Econ 104 Project 3

2023-12-06

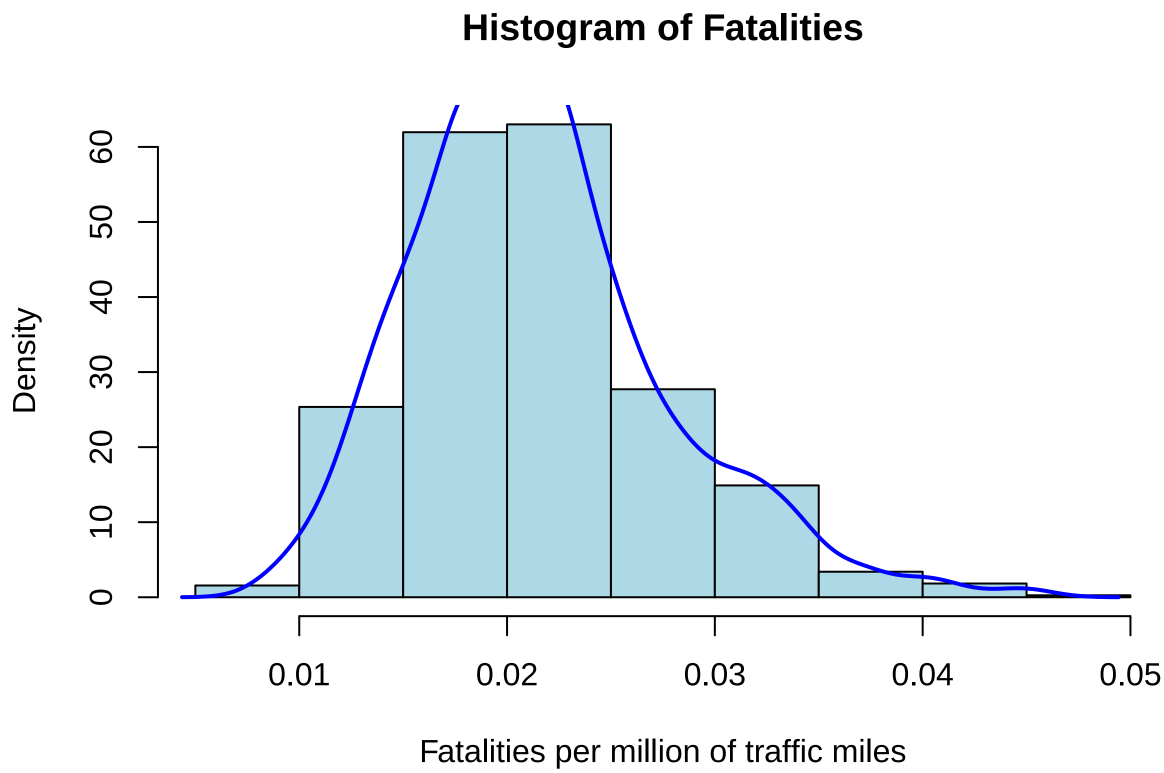
Authors: Sia Phulambrikar, Ahnaf Tamid, Sofia Giorgi, Michael Sorooshian

(a) Briefly discuss the question you are trying to answer with your model.

The dataset shows US panel data from 1983-1997. Using USSeatBelts, we are trying to answer how the number of fatalities per million of traffic miles (fatalities) is affected by seatbelt usage rate (seatbelt), whether there is a 65 mile per hour speed limit (speed65), whether there is a maximum of 0.08 blood alcohol content (alcohol), the median per capita income (income), and mean age (age). USSeatBelts can be found in the AER library: https://cran.r-project.org/web/packages/AER/AER.pdf

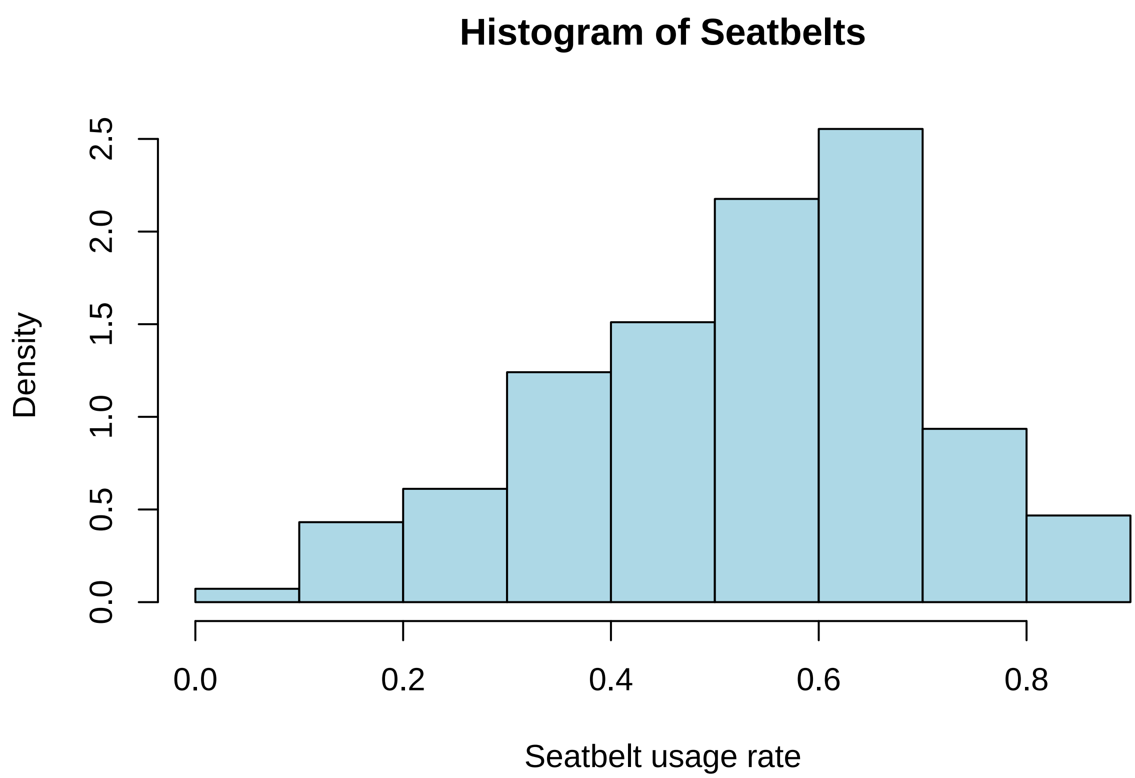
(b) Provide a descriptive analysis of your variables. This should include relevant figures with comments including some graphical depiction of individual heterogeneity.

```
hist(USSeatBelts[,"fatalities"],prob=TRUE,col="lightblue",main="Histogram of Fatalities",
     xlab="Fatalities per million of traffic miles")
lines(density(USSeatBelts[,"fatalities"]),col="blue",lwd=2)
```



Here, we have a histogram of fatalities, which is almost normally distributed but with a slight skew to the right, as can be seen by its tail.

```
hist(USSeatBelts[,"seatbelt"],prob=TRUE,col="lightblue",main="Histogram of Seatbelts",
     xlab="Seatbelt usage rate")
```



The histogram of seatbelt usage rate has a skew to the left.

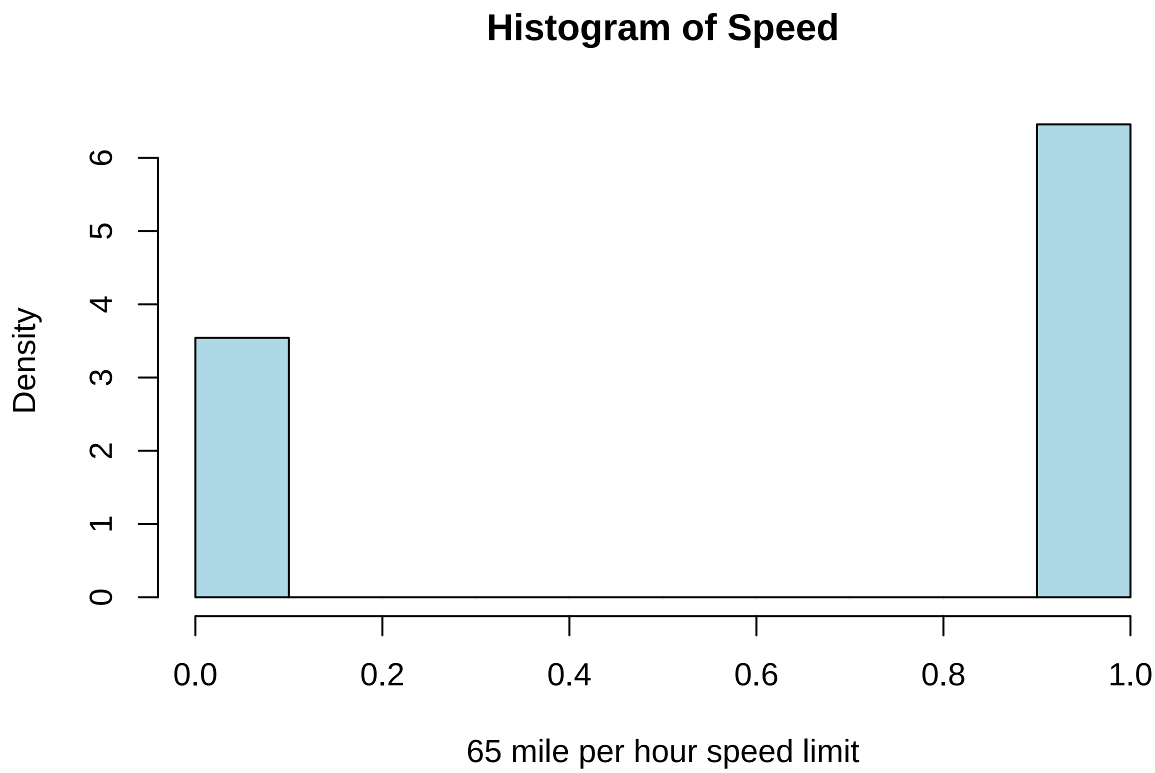
```r
tail(USSeatBelts[,"seatbelt"])
```

```
## [1] 0.66 0.67 0.70 0.71 0.72 0.75
```

```r
summary(USSeatBelts[,"speed65"])
```

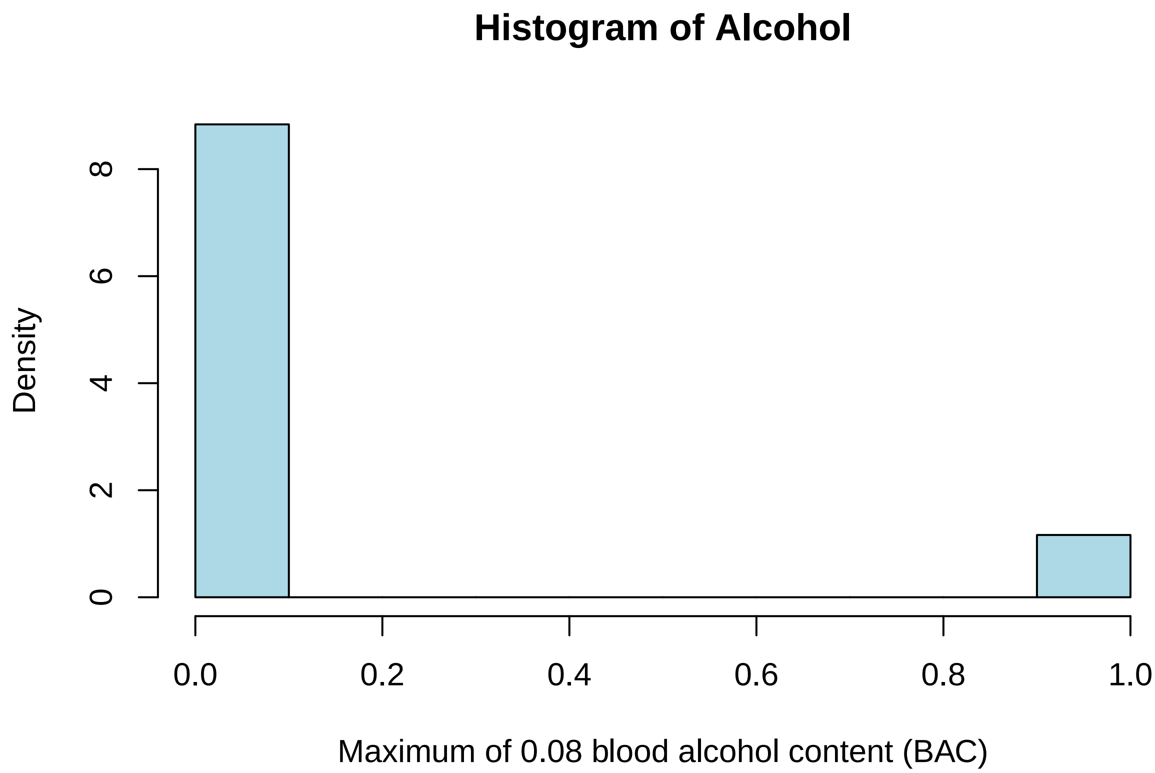
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  1.0000  0.6458  1.0000  1.0000
```

```r
hist(USSeatBelts[,"speed65"],prob=TRUE,col="lightblue",main="Histogram of Speed",
     xlab="65 mile per hour speed limit")
```



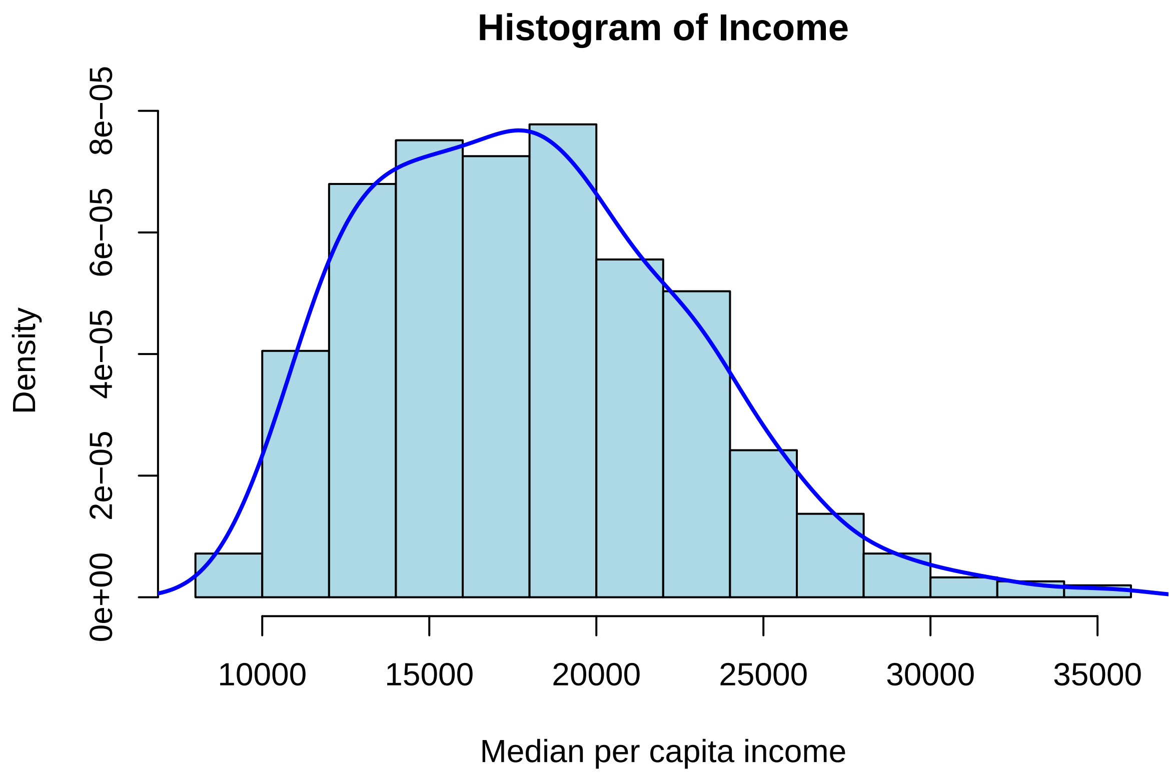
The 65 mile per hour speed limit is binary as either a 1 (there is a 65mph speed limit) or 0 (no 65 mph speed limit). Therefore, the graph will innately not be normally distributed, but this histogram does show that there is about twice as much density for the value of 1.

```
hist(USSeatBelts[,"alcohol"],prob=TRUE,col="lightblue",main="Histogram of Alcohol",
     xlab="Maximum of 0.08 blood alcohol content (BAC)")
```



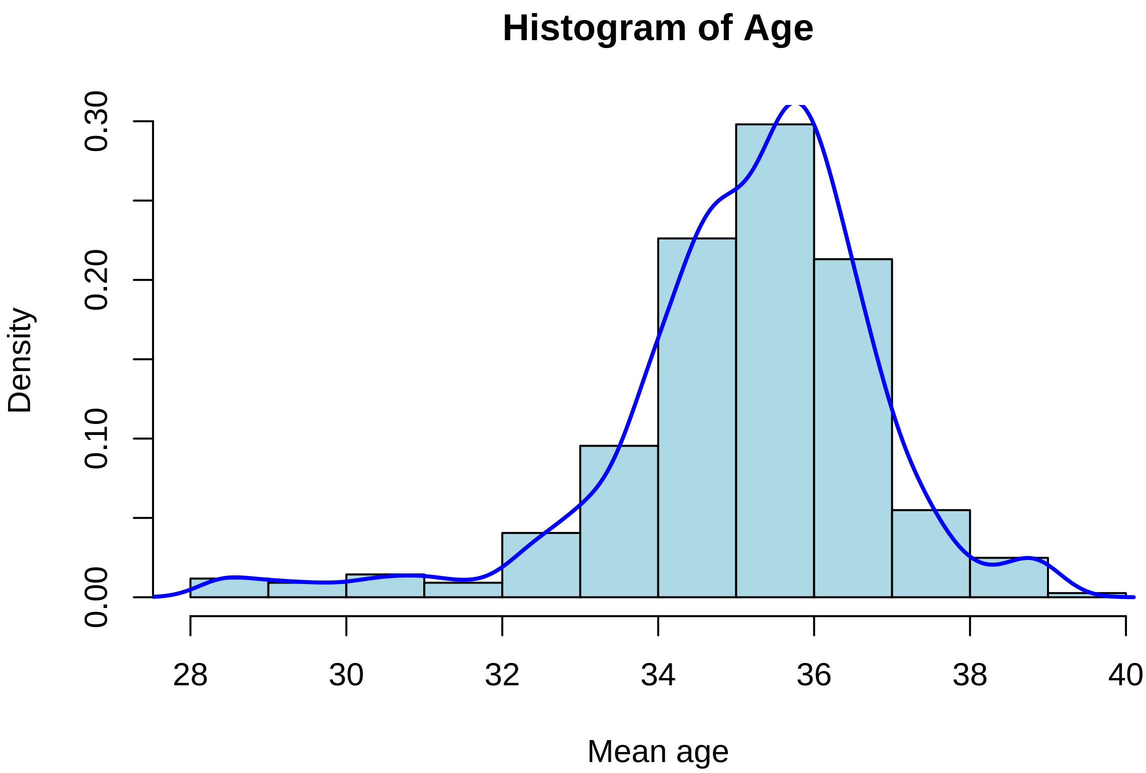
Similar to speed, the maximum of 0.08 BAC is a binary variable of either 1 (there is a maximum of 0.08 BAC) or 0 (there is not a maximum of 0.08 BAC). In this dataset, there is a disproportionate amount of density on 0.

```
hist(USSeatBelts[,"income"],prob=TRUE,col="lightblue",main="Histogram of Income",
     xlab="Median per capita income")
lines(density(USSeatBelts[,"income"]),col="blue",lwd=2)
```



The median per capita income is skewed to the right and has large tails.

```
hist(USSeatBelts[,"age"],prob=TRUE,col="lightblue",main="Histogram of Age",xlab="Mean age")
lines(density(USSeatBelts[,"age"]),col="blue",lwd=2)
```



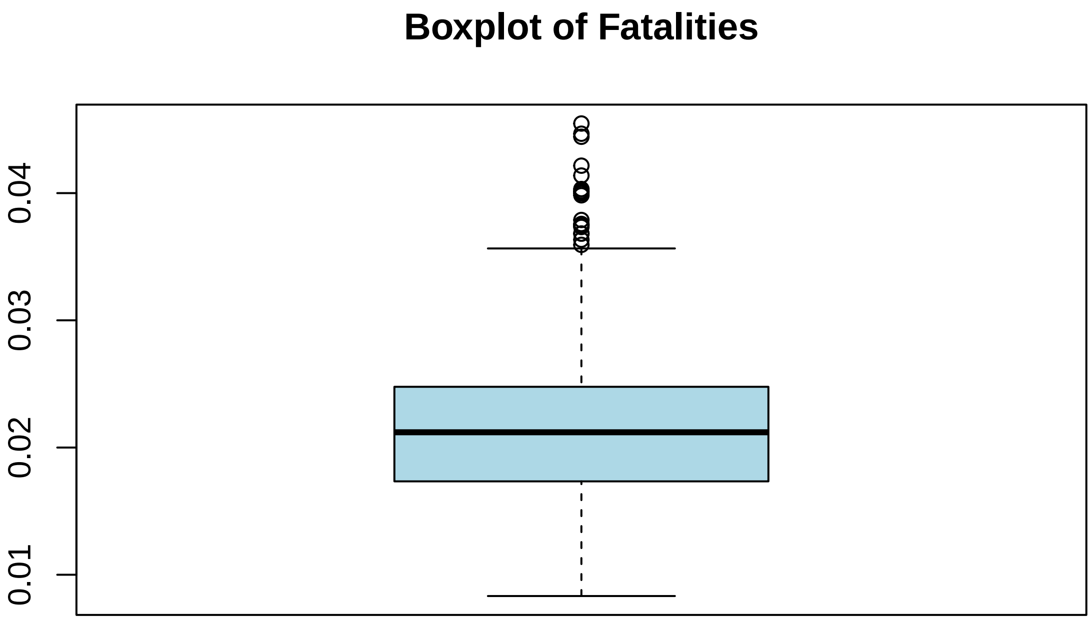
The mean age is roughly normally distributed but clearly skewed to the left.

```r
USSeatBelts_vars <- USSeatBelts[, c("fatalities", "seatbelt", "speed65","alcohol","income","age")]
summary(USSeatBelts_vars)
```

```
##    fatalities          seatbelt        speed65          alcohol      
##  Min.   :0.008327   Min.   :0.0600   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.017341   1st Qu.:0.4200   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :0.021199   Median :0.5500   Median :1.0000   Median :0.0000  
##  Mean   :0.021490   Mean   :0.5289   Mean   :0.6458   Mean   :0.1163  
##  3rd Qu.:0.024774   3rd Qu.:0.6500   3rd Qu.:1.0000   3rd Qu.:0.0000  
##  Max.   :0.045470   Max.   :0.8700   Max.   :1.0000   Max.   :1.0000  
##                     NA's   :209                                       
##      income           age       
##  Min.   : 8372   Min.   :28.23  
##  1st Qu.:14266   1st Qu.:34.39  
##  Median :17624   Median :35.39  
##  Mean   :17993   Mean   :35.14  
##  3rd Qu.:21080   3rd Qu.:36.13  
##  Max.   :35863   Max.   :39.17  
## 
```

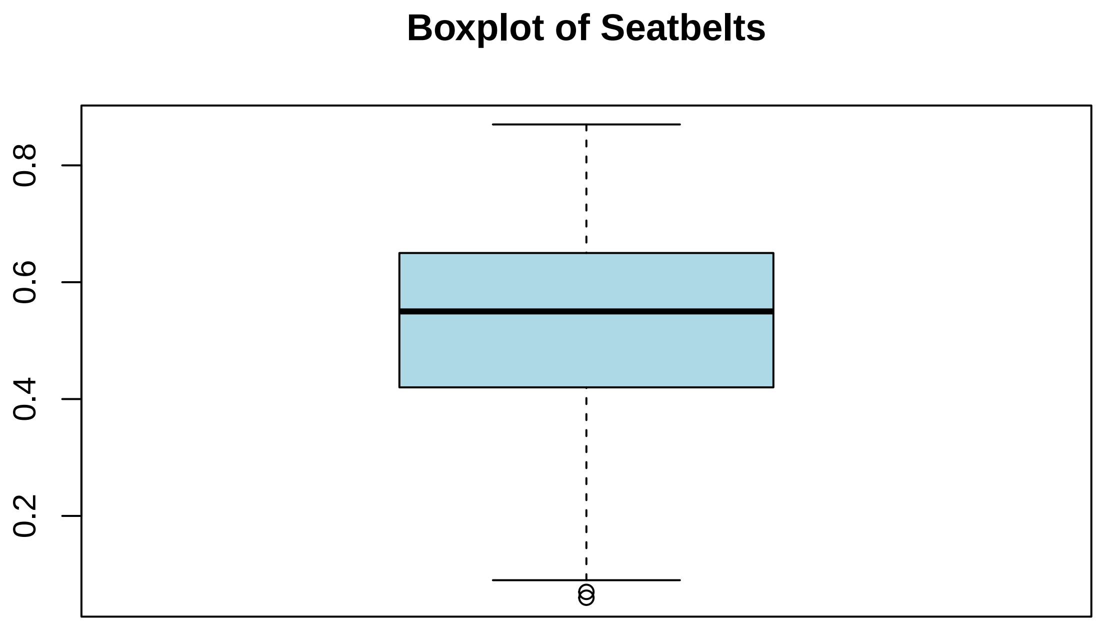
Here is a summary of the variables, which are further explored in the boxplots below.

```r
boxplot(USSeatBelts[,"fatalities"], main = "Boxplot of Fatalities", col="lightblue")
```



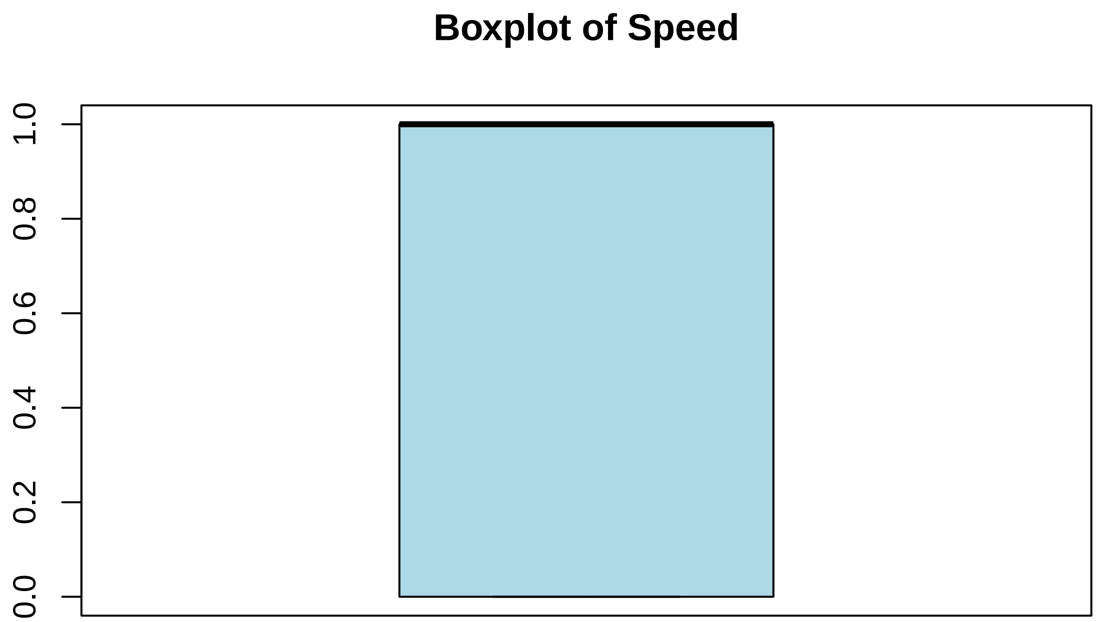
Fatalities has a minimum of approximately 0.008, a median of 0.02, and a maximum of 0.05. The series of points beyond the third quartile further shows its skew to the right.

```r
boxplot(USSeatBelts[,"seatbelt"], main = "Boxplot of Seatbelts", col="lightblue")
```



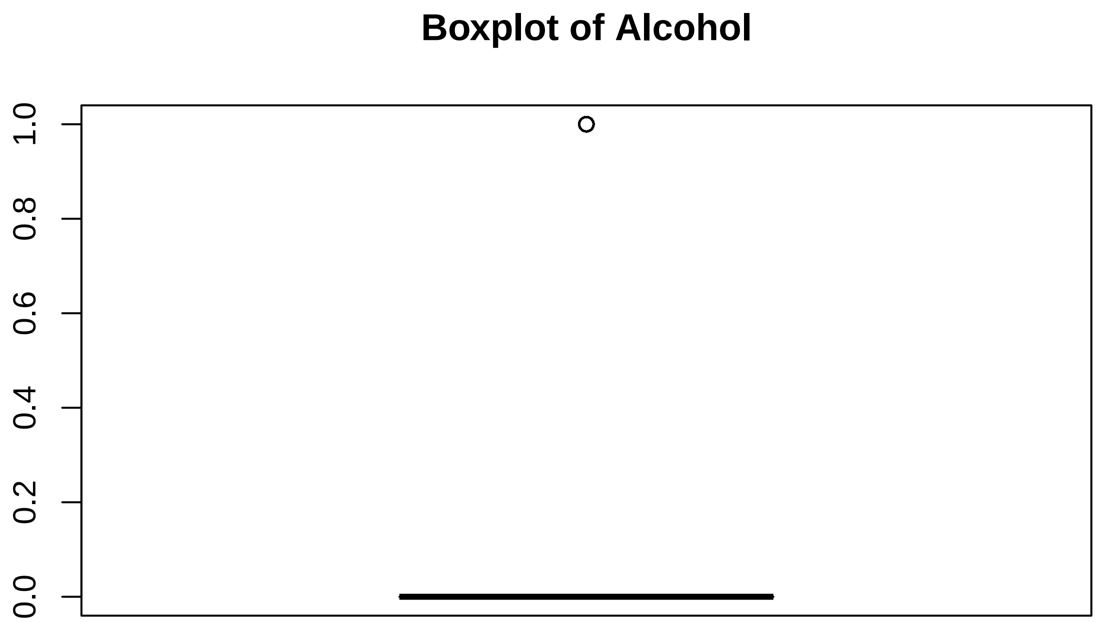
Seatbelts has a denser tail, with a minimum of about 0.06, median of 0.55, and maximum of 0.87. Its “NA” values have been omitted.

```
boxplot(USSeatBelts[,"speed65"], main = "Boxplot of Speed", col="lightblue")
```



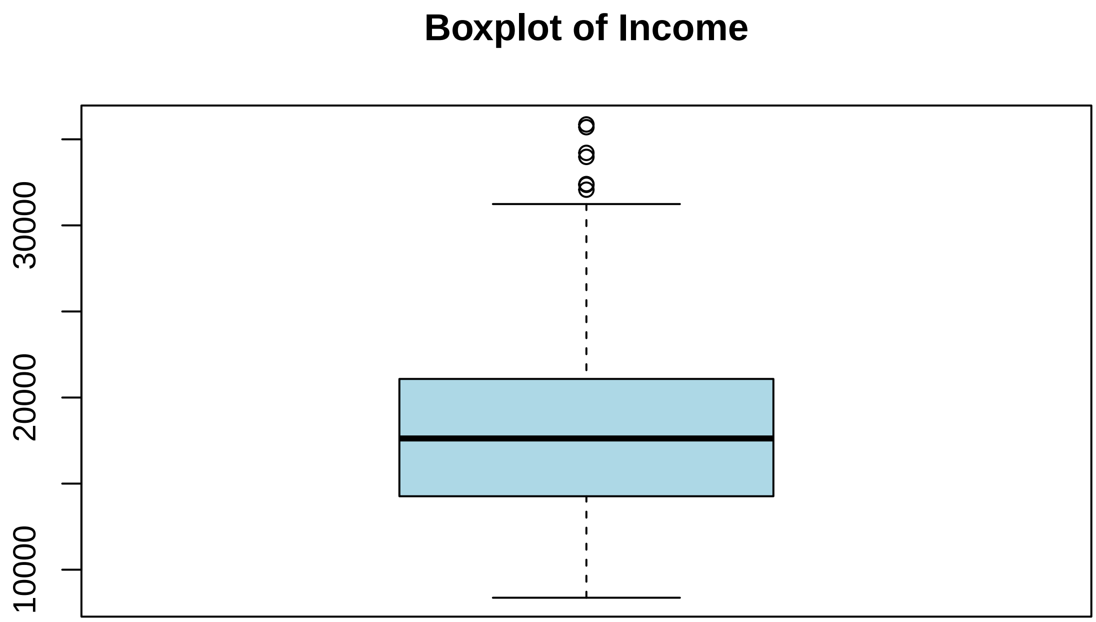
The boxplot of speed is interesting, and it is shown that the median is 1 while the mean is 0.65.

```
boxplot(USSeatBelts[,"alcohol"], main = "Boxplot of Alcohol", col="lightblue")
```



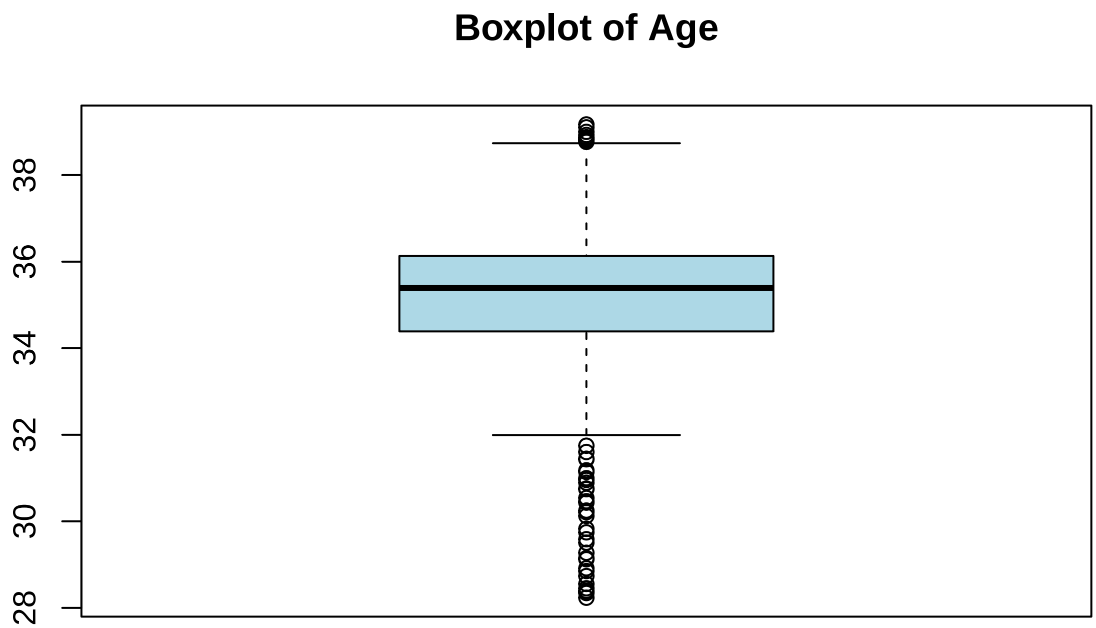
The boxplot of alcohol is the opposite of the speed boxplot, and shows that the median is 0. The data summary states that the mean is 0.11– much closer to 0 than 1.

```
boxplot(USSeatBelts[,"income"], main = "Boxplot of Income", col="lightblue")
```



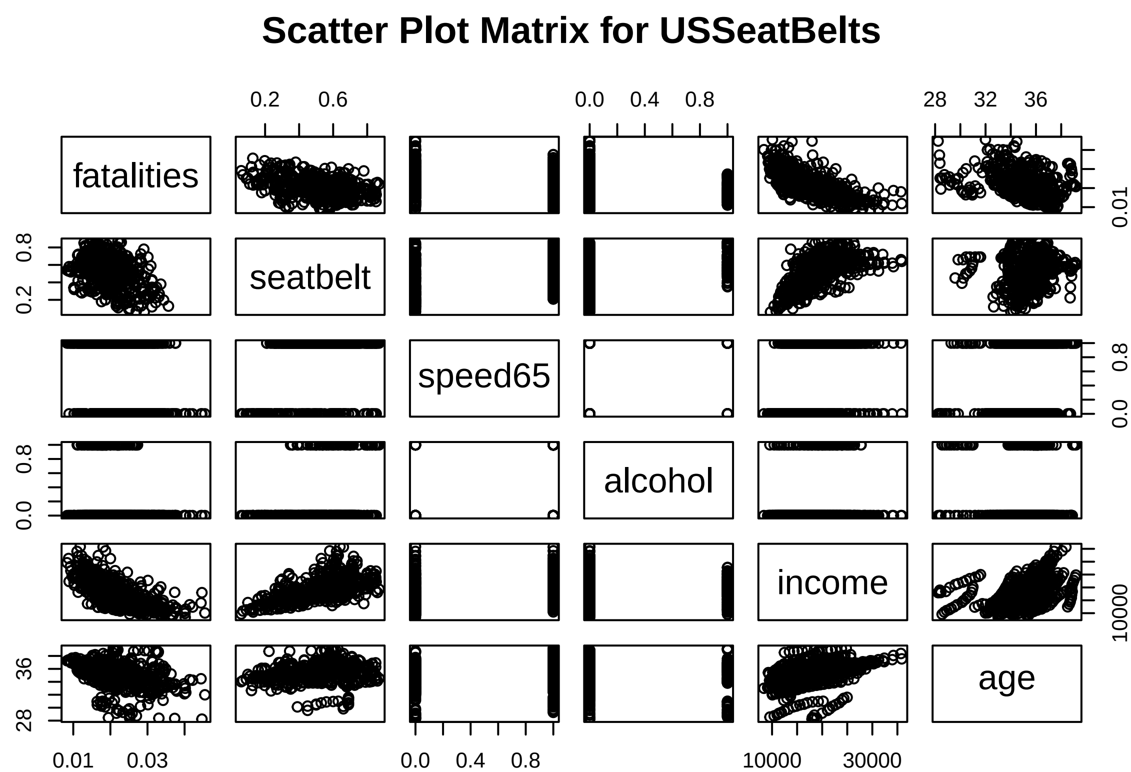
Income has a boxplot reminiscent of its right-skew. Its minimum is about 8,372, mean is 17,993, median is 17,624, and maximum is 35,863.

```
boxplot(USSeatBelts[,"age"], main = "Boxplot of Age", col="lightblue")
```



Age has a minimum of 28, with mean and median of 35 and maximum of 39, showing that the ages shown in this dataset were not considerably varied.

```r
pairs(USSeatBelts[, c("fatalities", "seatbelt", "speed65","alcohol","income","age")],
      main = "Scatter Plot Matrix for USSeatBelts")
```



Individual heterogeneity can be detected in this scatterplot matrix. Seatbelt and age, for instance, seem to separate into 2 distinct groups (an indicator of heterogeneity). The same thing can be found between fatalities and age, and income and age.

```r
matrix <- cor(USSeatBelts_vars)
print(matrix)
```

```
##            fatalities seatbelt    speed65     alcohol     income         age
## fatalities  1.0000000       NA -0.2818365 -0.16983803 -0.7035576 -0.37541306
## seatbelt           NA        1         NA          NA         NA          NA
## speed65    -0.2818365       NA  1.0000000  0.19203024  0.3616334  0.18895898
## alcohol    -0.1698380       NA  0.1920302  1.00000000  0.1218024 -0.05466039
## income     -0.7035576       NA  0.3616334  0.12180241  1.0000000  0.40752738
## age        -0.3754131       NA  0.1889590 -0.05466039  0.4075274  1.00000000
```

This correlation matrix shows us that many of the variables are slightly negatively correlated. Only income and age have a slightly notable positive correlation (higher income is associated with higher age). Income and fatalities have the most significant correlation, which is negative– meaning that higher income individuals had less fatalities.

c) Pooled Model

```r
pdata <- pdata.frame(USSeatBelts, index = c("state", "year"))
```

```r
pooled_model <- plm(fatalities ~ seatbelt + speed65 + alcohol + income + age,
                    data =USSeatBelts, model = "pooling")
crse<- coeftest(pooled_model, vcov=vcovHC(pooled_model,
type="HC0",cluster="group"))
stargazer(pooled_model, crse, column.labels = c("\\textit{Pooled}", "\\textit{Pooled(prse)}"),
model.names = FALSE,type = "text")
```

```
##
## ===================================================
##                        Dependent variable:
##              --------------------------------------
##                     fatalities
##                       Pooled          Pooled(prse)
##                         (1)                (2)
## ---------------------------------------------------
## seatbelt              0.002*             0.002
##                      (0.001)            (0.003)
##
## speed65              -0.00003           -0.00003
##                      (0.0004)           (0.001)
##
## alcohol             -0.002***           -0.002**
##                      (0.0005)           (0.001)
##
## income             -0.00000***        -0.00000***
##                     (0.00000)          (0.00000)
##
## age                  -0.0001            -0.0001
##                      (0.0001)           (0.0004)
##
## Constant             0.038***           0.038***
##                      (0.004)            (0.014)
##
## ---------------------------------------------------
## Observations           556
## R2                    0.493
## Adjusted R2           0.489
## F Statistic  107.140*** (df = 5; 550)
## ===================================================
## Note:                  *p<0.1; **p<0.05; ***p<0.01
```

From the Summary we see that with a one unit increase in seatbelt we see an increase in fatalities. Also we see that with a one unit increase in alcoholyes we see that there is an decrease in fatalities. This suggests that there is possibly time-invariant individual characteristics and/or Heterogeneity in individual-specific effects. We will official check this once we do the F-test below.

Fixed Effects Model

One Way Time Effects Model

```r
pdata <- pdata.frame(USSeatBelts, index = c("state", "year"))
```

```r
fixed_effects_model.time <- plm(fatalities ~ seatbelt + speed65 + alcohol + income + age,
                                data =USSeatBelts, model = "within", effect="time")

summary(fixed_effects_model.time)
```

```
## Oneway (time) effect Within Model
##
## Call:
## plm(formula = fatalities ~ seatbelt + speed65 + alcohol + income +
##     age, data = USSeatBelts, effect = "time", model = "within")
##
## Unbalanced Panel: n = 51, T = 8-15, N = 556
##
## Residuals:
##        Min.     1st Qu.      Median     3rd Qu.        Max.
## -0.00973514 -0.00231823 -0.00034668  0.00193244  0.01403237
##
## Coefficients:
##             Estimate  Std. Error  t-value  Pr(>|t|)
## seatbelt  3.1724e-03  1.3329e-03   2.3801   0.01766 *
## speed65   2.7282e-04  5.7436e-04   0.4750   0.63498
## alcohol  -1.9171e-03  4.6592e-04  -4.1148 4.486e-05 ***
## income   -8.4331e-07  5.6427e-08 -14.9451 < 2.2e-16 ***
## age      -1.1160e-04  1.1324e-04  -0.9856   0.32479
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    0.010581
## Residual Sum of Squares: 0.0066692
## R-Squared:      0.36971
## Adj. R-Squared: 0.34737
## F-statistic: 62.8812 on 5 and 536 DF, p-value: < 2.22e-16
```

This model provides us with estimates that make us doubt it's fit. The estimates show that one unit increase in seat-belt will actually increase fatalities, while alcohol will not. The problem with this model is that it doesn't take into the factor the individual effects.

One Way Individual Effects Model

```r
fixed_effects_model <- plm(fatalities ~ seatbelt + speed65 + alcohol + income + age,
                                data =USSeatBelts, model = "within")

summary(fixed_effects_model)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = fatalities ~ seatbelt + speed65 + alcohol + income +
##     age, data = USSeatBelts, model = "within")
##
## Unbalanced Panel: n = 51, T = 8-15, N = 556
```

```
##
## Residuals:
##        Min.     1st Qu.      Median     3rd Qu.        Max.
## -0.00582760 -0.00108314 -0.00018041  0.00102833  0.00714667
##
## Coefficients:
##             Estimate  Std. Error t-value Pr(>|t|)
## seatbelt -7.2974e-03  1.1251e-03 -6.4861 2.12e-10 ***
## speed65  -7.7449e-04  3.2580e-04 -2.3772 0.017820 *
## alcohol  -1.2168e-03  3.7844e-04 -3.2154 0.001387 **
## income   -4.7302e-07  6.2451e-08 -7.5743 1.76e-13 ***
## age       2.9538e-04  3.6159e-04  0.8169 0.414377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    0.005078
## Residual Sum of Squares: 0.0017099
## R-Squared:      0.66328
## Adj. R-Squared: 0.62624
## F-statistic: 196.984 on 5 and 500 DF, p-value: < 2.22e-16
```

This model doesn't make the most sense. From interpreting the estimates we see that the effect of fatality is the same with someone going 65 mile per hour and its the same for someone that wears seatbelts. The reason for the difference is that the Fixed Effects model doesn't take the time-invariant variables into factor which could cause the biased and inconsistent estimates.

Ftest

```r
pFtest(fixed_effects_model.time, pooled_model)
```

```
##
##  F test for time effects
##
## data:  fatalities ~ seatbelt + speed65 + alcohol + income + age
## F = 2.5427, df1 = 14, df2 = 536, p-value = 0.001548
## alternative hypothesis: significant effects
```

The F test for the timed fixed effects and pooled model, infers that we should reject the H_0: Pooled model. So we should use the Oneway-time Fixed Effects model

```r
pFtest(fixed_effects_model, pooled_model)
```

```
##
##  F test for individual effects
##
## data:  fatalities ~ seatbelt + speed65 + alcohol + income + age
## F = 31.595, df1 = 50, df2 = 500, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

From the F test we can conclude that we should reject the H_0: Pooled model. So we should use the Oneway-Individual Fixed Effects model

Since our model includes both the timed fixed effect and individual effect. We will use the two way model.

Twoway Effects within Fixed effects

```r
fixed_effects_model.twoway <- plm(fatalities ~ seatbelt + speed65 + alcohol + income + age,
                              data =USSeatBelts, model = "within", effect = "twoway")

summary(fixed_effects_model.twoway)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = fatalities ~ seatbelt + speed65 + alcohol + income +
##     age, data = USSeatBelts, effect = "twoway", model = "within")
##
## Unbalanced Panel: n = 51, T = 8-15, N = 556
##
## Residuals:
##        Min.     1st Qu.      Median     3rd Qu.        Max.
## -0.00461323 -0.00084171 -0.00010816  0.00074164  0.00741094
##
## Coefficients:
##             Estimate  Std. Error t-value Pr(>|t|)
## seatbelt -3.5745e-03  1.1314e-03 -3.1592 0.001681 **
## speed65  -8.1827e-04  4.1955e-04 -1.9503 0.051710 .
## alcohol  -6.3520e-04  3.4838e-04 -1.8233 0.068872 .
## income    4.0015e-07  1.4099e-07  2.8382 0.004727 **
## age       1.2219e-03  3.7574e-04  3.2519 0.001226 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    0.0013806
## Residual Sum of Squares: 0.0012768
## R-Squared:      0.075153
## Adj. R-Squared: -0.056153
## F-statistic: 7.89843 on 5 and 486 DF, p-value: 3.6264e-07
```

The model above differences across time and individuals. This is the best of the models but even then we see that speeding past 65 will decrease the likelihood of fatality more than wearing seatbelts will. This doens't make sense. We are going to model the Random Effects model and compare it with this model.

Random Effects Model

```r
pdata <- pdata.frame(USSeatBelts, index = c("state", "year"))

random_effects_model <- plm(fatalities ~ seatbelt + speed65 + alcohol + income + age,
                            data =USSeatBelts, model = "random",  effect="twoway")

summary(random_effects_model)
```

```
## Twoways effects Random Effect Model
##    (Swamy-Arora's transformation)
##
```

```
## Call:
## plm(formula = fatalities ~ seatbelt + speed65 + alcohol + income +
##     age, data = USSeatBelts, effect = "twoway", model = "random")
##
## Unbalanced Panel: n = 51, T = 8-15, N = 556
##
## Effects:
##                     var   std.dev share
## idiosyncratic 2.627e-06 1.621e-03 0.209
## individual    9.328e-06 3.054e-03 0.743
## time          6.046e-07 7.775e-04 0.048
## theta:
##            Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## id    0.8155882 0.8258055 0.8485676 0.8423791 0.8543800 0.8642431
## time  0.2308456 0.6442915 0.7197868 0.6906790 0.7197868 0.7197868
## total 0.2298993 0.6346767 0.6912813 0.6731465 0.7044729 0.7071259
##
## Residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -0.010816 -0.002906 -0.000623 -0.000202  0.002251  0.013227
##
## Coefficients:
##                Estimate  Std. Error z-value Pr(>|z|)
## (Intercept)  3.7016e-02  4.5954e+00  0.0081   0.9936
## seatbelt    -4.2509e-03  6.3168e-01 -0.0067   0.9946
## speed65     -1.1036e-03  2.2039e-01 -0.0050   0.9960
## alcohol     -1.1853e-03  2.0369e-01 -0.0058   0.9954
## income      -4.3552e-07  3.8325e-05 -0.0114   0.9909
## age         -1.4771e-04  1.3681e-01 -0.0011   0.9991
##
## Total Sum of Squares:    0.014039
## Residual Sum of Squares: 0.0082279
## R-Squared:      0.42731
## Adj. R-Squared: 0.4221
## Chisq: 0.000535645 on 5 DF, p-value: 1
```

The Random Effects model considers the time-invariant variables in the model and therefore we get more accurate results. From the interpretation of the estimates, we can tell that wearing seat-belt will best determinant of decreasing fatalities. Whereas the other estimates of the variables show that they are less likely to decrease the likelihood of fatalities. We will perform the Huasman Test to deterimine which model is best.

Perform diagnostic test

Hausman Test (Fixed Effects Model vs Random Effects Model)

```r
hausman_test <- phtest(fixed_effects_model, random_effects_model)
print(hausman_test)
```

```
##
##  Hausman Test
##
```

```
## data:  fatalities ~ seatbelt + speed65 + alcohol + income + age
## chisq = 3.5207e-05, df = 5, p-value = 1
## alternative hypothesis: one model is inconsistent
```

In Conclusion: H_0: REM H_1: FEM

Our p-value = 1, comparing it the with the significance level of 0.05, we fail to reject the H_0: REM and conclude that Random Effects Model is the best model fit for this data. This most likely suggests that there are time-invariant unobserved factors that affect the depedent variable, which is fatality in our case. All in all, Random Effects model provides the best and most efficient for our data.

Q2 Binary Dependent Variables

(a) Briefly discuss the question you are trying to answer.

We are trying to answer whether a person's credit card application will be accepted or rejected based on these 5 factors and they are: number of major derogatory reports(reports), their age(age), their income(income), whether they own a home or not(owner), and the number of dependents they have(dependents).

card: is the dependent variable. It signifies whether the application for credit card was accepted or rejected

owner: is an indicator variable. it signifies whether the applicant owns a home or not.

reports: is a continuous variable. it signifies how many major derogatory reports is against the applicant

age: is a continuous variable. it signifies the age of the owner plus twelfths of a year

income: is a continuous variable. it signifies the yearly income(in USD 10,000) of the applicant.

dependents: is a continuous variable. it signifies the number of dependents the applicant has.

Source:

The CreditCard dataset can be found in the AER package. Main Reference: Greene, W.H. (2003). Econometric Analysis, 5th edition. Upper Saddle River, NJ: Prentice Hall.

This dataset consists of Cross-Section data on the credit history for a sample of applicants for a type of credit card. The data frame contains 1,319 observations on 12 variables.

(b) Descriptive Analysis of Variables

```r
sum(is.na(CreditCard))
```

```
## [1] 0
```

```r
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
```

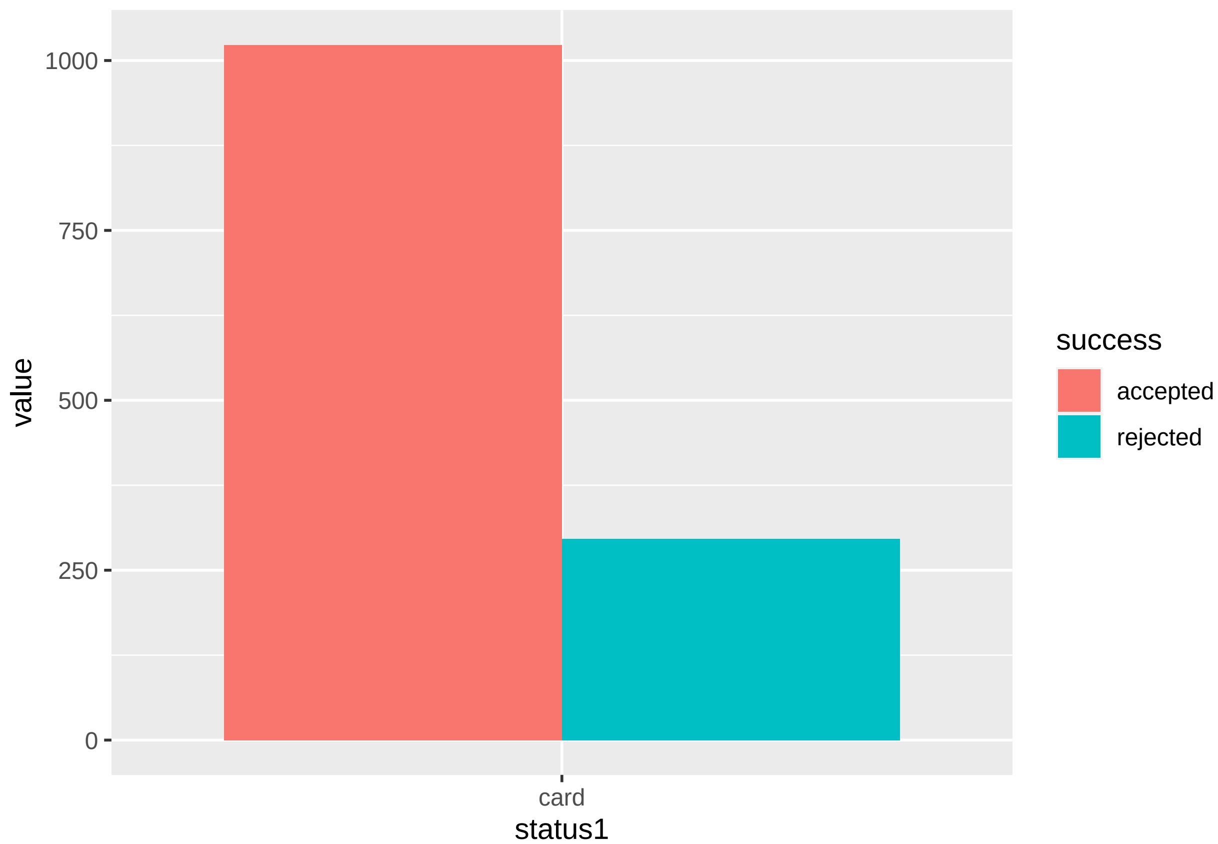
```
##      share             expenditure        owner     selfemp      dependents
##  Min.   :0.0001091   Min.   :   0.000   no :738   no :1228   Min.   :0.0000
##  1st Qu.:0.0023159   1st Qu.:   4.583   yes:581   yes:  91   1st Qu.:0.0000
##  Median :0.0388272   Median : 101.298                        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
```

```
CreditCard_vars <- CreditCard[, c("card","reports","age","income","owner","dependents")]
summary(CreditCard_vars)
```

```
##   card          reports              age               income          owner
##  no : 296   Min.   : 0.0000   Min.   : 0.1667   Min.   : 0.210   no :738
##  yes:1023   1st Qu.: 0.0000   1st Qu.:25.4167   1st Qu.: 2.244   yes:581
##             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
##    dependents
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :1.0000
##  Mean   :0.9939
##  3rd Qu.:2.0000
##  Max.   :6.0000
```

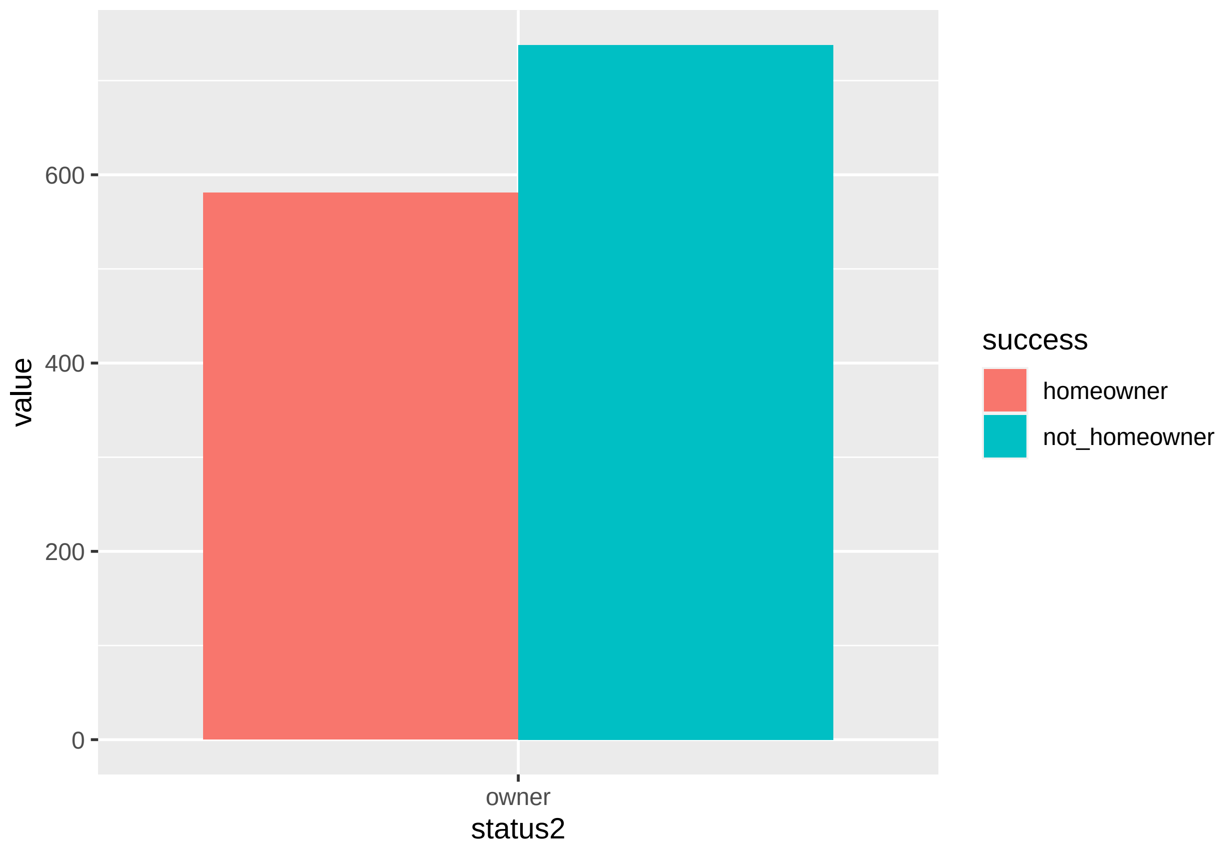
Histograms

```
status1 <- c("card")
rejected <- c(296)
accepted <- c(1023)

tata <- data.frame(status1, rejected, accepted)

tata %>%
  gather(key="success", value = value, -status1) %>%
  ggplot(aes(y = value, x= status1, fill=success)) +
  geom_bar(position = "dodge", stat = "identity")
```



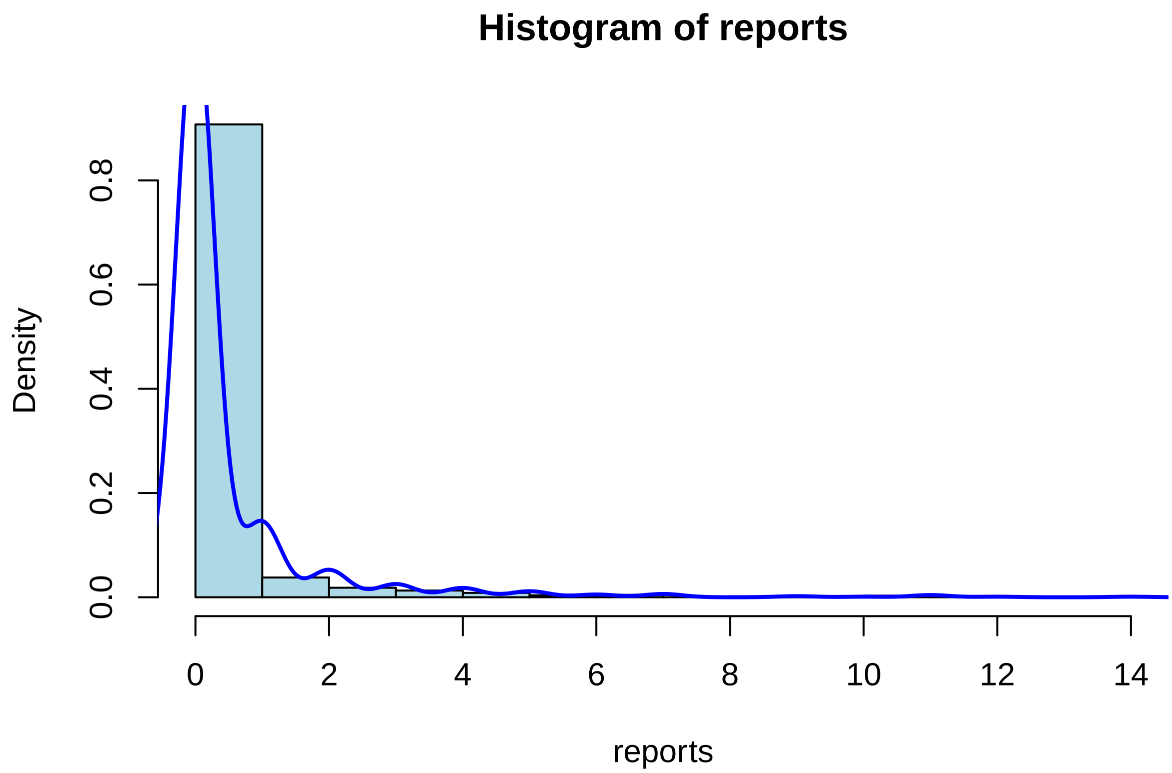
The data shows that most credit card application from the sample of applicants were accepted.

```
status2 <- c("owner")
not_homeowner <- c(738)
homeowner <- c(581)

tata2 <- data.frame(status2, not_homeowner, homeowner)

tata2 %>%
  gather(key="success", value = value, -status2) %>%
  ggplot(aes(y = value, x= status2, fill=success)) +
  geom_bar(position = "dodge", stat = "identity")
```



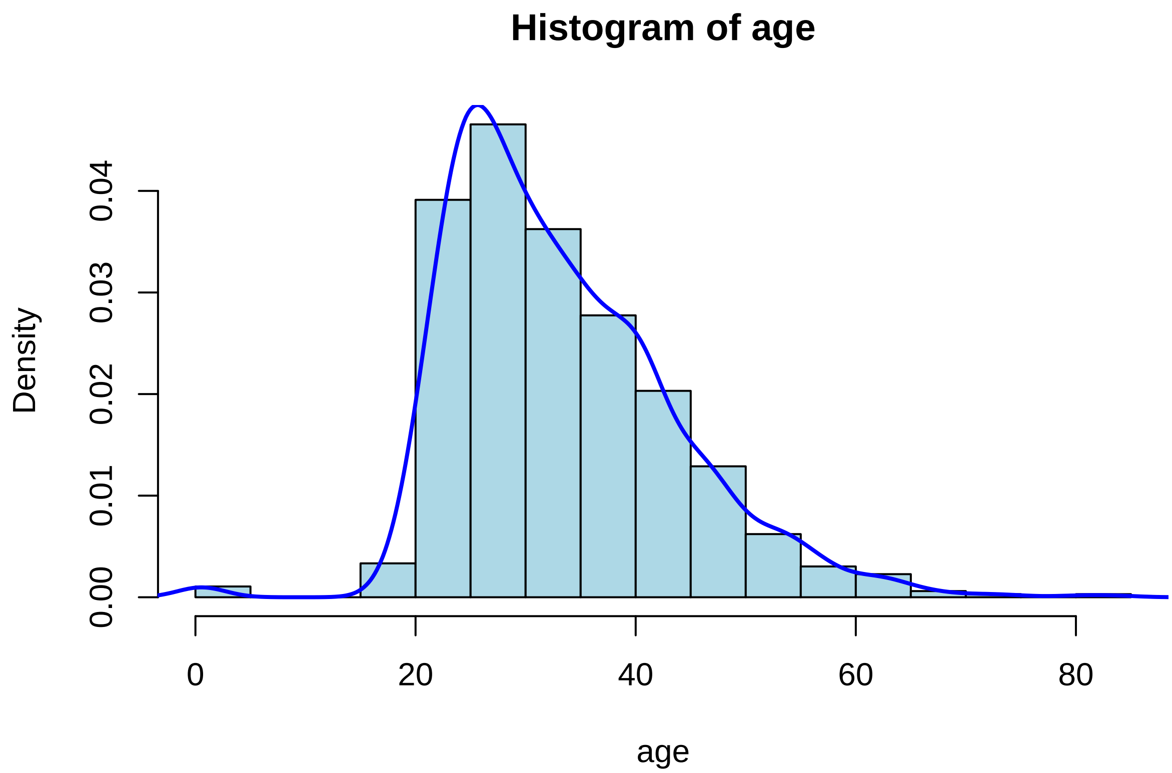
This shows that there's a higher percentage of applicants that do not own a home. From this we can sort of deduce that homeownership doesn't play a large part on whether a credit card application will be accepted or not because it is almost a split between the pool of applicants on whether they own a home or not, and we see that most creditcard applications are accepted.

```
hist(CreditCard[,"reports"], prob = TRUE, col="lightblue", main = "Histogram of reports",
     xlab = "reports")
lines(density(CreditCard[,"reports"]), col = "blue", lwd=2)
```



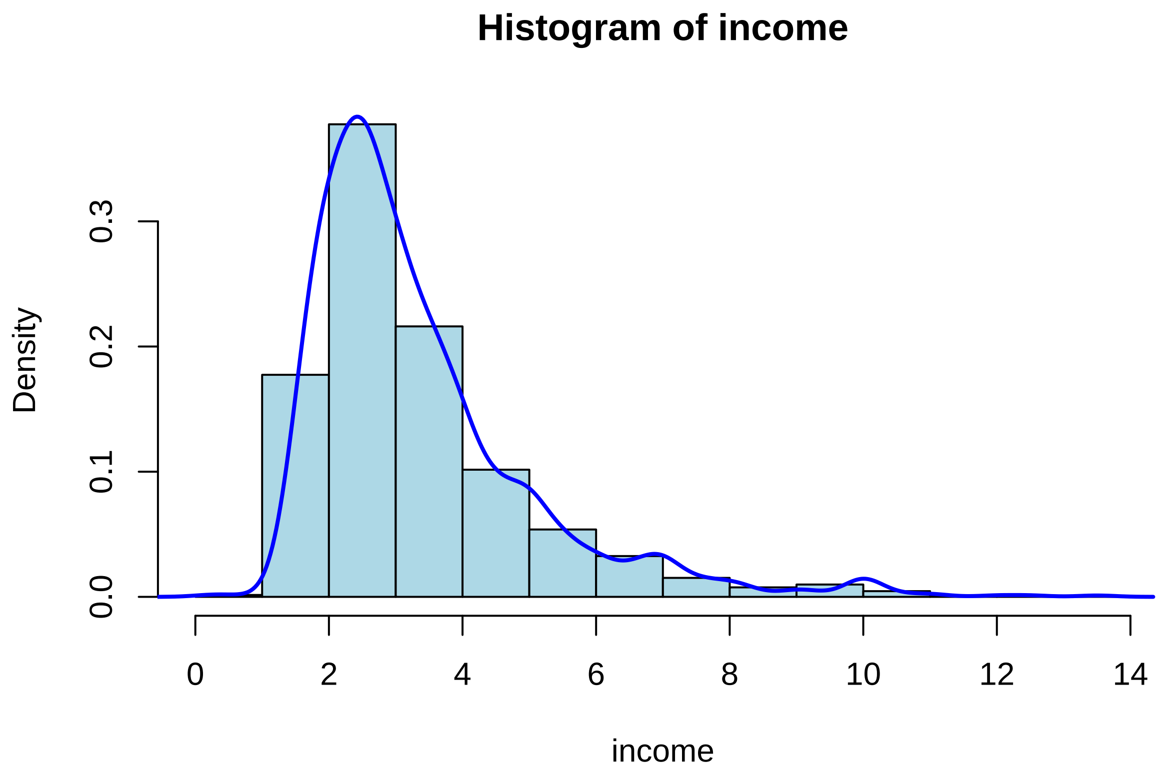
The number of major derogatory reports were very much skewed right. From this histogram we can slightly intuitively infer that the amount of reports affects whether application for creditcard is accepted or not, since both variables are at the extreme end of each other.

```
hist(CreditCard[,"age"], prob = TRUE, col="lightblue", main = "Histogram of age",
     xlab = "age")
lines(density(CreditCard[,"age"]), col = "blue", lwd=2)
```



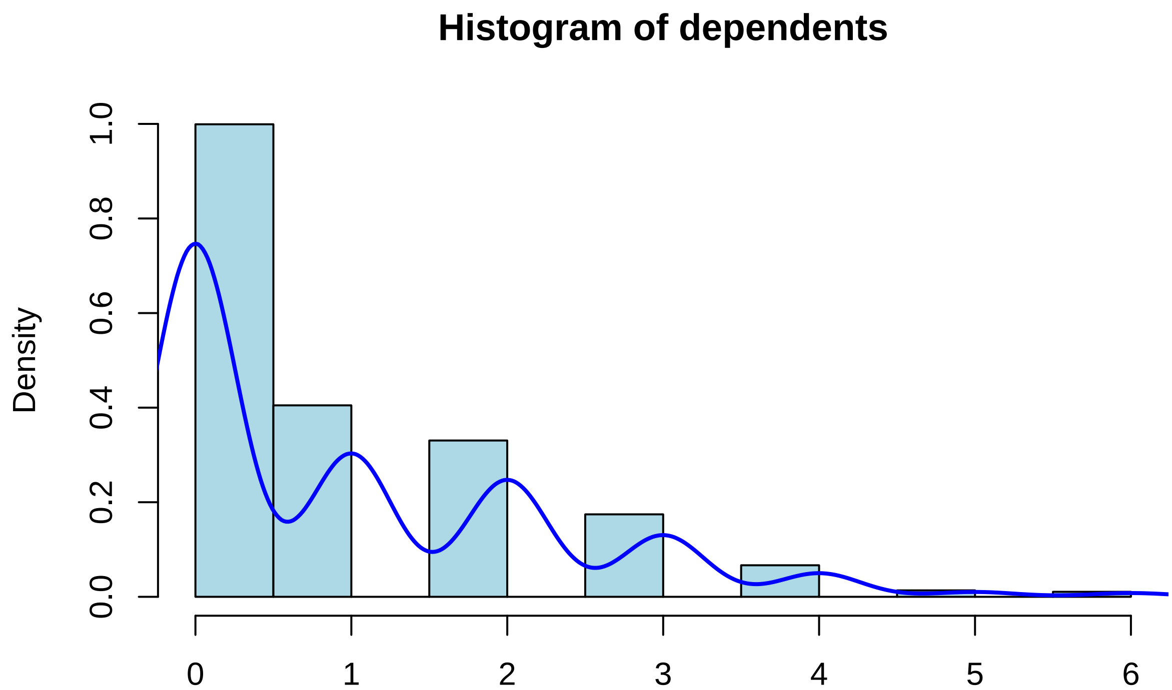
The histogram closely resembles a bell curve. The age of the applicants were around the age of 30. The oldest of the applicants were around their 80's.

```
hist(CreditCard[,"income"], prob = TRUE, col="lightblue", main = "Histogram of income",
     xlab = "income")
lines(density(CreditCard[,"income"]), col = "blue", lwd=2)
```



The histogram looks to be skewed right, because we have outliers that make a lot more than the average group of people. The yearly income of the applicants were around 20,000 dollar to 30,000. The outliers make about 100,000 per year.

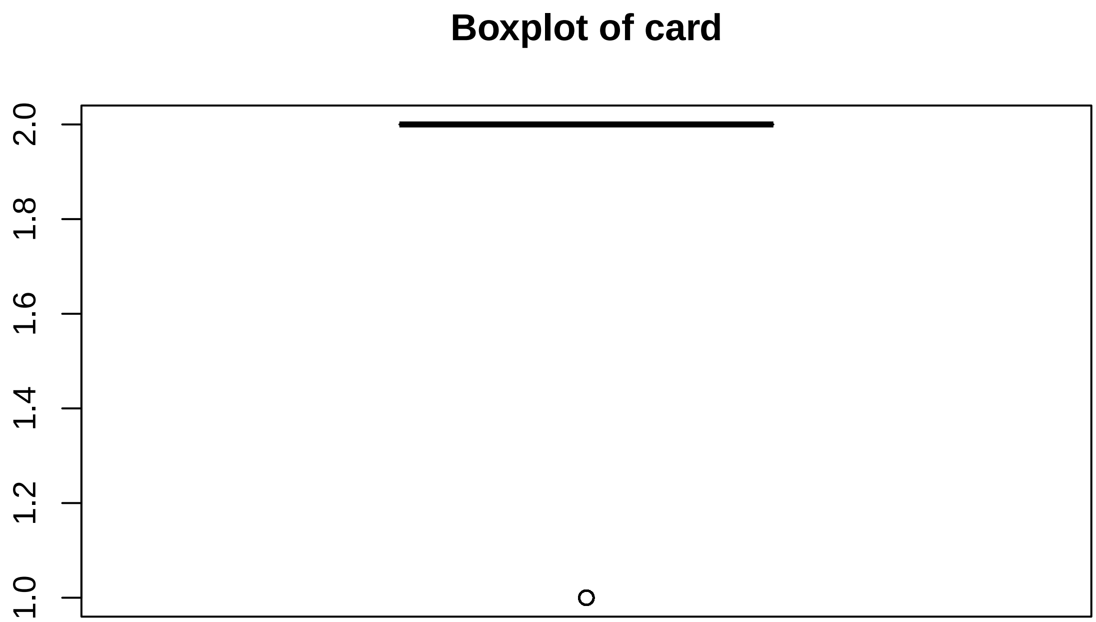
```
hist(CreditCard[,"dependents"], prob = TRUE, col="lightblue", main = "Histogram of dependents",
     xlab = "")
lines(density(CreditCard[,"dependents"]), col = "blue", lwd=2)
```



The histogram is skewed right. Dependents usually means children so it makes sense it is skewed left. Most people don't tend to have more than 1 or 2 children. The max amount of dependents were 6 which is the outlier.

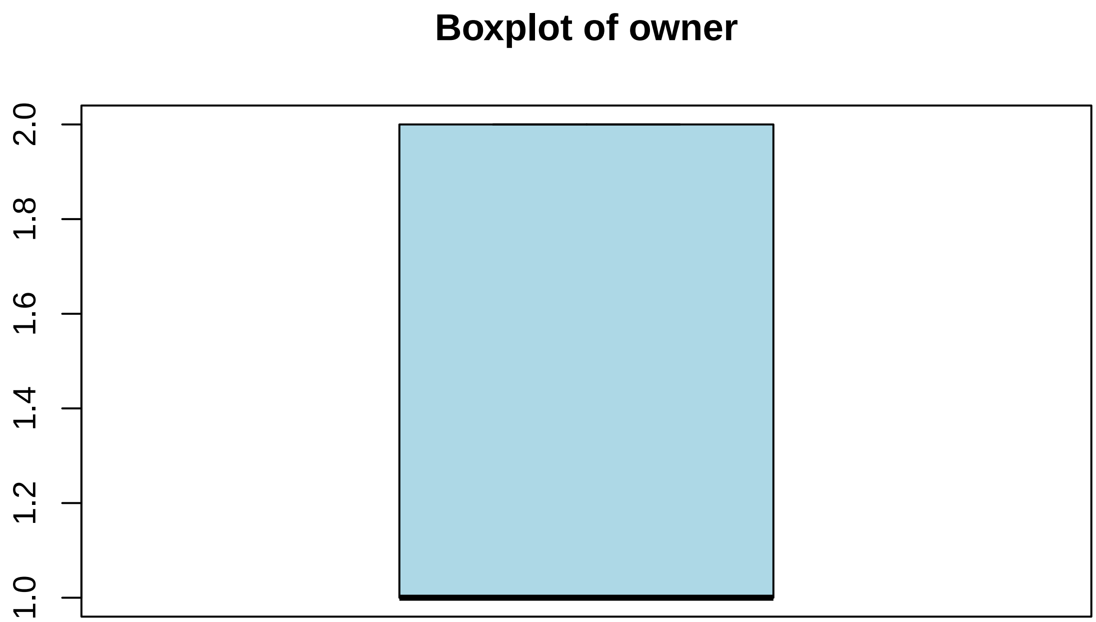
Boxplots

```r
boxplot(CreditCard[,"card"], main = "Boxplot of card", col="lightblue")
```



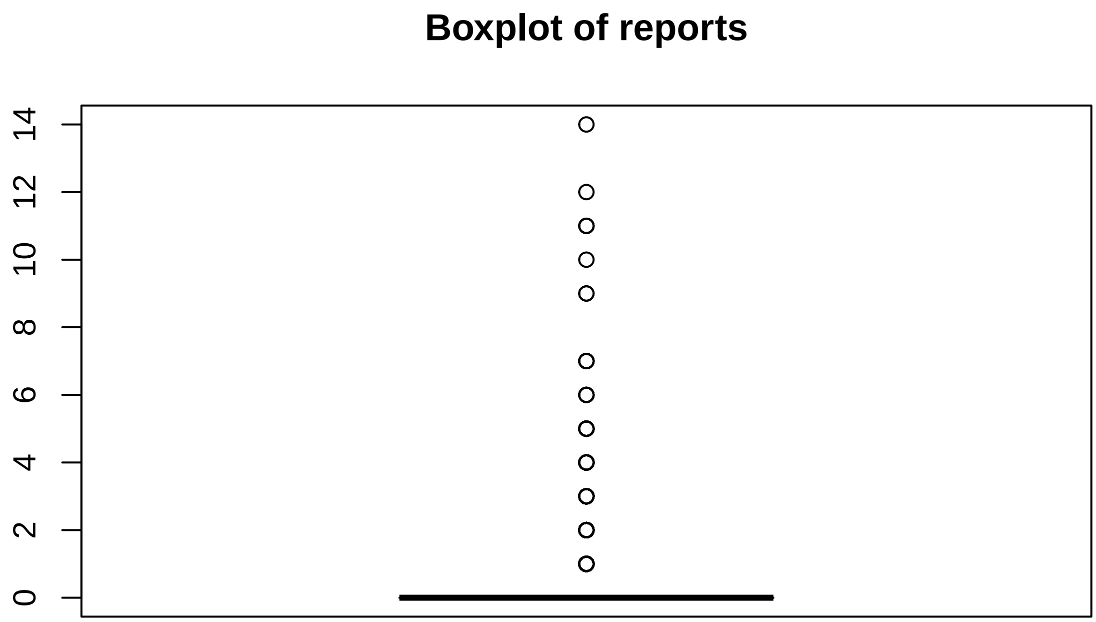
The totally squeezed boxplot suggests that the IQR and whiskers are very short. It means there is very low variability, since the data is highly concentrated in a narrow range. This suggests that most creditcard application are accepted.It suggests a negatively skewed distribution.

```
boxplot(CreditCard[,"owner"], main = "Boxplot of owner", col="lightblue")
```



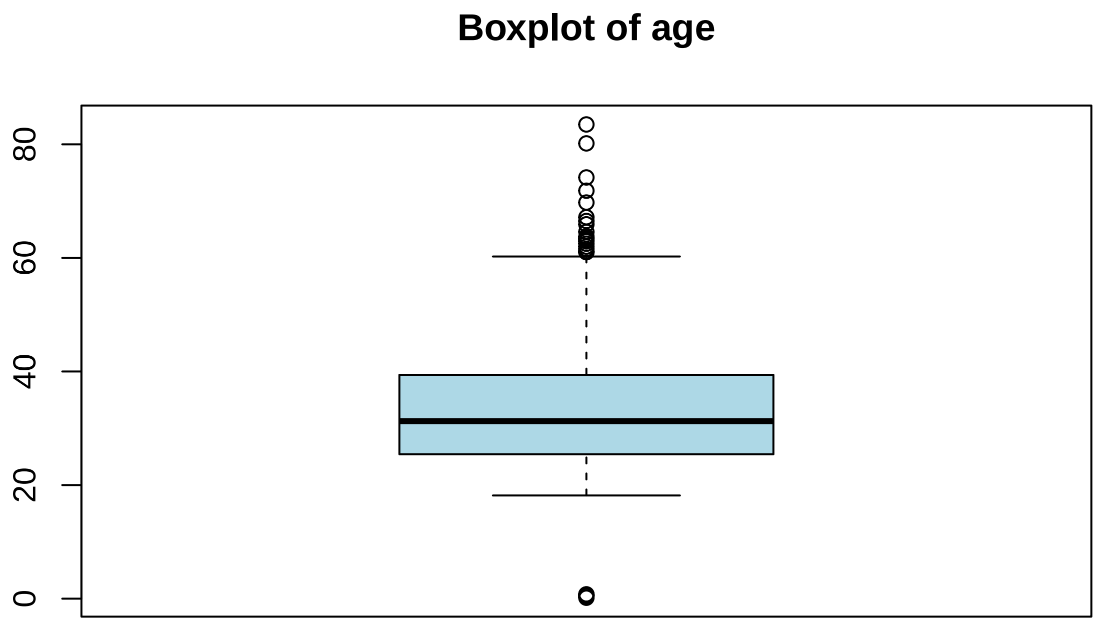
The box plot shows that the IQR is very large, which suggests there is a good amount of spread and variability. The median is in the extreme end, which suggests that the distribution is positively skewed.

```r
boxplot(CreditCard[,"reports"], main = "Boxplot of reports", col="lightblue")
```



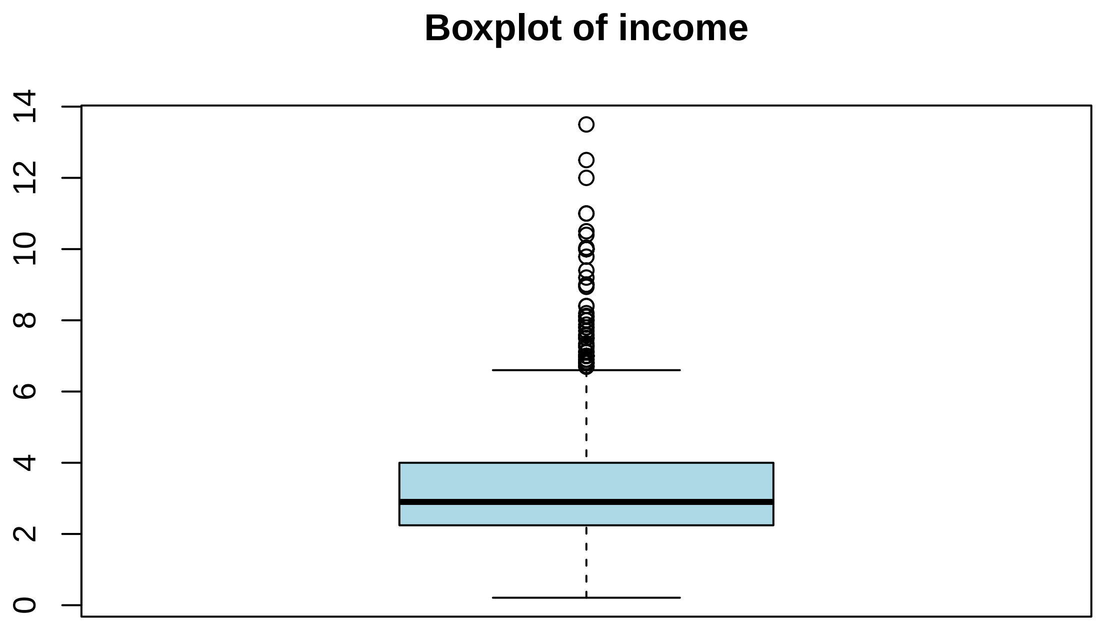
The totally squeezed boxplot suggests that the IQR and whiskers are very short. It means there is very low variability, since the data is highly concentrated in a narrow range. This suggests that the number of major derogatory reports are very low. The median is towards the bottom extreme which suggests a positively skewed distribution.

```
boxplot(CreditCard[,"age"], main = "Boxplot of age", col="lightblue")
```



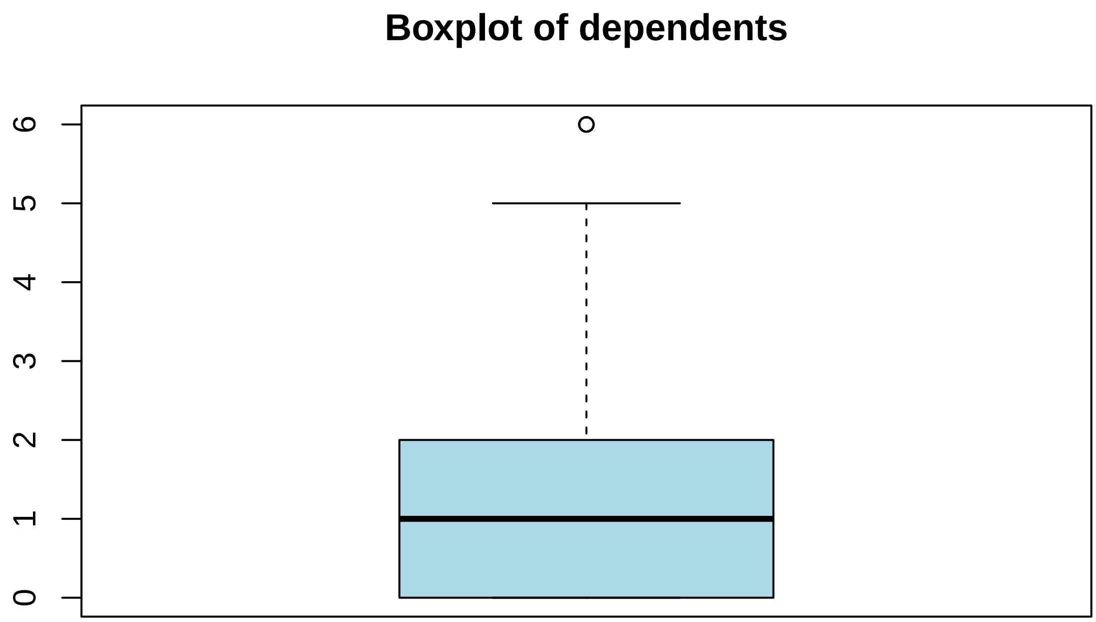
The IQR/the box is around the range of 20 to 40 with the median being around 30, this suggests that there is low variability among age. There whisker is longer on the upper end which means there are outliers that are older in age.

```
boxplot(CreditCard[,"income"], main = "Boxplot of income", col="lightblue")
```



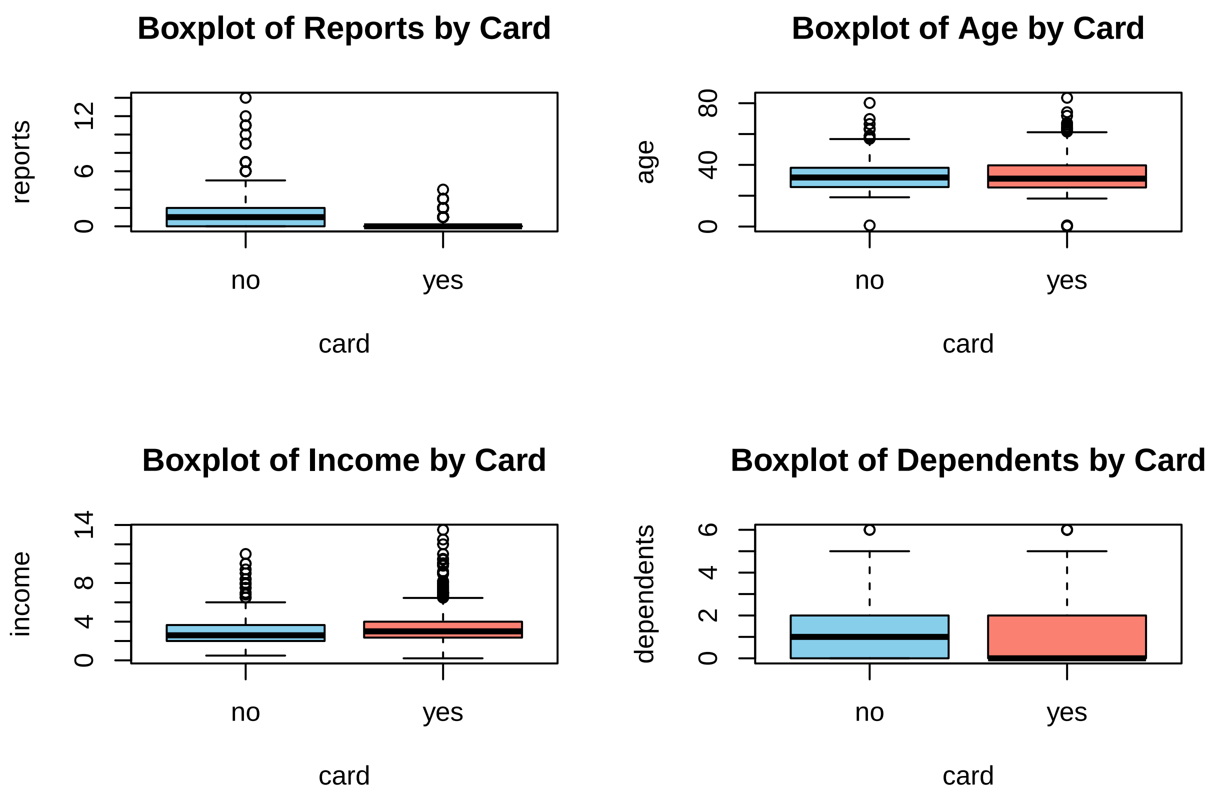
The IQR is pretty narrow which falls between the 20,000 to 40,000 range, this suggests there is low variability. There are potential outliers starting from the 60,000 to 100,000 range. The whiskers range from 0 to 60,000.

```
boxplot(CreditCard[,"dependents"], main = "Boxplot of dependents", col="lightblue")
```

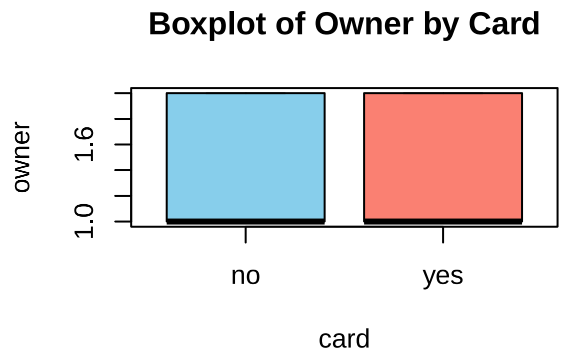


The IQR ranges from 0 to 2 dependents, while the median is set around 1 dependents. This means there is low variability among the dependents. The whisker ranges up to 5. There is a potential outlier at 6 dependents.

```
par(mfrow = c(2, 2))
boxplot(reports ~ card, data = CreditCard, main = "Boxplot of Reports by Card",
        col = c("skyblue", "salmon"))
boxplot(age ~ card, data = CreditCard, main = "Boxplot of Age by Card",
        col = c("skyblue", "salmon"))
boxplot(income ~ card, data = CreditCard, main = "Boxplot of Income by Card",
        col = c("skyblue", "salmon"))
boxplot(dependents ~ card, data = CreditCard, main = "Boxplot of Dependents by Card",
        col = c("skyblue", "salmon"))
```




```r
boxplot(owner ~ card, data = CreditCard, main = "Boxplot of Owner by Card",
        col = c("skyblue", "salmon"))
```



For Boxplot of Reports by Card

We see that the median of application being accepted when the the number of report is close to 0. The IQR is very narrow but still towards the bottom of the range. This suggests that the likelihood of the card being rejected increases as the number of reports increases.

For the Boxplot of Age by Card

We see that the correlation between Age and whether the credit card application is not very closely related. The IQR is around the same age so is the median for both situation where the application is accepted or rejected.

For the Boxplot of Income by Card

The relation between Income and whether application is accepted is slightly related. We see that the IQR and median for the accepted application is a bit higher as income increases and the IQR and median is slightly lower as the income decreases. We also see that the outliers for the accepted application is much higher with outlier of income. Which means higher income increases the likelihood of the application being accepted. The reason for the outliers is that the data is not extensive enough to include people with higher income applicants, if we had enough data on higher income individuals we would see that the application being accepted greatly increases with the increase in income.

For the Boxplot of Dependents by Card

We can see that the median for the application being accepted is toward the bottom which suggests that lower amount of dependents increases the likelihood of the application being accepted. The IQR range is the same for both meaning the variability of whether the application is accepted or not based on dependents is around the same. The median for the application being rejected is slightly higher, which slightly suggests that as the number of dependents increase so the likelihood of the application being rejected.

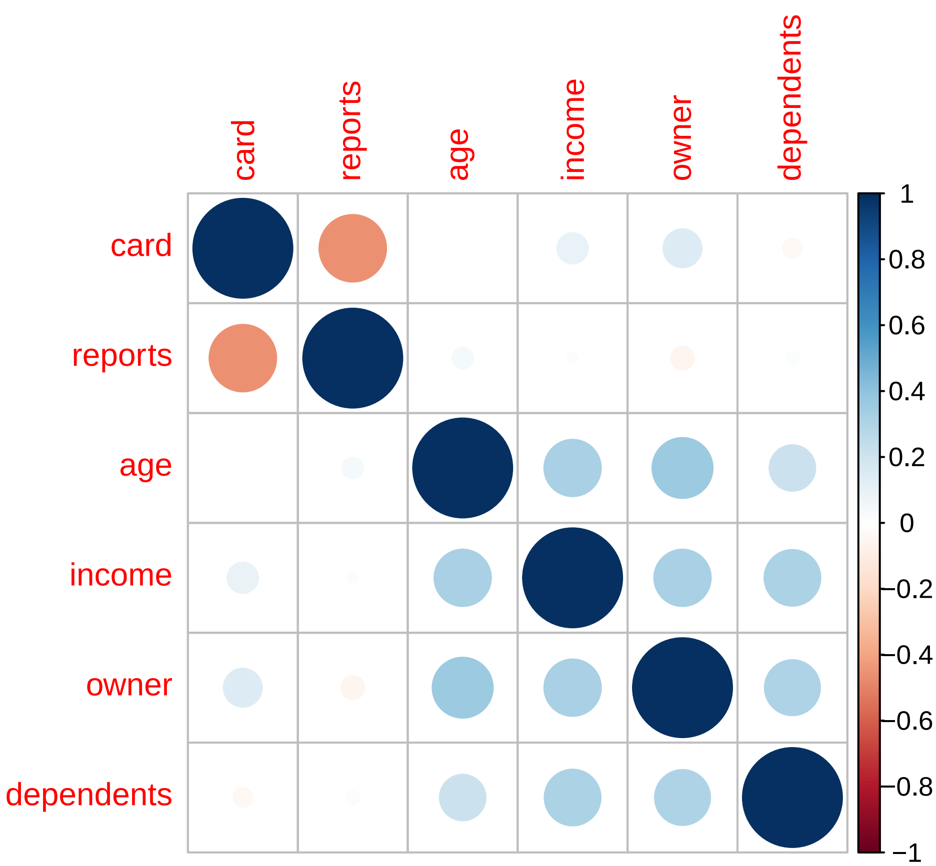
For the Boxplot of Owner by Card

The IQR is both very large and around the same. Which means there is large variability. The median is both towards the bottom end, this suggests that the ownership of home might not seriously affect whether your application is rejected or accepted.

```r
CreditCard_vars_numeric <- as.data.frame(lapply(CreditCard_vars, as.numeric))
matrix <- cor(CreditCard_vars_numeric)
print(matrix)
```

```
##                     card     reports          age     income       owner
## card        1.0000000000 -0.45257686 0.0005368538 0.09430752  0.14782578
## reports    -0.4525768570  1.00000000 0.0440885132 0.01102287 -0.05357042
## age         0.0005368538  0.04408851 1.0000000000 0.32465320  0.36774912
## income      0.0943075202  0.01102287 0.3246531987 1.00000000  0.32477622
## owner       0.1478257752 -0.05357042 0.3677491218 0.32477622  1.00000000
## dependents -0.0361263878  0.01973090 0.2121464324 0.31760130  0.30918973
##             dependents
## card       -0.03612639
## reports     0.01973090
## age         0.21214643
## income      0.31760130
## owner       0.30918973
## dependents  1.00000000
```

```r
corrplot(matrix)
```



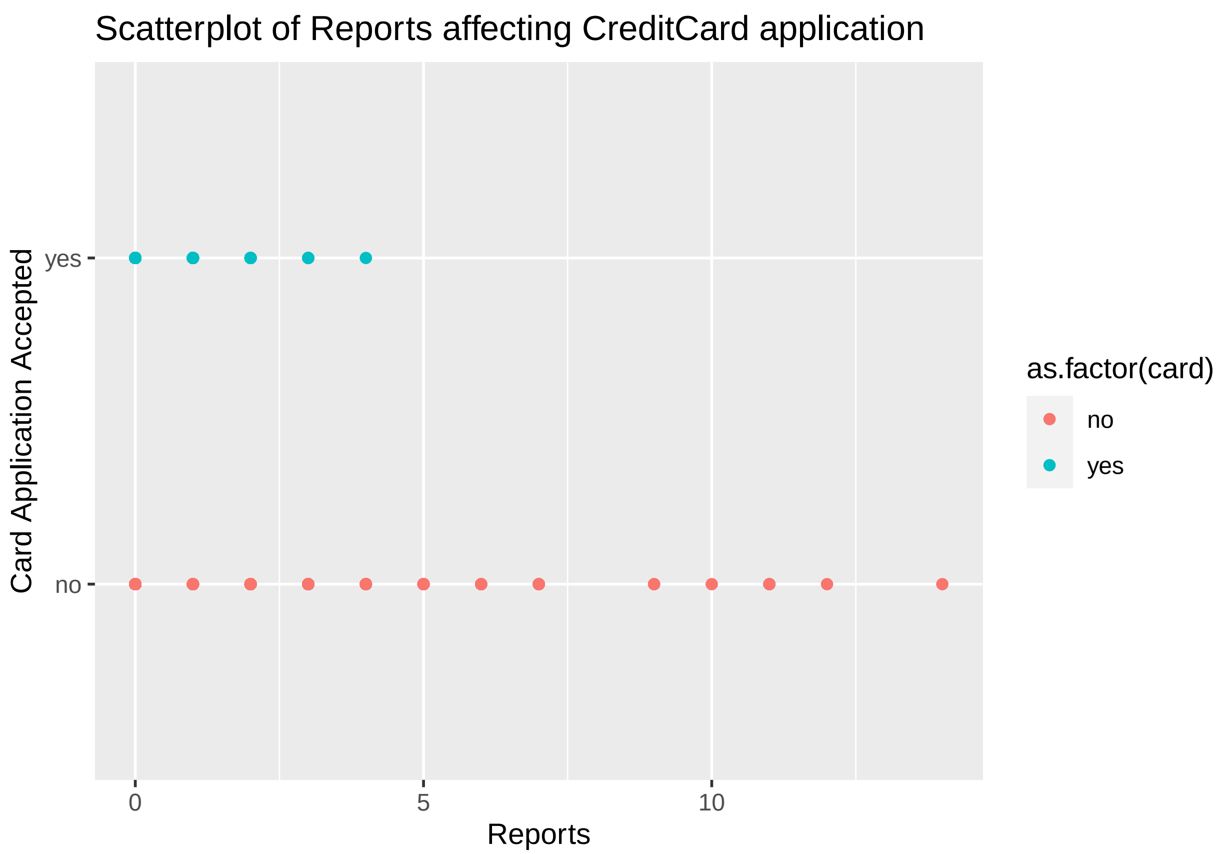
The Correlation plot shows which factors can affect whether the application is accepted or rejected. We can rank them based on how much they affect the application.

1) Reports: This is negatively correlated, since the number of reports increases the likelihood application getting rejected also increases
2) Owner: This is positively correlated, the chances of application getting accepted is higher if the applicant owns a home.
3) Income: This is slightly positively correlated. As the income of the individual increases so does the likelihood of the application being accepted.
4) dependents : This is slightly negatively correlated. The higher number the number of dependents an applicant has, the slightly lower chance they will have of getting their application getting accepted.
5) Age: is barely positively correlated. The older the applicant the likelihood of application being accepted.

Note: The data mostly consists of applicants in their 30's, and that already increases the chances of application being accepted due to the likelihood that people around that age have higher income and more likely to own a home. The data doesn't have much data on applicants that are much younger around the age of 18-20.

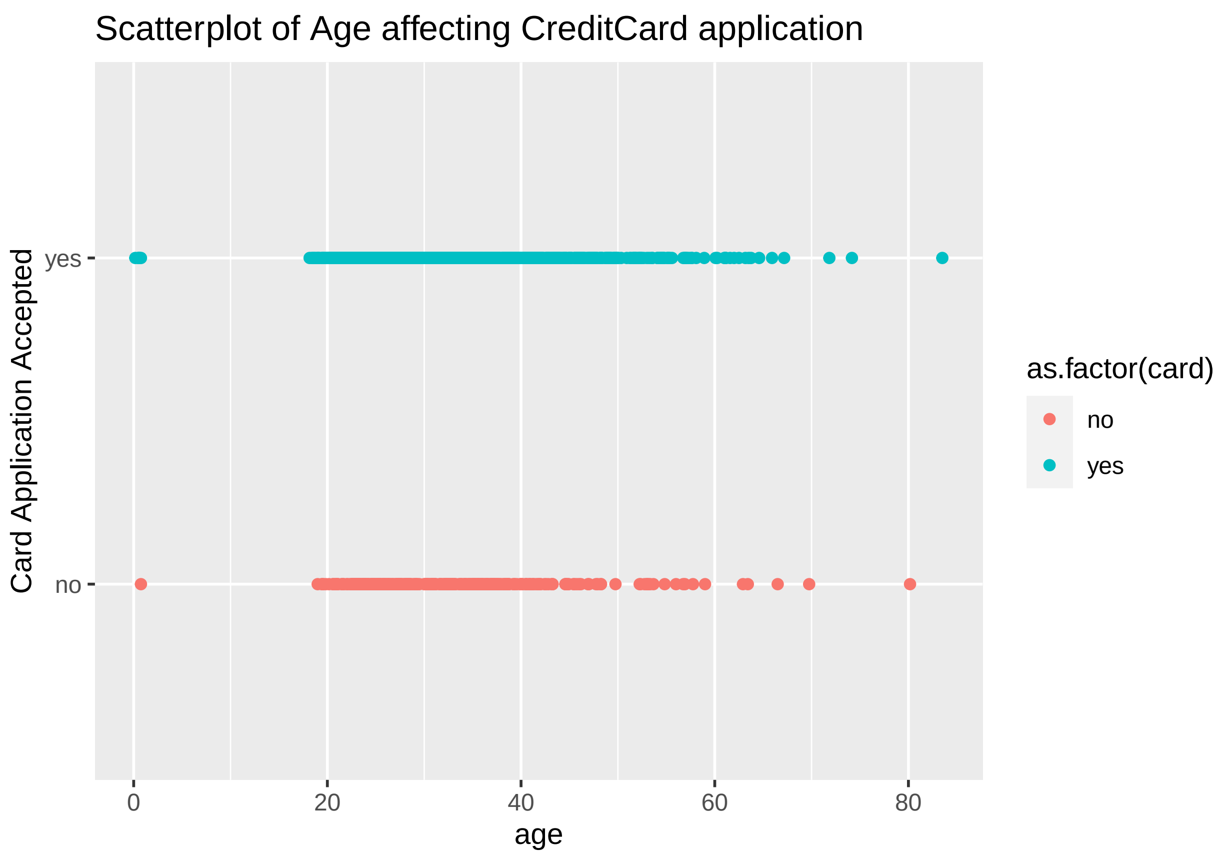
Scatterplots

```r
ggplot(CreditCard, aes(x = reports, y = card)) +
  geom_point(aes(color = as.factor(card))) +
  labs(title = "Scatterplot of Reports affecting CreditCard application",
       x = "Reports", y = "Card Application Accepted")
```



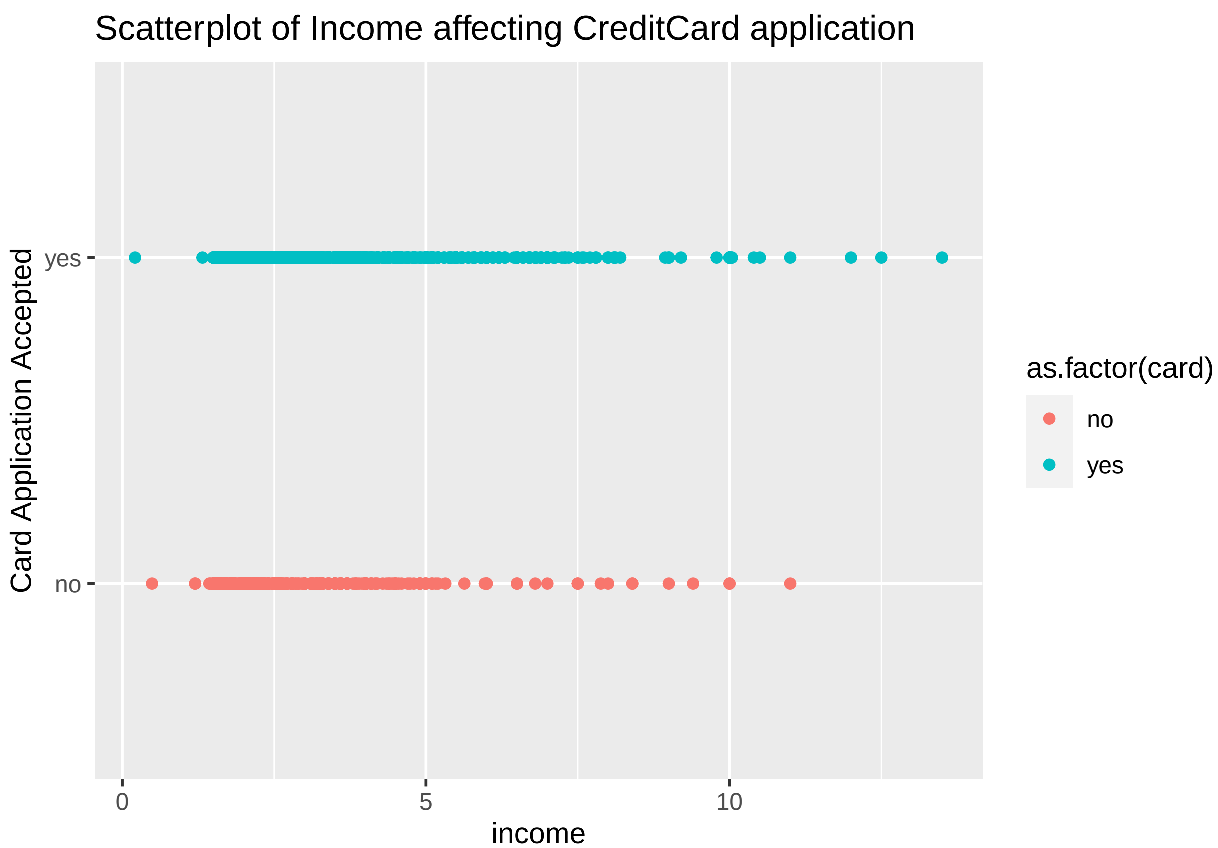
The scatterplot shows that after a certain amount of major derogatory reports the likelyhood of the application getting rejected increases by a lot.

```
ggplot(CreditCard, aes(x = age, y = card)) +
  geom_point(aes(color = as.factor(card))) +
  labs(title = "Scatterplot of Age affecting CreditCard application",
       x = "age", y = "Card Application Accepted")
```



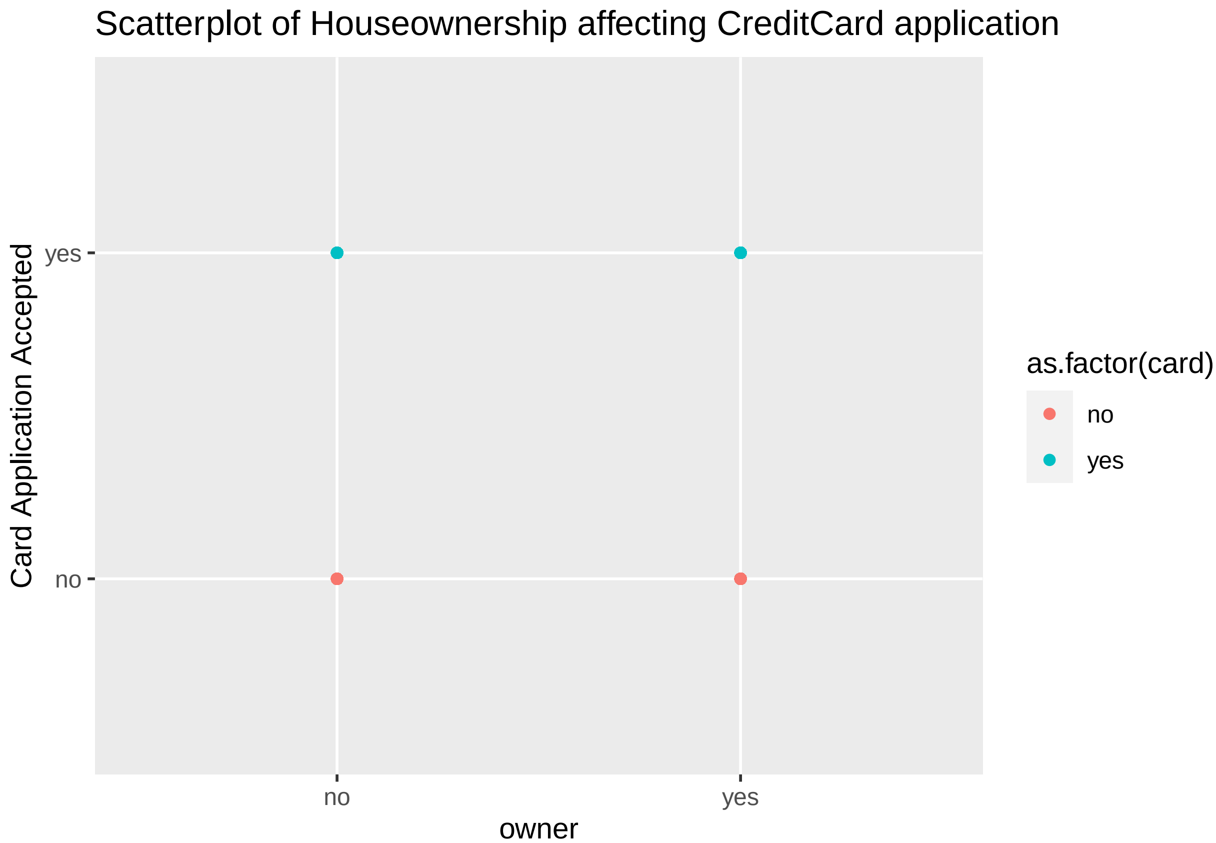
The scatterplot shows that age doesn't really play much of factor on whether the application will be accepted or not. Although we do see that the oulier age of 80 and applicants application was accepted.

```
ggplot(CreditCard, aes(x = income, y = card)) +
  geom_point(aes(color = as.factor(card))) +
  labs(title = "Scatterplot of Income affecting CreditCard application",
       x = "income", y = "Card Application Accepted")
```



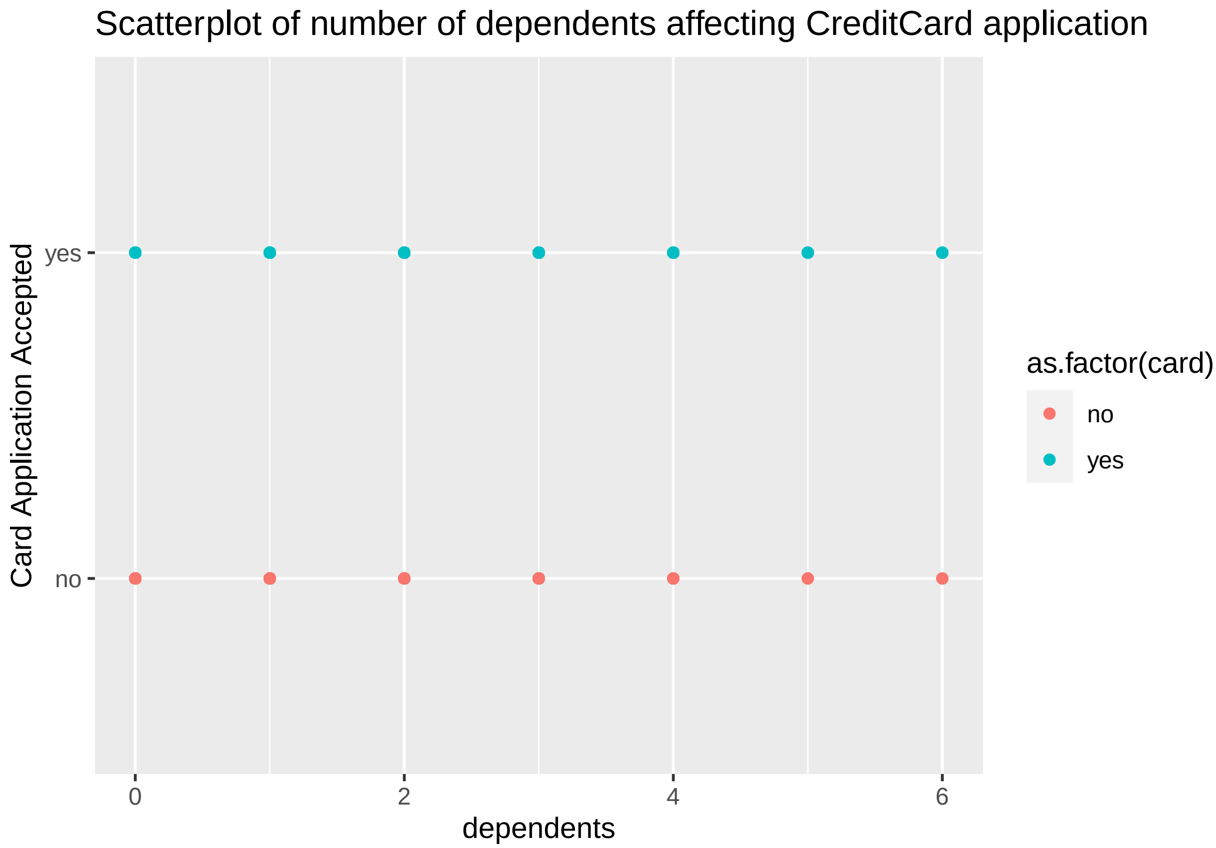
The Scatter plot shows that as income increases the likelihood of application being accepted also increases. We also see that from the income of 20,000 to 50,000 the application process isn't affected much.

```
ggplot(CreditCard, aes(x = owner, y = card)) +
  geom_point(aes(color = as.factor(card))) +
  labs(title = "Scatterplot of Houseownership affecting CreditCard application",
       x = "owner", y = "Card Application Accepted")
```



The scatterplot isn't a great representation of whether owning a house will affect the application process.Although we can infer that owning a home isn't necessary for credit card application to be accepted. From the scatterplot it seems like it won't affect it as much, but that is not what we see from the correlation matrix.

```
ggplot(CreditCard, aes(x = dependents, y = card)) +
  geom_point(aes(color = as.factor(card))) +
  labs(title = "Scatterplot of number of dependents affecting CreditCard application",
       x = "dependents", y = "Card Application Accepted")
```



The scatterplot shows that there is still possibility of application being accepted if the numnber of dependents are very high.

c) Fit the three models below, and identify which model is your preferred one and why. Make sure to include statistical diagnostics to support your conclusion, and to comment on your findings.

Linear Probability Model

First, we run a Linear Probability model. We can find the marginal effects using the command `margins(lpm)`.

```
df <- CreditCard_vars
df$cardnum <- as.numeric(df$card) - 1
lpm <- lm(cardnum ~ reports + age + income + owner + dependents, data = df)
coeftest(lpm, vcov = hccm(lpm,type="hc1"))
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)  0.7951062  0.0366631  21.6868 < 2.2e-16 ***
## reports     -0.1373363  0.0114298 -12.0156 < 2.2e-16 ***
## age         -0.0017433  0.0011051  -1.5775 0.1149115
## income       0.0240911  0.0071381   3.3750 0.0007598 ***
## owneryes     0.1145024  0.0224205   5.1070 3.755e-07 ***
```

```
## dependents  -0.0306343  0.0092924  -3.2967 0.0010045 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
margins(lpm)
```

```
## Average marginal effects

## lm(formula = cardnum ~ reports + age + income + owner + dependents,     data = df)

##   reports       age  income dependents owneryes
##   -0.1373 -0.001743 0.02409   -0.03063   0.1145
```

Probit Model

Next, we run a Probit model. We can find the marginal effects using the command `margins(mod.probit)`.

```r
mod.probit <- glm(cardnum ~ reports + age + income + owner + dependents, data = df,
                  family=binomial(link="probit"))
margins(mod.probit)
```

```
## Average marginal effects

## glm(formula = cardnum ~ reports + age + income + owner + dependents,     family = binomial(link = "pr

##   reports       age  income dependents owneryes
##   -0.1734 -0.001283 0.02744   -0.03006   0.1064
```

Logit Model

Finally, we run a Logit model. We can find the marginal effects using the command `margins(mod.logit)`.

```r
mod.logit <- glm(cardnum ~ reports + age + income + owner + dependents, data = df,
                 family=binomial(link="logit"))
margins(mod.logit)
```

```
## Average marginal effects

## glm(formula = cardnum ~ reports + age + income + owner + dependents,     family = binomial(link = "lo

##   reports      age  income dependents owneryes
##   -0.1676 -0.00127 0.03264    -0.0322   0.1024
```

Evaluating the models

We can evaluate models using AIC and BIC criteria, and select the model specification that results in the lowest AIC/BIC.

```r
AIC(lpm, mod.probit, mod.logit)
```

```
##            df      AIC
## lpm         7 1101.133
## mod.probit  6 1071.710
## mod.logit   6 1068.559
```

```r
BIC(lpm, mod.probit, mod.logit)
```

```
##            df      BIC
## lpm         7 1137.425
## mod.probit  6 1102.818
## mod.logit   6 1099.667
```

The lowest AIC and BIC is given by the Logit Model. Thus, we prefer the logit model in explaining the credit card probability.