

---

# Credit Card Data Report

21-12-2020

---

By

Sarah Merin John , 20BDA06

## Introduction

In today's world we are on an express train to a cashless society. Credit cards are an excellent tool for making payments in a cashless society. However, banks do not approve all credit card applications. They consider multiple factors while deciding whether to approve card applications such as credit history ,income , expenditure , age , owner of the house etc.

Using this Credit Card dataset we can understand how each of these variables affect one in being approved for getting a Credit Card.

## Context

Cross-section data on the credit history for a sample of applicants for a type of credit card.

## Description of Data

Format A data frame containing 1,319 observations on 15 variables.

The contents of individual variables are as follows :

- **card** : Factor. Was the application for a credit card accepted?
  - **cardN** : Numerical value of card in binary
  - **reports** : Number of major derogatory reports
  - **age** : Age in years plus twelfths of a year.
  - **income** : Yearly income (in USD 10,000).
-

- **USD** : Income after conversion (in USD 10,000).
- **share** : Ratio of monthly credit card expenditure to yearly income.
- **expenditure** : Average monthly credit card expenditure
- **owner** : Factor. Does the individual own their home?
- **ownerN** : Numerical value of owner in binary
- **selfemp** : Factor. Is the individual self-employed?
- **dependents** : Number of dependents.
- **months** : Months living at current address.
- **majorcards** : Number of major credit cards held.
- **active** : Number of active credit accounts.

## Notes

According to Greene (2003, p. 952) **dependents** equals **1 + number of dependents**

## Acknowledgments

This dataset was originally published alongside the 5th edition of William Greene's book *Econometric Analysis*.

## Output

### 1) Importing Data

```
Datafinal= read.csv("C:/Users/SARAH/Desktop/CreditCard1.csv")
```

```
head(Datafinal,n=5)
```

Serial	card	cardN	reports	age	income	USD	share	expenditure	selfemp	dependents	months	majorcards	active	ownerN	owner
1	yes	1	0	38	4.5200	45200	0.033269910	124.983300	0	3	54	1	12	1	yes
2	yes	1	0	33	2.4200	24200	0.005216942	9.854167	0	3	34	1	13	0	no
3	yes	1	0	34	4.5000	45000	0.004155556	15.000000	0	4	58	1	5	1	yes
4	yes	1	0	31	2.5400	25400	0.065213780	137.869200	0	0	25	1	7	0	no
5	yes	1	0	32	9.7867	97867	0.067050590	546.503300	0	2	64	1	5	1	yes

## Exploratory Data Analysis

### 2) Missing Data

Checking if any missing value is there in this dataset . And we can see , there is no missing value in this dataset .

```
sum(is.na(Datafinal))
```

```
0
```

### 3) Calculating Descriptive Statistics using `summary()`

The descriptive statistics help determine the distribution of numerical variables. In this table, we can find a simple statistical report for each attribute. For instance, the maximum age is 84 .

```
summary(Datafinal)
```

Serial	card	cardN	reports	age
Min. : 1.0	no : 296	Min. : 0.0000	Min. : 0.0000	Min. : 0.00
1st Qu.: 330.5	yes:1023	1st Qu.: 1.0000	1st Qu.: 0.0000	1st Qu.: 25.00
Median : 660.0		Median : 1.0000	Median : 0.0000	Median : 31.00
Mean : 660.0		Mean : 0.7756	Mean : 0.4564	Mean : 33.25
3rd Qu.: 989.5		3rd Qu.: 1.0000	3rd Qu.: 0.0000	3rd Qu.: 39.00
Max. : 1319.0		Max. : 1.0000	Max. : 14.0000	Max. : 84.00
income	USD	share	expenditure	
Min. : 0.210	Min. : 2100	Min. : 0.0001091	Min. : 0.000	
1st Qu.: 2.244	1st Qu.: 22438	1st Qu.: 0.0023159	1st Qu.: 4.583	
Median : 2.900	Median : 29000	Median : 0.0388272	Median : 101.298	
Mean : 3.365	Mean : 33654	Mean : 0.0687322	Mean : 185.057	
3rd Qu.: 4.000	3rd Qu.: 40000	3rd Qu.: 0.0936168	3rd Qu.: 249.036	
Max. : 13.500	Max. : 135000	Max. : 0.9063205	Max. : 3099.505	
selfemp	dependents	months	majorcards	
Min. : 0.00000	Min. : 0.0000	Min. : 0.00	Min. : 0.0000	
1st Qu.: 0.00000	1st Qu.: 0.0000	1st Qu.: 12.00	1st Qu.: 1.0000	
Median : 0.00000	Median : 1.0000	Median : 30.00	Median : 1.0000	
Mean : 0.06899	Mean : 0.9939	Mean : 55.27	Mean : 0.8173	
3rd Qu.: 0.00000	3rd Qu.: 2.0000	3rd Qu.: 72.00	3rd Qu.: 1.0000	
Max. : 1.00000	Max. : 6.0000	Max. : 540.00	Max. : 1.0000	
active	ownerN	owner		
Min. : 0.000	Min. : 0.0000	no : 738		
1st Qu.: 2.000	1st Qu.: 0.0000	yes: 581		
Median : 6.000	Median : 0.0000			
Mean : 6.997	Mean : 0.4405			
3rd Qu.: 11.000	3rd Qu.: 1.0000			
Max. : 46.000	Max. : 1.0000			

### 4) Data Distributions

The below list shows the median values for "Age", "Income", and "Share" respectively. The median is the value at the center of the distribution, therefore 50% of the observations in the distribution will have values above the median and 50% will have values below.

For example : The median of age here is age here is 31 , so half the observations lie above the age of 31 and half of the observations lie below the age of 31.

```

a=median(Datafinal$age)
i=median(Datafinal$income)
s=median(Datafinal$share)
print(a)
print(i)
print(s)

```

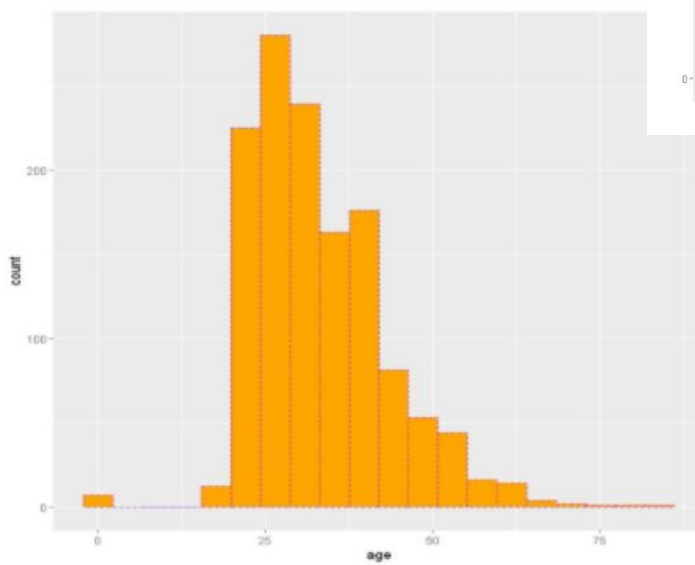
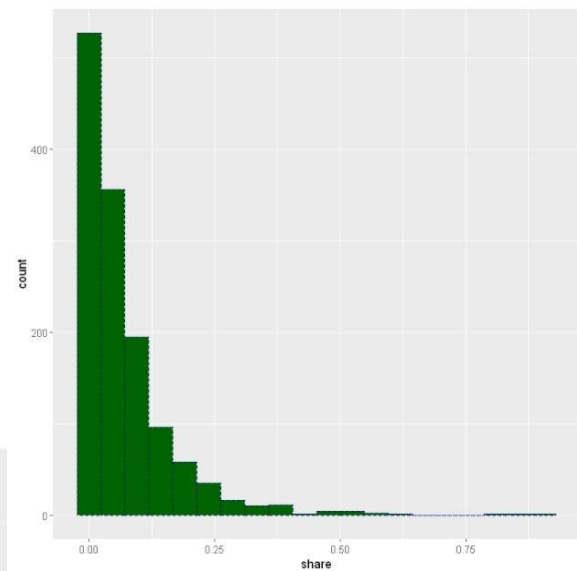
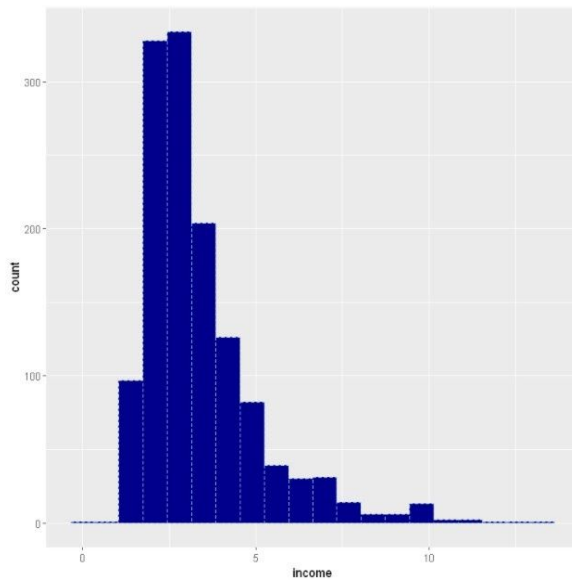
```

[1] 31
[1] 2.9
[1] 0.03882722

```

## 5) Normalized histograms (Distribution)

- ❖ income Distribution (Navy Blue)
- ❖ age Distribution (Orange)
- ❖ share Distribution (Dark Green)



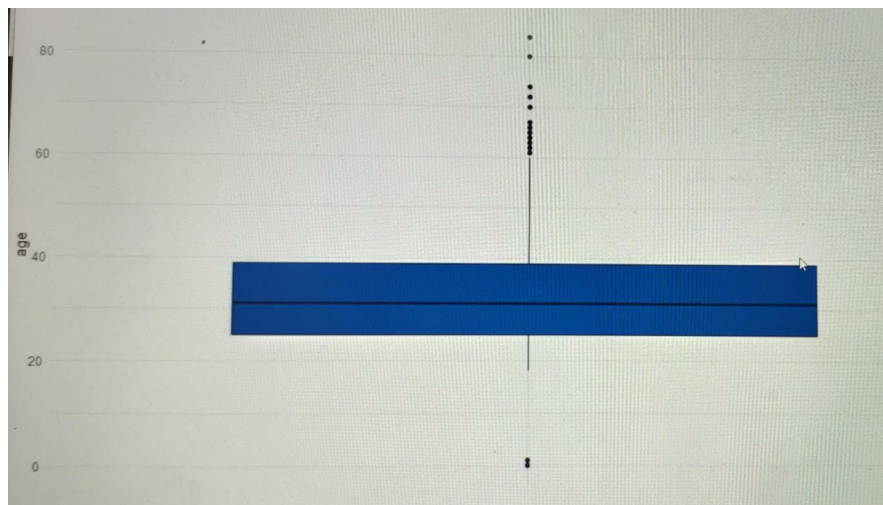
The above histograms shows the frequency distribution for income , age and share that have applied for a credit card.

The **income histogram** is a right skewed histogram . A distribution skewed to the right is also called a positively skewed histogram. We can see here the peak is higher in the range of 2 to 3. This shows that most of the people who have applied for a credit card have an income range of 2 to 3 (20,000 to 30,000 dollars).

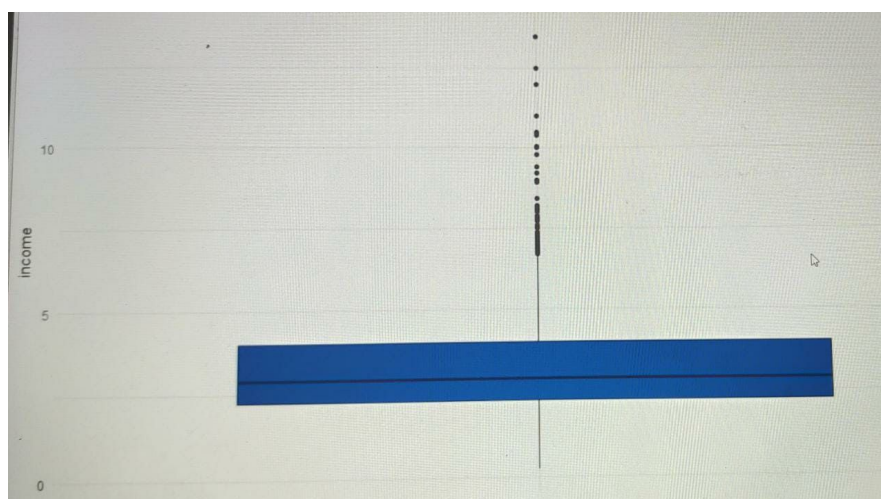
The **age histogram** is a right skewed histogram. We can understand from here that the age group from 20 to 35 have applied for the credit card the most , which shows that the young adult is applying most for credit cards.

The frequency distribution for **share** is also a perfectly right skewed distribution

## 6) Outliers



Using variable " age "



**Using variable " income "**

The main part of the box plot shows where the middle portion of the data is: the interquartile range. At the ends of the box, you'll find the first quartile (the 25% mark) and the third quartile (the 75% mark).

Using this we can understand the outlier values. Only the data that lies within Lower and upper limit are statistically considered normal and thus can be used for further observation or study. In the case of the box plot of age the lower limit is 25 and upper limit is 39. So any value that will be more than the upper limit or lesser than the lower limit will be the outliers. Here we can see that above the age of 60 all the values are outliers and can be discarded from the entire series so that analysis made on this series is not influenced by these extreme values.

Similarly in the box plot of income , the values from above 5.5 are extreme values.

## 7) Chi Square test

```
> chisq.test(table(Datafinal1$card,Datafinal1$owner), correct = FALSE)

Pearson's Chi-squared test

data:  table(Datafinal1$card, Datafinal1$owner)
X-squared = 28.823, df = 1, p-value = 7.929e-08

> summary(table(Datafinal1$card,Datafinal1$owner))
Number of cases in table: 1319
Number of factors: 2
Test for independence of all factors:
    Chisq = 28.823, df = 1, p-value = 7.929e-08
> |
```

Chi-squared test in R can be used to test if two categorical variables are dependent, by means of a contingency table.

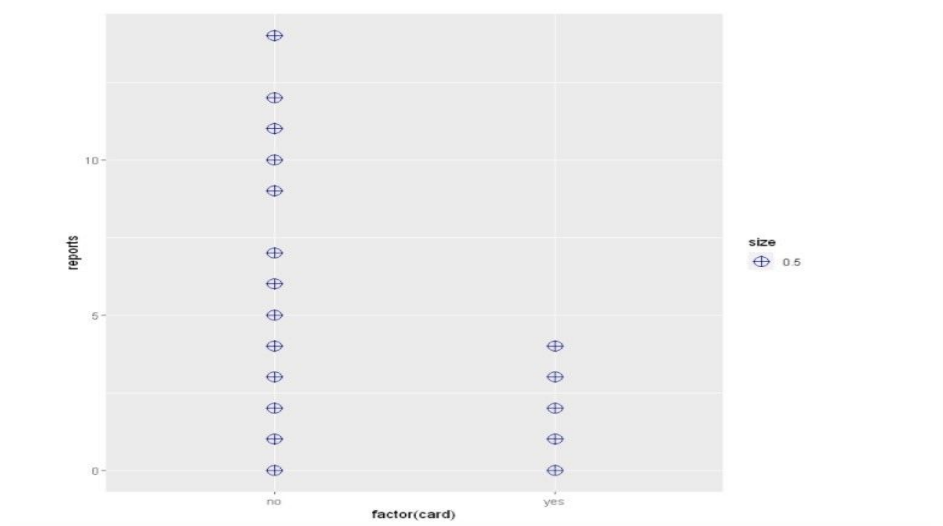
**Null hypothesis :** Card and owner are dependent.

**Using p-value :** Since the p-value is more than 0.05 we can reject the null hypothesis and infer that card and owner are independent variables.

**Using chi-square value :** For 2 x 2 contingency tables with 2 degrees of freedom (d.o.f), if the Chi-Squared calculated is greater than 3.841 (critical value), we reject the null hypothesis that the variables are independent. Here the chi-square value is 28.823 hence we reject the null hypothesis and conclude that card and owner are independent.

In conclusion getting a credit card is not dependent on whether you own a house or not.

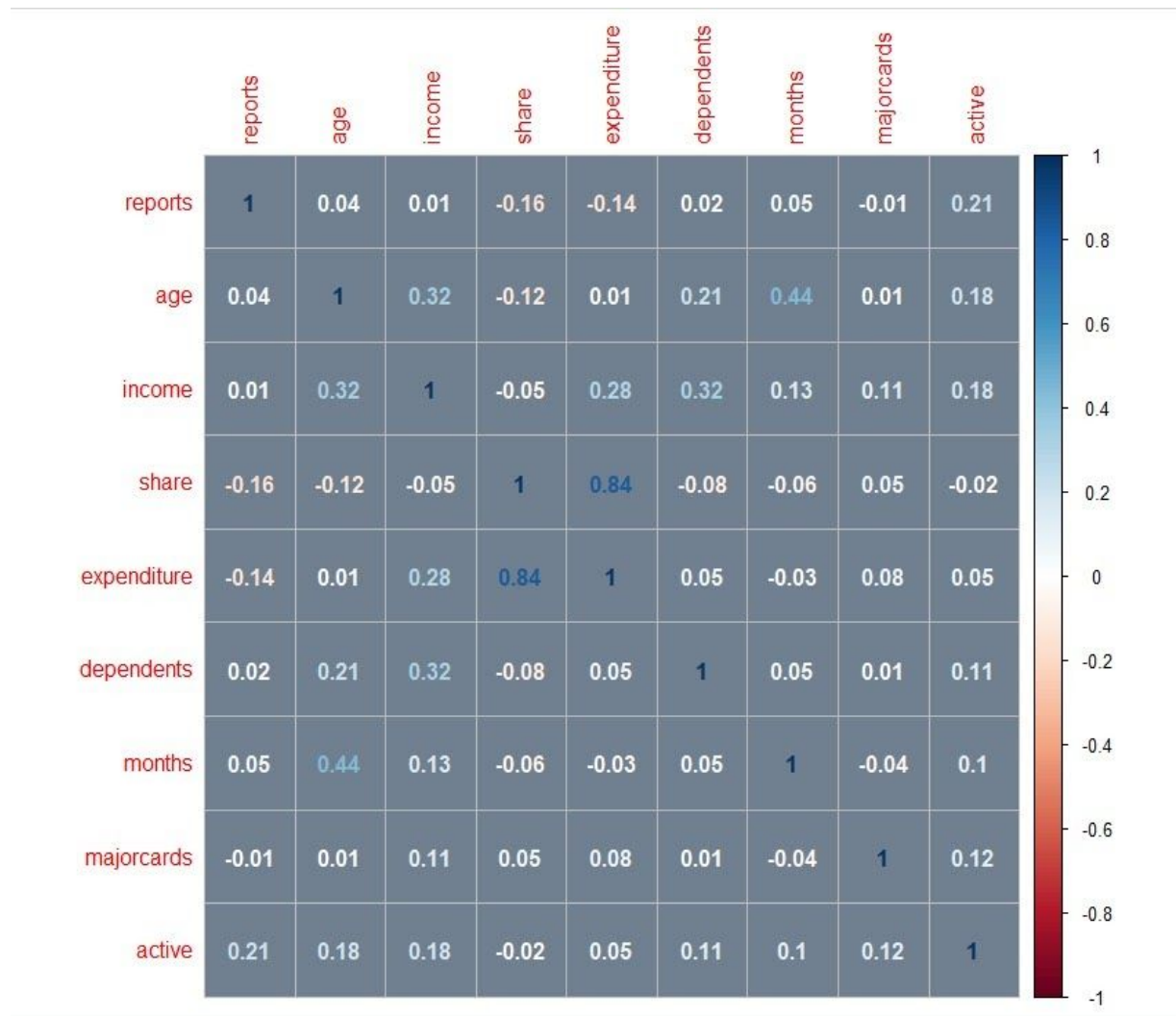
## 8) Scatter Plot



Derogatory reports are a major contributor to credit decisions. A major derogatory report results when a credit account that is being monitored by the credit reporting agency is more than 60 days late in payment. A minor derogatory report is generated when an account is 30 days delinquent. This scatter plot shows us that the probability of getting a credit card reduces as the number of derogatory reports increase. Here, you can see if the


number of derogatory reports are more than 5 then there is no possibility of getting a credit card. The evaluation of getting a credit card is based on other variables if the derogatory reports are less than 5.

## 9) Correlation Heatmap



Correlation is a statistical measure that expresses the extent to which two variables change together at a constant rate. In statistics, correlation coefficients are used to determine how strong a relationship is between two variables, and a heatmap indicating the same is called





the correlation heatmap. Although there are several kinds of correlation coefficients, I used Pearson's R to analyze the data. So, the correlation coefficient can take values from -1 to +1. The closer the value is towards 1, the more positively correlated it is.

As we can see here income and age are correlated with each other, their correlation values are 0.32. Also share and expenditure are highly correlated with each other and the value is 0.84.

The shades of blue (light to dark) above value 0 define how best the two variables are correlated, for example the share and expenditure. Whereas the values below 0 to -1 are least correlated.

## Conclusion

Through using different statistical and visualization approaches we have drawn various conclusions regarding the Credit Card dataset. From the initial analysis we are able to conclude that the most significant variables are income and reports (number of derogatory reports). We have seen how these have a direct impact on the approval for a credit card. While being an owner has no influence on getting approval for a credit card, the number of derogatory reports has a major impact on getting a credit card.

## Appendix- R Code

```

Datafinal= read.csv("C:/Users/SARAH/Desktop/CreditCard1.csv")
head(Datafinal,n=5)
sum(is.na(Datafinal))
summary(Datafinal)
a=median(Datafinal$age)
i=median(Datafinal$income)
s=median(Datafinal$share)
print(a)
print(i)
print(s)

library(ggplot2) #Histogram
ggplot(Datafinal)+geom_histogram(aes(x=age),bins=20,col='purple',fill='orange',linetype="dashed")
ggplot(Datafinal)+geom_histogram(aes(x=income),bins=20,col='lightblue',fill="darkblue",linetype="dashed")
ggplot(Datafinal)+geom_histogram(aes(x=share),bins=20,col='darkblue',fill='darkgreen',linetype="dashed")

#Box Plot
ggplot(Datafinal) +aes(x = "", y = age) +geom_boxplot(fill = "#0c4c8a") +theme_minimal()
ggplot(Datafinal) + aes(x = "", y = income) + geom_boxplot(fill = "#0c4c8a")+theme_minimal()

#chi-square test
chisq.test(table(Datafinal$card,Datafinal$owner), correct = FALSE)
summary(table(Datafinal$card,Datafinal$owner))

#Scatter Plot
ggplot(Datafinal,aes(x=factor(card),y=reports))+geom_point(color="Navyblue",shape=10,aes(size=0.5))

#Correlation matrix
library(hmsic)
library(corrplot)
mydata<-Datafinal %>%
select(reports,age,income,share,expenditure,dependents,months,majorcards,active)
mydata.rcorr=rcorr(as.matrix(mydata))
mydata.coeff=mydata.rcorr$r

```