Module 1

Semih Barutcu

6/1/2020

0

I used tidyverse package family to analyze the data.

```
library(pacman)
p_load(tidyverse, lubridate, skimr, summarytools, autoEDA, visdat, C50)
```

I saved 3 datasets as csv from excel spreadsheets source.

```
transactions <- read.csv("Transactions.csv", header = T, skip = 1)

cdemographics <- read.csv("CustomerDemographic.csv", header = T, skip = 1)

caddress <- read.csv("CustomerAddress.csv", header = T, skip = 1)

newcustomer <- read.csv("NewCustomerList.csv", header = T, skip = 1)
```

I arranged dates using lubridate package function mdy() to be able to use date features for my analyses.

```
transactions$transaction_date <- mdy(transactions$transaction_date)
cdemographics$DOB <- mdy(cdemographics$DOB)
newcustomer$DOB <- mdy(newcustomer$DOB)
```

Changing chr(character) variables to factor is applied using lapply() fuction after I listed these columns. Categorical data is much more useful to explore the data. I also removed "$" sign from standard_cost variable to be able to get proper statistics as numeric values.

```
cols1 <- c("order_status", "brand", "product_line", "product_class", "product_size", "standard_cost")
transactions[cols1] <- lapply(transactions[cols1], factor)

cols2 <- c("gender","job_title", "job_industry_category", "wealth_segment", "deceased_indicator", "owns_
cdemographics[cols2] <- lapply(cdemographics[cols2], factor)

cols3 <- c("address", "postcode","state", "country")
caddress[cols3] <- lapply(caddress[cols3], factor)

cols4 <- c("gender","job_title", "job_industry_category", "wealth_segment", "deceased_indicator", "owns_
newcustomer[cols4] <- lapply(newcustomer[cols4], factor)

# Nested gsub() function. First remove $ sign and after remove commas if exists
transactions$standard_cost <- as.numeric(gsub(",", "",gsub("\\$", "", transactions$standard_cost)))
```

1

All summary statistics are listed below.

All transactions were happened in 2017. 360 of the total 20000 transactions are missing online_order information. 179 of the orders were cancelled. 197 of the transactions are without a brand, product_line, product_class, product_size, standard_cost and product_first_sold_date.

3 of 4000 total observations are misidentified as F, Femal and M. There are 88 observations with gender U and 87 of observations do not have tenure information. 88 of customers do not have date of birth information. Job title is missing for 506 persons and job industry category is missing for 656.

New South Wales and Victoria states used with both full names and abbrevations. All 3999 address records are from Australia. 3 addresses are used for 2 times.

```r
summary(transactions)
```

```
##  transaction_id    product_id      customer_id     transaction_date    
##  Min.   :    1   Min.   :  0.00   Min.   :   1.0   Min.   :2017-01-01  
##  1st Qu.: 5001   1st Qu.: 18.00   1st Qu.: 857.8   1st Qu.:2017-04-01  
##  Median :10000   Median : 44.00   Median :1736.0   Median :2017-07-03  
##  Mean   :10000   Mean   : 45.36   Mean   :1738.2   Mean   :2017-07-01  
##  3rd Qu.:15000   3rd Qu.: 72.00   3rd Qu.:2613.0   3rd Qu.:2017-10-02  
##  Max.   :20000   Max.   :100.00   Max.   :5034.0   Max.   :2017-12-30  
##                                                                        
##  online_order       order_status              brand        product_line  
##  Mode :logical   Approved :19821                 : 197            :  197  
##  FALSE:9811      Cancelled:  179   Giant Bicycles:3312   Mountain:  423  
##  TRUE :9829                        Norco Bicycles:2910   Road    : 3970  
##  NA's :360                         OHM Cycles    :3043   Standard:14176  
##                                    Solex         :4253   Touring : 1234  
##                                    Trek Bicycles :2990                   
##                                    WeareA2B      :3295                   
##  product_class  product_size     list_price      standard_cost    
##        :  197         :  197   Min.   :  12.01   Min.   :   7.21  
##  high  : 3013   large : 3976   1st Qu.: 575.27   1st Qu.: 215.14  
##  low   : 2964   medium:12990   Median :1163.89   Median : 507.58  
##  medium:13826   small : 2837   Mean   :1107.83   Mean   : 556.05  
##                                3rd Qu.:1635.30   3rd Qu.: 795.10  
##                                Max.   :2091.47   Max.   :1759.85  
##                                                  NA's   :197      
##  product_first_sold_date
##  Min.   :33259          
##  1st Qu.:35667          
##  Median :38216          
##  Mean   :38200          
##  3rd Qu.:40672          
##  Max.   :42710          
##  NA's   :197            
```

```r
summary(cdemographics)
```

```
##   customer_id    first_name         last_name            gender    
##  Min.   :   1   Length:4000        Length:4000        F     :   1  
##  1st Qu.:1001   Class :character   Class :character   Femal :   1  
##  Median :2000   Mode  :character   Mode  :character   Female:2037  
##  Mean   :2000                                         M     :   1  
##  3rd Qu.:3000                                         Male  :1872  
##  Max.   :4000                                         U     :  88  
## 
```

```
##  past_3_years_bike_related_purchases      DOB
##  Min.   : 0.00                      Min.   :1931-10-23
##  1st Qu.:24.00                      1st Qu.:1968-01-25
##  Median :48.00                      Median :1977-07-25
##  Mean   :48.89                      Mean   :1977-07-25
##  3rd Qu.:73.00                      3rd Qu.:1987-02-28
##  Max.   :99.00                      Max.   :2002-03-11
##                                     NA's   :88
##                                job_title          job_industry_category
##                                     : 506   Manufacturing     :799
##  Business Systems Development Analyst:  45   Financial Services:774
##  Social Worker                       :  44   n/a               :656
##  Tax Accountant                      :  44   Health            :602
##  Internal Auditor                    :  42   Retail            :358
##  Legal Assistant                     :  41   Property          :267
##  (Other)                             :3278   (Other)           :544
##            wealth_segment deceased_indicator   default          owns_car
##  Affluent Customer: 979   N:3998            Length:4000        No :1976
##  High Net Worth   :1021   Y:   2            Class :character   Yes:2024
##  Mass Customer    :2000                     Mode  :character
##
##
##
##
##      tenure
##  Min.   : 1.00
##  1st Qu.: 6.00
##  Median :11.00
##  Mean   :10.66
##  3rd Qu.:15.00
##  Max.   :22.00
##  NA's   :87
```

```r
summary(caddress)
```

```
##   customer_id                      address        postcode
##  Min.   :   1   3 Mariners Cove Terrace:   2   2170   :  31
##  1st Qu.:1004   3 Talisman Place       :   2   2145   :  30
##  Median :2004   64 Macpherson Junction :   2   2155   :  30
##  Mean   :2004   0 3rd Road             :   1   2153   :  29
##  3rd Qu.:3004   0 American Ash Parkway :   1   2560   :  26
##  Max.   :4003   0 Arapahoe Court       :   1   2770   :  26
##                 (Other)                :3990   (Other):3827
##              state           country     property_valuation
##  New South Wales:  86   Australia:3999   Min.   : 1.000
##  NSW            :2054                    1st Qu.: 6.000
##  QLD            : 838                    Median : 8.000
##  VIC            : 939                    Mean   : 7.514
##  Victoria       :  82                    3rd Qu.:10.000
##                                          Max.   :12.000
##
```

```r
summary(newcustomer)
```

```
##   first_name         last_name            gender
```

```
##  Length:1000        Length:1000        Female:513
##  Class :character   Class :character   Male  :470
##  Mode  :character   Mode  :character   U     : 17
##
##
##
##
##  past_3_years_bike_related_purchases      DOB
##  Min.   : 0.00                       Min.   :1938-06-08
##  1st Qu.:26.75                       1st Qu.:1957-10-09
##  Median :51.00                       Median :1972-03-24
##  Mean   :49.84                       Mean   :1971-04-20
##  3rd Qu.:72.00                       3rd Qu.:1983-04-12
##  Max.   :99.00                       Max.   :2002-02-27
##                                      NA's   :17
##                 job_title          job_industry_category
##                      :106   Financial Services:203
##  Associate Professor  : 15   Manufacturing     :199
##  Environmental Tech   : 14   n/a               :165
##  Software Consultant  : 14   Health            :152
##  Chief Design Engineer: 13   Retail            : 78
##  Assistant Manager    : 12   Property          : 64
##  (Other)              :826   (Other)           :139
##           wealth_segment deceased_indicator owns_car      tenure
##  Affluent Customer:241    N:1000            No :507   Min.   : 0.00
##  High Net Worth   :251                      Yes:493   1st Qu.: 7.00
##  Mass Customer    :508                                Median :11.00
##                                                       Mean   :11.39
##                                                       3rd Qu.:15.00
##                                                       Max.   :22.00
##
##             address       postcode   state          country
##  0 Bay Drive     :  1   2145   :  9   NSW:506   Australia:1000
##  0 Dexter Parkway:  1   2232   :  9   QLD:228
##  0 Emmet Trail   :  1   2148   :  7   VIC:266
##  0 Esker Avenue  :  1   2168   :  7
##  0 Express Lane  :  1   2750   :  7
##  0 Kipling Way   :  1   3029   :  7
##  (Other)         :994   (Other):954
##  property_valuation       X               X.1              X.2
##  Min.   : 1.000     Min.   :0.4000   Min.   :0.4000   Min.   :0.4000
##  1st Qu.: 6.000     1st Qu.:0.5700   1st Qu.:0.6400   1st Qu.:0.7083
##  Median : 8.000     Median :0.7500   Median :0.8375   Median :0.9375
##  Mean   : 7.397     Mean   :0.7468   Mean   :0.8372   Mean   :0.9408
##  3rd Qu.: 9.000     3rd Qu.:0.9200   3rd Qu.:1.0100   3rd Qu.:1.1250
##  Max.   :12.000     Max.   :1.1000   Max.   :1.3750   Max.   :1.7188
##
##       X.3              X.4              Rank            Value
##  Min.   :0.3400   Min.   :   1.0   Min.   :   1.0   Min.   :0.3400
##  1st Qu.:0.6500   1st Qu.: 250.0   1st Qu.: 250.0   1st Qu.:0.6495
##  Median :0.8500   Median : 500.0   Median : 500.0   Median :0.8600
##  Mean   :0.8686   Mean   : 498.8   Mean   : 498.8   Mean   :0.8817
##  3rd Qu.:1.0600   3rd Qu.: 750.2   3rd Qu.: 750.2   3rd Qu.:1.0750
##  Max.   :1.7188   Max.   :1000.0   Max.   :1000.0   Max.   :1.7188
```

```
##
```

I checked addresses below which exists 2 times in the data. They have different postcodes and customer IDs.

```r
caddress %>% filter(address == "3 Mariners Cove Terrace")
```

```
##   customer_id                 address postcode state   country
## 1        2333 3 Mariners Cove Terrace     3108   VIC Australia
## 2        2985 3 Mariners Cove Terrace     2216   NSW Australia
##   property_valuation
## 1                 10
## 2                 10
```

```r
caddress %>% filter(address == "3 Talisman Place")
```

```
##   customer_id          address postcode state   country property_valuation
## 1         737 3 Talisman Place     4811   QLD Australia                  2
## 2        2475 3 Talisman Place     4017   QLD Australia                  5
```

```r
caddress %>% filter(address == "64 Macpherson Junction")
```

```
##   customer_id                address postcode state   country
## 1        2320 64 Macpherson Junction     2208   NSW Australia
## 2        3540 64 Macpherson Junction     4061   QLD Australia
##   property_valuation
## 1                 11
## 2                  8
```

Gender and state variables corrections have been made below. I used factor function to get corrected categories.

```r
cdemographics$gender[cdemographics$gender == "Femal" | cdemographics$gender == "F"] <- "Female"

cdemographics$gender[cdemographics$gender == "M"] <- "Male"

cdemographics$gender <- factor(cdemographics$gender)

caddress$state[caddress$state == "New South Wales"] <- "NSW"

caddress$state[caddress$state == "Victoria"] <- "VIC"

caddress$state <- factor(caddress$state)

summary(cdemographics$gender)
```

```
## Female   Male      U
##   2039   1873     88
```

```r
summary(caddress$state)
```

```
##  NSW  QLD  VIC
## 2140  838 1021
```

Age variable is added to cdemographics and newcustomer datasets.

```r
cdemographics$age <- 2020 - year(cdemographics$DOB)
newcustomer$age <- 2020 - year(newcustomer$DOB)
```

Summaries of new age columns can be seen below.

```r
summary(cdemographics$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   18.00   33.00   43.00   42.94   52.00   89.00      88
```

```r
summary(newcustomer$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   18.00   37.00   48.00   49.21   63.00   82.00      17
```

2 EDA (Exploratary Data Analysis)

I started to investigate datasets with using automatic Exploratary Data Analysis tools.

dfsummary

```r
cdemographics %>% dfSummary() %>% view()
```

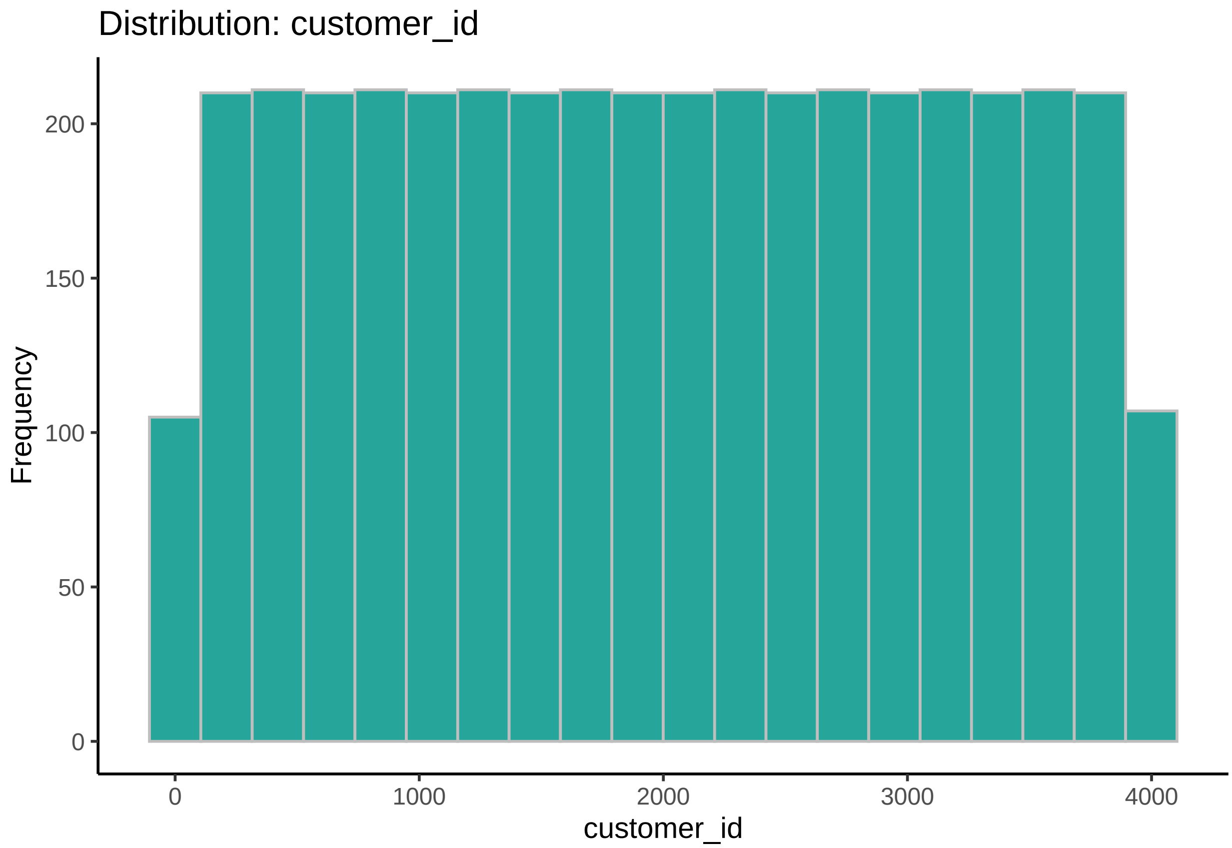
```
## Switching method to 'browser'

## Output file written: C:\Users\sbaru\AppData\Local\Temp\RtmpukPlY7\file6b0c4dbab2c.html
```

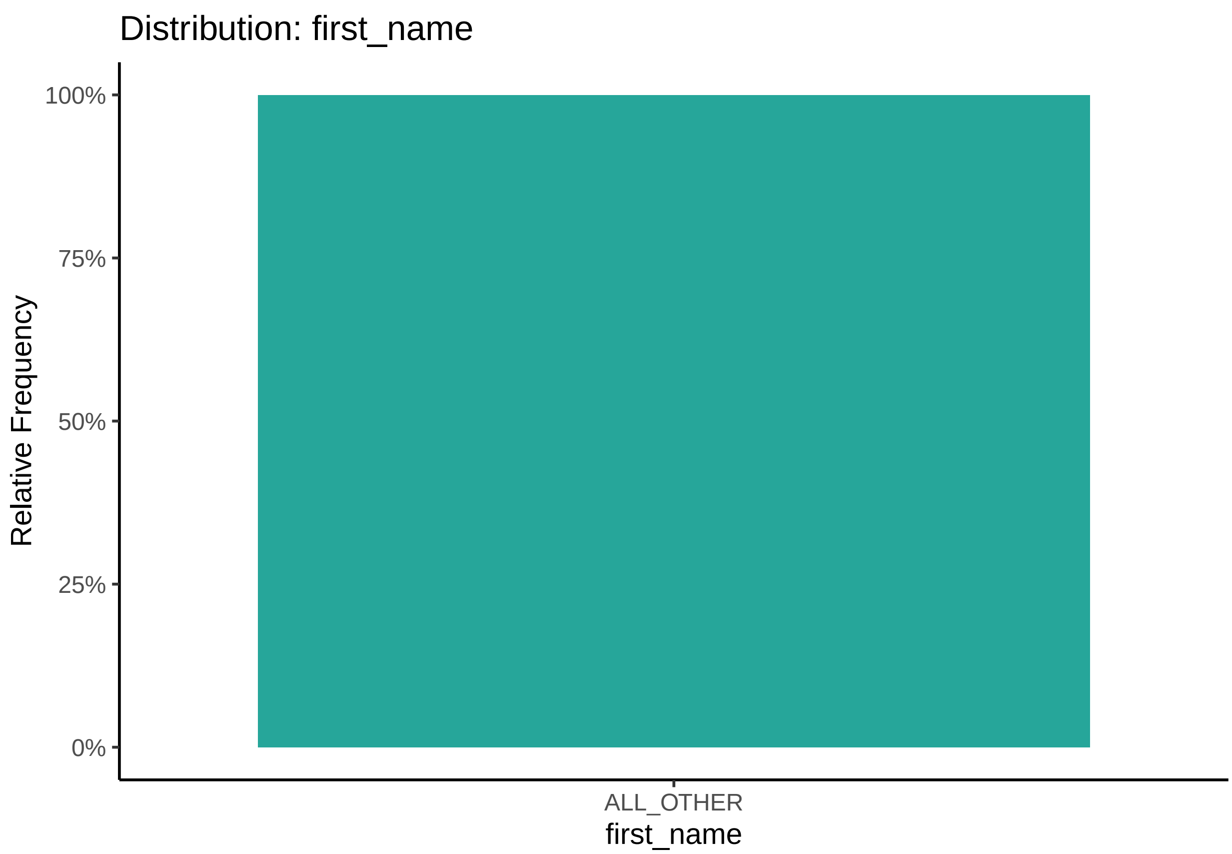
autoEDA

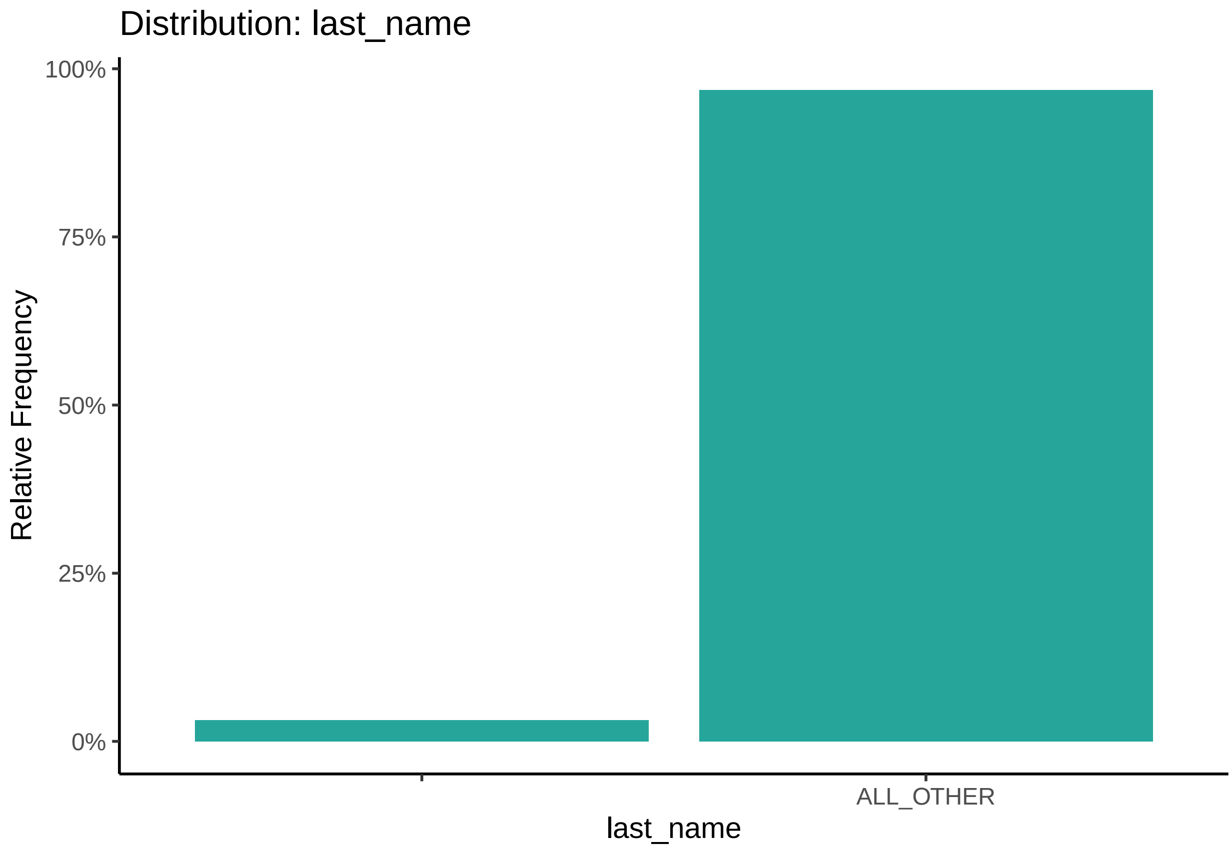
I arranged the code below as echo = F because it produces a graph for every column of datasets and make it the report hard to read. I use it as a prior investigarion. Graphs, which make sense to me, are going to be plotted after auto EDA part.

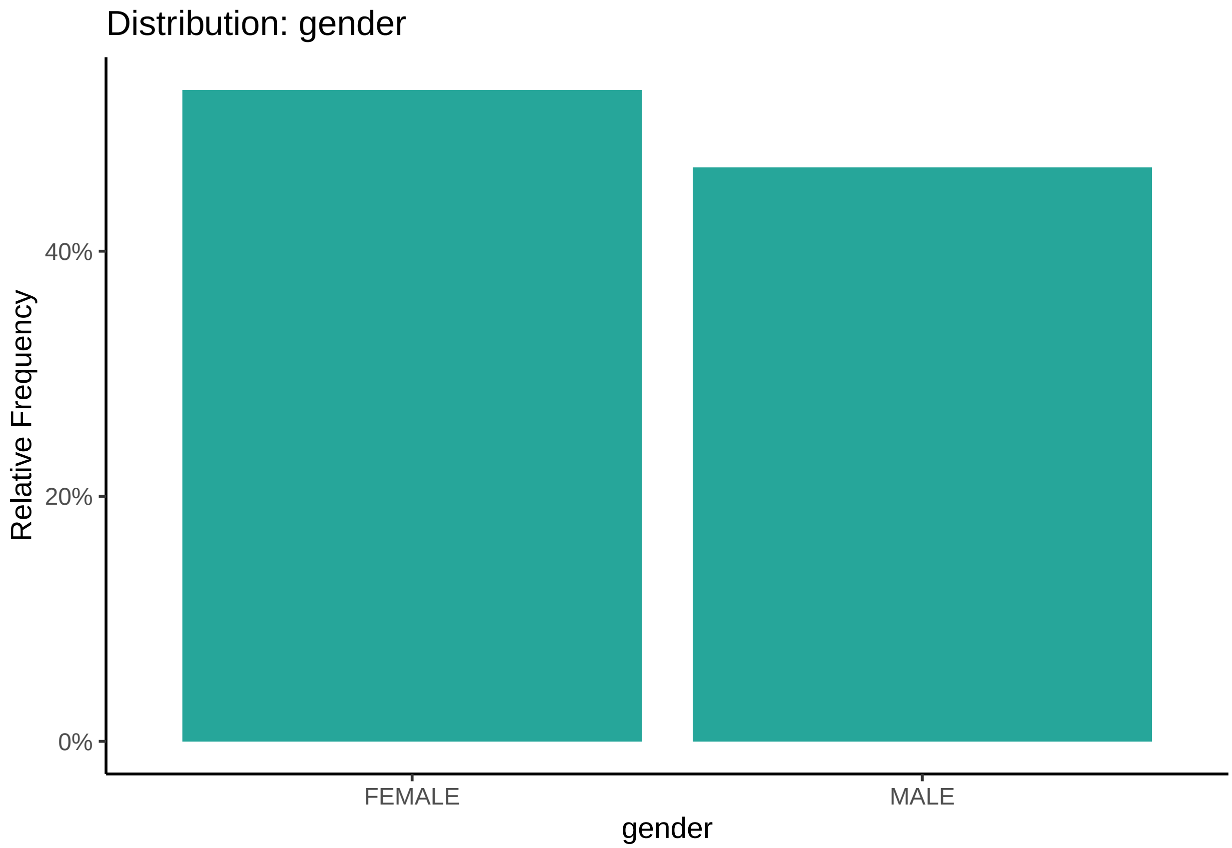
0th product have the most transactions record and its range shows a different trend than remainings. It has biggest price range between all the products.

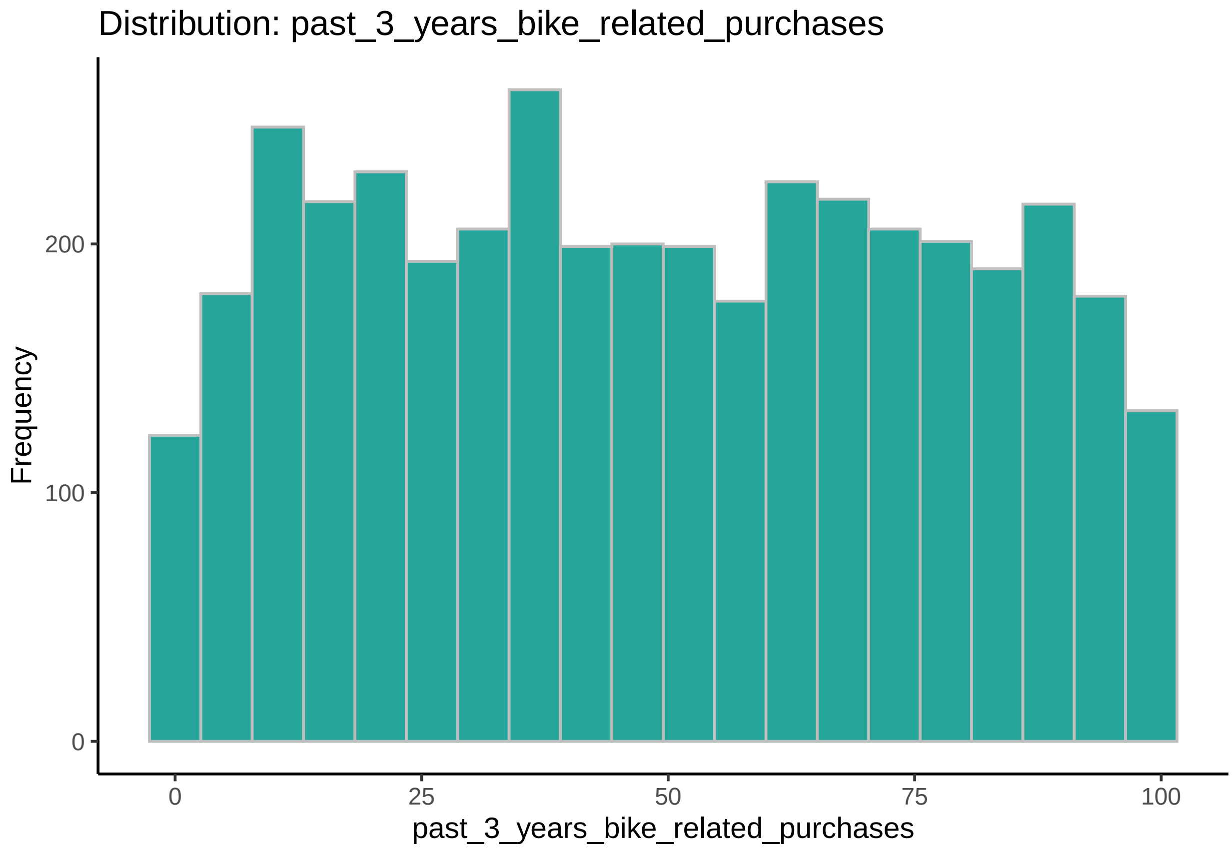
```
## Loading required package: RColorBrewer

## autoEDA | Setting color theme
## autoEDA | Removing constant features
## autoEDA | 0 constant features removed
## autoEDA | 0 zero spread features removed
## autoEDA | Removing features containing majority missing values
## autoEDA | 0 majority missing features removed
## autoEDA | Cleaning data
## autoEDA | Correcting sparse categorical feature levels
## autoEDA | Performing univariate analysis
## autoEDA | Visualizing data
```

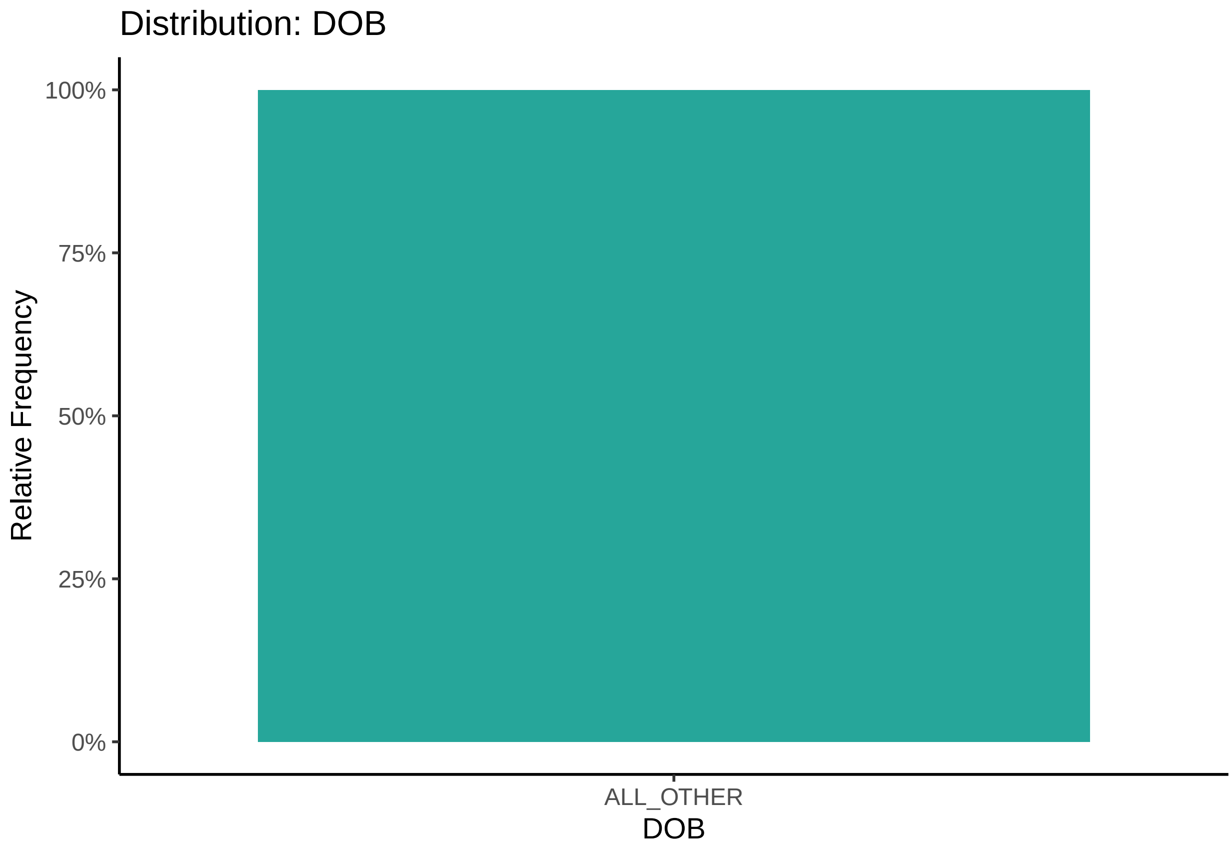

Distribution: customer_id
Frequency
0
50
100
150
200
0
1000
2000
3000
4000
customer_id

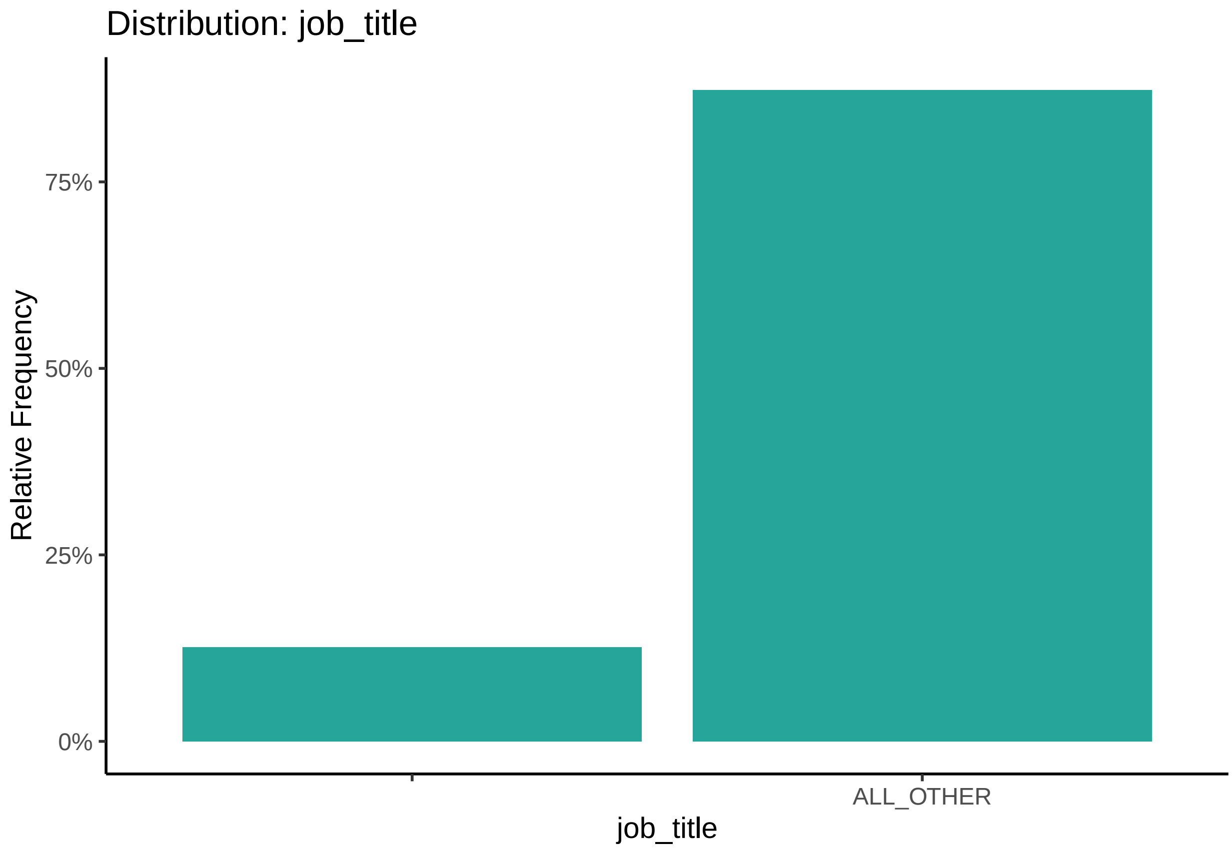

Distribution: first_name
100%
75%
50%
25%
0%
Relative Frequency
ALL_OTHER
first_name

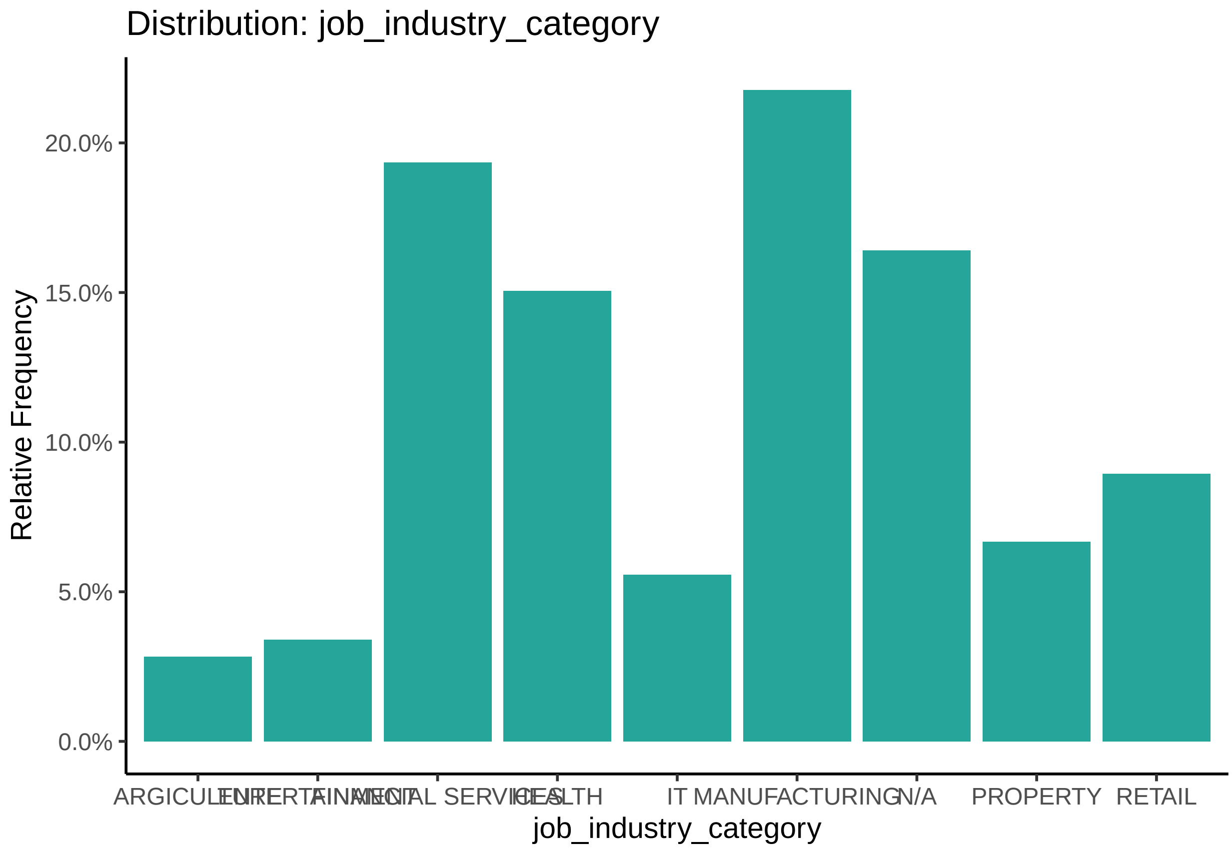

Distribution: last_name
100%
75%
50%
25%
0%
Relative Frequency
ALL_OTHER
last_name

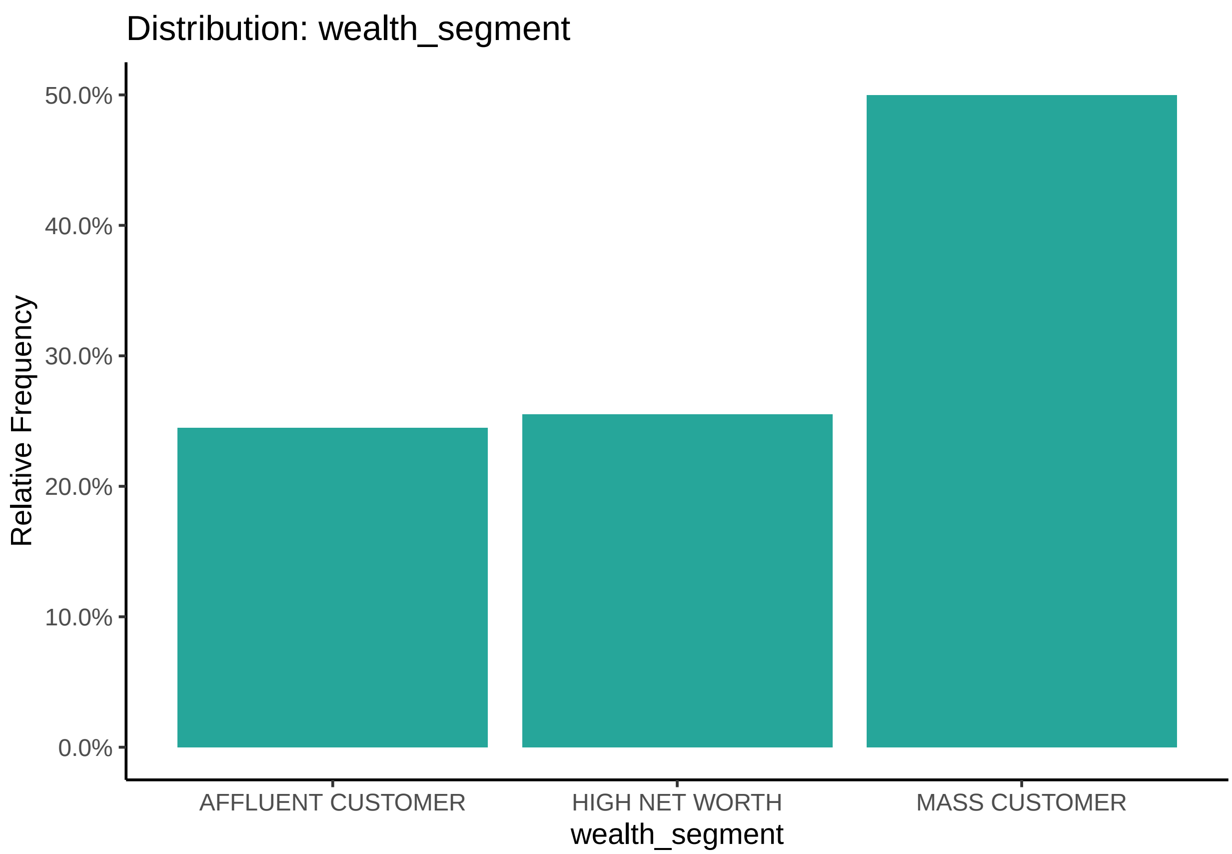

Distribution: gender
Relative Frequency
40%
20%
0%
FEMALE
MALE
gender

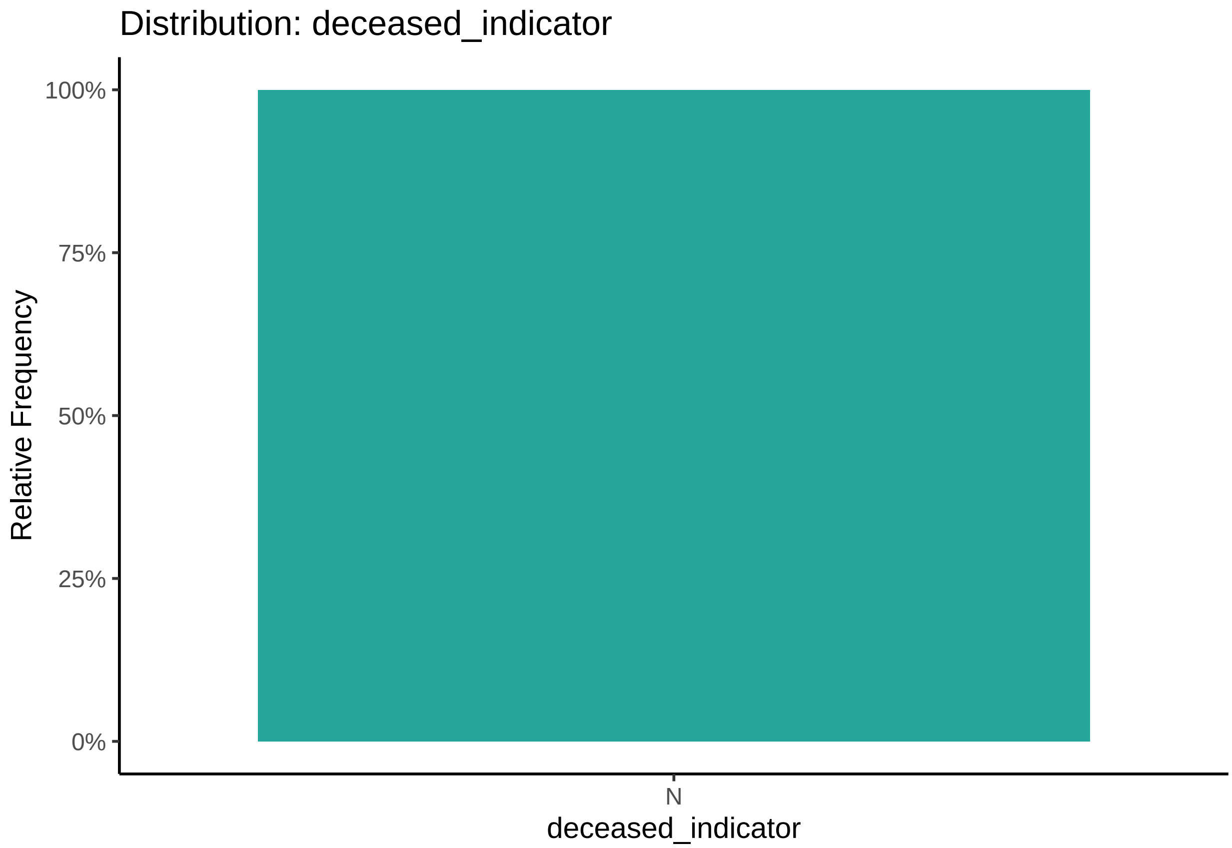

Distribution: past_3_years_bike_related_purchases
Frequency
200
100
0
0
25
50
75
100
past_3_years_bike_related_purchases

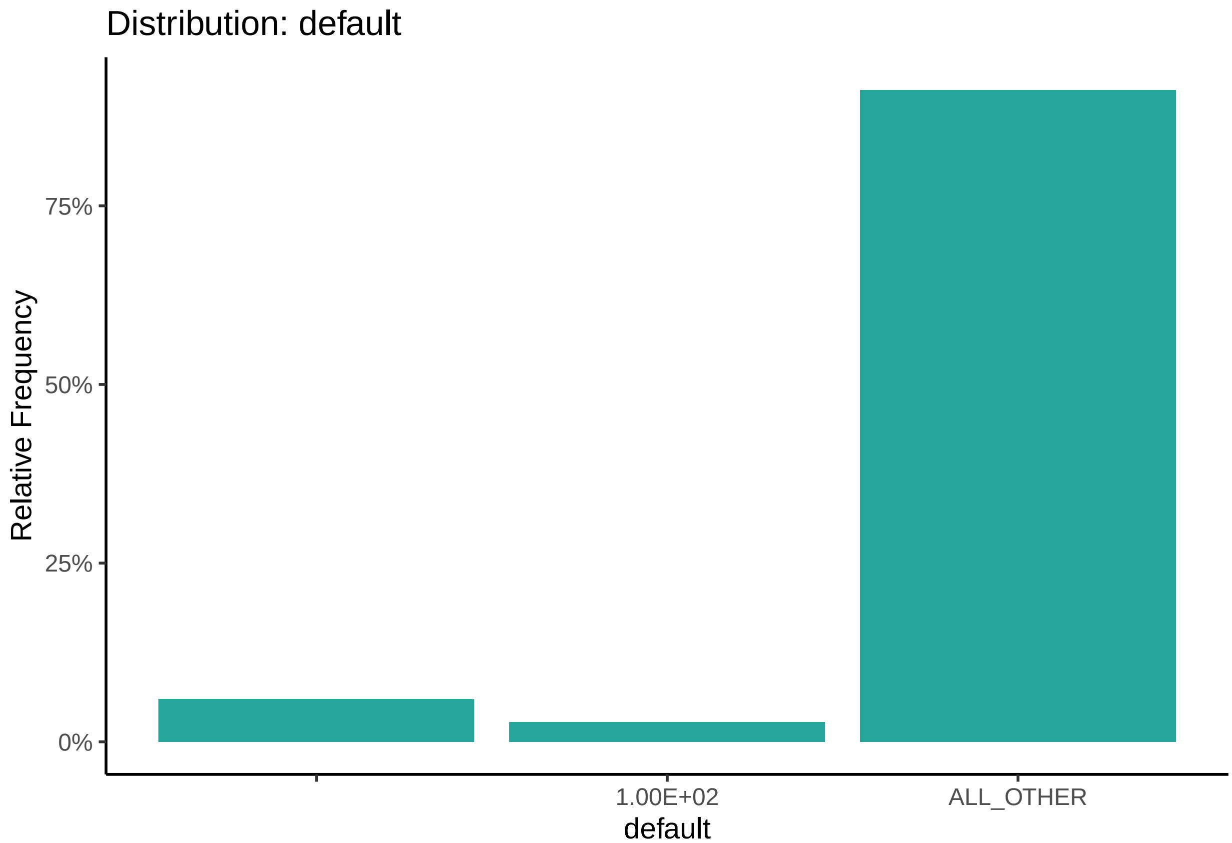

Distribution: DOB
100%
75%
50%
25%
0%
Relative Frequency
ALL_OTHER
DOB

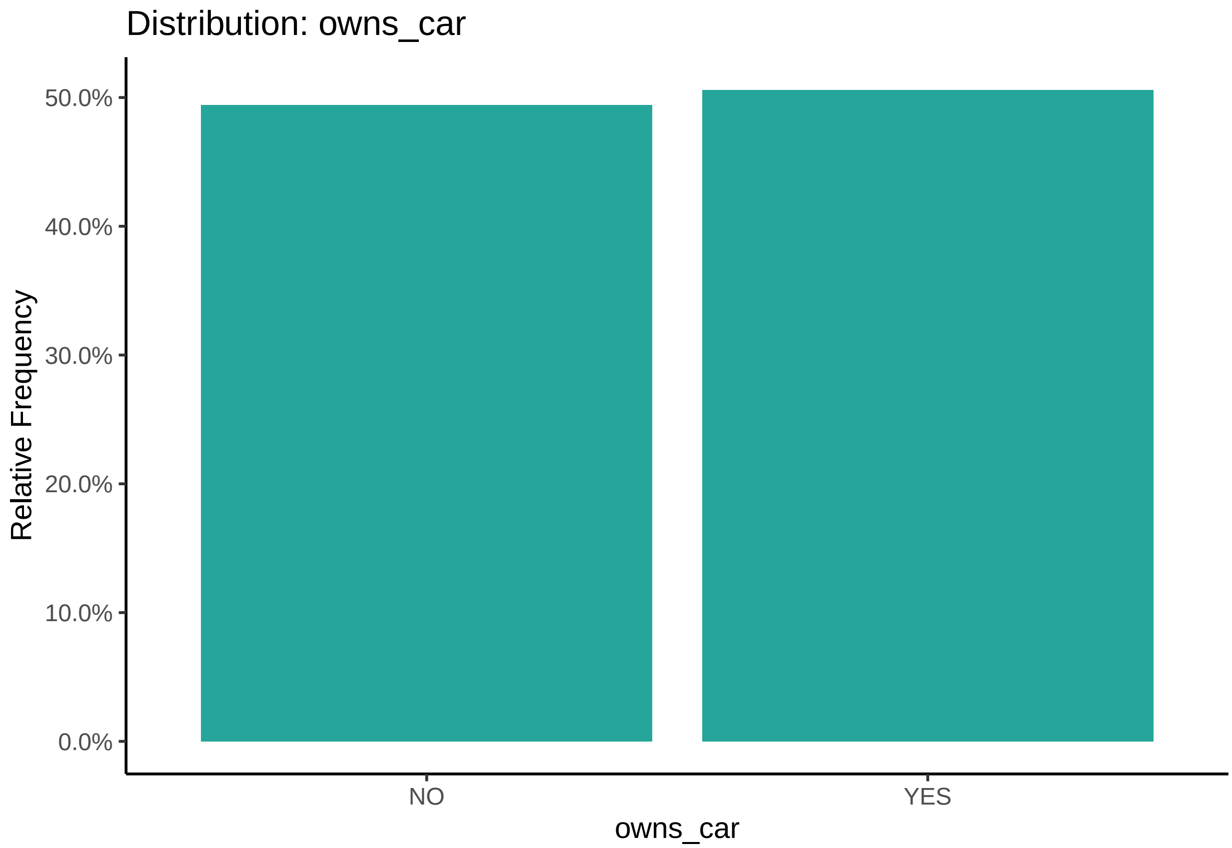

Distribution: job_title
Relative Frequency
75%
50%
25%
0%
ALL_OTHER
job_title

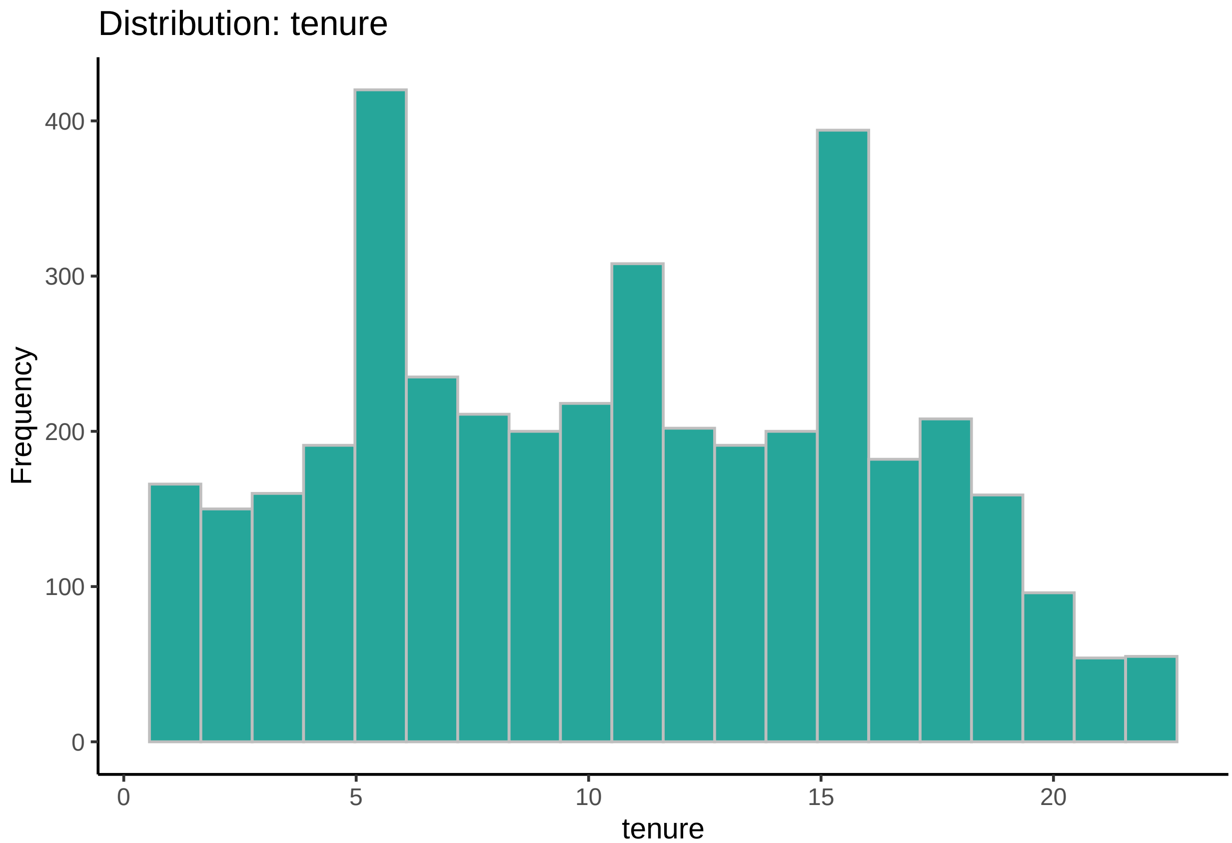

Distribution: job_industry_category
Relative Frequency
20.0%
15.0%
10.0%
5.0%
0.0%
HEALTH
IT
MANUFACTURING
N/A
PROPERTY
RETAIL
job_industry_category


Distribution: wealth_segment
Relative Frequency
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
AFFLUENT CUSTOMER
HIGH NET WORTH
MASS CUSTOMER
wealth_segment

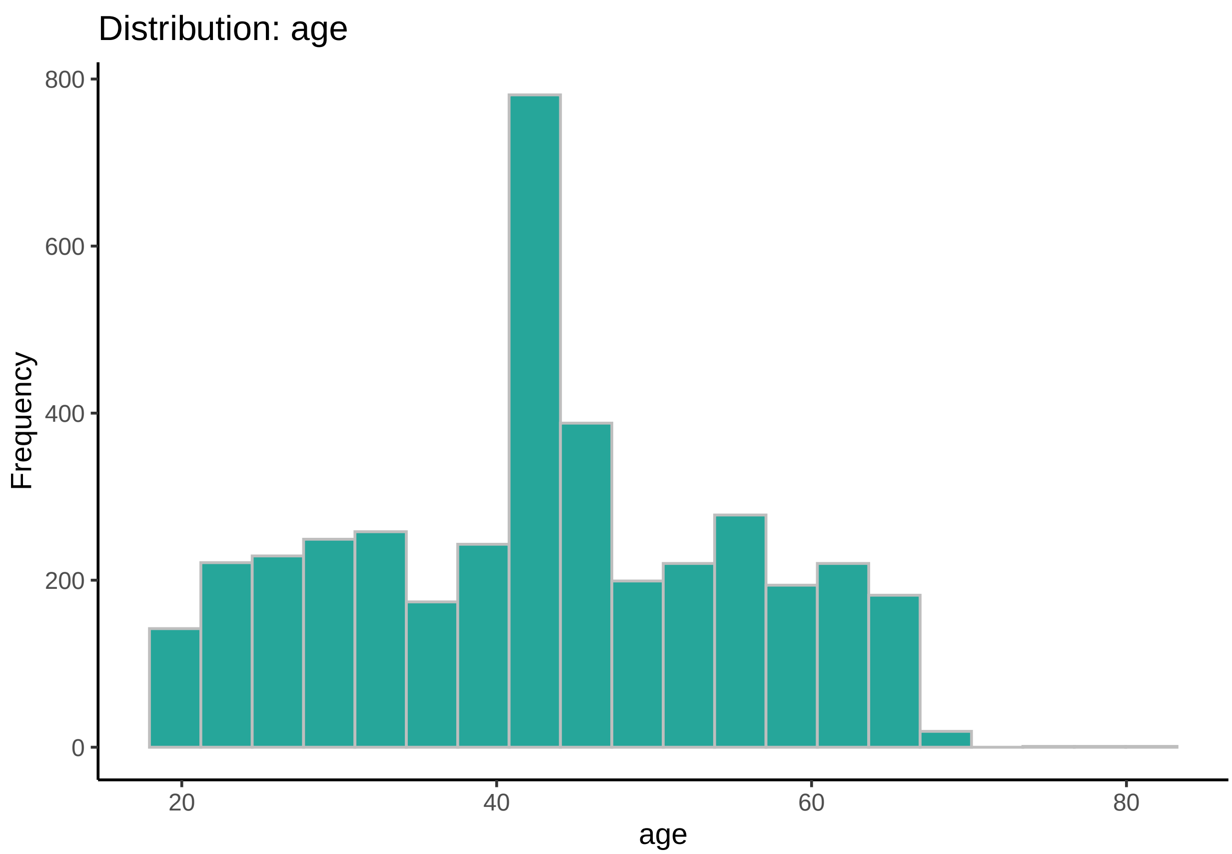

Distribution: deceased_indicator
100%
75%
50%
25%
0%
Relative Frequency
N
deceased_indicator


Distribution: default
Relative Frequency
75%
50%
25%
0%
1.00E+02
ALL_OTHER
default


Distribution: owns_car
Relative Frequency
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
NO
YES
owns_car



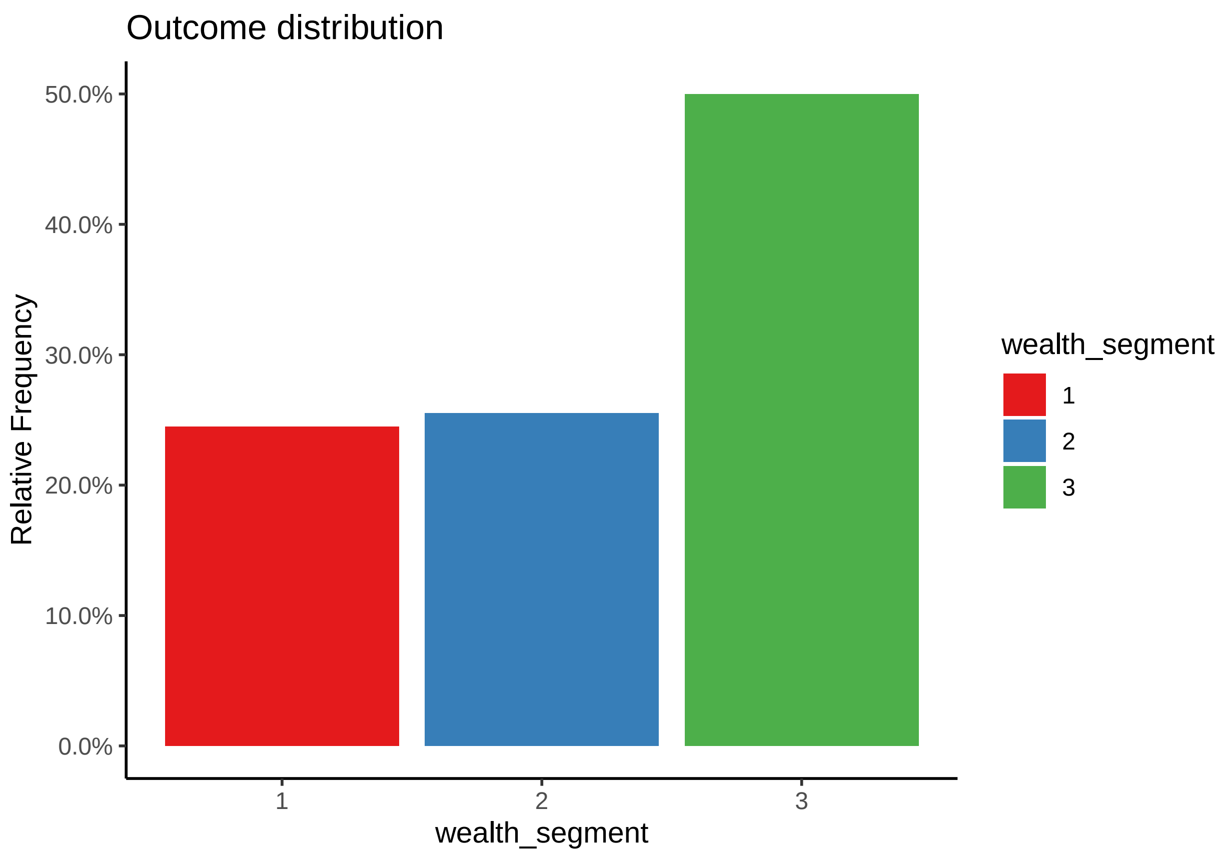
Distribution: tenure
Frequency
400
300
200
100
0
0
5
10
15
20
tenure

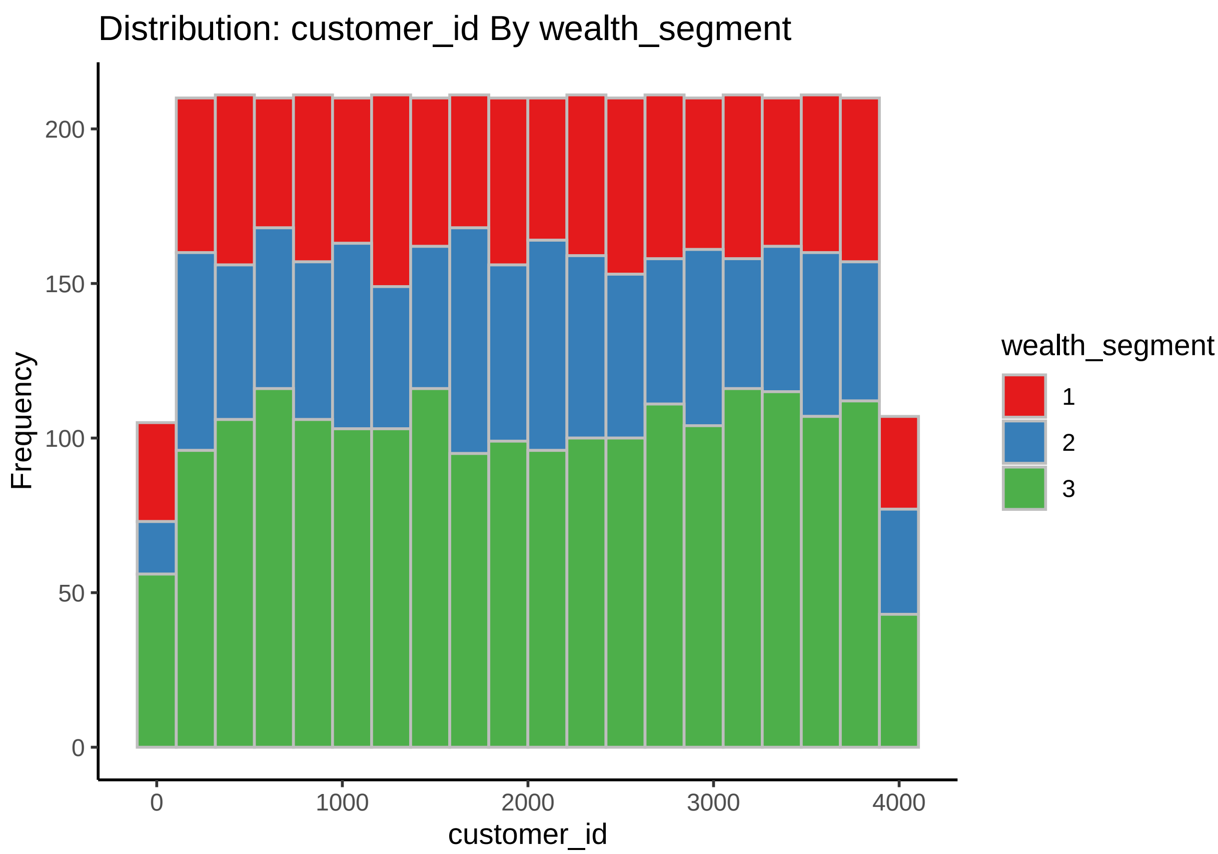



```
##                                Feature Observations FeatureClass FeatureType
## 1                          customer_id         4000      numeric  Continuous
## 2                           first_name         4000    character Categorical
## 3                            last_name         4000    character Categorical
## 4                               gender         4000    character Categorical
## 5  past_3_years_bike_related_purchases         4000      numeric  Continuous
## 6                                  DOB         4000    character Categorical
## 7                            job_title         4000    character Categorical
## 8                 job_industry_category         4000    character Categorical
## 9                       wealth_segment         4000    character Categorical
## 10                  deceased_indicator         4000    character Categorical
## 11                             default         4000    character Categorical
## 12                            owns_car         4000    character Categorical
## 13                              tenure         4000      numeric  Continuous
## 14                                 age         4000      numeric  Continuous
##    PercentageMissing PercentageUnique ConstantFeature ZeroSpreadFeature
## 1               0.00           100.00              No                No
## 2               0.00            78.47              No                No
## 3               0.00            93.15              No                No
## 4               0.00             0.08              No                No
## 5               0.00             2.50              No                No
## 6               2.20            86.20              No                No
## 7               0.00             4.90              No                No
## 8               0.00             0.25              No                No
## 9               0.00             0.08              No                No
## 10              0.00             0.05              No                No
```

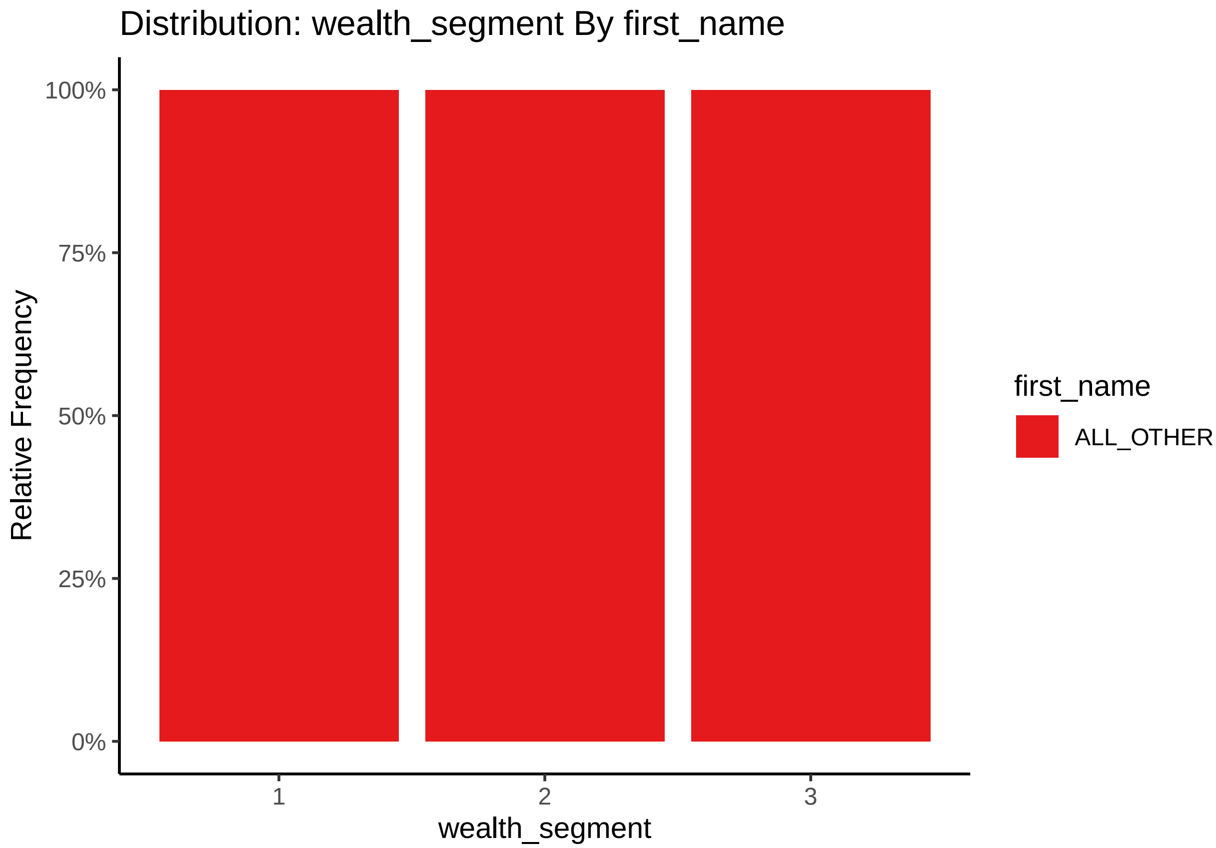
```
## 11             0.00            2.55            No                No
## 12             0.00            0.05            No                No
## 13             2.17            0.58            No                No
## 14             2.20            1.40            No                No
##    LowerOutliers UpperOutliers ImputationValue MinValue FirstQuartile Median
## 1              0             0          2000.5        1       1000.75 2000.5
## 2              0             0       ALL_OTHER        0          0.00    0.0
## 3              0             0       ALL_OTHER        0          0.00    0.0
## 4              0             0          FEMALE        0          0.00    0.0
## 5              0             0              48        0         24.00   48.0
## 6              0             0         MISSING        0          0.00    0.0
## 7              0             0       ALL_OTHER        0          0.00    0.0
## 8              0             0   MANUFACTURING        0          0.00    0.0
## 9              0             0   MASS CUSTOMER        0          0.00    0.0
## 10             0             0               N        0          0.00    0.0
## 11             0             0       ALL_OTHER        0          0.00    0.0
## 12             0             0             YES        0          0.00    0.0
## 13             0             0              11        1          6.00   11.0
## 14             0             2              43       18         33.00   43.0
##       Mean          Mode ThirdQuartile MaxValue LowerOutlierValue
## 1  2000.50             1       3000.25     4000           -1998.5
## 2     0.00           MAX          0.00        0               0.0
## 3     0.00                        0.00        0               0.0
## 4     0.00        FEMALE          0.00        0               0.0
## 5    48.89            16         73.00       99             -49.5
## 6     0.00    1978-01-30          0.00        0               0.0
## 7     0.00                        0.00        0               0.0
## 8     0.00 MANUFACTURING          0.00        0               0.0
## 9     0.00 MASS CUSTOMER          0.00        0               0.0
## 10    0.00             N          0.00        0               0.0
## 11    0.00                        0.00        0               0.0
## 12    0.00           YES          0.00        0               0.0
## 13   10.66             7         15.00       22              -7.5
## 14   42.94            42         52.00       89               4.5
##    UpperOutlierValue
## 1             5999.5
## 2                0.0
## 3                0.0
## 4                0.0
## 5              146.5
## 6                0.0
## 7                0.0
## 8                0.0
## 9                0.0
## 10               0.0
## 11               0.0
## 12               0.0
## 13              28.5
## 14              80.5

## autoEDA | Setting color theme
## autoEDA | Removing constant features
## autoEDA | 0 constant features removed
## autoEDA | Removing zero spread features
```

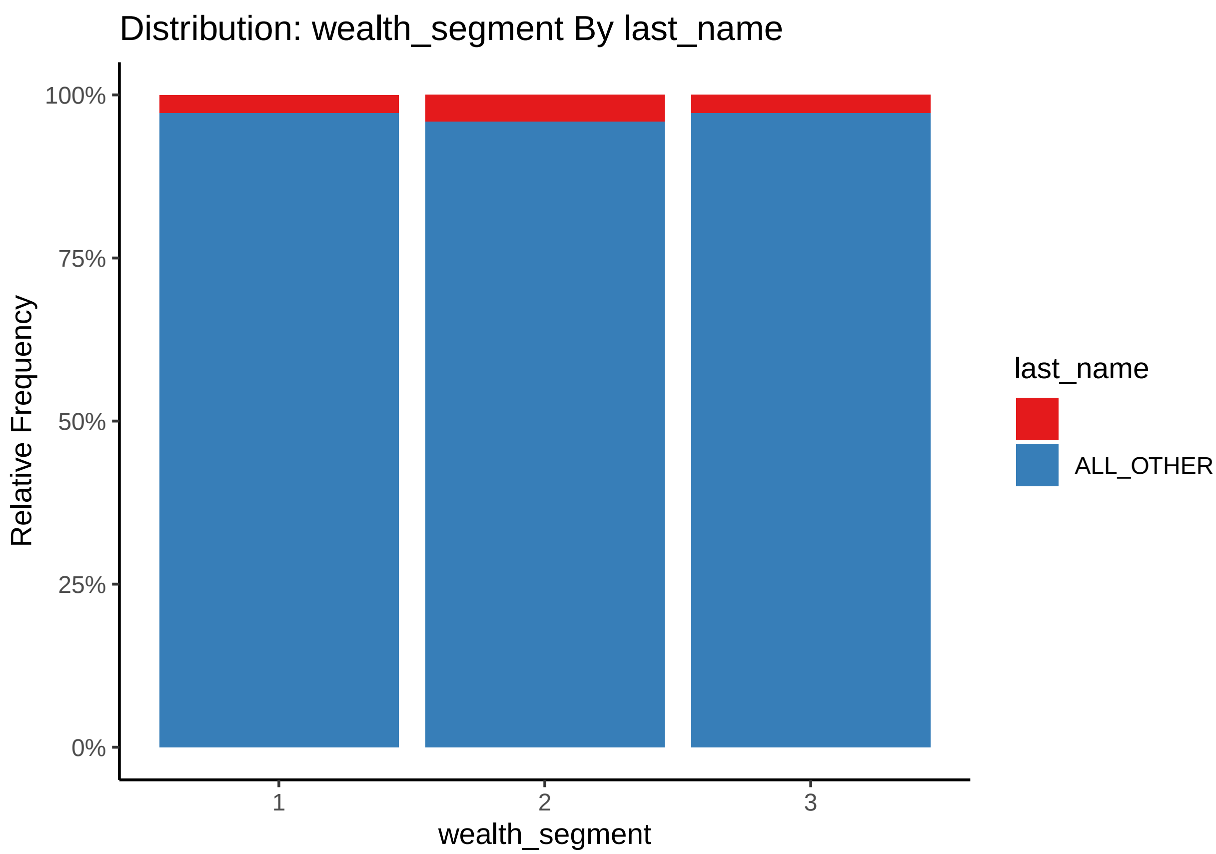
```
## autoEDA | 0 zero spread features removed
## autoEDA | Removing features containing majority missing values
## autoEDA | 0 majority missing features removed
## autoEDA | Cleaning data
## autoEDA | Correcting sparse categorical feature levels
## autoEDA | Sorting features
## autoEDA | Multi-class classification outcome detected
## autoEDA | Calculating feature predictive power
## autoEDA | Visualizing data
```

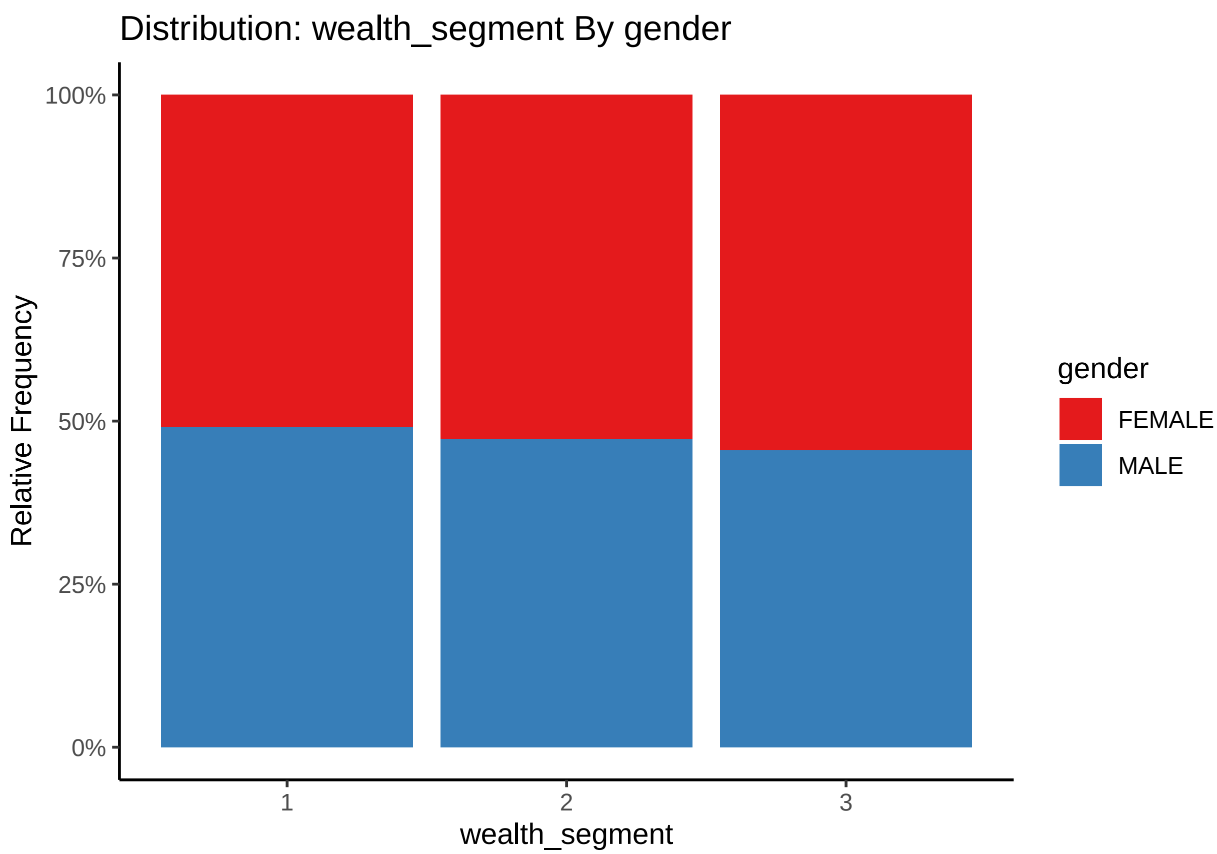


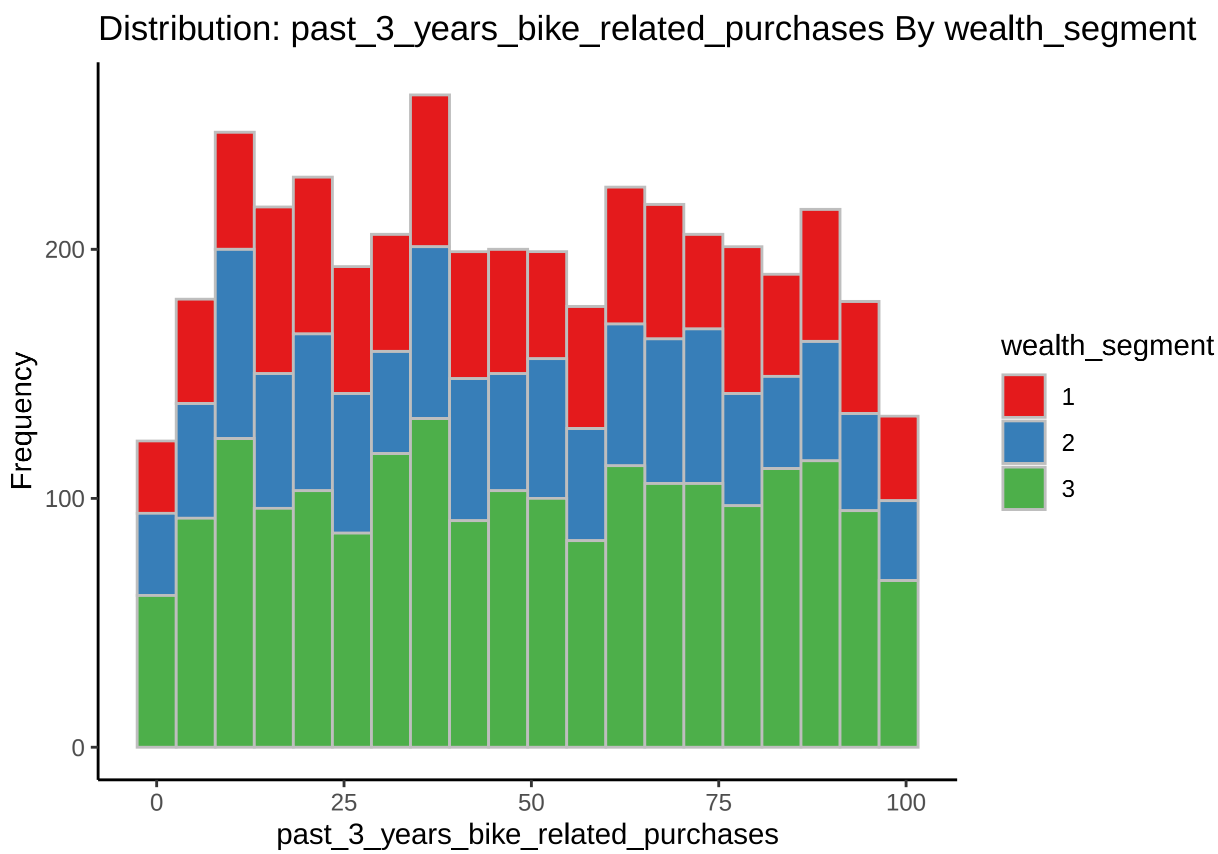


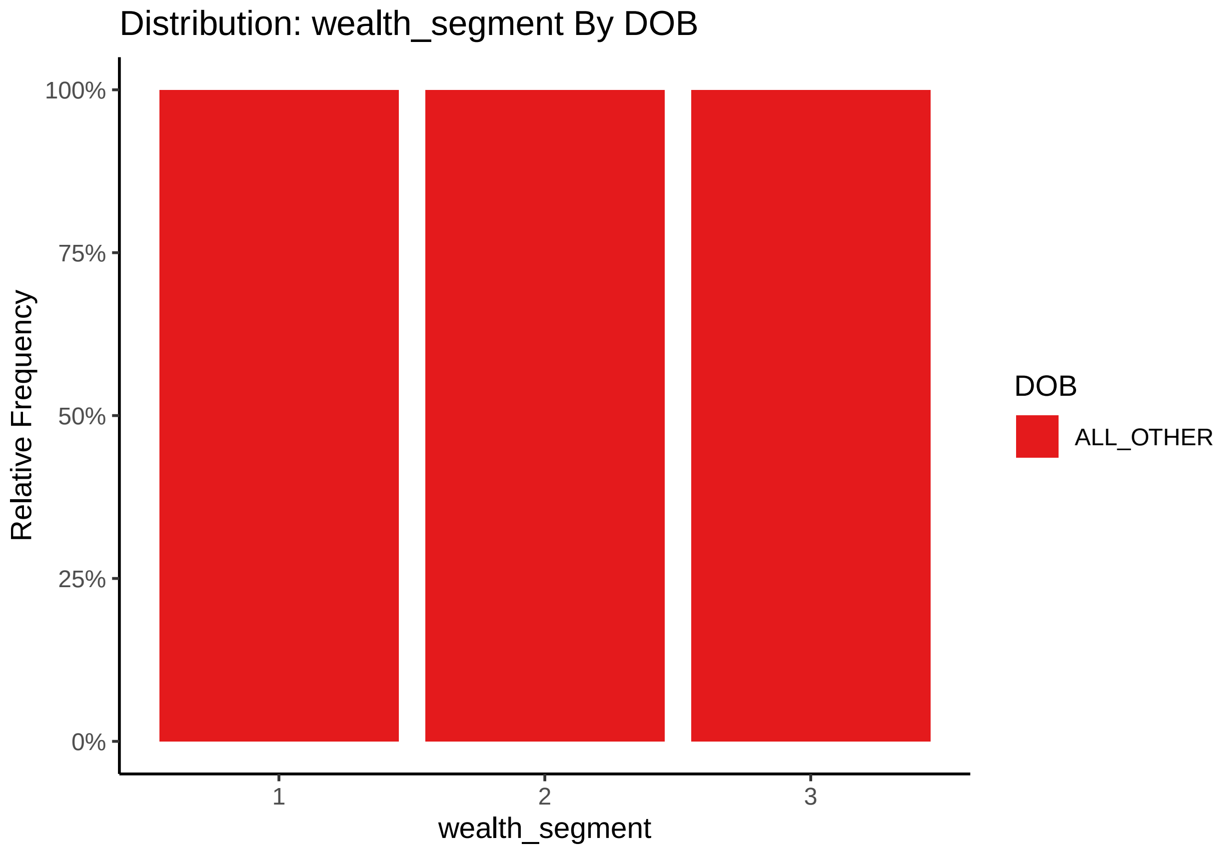
Distribution: customer_id By wealth_segment
Frequency
200
150
100
50
0
0
1000
2000
3000
4000
customer_id
wealth_segment
1
2
3

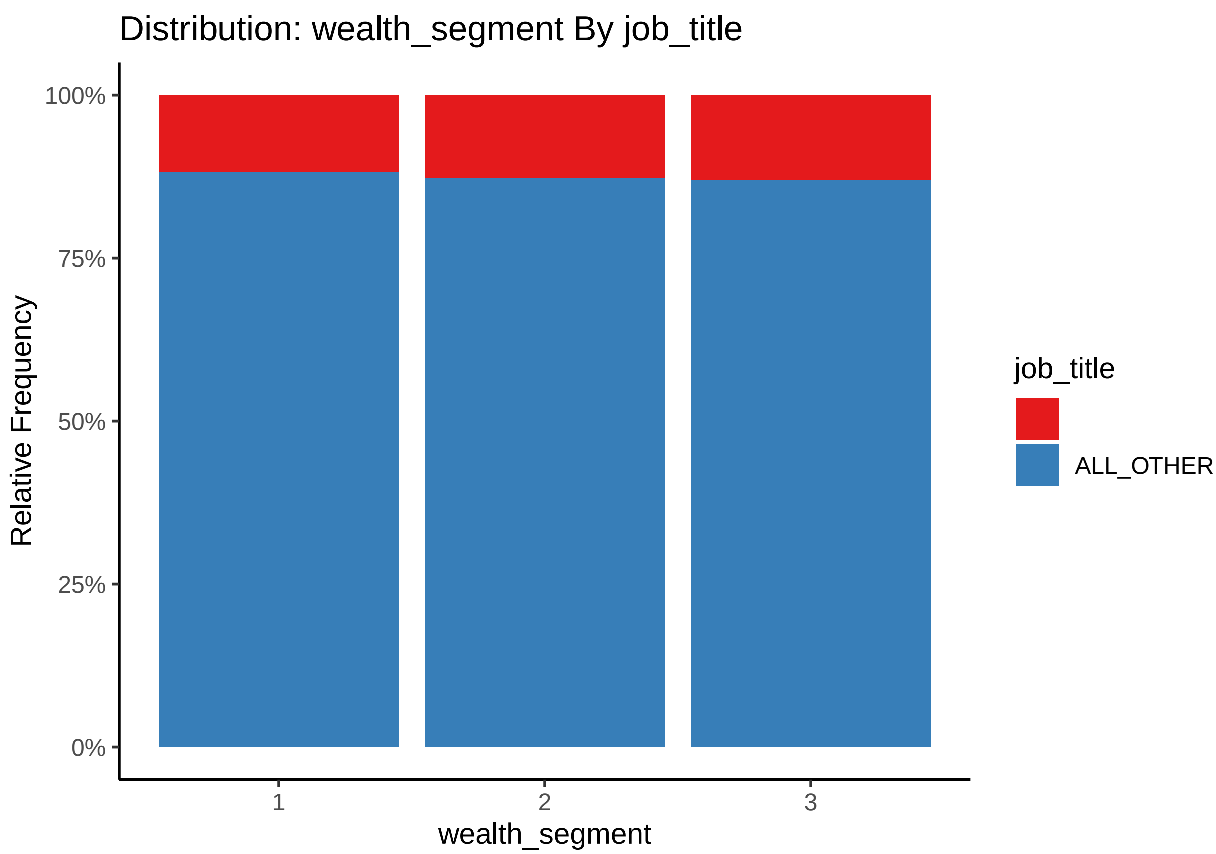

Distribution: wealth_segment By first_name
Relative Frequency
100%
75%
50%
25%
0%
1
2
3
wealth_segment
first_name
ALL_OTHER

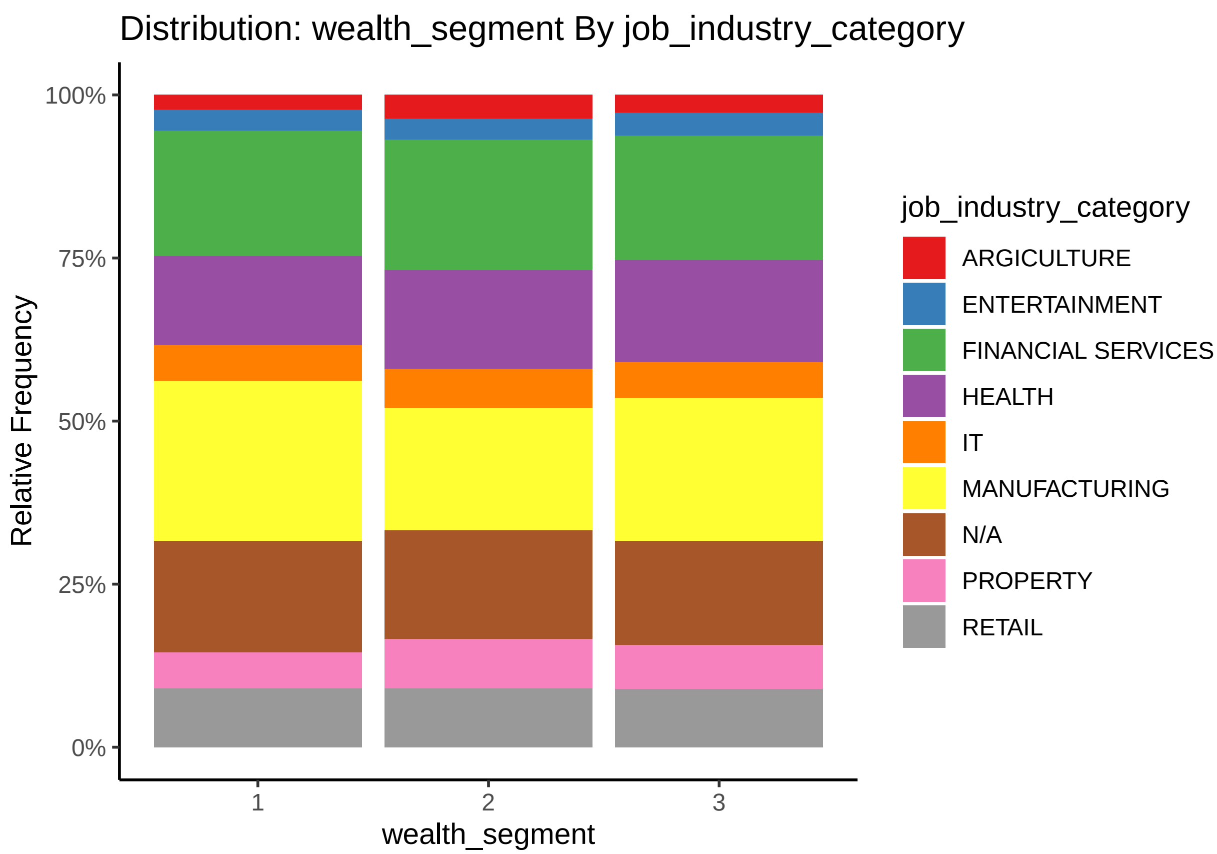

Distribution: wealth_segment By last_name
Relative Frequency
100%
75%
50%
25%
0%
1
2
3
wealth_segment
last_name
ALL_OTHER


Distribution: wealth_segment By gender
Relative Frequency
100%
75%
50%
25%
0%
1
2
3
wealth_segment
gender
FEMALE
MALE

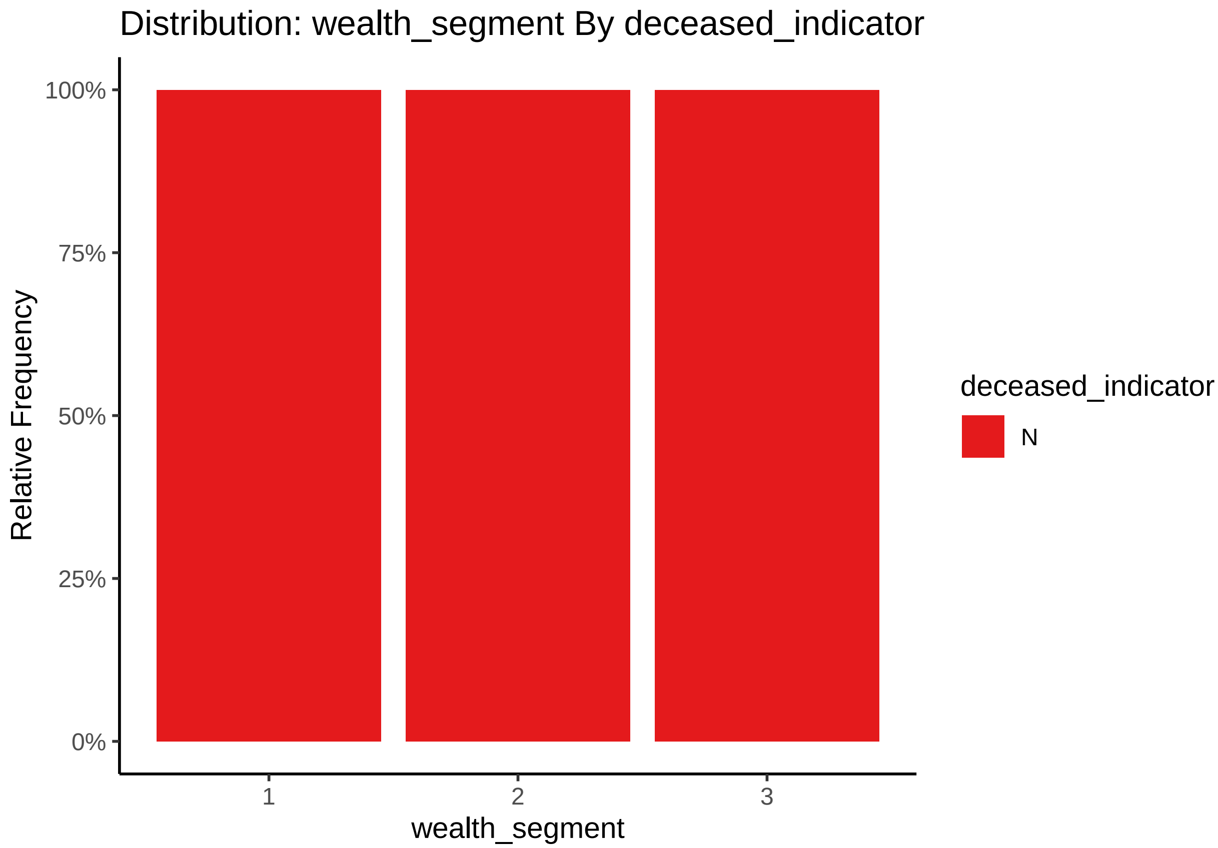

Distribution: past_3_years_bike_related_purchases By wealth_segment
Frequency
200
100
0
0
25
50
75
100
past_3_years_bike_related_purchases
wealth_segment
1
2
3


Distribution: wealth_segment By DOB
100%
75%
50%
25%
0%
Relative Frequency
1
2
3
wealth_segment
DOB
ALL_OTHER

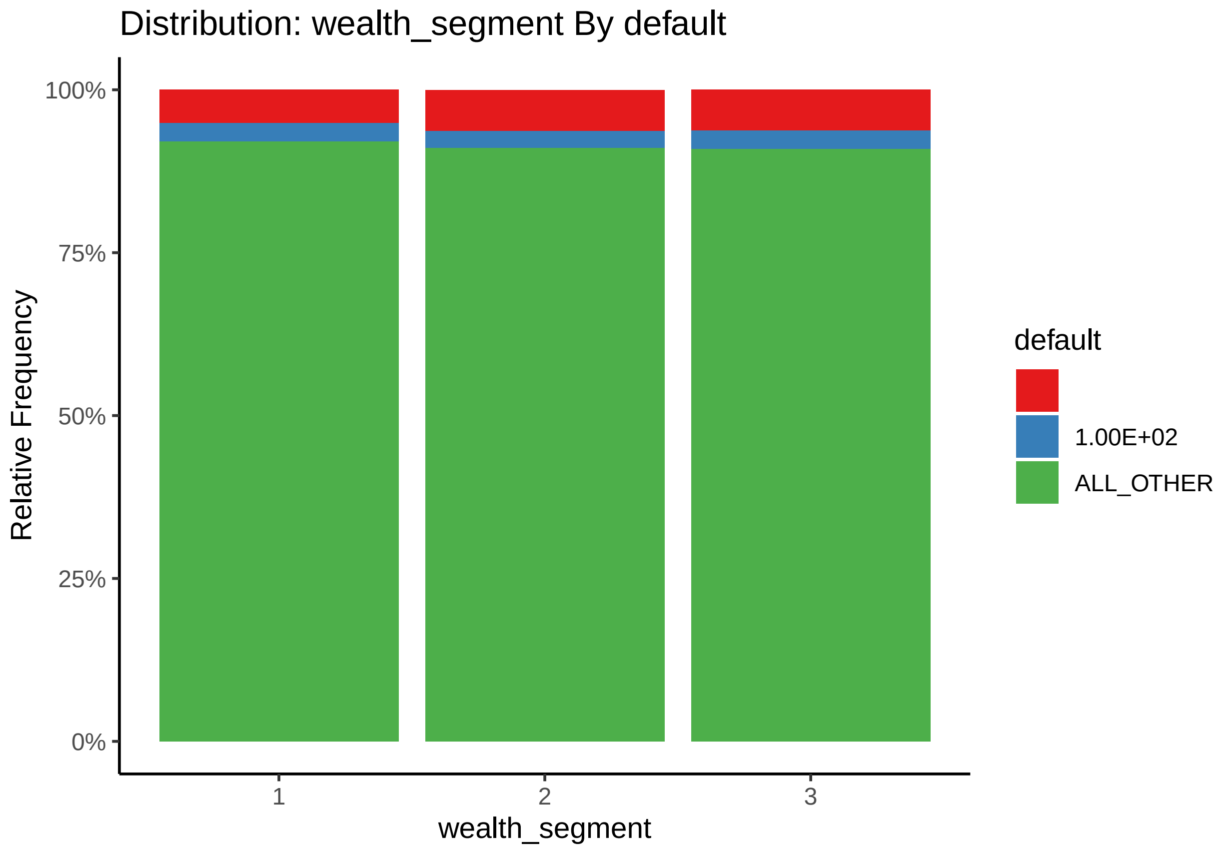

Distribution: wealth_segment By job_title
100%
75%
50%
25%
0%
Relative Frequency
1
2
3
wealth_segment
job_title
ALL_OTHER

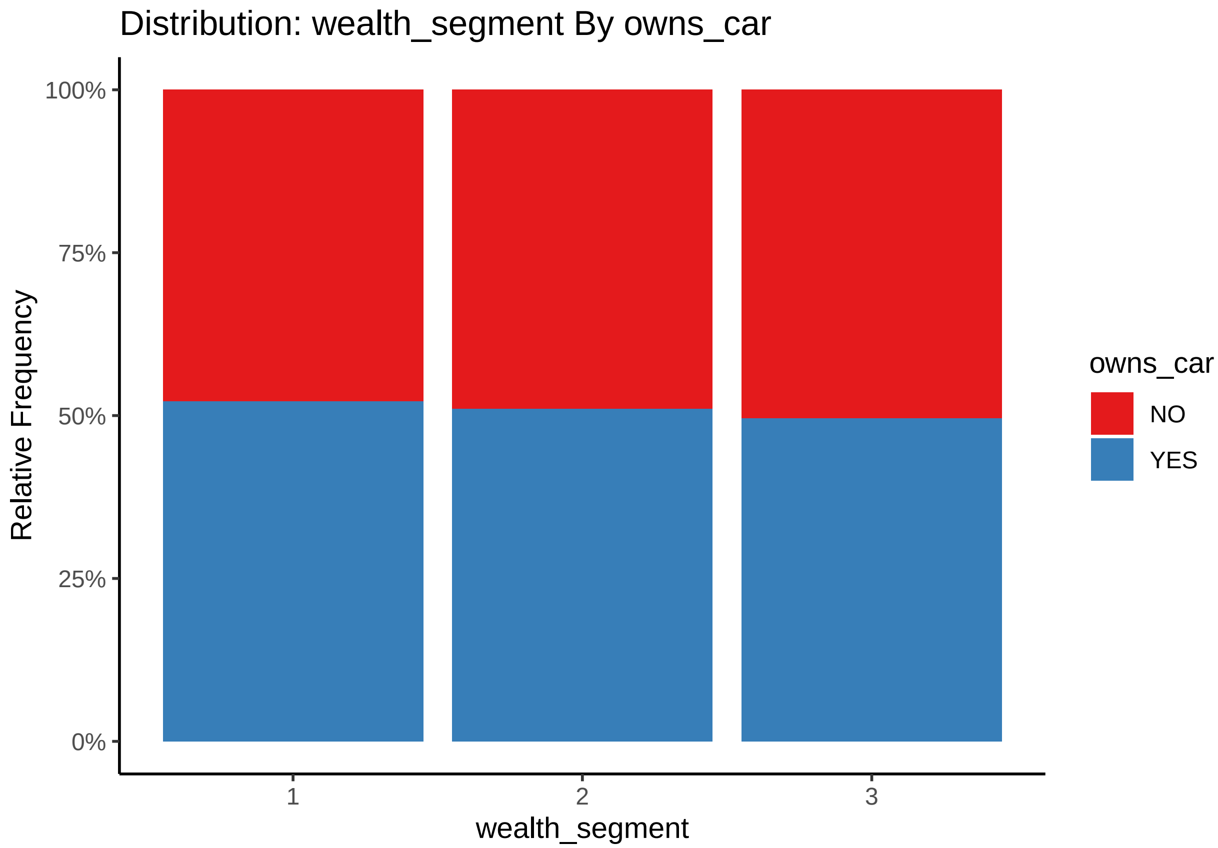


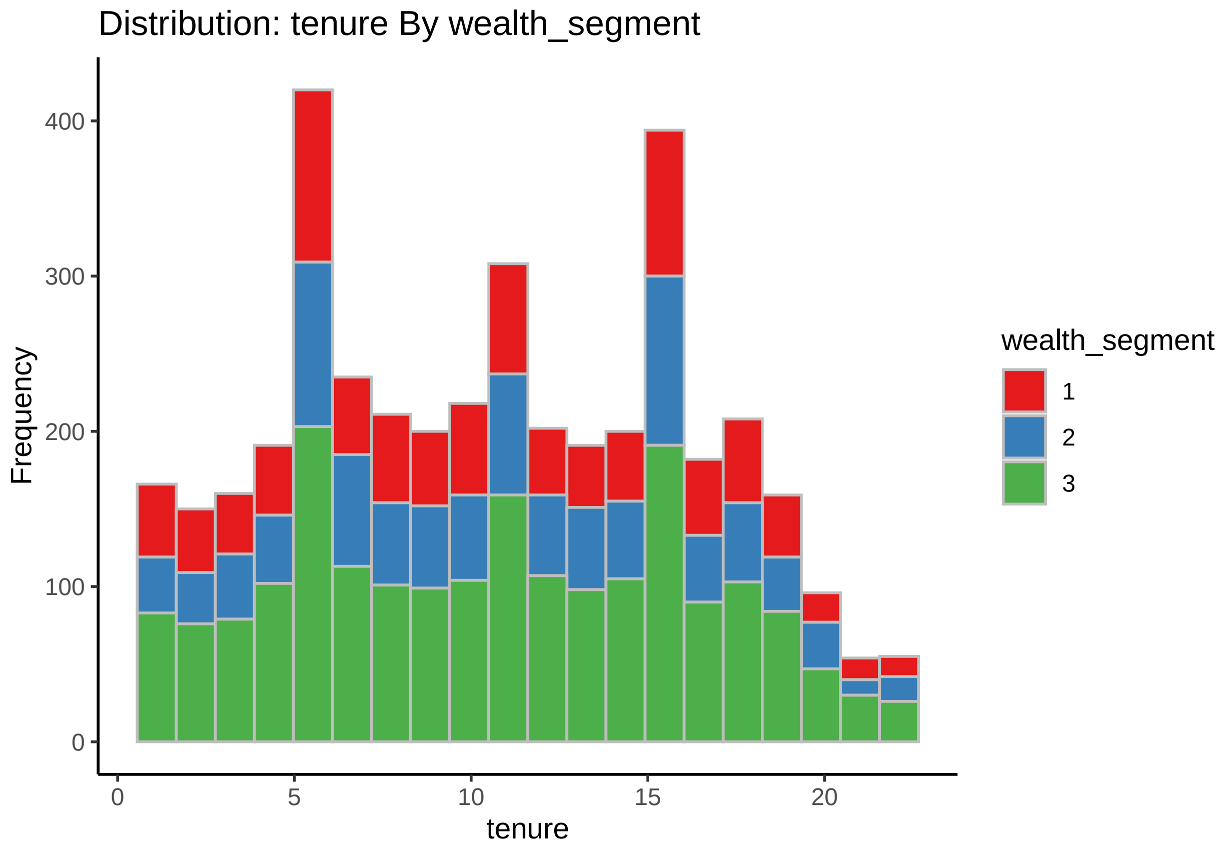
Distribution: wealth_segment By job_industry_category
Relative Frequency
100%
75%
50%
25%
0%
1
2
3
wealth_segment
job_industry_category
ARGICULTURE
ENTERTAINMENT
FINANCIAL SERVICES
HEALTH
IT
MANUFACTURING
N/A
PROPERTY
RETAIL



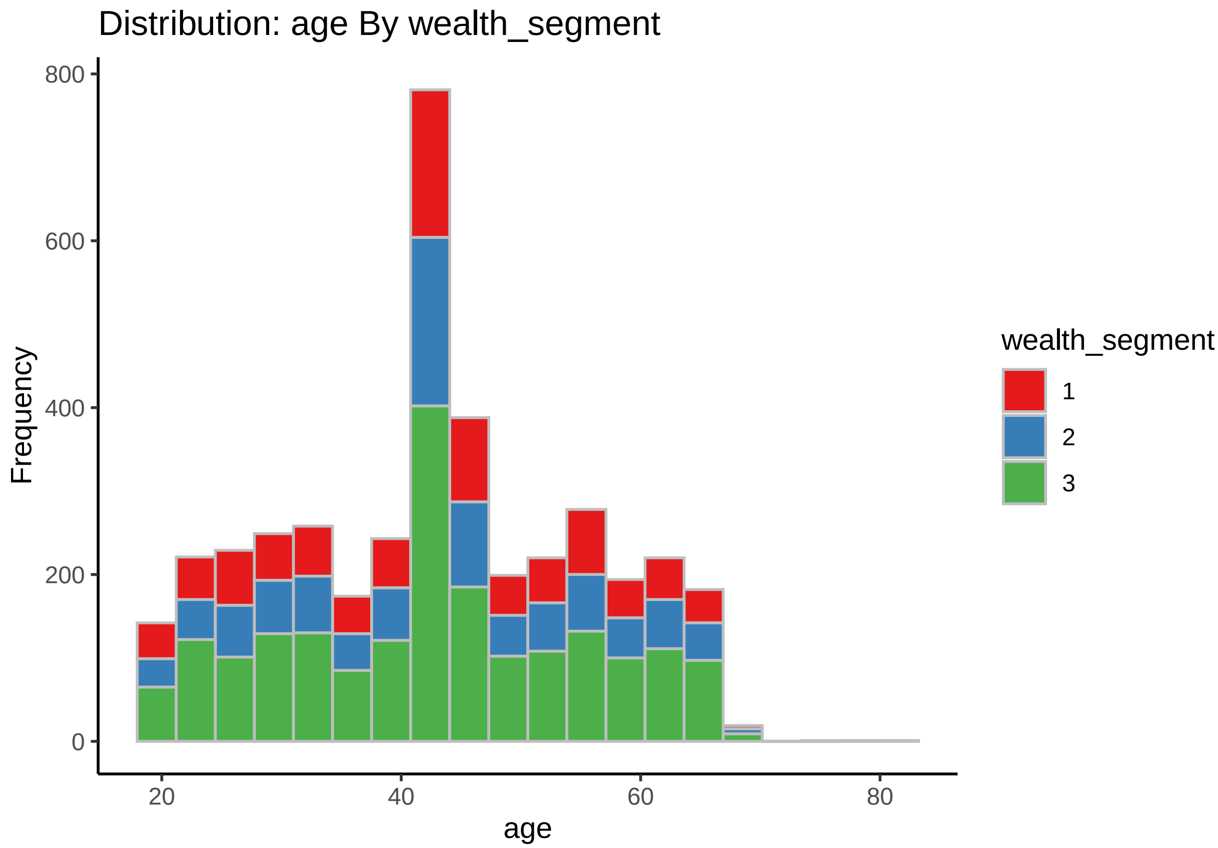
Distribution: wealth_segment By deceased_indicator
100%
75%
50%
25%
0%
Relative Frequency
1
2
3
wealth_segment
deceased_indicator
N


Distribution: wealth_segment By default
Relative Frequency
100%
75%
50%
25%
0%
1
2
3
wealth_segment
default
1.00E+02
ALL_OTHER

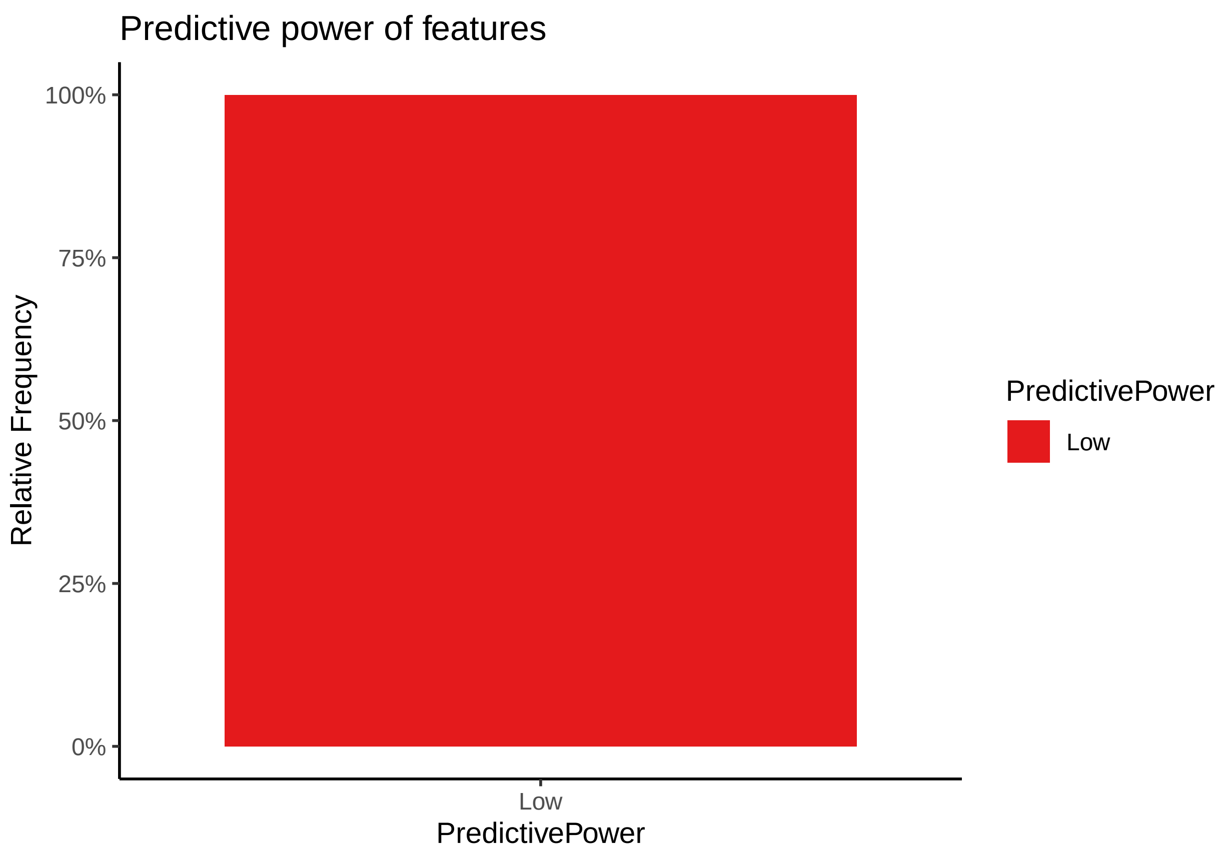

Distribution: wealth_segment By owns_car
Relative Frequency
100%
75%
50%
25%
0%
1
2
3
wealth_segment
owns_car
NO
YES



Distribution: tenure By wealth_segment
Frequency
400
300
200
100
0
0
5
10
15
20
tenure
wealth_segment
1
2
3



Distribution: age By wealth_segment
Frequency
800
600
400
200
0
20
40
60
80
age
wealth_segment
1
2
3

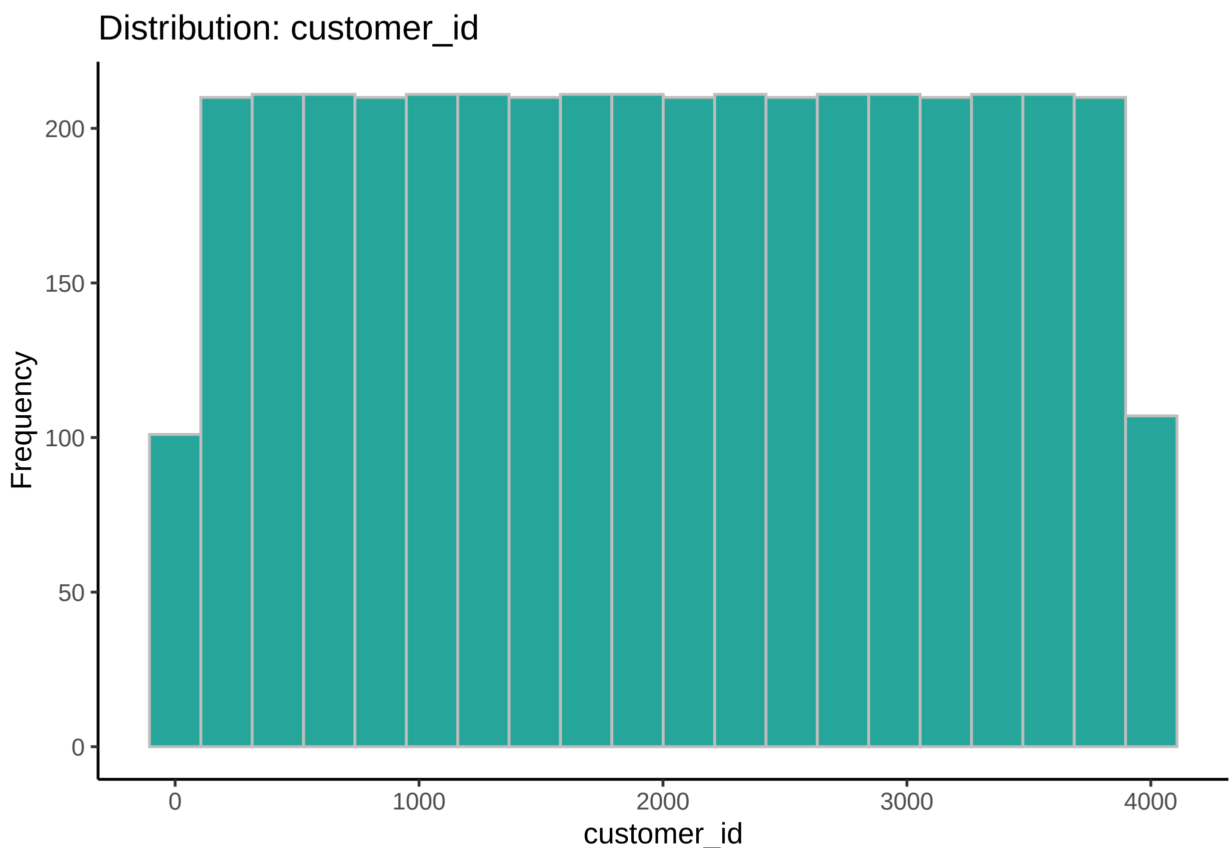


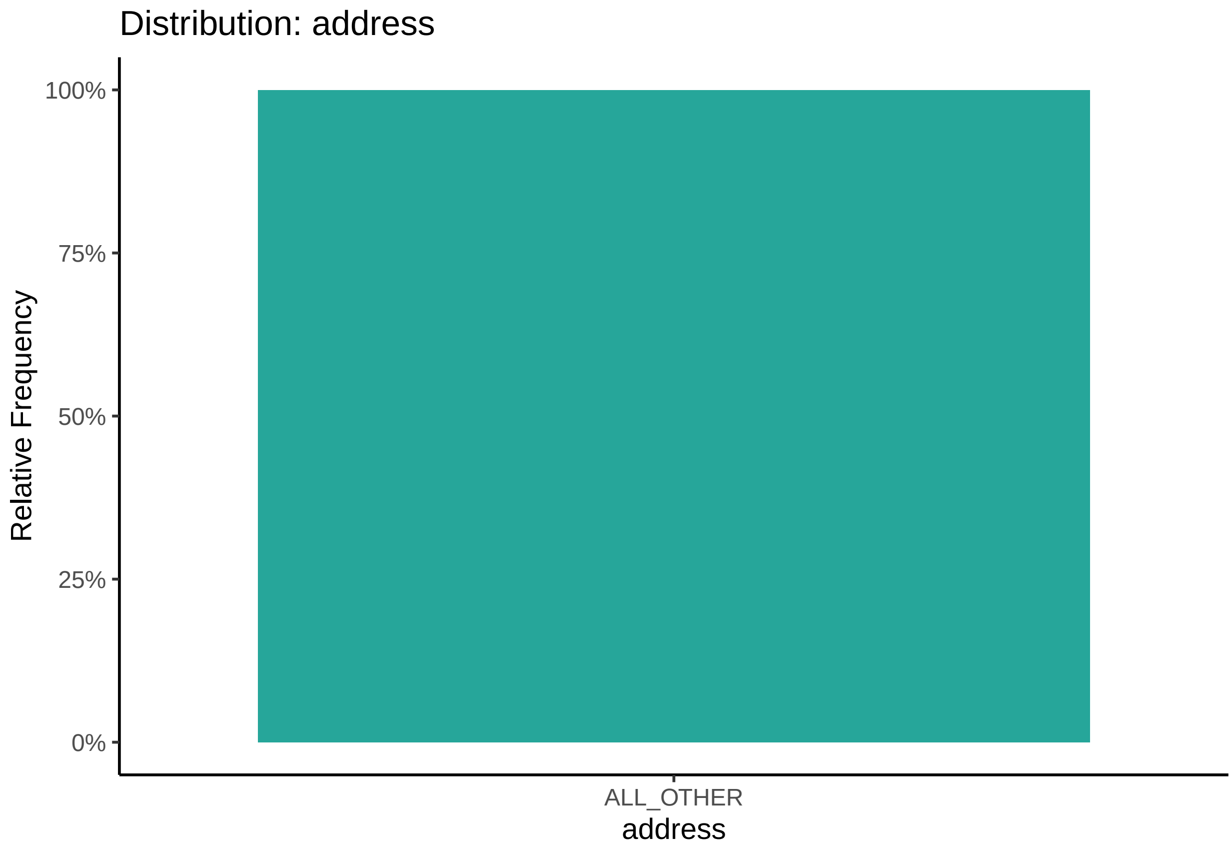

```
##                              Feature Observations FeatureClass FeatureType
## 1                                age         4000      numeric  Continuous
## 2                        customer_id         4000      numeric  Continuous
## 3                 deceased_indicator         4000    character Categorical
## 4                            default         4000    character Categorical
## 5                                DOB         4000         Date  Continuous
## 6                         first_name         4000    character Categorical
## 7                             gender         4000    character Categorical
## 8              job_industry_category         4000    character Categorical
## 9                          job_title         4000    character Categorical
## 10                         last_name         4000    character Categorical
## 11                          owns_car         4000    character Categorical
## 12 past_3_years_bike_related_purchases         4000      numeric  Continuous
## 13                            tenure         4000      numeric  Continuous
## 14                    wealth_segment         4000    character Categorical
##    PercentageMissing PercentageUnique ConstantFeature ZeroSpreadFeature
## 1               2.20             1.40              No                No
## 2               0.00           100.00              No                No
## 3               0.00             0.05              No                No
## 4               0.00             2.55              No                No
## 5               2.20            86.20              No                No
## 6               0.00            78.47              No                No
## 7               0.00             0.08              No                No
## 8               0.00             0.25              No                No
## 9               0.00             4.90              No                No
## 10              0.00            93.15              No                No
```

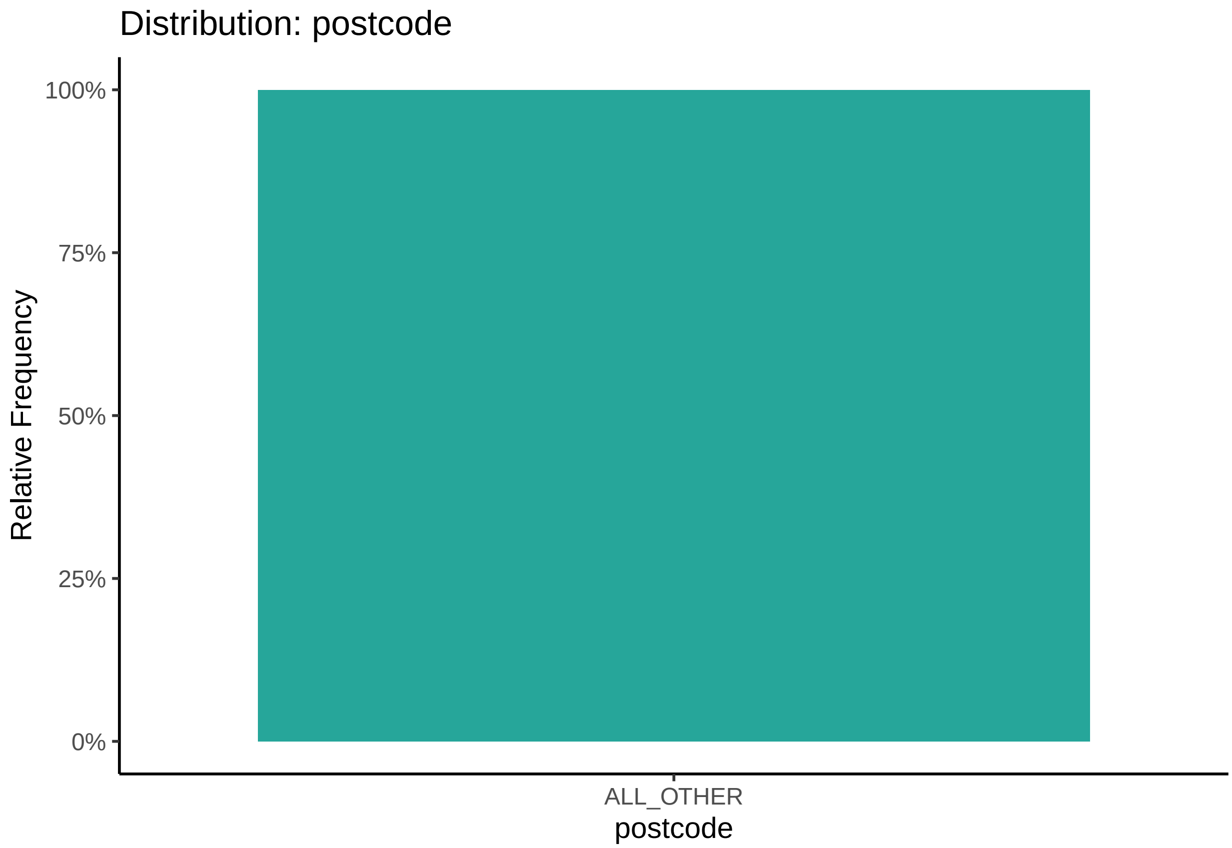
```
## 11             0.00            0.05             No                No
## 12             0.00            2.50             No                No
## 13             2.17            0.58             No                No
## 14             0.00            0.08             No                No
##    LowerOutliers UpperOutliers ImputationValue MinValue FirstQuartile Median
## 1              0             2              43       18         33.00   43.0
## 2              0             0          2000.5        1       1000.75 2000.5
## 3              0             0               N        0          0.00    0.0
## 4              0             0       ALL_OTHER        0          0.00    0.0
## 5              0             0               0        0          0.00    0.0
## 6              0             0       ALL_OTHER        0          0.00    0.0
## 7              0             0          FEMALE        0          0.00    0.0
## 8              0             0   MANUFACTURING        0          0.00    0.0
## 9              0             0       ALL_OTHER        0          0.00    0.0
## 10             0             0       ALL_OTHER        0          0.00    0.0
## 11             0             0             YES        0          0.00    0.0
## 12             0             0              48        0         24.00   48.0
## 13             0             0              11        1          6.00   11.0
## 14             0             0   MASS CUSTOMER        0          0.00    0.0
##       Mean          Mode ThirdQuartile MaxValue LowerOutlierValue
## 1    42.94            42         52.00       89               4.5
## 2  2000.50             1       3000.25     4000           -1998.5
## 3     0.00             N          0.00        0               0.0
## 4     0.00                        0.00        0               0.0
## 5     0.00    1978-01-30          0.00        0               0.0
## 6     0.00           MAX          0.00        0               0.0
## 7     0.00        FEMALE          0.00        0               0.0
## 8     0.00 MANUFACTURING          0.00        0               0.0
## 9     0.00                        0.00        0               0.0
## 10    0.00                        0.00        0               0.0
## 11    0.00           YES          0.00        0               0.0
## 12   48.89            16         73.00       99             -49.5
## 13   10.66             7         15.00       22              -7.5
## 14    0.00 MASS CUSTOMER          0.00        0               0.0
##    UpperOutlierValue PredictivePowerPercentage PredictivePower
## 1               80.5                         1             Low
## 2             5999.5                         4             Low
## 3                0.0                         0             Low
## 4                0.0                         0             Low
## 5                0.0                         0             Low
## 6                0.0                         0             Low
## 7                0.0                         2             Low
## 8                0.0                         2             Low
## 9                0.0                         0             Low
## 10               0.0                         1             Low
## 11               0.0                         1             Low
## 12             146.5                         5             Low
## 13              28.5                         2             Low
## 14               0.0                         0             Low
```

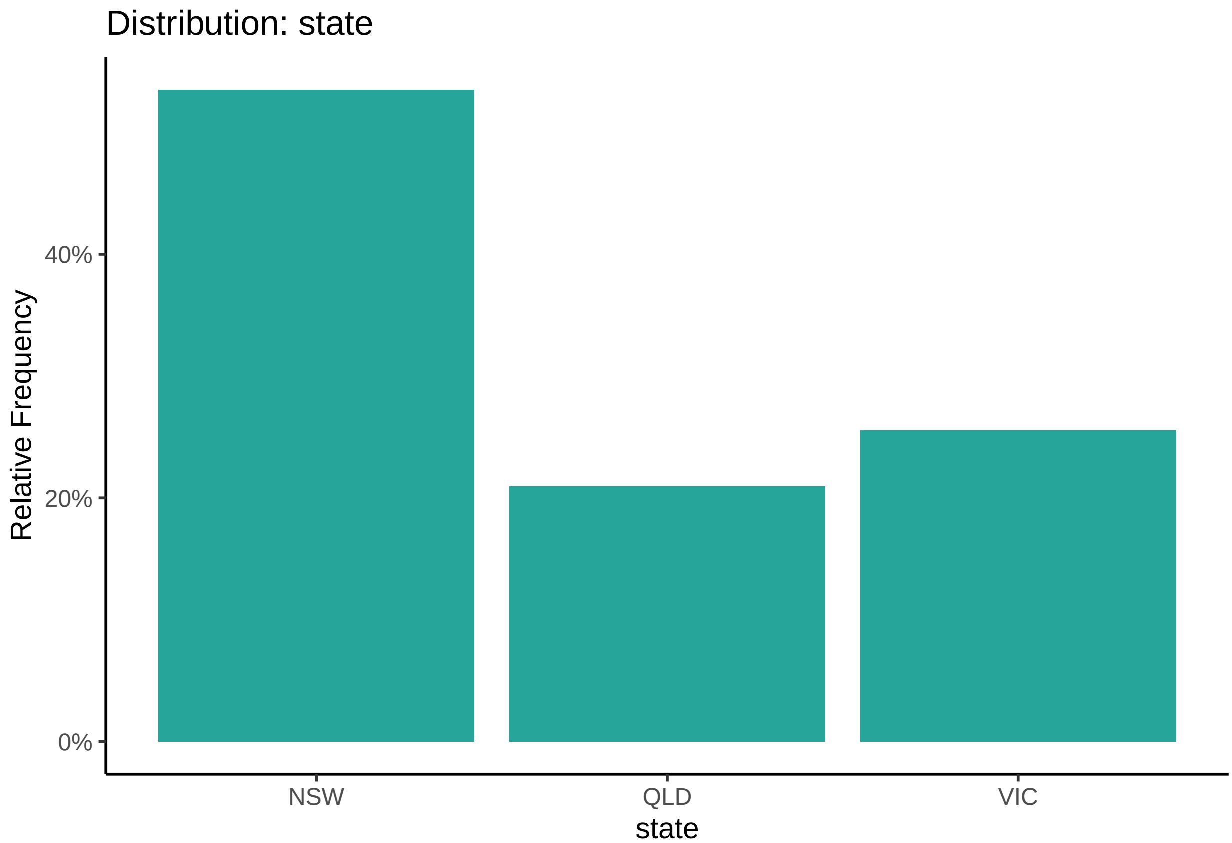
```
## autoEDA | Setting color theme
## autoEDA | Removing constant features
## autoEDA | 1 constant feature removed
## autoEDA | 0 zero spread features removed
```

```
## autoEDA | Removing features containing majority missing values
## autoEDA | 0 majority missing features removed
## autoEDA | Cleaning data
## autoEDA | Correcting sparse categorical feature levels
## autoEDA | Performing univariate analysis
## autoEDA | Visualizing data
```




Distribution: address
100%
75%
50%
25%
0%
Relative Frequency
ALL_OTHER
address


Distribution: postcode
Relative Frequency
100%
75%
50%
25%
0%
ALL_OTHER
postcode

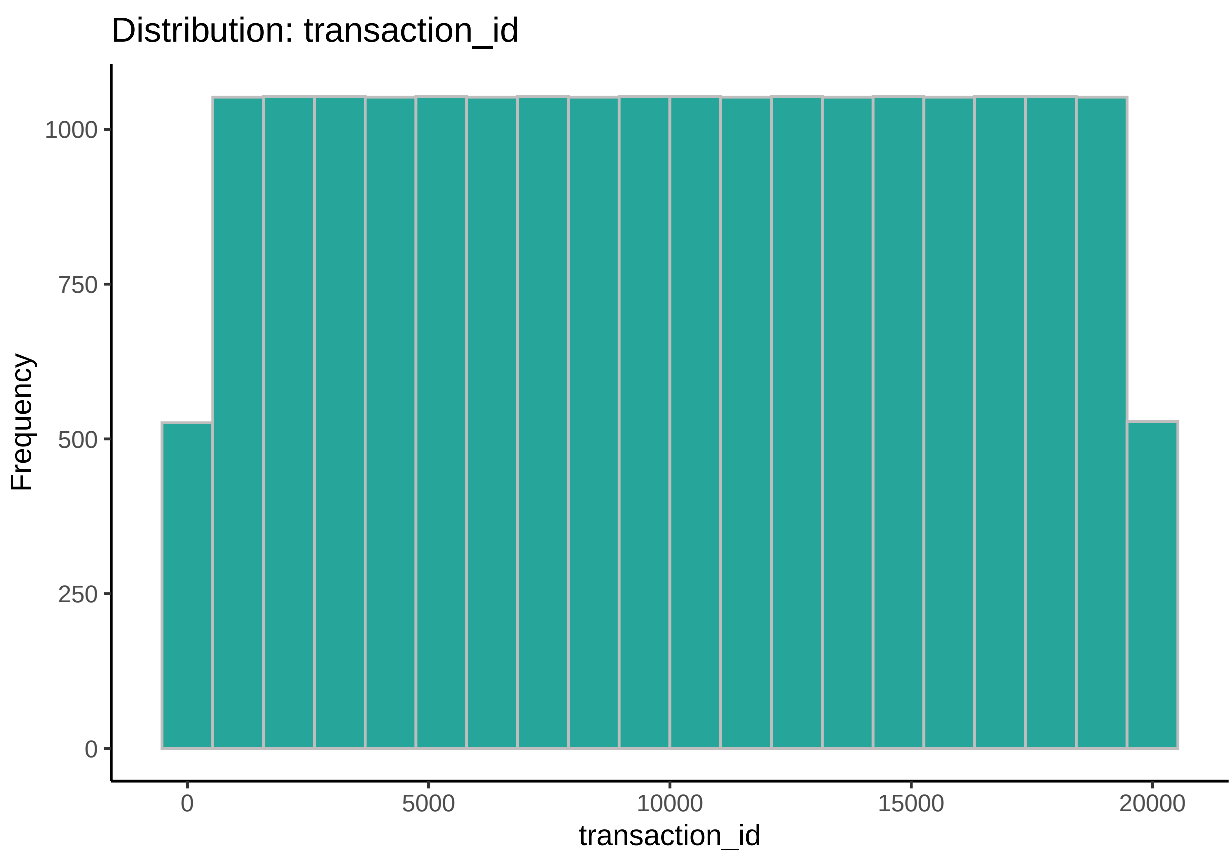


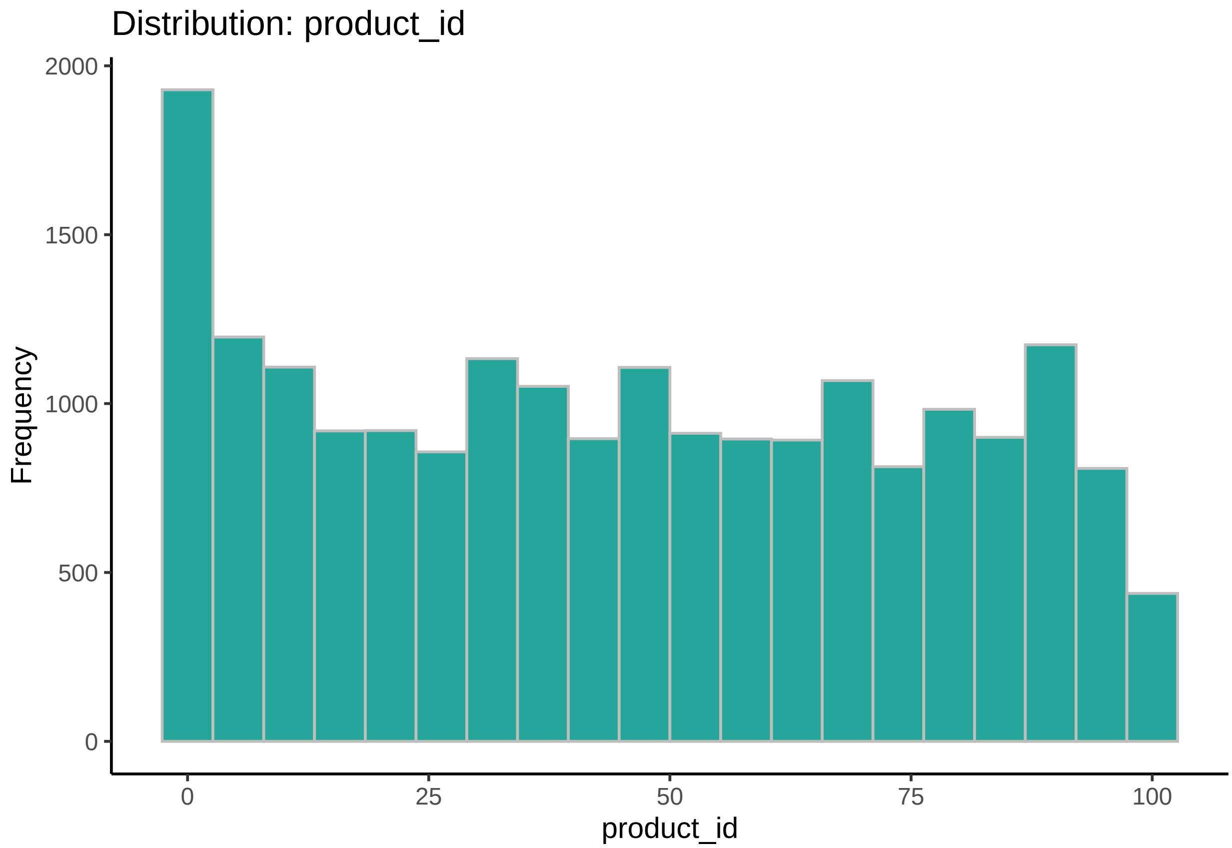
Distribution: state
Relative Frequency
40%
20%
0%
NSW
QLD
VIC
state

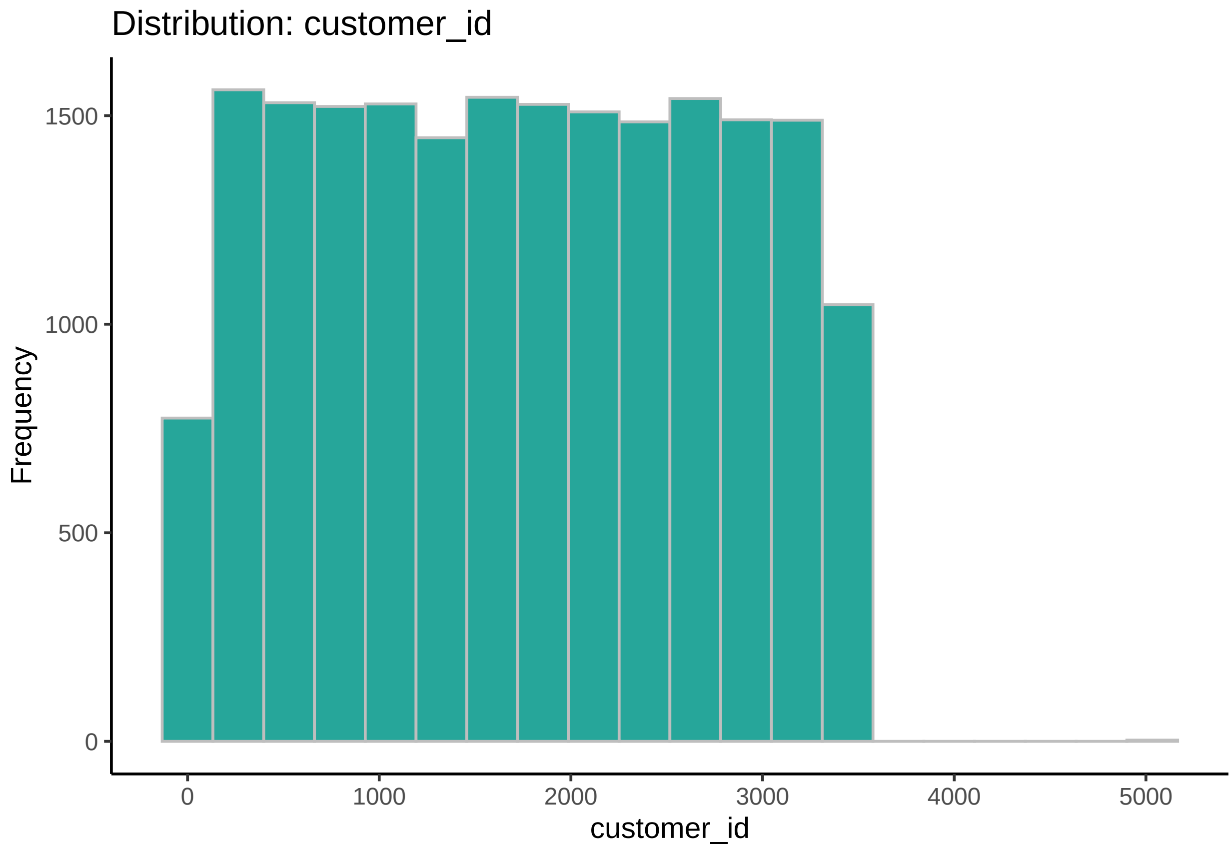

```
##               Feature Observations FeatureClass FeatureType PercentageMissing
## 1         customer_id         3999      numeric  Continuous                 0
## 2             address         3999    character Categorical                 0
## 3            postcode         3999    character Categorical                 0
## 4               state         3999    character Categorical                 0
## 5  property_valuation         3999    character Categorical                 0
##   PercentageUnique ConstantFeature ZeroSpreadFeature LowerOutliers
## 1           100.00              No                No             0
## 2            99.92              No                No             0
## 3            21.83              No                No             0
## 4             0.08              No                No             0
## 5             0.30              No                No             0
##   UpperOutliers ImputationValue MinValue FirstQuartile Median    Mean
## 1             0            2004        1        1004.5   2004 2003.99
## 2             0       ALL_OTHER        0           0.0      0    0.00
## 3             0       ALL_OTHER        0           0.0      0    0.00
## 4             0             NSW        0           0.0      0    0.00
## 5             0               9        0           0.0      0    0.00
##                        Mode ThirdQuartile MaxValue LowerOutlierValue
## 1                         1        3003.5     4003             -1994
## 2 3 MARINERS COVE TERRACE             0.0        0                 0
## 3                      2170           0.0        0                 0
## 4                       NSW           0.0        0                 0
## 5                         9           0.0        0                 0
##   UpperOutlierValue
## 1              6002
```

```
## 2                 0
## 3                 0
## 4                 0
## 5                 0
```

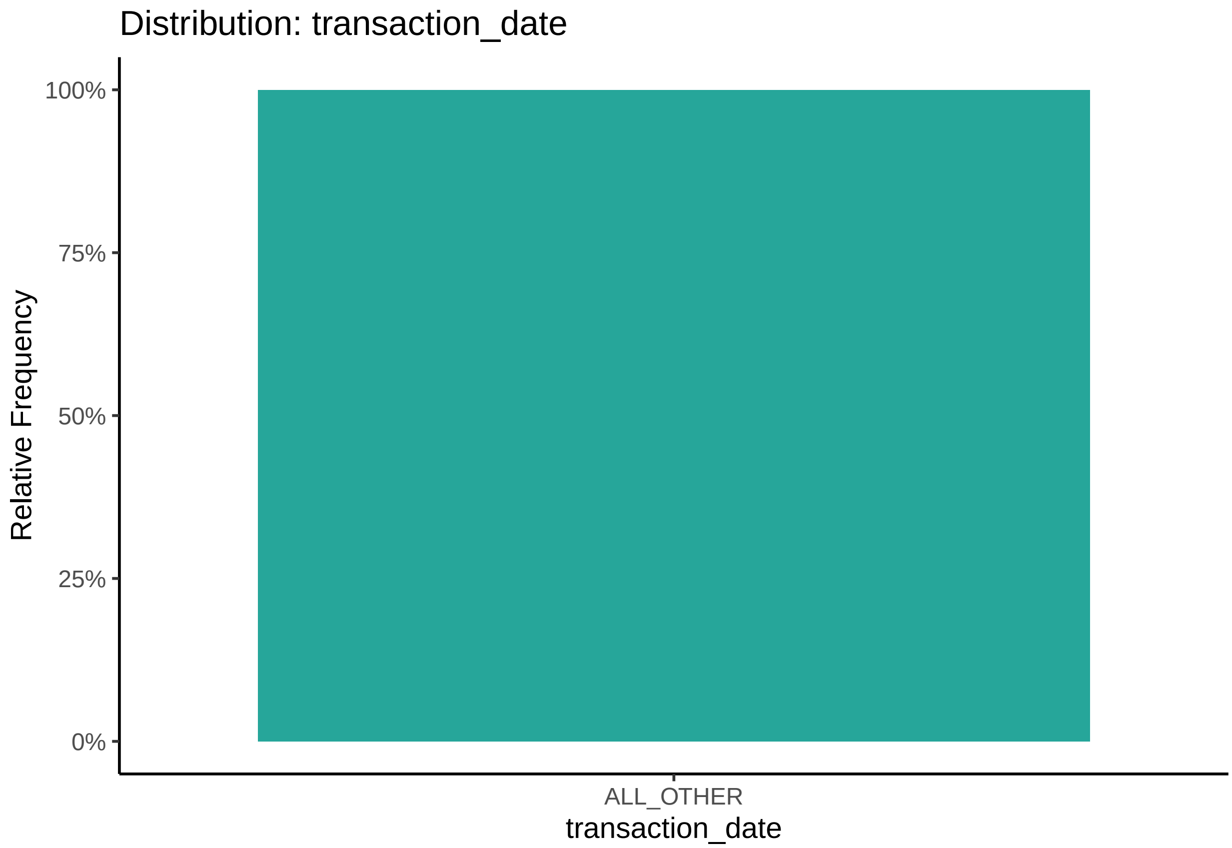
```
## autoEDA | Setting color theme
## autoEDA | Removing constant features
## autoEDA | 0 constant features removed
## autoEDA | 0 zero spread features removed
## autoEDA | Removing features containing majority missing values
## autoEDA | 0 majority missing features removed
## autoEDA | Cleaning data
## autoEDA | Correcting sparse categorical feature levels
## autoEDA | Performing univariate analysis
## autoEDA | Visualizing data
```



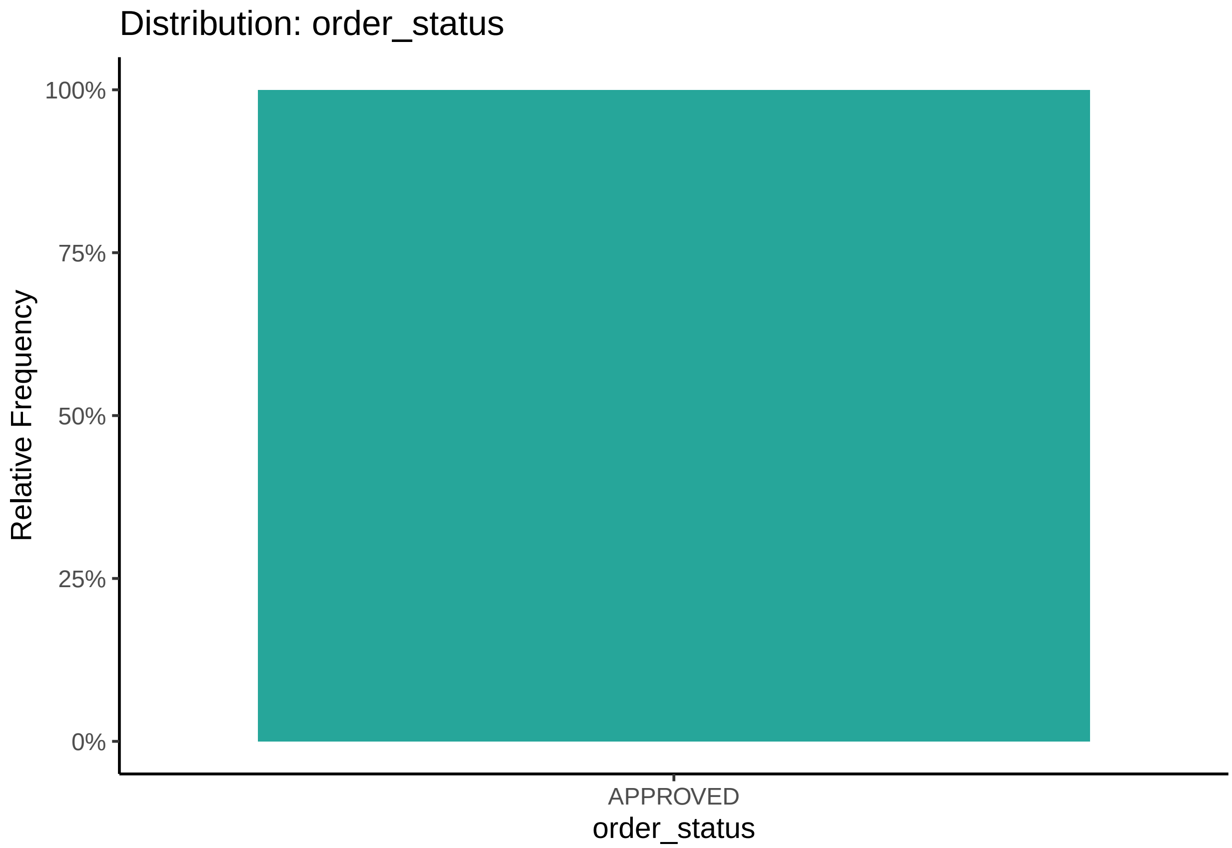

Distribution: product_id
Frequency
2000
1500
1000
500
0
0
25
50
75
100
product_id

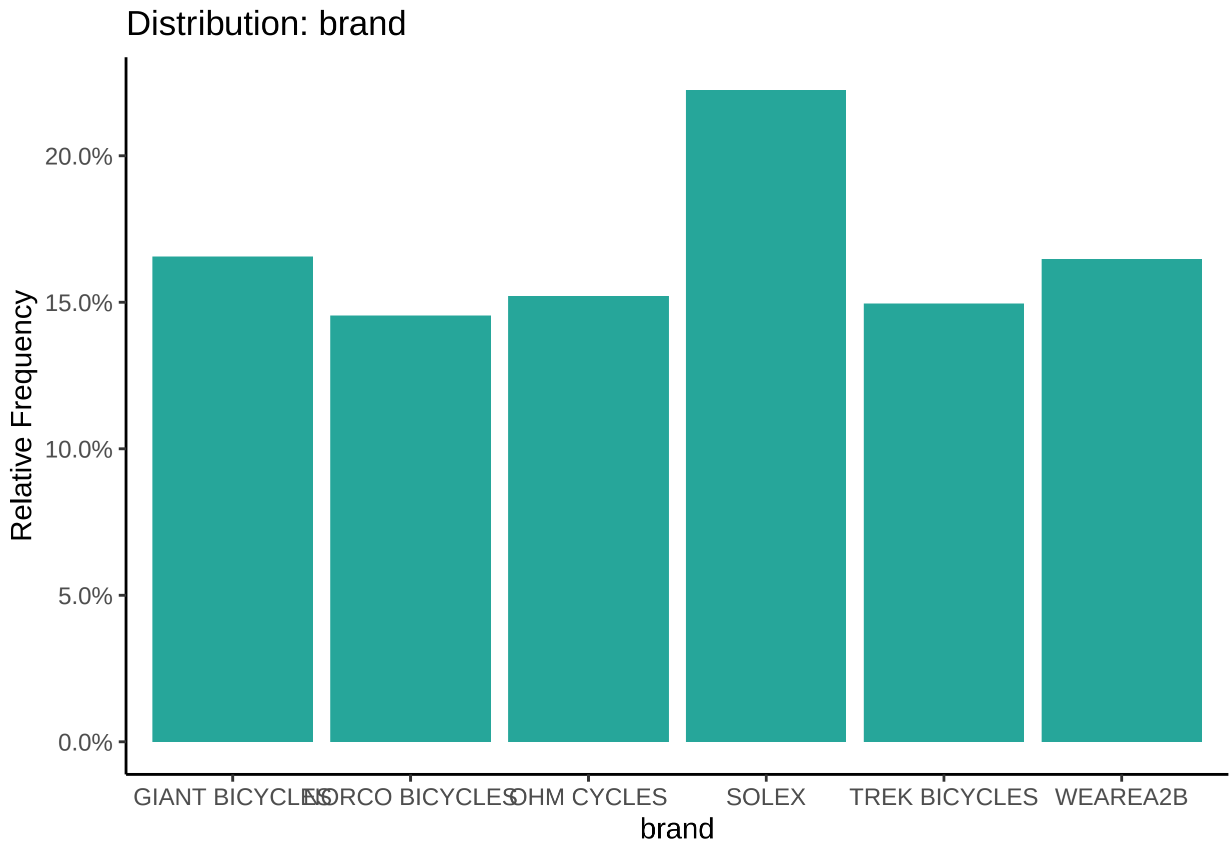


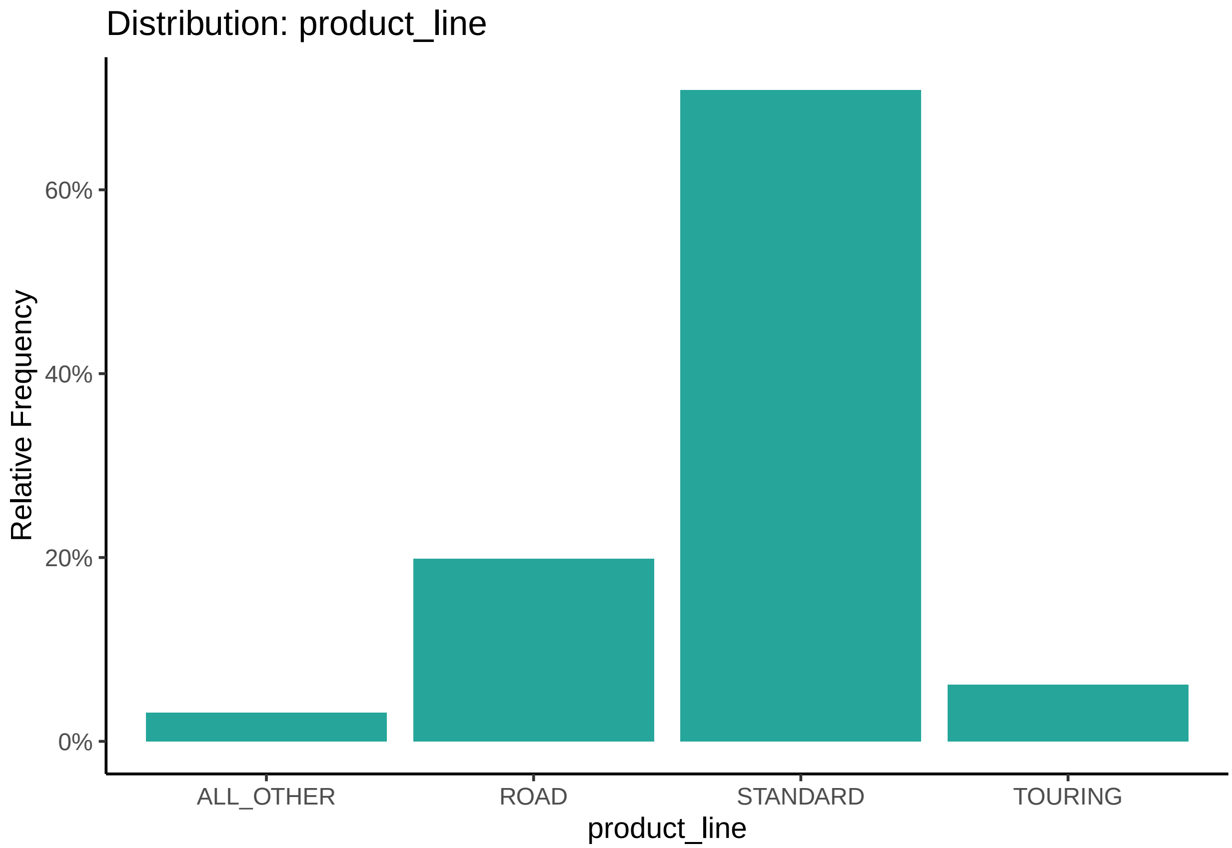
Distribution: customer_id
Frequency
1500
1000
500
0
0
1000
2000
3000
4000
5000
customer_id

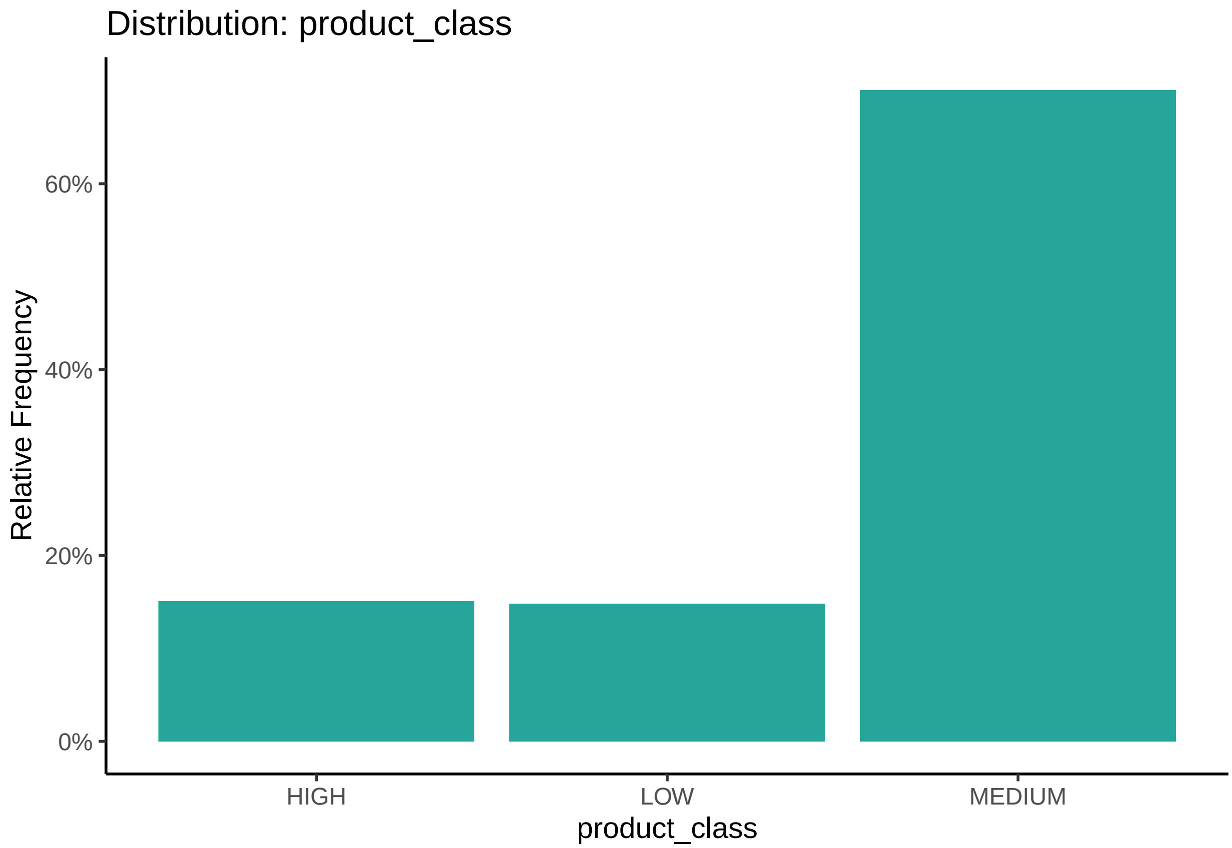

Distribution: transaction_date
100%
75%
50%
25%
0%
Relative Frequency
ALL_OTHER
transaction_date


Distribution: online_order
Relative Frequency
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
FALSE
TRUE
online_order

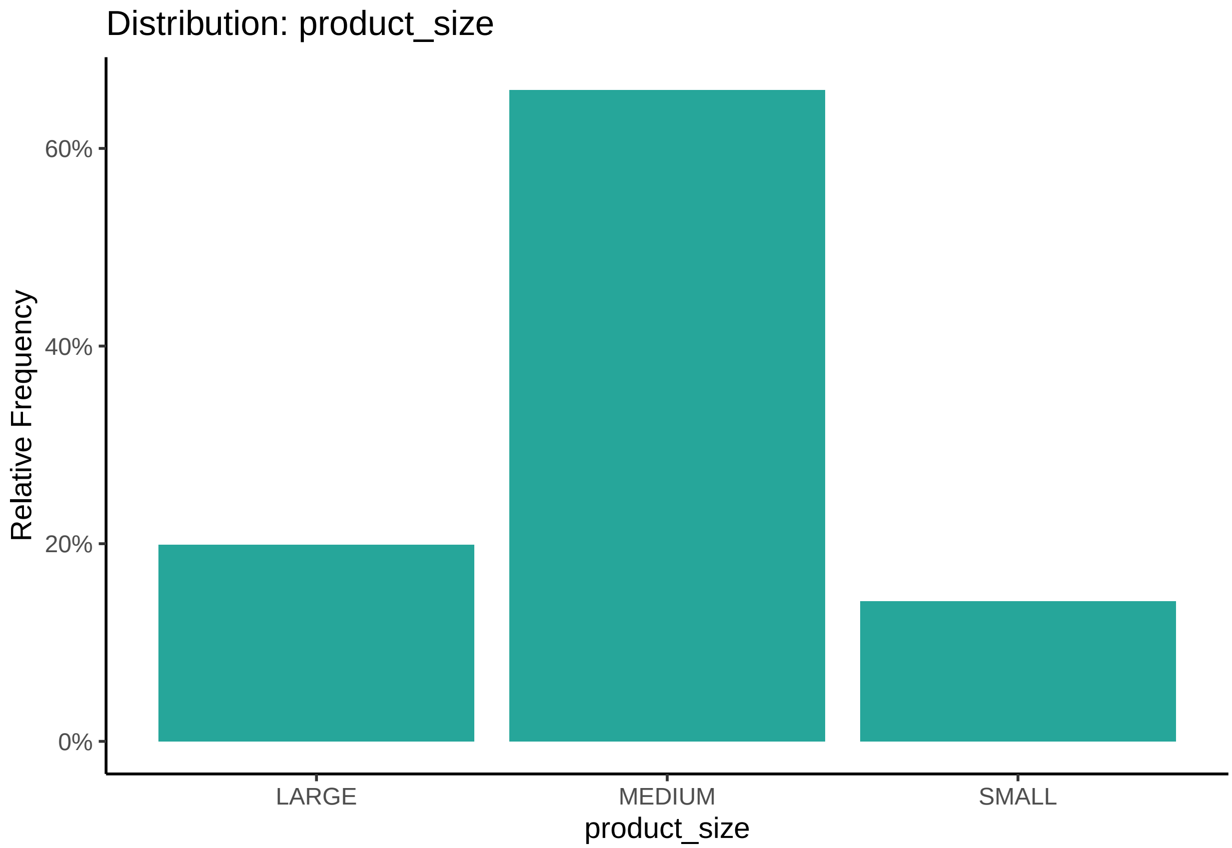

Distribution: order_status
Relative Frequency
100%
75%
50%
25%
0%
APPROVED
order_status


Distribution: brand
Relative Frequency
20.0%
15.0%
10.0%
5.0%
0.0%
GIANT BICYCLES
NORCO BICYCLES
OHM CYCLES
SOLEX
TREK BICYCLES
WEAREA2B
brand

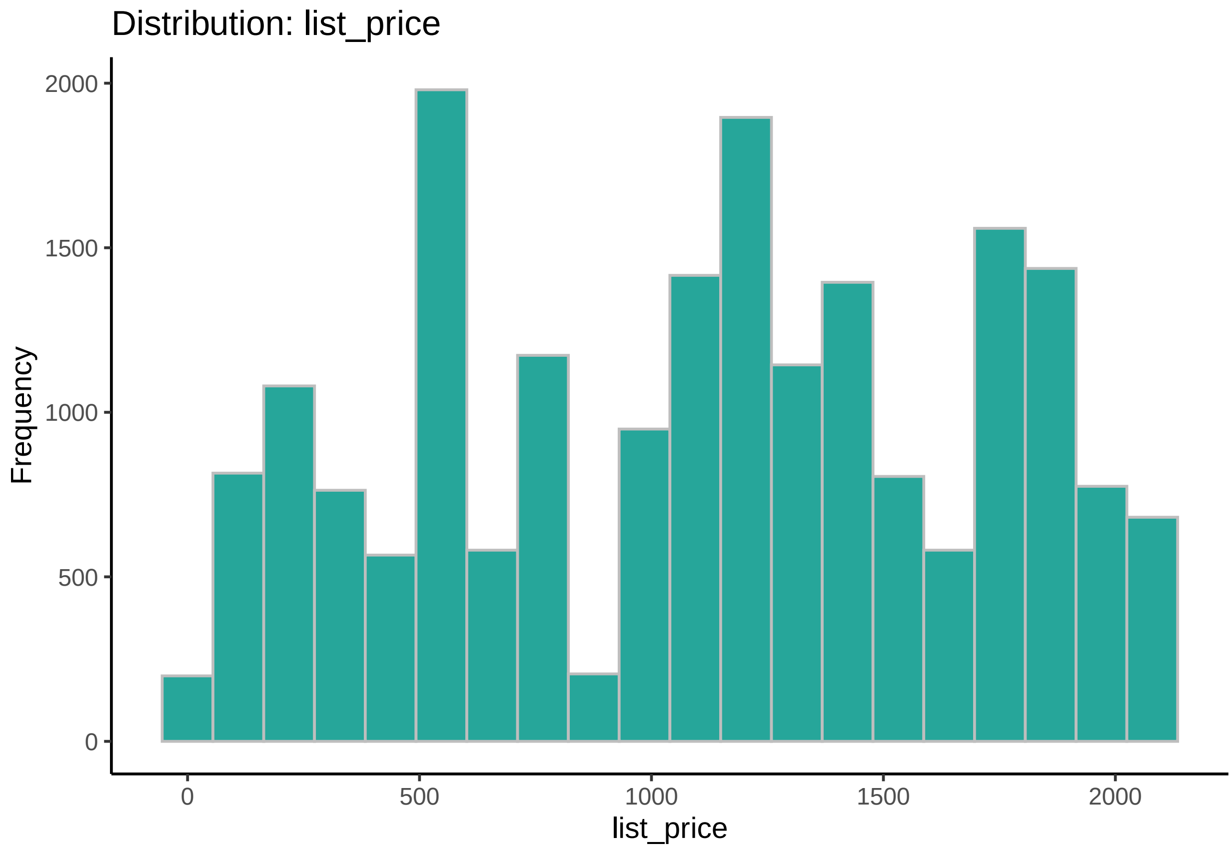

Distribution: product_line
Relative Frequency
60%
40%
20%
0%
ALL_OTHER
ROAD
STANDARD
TOURING
product_line

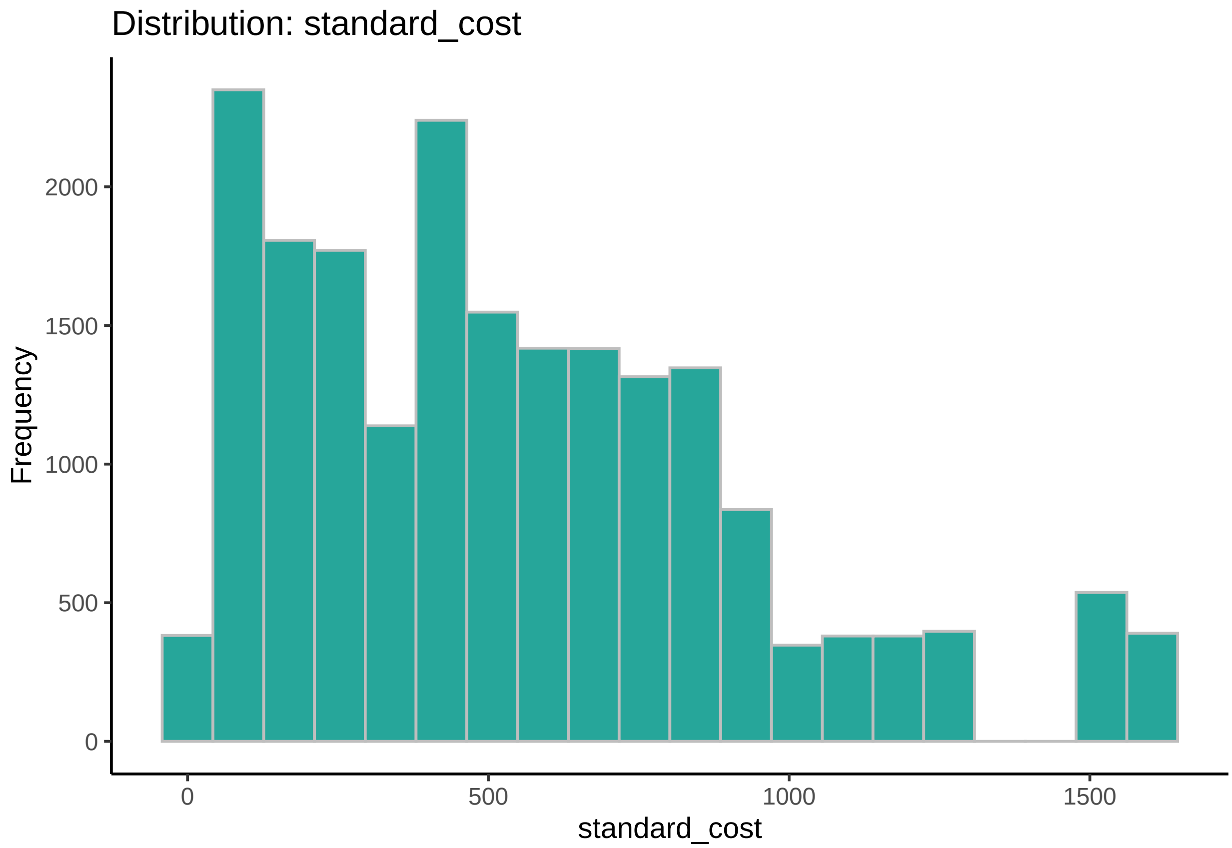


Distribution: product_class
Relative Frequency
60%
40%
20%
0%
HIGH
LOW
MEDIUM
product_class

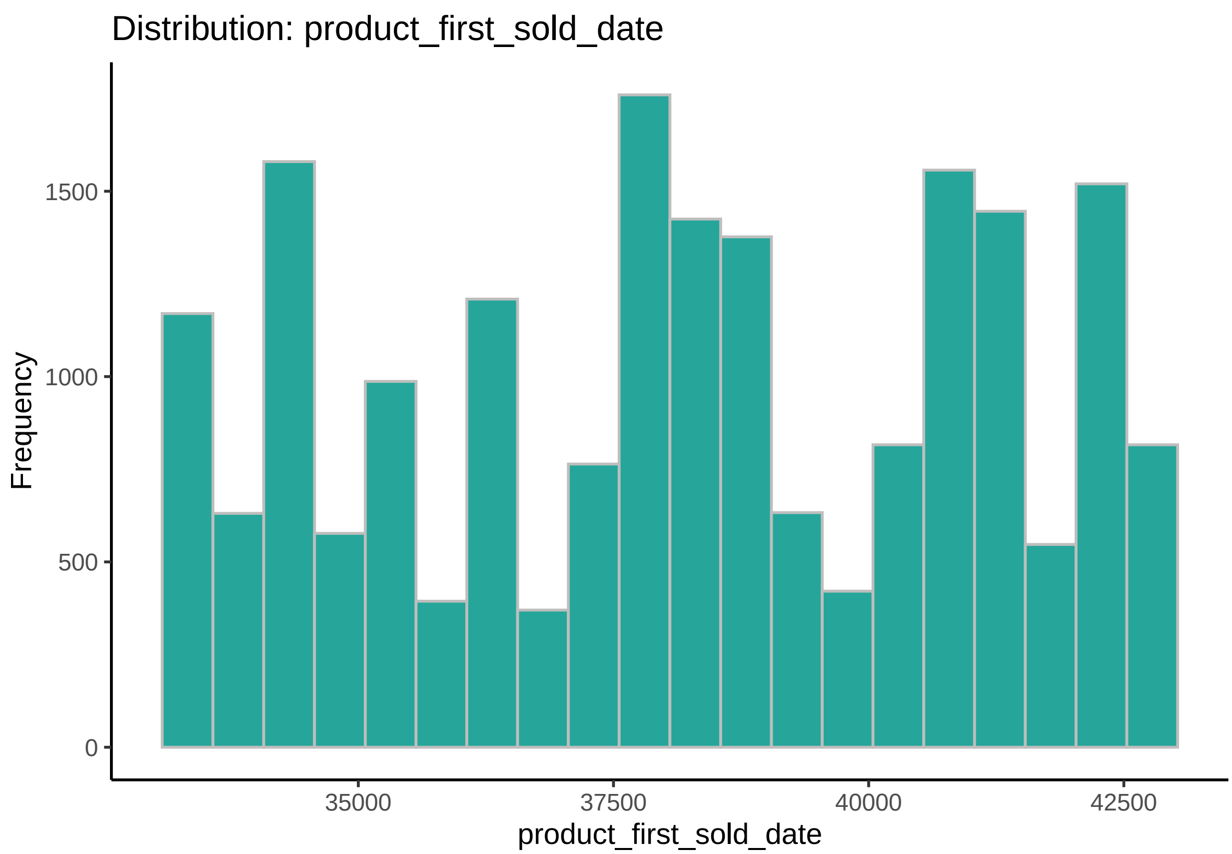


Distribution: product_size
Relative Frequency
60%
40%
20%
0%
LARGE
MEDIUM
SMALL
product_size


Distribution: list_price
Frequency
2000
1500
1000
500
0
0
500
1000
1500
2000
list_price



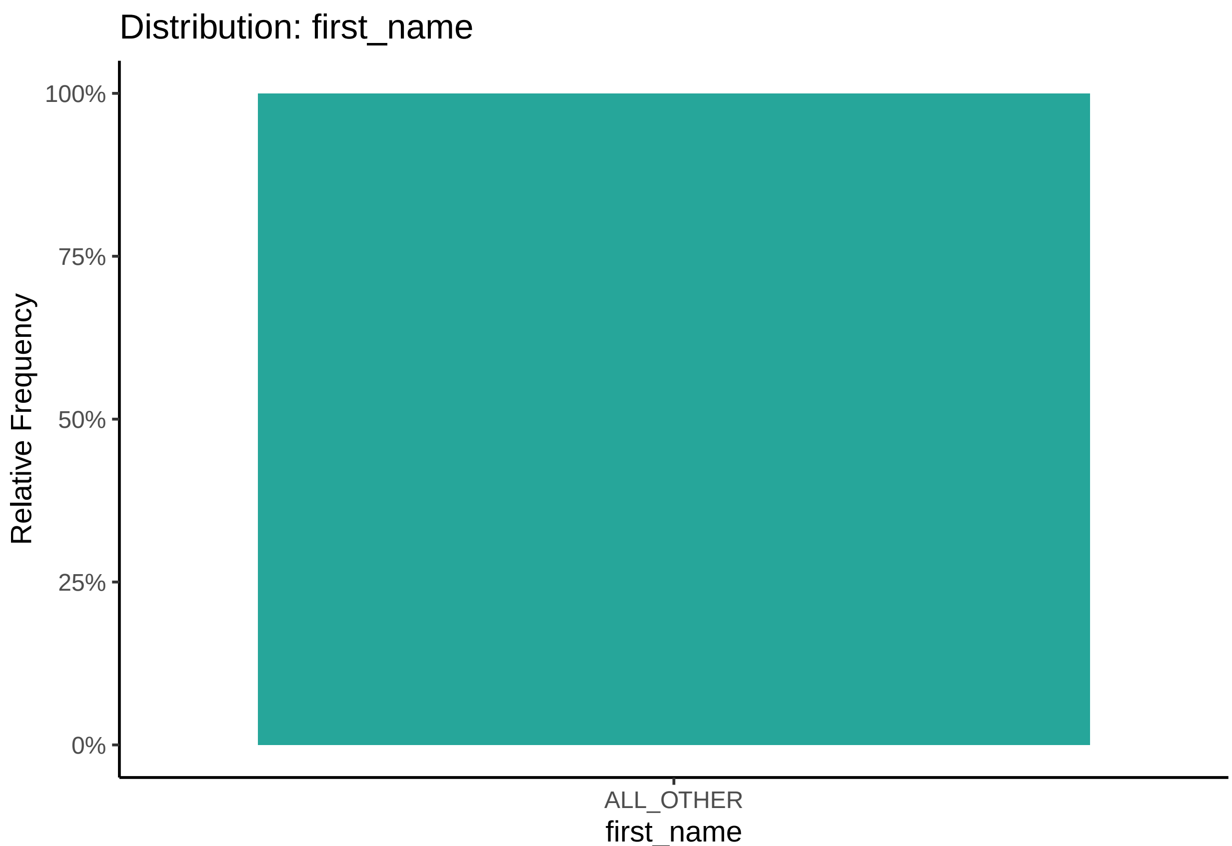
Distribution: standard_cost
Frequency
2000
1500
1000
500
0
0
500
1000
1500
standard_cost

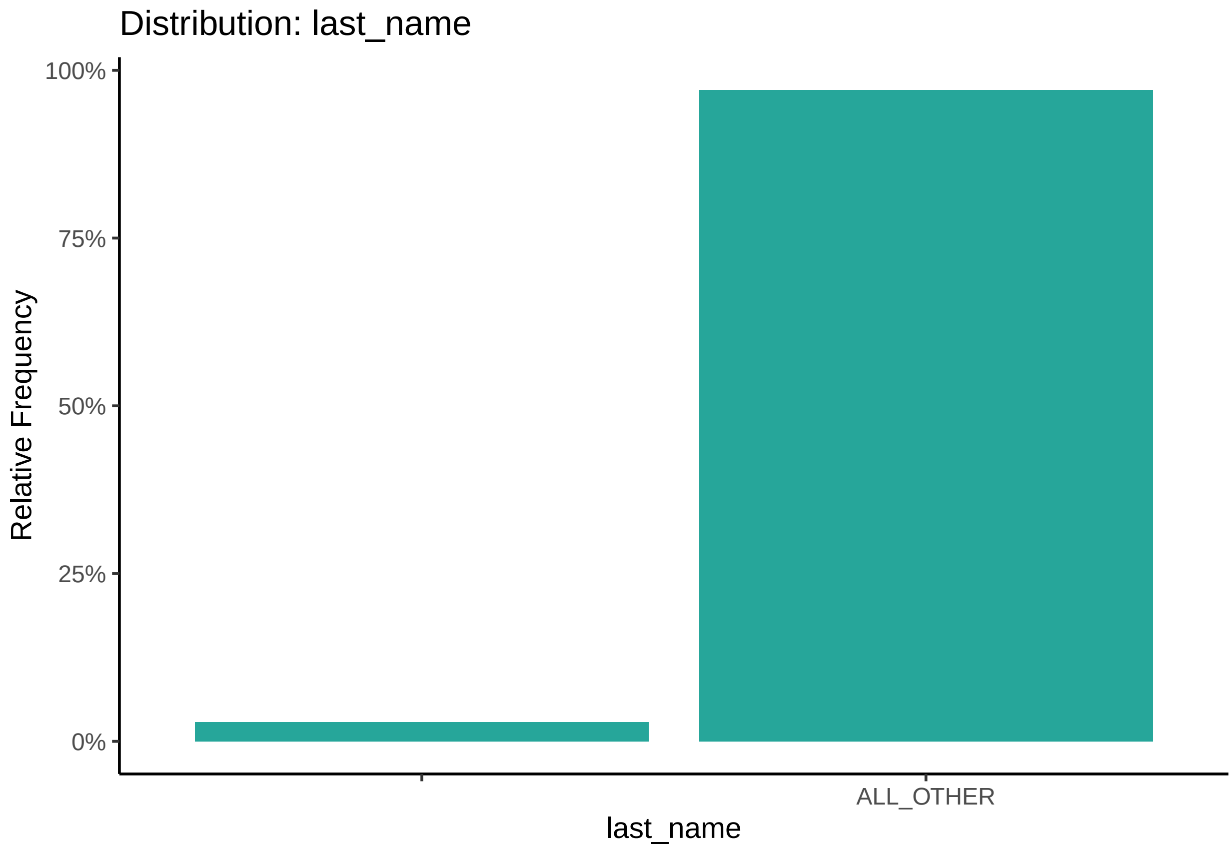


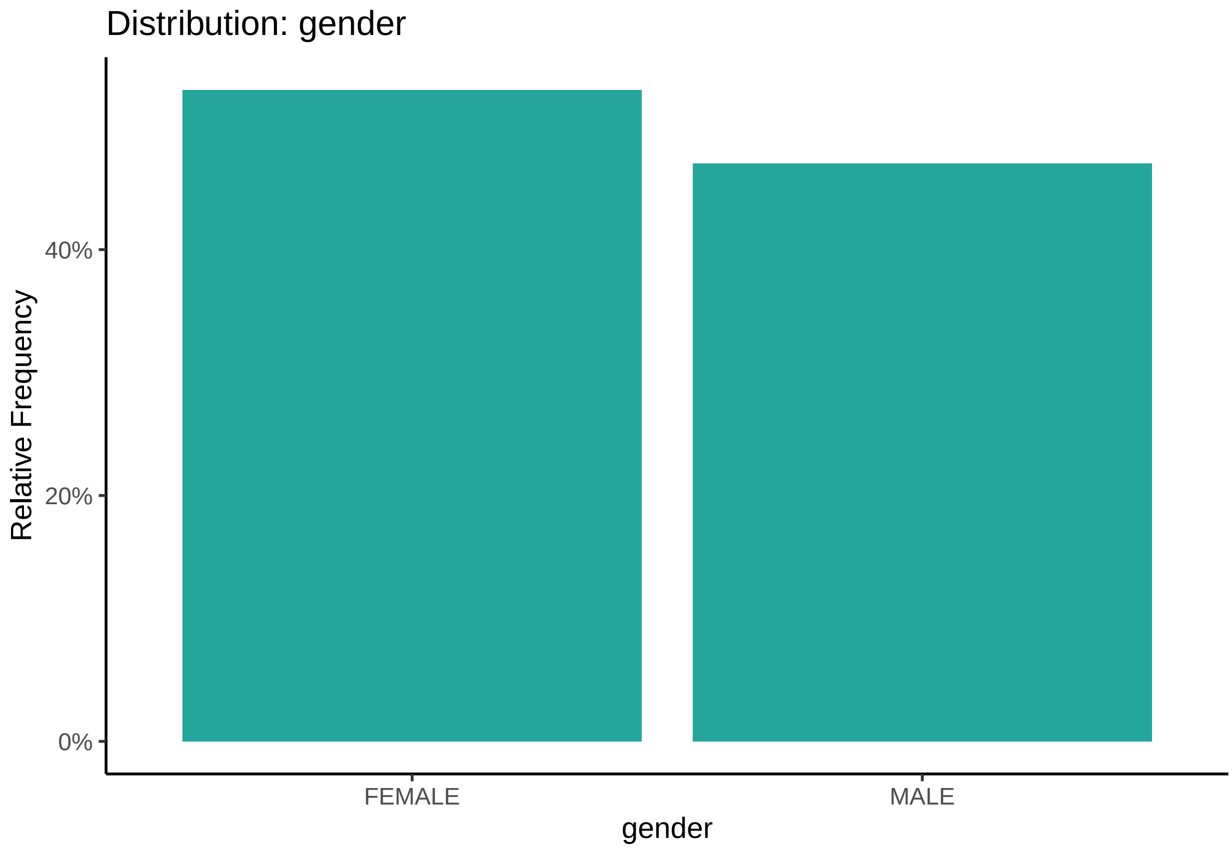

```
##                    Feature Observations FeatureClass FeatureType
## 1           transaction_id        20000      numeric  Continuous
## 2               product_id        20000      numeric  Continuous
## 3              customer_id        20000      numeric  Continuous
## 4         transaction_date        20000    character Categorical
## 5             online_order        20000    character Categorical
## 6             order_status        20000    character Categorical
## 7                    brand        20000    character Categorical
## 8             product_line        20000    character Categorical
## 9            product_class        20000    character Categorical
## 10            product_size        20000    character Categorical
## 11              list_price        20000      numeric  Continuous
## 12           standard_cost        20000      numeric  Continuous
## 13 product_first_sold_date        20000      numeric  Continuous
##    PercentageMissing PercentageUnique ConstantFeature ZeroSpreadFeature
## 1               0.00           100.00              No                No
## 2               0.00             0.50              No                No
## 3               0.00            17.47              No                No
## 4               0.00             1.82              No                No
## 5               1.80             0.01              No                No
## 6               0.00             0.01              No                No
## 7               0.00             0.04              No                No
## 8               0.00             0.03              No                No
## 9               0.00             0.02              No                No
## 10              0.00             0.02              No                No
## 11              0.00             1.48              No                No
```

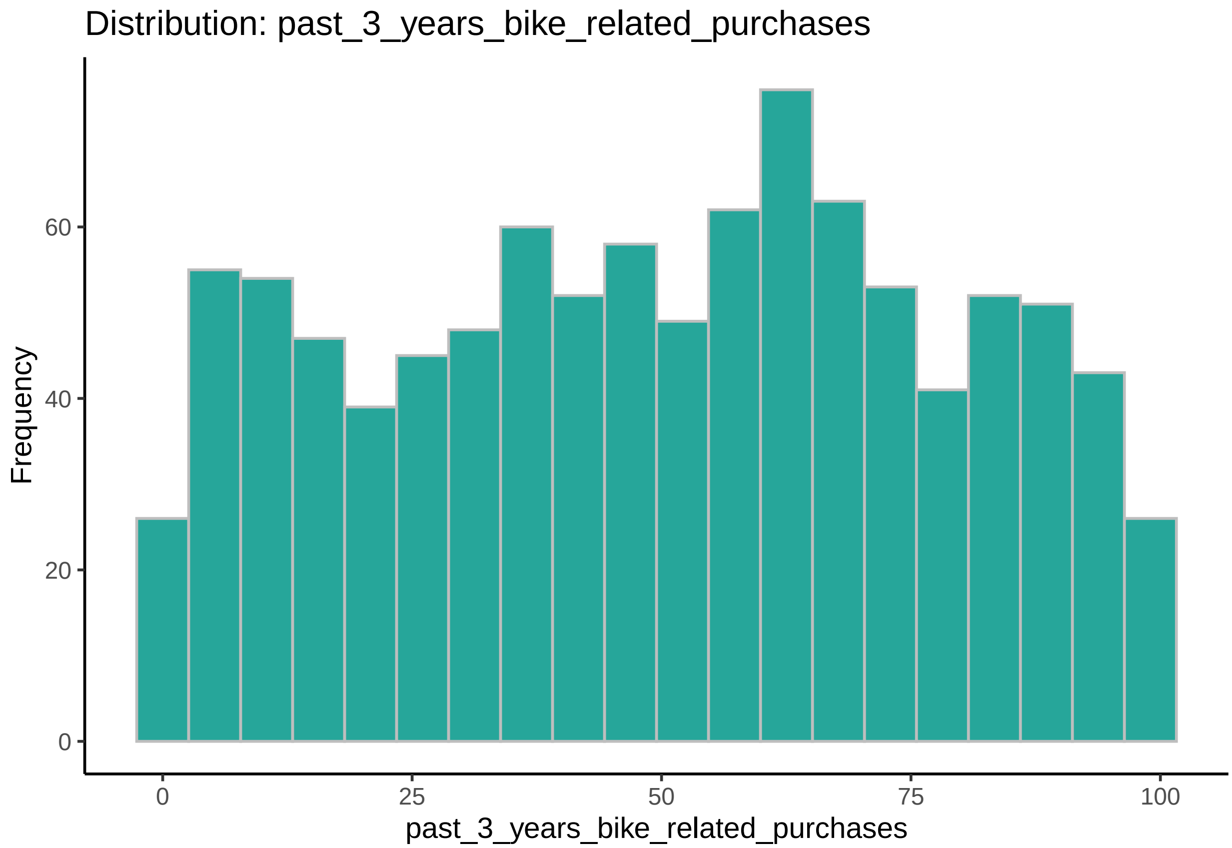
```
## 12             0.98             0.52              No                 No
## 13             0.98             0.50              No                 No
##    LowerOutliers UpperOutliers ImputationValue MinValue FirstQuartile   Median
## 1              0             0         10000.5     1.00       5000.75 10000.50
## 2              0             0              44     0.00         18.00    44.00
## 3              0             0            1736     1.00        857.75  1736.00
## 4              0             0       ALL_OTHER     0.00          0.00     0.00
## 5              0             0         MISSING     0.00          0.00     0.00
## 6              0             0        APPROVED     0.00          0.00     0.00
## 7              0             0           SOLEX     0.00          0.00     0.00
## 8              0             0       ALL_OTHER     0.00          0.00     0.00
## 9              0             0          MEDIUM     0.00          0.00     0.00
## 10             0             0          MEDIUM     0.00          0.00     0.00
## 11             0             0         1163.89    12.01        575.27  1163.89
## 12             0           195          507.58     7.21        215.14   507.58
## 13             0             0           38216 33259.00      35667.00 38216.00
##        Mean       Mode ThirdQuartile MaxValue LowerOutlierValue
## 1  10000.50          1      15000.25 20000.00         -9998.500
## 2     45.36          0         72.00   100.00           -63.000
## 3   1738.25       1068       2613.00  5034.00         -1775.125
## 4      0.00 2017-02-14          0.00     0.00             0.000
## 5      0.00       TRUE          0.00     0.00             0.000
## 6      0.00   APPROVED          0.00     0.00             0.000
## 7      0.00      SOLEX          0.00     0.00             0.000
## 8      0.00   STANDARD          0.00     0.00             0.000
## 9      0.00     MEDIUM          0.00     0.00             0.000
## 10     0.00     MEDIUM          0.00     0.00             0.000
## 11  1107.83    2091.47       1635.30  2091.47         -1014.775
## 12   556.05     388.92        795.10  1759.85          -654.800
## 13 38199.78      33879      40672.00 42710.00         28159.500
##    UpperOutlierValue
## 1          29999.500
## 2            153.000
## 3           5245.875
## 4              0.000
## 5              0.000
## 6              0.000
## 7              0.000
## 8              0.000
## 9              0.000
## 10             0.000
## 11          3225.345
## 12          1665.040
## 13         48179.500

## autoEDA | Setting color theme
## autoEDA | Removing constant features
## autoEDA | 2 constant features removed
## autoEDA | 0 zero spread features removed
## autoEDA | Removing features containing majority missing values
## autoEDA | 0 majority missing features removed
## autoEDA | Cleaning data
## autoEDA | Correcting sparse categorical feature levels
## autoEDA | Performing univariate analysis
```

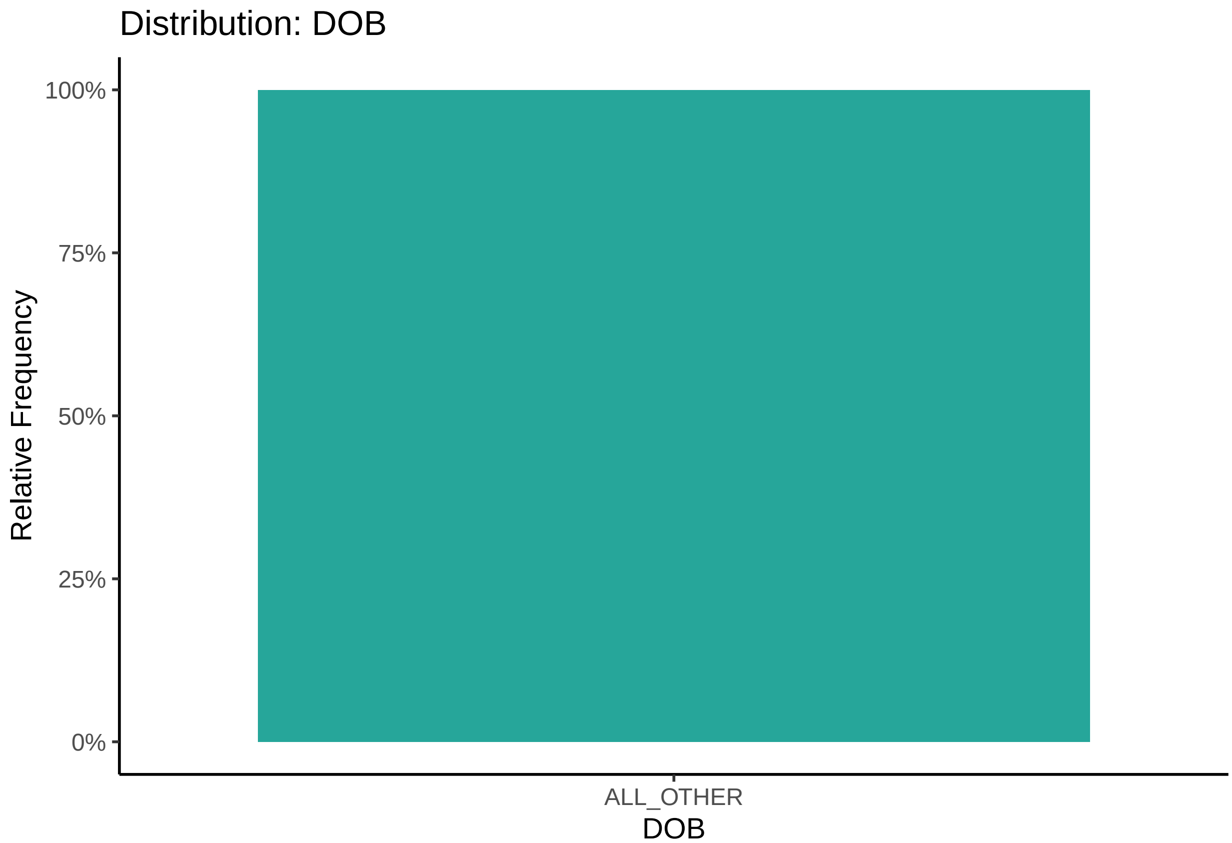
```
## autoEDA | Visualizing data
```

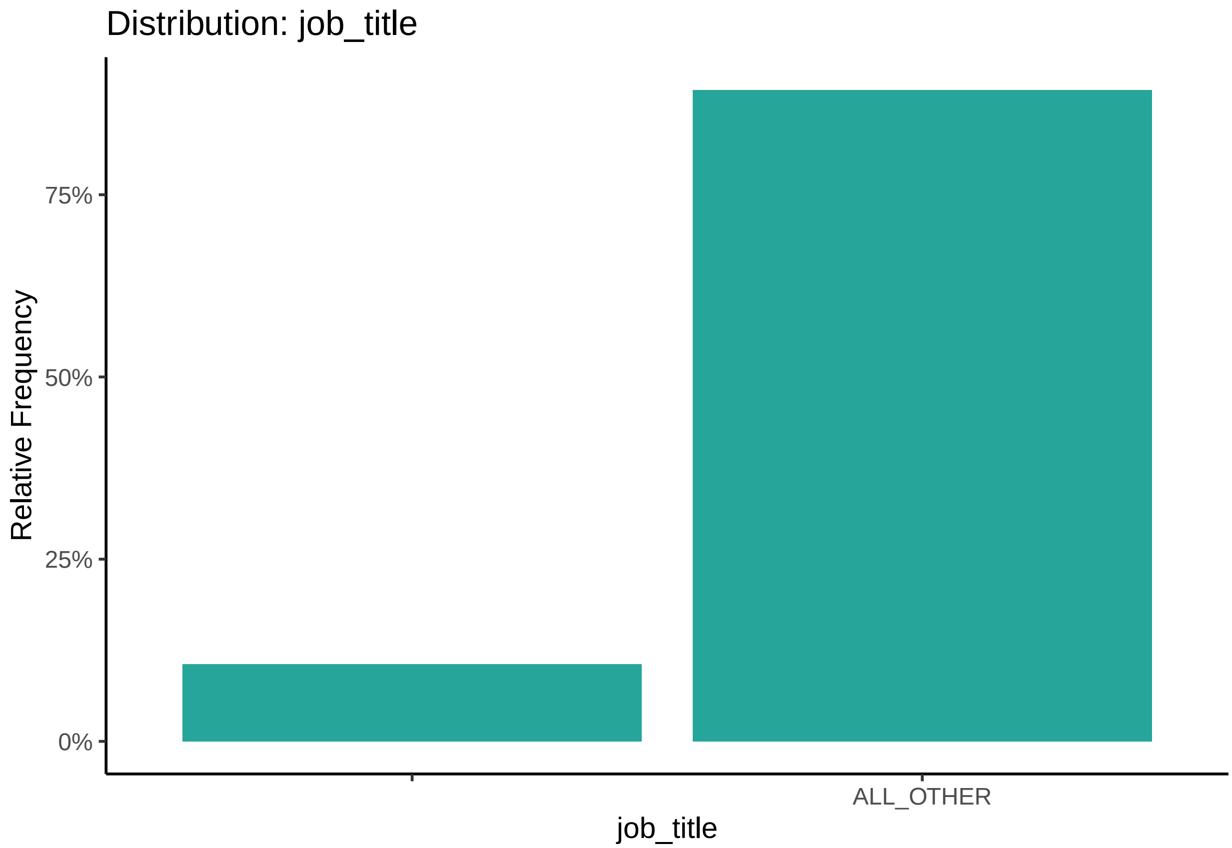


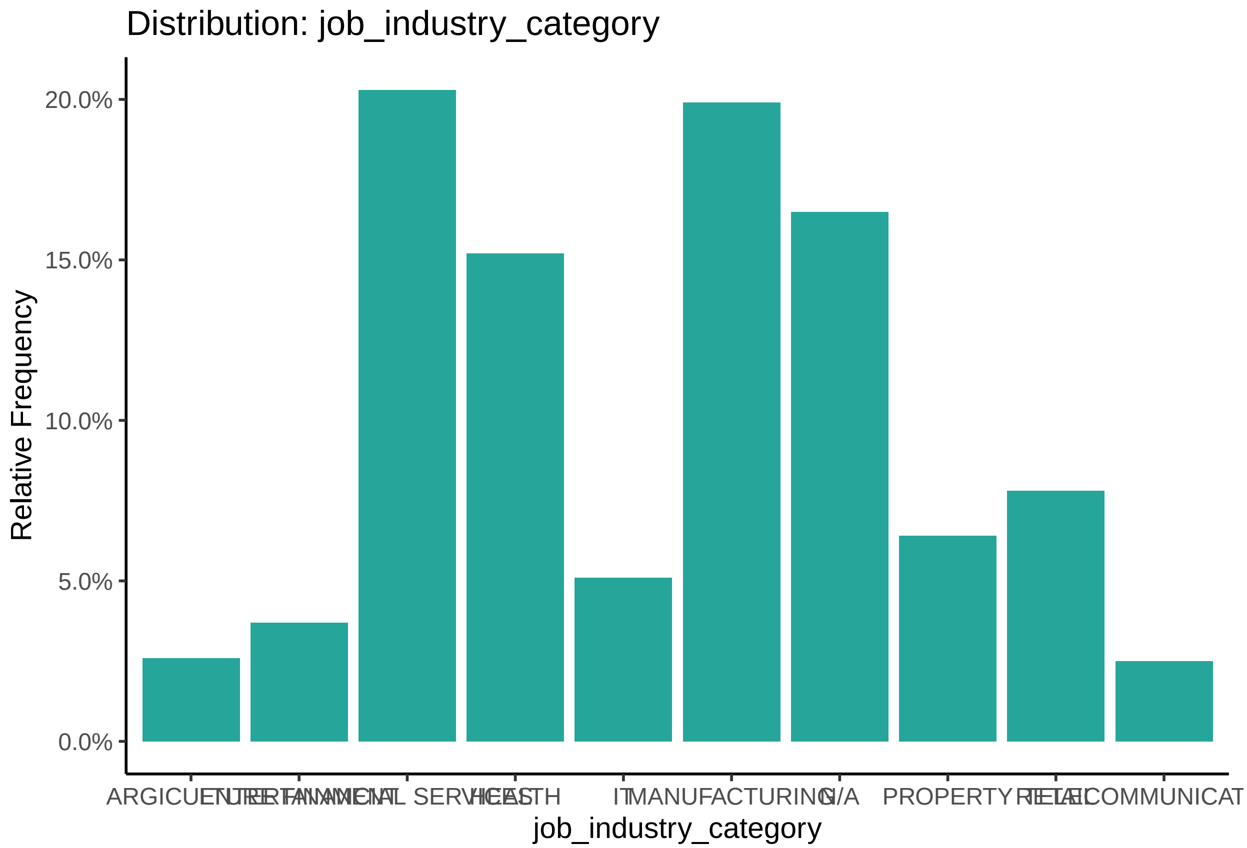

Distribution: last_name
100%
75%
50%
25%
0%
Relative Frequency
ALL_OTHER
last_name

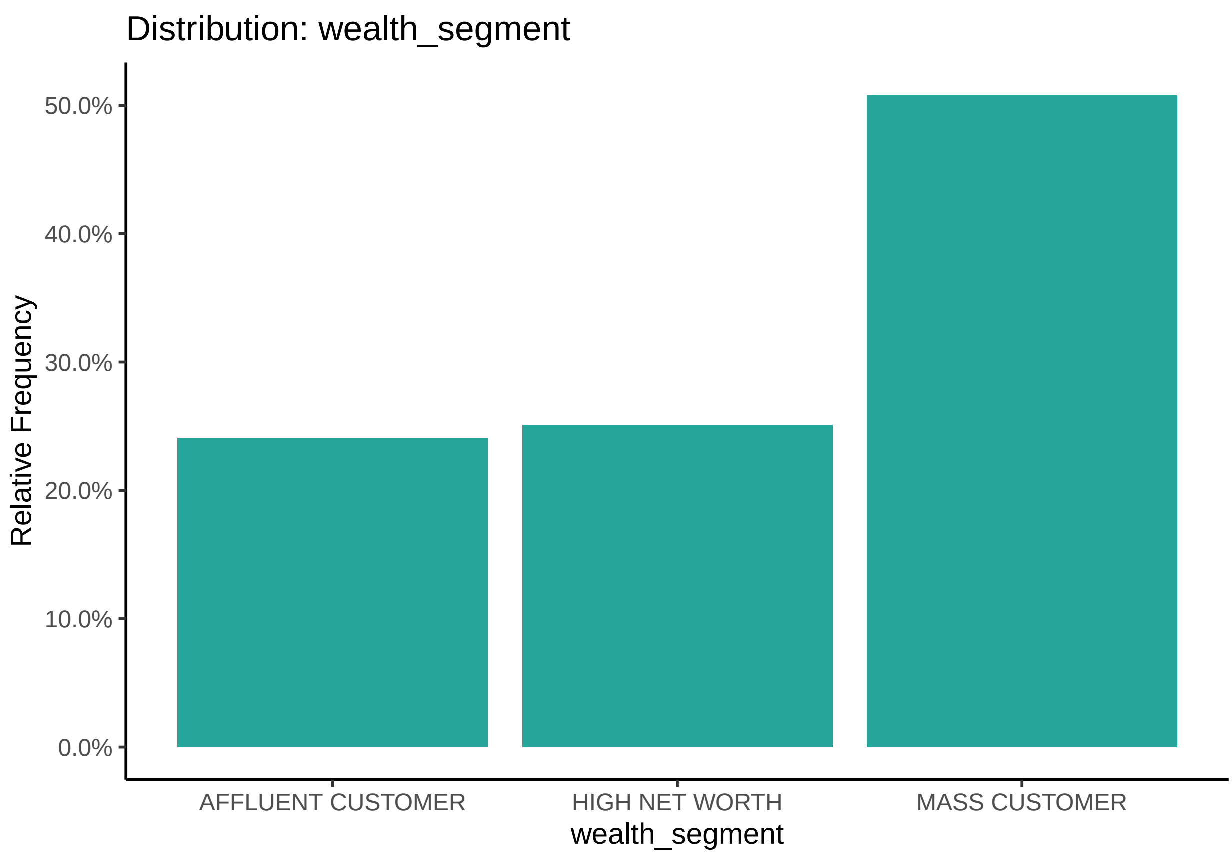

Distribution: gender
Relative Frequency
40%
20%
0%
FEMALE
MALE
gender

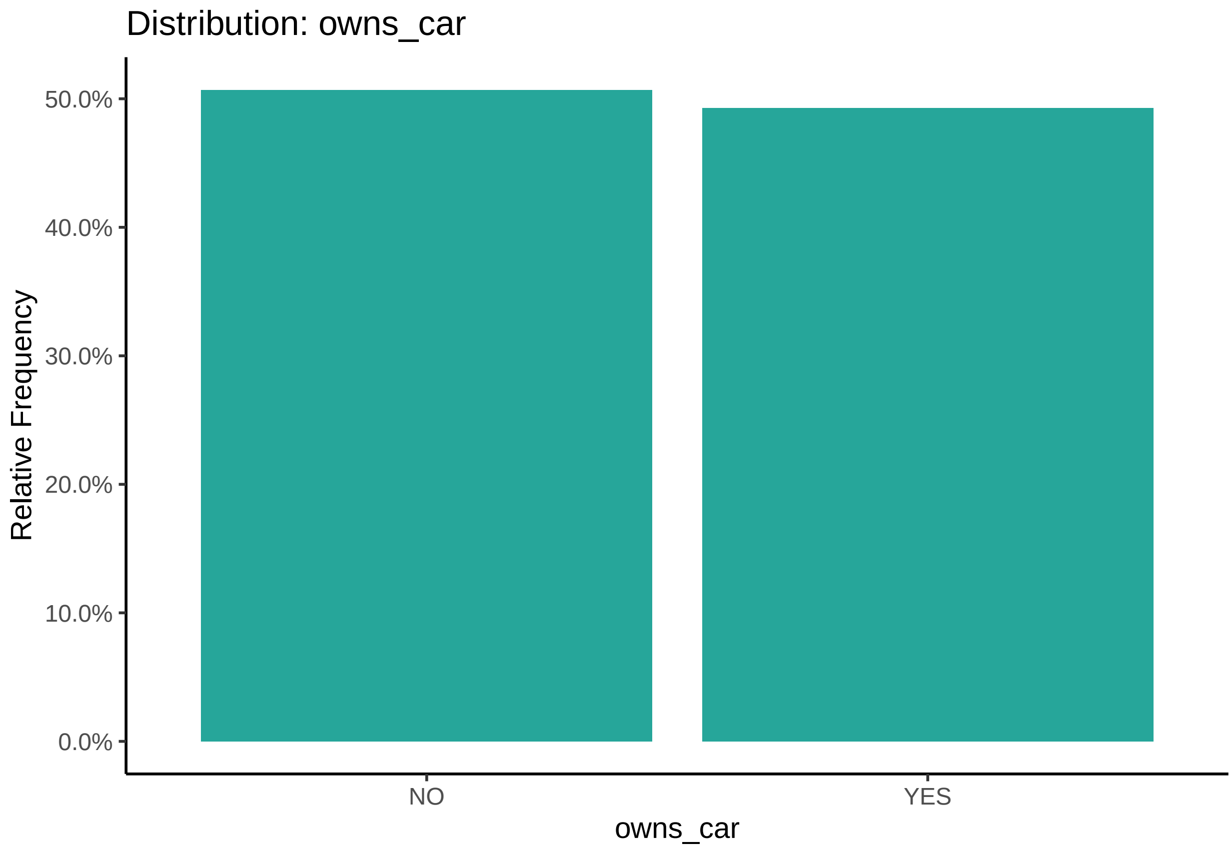

Distribution: past_3_years_bike_related_purchases
Frequency
0
20
40
60
0
25
50
75
100
past_3_years_bike_related_purchases

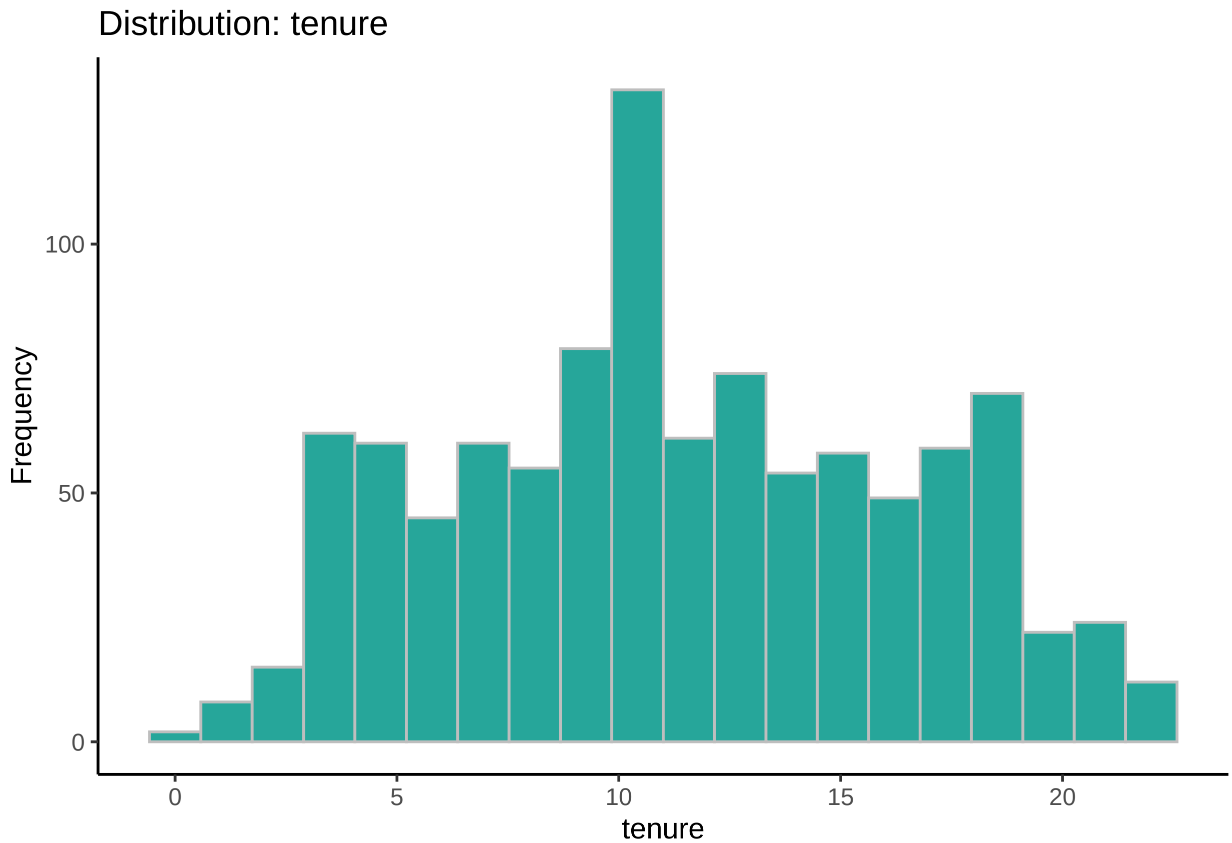

Distribution: DOB
100%
75%
50%
25%
0%
Relative Frequency
ALL_OTHER
DOB


Distribution: job_title
Relative Frequency
75%
50%
25%
0%
ALL_OTHER
job_title

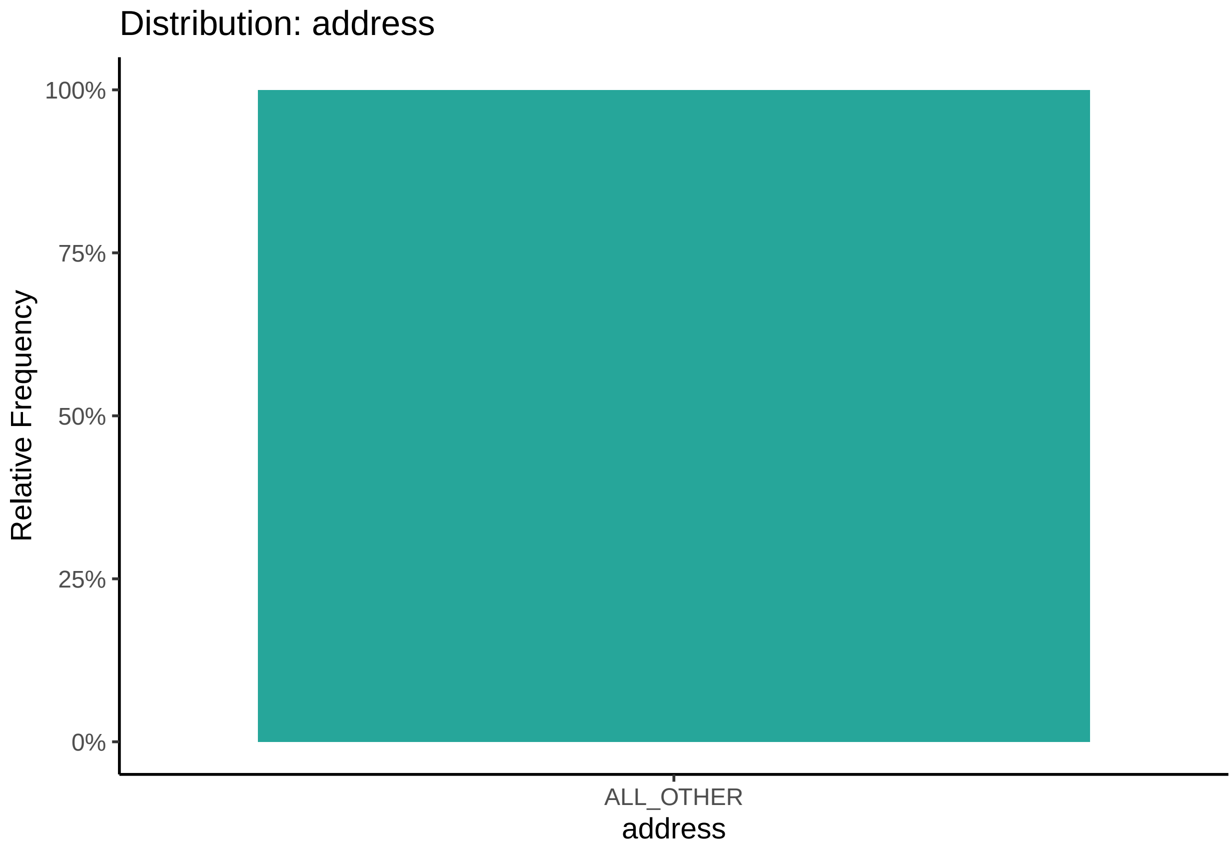

Distribution: job_industry_category
Relative Frequency
20.0%
15.0%
10.0%
5.0%
0.0%
ARGICULTURE
ENTERTAINMENT
FINANCIAL SERVICES
HEALTH
IT
MANUFACTURING
N/A
PROPERTY
RETAIL
TELECOMMUNICAT
job_industry_category

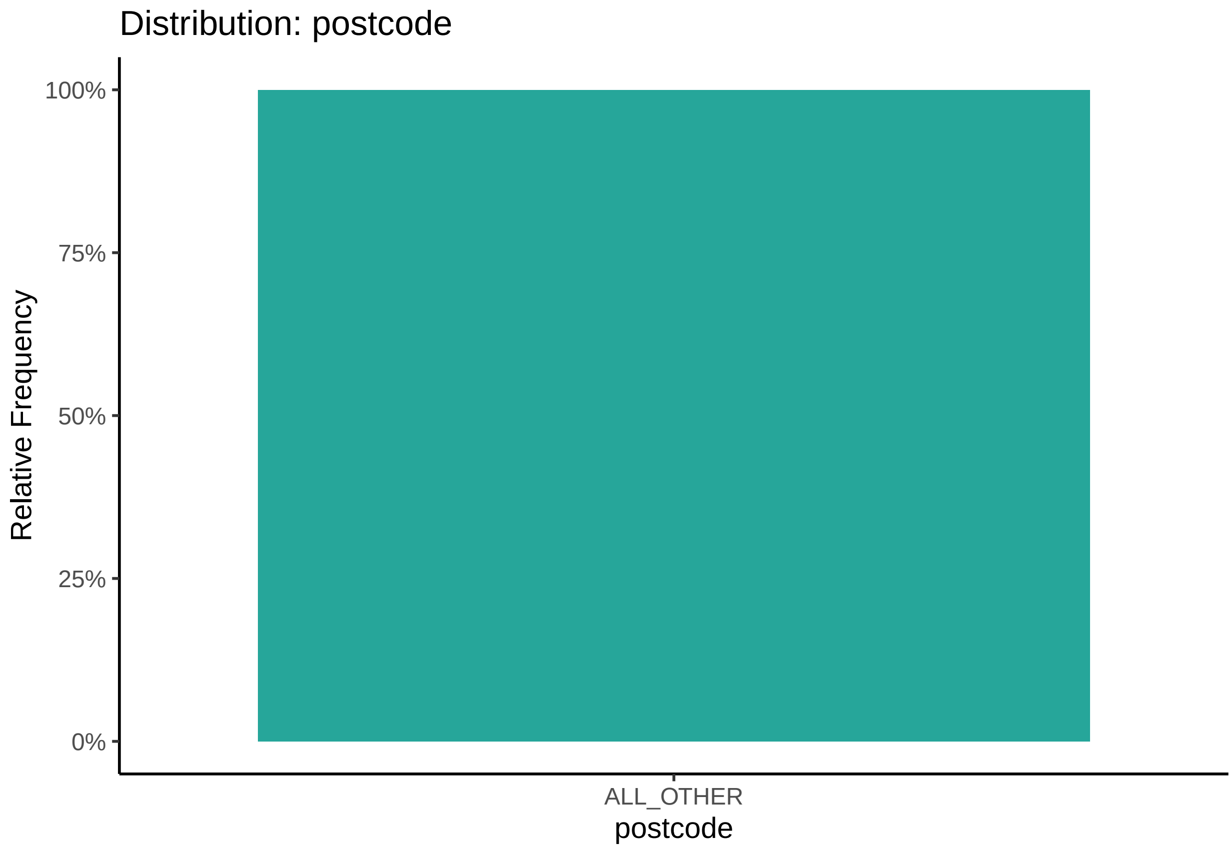

Distribution: wealth_segment
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
Relative Frequency
AFFLUENT CUSTOMER
HIGH NET WORTH
MASS CUSTOMER
wealth_segment

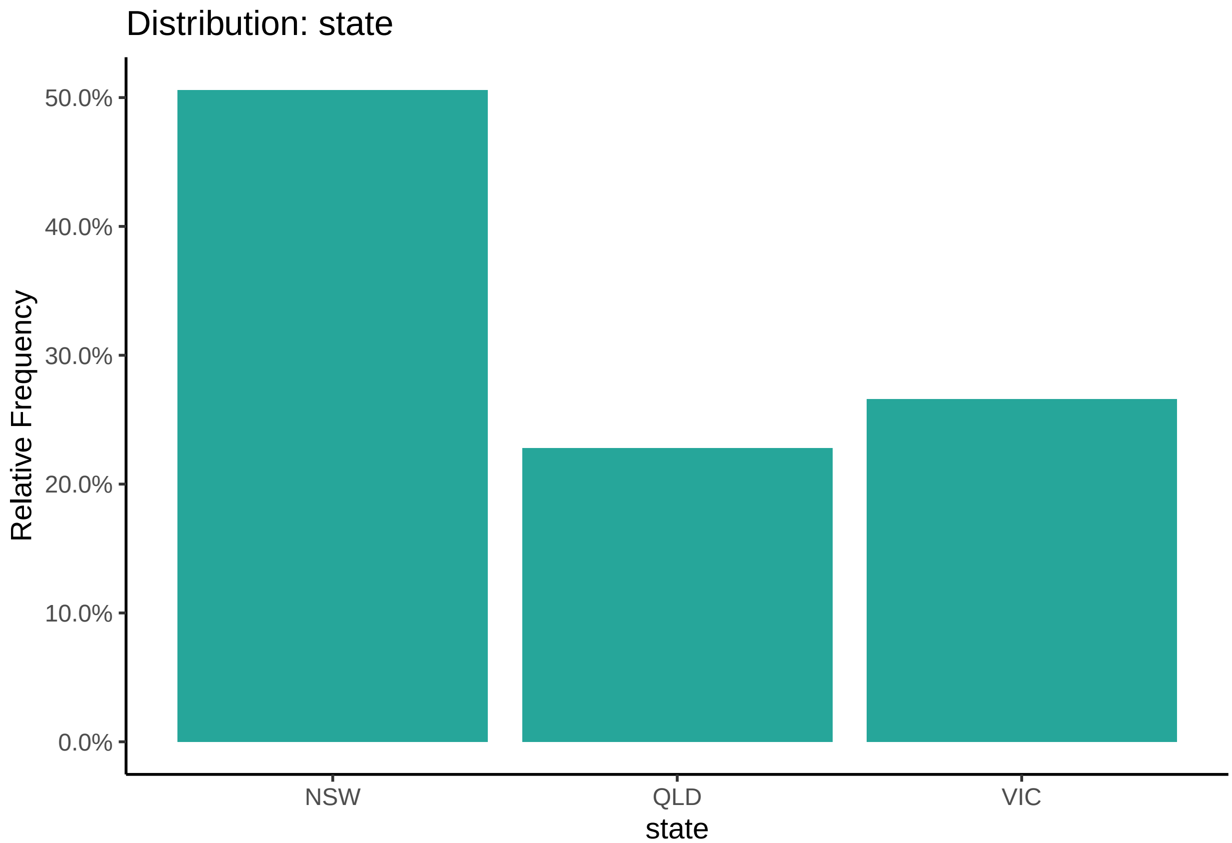

Distribution: owns_car
Relative Frequency
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
NO
YES
owns_car



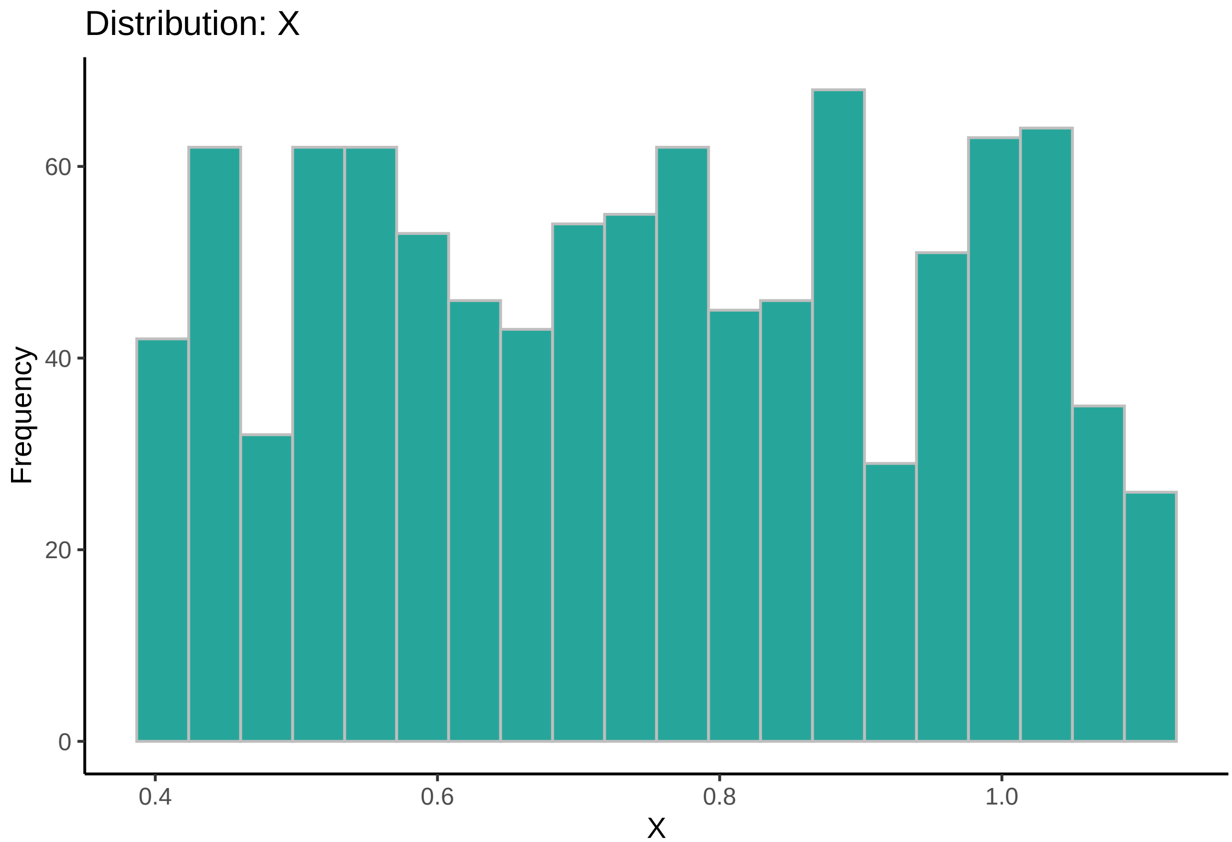
Distribution: tenure
Frequency
100
50
0
0
5
10
15
20
tenure

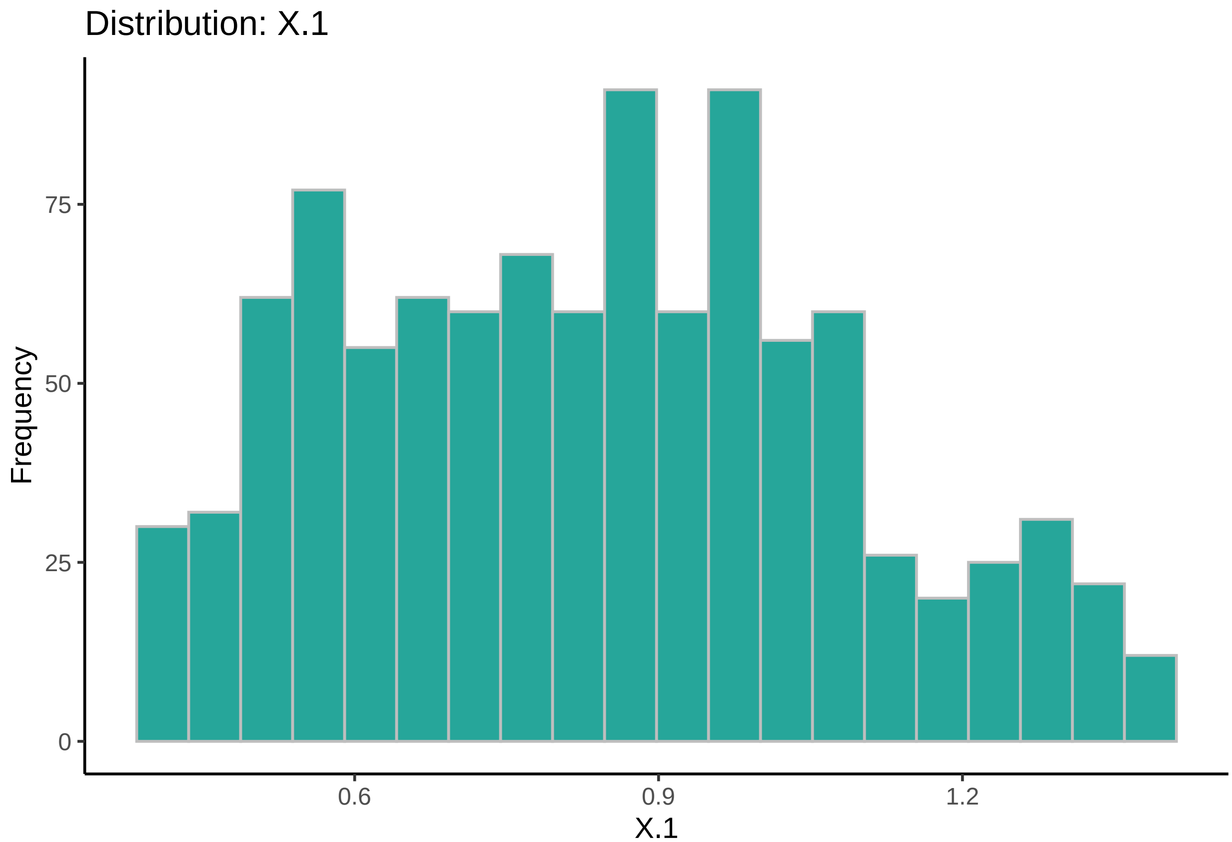

Distribution: address
100%
75%
50%
25%
0%
Relative Frequency
ALL_OTHER
address


Distribution: postcode
Relative Frequency
100%
75%
50%
25%
0%
ALL_OTHER
postcode

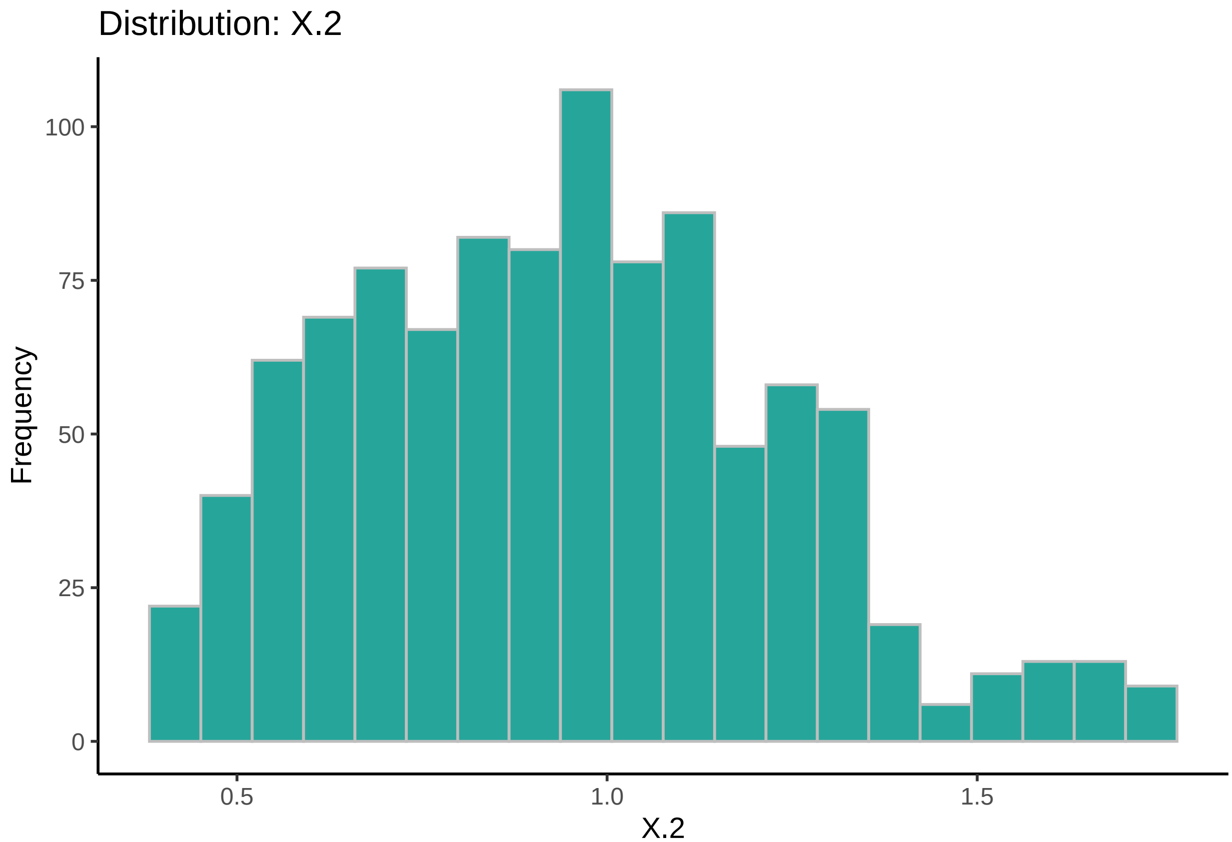

Distribution: state
Relative Frequency
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
NSW
QLD
VIC
state

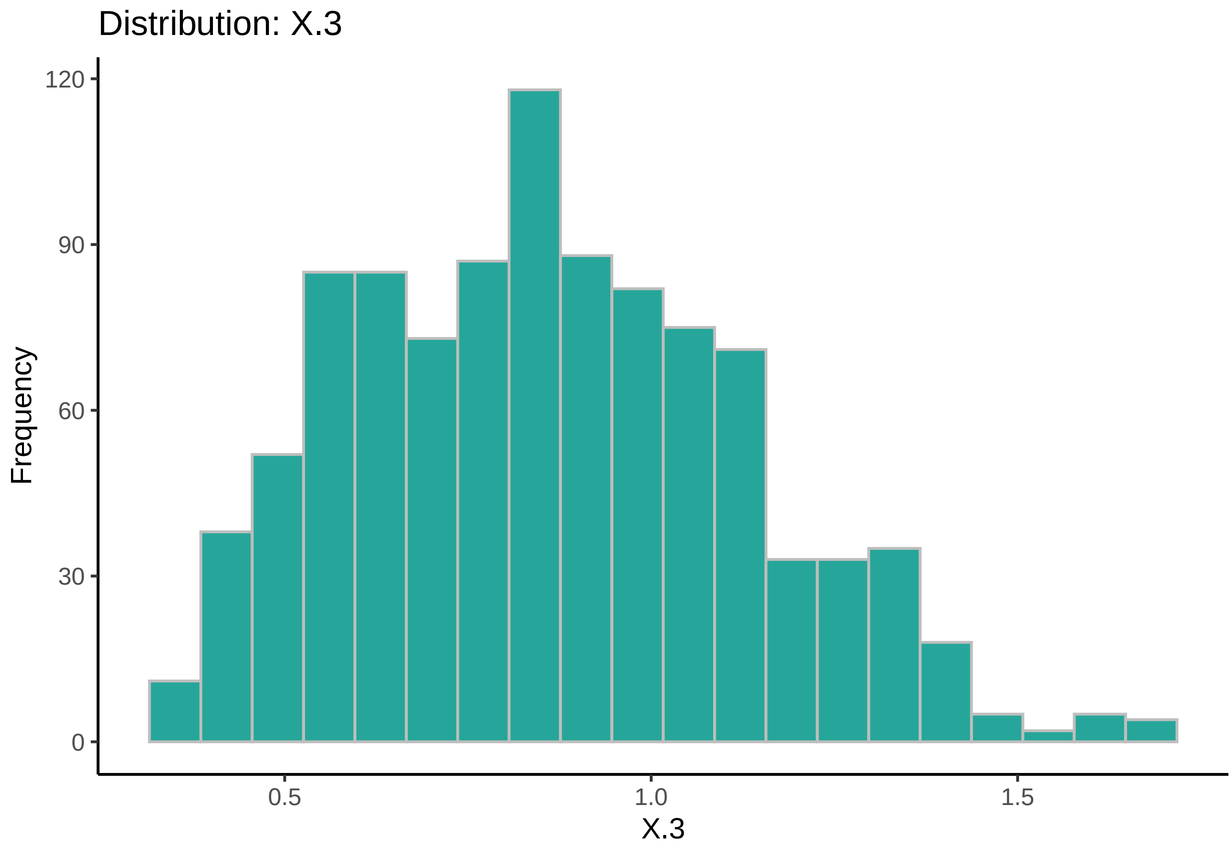
Distribution: property_valuation
Relative Frequency
15.0%
10.0%
5.0%
0.0%
1
10
11
12
2
3
4
5
6
7
8
9
property_valuation


Distribution: X
Frequency
60
40
20
0
0.4
0.6
0.8
1.0
X

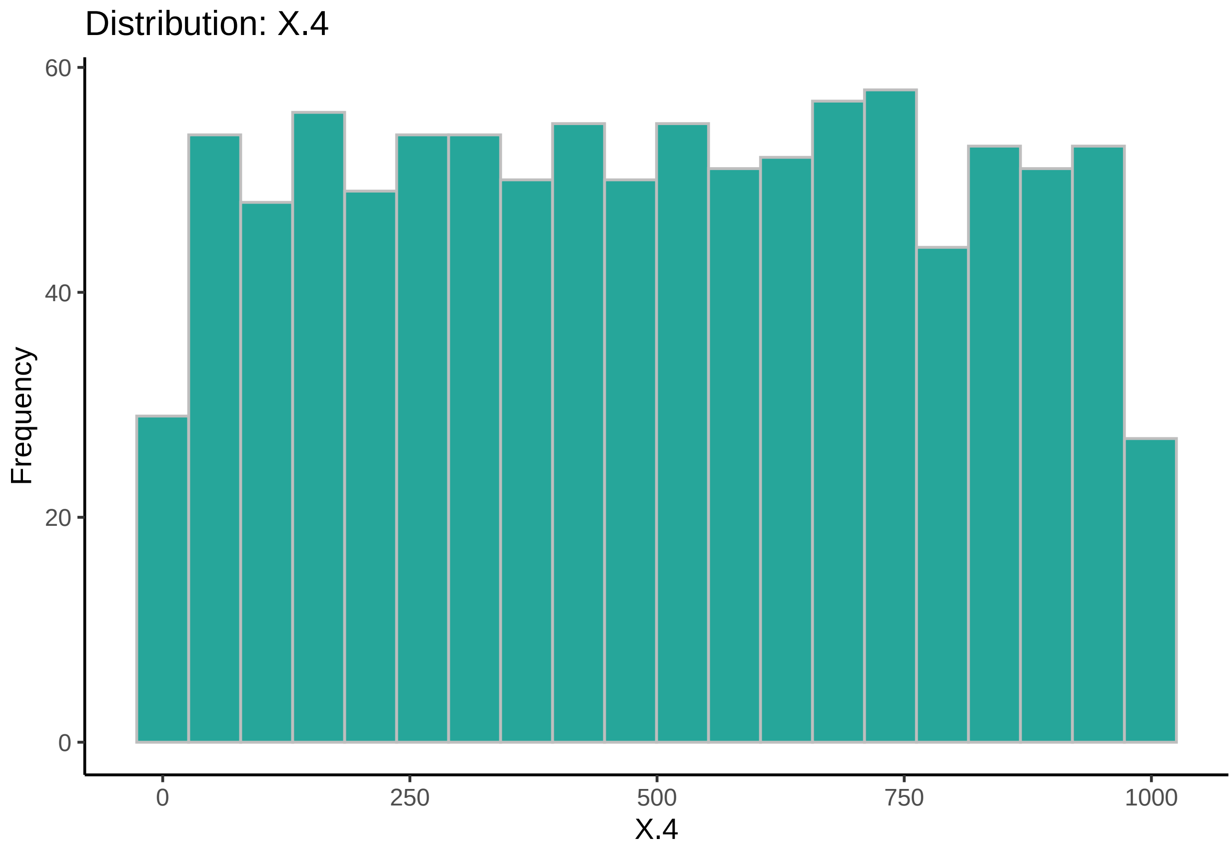


Distribution: X.1
Frequency
75
50
25
0
0.6
0.9
1.2
X.1

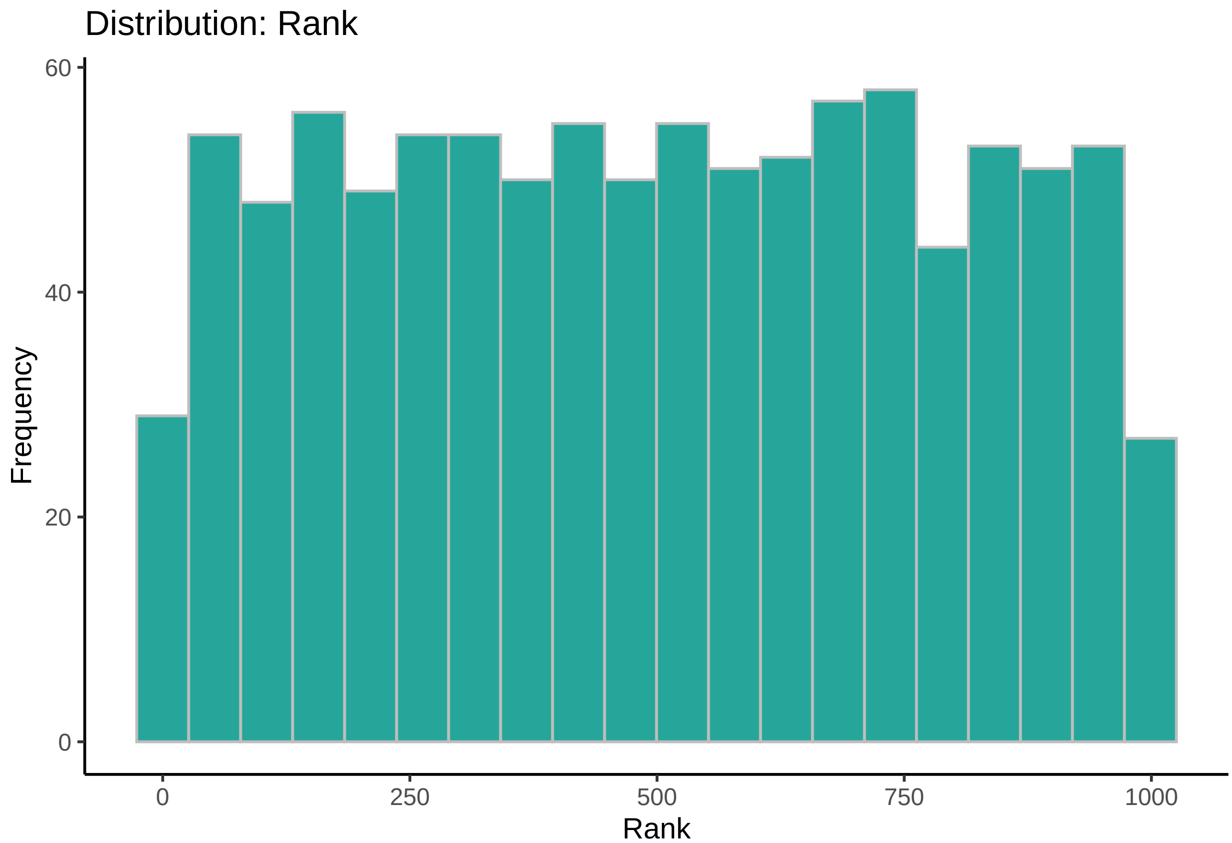

Distribution: X.2
100
75
50
25
0
Frequency
0.5
1.0
1.5
X.2

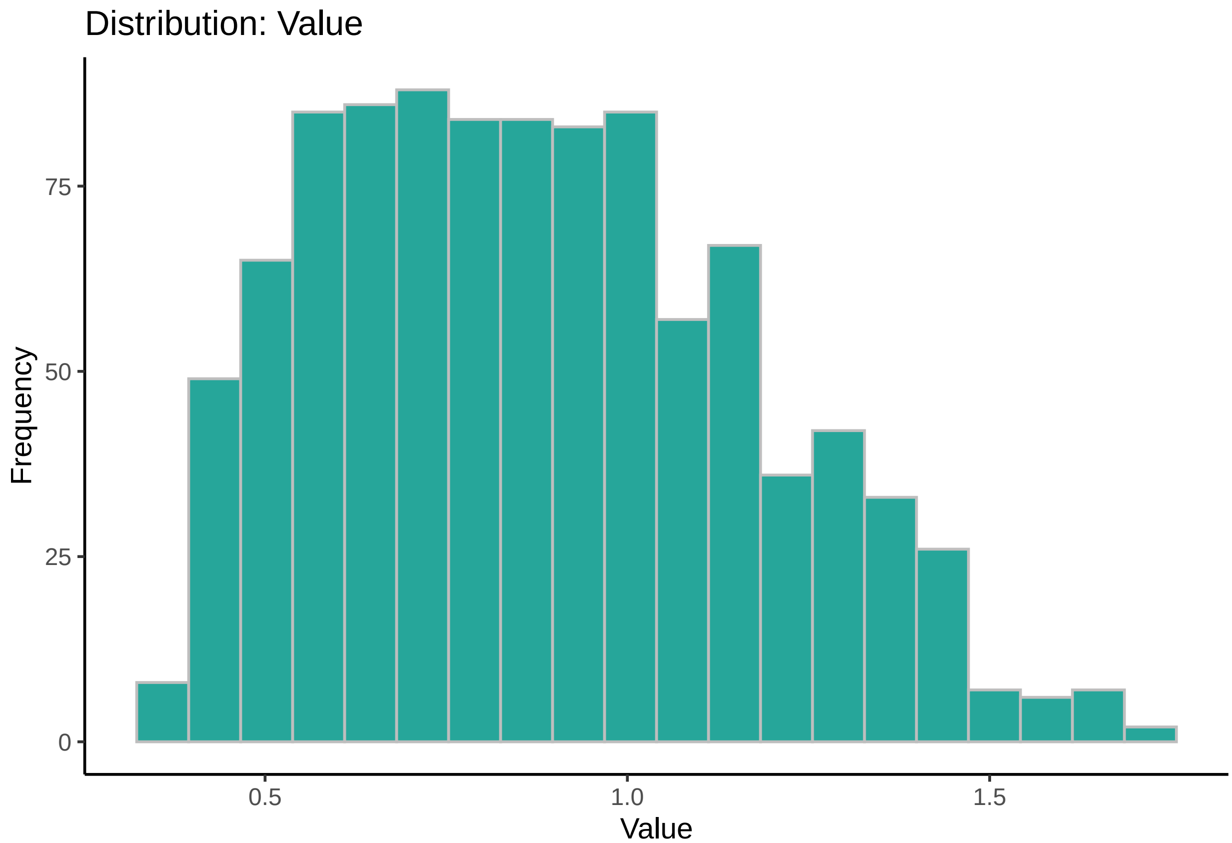


Distribution: X.3
Frequency
120
90
60
30
0
0.5
1.0
1.5
X.3

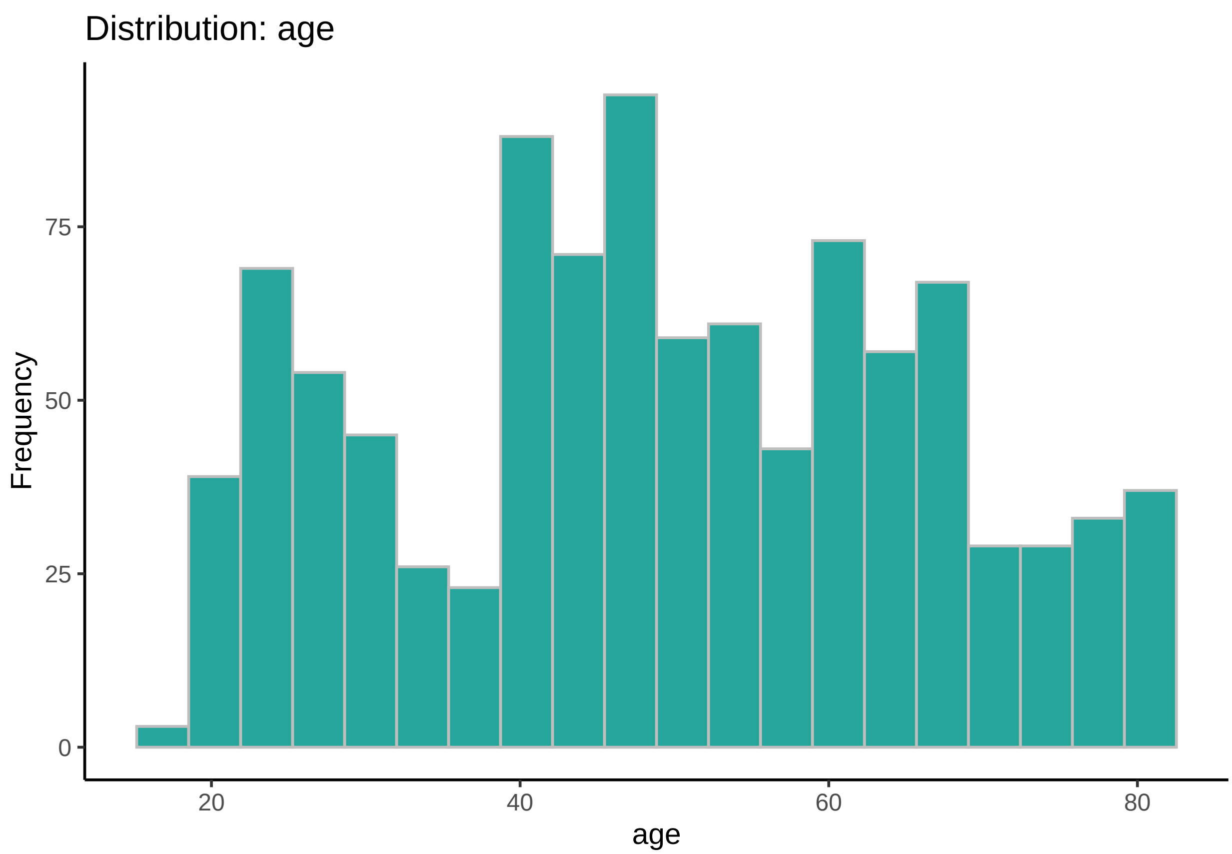


Distribution: X.4
60
40
20
0
Frequency
0
250
500
750
1000
X.4


Distribution: Rank
Frequency
Rank
0
20
40
60
0
250
500
750
1000



Distribution: Value
Frequency
75
50
25
0
0.5
1.0
1.5
Value




```
##                                   Feature Observations FeatureClass FeatureType
## 1                              first_name         1000    character Categorical
## 2                               last_name         1000    character Categorical
## 3                                  gender         1000    character Categorical
## 4  past_3_years_bike_related_purchases         1000      numeric  Continuous
## 5                                     DOB         1000    character Categorical
## 6                               job_title         1000    character Categorical
## 7                   job_industry_category         1000    character Categorical
## 8                          wealth_segment         1000    character Categorical
## 9                                owns_car         1000    character Categorical
## 10                                 tenure         1000      numeric  Continuous
## 11                                address         1000    character Categorical
## 12                               postcode         1000    character Categorical
## 13                                  state         1000    character Categorical
## 14                     property_valuation         1000    character Categorical
## 15                                      X         1000      numeric  Continuous
## 16                                    X.1         1000      numeric  Continuous
## 17                                    X.2         1000      numeric  Continuous
## 18                                    X.3         1000      numeric  Continuous
## 19                                    X.4         1000      numeric  Continuous
## 20                                   Rank         1000      numeric  Continuous
## 21                                  Value         1000      numeric  Continuous
## 22                                    age         1000      numeric  Continuous
##    PercentageMissing PercentageUnique ConstantFeature ZeroSpreadFeature
## 1                0.0             94.0              No                No
## 2                0.0             96.2              No                No
```

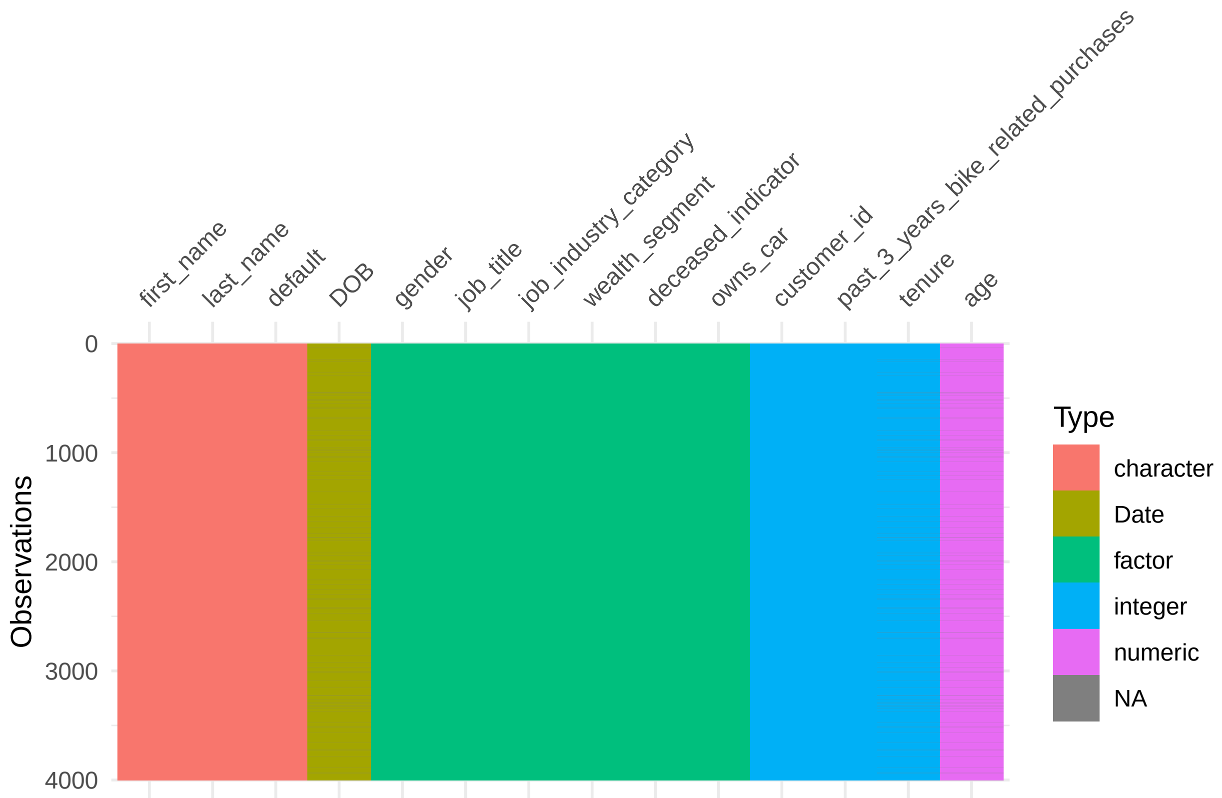
```
## 3              0.0             0.3             No                No
## 4              0.0            10.0             No                No
## 5              1.7            95.9             No                No
## 6              0.0            18.5             No                No
## 7              0.0             1.0             No                No
## 8              0.0             0.3             No                No
## 9              0.0             0.2             No                No
## 10             0.0             2.3             No                No
## 11             0.0           100.0             No                No
## 12             0.0            52.2             No                No
## 13             0.0             0.3             No                No
## 14             0.0             1.2             No                No
## 15             0.0             7.1             No                No
## 16             0.0            12.9             No                No
## 17             0.0            18.3             No                No
## 18             0.0            31.7             No                No
## 19             0.0            32.4             No                No
## 20             0.0            32.4             No                No
## 21             0.0            31.9             No                No
## 22             1.7             6.6             No                No
##    LowerOutliers UpperOutliers    ImputationValue MinValue FirstQuartile Median
## 1              0             0          ALL_OTHER     0.00     0.0000000   0.00
## 2              0             0          ALL_OTHER     0.00     0.0000000   0.00
## 3              0             0             FEMALE     0.00     0.0000000   0.00
## 4              0             0                 51     0.00    26.7500000  51.00
## 5              0             0            MISSING     0.00     0.0000000   0.00
## 6              0             0          ALL_OTHER     0.00     0.0000000   0.00
## 7              0             0 FINANCIAL SERVICES     0.00     0.0000000   0.00
## 8              0             0      MASS CUSTOMER     0.00     0.0000000   0.00
## 9              0             0                 NO     0.00     0.0000000   0.00
## 10             0             0                 11     0.00     7.0000000  11.00
## 11             0             0          ALL_OTHER     0.00     0.0000000   0.00
## 12             0             0          ALL_OTHER     0.00     0.0000000   0.00
## 13             0             0                NSW     0.00     0.0000000   0.00
## 14             0             0                  9     0.00     0.0000000   0.00
## 15             0             0               0.75     0.40     0.5700000   0.75
## 16             0             0             0.8375     0.40     0.6400000   0.84
## 17             0             0             0.9375     0.40     0.7082812   0.94
## 18             0             7               0.85     0.34     0.6500000   0.85
## 19             0             0                500     1.00   250.0000000 500.00
## 20             0             0                500     1.00   250.0000000 500.00
## 21             0             3               0.86     0.34     0.6495313   0.86
## 22             0             0                 48    18.00    37.0000000  48.00
##      Mean               Mode ThirdQuartile   MaxValue LowerOutlierValue
## 1    0.00             DORIAN         0.000    0.00000        0.00000000
## 2    0.00                            0.000    0.00000        0.00000000
## 3    0.00             FEMALE         0.000    0.00000        0.00000000
## 4   49.84                 60        72.000   99.00000      -41.12500000
## 5    0.00         1941-07-21         0.000    0.00000        0.00000000
## 6    0.00                            0.000    0.00000        0.00000000
## 7    0.00 FINANCIAL SERVICES         0.000    0.00000        0.00000000
## 8    0.00      MASS CUSTOMER         0.000    0.00000        0.00000000
## 9    0.00                 NO         0.000    0.00000        0.00000000
## 10  11.39                  9        15.000   22.00000       -5.00000000
```

```
## 11   0.00        0 BAY DRIVE         0.000    0.00000         0.00000000
## 12   0.00               2145         0.000    0.00000         0.00000000
## 13   0.00                NSW         0.000    0.00000         0.00000000
## 14   0.00                  9         0.000    0.00000         0.00000000
## 15   0.75                0.6         0.920    1.10000         0.04500000
## 16   0.84               0.75         1.010    1.37500         0.08500000
## 17   0.94             0.8625         1.125    1.71875         0.08320312
## 18   0.87               0.85         1.060    1.71875         0.03500000
## 19 498.82                760       750.250 1000.00000      -500.37500000
## 20 498.82                760       750.250 1000.00000      -500.37500000
## 21   0.88             0.6375         1.075    1.71875         0.01132813
## 22  49.21                 46        63.000   82.00000        -2.00000000
##    UpperOutlierValue
## 1           0.000000
## 2           0.000000
## 3           0.000000
## 4         139.875000
## 5           0.000000
## 6           0.000000
## 7           0.000000
## 8           0.000000
## 9           0.000000
## 10         27.000000
## 11          0.000000
## 12          0.000000
## 13          0.000000
## 14          0.000000
## 15          1.445000
## 16          1.565000
## 17          1.750078
## 18          1.675000
## 19       1500.625000
## 20       1500.625000
## 21          1.713203
## 22        102.000000
```

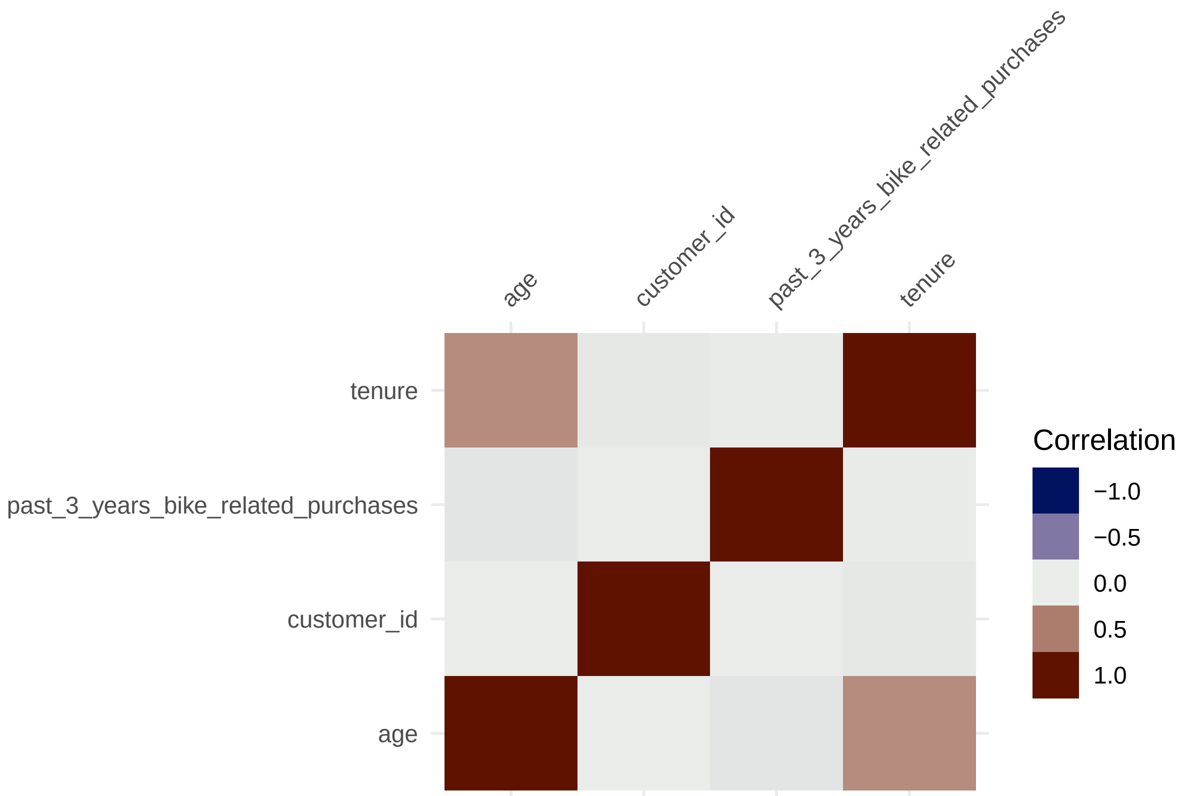
visdat

We can see that age(relatedly DOB) and tenure are missing for some customers. They are somewhat correalted also, we can see this from correlation plot. X columns which are nameless columns on newcustomer table are strongly correlated each other but we don't know about what they are measuring and also we don't have a similar past data about these features.

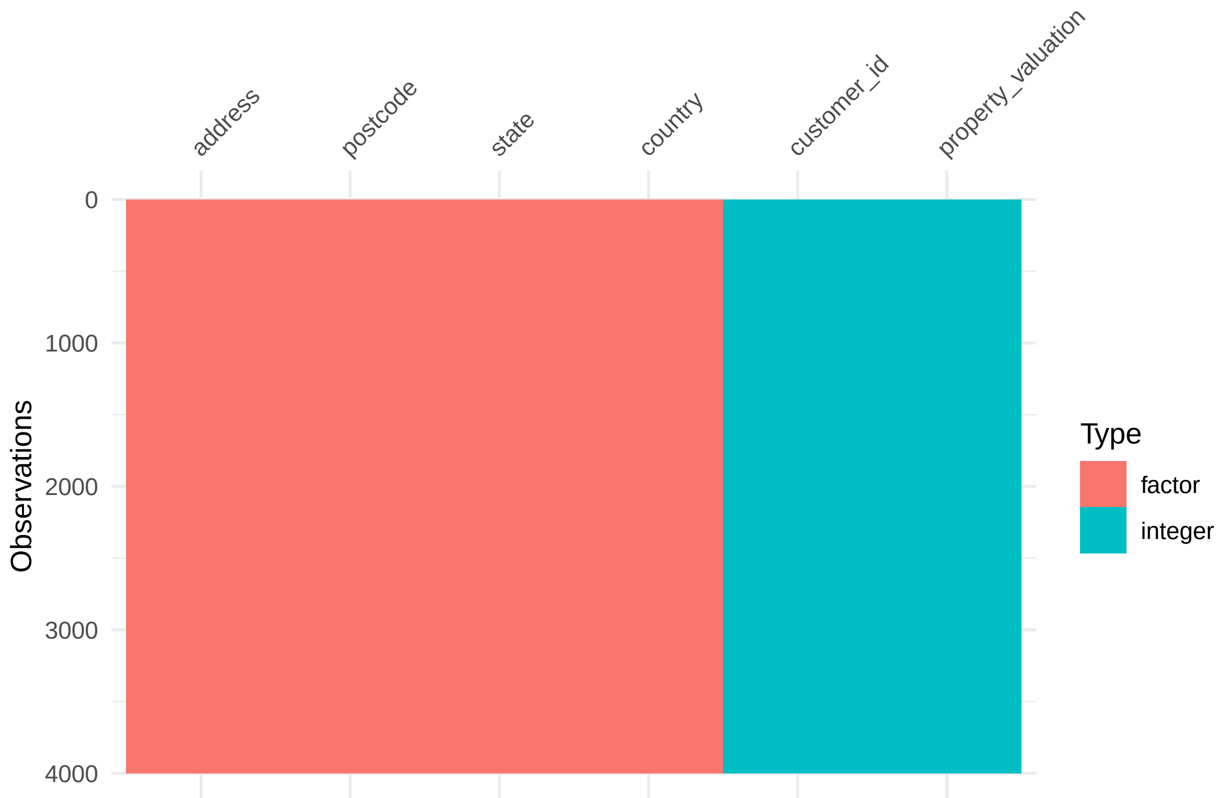
```
vis_dat(cdemographics)
```



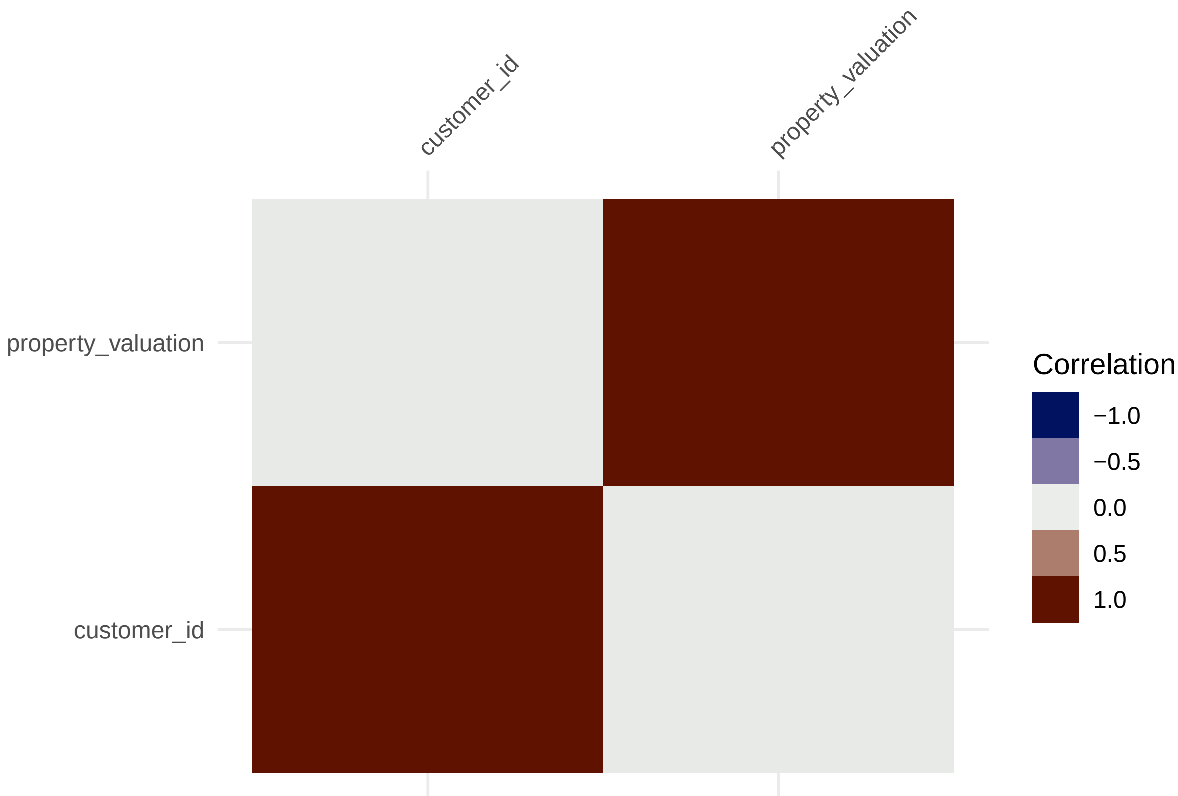

```
cdemographics %>% select_if(is.numeric) %>% vis_cor()
```



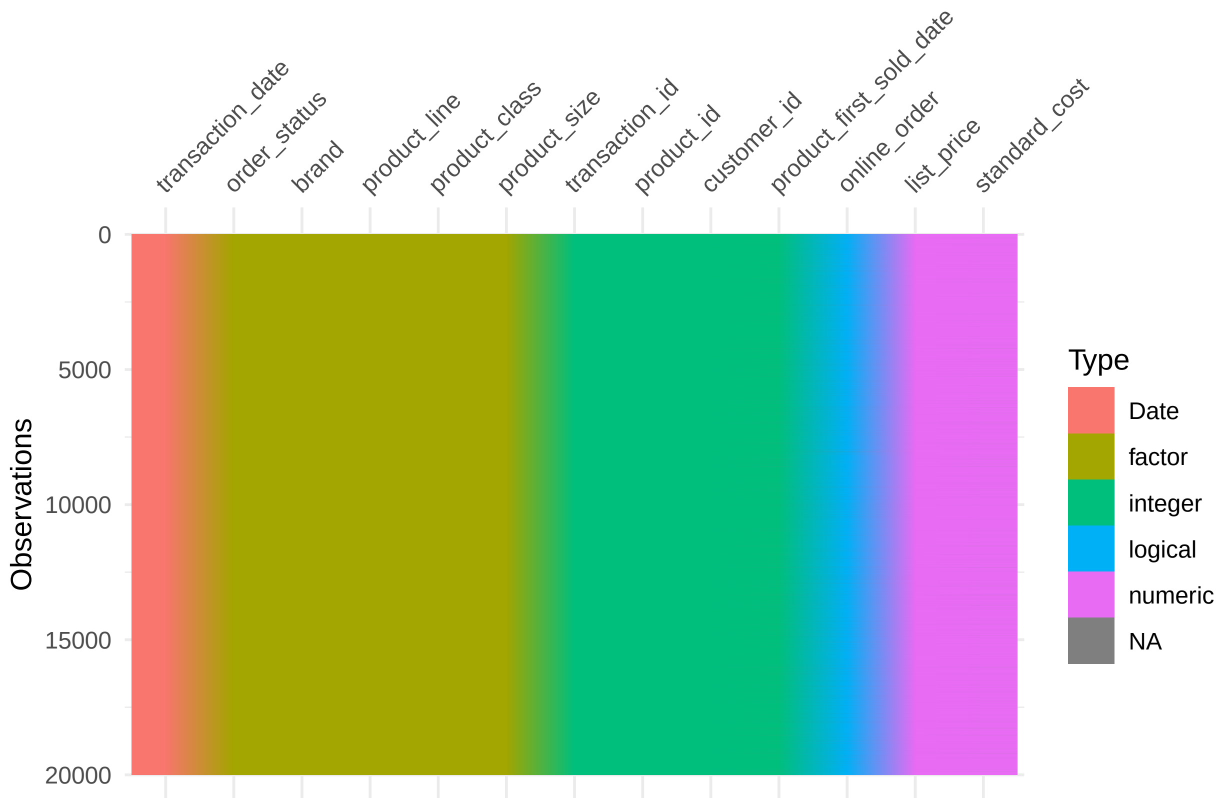

```
vis_dat(caddress)
```



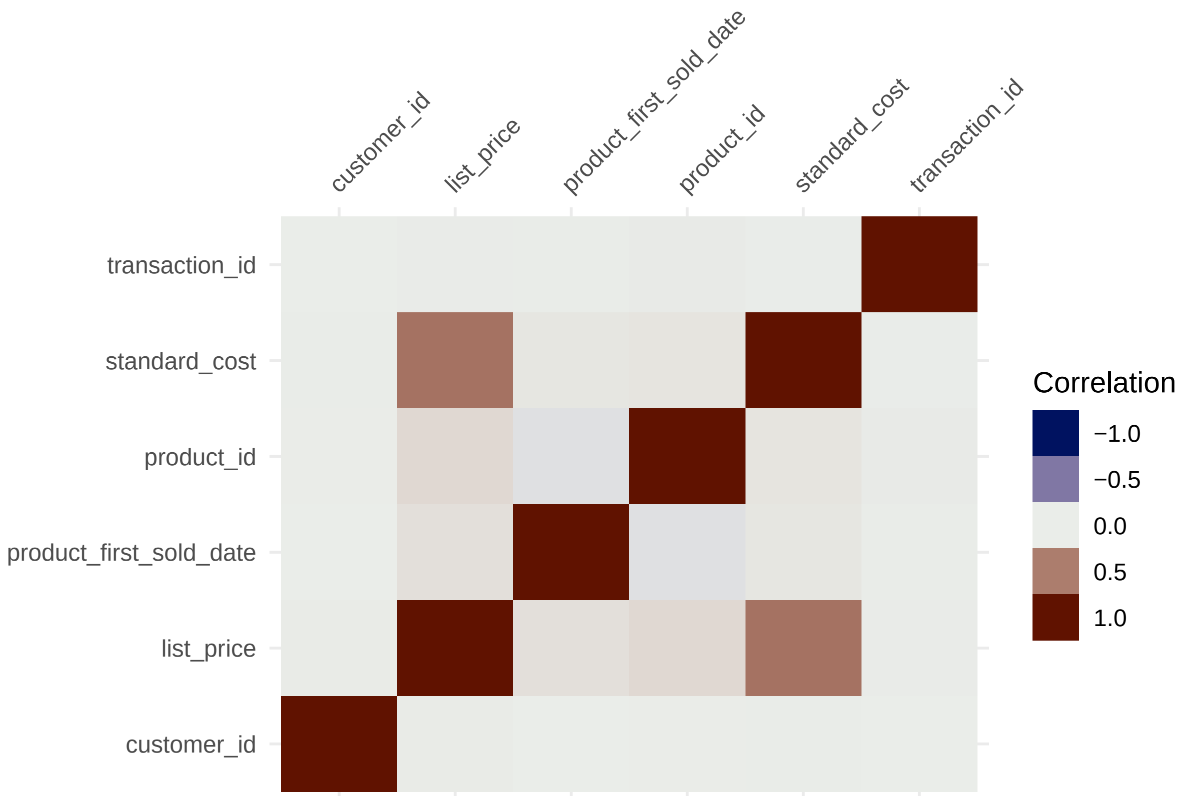

```r
caddress %>% select_if(is.numeric) %>% vis_cor()
```



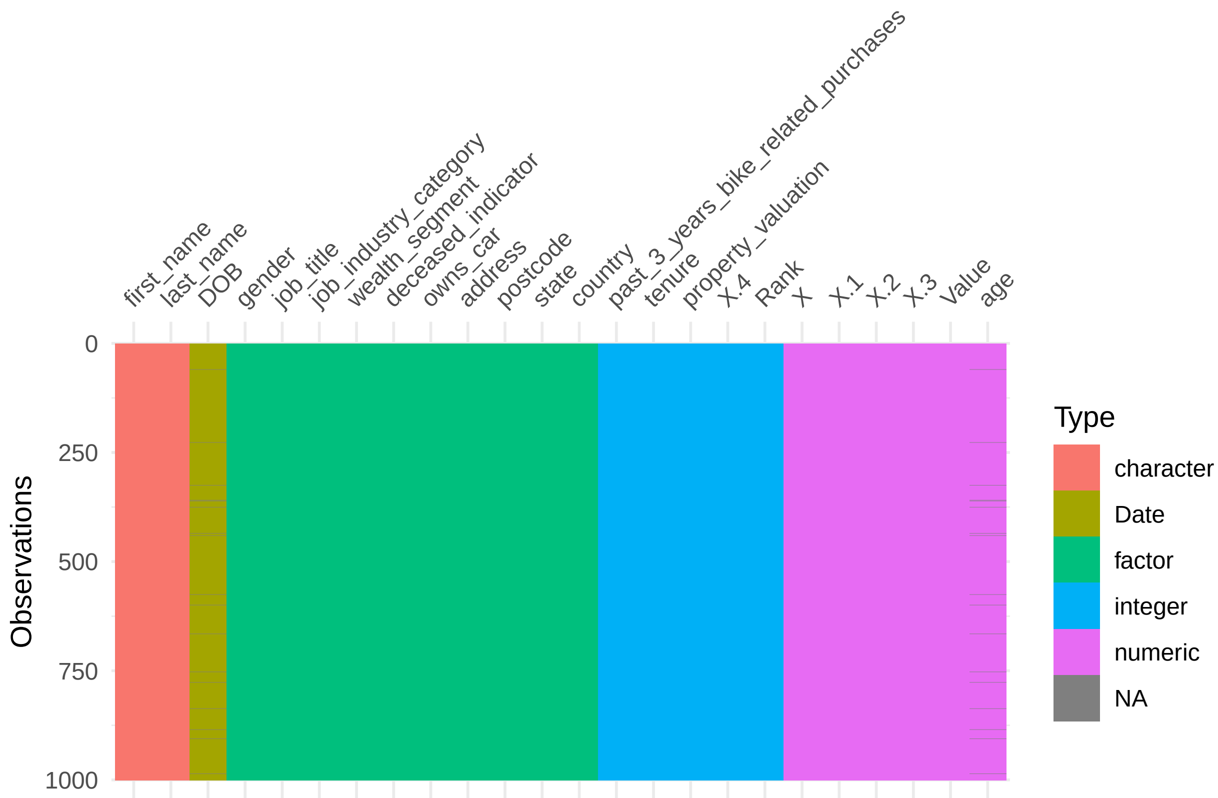

```
vis_dat(transactions)
```



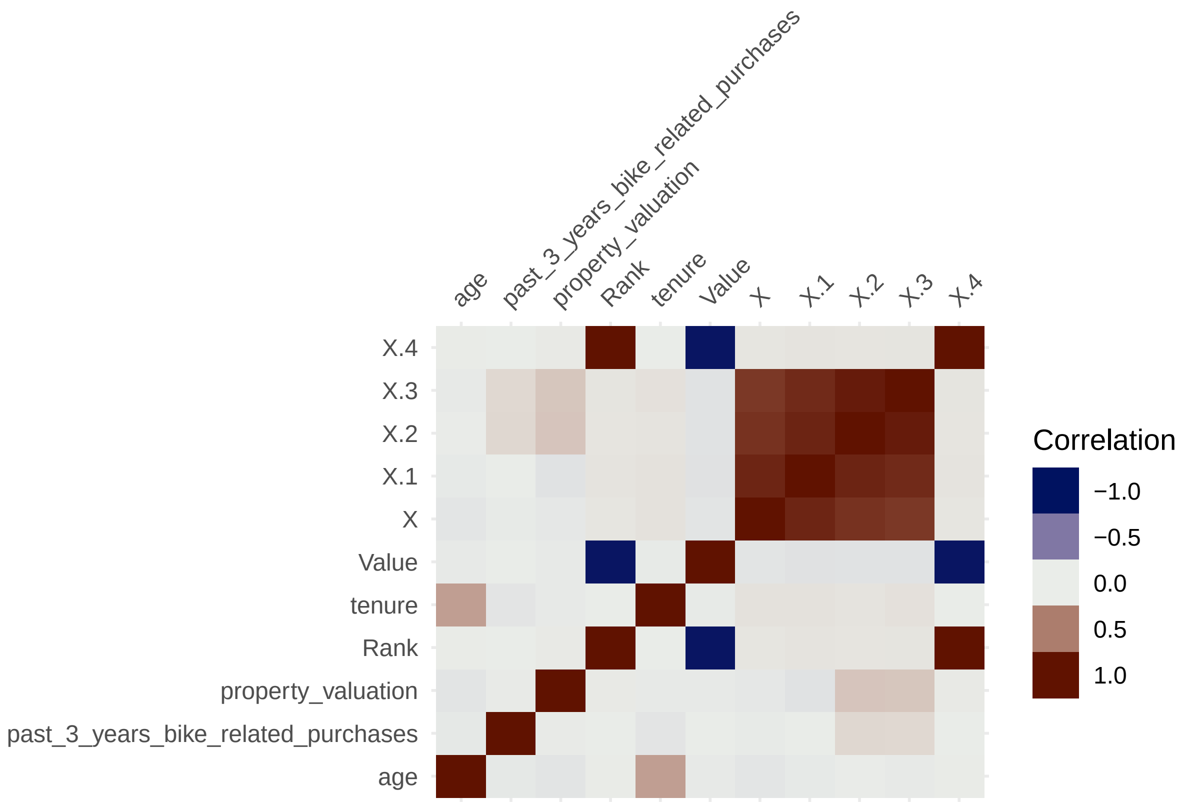

```
transactions %>% select_if(is.numeric) %>% vis_cor()
```




```
vis_dat(newcustomer)
```



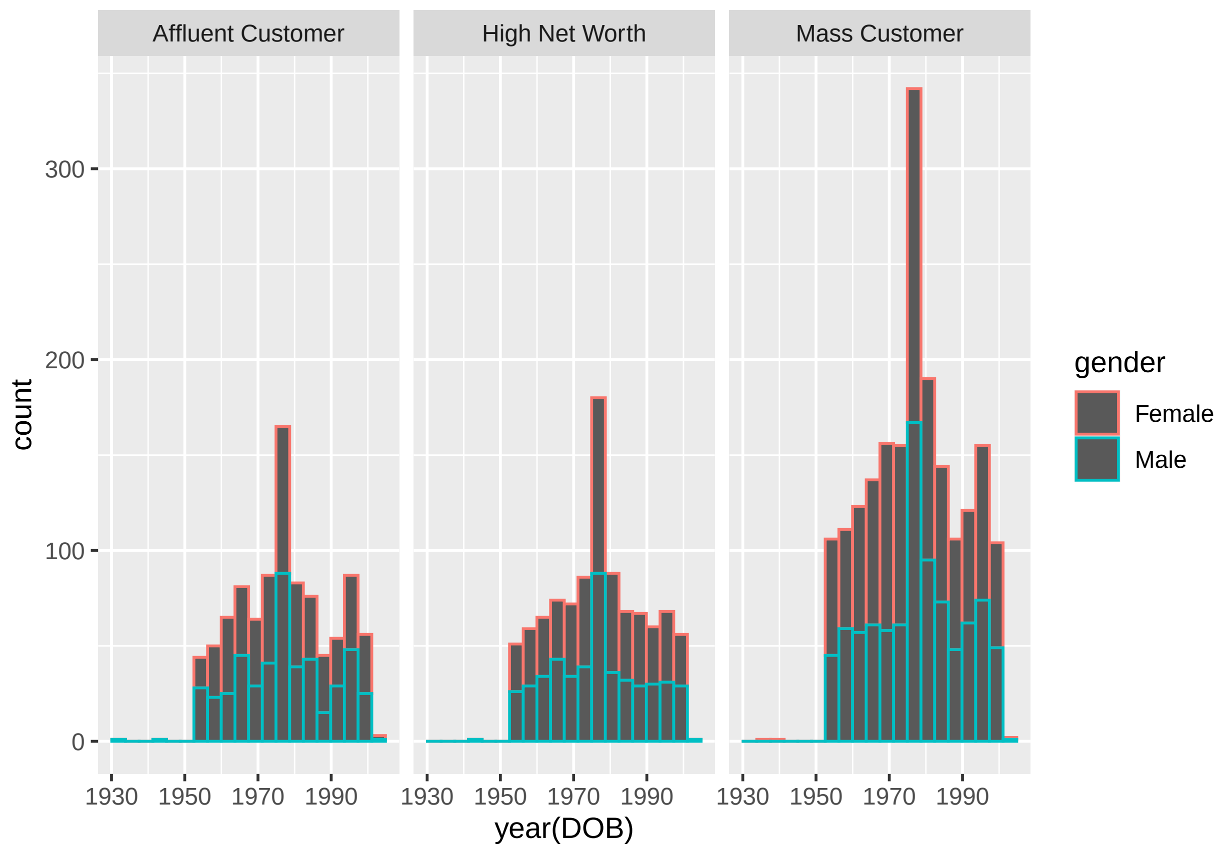

```
newcustomer %>% select_if(is.numeric) %>% vis_cor()
```



Selected Graphs and Tables

This graphic shows date of birth of customers accordingly their sexes. Customers are grouped by their wealth segments. Spreads look normally distributed.

```
cdemographics %>%
  filter(!is.na(DOB)) %>%
  ggplot(aes(year(DOB), color=gender)) +
  geom_histogram(bins=20) +
  facet_wrap(~wealth_segment)
```



I observed that 88 customers gender is marked as U while they do not have a determined date of birth(DOB). Also, only one of them have tenure information.

```
cdemographics %>%
  filter(is.na(DOB) | is.na(tenure)) %>%
  group_by(wealth_segment) %>%
  summarise(total = n(),
            proportion = total / 88)
```

```
## # A tibble: 3 x 3
##   wealth_segment    total proportion
##   <fct>             <int>      <dbl>
## 1 Affluent Customer    17      0.193
## 2 High Net Worth       25      0.284
## 3 Mass Customer        46      0.523
```

We can see that different brands are obtained for the 0th product and their prices are varied. Product_id variable is not consistent results to analyse.

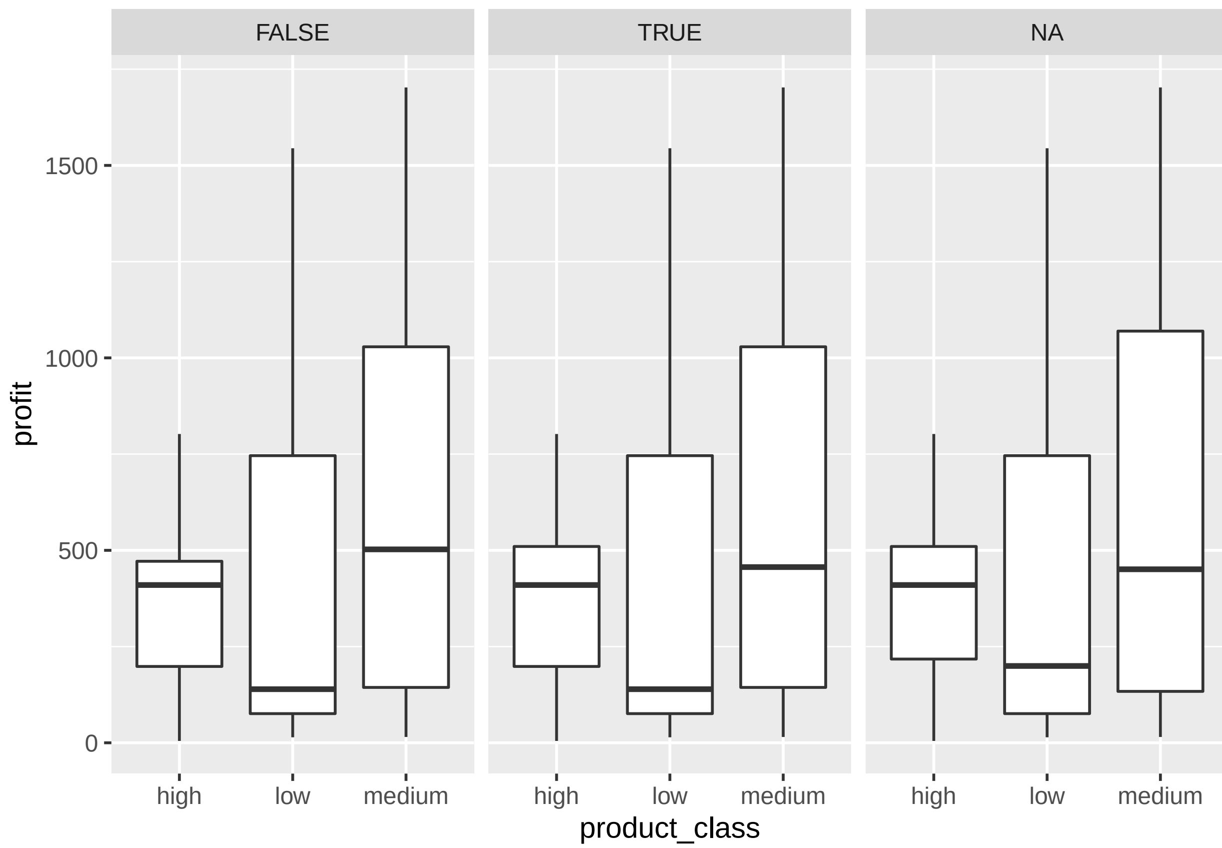
```
transactions %>%
  group_by(product_id, brand) %>%
  summarise(total = n(), avg=mean(list_price), min=min(list_price), max=max(list_price)) %>%
  arrange(product_id) %>%
  head()
```

```
## # A tibble: 6 x 6
## # Groups:   product_id [1]
##   product_id brand            total   avg   min   max
```

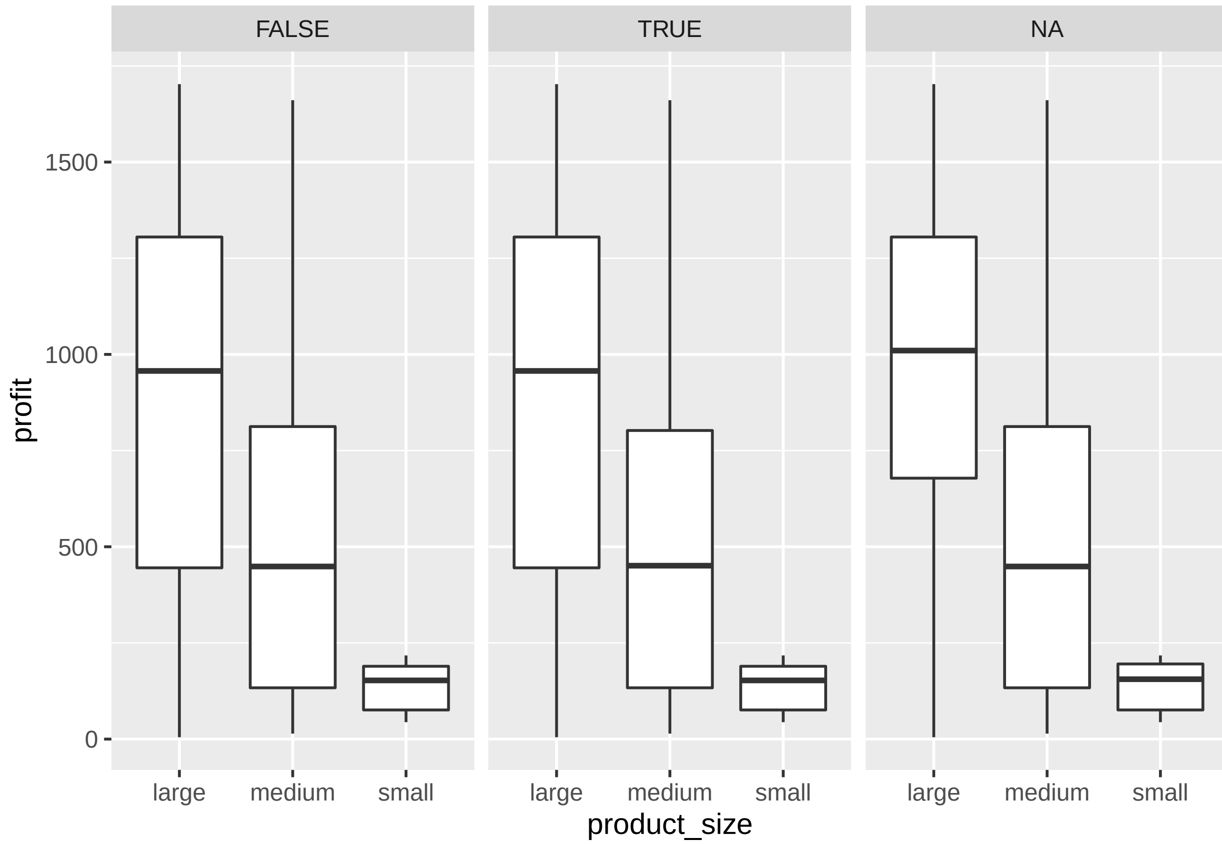
```
##         <int> <fct>            <int> <dbl> <dbl> <dbl>
## 1           0 ""                 197 1091.  16.1 2086.
## 2           0 "Giant Bicycles"   105  382. 231.   570.
## 3           0 "Norco Bicycles"   241  448. 360.   544.
## 4           0 "OHM Cycles"       242  152.  12.0  743.
## 5           0 "Solex"            276  255.  71.5  478.
## 6           0 "Trek Bicycles"    221  440. 291.   534.
```

Profit variable has been added to Transactions table. Profit is calculated by difference between list_price and standard_cost.

```
transactions <- transactions %>% mutate(profit = list_price - standard_cost)

transactions %>% filter(!is.na(profit)) %>% ggplot(aes(product_class, profit)) +
  geom_boxplot() +
  facet_wrap(~online_order)
```




```
transactions %>% filter(!is.na(profit)) %>% ggplot(aes(product_size, profit)) +
  geom_boxplot() +
  facet_wrap(~online_order)
```



Joining transactions and cdemographics table made possible to observe wealth_segment spread.

```
transactions %>%
  summarize(total_active_customers = n_distinct(customer_id)
            )
```

```
##   total_active_customers
## 1                   3494
```

```
transactions %>% filter(!is.na(profit)) %>%
  group_by(customer_id) %>%
  summarise(total_order= n(),
            total_profit=sum(profit),
            avg_profit = sum(profit) / n()) %>%
  arrange(desc(total_order)) %>%
  head()
```

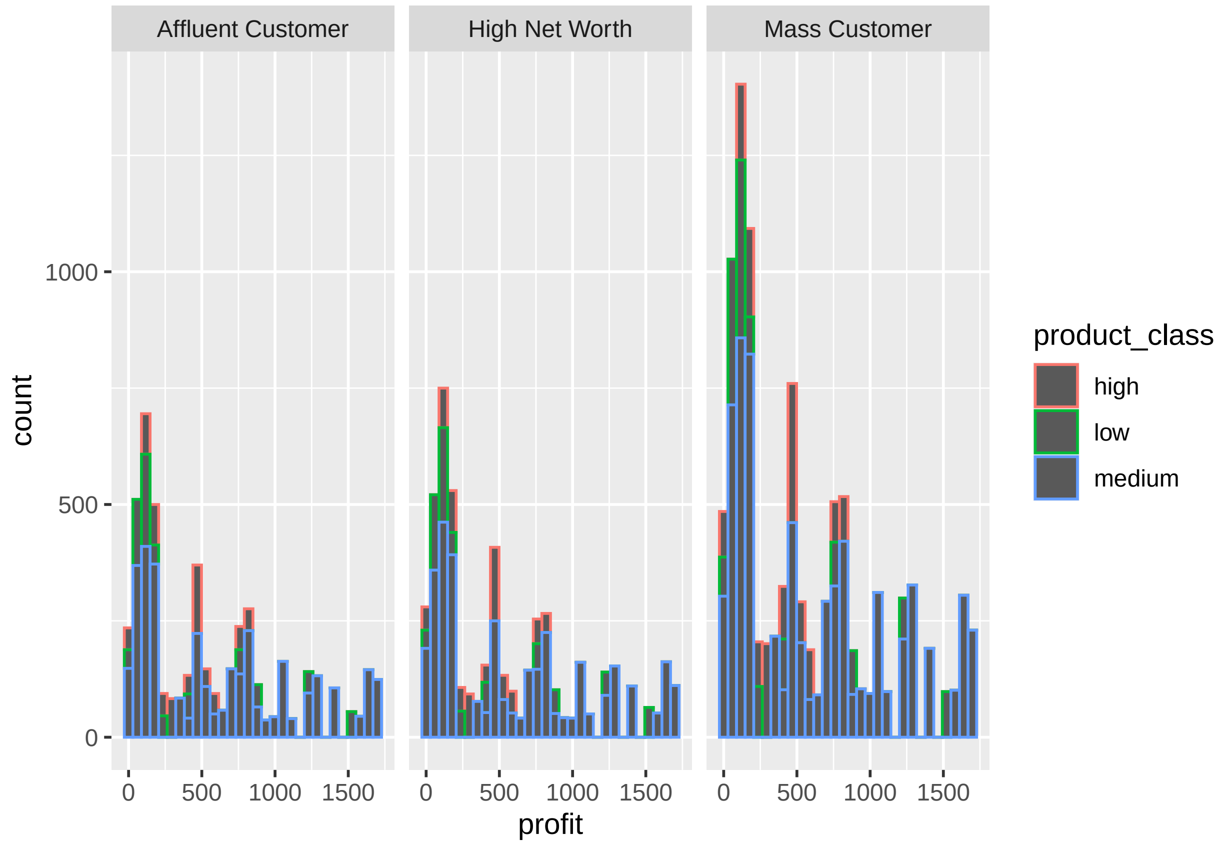
```
## # A tibble: 6 x 4
##   customer_id total_order total_profit avg_profit
##         <int>       <int>        <dbl>      <dbl>
## 1        1068          14        4842.       346.
## 2        2183          14        6513.       465.
## 3        2476          14        7493.       535.
## 4         637          13        5402.       416.
## 5        1129          13        6791.       522.
## 6        1140          13        8533.       656.
```

```r
# wealth segment statistics
transactions %>%
  left_join(cdemographics, by="customer_id") %>%
  filter(!is.na(profit)) %>%
  filter(!is.na(wealth_segment)) %>%
  group_by(wealth_segment) %>%
  summarise(total_customer = n_distinct(customer_id),
            total_order= n(),
            order_per_customer = n() / n_distinct(customer_id),
            total_profit = sum(profit),
            avg_profit = sum(profit)/n()
            )
```

```
## # A tibble: 3 x 6
##   wealth_segment total_customer total_order order_per_custo~ total_profit
##   <fct>                   <int>       <int>            <dbl>        <dbl>
## 1 Affluent Cust~            851        4810             5.65     2678011.
## 2 High Net Worth            895        5046             5.64     2770520.
## 3 Mass Customer            1747        9944             5.69     5481484.
## # ... with 1 more variable: avg_profit <dbl>
```

```r
transactions %>% filter(!is.na(profit)) %>%
  left_join(cdemographics, by="customer_id") %>%
  filter(!is.na(wealth_segment)) %>%
  ggplot(aes(profit, color = product_class)) +
  geom_histogram() +
  facet_wrap(~ wealth_segment)
```

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




```
transactions_grouped <- transactions %>%
  group_by(customer_id) %>%
  summarise(total_order= n(),
            total_profit = sum(profit),
            avg_profit = sum(profit)/n()
            )
transactions_grouped %>% group_by(total_order) %>% summarise(n = n())
```

```
## # A tibble: 14 x 2
##    total_order     n
##          <int> <int>
##  1           1    49
##  2           2   202
##  3           3   361
##  4           4   499
##  5           5   601
##  6           6   569
##  7           7   476
##  8           8   311
##  9           9   207
## 10          10   112
## 11          11    60
## 12          12    28
## 13          13    16
## 14          14     3
```

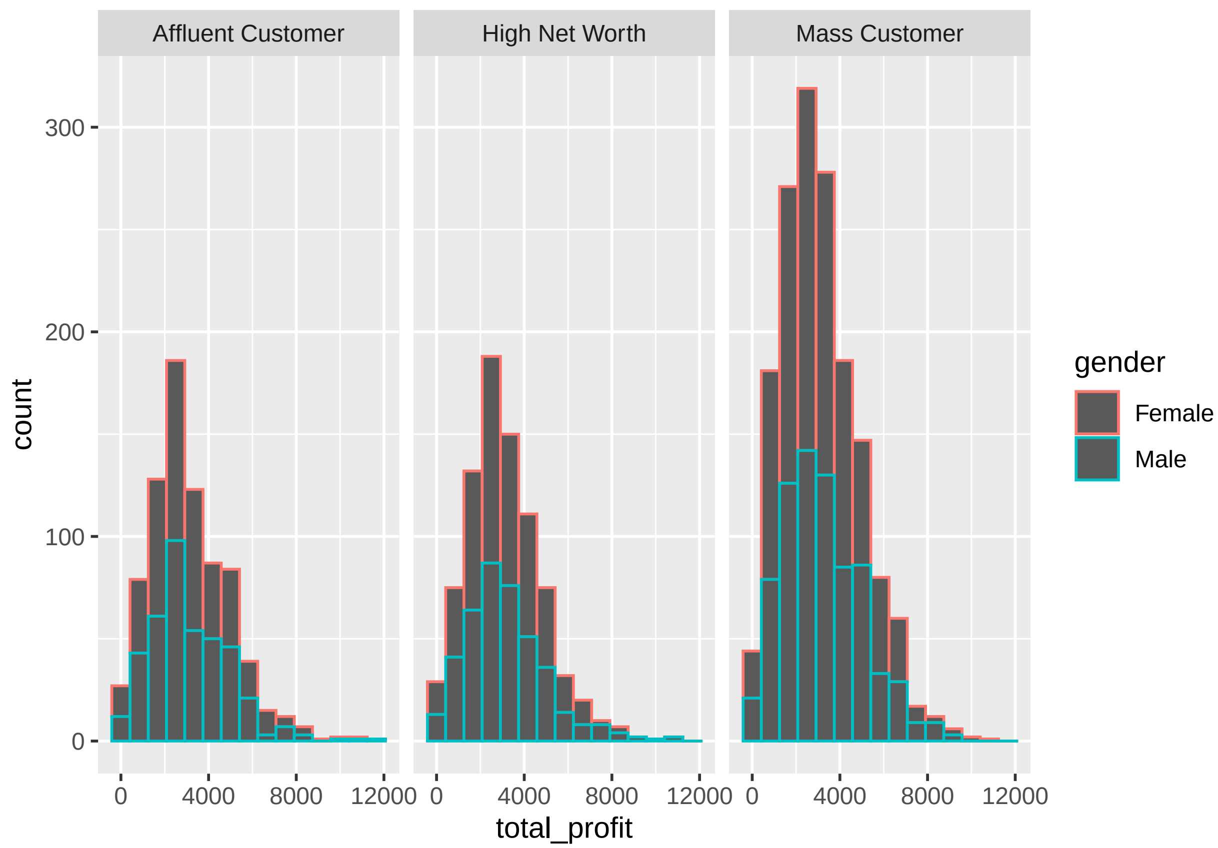
3

New customers should be categorized subject to given customer demographics data and related datasets. We can join tables to add new features to explore on cdemographics dataset. Firstly, I am going to focus decision tree models.

Preparing the data

I left-joined cdemographics and caddress tables and selected all columns that we can make predictions. I started to learn the data with sampling. 3126 of 3908 observation are attended as train and remainings are test.

```
# Join all the tables to be able to reach more features
training_set <- cdemographics %>%
  left_join(caddress, by="customer_id") %>%
  left_join(transactions_grouped, by="customer_id") %>%
  # job_title and job_industry_category
  select(total_profit, total_order, wealth_segment, gender, past_3_years_bike_related_purchases,
          owns_car, tenure, age, property_valuation) %>%
  drop_na()

training_set %>% ggplot(aes(total_profit, color=gender)) + geom_histogram(bins=15) + facet_wrap(~wealth_
```




```
#set.seed(123)
train_sample <- sample(nrow(training_set), round(nrow(training_set)*0.8))


train <- training_set[train_sample, ]
```

```r
test  <- training_set[-train_sample, ]
```

We can see below that training and test datasets have similar proportion of wealth_segments

```r
prop.table(table(train$wealth_segment))
```

```
##
## Affluent Customer    High Net Worth     Mass Customer
##         0.2491296         0.2618956         0.4889749
```

```r
prop.table(table(test$wealth_segment))
```

```
##
## Affluent Customer    High Net Worth     Mass Customer
##         0.2306502         0.2430341         0.5263158
```

```r
lm1 <- lm(total_profit~.,train)
```

```r
summary(lm1)
```

```
##
## Call:
## lm(formula = total_profit ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3823.3  -814.0  -100.1   722.7  6232.1
##
## Coefficients:
##                                    Estimate Std. Error t value
## (Intercept)                       -118.6573   141.7325  -0.837
## total_order                        548.4341    10.5823  51.826
## wealth_segmentHigh Net Worth       -59.3131    68.1626  -0.870
## wealth_segmentMass Customer        -56.2541    59.9718  -0.938
## genderMale                          30.2904    48.8043   0.621
## past_3_years_bike_related_purchases   2.9735     0.8514   3.493
## owns_carYes                         70.9106    48.7495   1.455
## tenure                               3.8717     4.7614   0.813
## age                                  0.3760     2.1454   0.175
## property_valuation                 -10.2422     8.6834  -1.180
##                                               Pr(>|t|)
## (Intercept)                                   0.402562
## total_order                       < 0.0000000000000002 ***
## wealth_segmentHigh Net Worth                  0.384288
## wealth_segmentMass Customer                   0.348328
## genderMale                                    0.534885
## past_3_years_bike_related_purchases           0.000487 ***
## owns_carYes                                   0.145905
## tenure                                        0.416211
## age                                           0.860891
## property_valuation                            0.238300
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1237 on 2575 degrees of freedom
## Multiple R-squared:  0.5127, Adjusted R-squared:  0.511
```

```
## F-statistic:   301 on 9 and 2575 DF,  p-value: < 0.00000000000000022
```