# Assignment 2

Eun Ho Kim(12455415), Younjoo Mo (12475440)

6/13/2020

## Question 1

```
library(tidyverse)
```

```
## -- Attaching packages -----------------------------------------------------------------------------------------

## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  3.0.0     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts -------------------------------------------------------------------------------- ti
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
load('credit.rda')
```

Before we start, if we look at the credit table, installment_rate, residence_since, existing_credits and num_dependents are coded with numbers like 1,2,3. . . ., which are inconsistent with the data description. Therefore, we want to change these codes to be same as data description.

```
credit$installment_rate=recode(credit$installment_rate,'1'='A81','2'='A82', '3'='A83','4'='A84')
credit$residence_since=recode(credit$residence_since,'1'='A71','2'='A73','3'='A74','4'='A75')
credit$existing_credits=recode(credit$existing_credits,'1'='A161','2'='A162','3'='A163','4'='A164')
credit$num_dependents=recode(credit$num_dependents, '1'='A181' , '2'='A182')
```

Also, we have to determine whether there exist N/A on data description

```
sum(is.na(credit))
```

```
## [1] 0
```

There is no N/A on data description, so we can just analyze data.

Now, we want to know which variables are categorical and which are numerical.

```
sapply(credit,class,simplify = TRUE, USE.NAMES = TRUE)
```

```
##     checking_status            duration      credit_history             purpose
##            "factor"           "integer"            "factor"            "factor"
##       credit_amount             savings          employment    installment_rate
##           "integer"            "factor"            "factor"         "character"
##     personal_status       other_parties     residence_since property_magnitude
##            "factor"            "factor"         "character"            "factor"
##                 age other_payment_plans             housing    existing_credits
##           "integer"            "factor"            "factor"         "character"
##                 job      num_dependents           telephone      foreign_worker
```

```
##           "factor"        "character"           "factor"            "factor"
##              class
##        "character"
```

Here, “factor” and “character” are categorical variables, “interger” is numerical variable.

By using summary function, we can analyze variables by numerically.

```r
summary(credit)
```

```
##  checking_status    duration    credit_history    purpose    credit_amount
##  A11:274         Min.   : 4.0   A30: 40        A43    :280   Min.   :  250
##  A12:269         1st Qu.:12.0   A31: 49        A40    :234   1st Qu.: 1366
##  A13: 63         Median :18.0   A32:530        A42    :181   Median : 2320
##  A14:394         Mean   :20.9   A33: 88        A41    :103   Mean   : 3271
##                  3rd Qu.:24.0   A34:293        A49    : 97   3rd Qu.: 3972
##                  Max.   :72.0                  A46    : 50   Max.   :18424
##                                                (Other): 55
##  savings   employment installment_rate   personal_status other_parties
##  A61:603   A71: 62    Length:1000        A91: 50         A101:907
##  A62:103   A72:172    Class :character   A92:310         A102: 41
##  A63: 63   A73:339    Mode  :character   A93:548         A103: 52
##  A64: 48   A74:174                       A94: 92
##  A65:183   A75:253
##
##
##  residence_since    property_magnitude      age        other_payment_plans
##  Length:1000        A121:282           Min.   :19.00   A141:139
##  Class :character   A122:232           1st Qu.:27.00   A142: 47
##  Mode  :character   A123:332           Median :33.00   A143:814
##                     A124:154           Mean   :35.55
##                                        3rd Qu.:42.00
##                                        Max.   :75.00
##
##  housing    existing_credits     job      num_dependents     telephone
##  A151:179   Length:1000        A171: 22   Length:1000        A191:596
##  A152:713   Class :character   A172:200   Class :character   A192:404
##  A153:108   Mode  :character   A173:630   Mode  :character
##                                A174:148
##
##
##
##  foreign_worker    class
##  A201:963       Length:1000
##  A202: 37       Class :character
##                 Mode  :character
##
##
##
##
```

But this omit the details of purpose,installment_rate,residence_since, existing_credits and num_dependents. So we also added details of them.

```r
summary(credit$purpose)
```

```
##  A40  A41 A410  A42  A43  A44  A45  A46  A48  A49
```

```
##   234   103    12   181   280    12    22    50     9    97
```

```r
table(credit$installment_rate)
```

```
##
## A81 A82 A83 A84
## 136 231 157 476
```

```r
table(credit$residence_since)
```

```
##
## A71 A73 A74 A75
## 130 308 149 413
```

```r
table(credit$existing_credits)
```

```
##
## A161 A162 A163 A164
##  633  333   28    6
```

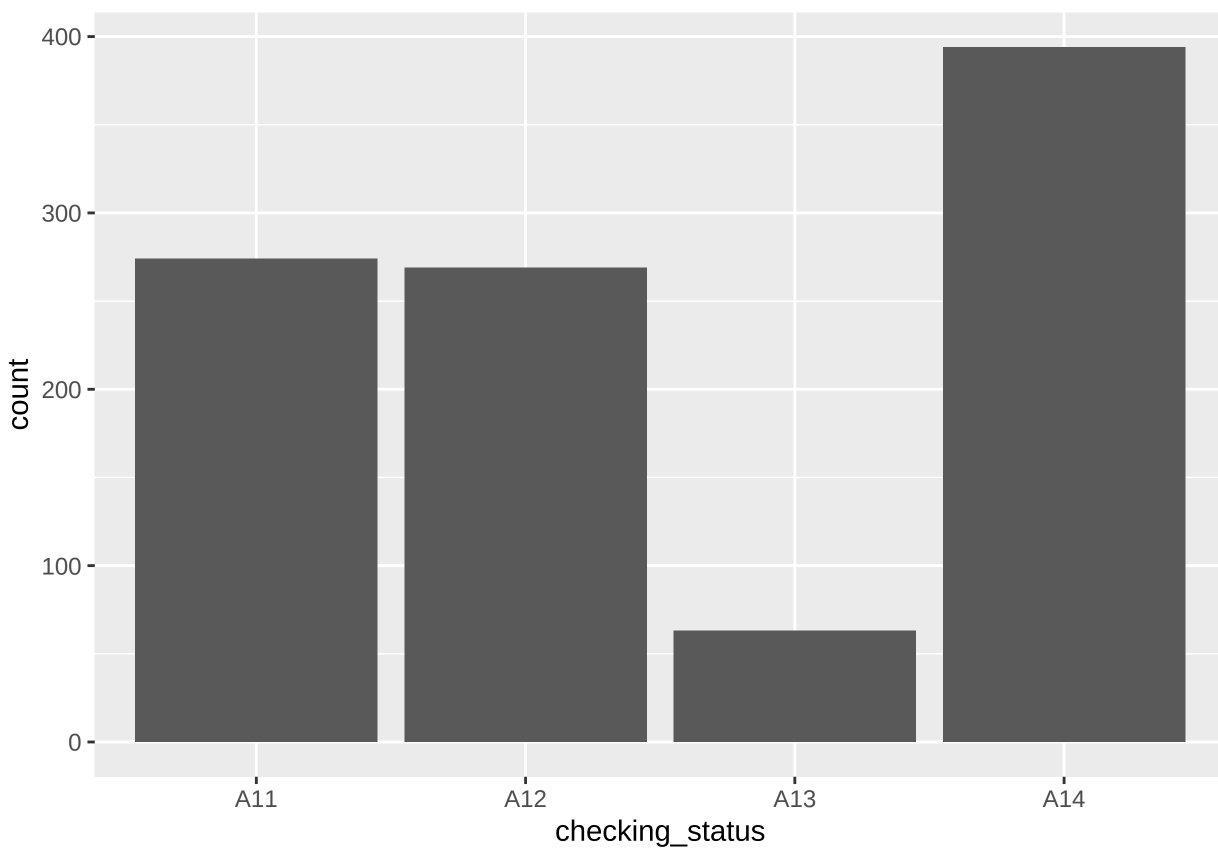
```r
table(credit$num_dependents)
```

```
##
## A181 A182
##  845  155
```

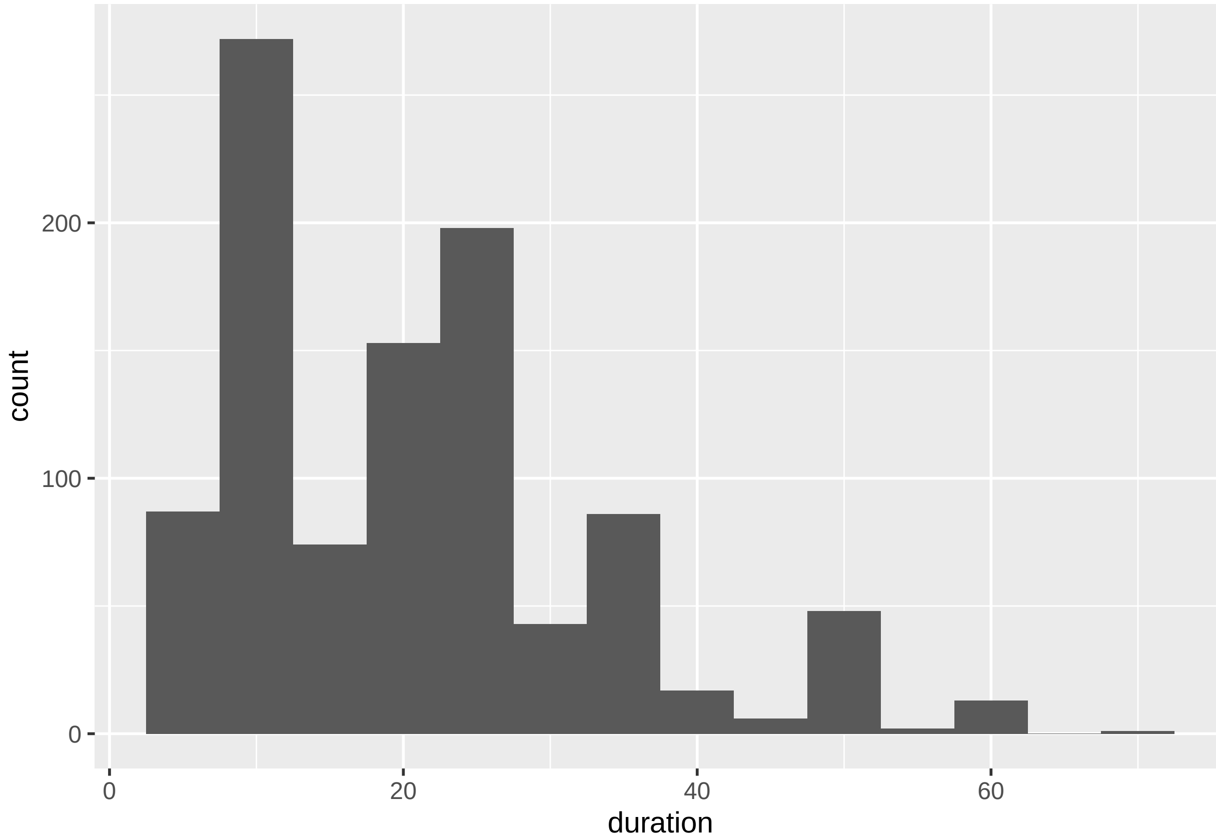
Then, we will analyze variables graphically by using ggplot function.

```r
a=ggplot(data=credit)
```

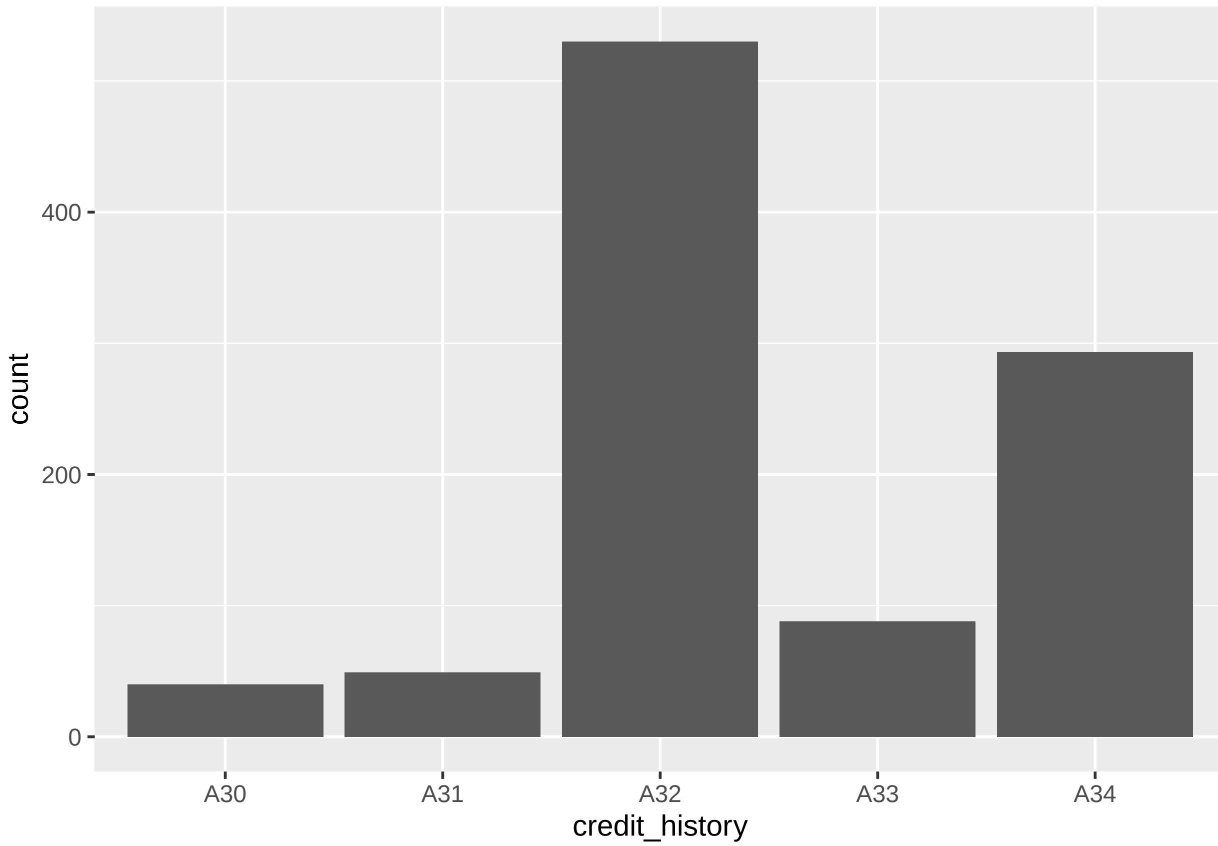
```r
a+geom_bar(aes(checking_status))
```



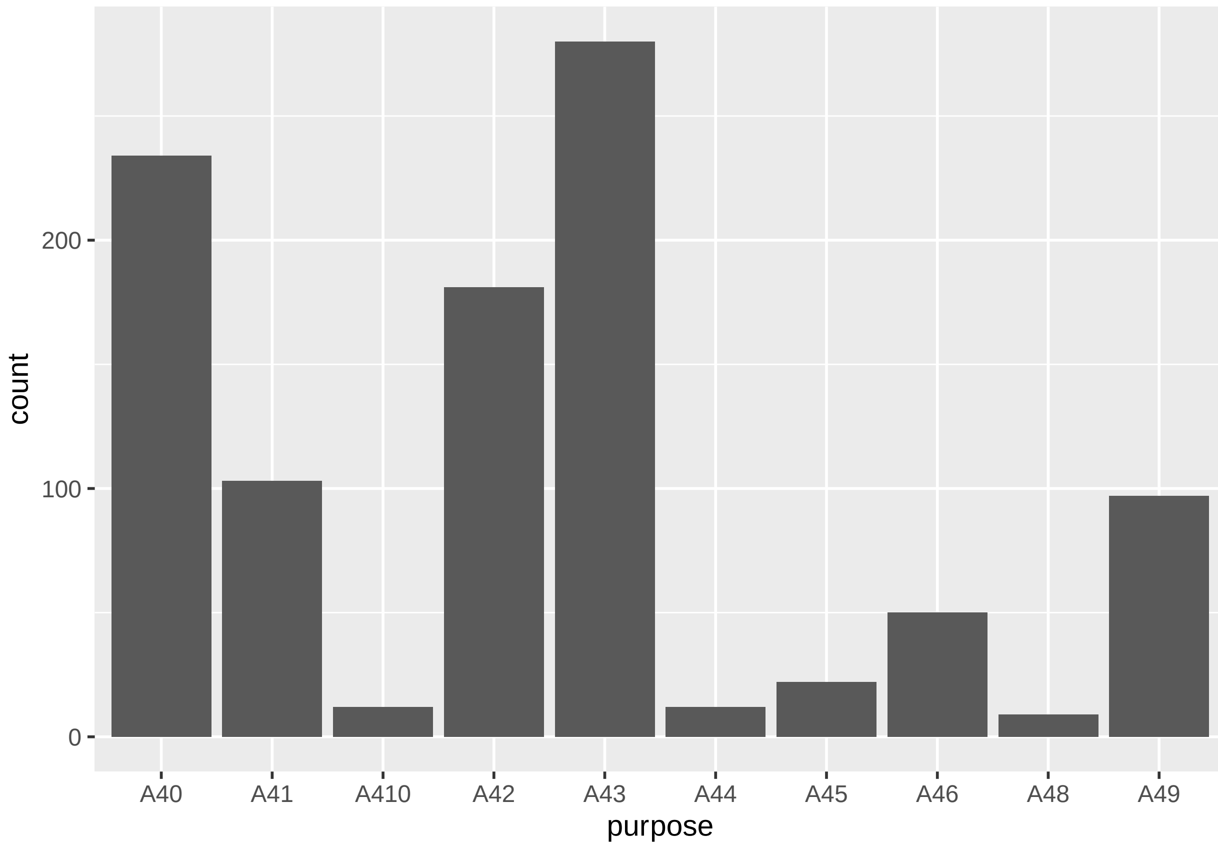
```r
a+geom_histogram(aes(duration), binwidth = 5)
```



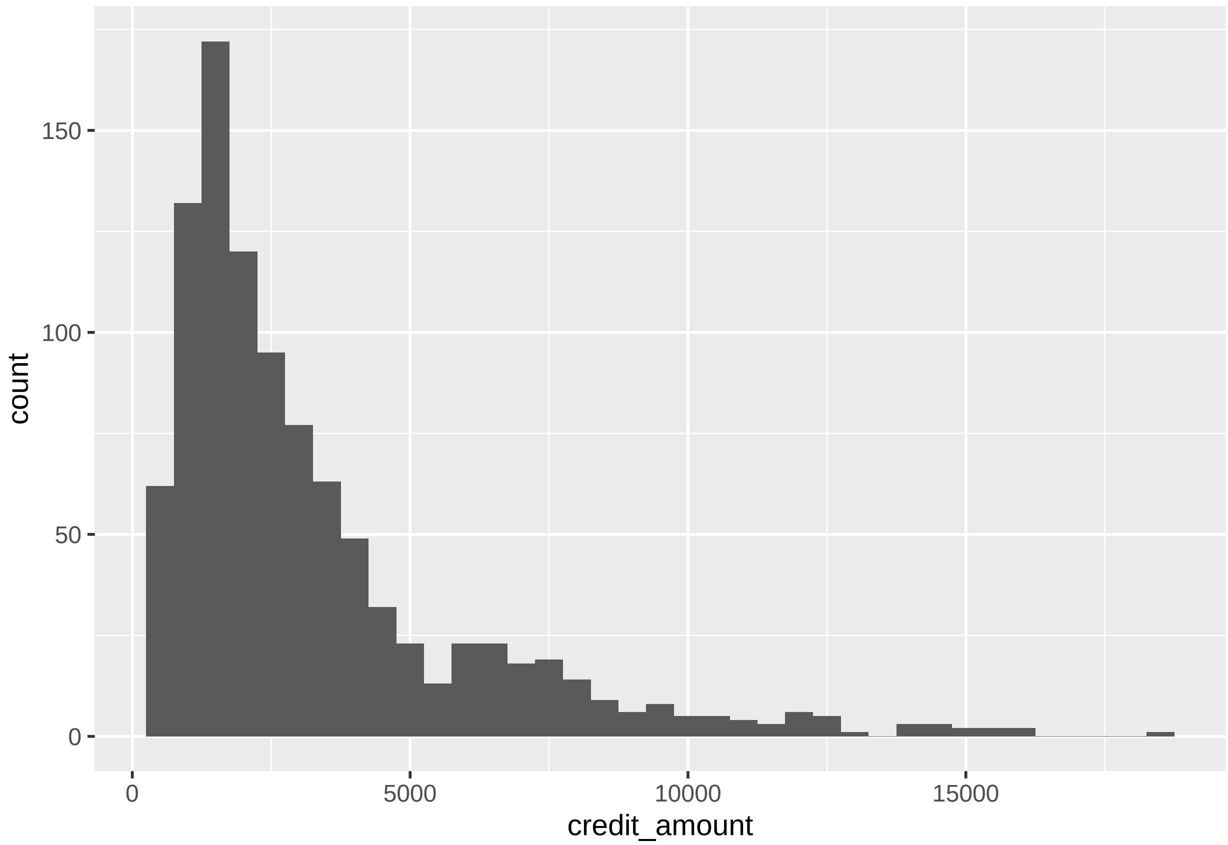

```r
a+geom_bar(aes(credit_history))
```



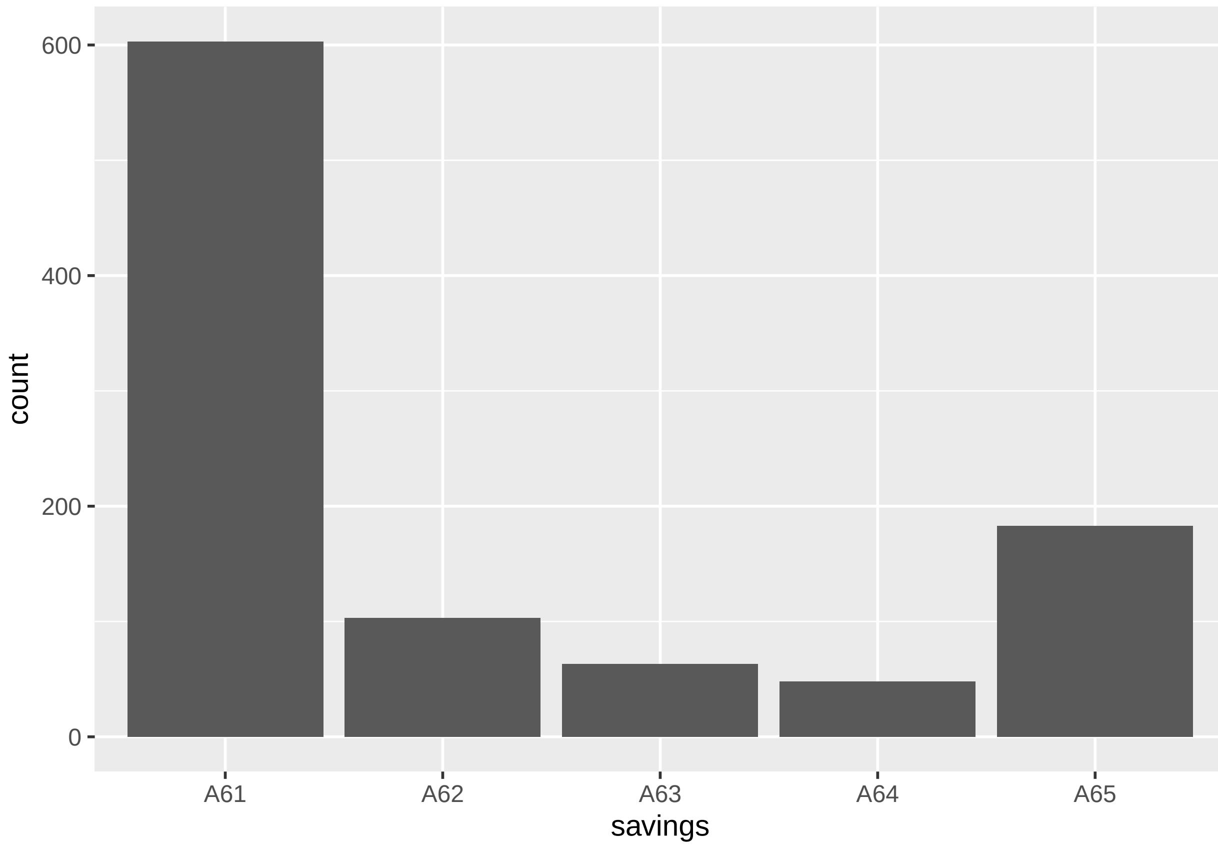

```
a+geom_bar(aes(purpose))
```



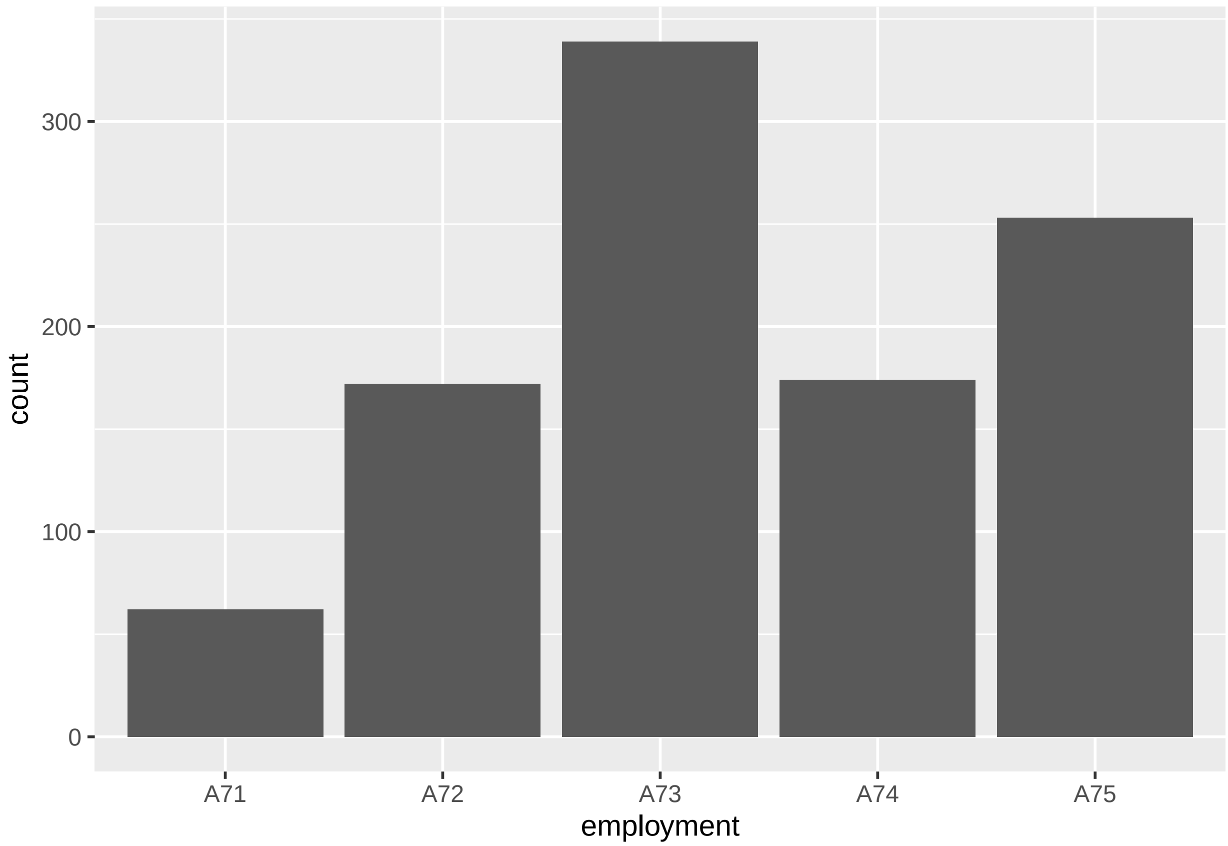

```
a+geom_histogram(aes(credit_amount), binwidth = 500)
```



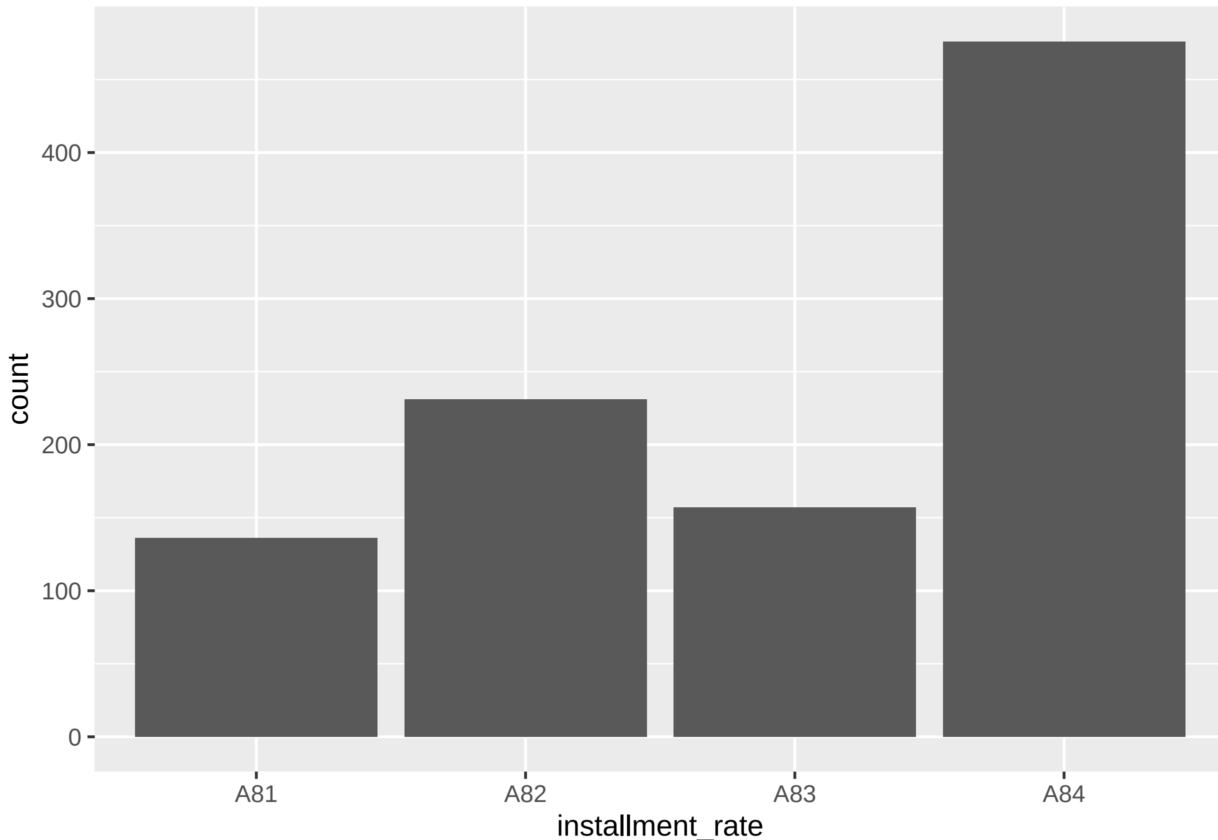

```
a+geom_bar(aes(savings))
```



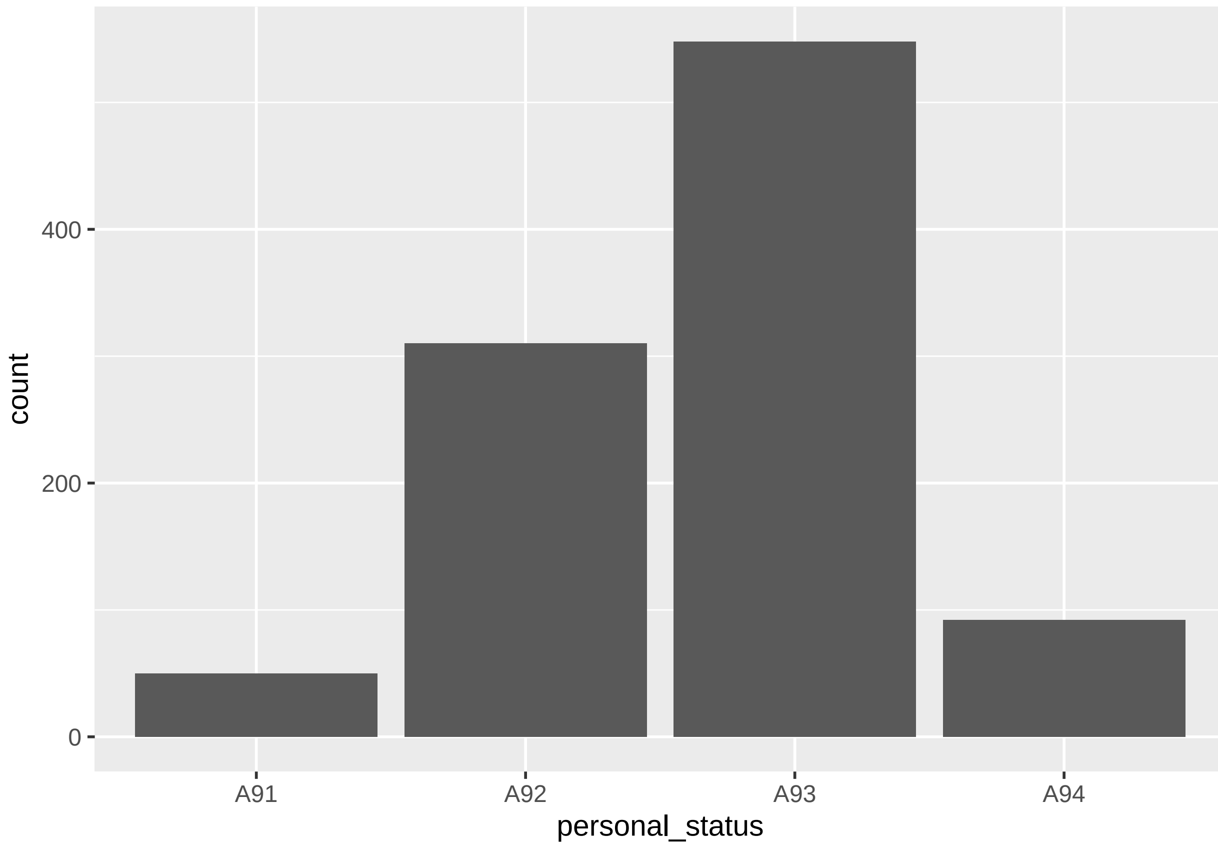

```
a+geom_bar(aes(employment))
```



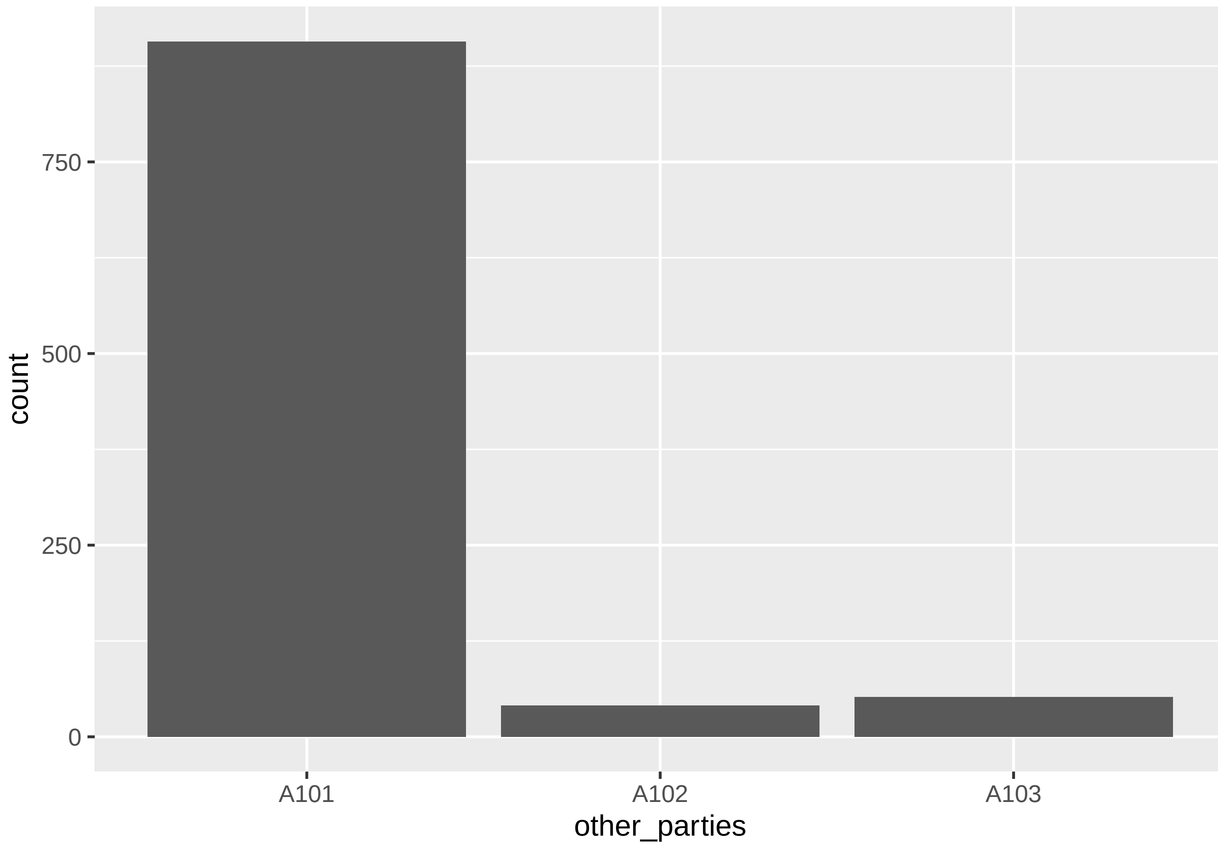

```
a+geom_bar(aes(installment_rate))
```



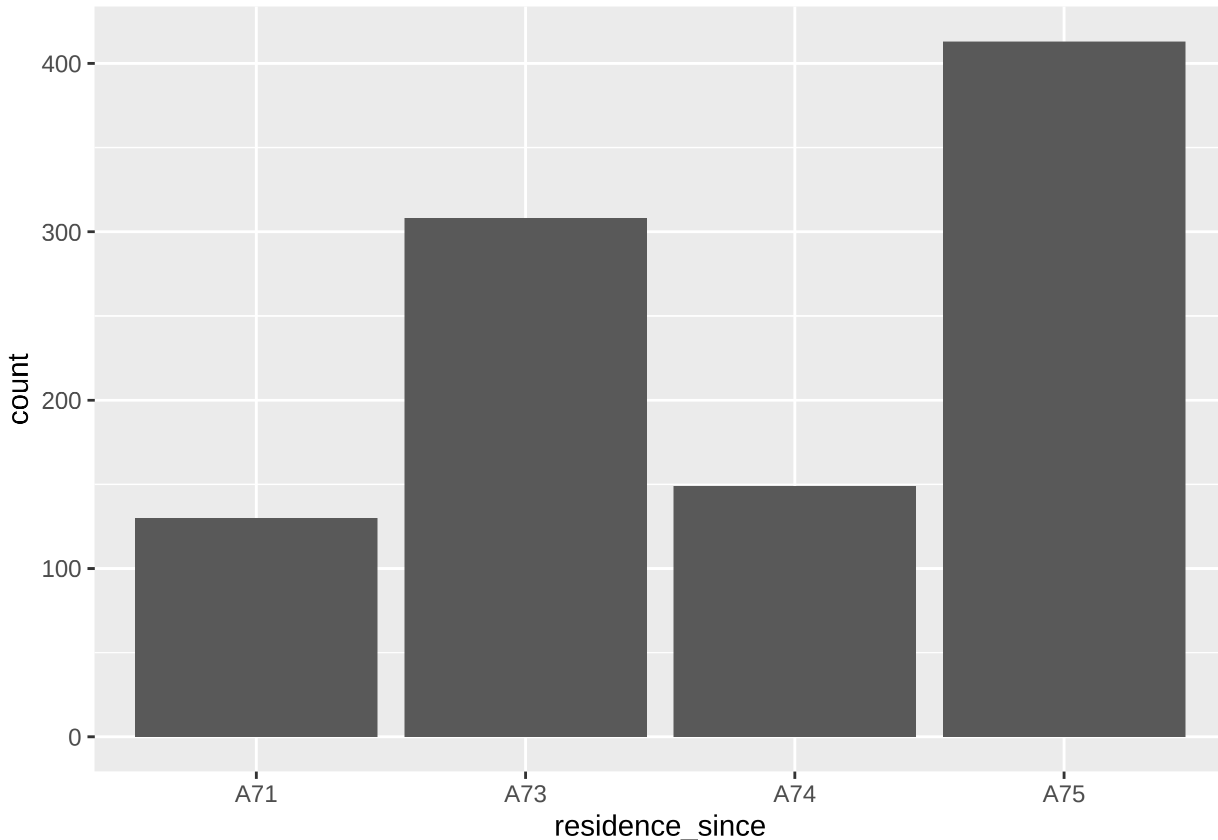

```
a+geom_bar(aes(personal_status))
```



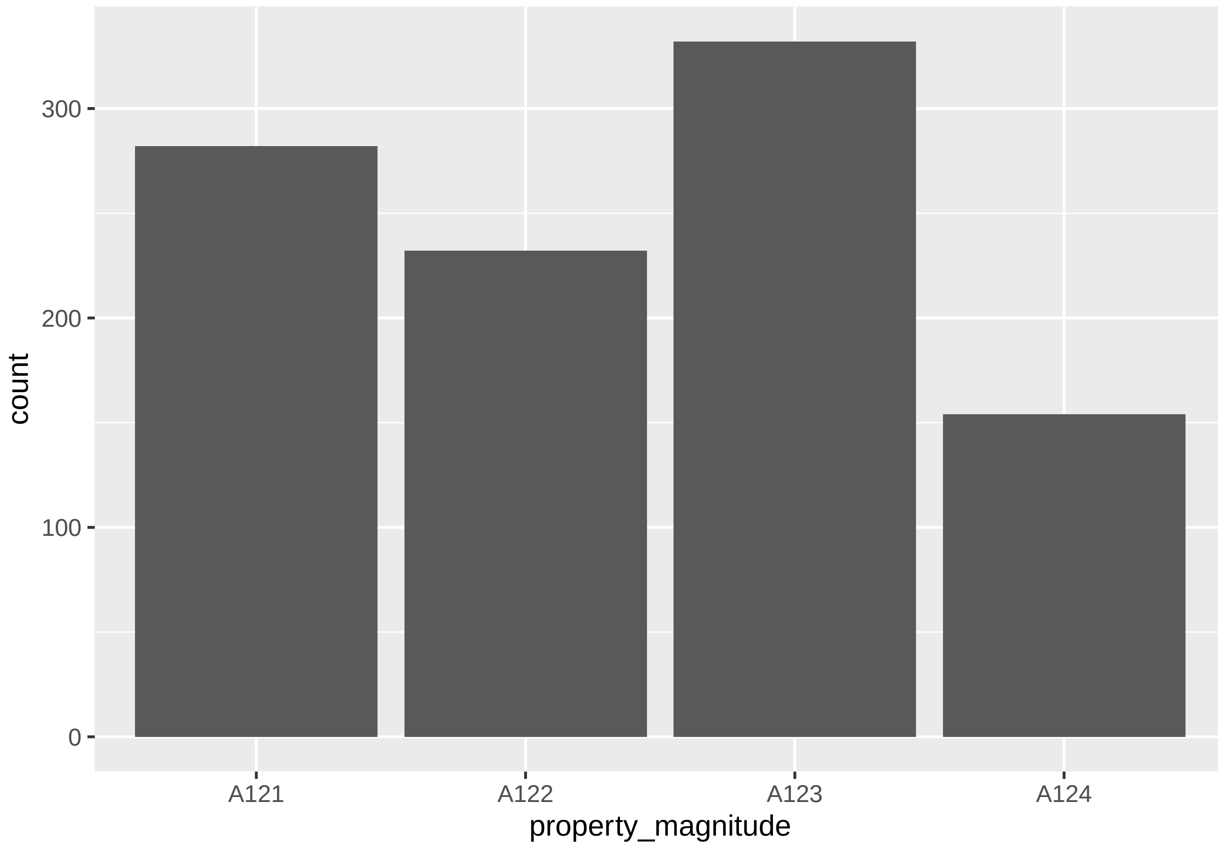

```
a+geom_bar(aes(other_parties))
```



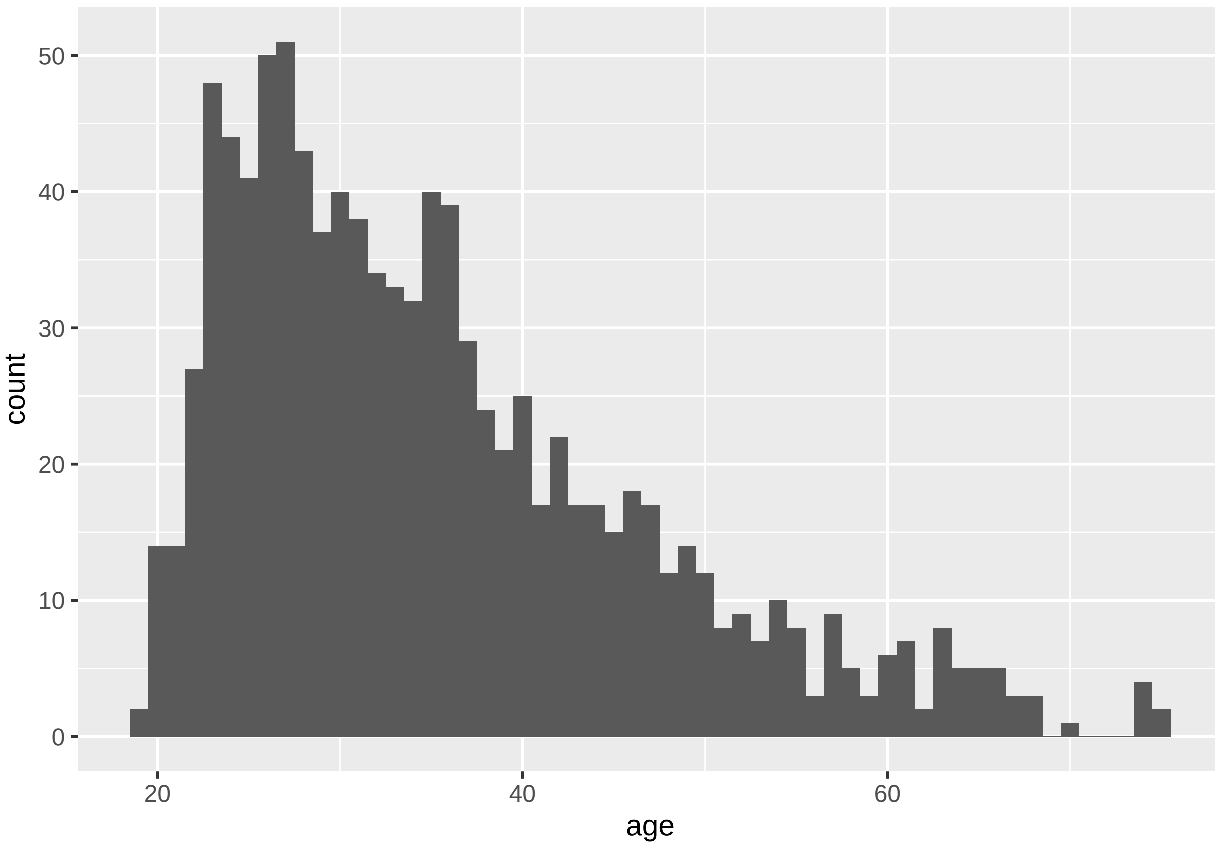

```
a+geom_bar(aes(residence_since))
```



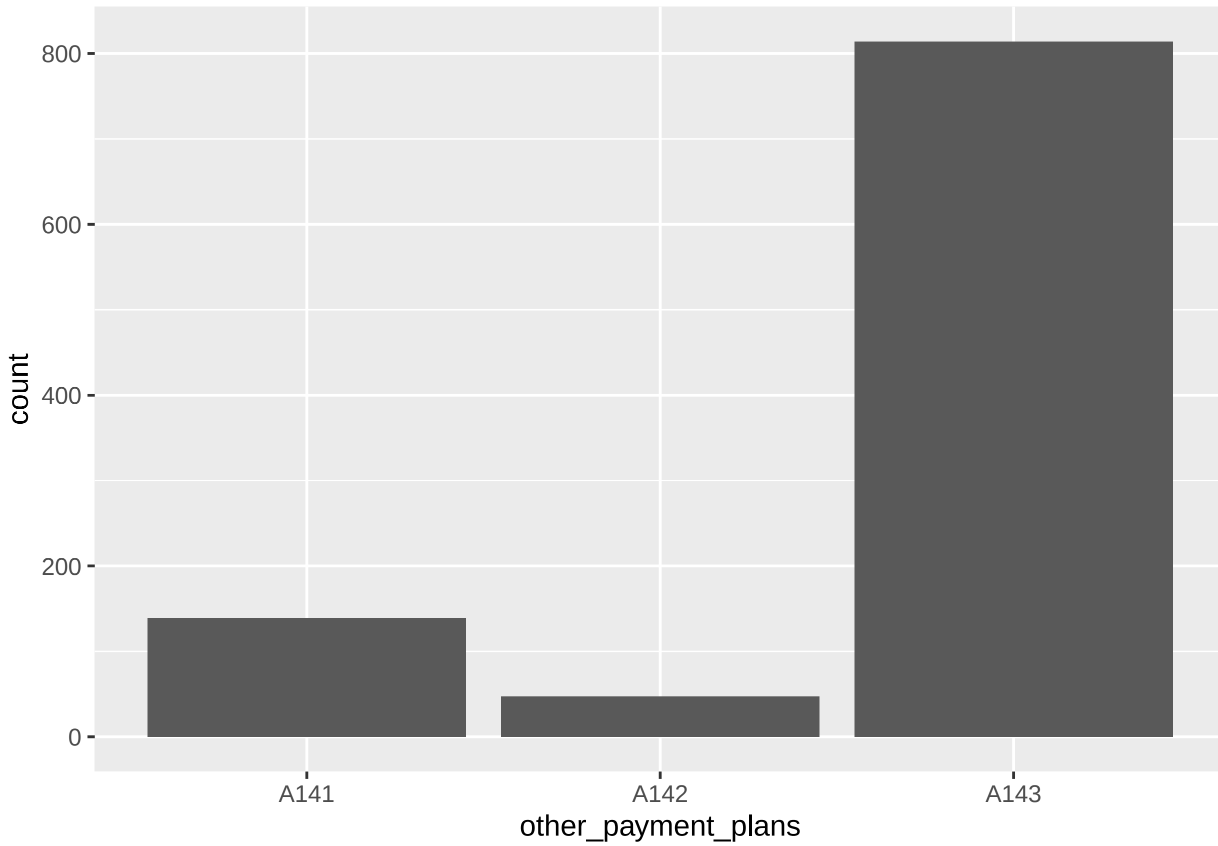

```
a+geom_bar(aes(property_magnitude))
```



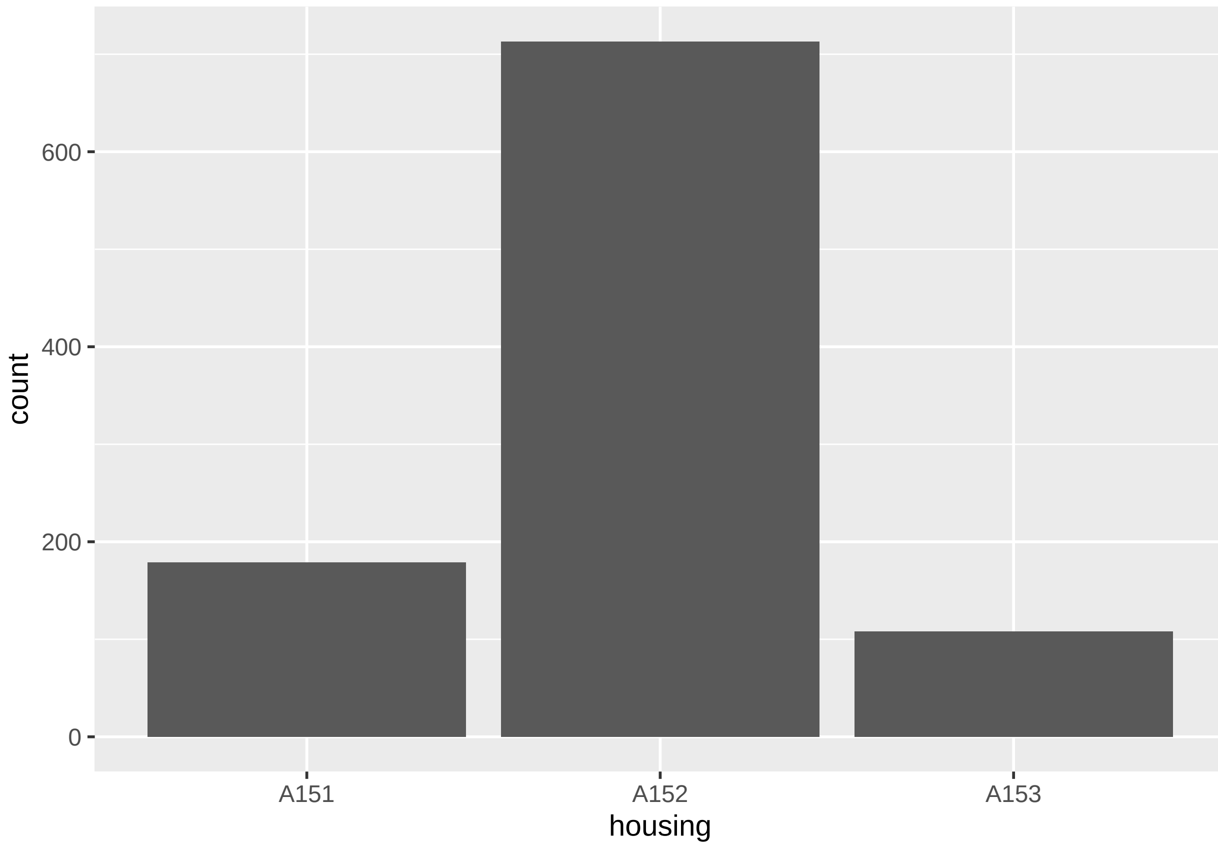

```r
a+geom_histogram(aes(age), binwidth = 1)
```



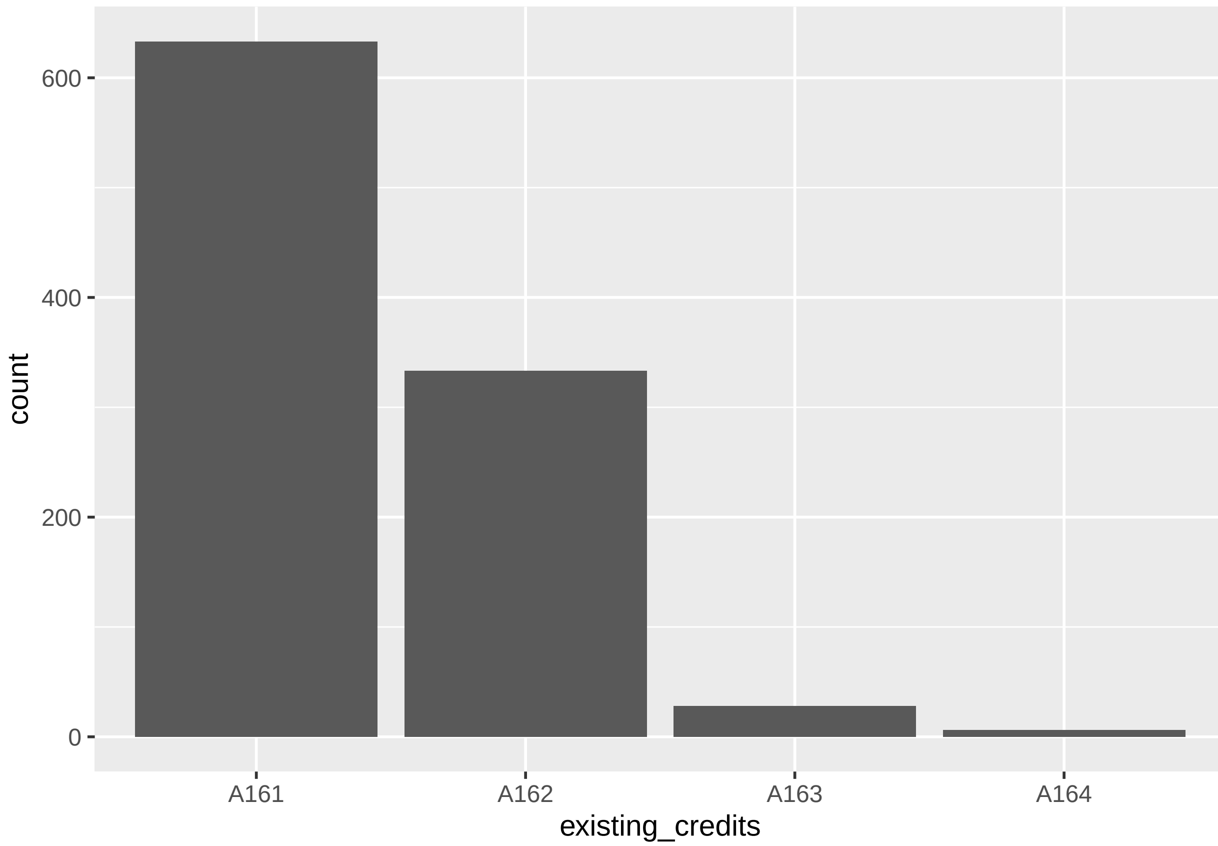

```r
a+geom_bar(aes(other_payment_plans))
```



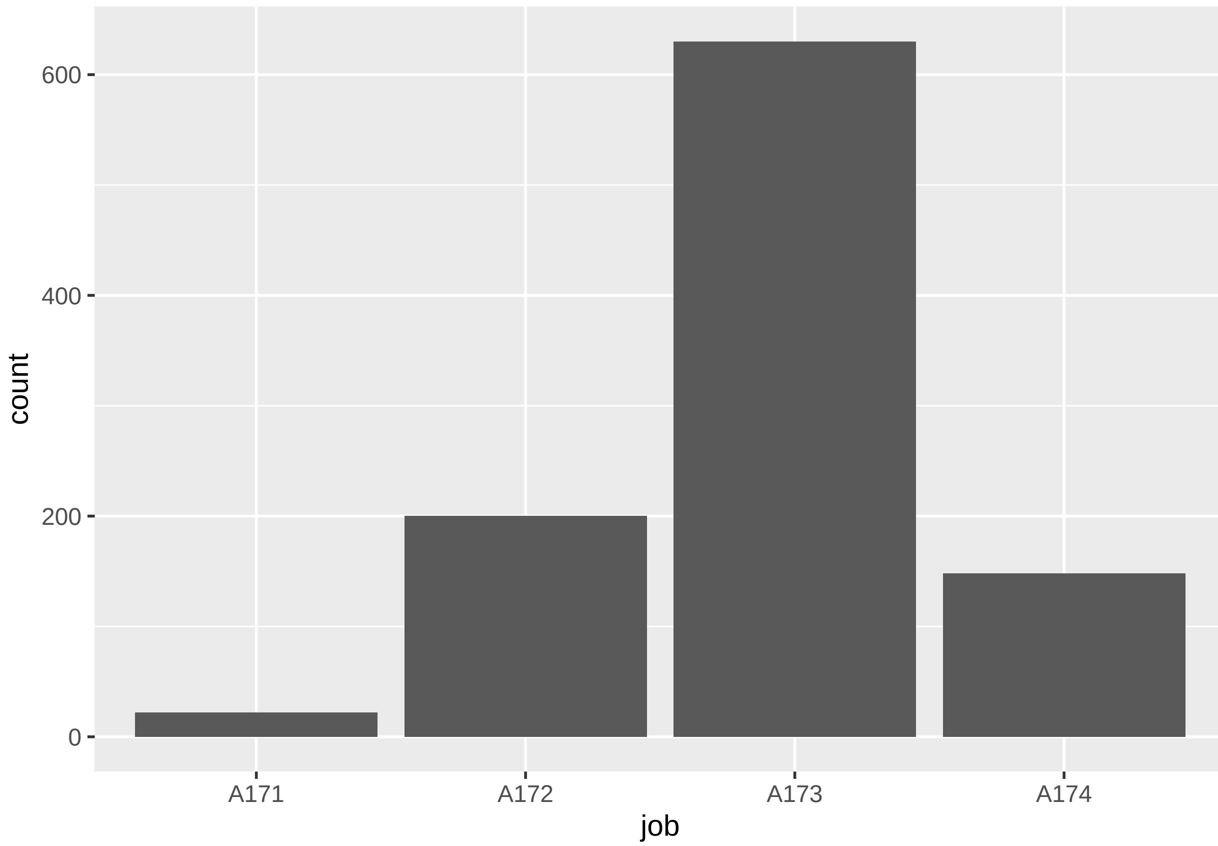

```
a+geom_bar(aes(housing))
```



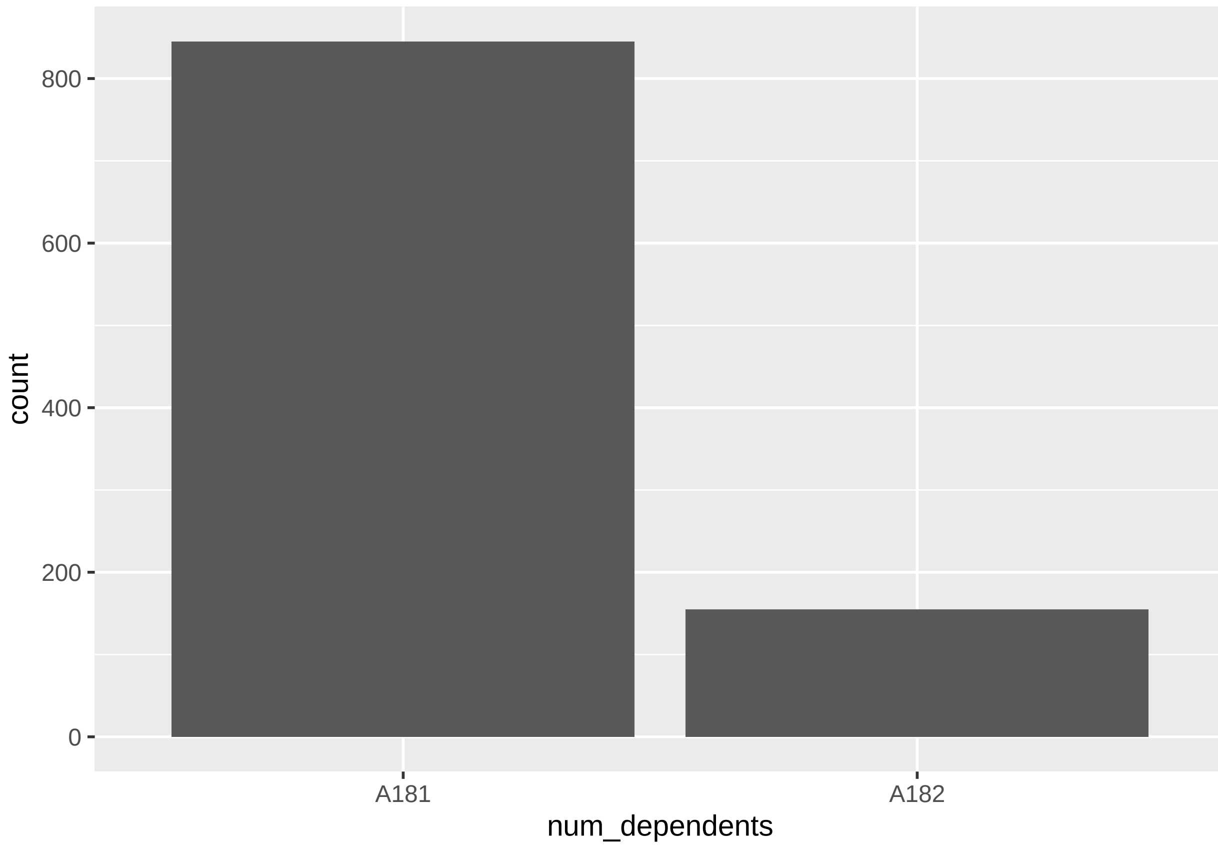

```r
a+geom_bar(aes(existing_credits))
```



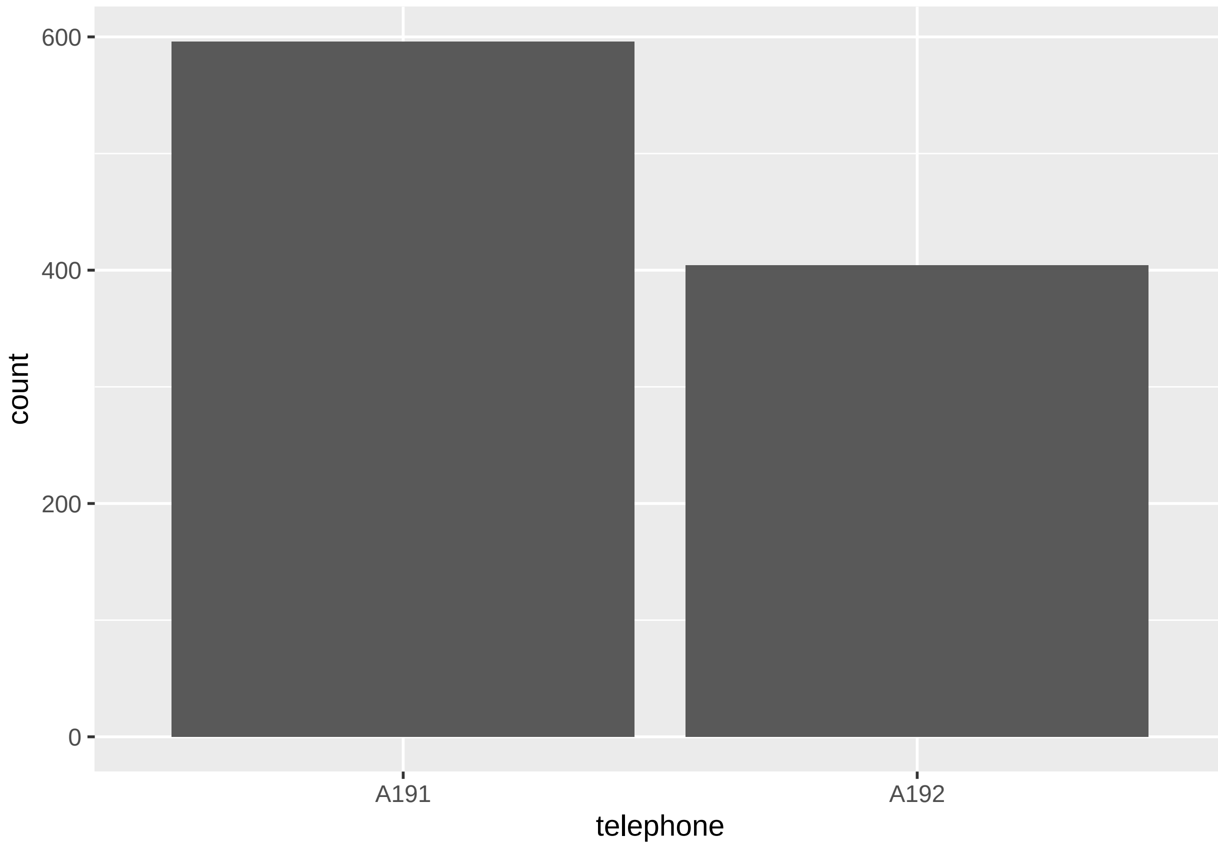

```
a+geom_bar(aes(job))
```



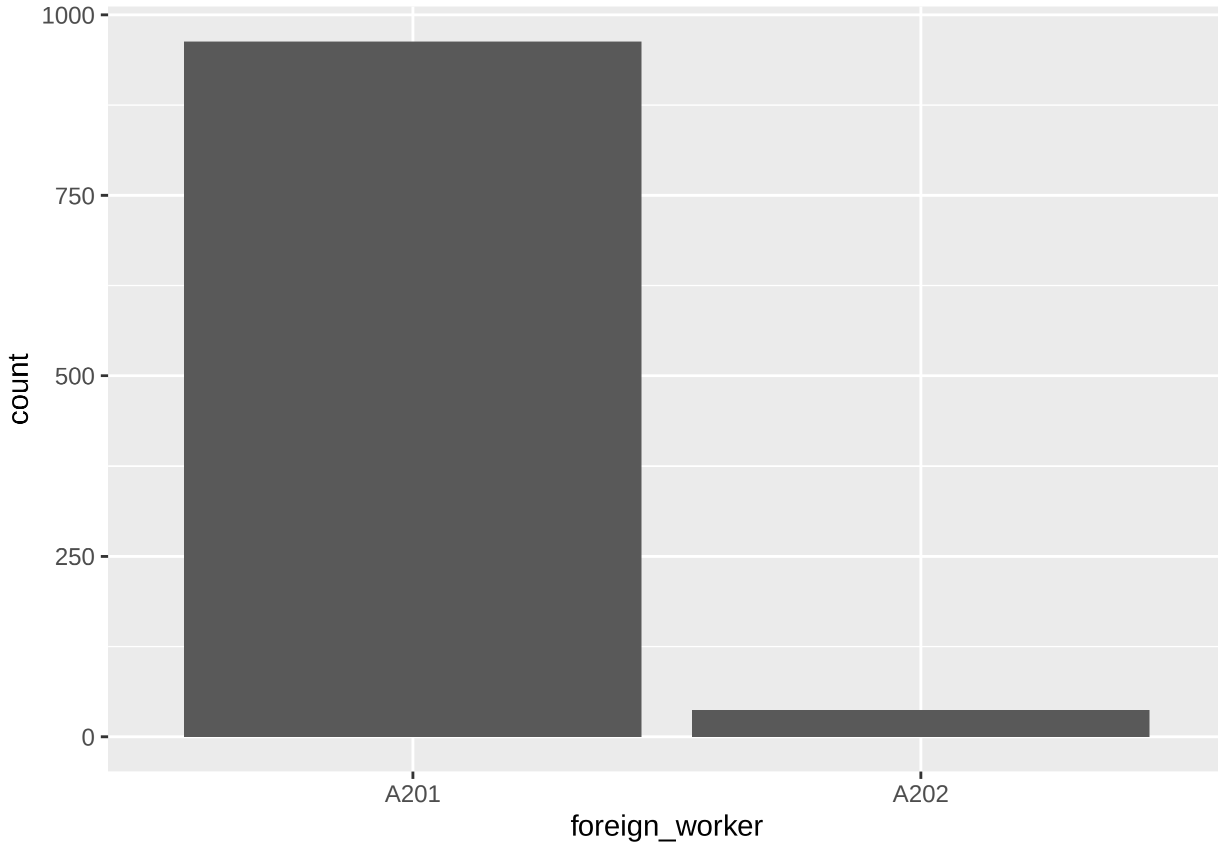

```
a+geom_bar(aes(num_dependents))
```



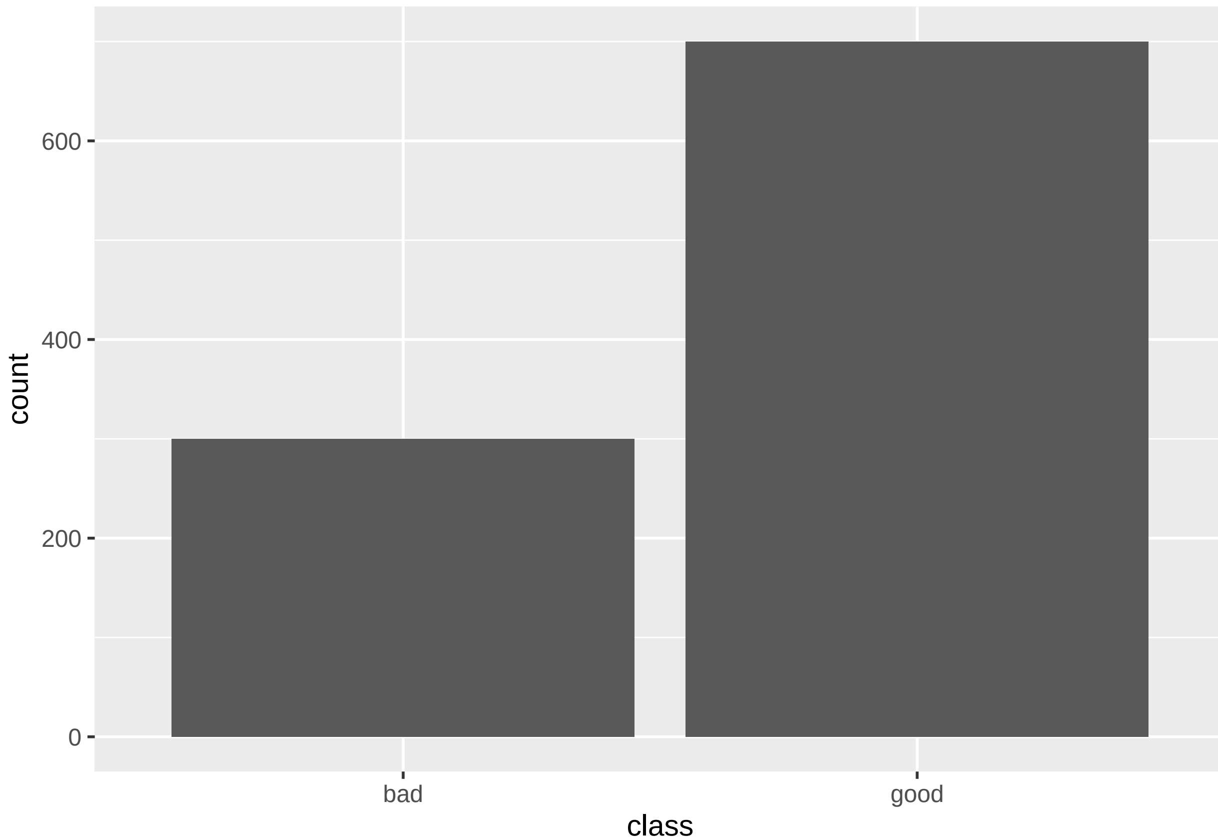

```
a+geom_bar(aes(telephone))
```




```
a+geom_bar(aes(foreign_worker))
```



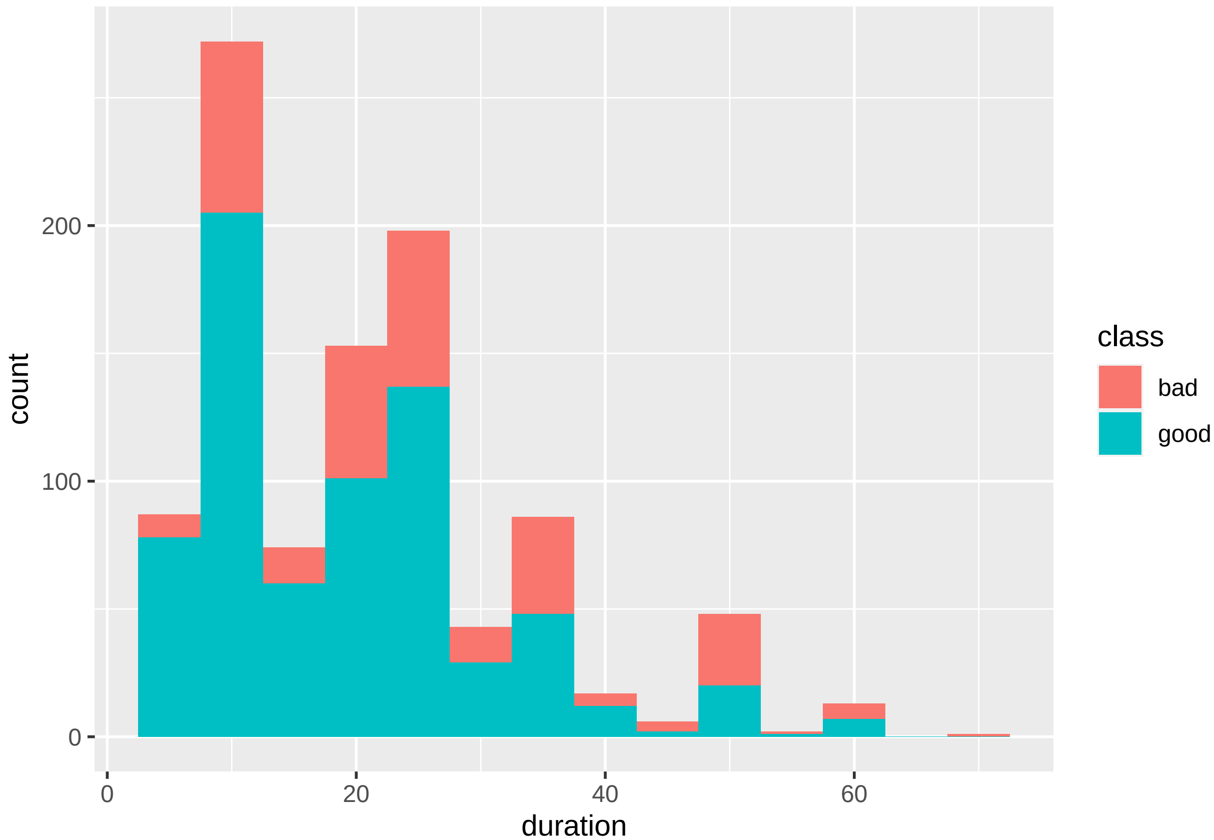

```
a+geom_bar(aes(class))
```




## Question 2

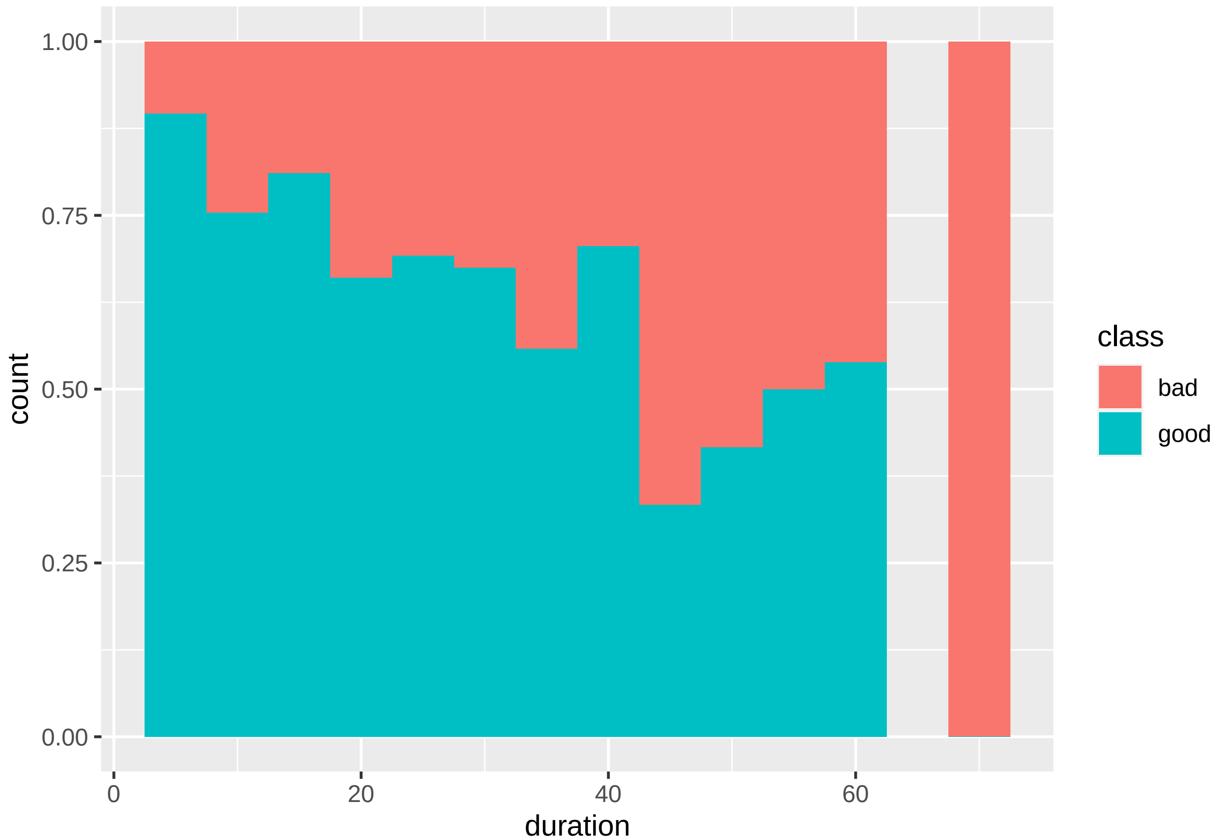
Duration

```
a+geom_histogram(aes(x=duration, fill=class), position='stack', binwidth = 5)
```

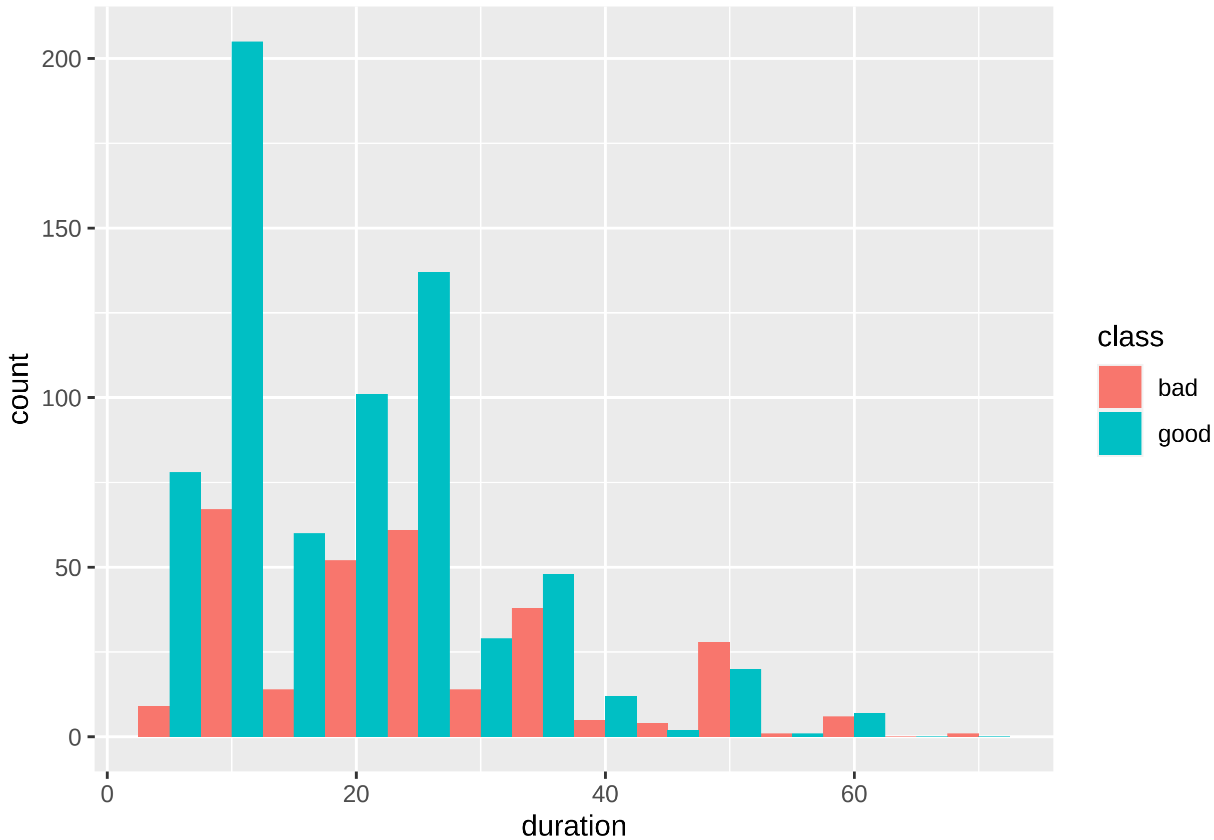



```
a+geom_histogram(aes(x=duration, fill=class), position='fill', binwidth = 5)
```

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




```
a+geom_histogram(aes(x=duration, fill=class), position='dodge', binwidth = 5)
```



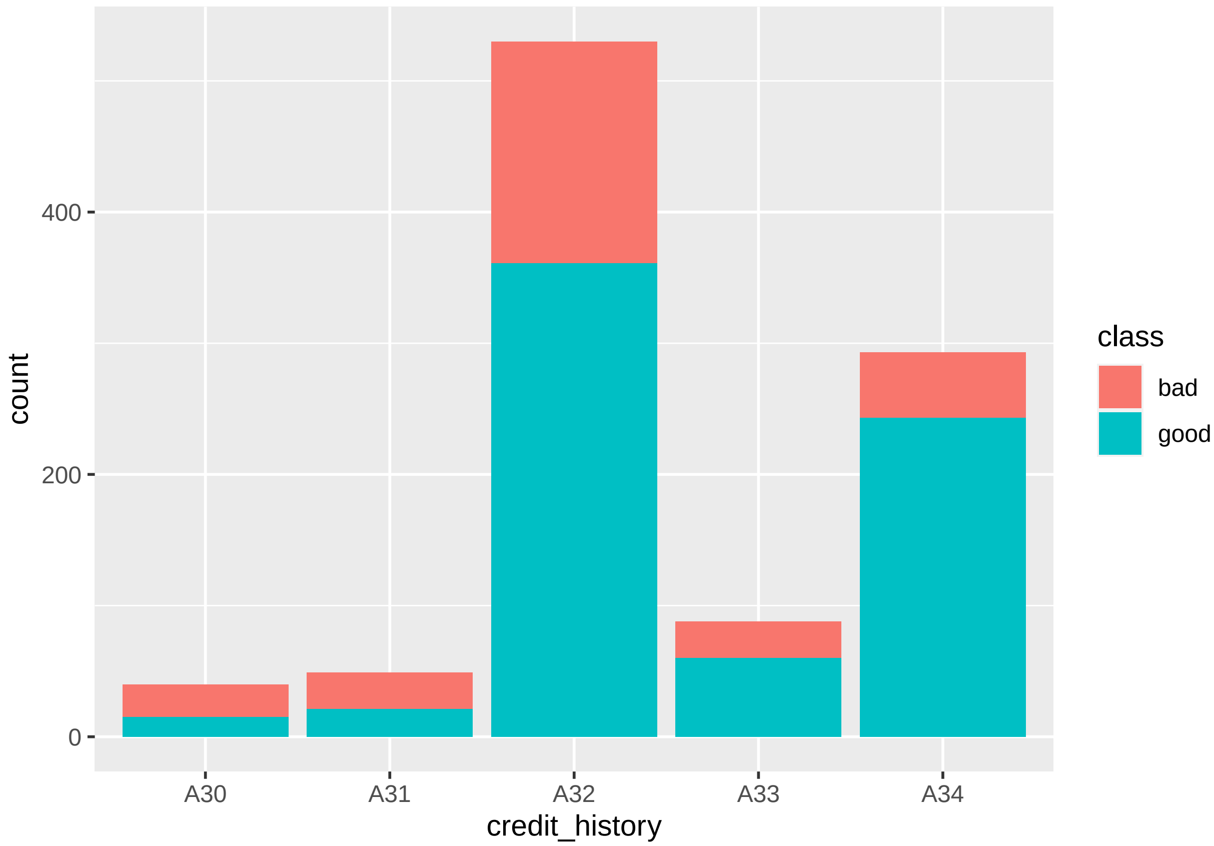

```
table(credit$class,credit$duration)
```

```
##
##         4   5   6   7   8   9  10  11  12  13  14  15  16  18  20  21  22  24
##   bad   0   0   9   0   1  14   3   0  49   0   1  12   1  42   1   9   0  56
##   good  6   1  66   5   6  35  25   9 130   4   3  52   1  71   7  21   2 128
##
##        26  27  28  30  33  36  39  40  42  45  47  48  54  60  72
##   bad   0   5   1  13   1  37   1   1   3   4   0  28   1   6   1
##   good  1   8   2  27   2  46   4   0   8   1   1  20   1   7   0
```

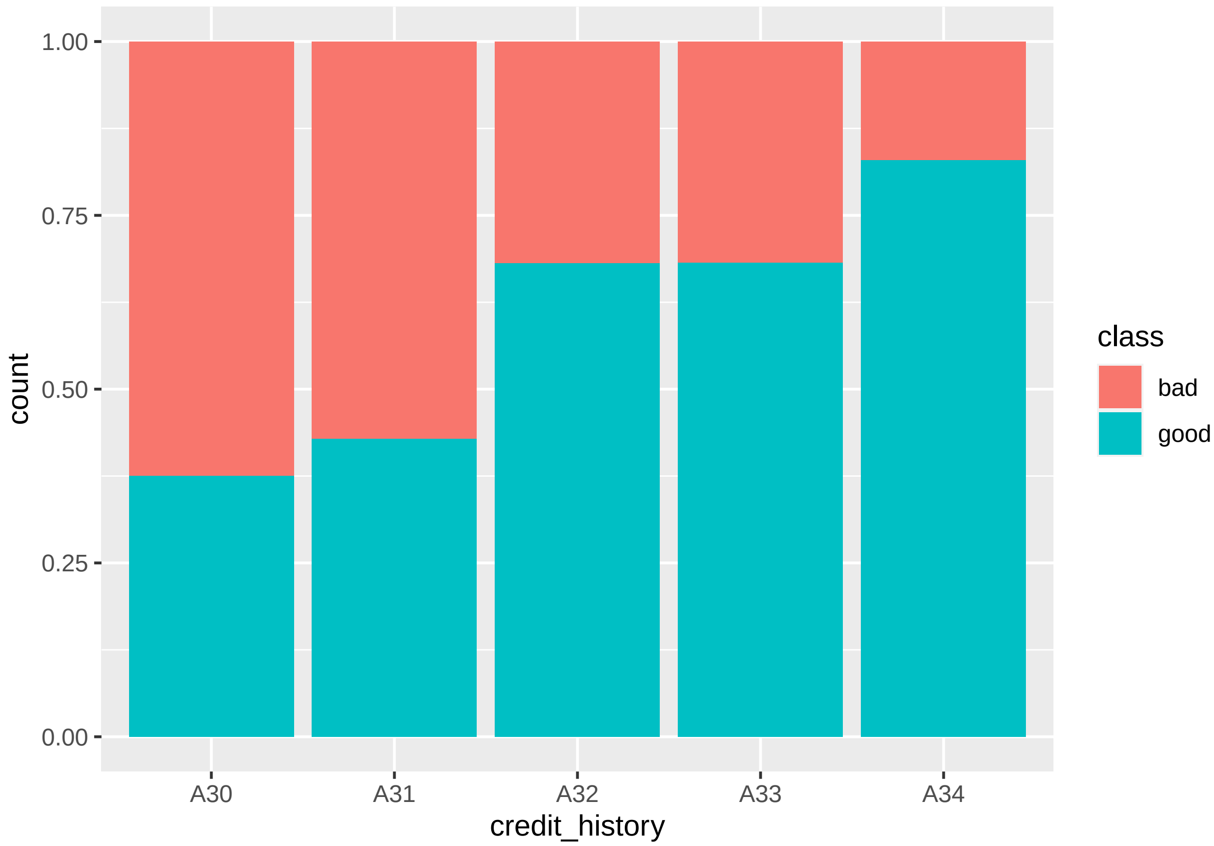
Until the duration of 40, the proportion of 'good' class is larger than 50 percent. After the duration of 40, the proportion of class 'good' decreased, but we have to notice that the number of sample is much smaller comparing with duration of 0~40 ( except duration around 50). Duration between 5~15 has the largest number of sample and also class 'good'.

credit_history

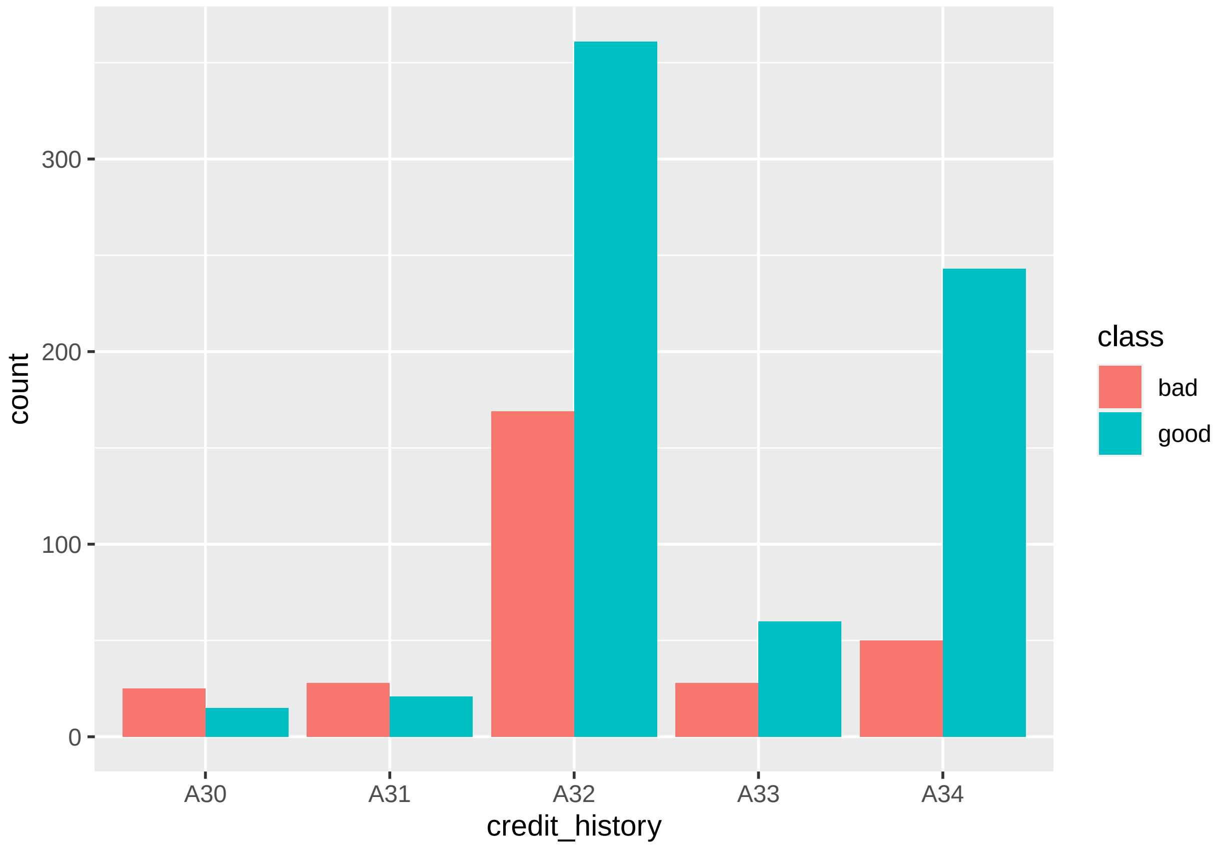
```
a+geom_bar(aes(x=credit_history, fill=class), position='stack')
```




```
a+geom_bar(aes(x=credit_history, fill=class), position='fill')
```




```r
a+geom_bar(aes(x=credit_history, fill=class), position='dodge')
```



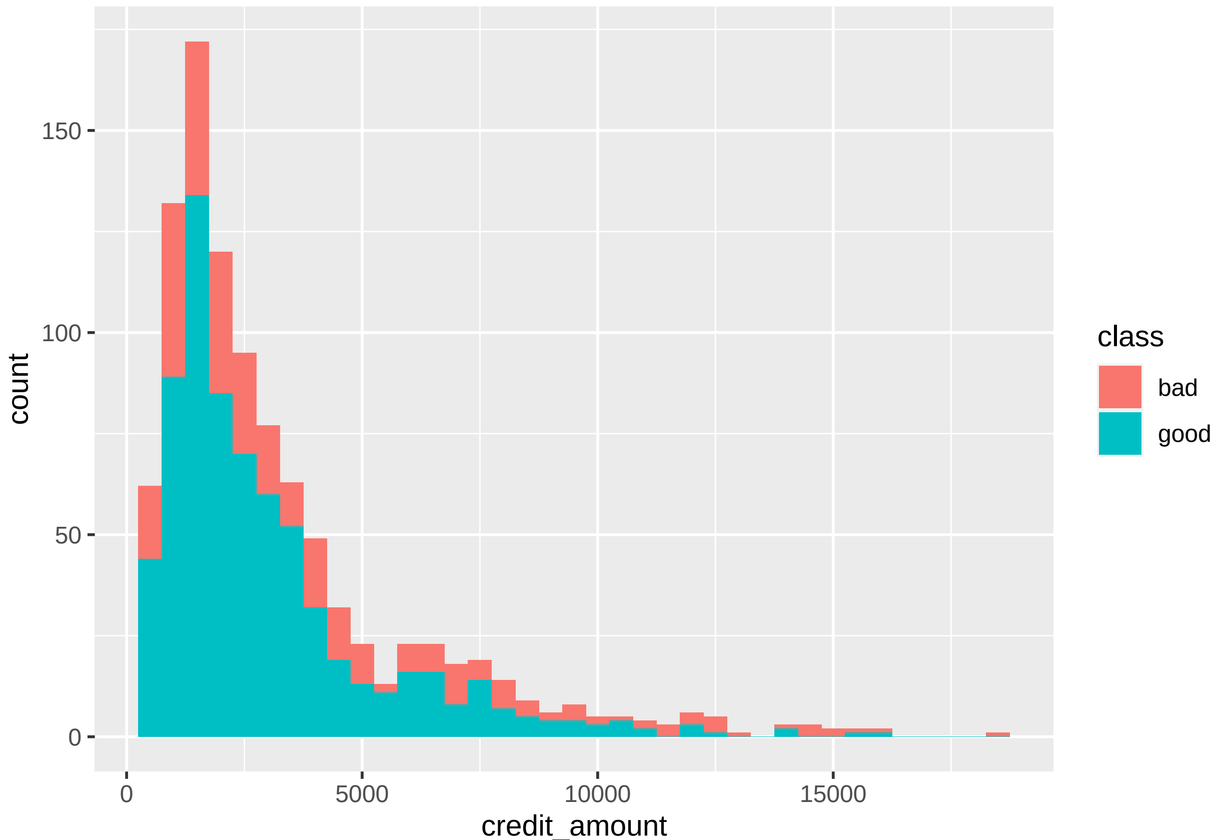

```
table(credit$credit_history)
```

```
##
## A30 A31 A32 A33 A34
##  40  49 530  88 293
```

The proportion of class 'good' monotonly increased from A30 to A34. It is reasonable because people who paid back duly normally have good credit and good class. More than 55 percent of consumer who delayed in paying off or having critical account has 'bad' class.

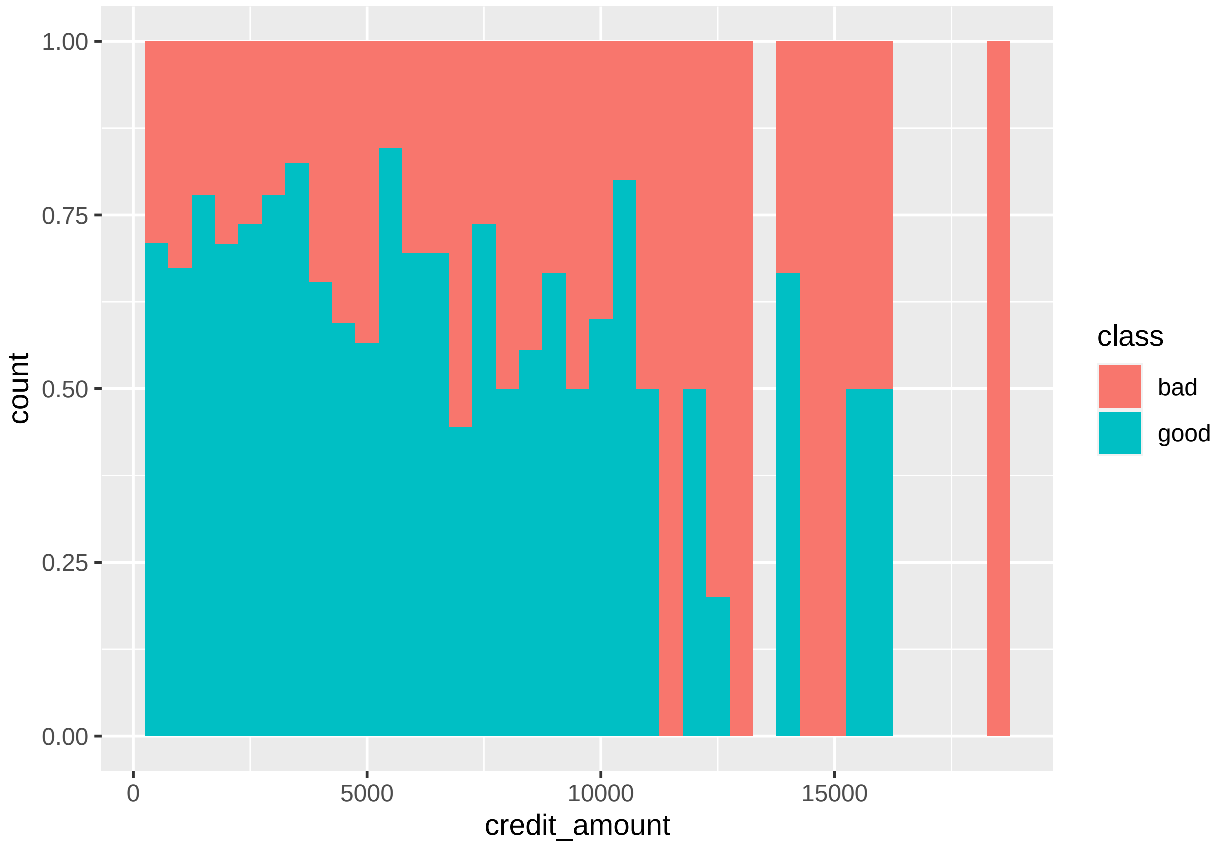
credit_amount

```
a+geom_histogram(aes(x=credit_amount, fill=class), position='stack', binwidth = 500)
```

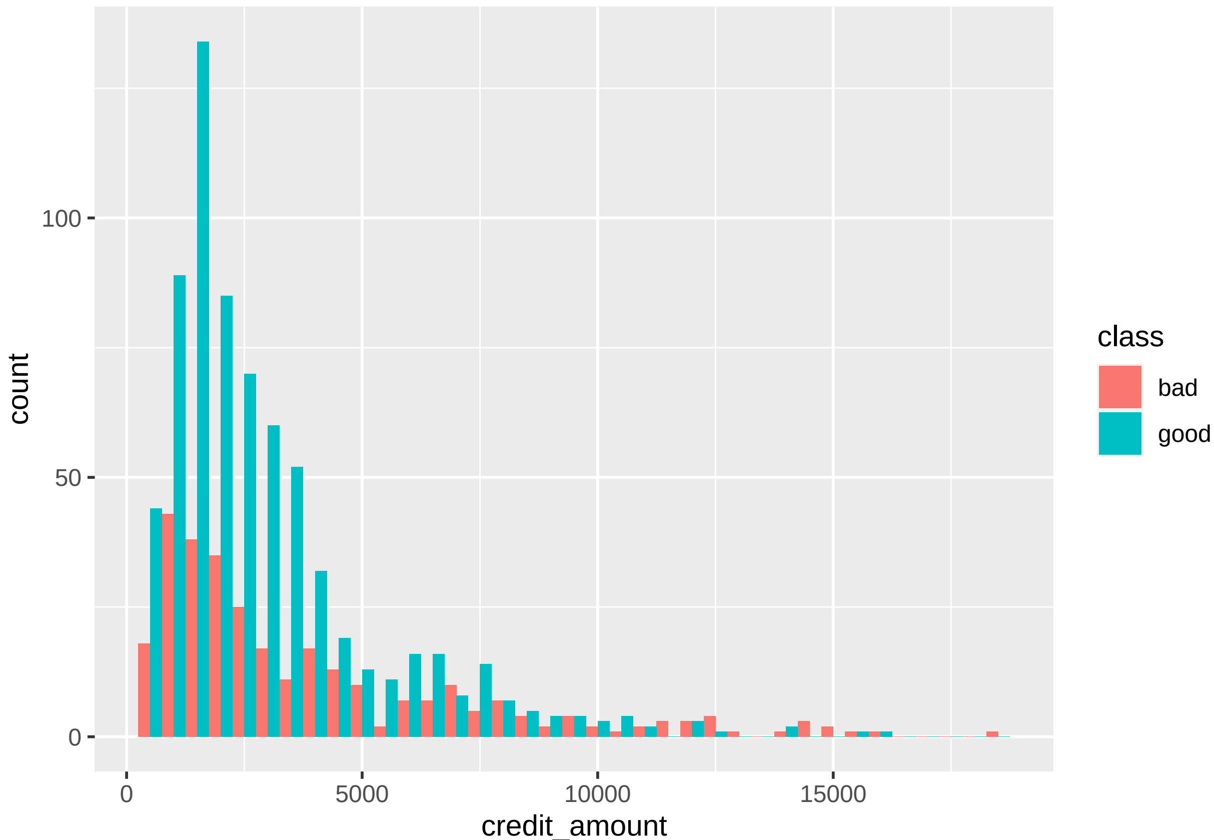



```
a+geom_histogram(aes(x=credit_amount, fill=class), position='fill', binwidth = 500)
```

```
## Warning: Removed 10 rows containing missing values (geom_bar).
```



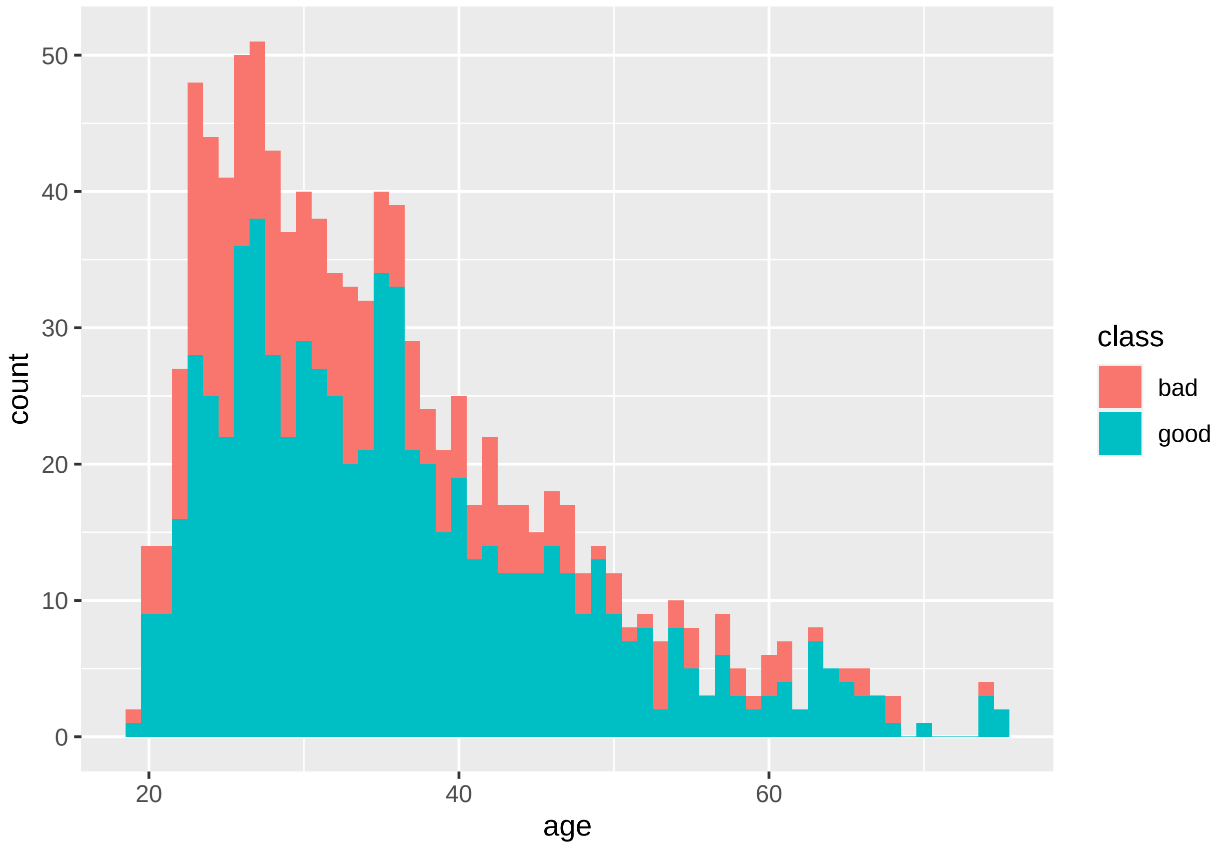

```
a+geom_histogram(aes(x=credit_amount, fill=class), position='dodge', binwidth = 500)
```




By looking at the normalized histogram of credit_amount, we found the tendency for people with low credit amount are more likely to have good class.

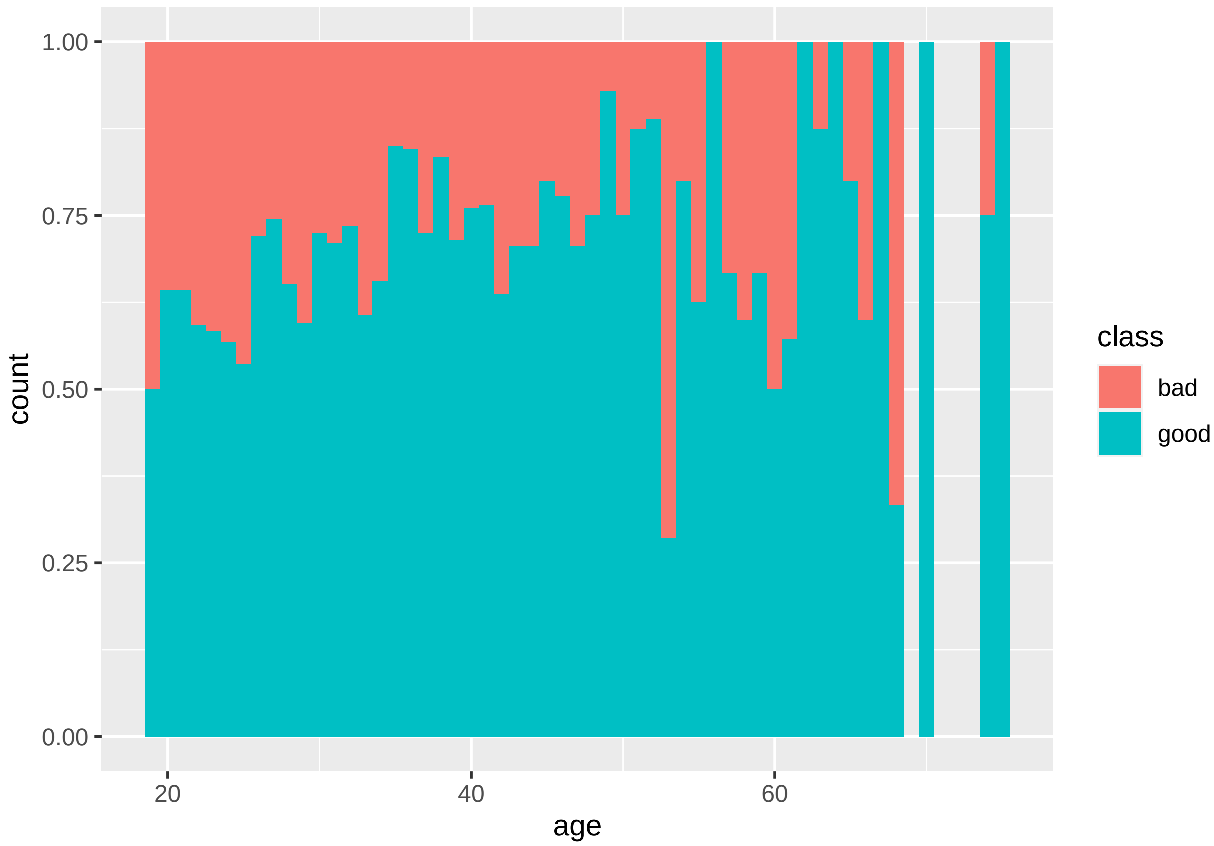
Age

```
a+geom_histogram(aes(x=age, fill=class), position='stack', binwidth = 1)
```

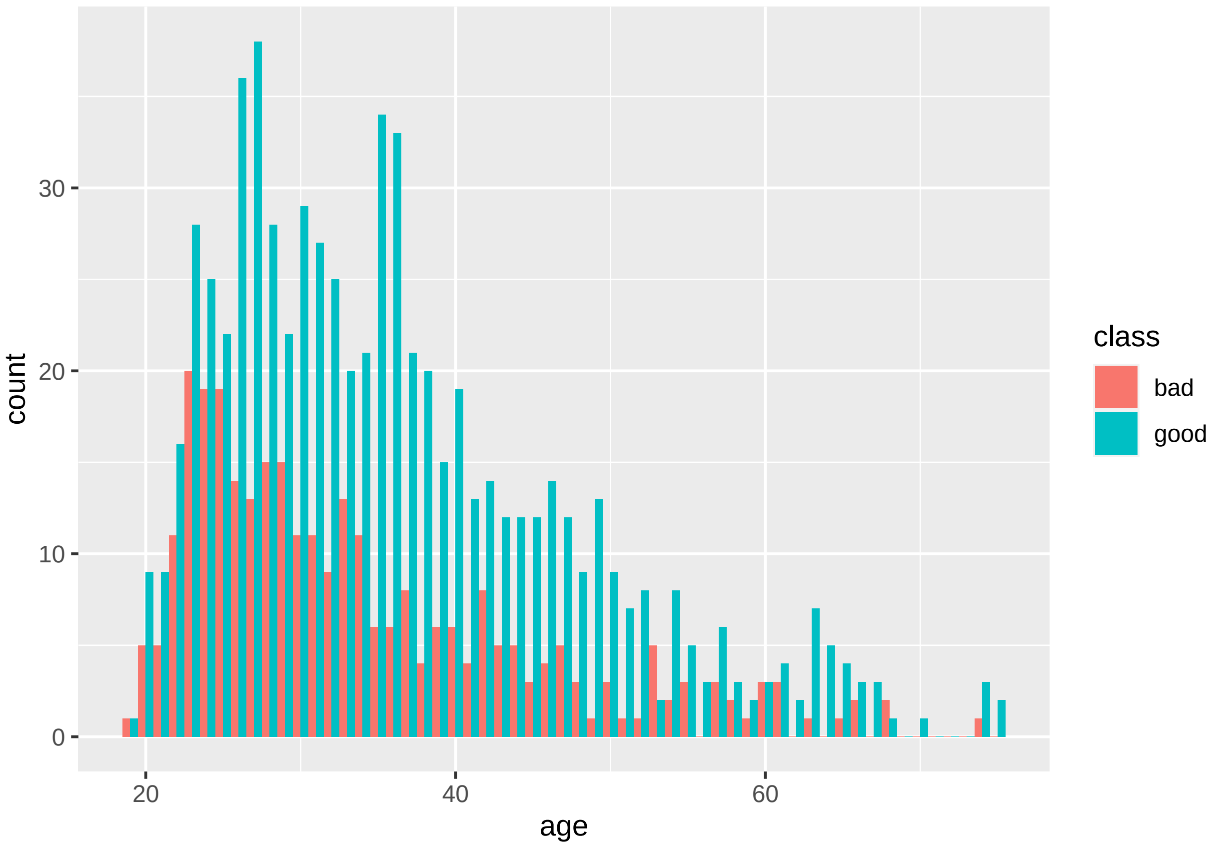



```
a+geom_histogram(aes(x=age, fill=class), position='fill', binwidth = 1)
```

```
## Warning: Removed 8 rows containing missing values (geom_bar).
```



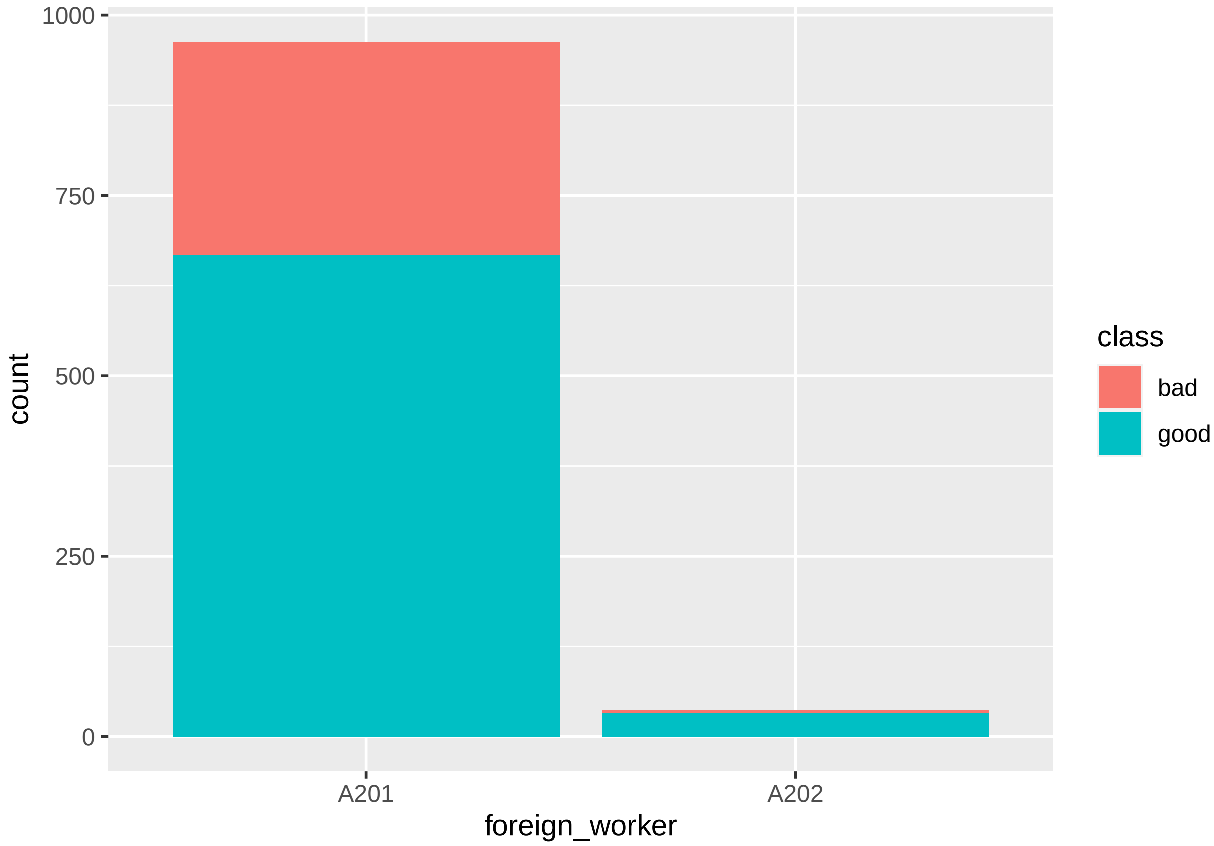

```
a+geom_histogram(aes(x=age, fill=class), position='dodge', binwidth = 1)
```



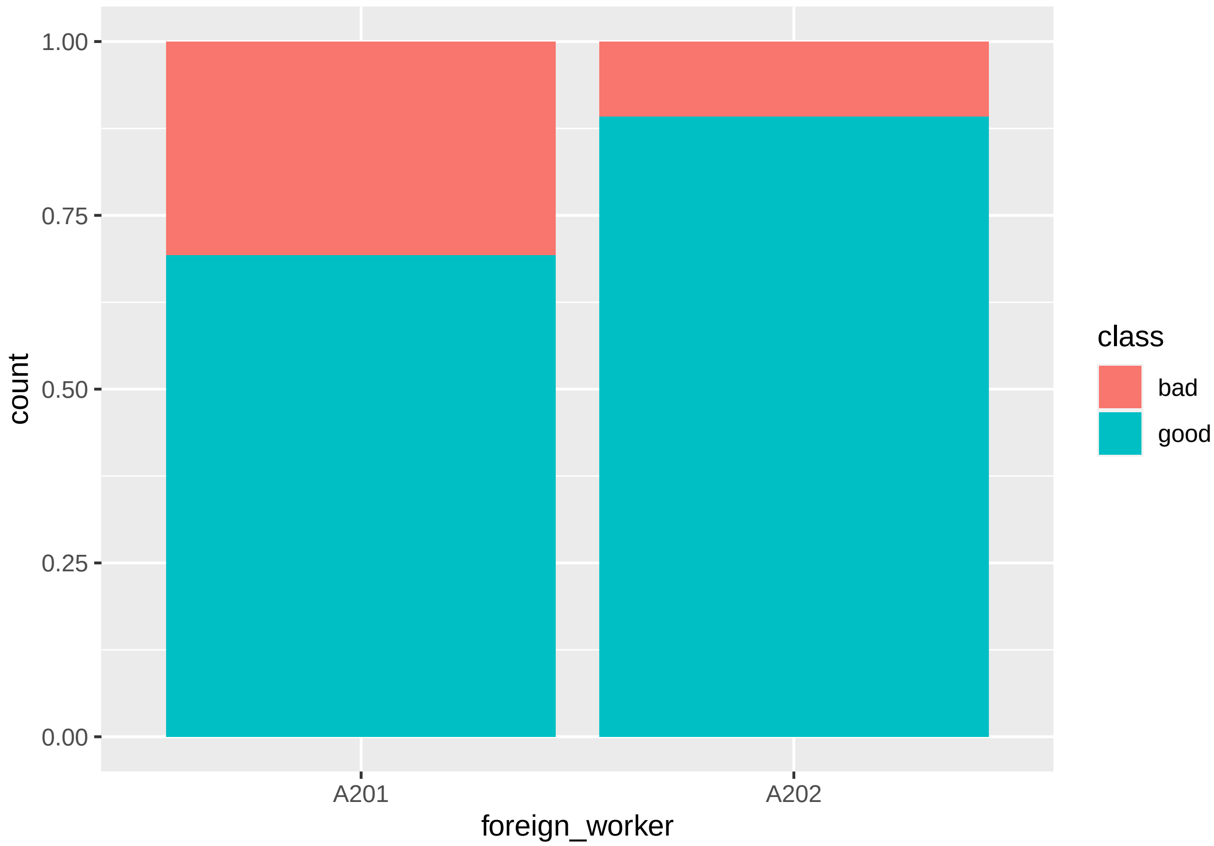

It was difficult to discern the proportion of class(good/bad ) varies across age by using the histogram of age with class overlay. But by observing normalized histogram, we were able to find out the tendency for people with higher age have higher proportion of good class

Foreign worker

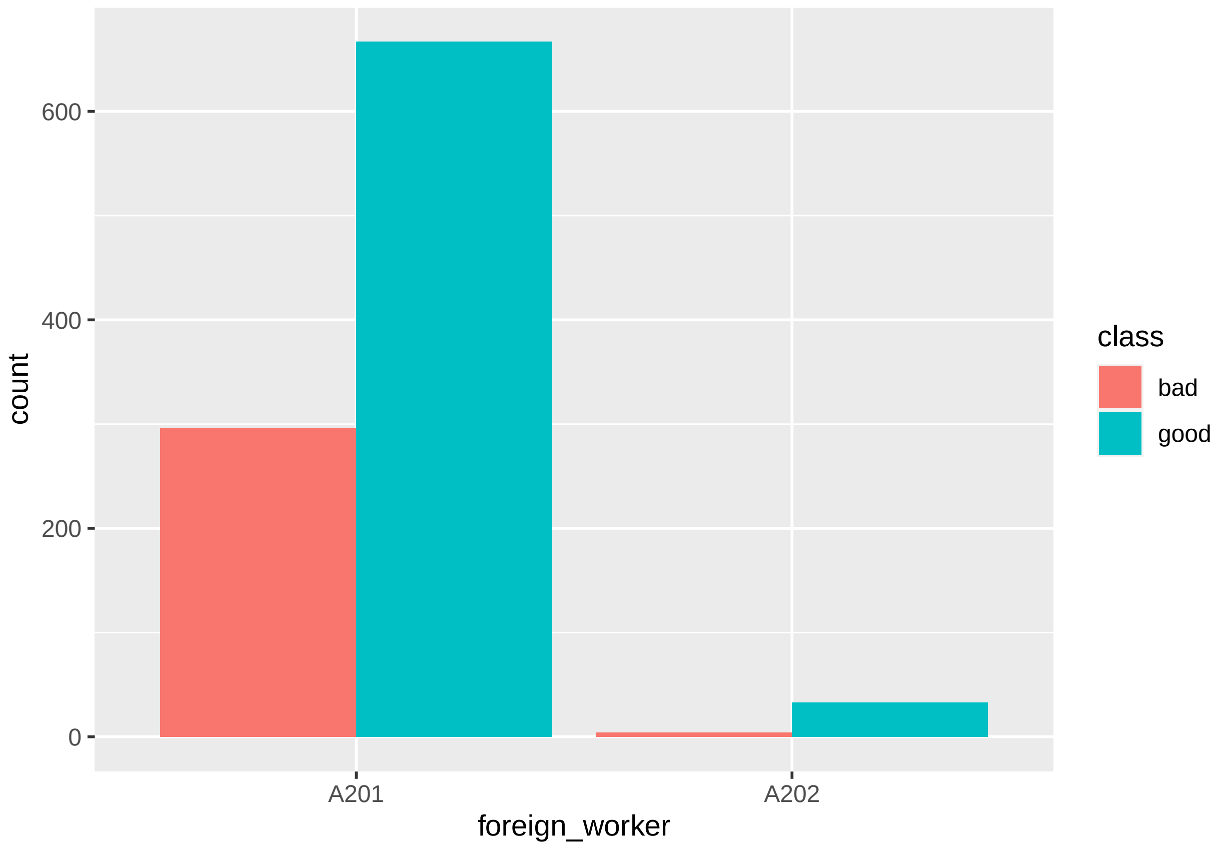
```
a+geom_bar(aes(x=foreign_worker, fill=class), position='stack')
```




```
a+geom_bar(aes(x=foreign_worker, fill=class), position='fill')
```




```
a+geom_bar(aes(x=foreign_worker, fill=class), position='dodge')
```




```
table(credit$foreign_worker)
```

```
##
## A201 A202
##  963   37
```

We were able to notice the tendency for people with A202 to have good class.

## Question 3

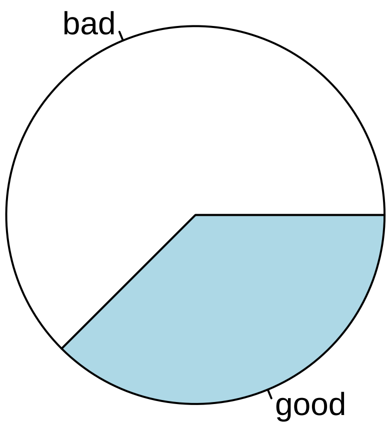
Contingency table and pie chart for credit history

```
mytable_class_credit_history=table(credit$class, credit$credit_history)
mytable_class_credit_history
```

```
##
##        A30 A31 A32 A33 A34
##   bad   25  28 169  28  50
##   good  15  21 361  60 243
```

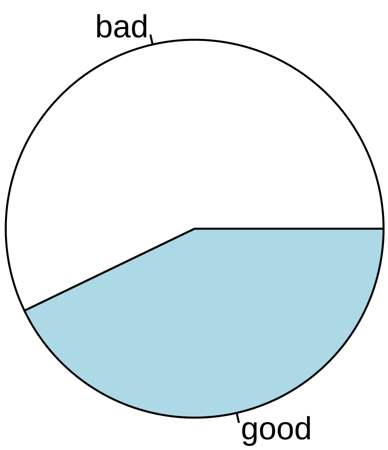
```
pie(table(filter(credit, credit_history=='A30')$class), main='A30:delay in paying off in the past')
```

**A30:delay in paying off in the past**

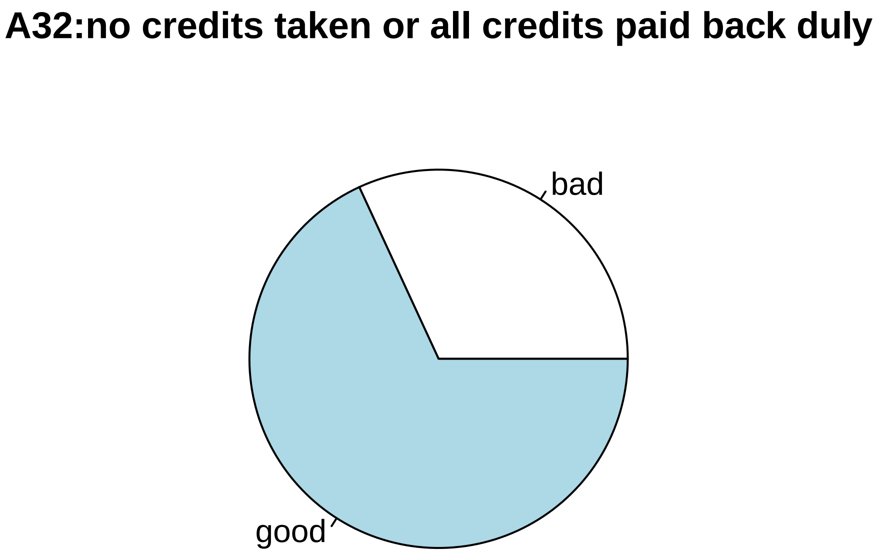



```r
pie(table(filter(credit, credit_history=='A31')$class), main='A31:critical account')
```

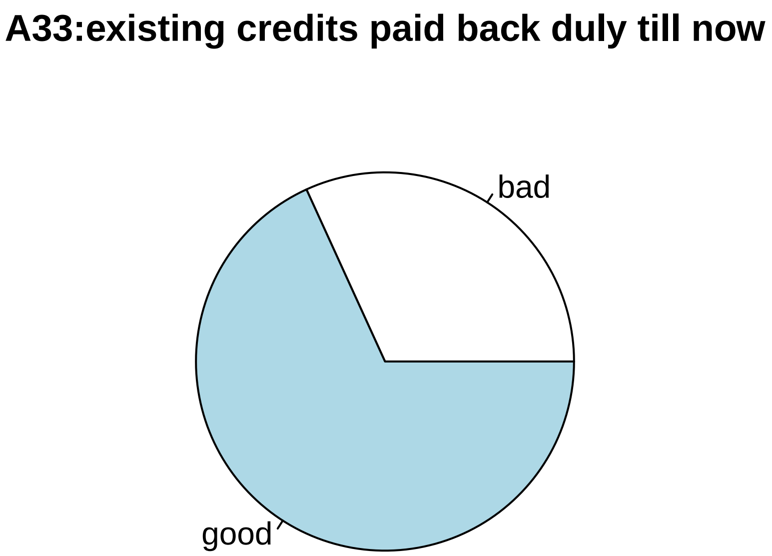
**A31:critical account**




```r
pie(table(filter(credit, credit_history=='A32')$class), main='A32:no credits taken or all credits paid
```

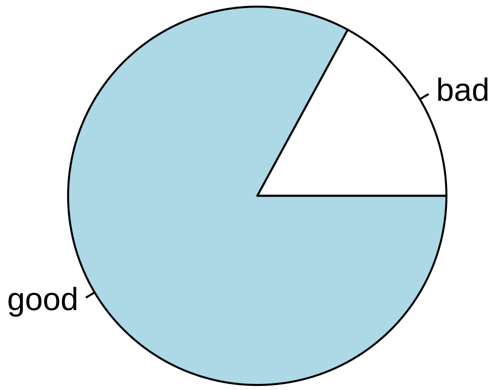



```
pie(table(filter(credit, credit_history=='A33')$class), main='A33:existing credits paid back duly till
```




```
pie(table(filter(credit, credit_history=='A34')$class), main='A34:all credits at this bank paid back du
```

## A34:all credits at this bank paid back duly

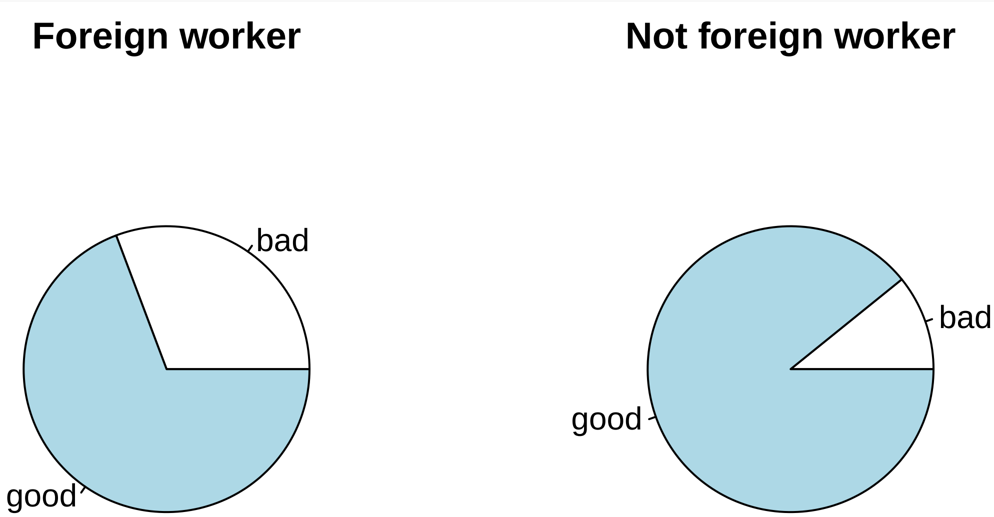



Contingency table and pie chart for foreign worker

```r
mytable_class_foregin_worker=table(credit$class, credit$foreign_worker)
mytable_class_foregin_worker
```

```
##
##         A201 A202
##   bad    296    4
##   good   667   33
```

```r
par(mfrow=c(1,2))
pie(table(filter(credit, foreign_worker=='A201')$class), main='Foreign worker')
pie(table(filter(credit, foreign_worker=='A202')$class), main='Not foreign worker')
```




Conditional probability of a customer being a good payer, given each of predictors individually

```r
prop.table(mytable_class_credit_history,2)
```

```
##
##              A30       A31       A32       A33       A34
##   bad  0.6250000 0.5714286 0.3188679 0.3181818 0.1706485
##   good 0.3750000 0.4285714 0.6811321 0.6818182 0.8293515
```

Probability of good class given that credit history is A30(delay in paying off in the past): 0.3750000

Probability of good class given that credit history is A31(critical account) : 0.4285714

Probability of good class given that credit history is A32(no credits taken or all credits paid back duly) : 0.6811321

Probability of good class given that credit history is A33(existing credits paid back duly till now) : 0.6818182

Probability of good class given that credit history is A34(all credits at this bank paid back duly) : 0.8293515

```
prop.table(mytable_class_foregin_worker,2)
```

```
##
##              A201      A202
##   bad  0.3073728 0.1081081
##   good 0.6926272 0.8918919
```

Probability of good class given that consumer is foreign worker :0.6926272

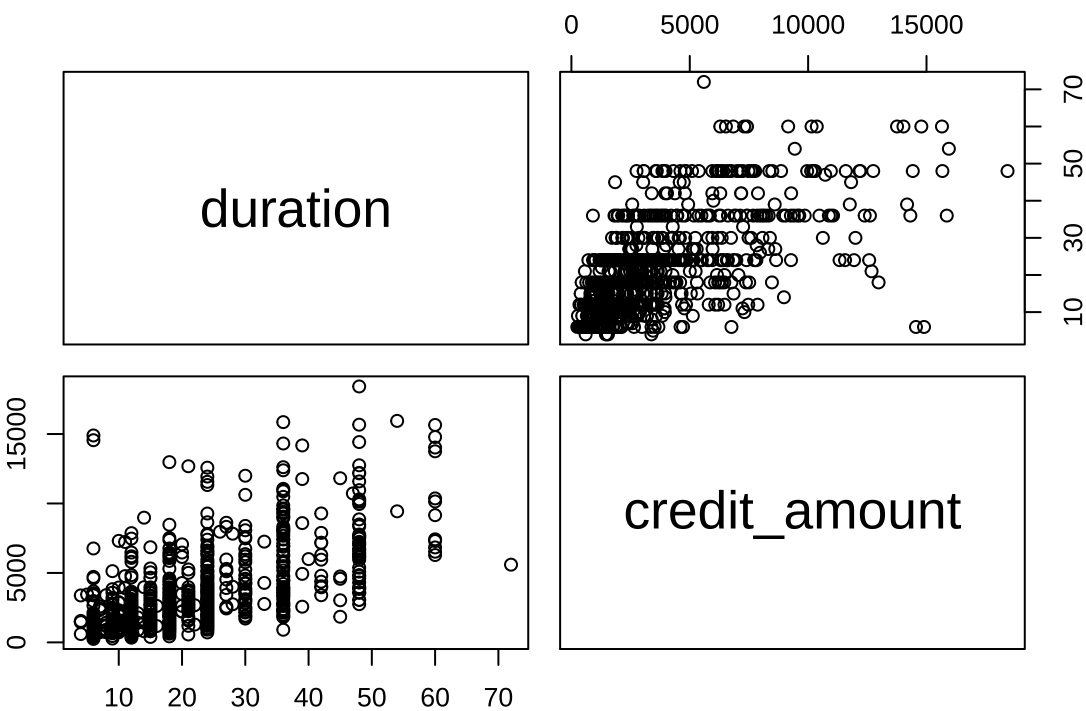
Probability of good class given that consumer is not foreign worker :0.8918919

## Question 4

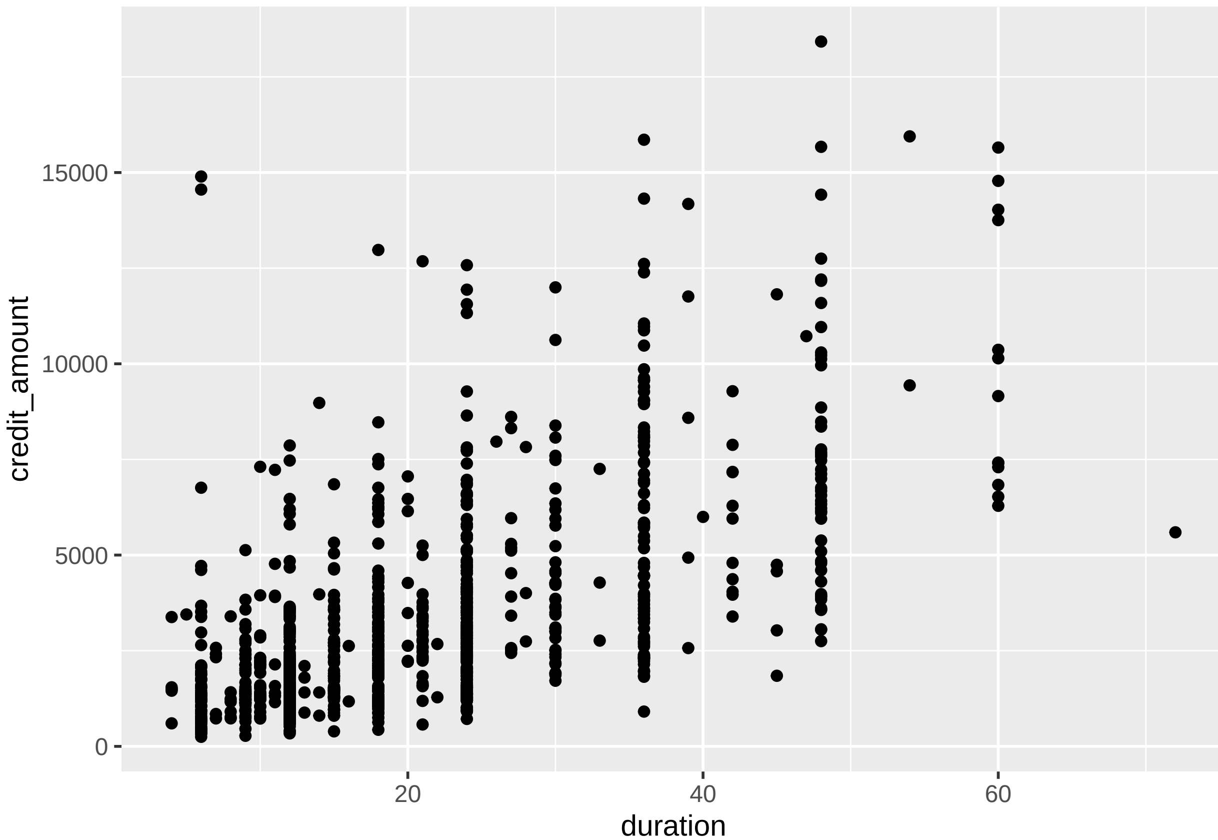
```
cor(select(credit, duration, credit_amount))
```

```
##                duration credit_amount
## duration      1.0000000     0.6249842
## credit_amount 0.6249842     1.0000000
```

```
pairs(~duration+credit_amount, data=credit)
```




```
a + geom_point(aes(x=duration, y=credit_amount))
```



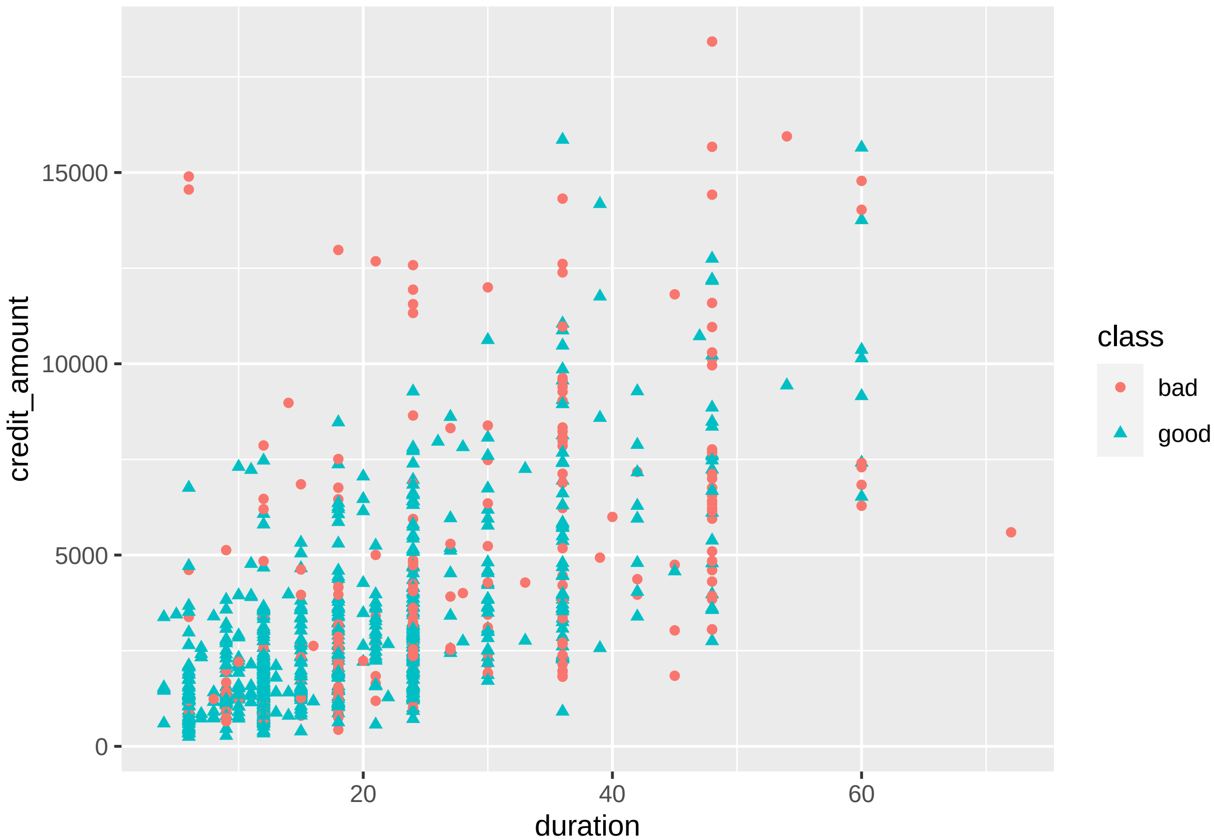

### Question a

It is certainly true that if the duration increases, the range of the credit_amount and also frequency of credit amount higher than 10,000 increases. Also, correlation between duration and credit_amount is 0.6249842 which is close to 1 rather than 0. Therefore, we can conclude that the two variables are correlated but not too strongly.

We will not drop duration or credit_amount. Although using correlated variables can wrongly emphasize more data inputs and creates unreliable results, but duration or credit amount is not highly strongly correlated. Moreover, there is no sufficient reasaon to completely exclude the variable from the model and hypothesis test can be used to determine other factors.

### Question b

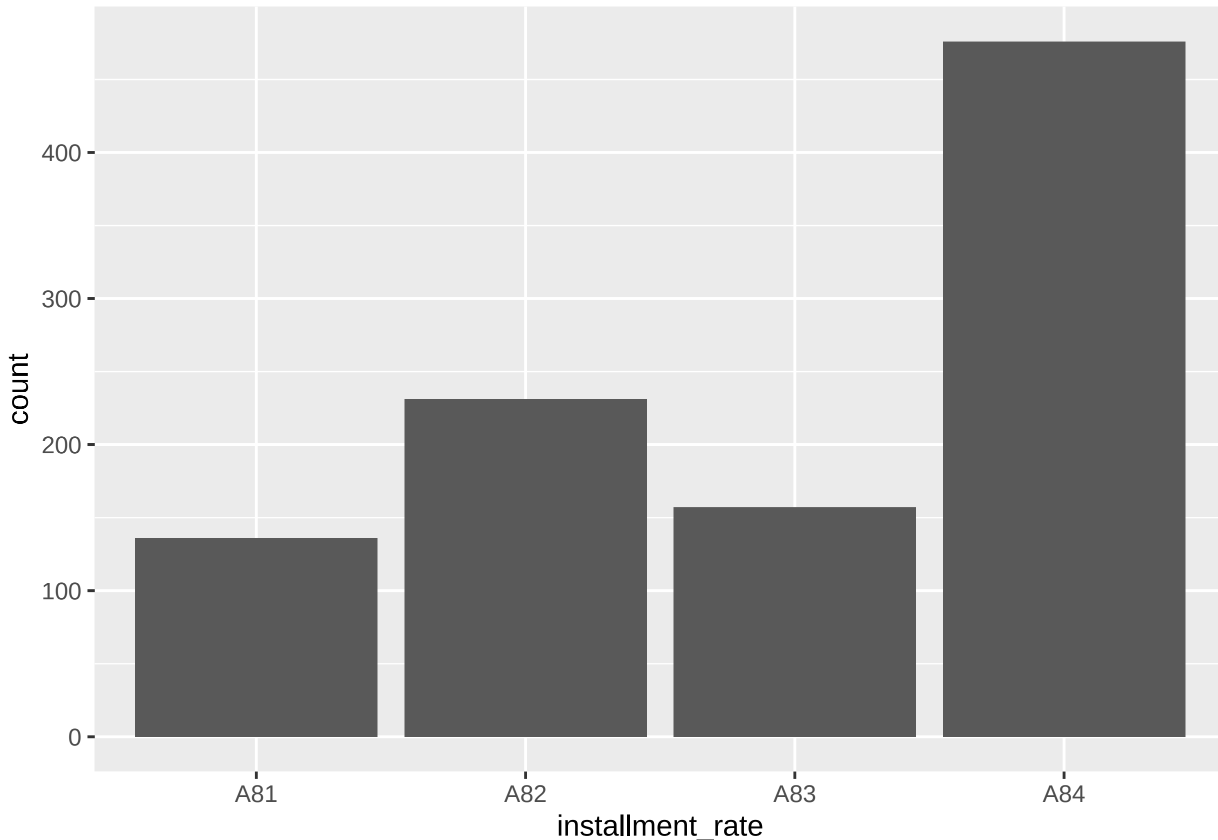
```
a + geom_point(aes(x=duration,y=credit_amount, shape = class,
color = class))
```




-There exists multivariate relationship among two variables and class. People with both low credit amount and duration tend to have higher proportion of good class.

## Question 5

```
a+geom_bar(aes(installment_rate))
```




After a small increase in the number of people in A82(installment rate between 25 and 35) compared with the sub group A81, the number of people decreased in A83(installment rate between 20 and 25) . We first guessed as installment rate decreases, the number of people will decrease. So we predicted that the number of people in last subgroup A84 should decrease. But the box plot gave us surprising result. The number of people dramatically increased as the installment rate was less than 20 . Therefore, we chose this variable, installment_rate which we want to further investigate in. We were curious about the last sub group "Why did the number of people sharply increased when installment rate is less than 20?"

## Question 6

The interesting fact that we found in the dataset is about personal stauts. If we first look at the data description, personal status is divided by 5 standards. A91:Male:divorced, A92:female:divorced or married, A93:male:single:, A94:male:married, A95:female: single. However, it is not that logical. For male, there are three indicator which are 1.divorced, 2.married or 3.single. For female, there are only two indicator which are 1.divorced/married and 2.single. There is no reason to combine divorced female and married female in same category. Divorced female and married female can have different class and they can influence to other variables significantly.

Moreover, if we look at the table,

```r
table(credit$personal_status)
```

```
##
## A91 A92 A93 A94
##  50 310 548  92
```

A95, which is female:single is missing. It is impossible that there is no female:single consumer in German credit. Therefore, we can make two assumption. First, female single is combined with A92: female: divorced or married. Secondly, someone made mistake while collecting data and female:single spread out to all indicators:A91~A94.