Analysis of Social Media Impact on Individuals using Logistic Regression.

Principal Component Analysis on Football Players Attributes.

Time Series Analysis on Supply,Transformation and Consumption of Electricity

I. LOGISTIC REGRESSION

In the previous Assignment, we demonstrated the use of Multiple Regression to Predict the Dependent Variable which was Continuous. However, in a real word situation Dependent Variable could be Categorical. Multiple Regression is not suitable when the dependent Variable is Categorical. In this case, Logistic regression allows you to predict a categorical variable of two or more categories. In this regression, we predict the class or category of an observation which is also called classification. Independent variables could be of Qualitative as well as Quantitative type. Depending on the Number of Categories of Dependent Variable Logistic Regression is classified as Binary Logistic Regression with 2 categories and Multinomial Logistic Regression with more than 2 categories.

II. STATEMENT OF OBJECTIVES AND DESCRIPTION OF VARIABLES

Social Media has evolved significantly over the past decade helping mankind connect easily and rapidly. Every Technological Advances has its Merits and Demerits, similarly, Social Media has also created Positive and Negative Impact on individuals. Using the Logistic Regression model, we would be predicting the effect of Social Media on an individual.

The Dataset is taken from the Pew Research Center Website. Initial Dataset was containing missing values which were removed before starting the regression process.

Dependent Variable:

What effect would you say social media has had on you?
Categories:
1) Positive (1)
2) Negative (2)

Variable is named as Positive/Negative in the dataset.

Independent Variables:

Continuous Variables:

1) Age Range (11-30)

Categorical Variables:

2) Does social media make you feel more ...?
1) Authentic (1)
2) Fake (2)

Variable is named as Authentic/Fake in the dataset.

3) Does social media make you feel more…?
1) Confident (1)
2) Insecure (2)

Variable is named as Confident / Insecure in the dataset.

4) Does social media make you feel more…?
1) Reserved (1)
2) Outgoing (2)

Variable is named as Reserved/Outgoing in the dataset.

5) Does social media make you feel more…?

1) Included (1)
2) Excluded (2)

Variable is Named as Included/Excluded in the dataset.

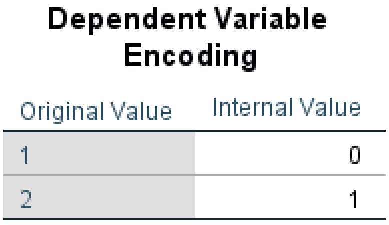
Data source:
https://www.pewresearch.org/internet/dataset/teens-and-tech-survey-2018/

III. ASSUMPTIONS UNDERTAKEN

- **Assumption 1: Mutual Exclusivity of Dependent Variable.**

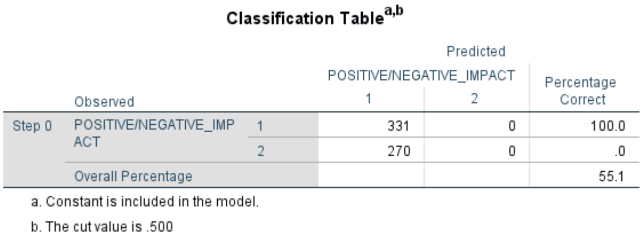
Mutual Exclusivity can be explained as the presence of one category implies the absence of another category. Dependent Variable has two values Positive and Negative quoted as (1 and 2) respectively. There shouldn't be a single observation which contains both the value.

As per the sample of 744 individuals, 331 individuals have reported that social media had a Positive Impact on them which is coded as 1 and the remaining 270 had a Negative Impact coded as 2. There is no such observation where both 1 and 2 are existing.



Dependent Variable Encoding

Original Value	Internal Value
1	0
2	1



Classification Table[a,b]

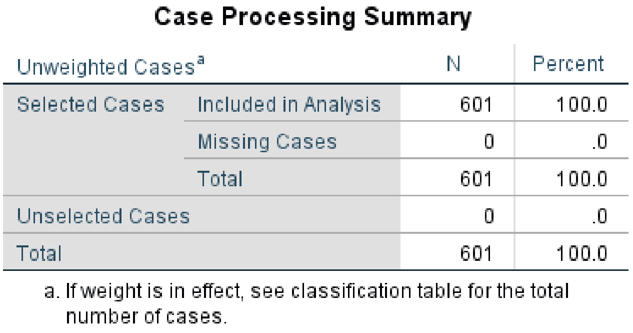
	Observed		Predicted		
			POSITIVE/NEGATIVE_IMPACT		Percentage Correct
			1	2	
Step 0	POSITIVE/NEGATIVE_IMPACT	1	331	0	100.0
		2	270	0	.0
	Overall Percentage				55.1

a. Constant is included in the model.
b. The cut value is .500

- **Assumption 2: Sample Size:**

This assumption is considered to verify whether we have sufficient observations in each variable to perform Logistic Regression. If we are having small samples with a large number of predictors, it may create a problem for analysis [1]. It works best when we have a sample size equal to or more than 20 cases per predictor.

Here, we have 5 predictors and each of them is having 601 observations which are sufficient to perform the Logistic Regression.



Case Processing Summary

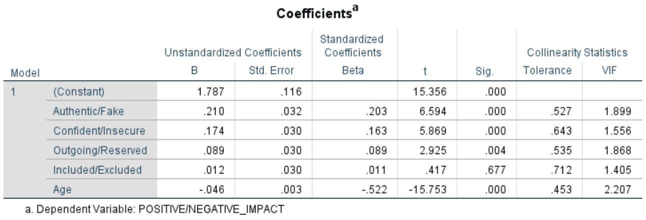
Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	601	100.0
	Missing Cases	0	.0
	Total	601	100.0
Unselected Cases		0	.0
Total		601	100.0

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

- **Assumption 3: Absence for Multicollinearity:**

VIF is a measure of how much common variance is present among independent variables. It is used to check the multicollinearity which is not evident in the correlation Matrix. We need to give attention to a predictor that has VIF greater than 10 as it could be collinear with other predictors. Tolerance is the inverse of the Variance Inflation Factor.

As per the below table assumption of Multicollinearity is met. There are no variables which are having VIF greater than 10.



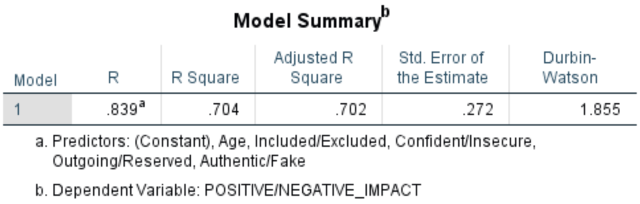
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	1.787	.116		15.356	.000		
	Authentic/Fake	.210	.032	.203	6.594	.000	.527	1.899
	Confident/Insecure	.174	.030	.163	5.869	.000	.643	1.556
	Outgoing/Reserved	.089	.030	.089	2.925	.004	.535	1.868
	Included/Excluded	.012	.030	.011	.417	.677	.712	1.405
	Age	-.046	.003	-.522	-15.753	.000	.453	2.207

a. Dependent Variable: POSITIVE/NEGATIVE_IMPACT

- **Assumption 4: Autocorrelation check between Errors (Independence of Errors)**

As per this assumption error terms (Residuals) shouldn't be correlated. To test this assumption, we use Durbin–Watson statistics. Durbin–Watson statistics value should be in the range of 1 to 3. Better if it is close to 2. Durbin Watson Statistics value is 1.855 hence the assumption of Independence of Residuals is held.



Model Summary[b]

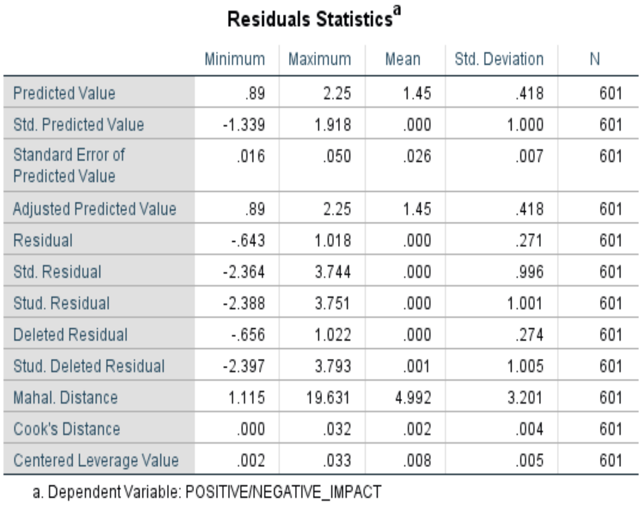
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.839[a]	.704	.702	.272	1.855

a. Predictors: (Constant), Age, Included/Excluded, Confident/Insecure, Outgoing/Reserved, Authentic/Fake

b. Dependent Variable: POSITIVE/NEGATIVE_IMPACT

- **Assumption 5: Outliers (No influential Data Points)**

To check this assumption, we use Cook's Distance. Cooks distance shouldn't be greater than or equal to 1. As per the Residual Statistics, the Min and Max Cook's distance is 0.000 and 0.032 respectively hence proving that there are no outliers.



Residuals Statistics[a]

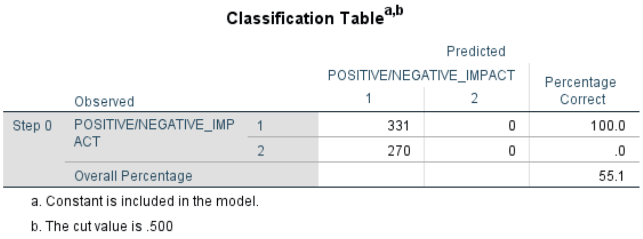
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	.89	2.25	1.45	.418	601
Std. Predicted Value	-1.339	1.918	.000	1.000	601
Standard Error of Predicted Value	.016	.050	.026	.007	601
Adjusted Predicted Value	.89	2.25	1.45	.418	601
Residual	-.643	1.018	.000	.271	601
Std. Residual	-2.364	3.744	.000	.996	601
Stud. Residual	-2.388	3.751	.000	1.001	601
Deleted Residual	-.656	1.022	.000	.274	601
Stud. Deleted Residual	-2.397	3.793	.001	1.005	601
Mahal. Distance	1.115	19.631	4.992	3.201	601
Cook's Distance	.000	.032	.002	.004	601
Centered Leverage Value	.002	.033	.008	.005	601

a. Dependent Variable: POSITIVE/NEGATIVE_IMPACT

IV. UNDERSTANDING AND BUILDING THE MODEL:

Analyze -> Regression -> Binary Logistic option from SPSS is used to define the Dependent Variable, Independent Variables and process the regression.

Block 0:



Classification Table[a,b]

Observed			Predicted		
			POSITIVE/NEGATIVE_IMPACT		Percentage Correct
			1	2	
Step 0	POSITIVE/NEGATIVE_IMPACT	1	331	0	100.0
		2	270	0	.0
	Overall Percentage				55.1

a. Constant is included in the model.

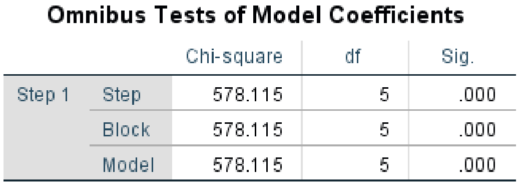
b. The cut value is .500

Block 0 is called a Null / Trivial Model. This shows the Social Media Impact on individuals without considering the Independent Variable. In the further regression process, the model adds Independent Variables which should result in an improved percentage of classification. It indicates that out of 601 observations, 331 people recorded Positive impact while 270 recorded Negative. This model shows an accuracy of 55.1 %.

Block 1:

This block is developed using all the Independent Variables selected. Now we will see tests and Results of Block 1.

Omnibus Test:



Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	578.115	5	.000
	Block	578.115	5	.000
	Model	578.115	5	.000

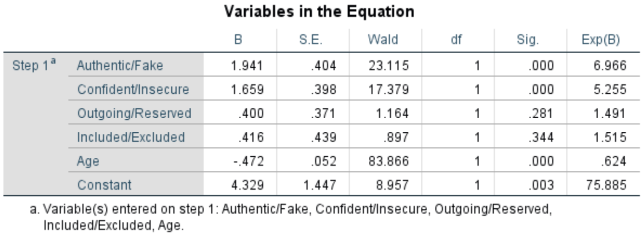
This test is used to check whether the Block 1 model has improved from the Block 0 base model. Here we have a null Hypothesis as all **β** coefficients are equal to zero and an Alternative Hypothesis as no **β** are equal to zero.

H_0: β1 = β2 = …. = $β_K$ = 0

H_1: Not all βs are equal to 0

Looking at Sig. values which are less than 0.05 for all Step, Block, and Model we reject the Null Hypothesis.

Variables in the Equation:



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Authentic/Fake	1.941	.404	23.115	1	.000	6.966
	Confident/Insecure	1.659	.398	17.379	1	.000	5.255
	Outgoing/Reserved	.400	.371	1.164	1	.281	1.491
	Included/Excluded	.416	.439	.897	1	.344	1.515
	Age	-.472	.052	83.866	1	.000	.624
	Constant	4.329	1.447	8.957	1	.003	75.885

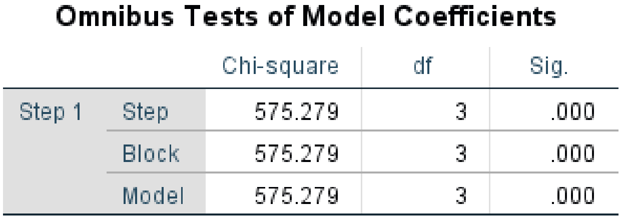
a. Variable(s) entered on step 1: Authentic/Fake, Confident/Insecure, Outgoing/Reserved, Included/Excluded, Age.

As we can see here, the sig value for the Outgoing/Reserved and Included/Excluded variable is greater than 0.05 which indicates that those two variables are not making a significant impact on the model's class prediction. We removed those variables and reprocessed the Logistic regression.

V. MODEL SUMMARY

This is a Model Summary with 3 Explanatory Variables.

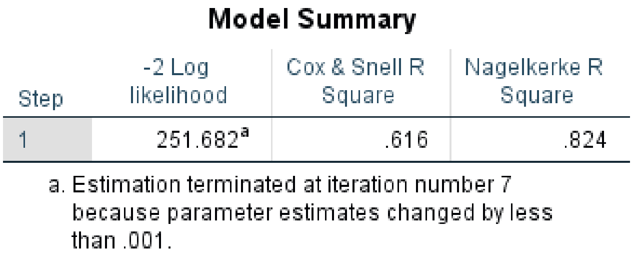
Omnibus Test:



Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	575.279	3	.000
	Block	575.279	3	.000
	Model	575.279	3	.000

We reject the null hypothesis, as all the sig. values are less than 0.05. No **β** is equal to zero.



Model Summary

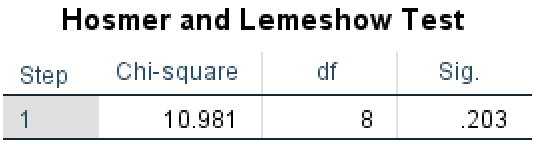
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	251.682[a]	.616	.824

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

-2 Log Likelihood is similar to Residual Standard Error value from Multiple Linear Regression. This value should be as low as possible. This model has 251.682 -2 Log Likelihood.

We cannot compute the R Square value as we do it in Multiple Regression. In Logistic Regression, we have Cox & Snell R Square and Nagelkerke R Square which are the pseudo-R square. They describe the amount of variation in the dependent variable which is predicted by the predictor variables. Theoretically maximum possible value for them is 1 if the relationship is perfect else it would be 0. Here 61.6 % and 82.4% variability in Social Media Impact is explained by predictors.

Hosmer and Lemeshow Test:

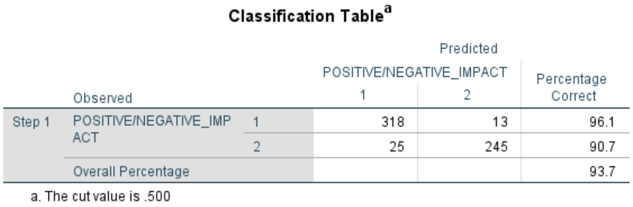


Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	10.981	8	.203

This is used to test the goodness of fit. If the sig. value is less than 0.05 model is poor fit else it is a better fit. Here 0.203 is greater than 0.05 which indicated that the model is a better fit.

Classification Table:



Classification Table[a]

			Predicted		
			POSITIVE/NEGATIVE_IMPACT		Percentage Correct
	Observed		1	2	
Step 1	POSITIVE/NEGATIVE_IMPACT	1	318	13	96.1
		2	25	245	90.7
	Overall Percentage				93.7

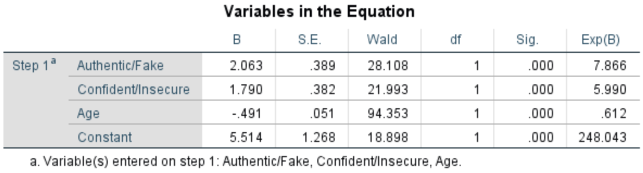
a. The cut value is .500

From this table, we can get many different accuracy insights for the model such as Specificity, Sensitivity, PAC.

93.7 Value in the table indicates the Percentage Accuracy in Classification. It has predicted 93.7 cases correctly.

Out of 331observations that were reported as Positive Impact Model has predicted 318 Correctly as Positive and 13 Incorrectly as Negative which implies the accuracy of 96.1%. 245 were correctly classified as Negative and 25 Incorrectly out of 270 observations which 90.7 % accurate. 96.1% is the specificity of a model whereas 90.7 % is the sensitivity of a model. The model should be able to correctly classify Observations in both the direction that is if the model is good at classify Positive Impact cases then it should also be good at predicting Negative Impact cases as well.

Variables in the Equation:



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Authentic/Fake	2.063	.389	28.108	1	.000	7.866
	Confident/Insecure	1.790	.382	21.993	1	.000	5.990
	Age	-.491	.051	94.353	1	.000	.612
	Constant	5.514	1.268	18.898	1	.000	248.043

a. Variable(s) entered on step 1: Authentic/Fake, Confident/Insecure, Age.

This table is used to find out the contribution of each predictor variable in predicting the class of Social Media Impact. The P values (.sig) implies that all variables are contributing significantly to model prediction as all the variables are having sig. value less than 0.05.

The Wald Value is similar to t statistics in Multiple Regression. It adheres to the Hypothesis that no variables are having a significant impact on the outcome that is all the **β =0**

B values are identical to β coefficients from Multiple linear regression. The sign of the B value indicates the relationship of that independent variable with the dependent variable. The negative sign indicates that the variable is inversely proportional to the dependent variable and vice versa. **B** values are filled into Logistic regression Equation which predicts the Social Media Impact on Individual using Age, Authentic\Fake and Confident\Insecure.

The regression equation is:

Log(p/1-p) = 5.514 + 2.063(Authentic\Fake) + 1.790(Confident\Insecure) – 0.491 (Age)

The Left-hand side is called Logit or Log odds. If we increase (Authentic\Fake) by unit it changes the Log odds by 2.063as it's categorical variable it can take only 1and 2 as an input. Age is a Continuous variable and it is inversely proportional due to which 1 unit increase in Age will reduce the Log odds by 0.491. P is the probability of Positive Social Media Impact and 1-P is the probability of not having a Positive Social Media Impact.

Odds Ratio:

Exp(B) values for each variable in table represents the Odds Ratio of that variable. It can be interpreted as Odds of having positive Impact is 7.866 times higher if person reports as Social Media makes them more Authentic when all the other factors are equal. If the odds ratio is less than 1 it means that independent variable is inversely proportional to Dependent Variable. Here if Age increases by 2 unit it will multiply 0.612*0.612 = 0.3745 which is reducing the odds of having Positive impact.

VI. CONCLUSION

Logistic Regression was implemented in order to predict the Impact of Social Media on Individuals with the help of factors such as Age and Four questions related to social media effect on Individual. Out of those Four questions, questions related to Authentic\Fake, Confident\Insecure were significant during model building process. Implemented model exhibited 93.7 % Overall Accuracy and 96.1% Specificity and 90.7% Sensitivity.

I. Principal Component Analysis

Datasets gathered in the real-world setting often contain a large number of attributes that could be in hundreds or thousands. Performing analysis on such extensive datasets becomes a cumbersome task. If there 10 attributes, the correlation matrix to understand interrelationship will be 10 by 10, it becomes difficult to understand relationships with 10 variables, and as the number increases it becomes more difficult. Principal Component Analysis is a technique used to reduce the number of variables when there is a large number of variables in a dataset which helps perform analysis. The variables in such a dataset are often correlated and could be duplicate, PCA transforms these correlated variables into a significantly small number of uncorrelated variables. These transformed variables retain as much possible information from original variables. In PCA the original variables are transformed into a smaller set of linear combinations, which contains all the variance from the original variables.

II. Steps involved in Principal Component Analysis:

1) Select a Suitable and Appropriate Dataset

2) Extract the Factors

3) Rotate the Factors

4) Interpret the Results

5) Compute the Factor Scores

1. Selection of Suitable and Appropriate Dataset

This includes multiple assumptions to check whether the dataset is suitable to conduct PCA. These tests mainly aim to check reliable Sample Size and strength of relationship among variables.

Assumption 1: Sample Size

There is a difference of opinion among authors in selecting sample size for PCA. In general, the larger the sample, is better. When there is a smaller number of samples, correlation coefficients among the variables are less reliable. There is an exception to this when there are strong correlations among variables and a few distinct factors.

The generalized approach ranges from at least 5 observations per variable and at least 100 in total to 10-20 per variable.

Assumption 2: Strength of the Relationship (Correlation)

i) Here we look at correlation coefficients. There should be a significant number of coefficient greater than 0.3

ii) **Bartlett’s Test of Sphericity**

Here $H_{0:}$ Correlation between Variables is zero. And

H_1: Variables are Correlated.

We want to reject the Null Hypothesis i.e.; the significance value should be less than 0.05

iii) **Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy**

This measures the extent to which correlation between pairs of variables can be explained by other variables. The maximum value for this is 1. For good factor analysis, this value should be greater than 0.6.

2. Factor Extraction

This step involves deciding the number of factors that best describe the relationship between the variables.

This can be dependent on two factors:

1) Rationalizing the number of factors, Determining the smallest number of factors that can be used.

2) Selecting factors based on Maximum possible variance explained.

Eigenvalue:

The eigenvalue for a particular component measures the variance in all the variables which are accounted for by that component.

PCA focuses on explaining as much of the total variance possible from variables, using as few components. Every new Principal Component created needs to be uncorrelated with the previously created component.

Several techniques help to decide the factor extraction.

Kaiser’s Criterion:

According to Kaiser’s Criterion, we select the factors\components which are having an eigenvalue greater than 1.

Catell’s Scree Plot:

This is a graphical representation of components and their eigenvalues measured. We keep the factors that are above the break or elbow point.

3. Rotate the Factors:

This step is focused on interpreting the factors, to make this step easier Rotation technique is used. Rotation maximizes the high correlation among factors and variables and reduces the low. There are two ways to do this if factors are kept uncorrelated then orthogonal else oblique if factors are allowed to correlate.

Varimax Orthogonal rotation simplifies the columns of the loading matrix so that each component is comprised of a limited set of variables.

4. Interpret the Result:

The simplified structure is not achieved compulsorily. We can name the factors based on the content of the high loading from each factor to check if they fit conceptually and can be denoted by the name.

5. Compute Factor Scores:

If desired we can compute and save the scores of each component obtained during PCA. This can be done in multiple ways such as regression, bartlett and Anderson-Rubin. Using Factor scores generated from these 5 components, we can perform further statistical analysis such as Regression, Classification.

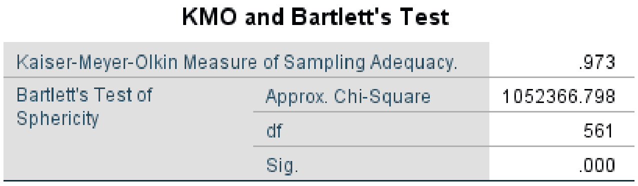
III. Example:

The dataset contains details of FIFA 2020 Players Ratings. There are 34 different attributes on which players are given ratings from0-100.

Assumptions:

The dataset contains 18000 records, the assumption of Sample size is fulfilled.

The majority of the correlation among variables is greater than 0.3. As the Correlation Table was large in size it is not attached here.



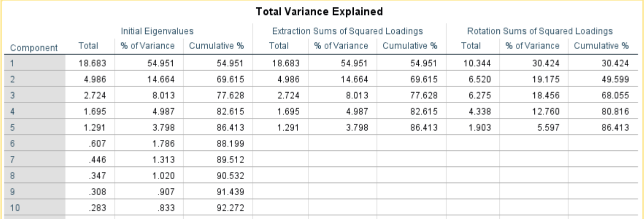
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.973
Bartlett's Test of Sphericity	Approx. Chi-Square	1052366.798
	df	561
	Sig.	.000

Bartlett's Test of Sphericity is significant. The p-value of less than 0.05 due to which we reject the null hypothesis of zero correlation.

KMO measure is having a value of 0.973. This indicates that the correlation between the pair of variables can be explained strongly by other variables.

Total Variance Explained:



Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	18.683	54.951	54.951	18.683	54.951	54.951	10.344	30.424	30.424
2	4.986	14.664	69.615	4.986	14.664	69.615	6.520	19.175	49.599
3	2.724	8.013	77.628	2.724	8.013	77.628	6.275	18.456	68.055
4	1.695	4.987	82.615	1.695	4.987	82.615	4.338	12.760	80.816
5	1.291	3.798	86.413	1.291	3.798	86.413	1.903	5.597	86.413
6	.607	1.786	88.199						
7	.446	1.313	89.512						
8	.347	1.020	90.532						
9	.308	.907	91.439						
10	.283	.833	92.272						

Total Column denotes the eigenvalue for each component. Which is a measure of variance in all the variables accounted for that component.

% of Variance Column denotes the percentage of Variance from original variables is incorporated in a particular component.

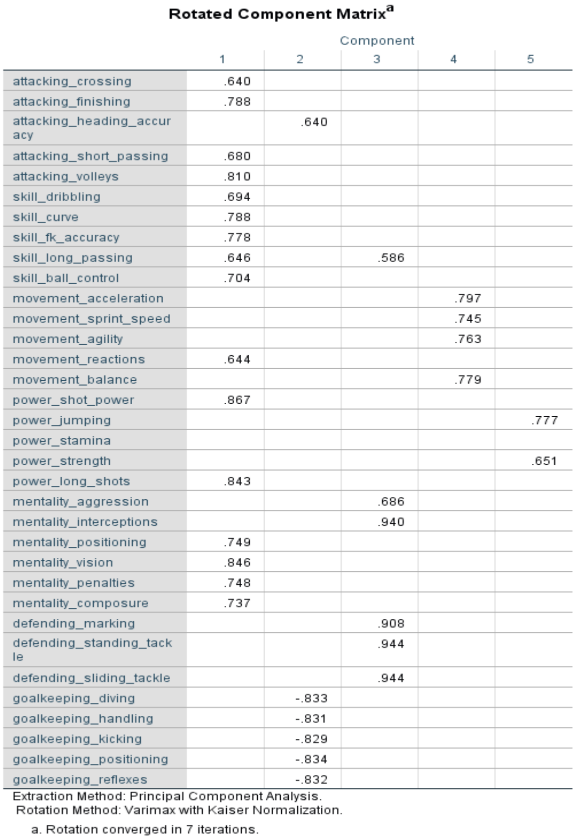
Cumulative % is an addition of subsequent component variance percentage.

Here 5 components with an eigenvalue greater than 1 have explained 86.413% of the total variance from original attributes.

There could be an argument in deciding, whether to keep the 5th component or remove it as it is having an eigenvalue close to 1 and the rest are quite high.

Rotated Component Matrix:

This table shows us how each of the underlying variables correlates with an individual component. As we have asked to suppress the correlations less than 0.5 it shows us only variables with high correlations with the respective component. Now we can try to interpret the relationship between variables of the high correlation for each component and name the component. The initial look at Dataset indicates that variables can be grouped as a type of skill or attribute such as Attacking, Movement, Skills, Power, Mentality, Defending and Goalkeeping. After processing PCA, the first component highly correlates with characteristics of an Attacker\Forward and Midfield Players. The second Component exhibits the characteristics of the Goal Keeper. The third Component is having variables related to Defenders.4th and 5th component exhibits common attributes among football players.



Rotated Component Matrix[a]

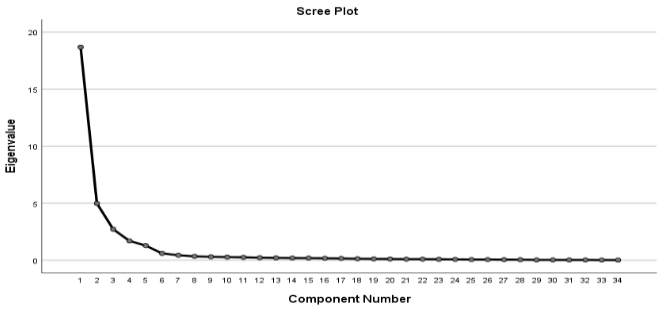
	Component				
	1	2	3	4	5
attacking_crossing	.640				
attacking_finishing	.788				
attacking_heading_accuracy		.640			
attacking_short_passing	.680				
attacking_volleys	.810				
skill_dribbling	.694				
skill_curve	.788				
skill_fk_accuracy	.778				
skill_long_passing	.646		.586		
skill_ball_control	.704				
movement_acceleration				.797	
movement_sprint_speed				.745	
movement_agility				.763	
movement_reactions	.644				
movement_balance				.779	
power_shot_power	.867				
power_jumping					.777
power_stamina					
power_strength					.651
power_long_shots	.843				
mentality_aggression			.686		
mentality_interceptions			.940		
mentality_positioning	.749				
mentality_vision	.846				
mentality_penalties	.748				
mentality_composure	.737				
defending_marking			.908		
defending_standing_tackle			.944		
defending_sliding_tackle			.944		
goalkeeping_diving		-.833			
goalkeeping_handling		-.831			
goalkeeping_kicking		-.829			
goalkeeping_positioning		-.834			
goalkeeping_reflexes		-.832			

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 7 iterations.

Scree Plot:

Here x-axis represents Component Numbers and Y-axis represents respective Eigenvalues. There is a straight line from the 34th to 6th component and then there is a break or elbow in the plot. This helps us selecting components after the breaking point which are 5-1.



IV. Conclusion:

PCA used to reduce the dimensions of 34 Football Player attributes, which reduced the number of attributes to 5 based on Player Paying Role and common characteristics. 34 Correlated variables got transformed into 5 uncorrelated components. Scores of these components can be used for further statistical analysis.

I. Time Series Analysis

Aim & Objective:

Time Series Analysis plays an important role when it comes to predicting Electricity Consumption. This Time Series Analysis aims to predict the Electricity Supply, Transformation and Consumption for the Nation France for at least 4 months in advance. The analysis would help electricity generating and distributing companies to increase their efficiency and revenue by planning according to the requirement.

Data Description:

The data set contains continuous monthly data from January 2016 to October 2020. The record contains the Electricity Supply, Transformation and Consumption value in Gigawatt-hour.

Link:
https://ec.europa.eu/eurostat/databrowser/view/NRG_CB_EM__custom_391273/default/table?lang=en

Data Cleaning:

Before starting any analysis, it is important to have data in a clean format that is without null values, noise, outliers for better analysis accuracy.

- Converted Time Column from Character type to Month and Date Format using **as. Date ()** function of R.
- The dataset which was in CSV format imported in R using the **read.csv()** function.
- Data in csv was aligned row-wise, converted that to column-wise by taking the transpose of the rows using **t()**.

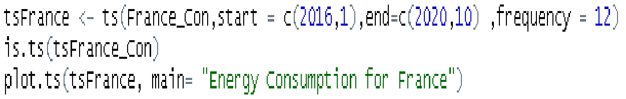
Variables:

Dependent Variable: Supply, Transformation and Consumption of Electricity in Gigawatt-hour. 1 Gigawatt is equal to 1000 Megawatts.

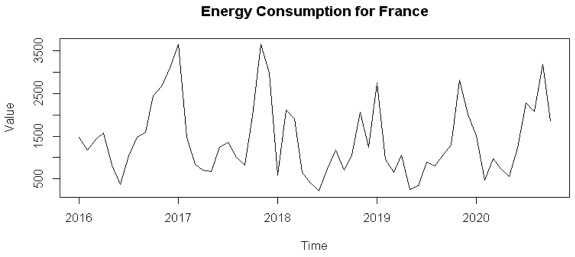
Independent Variable: Months from (January 2016-October 2020).

Time Series Model:

To begin the analysis, the first step which we need to do is to transform the data frame into a time-series format using ts() and to confirm whether we have successfully transformed the data frame we use **is.ts()**.



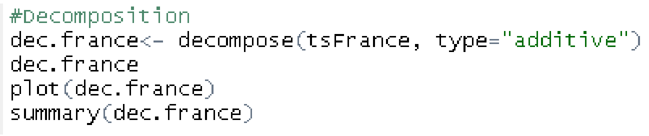
```
tsFrance <- ts(France_Con,start = c(2016,1),end=c(2020,10) ,frequency = 12)
is.ts(tsFrance_Con)
plot.ts(tsFrance, main= "Energy Consumption for France")
```



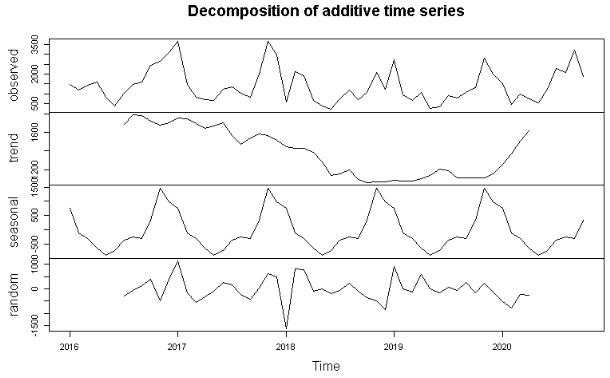
II. Time Series Transformation:

Time Series shows characteristics such as Trend, Seasonality and Random Fluctuations. From the above graph, we can make out that there is seasonality, slight downward trend and random fluctuations.

We have a seasonally decomposed time series. For which we used **decompose().** Here we have used an additive model as sizes of the seasonal fluctuation are about the same size between earlier and later periods**.**

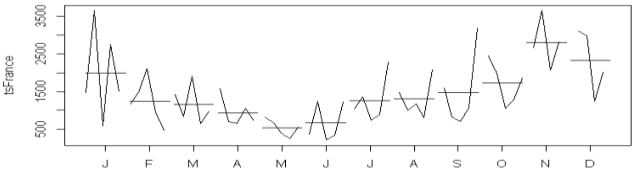


```
#Decomposition
dec.france<- decompose(tsFrance, type="additive")
dec.france
plot(dec.france)
summary(dec.france)
```

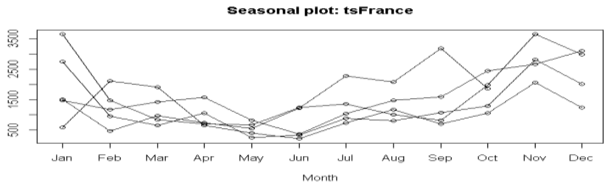


The above graph indicates the Trend, Seasonal and random fluctuations within time-series.

Month Plot: This gives us a graph of each month's electricity consumption from the year (2016-2020) and their averages.



Seasonal Plot:

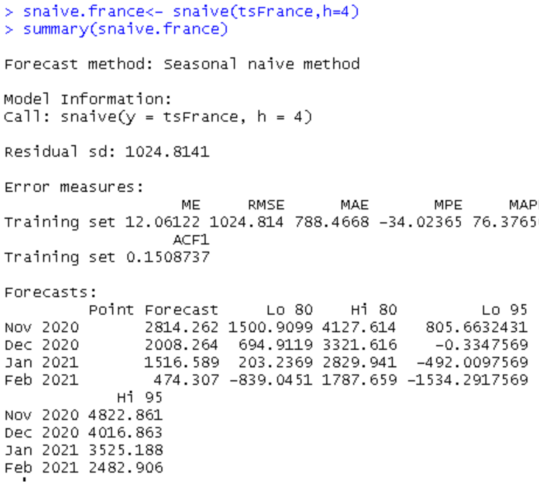


III. MODEL BUILDING:

1) Seasonal Naïve Method:

As the Time Series is exhibiting Seasonal Characteristic, we choose to perform Seasonal Naïve than Naïve method. In this Forecast, values are set to the last observed value from the same season of the year.

The Seasonal Naïve model gives **Residual Sd: 1024.81** and **MAPE of 76.37**.



```
> snaive.france<- snaive(tsFrance,h=4)
> summary(snaive.france)

Forecast method: Seasonal naive method

Model Information:
Call: snaive(y = tsFrance, h = 4)

Residual sd: 1024.8141

Error measures:
                   ME     RMSE      MAE       MPE     MAPE MASE
Training set 12.06122 1024.814 788.4668 -34.02365 76.37656    1
                  ACF1
Training set 0.1508737

Forecasts:
         Point Forecast      Lo 80    Hi 80          Lo 95
Nov 2020       2814.262 1500.9099 4127.614   805.6632431
Dec 2020       2008.264  694.9119 3321.616    -0.3347569
Jan 2021       1516.589  203.2369 2829.941  -492.0097569
Feb 2021        474.307 -839.0451 1787.659 -1534.2917569
            Hi 95
Nov 2020 4822.861
Dec 2020 4016.863
Jan 2021 3525.188
Feb 2021 2482.906
```

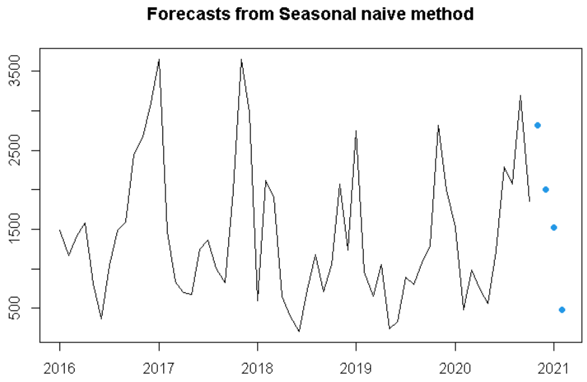
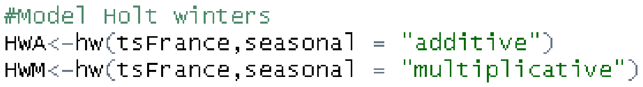


Figure: Forecast from Seasonal Naive

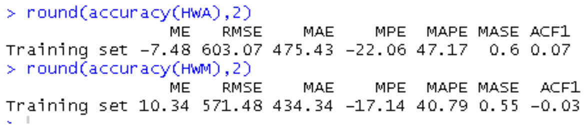
2) Holts Winters Model:

The forecasting method used in the exponential smoothing model takes weighted averages of past observations. Holts Winters model is one of the exponential smoothing models which helps to forecast time series with level, Trend, Seasonality.

Here we have compared both Holts Winters Additive and Multiplicative Model with consideration of all three Level, trend and Seasonality.

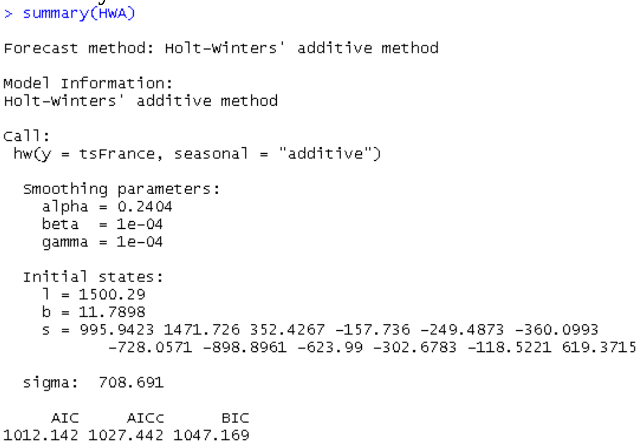


```
#Model Holt winters
HWA<-hw(tsFrance,seasonal = "additive")
HWM<-hw(tsFrance,seasonal = "multiplicative")
```



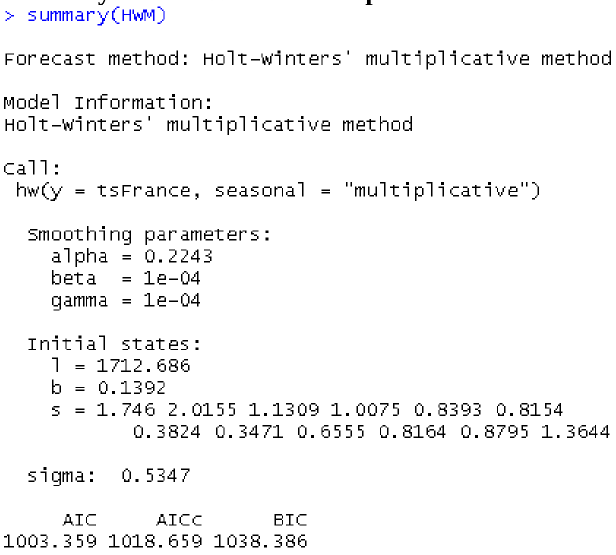
```
> round(accuracy(HWA),2)
                ME   RMSE    MAE    MPE  MAPE MASE ACF1
Training set -7.48 603.07 475.43 -22.06 47.17  0.6 0.07
> round(accuracy(HWM),2)
                ME   RMSE    MAE    MPE  MAPE MASE  ACF1
Training set 10.34 571.48 434.34 -17.14 40.79 0.55 -0.03
```

Summary of Holt Winters Additive Model:



```
> summary(HWA)

Forecast method: Holt-Winters' additive method

Model Information:
Holt-Winters' additive method

Call:
 hw(y = tsFrance, seasonal = "additive")

  Smoothing parameters:
    alpha = 0.2404
    beta  = 1e-04
    gamma = 1e-04

  Initial states:
    l = 1500.29
    b = 11.7898
    s = 995.9423 1471.726 352.4267 -157.736 -249.4873 -360.0993
           -728.0571 -898.8961 -623.99 -302.6783 -118.5221 619.3715

  sigma:  708.691

     AIC     AICc      BIC
1012.142 1027.442 1047.169
```

Summary of Holt Winters Multiplicative Model:



```
> summary(HWM)

Forecast method: Holt-Winters' multiplicative method

Model Information:
Holt-Winters' multiplicative method

Call:
 hw(y = tsFrance, seasonal = "multiplicative")

  Smoothing parameters:
    alpha = 0.2243
    beta  = 1e-04
    gamma = 1e-04

  Initial states:
    l = 1712.686
    b = 0.1392
    s = 1.746 2.0155 1.1309 1.0075 0.8393 0.8154
           0.3824 0.3471 0.6555 0.8164 0.8795 1.3644

  sigma:  0.5347

     AIC     AICc      BIC
1003.359 1018.659 1038.386
```

Forecast Plot for Additive and Multiplicative Model:

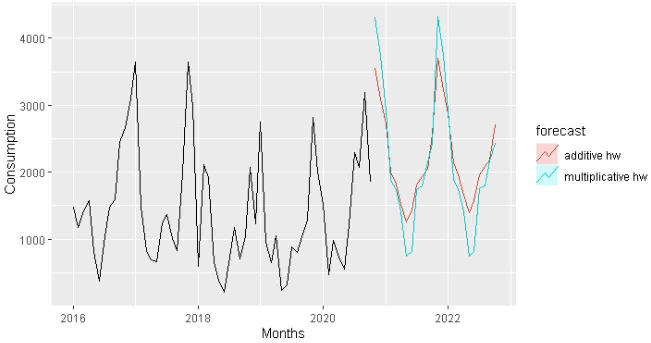


Figure Holts Winter Forecast Plot

Here Additive model should be the initial selection as there is no sign of Multiplicative Seasonal Fluctuations. But after comparing the result from the model, we found that Multiplicative Model is giving slightly more accuracy than the additive model. Additive Model is **having RMSE 603.07, MAPE 47.17 and AIC OF 1012.14** whereas the Multiplicative model is having **RSME: 571.4448 MAPE:40.79 AIC: 1003.35.**

When comparing Seasonal Naïve Model with Holt-Winters Models, Holt-Winters models are significantly better than Seasonal Naïve for forecasting.

3)ARIMA(Auto Regressive Integrated Moving Average) :

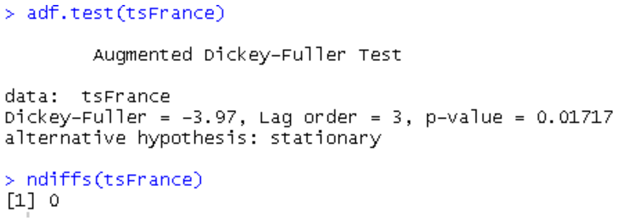
ARIMA model is a slight different approach for prediction, which is using past actual values of the series and recent errors of the predictions. AR stands for Auto Regression which is a lags of the stationarize series, MA stands for Moving Average,which is a Lags of the Forecast Error and if the time series needs to be differenced to make it stationay is called integrated.

Augumented Dickey- Fuller Test:

Time series needs to be stationary in order to implement the ARIMA model. Generally, if the time series exhibits Trend and Seasonality we need to stationaries such series. In order to find out whether Time Series is Stationary or Not we use the ADF test. Here Null Hypothesis is that time-series data is not stationary and the Alternative hypothesis is that it is stationary. A significant p-value indicates that the time series is stationary and doesn't require differencing. From the below figure we can find that the p-value 0.01717 is less than 0.05 which is significant. That's why we reject the Null Hypothesis.

Ndiffs():

In order to find the number of times time-series needs to be differenced to become stationary, we use ndiffs()Function. Here the Value of ndiffs is 0 which also indicates that the **d value** in the ARIMA will be 0 and the time series doesn't need to be integrated.



```
> adf.test(tsFrance)

	Augmented Dickey-Fuller Test

data:  tsFrance
Dickey-Fuller = -3.97, Lag order = 3, p-value = 0.01717
alternative hypothesis: stationary

> ndiffs(tsFrance)
[1] 0
```

Figure: ADF Test & Ndiffs

ACF & PACF:

When we look at the time series, we can find out that, there could be a level of correlation between time-series value and time series value one period back. ACF plot represents the correlation of the series with its own past values. The dotted line represents correlations that are significant, not equal to zero. From the Figure, we can find out that there is a significant correlation at a first and second lag.

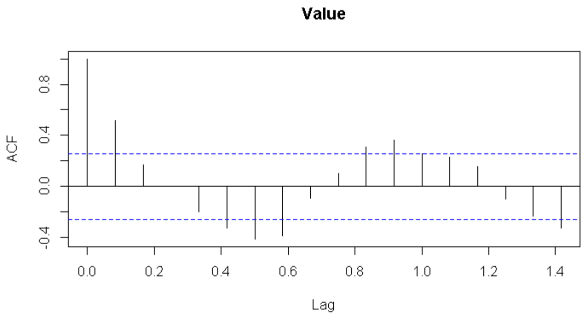


Figure: ACF

PACF is the Partial Autocorrelation Function. ACF & PACF helps us decide the value of p and q in the ARIMA model. There shouldn't be a significant correlation and partial correlations at the beginning of the series.

Looking at the ACF plot we can see that there is a significant correlation greater than zero at a first and second lag. PACF Plot shows correlation at a first lag. We mainly focus on the correlation of the initial lags at the start of the time series than odd significant correlation lags in between. So, we decide the value of **p as 2 and q as 1**.

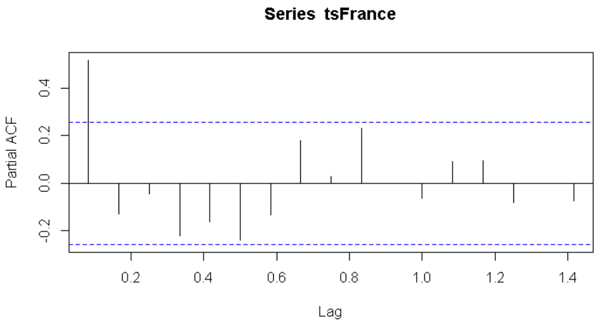
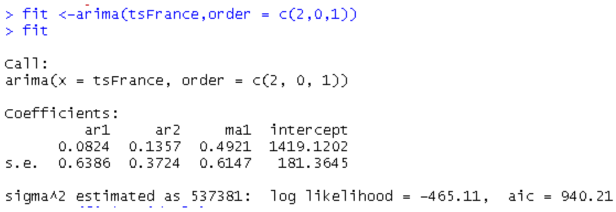


Figure: PACF

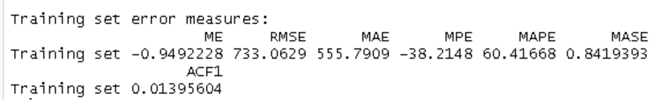
Model:



```
> fit <-arima(tsFrance,order = c(2,0,1))
> fit

Call:
arima(x = tsFrance, order = c(2, 0, 1))

Coefficients:
         ar1     ar2     ma1  intercept
      0.0824  0.1357  0.4921  1419.1202
s.e.  0.6386  0.3724  0.6147   181.3645

sigma^2 estimated as 537381:  log likelihood = -465.11,  aic = 940.21
```

Figure: ARIMA Model with (2,0,1)

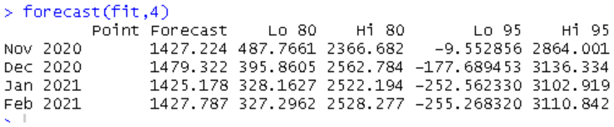
Summary(fit):



```
Training set error measures:
                     ME     RMSE      MAE      MPE     MAPE      MASE
Training set -0.9492228 733.0629 555.7909 -38.2148 60.41668 0.8419393
                   ACF1
Training set 0.01395604
```

Figure: Summary of ARIMA Model

Forecast:



```
> forecast(fit,4)
         Point Forecast    Lo 80    Hi 80       Lo 95    Hi 95
Nov 2020       1427.224 487.7661 2366.682   -9.552856 2864.001
Dec 2020       1479.322 395.8605 2562.784 -177.689453 3136.334
Jan 2021       1425.178 328.1627 2522.194 -252.562330 3102.919
Feb 2021       1427.787 327.2962 2528.277 -255.268320 3110.842
```

Figure Forecast ARIMA Model

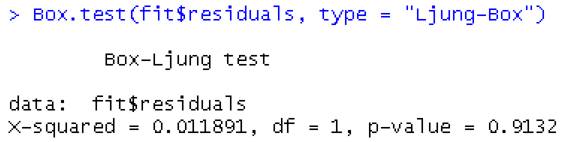
Ljung-Box Test :

This is a test of Residuals, there shouldn’t be any pattern left in residuals. Autocorrelations shouldn't be significant for residuals at any lag. Ljung Box Test is used to find whether autocorrelation of residuals does not differ from zero.

Here Null Hypothesis is that Autocorrelations are not significant and the model fits the data well.

The alternative hypothesis is that there is significant Autocorrelation among residuals and the ARIMA model doesn’t fit the data well.

The non-significant test result suggests that autocorrelations don’t differ from zero and the ARIMA model appears to fit the data well.



```
> Box.test(fit$residuals, type = "Ljung-Box")

	Box-Ljung test

data:  fit$residuals
X-squared = 0.011891, df = 1, p-value = 0.9132
```

Figure Box Ljung Test

From the above figure we can see that the p-value is non-significant that’s why we fail to reject the Null Hypothesis.

ARIMA Forecast Plot:

In the below Figure, Blue Dots represents the predicted value 4 periods ahead along with the interval.

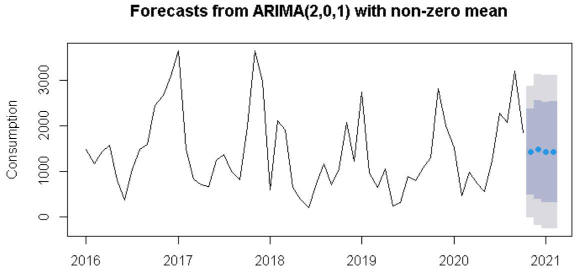


Figure: ARIMA Forecast

Check Residual Function Plot:

checkresiduals function provides the below figures in which ACF plot shows that there aren’t correlations greater than 0.2, Histogram indicates that residuals are normally distributed.

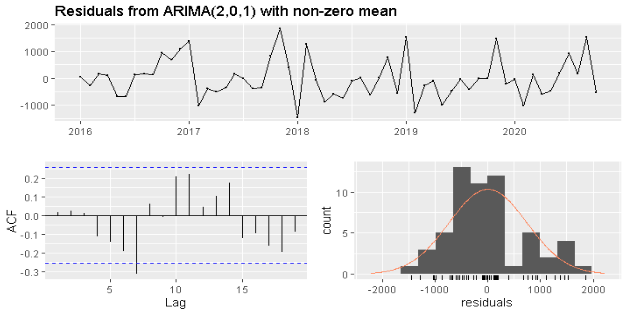


Figure: checkresiduals plot

Q-Q Plot of Residuals:

Q-Q Plot to check whether Residuals are Normally Distributed. From the figure, we can see that at the end residuals are not quite normally distributed.

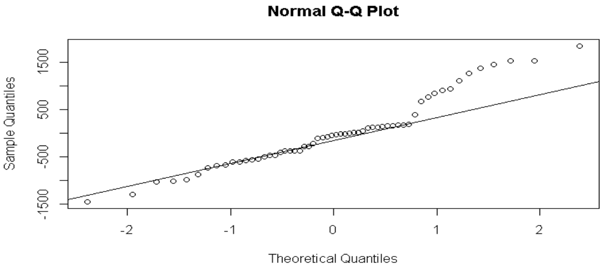


Figure: Q-Q Plot of the residuals

Results:

ARIMA model with (2,0,1) coefficients obtained predictions with an AIC value of 940.21 and RMSE 733.06. Which is better in comparison to the other three models.

Conclusion:

Time Series Analysis was performed to forecast the Electricity Supply, Transformation and Consumption for the Nation France. Seasonal Naïve, Holts Winter Additive & Multiplicative and ARIMA model with (2,0,1) coefficients were implemented on 58 months of historical data. The seasonal Naïve model performed least well in comparison to the other four models. ARIMA model was best with the least AIC value of 940.21

VII. References

[1] Pallant, Julie. Ebook: SPSS Survival Manual: a Step by Step Guide to Data Analysis Using IBM SPSS, McGraw-Hill Education, 2020