Credit Card Acceptance Model

Introduction

This project will analyze the cross-section data on the credit history for a sample of applicants for a type of credit card. I will see the influence of different variables on whether or not an individual was accepted for a credit card. This data is from cran.r-project.org and was done by William H. Greene. We are evaluating whether or not someone was accepted for a credit card so it only holds two values - yes or no.

Explanation of variables

The Y value is whether or not an applicant was accepted and it holds two values: yes or no.

A data frame containing 1,319 observations on 12 variables.

card Factor. Was the application for a credit card accepted?

reports Number of major derogatory reports.

age Age in years plus twelfths of a year.

income Yearly income (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?

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.

Brief Overall Summary Statistics for the data

```
library(ISLR)
library(AER)
data("CreditCard")
CreditCard = data.frame(CreditCard)
summary(CreditCard)
```

```
##   card         reports           age              income
##  no : 296   Min.   : 0.0000   Min.   : 0.1667   Min.   : 0.210
##  yes:1023   1st Qu.: 0.0000   1st Qu.:25.4167   1st Qu.: 2.244
##             Median : 0.0000   Median :31.2500   Median : 2.900
##             Mean   : 0.4564   Mean   :33.2131   Mean   : 3.365
##             3rd Qu.: 0.0000   3rd Qu.:39.4167   3rd Qu.: 4.000
##             Max.   :14.0000   Max.   :83.5000   Max.   :13.500
##      share            expenditure       owner     selfemp      dependents
##  Min.   :0.0001091   Min.   :   0.000   no :738   no :1228   Min.   :0.0000
```

```
##  1st Qu.:0.0023159   1st Qu.:   4.583   yes:581   yes:  91   1st Qu.:0.0000
##  Median :0.0388272   Median : 101.298                         Median :1.0000
##  Mean   :0.0687322   Mean   : 185.057                         Mean   :0.9939
##  3rd Qu.:0.0936168   3rd Qu.: 249.036                         3rd Qu.:2.0000
##  Max.   :0.9063205   Max.   :3099.505                         Max.   :6.0000
##      months         majorcards         active
##  Min.   :  0.00   Min.   :0.0000   Min.   : 0.000
##  1st Qu.: 12.00   1st Qu.:1.0000   1st Qu.: 2.000
##  Median : 30.00   Median :1.0000   Median : 6.000
##  Mean   : 55.27   Mean   :0.8173   Mean   : 6.997
##  3rd Qu.: 72.00   3rd Qu.:1.0000   3rd Qu.:11.000
##  Max.   :540.00   Max.   :1.0000   Max.   :46.000
```

```
library(ggplot2)
```

This table of summary statistics provides a brief overview of the data we are presented with. We are provided with the min., 1st quartile, median, mean, 3rd quartile, and max of each variable. There are three variables that hold yes and no values: the output - card, owner(does individual own their home), and selfemp(is the individual self employed). Ialso see that most people who own major cards only have 0 or 1. Another thing that stood out to me when briefly looking at this data overview was that there may be some outliers in the dataset. For the average monthly credit card expenditure, I notice that the mean is 185.057 however the maximum value in that data is 3099.505. This dataset will be interesting to analyze and I will see what conclusions I can draw from it through deeper analysis.

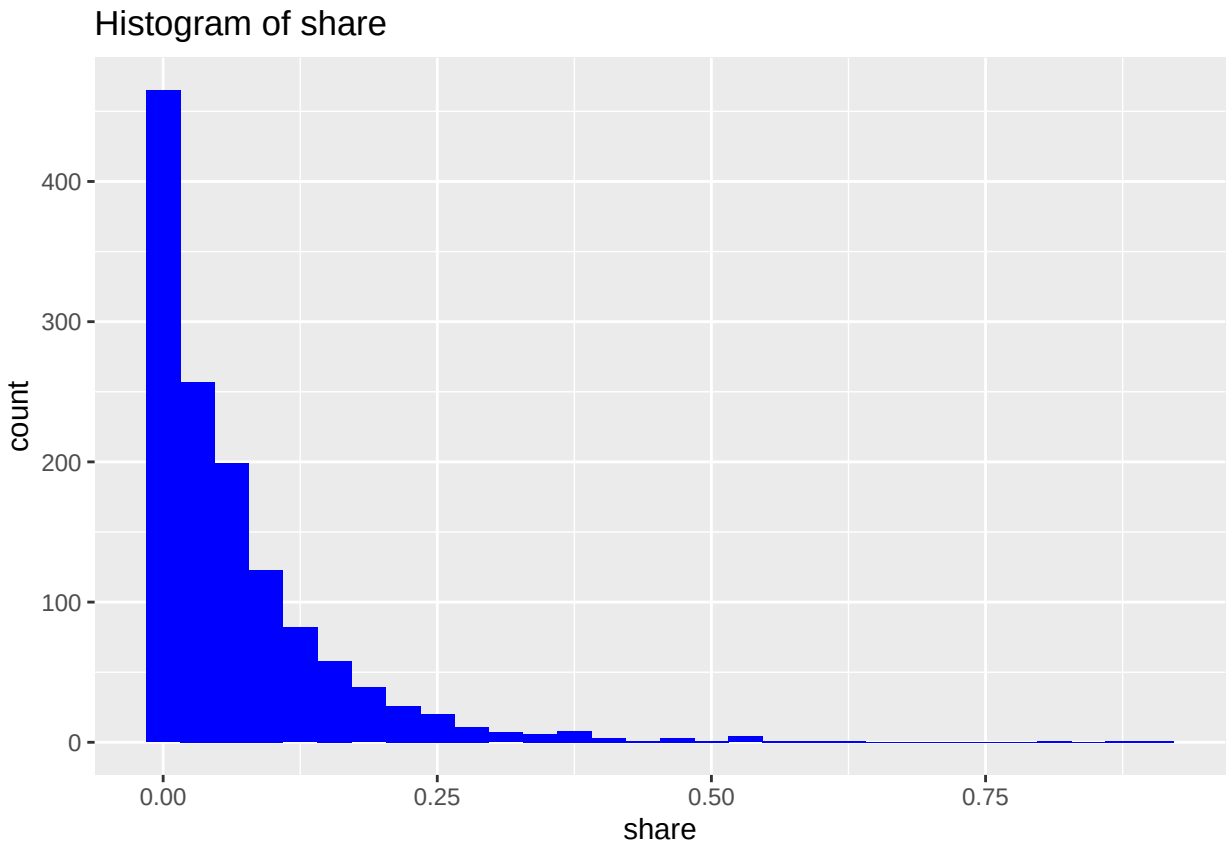
Analysis of four x variables

When looking at the x values provided, I believe the following can better explain the y: share, reports, majorcards, and active. I chose those four x values because intuitively, I assume that a negative impact on those four would negatively impact whether or not an individual gets a credit card so we would be able to see a correlation between them.

Analysis of share

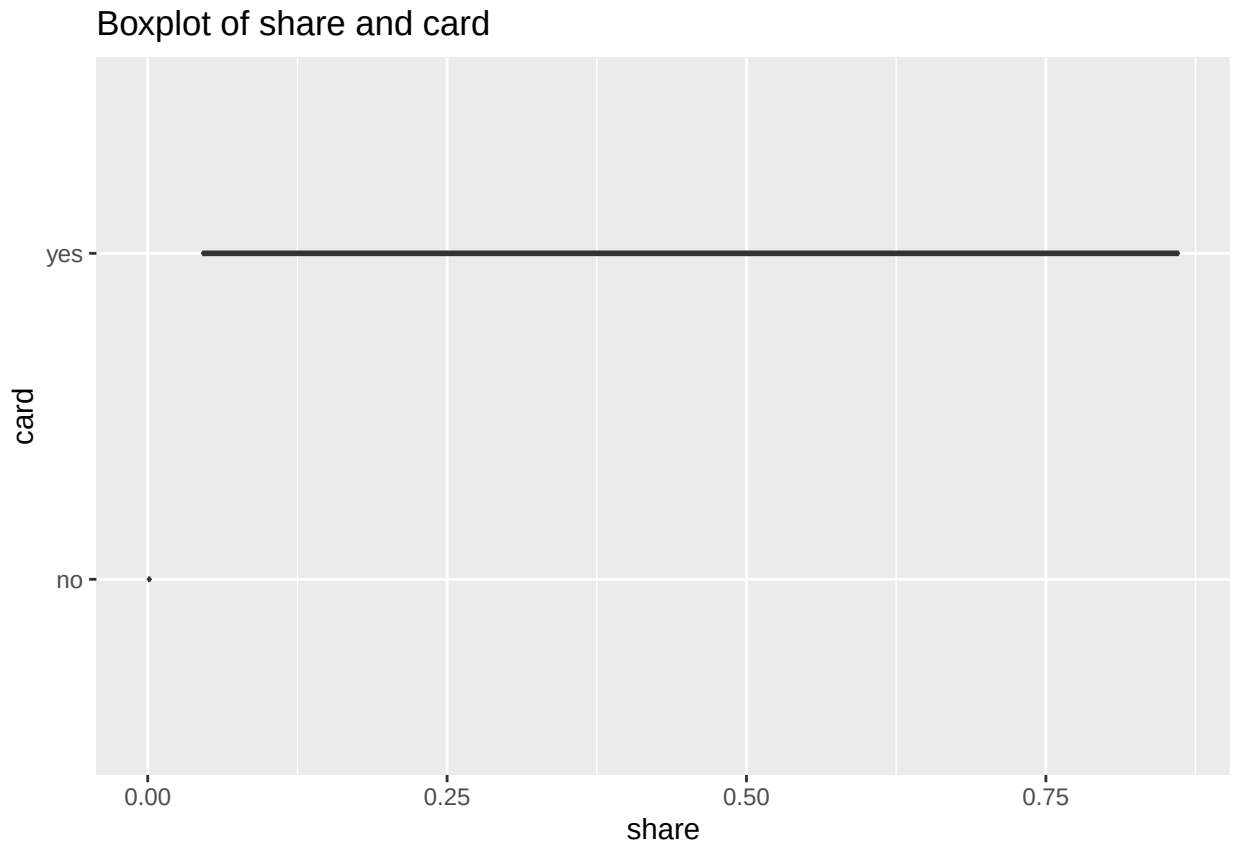
```
ggplot(CreditCard, aes(share), color = "blue")+
  geom_histogram(fill="blue") + ggtitle("Histogram of share")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

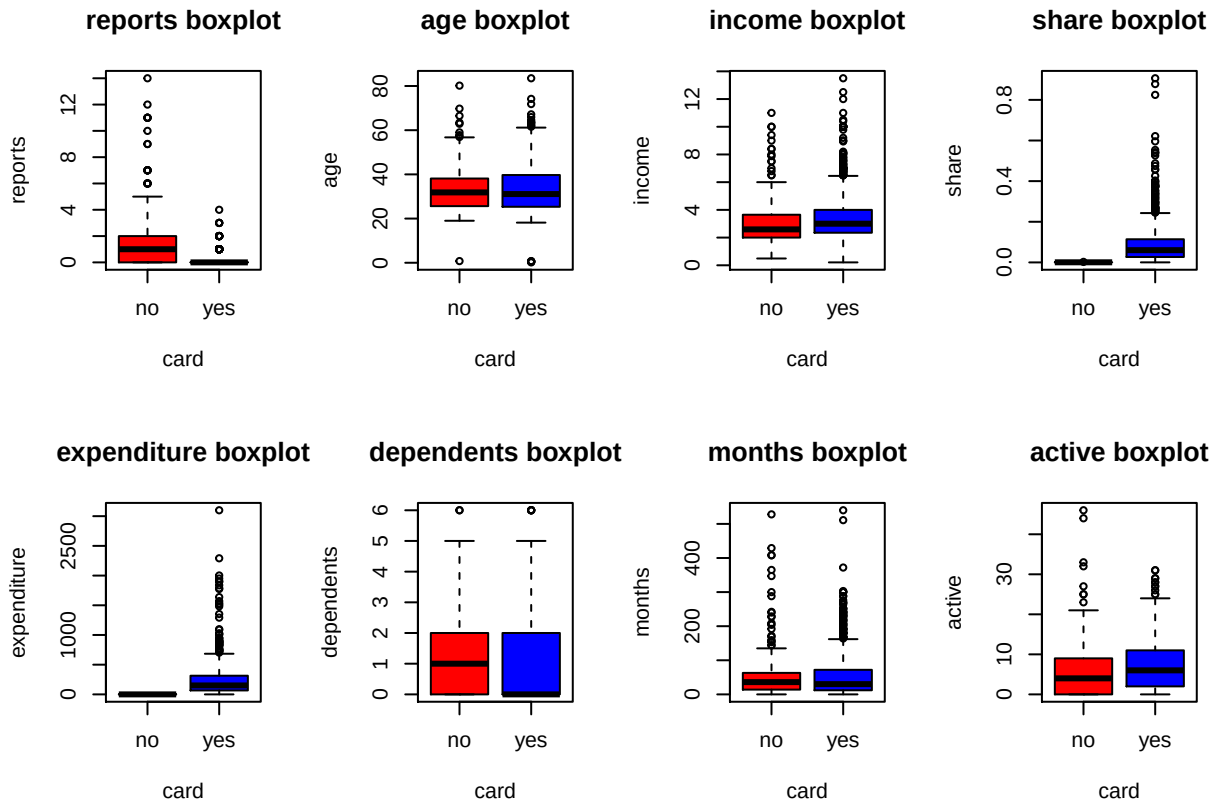


In the histogram above, I am analyzing a single variable, share, which is the ratio of monthly credit card expenditure to income. I notice that the histogram is right skewed. This says that people typically do not spend all of their income on credit card purchases. In fact, the data indicates that many people do not spend any money charged to a credit card. However, a good portion of the data indicates that many others do spend part of their income on credit card expenditures.

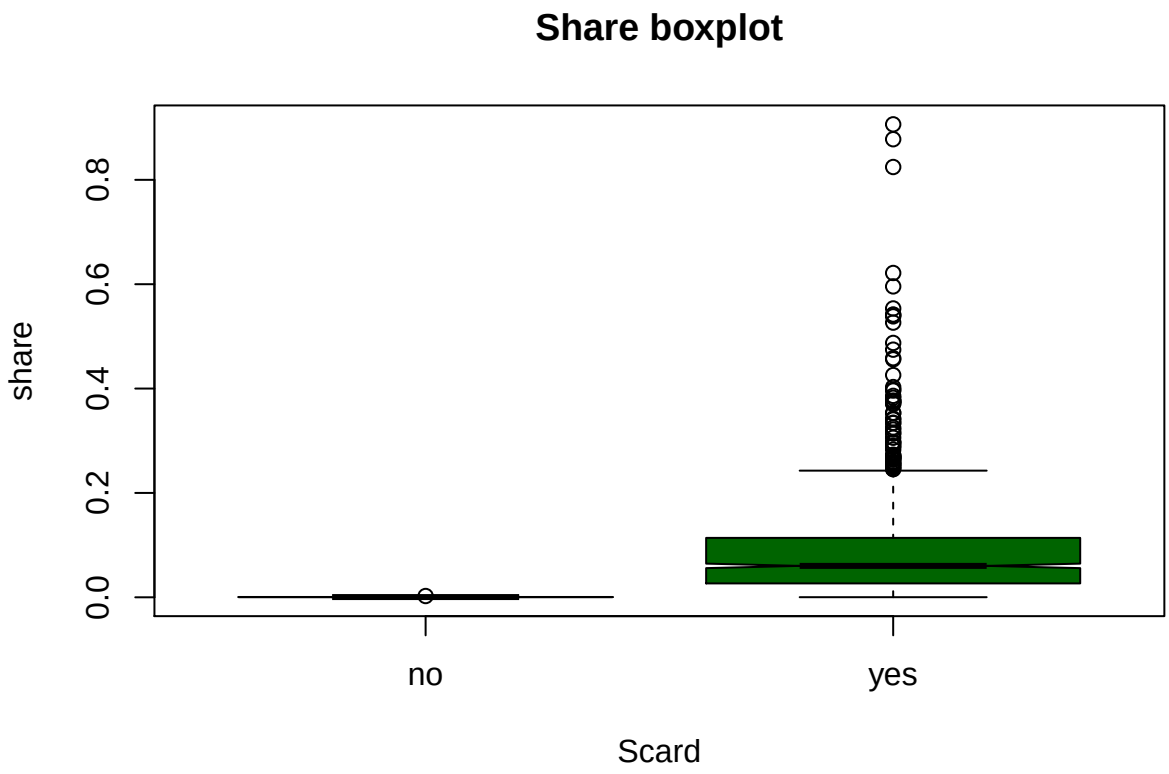
```
ggplot(CreditCard, aes(x=share, y=card)) +
  geom_boxplot() + ggtitle("Boxplot of share and card")
```




```
attach(CreditCard)
par(mfrow=c(2,4))
for(i in names(CreditCard)){
  # excluding the card variable and others categorical variables
  if( grepl(i, pattern="^card|owner|selfemp|name|majorcards")){ next}
  boxplot(eval(parse(text=i)) ~ card, ylab=i, main =paste(i, "boxplot"),
          col=c("red", "blue"))
}
```




```
share_boxplot = boxplot(share~card, data=CreditCard, notch=TRUE,
  col=(c("gold","darkgreen")),
  main="Share boxplot", xlab="Scard")
```



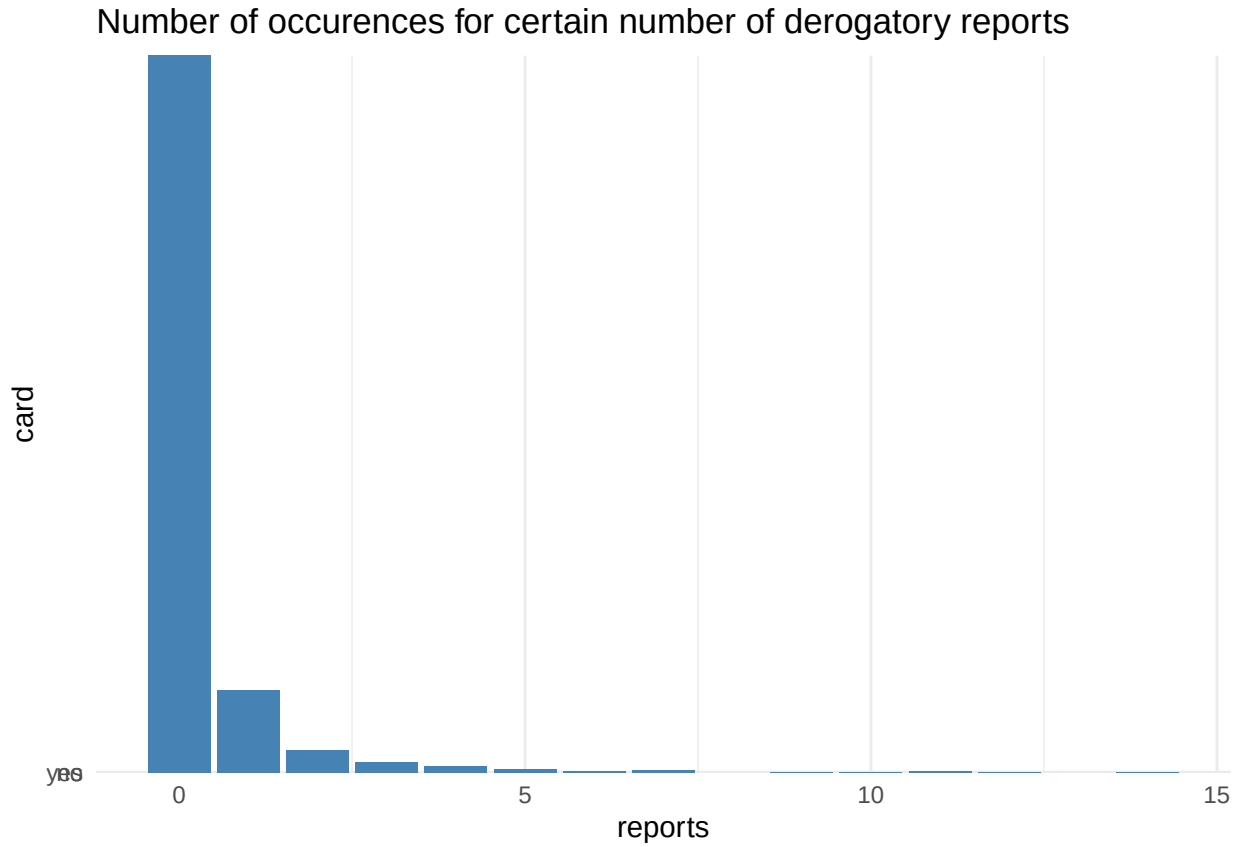

```
summary(share_boxplot)
```

```
##       Length Class  Mode
## stats 10     -none- numeric
## n      2     -none- numeric
## conf   4     -none- numeric
## out   62     -none- numeric
## group 62     -none- numeric
## names  2     -none- character
```

When looking at this graph, I notice that there were many credit cards approved even though the share of the monthly card expenditure and yearly income was in a wide range. However, this data also told me that everyone who got rejected for a credit card had a low share which is interesting to see because it seems that a low share is a good thing which means that people are not spending too money on their credit card in relation to their income.

Analysis of reports

```
ggplot(CreditCard, aes(x=reports, y=card)) +
  geom_bar(stat="identity", fill="steelblue")+
  theme_minimal()+ ggtitle("Number of occurences for certain number of derogatory reports")
```

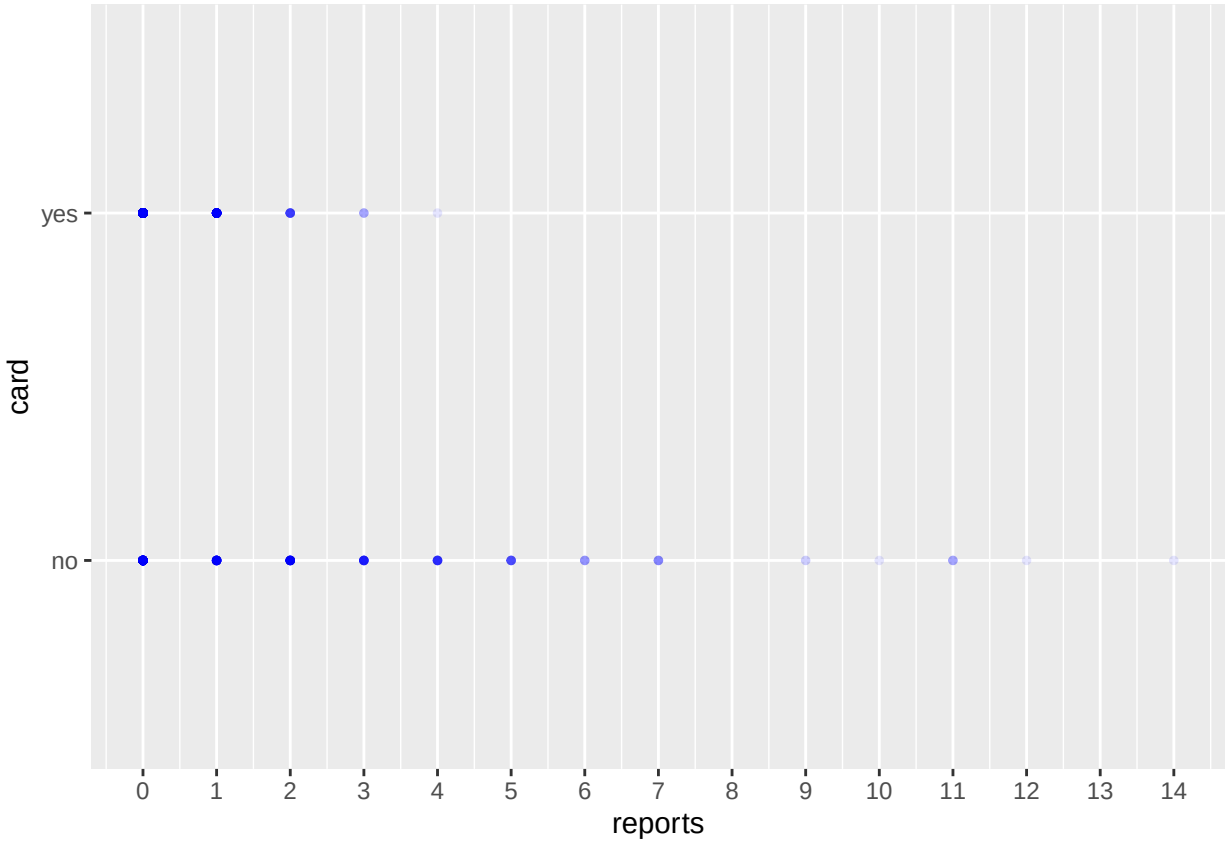



```
labs(y="count")
```

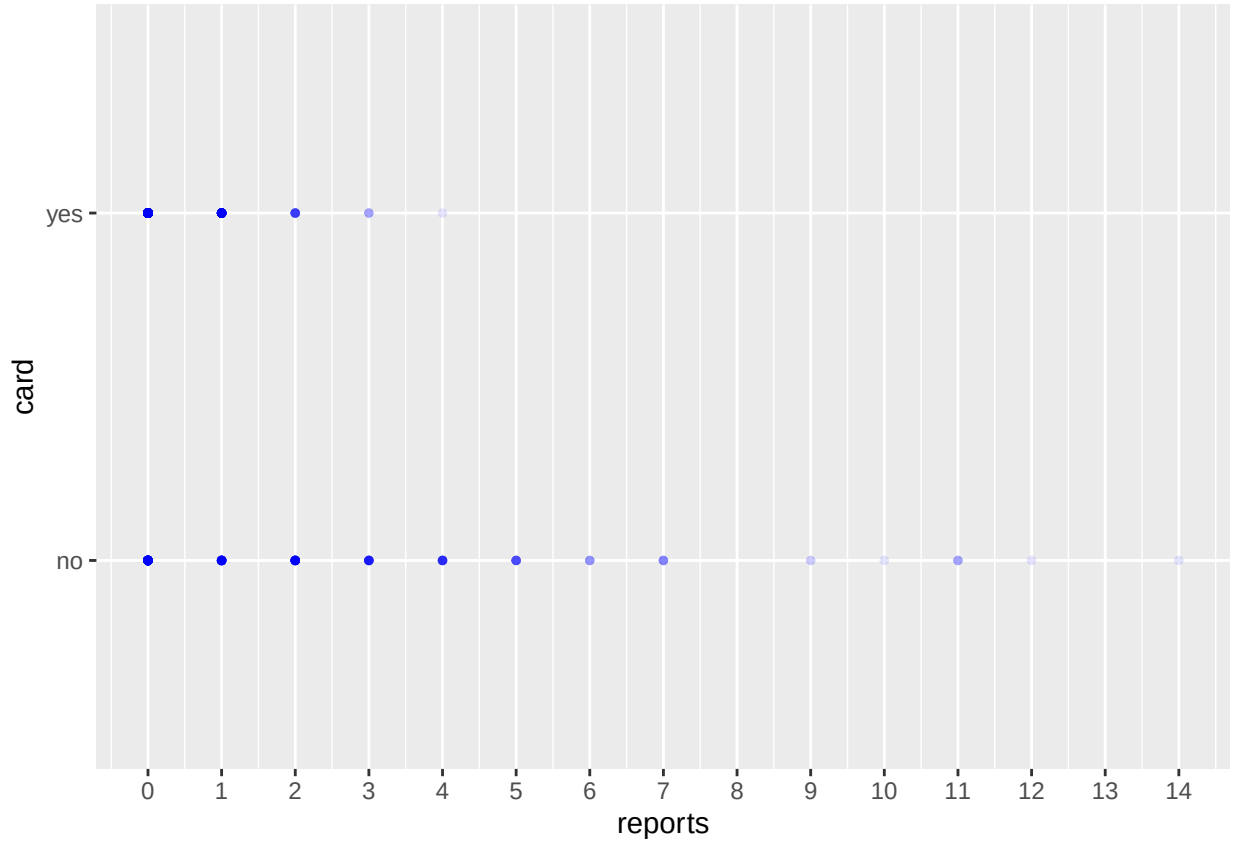
```
## $y
## [1] "count"
##
## attr(,"class")
## [1] "labels"
```

This barplot tells us that most people that applied for credit cards did not have any derogatory reports. The y-axis count tells us the number of observations of the derogatory reports that we see. This tells us that most people applying for a credit card did not commit crimes or do anything illegal since those that did may be deterred from opening a credit card if they committed a crime.

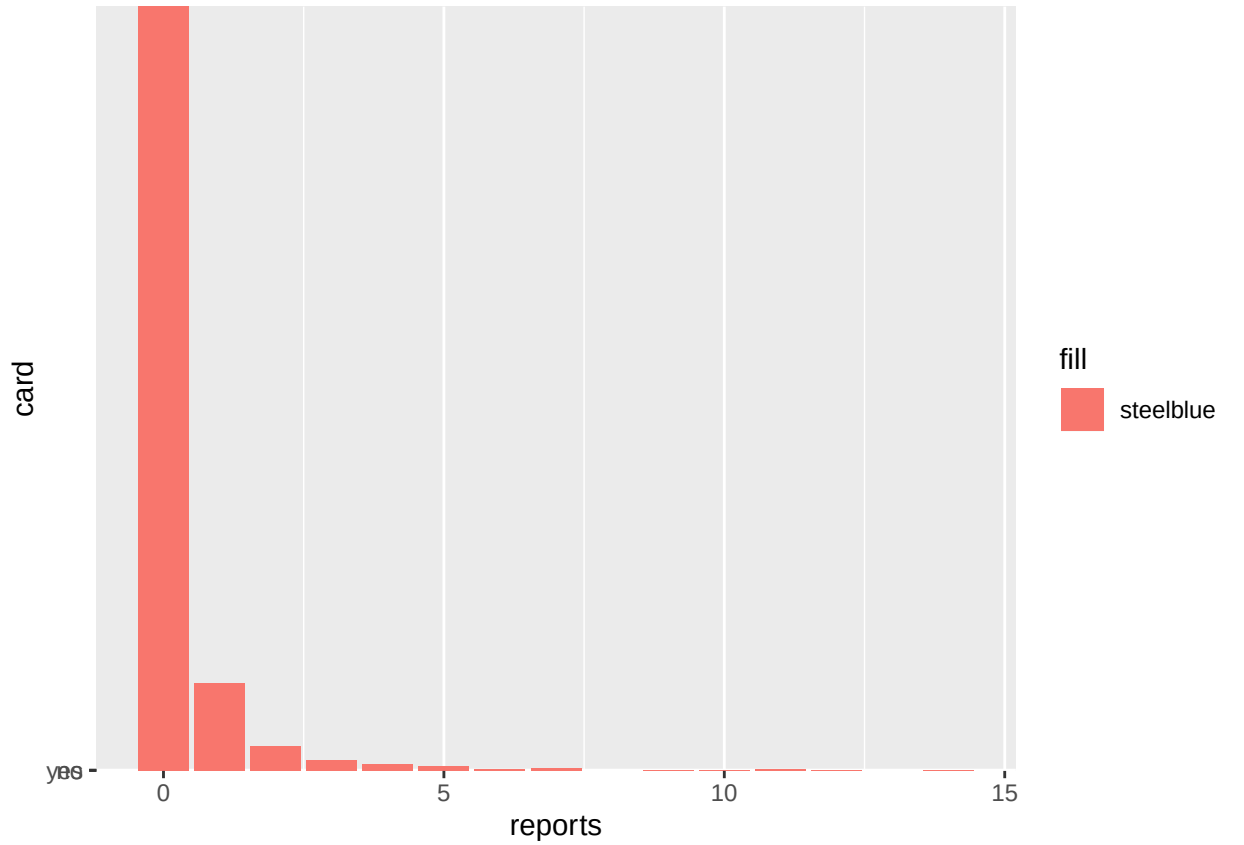
```
ggplot(CreditCard, aes(x=reports, y=card))+
  geom_point(color='blue', size = 1, alpha = 0.1) + scale_x_continuous(breaks = seq(0, 14, by = 1)) +
  labs(y="card", x="reports")
```




```
#stacked bar plot
ggplot(CreditCard, aes(x=reports, y=card))+
  geom_point(color='blue', size = 1, alpha = 0.1) + scale_x_continuous(breaks = seq(0, 14, by = 1)) +
  labs(y="card", x="reports")
```



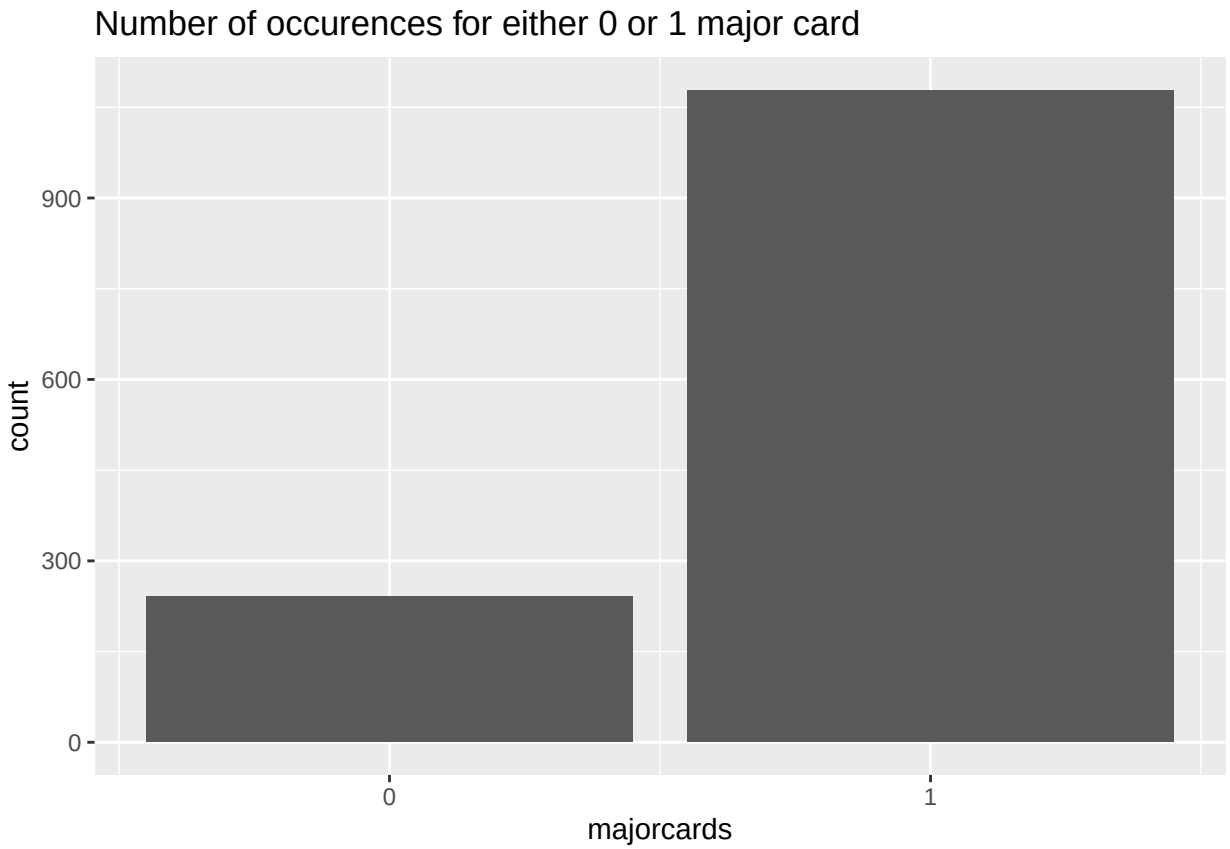

```
library(ggplot2)
# Stacked
ggplot(CreditCard, aes(fill='steelblue', y=card, x=reports)) +
    geom_bar(position="stack", stat="identity")
```



When looking at the plot I see that most people who get approved for a card typically do not have any derogatory reports. The shade of the points indicate the density so for example, if a point was more blue, that means it had many more points at that spot than another that was lighter. Also, I see that people who got rejected for a credit card had low reports as well. However, people with a high number of derogatory reports got rejected for a credit card as well. Most people who got approved for a credit card had less than 3 derogatory reports.

Analysis of majorcards

```
ggplot(CreditCard, aes(majorcards))+ scale_x_continuous(breaks=c(0,1)) +
  geom_bar() + ggtitle("Number of occurences for either 0 or 1 major card")
```




```
  # labs(y="count")
```

We see that most people who applied for a credit card only owned 1 major credit card already. More than a third of the people who applied with already 1 credit card, applied without having any major credit card at all. We will need to do further research to see whether or not it affected whether or not someone was approved for a credit card.

```
ggplot(CreditCard, aes(x=majorcards, y=card))+
  geom_point(color='blue', size = 1, alpha = 0.1) + scale_x_continuous(breaks=c(0,1))
```

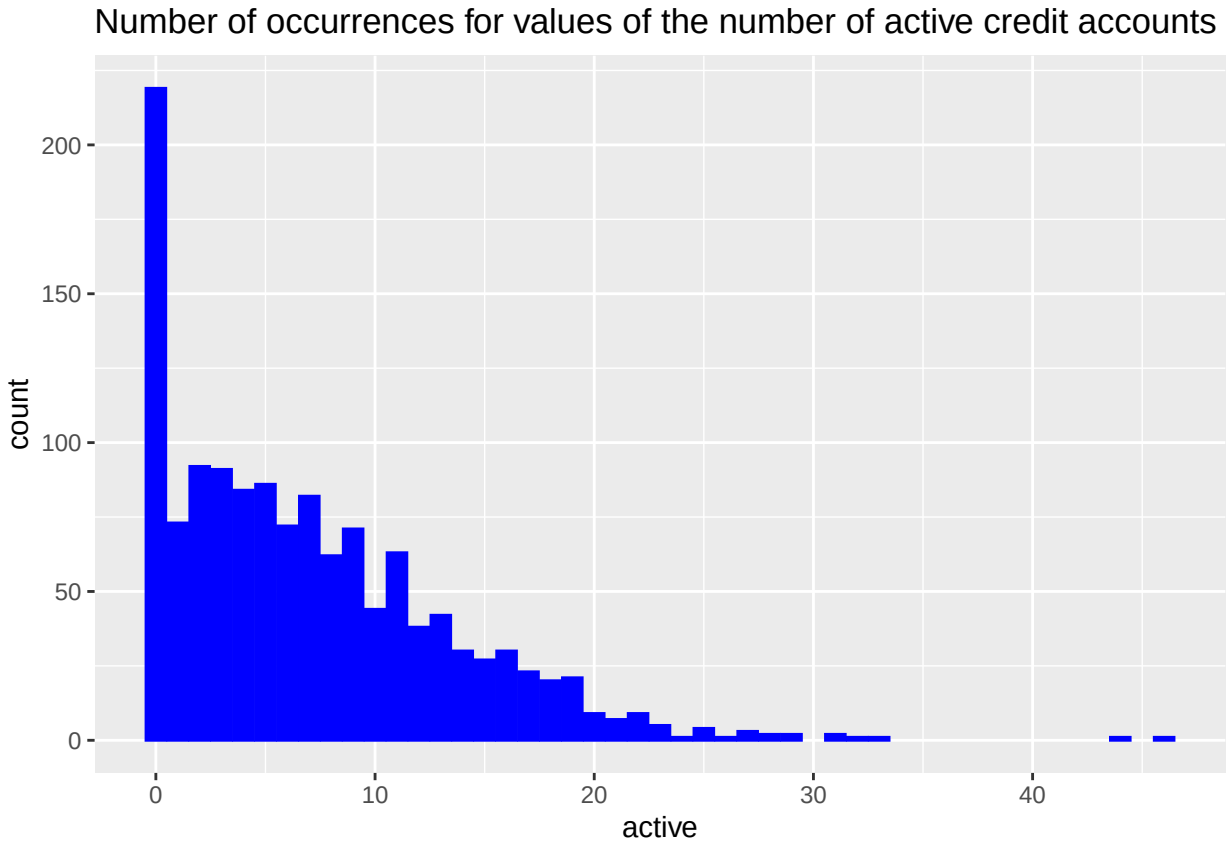

```
  labs(y="card", x="majorcards")
```

```
## $y
## [1] "card"
##
## $x
## [1] "majorcards"
##
## attr(,"class")
## [1] "labels"
```

This graph shows me that there is typically an equal distribution between people who get approved for a

Analysis of Active

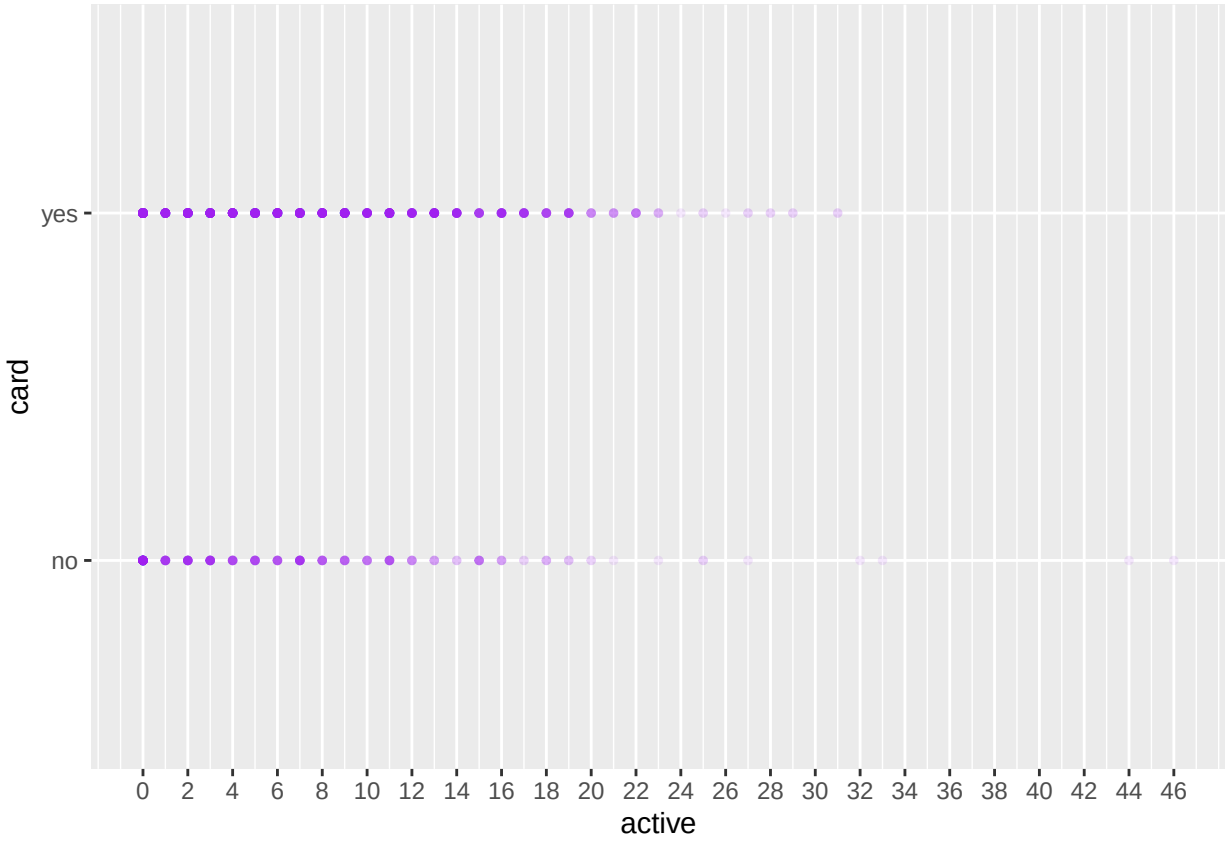
```
ggplot(CreditCard, aes(active))+
  geom_bar(color = "blue", fill = "blue") + ggtitle("Number of occurrences for values of the number of a
```




```
  # labs(y="count")
```

This graph tells us the distribution of the number of active credit accounts for the dataset we have. M

```
ggplot(CreditCard, aes(x=active, y=card))+
  geom_point(color='purple', size = 1, alpha = 0.1) +scale_x_continuous(breaks = seq(0, 46, by = 2)) +
  labs(y="card", x="active")
```



This graph indicates to me that the number of active credit accounts may not greatly influence the acceptance of the credit card. The distribution for the range from 0 to 6 of active credit accounts look very similar for yes and no in cards. Furthermore, we see from the graph that a good number of individuals with a higher number of active credit accounts, from 10 to 20, even get approved for the credit card. In fact, if we look at a high number like 16 active credit accounts, we see that more people got accepted, rather than rejected, for the credit card.