This is my quick data analytical on the Credit data. Balance is the Dependent Variable and others are Feature

```
library(ISLR)
data(Credit)
summary(Credit)
```

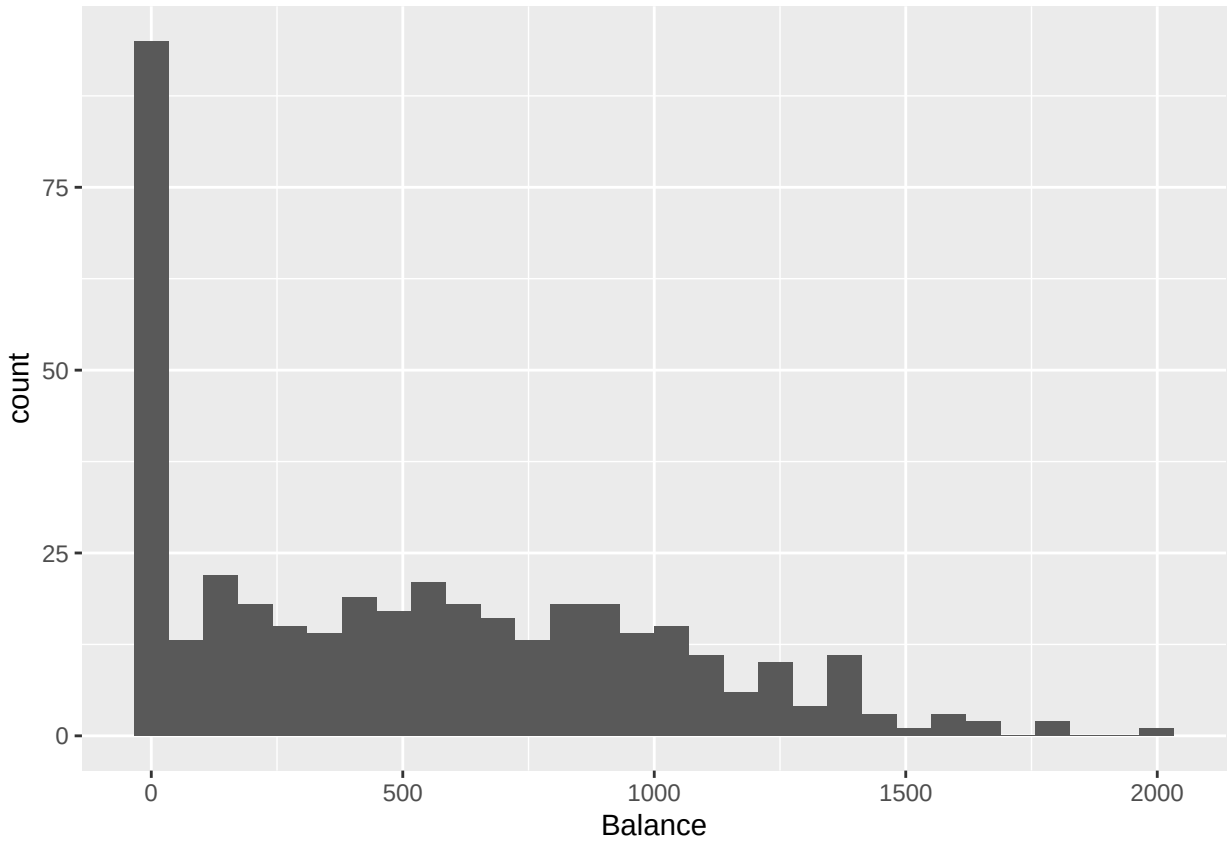
```
##        ID             Income           Limit           Rating
##  Min.   :  1.0   Min.   : 10.35   Min.   :  855   Min.   : 93.0
##  1st Qu.:100.8   1st Qu.: 21.01   1st Qu.: 3088   1st Qu.:247.2
##  Median :200.5   Median : 33.12   Median : 4622   Median :344.0
##  Mean   :200.5   Mean   : 45.22   Mean   : 4736   Mean   :354.9
##  3rd Qu.:300.2   3rd Qu.: 57.47   3rd Qu.: 5873   3rd Qu.:437.2
##  Max.   :400.0   Max.   :186.63   Max.   :13913   Max.   :982.0
##      Cards             Age          Education        Gender    Student
##  Min.   :1.000   Min.   :23.00   Min.   : 5.00    Male :193   No :360
##  1st Qu.:2.000   1st Qu.:41.75   1st Qu.:11.00   Female:207   Yes: 40
##  Median :3.000   Median :56.00   Median :14.00
##  Mean   :2.958   Mean   :55.67   Mean   :13.45
##  3rd Qu.:4.000   3rd Qu.:70.00   3rd Qu.:16.00
##  Max.   :9.000   Max.   :98.00   Max.   :20.00
##  Married              Ethnicity      Balance
##  No :155   African American: 99   Min.   :   0.00
##  Yes:245   Asian           :102   1st Qu.:  68.75
##            Caucasian       :199   Median : 459.50
##                                   Mean   : 520.01
##                                   3rd Qu.: 863.00
##                                   Max.   :1999.00
```

```
 # deleting ID column
Credit$ID <- NULL

 # Exploratory Data Analysis

library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.2
```

```
ggplot(data = Credit, aes(x = Balance)) +
  geom_histogram()
```

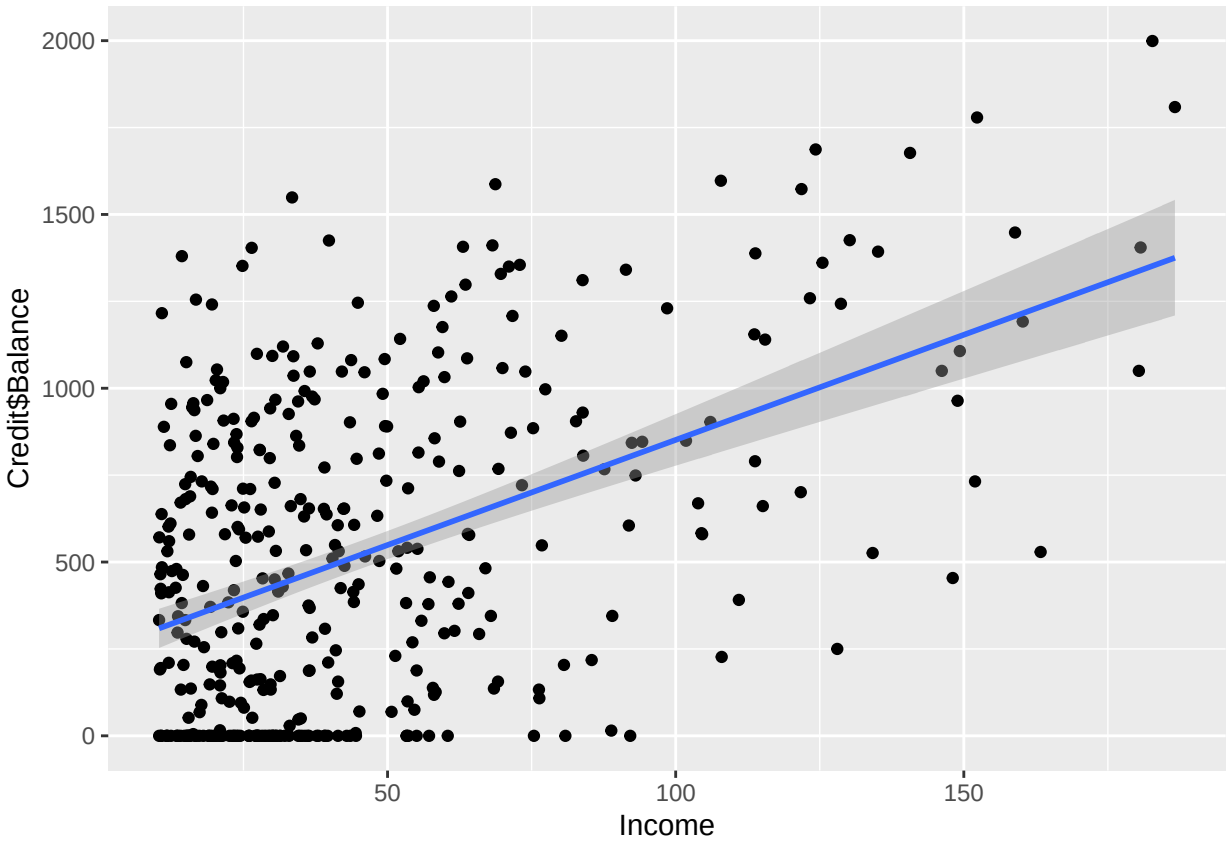
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



For Numeric variables

```
vars.numeric <- colnames(Credit[ ,c(1:6)])

for (i in vars.numeric) {
  plot <- ggplot(data = Credit, aes(x = Credit[ ,i], y = Credit$Balance)) +
    geom_point() +
    geom_smooth(method = "lm") +
    labs(x = i)
    print(plot)
}
```

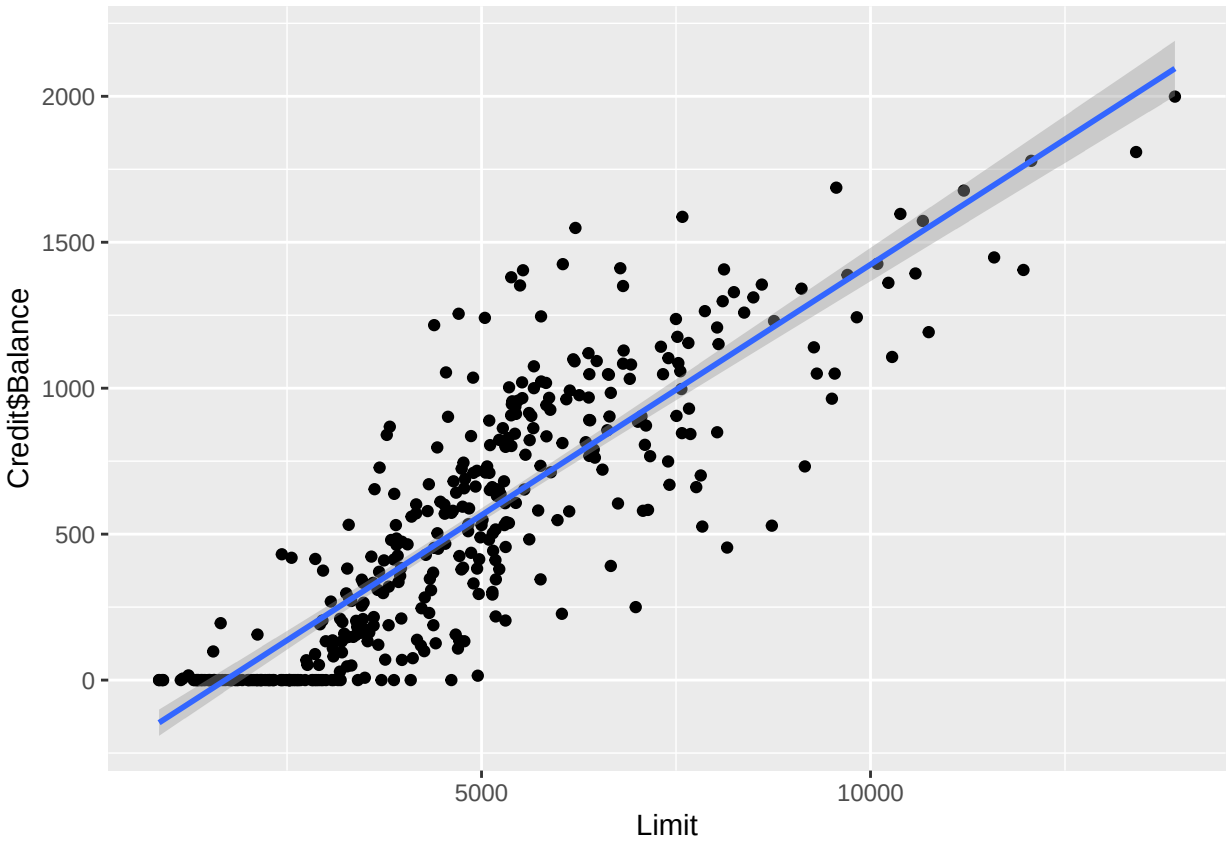
```
## Warning: Use of `Credit$Balance` is discouraged. Use `Balance` instead.

## Warning: Use of `Credit$Balance` is discouraged. Use `Balance` instead.

## `geom_smooth()` using formula 'y ~ x'
```




```
## Warning: Use of `Credit$Balance` is discouraged. Use `Balance` instead.
```

```
## Warning: Use of `Credit$Balance` is discouraged. Use `Balance` instead.
```

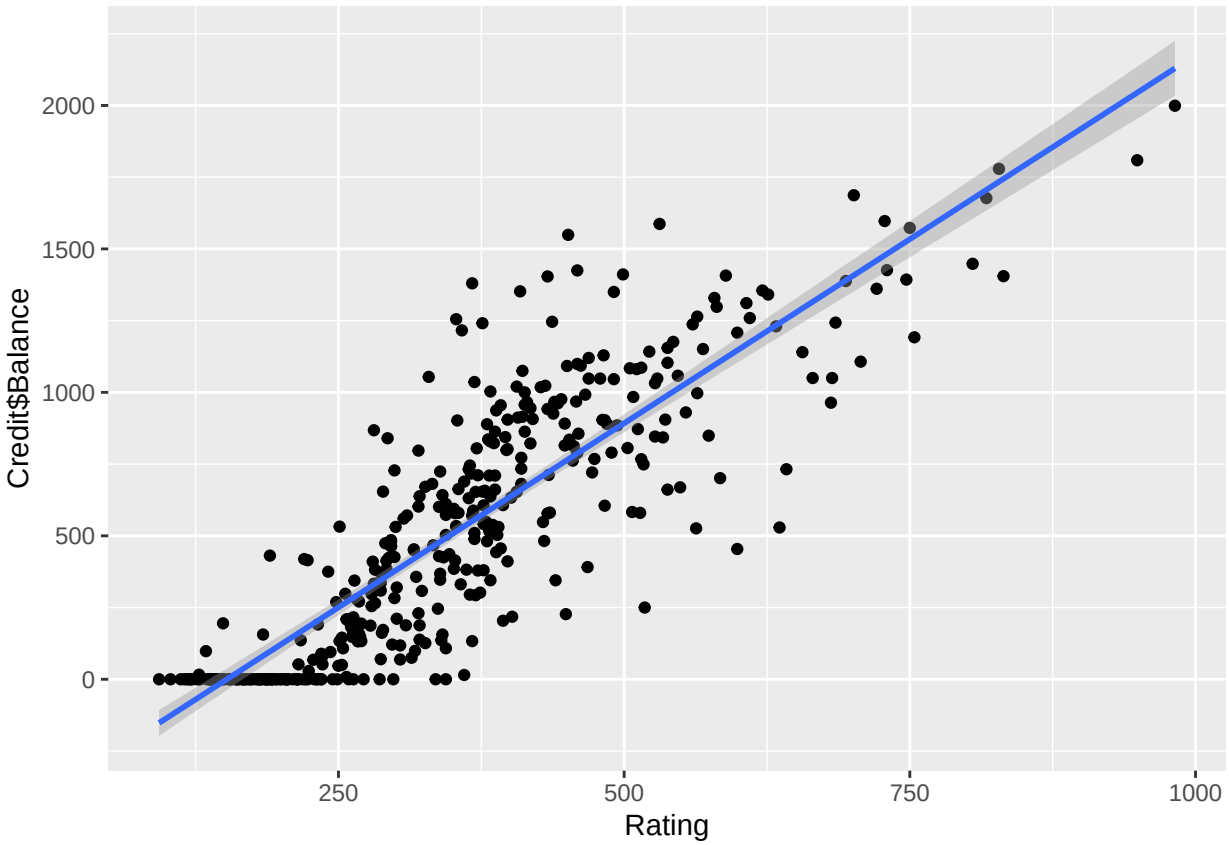
```
## `geom_smooth()` using formula 'y ~ x'
```




```
## Warning: Use of `Credit$Balance` is discouraged. Use `Balance` instead.
```

```
## Warning: Use of `Credit$Balance` is discouraged. Use `Balance` instead.
```

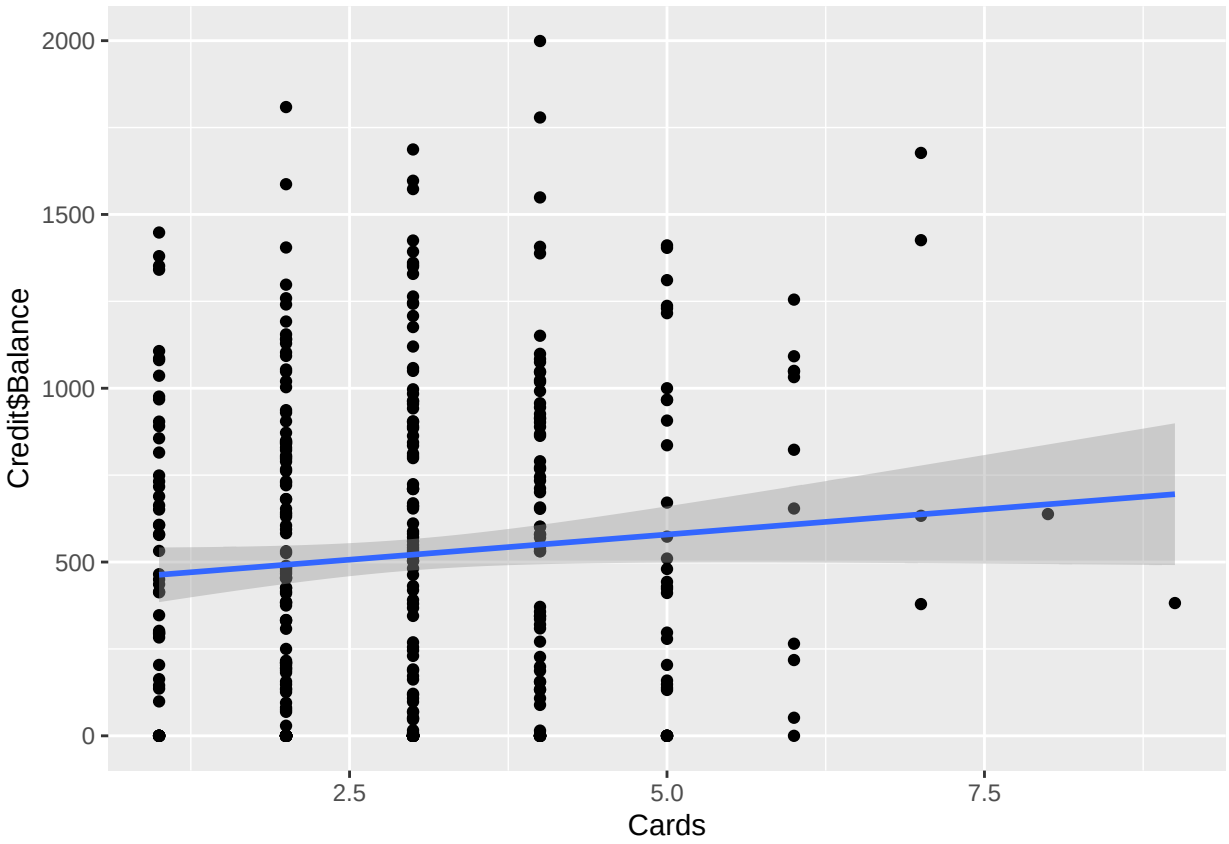
```
## `geom_smooth()` using formula 'y ~ x'
```




```
## Warning: Use of `Credit$Balance` is discouraged. Use `Balance` instead.
```

```
## Warning: Use of `Credit$Balance` is discouraged. Use `Balance` instead.
```

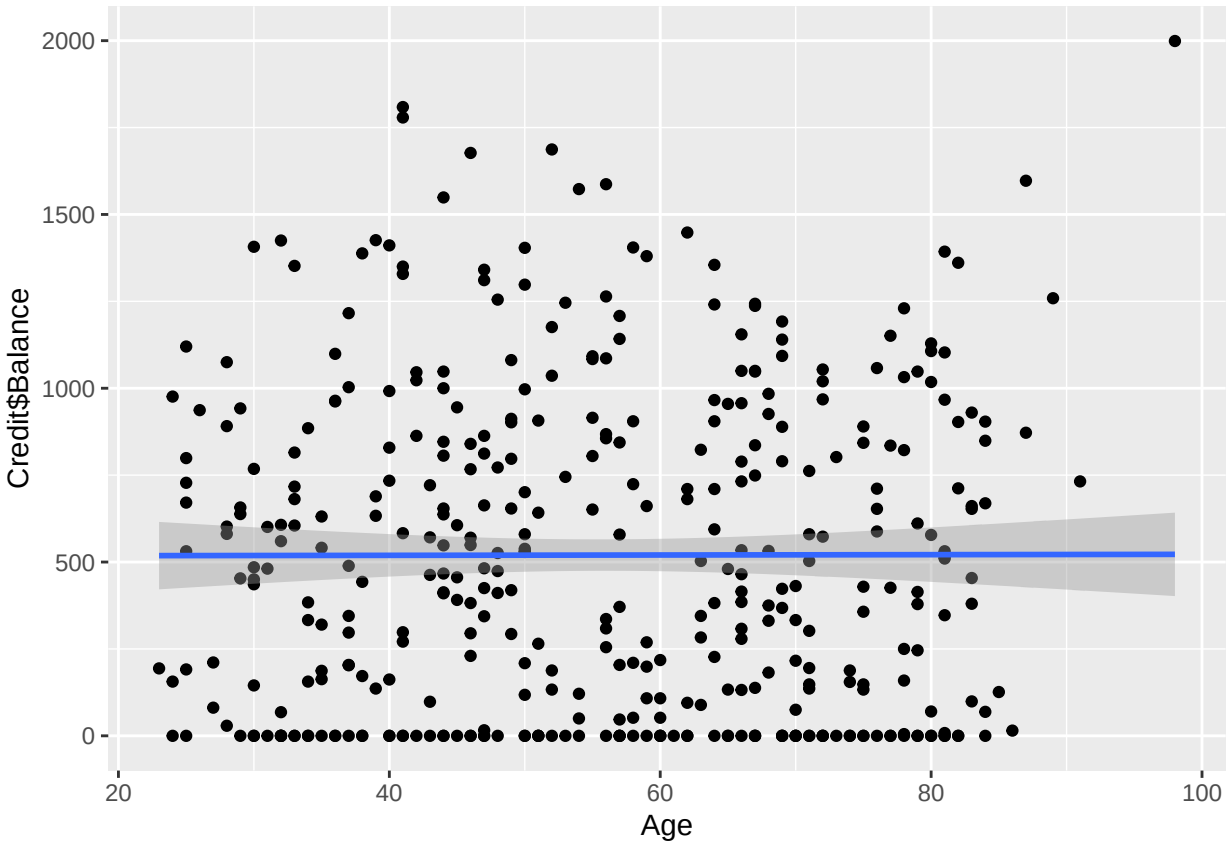
```
## `geom_smooth()` using formula 'y ~ x'
```



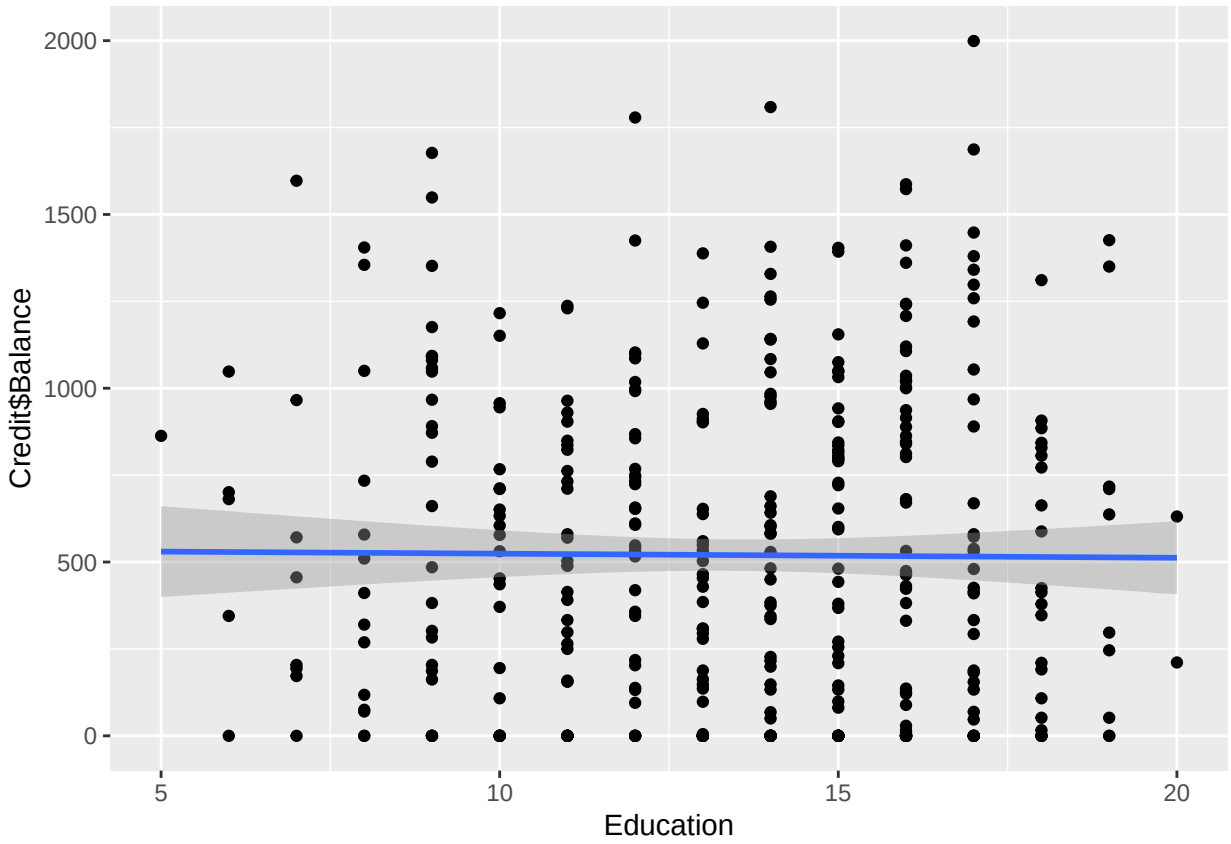

```
## Warning: Use of `Credit$Balance` is discouraged. Use `Balance` instead.
```

```
## Warning: Use of `Credit$Balance` is discouraged. Use `Balance` instead.
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
## Warning: Use of `Credit$Balance` is discouraged. Use `Balance` instead.

## Warning: Use of `Credit$Balance` is discouraged. Use `Balance` instead.

## `geom_smooth()` using formula 'y ~ x'
```

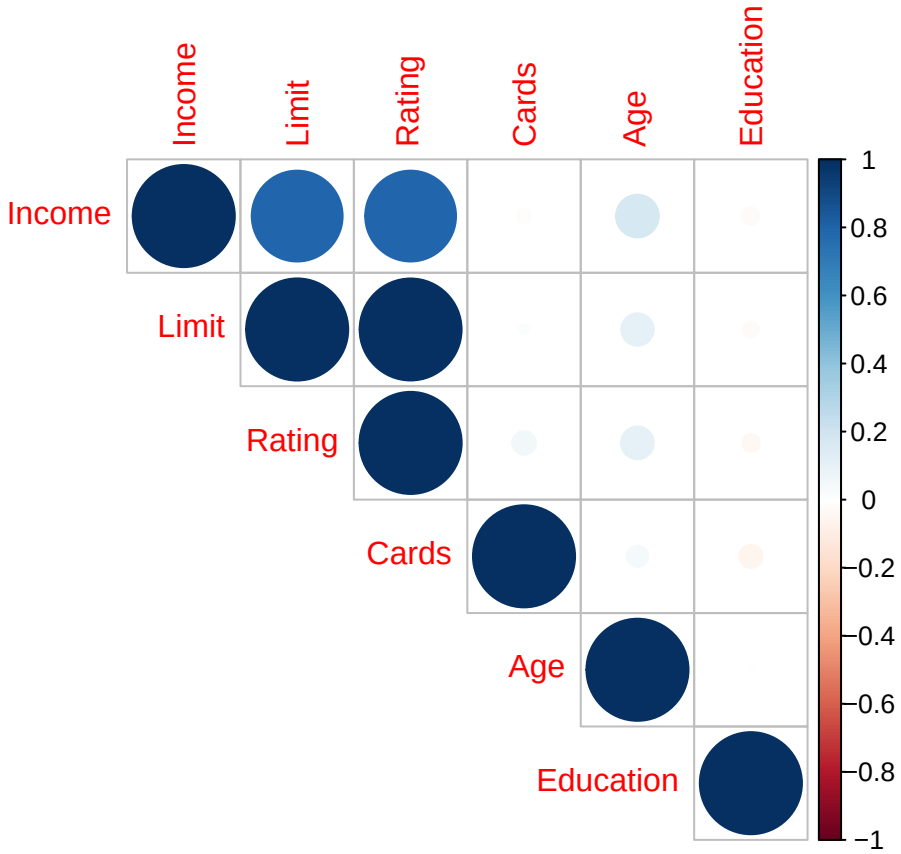


CORRELATION ANALYSIS BETWEEN NUMERICAL DATA

```
cortable <- cor(Credit[, vars.numeric])
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
corrplot(cortable, type = "upper")
```




```
cortable
```

```
##                 Income       Limit      Rating       Cards         Age
## Income      1.00000000  0.79208834  0.79137763 -0.01827261 0.175338403
## Limit       0.79208834  1.00000000  0.99687974  0.01023133 0.100887922
## Rating      0.79137763  0.99687974  1.00000000  0.05323903 0.103164996
## Cards      -0.01827261  0.01023133  0.05323903  1.00000000 0.042948288
## Age         0.17533840  0.10088792  0.10316500  0.04294829 1.000000000
## Education -0.02769198 -0.02354853 -0.03013563 -0.05108422 0.003619285
##               Education
## Income     -0.027691982
## Limit      -0.023548534
## Rating     -0.030135627
## Cards      -0.051084217
## Age         0.003619285
## Education   1.000000000
```

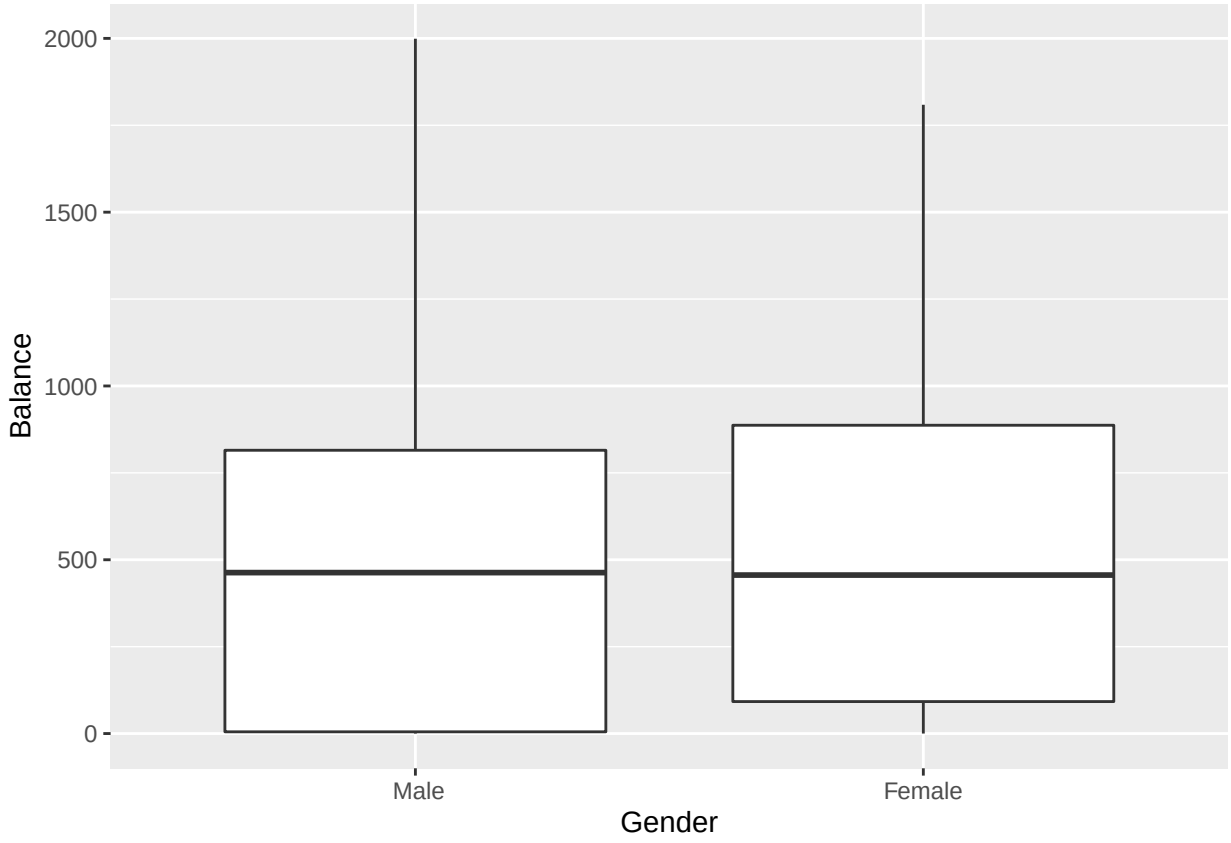
A strong correlation betwween Limit and Rating detected. To prevent collinearity, we can delete one of two variable

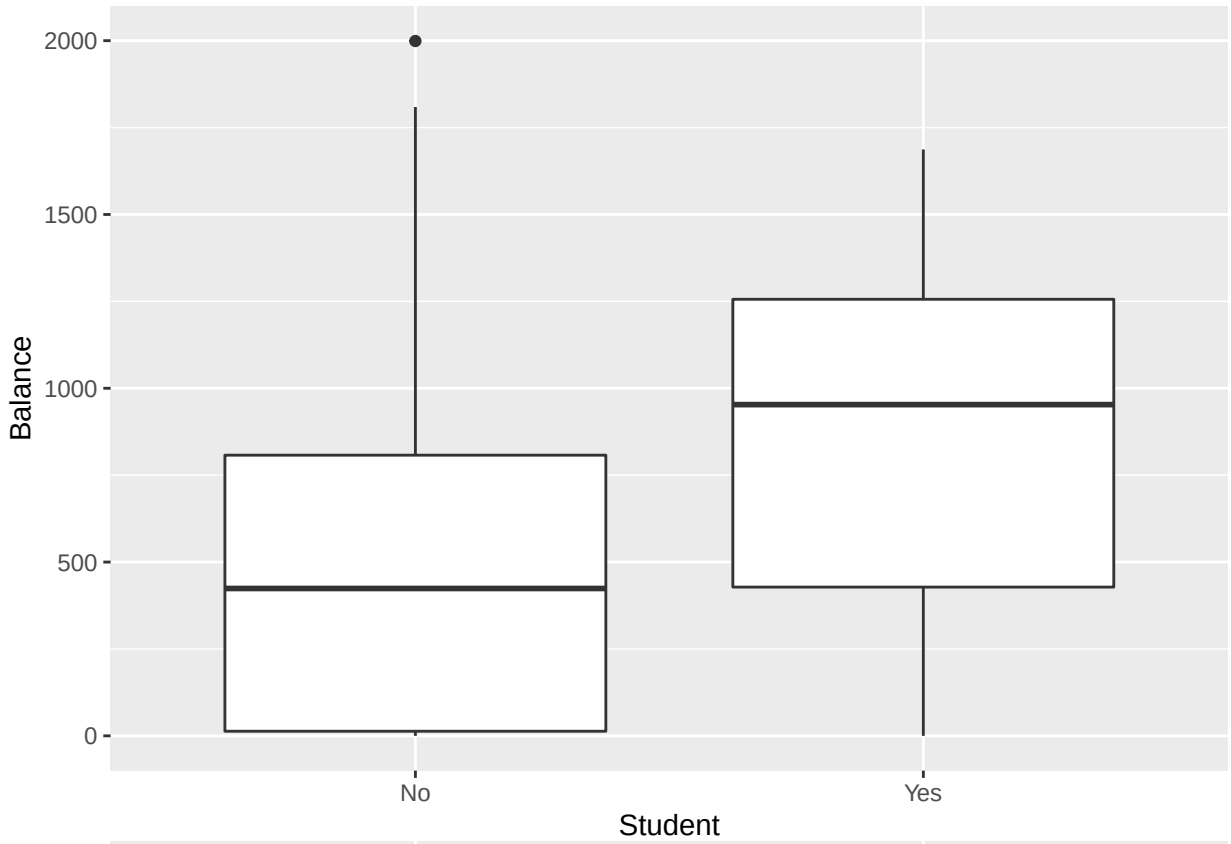
```
Credit$Limit <- NULL
```

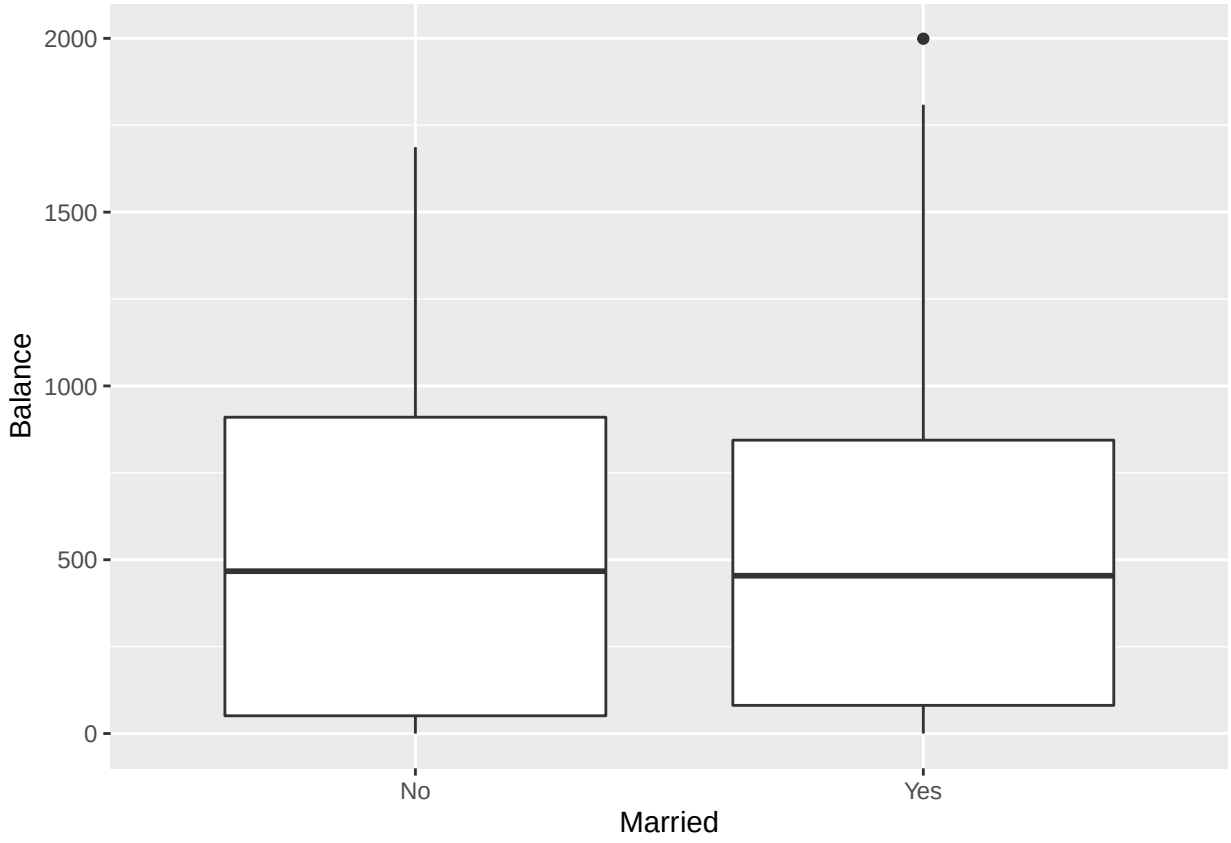
TARGET VARIABLE AND CATEGORICAL VARIABLE

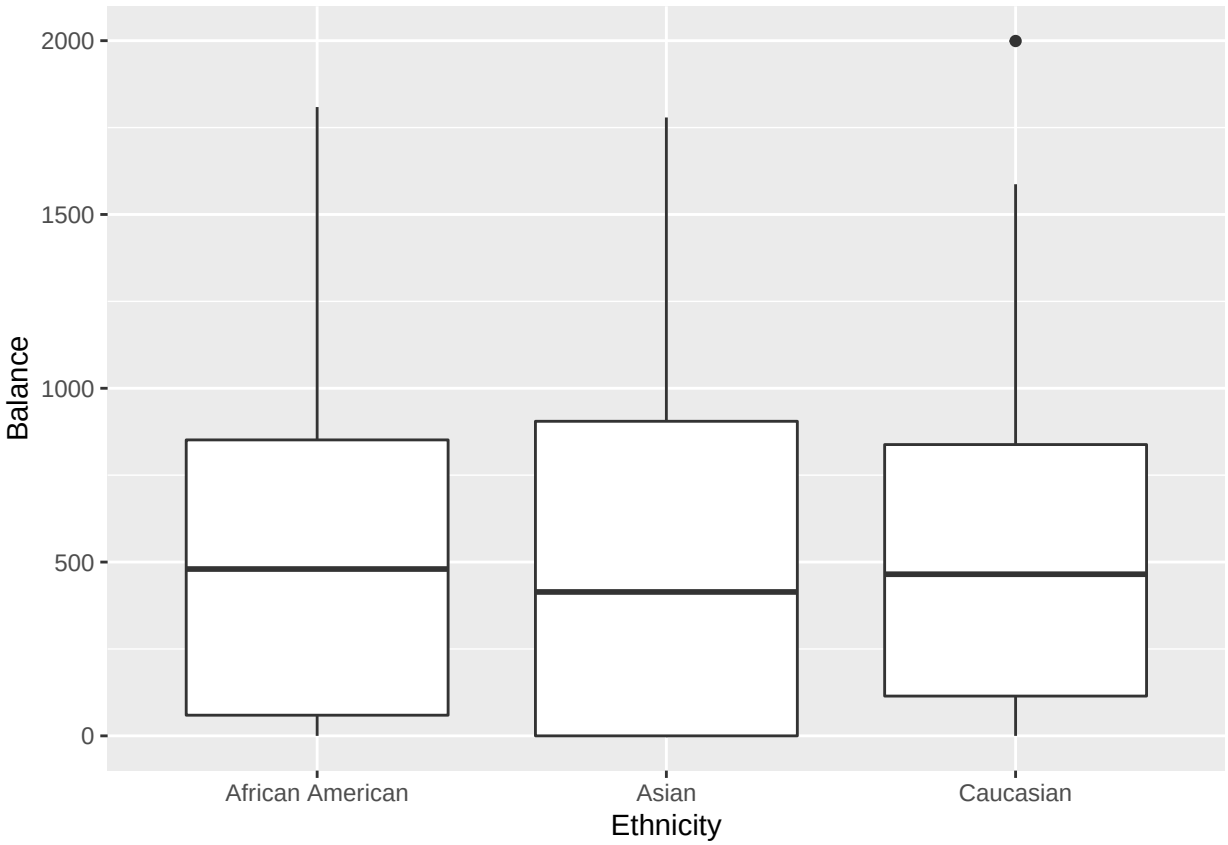
```
var.categorical <- colnames(Credit[, c("Gender", "Student", "Married", "Ethnicity")])
for (i in var.categorical) {
  plot <- ggplot(data = Credit, aes(x = Credit[, i], y = Balance)) +
    geom_boxplot()+
    labs(x = i)
```

```
  print(plot)
}
```




2000
1500
1000
500
0
Balance
No
Yes
Student


2000
1500
1000
500
0
Balance
No
Yes
Married



There are not so strong differences between Gender, Married and Ethnicity and Balance.

TASK 2: SELECT INTERACTION We should focus on the important variable first: Student, Income and Rating. That is easier for us