R Notebook

Code ▾

Libraries, Data bases and Data Normalization

Hide

```
library("tidyverse")
library("dplyr")
library("tidyr")
library("stringr")
library("lubridate")
library("readr")
library("ggplot2")
library("scales")
```

Hide

```
account <- read.csv2('C:/Users/thami/OneDrive/Desktop/Berka/account.asc', sep = ';', stringsA
sFactor = FALSE)
```

Hide

```
  account %>% mutate(frequency = if_else(frequency == "POPLATEK MESICNE","Monthly Issuance",
                                       if_else(frequency == "POPLATEK TYDNE", "Weekly Issuanc
e",
                                               if_else(frequency == "POPLATEK PO OBRATU", "Is
suance After Transaction","")))) %>%
  mutate(date = ymd((str_c("19",date)))) -> account

account <- rename(account, date_account = date)
```

Hide

```
View(account)
```

Hide

```
client <- read.csv2('C:/Users/thami/OneDrive/Desktop/Berka/client.asc', sep = ';', stringsAsF
actor = FALSE)
```

Hide

```
  client %>% mutate(year_birth = str_c("19",str_sub(birth_number,1,2)),
                    month_birth = str_sub(birth_number,3,4),
                    day_birth = str_sub(birth_number,5,6)) %>%
    mutate(client_sex = if_else(month_birth > 50 , "F","M")) %>%
    mutate(month_birth = if_else(client_sex == "M" ,month_birth, ifelse((as.numeric(month_bir
th) - 50) < 10 ,str_c("0",(as.numeric(month_birth) - 50)),(as.numeric(month_birth) - 50)))) %
>%
    mutate(birth_date = ymd(str_c(year_birth,month_birth,day_birth, sep = "-"))) %>%
    select(client_id,birth_date,client_sex,district_id) -> client
```

Hide

```
View(client)
str(client)
```

```
'data.frame':	5369 obs. of  4 variables:
 $ client_id  : int  1 2 3 4 5 6 7 8 9 10 ...
 $ birth_date : Date, format: "1970-12-13" "1945-02-04" ...
 $ client_sex : chr  "F" "M" "F" "M" ...
 $ district_id: int  18 1 1 5 5 12 15 51 60 57 ...
```

Hide

```
disposition <- read.csv2('C:/Users/thami/OneDrive/Desktop/Berka/disp.asc', sep = ';', stringsAsFactor = FALSE)
```

Hide

```
order <- read.csv2('C:/Users/thami/OneDrive/Desktop/Berka/order.asc', sep = ';', stringsAsFactor = FALSE)
```

Hide

```
order %>% mutate( tp_payment = if_else(k_symbol == "POJISTNE", "Insurrance",
                                if_else(k_symbol == "SIPO", "Household Payment",
                                if_else(k_symbol == "LEASING", "Leasing",
                                if_else(k_symbol == "UVER", "Loan Payment","Other"))))) %>%
     select(-k_symbol) -> order
```

Hide

```
transaction <- read.csv2('C:/Users/thami/OneDrive/Desktop/Berka/trans.asc', sep = ';', stringsAsFactor = FALSE)
```

Hide

```
transaction %>%
    mutate(date = ymd((str_c("19",date)))) %>%
    mutate(type = if_else(type == "PRIJEM", "Credit",
    if_else(type == "VYDAJ","Withdrawal",if_else(type == "VYBER","Withdrawal","")))) %>%
    mutate(operation = if_else( operation == "VYBER KARTOU" , "Credit Card Withdrawal",
                       if_else( operation == "VKLAD", "Credit in Cash",
                       if_else( operation == "PREVOD Z UCTU", "Collection from Another Bank",
                       if_else( operation == "VYBER", "Withdrawal in Cash",
                       if_else( operation == "PREVOD NA UCET", "Remittance to Another Ban
k","")))))) %>%
    mutate(tp_payment = if_else(k_symbol == "POJISTNE", "Insurrance",
                        if_else(k_symbol == "SLUZBY", "Payment for Statement",
                        if_else(k_symbol == "UROK", "Interest Credited",
                        if_else(k_symbol == "SANKC. UROK", "Sanction Interest if Negative",
                        if_else(k_symbol == "SIPO", "Household Payment",
                        if_else(k_symbol == "DUCHOD", "Old-Age Pension",
                        if_else(k_symbol == "UVER", "Loan Payment",""))))))))) %>%
    select(-k_symbol)-> transaction

    transaction <- rename(transaction, date_trans = date )
```

Hide

```
loan <- read.csv2('C:/Users/thami/OneDrive/Desktop/Berka/loan.asc', sep = ';', stringsAsFacto
r = FALSE)
```

Hide

```
 loan %>%
    mutate(date = ymd((str_c("19",date)))) %>%
    mutate(status_descr = if_else(status == "A","A. Contract Finished, no problems",
                          if_else(status == "B","B. Contract Finished, Loan not Payed",
                          if_else(status == "C","C. Running Contract, OK so far",
                          if_else(status == "D","D. Running Contract, Client in Debt","")))))
%>%
    mutate(status_descr = as.factor(status_descr))-> loan

  loan <- rename(loan, date_loan = date)
```

Hide

```
summary(card)
```

```
    card_id          disp_id        card_type       issued          
 Min.   :   1.0   Min.   :    9   junior :145   Min.   :1993-11-07  
 1st Qu.: 229.8   1st Qu.: 1387   classic:659   1st Qu.:1997-01-25  
 Median : 456.5   Median : 2938   gold   : 88   Median :1998-01-06  
 Mean   : 480.9   Mean   : 3512                 Mean   :1997-09-19  
 3rd Qu.: 684.2   3rd Qu.: 4460                 3rd Qu.:1998-08-05  
 Max.   :1247.0   Max.   :13660                 Max.   :1998-12-29  
```

Hide

```
View(card)
```

Hide

```
account <- tibble(account)
  client <- tibble(client)
  disposition <- tibble(disposition)
  order <- tibble(order)
  transaction <- tibble(transaction)
  loan <- tibble(loan)
  card <- tibble(card)
  district <- tibble(district)
```

Data Mining and Analysis

qtde_client <int>
5369

1 row

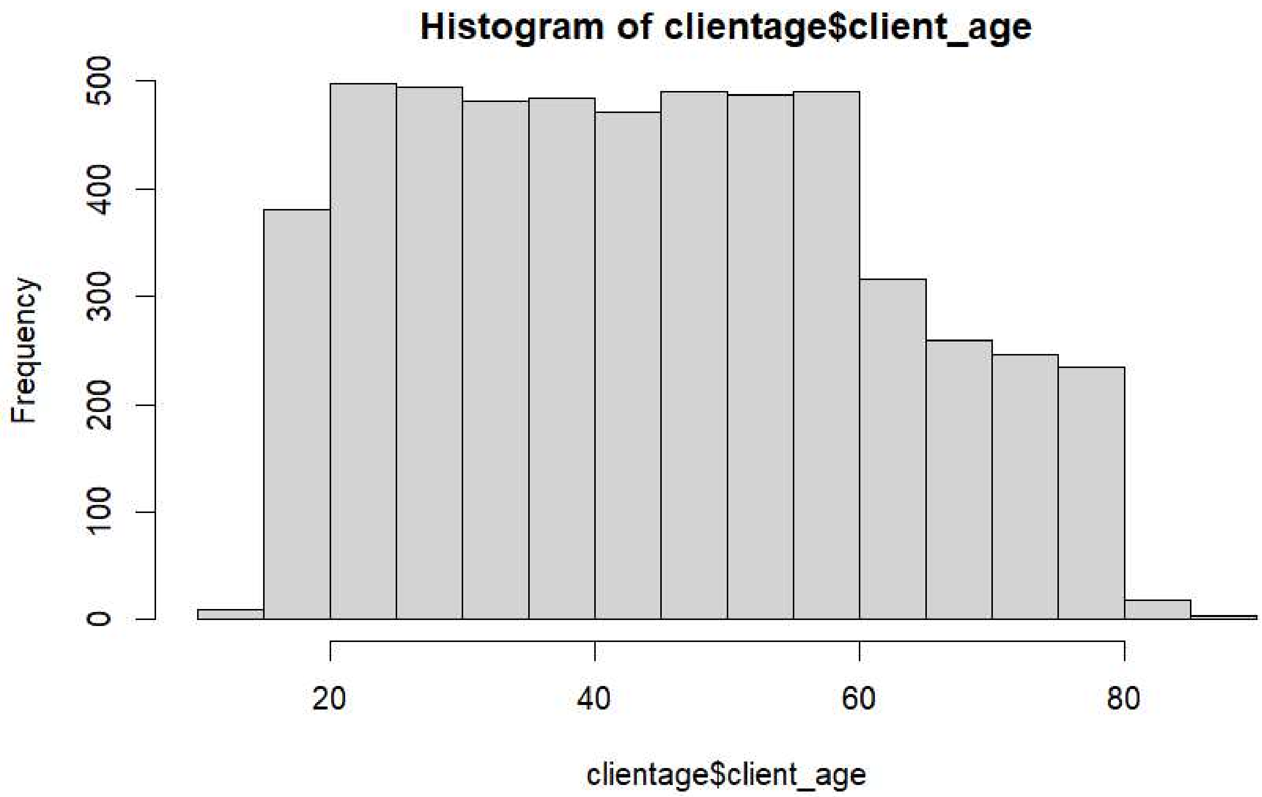
client_sex <chr>	n <int>
F	2645
M	2724

2 rows

Hide

```
clientage = client %>% mutate(client_age = year(as.period(interval(start = birth_date, end =
dbdate))))
summary(clientage$client_age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   11.0    30.0    44.0    44.8    58.0    87.0 
```



Hide

```
account %>%
    left_join(disposition, by = 'account_id') %>%
    left_join(district, by = 'district_id') %>%
    rename(account_district_name = district_name, account_region = region, account_district_i
d = district_id) %>%
    left_join(client, by = 'client_id') %>%
    left_join(district, by = 'district_id') %>%
    rename(client_district_name = district_name, client_region = region, client_district_id =
district_id) %>%
    select(account_id,
           frequency,
           date_account,
           account_district_id,
           account_district_name,
           account_region,
           disp_id,
           client_type,
           client_id,
           birth_date,
           client_sex,
           client_district_id,
           client_district_name,
           client_region) -> tb_account_client
```

Hide

```
client_district <- left_join(client, district, by = 'district_id')
View(client_district)
```

district_name	n
<chr>	<int>
Hl.m. Praha	663
Ostrava - mesto	180
Karvina	169
Brno - mesto	155
Zlin	109
Olomouc	104
Frydek - Mistek	86
Nachod	76
Usti nad Orlici	73
Kolin	71

1-10 of 77 rows Previous **1** 2 3 4 5 6 ... 8 Next

Hide

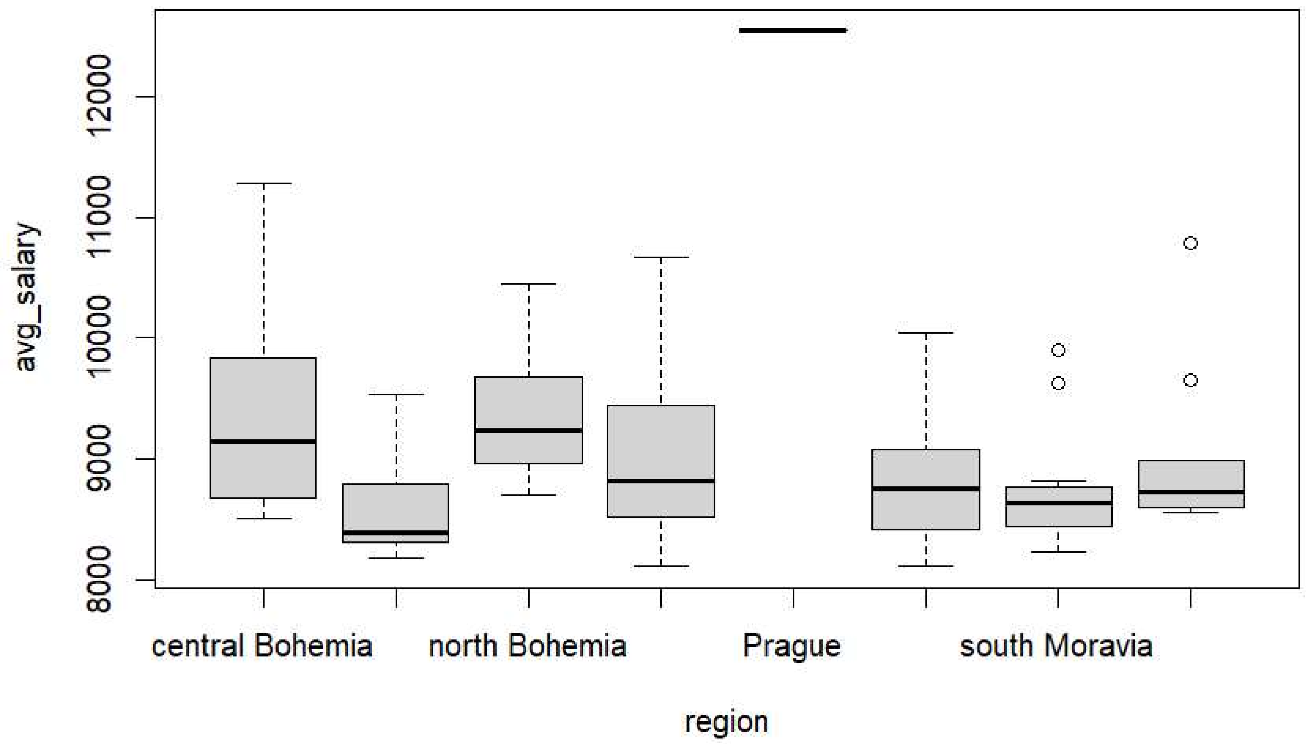
```
client_district %>% count(region) %>% arrange(desc(n))
```

region	n
<chr>	<int>
south Moravia	937
north Moravia	920
central Bohemia	664
Prague	663
east Bohemia	660
north Bohemia	561
west Bohemia	515
south Bohemia	449

8 rows

Hide

```
boxplot(avg_salary ~ region, data = district)
```



Hide

```
cli_dist_disp <- left_join(client_district, disposition, by = 'client_id')
View(cli_dist_disp)
```

Hide

```
client_card_all <- full_join(cli_dist_disp, card, by = 'disp_id')
View(client_card_all)
```

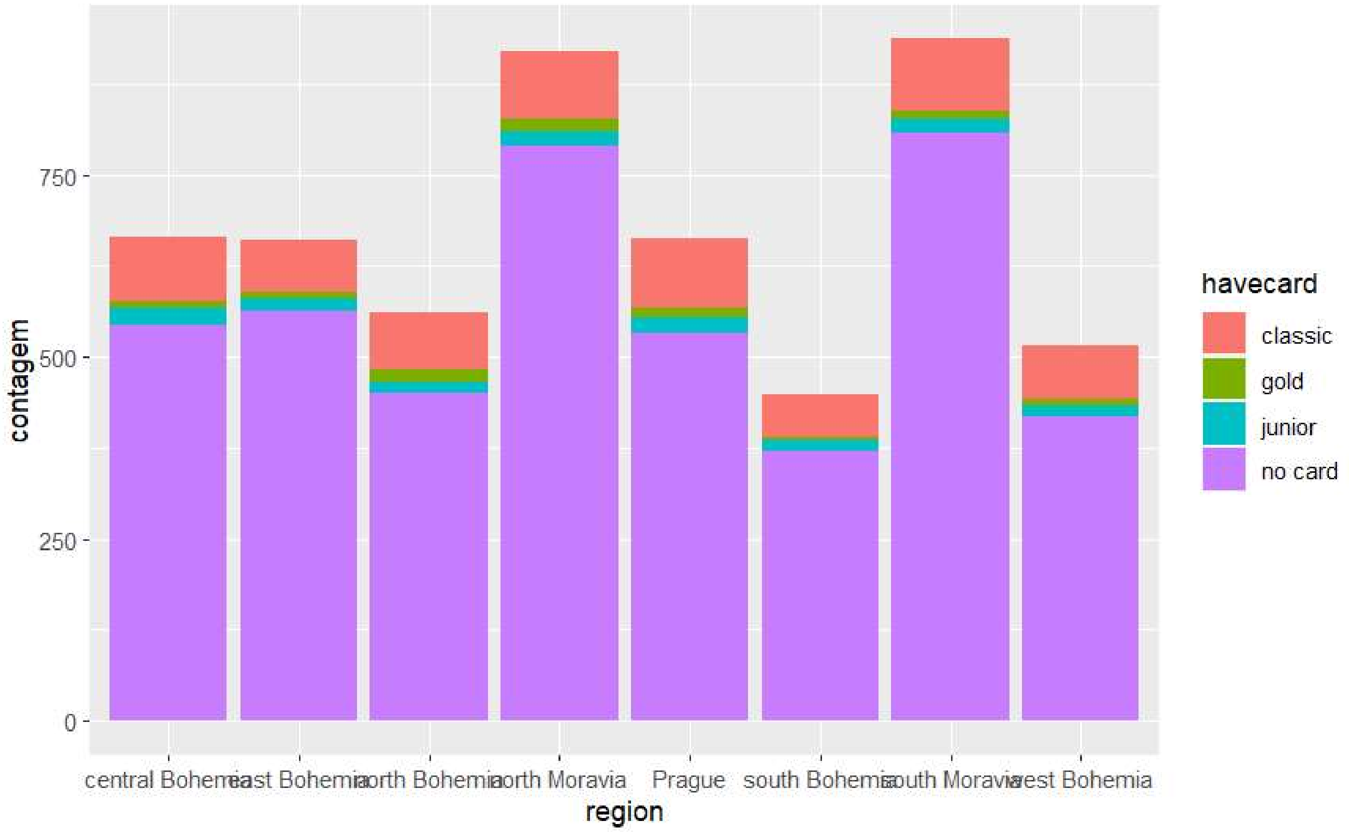
Hide

```
client_card_all <- mutate(client_card_all, havecard = if_else(is.na(card_id), 'no card', as.c
haracter(card_type)))
View(client_card_all)
```

Hide

```
client_card_all %>% mutate(contagem = 1) %>% group_by(region, havecard) %>% summarise(contage
m = sum(contagem)) %>%
  ggplot(aes(x = region, y = contagem, fill = havecard)) + geom_bar(stat = "identity")
```

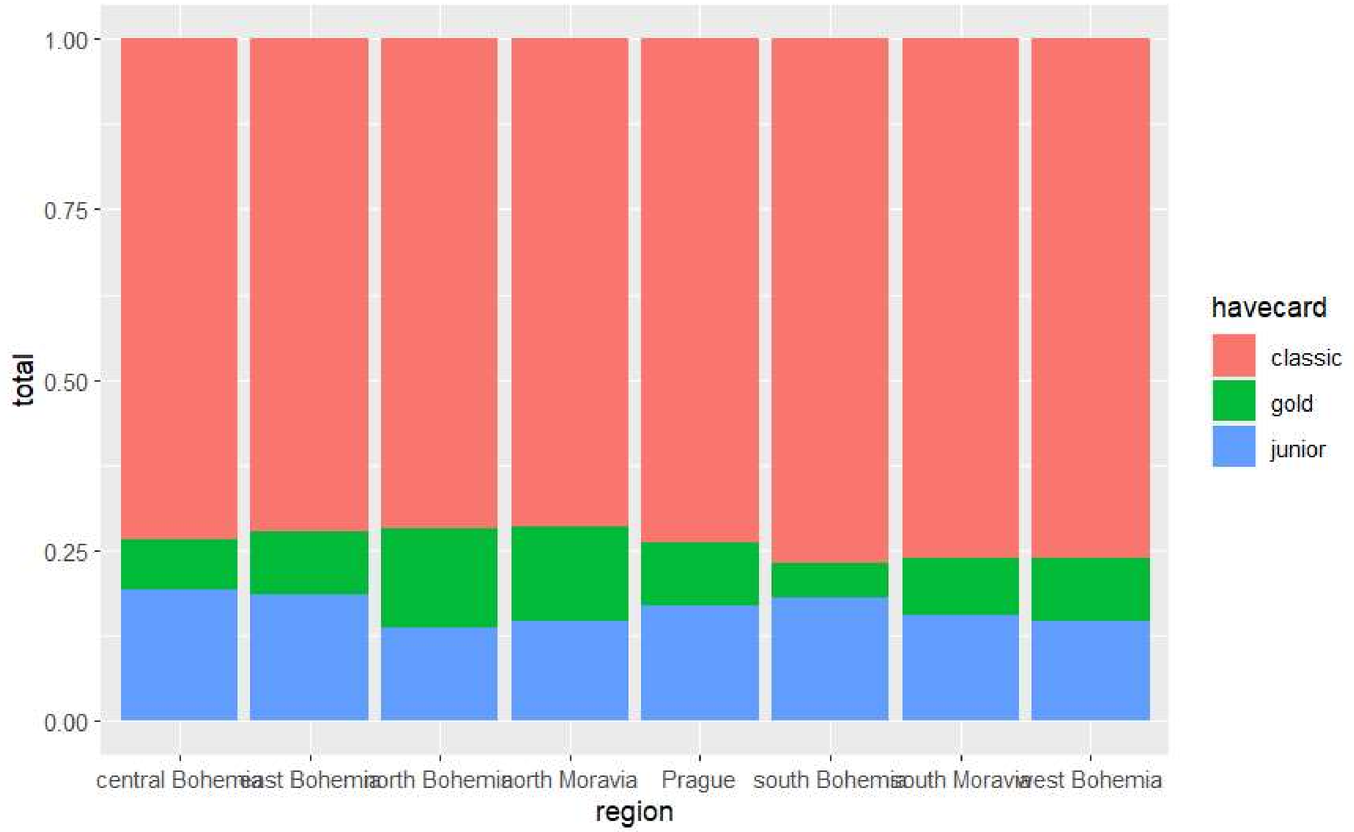
```
`summarise()` has grouped output by 'region'. You can override using the `.groups` argument.
```



Hide

```
client_card_all %>% filter(havecard == 'classic' | havecard == 'gold' | havecard == 'junior')
%>% mutate(contagem = 1) %>%
  group_by(region, havecard) %>% summarise(total = sum(contagem)) %>% ggplot(aes(x = region,
y = total, fill = havecard)) + geom_bar(stat = "identity", position = 'fill')
```

```
`summarise()` has grouped output by 'region'. You can override using the `.groups` argument.
```



Hide

```
client_loan <- inner_join(cli_dist_disp, loan, by = 'account_id')

client_loan <- filter(client_loan, client_type == 'OWNER')

View(client_loan)
```

Hide

```
client_loan_all <- full_join(cli_dist_disp, loan, by = 'account_id')

client_loan_all <- filter(client_loan_all, client_type == 'OWNER') # Filtragem por titular po
r 'Owner', somente owners poder pedir Loan

client_loan_all <- mutate(client_loan_all, haveloan = if_else(is.na(loan_id), 'FALSE', 'TRU
E')) # Identificando que tem LOAN ID e quem não tem

client_loan_all <- mutate(client_loan_all, haveloan2 = if_else(is.na(loan_id), 'no loan', as.
character(status_descr))) # Identificando quem tem loan e qual o status e quem não tem loan

View(client_loan_all)
```

Hide

```
tb_account_client %>%
    left_join(card, by = 'disp_id') %>%
    left_join(loan, by = 'account_id') -> tb_account_client_card_loan

View(tb_account_client_card_loan)
```

Hide

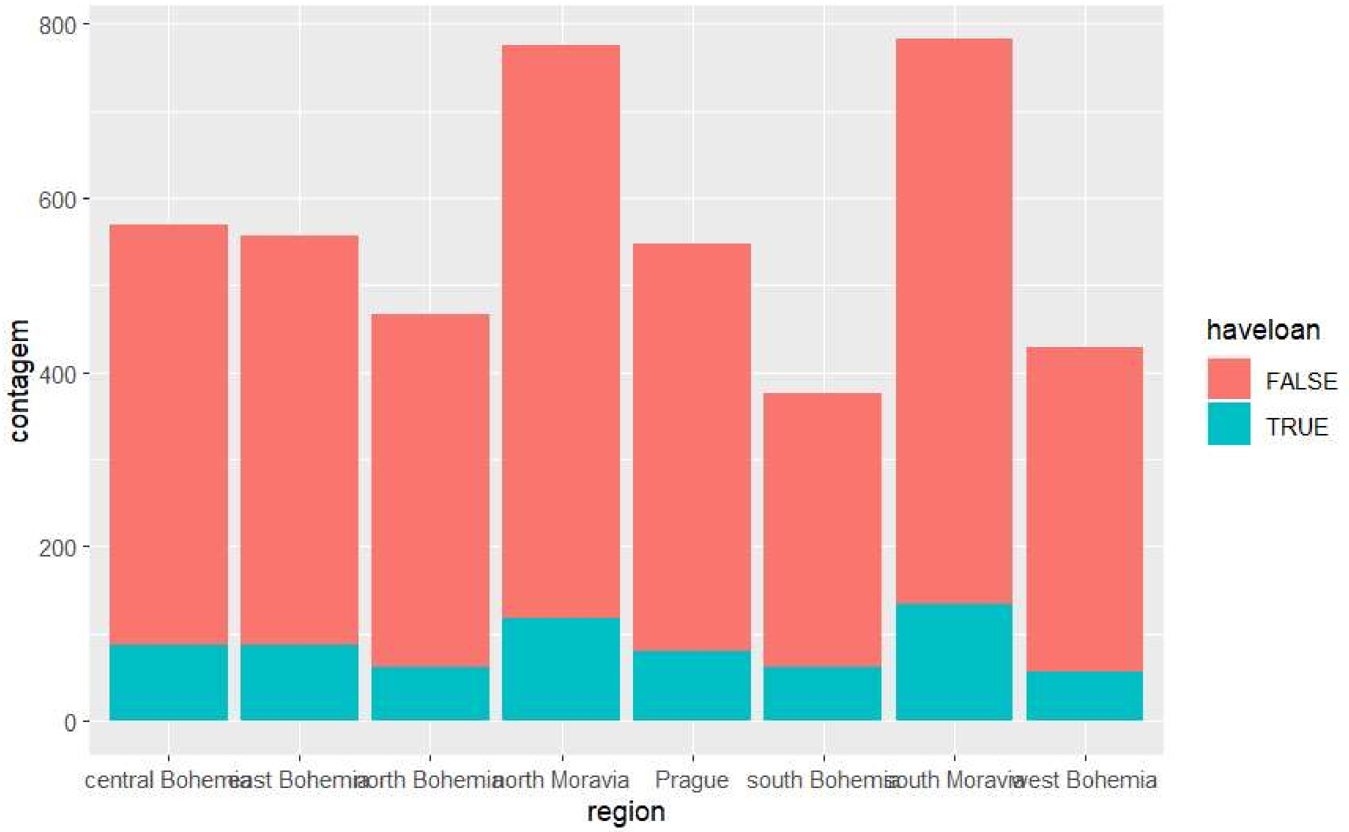
```
group_by (loan) %>%
  summarise (qtde_loan = n())
```

```
                                                        qtde_loan
                                                            <int>
                                                              682
1 row
```

Hide

```
client_loan_all %>% mutate(contagem = 1) %>% group_by(region, haveloan) %>% summarise(contage
m = sum(contagem)) %>%
  ggplot(aes(x = region, y = contagem, fill=haveloan)) + geom_bar(stat = "identity")
```

```
`summarise()` has grouped output by 'region'. You can override using the `.groups` argument.
```



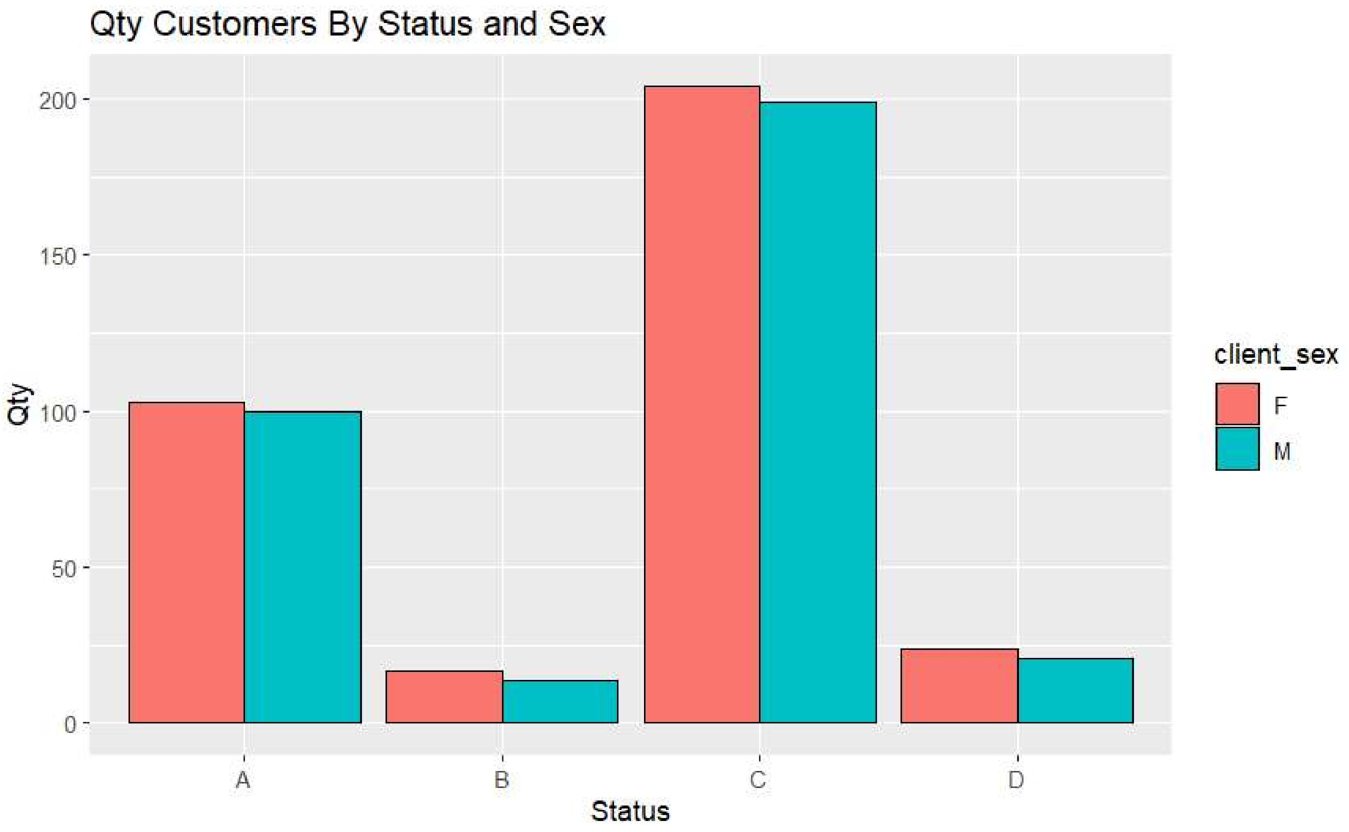
Hide

```
summary(loan)
```

```
    loan_id      account_id      date_loan              amount          duration
 Min.   :4959   Min.   :    2   Min.   :1993-07-05   Min.   :  4980   Min.   :12.00
 1st Qu.:5578   1st Qu.: 2967   1st Qu.:1995-07-04   1st Qu.: 66732   1st Qu.:24.00
 Median :6176   Median : 5738   Median :1997-02-06   Median :116928   Median :36.00
 Mean   :6172   Mean   : 5824   Mean   :1996-09-29   Mean   :151410   Mean   :36.49
 3rd Qu.:6752   3rd Qu.: 8686   3rd Qu.:1997-12-12   3rd Qu.:210654   3rd Qu.:48.00
 Max.   :7308   Max.   :11362   Max.   :1998-12-08   Max.   :590820   Max.   :60.00
   payments            status                                       status_descr
 Length:682         Length:682         A. Contract Finished, no problems   :203
 Class :character   Class :character   B. Contract Finished, Loan not Payed: 31
 Mode  :character   Mode  :character   C. Running Contract, OK so far      :403
                                       D. Running Contract, Client in Debt : 45
```

Hide

```
ggplot (data = filter(tb_account_client_card_loan,
                     client_type == 'OWNER' &
                     !is.na(loan_id == FALSE)),
         aes(x = status)) +
 geom_bar (mapping =  aes (fill = client_sex),
           position = 'dodge' ,
           color = 'black') +
 ggtitle('Qty Customers By Status and Sex') +
 xlab('Status') +
 ylab('Qty')
```



Hide

```
filter(tb_account_client_card_loan, client_type == 'OWNER' & !is.na(loan_id == TRUE)) %>%
   group_by(status, client_sex) %>%
   summarise(qtde = n())
```

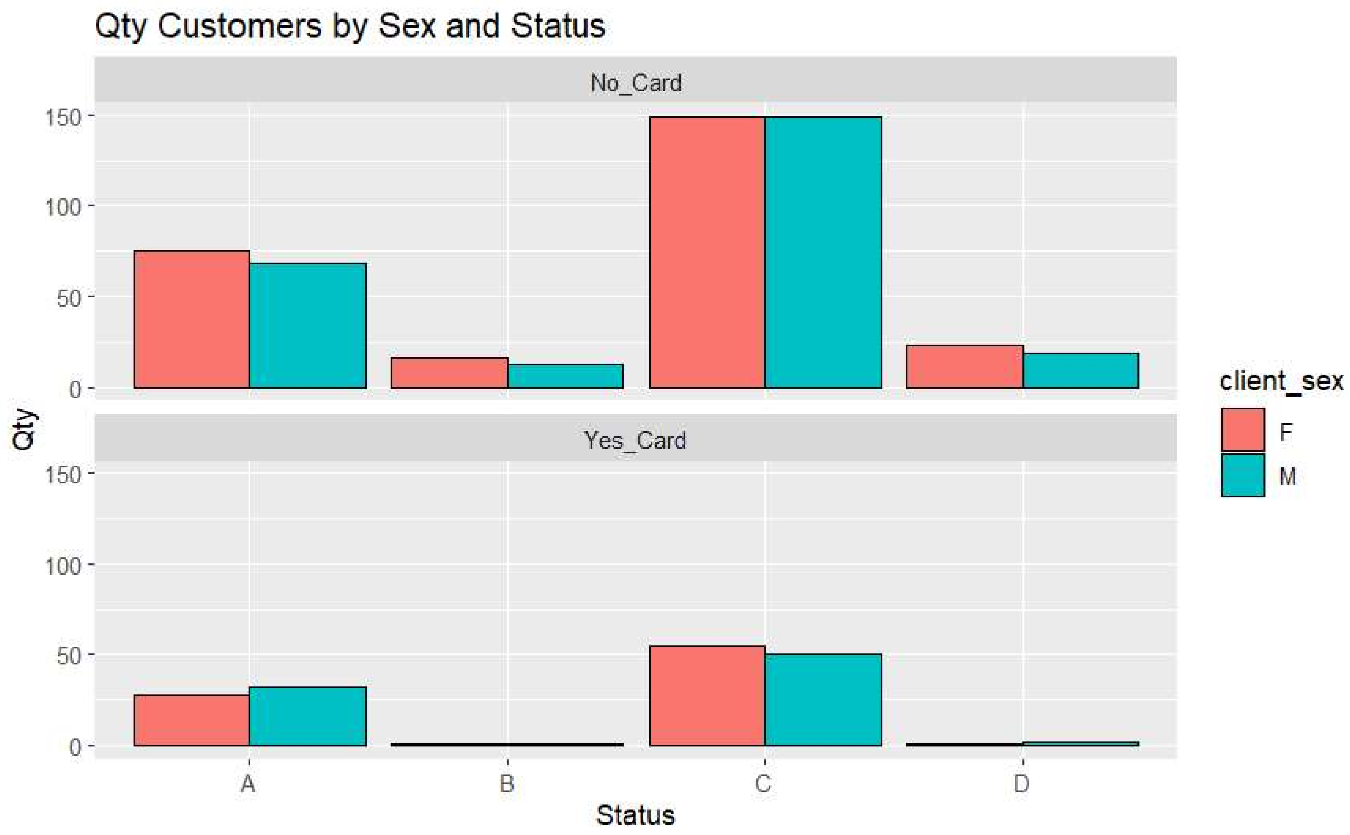
```
`summarise()` has grouped output by 'status'. You can override using the `.groups` argument.
```

status <chr>	client_sex <chr>	qtde <int>
A	F	103
A	M	100
B	F	17
B	M	14
C	F	204
C	M	199
D	F	24
D	M	21

8 rows

Hide

```
  mutate(tb_account_client_card_loan,
         status_card = ifelse (is.na(card_id) == TRUE,
                               'No_Card', 'Yes_Card')) -> tb_account_client_card_loan
```



Hide

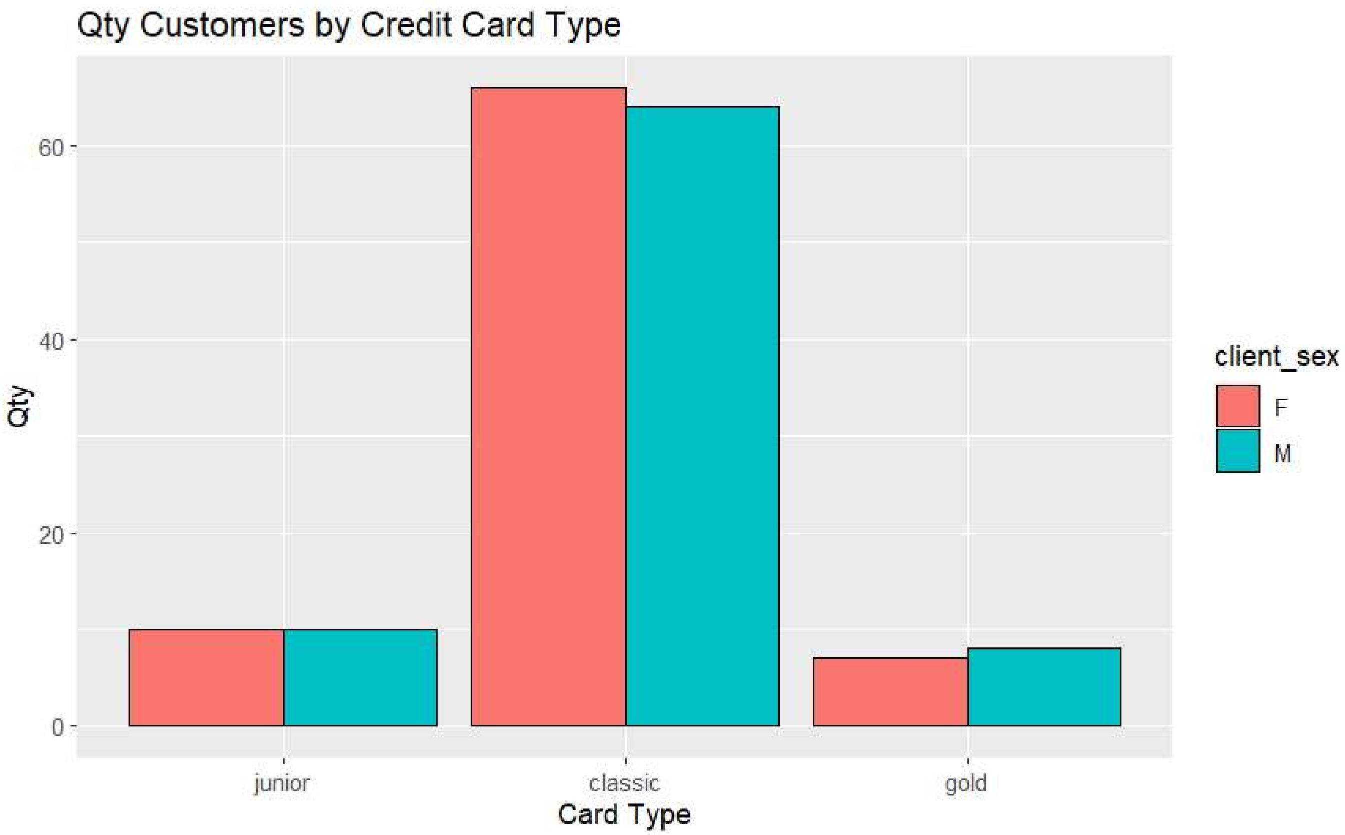
```
filter(tb_account_client_card_loan, status %in% c('A','C'))  %>%
  group_by (status_card) %>%
  summarise (qtde = n())
```

status_card <chr>	qtde <int>
No_Card	586
Yes_Card	165

2 rows

Hide

```
ggplot (data = filter (tb_account_client_card_loan, status %in% c('A','C') &
                         status_card == 'Yes_Card'),
        aes(x = card_type)) +
  geom_bar (mapping =  aes (fill = client_sex),
            position = 'dodge',
            color = 'black') +
  ggtitle('Qty Customers by Credit Card Type') +
  xlab('Card Type')+
  ylab('Qty')
```



Hide

```
  filter(tb_account_client_card_loan, status %in% c('A','C') & status_card == 'Yes_Card')  %
>%
  group_by (card_type) %>%
  summarise (qtde = n())
```

card_type <ord>	qtde <int>
junior	20
classic	130
gold	15

3 rows

Hide

```
filter(tb_account_client_card_loan, status %in% c('B','D'))  %>%
  group_by (status_card) %>%
  summarise (qtde = n())
```

status_card <chr>	qtde <int>
No_Card	71
Yes_Card	5

2 rows

Hide

```
loan %>%
  filter(status == 'B') %>%
  mutate(date_end = date_loan + months(duration)) -> rec_loan

View(rec_loan)

str(transaction)
```

```
tibble [1,056,320 × 10] (S3: tbl_df/tbl/data.frame)
 $ trans_id  : int [1:1056320] 695247 171812 207264 1117247 579373 771035 452728 725751 49721
1 232960 ...
 $ account_id: int [1:1056320] 2378 576 704 3818 1972 2632 1539 2484 1695 793 ...
 $ date_trans: Date[1:1056320], format: "1993-01-01" "1993-01-01" ...
 $ type      : chr [1:1056320] "Credit" "Credit" "Credit" "Credit" ...
 $ operation : chr [1:1056320] "Credit in Cash" "Credit in Cash" "Credit in Cash" "Credit in
Cash" ...
 $ amount    : num [1:1056320] 700 900 1000 600 400 1100 600 1100 200 800 ...
 $ balance   : chr [1:1056320] "700.00" "900.00" "1000.00" "600.00" ...
 $ bank      : chr [1:1056320] "" "" "" "" ...
 $ account   : int [1:1056320] NA NA NA NA NA NA NA NA NA NA ...
 $ tp_payment: chr [1:1056320] "" "" "" "" ...
```

Hide

```
transaction$amount <- as.numeric(transaction$amount)

transaction %>%
  filter(tp_payment == 'Loan Payment') %>%
  group_by(account_id) %>%
  summarise(total_payed = sum(amount)) -> loan_payment
View(loan_payment)

transaction %>%
  filter(tp_payment == 'Loan Payment') %>%
  group_by(account_id) %>%
  count(account_id) %>%
  rename(parc_payed = n)-> loan_qntd

transaction %>%
  filter(tp_payment == 'Loan Payment') %>%
  group_by(account_id,amount) %>%
  count(account_id) %>%
  select(-n ) %>%
  rename(parc= amount)-> loan_parc

rec_loan %>%
  left_join(loan_payment, by = 'account_id') %>%
  left_join(loan_qntd, by = 'account_id') %>%
  left_join(loan_parc, by = 'account_id') %>%
  mutate(parc_overdue = duration - parc_payed ) %>%
  mutate(value_overdue = parc_overdue * parc) %>%
  arrange(desc(value_overdue)) -> rec_loan

rec_loan %>% group_by(account_id) %>% summarise(date_end_loan = max(date_end)) -> clients_loa
n

transaction %>%
  inner_join(clients_loan, by = 'account_id') %>%
  filter(date_trans >= date_end_loan) %>%
  group_by(account_id,type) %>%
  summarise(value = sum(amount)) %>%
  spread(key = type, value = value) %>%
  mutate(total_after = Credit - Withdrawal) -> values_after
```

```
`summarise()` has grouped output by 'account_id'. You can override using the `.groups` argume
nt.
```

Hide

```
rec_loan %>%
  left_join(values_after, by = 'account_id') %>%
  mutate(analise_1 = if_else(total_after >= value_overdue,"Can Payment","Can't Pay")) %>%
  mutate(analise_2 = if_else(total_after >= 0,"Can Pay","Can't Pay")) -> rec_loan

max <- rec_loan

rec_loan %>%
  group_by(analise_2) %>%
  summarise(max_value = sum(value_overdue))
```

analise_2 <chr>	max_value <dbl>
Can't Pay	443993.3
Can Pay	567835.5

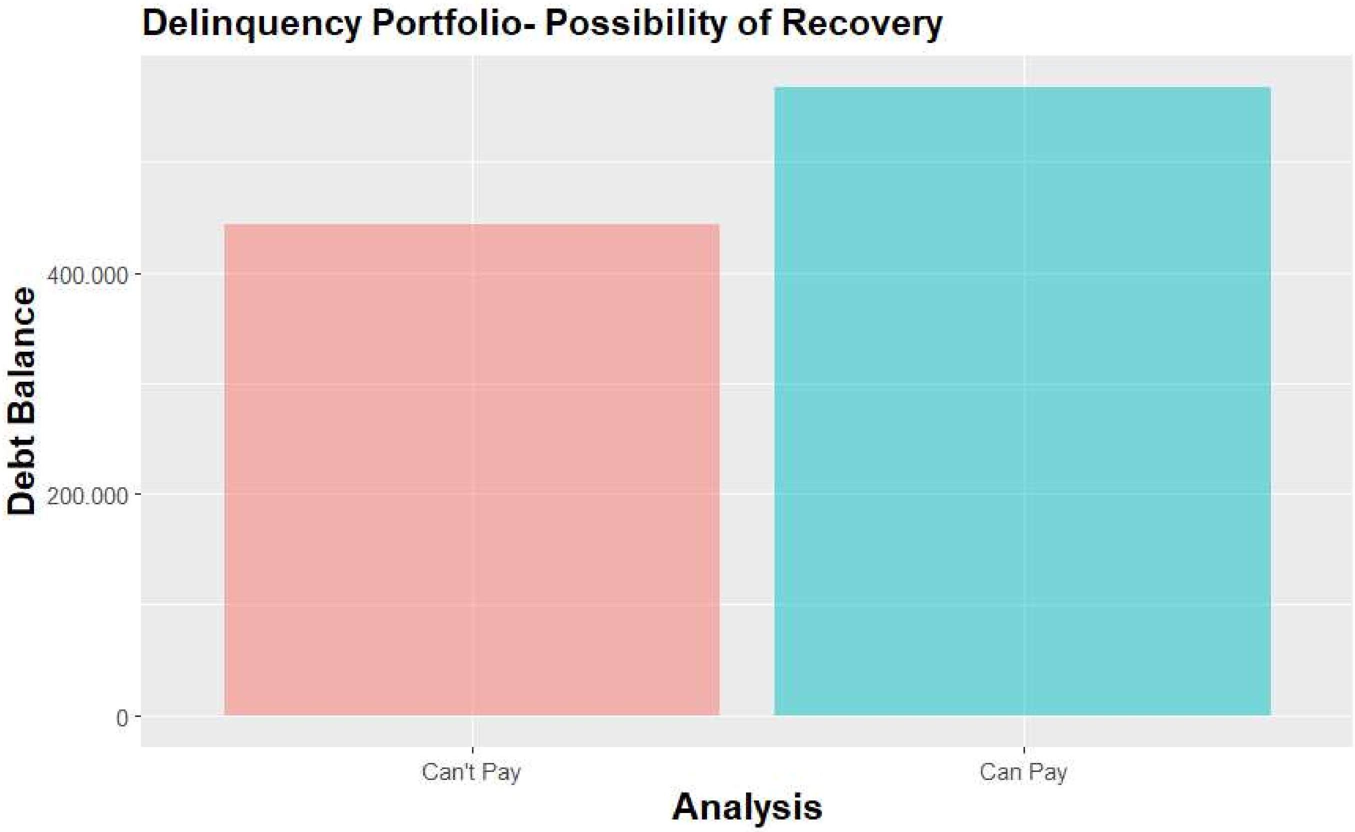
2 rows

Hide

```
NA
NA
NA
```

Hide

```
rec_loan %>%
  group_by(analise_2) %>%
  ggplot(mapping = aes(x = analise_2, y = value_overdue, fill = analise_2)) +
  geom_bar(alpha = 1/2, stat = "identity", show.legend = FALSE) +
  scale_y_continuous(labels = comma_format(big.mark = ".",
                                           decimal.mark = ",")) +
  labs(title = "Delinquency Portfolio- Possibility of Recovery",
       x = "Analysis",
       y = "Debt Balance",
       subtitle = NULL) +
  theme(plot.title = element_text(size=14, face="bold"),
        axis.title.x = element_text(size=14, face="bold"),
        axis.title.y = element_text(size=14, face="bold"))
```


Delinquency Portfolio- Possibility of Recovery
Debt Balance
400.000
200.000
0
Can't Pay
Can Pay
Analysis