A Statistical Analysis of Credit and Debit Card Usage Patterns

Dr. Nileema Bhalerao

Associate Professor, Department of Statistics, Fergusson College (Autonomous) Pune

ABSTRACT

This analysis deals with finding relation between encountering credit and debit card frauds and different age groups, gender, area of residence, monthly income, credit card limit and several other factors. It also involves fitting a prediction model on the credit card usage pattern of individuals. It includes studying a times series model for credit and debit card usage over the last decade and developing a forecasting model to predict the forecast using the previous data available.

INTRODUCTION

The world is turning cashless and Debit and credit cards are two of the most commonly used payment cards in the world. They both have a series of numbers embossed or printed along with the cardholder’s name on the front. Each has a magnetic stripe on the back, a special security code, and an embedded microchip on the front that encrypts key personal and financial information related to the cardholder and the related account. Credit cards give you access to a line of credit issued by a bank and thus provides us a flexibility to make purchases and pay for itlater which is the biggest advantage for people turning towards using credit cards. Whereas when a purchase is made through debit card the money is debited from ones account at that very moment. Through the introduction of credit and debit card people don’t have to worry about carrying cash everywhere and thus limiting their transaction amount. With the introduction of UPI the flexibility of doing cashless transactions is achieved even more. All the transactions can be made at the ease of the fingertips within the smartphone. Even the smallest of a transactionis made through UPI right from paying to the auto-driver or purchasing grocery or shopping online. But with the increasing number of online transactions there also is an increased risk of encountering frauds. Fraud’ in card transactions is unauthorized and unwanted usage of an account by someone other than the owner of that account. The first universal credit card which could be used at a variety of establishments, was introduced by the Diners’ Club, Inc., in 1950. Another major card of this type, known as a travel and entertainment card, was established by theAmerican Express Company in 1958 where as the first debit card was introduced in 1982 in Canada by Saskatchewan Credit Unions. But the fashion of making transactions by a card was not much popular back then now through digitalization and the world on the verge of turning cashless the use plastic money (credit and debit card) has increased a lot. Through this paper i would like to shed light on if there is any relation between the different age groups using credit and debit card and the chances of them encountering a fraud or is fraud related to specific age group or a specific area of residence and also many other factors.

OBJECTIVES

1. To analyze credit card usage pattern of individuals.
2. To fit a prediction model using logistic repression on the basis of the usagepattern for detection of fraud.
3. To fit a Time series on credit and debit card transactions per month Toidentify seasonal component, trend component, irregular component.
4. To develop a forecasting model using ARIMA technique to predict a forecastfor a given year using up the data of previous.

DATA COLLECTION

This project is executed using with two datasets.

The primary data collected through GOOGLE FORM. The data is collected for various aspects such as age, educational level, job profile, use of credit card , debitcard , the area of residence, monthly income, frauds encountered and it's types, etc. Another important thing to note is that the volunteers were not disclosed to disclose their personal information like Phone Number, Email ID, Bank Account number, CVV, Credit/Debit Card number, even the names of te volunteers was not recorded. This confidentiality gave the volunteers a sense of reassurance that their data will not be misused.

Out of the total 514 observations, **466** were legitimate transactions while **47** werefraud transactions. The total observations recorded were **514** out of this close to **80%** were debit cardusers, **7%** were Credit card users and **13%** used both Credit and Debit cards.The secondary data has been obtained through the official website of RBI. This data was compiled in such a way that the Monthly transactions data from April 2011 to February 2022 was procured. This data consists of Number of outstandingcards per month, Number of transactions of Credit cards(at POS and at ATM)and Number of transactions of Debit cards(at POS and at ATM). We use this data to fit Time Series on Number of Transactions for both Credit and Debit Cards.

SOFTWARES USED

R software

Excel

R-Markdown

EXPLORATORY DATA ANALYSIS

Pie chart: A pie chart is a circular statistical graphic, which is divided into slicesto illustrate numerical proportion. In a pie chart, the arc length of each slice is proportional to the quantity it represents.

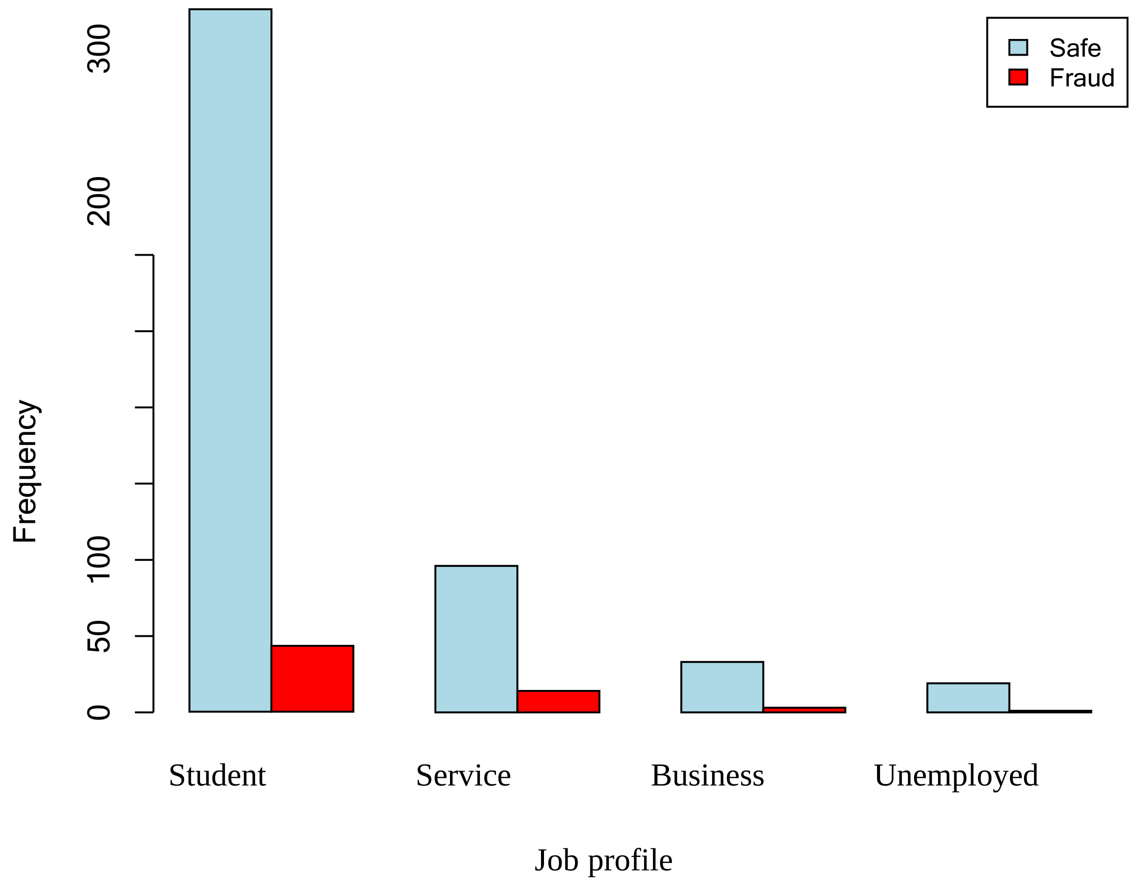
Histogram: A histogram is a bar graph-like representation of data that buckets arange of outcomes into columns along the x-axis. The y-axis represents the numbercount or percentage of occurrences in the data for each column and can be usedto visualize data distributions.

Sub-divided bar graph: Sub-divided bar diagrams are those diagrams which simultaneously present, total values as well as part values of a set of data. Differentparts of a bar must be shown in the same order for all bars of a diagram.

Venn Diagram: A Venn diagram uses overlapping circles or other shapes to illustrate the logical relationships between two or more sets of items.

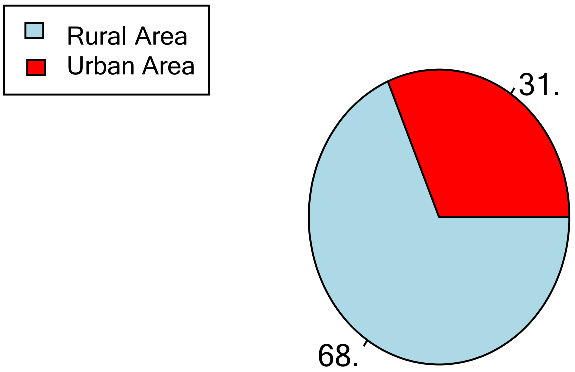
***EXPLORATORY DATA ANALYSIS FOR THE FORM DATA*: -** 1.

1.Fraud among Job profiles



Interpretation-We can say that the maximum frauds which are faced by acategory are students.

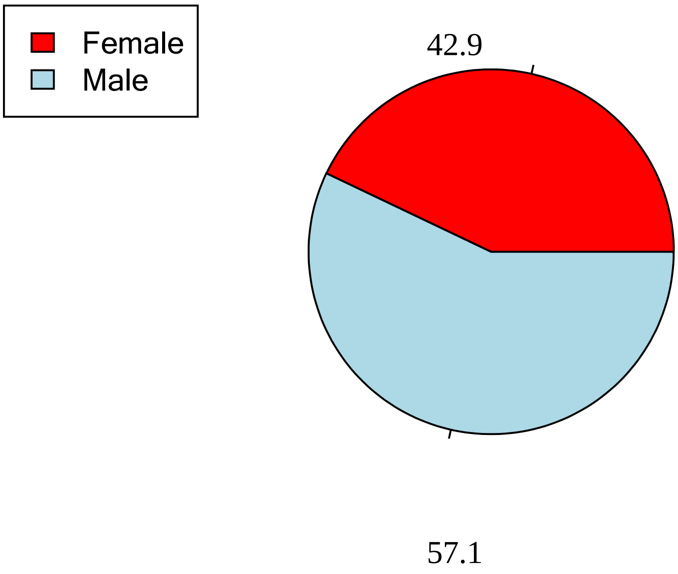
2. Pie chart for Area of residence for Credit card users



Conclusion- We can observe that 31.4% of the Credit card users in our data-setare residing in Rural Areas while 68.6 percent are residing in Urban Areas

3. *Pie Chart for Gender of Debit Card users*

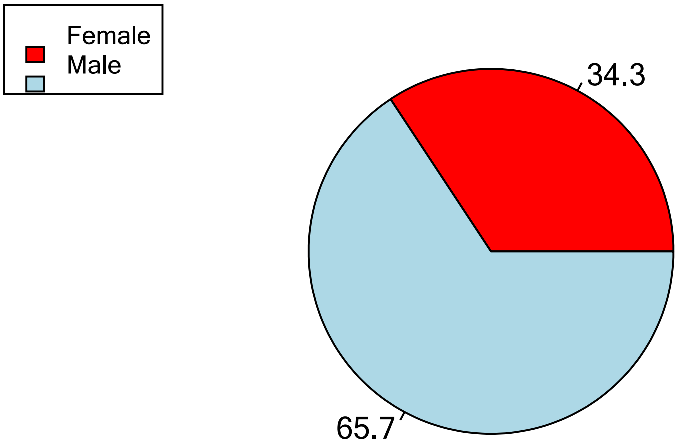
Gender of Debit card users



Conclusion- We can observe that 42.9% of the Debit card users in our data-setare Females and 57.1 percent are Males

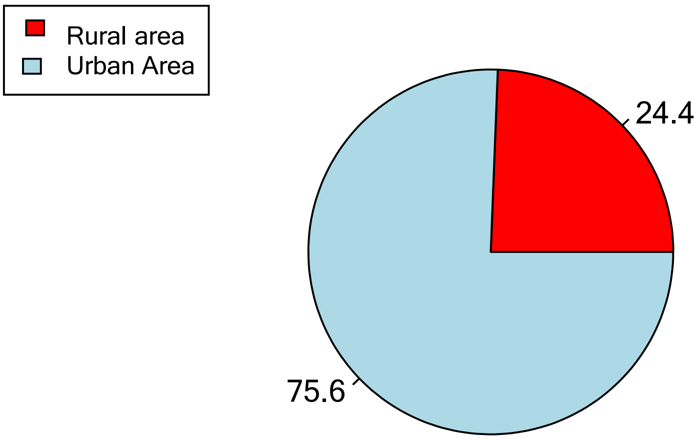
4. Pie Chart for Gender of Credit card users

Gender of Credit card users



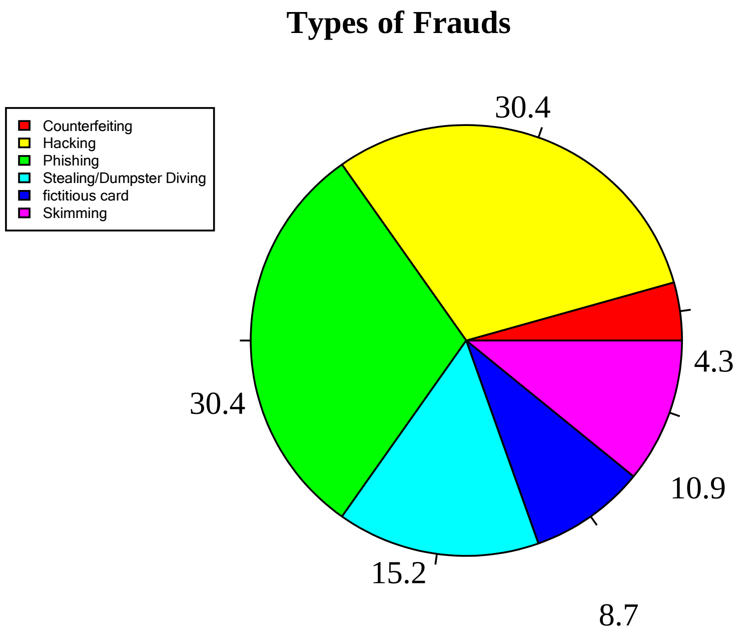
Conclusion- We can observe that 65.7% of the Credit card users in our data-setare Males and 34.1 percent are Females

5 Pie chart for Area of residence of Debit card users



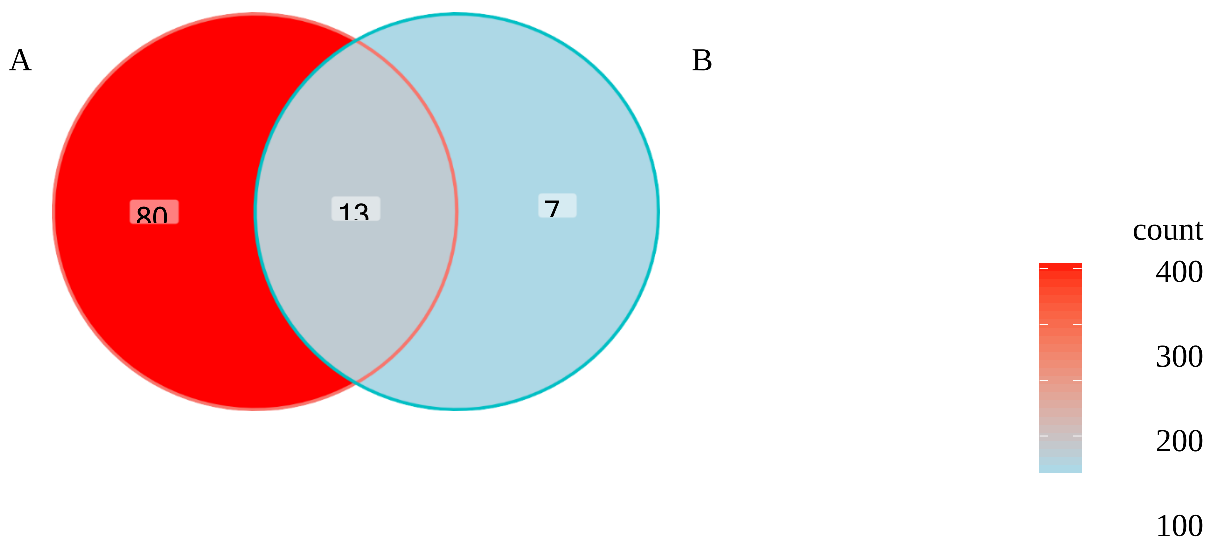
Conclusion- We can observe that 24.4% of the Credit card users in our data-setare residing in Rural Areas while 75.6 percent are residing in Urban Areas

6 Pie Chart for types of frauds



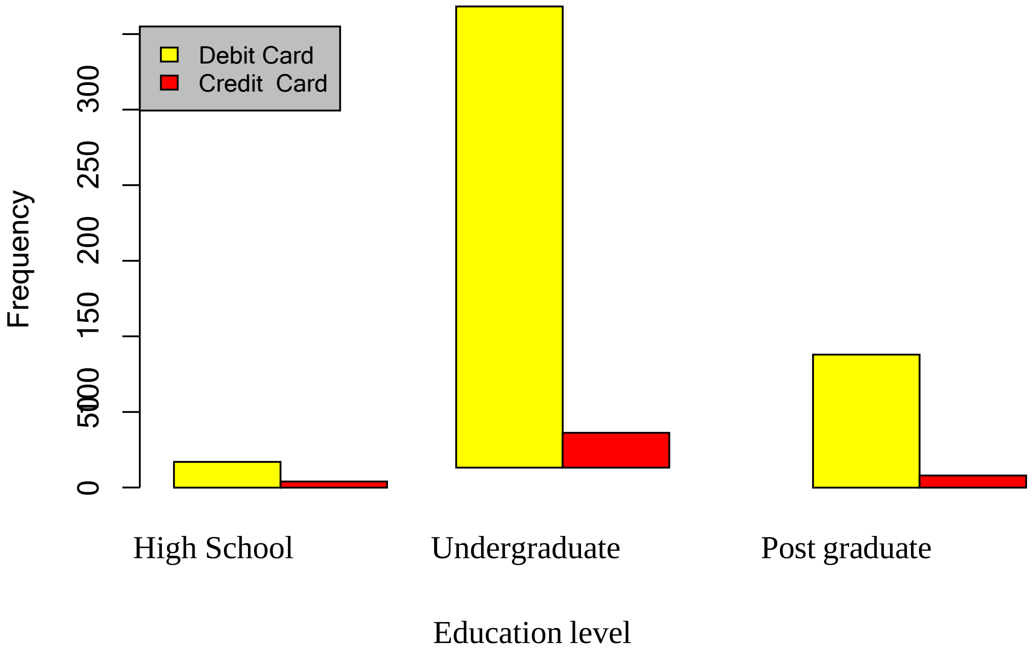
Conclusion- We can observe that the majority of the fraud type is Phishingand Hacking which account for 30.4% of the total frauds each, the next prevalent fraud type is Stealing/dumpster diving.

7 Venn Diagram for Debit and Credit Card users



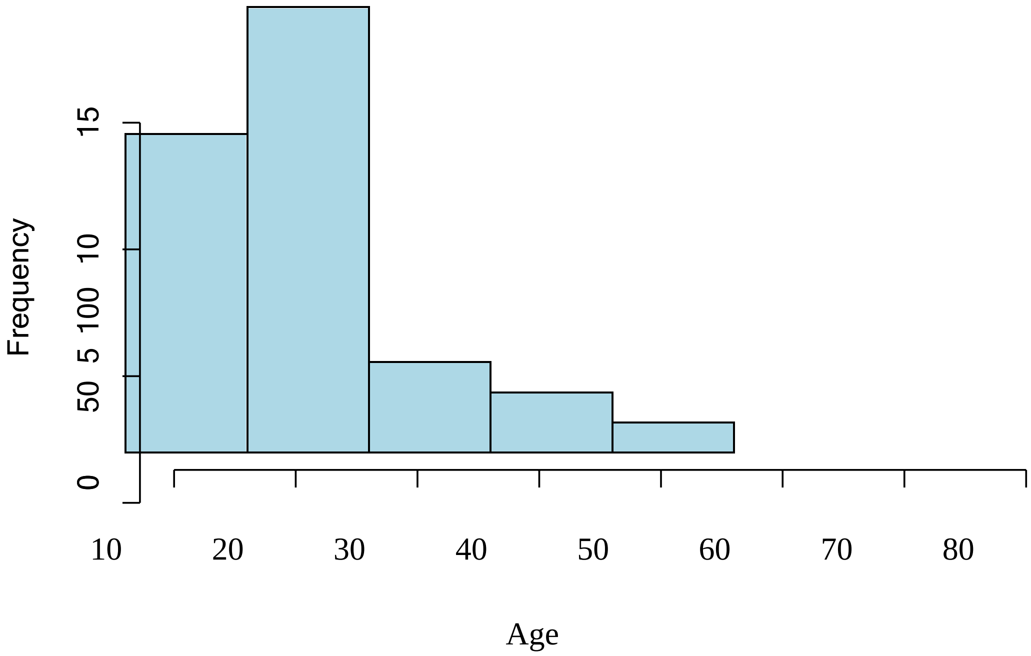
Conclusion- Here, we can observe that close to 80% of the people in our data-set use Debit cards. Whilst only 7% use a Credit Card. However, the percentage of people using both Debit and Credit Card is 13%.

8 Education levels among Debit and Credit card users



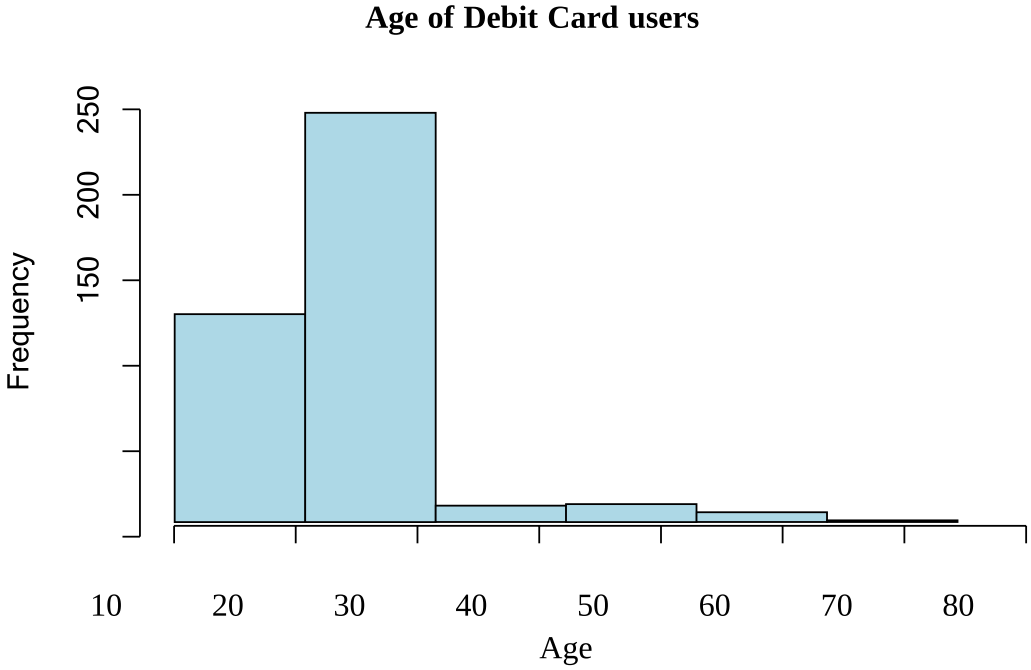
Conclusion- Here, we can observe that in both Credit and Debit card cate-gories, the people with almost undergraduate degree are using Debit/Credit card more. And most importantly, the penetration of Credit/Debit card in the people having almost High School qualification is the least.

9 Histogram for the age of Credit Card users



Conclusion- Here, we can observe that for Credit card users the Modal agegroup is 20-30. However, an important thing to observe here is that this is a sample of only 500 observations and so, there might be some deviation from the Age of the Population under study.

10 Histogram for age of Debit Card users



Conclusion- Here, we can observe that for Credit card users the Modal agegroup is 20-30. However, an important thing to observe here is that this is a sample of only 500 observations and so, there might be some deviation from the Age of the Population under study.

LOGISTIC REGRESSION

Logistic regression is used to predict the class (or category) of individuals based onone or multiple predictor variables (x). It is used to model a binary outcome, that is a variable, which can have only two possible values: 0 or 1, yes or no, diseased or non-diseased. Here we use Binary Logistic Regression Model- Used when the response is binary (i.e., it has two possible outcomes). The cracking example givenabove would utilize binary logistic regression. Other examples of binary responsescould include passing or failing a test, responding yes or no on a survey, and havinghigh or low blood pressure. The model for logistic regression is given as

$$Y = \pi(x) + \varepsilon$$

where,

$$\pi(x) = \frac{e^{\beta_0+\beta 1*x1+...+\beta 9*x9}}{1+e^{\beta_0+\beta 1*x1+...+\beta 9*x9}}$$

and $\varepsilon \sim \mathrm{B}(\pi(x))$

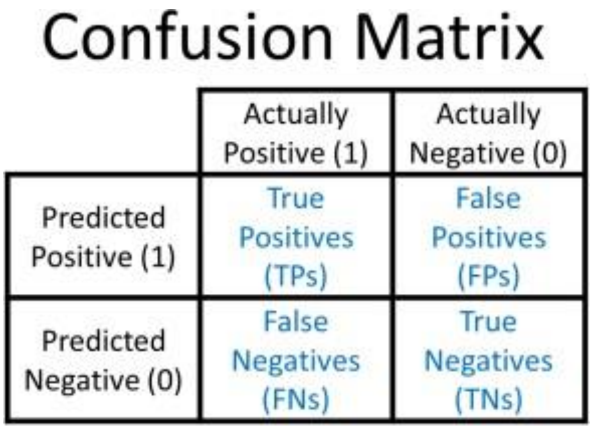
Note that- $\beta_0, \beta_1, .., \beta_9$ are the regression coefficients$

We are performing Logistic regression on the Debit Card users data-set toobtain a prediction model for FRAUD with the following regressors and response variable.

Logistic regression belongs to a family, named Generalized Linear Model (GLM),developed for extending the linear regression model to other situations. Other synonyms are binary logistic regression, binomial logistic regression and logit model. Logistic regression does not return directly the class of observations. It allows us to estimate the probability (p) of class membership. The probability will range between 0 and 1. You need to decide the threshold probability at which the category flips from one to the other. By default, this is set to p = 0.5, but in reality it should be settled based on the analysis purpose.

x	y
Fraud	Y
Gender	X1
Age	X2
Education	X3
Income	X4
Job	X5
Debit card limit	X6
Debit card frequency	X7
Debit card expense	X8

This data is now segregated into 80% Training and 20% Test data-set. A logistic regression model is fitted on the Training data-set, and using it we can proceed to predict the values of the Test data-set. Now, we proceed for defining the Confusion matrix



Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

From the Confusion Matrix it is clear that the accuracy of our model is 90.9%.

	0	1
0	68	5
1	2	1

	x
Accuracy	0.9078947
Kappa	0.1790123
Accuracy Lower	0.8193921
Accuracy Upper	0.9621619
Accuracy Null	0.9210526
Accuracy P Value	0.7498360
Mcnemar P Value	0.4496918

s.deviance	X.Residual.Deviance.
122.2744	Residual Deviance

s.null.deviance	X.Null.Deviance.
170.016	Null Deviance

s.aic	X.AIC.
178.2744	AIC

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-9.1153417	2.9813420	-3.0574626	0.0022322
age	0.0734748	0.0487571	1.5069553	0.1318221
gender Male	1.9204556	0.6859269	2.7997962	0.0051135
Education Post Graduate	0.9510107	1.3443669	0.7074042	0.4793154
Education Undergraduate	-0.4151767	1.2448471	-0.3335162	0.7387446
income25L-50L	-18.9201428	7576.6705050	-0.0024972	0.9980076
income3.5L-5L	-0.5584203	1.3071354	-0.4272092	0.6692270
income50L and above	-19.2353668	7462.0458989	-0.0025778	0.9979432
income5L-10L	-1.7021590	1.6244119	-1.0478617	0.2947023
Income below 3.5L	-1.2906616	1.3594752	-0.9493823	0.3424262
job Private	0.6211563	1.3598079	0.4567971	0.6478169
job Public	-17.4694146	7283.2385243	-0.0023986	0.9980862
job Student	2.0770289	1.3031134	1.5938972	0.1109591
job Unemployed	-17.2038913	2299.7799357	-0.0074807	0.9940313
Area Urban area	0.4298246	0.5767835	0.7452095	0.4561451
Debit_limit1L-2L	1.0340945	1.6814589	0.6149984	0.5385558
Debit_limit20K-50K	-0.9914687	0.7324723	-1.3535922	0.1758665
Debit_limit2L-3L	-20.4080638	7332.8892422	-0.0027831	0.9977794
Debit_limit3L and above	-19.7903227	3298.2558275	-0.0060002	0.9952125
Debit_limit50K-1L	-0.4369473	0.7569621	-0.5772380	0.5637787
Debit limit I don't use a debit card	1.9228500	1.2089507	1.5905116	0.1117195
debit frequency	0.8001439	0.2634011	3.0377388	0.0023836
Debit_expense10k-20k	-0.1797182	0.7749748	-0.2319020	0.8166141

Debit_expense20k-30k	2.6554254	0.8080932	3.2860387	0.0010161
Debit_expense30k-40k	-17.0002945	4301.7367218	-0.0039520	0.9968468
Debit_expense40k-50k	-19.0256786	4882.1044630	-0.0038970	0.9968906
Debit_expense50k and above	0.2608816	1.5210708	0.1715118	0.8638213
Debit expense Don't use a debit card	0.8510787	1.1552122	0.7367293	0.4612870

From the above result we can conclude that $(Null\ Deviance - Residual\ Deviance) >$ $\chi^2_{9;0.05}$ hence, we may conclude that the regression model is significant at 5% LOS.

Further we get to know that, the *following regressors are significant* - Debit cardlimit, Debit card expense, Debit card usage Frequency, Job, Gender.

Now, proceed to obtain the confusion matrix and get the Accuracy of ourpredictions

TIME SERIES ANALYSIS

Decomposing the Time Series

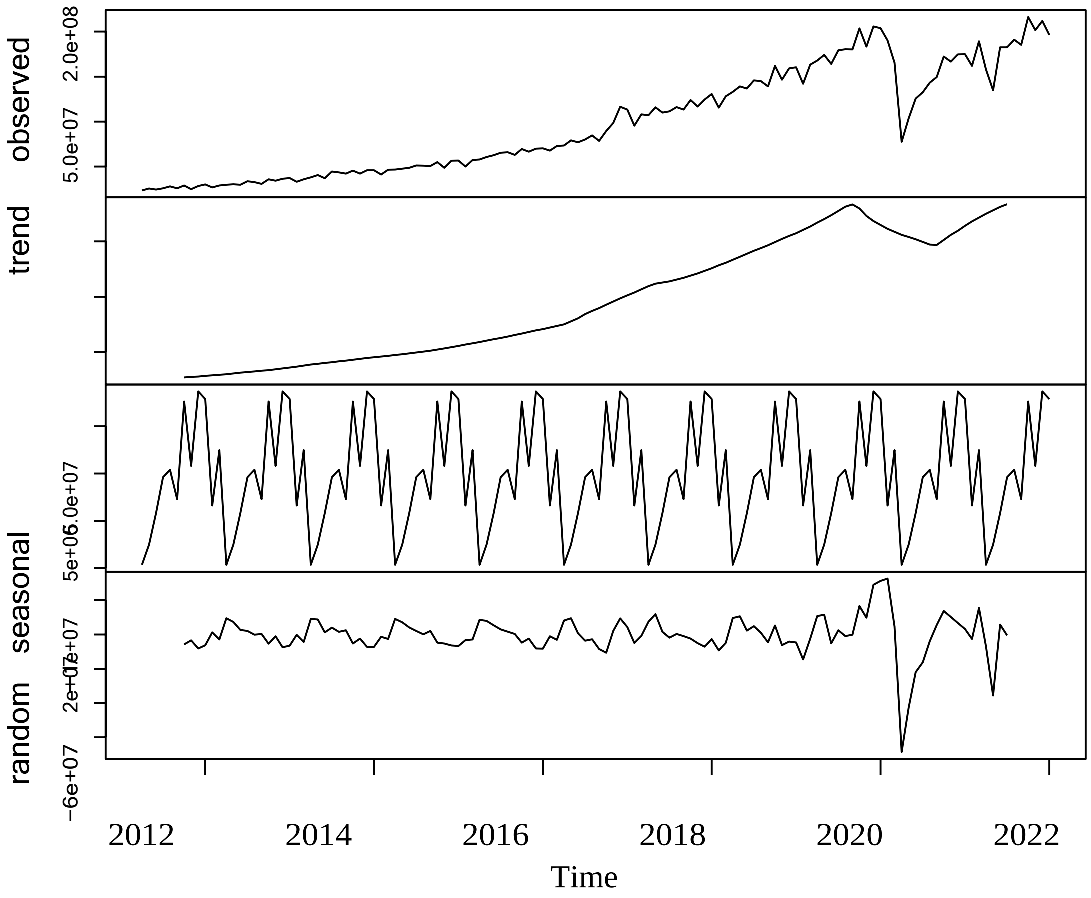
Time series arise as recordings of processes which vary over time. Time series plotdisplays the values of the process output in the order in which the values occur. A recording can either be a continuous trace or a set of discrete observations. We willconcentrate on the case where observations are made at discrete equally spaced times. By appropriate choice of origin and scale we can take the observation timesto be 1, 2, . . . T and we can denote the observations by Y1, Y2, .. . , YT .Akey analyzing a time series is to understand the form of any underlying patternof the data ordered over time. The pattern potentially consists of several different components, all of which combine to yield the observed values of the time series. There are 4 components of time series Trend, Seasonality, Cyclical Component and Random Component.

Trend: Long-term, gradual increasing, decreasing or stagnating tendency of the variable Y.

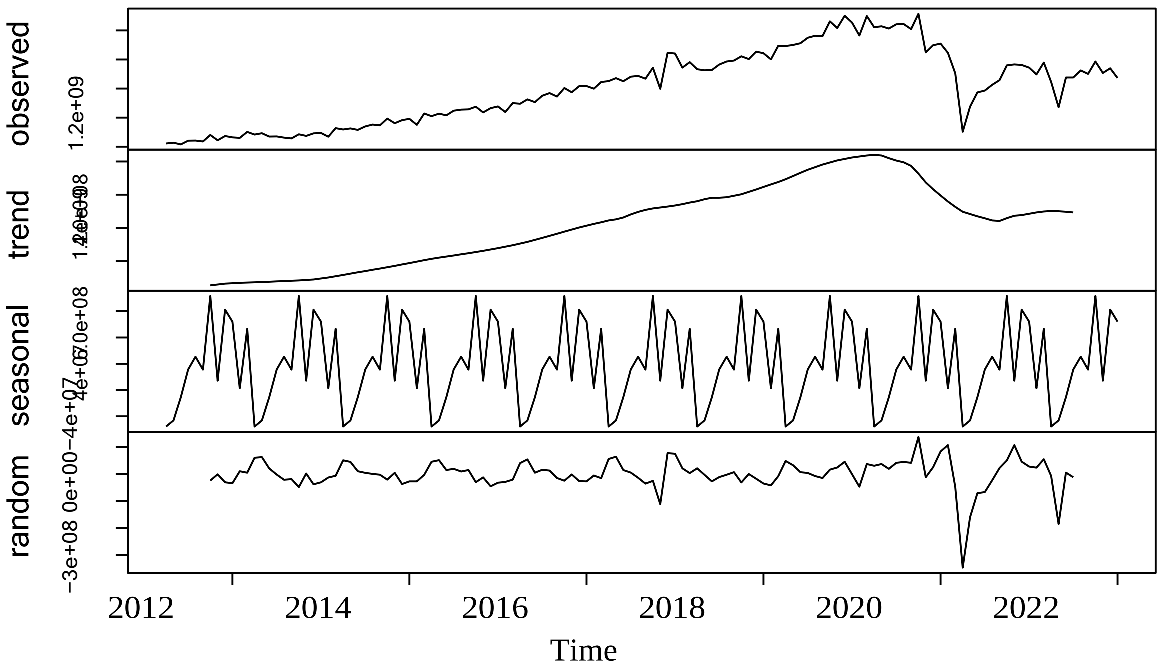
Seasonality: Regular, relatively short-term (yearly) repetitive up and down fluc-tuations of the variable Y depending on the season. **Cyclical Component:** Agradual, long-term, up and down potentially irregular swings of the variable Y. **Random Component:** A random increase or decrease of the variable Y for aspecific time period.

The data which we have is Monthly Debit and Credit Card transactions per month from April 2011 to Feb 2022, first we will decompose the Time Series for Debit Cards, Credit cards

Decomposition of additive time series



Decomposition of additive time series



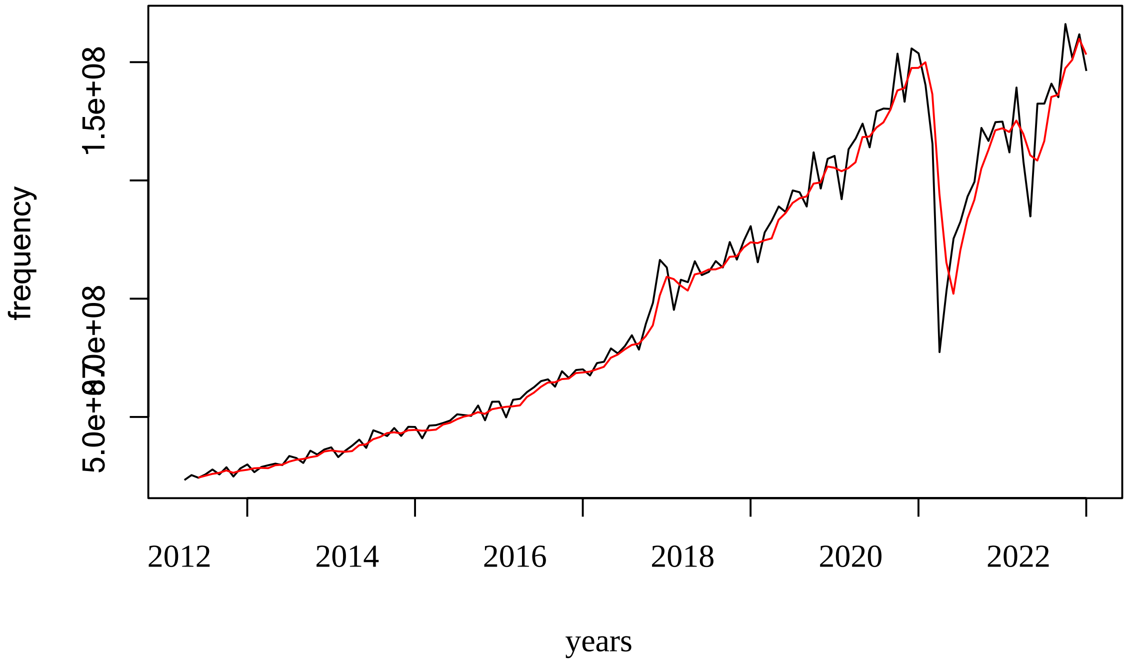
Conclusion: From the above graph it is clearly visible that the Number of transactions has an *increasing secular trend*, also one can observe a seasonal pattern from the graph, this may be attributed to Festivals like *Diwali*(High spending pattern is observed), and a reduce in Number of transactions is observed in the months of *February* which can be attributed to *Financial Year ending*, where all of the banks are closing their books and the failure rate of transactions grow.

Also, an important fact to note here is that in the 4th graph, we observe that there is a sharp decrease in transactions, this can be attributed to the *NationwideLock-down imposed attributed to the COVID-19 out-break.*

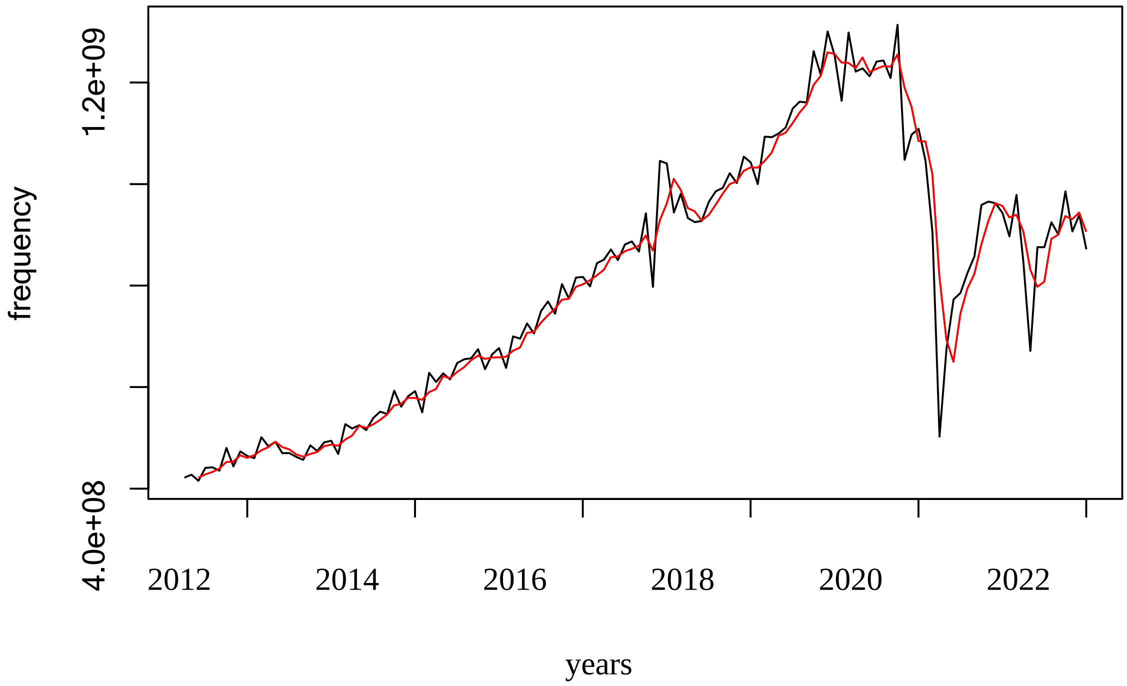
Also, in the 4th graph itself one can observe that there is a sharp irregular increasein the Number of transactions around Nov 2016. This can be attributed to the *Demonetization exercise carried out by the Govt. of India*.

Now, we look to fit a 3 period Moving Average, the following is a brief look backon Moving averages.

3 Monthly moving average for Credit card transactions



3 Monthly moving average for Debit card transactions



However, this method is not that useful when it comes to prediction on the timeseries is concerned.

We can use the Holt-Winters Triple exponential smoothing model to predict the data.

HOLT-WINTERS TRIPLE EXPONENTIAL SMOOTHING-

Here is a brief recall on **Exponential Smoothing -**

Triple exponential smoothing is used to handle the time series data containing a seasonal component. This method is based on three smoothing equations: sta- tionary component, trend, and seasonal. Both seasonal and trend can be additive or multiplicative. The three aspects of the time series behavior—value, trend, and seasonality—are expressed as three types of exponential smoothing, so Holt- Winters is called triple exponential smoothing. The model predicts a current or future value by computing the combined effects of these three influences. The model requires several parameters: one for each smoothing (α, β, γ), the length ofa season, and the number of periods in a season.

To perform Holt's triple exponential smoothing, we divide the data into 2 parts, the training set and the test set. The training set is from *April 2011 to July 2021*, and the test set is from *August 2021 to Feb 2022*.

	CC	DC
alpha	0.9244575	0.8947077
beta	0.0000000	0.0000000
gamma	1.0000000	0.3572679

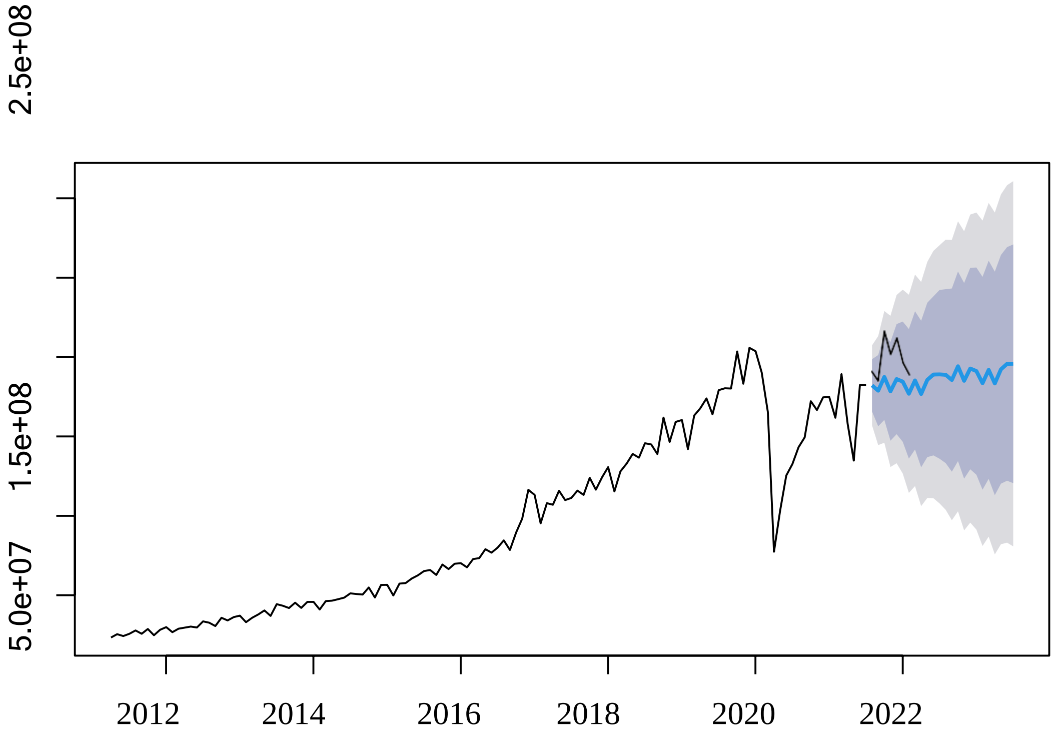
Interpretation- Here we can observe that in the case of Credit card transactions, α=0.9244575, β=0 and γ=1.

And in the case of Debit cards transactions, we observe that α=0.8947077, β=0and γ=0.3572679.

Now, we will check for the accuracy for both of our models,

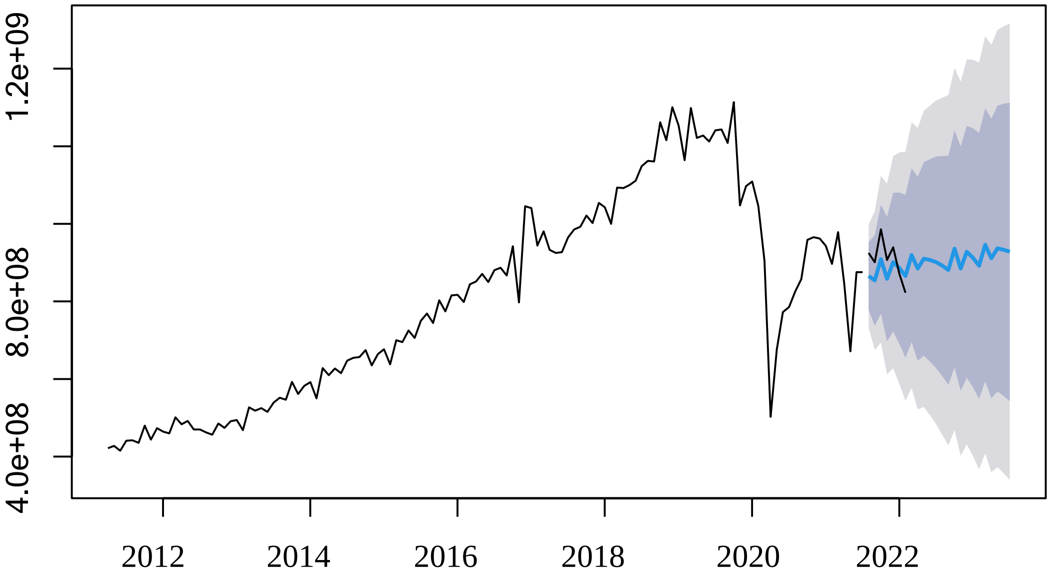
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	877408.9	12835344	6414093	0.3460736	5.858196	0.8357226	0.0137645
Test set	16588830.3	18548828	16588830	8.1567590	8.156759	2.1614374	NA

Forecasts from Holt Winters



	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	1687629	67718982	36941432	-0.1095647	4.624443	0.8522335	0.0243148
Test set	30339958	49922656	46662498	3.1359434	5.092340	1.0764970	NA

Forecasts from Holt Winters



In this plot we can observe that the observed values lie in the 90% confidence bands of our predictions for both Credit and Debit Card transactions, hence we can say our predictions are accurate. This can be further proved by the value of RMSE for both the observations.

Stationarity and ACF, PACF plots

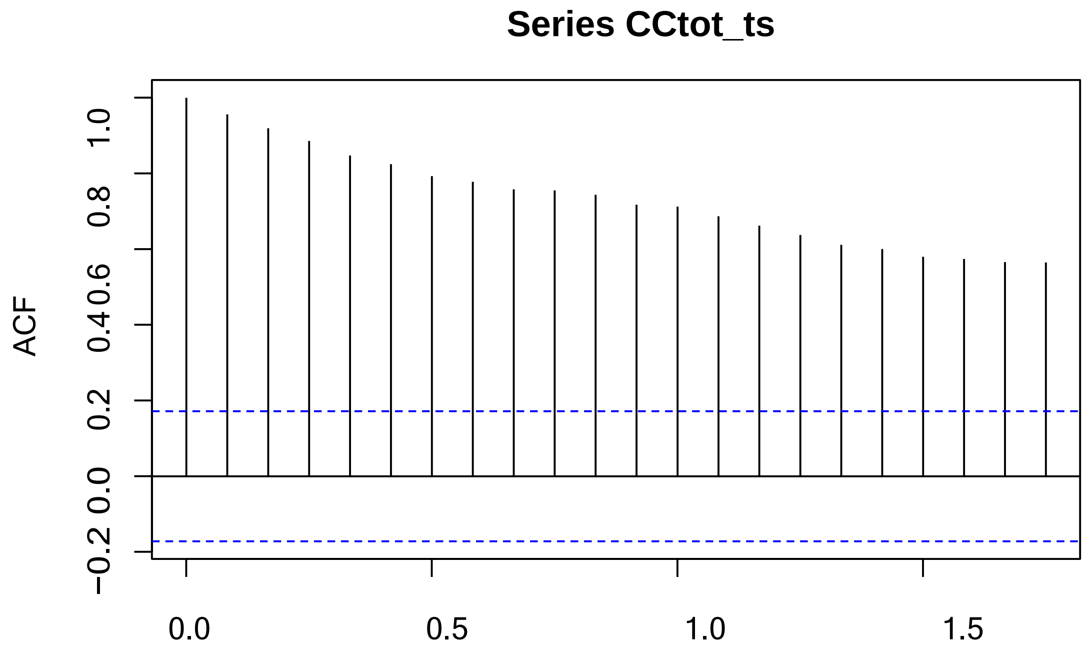
We will now check the stationarity of both the Time Series, before that, let usrecall **Stationarity of a Time series-**
In the most intuitive sense, stationarity means that the statistical properties of aprocess generating a time series do not change over time i.e the time series shows a constant mean and variance.
We perform the **KPSS test** for stationarity for both datasets-
H_0: The time series is stationary.
H_1: The time series is non-stationary.

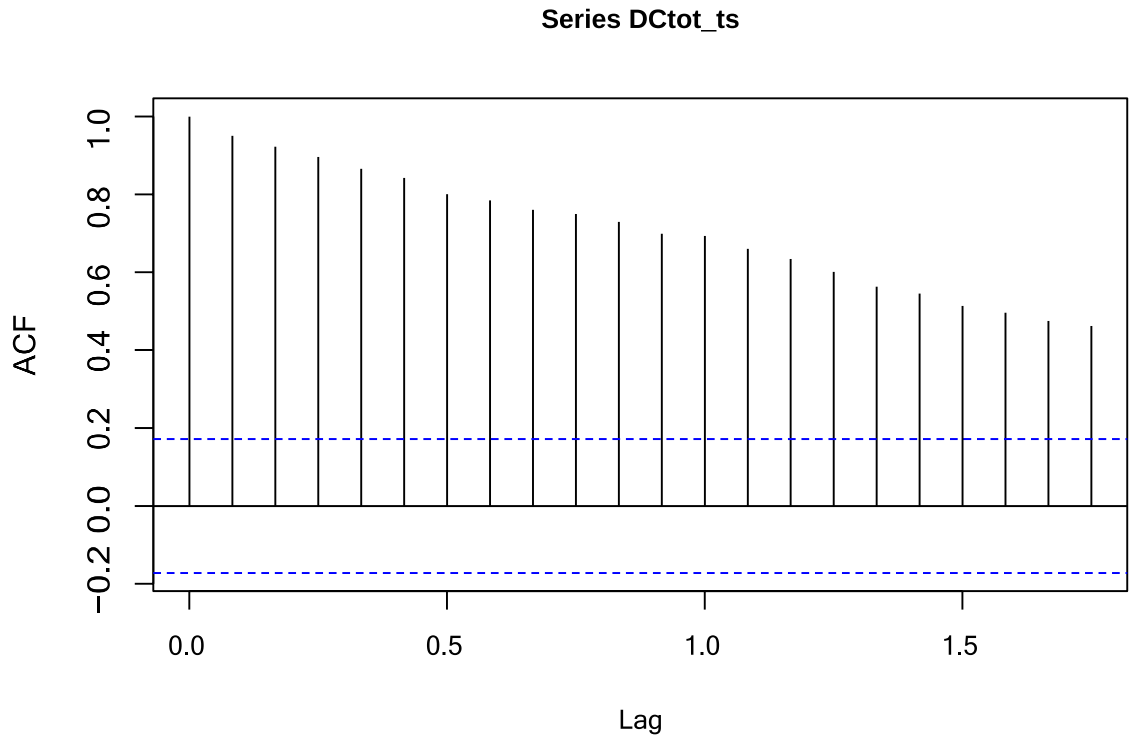
KPSS Level = 2.5743, Truncation lag parameter = 4, p-value = 0.01

KPSS Level = 1.9882, Truncation lag parameter = 4, p-value = 0.01

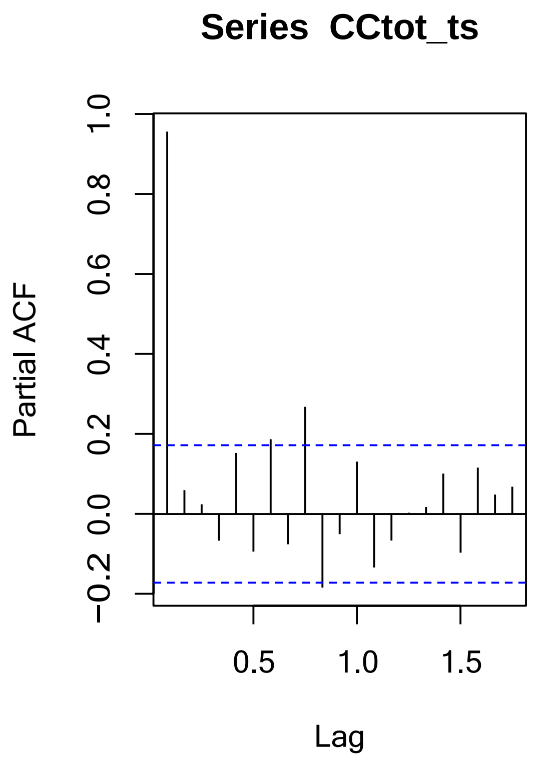
We can observe that the both the time series is non-stationary.
Now we proceed to plot the Partial Auto-correlation functions, and Auto- Correlation Function this will help us identify whether the time series has White Noise.

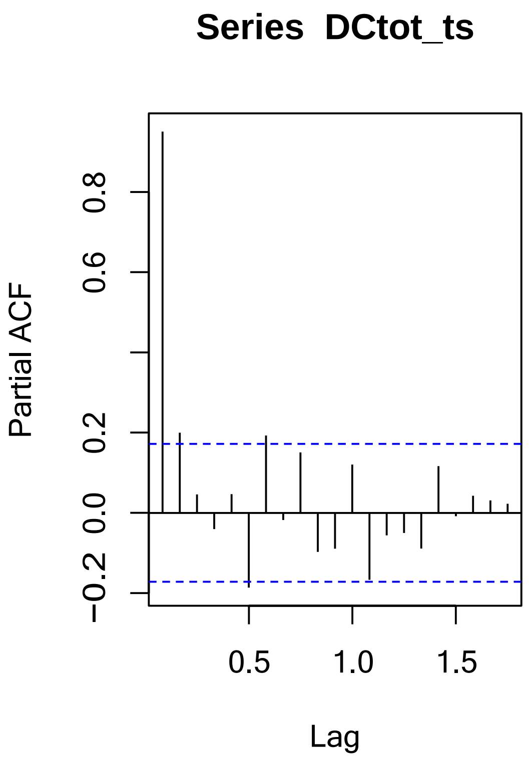
White Noise-A time series is white noise if the variables are independent and identically distributed with a mean of zero.





Interpretation- We can conclude from the plots that that we will get an AR(II) component in the model, however there is a chance that the model will be a mixed Model . We can also say that the time series is stationary





From both of the graphs we can observe that most of the points lie inside the Auto-correlation band. We now look to prove this by performing the *L-Jung Box test*

Ljung Box Test- The Ljung-Box test, named after statisticians Greta M. Ljung and George E.P. Box, is a statistical test that checks if autocorrelation exists in a time series.

The Ljung-Box test is used widely in econometrics and in other fields in which time series data is common.

H_0 : The residuals are independently distributed.

H_1 : The residuals are not independently distributed; they exhibit serial correla-tion.

```
Box-Ljung

test

X-squared = 1.7094, df = 5, p-value = 0.8877


Box-Ljung

test##

X-squared = 7.4687, df = 5, p-value = 0.188
```

From both the results we can accept H_0 at 5% LOS. Hence, we have proved that the time series have iid residuals.

ARIMA

Now, we proceed to fit the ARIMA model on the datasets, **ARIMA-**An auto-regressive integrated moving average, or ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predictfuture trends.

A statistical model is auto-regressive if it predicts future values based on past values. For example, an ARIMA model might seek to predict a stock's future

prices based on its past performance or forecast a company's earnings based on past periods.

ARIMA Parameters-

Each component in ARIMA functions as a parameter with a standard notation.For ARIMA models, a standard notation would be ARIMA with p, d, and q, whereinteger values substitute for the parameters to indicate the type of ARIMA model

used. The parameters can be defined as:

p: the number of lag observations in the model; also known as the lag order.

d: the number of times that the raw observations are differenced; also

known asthe degree of differencing.

q: the size of the moving average window; also known as the order of the

movingaverage.

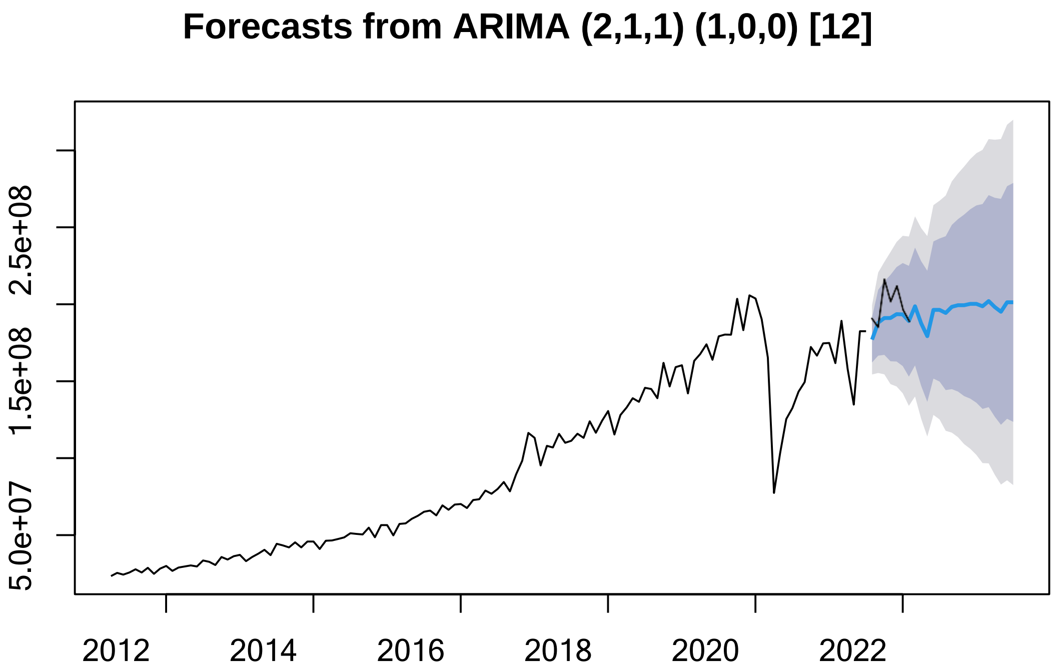
Just like in Holt-Winters Triple exponential smooting model, we will divide

thedata into 80% training and 20% test set.

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

Coefficients:

            ar1      ar2      ma1

        sar1-0.8735  -0.2777  0.9278

0.3573

s.e.      0.0949   0.0927  0.0391

0.0939

sigma^2 = 1.314e+14:  log likelihood = 2173.03

AIC=4356.07     AICc=4356.58   BIC=4370.13
```

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	1043549	11231532	6056698	0.7819785	6.289363	0.7891559	
	0.0071757						

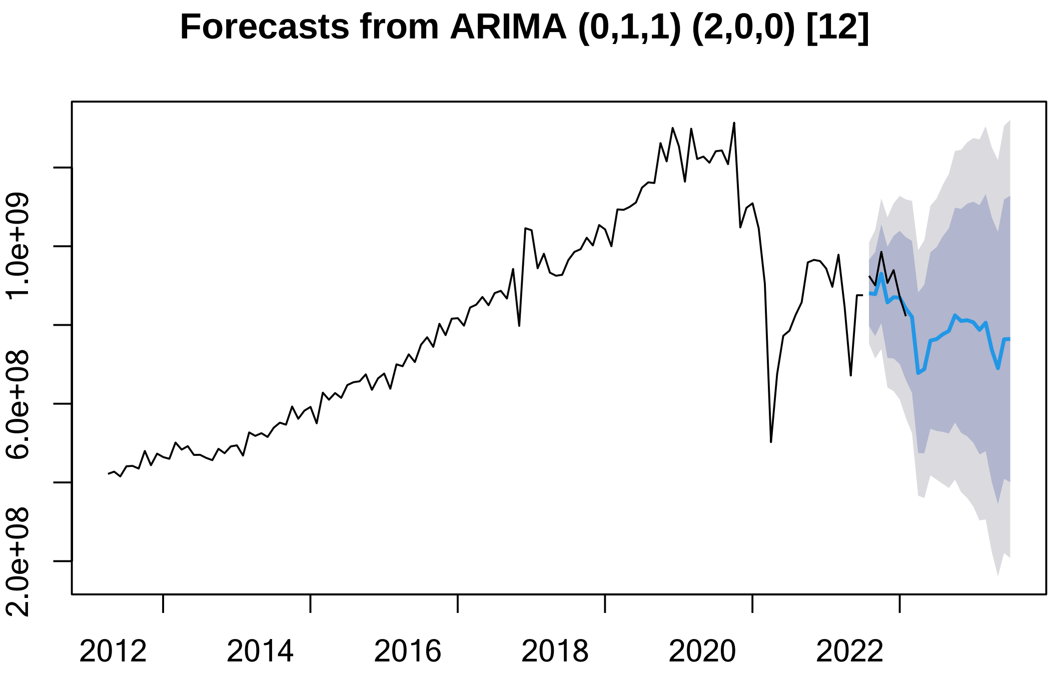
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Test set	9686806	13553477	10540955	4.6648044	5.125538	1.3734311	NA



Forecasts from ARIMA (2,1,1) (1,0,0) [12]

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


Coefficients:
           ma1     sar1     sar2
       -0.2201   0.2223   0.2796
s.e.    0.0981   0.0900   0.1243



sigma^2 = 4.313e+15:  log likelihood = -2388.64

AIC=4785.28   AICc=4785.62   BIC=4796.53
```

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	1334098	64606313	33550246	-0.1077849	4.349584	0.7739993	0.0211047
Test set	31836934	43182929	37314598	3.3617896	4.027755	0.8608424	NA



Now, from the graph we can say that our model is a good fit, because all of ourpredictions lie in the 95% confidence bands.

And, the accuracy of the plot can be measured using the RMSE value. The assumption which we made from the AR(II) plot is right.

CONCLUSION-

This paper started with Exploratory Data analysis of the primary data, where we can understand about the characteristics of that data. We used Pie charts, Multiple Bar plots, Histograms to give a visual impression of the data. This helped in identifying measures such as, the percentage of males and females, percentage of people having a particular job profile, percentage of fraud cases, etc.

Then we moved on to testing of hypothesis, where we initially test for independence between attributes such as Gender and Fraud, Job profile and fraud. We observed that these attributes are independent to Fraud. However, the attributes such as, Debit card frequency, Debit card limit are dependent on Fraud.

Then we used the technique of Logistic Regression for prediction of fraud, we observed that our model has an accuracy of about 90%.

Then we proceeded for Time Series analysis which is performed on data obtained from the RBI Website. We decomposed the time series, highlighted key events and tried to give an explanation about them. Apart from this, we prepared a Holt Winters prediction model for both the datasets, we observe that our predictions lie in the 95% prediction bands. Then we also test for stationarity, Ljung-box test on the time series, we observed that both the Time series are not stationary and both of them have independent residuals.

Then we fitted the ARIMA Model to the Time series. For this also, we found that our predictions lie in the 95% of the prediction bands.

REFERENCES

1. Non-Parametric Statistics-For the Behavioural Sciences- Sidney Siegul
2. https://www.investopedia.com/terms/
3. https://www.r-project.org/about.html
4. https://www.computerhope.com/jargon/e/excel.htm
5. https://rmarkdown.rstudio.com/
6. Introduction to Time Series and Forecasting - Douglas Montgomery
7. https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/
8. R. Markdown: The Definitive Guide - Garrett Grolemund, Joseph J. Allaire and Yihui Xie