Aston Business School
Birmingham

PORTFOLIO ASSIGNMENT

Academic Year 2022/23

Module Code: BNM863

Module Name: Business Analytics in Practice

Module Leader: Dr Shubhadeep Mukherjee

Portfolio Title:

A portfolio of data analysis tasks using predictive and marketing analytics methods

Contents

Portfolio no. 1

Part A

Objective

The goal of the study is to collect knowledge about the factors affecting Fresco customers' spending patterns and to create a predictive model that would group consumers into various spending groups. We can improve marketing strategy and maximize promotional efforts by comprehending the factors that influence purchasing decisions.

Methodology:

Since this is a supervised learning, we aim to run multinomial logistic regression The analysis involved data cleaning, creation of dummy variables, The analysis then proceeded by iteratively refining the model to achieve a parsimonious and statistically significant solution. This involved including the relevant independent variables and assessing their significance in predicting customer spending behaviour. By considering variables such as value products, top Fresco products, age, gender, and store type, we sought to uncover the key drivers that impacted customers' decisions to spend at Fresco.

Data Summary:

The dataset used in this analysis contains customer spending information and other variables. To facilitate the analysis, we transformed the spender type into dummy variables: Low Spenders (assigned a value of 2.00), Medium Spenders (assigned a value of 1.00), and High Spenders (assigned a value of 0.00). Additionally, independent variables such as age, gender and store types were also converted into dummy variables to facilitate analysis. After removing all the independent variables due to their insignificance, Valued products and Top Fresco products are used to get the best outcome.

Results and Recommendations

The final model's accuracy in classifying clients into spending categories was 91.2% overall, which is a respectable level of accuracy. With significant factors included, the logistic regression equation demonstrated a high level of significance (p 0.05), indicating that the model is successful in predicting spending trends. Based on the analysis, we recommend focusing marketing efforts on high and medium spenders, as these groups had the highest percentage correct in classification. To maximize marketing impact, we suggest targeted promotions and campaigns tailored to these influential factors.

Conclusion

The study has given us important information about Fresco's client spending patterns. The marketing management team may make data-driven decisions, optimise resource allocation, and improve consumer engagement by using the built predictive model. The results emphasise the value of budget-friendly goods, premium Fresco goods, and focused marketing initiatives in encouraging consumer purchasing. The outcomes pave the way for more successful tactics and higher levels of client pleasure.

Part B

First of all creating dummy variables, for the spender type since it will be the dependent variable, since the data has 3 outcomes it is regarded as Multinomial logistic regression. The Values assigned are as follows.

Low Spenders	2.00
Medium Spenders	1.00
High Spenders	0.00

For the independent variable the data should be dichotomous and n-1 variables should be chosen hence convenient store and Superstore are chosen. To take control over the data Gender is also converted to dummy and assigned the values as follows.

Convenient	1 0
Superstore	0 1
Online	0 0

Male	0
Female	1

Now the data is ready for further analysis in SPSS

Spender type	Gender	Age	ConvenientStores	Superstore	ValueProducts	BrandProducts	TopFrescoProducts
1.00	.00	26.00	1.00	.00	8.00	2.00	1.00
1.00	1.00	33.00	.00	1.00	6.00	5.00	1.00
.00	.00	56.00	.00	.00	35.00	8.00	12.00
2.00	.00	27.00	1.00	.00	.00	1.00	1.00
.00	1.00	55.00	.00	.00	38.00	18.00	20.00
2.00	1.00	20.00	.00	1.00	4.00	2.00	.00
.00	.00	39.00	.00	.00	40.00	9.00	15.00
2.00	.00	24.00	1.00	.00	5.00	.00	1.00
1.00	.00	37.00	.00	1.00	10.00	9.00	7.00
1.00	.00	38.00	.00	.00	17.00	6.00	7.00
1.00	.00	49.00	.00	.00	15.00	7.00	6.00
1.00	.00	33.00	.00	1.00	16.00	5.00	3.00
1.00	.00	36.00	.00	1.00	18.00	7.00	6.00
2.00	.00	22.00	.00	1.00	5.00	2.00	2.00
1.00	1.00	37.00	.00	1.00	13.00	4.00	4.00
1.00	.00	40.00	.00	1.00	15.00	8.00	8.00
1.00	1.00	41.00	.00	.00	17.00	18.00	7.00
2.00	.00	19.00	.00	1.00	5.00	4.00	1.00
1.00	1.00	36.00	.00	1.00	7.00	8.00	8.00
1.00	.00	39.00	.00	1.00	8.00	7.00	4.00
1.00	.00	23.00	1.00	.00	5.00	3.00	3.00
2.00	1.00	21.00	1.00	.00	8.00	2.00	1.00
1.00	.00	21.00	1.00	.00	7.00	9.00	4.00

Figure 1

To check the linearity of the model, ln of Age, Value Products And top fresco products are executed by computing variables. After that multinomial logistic regression is ran to see the assumption by looking at the significance level. The following execution has been found for both models.

Parameter Estimates

Dummy_SB[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
High Spenders	Intercept	-7.370	2.117	12.120	1	<.001			
	Age * LN_AGE	.015	.009	2.763	1	.096	1.016	.997	1.034
	Value Products * LN_VP	.024	.021	1.299	1	.254	1.024	.983	1.068
	Brand Products * LN_BP	.038	.037	1.046	1	.306	1.039	.966	1.117
	Top Fresco Products * LN_TFP	.138	.059	5.468	1	.019	1.148	1.023	1.288
Low Spenders	Intercept	7.454	3.289	5.138	1	.023			
	Age * LN_AGE	-.066	.041	2.543	1	.111	.936	.864	1.015
	Value Products * LN_VP	-.067	.077	.742	1	.389	.936	.804	1.089
	Brand Products * LN_BP	-.154	.124	1.545	1	.214	.857	.673	1.093
	Top Fresco Products * LN_TFP	-.111	.123	.816	1	.366	.895	.702	1.139

a. The reference category is: Medium Spenders.

Figure 2

In the above figure we can suggest that all the variables are insignificant($p<0.05$) in both the model refers that the assumption of linearity is satisfied. None of the variable will not violate the linear regression.

Model Adequacy: Its turn to check the model adequacy and run the regression multiple times to eliminate the variable that are insignificant in both of the models and search for parsimonious model on which the variables are more significant.

Case Processing Summary

		N	Marginal Percentage
Dummy_SB	High Spenders	27	36.0%
	Medium Spenders	30	40.0%
	Low Spenders	18	24.0%
Convenient Stores	.00	55	73.3%
	1.00	20	26.7%
Gender	Male	39	52.0%
	Female	36	48.0%
Superstore	.00	41	54.7%
	1.00	34	45.3%
Valid		75	100.0%
Missing		0	
Total		75	
Subpopulation		75[a]	

a. The dependent variable has only one value observed in 75 (100.0%) subpopulations.

Figure 3

The data above explains the marginal percentage of the whole data and explaining the data is fully cleaned hence no missing values.

Pseudo R-Square

Cox and Snell	.785
Nagelkerke	.888
McFadden	.714

Figure 4

The Pseudo R-square shows that all the values are close to 1 hence it is assumed that the model is a good fit.

Likelihood Ratio Tests

Effect	Model Fitting Criteria: -2 Log Likelihood of Reduced Model	Likelihood Ratio Tests: Chi-Square	df	Sig.
Intercept	46.186[a]	.000	0	.
Value Products	50.003	3.817	2	.148
Top Fresco Products	56.074	9.889	2	.007
Age	51.288	5.102	2	.078
Brand Products	48.375	2.189	2	.335
Convenient Stores	46.770	.584	2	.747
Gender	46.612	.426	2	.808
Superstore	46.453	.267	2	.875

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

Figure 5

The null hypothesis tested is that all parameters of the omitted effect are equal to 0. The significance level (Sig.) indicates the probability of obtaining the observed chi-square statistic under the null hypothesis. In this specific output, the significance levels are show in figure 5 .

Parameter Estimates

Dummy_SB[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
High Spenders	Intercept	-25.960	3.136	68.548	1	<.001			
	Value Products	.086	.083	1.069	1	.301	1.090	.926	1.284
	Top Fresco Products	.428	.183	5.479	1	.019	1.535	1.072	2.197
	Age	.071	.054	1.747	1	.186	1.074	.966	1.193
	Brand Products	.101	.140	.528	1	.468	1.107	.842	1.455
	[Convenient Stores=.00]	16.612	.000	.	1	.	16391754.689	16391754.689	16391754.689
	[Convenient Stores=1.00]	0[b]	.	.	0	.	.	.	.
	[Gender=.00]	-.355	1.169	.092	1	.761	.701	.071	6.928
	[Gender=1.00]	0[b]	.	.	0	.	.	.	.
	[Superstore=.00]	.644	1.241	.270	1	.604	1.904	.167	21.668
	[Superstore=1.00]	0[b]	.	.	0	.	.	.	.
Low Spenders	Intercept	18.342	4862.752	.000	1	.997			
	Value Products	-.403	.318	1.602	1	.206	.668	.358	1.247
	Top Fresco Products	-.533	.498	1.147	1	.284	.587	.221	1.557
	Age	-.290	.210	1.902	1	.168	.748	.495	1.130
	Brand Products	-.385	.325	1.399	1	.237	.681	.360	1.288
	[Convenient Stores=.00]	-7.398	4862.755	.000	1	.999	.001	.000	.[c]
	[Convenient Stores=1.00]	0[b]	.	.	0	.	.	.	.
	[Gender=.00]	.747	1.278	.341	1	.559	2.110	.172	25.848
	[Gender=1.00]	0[b]	.	.	0	.	.	.	.
	[Superstore=.00]	-5.451	4862.757	.000	1	.999	.004	.000	.[c]
	[Superstore=1.00]	0[b]	.	.	0	.	.	.	.

a. The reference category is: Medium Spenders.

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.

Figure 6

The significance is level of all the variables are given in the parameter estimates superstore shows the highest insignificance in both models. Therefore, it must be removed to step forward to get the model of parsimony.

Parameter Estimates

Dummy_SB[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
High Spenders	Intercept	-25.613	3.050	70.524	1	<.001			
	Value Products	.090	.080	1.280	1	.258	1.094	.936	1.280
	Top Fresco Products	.428	.182	5.502	1	.019	1.534	1.073	2.193
	Age	.067	.052	1.706	1	.192	1.070	.967	1.183
	Brand Products	.126	.129	.952	1	.329	1.134	.881	1.460
	[Convenient Stores=.00]	16.230	.000	.	1	.	11179208.946	11179208.946	11179208.946
	[Convenient Stores=1.00]	0[b]	.	.	0	.	.	.	.
	[Gender=.00]	-.094	1.036	.008	1	.928	.910	.119	6.942
	[Gender=1.00]	0[b]	.	.	0	.	.	.	.
Low Spenders	Intercept	12.892	7.002	3.390	1	.066			
	Value Products	-.403	.318	1.602	1	.206	.668	.358	1.247
	Top Fresco Products	-.533	.498	1.147	1	.284	.587	.221	1.557
	Age	-.290	.210	1.902	1	.168	.748	.495	1.130
	Brand Products	-.385	.325	1.399	1	.237	.681	.360	1.288
	[Convenient Stores=.00]	-1.947	1.888	1.064	1	.302	.143	.004	5.768
	[Convenient Stores=1.00]	0[b]	.	.	0	.	.	.	.
	[Gender=.00]	.747	1.278	.341	1	.559	2.110	.172	25.851
	[Gender=1.00]	0[b]	.	.	0	.	.	.	.

a. The reference category is: Medium Spenders.

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

Figure 7

After removing the superstore other variables significance has increased i.e values come near to 0.05. Now Gender has the highest insignificance in both the models. So, it has to be removed to get closer to parsimonious model.

Parameter Estimates

Dummy_SB[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
High Spenders	Intercept	-25.677	3.034	71.621	1	<.001			
	Value Products	.089	.078	1.293	1	.255	1.093	.938	1.273
	Top Fresco Products	.428	.183	5.474	1	.019	1.535	1.072	2.197
	Age	.067	.051	1.716	1	.190	1.069	.967	1.181
	Brand Products	.130	.122	1.125	1	.289	1.138	.896	1.446
	[Convenient Stores=.00]	16.265	.000	.	1	.	11586256.156	11586256.156	11586256.156
	[Convenient Stores=1.00]	0[b]	.	.	0	.	.	.	.
Low Spenders	Intercept	14.271	6.900	4.277	1	.039			
	Value Products	-.442	.325	1.848	1	.174	.643	.340	1.216
	Top Fresco Products	-.567	.477	1.415	1	.234	.567	.223	1.444
	Age	-.309	.206	2.249	1	.134	.734	.490	1.099
	Brand Products	-.393	.326	1.454	1	.228	.675	.356	1.279
	[Convenient Stores=.00]	-2.203	1.900	1.345	1	.246	.110	.003	4.575
	[Convenient Stores=1.00]	0[b]	.	.	0	.	.	.	.

a. The reference category is: Medium Spenders.

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

Figure 8

Other variables now get closer to significance level and Brand products shows the most insignificant value. By eliminating it, the model gets more better.

Parameter Estimates

Dummy_SB[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
High Spenders	Intercept	-25.288	3.000	71.054	1	<.001			
	Value Products	.143	.068	4.373	1	.037	1.154	1.009	1.319
	Top Fresco Products	.419	.181	5.350	1	.021	1.520	1.066	2.169
	Age	.074	.049	2.259	1	.133	1.077	.978	1.186
	[Convenient Stores=.00]	16.041	.000	.	1	.	9254903.779	9254903.779	9254903.779
	[Convenient Stores=1.00]	0[b]	.	.	0	.	.	.	.
Low Spenders	Intercept	15.066	6.617	5.185	1	.023			
	Value Products	-.516	.300	2.966	1	.085	.597	.332	1.074
	Top Fresco Products	-.825	.467	3.125	1	.077	.438	.175	1.094
	Age	-.347	.191	3.292	1	.070	.707	.486	1.028
	[Convenient Stores=.00]	-2.365	1.783	1.760	1	.185	.094	.003	3.093
	[Convenient Stores=1.00]	0[b]	.	.	0	.	.	.	.

a. The reference category is: Medium Spenders.

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

Figure 9

After eliminating Brand products most of the values come near towards significance. However Convenient Stores value is still insignificant.

Parameter Estimates

Dummy_SB[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
High Spenders	Intercept	-6.638	1.862	12.704	1	<.001			
	Value Products	.172	.070	6.087	1	.014	1.188	1.036	1.362
	Top Fresco Products	.454	.166	7.450	1	.006	1.575	1.137	2.182
Low Spenders	Intercept	4.994	1.653	9.127	1	.003			
	Value Products	-.452	.171	7.028	1	.008	.636	.455	.889
	Top Fresco Products	-.645	.248	6.762	1	.009	.525	.323	.853

a. The reference category is: Medium Spenders.

Figure 10

All the variables become less than 0.05 and we can say it as the parsimonious model. Every time when running the model to find the final model, Case process summary, Pseudo R-Square and classification has been changed. For the parsimonious model, Final out puts are given as Figure 11, 12 & 13.The convenient store is removed at the last because it has the significance only from low spenders model.

For High |Spenders

$\ln[(P/(1-P))$ = exp(-6.638 + 0.172 * 1.188 + 0.454 * 1.575)

For Low Spenders:

$\ln[(P/(1-P))$ = exp(4.994 - 0.452 * .455 - 0.645 * 0.323)

Case Processing Summary

		N	Marginal Percentage
Dummy_SB	High Spenders	27	36.0%
	Medium Spenders	30	40.0%
	Low Spenders	18	24.0%
Valid		75	100.0%
Missing		0	
Total		75	
Subpopulation		65[a]	

a. The dependent variable has only one value observed in 63 (96.9%) subpopulations.

Figure 11

Model Fitting Information

Model	Model Fitting Criteria -2 Log Likelihood	Likelihood Ratio Tests Chi-Square	df	Sig.
Intercept Only	158.750			
Final	61.853	96.897	4	<.001

Figure 13

Pseudo R-Square

Cox and Snell	.725
Nagelkerke	.820
McFadden	.600

Figure 12

Classification

Observed	Predicted High Spenders	Medium Spenders	Low Spenders	Percent Correct
High Spenders	23	4	0	85.2%
Medium Spenders	1	24	5	80.0%
Low Spenders	0	4	14	77.8%
Overall Percentage	32.0%	42.7%	25.3%	81.3%

Figure 13

From the classification table High Spenders and Medium Spenders have the highest percentage correct therefore low spenders are removed by selecting cases and filtered out lower spender.

Since the model is now more specific to check multicollinearity df beta and cooks distance , **Linear Regression** is now executed.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.881[a]	.777	.754	.38240

a. Predictors: (Constant), Top Fresco Products, Superstore, Gender, Age, Value Products, Brand Products, Convenient Stores

Figure 14

The model has an R-squared value of 0.777, which means that approximately 77.7% of the variability in the dependent variable can be explained by the independent variables included in the model. The adjusted R-squared value, taking into account the number of predictors and sample size, is 0.754. This adjusted value provides a more reliable estimate of the model's explanatory power. The standard error of the estimate is 0.38240, which represents the average distance between the observed values and the predicted values by the model

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	2.064	.250		8.272	<.001		
	Gender	-.048	.100	-.032	-.483	.631	.778	1.285
	Age	-.016	.005	-.286	-3.187	.002	.413	2.420
	Convenient Stores	.251	.187	.145	1.343	.184	.286	3.501
	Superstore	.045	.140	.029	.320	.750	.400	2.497
	Value Products	-.012	.007	-.185	-1.773	.081	.306	3.273
	Brand Products	-.023	.013	-.184	-1.790	.078	.313	3.191
	Top Fresco Products	-.036	.013	-.250	-2.730	.008	.398	2.511

a. Dependent Variable: Dummy_SB

Figure 15

Both the Collinearity Tolerance and Statistics VIF are less than 10 and greater than 0.1 respectively.

Statistics

Analog of Cook's influence statist

N	Valid	57
	Missing	18
Mean		.0481086
Median		.0031534
Std. Deviation		.11493170
Range		.62625
Minimum		.00000
Maximum		.62625

Figure 16

Cooks distance also approves the models adequacy which is less than 1 .

Statistics

Normalized residual

N	Valid	57
	Missing	0
Mean		-.0415360
Median		.0901750
Std. Deviation		.86838617
Range		5.78963
Minimum		-4.10444
Maximum		1.68518

Figure 17

Most of the values are within -2.5 to 2.5 hence aligns with the models adequacy.

Statistics

DFBETA for constant

N	Valid	57
	Missing	0
Mean		.0016091
Median		.0610566
Std. Deviation		.26542247
Range		1.64448
Minimum		-1.42145
Maximum		.22304

Statistics

DFBETA for Value Products

N	Valid	57
	Missing	0
Mean		-.0001474
Median		-.0007753
Std. Deviation		.00882143
Range		.06398
Minimum		-.02506
Maximum		.03892

Statistics

DFBETA for Top Fresco Products

N	Valid	57
	Missing	0
Mean		.0001424
Median		-.0030697
Std. Deviation		.02218016
Range		.13170
Minimum		-.02909
Maximum		.10261

FIGURE 15

Df beta for all the variables are less than 1 which refers.

Now Running the Binary logistics for the final model classification table executed is given below. It concludes the model achieved an overall percentage correct of 52.6%, which indicates the proportion of correctly classified cases across both categories.

Classification Table[a,b]

			Predicted		
			Dummy_SB		
	Observed		High Spenders	Medium Spenders	Percentage Correct
Step 0	Dummy_SB	High Spenders	0	27	.0
		Medium Spenders	0	30	100.0
	Overall Percentage				52.6

a. Constant is included in the model.

b. The cut value is .500

Figure 16

Classification Table[a]

			Predicted		
			Dummy_SB		
	Observed		High Spenders	Medium Spenders	Percentage Correct
Step 1	Dummy_SB	High Spenders	23	4	85.2
		Medium Spenders	1	29	96.7
	Overall Percentage				91.2

a. The cut value is .500

Figure 17

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Value Products	-.173	.070	6.099	1	.014	.841
	Top Fresco Products	-.454	.167	7.410	1	.006	.635
	Constant	6.649	1.869	12.661	1	<.001	772.244

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

Figure 18

Both the expected B is positive hence showing positive relationship.

The logistic regression equation based on the variables is:

$\ln(P / (1 - P)) = 6.649 - 0.173 * .841 - 0.454 * 0.635$

$\ln(P/(1-P)) = 6.215$

$P = 94.4\ \%$

Probability and the overall percentage in figure 17 show very less deviation hence it can be said that it is a fit model and prediction is very up to the mark.

Conclusion

Based on the analysis, we recommend focusing marketing efforts on high and medium spenders, as these groups had the highest percentage correct in classification. To maximize marketing impact, we suggest targeted promotions and campaigns tailored to these influential factors

Portfolio-2

Objective:

Performing segmentation analysis on a dataset including customer information from a UK bank is the goal of this work. Through cluster analysis, the investigation seeks to identify trends and patterns in the data that can be used to develop financial products and promotions catered to client categories.

Analyze

To analyse the data into SPSS, it needs to be data's to converted to Dummy variables as shown in the table given below. The SPSS analyse has been ran for 5 times. 1st time with normal cluster then assigned with Wards and Furthers with 3 and 4 Clusters respectively.

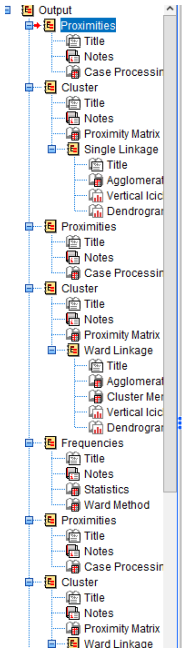
Gender		**Marital Status**	
M	0	Single	0
F	1	married	1
		Divorces	2
Housing			
Rent	0	**Credit Risk**	
own	1	Low	0
others	2	High	1
job			
unskilled	0		
skilled	1		
managem	2		
unemploye	3		

Figure1

Gender	Marital Statu	Marital Statu	Age	Age group	Housing	Housing	Job	Job	Credit Risk	Credit Risk
0	Single	0	23	Young	Own	1	Unskilled	0	Low	0
0	Divorced	2	32	Young	Own	1	Skilled	1	High	1
0	Single	0	38	Senior	Own	1	Management	2	High	1
0	Single	0	36	Senior	Own	1	Unskilled	0	High	1
0	Single	0	31	Young	Rent	0	Skilled	1	Low	0
0	Married	1	25	Young	Own	1	Skilled	1	Low	0
0	Married	1	26	Young	Own	1	Unskilled	0	Low	0
0	Single	0	27	Young	Own	1	Unskilled	0	Low	0
0	Single	0	25	Young	Own	1	Skilled	1	High	1
1	Divorced	2	43	Senior	Own	1	Skilled	1	High	1
0	Single	0	32	Young	Rent	0	Management	2	High	1
0	Single	0	34	Young	Rent	0	Unskilled	0	Low	0
0	Married	1	26	Young	Own	1	Skilled	1	Low	0
0	Single	0	44	Senior	Own	1	Skilled	1	High	1
0	Single	0	46	Senior	Own	1	Unskilled	0	Low	0
0	Divorced	2	39	Senior	Own	1	Management	2	Low	0
1	Divorced	2	25	Young	Own	1	Skilled	1	High	1
0	Single	0	31	Young	Own	1	Skilled	1	Low	0
0	Single	0	47	Senior	Own	1	Skilled	1	Low	0
1	Divorced	2	23	Young	Rent	0	Skilled	1	High	1
1	Divorced	2	22	Young	Own	1	Skilled	1	High	1
1	Divorced	2	26	Young	Rent	0	Skilled	1	High	1
0	Married	1	19	Young	Own	1	Skilled	1	High	1
1	Divorced	2	27	Young	Own	1	Management	2	High	1
0	Single	0	39	Senior	Rent	0	Unskilled	0	Low	0
0	Single	0	26	Young	Own	1	Skilled	1	Low	0
0	Single	0	50	Senior	Other	2	Skilled	1	High	1
0	Single	0	34	Young	Other	2	Skilled	1	Low	0
0	Single	0	23	Young	Rent	0	Skilled	1	Low	0
0	Single	0	23	Young	Own	1	Skilled	1	Low	0
0	Single	0	46	Senior	Other	2	Skilled	1	High	1
0	Single	0	35	Senior	Own	1	Skilled	1	Low	0
1	Divorced	2	28	Young	Own	1	Skilled	1	Low	0
0	Single	0	25	Young	Rent	0	Skilled	1	Low	0
1	Divorced	2	36	Senior	Rent	0	Skilled	1	High	1
0	Single	0	41	Senior	Own	1	Unskilled	0	Low	0
0	Divorced	2	54	Senior	Own	1	Skilled	1	High	1
1	Divorced	2	43	Senior	Own	1	Unskilled	0	Low	0
0	Married	1	33	Young	Own	1	Skilled	1	High	1
0	Single	0	34	Young	Own	1	Skilled	1	High	1
0	Single	0	39	Senior	Other	2	Unemployed	3	Low	0

Figurfe 2

The date is now ready to for further SPSS analysis. Iterative the process of clustering have to be done to get the best equal distribution of cluster. Therefore first of all normally run without any constrain,Hierarchical cluster to analyze the data which are given below.



Proximities

Case Processing Summary[a]

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

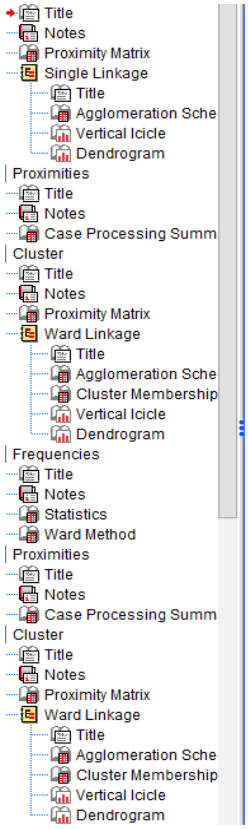
a. Euclidean Distance used

Cluster

Proximity Matrix

Case	Euclidean Distance															
	1: .98	2: .06	3: .25	4: .05	5: .98	6: .37	7: .55	8: .79	9: .16	10: .26	11: .64	12: .34	13: .98	14: .23	15: .47
1: .98	.000	1.476	1.528	1.028	.813	.621	.505	.113	1.160	1.812	1.418	.555	.606	1.213	.436
2: .06	1.476	.000	1.460	1.079	1.574	1.160	1.194	1.465	1.078	1.023	1.219	1.559	1.151	1.153	1.503
3: .25	1.528	1.460	.000	1.110	1.382	1.498	1.581	1.567	1.095	1.750	1.207	1.581	1.485	.594	1.444
4: .05	1.028	1.079	1.110	.000	1.282	1.189	1.133	1.019	.604	1.472	.998	1.129	1.182	.578	1.021
5: .98	.813	1.574	1.382	1.282	.000	.912	.944	.831	1.217	1.877	1.145	.606	.903	1.191	.820
6: .37	.621	1.160	1.498	1.189	.912	.000	.365	.628	1.197	1.554	1.405	.828	.132	1.269	.731
7: .55	.505	1.194	1.581	1.133	.944	.365	.000	.514	1.264	1.581	1.502	.739	.338	1.287	.624
8: .79	.113	1.465	1.567	1.019	.831	.628	.514	.000	1.157	1.796	1.403	.547	.613	1.223	.401
9: .16	1.160	1.078	1.095	.604	1.217	1.197	1.264	1.157	.000	1.490	.723	1.290	1.228	.682	1.224
10: .26	1.812	1.023	1.750	1.472	1.877	1.554	1.581	1.796	1.490	.000	1.591	1.863	1.550	1.500	1.783
11: .64	1.418	1.219	1.207	.998	1.145	1.405	1.502	1.403	.723	1.591	.000	1.337	1.402	.902	1.445
12: .34	.555	1.559	1.581	1.129	.606	.828	.739	.547	1.290	1.863	1.337	.000	.807	1.271	.576
13: .98	.606	1.151	1.485	1.182	.903	.132	.338	.613	1.228	1.550	1.402	.807	.000	1.255	.711
14: .23	1.213	1.153	.594	.578	1.191	1.269	1.287	1.223	.682	1.500	.902	1.271	1.255	.000	1.114
15: .47	.436	1.503	1.444	1.021	.820	.731	.624	.401	1.224	1.783	1.445	.576	.711	1.114	.000
16: .67	1.265	1.069	1.703	1.588	1.246	.731	.908	1.248	1.514	1.459	1.531	1.348	.718	1.530	1.247
17: 1.00	1.843	1.069	1.816	1.562	1.855	1.607	1.629	1.838	1.456	.483	1.543	1.930	1.608	1.578	1.890
18: .03	.406	1.453	1.288	1.070	.678	.551	.626	.431	1.124	1.772	1.312	.638	.535	1.073	.436
19: .54	1.056	1.741	1.075	1.391	.884	1.113	1.123	1.097	1.445	1.973	1.576	1.108	1.096	1.098	.856
20: .03	1.929	1.289	1.949	1.674	1.880	1.708	1.735	1.932	1.705	.918	1.688	1.837	1.694	1.751	1.990

Figure1



Single Linkage

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	121	145	.020	0	0	19
2	59	307	.040	0	0	4
3	8	272	.041	0	0	53
4	59	377	.048	2	0	14
5	76	415	.048	0	0	17
6	242	333	.049	0	0	217
7	98	218	.051	0	0	103
8	315	394	.051	0	0	28
9	36	248	.051	0	0	115
10	132	141	.057	0	0	37
11	211	243	.059	0	0	13
12	146	374	.061	0	0	290
13	211	362	.062	11	0	33
14	47	59	.063	0	4	18
15	297	341	.064	0	0	36
16	165	246	.071	0	0	17
17	76	165	.072	5	16	56
18	47	413	.076	14	0	24
19	121	166	.076	1	0	21
20	274	409	.076	0	0	54
21	33	121	.077	0	19	23
22	179	417	.077	0	0	238
23	33	57	.078	21	0	40
24	47	376	.079	18	0	47
25	87	399	.081	0	0	28
26	126	368	.082	0	0	82
27	18	264	.082	0	0	55
28	87	315	.083	25	8	90

Figure 2

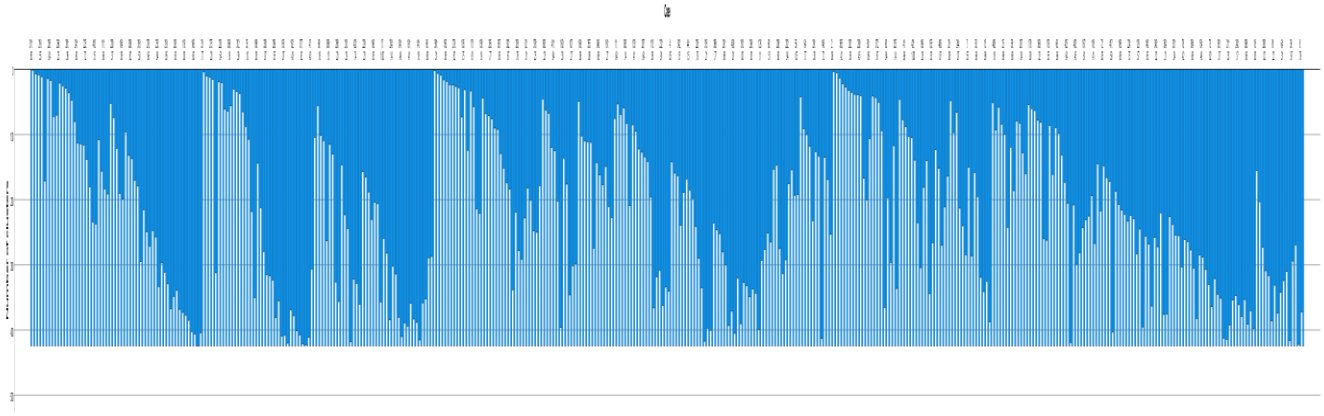


Figure 3

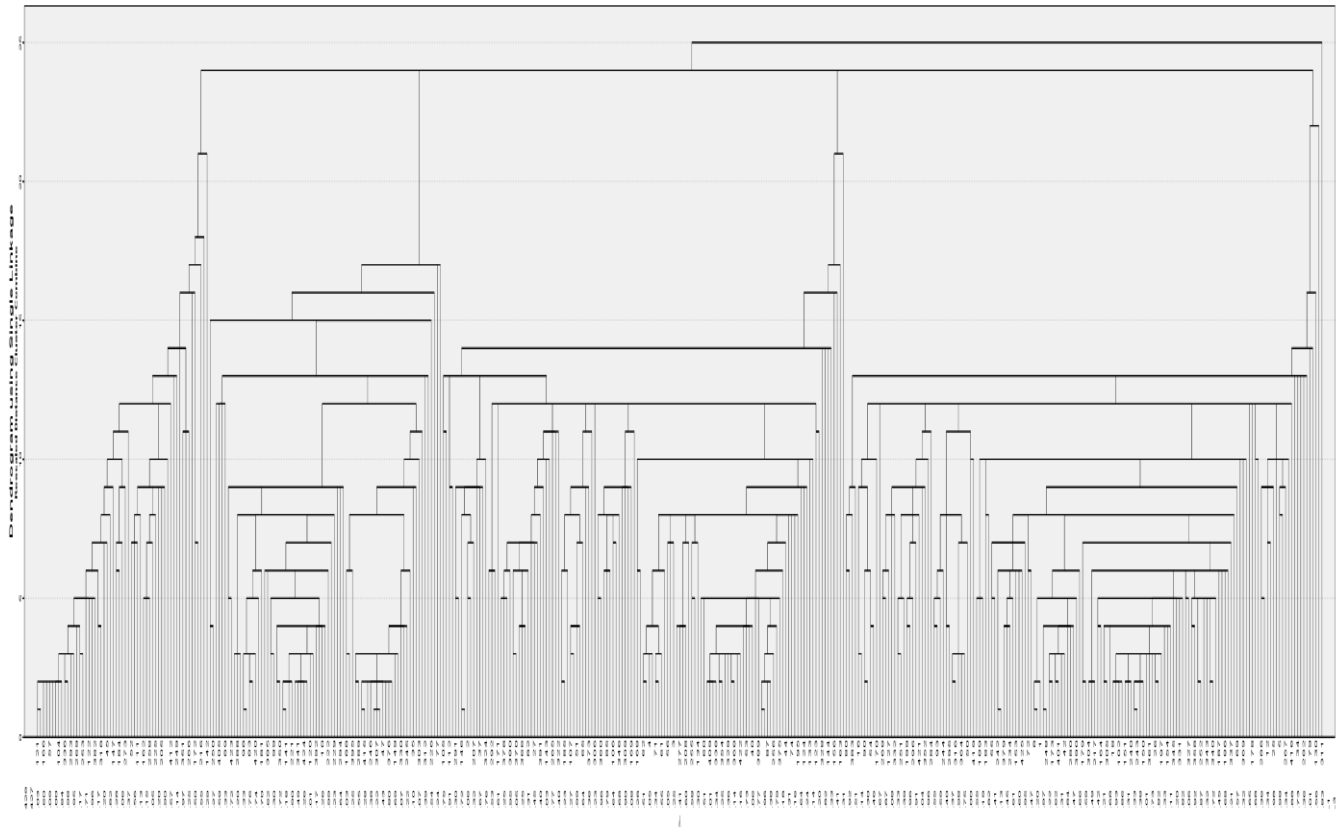
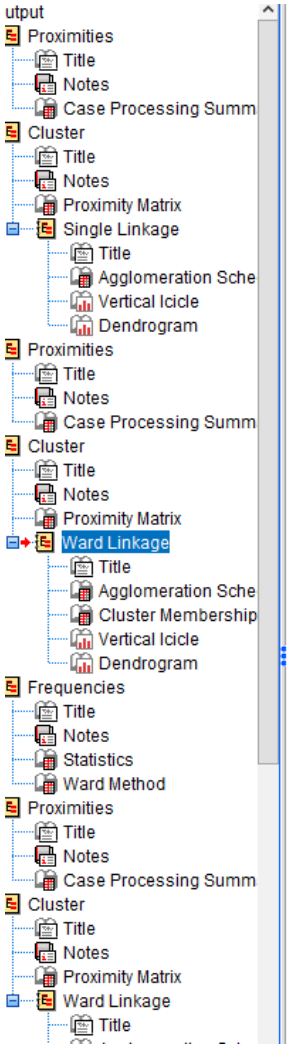


Figure 4

From the above diagram we can assume that the good distribution finds in between cluster number 3 and 4 .

Now, analysing the data with wards method to check the frequency distribution.



Ward Linkage

Agglomeration Schedule

Stage	Cluster Combined Cluster 1	Cluster 2	Coefficients	Stage Cluster First Appears Cluster 1	Cluster 2	Next Stage
1	121	145	.010	0	0	47
2	59	307	.030	0	0	15
3	8	272	.051	0	0	63
4	76	415	.075	0	0	60
5	242	333	.099	0	0	289
6	98	218	.125	0	0	108
7	315	394	.150	0	0	70
8	36	248	.175	0	0	196
9	132	141	.204	0	0	97
10	211	243	.234	0	0	12
11	146	374	.264	0	0	327
12	211	362	.296	10	0	68
13	297	341	.328	0	0	65
14	165	246	.363	0	0	60
15	59	377	.399	2	0	39
16	274	409	.437	0	0	90
17	179	417	.476	0	0	313
18	87	399	.516	0	0	70
19	126	368	.557	0	0	72
20	18	264	.598	0	0	96
21	160	185	.640	0	0	191
22	103	252	.681	0	0	162
23	191	424	.723	0	0	89
24	300	338	.765	0	0	90
25	102	398	.809	0	0	76
26	310	354	.853	0	0	158
27	376	413	.897	0	0	80
28	6	316	.941	0	0	92
29	81	173	.986	0	0	163
30	57	92	1.032	0	0	83
31	204	350	1.079	0	0	214

Double-click to activate

Figure 5

The above figure shows the ward linkage that comes out before dendrogram and vertical Icicle.

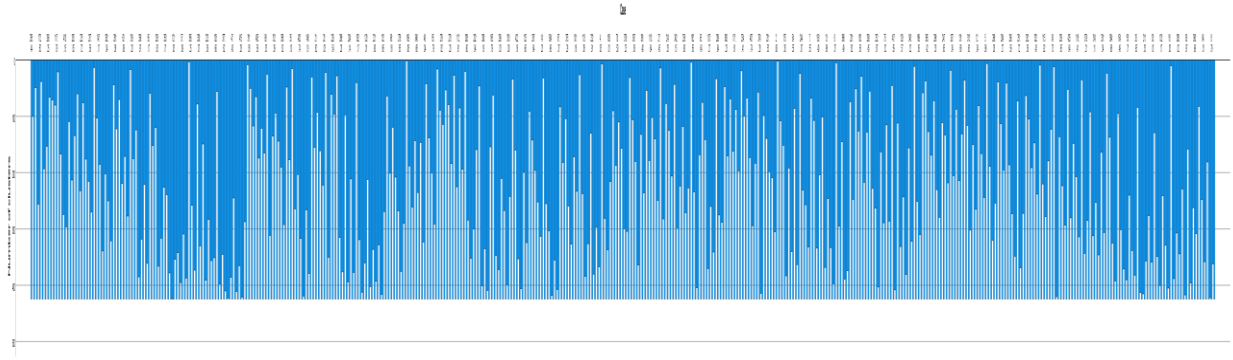


Figure 6

It show the vertical ICICLE for wards method and 4 clusters.

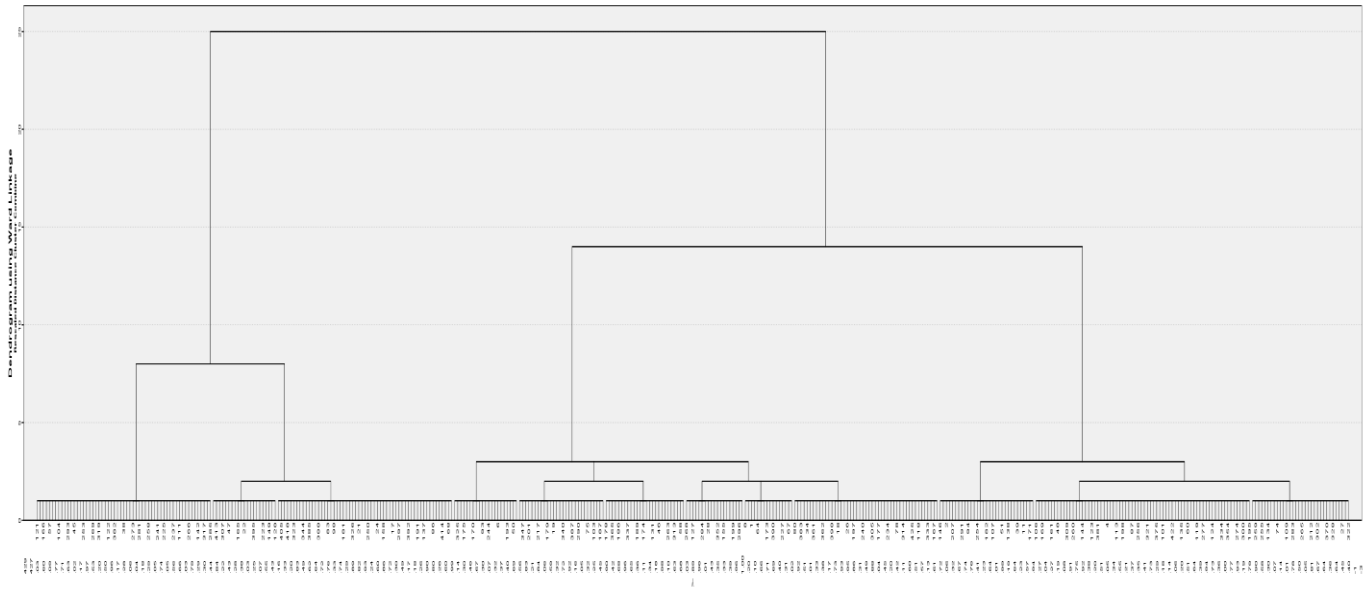


Figure 7

In the given dendogram on figure 7 the cluster is found not to be equally distributed.

Statistics

Ward Method

N	Valid	425
	Missing	0
Mode		1
Range		3
Minimum		1
Maximum		4

Ward Method

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	157	36.9	36.9	36.9
	2	133	31.3	31.3	68.2
	3	78	18.4	18.4	86.6
	4	57	13.4	13.4	100.0
	Total	425	100.0	100.0	

Figure 8

The charts on figure 8 explains The dataset had 425 customer records. The analysis grouped the records into four clusters. Cluster 1: It has 157 records, accounting for 36.9% of the total and valid records. Cluster 2: It has 133 records, representing 31.3% of the total and valid records. Cluster 3: It has 78 records, making up 18.4% of the total and valid records. Cluster 4: It has 57 records, accounting for 13.4% of the total and valid records.

Analyse with 3 Cluster ward method, The dendrogram seems a bit equally distributed compared to the previous ones.

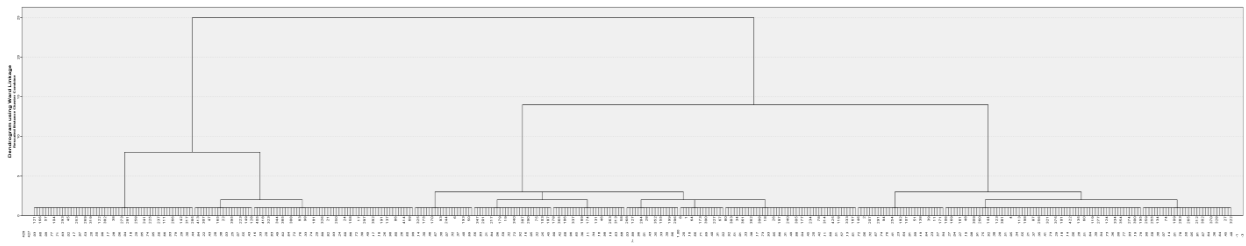


Figure 9

Ward Method

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	157	36.9	36.9	36.9
	2	133	31.3	31.3	68.2
	3	135	31.8	31.8	100.0
	Total	425	100.0	100.0	

Figure 10

The table above explains Cluster 1 consists of 157 records, which accounts for 36.9% of the total and valid records. This means that approximately 36.9% of the customers in the dataset belong to Cluster 1. Cluster 2 contains 133 records, representing 31.3% of the total and valid records. Thus, around 31.3% of the customers fall into Cluster 2. Cluster 3 comprises 135 records, making up 31.8% of the total and valid records. This indicates that roughly 31.8% of the customers belong to Cluster 3.

With Furthest method and 4 cluster the dendrogram in figure 11 is very unequally distributed.

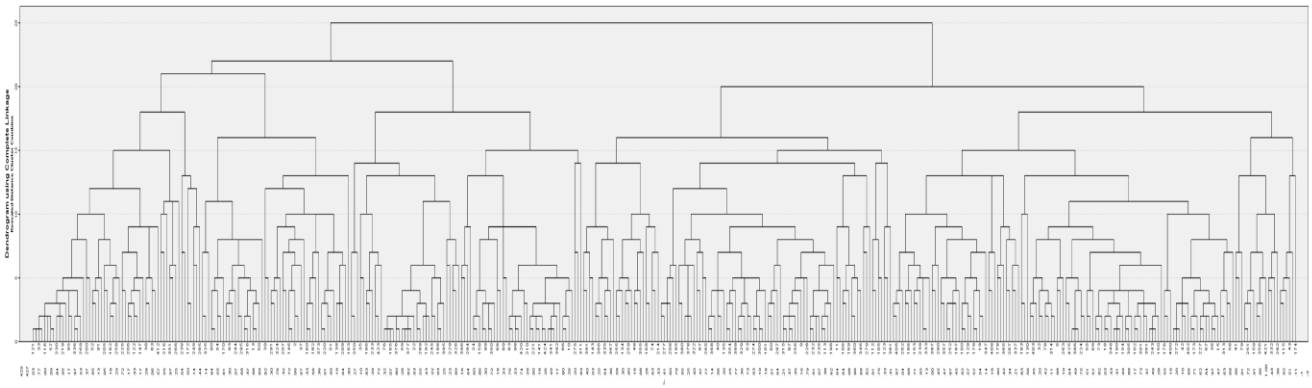


Figure 11.

Complete Linkage

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	240	56.5	56.5	56.5
	2	50	11.8	11.8	68.2
	3	78	18.4	18.4	86.6
	4	57	13.4	13.4	100.0
	Total	425	100.0	100.0	

Figure 12

The frequency column shows that it is unequally distributed, hence assumed it is not a good cluster.

Moving forward with the Furthest method with 3 clusters :

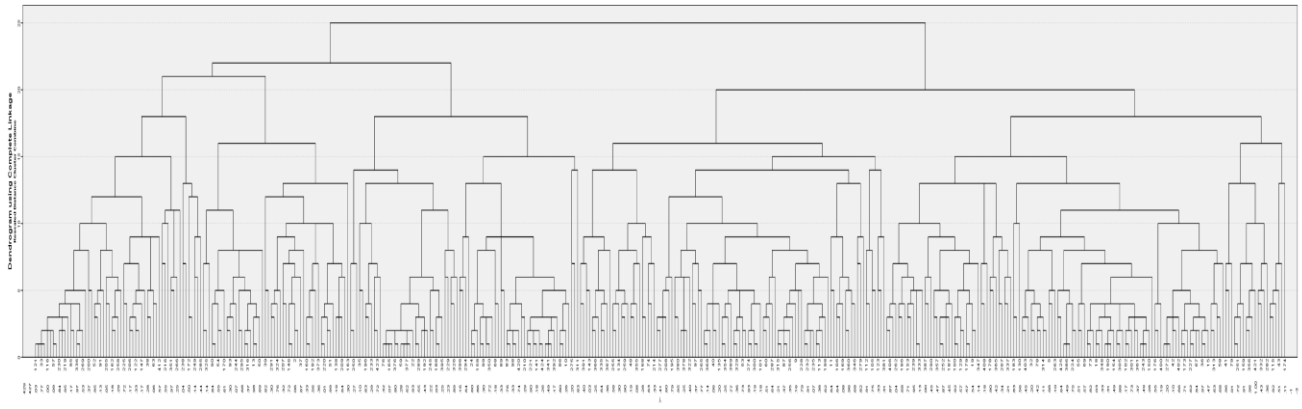


Figure 13

The frequency of the cluster is not equally distributed in this dendrogram in figure 13

Complete Linkage

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	240	56.5	56.5	56.5
	2	107	25.2	25.2	81.6
	3	78	18.4	18.4	100.0
	Total	425	100.0	100.0	

Figure 14

The descriptive also suggest that the data is not equally Distributed among the clusters.

Hence we can conclude from all the cluster analysis Wards Method with cluster 3 executes the best distribution model.

Therefore from the data of SPSS putting it into Excel file we can create Pivot table to get more details about the cluster analysis.

Report Summary

Wards method with 3 cluster

Row Labels	High	Low	Grand Total
1		**157**	**157**
Management		**22**	**22**
Other		5	5
Own		16	16
Rent		1	1
Skilled		**99**	**99**
Other		12	12
Own		76	76
Rent		11	11
Unemployed		**3**	**3**
Other		1	1
Own		2	2
Unskilled		**33**	**33**
Other		1	1
Own		28	28
Rent		4	4
2	**133**		**133**
Management	**21**		**21**
Other	3		3
Own	15		15
Rent	3		3
Skilled	**83**		**83**
Other	19		19
Own	52		52
Rent	12		12
Unemployed	**1**		**1**
Own	1		1
Unskilled	**28**		**28**
Other	1		1
Own	22		22
Rent	5		5
3	**78**	**57**	**135**
Management	**7**	**4**	**11**
Other	2	1	3
Own	5	2	7
Rent		1	1
Skilled	**52**	**37**	**89**
Other	2		2
Own	27	29	56
Rent	23	8	31
Unemployed	**4**	**3**	**7**
Other	1		1
Own	1	2	3
Rent	2	1	3
Unskilled	**15**	**13**	**28**
Other	3	1	4
Own	8	6	14
Rent	4	6	10
Grand Total	**211**	**214**	**425**

Figure 15

In cluster one of Wards method with 3 cluster given in figure 15, in cluster 1 the total number of cases is 157 where all of them belongs to low credit risk group. In Cluster 2, 133 cases has been observed where all of them belongs to High credit risk.

Finally in cluster 3, the number of cases that has been observed is 135 among them 78 belongs to HIGH credit risk and 57 belongs to low credit risk. Here 11 of the people works in Manage ment where 7 and 4 belongs high and low credit risk. Among 11 people 7 of them belongs to their own house, 5 an 2 people belongs to high and low credit risk respectively. People who are homeless , 2 people and 1 people belongs to high and low credit risk respectively. One of them lives in rented house who lives in rented house.

Among skilled workers, total 89 cases has been recorded among them 52 belongs to high credit risk and low credit risk respectively. Moreover, who owns house refers to have low credit risk, since 29 cases has been recorded as low and 27 has been recorded as high credit risk.

Total 28 cases of Unskilled workers has been recorded where 15 and 13 belongs to high and low credit risk.

Overall, in cluster 1 we can conclude those who owns house belongs to low credit risk. In cluster most of them belongs to high credit risk. Cluster 3 influences more towards high credit risk. However overall 214 cases observed as high credit risk and 211 files belongs to low credit risk.

Count of complete 3	Column Labels		
Row Labels	High	Low	Grand Total
1	**104**	**136**	**240**
Management	**17**	**20**	**37**
Other	3	5	8
Own	12	14	26
Rent	2	1	3
Skilled	**65**	**85**	**150**
Other	19	11	30
Own	38	64	102
Rent	8	10	18
Unemployed	**1**	**3**	**4**
Other		1	1
Own	1	2	3
Unskilled	**21**	**28**	**49**
Other	1	1	2
Own	16	23	39
Rent	4	4	8
2	**29**	**78**	**107**
Management	**4**	**6**	**10**
Other		1	1
Own	3	4	7
Rent	1	1	2
Skilled	**18**	**51**	**69**
Other		1	1
Own	14	41	55
Rent	4	9	13
Unemployed		**3**	**3**
Own		2	2
Rent		1	1
Unskilled	**7**	**18**	**25**
Other		1	1
Own	6	11	17
Rent	1	6	7
3	**78**		**78**
Management	**7**		**7**
Other	2		2
Own	5		5
Skilled	**52**		**52**
Other	2		2
Own	27		27
Rent	23		23
Unemployed	**4**		**4**
Other	1		1
Own	1		1
Rent	2		2
Unskilled	**15**		**15**
Other	3		3
Own	8		8
Rent	4		4
Grand Total	**211**	**214**	**425**

Figure 16

For Cluster 1 Total cases are 240, 104 cases classified as "High" credit risk and 136 cases classified as "Low" credit risk. In Management category, there are 37 cases in total, with 17 cases classified as "High" credit risk and 20 cases classified as "Low" credit risk. In skilled workers, 65 cases are classified as "High" credit risk, and the remaining 85 cases are classified as "Low" credit risk. In unemployment category only 4 of them has been recorded where 3 and has been recorded as low and high credit risk. Among the 49 unskilled people 28 and 21 are recorded as low and high credit risk. In cluster two 78 and 29 cases are seen as low And high credit risk respectively.

In Cluster 3, among 78 of the cases observed as all of them belongs to high credit risk.

In conclusion we can say that, in cluster 3 those who owns house and works as skilled worker are exposed to high credit risk.

AIM: The task of the protfolio is to conduct a conjoint analysis that will help to understand how consumers make choices and determine the relative significance of different product attributes. The goal is to analyse the data using linear regression to calculate utilities for each level factors, where utilities represent the relative importance of each level in influencing consumers preferences. Steps that were taken are described thoroughly with steps.

Features and Product Combination:

Price, Screen Size, Camera Quality and Battery life are taken. Different levels are as follows.

Prices are £800, 1000,1200 respectively, whereas Screen Sizes are 5 inches and 6 inches. Camera Qualities are 48MP and 64MP. Finally Battery life are 4000maH,5500maH, 6000maH. Therefore total 36 combinations have been made. These are the most important features that individuals look for in decision making while purchasing a new phone.

After creating the product combinations in the excel file now it is the turn to create dummy variables to run the linear regression. So for the regression model is well behaved and provides meaningful results the highest values are assigned to eliminate the multicollinearity as given below.

Price	1000	1200
800	0	0
1000	1	0
1200	0	1

Screensize	6
5	0
6	1

Camera Quallity	64
48	0
64	1

Battery	5500	6000
4000	0	0
5500	1	0
6000	0	1

Fig:1

After assigning the values of all this into dummy variables it becomes like attachment below.

	Product Combination	p1000	p1200	S6	CQ64	BL5500	BL600
36	36. $1200, 6 inches, 64 MP, 6000 mAh	0	1	1	1	0	1
35	35. $1200, 6 inches, 64 MP, 5500 mAh	0	1	1	1	1	0
34	34. $1200, 6 inches, 64 MP, 4000 mAh	0	1	1	1	0	0
33	33. $1200, 6 inches, 48 MP, 6000 mAh	0	1	1	0	0	1
32	32. $1200, 6 inches, 48 MP, 5500 mAh	0	1	1	0	1	0
31	31. $1200, 6 inches, 48 MP, 4000 mAh	0	1	1	0	0	0
30	30. $1200, 5 inches, 64 MP, 6000 mAh	0	1	0	1	0	1
29	29. $1200, 5 inches, 64 MP, 5500 mAh	0	1	0	1	1	0
28	28. $1200, 5 inches, 64 MP, 4000 mAh	0	1	0	1	0	0
27	27. $1200, 5 inches, 48 MP, 6000 mAh	0	1	0	0	0	1
26	26. $1200, 5 inches, 48 MP, 5500 mAh	0	1	0	0	1	0

Figure2

Making this into ranking 1-36 in google forms questionaries I have send the questionaries friends and families to hrank among the 36 product combinations and received results from 7 of my friends who are students. I have connected it with the excel file and take average data. And sort it from lowest to highest. And make a separate column named Average ranking and since SPSS does it in opposite way another column has been created for SPSS ranking as given below.

Product Combination	p1000	p1200	S6	CQ64	BL5500	BL600	Average Rank	SPSS RANK
2. $800, 5 inches, 48 MP, 5500 mAh	0	0	0	0	1	0	1	36
4. $800, 5 inches, 64 MP, 4000 mAh	0	0	0	1	0	0	2	35
1. $800, 5 inches, 48 MP, 4000 mAh	0	0	0	0	0	0	3	34
7. $800, 6 inches, 48 MP, 4000 mAh	0	0	1	0	0	0	4	33
11. $800, 6 inches, 64 MP, 5500 mAh	0	0	1	1	1	0	5	32
10. $800, 6 inches, 64 MP, 4000 mAh	0	0	1	1	0	0	6	31
9. $800, 6 inches, 48 MP, 6000 mAh	0	0	1	0	0	1	7	30
3. $800, 5 inches, 48 MP, 6000 mAh	0	0	0	0	0	1	8	29
5. $800, 5 inches, 64 MP, 5500 mAh	0	0	0	1	1	0	9	28
15. $1000, 5 inches, 48 MP, 6000 mAh	1	0	0	0	0	1	10	27
8. $800, 6 inches, 48 MP, 5500 mAh	0	0	1	0	1	0	11	26
21. $1000, 6 inches, 48 MP, 6000 mAh	1	0	1	0	0	1	12	25
6. $800, 5 inches, 64 MP, 6000 mAh	0	0	0	1	0	1	13	24
14. $1000, 5 inches, 48 MP, 5500 mAh	1	0	0	0	1	0	14	23
19. $1000, 6 inches, 48 MP, 4000 mAh	1	0	1	0	0	0	15	22
13. $1000, 5 inches, 48 MP, 4000 mAh	1	0	0	0	0	0	16	21
16. $1000, 5 inches, 64 MP, 4000 mAh	1	0	0	1	0	0	17	20

Figure 3

Now the data is ready to run for regression, where SPSS rank is the dependant variable and Products dummy variables on the independent variables.

Linear Regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.932[a]	.869	.842	4.19313

a. Predictors: (Constant), BL600, CQ64, S6, p1200, p1000, BL5500

Figure 4

Putting the data in the spss and running the linear regression it has been found that the model is fit that gives the output of R value of 0.932, indicating that around 93.2% of the outcome variable's variability is explained by the predictors. R square value is 0.869, meaning that approximately 86.9% of the variability can be accounted. R square suggest 86.9% of the variability can be accounted, as model summary given on figure 4.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	36.111	1.849		19.530	<.001
	p1000	-10.833	1.712	-.492	-6.328	<.001
	p1200	-21.917	1.712	-.995	-12.803	<.001
	S6	-5.000	1.398	-.241	-3.577	.001
	CQ64	-4.556	1.398	-.219	-3.259	.003
	BL5500	-2.083	1.712	-.095	-1.217	.233
	BL600	-3.667	1.712	-.166	-2.142	.041

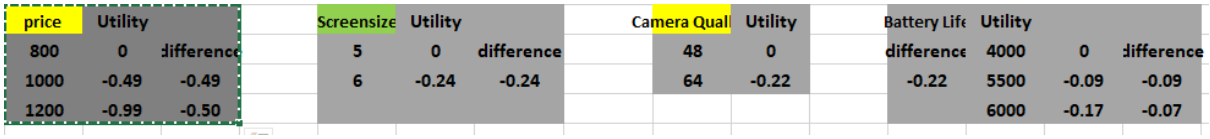
a. Dependent Variable: spss rank

Figure 5

The standardized coefficients (Beta) represents the relative importance of each predictor in explaining the dependent variable. Here the negative standardized

coefficients suggest that higher values of p1000, p1200, S6, and CQ64 are associated with lower values of the dependent variable (spss rank) indicating a less favorable outcome.

Calculating Utility Differences



price	Utility	
800	0	difference
1000	-0.49	-0.49
1200	-0.99	-0.50

Screensize	Utility	
5	0	difference
6	-0.24	-0.24

Camera Quali	Utility
48	0
64	-0.22

Battery Life	Utility		
difference	4000	0	difference
-0.22	5500	-0.09	-0.09
	6000	-0.17	-0.07

Figure 6

As shown in figure 6, putting the beta values to calculate the difference in utility for the four features defines that as the prices increases customers perceive the product to be less valuable. This goes same for screen sizes as well as it increases customer thinks its not easy to carry hench diminishing the preference. regarding camera quality, there is a diminishing utility, indicating that customers may not significantly prioritize small improvements in camera quality. Lastly, as battery life extends, there is a diminishing utility, suggesting that customers may not place as much value on additional battery life beyond a certain point.

Calculating the Summation of utility.

Product Combination	p1000	p1200	S6	CQ64	BL5500	BL600	Summation	Utility Rank
1. $800, 5 inches, 48 M	0	0	0	0	0	0	0	1
2. $800, 5 inches, 48 M	0	0	0	0	-0.09454	0	=SUM(B3:G3)	2
3. $800, 5 inches, 48 M	0	0	0	0	0	-0.16639	SUM(number1, [number2], ...)	
4. $800, 5 inches, 64 M	0	0	0	-0.21926	0	0	-0.219263878	4
7. $800, 6 inches, 48 M	0	0	-0.24066	0	0	0	-0.240655476	5
5. $800, 5 inches, 64 M	0	0	0	-0.21926	-0.09454	0	-0.313802277	6
8. $800, 6 inches, 48 M	0	0	-0.24066	0	-0.09454	0	-0.335193875	7
6. $800, 5 inches, 64 M	0	0	0	-0.21926	0	-0.16639	-0.38565146	8
9. $800, 6 inches, 48 M	0	0	-0.24066	0	0	-0.16639	-0.407043058	9
10. $800, 6 inches, 64 M	0	0	-0.24066	-0.21926	0	0	-0.459919353	10
13. $1000, 5 inches, 48	-0.4916	0	0	0	0	0	-0.491599676	11
11. $800, 6 inches, 64 M	0	0	-0.24066	-0.21926	-0.09454	0	-0.554457753	12
14. $1000, 5 inches, 48	-0.4916	0	0	0	-0.09454	0	-0.586138076	13
12. $800, 6 inches, 64 M	0	0	-0.24066	-0.21926	0	-0.16639	-0.626306936	14
15. $1000, 5 inches, 48	-0.4916	0	0	0	0	-0.16639	-0.657987259	15
16. $1000, 5 inches, 64	-0.4916	0	0	-0.21926	0	0	-0.710863554	16

Figure 7

Assigning the values of Standardized coefficient Beta into the values of 1 as shown in the table above the sum of utility is carried out and sorted out from smallest to largest where utility rank is assigned from 1-36 as the sums value increases in a separate column.

Calculating Correlation Coefficient of Utility Ranking and Average Ranking

Correlations

		Average Rank	Utility Rank
Average Rank	Pearson Correlation	1	.932**
	Sig. (2-tailed)		<.001
	N	36	36
Utility Rank	Pearson Correlation	.932**	1
	Sig. (2-tailed)	<.001	
	N	36	36

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

Figure 8

Putting down the values of utility ranking and average ranking is the spss, correlation analyse has been implemented. It tells that average rank fits 93% to the utility rank. The person correlation suggest that there is a strong positive relationship between the two variables. Overall, this indicates that higher-ranked items generally offer greater utility, and lower-ranked items have lower utility.

Portfolio-4

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

Yes the given data have a variation is trend which is positive and increases as the numbers of years has been increased with swift around trend that is seen in the figure given below for the 11 years. There are spikes are observed in certain intervals so it can be concluded that the data have seasonal component. As a reference a picture is given below.

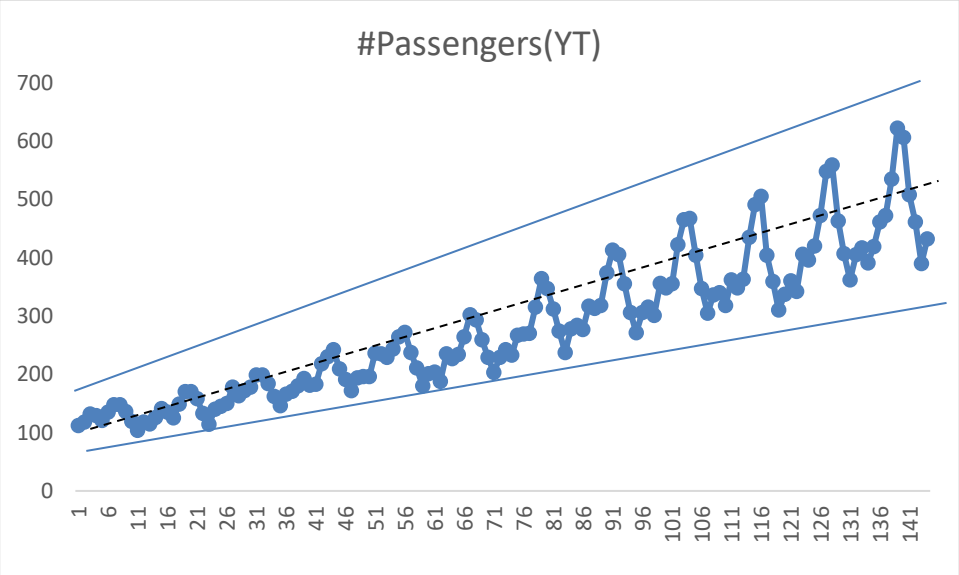


Figure 1

- **How many seasons can be recognised in this data set?**

  In the given data set 12 peaks can be recognized, hence it can be concluded that it consists of 12 seasons.

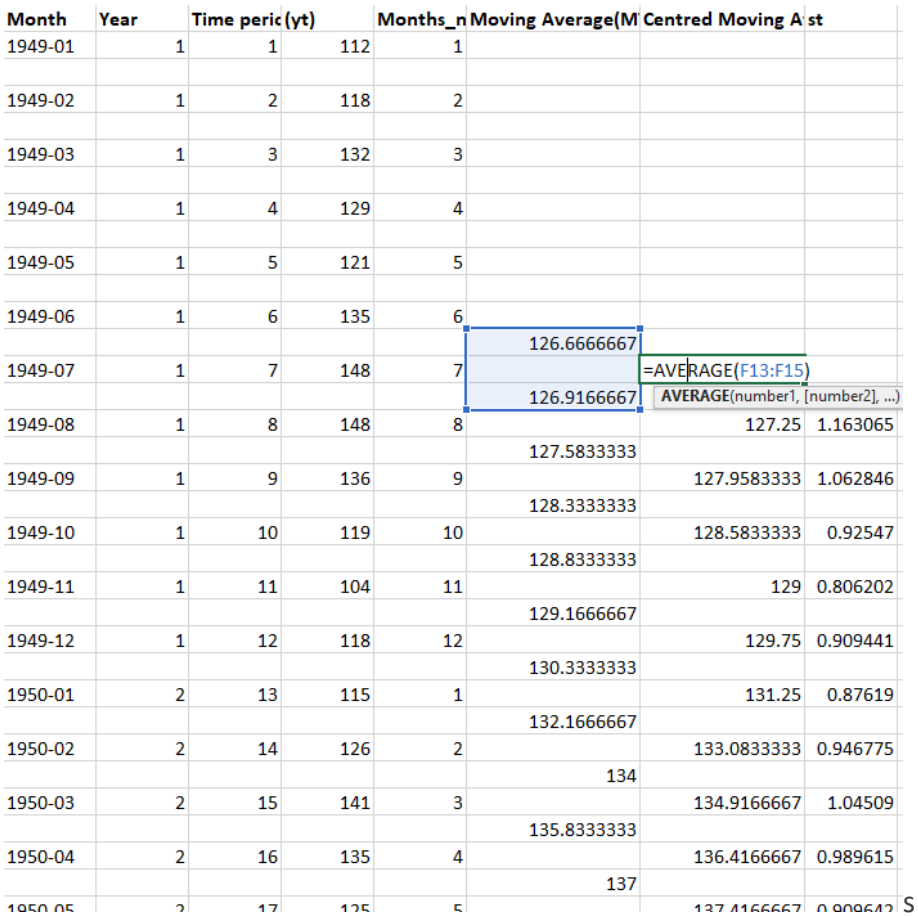
- **Calculate appropriate moving averages for this data set to smooth out the trend. Then calculate the seasonal components values. Provide an interpretation for the seasonal factor values.**

Calculating moving average is just to average the number of passengers in year 1and continue it till cell (F253) since we have to work with the data 11 years and forecast for next 6 months. As the reference given below.

Month	Year	Time peric	(yt)	Months_n	Moving Average(M	cmat	st
1949-01	1	1	112	1			
1949-02	1	2	118	2			
1949-03	1	3	132	3			
1949-04	1	4	129	4			
1949-05	1	5	121	5			
1949-06	1	6	135	6			
					=AVERAGE(D2:D24)		
1949-07	1	7	148	7	AVERAGE(**number1**, [number2], ...)		1.167269
					126.9166667		
1949-08	1	8	148	8		127.25	1.163065
					127.5833333		
1949-09	1	9	136	9		127.9583	1.062846
					128.3333333		
1949-10	1	10	119	10		128.5833	0.92547
					128.8333333		
1949-11	1	11	104	11		129	0.806202
					129.1666667		
1949-12	1	12	118	12		129.75	0.909441
					130.3333333		
1950-01	2	13	115	1		131.25	0.87619
					132.1666667		
1950-02	2	14	126	2		133.0833	0.946775
					134		
1950-03	2	15	141	3		134.9167	1.04509
					135.8333333		

Figure 2

From Moving average we can calculate the centred moving average as given below. So that the seasonal components can be calculated.



Month	Year	Time peric	(yt)	Months_n	Moving Average(M	Centred Moving A	st
1949-01	1	1	112	1			
1949-02	1	2	118	2			
1949-03	1	3	132	3			
1949-04	1	4	129	4			
1949-05	1	5	121	5			
1949-06	1	6	135	6			
					126.6666667		
1949-07	1	7	148	7		=AVERAGE(F13:F15)	
					126.9166667	AVERAGE(number1, [number2], ...)	
1949-08	1	8	148	8		127.25	1.163065
					127.5833333		
1949-09	1	9	136	9		127.9583333	1.062846
					128.3333333		
1949-10	1	10	119	10		128.5833333	0.92547
					128.8333333		
1949-11	1	11	104	11		129	0.806202
					129.1666667		
1949-12	1	12	118	12		129.75	0.909441
					130.3333333		
1950-01	2	13	115	1		131.25	0.87619
					132.1666667		
1950-02	2	14	126	2		133.0833333	0.946775
					134		
1950-03	2	15	141	3		134.9166667	1.04509
					135.8333333		
1950-04	2	16	135	4		136.4166667	0.989615
					137		
1950-05	2	17	125	5		137.4166667	0.909642

S

Figure 3

Seasonal component is calculated by **S(t)=** Centred moving average/ Passenger (YT)

Month	Year	Time perio	(yt)	Months_n	Moving Average(M	Centred Moving A	st
1949-01	1	1	112	1			
1949-02	1	2	118	2			
1949-03	1	3	132	3			
1949-04	1	4	129	4			
1949-05	1	5	121	5			
1949-06	1	6	135	6			
					126.6666667		
1949-07	1	7	148	7		126.7916667	=D14/G14
					126.9166667		
1949-08	1	8	148	8		127.25	1.163065
					127.5833333		
1949-09	1	9	136	9		127.9583333	1.062846
					128.3333333		
1949-10	1	10	119	10		128.5833333	0.92547
					128.8333333		
1949-11	1	11	104	11		129	0.806202
					129.1666667		
1949-12	1	12	118	12		129.75	0.909441
					130.3333333		
1950-01	2	13	115	1		131.25	0.87619
					132.1666667		
1950-02	2	14	126	2		133.0833333	0.946775
					134		
1950-03	2	15	141	3		134.9166667	1.04509
					135.8333333		
1950-04	2	16	135	4		136.4166667	0.989615
					137		
1950-05	2	17	125	5		137.4166667	0.909642
					137.8333333		

Figure 4

After the trend smoothed out the data is now ready for decomposition.

Month	Year1	year 2	Year3	Year4	Year5	Year6	Year7	Year8	Year9	Year10	Year11	Mean	typical
1		0.87619	0.922832	0.933788	0.908108	0.894737	0.924252	0.916252	0.904523	0.906063	0.894317	0.908106	0.91000371
2		0.946775	0.940193	0.966659	0.897025	0.815766	0.87375	0.880997	0.852691	0.841455	0.839951	0.885526	0.88737650
3		1.04509	1.099897	1.020939	1.068276	1.011841	0.984786	0.9949	0.995456	0.953887	0.985736	1.016081	1.01820370
4		0.989615	0.993145	0.946199	1.054206	0.970431	0.977441	0.972805	0.962989	0.915789	0.951161	0.973378	0.97541198
5		0.909642	1.032	0.945329	1.021941	0.993103	0.969479	0.979969	0.973937	0.953486	0.998811	0.97777	0.97981283
6		1.073874	1.052735	1.113191	1.081402	1.11041	1.117186	1.143439	1.149342	1.141857	1.109283	1.109272	1.11158981
7	1.167269	1.206387	1.162044	1.161372	1.171598	1.255717	1.273841	1.253256	1.258599	1.285901		1.219598	1.22214663
8	1.163065	1.187427	1.146423	1.211514	1.207101	1.201025	1.199309	1.220492	1.258054	1.316247		1.211066	1.21359610
9	1.062846	1.084358	1.048682	1.033587	1.053528	1.047876	1.063939	1.061418	1.085535	1.045278		1.058705	1.06091684
10	0.92547	0.896126	0.916117	0.926061	0.939518	0.915085	0.922042	0.906555	0.931752	0.919727		0.919845	0.92176703
11	0.806202	0.752268	0.820033	0.817426	0.801931	0.800789	0.787375	0.795791	0.818243	0.78539		0.798545	0.80021323
12	0.909441	0.904929	0.921369	0.909197	0.891188	0.890617	0.910108	0.889319	0.899297	0.845406		0.897087	0.89896164
											Total	11.97498	12
										Correction	Factor	1.002089	

Figure 5

The Figure displays the seasonal factors for every month in all year, indicating the relative impact of seasonality on the number of passengers in a multiplicative model.

First, average for each month across all seasons is calculated and anticipated that the average totals to 12 as it is expected that the sum of the seasonal factor should be equal to the number of seasons. The average is manipulated in order to get the sum up to 12 by multiplying the correction factor with each mean seasonal factor. Correction Factor is calculated by dividing 12 by the total of all seasonal factors.

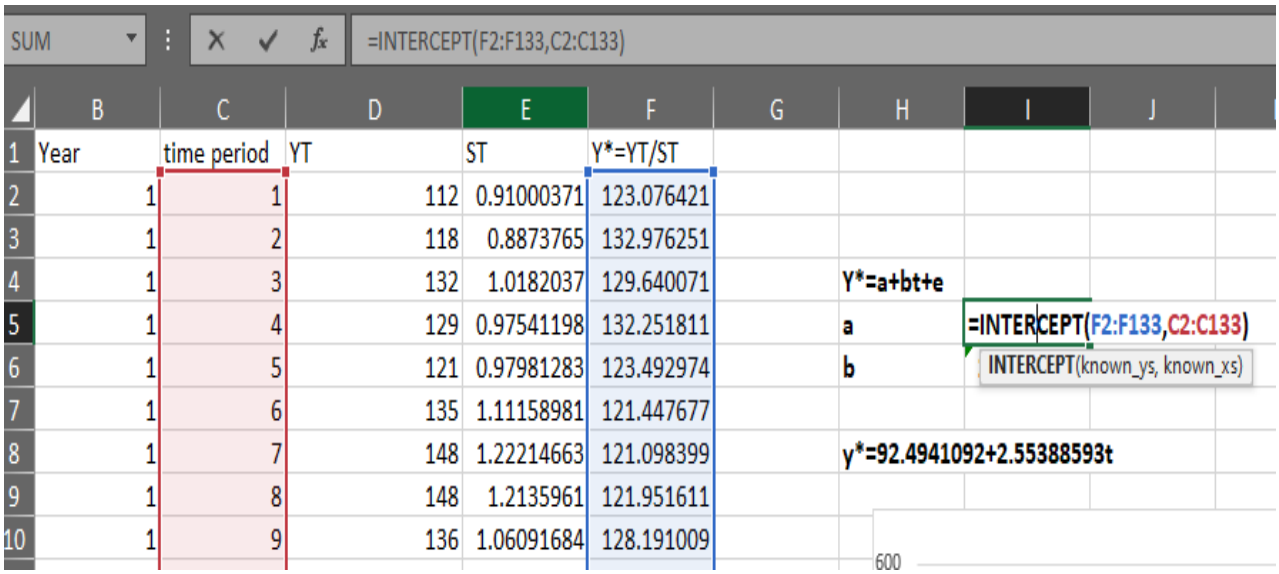
From the Typical Values we can say that the value has increased from to 1.22 in July that is it has increased by 22% and decreased by 20% in November. Like this we can compute the seasonal changes for each month.

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

  Multiplicative model, as discussed it is suitable to be fitted into data when the variations show a proportional shift around the trendline.

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

  To forecast the data for 1960, the slope and intercept has to be executed from the above data and that intends to become like y*=a+bt+e where a is the intercept and b is the slope. The figure is given below.



Intercept is 92.494 whereas the slope is 2.55. t is the time period .For the forecasted value, y*=92.49+2.55t is assigned to excel and the typical value are assigned to de-seasonalize the data and forecast smoothly. As the figure shown below.

1960-01	12	133	=F134*E134	0.91000371	432.1609
1960-02	12	134	386	0.8873765	434.7148
1960-03	12	135	445	1.0182037	437.2687
1960-04	12	136	429	0.97541198	439.8226
1960-05	12	137	433	0.97981283	442.3765
1960-06	12	138	495	1.11158981	444.9304
1960-07	12	139	547	1.22214663	447.4843
1960-08	12	140	546	1.2135961	450.0381
1960-09	12	141	480	1.06091684	452.5920
1960-10	12	142	420	0.92176703	455.1459
1960-11	12	143	366	0.80021323	457.6998
1960-12	12	144	414	0.89896164	460.2537

Column1	Column2	Column3
Forecasted Values		Actual Values
393		417
386		391
445		419
429		461
433		472
495		535
547		622
546		606
480		508
420		461
366		390
414		432

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

Calculate Mean Absolute Error (MAE) Mean Square Error (MSE):

To calculate the Mean Absolute value and mean square error for your forecasts. We have to find the difference between Actual and forecasted value. The negative value of the difference would the mitigated by taking absolute function in the excel and follow towards the formula.

Mean Absolute Error (MAE):

MAE = (1/n) * Σ|Actual - Forecast|

Mean Square Error (MSE):

MSE = (1/n) * Σ(Actual - Forecast)^2

Here,

n represents the number of data points (forecasts) being evaluated.

Now Calculating the data we get.

Error	Absolute Error	Error Square	Mae	MSE
24	24	563	**34**	**1503**
5	5	28		
-26	26	688		
32	32	1023		
39	39	1486		
40	40	1634		
75	75	5641		
60	60	3580		
28	28	775		
41	41	1719		
24	24	564		
18	18	333		

Excel File:

Fibal.xlsx

Portfolio 5

DATA Objective:

The aim of the task is to analyze and summarize the provided dataset, which includes the number of COVID-19 cases and deaths in the UK daily from January 1, 2020, to June 14, 2020.

Analysis:

1- Determining data:

There was an outlier value -525 in the data, we converted this value by taking the mean of above and lower 4 values **2921.625**

3450
3534
2711
2412
-525
2615
3287
2959
2405
2921.625

Figure 1

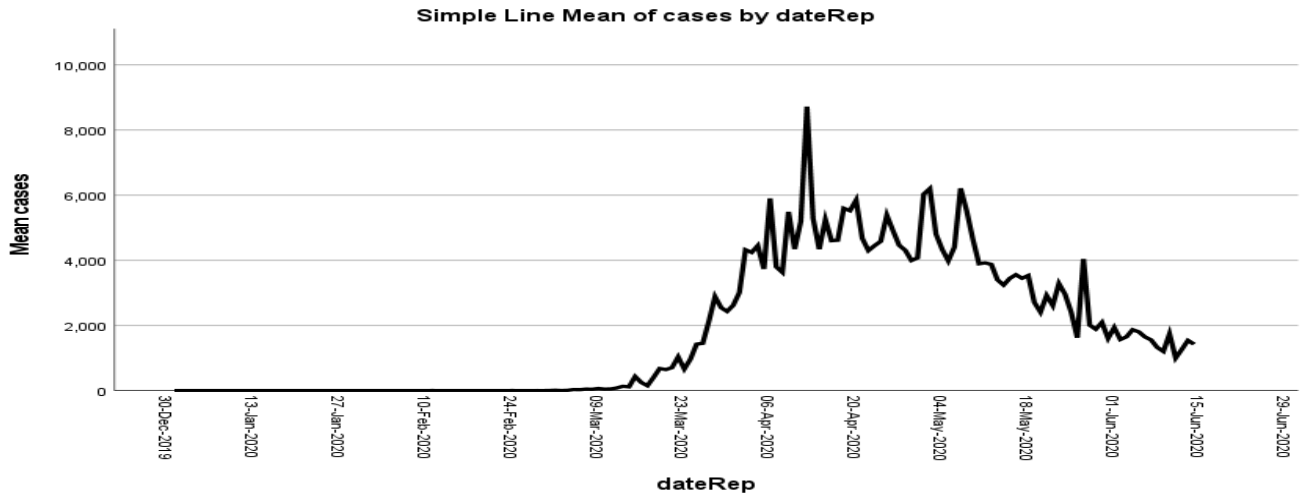


Figure 2

- 1st checked the data my making a simple line graph found that there is many unnecessary data given, therefore decided to reduce the number of data cases.
- Now the data is taken from cases 93 till case 174.

2- Stationary vs non-stationary data:

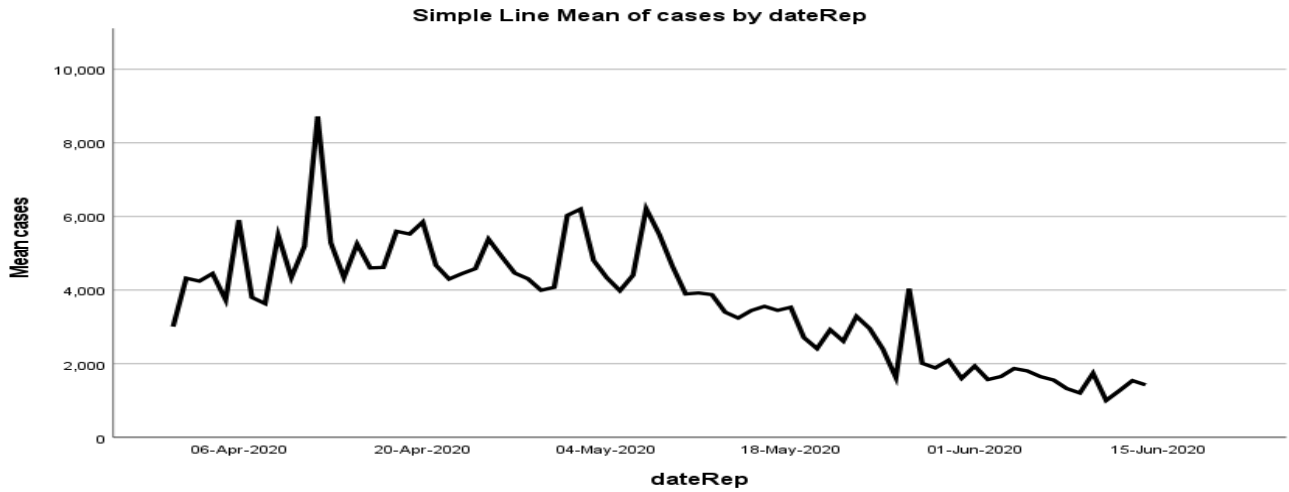


Figure 3

- There is a declining trend can be seen in the graph, telling that this data is non-stationary.

Case Processing Summary

		cases
Series Length		82
Number of Missing Values	User-Missing	0
	System-Missing	7[a]
Number of Valid Values		75
Number of Computable First Lags		74

a. Some of the missing values are imbedded within the series.

Figure4

- There are 82 no. of cases, 7 missing values, and 75 valid values.

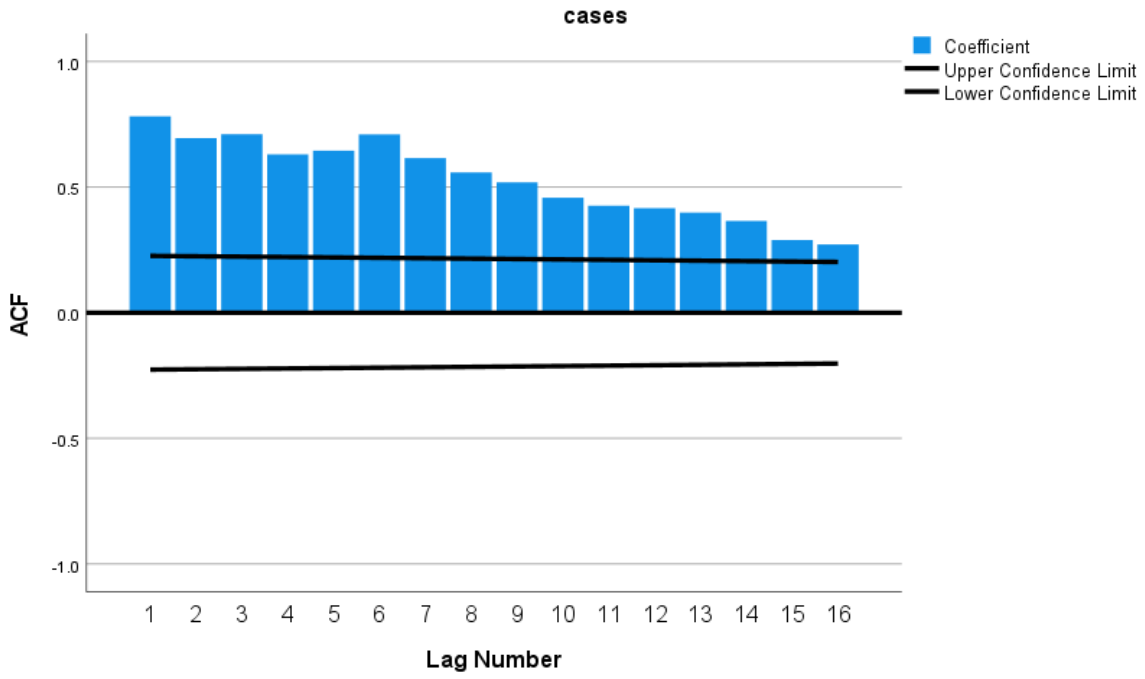


Figure 5

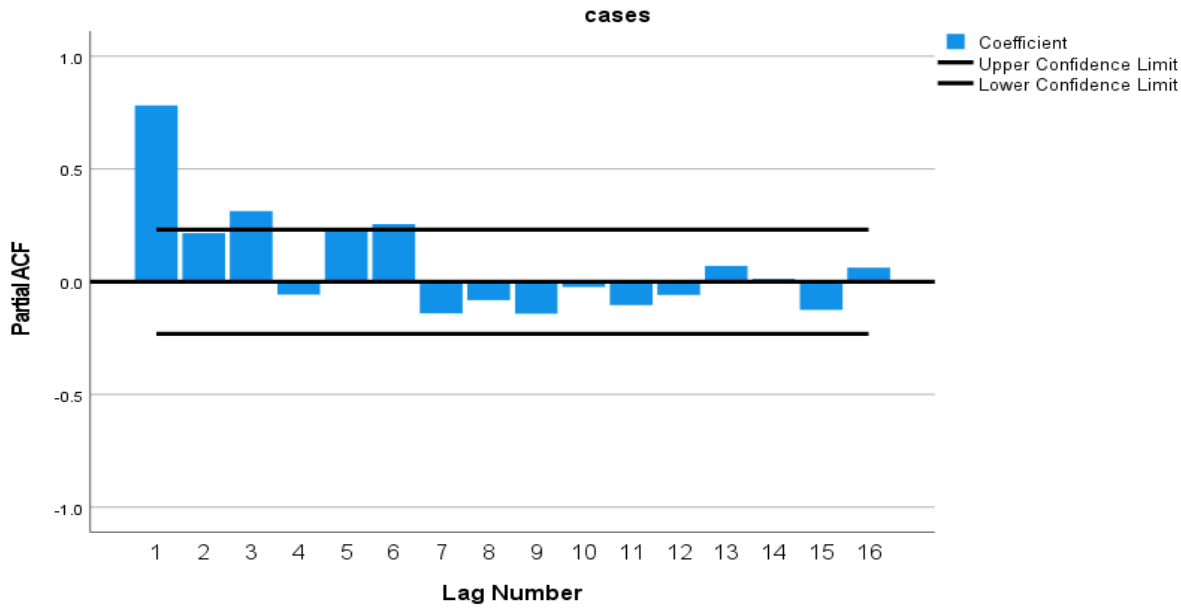


Figure 6

In the ACF graph, decline trend is stating that data in non-stationary.

Differencing:

Now, differencing is required in order to reduce the no. of lags.

We do difference generally by 1.

Case Processing Summary

		cases
Series Length		82
Number of Missing Values	User-Missing	0
	System-Missing	7[a]
Number of Valid Values		75
Number of Values Lost Due to Differencing		1
Number of Computable First Lags After Differencing		73

a. Some of the missing values are imbedded within the series.

Figure7

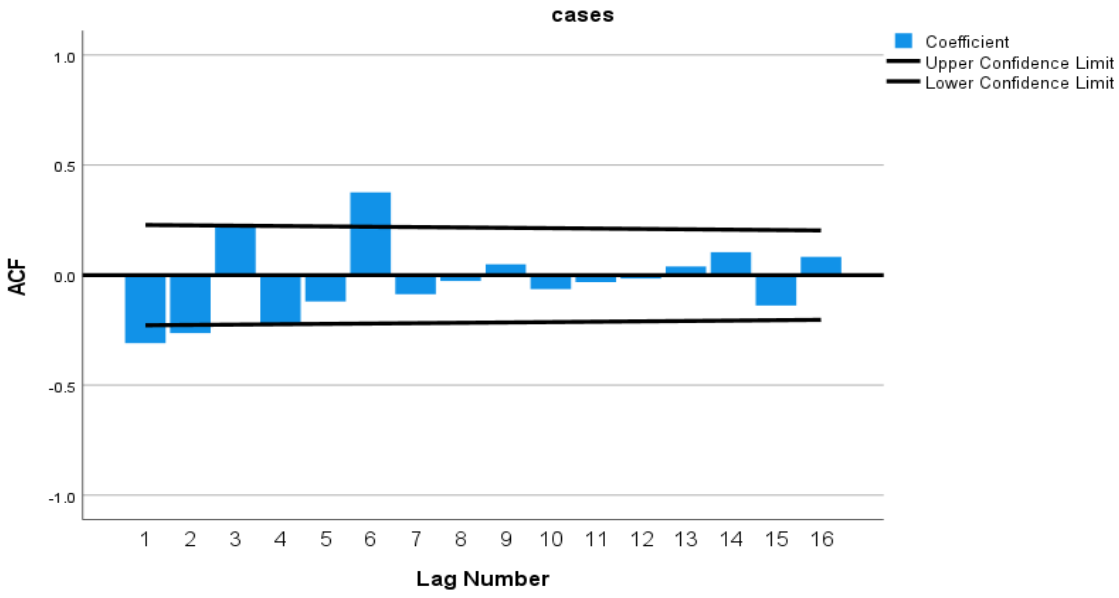


Figure 8

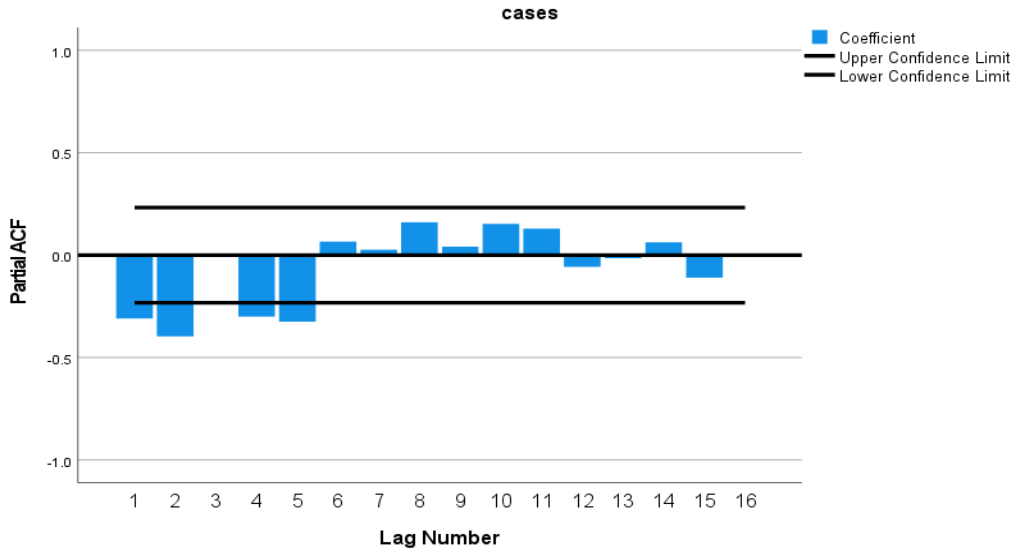


Figure 9

counting lags in both ACF and PACF

ACF = 6

Partial ACF = 5

Step 6: Creating ARIMA(5, 1, 6) Model

For ARIMA model, we need to determine the lags *p* and *q* in the ARIMA (*p, d, q*) model.

Whereas,

q = ACF = moving average

d = how many times the data needs to be differenced to produce a stationary series

p = Partial ACF = autoregressive model

By Differencing we got the following values:

q = 6

d = 1

p = 5

Model Description

			Model Type
Model ID	cases	Model_1	ARIMA(5,1,6)

Model Statistics

Model	Number of Predictors	Model Fit statistics: Stationary R-squared	RMSE	MAE	Ljung-Box Q(18): Statistics	DF	Sig.	Number of Outliers
cases-Model_1	0	.417	821.831	576.655	1.993	7	.960	0

Ljung-BoxQ values should be > 0.05 of an adequate model. Ljung-Box Q value is **0.960.** There fore proved, our ARIMA(5,1,6) Model is adequate.

- RMSE = 821.831
- MAE = 576.655

ARIMA Model Parameters

					Estimate	SE	t	Sig.
cases-Model_1	cases	No Transformation	AR	Lag 1	-.345	.388	-.889	.377
				Lag 2	-.260	.257	-1.014	.315
				Lag 3	-.045	.250	-.179	.858
				Lag 4	-.289	.243	-1.191	.238
				Lag 5	-.137	.309	-.443	.659
			Difference		1			
			MA	Lag 1	.245	.382	.643	.523
				Lag 2	.213	.354	.602	.550
				Lag 3	.022	.269	.082	.935
				Lag 4	-.175	.273	-.640	.524
				Lag 5	-.017	.319	-.055	.957
				Lag 6	-.356	.207	-1.722	.090

Significant lag < 0.05

Insignificant lag > 0.05

The lags in AR are insignificant as they are > 0.05.

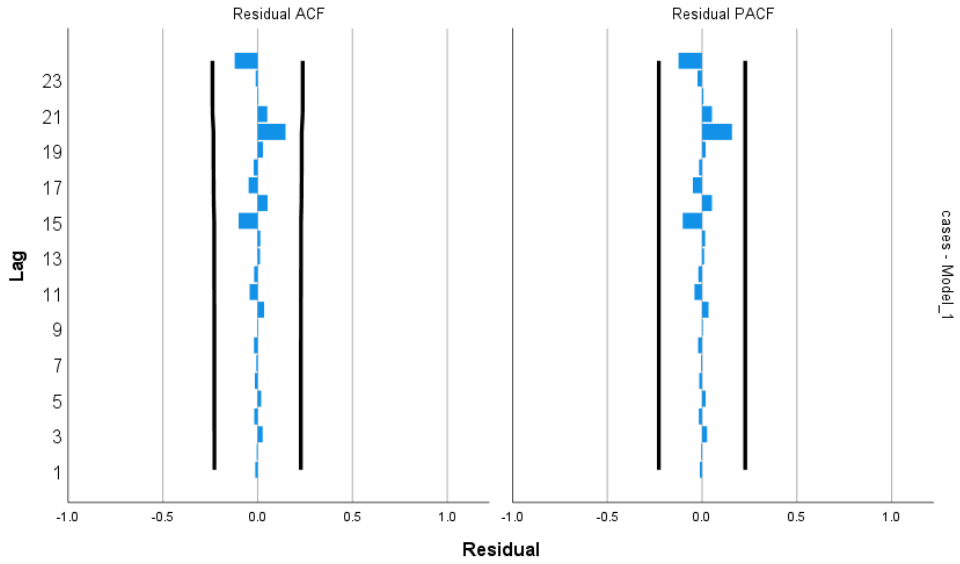
In MA also all lags are insignificant. hence we need to remove these lags and go for a small model.

Forecast

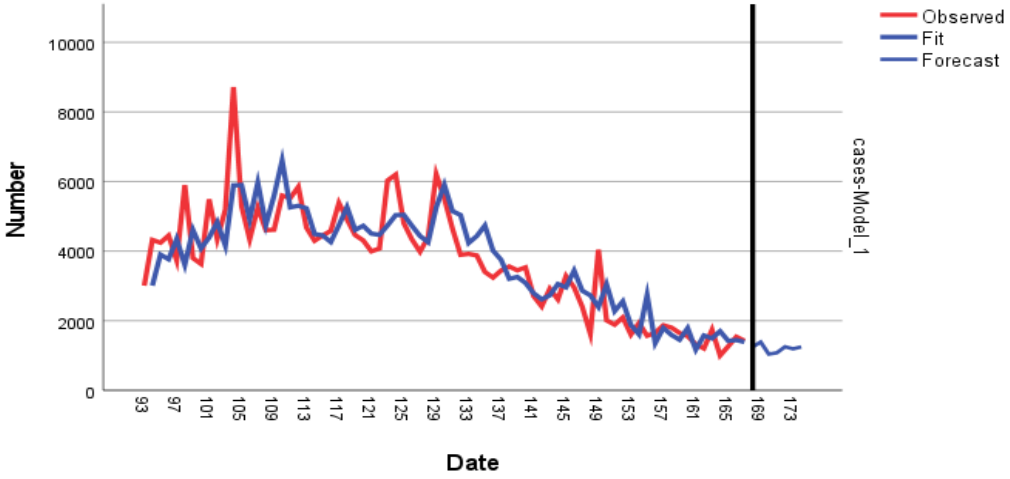
Model		168	169	170	171	172	173	174
cases-Model_1	Forecast	1246	1390	1041	1080	1253	1194	1249
	UCL	2830	3101	2767	2878	3091	3087	3523
	LCL	-337	-321	-684	-718	-585	-698	-1025

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.

The above forecast table shows the forecasted values of given 7 days in the question from 168 to 174 cases.



According to Residual ACF and Residual PACF graph, all lags are within the significant interval hence, this model predicting their time series behaviour well.



Creating small ARIMA(5, 1, 5) Model. For creating a smaller and better model we need to reduce the no. of lags which values are < 0.05

Model Description

			Model Type
Model ID	cases	Model_1	ARIMA(5,1,5)

Figure 15

Model Fit

					Percentile						
Fit Statistic	Mean	SE	Minimum	Maximum	5	10	25	50	75	90	95
Stationary R-squared	.408	.	.408	.408	.408	.408	.408	.408	.408	.408	.408
R-squared	.759	.	.759	.759	.759	.759	.759	.759	.759	.759	.759
RMSE	821.903	.	821.903	821.903	821.903	821.903	821.903	821.903	821.903	821.903	821.903
MAPE	16.447	.	16.447	16.447	16.447	16.447	16.447	16.447	16.447	16.447	16.447
MaxAPE	77.048	.	77.048	77.048	77.048	77.048	77.048	77.048	77.048	77.048	77.048
MAE	566.097	.	566.097	566.097	566.097	566.097	566.097	566.097	566.097	566.097	566.097
MaxAE	2836.208	.	2836.208	2836.208	2836.208	2836.208	2836.208	2836.208	2836.208	2836.208	2836.208
Normalized BIC	14.005	.	14.005	14.005	14.005	14.005	14.005	14.005	14.005	14.005	14.005

Figure 16

Model Statistics

Model	Number of Predictors	Model Fit statistics: Stationary R-squared	Model Fit statistics: RMSE	Model Fit statistics: MAE	Ljung-Box Q(18): Statistics	Ljung-Box Q(18): DF	Ljung-Box Q(18): Sig.	Number of Outliers
cases-Model_1	0	.408	821.903	566.097	4.434	8	.816	0

Figure 17

- RMSE = 821.903
- MAE = 566.097
- Ljung-Box = 0.816

ARIMA Model Parameters

					Estimate	SE	t	Sig.
cases-Model_1	cases	No Transformation	AR	Lag 1	-.679	.269	-2.523	.014
				Lag 2	-.054	.270	-.201	.841
				Lag 3	-.083	.240	-.348	.729
				Lag 4	-.359	.201	-1.785	.079
				Lag 5	-.526	.146	-3.597	<.001
			Difference		1			
			MA	Lag 1	-.083	.307	-.271	.787
				Lag 2	.663	.236	2.814	.006
				Lag 3	-.049	.365	-.135	.893
				Lag 4	-.386	.217	-1.779	.080
				Lag 5	-.283	.270	-1.049	.298

Figure 18

Values >0.05 are lags in our model. We reduce them but also keep an eye on Ljung box value, if that decreases, we choose the previous model.

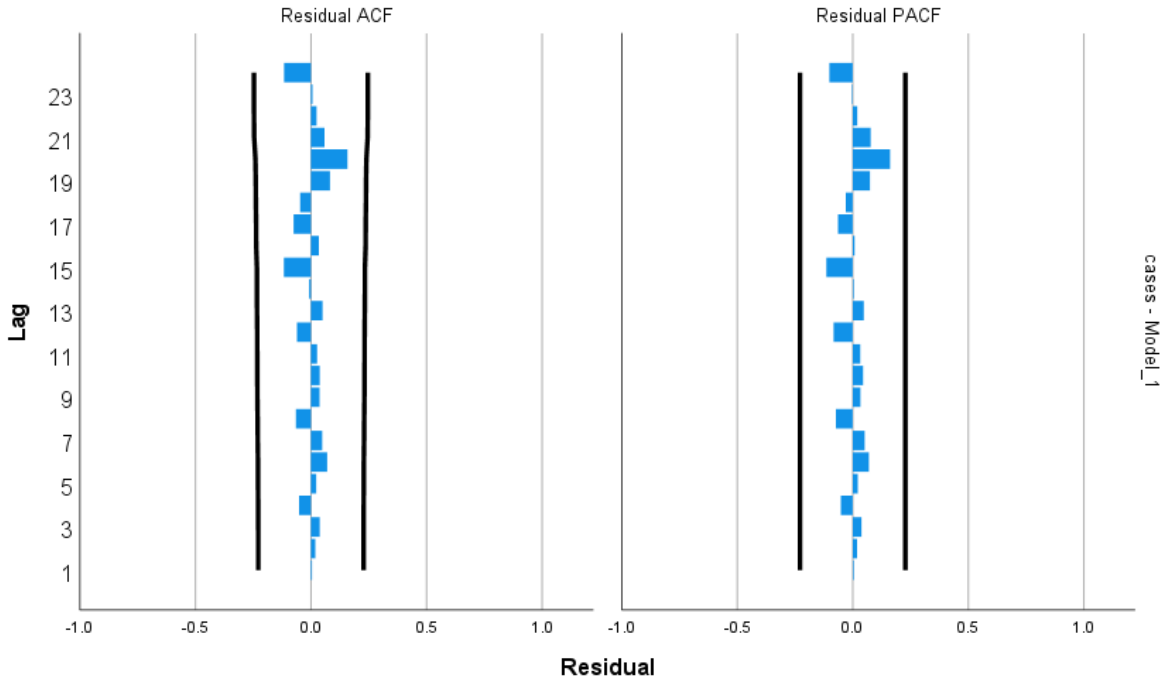


Figure 19

All lags are under the significant interval, which is a good behaviour.

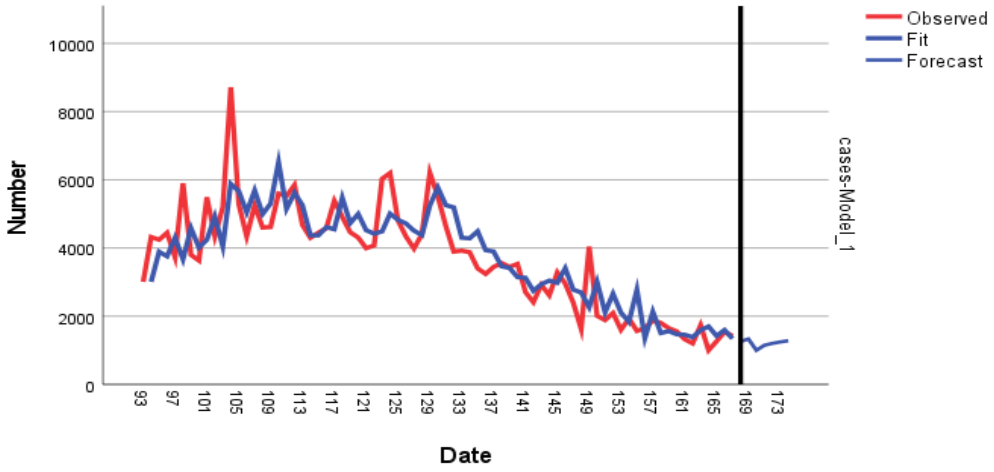


Figure 20

Create a smaller ARIMA(4, 1, 5)

Model Description

			Model Type
Model ID	cases	Model_1	ARIMA(4,1,5)

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	.388	828.859	563.083	14.468	9	.107	0

Figure 21

- RMSE = 828.859
- MAE = 563.083
- Ljung-Box Q = 0.107

It can be seen that the Ljung-Box value has decreased; hence we will go with the previous model.

Portfolio 6 report

Aim

The aim of the portfolio is to forecast the exchange rate for August 8, 2020, using artificial neural network in IBM SPSS.

Procedure:

- Analysing the data for US US/UK currency from January 4, 2010, to August 7, 2020, using graphs.
- Selecting appropriate inputs and outputs for the neural network model and justifying the choices based on analysis.
- Training a suitable neural network model and reporting tables and diagrams generated
- Interpreting the results of the model
- Providing a table of preferences for the input values and interpreting it.
- Using plots where necessary to support data analysis.
- Reporting the one-step-ahead forecast, i.e., the exchange rate for August 8, 2020.

Analyse

Figure 1

Plotting the graph of Exchange rate vs observation date, highest peak is noticed around 11th Aug 2014. Before this an average trend was shown , and then a decreasing trend is shown till April 2019. The lowest peak is noticed in 11-May-2020.

In SPSS ACF has been run where a model summary has been executed as follows.

Model Description

Model Name		MOD_2
Series Name	1	DEXUSUK
Transformation		None
Non-Seasonal Differencing		0
Seasonal Differencing		0
Length of Seasonal Period		No periodicity
Maximum Number of Lags		16
Process Assumed for Calculating the Standard Errors of the Autocorrelations		Independence(white noise)[a]
Display and Plot		All lags

Applying the model specifications from MOD_2

a. Not applicable for calculating the standard errors of the partial autocorrelations.

Figure 2

From the above figure we can say that It does not involve any transformations or differencing, even thers no seasonal pattern noticed. The model consists of 16 lags where a decreasing trend line can be observered in the figure 5. he standard errors of the autocorrelations are assumed to be calculated under the assumption of independence (white noise).

Case Processing Summary

		DEXUSUK
Series Length		2766
Number of Missing Values	User-Missing	0
	System-Missing	111[a]
Number of Valid Values		2655
Number of Computable First Lags		2546

a. Some of the missing values are imbedded within the series.

Figure 3

The series has a total length of 2766 data points. Out of these, there are no missing values However there are 111 missing values marked system-missing. This means that these values are missing due to some system-related issue or data collection problem. Excluding the missing values, there are still 2655 valid values available for analysis. It's important to note that some of the missing values are embedded within the series, the summary mentions that there are 2546 computable first lags.

Autocorrelations

Series: DEXUSUK

Lag	Autocorrelation	Std. Error[a]	Box-Ljung Statistic Value	df	Sig.[b]
1	.956	.019	2534.158	1	.000
2	.954	.019	5059.505	2	.000
3	.952	.019	7576.501	3	.000
4	.951	.019	10087.456	4	.000
5	.953	.019	12598.925	5	.000
6	.948	.019	15096.104	6	.000
7	.946	.019	17581.177	7	.000
8	.945	.019	20060.316	8	.000
9	.944	.019	22533.628	9	.000
10	.941	.019	24993.521	10	.000
11	.938	.019	27440.985	11	.000
12	.937	.019	29881.275	12	.000
13	.937	.019	32321.084	13	.000
14	.936	.019	34755.842	14	.000
15	.933	.019	37177.602	15	.000
16	.930	.019	39584.708	16	.000

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

Figure 4

In this summary, the autocorrelation values range from 0.930 to 0.956 for lags 1 to 16The **null hypothesis** indicated that there is no autocorrelation in the series is tested by the **Ljung-Box** statistic. According to this presumption, the data points are independent and are not affected by any regular patterns or trends.

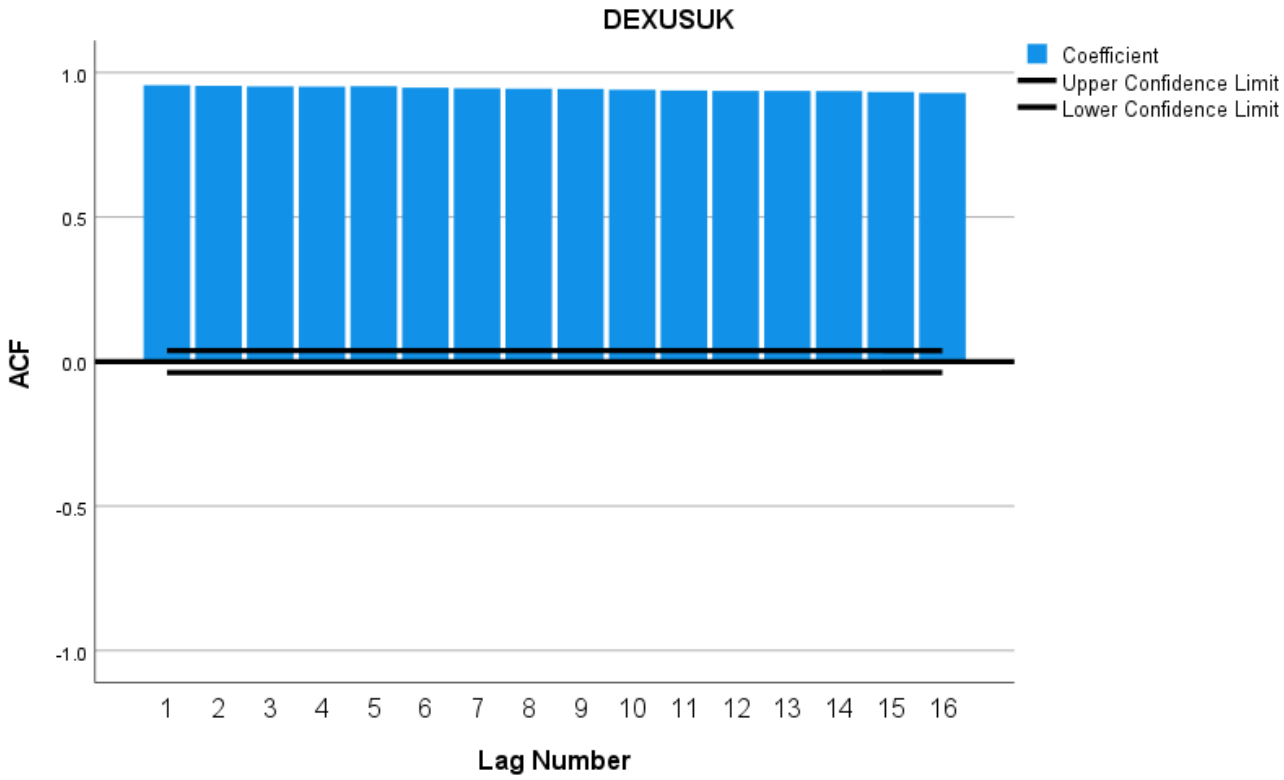


Figure 5

Figure 5 shows all the spikes are significant in this ACF as it crosses the upper confidence limit.

Partial Autocorrelations

Series: DEXUSUK

Lag	Partial Autocorrelation	Std. Error
1	.956	.019
2	.465	.019
3	.290	.019
4	.213	.019
5	.207	.019
6	.083	.019
7	.047	.019
8	.055	.019
9	.057	.019
10	.012	.019
11	.002	.019
12	.009	.019
13	.043	.019
14	.034	.019
15	.003	.019
16	-.021	.019

Figure 6

The partial autocorrelation is strongest at lag 1 (0.956) and gradually decreases as the lag increases. lag 16, there is a slight negative partial autocorrelation (-0.021), indicating a weak inverse relationship.

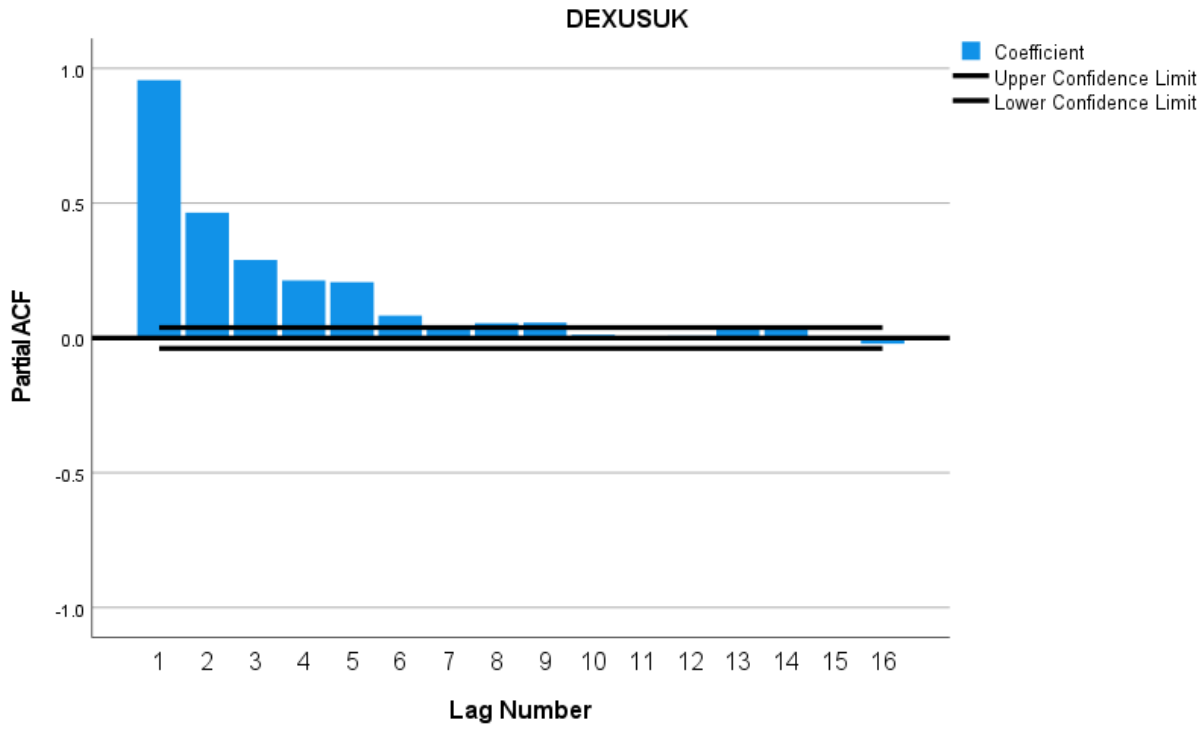


Figure 7

The diagram shows the autoregressive model of order 9 :[9,0,0]

In this partial ACF, we have found 9 significant lags which is crossing the upper confidence limit' hence these 2 two plots confirms that they are suitable for ARIMA model to fit in this analysis.

Neural Network Analysis

Now we will go through multilayer perception

Case Processing Summary

		N	Percent
Sample	Training	912	52.0%
	Testing	415	23.7%
	Holdout	426	24.3%
Valid		1753	100.0%
Excluded		1013	
Total		2766	

Figure 8

We simply break it down to Summary breakdown:

- Training: 912 cases, representing 52.0% of the sample.
- Testing: 415 cases, representing 23.7% of the sample.
- Holdout: 426 cases, representing 24.3% of the sample.
- Valid cases: 1753 cases, accounting for 100% of the valid cases in the dataset.
- Excluded cases: 1013 cases were excluded from the analysis.
- Total cases: 2766 cases were initially included in the dataset.
-

Network Information

Input Layer	Covariates	1	yt-1
		2	yt-2
		3	yt-3
		4	yt-4
		5	yt5
		6	yt6
		7	yt7
		8	yt8
		9	yt9
	Number of Units[a]		9
	Rescaling Method for Covariates		Standardized
Hidden Layer(s)	Number of Hidden Layers		1
	Number of Units in Hidden Layer 1[a]		7
	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

Figure 9

The neural network model's architecture and properties are described by the network information. Nine covariates reflecting lagged values of the dependent variable, ranging from yt-1 to yt-9, make up the input layer. These covariates have been rescaled to have a mean of 0 and a standard deviation of 1, which is known as standardization.

The network has one hidden layer with seven units. The sigmoid function, which converts the input values to a range between 0 and 1, is the activation function employed in this hidden layer.

The output layer is responsible for predicting the value of the dependent variable yt. It has a single unit and uses the identity function as its activation function, which simply passes the input through. The dependent variable's scale is also standardised.

The error function used in the network is the sum of squares, which calculates the squared difference between the predicted values and the actual values to measure the overall prediction error.

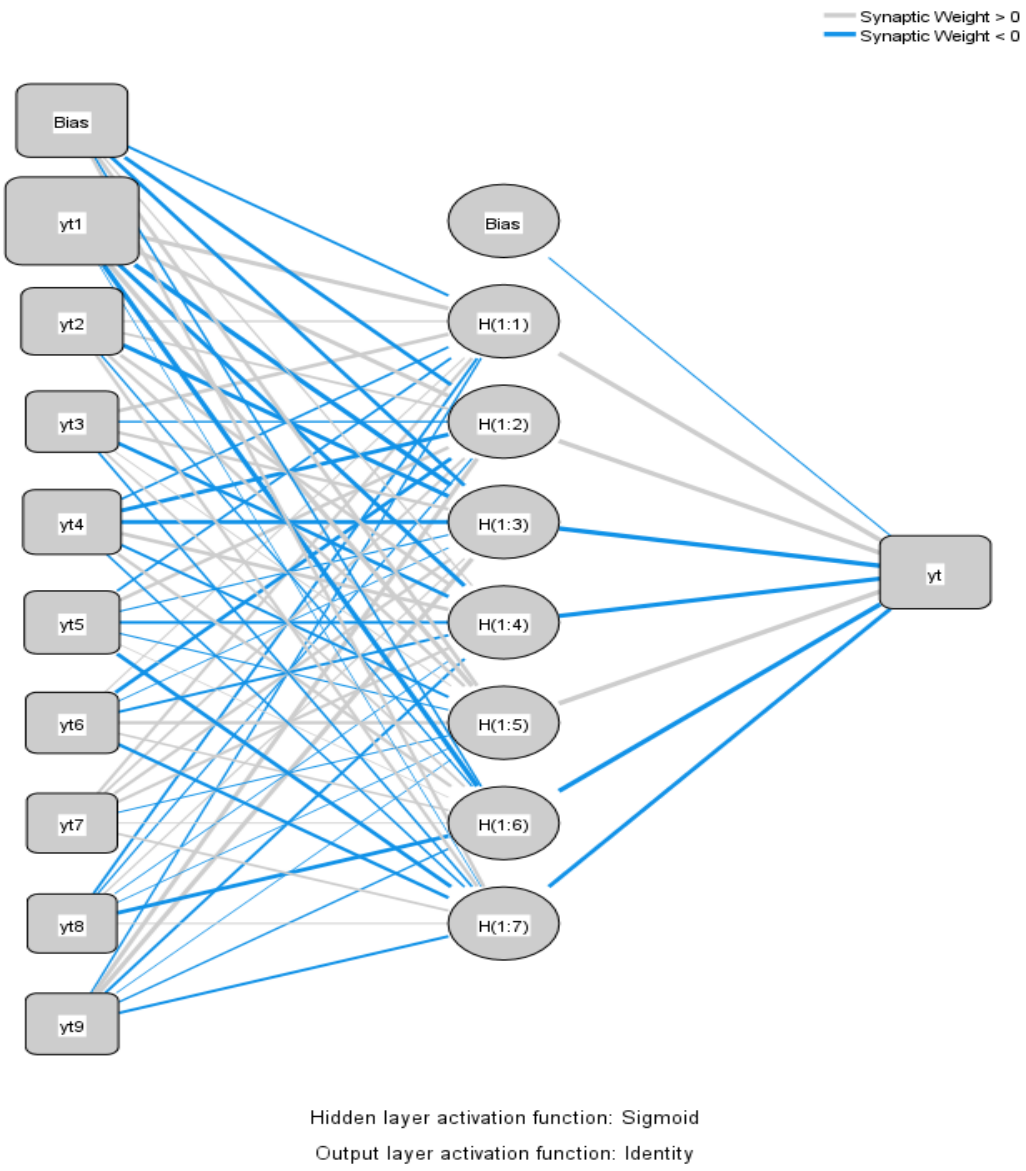


Figure- 10

There is one hidden layer in figure 10. 7 hidden neurons. Grayarcs(Synaptic weights): shows positive whereas blue arcs reflects to positive.

Model Summary

Training	Sum of Squares Error	1.972
	Relative Error	.004
	Stopping Rule Used	1 consecutive step(s) with no decrease in error[a]
	Training Time	0:00:00.03
Testing	Sum of Squares Error	.738
	Relative Error	.004
Holdout	Relative Error	.003

Dependent Variable: yt

a. Error computations are based on the testing sample.

Figure 11

Training: Sum of Squares Error: 1.972 Relative Error: 0.00

Testing: Sum of Squares Error: 0.738 Relative Error: 0.004

Holdout: Relative Error: 0.003 Dependent Variable: yt

Normalized Importance

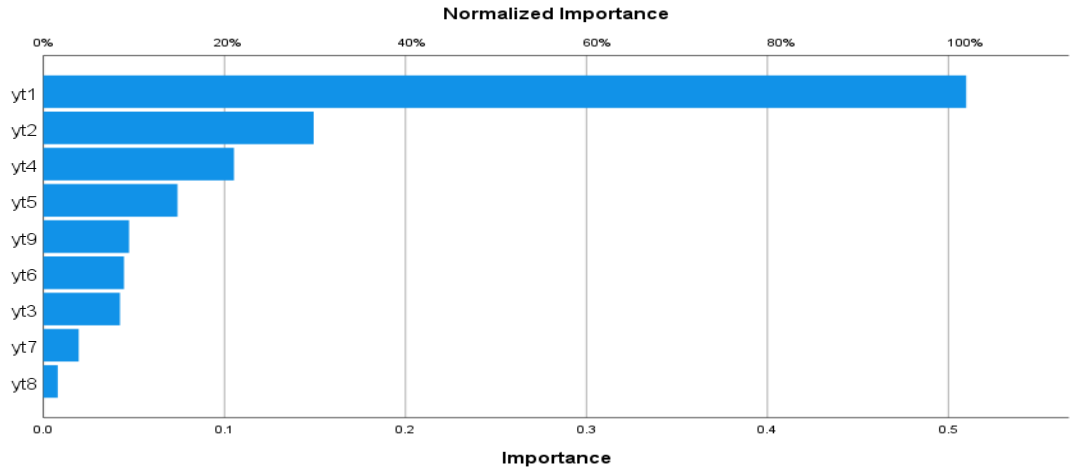
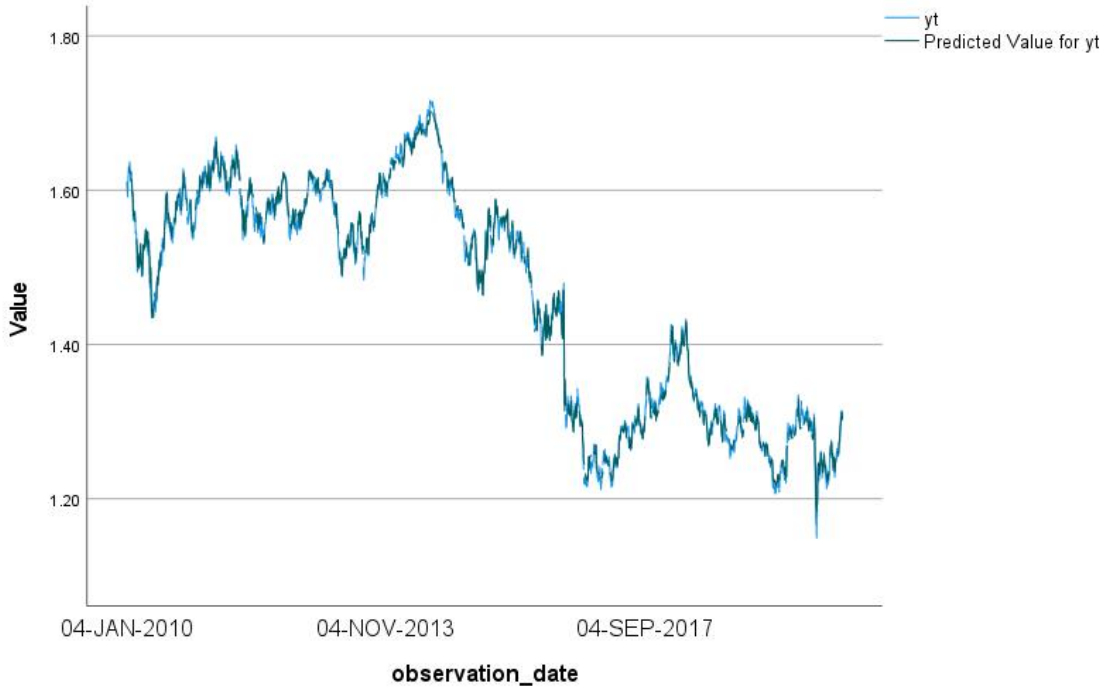


Figure 11

In Figure 11, yt1 refers to the most important bar which decreases exponentially from yt2 till yt8 . Indicating the least importance of among the all.

Original VS Predicted value



-Blue line is the predicted one

-Green line is the predicted one

Figure 13

The figure above indicates that predicted value fits the original model with a very little deviations in certain places.

	SNo	observation_date	DEXUSUK	yt9	yt8	yt7	yt6	yt5	yt4	yt3	yt2	yt1	yt	MLP_PredictedValue
2756	2756.00	2020-07-27	1.29	1.25	1.26	1.26	1.26	1.27	1.27	1.27	1.28	1.28	1.29	1.28
2757	2757.00	2020-07-28	1.30	1.26	1.26	1.26	1.27	1.27	1.27	1.28	1.28	1.29	1.30	1.29
2758	2758.00	2020-07-29	1.30	1.26	1.26	1.27	1.27	1.27	1.28	1.28	1.29	1.30	1.30	1.30
2759	2759.00	2020-07-30	1.30	1.26	1.27	1.27	1.27	1.28	1.28	1.29	1.30	1.30	1.30	1.30
2760	2760.00	2020-07-31	1.31	1.27	1.27	1.27	1.28	1.28	1.29	1.30	1.30	1.30	1.31	1.30
2761	2761.00	2020-08-03	1.31	1.27	1.27	1.28	1.28	1.29	1.30	1.30	1.30	1.31	1.31	1.31
2762	2762.00	2020-08-04	1.31	1.27	1.28	1.28	1.29	1.30	1.30	1.30	1.31	1.31	1.31	1.31
2763	2763.00	2020-08-05	1.31	1.28	1.28	1.29	1.30	1.30	1.30	1.31	1.31	1.31	1.31	1.30
2764	2764.00	2020-08-06	1.31	1.28	1.29	1.30	1.30	1.30	1.31	1.31	1.31	1.31	1.31	1.31
2765	2765.00	2020-08-07	1.30	1.29	1.30	1.30	1.30	1.31	1.31	1.31	1.31	1.31	1.30	1.32
2766	2766.00	2020-08-08	.	1.30	1.30	1.30	1.31	1.31	1.31	1.31	1.31	1.30	.	1.30

Figure 14

Finally,after the analysis SPSS created a column of Predicted values where the forecasted value of 8th August 2020 which is 1.30.

Final Output:

Output2 f ina
l.spv