Government of India
Ministry of Science & Technology
Department of Science & Technology
Innovation in Science Pursuit for Inspired Research (INSPIRE)

IVR Number: 201900004249

Name: Gunavant Prakash Thakare

-:Title of Project:-

Credit And Debit Card Fraud Detection Model Using Time Series

Format for **Research Project Report**

(To be filled by INSPIRE Scholar)

1. Cover Page (indicating title of project, name of scholar with IVR no):
2. Project Completion Certificate (in format as above):
3. Acknowledgement:
4. Title of Research Project:
5. Aim/Objectives:
6. Introduction of Research Topic:
7. Theoretical Framework of Topic:
8. Profile of Organization/Research Lab:
9. Methodology Followed:
10. Analysis and Interpretation of Research Project:
11. Conclusion & Suggestion of Research Project:
12. Result(s) Achieved:
13. References:
14. Declaration by the Scholar: I Gunavant Prakash Thakare (Full name) hereby declare that the details/facts mentioned above are true to the best of my knowledge and I solely be held responsible in case of any discrepancies found in the details mentioned above.

(Signature of Scholar)

Date: **10/01/2022**
Place: **Malegaon**

Note: 1. Kindly upload the Research Project Report as single PDF file whose size should not exceed 5 MB on Applicant's web-portal www.online-inspire.gov.in

2. Please do not send the hard copy of above document to DST.

****

1.Acknowledgement

Date: 10/01/2022
Place: Malegaon

I feel great pleasure in expressing my gratitude to my Professor **Ansari Mohammad Saeed** for their precious suggestions and guidance for the completion of my Mentorship project activity titled "**Credit And Debit Card Fraud Detection Using Time Series"**, I am really very thankful to them. Secondly I would also like to thank my parents and friends who helped me a lot in finalizing this project within the limitedtime frame.

Finally I would like to thank God, for letting me through all the difficulties. I have experienced your guidance day by day. You are the one who let me finish my Project. I will keep on trusting you for my future

Name Student
Gunavant Prakash Thakare

-: Title of Project :-

Credit And Debit Card Fraud Detection Using Time Series

OBJECTIVES

1. To analyse what percentage of Debit/Credit card users are more prone to Fraudulent transactions.
2. To analyse whether the usage of Credit/Debit cards has decreased since the onset of Unified Payments Service(UPI).
3. To fit a prediction model using logistic repression on the basis of the usage pattern for detection of fraud, plotting the ROC curve for checking the goodness of the model.
4. To fit a Time series on credit and debit card transactions per month. To identify seasonal component, trend component, irregular component.
5. To fit a Holt-Winters Model on the data and forecast the time series for the future.
6. To develop a forecasting model using ARIMA technique to predict a forecast for a given year using up the data of previous.
7. To compare which model is better at prediction using the RMSE values of the two predictions.

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 it later 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 transaction is 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 the American 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 our project we 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. We would also study about the usage of credit and debit card is over a period of time and develop a forecasting model using.

DATA COLLECTION

This project is performed using with two datasets.

The primary data collected through GOOGLE FORM and the data is collected for various aspects such as age, educational level, job profile, use of credit card , debit card , 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 the 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 **48** were fraud transactions.

The total observations recorded were **514** out of this close to **80%** were debit card users, **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 outstanding cards 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

R is a programming language for statistical computing and graphics supported by the R Core Team and the R Foundation for Statistical Computing. Created by statisticians Ross Ihaka and Robert Gentleman, R is used among data miners and statisticians for data analysis and developing statistical software.

R-Software was used to perform Exploratory Data Analysis, Time Series Analysis, Logistic Regression and Testing of Hypothesis.

Excel

Excel is a spreadsheet program from Microsoft and a component of its Office product group for business applications. Microsoft Excel enables users to format, organize and calculate data in a spreadsheet.

By organizing data using software like Excel, data analysts and other users can make information easier to view as data is added or changed. Excel contains a large number of boxes called cells that are ordered in rows and columns. Data is placed in these cells.

Excel was used mainly in sorting the data, different function from excel such as vlookup(), sort() were utilised. It was also used in compiling the Time Series data-set.

R-Markdown R Markdown is a flexible type of document that allows you to seamlessly combine executable R code, and its output, with text in a single document. These documents can be readily converted to multiple static and dynamic output formats, including PDF (.pdf), Word (.docx), and HTML (.html). The benefit of a well-prepared R Markdown document is full reproducibility. This also means that, you are able to add more data to your analysis, you will be able to recompile the report without making any changes in the actual document. The rmarkdown package comes pre-installed with RStudio, so no action is necessary.

MOTIVATION

In the past couple of decades the usage of credit/debit cards has seen a dramatic increase, this has majorly due to the following factors: Convenience, Real time transactions, robust security features. There are many people who find the usage of cards Convenient and tend to overspend, at the same time there are people who use cards as a tool for for controlling their spending.

Fraud is defined as a wrongful or criminal deception intended to result in financial or personal gain. Card frauds are very prevalent among Card users, this involves fraud by different means such as Skimming, Phishing, Hacking and making Fictitious Cards. The motivation of this project is to ensure that Card users should get maximum protection from Frauds. This is important because as technologies evolve, the Fraudsters also develop their techniques and come up with advanced techniques to steal our wealth.

Also, another important component of our project is to predict the Number of Transactions. As we witness the onset of UPI we can see people are preferring UPI transactions more than Cards primarily because of convenience, ease of usage, simple process. Further, due to COVID there has been a greater demand for Contactless payment methods, here UPIs transactions have really emerged as a robust method for Online transactions. However, people still perform card transactions for withdrawing Cash, paying at Petrol pumps and Online Shopping. Therefore, we will study the the Trend of Debit/Credit Card transactions from April 2011 to May 22 we will also try to provide reasoning some reasoning from the empirical data about the behavior of the Time Series.

EXPLORATORY DATA ANALYSIS

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

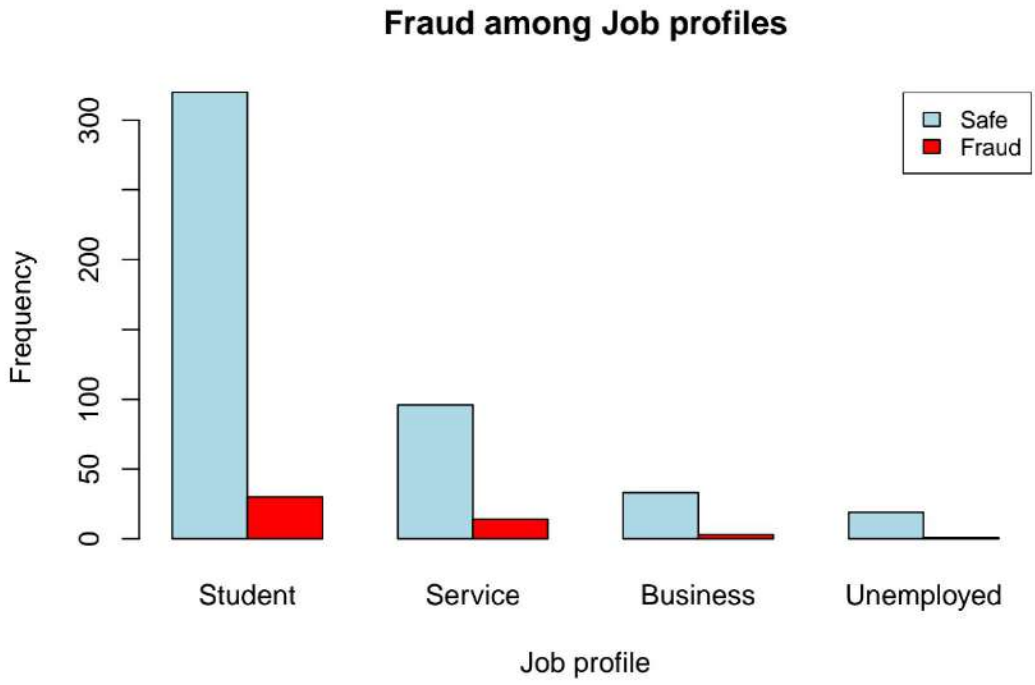
Histogram: A histogram is a bar graph-like representation of data that buckets a range of outcomes into columns along the x-axis. The y-axis represents the number count or percentage of occurrences in the data for each column and can be used to visualize data distributions.

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. Different parts 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.

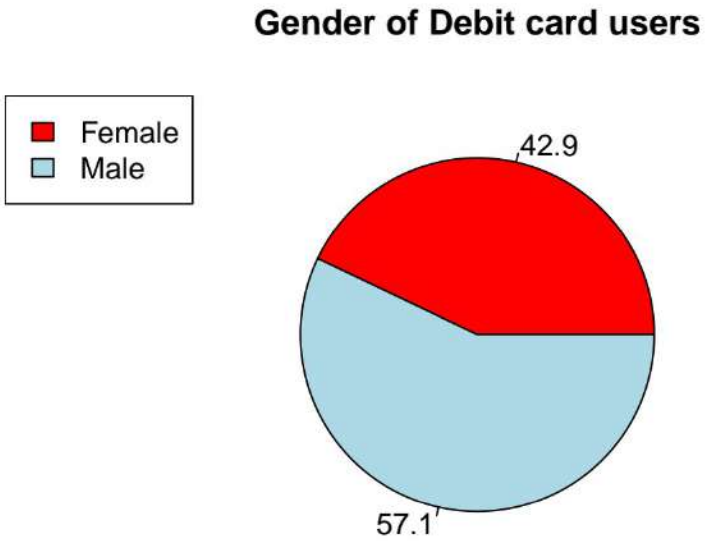
EXPLORATORY DATA ANALYSIS FOR THE PRIMARY DATA:

1. ***Plot for Fraud and Job profile***



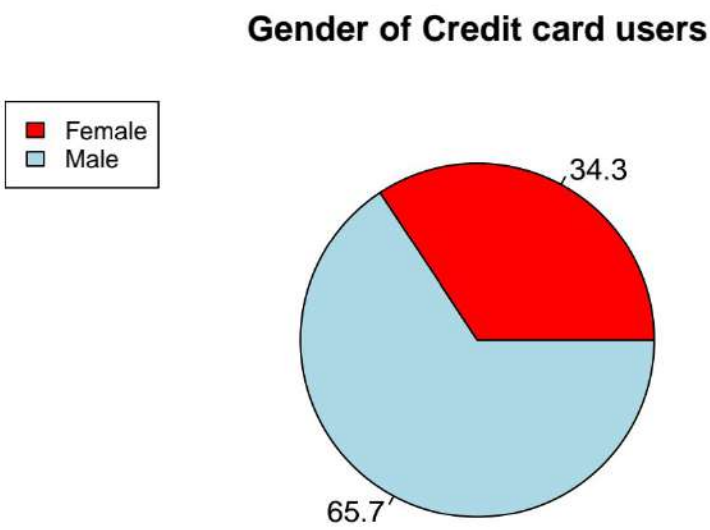
Interpretation-We can say that the maximum frauds which are faced by a category is Student.

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



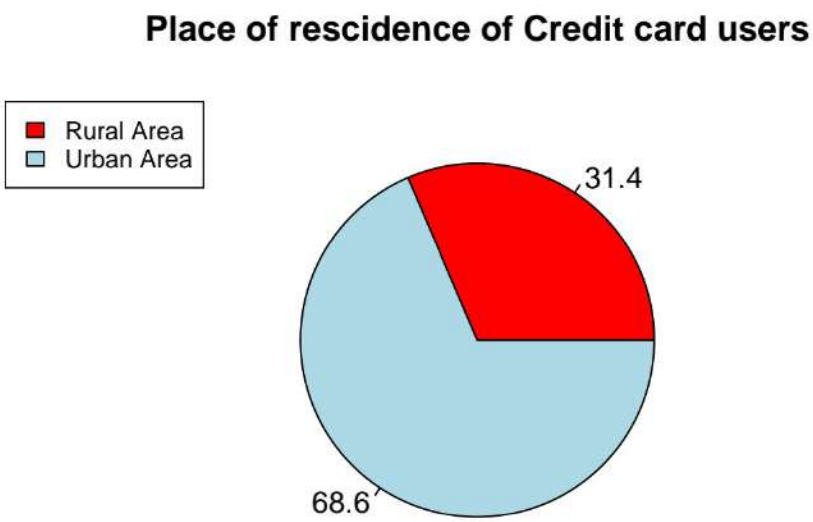
Conclusion- We can observe that 42.9% of the Debit card users in our data-set are Females and 57.1% are Males

3. **Pie Chart for Gender of Credit card users**



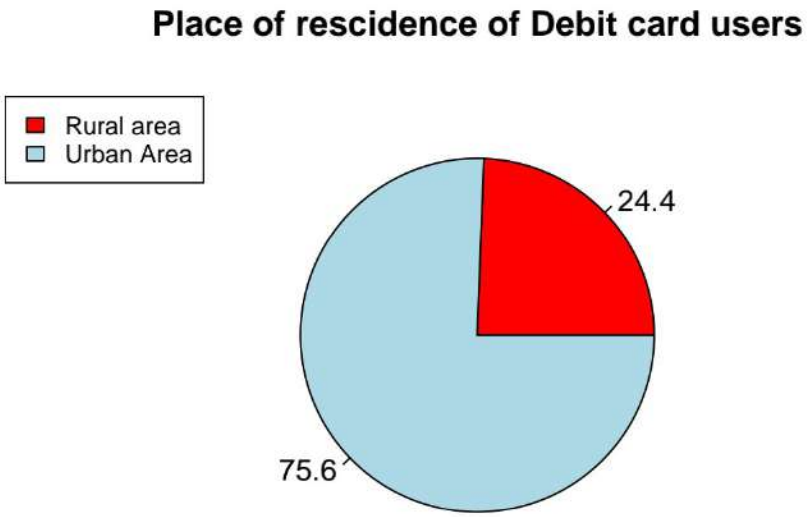
Conclusion- We can observe that 65.7% of the Credit card users in our data-set are Males and 34.1% are Females

4. **Pie chart for Area of residence for Credit card users**



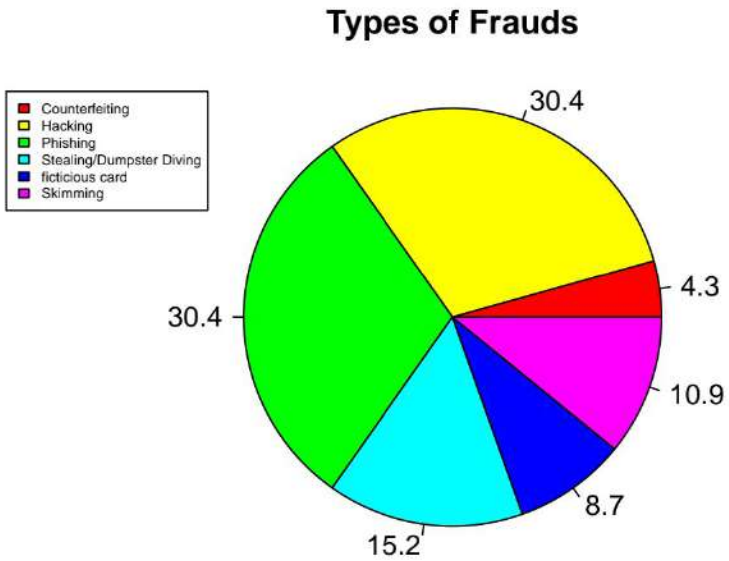
Conclusion- We can observe that 31.4% of the Credit card users in our data-set are residing in Rural Areas while 68.6 percent are residing in Urban Areas

5. **Pie chart for Area of residence of Debit card users**



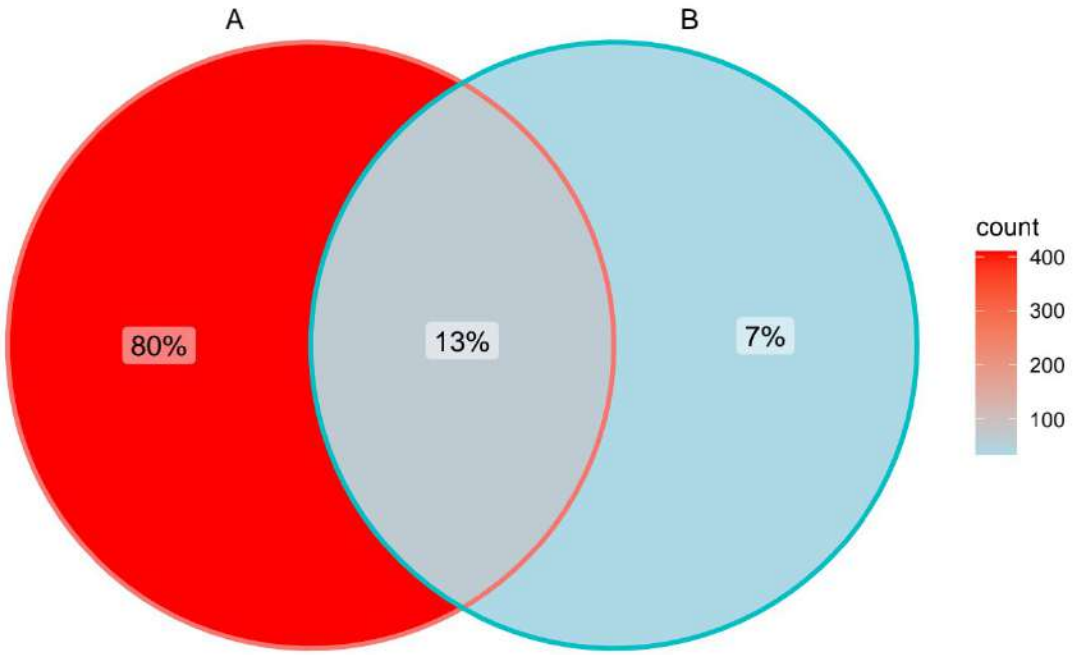
Conclusion: We can observe that 24.4% of the Credit card users in our data-set are residing in Rural Areas while 75.6 percent are residing in Urban Areas

6. **Pie Chart for types of frauds**



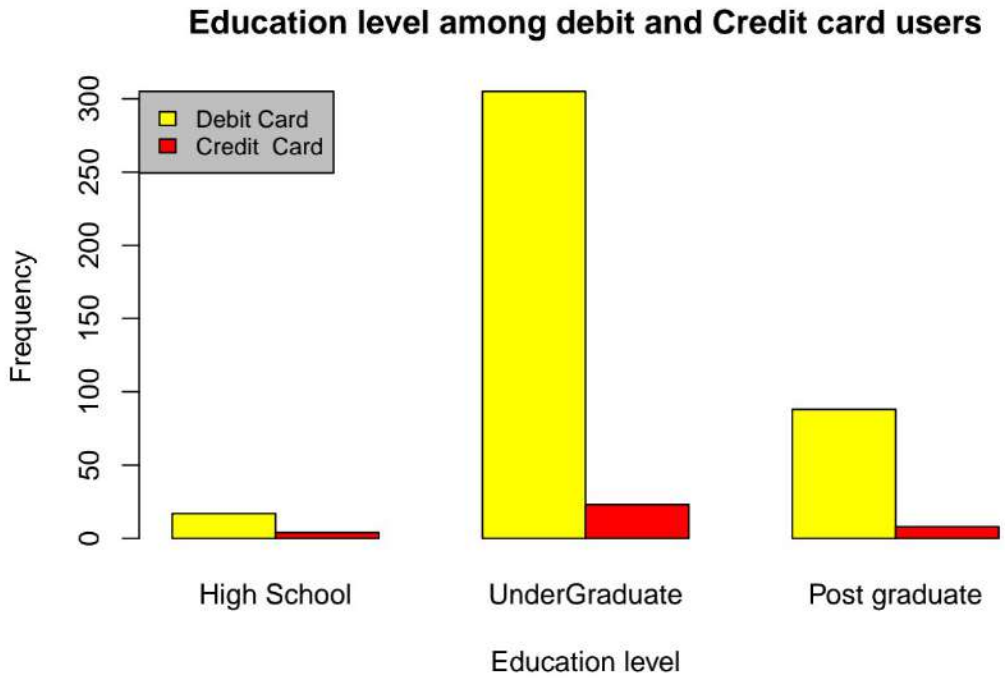
Conclusion: We can observe that the majority of the fraud type is Phishing and Hacking which account for 30.4% of the total frauds each, the next prevalent fraud type is Stealing/dumpster diving.

7. **Venn Diagram for Debit and Credit Card users**



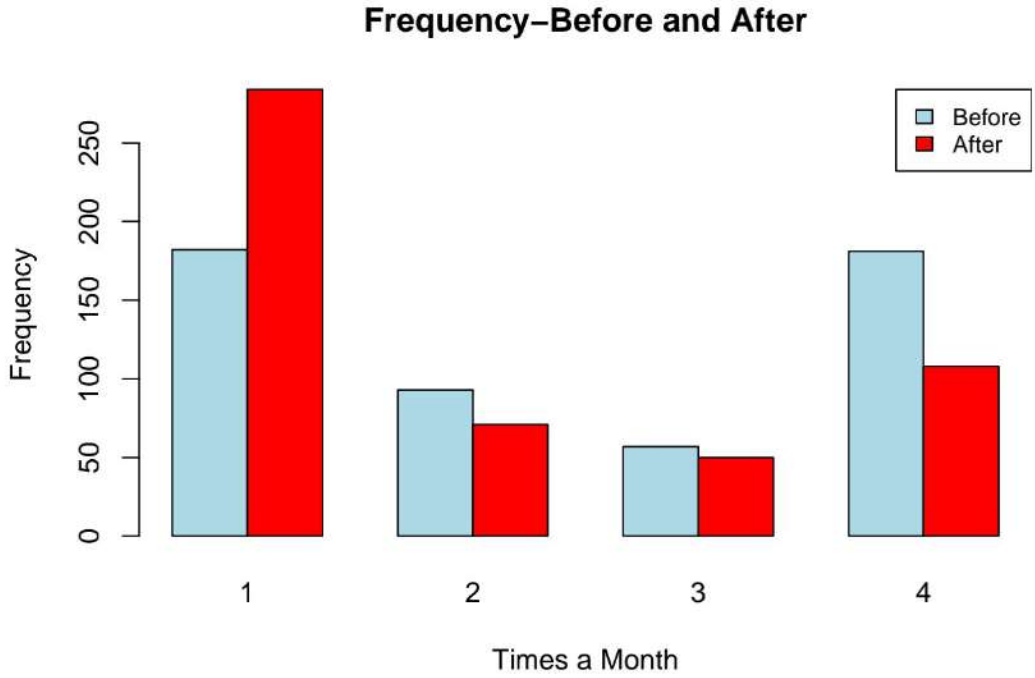
Conclusion: Here, We can observe that close to 80% of the people in our data-set use Debit cards. Whilst only 7% use a Credit Card. However, the percentage of people using both Debit and Credit Card is 13%.

8. **Education levels among Debit and Credit card users**



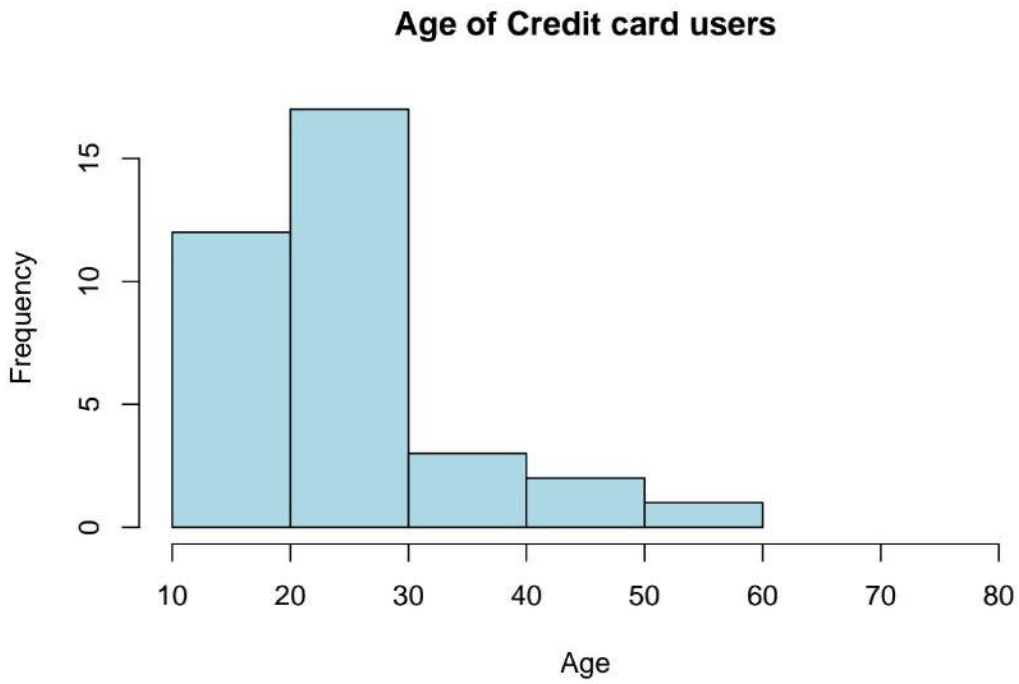
Conclusion: Here, We can observe that in both Credit and Debit card categories, the people with atmost undergraduate degree are using Debit/Credit card more. And most importantly, the penetration of of Credit/Debit card in the people having utmost High School qualification is the least.

9. **Multiple Bar Graph for Usage Frequency of Cards Before and After introduction of UPI-**



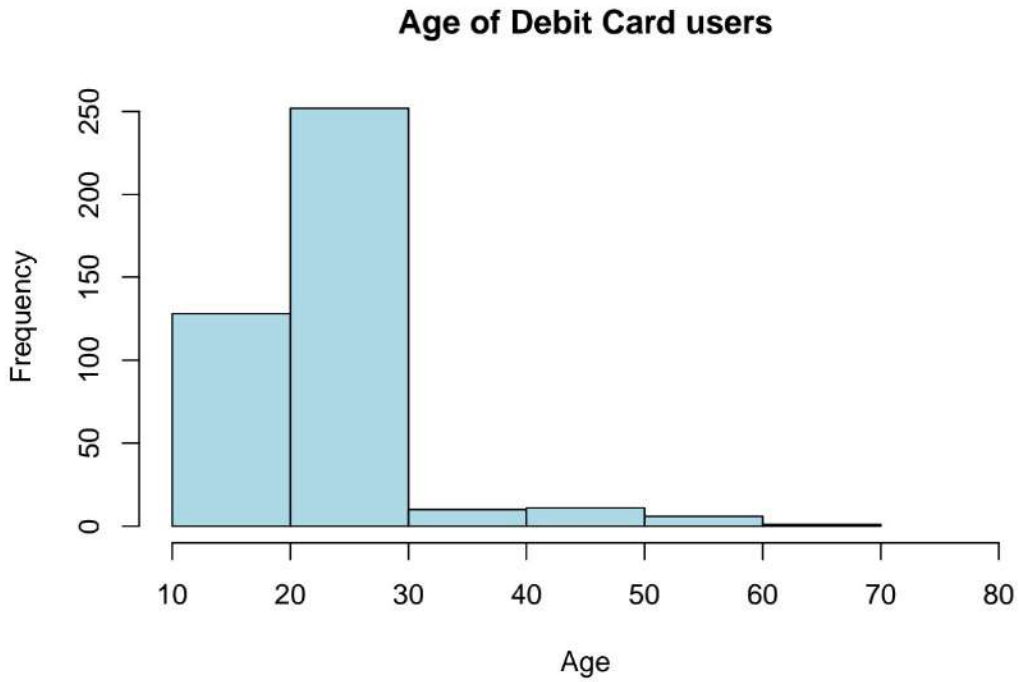
Conclusion- Here, we can observe that usage of Credit/Debit transactions has decreased since the introduction of UPI

10. **Histogram for the age of Credit Card users**



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

11. **Histogram for age of Debit Card users**



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

TESTING OF HYPOTHESIS

In this section, we perform the following tests on the primary data

1. Chi Square Test of Independence of Attributes
2. The Wilcoxon sign ranked test
3. Risk ratios

1.Chi Square test of Independence of attributes

i) **Association between Area of residence and fraud**

H_0 : The attributes are independent

H_1 : The attributes are not independent

Area/Fraud	0	1
Urban Area	358	33
Rural Area	108	14

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  conTable
## X-squared = 0.69708, df = 1, p-value = 0.4038
```

We observe that, $\chi^2_{calc} < \chi^2_{table}$ hence we accept H_0 at 5% LOS

ii) **Association between Gender and fraud**

H_0 : The attributes are independent.

H_1 : The attributes are not independent.

Gender/Fraud	0	1
Male	278	34
Female	188	13

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  conTable
## X-squared = 2.3747, df = 1, p-value = 0.1233
```

We observe that, $\chi^2_{calc} < \chi^2_{table}$ hence we accept H_0 at 5% LOS

iii) **Association between job profile and fraud**

H_0 : The attriutes are independent

H_1 : The attributes are not independent

Job Profile/Fraud	0	1
Business/Self employed	31	3
Private	93	14
Public	3	0
Student	320	30
Unemployed	19	0

```
##
```

```
##  Pearson's Chi-squared test
##
## data:  conTable
## X-squared = 4.3481, df = 4, p-value = 0.3609
```

We observe that, $\chi^2_{calc} < \chi^2_{table}$ hence we accept H_0 at 5% LOS

iv) **Association between Monthly Credit Card expense and Fraud**

H_0 : The attriutes are independent

H_1 : The attributes are not independent

Monthly Credit Card Expense/Fraud	0	1
0-10k	69	7
10k-20k	26	5
20k-30k	10	5
30k-40k	10	2
40k-50k	5	1
50k and above	2	0
Do not use a Credit card	344	27

```
##
##  Pearson's Chi-squared test
##
## data:  conTable
## X-squared = 15.341, df = 6, p-value = 0.01776
```

We observe that, $\chi^2_{calc} > \chi^2_{table}$, hence we REJECT H_0 at 5% LOS

v) **Association between Monthly Debit Card expense and Fraud**

H_0 : The attributes are independent

H_1 : The attributes are not independent

List of users

Debit Card Limit/Fraud	0	1
0-10k	308	20
10k-20k	72	11
20k-30k	24	9
30k-40k	16	2
40k-50k	6	0
50k and above	6	0
I do not use a Debit card	34	4

```
##
##  Pearson's Chi-squared test
##
## data:  conTable
## X-squared = 19.369, df = 6, p-value = 0.003584
```

We observe that, $\chi^2_{calc} > \chi^2_{table}$, hence we REJECT H_0 at 5% LOS

vi) **Association between Debit card usage frequency and Fraud**

H_0 : The attriutes are independent

H_1 : The attributes are not independent

List of users

Debit Card Frequency/Fraud	0	1
Very less	49	4
Less	115	5
Sometimes	149	15
Often	76	9
Very often	59	13

```
##
##  Pearson's Chi-squared test
##
## data:  conTable
## X-squared = 10.645, df = 4, p-value = 0.03086
```

We observe that, $\chi^2_{calc} > \chi^2_{table}$, hence we REJECT H_0 at 5% LOS

Further we can also observe that the Odds of fraud associated with Very Often is the highest $13/59 = 0.22$

2. Wilcoxon Sign ranked test

Wilcoxon signed rank test is used to compare two related samples, matched samples, or to conduct a paired difference test of repeated measurements on a single sample to assess whether their population mean ranks differ. It is used to compare two sets of scores that come from the same participants. This can occur when we wish to investigate any change in scores from one time point to another, or when individuals are subjected to more than one condition.

$W=\sum_{i=1}^{N_r}[\text{sgn}(x_{2,i}-x_{1,i})\cdot R_i]$

Here we first perform the **Shapiro- Willik Test of Normality** on both of the variables

H_0 : The variable is Normally distributed

H_1 : The variable is not Normally distributed

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.78238, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  y
## W = 0.71891, p-value < 2.2e-16
```

We observe that, $W_{calc} > W_{table}$, hence we REJECT H_0 at 5% LOS.

The variables are not Normally distributed and hence we go for Non-parametric test.

H_0 : The Debit/Credit Card usage per month was the same before and after

introduction of UPI.

H_1 : The Debit/Credit Card usage per month was more before the introduction of UPI than after the introduction of UPI.

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  x and y
## V = 27670, p-value = 1.735e-13
## alternative hypothesis: true location shift is greater than 0
```

We observe that, $V_{calc} > V_{table}$, hence we REJECT H_0 at 5% LOS. Hence we can say that Usage of Debit/Credit card was **more** before introduction of UPI.

3. Risk Ratios

Let us have a brief look at Risk Ratios

This can be said to be a ratio of the probabilities of risk in one group compared to the possibilities of an occurrence of risk in another group. It is commonly taken into use to present the outcomes of various groups. These are also termed as a relative risk

Risk Ratio- $R = \frac{A/(A+B)}{C/(C+D)}$

Risk Ratio for Gender and Fraud-

Gender/Fraud	0	1
Male	278	34
Female	188	13

Risk Ratio=$\frac{34/(34+278)}{13/(13+188)} = 1.684911$

Conclusion- As RR>1 we can say that Males are 1.684911 times as prone to Credit/ Debit Card Fraud as Females.

Risk Ratio for Urban and Rural Areas-

List of users

Area/Fraud	0	1
Urban Area	358	33
Rural Area	108	14

Risk Ratio=$\frac{33/(33+358)}{14/(14+108)} = 0.7354768$

Conclusion- As RR>1 we can say that people living in Urban Areas are 0.7354768 times as prone to Credit/Debit Card Fraud as people living in Rural

Areas.

LOGISTIC REGRESSION

Logistic regression is used to predict the class (or category) of individuals based on one 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 given above would utilize binary logistic regression. Other examples of binary responses could include passing or failing a test, responding yes or no on a survey, and having high 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 $\varepsilon \sim \mathsf{B}(\pi(x))$

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

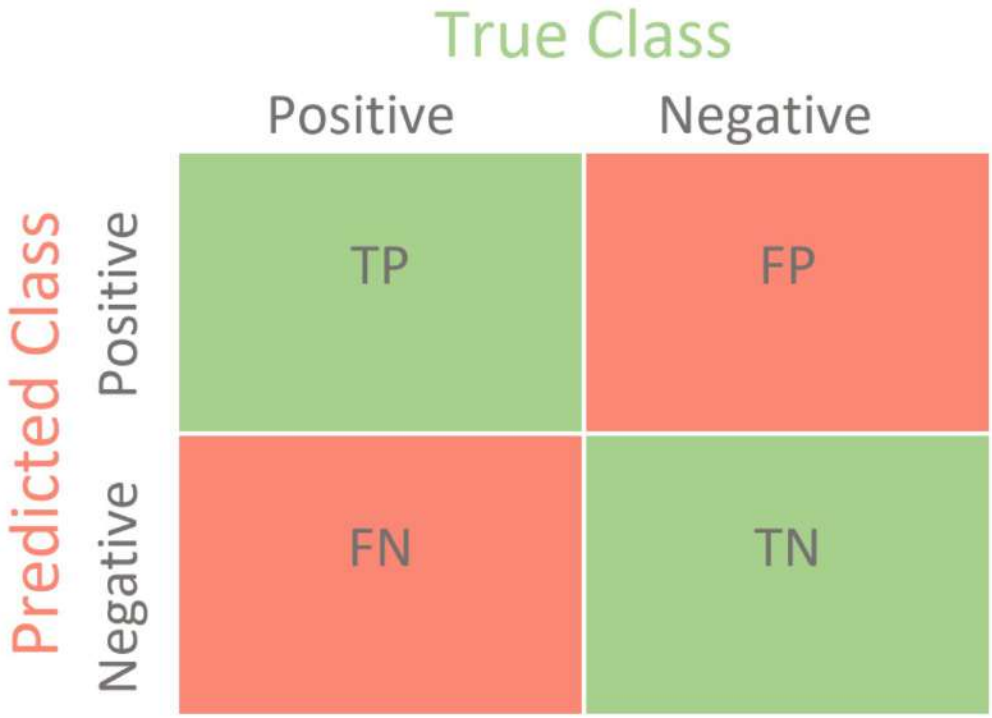
We are performing Logistic regression on the Debit Card users data-set to obtain a prediction model for FRAUD with the following regressors and response variable.

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
AREA	X9

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**



A **Confusion Matrix** is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

The following are some of the Terminologies of Confusion Matrix-

1. *True Positives (TP):* These are cases in which we predicted yes, and they have encountered fraud.
2. *True Negatives (TN):* We predicted no, and they don't have encountered fraud.
3. *False Positives (FP):* We predicted yes, but they have not encountered fraud. (Also known as a *"Type I error."*)
4. *False Negatives (FN):* We predicted no, but they actually have encountered fraud. (Also known as a *"Type II error."*)
5. *Accuracy:* Overall, how often is the classifier correct? $(TP + TN)/TOTAL$
6. *Mis-classification Rate:* Overall, how often is it wrong? $(FP + FN)/TOTAL$
7. *Mathhew's Correlation coefficient:* Matthews correlation coefficient (MCC)

is a metric we can use to assess the performance of a classification model. 8. *ROC Curve:* An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate, False Positive Rate

Null deviance	Residual Deviance	AIC
170.016	122.274	178.2744

	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
genderMale	1.9204556	0.6859269	2.7997962	0.0051135
educationPost Graduate	0.9510107	1.3443669	0.7074042	0.4793154
educationUndergraduate	-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
incomebelow 3.5L	-1.2906616	1.3594752	-0.9493823	0.3424262
jobPrivate	0.6211563	1.3598079	0.4567971	0.6478169
jobPublic	-17.4694146	7283.2385243	-0.0023986	0.9980862
jobStudent	2.0770289	1.3031134	1.5938972	0.1109591
jobUnemployed	-17.2038913	2299.7799357	-0.0074807	0.9940313
AreaUrban 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

	Estimate	Std. Error	z value	Pr(>|z|)
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_limitI 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_expenseDon't use a debit card	0.8510787	1.1552122	0.7367293	0.4612870

Note that-

1. Gender is a factor with two levels, but glm() is showing us only one coefficients associated with Gender. This is because of the term dummy variable trap. To see the effect of k categories you only need (k$ -$1) dummies.

2.From the above result we can conclude that $(NullDeviace - ResidualDeviance) > \chi^2_{9;0.05}$ 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 card limit, Debit card expense, Debit card usage Frequency, Job, Gender.

3. We perform the **Wald's Test** to check (2) $H_0 : \beta_i = 0$ vs. $H_1 : \beta_i \neq 0$ Now, we proceed to obtain the confusion matrix and get the Accuracy of our predictions

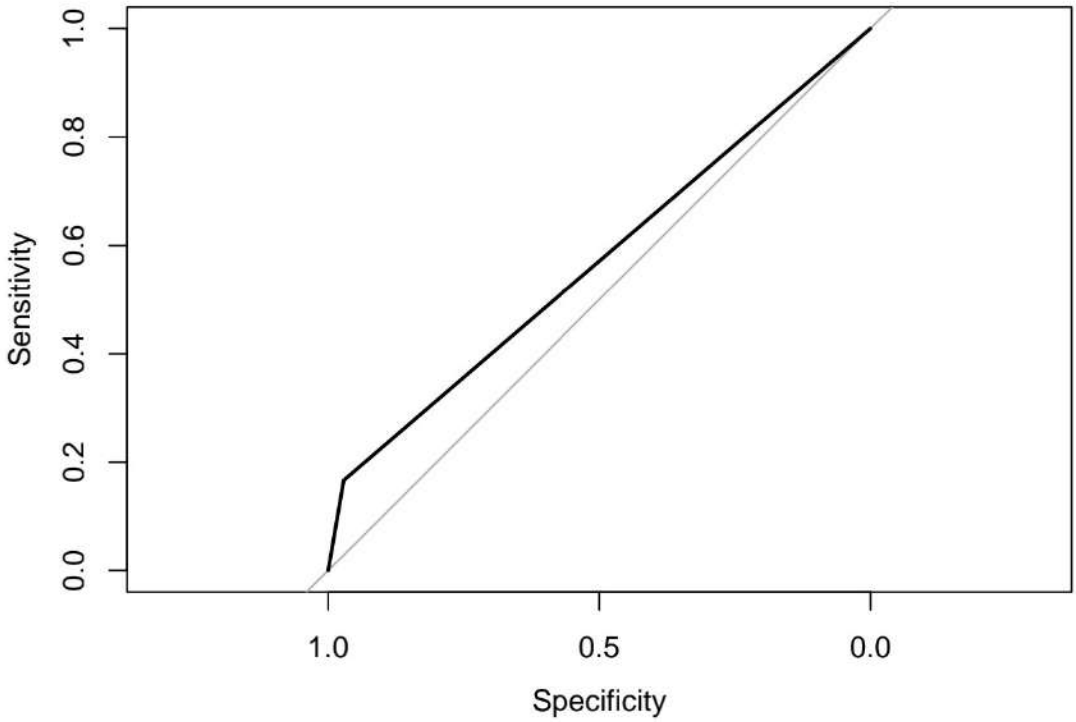
```
## New names:
## * `` -> ...1
## * `` -> ...2
## * `` -> ...3
```

```
## * `` -> ...4
## * `` -> ...5
## * ...
```

Prediction/Reference	0	1
0	68	5
1	2	1

	Metrics	Values
	Accuracy	0.9079
	95% CI	(0.8094 , 0.9622)
	Sensitivity	0.9714
	Specificity	0.1667
	AUC value	0.56

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```



Here we are considering the **ROC curve-**An ROC curve (Receiver Operating Characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate, False positive rate. **AUC Curve** AUC is an effective way to summarize the overall diagnostic accuracy of the test. It takes values from 0 to 1, where a value of 0 indicates a perfectly inaccurate test and a value of 1 reflects a perfectly accurate test. A value of 0.5 for AUC indicates that the ROC curve will fall on the diagonal (i.e., 45 degree line) and hence suggests that the diagnostic test has no discriminatory ability. The AUC value obtained by us is **0.56** which means that our model has a very slight discriminatory ability. Further, we obtain the value of **Mathew's Correlation coefficient**=0.19124 which implies that our prediction model is better than random method of classification.

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

TIME SERIES ANALYSIS

1. Decomposing the Time Series

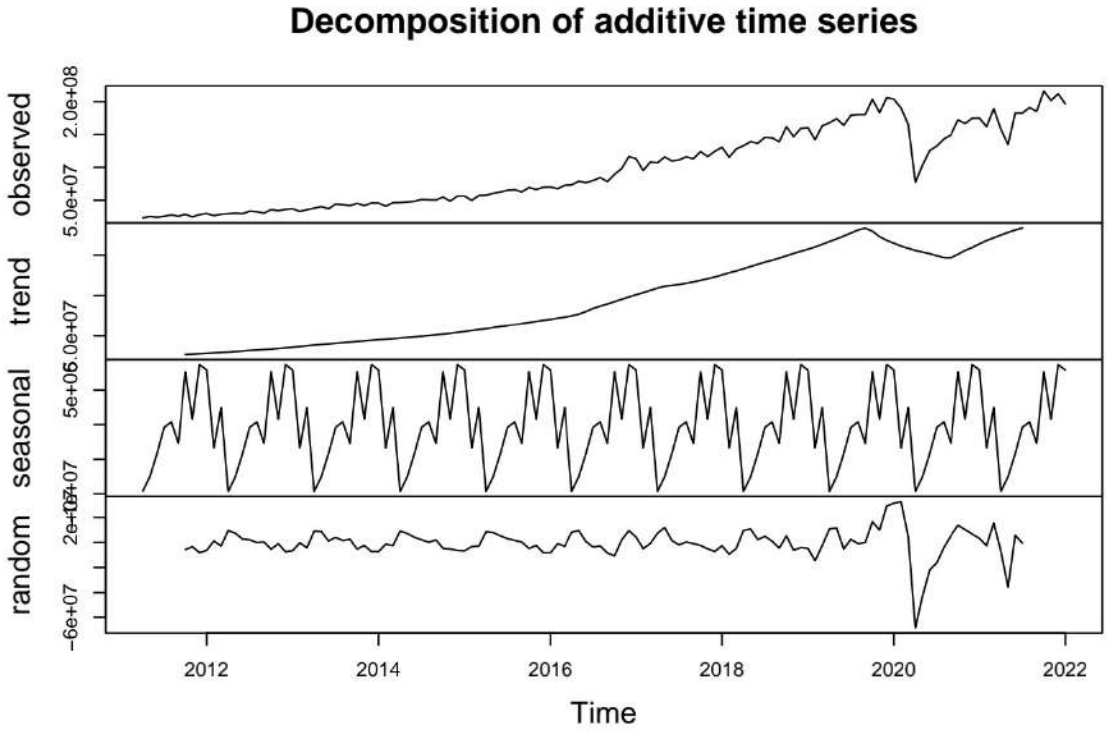
Time series arise as recordings of processes which vary over time. Time series plot displays the values of the process output in the order in which the values occur. A recording can either be a continuous trace or a set of discrete observations. We will concentrate on the case where observations are made at discrete equally spaced times. By appropriate choice of origin and scale we can take the observation times to be 1, 2, . . . T and we can denote the observations by Y1, Y2, .. . , YT .A key analyzing a time series is to understand the form of any underlying pattern of 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.

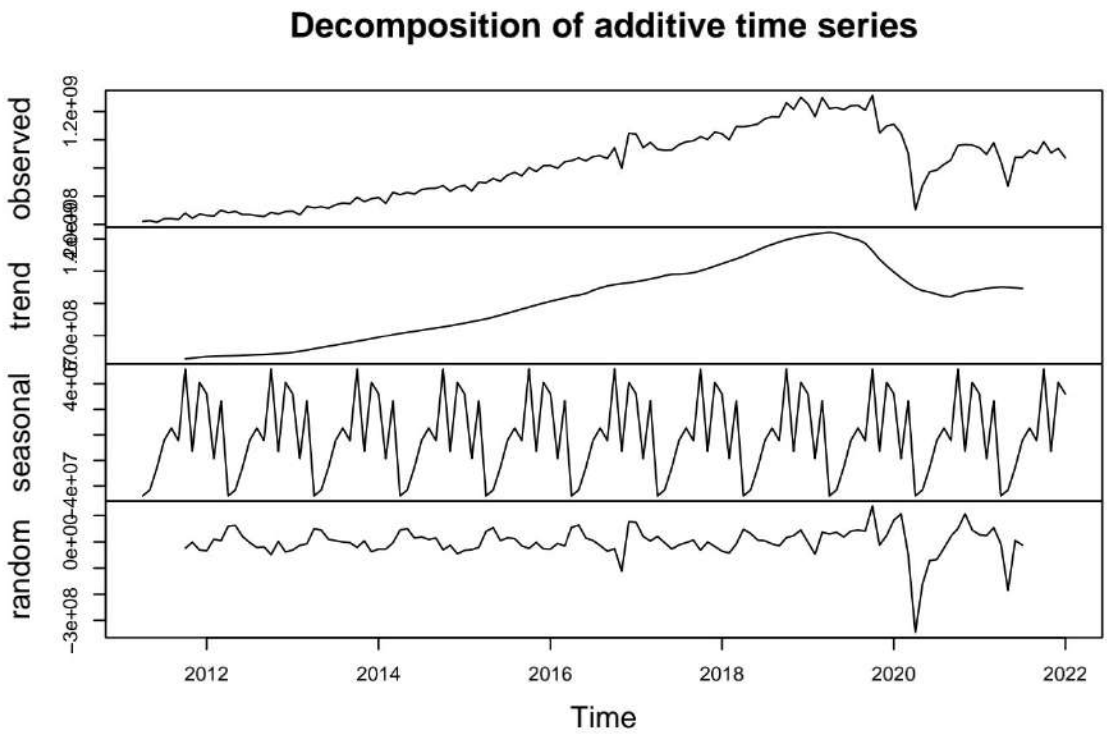
Trend: Long-term, gradual increasing, decreasing or stagnating tendency of the variable Y.

Seasonality: Regular, relatively short-term (yearly) repetitive up and down fluctuations of the variable Y depending on the season. **Cyclical Component:** A gradual, long-term, up and down potentially irregular swings of the variable Y.

Random Component: A random increase or decrease of the variable Y for a specific 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 the Time Series for Debit Cards, Credit cards





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 maybe 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 *Nationwide Lock-down imposed attributed to the COVID-19 out-break.*

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

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

Moving averages-Moving Averages gives a trend with a fair degree of accuracy, in this method we take the Arithmetic mean of observations for a certain span of time. The formula is given as follows for 3 Monthly MA-

$\hat{Y} = \frac{\sum_{i=t-1}^{t+1} Y_i}{3}$

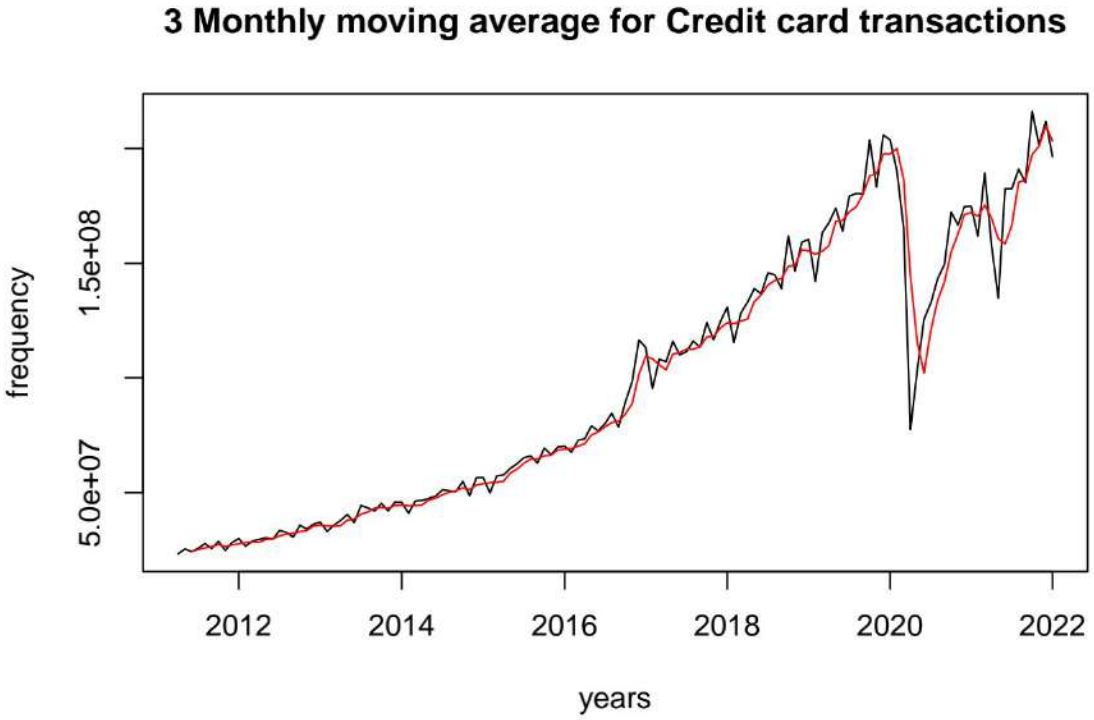
Demerits- 1. In non-periodic data, this method is less effective.

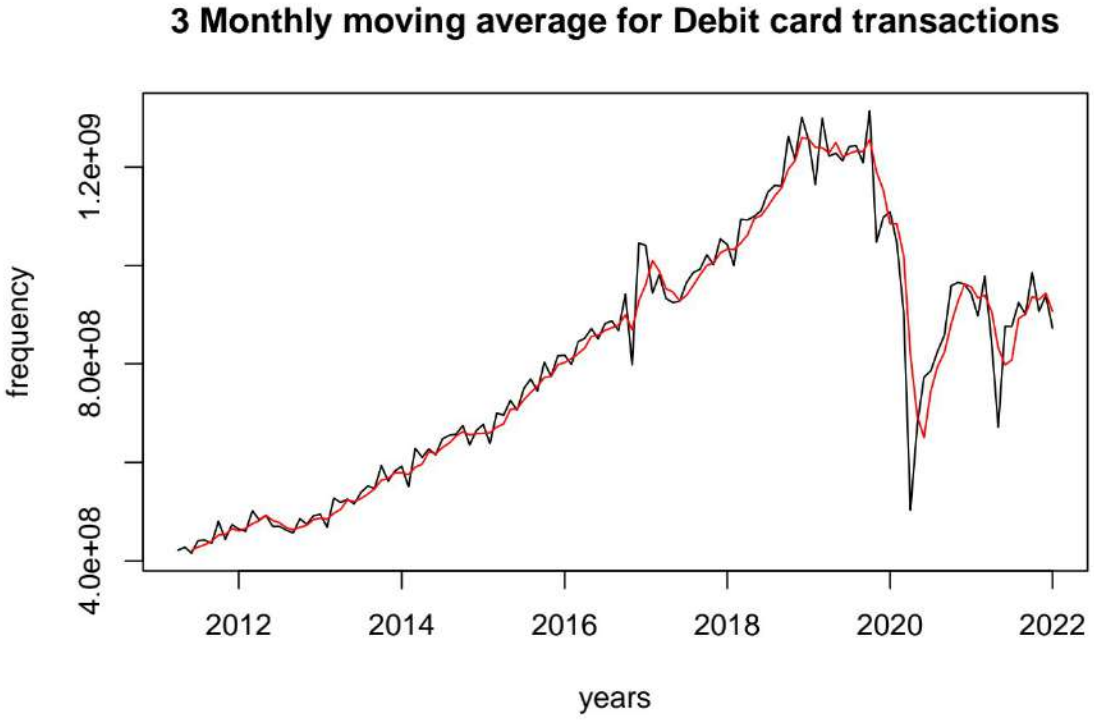
2. Selection of proper 'period' fo computing moving average is difficult.

3. Values of first few months and last few months is not found.

4. We can't predict any values because selection of k is subjective.

The following is the plot for SMAs, one can observe that the seasonal components are smoothened and an increasing trend is visible.





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

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

2. 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: stationary 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 of a 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.*

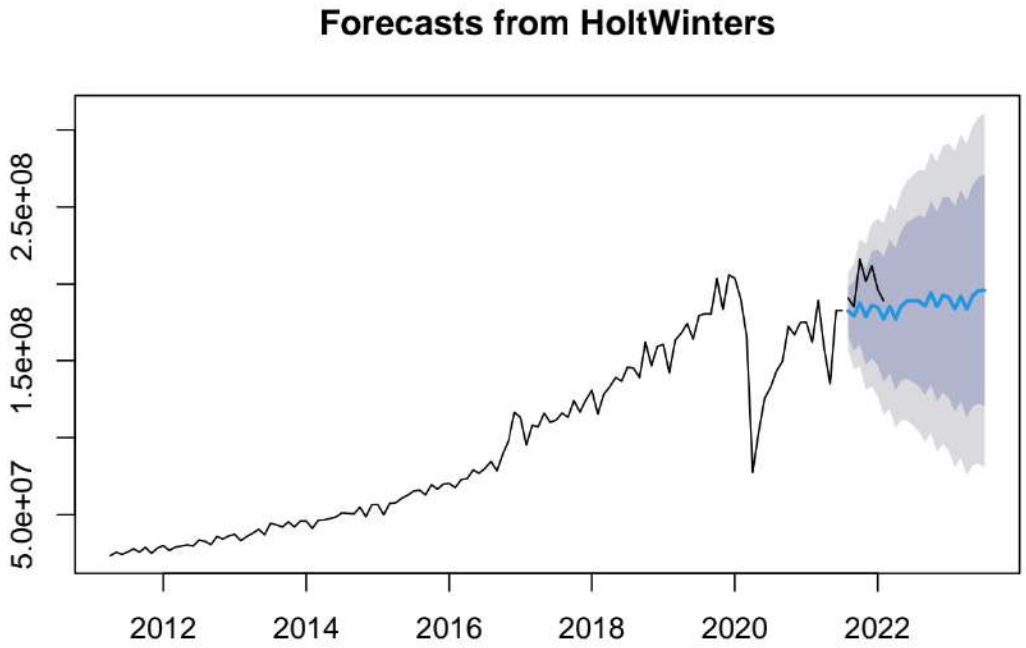
	CC	DC
alpha	0.9244575	0.8947077
beta	0.0000000	0.0000000
gamma	1.0000000	0.3572679

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

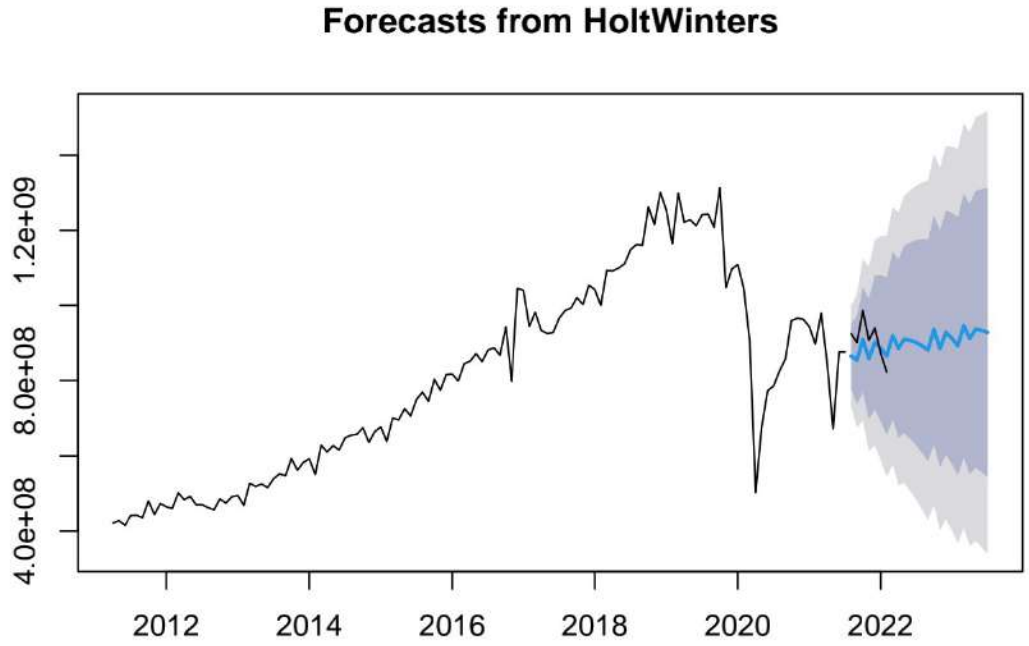
And in the case of Debit cards transactions, we observe that α=0.8947077, β=0 and γ=0.3572679.

Now, we will check for the accuracy for both of our Holt-Winters 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



	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



In this plot we can observe that the observed values lie in the 90% confidence bands of our predictions for both Crdit an 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.

3. Stationarity and ACF, PACF plots

We will now check the stationarity of both the Time Series, before that, let us recall **Stationarity of a Time series-**

In the most intuitive sense, stationarity means that the statistical properties of a process 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.

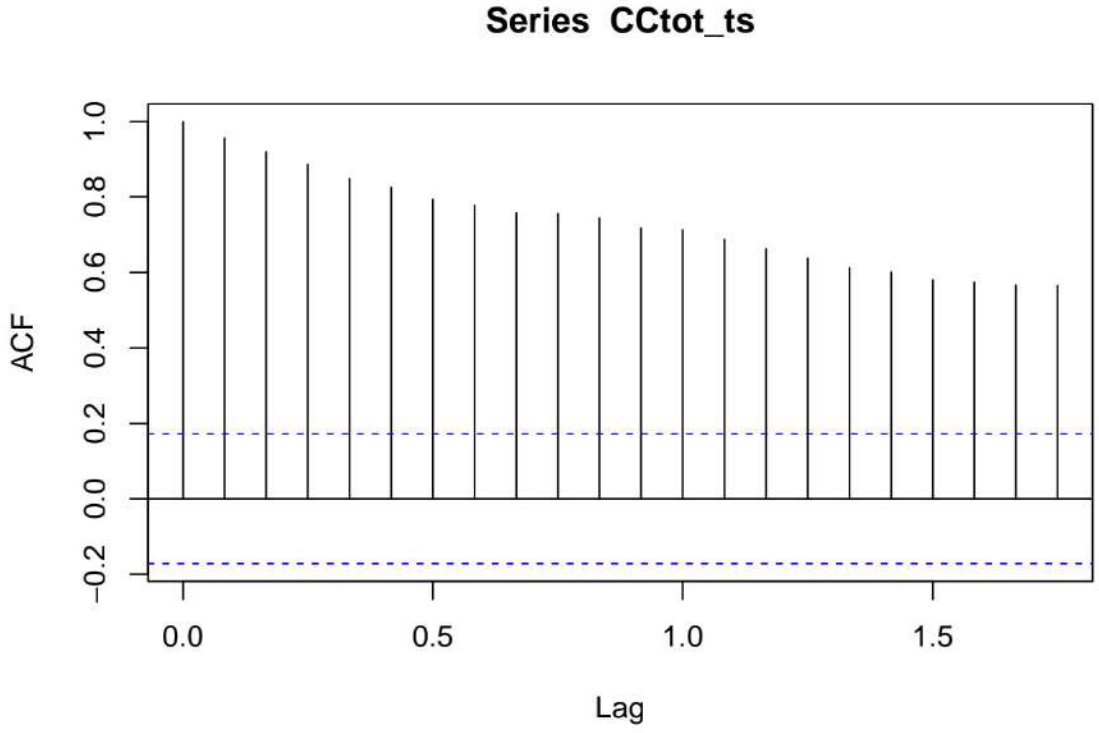
##

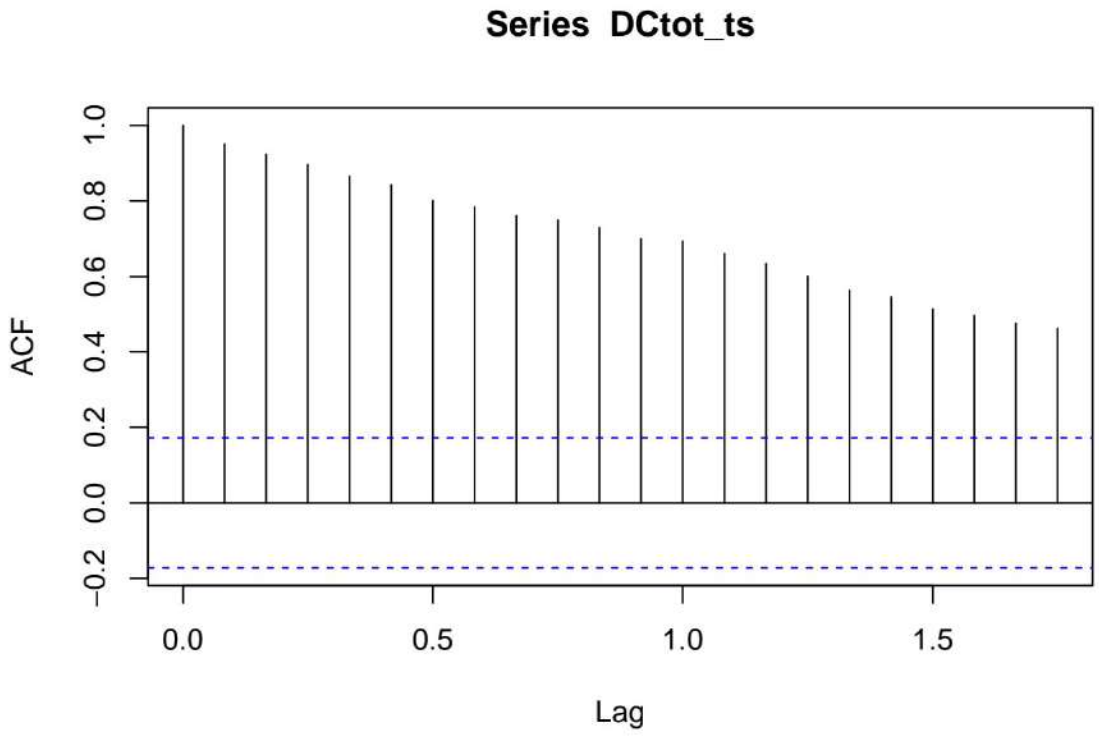
```
##  KPSS Test for Level Stationarity
##
## data:  CCtot_ts
## KPSS Level = 2.5743, Truncation lag parameter = 4, p-value = 0.01

##
##  KPSS Test for Level Stationarity
##
## data:  DCtot_ts
## KPSS Level = 1.9882, Truncation lag parameter = 4, p-value = 0.01
```

We can observe that the both the time series is non-stationary.

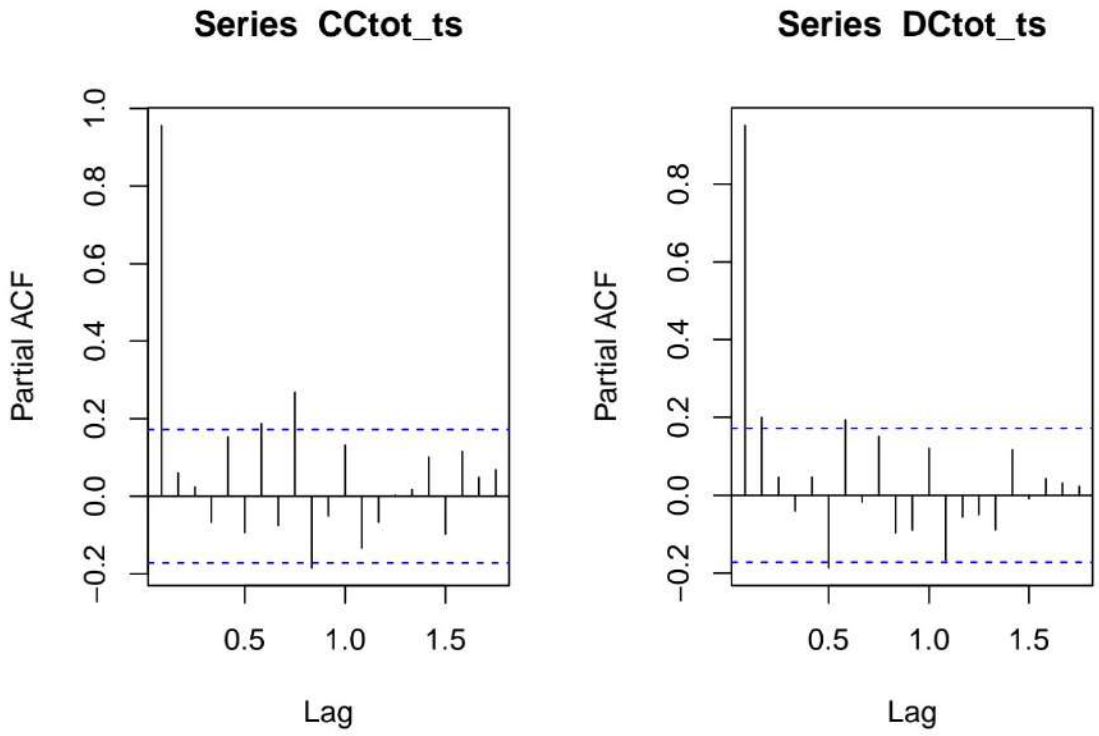
Now we proceed to plot the Partial Autocorrelation functions, and Auto Correlation Function this will help us identify whether the time series has White Noise.

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 non stationary.



From both of the graphs we can observe that most of the points lie inside the Autocorrelation 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 ACFs are insignificant.

H_1 : The residuals ACFs are significant.

Test Statistic-

$$Q_{LB} = n(n+2)\sum_{j=1}^{m} \frac{r_j^2}{n-j}$$

```
##
##  Box-Ljung test
```

```
##
## data:  mymodel$resid
## X-squared = 1.7094, df = 5, p-value = 0.8877

##
##  Box-Ljung test
##
## data:  mymodel$resid
## 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, i.e The residual AutoCorrelations are insignificant.

4. Arima

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

A statistical model is autoregressive 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, where integer 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 as

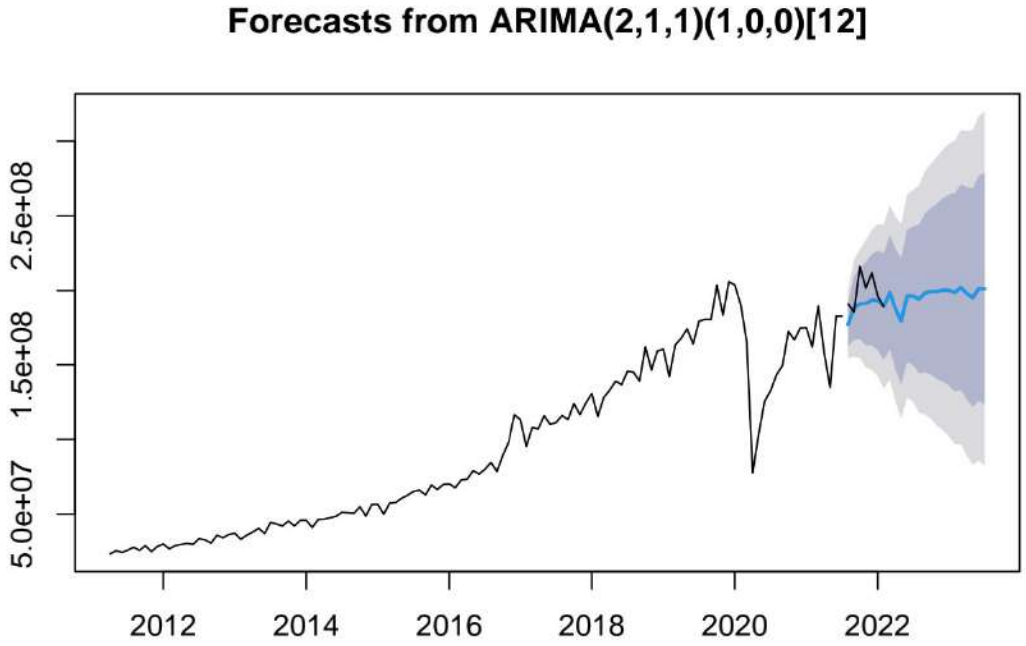
the degree of differencing.

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

Just like in Holt-Winters Triple exponential smoothing model, we will divide the data into 80% training and 20% test set.

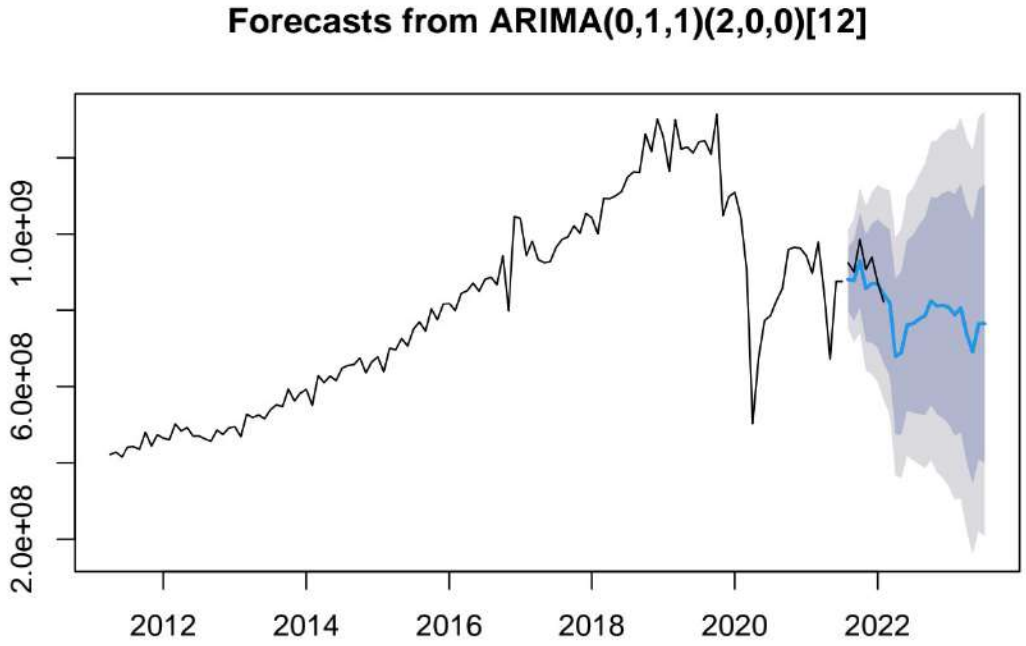
```
## Series: CCtot_ts1
## ARIMA(2,1,1)(1,0,0)[12]
##
## Coefficients:
##           ar1      ar2     ma1    sar1
##       -0.8735  -0.2777  0.9278  0.3573
## s.e.   0.0949   0.0927  0.0391  0.0939
##
## sigma^2 = 1.314e+14:  log likelihood = -2173.03
## AIC=4356.07   AICc=4356.58   BIC=4370.13
```

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	1043549	11231532	6056698	0.7819785	6.289363	0.7891559	0.0071757
Test set	9686806	13553477	10540955	4.6648044	5.125538	1.3734311	NA




```
## Series: DCtot_ts1
## 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 our predictions lie in the 95% confidence bands. Here, we can observe that Credit card ARIMA Model has coefficients (2,1,1)(1,0,0)[12] and for Debit cards it has coefficients (0,1,1)(2,0,0)[12]. Further, this is an example of SARIMA Model

Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component. It adds three new hyperparameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

From the Holt-Winters and SARIMA Model, we can conclude from the both the models that the SARIMA model is a better fit to the data, because the RMSEs are smaller for that than the corresponding RMSEs of Holt winters model.

The in-depth study of this model is out of the scope of this project. It is just used for prediction purposes.

CONCLUSION-

This project begins with Exploratory Data analysis of the Primary data, where we we can understand about the characteristics of that the data. We use Pie charts, Multiple Barplots, Histograms to give a visual impression of the data. This helps us 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 move on to testing of hypothesis, where we initially test for Independence between attributes such as Gender and Fraud, Job profile and fraud. We observe that these attributes are independent to Fraud. However, the attributes such as, Debit card frequency, Debit card limit are dependent on Fraud. Further we use the Wilcoxon sign ranked test to test whether Debit/Credit card usage has decreased after the introduction of UPI payment. And then we find risk ratios to highlight which subgroup in our sample are more prone to card frauds. We get the result that people living in Rural areas are more prone to fraud than Urban Areas, and that Males are more prone to fraud than Women. The we use the technique of Logistic Regression for prediction of fraud, we observe that our model has an accuracy of about 90%. However, after plotting the ROC curve we observe that the Area under the curve is 0.56, which means that our prediction model is just slightly better than random allocation techique.

Then we proceed for Time Series analysis which is performed on data obtained from the RBI Website. We decompose the time series, highlight key events and try to give an explanation about these events. Apart from this, we prepare 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 can observe that both the Time series are not stationary and and both of them have independent residuals.

Then we fit the ARIMA Model to the Time series, we observe that the model is

a SARIMA Model and here also we find that our predictions lie in the 95% of the prediction bands. Finally from the RMSEs of both Holt-Winter's and SARIMA model we can say that SARIMA Model is much better because of the lower RMSE value.

LIMITATIONS AND SCOPE-

1. The **sample size 514** which we are using for analysis is relatively small as the size of total debit and credit card users in India are **985.32 million**
2. This means that that there is a chance that the results might be skewed
3. The prediction model is a fundamental model for Frauds, there are other models using Machine learning techniques like ANN, KNN, Naive Bayes algorithm which can be performed.
4. There are far less parameters for Fraud detection model considered in the model, Models such as Bank account number, Debit/Credit card number, IFSC number can be very significant for fraud detection
5. There is ample scope for Prediction model for Frauds for Credit/Debit Cards, this is very important because Card frauds are very frequent.

REFERENCES

1. Non Parametric Statistics-FOR THE BEHAVIORAL 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