Aston Business School
Birmingham

Business in Practice Portfolios

Table of Contents

Portfolio 1: Logistic Regression

The aim of this project is to build a model that will be able to determine the spending type of potential customers. Potential customers fall into three main categories:

- Low Spender
- Medium Spender
- High Spender

Since there are 3 outcomes to predict, we are going to build a multinomial logistic regression model where the spender variable (dependent) outcomes are assigned to dummy variables 0 (low), 1(medium), and 2(high).

It is a multinomial regression model therefore before we start building our model we need to create a reference group to the model. A reference group is selected by the highest likelihood of occurrence. We can see from the below table that most frequent spender type in our data is medium spender. Therefore, I will select medium spender as the reference.

Spender

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

Linearity Assumption

Before we go ahead and create the model, the first thing we need to do is check whether the assumptions of the logistic regression model are met in order to avoid creating any biased model.

The first assumption that we need to test is the linearity assumption. We do so by creating the natural log variables for all continuous independent variables and run a multinomial logistic regression model between the interactions of the natural log variables and continuous variables and assigning spender as the dependent variable. After we run our model. We want to see whether the interactions between our independent variables are significant or not. If they are all insignificant then our models, follow the linearity assumptions.

Parameter Estimates

Spender[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B) Lower Bound	Upper Bound
Low Spender	Intercept	14.824	3.919	14.306	1	.000			
	Age * Ln_Age	-.081	.042	3.684	1	.055	.922	.849	1.002
	Value Products * Ln_valueProducts	-.091	.080	1.280	1	.258	.913	.781	1.069
	Brand Products * Ln_BrandProducts	-.192	.129	2.206	1	.137	.826	.641	1.063
	Top Fresco Products * Ln_TopFrescoProducts	-.249	.136	3.364	1	.067	.780	.597	1.017
Medium spender	Intercept	7.370	2.117	12.120	1	.000			
	Age * Ln_Age	-.015	.009	2.763	1	.096	.985	.967	1.003
	Value Products * Ln_valueProducts	-.024	.021	1.299	1	.254	.976	.937	1.017
	Brand Products * Ln_BrandProducts	-.038	.037	1.046	1	.306	.963	.895	1.035
	Top Fresco Products * Ln_TopFrescoProducts	-.138	.059	5.468	1	.019	.871	.776	.978

a. The reference category is: High Spender.

As we can see from the above table, the significance in both models for the interactions are all greater (>) than 0.05 which means they are all insignificant, therefore the linearity assumption is met.

Estimating the model and achieving a parsimonious model

After we have confirmed that the linearity assumption is met, we can go ahead and build the multinomial logistic model where we assign age, value product, brand product, top fresco product, gender, and store type will be our independent variables and spending type as our dependent variable.

Parameter Estimates

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

a. The reference category is: Medium spender.

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

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

We need to check from the above model if all independent variables(IV's) have strong explanatory power. To do so we look at the walds statistic coefficients. IV's where the significance of the coefficient is higher than 0.05 are insignificant and need to be removed from the model. We begin by removing the most insignificant variable and re-run it. We keep repeating the process until all IV's are significant(<0.05) and achieved a parsimonious model.

Parameter Estimates

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

a. The reference category is: Medium spender.

The above table is the most parsimonious model achieved after removing all insignificant variables.

Models Adequacy

Now that we have achieved the most parsimonious, we need to make sure our model is adequate. We do so by running several test:

1. **Multicollinearity**: there should be no collinearity between continuous independent variables.

Coefficients[a]

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

a. Dependent Variable: Spender

The above table shows that the tolerance for all three variables is greater than 0.1 and the VIF values are all greater than 10 which means that no multicollinearity is occurring between IV's.

2. **Examine standardised residuals(ZRE_1)**: Since it's a multinomial model, we have 2 models to examine the standardised residuals for. For the first model, we ran binary logistic regression between low and medium spender and retrieved the standardized residuals and then repeated the same steps to retrieve the standardized residuals between medium and high spenders.
<u>Results:</u> In both models, less than 5% had absolute values(ZRE_1) above 2 and less than 1% had absolute values(ZRE_1) above 2.5. Therefore, it passed the residual test.

3. **Cook's Distance**: The same process was repeated from the standardised residuals test.
<u>Results:</u> In both models **no** residuals have a Cook's distance over 1, therefore it passes.

4. **DFBetas:** The same process was repeated from the standardised residuals test. It measures the difference in each parameter estimate with and without the influential point.
<u>Results</u>**:** DFBetas is less than 1 for every independent variable.

Results of some of the outputs:

COO_1	ZRE_1	DFB0_1	DFB1_1	DFB2_1	DFB3_1	COO_2	ZRE_2	DFB0_2	DFB1_2	DFB2_2
.21418	1.27914	.04191	.00989	.01442	-.06926	.00002	-.04875	-.01119	.00010	.00011
.18567	.58777	-1.47129	.06900	-.01168	-.01245	.00003	-.05639	-.01331	.00009	.00018
.	.	.	.	.	.	.00012	.07956	-.02224	.00012	.00060
.00664	-.22397	-.14001	-.00049	.01502	.01429	.	.	.	.	.
.	.	.	.	.	.	.00000	.01235	-.00088	.00000	.00002
.00035	-.10928	-.07743	.00195	.00212	.00463	.	.	.	.	.
.	.	.	.	.	.	.00007	.05961	-.01086	-.00006	.00044
.00852	-.33336	-.27596	.00482	.01075	.02142	.	.	.	.	.
.00000	.02624	-.00879	.00026	.00019	.00048	.00527	-.31623	-.17488	.00193	.00278
.00000	.00649	-.00068	.00002	.00002	.00004	.02258	-.55134	-.28689	.00465	-.00083
.00000	.00210	-.00010	.00000	.00000	.00000	.02458	-.61058	-.14258	-.00144	.00238
.00006	.05604	-.02967	.00078	.00125	.00123	.00128	-.17908	-.09029	.00095	.00022
.00000	.01008	-.00147	.00004	.00006	.00008	.01565	-.44217	-.26985	.00406	-.00116

All four assumptions are satisfied; therefore, the model is adequate.

Goodness of fit

Here we ran a few tests to check how well our model fits a given set of observations, or how well it predicts them. It is important to have a model that predicts well if we want to deploy in the supermarket.

1.

Pseudo R-Square

Cox and Snell	.767
Nagelkerke	.868
McFadden	.676

Pseudo R-square is one of the tests we run to check its goodness of fit and we can see that Cox and Snell's as well as Nagelkerke's test are close to 1. This indicated that our model is very good.

2. Hosmer and Lemeshow's test: we can see the significance of the test is well above 0.05, which means our model is really good.

Hosmer and Lemeshow Test

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

3. Classification Table

Classification

Observed	Predicted: Low Spender	Predicted: Medium spender	Predicted: High Spender	Predicted: Percent Correct
Low Spender	16	2	0	88.9%
Medium spender	4	23	3	76.7%
High Spender	0	4	23	85.2%
Overall Percentage	26.7%	38.7%	34.7%	82.7%

We can see from the above table that the overall accuracy of the model is 82.7% which is really good at prediction. Moreover, the accuracy of the model for correctly predicting that the potential customers are low, medium, or high spender is 88.9%, 76.7%, and 85.7% respectively, which is very accurate as well. Therefore, the overall goodness of fit of our model is really good.

Predictive model output interpretation

By looking at the Exp(B) values for our parsimonious model we can deduce the following:

1. **Low spender customers**:

   Age: For every unit increase in age there is a 27.6% reduction in the odds of the customer being a low spender with reference to medium spenders.

   Value Products: For every unit increase in the value product, there is a 29.7% reduction in the odds of the customer being a low spender with reference to medium spenders.

   Top fresco products: For every unit increase in Top fresco products, there is a 44.1 % reduction in the odds of the customer being a low spender with reference to medium spenders.

2. **High spender customers:**

   Age: For every unit increase in age there is an 8.7% rise in the odds of the customer being a high spender with reference to medium spenders.

   Value Products: For every unit increase in the value product, there is a 15.8% rise in the odds of the customer being a high spender with reference to medium spenders.

   Top fresco products: For every unit increase in Top fresco products, there is a 52.4 % rise in the odds of the customer being a high spender with reference to medium spenders.

Recommendations

- Since customers are more likely to be high spenders as they age, then considering creating specialized products targeting younger age groups would be a good way to increase revenues
- Customers are more likely to spend more on top fresco products, therefore increasing varieties of top fresco products and reducing brand products can increase revenue for the supermarket.

Portfolio 2: Conjoint Analysis

Introduction

The aim of this report is to conduct a conjoint analysis in order to identify people's most influential decision making based on a combination of attributes and levels concerning mobile phone features to launch a successful phone that suits most people's likings.

Attribute selection

As you can see from the table below, we have selected the most four common attributes customers look for when purchasing a mobile phone. These attributes consist of 3, 2, 2, 3 levels respectively, hence 36 combinations/products have been created in a survey and submitted to 10 people comprising of family and friends asking them to rank these combinations from best preferred (36) to least preferred (1).

Storage	Size	headphone jack	Price
64GB	5.5"	with jack	£499
128GB	6.7"	without jack	£799
256GB			£999

Regression Analysis

Before running the regression analysis, for every attribute with n levels, we had to create an n-1 dummy variable. We removed the first level for every attribute to avoid multicollinearity while running the regression model.

Dependent variable: The **average ranking** of our 10 respondents.

Independent variable: 128GB, 256GB, 6.7", without jack, £799, £999

Output:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.987[a]	.975	.970	1.83203

a. Predictors: (Constant), £999, Without Jack, 6.7", 256GB, 128GB, £799

The model achieves an R-square value of 0.975 which means our dummy variables is able to predict the rank order with 97.5% accuracy.

Model Analysis with Utility Graphs

Coefficients[a]

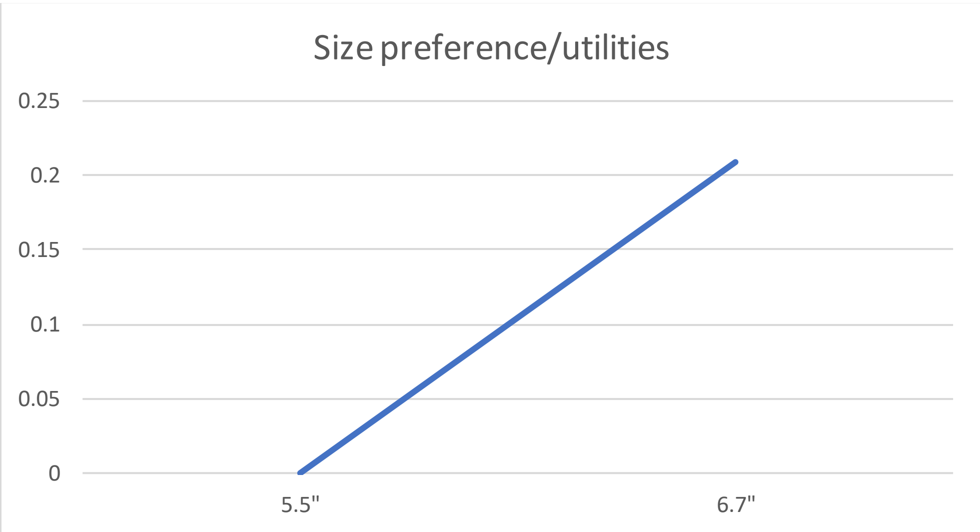
Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	31.083	.808		38.477	.000
	128GB	.833	.748	.038	1.114	.274
	256GB	.417	.748	.019	.557	.582
	6.7"	4.333	.611	.209	7.096	.000
	Without Jack	-6.333	.611	-.305	-10.371	.000
	£799	-12.750	.748	-.579	-17.047	.000
	£999	-23.250	.748	-1.055	-31.086	.000

a. Dependent Variable: Rank

From the table above we look at the standardized coefficient beta which gives us an outlook on how certain levels have an impact on people's preferences.

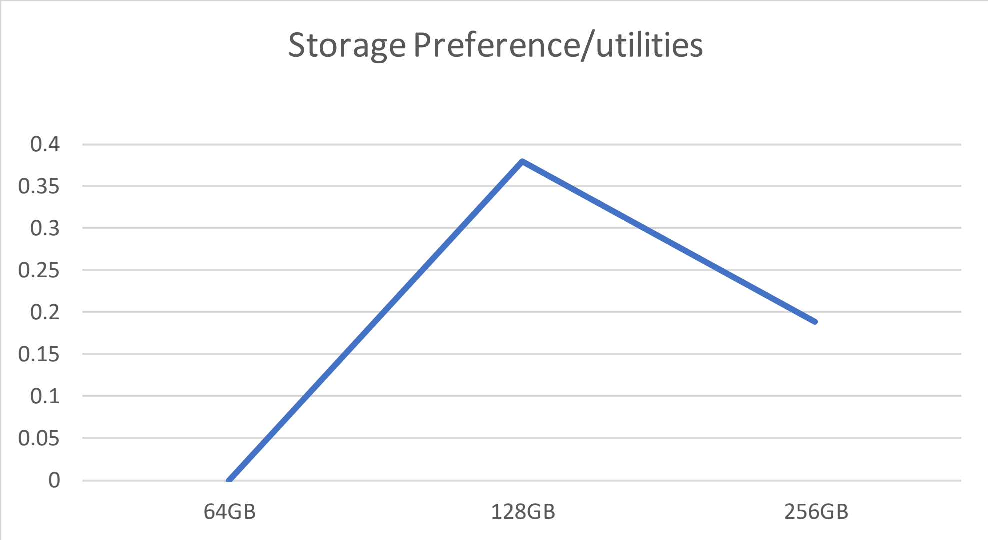
Screen Size

- 6'7 has the highest beta coefficient (0.209) with respect to all attributes. This reveals that choosing the screen size is the most important attribute they consider when choosing a phone.



Storage

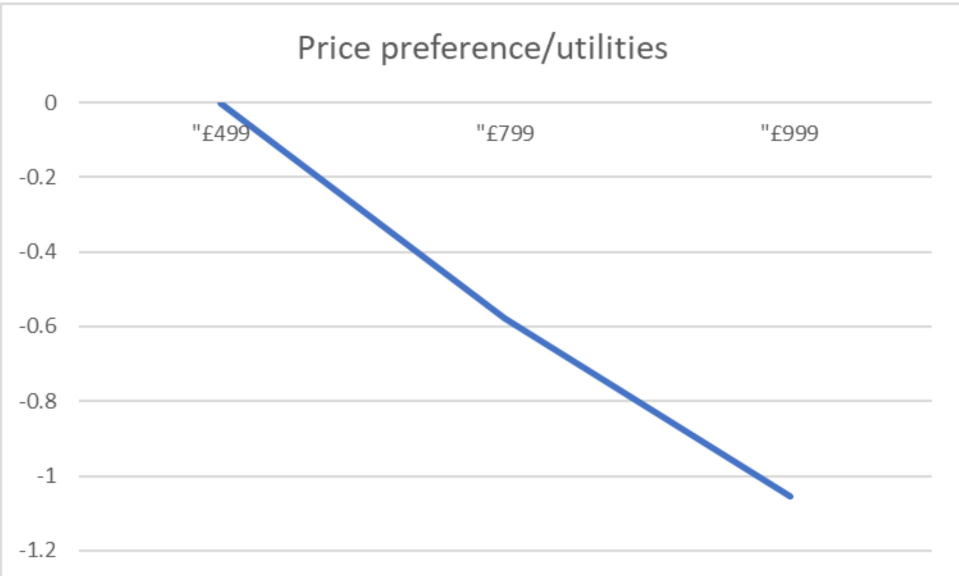
- 128GB has the highest coefficient beta for out of all levels under the storage attribute. This shows that people mostly prefer a phone with 128GB storage memory.
- As shown in the utility graph below there is a drastic downward shift in customers' decision/preference as phone's storage increases from 128GB to 256GB.



Price

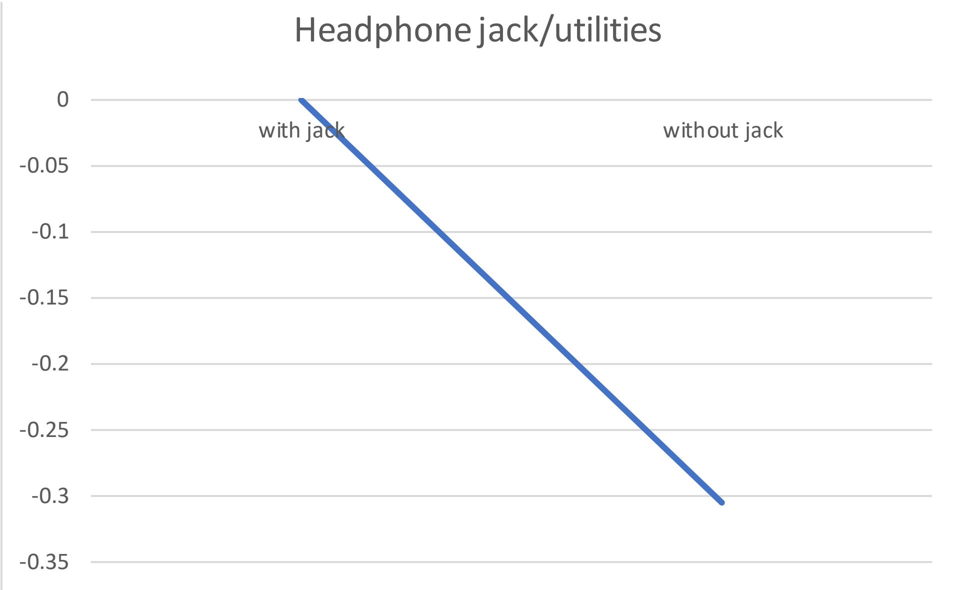
- £999 has the lowest beta coefficient (-1.055) with respect to all levels and all levels of other attributes. This reveals that high phone prices play a huge negative impact in changing customers' decisions in getting the phone.

- It can be seen from the below utility graph how there is a drastic negative change in customer's decisions in getting the phone as price increases.



Headphone Jack

- Below graph shows how not having a headphone jack can have a negative change in customer's decisions in getting the phone.



Utility Value Combinations

The below table shows the utilities of all possible product combinations with all their different levels. We can observe from the below table by looking at the positive utility values that customers get influenced positively in their decisions mostly with phones that have a low price (£499), have a headphone jack built-in them, screen size of 6'7 and has a storage of 128GB.

Product Combinations	Sum of utilities
128GB,6.7",with jack,499	0.247
256GB,6.7",with jack,499	0.228
64GB,6.7",with jack,499	0.209
128GB,5.5",with jack,499	0.038
256GB,5.5",with jack,499	0.019
64GB,5.5",with jack,499	0
128GB,6.7",without jack,499	-0.058
256GB,6.7",without jack,499	-0.077
64GB,6.7",without jack,499	-0.096
128GB,5.5",without jack,499	-0.267
256GB,5.5",without jack,499	-0.286
64GB,5.5",without jack,499	-0.305
128GB,6.7",with jack,799	-0.332
256GB,6.7",with jack,799	-0.351
64GB,6.7",with jack,799	-0.37
128GB,5.5",with jack,799	-0.541
256GB,5.5",with jack,799	-0.56
64GB,5.5",with jack,799	-0.579
128GB,6.7",without jack,799	-0.637
256GB,6.7",without jack,799	-0.656
64GB,6.7",without jack,799	-0.675
128GB,6.7",with jack,999	-0.808
256GB,6.7",with jack,999	-0.827
64GB,6.7",with jack,999	-0.846
128GB,5.5",without jack,799	-0.846
256GB,5.5",without jack,799	-0.865
64GB,5.5",without jack,799	-0.884
128GB,5.5",with jack,999	-1.017
256GB,5.5",with jack,999	-1.036
64GB,5.5",with jack,999	-1.055
128GB,6.7",without jack,999	-1.113
256GB,6.7",without jack,999	-1.132
64GB,6.7",without jack,999	-1.151
128GB,5.5",without jack,999	-1.322
256GB,5.5",without jack,999	-1.341
64GB,5.5",without jack,999	-1.36

Correlation Analysis

Now, we are going to run a correlation analysis in order to see how significantly correlated the customer ranking and sum of utilities with each other.

Correlations

		Sum of utility	Rank
Sum of utility	Pearson Correlation	1	.987**
	Sig. (2-tailed)		.000
	N	36	36
Rank	Pearson Correlation	.987**	1
	Sig. (2-tailed)	.000	
	N	36	36

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

There is a strong significant correlation of 0.987 between the ranking and the utilities, therefore this shows that our utility estimates are accurate.

Portfolio 3: Clustering Analysis

Introduction

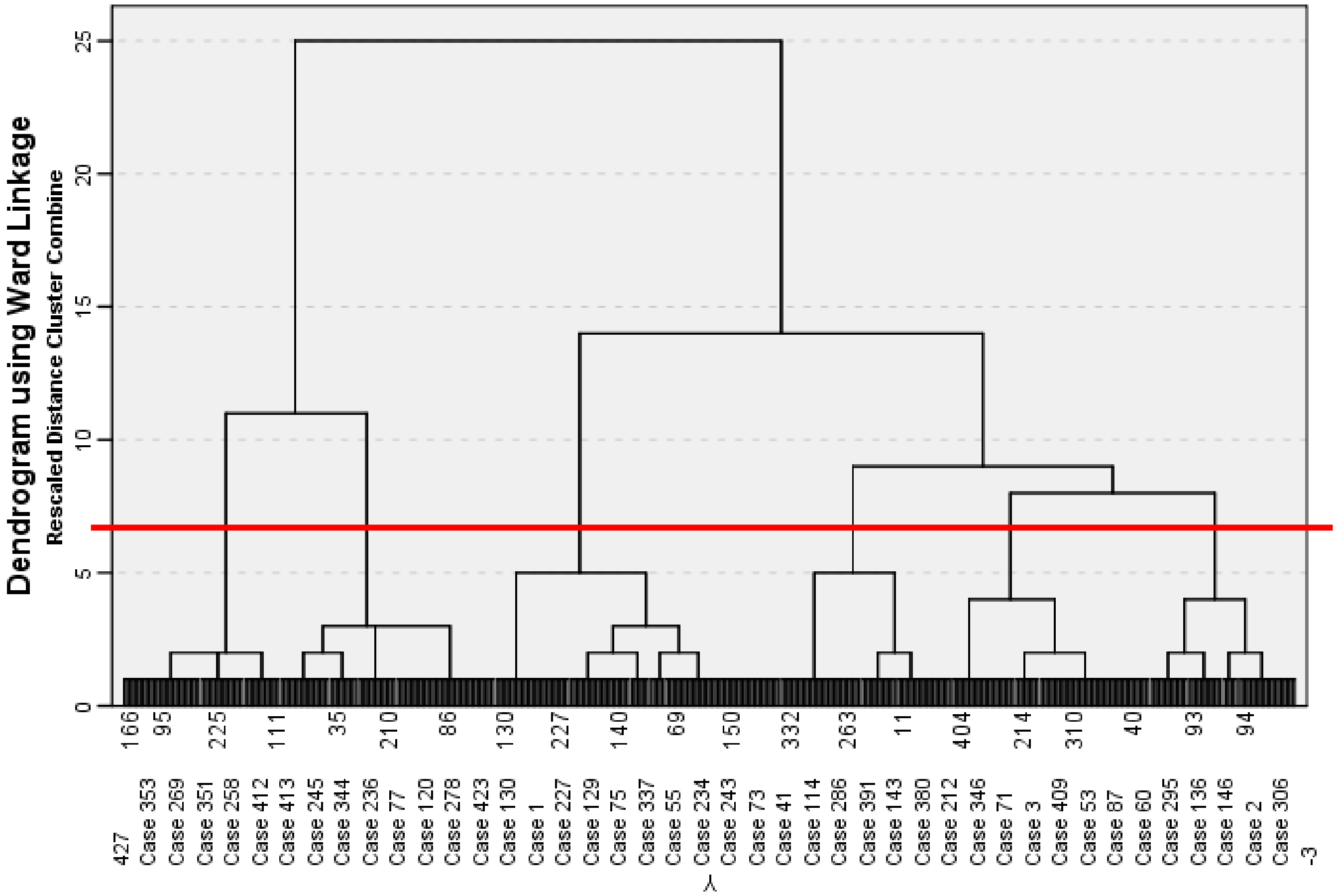
The aim of this report is to help a UK's bank development team to undertake a segmentation analysis in order to identify trends and patterns in a sample of records collected from their customers. Cluster analysis has been used to identify hidden segments of customers by using SPSS to create a different number of clusters using two different methods. The best clustering of customers will be based on how evenly distributed the clusters are for different customer groups.

Variable	Current Account	Saving Account	Months Customer	Months Employed	Age	Job	Credit Risk	Gender	Marital Status	Housing
Type	Cont.	Cont.	Cont.	Cont.	Cont.	Categ.	Categ.	Categ.	Categ.	Categ.

In the beginning, the input parameters used for cluster analysis were only the continuous features since SPSS only accepts numerical values. However, when running the clustering analysis using agglomerative hierarchical clustering methods, the model was producing a huge disparity in the distribution of the clusters which created a biased variance, even while using different algorithms. Therefore, we remodeled our clustering analysis by including all the five categorical variables, but since SPSS doesn't accept categorical variables we converted all the categorical to numeric by assigning them to a correspondent number.

Method 1

For the first model, the clustering method used was Wards method using Euclidean distance to measure proximity. Dendrogram was used to estimate the number of clusters to select as part of primary analysis.



From the above dendrogram, the red line has been able to slice horizontally the dendrogram into 6 clusters. Therefore, we shall start our iteration with 6 clusters and downward and see which clusters have values with fair distributions to each other.

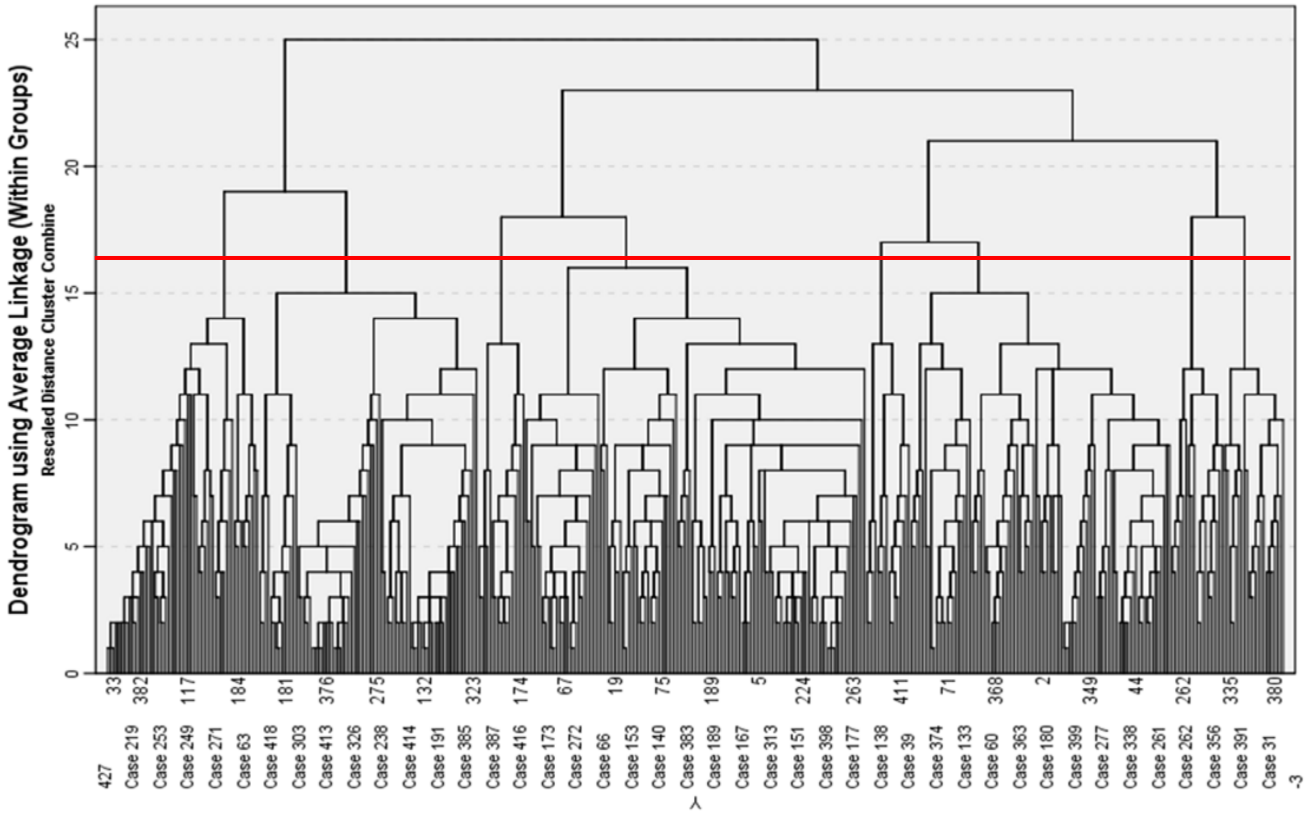
Frequencies Tables

<table>
<tr><th>Number of Clusters</th><th>Wards Frequency Table</th><th>Interpretation</th></tr>
<tr><td>Group 6</td><td><table>
<tr><th></th><th></th><th>Frequency</th><th>Percent</th></tr>
<tr><td>Valid</td><td>1</td><td>103</td><td>24.2</td></tr>
<tr><td></td><td>2</td><td>53</td><td>12.5</td></tr>
<tr><td></td><td>3</td><td>72</td><td>16.9</td></tr>
<tr><td></td><td>4</td><td>62</td><td>14.6</td></tr>
<tr><td></td><td>5</td><td>78</td><td>18.4</td></tr>
<tr><td></td><td>6</td><td>57</td><td>13.4</td></tr>
</table></td><td>The proportion of values in all clusters are nearly similar to each, however, it is not feasible for the bank to invest in curating 6 different</td></tr>
</table>

<table>
<tr><td></td><td></td><td>products. As it requires a lot of capital.</td></tr>
<tr><td>Group 5</td><td><table><tr><th></th><th></th><th>Frequency</th><th>Percent</th></tr><tr><td>Valid</td><td>1</td><td>103</td><td>24.2</td></tr><tr><td></td><td>2</td><td>125</td><td>29.4</td></tr><tr><td></td><td>3</td><td>62</td><td>14.6</td></tr><tr><td></td><td>4</td><td>78</td><td>18.4</td></tr><tr><td></td><td>5</td><td>57</td><td>13.4</td></tr></table></td><td>Same interpretation applies according to Group6.</td></tr>
<tr><td>Group 4</td><td><table><tr><th></th><th></th><th>Frequency</th><th>Percent</th></tr><tr><td>Valid</td><td>1</td><td>103</td><td>24.2</td></tr><tr><td></td><td>2</td><td>187</td><td>44.0</td></tr><tr><td></td><td>3</td><td>78</td><td>18.4</td></tr><tr><td></td><td>4</td><td>57</td><td>13.4</td></tr></table></td><td>A huge proportion are clustered around 2.There’s disproportionality in this group.</td></tr>
<tr><td>Group 3</td><td><table><tr><th></th><th></th><th>Frequency</th><th>Percent</th></tr><tr><td>Valid</td><td>1</td><td>103</td><td>24.2</td></tr><tr><td></td><td>2</td><td>187</td><td>44.0</td></tr><tr><td></td><td>3</td><td>135</td><td>31.8</td></tr></table></td><td>Values are not quite evenly distributed, most values are clustered around cluster 2. But it could be worth a look for the bank to invest in newly curated products.</td></tr>
<tr><td>Group 2</td><td><table><tr><th></th><th></th><th>Frequency</th><th>Percent</th></tr><tr><td>Valid</td><td>1</td><td>290</td><td>68.2</td></tr><tr><td></td><td>2</td><td>135</td><td>31.8</td></tr></table></td><td>There is disproportionality in this group of clusters. A huge proportion belongs to one cluster while a few belong to the other one.</td></tr>
</table>

Method 2

Now for the second model, the clustering method I used is the Average Linkage (Within Group) method using Euclidean distance to measure proximity. Dendrogram was used to estimate the number of clusters to select as part of primary analysis.



From the above dendrogram, it can be seen that the red line has been able to slice horizontally the dendrogram into 8 clusters. Therefore, we shall start our iteration with 8 clusters and downward and see which cluster has the best evenly proportion.

<table>
<tr><th>Number of Clusters</th><th>Average Linkage (Within Group) Frequency Table</th><th>Interpretation</th></tr>
<tr><td>Group 8</td><td>
<table>
<tr><th></th><th></th><th>Frequency</th><th>Percent</th></tr>
<tr><td>Valid</td><td>1</td><td>123</td><td>28.9</td></tr>
<tr><td></td><td>2</td><td>94</td><td>22.1</td></tr>
<tr><td></td><td>3</td><td>17</td><td>4.0</td></tr>
<tr><td></td><td>4</td><td>79</td><td>18.6</td></tr>
<tr><td></td><td>5</td><td>16</td><td>3.8</td></tr>
<tr><td></td><td>6</td><td>22</td><td>5.2</td></tr>
<tr><td></td><td>7</td><td>19</td><td>4.5</td></tr>
<tr><td></td><td>8</td><td>55</td><td>12.9</td></tr>
</table>
</td><td>Cluster 3,5,6, and 7 have very low values <5%. It is not feasible for the bank to invest in new products for a small cluster of people.</td></tr>
</table>

<table>
<tr><td>Group 7</td><td><table><tr><th></th><th></th><th>Frequency</th><th>Percent</th></tr><tr><td>Valid</td><td>1</td><td>123</td><td>28.9</td></tr><tr><td></td><td>2</td><td>110</td><td>25.9</td></tr><tr><td></td><td>3</td><td>17</td><td>4.0</td></tr><tr><td></td><td>4</td><td>79</td><td>18.6</td></tr><tr><td></td><td>5</td><td>22</td><td>5.2</td></tr><tr><td></td><td>6</td><td>19</td><td>4.5</td></tr><tr><td></td><td>7</td><td>55</td><td>12.9</td></tr></table></td><td>Cluster 3,5,6, have very low values <=5%. It is not feasible for the bank to invest in new products for small clusters of people.</td></tr>
<tr><td>Group 6</td><td><table><tr><th></th><th></th><th>Frequency</th><th>Percent</th></tr><tr><td>Valid</td><td>1</td><td>123</td><td>28.9</td></tr><tr><td></td><td>2</td><td>110</td><td>25.9</td></tr><tr><td></td><td>3</td><td>17</td><td>4.0</td></tr><tr><td></td><td>4</td><td>79</td><td>18.6</td></tr><tr><td></td><td>5</td><td>41</td><td>9.6</td></tr><tr><td></td><td>6</td><td>55</td><td>12.9</td></tr></table></td><td>Cluster 3 and 5 have relatively low values. It's not feasible for a bank to invest in 6 curated products and two of them are for a low proportion of groups</td></tr>
<tr><td>Group 5</td><td><table><tr><th></th><th></th><th>Frequency</th><th>Percent</th></tr><tr><td>Valid</td><td>1</td><td>140</td><td>32.9</td></tr><tr><td></td><td>2</td><td>110</td><td>25.9</td></tr><tr><td></td><td>3</td><td>79</td><td>18.6</td></tr><tr><td></td><td>4</td><td>41</td><td>9.6</td></tr><tr><td></td><td>5</td><td>55</td><td>12.9</td></tr></table></td><td>Values are not distributed evenly for all clusters . Some have low while some have a high proportion of values.</td></tr>
<tr><td>Group 4</td><td><table><tr><th></th><th></th><th>Frequency</th><th>Percent</th></tr><tr><td>Valid</td><td>1</td><td>140</td><td>32.9</td></tr><tr><td></td><td>2</td><td>110</td><td>25.9</td></tr><tr><td></td><td>3</td><td>134</td><td>31.5</td></tr><tr><td></td><td>4</td><td>41</td><td>9.6</td></tr></table></td><td>Cluster 1,2 and 3 have values that are similar in proprtion to each other, while cluster 4 have relatively low proportion of people that belong to it</td></tr>
<tr><td>Group 3</td><td><table><tr><th></th><th></th><th>Frequency</th><th>Percent</th></tr><tr><td>Valid</td><td>1</td><td>140</td><td>32.9</td></tr><tr><td></td><td>2</td><td>151</td><td>35.5</td></tr><tr><td></td><td>3</td><td>134</td><td>31.5</td></tr></table></td><td>This group is an ideal one from the bank. Cluster 1,2 and 3 are all evenly distributed to each other. It would be worth it for the bank to invest in 3 different products for these clusters.</td></tr>
<tr><td>Group 2</td><td><table><tr><th></th><th></th><th>Frequency</th><th>Percent</th></tr><tr><td>Valid</td><td>1</td><td>291</td><td>68.5</td></tr><tr><td></td><td>2</td><td>134</td><td>31.5</td></tr></table></td><td>There is disproprtionality in this group of clusters. A huge proportion belongs to one cluster while a few belong to the other one.</td></tr>
</table>

Final Decision

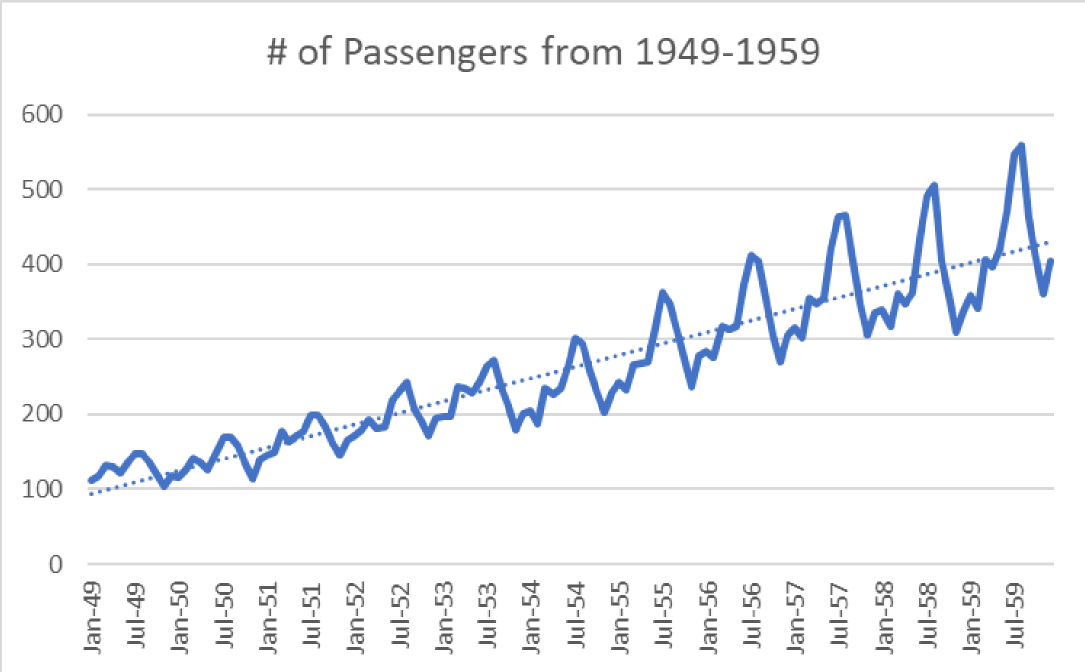
After interpreting both clustering methods, it is advisable for the bank to invest in the development of segment-specific financial products for group 3 of the Average Linkage (Within Group) method since this group is the most ideal clustered group. Cluster 1,2 and 3 have values that are evenly distributed to each other. It would be worth it and less costly for the bank to invest in 3 different products for these 3 segments of people.

Portfolio 4: Time Series Forecasting

Aim-

The aim of this report is to apply forecasting with the decomposition technique on time series data, where we must predict the monthly number of passengers for an American airline company for the year 1960. We are given time series data consisting of a monthly number of passengers using the airline from the year 1949-to 1960.

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

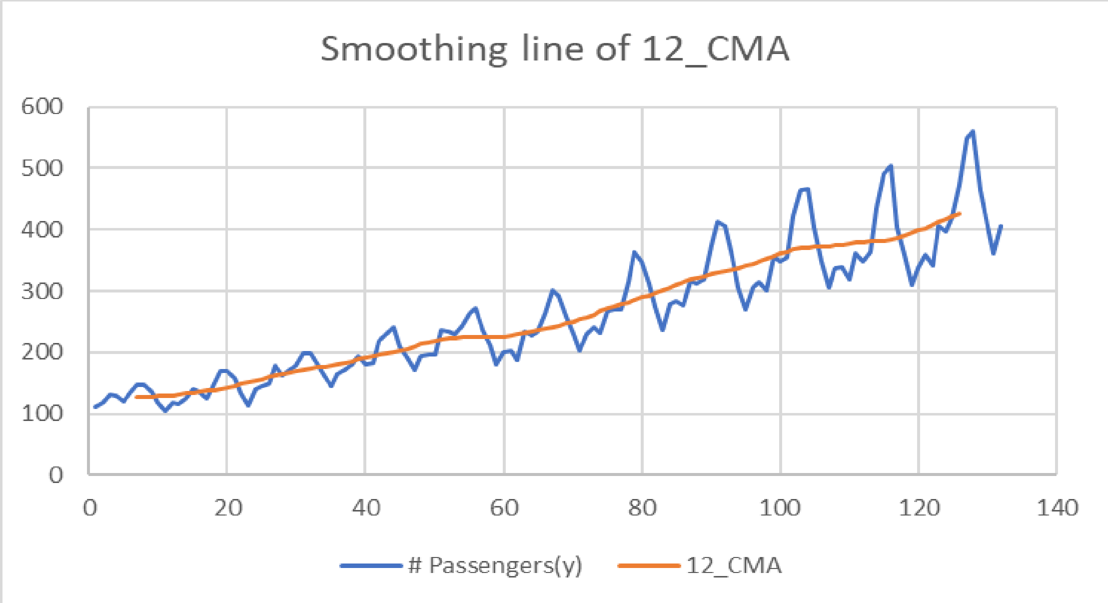


We can see from the graph there is an upward trend, the average number of passengers is increasing over time. Moreover, we can see there is a seasonal component where the number of passengers **peaks in July** during summertime where most people travel for vacation during this period, and it is at its **lowest during January** time, people are less likely to travel during this period.

How many seasons can be recognised in this data set?

Since number of passengers is being recorded monthly over a one-year period, it means there are **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.



Since there are 12 seasons a year and the number of seasons is an even number, I have used the 12 central moving average to smooth and capture the average change in passengers over time and we can see from the above graph that it's moving in an upward trend.

After calculating s_t=Yt/m_t (refer to excel sheet) we can see that no two values from the same month but different season is the same,For example, sept-1949 s_t= 1.062846 and sept=1950 s_t=1.084358. We need to adjust them and come up with the seasonal factor for them to have similar values.

After performing the necessary calculations to obtain the seasonal factor (please refer to the excel sheet to check the whole process of obtaining these results) I have come up with the following outcome:

Month	SF	% impact
Jan	0.910004	-9.00%
Feb	0.887377	-11.26%
Mar	1.018204	1.82%
Apr	0.975412	-2.46%
May	0.979813	-2.02%
Jun	1.11159	11.16%
Jul	1.222147	22.21%
Aug	1.213596	21.36%
Sep	1.060917	6.09%
Oct	0.921767	-7.82%
Nov	0.800213	-19.98%
Dec	0.898962	-10.10%

We can see from the above table how the seasonal factor affects the number of passengers. For example, for the month of January seasonal factor of 0.91 means that the number of passengers will be 9% below the average level of monthly number of passengers. For the month of July seasonal factor of 1.222147 means, the number of passengers will be 22.21% above the average level of monthly number of passengers and so on.

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

The multiplicative model is best suited for this data set since the seasonal amplitude variation is not constant. We can see from the first graph that there is an increase in variation in the seasonal amplitude, therefore we chose the multiplicative model.

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

After retrieving with the following intercept= 92.49 and it's coefficient= 2.55 we were able to come up with the following forecasted DE-seasonalised (Y_t*) equation Y_t*= 92.49+2.55*t were t is the time period. After coming up with the de-seasonalised forecasted number of passengers for every month for year 1960, we will have to multiply the forecasted Y_t* with its seasonal factor in order to make it seasonal. After we doing that for all the months for 1960 we came up with the following forecasted numbers passengers.

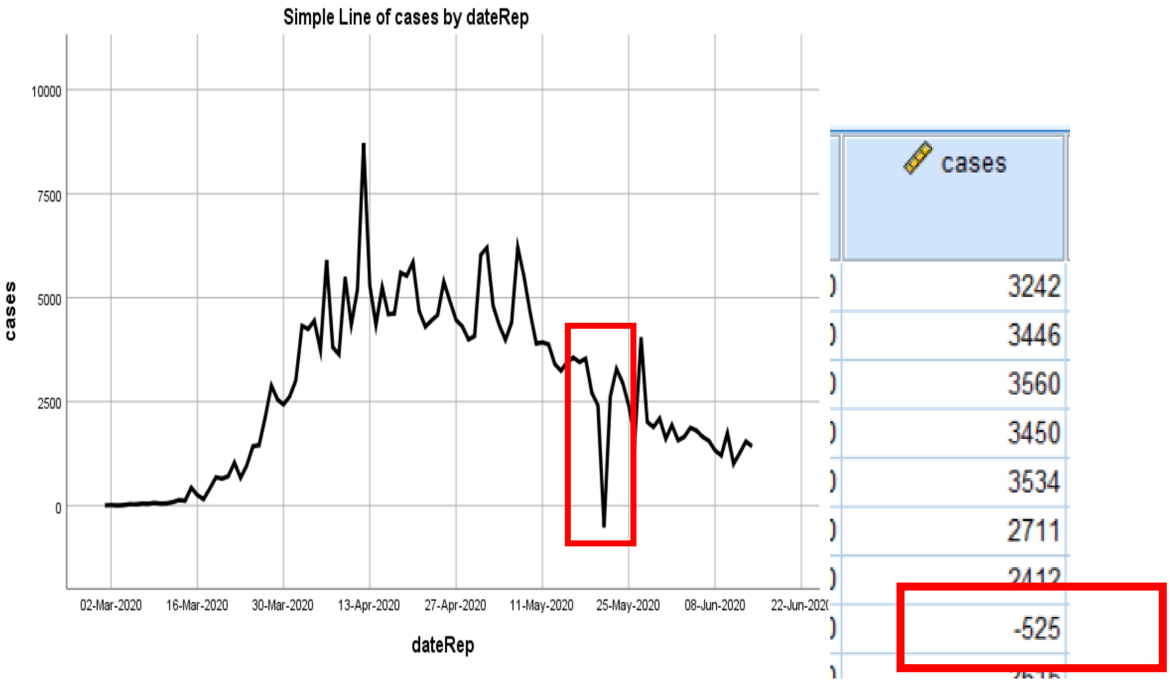
Actual passenger	**Predicted Passengers**
417	393
391	386
419	445
461	429
472	433
535	495
622	547
606	546
508	480
461	420
390	366
432	414

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

Actual passenger	Predicted Passengers	Abs Error	Abs squared error
417	393	24	576.0
391	386	5	25.0
419	445	26	676.0
461	429	32	1024.0
472	433	39	1521.0
535	495	40	1600.0
622	547	75	5625.0
606	546	60	3600.0
508	480	28	784.0
461	420	41	1681.0
390	366	24	576.0
432	414	18	324.0
		MAE	34.33333333
		MSE	1501.0

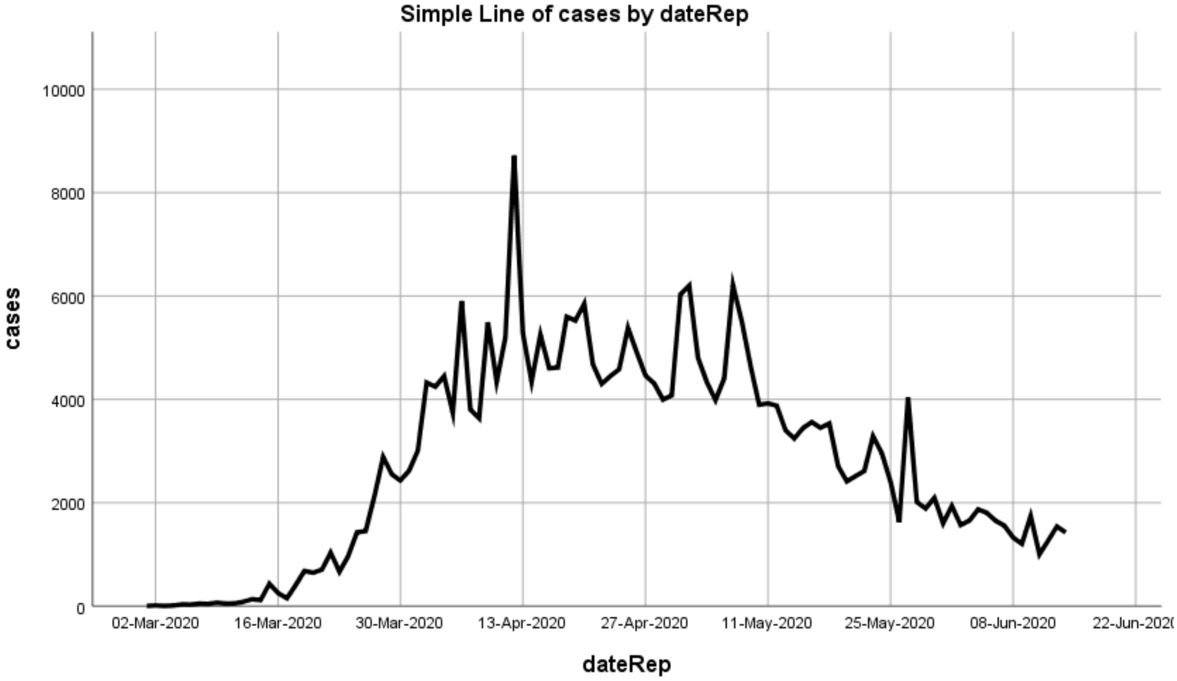
Portfolio 5: ARIMA

Is your time series stationary? Explain it by both using the plot of time series data as well as the applying the autocorrelation method. Does this time series require differencing? Why?



Since cases from 31st Dec 2019 up to 1st of March are close to 0, I have removed them for better interpretation and better forecasting. At first, when we plotted the time series to check whether it is stationary or not we noticed there is a huge dip in cases around may time. After going to check that data we saw there is a negative value in the number of cases which must have been an error. Instead of deleting the value, we imputed it by interpolating the previous and subsequent values of the negative case.

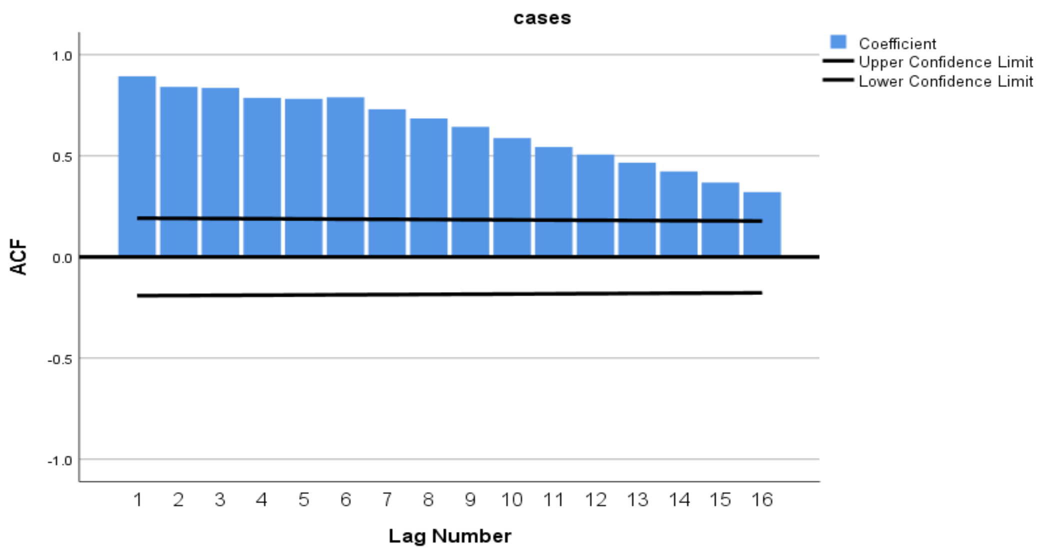
After treating the error, we plotted the time series and analyzed it again.



We can see from the above time-series graph that the number of cases increases significantly overtime before it starts to decrease at a slower rate. Therefore, the above time series is **non-stationary** since the mean cases overtime is not constant.

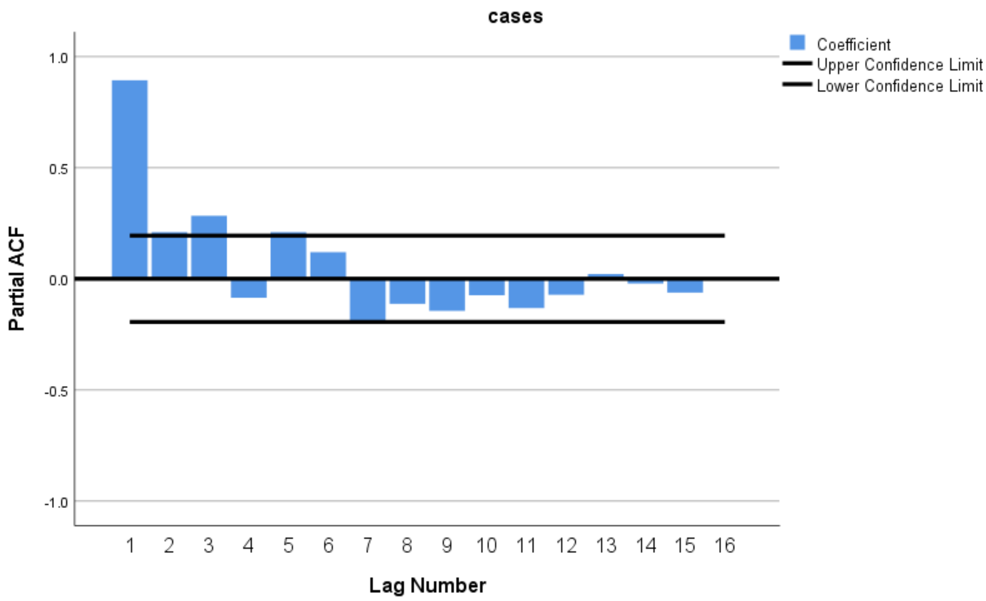
Another way to confirm if our time series is non-stationary is by plotting the autocorrelation and partial autocorrelation functions of the time series.

ACF plot



As we can see from the above plot, all the lags in the ACF plot are significant since they are all above the horizontal line, while there is a gradual decay in the spikes. This indicates a non-stationary series.

PACF plot

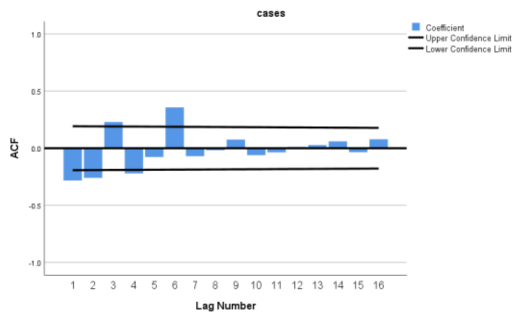


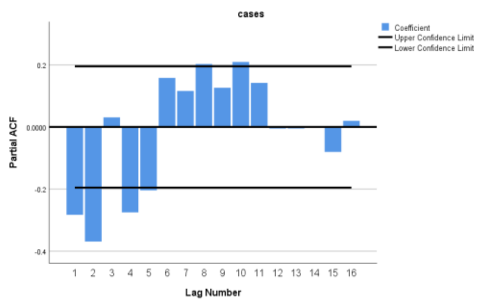
We can see from the PACF plot that the first spike is extremely significant, where the PACF coefficient is close to one, this indicates that it is non-stationary. We can also see that lag 2 and 3 are slightly significant and the rest of the lags are insignificant.

After seeing the ACF and PACF plots we can conclude with high certainty that the time series is **nonstationary**. Therefore, **differencing** is **required** in order to stabilize the mean of the time series and make it stationary.

Estimating ARIMA models with Justifications

After transforming the time series from non-stationary to stationary through differencing, we can select our ARIMA model parameters by looking at our new differenced ACF and PACF plots, where the ACF plot shows us the number of terms in the moving average (q) and PACF shows us the number of lags in autoregressive (p).





From the above plots, we can observe for the ACF that lags 1,2,3,4 and 6 are significant in the MA(q) therefore I will assign q= 6. Additionally, for PACF we can observe as well that lags 1,2,4 and 5 are also significant, there I will assign p=5. Hence, the values for my ARIMA model will initially be assigned to ARIMA(p=4, d=1, q=2) where d stands for differencing and it is assigned to one because we only differenced our times series once.

Diagnostic check

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	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	.418	444.714	2.926	7	.892	0

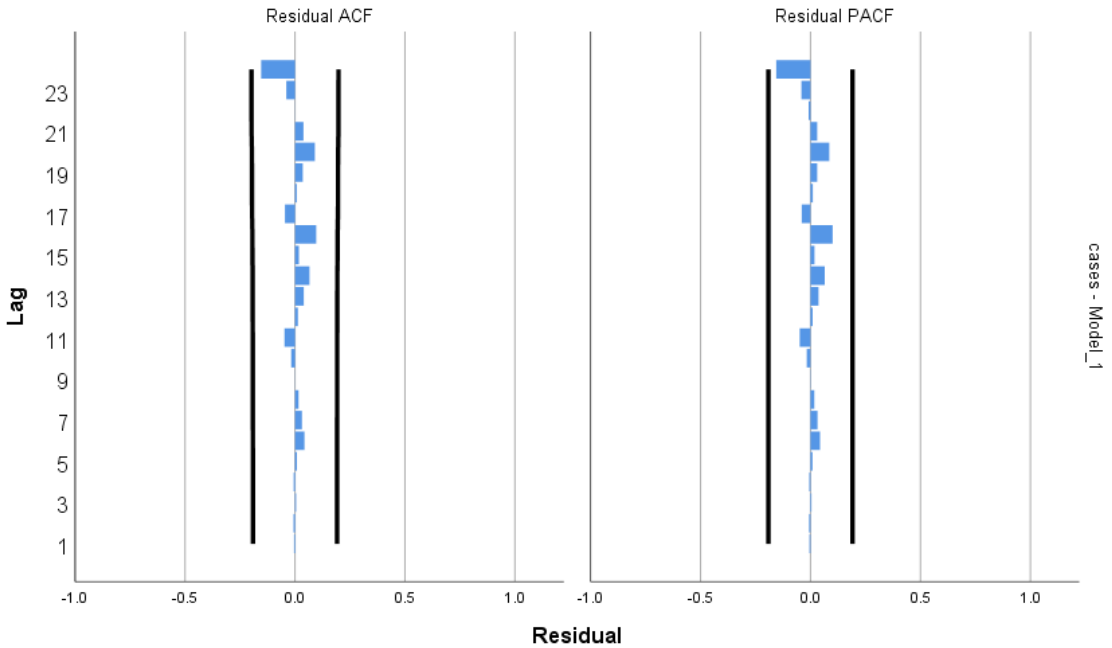
ARIMA Model Parameters

					Estimate	SE	t	Sig.
cases-Model_1	cases	No Transformation	Constant		6.520	48.014	.136	.892
			AR	Lag 1	-.082	.466	-.177	.860
				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	.466	.456	1.023	.309
				Lag 2	.154	.455	.339	.735
				Lag 3	.099	.272	.362	.718
				Lag 4	-.108	.285	-.378	.707
				Lag 5	-.063	.303	-.208	.836
				Lag 6	-.404	.182	-2.213	.029

- By looking at the model statistic table we can see that the Ljung-Box Q test has a significant value(=0.892) of greater than 5% which means that our model is adequate.
- By looking at the ARIMA model parameters we can look at the significance of our lags in AR and MA values. It appears that all lags for AR(p) and all lags for MA(q) except MA(5) are insignificant since their significance values are greater than 0.05. Since we have insignificance lags, we can't consider it as our best model yet, we are going to have to reduce the number of lags later until we reach our most parsimonious model.

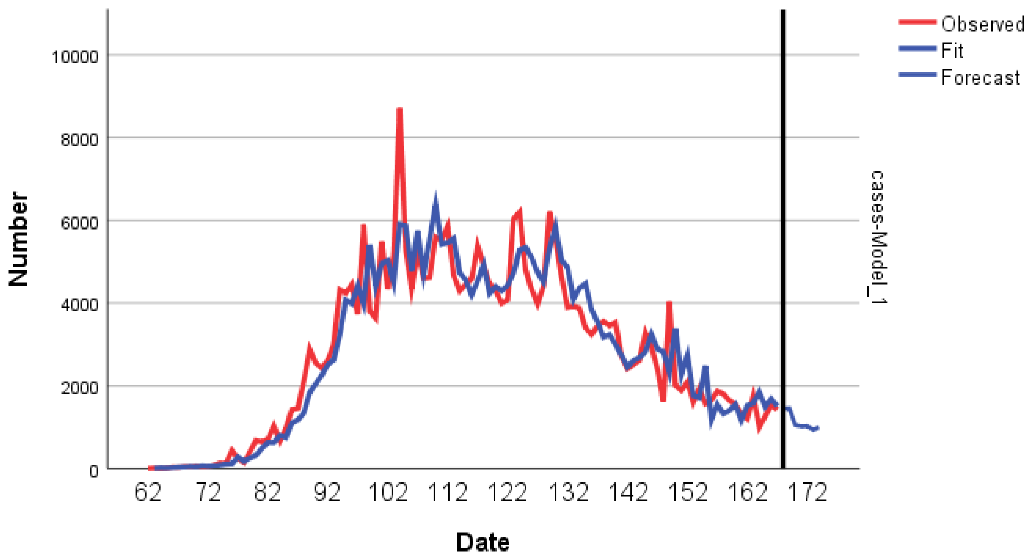
Residual Plots

The below table shows the ACF and PACF residual plots, we can see that all the lags are between the significance intervals. The residual plots are behaving like white noise, so we can say that this model is predicting the behaviour of the time series well.



Goodness of fit

The below graph shows the observed values against the fit values. We can see that the red and blue line almost coincides but not quite well. This could be from having some insignificance lags, because having insignificance lags can have an impact on the standard error of our model, therefore a parsimonious model is desired.



Parsimonious model

The aim is to achieve a parsimonious model where all the lags in our model are significant (<0.05) while still passing the diagnostic test. After removing the insignificant lags and re-running different models in order to achieve a parsimonious model, I ended up with the following ARIMA(p=4,d=1,q=5) model.

Model Description

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

Model Statistics

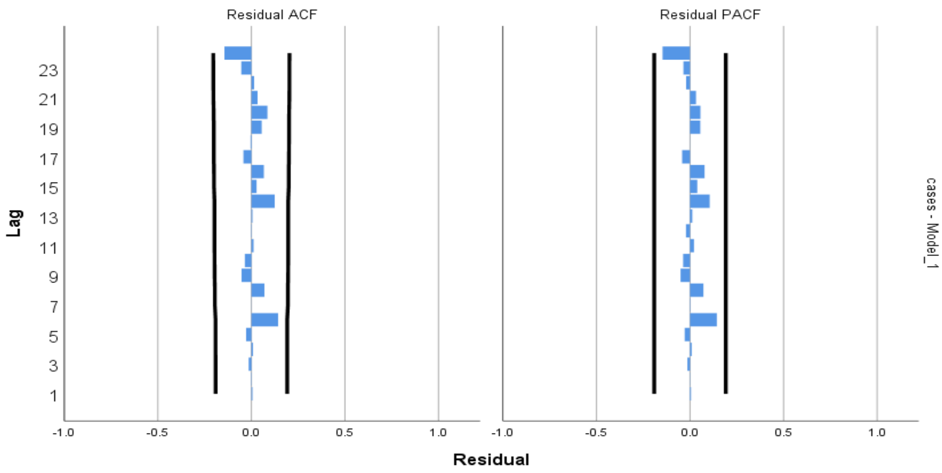
Model	Number of Predictors	Model Fit statistics		Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	MAE	Statistics	DF	Sig.	
cases-Model_1	0	.402	443.230	6.285	9	.711	0

ARIMA Model Parameters

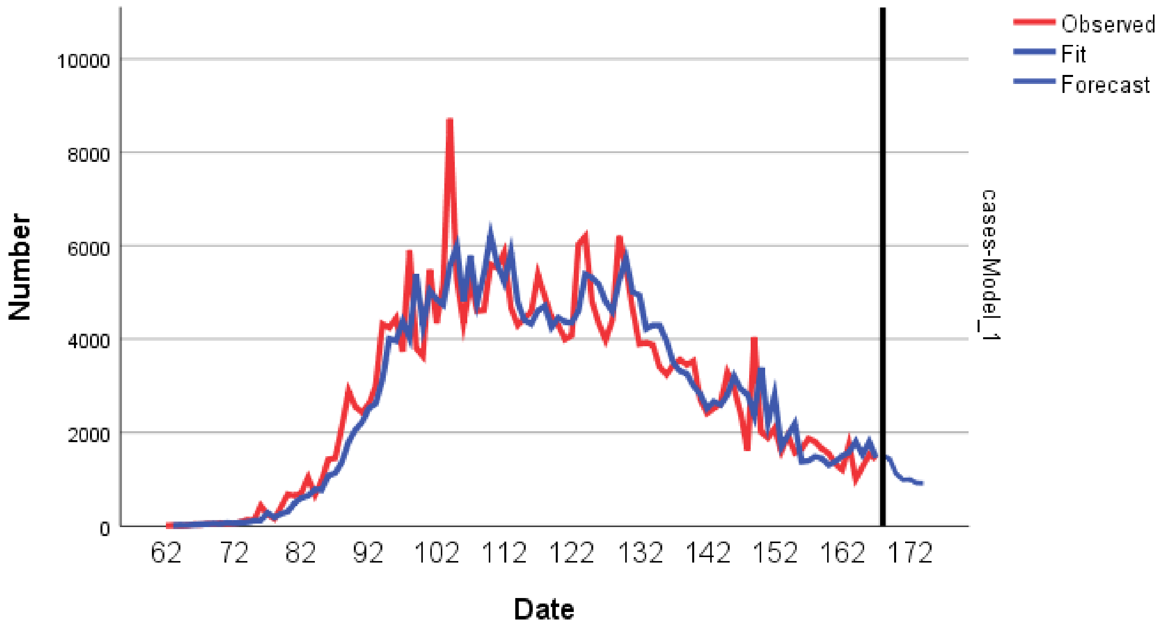
					Estimate	SE	t	Sig.
cases-Model_1	cases	No Transformation	Constant		7.121	45.262	.157	.875
			AR	Lag 1	.454	.217	2.092	.039
				Lag 2	-.410	.231	-1.778	.079
				Lag 3	.496	.213	2.334	.022
				Lag 4	-.290	.190	-1.528	.130
			Difference		1			
			MA	Lag 1	1.039	.365	2.848	.005
				Lag 2	-.396	.348	-1.138	.258
				Lag 3	.379	.340	1.114	.268
				Lag 4	-.236	.370	-.638	.525
				Lag 5	-.307	.202	-1.525	.131

Interpretation:

- Ljung-Box Q statistics have a significance value of >0.05 which shows that our model is adequate.
- After trying and testing different models to try and get all lags to be significant, this was the best significant lags I could obtain from spss since the software doesn't allow you to remove lags from the middle. Therefore, I have chosen the best lags based on the lowest achieved MAE.
- The residual plots are behaving like white noise, so we can say that this model is predicting the behaviour of the time series well.



- We can see that the red and blue coincides better than the previous model, since all lags are significant this time, an error will be reduced.



Forecasting

ARIMA(5,1,6)

Forecast

Model		168	169	170	171	172	173	174
cases-Model_1	Forecast	1449	1459	1054	1019	1031	937	1001

ARIMA(4,1,5)

Forecast

Model		168	169	170	171	172	173	174
cases-Model_1	Forecast	1508	1439	1114	987	998	919	916

Portfolio 6: ANN

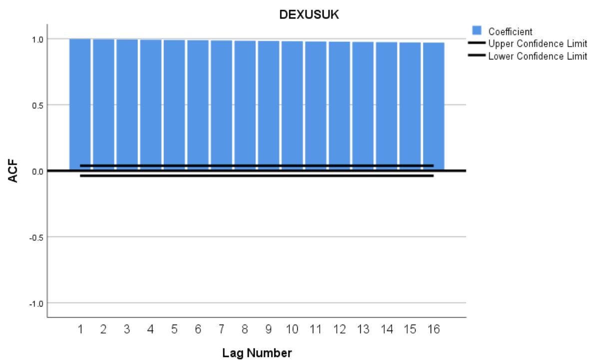
Introduction

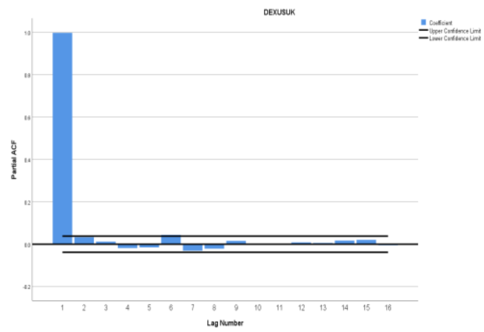
The aim of this report is to apply an artificial neural network algorithm (ANN) on time-series data set that includes the daily exchange rate from January 4, 2010, until August 7, 2020.
We have to predict the value of £/$ for August 8^{th,} 2020, using ANN.

Imputing missing values

The original dataset contained multiple missing values, the way we imputed them is by aggregating the previous 4 days and dividing them by 4.

Model Selection





To analyze the time series, we ran an autocorrelation in order to indicate the autoregressive lags. For ACF, we can see that all the lags are significant while for Partial ACF lag 1 and 6 are significant. Thus, the model that suits this time series data is ARIMA(6,0,0) or AR(6).

We are going to feed our artificial neural network with 6 inputs since there are 6 lags exceeding significance AR(6) and we are going to have 1 output as the values of UK/US exchange. In other words, the UK/US exchange rate at time Yt is going to be our output, and the UK/US exchange rate at time Yt-6, Yt-5, Yt-4, Yt-3, Yt-2, Yt-1 as our input.

Autoregressive model:

$Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + a_3 Y_{t-3} + a_4 Y_{t-4} + a_5 Y_{t-5} + a_6 Y_{t-6} + \epsilon$

SPSS Results

These are the set of inputs we included in our SPSS model to run our neural network model:

Case Processing Summary

		N	Percent
Sample	Training	1363	49.4%
	Testing	725	26.3%
	Holdout	671	24.3%
Valid		2759	100.0%
Excluded		7	
Total		2766	

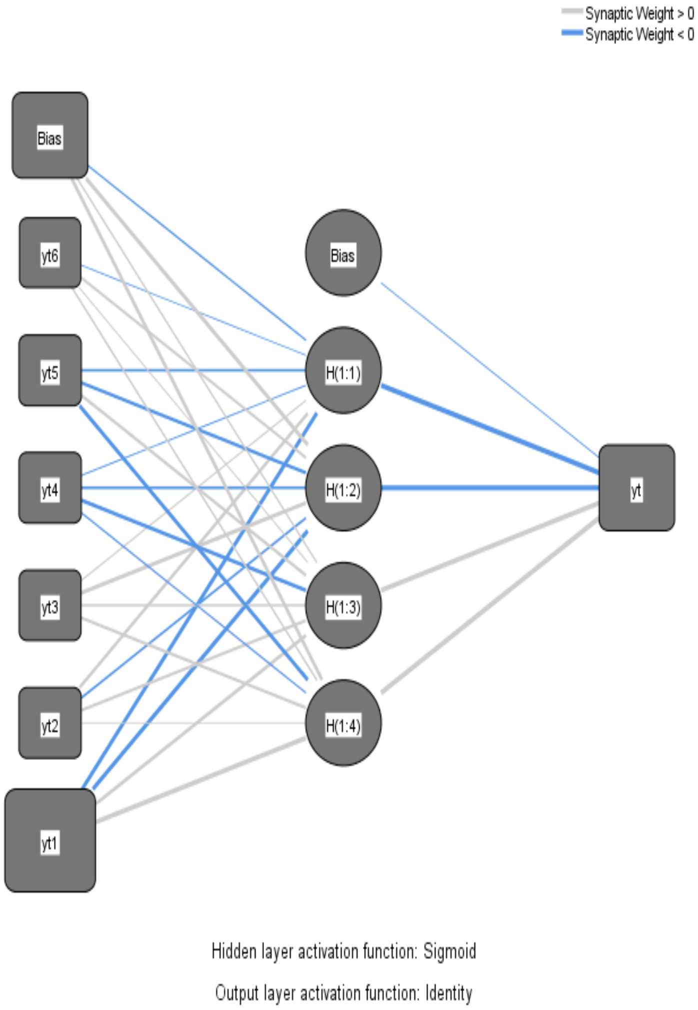
- In partitioning our data set, we have assigned 50% of our data to training, 25% to testing, and 25% of our data going to holdout in order to get an honest estimate of our predictive model.
- There are 7 empty values in our dataset that were excluded by the model, which left us with 2766 records.

Network Information

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

a. Excluding the bias unit

- Standardized methods have been used to rescale our input and output layers for our model to understand the data better.
- Input layer has 6 units while output layer has 1 unit.
- We chose 1 hidden layer with 4 units in the hidden layer that have been automatically chosen by the model.
- The activation function that has been chosen is sigmoid.
- The error function has been defined as the sum of squares.



We let the software to select the number of hidden units, so it selects 4 hidden neurons in the hidden layer as you can see from the network diagram. There is also one bias unit both in the input layer and hidden layer.

The best weights obtained from the training are also reported as part of the output of the model. These weights can be negative or positive on some arcs connecting nodes to the next layer nodes. The blue lines indicate that the synaptic weight that's been assigned is negative while the grey lines are positive weights. Therefore NNAR(6,4) is the neural network describing this time series data. Now that we trained the model, we can forecast the exchange rate for August 8,2020.

Parameter Estimates

Predictor		Predicted				
		Hidden Layer 1				Output Layer
		H(1:1)	H(1:2)	H(1:3)	H(1:4)	yt
Input Layer	(Bias)	-.111	.635	.114	.499	
	yt6	-.027	.201	.073	.143	
	yt5	-.174	-.228	.222	-.584	
	yt4	-.067	-.174	-.413	-.088	
	yt3	.078	.421	.148	.217	
	yt2	.229	-.165	.203	.026	
	yt1	-.638	-.785	.335	.882	
Hidden Layer 1	(Bias)					-.028
	H(1:1)					-1.786
	H(1:2)					-1.751
	H(1:3)					1.444
	H(1:4)					2.000

- These are the optimal synaptic weights that have been assigned to the neural network function. As we can see some of the weights are negative and some are positive.

Model Summary

Training	Sum of Squares Error	2.540
	Relative Error	.004
	Stopping Rule Used	1 consecutive step(s) with no decrease in error[a]
	Training Time	0:00:00.02
Testing	Sum of Squares Error	1.213
	Relative Error	.003
Holdout	Relative Error	.003

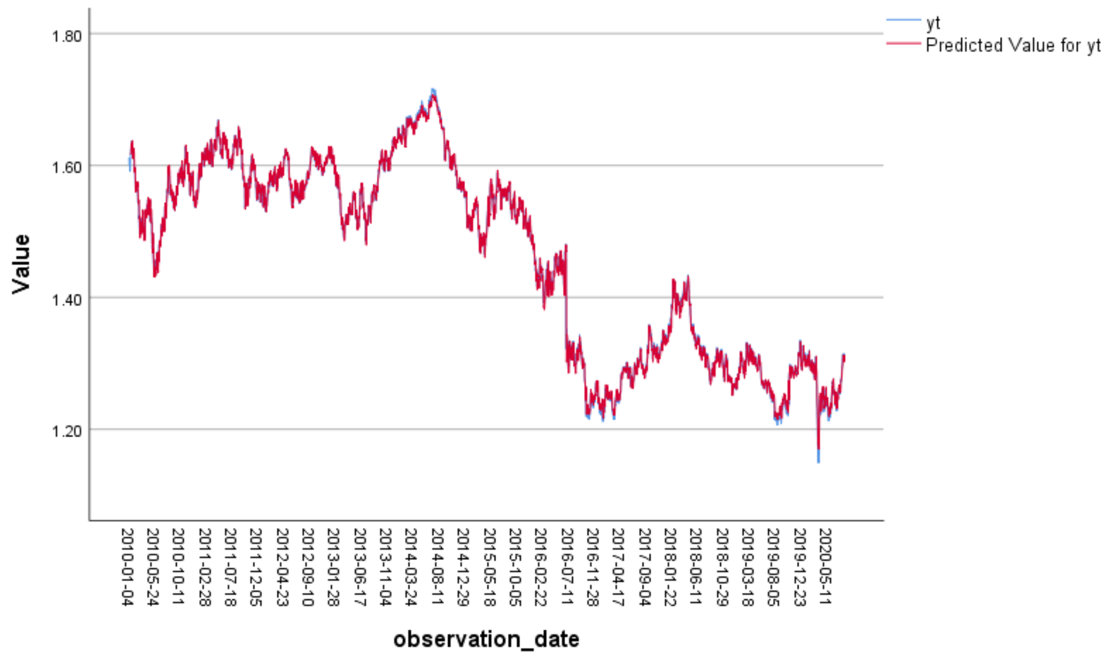
Dependent Variable: yt

a. Error computations are based on the testing sample.

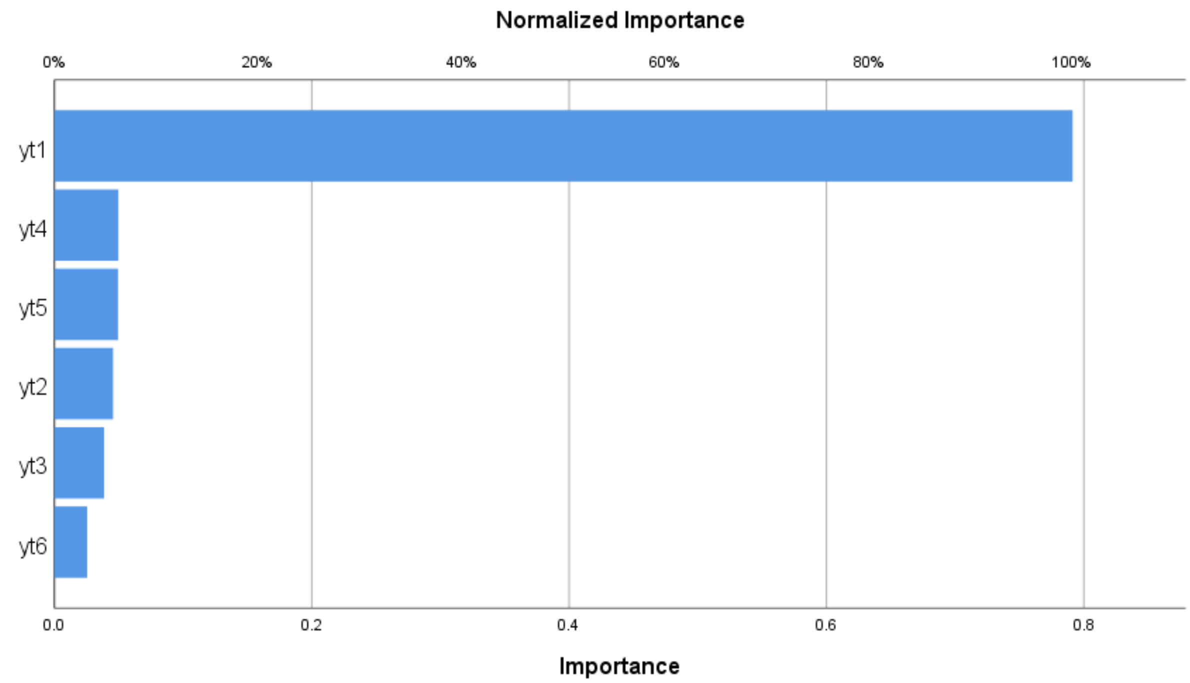
As we can see from the model summary table, the relative error is less than 1% in the training, testing and holdout set which shows that the model NNAR(6,4) works and fits perfectly in predicting the UK/US exchange rates.

Predicted Result achieved:

1.3025 is the predicted value for August 8th, 2020.



From the above line graph, the blue lines show the plots of the original time series while the red line shows the predicted UK/US rates of the time series. We can see that the two plot overlaps perfectly with each other which shows that the model is predicting well.



The above bar graph shows the importance of each independent variable in helping the model predicts. As we can see Yt-1 is the most important independent variable with its importance scale being close to 100% while Yt-6 and Yt-3 being the least important with importance scale being 3.2% and 4.9% respectively.