project

Atharva Janardan Rajadhyaksha | Srushti Sanjay Kharat | Niharika Dhapola

2023-05-10

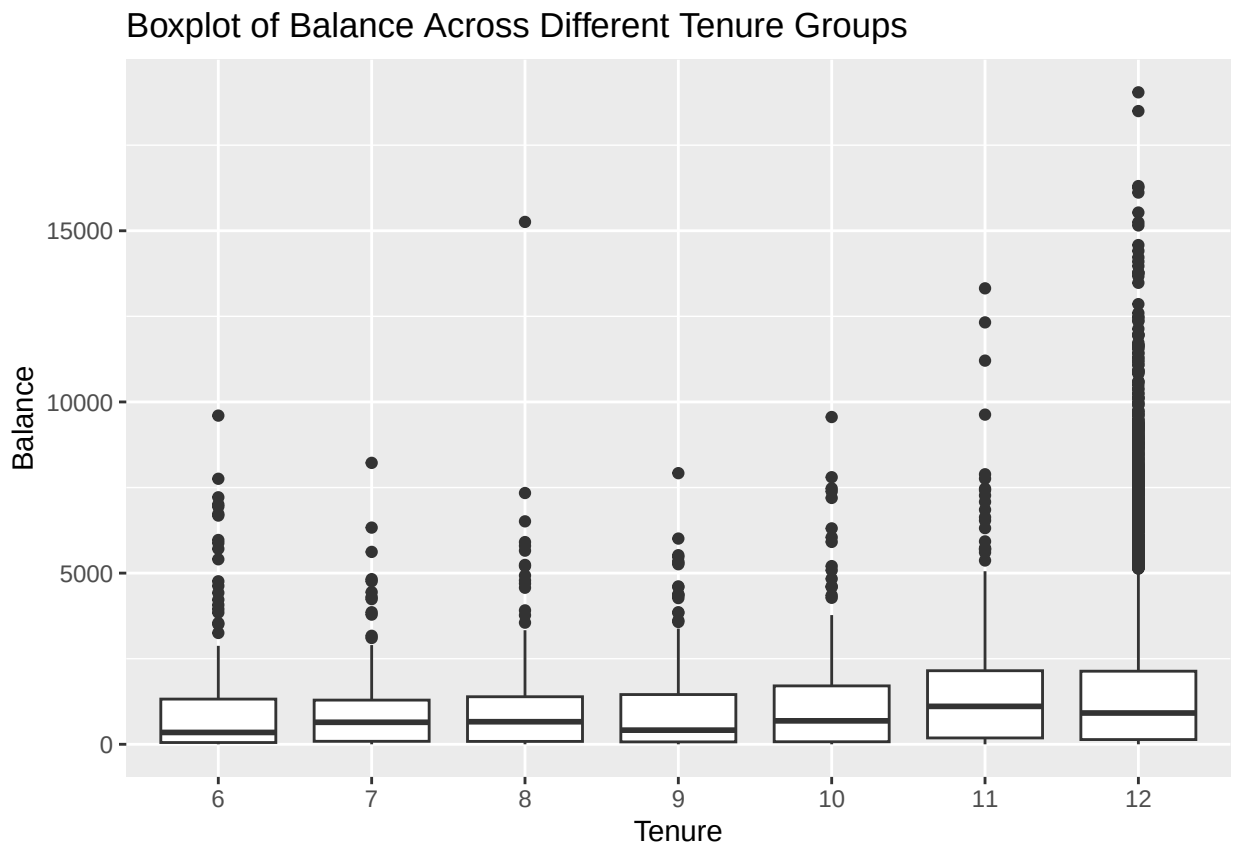
```
suppressPackageStartupMessages(library(tidyverse))
library(ggplot2)
library(tidyverse)
library(cluster)
suppressPackageStartupMessages(library(factoextra))
```

```
#Reading data from a csv file.
cc_data <- read.csv("CC GENERAL.csv")
```

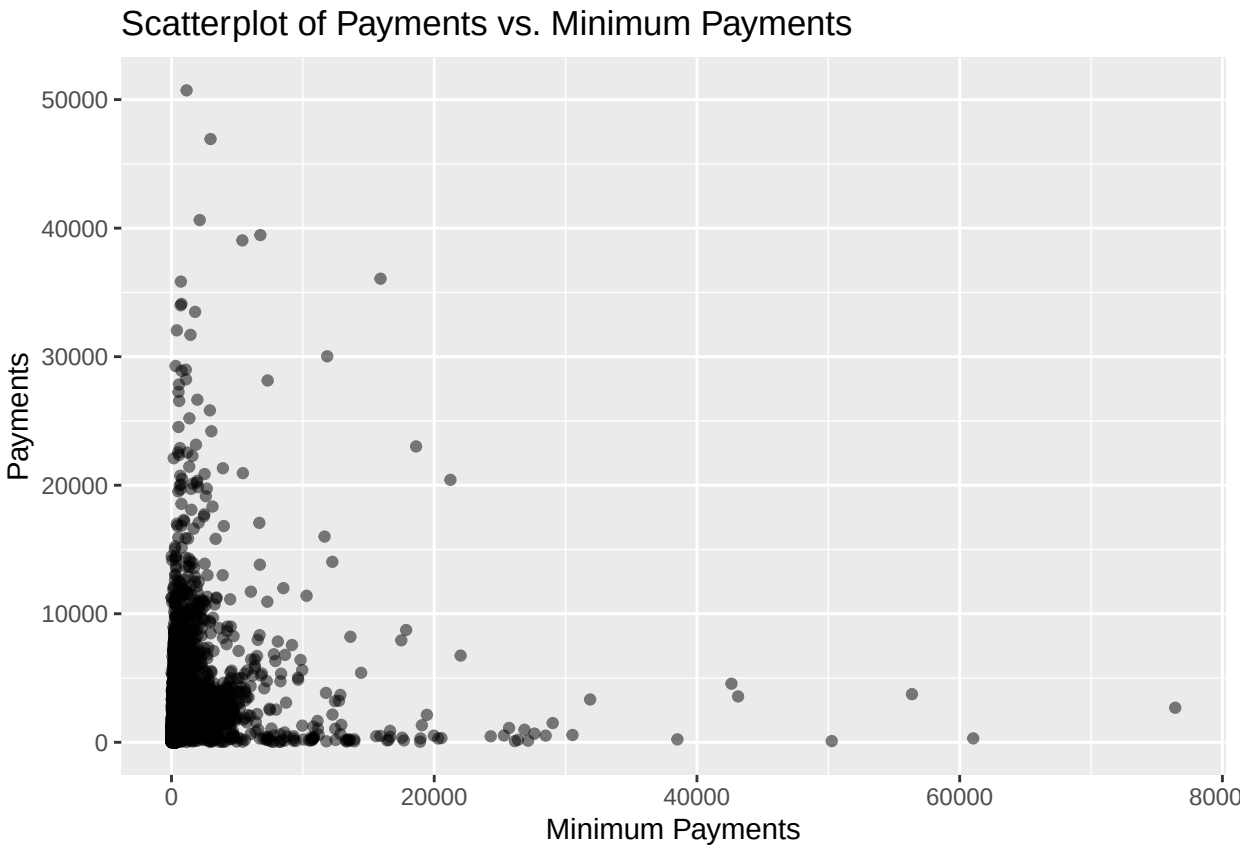
```
#Imputing the data. We have replaced NA values with the median of that particular column
cc_imputed <- cc_data %>%
  select(-CUST_ID) %>%
  mutate(across(everything(), ~ifelse(is.na(.), median(., na.rm = TRUE), .)))
```

```
#Scaling the data to have a mean of 0 and sd of 1
cc_scaled <- cc_imputed %>%
  mutate(across(everything(), scale))
```

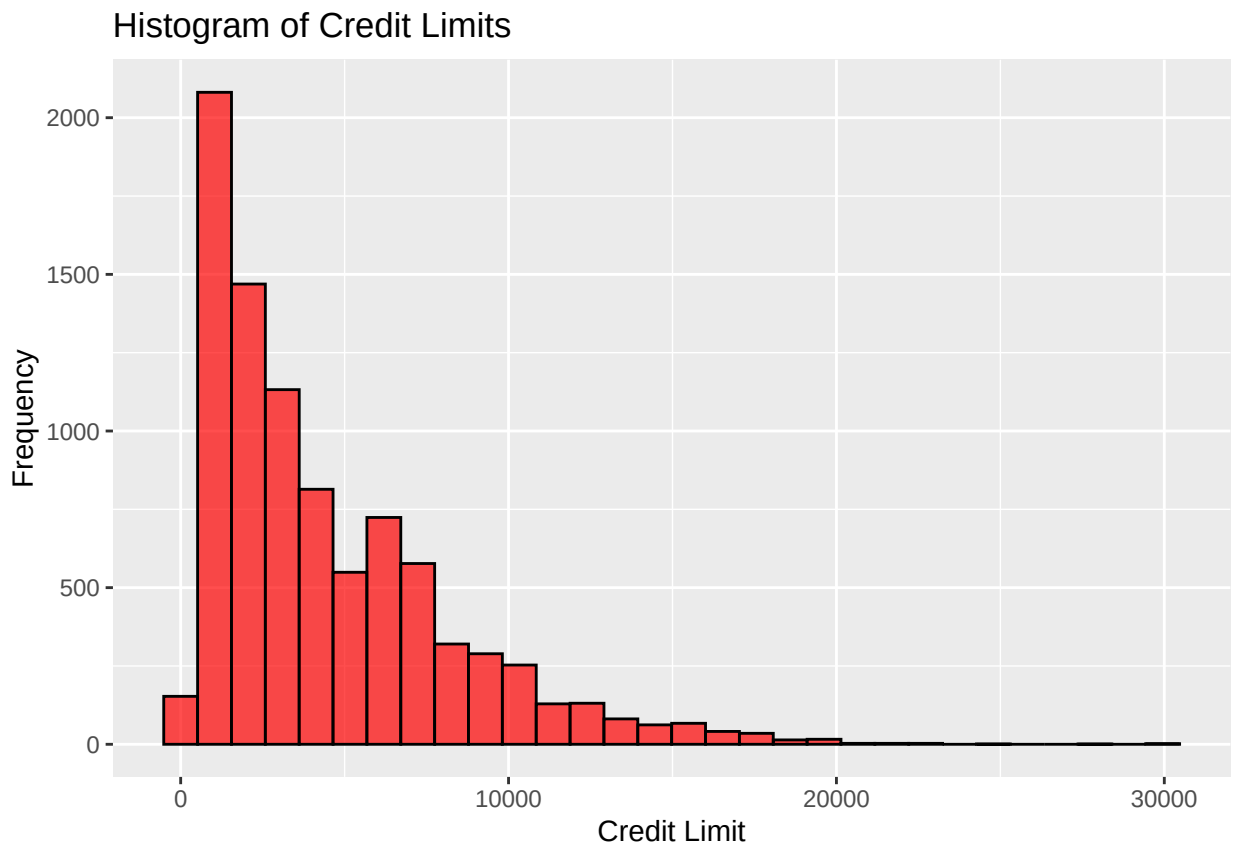
```
#Visualizing boxplots
ggplot(cc_imputed, aes(x = factor(TENURE), y = BALANCE)) +
  geom_boxplot() +
  ggtitle("Boxplot of Balance Across Different Tenure Groups") +
  xlab("Tenure") +
  ylab("Balance")
```



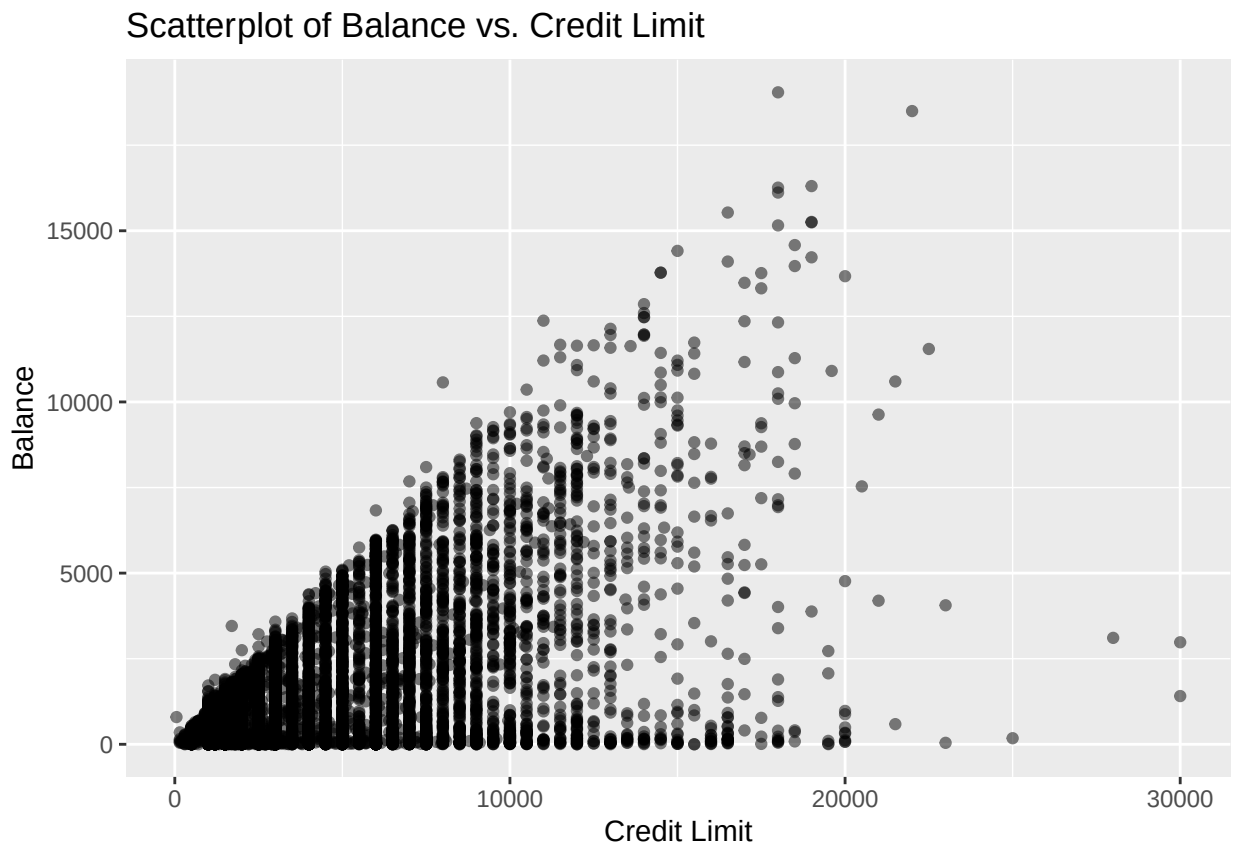

```
ggplot(cc_imputed, aes(x = MINIMUM_PAYMENTS, y = PAYMENTS)) +
  geom_point(alpha = 0.5) +
  ggtitle("Scatterplot of Payments vs. Minimum Payments") +
  xlab("Minimum Payments") +
  ylab("Payments")
```




```
ggplot(cc_imputed, aes(x = CREDIT_LIMIT)) +
  geom_histogram(bins = 30, fill = "red", color = "black", alpha = 0.7) +
  ggtitle("Histogram of Credit Limits") +
  xlab("Credit Limit") +
  ylab("Frequency")
```

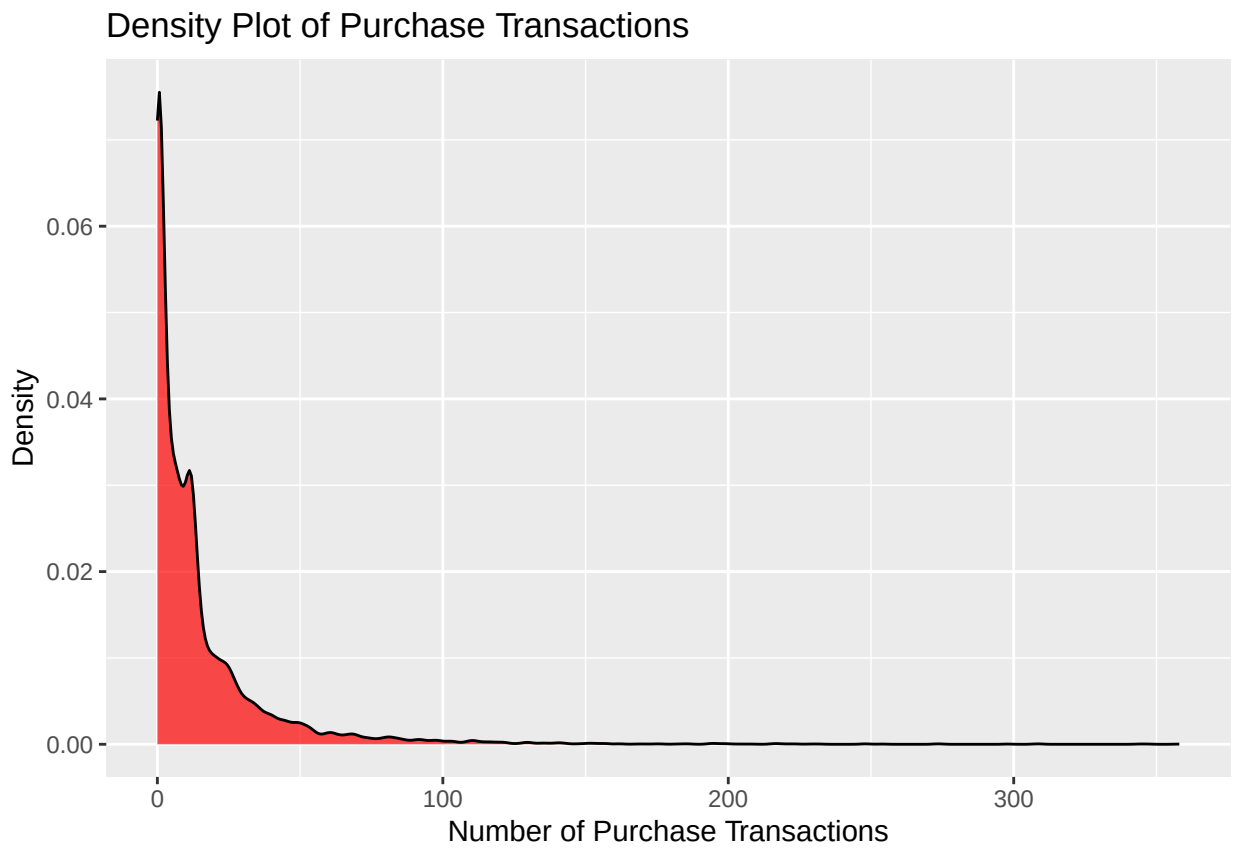



```
ggplot(cc_imputed, aes(x = CREDIT_LIMIT, y = BALANCE)) +
  geom_point(alpha = 0.5) +
  ggtitle("Scatterplot of Balance vs. Credit Limit") +
  xlab("Credit Limit") +
  ylab("Balance")
```

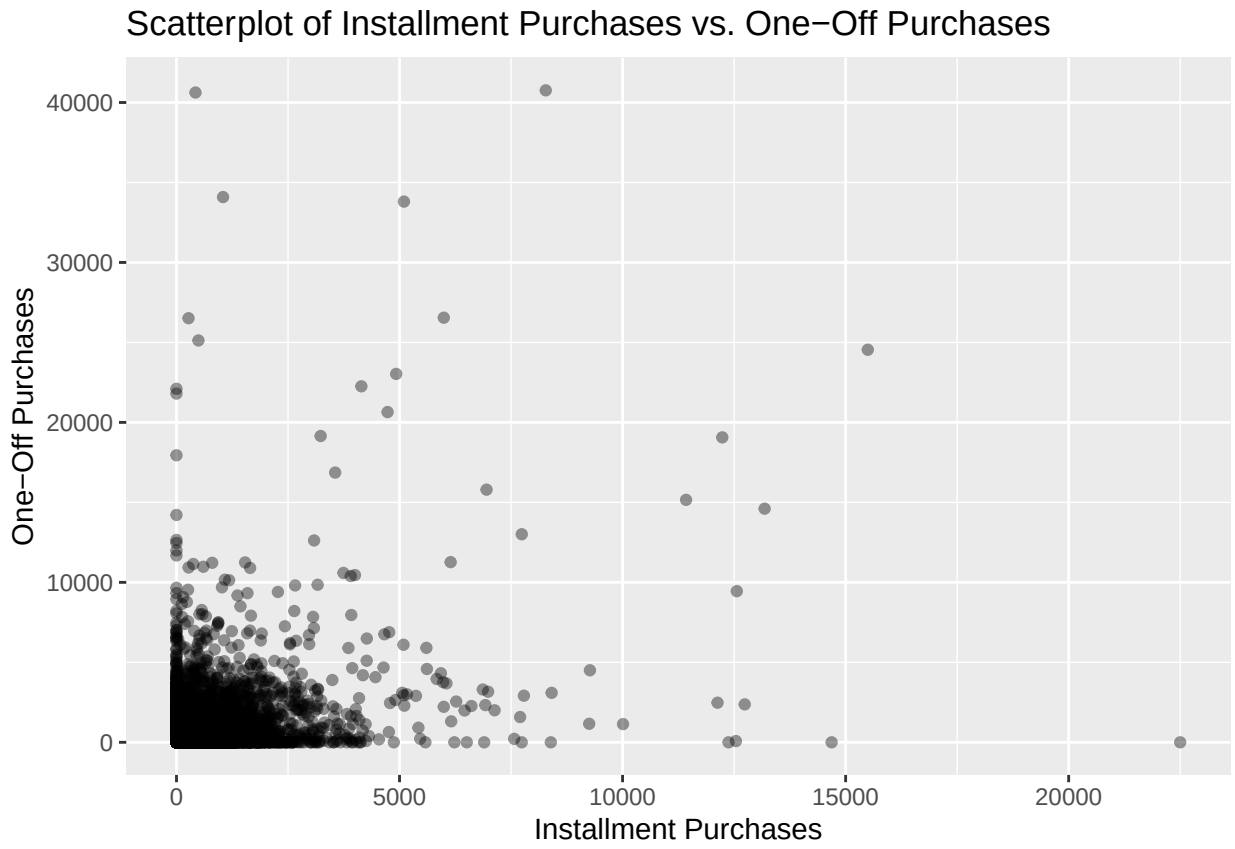



```
ggplot(cc_imputed, aes(x = PURCHASES_TRX, y = ..density..)) +
  geom_density(fill = "red", color = "black", alpha = 0.7) +
  ggtitle("Density Plot of Purchase Transactions") +
  xlab("Number of Purchase Transactions") +
  ylab("Density")
```

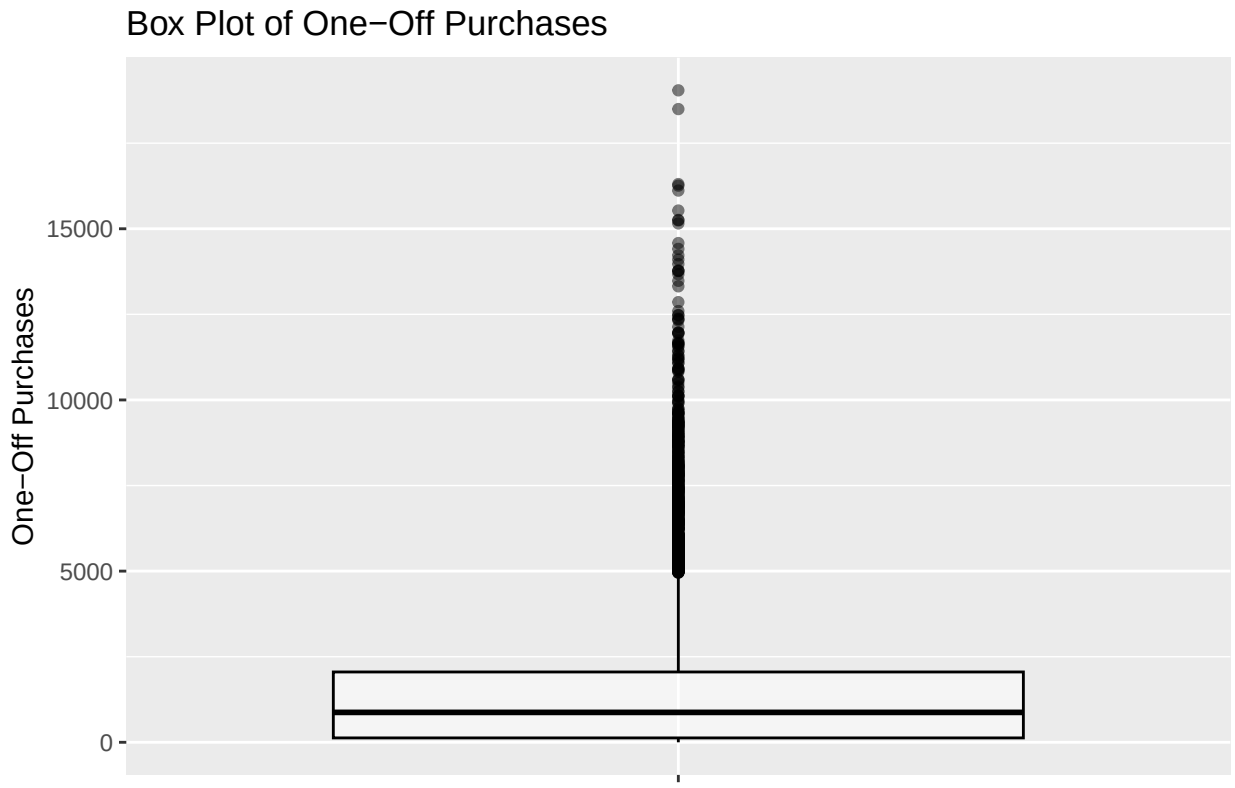
```
## Warning: The dot-dot notation (‘..density..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(density)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```




```
ggplot(cc_imputed, aes(x = INSTALLMENTS_PURCHASES, y = ONEOFF_PURCHASES)) +
  geom_point(alpha = 0.4) +
  ggtitle("Scatterplot of Installment Purchases vs. One-Off Purchases") +
  xlab("Installment Purchases") +
  ylab("One-Off Purchases")
```




```
ggplot(cc_imputed, aes(x = "", y = BALANCE)) +
  geom_boxplot(color = "black", alpha = 0.5) +
  ggtitle("Box Plot of One-Off Purchases") +
  xlab("") +
  ylab("One-Off Purchases")
```

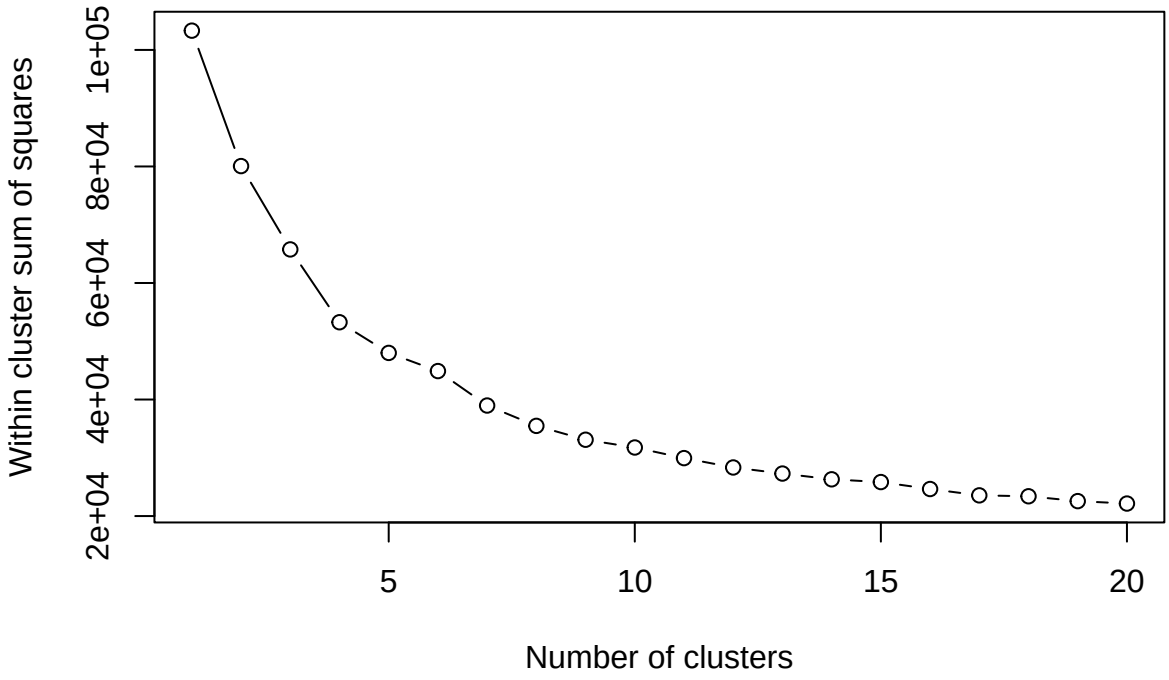



```
data <- select(cc_data, -c(CUST_ID))
data <- na.omit(data)
```

```
#Scaling the data
scaled_data <- scale(data)
# We will be extracting 5 principal components from the data
n_comp <- 5
#Performing PCA and storing the results in a dataframe
pca_result <- prcomp(scaled_data, center = TRUE, scale = TRUE)
pca_data <- as.data.frame(pca_result$x[, 1:n_comp])
head(pca_data)
```

```
##          PC1        PC2         PC3         PC4         PC5
## 1 -1.6962971 -1.1225190  0.49153311  0.71947913  0.07982586
## 2 -1.2156104  2.4354968  0.69461763 -0.09883702  0.80297229
## 3  0.9357991 -0.3851793 -0.02595178  1.29376862 -1.98717027
## 5 -1.6145448 -0.7245442  0.27234236  1.08605360 -0.42778877
## 6  0.2236877 -0.7835645 -1.18436576  0.72131105  0.80119651
## 7  6.2652350 -0.6094139  2.08544348 -0.57775168 -0.96556154
```

```
#Performing WSS on the data and plotting the results.
wss <- numeric(20)
for (i in 1:20) {
  wss[i] <- sum(kmeans(pca_data, centers = i)$withinss)
}
plot(1:20, wss, type = "b", xlab = "Number of clusters", ylab = "Within cluster sum of squares")
```



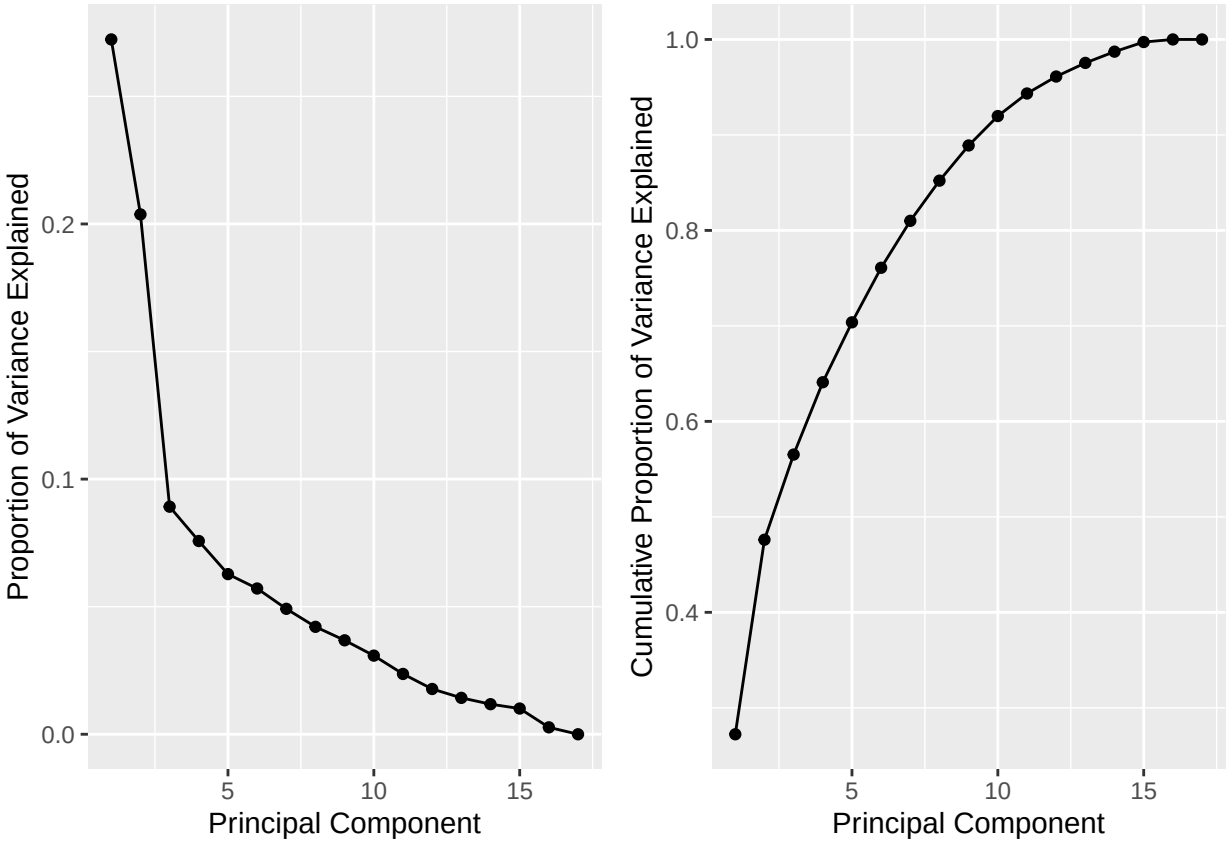

```
#Creating a tibble of the information about the PCA results found earlier.
#PVE is the proportion of variance explained by each principal component.
PVE <- tibble(
  PC=1:length(pca_result$sdev),
  Var=pca_result$sdev^2,
  PVE=Var/sum(Var),
  CumPVE=cumsum(PVE)
)
PVE
```

```
## # A tibble: 17 x 4
##       PC       Var         PVE CumPVE
##    <int>     <dbl>       <dbl>  <dbl>
##  1     1 4.63      0.272        0.272
##  2     2 3.46      0.204        0.476
##  3     3 1.52      0.0892       0.565
##  4     4 1.29      0.0757       0.641
##  5     5 1.07      0.0628       0.704
##  6     6 0.971     0.0571       0.761
##  7     7 0.836     0.0492       0.810
##  8     8 0.716     0.0421       0.852
##  9     9 0.626     0.0368       0.889
## 10    10 0.524     0.0308       0.920
## 11    11 0.402     0.0236       0.943
```

```
## 12    12 0.302     0.0177       0.961
## 13    13 0.243     0.0143       0.975
## 14    14 0.200     0.0118       0.987
## 15    15 0.171     0.0101       0.997
## 16    16 0.0461    0.00271      1.00
## 17    17 0.0000117 0.000000690  1
```

```
#1/number of var
# Plotting a comparative grid of PVE vs Cumulative PVE using qplot
cowplot::plot_grid(
qplot(data=PVE,x=PC,y=PVE,geom=c("point","line"),
      xlab = "Principal Component",
      ylab = "Proportion of Variance Explained"),
qplot(data=PVE,x=PC,y=CumPVE,geom=c("point","line"),
      xlab = "Principal Component",
      ylab = "Cumulative Proportion of Variance Explained")
)
```

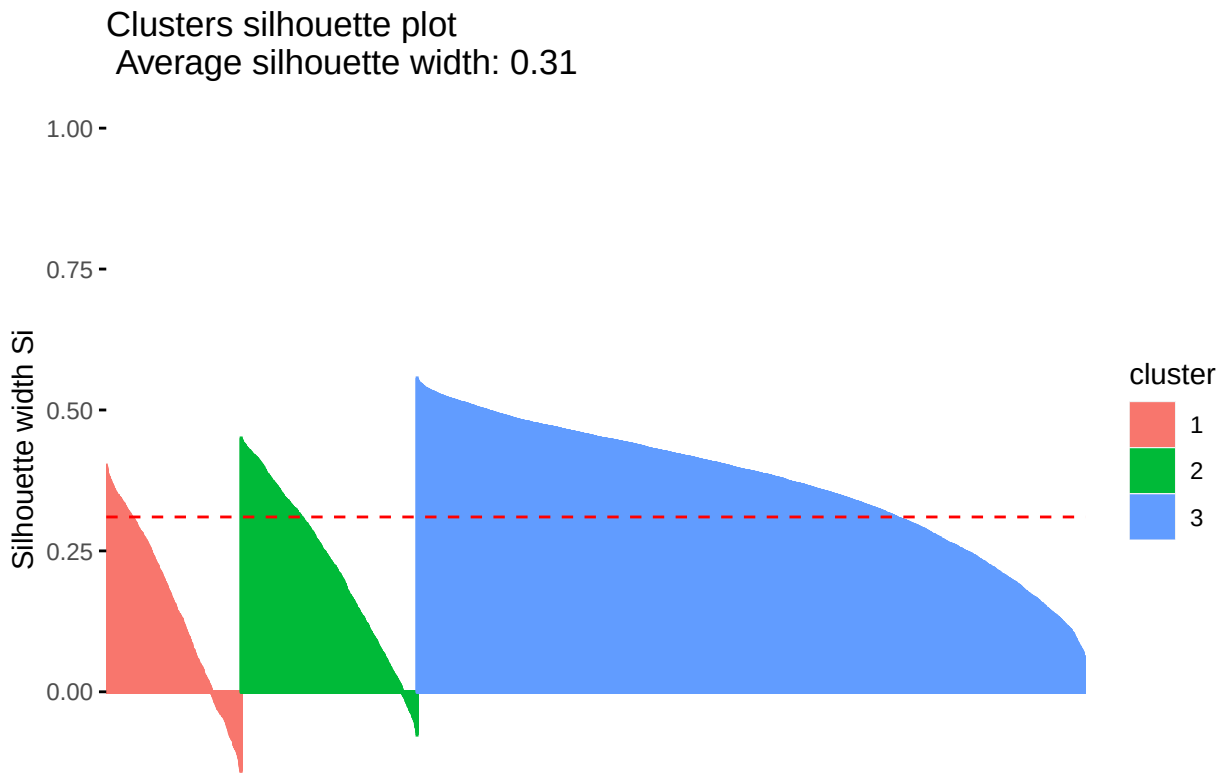
```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



```r
#Performing k-means clustering on the principal components obtained earlier.
# We have set the number of clusters "k" to 3
k <- 3
kmeans_result <- kmeans(pca_data, centers = k)
clusters <- kmeans_result$cluster
#Calculating the quality of clustering using silhouette
sil <- silhouette(clusters, dist(pca_data))
```

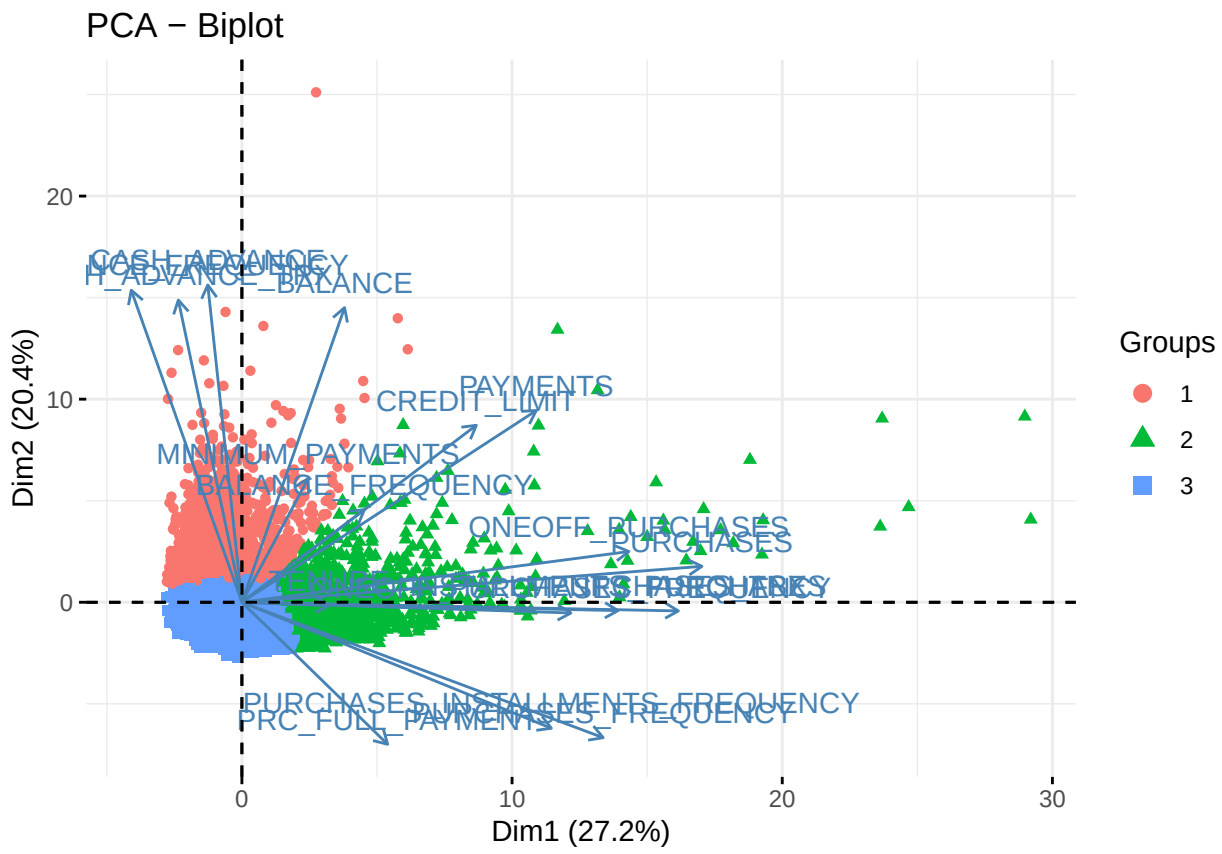
```r
#Visualizing the silhouette widths in a plot.
fviz_silhouette(sil)
```

```
##   cluster size ave.sil.width
## 1       1 1189          0.15
## 2       2 1552          0.22
## 3       3 5895          0.37
```




```r
km_out <- kmeans(pca_data[, 1:2], centers = k)

# Plot the biplot (PC1 vs PC2) with colored clusters
fviz_pca_biplot(pca_result, axes = c(1, 2), geom = "point", habillage = km_out$cluster,
                ggtheme = theme_minimal())
```


PCA − Biplot
Dim2 (20.4%)
Dim1 (27.2%)
20
10
0
0
10
20
30
CASH_ADVANCE
H_ADVANCE_BALANCE
PAYMENTS
CREDIT_LIMIT
MINIMUM_PAYMENTS
BALANCE_FREQUENCY
ONEOFF_PURCHASES
_PURCHASES
PURCHASES_INSTALLMENTS_FREQUENCY
PRC_FULL_PAYMENT
Groups
1
2
3