R_5_visuals_ggplot

Mike Montes

2024-01-19

```
knitr::opts_chunk$set(echo = TRUE)
```

1 Read the dataset in and call it 'loan.'

```
library(readr)
loan <- read_csv("C:/Users/mmsax/Downloads/loan.csv")
```

```
## Rows: 10000 Columns: 11
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): term, grade, emp_length, home_ownership, verification_status, loan_...
## dbl (5): id, loan_amnt, int_rate, installment, annual_inc
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(loan, 5)
```

```
## # A tibble: 5 x 11
##        id loan_amnt term    int_rate installment grade emp_length home_ownership
##     <dbl>     <dbl> <chr>      <dbl>       <dbl> <chr> <chr>      <chr>
## 1 1077501      5000 36 mon~     10.6        163.  B     10+ years  RENT
## 2 1077430      2500 60 mon~     15.3         59.8 C     < 1 year   RENT
## 3 1077175      2400 36 mon~     16.0         84.3 C     10+ years  RENT
## 4 1076863     10000 36 mon~     13.5        339.  C     10+ years  RENT
## 5 1075358      3000 60 mon~     12.7         67.8 B     1 year     RENT
## # i 3 more variables: annual_inc <dbl>, verification_status <chr>,
## #   loan_status <chr>
```

```
colnames(loan)
```

```
##  [1] "id"                 "loan_amnt"          "term"
##  [4] "int_rate"           "installment"        "grade"
##  [7] "emp_length"         "home_ownership"     "annual_inc"
## [10] "verification_status" "loan_status"
```

```
str(loan)
```

```
## spc_tbl_ [10,000 x 11] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ id                 : num [1:10000] 1077501 1077430 1077175 1076863 1075358 ...
##  $ loan_amnt          : num [1:10000] 5000 2500 2400 10000 3000 ...
##  $ term               : chr [1:10000] "36 months" "60 months" "36 months" "36 months" ...
##  $ int_rate           : num [1:10000] 10.6 15.3 16 13.5 12.7 ...
##  $ installment        : num [1:10000] 162.9 59.8 84.3 339.3 67.8 ...
##  $ grade              : chr [1:10000] "B" "C" "C" "C" ...
##  $ emp_length         : chr [1:10000] "10+ years" "< 1 year" "10+ years" "10+ years" ...
##  $ home_ownership     : chr [1:10000] "RENT" "RENT" "RENT" "RENT" ...
##  $ annual_inc         : num [1:10000] 24000 30000 12252 49200 80000 ...
##  $ verification_status: chr [1:10000] "Verified" "Source Verified" "Not Verified" "Source Verified"
##  $ loan_status        : chr [1:10000] "Fully Paid" "Charged Off" "Fully Paid" "Fully Paid" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   id = col_double(),
##   ..   loan_amnt = col_double(),
##   ..   term = col_character(),
##   ..   int_rate = col_double(),
##   ..   installment = col_double(),
##   ..   grade = col_character(),
##   ..   emp_length = col_character(),
##   ..   home_ownership = col_character(),
##   ..   annual_inc = col_double(),
##   ..   verification_status = col_character(),
##   ..   loan_status = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

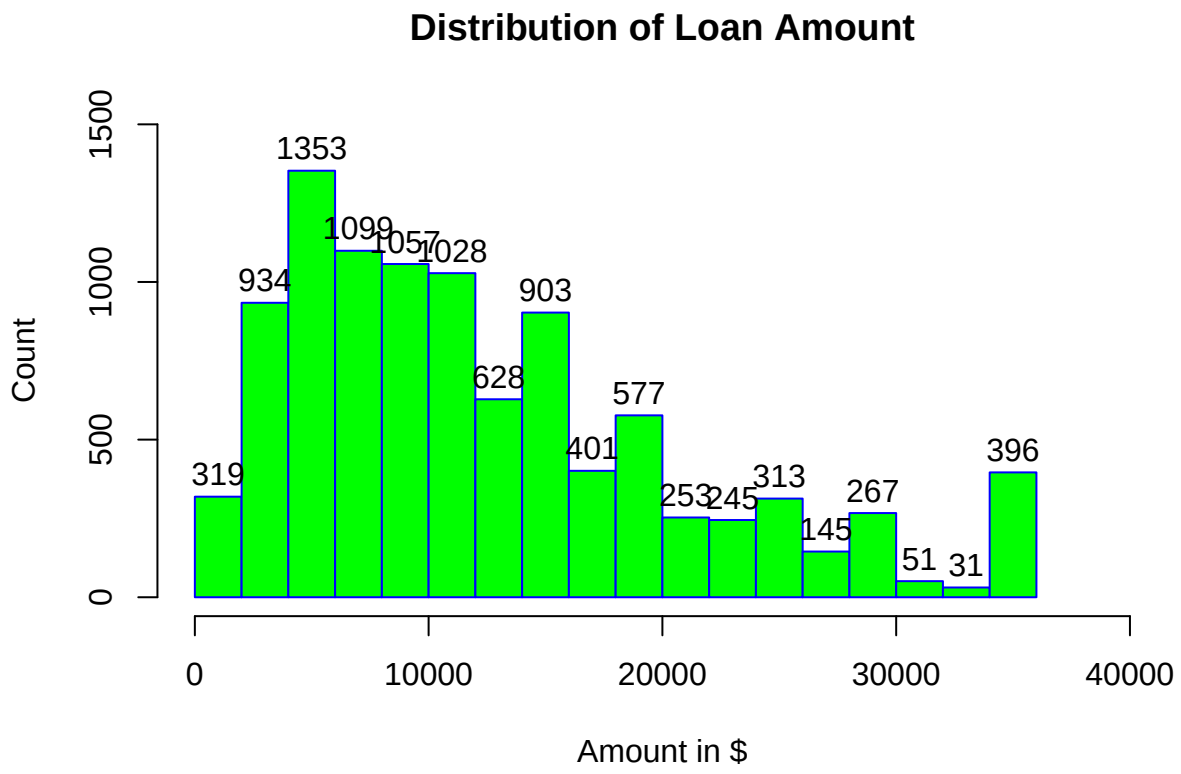
2 Plot histogram and density of loan_amnt using basic.

Using extra commands to learn them

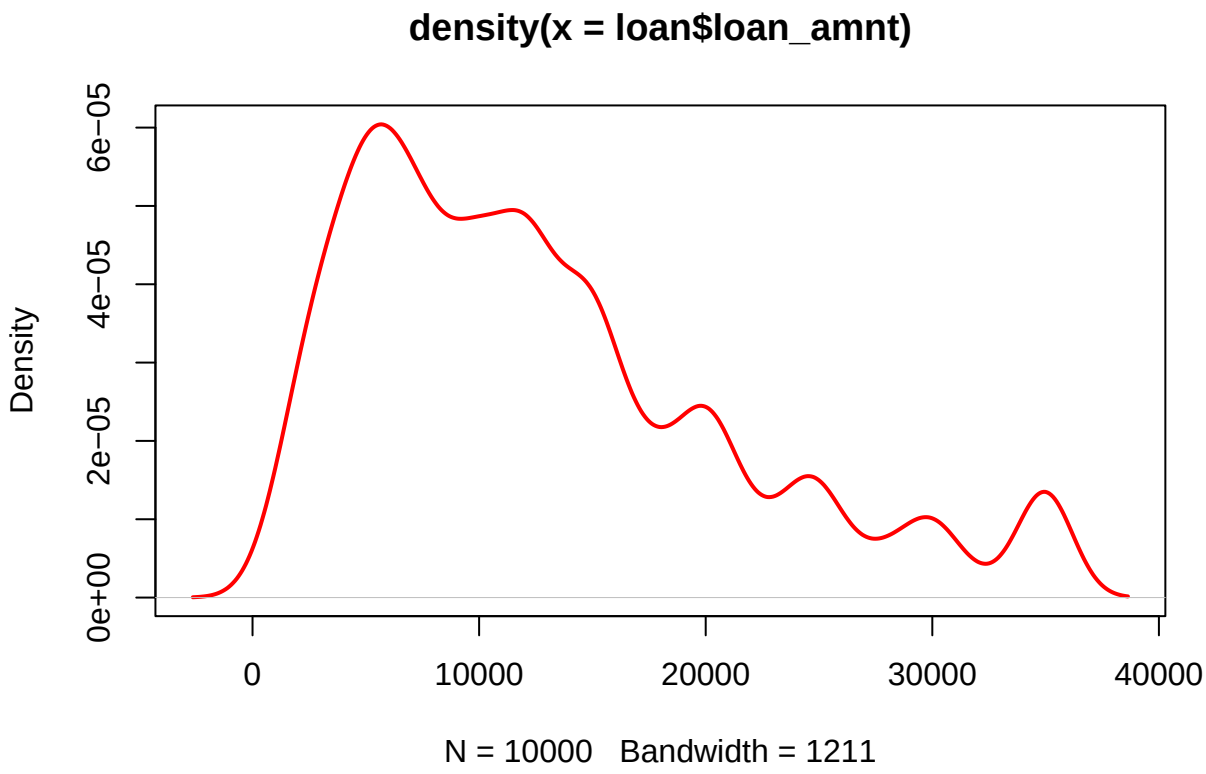
```
?plot
```

```
## starting httpd help server ... done
```

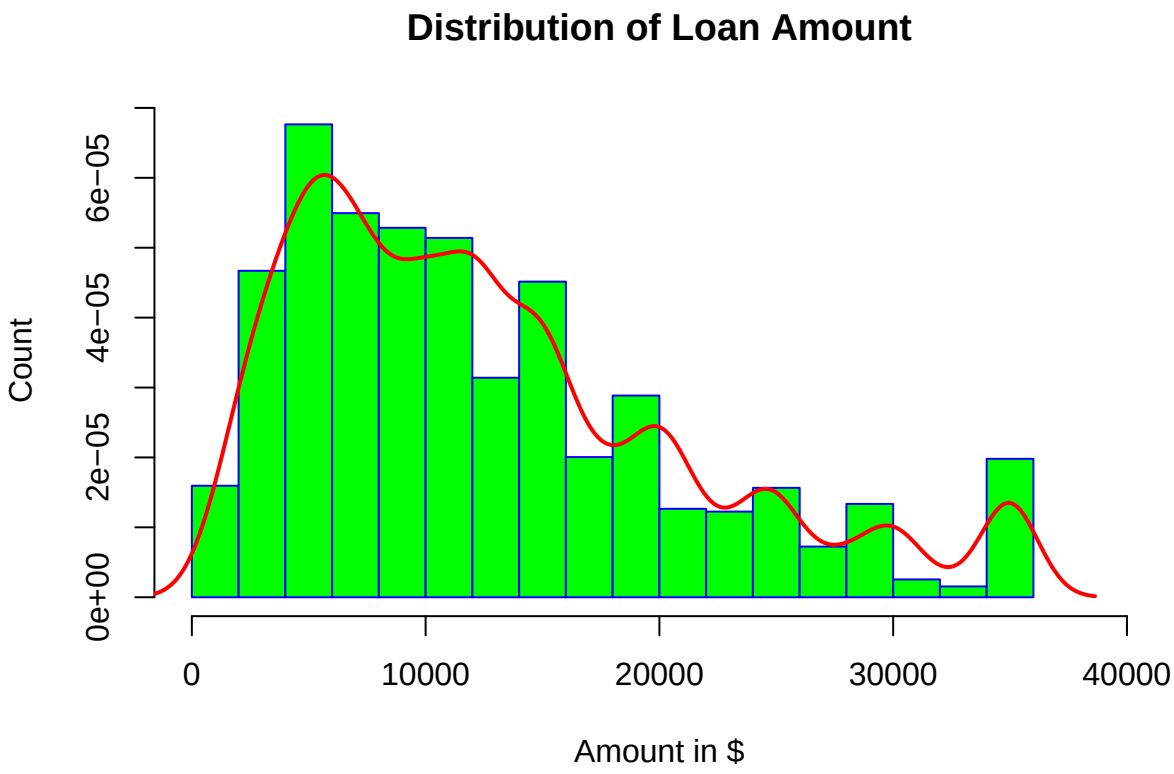
```
?hist

### Histogram
hist(loan$loan_amnt,
     breaks = 'Sturges',
     main = 'Distribution of Loan Amount',
     xlab = 'Amount in $',
     ylab = 'Count',
     xlim = c(0, 40000),
     ylim = c(0, 1500),
     col = 'green',
     border = 'blue',
     labels = TRUE)
```




```
### Density
d <- density(loan$loan_amnt)
plot(d,
     col = 'red',
     lwd = 2)
```




```
### Histogram with density overlaid
hist(loan$loan_amnt,
     breaks = 'Sturges',
     main = 'Distribution of Loan Amount',
     prob = TRUE,
     xlab = 'Amount in $',
     ylab = 'Count',
     xlim = c(0, 40000),
     col = 'green',
     border = 'blue')
# not sure why I can't have y-limit here
lines(density(loan$loan_amnt),
      lwd = 2,
      col = 'red')
```



3 Histogram and density of loan_amnt with vertical line for mean using ggplot2.

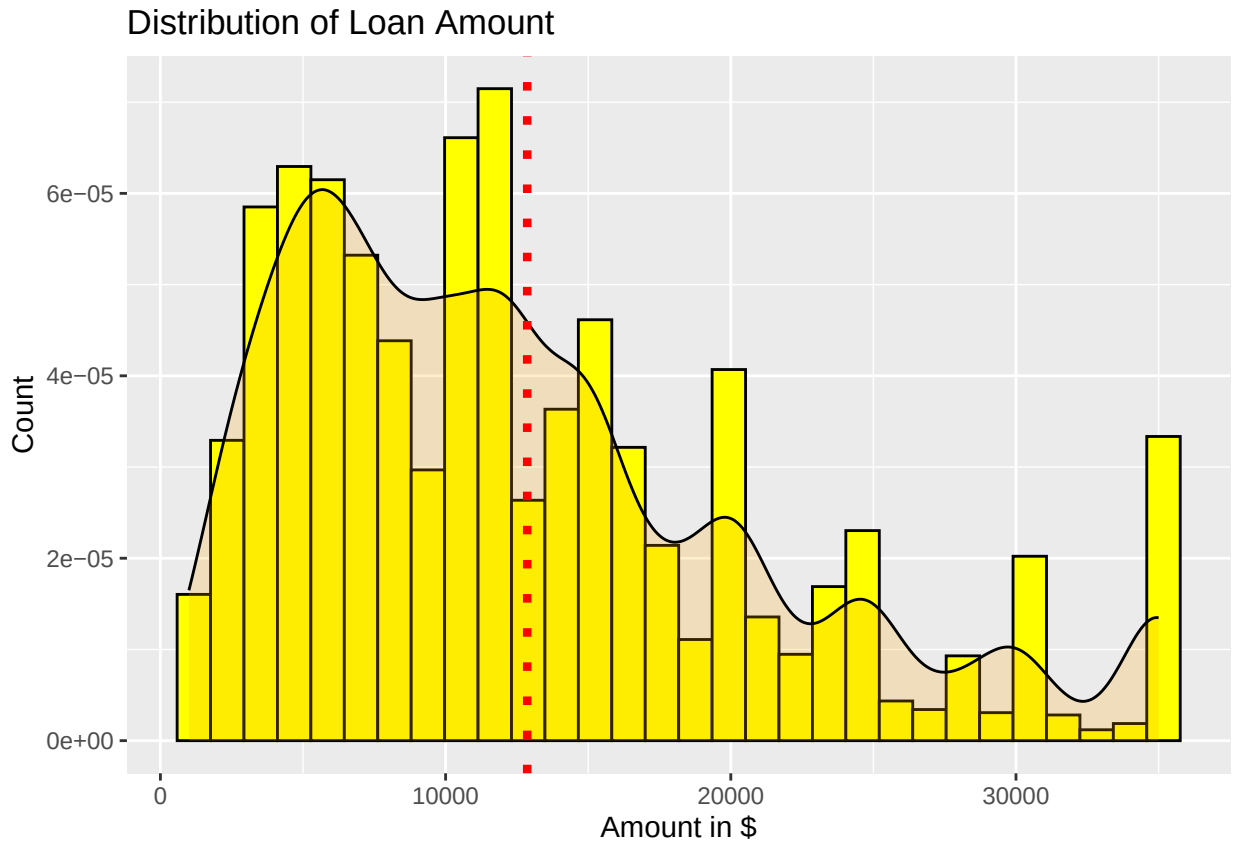
```
?linetype
```

```
## No documentation for 'linetype' in specified packages and libraries:
## you could try '??linetype'
```

```
library(ggplot2)

ggplot(data = loan,
       aes(x = loan_amnt)) +
  geom_histogram(aes(y = ..density..),
                 color = 'black',
                 fill = 'yellow') +
  geom_density(alpha = 0.2,
               fill = 'orange') +
  geom_vline(aes(xintercept = mean(loan_amnt)),
             color = 'red',
             linetype = 3,
             size = 1.5) +
  ggtitle('Distribution of Loan Amount') +
```

```
  xlab('Amount in $') +
  ylab('Count')
```

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

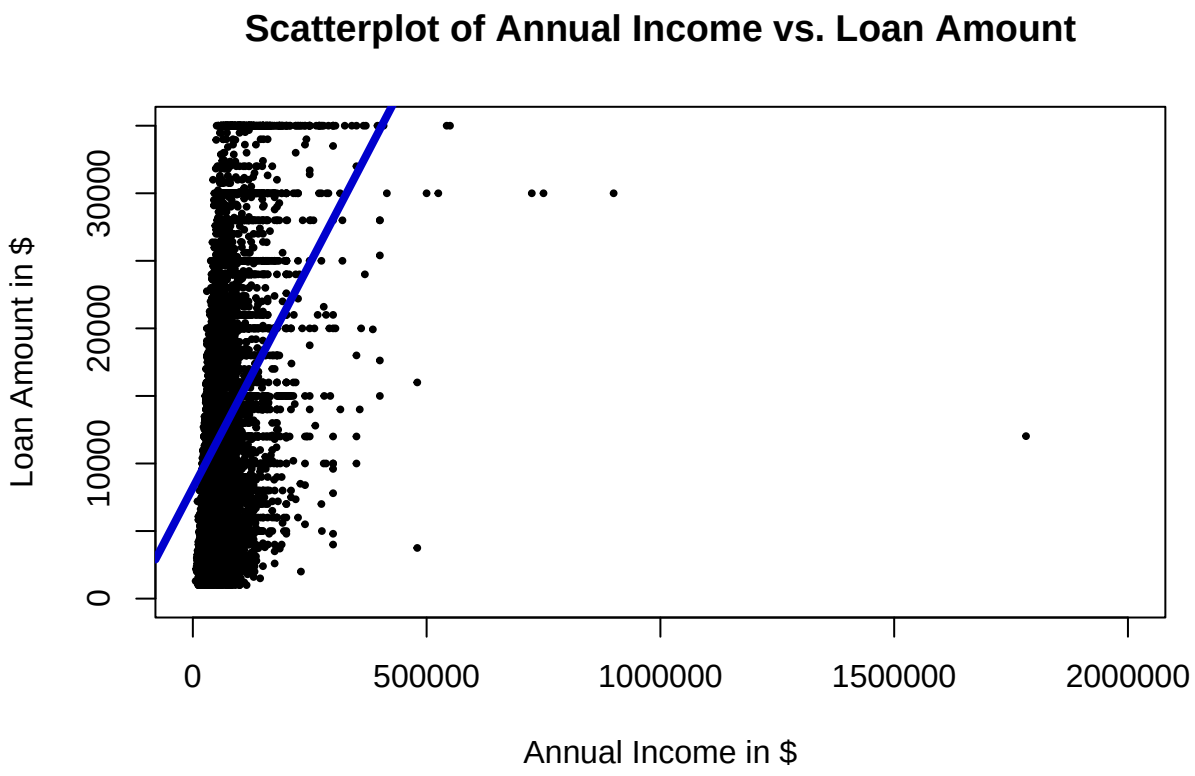



```
# I tried to label mean line but couldn't.
#ggplot(df, aes(x=x, y=y)) +
#geom_point() +
 # geom_vline(xintercept=10) +
  #annotate("text", x=9, y=20, label="Some text", angle=90, size=15, color="blue")
```

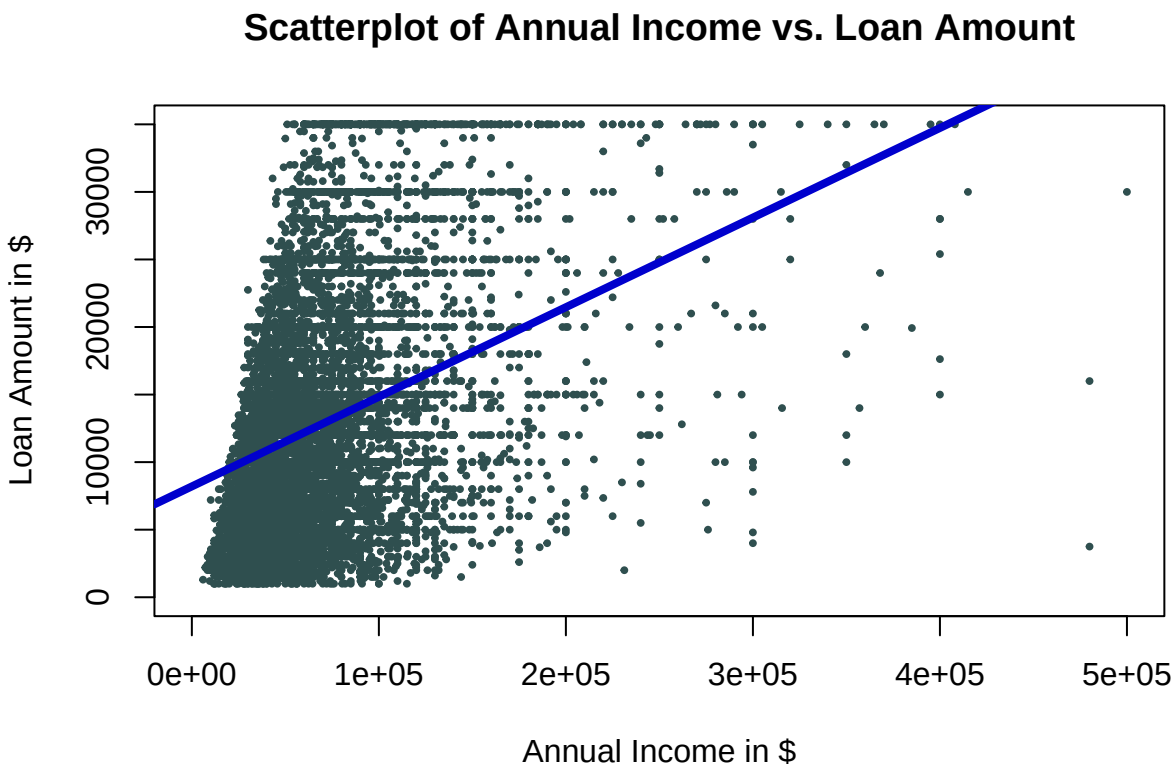
4 The scatter plot of loan_amnt (y-axis) vs. annual_inc (x-axis) + the trend line using basic graphics.

```
?plot

plot(loan$annual_inc, loan$loan_amnt,
     main = 'Scatterplot of Annual Income vs. Loan Amount',
     xlab = 'Annual Income in $',
     ylab = "Loan Amount in $",
     xlim = c(0, 2000000),
```

```
     ylim = c(0, 35000),
     pch = 20,
     cex = 0.6)
abline(lm(loan_amnt ~ annual_inc,
          data = loan),
       lwd = 4,
       col = 'blue3')
```



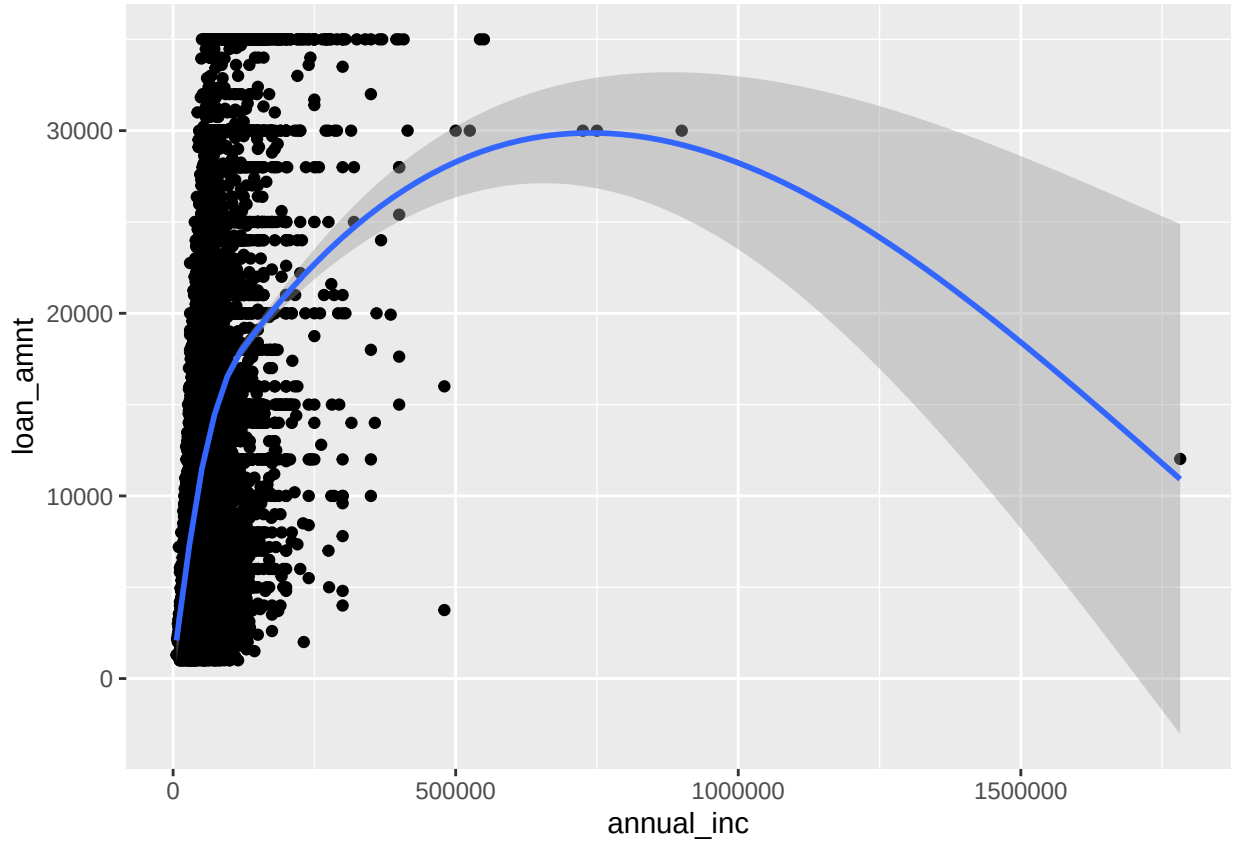

```
# X-axis zoomed 4x
plot(loan$annual_inc, loan$loan_amnt,
     main = 'Scatterplot of Annual Income vs. Loan Amount',
     xlab = 'Annual Income in $',
     ylab = "Loan Amount in $",
     xlim = c(0, 500000),
     ylim = c(0, 35000),
     pch = 20,
     col = 'darkslategrey',
     cex = 0.6)
abline(lm(loan_amnt ~ annual_inc,
          data = loan),
       lwd = 4,
       col = 'blue3')
```



5 Scatter plot of loan_amnt vs.annual_inc with trend line using ggplot2.

```
ggplot(data = loan,
       aes(x = annual_inc, y = loan_amnt)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```




```
# come back and add to parameters and make a new graph removing the outliers
```

6 Barplot of term and grade on the same barplot using basic.

```
# determine if term and grade are numericals
str(loan)
```

```
## spc_tbl_ [10,000 x 11] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ id                 : num [1:10000] 1077501 1077430 1077175 1076863 1075358 ...
##  $ loan_amnt          : num [1:10000] 5000 2500 2400 10000 3000 ...
##  $ term               : chr [1:10000] "36 months" "60 months" "36 months" "36 months" ...
##  $ int_rate           : num [1:10000] 10.6 15.3 16 13.5 12.7 ...
##  $ installment        : num [1:10000] 162.9 59.8 84.3 339.3 67.8 ...
##  $ grade              : chr [1:10000] "B" "C" "C" "C" ...
##  $ emp_length         : chr [1:10000] "10+ years" "< 1 year" "10+ years" "10+ years" ...
##  $ home_ownership     : chr [1:10000] "RENT" "RENT" "RENT" "RENT" ...
##  $ annual_inc         : num [1:10000] 24000 30000 12252 49200 80000 ...
##  $ verification_status: chr [1:10000] "Verified" "Source Verified" "Not Verified" "Source Verified"
##  $ loan_status        : chr [1:10000] "Fully Paid" "Charged Off" "Fully Paid" "Fully Paid" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   id = col_double(),
##   ..   loan_amnt = col_double(),
```

```
##   ..   term = col_character(),
##   ..   int_rate = col_double(),
##   ..   installment = col_double(),
##   ..   grade = col_character(),
##   ..   emp_length = col_character(),
##   ..   home_ownership = col_character(),
##   ..   annual_inc = col_double(),
##   ..   verification_status = col_character(),
##   ..   loan_status = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

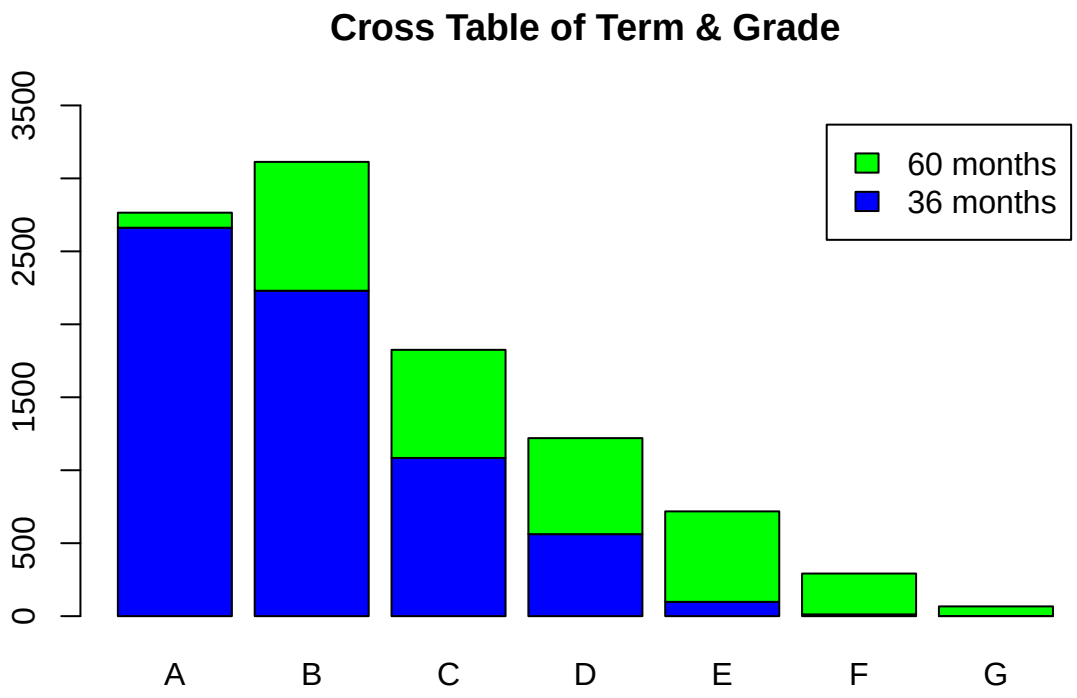
```
loan$term <- as.factor(loan$term)
loan$grade <- as.factor((loan$grade))

# determine number of levels to create colors list
levels(loan$term)
```

```
## [1] "36 months" "60 months"
```

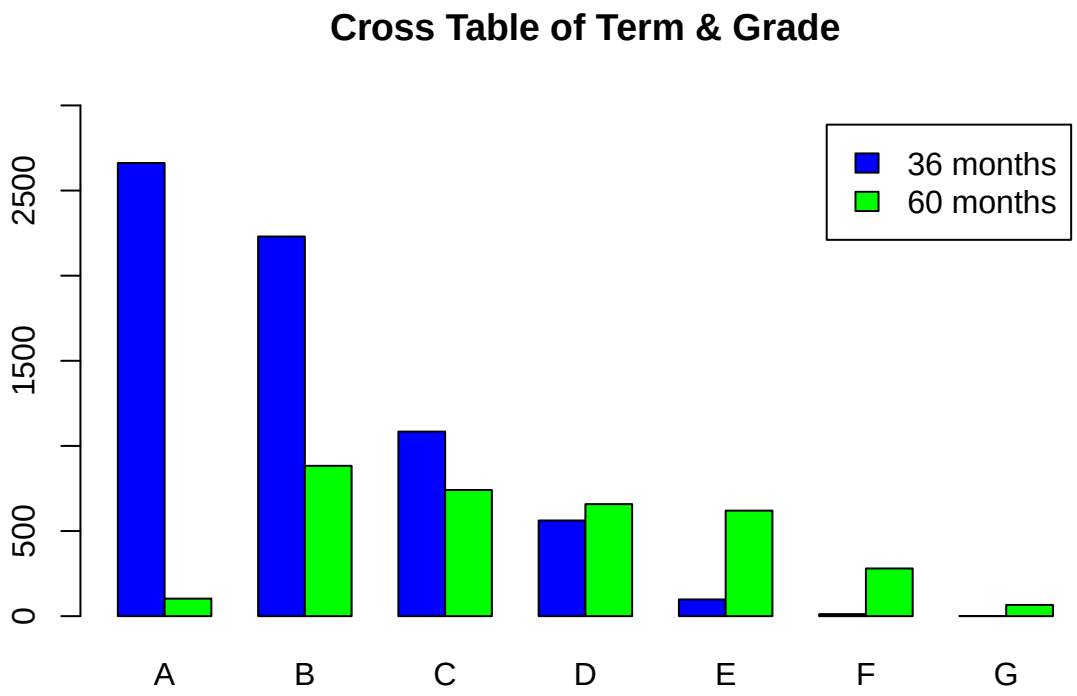
```
levels(loan$grade)
```

```
## [1] "A" "B" "C" "D" "E" "F" "G"
```

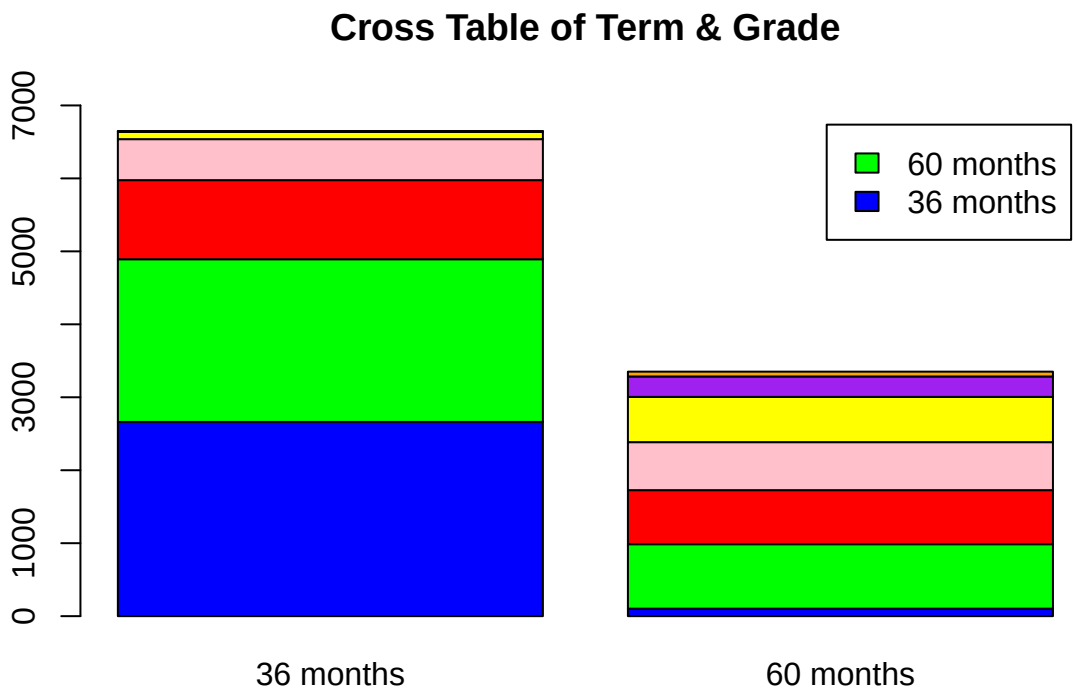
```
# Grade on x-axis
freq_table_term_grade <- xtabs(~ term + grade,
                               data = loan)
barplot(freq_table_term_grade,
        main = 'Cross Table of Term & Grade',
        legend = rownames(freq_table_term_grade),
        col = c('blue', 'green'),
        ylim = c(0, 3500))
```



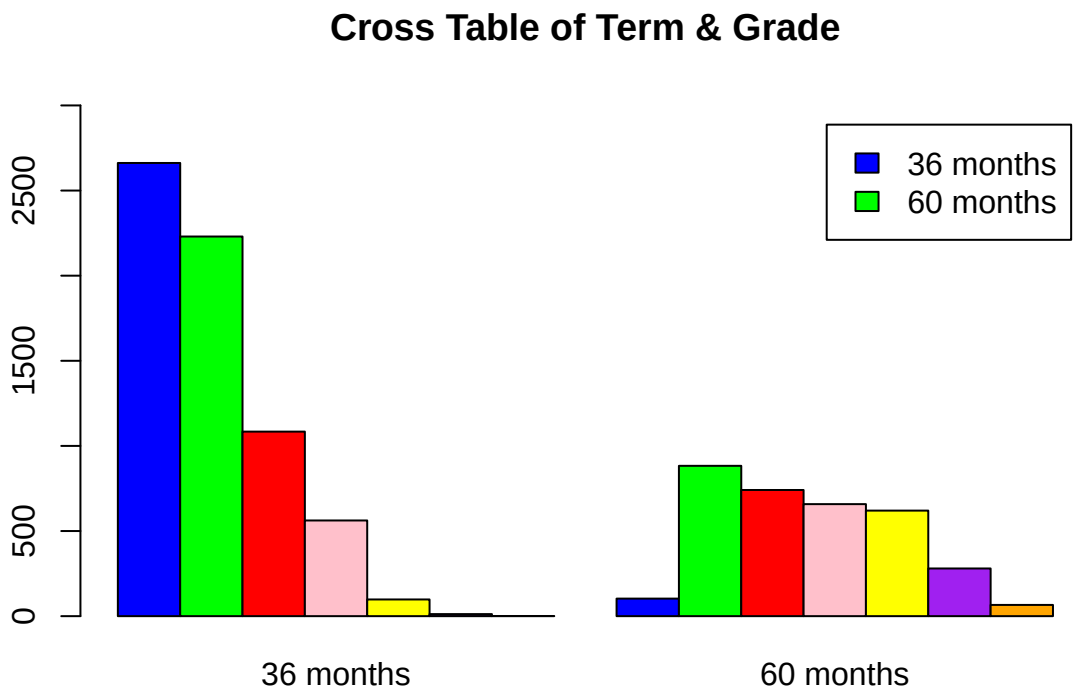

```
barplot(freq_table_term_grade,
        main = 'Cross Table of Term & Grade',
        legend = rownames(freq_table_term_grade),
        col = c('blue', 'green'),
        beside = TRUE,
        ylim = c(0, 3000))
```




```
# Term on x-axis
freq_table_term_grade_2 <- xtabs(~ grade + term,
                                 data = loan)

barplot(freq_table_term_grade_2,
        main = 'Cross Table of Term & Grade',
        legend = rownames(freq_table_term_grade),
        col = c('blue', 'green', 'red', 'pink', 'yellow', 'purple', 'orange'),
        ylim = c(0, 7000))
```

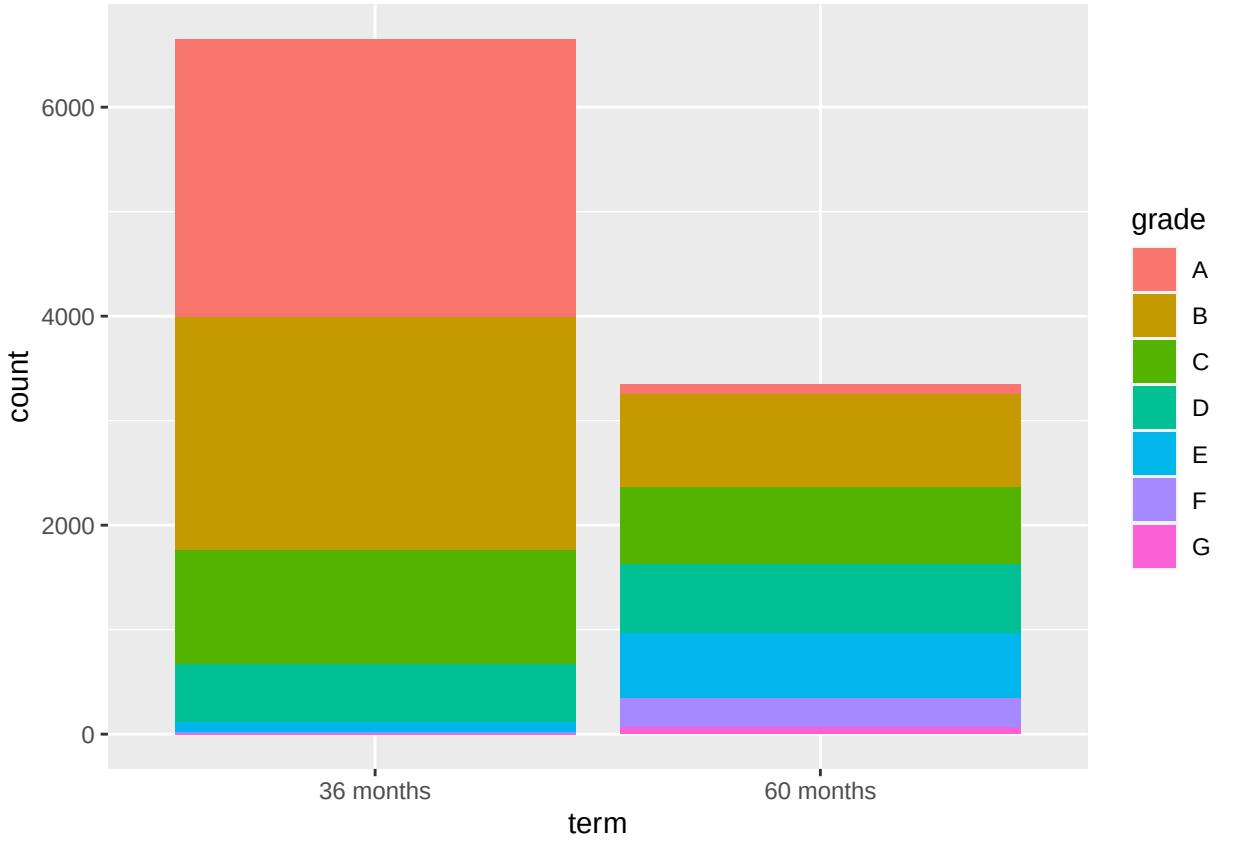



```
barplot(freq_table_term_grade_2,
        main = 'Cross Table of Term & Grade',
        legend = rownames(freq_table_term_grade),
        col = c('blue', 'green', 'red', 'pink', 'yellow', 'purple', 'orange'),
        beside = TRUE,
        ylim = c(0, 3000))
```

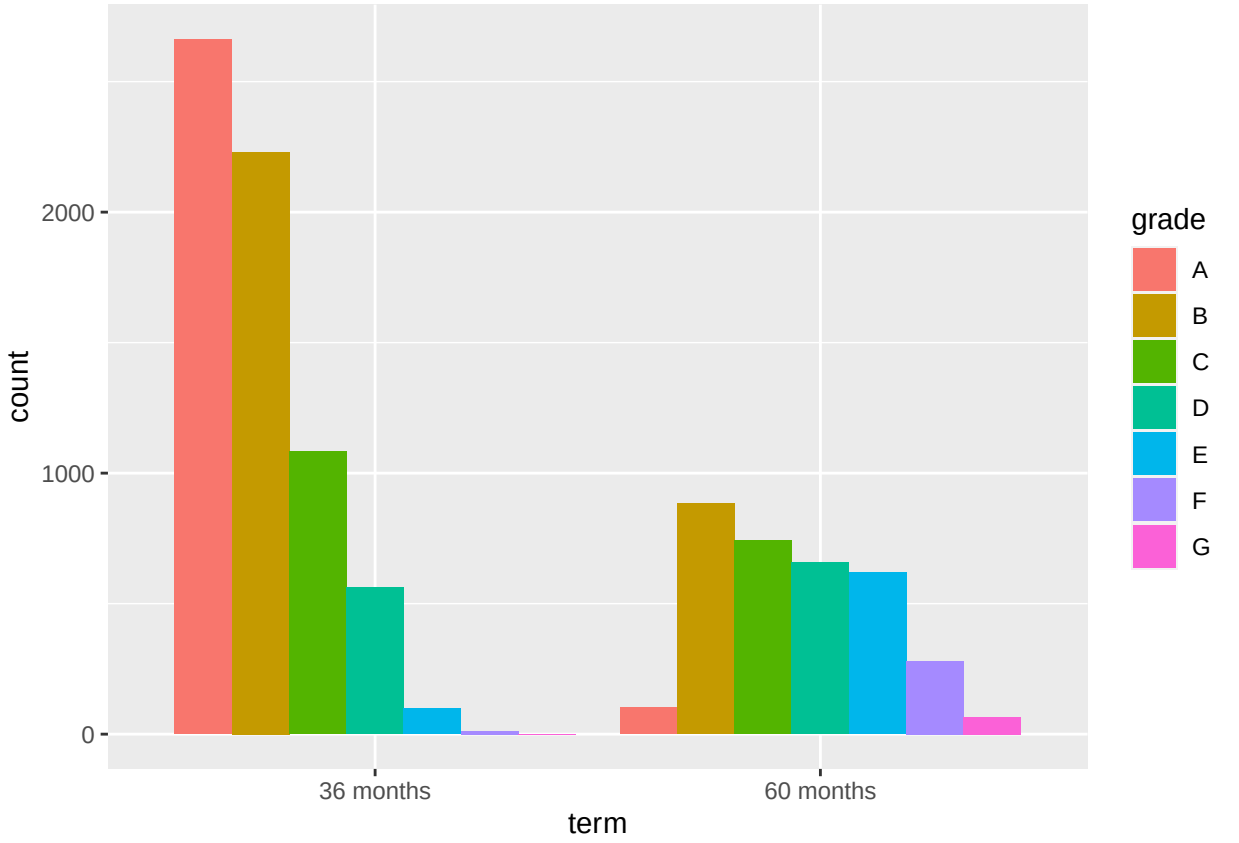


7 Barplot of term and grade on the same barplot using ggplot2

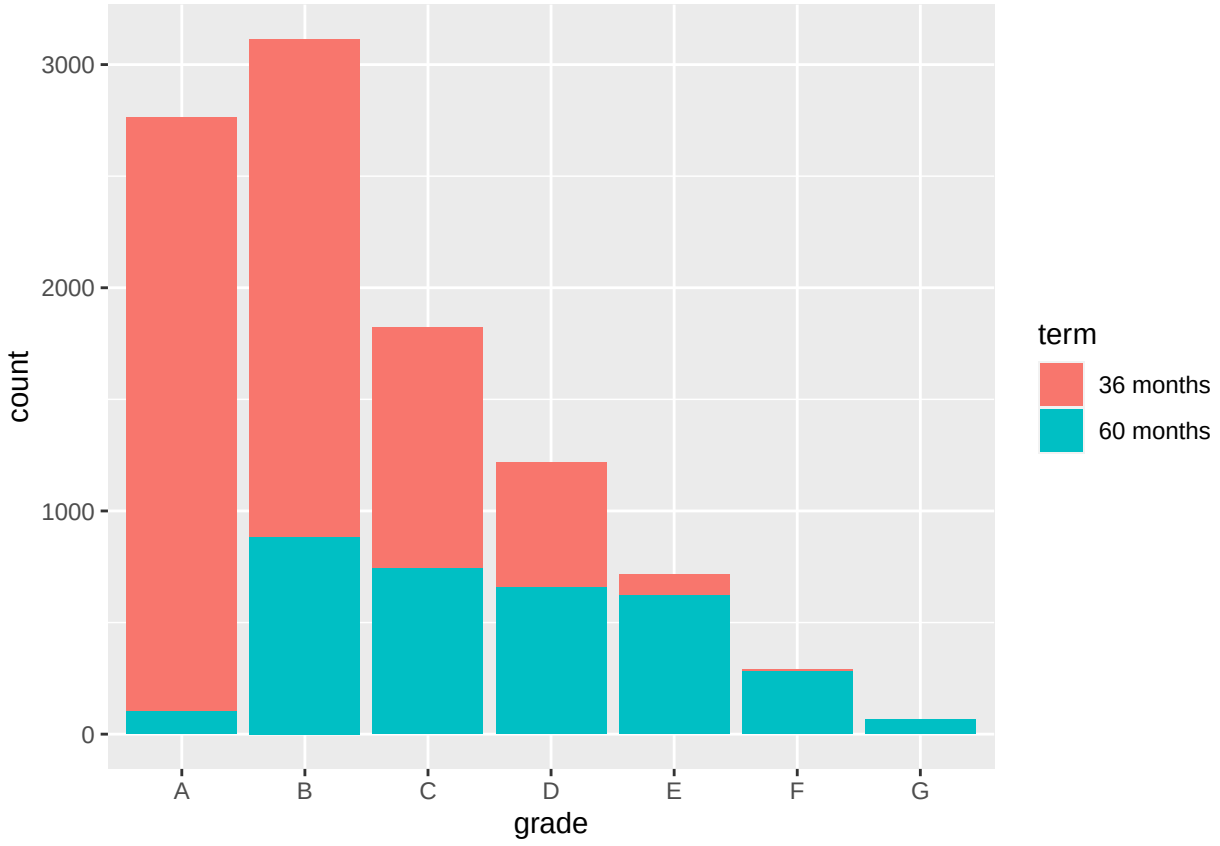
```
# By term
ggplot(data = loan,
       aes(x = term, y = ..count..)) +
  geom_bar(aes(fill = grade))
```



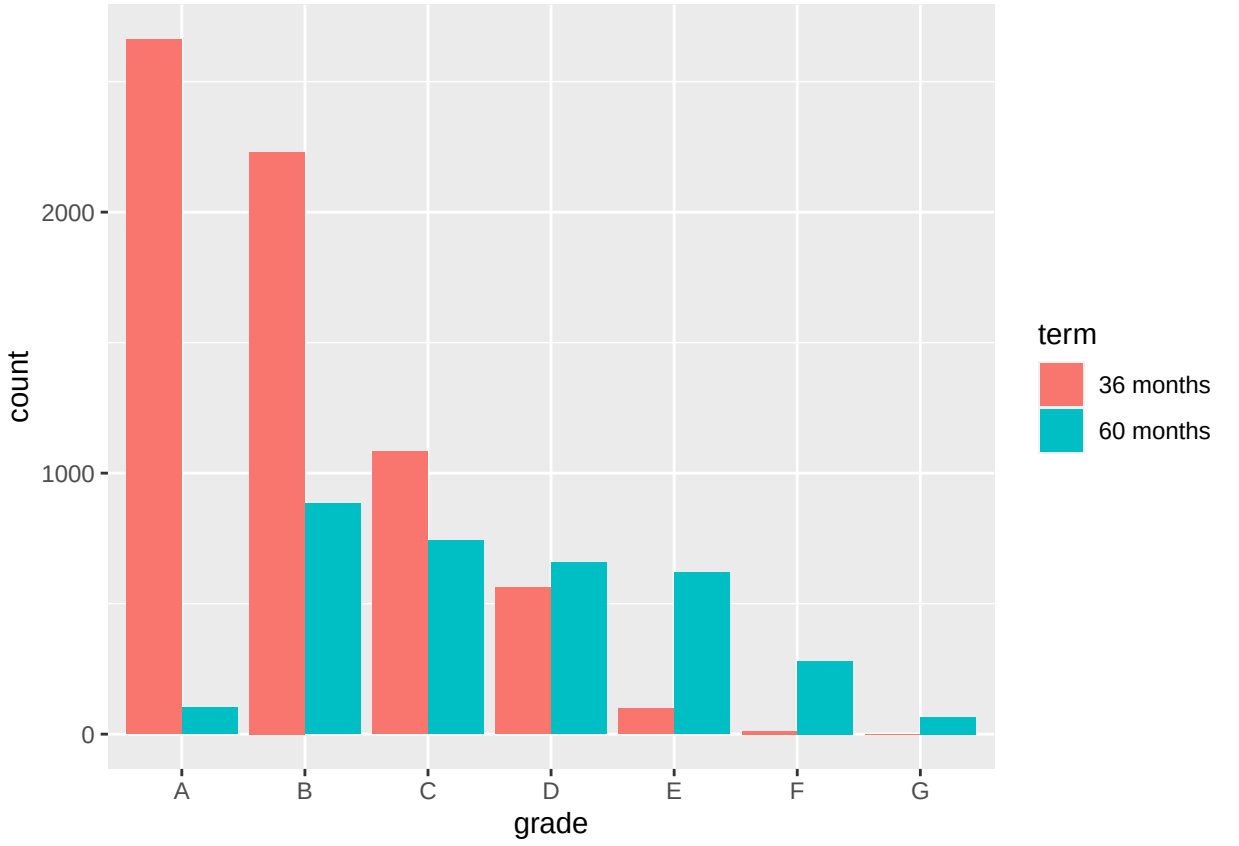

```
ggplot(data = loan,
       aes(x = term, y = ..count..)) +
  geom_bar(aes(fill = grade),
           position = 'dodge')
```




```
# By grade
ggplot(data = loan,
       aes(x = grade, y = ..count..)) +
  geom_bar(aes(fill = term))
```




```
ggplot(data = loan,
       aes(x = grade, y = ..count..)) +
  geom_bar(aes(fill = term),
           position = 'dodge')
```



8 Boxplot loan_amnt vs.term and save as 'loanterm.jpg' using basic graphics.

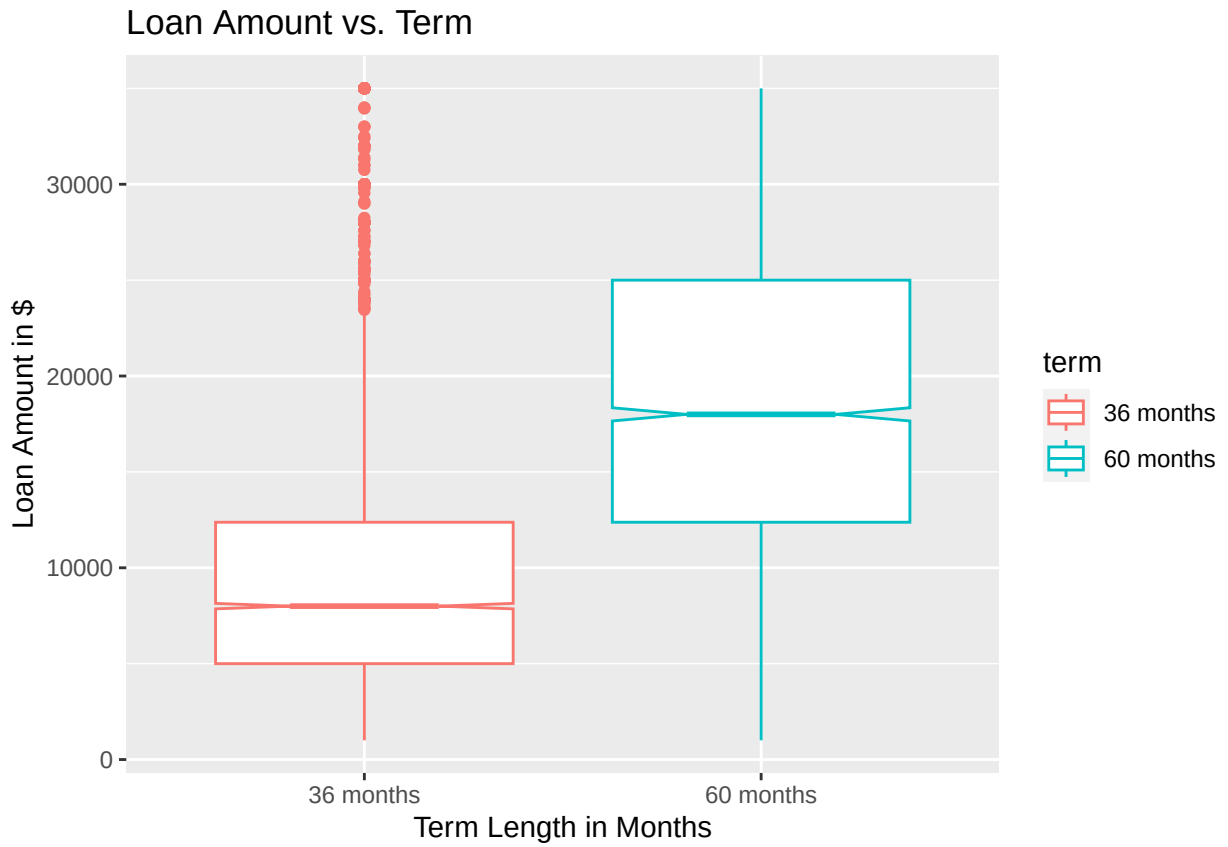
```
jpeg("C:\\Users\\mmsax\\Desktop\\_MU\\GDSCI_502_R\\Week05\\loanterm.jpg")

boxplot(loan_amnt ~ term,
        data = loan,
        notch = TRUE,
        col = c('blue'),
        main = 'Loan Amount by Term',
        xlab = 'Term',
        ylim = c(0, 35000))

dev.off()
```

```
## pdf
##   2
```

9 Boxplot loan_amnt vs.term and save as 'loanterm.jpg' using ggplot2.

State differences between loan amount with respect to term.

```
ggplot(data = loan,
       aes(x = term,
           y = loan_amnt)) +
  geom_boxplot(aes(col = term),
               notch = TRUE) +
  ggtitle('Loan Amount vs. Term') +
  xlab('Term Length in Months') +
  ylab('Loan Amount in $')
```




```
ggsave("C:\\Users\\mmsax\\OneDrive\\Desktop\\_MU\\_GDSCI_502_R\\Week05\\loanterm2.jpg", width = 20, hei

print('There is a significant diffence between the average as well as the inner two quartiles of loan a
```

```
## [1] "There is a significant diffence between the average as well as the inner two quartiles of loan a
```