TERMINAL ASSIGNMENT BASED ASSESSMENT

Kirubakaran Balaraman
School of Computing
National College of Ireland
Dublin, Ireland
x19241658@student.ncirl.ie

I. Main Objective

The project focuses on choosing data from acceptable sources provided and work on a Time series model, a binary logistic regression model and also show the understanding of the Principle component analysis.

II. Time Series Analysis

A. *Objective*

The aim of this study is to analyse a time series data and to find the optimum model for forecasting the data. This study focuses on analysing the Broad Economic Categories (BEC) trade value data for European Union countries and forecast the trade value for next three years.

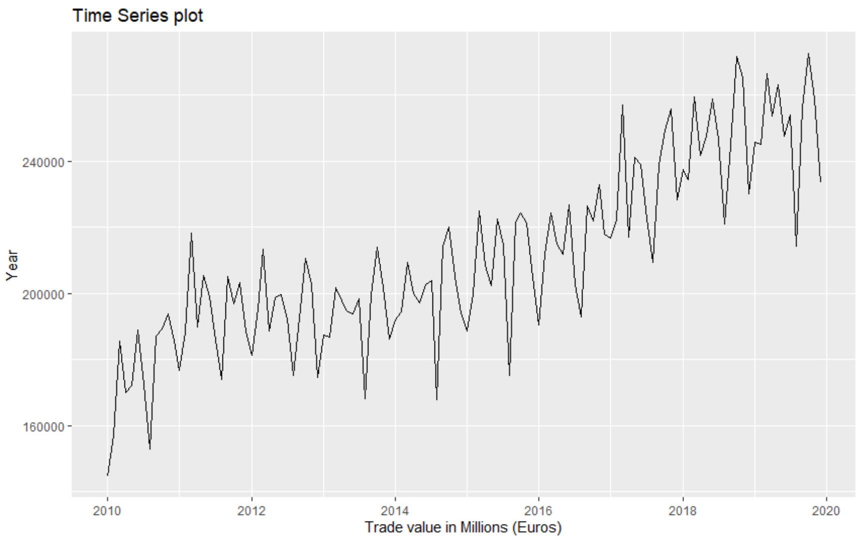
B. *Data Sourcing and Transformation*

The data chosen for this study is EU27 countries trade value data for BEC from the year 2010 to 2019. The frequency of the data is monthly consisting of 120 entries and is sourced from the Eurostat website (https://ec.europa.eu/eurostat/web/products-datasets/-/ext_st_eu27_2020bec).

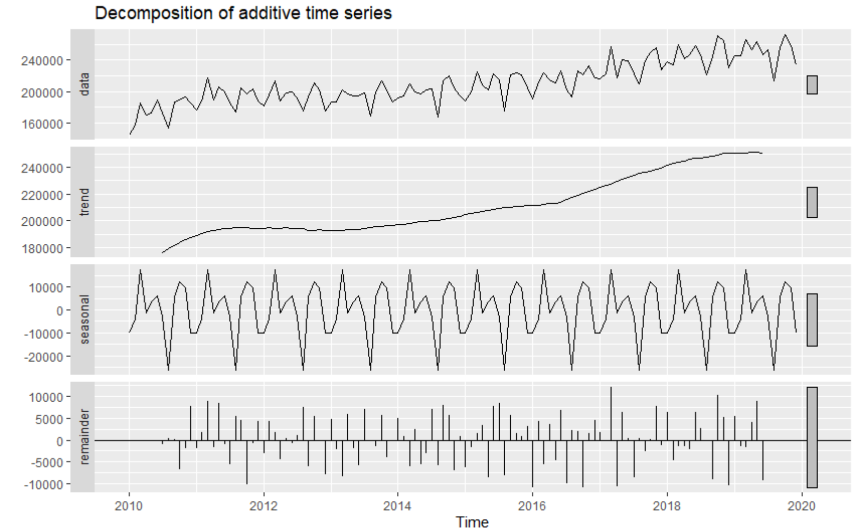
The data is then imported in R, cleaned and a time series is generated using the ts() function for further analysis to be carried out.

C. *Analysis of the time series*

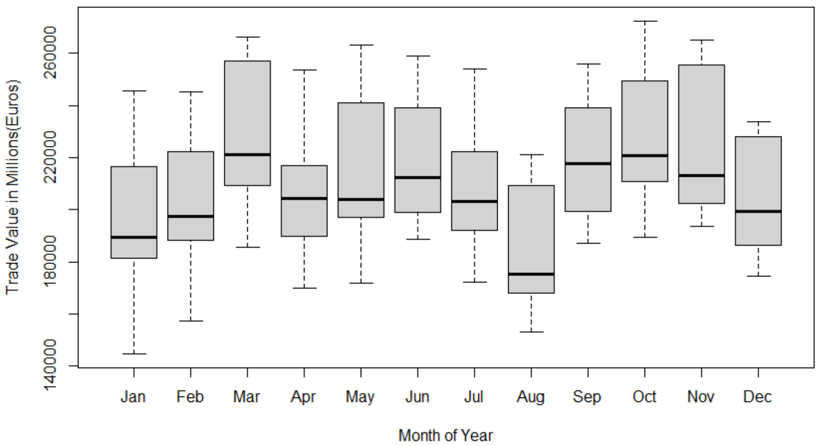
The time series is plotted in R using the plot() function to get the initial understanding about the data.



From the graph we can infer that the data follow an upward trend and also seasonlaity appears in the data. The trend, seasonality, random part of the data can be viewed clearly by doing additive decomposing. The additive decomposing is chosen as the variance of the data seems to almost same throught the period and there is no gradual increase in the variance.



From this we can clearly see that there is an upward trend and seasonlity exists such that trade value increases during the start fof the year and reaches a negative spike durin the mid of every year. The below boxplot shows a better understanding of the seasonality. There is a positive spike in the month of march and a negative spike in the month of august every year.



D. *Fitting model and forecasting*

Simple Exponential Model:

Simple Exponential model is applied to our data to check for its performance. When it is fit in our time series it didn’t perform well as our data has seasonal and trend data. Simple exponential model doesn’t have those components. We can see the RMSE value is too high with a value of 15088.33 and the AIC is 2889.

```
Forecast method: Simple exponential smoothing

Model Information:
Simple exponential smoothing

Call:
 ses(y = fd, h = 3 * 12)

  Smoothing parameters:
    alpha = 0.197

  Initial states:
    l = 169371.7751

  sigma:  15215.66

     AIC     AICc      BIC
2889.701 2889.908 2898.064

Error measures:
                   ME     RMSE      MAE      MPE     MAPE     MASE        ACF1
Training set 3390.995 15088.33 11890.02 1.161008 5.692569 1.116018 -0.01051494
```

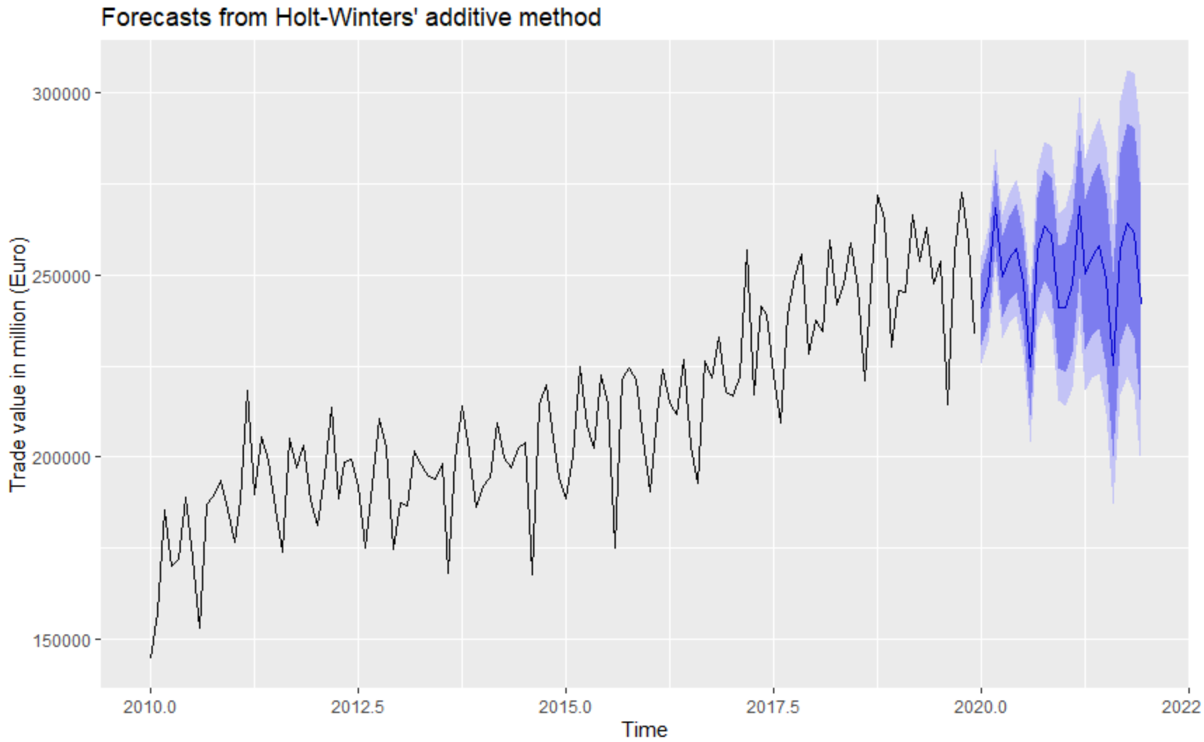
w

Holt's Winters model:

Holt's Winters model is a triple exponential model which can fit timeseries with trend, level and seasonal components. For our time series we have trend and seasonal components. We will se the R function hw() to fit the timeseries and do forecasts for 3 periods with additive method.

```
Call:
 hw(y = fd, seasonal = "additive")

  Smoothing parameters:
    alpha = 0.2527
    beta  = 0.027
    gamma = 1e-04

  Initial states:
    l = 174889.7231
    b = 1580.4408
    s = -10083.71 9644.297 12327.55 5557.296 -26254.35 -2975.378
           6374.429 3499.3 -1383.87 17392.21 -3790.291 -10307.48

  sigma:  7620.193

     AIC     AICc      BIC
2736.581 2742.581 2783.968

Error measures:
                    ME     RMSE      MAE        MPE    MAPE      MASE       ACF1
Training set -473.3179 7094.013 5730.743 -0.3284165 2.77075 0.5378977 -0.1932027
```

The model is fit to the time series and it has an AIC value of 2736.591 with RMSE of 7094.013. We will compare these with other models to find the optimum one. The below graph plots the forecasts generated from the Holt's Winters model.



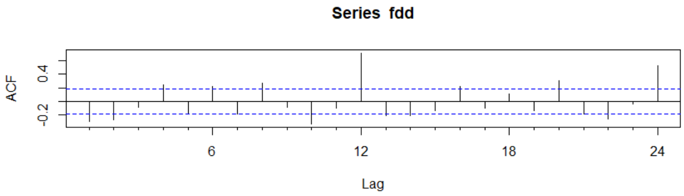
Seasonal ARIMA model:

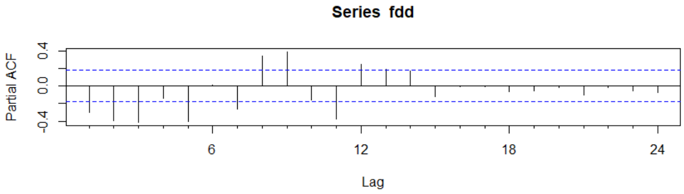
Seasonal Auto-Regressive Integrated Moving Average (SARIMA) is a form of ARIMA which supports seaonsal data in a time series. Where the ARIMA model have three orders p (autoregression), d (difference) and q (moving average)regression for the trend, the SARIMA model has those included along with P, Q, D for seasonal amd m (timesteps in single seasonal period).

Before applying the model, the stationarity of the time series should be cheked. As our time series has seasonal and trend components it is not stationary and we test this using Dickey-Fuller test. We need a p value < 0.05 to reject the null hypothesis which states the series is not stationary. We do 1 order differencing to our data and run Dickey- Fuller test in R using acf.test().

```
> ndiffs(fd)
[1] 1
> adf.test(fdd)#pass

        Augmented Dickey-Fuller Test

data:  fdd
Dickey-Fuller = -12.554, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

We can see that the d value is 1 for the ARIMA model. We can find p and q values from ACF and PACF plots.



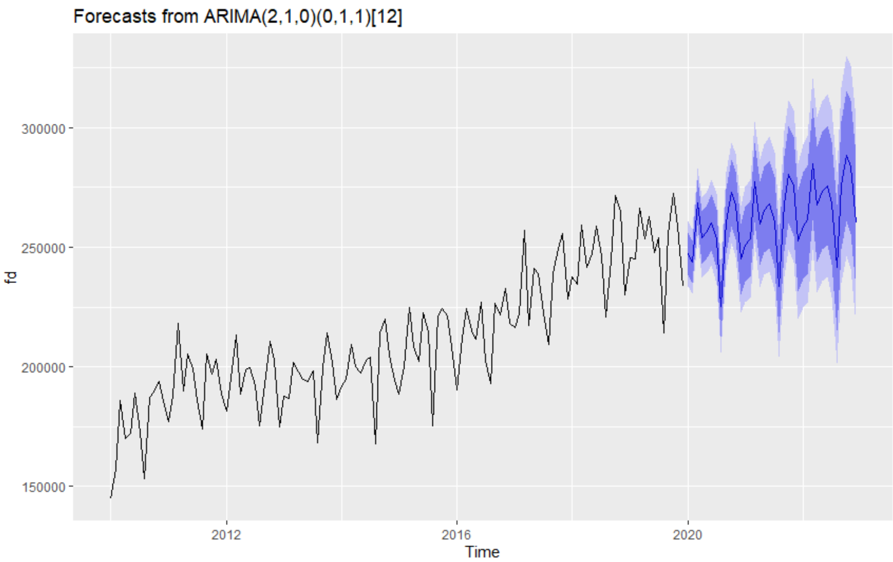


Finally the SARIMA model is run in R using Arima() function with (p,q,d) = (2,1,0) and (P,Q,D)M = (0,1,1)12

```
ARIMA(2,1,0)(0,1,1)[12]

Coefficients:
          ar1      ar2     sma1
      -0.9638  -0.6095  -0.7375
s.e.   0.0761   0.0767   0.1249

sigma^2 estimated as 48216780:  log likelihood=-1102.18
AIC=2212.35   AICc=2212.74   BIC=2223.04

Training set error measures:
                    ME     RMSE      MAE        MPE     MAPE      MASE        ACF1
Training set -987.2801 6464.354 4921.297 -0.5772426 2.305089 0.4619216 -0.0592222
>
```

Also, Ljung test came up with a p-value>0.05 suggesting zero autocorrelations and the model fits the data well.

```
        Box-Ljung test

data:  ar.fd$residuals
X-squared = 0.43148, df = 1, p-value = 0.5113
```



The SARIMA model performed well with a lower AIC value of 2212.35 and RMSE of 6464.354 than Holt's Winter model.

E. Conclusion

The forecast for the time series for 3 years has been generated and plotted with a AIC value of 2212.35. Seasonal ARIMA model performed better than the other models capturing the trend and the seasonality in the series with a RMSE of 6464.354.

III. BINARY LOGISTIC REGRESSION

A. Objective

The attitude of a person towards their country is influenced by different factors. The overall satisfaction may change for each person based on their opinions on problems and merits that exists in their country. This study foccuses on predicting the country satisfaction of an individual based on the current economic situation, whether an individual is satisfied with the way the democracy is working, the acceptance of homosexuality by the society and their age.

B. Data Source

The dataset is sourced from the spring 2019 survey conducted by Pew Research Centre on Global attitude and trends of the people.It is available in the pew research centre website https://www.pewresearch.org/global/dataset/spring-2019-survey-data/. It involves a questionnaire with lot of questions about the people attitude towards several things. For the purpose of the project, the following questions are selected to understand about the attitude towards their country.

Variable	Question and response	Type	Dependent/Independent
country_satis	Overall, are you satisfied or dissatisfied with the way things are going in our country today? Options – 1: Satisfied, 2: Dissatisfied	Dichotomous	Dependent
econ_sit	How would you describe the current economic situation? Options – 1: Very Good, 2: Somewhat Good, 3: Somewhat Bad, 4: Very Bad	Categorical	Independent
satisfied_democracy	How satisfied are you with the way democracy is working in our country? Options – 1: Very satisfied, 2: Somewhat satisfied,3: Not too satisfied,4: Not at all satisfied	Categorical	Independent
homosexuality	Do you think homosexuality is accepeted by the society in your country? Options – 1: Yes, 2: No, 3: Don't know, 4: Refused	Categorical	Independent
age	How old were you at your last birthday? Options – Fill Age or Fill 97: 97 or older, 98: Don't know, 99: Refused	Continuous	Independent

C. Data cleaning and Transformation

- The data was downloaded as a .csv file from the source and imported in R to further clean and transform the data to be suitable for the analysis.
- The raw data contained more than 100 columns and the needed columns were chosen and stored in a dataframe.
- The columns are then checked for null values, cleaned and exported into a .csv file to be used in SPSS for further analysis.

D. *Checking for assumptions*

i. **Sample size :** The sample used in our project consists of 996 records which satisfies the sample size assumption.

ii. **Dichotomous dependent variable**: The dependent variable country_satis is dichotomous and so takes only two values.

iii. **Absence of multicollinearity:** There are no multicollinearities between the independent variables use in our study. The pearsons coefficients are not too high to prove the existence of multicollinearity.

Correlations

		COUNTRY_SATIS	ECON_SIT	SATISFIED_DEMOCRACY	HOMOSEXUALITY	AGE
COUNTRY_SATIS	Pearson Correlation	1	.360**	.402**	-.047	.134**
	Sig. (2-tailed)		.000	.000	.136	.000
	N	996	996	996	996	996
ECON_SIT	Pearson Correlation	.360**	1	.390**	-.002	.013
	Sig. (2-tailed)	.000		.000	.943	.686
	N	996	996	996	996	996
SATISFIED_DEMOCRACY	Pearson Correlation	.402**	.390**	1	-.092**	.122**
	Sig. (2-tailed)	.000	.000		.004	.000
	N	996	996	996	996	996
HOMOSEXUALITY	Pearson Correlation	-.047	-.002	-.092**	1	.135**
	Sig. (2-tailed)	.136	.943	.004		.000
	N	996	996	996	996	996
AGE	Pearson Correlation	.134**	.013	.122**	.135**	1
	Sig. (2-tailed)	.000	.686	.000	.000	
	N	996	996	996	996	996

iv. **Dependent variable outcome:** There are no multicollinearities between the independent variables use in our study. The pearsons coefficients are not too high to prove the existence of multicollinearity.

E. *Building the model*

The binary logistic regression is carried out in SPSS and the outputs are analyzed and evaluated to form the logistic regression equation. The baseline model with no independent varaible showed a classification accuracy of 79.3%.

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted COUNTRY_SATIS Not Satisfied	Satisfied	Percentage Correct
Step 0	COUNTRY_SATIS	Not Satisfied	790	0	100.0
		Satisfied	206	0	.0
	Overall Percentage				79.3

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	-1.344	.078	295.213	1	.000	.261

We would build the model and try to improve the classification accuracy and do approriate tests to validate our findings.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	250.651	10	.000
	Block	250.651	10	.000
	Model	250.651	10	.000

All the predicted variables are included in the prediction of the model in Block 1 and the Omnibus test is conducted to check if the accuracy of the model increases with the inclusion of the predictor vatiables. From the above figure we can see that the p value < 0.05 which suggests that the model performance is improved.

We can check the fit of the model to our data by checking the signicance value in Hosmer Lemeshow test. As p>0.05, we can assume that there is a good fit to the data.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.487	8	.593

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	764.717[a]	.222	.348

From the model summary, we can see the Pseudo R square value, Nagelkerke R square, which shows that 34.8 % of the variance in the dependent variable is explained by the model.

From the below classification table, it can be seen that the accuracy of the model improved from 79.3% of baseline model to 82.8%.

Classification Table[a]

			Predicted		
			COUNTRY_SATIS		Percentage Correct
	Observed		Not Satisfied	Satisfied	
Step 1	COUNTRY_SATIS	Not Satisfied	749	41	94.8
		Satisfied	130	76	36.9
	Overall Percentage				82.8

a. The cut value is .500

We will look into the variables in the equation table to check the coefficients for the predictor variables and the odds ratio of each of the varaibles.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	ECON_SIT			65.525	3	.000	
	ECON_SIT(1)	2.932	.575	25.968	1	.000	18.758
	ECON_SIT(2)	2.122	.482	19.343	1	.000	8.346
	ECON_SIT(3)	.678	.504	1.812	1	.178	1.970
	SATISFIED_DEMOCRACY			66.393	3	.000	
	SATISFIED_DEMOCRACY(1)	2.512	.411	37.429	1	.000	12.334
	SATISFIED_DEMOCRACY(2)	1.752	.263	44.419	1	.000	5.763
	SATISFIED_DEMOCRACY(3)	.766	.273	7.867	1	.005	2.150
	HOMOSEXUALITY			8.500	3	.037	
	HOMOSEXUALITY(1)	-2.516	.880	8.169	1	.004	.081
	HOMOSEXUALITY(2)	-2.599	.928	7.846	1	.005	.074
	HOMOSEXUALITY(3)	-2.756	1.056	6.810	1	.009	.064
	AGE	-.017	.005	12.630	1	.000	.983
	Constant	-.539	1.047	.265	1	.607	.584

a. Variable(s) entered on step 1: ECON_SIT, SATISFIED_DEMOCRACY, HOMOSEXUALITY, AGE.

The table suggests that the ECON_SIT(3) doesn't contribute to the prediction as it is not significant enough. The coefficients of remaining predictors will help us build the logistic regression equation.

F. Results and interpretation

The variables in the equation table contains Exp(B) values for each predictors. The result shows that if a person is very satisfied with the economic situation of theie country, then the odds that they answer as satisfied to the overall satisfaction is 18.758 times greater than a person who responds not satisfied. This means that the economic situation plays a major role in overall satisfaction.

Also, person having good opinion about their country's democracy has 12.334 time more odds to respond as satisfied, than those who don't have a good opinion about their democracy.

The odds ratio is interpreted for all the predictors in such a way. The logistic regression equation can be formed with the help of the coefficients as follows:

$$E(y)=e^{-0.539+2.9(x1)+2.1(x2)+2.5(x3)+1.75(x4)+0.76(x5)-2.5(x6)-2.6(x7)-2.7(x8)-0.017(x9)} / (1+ e^{-0.539+2.9(x1)+2.1(x2)+2.5(x3)+1.75(x4)+0.76(x5)-2.5(x6)-2.6(x7)-2.7(x8)-0.017(x9)})$$

G. *Conclusion*

The logistic regression model is built with four predictor variables which helps us predict whether a person is satisfied with their own country with a accuracy of 82.8 %. The odds of a person satisfied with their own country is majorly dependent upon their satisfaction with the economic situation and the way the democracy works.

IV. Principle Component Analysis

A. *Introduction*

Statistical analysis bacame a major part of today's world. With lot of data available around the world, analysing and exploring them will help us uncover many patterns and insights that helps us in many ways. When doing such an analysis often time we are left with huge data with lot of atttributes to take into account for predicting the dependent variable. In such cases where there are hundreds of attriubutes, we cannot use all of them and we will be needing a technique to reduce the number of attributes without the cost of losing valuable information. This concept is called the Dimensionality reduction. In this study, we will look into a dimensionality reduction technique called Principle Component Analysis.

B. *Principle Component analysis*

Principle Component Analysis (PCA) is a technique to extract few components from a high dimensional dataset without losing much data. The columns of a dataset are referred to as dimensions or attributes. For forming the components weights are chosen such that it maximizes the variance explained and also the components are uncorrelated with each other.

C. *Data Source and Description*

The understanding of the PCA is experimented with the UCI heart disease dataset sourced from Kaggle site which can be accessed through https://www.kaggle.com/ronitf/heart-disease-uci. The dataset is a small dataset consisting of 14 columns and 303 rows. For the purpose of this study we are chossing only the continuous variables from those columns which includes:

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	age	Numeric	2	0	Age	None	None	8	Right	Scale	Input
2	trestbps	Numeric	3	0	Resting blood pressure	None	None	8	Right	Scale	Input
3	chol	Numeric	3	0	Serum cholestrol	None	None	8	Right	Scale	Input
4	thalach	Numeric	3	0	Maximum achieved Heart Rate	None	None	8	Right	Scale	Input
5	oldpeak	Numeric	3	1	ST depression induced by exercise	None	None	8	Right	Scale	Input

D. *Extracting the Principle Components*

The data is imported in SPSS and PCA is carried and components are extracted with eigen values more than 1. We can tell that the data is suitable for factor analyis with the help of following tests.

Factor Analysis

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.557
Bartlett's Test of Sphericity	Approx. Chi-Square	144.172
	df	10
	Sig.	.000

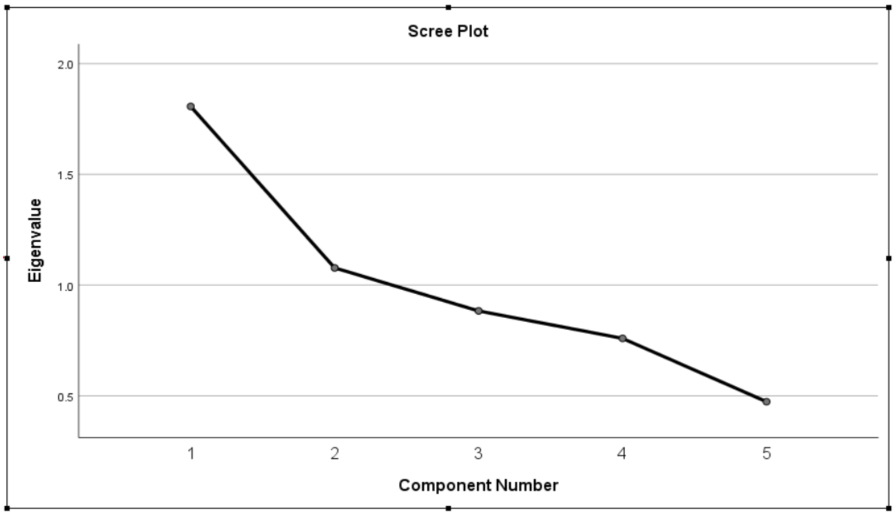
The Barlett's test of Sphericity has a significance value < 0.05 which suggests that we can reject the null hypothesis stating there is no corrleation struture. So we can say that there is correlation to do factor analysis. KMO test with a value > 0.5 shows that more than 50 % correlation between variables can be explained by other variable.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.807	36.132	36.132	1.807	36.132	36.132	1.584	31.690	31.690
2	1.078	21.551	57.682	1.078	21.551	57.682	1.300	25.993	57.682
3	.883	17.668	75.350						
4	.759	15.183	90.533						
5	.473	9.467	100.000						

Extraction Method: Principal Component Analysis.

The table above shows that there are total 5 components that were extracted in which the top two has the eigen values more than 1. These two components are called the priniciple components and will help to explain 57.68% of the variance in the data. We can view this better with a scree plot to check if the components are above the 1 eigen value threshold.



We used the default varimax rotation method, which results in uncorrleated components. Thus we can use the rotated component matrix below to decide on the variables and form groups.

Rotated Component Matrix[a]

	Component	
	1	2
age	.574	.515
trestbps		.645
chol		.781
thalach	-.843	
oldpeak	.705	

From the matrix, we can interpret that the age, thalach and oldpeak are highly correlated, whereas trestbps and chol are correlated in the other component. The component 1 includes the age, maximum heart rate achieved and depression and the component 2 has the resting blood pressure and cholestrol.

E. Conclusion

In our study of the heart disease dataset, we were able to reduce the dimension from 5 to 2 principle components which explained 57.68% variance. This study only focussed on a small dataset to show the understanding of the concepts. This can be implemented in large scale when working on complex datasets with enormous number of attributes to find priniciple components wich can be used for the analysis.