Project

On

R

Programming language

On

Credit Card Customers

Presented by

Sajith Vellappillil

Under the guidance of

Mr. Hamid Rajaee

Index:

INTRODUCTION

Today massive data is collected by business, industries, and governments everyday. These enterprises must be able to not only collect and store data but also analyze it in an environment for statistical computing and graphics which would serve as a base to make strategic and informed decisions that can increase their profitability and solve real-life problems.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. The R language is widely used among statisticians and data miners for developing statistical software and data analysis.

PROJECT OUTLINE

This project is on the Customer's Credit Card details in which there are a number of customers leaving their credit card services. We would perform Data Analysis and examine trends & correlations within our data.

PROJECT - DETAILS

GOALS:

1. **Clean the data to make it meaningful for statistical computing.**
2. **Examine trends and correlations within the data**
3. **Determine which features are most important for a churning customer.**

DATA SUMMARY

- Credit Card Customer Data with originally 21 variables (columns) and 10127 observations (rows).
- Cleaned (Wrangled) data consist of 9 variables and 7081 observations

VARIABLES AND TARGET

1. **Customer_Age — (Feature, int,continuous)** Age of the Customer
2. **Gender — (Feature, chr, discrete)** Sex of the Customer
3. **Dependent_count — (Feature, int, continuous)** number of dependents a user has. To analyse how many people are dependent on a credit card user for financial support. A higher count tells us that the expenditures can be high.
4. **Education_Level — (Feature, char, discrete)** Education Level of the Customer
5. **Marital_Status — (Feature, char, discrete)** Martial Status of the Customer
6. **Income_Category — (Feature, char, discrete)** Income Category of Customer
7. **Card_Category — (Feature, char, discrete)** Card Category of Customer
8. **Credit_Limit — (Feature, dbl,continuous)** Client Number
9. **Customer Status — (TARGET, discrete, binary): Inactive (**customer exits the company) **or Active (**customer stays in company)

DATA WRANGLING

Reading the data file & install required libraries. Check for any problems with the dataset.

```
> #---------------------------
> #READING THE DATA
> #---------------------------
> pro <- read.csv(file.choose())
> dim(pro)#to get the shape of the original data
[1] 10127    21
> str(pro)# to get the structure of the data
'data.frame':   10127 obs. of  21 variables:
 $ CLIENTNUM              : int  768805383 818770008 713982108 769911858 709106358 713061558 810347208 818906208 710930508 719661558 ...
 $ Attrition_Flag         : chr  "Existing Customer" "Existing Customer" "Existing Customer" "Existing Customer" ...
 $ Customer_Age           : int  45 49 51 40 40 44 51 32 37 48 ...
 $ Gender                 : chr  "M" "F" "M" "F" ...
 $ Dependent_count        : int  3 5 3 4 3 2 4 0 3 2 ...
 $ Education_Level        : chr  "High School" "Graduate" "Graduate" "High School" ...
 $ Marital_Status         : chr  "Married" "Single" "Married" "Unknown" ...
 $ Income_Category        : chr  "$60K - $80K" "Less than $40K" "$80K - $120K" "Less than $40K" ...
 $ Card_Category          : chr  "Blue" "Blue" "Blue" "Blue" ...
 $ Months_on_book         : int  39 44 36 34 21 36 46 27 36 36 ...
 $ Total_Relationship_Count: int  5 6 4 3 5 3 6 2 5 6 ...
 $ Months_Inactive_12_mon : int  1 1 1 4 1 1 1 2 2 3 ...
 $ Contacts_Count_12_mon  : int  3 2 0 1 0 2 3 2 0 3 ...
 $ Credit_Limit           : num  12691 8256 3418 3313 4716 ...
 $ Total_Revolving_Bal    : int  777 864 0 2517 0 1247 2264 1396 2517 1677 ...
 $ Avg_Open_To_Buy        : num  11914 7392 3418 796 4716 ...
 $ Total_Amt_Chng_Q4_Q1   : num  1.33 1.54 2.59 1.4 2.17 ...
 $ Total_Trans_Amt        : int  1144 1291 1887 1171 816 1088 1330 1538 1350 1441 ...
 $ Total_Trans_Ct         : int  42 33 20 20 28 24 31 36 24 32 ...
 $ Total_Ct_Chng_Q4_Q1    : num  1.62 3.71 2.33 2.33 2.5 ...
 $ Avg_Utilization_Ratio  : num  0.061 0.105 0 0.76 0 0.311 0.066 0.048 0.113 0.144 ...
>
```

```
head(pro,20)# get first 20 observations

tail(pro,20)# get last 20 observations

pro1 <- pro# make a copy
```

Checking for missing values and assigning NA to missing values and 'Unknown' values

```
> pro1[pro1=='']<-NA #assign NA to missing values
> pro1[pro1=='Unknown']<-NA  #assign NA to 'Unknown' value
> sum(is.na(pro1))    # check total NA values
[1] 3380
> sum(is.na(pro1$Marital_Status))
[1] 749
> sum(is.na(pro1$Education_Level))
[1] 1519
> sum(is.na(pro1$Income_Category))
[1] 1112
> 749+1519+1112
[1] 3380
>
```

Copying dataset and slicing for our analysis

- Target Variable 'Attrition_Flag' renamed to 'Customer Status' using colnames
- Renamed values of Target Variable to 'Active' and 'Inactive'
- We removed any rows/entries with a "Unknown"/NA value.

```
> pro1 <- na.omit(pro1) # get all observations except NA
> pro1 <- pro1[,-c(1, 10:13,15:23)] # selected the columns we care about
> dim(pro1)
[1] 7081    9
> colnames(pro1)[1] <- c("Customer Status")
> pro1$`Customer Status`[pro1$`Customer Status`=='Attrited Customer']<- "Inactive"
> pro1$`Customer Status`[pro1$`Customer Status`=='Existing Customer']<- "Active"
> str(pro1)
'data.frame':   7081 obs. of  9 variables:
 $ Customer Status: chr  "Active" "Active" "Active" "Active" ...
 $ Customer_Age   : int  45 49 51 40 44 37 48 56 57 48 ...
 $ Gender         : chr  "M" "F" "M" "M" ...
 $ Dependent_count: int  3 5 3 3 2 3 2 1 2 4 ...
 $ Education_Level: chr  "High School" "Graduate" "Graduate" "Uneducated" ...
 $ Marital_Status : chr  "Married" "Single" "Married" "Married" ...
 $ Income_Category: chr  "$60K - $80K" "Less than $40K" "$80K - $120K" "$60K - $80K" ...
 $ Card_Category  : chr  "Blue" "Blue" "Blue" "Blue" ...
 $ Credit_Limit   : num  12691 8256 3418 4716 4010 ...
>
```

We see here we initially had 10,127 rows & 21 columns, we now have 7081 rows and 9 columns.

Exploratory Data Analysis (EDA)

1. Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to maximize insight into a data set;
2. uncover underlying structure;
3. extract important variables;
4. detect outliers and anomalies;
5. test underlying assumptions;
6. develop parsimonious models; and
7. determine optimal factor settings.

In this report we will discuss and employ EDA in the form of 10 questions and try to understand data more in depth.

Question 1 – How much is the average of numeric values for all the customers and what is your findings?

Multivariate analysis → 1 Categorical and 3 Numerical

Syntax & Result

```
> aggregate(pro1[c(2,4,9)],pro1[1] ,mean) # finding mean for all numeric variables from dataset.
  Customer Status Customer_Age Dependent_count Credit_Limit
1          Active     46.31736        2.331434     8555.099
2        Inactive     46.51033        2.371968     8158.580
```

We can evidently see that Active Customers has higher mean credit limits than Inactive customers. The mean for other variables does not have a significant difference. Hence, we can assume that credit limit has an inverse relation with the customer exits. i.e. higher the credit limit, lower would be the customer exit.

Question 2 – Is there any relation between Customer Status and Income_Category?

Chisq Test is done to prove relationship between 2 Categorical variables

Syntax & Result

```
> #We apply the chisq.test function to the contingency table Chi_tbl.
> Chi_tbl <- table(pro1$`Customer Status`,pro1$Income_Category)
> Chi_tbl

           $120K + $40K - $60K $60K - $80K $80K - $120K Less than $40K
  Active       470        1208         958         1013            2319
  Inactive     102         204         145          189             473
> chisq.test(Chi_tbl)

        Pearson's Chi-squared test

data:  Chi_tbl
X-squared = 12.312, df = 4, p-value = 0.01518
```

Null hypothesis: Customer status is independent of Income_Category,i.e; Customer status is not effected by Income_Category.

Here, the p-value 0.1518 < 0.05 significance level.

It indicates strong evidence against the null hypothesis, therefore, we reject the null hypothesis and accept the alternative hypothesis.

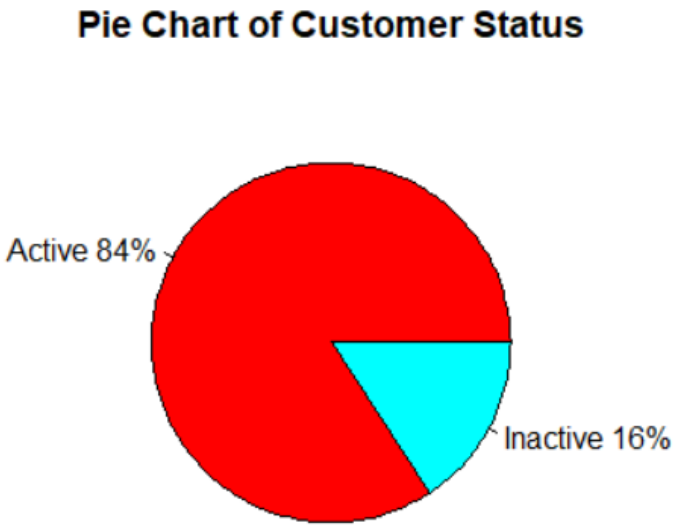
Question 3 – What is the distribution of target (Customer Status)?

Syntax

```
#since target is categorical variable, in univariate Analysis for summarization I will find frequency and
#for visualization we use: pie chart


# Pie Chart with Percentages
count<-table(pro1$`Customer Status`)
count
lbls <- c("Active", "Inactive")
pct <- round(count/sum(count)*100)
lbls <- paste(lbls, pct) # adding percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(count,labels = lbls, col=rainbow(length(lbls)),
    main="Pie Chart of Customer Status")
```

Result



As we have 84% active customers and 16% inactive (exited) customers, we are dealing with unbalanced data.

Active 84% → 5948 Customers

Inactive 16% → 1133 Customers (This is an alarming number of exits)

Question 4 – What is the distribution of Gender, Education_Level, Marital Status, Income_Category and Card_Category in terms of Customer Status?

Syntax & Result

```
> table(pro1$`Customer Status`,pro1$Gender) #

             F    M
  Active   2799 3169
  Inactive  576  537
> table(pro1$`Customer Status`,pro1$Education_Level)

           College Doctorate Graduate High School Post-Graduate Uneducated
  Active       712       283     2185        1416           354       1018
  Inactive     132        75      406         237            77        186
> table(pro1$`Customer Status`,pro1$Marital_Status)

           Divorced Married Single
  Active        477    3035   2456
  Inactive       92     529    492
> table(pro1$`Customer Status`,pro1$Income_Category)

           $120K + $40K - $60K $60K - $80K $80K - $120K Less than $40K
  Active         470        1208         958          1013           2319
  Inactive       102         204         145           189            473
> table(pro1$`Customer Status`,pro1$Card_Category)

           Blue Gold Platinum Silver
  Active   5564   68        8    328
  Inactive 1034   13        3     63
>
```

Question 5 -Is there any outliers in the data?

Syntax

```
> #Creating a function to find outliers based n Mean and Standard deviation
> notout<-function(x){
+   print("summary before applying this method ")
+   print(summary(x))
+   M1<-mean(x,na.rm = TRUE)
+   S1<-sd(x,na.rm=TRUE)
+   low1<-M1-3*S1
+   up1<-M1+3*S1
+   x[x<low1]<-NA
+   x[x>up1]<-NA
+   print("summary after applying this method ")
+   print(summary(x))
+   return(x)
+ }
>
> pro1$Credit_Limit<-notout(pro1$Credit_Limit)
[1] "summary before applying this method "
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1438    2498    4287    8493   10729   34516
[1] "summary after applying this method "
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1438    2498    4287    8493   10729   34516
>
> pro1$Dependent_count<-notout(pro1$Dependent_count)
[1] "summary before applying this method "
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   2.000   2.338   3.000   5.000
[1] "summary after applying this method "
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   2.000   2.338   3.000   5.000
>
```

We applied the function on variables Credit_Limit and Dependent_Count. No outliers were found.

Question 6 – What is the Customer Status in relation to each of the below variables?

A) Income_Category

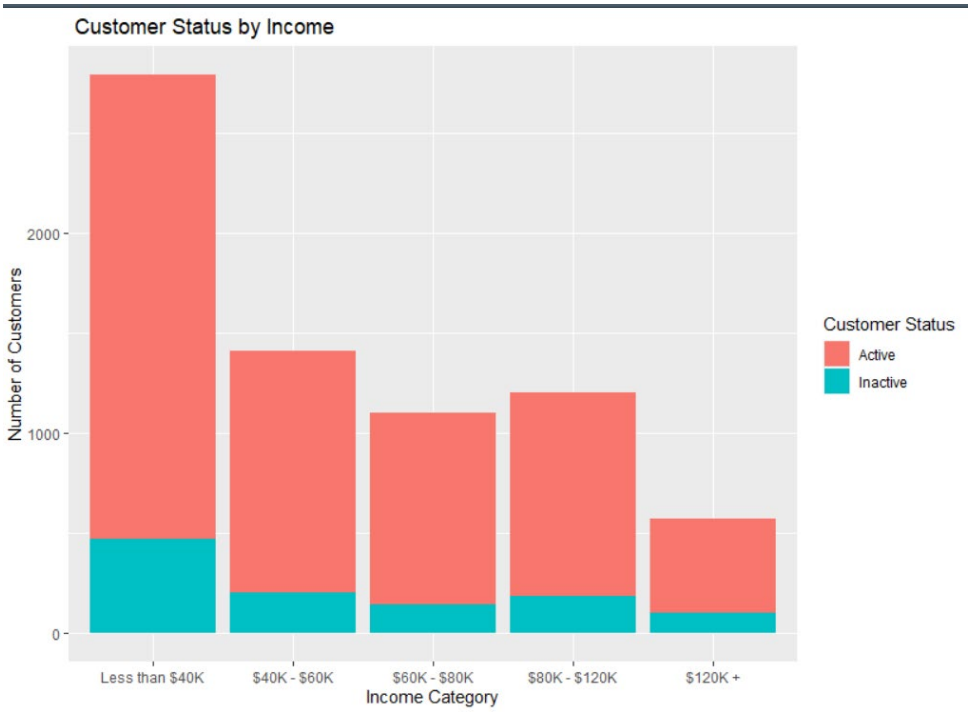
Bivariate Analysis→ Categorical Vs. Categorical

Syntax

```
#converting to factor and adding levels
pro1$Income_Category <- factor(pro1$Income_Category,
                               levels = c("Less than $40K","$40K - $60K","$60K - $80K","$80K - $120K","$120K +"))

ggplot(pro1 , aes(x = Income_Category)) +
  geom_bar(aes(fill = `Customer Status`)) +
  xlab("Income Category")+ ylab("Number of Customers")+
  ggtitle(" Customer Status by Income" )
```

Result



We see that the inactive customers fall within the range of Less than $40K,however majority of our active/inactive customers fall in the same range.

B) Education_Level

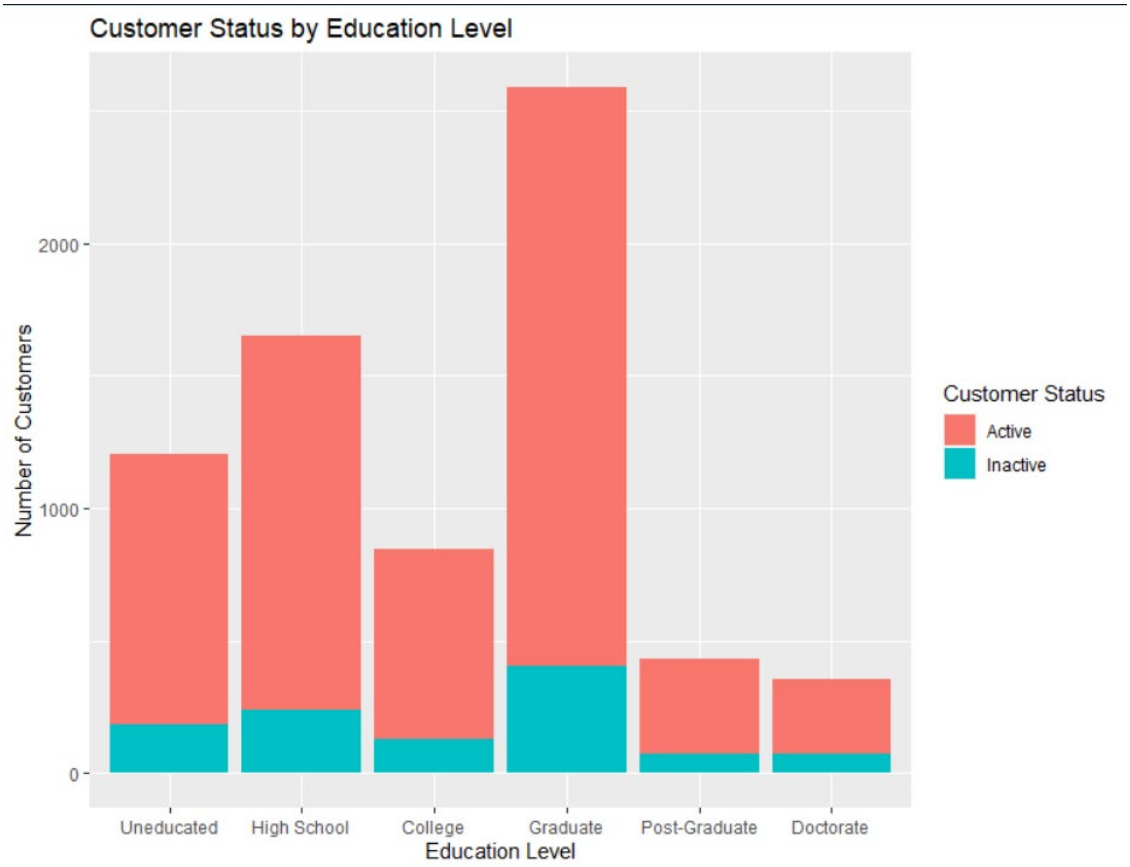
Bivariate Analysis→ Categorical Vs. Categorical

Syntax

```
#converting to factor and adding levels
pro1$Education_Level <- factor(pro1$Education_Level,
                               levels = c("Uneducated","High School", "College","Graduate","Post-Graduate","Doctorate"))

ggplot(pro1, aes(x = Education_Level)) +
  geom_bar(aes(fill = `Customer Status`)) +
  ylab("Number of Customers") +
  xlab("Education Level") + ggtitle("Customer Status by Education Level" )
```

Result



Most of our active customers are graduates & have a high school degree.

C) Marital_Status

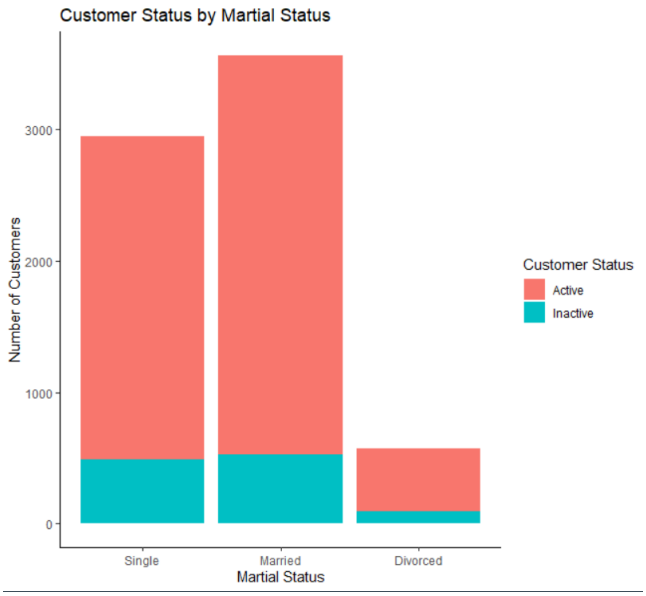
Bivariate Analysis→ Categorical Vs. Categorical

Syntax

```
#converting to factor and adding levels
pro1$Marital_Status <- factor(pro1$Marital_Status,
                             levels = c("Single", "Married","Divorced"))

ggplot(pro1 , aes(x = Marital_Status)) +
  geom_bar(aes(fill = `Customer Status`), position = position_stack(reverse = FALSE)) +
  theme(legend.position = "top") + theme_classic() + ylab("Number of Customers") +
  xlab("Martial Status") + ggtitle("Customer Status by Martial Status" )
```

Result



Martial Status of Customers, very small proportion are divorced, majority are married & single.

D) Card_Category

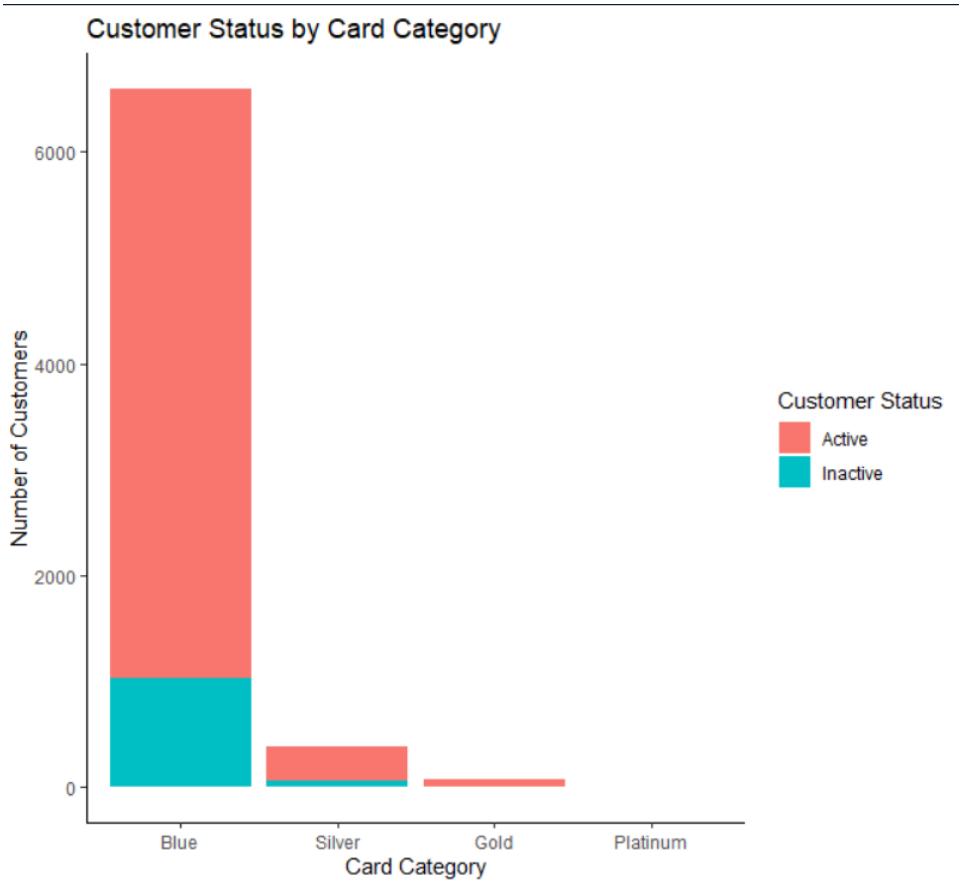
Bivariate Analysis→ Categorical Vs. Categorical

Syntax

```
#converting to factor and adding levels
pro1$Card_Category <- factor(pro1$Card_Category,
                             levels = c("Blue", "Silver","Gold","Platinum"))

ggplot(pro1 , aes(x = Card_Category)) +
  geom_bar(aes(fill = `Customer Status`), position = position_stack(reverse = FALSE)) +
  theme(legend.position = "top") + theme_classic() + ylab("Number of Customers") +
  xlab("Card Category") + ggtitle("Customer Status by Card Category" )
```

Result



Blue Card is the most significant Card Category for our active & inactive customers.

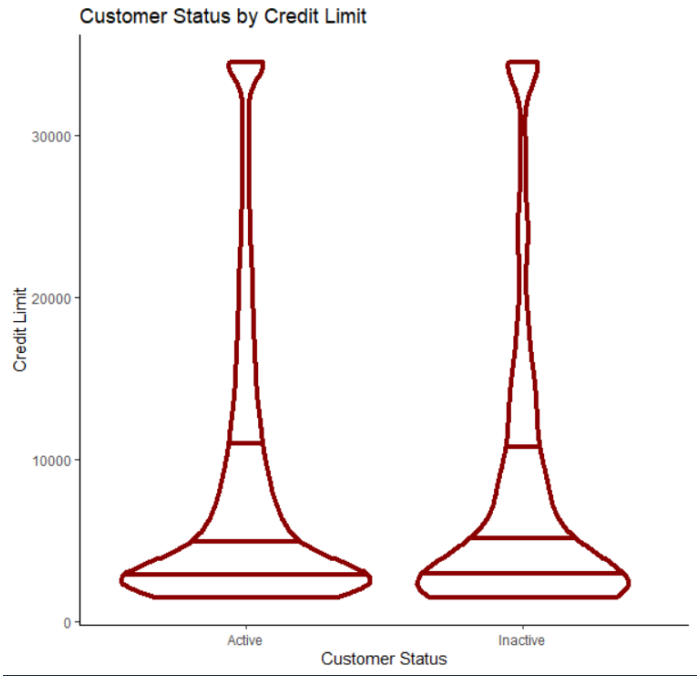
Question 7 – What are the Quantile values for Customer Status by Credit_Limit?

Bivariate Analysis→Categorical vs. Continuous

Syntax

```
ggplot(pro1 , aes(`Customer Status`,Credit_Limit,color= Credit_Limit)) +
  geom_violin(draw_quantiles = c(0.25,0.5,0.75),colour="dark red",size=1.5) +
  theme_classic() +xlab("Customer Status") + ylab("Credit Limit") +
  ggtitle("Customer Status by Credit Limit" )
```

Result



There is a larger spread of active customers. The red horizontal lines are quantiles.

First line from bottom is 25th percentile or Q1

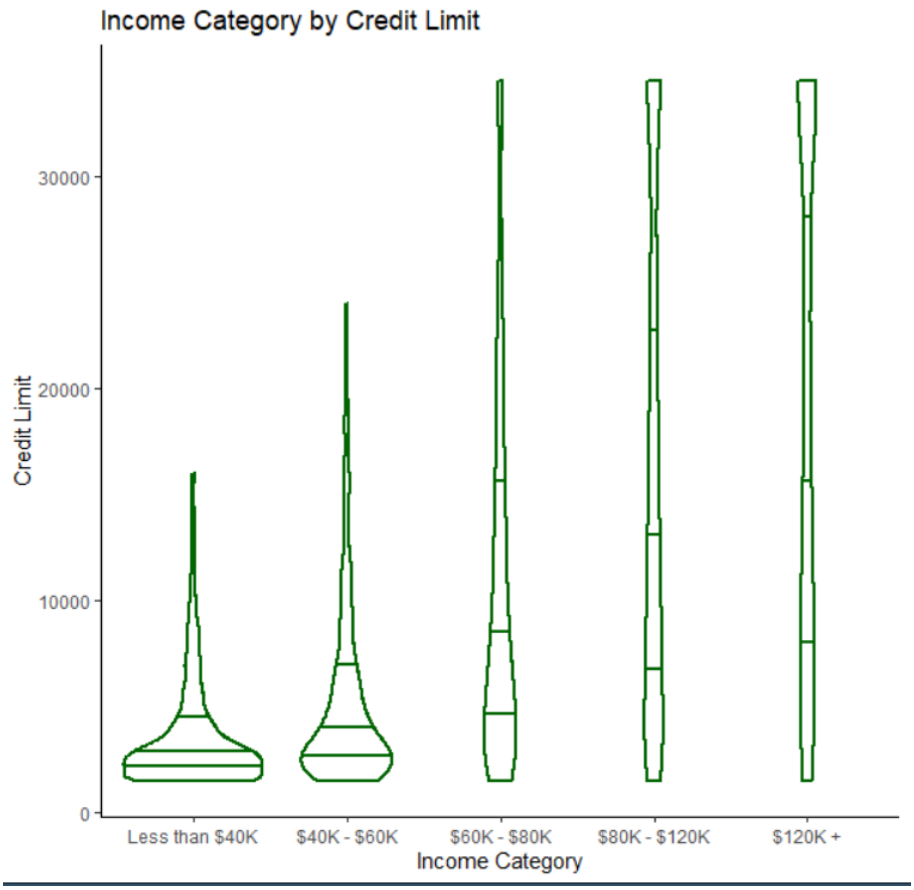
Question 8: What is the correlation between Income_Category and Credit_Limit?

Bivariate Analysis→Categorical vs. Continuous

Syntax

```
ggplot(pro1 , aes(Income_Category,Credit_Limit,color= Credit_Limit)) +
  geom_violin(draw_quantiles = c(0.25,0.5,0.75),colour="dark green",size=1.)+
  theme_classic()+xlab("Income Category") + ylab("Credit Limit") +
  ggtitle("Income Category by Credit Limit" )
```

Result



We can see here that the higher income category has a direct correlation to a higher credit limit

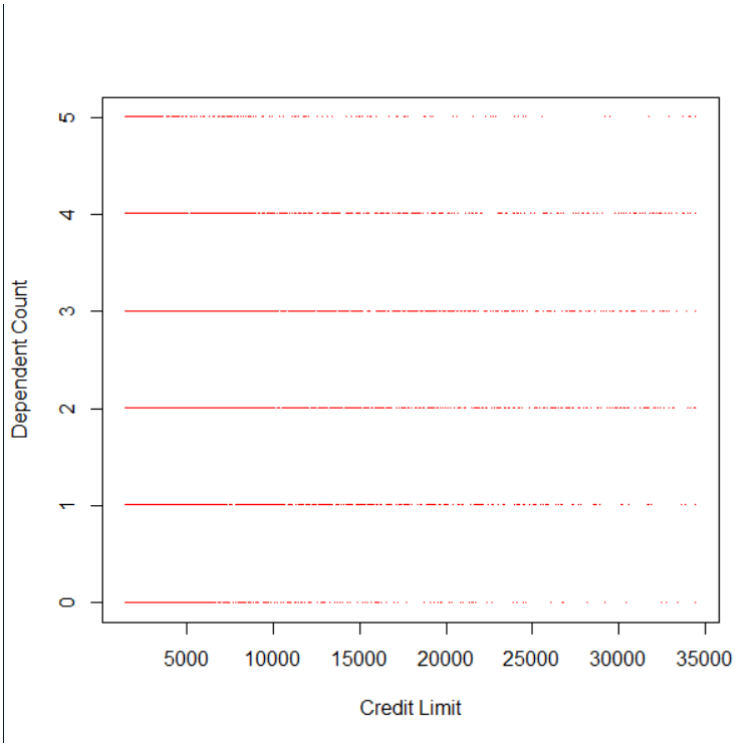
Question 9 - Question 9: What is the relation between Credit_Limit and Dependent_count?

Bivariate Analysis

Syntax

```
plot(x = pro1$Credit_Limit, y = pro1$Dependent_count,
     pch = 46,
     xlab = "Credit Limit", ylab = "Dependent Count", col = "red")
colors()
```

Result



The variable Credit_Limit is independent of Dependent_Count

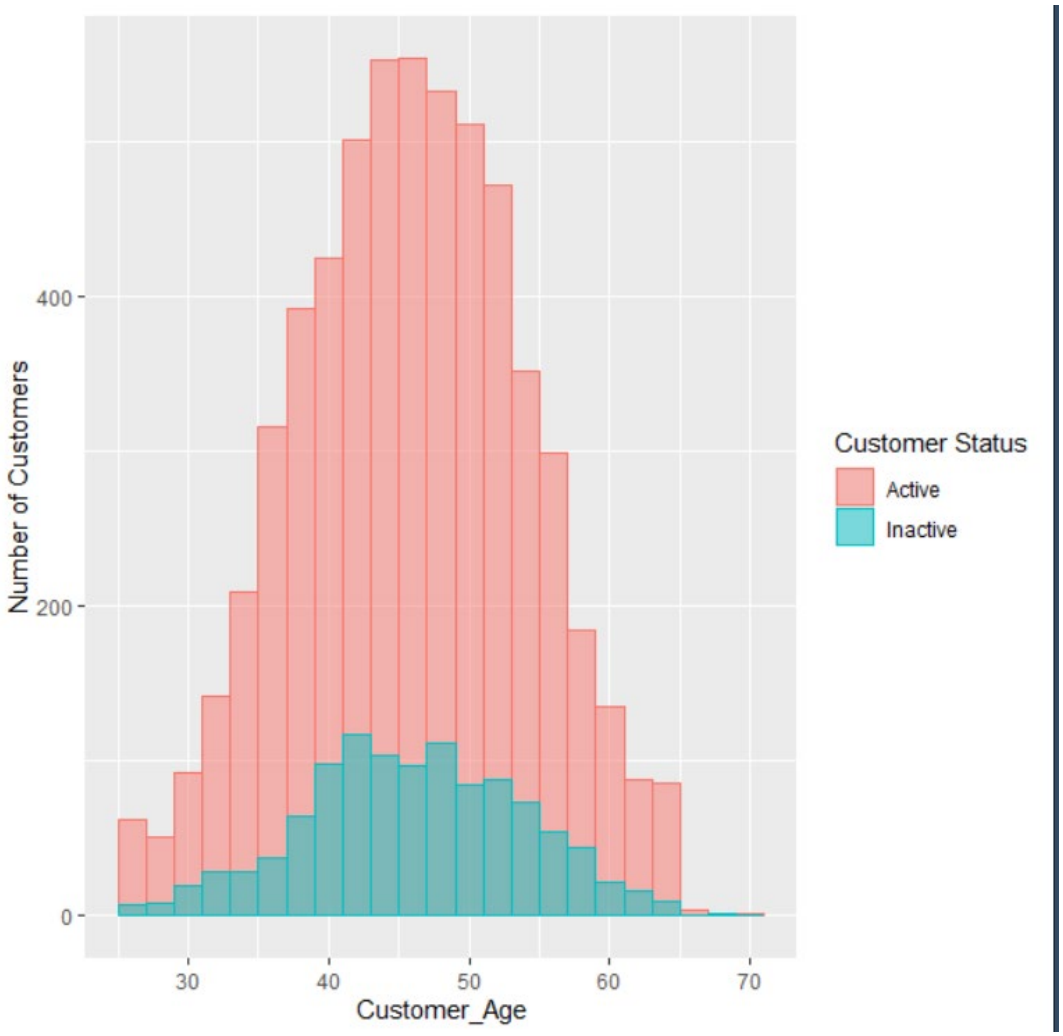
Question 10: Analyze using histogram Customer_Age vs Customer Status?

Bivariate Analysis

Syntax

```
ggplot(pro1, aes(x=Customer_Age, fill=`Customer Status`, color=`Customer Status`)) +
  geom_histogram(position="identity",binwidth = 2, alpha=0.5) + ylab("Number of Customers")
```

Result



THANK YOU