Part A -Time Series Analysis

I. Introduction to Time Series Analysis

Time series analysis is studied for visualizing and understanding these dependencies in the data observed from the past, and to use them to predict the data in the future. Huge data gathered from observations that are recorded in a particular sequence with time like interest rates changing in a week, regular fluctuating stock prices, monthly tourism in a country, annual turnover numbers, and so on. We have a never-ending list of topics where time series can be studied. Tourism has a very significant impact on the economy of any country. This substantial impact is enough to promote the investigation on the number of visitors in a country and to make a detailed forecast to prepare for the future.

II. Statement Of Objectives And Description Of Data

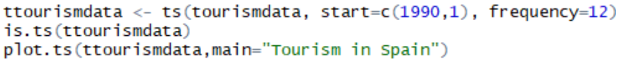
In this project, time series analysis is performed to predict the tourism rate in Spain for the upcoming years according to the net occupancy rate of rooms in hotels. This analysis aims to prepare the country economically for the coming years.

The data is sourced from the Eurostat website (https://ec.europa.eu/eurostat) which consists of the monthly data on the net occupancy rate of rooms in hotels in Spain from January 1990 till October 2020.

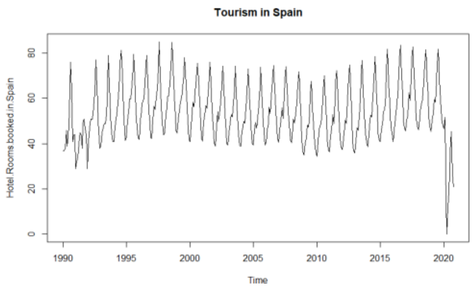
III. Time Series Analysis

A. Plotting Time Series Graph

The first step to plot the time series graph is very important to understand the pattern of the data, which helps to categorize the time series into trend, seasonal, or any other pattern. According to the pattern observed the forecasting model can be chosen. In the below graph, it is observed that there is a regularly repeating pattern of highs and lows in tourism recorded over the years in Spain, which can be determined as the pattern of seasonality. Tourism in Spain has a horizontal pattern until April month of 2020 but falls drastically after that period. This explains that tourism was affected badly due to the pandemic situation over the world.

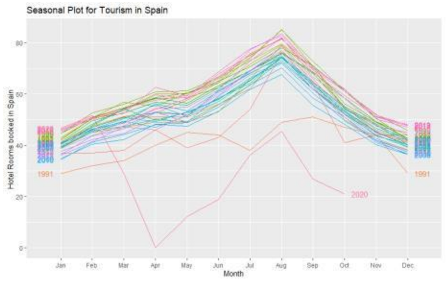


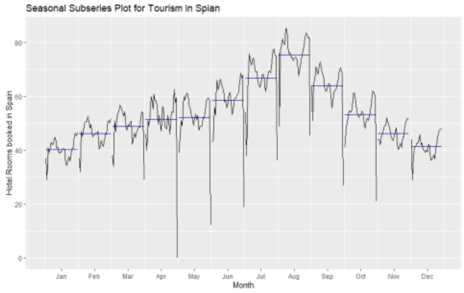
```
ttourismdata <- ts(tourismdata, start=c(1990,1), frequency=12)
is.ts(ttourismdata)
plot.ts(ttourismdata,main="Tourism in Spain")
```



B. Seasonality

Seasonal plot and seasonal subseries plot enables the seasonal pattern to be observed clearly. The seasonal plot shows the data plotted for each season, here it shows the data plotted against each month and explains that the net the occupancy rate of rooms in hotels in Spain goes high in August for almost every year with very few exceptions. Similarly, the alternative seasonal subseries plot shows the mean values for each month denoted by a horizontal line. This again shows that tourism in Spain increases gradually towards August and then again have a gradual decrease.





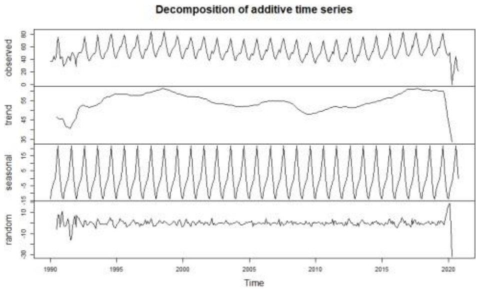
C. Seasonal Decomposition

The method of Seasonal Decomposition decomposes a series into a seasonal component, trend component, observed component, and a random component as shown below. Seasonal decomposition can be additive or multiplicative. The components given are summed up to give the values of the time series in an additive model.

$$Y_t = Trend_t + Seasonal_t + Irregular_t$$

where the observations at time t are the sum of contributions of the trend, seasonal effect, and irregularity in the data.

The additive model for seasonal decomposition is used where the seasonal variations do not depend on the changing time. Since, in the considered time series data, the seasonal variations in the early period are the same as the size of the seasonal variations in the post periods, an additive model is appropriate here.



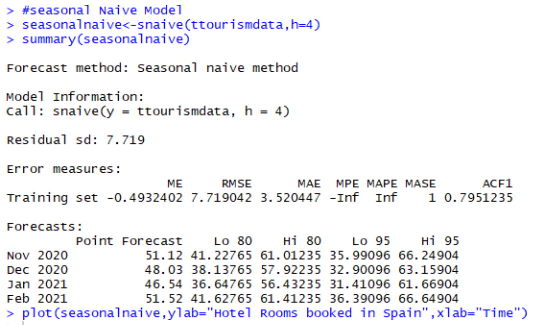
IV. Model Building

A. Model 1: Seasonal Naïve Model

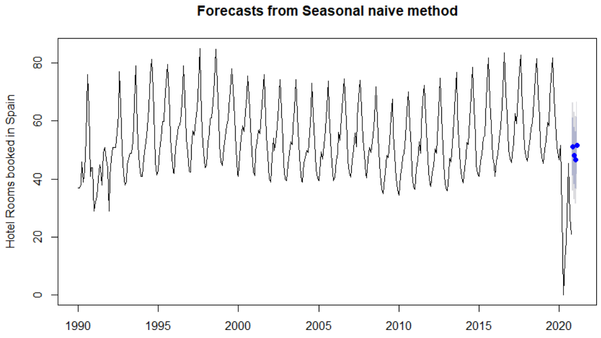
This model is useful for series with high seasonality. Here, forecasted values equal to the last recorded value from the same season of the year, like in the below time series the forecasted values are equal to the last given value for the same month of the previous year.

$$y_{T+h|T}=y_{T+h-m(k+1)},$$

where m is the seasonal period and k is the integer part of (h-1)/m



```
> #seasonal Naive Model
> seasonalnaive<-snaive(ttourismdata,h=4)
> summary(seasonalnaive)

Forecast method: Seasonal naive method

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

Residual sd: 7.719

Error measures:
                     ME     RMSE      MAE  MPE MAPE MASE      ACF1
Training set -0.4932402 7.719042 3.520447 -Inf  Inf    1 0.7951235

Forecasts:
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Nov 2020          51.12 41.22765 61.01235 35.99096 66.24904
Dec 2020          48.03 38.13765 57.92235 32.90096 63.15904
Jan 2021          46.54 36.64765 56.43235 31.41096 61.66904
Feb 2021          51.52 41.62765 61.41235 36.39096 66.64904
> plot(seasonalnaive,ylab="Hotel Rooms booked in Spain",xlab="Time")
```



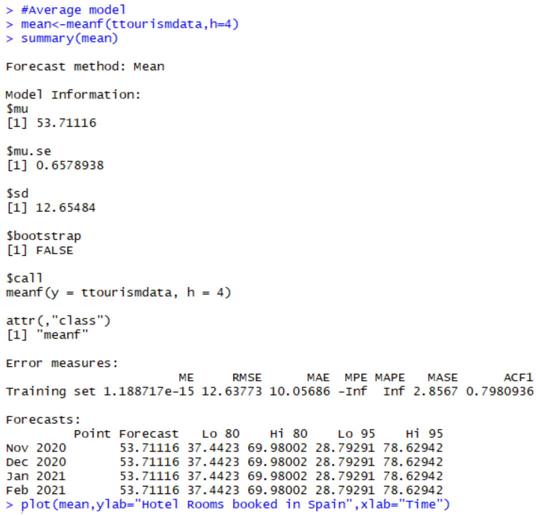
From the above result, the seasonal naïve is forecasting up to 4 periods with RMSE value 7.72.

B. *Model 2: Average Model*

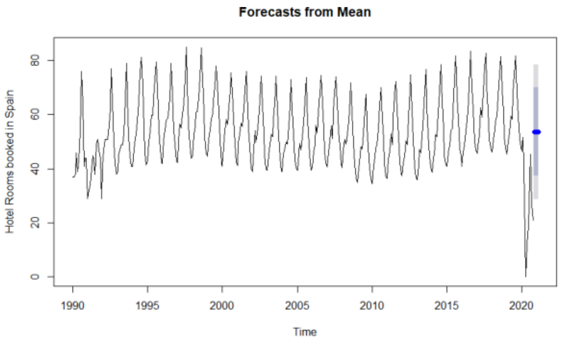
In this model, the forecasted values equal to the average or mean of the past data. Below is the equation for the average model,

$$y_{T+h|T} =(y_1+\cdots+y_T)/T$$

where $y_1 \ldots y_T$ is the past data.



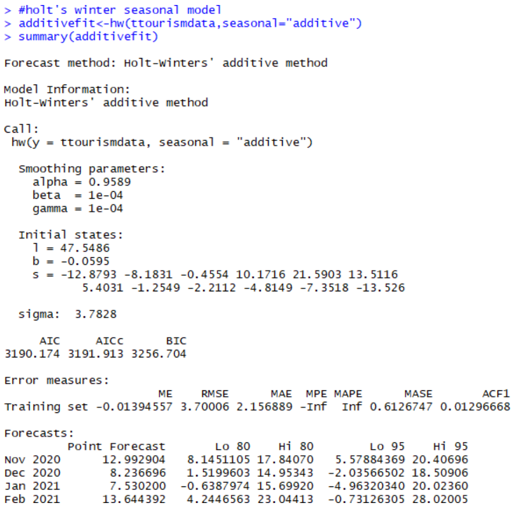
```
> #Average model
> mean<-meanf(ttourismdata,h=4)
> summary(mean)

Forecast method: Mean

Model Information:
$mu
[1] 53.71116

$mu.se
[1] 0.6578938

$sd
[1] 12.65484

$bootstrap
[1] FALSE

$call
meanf(y = ttourismdata, h = 4)

attr(,"class")
[1] "meanf"

Error measures:
                       ME     RMSE      MAE  MPE MAPE   MASE      ACF1
Training set 1.188717e-15 12.63773 10.05686 -Inf  Inf 2.8567 0.7980936

Forecasts:
         Point Forecast   Lo 80    Hi 80    Lo 95    Hi 95
Nov 2020       53.71116 37.4423 69.98002 28.79291 78.62942
Dec 2020       53.71116 37.4423 69.98002 28.79291 78.62942
Jan 2021       53.71116 37.4423 69.98002 28.79291 78.62942
Feb 2021       53.71116 37.4423 69.98002 28.79291 78.62942
> plot(mean,ylab="Hotel Rooms booked in Spain",xlab="Time")
```



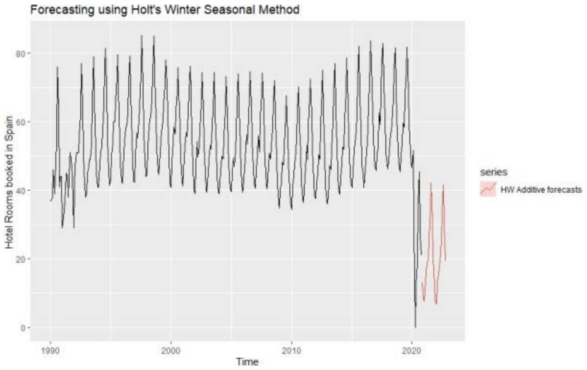
The given average model obtains the four predicted values with an RMSE of 12.64.

C. *Model 3:Holt's Winter Seasonal Model*

If a time series has an additive property of increasing and decreasing seasonality and trend, in such cases this model can be used to make short-term forecasts. Here, the seasonal model provides the forecasts using the equation given below and the three smoothing parameters α, β, and γ, having values between 0 and 1.



```
> #holt's winter seasonal model
> additivefit<-hw(ttourismdata,seasonal="additive")
> summary(additivefit)

Forecast method: Holt-Winters' additive method

Model Information:
Holt-Winters' additive method

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

  Smoothing parameters:
    alpha = 0.9589
    beta  = 1e-04
    gamma = 1e-04

  Initial states:
    l = 47.5486
    b = -0.0595
    s = -12.8793 -8.1831 -0.4554 10.1716 21.5903 13.5116
           5.4031 -1.2549 -2.2112 -4.8149 -7.3518 -13.526

  sigma:  3.7828

     AIC     AICc      BIC
3190.174 3191.913 3256.704

Error measures:
                      ME    RMSE      MAE  MPE MAPE      MASE       ACF1
Training set -0.01394557 3.70006 2.156889 -Inf  Inf 0.6126747 0.01296668

Forecasts:
         Point Forecast      Lo 80    Hi 80       Lo 95    Hi 95
Nov 2020      12.992904  8.1451105 17.84070  5.57884369 20.40696
Dec 2020       8.236696  1.5199603 14.95343 -2.03566502 18.50906
Jan 2021       7.530200 -0.6387974 15.69920 -4.96320340 20.02360
Feb 2021      13.644392  4.2446563 23.04413 -0.73126305 28.02005
```



This method has two types of variations, that have seasonal components with different natures, additive and multiplicative. The additive is preferred when we have the seasonal fluctuations mostly constant over the time series, whereas multiplicative is preferred in the case of time series having the seasonal fluctuations proportional to the time level. Hence, in the above plot, Holt's winter seasonal method is forecasted by additive type. The summary function returns the value for α, β, and γ as 0.96,1e-04 and 1e-04 respectively.It also gives the RMSE value of 3.7 and AICc of 3191.9.

D. Model 4: Seasonal ARIMA Model

Autoregressive Integrated Moving Average(ARIMA) and SARIMA are the most widely used approaches in forecasting time series data. These models provide a statistical approach applied to a time series having an irregular component, which allows for the component with non-zero autocorrelations. These models are built to fit stationary data. A seasonal ARIMA(SARIMA) is formed by including an additional seasonal component in the ARIMA model. As the time series for tourism in Spain include seasonal as well as non-seasonal components we use Seasonal ARIMA for forecasting.

$$\text{ARIMA}\quad \underbrace{(p,d,q)}_{\substack{\uparrow \\ \text{Non-seasonal part} \\ \text{of the model}}}\quad \underbrace{(P,D,Q)_m}_{\substack{\uparrow \\ \text{Seasonal part of} \\ \text{of the model}}}$$

where m is the number of observations per year,
p represents the trend auto-regressive element,
d represents the trend differencing element,
q represents the trend moving average element
and P, D, Q are represented similarly for the
seasonal component.

There are certain steps to perform the ARIMA model:

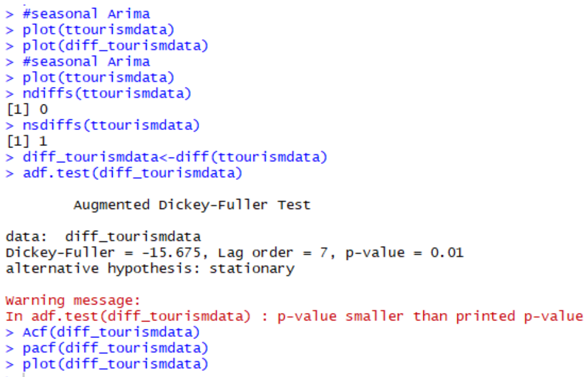
1. Visualize the Time Series:

It is important to analyze the time series pattern before building the model which is mentioned in above Part III-A.

2. Stationarizing the series

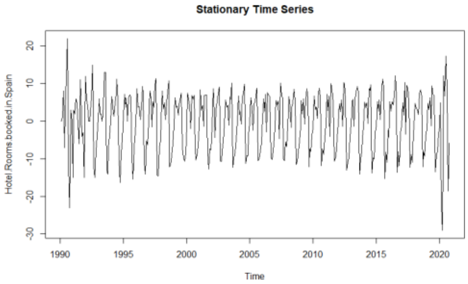
This step is to check if the series is stationary or not, this can be checked by using a Dickey-Fuller test where a significant result suggests stationarity.

Differencing: If the time series is not stationary that is the p-value for the Dickey-Fuller test is not significant then we have certain approaches for making non-stationary data stationary like differencing. Differencing is used to model the differences of the terms rather than the actual term, which is given by the diff() function. The function ndiffs() and nsdiffs() can be used to determine the number of seasonal and ordinal differences(d/D) as given below.



```
> #seasonal Arima
> plot(ttourismdata)
> plot(diff_tourismdata)
> #seasonal Arima
> plot(ttourismdata)
> ndiffs(ttourismdata)
[1] 0
> nsdiffs(ttourismdata)
[1] 1
> diff_tourismdata<-diff(ttourismdata)
> adf.test(diff_tourismdata)

        Augmented Dickey-Fuller Test

data:  diff_tourismdata
Dickey-Fuller = -15.675, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(diff_tourismdata) : p-value smaller than printed p-value
> Acf(diff_tourismdata)
> pacf(diff_tourismdata)
> plot(diff_tourismdata)
```

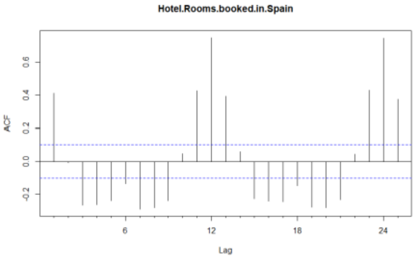
The **Dickey-Fuller** test shows a significant p-value almost equal to 0.00 for the first order of differencing as seen in the above screen capture. After plotting the differenced data we get the stationary plot shown below.

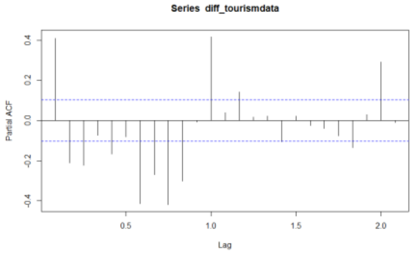


3. Finding Optimal Parameters

The parameters p, q, P, Q can be obtained using ACF and PACF plots where ACF is a plot of total auto-correlation function and PACF is of the partial autocorrelation function.

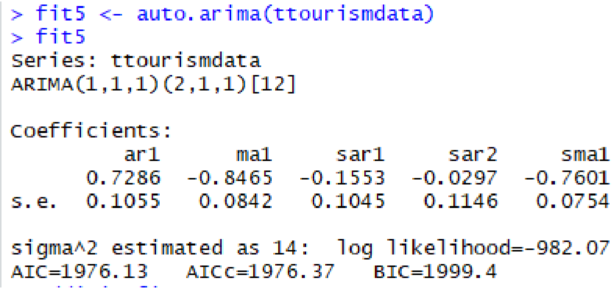
The ACF graph has a cut-off on the curve after 1st lag and so it might be the MA(1) process. The blue line indicates significantly different values than 0.





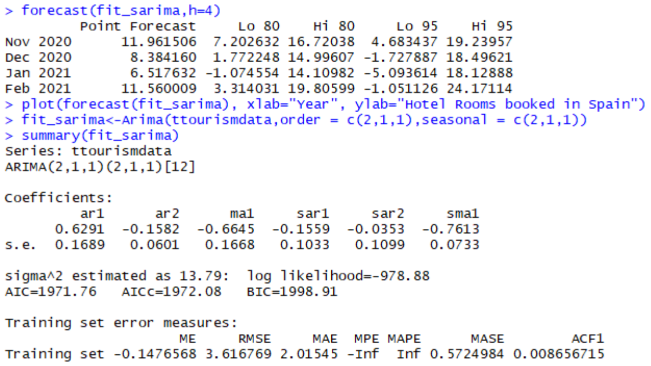
4. Fitting ARIMA Model

With the parameters found in previous sections, the ARIMA model is built by checking different combinations of p,d,q. The one with the lowest AICc should be chosen as the best model. The auto.arima() function can also be used to determine the model which possibly has the best fit. This might be incorrect and hence by checking various combinations the best-fit model can be chosen.



```
> fit5 <- auto.arima(ttourismdata)
> fit5
Series: ttourismdata
ARIMA(1,1,1)(2,1,1)[12]

Coefficients:
         ar1      ma1     sar1     sar2     sma1
      0.7286  -0.8465  -0.1553  -0.0297  -0.7601
s.e.  0.1055   0.0842   0.1045   0.1146   0.0754

sigma^2 estimated as 14:  log likelihood=-982.07
AIC=1976.13   AICc=1976.37   BIC=1999.4
```

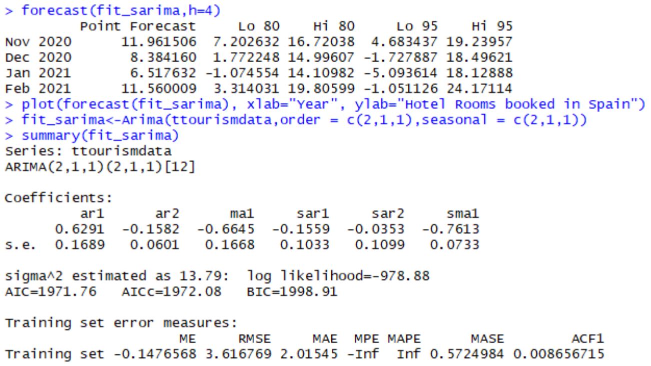
After checking various models with different parameters, the final model ARIMA(2,1,1)(2,1,1)[12] is selected with AICc 1972.08 and RMSE having value 3.62 as given below.



```
> forecast(fit_sarima,h=4)
         Point Forecast       Lo 80     Hi 80      Lo 95    Hi 95
Nov 2020      11.961506  7.202632 16.72038  4.683437 19.23957
Dec 2020       8.384160  1.772248 14.99607 -1.727887 18.49621
Jan 2021       6.517632 -1.074554 14.10982 -5.093614 18.12888
Feb 2021      11.560009  3.314031 19.80599 -1.051126 24.17114
> plot(forecast(fit_sarima), xlab="Year", ylab="Hotel Rooms booked in Spain")
> fit_sarima<-Arima(ttourismdata,order = c(2,1,1),seasonal = c(2,1,1))
> summary(fit_sarima)
Series: ttourismdata
ARIMA(2,1,1)(2,1,1)[12]

Coefficients:
         ar1      ar2      ma1     sar1     sar2     sma1
      0.6291  -0.1582  -0.6645  -0.1559  -0.0353  -0.7613
s.e.  0.1689   0.0601   0.1668   0.1033   0.1099   0.0733

sigma^2 estimated as 13.79:  log likelihood=-978.88
AIC=1971.76   AICc=1972.08   BIC=1998.91

Training set error measures:
                     ME     RMSE     MAE  MPE MAPE      MASE        ACF1
Training set -0.1476568 3.616769 2.01545 -Inf  Inf 0.5724984 0.008656715
```

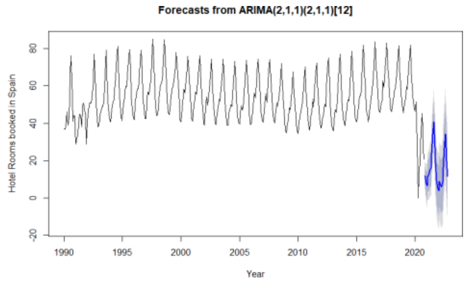
5. Forecasting and plotting

As the final ARIMA model is ready, the future values can be predicted and plotted using the model as shown below.



```
> forecast(fit_sarima,h=4)
         Point Forecast       Lo 80     Hi 80      Lo 95    Hi 95
Nov 2020      11.961506  7.202632 16.72038  4.683437 19.23957
Dec 2020       8.384160  1.772248 14.99607 -1.727887 18.49621
Jan 2021       6.517632 -1.074554 14.10982 -5.093614 18.12888
Feb 2021      11.560009  3.314031 19.80599 -1.051126 24.17114
> plot(forecast(fit_sarima), xlab="Year", ylab="Hotel Rooms booked in Spain")
> fit_sarima<-Arima(ttourismdata,order = c(2,1,1),seasonal = c(2,1,1))
> summary(fit_sarima)
Series: ttourismdata
ARIMA(2,1,1)(2,1,1)[12]

Coefficients:
         ar1      ar2      ma1     sar1     sar2     sma1
      0.6291  -0.1582  -0.6645  -0.1559  -0.0353  -0.7613
s.e.  0.1689   0.0601   0.1668   0.1033   0.1099   0.0733

sigma^2 estimated as 13.79:  log likelihood=-978.88
AIC=1971.76   AICc=1972.08   BIC=1998.91

Training set error measures:
                     ME     RMSE     MAE  MPE MAPE      MASE        ACF1
Training set -0.1476568 3.616769 2.01545 -Inf  Inf 0.5724984 0.008656715
```

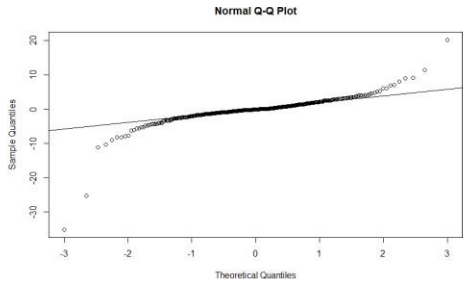
Comparing the RMSE values of various models performed above the SARIMA Model has the lowest RMSE value of 3.62 and hence we can choose this as the optimum model for forecasting time series on the tourism data in Spain.



V. Evaluating The Model Fit

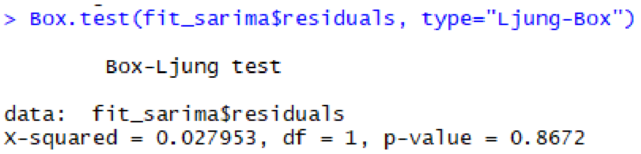
A. *Normal Q-Q Plot*

The residuals are normally distributed with a few outliers in the below Normal Q-Q plot.



B. *Ljung-Box test*

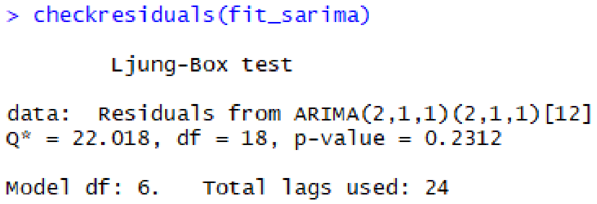
The Ljung-Box test is used to test such that the autocorrelations are all zero. The p-value is non-significant which suggests that the autocorrelations don't differ from zero.



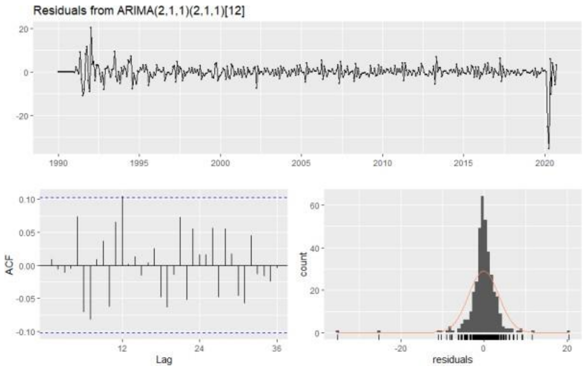
```
> Box.test(fit_sarima$residuals, type="Ljung-Box")

        Box-Ljung test

data:  fit_sarima$residuals
X-squared = 0.027953, df = 1, p-value = 0.8672
```

C. *Check residuals*

According to the ACF plot below, the residual autocorrelations do not differ significantly from 0 so we can say that the model is correctly specified. Also, the residuals are normally distributed and have constant variance and so this model can be considered as one of the best-fit models.



```
> checkresiduals(fit_sarima)

        Ljung-Box test

data:  Residuals from ARIMA(2,1,1)(2,1,1)[12]
Q* = 22.018, df = 18, p-value = 0.2312

Model df: 6.   Total lags used: 24
```



VI. Conclusion

Time Series analysis on four different models resulted in different forecasts for the next four periods with different accuracies. Comparing the models, we can conclude that the Seasonal ARIMA model is the best model for forecasting the data on Tourism in Spain for the upcoming years. The Tourism in Spain is forecasted to have comparatively less tourism in the upcoming years than the previous years due pandemic situation showing adverse affect.

VII. References

[1] Papatheodorou, Andreas & Song, Haiyan. (2005). International Tourism Forecasts: Time-Series Analysis of World and Regional Data. Tourism Economics. 11. 11-23. 10.5367/0000000053297167.

Part B - Logistic Regression Analysis

I. Introduction to Logistic regression analysis

Logistic Regression is an extended version of linear regression. This regression model is used for the binary classification of variables. Here, the dependent variable must be dichotomous, that is, variables with two classes only. The independent variables can be either be continuous or categorical. It uses probability as the base for performing predictive analysis.

II. STATEMENT OF OBJECTIVES AND DESCRIPTION OF VARIABLES

This project aims to analyze the factors that are affecting owning a mobile phone using the logistic regression model. The model uses the research data collected using the public opinion polling conducted by the Pew Research Center. The variables are in the form of questions that are answered in categorical forms. The answer to the question asked whether a person owns a mobile phone or not is predicted with this model using various factors like one's present age, the level of education, the standard of living, ability to read English, gender, and that mobile phones are good or bad for the society.

The data is sourced from the Pew Research Center site(https://www.pewresearch.org/download-datasets/). Before starting the analysis all the necessary factors are merged into one dataset and the missing values are removed from all the columns for better predictions.

Below is the description of all the variables that are used for analyzing the logistic regression:

A. *Dependent variable:*

1. Name: Do you own a mobile phone yes or no?
   Type: Dichotomous categorical
   Categories: Yes(1), No(2)

B. *Independent variables:*

1. Name: How old were you at your last birthday?
   Type: Continuous
2. Name: Are you a high school graduate?
   Type: Dichotomous categorical
   Categories: Not a High-School Graduate(1), High-School Graduate(2)
3. Name: Adding all the advantages and disadvantages of mobile phones would you say mobile phones have mostly been a good thing or a bad thing for society?
   Type: Dichotomous categorical
   Categories: Good thing(1), Bad thing(2)
4. Name: Compared to your parents when they were the age you are now do you think your own standard of living now is better or worse as theirs was?
   Type: Dichotomous categorical
   Categories: Better(1),Worse(2)
5. Name: Can you read at least some English yes or no?
   Type: Dichotomous categorical
   Categories: Yes(1),No(2)
6. Name: Gender
   Type: Dichotomous categorical
   Categories: Yes(1),No(2)

III. ASSUMPTIONS

One part of the modeling involves checking to ensure that the information you want to analyze can be evaluated using logistic regression when you want to use logistic regression to analyze the results. You need to do this because the logistic regression is fair to be applied only if the data meets all the assumptions needed for logistic regression.

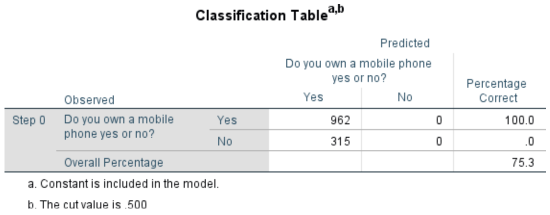
A. *Assumption 1:Dependent variables to be Mutually Exclusive*

The first assumption for logistic regression analysis is that the dependent variable should be mutually exclusive which explains that if the dependent variable for observation has the value of Yes then it cannot contain the value of No. That means no observation can consist of both the values of Yes and No.

Dependent Variable Encoding

Original Value	Internal Value
Yes	0
No	1

The below classification table shows that there are 962 observations are categorized into the Yes category with an internal value of 0 and 315 observations are categorized into the No category having an internal value of 1. These values sum up t the total sample observations taken in this model for analysis which explains that there are no observations that are categorized in both the categories resulting in a mutually exclusive dependent variable.



Classification Table[a,b]

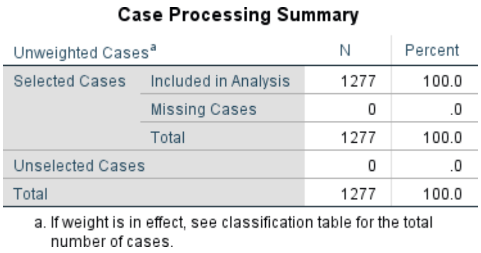
	Observed		Predicted: Do you own a mobile phone yes or no? Yes	Predicted: Do you own a mobile phone yes or no? No	Percentage Correct
Step 0	Do you own a mobile phone yes or no?	Yes	962	0	100.0
		No	315	0	.0
	Overall Percentage				75.3

a. Constant is included in the model.

b. The cut value is .500

B. *Assumption 2: Sample Size*

This assumption is to determine the sample size. The sample size for logistic regression is possibly large to give the best results. The small sample size with large independent variables might cause problems for analyzing the model. The sample size of 20 cases per independent variable must be considered for getting appropriate results from the logistic regression model. This model has 1277 selected cases for modeling which are well sufficient to satisfy the logistic regression model.



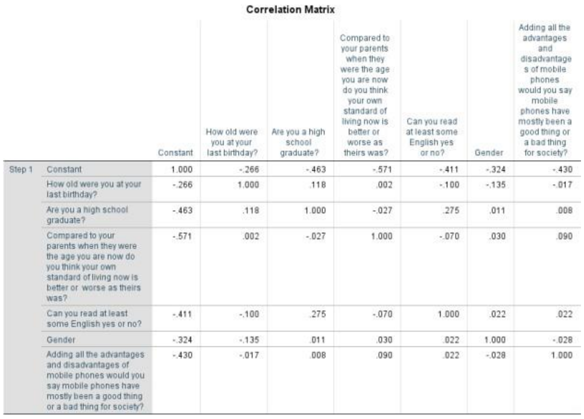
Case Processing Summary

Unweighted Cases^{a}		N	Percent
Selected Cases	Included in Analysis	1277	100.0
	Missing Cases	0	.0
	Total	1277	100.0
Unselected Cases		0	.0
Total		1277	100.0

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

C. Assumption 3: Absence of Multicollinearity

Multicollinearity in variables occurs when two or more independent variables are strongly related to each other. This becomes problematic in estimating the logistic regression model and cannot easily determine which independent variable is contributing to the variance seen in the dependent variable.



Correlation Matrix

		Constant	How old were you at your last birthday?	Are you a high school graduate?	Compared to your parents when they were the age you are now do you think your own standard of living now is better or worse as theirs was?	Can you read at least some English yes or no?	Gender	Adding all the advantages and disadvantages of mobile phones would you say mobile phones have mostly been a good thing or a bad thing for society?
Step 1	Constant	1.000	-.266	-.463	-.571	-.411	-.324	-.430
	How old were you at your last birthday?	-.266	1.000	.118	.002	-.100	-.135	-.017
	Are you a high school graduate?	-.463	.118	1.000	-.027	.275	.011	.008
	Compared to your parents when they were the age you are now do you think your own standard of living now is better or worse as theirs was?	-.571	.002	-.027	1.000	-.070	.030	.090
	Can you read at least some English yes or no?	-.411	-.100	.275	-.070	1.000	.022	.022
	Gender	-.324	-.135	.011	.030	.022	1.000	-.028
	Adding all the advantages and disadvantages of mobile phones would you say mobile phones have mostly been a good thing or a bad thing for society?	-.430	-.017	.008	.090	.022	-.028	1.000

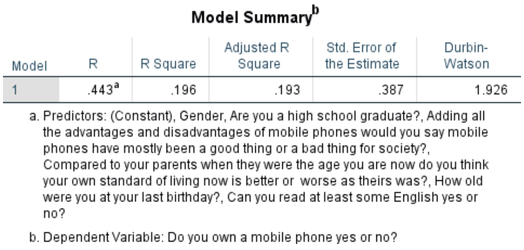
Multicollinearity can be checked in either 1) using correlation matrix for all variables and if the correlation coefficient value is not in between -0.70 to 0.70 then the predictors are said to be multicollinear. 2) An easier way is to check the VIF values to be less than 10.

The above table shows that the correlation between variables is between the range -0.70 to 0.70 and hence the variables are not considered to be multicollinear.

D. Assumption 4:Independece of errors

This assumption states that there must be independent of the residual values. This suggests that there should be no connection between the error terms. Durbin-Watson statistics can be used to verify this assumption.

Durbin-Watson statistics are used to verify the auto-correlation between the residual values ranging from 1 to 3 and determines good statistics when close to 2. For our model, the Durbin-Watson numbers have a value of 1.926, explaining that there is far less auto-correlation between the residuals.



Model Summary^{b}

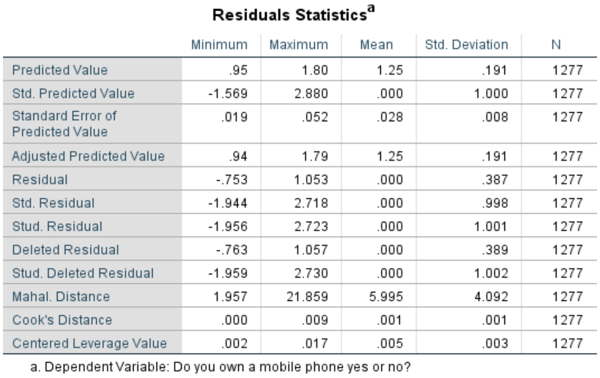
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.443^{a}	.196	.193	.387	1.926

a. Predictors: (Constant), Gender, Are you a high school graduate?, Adding all the advantages and disadvantages of mobile phones would you say mobile phones have mostly been a good thing or a bad thing for society?, Compared to your parents when they were the age you are now do you think your own standard of living now is better or worse as theirs was?, How old were you at your last birthday?, Can you read at least some English yes or no?

b. Dependent Variable: Do you own a mobile phone yes or no?

E. Assumption 5: No significant outliers

In the model, any significant outliers, highly influential points, or high leverage points must not be present. Our model can be influenced by these data points which can bias the expected performance. It will make the data less representative overall and might cause differences in the overall predictions. The outliers can be tested in SPSS using Cook's Distance.



Residuals Statistics^{a}

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	.95	1.80	1.25	.191	1277
Std. Predicted Value	-1.569	2.880	.000	1.000	1277
Standard Error of Predicted Value	.019	.052	.028	.008	1277
Adjusted Predicted Value	.94	1.79	1.25	.191	1277
Residual	-.753	1.053	.000	.387	1277
Std. Residual	-1.944	2.718	.000	.998	1277
Stud. Residual	-1.956	2.723	.000	1.001	1277
Deleted Residual	-.763	1.057	.000	.389	1277
Stud. Deleted Residual	-1.959	2.730	.000	1.002	1277
Mahal. Distance	1.957	21.859	5.995	4.092	1277
Cook's Distance	.000	.009	.001	.001	1277
Centered Leverage Value	.002	.017	.005	.003	1277

a. Dependent Variable: Do you own a mobile phone yes or no?

Cook’s Distance value must not exceed 1. The points for which the value is exceeding 1 are considered to be outliers. In this data model, the minimum value of Cook’s Distance is .000 and the maximum is .009. Hence, our data model does not consist of any outliers or influential data points that need to be removed from the dataset.

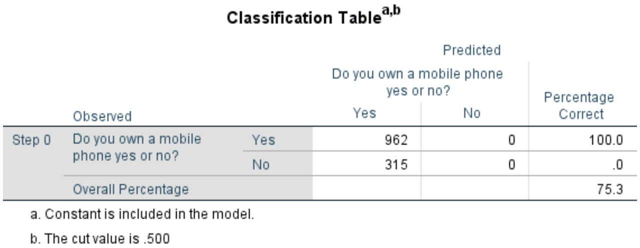
IV. UNDERSTANDING AND BUILDING MODEL

A. Model 1

After all the assumptions are met, the model is built using all the independent transformed variables.

Block 0:

Block 0 is the null model which is built without the independent variables. That is owning a mobile phone is predicted without any predictors as shown below. Further blocks are built adding independent variables to improve the classification accuracy.

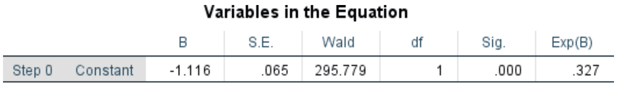


Classification Table^{a,b}

			Predicted		
			Do you own a mobile phone yes or no?		Percentage Correct
	Observed		Yes	No	
Step 0	Do you own a mobile phone yes or no?	Yes	962	0	100.0
		No	315	0	.0
	Overall Percentage				75.3

a. Constant is included in the model.

b. The cut value is .500



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-1.116	.065	295.779	1	.000	.327

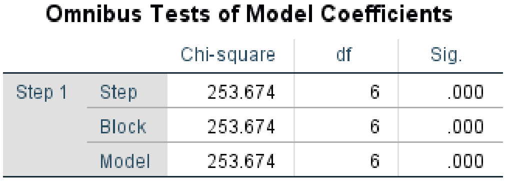
Block 1:

The global hypothesis test is conducted, to check if any of the regression coefficients are other than 0. Using the 0.05 significance level. The null hypothesis is such that all the coefficients of independent variables are equal to 0. The test is such that the statistics are distributed around X^2 with df(degree of freedom) equals to the number of the independent variables.

H_0: $b_1=b_2=b_3=b_4=b_5=b_6=0$

H_1: Not all b’s are 0.

The below Omnibus Tests of model coefficients tell that the p-value is significant and hence we reject the null hypothesis stating that at least one of the coefficients of independent variables is not equal to zero.



Omnibus Tests of Model Coefficients

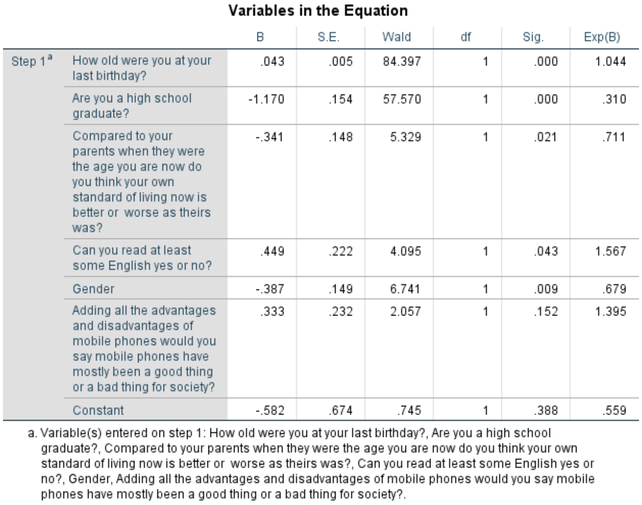
		Chi-square	df	Sig.
Step 1	Step	253.674	6	.000
	Block	253.674	6	.000
	Model	253.674	6	.000

Next, performing the individual hypothesis where individual regression coefficients are being examined. The coefficients of each independent variable are checked.

$H_{0:}b_k=0$

$H_{1:}b_k\neq0$

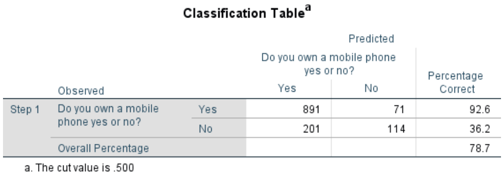
In the below table, the p-value for all the variables is examined and the null hypothesis states that the variables have the regression coefficient equal to zero. The p-value for all the variables except for the Good_Or_Bad_For_Society is significant, hence we reject the null hypothesis for these variables, and accept the null hypothesis only for the mentioned one variable with a significance value of .152.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	How old were you at your last birthday?	.043	.005	84.397	1	.000	1.044
	Are you a high school graduate?	-1.170	.154	57.570	1	.000	.310
	Compared to your parents when they were the age you are now do you think your own standard of living now is better or worse as theirs was?	-.341	.148	5.329	1	.021	.711
	Can you read at least some English yes or no?	.449	.222	4.095	1	.043	1.567
	Gender	-.387	.149	6.741	1	.009	.679
	Adding all the advantages and disadvantages of mobile phones would you say mobile phones have mostly been a good thing or a bad thing for society?	.333	.232	2.057	1	.152	1.395
	Constant	-.582	.674	.745	1	.388	.559

a. Variable(s) entered on step 1: How old were you at your last birthday?, Are you a high school graduate?, Compared to your parents when they were the age you are now do you think your own standard of living now is better or worse as theirs was?, Can you read at least some English yes or no?, Gender, Adding all the advantages and disadvantages of mobile phones would you say mobile phones have mostly been a good thing or a bad thing for society?.



Classification Table[a]

			Predicted		
			Do you own a mobile phone yes or no?		Percentage Correct
	Observed		Yes	No	
Step 1	Do you own a mobile phone yes or no?	Yes	891	71	92.6
		No	201	114	36.2
	Overall Percentage				78.7

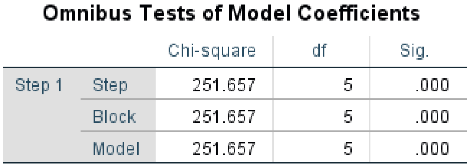
a. The cut value is .500

To obtain better results from the model, the non-significant predictors are to be removed from the model (value with the greatest p-value), and hence we will run Model 2 with the updated variables.

B. Model 2

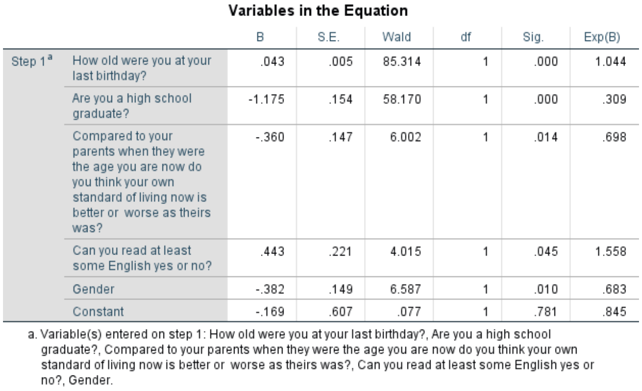
In this model, the independent variable Good_Or_Bad_For_Society is removed as it is the one with the non-significant p-value in the previous model built and run the analysis again.

Below are the findings for Model 2:



Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	251.657	5	.000
	Block	251.657	5	.000
	Model	251.657	5	.000



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	How old were you at your last birthday?	.043	.005	85.314	1	.000	1.044
	Are you a high school graduate?	-1.175	.154	58.170	1	.000	.309
	Compared to your parents when they were the age you are now do you think your own standard of living now is better or worse as theirs was?	-.360	.147	6.002	1	.014	.698
	Can you read at least some English yes or no?	.443	.221	4.015	1	.045	1.558
	Gender	-.382	.149	6.587	1	.010	.683
	Constant	-.169	.607	.077	1	.781	.845

a. Variable(s) entered on step 1: How old were you at your last birthday?, Are you a high school graduate?, Compared to your parents when they were the age you are now do you think your own standard of living now is better or worse as theirs was?, Can you read at least some English yes or no?, Gender.

The global, as well as individual hypothesis on this model, was checked again. The p-value of all the independent variables is significant and hence we can consider this as the final model with optimized R^2 value.

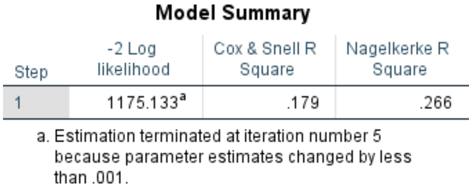
V. MODEL SUMMARY

As per our analysis, **Model 2** is considered to be the final model to predict if the person owns a mobile or not.

Below is the summarized analysis of our final model concerning the tables mentioned above in Model 2:

Model Summary:

The Model Summary table is the first table of importance. The Cox and Snell R square value .179 and Nagelkerke R square value .266 as obtained in the below summary table are similar to the R square value in multiple regression. These are known as pseudo R square statistics. They define the variance in the dependent variable predicted by independent variables. In theory, the maximum value is 1 when the relationship is perfect and 0 when there is no relationship. Another parameter explained in this table is -2 Log likelihood which helps in comparing the model, the lower the value -2 Log likelihood the better the model fit.



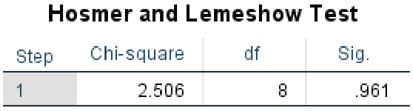
Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1175.133[a]	.179	.266

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test:

This is the test for model fit where a significance level less than 0.0 (p-value<0.05) is indicated by the poor fit. The below table shows the test results stating that the model has a good fit.

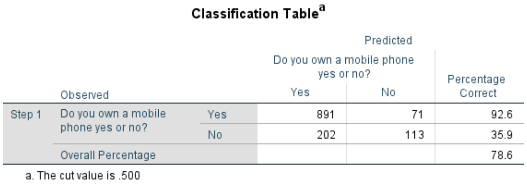


Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	2.506	8	.961

Classification table :

The classification table shows the overall percentage accuracy in the model which is 78.6% as well as the individual correct percentages of the classified variables. The "Yes" category of the dependent variable "Do you own a mobile phone or not?" is classified 92.6% correctly whereas the "No" category is classified 35.9% correctly.



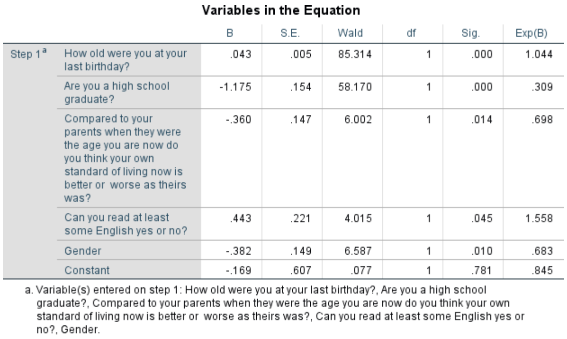
Classification Table[a]

			Predicted		
			Do you own a mobile phone yes or no?		Percentage Correct
	Observed		Yes	No	
Step 1	Do you own a mobile phone yes or no?	Yes	891	71	92.6
		No	202	113	35.9
	Overall Percentage				78.6

a. The cut value is .500

Variables in the equation:

Variables in the equation are the key output of the model. The significant value for all the variables states that they contribute significantly to predict the dependent variable. The Wald value represents the test statistic that tests the null hypothesis that β = 0. B values are similar to the β values in multiple regression.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	How old were you at your last birthday?	.043	.005	85.314	1	.000	1.044
	Are you a high school graduate?	-1.175	.154	58.170	1	.000	.309
	Compared to your parents when they were the age you are now do you think your own standard of living now is better or worse as theirs was?	-.360	.147	6.002	1	.014	.698
	Can you read at least some English yes or no?	.443	.221	4.015	1	.045	1.558
	Gender	-.382	.149	6.587	1	.010	.683
	Constant	-.169	.607	.077	1	.781	.845

a. Variable(s) entered on step 1: How old were you at your last birthday?, Are you a high school graduate?, Compared to your parents when they were the age you are now do you think your own standard of living now is better or worse as theirs was?, Can you read at least some English yes or no?, Gender.

The value B is the coefficient by which each predictor computes the dependent variable if all the remaining predictors are constant in the regression equation.

β_0 = -.169 (constant)

β_1=.043, an increase in the age of a person by 1 unit increases the Log odds of owning a mobile phone by .043.

β_2=-1.175, an increase in high-school graduates by 1 unit reduces the Log odds of owning a mobile phone by 1.175.

β_3=-.360, an increase in the standard of living by 1 unit reduces the Log odds of owning a mobile phone by .360.

β_4=.443, an increase in the ability to read English by 1 unit increases the Log odds of owning a mobile phone by .443.

β_5 =-.382, an increase in the gender category by 1 unit reduces the Log odds of owning a mobile phone by .382.

Final Regression equation:

Lop(p/1-p) = (-.169) + .043 (Age) -1.175 (High-School graduate) -.360 (Standard of living) + .443 (Ability to read English) -.382(Gender)

Odds Ratio:

The Exp(B) values are the odds ratio for each variable in the equation. As the odds ratio increases the probability of the occurring outcome increases. It can be explained as the odds that a person is owning a mobile phone is reduced by (0.698*0.698) if a person's standard of living increases by 2 units. Also, the odds that a person owns a mobile phone is 1.558 times higher if a person can read English when all other factors are equal.

VI. CONCLUSION

After performing the logistic regression analysis on the dataset to know if a person owns a mobile phone or not, it can be stated that the probability of a person owning a mobile phone is the largest for those who can read English. The other factors like age, the standard of living, education, and gender also help in defining the probability of owning a mobile phone.

VII. REFERENCES

[1] Peng, Joanne & Lee, Kuk & Ingersoll, Gary. (2002). An Introduction to Logistic Regression Analysis and Reporting. Journal of Educational Research - J EDUC RES. 96. 3-14. 10.1080/00220670209598786.

Part C: Principal Component Analysis

I. Introduction to Principal Component Analysis

PCA is the dimensionality reduction approach that is applied where the high dimensional datasets are to be transformed into smaller ones keeping almost all the information provided in the large dataset. Lowering the number of variables would naturally affect the accuracy, but this is the trick of dimensionality reduction to trade a little accuracy for making the datasets simpler and easier. These datasets are very much easier to explore and visualize and also analyzing such data becomes easier and faster.

II. Dataset Chosen for Illustrating Principal Component Analysis

The dataset selected to explain PCA is sourced from Kaggle (https://www.kaggle.com/). This dataset contains all the outfield player's data from every season in Major League Soccer(MLS). This dataset contains data for 15076 players identified as rows and 28 columns with player's attributes of which 20 columns are considered in performing PCA.

III. Steps Involved in Performing Principal Component Analysis with Example

1) *Selecting and measuring a set of variables :*

Sample Size:

Correlation coefficients tend to be less reliable when the sample size is small. PCA should be ideally performed on the dataset with more than 20 columns. Generally, it is recommended that there should be at least 5 observations for each variable. In the example of all players data in MLS, the sample size is 15076 with 20 which is adequate for performing PCA and hence fulfill the assumption of sample size.

The relationship between the variables:

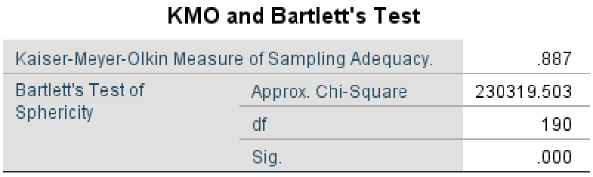
This step aims to study if there is any relationship between the variables. Sometimes, the variables are highly correlated with each other and contain reductant data such correlations need to be identified with the help of a correlation/covariance matrix.

Bartlett's Test of significance is for testing the null hypothesis that all the variables in the dataset have zero correlations. In this test, p-values should be significant(P<0.05) to reject the null hypothesis.

The below table shows that Bartlett's Test of Sphericity is significant, hence rejecting the null hypothesis and stating that the variables for all the players in MLS have at least some correlation and is not equal to zero.

Kaiser-Meyer-Olkin(KMO) is the measure of sampling adequacy. This has a value between 0 and 1, and values close to 1 are considered to be better. For a good PCA, KMO should exceed 0.5.

After performing PCA on the players' data the KMO is obtained to be 0.887 which is pretty more than 0.5 and is considered to be good sampling adequacy.



KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.887
Bartlett's Test of Sphericity	Approx. Chi-Square	230319.503
	df	190
	Sig.	.000

2) *Extracting the factors*

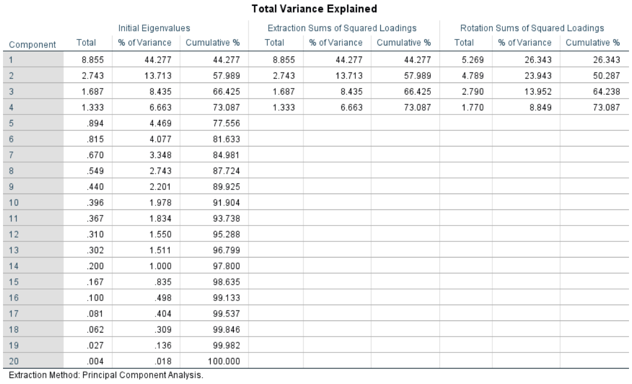
This step involves selecting the number of factors that best decides the relationship within the variables. Extracting factors in PCA is the "first pass" in finding the appropriate factors.

Factor extraction has two conflicting requirements::

1. Rationalize the number of factors, keep it as small as possible
2. Maximize the total variance explained

Eigenvalue:

The eigenvalue is the variance measuring factor of the specific components in PCA. The components in PCA must be selected such that the total variance is maximized. The below table shows the eigenvalue for each component. There are four components which are having an eigenvalue greater than 1. They explain the total variance of 73.08% when considering all four components. The third and fourth components have eigenvalues very close, so there can be an argument in whether both the components should be considered in the analysis or only the top three must be taken. The only first component explains a very high eigenvalue equal to 8.855 and 44.27 % of the variance.



Total Variance Explained

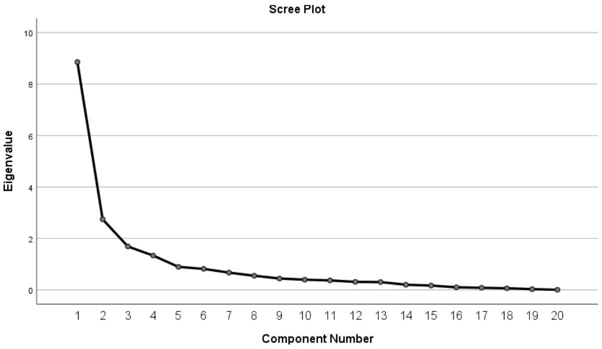
	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	8.855	44.277	44.277	8.855	44.277	44.277	5.269	26.343	26.343
2	2.743	13.713	57.989	2.743	13.713	57.989	4.789	23.943	50.287
3	1.687	8.435	66.425	1.687	8.435	66.425	2.790	13.952	64.238
4	1.333	6.663	73.087	1.333	6.663	73.087	1.770	8.849	73.087
5	.894	4.469	77.556						
6	.815	4.077	81.633						
7	.670	3.348	84.981						
8	.549	2.743	87.724						
9	.440	2.201	89.925						
10	.396	1.978	91.904						
11	.367	1.834	93.738						
12	.310	1.550	95.288						
13	.302	1.511	96.799						
14	.200	1.000	97.800						
15	.167	.835	98.635						
16	.100	.498	99.133						
17	.081	.404	99.537						
18	.062	.309	99.846						
19	.027	.136	99.982						
20	.004	.018	100.000						

Extraction Method: Principal Component Analysis.

There are a few techniques that may help in PCA:

- **Kaiser's Criterion**: In Kaiser's Criterion, factors with eigenvalues equal to 1 or more are selected.

- **Catell’s scree test**: In this test, each eigenvalue of the factors are plotted. The factors that are above the elbow or breakpoint in the plot are selected.

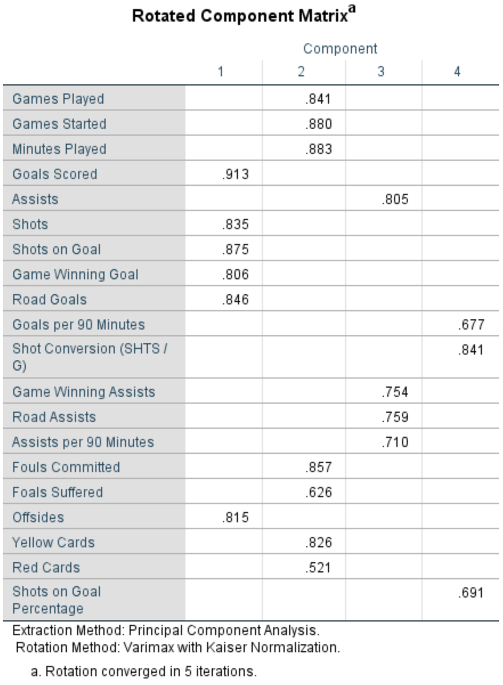


In the above scree plot, the first four factors having eigenvalues more than 1 are extracted for analysis. This satisfies the Kaiser’s Criterion as well as Catell’s scree test.

3) Rotate the factors:

This step is undertaken so that the factors become more interpretable without affecting their original properties. Rotating the factors is done to maximize the correlation between the variables and also reduces the low ones. The varimax rotation purifies the columns of the given matrix, such that each component is defined by limited variables.

The below table explains the rotated component matrix for PCA of the player’s data in MLS. It explains each component and categorizes the variables into four different components according to the correlation between them. The table has eliminated the factors that have a correlation coefficient of less than 0.5 as such factors are suppressed.



Rotated Component Matrix[a]

	Component			
	1	2	3	4
Games Played		.841		
Games Started		.880		
Minutes Played		.883		
Goals Scored	.913			
Assists			.805	
Shots	.835			
Shots on Goal	.875			
Game Winning Goal	.806			
Road Goals	.846			
Goals per 90 Minutes				.677
Shot Conversion (SHTS / G)				.841
Game Winning Assists			.754	
Road Assists			.759	
Assists per 90 Minutes			.710	
Fouls Committed		.857		
Foals Suffered		.626		
Offsides	.815			
Yellow Cards		.826		
Red Cards		.521		
Shots on Goal Percentage				.691

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

a. Rotation converged in 5 iterations.

4) Interpreting the result:

After all the steps, an important step of analysis is to interpret the result. The analysis should make sense. Not necessarily all the PCA fit well and give good interpretations.

The four components in the above-rotated matrix can be named with the respective factors present in them. The 1st component consists of the factors explaining the data related to goals made by the player, so we can name the component as “Different goal statistics”. The 2nd component explains the overall performance of the player for the games he played and so this can be named “Player Performance”. The third component is all about statistics for assists made by the player, which can be therefore named as “Statistics for assists”. The last component in the below matrix, somewhat explain shots statistics and so can be named as “shots statistics”.

IV. Conclusion:

Principal component analysis successfully reduced the dimension of the soccer dataset from 20 player attributes reducing it to 4.