Personal Loan

Assel_Kassenova, Liu Xiuting, Pouroullis Bofilatos Eleni, Yin Haoran, Zhou Ziyu

5/8/2023

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(ggplot2)
library(rpart)
library(arules)
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
##
## Attaching package: 'arules'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
library(cluster)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
```

```
##   method from
##   +.gg   ggplot2
```

```
library(e1071)
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

PART 1: DATA PREPARATION

```
df <- read_csv("Bank_Personal_loan_Modelling.csv")
```

```
## Rows: 5000 Columns: 14
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (14): ID_Customer, Age, Experience, Income, ZIP Code, Family, CCAvg, Edu...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
dim(df)
```

```
## [1] 5000   14
```

Dataset has 5000 rows and 14 columns (variables)

```
colnames(df)
```

```
##  [1] "ID_Customer"        "Age"                "Experience"
##  [4] "Income"             "ZIP Code"           "Family"
##  [7] "CCAvg"              "Education"          "Mortgage"
## [10] "Personal Loan"      "Securities Account" "CD Account"
## [13] "Online"             "CreditCard"
```

The column names are customer ID (a unique identifier for each row in this dataset) Age, experience, income, zip code, family, CCavg, education, mortgage, personal loan, securities account, cd account, online and creditcard.

The dataset includes demographic and financial information of bank customers, and the target variable is whether a customer accepted the offer of a personal loan.

THE PROBLEM STATEMENT

for this dataset is to identify factors that influence customers' acceptance of personal loans and develop a predictive model to target potential loan customers.

```
# Clean the dataset:

sum(is.na(df))
```

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

```
# NO NA values in data set

# Check for duplicates -- none
sum(duplicated(df)) # No duplicates found
```

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

```
#Check for any impossible values:
summary1 <- summary(df)

#Experience has negative values ==> remove any negatives.
df2 <- df[df$Experience >= 0,]
```

We will remove the columns that will not assist us in building a predictive model: Customer ID and ZIP Code are both categorical variables and will be removed. CD Account & Online: Unclear as to what this variable is in the description so we will remove it.

```
colnames(df2)
```

```
##  [1] "ID_Customer"        "Age"                "Experience"
##  [4] "Income"             "ZIP Code"           "Family"
##  [7] "CCAvg"              "Education"          "Mortgage"
## [10] "Personal Loan"      "Securities Account" "CD Account"
## [13] "Online"             "CreditCard"
```

```
df2 <- df2[, -c(1, 5, 12, 13)]
colnames(df2)
```

```
##  [1] "Age"                "Experience"         "Income"
##  [4] "Family"             "CCAvg"              "Education"
##  [7] "Mortgage"           "Personal Loan"      "Securities Account"
## [10] "CreditCard"
```

```
df3 <- df2
df3
```

```
## # A tibble: 4,948 x 10
##      Age Experience Income Family CCAvg Education Mortgage `Personal Loan`
##    <dbl>      <dbl>  <dbl>  <dbl> <dbl>     <dbl>    <dbl>            <dbl>
##  1    25          1     49      4   1.6         1        0                0
##  2    45         19     34      3   1.5         1        0                0
##  3    39         15     11      1   1           1        0                0
##  4    35          9    100      1   2.7         2        0                0
##  5    35          8     45      4   1           2        0                0
##  6    37         13     29      4   0.4         2      155                0
##  7    53         27     72      2   1.5         2        0                0
##  8    50         24     22      1   0.3         3        0                0
##  9    35         10     81      3   0.6         2      104                0
## 10    34          9    180      1   8.9         3        0                1
## # i 4,938 more rows
## # i 2 more variables: `Securities Account` <dbl>, CreditCard <dbl>
```

```
colnames(df3)
```

```
##  [1] "Age"                "Experience"         "Income"
##  [4] "Family"             "CCAvg"              "Education"
##  [7] "Mortgage"           "Personal Loan"      "Securities Account"
## [10] "CreditCard"
```

PART 2: Descriptive analytics with visualizations

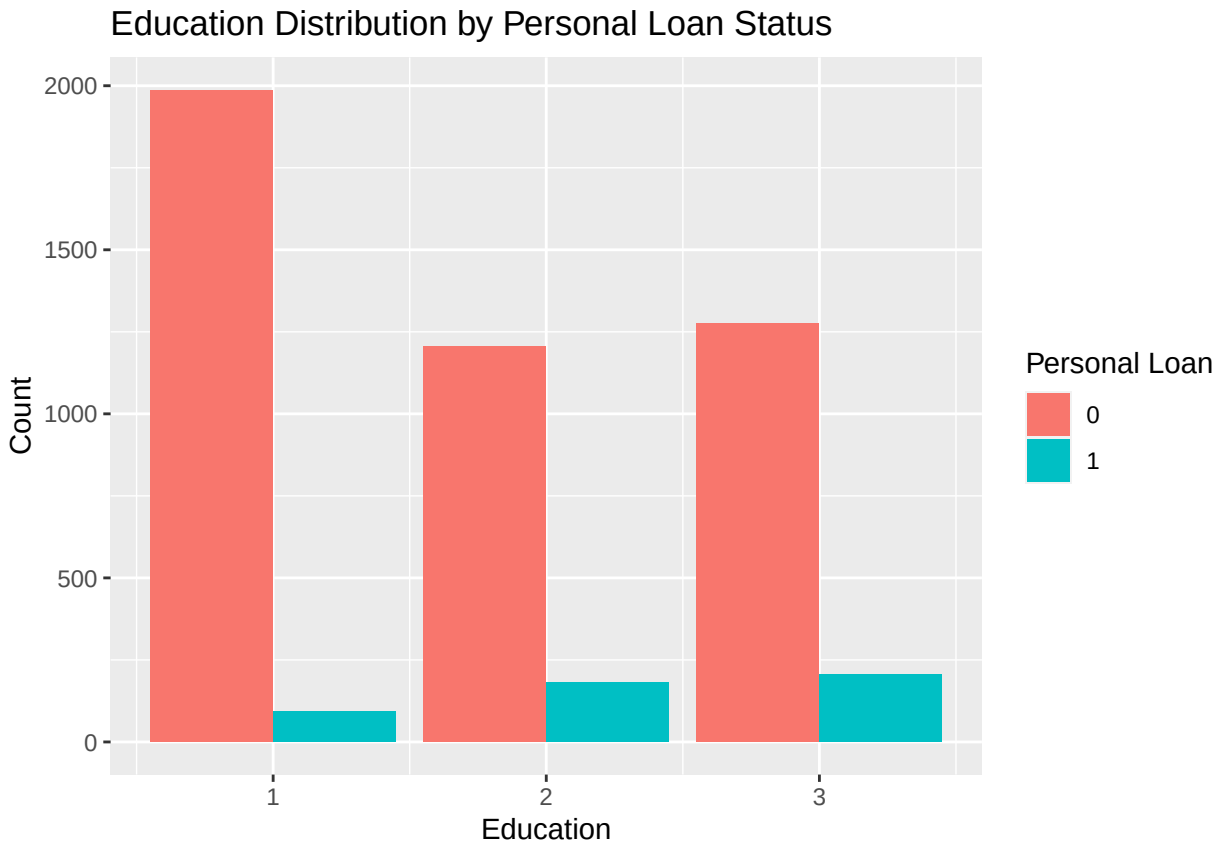
Let's look at a few of the variables and their relationships: Compare the distribution of Education between customers who accepted and did not accept the personal loan offer

The Link To our Vizualization Dashboard:

https://ad699.netlify.app/

```
df3 %>%
  group_by(`Personal Loan`, Education) %>%
  summarise(count = n()) %>%
  ggplot(aes(x = Education, y = count, fill = factor(`Personal Loan`))) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Education Distribution by Personal Loan Status",
       x = "Education", y = "Count", fill = "Personal Loan")
```

```
## `summarise()` has grouped output by 'Personal Loan'. You can override using the
## `.groups` argument.
```



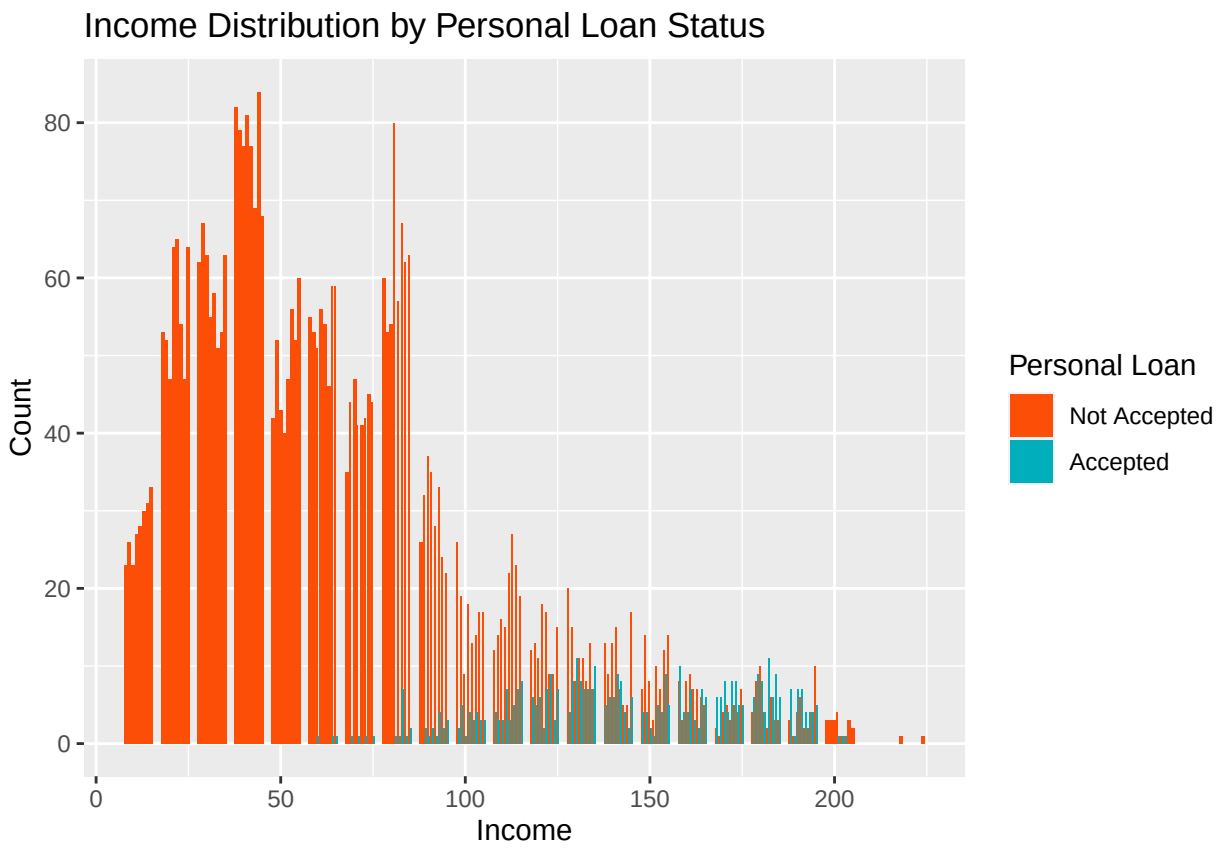
Level 1: Has the highest count, unaccepted and lowest count accepted, Level 3 contrarily has the lowest count unaccepted and highest count accepted. Education Level. 1: Undergrad; 2: Graduate;3: Advanced/Professional

```
# Income Distribution

df3 %>%
  group_by(`Personal Loan`, Income) %>%
  summarise(count = n()) %>%
  ggplot(aes(x = Income, y = count, fill = factor(`Personal Loan`))) +
```

```
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Income Distribution by Personal Loan Status",
       x = "Income", y = "Count", fill = "Personal Loan") +
  scale_fill_manual(values = c("0" = "#FC4E07", "1" = "#00AFBB"),
                    name = "Personal Loan", labels = c("Not Accepted", "Accepted"))
```

```
## `summarise()` has grouped output by 'Personal Loan'. You can override using the
## `.groups` argument.
```



We can see that lower income customers are more likely to not accept the personal loan offer. Higher income customers are more likely to accept, because they may have a greater ability to repay the loan or have a higher credit score.

The interest rate offered on a personal loan can be an important factor in the customer's decision to accept or decline the loan offer. It is possible that customers with lower incomes may be offered higher interest rates, which could make the loan less attractive to them. On the other hand, customers with higher incomes may be offered lower interest rates, which could make the loan more attractive to them

Compare the distribution of Securities Account between customers who accepted and did not accept the personal loan offer

```
# Securities Account Distribution by Personal Loan Status
df3 %>%
  group_by(`Personal Loan`, `Securities Account`) %>%
  summarise(count = n(), .groups = "keep") %>%
  mutate(`Personal Loan` = ifelse(`Personal Loan` == 0, "Not Accepted", "Accepted")) %>%
  ggplot(aes(x = factor(`Securities Account`), y = count, fill = `Personal Loan`)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Securities Account Distribution by Personal Loan Status",
```

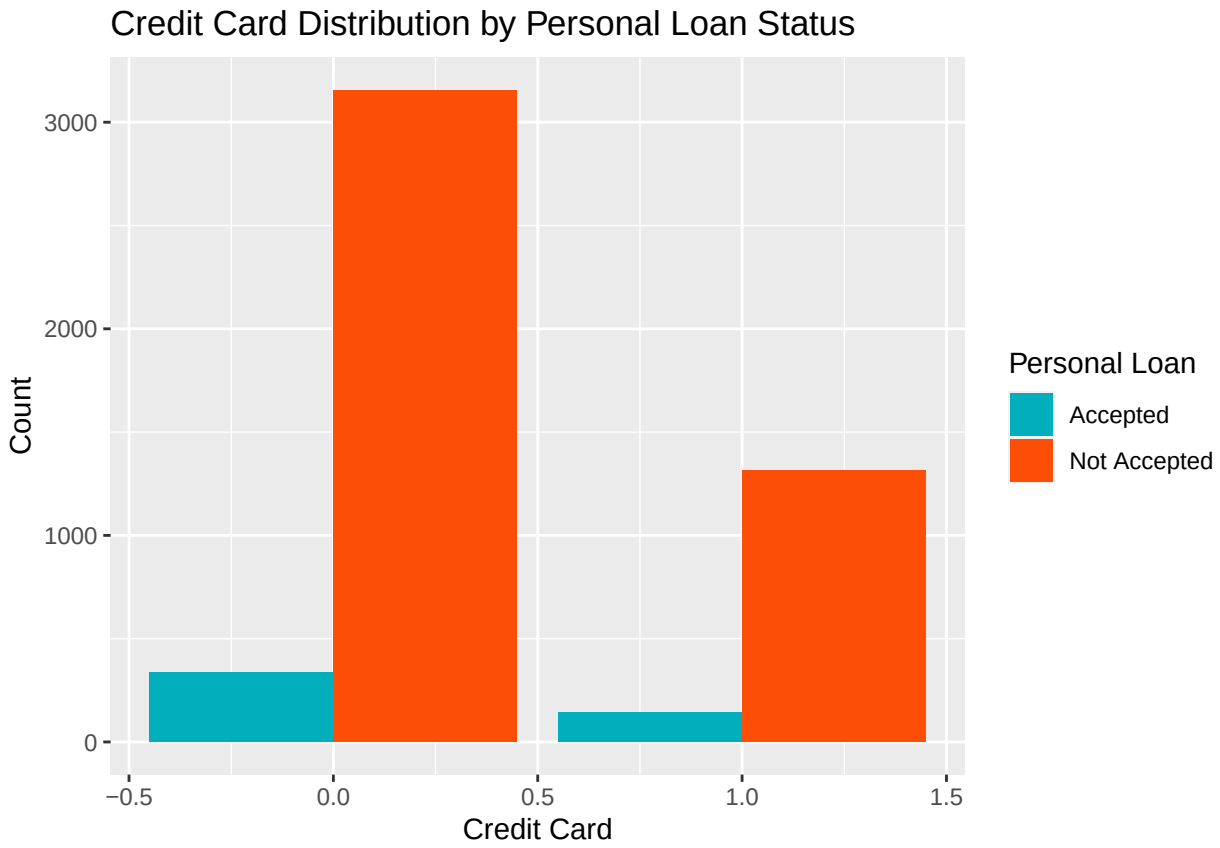
```
      x = "Securities Account", y = "Count") +
  scale_fill_manual(values = c("Accepted" = "#00AFBB", "Not Accepted" = "#FC4E07"))
```

A securities account typically does not have a direct impact on whether a person is approved or denied for a personal loan. We can consider not including this variable in our model.

```
# Credit Card Distribution by Personal Loan Status

df3 %>%
  group_by(`Personal Loan`, CreditCard) %>%
  summarise(count = n()) %>%
  mutate(`Personal Loan` = ifelse(`Personal Loan` == 0, "Not Accepted", "Accepted")) %>%
  ggplot(aes(x = CreditCard, y = count, fill = `Personal Loan`)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Credit Card Distribution by Personal Loan Status",
       x = "Credit Card", y = "Count", fill = "Personal Loan") +
  scale_fill_manual(values = c("Not Accepted" = "#FC4E07", "Accepted" = "#00AFBB"))
```

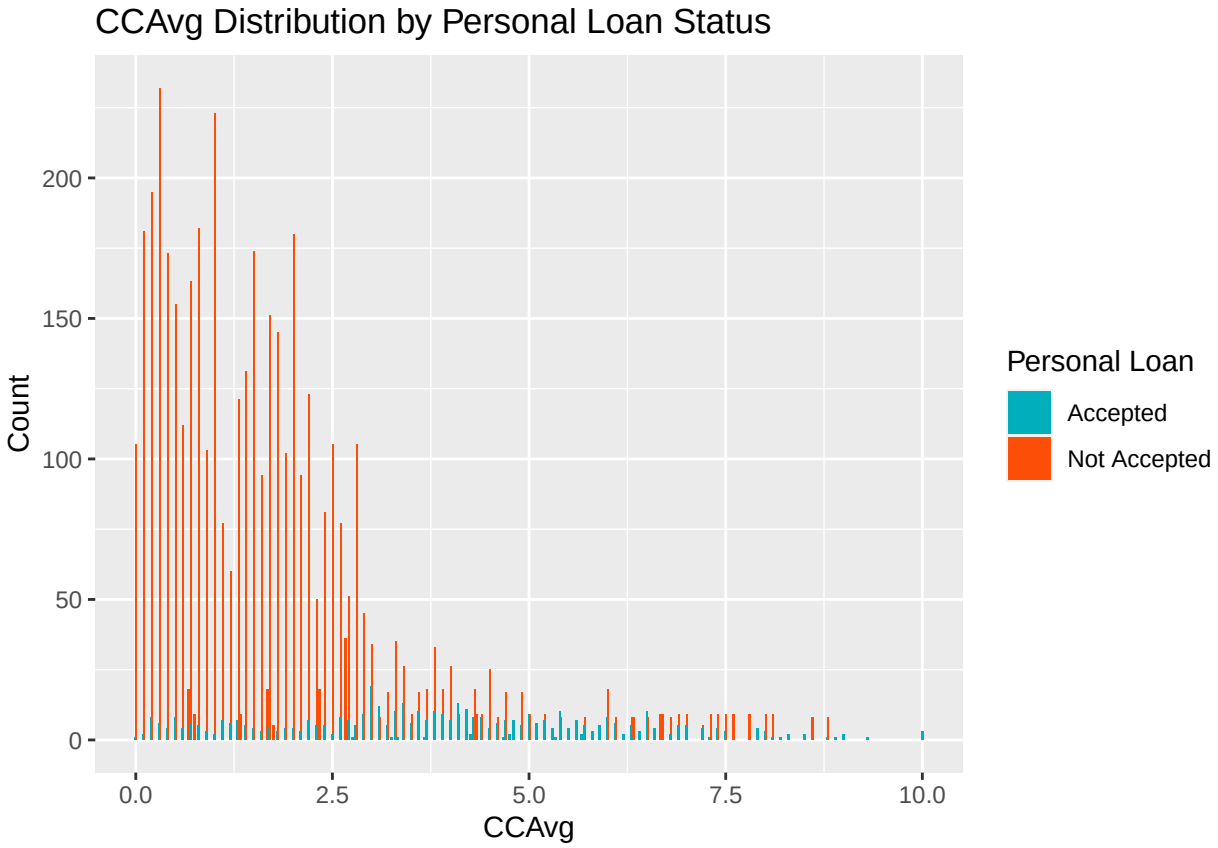
```
## `summarise()` has grouped output by 'Personal Loan'. You can override using the
## `.groups` argument.
```



Most customers that do not have a credit card did not accept the loan offer. Additionally, customers who do not have a credit card may simply be less interested in borrowing money or may prefer to use other means of financing One possible explanation could be that individuals who do not have a credit card may have a lower credit score, which may make it harder for them to get approved for loans from other financial institutions. Thus, they may be more likely to accept a loan offer from the bank even if the interest rates are higher On the other hand, individuals with credit cards may have a higher credit score and may have access to other loan options with better terms and lower interest rates, making them less likely to accept the loan offer from the bank.

```
df3 %>%
  group_by(`Personal Loan`, CCAvg) %>%
  summarise(count = n()) %>%
  mutate(`Personal Loan` = ifelse(`Personal Loan` == 0, "Not Accepted", "Accepted")) %>%
  ggplot(aes(x = CCAvg, y = count, fill = `Personal Loan`)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "CCAvg Distribution by Personal Loan Status", x = "CCAvg",
       y = "Count", fill = "Personal Loan") +
  scale_fill_manual(values = c("Not Accepted" = "#FC4E07", "Accepted" = "#00AFBB"))
```

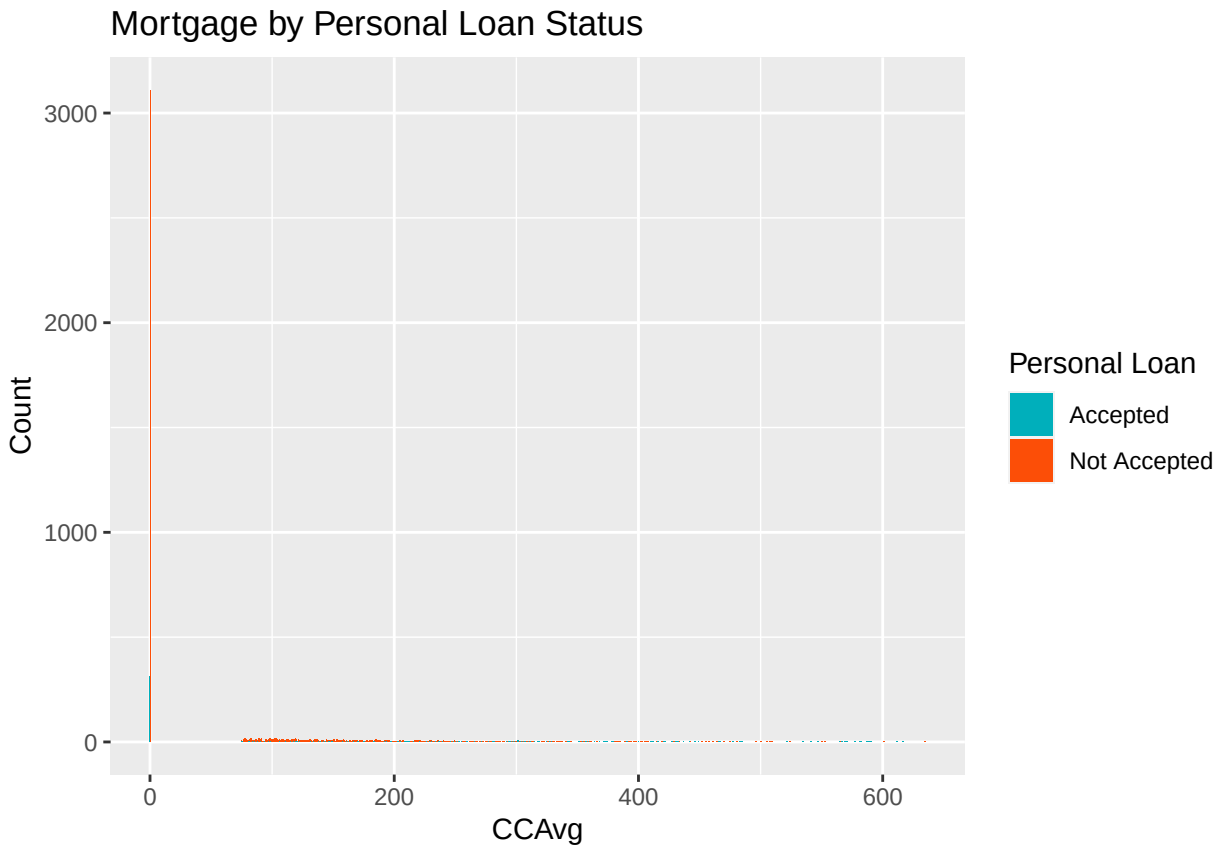
```
## `summarise()` has grouped output by 'Personal Loan'. You can override using the
## `.groups` argument.
```



Those who did not accept has a lower CCavg:Average spending on credit cards per month. This relationship may indicate that customers who spend more on their credit cards may have a higher income or a better credit score, which makes them more eligible for a personal loan offer. Alternatively, it may be that customers who accepted the personal loan offer did so in order to pay off their credit card debt, which could explain the higher credit card spending.

```
df3 %>%
  group_by(`Personal Loan`, Mortgage) %>%
  summarise(count = n()) %>%
  mutate(`Personal Loan` = ifelse(`Personal Loan` == 0, "Not Accepted", "Accepted")) %>%
  ggplot(aes(x = Mortgage, y = count, fill = `Personal Loan`)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Mortgage by Personal Loan Status",
       x = "CCAvg", y = "Count", fill = "Personal Loan") +
  scale_fill_manual(values = c("Not Accepted" = "#FC4E07", "Accepted" = "#00AFBB"))
```

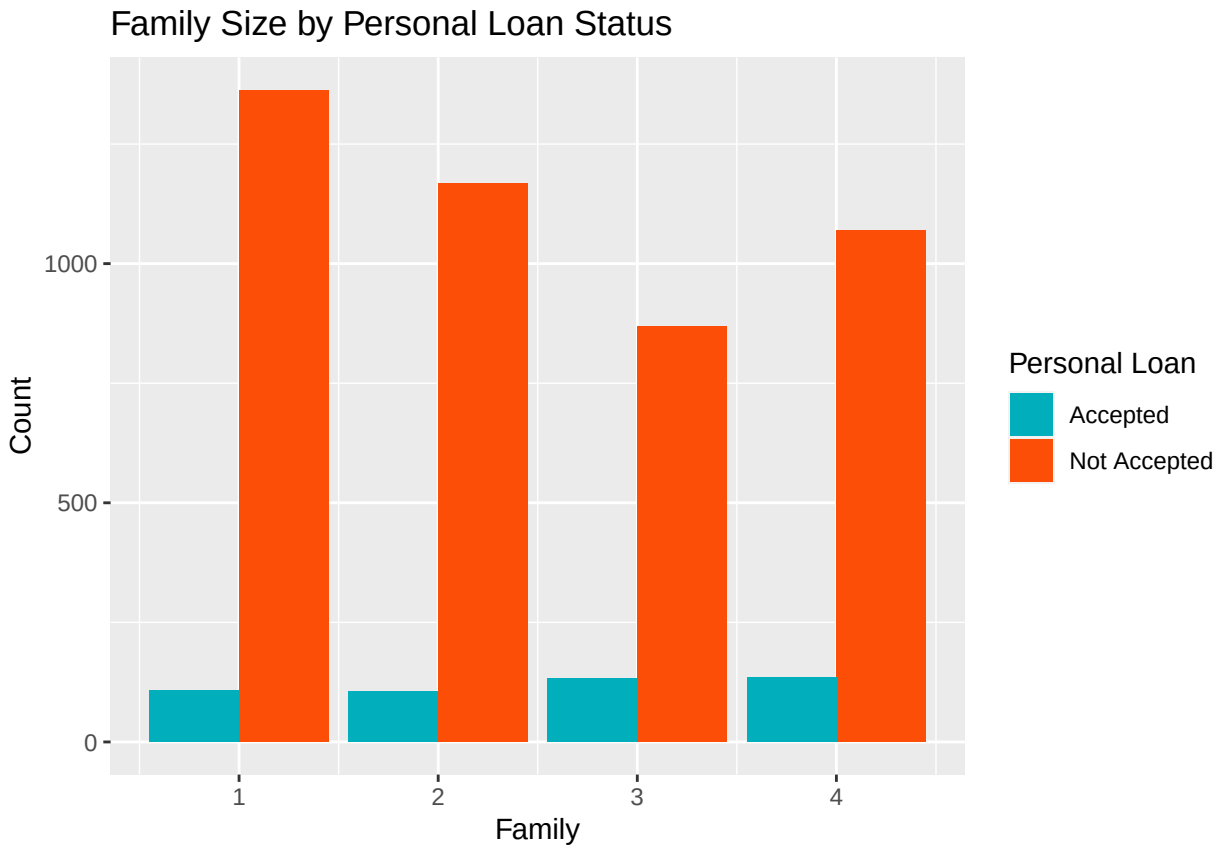
```
## `summarise()` has grouped output by 'Personal Loan'. You can override using the
## `.groups` argument.
```



majority of customers in the dataset did not have a mortgage, hence the low count.

```
df3 %>%
  group_by(`Personal Loan`, Family) %>%
  summarise(count = n()) %>%
  mutate(`Personal Loan` = ifelse(`Personal Loan` == 0, "Not Accepted", "Accepted")) %>%
  ggplot(aes(x = Family, y = count, fill = `Personal Loan`)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Family Size by Personal Loan Status",
       x = "Family", y = "Count", fill = "Personal Loan") +
  scale_fill_manual(values = c("Not Accepted" = "#FC4E07", "Accepted" = "#00AFBB"))
```

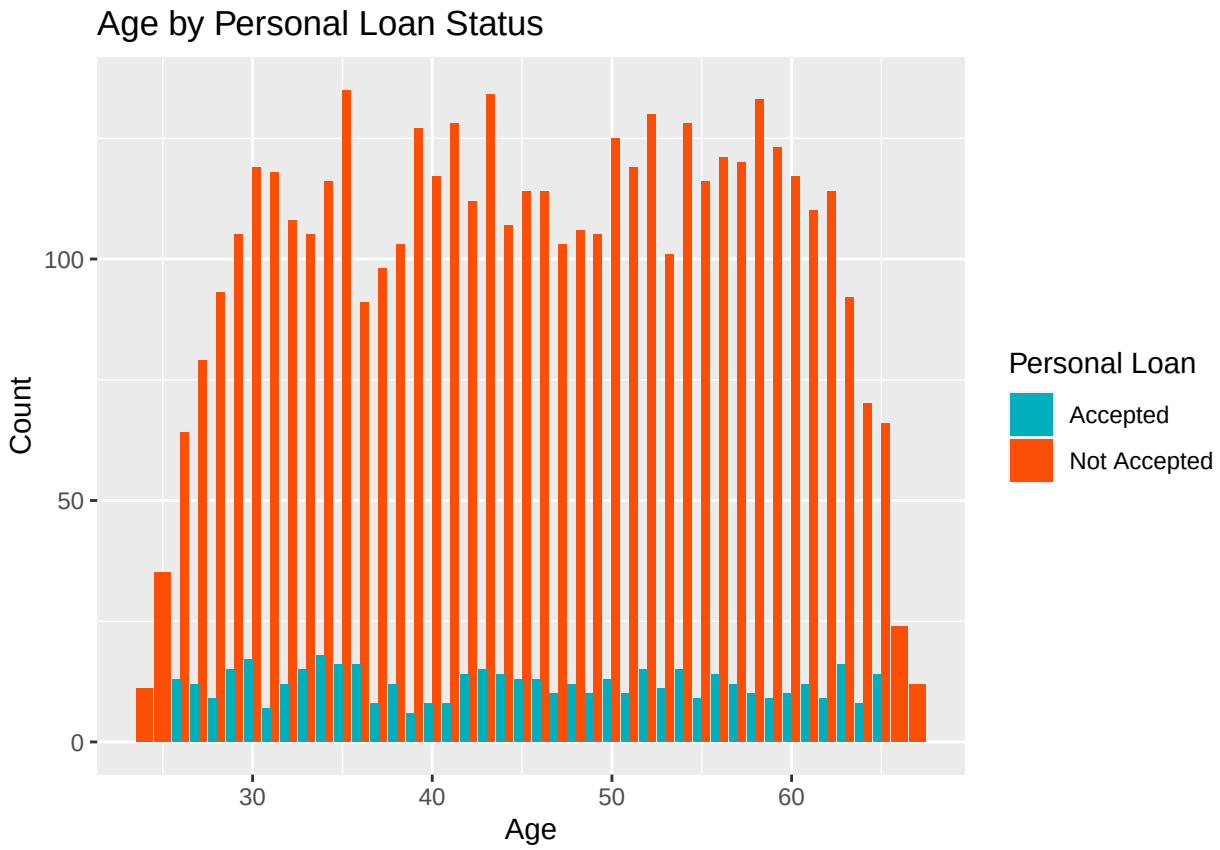
```
## `summarise()` has grouped output by 'Personal Loan'. You can override using the
## `.groups` argument.
```



It appears that individuals with larger families (more than three members) were more likely to accept personal loans compared to those with smaller families (less than or equal to three members) individuals with larger families may have more financial responsibilities and expenses, such as children's education or healthcare, and may therefore have a greater need for additional financial resources

```
df3 %>%
  group_by(`Personal Loan`, Age) %>%
  summarise(count = n()) %>%
  mutate(`Personal Loan` = ifelse(`Personal Loan` == 0, "Not Accepted", "Accepted")) %>%
  ggplot(aes(x = Age, y = count, fill = `Personal Loan`)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Age by Personal Loan Status",
       x = "Age", y = "Count", fill = "Personal Loan") +
  scale_fill_manual(values = c("Not Accepted" = "#FC4E07", "Accepted" = "#00AFBB"))
```

```
## `summarise()` has grouped output by 'Personal Loan'. You can override using the
## `.groups` argument.
```




```
df3 %>%
  group_by(`Personal Loan`, Experience) %>%
  summarise(count = n()) %>%
  mutate(`Personal Loan` = ifelse(`Personal Loan` == 0, "Not Accepted", "Accepted")) %>%
  ggplot(aes(x = Experience, y = count, fill = `Personal Loan`)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Years Experience by Personal Loan Status",
       x = "Experience", y = "Count", fill = "Personal Loan") +
  scale_fill_manual(values = c("Not Accepted" = "#FC4E07", "Accepted" = "#00AFBB"))
```

```
## `summarise()` has grouped output by 'Personal Loan'. You can override using the
## `.groups` argument.
```

AGE + Experience From the graphs, it appears that there are no significant differences in the distribution of ages and experiences between those who accepted and those who did not accept personal loans. This suggests that age and experience may not be strong predictors of whether an individual is likely to accept a personal loan or not.

```
colnames(df3)
```

```
##  [1] "Age"                "Experience"         "Income"
##  [4] "Family"             "CCAvg"              "Education"
##  [7] "Mortgage"           "Personal Loan"      "Securities Account"
## [10] "CreditCard"
```

Check correlation of variables:

```
cor(df3[,c("Age", "Income","CCAvg","Mortgage","Education",
           "Experience", "Family", "CreditCard", "Securities Account")])
```

```
##                            Age       Income        CCAvg     Mortgage
## Age                 1.0000000000 -0.058005788 -0.050878873 -0.015183972
## Income             -0.0580057884  1.000000000  0.646178154  0.206920864
## CCAvg              -0.0508788729  0.646178154  1.000000000  0.109904537
## Mortgage           -0.0151839719  0.206920864  0.109904537  1.000000000
## Education           0.0462218656 -0.187992205 -0.133939348 -0.032558771
## Experience          0.9941014890 -0.049244828 -0.048938962 -0.013459294
## Family             -0.0392786041 -0.155665764 -0.107230226 -0.020418837
## CreditCard          0.0074956879 -0.004493256 -0.007376833 -0.006909691
## Securities Account  0.0004958185 -0.002327176  0.012476822 -0.003717143
##                       Education    Experience      Family   CreditCard
## Age                 0.046221866  0.9941014890 -0.03927860  0.007495688
```

```
## Income             -0.187992205 -0.0492448284 -0.15566576 -0.004493256
## CCAvg              -0.133939348 -0.0489389617 -0.10723023 -0.007376833
## Mortgage           -0.032558771 -0.0134592936 -0.02041884 -0.006909691
## Education           1.000000000  0.0182432552  0.06403235 -0.012604345
## Experience          0.018243255  1.0000000000 -0.04560995  0.008876095
## Family              0.064032351 -0.0456099542  1.00000000  0.012905354
## CreditCard         -0.012604345  0.0088760955  0.01290535  1.000000000
## Securities Account -0.007508034 -0.0004570698  0.02015521 -0.017030269
##                    Securities Account
## Age                     0.0004958185
## Income                 -0.0023271758
## CCAvg                   0.0124768215
## Mortgage               -0.0037171431
## Education              -0.0075080343
## Experience             -0.0004570698
## Family                  0.0201552065
## CreditCard             -0.0170302693
## Securities Account      1.0000000000
```

Age and Experience have a very strong positive correlation of 0.994, indicating that they are highly correlated. Income and CCAvg have a strong positive correlation of 0.646, indicating that people with higher incomes tend to have higher credit card spending. None of the variables have a very strong correlation with Securities Account, with the highest correlation being only 0.012.

```
#Remove Age variable due to multicollinearity concerns:
df3 <- df3[, !colnames(df3) %in% c("Age", "Securities Account")]
```

Scatterplot Matrix Convert relevant variables to the appropriate data types

```
df3$Education <- as.factor(df3$Education)
#df3$Personal Loan <- as.factor(df3$'Personal Loan')
colnames(df3)
```

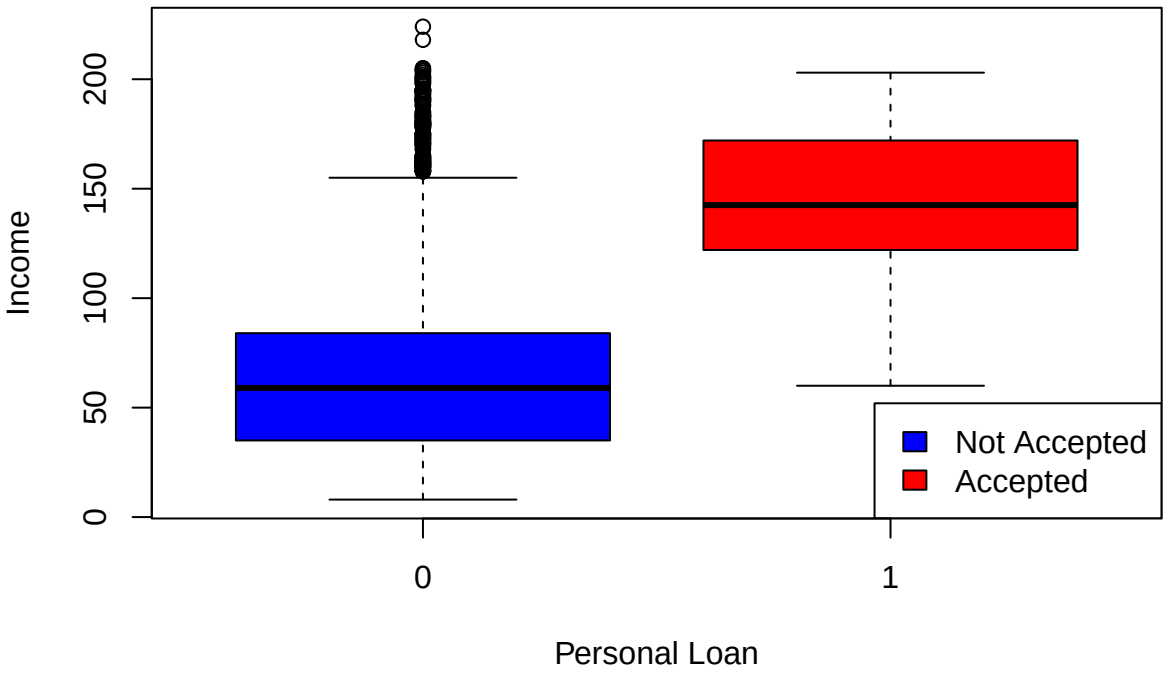
```
## [1] "Experience"    "Income"        "Family"        "CCAvg"
## [5] "Education"     "Mortgage"      "Personal Loan" "CreditCard"
```

```
# Plot the pairs using ggpairs
ggpairs(df3[, c("Income", "CCAvg", "Mortgage", "Education", "Personal Loan",
                "Experience", "Family","CreditCard")],
        columns = c(1:8),
        mapping = aes(fill = 'Personal Loan', color = 'Personal.Loan'))
```

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

CCAvg and Income have a moderate to strong positive correlation Overall, the scatterplot matrix suggests that there may not be strong linear relationships between the independent variables and the target variable such that multicollinearity is present. Multicollinearity may not be a major concern, as there are no strong pairwise correlations between the independent variables.

```
#Boxplot
boxplot(Income ~ `Personal Loan`, data = df3, col = c("blue", "red"),
        xlab = "Personal Loan", ylab = "Income")
plot.window(xlim = c(0.9, 2.1), ylim = c(0, 250), log = "", yaxs = "i")
legend("bottomright", legend = c("Not Accepted", "Accepted"), fill = c("blue", "red"))
```



Higher incomes more likely to accept personal loan.

PART 3: Building a model

Use logistic regression because the dependent variable (personal loan) is binary.

```
#a: Logistic Regression
model <- glm(`Personal Loan` ~ Experience + Income + Family +
               CCAvg + Education + `Mortgage` + `CreditCard`, data = df3, family = "binomial")

summary(model)
```

```
##
## Call:
## glm(formula = `Personal Loan` ~ Experience + Income + Family +
##     CCAvg + Education + Mortgage + CreditCard, family = "binomial",
##     data = df3)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.344e+01  5.522e-01 -24.346  < 2e-16 ***
## Experience   7.955e-03  6.432e-03   1.237    0.216
## Income       5.952e-02  2.718e-03  21.894  < 2e-16 ***
## Family       5.980e-01  7.167e-02   8.345  < 2e-16 ***
## CCAvg        1.723e-01  4.092e-02   4.211 2.55e-05 ***
## Education2   3.927e+00  2.533e-01  15.504  < 2e-16 ***
## Education3   3.994e+00  2.483e-01  16.086  < 2e-16 ***
## Mortgage     9.598e-04  5.644e-04   1.701    0.089 .
## CreditCard   2.381e-02  1.604e-01   0.148    0.882
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 3151.5  on 4947  degrees of freedom
## Residual deviance: 1321.2  on 4939  degrees of freedom
## AIC: 1339.2
##
## Number of Fisher Scoring iterations: 8
```

Look for high p values: Credit Card has a high p-value of 0.888, which means that it is not statistically significant in predicting the response variable. We can therefore remove it

```
model2 <- glm(`Personal Loan` ~Income + Family +
               CCAvg + Education, data = df3, family = "binomial")

summary(model2)
```

```
##
## Call:
## glm(formula = `Personal Loan` ~ Income + Family + CCAvg + Education,
##     family = "binomial", data = df3)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -13.20524    0.52168 -25.313  < 2e-16 ***
## Income        0.05986    0.00270  22.167  < 2e-16 ***
## Family        0.60018    0.07171   8.370  < 2e-16 ***
## CCAvg         0.16216    0.04058   3.996 6.44e-05 ***
## Education2    3.89841    0.25136  15.509  < 2e-16 ***
## Education3    3.96062    0.24593  16.104  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3151.5  on 4947  degrees of freedom
## Residual deviance: 1325.6  on 4942  degrees of freedom
## AIC: 1337.6
##
## Number of Fisher Scoring iterations: 8
```

the remaining variables in the model are all statistically significant (indicated by their low p-values).

```
set.seed(2410)
train.index <- sample(nrow(df3), 0.6 * nrow(df3))
train <- df3[train.index, ]
test <- df3[-train.index, ]

model_final <- glm(`Personal Loan` ~Income + Family +
                     CCAvg + Education,
                   family = "binomial", data = train)

predicted <- predict(model_final, newdata = test, type = "response")

# Converting predicted probabilities to binary outcomes
predicted <- ifelse(predicted >= 0.5, 1, 0)
```

```r
# Evaluating the model
confusion_matrix <- table(predicted, test$`Personal Loan`)
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)

confusion_matrix
```

```
##
## predicted    0    1
##         0 1774   54
##         1   31  121
```

accuracy #95.70%

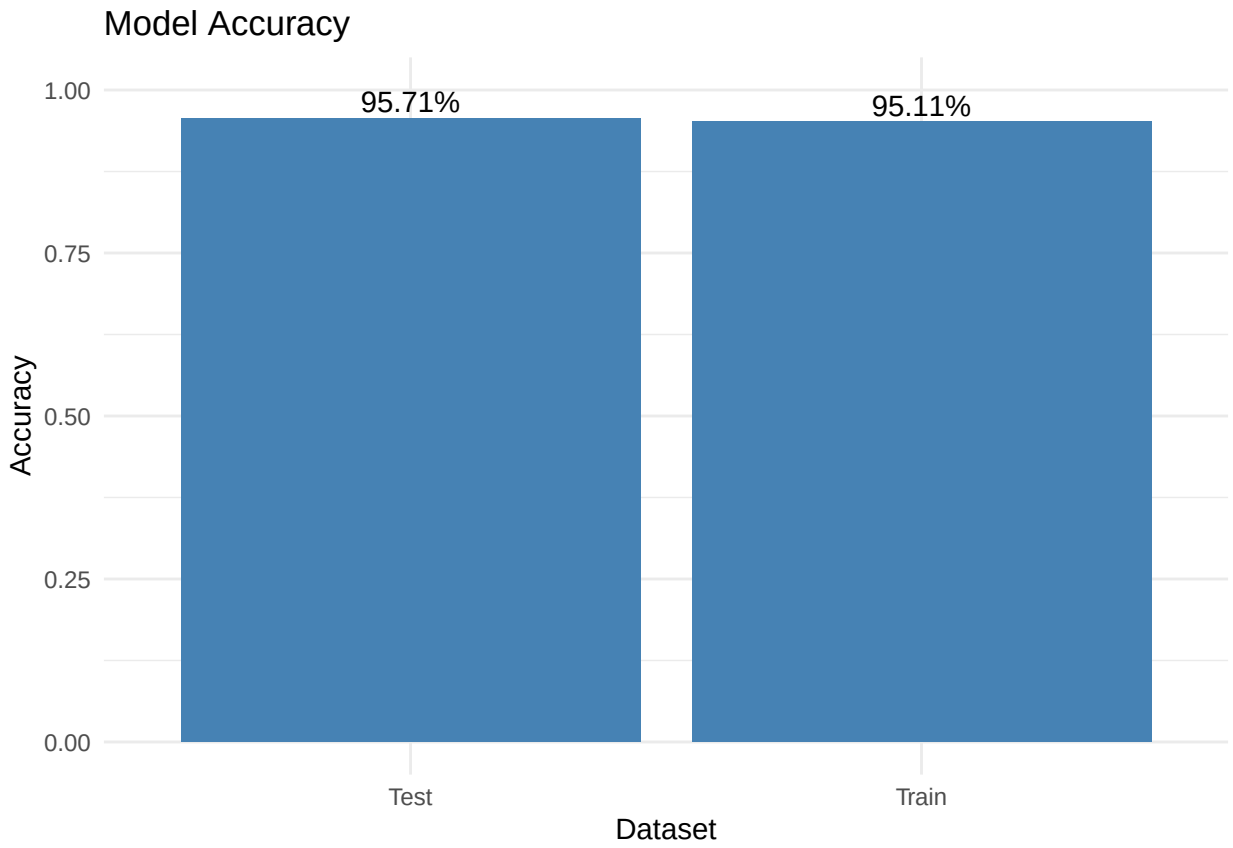
```r
# Train data accuracy
train_predicted <- predict(model_final, newdata = train, type = "response")
train_predicted <- ifelse(train_predicted >= 0.5, 1, 0)
train_confusion_matrix <- table(train_predicted, train$`Personal Loan`)
train_accuracy <- sum(diag(train_confusion_matrix)) / sum(train_confusion_matrix)

# Test data accuracy
test_predicted <- predict(model_final, newdata = test, type = "response")
test_predicted <- ifelse(test_predicted >= 0.5, 1, 0)
test_confusion_matrix <- table(test_predicted, test$`Personal Loan`)
test_accuracy <- sum(diag(test_confusion_matrix)) / sum(test_confusion_matrix)

accuracy_df <- data.frame(Dataset = c("Train", "Test"),
                          Accuracy = c(train_accuracy, test_accuracy))
```

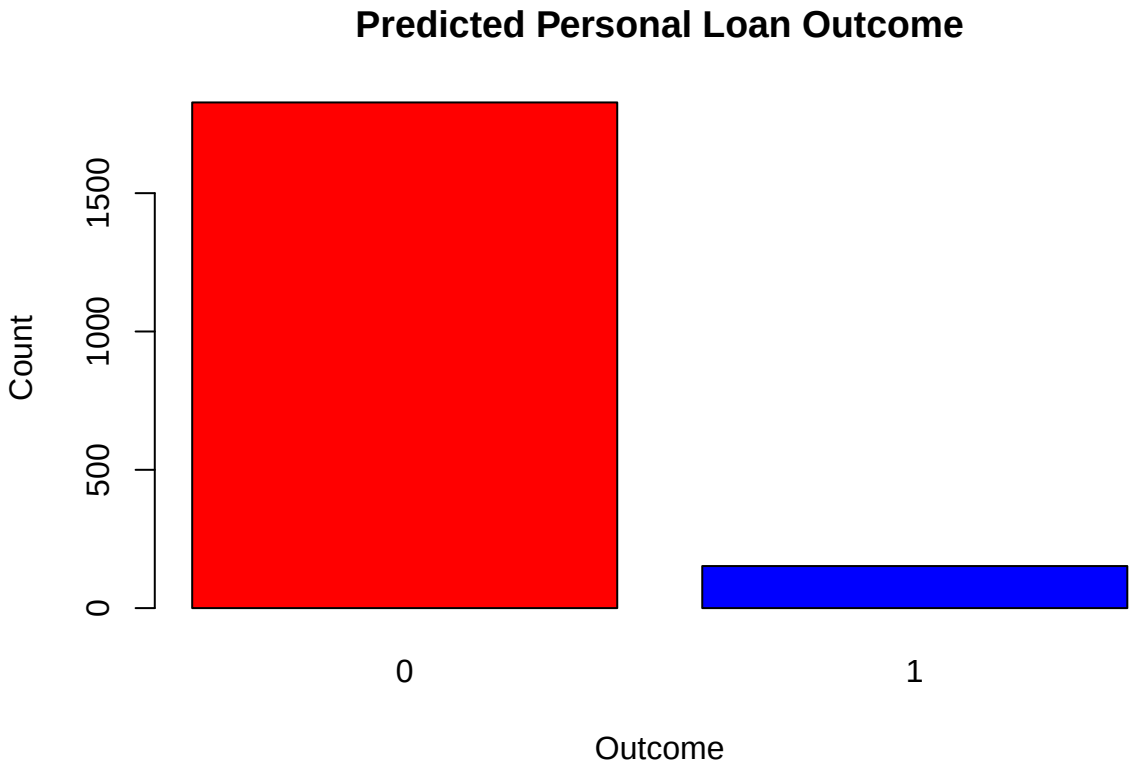
```r
accuracy_plot <- ggplot(accuracy_df, aes(x = Dataset, y = Accuracy)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Model Accuracy",
       x = "Dataset",
       y = "Accuracy") +
  ylim(0, 1) +
  geom_text(aes(label = paste0(round(Accuracy * 100, 2), "%")), vjust = -0.3, size = 4) +
  theme_minimal()

accuracy_plot
```




```
# Count the occurrences of predicted 0 and 1
predicted_table <- table(predicted)

# Plot the counts as a barplot
barplot(predicted_table,
        main = "Predicted Personal Loan Outcome",
        xlab = "Outcome",
        ylab = "Count",
        col = c("red", "blue"))
```



Seems to be imbalanced towards predicting not accept outcome.

Now lets predict a customer:

```
new_customer <- data.frame(
  Income = 80000,
  Family = 3,
  CCAvg = 2,
  Education = 1
)
```

```
# Convert 'Education' variable to factor in new customer data
new_customer$Education <- factor(new_customer$Education)

# Make the prediction
prediction <- predict(model2, newdata = new_customer, type = "response")
prediction
```

```
## 1
## 1
```

This customer will accept the personal loan offer. Good to see it is predicting an outcome fo 1 given the imbalance.

Cross Validation

Logistic Regression Model

```
set.seed(2410)

df3$`Personal Loan` <- as.factor(df3$`Personal Loan`) # Convert outcome variable to a factor
```

```
train_control <- trainControl(method="cv", number=10)
cv_model <- train(`Personal Loan` ~Income + Family +
                    CCAvg + Education, data=df3, method="glm", family="binomial", trControl=train_contro

accuracy <- cv_model$results$Accuracy
kappa <- cv_model$results$Kappa
accuracy
```

```
## [1] 0.9541197
```

```
kappa
```

```
## [1] 0.7068326
```

Acc:0.95412 Kappa: 0.7068 a moderate level of agreement between the model's predictions and the actual outcomes.

K-Nearest Neighbors

```
train.index <- sample(nrow(df3), 0.6 * nrow(df3))
train <- df3[train.index, ]
test <- df3[-train.index, ]


train_norm <- train
train_norm[, c("Income", "Family", "CCAvg")] <- scale(train_norm[, c("Income", "Family", "CCAvg")])

valid_norm <- test
valid_norm[, c("Income", "Family", "CCAvg")] <- scale(valid_norm[, c("Income", "Family", "CCAvg")])

train_norm <- data.frame(train_norm)
valid_norm <- data.frame(valid_norm)

colnames(train_norm)
```

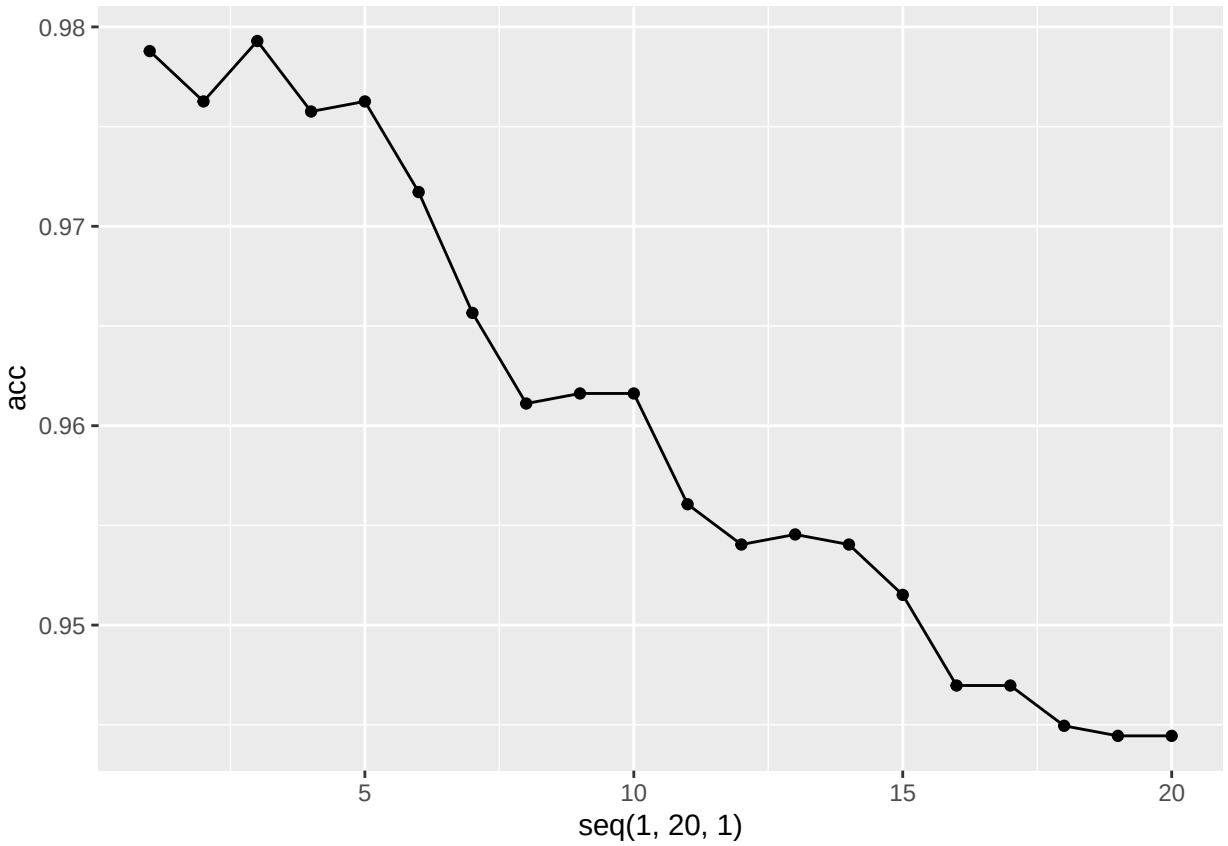
```
## [1] "Experience"    "Income"        "Family"        "CCAvg"
## [5] "Education"     "Mortgage"      "Personal.Loan" "CreditCard"
```

We exclude Education as it is a categorical variable. Find the best value of K

```
library(class)
set.seed(2410)

acc <- c()
for (k in seq(1, 20, 1)) {
  m <- knn(train = train_norm[, c(1:4, 7, 8)], test = valid_norm[, c(1:4,7,8)],
           cl = train_norm[, 7], k = k)
  acc <- c(acc, mean(valid_norm[, 7] == m))
}

ggplot() + geom_point(aes(x = seq(1, 20, 1), y = acc)) +
  geom_line(aes(x = seq(1, 20, 1), y = acc))
```




```
model_knn <- knn(train = train_norm[, c(1:4, 7, 8)],
                 test = valid_norm[, c(1:4, 7, 8)], cl = train_norm[, 7], k = 3)

sum(model_knn == valid_norm[, 7]) / nrow(valid_norm)
```

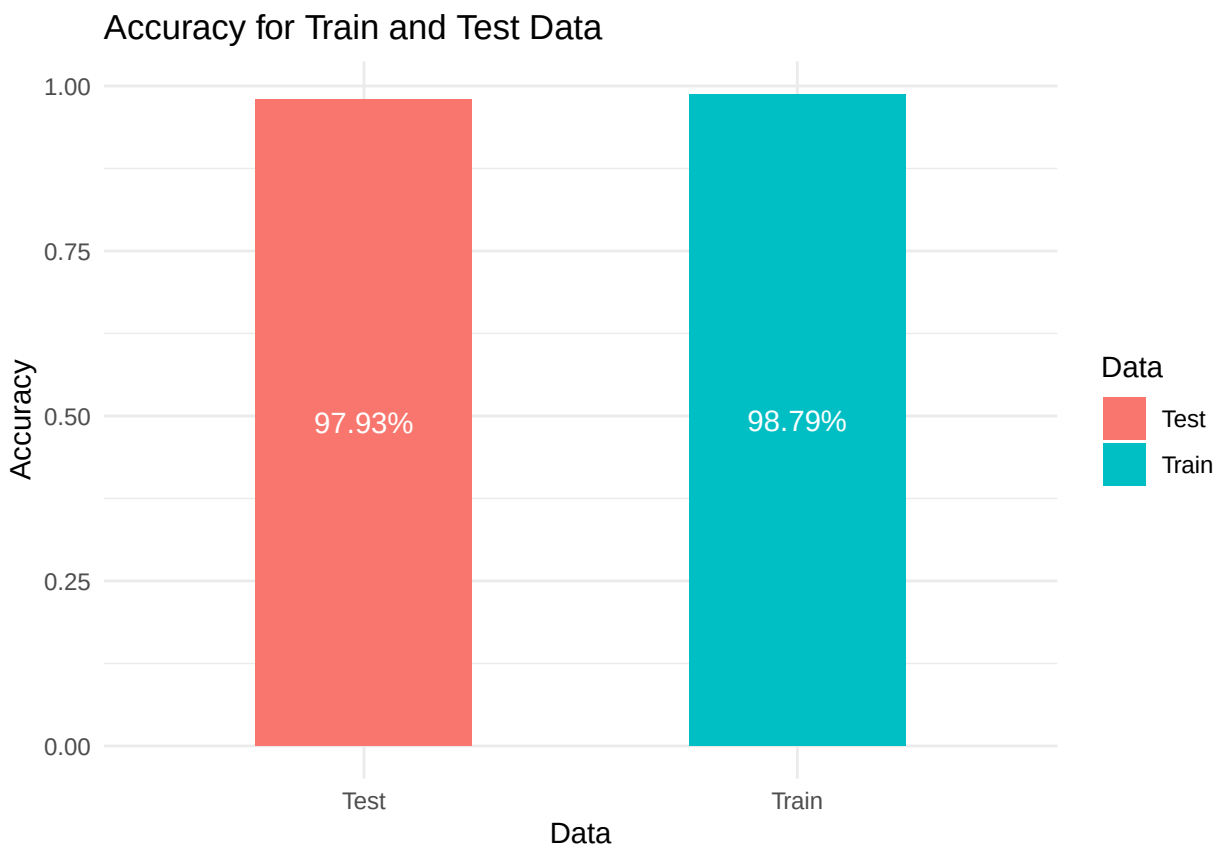
```
## [1] 0.9792929
```

Accuracy: 97.92%

```
train_accuracy <- sum(train_norm[, 7] == knn(train_norm[, c(1:4, 7, 8)], train_norm[, c(1:4, 7, 8)], cl
test_accuracy <- sum(valid_norm[, 7] == model_knn) / nrow(valid_norm)

accuracy_df <- data.frame(Data = c("Train", "Test"),
                          Accuracy = c(train_accuracy, test_accuracy))


ggplot(accuracy_df, aes(x = Data, y = Accuracy, fill = Data)) +
  geom_bar(stat = "identity", width = 0.5) +
  geom_text(aes(label = paste0(round(Accuracy * 100, 2), "%")),
            position = position_stack(vjust = 0.5),
            color = "white", size = 4) +
  labs(x = "Data", y = "Accuracy", fill = "Data") +
  ggtitle("Accuracy for Train and Test Data") +
  theme_minimal()
```

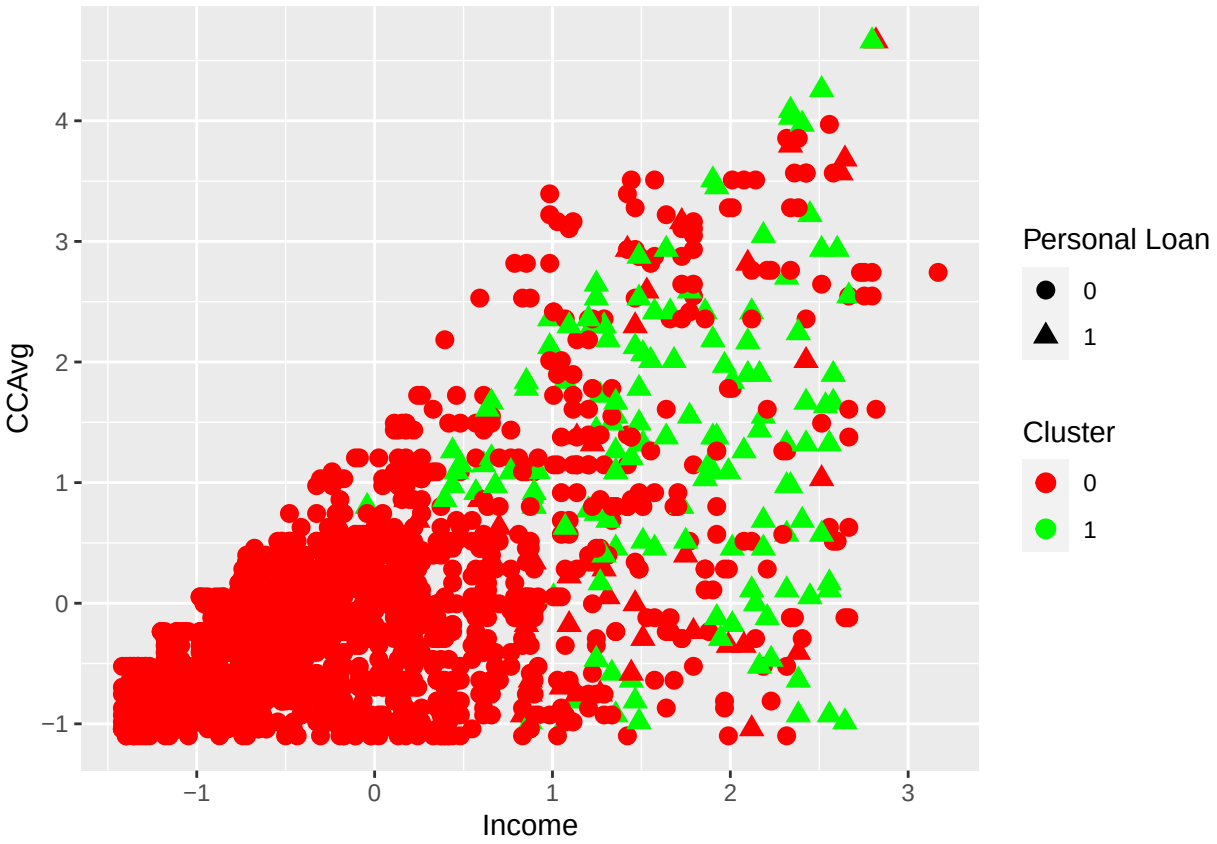



```r
# Visualize the clusters:
k <- 3
model_knn <- knn(train = train_norm[, c(1:4, 7, 8)],
                 test = valid_norm[, c(1:4, 7, 8)], cl = train_norm[, 7], k = k)

valid_clusters <- data.frame(valid_norm[, c(1:4, 7, 8)],
                             Cluster = factor(model_knn),
                             Personal_Loan = valid_norm[, 7]) # Include the Personal Loan variable


cluster_plot <- ggplot(valid_clusters, aes(x = Income, y = CCAvg)) +
  geom_point(aes(color = Cluster, shape = Personal_Loan), size = 3) +
  labs(x = "Income", y = "CCAvg", color = "Cluster", shape = "Personal Loan") +
  scale_shape_manual(values = c(16, 17)) + # Customize the shape values
  scale_color_manual(values = c("#FF0000", "#00FF00")) # Customize the color values

print(cluster_plot)
```



Cluster Description:

With k =3 for KNN, 2 clusters are formed. Cluster 0: Lower Income, Lower CCAvg, and ultimately most likely to NOT accept Personal Loan offer Cluster 1: Higher Income, Higher CCAVg, More likely to accept Personal Loan.

PART 4:BUSINESS INSIGHTS

Based on the work done above and bearing our problem statement in mind there are some key business insights we have identified.

1: Identification of Key Factors

Through the analysis, the key factors that significantly influence customers' acceptance of personal loans are identified. This insight can help businesses understand the main drivers behind customers' decision-making process and focus their efforts on those factors to improve loan acceptance rates. Key Factors Identified:Income level, Credit History, employment history, Age and experience.

2: Customer Segmentation

Customer Segmentation: The analysis can uncover different segments of customers based on their likelihood of accepting personal loans. By understanding these segments, businesses can tailor their marketing and communication strategies to target each segment effectively. This segmentation can be based on various characteristics such as demographics,income levels, employment status, or credit history.

3: Predictive Model

Developing a predictive model allows businesses to forecast the likelihood of a customer acceptinga personal loan. Best Model we suggest is the logistic regression model which best suits this particular data set as well as our problem statement enables proactive targeting of potential loan customers, leading to more effective

marketing campaigns and resource allocation.By accurately identifying potential loan customers, businesses can streamline their efforts and improve conversion rates.

4: Risk Assessment

The analysis can also provide insights into the risk associated with granting personal loans to different customers. By analyzing historical data and customer attributes, businesses can assess the creditworthiness of potential borrowers and make informed decisions to mitigate risks. This insight helps optimize the loan approval process and minimize default rates.This is something to consider for future use of this analysis as it currently predicts loan acceptance and not loan approval.

5: Product Development and Offerings

Understanding the factors that influence customers' acceptance of personal loans can guide product development efforts.By identifying customer preferences, pain points, and specific needs, businesses can tailor their loan products and offerings to better align with customer expectations.This insight can help differentiate their loan products from competitors and enhance customer satisfaction.

6: Marketing Strategy Optimization

Insights from the analysis can contribute to optimizing marketing strategies for personal loans. By identifying influential factors, businesses can prioritize marketing channels, messages, and promotional activities that resonate with potential loan customers.This insight helps allocate marketing budgets more effectively, resulting in higher customer engagement and conversion rates.