Train and evaluate models with tidymodels

2022-11-19

# Introduction

The goal of this analysis is to **predict with precision the probability of customer churn** and consequently *help in customer retention and mitigation of the risks related to customer churn.*

## Business Value

Below are some the business values that this analysis will create:

1. Reduce churn rate by detecting early clients who are about to churn.

2. Increase customer retention.

3. Understand customer base better based on their activity.

## Data-set

The data set used in this project was sourced from client transaction data recorded since the inception of the banking institution.Based in this data set,informative feature sets were created which were subsequently used in the predictive model.

### Inclusion Criteria

The clients records included in this study were selected from those who were determined to have done at least one transaction within the banking platform.

### Exclusion Criteria

Clients who on boarded the digital platform of the bank and did not make any transaction.

## Project setup and data loading

library(tidymodels);library(feather);library(tidyverse);library(janitor);library(flextable);library(SmartEDA)  
dataset <- readxl::read\_excel("E:/New folder/denis/old/data/rawdata/base\_.xlsx") %>%   
 mutate\_at(c("Min\_LoanTaken","Max\_LoanTaken"),as.numeric)

# Exploratory Data Analysis

## Data Overview

ExpData(dataset) %>%   
 flextable() %>%   
 autofit()

| Descriptions | Value |
| --- | --- |
| Sample size (nrow) | 32506 |
| No. of variables (ncol) | 39 |
| No. of numeric/interger variables | 36 |
| No. of factor variables | 0 |
| No. of text variables | 1 |
| No. of logical variables | 0 |
| No. of identifier variables | 1 |
| No. of date variables | 2 |
| No. of zero variance variables (uniform) | 1 |
| %. of variables having complete cases | 66.67% (26) |
| %. of variables having >0% and <50% missing cases | 33.33% (13) |
| %. of variables having >=50% and <90% missing cases | 0% (0) |
| %. of variables having >=90% missing cases | 0% (0) |

**Observations and inferences.**

* The data set contains 32,506 records of customer transactions with 39 features.
* Only 66.67% of the cases have complete records meaning the remaining 33.33% have missing values.This is normal with most real life data sets that contain missing values.

## Uni-variate Data Analysis.

### Target Variable

dataset %>%   
 group\_by(Churn) %>%   
 tally(name = "Count") %>%   
 mutate(Percent = paste(round(Count/sum(Count),2)\*100,"%")) %>%   
 adorn\_totals('row') %>%   
 flextable() %>%   
 autofit()

| Churn | Count | Percent |
| --- | --- | --- |
| Churn | 4,315 | 13 % |
| Not Churn | 28,191 | 87 % |
| Total | 32,506 | - |

**Observations and Inferences**

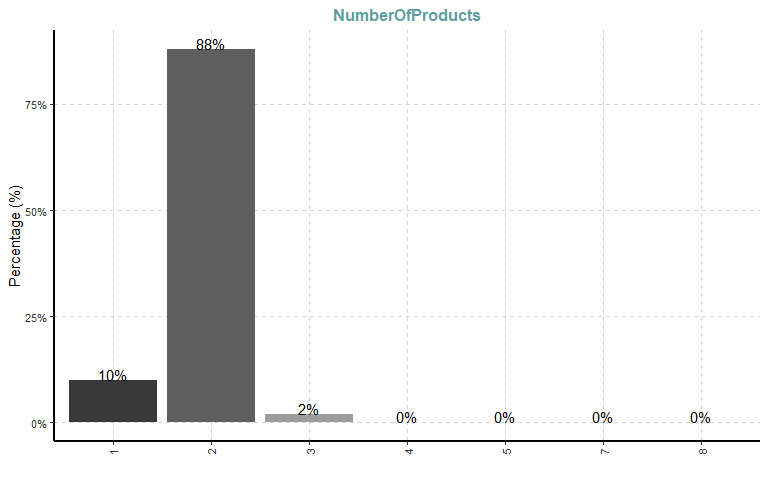
* The institution has suffered a 13% churn rate overtime.

### Explanatory Variables

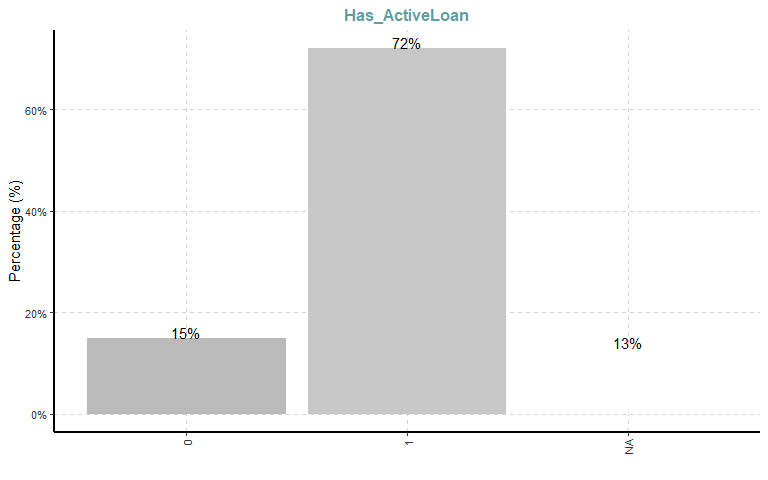
#### Categorical Variables

dataset %>% select(-Churn) %>% ExpCatViz()

## [[1]]



##   
## [[2]]



ExpCTable(dataset %>% select(NumberOfProducts,Has\_ActiveLoan)) %>%   
 flextable() %>% autofit()

| Variable | Valid | Frequency | Percent | CumPercent |
| --- | --- | --- | --- | --- |
| NumberOfProducts | 1 | 3,128 | 9.62 | 9.62 |
| NumberOfProducts | 2 | 28,540 | 87.80 | 97.42 |
| NumberOfProducts | 3 | 789 | 2.43 | 99.85 |
| NumberOfProducts | 4 | 44 | 0.14 | 99.99 |
| NumberOfProducts | 5 | 3 | 0.01 | 100.00 |
| NumberOfProducts | 7 | 1 | 0.00 | 100.00 |
| NumberOfProducts | 8 | 1 | 0.00 | 100.00 |
| NumberOfProducts | TOTAL | 32,506 |  |  |
| Has\_ActiveLoan | 0 | 4,788 | 14.73 | 14.73 |
| Has\_ActiveLoan | 1 | 23,403 | 72.00 | 86.73 |
| Has\_ActiveLoan | NA | 4,315 | 13.27 | 100.00 |
| Has\_ActiveLoan | TOTAL | 32,506 |  |  |

**Observation & inferences**

* *Number of products* : Majority of clients (88%) subscribed to at least two of the financial institution’s digital products while a very small percentage subscribed to more than 2 products.
* *Has\_Active\_Loan* : 72% of the clients were still servicing loans with the institution.

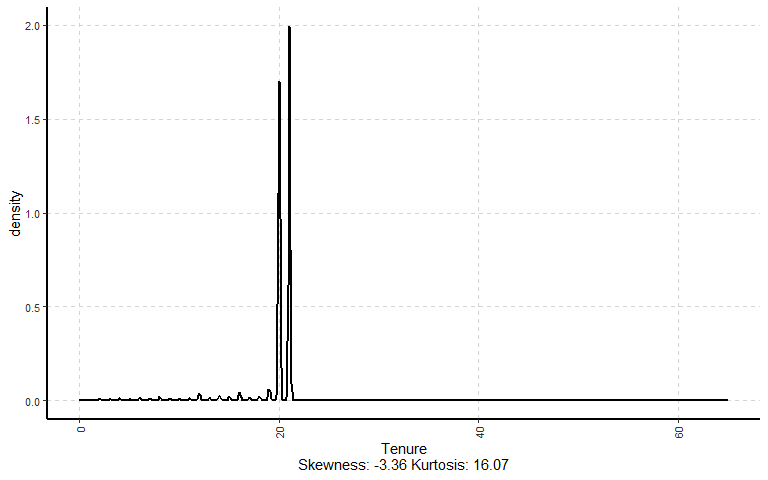
#### Numeric Variables

num\_vars <- dataset %>% select(-c(Churn,ClientID,NumberOfProducts,Has\_ActiveLoan))  
num\_vars %>%   
 # select(Tenure,No\_of\_transactions) %>%   
 ExpNumStat() %>% select(-c(Group,NegInf,PosInf,Per\_of\_Missing,CV,nNeg,nZero,NA\_Value,nPos)) %>% #view()  
 flextable(theme\_fun = set\_flextable\_defaults(font.size = 12,padding = 0)) %>% autofit()

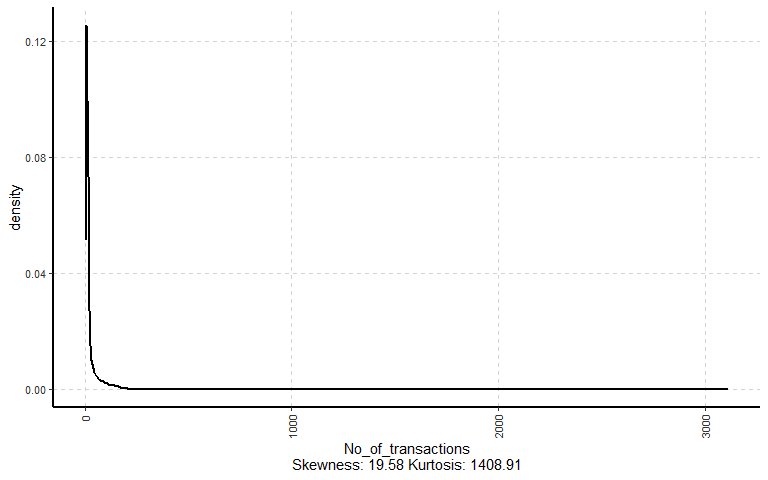
| Vname | TN | sum | min | max | mean | median | SD | IQR | Skewness | Kurtosis |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ave\_LoanTaken | 32,506 | 219,381,812.83 | 992.250 | 1,200,000.000 | 7,781.981 | 5,264.000 | 19,061.802 | 9,705.770 | 39.630 | 2,063.211 |
| Average\_DepositsMade | 32,506 | 48,990,493.92 | 0.000 | 275,108.800 | 1,507.122 | 890.286 | 2,637.840 | 1,687.637 | 41.896 | 3,795.501 |
| Days\_Since\_LastTransaction | 32,506 | 13,512,406.00 | 2.000 | 751.000 | 415.690 | 567.000 | 236.566 | 430.000 | -0.604 | -1.302 |
| DaysSince\_Last\_Loan | 32,506 | 14,090,399.00 | 1.000 | 758.000 | 499.819 | 610.000 | 210.870 | 153.000 | -1.440 | 0.413 |
| deposits\_12 | 32,506 | 19,756,971.75 | 0.000 | 4,000,000.000 | 607.795 | 0.000 | 30,115.963 | 0.000 | 88.129 | 10,167.130 |
| deposits\_3 | 32,506 | 17,515,829.06 | 0.000 | 4,000,000.000 | 538.849 | 0.000 | 29,269.198 | 0.000 | 93.794 | 11,308.800 |
| deposits\_6 | 32,506 | 17,712,001.06 | 0.000 | 4,000,000.000 | 544.884 | 0.000 | 29,270.402 | 0.000 | 93.781 | 11,306.861 |
| First\_Loan\_Last\_Loan\_diff | 32,506 | -11,016,050.64 | -600,000.000 | 1,272,500.000 | -390.765 | 0.000 | 14,254.156 | 0.000 | 61.520 | 5,535.873 |
| First\_Loan\_Last\_Loan\_Ratio | 32,506 | 27,700.62 | 0.032 | 159.439 | 0.983 | 1.000 | 1.161 | 0.000 | 107.632 | 13,426.885 |
| FirstLoanTaken | 32,506 | 218,767,007.16 | 590.000 | 1,300,000.000 | 7,760.172 | 5,040.000 | 23,958.970 | 9,296.000 | 37.994 | 1,729.219 |
| last\_deposit\_amount | 32,506 | 171,281,930.14 | 0.000 | 1,434,950.000 | 5,269.240 | 1,120.000 | 12,464.793 | 5,798.750 | 50.565 | 5,437.302 |
| Last\_Loan\_First\_Loan\_Ratio | 32,506 | 32,030.50 | 0.006 | 31.000 | 1.136 | 1.000 | 0.934 | 0.000 | 15.310 | 322.351 |
| last\_withdrawal\_amount | 32,506 | 406,870,976.06 | 0.000 | 2,480,237.000 | 12,516.796 | 6,272.000 | 36,721.872 | 13,440.000 | 40.471 | 2,149.640 |
| LastLoanTaken | 32,506 | 229,783,057.80 | 1,100.000 | 1,200,000.000 | 8,150.937 | 5,488.000 | 20,717.717 | 9,968.000 | 39.354 | 1,961.209 |
| Max\_ArrearsDays | 32,506 | 10,990,129.00 | -28.000 | 615.000 | 389.845 | 573.000 | 260.904 | 578.000 | -0.668 | -1.425 |
| Max\_DepositsMade | 32,506 | 255,172,590.62 | 0.000 | 4,000,000.000 | 7,850.015 | 4,145.000 | 37,288.562 | 8,960.000 | 58.958 | 4,876.754 |
| Max\_LoanTaken | 32,506 | 221,291,543.12 | 590.000 | 1,200,000.000 | 7,849.723 | 5,488.000 | 19,677.546 | 8,780.000 | 38.003 | 1,881.078 |
| Max\_WithdrawalMade | 32,506 | 251,339,561.87 | 0.000 | 4,000,000.000 | 7,732.098 | 4,070.000 | 36,663.375 | 8,960.000 | 59.653 | 5,074.726 |
| Min\_LoanTaken | 32,506 | 217,099,535.84 | 1,100.000 | 1,250,000.000 | 7,701.023 | 4,816.000 | 23,379.553 | 9,880.000 | 39.812 | 1,865.386 |
| no\_of\_LoanTaken | 32,506 | 78,511.00 | 1.000 | 40.000 | 2.785 | 1.000 | 3.911 | 1.000 | 2.866 | 8.505 |
| No\_of\_transactions | 32,506 | 720,251.00 | 1.000 | 3,114.000 | 22.157 | 8.000 | 37.672 | 13.000 | 19.584 | 1,408.912 |
| No\_of\_transactions\_Deposits | 32,506 | 202,110.00 | 0.000 | 939.000 | 6.218 | 2.000 | 13.166 | 5.000 | 18.756 | 951.597 |
| No\_of\_transactions\_Withdrawals | 32,506 | 506,555.00 | 0.000 | 2,161.000 | 15.583 | 6.000 | 25.811 | 7.000 | 19.920 | 1,475.522 |
| Tenure | 32,506 | 636,175.00 | 0.000 | 65.000 | 19.571 | 20.000 | 3.270 | 1.000 | -3.363 | 16.068 |
| TotalDepositsMade | 32,506 | 1,325,371,994.74 | 0.000 | 15,869,967.980 | 40,773.149 | 6,272.500 | 173,758.386 | 26,498.750 | 39.236 | 2,721.838 |
| transactions\_made\_12 | 32,506 | 278,590.00 | 0.000 | 665.000 | 8.570 | 0.000 | 22.027 | 3.000 | 4.882 | 58.922 |
| transactions\_made\_3 | 32,506 | 64,502.00 | 0.000 | 408.000 | 1.984 | 0.000 | 6.769 | 0.000 | 11.090 | 443.807 |
| transactions\_made\_6 | 32,506 | 128,083.00 | 0.000 | 579.000 | 3.940 | 0.000 | 11.949 | 0.000 | 7.830 | 206.111 |
| WithdrawalMade | 32,506 | 1,483,086,187.43 | 0.000 | 16,171,410.720 | 45,624.998 | 10,960.000 | 175,183.463 | 35,840.000 | 39.289 | 2,794.307 |
| withdrawals\_12 | 32,506 | 92,470,697.40 | 0.000 | 700,000.000 | 2,844.727 | 0.000 | 7,107.285 | 3,472.000 | 31.431 | 2,876.845 |
| withdrawals\_3 | 32,506 | 46,664,953.64 | 0.000 | 220,000.000 | 1,435.580 | 0.000 | 4,540.592 | 0.000 | 7.132 | 181.032 |
| withdrawals\_6 | 32,506 | 65,115,775.04 | 0.000 | 220,000.000 | 2,003.192 | 0.000 | 5,232.498 | 0.000 | 5.338 | 103.307 |

num\_vars %>% ExpNumViz(type = 2)

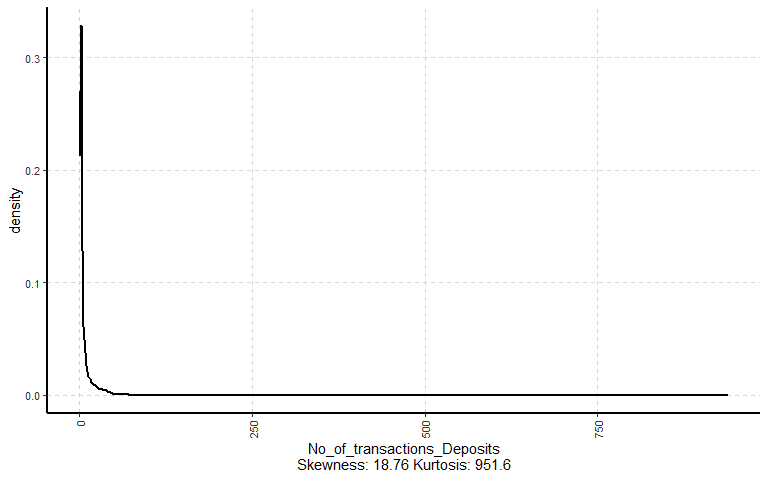
## [[1]]



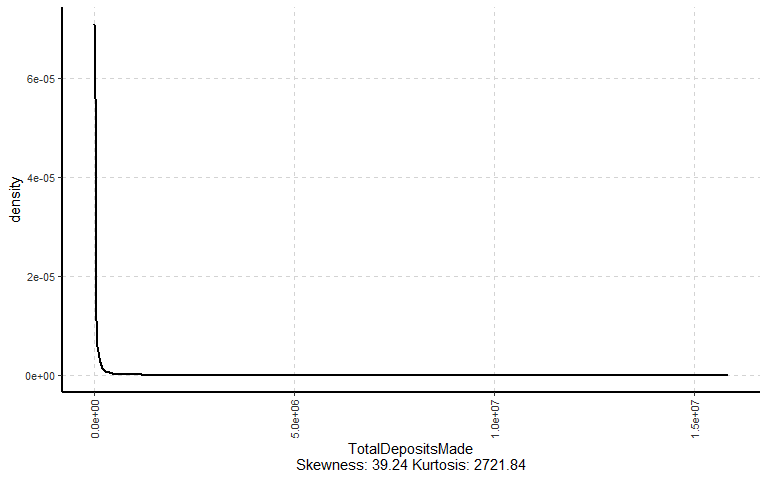
##   
## [[2]]



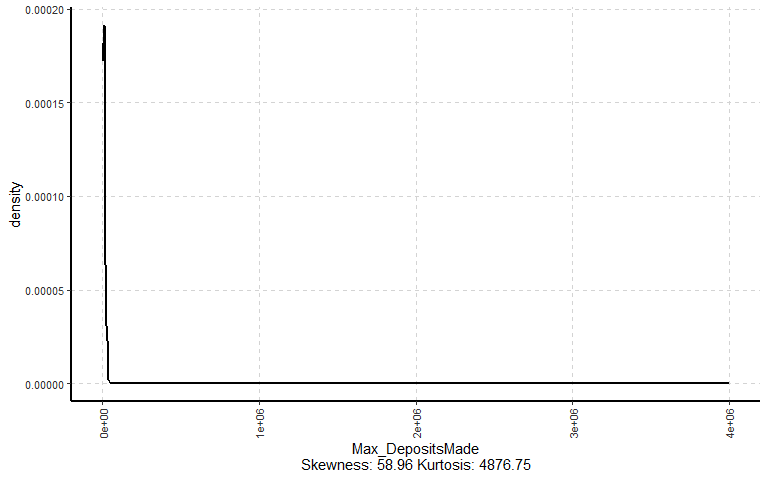
##   
## [[3]]



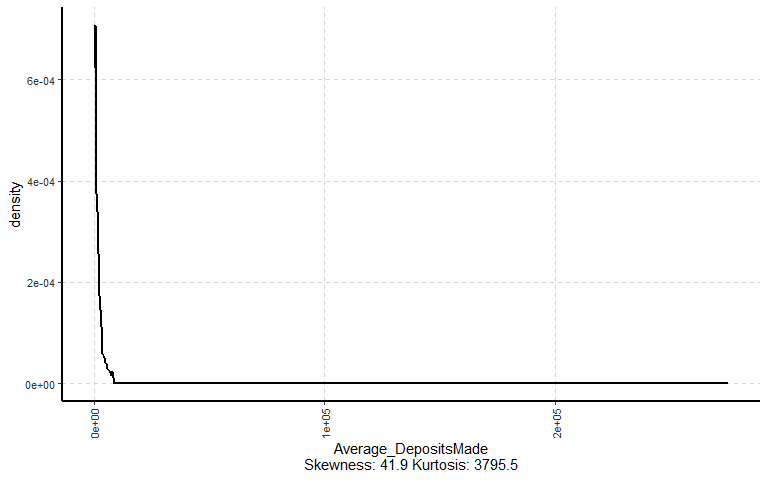
##   
## [[4]]



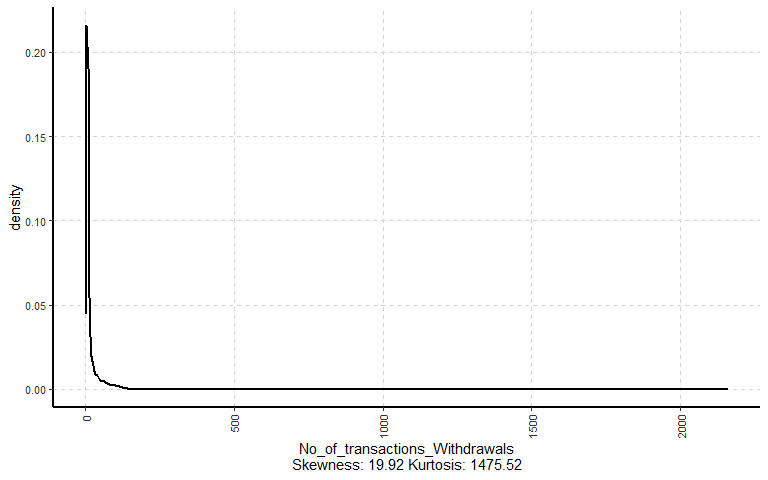
##   
## [[5]]



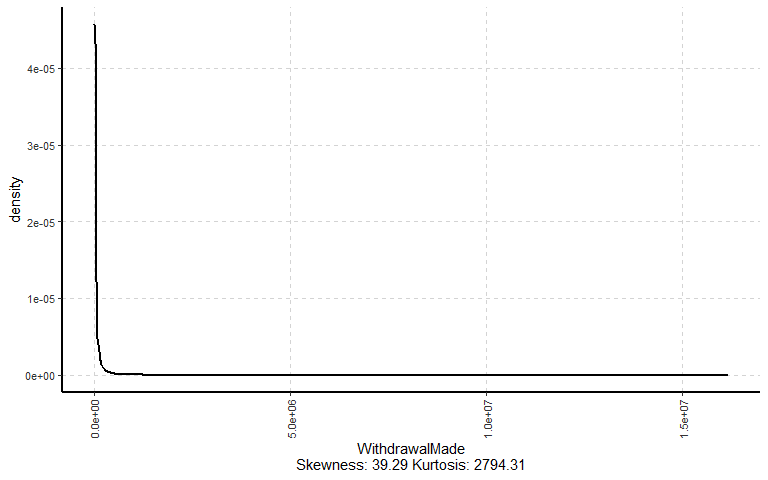
##   
## [[6]]



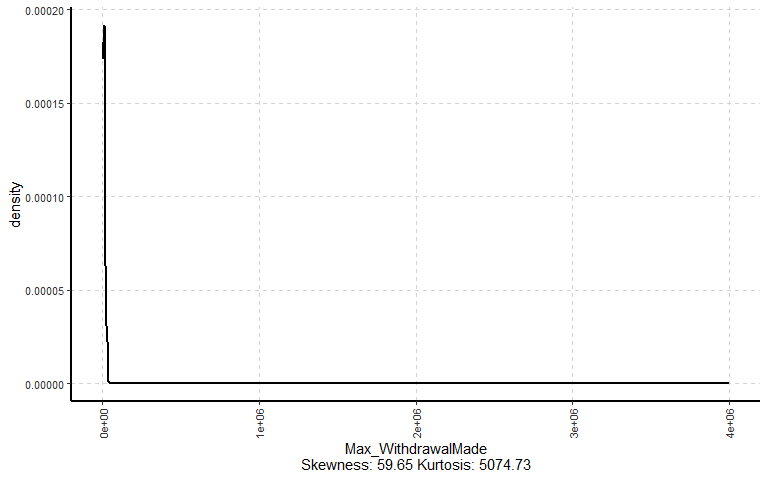
##   
## [[7]]



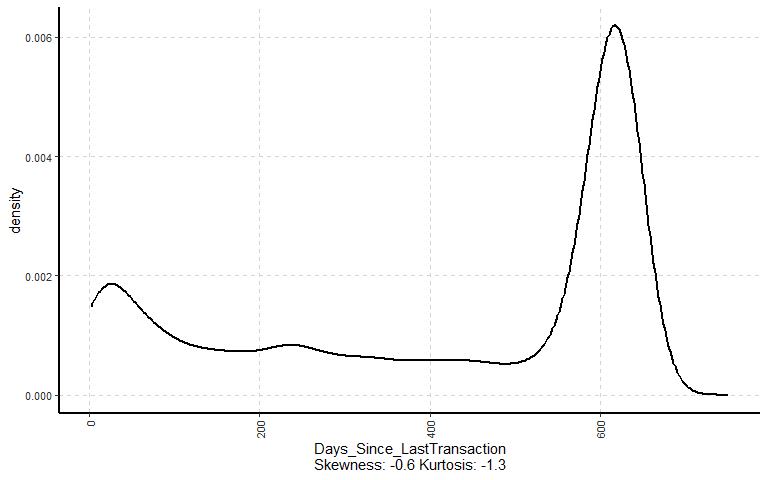
##   
## [[8]]



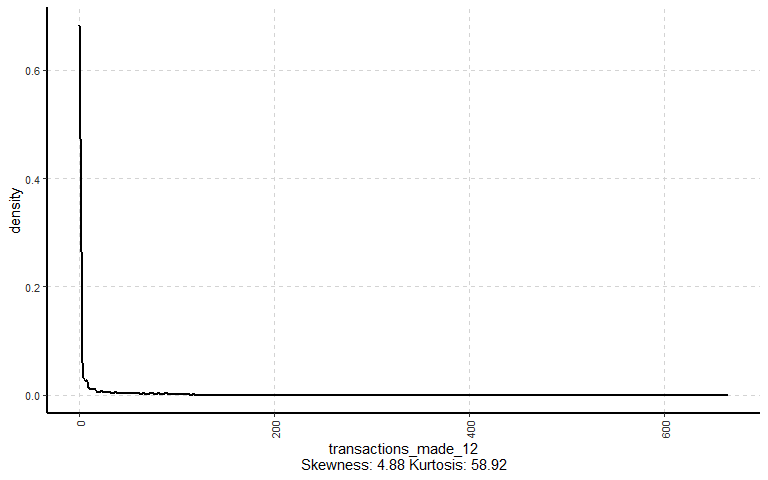
##   
## [[9]]



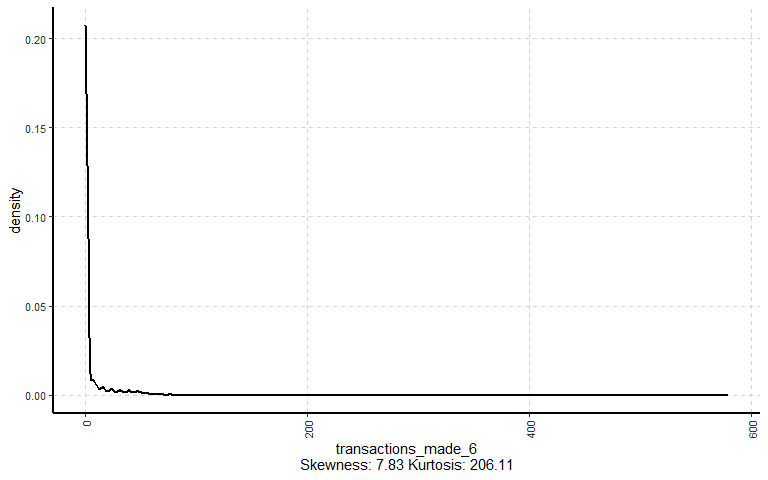
##   
## [[10]]



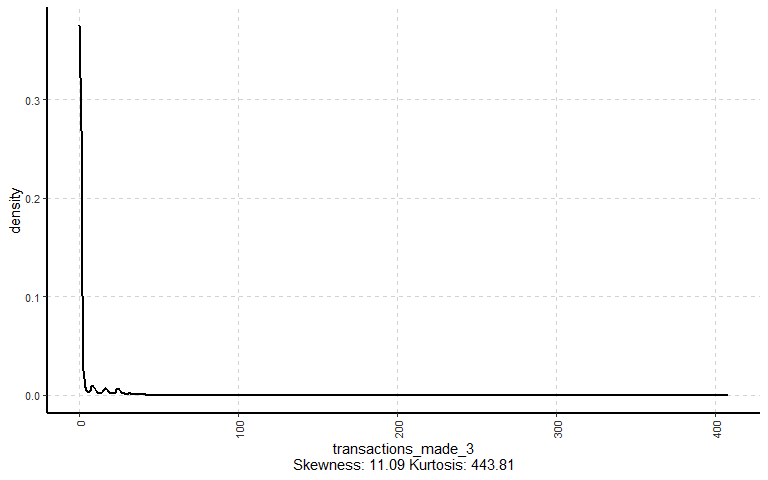
##   
## [[11]]



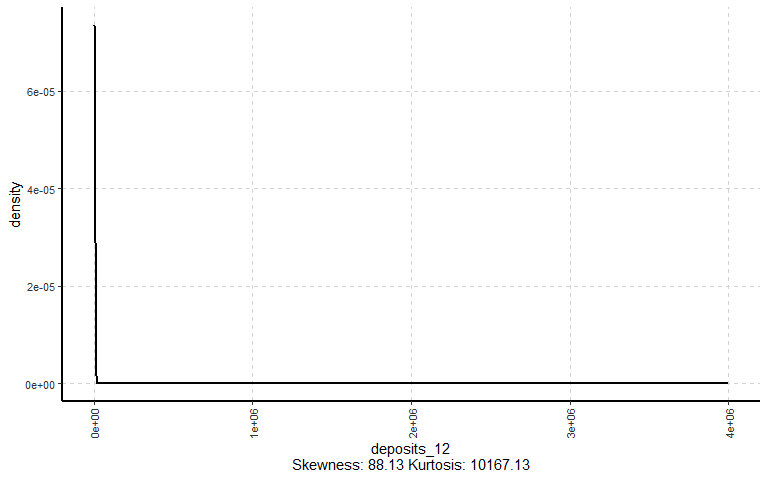
##   
## [[12]]



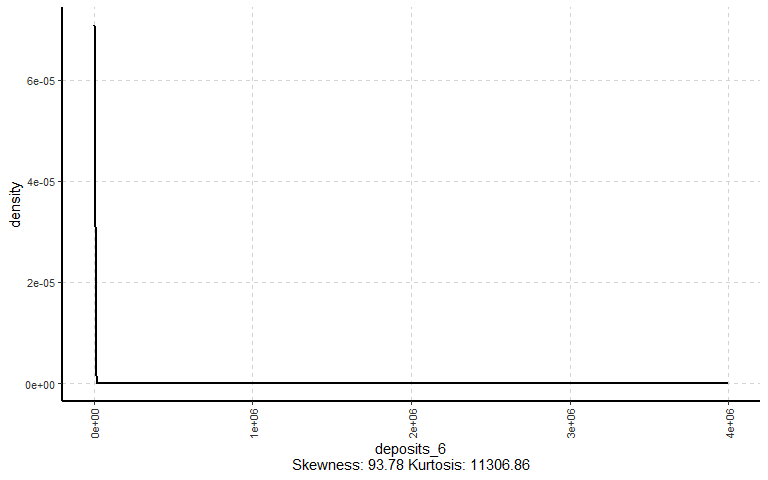
##   
## [[13]]



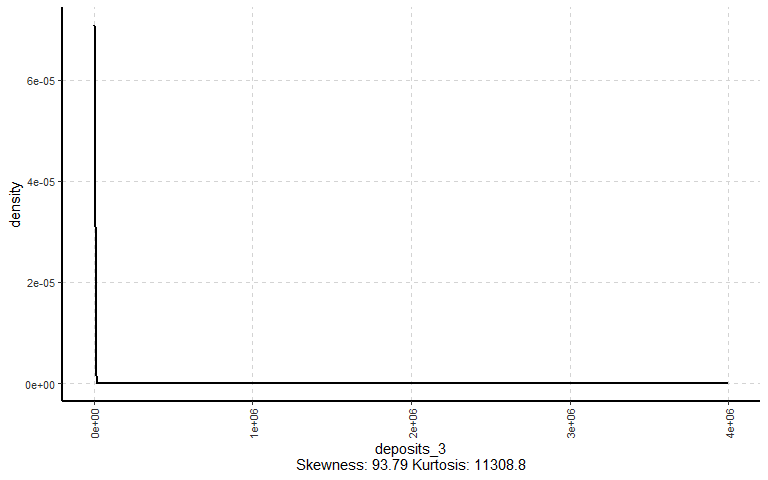
##   
## [[14]]



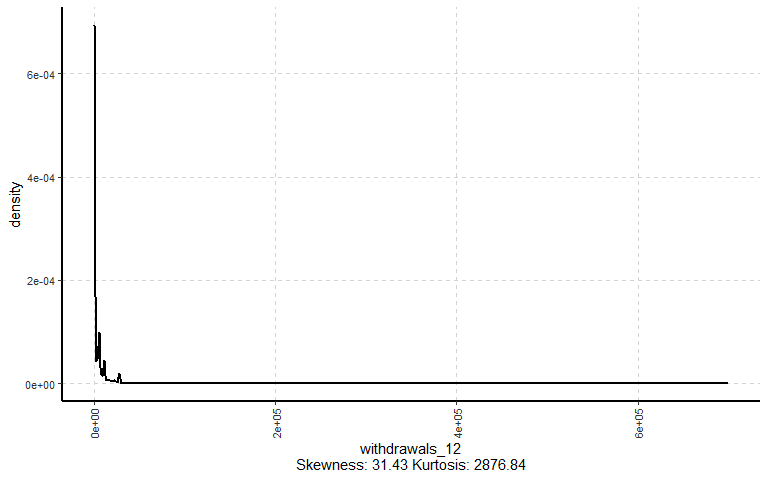
##   
## [[15]]



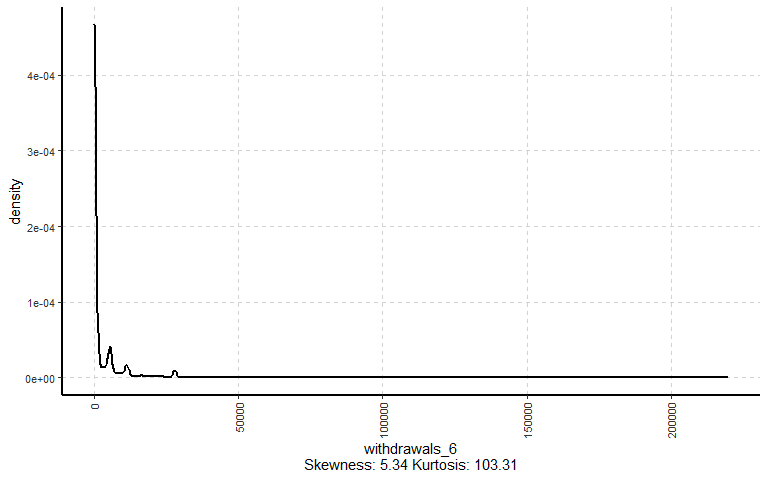
##   
## [[16]]



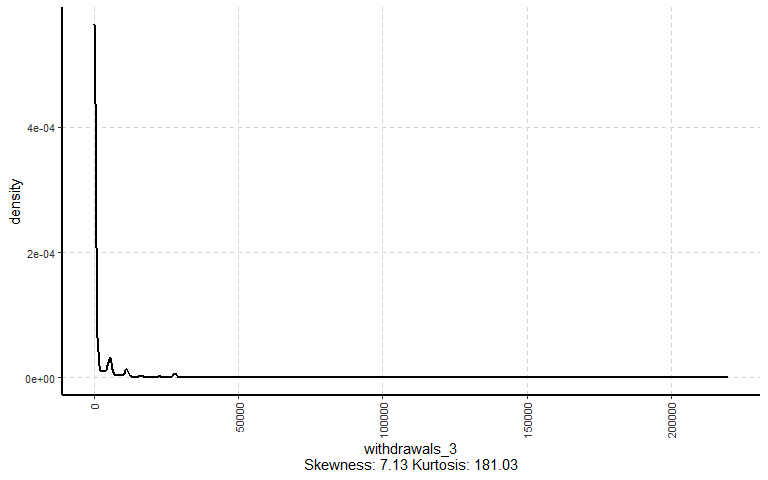
##   
## [[17]]



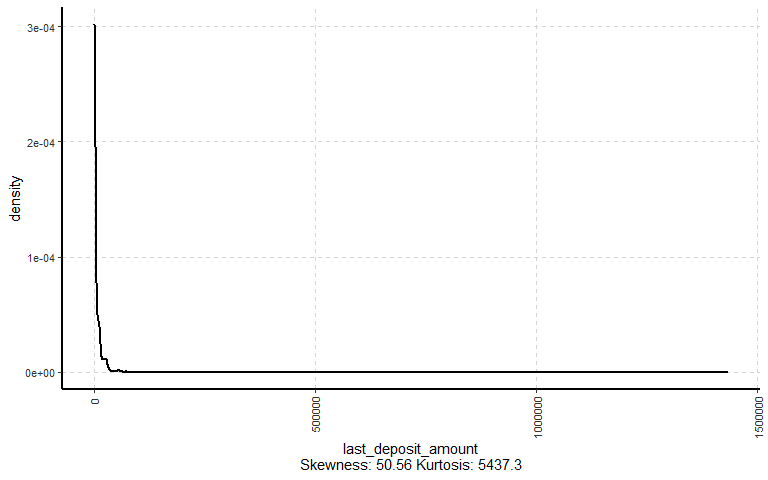
##   
## [[18]]



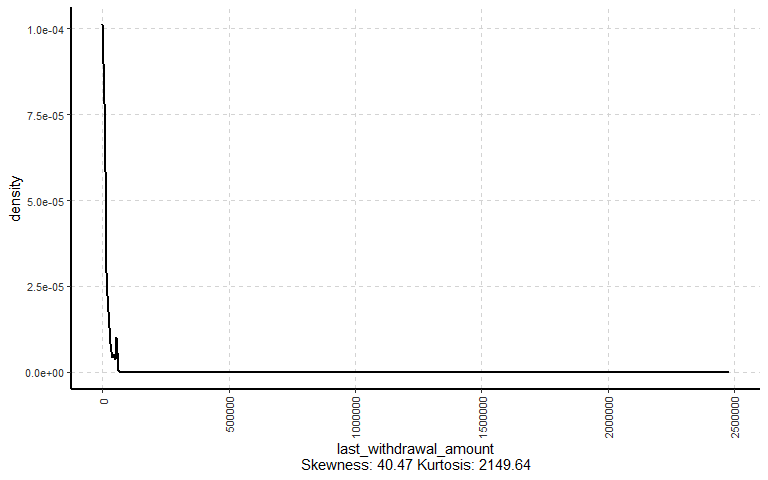
##   
## [[19]]



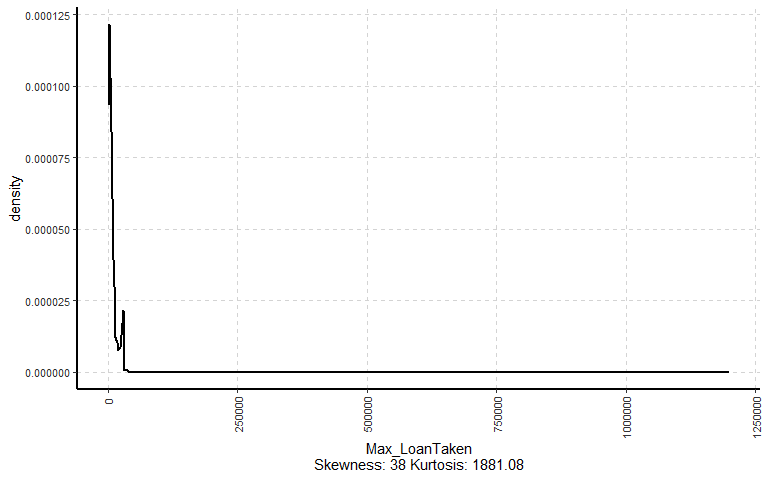
##   
## [[20]]



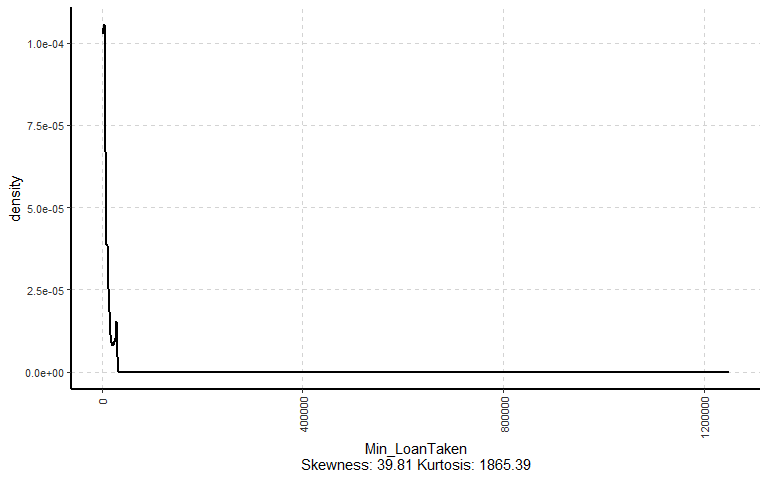
##   
## [[21]]



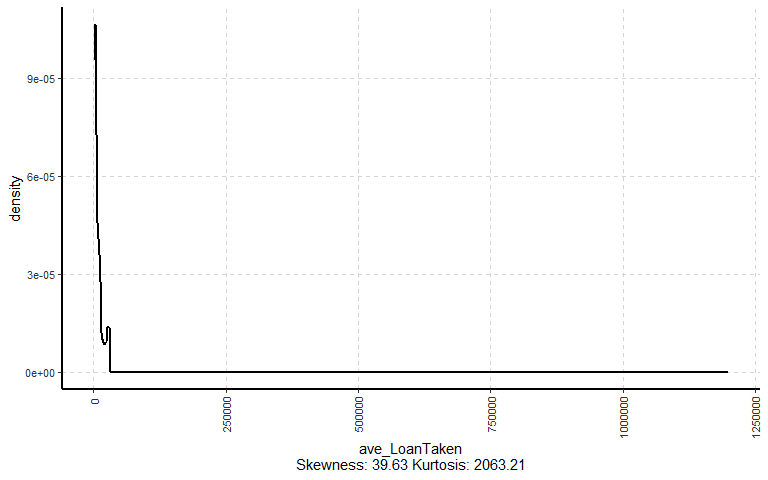
##   
## [[22]]



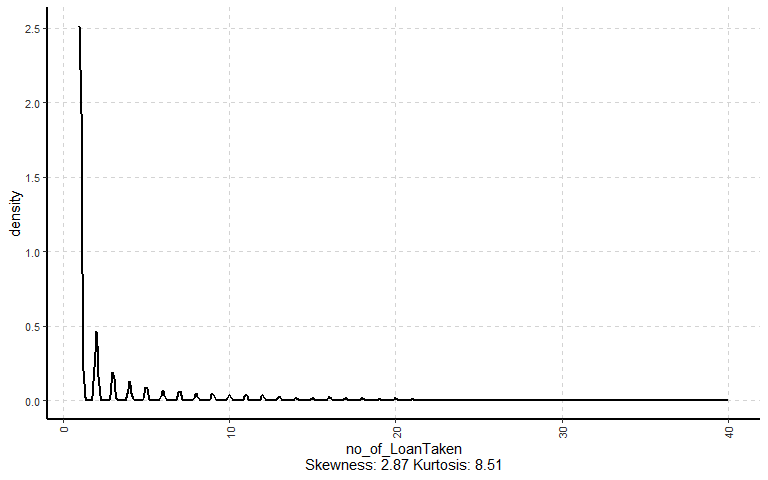
##   
## [[23]]



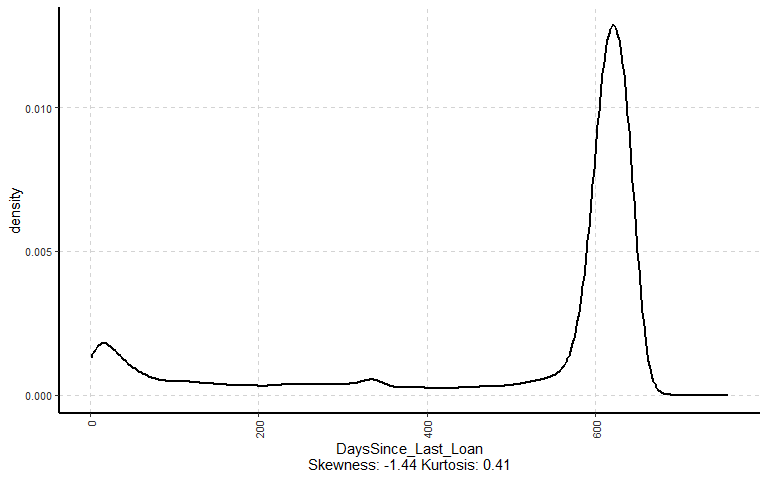
##   
## [[24]]



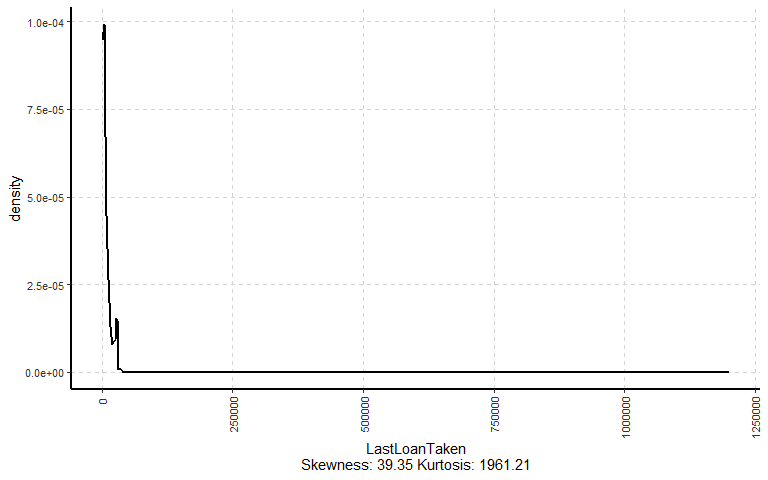
##   
## [[25]]



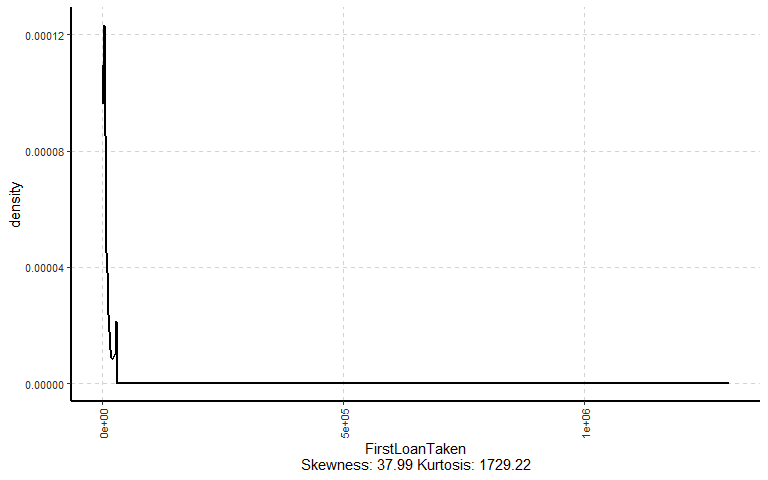
##   
## [[26]]



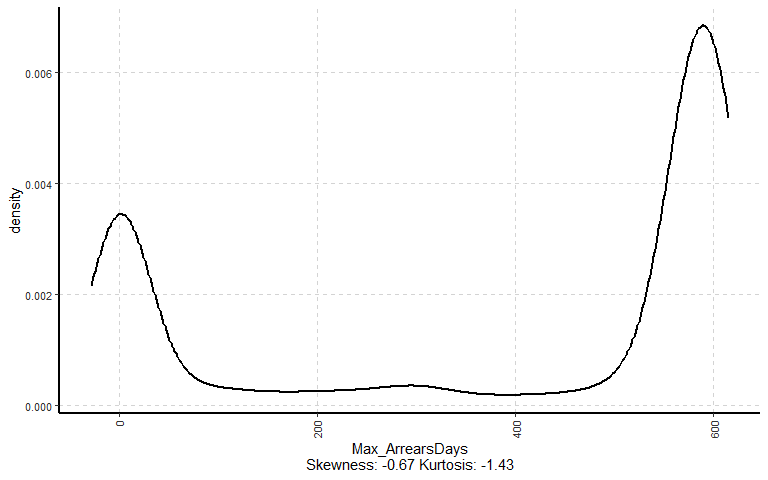
##   
## [[27]]



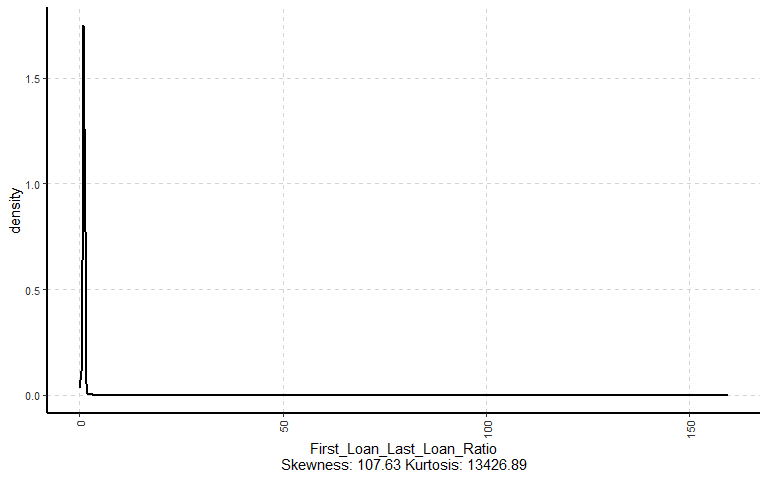
##   
## [[28]]



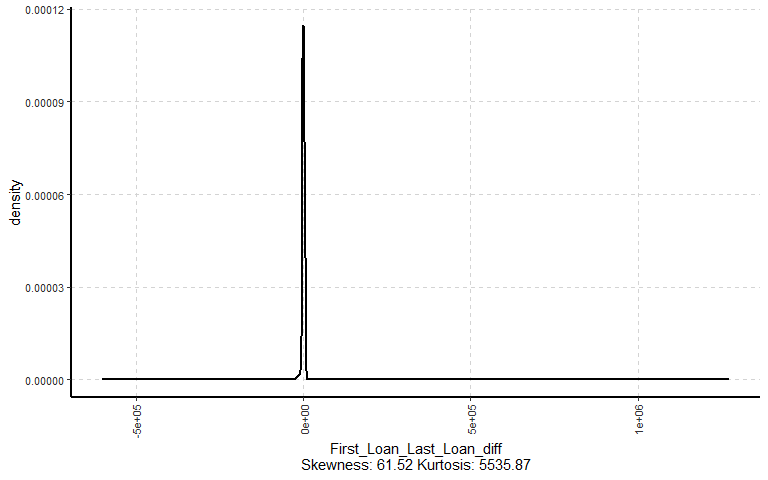
##   
## [[29]]



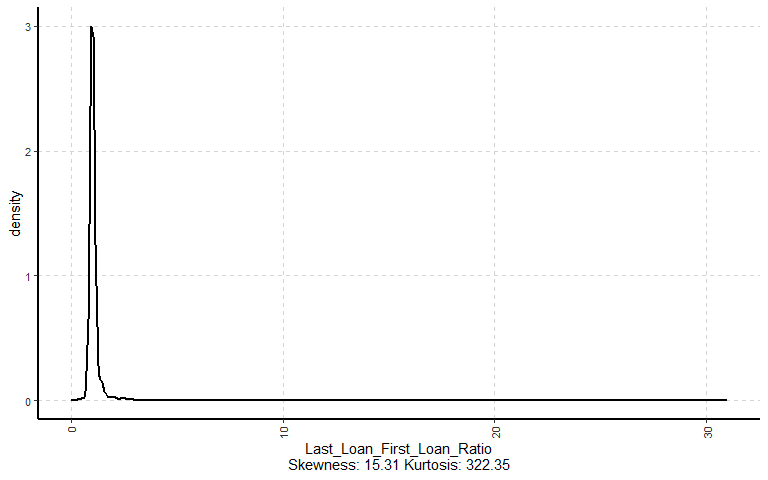
##   
## [[30]]



##   
## [[31]]



##   
## [[32]]



**Observations and inferences**

.

## Bivariate Distribution

### Categorical Variables Vs Target Variables

dataset %>%   
 select(c(Churn,NumberOfProducts,Has\_ActiveLoan)) %>%   
 ExpCTable(Target = 'Churn',) %>% flextable() %>% autofit()

| VARIABLE | CATEGORY | Churn:Churn | Churn:Not Churn | TOTAL |
| --- | --- | --- | --- | --- |
| NumberOfProducts | 1 | 3,128 | 0 | 3,128 |
| NumberOfProducts | 2 | 1,152 | 27,388 | 28,540 |
| NumberOfProducts | 3 | 35 | 754 | 789 |
| NumberOfProducts | 4 | 0 | 44 | 44 |
| NumberOfProducts | 5 | 0 | 3 | 3 |
| NumberOfProducts | 7 | 0 | 1 | 1 |
| NumberOfProducts | 8 | 0 | 1 | 1 |
| NumberOfProducts | TOTAL | 4,315 | 28,191 | 32,506 |
| Has\_ActiveLoan | 0 | 0 | 4,788 | 4,788 |
| Has\_ActiveLoan | 1 | 0 | 23,403 | 23,403 |
| Has\_ActiveLoan | NA | 4,315 | 0 | 4,315 |
| Has\_ActiveLoan | TOTAL | 4,315 | 28,191 | 32,506 |

dataset %>%   
 select(c(Churn,NumberOfProducts,Has\_ActiveLoan)) %>%   
 mutate(Has\_ActiveLoan = ifelse(Has\_ActiveLoan == 0 | is.na(Has\_ActiveLoan),"No","Yes")) %>%   
 reshape2::melt('Churn') %>%   
 group\_by(Churn,variable,value) %>% summarise(Count = n()) %>% ungroup() %>%   
 # group\_by(variable,value) %>%   
 # mutate(Percent = paste0( round(Count/sum(Count),2)\*100,"%")) %>%   
 reshape2::melt(c("Churn","variable","value")) %>% clean\_names() %>%   
 reshape2::dcast(variable+value~churn,value.var = "value\_2") %>%   
 mutate\_at(c("Churn","Not Churn"),.funs = function(x)ifelse(is.na(x),0,x))%>%   
 split(.$variable) %>% map\_df(~adorn\_totals(.,"row")%>%   
 mutate(Total = (as.numeric(Churn) + as.numeric(`Not Churn`)),  
 Churn\_Rate = paste(round(Churn/Total,2)\*100,"%"))) %>%   
 flextable() %>% autofit() %>% merge\_v(j = 1,target = 'variable')

## `summarise()` has grouped output by 'Churn', 'variable'. You can override using  
## the `.groups` argument.

| variable | value | Churn | Not Churn | Total | Churn\_Rate |
| --- | --- | --- | --- | --- | --- |
| NumberOfProducts | 1 | 3,128 | 0 | 3,128 | 100 % |
| 2 | 1,152 | 27,388 | 28,540 | 4 % |
| 3 | 35 | 754 | 789 | 4 % |
| 4 | 0 | 44 | 44 | 0 % |
| 5 | 0 | 3 | 3 | 0 % |
| 7 | 0 | 1 | 1 | 0 % |
| 8 | 0 | 1 | 1 | 0 % |
| Total | - | 4,315 | 28,191 | 32,506 | 13 % |
| Has\_ActiveLoan | No | 4,315 | 4,788 | 9,103 | 47 % |
| Yes | 0 | 23,403 | 23,403 | 0 % |
| Total | - | 4,315 | 28,191 | 32,506 | 13 % |

**Observations and inferences**

* Majority of clients ## Build models

Let’s create a [**model specification**](https://www.tmwr.org/models.html) for each model we want to try:

To set up your modeling code, consider using the [parsnip addin](https://parsnip.tidymodels.org/reference/parsnip_addin.html) or the [usemodels](https://usemodels.tidymodels.org/) package.

Now let’s build a [**model workflow**](https://www.tmwr.org/workflows.html) combining each model specification with a data preprocessor:

If your feature engineering needs are more complex than provided by a formula like sex ~ ., use a [recipe](https://www.tidymodels.org/start/recipes/). [Read more about feature engineering with recipes](https://www.tmwr.org/recipes.html) to learn how they work.