car_insurance

416623

2025-08-21

Data source: https://www.kaggle.com/datasets/sagnik1511/car-insurance-data/data

Context The company has shared its annual car insurance data. Now, you have to find out the real customer behaviors over the data.

The work will focus on exploring the data (EDA) and then using popular machine learning models to predict the occurrence of a claim. We will assume that explanatory variables translate into the likelihood of claims occurring.

Basic first steps

```
## 'data.frame':    10000 obs. of  19 variables:
##  $ ID                 : int  569520 750365 199901 478866 731664 877557 930134 461006 68366 445911 ...
##  $ AGE                : chr  "65+" "16-25" "16-25" "16-25" ...
##  $ GENDER             : chr  "female" "male" "female" "male" ...
##  $ RACE               : chr  "majority" "majority" "majority" "majority" ...
##  $ DRIVING_EXPERIENCE : chr  "0-9y" "0-9y" "0-9y" "0-9y" ...
##  $ EDUCATION          : chr  "high school" "none" "high school" "university" ...
##  $ INCOME             : chr  "upper class" "poverty" "working class" "working class" ...
##  $ CREDIT_SCORE       : num  0.629 0.358 0.493 0.206 0.388 ...
##  $ VEHICLE_OWNERSHIP  : num  1 0 1 1 1 1 0 0 0 1 ...
##  $ VEHICLE_YEAR       : chr  "after 2015" "before 2015" "before 2015" "before 2015" ...
##  $ MARRIED            : num  0 0 0 0 0 0 1 0 1 0 ...
##  $ CHILDREN           : num  1 0 0 1 0 1 1 1 0 1 ...
##  $ POSTAL_CODE        : int  10238 10238 10238 32765 32765 10238 10238 10238 10238 32765 ...
##  $ ANNUAL_MILEAGE     : num  12000 16000 11000 11000 12000 13000 13000 14000 13000 11000 ...
##  $ VEHICLE_TYPE       : chr  "sedan" "sedan" "sedan" "sedan" ...
##  $ SPEEDING_VIOLATIONS: int  0 0 0 0 2 3 7 0 0 0 ...
##  $ DUIS               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PAST_ACCIDENTS     : int  0 0 0 0 1 3 3 0 0 0 ...
##  $ OUTCOME            : num  0 1 0 0 1 0 0 1 0 1 ...
```

We can see our data has very clear structure. We will start with dropping some unnecesary column - it is definately *ID*, *POSTAL_CODE*. I will also drop *RACE* columns, as I don't want to discuss potential race effect on insurance result.

As remaining columns we get:

```
##  [1] "AGE"                 "GENDER"              "DRIVING_EXPERIENCE"
##  [4] "EDUCATION"           "INCOME"              "CREDIT_SCORE"
##  [7] "VEHICLE_OWNERSHIP"   "VEHICLE_YEAR"        "MARRIED"
## [10] "CHILDREN"            "ANNUAL_MILEAGE"      "VEHICLE_TYPE"
## [13] "SPEEDING_VIOLATIONS" "DUIS"                "PAST_ACCIDENTS"
## [16] "OUTCOME"
```

Pretty much self explaining, apart from one, it is DUIS - Driving Under the Influence (at least i'll assume that's it).

Before additional 'factoring' we will check for potential *NA* values.

```
na_values<-function(x){
  sum(as.vector(is.na(x)))
}

apply(df,2,na_values)
```

```
##                 AGE              GENDER  DRIVING_EXPERIENCE           EDUCATION
##                   0                   0                   0                   0
##              INCOME        CREDIT_SCORE   VEHICLE_OWNERSHIP        VEHICLE_YEAR
##                   0                 982                   0                   0
##             MARRIED            CHILDREN      ANNUAL_MILEAGE        VEHICLE_TYPE
##                   0                   0                 957                   0
## SPEEDING_VIOLATIONS                DUIS      PAST_ACCIDENTS             OUTCOME
##                   0                   0                   0                   0
```

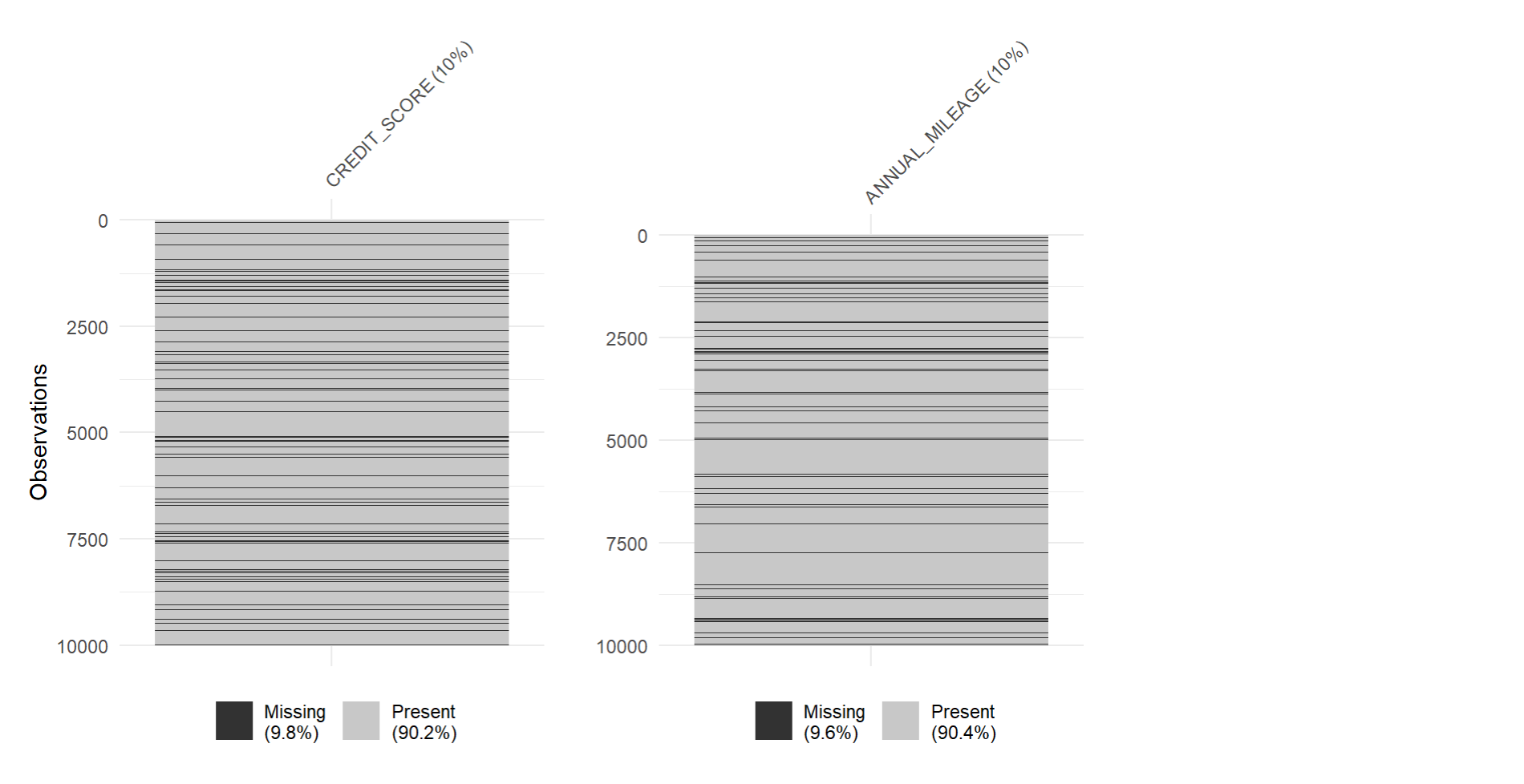
We can see some lacks in two variables, it is CREDIT_SCORE and ANNUAL_MILEAGE. Both variables are numerical (among many categorical variables). For now will skip this problem - yet we will see how these 'NA's correlate.

Amount of observations where both variables are NA:

```
nrow(df%>%dplyr::filter(is.na(CREDIT_SCORE) & is.na(ANNUAL_MILEAGE)))
```

```
## [1] 88
```

Lacks are pretty separated.



Now we will move to potential factoring in categorical (for now character) variables. At the same time will check their distribution in the data.

```
## AGE
## 16-25 26-39 40-64   65+
##  2016  3063  2931  1990
## GENDER
## female   male
##   5010   4990
## DRIVING_EXPERIENCE
##   0-9y 10-19y 20-29y   30y+
##   3530   3299   2119   1052
## EDUCATION
## high school        none  university
##        4157        1915        3928
## INCOME
##  middle class       poverty   upper class working class
##          2138          1814          4336          1712
## VEHICLE_YEAR
##  after 2015 before 2015
##        3033        6967
## VEHICLE_TYPE
##      sedan sports car
##       9523        477
```

Fortunately, we do not see any bigger disproportions, apart from last one indicating type of car. Nothing unexpected - sports cars are rare, even under insurance cooperations.

We'll factor some variables (it is not necesarry, but we'll do it)

```
df$AGE = factor(df$AGE,levels=c("16-25","26-39","40-64","65+"),ordered=TRUE)
df$DRIVING_EXPERIENCE=factor(df$DRIVING_EXPERIENCE, levels=c("0-9y","10-19y","20-29y","30y+"),ordered = TRUE)
df$EDUCATION=factor(df$EDUCATION,levels=c("none","high school","university"),ordered=TRUE)
df$INCOME=factor(df$INCOME, c("poverty","working class","middle class","upper class"),ordered=TRUE)
```

For now it will be enough.

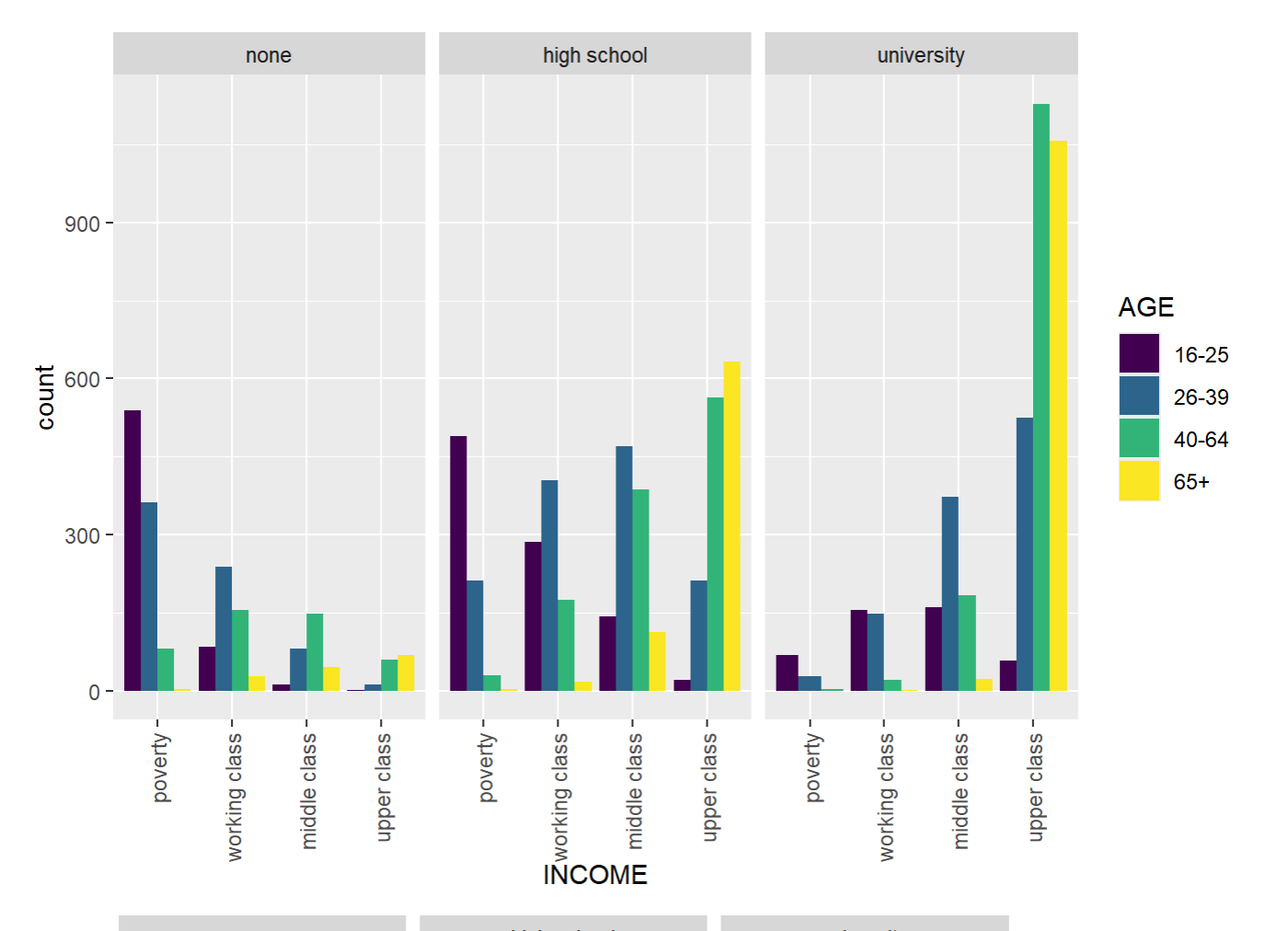
Analysis

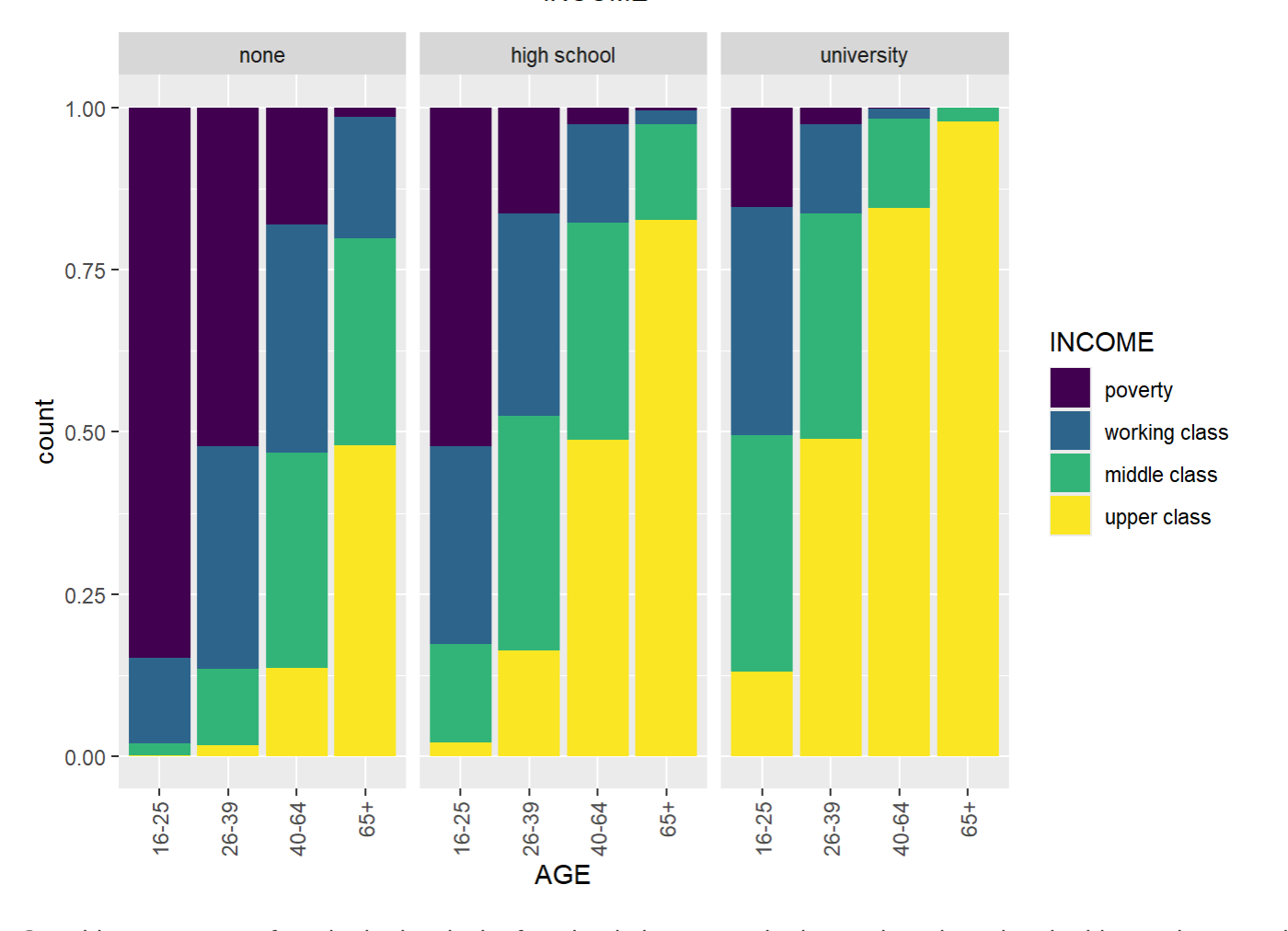
As at the end the work we are going to perform prediction - analysis won't be very long (especially if we have many columns). We will try to get some bigger knowledge about insurance company clients.

We can see some very regular structure in terms of ages and genders of insurance company clients. As to be predicted we can see more middle-aged people in comparison to youths and elderly.

```
##        DRIVING_EXPERIENCE
## AGE     0-9y 10-19y 20-29y 30y+
##   16-25 2016      0      0    0
##   26-39  683   2380      0    0
##   40-64  533    587   1811    0
##   65+    298    332    308 1052
```

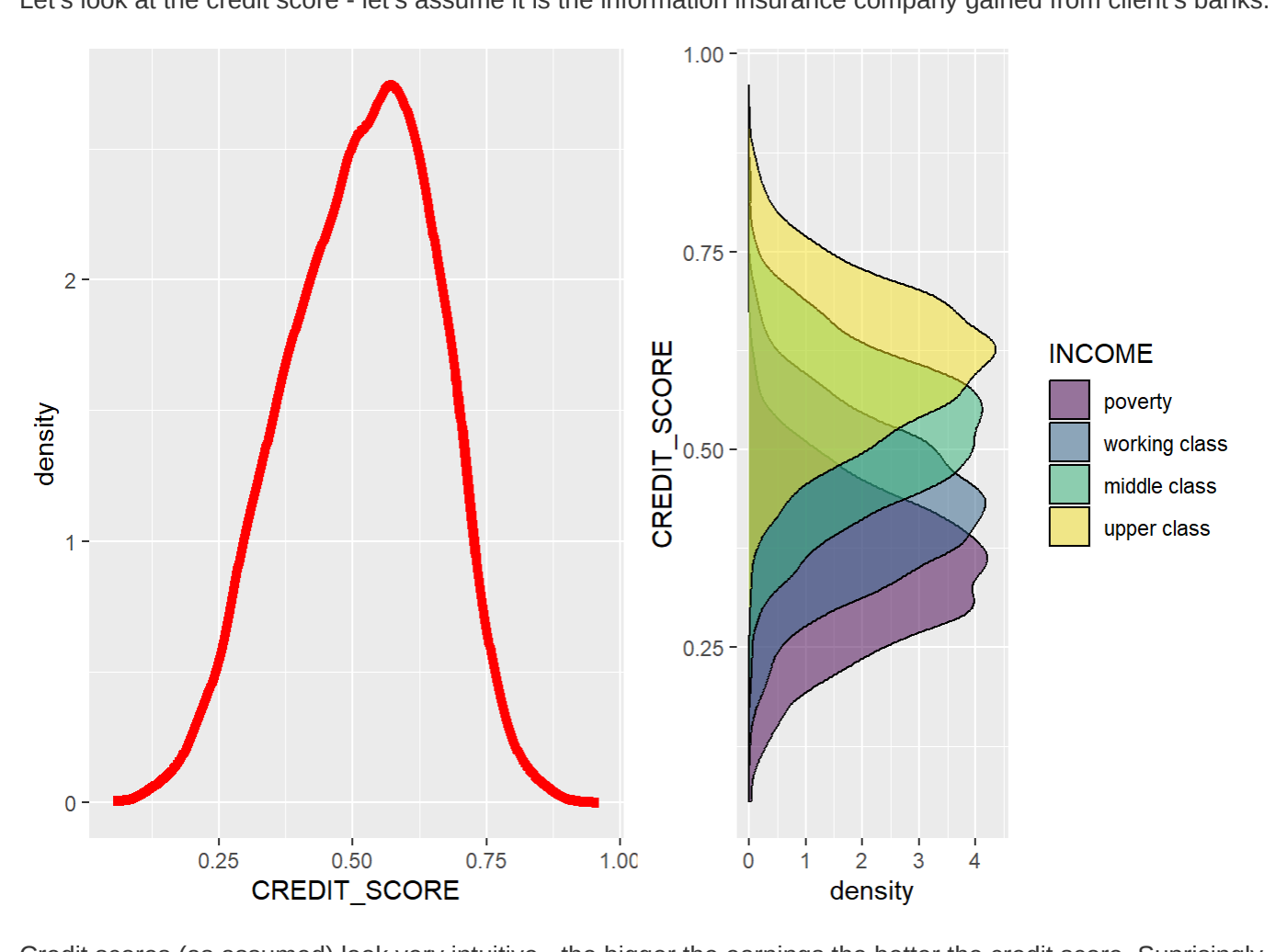
As natural the older the clients the bigger experience they may get. Interestingly we can see some significant amount of 65+ who got their experience lately.





One thing we can get from both plots is the fact that in long-term the better the eductation the bigger chance to be in better class. Intuitively it seems natural.

Let's look at the credit score - let's assume it is the information insurance company gained from client's banks.



Credit scores (as assumed) look very intuitive - the bigger the earnings the better the credit score. Suprisingly - distributions look very gaussian.

If it comes to variable VEHICLE_OWNERSHIP - it may be indicator of some other person behind a client - moral hazard things. We may see some correlation between ownership and outcome (dependencies with Y variable will be discussed later).

1 indicates ownership, 0 otherwise.

```
##
##    0    1
## 3030 6970
```

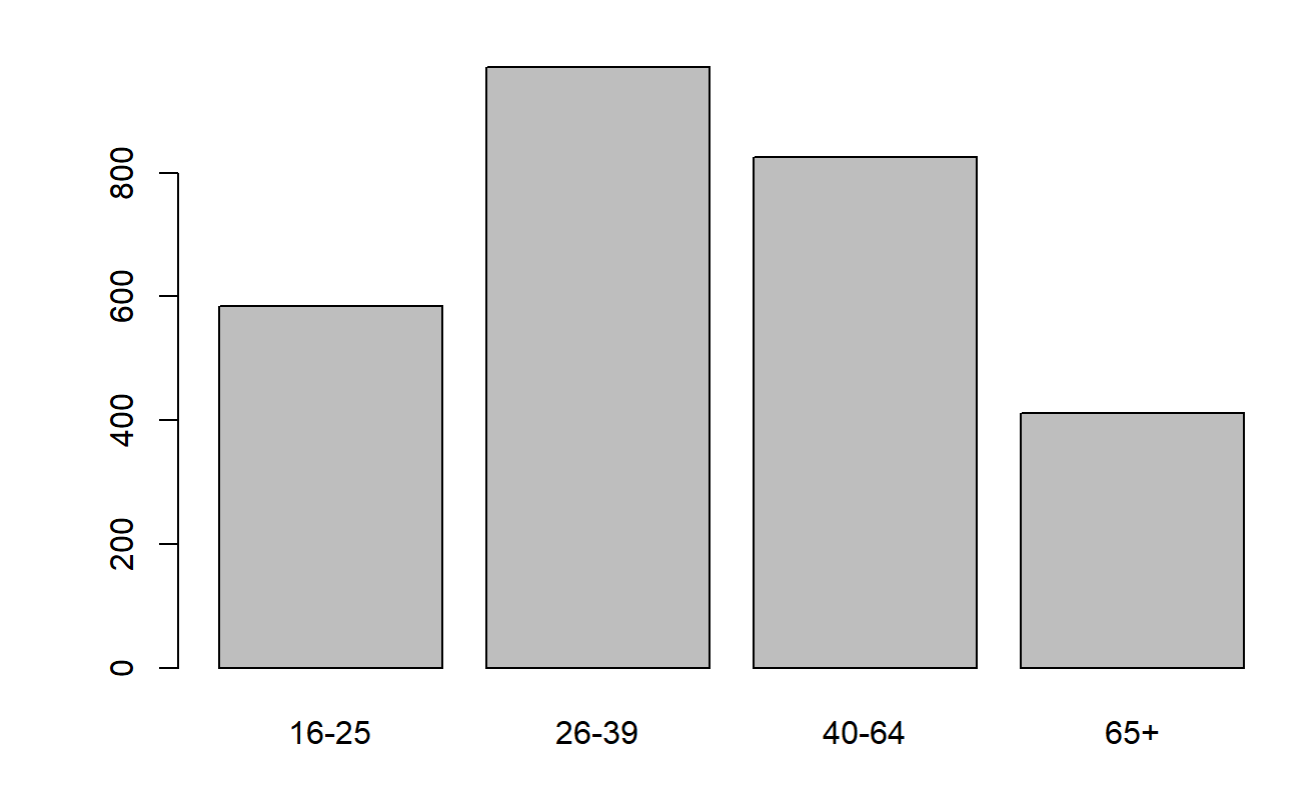
Let's look at numbers about marriage and children. First variable: 1 indicates client being married, 0 otherwise. Second variable: 1 indicates having children.

```
table(df[c("MARRIED","CHILDREN")])
```

```
##        CHILDREN
## MARRIED    0    1
##       0 2226 2792
##       1  886 4096
```

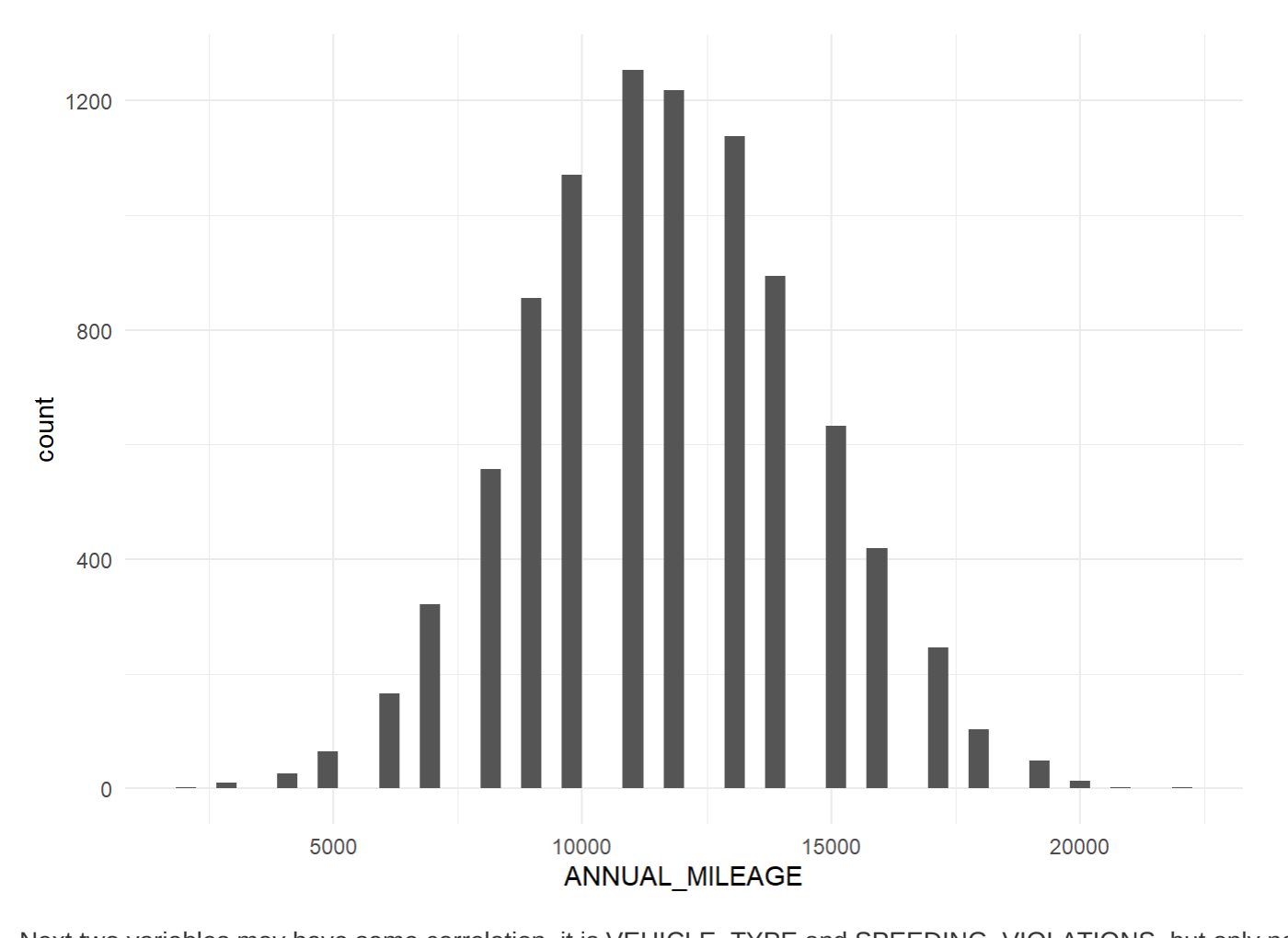
Suprisingly we can see many not married with children clients. Let's look at age distribution for those people.

```
barplot(table(df%>%filter(MARRIED==0 & CHILDREN==1)%>%
  dplyr::select(AGE)))
```



It looks like it isn't that much connected to the fact that some people may be divorced. Very interesting. Another story if MARRIED in our data is an indicator of legal marriage or just having a partner for very long time - if the second one is the correct one - then interpretation for the data may differ.

We will interpret ANNUAL_MILEAGE as of how much client drives the hole year - doesn't matter if insured car is just for commute or for work-wise purpose (indicator of how much does client drives).



Next two variables may have some correlation, it is VEHICLE_TYPE and SPEEDING_VIOLATIONS, but only naively. Let's look at the numbers.

```
divide_by_sum<-function(x){ x/sum(x)}
v<-table(df$VEHICLE_TYPE, df$SPEEDING_VIOLATIONS)
probs<-apply(v,1,divide_by_sum)
knitr::kable(data.frame(probs)%>%rownames_to_column(var="amount_of_speed_val"),align = "c",caption = "Frequencies of speed_val for given car")
```

Frequencies of speed_val for given car

amount_of_speed_val	sedan	sports.car
0	0.5039378	0.4800839
1	0.1533130	0.1761006
2	0.1155098	0.1278826
3	0.0829571	0.0838574
4	0.0528195	0.0566038
5	0.0324478	0.0209644
6	0.0187966	0.0188679
7	0.0143862	0.0062893
8	0.0075606	0.0062893
9	0.0050404	0.0020964
10	0.0048304	0.0083857
11	0.0029402	0.0041929
12	0.0021002	0.0000000
13	0.0011551	0.0020964
14	0.0004200	0.0020964
15	0.0008401	0.0000000
16	0.0003150	0.0020964
17	0.0003150	0.0000000
18	0.0001050	0.0000000
19	0.0001050	0.0020964
22	0.0001050	0.0000000

We can't see any specific differences.

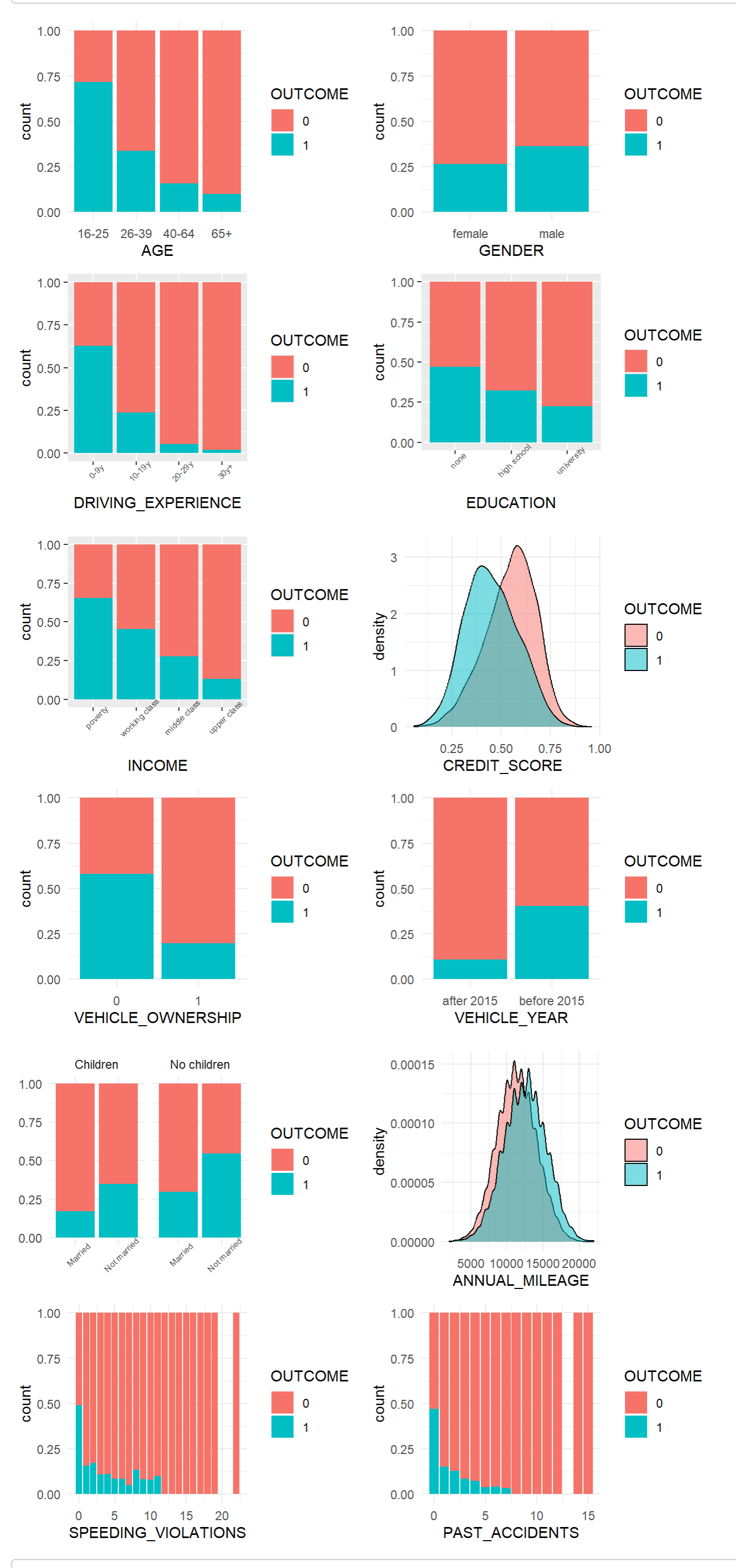
We will skip last two variables for Y variable comparison in next chapter.

Outcome variable

Value count on OUTCOME after dropping NA records:

```
table(drop_na(df)$OUTCOME)
```

```
##
##    0    1
## 5613 2536
```




```
##                 OUTCOME
## PAST_ACCIDENTS     0    1
##              0  2961 2623
##              1  1515  268
##              2   961  143
##              3   591   55
##              4   339   27
##              5   223    9
##              6   138    6
##              7    59    2
##              8    41    0
##              9    19    0
##             10     9    0
##             11     7    0
##             12     2    0
##             14     1    0
##             15     1    0
```

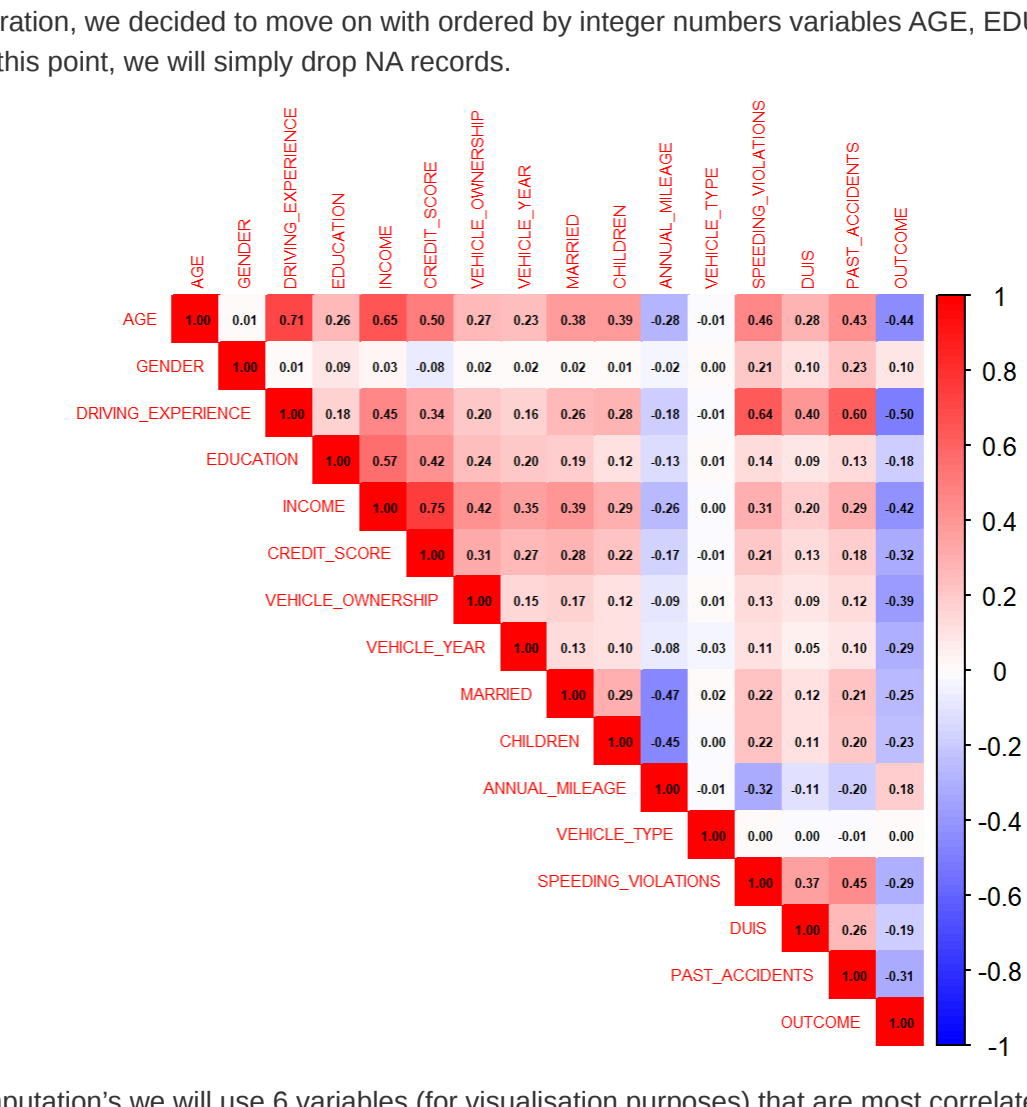
Results if it comes to future claims looks clear in most variables. The more experienced driver - the less chance of claim; on average the higher the education - we get less claims; if the car is not owned by the client - chances of claims rises; older cars correlate with claims; people with children drive less risky, more confidentally; little difference in annual_mileage for claims and not claims observations; legal violations do not correlate positively very much with outcome variable - suprisingly the more of them, the less frequent claims get.

PCA

To sum up, we will drop our variables to components. PCA is the statistical technique that reduces about of dimentions by creating synthetic variables with linear combinations of already existing ones. Before that we need to do some preprocessing like numerical ordering or one-hot encoding.

```
df1$GENDER<-recode(df1$GENDER,'female'=0, 'male'=1)
levels(df1$DRIVING_EXPERIENCE)<-c(1:4)
df1$DRIVING_EXPERIENCE <- as.double(as.character(df1$DRIVING_EXPERIENCE))
levels(df1$AGE)<-c(1:4)
df1$AGE <- as.double(as.character(df1$AGE))
levels(df1$EDUCATION)<-c(1:3)
df1$EDUCATION <- as.double(as.character(df1$EDUCATION))
levels(df1$INCOME)<-c(1:4)
df1$INCOME <- as.double(as.character(df1$INCOME))
df1$VEHICLE_YEAR<-recode(df1$VEHICLE_YEAR,'after 2015'=1, 'before 2015'=0)
df1$VEHICLE_TYPE<-recode(df1$VEHICLE_TYPE,'sports car'=1, 'sedan'=0)
```

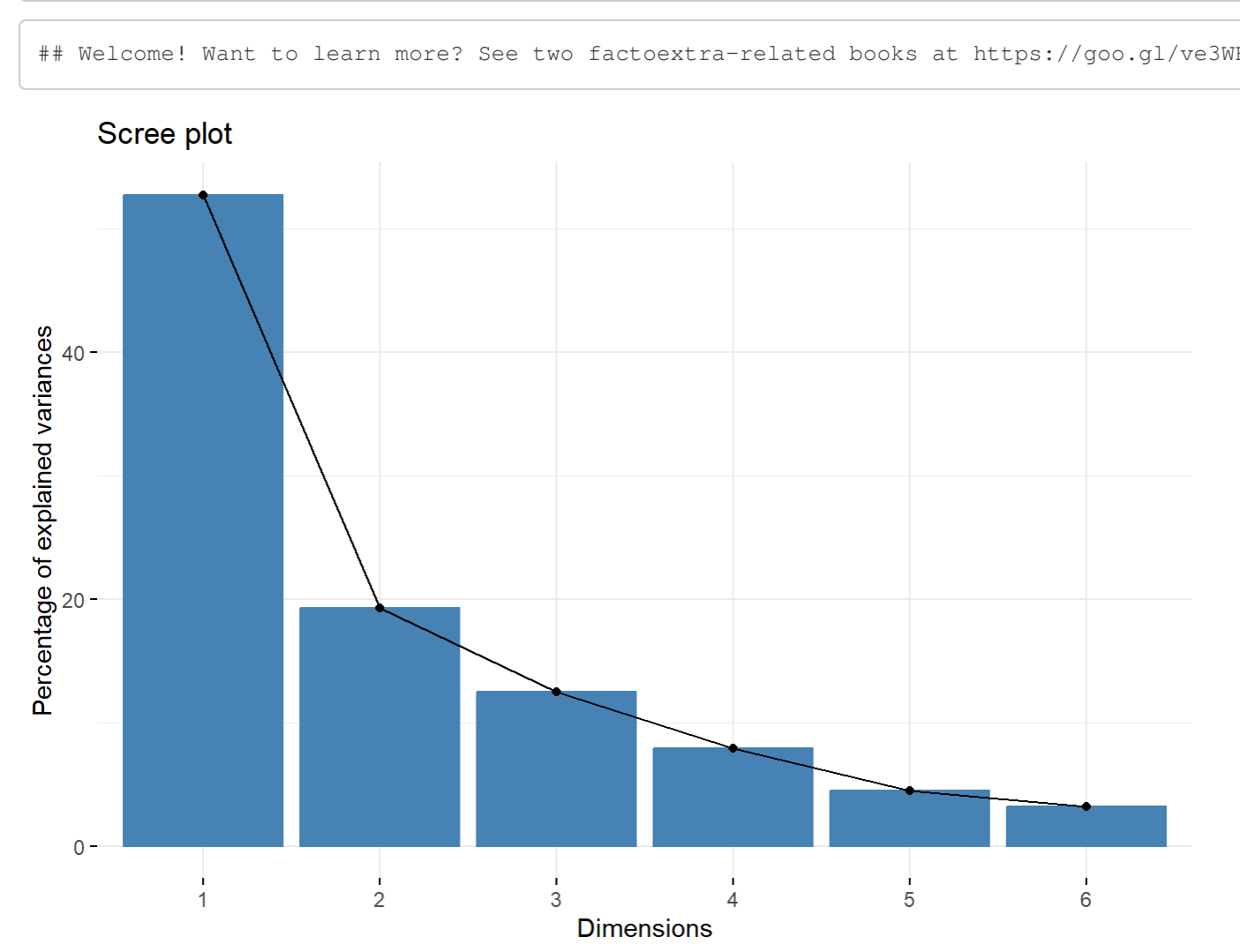
After consideration, we decided to move on with ordered by integer numbers variables AGE, EDUCATION, DRIVING_EXPERIENCE and INCOME. At this point, we will simply drop NA records.

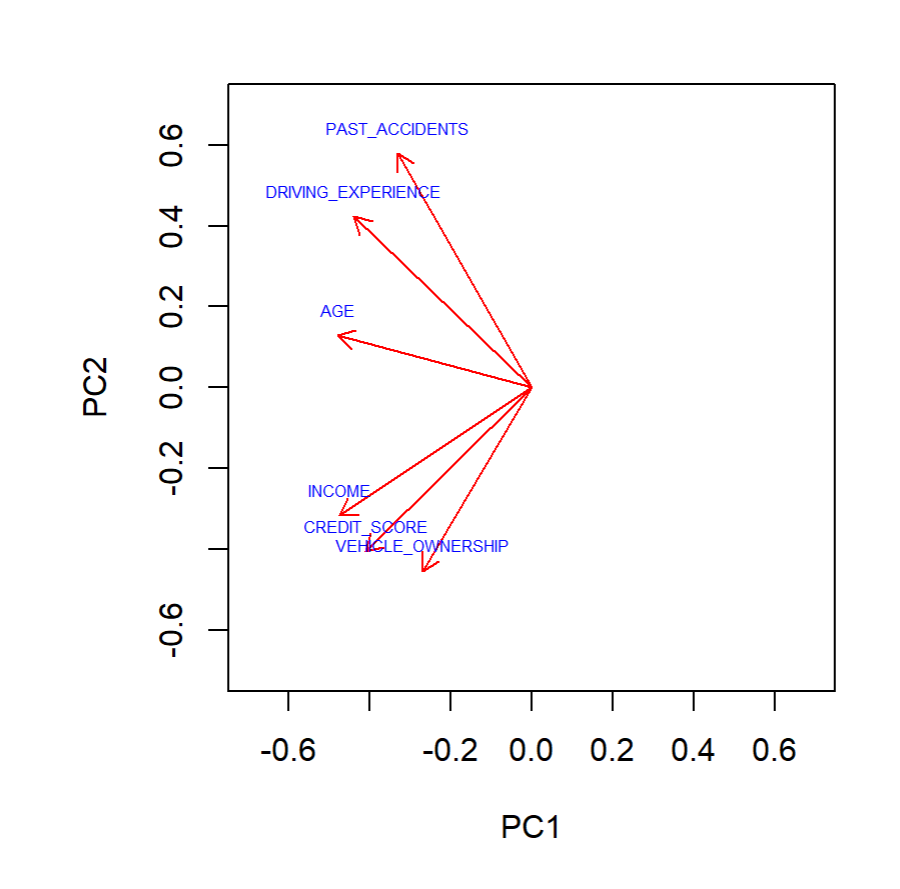


For PCA computation's we will use 6 variables (for visualisation purposes) that are most correlated with Y, it is: AGE, DRIVING_EXPERIENCE, INCOME, VEHICLE_OWNERSHIP, CREDIT_SCORE and PAST_ACCIDENTS (even though initial analysis did not find much of effect). We need to take into account the fact of many 0/1 variables - they are not very convenient for linear purposes. For PCA - we scale the data, so they are mean=0 and sd=1.

```
pca<-prcomp(subset(df1, select = c(AGE, DRIVING_EXPERIENCE, INCOME, VEHICLE_OWNERSHIP, CREDIT_SCORE, PAST_ACCIDEN
TS)), center = TRUE, scale. = TRUE)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```





Biplot and scatterplot both speak for itself - variables ale negatively correlated with claim, therefore we can see more of Y=1 on the right side of the graph.