BUSINESS ANALYTICS IN PRACTICE – PORTFOLIO TASKS

MARCH 25, 2022

STUDENT NUMBER: 210197062

Contents

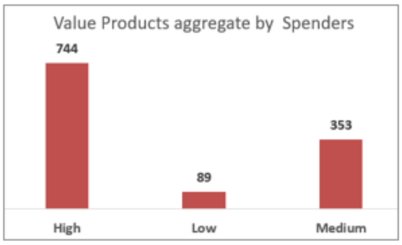
Aston University
Birmingham

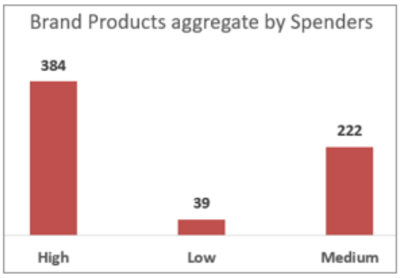
1. Logistic regression for classification

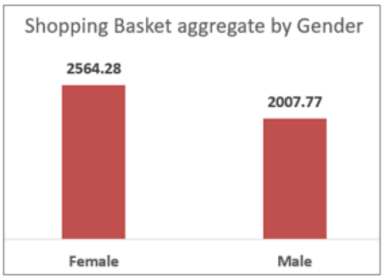
This report aims to provide deep insight into the spending pattern of customers by attempting to create a model which predicts the spending category to which he/she belongs to, based on various factors such as age, gender, type of store, the type of product preferred.

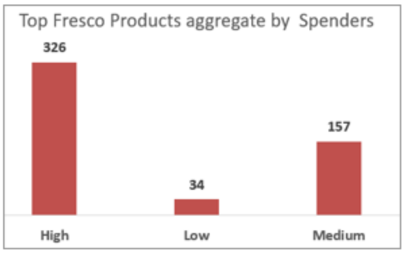
This model shall be a critical game changer to Fresco with regards to revenue generation and customer acquisition and retention as Fresco can concentrate on specific demographics and area which customer are more aligned towards. An example of this is Fresco can curate specific deals by choosing the right type of products - Fresco Top / Brand / Value to attract more high spending customers to generate more revenue. Many more strategies such as above can be employed based on the predictive model which can be crucial to Fresco, hence it is essential that a carefully considered model is created to ensure business value is generated as an outcome of this project

1.1.Summary









From the graphs on the right side, we can understand the relation between the individual variables to the dependent variable – Customer spend pattern. When all these variables are simultaneously considered to form the predictive model, we can create a odds calculator to calculate the probability of a customer being Low / Medium / High spender based on the variables – Age of customer, Value products and Top Fresco product. The final list of variables has been considered based on statistical analysis and process of elimination to retain the most relevant.

1.2. logistic regression – model parameters

For the classification analysis modelling has been performed using logistic regression.

Logistic regression is a supervised machine learning method where dependent and independent variables are defined with hyperparameter Total groups.

1.2.1. Predictor variables

Variables	Remarks
Age	Part of final parsimonious model
Value Products	Part of final parsimonious model
Brand Products	Not part of final parsimonious model
Top Fresco products	Part of final parsimonious model
Gender	Not part of final parsimonious model
Store type	Not part of final parsimonious model

1.2.2. Target variable

Spender type – Low, Medium & High, this variable is derived from another spender basket variable based on the spend levels.

1.3. Baseline category selection

Out of all the available categories, medium category is selected as the reference as it is the most frequently occurring category under spender type

New Spendings Target

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

1.4. Linearity of logistic regression model

The second order and interaction terms of the log converted continuous variables created and model is fitted to check the significance of the variables, here if the significance is greater than 0.05 for the interaction term the assumption of linearity holds good

Parameter Estimates

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

a. The reference category is: Medium.

1.5. Parsimonious model

We include all the variables into our consideration for the building of the model, post our first analysis we eliminate the variable with the highest significance out of the variables having significance greater than 0.05. This is done iteratively until we end up with a model consisting of all significant variables i.e., p value less than 0.05.

The final parsimonious model is as below

Parameter Estimates

New Spendings Target[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B) Lower Bound	Upper Bound
Low	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	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.

As we can see, the significant variables that end in our most parsimonious model are

1. Age
2. Value products
3. Top Fresco products

Ranked in the order of least likely to affect the high spending customer to most likely, keeping medium category as the reference.

1.6. Adequacy tests for the most parsimonious model

1.6.1. Multicollinearity

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: New Spendings Target

Check: VIF to be less than 10

Result: Pass

Check: Tolerance to be greater than 0.1

Result: Pass

1.6.2. Cooks distance

Check: Distance should be less than 1 for the models

Result: All values are less than 1 for COO_1

1.6.3. Standardized residuals

Check: Less than 5% should have absolute value of above 2

Result: Pass, only 3 residuals are above 2

1.6.4. DFBeta's

Check: Less than 1 for all independent variables

Result: Pass

Final conclusion: All adequacy tests seems to be passing, the derived parsimonious model can be deemed adequate

1.7. Goodness of fit

1.7.1. Assumption check – Independence of error

Goodness-of-Fit

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

Check: Ratio of chi-square to df to be less than 2

Result: Pass, value much less than 2

1.7.2. Pseudo R squared

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	22.302[a]	.576	.785

a. Estimation terminated at iteration number 8 because parameter estimates changed by less than .001.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	29.991[a]	.576	.768

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Interpretation: Nagelkerke close to 1, Cox & Snell not close to 1. This is a high-performance model

1.7.3. Hosmer and Lemeshow's test

Hosmer and Lemeshow Test

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

Hosmer and Lemeshow Test

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

Interpretation: Significance is higher than 0.05. Hence this is a high-performance model

1.7.4. Classification accuracy

Classification

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

The accuracy is of very good quality i.e., 82.7%

1.8. Interpretation of predictive model output

The effect of relevant factors for

Customer being a low spender

Age – a unit increase in age shall result in a decreased odd of 0.724 (-27.6%) times the age of the customer, being a low spender with reference being medium spender

Value product - a unit increase in value product shall result in a decreased odd of 0.703(-29.7%) times value product, being a low spender with reference being medium spender

Top Fresco product - a unit increase in age shall result in a decreased odd of 0.559 (-44.1%) times value product, being a low spender with reference being medium spender

Customer being a high spender

Age – a unit increase in age shall result in an increased odd of 1.087 (+8.7%) times the age of the customer, being a high spender with reference being medium spender

Value product - a unit increase in value product shall result in an increased odd of 1.158 (+15.8%) times value product, being a high spender with reference being medium spender

Top Fresco product - a unit increase in value product shall result in an increased odd of 1.524 (+52.4%) times value product, being a high spender with reference being medium spender

1.9. Strategy and Recommendations

1.9.1. **targeting mid and high age** – Due to increase in odds of being a high spender increasing with age, higher age groups can be targeted with more options curation in Top Fresco products and value products, the increase in odds is however not very high compared to other factors

1.9.2. **Increased Value products** – A slight increase in odds – 1.4% of being a high spender is noticed in improvement of value products, hence this may not be the most important variable for attracting customers but significant.

1.9.3. **Increased Top Fresco product** – Most important variable to attract high spenders, working on the Top Fresco line is important for Fresco to ensure good revenue generation.

2. Conjoint analysis

Conjoint analysis study is a product research strategy employed by companies to understand more about consumer preferences and trends and sensitivity to changes in the product specification offerings.

In this report we discuss about a smartphone company looking to optimize its product offering based on the consumer preferences and try to get a deeper insight into the various options being offered under each category to optimize its strategy and successfully gain market share and stay competitive in the market.

2.1. Steps undertaken

2.1.1. Feature selection

After analysis below 4 attributes with respective level are selected

Attributes	**Levels**
Camera	12 Mega Pixel [12MP], 16 Mega Pixel [16MP]
Memory	4 Giga byte [4GB], 6 Giga byte [6GB], 8 Giga byte [8GB] 12 Giga byte [12GB]
Storage	128 Giga byte [128GB], 256 Giga byte [256GB] 512 Giga byte [512GB]
Android version	11.0 [11], 12.0 [12]

2.1.2. Data collection

10 participants – comprising of my friends and family, were asked to rank the various combination of the specification in an excel sheet. Total – 2 x 4 x 3 x 2 = **48 products** Were ranked and taken into consideration

2.1.3. Input data preparation for regression analysis

In this step we create separate columns for each level of the attribute and denote it by 1 – if level part of the product & 0 if level not part of product, to avoid multicollinearity issue while running regression algorithm, the first level of all attributes is dropped from the model.

2.1.4. Regression analysis

Dependent variable: Rank

Independent variable: 16MP, 6GB, 8GB, 12GB, 256GB, 512GB, 12

Regression analysis is done for 10 datasets consisting of the rank order collected from the participants and the standardized coefficient beta is collected. The objective of running the regression analysis is to find the standardized coefficient beta for the coefficients, ultimately what we are achieving here is we are assigning weights to the various variables (levels of each attribute with datum as 1st level). By doing this we can easily understand the priority of each of the levels with respect to each attribute also with respect other levels. The adjusted R-Square for all models is averaging to 0.98 which shows the model is able to successfully predict to a very good percentage the value of ranks.

2.1.5. Data segregation

Averaging is done for each of the coefficient β to find the aggregate value for each of the level representative of the collective participant list.

2.2. Analysis output and interpretation

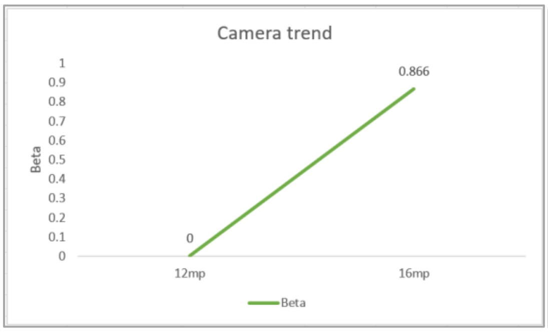
After successfully applying the above steps the aggregate value for each of the level is as follows

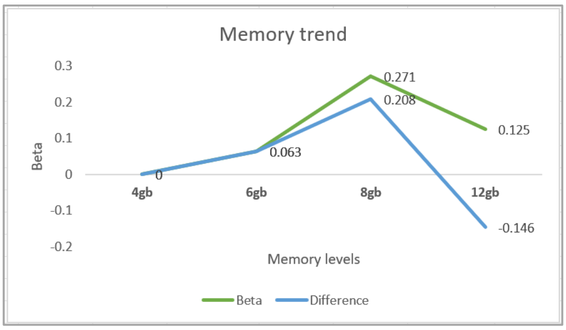
Camera	Beta avg	Difference	Memory	Beta avg	Difference	Storage	Beta avg	Difference	Android version	Beta avg
12mp	0		4gb	0		128gb	0		11	0
16mp	0.866	0.866	6gb	0.063	0.063	256gb	0.476	0.476	12	0
			8gb	0.271	0.208	512gb	0.136	-0.34		
			12gb	0.125	-0.146					

Level	Beta avg	Customer preference rank
16MP	0.86621	1
256GB	0.47639	2
8GB	0.27089	3
512GB	0.13611	4
12GB	0.12503	5
6GB	0.06251	6
12	0	Nil

2.2.1. <u>Camera</u>

β for 16MP is the highest out of all the levels and with respect to other levels under the attribute Camera. This is indicative of the fact that camera is the major factor of consideration for smartphone selection for the focus group

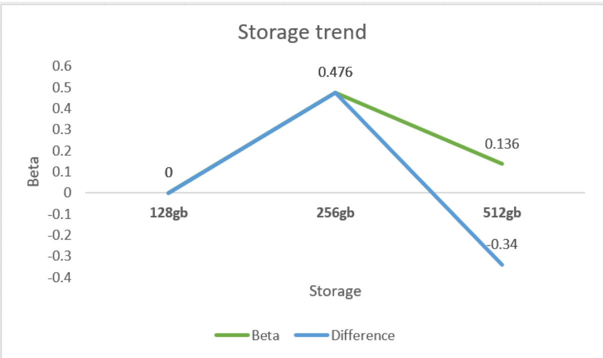




2.2.2. <u>Memory</u>

β is the high for 8gb under the attribute Memory showing this is the most preferred memory capacity preferred for the smartphone, overall, this level ranks 3rd most favourite considering all levels

As seen in the graph, the shift in difference while moving to a higher specification from 8gb to 12gb shows a dip.



2.2.3. <u>Storage</u>

With utility value of 0.476 storage capacity of 256GB ranks the second highest compared to all levels, and it ranks the highest under the storage attribute. This shows that people are concerned considerably about the storage space of their smartphones

There is a major dip in customer preference from 256gb to 512gb, **<u>Android version</u>**

We can see here the utility weight is 0 for android version, this is mainly because customer to identify any improved value with either of the options available and perceive both the versions to be similar in nature.

2.3. <u>Consumer preference fluctuations</u>

Regarding storage space, even though there is a better option i.e., 512GB, they are not interested in this and lean more towards 256GB option, one of the reasons this can be attributed is because of absence of this configuration from majority of the available smartphones and since the population is now moving to cloud this increase may be considered unnecessary by the customers

Regarding memory option, 12GB option provides an increased configuration to the smartphone, however this is not perceived as added value by the customer, which can be interpreted as customer mostly not being aware of the uses of having a higher memory or maybe finding it as a useless feature to have since smartphones are already quite fast with 8GB memory.

Here it is worth noting that regarding smartphone configurations user demographic matters a lot for the consideration and providing weights to the various configurations. For example, older age group or light users of smartphone go for generic options and won't prefer to go with high end options which is generally preferred by heavy users and younger generation.

2.4. Utility value for product combinations

Below is the table of utilities for all possible product combination for the 4 attributes with its respective levels considered. After running correlation analysis, it has been found the aggregate correlation coefficient considering all 10 datasets to be equal to 0.99 which shows strong significant correlation between the ranking and the utility values.

Product Variations	Sum of utilities	Product Variations	Sum of utilities	Product Variations	Sum of utilities
12MP, 4GB, 128GB, 11.0	0	12MP, 8GB, 512GB, 11.0	0.407	16MP, 6GB, 256GB, 11.0	1.405
12MP, 4GB, 128GB, 12.0	0	12MP, 8GB, 512GB, 12.0	0.407	16MP, 6GB, 256GB, 12.0	1.405
12MP, 4GB, 256GB, 11.0	0.476	12MP, 12GB, 128GB, 11.0	0.125	16MP, 6GB, 512GB, 11.0	1.065
12MP, 4GB, 256GB, 12.0	0.476	12MP, 12GB, 128GB, 12.0	0.125	16MP, 6GB, 512GB, 12.0	1.065
12MP, 4GB, 512GB, 11.0	0.136	12MP, 12GB, 256GB, 11.0	0.601	16MP, 8GB, 128GB, 11.0	1.137
12MP, 4GB, 512GB, 12.0	0.136	12MP, 12GB, 256GB, 12.0	0.601	16MP, 8GB, 128GB, 12.0	1.137
12MP, 6GB, 128GB, 11.0	0.063	12MP, 12GB, 512GB, 11.0	0.261	16MP, 8GB, 256GB, 11.0	1.613
12MP, 6GB, 128GB, 12.0	0.063	12MP, 12GB, 512GB, 12.0	0.261	16MP, 8GB, 256GB, 12.0	1.613
12MP, 6GB, 256GB, 11.0	0.539	16MP, 4GB, 128GB, 11.0	0.866	16MP, 8GB, 512GB, 11.0	1.273
12MP, 6GB, 256GB, 12.0	0.539	16MP, 4GB, 128GB, 12.0	0.866	16MP, 8GB, 512GB, 12.0	1.273
12MP, 6GB, 512GB, 11.0	0.199	16MP, 4GB, 256GB, 11.0	1.342	16MP, 12GB, 128GB, 11.0	0.991
12MP, 6GB, 512GB, 12.0	0.199	16MP, 4GB, 256GB, 12.0	1.342	16MP, 12GB, 128GB, 12.0	0.991
12MP, 8GB, 128GB, 11.0	0.271	16MP, 4GB, 512GB, 11.0	1.002	16MP, 12GB, 256GB, 11.0	1.467
12MP, 8GB, 128GB, 12.0	0.271	16MP, 4GB, 512GB, 12.0	1.002	16MP, 12GB, 256GB, 12.0	1.467
12MP, 8GB, 256GB, 11.0	0.747	16MP, 6GB, 128GB, 11.0	0.929	16MP, 12GB, 512GB, 11.0	1.127
12MP, 8GB, 256GB, 12.0	0.747	16MP, 6GB, 128GB, 12.0	0.929	16MP, 12GB, 512GB, 12.0	1.127

3. Clustering analysis

This report aims to support a UK banks product development team by helping them create segments from their customer base based on trends and insights – Dependent variables, to curate targeted products for these customers.

Clustering analysis is conducted on the provided dataset to identify the hidden segments of customer base using statistical software SPSS to create different numbers of clusters using multiple methods, ultimately an ideal number of customer group is decided with justification.

3.1. Step 1: Data preparation

The input for clustering analysis conducted using SPSS software shall consider only numerical variables. When the model is run without considering the categorical variables present in our data, biased frequency proportion is obtained i.e., values being concentrated into 1 group making other groups sparse. Hence it is necessary that we convert the categorical variables into numeric by assigning numbers to the various levels.

The provided dataset consists of 5 categorical variables. We convert these variables into numerical by assigning numerical values for the various categories.

We also standardize the values from a scale of 0 to 1 also as part of our data preparation process since the scales of the variables vary from 1000s to 10s, this is done to improve the performance of the clustering model.

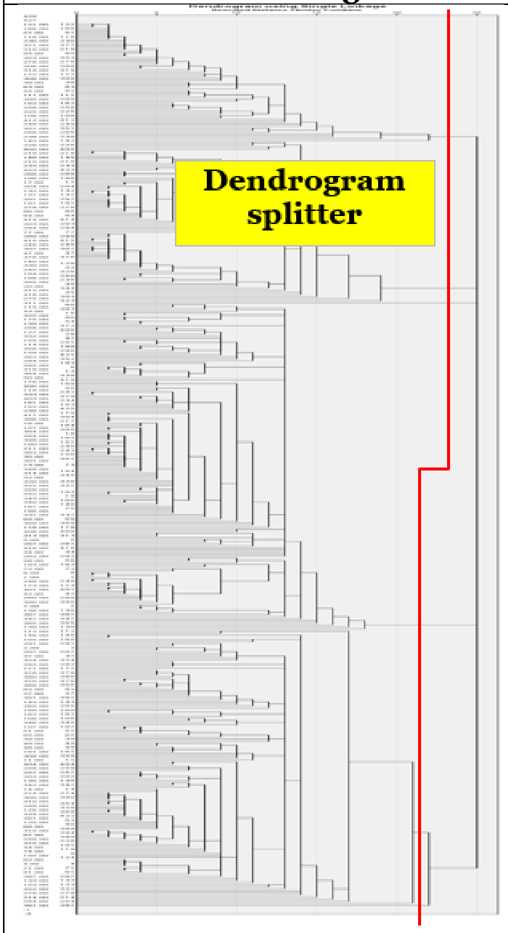
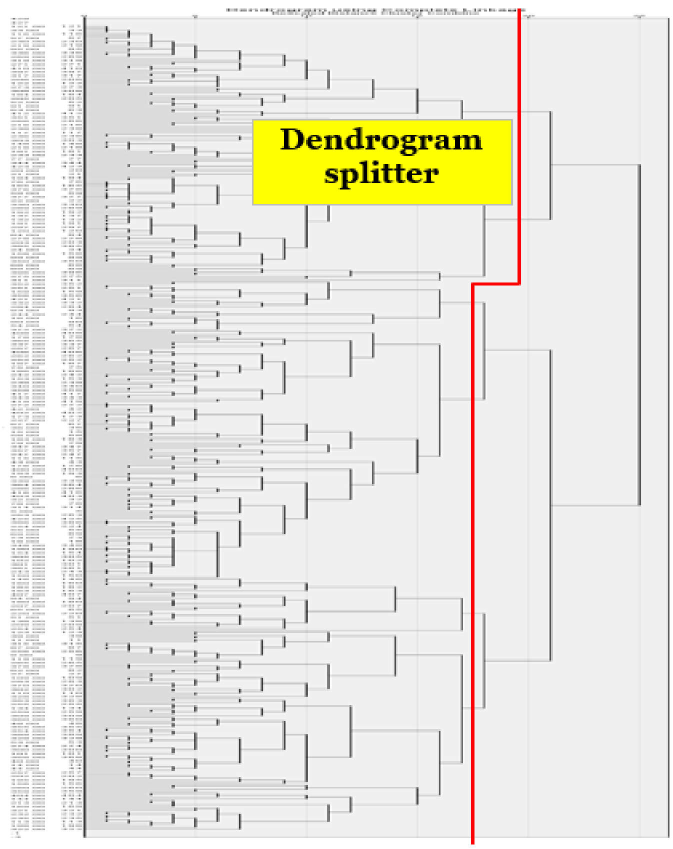
3.2. Step 2: Model parameter selection

Modelling parameters selected

Model 1 – Nearest neighbour		**Model 2 – Furthest neighbour**	
Parameter	**Value**	**Parameter**	**Value**
<u>Cluster method</u>	Nearest neighbour	<u>Cluster method</u>	Furthest neighbour
<u>Measurement distance</u>	Euclidean distance	<u>Measurement distance</u>	Euclidean distance
<u>Value transformation</u>	0 – 1 Standardization	<u>Value transformation</u>	0 – 1 Standardization

3.3. Step 3: Dendrogram analysis for cluster estimation

The output that is gained after running the model with above parameters is a dendrogram which is of importance to determine the total clusters that can be considered as part of the primary analysis

Model 1 – Nearest neighbour	**Model 2 – Furthest neighbour**
	
Output interpretation	**Output interpretation**
The dendrogram has been split in this way to include 6 groups, this seems logical visually as this seems to include the clusters with fair proportion, this shall be the input for next iterative step	The dendrogram has been split in this way to include 6 groups, this seems logical visually as this seems to include the clusters with fair proportion, this shall be the input for next iterative step

3.4. Step 4: Iteratively creating clusters from step 3 output

6 clusters is considered as starting point from which the number of clusters is reduced. The frequency of proportions in each of these clusters is then analysed in step 5 to decide the best clusters

Model 1 – Nearest neighbour

6 Group

Group	Percent
Group 1	36.9
Group 2	31.1
Group 3	18.1
Group 4	13.4
Group 5	0.2
Group 6	0.2
Total	100.0

5 Group

Group	Percent
Group 1	36.9
Group 2	31.3
Group 3	18.1
Group 4	13.4
Group 5	0.2
Total	100.0

4 Group

Group	Percent
Group 1	**36.9**
Group 2	**31.3**
Group 3	**18.4**
Group 4	**13.4**
Total	100.0

3 Group

Group	Percent
Group 1	36.9
Group 2	31.3
Group 3	31.8
Total	100.0

2 Group

Group	Percent
Group 1	68.2
Group 2	31.8
Total	100.0

Model 2 – Furthest neighbour

6 Group

Group	Percent
Group 1	36.9
Group 2	6.6
Group 3	24.7
Group 4	16.7
Group 5	13.4
Group 6	1.6
Total	100.0

5 Group

Group	Percent
Group 1	36.9
Group 2	6.6
Group 3	24.7
Group 4	18.4
Group 5	13.4
Total	100.0

4 Group

Group	Percent
Group 1	**36.9**
Group 2	**31.3**
Group 3	**18.4**
Group 4	**13.4**
Total	100.0

3 Group

Group	Percent
Group 1	68.2
Group 2	18.4
Group 3	13.4
Total	100.0

2 Group

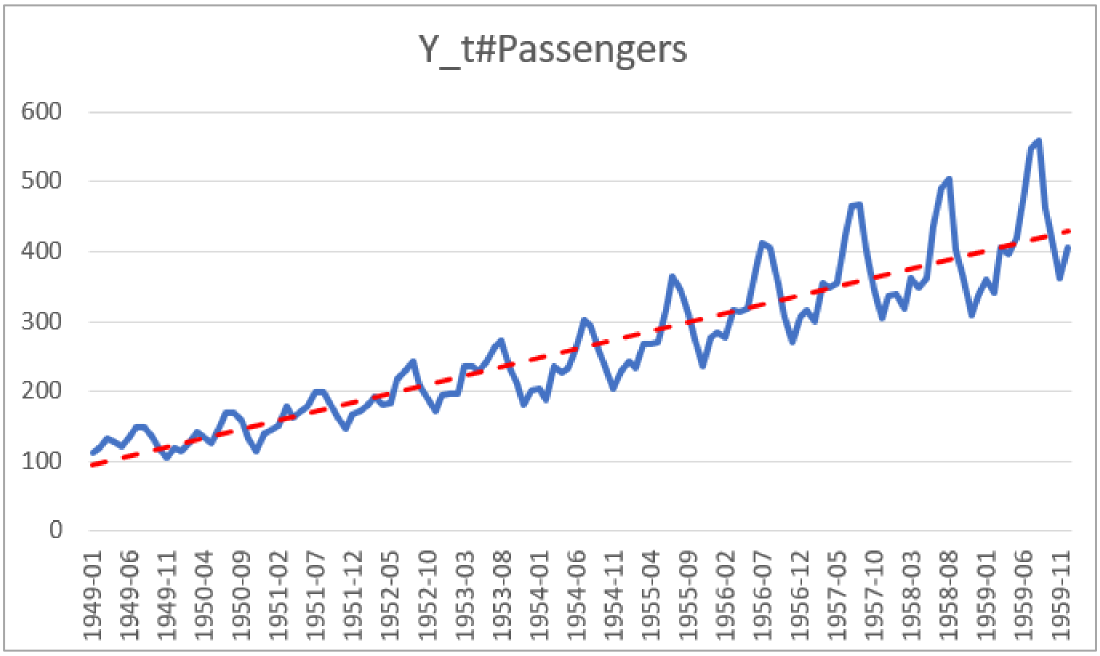
Group	Percent
Group 1	68.2
Group 2	31.8
Total	100.0

3.5. Step 5: Selecting the best number of cluster groups

Based on the above frequency table starting from 6 clusters until 2 clusters. We can see that for clusters 5 and 6 there are values less than 10%. This is not ideal seeing from the perspective of the bank as they will end up developing a product for a relatively small population which may not be financially feasible.

Hence, we select the total clusters to be 4 as surprisingly for both methods considered i.e., Nearest neighbour and Furthest neighbour. The proportion of values residing in all 4 groups are distributed fairly with less bias towards a single group.

4. Time series



4.1. Trend

The time series data taken into consideration has a trend as it is found to have a consistently increasing mean over time.

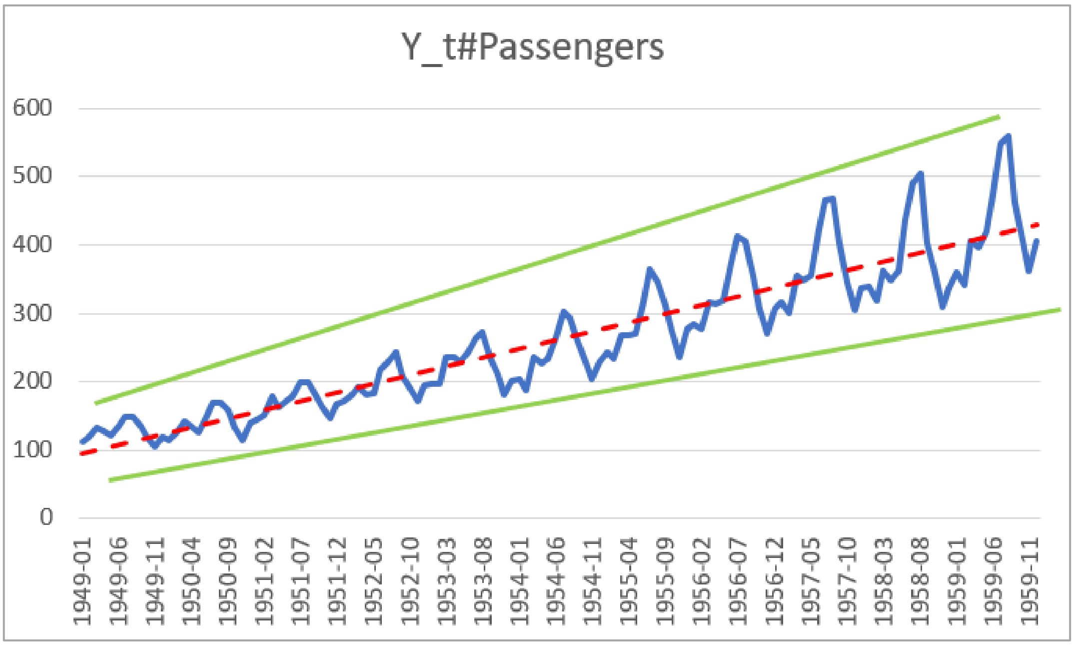
4.2. Seasonal component

There is also a seasonal component which is apparent by the cycles observed in the time series with fluctuation at the same frequency.

4.3. Total seasons

1 – January	5 – May	9 - September
2 – February	6 – June	10 – October
3 – March	7 – July	11 - November
4 – April	8 – August	12 - December

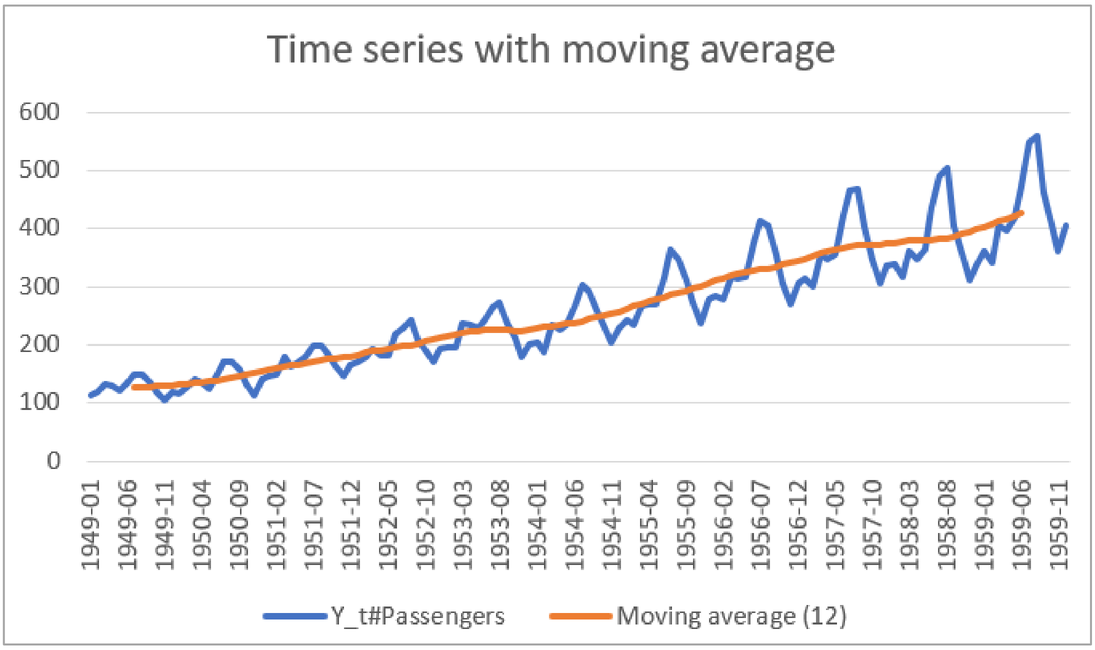
4.4. Model type interpretation



As we can see from Graph 2. The magnitude of seasonal component is changing over time, that is the moving average is increasing (Graph 3), shown by green line. Hence, we can conclude this is a **multiplicative** model.

4.5. Moving average and seasonal value

The moving averages for the time series is calculated by considering time period of 12. The graph obtained with the moving average value is shown below with an upward trend



Seasonal value that is impacting every month is shown in the table below

Months	Typical SF	% Impact
January	0.910004	-9%
February	0.887377	-11%
March	1.018204	2%
April	0.975412	-2%
May	0.979813	-2%
June	1.11159	11%
July	1.222147	22%
August	1.213596	21%
September	1.060917	6%
October	0.921767	-8%
November	0.800213	-20%
December	0.898962	-10%

The interpretation of seasonal factor is as below

1. The months January, February, April, May, October, November, December have a negative impact due to seasonal factor, that is without this seasonal effect the number of passengers would experience an increase

2. The months March, June, July, August, September has a positive impact due to seasonal factor, that is with this seasonal effect the performance is found to increase with regards to number of passengers

3. The most negatively affected month due to seasonal factor is November, where around 20% of total passenger number is affected by this

4. The most positively affected month due to seasonal factor is July, where around 22% of total passenger gain is experienced due to seasonal factor

4.6. Forecasting the number for year 1960

The seasonal factor is used to derive the de-seasonalized data for the historic data, using this de-seasonalized passenger data as dependent variable – y and the time period as independent variable – x, the intercept and the slope is calculated based on the best fit regression line. The intercept and slope are then used to calculate the predicted value of the de-seasonalized number of passengers for the next in series time period – x. Post this the seasonalized prediction is calculated by multiplying the seasonal factor to the de-seasonalized prediction. The predicted output table is as below

Month	Prediction	Month	Prediction	Month	Prediction
1960-01	393	1960-05	433	1960-09	480
1960-02	386	1960-06	495	1960-10	420
1960-03	445	1960-07	547	1960-11	366
1960-04	429	1960-08	546	1960-12	414

4.7. Time series forecasting performance

Mean absolute error – is calculated by dividing the summation of absolute difference over the total observation. When this metric is calculated the result is obtained to be **34**

Mean squared error – is calculated by taking the square root of the division of absolute error squared by total observation. The output of this metric calculation is found to be - **1501**

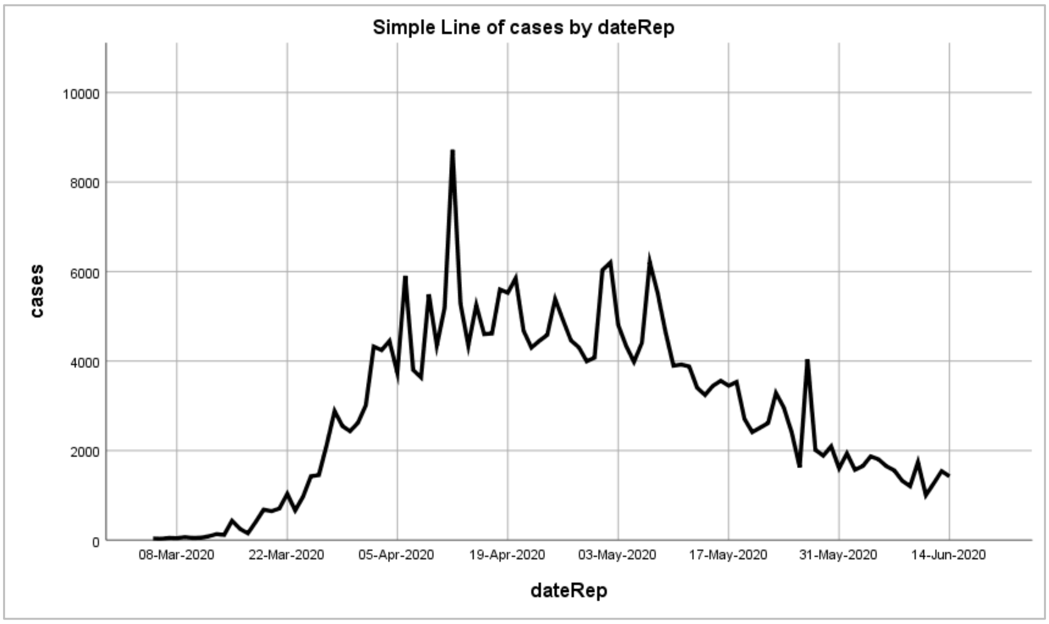
5. ARIMA models

To create a time series model and successfully predict the number of covid cases in the UK from June 15th - 2020 to June 21st - 2020 – 7 days.

The report shall outline various steps involved in validating the dataset and providing justification to the iterative steps involved in selecting the best ARIMA model for prediction.

5.1. Stationarity analysis

5.1.1. Visual inspection of time series graph



Looking at this graph it can be interpreted that the trend of cases i.e., mean of the time series is upward at the beginning of time period with relatively stationary at the middle and a downward trend towards the end of the time period. This changing trend is indicative of a non-stationary behaviour of the time series.

5.1.2. Handling error in data

Upon initial inspection it was found that there is a data point with negative value for cases. This is an anomaly in the data as logically it is not possible to have negative value for the cases. We handle by performing a linear interpolation to ensure correct continuity in the graph

5.1.3. Autocorrelation plots to understand stationarity

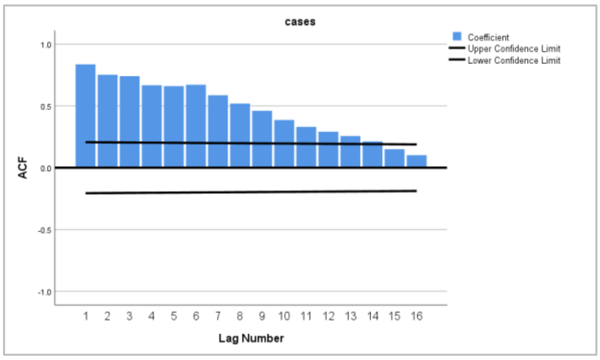


Figure 2 - ACF plot

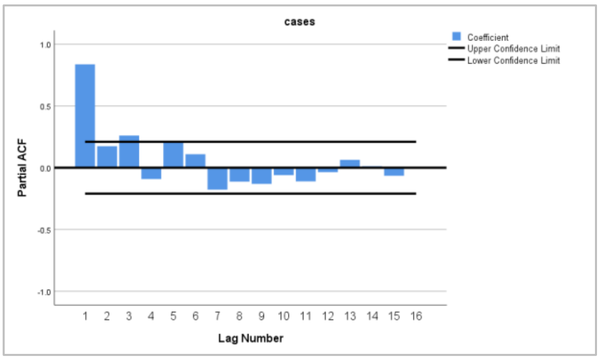


Figure 3 - PACF plot

By studying the ACF plot it is apparent that all the spikes are significant and there is a gradual decrease

Studying the PACF plot we can see the first spike is very significant – close to 1, from the next leg they are all 0

These 2 observations are indicative of the fact that the time series data is not stationary

5.2. Treating nonstationary data

It is important for our model to be stationary to successfully be able to predict the future number of cases, since stationarity in a general sense means statistical properties of the of the time series generator is not changing over time.

We can make a time series data stationary using a process called as differencing, what this does is it stabilizes the mean of the time series data by cancelling the changes at the level of time series. This hereby eliminated seasonality and trends in the time series data.

5.3. The ARIMA model

5.3.1. Parameter selection

ARIMA model has 3 key parameters – p, q and d and is generally defined by ARIMA(p, d, q)

p – Order of auto regressive part

d – Number of differences required for making model stationary

q – Order of moving averages

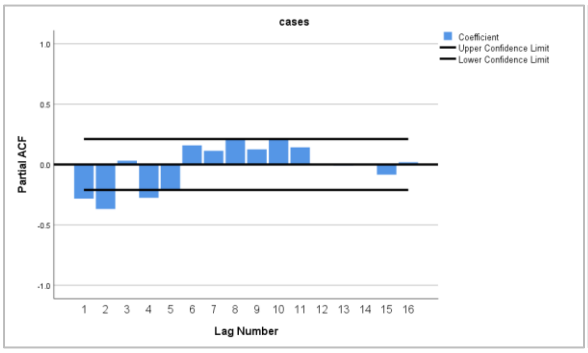


Figure 4 - Differenced PACF plot

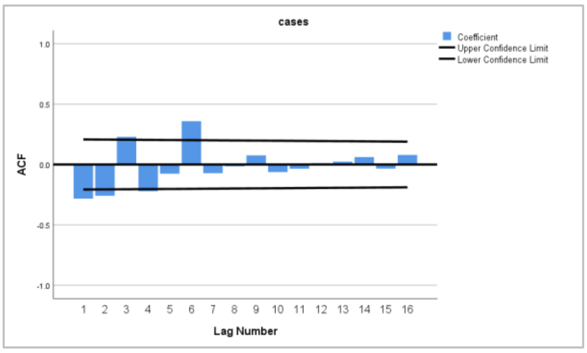


Figure 5 - Differenced ACF plot

From PACF plot we can see the order is 4 as there are 4 values of lag that are significant, hence the p value that can be considered is equal to 4

From ACF plot we can see the order is 6 as there are 6 values of lag that are significant, hence the q value that can be considered is equal to 6

The time series has been differenced once to introduce stationarity, hence the d value that can be considered is equal to 1

5.4. Best model selection

Iteratively the values for the p and q are reduced to understand model behaviour with respect to mean absolute error, Ljung box statistic for model adequacy and residual plots.

Below was the observation for the iterative process

Model	p	d	q	Ljung box	MAE	Final p significant ?	Final q significant ?	Residual significance	Remarks
ARIMA	4	1	6	0.937	459	Yes	No	Insignificant	
ARIMA	4	1	5	0.728	456.3	Yes	Yes	Insignificant	
ARIMA	4	1	4	0.464	461.6	No	Yes	Insignificant	
ARIMA	4	1	3	0.075	472	Yes	No	Significant	Reject due to residual significance
ARIMA	3	1	6	0.953	461.9	No	Yes	Insignificant	
ARIMA	3	1	5	0.518	458.88	Yes	No	Insignificant	
ARIMA	3	1	4	0.034	490.12	No	No	Significant	Reject due to residual significance and adequacy fail
ARIMA	2	1	6	0.955	458.7	Yes	No	Insignificant	
ARIMA	2	1	5	0.63	451	Yes	No	Insignificant	Lowest MAE value out of all models
ARIMA	2	1	4	0.088	488.91	No	Yes	Significant	Reject due to residual significance

5.4.1. Interpretation and model selection

From the above table it can be observed that the lowest MAE has been obtained for ARIMA (2,1,5) model, also this model is found to satisfy the model adequacy with Ljung box statistic value greater than 0.05, and residuals are found to be in the insignificant range.

5.5. Justifying model selection

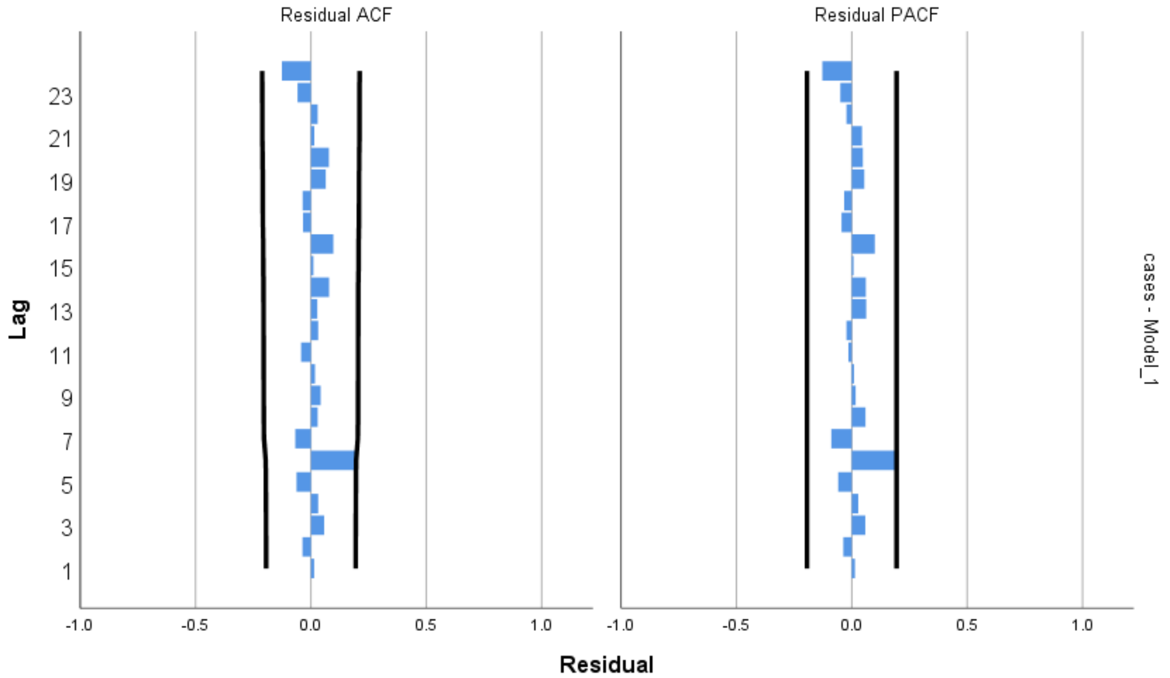
5.5.1. Ljung box test for adequacy

Model Statistics

		Model Fit statistics			Ljung-Box Q(18)			
Model	Number of Predictors	Stationary R-squared	RMSE	MAE	Statistics	DF	Sig.	Number of Outliers
cases-Model_1	0	.380	699.340	451.386	8.917	11	.630	0

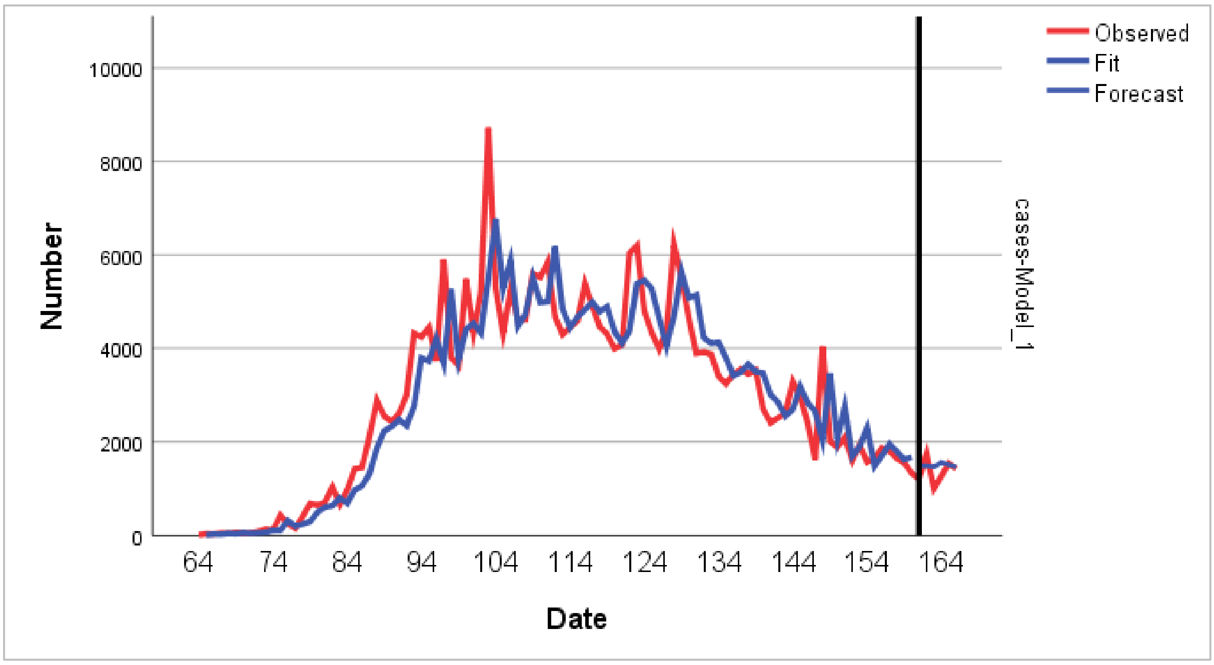
As we can see from the table, Ljung box statistic is above the threshold 0.05 hence confirming adequacy for the model

5.5.2. Residual plot for ACF and PACF



As we can observe from the graph the residuals fall under the insignificant range, hence confirming this to be a good model

5.5.3. Observed vs Fit graph



As we can see from the above graph we can see that there is a good amount of fit observed against the observed data, also the model was tested for forecasting. Which is denoted by the line

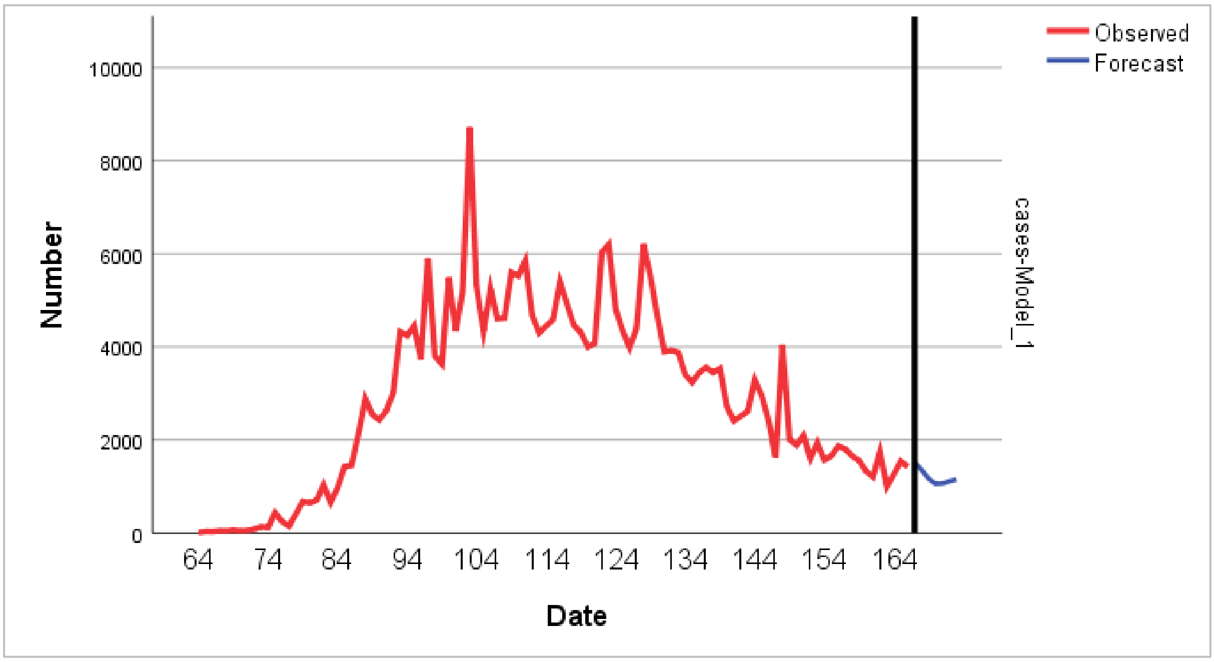
5.6. Estimation of covid cases from June 15 to June 21 – 2021

Estimated values table

Date	Cases
15-Jun-20	1511
16-Jun-20	1357
17-Jun-20	1164
18-Jun-20	1055
19-Jun-20	1063
20-Jun-20	1113
21-Jun-20	1149

Prediction quality

The prediction obtained is of moderate quality especially due to the high MAE of approximately 400 cases. However, the model is found to fit with an acceptable range of overlap (observed vs fit graph section 3.3) and continuity as seen below



5.7. More parsimonious model

Due to the shortcomings of the SPSS software i.e., inability to remove lags in between, we end up with model consisting of insignificant variables as seen below

ARIMA Model Parameters

					Estimate	SE	t	Sig.
cases-Model_1	cases	No Transformation	AR	Lag 1	.645	.285	2.264	.026
				Lag 2	-.337	.240	-1.404	.164
			Difference		1			
			MA	Lag 1	1.248	.275	4.534	.000
				Lag 2	-.515	.382	-1.348	.181
				Lag 3	-.089	.191	-.469	.640
				Lag 4	.319	.170	1.878	.063
				Lag 5	-.407	.120	-3.394	.001

The most parsimonious model if allowed to eliminate the respective lags would then be

For Autoregressive part – Lag1 (p = 1)

For Moving average part – Lag1, Lag2 (q = 2)

6. Artificial Neural Network

The report aims to predict the exchange rate for August 8, 2020 using Artificial Neural Network – Supervised Machine learning algorithm.

The key steps and its interpretation are also highlighted in this report to justify decisions and the quality of the model is also assessed.

6.1. Data quality

Post analysis around 110 values were found to be missing from the data, this amounts to around 4% of the data. Hence it is crucial that we somehow handle this to ensure good performance of the model and ensure data sanctity is maintained.

The missing data was then imputed using linear interpolation, by averaging the previous value and the next value to match the trend and ensure time series is continuous.

6.2. Input and output variable specification

The input variables are selected based on an analysis conducted on the autocorrelation plots.

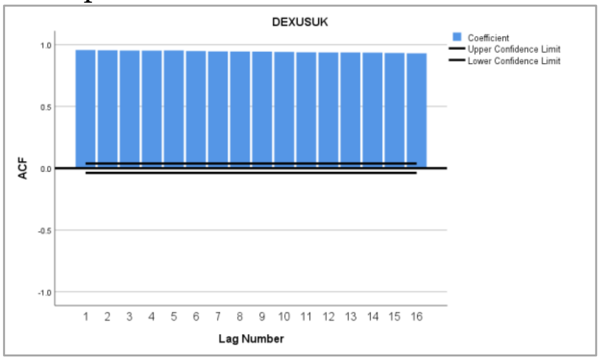


Figure 1

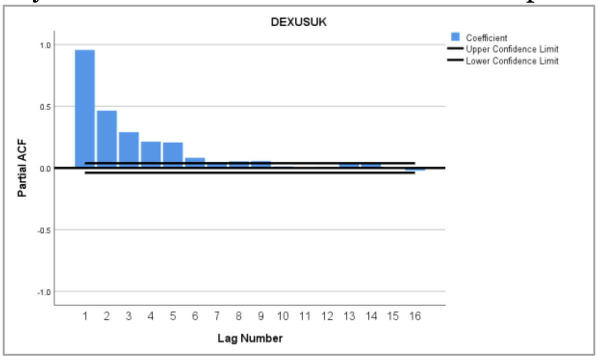


Figure 2

Looking at the above correlation plots we can come to conclusion regarding input values; all values are significant in the ACF plot and there is a gradual decrease in the trend. Looking at the PACF plot we can understand that there are 6 lag values which are significant, these lag values contribute to the input for the neural network.

The input variables that are considered based on this are Y_{t-6}, Y_{t-5}, Y_{t-4}, Y_{t-3}, Y_{t-2}, Y_{t-1}. These inputs are the lagged values of the same data lagged by the value of 6, 5, 4, 3, 2, 1 respectively.

Since this is a supervised learning method, we set the target variable as the variable the exchange rate itself.

6.3. Model training and output interpretation

Multilayer perceptron model is trained with 1 hidden layer for the prediction, the ratio of the data split between training, testing and holdout is set to be 2:1:1

6.3.1. Output tables

Case Processing Summary

		N	Percent
Sample	Training	1414	51.3%
	Testing	675	24.5%
	Holdout	670	24.3%
Valid		2759	100.0%
Excluded		12	
Total		2771	

This table shows the actual split of the data between Training, Testing and the holdout operations. The total values that are excluded due to missing are 739

Network Information

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

a. Excluding the bias unit

Table 2

The above table summarizes the key parameters of the machine learning model, such as the total hidden layers, activation function for hidden layer, total units in input, hidden and output layer etc. and shows the scaling method used for handling the inputs

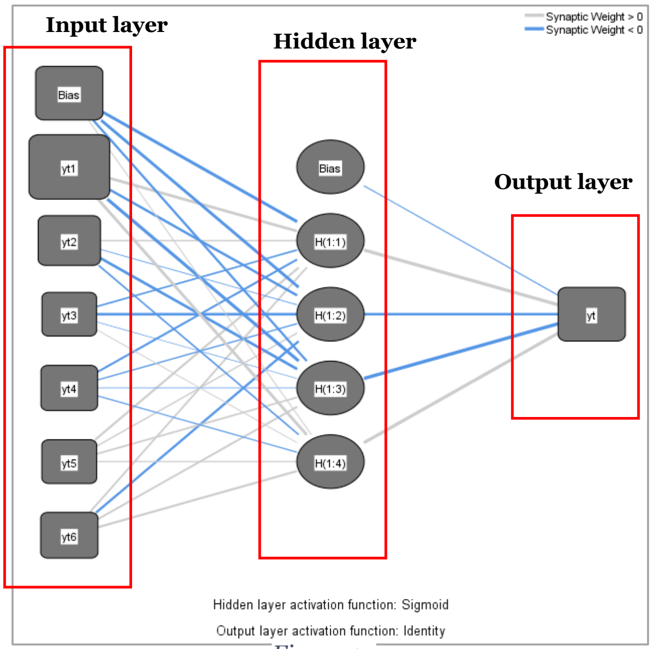


Figure 3

Visual representation of a Multilayer perceptron, this diagram shows the intricate ways in which the neural network relates to the input and the output to provide the optimal output

Model Summary

Training	Sum of Squares Error	2.563
	Relative Error	.004
	Stopping Rule Used	1 consecutive step(s) with no decrease in error[a]
	Training Time	0:00:00.01
Testing	Sum of Squares Error	1.096
	Relative Error	.003
Holdout	Relative Error	.004

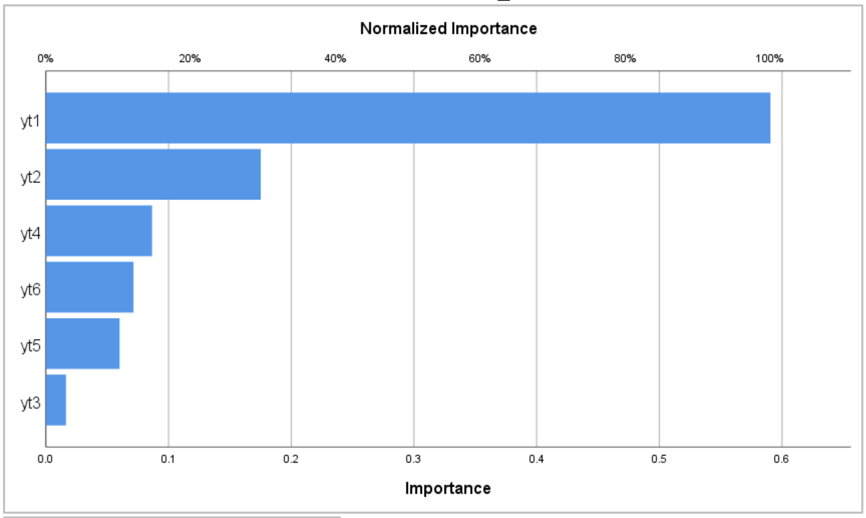
Dependent Variable: yt

a. Error computations are based on the testing sample.

this is a crucial table that talks about the performance of the created model. The created model has around 4% of error for training, testing and holdout which is good performance

Figure 4

6.4. Variable importance



Independent Variable Importance

	Importance	Normalized Importance
yt-6	.071	12.1%
yt-5	.060	10.2%
yt-4	.087	14.6%
yt-3	.016	2.8%
yt-2	.175	29.6%
yt-1	.591	100.0%

Based on the importance value identified for the input variables by the multilayer perceptron it can be interpreted that the most importance input variable is the first lag Y_{t-1} followed by Y_{t-2}, Then Y_{t-4}, Y_{t-6}, Y_{t-5}, Y_{t-3} . This means the strongest variable are more contributing to the predictive performance of the model.

6.5. Validating model performance

The model performance can be evaluated by checking its prediction power against existing values. This is done by visually overlapping the observed values against fit values

From the above graph it can be said that the overlap of the predicted against observed values are very apparent. Hence it can be concluded that the model is good and has strong predictive power.

6.6. One step ahead forecast

the prediction value of exchange rate for **August-8-2020** is **1.3044**