```r
#Libraries
library(tidyverse)
```

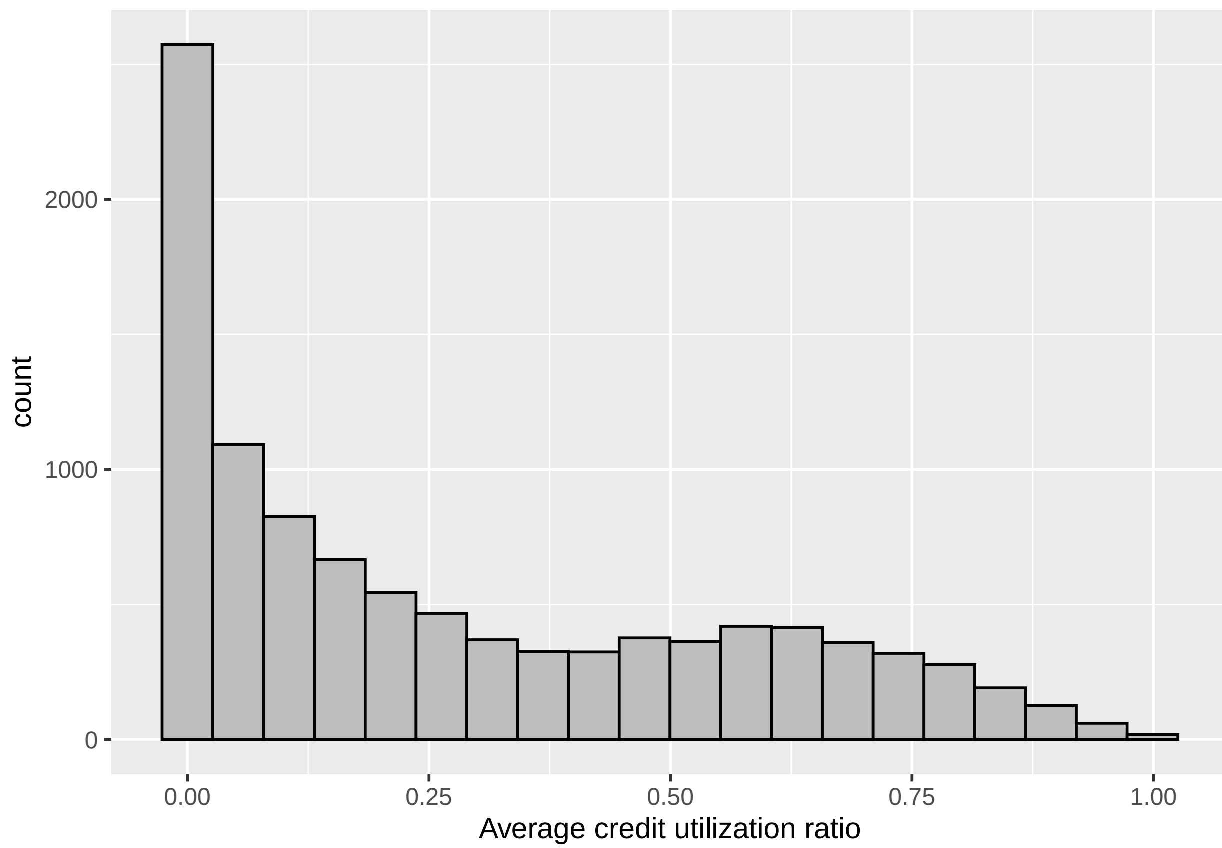
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ 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 error
```

```r
library(dplyr)
library(rpart)
library(partykit)
```

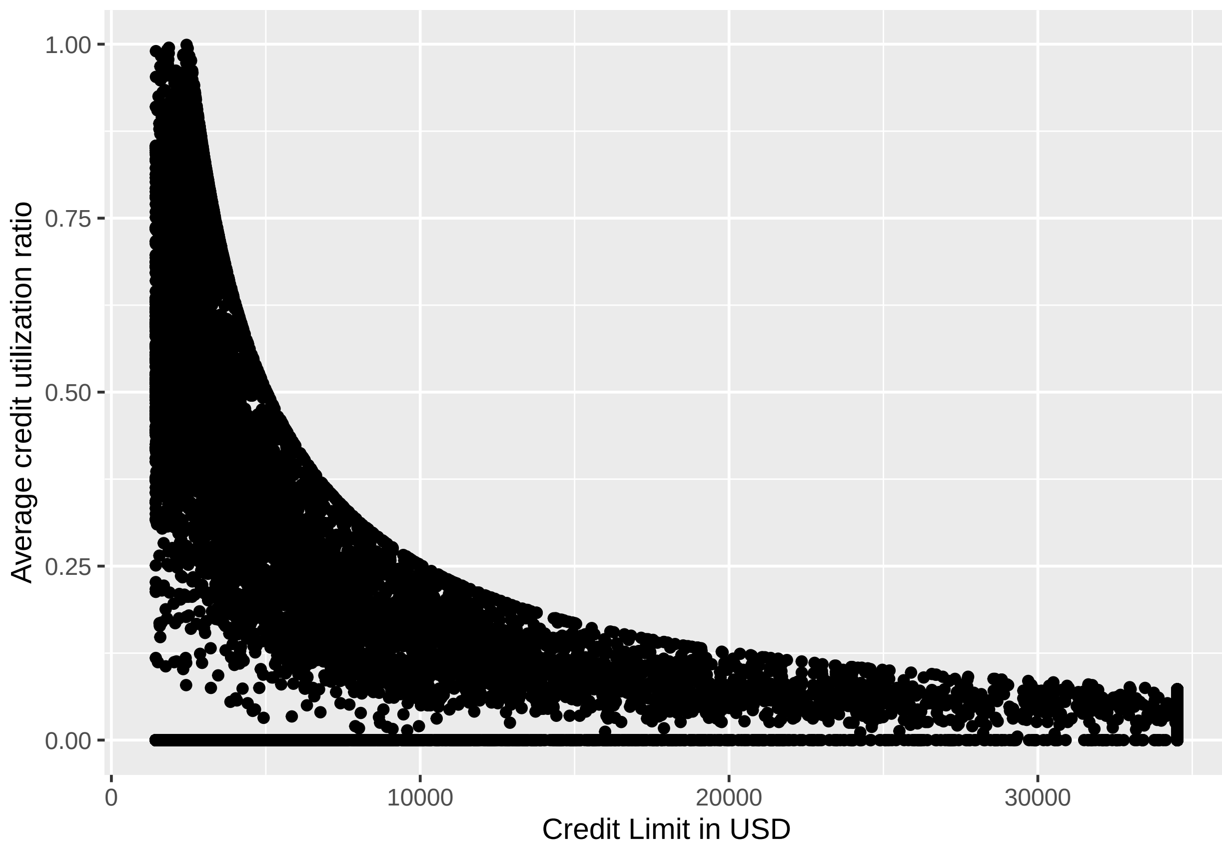
```
## Loading required package: grid
## Loading required package: libcoin
## Loading required package: mvtnorm
```

```r
#Cleaning data for use
credit_data <- read.csv("credit_card.csv")
customer_data <- read.csv("customer.csv")
total_raw_data <- left_join(customer_data, credit_data, by = 'client_num')

client_data <- total_raw_data %>% select(-state_cd, -zipcode) %>%
  mutate(ratio_usage = case_when(avg_utilization_ratio <= 0.3 ~ 'Low',
                                 avg_utilization_ratio > 0.3 ~ 'High'))
```

```r
#Visualizations

client_data %>% ggplot(aes(x = avg_utilization_ratio)) +
  geom_histogram(color = 'black', fill = 'grey', bins = 20) +
  labs(x = 'Average credit utilization ratio')
```



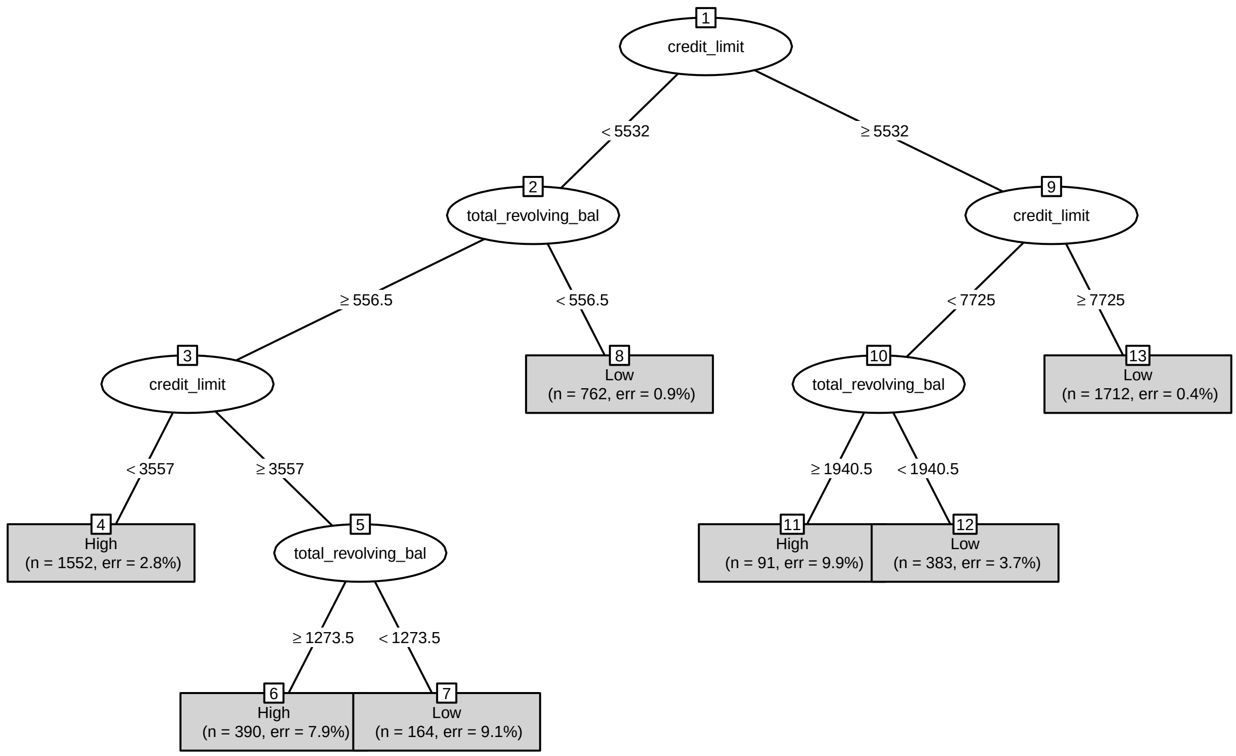

```r
client_data %>% ggplot(aes(x = credit_limit, y = avg_utilization_ratio)) +
  geom_point() +
  labs(x = "Credit Limit in USD", y = "Average credit utilization ratio")
```




```
# Decision tree model to predict delinquency

set.seed(172)

testing_data <- sample_n(client_data, size = round(0.5 * nrow(client_data)))
training_data <- anti_join(client_data, testing_data)
```

```
## Joining with ‘by = join_by(client_num, customer_age, gender, dependent_count,
## education_level, marital_status, car_owner, house_owner, personal_loan,
## contact, customer_job, income, cust_satisfaction_score, card_category,
## annual_fees, activation_30_days, customer_acq_cost, week_start_date, week_num,
## qtr, current_year, credit_limit, total_revolving_bal, total_trans_amt,
## total_trans_vol, avg_utilization_ratio, use_chip, exp_type, interest_earned,
## delinquent_acc, ratio_usage)‘
```

```
tree <- rpart(ratio_usage ~ total_revolving_bal + credit_limit + income +
                interest_earned + personal_loan, data = training_data)
plot(as.party(tree), gp = gpar(cex = 0.5), type = 'simple')
```




```r
#Use tree for predictions

predictions <- predict(tree, newdata = testing_data, type = 'class')

confusion_matrix <- table(predictions, testing_data$ratio_usage)

confusion_matrix
```

```
##
## predictions High  Low
##        High 1828  114
##        Low    46 3066
```

```r
#Compute relevant rates

Accuracy <- (1828 + 3066) / (1828 + 114 + 46 + 3066)

Sensitivity <- 1828 / (1828 + 46)
Specificity <- 3066 / (3066 + 114)
FP_rate <- 114 / (114 + 3066)
FN_rate <- 46 / (1828 + 46)
```