Credit Risk Analysis

Suresh L. Paul

2022-11-29

1 Research question

Explain the research question. Be concise and explain what your project is about.

In this project, I investigate the relationship between certain key loan characteristics in two credit risk datasets. The first dataset (*crisk*) is a smaller dataset (approximately 35,000 observations) whereas the second dataset (*credit_risk*) is larger simulated dataset (800,000 observations). I establish the relationship between Loan Grade (*loan_grade*) and (i) Loan Interest (*loan_int_rate*), (ii) Personal Income (*person_income*), and (iii) the ratio of Loan Amount to the Personal Income (*loan_percent_income*). Further, I also investigate the relationship between Prior default Status (*cb_person_default_on_file*) and Loan Interest Rate (*loan_int_rate*) sliced across the categories of Loan Grade (*loan_grade*), Home Ownership (*person_home_ownership*), and Loan Intent (*loan_intent*).

2 Data analysis

The smaller dataset (*crisk*) is directly downloaded from the source website. The larger dataset (*credit_risk*) is assimilated by appending three separate datasets - (*credit_risk1*), (*credit_risk2*), and (*credit_risk3*) respectively. The description of variables are as follows: *loan_number* is a unique loan identifier, *person_age* gives the age of the loan recipient, *person_income* is the income of the loan recipient in US dollars, *person_home_ownership* denotes the home ownership status of the loan recipient (i.e., RENT, OWN, MORTGAGE, OTHER), *person_emp_length* length of employment of the loan recipient in months, *loan_intent* provides the purpose of the loan (PERSONAL, EDUCATION, MEDICAL, VENTURE, HOME IMPROVEMENT, and DEBT CONSOLIDATION), *loan_grade* gives the credit rating/grade of loan (i.e., A, B, C, D, E, F, and G), *loan_amnt* is the total loan amount in US dollars, *loan_int_rate* is the effective loan interest rate, *loan_status* is the status of the current loan (0 is active, 1 is default), *cb_person_default_on_file* is indicator that mentions whether the loan recipient has defaulted on a loan previously (Yes/No), *cb_person_cred_hist_length* is the length of credit history, and finally *loan_percent_income* is the total loan amount as a percentage of income.

2.1 Exploratory data analysis

This subsection documents the initial diagnostic evaluation of all datasets.

2.1.1 Loading Libraries

First, I start with installing and loading all required *R* libraries. They include,

- *data.table()*

- *tidyverse()*, which includes *ggplot2()* and *dplyr()*
- *tidyquant()*
- *janitor()*
- *plotly()*
- *gganimate()*
- *gifski()*
- *patchwork()*

Also, the path to datasets, namely *gitpath* and *dirpath*, are initialized.

```
# load all libraries ...

library(data.table)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2

## Warning: package 'ggplot2' was built under R version 4.2.2

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::between()   masks data.table::between()
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x purrr::transpose() masks data.table::transpose()
```

```
library(tidyquant)
```

```
## Loading required package: lubridate

## Warning: package 'lubridate' was built under R version 4.2.2

## Loading required package: timechange

## Warning: package 'timechange' was built under R version 4.2.2

##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:data.table':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year
##
## The following objects are masked from 'package:base':
##
```

```
##     date, intersect, setdiff, union
##
## Loading required package: PerformanceAnalytics
## Loading required package: xts
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
##
##
## Attaching package: 'xts'
##
## The following objects are masked from 'package:dplyr':
##
##     first, last
##
## The following objects are masked from 'package:data.table':
##
##     first, last
##
##
## Attaching package: 'PerformanceAnalytics'
##
## The following object is masked from 'package:graphics':
##
##     legend
##
## Loading required package: quantmod
## Loading required package: TTR
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(plotly)
```

```
## Warning: package 'plotly' was built under R version 4.2.2
```

```
##
## Attaching package: 'plotly'
##
## The following object is masked from 'package:ggplot2':
##
##     last_plot
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following object is masked from 'package:graphics':
##
##     layout
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(gganimate)
```

```
## Warning: package 'gganimate' was built under R version 4.2.2
```

```
library(gifski)
```

```
## Warning: package 'gifski' was built under R version 4.2.2
```

```
library(patchwork)
```

```
## Warning: package 'patchwork' was built under R version 4.2.2
```

```
#-------------------------------------------------------------------------------
# initialize path to files ...
#-------------------------------------------------------------------------------


# Github branch path
gitpath <- "https://raw.githubusercontent.com/sureshlazaruspaul/"

# Github path to files
dirpath <- "/BUS662-practice-datasets/main/credit-risk/"
```

2.1.2 Importing strategy and Descriptive Summary

All datasets are imported using the *fread()* function from the *data.table* library package and imported from online sources.

```
#-------------------------------------------------------------------------------
# Credit Risk
# - sample 1: small dataset
#-------------------------------------------------------------------------------

crisk <- fread(paste0(gitpath, dirpath, "crisk.csv"))

#-------------------------------------------------------------------------------
# Credit Risk
```

```
# - sample 2: large dataset
#------------------------------------------------------------------------------

# read first csv file ...
credit_risk1 <- fread(paste0(gitpath, dirpath, "credit_risk1.csv"))
credit_risk2 <- fread(paste0(gitpath, dirpath, "credit_risk2.csv"))
credit_risk3 <- fread(paste0(gitpath, dirpath, "credit_risk3.csv"))

# create an empty dataframe
credit_risk = data.frame()

# use rbind() to append the imported data to the empty dataframe
credit_risk <- rbind(credit_risk, credit_risk1)
credit_risk <- rbind(credit_risk, credit_risk2)
credit_risk <- rbind(credit_risk, credit_risk3)

#remove credit_risk1 and credit_risk2
rm("credit_risk1", "credit_risk2", "credit_risk3")
```

2.1.2.1 Import

2.1.2.2 Describe

- *head()*

```
head(crisk)
```

```
##    person_age person_income person_home_ownership person_emp_length loan_intent
## 1:         22         59000                  RENT               123    PERSONAL
## 2:         21          9600                   OWN                 5   EDUCATION
## 3:         25          9600              MORTGAGE                 1     MEDICAL
## 4:         23         65500                  RENT                 4     MEDICAL
## 5:         24         54400                  RENT                 8     MEDICAL
## 6:         21          9900                   OWN                 2     VENTURE
##    loan_grade loan_amnt loan_int_rate loan_status loan_percent_income
## 1:          D     35000         16.02           1                0.59
## 2:          B      1000         11.14           0                0.10
## 3:          C      5500         12.87           1                0.57
## 4:          C     35000         15.23           1                0.53
## 5:          C     35000         14.27           1                0.55
## 6:          A      2500          7.14           1                0.25
##    cb_person_default_on_file cb_person_cred_hist_length
## 1:                         Y                          3
## 2:                         N                          2
## 3:                         N                          3
## 4:                         N                          2
## 5:                         Y                          4
## 6:                         N                          2
```

```
head(credit_risk)
```

```
##    loan_number person_age person_income person_home_ownership person_emp_length
## 1:           1         29         65000              MORTGAGE                 5
## 2:           2         36         76000                   OWN                 3
## 3:           3         23         83000                  RENT                10
## 4:           4         28         51000                  RENT                 3
## 5:           5         24         78000                  RENT                 3
## 6:           6         24         54000                  RENT                 8
##    loan_intent loan_grade loan_amnt cb_person_cred_hist_length
## 1:     MEDICAL          C      5000                          2
## 2:   EDUCATION          A     12250                         11
## 3:    PERSONAL          E     11200                          3
## 4:    PERSONAL          A      2400                          2
## 5:   EDUCATION          E      6000                          3
## 6:     VENTURE          A     15000                         10
##    loan_percent_income loan_int_rate loan_status cb_person_default_on_file
## 1:                0.08     13.119272           0                        N
## 2:                0.16      8.552626           0                        N
## 3:                0.13     18.925358           1                        Y
## 4:                0.05      6.874845           0                        Y
## 5:                0.08     17.018009           1                        Y
## 6:                0.28      7.718234           1                        N
```

- *tail()*

```
tail(crisk)
```

```
##    person_age person_income person_home_ownership person_emp_length
## 1:         52         64500                  RENT                 0
## 2:         57         53000              MORTGAGE                 1
## 3:         54        120000              MORTGAGE                 4
## 4:         65         76000                  RENT                 3
## 5:         56        150000              MORTGAGE                 5
## 6:         66         42000                  RENT                 2
##        loan_intent loan_grade loan_amnt loan_int_rate loan_status
## 1:       EDUCATION          B      5000         11.26           0
## 2:        PERSONAL          C      5800         13.16           0
## 3:        PERSONAL          A     17625          7.49           0
## 4: HOMEIMPROVEMENT          B     35000         10.99           1
## 5:        PERSONAL          B     15000         11.48           0
## 6:         MEDICAL          B      6475          9.99           0
##    loan_percent_income cb_person_default_on_file cb_person_cred_hist_length
## 1:                0.08                        N                         20
## 2:                0.11                        N                         30
## 3:                0.15                        N                         19
## 4:                0.46                        N                         28
## 5:                0.10                        N                         26
## 6:                0.15                        N                         30
```

```
tail(credit_risk)
```

```
##    loan_number person_age person_income person_home_ownership person_emp_length
## 1:      799995         46         40000                   OWN                 5
```

```
## 2:      799996         32         68000              MORTGAGE                1
## 3:      799997         24         37000              MORTGAGE               14
## 4:      799998         29         56000                   OWN                8
## 5:      799999         23         43000                  RENT               11
## 6:      800000         30         83000                   OWN                4
##    loan_intent loan_grade loan_amnt cb_person_cred_hist_length
## 1:   EDUCATION          A      3000                          29
## 2:   EDUCATION          A     12000                           4
## 3:   EDUCATION          A      9000                          10
## 4:   EDUCATION          D      9500                           4
## 5:    PERSONAL          A     19000                           9
## 6:     VENTURE          D     21000                           9
##    loan_percent_income loan_int_rate loan_status cb_person_default_on_file
## 1:                0.07      7.343923           0                        Y
## 2:                0.18      8.154219           0                        N
## 3:                0.24      8.440983           0                        N
## 4:                0.17     16.288348           1                        N
## 5:                0.44      7.360357           0                        N
## 6:                0.25     14.351446           0                        N
```

- *glimpse()*

```
glimpse(crisk)
```

```
## Rows: 32,581
## Columns: 12
## $ person_age                <int> 22, 21, 25, 23, 24, 21, 26, 24, 24, 21, 22,~
## $ person_income             <int> 59000, 9600, 9600, 65500, 54400, 9900, 7710~
## $ person_home_ownership     <chr> "RENT", "OWN", "MORTGAGE", "RENT", "RENT", ~
## $ person_emp_length         <int> 123, 5, 1, 4, 8, 2, 8, 5, 8, 6, 6, 2, 2, 4,~
## $ loan_intent               <chr> "PERSONAL", "EDUCATION", "MEDICAL", "MEDICA~
## $ loan_grade                <chr> "D", "B", "C", "C", "C", "A", "B", "B", "A"~
## $ loan_amnt                 <int> 35000, 1000, 5500, 35000, 35000, 2500, 3500~
## $ loan_int_rate             <dbl> 16.02, 11.14, 12.87, 15.23, 14.27, 7.14, 12~
## $ loan_status               <int> 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0~
## $ loan_percent_income       <dbl> 0.59, 0.10, 0.57, 0.53, 0.55, 0.25, 0.45, 0~
## $ cb_person_default_on_file <chr> "Y", "N", "N", "N", "Y", "N", "N", "N", "N"~
## $ cb_person_cred_hist_length <int> 3, 2, 3, 2, 4, 2, 3, 4, 2, 3, 4, 2, 2, 4, 4~
```

```
glimpse(credit_risk)
```

```
## Rows: 800,000
## Columns: 13
## $ loan_number               <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, ~
## $ person_age                <int> 29, 36, 23, 28, 24, 24, 25, 30, 26, 32, 26,~
## $ person_income             <dbl> 65000, 76000, 83000, 51000, 78000, 54000, 2~
## $ person_home_ownership     <chr> "MORTGAGE", "OWN", "RENT", "RENT", "RENT", ~
## $ person_emp_length         <int> 5, 3, 10, 3, 3, 8, 3, 1, 1, 5, 0, 3, 2, 11,~
## $ loan_intent               <chr> "MEDICAL", "EDUCATION", "PERSONAL", "PERSON~
## $ loan_grade                <chr> "C", "A", "E", "A", "E", "A", "C", "D", "A"~
## $ loan_amnt                 <int> 5000, 12250, 11200, 2400, 6000, 15000, 3000~
## $ cb_person_cred_hist_length <int> 2, 11, 3, 2, 3, 10, 3, 10, 4, 3, 7, 3, 4, 8~
```

```
## $ loan_percent_income        <dbl> 0.08, 0.16, 0.13, 0.05, 0.08, 0.28, 0.14, 0~
## $ loan_int_rate              <dbl> 13.119272, 8.552626, 18.925358, 6.874845, 1~
## $ loan_status                <int> 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1~
## $ cb_person_default_on_file  <chr> "N", "N", "Y", "Y", "Y", "N", "N", "Y", "N"~
```

- *str()*

```
str(crisk)
```

```
## Classes 'data.table' and 'data.frame':   32581 obs. of  12 variables:
##  $ person_age                : int  22 21 25 23 24 21 26 24 24 21 ...
##  $ person_income             : int  59000 9600 9600 65500 54400 9900 77100 78956 83000 10000 ...
##  $ person_home_ownership     : chr  "RENT" "OWN" "MORTGAGE" "RENT" ...
##  $ person_emp_length         : int  123 5 1 4 8 2 8 5 8 6 ...
##  $ loan_intent               : chr  "PERSONAL" "EDUCATION" "MEDICAL" "MEDICAL" ...
##  $ loan_grade                : chr  "D" "B" "C" "C" ...
##  $ loan_amnt                 : int  35000 1000 5500 35000 35000 2500 35000 35000 35000 1600 ...
##  $ loan_int_rate             : num  16 11.1 12.9 15.2 14.3 ...
##  $ loan_status               : int  1 0 1 1 1 1 1 1 1 1 ...
##  $ loan_percent_income       : num  0.59 0.1 0.57 0.53 0.55 0.25 0.45 0.44 0.42 0.16 ...
##  $ cb_person_default_on_file : chr  "Y" "N" "N" "N" ...
##  $ cb_person_cred_hist_length: int  3 2 3 2 4 2 3 4 2 3 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
str(credit_risk)
```

```
## Classes 'data.table' and 'data.frame':   800000 obs. of  13 variables:
##  $ loan_number               : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ person_age                : int  29 36 23 28 24 24 25 30 26 32 ...
##  $ person_income             : num  65000 76000 83000 51000 78000 54000 21000 100000 37000 37000 ...
##  $ person_home_ownership     : chr  "MORTGAGE" "OWN" "RENT" "RENT" ...
##  $ person_emp_length         : int  5 3 10 3 3 8 3 1 1 5 ...
##  $ loan_intent               : chr  "MEDICAL" "EDUCATION" "PERSONAL" "PERSONAL" ...
##  $ loan_grade                : chr  "C" "A" "E" "A" ...
##  $ loan_amnt                 : int  5000 12250 11200 2400 6000 15000 3000 7000 6000 4800 ...
##  $ cb_person_cred_hist_length: int  2 11 3 2 3 10 3 10 4 3 ...
##  $ loan_percent_income       : num  0.08 0.16 0.13 0.05 0.08 0.28 0.14 0.07 0.16 0.13 ...
##  $ loan_int_rate             : num  13.12 8.55 18.93 6.87 17.02 ...
##  $ loan_status               : int  0 0 1 0 1 1 1 0 0 0 ...
##  $ cb_person_default_on_file : chr  "N" "N" "Y" "Y" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

- *class()*

```
class(crisk)
```

```
## [1] "data.table" "data.frame"
```

```
class(credit_risk)
```

```
## [1] "data.table" "data.frame"
```

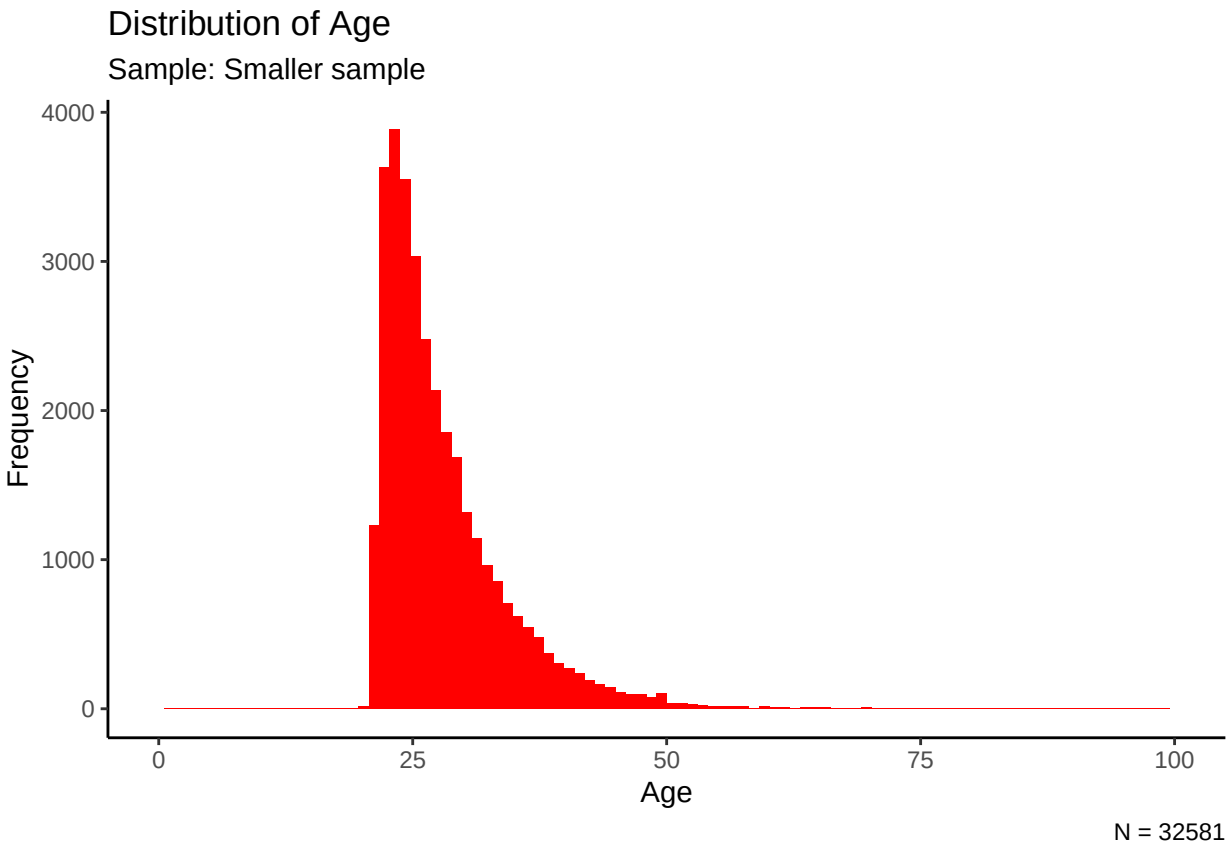
2.1.2.3 Descriptive plots - histogram()

- small dataset *crisk*
- variable: *person_age*

```
# canvas
canvas <- crisk %>%
  ggplot() +
  theme_classic() +
  xlim(0,100) +
  labs(
    title = "Distribution of Age",
    subtitle = "Sample: Smaller sample",
    caption = paste("N =", nrow(crisk)),
    x = "Age",
    y = "Frequency"
  )

# plot
canvas +
  geom_histogram(
    aes(
      x=person_age
    ),
    bins = 100,
    fill = "red"
  )
```

```
## Warning: Removed 5 rows containing non-finite values (‘stat_bin()‘).
```

```
## Warning: Removed 2 rows containing missing values (‘geom_bar()‘).
```

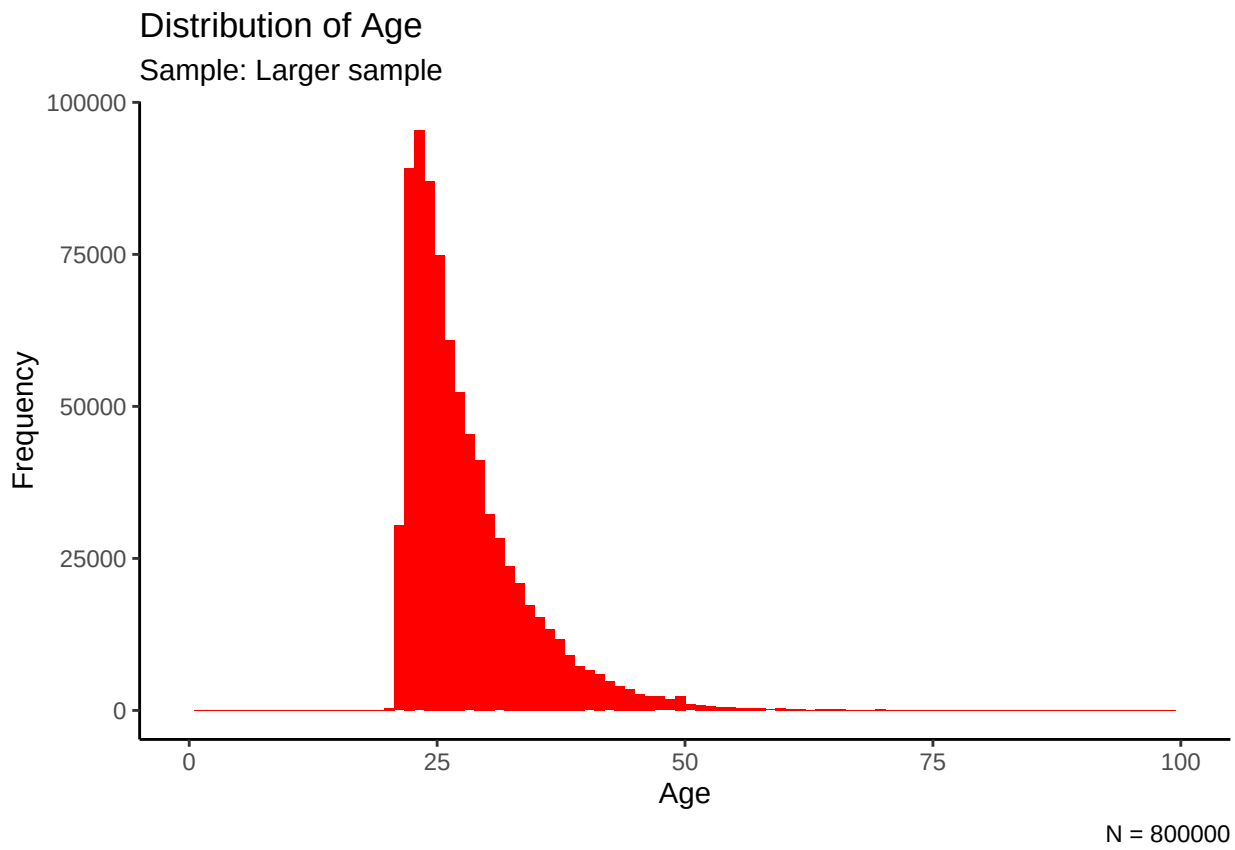


Statistic	Value
Min	20
P1	21
P5	22
P10	22
Q1 = P25	23
Mean	28
Median	26
Q3 = P75	30
P90	36
P95	40
P99	50
Max	144
Std. Dev	6

- large dataset *credit_risk*
- variable: *person_age*

```r
# canvas
canvas <- credit_risk %>%
  ggplot() +
  theme_classic() +
  xlim(0,100) +
  labs(
    title = "Distribution of Age",
    subtitle = "Sample: Larger sample",
    caption = paste("N =", nrow(credit_risk)),
    x = "Age",
    y = "Frequency"
  )

# plot
canvas +
  geom_histogram(
    aes(
      x=person_age
    ),
    bins = 100,
    fill = "red"
  )
```

```
## Warning: Removed 2 rows containing missing values (‘geom_bar()‘).
```



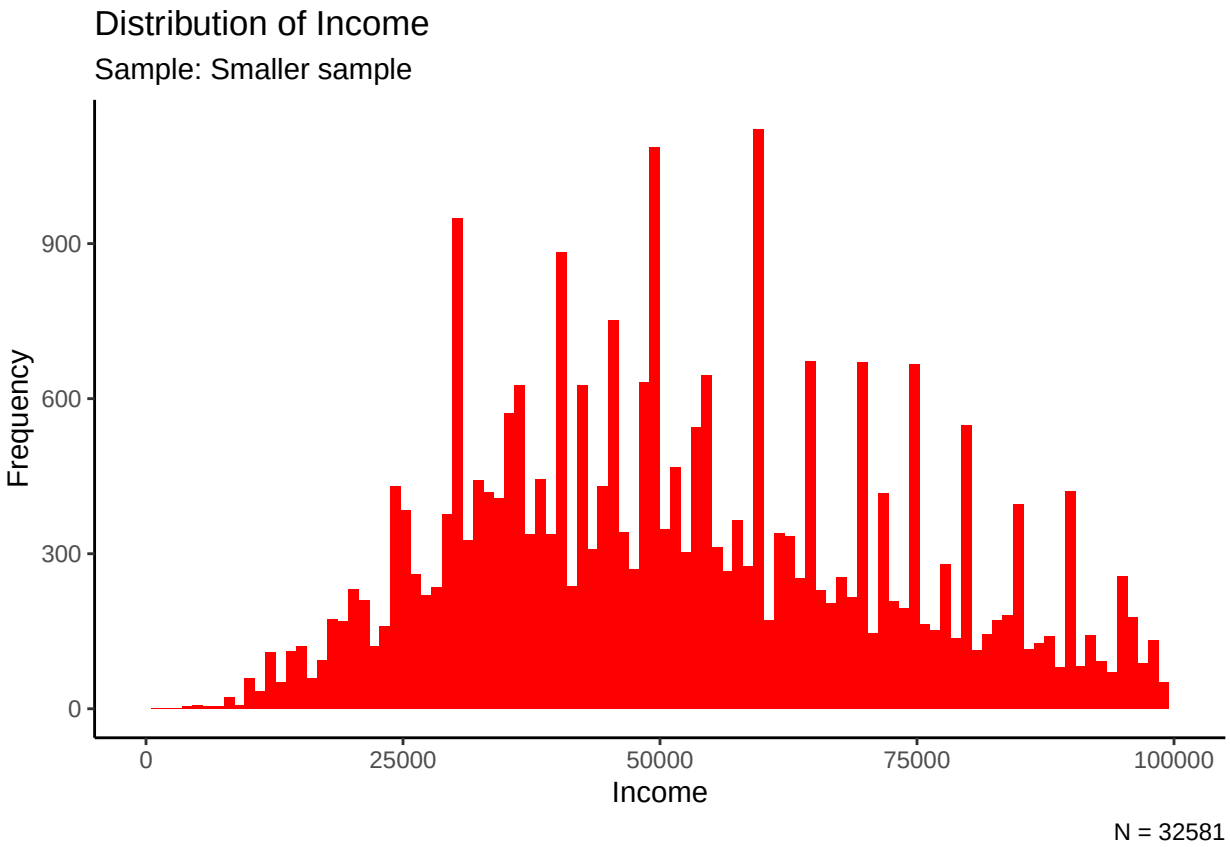
Statistic	Value
Min	20
P1	21
P5	22
P10	22
Q1 = P25	23
Mean	28
Median	26
Q3 = P75	30
P90	36
P95	40
P99	50
Max	94
Std. Dev	6

- small dataset *crisk*
  - variable: *person_income*

```
# canvas
canvas <- crisk %>%
  ggplot() +
  theme_classic() +
  xlim(0,100000) +
  labs(
    title = "Distribution of Income",
    subtitle = "Sample: Smaller sample",
    caption = paste("N =", nrow(crisk)),
    x = "Income",
    y = "Frequency"
  )

# plot
canvas +
  geom_histogram(
    aes(
      x=person_income
    ),
    bins = 100,
    fill = "red"
  )
```

```
## Warning: Removed 4207 rows containing non-finite values (`stat_bin()`).
```

```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```



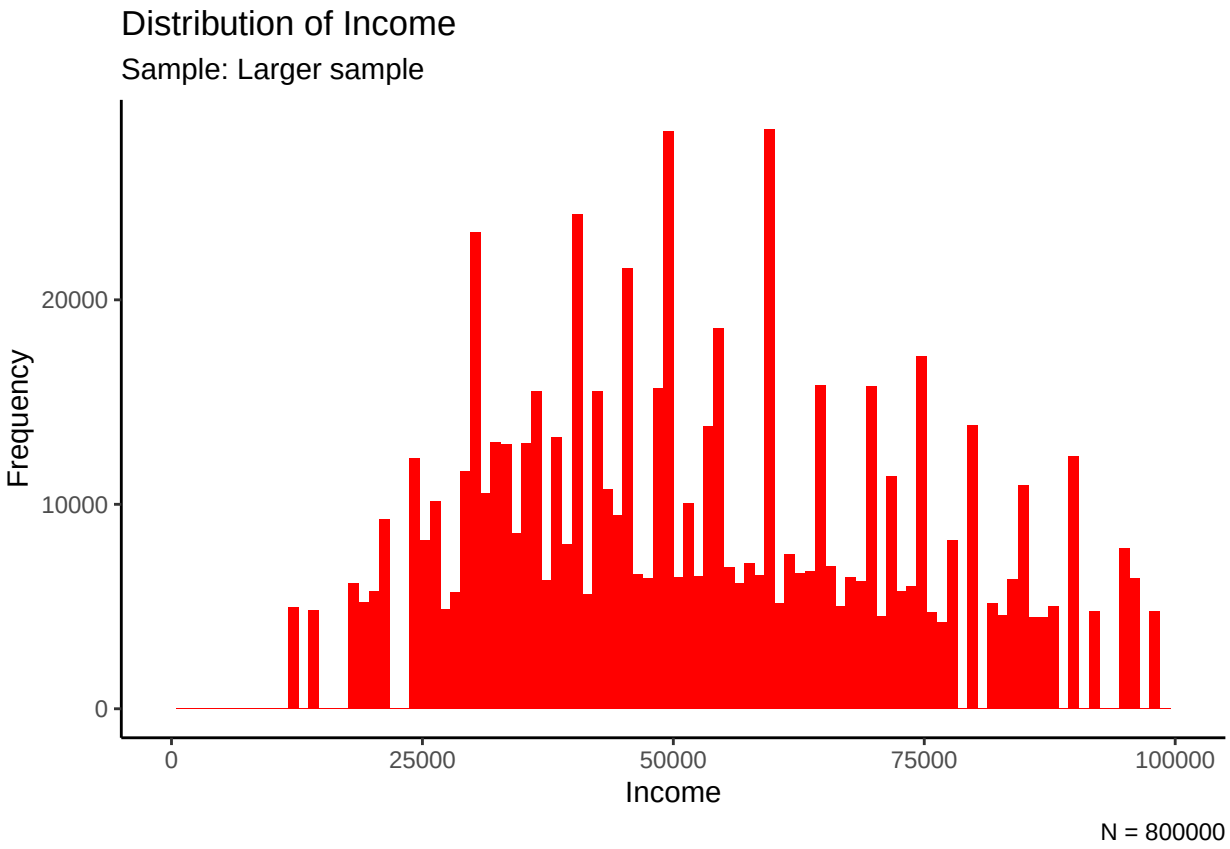
Statistic	Value
Min	4000
P1	1.44×10^4
P5	2.288×10^4
P10	2.859×10^4
Q1 = P25	3.85×10^4
Mean	6.6075×10^4
Median	5.5×10^4
Q3 = P75	7.92×10^4
P90	1.10004×10^5
P95	1.38×10^5
P99	2.252×10^5
Max	6×10^6
Std. Dev	6.1983×10^4

- large dataset *credit_risk*
  - variable: *person_income*

```r
# canvas
canvas <- credit_risk %>%
  ggplot() +
  theme_classic() +
  xlim(0,100000) +
  labs(
    title = "Distribution of Income",
    subtitle = "Sample: Larger sample",
    caption = paste("N =", nrow(credit_risk)),
    x = "Income",
    y = "Frequency"
  )

# plot
canvas +
  geom_histogram(
    aes(
      x=person_income
    ),
    bins = 100,
    fill = "red"
  )
```

```
## Warning: Removed 85181 rows containing non-finite values (‘stat_bin()‘).
```

```
## Warning: Removed 2 rows containing missing values (‘geom_bar()‘).
```

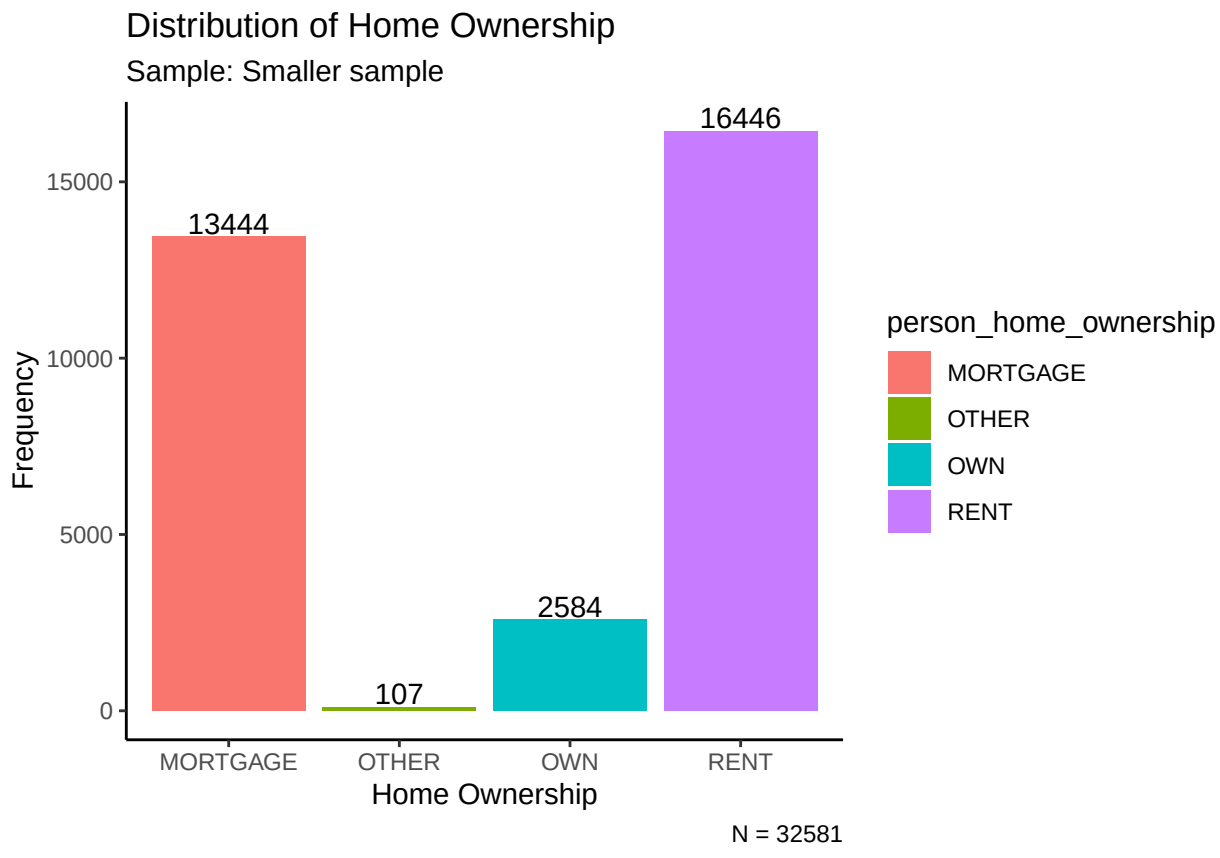


Statistic	Value
Min	1.2×10^4
P1	1.44×10^4
P5	2.4×10^4
P10	2.88×10^4
Q1 = P25	3.84×10^4
Mean	6.2312×10^4
Median	5.5×10^4
Q3 = P75	7.7×10^4
P90	1.05×10^5
P95	1.25×10^5
P99	2×10^5
Max	2.5×10^5
Std. Dev	3.5248×10^4

- small dataset *crisk*
  - variable: *person_home_ownership*

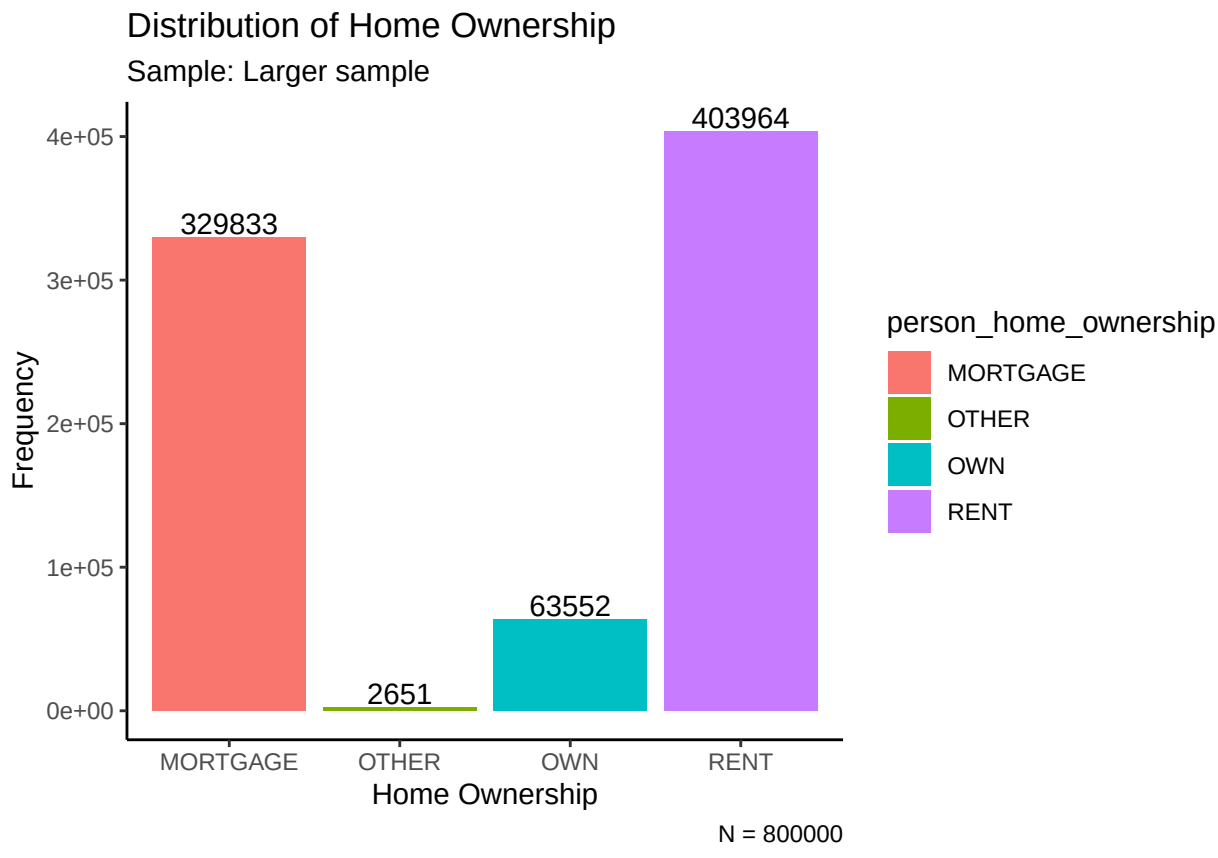
```
# canvas
canvas <- crisk %>%
  ggplot(
    aes(
      x=person_home_ownership,
      fill=person_home_ownership
    )
  ) +
  geom_text(stat = "count",
            aes(label = ..count..),
            vjust = -0.15) +
  theme_classic() +
  labs(
    title = "Distribution of Home Ownership",
    subtitle = "Sample: Smaller sample",
    caption = paste("N =", nrow(crisk)),
    x = "Home Ownership",
    y = "Frequency"
  )

# plot
canvas + geom_bar()
```

```
## Warning: The dot-dot notation (‘..count..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(count)‘ instead.
```



- large dataset *credit_risk*
  - variable: *person_home_ownership*

```
# canvas
canvas <- credit_risk %>%
  ggplot(
    aes(
      x=person_home_ownership,
      fill=person_home_ownership
    )
  ) +
  geom_text(stat = "count",
            aes(label = ..count..),
            vjust = -0.15) +
  theme_classic() +
  labs(
    title = "Distribution of Home Ownership",
    subtitle = "Sample: Larger sample",
    caption = paste("N =", nrow(credit_risk)),
    x = "Home Ownership",
    y = "Frequency"
  )

# plot
canvas + geom_bar()
```

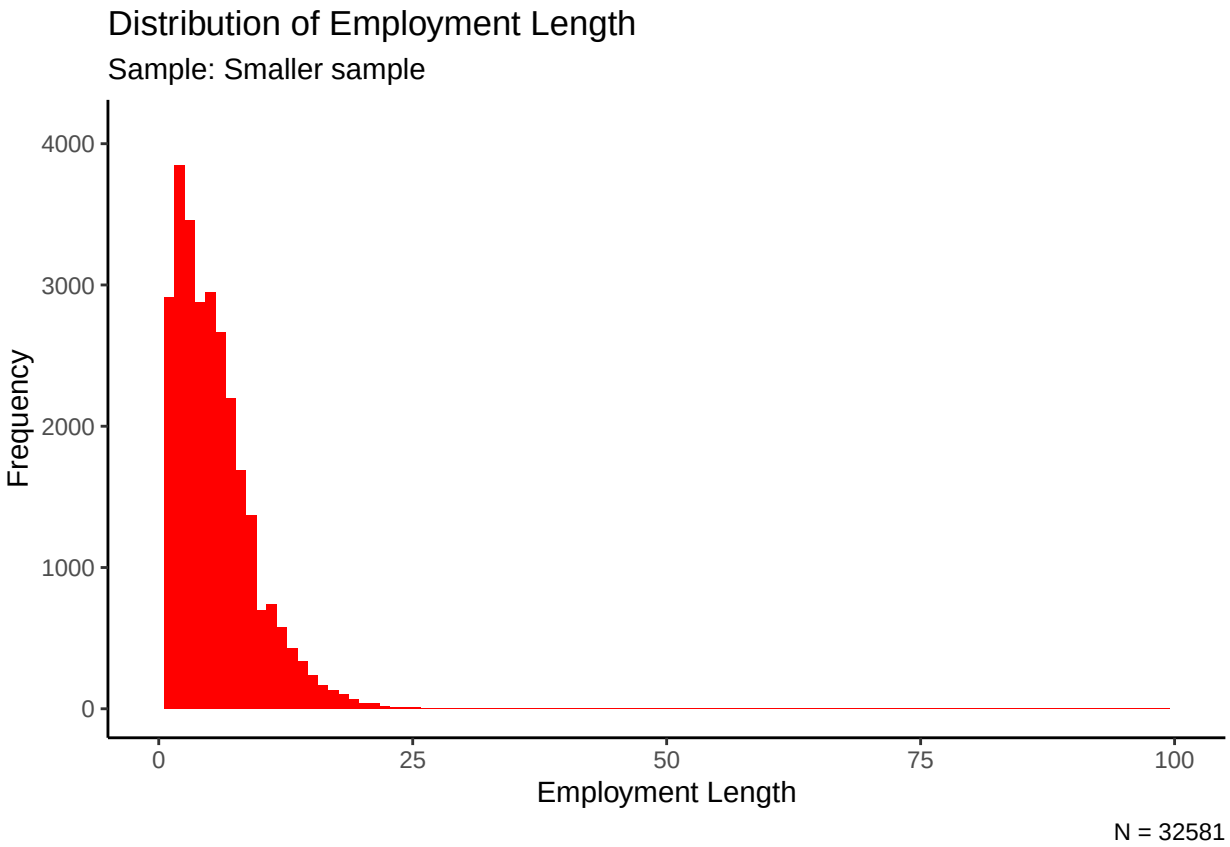


- small dataset *crisk*
- variable: *person_emp_length*

```
# canvas
canvas <- crisk %>%
  ggplot() +
  theme_classic() +
  xlim(0,100) +
  labs(
    title = "Distribution of Employment Length",
    subtitle = "Sample: Smaller sample",
    caption = paste("N =", nrow(crisk)),
    x = "Employment Length",
    y = "Frequency"
  )

# plot
canvas +
  geom_histogram(
    aes(
      x=person_emp_length
    ),
    bins = 100,
    fill = "red"
  )
```

```
## Warning: Removed 897 rows containing non-finite values (`stat_bin()`).
```

```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```

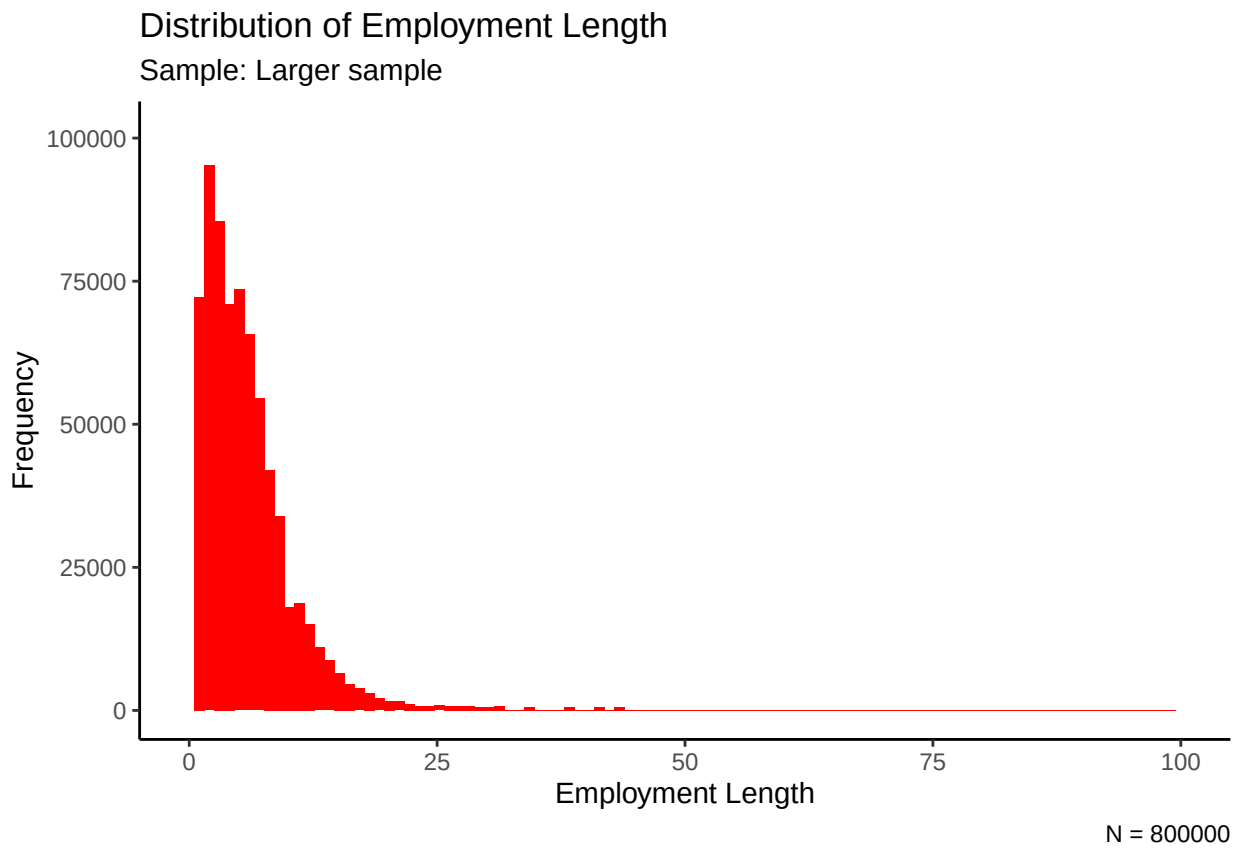


Statistic	Value
Min	0
P1	0
P5	0
P10	0
Q1 = P25	2
Mean	5
Median	4
Q3 = P75	7
P90	10
P95	13
P99	18
Max	123
Std. Dev	4

- large dataset *credit_risk*
- variable: *person_emp_length*

```r
# canvas
canvas <- credit_risk %>%
  ggplot() +
  theme_classic() +
  xlim(0,100) +
  labs(
    title = "Distribution of Employment Length",
    subtitle = "Sample: Larger sample",
    caption = paste("N =", nrow(credit_risk)),
    x = "Employment Length",
    y = "Frequency"
  )

# plot
canvas +
  geom_histogram(
    aes(
      x=person_emp_length
    ),
    bins = 100,
    fill = "red"
  )
```

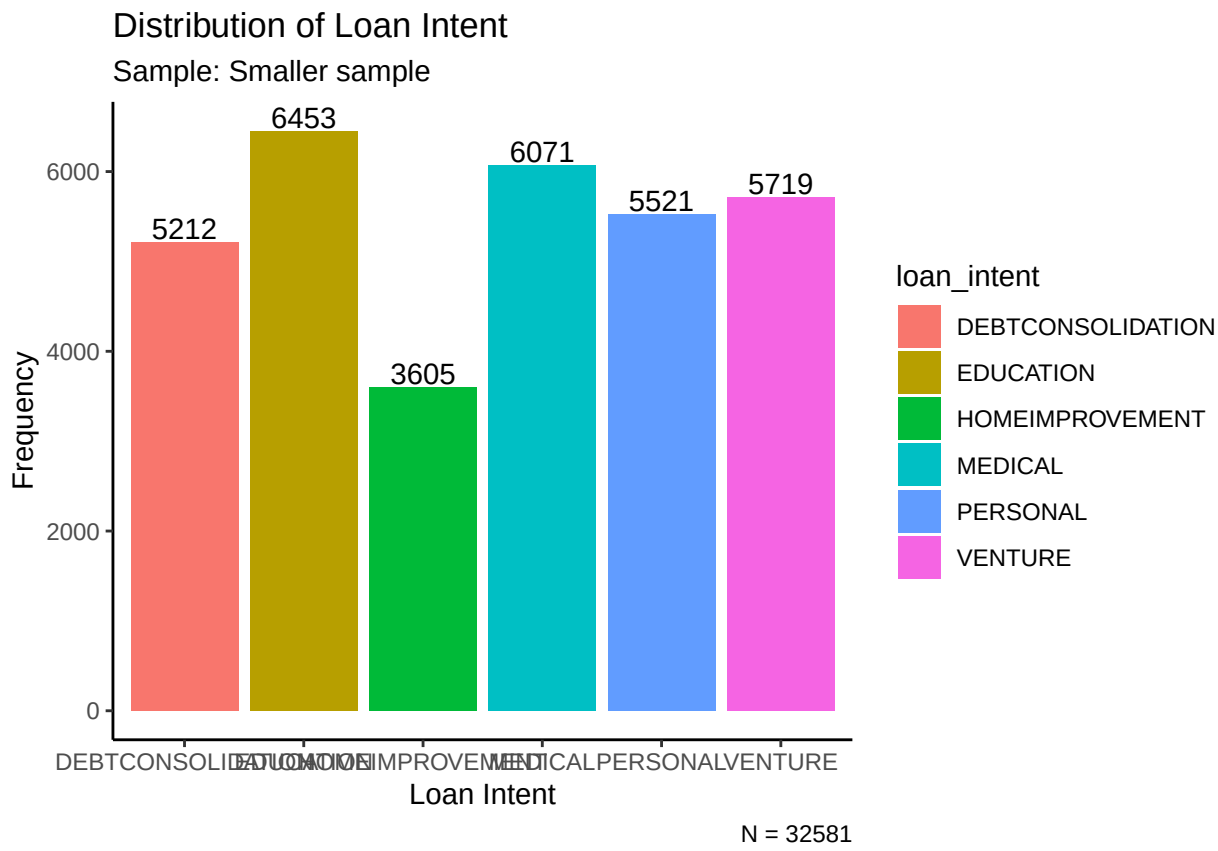
```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```



Statistic	Value
Min	0
P1	0
P5	0
P10	0
Q1 = P25	2
Mean	5
Median	4
Q3 = P75	7
P90	11
P95	14
P99	24
Max	43
Std. Dev	5

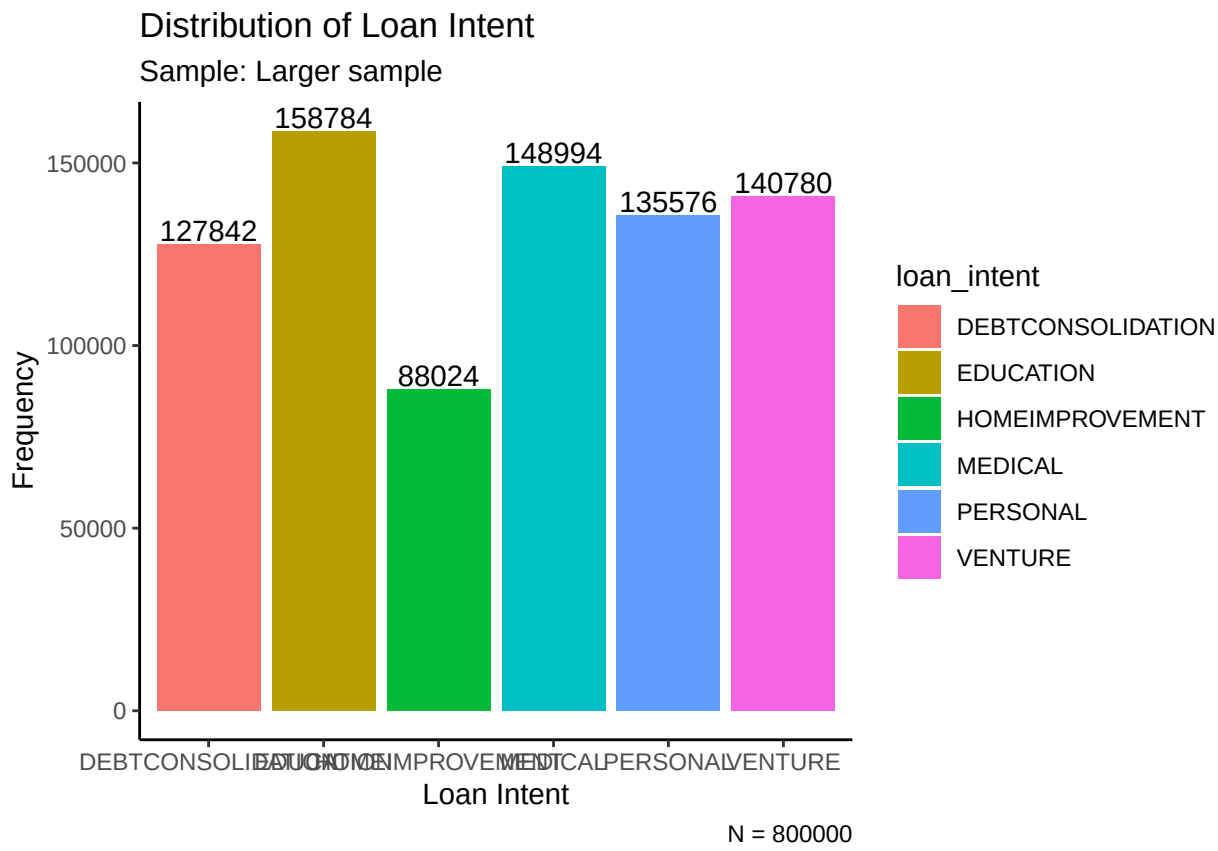
- small dataset *crisk*
- variable: *loan_intent*

```
# canvas
canvas <- crisk %>%
  ggplot(
    aes(
      x=loan_intent,
      fill=loan_intent
    )
  ) +
  geom_text(stat = "count",
            aes(label = ..count..),
            vjust = -0.15) +
  theme_classic() +
  labs(
    title = "Distribution of Loan Intent",
    subtitle = "Sample: Smaller sample",
    caption = paste("N =", nrow(crisk)),
    x = "Loan Intent",
    y = "Frequency"
  )

# plot
canvas + geom_bar()
```



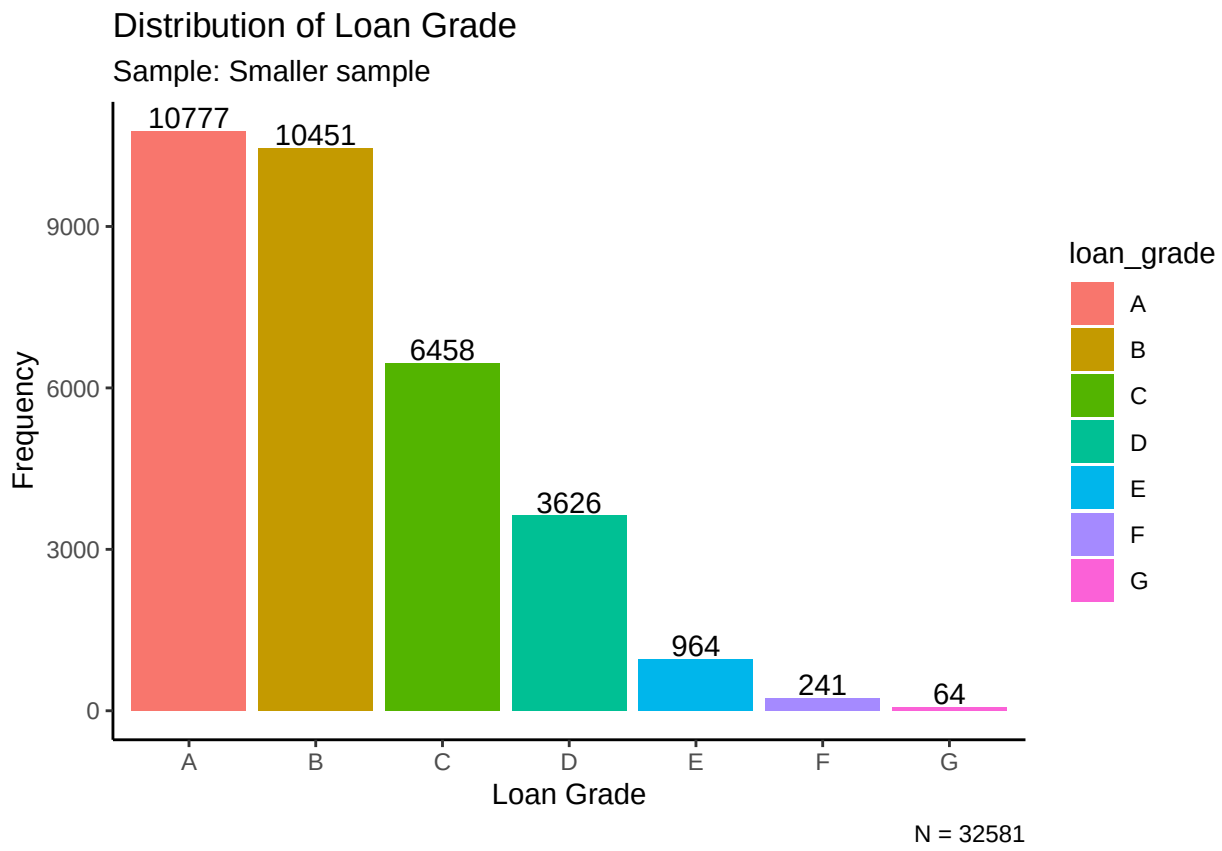
- large dataset *credit_risk*
- variable: *loan_intent*

```
# canvas
canvas <- credit_risk %>%
  ggplot(
    aes(
      x=loan_intent,
      fill=loan_intent
    )
  ) +
  geom_text(stat = "count",
            aes(label = ..count..),
            vjust = -0.15) +
  theme_classic() +
  labs(
    title = "Distribution of Loan Intent",
    subtitle = "Sample: Larger sample",
    caption = paste("N =", nrow(credit_risk)),
    x = "Loan Intent",
    y = "Frequency"
  )

# plot
canvas + geom_bar()
```



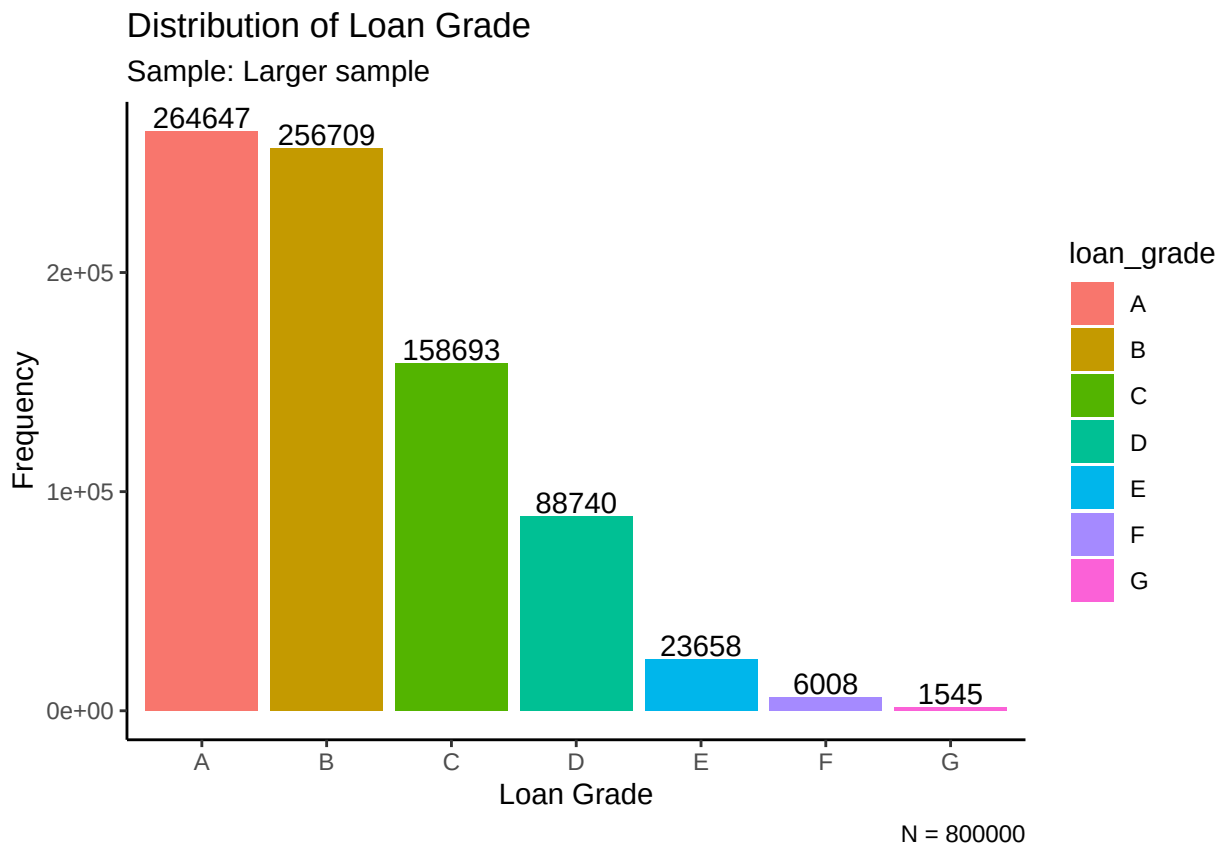
- small dataset *crisk*
- variable: *loan_grade*

```
# canvas
canvas <- crisk %>%
  ggplot(
    aes(
      x=loan_grade,
      fill=loan_grade
    )
  ) +
  geom_text(stat = "count",
            aes(label = ..count..),
            vjust = -0.15) +
  theme_classic() +
  labs(
    title = "Distribution of Loan Grade",
    subtitle = "Sample: Smaller sample",
    caption = paste("N =", nrow(crisk)),
    x = "Loan Grade",
    y = "Frequency"
  )

# plot
canvas + geom_bar()
```



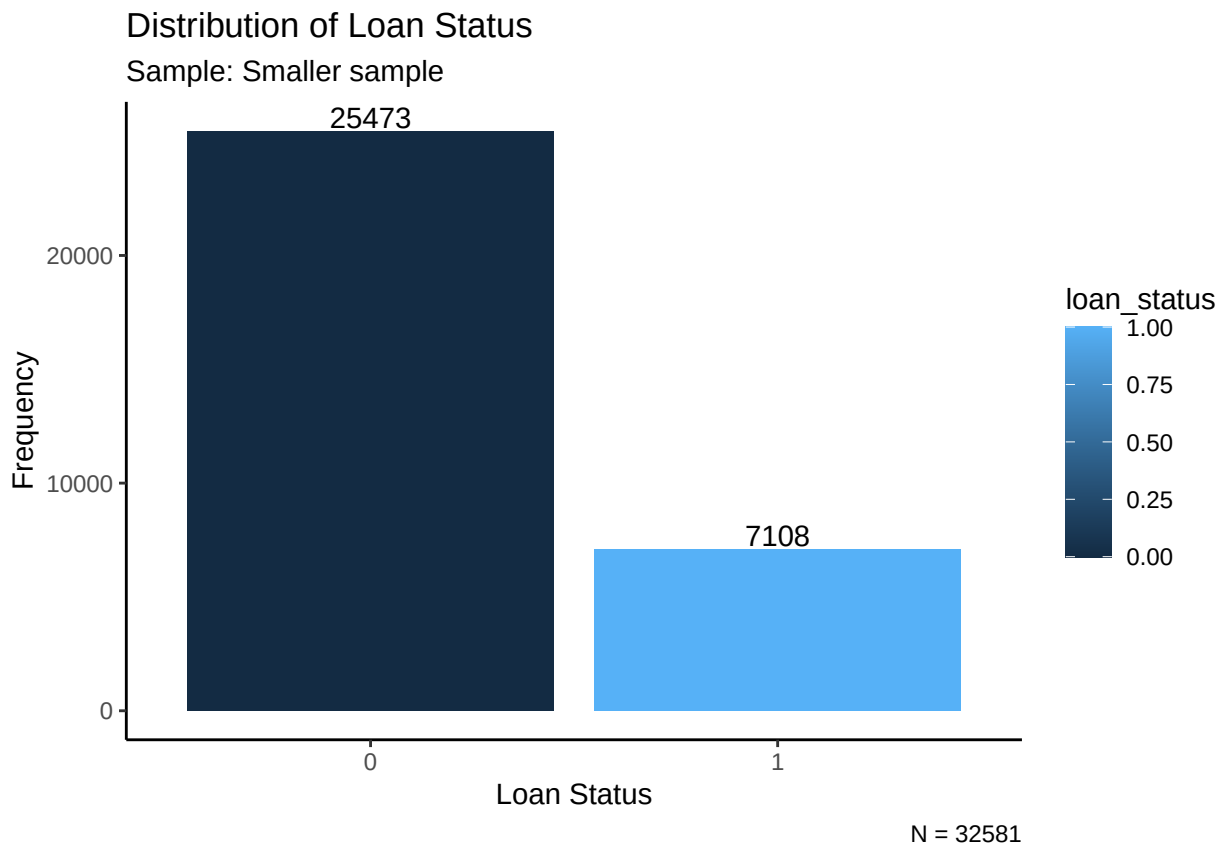
- large dataset *credit_risk*
- variable: *loan_grade*

```
# canvas
canvas <- credit_risk %>%
  ggplot(
    aes(
      x=loan_grade,
      fill=loan_grade
    )
  ) +
  geom_text(stat = "count",
            aes(label = ..count..),
            vjust = -0.15) +
  theme_classic() +
  labs(
    title = "Distribution of Loan Grade",
    subtitle = "Sample: Larger sample",
    caption = paste("N =", nrow(credit_risk)),
    x = "Loan Grade",
    y = "Frequency"
  )

# plot
canvas + geom_bar()
```



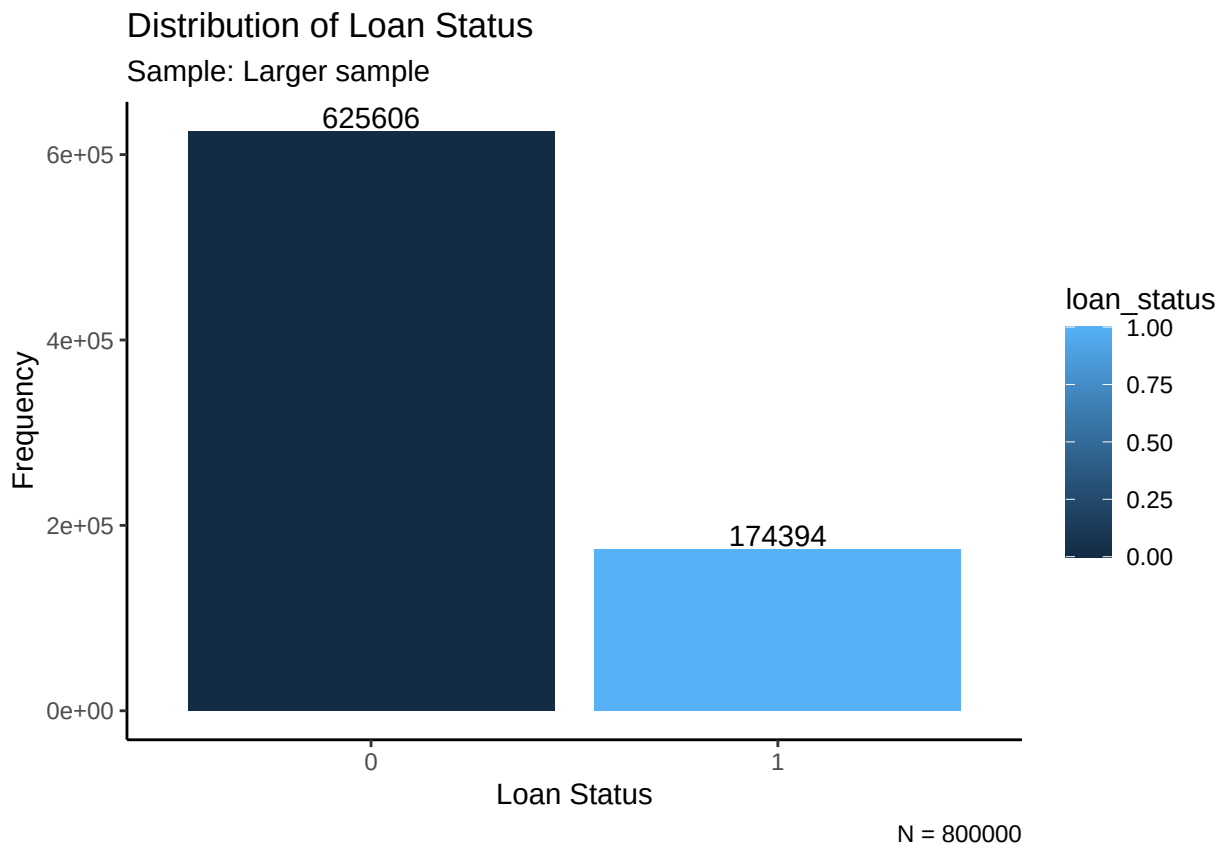
- small dataset *crisk*
- variable: *loan_status*

```
# canvas
canvas <- crisk %>%
  ggplot(
    aes(
      x=as.factor(loan_status),
      fill=loan_status
    )
  ) +
  geom_text(stat = "count",
            aes(label = ..count..),
            vjust = -0.15) +
  theme_classic() +
  labs(
    title = "Distribution of Loan Status",
    subtitle = "Sample: Smaller sample",
    caption = paste("N =", nrow(crisk)),
    x = "Loan Status",
    y = "Frequency"
  )

# plot
canvas + geom_bar()
```



- large dataset *credit_risk*
- variable: *loan_status*

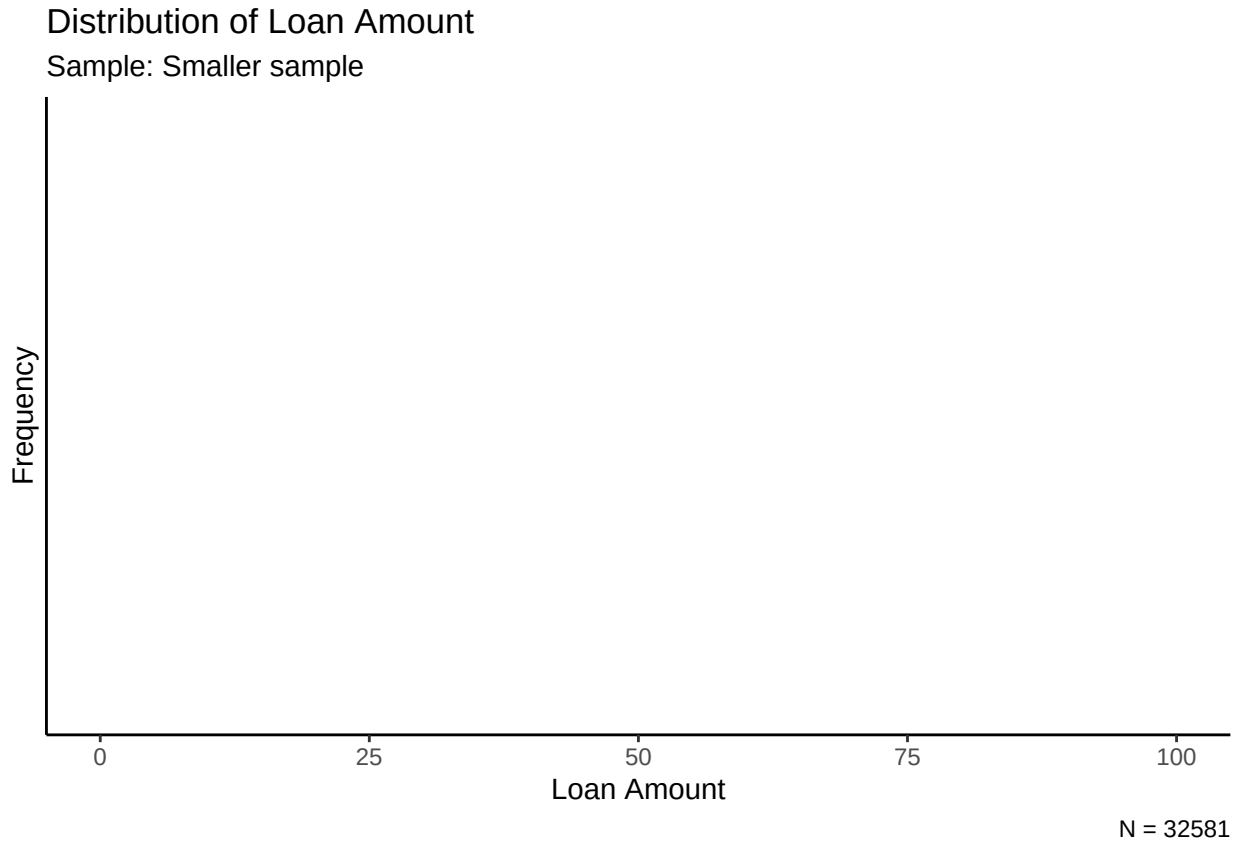
```
# canvas
canvas <- credit_risk %>%
  ggplot(
    aes(
      x=as.factor(loan_status),
      fill=loan_status
    )
  ) +
  geom_text(stat = "count",
            aes(label = ..count..),
            vjust = -0.15) +
  theme_classic() +
  labs(
    title = "Distribution of Loan Status",
    subtitle = "Sample: Larger sample",
    caption = paste("N =", nrow(credit_risk)),
    x = "Loan Status",
    y = "Frequency"
  )

# plot
canvas + geom_bar()
```



- small dataset *crisk*
- variable: *loan_amnt*

```
# canvas
canvas <- crisk %>%
  ggplot() +
  theme_classic() +
  xlim(0,100) +
  labs(
    title = "Distribution of Loan Amount",
    subtitle = "Sample: Smaller sample",
    caption = paste("N =", nrow(crisk)),
    x = "Loan Amount",
    y = "Frequency"
  )

# plot
canvas +
  geom_histogram(
    aes(
      x=loan_amnt
    ),
    bins = 100,
    fill = "red"
  )
```

```
## Warning: Removed 32581 rows containing non-finite values (`stat_bin()`).
```

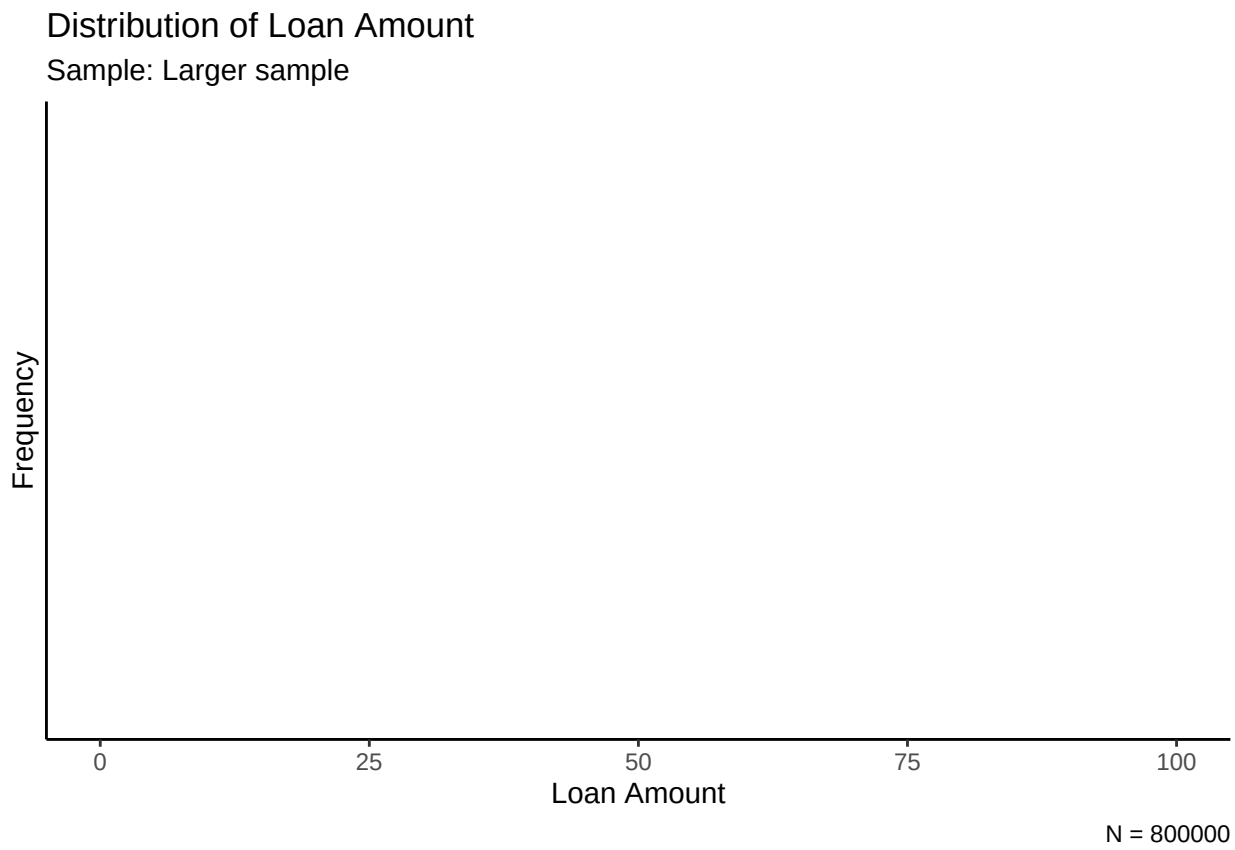


Statistic	Value
Min	500
P1	1000
P5	2000
P10	3000
Q1 = P25	5000
Mean	9589
Median	8000
Q3 = P75	1.22×10^4
P90	1.9×10^4
P95	2.4×10^4
P99	2.98×10^4
Max	3.5×10^4
Std. Dev	6322

- large dataset *credit_risk*
- variable: *loan_amnt*

```r
# canvas
canvas <- credit_risk %>%
  ggplot() +
  theme_classic() +
  xlim(0,100) +
  labs(
    title = "Distribution of Loan Amount",
    subtitle = "Sample: Larger sample",
    caption = paste("N =", nrow(credit_risk)),
    x = "Loan Amount",
    y = "Frequency"
  )

# plot
canvas +
  geom_histogram(
    aes(
      x=loan_amnt
    ),
    bins = 100,
    fill = "red"
  )
```

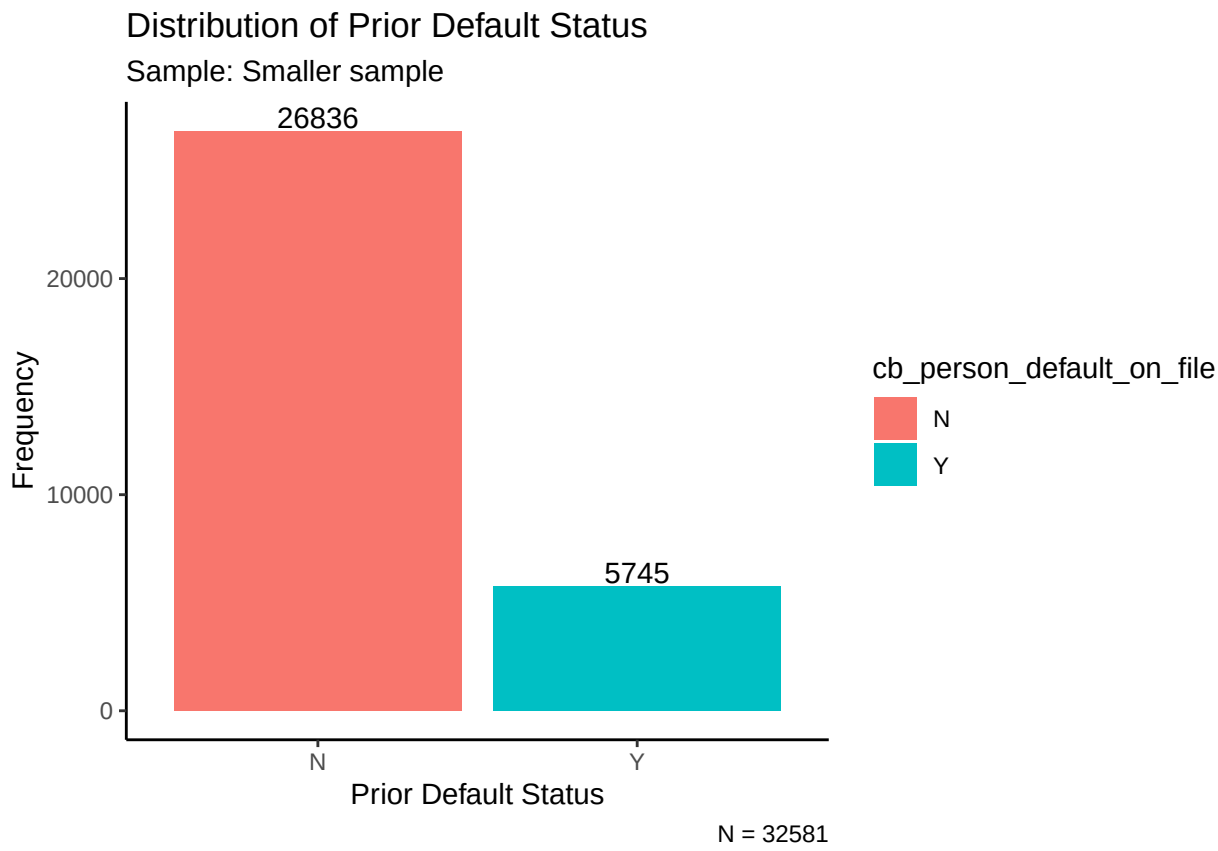
```
## Warning: Removed 800000 rows containing non-finite values (‘stat_bin()‘).
```



Statistic	Value
Min	1000
P1	1000
P5	2000
P10	3000
Q1 = P25	5000
Mean	9501
Median	8000
Q3 = P75	1.2×10^4
P90	1.85×10^4
P95	2.4×10^4
P99	2.8×10^4
Max	3.5×10^4
Std. Dev	6279

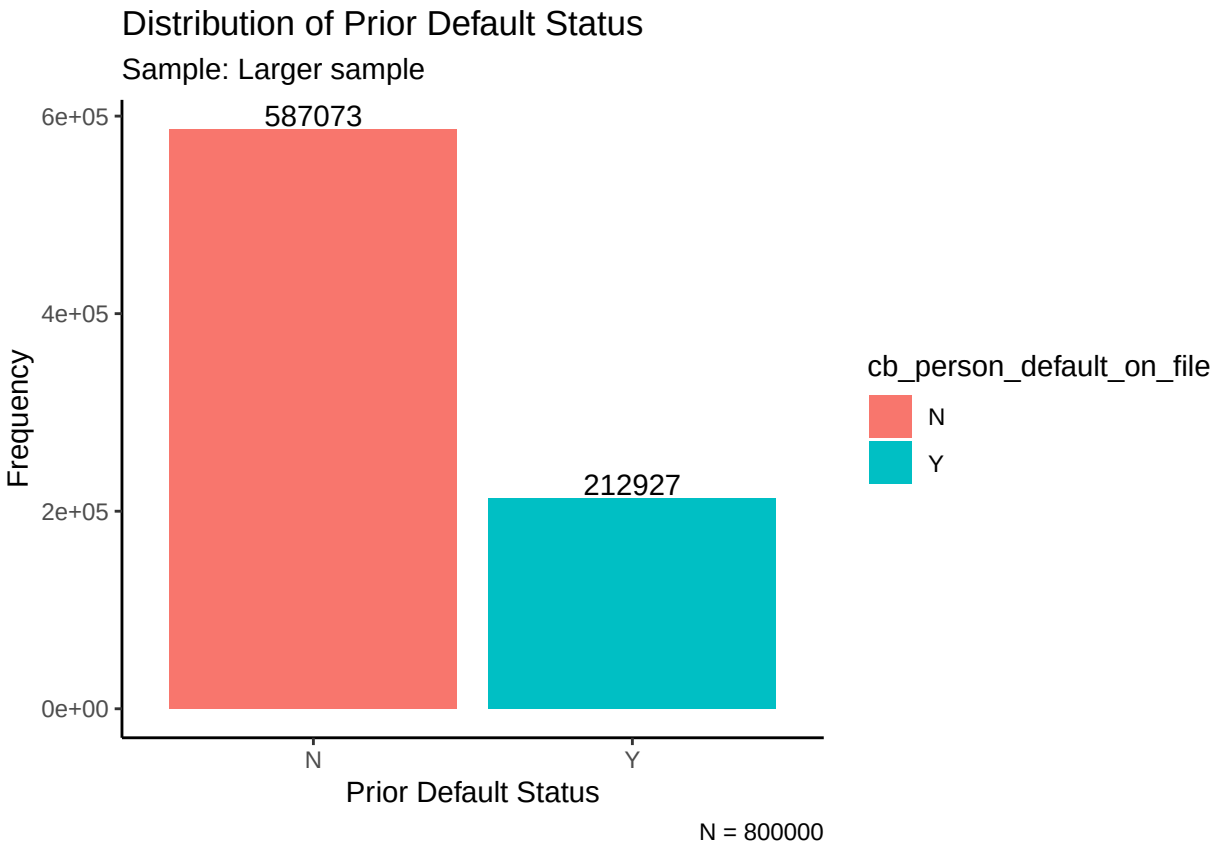
- small dataset *crisk*
- variable: *cb_person_default_on_file*

```
# canvas
canvas <- crisk %>%
  ggplot(
    aes(
      x=cb_person_default_on_file,
      fill=cb_person_default_on_file
    )
  ) +
  geom_text(stat = "count",
            aes(label = ..count..),
            vjust = -0.15) +
  theme_classic() +
  labs(
    title = "Distribution of Prior Default Status",
    subtitle = "Sample: Smaller sample",
    caption = paste("N =", nrow(crisk)),
    x = "Prior Default Status",
    y = "Frequency"
  )

# plot
canvas + geom_bar()
```



- large dataset *credit_risk*
- variable: *cb_person_default_on_file*

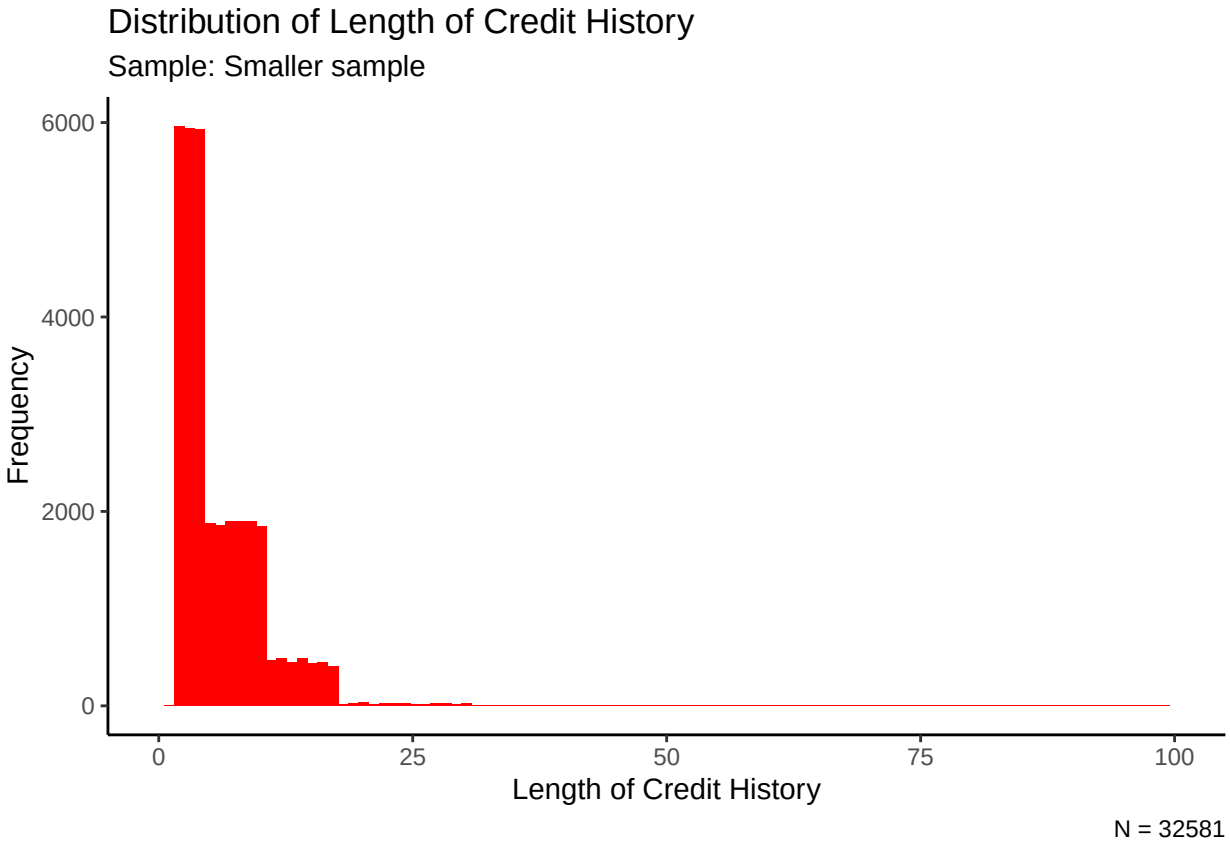
```
# canvas
canvas <- credit_risk %>%
  ggplot(
    aes(
      x=cb_person_default_on_file,
      fill=cb_person_default_on_file
    )
  ) +
  geom_text(stat = "count",
            aes(label = ..count..),
            vjust = -0.15) +
  theme_classic() +
  labs(
    title = "Distribution of Prior Default Status",
    subtitle = "Sample: Larger sample",
    caption = paste("N =", nrow(credit_risk)),
    x = "Prior Default Status",
    y = "Frequency"
  )

# plot
canvas + geom_bar()
```



- small dataset *crisk*
- variable: *cb_person_cred_hist_length*

```
# canvas
canvas <- crisk %>%
  ggplot() +
  theme_classic() +
  xlim(0,100) +
  labs(
    title = "Distribution of Length of Credit History",
    subtitle = "Sample: Smaller sample",
    caption = paste("N =", nrow(crisk)),
    x = "Length of Credit History",
    y = "Frequency"
  )

# plot
canvas +
  geom_histogram(
    aes(
      x=cb_person_cred_hist_length
    ),
    bins = 100,
    fill = "red"
  )
```

```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```

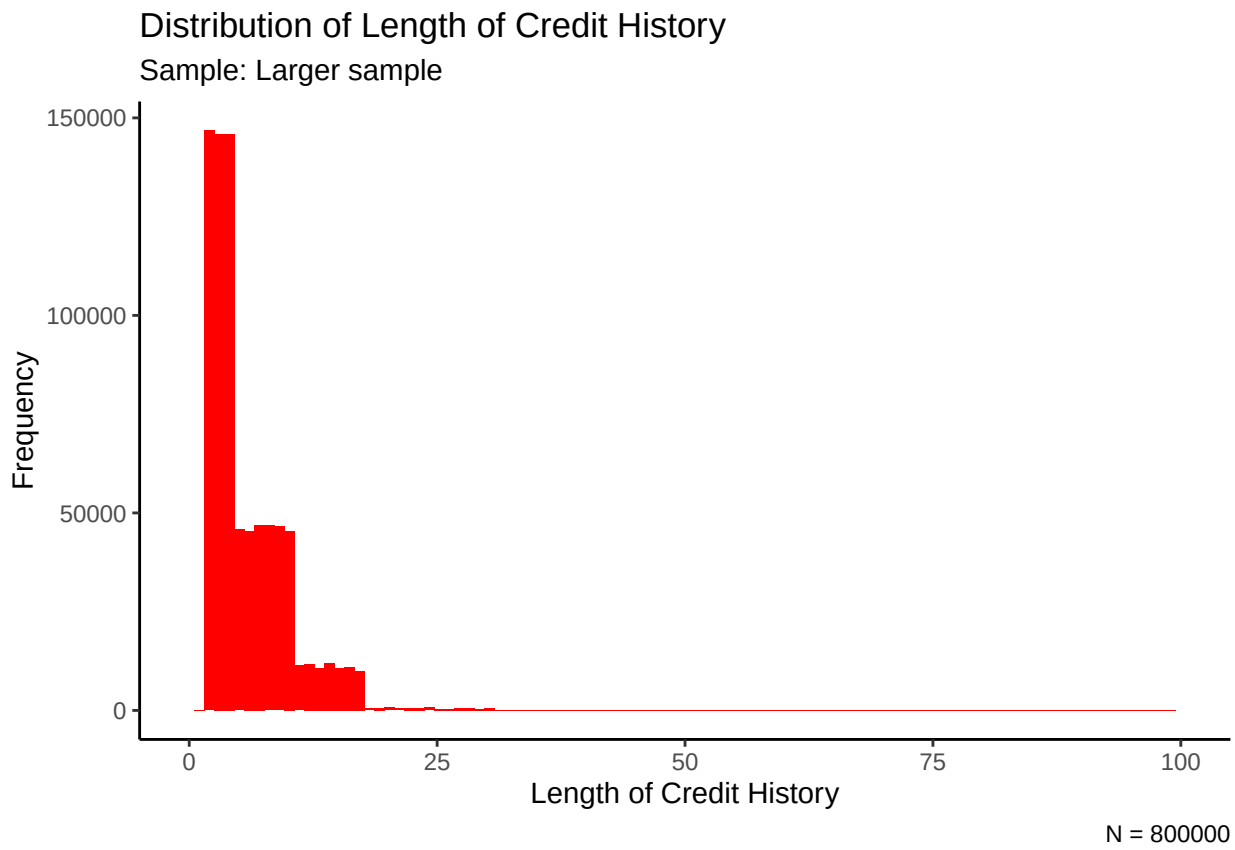


Statistic	Value
Min	2
P1	2
P5	2
P10	2
Q1 = P25	3
Mean	6
Median	4
Q3 = P75	8
P90	11
P95	14
P99	17
Max	30
Std. Dev	4

- large dataset *credit_risk*
- variable: *cb_person_cred_hist_length*

```r
# canvas
canvas <- credit_risk %>%
  ggplot() +
  theme_classic() +
  xlim(0,100) +
  labs(
    title = "Distribution of Length of Credit History",
    subtitle = "Sample: Larger sample",
    caption = paste("N =", nrow(credit_risk)),
    x = "Length of Credit History",
    y = "Frequency"
  )

# plot
canvas +
  geom_histogram(
    aes(
      x=cb_person_cred_hist_length
    ),
    bins = 100,
    fill = "red"
  )
```

```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```

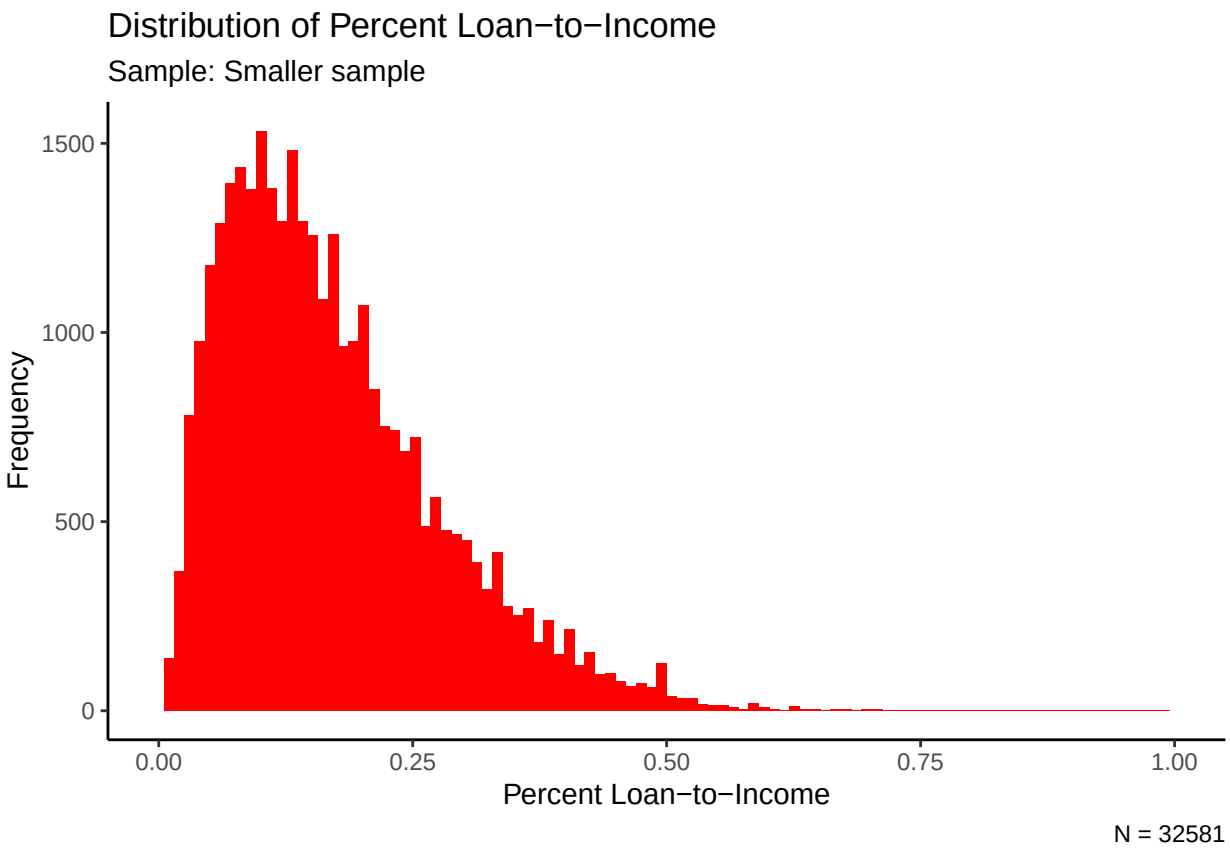


Statistic	Value
Min	2
P1	2
P5	2
P10	2
Q1 = P25	3
Mean	6
Median	4
Q3 = P75	8
P90	11
P95	14
P99	17
Max	30
Std. Dev	4

- small dataset *crisk*
- variable: *loan_percent_income*

```
# canvas
canvas <- crisk %>%
  ggplot() +
  theme_classic() +
  xlim(0,1) +
  labs(
    title = "Distribution of Percent Loan-to-Income",
    subtitle = "Sample: Smaller sample",
    caption = paste("N =", nrow(crisk)),
    x = "Percent Loan-to-Income",
    y = "Frequency"
  )

# plot
canvas +
  geom_histogram(
    aes(
      x=loan_percent_income
    ),
    bins = 100,
    fill = "red"
  )
```

```
## Warning: Removed 2 rows containing missing values (‘geom_bar()‘).
```



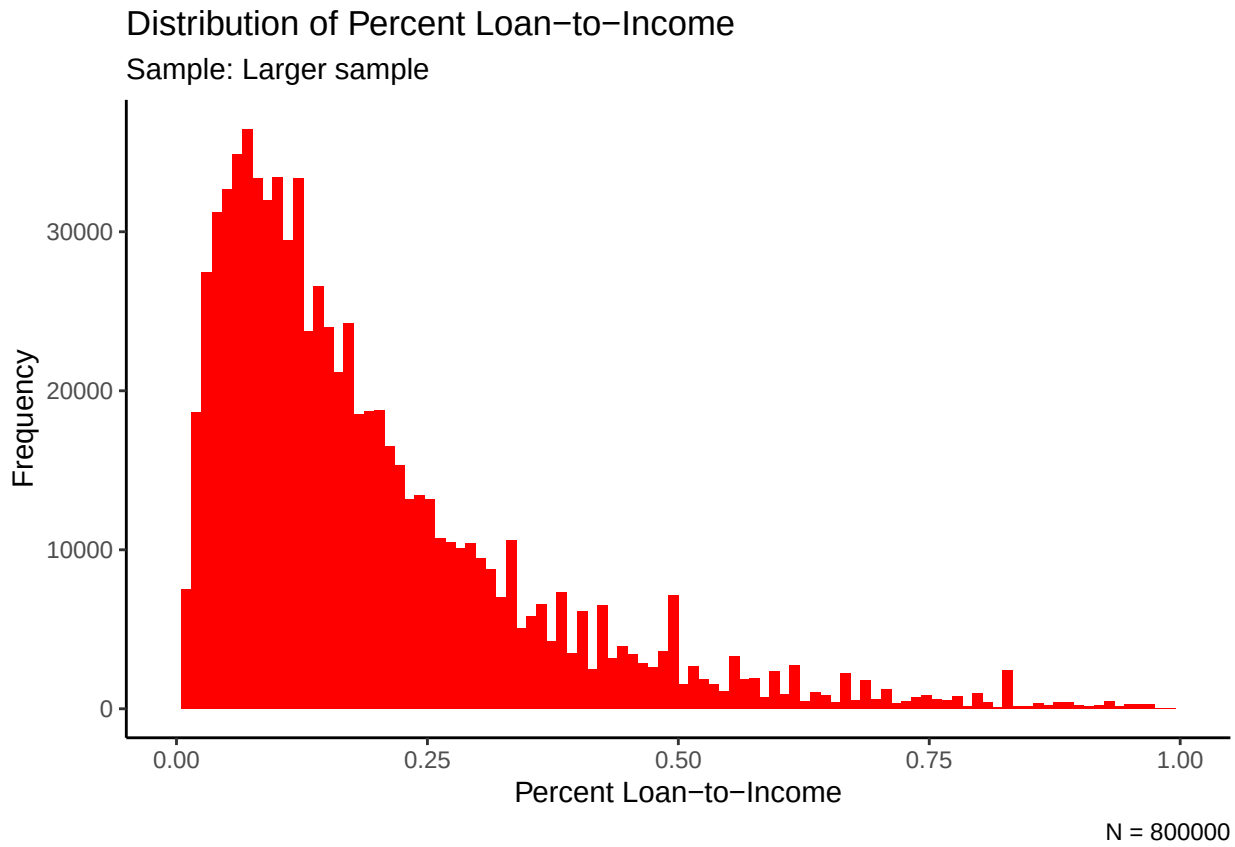
Statistic	Value
Min	0
P1	2
P5	4
P10	5
Q1 = P25	9
Mean	17.02
Median	15
Q3 = P75	23
P90	32
P95	38
P99	50
Max	83
Std. Dev	10.678

- large dataset *credit_risk*
- variable: *loan_percent_income*

```r
# canvas
canvas <- credit_risk %>%
  ggplot() +
  theme_classic() +
  xlim(0,1) +
  labs(
    title = "Distribution of Percent Loan-to-Income",
    subtitle = "Sample: Larger sample",
    caption = paste("N =", nrow(credit_risk)),
    x = "Percent Loan-to-Income",
    y = "Frequency"
  )

# plot
canvas +
  geom_histogram(
    aes(
      x=loan_percent_income
    ),
    bins = 100,
    fill = "red"
  )
```

```
## Warning: Removed 5530 rows containing non-finite values (‘stat_bin()‘).
```

```
## Warning: Removed 2 rows containing missing values (‘geom_bar()‘).
```



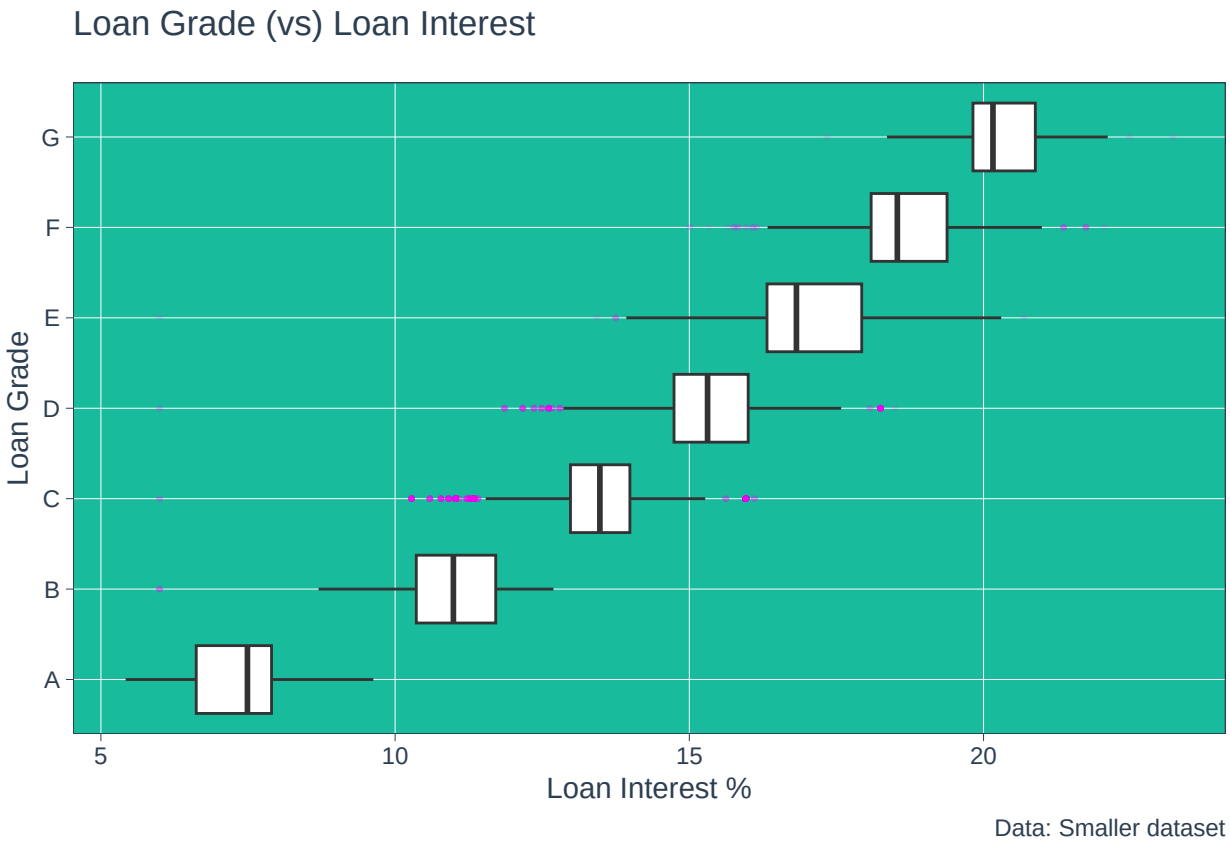
Statistic	Value
Min	0
P1	2
P5	3
P10	4
Q1 = P25	8
Mean	20.028
Median	14
Q3 = P75	26
P90	42
P95	56
P99	93
Max	292
Std. Dev	18.88

2.2 Empirical data analysis

2.2.1 Relationship between Loan Grade and Loan Interest

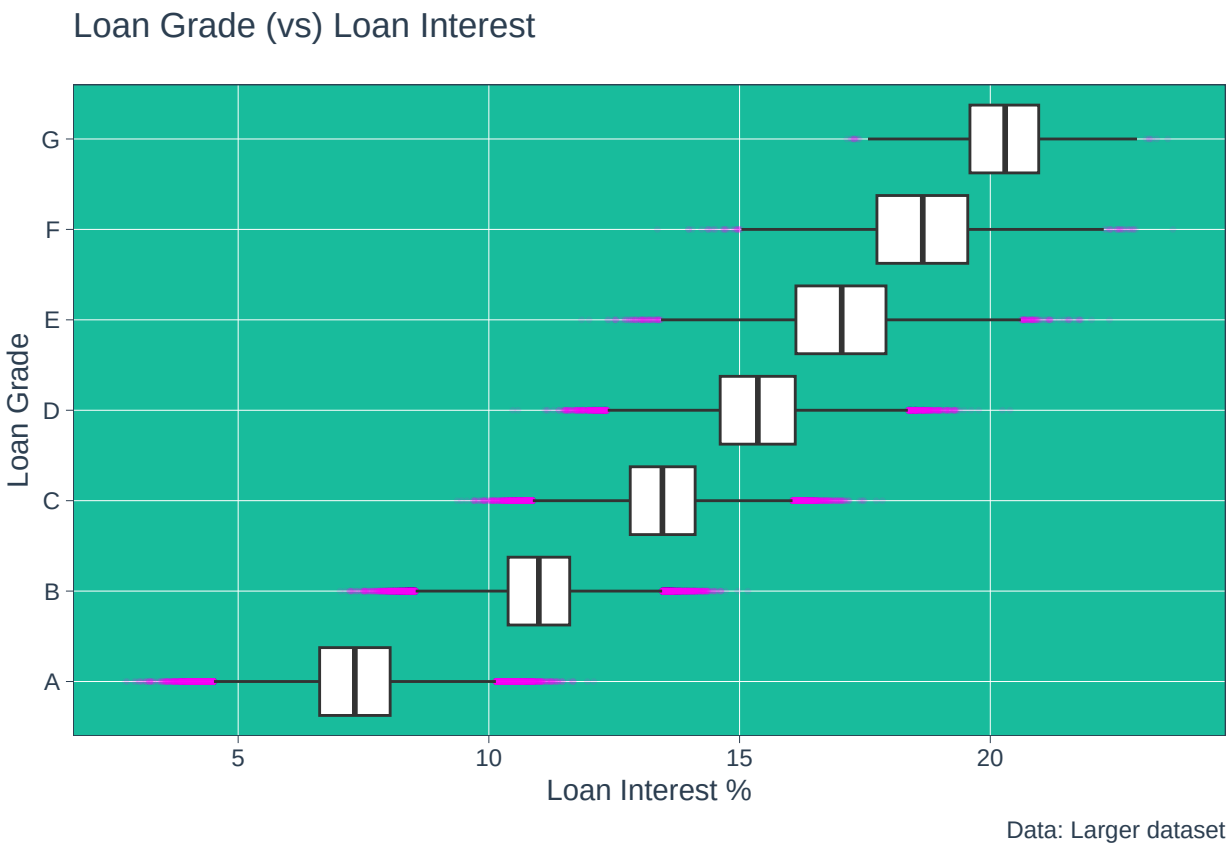
Conventional wisdom would dictate that better interest rate would be reserved for better grade loans. The following plot gives the relationship between the loan grade and its corresponding loan interest rate offered.

```
# create canvas
canvas <- crisk %>%
  ggplot(
    mapping = aes(
      x = loan_grade,
      y = loan_int_rate
    )
  ) +
  theme_tq_green() +
  labs(
    title = "Loan Grade (vs) Loan Interest",
    subtitle = "",
    caption = "Data: Smaller dataset",
    x = "Loan Grade",
    y = "Loan Interest %"
  )

# box plot with options ...
canvas + geom_boxplot(na.rm = TRUE, outlier.color = "magenta",
                      outlier.size = 0.5,
                      outlier.alpha = 0.1) -> boxplot1

boxplot1 + coord_flip() # flip x and y axis
```



From the plot, we observe that ...

```
# create canvas
canvasa <- credit_risk %>%
  ggplot(
    mapping = aes(
      x = loan_grade,
      y = loan_int_rate
    )
  ) +
  theme_tq_green() +
  labs(
    title = "Loan Grade (vs) Loan Interest",
    subtitle = "",
    caption = "Data: Larger dataset",
    x = "Loan Grade",
    y = "Loan Interest %"
  )

# box plot with options ...
canvasa + geom_boxplot(na.rm = TRUE, outlier.color = "magenta",
                       outlier.size = 0.5,
                       outlier.alpha = 0.1) -> boxplot1a

boxplot1a + coord_flip() # flip x and y axis
```



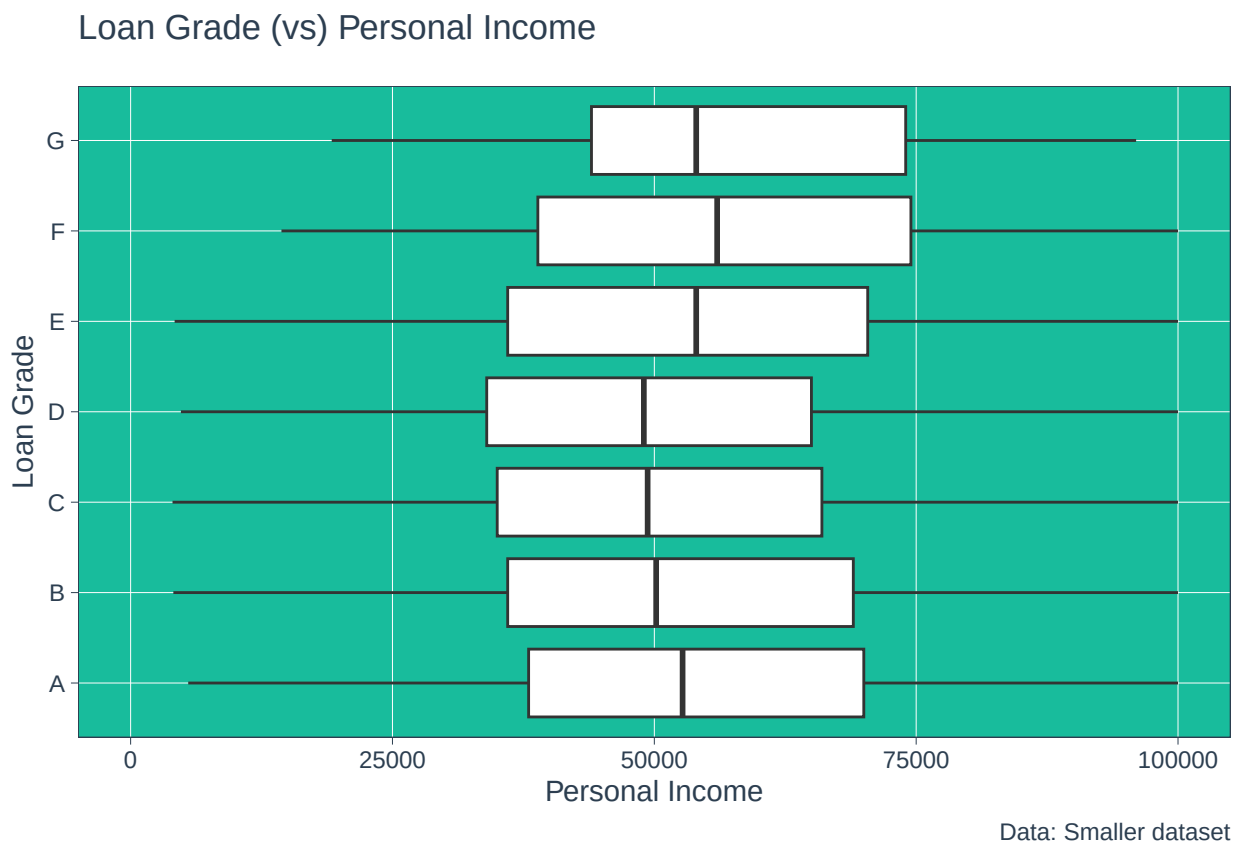
From the plot, we observe that ...

2.2.2 Relationship between Loan Grade and Personal Income

It also plausible to assume that better grade of loans are reserved for high-earning individuals. Therefore, in the next plot, we investigate if personal income and loan grade are related.

```
# create canvas
canvas1 <- crisk %>%
  ggplot(
    mapping = aes(
      x = loan_grade,
      y = person_income
    )
  ) +
  theme_tq_green() +
  scale_y_continuous(limits = c(0, 100000)) +
  labs(
    title = "Loan Grade (vs) Personal Income",
    subtitle = "",
    caption = "Data: Smaller dataset",
    x = "Loan Grade",
    y = "Personal Income"
  )

canvas1 +
```

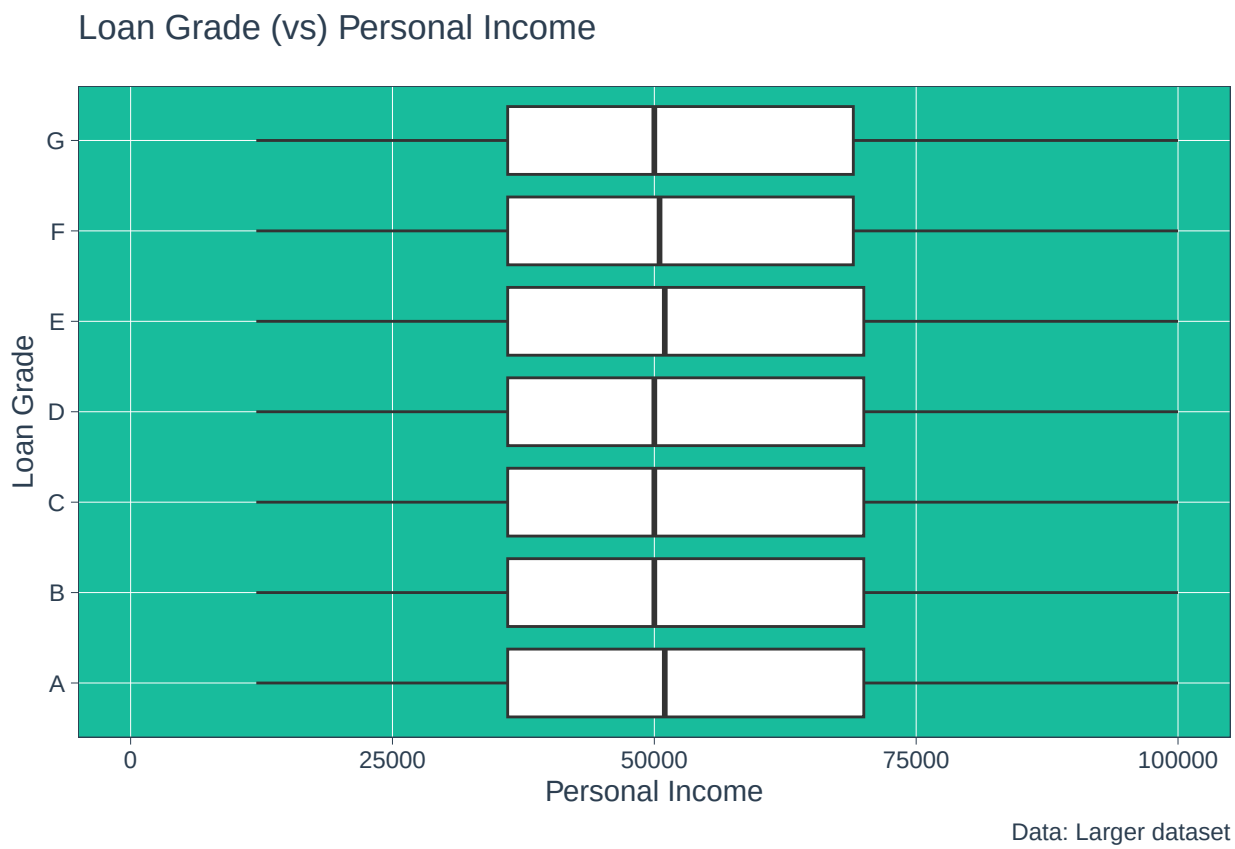
```
  geom_boxplot(na.rm = TRUE,
               outlier.color = "magenta",
               outlier.size = 0.5,
               outlier.alpha = 0.1) -> boxplot2

boxplot2 + coord_flip()
```



From the plot, we observe that ...

```
# create canvas
canvas1a <- credit_risk %>%
  ggplot(
    mapping = aes(
      x = loan_grade,
      y = person_income
    )
  ) +
  theme_tq_green() +
  scale_y_continuous(limits = c(0, 100000)) +
  labs(
    title = "Loan Grade (vs) Personal Income",
    subtitle = "",
    caption = "Data: Larger dataset",
    x = "Loan Grade",
    y = "Personal Income"
  )
```

```
canvas1a +
  geom_boxplot(na.rm = TRUE,
               outlier.color = "magenta",
               outlier.size = 0.5,
               outlier.alpha = 0.1) -> boxplot2a

boxplot2a + coord_flip()
```



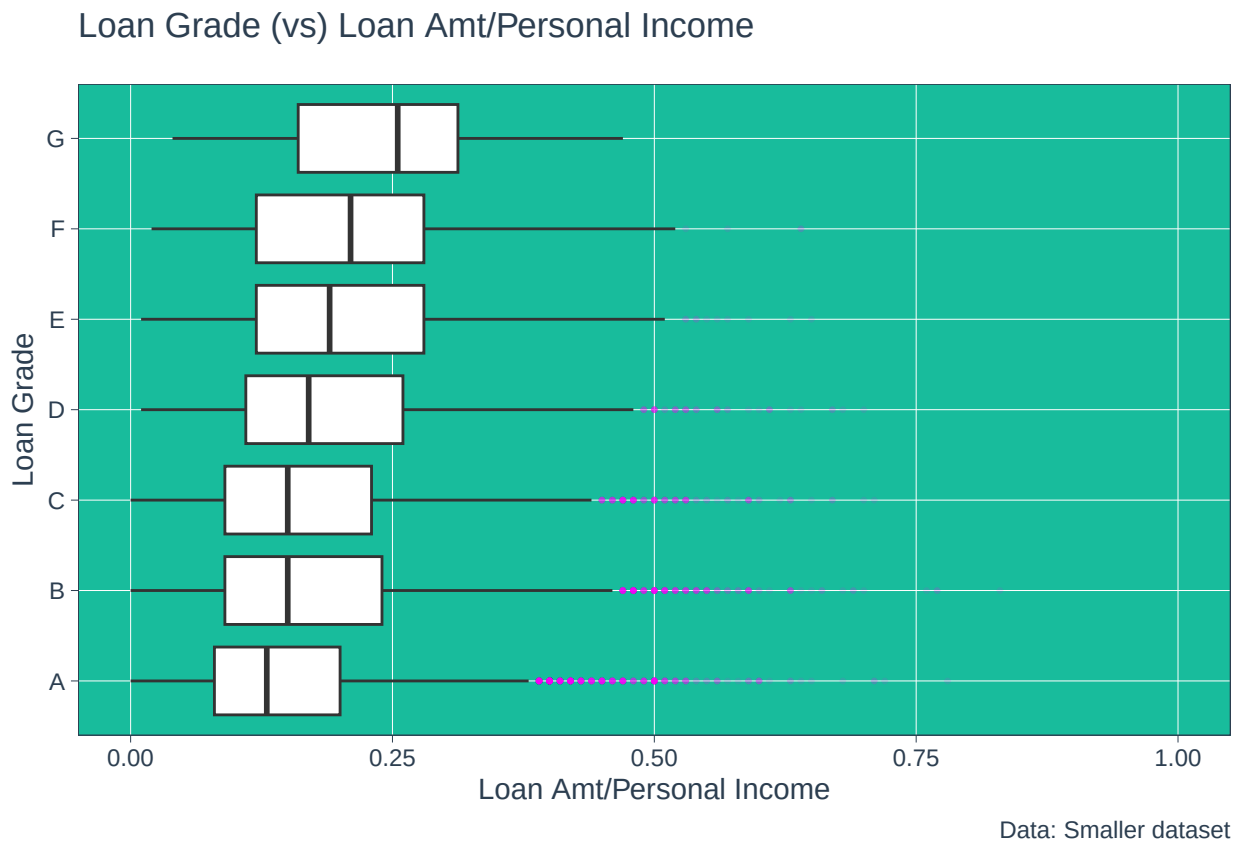
From the plot, we observe that ...

2.2.3 Relationship between Loan Grade and Loan Amt/Personal Income

Next, we investigate whether loan grade is related to the share of the loan amount to personal income.

```
# create canvas
canvas2 <- crisk %>%
  ggplot(
    mapping = aes(
      x = loan_grade,
      y = loan_percent_income
    )
  ) +
  theme_tq_green() +
  scale_y_continuous(limits = c(0, 1)) +
```

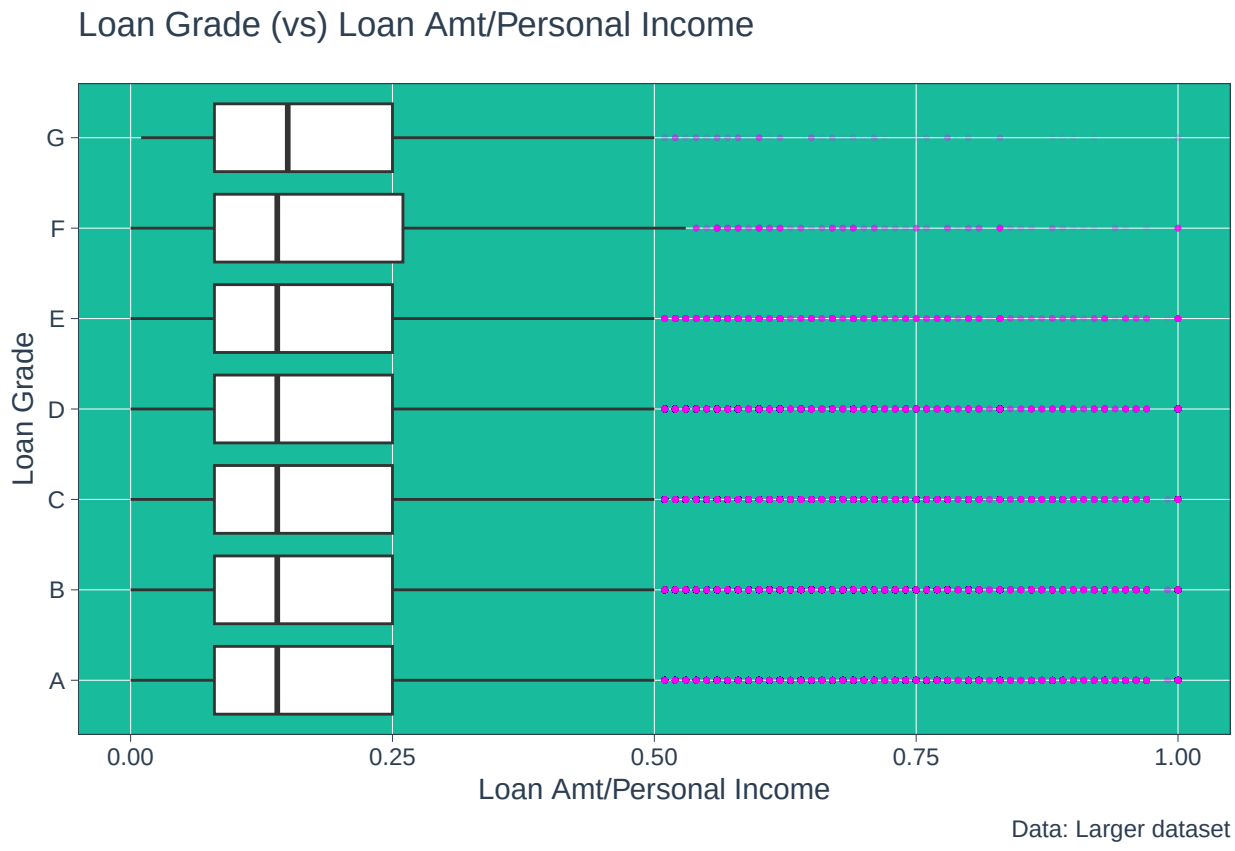
```
  labs(
    title = "Loan Grade (vs) Loan Amt/Personal Income",
    subtitle = "",
    caption = "Data: Smaller dataset",
    x = "Loan Grade",
    y = "Loan Amt/Personal Income"
  )

canvas2 +
  geom_boxplot(na.rm = TRUE,
               outlier.color = "magenta",
               outlier.size = 0.5,
               outlier.alpha = 0.1) -> boxplot3

boxplot3 + coord_flip()
```



From the plot, we observe that ...

```
# create canvas
canvas2a <- credit_risk %>%
  ggplot(
    mapping = aes(
      x = loan_grade,
      y = loan_percent_income
    )
  ) +
```

```
  theme_tq_green() +
  scale_y_continuous(limits = c(0, 1)) +
  labs(
    title = "Loan Grade (vs) Loan Amt/Personal Income",
    subtitle = "",
    caption = "Data: Larger dataset",
    x = "Loan Grade",
    y = "Loan Amt/Personal Income"
  )


canvas2a +
  geom_boxplot(na.rm = TRUE,
               outlier.color = "magenta",
               outlier.size = 0.5,
               outlier.alpha = 0.1) -> boxplot3a

boxplot3a + coord_flip()
```



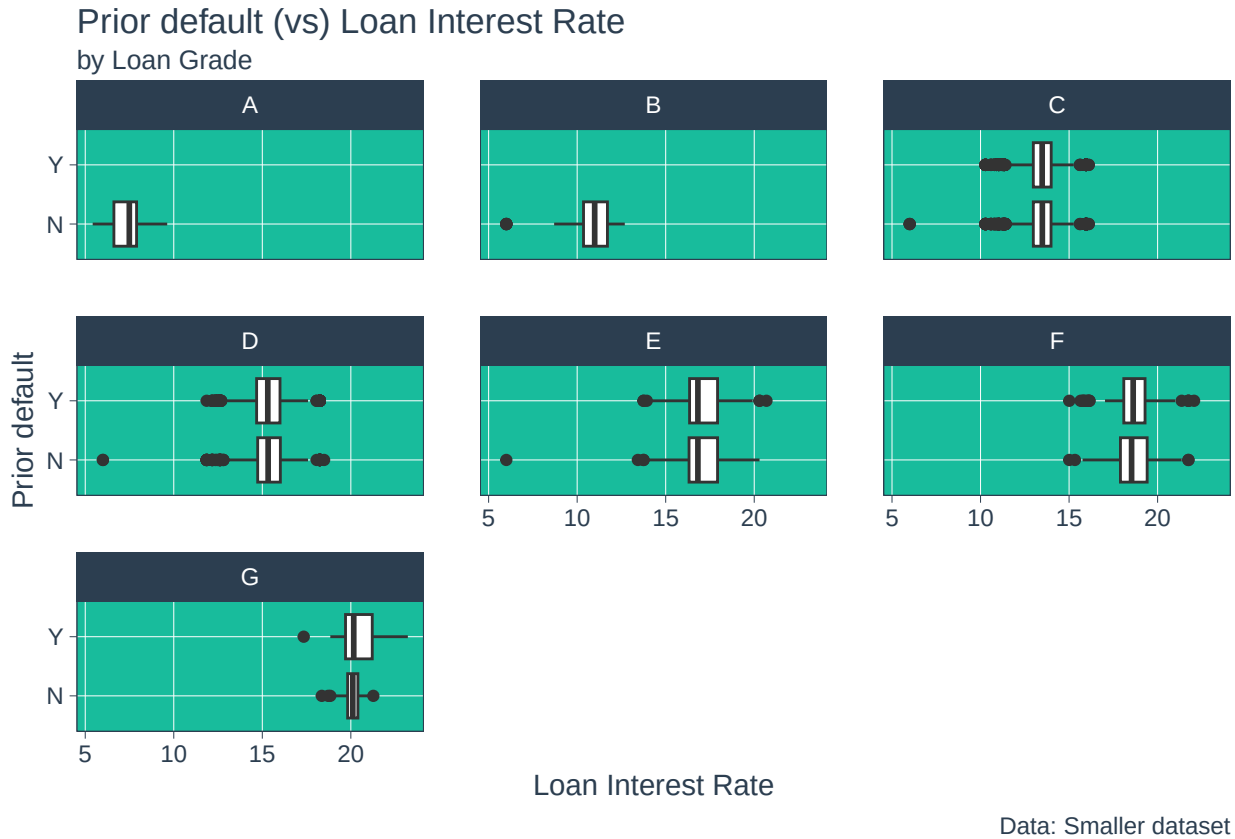
From the plot, we observe that ...

2.2.4 Relationship between Loan Grade, Prior default Status, and Loan Interest Rate

Here, we analyze whether the relationship between Prior default Status and Loan Interest Rate vary by the quality of the loan offered (Loan Grade).

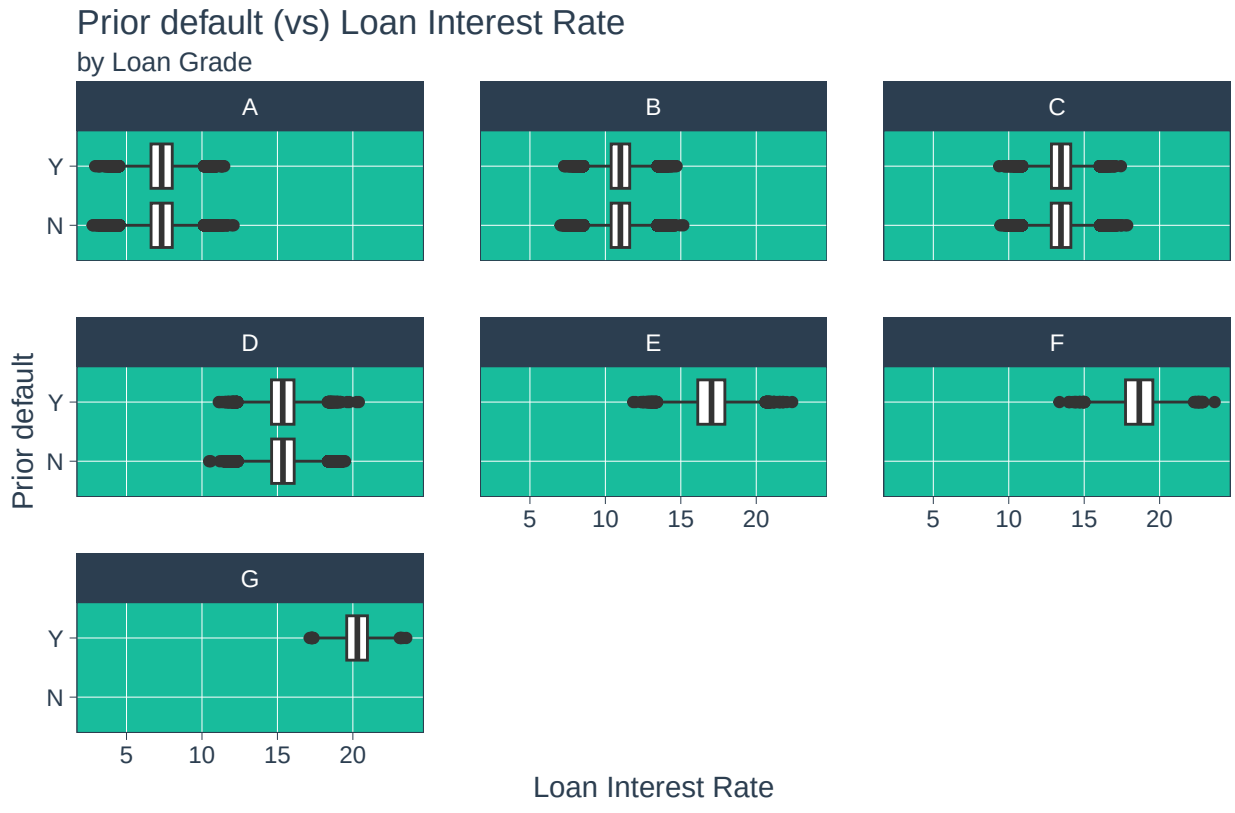
```
# create canvas
canvas4 <- crisk %>% ggplot(
  mapping = aes(
    x = cb_person_default_on_file,
    y = loan_int_rate
  )) +
  theme_tq_green() +
  labs(
    title = "Prior default (vs) Loan Interest Rate",
    subtitle = "by Loan Grade",
    caption = "Data: Smaller dataset",
    x = "Prior default", y = "Loan Interest Rate"
  )

#-------------------------------------------------------------------------------
# Prior Default Status on Loan Interest Rate by Loan Grade
#-------------------------------------------------------------------------------

canvas4 + geom_boxplot() + coord_flip() +
  facet_wrap(~crisk$loan_grade) # sep plots
```

```
## Warning: Removed 3116 rows containing non-finite values (`stat_boxplot()`).
```



From the plot, we observe that ...

```
# create canvas
canvas4a <- credit_risk %>% ggplot(
  mapping = aes(
    x = cb_person_default_on_file,
    y = loan_int_rate
    )
  ) +
  theme_tq_green() +
  labs(
    title = "Prior default (vs) Loan Interest Rate",
    subtitle = "by Loan Grade",
    caption = "Data: Larger dataset",
    x = "Prior default", y = "Loan Interest Rate"
  )

#-------------------------------------------------------------------------------
# Prior Default Status on Loan Interest Rate by Loan Grade
#-------------------------------------------------------------------------------

canvas4a + geom_boxplot() + coord_flip() +
  facet_wrap(~credit_risk$loan_grade) # sep plots
```

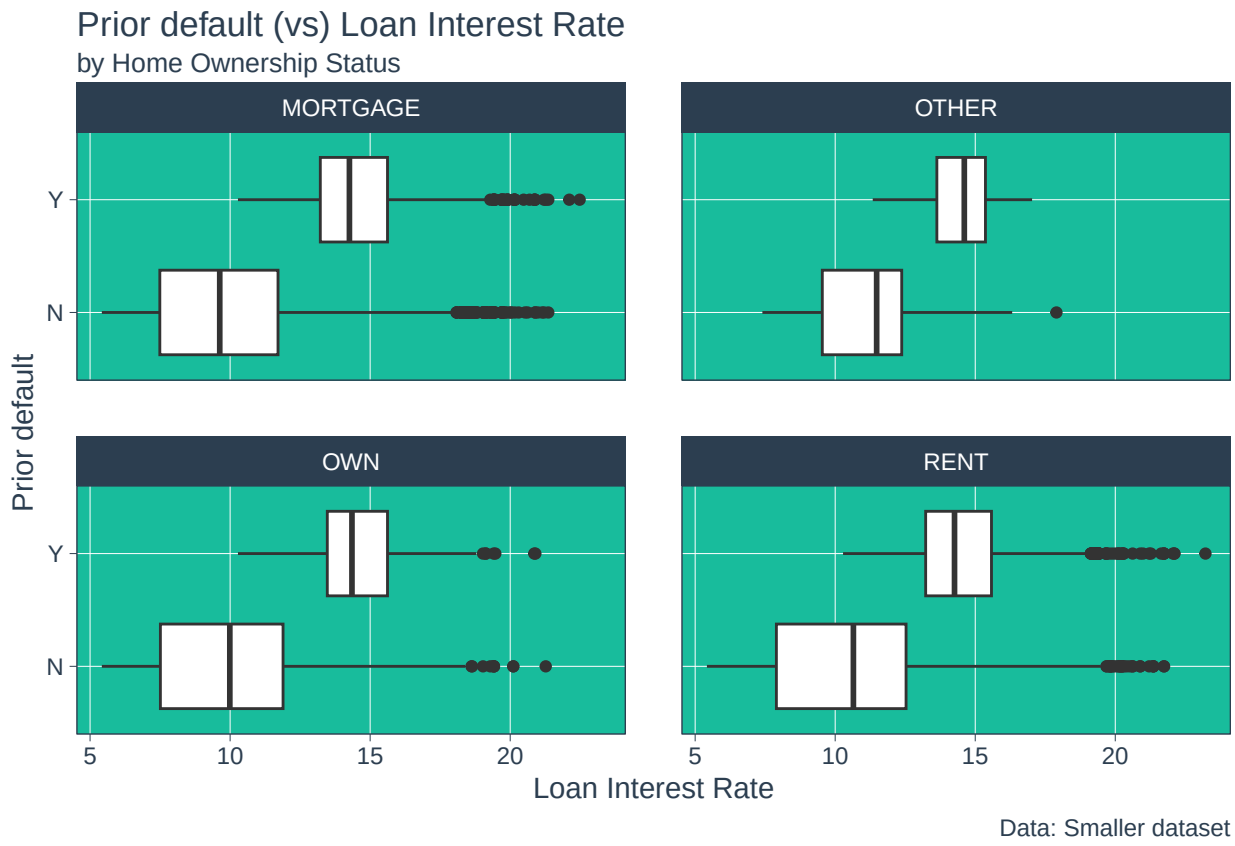


From the plot, we observe that ...

2.2.5 Relationship between Home Ownership, Prior default Status, and Loan Interest Rate

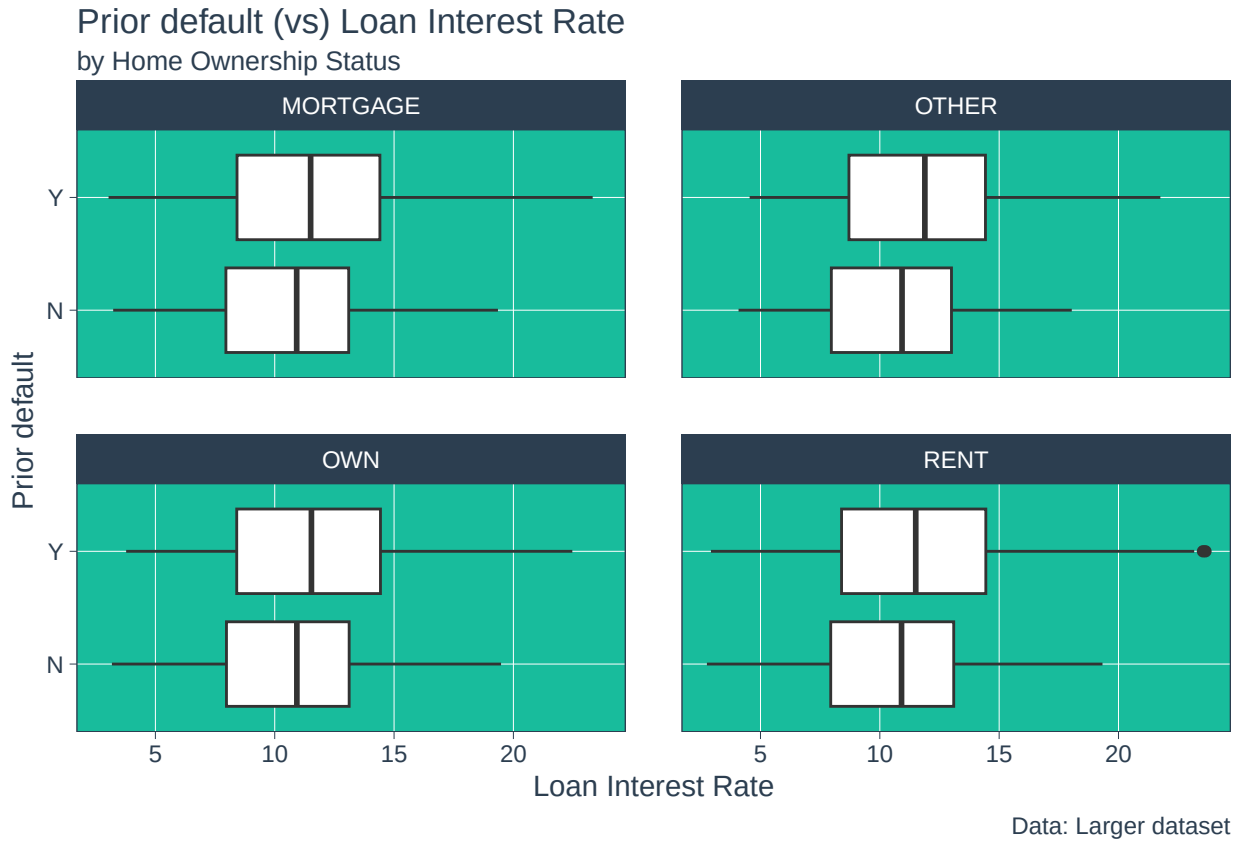
```
#-------------------------------------------------------------------------------
# Prior Default Status on Loan Interest Rate by Home Ownership
#-------------------------------------------------------------------------------
# create canvas
canvas5 <- crisk %>% ggplot(
  mapping = aes(
    x = cb_person_default_on_file,
    y = loan_int_rate
  )) +
  theme_tq_green() +
  labs(
    title = "Prior default (vs) Loan Interest Rate",
    subtitle = "by Home Ownership Status",
    caption = "Data: Smaller dataset",
    x = "Prior default", y = "Loan Interest Rate"
  )

canvas5 + geom_boxplot() + coord_flip() +
  facet_wrap(~crisk$person_home_ownership) # sep plots
```

```
## Warning: Removed 3116 rows containing non-finite values (`stat_boxplot()`).
```



From the plot, we observe that ...

```
# create canvas
canvas5 <- credit_risk %>% ggplot(
  mapping = aes(
    x = cb_person_default_on_file,
    y = loan_int_rate
  )) +
  theme_tq_green() +
  labs(
    title = "Prior default (vs) Loan Interest Rate",
    subtitle = "by Home Ownership Status",
    caption = "Data: Larger dataset",
    x = "Prior default", y = "Loan Interest Rate"
  )

canvas5 + geom_boxplot() + coord_flip() +
  facet_wrap(~credit_risk$person_home_ownership) # sep plots
```



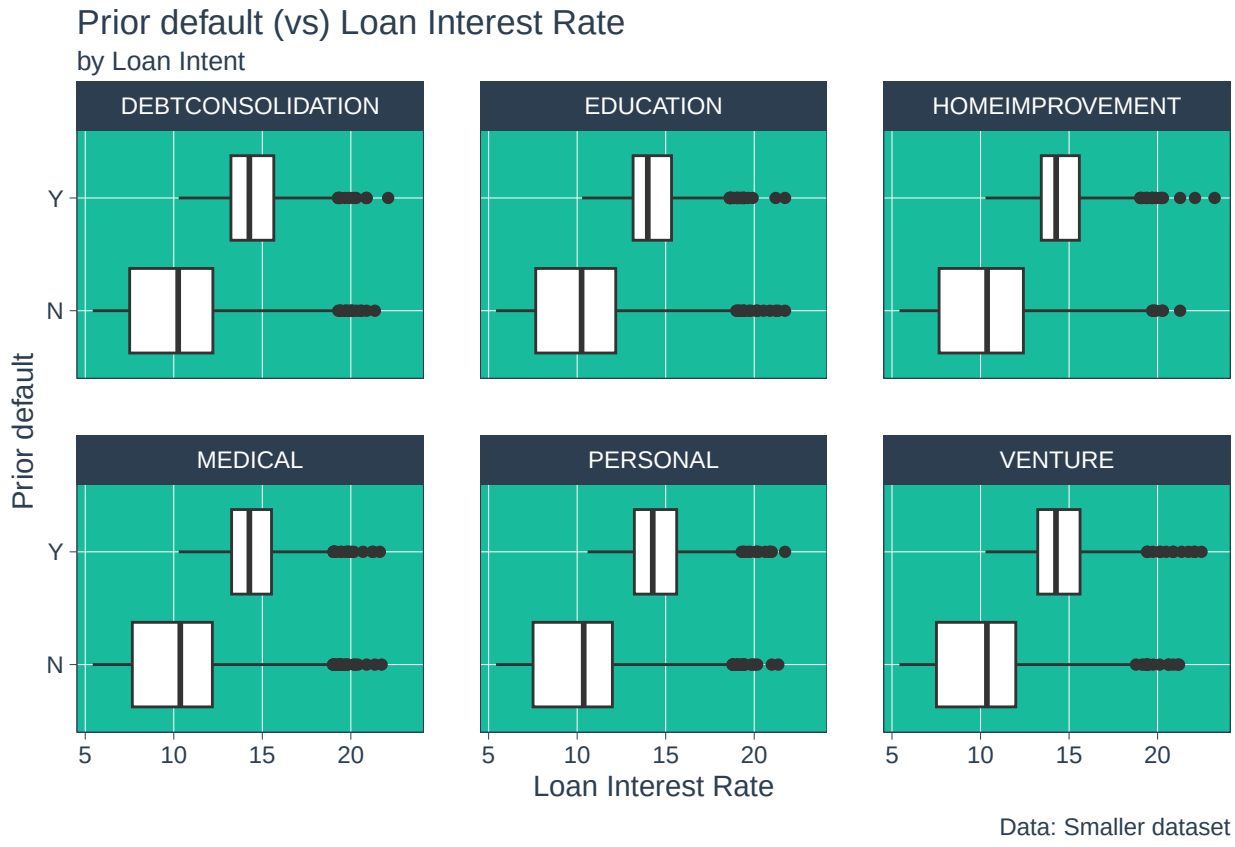
From the plot, we observe that ...

2.2.6 Relationship between Loan Intent, Prior default Status, and Loan Interest Rate

```
#-------------------------------------------------------------------------
# Prior Default Status on Loan Interest Rate by Home Ownership
#-------------------------------------------------------------------------
```

```
# create canvas
canvas6 <- crisk %>% ggplot(
  mapping = aes(
    x = cb_person_default_on_file,
    y = loan_int_rate
  )) +
  theme_tq_green() +
  labs(
    title = "Prior default (vs) Loan Interest Rate",
    subtitle = "by Loan Intent",
    caption = "Data: Smaller dataset",
    x = "Prior default", y = "Loan Interest Rate"
  )

canvas6 + geom_boxplot() + coord_flip() +
  facet_wrap(~crisk$loan_intent) # sep plots
```

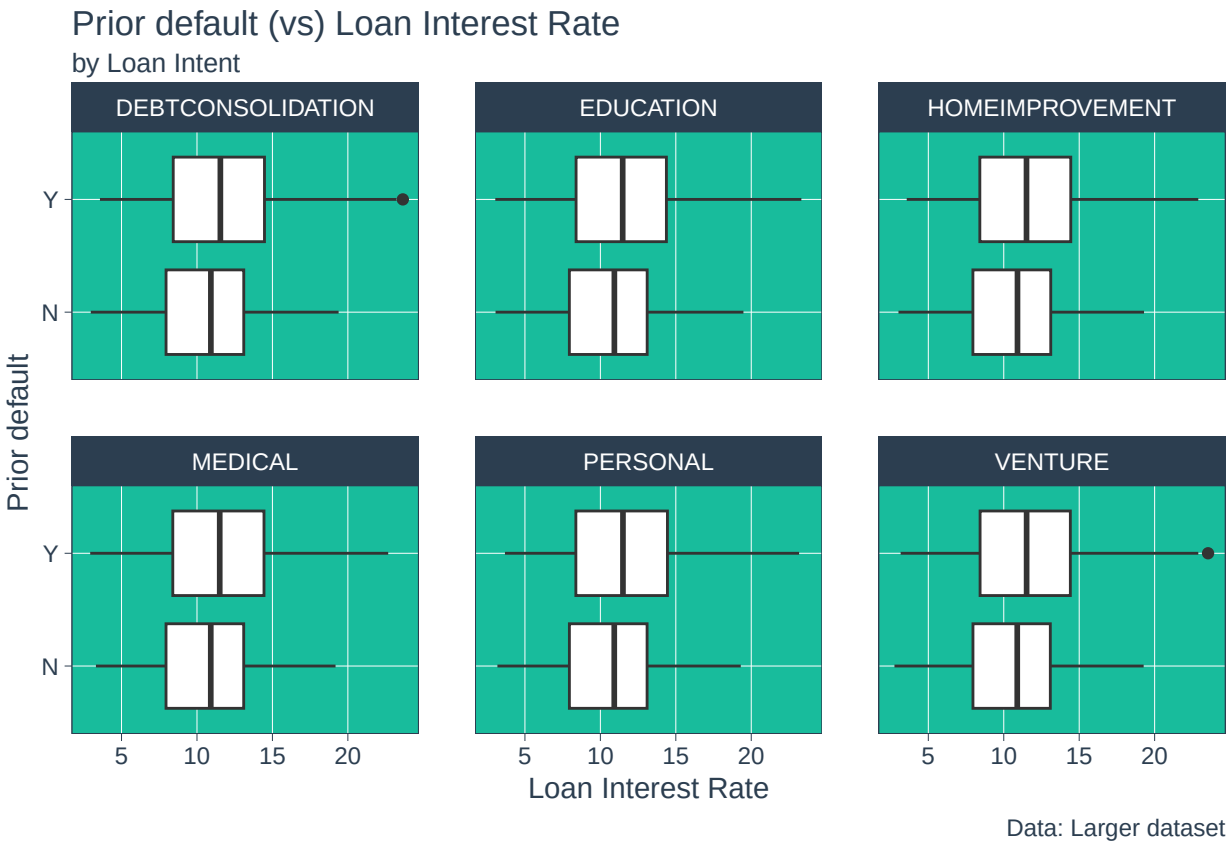
```
## Warning: Removed 3116 rows containing non-finite values (‘stat_boxplot()‘).
```



From the plot, we observe that ...

```
# create canvas
canvas6 <- credit_risk %>% ggplot(
  mapping = aes(
    x = cb_person_default_on_file,
```

```
    y = loan_int_rate
  )) +
  theme_tq_green() +
  labs(
    title = "Prior default (vs) Loan Interest Rate",
    subtitle = "by Loan Intent",
    caption = "Data: Larger dataset",
    x = "Prior default", y = "Loan Interest Rate"
  )

canvas6 + geom_boxplot() + coord_flip() +
  facet_wrap(~credit_risk$loan_intent) # sep plots
```



From the plot, we observe that ...