Bivariate Analysis For Banking Data

Ada Lazuli

2022-09-05

Contents

Libraries

```
library(ggplot2)
library(ggcorrplot)
library(dplyr)
library(tidyverse)
```

Helper Functions

```
box_cat_plt <- function(df, x, y, xname, yname) {
  title <- paste("Boxplot Comparing", xname, "to", yname)
  ggplot(df, aes(x = x, color = y,  fill = y)) + geom_boxplot(alpha = 0.5) + coord_flip() +
    xlab(xname) +
    theme(axis.ticks.y = element_blank(), axis.text.y = element_blank(),
          axis.title.y = element_blank()) +
    ggtitle(title) +
    labs(fill = yname, color = yname)
```

```
}

scatter_comp_plt <- function(df, x, y, color, colorname){
  title <- paste("Scatter Plot of AGE by Limit with", colorname, "shading")
  ggplot(df, aes(x = x, y = y, color = color)) +
    geom_point(alpha = 0.3) +
    xlab("Age") +
    ylab("Credit Limit") +
    ggtitle(title)
}

chi_matrix <- function(df) {
  rtn <- c()
  for(x in colnames(df)){
    temp <- c()
    for (y in colnames(df)) {
      temp <- c(temp, chisq.test(df[[x]], df[[y]])$p.value)
    }
    rtn <- rbind(rtn, temp)
  }

  rtn <- as.data.frame(rtn)
  colnames(rtn) <- colnames(df)
  rownames(rtn) <- colnames(df)

  return(rtn)
}

chi_vector <- function(df, target) {
  rtn <- c()
  status <- c()
  for(x in colnames(df)){
    val <-  round(chisq.test(df[[x]], target)$p.value,digits = 4)
    rtn <- c(rtn, val)
    status <- c(status, if(val < 0.01) "Passed" else "Failed")
  }
  rtn <- cbind(rtn,status)
  rtn <- cbind(rtn, colnames(df))
  rtn <- as.data.frame(rtn)
  colnames(rtn) <- c("pvalue", "Result", "feature")
  return(rtn)
}

extract_month <- function(x) {
  if (x == "PAY_0"){
    return("9")
  } else if (x == "PAY_2") {
    return("8")
  }else if (x == "PAY_3") {
    return("7")
  }else if (x == "PAY_4") {
    return("6")
  }else if (x == "PAY_5") {
```

```
    return("5")
  }else if (x == "PAY_6") {
    return("4")
  } else if (x == "BILL_AMT1") {
    return("9")
  } else if (x == "BILL_AMT2") {
    return("8")
  }else if (x == "BILL_AMT3") {
    return("7")
  }else if (x == "BILL_AMT4") {
    return("6")
  }else if (x == "BILL_AMT5") {
    return("5")
  }else if (x == "BILL_AMT6") {
    return("4")
  }else if (x == "PAY_AMT1") {
    return("9")
  }else if (x == "PAY_AMT2") {
    return("8")
  }else if (x == "PAY_AMT3") {
    return("7")
  }else if (x == "PAY_AMT4") {
    return("6")
  }else if (x == "PAY_AMT5") {
    return("5")
  }else if (x == "PAY_AMT6") {
    return("4")
  }

}
```

Data Loading

```
df <- read.csv("credit card default data set.csv")
df$default <- factor(df$default.payment.next.month, labels = c("No", "Yes"))
df$default.payment.next.month <- NULL
```

Pre-processing

The following features are categorical and need to be converted into factor types:

1. SEX
2. EDUCATION
3. MARRIAGE

```
df$SEX <- factor(df$SEX, labels = c("Male", "Female"))
df$EDUCATION <- factor(df$EDUCATION, labels = c("Others", "Graduate School", "Univserity", "High School
df$MARRIAGE <- factor(df$MARRIAGE, labels = c("Others", "Married", "Single","Others"))
```

Bivariate Analysis (Non-Payment Features)

Overview

The dataset has **30000** rows and **25** features. The names of the features are:

ID, LIMIT_BAL, SEX, EDUCATION, MARRIAGE, AGE, PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6, BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6, PAY_AMT1, PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5, PAY_AMT6, default

```
glimpse(df)
```

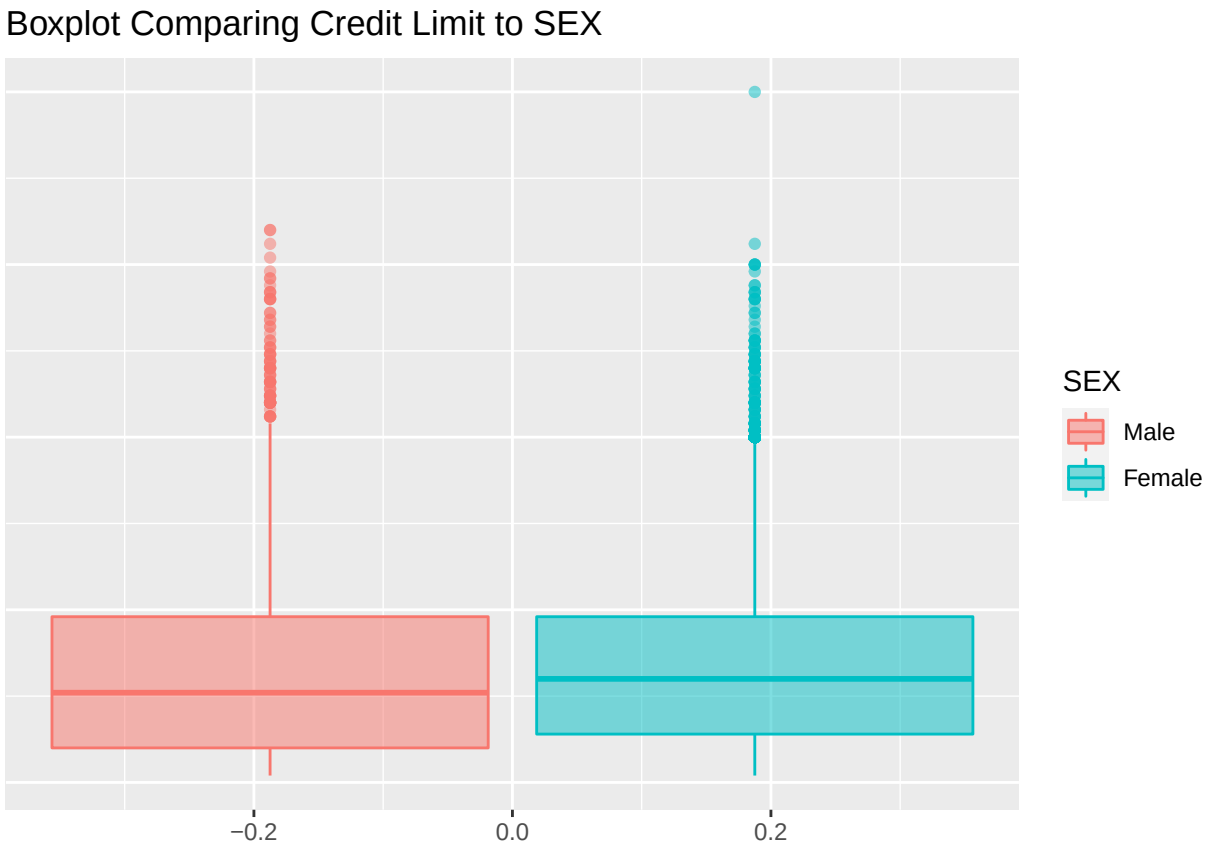
```
## Rows: 30,000
## Columns: 25
## $ ID        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1~
## $ LIMIT_BAL <int> 20000, 120000, 90000, 50000, 50000, 50000, 500000, 100000, 1~
## $ SEX       <fct> Female, Female, Female, Female, Male, Male, Male, Female, Fe~
## $ EDUCATION <fct> Univserity, Univserity, Univserity, Univserity, Univserity, ~
## $ MARRIAGE  <fct> Married, Single, Single, Married, Married, Single, Single, S~
## $ AGE       <int> 24, 26, 34, 37, 57, 37, 29, 23, 28, 35, 34, 51, 41, 30, 29, ~
## $ PAY_0     <int> 2, -1, 0, 0, -1, 0, 0, 0, 0, -2, 0, -1, -1, 1, 0, 1, 0, 0, 1~
## $ PAY_2     <int> 2, 2, 0, 0, 0, 0, 0, -1, 0, -2, 0, -1, 0, 2, 0, 2, 0, 0, -2,~
## $ PAY_3     <int> -1, 0, 0, 0, -1, 0, 0, -1, 2, -2, 2, -1, -1, 2, 0, 0, 2, 0, ~
## $ PAY_4     <int> -1, 0, 0, 0, 0, 0, 0, 0, 0, -2, 0, -1, -1, 0, 0, 0, 2, -1, -~
## $ PAY_5     <int> -2, 0, 0, 0, 0, 0, 0, 0, 0, -1, 0, -1, -1, 0, 0, 0, 2, -1, -~
## $ PAY_6     <int> -2, 2, 0, 0, 0, 0, 0, -1, 0, -1, -1, 2, -1, 2, 0, 0, 2, -1, ~
## $ BILL_AMT1 <int> 3913, 2682, 29239, 46990, 8617, 64400, 367965, 11876, 11285,~
## $ BILL_AMT2 <int> 3102, 1725, 14027, 48233, 5670, 57069, 412023, 380, 14096, 0~
## $ BILL_AMT3 <int> 689, 2682, 13559, 49291, 35835, 57608, 445007, 601, 12108, 0~
## $ BILL_AMT4 <int> 0, 3272, 14331, 28314, 20940, 19394, 542653, 221, 12211, 0, ~
## $ BILL_AMT5 <int> 0, 3455, 14948, 28959, 19146, 19619, 483003, -159, 11793, 13~
## $ BILL_AMT6 <int> 0, 3261, 15549, 29547, 19131, 20024, 473944, 567, 3719, 1391~
## $ PAY_AMT1  <int> 0, 0, 1518, 2000, 2000, 2500, 55000, 380, 3329, 0, 2306, 218~
## $ PAY_AMT2  <int> 689, 1000, 1500, 2019, 36681, 1815, 40000, 601, 0, 0, 12, 99~
## $ PAY_AMT3  <int> 0, 1000, 1000, 1200, 10000, 657, 38000, 0, 432, 0, 50, 8583,~
## $ PAY_AMT4  <int> 0, 1000, 1000, 1100, 9000, 1000, 20239, 581, 1000, 13007, 30~
## $ PAY_AMT5  <int> 0, 0, 1000, 1069, 689, 1000, 13750, 1687, 1000, 1122, 3738, ~
## $ PAY_AMT6  <int> 0, 2000, 5000, 1000, 679, 800, 13770, 1542, 1000, 0, 66, 364~
## $ default   <fct> Yes, Yes, No, No, No, No, No, No, No, No, No, No, No, Yes, N~
```

LIMIT_BAL v.s. Categorical Features

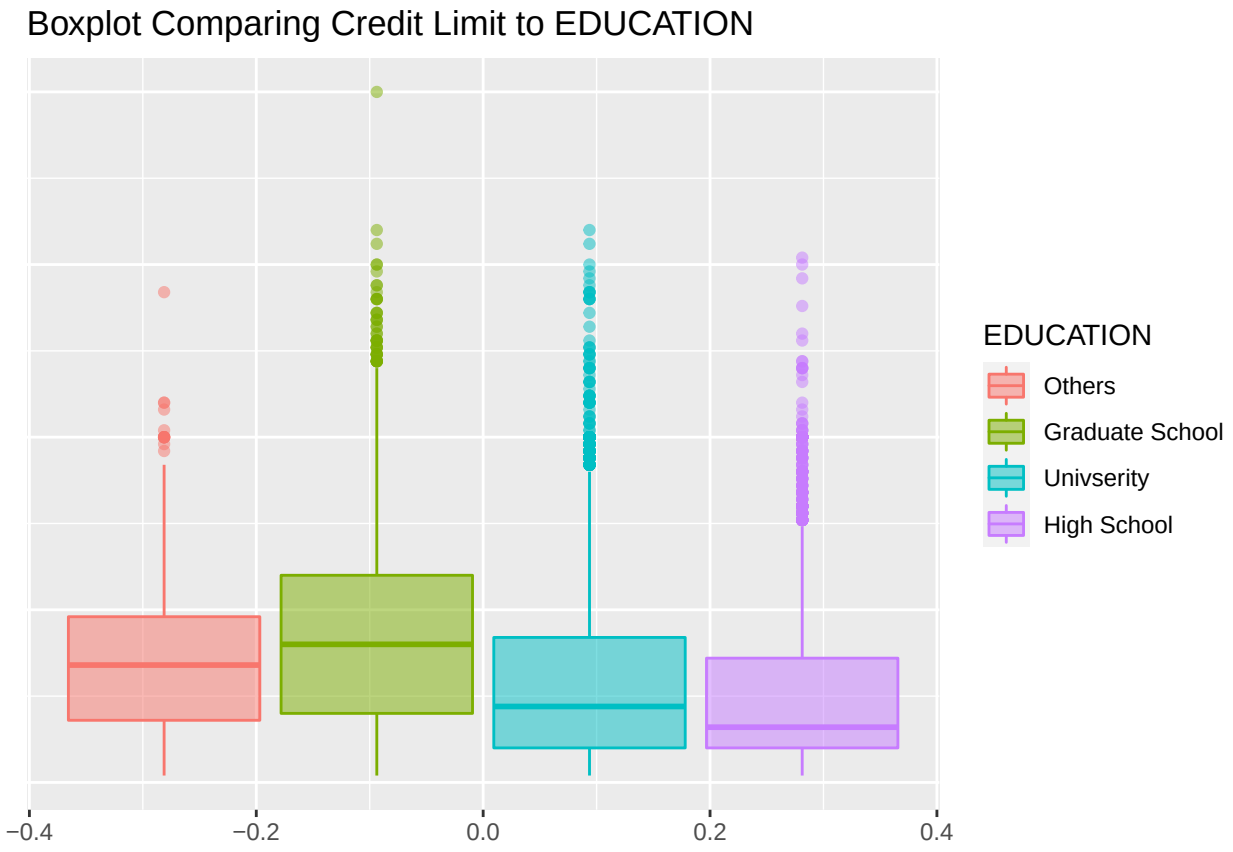
The P-value for chi-squared test for independence between LIMIT_BAL and each category is:

1. LIMIT_BAL vs. SEX: $7.7475724 \times 10^{-109}$
2. LIMIT_BAL vs. EDUCATION: 0
3. LIMIT_BAL vs. MARRIAGE: $1.1440653 \times 10^{-51}$
4. LIMIT_BAL vs. Default: $6.7793334 \times 10^{-161}$

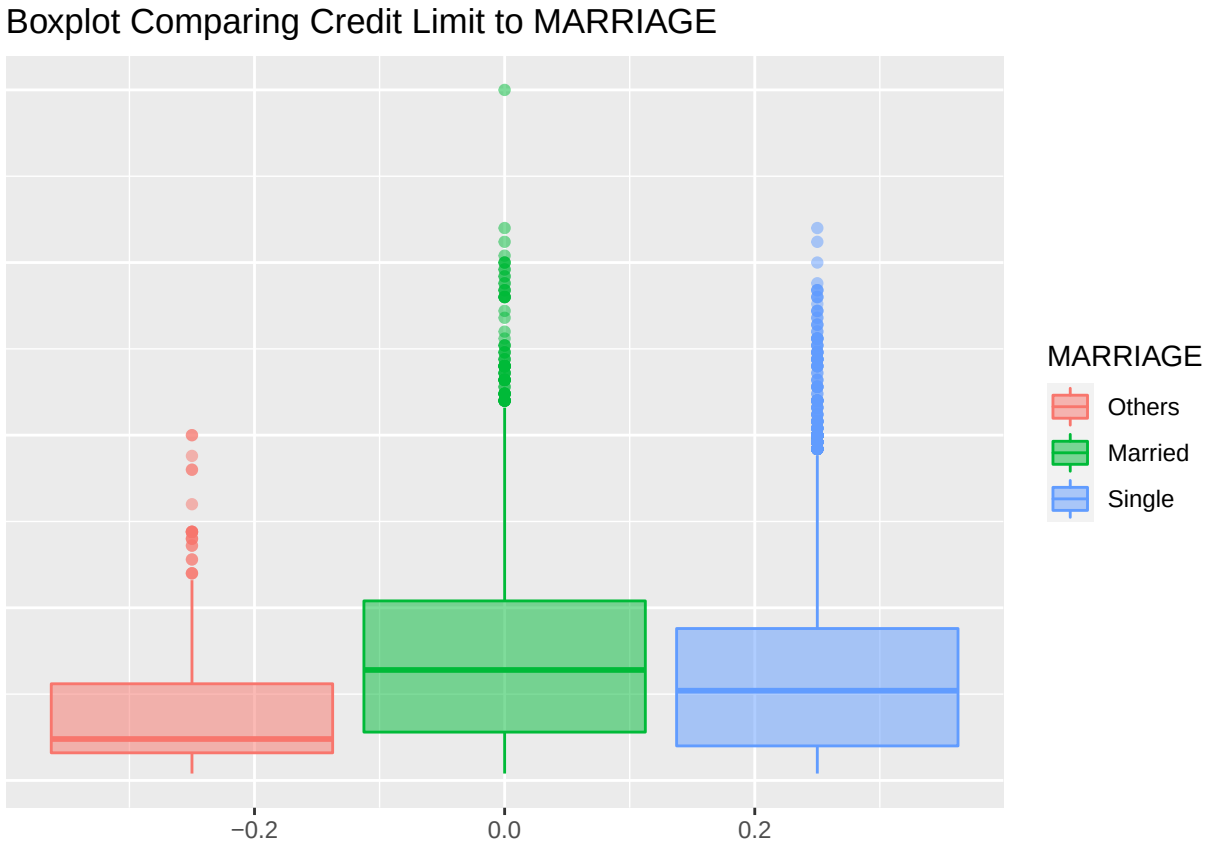
```
box_cat_plt(df, df$LIMIT_BAL, df$SEX, "Credit Limit", "SEX")
```




```
box_cat_plt(df, df$LIMIT_BAL, df$EDUCATION, "Credit Limit", "EDUCATION")
```




```
box_cat_plt(df, df$LIMIT_BAL, df$MARRIAGE, "Credit Limit", "MARRIAGE")
```

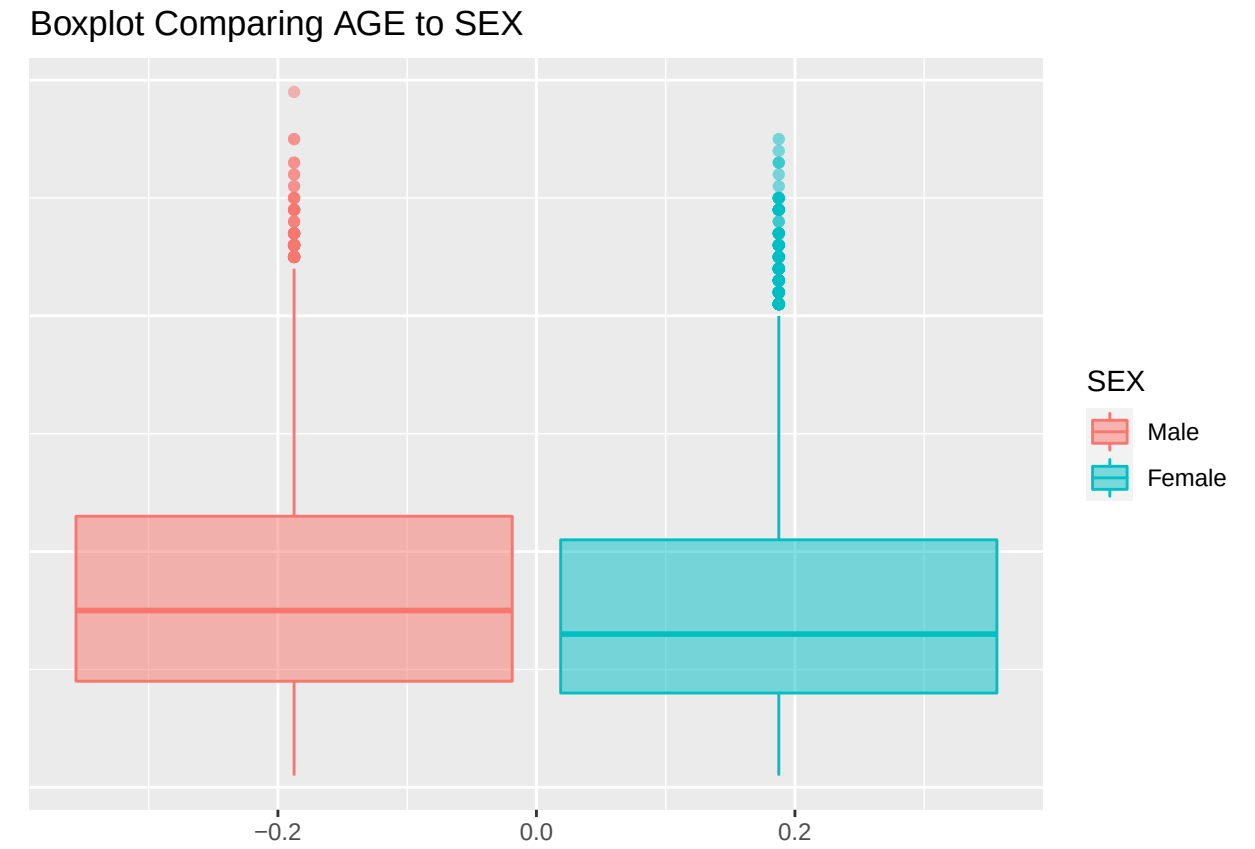


AGE v.s. Categorical Features

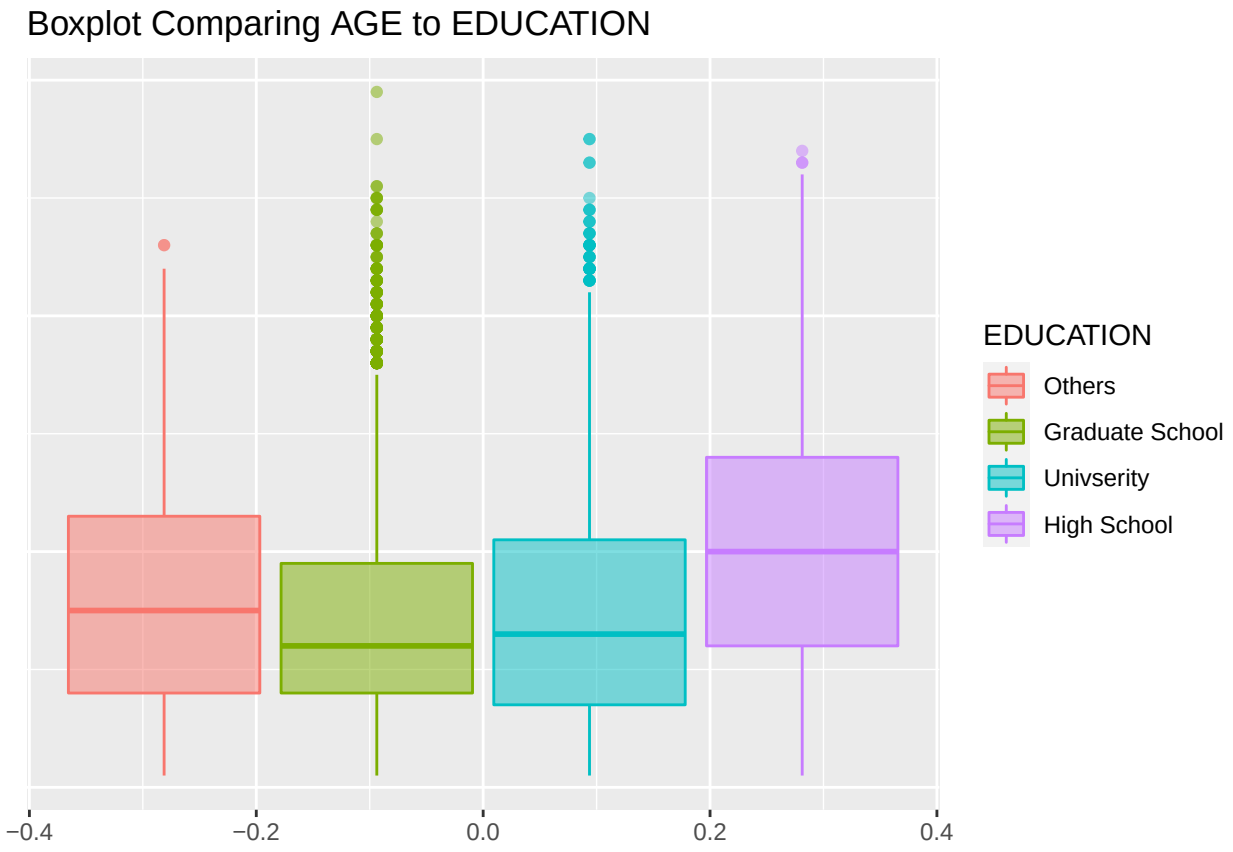
The P-value for chi-squared test for independence between AGE and each category is:

1. AGE vs. SEX: $5.1976868 \times 10^{-42}$
2. AGE vs. EDUCATION: 0
3. AGE vs. MARRIAGE: 0
4. AGE vs. Default: $5.6429915 \times 10^{-12}$

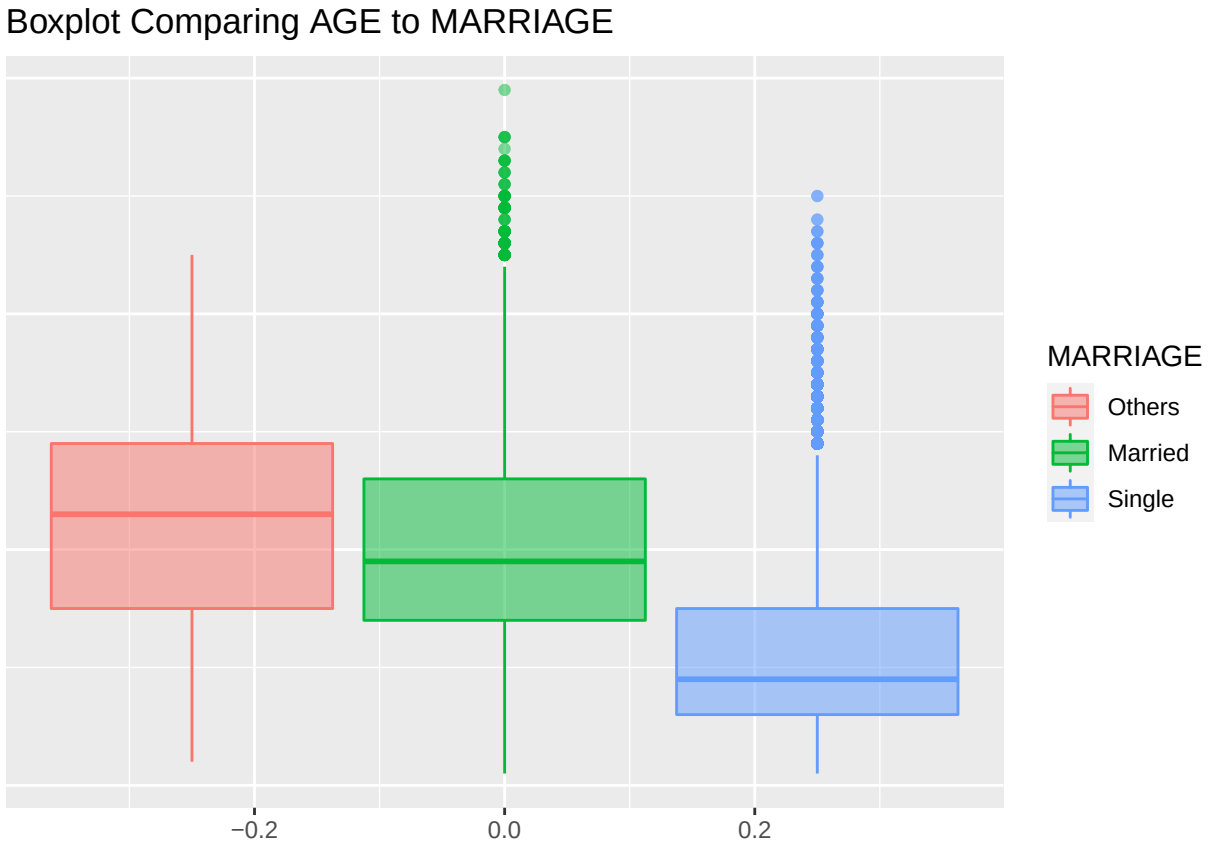
```
box_cat_plt(df, df$AGE, df$SEX, "AGE", "SEX")
```




```
box_cat_plt(df, df$AGE, df$EDUCATION, "AGE", "EDUCATION")
```



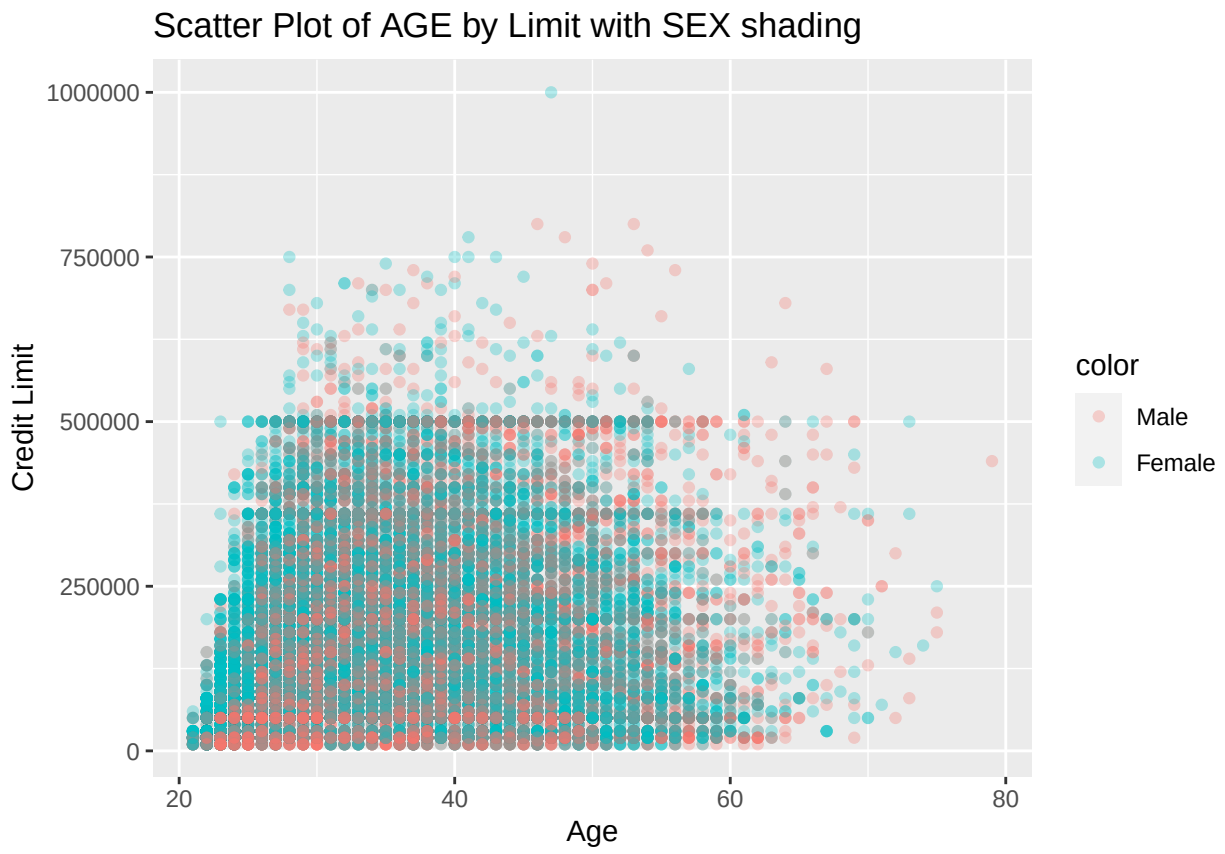

```
box_cat_plt(df, df$AGE, df$MARRIAGE, "AGE", "MARRIAGE")
```



AGE v.s. LIMIT_BAL

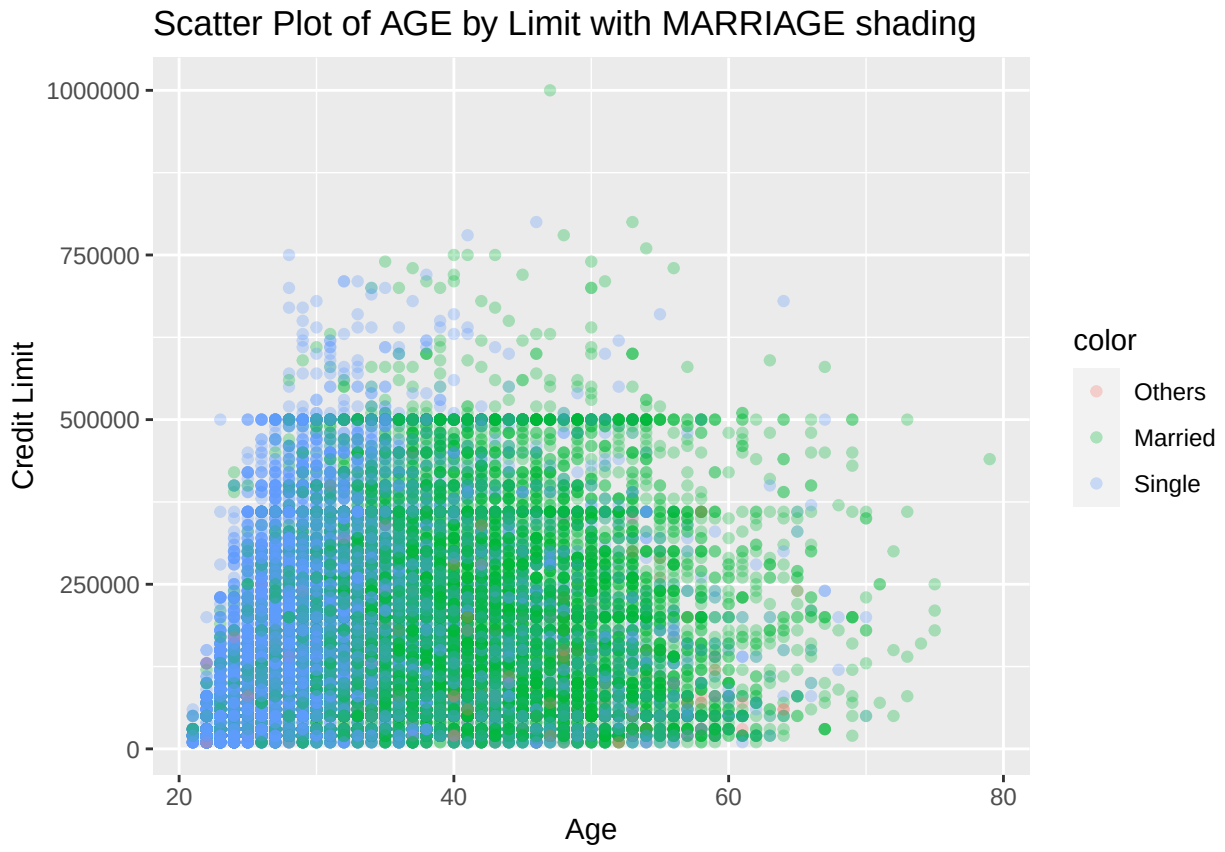
The correlation between AGE and LIMIT_BAL is 0.1447128

```
scatter_comp_plt(df, df$AGE, df$LIMIT_BAL, df$SEX, "SEX")
```

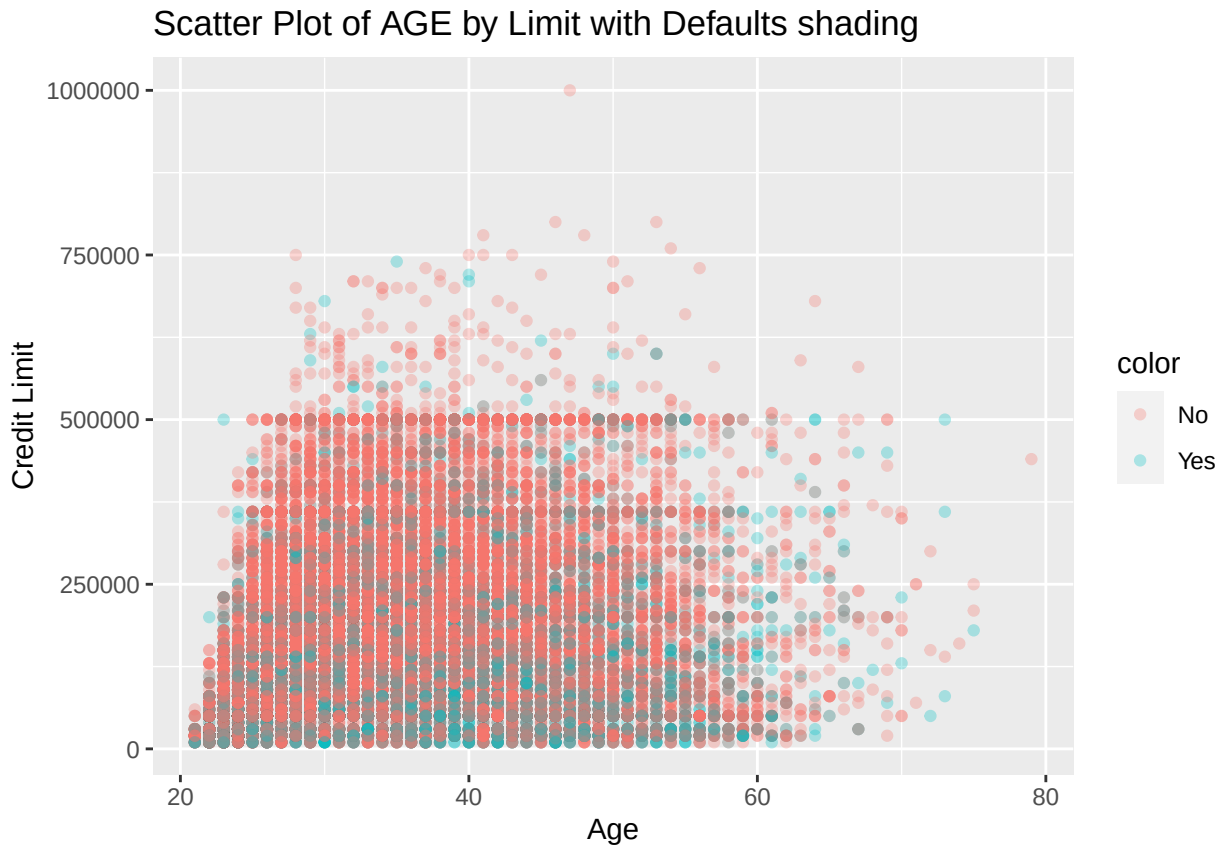



```
scatter_comp_plt(df, df$AGE, df$LIMIT_BAL, df$EDUCATION, "EDUCATION")
```


```
scatter_comp_plt(df, df$AGE, df$LIMIT_BAL, df$MARRIAGE, "MARRIAGE")
```




```
scatter_comp_plt(df, df$AGE, df$LIMIT_BAL, df$default, "Defaults")
```



Categorical Chi Squre Matrix

A matrix of the results for chi-squared testing was created to evalute the categorical features against each other for independence.

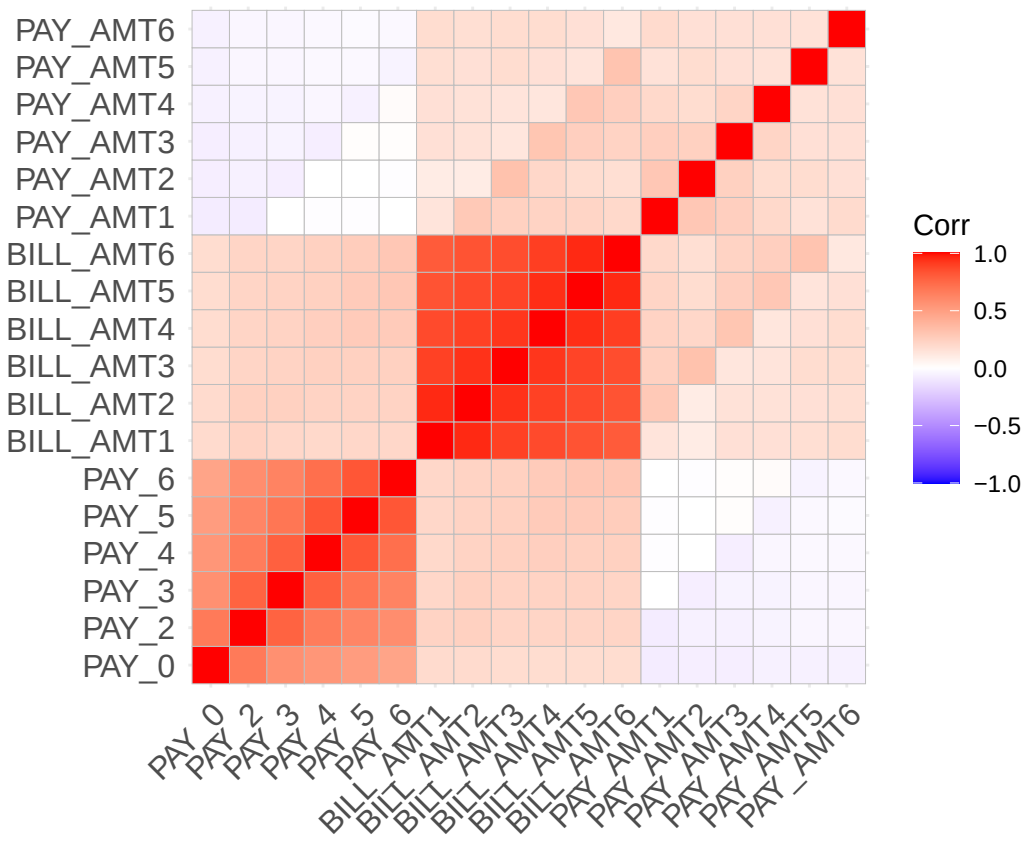
```
chi_matrix(df %>% select(SEX, MARRIAGE, EDUCATION, default))
```

```
##                    SEX      MARRIAGE     EDUCATION      default
## SEX       0.000000e+00  5.381729e-07  2.603103e-05 4.944679e-12
## MARRIAGE  5.381729e-07  0.000000e+00 6.334696e-232 7.790720e-07
## EDUCATION 2.603103e-05 6.334696e-232  0.000000e+00 1.495065e-34
## default   4.944679e-12  7.790720e-07  1.495065e-34 0.000000e+00
```

Bivariate Analysis (Payment Features)

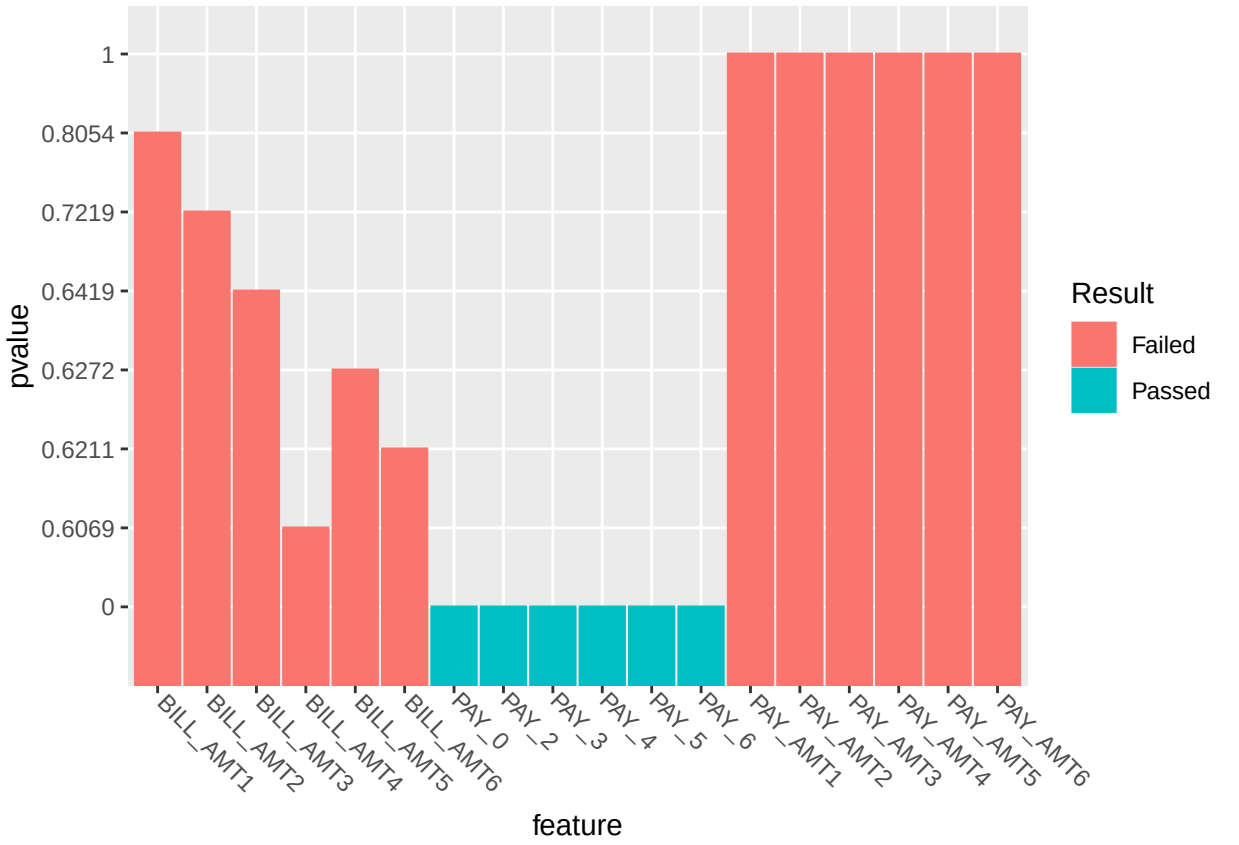
Correlation Matrix

```
ggcorrplot(cor(df%>% select(-AGE, -LIMIT_BAL, -SEX, -EDUCATION, -MARRIAGE, -default, -ID)))
```



Independence Testing

```
chiv <- chi_vector(df%>% select(-AGE, -LIMIT_BAL, -SEX, -EDUCATION, -MARRIAGE, -default, -ID), df$defaul
ggplot(chiv, aes(x = feature, y = pvalue, color = Result, fill = Result)) + geom_col() +
  theme(axis.text.x = element_text(angle = -45, vjust = 0.5, hjust = -0.05))
```



Bivariate Analysis (Time Series)

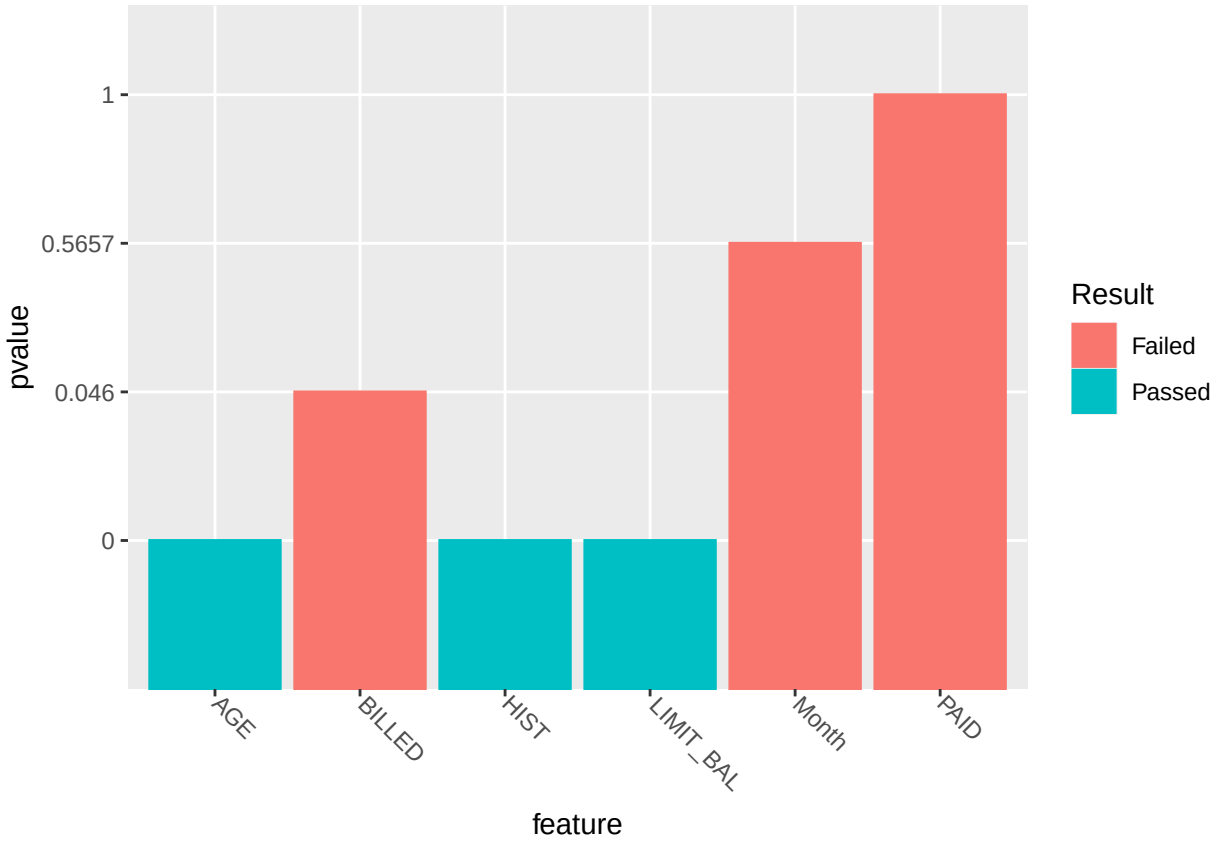
Pre-Processing

```
df$Index <- 1:nrow(df)

df_hist <- df %>% pivot_longer(c(PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6), "Month") %>%
  rename(HIST = value)

df_hist$Month <- sapply(df_hist$Month, extract_month)

df_billed <- df %>%   pivot_longer(c(BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6),
  rename(BILLED = value) %>%
  select(Month, BILLED, Index)

df_billed$Month <- sapply(df_billed$Month, extract_month)

df_paid <- df %>% pivot_longer(c(PAY_AMT1, PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5, PAY_AMT6), "Month") %
  rename(PAID = value) %>%
  select(Month, PAID, Index)

df_paid$Month <- sapply(df_paid$Month, extract_month)
```

```r
df_joined <- inner_join(x = df_billed, y = df_paid, by = c("Index", "Month")) %>%
inner_join(df_hist, by = c("Index", "Month")) %>%
select(Month, BILLED, PAID, HIST, LIMIT_BAL, SEX, EDUCATION, MARRIAGE, AGE, default)

df_sample <- df_joined %>% group_by(default) %>% slice_sample(prop = .3) %>% ungroup()
```

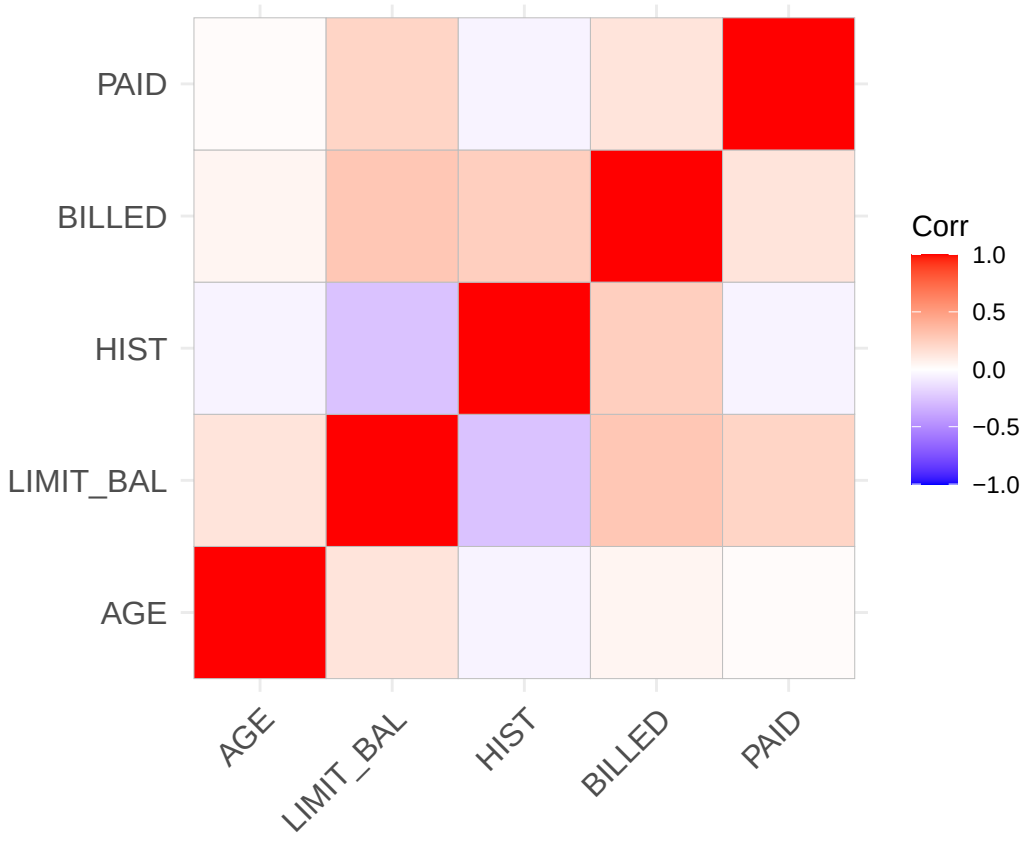
Quantitivative Feature Chi Square Eval

```r
chiv <- chi_vector(df_sample%>% select(-SEX, -EDUCATION, -MARRIAGE, -default), df_sample$default)
ggplot(chiv, aes(x = feature, y = pvalue, color = Result, fill = Result)) + geom_col() +
  theme(axis.text.x = element_text(angle = -45, vjust = 0.5, hjust = -0.05))
```



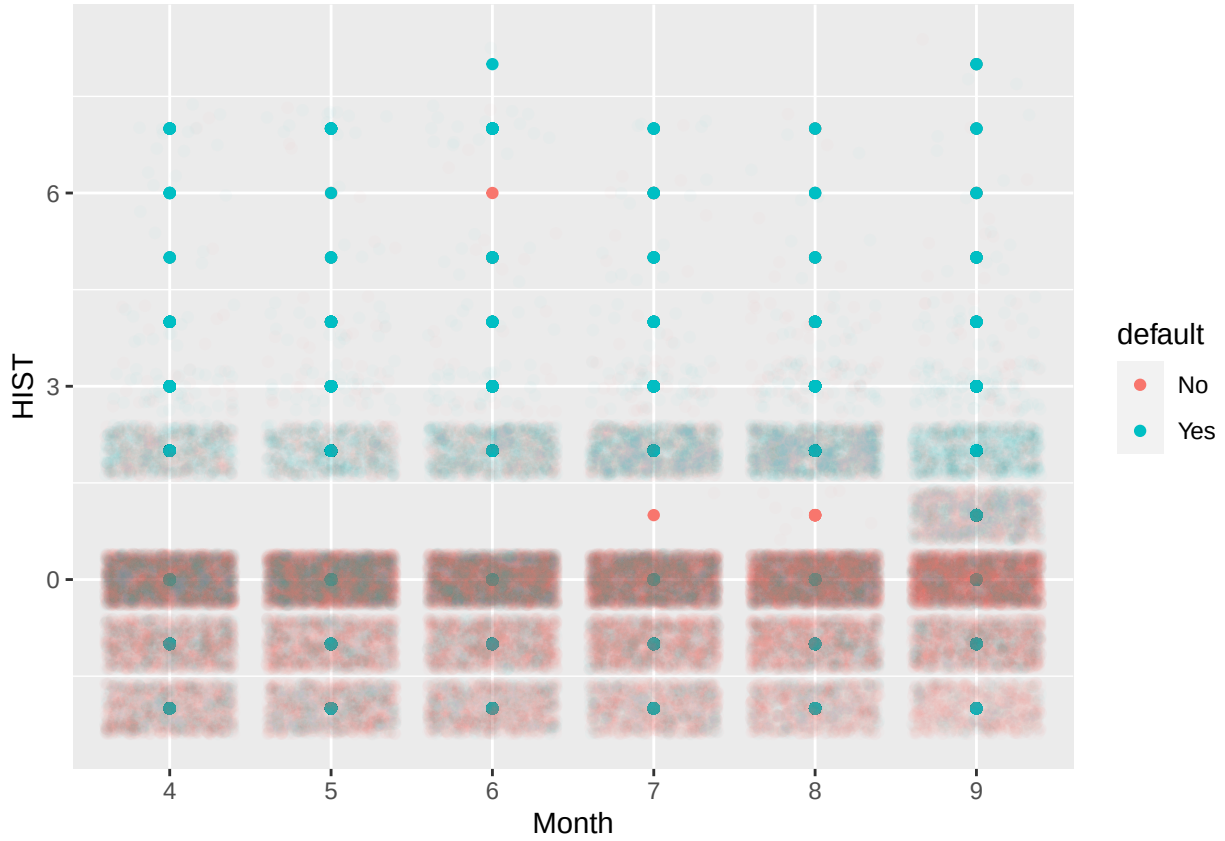
Qualitative Variable Correlation

```r
ggcorrplot(cor(df_sample %>% select(AGE, LIMIT_BAL, HIST, BILLED, PAID)))
```



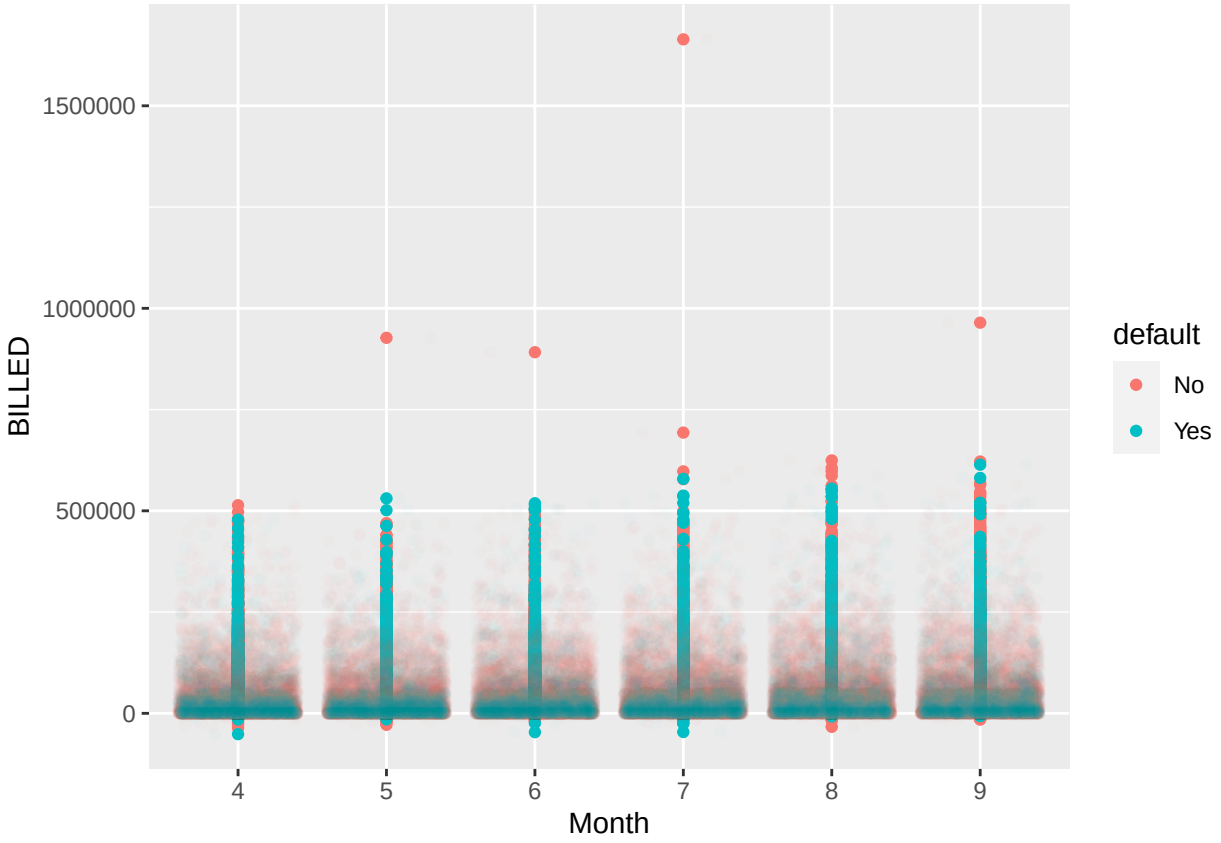
HIST over Time

```
ggplot(df_sample, aes(x = Month, y = HIST, color = default)) + geom_point() + geom_jitter(alpha = .02)
```



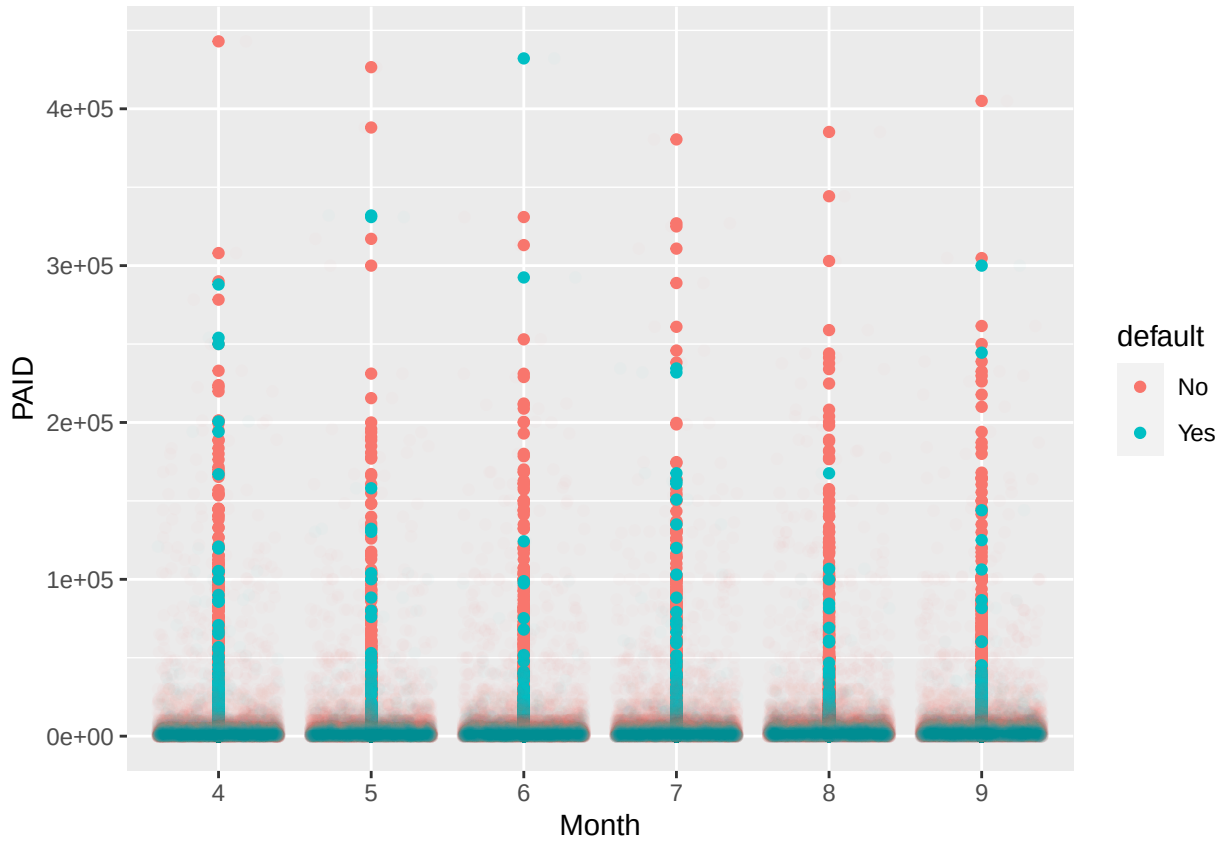
BILLED over Time

```
ggplot(df_sample, aes(x = Month, y = BILLED, color = default)) + geom_point() + geom_jitter(alpha = .02
```



Paid over Time

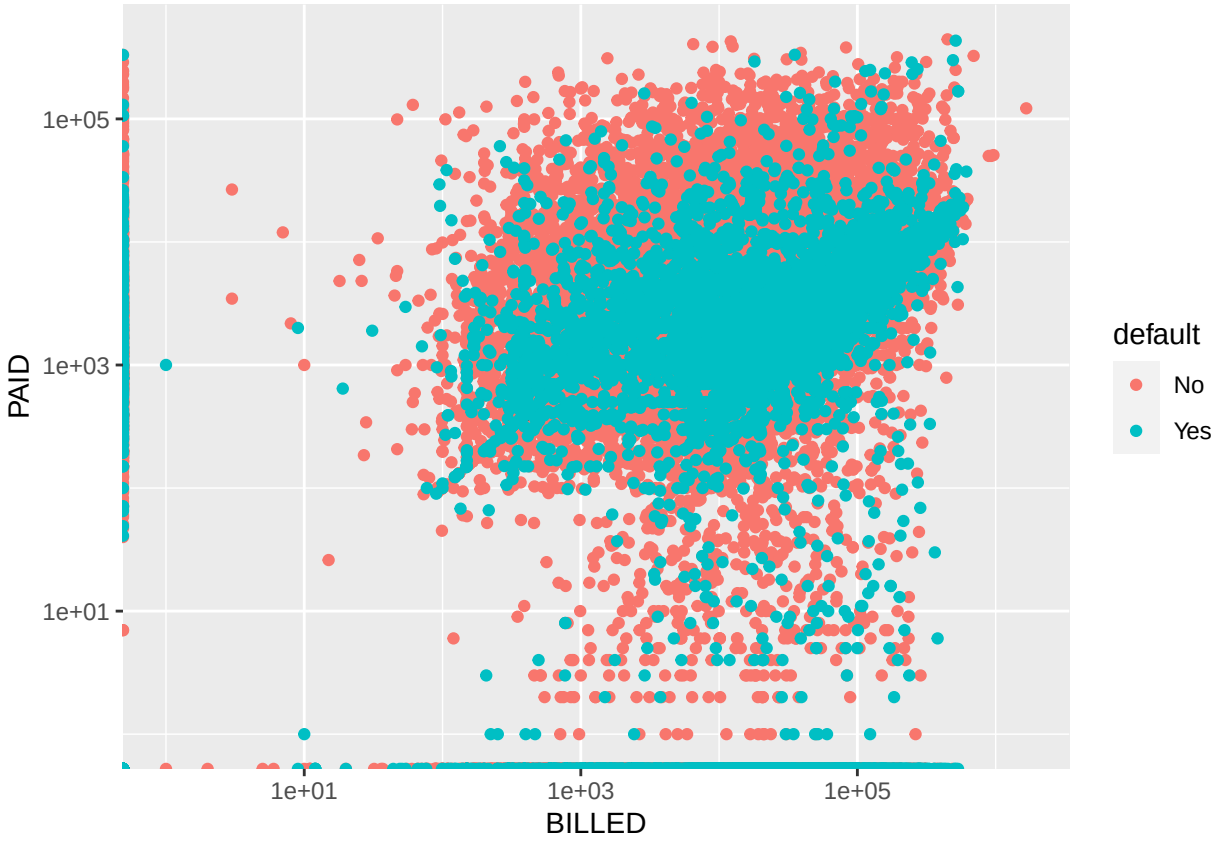
```
ggplot(df_sample, aes(x = Month, y = PAID, color = default)) + geom_point() + geom_jitter(alpha = .02)
```



PAID vs BILLED

```
ggplot(df_sample, aes(x = BILLED, y = PAID, color = default)) + geom_point()
```


```
ggplot(df_sample, aes(x = BILLED, y = PAID, color = default)) + geom_point() + scale_y_log10() + scale_
```


1e+05
1e+03
1e+01
PAID
1e+01
1e+03
1e+05
BILLED
default
No
Yes