Credit risk EDA

Anupriya Kushwanshi

May 27, 2019

Insights: 1 is good and 2 is bad.

. 70% customers are good and 30% are bad.
. 50% of "good" customers do not have checking accounts and 45% defaulters have less than 0 Deutsche Mark in their checking accounts.
. 51% "good" customers paid existing credits till now and nearly 30% have critical account or other credits existing (not at this bank). More than 56% defaulters paid existing credits till now.
. For good customers, most common purpose for taking the loan is radio/television whereas for defaulters it's car.
. More than 50% good customers and nearly 70% bad customers have less than 100 Deutsche Mark in their accounts.
. Employment is almost well distributed among the customers with 1-4 year of work experience having a better share. Very few(~7.6%) are unemployed.
. 57% good customers and 49% bad customers single males.
. 91% good and bad customers have no other debtors.
. 31% good customers have properties in the form of real state and 33% in car. The proportion is almost same among the bad customers.
. 84% of good customers and 75% of bad customers have no other installment plans. . 75% good customers and 62% bad customers own a house.
. 64% good customers and 62% bad customers are skilled employees or officials.
. 95% good customers and 99% bad customers are foreign employees.

```
german_credit = read.table("http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/ger
man/german.data")
colnames(german_credit)=c("chk_acct","duration","credit_his","purpose","amount","saving_acct","p
resent_emp","installment_rate","sex","other_debtor","present_resid","property","age","other_inst
all","housing","n_credits","job","n_people","telephone","foreign","response")
head(german_credit)
```

```
##    chk_acct duration credit_his purpose amount saving_acct present_emp
## 1      A11        6        A34     A43   1169         A65         A75
## 2      A12       48        A32     A43   5951         A61         A73
## 3      A14       12        A34     A46   2096         A61         A74
## 4      A11       42        A32     A42   7882         A61         A74
## 5      A11       24        A33     A40   4870         A61         A73
## 6      A14       36        A32     A46   9055         A65         A73
##    installment_rate sex other_debtor present_resid property age
## 1                 4 A93         A101             4     A121  67
## 2                 2 A92         A101             2     A121  22
## 3                 2 A93         A101             3     A121  49
## 4                 2 A93         A103             4     A122  45
## 5                 3 A93         A101             4     A124  53
## 6                 2 A93         A101             4     A124  35
##    other_install housing n_credits  job n_people telephone foreign response
## 1           A143    A152         2 A173        1      A192    A201        1
## 2           A143    A152         1 A173        1      A191    A201        2
## 3           A143    A152         1 A172        2      A191    A201        1
## 4           A143    A153         1 A173        2      A191    A201        1
## 5           A143    A153         2 A173        2      A191    A201        2
## 6           A143    A153         1 A172        2      A192    A201        1
```

```
str(german_credit)
```

```
## 'data.frame':    1000 obs. of  21 variables:
##  $ chk_acct        : Factor w/ 4 levels "A11","A12","A13",..: 1 2 4 1 1 4 4 2 4 2 ...
##  $ duration        : int  6 48 12 42 24 36 24 36 12 30 ...
##  $ credit_his      : Factor w/ 5 levels "A30","A31","A32",..: 5 3 5 3 4 3 3 3 3 5 ...
##  $ purpose         : Factor w/ 10 levels "A40","A41","A410",..: 5 5 8 4 1 8 4 2 5 1 ...
##  $ amount          : int  1169 5951 2096 7882 4870 9055 2835 6948 3059 5234 ...
##  $ saving_acct     : Factor w/ 5 levels "A61","A62","A63",..: 5 1 1 1 1 5 3 1 4 1 ...
##  $ present_emp     : Factor w/ 5 levels "A71","A72","A73",..: 5 3 4 4 3 3 5 3 4 1 ...
##  $ installment_rate: int  4 2 2 2 3 2 3 2 2 4 ...
##  $ sex             : Factor w/ 4 levels "A91","A92","A93",..: 3 2 3 3 3 3 3 3 1 4 ...
##  $ other_debtor    : Factor w/ 3 levels "A101","A102",..: 1 1 1 3 1 1 1 1 1 1 ...
##  $ present_resid   : int  4 2 3 4 4 4 4 2 4 2 ...
##  $ property        : Factor w/ 4 levels "A121","A122",..: 1 1 1 2 4 4 2 3 1 3 ...
##  $ age             : int  67 22 49 45 53 35 53 35 61 28 ...
##  $ other_install   : Factor w/ 3 levels "A141","A142",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ housing         : Factor w/ 3 levels "A151","A152",..: 2 2 2 3 3 3 2 1 2 2 ...
##  $ n_credits       : int  2 1 1 1 2 1 1 1 1 2 ...
##  $ job             : Factor w/ 4 levels "A171","A172",..: 3 3 2 3 3 2 3 4 2 4 ...
##  $ n_people        : int  1 1 2 2 2 2 1 1 1 1 ...
##  $ telephone       : Factor w/ 2 levels "A191","A192": 2 1 1 1 1 2 1 2 1 1 ...
##  $ foreign         : Factor w/ 2 levels "A201","A202": 1 1 1 1 1 1 1 1 1 1 ...
##  $ response        : int  1 2 1 1 2 1 1 1 1 2 ...
```

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------
---------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.1       v purrr   0.2.5
## v tibble  2.0.1       v dplyr   0.8.0.1
## v tidyr   0.8.2       v stringr 1.3.1
## v readr   1.2.1       v forcats 0.3.0
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
## Warning: package 'tibble' was built under R version 3.5.2
```

```
## Warning: package 'tidyr' was built under R version 3.5.2
```

```
## Warning: package 'purrr' was built under R version 3.5.2
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

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

```
(nrow(filter(german_credit, chk_acct == 'A11' & response == 2))/nrow(filter(german_credit,respon
se == 2 )))*100
```

```
## [1] 45
```

```
(nrow(filter(german_credit, chk_acct == 'A14' & response == 1))/nrow(filter(german_credit,respon
se == 1 )))*100
```

```
## [1] 49.71429
```

```
german_credit%>%
ggplot(aes(response, fill=chk_acct))+geom_bar()
```
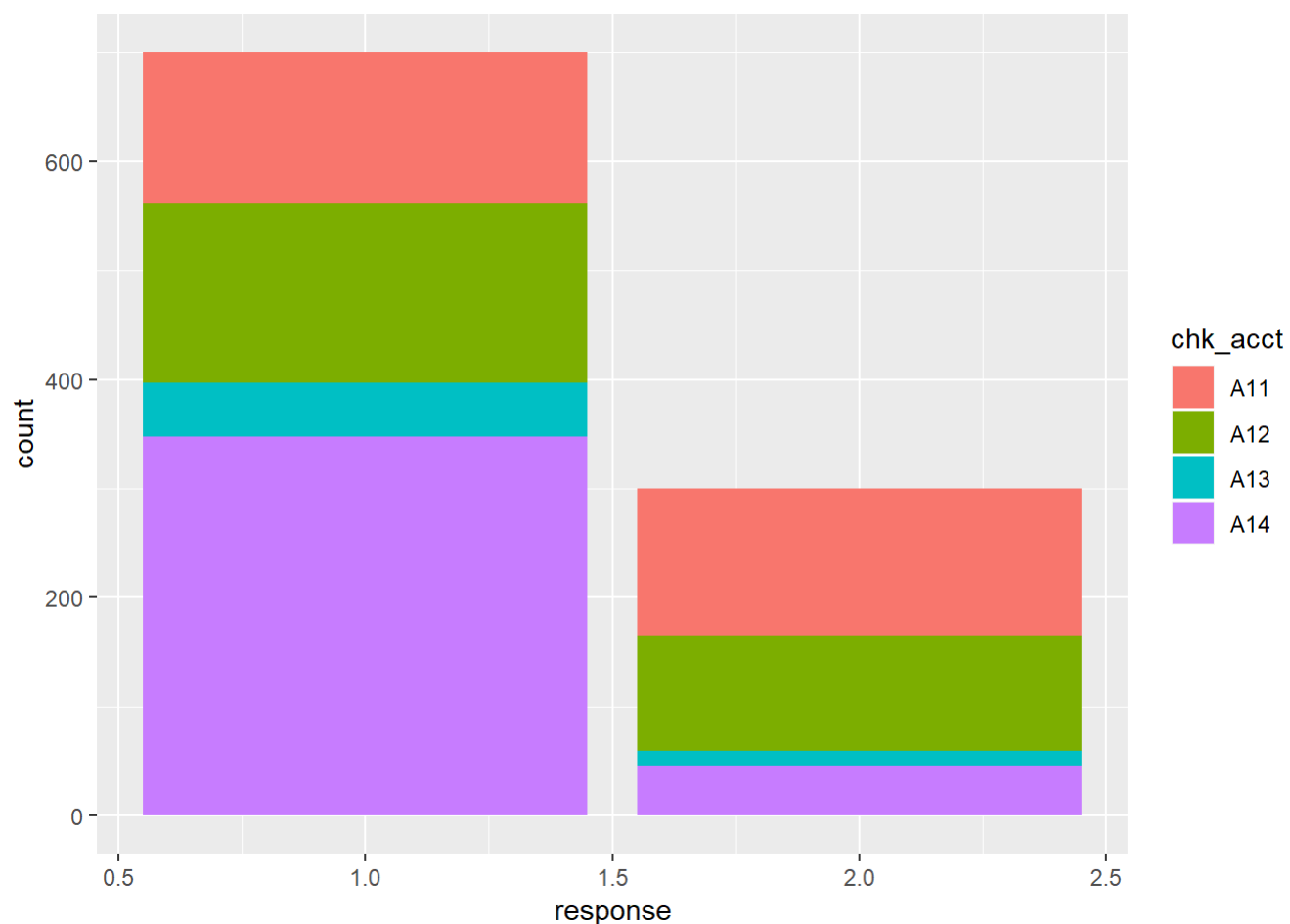
Status of existing checking account

A11 : … < 0 DM

A12 : 0 <= … < 200 DM

A13 : … >= 200 DM / salary assignments for at least 1 year
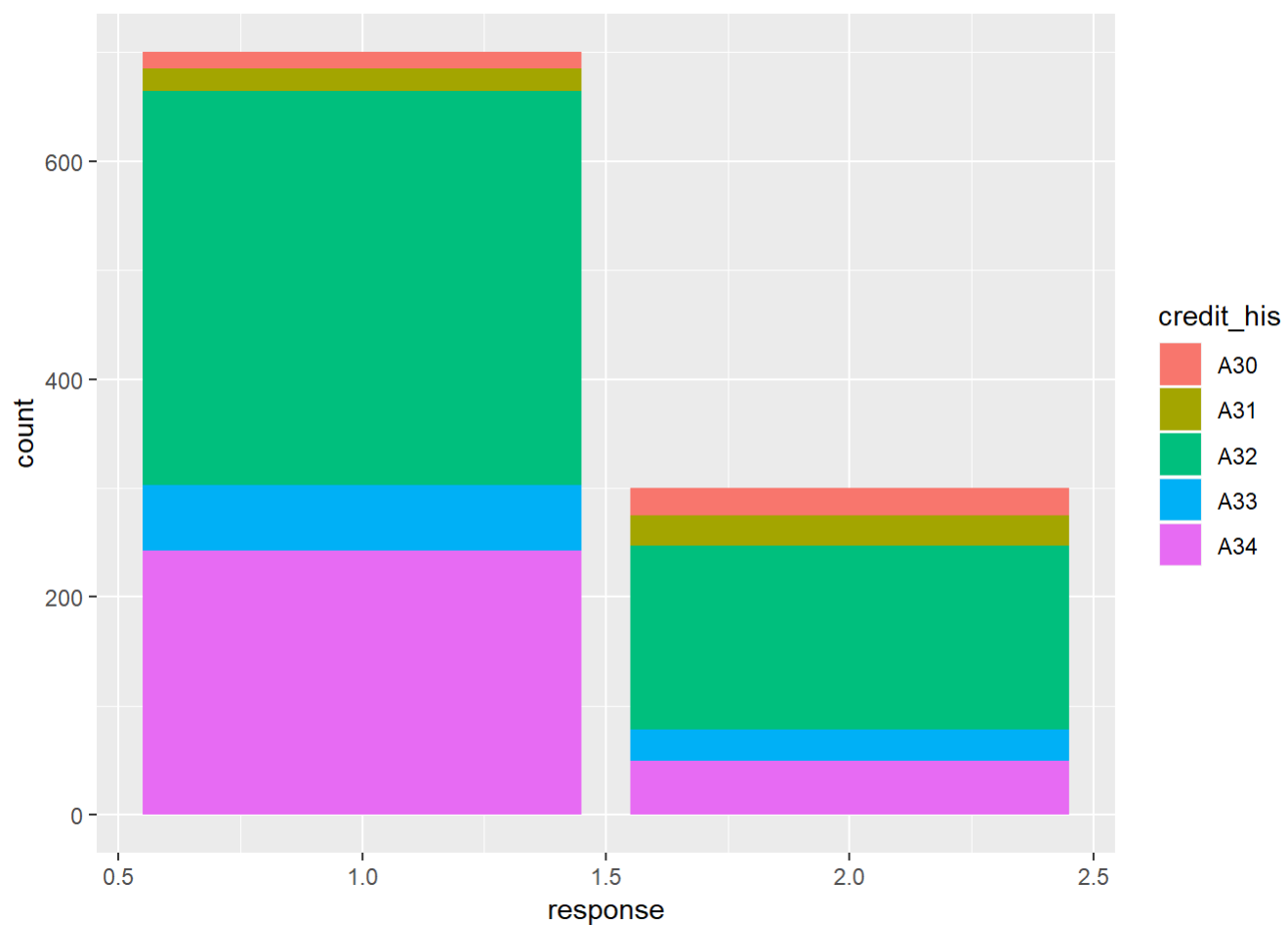
A14 : no checking account

```
(nrow(filter(german_credit, credit_his == 'A32' & response == 2))/nrow(filter(german_credit,resp
onse == 2 )))*100
```

```
## [1] 56.33333
```

```
(nrow(filter(german_credit, credit_his == 'A32' & response == 1))/nrow(filter(german_credit,resp
onse == 1 )))*100
```

```
## [1] 51.57143
```

```
german_credit%>%
ggplot(aes(response, fill=credit_his))+geom_bar()
```

Credit history

A30 : no credits taken/ all credits paid back duly

A31 : all credits at this bank paid back duly

A32 : existing credits paid back duly till now

A33 : delay in paying off in the past

A34 : critical account/ other credits existing (not at this bank)

```
german_credit%>%
ggplot(aes(response, fill=purpose))+geom_bar()
```
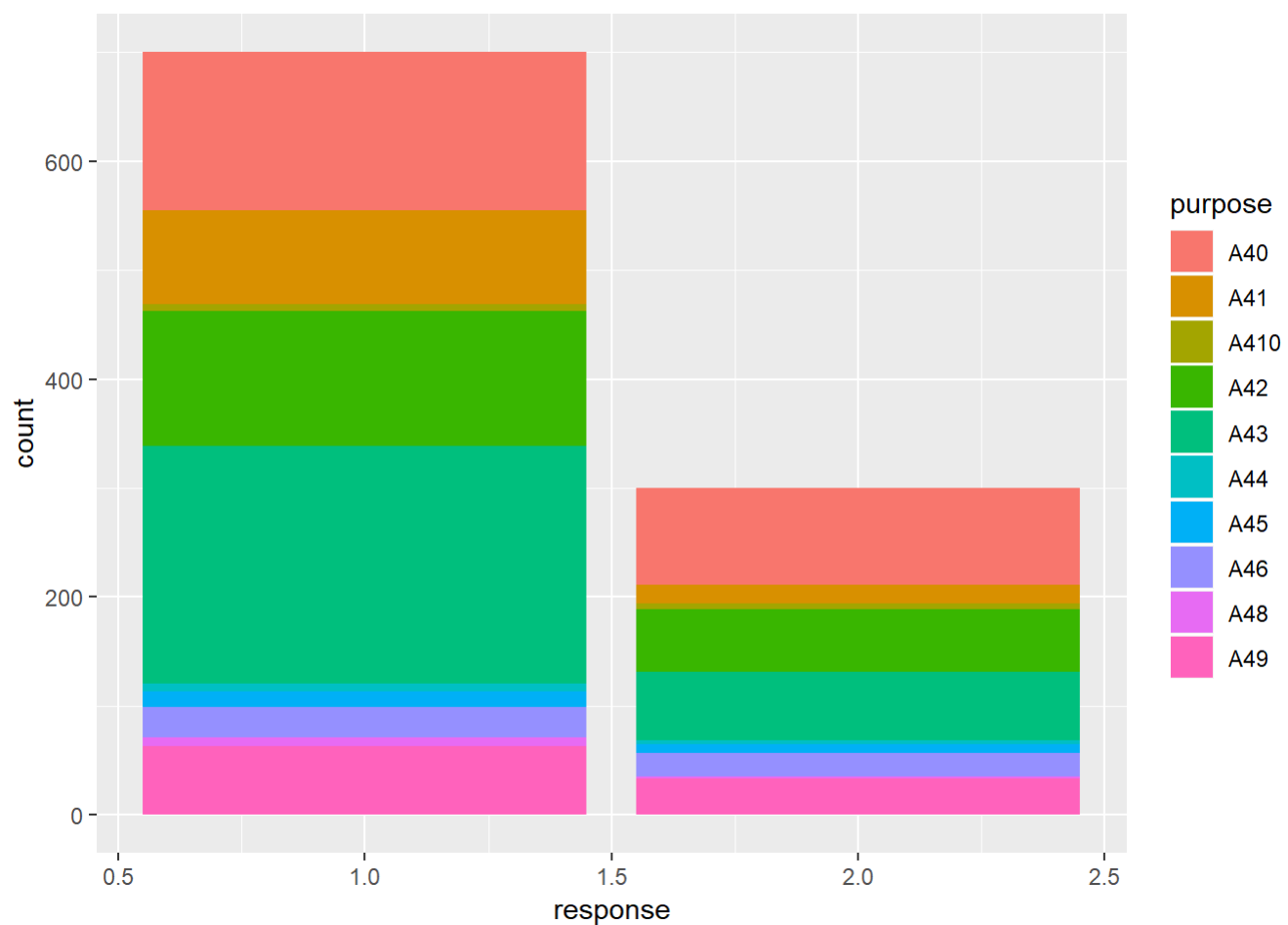
Purpose A40 : car (new) A41 : car (used) A42 : furniture/equipment A43 : radio/television A44 : domestic appliances A45 : repairs A46 : education A47 : (vacation - does not exist?) A48 : retraining A49 : business A410 : others

```
german_credit%>%
ggplot(aes(response, fill=saving_acct))+geom_bar()
```
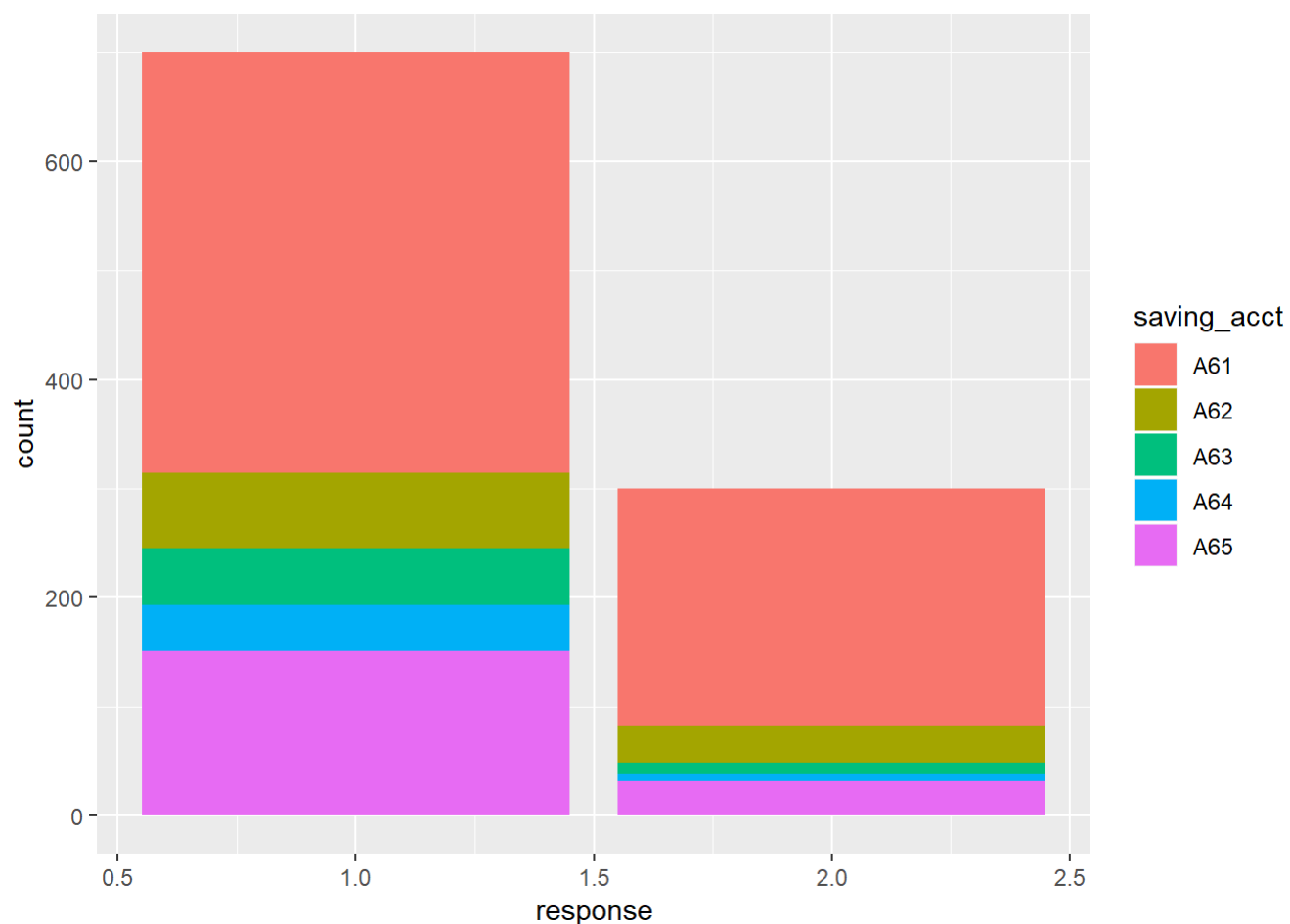
Savings account/bonds A61 : … < 100 DM A62 : 100 <= … < 500 DM A63 : 500 <= … < 1000 DM A64 : .. >= 1000
DM A65 : unknown/ no savings account

```
(nrow(filter(german_credit, present_emp == 'A71' & response == 2))/nrow(filter(german_credit,res
ponse == 2 )))*100
```

```
## [1] 7.666667
```

```
german_credit%>%
ggplot(aes(response, fill=present_emp))+geom_bar()
```
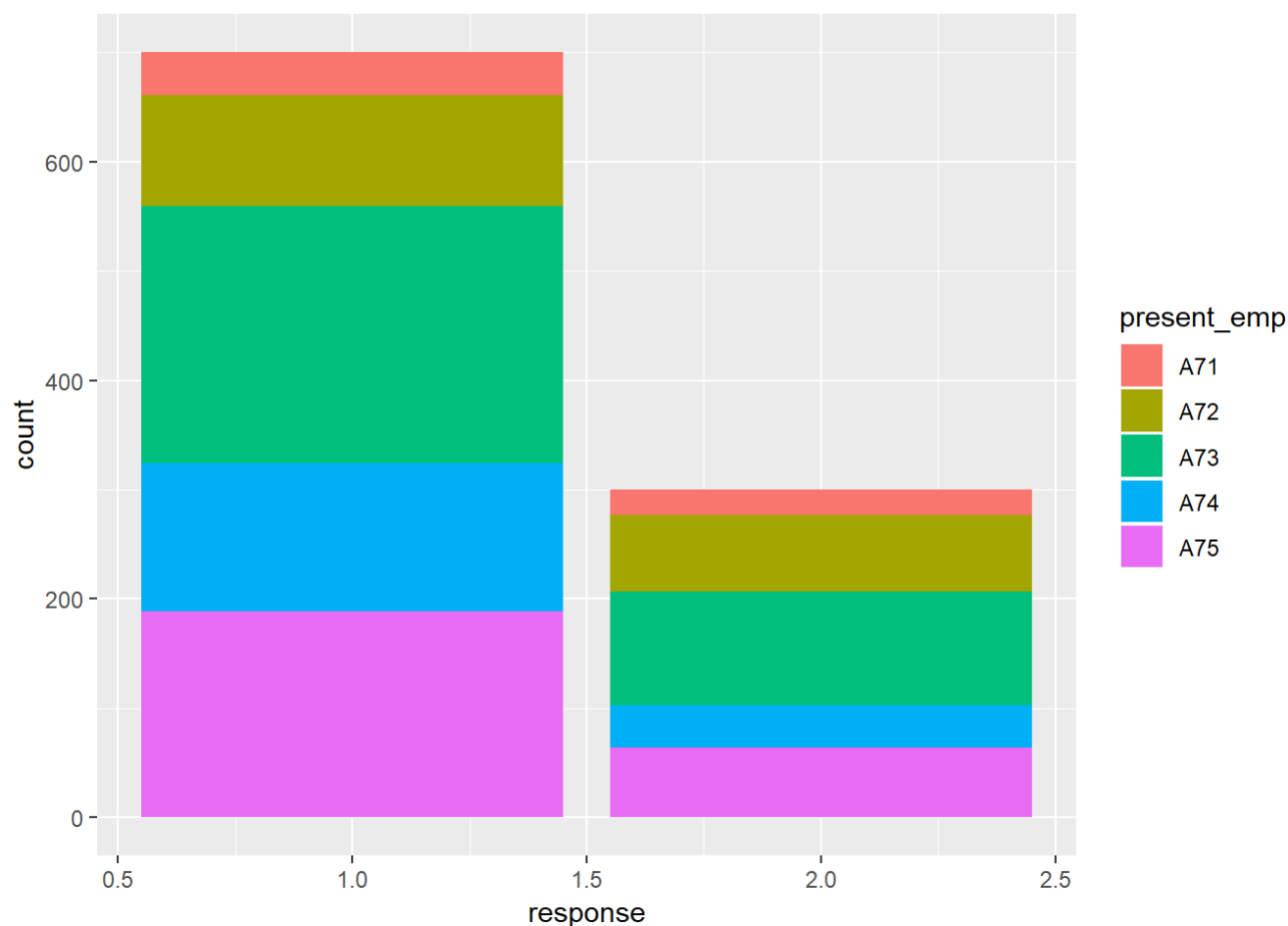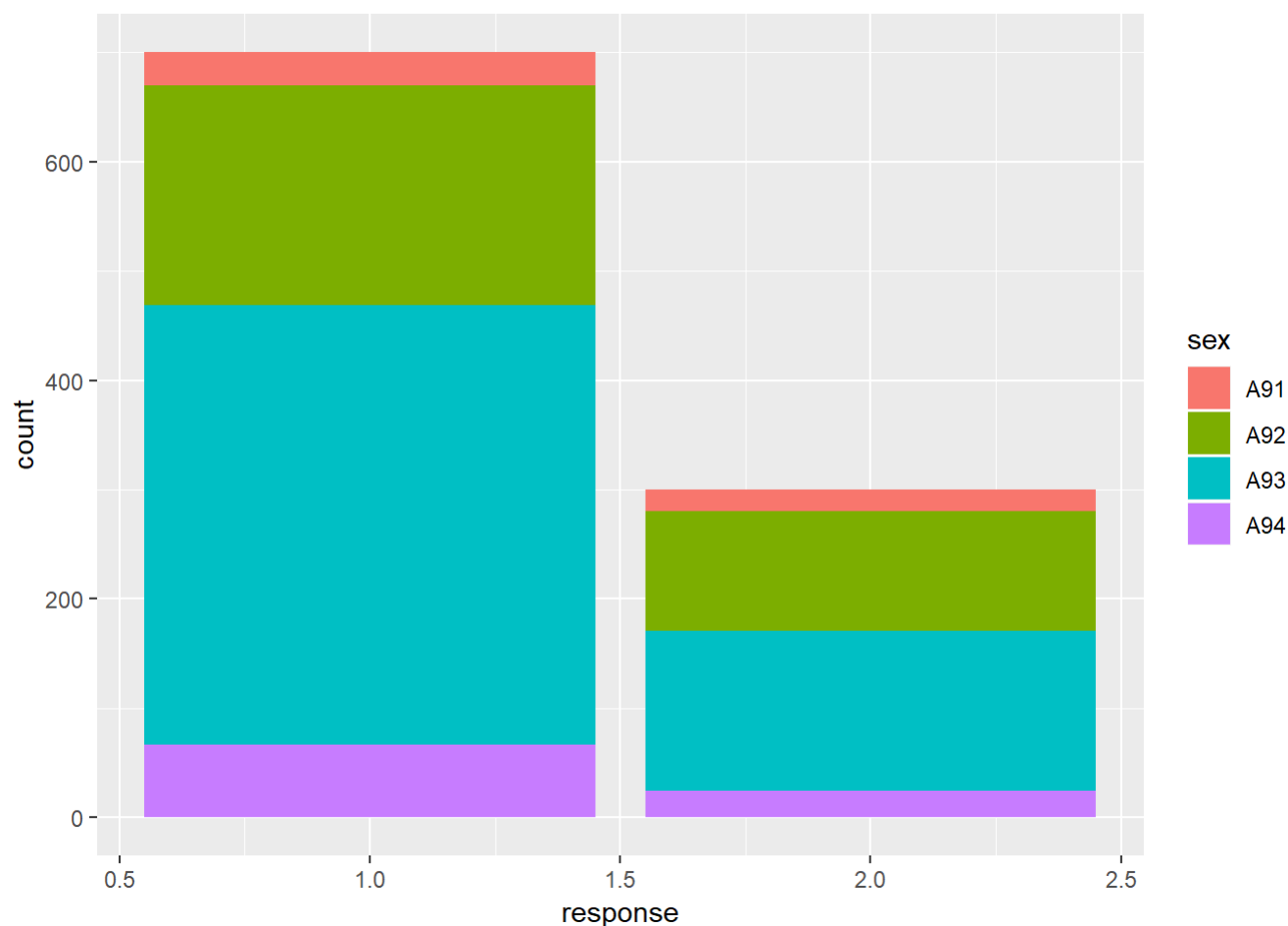
Present employment since A71 : unemployed A72 : … < 1 year A73 : 1 <= … < 4 years A74 : 4 <= … < 7 years A75 : .. >= 7 years

```
(nrow(filter(german_credit, sex == 'A93' & response == 2))/nrow(filter(german_credit,response ==
2 )))*100
```

```
## [1] 48.66667
```

```
(nrow(filter(german_credit, sex == 'A93' & response == 1))/nrow(filter(german_credit,response ==
1 )))*100
```

```
## [1] 57.42857
```

```
german_credit%>%
ggplot(aes(response, fill=sex))+geom_bar()
```

Personal status and sex A91 : male : divorced/separated A92 : female : divorced/separated/married A93 : male : single A94 : male : married/widowed A95 : female : single
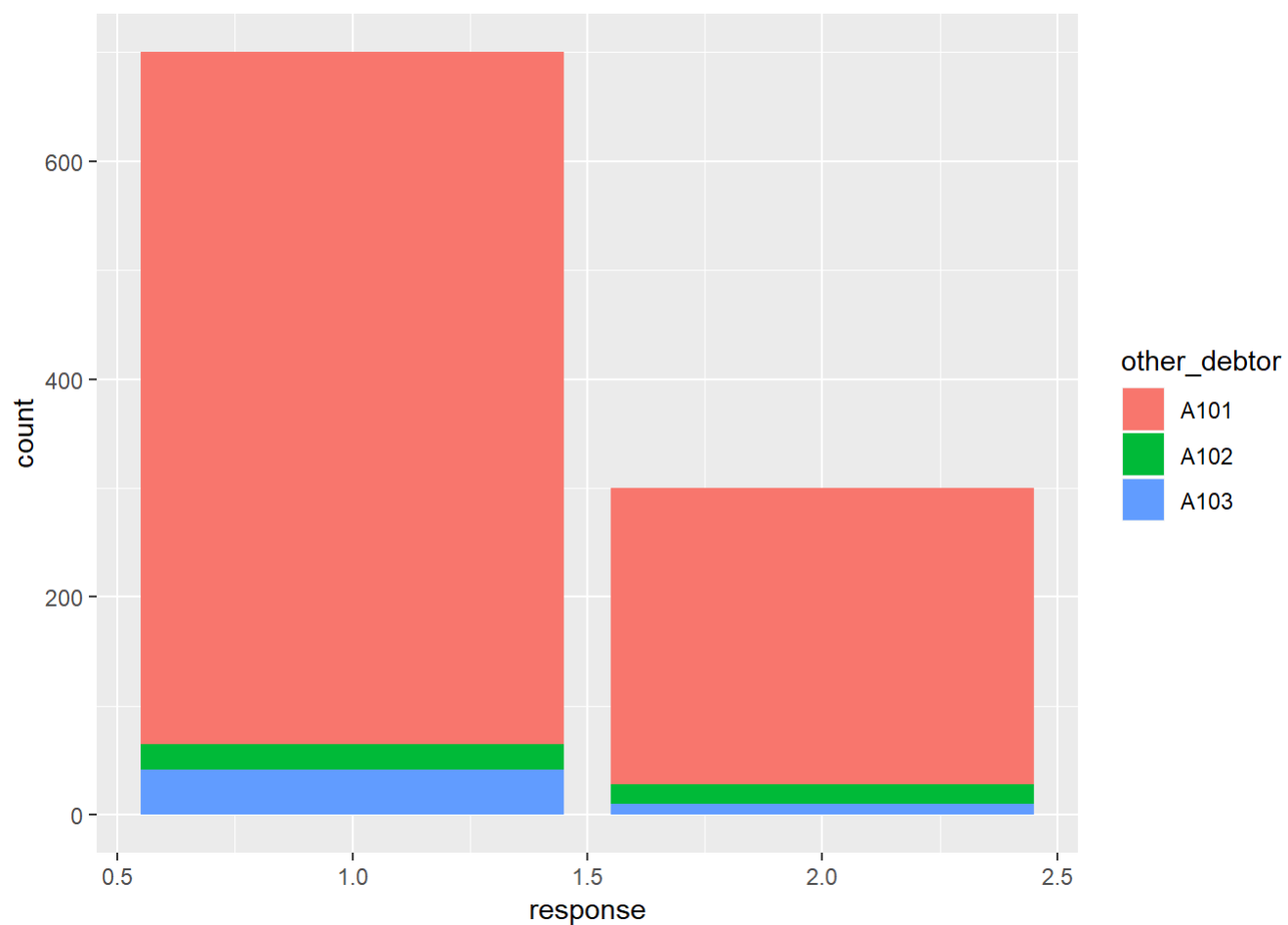
```
(nrow(filter(german_credit, other_debtor == 'A101' & response == 2))/nrow(filter(german_credit,r
esponse == 2 )))*100
```

```
## [1] 90.66667
```

```
(nrow(filter(german_credit, other_debtor == 'A101' & response == 1))/nrow(filter(german_credit,r
esponse == 1 )))*100
```

```
## [1] 90.71429
```

```
german_credit%>%
ggplot(aes(response, fill=other_debtor))+geom_bar()
```

Other debtors / guarantors A101 : none A102 : co-applicant A103 : guarantor
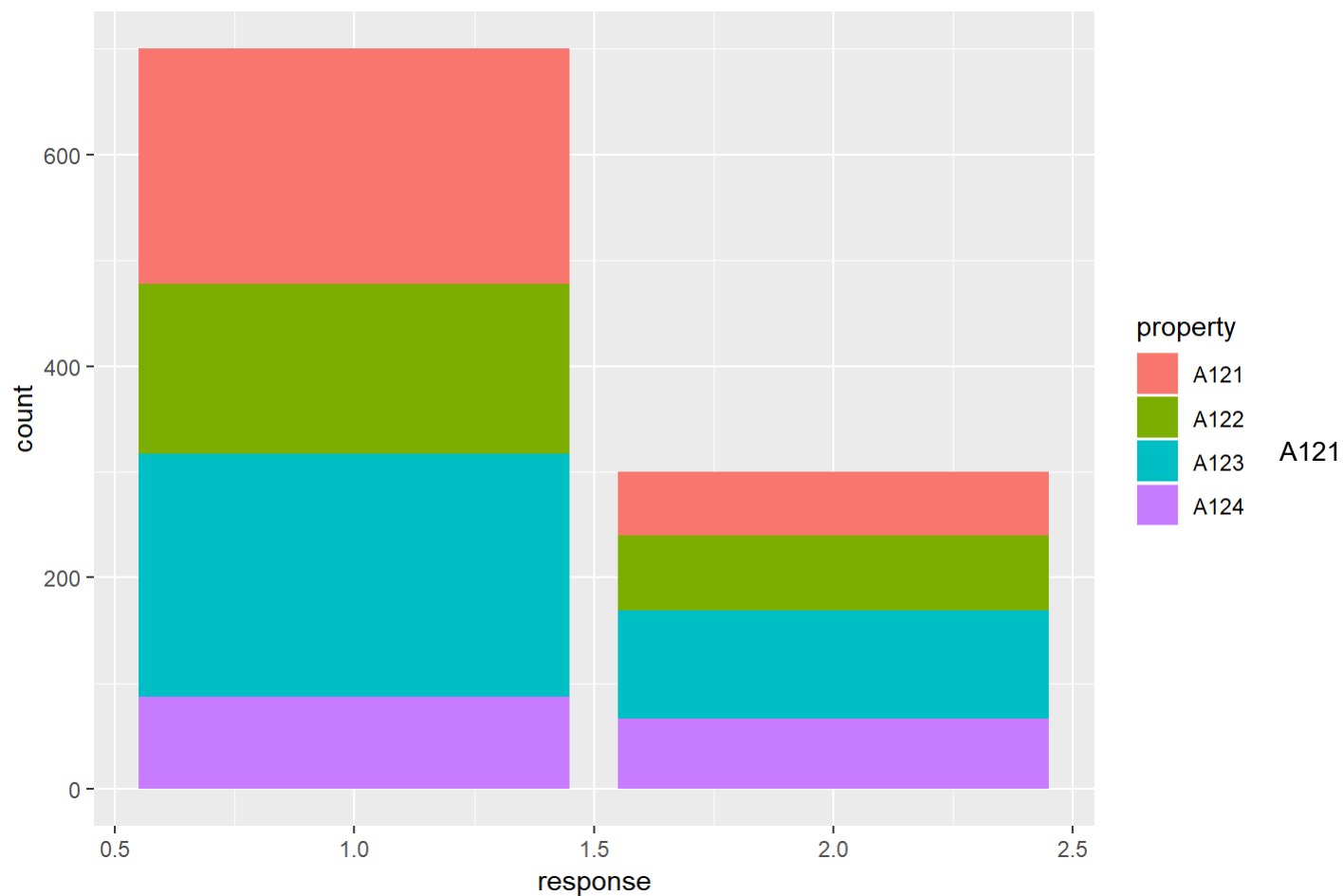
```
(nrow(filter(german_credit, property == 'A121' & response == 1))/nrow(filter(german_credit,respo
nse == 1 )))*100
```

```
## [1] 31.71429
```

```
(nrow(filter(german_credit, property == 'A123' & response == 1))/nrow(filter(german_credit,respo
nse == 1 )))*100
```
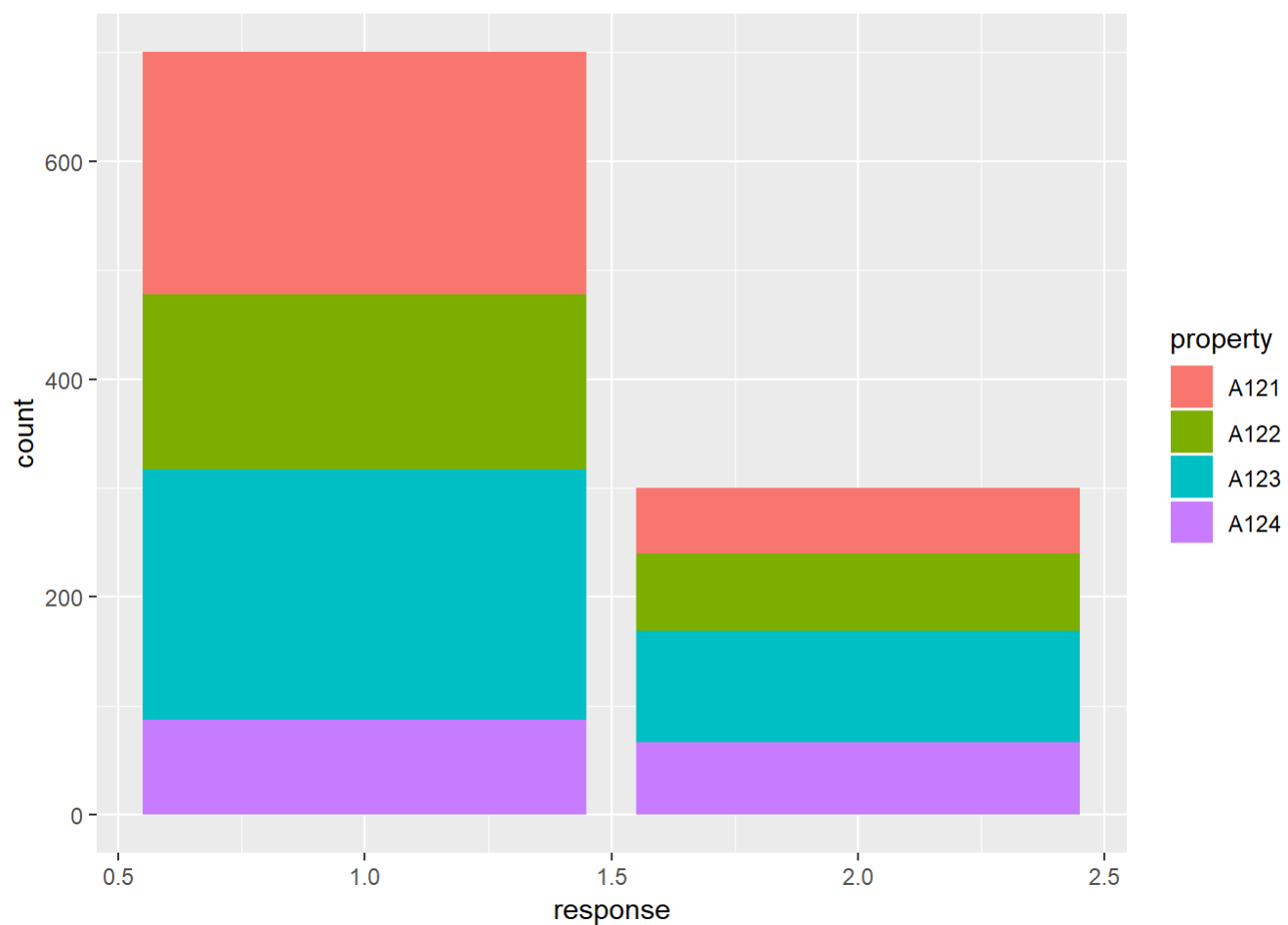
```
## [1] 32.85714
```

```
german_credit%>%
ggplot(aes(response, fill=property))+geom_bar()
```

: real estate A122 : if not A121 : building society savings agreement/ life insurance A123 : if not A121/A122 : car or other, not in attribute 6 A124 : unknown / no property

```
german_credit%>%
ggplot(aes(response, fill=property))+geom_bar()
```
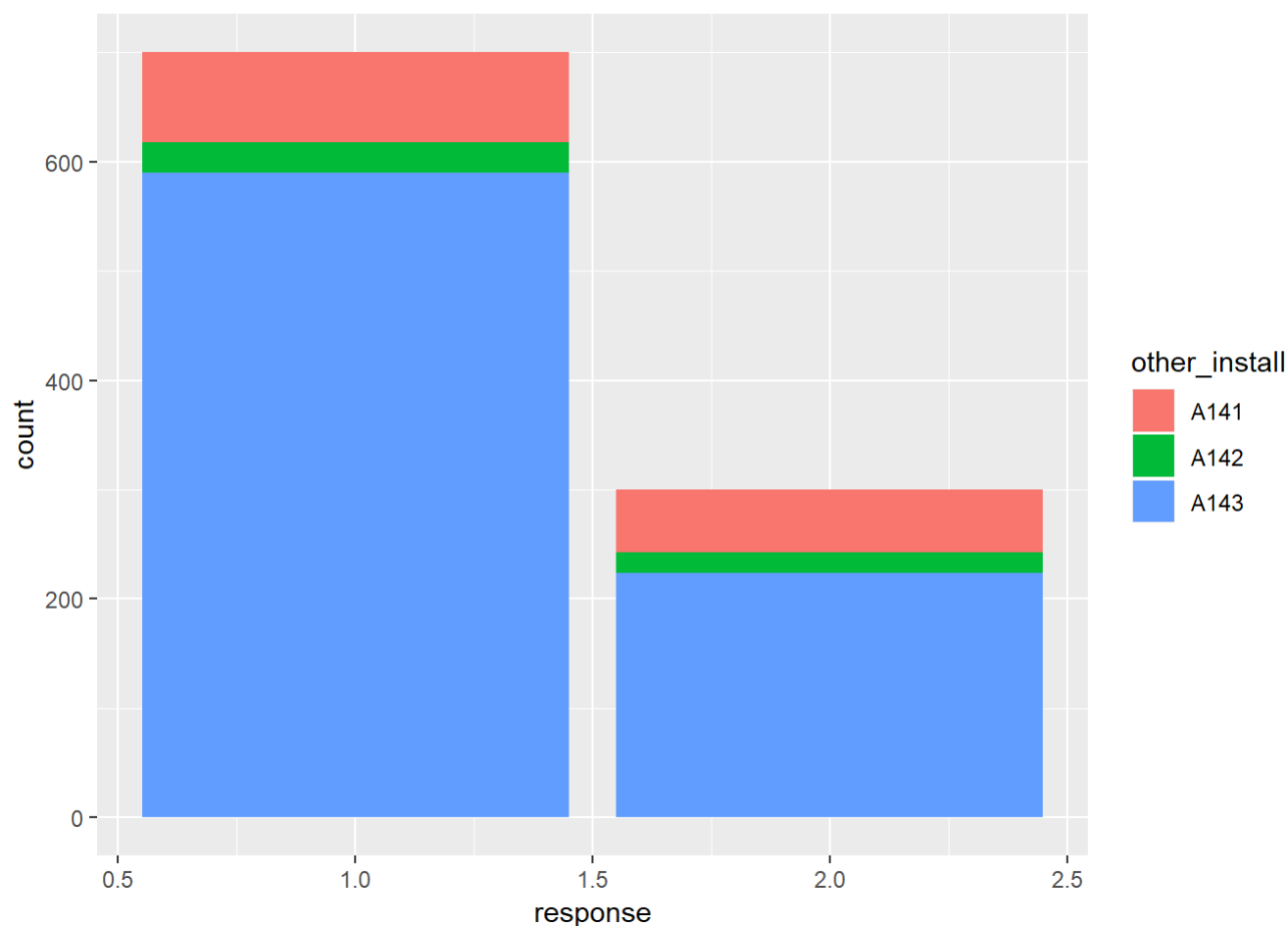
```
(nrow(filter(german_credit, other_install == 'A143' & response == 2))/nrow(filter(german_credit,
response == 2 )))*100
```

```
## [1] 74.66667
```

```
(nrow(filter(german_credit, other_install == 'A143' & response == 1))/nrow(filter(german_credit,
response == 1 )))*100
```

```
## [1] 84.28571
```

```
german_credit%>%
ggplot(aes(response, fill=other_install))+geom_bar()
```

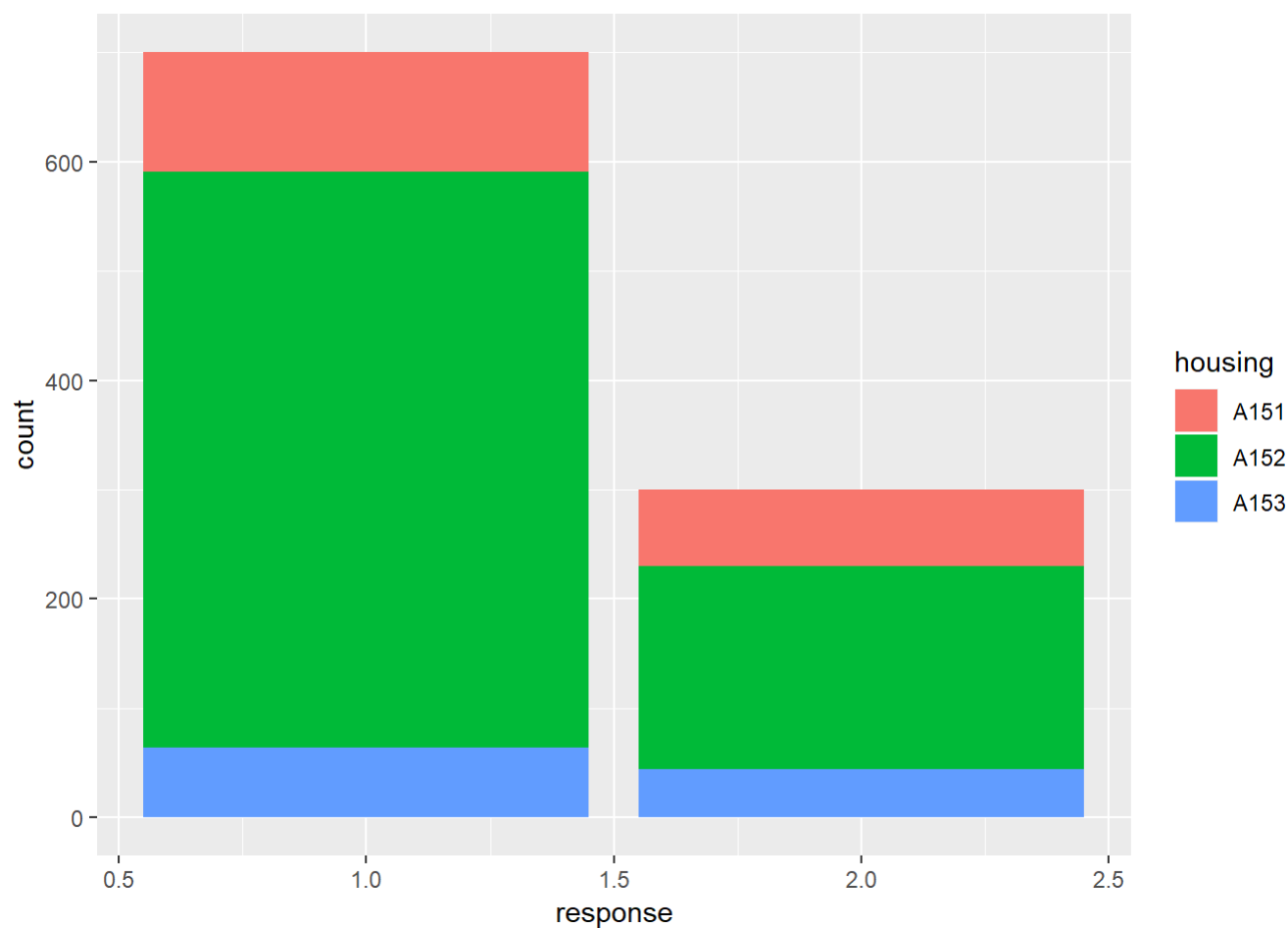Other installment plans A141 : bank A142 : stores A143 : none

```
(nrow(filter(german_credit, housing == 'A152' & response == 2))/nrow(filter(german_credit,respon
se == 2 )))*100
```

```
## [1] 62
```

```
(nrow(filter(german_credit, housing == 'A152' & response == 1))/nrow(filter(german_credit,respon
se == 1 )))*100
```

```
## [1] 75.28571
```

```
german_credit%>%
ggplot(aes(response, fill=housing))+geom_bar()
```

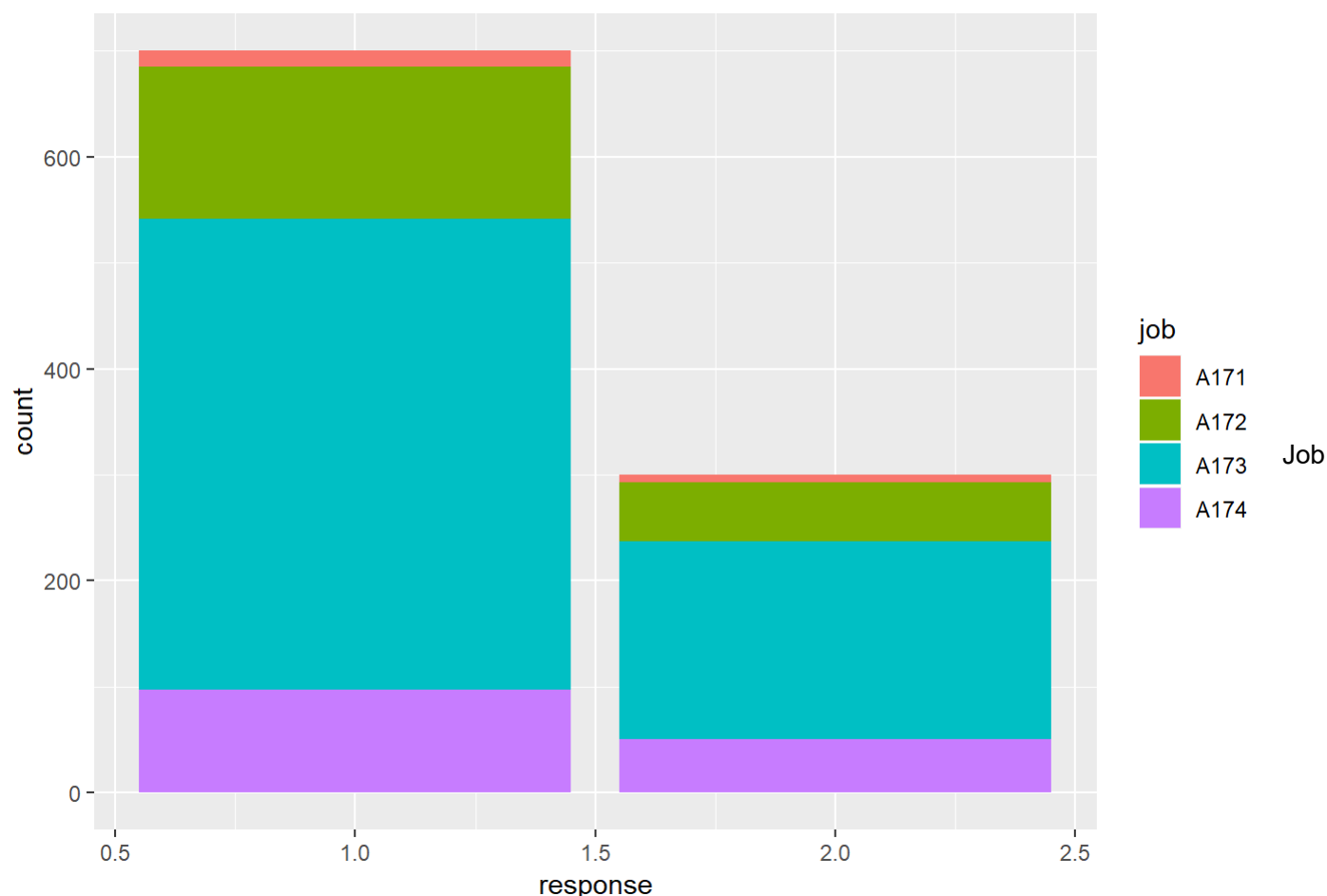Housing A151 : rent A152 : own A153 : for free

```
(nrow(filter(german_credit, job == 'A173' & response == 2))/nrow(filter(german_credit,response =
= 2 )))*100
```

```
## [1] 62
```

```
(nrow(filter(german_credit, job == 'A173' & response == 1))/nrow(filter(german_credit,response =
= 1 )))*100
```

```
## [1] 63.42857
```

```
german_credit%>%
ggplot(aes(response, fill=job))+geom_bar()
```

A171 : unemployed/ unskilled - non-resident A172 : unskilled - resident A173 : skilled employee / official A174 : management/ self-employed/ highly qualified employee/ officer
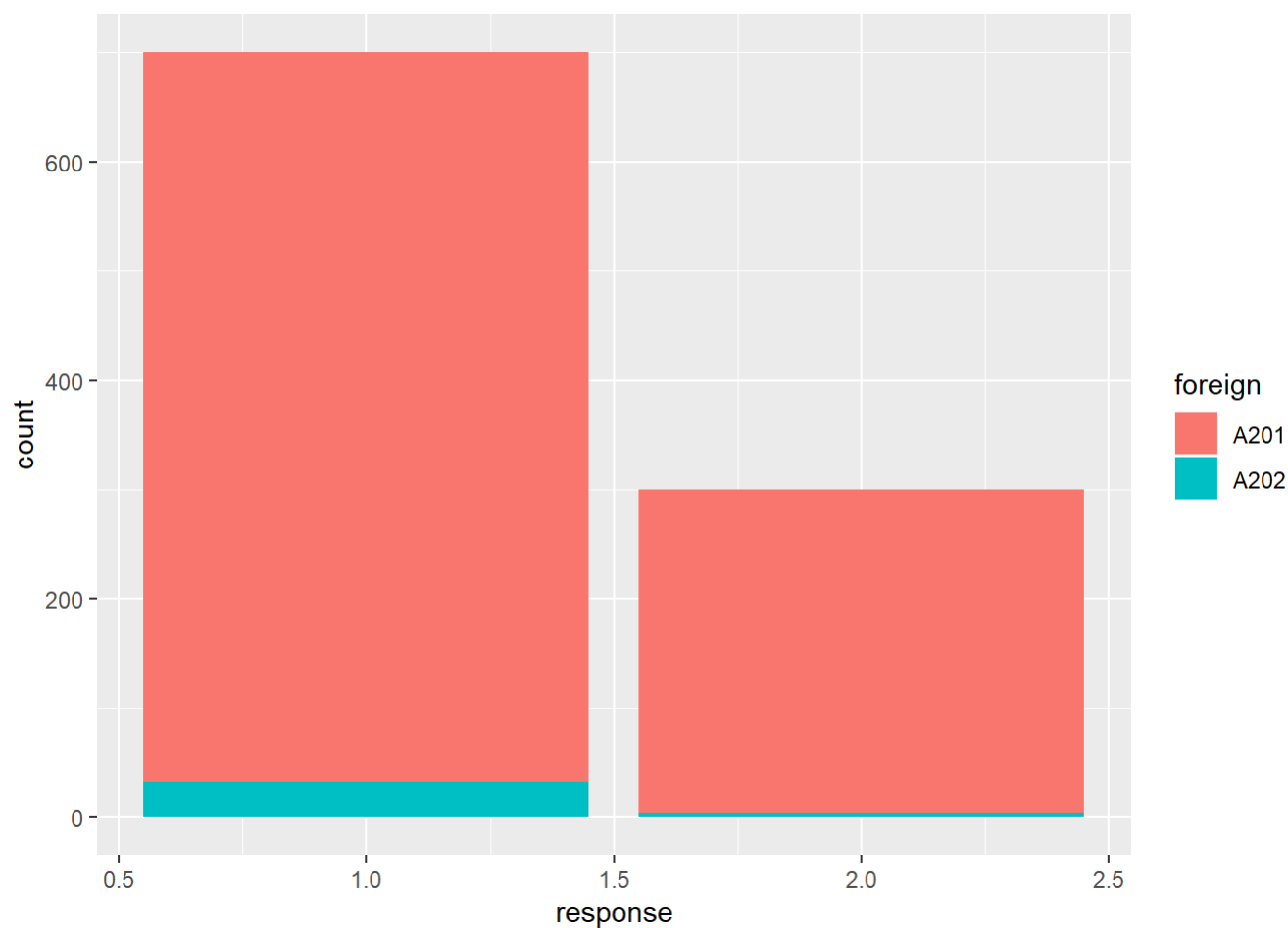
```
(nrow(filter(german_credit, foreign == 'A201' & response == 2))/nrow(filter(german_credit,respon
se == 2 )))*100
```

```
## [1] 98.66667
```

```
(nrow(filter(german_credit, foreign == 'A201' & response == 1))/nrow(filter(german_credit,respon
se == 1 )))*100
```

```
## [1] 95.28571
```

```
german_credit%>%
ggplot(aes(response, fill=foreign))+geom_bar()
```
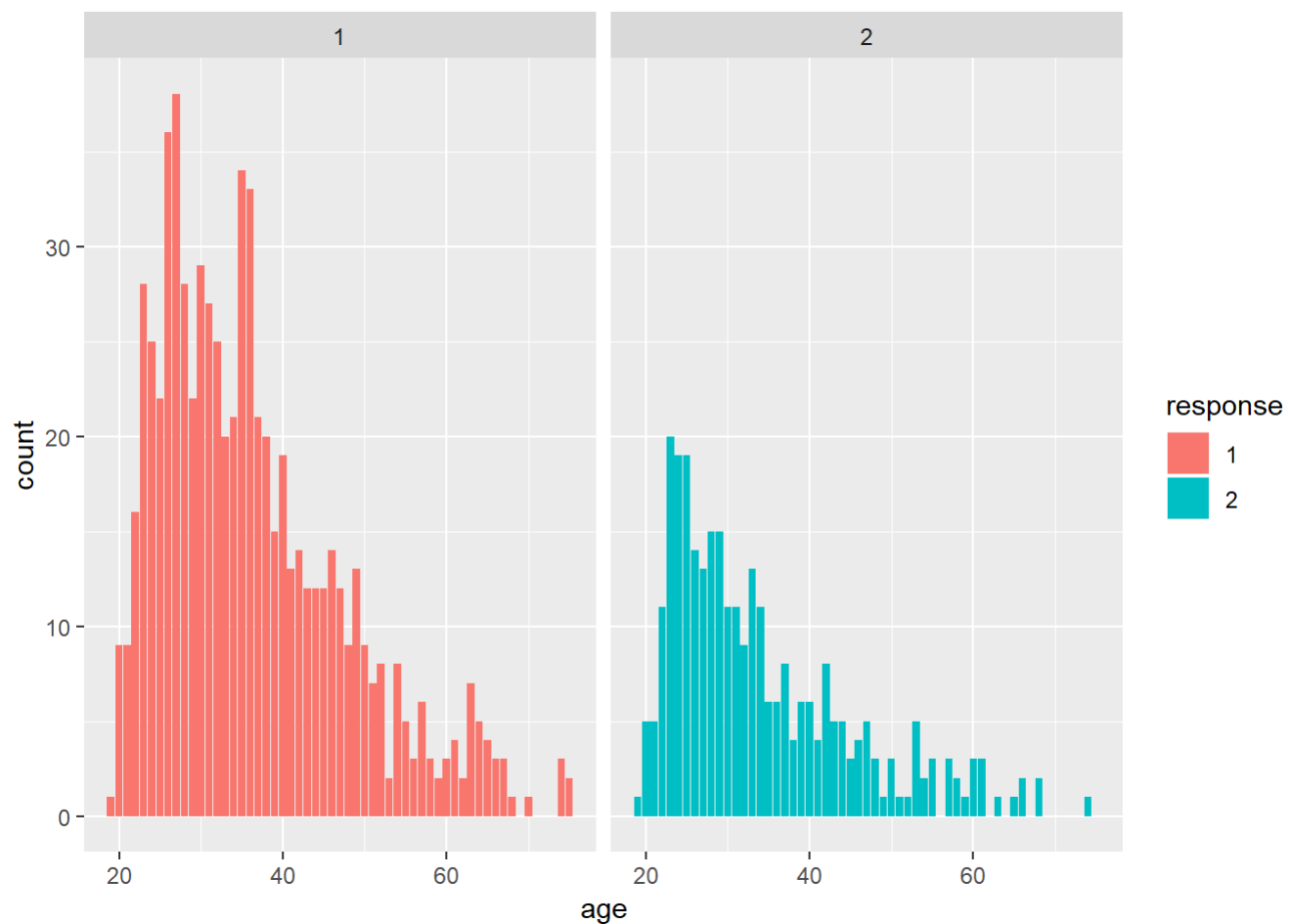
foreign worker A201 : yes A202 : no

```
(nrow(filter(german_credit,  response == 2))/nrow(german_credit))*100
```

```
## [1] 30
```

```
(nrow(filter(german_credit,  response == 1))/nrow(german_credit))*100
```

```
## [1] 70
```

```
german_credit$response <- as.factor(german_credit$response)
german_credit%>%
ggplot(aes(age))+geom_bar(aes(fill=response))+facet_wrap(~response)
```

```
german_credit%>%
ggplot(aes(amount))+geom_histogram(aes(fill=response))+facet_wrap(~response)
```

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

Credit risk EDA