Due Diligence Project

Joseph Annand

2024-04-14

## Libraries

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(caret)

## Warning: package 'caret' was built under R version 4.3.2

## Loading required package: lattice  
##   
## Attaching package: 'caret'  
##   
## The following object is masked from 'package:purrr':  
##   
## lift

library(writexl)

## Warning: package 'writexl' was built under R version 4.3.3

## Import Data

data <- read.csv("data/Customer\_Dataset\_Data.csv")  
data <- data %>% column\_to\_rownames(., var = "CustomerID")

# Drop irrelevant columns  
  
data <- data[, -c(7,18,19,20,21,23,24,25,26,27,28,29,30,31,32)]

Columns 1-5 are demographic-related variables. Columns 6-9 are occupation-related variables. Columns 10-13 and 15 are personal-finance-related variables. Column 14 is blank. Columns 16-22 are household-related variables. Columns 23-27 are car-related variables. Columns 28 and 29 are political variables. Columns 30-33 are credit-card-related variables. Columns 35-37, 39, 40, 43, 44, and 54 are quantitative service history variables. Columns 38, 41, 42, 45-53, and 55-59 are telecommunication-service-related variables.

## Recode 4 features: Convert monetary columns from character to numeric

data$VoiceOverTenure <- parse\_number(data$VoiceOverTenure)

## Warning: 3 parsing failures.  
## row col expected actual  
## 842 -- a number #NULL!  
## 2758 -- a number #NULL!  
## 3480 -- a number #NULL!

data$VoiceLastMonth <- parse\_number(data$VoiceLastMonth)  
  
data$EquipmentOverTenure <- parse\_number(data$EquipmentOverTenure)

## Warning: 3296 parsing failures.  
## row col expected actual  
## 3 -- a number $ -  
## 4 -- a number $ -  
## 5 -- a number $ -  
## 7 -- a number $ -  
## 8 -- a number $ -  
## ... ... ........ ......  
## See problems(...) for more details.

data$EquipmentLastMonth <- parse\_number(data$EquipmentLastMonth)

## Warning: 3296 parsing failures.  
## row col expected actual  
## 3 -- a number $ -  
## 4 -- a number $ -  
## 5 -- a number $ -  
## 7 -- a number $ -  
## 8 -- a number $ -  
## ... ... ........ ......  
## See problems(...) for more details.

data$DataOverTenure <- parse\_number(data$DataOverTenure)

## Warning: 3656 parsing failures.  
## row col expected actual  
## 1 -- a number $ -  
## 3 -- a number $ -  
## 4 -- a number $ -  
## 6 -- a number $ -  
## 7 -- a number $ -  
## ... ... ........ ......  
## See problems(...) for more details.

data$DataLastMonth <- parse\_number(data$DataLastMonth)

## Warning: 3656 parsing failures.  
## row col expected actual  
## 1 -- a number $ -  
## 3 -- a number $ -  
## 4 -- a number $ -  
## 6 -- a number $ -  
## 7 -- a number $ -  
## ... ... ........ ......  
## See problems(...) for more details.

data$VoiceOverTenure[is.na(data$VoiceOverTenure)] <- 0  
data$VoiceLastMonth[is.na(data$VoiceLastMonth)] <- 0  
  
data$EquipmentOverTenure[is.na(data$EquipmentOverTenure)] <- 0  
data$EquipmentLastMonth[is.na(data$EquipmentLastMonth)] <- 0  
  
data$DataOverTenure[is.na(data$DataOverTenure)] <- 0  
data$DataLastMonth[is.na(data$DataLastMonth)] <- 0

VoiceOverTenure, VoiceOverLastMonth, EquipmentOverTenure, EquipmentLastMonth, DataOverTenure, and DataLastMonth, HHIncome were all initially coded as character vectors with commas and dollar signs. All values in these columns were parsed for numbers and converted to numeric data type. NA values were converted to zero, so that the derived features could be calculated.

## Recode HHIncome feature

data$HHIncome <- parse\_number(data$HHIncome)

## Derived Feature 1: TotalOverTenure

data$TotalOverTenure <- data$VoiceOverTenure + data$EquipmentOverTenure + data$DataOverTenure

Determining a new feature that is the sum of the total amounts they have spent on phone, equipment, and data services over their tenure with the phone company allows for tracking of the most lucrative customers and find which variables are most related to how much customers spend in telecommunication services.

## Derived Feature 2: TotalByTenure

data$TotalByTenure <- data$TotalOverTenure / data$PhoneCoTenure

Expanding on the TotalOverTenure feature, the TotalByTenure feature calculates the average amount a customer spends on services per year with the phone company. With this, customers across the range of tenure with the phone company can be compared.

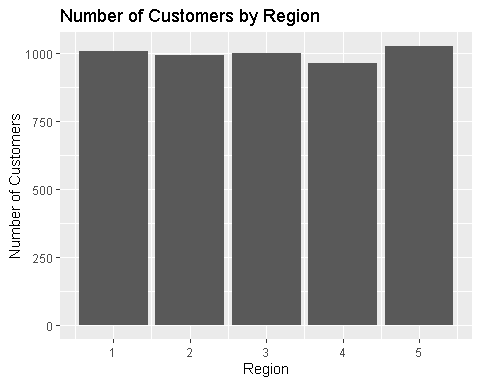
## Derived Feature 3: TotalLastMonth

data$TotalLastMonth <- data$VoiceLastMonth + data$EquipmentLastMonth + data$DataLastMonth

TotalLastMonth feature sums the amounts each customer paid in services in the past month. This data gives the most updated look at the total revenue each customer brings in.

## Breakdown of Customers by Region

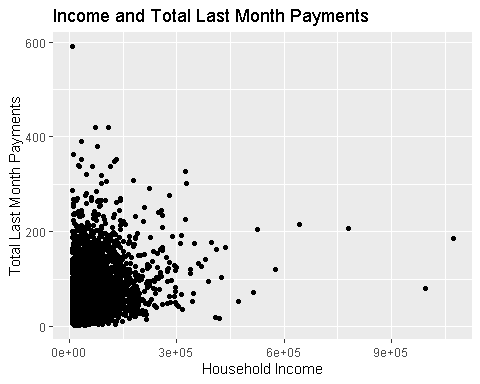
d <- ggplot(data = data, aes(Region)) +  
 geom\_bar() +  
 labs(x="Region", y="Number of Customers",  
 title = "Number of Customers by Region")  
d



The data from the Region feature was visualized using a bar chart to see the distribution of customers across the five regions. Customers are evenly distributed across the five regions.

## Relationship Between Income and Total Last Month

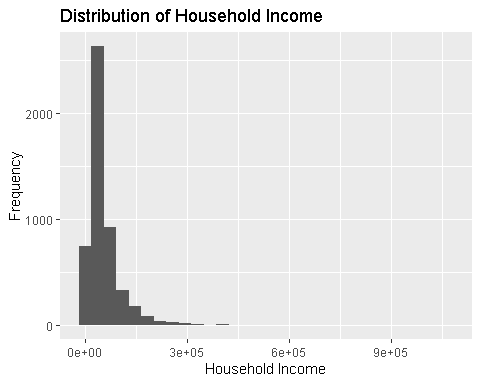
a <- ggplot(data = data, aes(HHIncome, TotalLastMonth)) +  
 geom\_point() +  
 labs(x="Household Income", y= "Total Last Month Payments",  
 title = "Income and Total Last Month Payments")  
a



## Normalize HHIncome

c <- ggplot(data = data, aes(HHIncome)) +  
 geom\_histogram() +  
 labs(x="Household Income", y= "Frequency",  
 title = "Distribution of Household Income")  
c

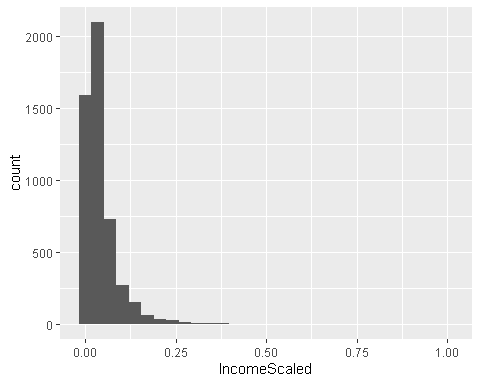
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



process <- preProcess(as.data.frame(data$HHIncome), method = c("range"))  
  
normIncome <- predict(process, as.data.frame(data$HHIncome))  
  
data$IncomeScaled <- normIncome$`data$HHIncome`

f <- ggplot(data = data, aes(IncomeScaled)) +  
 geom\_histogram()  
f

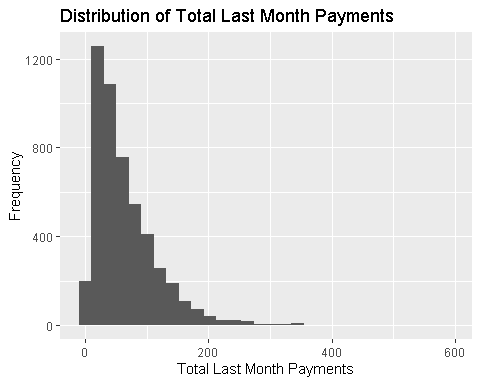
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## Normalize TotalLastMonth

e <- ggplot(data = data, aes(TotalLastMonth)) +  
 geom\_histogram() +  
 labs(x="Total Last Month Payments", y= "Frequency",  
 title = "Distribution of Total Last Month Payments")  
e

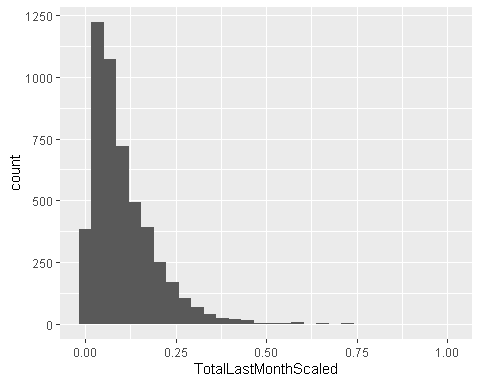
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



process <- preProcess(as.data.frame(data$TotalLastMonth), method = c("range"))  
  
normTLS<- predict(process, as.data.frame(data$TotalLastMonth))  
  
data$TotalLastMonthScaled <- normTLS$`data$TotalLastMonth`

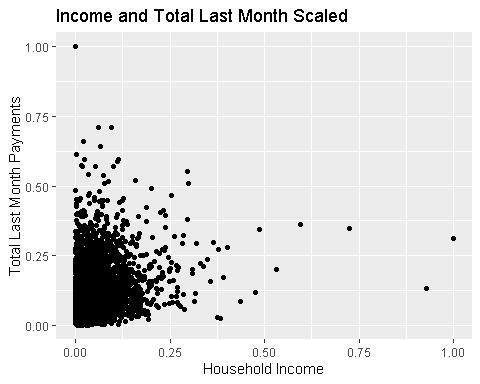
g = ggplot(data = data, aes(TotalLastMonthScaled)) +  
 geom\_histogram()  
g

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



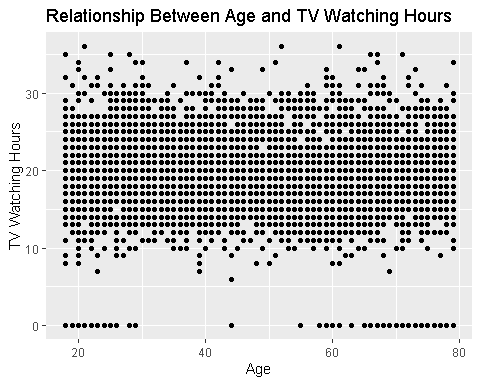
## Relationship Between Income and Total Last Month Normalized

h <- ggplot(data = data, aes(IncomeScaled, TotalLastMonthScaled)) +  
 geom\_point() +   
 labs(x="Household Income", y= "Total Last Month Payments",  
 title = "Income and Total Last Month Scaled")  
h



## Relationship between TV Watching Hours and Age

b <- ggplot(data = data, aes(Age, TVWatchingHours)) +  
 geom\_point() +  
 labs(x="Age", y= "TV Watching Hours",  
 title = "Relationship Between Age and TV Watching Hours")  
b



TV Watching Hours is plotted as a function of Age. There is no apparent correlation between the two variables.