EDA for OZNAL project

Filip Remšík, Filip Mojto

2025-04-30

Contents

Company Bankruptcy Prediction - EDA

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 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- 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(conflicted)
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.4.3
```

```
## corrplot 0.95 loaded
```

```
library(purrr)
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.4.3
```

```
## Loading required package: lattice
```

```
library(nnet)
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.4.3
```

```
## Type 'citation("pROC")' for a citation.
```

```
conflict_prefer("filter", "dplyr")
```

```
## [conflicted] Will prefer dplyr::filter over any other package.
```

```
conflict_prefer("lag", "dplyr")
```

```
## [conflicted] Will prefer dplyr::lag over any other package.
```

```
conflict_prefer("select", "dplyr")
```

```
## [conflicted] Will prefer dplyr::select over any other package.
```

Dataset

Load

```
paste("Loading dataset from '", getwd(), "'")
```

```
## [1] "Loading dataset from ' C:/Users/fmojt/Documents/DataAnalysisProjects/OZNAL-project/documentatio
```

```
df <- read_csv("../data/raw/data.csv", show_col_types = FALSE)
```

View

```
View(df)
```

Feature selection

Based on the hypothesis we formulated previously we will select the following features as our potential predictors.

```
TARGET = 'Bankrupt?'
# indices of predictors
PREDICTORS = c(2, 3, 4, 5, 6, 7, 8, 9, 10, 24, 25, 26, 27, 28, 29, 30, 31, 32)
PREDICTOR_NAMES <- names(df)[PREDICTORS]
cat("Predictors: \n")
```

```
## Predictors:
```

```
walk(PREDICTOR_NAMES, print)
```

```
## [1] "ROA(C) before interest and depreciation before interest"
## [1] "ROA(A) before interest and % after tax"
## [1] "ROA(B) before interest and depreciation after tax"
## [1] "Operating Gross Margin"
## [1] "Realized Sales Gross Margin"
## [1] "Operating Profit Rate"
## [1] "Pre-tax net Interest Rate"
## [1] "After-tax net Interest Rate"
## [1] "Non-industry income and expenditure/revenue"
```

```
## [1] "Per Share Net profit before tax (Yuan ¥)"
## [1] "Realized Sales Gross Profit Growth Rate"
## [1] "Operating Profit Growth Rate"
## [1] "After-tax Net Profit Growth Rate"
## [1] "Regular Net Profit Growth Rate"
## [1] "Continuous Net Profit Growth Rate"
## [1] "Total Asset Growth Rate"
## [1] "Net Value Growth Rate"
## [1] "Total Asset Return Growth Rate Ratio"
```

```
selected_data <- df %>%
  select(all_of(PREDICTORS), all_of(TARGET))

selected_data
```

```
## # A tibble: 6,819 x 19
##    ROA(C) before interest and de~1 ROA(A) before intere~2 ROA(B) before intere~3
##                              <dbl>                  <dbl>                  <dbl>
##  1                           0.371                  0.424                  0.406
##  2                           0.464                  0.538                  0.517
##  3                           0.426                  0.499                  0.472
##  4                           0.400                  0.451                  0.458
##  5                           0.465                  0.538                  0.522
##  6                           0.389                  0.415                  0.419
##  7                           0.391                  0.446                  0.436
##  8                           0.508                  0.571                  0.559
##  9                           0.489                  0.545                  0.543
## 10                           0.496                  0.551                  0.543
## # i 6,809 more rows
## # i abbreviated names:
## #   1: `ROA(C) before interest and depreciation before interest`,
## #   2: `ROA(A) before interest and % after tax`,
## #   3: `ROA(B) before interest and depreciation after tax`
## # i 16 more variables: `Operating Gross Margin` <dbl>,
## #   `Realized Sales Gross Margin` <dbl>, `Operating Profit Rate` <dbl>, ...
```

Feature transformation

In this section we will perform some additional preprocessing with data.

This blocks ensures that target is treated as factor so that it is later processed with some plots properly.

```
# target is now treated as factor (required in boxplots for proper binning)
selected_data[[TARGET]] <- as.factor(selected_data[[TARGET]])
```

Basic statistics & feature scale

After selecting predictors, we can now examine the dataset a bit more in detail. We will use function *summary()* which displays for each column the following:

1) **Minimum**
2) **1st quartile**
3) **Median**
4) **Mean**

5) **3rd quatle**

6) **Maximum**

From minimum and maximum value we can also deduce the **scale** of every column.

```
summary(selected_data)
```

```
##  ROA(C) before interest and depreciation before interest
##  Min.   :0.0000
##  1st Qu.:0.4765
##  Median :0.5027
##  Mean   :0.5052
##  3rd Qu.:0.5356
##  Max.   :1.0000
##  ROA(A) before interest and % after tax
##  Min.   :0.0000
##  1st Qu.:0.5355
##  Median :0.5598
##  Mean   :0.5586
##  3rd Qu.:0.5892
##  Max.   :1.0000
##  ROA(B) before interest and depreciation after tax Operating Gross Margin
##  Min.   :0.0000                                    Min.   :0.0000
##  1st Qu.:0.5273                                    1st Qu.:0.6004
##  Median :0.5523                                    Median :0.6060
##  Mean   :0.5536                                    Mean   :0.6079
##  3rd Qu.:0.5841                                    3rd Qu.:0.6139
##  Max.   :1.0000                                    Max.   :1.0000
##  Realized Sales Gross Margin Operating Profit Rate Pre-tax net Interest Rate
##  Min.   :0.0000              Min.   :0.0000        Min.   :0.0000
##  1st Qu.:0.6004              1st Qu.:0.9990        1st Qu.:0.7974
##  Median :0.6060              Median :0.9990        Median :0.7975
##  Mean   :0.6079              Mean   :0.9988        Mean   :0.7972
##  3rd Qu.:0.6138              3rd Qu.:0.9991        3rd Qu.:0.7976
##  Max.   :1.0000              Max.   :1.0000        Max.   :1.0000
##  After-tax net Interest Rate Non-industry income and expenditure/revenue
##  Min.   :0.0000              Min.   :0.0000
##  1st Qu.:0.8093              1st Qu.:0.3035
##  Median :0.8094              Median :0.3035
##  Mean   :0.8091              Mean   :0.3036
##  3rd Qu.:0.8095              3rd Qu.:0.3036
##  Max.   :1.0000              Max.   :1.0000
##  Per Share Net profit before tax (Yuan ¥)
##  Min.   :0.0000
##  1st Qu.:0.1704
##  Median :0.1797
##  Mean   :0.1844
##  3rd Qu.:0.1935
##  Max.   :1.0000
##  Realized Sales Gross Profit Growth Rate Operating Profit Growth Rate
##  Min.   :0.00000                         Min.   :0.0000
##  1st Qu.:0.02206                         1st Qu.:0.8480
##  Median :0.02210                         Median :0.8480
##  Mean   :0.02241                         Mean   :0.8480
##  3rd Qu.:0.02215                         3rd Qu.:0.8481
```

```
##  Max.   :1.00000                          Max.   :1.0000
##  After-tax Net Profit Growth Rate Regular Net Profit Growth Rate
##  Min.   :0.0000                   Min.   :0.0000
##  1st Qu.:0.6893                   1st Qu.:0.6893
##  Median :0.6894                   Median :0.6894
##  Mean   :0.6891                   Mean   :0.6892
##  3rd Qu.:0.6896                   3rd Qu.:0.6896
##  Max.   :1.0000                   Max.   :1.0000
##  Continuous Net Profit Growth Rate Total Asset Growth Rate
##  Min.   :0.0000                    Min.   :0.000e+00
##  1st Qu.:0.2176                    1st Qu.:4.860e+09
##  Median :0.2176                    Median :6.400e+09
##  Mean   :0.2176                    Mean   :5.508e+09
##  3rd Qu.:0.2176                    3rd Qu.:7.390e+09
##  Max.   :1.0000                    Max.   :9.990e+09
##  Net Value Growth Rate Total Asset Return Growth Rate Ratio Bankrupt?
##  Min.   :0.000e+00     Min.   :0.0000                       0:6599
##  1st Qu.:0.000e+00     1st Qu.:0.2638                       1: 220
##  Median :0.000e+00     Median :0.2640
##  Mean   :1.566e+06     Mean   :0.2642
##  3rd Qu.:0.000e+00     3rd Qu.:0.2644
##  Max.   :9.330e+09     Max.   :1.0000
```

From the above values we can observe that almost all the features have the same scale ranging from 0 to 1. Only *Total_asset_growth_rate* and *Net Value Growth Rate* are exceptions where maximum reaches billions.

We can also notice the target where **distribution** for each class (0 - not bankrupt and 1 - bankrupt) is **heavily imbalanced**.

From the scale and quartiles we can also see some **skewness** in data. We will examine this more later in this document.

Missing values

Now we will detect any missing values in the data.

```r
nrows <- nrow(df)


get_nan_percentage <- function(filter = FALSE) {
  nans <- df %>%
    summarise(across(everything(), ~round(sum(is.na(.)) / nrow(df) * 100, 2))) %>%
    rename_with(~paste0(., "_%"))

  if (filter) {
    # Keep only columns where NA percentage > 0
    nans <- nans %>%
      select(where(~ . > 0))
  }

  return(nans)
}

nan_percentage <- get_nan_percentage()
nan_percentage
```

```
## # A tibble: 1 x 96
##   `Bankrupt?_%` ROA(C) before interest and depreciation~1 ROA(A) before intere~2
##           <dbl>                                     <dbl>                  <dbl>
## 1             0                                         0                      0
## # i abbreviated names:
## #   1: `ROA(C) before interest and depreciation before interest_%`,
## #   2: `ROA(A) before interest and % after tax_%`
## # i 93 more variables:
## #   `ROA(B) before interest and depreciation after tax_%` <dbl>,
## #   `Operating Gross Margin_%` <dbl>, `Realized Sales Gross Margin_%` <dbl>,
## #   `Operating Profit Rate_%` <dbl>, `Pre-tax net Interest Rate_%` <dbl>, ...
```

```r
nan_cols <- nan_percentage %>%
  select(where(~ . > 0))

cat("No. of cols with missing values:", ncol(nan_cols))
```

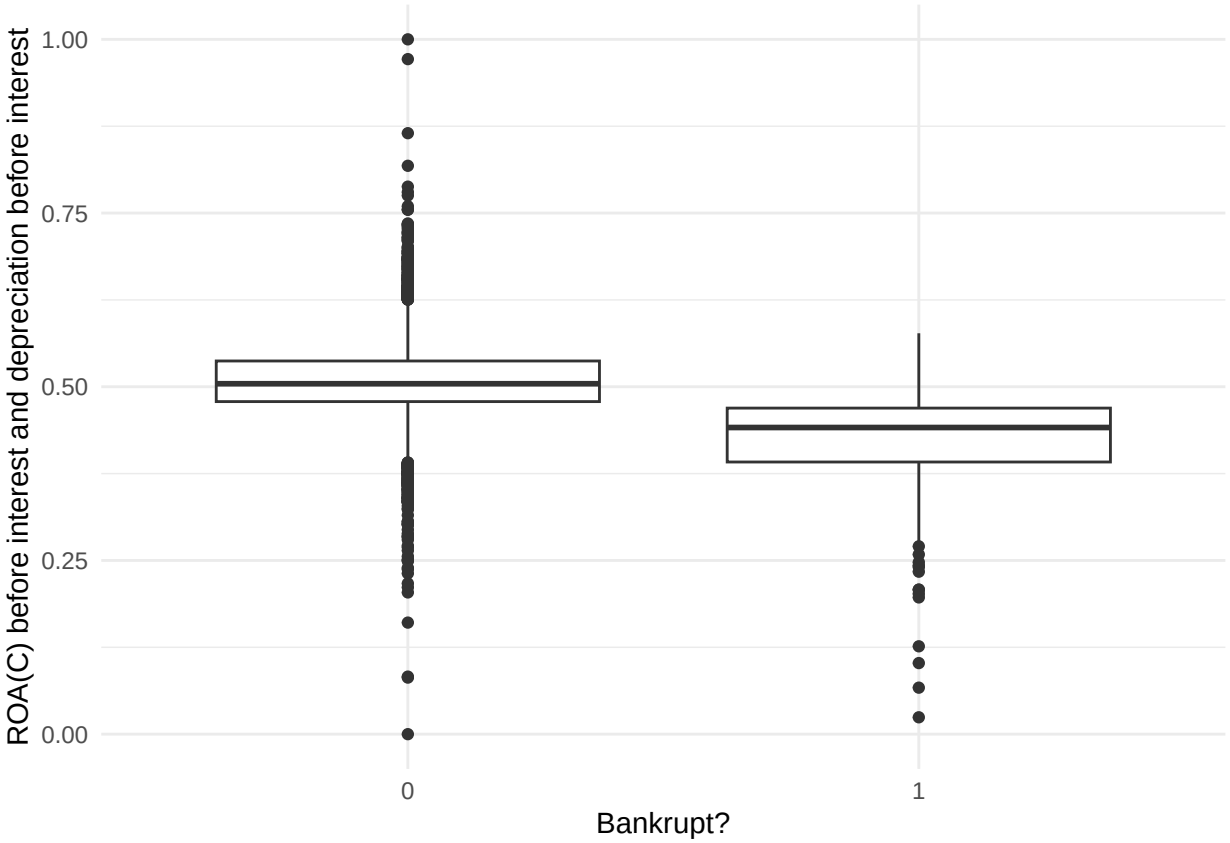
```
## No. of cols with missing values: 0
```

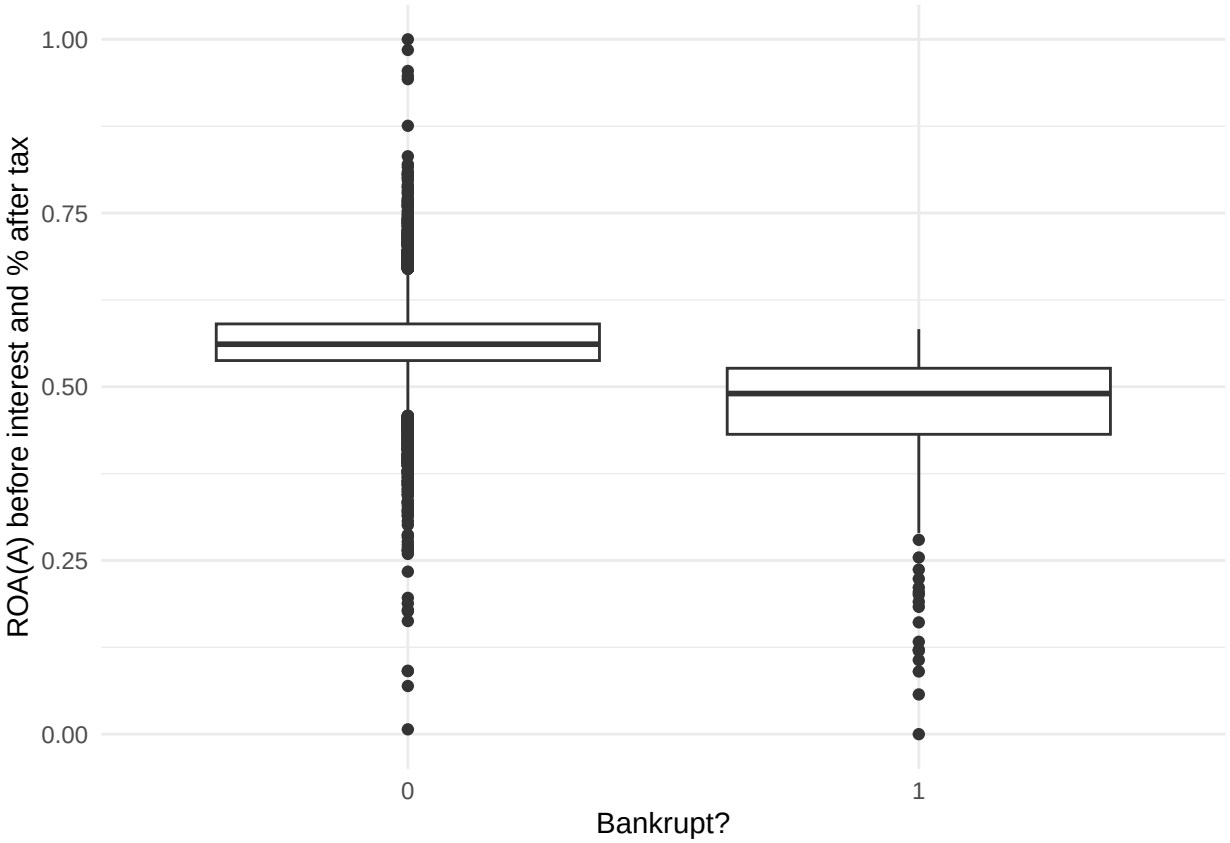
We have **detected no missing values**. Thus, no more action to be taken is required here.

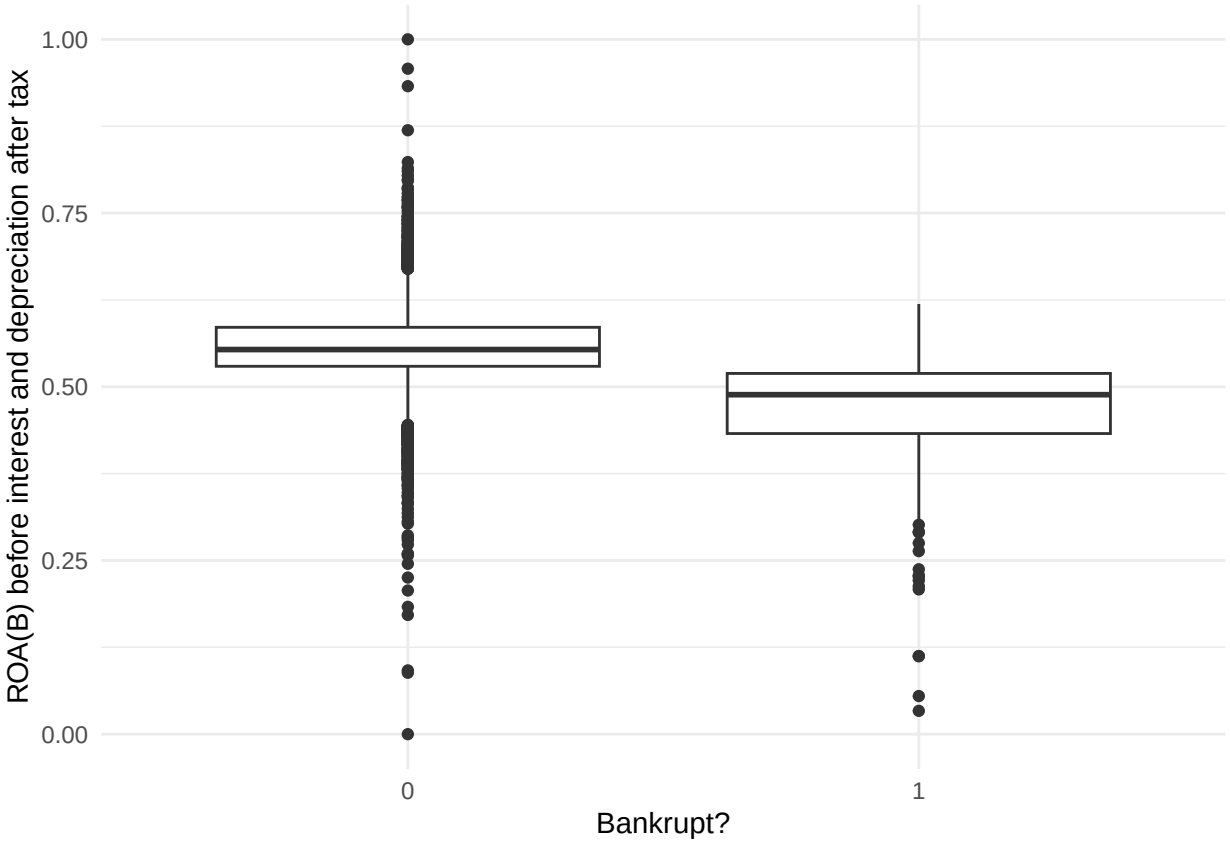
Outliers

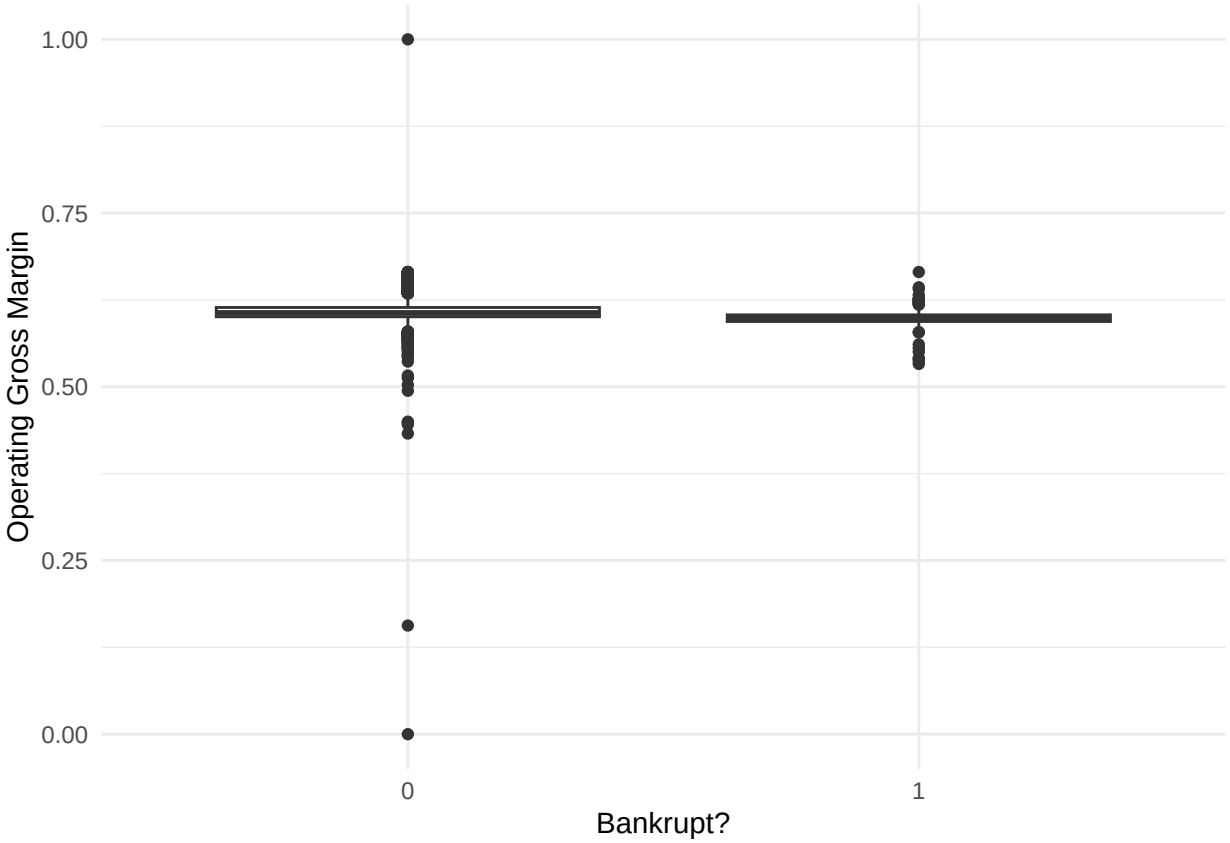
Now we will use plots to visualize outliers in our data. Boxplots are the ideal tool to properly distinguish outliers from the rest of the data.

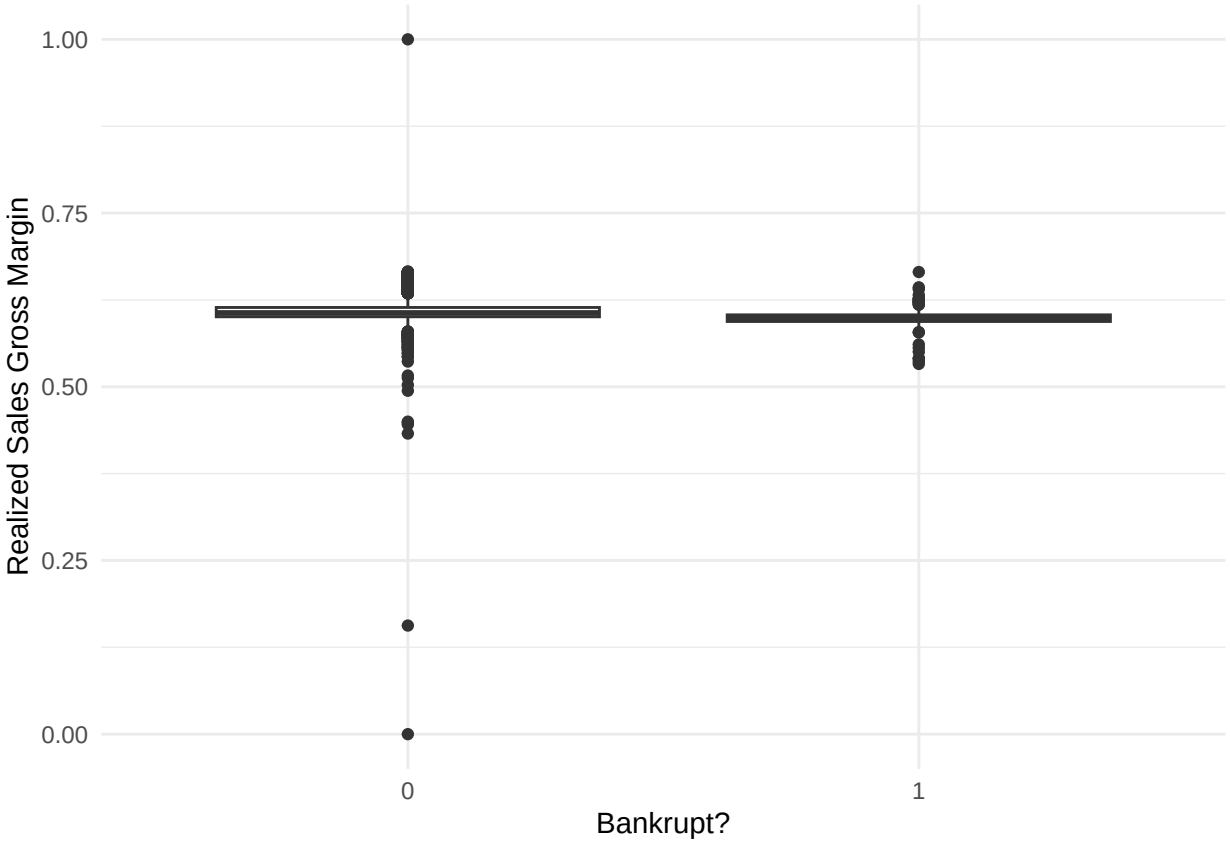
```r
show.boxplot <- function(data, cols, target){
  plots <- map(cols, ~ggplot(
    data,
    aes(x = .data[[target]], y = .data[[.x]])) +
      geom_boxplot(aes(group = .data[[target]])) +
      theme_minimal()
  )

  walk(plots, print)
}

selected_data[[TARGET]] <- as.factor(selected_data[[TARGET]])

show.boxplot(data = selected_data, cols = PREDICTOR_NAMES, target = TARGET)
```

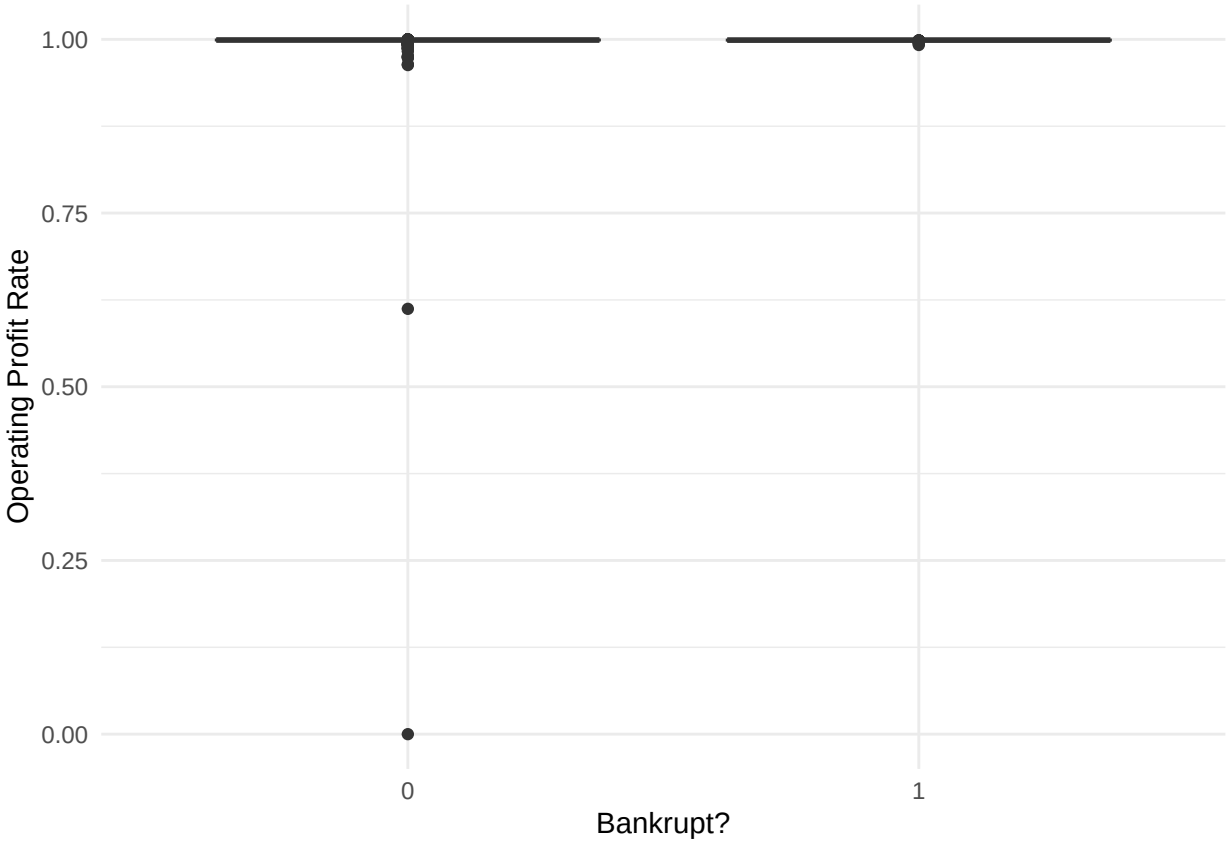

ROA(C) before interest and depreciation before interest
1.00
0.75
0.50
0.25
0.00
0
1
Bankrupt?

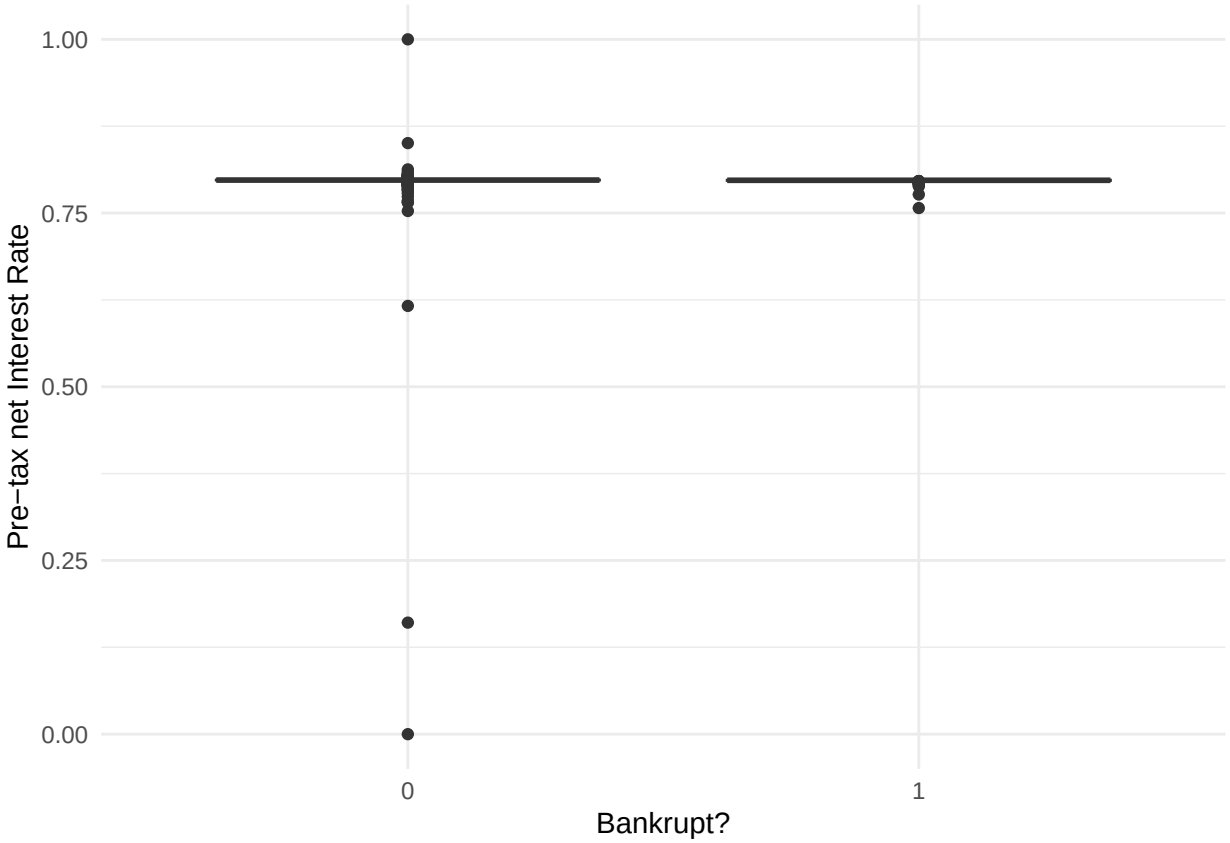

1.00
0.75
0.50
0.25
0.00
ROA(A) before interest and % after tax
0
1
Bankrupt?

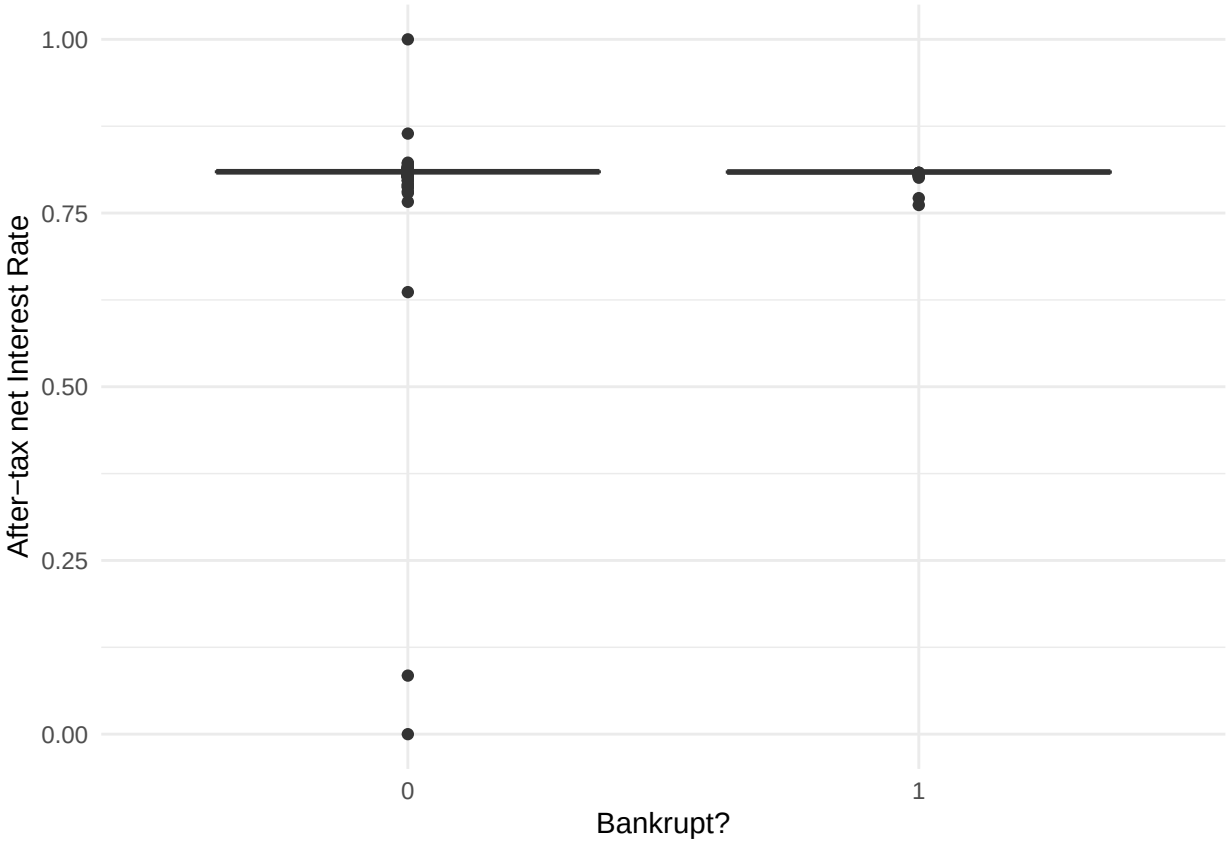

1.00
0.75
0.50
0.25
0.00
ROA(B) before interest and depreciation after tax
0
1
Bankrupt?

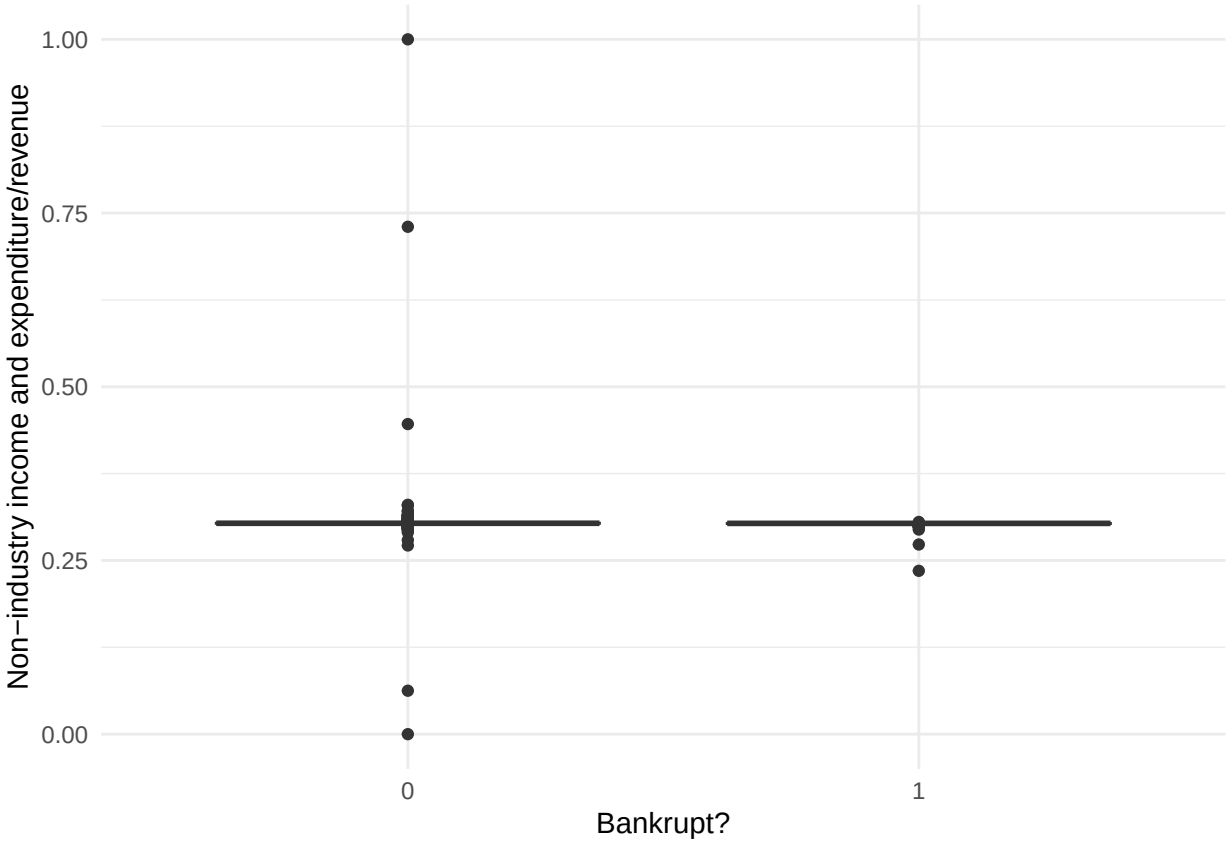

1.00
0.75
0.50
0.25
0.00
Operating Gross Margin
0
1
Bankrupt?

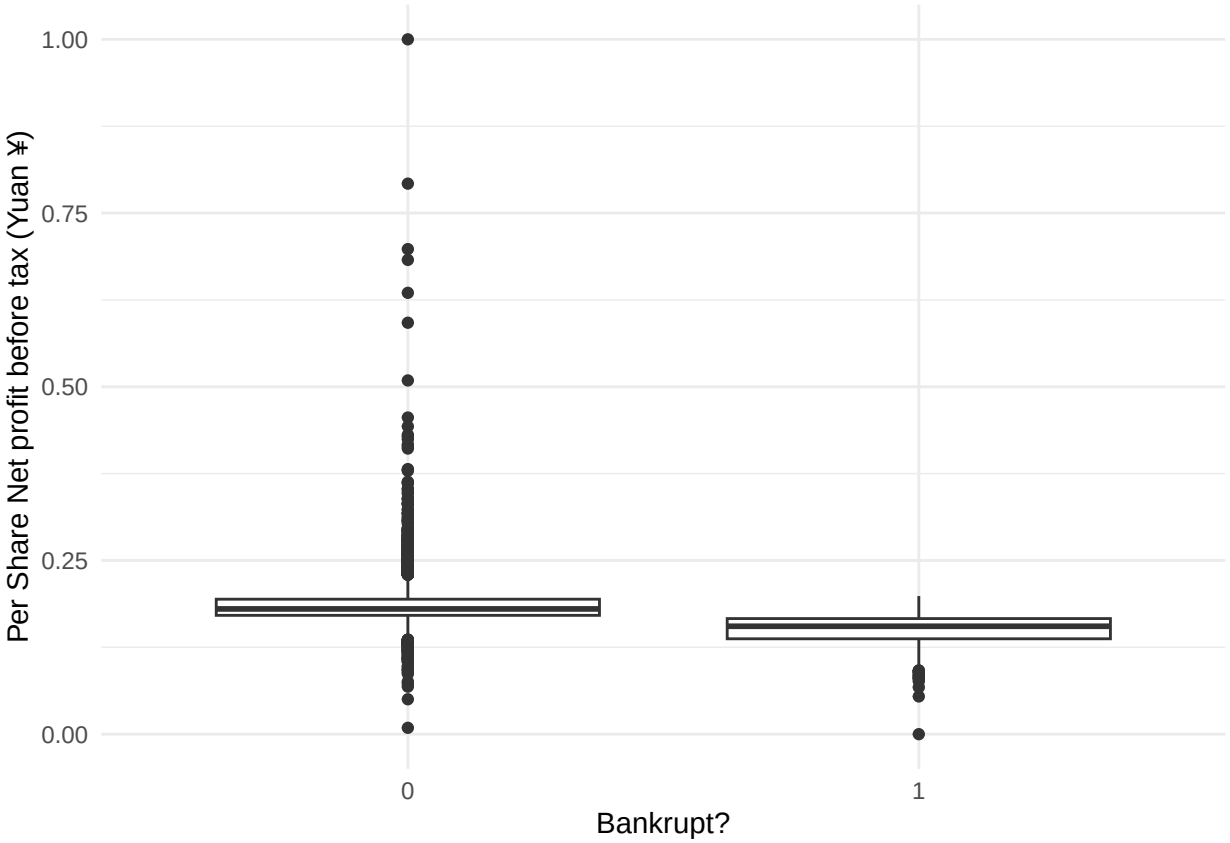

1.00
0.75
0.50
0.25
0.00
Realized Sales Gross Margin
0
1
Bankrupt?

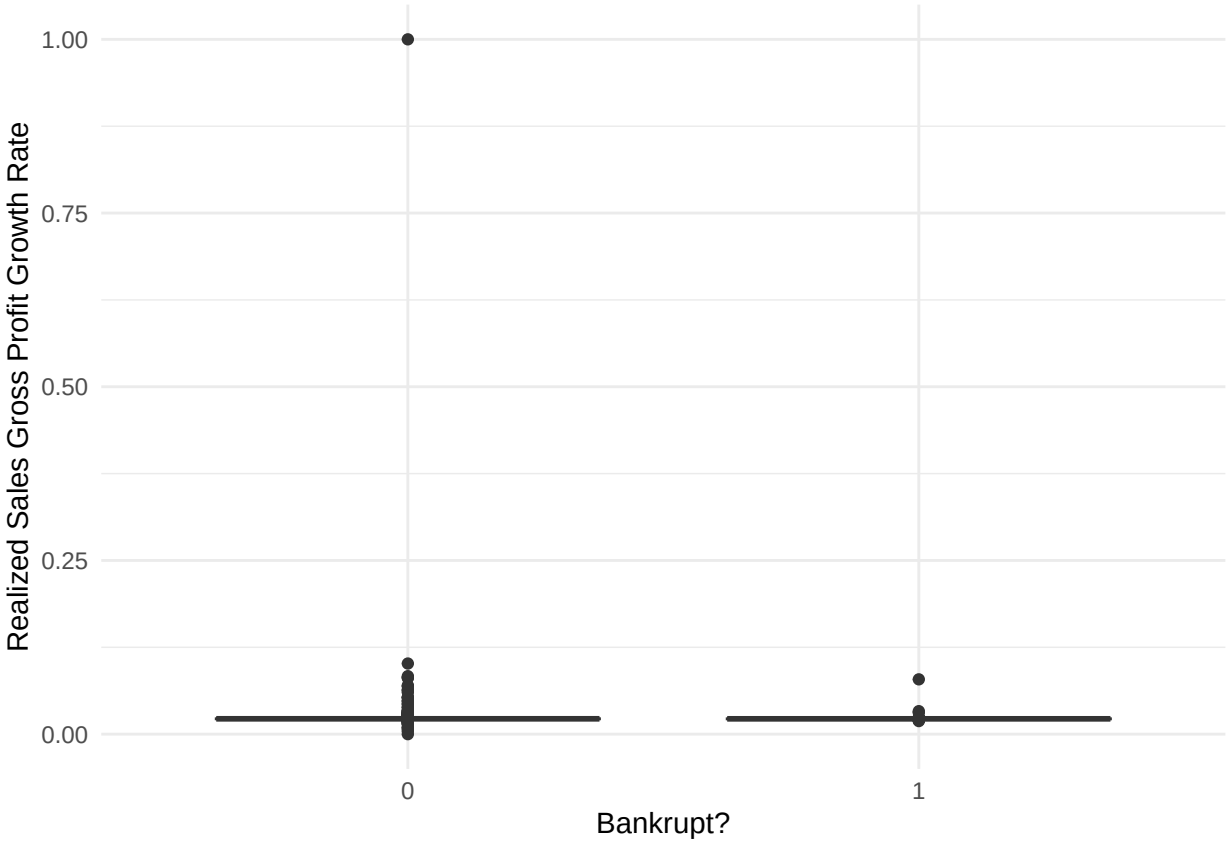

1.00
0.75
0.50
0.25
0.00
Operating Profit Rate
0
1
Bankrupt?

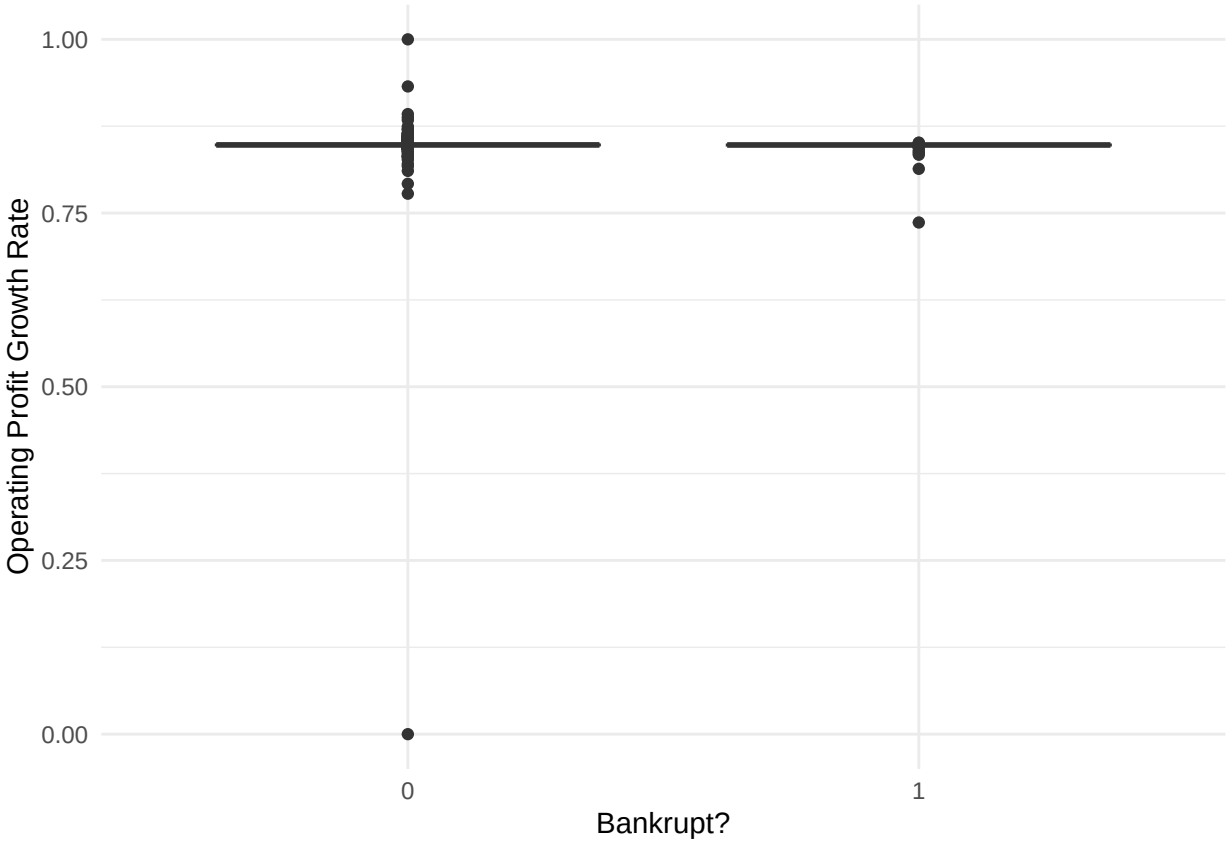

1.00
0.75
0.50
0.25
0.00
Pre−tax net Interest Rate
0
1
Bankrupt?

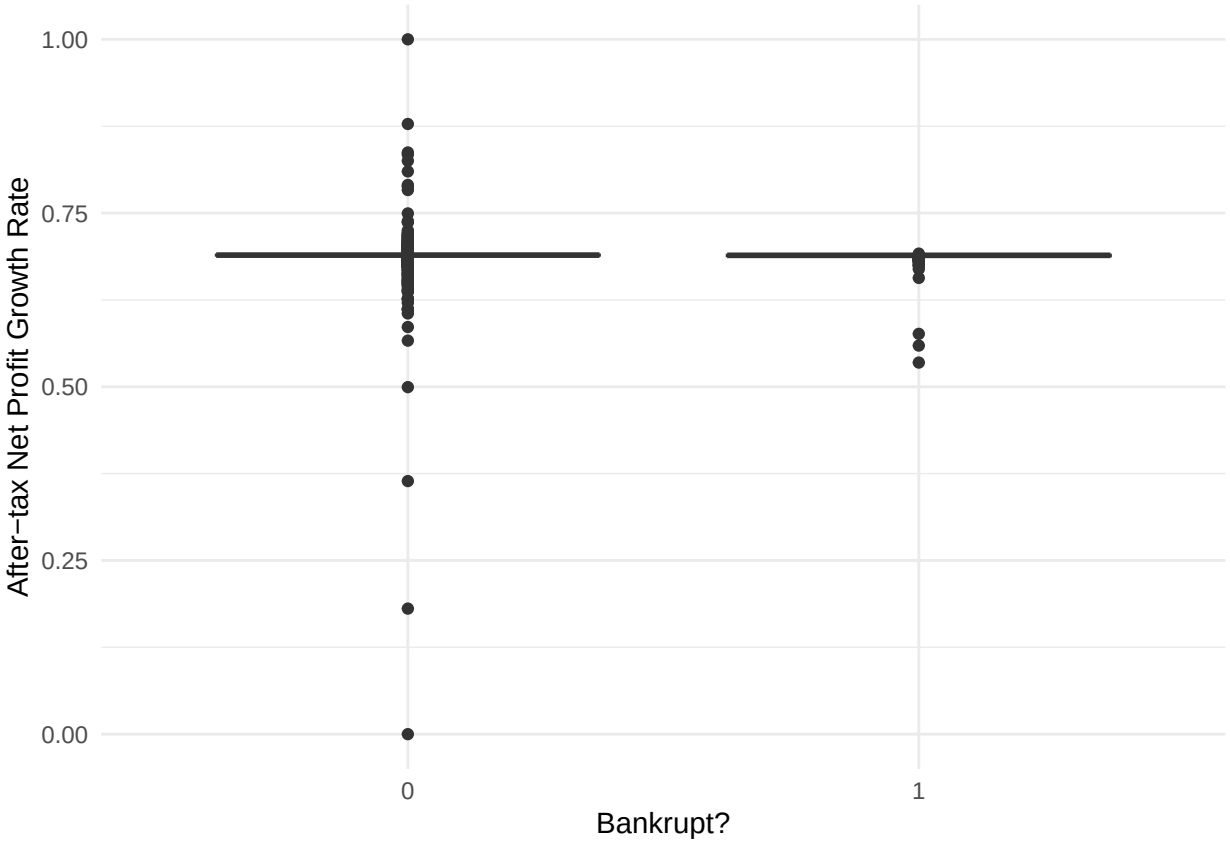

1.00
0.75
0.50
0.25
0.00
After−tax net Interest Rate
0
1
Bankrupt?

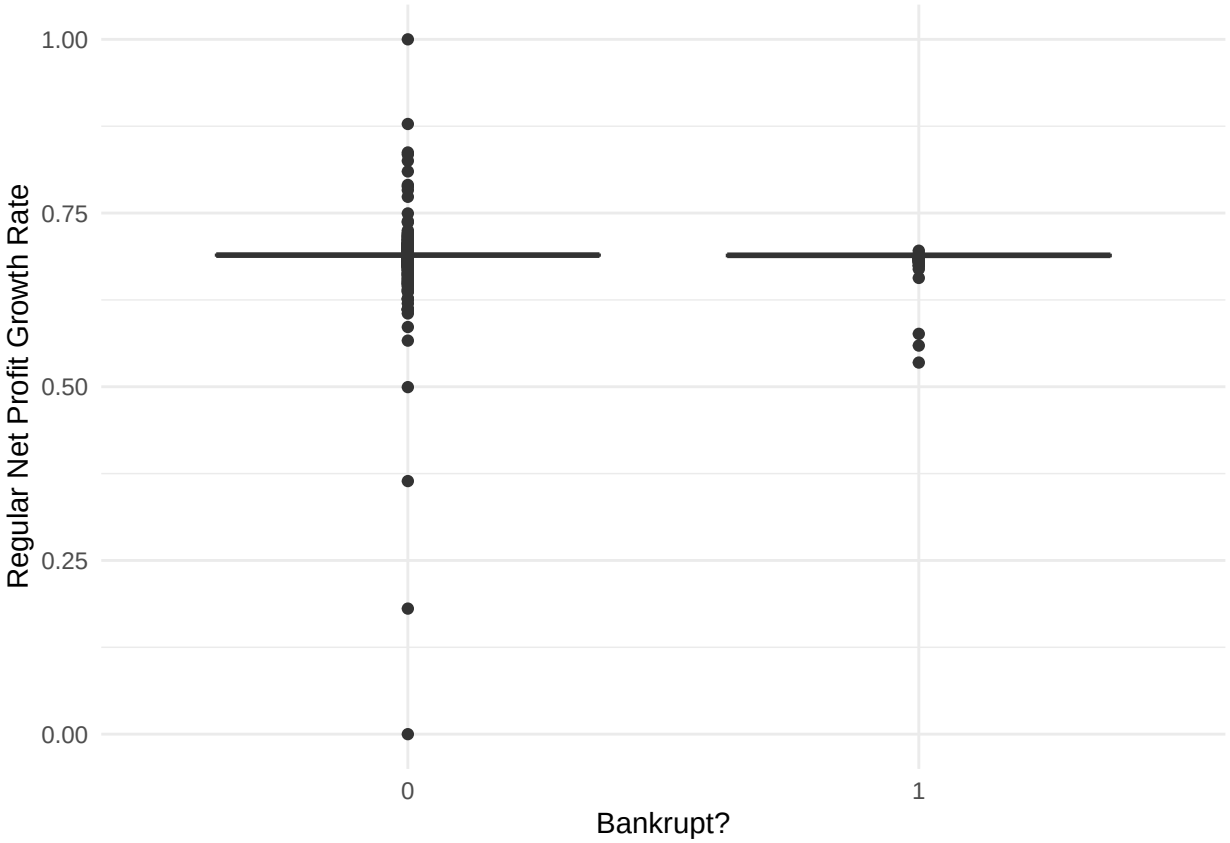

1.00
0.75
0.50
0.25
0.00
Non−industry income and expenditure/revenue
0
1
Bankrupt?

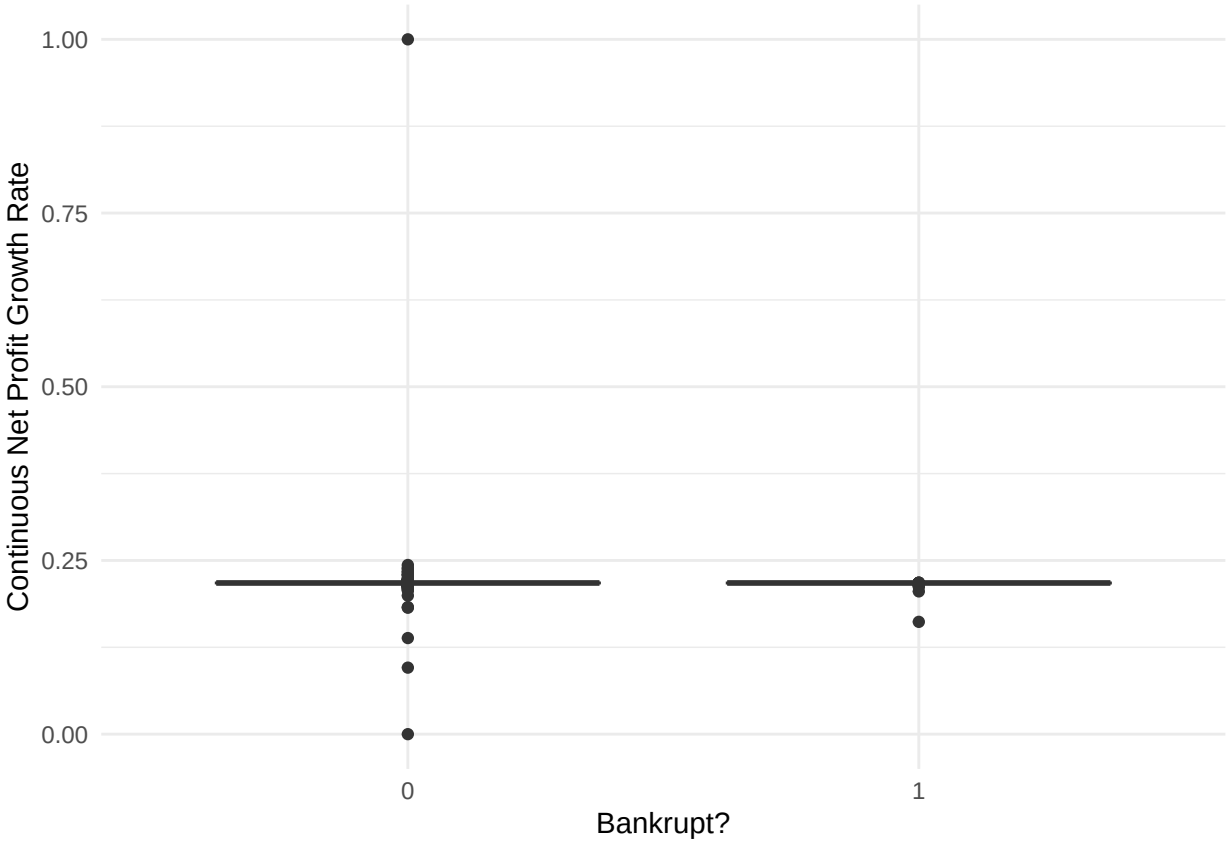

1.00
0.75
0.50
0.25
0.00
Per Share Net profit before tax (Yuan ¥)
0
1
Bankrupt?

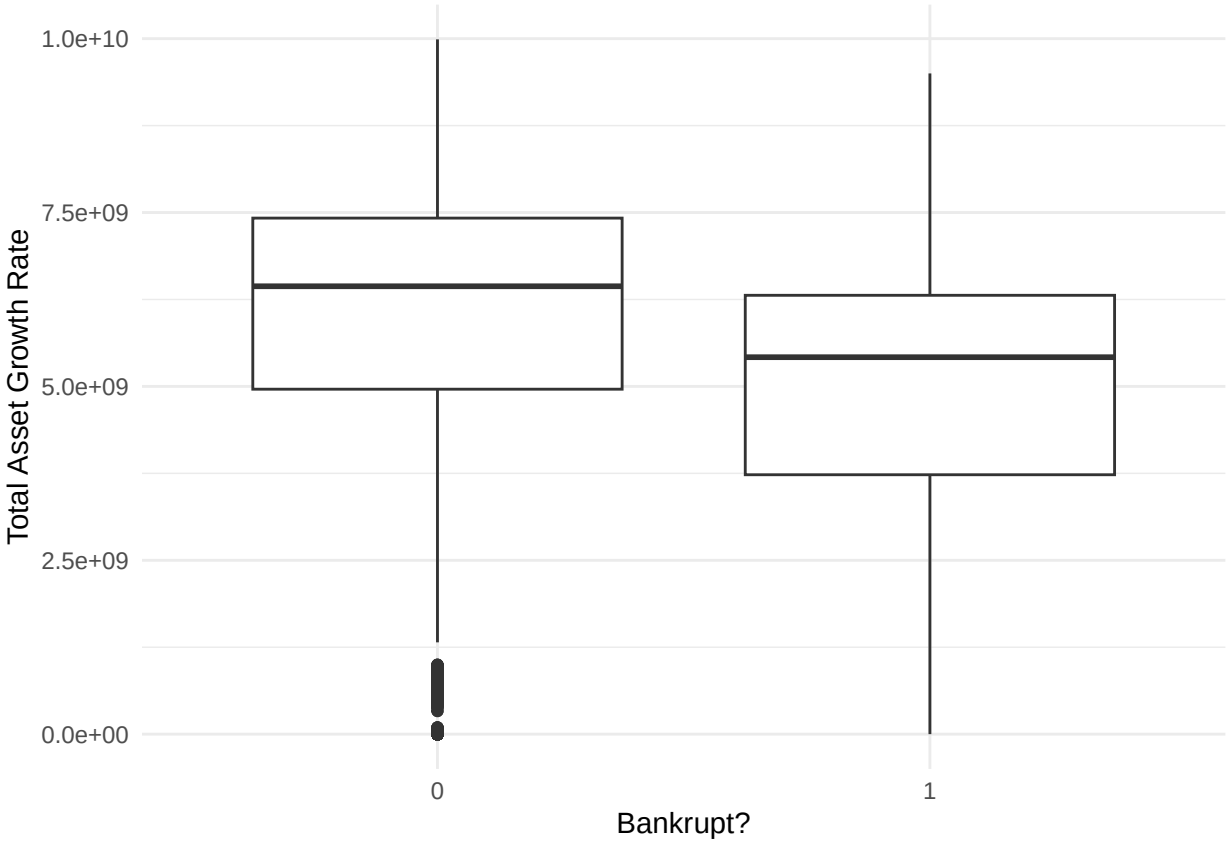

1.00
0.75
0.50
0.25
0.00
Realized Sales Gross Profit Growth Rate
0
1
Bankrupt?

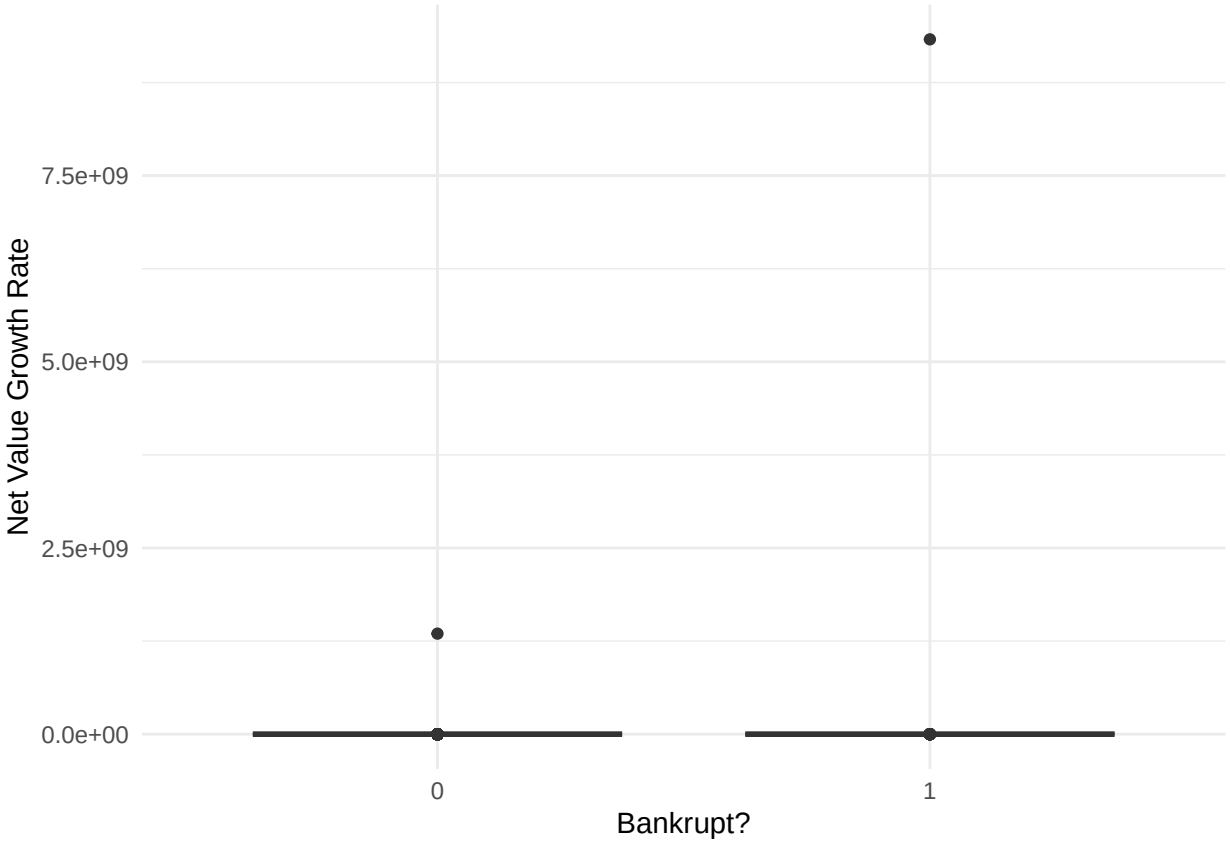

1.00
0.75
0.50
0.25
0.00
Operating Profit Growth Rate
0
1
Bankrupt?

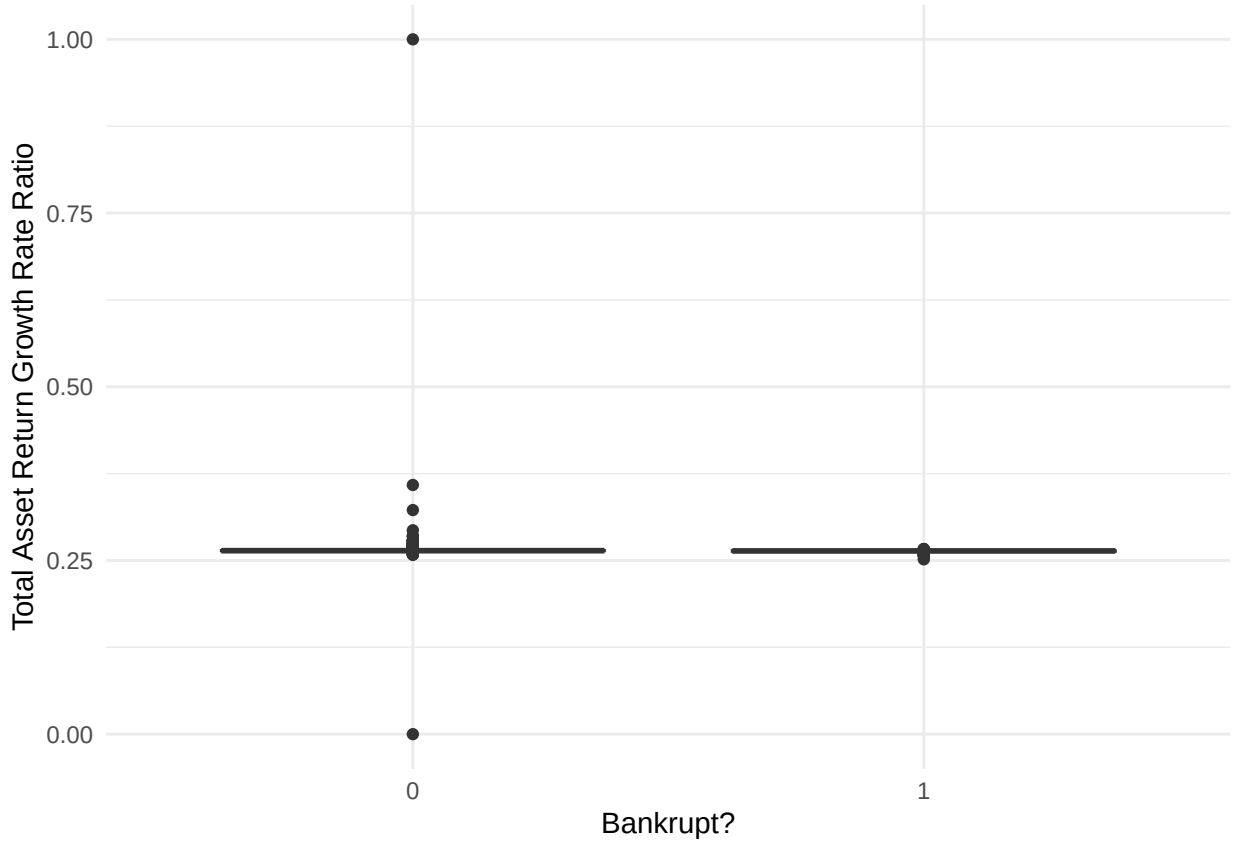

1.00
0.75
0.50
0.25
0.00
After−tax Net Profit Growth Rate
0
1
Bankrupt?


1.00
0.75
0.50
0.25
0.00
Regular Net Profit Growth Rate
0
1
Bankrupt?


1.00
0.75
0.50
0.25
0.00
Continuous Net Profit Growth Rate
0
1
Bankrupt?


1.0e+10
7.5e+09
5.0e+09
2.5e+09
0.0e+00
Total Asset Growth Rate
0
1
Bankrupt?


7.5e+09
5.0e+09
2.5e+09
0.0e+00
Net Value Growth Rate
0
1
Bankrupt?



From the plots above we have concluded that features:

1) *ROA(C)*
2) *ROA(A)*
3) *ROA(B)*
4) *Per Share Net profit before tax*
5) *After-tax Net Profit Growth Rate*
6) *Regular Net Profit Growth Rate*

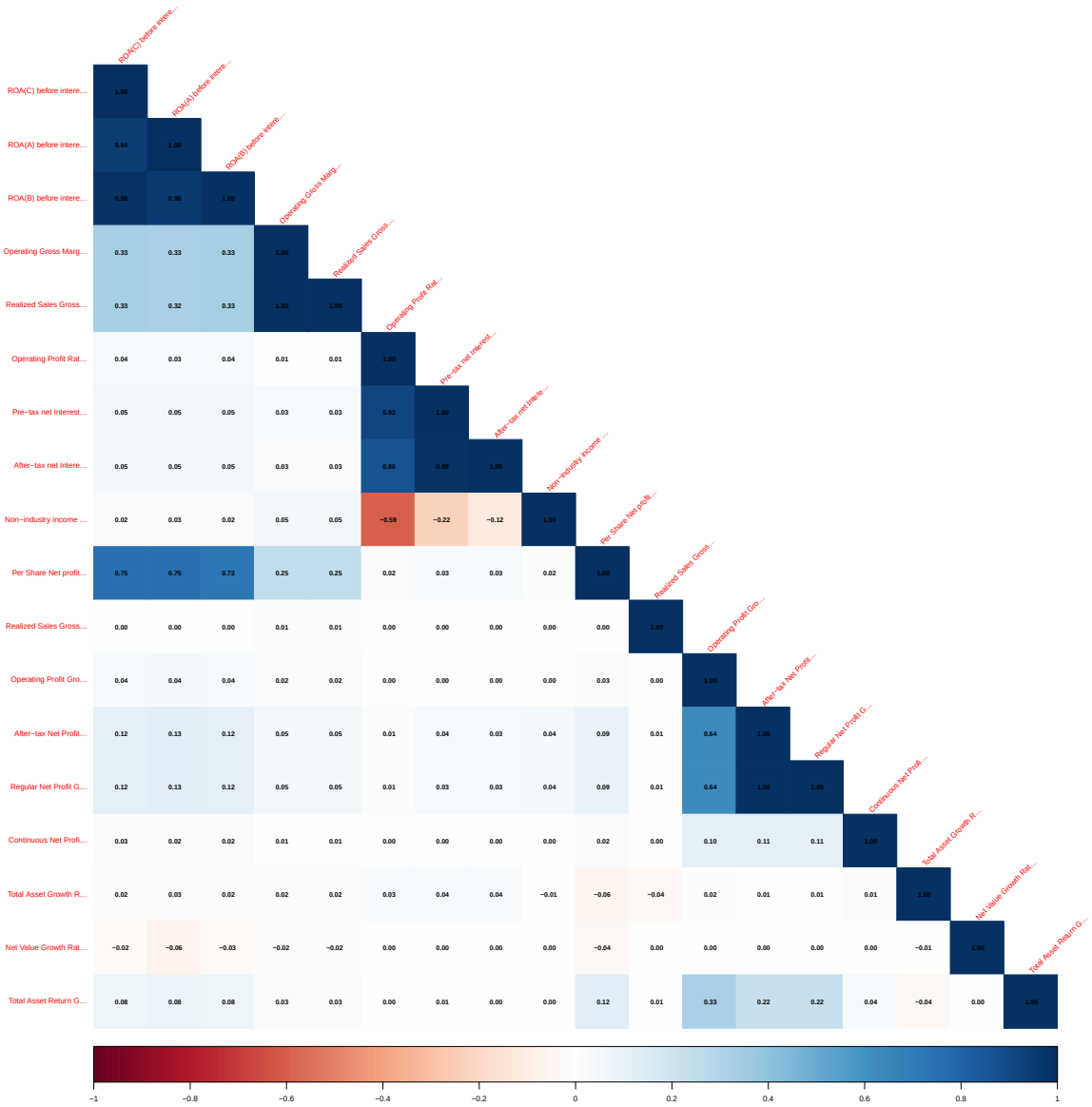
have substantial portion of outliers which will be needed to be dealt with later in data preprocessing. Some attributes like *Net Value Growth Rate* have low IQR.

Independence

For most of the classification models, the independence between individual predictors is crucial. This basically means their correlation is not significant. We can visually check that using **corrplot**.

```
show.corrplot <- function(matrix){
  p <- corrplot(
    matrix,
    method = 'color',
    type = 'lower',
    tl.cex = 0.6,         # much smaller axis text
    cl.cex = 0.6,         # smaller colorbar text
    addCoef.col = "black",
```

```
    number.cex = 0.5,     # smaller correlation coefficients
    tl.srt = 45
  )

  print(p)
}

# we need to truncate the predictor names so that the results are displayed
# better
cor_matrix <- cor(selected_data[,PREDICTOR_NAMES])
truncate_labels <- function(labels, max_length = 20) {
  sapply(labels, function(lab) {
    if (nchar(lab) > max_length) {
      paste0(substr(lab, 1, max_length), "...")  # ellipsis
    } else {
      lab
    }
  })
}

# applying truncation
colnames(cor_matrix) <- truncate_labels(colnames(cor_matrix))
rownames(cor_matrix) <- truncate_labels(rownames(cor_matrix))

show.corrplot(cor_matrix)
```




```
## $corr
##                          ROA(C) before intere... ROA(A) before intere...
## ROA(C) before intere...            1.0000000000            0.940123708
## ROA(A) before intere...            0.9401237080            1.000000000
## ROA(B) before intere...            0.9868494970            0.955740625
## Operating Gross Marg...            0.3347194403            0.326969108
## Realized Sales Gross...            0.3327553323            0.324955926
## Operating Profit Rat...            0.0357247305            0.032053341
## Pre-tax net Interest...            0.0534194181            0.053517782
## After-tax net Intere...            0.0492215053            0.049474030
## Non-industry income ...            0.0205006112            0.029648763
## Per Share Net profit...            0.7505635078            0.752578273
## Realized Sales Gross...            0.0005910674            0.003277138
## Operating Profit Gro...            0.0365105929            0.042207587
```

```
## After-tax Net Profit...          0.1150828235           0.125384033
## Regular Net Profit G...          0.1150395690           0.125871851
## Continuous Net Profi...          0.0252338116           0.024887245
## Total Asset Growth R...          0.0196350824           0.026976682
## Net Value Growth Rat...         -0.0219302238          -0.063970410
## Total Asset Return G...          0.0799056742           0.081981689
##                       ROA(B) before intere... Operating Gross Marg...
## ROA(C) before intere...           0.986849497           0.334719440
## ROA(A) before intere...           0.955740625           0.326969108
## ROA(B) before intere...           1.000000000           0.333748562
## Operating Gross Marg...           0.333748562           1.000000000
## Realized Sales Gross...           0.331755464           0.999518278
## Operating Profit Rat...           0.035212361           0.005744836
## Pre-tax net Interest...           0.053725531           0.032492729
## After-tax net Intere...           0.049952046           0.027175301
## Non-industry income ...           0.022366188           0.051437781
## Per Share Net profit...           0.722940059           0.247789488
## Realized Sales Gross...           0.002142249           0.014172027
## Operating Profit Gro...           0.036144239           0.022866904
## After-tax Net Profit...           0.117130075           0.054638554
## Regular Net Profit G...           0.117041510           0.053429778
## Continuous Net Profi...           0.024414357           0.009121179
## Total Asset Growth R...           0.022104093           0.016012790
## Net Value Growth Rat...          -0.026126724          -0.017448008
## Total Asset Return G...           0.079972127           0.026544637
##                       Realized Sales Gross... Operating Profit Rat...
## ROA(C) before intere...           0.332755332           0.035724731
## ROA(A) before intere...           0.324955926           0.032053341
## ROA(B) before intere...           0.331755464           0.035212361
## Operating Gross Marg...           0.999518278           0.005744836
## Realized Sales Gross...           1.000000000           0.005609531
## Operating Profit Rat...           0.005609531           1.000000000
## Pre-tax net Interest...           0.032231576           0.916447780
## After-tax net Intere...           0.026851472           0.862190707
## Non-industry income ...           0.051241649          -0.592005877
## Per Share Net profit...           0.246004126           0.020219389
## Realized Sales Gross...           0.014187935           0.000831085
## Operating Profit Gro...           0.022777798           0.004952375
## After-tax Net Profit...           0.054470153           0.011328235
## Regular Net Profit G...           0.053258691           0.011226598
## Continuous Net Profi...           0.009117152           0.001318140
## Total Asset Growth R...           0.016583407           0.034464841
## Net Value Growth Rat...          -0.017451246          -0.000207210
## Total Asset Return G...           0.026463425           0.003677239
##                       Pre-tax net Interest... After-tax net Intere...
## ROA(C) before intere...          0.0534194181          0.0492215053
## ROA(A) before intere...          0.0535177818          0.0494740304
## ROA(B) before intere...          0.0537255312          0.0499520459
## Operating Gross Marg...          0.0324927287          0.0271753008
## Realized Sales Gross...          0.0322315762          0.0268514724
## Operating Profit Rat...          0.9164477799          0.8621907072
## Pre-tax net Interest...          1.0000000000          0.9863790243
## After-tax net Intere...          0.9863790243          1.0000000000
## Non-industry income ...         -0.2200446550         -0.1152110385
```

```
## Per Share Net profit...          0.0340456171          0.0306213480
## Realized Sales Gross...          0.0012460443          0.0012259792
## Operating Profit Gro...          0.0039090972          0.0029624046
## After-tax Net Profit...          0.0351502735          0.0312226845
## Regular Net Profit G...          0.0349143094          0.0309638919
## Continuous Net Profi...          0.0030128650          0.0025654533
## Total Asset Growth R...          0.0376325965          0.0370659868
## Net Value Growth Rat...         -0.0009979886         -0.0008578239
## Total Asset Return G...          0.0050042235          0.0043788603
##                       Non-industry income ... Per Share Net profit...
## ROA(C) before intere...          0.0205006112          0.7505635078
## ROA(A) before intere...          0.0296487634          0.7525782732
## ROA(B) before intere...          0.0223661885          0.7229400587
## Operating Gross Marg...          0.0514377810          0.2477894885
## Realized Sales Gross...          0.0512416489          0.2460041263
## Operating Profit Rat...         -0.5920058772          0.0202193894
## Pre-tax net Interest...         -0.2200446550          0.0340456171
## After-tax net Intere...         -0.1152110385          0.0306213480
## Non-industry income ...          1.0000000000          0.0192793939
## Per Share Net profit...          0.0192793939          1.0000000000
## Realized Sales Gross...          0.0004836086         -0.0004959322
## Operating Profit Gro...         -0.0041996545          0.0297820794
## After-tax Net Profit...          0.0431788189          0.0902225554
## Regular Net Profit G...          0.0429513305          0.0903675642
## Continuous Net Profi...          0.0028549405          0.0244478466
## Total Asset Growth R...         -0.0082237311         -0.0552984947
## Net Value Growth Rat...         -0.0015048711         -0.0376294406
## Total Asset Return G...          0.0011145247          0.1201349571
##                       Realized Sales Gross... Operating Profit Gro...
## ROA(C) before intere...          0.0005910674           0.036510593
## ROA(A) before intere...          0.0032771376           0.042207587
## ROA(B) before intere...          0.0021422492           0.036144239
## Operating Gross Marg...          0.0141720269           0.022866904
## Realized Sales Gross...          0.0141879348           0.022777798
## Operating Profit Rat...          0.0008310850           0.004952375
## Pre-tax net Interest...          0.0012460443           0.003909097
## After-tax net Intere...          0.0012259792           0.002962405
## Non-industry income ...          0.0004836086          -0.004199655
## Per Share Net profit...         -0.0004959322           0.029782079
## Realized Sales Gross...          1.0000000000           0.002192332
## Operating Profit Gro...          0.0021923316           1.000000000
## After-tax Net Profit...          0.0064700515           0.639394403
## Regular Net Profit G...          0.0064442338           0.636792820
## Continuous Net Profi...          0.0007466083           0.100821405
## Total Asset Growth R...         -0.0351155875           0.015553436
## Net Value Growth Rat...         -0.0006979139          -0.004534360
## Total Asset Return G...          0.0058427712           0.326720441
##                       After-tax Net Profit... Regular Net Profit G...
## ROA(C) before intere...           0.115082824           0.115039569
## ROA(A) before intere...           0.125384033           0.125871851
## ROA(B) before intere...           0.117130075           0.117041510
## Operating Gross Marg...           0.054638554           0.053429778
## Realized Sales Gross...           0.054470153           0.053258691
## Operating Profit Rat...           0.011328235           0.011226598
```

```
## Pre-tax net Interest...           0.035150274           0.034914309
## After-tax net Intere...           0.031222685           0.030963892
## Non-industry income ...           0.043178819           0.042951331
## Per Share Net profit...           0.090222555           0.090367564
## Realized Sales Gross...           0.006470051           0.006444234
## Operating Profit Gro...           0.639394403           0.636792820
## After-tax Net Profit...           1.000000000           0.996186193
## Regular Net Profit G...           0.996186193           1.000000000
## Continuous Net Profi...           0.113051390           0.112903712
## Total Asset Growth R...           0.008039462           0.008910851
## Net Value Growth Rat...          -0.003660496          -0.003649268
## Total Asset Return G...           0.223918939           0.223105724
##                       Continuous Net Profi... Total Asset Growth R...
## ROA(C) before intere...          0.0252338116           0.019635082
## ROA(A) before intere...          0.0248872448           0.026976682
## ROA(B) before intere...          0.0244143572           0.022104093
## Operating Gross Marg...          0.0091211788           0.016012790
## Realized Sales Gross...          0.0091171517           0.016583407
## Operating Profit Rat...          0.0013181404           0.034464841
## Pre-tax net Interest...          0.0030128650           0.037632596
## After-tax net Intere...          0.0025654533           0.037065987
## Non-industry income ...          0.0028549405          -0.008223731
## Per Share Net profit...          0.0244478466          -0.055298495
## Realized Sales Gross...          0.0007466083          -0.035115588
## Operating Profit Gro...          0.1008214050           0.015553436
## After-tax Net Profit...          0.1130513905           0.008039462
## Regular Net Profit G...          0.1129037121           0.008910851
## Continuous Net Profi...          1.0000000000           0.010175352
## Total Asset Growth R...          0.0101753525           1.000000000
## Net Value Growth Rat...         -0.0002704139          -0.008688357
## Total Asset Return G...          0.0361980100          -0.037135519
##                       Net Value Growth Rat... Total Asset Return G...
## ROA(C) before intere...         -0.0219302238           0.079905674
## ROA(A) before intere...         -0.0639704100           0.081981689
## ROA(B) before intere...         -0.0261267244           0.079972127
## Operating Gross Marg...         -0.0174480079           0.026544637
## Realized Sales Gross...         -0.0174512459           0.026463425
## Operating Profit Rat...         -0.0002072100           0.003677239
## Pre-tax net Interest...         -0.0009979886           0.005004224
## After-tax net Intere...         -0.0008578239           0.004378860
## Non-industry income ...         -0.0015048711           0.001114525
## Per Share Net profit...         -0.0376294406           0.120134957
## Realized Sales Gross...         -0.0006979139           0.005842771
## Operating Profit Gro...         -0.0045343597           0.326720441
## After-tax Net Profit...         -0.0036604962           0.223918939
## Regular Net Profit G...         -0.0036492682           0.223105724
## Continuous Net Profi...         -0.0002704139           0.036198010
## Total Asset Growth R...         -0.0086883569          -0.037135519
## Net Value Growth Rat...          1.0000000000          -0.004652796
## Total Asset Return G...         -0.0046527960           1.000000000
##
## $corrPos
##                     xName                 yName  x  y          corr
## 1   ROA(C) before intere... ROA(C) before intere...  1 18  1.0000000000
```

```
## 2   ROA(C) before intere... ROA(A) before intere...  1 17  0.9401237080
## 3   ROA(C) before intere... ROA(B) before intere...  1 16  0.9868494970
## 4   ROA(C) before intere... Operating Gross Marg...  1 15  0.3347194403
## 5   ROA(C) before intere... Realized Sales Gross...  1 14  0.3327553323
## 6   ROA(C) before intere... Operating Profit Rat...  1 13  0.0357247305
## 7   ROA(C) before intere... Pre-tax net Interest...  1 12  0.0534194181
## 8   ROA(C) before intere... After-tax net Intere...  1 11  0.0492215053
## 9   ROA(C) before intere... Non-industry income ...  1 10  0.0205006112
## 10  ROA(C) before intere... Per Share Net profit...  1  9  0.7505635078
## 11  ROA(C) before intere... Realized Sales Gross...  1  8  0.0005910674
## 12  ROA(C) before intere... Operating Profit Gro...  1  7  0.0365105929
## 13  ROA(C) before intere... After-tax Net Profit...  1  6  0.1150828235
## 14  ROA(C) before intere... Regular Net Profit G...  1  5  0.1150395690
## 15  ROA(C) before intere... Continuous Net Profi...  1  4  0.0252338116
## 16  ROA(C) before intere... Total Asset Growth R...  1  3  0.0196350824
## 17  ROA(C) before intere... Net Value Growth Rat...  1  2 -0.0219302238
## 18  ROA(C) before intere... Total Asset Return G...  1  1  0.0799056742
## 19  ROA(A) before intere... ROA(A) before intere...  2 17  1.0000000000
## 20  ROA(A) before intere... ROA(B) before intere...  2 16  0.9557406253
## 21  ROA(A) before intere... Operating Gross Marg...  2 15  0.3269691085
## 22  ROA(A) before intere... Realized Sales Gross...  2 14  0.3249559264
## 23  ROA(A) before intere... Operating Profit Rat...  2 13  0.0320533406
## 24  ROA(A) before intere... Pre-tax net Interest...  2 12  0.0535177818
## 25  ROA(A) before intere... After-tax net Intere...  2 11  0.0494740304
## 26  ROA(A) before intere... Non-industry income ...  2 10  0.0296487634
## 27  ROA(A) before intere... Per Share Net profit...  2  9  0.7525782732
## 28  ROA(A) before intere... Realized Sales Gross...  2  8  0.0032771376
## 29  ROA(A) before intere... Operating Profit Gro...  2  7  0.0422075868
## 30  ROA(A) before intere... After-tax Net Profit...  2  6  0.1253840331
## 31  ROA(A) before intere... Regular Net Profit G...  2  5  0.1258718506
## 32  ROA(A) before intere... Continuous Net Profi...  2  4  0.0248872448
## 33  ROA(A) before intere... Total Asset Growth R...  2  3  0.0269766817
## 34  ROA(A) before intere... Net Value Growth Rat...  2  2 -0.0639704100
## 35  ROA(A) before intere... Total Asset Return G...  2  1  0.0819816887
## 36  ROA(B) before intere... ROA(B) before intere...  3 16  1.0000000000
## 37  ROA(B) before intere... Operating Gross Marg...  3 15  0.3337485618
## 38  ROA(B) before intere... Realized Sales Gross...  3 14  0.3317554638
## 39  ROA(B) before intere... Operating Profit Rat...  3 13  0.0352123605
## 40  ROA(B) before intere... Pre-tax net Interest...  3 12  0.0537255312
## 41  ROA(B) before intere... After-tax net Intere...  3 11  0.0499520459
## 42  ROA(B) before intere... Non-industry income ...  3 10  0.0223661885
## 43  ROA(B) before intere... Per Share Net profit...  3  9  0.7229400587
## 44  ROA(B) before intere... Realized Sales Gross...  3  8  0.0021422492
## 45  ROA(B) before intere... Operating Profit Gro...  3  7  0.0361442385
## 46  ROA(B) before intere... After-tax Net Profit...  3  6  0.1171300748
## 47  ROA(B) before intere... Regular Net Profit G...  3  5  0.1170415103
## 48  ROA(B) before intere... Continuous Net Profi...  3  4  0.0244143572
## 49  ROA(B) before intere... Total Asset Growth R...  3  3  0.0221040931
## 50  ROA(B) before intere... Net Value Growth Rat...  3  2 -0.0261267244
## 51  ROA(B) before intere... Total Asset Return G...  3  1  0.0799721270
## 52  Operating Gross Marg... Operating Gross Marg...  4 15  1.0000000000
## 53  Operating Gross Marg... Realized Sales Gross...  4 14  0.9995182781
## 54  Operating Gross Marg... Operating Profit Rat...  4 13  0.0057448359
## 55  Operating Gross Marg... Pre-tax net Interest...  4 12  0.0324927287
```

```
## 56  Operating Gross Marg... After-tax net Intere...  4 11  0.0271753008
## 57  Operating Gross Marg... Non-industry income ...  4 10  0.0514377810
## 58  Operating Gross Marg... Per Share Net profit...  4  9  0.2477894885
## 59  Operating Gross Marg... Realized Sales Gross...  4  8  0.0141720269
## 60  Operating Gross Marg... Operating Profit Gro...  4  7  0.0228669035
## 61  Operating Gross Marg... After-tax Net Profit...  4  6  0.0546385535
## 62  Operating Gross Marg... Regular Net Profit G...  4  5  0.0534297784
## 63  Operating Gross Marg... Continuous Net Profi...  4  4  0.0091211788
## 64  Operating Gross Marg... Total Asset Growth R...  4  3  0.0160127899
## 65  Operating Gross Marg... Net Value Growth Rat...  4  2 -0.0174480079
## 66  Operating Gross Marg... Total Asset Return G...  4  1  0.0265446373
## 67  Realized Sales Gross... Realized Sales Gross...  5 14  1.0000000000
## 68  Realized Sales Gross... Operating Profit Rat...  5 13  0.0056095307
## 69  Realized Sales Gross... Pre-tax net Interest...  5 12  0.0322315762
## 70  Realized Sales Gross... After-tax net Intere...  5 11  0.0268514724
## 71  Realized Sales Gross... Non-industry income ...  5 10  0.0512416489
## 72  Realized Sales Gross... Per Share Net profit...  5  9  0.2460041263
## 73  Realized Sales Gross... Realized Sales Gross...  5  8  0.0141879348
## 74  Realized Sales Gross... Operating Profit Gro...  5  7  0.0227777984
## 75  Realized Sales Gross... After-tax Net Profit...  5  6  0.0544701528
## 76  Realized Sales Gross... Regular Net Profit G...  5  5  0.0532586912
## 77  Realized Sales Gross... Continuous Net Profi...  5  4  0.0091171517
## 78  Realized Sales Gross... Total Asset Growth R...  5  3  0.0165834072
## 79  Realized Sales Gross... Net Value Growth Rat...  5  2 -0.0174512459
## 80  Realized Sales Gross... Total Asset Return G...  5  1  0.0264634250
## 81  Operating Profit Rat... Operating Profit Rat...  6 13  1.0000000000
## 82  Operating Profit Rat... Pre-tax net Interest...  6 12  0.9164477799
## 83  Operating Profit Rat... After-tax net Intere...  6 11  0.8621907072
## 84  Operating Profit Rat... Non-industry income ...  6 10 -0.5920058772
## 85  Operating Profit Rat... Per Share Net profit...  6  9  0.0202193894
## 86  Operating Profit Rat... Realized Sales Gross...  6  8  0.0008310850
## 87  Operating Profit Rat... Operating Profit Gro...  6  7  0.0049523751
## 88  Operating Profit Rat... After-tax Net Profit...  6  6  0.0113282354
## 89  Operating Profit Rat... Regular Net Profit G...  6  5  0.0112265979
## 90  Operating Profit Rat... Continuous Net Profi...  6  4  0.0013181404
## 91  Operating Profit Rat... Total Asset Growth R...  6  3  0.0344648411
## 92  Operating Profit Rat... Net Value Growth Rat...  6  2 -0.0002072100
## 93  Operating Profit Rat... Total Asset Return G...  6  1  0.0036772389
## 94  Pre-tax net Interest... Pre-tax net Interest...  7 12  1.0000000000
## 95  Pre-tax net Interest... After-tax net Intere...  7 11  0.9863790243
## 96  Pre-tax net Interest... Non-industry income ...  7 10 -0.2200446550
## 97  Pre-tax net Interest... Per Share Net profit...  7  9  0.0340456171
## 98  Pre-tax net Interest... Realized Sales Gross...  7  8  0.0012460443
## 99  Pre-tax net Interest... Operating Profit Gro...  7  7  0.0039090972
## 100 Pre-tax net Interest... After-tax Net Profit...  7  6  0.0351502735
## 101 Pre-tax net Interest... Regular Net Profit G...  7  5  0.0349143094
## 102 Pre-tax net Interest... Continuous Net Profi...  7  4  0.0030128650
## 103 Pre-tax net Interest... Total Asset Growth R...  7  3  0.0376325965
## 104 Pre-tax net Interest... Net Value Growth Rat...  7  2 -0.0009979886
## 105 Pre-tax net Interest... Total Asset Return G...  7  1  0.0050042235
## 106 After-tax net Intere... After-tax net Intere...  8 11  1.0000000000
## 107 After-tax net Intere... Non-industry income ...  8 10 -0.1152110385
## 108 After-tax net Intere... Per Share Net profit...  8  9  0.0306213480
## 109 After-tax net Intere... Realized Sales Gross...  8  8  0.0012259792
```

```
## 110 After-tax net Intere... Operating Profit Gro...  8  7  0.0029624046
## 111 After-tax net Intere... After-tax Net Profit...  8  6  0.0312226845
## 112 After-tax net Intere... Regular Net Profit G...  8  5  0.0309638919
## 113 After-tax net Intere... Continuous Net Profi...  8  4  0.0025654533
## 114 After-tax net Intere... Total Asset Growth R...  8  3  0.0370659868
## 115 After-tax net Intere... Net Value Growth Rat...  8  2 -0.0008578239
## 116 After-tax net Intere... Total Asset Return G...  8  1  0.0043788603
## 117 Non-industry income ... Non-industry income ...  9 10  1.0000000000
## 118 Non-industry income ... Per Share Net profit...  9  9  0.0192793939
## 119 Non-industry income ... Realized Sales Gross...  9  8  0.0004836086
## 120 Non-industry income ... Operating Profit Gro...  9  7 -0.0041996545
## 121 Non-industry income ... After-tax Net Profit...  9  6  0.0431788189
## 122 Non-industry income ... Regular Net Profit G...  9  5  0.0429513305
## 123 Non-industry income ... Continuous Net Profi...  9  4  0.0028549405
## 124 Non-industry income ... Total Asset Growth R...  9  3 -0.0082237311
## 125 Non-industry income ... Net Value Growth Rat...  9  2 -0.0015048711
## 126 Non-industry income ... Total Asset Return G...  9  1  0.0011145247
## 127 Per Share Net profit... Per Share Net profit... 10  9  1.0000000000
## 128 Per Share Net profit... Realized Sales Gross... 10  8 -0.0004959322
## 129 Per Share Net profit... Operating Profit Gro... 10  7  0.0297820794
## 130 Per Share Net profit... After-tax Net Profit... 10  6  0.0902225554
## 131 Per Share Net profit... Regular Net Profit G... 10  5  0.0903675642
## 132 Per Share Net profit... Continuous Net Profi... 10  4  0.0244478466
## 133 Per Share Net profit... Total Asset Growth R... 10  3 -0.0552984947
## 134 Per Share Net profit... Net Value Growth Rat... 10  2 -0.0376294406
## 135 Per Share Net profit... Total Asset Return G... 10  1  0.1201349571
## 136 Realized Sales Gross... Realized Sales Gross... 11  8  1.0000000000
## 137 Realized Sales Gross... Operating Profit Gro... 11  7  0.0021923316
## 138 Realized Sales Gross... After-tax Net Profit... 11  6  0.0064700515
## 139 Realized Sales Gross... Regular Net Profit G... 11  5  0.0064442338
## 140 Realized Sales Gross... Continuous Net Profi... 11  4  0.0007466083
## 141 Realized Sales Gross... Total Asset Growth R... 11  3 -0.0351155875
## 142 Realized Sales Gross... Net Value Growth Rat... 11  2 -0.0006979139
## 143 Realized Sales Gross... Total Asset Return G... 11  1  0.0058427712
## 144 Operating Profit Gro... Operating Profit Gro... 12  7  1.0000000000
## 145 Operating Profit Gro... After-tax Net Profit... 12  6  0.6393944028
## 146 Operating Profit Gro... Regular Net Profit G... 12  5  0.6367928200
## 147 Operating Profit Gro... Continuous Net Profi... 12  4  0.1008214050
## 148 Operating Profit Gro... Total Asset Growth R... 12  3  0.0155534360
## 149 Operating Profit Gro... Net Value Growth Rat... 12  2 -0.0045343597
## 150 Operating Profit Gro... Total Asset Return G... 12  1  0.3267204408
## 151 After-tax Net Profit... After-tax Net Profit... 13  6  1.0000000000
## 152 After-tax Net Profit... Regular Net Profit G... 13  5  0.9961861926
## 153 After-tax Net Profit... Continuous Net Profi... 13  4  0.1130513905
## 154 After-tax Net Profit... Total Asset Growth R... 13  3  0.0080394625
## 155 After-tax Net Profit... Net Value Growth Rat... 13  2 -0.0036604962
## 156 After-tax Net Profit... Total Asset Return G... 13  1  0.2239189393
## 157 Regular Net Profit G... Regular Net Profit G... 14  5  1.0000000000
## 158 Regular Net Profit G... Continuous Net Profi... 14  4  0.1129037121
## 159 Regular Net Profit G... Total Asset Growth R... 14  3  0.0089108509
## 160 Regular Net Profit G... Net Value Growth Rat... 14  2 -0.0036492682
## 161 Regular Net Profit G... Total Asset Return G... 14  1  0.2231057243
## 162 Continuous Net Profi... Continuous Net Profi... 15  4  1.0000000000
## 163 Continuous Net Profi... Total Asset Growth R... 15  3  0.0101753525
```

```
## 164 Continuous Net Profi... Net Value Growth Rat... 15  2 -0.0002704139
## 165 Continuous Net Profi... Total Asset Return G... 15  1  0.0361980100
## 166 Total Asset Growth R... Total Asset Growth R... 16  3  1.0000000000
## 167 Total Asset Growth R... Net Value Growth Rat... 16  2 -0.0086883569
## 168 Total Asset Growth R... Total Asset Return G... 16  1 -0.0371355194
## 169 Net Value Growth Rat... Net Value Growth Rat... 17  2  1.0000000000
## 170 Net Value Growth Rat... Total Asset Return G... 17  1 -0.0046527960
## 171 Total Asset Return G... Total Asset Return G... 18  1  1.0000000000
##
## $arg
## $arg$type
## [1] "lower"
```

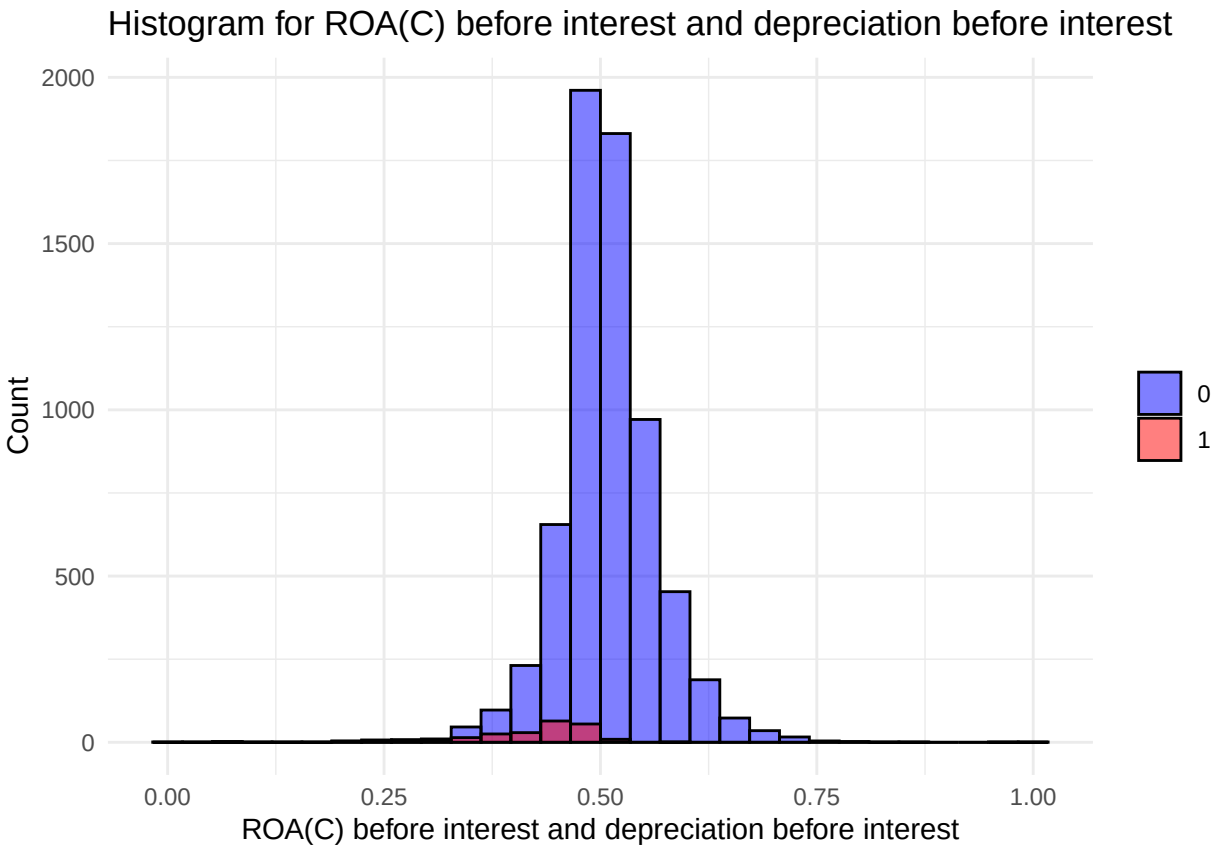
After checking the matrix we can see there are some **significant dependence** between attribues like ROA(C), ROA(B), etc.

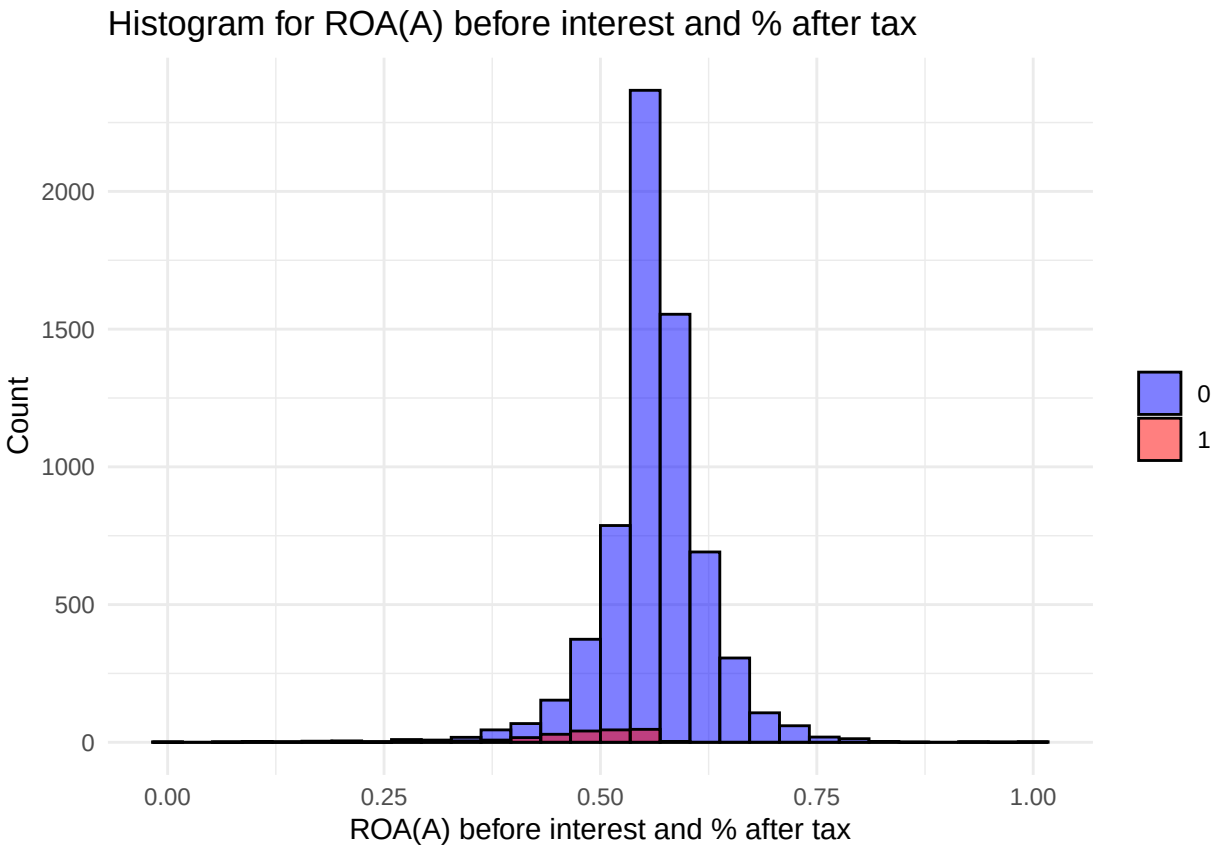
Distribution

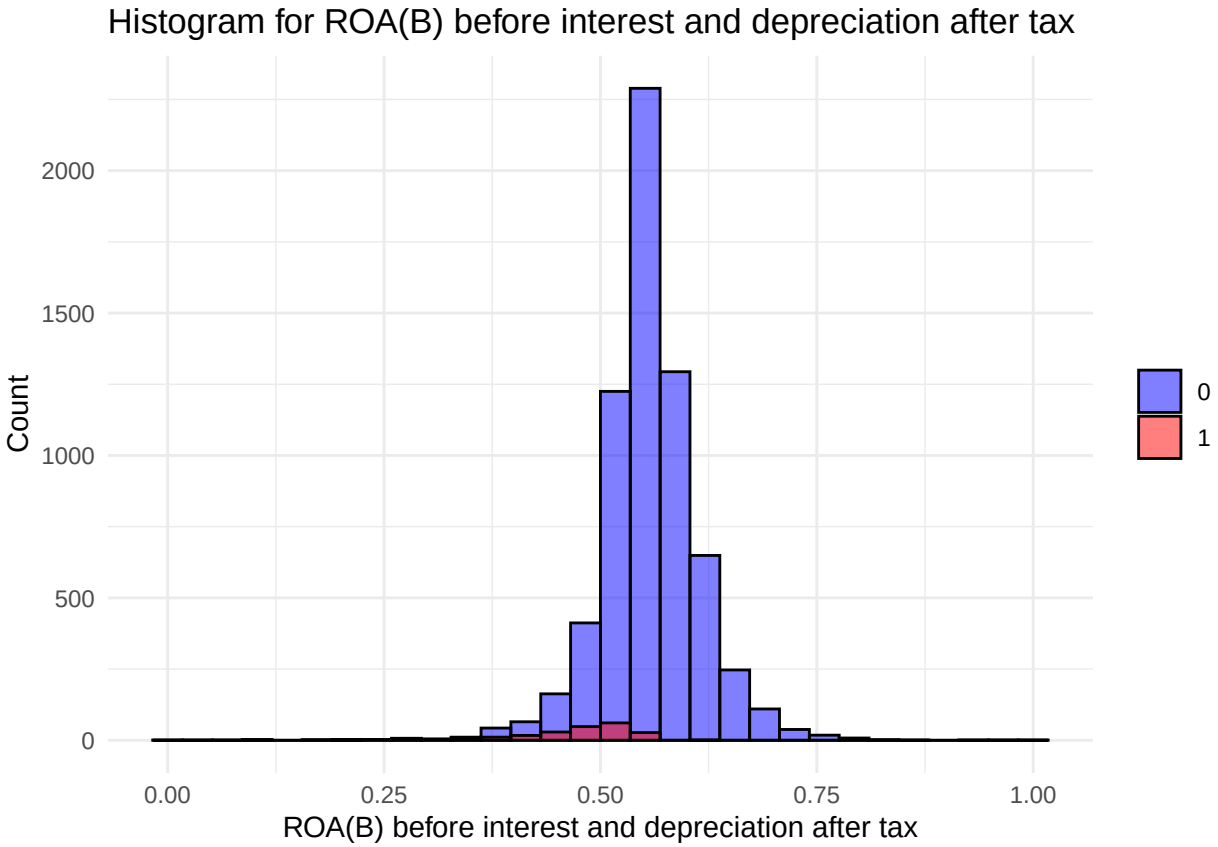
Class counts

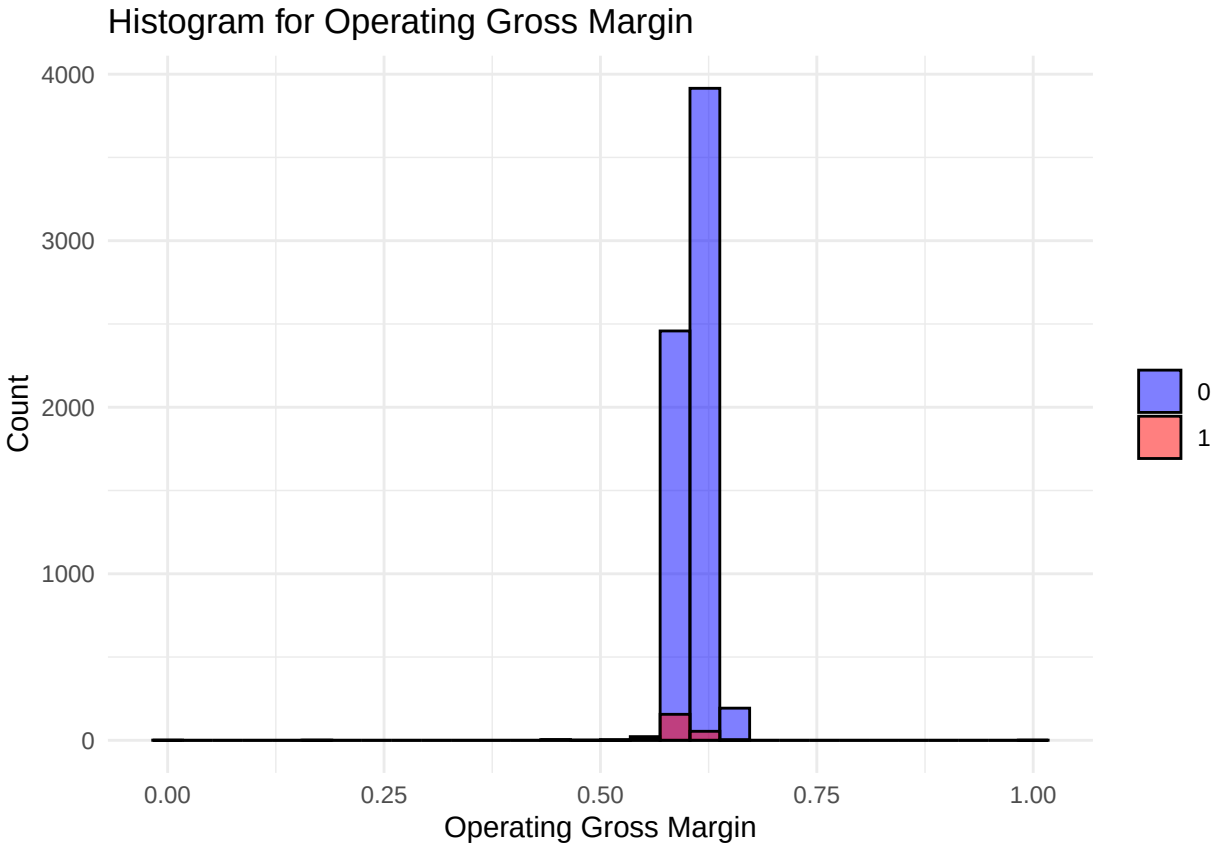
First we are going to utilize **Histogram** to display the distribution for individual classes.

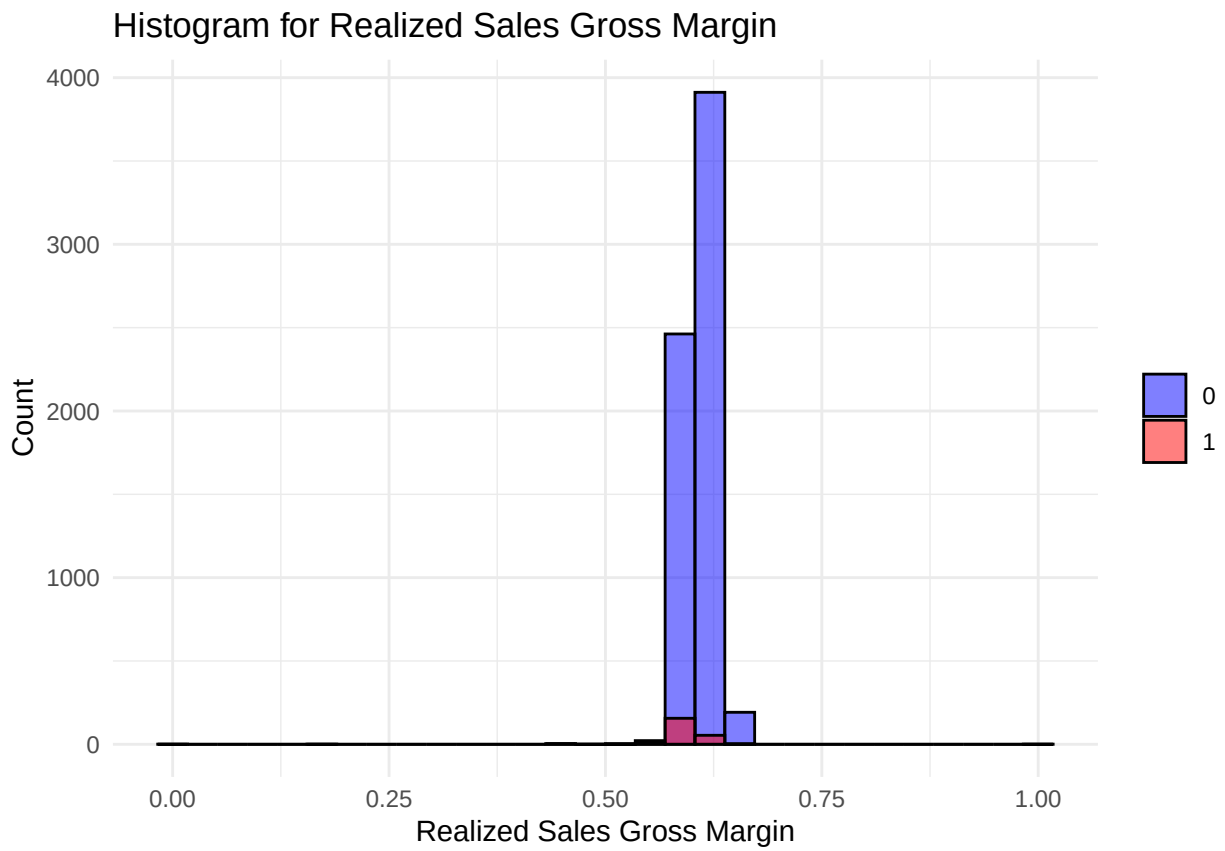
```
show.histogram <- function(df, cols, target) {

  plot_list <- map(cols, ~ {
    ggplot(data = df, aes(x = .data[[.x]], fill = as.factor(.data[[target]]))) +
      geom_histogram(position = "identity", alpha = 0.5, bins = 30, color = "black") +
      scale_fill_manual(values = c('blue', 'red')) +  # Customize fill colors
      labs(title = paste("Histogram for", .x), x = .x, y = "Count") +
      theme_minimal() +
      theme(legend.title = element_blank())
  })

  walk(plot_list, print)
}

# Usage example:
show.histogram(selected_data, PREDICTOR_NAMES, TARGET)
```

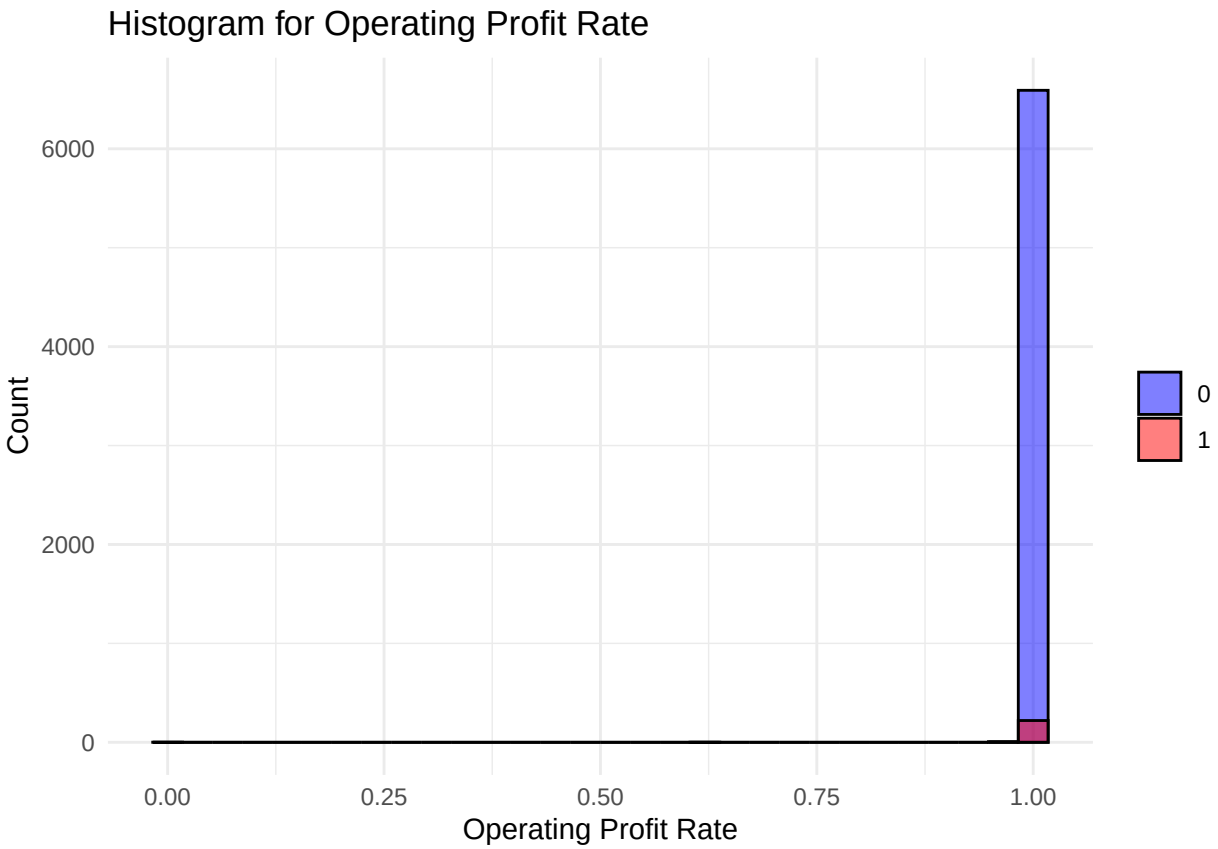

Histogram for ROA(C) before interest and depreciation before interest
Count
2000
1500
1000
500
0
0.00
0.25
0.50
0.75
1.00
ROA(C) before interest and depreciation before interest
0
1

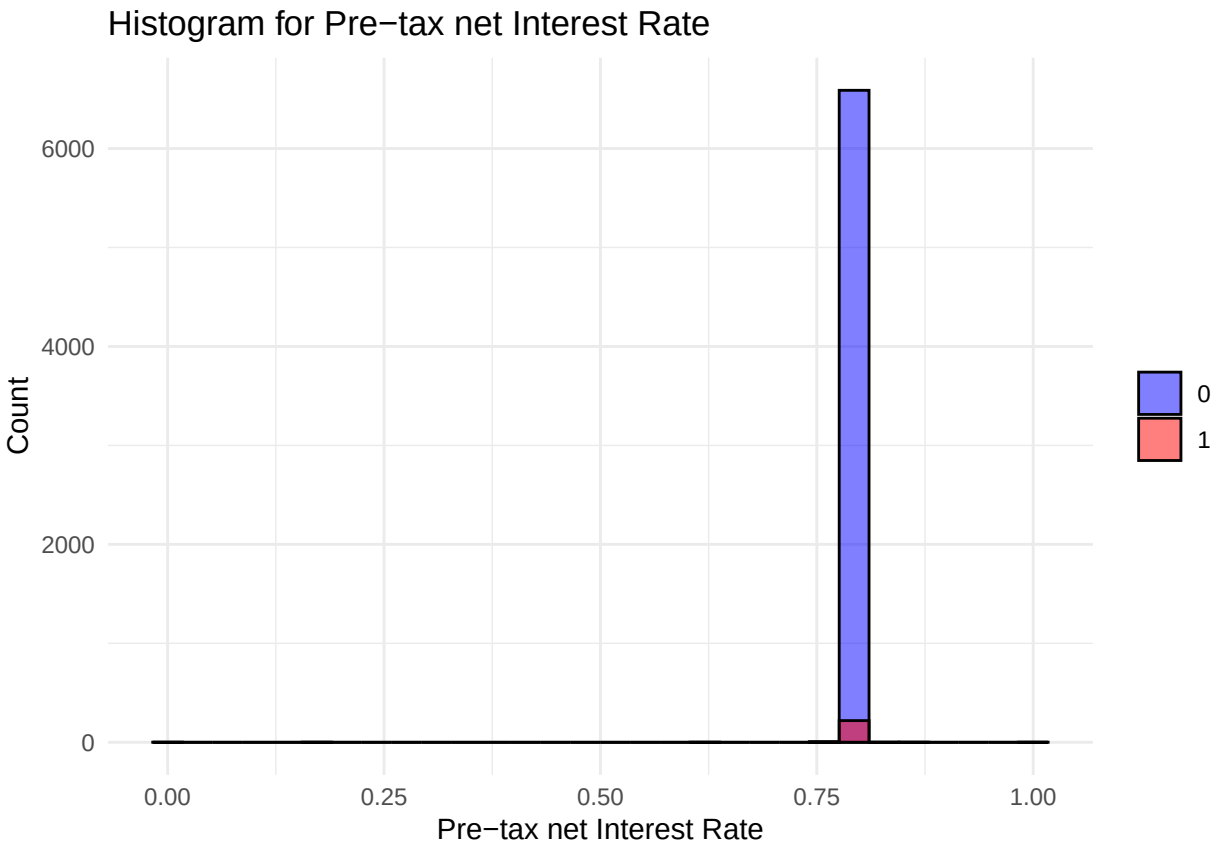

Histogram for ROA(A) before interest and % after tax
Count
2000
1500
1000
500
0
0.00
0.25
0.50
0.75
1.00
ROA(A) before interest and % after tax
0
1

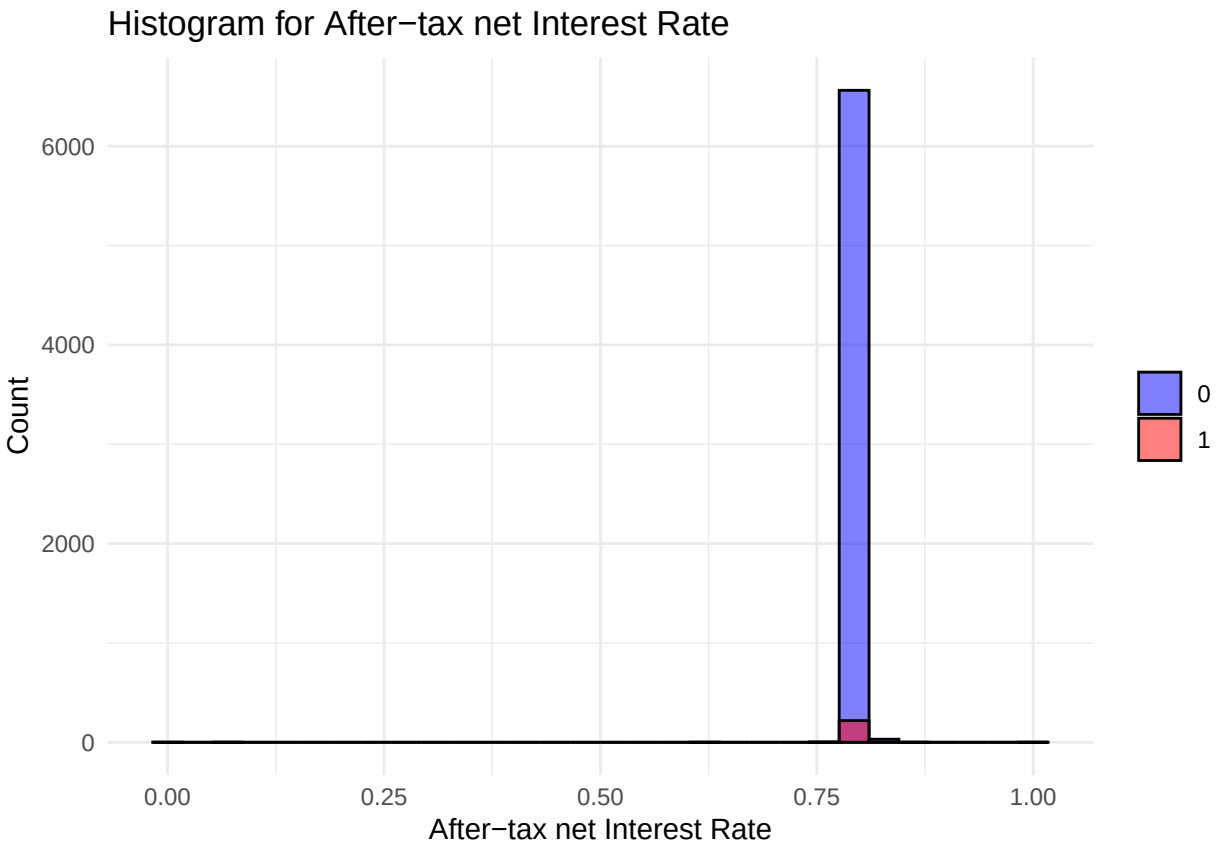

Histogram for ROA(B) before interest and depreciation after tax
Count
2000
1500
1000
500
0
0.00
0.25
0.50
0.75
1.00
ROA(B) before interest and depreciation after tax
0
1

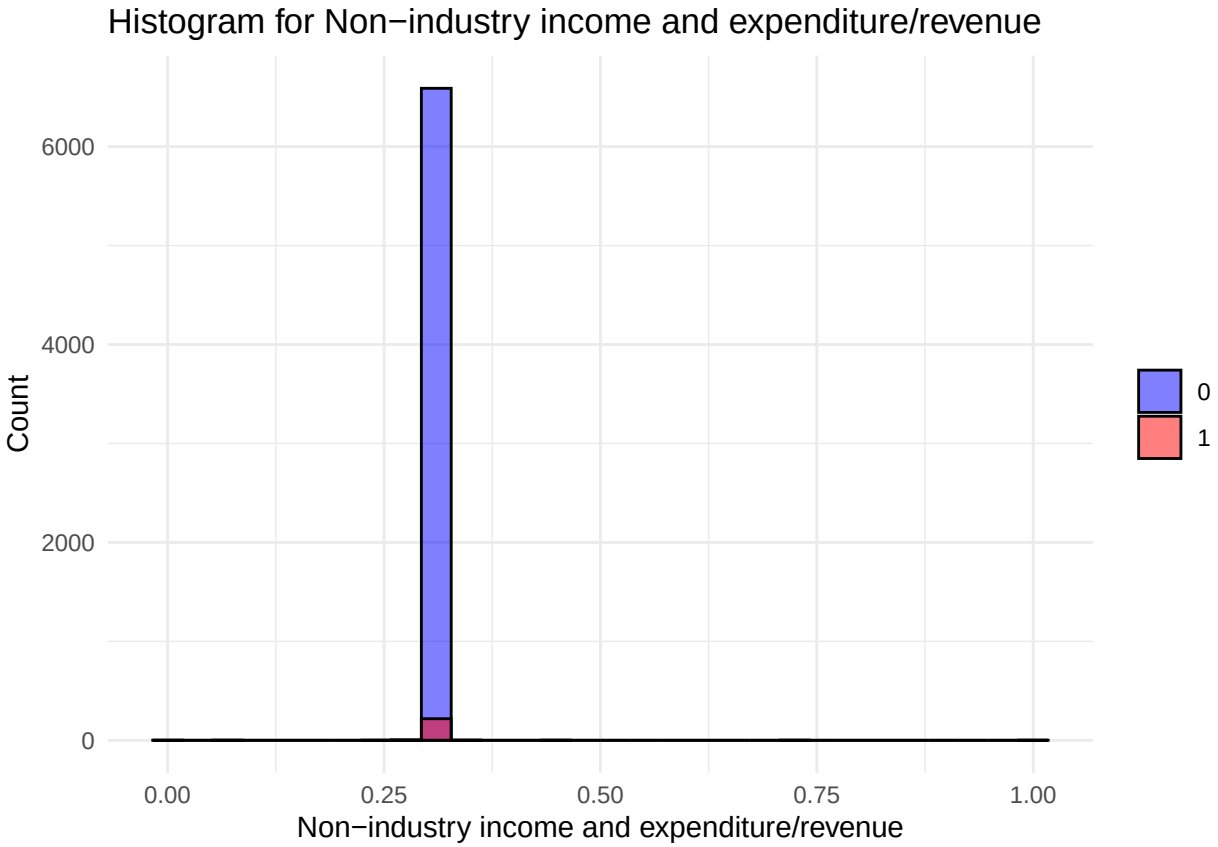

Histogram for Operating Gross Margin
Count
4000
3000
2000
1000
0
0.00
0.25
0.50
0.75
1.00
Operating Gross Margin
0
1

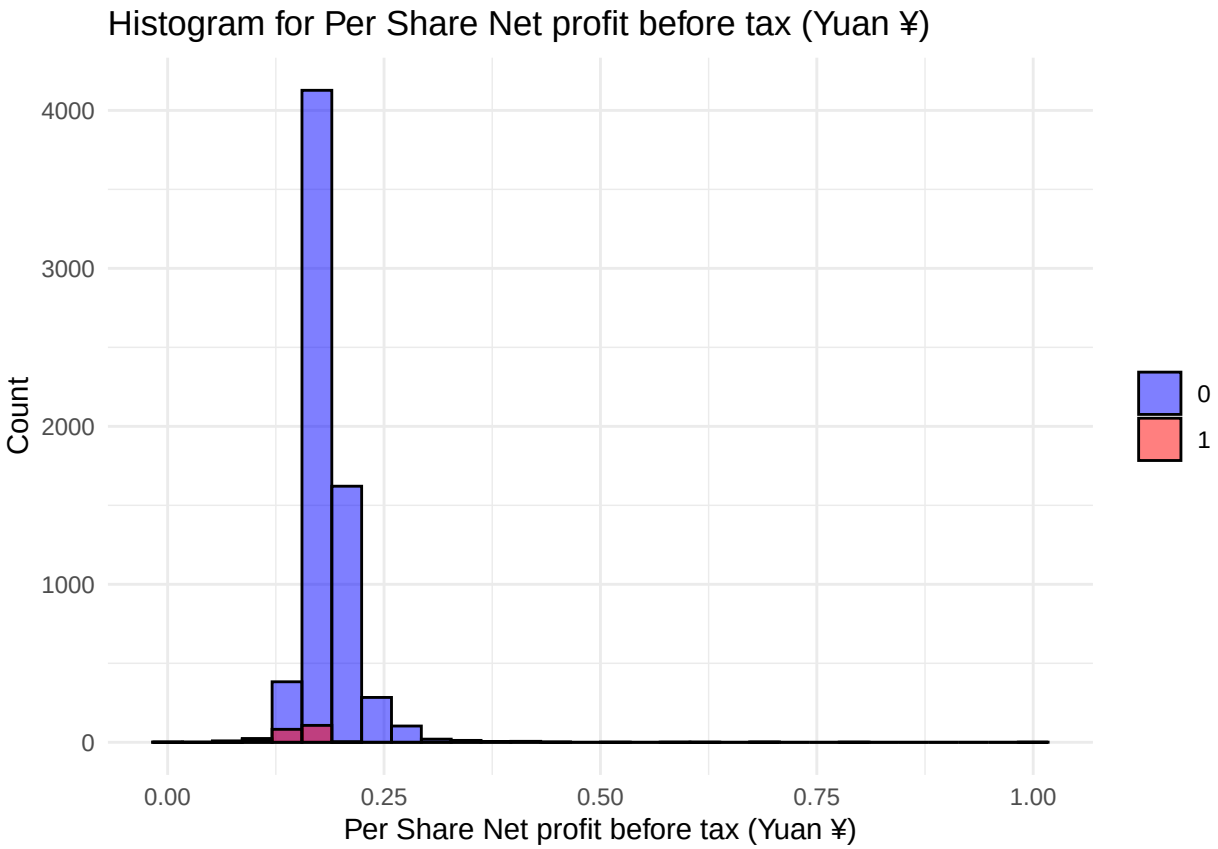

Histogram for Realized Sales Gross Margin
Count
4000
3000
2000
1000
0
0.00
0.25
0.50
0.75
1.00
Realized Sales Gross Margin
0
1

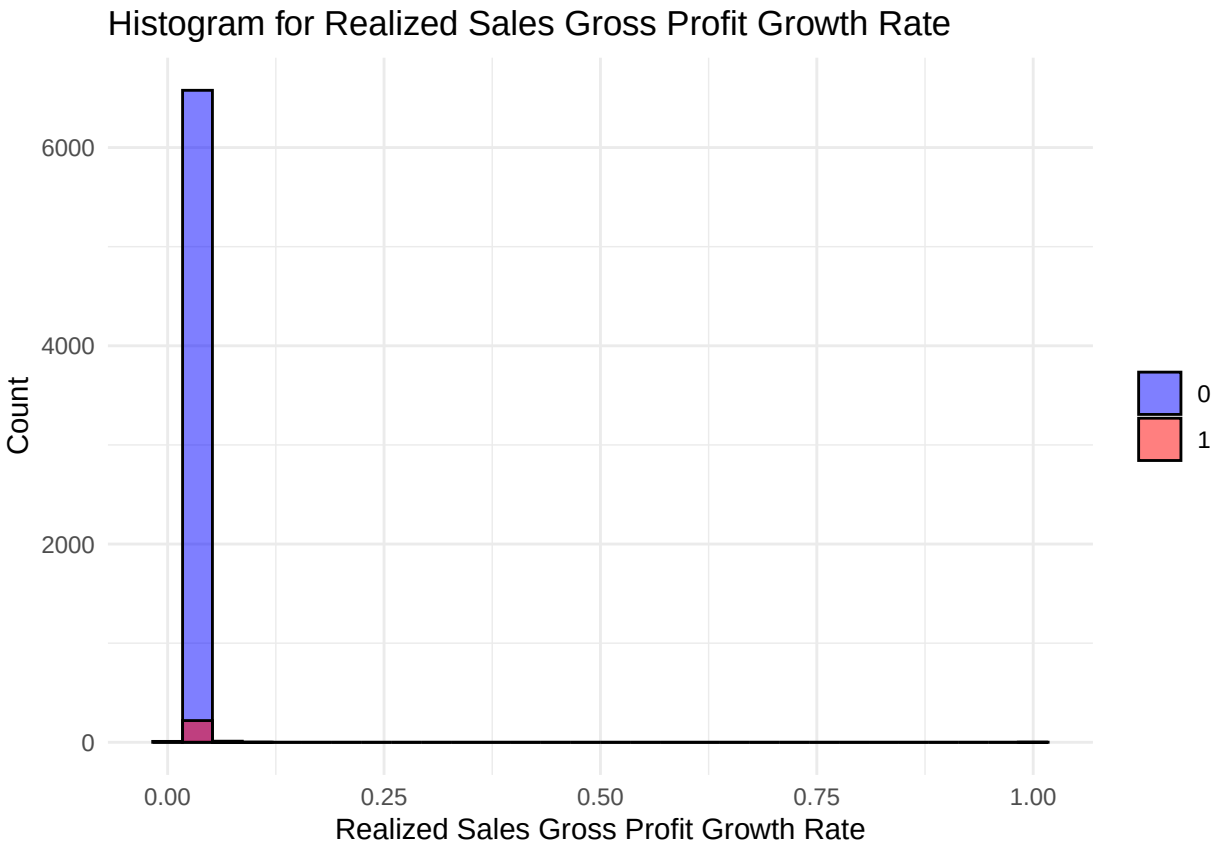

Histogram for Operating Profit Rate
Count
6000
4000
2000
0
0.00
0.25
0.50
0.75
1.00
Operating Profit Rate
0
1

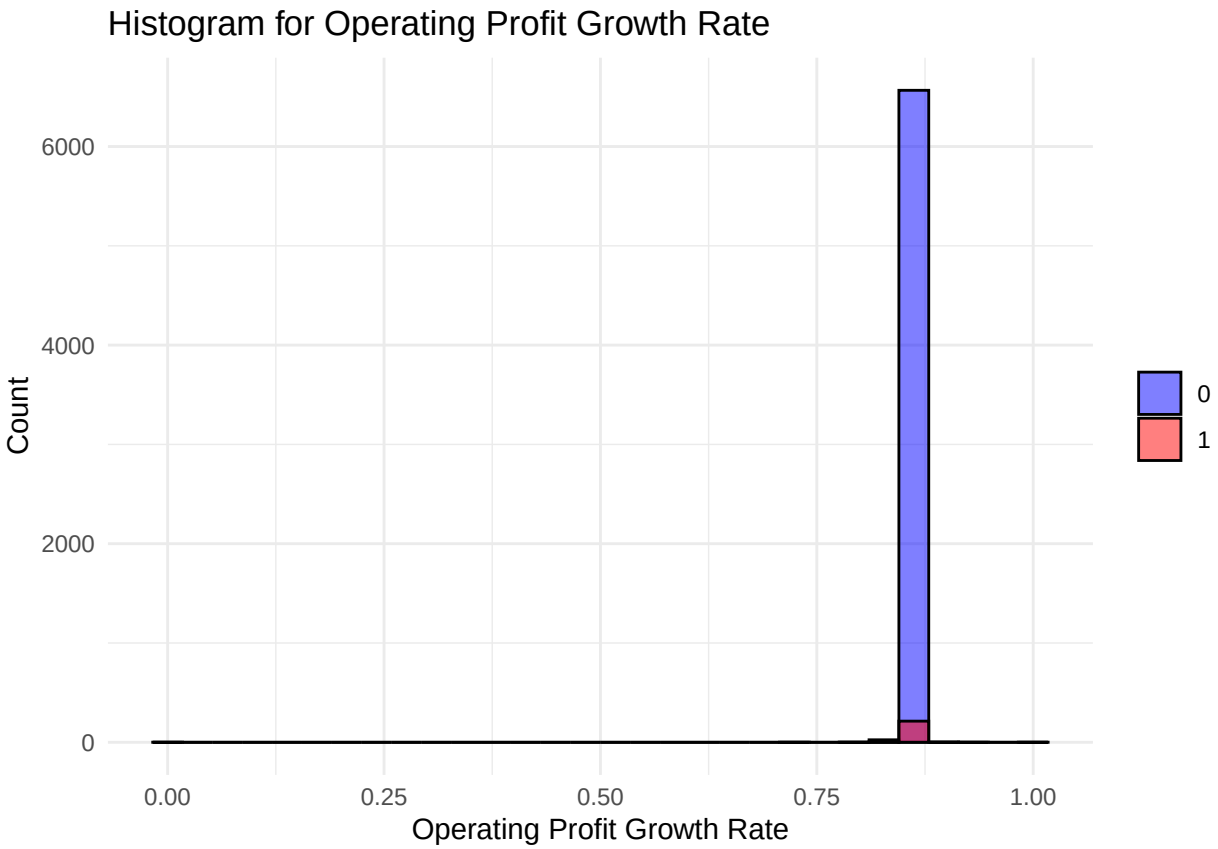

Histogram for Pre−tax net Interest Rate
Count
6000
4000
2000
0
0.00
0.25
0.50
0.75
1.00
Pre−tax net Interest Rate
0
1

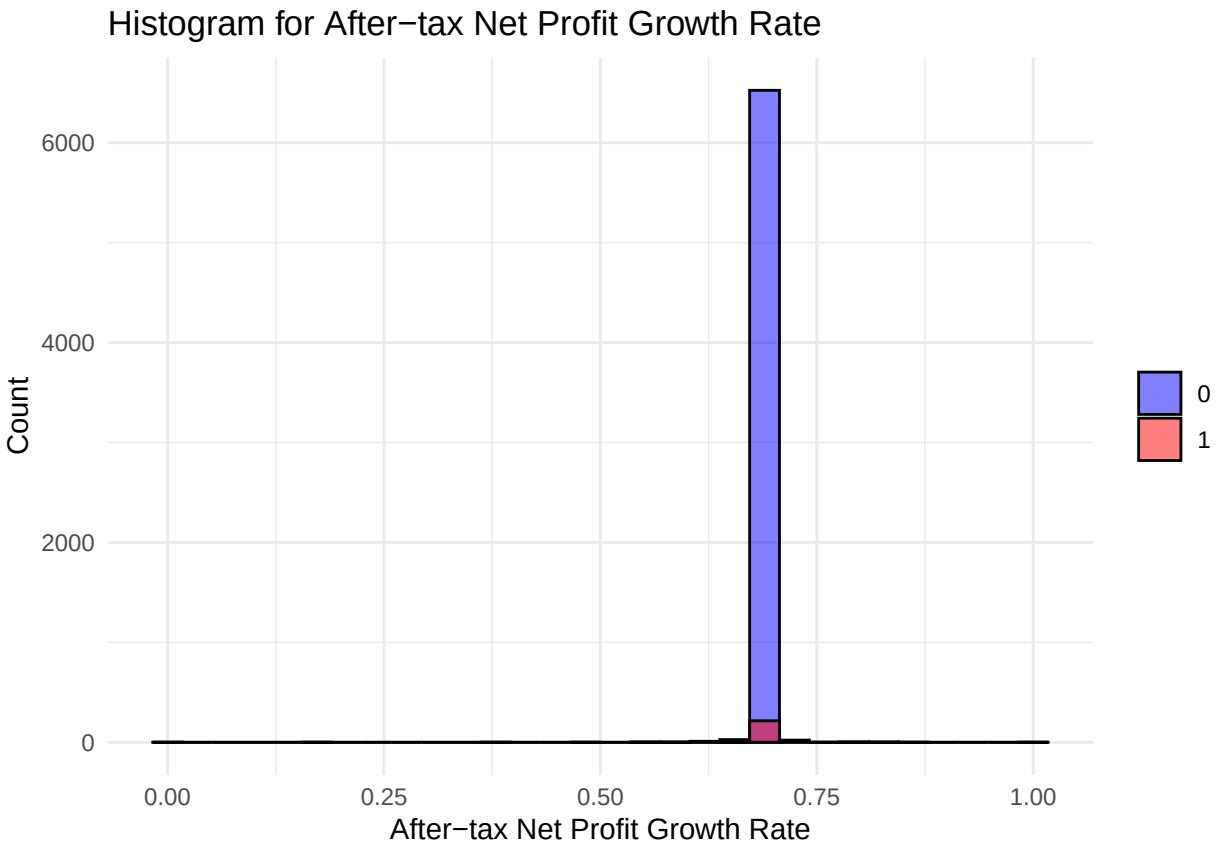

Histogram for After−tax net Interest Rate
Count
6000
4000
2000
0
0.00
0.25
0.50
0.75
1.00
After−tax net Interest Rate
0
1

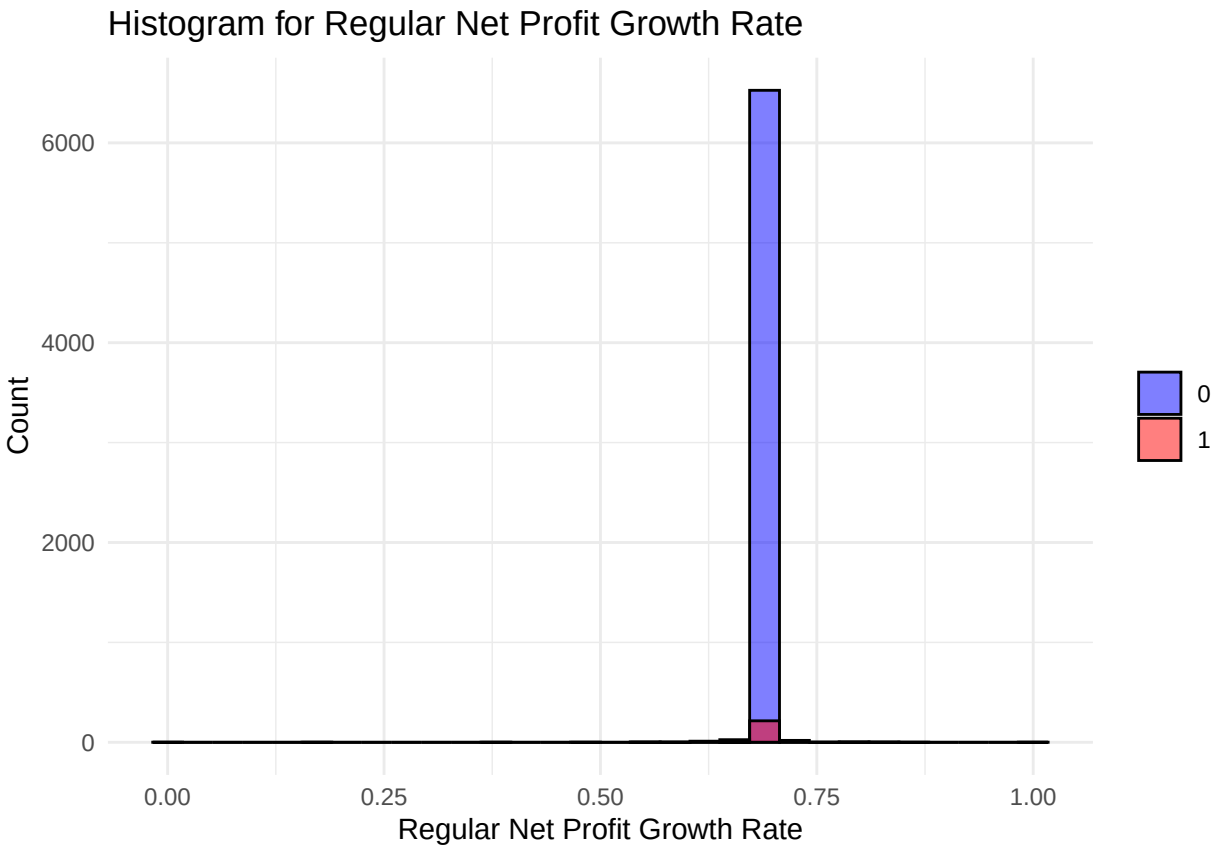

Histogram for Non−industry income and expenditure/revenue
Count
6000
4000
2000
0
0.00
0.25
0.50
0.75
1.00
Non−industry income and expenditure/revenue
0
1

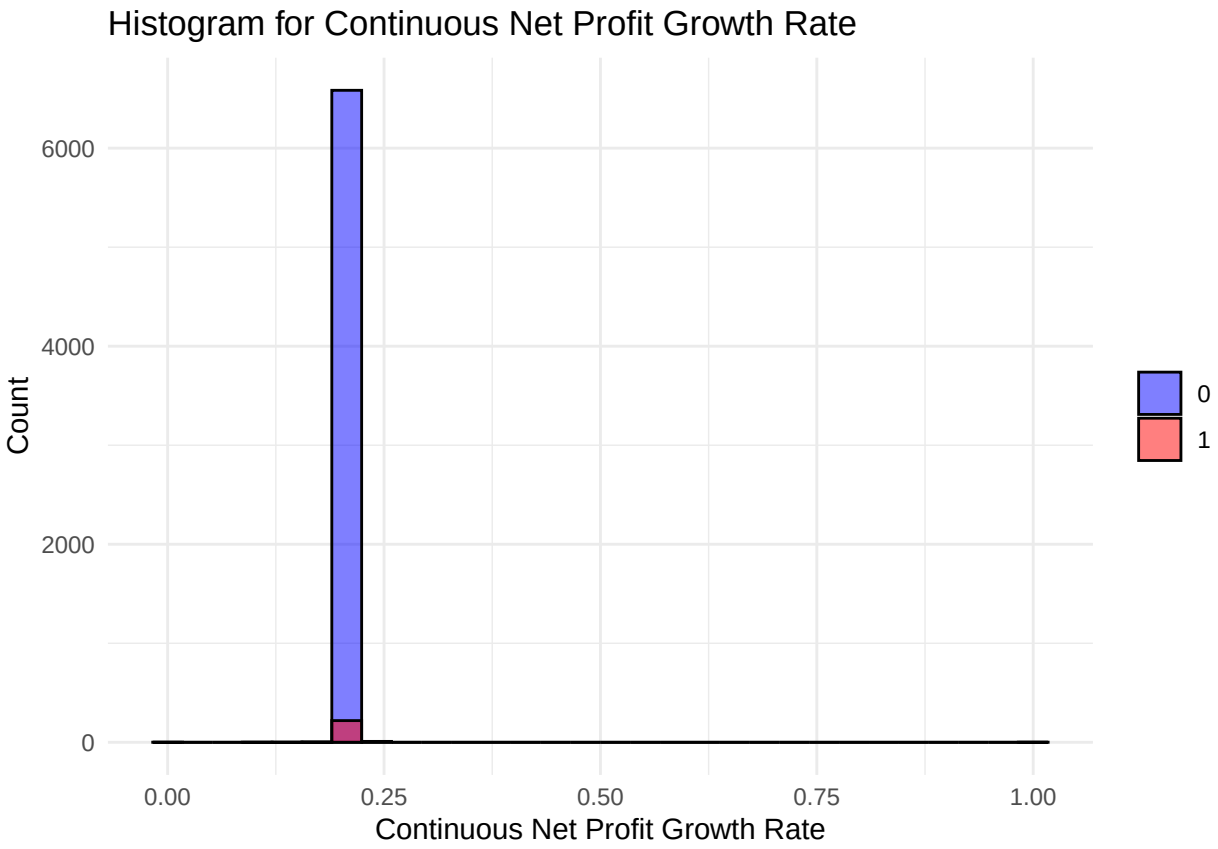

Histogram for Per Share Net profit before tax (Yuan ¥)
Count
4000
3000
2000
1000
0
0.00
0.25
0.50
0.75
1.00
Per Share Net profit before tax (Yuan ¥)
0
1

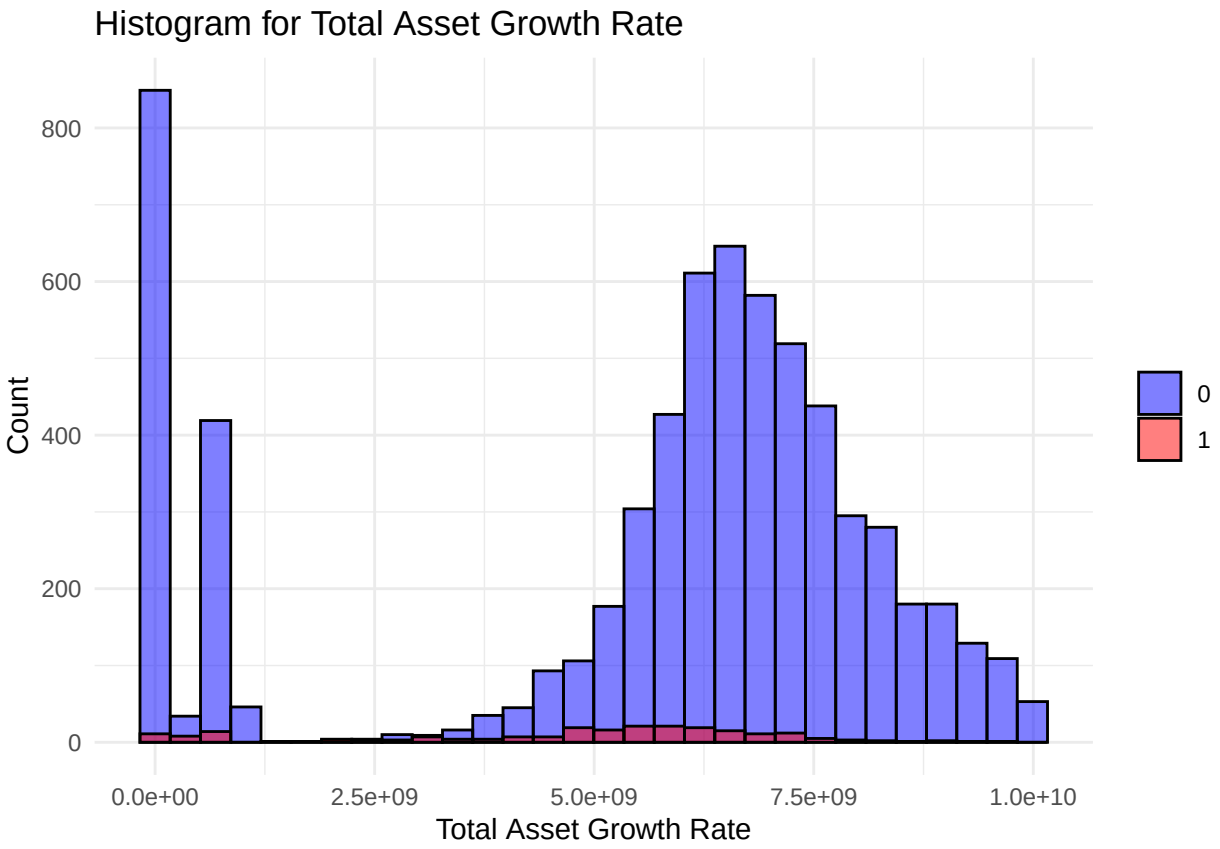

Histogram for Realized Sales Gross Profit Growth Rate
Count
6000
4000
2000
0
0.00
0.25
0.50
0.75
1.00
Realized Sales Gross Profit Growth Rate
0
1

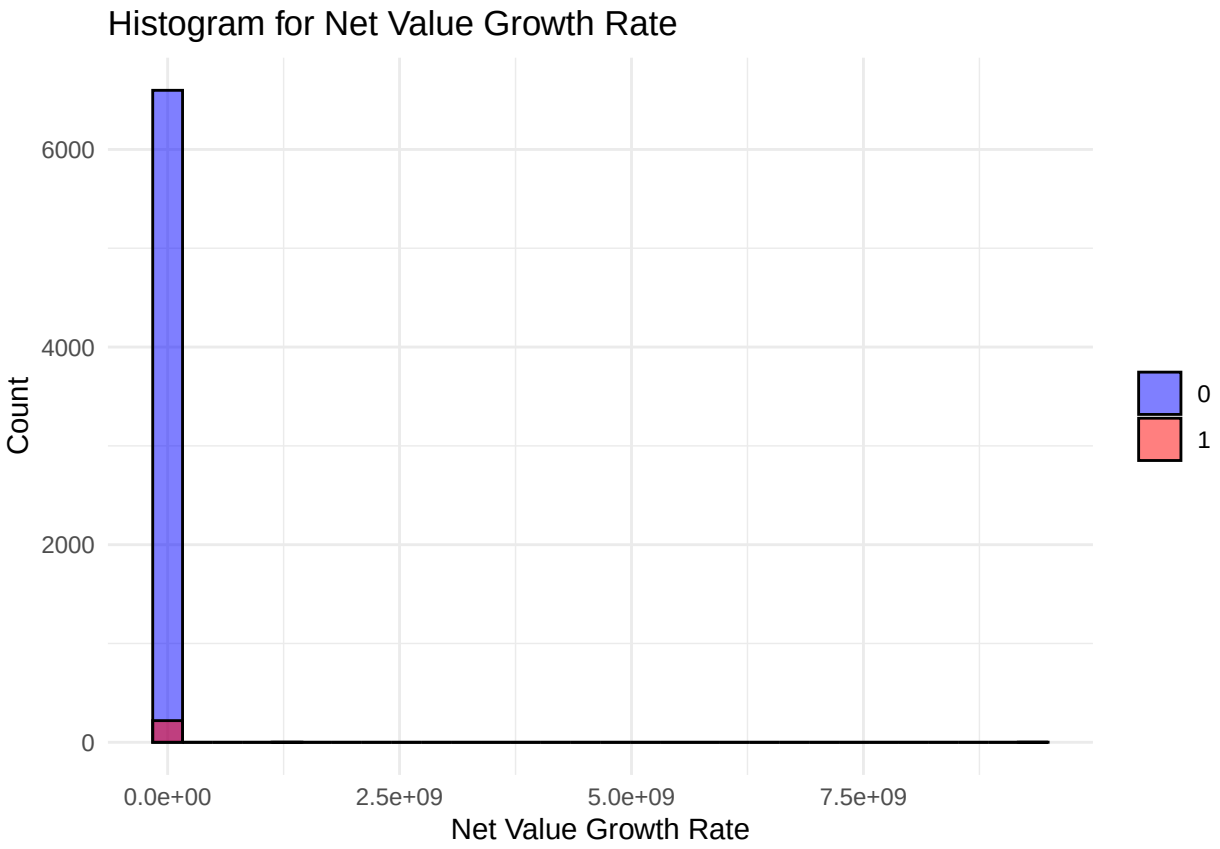

Histogram for Operating Profit Growth Rate
Count
6000
4000
2000
0
0.00
0.25
0.50
0.75
1.00
Operating Profit Growth Rate
0
1

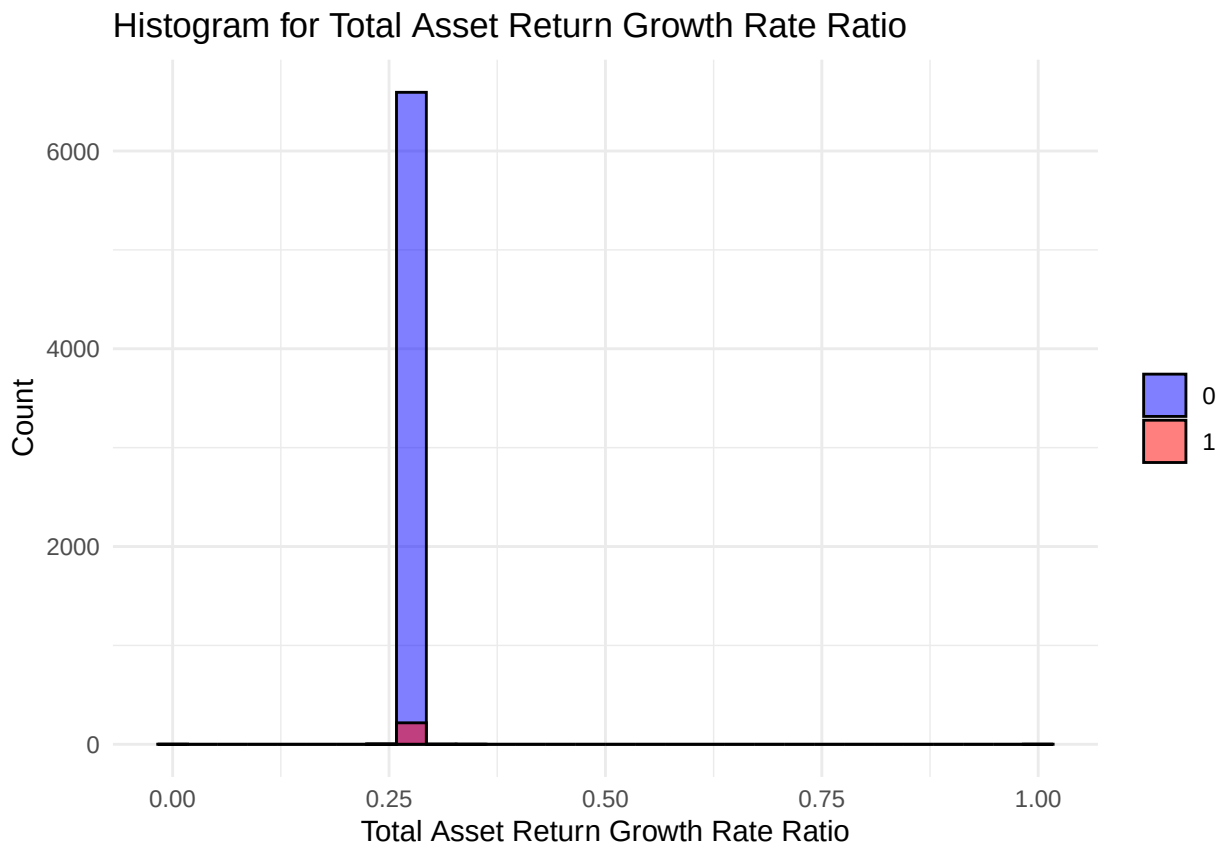

Histogram for After−tax Net Profit Growth Rate
Count
6000
4000
2000
0
0.00
0.25
0.50
0.75
1.00
After−tax Net Profit Growth Rate
0
1


Histogram for Regular Net Profit Growth Rate
Count
6000
4000
2000
0
0.00
0.25
0.50
0.75
1.00
Regular Net Profit Growth Rate
0
1


Histogram for Continuous Net Profit Growth Rate
Count
6000
4000
2000
0
0.00
0.25
0.50
0.75
1.00
Continuous Net Profit Growth Rate
0
1


Histogram for Total Asset Growth Rate
Count
800
600
400
200
0
0.0e+00
2.5e+09
5.0e+09
7.5e+09
1.0e+10
Total Asset Growth Rate
0
1


Histogram for Net Value Growth Rate
Count
6000
4000
2000
0
0.0e+00
2.5e+09
5.0e+09
7.5e+09
Net Value Growth Rate
0
1

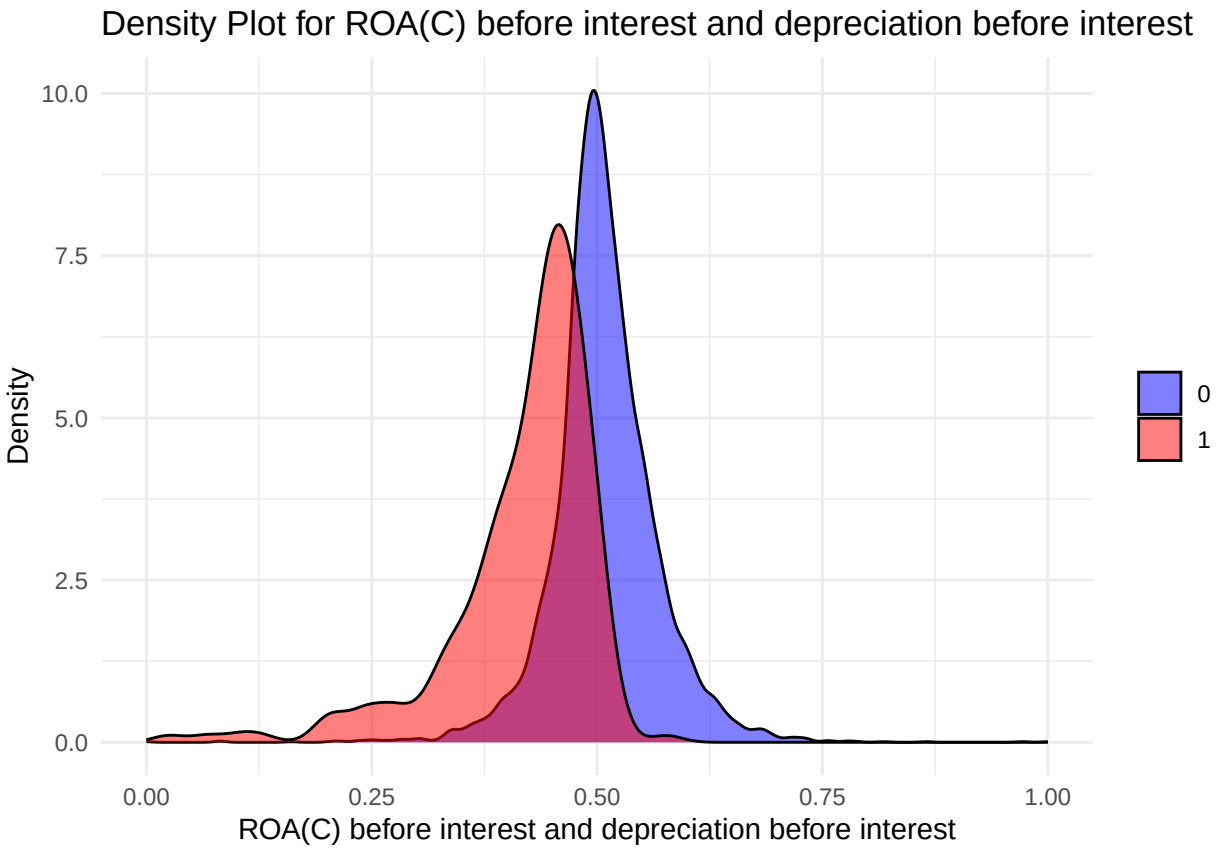


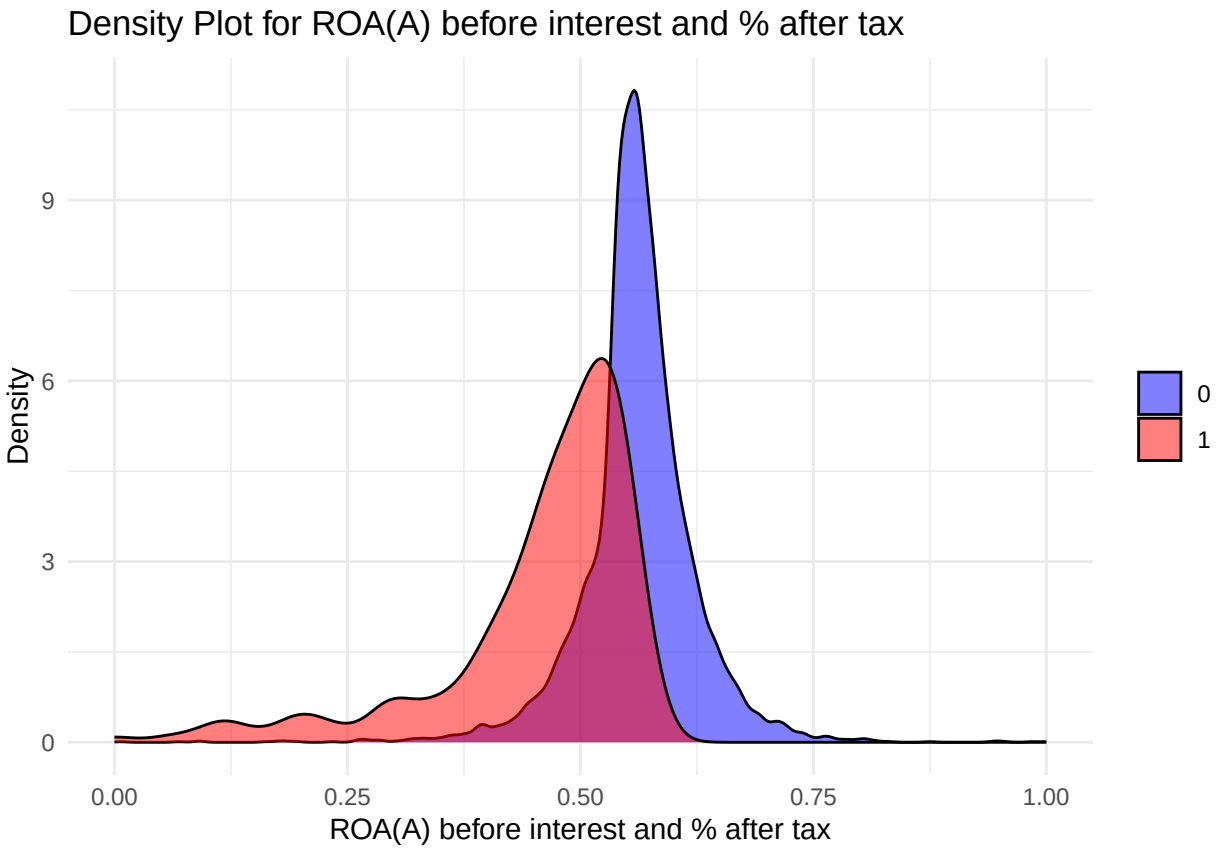
The histogram only confirmed that class 1 is highly imbalanced across all predictors.

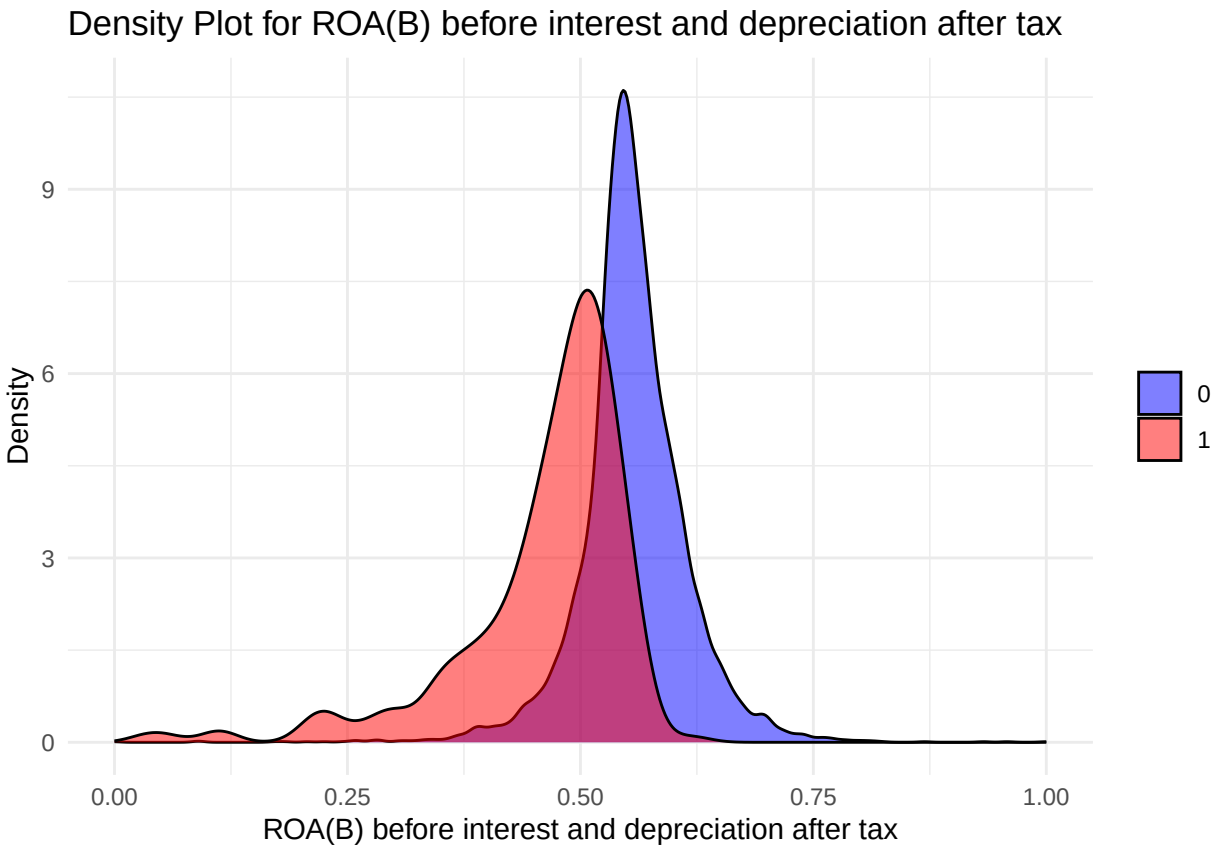
Density

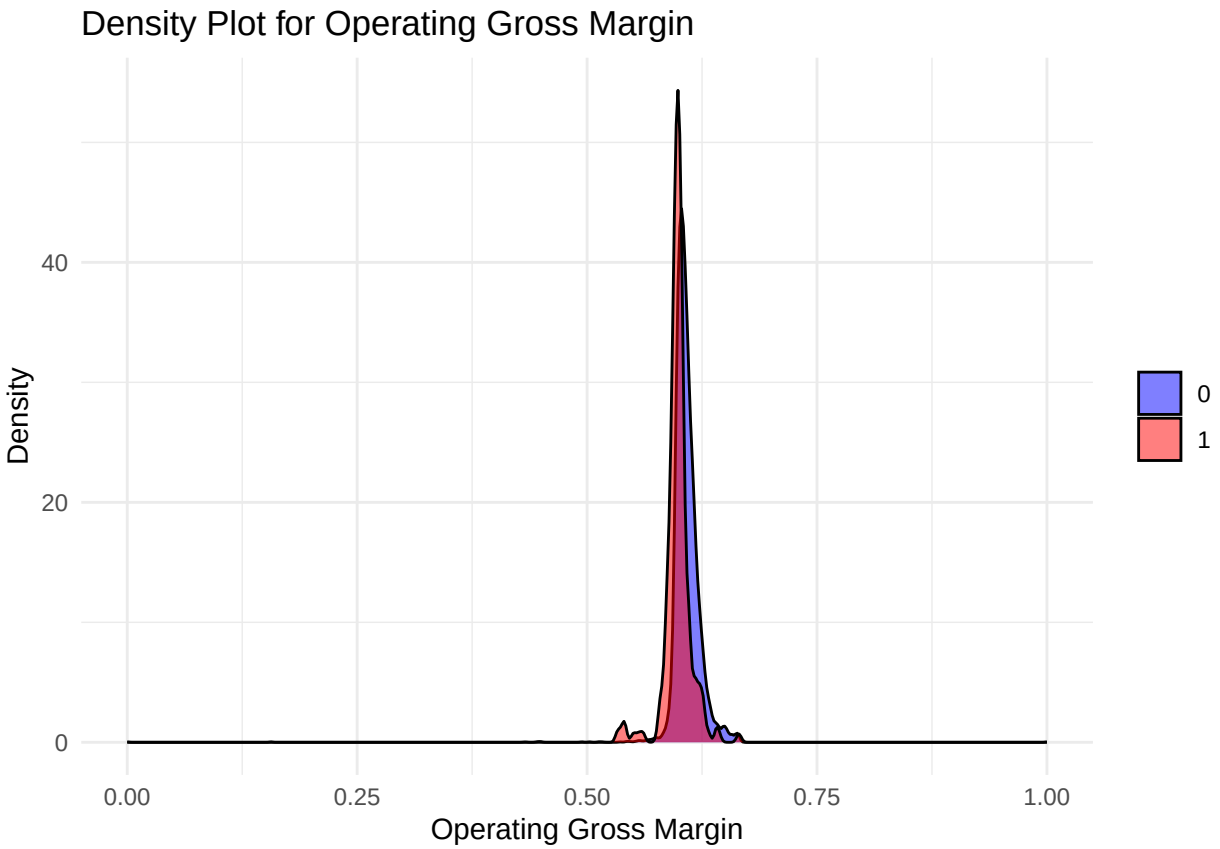
Now we can replace histogram with **Density Plot** to better catch the distribution shape for each predictor.

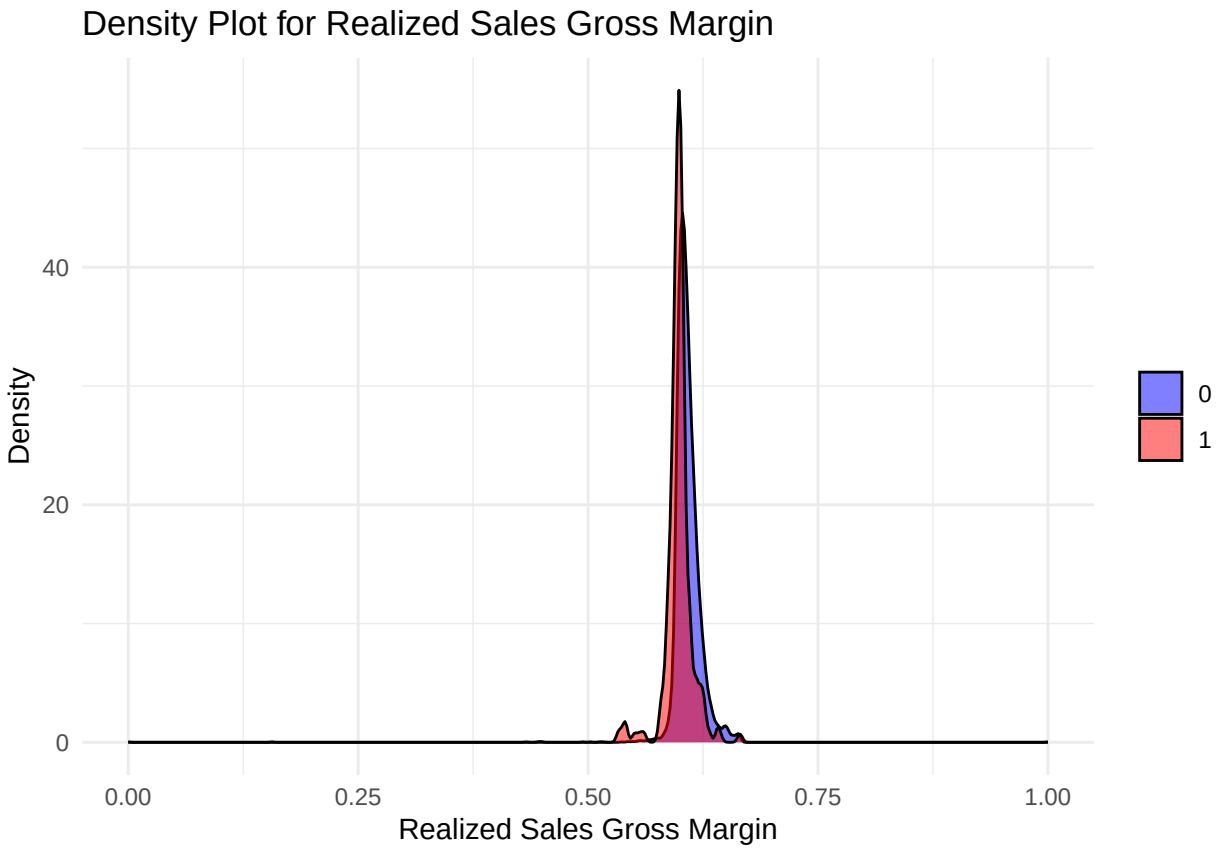
```
show.densityplot <- function(df, cols, target) {

  plot_list <- map(cols, ~ {
    ggplot(data = df, aes(x = .data[[.x]], fill = as.factor(.data[[target]]))) +
      geom_density(alpha = 0.5) +  # Adjust transparency for overlapping
      scale_fill_manual(values = c('blue', 'red')) +  # Customize fill colors
      labs(title = paste("Density Plot for", .x), x = .x, y = "Density") +
      theme_minimal() +
      theme(legend.title = element_blank())
  })

  walk(plot_list, print)
}

# Usage example:
show.densityplot(selected_data, PREDICTOR_NAMES, TARGET)
```

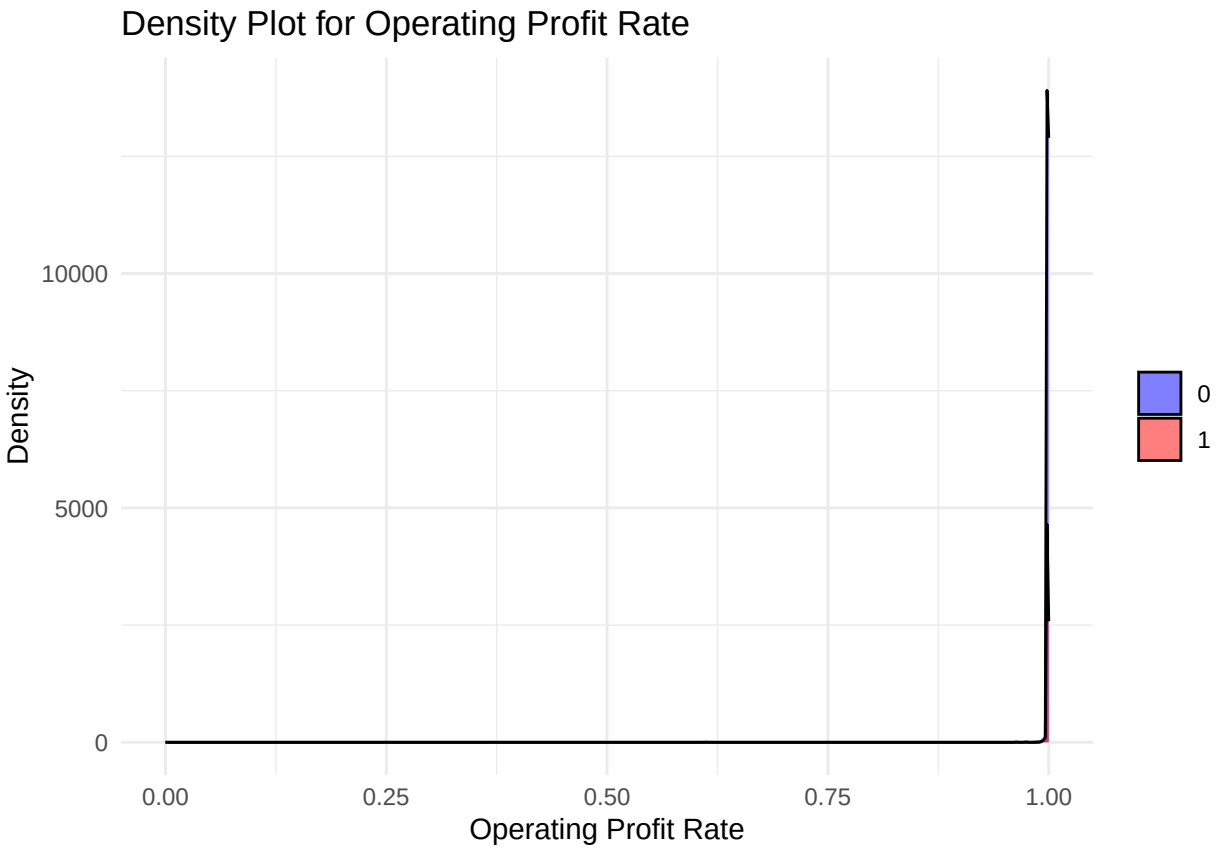

Density Plot for ROA(C) before interest and depreciation before interest
Density
10.0
7.5
5.0
2.5
0.0
0.00
0.25
0.50
0.75
1.00
ROA(C) before interest and depreciation before interest
0
1

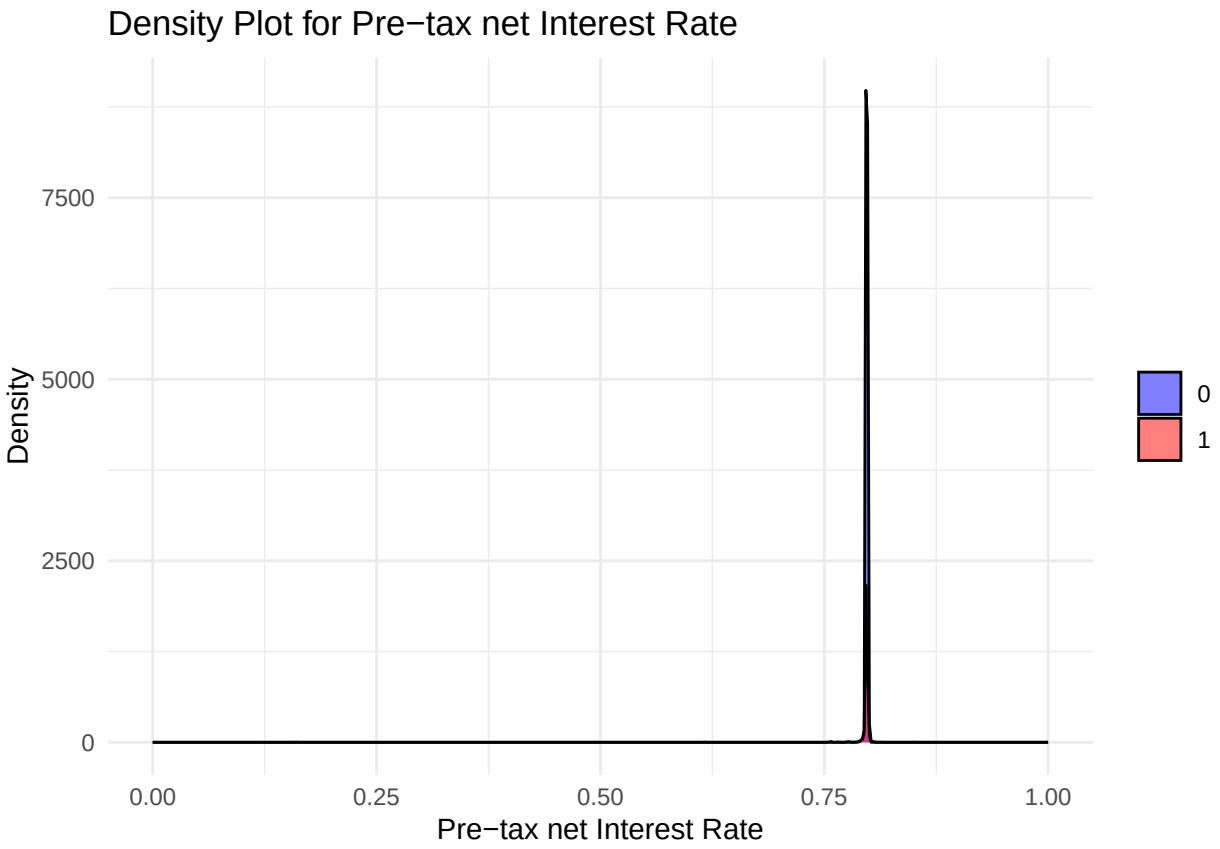

Density Plot for ROA(A) before interest and % after tax
Density
9
6
3
0
0.00
0.25
0.50
0.75
1.00
ROA(A) before interest and % after tax
0
1

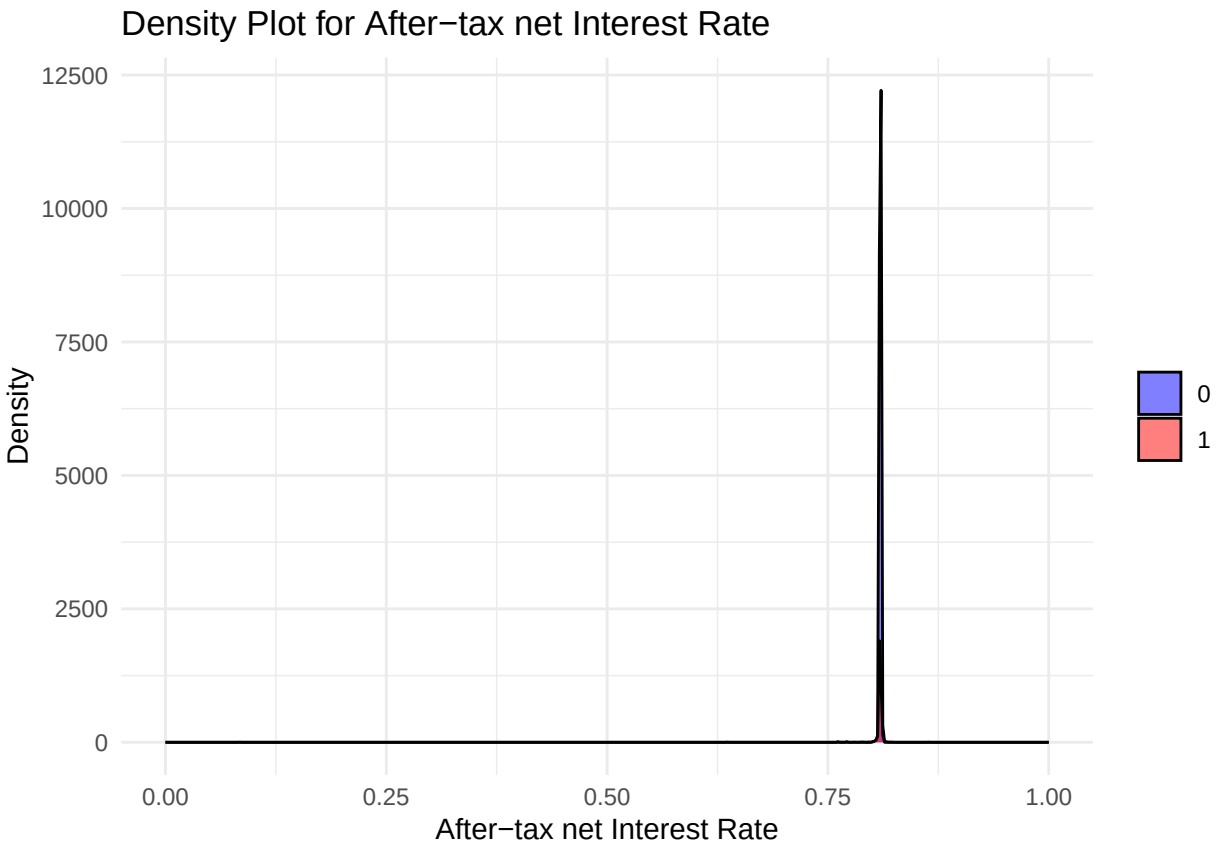

Density Plot for ROA(B) before interest and depreciation after tax
Density
9
6
3
0
0.00
0.25
0.50
0.75
1.00
ROA(B) before interest and depreciation after tax
0
1

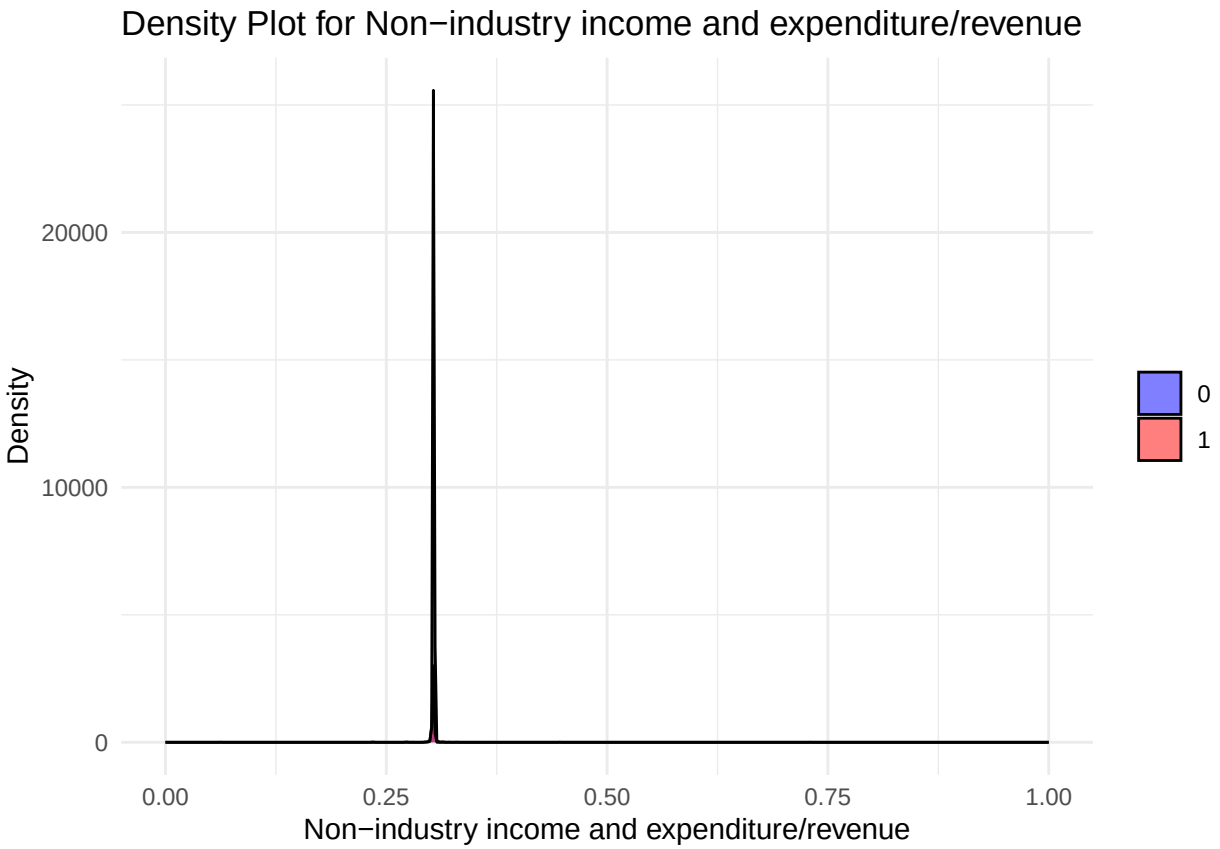

Density Plot for Operating Gross Margin
Density
40
20
0
0.00
0.25
0.50
0.75
1.00
Operating Gross Margin
0
1

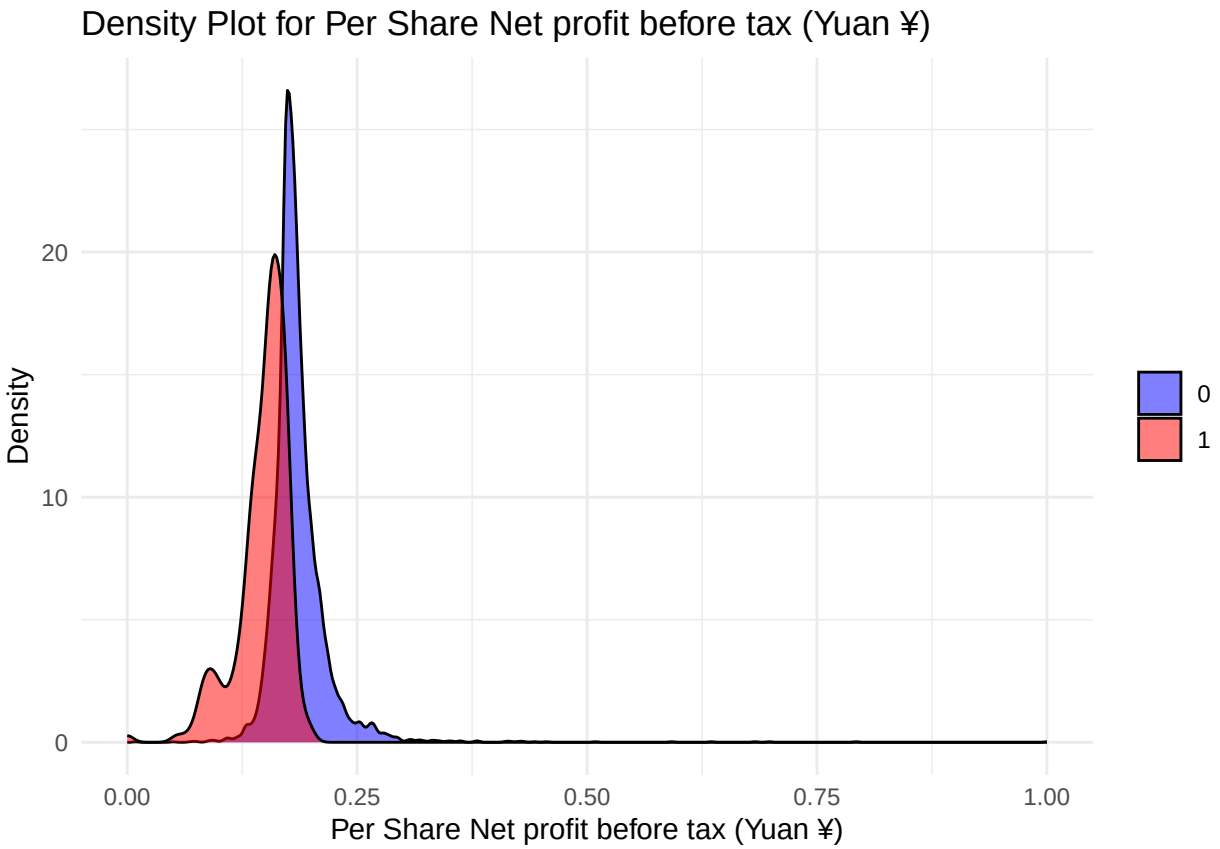

Density Plot for Realized Sales Gross Margin
Density
40
20
0
0.00
0.25
0.50
0.75
1.00
Realized Sales Gross Margin
0
1

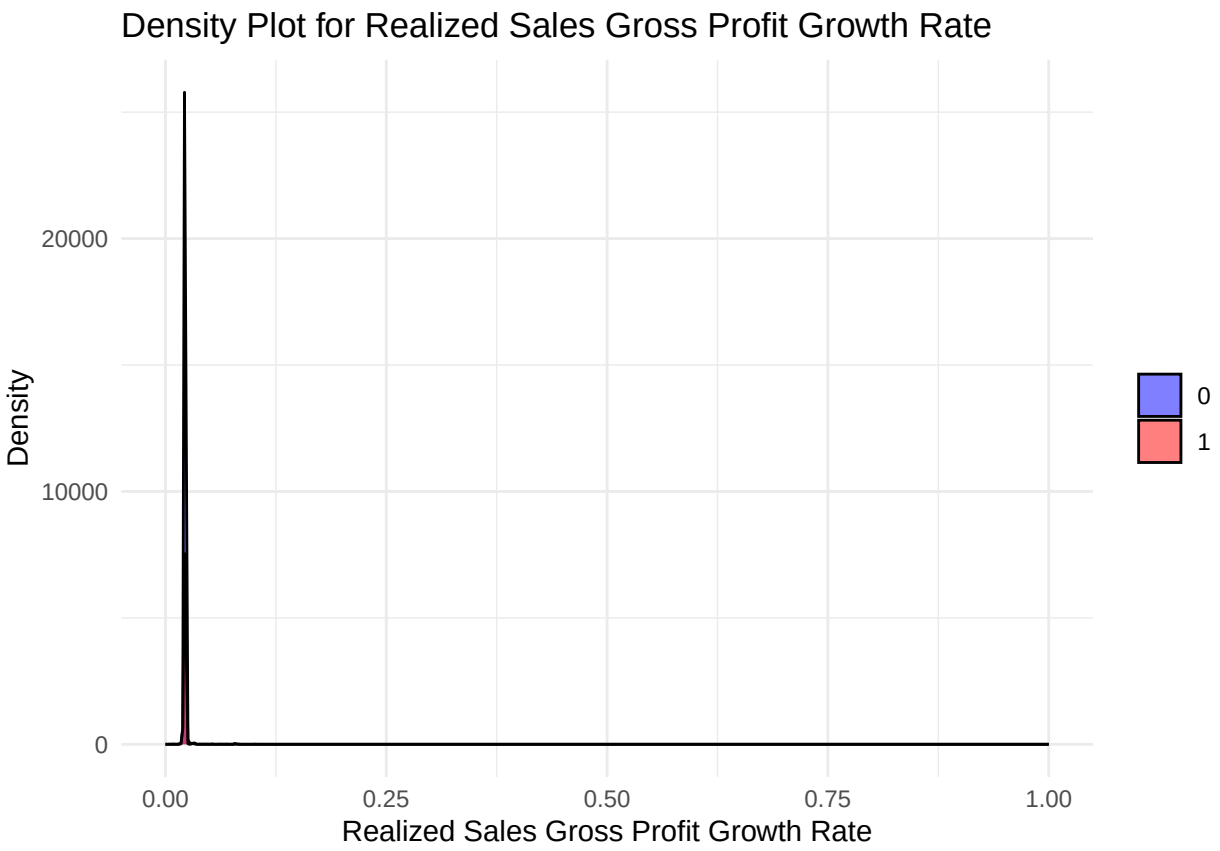

Density Plot for Operating Profit Rate
Density
10000
5000
0
0.00
0.25
0.50
0.75
1.00
Operating Profit Rate
0
1

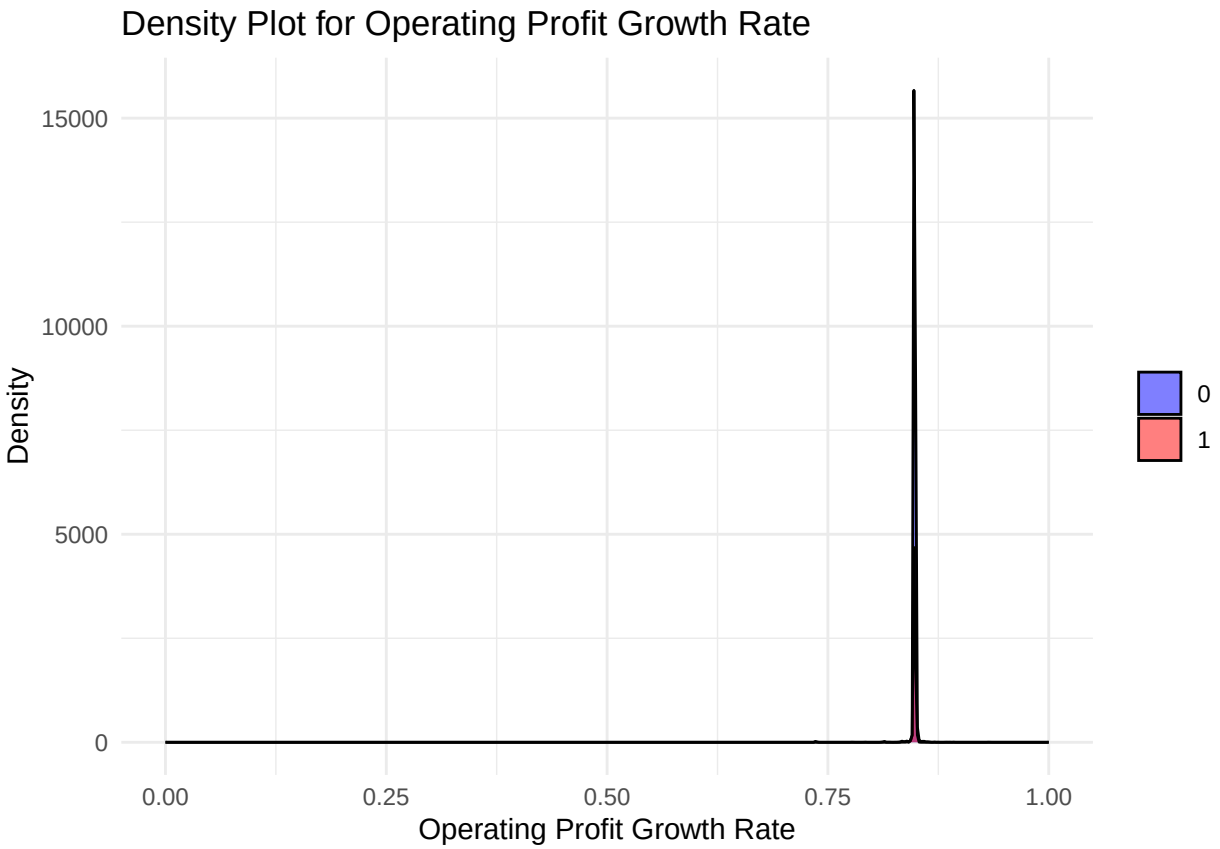

Density Plot for Pre−tax net Interest Rate
Density
7500
5000
2500
0
0.00
0.25
0.50
0.75
1.00
Pre−tax net Interest Rate
0
1

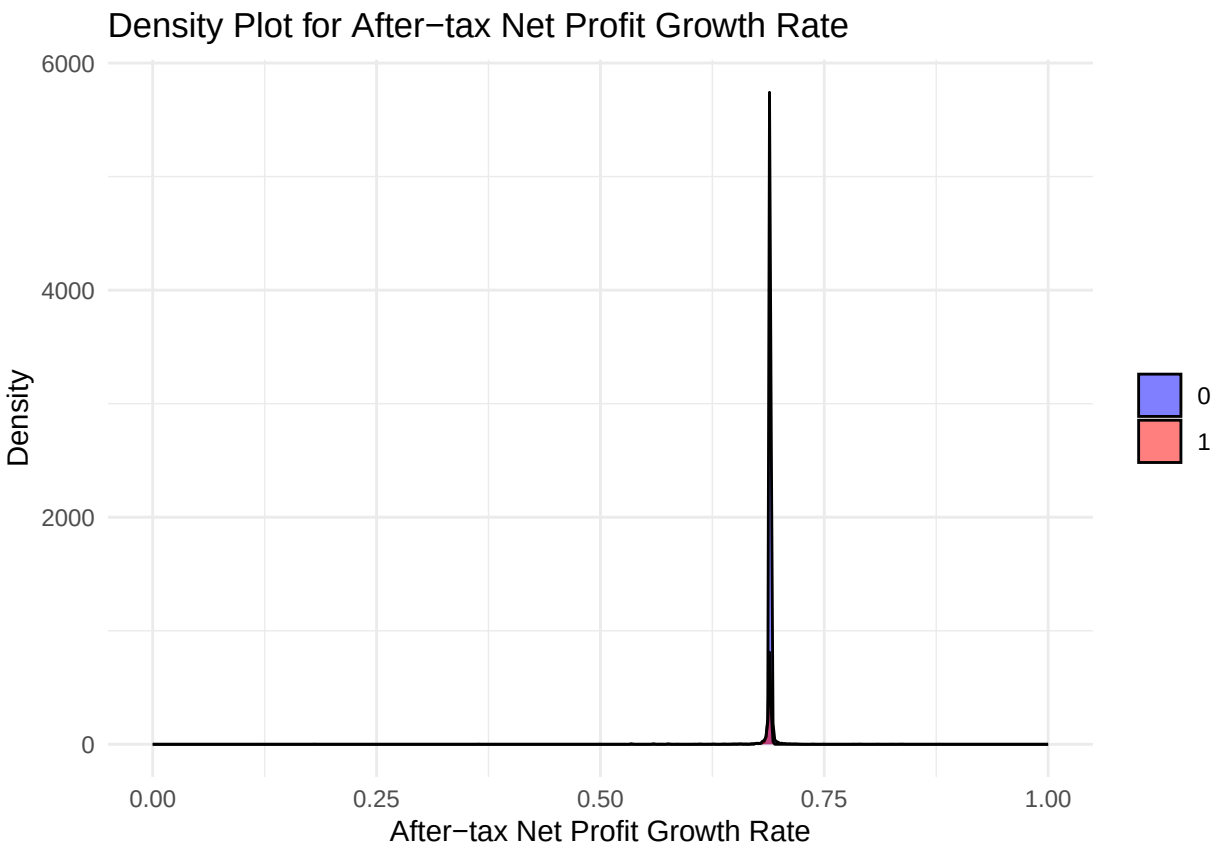

Density Plot for After−tax net Interest Rate
12500
10000
7500
5000
2500
0
Density
0.00
0.25
0.50
0.75
1.00
After−tax net Interest Rate
0
1

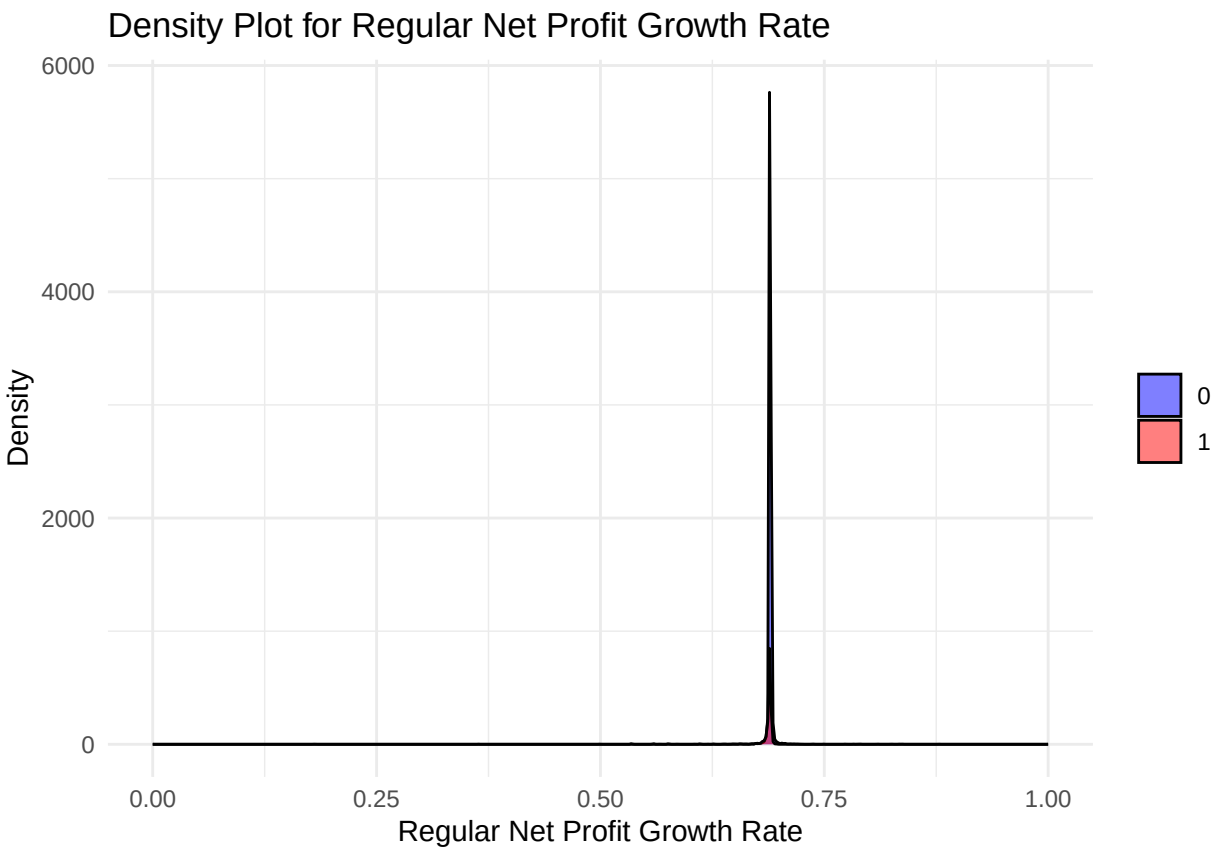

Density Plot for Non−industry income and expenditure/revenue
Density
20000
10000
0
0.00
0.25
0.50
0.75
1.00
Non−industry income and expenditure/revenue
0
1

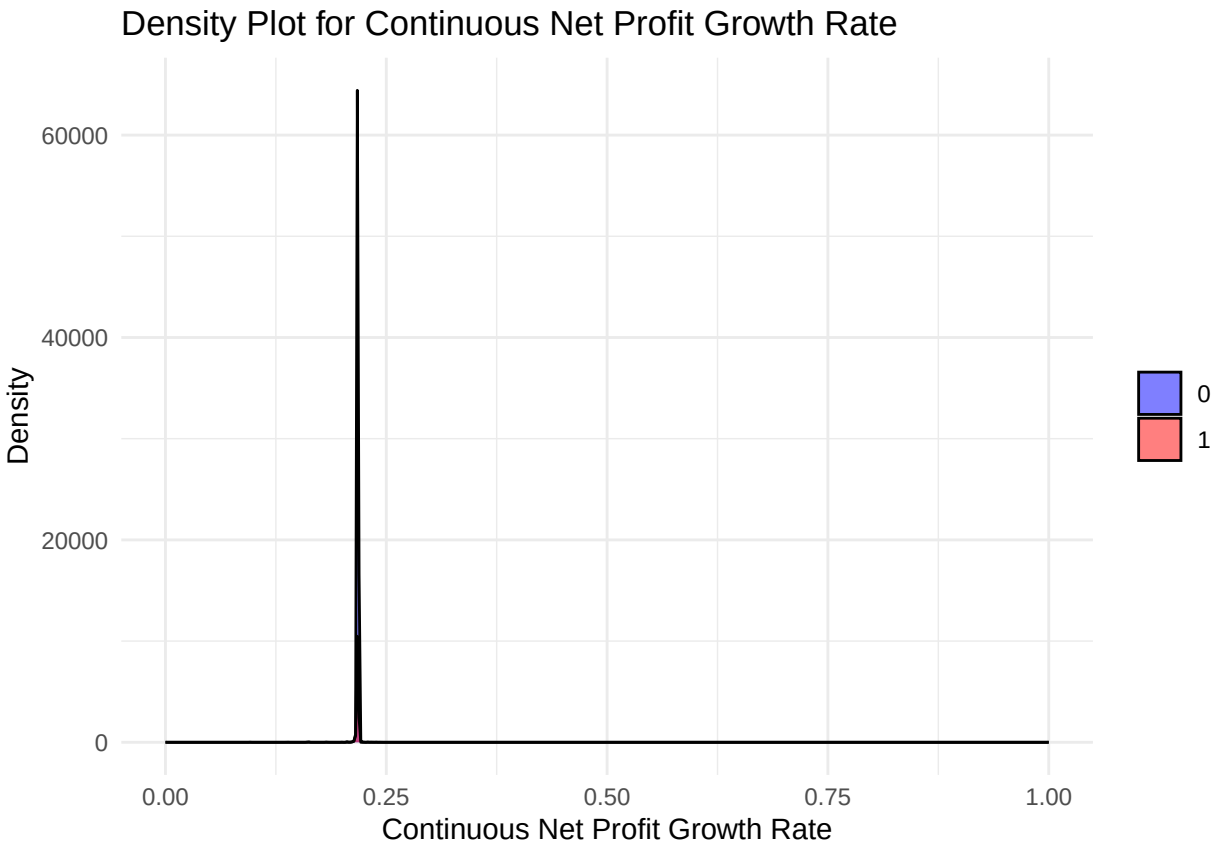

Density Plot for Per Share Net profit before tax (Yuan ¥)
Density
20
10
0
0.00
0.25
0.50
0.75
1.00
Per Share Net profit before tax (Yuan ¥)
0
1

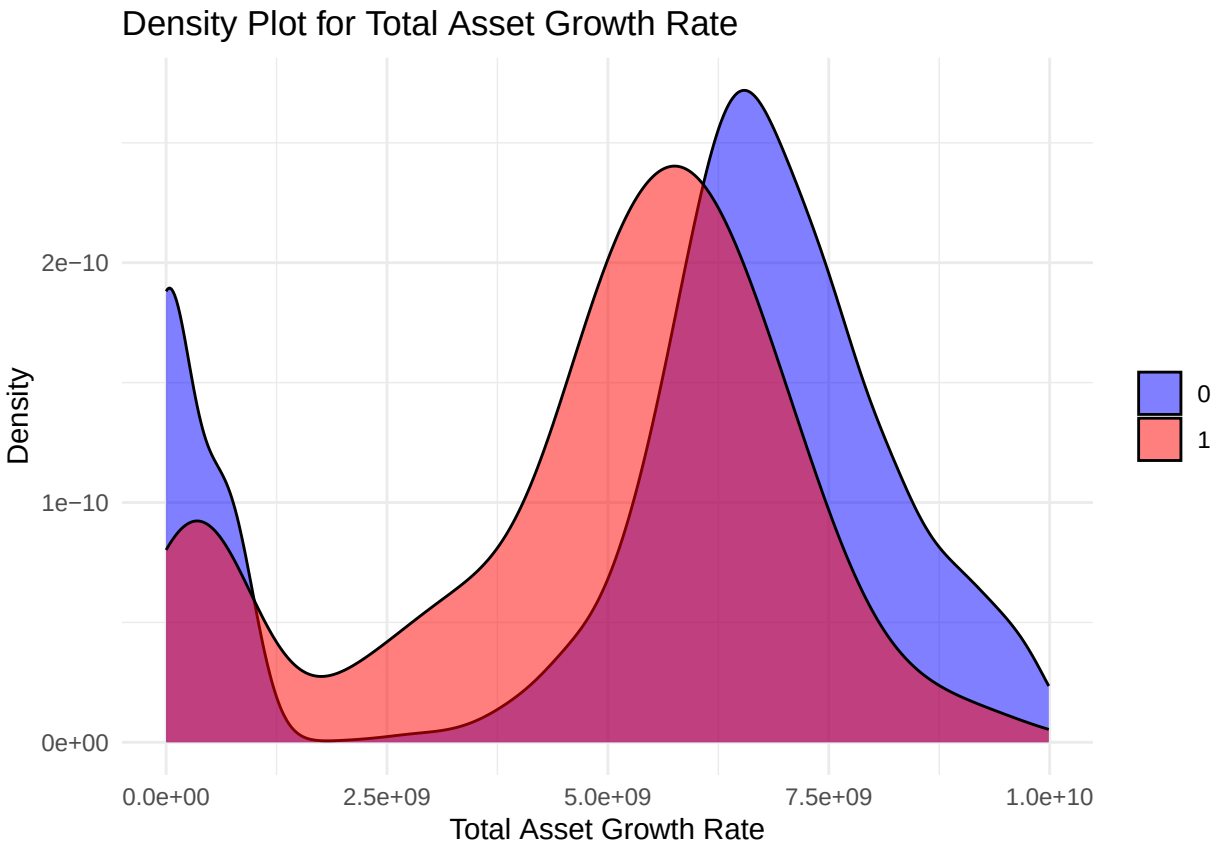

Density Plot for Realized Sales Gross Profit Growth Rate
Density
20000
10000
0
0.00
0.25
0.50
0.75
1.00
Realized Sales Gross Profit Growth Rate
0
1

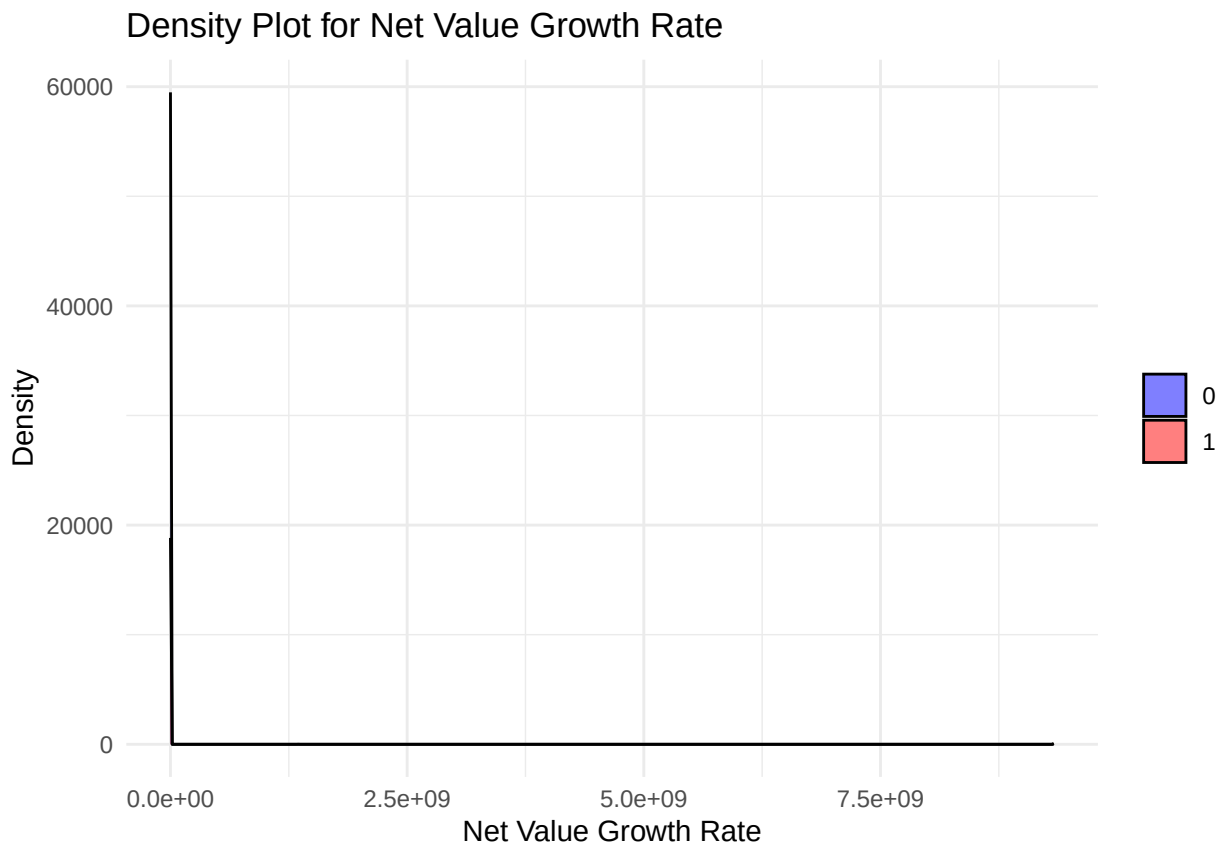

Density Plot for Operating Profit Growth Rate
15000
10000
5000
0
Density
0.00
0.25
0.50
0.75
1.00
Operating Profit Growth Rate
0
1

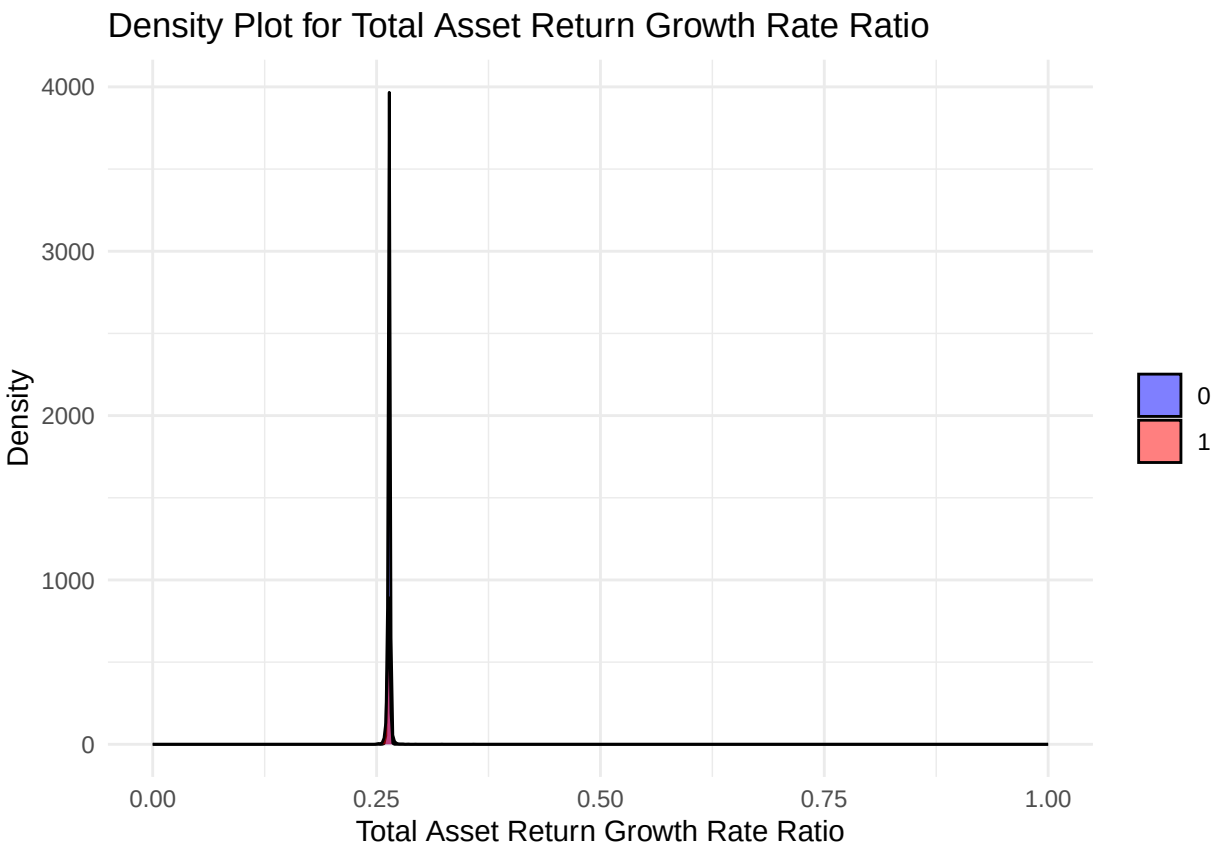

Density Plot for After−tax Net Profit Growth Rate
6000
4000
2000
0
Density
0.00
0.25
0.50
0.75
1.00
After−tax Net Profit Growth Rate
0
1


Density Plot for Regular Net Profit Growth Rate
6000
4000
2000
0
Density
0.00
0.25
0.50
0.75
1.00
Regular Net Profit Growth Rate
0
1


Density Plot for Continuous Net Profit Growth Rate
60000
40000
20000
0
Density
0.00
0.25
0.50
0.75
1.00
Continuous Net Profit Growth Rate
0
1


Density Plot for Total Asset Growth Rate
Density
2e−10
1e−10
0e+00
0.0e+00
2.5e+09
5.0e+09
7.5e+09
1.0e+10
Total Asset Growth Rate
0
1

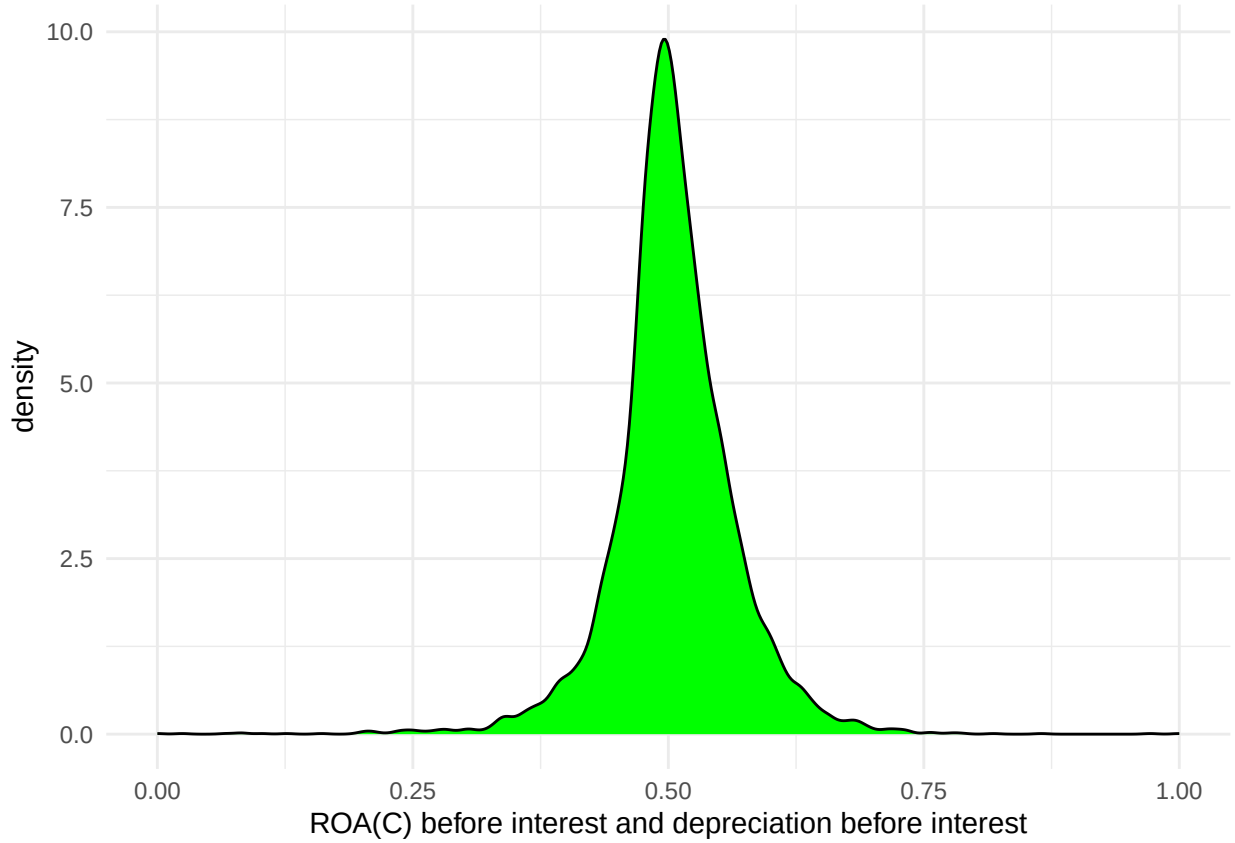

Density Plot for Net Value Growth Rate
60000
40000
20000
0
Density
0.0e+00
2.5e+09
5.0e+09
7.5e+09
Net Value Growth Rate
0
1

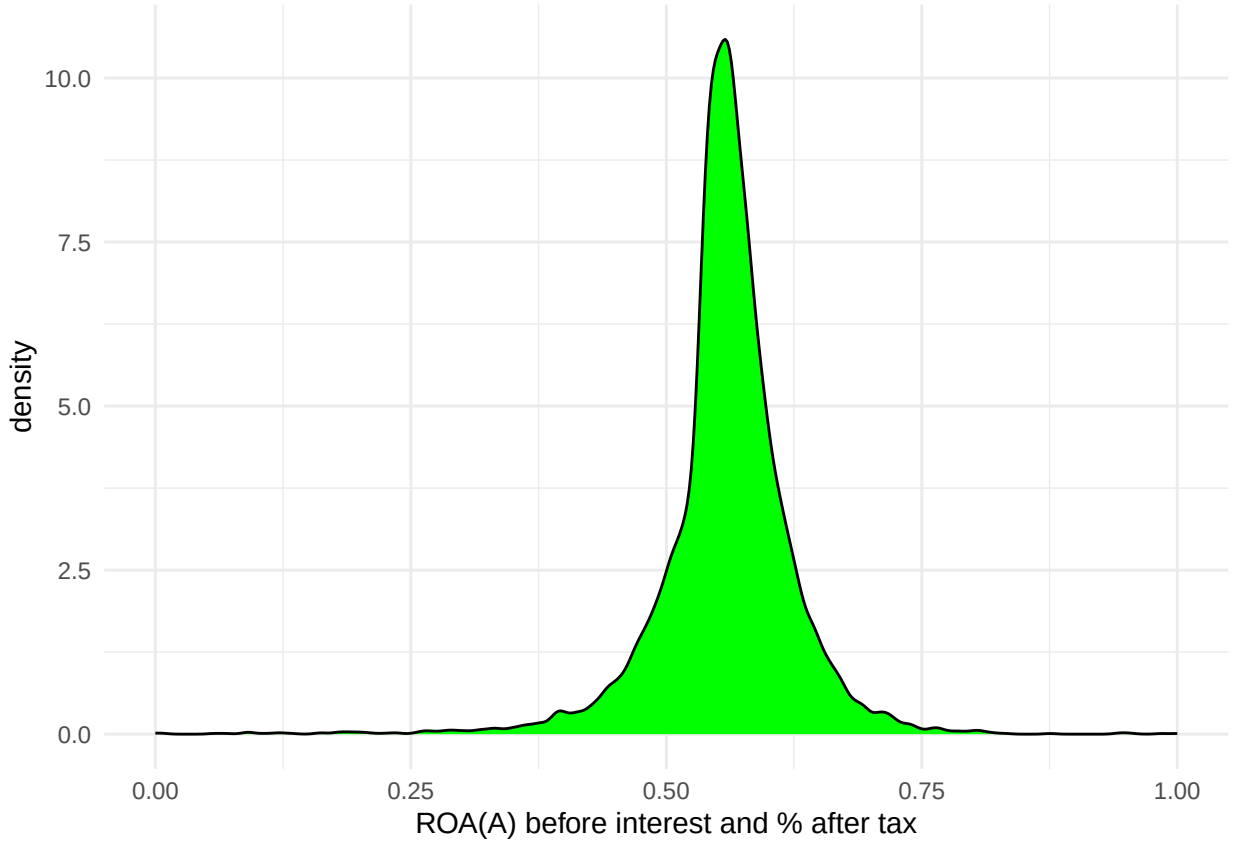


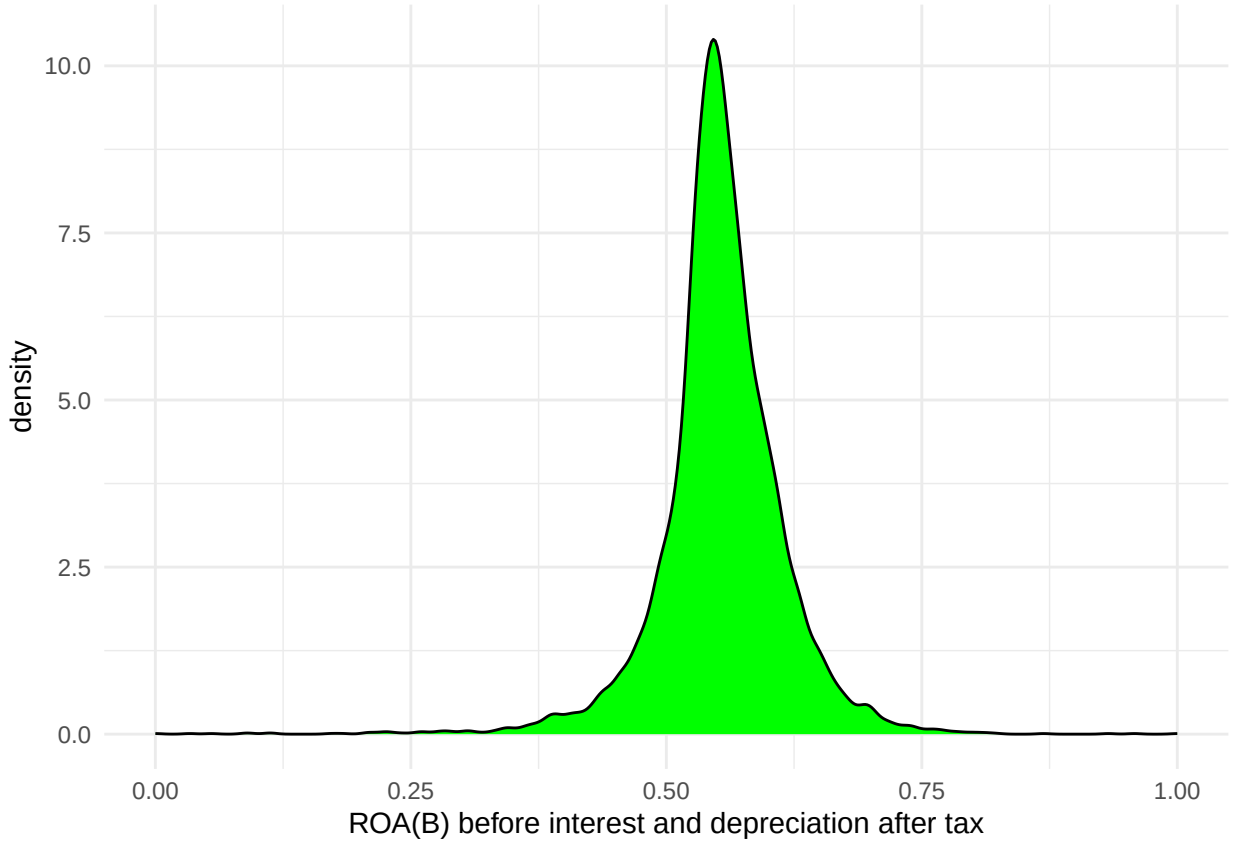
Normality

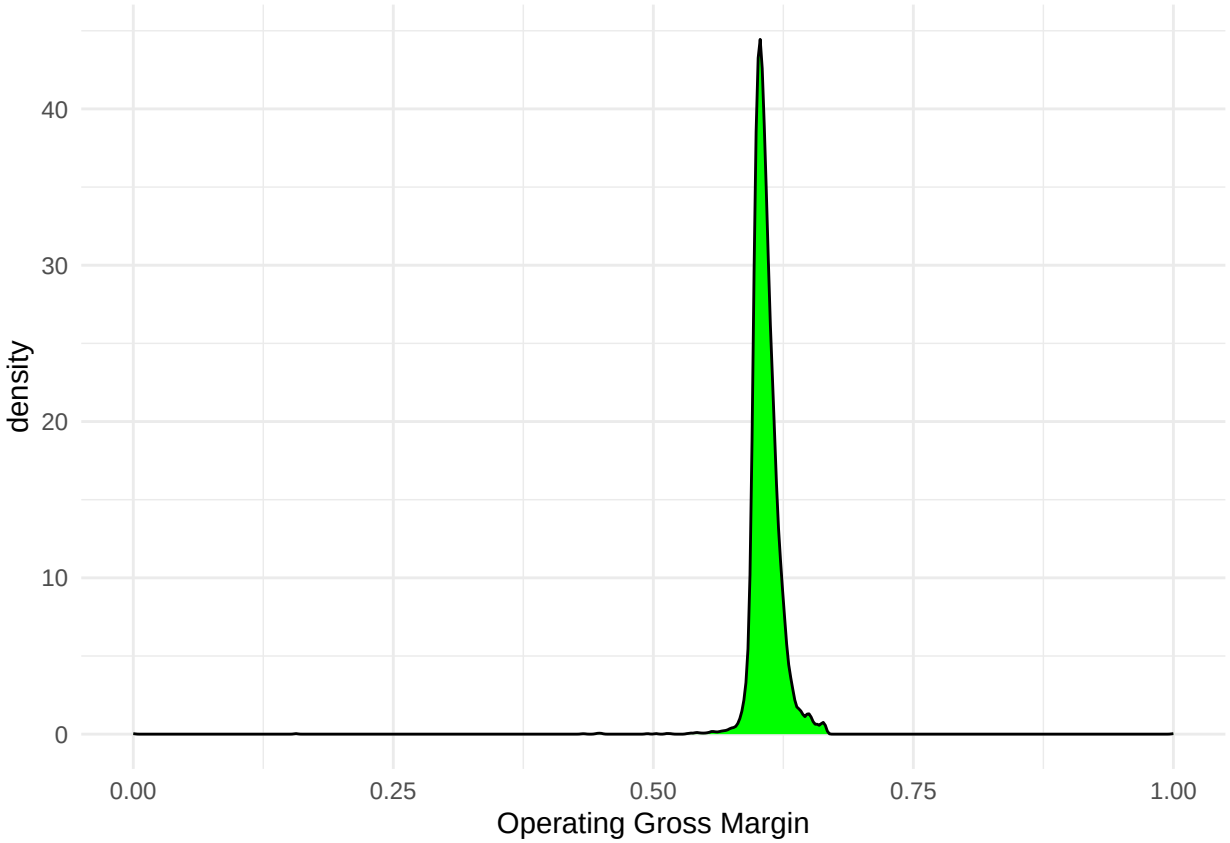
We will also apply the density plot to **visualize density for each predictor**.

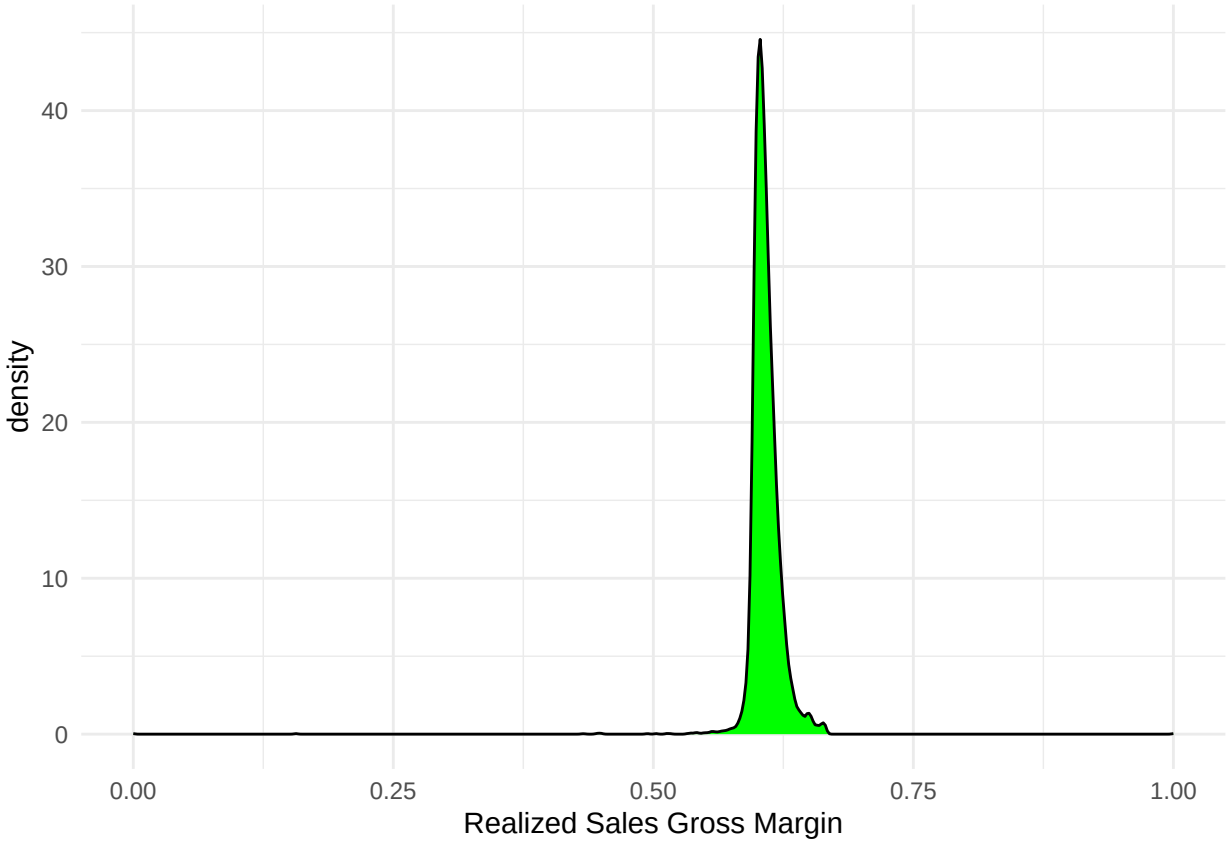
```
show.scatterplot <- function(df, cols){

  plot_list <- map(cols, ~ ggplot(
    data = df,
    mapping = aes(x = .data[[.x]])
  ) + geom_density(fill = 'green') +
    theme_minimal())

  walk(plot_list, print)
}

show.scatterplot(selected_data, PREDICTOR_NAMES)
```

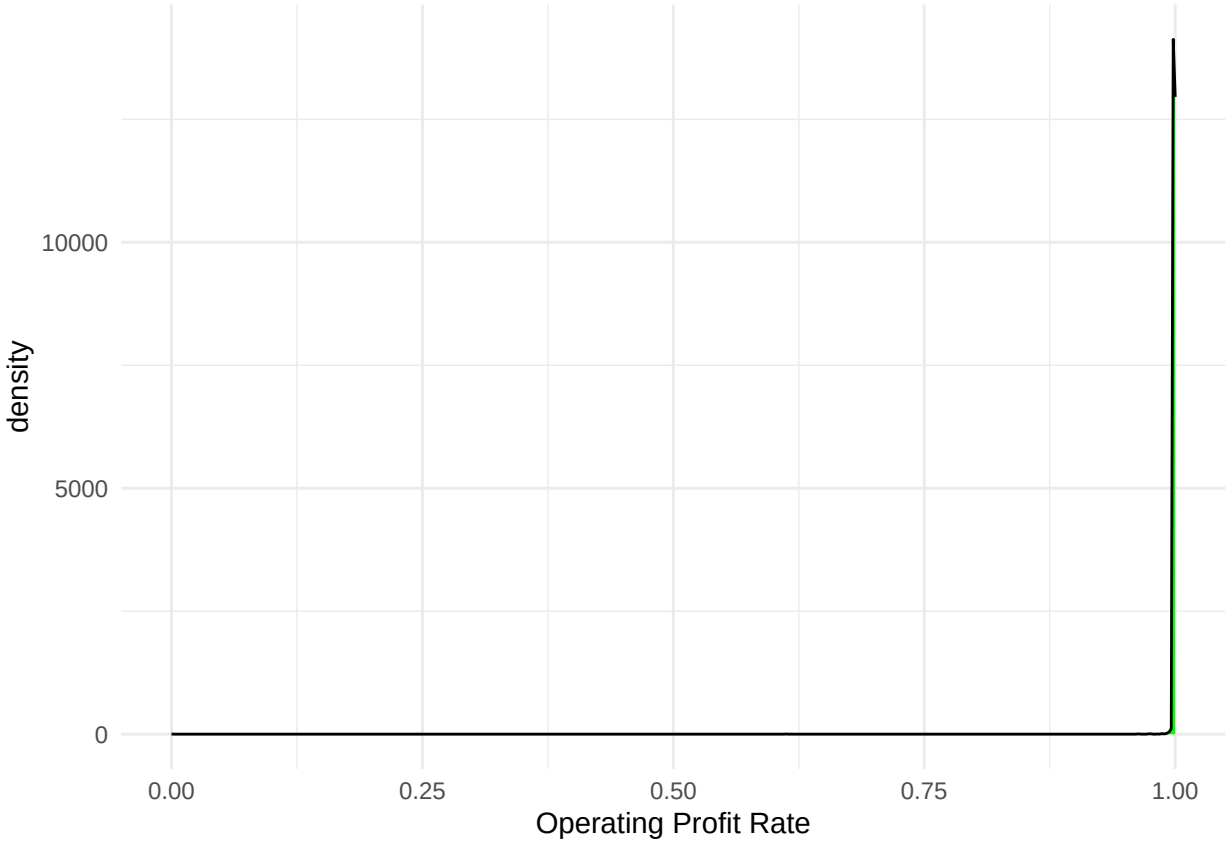

10.0
7.5
5.0
2.5
0.0
density
0.00
0.25
0.50
0.75
1.00
ROA(C) before interest and depreciation before interest

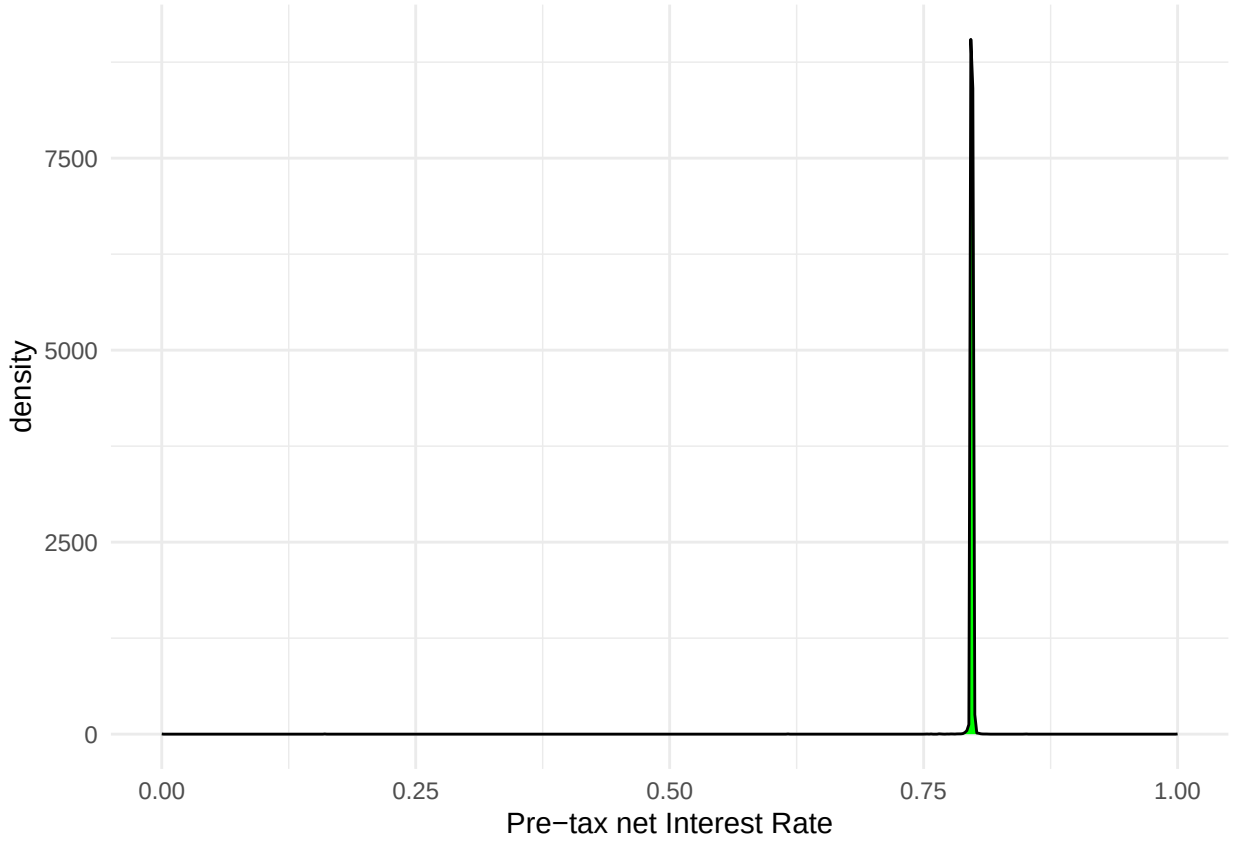

10.0
7.5
5.0
2.5
0.0
density
0.00
0.25
0.50
0.75
1.00
ROA(A) before interest and % after tax

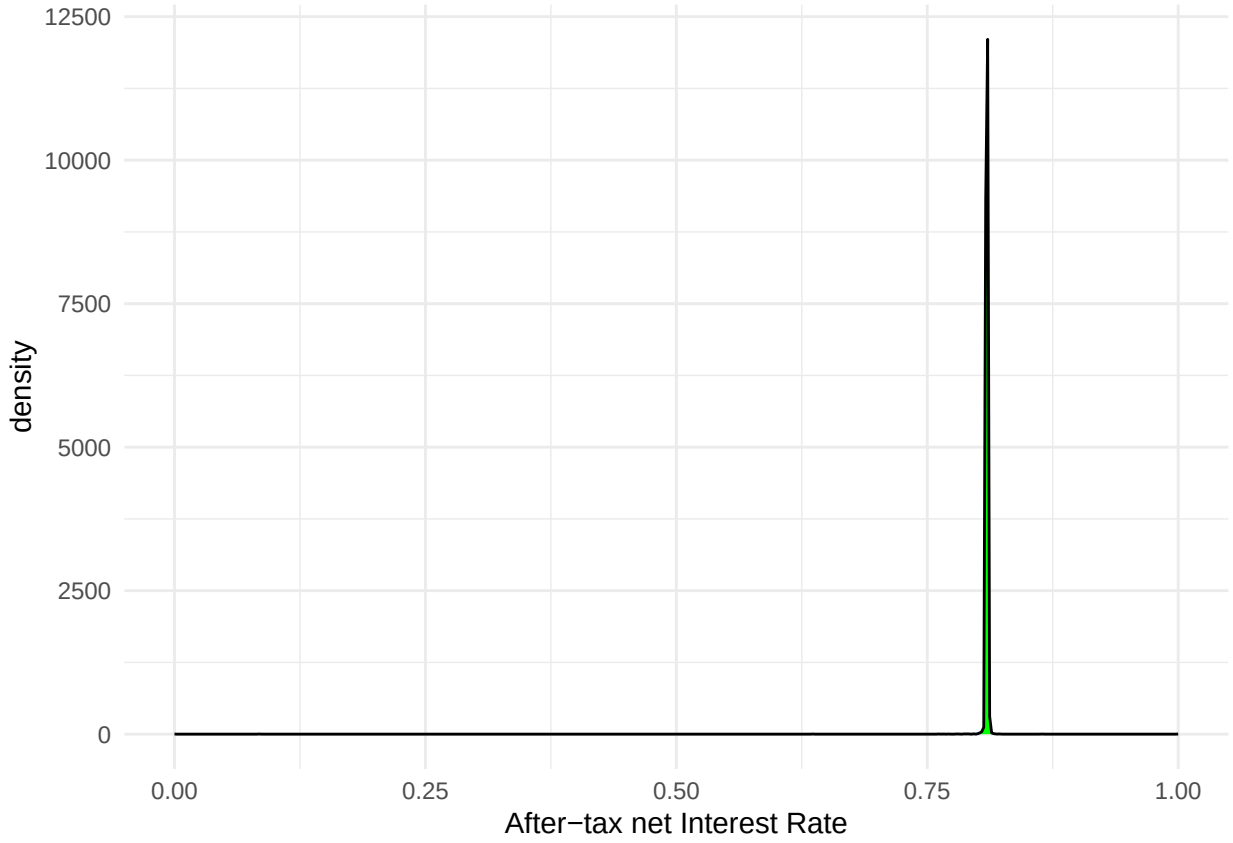

10.0
7.5
5.0
2.5
0.0
density
0.00
0.25
0.50
0.75
1.00
ROA(B) before interest and depreciation after tax

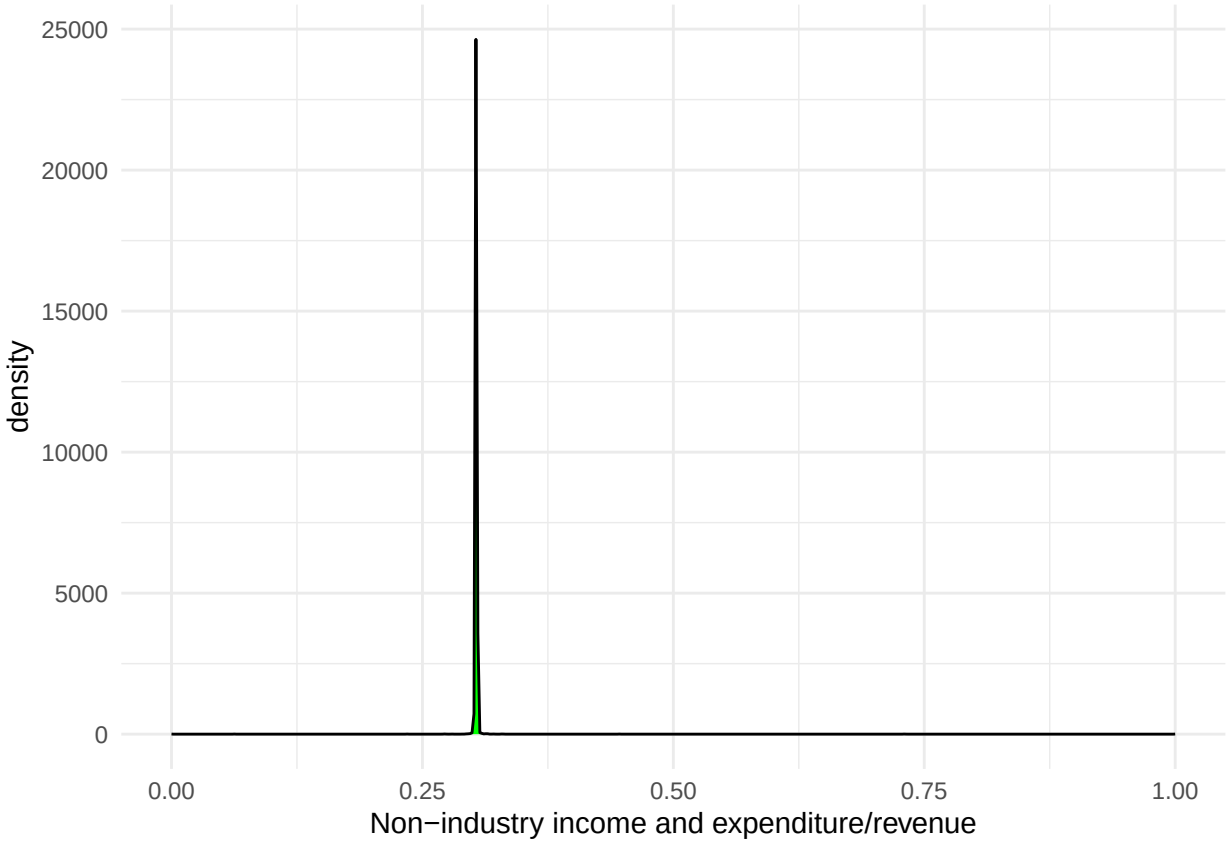

40
30
20
10
0
density
0.00
0.25
0.50
0.75
1.00
Operating Gross Margin

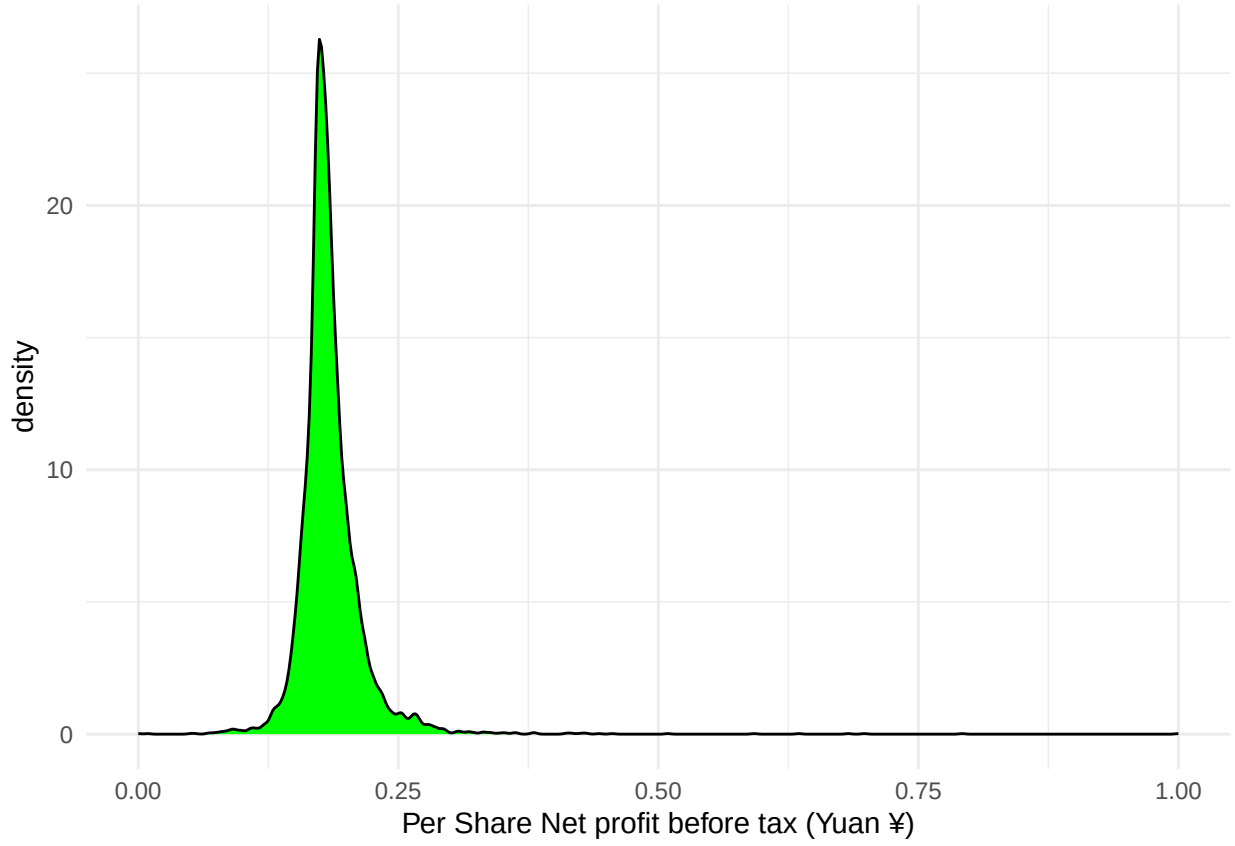

density
40
30
20
10
0
0.00
0.25
0.50
0.75
1.00
Realized Sales Gross Margin

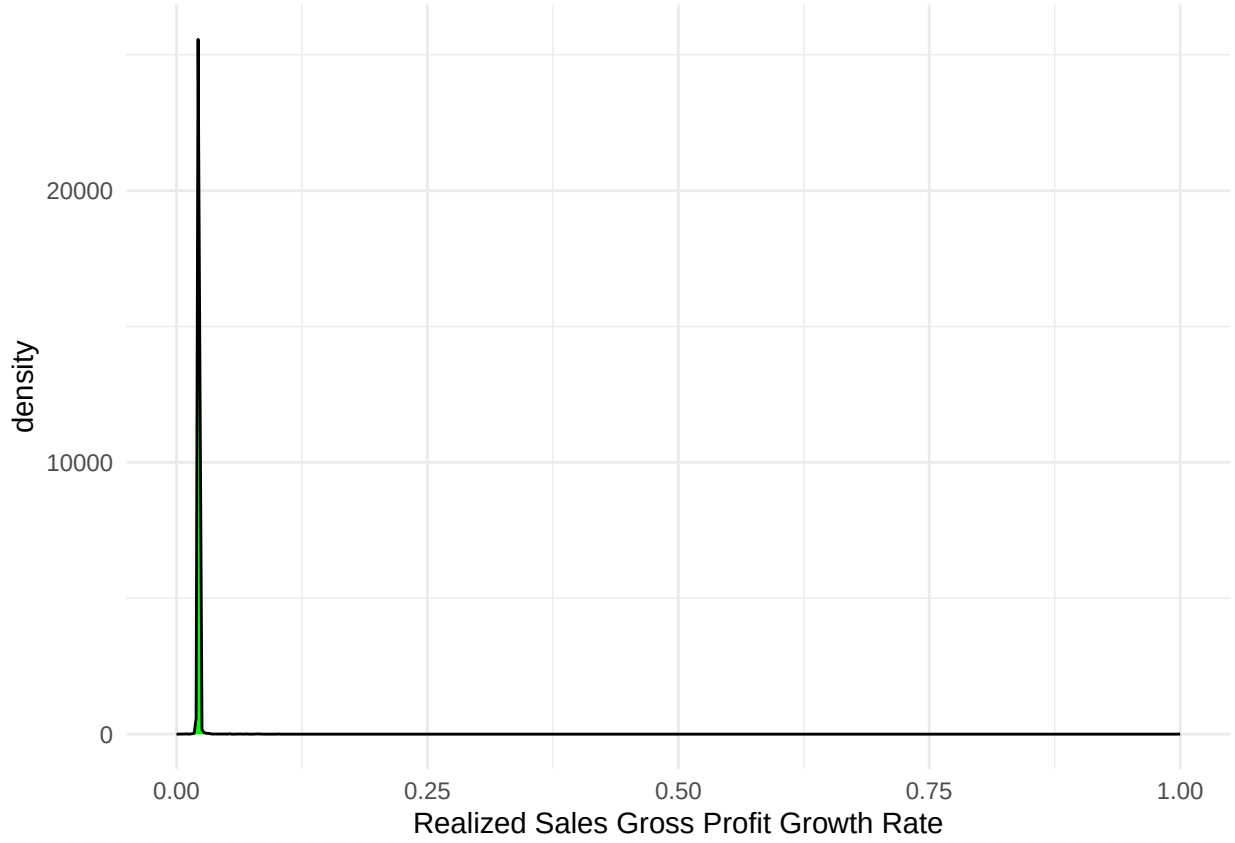

density
10000
5000
0
0.00
0.25
0.50
0.75
1.00
Operating Profit Rate

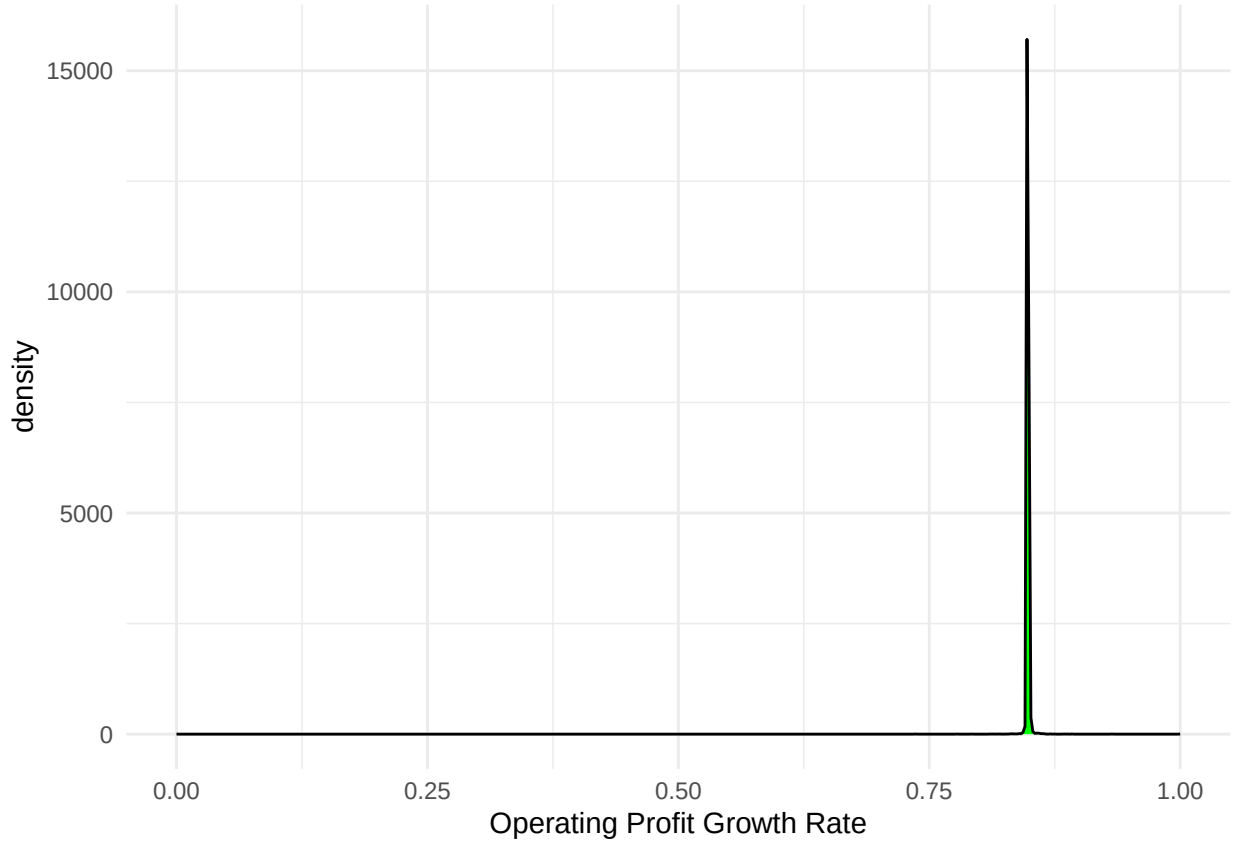

7500
5000
2500
0
density
0.00
0.25
0.50
0.75
1.00
Pre−tax net Interest Rate

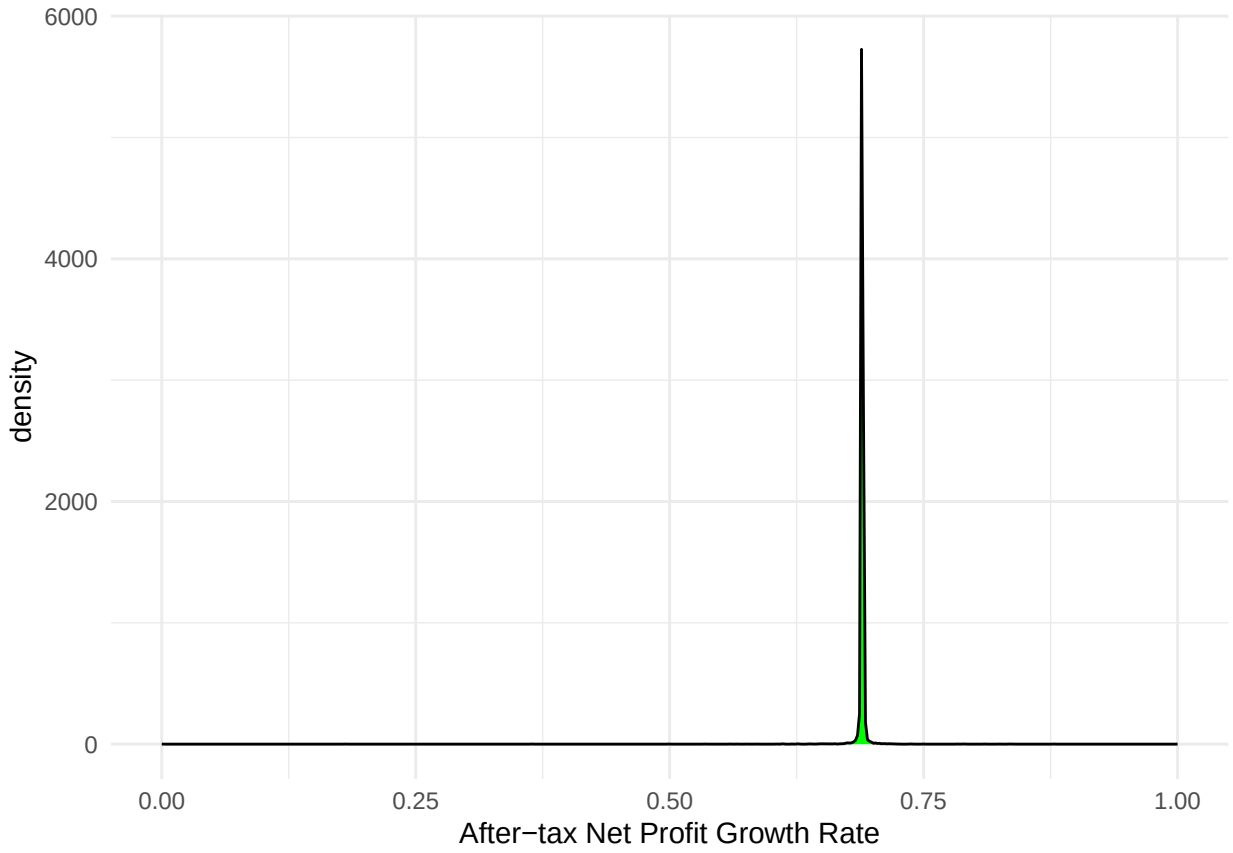

12500
10000
7500
5000
2500
0
density
0.00
0.25
0.50
0.75
1.00
After−tax net Interest Rate

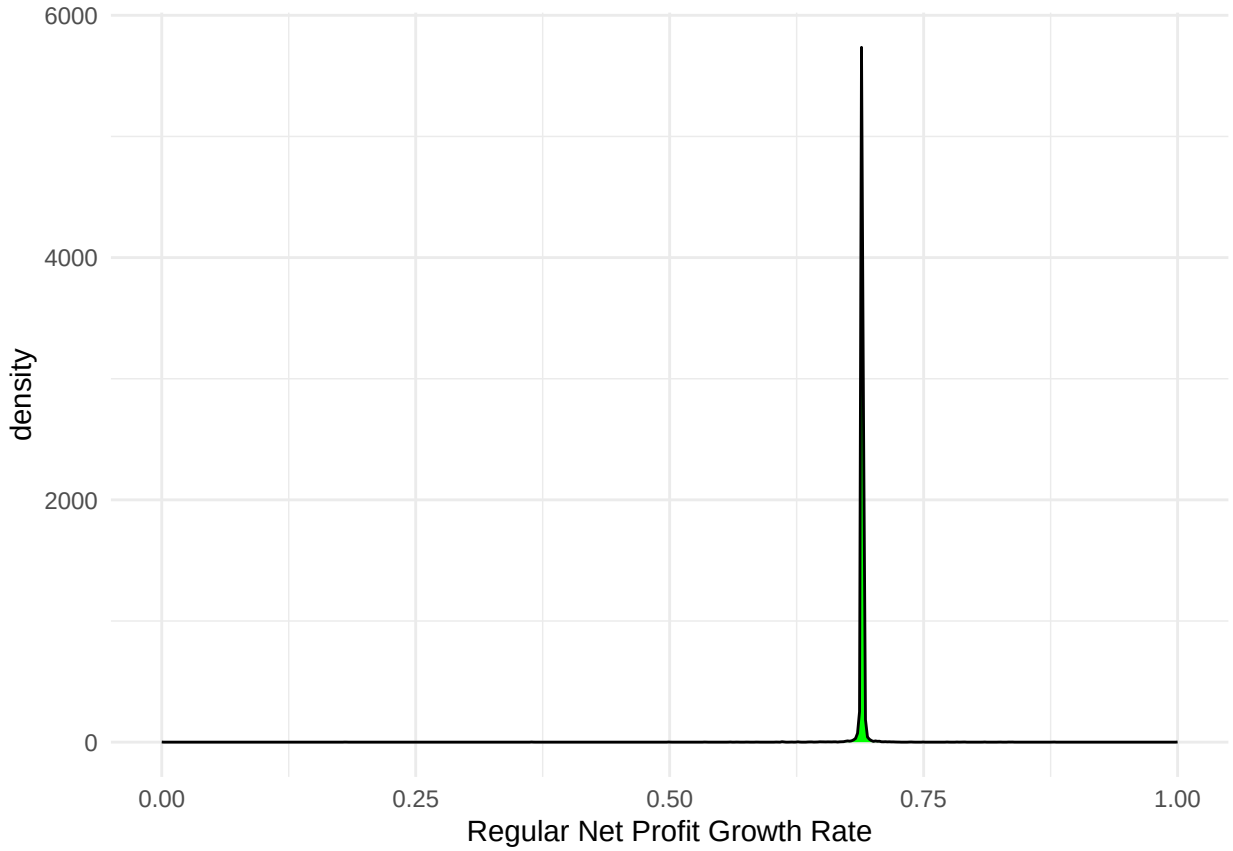

25000
20000
15000
10000
5000
0
density
0.00
0.25
0.50
0.75
1.00
Non−industry income and expenditure/revenue

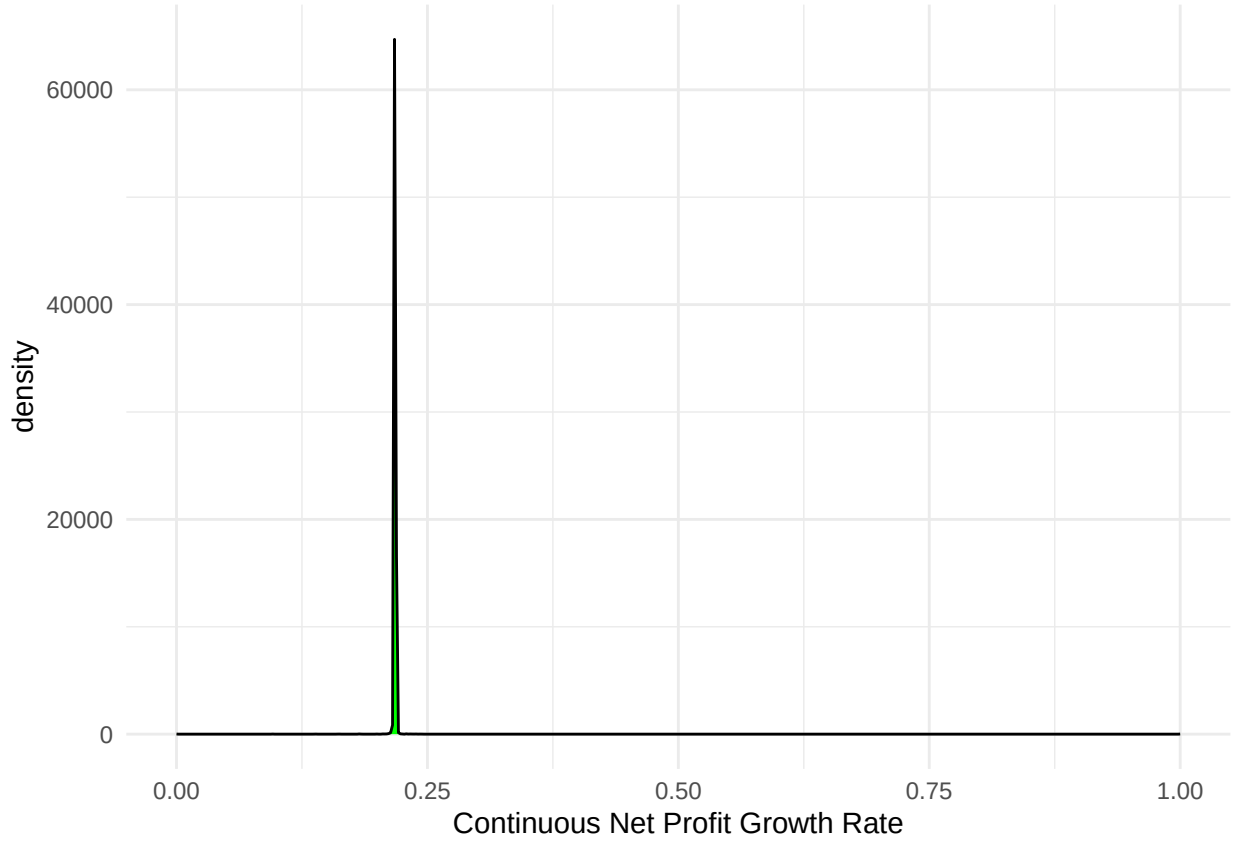

density
20
10
0
0.00
0.25
0.50
0.75
1.00
Per Share Net profit before tax (Yuan ¥)

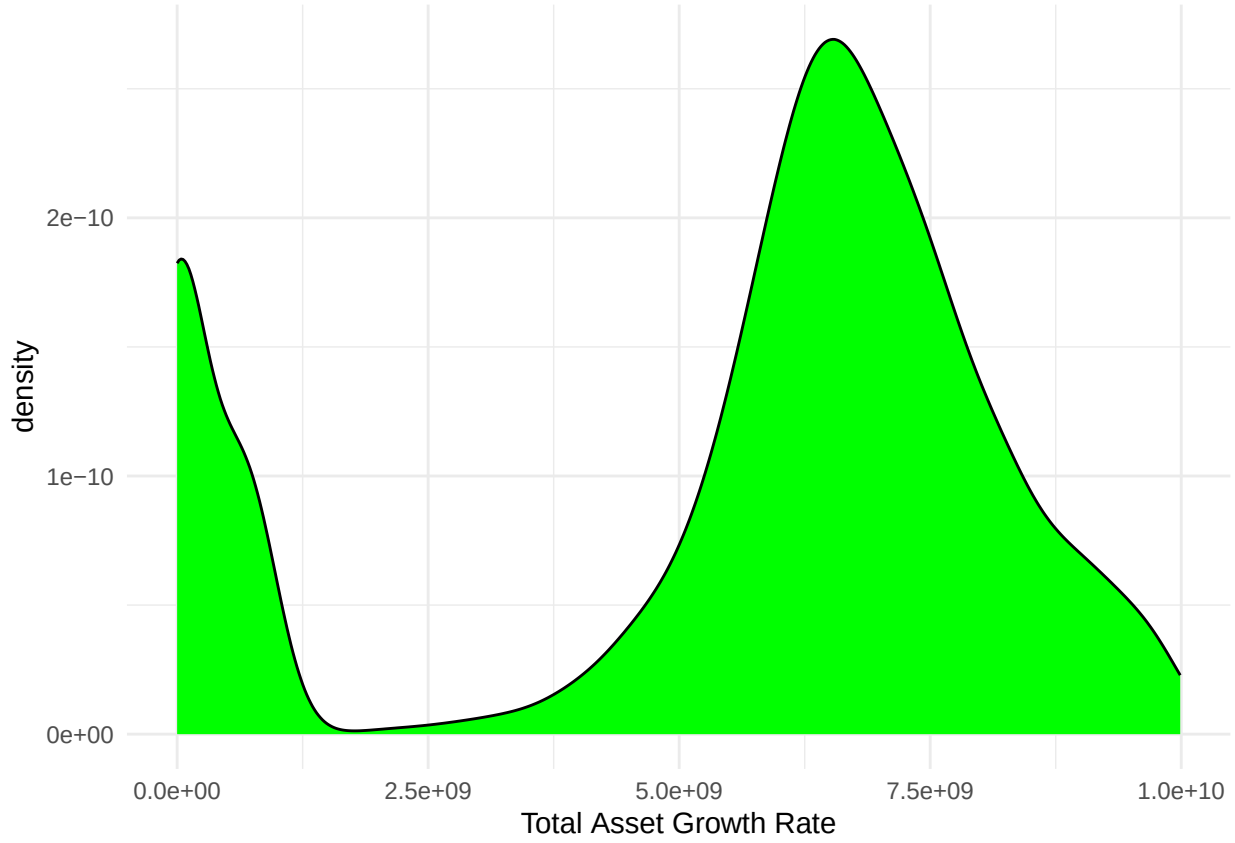

density
20000
10000
0
0.00
0.25
0.50
0.75
1.00
Realized Sales Gross Profit Growth Rate

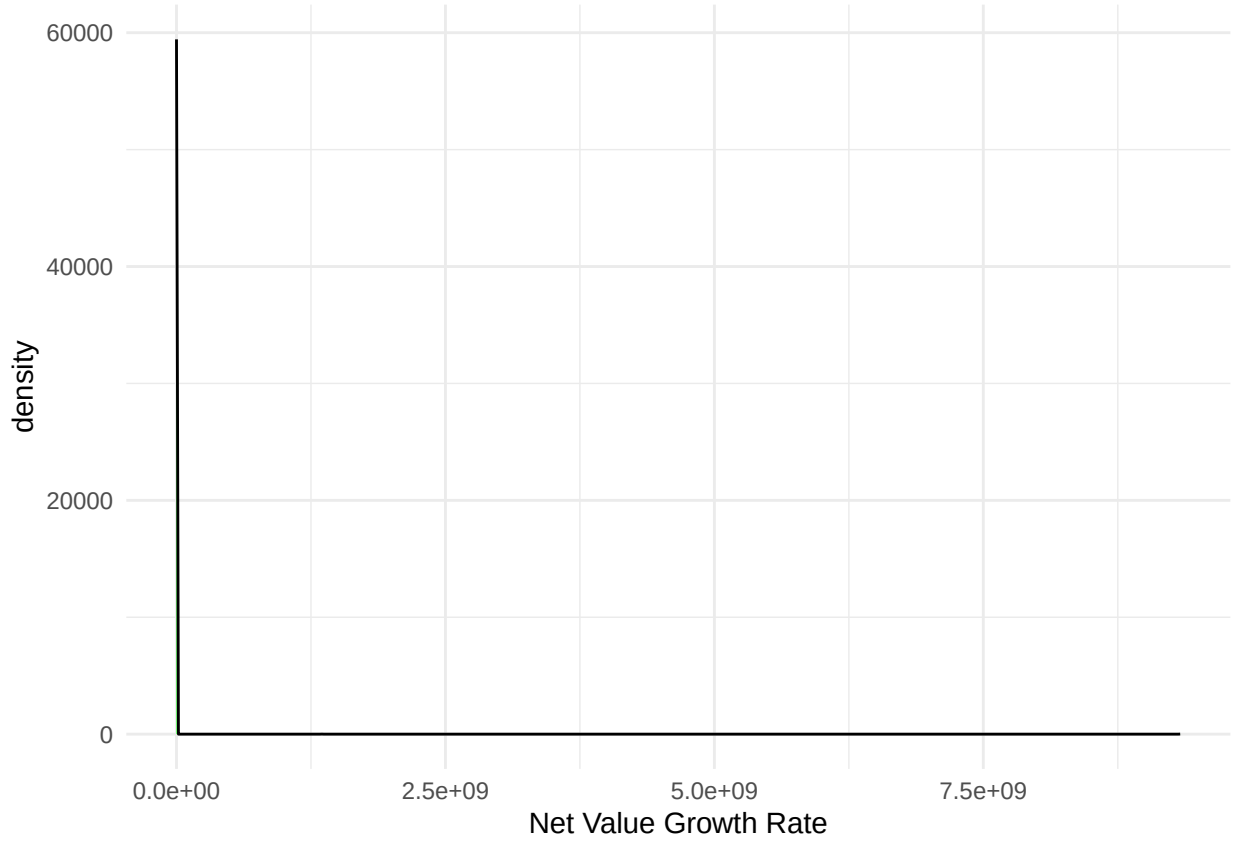

15000
10000
5000
0
density
0.00
0.25
0.50
0.75
1.00
Operating Profit Growth Rate

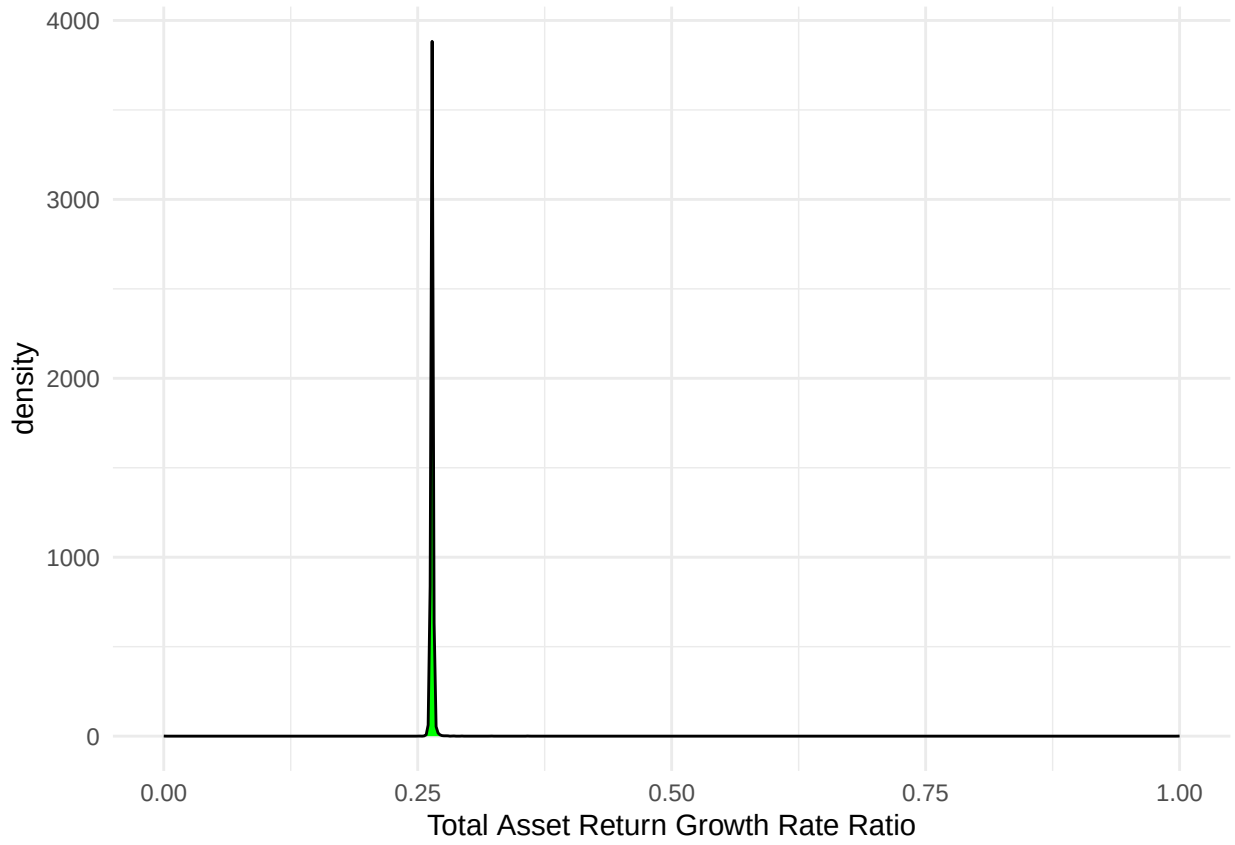

6000
4000
2000
0
density
0.00
0.25
0.50
0.75
1.00
After−tax Net Profit Growth Rate


6000
4000
2000
0
density
0.00
0.25
0.50
0.75
1.00
Regular Net Profit Growth Rate


60000
40000
20000
0
density
0.00
0.25
0.50
0.75
1.00
Continuous Net Profit Growth Rate


density
2e−10
1e−10
0e+00
0.0e+00
2.5e+09
5.0e+09
7.5e+09
1.0e+10
Total Asset Growth Rate


60000
40000
20000
0
density
0.0e+00
2.5e+09
5.0e+09
7.5e+09
Net Value Growth Rate



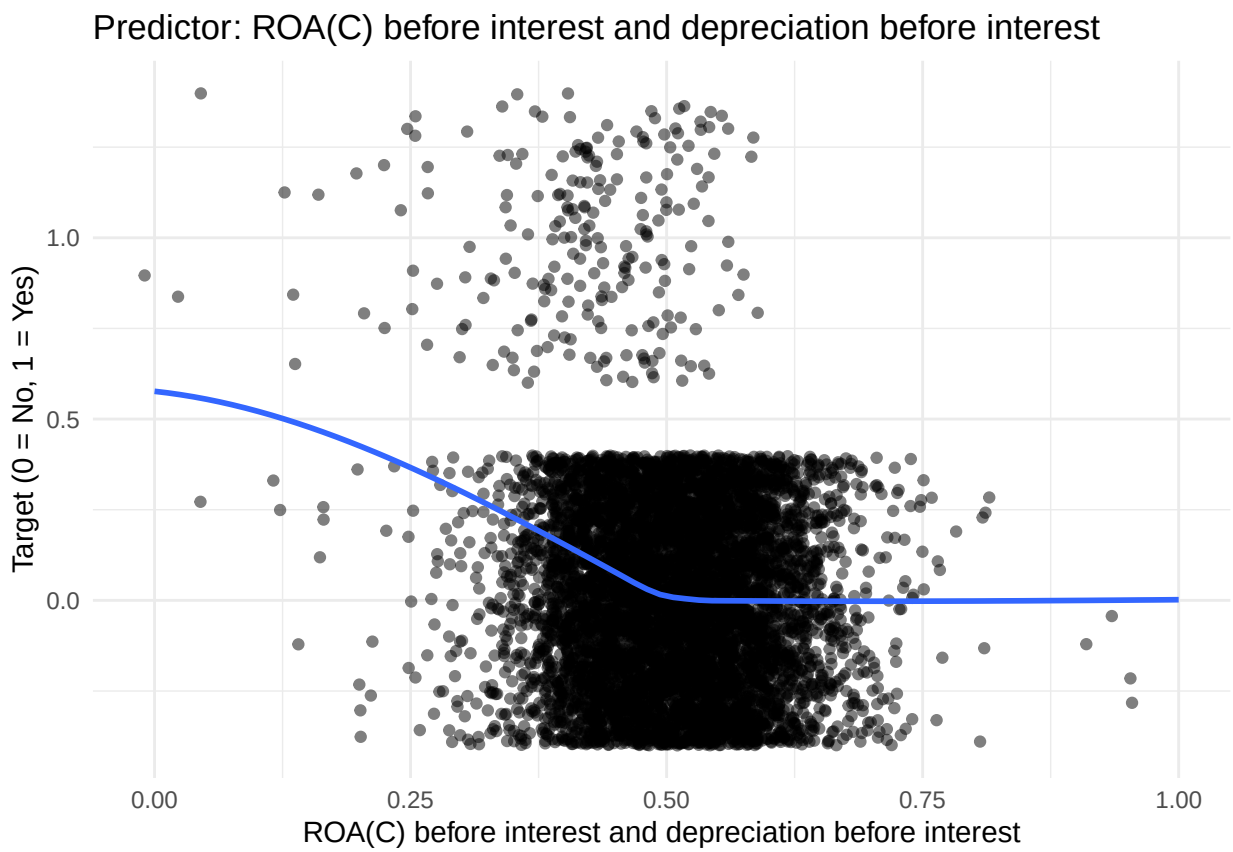
From the plots above we have detected for the most of the predictors **very high Kurtosis** which means sharp, tall peaks and heavy tails in the distributions. This implies most of the values are concentrated very close to the mean/median and there are also extreme outliers - relatively few but very far from the center.

Linearity

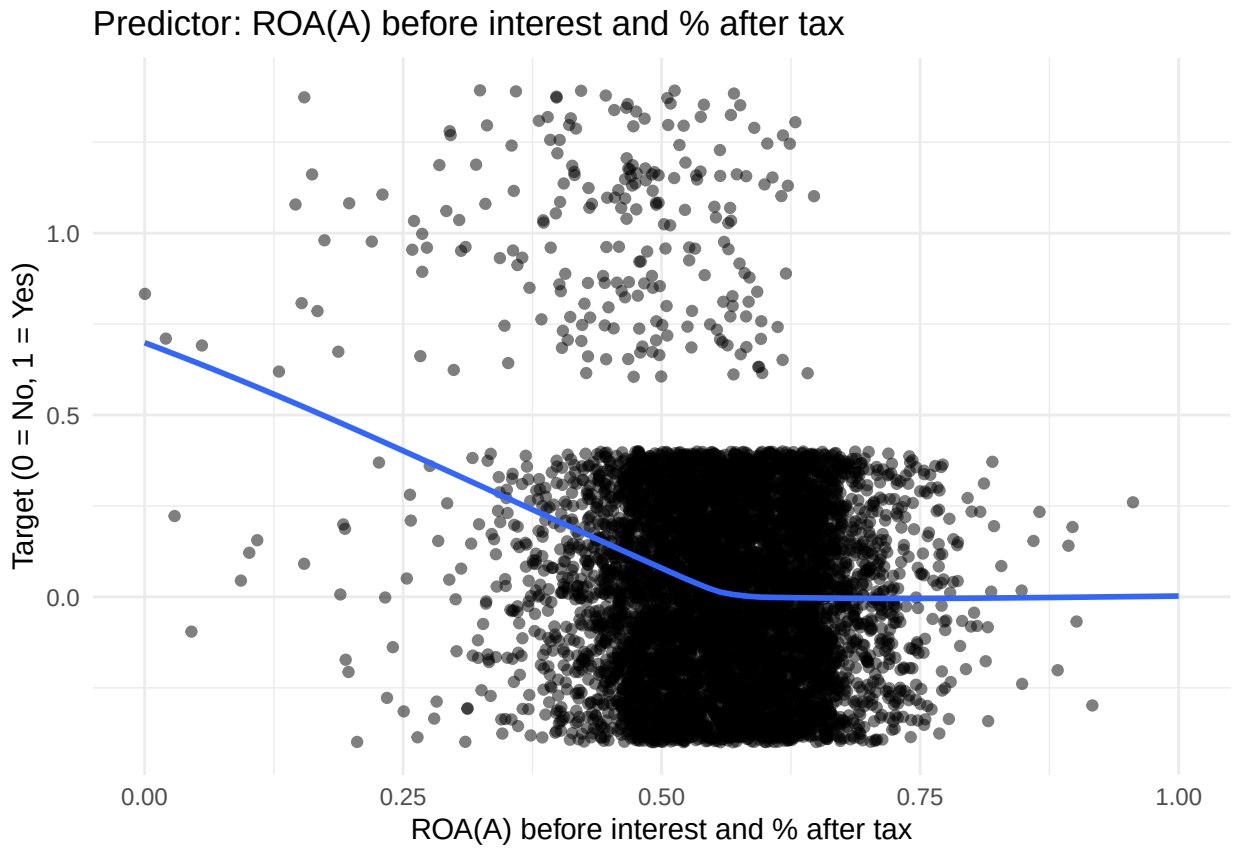
Now we can use scatterplot and try to apply the **smoothing line** and attempt to **catch the linearity between target and predictors**. If the line resembles a curve or a cluster the relationship is more complicated than linear.

```
show.curve.scatterplot <- function(df, cols, target){
  # Coerce target safely: factor → character → numeric
  df <- df %>%
    mutate(y_numeric = as.numeric(as.character(.data[[target]])))

  plot_list <- map(cols, ~ ggplot(df, aes(x = .data[[.x]], y = y_numeric)) +
    geom_jitter(width = 0.1, alpha = 0.5) +
    geom_smooth(method = "loess", se = FALSE) +
    labs(title = paste("Predictor:", .x), y = "Target (0 = No, 1 = Yes)") +
    theme_minimal()
  )

  walk(plot_list, print)
}

show.curve.scatterplot(df = selected_data, cols = PREDICTOR_NAMES, target = TARGET)
```

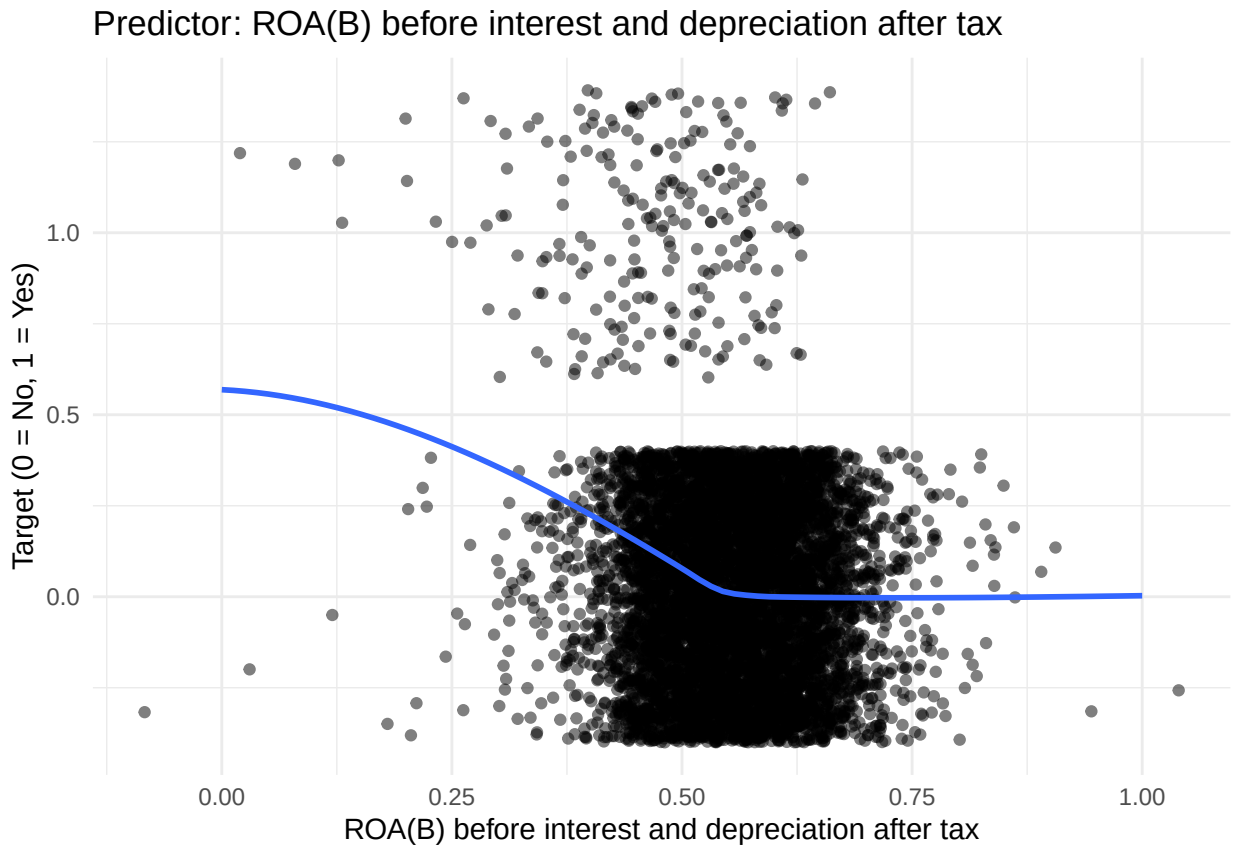
```
## `geom_smooth()` using formula = 'y ~ x'
```



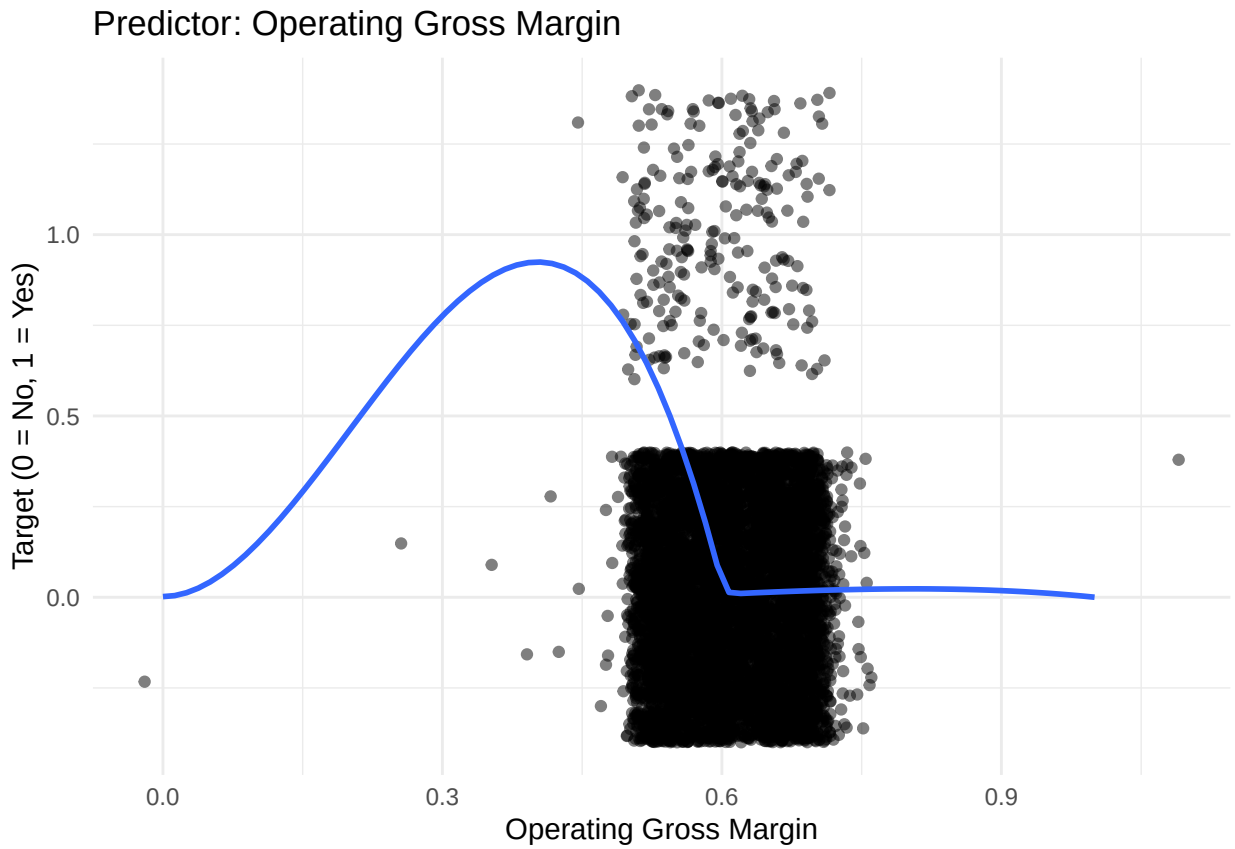

```
## `geom_smooth()` using formula = 'y ~ x'
```



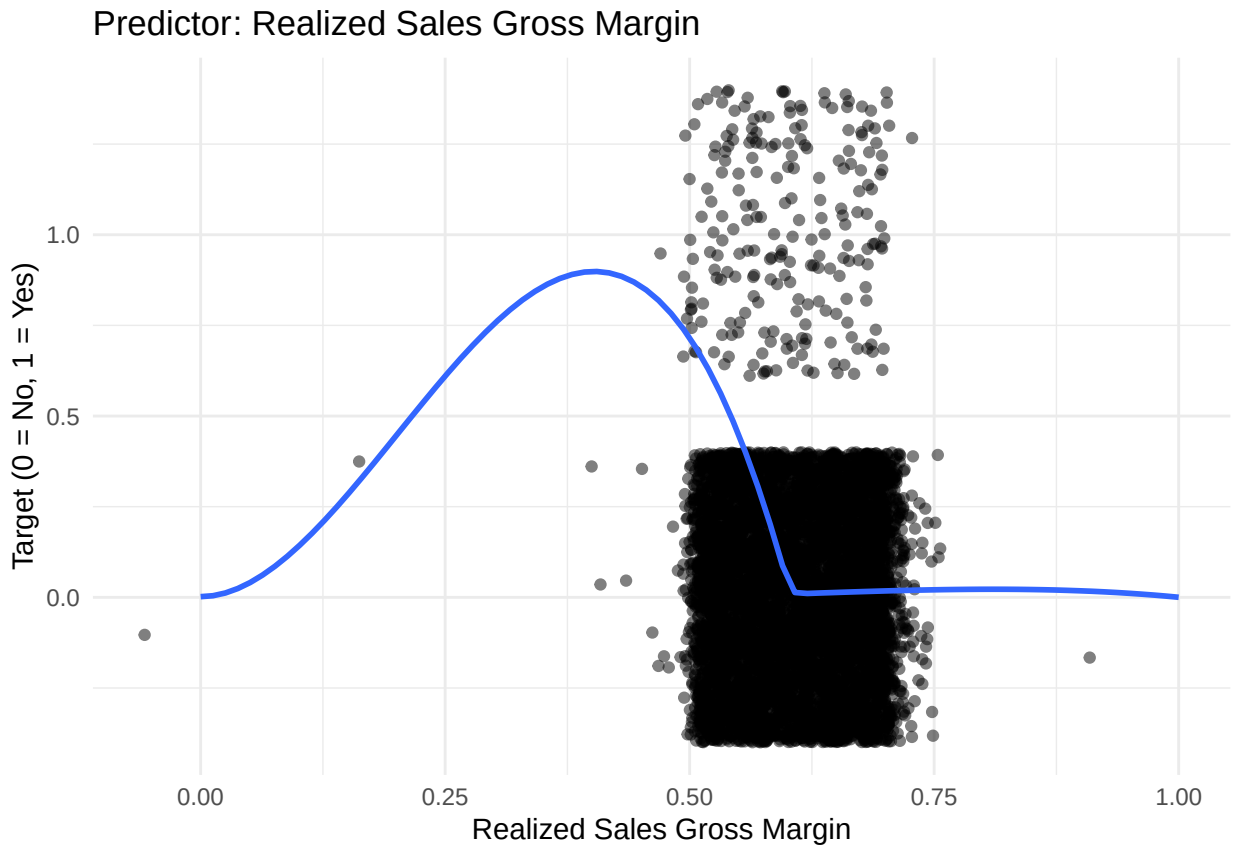

```
## `geom_smooth()` using formula = 'y ~ x'
```



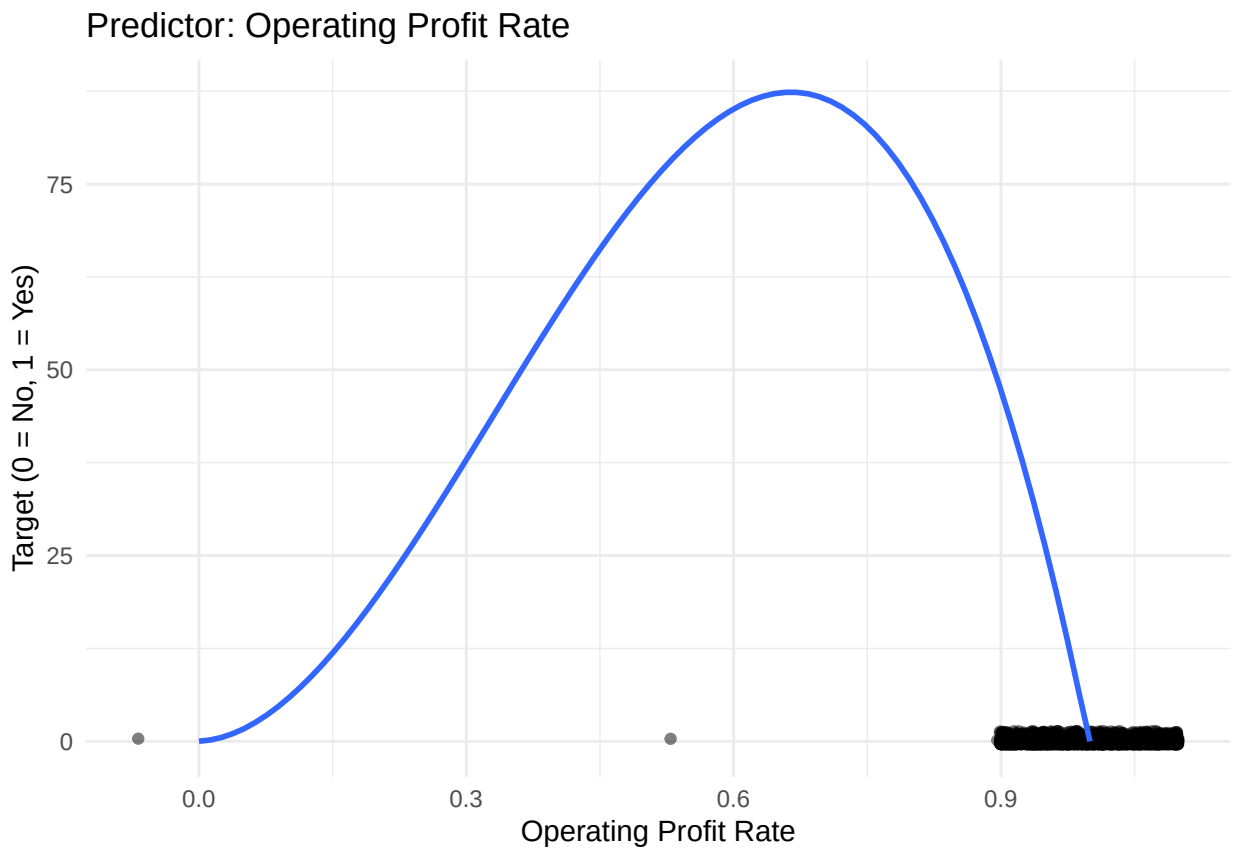

```
## `geom_smooth()` using formula = 'y ~ x'
```



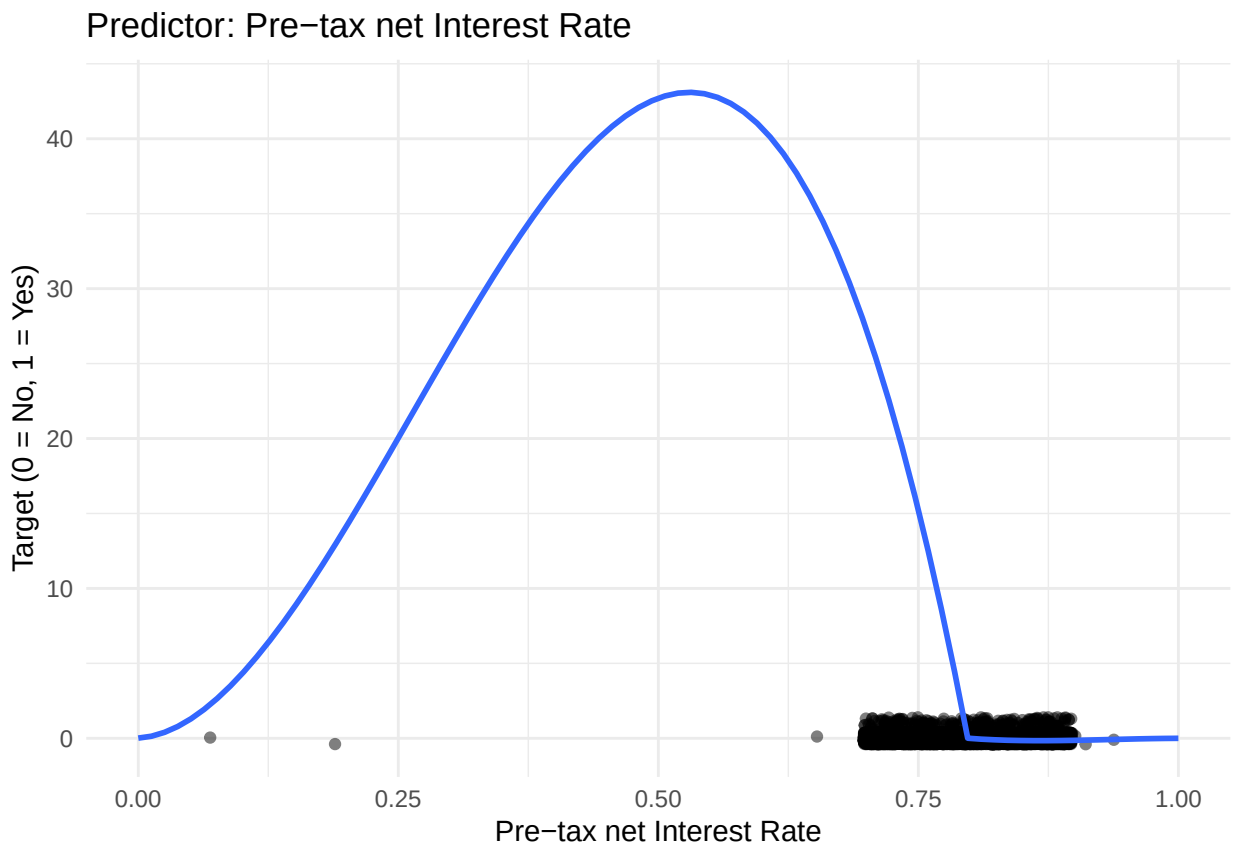

```
## `geom_smooth()` using formula = 'y ~ x'
```



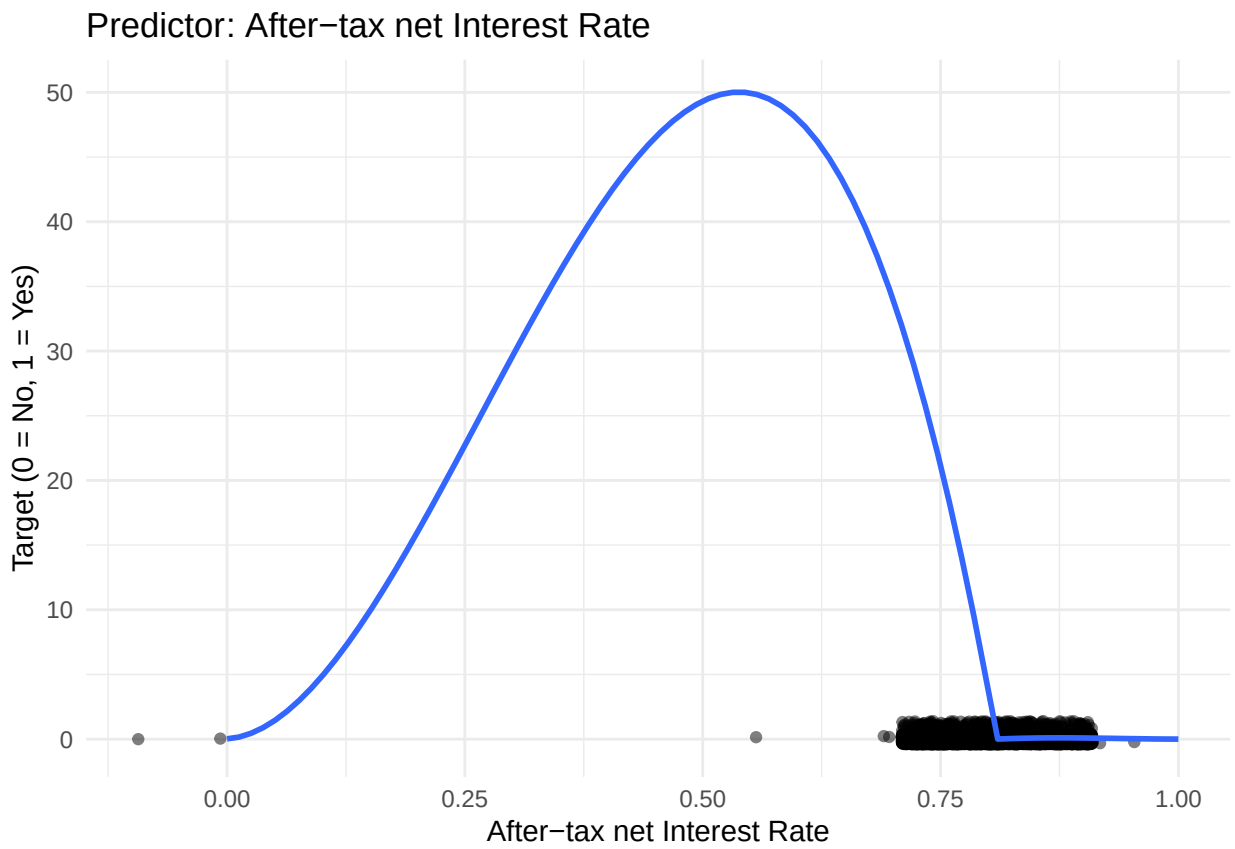

```
## `geom_smooth()` using formula = 'y ~ x'
```



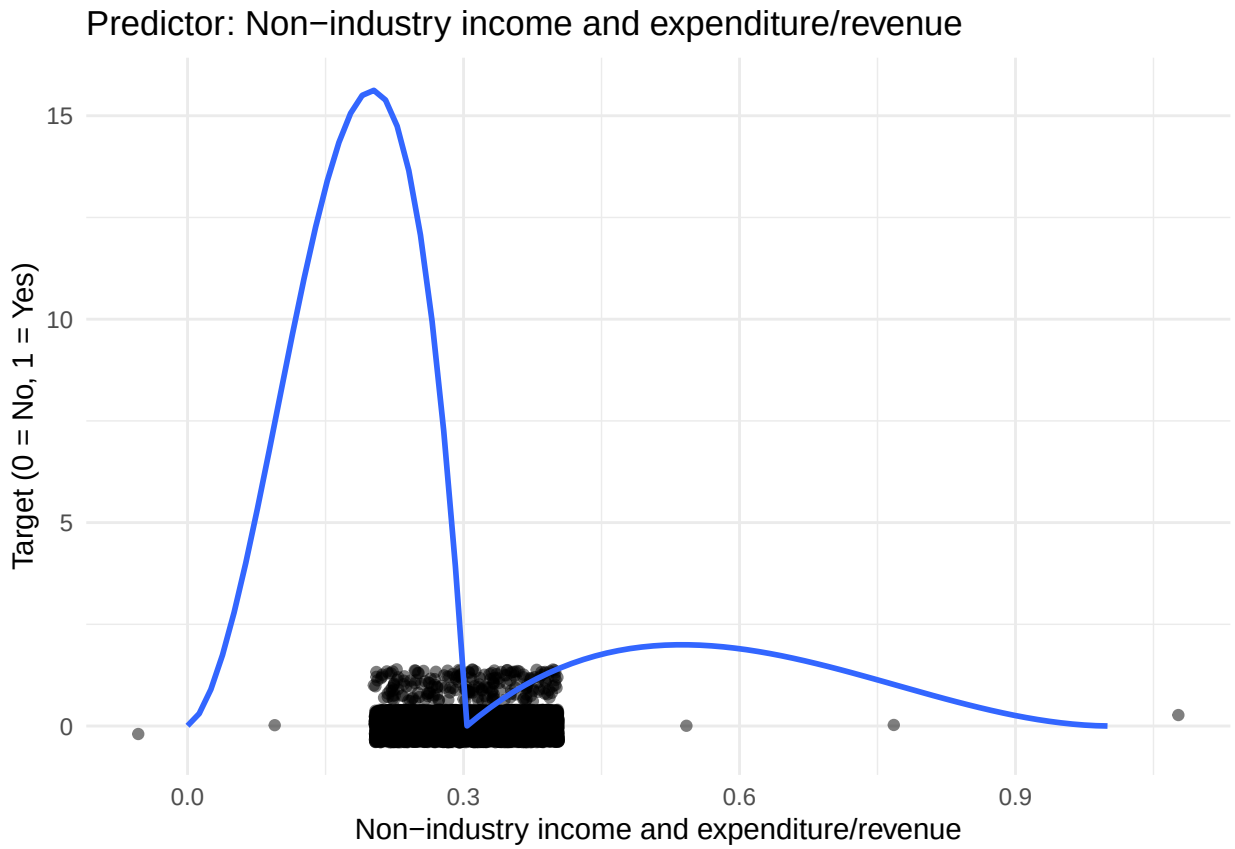

```
## `geom_smooth()` using formula = 'y ~ x'
```



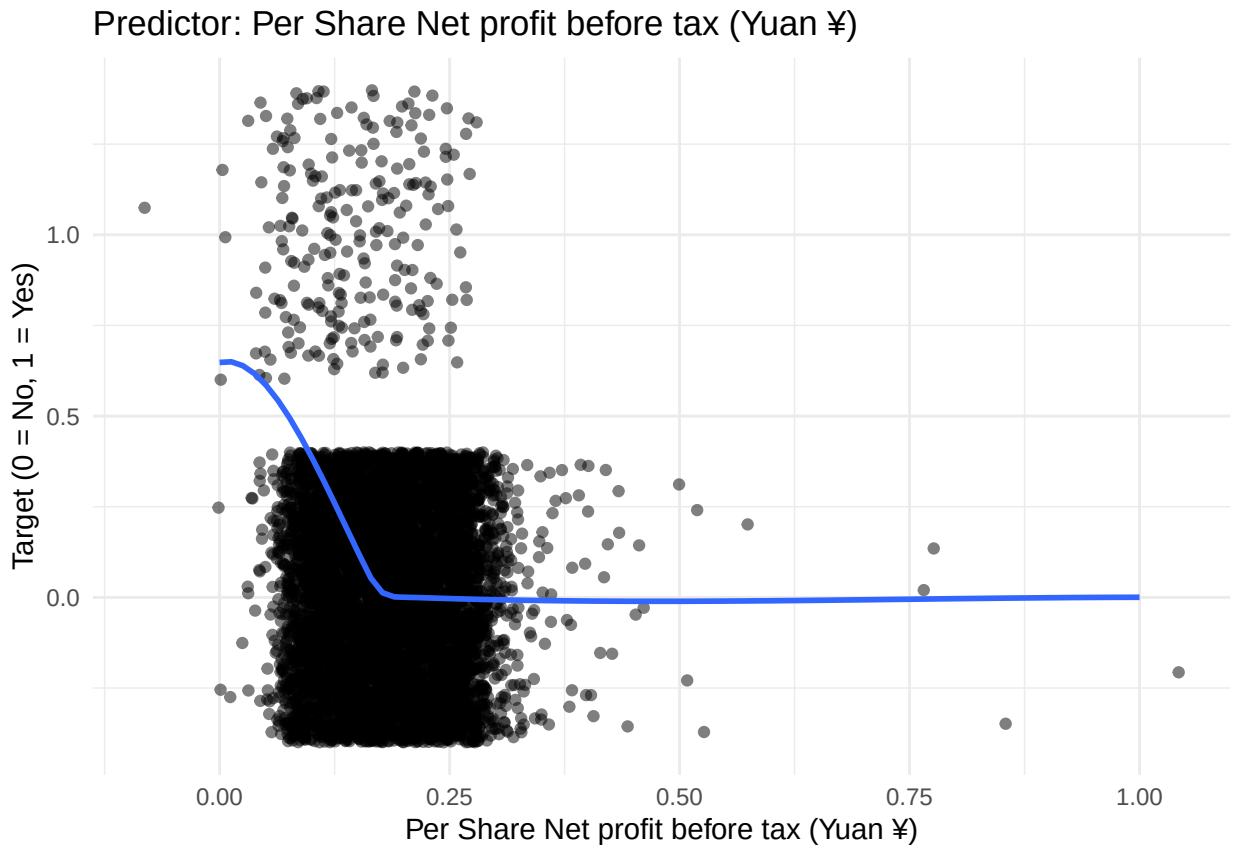

```
## `geom_smooth()` using formula = 'y ~ x'
```



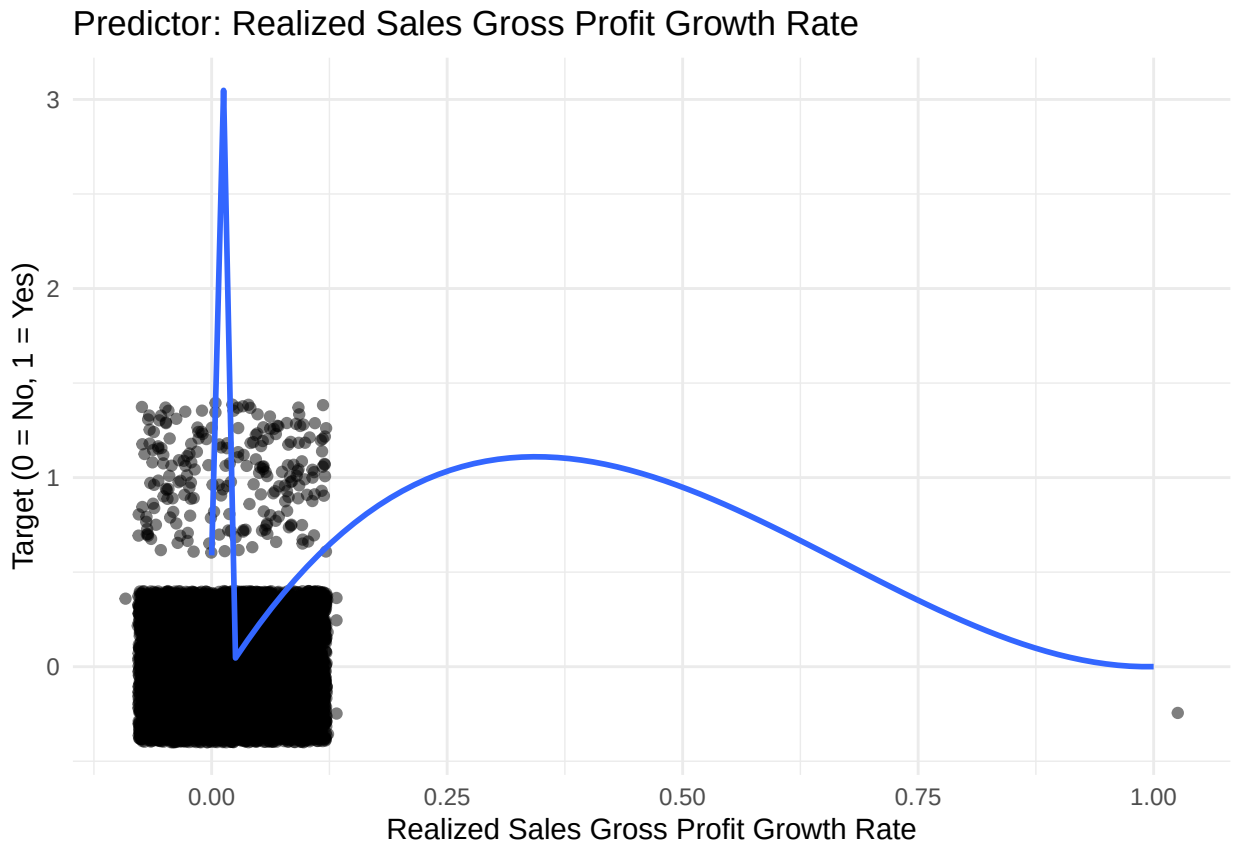

```
## `geom_smooth()` using formula = 'y ~ x'
```



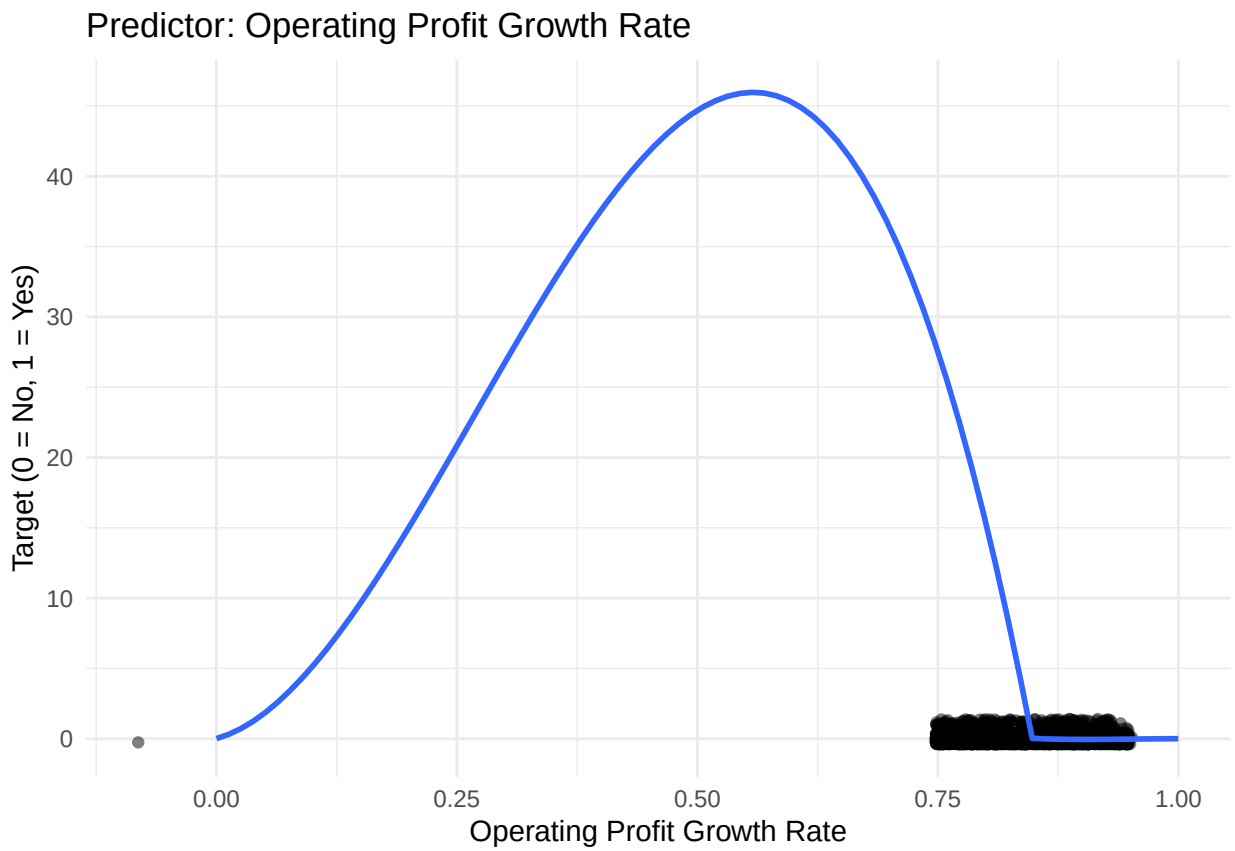

```
## `geom_smooth()` using formula = 'y ~ x'
```



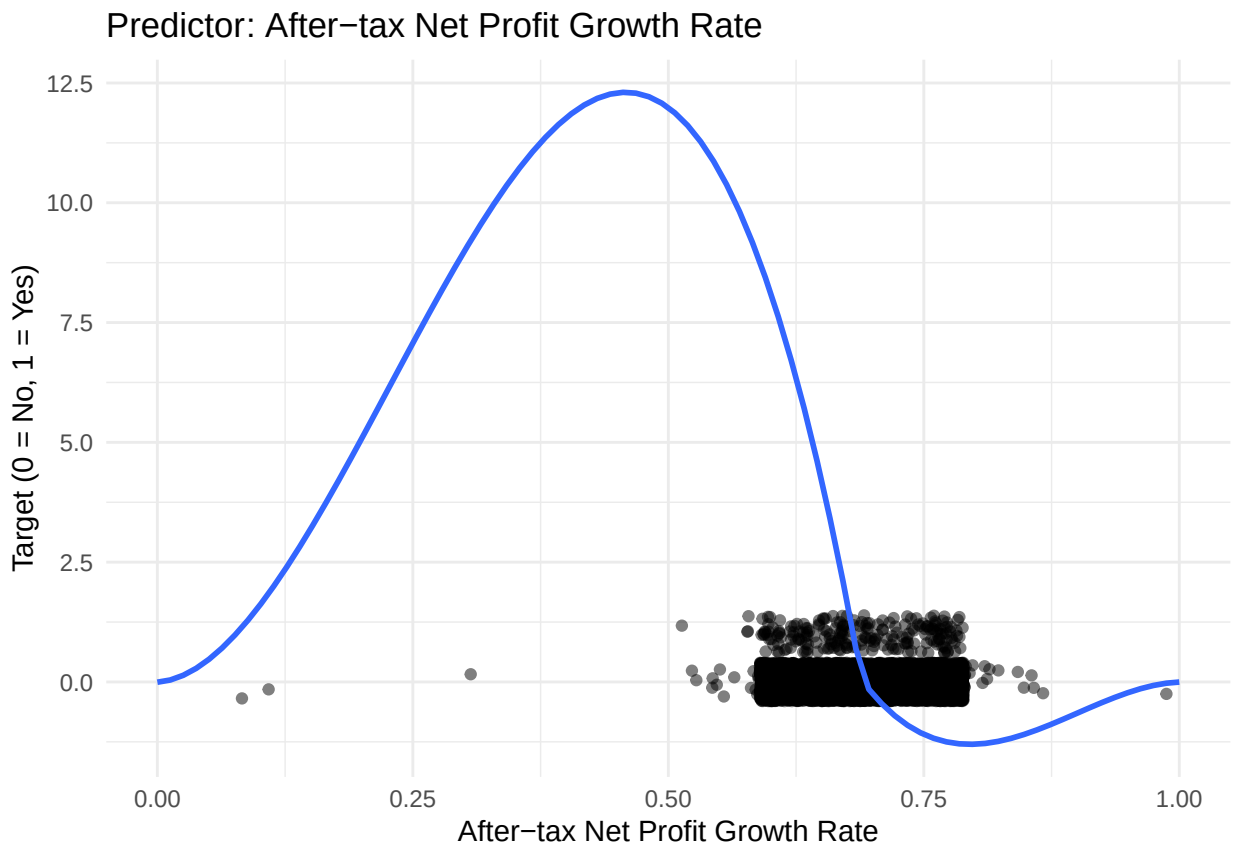

```
## `geom_smooth()` using formula = 'y ~ x'
```



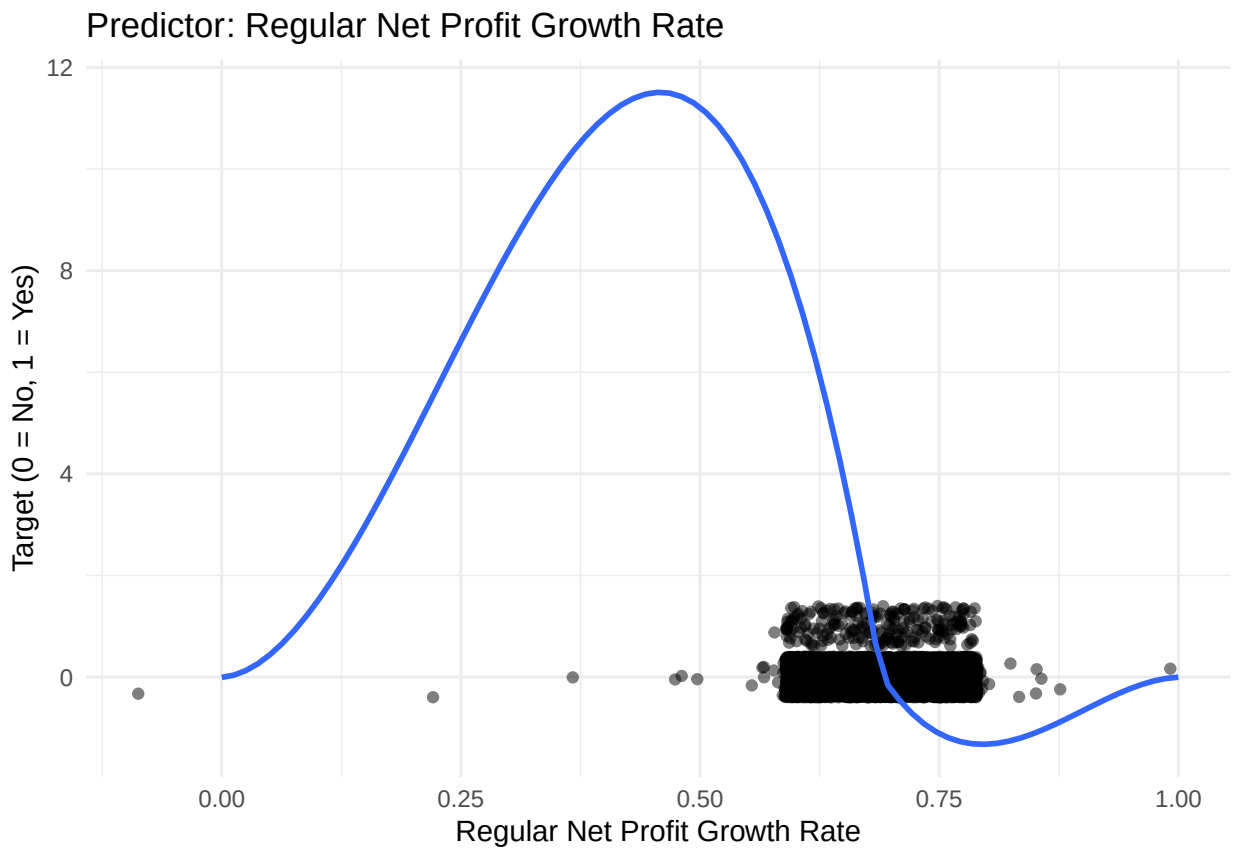

```
## `geom_smooth()` using formula = 'y ~ x'
```



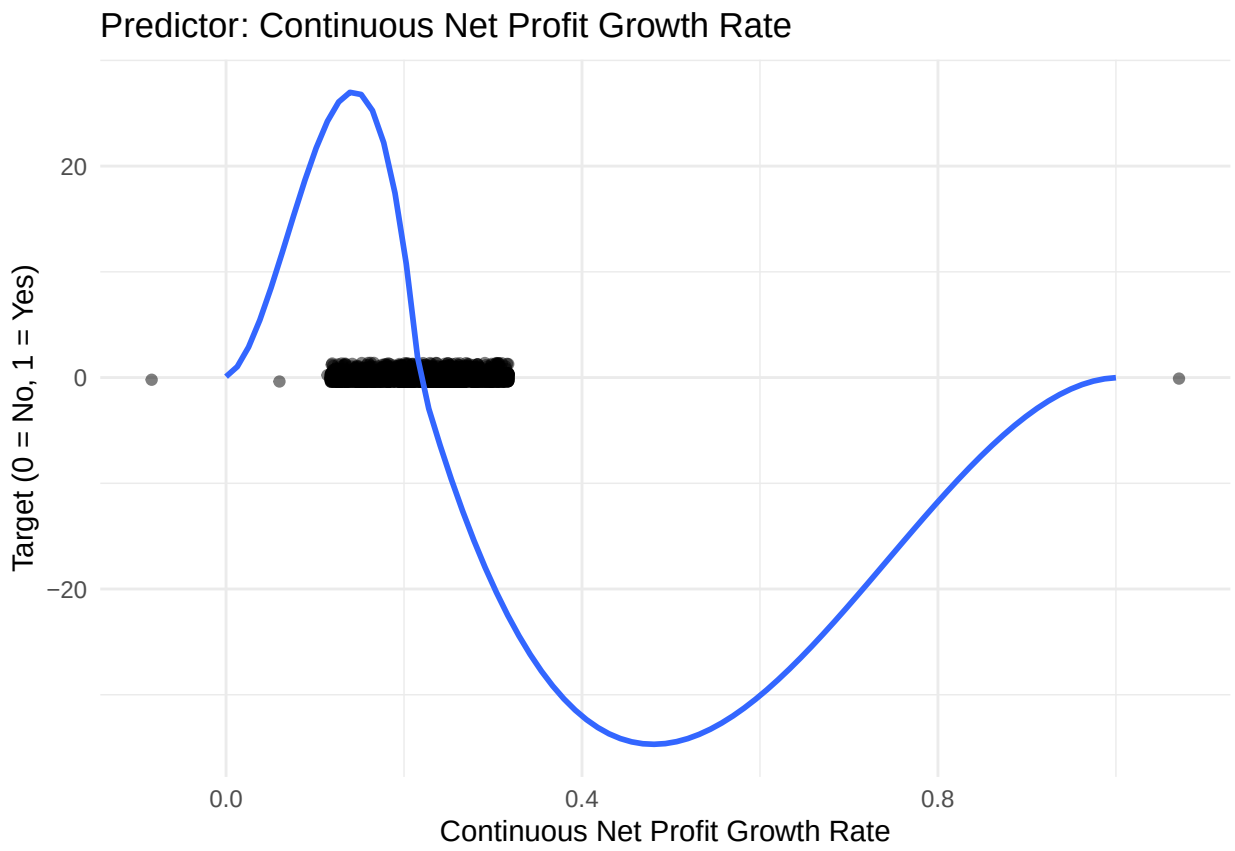

```
## `geom_smooth()` using formula = 'y ~ x'
```



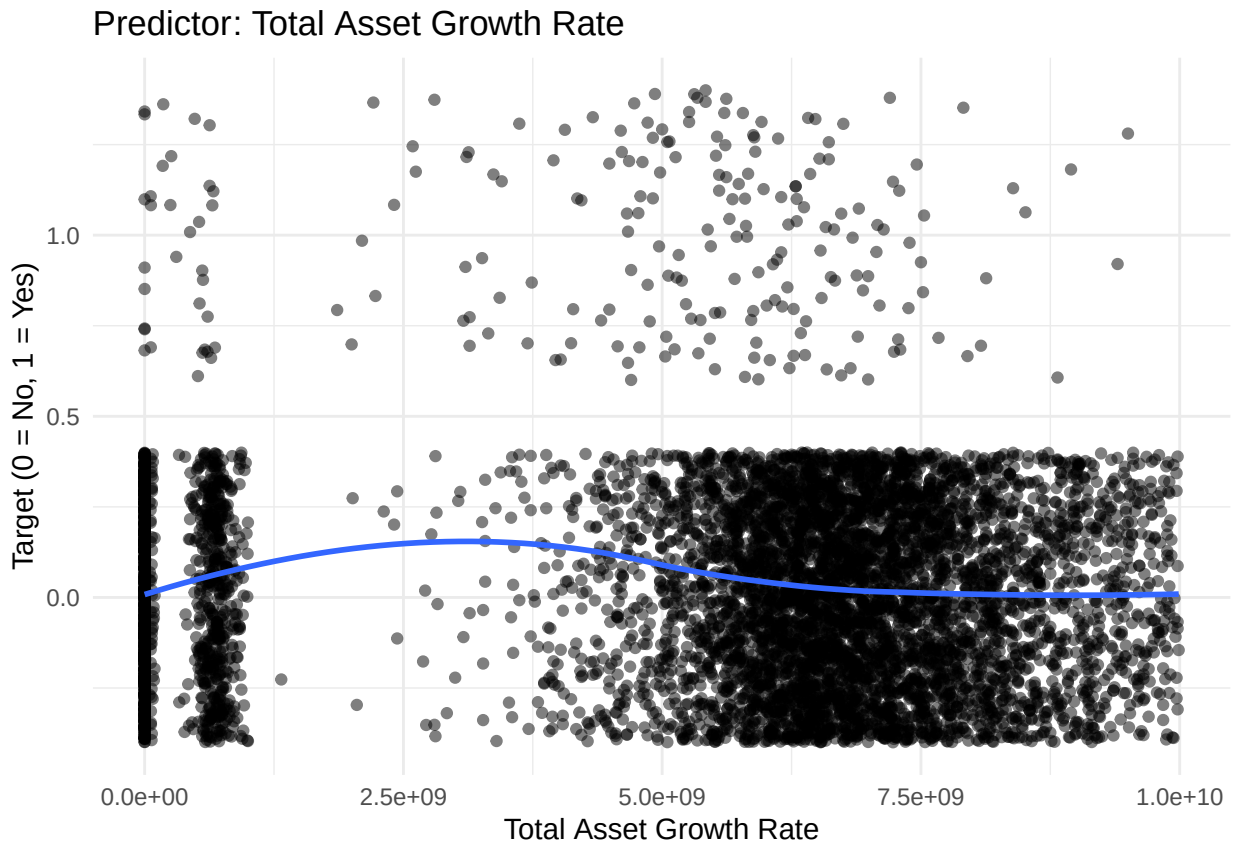

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
## `geom_smooth()` using formula = 'y ~ x'
```




```
## `geom_smooth()` using formula = 'y ~ x'
```



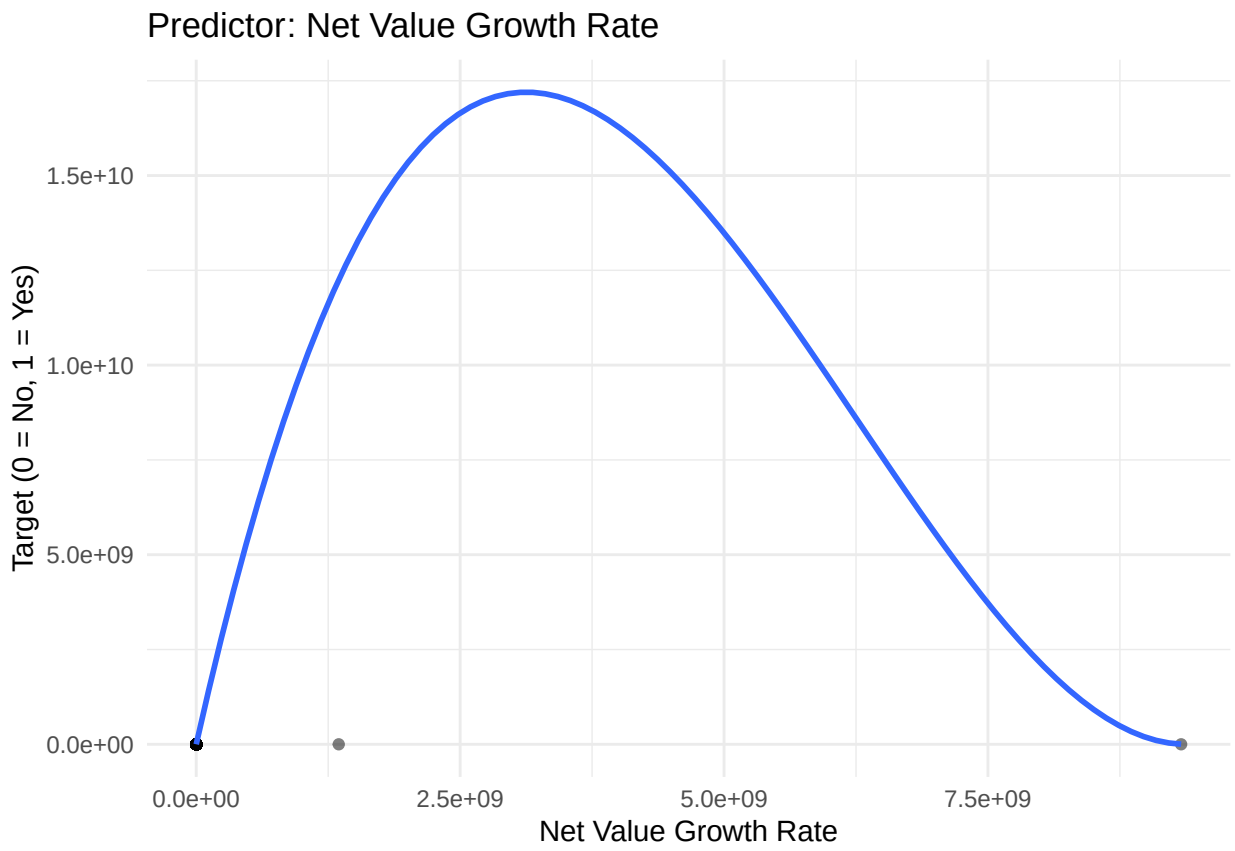

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : pseudoinverse used at -4.665e+07
```

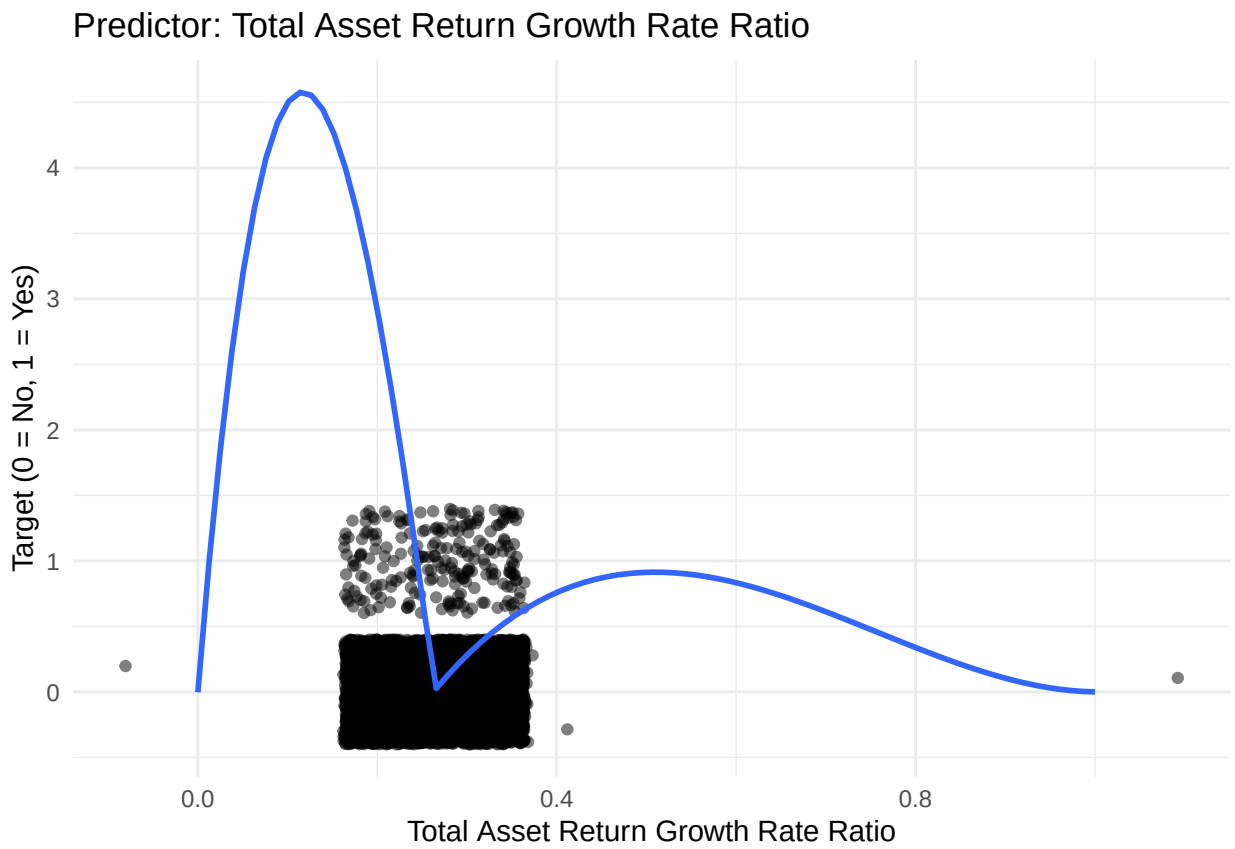
```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : neighborhood radius 4.665e+07
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : reciprocal condition number 5.3318e-15
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : There are other near singularities as well. 8.7922e+19
```




```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# show.scatterplot_with_sigmoid <- function(df, cols, target) {
#
#   plot_list <- map(cols, ~ ggplot(
#     data = df,
#     mapping = aes(x = .data[[.x]], y = .data[[target]])
#   ) +
#     geom_point(color = 'blue', alpha = 0.5) +  # Scatter plot points
#     geom_smooth(method = "glm",
#                 method.args = list(family = "binomial"),  # Logistic regression (sigmoid)
#                 se = FALSE, color = 'red') +  # Logistic regression line
#     theme_minimal())
#
#   walk(plot_list, print)
# }
#
# show.scatterplot_with_sigmoid(selected_data, PREDICTOR_NAMES, TARGET)
```

From the plots we can see that trying to draw a linear relationships for most of the predictors fails as one target class is highly underrepresented. In some cases, model would prefer one class only.

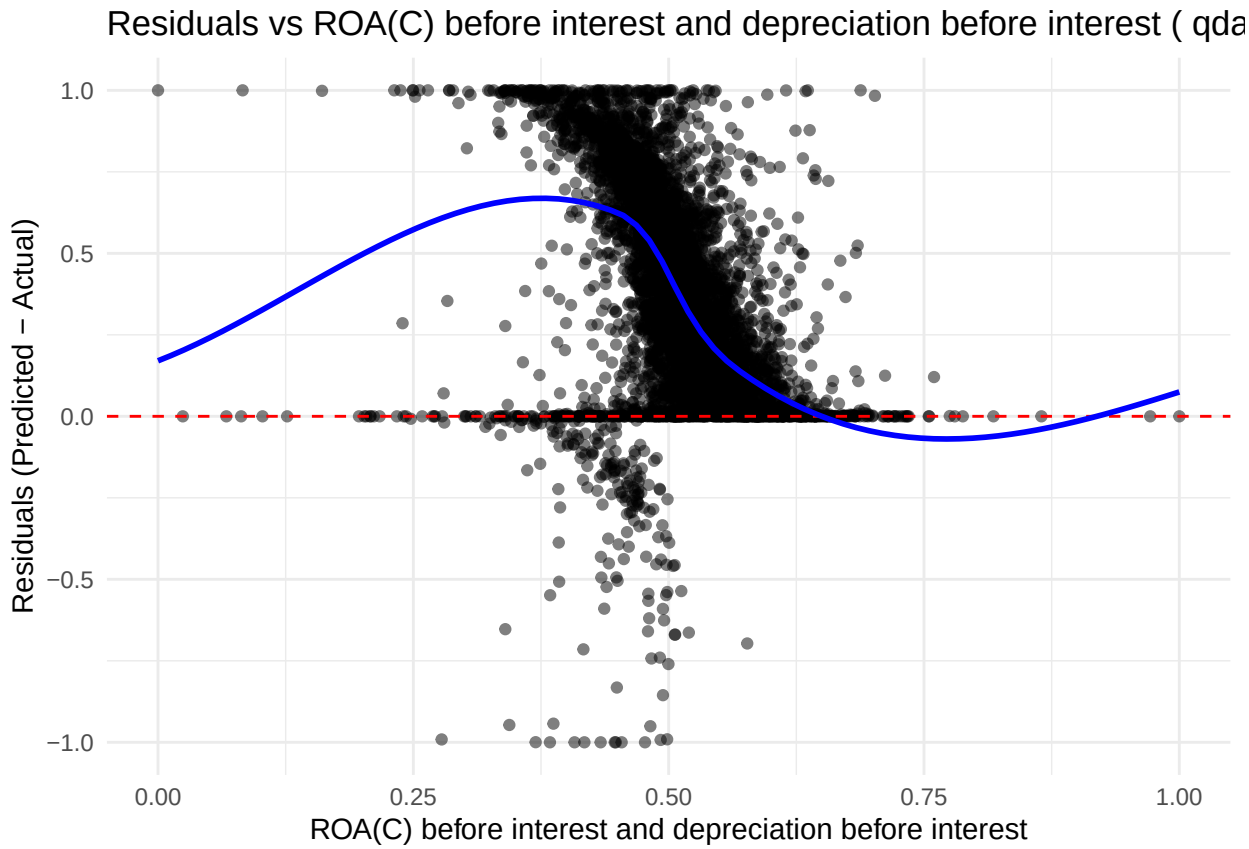
Homoscedasticity

LDA/QDA requires homoscedasticity in predictors' distributions. To test this out, we need to fit LDA or QDA model and display the residuals using **Scatterplot**.

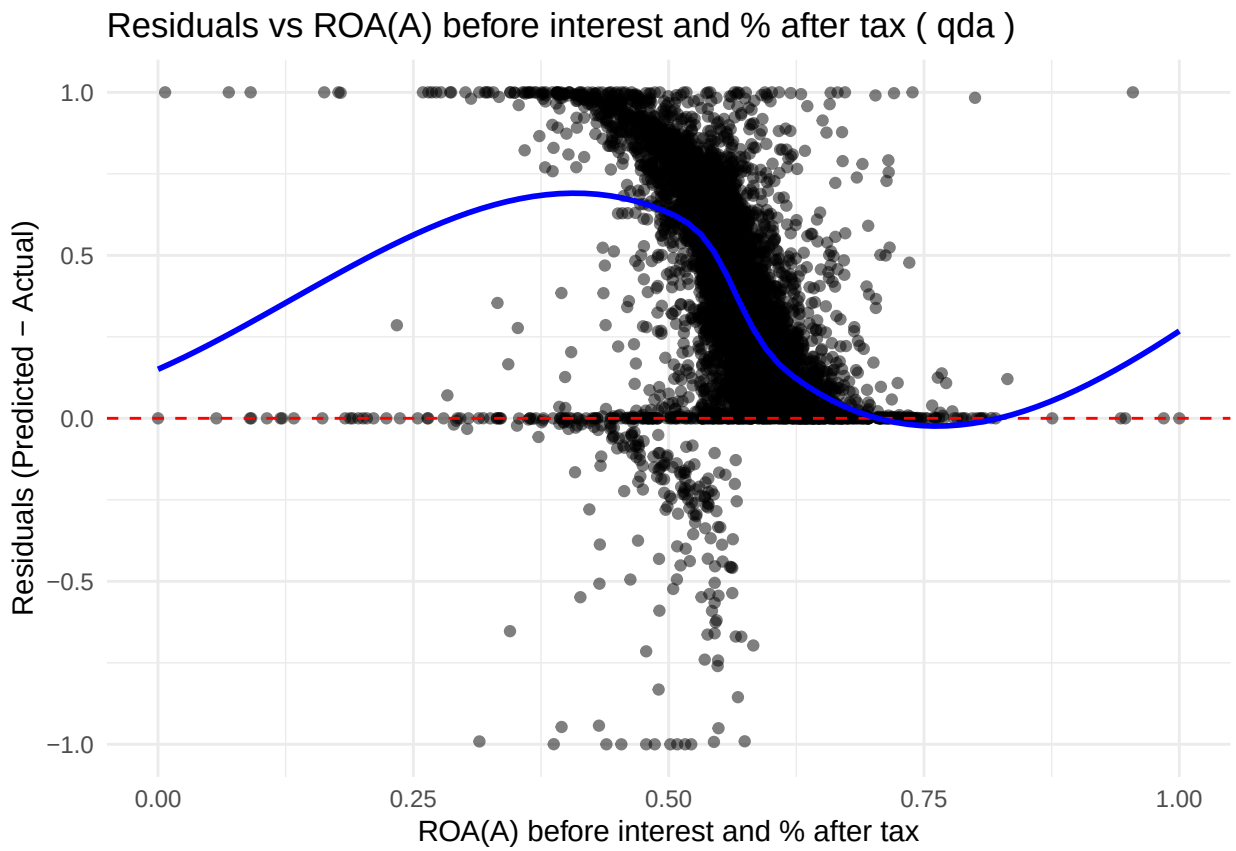
```
library(MASS)
library(ggplot2)
```

```r
library(dplyr)
library(purrr)

# Your target variable
TARGET <- "Bankrupt?"

# Make sure the target is a factor
selected_data[[TARGET]] <- as.factor(selected_data[[TARGET]])

# Construct a valid formula with backticks
model_formula <- as.formula(paste0("`", TARGET, "` ~ ."))

# Fit QDA
qda_model <- qda(model_formula, data = selected_data)

# Store predictions
selected_data$qda_pred <- predict(qda_model)$posterior[,2]

# Define residual plotting function
plot_residuals <- function(model_type) {
  walk(PREDICTOR_NAMES, ~{
    residuals <- selected_data[[paste0(model_type, "_pred")]] -
                 as.numeric(as.character(selected_data[[TARGET]]))

    p <- ggplot(selected_data, aes(x = .data[[.x]], y = residuals)) +
      geom_point(alpha = 0.5) +
      geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
      geom_smooth(method = "loess", color = "blue", se = FALSE) +
      labs(title = paste("Residuals vs", .x, "(", model_type, ")"),
           x = .x,
           y = "Residuals (Predicted - Actual)") +
      theme_minimal()
    print(p)
  })
}

# Plot residuals for LDA and QDA
plot_residuals("qda")
```

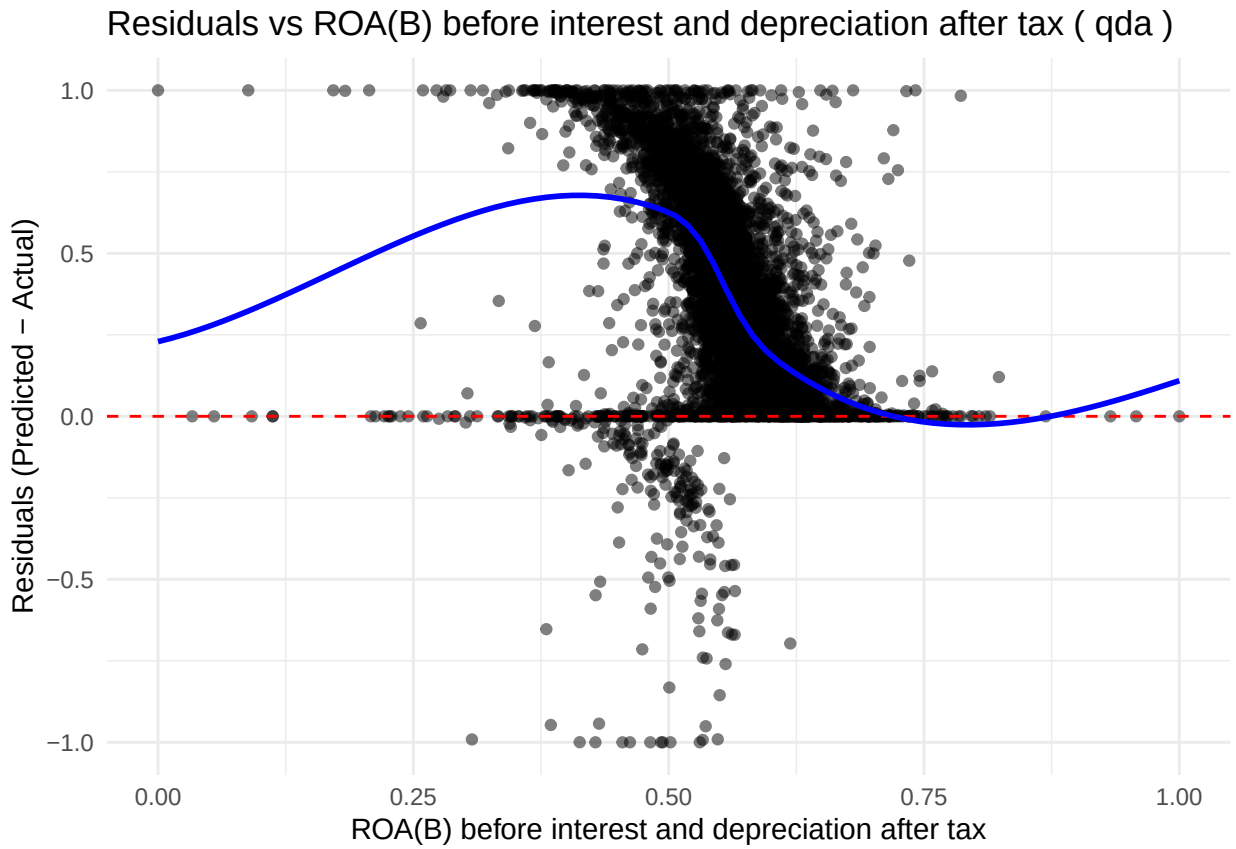
```
## `geom_smooth()` using formula = 'y ~ x'
```



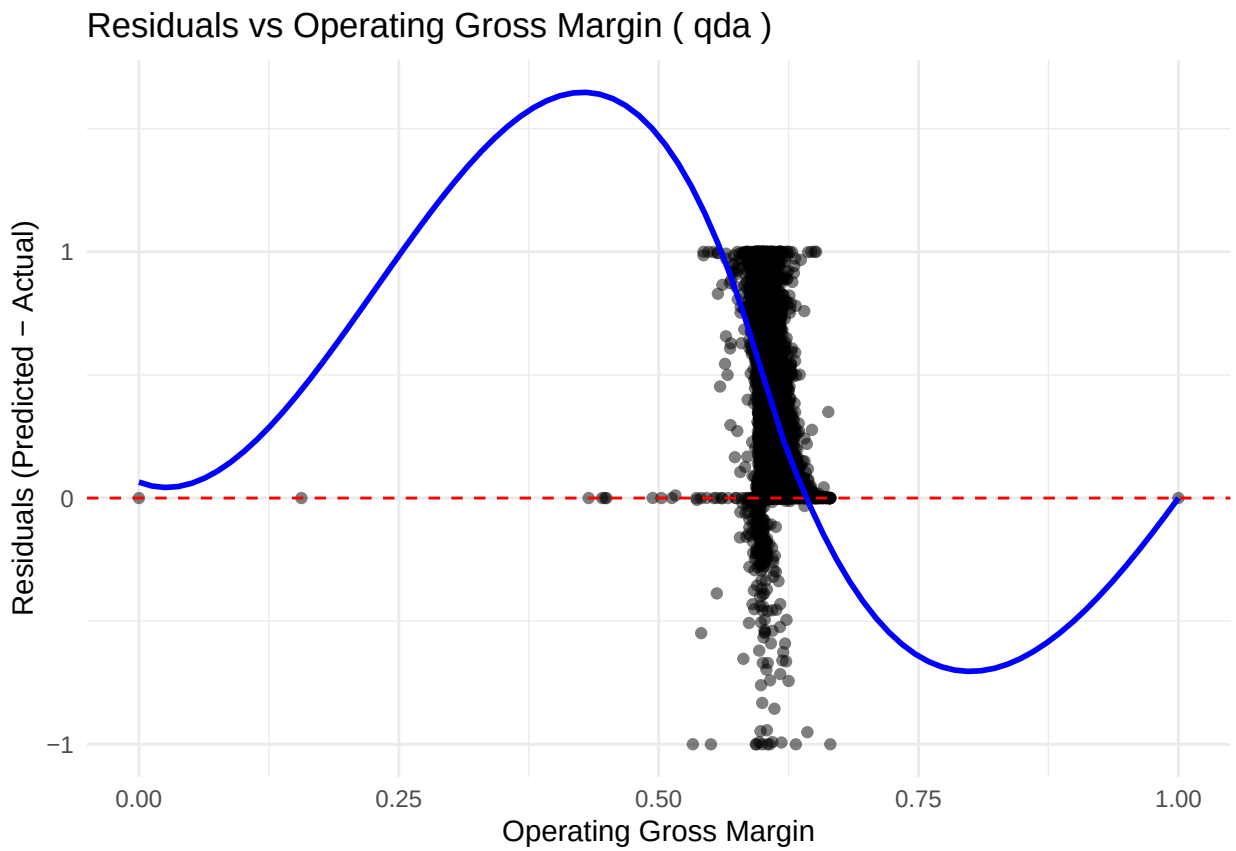

```
## `geom_smooth()` using formula = 'y ~ x'
```



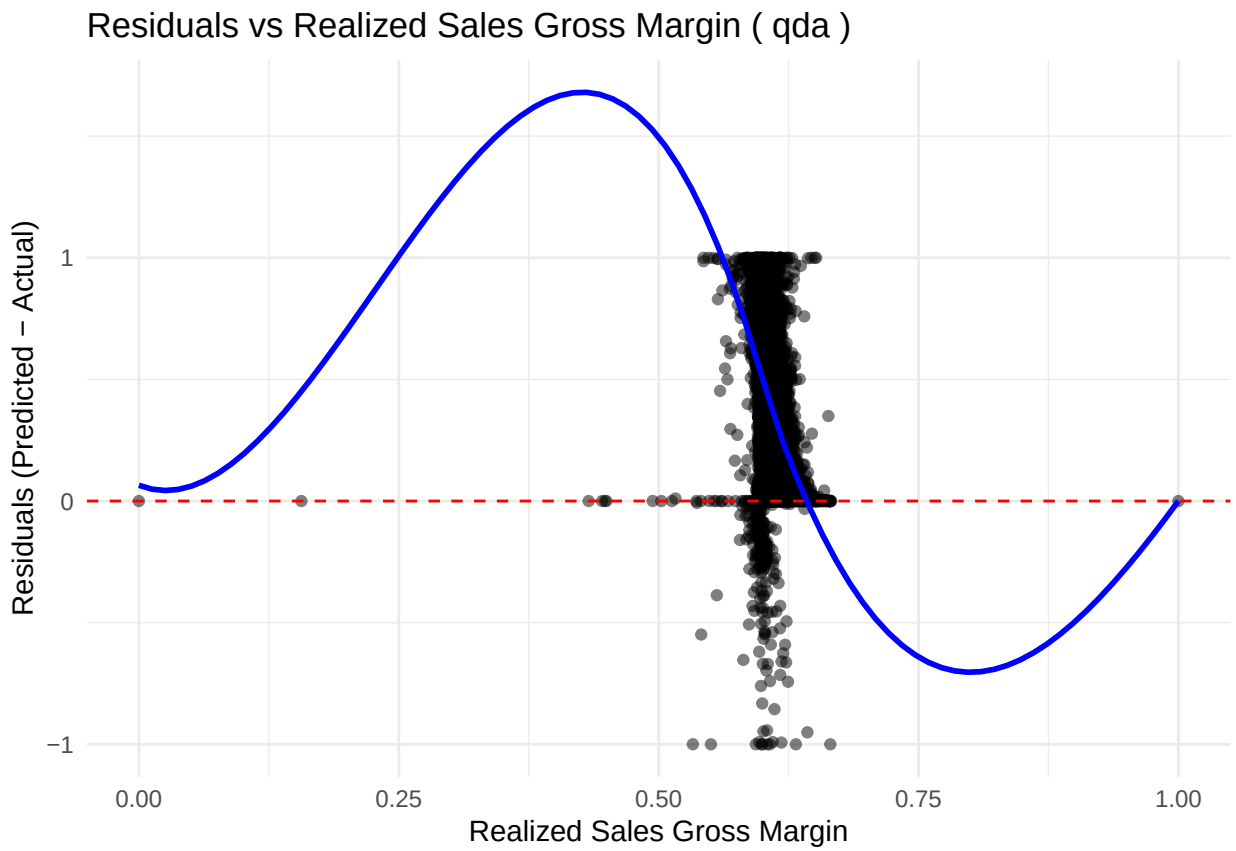

```
## `geom_smooth()` using formula = 'y ~ x'
```



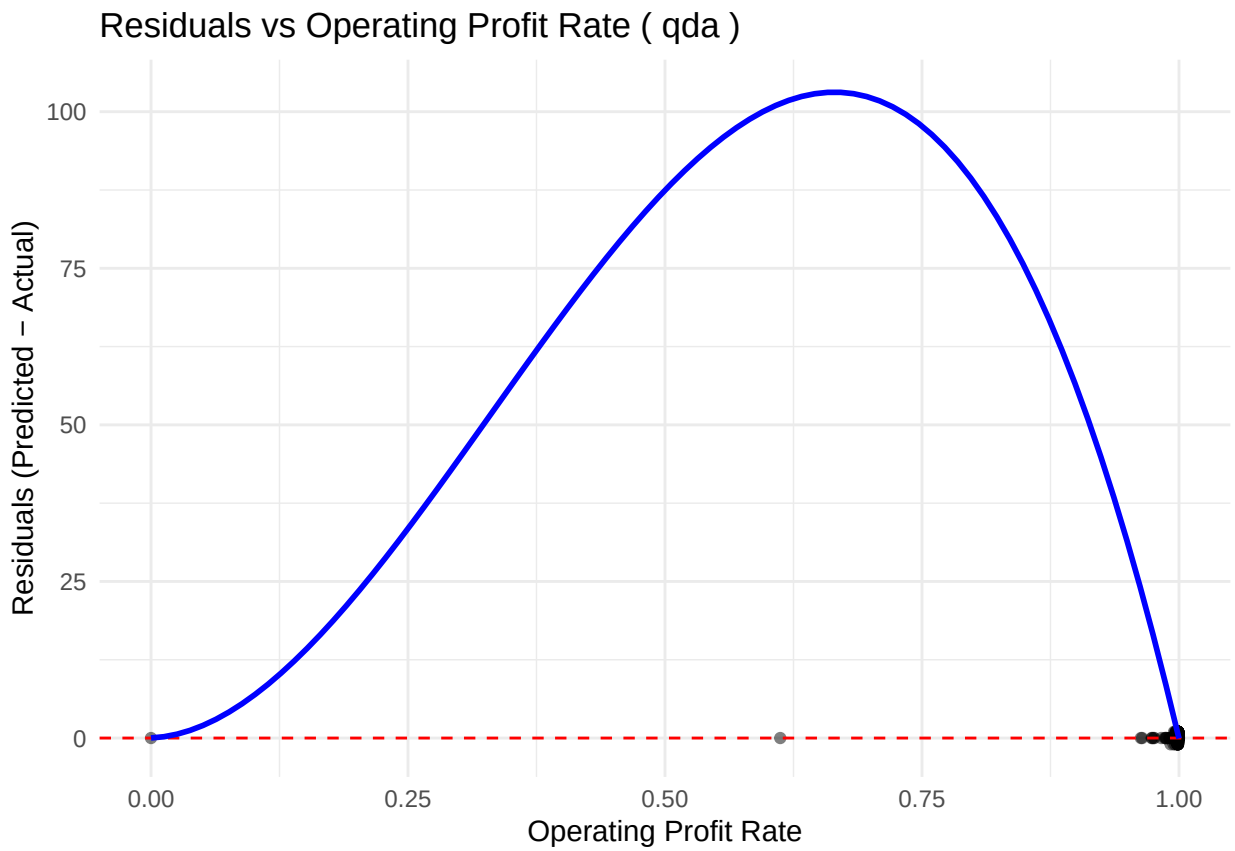

```
## `geom_smooth()` using formula = 'y ~ x'
```



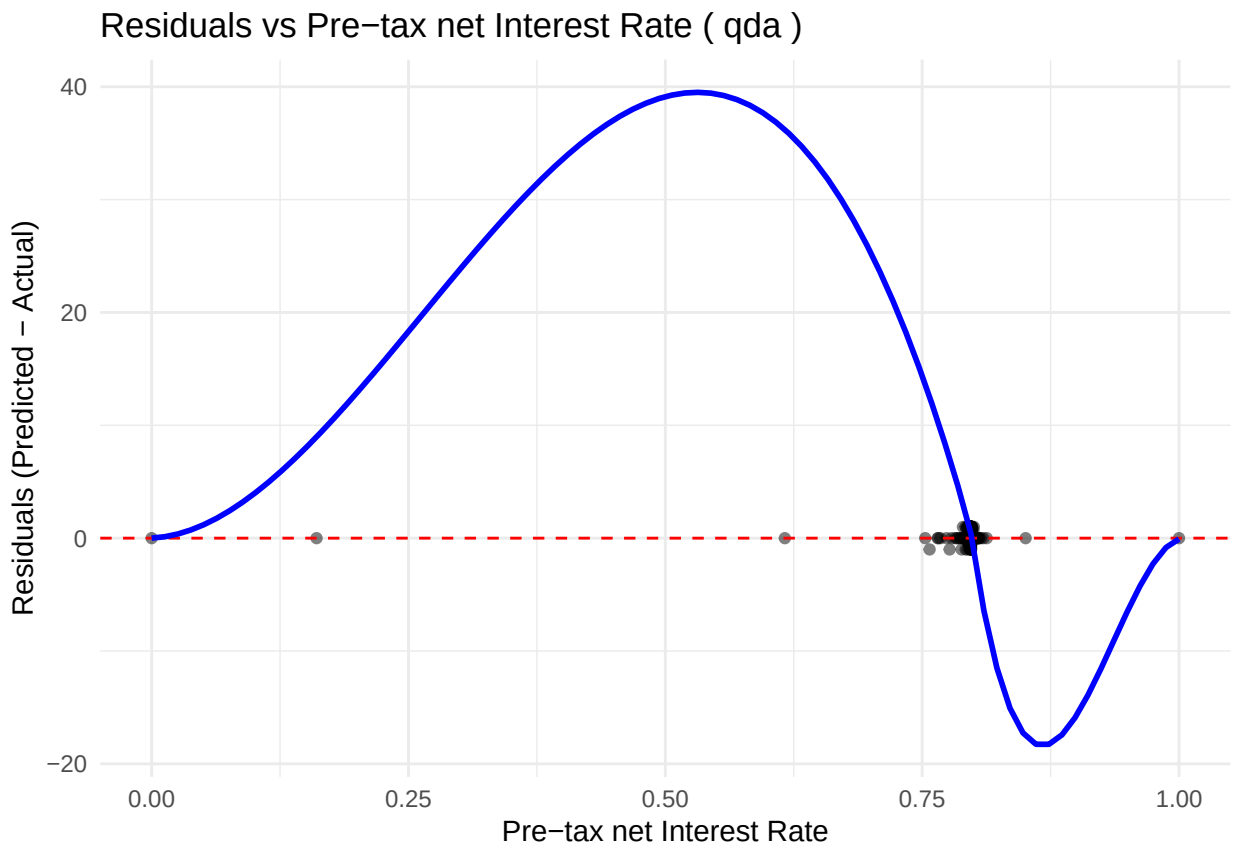

```
## `geom_smooth()` using formula = 'y ~ x'
```



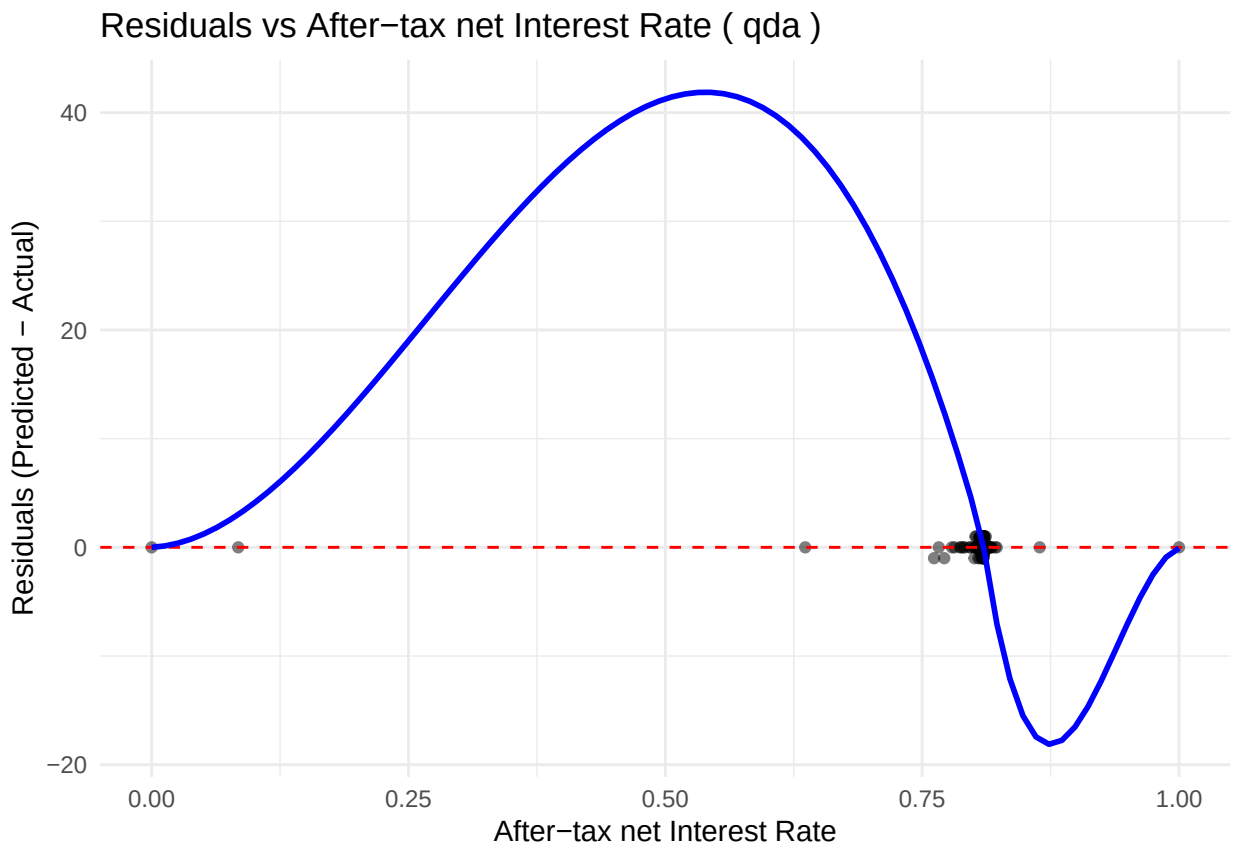

```
## `geom_smooth()` using formula = 'y ~ x'
```



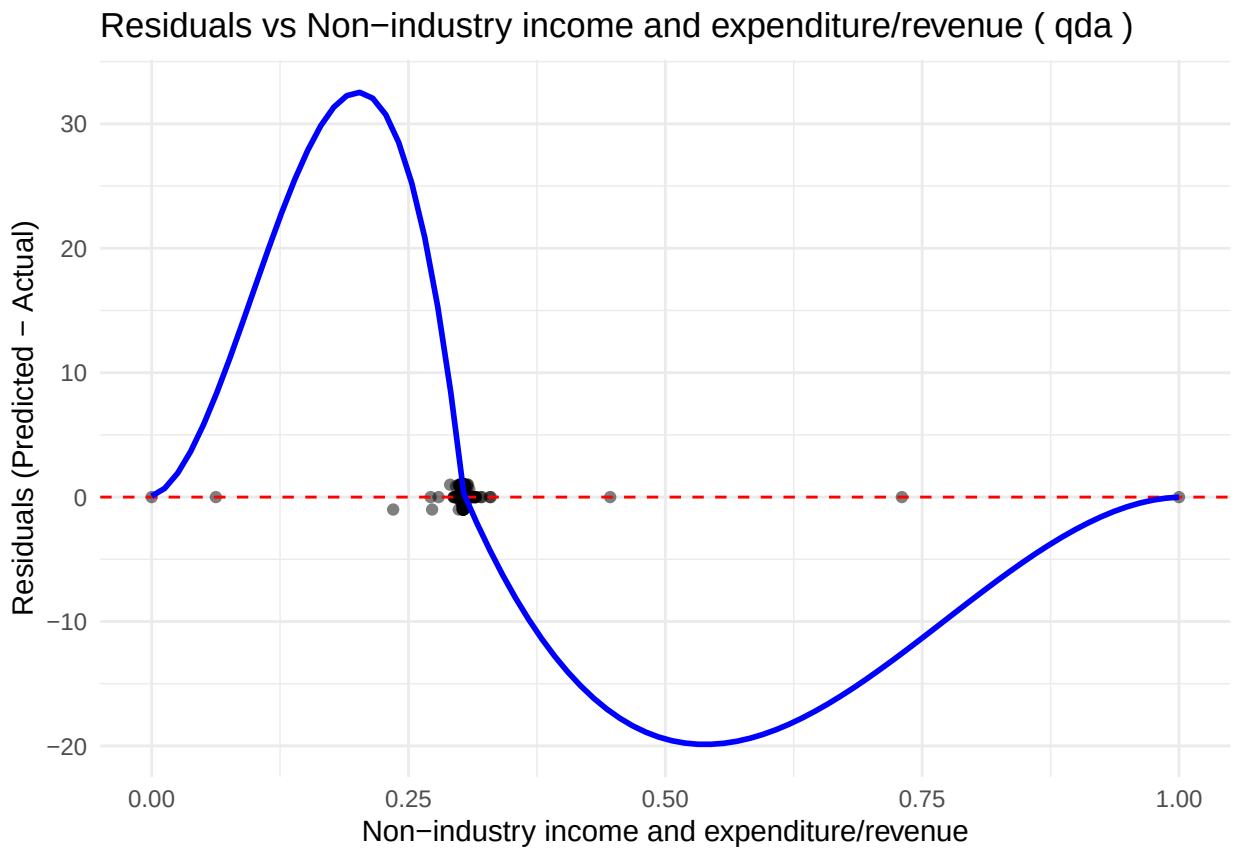

```
## `geom_smooth()` using formula = 'y ~ x'
```



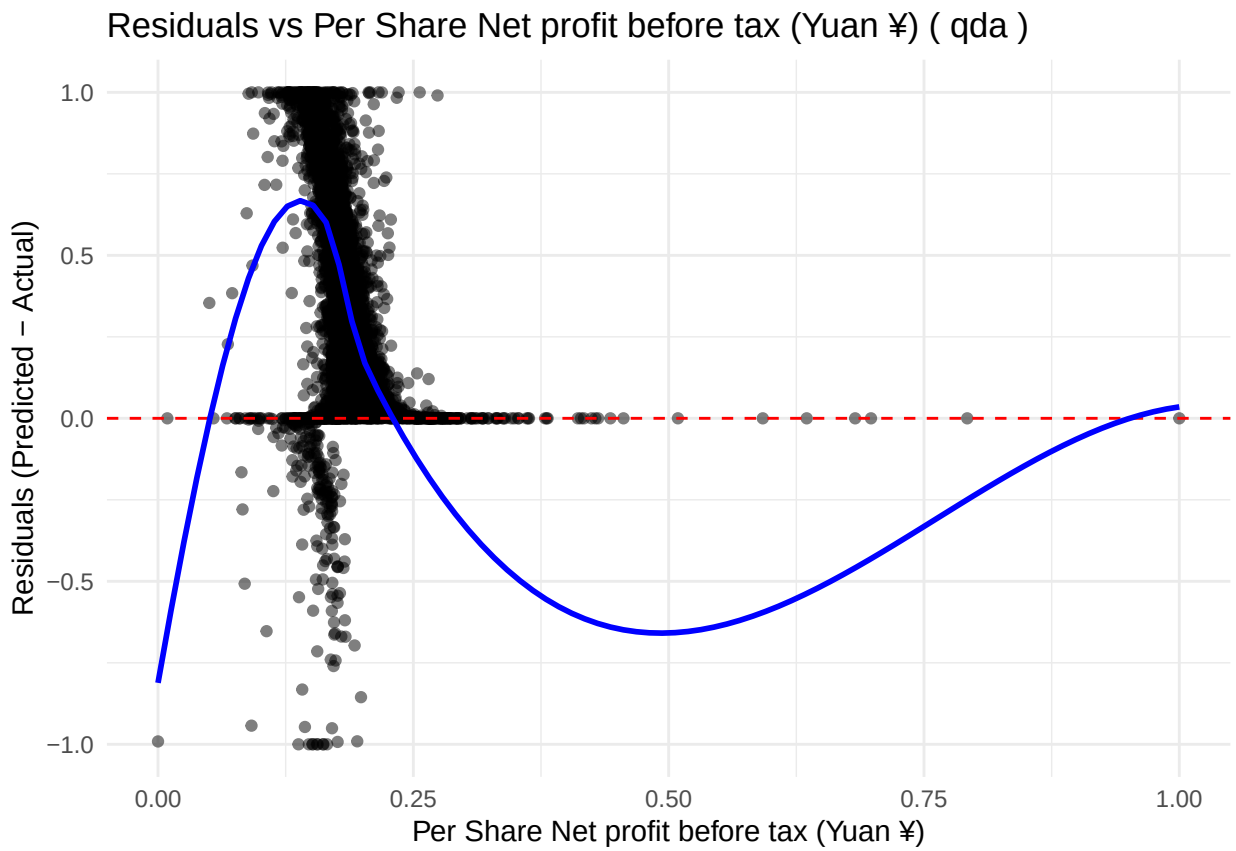

```
## `geom_smooth()` using formula = 'y ~ x'
```



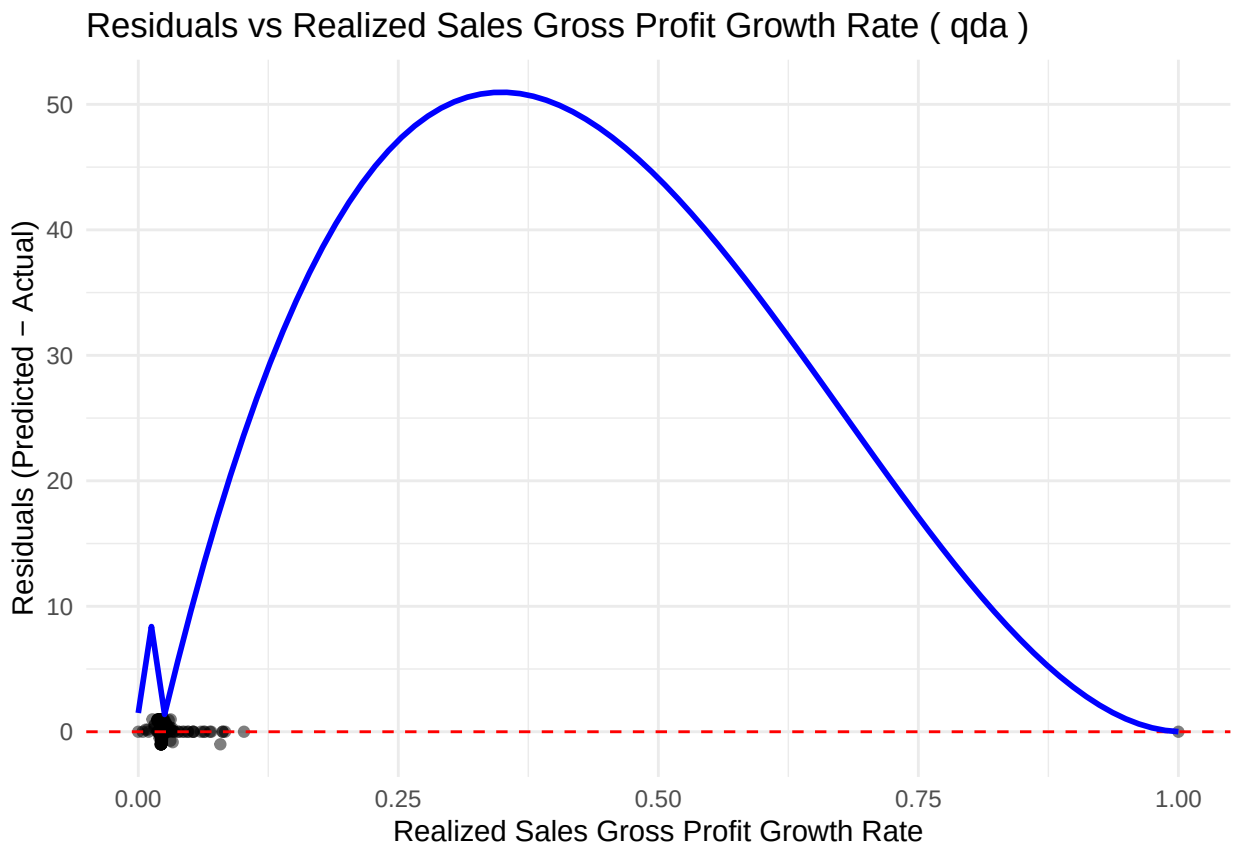

```
## `geom_smooth()` using formula = 'y ~ x'
```



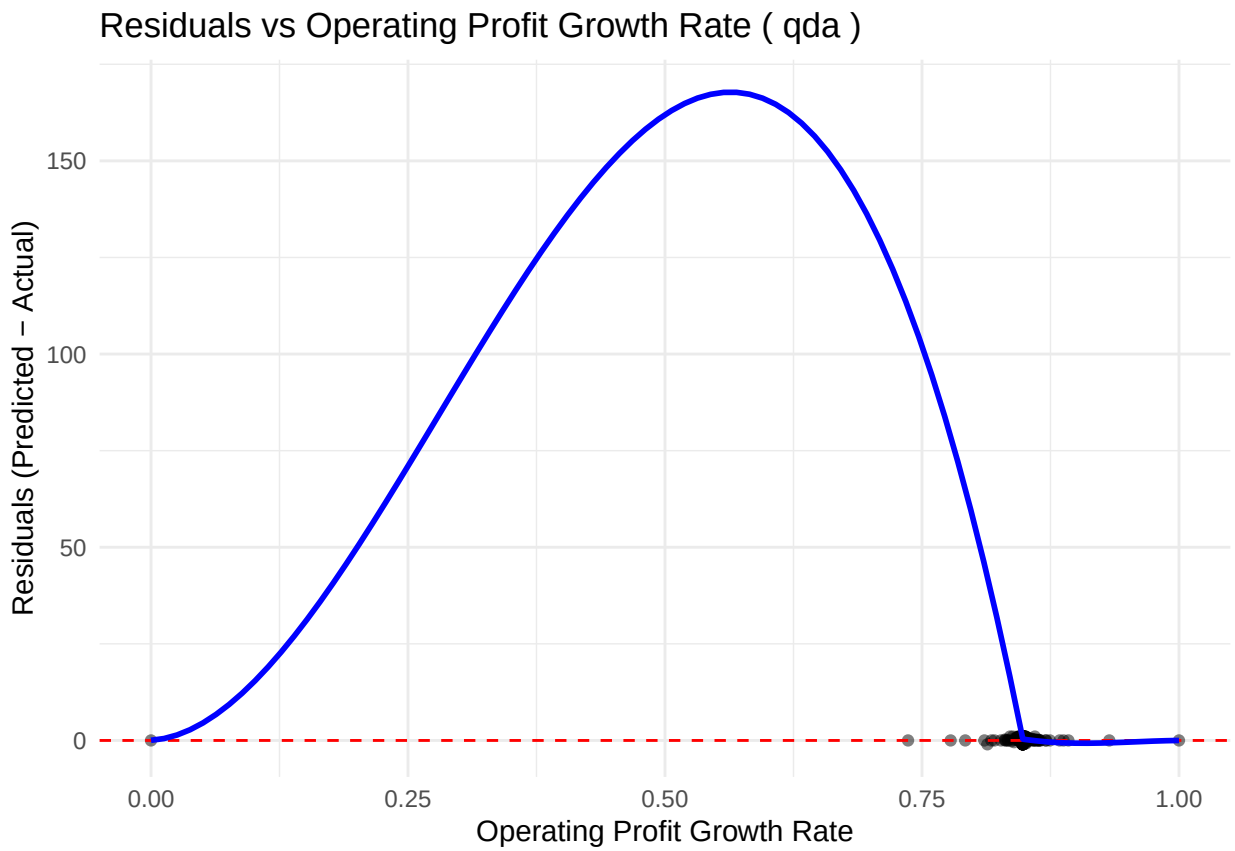

```
## `geom_smooth()` using formula = 'y ~ x'
```



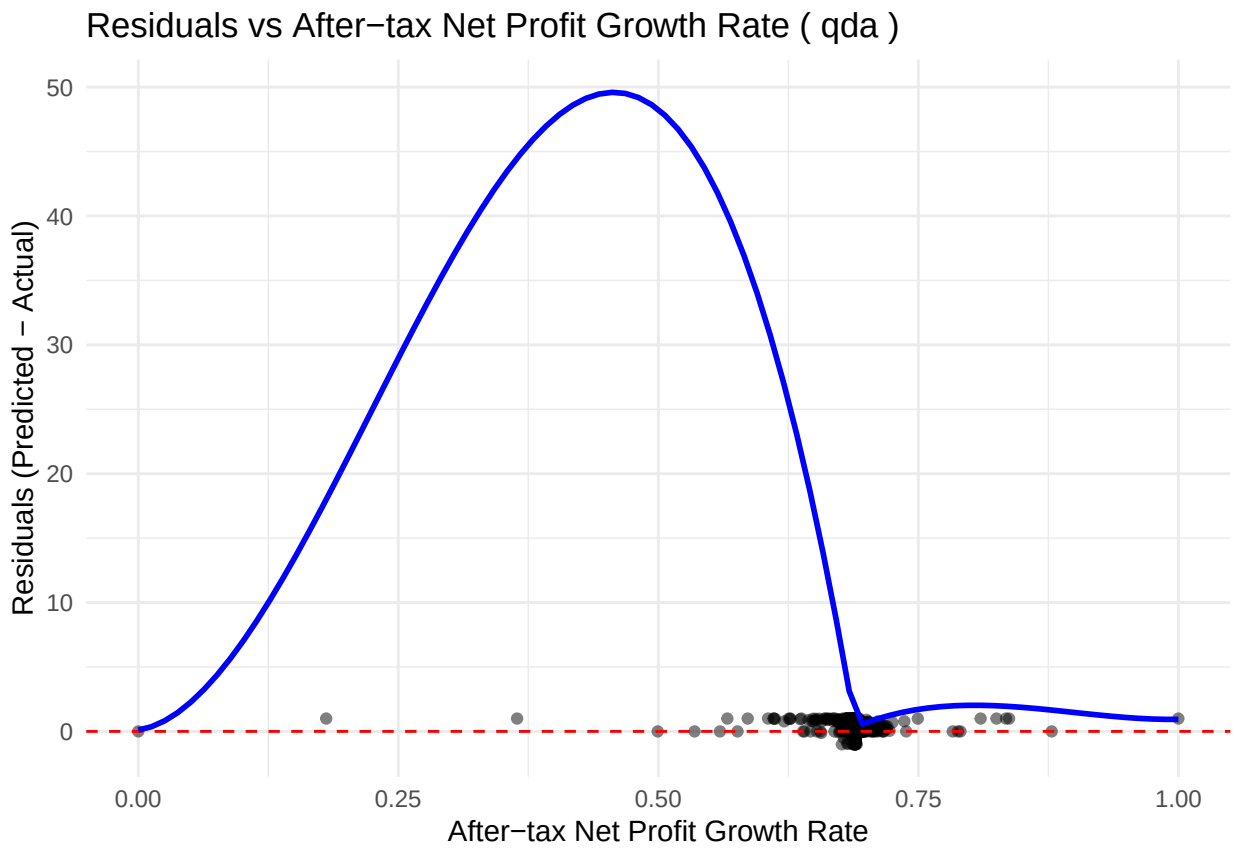

```
## `geom_smooth()` using formula = 'y ~ x'
```



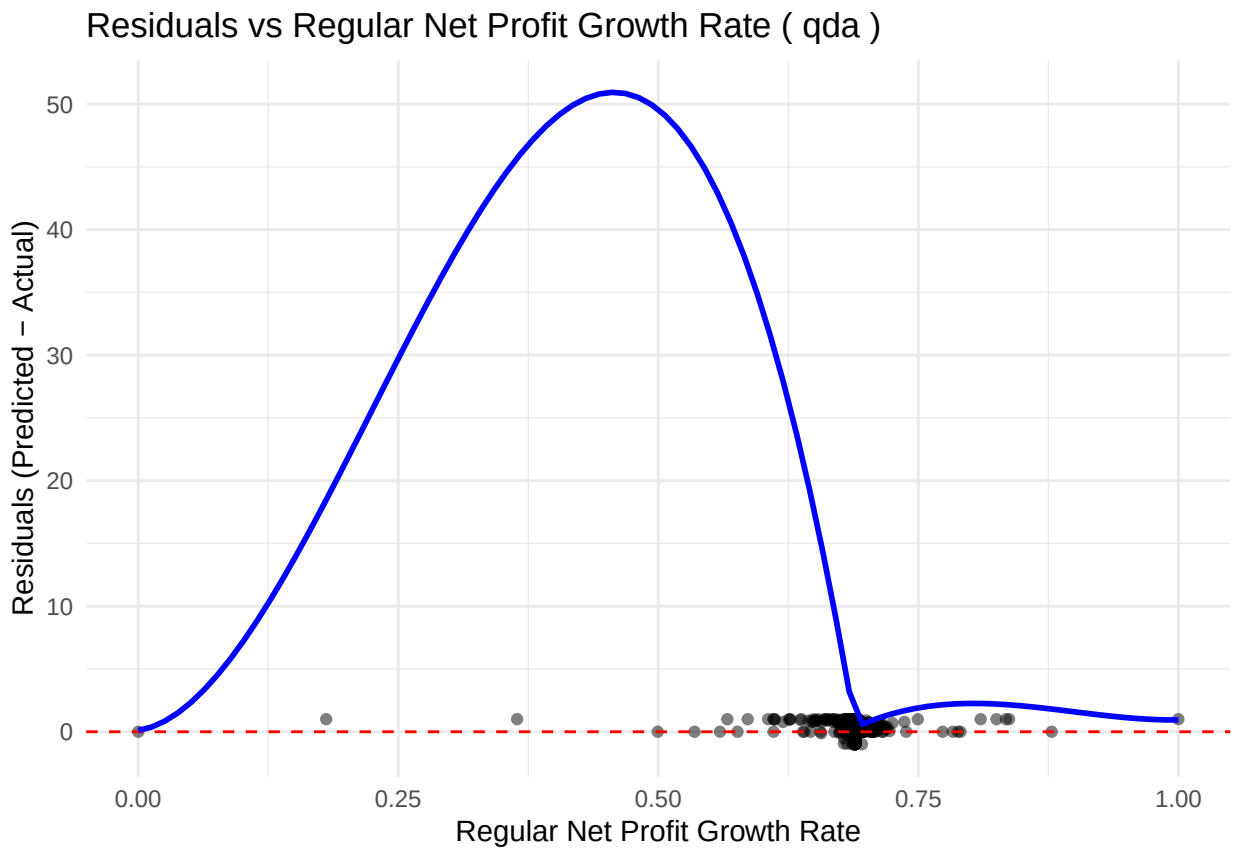

```
## `geom_smooth()` using formula = 'y ~ x'
```



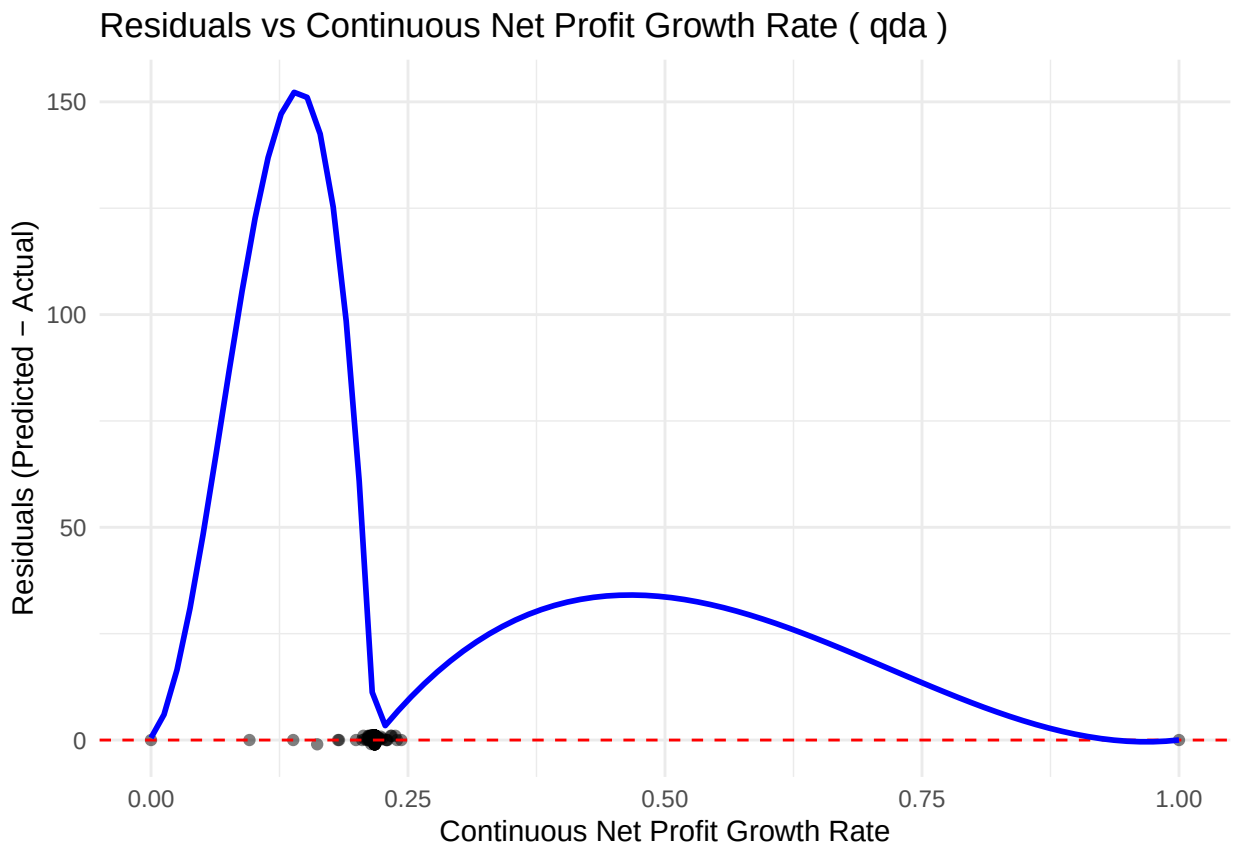

```
## `geom_smooth()` using formula = 'y ~ x'
```



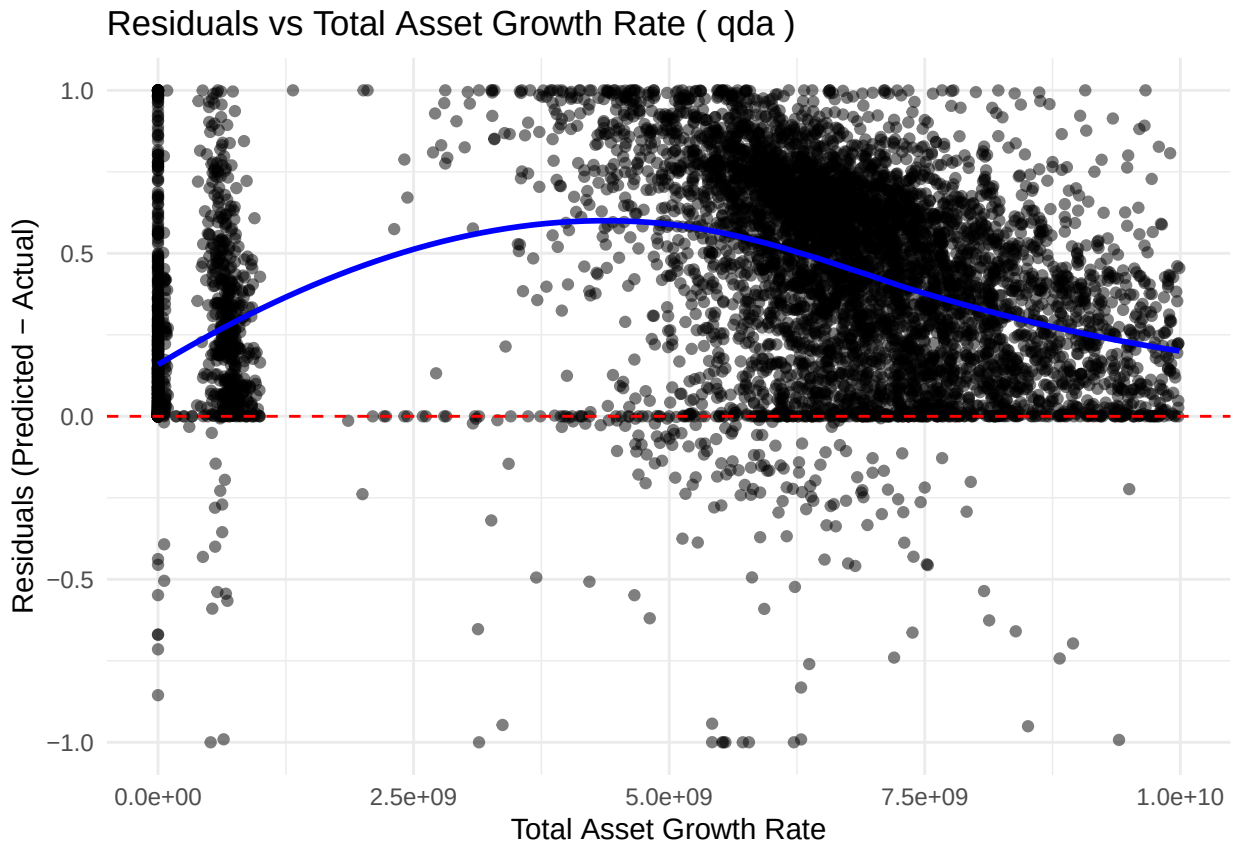

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
## `geom_smooth()` using formula = 'y ~ x'
```




```
## `geom_smooth()` using formula = 'y ~ x'
```



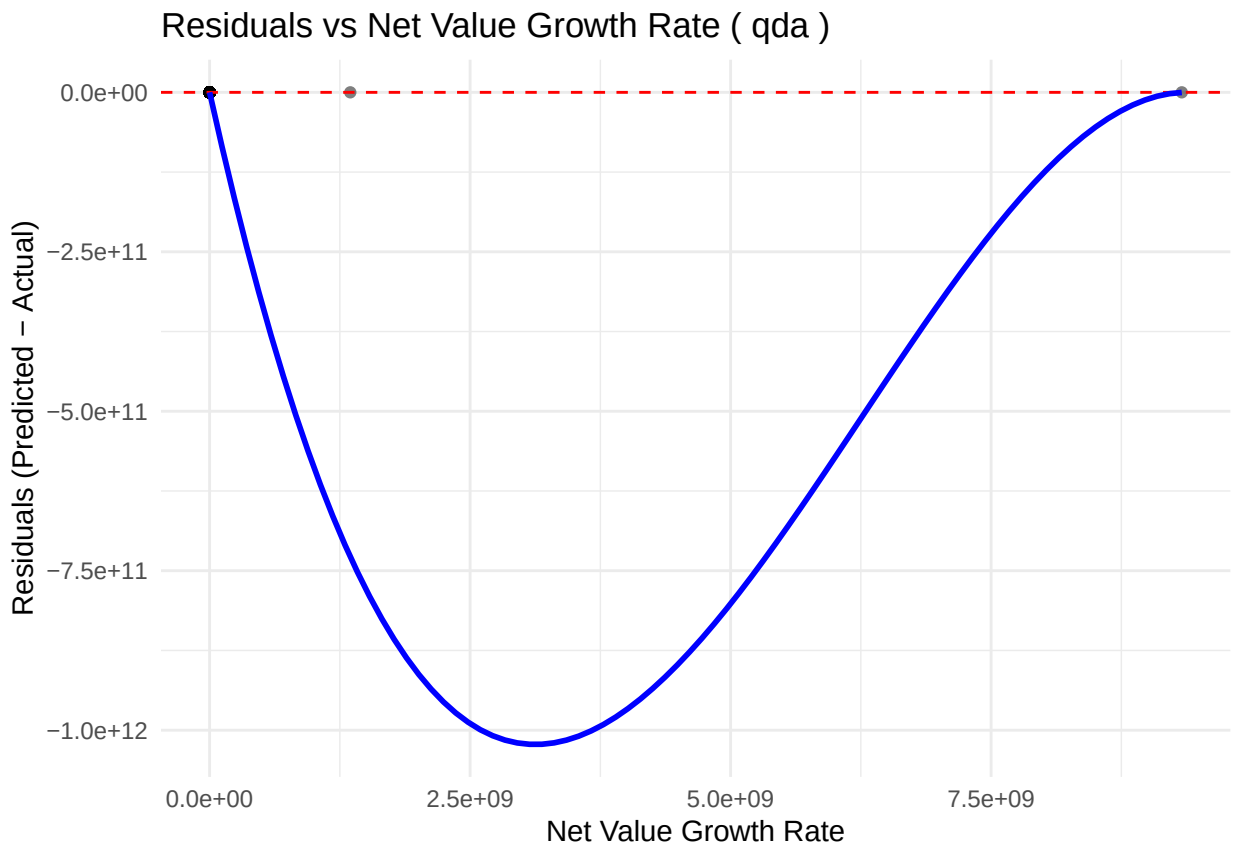

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : pseudoinverse used at -4.665e+07
```

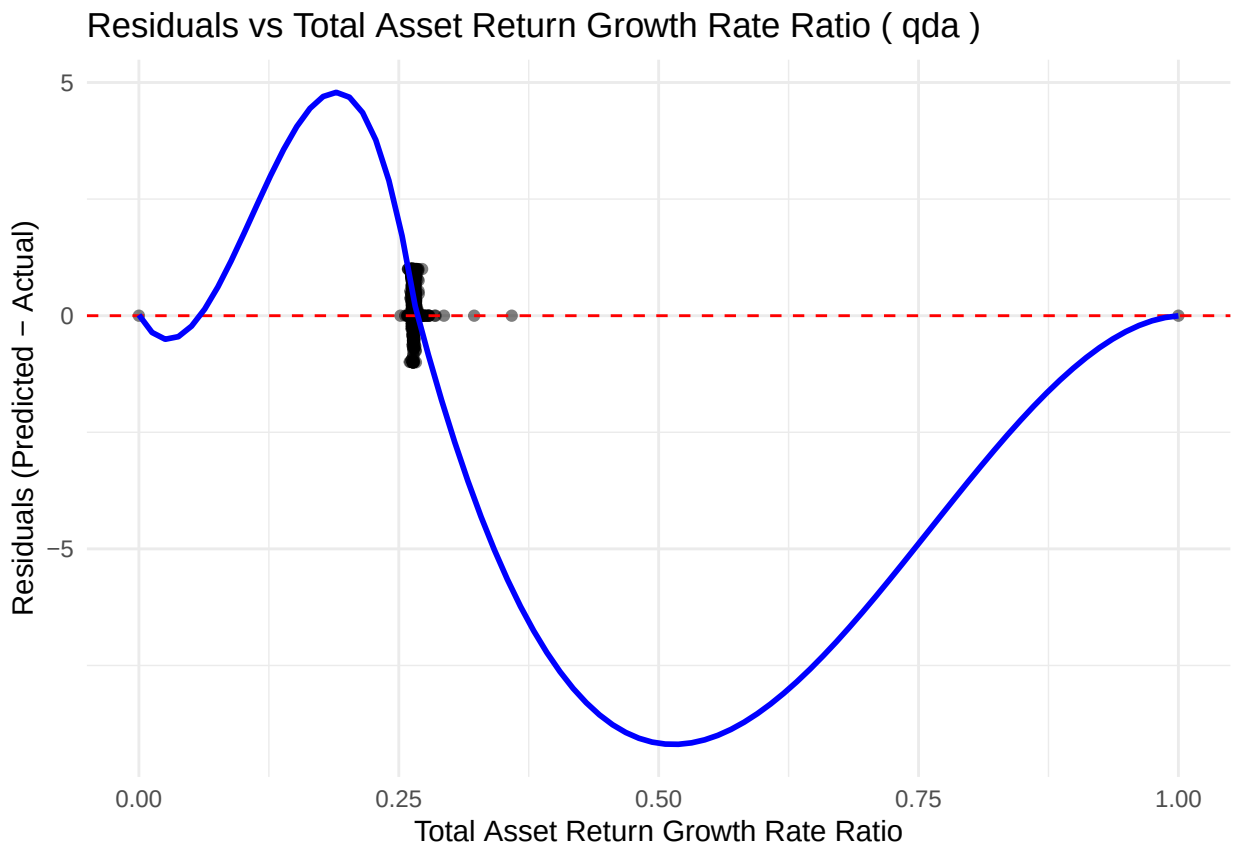
```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : neighborhood radius 4.665e+07
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : reciprocal condition number 5.3318e-15
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : There are other near singularities as well. 8.7922e+19
```




```
## `geom_smooth()` using formula = 'y ~ x'
```



We can clearly observe a pattern for many predictors. This pattern involves residuals getting more significant at certain point (clusters). This implies **heteroscedasticity** for the mentioned predictors.

Saving the observations

After finishing EDA we save some important observations for better accessibility.

```
library(knitr)

kable_correlation <- kable(cor_matrix, format = "markdown")
writeLines(kable_correlation, "../output/tables/independence_matrix.md")
```

Conclusion & Recommendation

By performing EDA, we have performed the following actions:

1) Feature Selection
2) Checking various metrics of basic statistics and scale for all predictors
3) Detecting missing values
4) Detecting outliers
5) Detecting the independence between predictors
6) Observing the distribution for each predictor

During the next phase, **Data Preprocessing**, we recommend to do the following actions:

As some models are sensitive to outliers, we recommend **removing the outliers** if their portion is not substantial. Otherwise we recommend **imputing a placeholder value** (median, mean, etc.).

We recommend removing any attributes that have correlation **higher than 0.7**. Please check the independence matrix in this document.

Some predictors **did not fulfill the condition of homoscedasticity**. Their residuals clearly followed a certain pattern of getting significant at certain point on x axis which clearly looked like clusters. Thus this condition is not fulfilled.

Picking the right Model

- We have also detected high imbalance between target classes (class 0 was much more populated compared to bankrupted companies).
- We recommend some actions to be taken here for each model separately.

Logistic Regression

- Simple, interpretable, and fast.
- Because of high imbalance between target classes, it can be biased towards the majority class. We recommend using **class weights**.
- We could also apply **regularization (like L1, L2, etc.)** to improve performance.

Support Vector Machines (SVM)

- Powerful classifiers, especially for high-dimensional datasets.
- SVMs are sensitive to **class imbalance**. However, **class weighting** can help.
- Using **RBF kernel** might help capture more complex (non-linear) decision boundaries
- Computationally expensive with large datasets.

Linear Discriminant Analysis (LDA) / Quadratic Discriminant Analysis (QDA)

- They expect **normally distributed** data within each class.
- **Highly skewed data** can degrade their performance.
- QDA can **deal with non-linear relationships.**

Random Forests

- Generally **perform better on imbalanced dataset**.
- They can **handle non-linear relationships**.
- They are **computationally expensive**.
- **Parameter tunning** is recommended.

We definitely recommend trying this model!