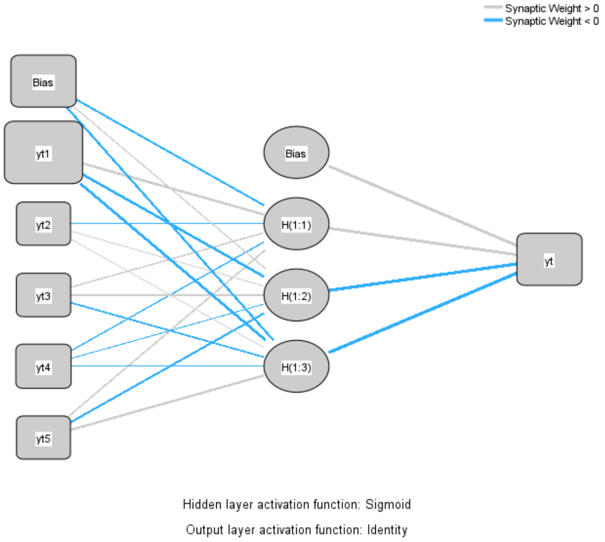
We have used the multilayer perceptron that includes an input layer, an output layer, and one or more hidden layers that have hidden neurons with sigmoid activation functions. For this time series study, we employed a custom architecture where we used single hidden layer perceptron, the program assisted in choosing the optimal hidden neurons and the number of units was set to Automatically compute. Since the Sigmoid connects the weighted sums of units in one layer to the values of units in the subsequent layer, we utilized it for the input layer and the identity activation function for the output layer. Batch training was used since it determines the network that assists in handling the records by utilizing all the data from the training dataset. Using the Network Information table, we provided a description of the neural network model for this time series. The network information diagram shows that, as anticipated, there are 5 units in the input layer, 3 units in the hidden layer (Automatically Compute), and 1 unit in the output layer.



Network Information

Input Layer	Covariates	1	yt-1
		2	yt-2
		3	yt-3
		4	yt-4
		5	yt-5
	Number of Units[a]		5
	Rescaling Method for Covariates		Standardized
Hidden Layer(s)	Number of Hidden Layers		1
	Number of Units in Hidden Layer 1[a]		3
	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

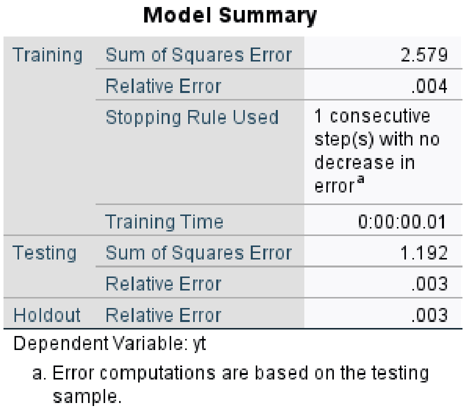


Hidden layer activation function: Sigmoid

Output layer activation function: Identity

A basic neural network architecture consisting of an input layer, one hidden layer, and an output layer is depicted in the plot. The input layer feeds the hidden layer with five nodes (yt1 to yt5) representing input variables and one bias node. Three neurons (H(1:1) to H(1:3)) make up the hidden layer. Each neuron receives input from every input node in addition to an extra bias. The hidden layer is given a sigmoid activation function, which gives the model non-linearity.

Model Summary & Parameters

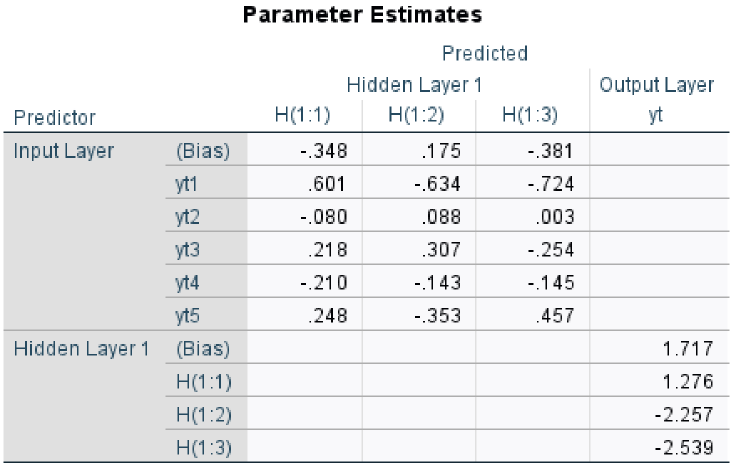


Model Summary

Training	Sum of Squares Error	2.579
	Relative Error	.004
	Stopping Rule Used	1 consecutive step(s) with no decrease in error[a]
	Training Time	0:00:00.01
Testing	Sum of Squares Error	1.192
	Relative Error	.003
Holdout	Relative Error	.003

Dependent Variable: yt

a. Error computations are based on the testing sample.



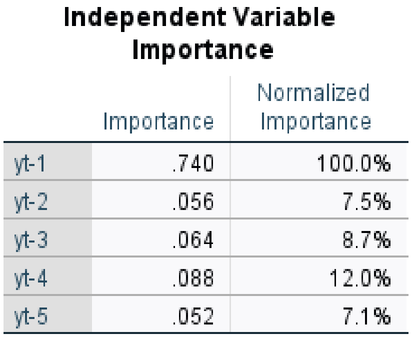
Parameter Estimates

Predictor		Predicted			
		Hidden Layer 1			Output Layer
		H(1:1)	H(1:2)	H(1:3)	yt
Input Layer	(Bias)	-.348	.175	-.381	
	yt1	.601	-.634	-.724	
	yt2	-.080	.088	.003	
	yt3	.218	.307	-.254	
	yt4	-.210	-.143	-.145	
	yt5	.248	-.353	.457	
Hidden Layer 1	(Bias)				1.717
	H(1:1)				1.276
	H(1:2)				-2.257
	H(1:3)				-2.539

The specifics of our neural network are explained in the model summary. Our Initial model performs well because its relative error is approximately 0.4% in the training, testing, and holdout phases, and 0.3% in the holdout phase. Smaller error values are indicative of a better model. In particular, the holdout phase exhibits superior fitting and forecasting, which is why NNAR is a very useful tool for exchange rate forecasting.

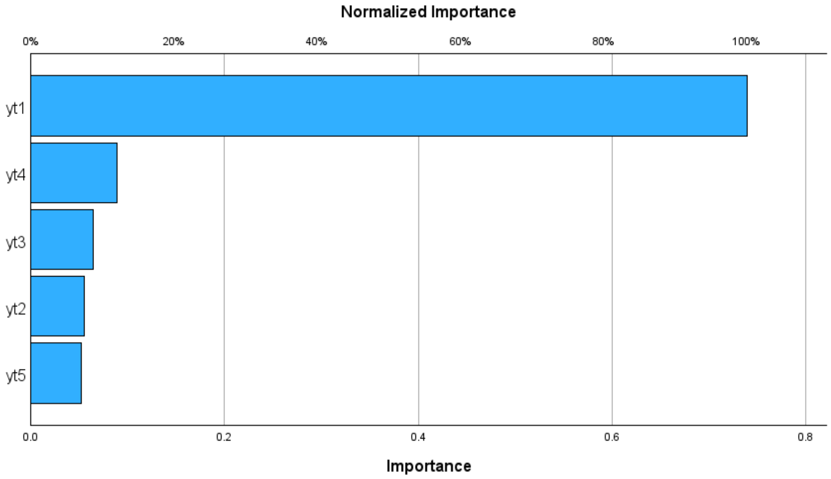
Independent Variable Information

The sensitivity analysis was done to find out how important each predictor was in determining the neural network, and the results are shown in the table and chart below. Yt-1 is the most significant predictor, according to the output. Yt-4 also makes a significant contribution to the forecasting process followed by Yt3 & Yt2, and YT-5 is the least significant input.

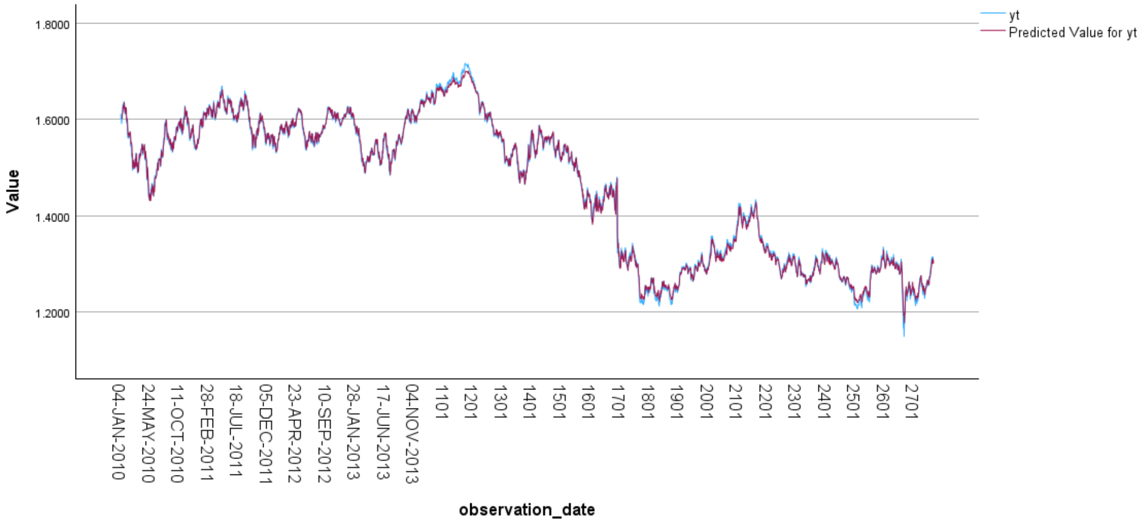


Independent Variable Importance

	Importance	Normalized Importance
yt-1	.740	100.0%
yt-2	.056	7.5%
yt-3	.064	8.7%
yt-4	.088	12.0%
yt-5	.052	7.1%



Graph



The graph compares the actual values of 'yt' against model predictions over time. The data spans from January 2010 to August 2020, with notable variations possibly reflecting impactful events during this period.

Final Forecast for August 8, 2020

Hence to report the one-step-ahead forecast, the exchange rate for August 8, 2020. We used all the above analysis which led to our final forecast which is **$1.3009.**

Portfolio task 4

Logistic Regression Analysis

Part A

Aim

To predict the value of a single customer's shopping basket, our primary goal in this report is to analyze data from Fresco Supermarket and find trends and patterns in a sample of weekly data collected for several of their loyalty cardholders over a 26-week period. We will also make any necessary adjustments. There are 75 samples in all. First, all independent factors influencing our dependent variable were looked at, along with the dependent and independent variables.

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

The shopping basket was divided into two groups: low & medium. This served as our dependent variable, allowing us to determine the potential customer's spending patterns. After that, we estimated our model and proceeded with our analysis using the parsimonious model. We performed a multicollinearity test, residual analysis, observed standardized residuals, computed the cook distance and Deta, and observed the results to assess the suitability of our final model. Lastly, we examined the pseudo r square, Hosmer and Lemeshow's tests, and classification accuracy to determine the goodness of fit. We discovered that each category's accuracy was comparable. After passing the adequacy criteria, we found a parsimonious model that can be applied to classification and prediction.

Methods

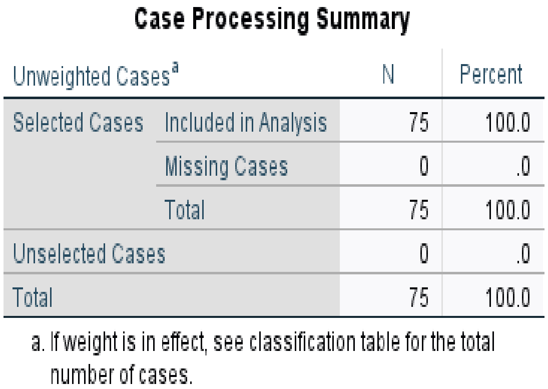
Binary logistic regression was employed in the analysis because the dataset has a categorical dependent variable with 2 levels. This approach assisted us in analyzing if the potential customer's spending is influenced by variables like their age, gender, frequency of shopping per week, and shopping basket price.

Results

Low Spender (1)	High Spender (0)
Shopping basket value < £50	Shopping basket Value > £50
95.3 % correctly predicted	93.8 % correctly predicted

Part B

Binary logistic regression will be used in this report since our dependent variable contains two possible outcomes such as Low Spender & High Spender. The customers' gender, age, frequency of shopping per week, and shopping basket price are our independent variables. We looked at the highest frequency to determine our reference category; from the table below, we found that the medium spender had the highest frequency; as a result, we have used that as our reference category in our analysis.



Case Processing Summary

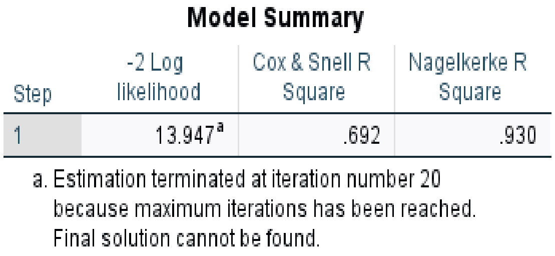
Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	75	100.0
	Missing Cases	0	.0
	Total	75	100.0
Unselected Cases		0	.0
Total		75	100.0

a. If weight is in effect, see classification table for the total number of cases.

Logistic Regression 1st Model

The table shows that there are 75 examples in the dataset overall, and every single one of them gets analyzed (100%). No cases have been left out of the analysis or excluded, and there are no missing cases. When using processes such as logistic regression, the output usually includes this summary table, which indicates the number of cases being analyzed and any missing data.

Model Estimation



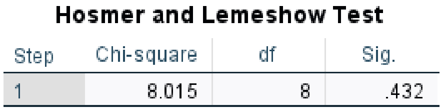
Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	13.947[a]	.692	.930

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

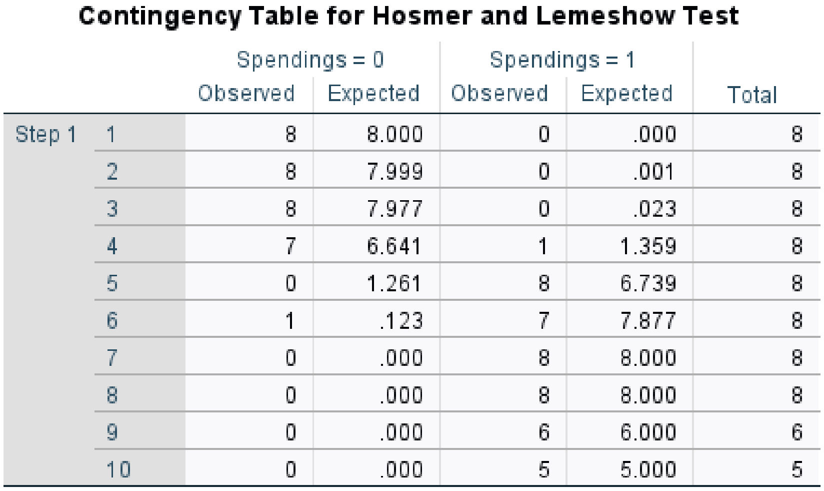
The Model summary has a Nagelkerke R Square of 0.930, a Cox & Snell R Square value of 0.692, and a -2 Log probability of 13.947. The footnote states that because the maximum number of iterations was reached without a definitive solution for the model being found, the estimation was stopped at iteration number 20. This implies that there might be a problem with the convergence of the model.

Hosmer and Lemeshow Test for 1st Model



Hosmer and Lemeshow Test

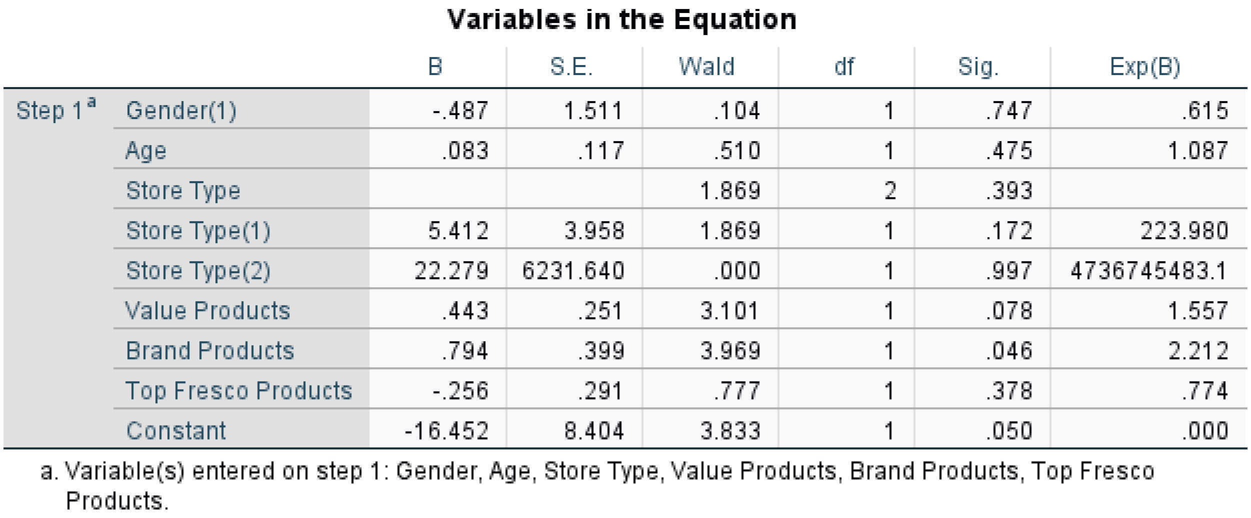
Step	Chi-square	df	Sig.
1	8.015	8	.432



Contingency Table for Hosmer and Lemeshow Test

		Spendings = 0		Spendings = 1		
		Observed	Expected	Observed	Expected	Total
Step 1	1	8	8.000	0	.000	8
	2	8	7.999	0	.001	8
	3	8	7.977	0	.023	8
	4	7	6.641	1	1.359	8
	5	0	1.261	8	6.739	8
	6	1	.123	7	7.877	8
	7	0	.000	8	8.000	8
	8	0	.000	8	8.000	8
	9	0	.000	6	6.000	6
	10	0	.000	5	5.000	5

A goodness-of-fit test for logistic regression models is the Hosmer and Lemeshow Test. In this instance, the test produces a significance (Sig.) of.432 and a Chi-square value of 8.015 with 8 degrees of freedom. With a high p-value, the model's predictions and actual values do not deviate significantly from one another. The observed and predicted counts of instances for low spenders (Spending = 0) and high spenders (Spending = 1) across 10 groups are shown in the supporting Contingency Table for the Hosmer and Lemeshow Test.



Variables in the Equation

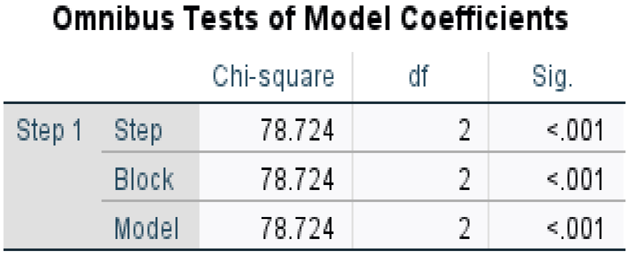
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Gender(1)	-.487	1.511	.104	1	.747	.615
	Age	.083	.117	.510	1	.475	1.087
	Store Type			1.869	2	.393	
	Store Type(1)	5.412	3.958	1.869	1	.172	223.980
	Store Type(2)	22.279	6231.640	.000	1	.997	4736745483.1
	Value Products	.443	.251	3.101	1	.078	1.557
	Brand Products	.794	.399	3.969	1	.046	2.212
	Top Fresco Products	-.256	.291	.777	1	.378	.774
	Constant	-16.452	8.404	3.833	1	.050	.000

a. Variable(s) entered on step 1: Gender, Age, Store Type, Value Products, Brand Products, Top Fresco Products.

In the First Model we can see that most of the variables are above the significance level 0.05%. Hence, we need to eliminate them one by one to get a final model with all the significant variables.

<u>Final Model</u>

Model Estimation

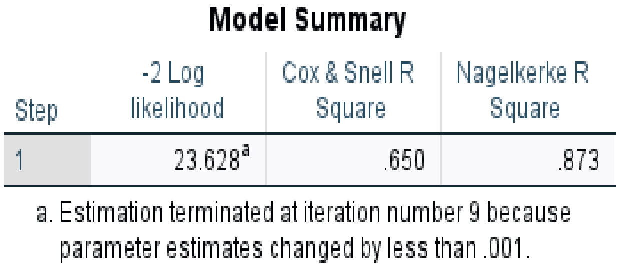


Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	78.724	2	<.001
	Block	78.724	2	<.001
	Model	78.724	2	<.001

The model coefficients are all statistically significant, according to the Omnibus Tests of Model Coefficients table. We may infer that the collection of independent variables included in the model at Step 1 meaningfully contributes to the prediction of the dependent variable with a Chi-square statistic of 78.724 at 2 degrees of freedom and a significance level of less than 0.001. This demonstrates that the model is useful in differentiating between the two spending categories in the Fresco Supermarket sample data (high versus low spenders).

The logistic regression model's fit to the data is indicated by the table, which displays the model's -2 Log likelihood of 23.628 at Step 1. A significant correlation between the predictors and the binary outcome is suggested by the Cox & Snell R Square value of.650 and the Nagelkerke R Square value of.873.

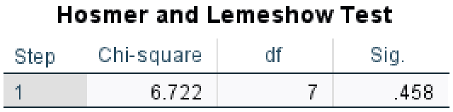


Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	23.628[a]	.650	.873

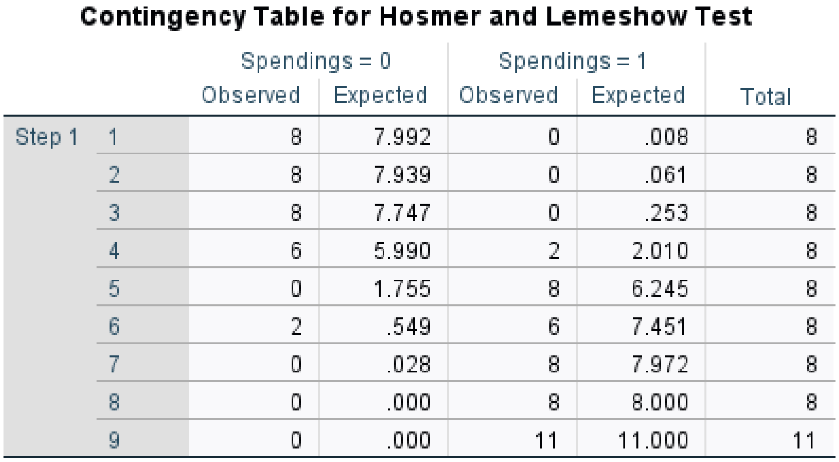
a. Estimation terminated at iteration number 9 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test for Final Model



Hosmer and Lemeshow Test

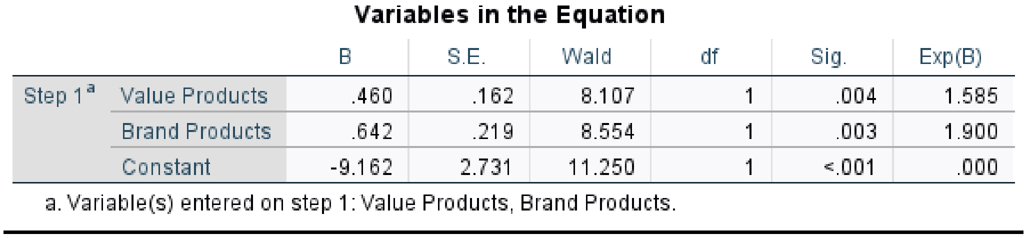
Step	Chi-square	df	Sig.
1	6.722	7	.458



Contingency Table for Hosmer and Lemeshow Test

		Spendings = 0		Spendings = 1		
		Observed	Expected	Observed	Expected	Total
Step 1	1	8	7.992	0	.008	8
	2	8	7.939	0	.061	8
	3	8	7.747	0	.253	8
	4	6	5.990	2	2.010	8
	5	0	1.755	8	6.245	8
	6	2	.549	6	7.451	8
	7	0	.028	8	7.972	8
	8	0	.000	8	8.000	8
	9	0	.000	11	11.000	11

A Chi-square value of 6.722 with 7 degrees of freedom and a significance level of.458 is indicated by the Hosmer and Lemeshow Test table and the related Contingency Table. Given that a high p-value denotes no significant difference between the observed and model-predicted values, this result reflects a strong match between the model and the observed data. The distribution of observed against expected frequencies among groups is further detailed in the Hosmer and Lemeshow Test Contingency Table, which supports the appropriateness of the model by demonstrating near alignment.



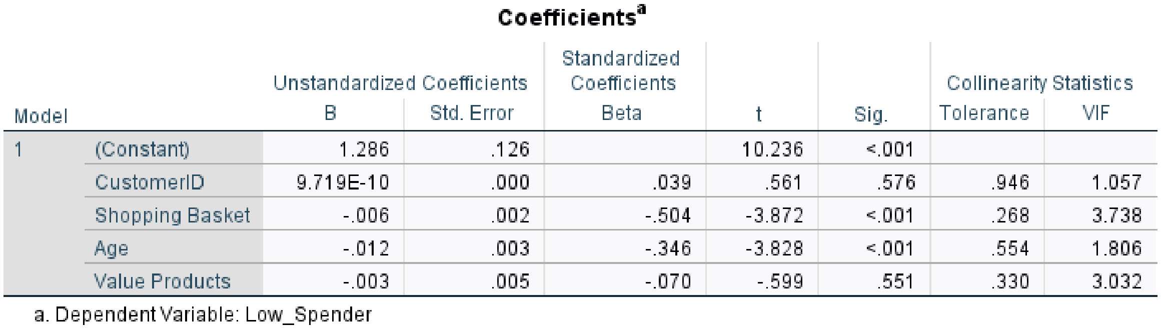
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Value Products	.460	.162	8.107	1	.004	1.585
	Brand Products	.642	.219	8.554	1	.003	1.900
	Constant	-9.162	2.731	11.250	1	<.001	.000

a. Variable(s) entered on step 1: Value Products, Brand Products.

After elimination of all the nonsignificant variables from the first model we are left with Value Products & Brand Products which are significant with the values less than 0.05% significance level showing that our final model is adequate

Regression Analysis

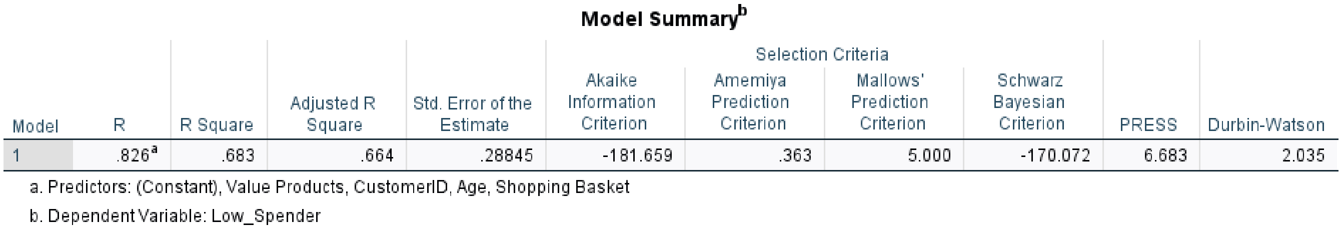


Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	1.286	.126		10.236	<.001		
	CustomerID	9.719E-10	.000	.039	.561	.576	.946	1.057
	Shopping Basket	-.006	.002	-.504	-3.872	<.001	.268	3.738
	Age	-.012	.003	-.346	-3.828	<.001	.554	1.806
	Value Products	-.003	.005	-.070	-.599	.551	.330	3.032

a. Dependent Variable: Low_Spender

We can observe that VIF is <10 and Tolerance is >0.1 by looking at the collinearity statistics which shows that independent variables have no multicollinearity.



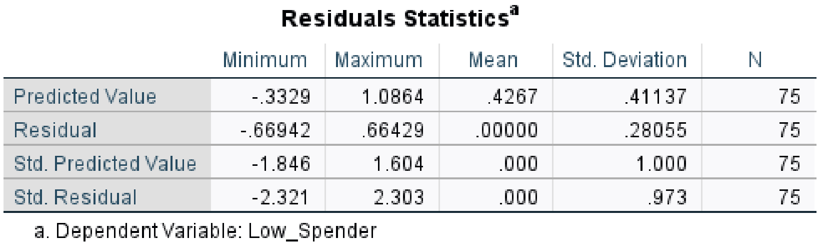
Model Summary[b]

					Selection Criteria					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Akaike Information Criterion	Amemiya Prediction Criterion	Mallows' Prediction Criterion	Schwarz Bayesian Criterion	PRESS	Durbin-Watson
1	.826[a]	.683	.664	.28845	-181.659	.363	5.000	-170.072	6.683	2.035

a. Predictors: (Constant), Value Products, CustomerID, Age, Shopping Basket

b. Dependent Variable: Low_Spender

Based on the Model Summary table, the regression model's R Square value is.683, meaning it can account for roughly 68.3% of the variation in the dependent variable 'Low Spender'. A more precise indicator of model fit is the Adjusted R Square of.664, which accounts for the number of predictors in the model. The model's residuals appear to lack significant autocorrelation, as indicated by the Durbin-Watson value of 2.035.



Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-.3329	1.0864	.4267	.41137	75
Residual	-.66942	.66429	.00000	.28055	75
Std. Predicted Value	-1.846	1.604	.000	1.000	75
Std. Residual	-2.321	2.303	.000	.973	75

a. Dependent Variable: Low_Spender

Portfolio Task 5

Cluster analysis.

Aim

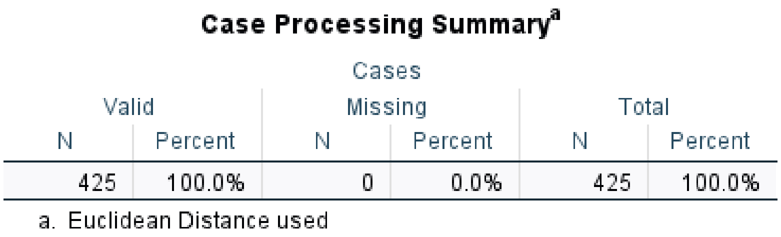
This study, initiated by a prominent UK bank, aims to utilize a detailed dataset covering aspects like gender, age, financial balances, marital status, homeownership, employment status, customer tenure, employment history, and credit risk. The objective is to analyze and segment the customers into distinct groups based on these varied data points. Through such segmentation, we can uncover key trends and enable the bank's product development team to craft specialized financial products that cater to the unique needs of each segment. This process will involve thorough data preparation, exploration, and the use of sophisticated analytics to create actionable and strategic segments for product innovation and targeted marketing.

Methods

Given that market segmentation and cluster analysis have the same objectives, we opted to perform the market segmentation in this study using Hierarchical Cluster Analysis. Since the goal is to obtain data for a group of bank customers, the customers in that group should have characteristics that are like one another but different from those of customers in other groups. The clustering approach can be used to do this based on the individuals' similarity and distance from one another. Ultimately, each bank client will be grouped into a cluster. An illustration of this would be a dendrogram.

Characteristics

We started the study by considering all the continuous and categorical factors, such as credit risk, age, gender, marital status, housing, Job, current account, savings account, months customer, and months employed. To use them in our first clustering, we turned all these variables' categorical values into numerical values. We assigned distinct client IDs to each case to improve our comprehension, which helped us analyze the closeness matrix and dendrogram. In this instance, the distance was measured using the Euclidean distance and the Within Linkage. Additionally, we normalized the data between 0 and 1, which is always better because it prevents the measurement unit from influencing our findings. The outcome of the first clustering method is shown below.

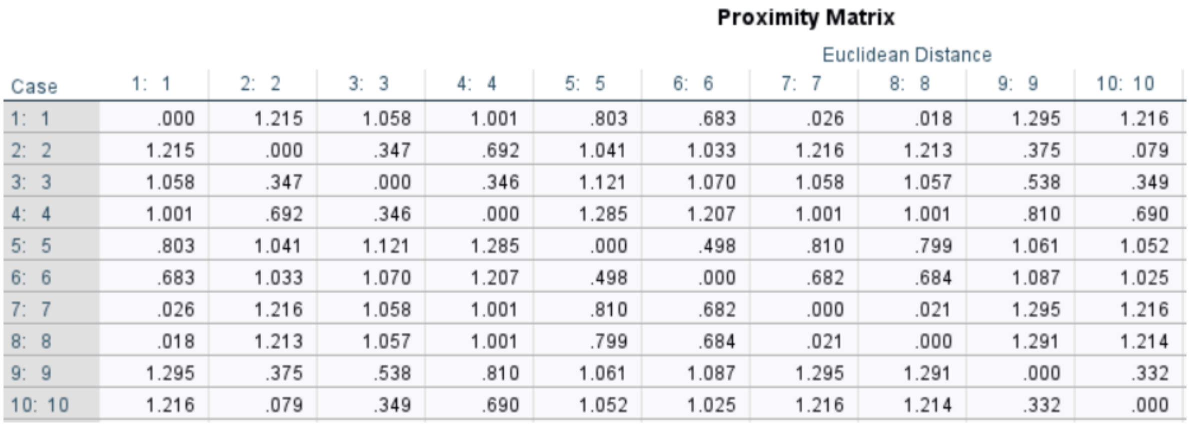


Case Processing Summary[a]

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

a. Euclidean Distance used

We can observe that we have a total of 425 cases with 0 number of missing cases.



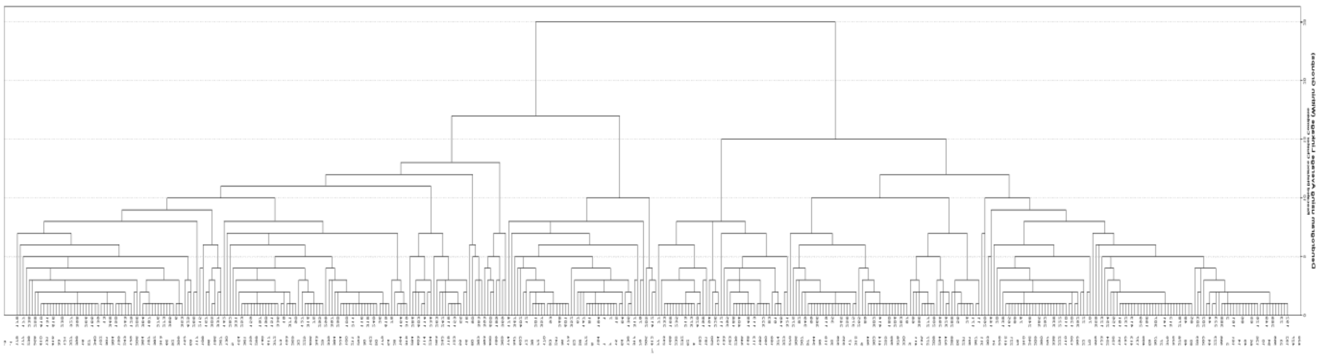
Proximity Matrix

	Euclidean Distance									
Case	1: 1	2: 2	3: 3	4: 4	5: 5	6: 6	7: 7	8: 8	9: 9	10: 10
1: 1	.000	1.215	1.058	1.001	.803	.683	.026	.018	1.295	1.216
2: 2	1.215	.000	.347	.692	1.041	1.033	1.216	1.213	.375	.079
3: 3	1.058	.347	.000	.346	1.121	1.070	1.058	1.057	.538	.349
4: 4	1.001	.692	.346	.000	1.285	1.207	1.001	1.001	.810	.690
5: 5	.803	1.041	1.121	1.285	.000	.498	.810	.799	1.061	1.052
6: 6	.683	1.033	1.070	1.207	.498	.000	.682	.684	1.087	1.025
7: 7	.026	1.216	1.058	1.001	.810	.682	.000	.021	1.295	1.216
8: 8	.018	1.213	1.057	1.001	.799	.684	.021	.000	1.291	1.214
9: 9	1.295	.375	.538	.810	1.061	1.087	1.295	1.291	.000	.332
10: 10	1.216	.079	.349	.690	1.052	1.025	1.216	1.214	.332	.000

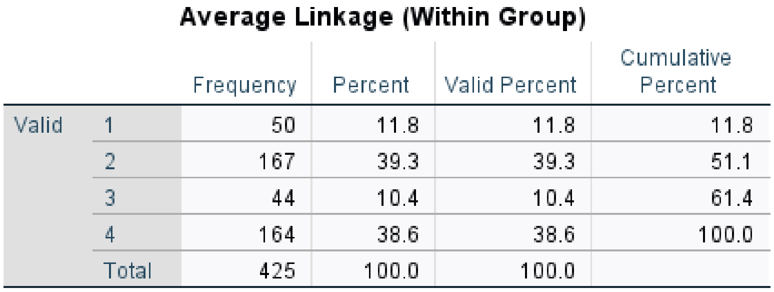
The Proximity matrix represents the Euclidean distances between different cases. Euclidean distance is a measure of the true straight-line distance between two points in Euclidean space. Cases 1 and 2 are separated by 1.215, indicating a considerable degree of dissimilarity. Between Cases 1 and 8, the smallest non-zero distance in the matrix is.018, suggesting that these two are the most similar of all the pairs.

Dendrogram Analysis

Within Group Linkage Method



Using the dendrogram below as a guide, we looked at several clusters. Then, using our analysis and our imaginary line as a guide, we decided that, given the distribution, 4 clusters would be a suitable option.

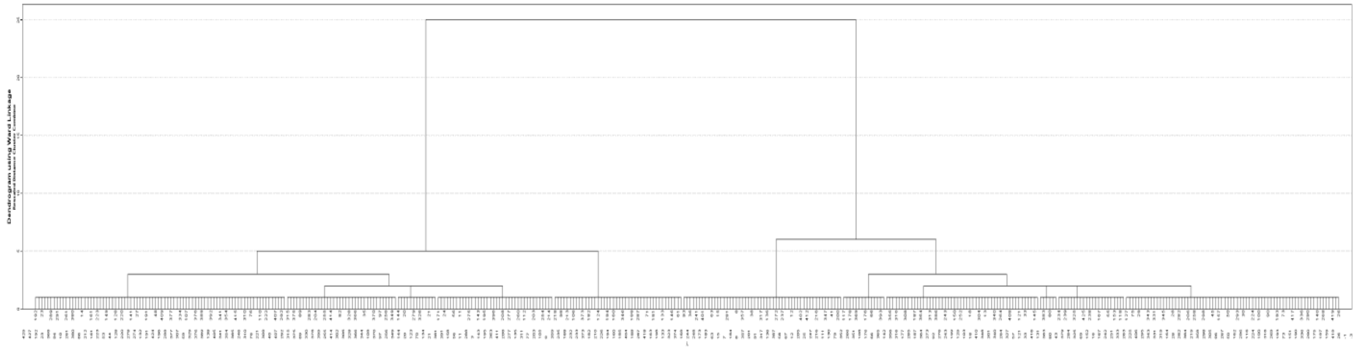


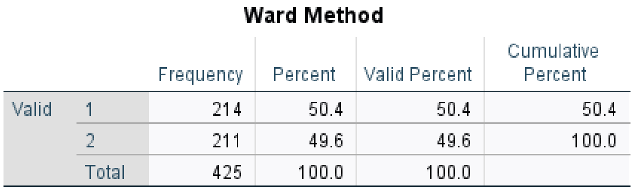
Average Linkage (Within Group)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	50	11.8	11.8	11.8
	2	167	39.3	39.3	51.1
	3	44	10.4	10.4	61.4
	4	164	38.6	38.6	100.0
	Total	425	100.0	100.0	

According to the average Linkage within group method, our clusters were divided into 4. However, we can't see any equal distribution among them in this method.

Ward Method





Ward Method

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	214	50.4	50.4	50.4
	2	211	49.6	49.6	100.0
	Total	425	100.0	100.0	

We also used Ward method and changed the number of clusters to 2 as it looks like a good Method since it is equally distributing the cases into 2 clusters.

Result

Following a review of the frequency tables for both approaches, we determined that the Ward Method should have 2 total clusters because the percentage of values in each cluster is evenly distributed. This will help the bank's product development team create financial products and promotions that are tailored to the needs of the target market.

Portfolio Task 6

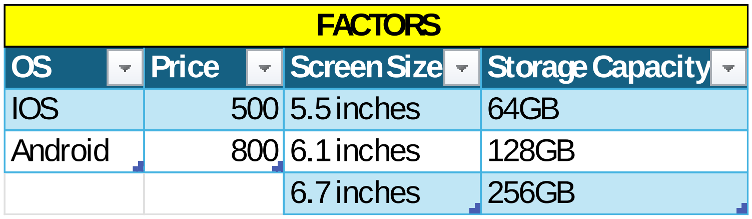
Conjoint analysis study

INTRODUCTION

The purpose of this study is to identify the features that consumers value most when making a mobile phone purchase, as well as their preferences for buying new phones. This data helps us with the launch of a new mobile phone by giving us insight into what consumers anticipate. We employed conjoint analysis, a multivariate technique designed to assess preferences for any kind of product, in this investigation. Another common application of conjoint analysis is in the design of new or expanded products. In our instance, this approach was utilized to find out what the preferences of the public were for the introduction of a new cell phone.

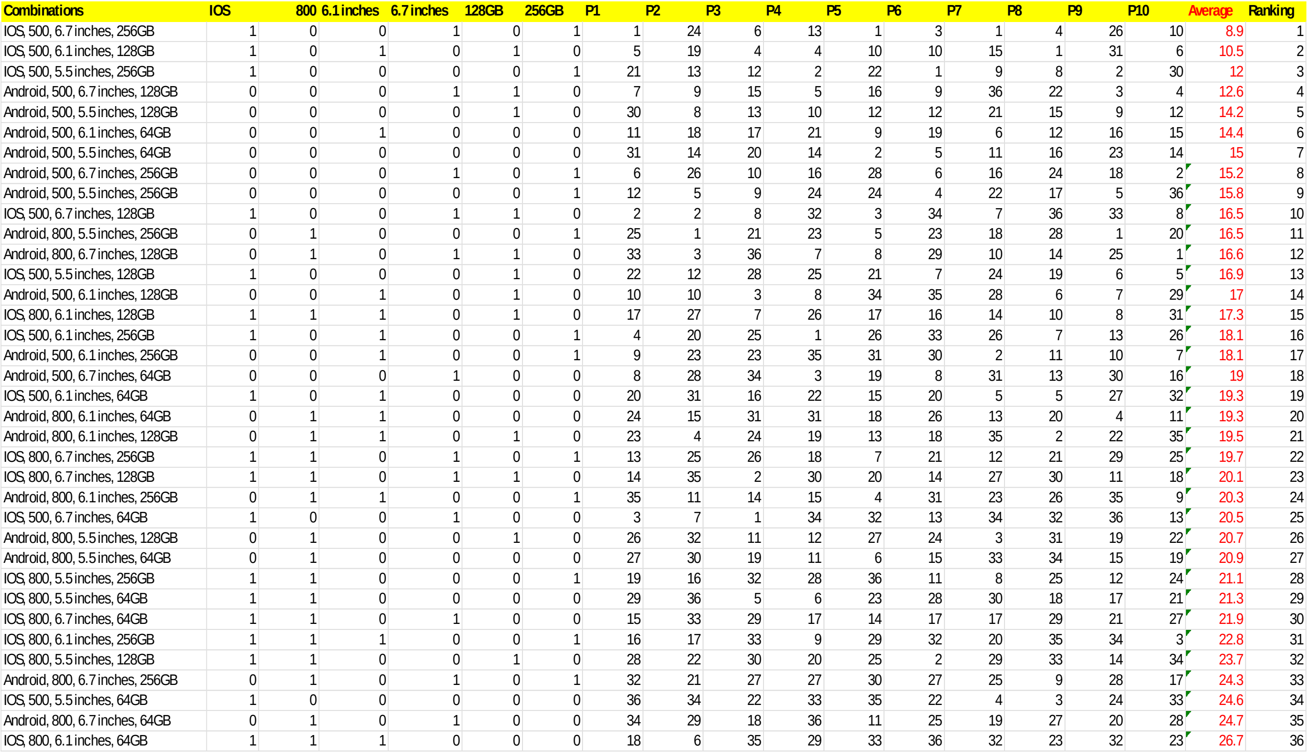
FACTORS:

Table 1 displays four components, each with a level of 3, 3, 2, 2, and so on. Because consumers look at these aspects before making a purchase, they are considered relevant. Next, we used to combine the product into 36 distinct (3x3x2x2) combinations. 10 friends & Family ranked these product combinations according to their preferences; the results were then combined with 1 being the best and 36 being the worst, the rankings were provided on a scale of 1 to 36.



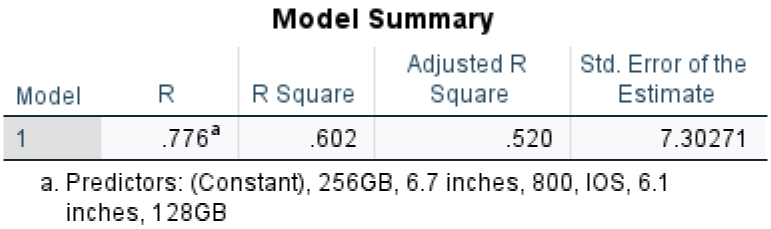
FACTORS			
OS	Price	Screen Size	Storage Capacity
IOS	500	5.5 inches	64GB
Android	800	6.1 inches	128GB
		6.7 inches	256GB

We created dummy variables in Excel using Power query Editor and combined the factors into 36 different combinations. Each combination was ranked from 1-36 by 10 friends/family according to their preferences. further analyses were done using SPSS.



Combinations	IOS	800	6.1 inches	6.7 inches	128GB	256GB	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Average	Ranking
IOS, 500, 6.7 inches, 256GB	1	0	0	1	0	1	1	24	6	13	1	3	1	4	26	10	8.9	1
IOS, 500, 6.1 inches, 128GB	1	0	1	0	1	0	5	19	4	4	10	10	15	1	31	6	10.5	2
IOS, 500, 5.5 inches, 256GB	1	0	0	0	0	1	21	13	12	2	22	1	9	8	2	30	12	3
Android, 500, 6.7 inches, 128GB	0	0	0	1	1	0	7	9	15	5	16	9	36	22	3	4	12.6	4
Android, 500, 5.5 inches, 128GB	0	0	0	0	1	0	30	8	13	10	12	12	21	15	9	12	14.2	5
Android, 500, 6.1 inches, 64GB	0	0	1	0	0	0	11	18	17	21	9	19	6	12	16	15	14.4	6
Android, 500, 5.5 inches, 64GB	0	0	0	0	0	0	31	14	20	14	2	5	11	16	23	14	15	7
Android, 500, 6.7 inches, 256GB	0	0	0	1	0	1	6	26	10	16	28	6	16	24	18	2	15.2	8
Android, 500, 5.5 inches, 256GB	0	0	0	0	0	1	12	5	9	24	24	4	22	17	5	36	15.8	9
IOS, 500, 6.7 inches, 128GB	1	0	0	1	1	0	2	2	8	32	3	34	7	36	33	8	16.5	10
Android, 800, 5.5 inches, 256GB	0	1	0	0	0	1	25	1	21	23	5	23	18	28	1	20	16.5	11
Android, 800, 6.7 inches, 128GB	0	1	0	1	1	0	33	3	36	7	8	29	10	14	25	1	16.6	12
IOS, 500, 5.5 inches, 128GB	1	0	0	0	1	0	22	12	28	25	21	7	24	19	6	5	16.9	13
Android, 500, 6.1 inches, 128GB	0	0	1	0	1	0	10	10	3	8	34	35	28	6	7	29	17	14
IOS, 800, 6.1 inches, 128GB	1	1	1	0	1	0	17	27	7	26	17	16	14	10	8	31	17.3	15
IOS, 500, 6.1 inches, 256GB	1	0	1	0	0	1	4	20	25	1	26	33	26	7	13	26	18.1	16
Android, 500, 6.1 inches, 256GB	0	0	1	0	0	1	9	23	23	35	31	30	2	11	10	7	18.1	17
Android, 500, 6.7 inches, 64GB	0	0	0	1	0	0	8	28	34	3	19	8	31	13	30	16	19	18
IOS, 500, 6.1 inches, 64GB	1	0	1	0	0	0	20	31	16	22	15	20	5	5	27	32	19.3	19
Android, 800, 6.1 inches, 64GB	0	1	1	0	0	0	24	15	31	31	18	26	13	20	4	11	19.3	20
Android, 800, 6.1 inches, 128GB	0	1	1	0	1	0	23	4	24	19	13	18	35	2	22	35	19.5	21
IOS, 800, 6.7 inches, 256GB	1	1	0	1	0	1	13	25	26	18	7	21	12	21	29	25	19.7	22
IOS, 800, 6.7 inches, 128GB	1	1	0	1	1	0	14	35	2	30	20	14	27	30	11	18	20.1	23
Android, 800, 6.1 inches, 256GB	0	1	1	0	0	1	35	11	14	15	4	31	23	26	35	9	20.3	24
IOS, 500, 6.7 inches, 64GB	1	0	0	1	0	0	3	7	1	34	32	13	34	32	36	13	20.5	25
Android, 800, 5.5 inches, 128GB	0	1	0	0	1	0	26	32	11	12	27	24	3	31	19	22	20.7	26
Android, 800, 5.5 inches, 64GB	0	1	0	0	0	0	27	30	19	11	6	15	33	34	15	19	20.9	27
IOS, 800, 5.5 inches, 256GB	1	1	0	0	0	1	19	16	32	28	36	11	8	25	12	24	21.1	28
IOS, 800, 5.5 inches, 64GB	1	1	0	0	0	0	29	36	5	6	23	28	30	18	17	21	21.3	29
IOS, 800, 6.7 inches, 64GB	1	1	0	1	0	0	15	33	29	17	14	17	17	29	21	27	21.9	30
IOS, 800, 6.1 inches, 256GB	1	1	1	0	0	1	16	17	33	9	29	32	20	35	34	3	22.8	31
IOS, 800, 5.5 inches, 128GB	1	1	0	0	1	0	28	22	30	20	25	2	29	33	14	34	23.7	32
Android, 800, 6.7 inches, 256GB	0	1	0	1	0	1	32	21	27	27	30	27	25	9	28	17	24.3	33
IOS, 500, 5.5 inches, 64GB	1	0	0	0	0	0	36	34	22	33	35	22	4	3	24	33	24.6	34
Android, 800, 6.7 inches, 64GB	0	1	0	1	0	0	34	29	18	36	11	25	19	27	20	28	24.7	35
IOS, 800, 6.1 inches, 64GB	1	1	1	0	0	0	18	6	35	29	33	36	32	23	32	23	26.7	36

Regression on SPSS

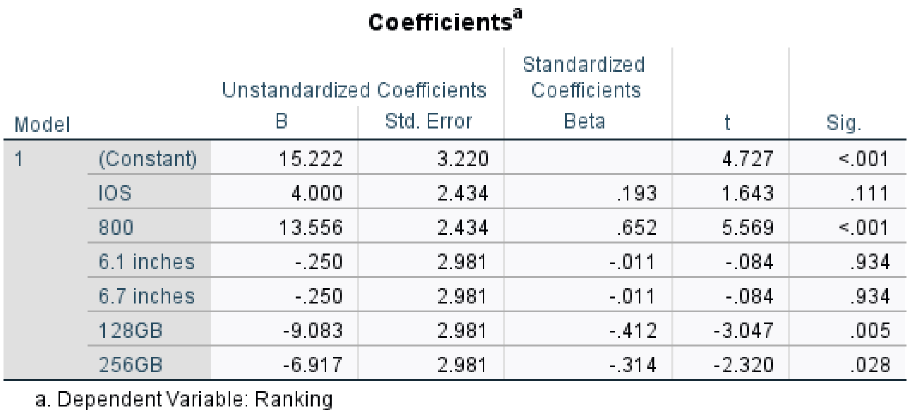


Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.776[a]	.602	.520	7.30271

a. Predictors: (Constant), 256GB, 6.7 inches, 800, IOS, 6.1 inches, 128GB

We can observe that our R Square is 0.602 which means that the dummy variables can predict the ranking accurately.



Coefficients[a]

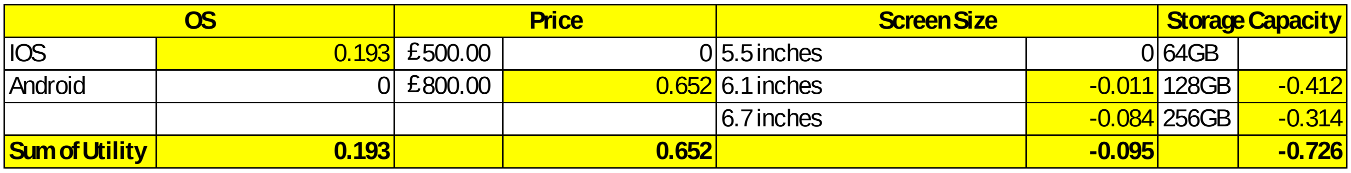
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	15.222	3.220		4.727	<.001
	IOS	4.000	2.434	.193	1.643	.111
	800	13.556	2.434	.652	5.569	<.001
	6.1 inches	-.250	2.981	-.011	-.084	.934
	6.7 inches	-.250	2.981	-.011	-.084	.934
	128GB	-9.083	2.981	-.412	-3.047	.005
	256GB	-6.917	2.981	-.314	-2.320	.028

a. Dependent Variable: Ranking

The Standardized Coefficients Beta Values indicate that more people would prefer buying an IOS over android since the value is positive (.193)

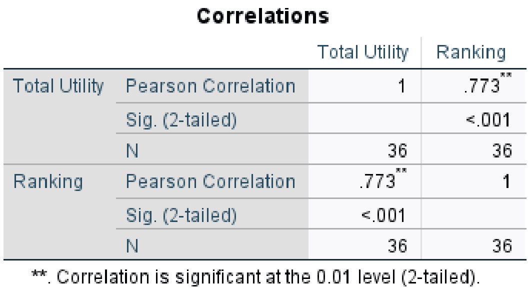
Also, the 128gb variant phones have a more negative value as compared to the 256gb variant indicating that people would like more storage in their phone for which they are willing to $800 (.652) having a positive standardized coefficients beta.

The 6.1- & 6.7-inches phones have the same negative beta values (-.011) stating that people are okay with any screen size if they get other features and desired price for their phone.



OS		Price		Screen Size		Storage Capacity	
IOS	0.193	£500.00	0	5.5 inches	0	64GB	
Android	0	£800.00	0.652	6.1 inches	-0.011	128GB	-0.412
				6.7 inches	-0.084	256GB	-0.314
Sum of Utility	**0.193**		**0.652**		**-0.095**		**-0.726**

We later calculated the utility figures obtained by taking the sum of utilities for each combination of products. This indicates to us if a mobile phone's factor offers us High utility (e.g., 0.845) or low utility (e.g., -0.496).

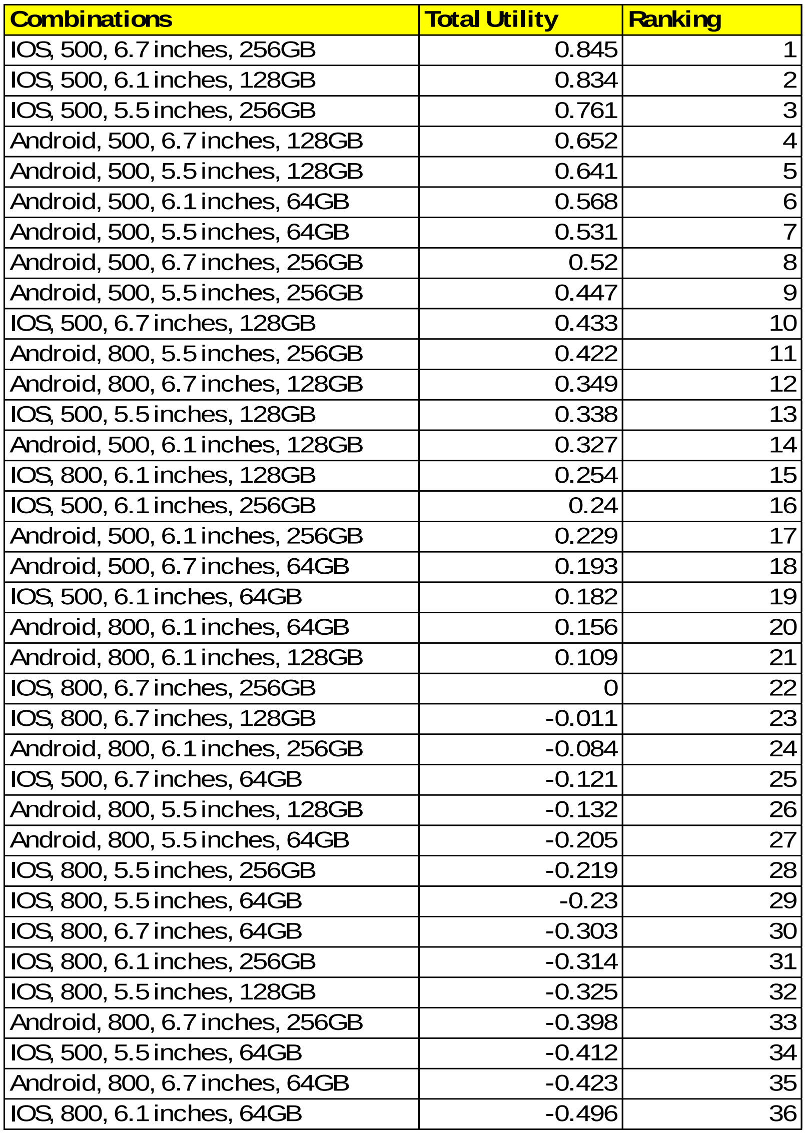


Correlations

		Total Utility	Ranking
Total Utility	Pearson Correlation	1	.773**
	Sig. (2-tailed)		<.001
	N	36	36
Ranking	Pearson Correlation	.773**	1
	Sig. (2-tailed)	<.001	
	N	36	36

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

To check the Correlations between Total Utility and Product Ranking we had to run the Pearson Correlation test with significance (2 tailed). We can see that there is a strong correlation between our variables hence our estimates for utilities are correct.



Combinations	Total Utility	Ranking
IOS, 500, 6.7 inches, 256GB	0.845	1
IOS, 500, 6.1 inches, 128GB	0.834	2
IOS, 500, 5.5 inches, 256GB	0.761	3
Android, 500, 6.7 inches, 128GB	0.652	4
Android, 500, 5.5 inches, 128GB	0.641	5
Android, 500, 6.1 inches, 64GB	0.568	6
Android, 500, 5.5 inches, 64GB	0.531	7
Android, 500, 6.7 inches, 256GB	0.52	8
Android, 500, 5.5 inches, 256GB	0.447	9
IOS, 500, 6.7 inches, 128GB	0.433	10
Android, 800, 5.5 inches, 256GB	0.422	11
Android, 800, 6.7 inches, 128GB	0.349	12
IOS, 500, 5.5 inches, 128GB	0.338	13
Android, 500, 6.1 inches, 128GB	0.327	14
IOS, 800, 6.1 inches, 128GB	0.254	15
IOS, 500, 6.1 inches, 256GB	0.24	16
Android, 500, 6.1 inches, 256GB	0.229	17
Android, 500, 6.7 inches, 64GB	0.193	18
IOS, 500, 6.1 inches, 64GB	0.182	19
Android, 800, 6.1 inches, 64GB	0.156	20
Android, 800, 6.1 inches, 128GB	0.109	21
IOS, 800, 6.7 inches, 256GB	0	22
IOS, 800, 6.7 inches, 128GB	-0.011	23
Android, 800, 6.1 inches, 256GB	-0.084	24
IOS, 500, 6.7 inches, 64GB	-0.121	25
Android, 800, 5.5 inches, 128GB	-0.132	26
Android, 800, 5.5 inches, 64GB	-0.205	27
IOS, 800, 5.5 inches, 256GB	-0.219	28
IOS, 800, 5.5 inches, 64GB	-0.23	29
IOS, 800, 6.7 inches, 64GB	-0.303	30
IOS, 800, 6.1 inches, 256GB	-0.314	31
IOS, 800, 5.5 inches, 128GB	-0.325	32
Android, 800, 6.7 inches, 256GB	-0.398	33
IOS, 500, 5.5 inches, 64GB	-0.412	34
Android, 800, 6.7 inches, 64GB	-0.423	35
IOS, 800, 6.1 inches, 64GB	-0.496	36

Conclusion

We may conclude from this analysis that conjoint analysis is a powerful technique that businesses can use to determine pricing in accordance with consumer preferences as well as market strategies prior to introducing a new mobile phone. For instance, The Combination with the lowest Utility was IOS, 800, 6.1 inches, 64GB. (-0.496) whereas the combination with the highest Utility was IOS, 500, 6.7 inches, 256GB (0.845). With this study we found out that people tend to buy cheap mobile phones with more features.