Cycling Accessories - Sales Analysis

Semih Barutcu

6/1/2020

Introduction

In this project, I used 4 tables to analyze a medium size bikes & cycling accessories organisation which are generated for KPMG Virtual Internship. This tables are consisting of last year sales (transactions), customer demogrophics (cdemographics), customer address information (caddress) and potential new customers (newcustomer). I wrote the names of datasets in parantheses as I saved.

I used tidyverse package family to analyze the data.

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

```
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)
```

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)))
```

First Look, Handling Incorrect Data and Feature Engineering

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.

```
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
```

```
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
```

```
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.

```
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
```

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

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

Age variable is added to cdemographics and newcustomer datasets.

```
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
```

Exploratory Data Analysis (EDA)

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

dfsummary

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

```
## Switching method to 'browser'
```

```
## Output file written: C:\Users\sbaru\AppData\Local\Temp\RtmpGgWLwo\file310c2dbd2b83.html
```

autoEDA

I arranged the code below as eval = 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 investigation. 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.

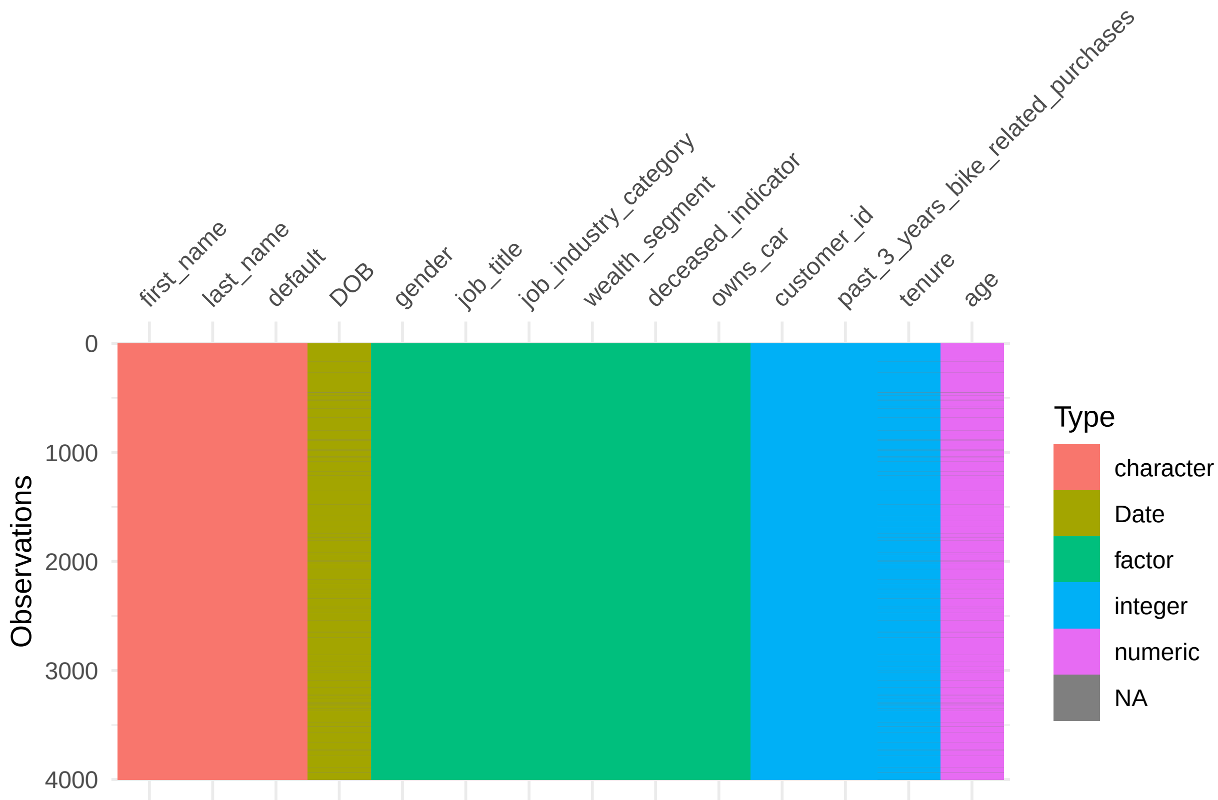
```
autoEDA(cdemographics)
autoEDA(cdemographics, y = "wealth_segment")

autoEDA(caddress)
autoEDA(transactions)
autoEDA(newcustomer)
```

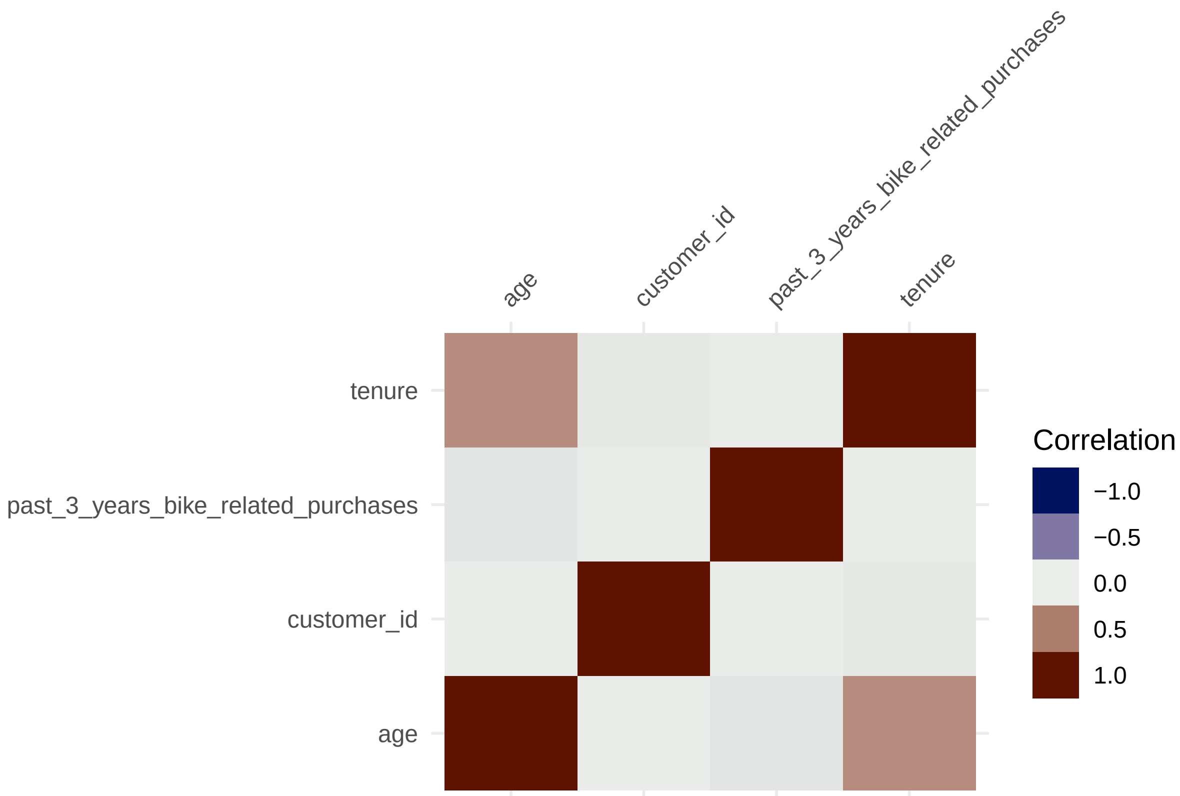
visdat

We can see that age(relatedly DOB) and tenure are missing for some customers. They are somewhat correlated 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.

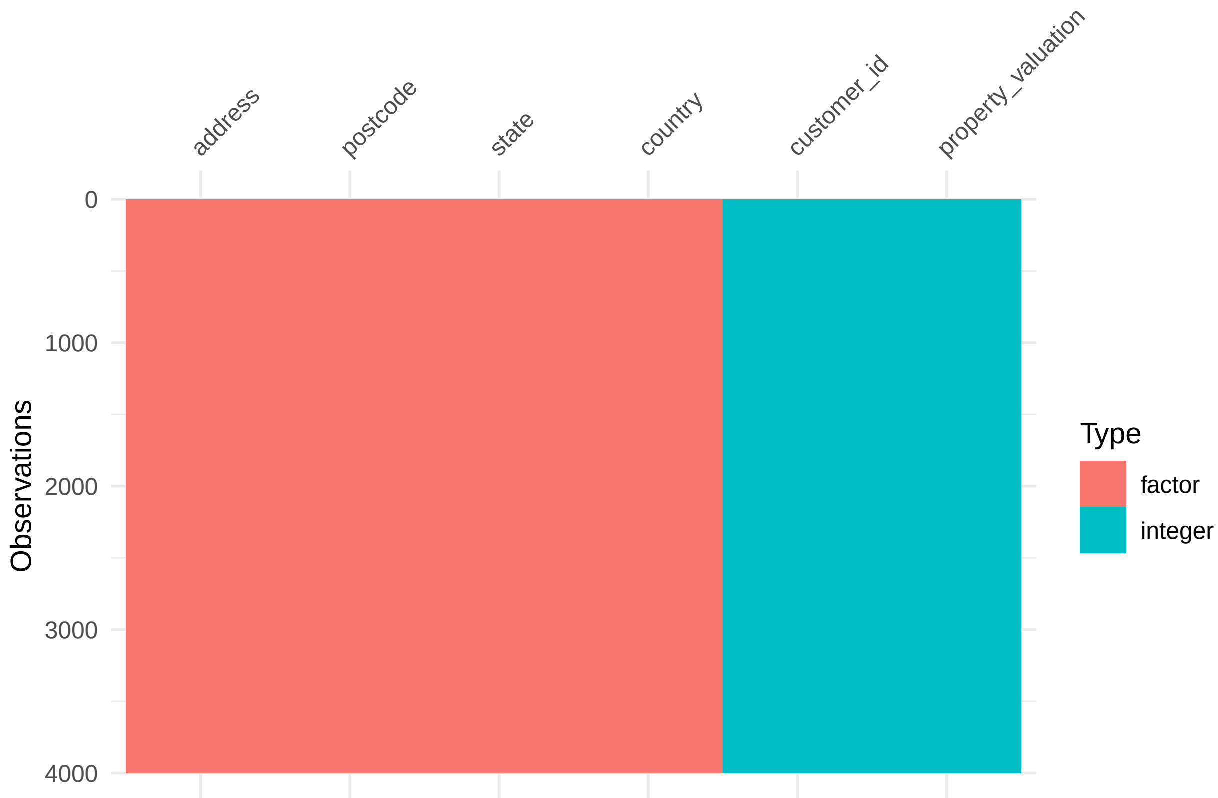
```
vis_dat(cdemographics)
```



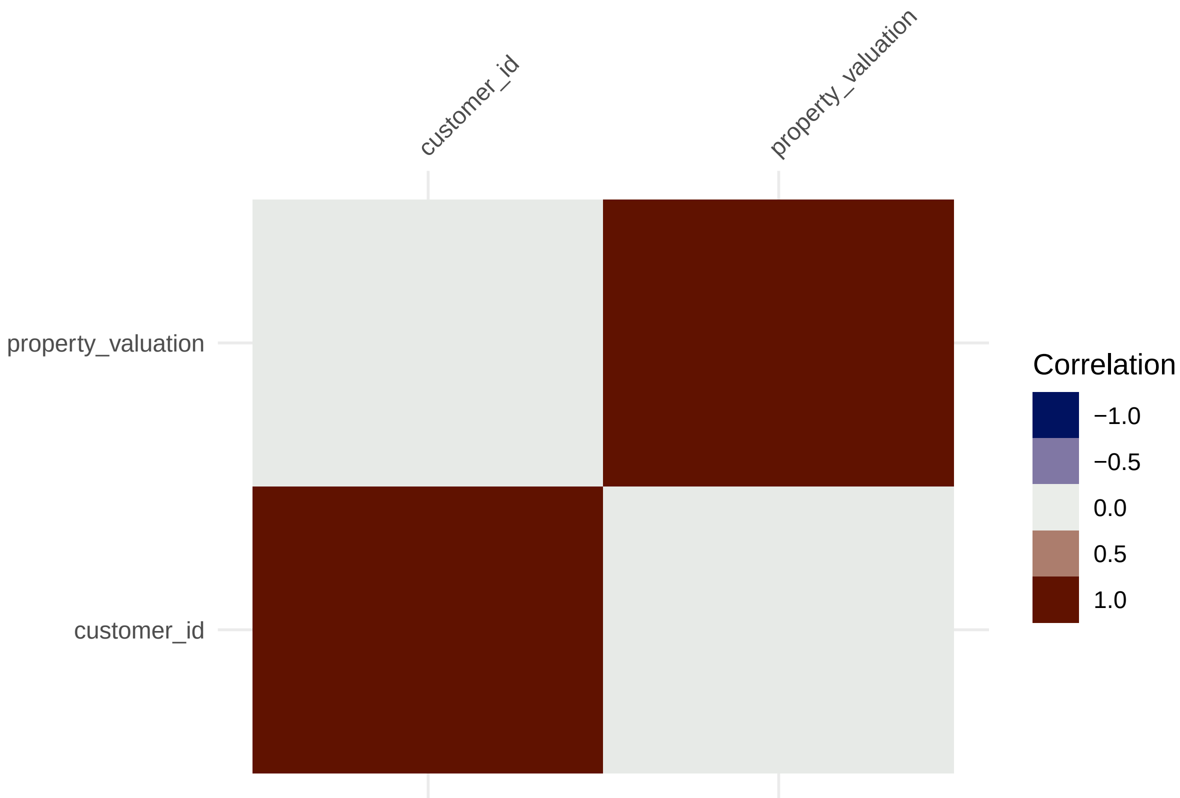

```
cdemographics %>% select_if(is.numeric) %>% vis_cor()
```



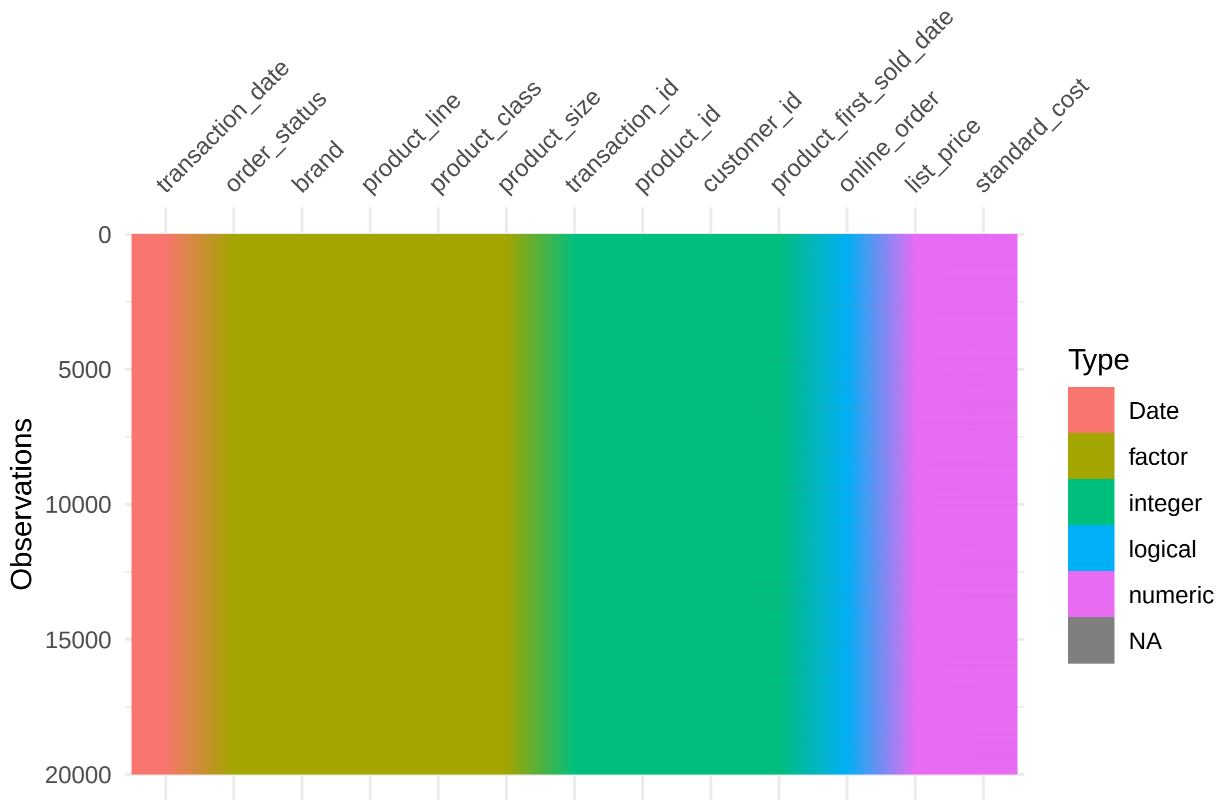

```
vis_dat(caddress)
```



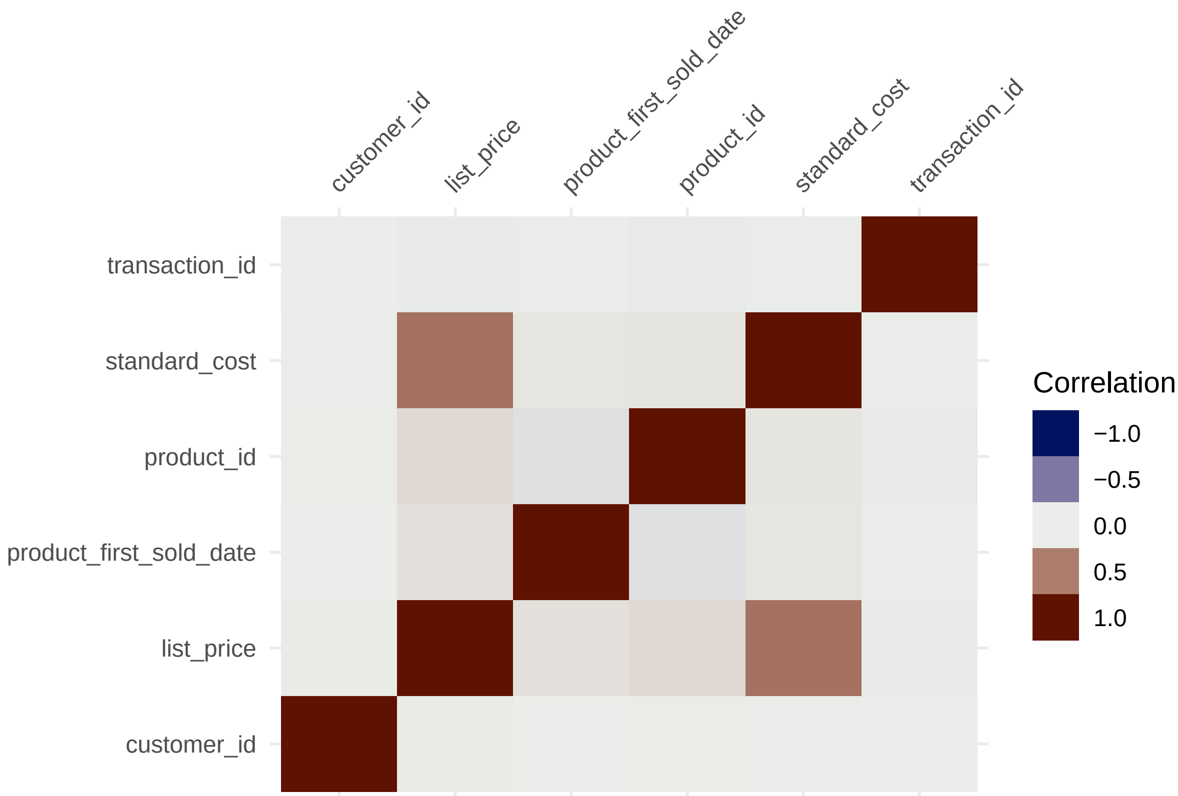

```
caddress %>% select_if(is.numeric) %>% vis_cor()
```



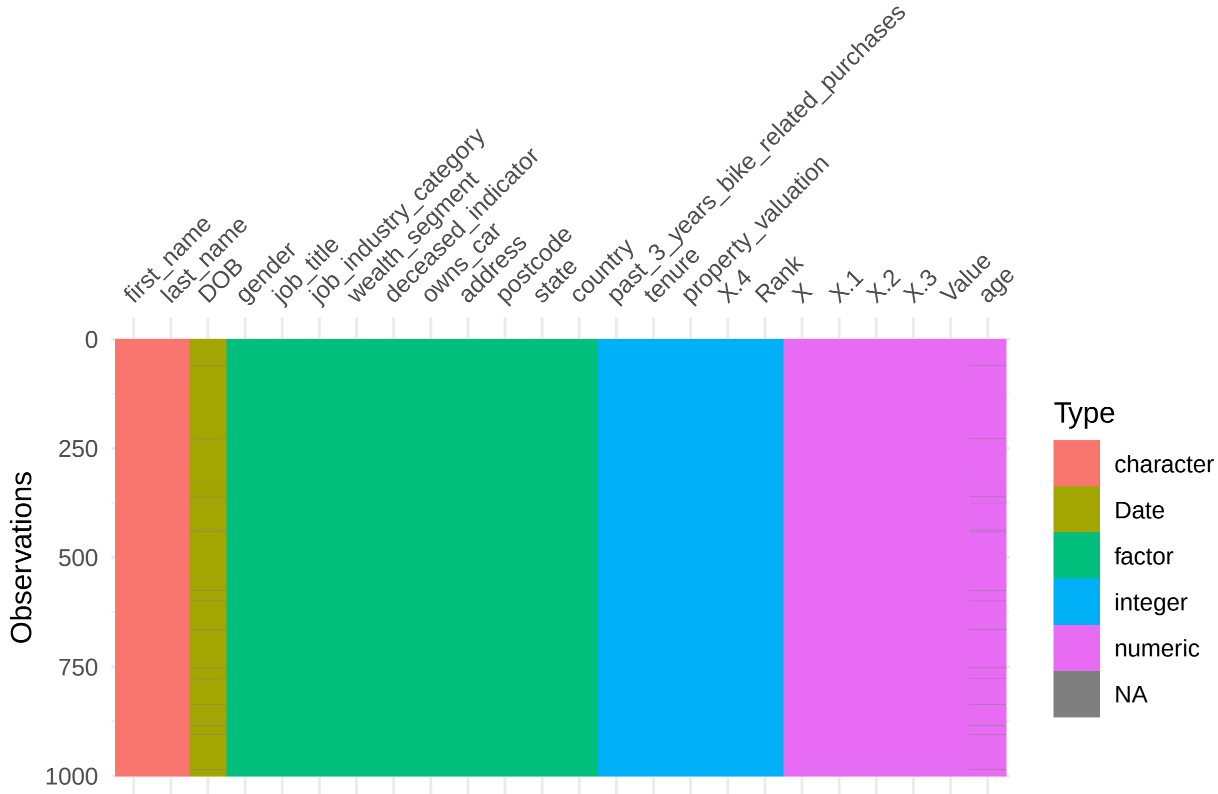

```
vis_dat(transactions)
```



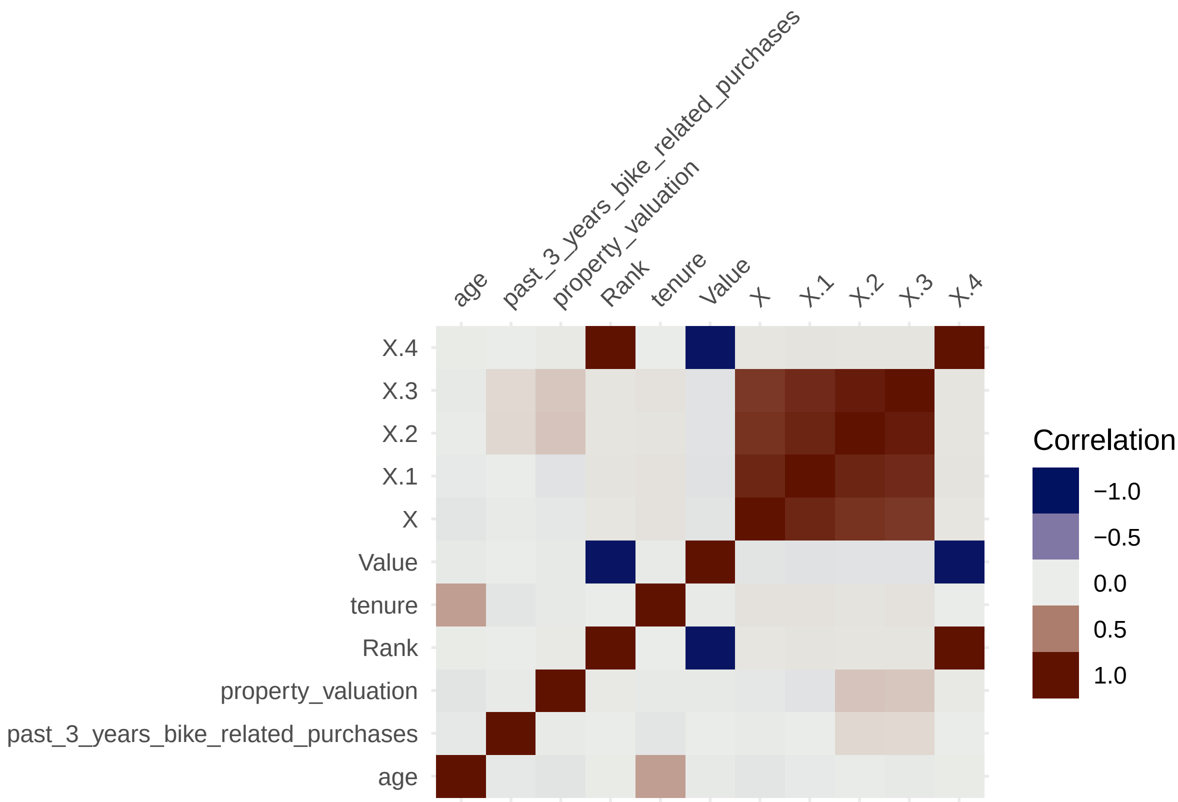

```
transactions %>% select_if(is.numeric) %>% vis_cor()
```




```
vis_dat(newcustomer)
```




```r
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), fill=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) %>%
  summarize(total = n(),
            proportion = total / 88)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)

## # 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()
```

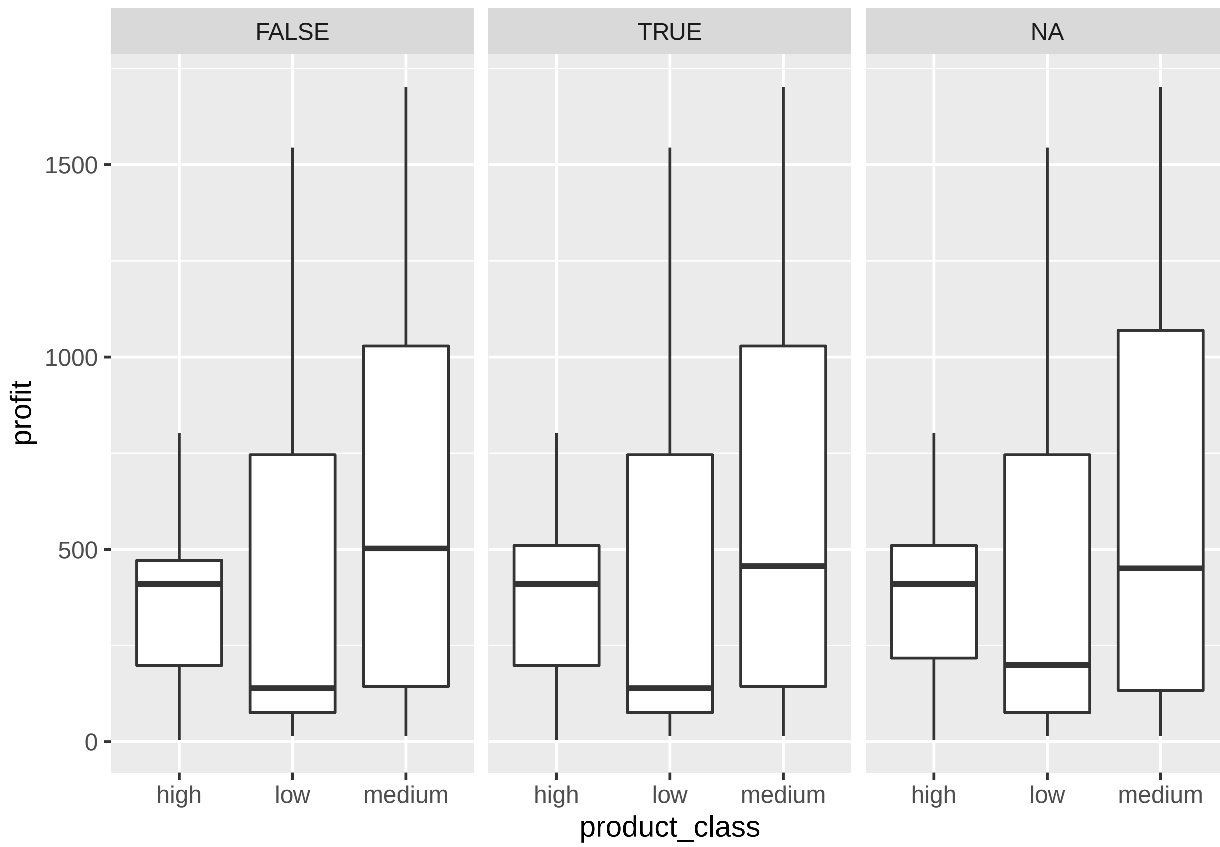
```
## `summarise()` regrouping output by 'product_id' (override with `.groups` argument)
```

```
## # 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.
```

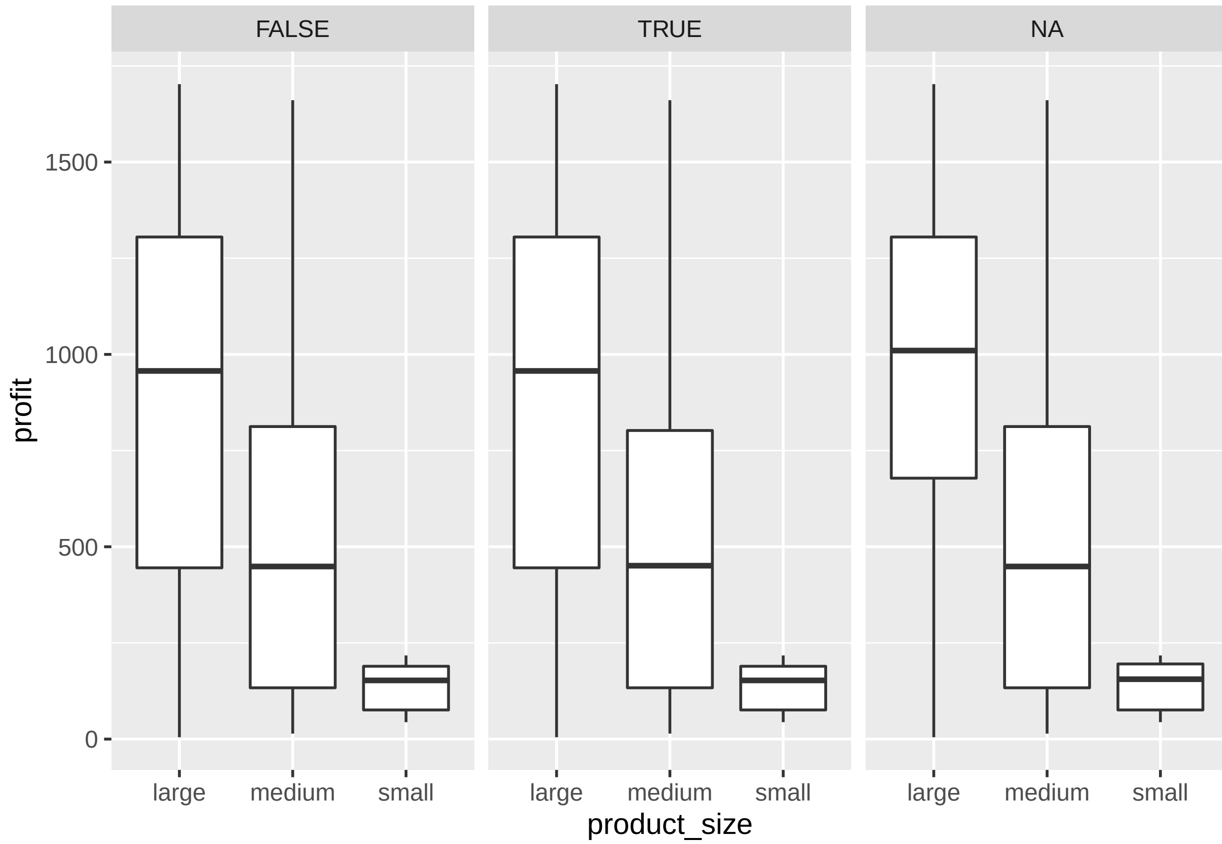
Profit variable has been added to Transactions table. Profit is calculated by difference between list_price and standard_cost.

```
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()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # 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.
```

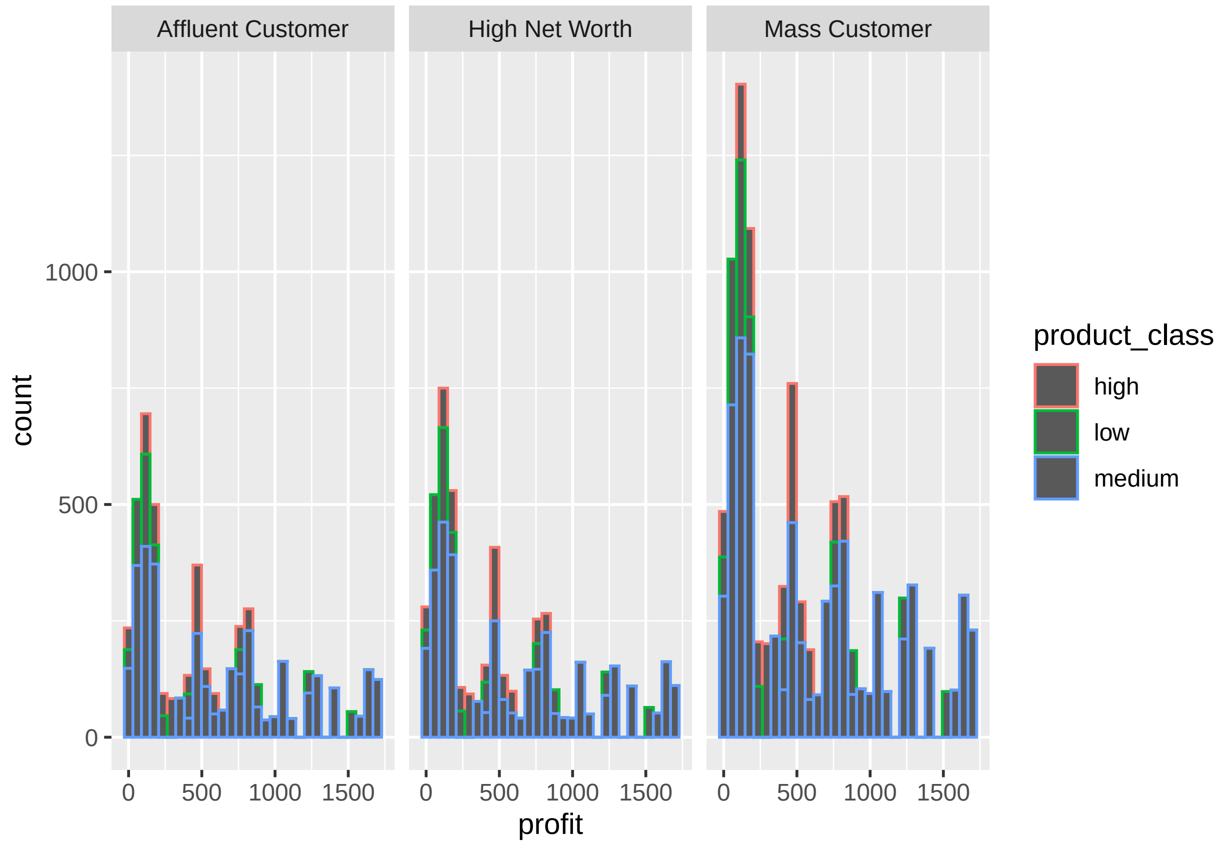
```r
# wealth segment statistics
transactions %>%
  left_join(cdemographics, by="customer_id") %>%
  filter(!is.na(profit)) %>%
  filter(!is.na(wealth_segment)) %>%
  group_by(wealth_segment) %>%
  summarise(total_customer = n_distinct(customer_id),
            total_order= n(),
            order_per_customer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

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # 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`.
```




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

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
transactions_grouped %>% group_by(total_order) %>% summarise(n = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)

## # 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
```

Modeling

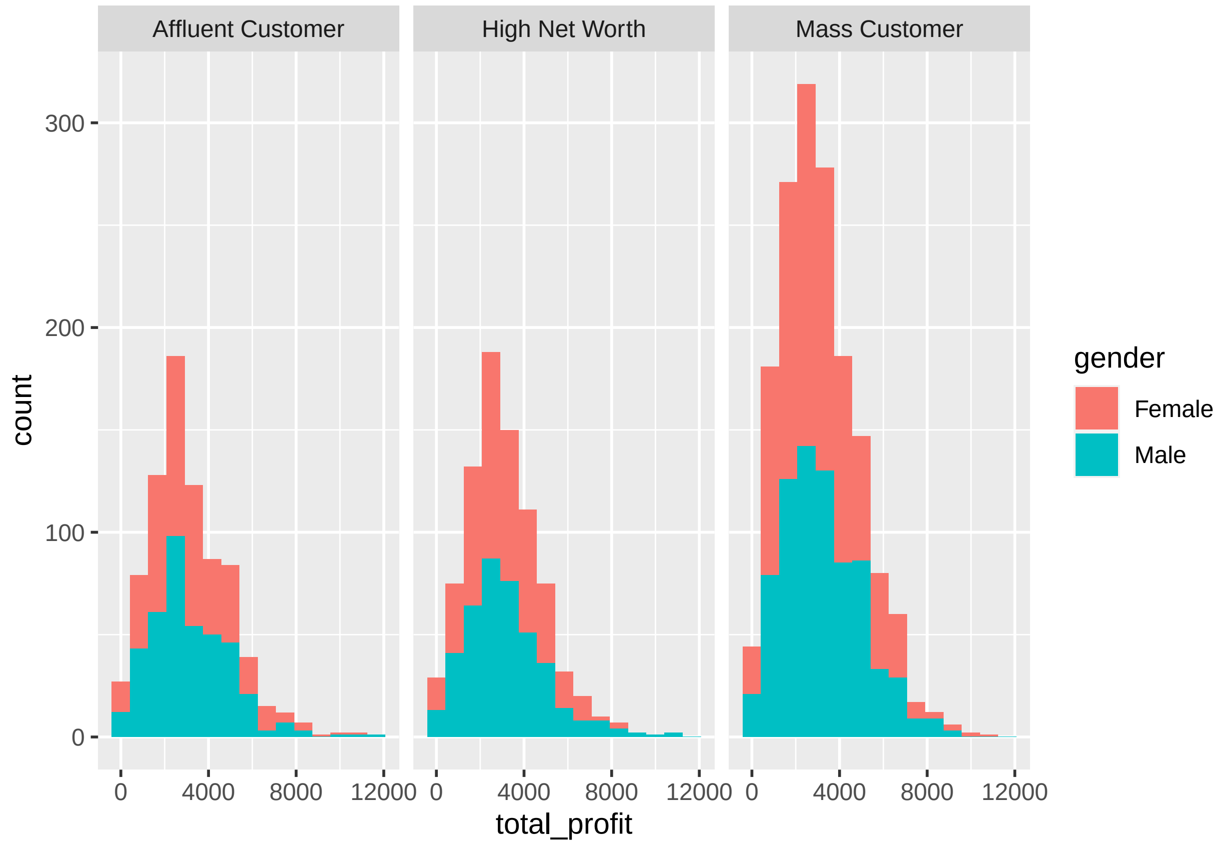
New customers should be categorized subject to given customer demographics data and related datasets. We can join tables to add new features to explore on cdemographics dataset. Firstly, I am going to focus decision tree models.

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, fill=gender)) + geom_histogram(bins=15) + facet_wrap(~wealth_s
```




```
#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
##         0.2460348         0.2553191         0.4986460
```

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

```
##
## Affluent Customer    High Net Worth     Mass Customer
##         0.2430341         0.2693498         0.4876161
```

After constructing a linear model, there isn't a significant predictor for the total profit.

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

```r
summary(lm1)
```

```
##
## Call:
## lm(formula = total_profit ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3931.0  -820.2   -96.3   734.1  6330.1
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         65.6988   140.5946   0.467   0.6403
## total_order                        543.3773    10.5081  51.711   <2e-16 ***
## wealth_segmentHigh Net Worth       -72.3436    68.4905  -1.056   0.2909
## wealth_segmentMass Customer        -14.6109    59.6995  -0.245   0.8067
## genderMale                          17.3147    48.5233   0.357   0.7212
## past_3_years_bike_related_purchases   1.6020     0.8516   1.881   0.0601 .
## owns_carYes                        112.5928    48.5273   2.320   0.0204 *
## tenure                               2.8324     4.7540   0.596   0.5514
## age                                 -1.0562     2.1435  -0.493   0.6222
## property_valuation                 -16.3705     8.5550  -1.914   0.0558 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1231 on 2575 degrees of freedom
## Multiple R-squared:  0.5106, Adjusted R-squared:  0.5088
## F-statistic: 298.4 on 9 and 2575 DF,  p-value: < 2.2e-16
```

Conclusion

After I tried a couple of machine algorithms, I believe this data was created randomly and hard to regularize with any model. While I couldn't explore any meaningful relationship between variables, this project will be a good resource for me with EDA part.