National College of Ireland

Project Submission Sheet – 2020/2021

Student Name: Omkar Ratnoji Tawade

Student ID: 19232136

Programme: MSCDAD_A **Year:** 2020-21

Module: Statistics for Data Analytics

Lecturer: Tony Delaney

Submission Due Date: 05-01-2021

Project Title: Terminal Assignment (TABA)

Word Count: 5333 words

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Omkar Ratnoji Tawade

Date: 04-01-2021

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

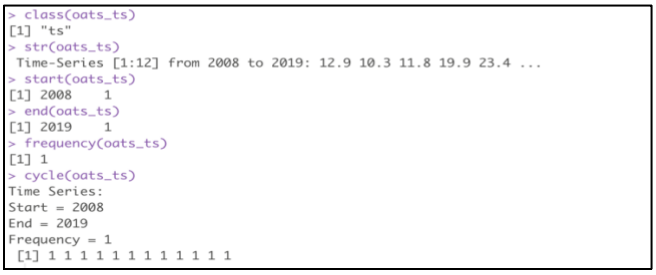
Statistics for Data Analytics

Terminal Assignment

Omkar Ratnoji Tawade
Data Analytics
National College of Ireland
Dublin, Ireland
x19232136@student.ncirl.ie

I. Time Series

For time series analysis, I have downloaded the Price of Oats per kg in the UK. The data was recorded for each year from 2008 to 2019. Now we will first explore the data. So, I have first imported the data file and converted it into the data frame. After importing the data, I have applied the time series function in which I passed the price column of the oats data frame and also initialized the timeline for the data which is from the year 2008 to 2019. Also, we will need to initialize the frequency equals to 1 here because the data is stored year wise. After time series is generated, we can also verify it by using the class function and you will get time series as the output. Start and End function will help to verify the timeline of the time series. Since our data is recorded year-wise so the output of the cycle function is equal to 1. We have now completed the structural overview of the data.



```
> class(oats_ts)
[1] "ts"
> str(oats_ts)
 Time-Series [1:12] from 2008 to 2019: 12.9 10.3 11.8 19.9 23.4 ...
> start(oats_ts)
[1] 2008    1
> end(oats_ts)
[1] 2019    1
> frequency(oats_ts)
[1] 1
> cycle(oats_ts)
Time Series:
Start = 2008
End = 2019
Frequency = 1
 [1] 1 1 1 1 1 1 1 1 1 1 1 1
```

Figure 1. Data exploration.

Now we will explore the data by using visuals. It will help to know about the variance in the data.

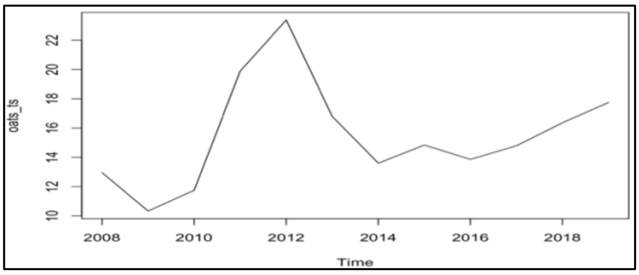


Figure 2. A plot of the time series.

By observing Figure 2, we can say that the general trend is an upward trend but there is a large fluctuation of the price of oats in 2012. Now we will observe the plot using abline. abline is a line that passes through the mean point of the price of each year. I have passed the regression model inside the abline function. In time-series data there is no independent variable. The price of oats is our dependent variable, and the time of this data set is passed as an independent variable.

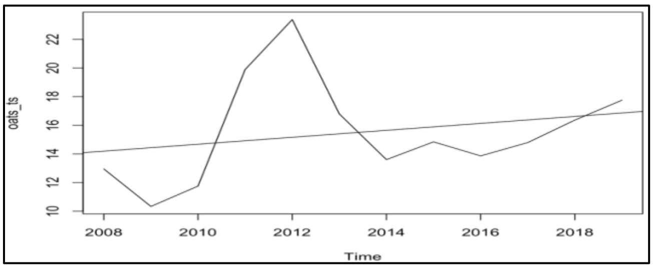


Figure 3. Plot of time series with abline.

By referring the Figure 3, we can say that there is more variance in the given time series data. Based on Figure 3, we can define variance as the difference between abline and the sparkline. Figure 3 suggests that variance is not standard or stationary. Since variance is not standard, we cannot apply the time series. Also, by referring the Figure 3 we can say that the mean is fluctuating, and it is not stationary. So, we will need to make a variance and mean stationary in order to apply the time series.

To remove the variance or to minimize the variance we can apply log transformation on the price of oats. We can now see the plot of the time series after applying log transformation.

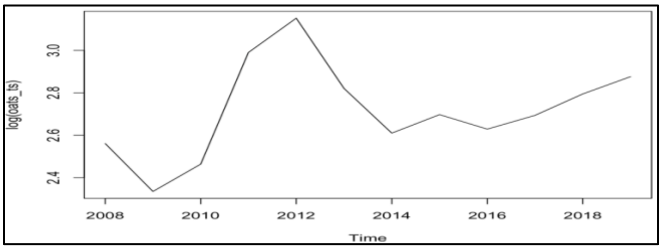


Figure 4. Log transformation to homogenize the variance.

By comparing Figure 3 and Figure 4, we can see a slight change in the data. Since our dataset is small so it will be difficult to see if we have homogenized the variance or not. Now we will homogenize the mean to make the data stationary.

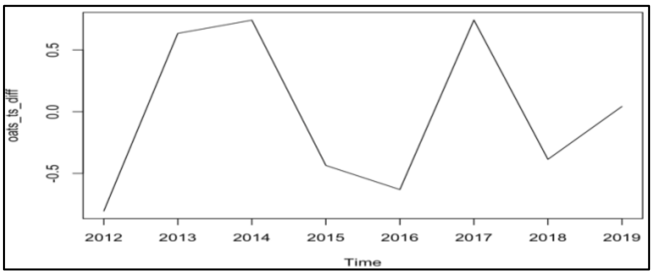


Figure 5. Differentiation to homogenize the mean.

After applying log transformation and differentiation on the data then the abline gets plotted at the center of the graph which suggests that data becomes stationary and now we can apply time series analysis on the data. We have completed the preprocessing of the data.

A. ARIMA Model:

We will apply the ARIMA model to our dataset. There are three parts of the ARIMA model. AR which stands for autoregressive, MA stands for moving average and I stand for Integration. So, the ARIMA model is defined as the Integration of autoregression of moving average. We can say there are three sub-models in ARIMA. By running the autoregressive model on the data, we get a p value. q value will be generated by running the moving average model and the d value is derived from the integration. We decide the p and q value by analyzing the plot of PACF and ACF respectively.

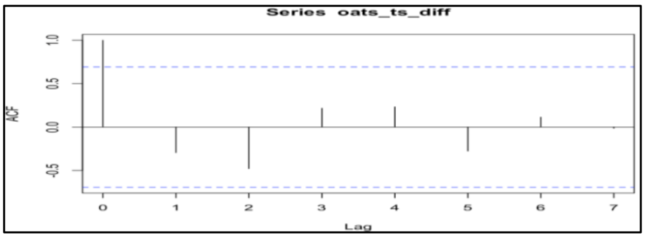


Figure 6. ACF plot

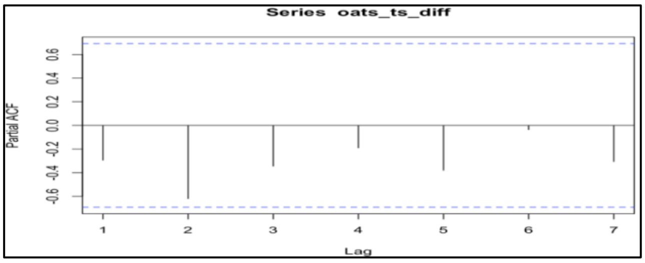


Figure 7. PACF plot

It is very important to find the perfect order which will be passed in the Arima model. Order is decided by autocorrelation, partial autocorrelation, and the number of differentiations performed on the time series. Autocorrelation of lag1 is the correlation between the time series and same time series offset by one step. ACF plots the graph between autocorrelation values and lags. If these values are small and lie between the blue dotted line as shown in Figure 6 then those values are not statistically significant. Based on Figure 6, we can say that the zeroth line is the only statistically significant value in the ACF plot. Partial autocorrelation is the correlation between the time series and lag version of itself after we subtract the correlation from the smaller lags. So, it is a correlation associated with just that particular lag. By comparing ACF and PACF for time series, we can deduce the model order. If the amplitude of ACF tails of at increasing lag and PACF cuts off after some lag p then we have the AR model. By observing the ACF and PACF plots of our study, we can say that it is an AR(2) model. The order of the ARIMA model will be (2,4,0) because we have applied four numbers of differentiation on the time series and based on ACF and PACF plots we concluded it is an AR(2) model. Now we can pass the time series in the ARIMA model. We will now integrate the output of the model.

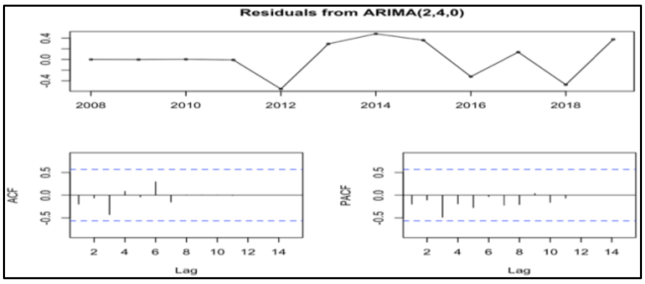
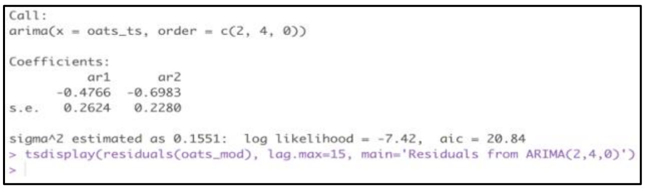


Figure 8. Model Residuals Plot.

Based on Figure 8, we can say that all lags are under the significance level which suggests that the order used in modeling is good. There are three main steps involved in the ARIMA model. The first step is model identification based on ACF and PACF. In our study, it is the AR model. The second step is to estimate parameters based on residuals. So, we get the coefficients of the AR model in our study.



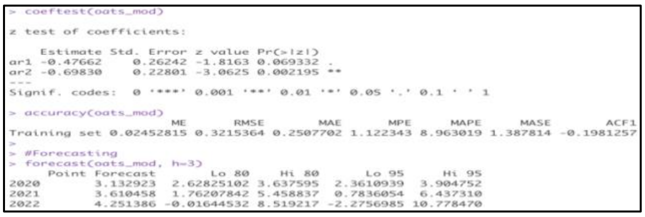
```
Call:
arima(x = oats_ts, order = c(2, 4, 0))

Coefficients:
          ar1      ar2
      -0.4766  -0.6983
s.e.   0.2624   0.2280

sigma^2 estimated as 0.1551:  log likelihood = -7.42,  aic = 20.84
> tsdisplay(residuals(oats_mod), lag.max=15, main='Residuals from ARIMA(2,4,0)')
> 
```

Figure 9. Model Coefficients.

The third step is diagnostic checking using the Ljung-Box test. The null hypothesis of the Ljung-Box test is that residuals are independent and identically distributed. In our study, the p-value of the Ljung Box test is 0.6741 which suggests that we will need to accept the null hypothesis and we can conclude that model is a good fit.



```
> coeftest(oats_mod)

z test of coefficients:

    Estimate Std. Error z value Pr(>|z|)
ar1 -0.47662    0.26242 -1.8163 0.069332 .
ar2 -0.69830    0.22801 -3.0625 0.002195 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> accuracy(oats_mod)
                     ME      RMSE       MAE      MPE     MAPE     MASE       ACF1
Training set 0.02452815 0.3215364 0.2507702 1.122343 8.963019 1.387814 -0.1981257
>
> #Forecasting
> forecast(oats_mod, h=3)
     Point Forecast       Lo 80    Hi 80      Lo 95     Hi 95
2020       3.132923  2.62825102 3.637595  2.3610939  3.904752
2021       3.610458  1.76207842 5.458837  0.7836054  6.437310
2022       4.251386 -0.01644532 8.519217 -2.2756985 10.778470
```

Figure 10. Model accuracy and Forecasting.

Based on Figure 10, we can say that there is only a coefficient significant in our model. Also, the RMSE of the model is about 32.15%. By observing these values, we can conclude that model is a good fit and used for forecasting. Figure 10 shows the forecasted values and also, we can observe the same in the graph.

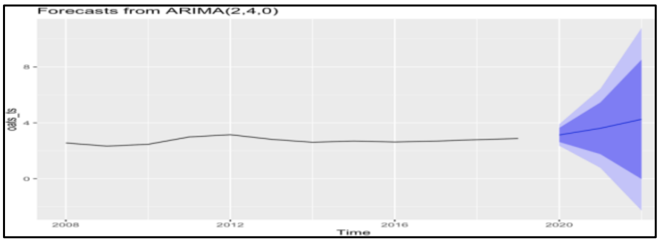


Figure 11. Forecasting based on the ARIMA model.

We have successfully forecasted the next period of the time series using the ARIMA model.

B. *Simple Exponential Smoothing:*

The simple exponential smoothing model does not have any trend or seasonality exist. The forecast equation of simple smoothing equation that says forecast equals the level of the time series. The level of the time series itself is the weighted average of level in the previous period and the value of time series at time t. I am assuming this model can be the best model among the three models because simple exponential smoothing is more suitable for data with no trend or seasonal pattern. We pass two parameters in the simple exponential smoothing function which are α and h. α ranges between 0 and 1. α is the weight given to each observation. In exponential smoothing recent observations are weighted more heavily than the past observations. Ideally, α=0.2 is passed in a simple exponential smoothing function. If we decrease the value of α then the model will weight historical data, more and if we increase the value of α then the model will weight recent observations more. So, it is very important to select the right α value. We will pass the sequence of α in a simple exponential smoothing function and then we will plot the value of α and RMSE to select the correct value of α.

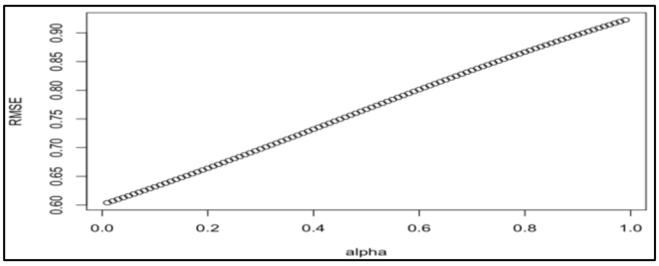


Figure 12. Plot of alpha against RMSE.

Based on Figure 12, we can say that less is the value of alpha, less is the RMSE value. But we cannot take the value of alpha equal to zero because it will weigh past data more. We will continue with alpha equal to 0.2. Now we will interpret the results of our model. We will first analyze the forecast plot.

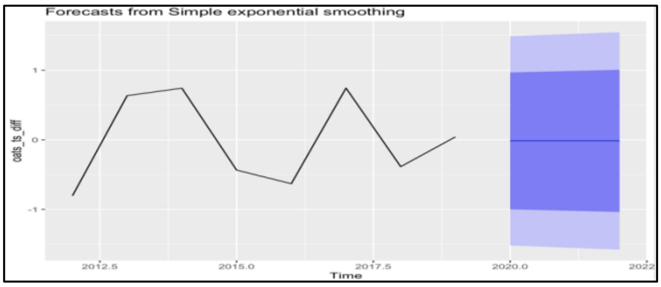
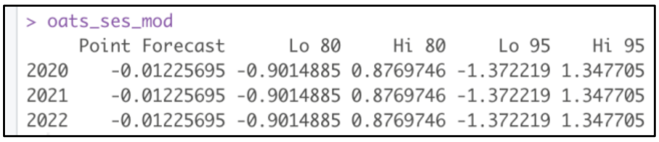


Figure 13. Forecast plot of simple exponential smoothing model.

We can see the interval of the forecasted value of the price of oats of 3 periods ahead and these intervals are not deviating much so we can say that our model is performing well. Also, the RMSE value of the model was 66.41% which suggests the model is a good fit. Figure 14 shows the forecasted value of the simple exponential smoothing model.

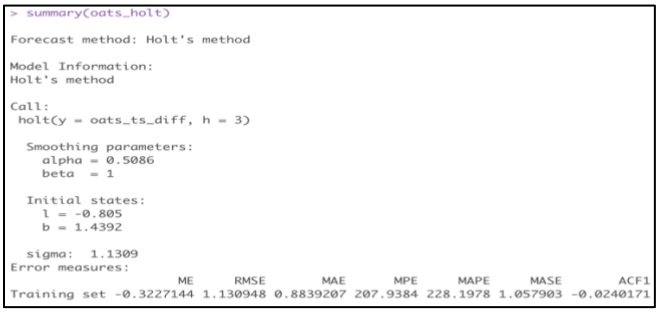


```
> oats_ses_mod
     Point Forecast       Lo 80     Hi 80     Lo 95    Hi 95
2020    -0.01225695 -0.9014885 0.8769746 -1.372219 1.347705
2021    -0.01225695 -0.9014885 0.8769746 -1.372219 1.347705
2022    -0.01225695 -0.9014885 0.8769746 -1.372219 1.347705
```

Figure 14. Forecasted value of simple exponential smoothing model.

C. *Holt's Model:*

Holt's method also belongs to the exponential smoothing group. It is used to handle the data with a linear trend. It not only smooths the trend and the slope by using smoothing constants but also provides more flexibility in selecting the rates at which trends and slopes are tracked. So, there is three-part of holt's model, the first part is the smoothing parameter in which we use alpha to weight the past and recent observation. The second part of the forecasting equation of holt's model is the smoothing of the trend in which we will be using a new parameter β. It also ranges from 0 to 1. We will use our processed time series in the holt's function and in this case, we are not going to pass the alpha and beta value so this value will be decided by the function itself. We will analyze the value of alpha and beta used by the function.



```
> summary(oats_holt)

Forecast method: Holt's method

Model Information:
Holt's method

Call:
 holt(y = oats_ts_diff, h = 3)

  Smoothing parameters:
    alpha = 0.5086
    beta  = 1

  Initial states:
    l = -0.805
    b = 1.4392

  sigma:  1.1309
Error measures:
                     ME     RMSE       MAE      MPE     MAPE     MASE       ACF1
Training set -0.3227144 1.130948 0.8839207 207.9384 228.1978 1.057903 -0.0240171
```

Figure 15. Summary of Holts Model.

Based on Figure 15, we can say that holt's function used alpha=0.5086 and beta=1 in the model. The trend smoothing parameter is very high. We will now observe the forecasting plot of holt's method.

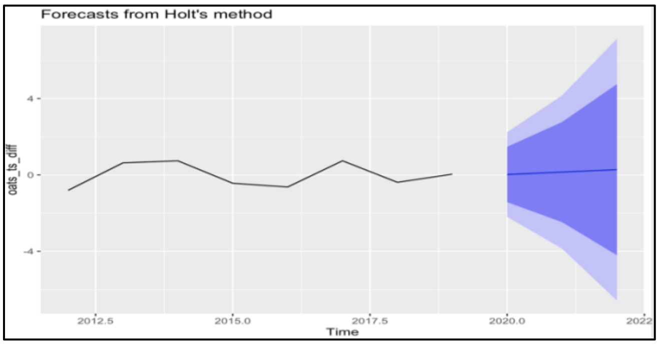
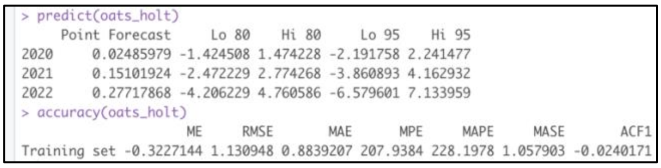


Figure 16. Plot of holt linear trend model.

By observing Figure 16, we can say that model is working well for forecasting the value. I also tried to generate models. using low magnitude alpha and beta values but the forecasting was not accurate and was deviating from the past observations. too much. We were able to predict the three periods ahead of the given time series.



```
> predict(oats_holt)
     Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
2020     0.02485979 -1.424508 1.474228 -2.191758 2.241477
2021     0.15101924 -2.472229 2.774268 -3.860893 4.162932
2022     0.27717868 -4.206229 4.760586 -6.579601 7.133959
> accuracy(oats_holt)
                     ME     RMSE       MAE      MPE     MAPE     MASE       ACF1
Training set -0.3227144 1.130948 0.8839207 207.9384 228.1978 1.057903 -0.0240171
```

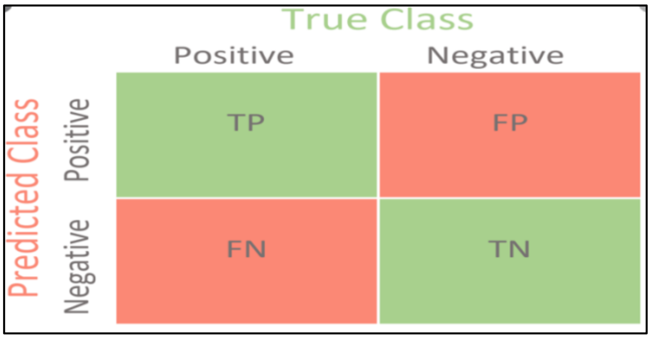
Figure 17. Forecasted values using the Holts model.

II. Logistic regression

Whenever the outcome of a variable is categorical or is discrete then we can use the logistic regression. Though logistic regression is working as a classifier but under the hood logistic regression uses the linear models. In logistic regression, we use the logistic function or the sigmoid function for the classification. We can pass a linear function into this sigmoid function. We can express the logistic regression equation as,

$$F(x) = \frac{1}{1 + e^{-(\beta o + B1x)}}$$

The above equation $\beta o + B1x$ is the linear regression equation. This means we can take our linear regression solution and place it into the sigmoid function. The sigmoid function takes in any value and outputs it to between 0 and 1. The graph of this equation is an S-shaped curve. Logistic regression is used for binary classification and also for multiclass classification. The convention of binary classification is to have two classes 0 and 1. After training data using the logistic function, we can evaluate model performance by analyzing the confusion matrix.



		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Figure 18. Confusion Matrix.

The confusion matrix is a cross table between the predicted condition or what you predicted the label to be versus the true condition or what the true label actually was. If the condition was positive and our model predicted it to be positive, then it comes under the true positive class. If the condition was positive and our model predicted it to be negative, then it is called false positive or also called a Type 1 error. If the condition was negative and our model predicted it to be negative, then it is called true negative and if our model predicted the negative condition as a positive condition then it is called false negative or also called Type 2 error. Based on this matrix, various parameters of the models such as accuracy, positive likely hood ratio, negative likely hood ratio, etc. are calculated. The importance of the confusion matrix and other various calculated metrics is that they are fundamentally comparing the predicted values versus the true values.

A. *Survey Description:*

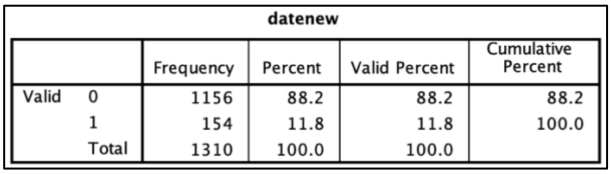
I have downloaded September 2005 – Online Dating survey data. It was a tracking survey. The survey covered the question about online dating and Hurricane Katrina donations. We will be focusing on the Online dating part. The survey was in the field from September to December 2005. Survey data was recorded with the help of a telephonic interview. According to the data file, 3215 respondents recorded their responses to this survey. The confidence level of the result based on this study was 95%. This suggests the amount of sampling error rate. Also, weights were included in the data file. Demographic weights were derived from the analysis of the Census Bureau's Annual Social and Economic Supplement (March 2004). So, the majority of the surveys are divided into three parts which are screener questions, main questionnaires, and demographics questions. The first three questions of the survey were related to the sentiment. After these questions, the next 15 questions were about the internet. There were about 17 questions asked about dating life.

B. *Research Question and Hypotheses:*

After observing all questions, I had decided on my research question. My research question is what factors influence individuals for online dating or the reason behind why induvial get involved in online dating? After deciding on the research question, I had made a few hypotheses based on survey questions. So, my first hypothesis is that if an individual is believing in trusting others then it is more likely he will use the online dating site. Q2 of the survey ask about trust. The second hypothesis is that if an individual is using the internet more and if he is performing a good amount of task on the internet then that individuals are likely to be using the online dating app. This hypothesis is derived from WEB-B questions. The third hypothesis is that if an individual thinks that whatever he finds on the web is reliable then that individual is likely to use the dating site. This hypothesis is derived from the Q17 of the survey. The next three hypotheses are based on demographic questions. The fourth hypothesis is that men use dating websites more than women comparatively. The fifth hypothesis and sixth hypothesis are that individuals who are postgraduate or the individual whose range around 20 to 30 are more likely to use online dating.

C. *Data preparation:*

Now we will explore our data. For logistic regression, our dependent variable should be categorical. In our case, Datenew which is our dependent variable has 3 labels which are yes, no, and refused/don't know. For better analysis, we are converting the Datenew into a dichotomous variable. We will execute this by filtering out the dataset by the Datenew variable.



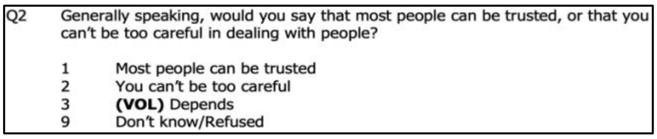
datenew

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	1156	88.2	88.2	88.2
	1	154	11.8	11.8	100.0
	Total	1310	100.0	100.0	

Figure 19. Frequency of dependent variable.

Figure 19 shows the frequency table of our dependent variable. Since we filtered our dataset based on our dependent variable, so we can see only two values in our dependent variables. Also, we have recoded our Datenew variable. Before recoding recodes value of Yes was 1 and for No, it was 2. But after recoding, I changed the recode value of yes to 1 and No to 2. I have converted this variable into a dichotomous variable. We had to follow this step because when we execute binary logistic regression in SPSS it considers two classes of output 0 and 1. If we had not recoded the Datenew variable then it would have considered 0 for Yes and 1 for No and the whole meaning of analysis would have changed or gone wrong. So, it very important to check the recodes of the dependent variable before applying logistic regression.

Now we will observe our independent variable. So, starting with Q2, which has asked the respondent whether he/she can trust others easily or not.

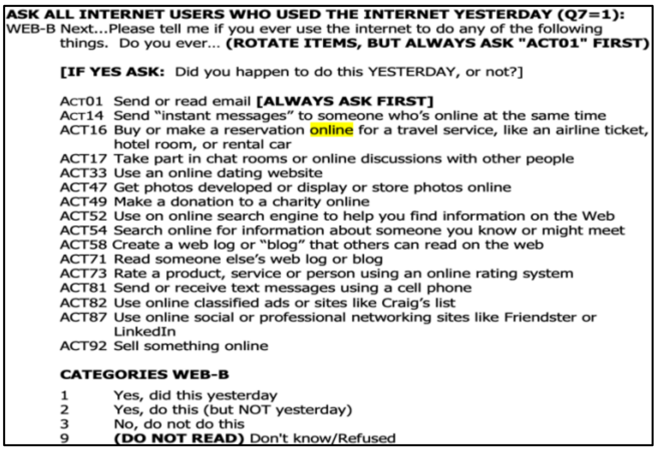


Q2 Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?

1 Most people can be trusted
2 You can't be too careful
3 **(VOL)** Depends
9 Don't know/Refused

Figure 20. Q2 of the survey.

To make this variable more helpful for the analysis, I have computed a new variable named Trust and I have coded it as 1 when the response in Q2 was 1 and 2. I have coded Trust as 0 when the Q2 response was 3. I have removed the cases where respondents marked Refused/Don't know in our independent variables. I am trying to make multiclass variables as dichotomous variables wherever possible, which will help in our analysis. I have also computed a separate new variable for our second independent variable based on the WEB-B question. It was a grid type question.



ASK ALL INTERNET USERS WHO USED THE INTERNET YESTERDAY (Q7=1):
WEB-B Next...Please tell me if you ever use the internet to do any of the following things. Do you ever... **(ROTATE ITEMS, BUT ALWAYS ASK "ACT01" FIRST)**

[IF YES ASK: Did you happen to do this YESTERDAY, or not?]

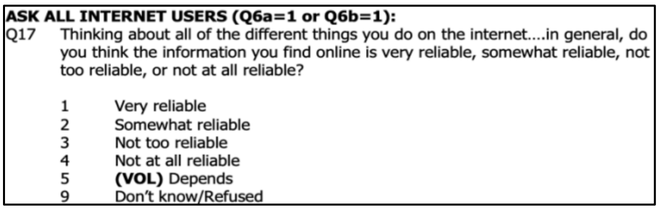
ACT01 Send or read email **[ALWAYS ASK FIRST]**
ACT14 Send "instant messages" to someone who's online at the same time
ACT16 Buy or make a reservation online for a travel service, like an airline ticket, hotel room, or rental car
ACT17 Take part in chat rooms or online discussions with other people
ACT33 Use an online dating website
ACT47 Get photos developed or display or store photos online
ACT49 Make a donation to a charity online
ACT52 Use on online search engine to help you find information on the Web
ACT54 Search online for information about someone you know or might meet
ACT58 Create a web log or "blog" that others can read on the web
ACT71 Read someone else's web log or blog
ACT73 Rate a product, service or person using an online rating system
ACT81 Send or receive text messages using a cell phone
ACT82 Use online classified ads or sites like Craig's list
ACT87 Use online social or professional networking sites like Friendster or LinkedIn
ACT92 Sell something online

CATEGORIES WEB-B

1 Yes, did this yesterday
2 Yes, do this (but NOT yesterday)
3 No, do not do this
9 **(DO NOT READ)** Don't know/Refused

Figure 21. WEB-B question from the survey.

This question will help us to know that for how many tasks that respondent uses the internet. So, I have computed a variable named Task which ranges from 0 to 8. I have simply recorded the number of mentioned tasks above for which the respondent uses the internet. Since there were eight tasks so our new variable has 8 levels. Our third independent variable is based on Q17 which asks the respondent that whether he/she think that whatever they search on the internet is reliable or not.

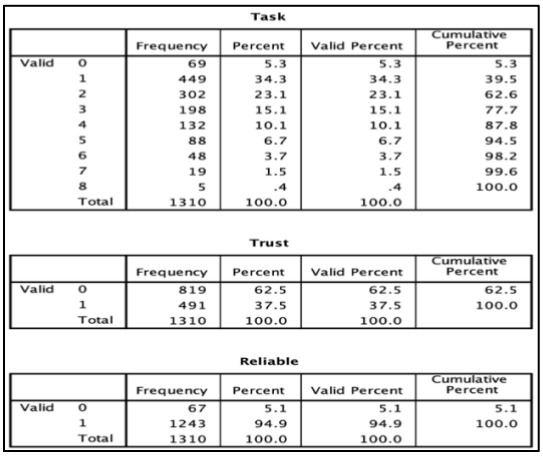


ASK ALL INTERNET USERS (Q6a=1 or Q6b=1):
Q17 Thinking about all of the different things you do on the internet....in general, do you think the information you find online is very reliable, somewhat reliable, not too reliable, or not at all reliable?

1 Very reliable
2 Somewhat reliable
3 Not too reliable
4 Not at all reliable
5 **(VOL)** Depends
9 Don't know/Refused

Figure 22. Q17 of the survey.

I have computed new variable name reliable based on this question and converted this multiclass variable into a dichotomous variable. So reliable variable has two classes which are reliable and not reliable.



Task

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	69	5.3	5.3	5.3
	1	449	34.3	34.3	39.5
	2	302	23.1	23.1	62.6
	3	198	15.1	15.1	77.7
	4	132	10.1	10.1	87.8
	5	88	6.7	6.7	94.5
	6	48	3.7	3.7	98.2
	7	19	1.5	1.5	99.6
	8	5	.4	.4	100.0
	Total	1310	100.0	100.0	

Trust

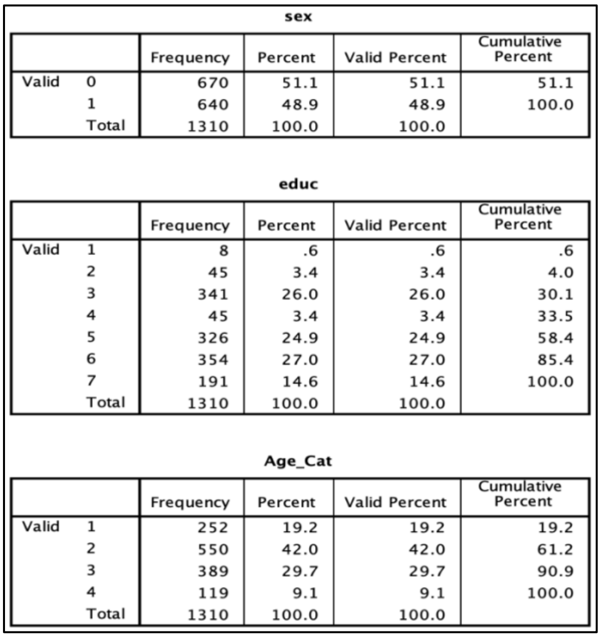
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	819	62.5	62.5	62.5
	1	491	37.5	37.5	100.0
	Total	1310	100.0	100.0	

Reliable

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	67	5.1	5.1	5.1
	1	1243	94.9	94.9	100.0
	Total	1310	100.0	100.0	

Figure 23. Frequency table of Task, Trust, and Reliable.

Now the other three remaining independent variables are demographic. The age variable was continuous, so I have converted it to a categorical variable. Sex is a categorical variable, and we have recoded females as 0 and 1 for males. Educ variable stores the level of education of the respondent and it is a multiclass variable. Figure 24 shows the data overview of demographic variables.



sex

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	670	51.1	51.1	51.1
	1	640	48.9	48.9	100.0
	Total	1310	100.0	100.0	

educ

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	8	.6	.6	.6
	2	45	3.4	3.4	4.0
	3	341	26.0	26.0	30.1
	4	45	3.4	3.4	33.5
	5	326	24.9	24.9	58.4
	6	354	27.0	27.0	85.4
	7	191	14.6	14.6	100.0
	Total	1310	100.0	100.0	

Age_Cat

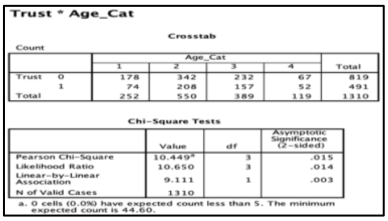
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	252	19.2	19.2	19.2
	2	550	42.0	42.0	61.2
	3	389	29.7	29.7	90.9
	4	119	9.1	9.1	100.0
	Total	1310	100.0	100.0	

Figure 24. Frequency table of sex and education

D. *Assumptions of Binary Logistic Regression:*

There are four assumptions we need to check before proceeding with the binary logistic regression.

1. The dependent variable should be always dichotomous. In our study, Datenew has two-level which are Yes and No.
2. There should be one or more independent variables that can be continuous or categorical. In our study, Task is a scale variable that ranges from 0 to 8, and the rest of the other five independent variables are categorical variables.
3. The third assumption is to check the independence of the observations. So, I have performed a Chi-square test on a categorical variable to check whether they are independent or not. According to the results of the Chi-square test, the Trust variable had a significant value less than 0.05 for sex, age, and Educ which suggests that we have enough evidence to reject the null hypothesis. The null hypothesis of the Chi-square test states that two categorical variables are independent. So, in our study, Trust has an association with sex, age, and Educ.



Trust * Age_Cat

Crosstab

Count

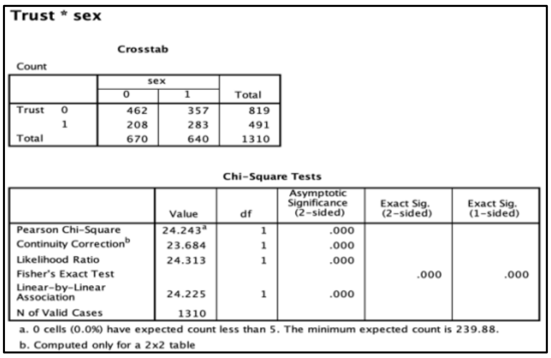
		Age_Cat				
		1	2	3	4	Total
Trust	0	178	342	232	67	819
	1	74	208	157	52	491
Total		252	550	389	119	1310

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	10.449[a]	3	.015
Likelihood Ratio	10.650	3	.014
Linear-by-Linear Association	9.111	1	.003
N of Valid Cases	1310		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 44.60.

Figure 25. Chi-square test between Trust and Age_Cat.



Trust * sex

Crosstab

Count

		sex		
		0	1	Total
Trust	0	462	357	819
	1	208	283	491
Total		670	640	1310

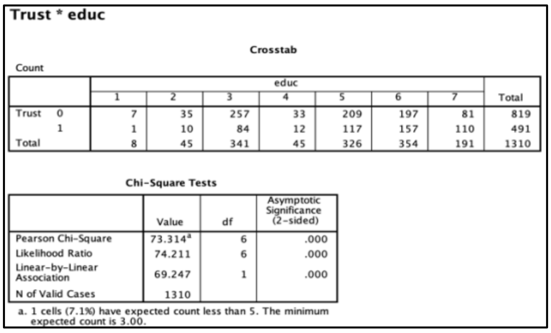
Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	24.243[a]	1	.000		
Continuity Correction[b]	23.684	1	.000		
Likelihood Ratio	24.313	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	24.225	1	.000		
N of Valid Cases	1310				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 239.88.

b. Computed only for a 2x2 table

Figure 26. Chi-square test between Trust and sex.



Trust * educ

Crosstab

Count

		educ							
		1	2	3	4	5	6	7	Total
Trust	0	7	35	257	33	209	197	81	819
	1	1	10	84	12	117	157	110	491
Total		8	45	341	45	326	354	191	1310

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	73.314[a]	6	.000
Likelihood Ratio	74.211	6	.000
Linear-by-Linear Association	69.247	1	.000
N of Valid Cases	1310		

a. 1 cells (7.1%) have expected count less than 5. The minimum expected count is 3.00.

Figure 27. Chi-square test between Trust and Educ.

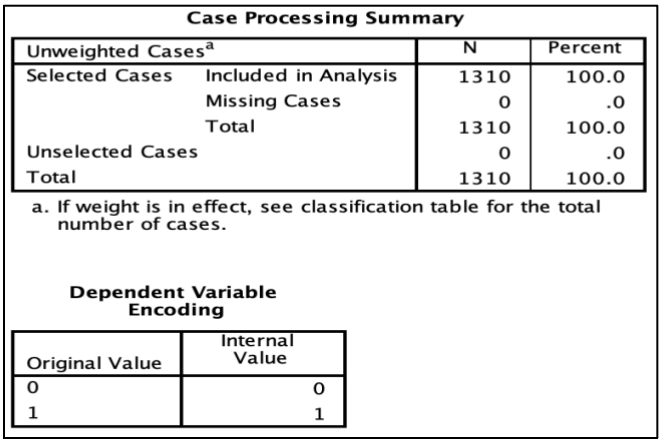
4. The fourth assumption is that there should a linear relationship between the scale variable and the logit transformation of the dependent variable. In our study, we have converted our scale variable which was age into the categorical variable. So, there is no scale variable in our model

E. *Model Building:*

Our independent and dependent variables are ready for producing a binary logistic model. Out of six independent variables, five independent variables are categorical, and one variable is a scale variable. While building the model, we can have a reference category for categorical variables. So, for the Trust variable, I have selected No as a reference category. For Age, I have selected 18-30 as a reference category. For the Reliable variable, I have selected Not reliable as a reference category. For sex, I have selected female as the reference category and for educ, I have selected Post-graduate training/professional school after college as the reference category. This method helps to interpret the result or output better. We can be able to compare which category helps the model to predict better. We will be using the Hosmer-Lemeshow-goodness of fit test to check the significance of the model. Now we can run the model and evaluate the model.

F. *Model Evaluation:*

When we run the binary logistic regression, we will be able to see always two blocks in the output. Block 0 suggests the null model. In the null model, all independent variables are excluded from the analysis, and the only intercept is involved in the model.



Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	1310	100.0
	Missing Cases	0	.0
	Total	1310	100.0
Unselected Cases		0	.0
Total		1310	100.0

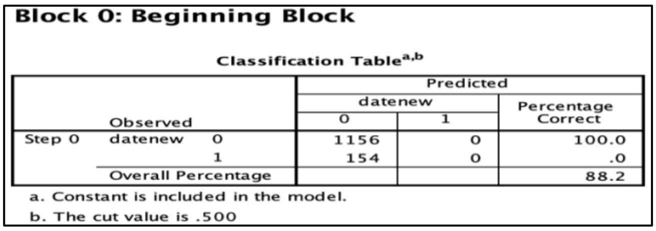
a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

Figure 28. Summary of data and dependent variable encoding.

As we have discussed the filtering of data earlier, so you can refer the Figure 28 to know about the number of cases filtered. Also, it is very important to check the dependent variable encoding table because if recodes of the dependent variable are other than 0 and 1 then the meaning of the analysis will change. Block 0 also contains the classification model produced only using the intercept and the efficiency of the model produced using intercept only is 88.2% which is quite good.



Block 0: Beginning Block

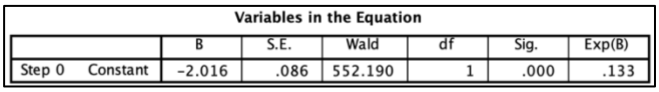
Classification Table[a,b]

Observed			Predicted: datenew 0	Predicted: datenew 1	Percentage Correct
Step 0	datenew	0	1156	0	100.0
		1	154	0	.0
	Overall Percentage				88.2

a. Constant is included in the model.

b. The cut value is .500

Figure 29. Classification table of the null model.

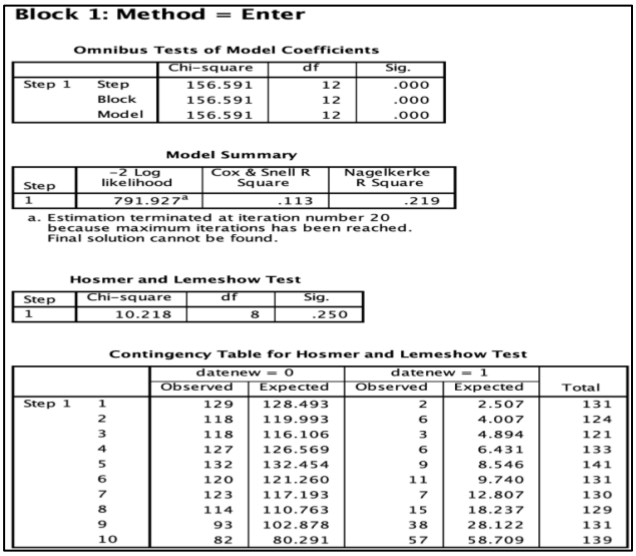


Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-2.016	.086	552.190	1	.000	.133

Figure 30. Variables in the equation of the null model.

Figure 30 suggests that constant is significant in the null model. Now we will see how our model performs when we add independent variables to the model. We will start by analyzing whether the model produced is significant or not.



Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	156.591	12	.000
	Block	156.591	12	.000
	Model	156.591	12	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	791.927[a]	.113	.219

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Hosmer and Lemeshow Test

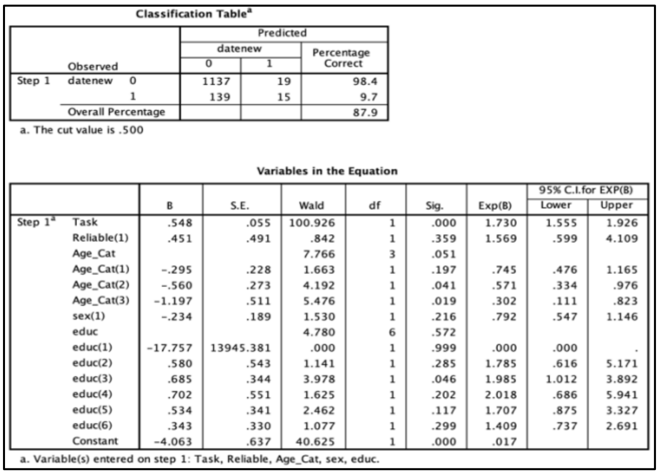
Step	Chi-square	df	Sig.
1	10.218	8	.250

Contingency Table for Hosmer and Lemeshow Test

		datenew = 0		datenew = 1		
		Observed	Expected	Observed	Expected	Total
Step 1	1	129	128.493	2	2.507	131
	2	118	119.993	6	4.007	124
	3	118	116.106	3	4.894	121
	4	127	126.569	6	6.431	133
	5	132	132.454	9	8.546	141
	6	120	121.260	11	9.740	131
	7	123	117.193	7	12.807	130
	8	114	110.763	15	18.237	129
	9	93	102.878	38	28.122	131
	10	82	80.291	57	58.709	139

Figure 31. Model significance, Model summary, the pseudo-R-square value of the model.

The omnibus test table suggests the significance of the model. By referring to Figure 31, we can see the significance value of the model under the Omni bust table is below 0.05 which suggests that the model produced is significant. Since we are performing the logistic regression so there will not be an R-square value like linear regression, but instead of that researcher like Nagelkerke and Cox and Snell introduced the pseudo-R-square value to estimate the variance in the model. So, by referring to Figure 31, we can say that our model can explain about 22% variation of the dependent variable based on the Nagelkerke R square method. Hosmer and Lemeshow Test are used to identify whether the data fits in the model or not. Figure 31 shows that the significance value of the Hosmer and Lemeshow test is above 0.05 which suggests that data fits well in the model. We will now analyze the classification table.



Classification Table[a]

Observed			Predicted: datenew 0	Predicted: datenew 1	Percentage Correct
Step 1	datenew	0	1137	19	98.4
		1	139	15	9.7
	Overall Percentage				87.9

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
Step 1[a]	Task	.548	.055	100.926	1	.000	1.730	1.555	1.926
	Reliable(1)	.451	.491	.842	1	.359	1.569	.599	4.109
	Age_Cat			7.766	3	.051			
	Age_Cat(1)	-.295	.228	1.663	1	.197	.745	.476	1.165
	Age_Cat(2)	-.560	.273	4.192	1	.041	.571	.334	.976
	Age_Cat(3)	-1.197	.511	5.476	1	.019	.302	.111	.823
	sex(1)	-.234	.189	1.530	1	.216	.792	.547	1.146
	educ			4.780	6	.572			
	educ(1)	-17.757	13945.381	.000	1	.999	.000	.000	.
	educ(2)	.580	.543	1.141	1	.285	1.785	.616	5.171
	educ(3)	.685	.344	3.978	1	.046	1.985	1.012	3.892
	educ(4)	.702	.551	1.625	1	.202	2.018	.686	5.941
	educ(5)	.534	.341	2.462	1	.117	1.707	.875	3.327
	educ(6)	.343	.330	1.077	1	.299	1.409	.737	2.691
	Constant	-4.063	.637	40.625	1	.000	.017		

a. Variable(s) entered on step 1: Task, Reliable, Age_Cat, sex, educ.

Figure 32. Classification table and Variables in the equation.

Based on Figure 32, we can say that 98.4% of respondents who said that they never had used the dating sites to meet new people were also predicted by the model that they never had visited any online dating website or other sites to meet new people. 9.7% of respondents who visited online dating sites were also precited by the model that they have visited the online dating site. The overall accuracy of the model in terms of classification is about 88%. The Wald test is used to check whether variables in the equation are significant or not. Based on Figure 32, we can say that the task, second category of age, third category of age, and third category of Educ are statistically significant. Also, this table is used to predict the probability of an event occurring based on one or more than one unit change in the independent variable when all other independent variables are kept constant. Based on Figure 32, we can say that the odds of a respondent using the dating website increases when the individuals use the internet for more activities.

G. *Conclusion:*

At the start of the analysis, we had proposed a few hypotheses, now we will check which of our hypothesis is true based on our model. So, my first hypothesis was about trusting others, but we had to remove the trust variable because it was having an association with other independent variables. Our second hypothesis came true which suggested that if individuals perform more activities on the internet then he is likely to be using the dating site. The third hypothesis is false because a reliable variable is not a significant variable in the equation. The fourth hypothesis turned to be false and based on Figure 32 we can say that women use dating sites more than men. The fifth hypothesis was turned out to be false, but it was not significant. It suggested that induvial whose education level is postgraduate are less likely to visit dating websites. The sixth hypothesis was true that individual whose age ranges between 18 and 30 are more like to visit dating websites than respondents whose age is more than this range. Finally, we can conclude that logistic regression was performed using Trust, Reliability, Task, Sex, Age, and Educ to predict the likelihood that respondents visit the online dating site. The model produced was statistically significant. The model explained 22% of the variance in respondents who use online dating sites and correctly classified 88% of the cases.

III. Principle Component Analysis

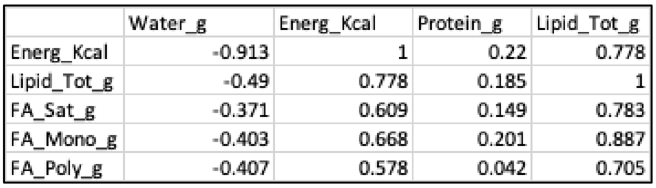
Before discussing about the PCA, we must know why we are using PCA. The machine learning method works great with huge datasets and having a large amount of data we can build a better predictive model. However, using a large amount of dataset is having its own drawbacks and the biggest drawback is the curse of dimensionality. Dimension is nothing but the features or attributes of the dataset. Suppose we produced multiple models using a different number of features of the same dataset then you can observe that as we add a greater number of features in the model then the model performance gets better and better. But this is not true completely, there is always a threshold of the number of features we can add to our model. If we keep on adding features after this threshold then model performance will start decreasing and the error rate increases. This is because features that keep on adding are sometimes irrelevant in the model and these features add confusion in the model and due to this model is not able to predict correctly. So, to avoid the curse of dimensionality or the problem of overfitting of the model, we use a technique called principle component analysis. PCA converts high dimensionality to low dimensionality. PCA is a dimension reduction technique that finds the correlation between the variable and thereby decreases the dimension but in this process, any important or significant information in the data is not altered or removed from the dataset. We find highly correlated feature in the dataset because these features cause an output which is bias. Also, these features are redundant feature because they do not account for your output.

A. *Process of PCA:*

The first step of PCA is the standardization of data. In this process, we scale our all variables such that their values range in fixed intervals. If we do not follow this process, then output using this variable will be biased because variables with a larger range will have a great influence on the output. It is a simple mathematical process where the variable value is subtracted by the mean and then divided by the standard deviation. The second step is computing the covariance matrix, it shows the correlation between different variables in the dataset. Using this matrix, we can be able to detect highly correlated variables and it is important to detect such variables because they make the model more biased and also impact the performance of the model. The third step is to calculate the Eigenvector and Eigenvalues. These are the mathematical values computed from your covariance matrix. These values determine the principle components of the dataset. We know that Eigenvectors and Eigenvalues are always computed in pairs. Dimension in the dataset will give us an idea about how many Eigenvectors we need to be calculated. Eigenvectors are calculated with the help of the covariance matrix. Eigenvectors identify in which direction or in which variable there is more variance in the dataset. Maximum variance stores the maximum information, so it is important to find this direction. In the fourth step, we will need to compute the principle components. After computing an eigenvector and eigenvalue, we will now arrange them in descending order. We mostly consider the components whose eigenvalue is greater than 1. We get a table in which components are arranged in descending order and along with that total variance explained by each component is also shown. In the last step, we can replace the variables of the dataset with the factors generated and thereby achieve the goal of reduction in dimension of the dataset.

B. *Example of PCA:*

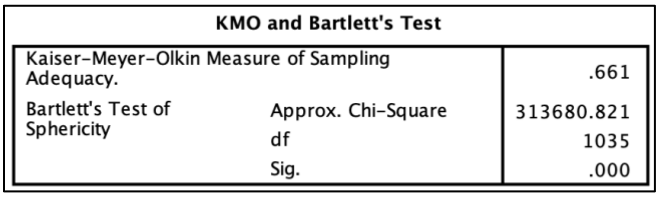
I have downloaded a dataset that was obtained from the database of USDA National Nutrient Database. This database contains the nutrients of various food groups. The dataset contains the values of minerals, vitamins, energy, macronutrient, and water. There are 46 such variables in the datasets. I will be implementing PCA on this dataset using SPSS. So, we analyze this process step by step now. The first step is to standardize all variables. In SPSS there is an option in the factor analysis tab called extraction. In that menu, if we checked the correlation matrix then it is equivalent to standardizing the variables. Now the second step is to compute the covariance matrix. As we are executing the PCA on 46 variables, so the matrix produced is very large. So, we will observe one case in which a variable is highly correlated with a couple of variables.



	Water_g	Energ_Kcal	Protein_g	Lipid_Tot_g
Energ_Kcal	-0.913	1	0.22	0.778
Lipid_Tot_g	-0.49	0.778	0.185	1
FA_Sat_g	-0.371	0.609	0.149	0.783
FA_Mono_g	-0.403	0.668	0.201	0.887
FA_Poly_g	-0.407	0.578	0.042	0.705

Figure 33. Part of the correlation matrix.

Figure 33 shows the part of the correlation matrix. By analyzing this matrix, we can say that Lipid_Tot_g is highly correlated with Engergy_Kcal, FA_Sat_g, FA_Mono_g, and FA_Poly_g. We can know which variables are highly correlated by analyzing the correlation matrix. Also, the correlation matrix is used for computing eigenvectors.

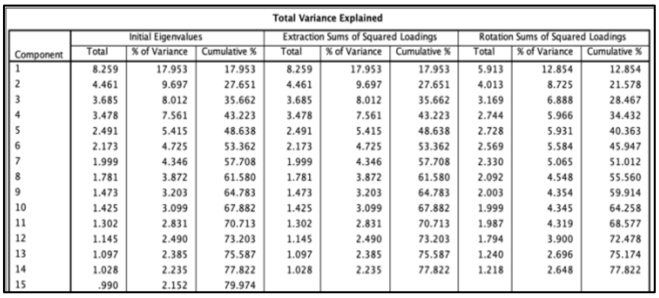


KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.661
Bartlett's Test of Sphericity	Approx. Chi-Square	313680.821
	df	1035
	Sig.	.000

Figure 34. KMNO and Bartlett's Test Result.

If the output of the Kaiser-Meyer-Olkin test is above 0.5 then it is said that the dataset is suited for factor analysis. Figure 34 shows that we obtain a 0.661 value in KMO which suggests that variables adequate for sampling.

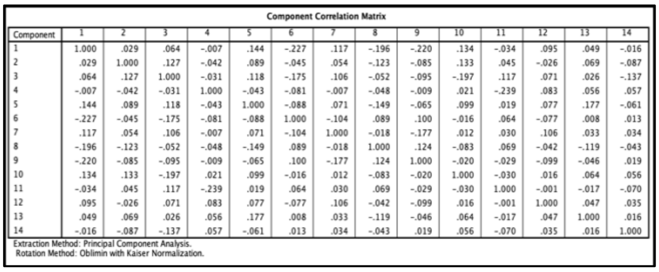


Total Variance Explained

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	8.259	17.953	17.953	8.259	17.953	17.953	5.913	12.854	12.854
2	4.461	9.697	27.651	4.461	9.697	27.651	4.013	8.725	21.578
3	3.685	8.012	35.662	3.685	8.012	35.662	3.169	6.888	28.467
4	3.478	7.561	43.223	3.478	7.561	43.223	2.744	5.966	34.432
5	2.491	5.415	48.638	2.491	5.415	48.638	2.728	5.931	40.363
6	2.173	4.725	53.362	2.173	4.725	53.362	2.569	5.584	45.947
7	1.999	4.346	57.708	1.999	4.346	57.708	2.330	5.065	51.012
8	1.781	3.872	61.580	1.781	3.872	61.580	2.092	4.548	55.560
9	1.473	3.203	64.783	1.473	3.203	64.783	2.003	4.354	59.914
10	1.425	3.099	67.882	1.425	3.099	67.882	1.999	4.345	64.258
11	1.302	2.831	70.713	1.302	2.831	70.713	1.987	4.319	68.577
12	1.145	2.490	73.203	1.145	2.490	73.203	1.794	3.900	72.478
13	1.097	2.385	75.587	1.097	2.385	75.587	1.240	2.696	75.174
14	1.028	2.235	77.822	1.028	2.235	77.822	1.218	2.648	77.822
15	.990	2.152	79.974						

Figure 35. Total Variance Explained Table.

By observing Figure 35, we can say that by only using fourteen components we can explain almost 78 percent of the variance in the model. We neglect the component whose eigenvalues are less than 1. For selecting correct rotation in the factor analysis, I have first used the Direct Oblimin method.



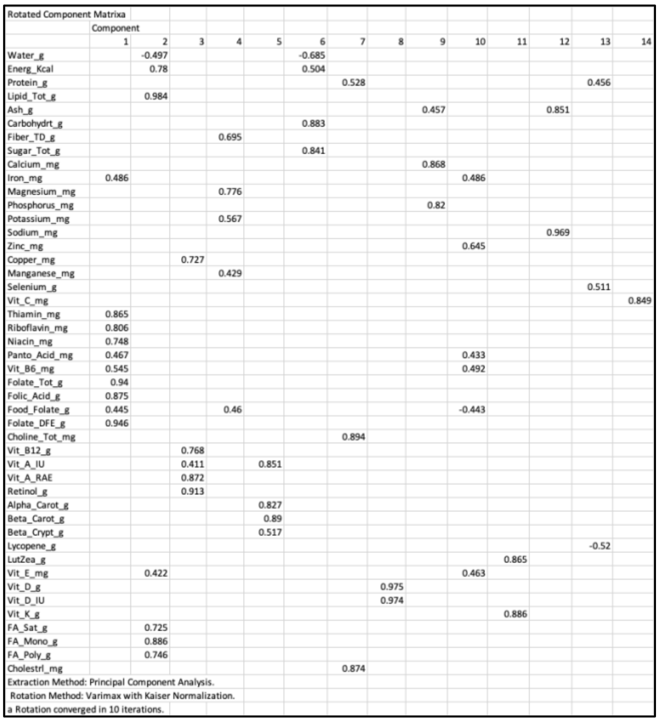
Component Correlation Matrix

Component	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1.000	.029	.064	-.007	.144	-.227	.117	-.196	-.220	.134	-.034	.095	.049	-.016
2	.029	1.000	.127	-.042	.089	-.045	.054	-.123	-.085	.133	.045	-.026	.069	-.087
3	.064	.127	1.000	-.031	.118	-.175	.106	-.052	-.095	-.197	.117	.071	.026	-.137
4	-.007	-.042	-.031	1.000	-.043	-.081	-.007	-.048	-.009	.021	-.239	.083	.056	.057
5	.144	.089	.118	-.043	1.000	-.088	.071	-.149	-.065	.099	.019	.077	.177	-.061
6	-.227	-.045	-.175	-.081	-.088	1.000	-.104	.089	.100	-.016	.064	-.077	.008	.013
7	.117	.054	.106	-.007	.071	-.104	1.000	-.018	-.177	.012	.030	.106	.033	.034
8	-.196	-.123	-.052	-.048	-.149	.089	-.018	1.000	.124	-.083	.069	-.042	-.119	-.043
9	-.220	-.085	-.095	-.009	-.065	.100	-.177	.124	1.000	-.020	-.029	-.099	-.046	.019
10	.134	.133	-.197	.021	.099	-.016	.012	-.083	-.020	1.000	-.030	.016	.064	.056
11	-.034	.045	.117	-.239	.019	.064	.030	.069	-.029	-.030	1.000	-.001	-.017	-.070
12	.095	-.026	.071	.083	.077	-.077	.106	-.042	-.099	.016	-.001	1.000	.047	.035
13	.049	.069	.026	.056	.177	.008	.033	-.119	-.046	.064	-.017	.047	1.000	.016
14	-.016	-.087	-.137	.057	-.061	.013	.034	-.043	.019	.056	-.070	.035	.016	1.000

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

Figure 36. Component Correlation using Direct Oblimin Method

By observing Figure 36, we can conclude that the value between different components is not above 0.32. So, I have decided to use the varimax rotation method for factor analysis.



Rotated Component Matrixa

	Component													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Water_g		-0.497				-0.685								
Energ_Kcal		0.78				0.504								
Protein_g							0.528						0.456	
Lipid_Tot_g		0.984												
Ash_g									0.457			0.851		
Carbohydrt_g						0.883								
Fiber_TD_g				0.695										
Sugar_Tot_g						0.841								
Calcium_mg									0.868					
Iron_mg	0.486									0.486				
Magnesium_mg				0.776										
Phosphorus_mg									0.82					
Potassium_mg				0.567										
Sodium_mg												0.969		
Zinc_mg										0.645				
Copper_mg			0.727											
Manganese_mg				0.429										
Selenium_g													0.511	
Vit_C_mg														0.849
Thiamin_mg	0.865													
Riboflavin_mg	0.806													
Niacin_mg	0.748													
Panto_Acid_mg	0.467									0.433				
Vit_B6_mg	0.545									0.492				
Folate_Tot_g	0.94													
Folic_Acid_g	0.875													
Food_Folate_g	0.445			0.46						-0.443				
Folate_DFE_g	0.946													
Choline_Tot_mg							0.894							
Vit_B12_g			0.768											
Vit_A_IU			0.411		0.851									
Vit_A_RAE			0.872											
Retinol_g			0.913											
Alpha_Carot_g					0.827									
Beta_Carot_g					0.89									
Beta_Crypt_g					0.517									
Lycopene_g													-0.52	
LutZea_g											0.865			
Vit_E_mg		0.422								0.463				
Vit_D_g								0.975						
Vit_D_IU								0.974						
Vit_K_g											0.886			
FA_Sat_g		0.725												
FA_Mono_g		0.886												
FA_Poly_g		0.746												
Cholestrl_mg							0.874							

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a Rotation converged in 10 iterations.

Figure 37. Rotated Component Matrix

The sample size of the dataset was 8791. So, I have decided to keep factor loading equal to 0.45. Based on this factor loading, we can see which variable comes under which particular factor by referring to Figure 37. Hence, we have successfully implemented the PCA method on our dataset.

REFERENCES

[1] Selling price of oats, Europe. Accessed on: December 22, 2020. [Online]. Available: https://ec.europa.eu/eurostat/web/main/data/database.

[2] Online Dating, USA. Accessed on December 18, 2020. [Online]. Available: https://www.pewresearch.org/internet/dataset/september-2005-online-dating/

[3] USDA National Nutrient Database, USA. Accessed on January 1, 2020. [Online]. Available: https://data.world/sharon/usda-nutrient-database-sr-28