Time Series, Binary Logistic Regression and Principal Component Analysis

Deepak Kumar Swain, x19216769@student.ncirl.ie, MSc Data Analytics, National College of Ireland

***Abstract*—In this paper, an attempt is made to use Time Series analysis, Logistic Regression analysis and Principal component analysis on necessary datasets using SPSS and R.**

TIME SERIES ANALYSIS:

OBJECTIVE

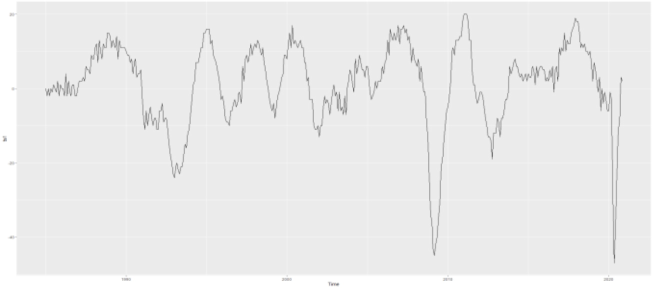
The objective of time series analysis is to understand the type of time series of the dataset and apply different time series models which are suitable for the data. Also, necessary comparisons need to be done on the basis of appropriateness of the model with the data, Root Mean Squared Errors (RMSE) and Akaike Information Criterion (AIC) values, the optimum model shall be used to perform forecasting for upcoming periods.

ABOUT THE DATASET:

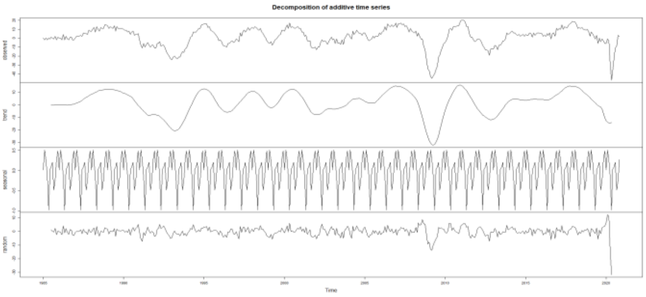
The dataset[1] for time series is collected from Europa website containing business and consumer survey details. The selected dataset is providing details about Industry/Business Climate Indicator. This dataset contains data from January 1985 to November 2020. For the time series analysis, a [1]particular column has been selected which shows the total production trend of European manufacturing industries on a monthly basis.

TIME SERIES ANALYSIS:

Let us plot the time series to understand the underlying pattern.



The above plot seems to be showing a horizontal patter which means the data fluctuate around a constant mean. It indicates that the time series is stationary. Also, the plot seems to be showing a sign of seasonality at certain point of times. The plot is also showing irregular fluctuations over the time periods. However, the plot is not showing any signs of trend. But the data can be decomposed to see if there is any trend component in the data. The above plot clearly shows that multiplicative decomposition is not appropriate. So, additive decomposition is used to see the seasonal, trend and irregular component.



The above is a plot of additive decomposition clearly shows the sign of seasonality and irregularity.But, there is no sign of trend here. Hence, the suitable models for the time series seem to be ‘ETS Model’, ARIMA Model and SARIMA model. But we will be applying other models just to see the perfomance of those models. There can be a possiblity that simple models like naïve or seasonal naïve model can perform better than the above mentioned models.

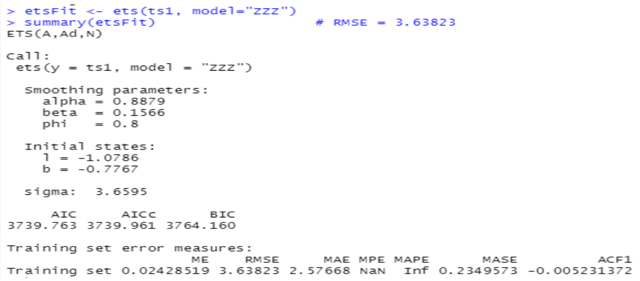
ETS (ERROR, TREND, SEASONAL) MODEL:

The ETS model is one of the most flexible models which can be used in simple exponential smoothing (time series without trend of seasonality), double exponential smoothing (time series with trend but no seasonality) and triple exponential smoothing (time series with trend and seasonality). The ets function in R is defined as below.

ets (TimeSeries, model= “Error Type, Trend Type, Seasonal Type”)

The available options for error type are: A – Additive, M – Multiplicative and Z – Automatic (Additive/Multiplicative). Whereas the available options for trend and seasonality are N- None, A- Additive, M- Multiplicative and Z- Automatic (N or A or M).

The easiest way of applying the ets model is to pass the model parameter as ZZZ and let the function decide the best parameter with optimum result. The ets model with parameter as ZZZ gave the following result.

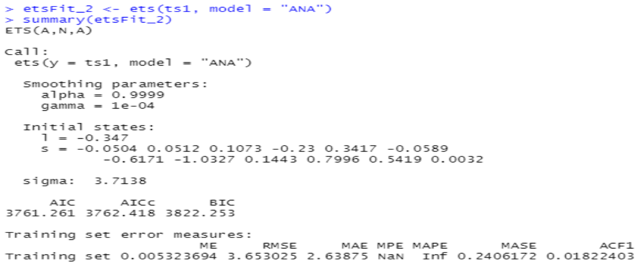


```
> etsFit <- ets(ts1, model="ZZZ")
> summary(etsFit)                    # RMSE = 3.63823
ETS(A,Ad,N)

Call:
 ets(y = ts1, model = "ZZZ")

  Smoothing parameters:
    alpha = 0.8879
    beta  = 0.1566
    phi   = 0.8

  Initial states:
    l = -1.0786
    b = -0.7767

  sigma:  3.6595

     AIC     AICc      BIC
3739.763 3739.961 3764.160

Training set error measures:
                     ME    RMSE     MAE MPE MAPE      MASE         ACF1
Training set 0.02428519 3.63823 2.57668 NaN  Inf 0.2349573 -0.005231372
```

The automatic parameterized ets model is giving RMSE value as 3.63823 which looks like a fair value but we can’t reach to the conclusion yet. Also, the model summary is showing the values of alpha, beta and phi. This is indicating the model parameter was taken as AAN (additive trend, no seasonality). Basically it is behaving like a Holt’s model. Earlier, we saw that there was no sign of trend but there was sign of seasonality in the data. So, ANN doesn’t seem to be appropriate even if it performs well. The appropriate ets

[1] https://ec.europa.eu/info/business-economy-euro/indicators-statistics/economic-databases/business-and-consumer-surveys/download-business-and-consumer-survey-data/time-series_en

model seems to ets model with parameter ANA which is considering additive seasonality. However, the permance is not as good as the model with ANN parameter since the RMSE value is showing 3.653025. Also, the Akaike Information Criterion (AIC) value for ANA is more than ANN(got this from ZZZ) model.



```
> etsFit_2 <- ets(ts1, model = "ANA")
> summary(etsFit_2)
ETS(A,N,A)

Call:
 ets(y = ts1, model = "ANA")

  Smoothing parameters:
    alpha = 0.9999
    gamma = 1e-04

  Initial states:
    l = -0.347
    s = -0.0504 0.0512 0.1073 -0.23 0.3417 -0.0589
           -0.6171 -1.0327 0.1443 0.7996 0.5419 0.0032

  sigma:  3.7138

     AIC     AICc      BIC
3761.261 3762.418 3822.253

Training set error measures:
                      ME     RMSE     MAE MPE MAPE      MASE       ACF1
Training set 0.005323694 3.653025 2.63875 NaN  Inf 0.2406172 0.01822403
```

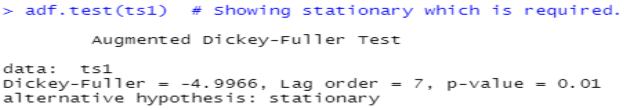
Multiplicative model hasn’t been considered for the time series as it is not appropriate.Also, the ets models with consideration of multiplicative parameters did not perform well as compared to models with additive parameters.

ARIMA MODEL:

Autoregressive Integrated Moving Average (ARIMA) model works well with stationary time series data or with the data that can be stationarized. ARIMA model is an integrated approach of taking both autoregressive and moving average into account. Autoregressive approach is a linear function of recent actual values. Wehereas, the moving average is a linear function of residuals. Both these methods do not pay much important on the data which are not recent. ARIMA model also has another factor called Integrated which takes trends into considaration. ARIMA model is define as ARIMA (p,d,q) where p represnts the value of autoregressiveness, q rpresents the moving average value and d represents the order of difference that needs to be used to make the time series stationary.

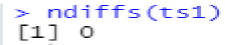
Earlier, we saw that the time series plot was looking like a stationary series. So, ARIMA model can be suitable for this kind of data. However, we did see some signs of seasonality in the time series. Lets see how the ARIMA model performs for the time series.

The key assumption in ARIMA model is that the time series should be stationary. To evaluate if the time series is stationary, Augmented Dickey-Fuller (ADF) test is performed which gave below results.



```
> adf.test(ts1)  # showing stationary which is required.

        Augmented Dickey-Fuller Test

data:  ts1
Dickey-Fuller = -4.9966, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary
```

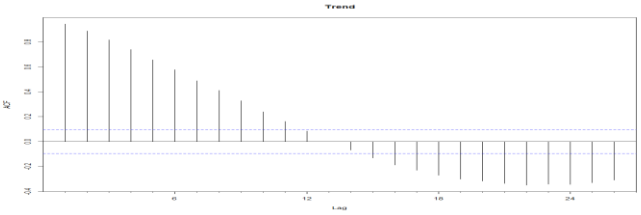
As we can see here, the p-value is significant which means the alernative hypohesis is true, i.e. , the time series is stationary. The difference function can be used to convert a time series to stationary by removing the trend and seasonality. In the earlier plot we saw that the time series was looking stationary but it is good to check if there is any need of taking higher order difference to make the series stationary. In R, diff() funtion is used to make the time series stationary by taking the difference of required order. To know which ordered difference was taken by the function we can use ndiff() function.

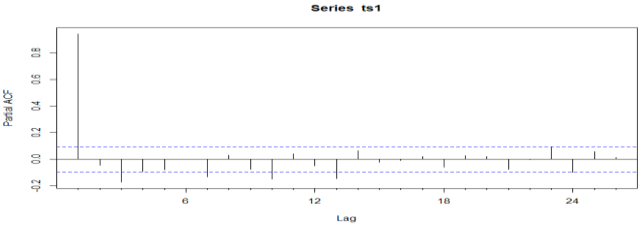


```
> ndiffs(ts1)
[1] 0
```

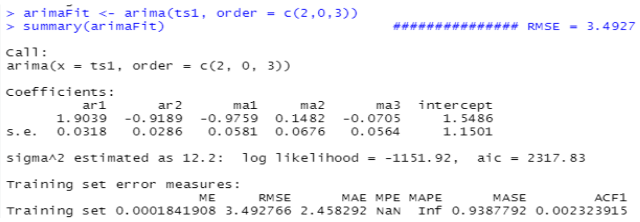
The above value shows that there was no difference taken by the function which proved that the time series is already in stationary state. So, the time series do not need to integrated anymore.

Now, lets have a look at the autocorrelation (ACF) plot which shows the correlation of time series with its own past values. The blue dashed lines in the following ACF plot shows whether the corrleations are drastically different from zero. As we can see in the ACF plot, there are significant spikes in the initial lags.



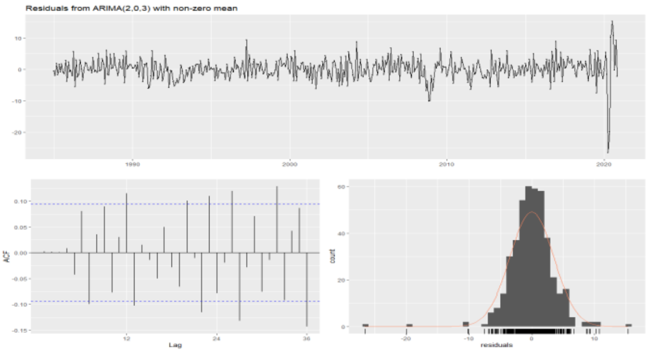


While applying the ARIMA model all the possible combination of p, d, q values were taken between 0 and 3. Also, for the selection of the p and q values, ACF and PCF plots were refered respectively. In the ACF plot, we can see spikes in the initial lags (lag 1, 2 and 3). Whereas in PCF plot, the spikes are at lag 1 and lag 3 (Later lags are also showing spikes but ARIMA model focuses more on initial lags). So, the suitable parameters for the models seems to be a combination of these numbers. Earlier, we have already found out the value of d as 0. All the above mentioned combinations of p, d and q values were used to find the optimum ARIMA model.

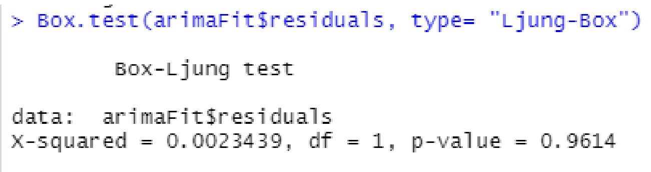


```
> arimaFit <- arima(ts1, order = c(2,0,3))
> summary(arimaFit)                              ############### RMSE = 3.4927

Call:
arima(x = ts1, order = c(2, 0, 3))

Coefficients:
         ar1      ar2      ma1     ma2      ma3  intercept
      1.9039  -0.9189  -0.9759  0.1482  -0.0705     1.5486
s.e.  0.0318   0.0286   0.0581  0.0676   0.0564     1.1501

sigma^2 estimated as 12.2:  log likelihood = -1151.92,  aic = 2317.83

Training set error measures:
                        ME     RMSE      MAE MPE MAPE      MASE        ACF1
Training set 0.0001841908 3.492766 2.458292 NaN  Inf 0.9387792 0.002323915
```

Finally, the optimum ARIMA model was received with p, d and q values as 2, 0 and 3 respectively with a root mean sqare value as 3.4927. So, ARIMA(2,0,3) model performs better than ets model. The residual plot in the folllowing ARIMA(2,0,3) model is showing like a normal distribution which makes the model a good fit.



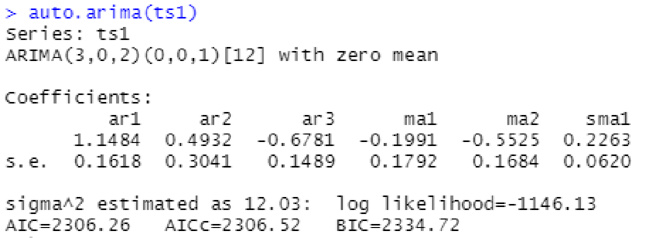
Also. In the ACF plot, there are no spikes beyond the threshold values in th initial lags. This shows that there is no high autocorrelation. An alternative test called ‘Ljung Box’ test is also perfomred to check if the model fits the data well.



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

        Box-Ljung test

data:  arimaFit$residuals
X-squared = 0.0023439, df = 1, p-value = 0.9614
```

As we can clearly see from the above result, the p- value is showing insignifiant value (p-value > 0.05). This indicates that the autocorrelations are close to zero. Hence, the ARIMA model seems to fit the data well.

However, when the AUTO.ARIMA function was used on the original time series to check the best possible values for (p,d,q), the following output was received.



```
> auto.arima(ts1)
Series: ts1
ARIMA(3,0,2)(0,0,1)[12] with zero mean

Coefficients:
         ar1     ar2      ar3      ma1      ma2    sma1
      1.1484  0.4932  -0.6781  -0.1991  -0.5525  0.2263
s.e.  0.1618  0.3041   0.1489   0.1792   0.1684  0.0620

sigma^2 estimated as 12.03:  log likelihood=-1146.13
AIC=2306.26   AICc=2306.52   BIC=2334.72
```

The optimum model suggested by auto ARIMA is showing the p, d,q values as 3,0,2 respectively. Along with the p,d and q values, it is also showing the parameters for seasonality as 0,0,1 with a seasonal frequency of 12. This is indicating that there is a presence of seasonal component in the time series which is not surprising as we encountered the signs of seasnality in the monthplot. So, it will be interesting apply the SARIMA model which takes the seasonality into consideration.

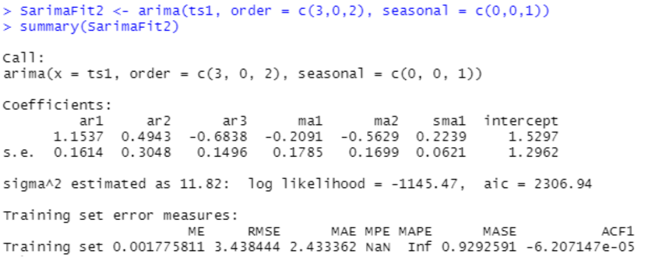
SARIMA MODEL:

Seasonal factors are taken into account within a seasonal ARIMa model. A SARIMA model in R is denoated as below.

$$\text{SARIMA } (p,d,q)\ (P,D,Q)_m$$

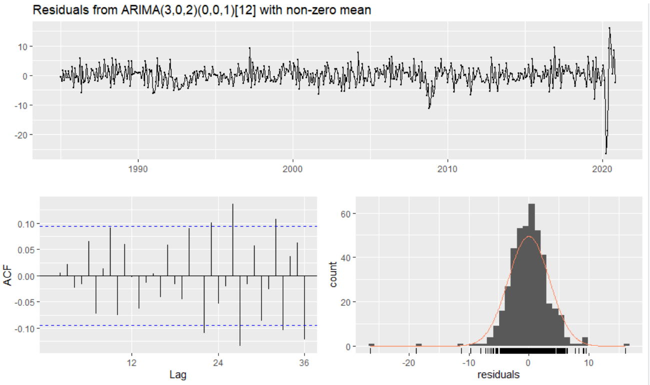
Where, the lower cased p, d and q represent the order for autoregressiveness, difference to convert a series to stationary and moving average order respectively. The uppercased P,D and Q represent the seasonal autoregressive order, seasonal difference order and seasonal moving averge order respectively. The parameter m represents the number of observations per year. For the current time series, the observation is taken on a monthly basis. Hence, the value of m should be 12.

To apply the optimum SARIMA model, auto.arima() funtion is used in R. The suggetion given by auto arima function was ARIMA(3,0,2)(0,0,1)[12]. Hence, the same parameters were given in the SARIMA model which gave following result.

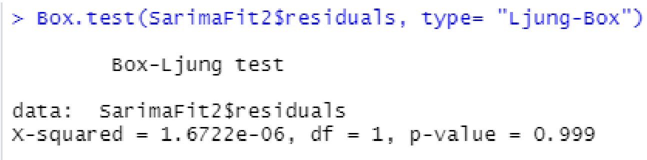


```
> SarimaFit2 <- arima(ts1, order = c(3,0,2), seasonal = c(0,0,1))
> summary(SarimaFit2)

Call:
arima(x = ts1, order = c(3, 0, 2), seasonal = c(0, 0, 1))

Coefficients:
         ar1     ar2      ar3      ma1      ma2    sma1  intercept
      1.1537  0.4943  -0.6838  -0.2091  -0.5629  0.2239     1.5297
s.e.  0.1614  0.3048   0.1496   0.1785   0.1699  0.0621     1.2962

sigma^2 estimated as 11.82:  log likelihood = -1145.47,  aic = 2306.94

Training set error measures:
                     ME     RMSE      MAE MPE MAPE      MASE          ACF1
Training set 0.001775811 3.438444 2.433362 NaN  Inf 0.9292591 -6.207147e-05
```

As we can see here the root mean squared value is showing as 3.4384 which is better than the ets and ARIMA models. Also, we can see that the AIC value is 2306.94 which is less than ETS and ARIMA models.



The residual and ACF plots are also indicating that the model is a good fit fot the data as the risidual plot shows the normal distribution of residuals and ACF plot doesn’t show any spikes at the initial lags. Also, the p-value in the box test showed insignifiant value.



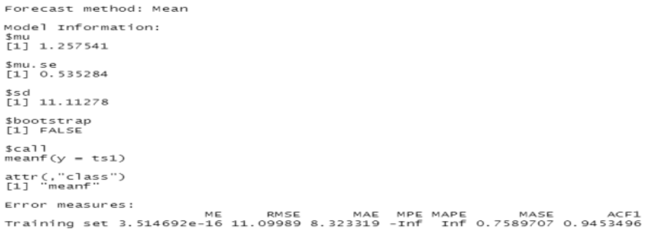
```
> Box.test(SarimaFit2$residuals, type= "Ljung-Box")

        Box-Ljung test

data:  SarimaFit2$residuals
X-squared = 1.6722e-06, df = 1, p-value = 0.999
```

Hence, SARIMA model outperforms the earlier two models with optimum perfomance.

Apart from these three model, few more models were applied to the time series irrespective of the appropriateness just to see how well they are performing. Infact, in certain scenarios, simple models like naïve and seasonal naive models can fit well with the time series.

MEAN MODEL:

The mean model is suitable for a time series which is stationary and random. Basically the mean model takes the mean of the pervious values to decide the future value. Since our time series is stationary, it can be a good option to apply mean model on it.

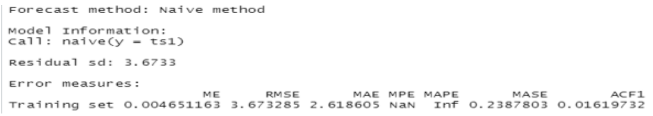


```
Forecast method: Mean

Model Information:
$mu
[1] 1.257541

$mu.se
[1] 0.535284

$sd
[1] 11.11278

$bootstrap
[1] FALSE

$call
meanf(y = ts1)

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

Error measures:
                       ME     RMSE      MAE  MPE MAPE      MASE      ACF1
Training set 3.514692e-16 11.09989 8.323319 -Inf  Inf 0.7589707 0.9453496
```

As we can see here, the RMSE value observed by mean model is 11.09989 which is not not a good value as compared to the earlier discussed models.

NAÏVE MODEL:

Naïve model forcasts the future values in a time series by taking the actual value of previous perios. Since it takes only last period’s data for forecasting, hitorical data points prior to the last period are of no use.

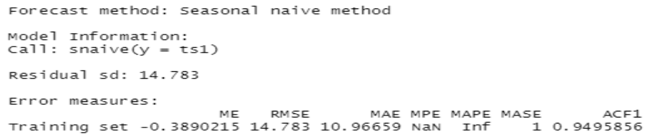


```
Forecast method: Naive method

Model Information:
Call: naive(y = ts1)

Residual sd: 3.6733

Error measures:
                      ME     RMSE      MAE MPE MAPE      MASE       ACF1
Training set 0.004651163 3.673285 2.618605 NaN  Inf 0.2387803 0.01619732
```

As we can see here, the naïve model is perfoming better than mean model but it is not as efficient as SARIMA model.

SEASONAL NAÏVE MODEL:

This model is an addon to the naïve model which takes the seasonal factor into consideration. Seasonal naïve model sets the forecast value same as the last observed value from the same season. We do have a sign of seasonality in our time series. So, it can be interesting to use snaive model for the time series.



```
Forecast method: Seasonal naive method

Model Information:
Call: snaive(y = ts1)

Residual sd: 14.783

Error measures:
                     ME   RMSE      MAE MPE MAPE MASE      ACF1
Training set -0.3890215 14.783 10.96659 NaN  Inf    1 0.9495856
```

As we can see here, the seasonal naïve model performs even worse than the mean model with an RMSE value 14.783.

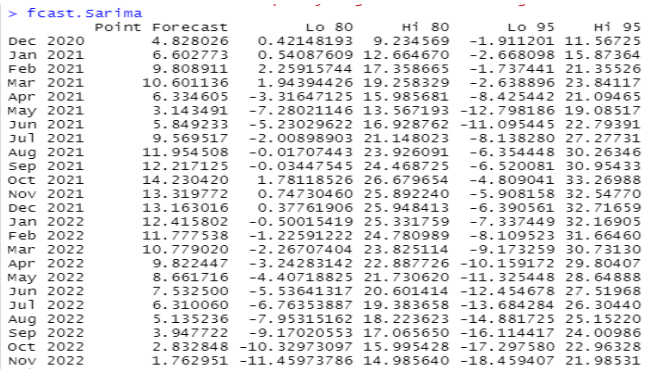
Apart from these models, few more models like Simple exponential, Holt's model and Holt-Winter's model were applied on the time series even though these were not the appropriate models for the current time series. These three models are generally used when there is a sign of trend in a time series. After applying several possible models, the following results are received.

Model	R Function	RMSE	AIC
SARIMA Model	arima(3,0,2)(0,0,1)[12]	3.438444	2306.94
ARIMA Model	arima(2,0,3)	3.492766	2317.83
Error, Trend, Seasonal	ets(ts,model='ZZZ') or ets(ts, model='ANN')	3.63823	3741.029
Error, Trend, Seasonal	ets(ts,model='ANA')	3.653025	3761.261
Holt-Winter Model	hw()	3.667565	3768.686
Simple Exponential Model	ses()	3.669027	3741.029
Holt's Model	holt()	3.669898	3745.234
Naive Model	naive()	3.673285	NA
Mean Model	meanf()	11.09989	NA
Seasonal Naïve Model	snaive()	14.783	NA

As we can see from the above details, SARIMA, ARIMa and ETS models are performing fairly better than other models. ETS(with model parameter as ANN), Holt-Winter, Simple exponential and Holt models are perfoming quite good as well but these models are not appropriate for the time series as the ets model did not take the seasnality parameter into consideration whereas the later models are appropriate for a time series that involves trends in it.

FINAL MODEL FOR FORECASTING:

As the time series which is being considered here is staionary with a sign of seasonality, it is no surprise that SARIMA model performed better than the other models. SARIMA model is quite suitable for these type of time series. Hence, this model will be used to do the forecasting for the time series.

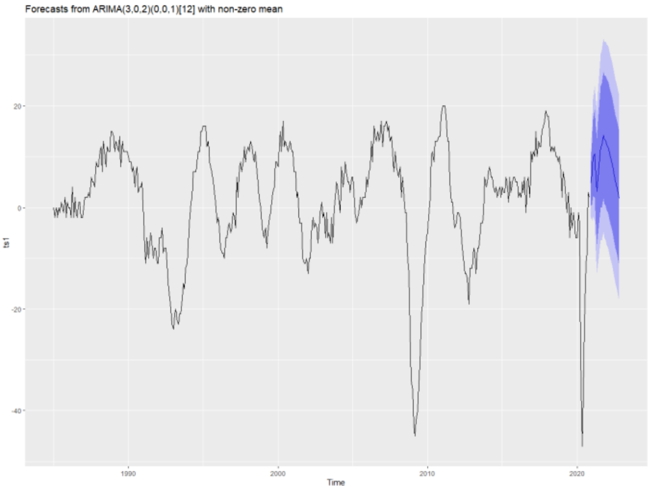


```
> fcast.Sarima
         Point Forecast        Lo 80     Hi 80      Lo 95    Hi 95
Dec 2020       4.828026   0.42148193  9.234569  -1.911201 11.56725
Jan 2021       6.602773   0.54087609 12.664670  -2.668098 15.87364
Feb 2021       9.808911   2.25915744 17.358665  -1.737441 21.35526
Mar 2021      10.601136   1.94394426 19.258329  -2.638896 23.84117
Apr 2021       6.334605  -3.31647125 15.985681  -8.425442 21.09465
May 2021       3.143491  -7.28021146 13.567193 -12.798186 19.08517
Jun 2021       5.849233  -5.23029622 16.928762 -11.095445 22.79391
Jul 2021       9.569517  -2.00898903 21.148023  -8.138280 27.27731
Aug 2021      11.954508  -0.01707443 23.926091  -6.354448 30.26346
Sep 2021      12.217125  -0.03447545 24.468725  -6.520081 30.95433
Oct 2021      14.230420   1.78118526 26.679654  -4.809041 33.26988
Nov 2021      13.319772   0.74730460 25.892240  -5.908158 32.54770
Dec 2021      13.163016   0.37761906 25.948413  -6.390561 32.71659
Jan 2022      12.415802  -0.50015419 25.331759  -7.337449 32.16905
Feb 2022      11.777538  -1.22591222 24.780989  -8.109523 31.66460
Mar 2022      10.779020  -2.26707404 23.825114  -9.173259 30.73130
Apr 2022       9.822447  -3.24283142 22.887726 -10.159172 29.80407
May 2022       8.661716  -4.40718825 21.730620 -11.325448 28.64888
Jun 2022       7.532500  -5.53641317 20.601414 -12.454678 27.51968
Jul 2022       6.310060  -6.76353887 19.383658 -13.684284 26.30440
Aug 2022       5.135236  -7.95315162 18.223623 -14.881725 25.15220
Sep 2022       3.947722  -9.17020553 17.065650 -16.114417 24.00986
Oct 2022       2.832848 -10.32973097 15.995428 -17.297580 22.96328
Nov 2022       1.762951 -11.45973786 14.985640 -18.459407 21.98531
```

The values showing under the level 'Point Forecast' show the time series values for the next three years along with monthly prediction values (December 2020 to November 2022). Apart from this column, four more columns are also shown which are sowing the convident/prediction interval values for 80% and 95%. LO80 and LO95 represent minimum values for the convidence intervals 80% and 95% whereas HI80 and HI95 represent the maximum values for the confidence intervals 80% and 95% respectively. Confidence interval of 80% tells that it is 80 percent sure that the estimated value will come with the given interval. Similarly, 95 percent confidence interval indicates that it is 95 percent confident that the estimated values would come with the given interval. The interesting part here is that the forecasted values are nothing but the mean of minimum and maximum values of the confidence interval.

Lets consider the forcasted value for December 2020 which is 4.828026. The mean of Lo80 and Hi80 is (0.42148193 + 9.234569)/2 = 4.8280254. Similarly, the average of Lo95 and Hi95 is (-1.911201+ 11.56725)/2 = 4.8280245. As we can see here the average of both the intervals are same as the forecasted value(ignoring the later decimal points).

The forecasted values can also be seen visually using autoplot.


Forecasts from ARIMA(3,0,2)(0,0,1)[12] with non-zero mean

The above plot shows the original time series data (black coloured) along with the forecasted data values (blue coloured). Also, we can see there are two types coloured areas shown with respect to the forecasted values. These are nothing but the prediction interval regions which is discussed earlier. The intervals showing in deep blue colour are the intervals with 80 percent confidence whereas the intervals showing in light blue colours are the intervals with 95 percent confidence.

BINARY LOGISTIC REGRESSION ANALYSIS:

OBJECTIVE:

The objective is to perform binary logistic regression analysis on a dataset using SPSS and R. In this analysis, the importance of dependent and independent variables should be clearly mentioned. Also, the assumptions based on which the model is build need to verified for the goodness of the fit. All the important measures and tests need to be checked and explained with proper justification.

ABOUT THE DATASET:

The dataset[2] is collected from ‘Pew Research Center’ website that contains information about people’s opinions, social matters. The selected dataset is about a political survey taken in March 2019 in the United States. The survey was taken among 1419 adults living in the United States. These people were interviews on a landline telephone or on a cell phone. A final weightage value has been given by combining all the samples. The following columns have been selected from the survey file to perform logistic regression analysis,

Details About Columns:

Name: q1

Description: This column represents one of the questions asked to te participants. The question is as follows.

“All in all, are you satisfied or dissatisfied with the way things are going in this country today?”

There were two options available for answering.

1-Satisfied, 2 – Dissatisfied

Name: q20

Description: This column asks the following question.

“Some people say they are basically content with the federal government, others say they are frustrated, and others say they are angry. Which of these best describes how you feel?”

The available options to answer this question were:

1-Baically content, 2- Frustrated, 3- Angry, 9- Don’t know

Name: q25

Description: This column is about the following question.

“How much of the time do you think you can trust the government in Washington to do what is right? Just about always, most of the time, or only some of the time? ”

The available options to answer the questions were:

1-Just about always,2- Most of the time, 3- Only some of the time, 4- Never, 9- Don’t know/Refused

Name: q47

Description: This asks the following question

“From what you have seen or heard about events in the new Congress, in general, do you think the Democrats in Congress are keeping the promises they made during the campaign, or not?”

The available options were:

1-Yes, keeping promises, 2- No, not keeping promises, 9- Don’t know

Name: q64

Description: This column provides the feedback for the following question.

“How fair do you think our present federal tax system is? ”

The available options were:

1-Very fair, 2- Moderately fair, 3- Not too fair, 4- Not fair at all, 9- Don’t know

Name: q70

Description: It shows the feedback for the following question.

“ Do you approve or disapprove of the tax law passed by Donald Trump and Congress in 2017?”

The available options were:

1-Approve, 2- Disapprove, 9-Don’t know

Name: q75

Description: It provides the feedback for the following question.

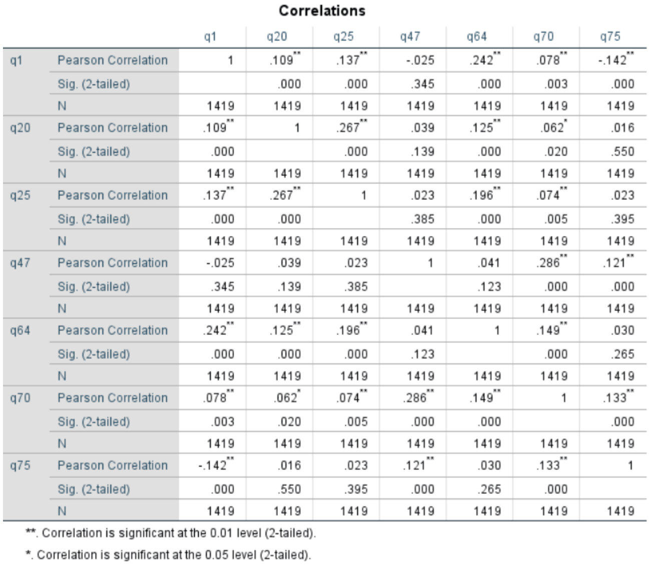
“How much, if anything, have you read or heard about allegations of misconduct by Donald Trump during his time in office or while he was running for president? ”

The available options to select were:

1-A lot, 2- A little, 3- othing at all, 9- Don’t know

VARIABLE SELECTION:

In a logical way, it seems that the main objective of the survey is to check if the people are satisfied with the way things are going in the United States. Also, the column q1 is having dichotomous values (1 and 2) and it seems to be suitable to take it as a dependent variable in the bi regression analysis. Other columns in the datasets seem to be impactful in deciding if a person is satified with the ongoing situations in the country. Lets have a look at the correlation matrix using SPSS.



Correlations

		q1	q20	q25	q47	q64	q70	q75
q1	Pearson Correlation	1	.109**	.137**	-.025	.242**	.078**	-.142**
	Sig. (2-tailed)		.000	.000	.345	.000	.003	.000
	N	1419	1419	1419	1419	1419	1419	1419
q20	Pearson Correlation	.109**	1	.267**	.039	.125**	.062*	.016
	Sig. (2-tailed)	.000		.000	.139	.000	.020	.550
	N	1419	1419	1419	1419	1419	1419	1419
q25	Pearson Correlation	.137**	.267**	1	.023	.196**	.074**	.023
	Sig. (2-tailed)	.000	.000		.385	.000	.005	.395
	N	1419	1419	1419	1419	1419	1419	1419
q47	Pearson Correlation	-.025	.039	.023	1	.041	.286**	.121**
	Sig. (2-tailed)	.345	.139	.385		.123	.000	.000
	N	1419	1419	1419	1419	1419	1419	1419
q64	Pearson Correlation	.242**	.125**	.196**	.041	1	.149**	.030
	Sig. (2-tailed)	.000	.000	.000	.123		.000	.265
	N	1419	1419	1419	1419	1419	1419	1419
q70	Pearson Correlation	.078**	.062*	.074**	.286**	.149**	1	.133**
	Sig. (2-tailed)	.003	.020	.005	.000	.000		.000
	N	1419	1419	1419	1419	1419	1419	1419
q75	Pearson Correlation	-.142**	.016	.023	.121**	.030	.133**	1
	Sig. (2-tailed)	.000	.550	.395	.000	.265	.000	
	N	1419	1419	1419	1419	1419	1419	1419

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

*. Correlation is significant at the 0.05 level (2-tailed).

As we can see in the above correlation matrix, no colums are highly correlated with each other which is good for deciding

[2] https://www.pewresearch.org/politics/dataset/march-2019-political-survey/

the independent variable. However, it is a bit tricky to decide which column to take as the dependent variable. To analyze further, lets have a look at another factor called significance level which is shown in the above matrix. The column q47 seems to have insignificant relation with most of the columns. Hence, it seems to be an irrelevant column for the analysis. Among the other columns, q70 is showing significance values with all the remaining columns whereas q1 is showing significance values with all the remaining columns except q47. As we saw earlier that q47 looks like an irrelevant column in the analysis, q1 and q70 seem to be a good candidate for the selection of dependent variable. The column q70 is a categorical variable but it is not dichotomous. Also, it is not appropriate to convert it to dichotomous (< 2 or >= 2) as it will change the meanng of the feedback. The column q1 is a dichotomous variable which is suitable to be treated as dependent variable. Hence, further analysis will be done by taking q1 as dependent variable and other five columns (excluding q47) as independent variables.

LOGISTIC REGRESSION MODEL USING SPSS & R:

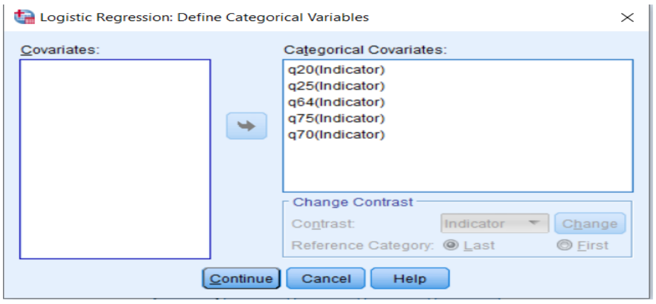
Logistic regression [1] is one of the most commonly used model for classification. It predicts the probability based on odds ratio to decide the class of the dependent variable. The linear regression model uses the linear equation Y= mX + C to predict the value of dependent variable and this can be of any numerical value. But, in case of logistic regression which deals with probabilities (response value must be between 0 and 1), linear eqution is not significant. Hence, logistic regression uses a different funcion called ‘Logit Function’ (Inverse of Sigmoid funtion) to calculate the probability of response variable.

$$\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

p = probability

$\frac{p}{1-p}$ = corresponding odds

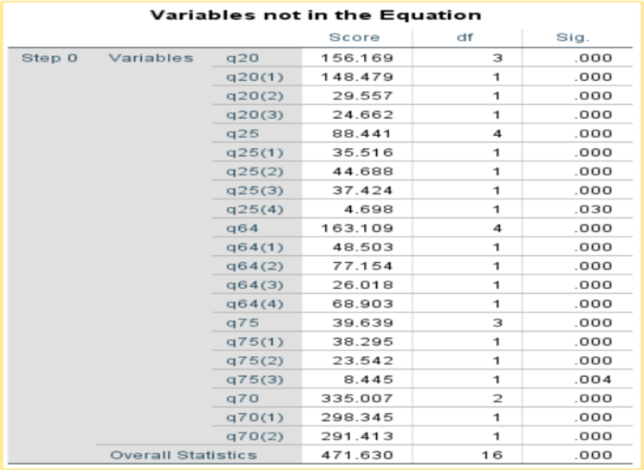
But there is one important aspect that needs to be considered while applying logistic regression model. If an independent variable is categorical and has more than two categories in it, the variable cannot be used directing in the equation as it is. The equation can only have the variables which are continuous or binary type. In the current dataset, all the independent variables are categorical with more than two categories. In SPSS, these categorical variables can be handled easily if we let the tool know which are the categorical features.



q20(Indicator)
q25(Indicator)
q64(Indicator)
q75(Indicator)
q70(Indicator)

Reference Category: ● Last ○ First

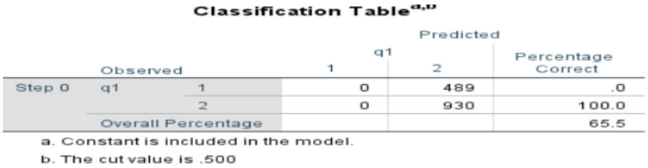
Once the variables are declared as categorical variables, SPSS creates the dummy columns for these features. As we can see in the following table, SPSS has created dummy columns for each categorical feature. One interesting thing that can be noticed here that one less dummy variable is created for each feature. For example, q20 have four categorical values(1,2,3,9) but the dummy columns are created as q20(1), q20(2) and q20(3). This is how the dummy column creation works. If the feature has K number of categories, (K-1) dummies will be created for this as this much of columns are enough to explain the feature. Also, it is done to reduce the correlations among the dummy columns.



Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	q20	156.169	3	.000
		q20(1)	148.479	1	.000
		q20(2)	29.557	1	.000
		q20(3)	24.662	1	.000
		q25	88.441	4	.000
		q25(1)	35.516	1	.000
		q25(2)	44.688	1	.000
		q25(3)	37.424	1	.000
		q25(4)	4.698	1	.030
		q64	163.109	4	.000
		q64(1)	48.503	1	.000
		q64(2)	77.154	1	.000
		q64(3)	26.018	1	.000
		q64(4)	68.903	1	.000
		q75	39.639	3	.000
		q75(1)	38.295	1	.000
		q75(2)	23.542	1	.000
		q75(3)	8.445	1	.004
		q70	335.007	2	.000
		q70(1)	298.345	1	.000
		q70(2)	291.413	1	.000
	Overall Statistics		471.630	16	.000

The dependent variable in the dataset has values has 1 and 2. SPSS encoded these values to binary values (1 as 0 and 2 as 1) to perform binary logistic regression. After running the model in SPSS, the first thing which is noticed is the ‘Null Model’ which runs the model without taking any explanatory variables into consideration.



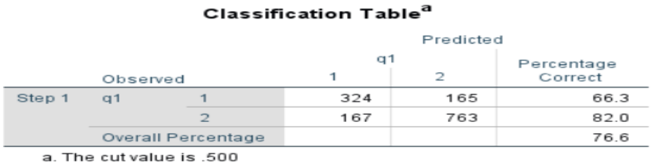
Classification Table[a,b]

			Predicted		
			q1		Percentage Correct
	Observed		1	2	
Step 0	q1	1	0	489	.0
		2	0	930	100.0
	Overall Percentage				65.5

a. Constant is included in the model.

b. The cut value is .500

The Null model is showing 65.5 percent accuracy in estimating the response value correctly. So, if we are going to include any predictors to esimate the resonse, the accuracy must be more than 65.5 percent. Else, it will indicate that the independent variables do not have significant impact on the model building. After including the independent variables, the following result received with an accuracy of 76.6 percent. It seems to be performing well. Here, the threshold probability value is taken as 0.5 which indicates that if the proability is less than 0.5, it will belong to the class 0. Else, it will come under class 1.

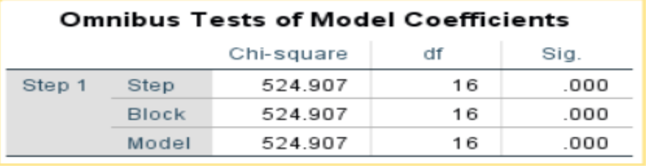


Classification Table[a]

			Predicted		
			q1		Percentage Correct
	Observed		1	2	
Step 1	q1	1	324	165	66.3
		2	167	763	82.0
	Overall Percentage				76.6

a. The cut value is .500

There is a test called ‘Omnibus test’ for model coefficents which explains whether coefficients of all the independent variables are zero or not. Through Chi-square distribution it decides the significance of each predictors.

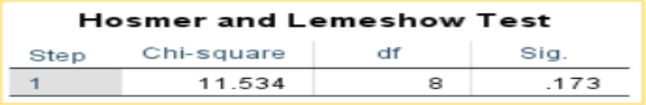


Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	524.907	16	.000
	Block	524.907	16	.000
	Model	524.907	16	.000

As wecan see here, the significance value is less than 0.05 which means it is not satisfying the null hypothesis which states that all the coefficients of the features are zero. Hence, the null hypothesis is rejected as per the Omnibus test.

A test called 'Hosmer & Lemshow test' is used to check if the model is a good fit with the data. It groups the cases into deciles of risk and then compares the observed probability with the expected probability within each decile.

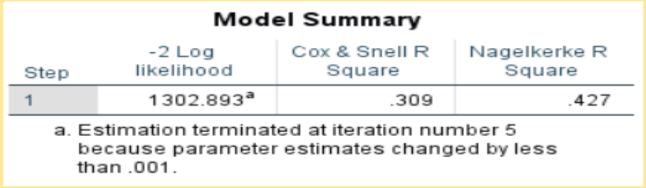


Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	11.534	8	.173

To measure the goodness of the model, the significance value is taken into consideration. If the value is more than 0.05, the model is considered to be good fit. Else, it is considered as a poor model. For the current dataset, significance value is showing 0.173 which indicates that the model is a good fit. However, we may not consider this as the best model as the value is not close to 1.

There is another factor called Pseudo R Square [2] statistics which is analogous to R squared measures in multilinear regression but the way of calculating the R squared value is different in both the cases. There are various methods of calculating the pseudo R squared values but two of the most widely used methods are 'Cox and Snell R Square' and 'Nagelkerke R Square'.



Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1302.893[a]	.309	.427

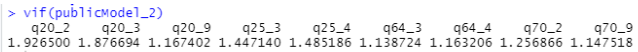
a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Both these R squared measures are compared with log likelihood for a baseline model. Theoritically, Cox & Snell's R square value ranges from 0 to a maximum value which is less than 1. Even for a perfect model, the Cox & Snell R square value does not rech to one. On the other hand, Negelkerke R square value , which is an adjusted method of Cox & Snell, ranges from 0 to 1. In several cases, Negelkerke's R square value is prefered over Cox & Snell's R square value. The current model is showing the Cox & Snell's and Nagelkerke's R squared value as 0.309 and 0.427. By looking at these value, we can say that it is fairly an ok model but still it cannot be considered s the most perfectly fit model.

So far the model showed acceptable results. However, it cannot be ignored that the model is built on certain assummtions.

The first assumption of the model is that the dependent variable outcomes should be mutually exclusive. For the current dataset, the response has two outcome, satisfied or not satisfied. This clearly shows that the outcomes are mutually exclusive as these are completey opposite of each other and do not have any commonality. Hence, the assumption doesn't seem to be violated here.

The second assumption is about multicollinearity. For a good model, the independent variables shouldn't be strongly related to each other. To measure the multicollinearity, a factor clled 'Variance Inflation Factor' (VIF) is used. If the vif values for the features are less than 5, this indicates that there is no strong correlation between the features.

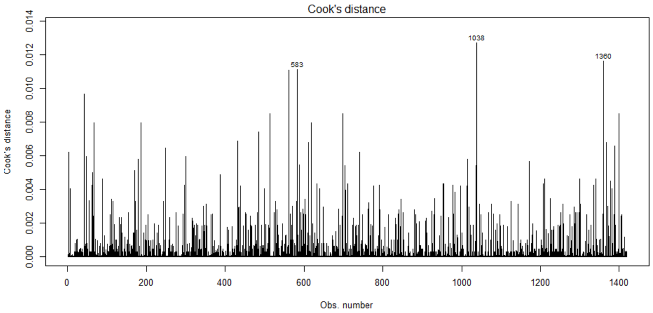


```
> vif(publicModel_2)
   q20_2    q20_3    q20_9    q25_3    q25_4    q64_3    q64_4    q70_2    q70_9 
1.926500 1.876694 1.167402 1.447140 1.485186 1.138724 1.163206 1.256866 1.147518 
```

As we can see above, the vif values for all the features are less than 5. Hence, there is no violation to the assumption.

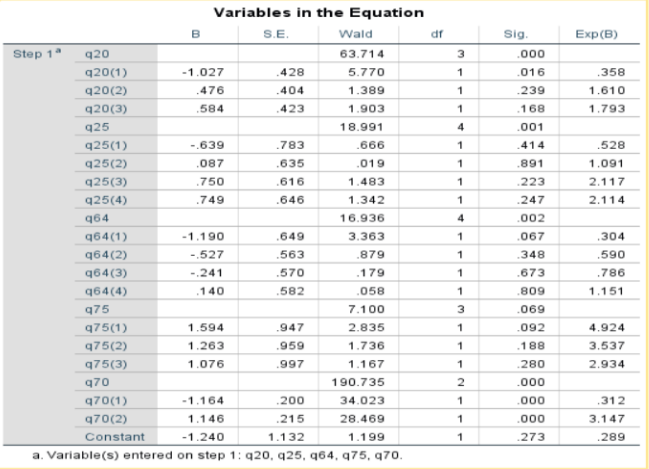
The third assumption is about sample size. Logistic regression works well with large samples. Ideally, if the number of cases should be atleast 20 times the number of predictors. For the current dataset, we have around 16 predictors (including the dummy columns). So the minimum suiatble cases for the model should be atleast 320 case. The dataset is having 1419 cases which is satisfying the assumption.

Like multiple linear regression, we can also plot diagnotics plots to analyze the goodness of the model. However, for a dichotomous response variable, it is not a good idea to rely on plots like residual plots. But we can check if there are any influential points by checking the cook's distance.



As we can see in the above plot, cook's distance values are not exceeding 0.014 of any of the observations. For violating the assumption, cook's distance should be more than 1. So, no violation to the assumption is seen here.

As we saw that the model is satisfying all the assumptions and also giving a fair result with an accuracy of 76.6 percent, we will hold onto this model. Let's discuss about one of the most important components in the model, 'Variables in the equation'.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	q20			63.714	3	.000	
	q20(1)	-1.027	.428	5.770	1	.016	.358
	q20(2)	.476	.404	1.389	1	.239	1.610
	q20(3)	.584	.423	1.903	1	.168	1.793
	q25			18.991	4	.001	
	q25(1)	-.639	.783	.666	1	.414	.528
	q25(2)	.087	.635	.019	1	.891	1.091
	q25(3)	.750	.616	1.483	1	.223	2.117
	q25(4)	.749	.646	1.342	1	.247	2.114
	q64			16.936	4	.002	
	q64(1)	-1.190	.649	3.363	1	.067	.304
	q64(2)	-.527	.563	.879	1	.348	.590
	q64(3)	-.241	.570	.179	1	.673	.786
	q64(4)	.140	.582	.058	1	.809	1.151
	q75			7.100	3	.069	
	q75(1)	1.594	.947	2.835	1	.092	4.924
	q75(2)	1.263	.959	1.736	1	.188	3.537
	q75(3)	1.076	.997	1.167	1	.280	2.934
	q70			190.735	2	.000	
	q70(1)	-1.164	.200	34.023	1	.000	.312
	q70(2)	1.146	.215	28.469	1	.000	3.147
	Constant	-1.240	1.132	1.199	1	.273	.289

a. Variable(s) entered on step 1: q20, q25, q64, q75, q70.

The first column (labeled as B) represent the coefficients of all the predictors. As expected, the variables q20, q25,q64,q70 and q75 do not have any coefficient values. This is because dummy clumns have been created for these column which made the original columns to be excluded. The 2nd and 4th columna are showing the standard errors and degree of freedom for each column. Due to the creation f dummy columns, the categorical columns are now showing df value as 1 instead of the original values. The third columns shows the Wald statistics values which is analogous to t test in linear regression. Based on the Wald test, the

$$\log\left(\frac{p(X)}{p(1-X)}\right) = -1.240 + (-1.027 * q20(1)) + (.476 * q20(2)) + (.584 * q20(3)) + (-.639 * q25(1)) + (.087 * q25(2)) + (.750 * q25(3)) + (.749 * q25(4)) + (-1.19 * q64(1)) + (-.527 * q64(2)) + (-.241 * q64(3)) + (.140 * q64(4)) + (1.594 * q75(1)) + (1.263 * q75(2)) + (1.076 * q75(3)) + (-1.164 * q70(1)) + (1.146 * q70(2))$$

The significance values shown in the ‘variables in the equation’ table are based on Wald statistics which decides which features have significant contribution to the model.

The last column in the table which is denoted as EXP(B) shows the values of odds ratios for each column. As we saw that all the features are categorical variables with more than two categories. To handle these categorical values, dummy columns are created. So, the original columns are not having any odds ratio values in the above shown table. These odds ratio values are reflecting for the dmmy columns which are nothing but the factors of the odds that the the value q1=1. In a dummy column, odds ratio for the value 1 can be calculated by dividing the ‘probabilty of occurance of 1’ by ‘the probability of no occurance of 1’.

Lets interpret the odds ratio value for the column ‘q20(1)’ which is showing as 0.358. This indicates that the odds of getting 1 correctly in ‘q20(1)’ (means the answer to the q20 as ‘Basically content’) compared to the odds of not getting 1 correctly.

PRINCIPAL COMPONENT ANALYSIS:

OBJECTIVE:

The objective of this analysis is to take a dataset having large number of columns (at least 20) and perform principal component analysis to reduce the number of variables to reasonably less number of components based on the percentage of explanatory variables. After finding the components, necessary justifications need to be provided. Also, the logical reason of variable groupings among the components need to be expressed.

ABOUT THE DATASET:

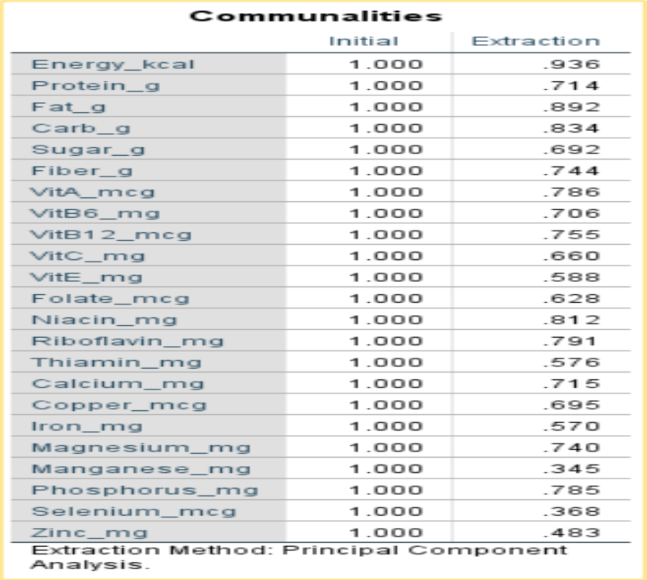
The dataset[3] [3] is collected from data world website. It contains the nutrient details from the United States Department of Agriculture. The available columns in the dataset is as below.

ID, FoodGroup, ShortDescrip, Descrip, CommonName, MfgName, ScientificName, Energy_kcal, Protein_g, Fat_g, Carb_g, Sugar_g, Fiber_g, VitA_mcg, VitB6_mg, VitB12_mcg, VitC_mg, VitE_mg, Folate_mcg, Niacin_mg, Riboflavin_mg, Thiamin_mg, Calcium_mg, Copper_mcg, Iron_mg, Magnesium_mg, Manganese_mg, Phosphorus_mg, Selenium_mcg, Zinc_mg

The first seven columns give the information about different products whereas the next 23 columns provide the details about different types of nutrient types along with the amount in each product. As we can notice here that the number of columns is very high which can create complexity in a model analysis. By looking at the date it seems that all the 38 columns are completely different and it is not easy to decide whether the number of columns can be reduced just by looking at the dataset. To tackle this kind of situation, ‘Principal Component Analysis’ concept is used.

MODEL ANALYIS:

Principal component analysis (PCA) is a dimensionality reduction process that transforms a large number of correlated variables to a fairly smaller uncorrelated variables called principal components. These components capture most of the information. Basically, the components act as a super variable that explains the data well by reducing the number of columns to components. It uses the eigenvalue concept to analyze the variance among the features and then forms the components by calculating the cumulative percentage of variance.



Communalities

	Initial	Extraction
Energy_kcal	1.000	.936
Protein_g	1.000	.714
Fat_g	1.000	.892
Carb_g	1.000	.834
Sugar_g	1.000	.692
Fiber_g	1.000	.744
VitA_mcg	1.000	.786
VitB6_mg	1.000	.706
VitB12_mcg	1.000	.755
VitC_mg	1.000	.660
VitE_mg	1.000	.588
Folate_mcg	1.000	.628
Niacin_mg	1.000	.812
Riboflavin_mg	1.000	.791
Thiamin_mg	1.000	.576
Calcium_mg	1.000	.715
Copper_mcg	1.000	.695
Iron_mg	1.000	.570
Magnesium_mg	1.000	.740
Manganese_mg	1.000	.345
Phosphorus_mg	1.000	.785
Selenium_mcg	1.000	.368
Zinc_mg	1.000	.483

Extraction Method: Principal Component Analysis.

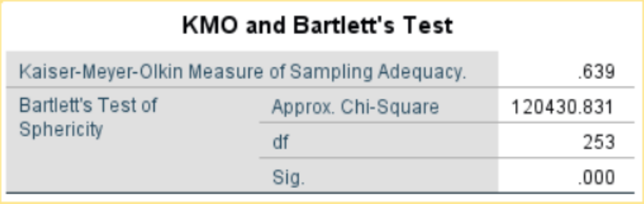
In the above communalites table, the ‘Initial’ columns shows the initial values for all the variables are taken as 1 whereas

[3] https://data.world/exercises/principal-components-exercise-1

the ‘Extraction’ shows the proportion of variance of each variables that can be explained by the components. As we can see, most of the variables have fairly high extraction values which is a good sign to proceed further with the analysis.

But before proceeding with further analysis, we need to check if the application of principal component analysis is appropriate for the data. We have two measure tests that explain this.

The first test is called ‘Bartlett’s Test of Sphericity’. This is a test on the null hypothesis that states that all true correlations between variables are zero. Bartlett’s test uses Chi-square statistics to find the significance level for the test. If the significance value is less than 0.05 (considered that the confidence interval is 95%), it means that the null hypothesis is turning out to be false. For the current dataset, significance value is coming as 0.000 which indicats that there are correlations between the variables in the dataset. So, we are good to proceed with the analysis.



KMO and Bartlett's Test

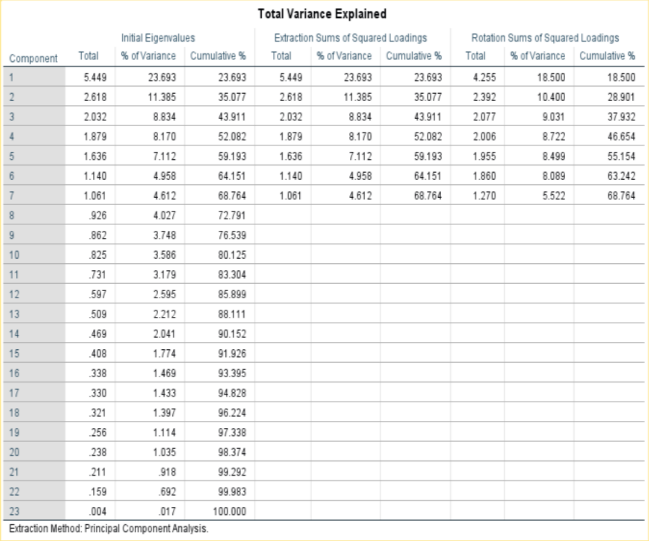
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.639
Bartlett's Test of Sphericity	Approx. Chi-Square	120430.831
	df	253
	Sig.	.000

The second test is called ‘Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO Measure)’. It is a measure to explain to what extent the correlation between pairs of variables can be explained by other variables. It is calculated with the following parameters.

$$KMO\ Adequacy = \frac{Sum\ of\ squared\ correlations}{(Sum\ of\ squared\ correlations) + (Sum\ of\ squared\ partial\ correlations)}$$

The maximum value for this is 1 which is the perfect value for the factor analysis wheras 0 is the minimum value. For a good factor analysis, the KMO test value should be more than 0.5. For the current model, KMO test value is showing as 0.639 which is a good situation for the factor analysis.

Now, lets have a look at the ‘Total Variance Explained’ table. As we can see in the following table, the first column is showing the component numbers. The second column is the showing the egenvalues for each components. Component 1 is showing an eigenvalue as 5.449 which is high as compared to other components whereas the components from 2 to 7 are showing fairly good eigenvalues.



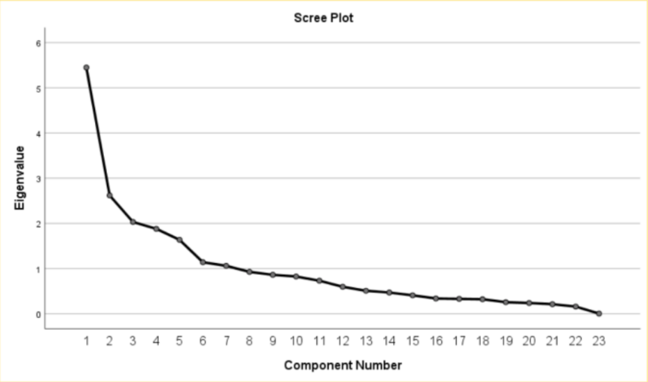
Total Variance Explained

	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	5.449	23.693	23.693	5.449	23.693	23.693	4.255	18.500	18.500
2	2.618	11.385	35.077	2.618	11.385	35.077	2.392	10.400	28.901
3	2.032	8.834	43.911	2.032	8.834	43.911	2.077	9.031	37.932
4	1.879	8.170	52.082	1.879	8.170	52.082	2.006	8.722	46.654
5	1.636	7.112	59.193	1.636	7.112	59.193	1.955	8.499	55.154
6	1.140	4.958	64.151	1.140	4.958	64.151	1.860	8.089	63.242
7	1.061	4.612	68.764	1.061	4.612	68.764	1.270	5.522	68.764
8	.926	4.027	72.791						
9	.862	3.748	76.539						
10	.825	3.586	80.125						
11	.731	3.179	83.304						
12	.597	2.595	85.899						
13	.509	2.212	88.111						
14	.469	2.041	90.152						
15	.408	1.774	91.926						
16	.338	1.469	93.395						
17	.330	1.433	94.828						
18	.321	1.397	96.224						
19	.256	1.114	97.338						
20	.238	1.035	98.374						
21	.211	.918	99.292						
22	.159	.692	99.983						
23	.004	.017	100.000						

Extraction Method: Principal Component Analysis.

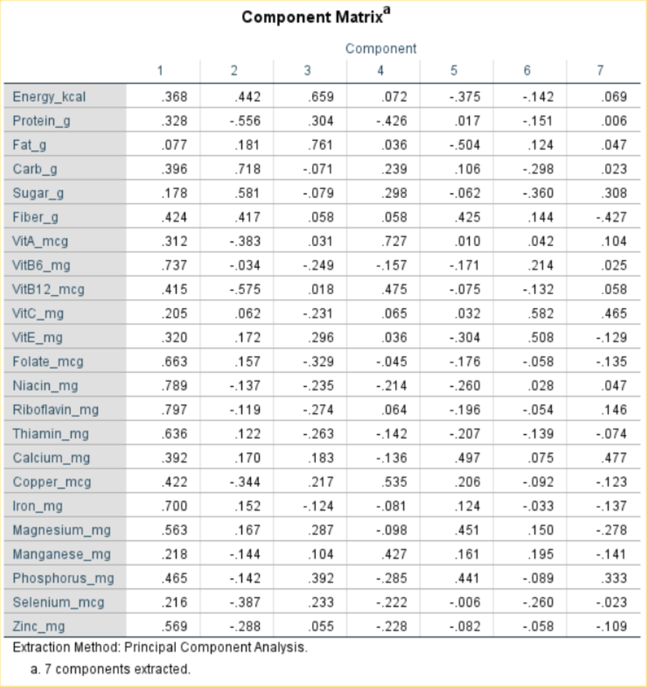
The third column provides the percentage of variances that are being explained by each components whereas the 4th column shows the cummulative values of percentage variances. From the above details, it is clearly seen that seven components are able to explain 68.764 percentage of variances of the features. This indicates that seven component creation should be enough for the principal component analysis. As a thumb rule, the components which have eigenvalues greater than one are considered to be good candidates to be included in the analysis.

But this is not enough to decide the number of components that need to be created. To get a better clarity on this, we can have a look at the following scree plot that explain to what extent a component is responsible in explaining the variance.



In general, scree means rocky debris lying on a slope or at the base of a hill or mountain. If we see in the above plot, it exactly depicts the similar visual. We can clearly see in this plot that the first component is having very high eigenvalues which is explaining arounf 23.5 percent of variance. Apart from this, there are six more components which are being pointed above the eigenvalue 1. Hence, it is a good idea to go with seven components as per the principal component analysis.

The next thing that needs o be paid attention is the component matix which shows the correlations of each columns with the respective components. Each item has a



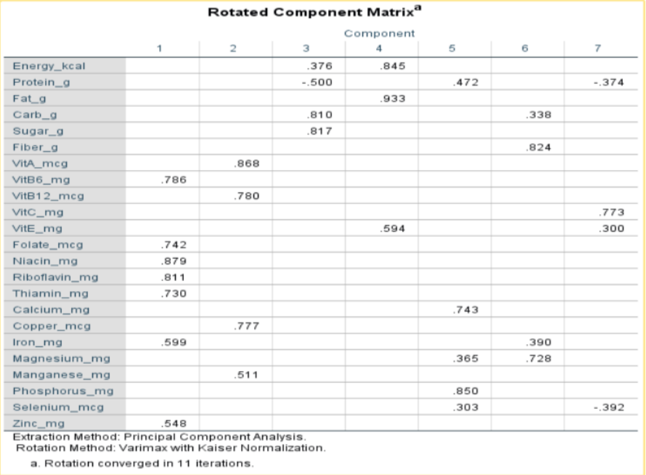
Component Matrix[a]

	Component						
	1	2	3	4	5	6	7
Energy_kcal	.368	.442	.659	.072	-.375	-.142	.069
Protein_g	.328	-.556	.304	-.426	.017	-.151	.006
Fat_g	.077	.181	.761	.036	-.504	.124	.047
Carb_g	.396	.718	-.071	.239	.106	-.298	.023
Sugar_g	.178	.581	-.079	.298	-.062	-.360	.308
Fiber_g	.424	.417	.058	.058	.425	.144	-.427
VitA_mcg	.312	-.383	.031	.727	.010	.042	.104
VitB6_mg	.737	-.034	-.249	-.157	-.171	.214	.025
VitB12_mcg	.415	-.575	.018	.475	-.075	-.132	.058
VitC_mg	.205	.062	-.231	.065	.032	.582	.465
VitE_mg	.320	.172	.296	.036	-.304	.508	-.129
Folate_mcg	.663	.157	-.329	-.045	-.176	-.058	-.135
Niacin_mg	.789	-.137	-.235	-.214	-.260	.028	.047
Riboflavin_mg	.797	-.119	-.274	.064	-.196	-.054	.146
Thiamin_mg	.636	.122	-.263	-.142	-.207	-.139	-.074
Calcium_mg	.392	.170	.183	-.136	.497	.075	.477
Copper_mcg	.422	-.344	.217	.535	.206	-.092	-.123
Iron_mg	.700	.152	-.124	-.081	.124	-.033	-.137
Magnesium_mg	.563	.167	.287	-.098	.451	.150	-.278
Manganese_mg	.218	-.144	.104	.427	.161	.195	-.141
Phosphorus_mg	.465	-.142	.392	-.285	.441	-.089	.333
Selenium_mcg	.216	-.387	.233	-.222	-.006	-.260	-.023
Zinc_mg	.569	-.288	.055	-.228	-.082	-.058	-.109

Extraction Method: Principal Component Analysis.

a. 7 components extracted.

loading that is corresponding to each of the seven components. For example, Energy_kcal has a correlation value of 0.368 with the 1st component. However, it is a bit complex to understand and interpret the matrix. To make is simpler, we can introduce the rotation technique along with supressing the small correlation values (here we have taken the value as 0.30 as it is most commonly used). After applying this technique, the matrix will be as below.



Rotated Component Matrix[a]

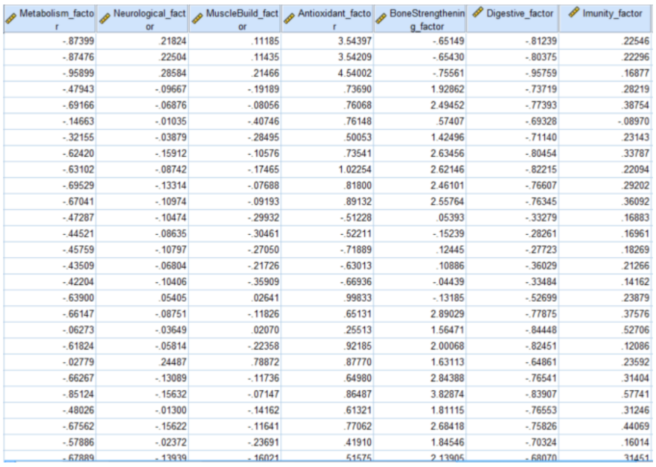
	Component						
	1	2	3	4	5	6	7
Energy_kcal			.376	.845			
Protein_g			-.500		.472		-.374
Fat_g				.933			
Carb_g			.810			.338	
Sugar_g			.817				
Fiber_g						.824	
VitA_mcg		.868					
VitB6_mg	.786						
VitB12_mcg		.780					
VitC_mg							.773
VitE_mg				.594			.300
Folate_mcg	.742						
Niacin_mg	.879						
Riboflavin_mg	.811						
Thiamin_mg	.730						
Calcium_mg					.743		
Copper_mcg		.777					
Iron_mg	.599					.390	
Magnesium_mg					.365	.728	
Manganese_mg		.511					
Phosphorus_mg					.850		
Selenium_mcg					.303		-.392
Zinc_mg	.548						

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

a. Rotation converged in 11 iterations.

Now, it looks easier to group the variable with respect to the respective components. As we can see from the above matrix, the columns ‘VitB6_mg’, ‘Folate_mcg’, ‘Niachin_mcg’, ‘Riboflavin_mg’, ‘Thiamin_mg’, ‘Iron_mg’ and ‘Zinc_mg’ are coming under component 1 as these are having better correlation values for component 1. Similary, other variables can be interpreted with the respective components.

However, the important thing to d here is giving names to these super variables with proper justifications. If we have a closer look at the variables that are belonging to componet 1 have similar effect on human body. All the columns are responsible for building cell metabolism in human body. Similarly, the variable names under component 2 are responsible for good neurological funtioning. The variables under component 3 are responsible for muscle building whereas the variables under component 4 act as antioxidants in the body. The variables under component 5 are resonsible for building strength in bones whereas the variables under component 6 are responsible for maintaining healty digestive system. Finally, the component 7 variables are responsible for building good imune system. Now the super variables can be used as features in several model analysis.



Metabolism_factor	Neurological_factor	MuscleBuild_factor	Antioxidant_factor	BoneStrengthening_factor	Digestive_factor	Imunity_factor
-.87399	.21824	.11185	3.54397	-.65149	-.81239	.22546
-.87476	.22504	.11435	3.54209	-.65430	-.80375	.22296
-.95899	.28584	.21466	4.54002	-.75561	-.95759	.16877
-.47943	-.09667	-.19189	.73690	1.92862	-.73719	.28219
-.69166	-.06876	-.08056	.76068	2.49452	-.77393	.38754
-.14663	-.01035	-.40746	.76148	.57407	-.69328	-.08970
-.32155	-.03879	-.28495	.50053	1.42496	-.71140	.23143
-.62420	-.15912	-.10576	.73541	2.63456	-.80454	.33787
-.63102	-.08742	-.17465	1.02254	2.62146	-.82215	.22094
-.69529	-.13314	-.07688	.81800	2.46101	-.76607	.29202
-.67041	-.10974	-.09193	.89132	2.55764	-.76345	.36092
-.47287	-.10474	-.29932	-.51228	.05393	-.33279	.16883
-.44521	-.08635	-.30461	-.52211	-.15239	-.28261	.16961
-.45759	-.10797	-.27050	-.71889	.12445	-.27723	.18269
-.43509	-.06804	-.21726	-.63013	.10886	-.36029	.21266
-.42204	-.10406	-.35909	-.66936	-.04439	-.33484	.14162
-.63900	.05405	.02641	.99833	-.13185	-.52699	.23879
-.66147	-.08751	-.11826	.65131	2.89029	-.77875	.37576
-.06273	-.03649	.02070	.25513	1.56471	-.84448	.52706
-.61824	-.05814	-.22358	.92185	2.00068	-.82451	.12086
-.02779	.24487	.78872	.87770	1.63113	-.64861	.23592
-.66267	-.13089	-.11736	.64980	2.84388	-.76541	.31404
-.85124	-.15632	-.07147	.86487	3.82874	-.83907	.57741
-.48026	-.01300	-.14162	.61321	1.81115	-.76553	.31246
-.67562	-.15622	-.11641	.77062	2.68418	-.75826	.44069
-.57886	-.02372	-.23691	.41910	1.84546	-.70324	.16014
-.67889	-.13939	-.16021	.51575	2.13905	-.68070	.31451

CONCLUSION:

In summary, we observed in time series analysis that the SARIMA model gave the optimum result as compared to other models. This is not surprising as the time series was stationary with a sign of seasonal effect which is suitable for SARIMA model.

On the otherhand, the logistic model analyis was able explain fairly the outcome of the response variable (person is satisfied or not with the situation in the United States). However, we saw that there were some dummy columns in the equation which were not significant.So, further improvements can be done on the model by removing the insignificant dummy columns and reapplying the logistic regression model.

In the end, principal component analysis was done on a dataset containing 23 correlated variables and then these variables were reduced to seven components with necessary measures and justification.

REFERENCE:

[1] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, "An Introduction to Statistical Learning With Applications in R", 8th ed, 2017, pp. 130-137, pp. 230-236

[2] Andy Field, “Discovering Statistics Using SPSS”, 3rd ed, SAGE Publications Inc, 2009, pp. 264-340