Final Project All Stages - David Suffolk

David Suffolk

8 August, 2019

Part 1

Dataset 1

Introduction

Understanding customer churn is an important part of business development strategy. This is especially true in businesses that depend on recurring payments (ex: gym memberships and other subscription services). In order to recruit and retain customers, businesses need to build operations that prevent customers from ending the service and understanding why previous customers left can help them to do that. Data Science can step in and help to identify features of a customer profile or the business's reputation that lead to retention and churn of clients.

One industry to focus on in this regard is the newspaper subscription service. While many may theorize that subscriptions decrease as online accessibility to news resources has increased, this may not paint the whole picture as to why people who do have subscriptions to newspapers eventually cancel or decide to continue.

Research questions

Are there any common elements of a customer profile in newspaper subscribers that can predict churn?
Is there a correlation between the source of the customer referral and churn?
If a customer pays less for a subscription, are they more than likely to remain or churn?
Do different types of subscription plans (ex: daily, Sunday only) have a correlation with churn?
Do rewards programs have any impact on attracting or retaining customers?

Approach

By researching customer demographics, we may be able to identify which customer profiles are more likely to unsubscribe from a newspaper. We may also be able to identify reasons why a particular customer no longer wishes to continue a subscription (ex: online availability, change in family/income, etc.).

How your approach addresses (fully or partially) the problem.

If we can identify aspects of a customer profile, we can start to predict their likelihood (or length of time) that they will be a newspaper subscriber. The resulting information could lead to resources being allocated to reaching out to the most likely customers with the highest longevity as a subscriber.

Data

“Newspaper churn”

This dataset includes churn data for a newspaper based in California. There are 10,000 customer profiles. Variables include household income, home ownership, ethnicity, number of children, year of residence, age range, language, full address, weekly fee, delivery period, Nielsen Prizm identifier, reward program, source, and whether they are still a current subscriber.

Due to the amount of entries and the context of this being one newspaper, the data set provides the ability to do a specific approach for a sample data of the larger newspaper industry.

The description does not specify the newspaper that the dataset is in reference to which limits the ability to identify characteristics of the newspaper itself that may be impacting churn.

Link to dataset: https://www.kaggle.com/leiyiting01/newspaper-churn

Required Packages

readxl
ggplot2
pROC
caTools
caret
Class
e1071
mlogit

Plots and Table Needs

Histogram
Probability Plot
AUC

Questions for future steps.

Are there any correlations between profile characteristics and subscription status that need to be investigated further?
What is the strength of each variable in predicting commitment of customer?

Dataset 2

Introduction

Understanding customer churn is an important part of business development strategy. This is especially true in businesses that depend on recurring payments (ex: gym memberships and other subscription services). In order to recruit and retain customers, businesses need to build operations that prevent customers from ending the service and understanding why previous customers left can help them to do that. Data Science can step in and help to identify features of a customer profile or the business's reputation that lead to retention and churn of clients.

The telecom industry is one such business format. Their products are often viewed as necessary utilities with few other options for clients. Additionally, it is not just one product being sold but several options and packages. This means that churn becomes much more than a straightforward concern of losing a customer but also a customer removing one of many services.

Research questions

What types of customers pay the most for services (sign up for the most packages)?
Are there elements of a customer profile that can predict churn?
What is the average lifespan (months/years using service) of a customer before they churn?
Is there a particular service that a customer has that is most closely associated with churn?
Is there a particular service that is least associated with churn? In other words, is there a service that customers use that may predict whether they are less likely to churn?

Approach

By researching the various services and how they are associated with customer profiles, we may be able to identify patterns of customers that discontinue the telecom service.

How your approach addresses (fully or partially) the problem.

If customer profiles and products can help predict churn, the business can understand which products have the best strength with retaining clients and what customer profiles would benefit more from marketing and retention resources.

Data

“Telco Customer Churn”

The data is an IBM Sample Dataset. It includes over 7,000 customers with demographic details and the services that they had purchased.

The variables include gender, senior citizen, partnered status, dependents, tenure, services used, contract type, paperless billing, payment method, monthly charges, total charges and churn status.

The source of the data is a little unclear. We do not know the month and year when this data was collected, the reasons for discontinuing service, or the location.

Link to dataset: https://www.kaggle.com/blastchar/telco-customer-churn

Required Packages

csv
ggplot2
pROC
caTools
caret
Class
e1071
mlogit

Plots and Table Needs

Histogram
Probability Plot
ROC

Questions for future steps

What are the most common combinations of services used by customers?
What is the average tenure of a customer?
Are there any customer demographics that correlate with short tenure and churn?

Dataset 3

Introduction

Understanding customer churn is an important part of business development strategy. This is especially true in businesses that depend on recurring payments (ex: gym memberships and other subscription services). In order to recruit and retain customers, businesses need to build operations that prevent customers from ending the service and understanding why previous customers left can help them to do that. Data Science can step in and help to identify features of a customer profile or the business's reputation that lead to retention and churn of clients.

While the banking business model does not rely solely on monthly payments but also on account balances. Banks also have insights into different demographics of a customer that can help build a customer profile (ex: salary, credit score, etc.). If churn can be predicted, banks can get ahead of customers that are flight risks to try and save the relationship.

Research questions

What elements of a customer profile can help predict churn?
What services can help to predict customer churn?
What is the average tenure of a customer that discontinues service?
Are there signals in account balance that can help predict customer churn?
What services are associated with customers who do continue with banking services?

Approach

By researching the various services and variables in a customer profile, we may be able to identify patterns of customers that discontinue the banking services.

How your approach addresses (fully or partially) the problem.

If customer profiles and products can help predict churn, the business can understand which products have the best strength with retaining clients and what customer profiles would benefit more from marketing and retention resources.

Data

"Bank Customer Churn Modeling"

The dataset is part of a project to predict customer churn in the context of banking. There are 10,000 customer profiles in the dataset.

The variables of the dataset include credit score, geography, gender, age, tenure, balance, number of products, credit card possession, active member status, salary, and if the customer exited the bank.

The dataset is vague about its origins and looks to be designed for projects that challenge data science students to build models that predict customer churn.

Link to dataset: https://www.kaggle.com/barelydedicated/bank-customer-churn-modeling

Required Packages

csv
ggplot2
pROC
caTools
caret
Class
e1071
mlogit

Plots and Table Needs

Histogram
Probability Plot
ROC

Questions for future steps

What are the most common attributes of churned customers?
What is the average number of products per customer?
Are there any customer demographics that correlate with short tenure and churn?

Part 2

Import Data File

I will be using the read_excel function from the readxl library to import the data from the excel spreadsheet source. For column names, I will replace all spaces with an underscore.

```
ChurnData <- read_excel("NewspaperChurn.xlsx")
names(ChurnData) <- gsub(" ","_",names(ChurnData))
summary(ChurnData)
```

```
##  SubscriptionID       HH_Income         Home_Ownership
##  Min.   :110001064   Length:15855       Length:15855
##  1st Qu.:150182445   Class :character   Class :character
##  Median :180333300   Mode  :character   Mode  :character
##  Mean   :164372917
##  3rd Qu.:180636209
##  Max.   :181554089
##   Ethnicity         dummy_for_Children Year_Of_Residence
##  Length:15855       Length:15855       Min.   : 1.00
##  Class :character   Class :character   1st Qu.: 4.00
##  Mode  :character   Mode  :character   Median :10.00
##                                        Mean   :13.55
##                                        3rd Qu.:21.00
##                                        Max.   :56.00
##   Age_range           Language           Address
##  Length:15855       Length:15855       Length:15855
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##     State               City               County
Zip_Code
##  Length:15855       Length:15855       Length:15855       Min.
:90603
##  Class :character   Class :character   Class :character   1st
Qu.:92627
##  Mode  :character   Mode  :character   Mode  :character   Median
:92688
##                                                           Mean
:92425
##                                                           3rd
Qu.:92806
##                                                           Max.
:92887
```

```
##   weekly_fee         Deliveryperiod     Nielsen_Prizm
##  Length:15855       Length:15855       Length:15855
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  reward_program    Source_Channel      Subscriber
##  Min.   :  0.000   Length:15855       Length:15855
##  1st Qu.:  0.000   Class :character   Class :character
##  Median :  0.000   Mode  :character   Mode  :character
##  Mean   :  1.101
##  3rd Qu.:  0.000
##  Max.   :353.000
```

For this analysis, I will be using the following variables. My biggest obstacle is the format of the values are in string format so my cleaning involved converting them to numeric representations. I will keep columns of either values in my final data frame.

Column Values

Household Income

18 levels of variables and all characters. Convert values to a numeric representation.

Home Ownership

2 levels (“RENTER” or “OWNER”). Converted to numeric values.

Children

2 levels (“Y” or “N”). Converted to numeric values.

Year of Residence

Numeric values. These are imported as characters and will need to be converted to numeric.

Age Range

12 levels and 108 blank entries. Converted to numeric values

Zip Code

Over 100 different string variables with no blank values. No changes needed.

Weekly Fee

14 levels with 186 blank entries. Converted to numeric values.

Subscriber

Binary values of “YES” or “NO” with no blank entries. This is the value I will be trying to predict and I have converted them to numeric values.

Here are the various variables mentioned above and how I have adapted them for analysis.

Convert Subscriber values to numbers.
Subscribers = 1
Non-subscribers = 2

```r
ChurnData$Subscriber_Number <- as.character(ChurnData$Subscriber)
ChurnData$Subscriber_Number[ChurnData$Subscriber_Number == "YES"] <-
"1"
ChurnData$Subscriber_Number[ChurnData$Subscriber_Number == "NO"] <-
"2"
ChurnData$Subscriber_Number <- as.numeric(ChurnData$Subscriber_Number)
ChurnData$Subscriber_Number[1:20]
```

```
##  [1] 2 1 1 2 1 2 2 2 2 1 2 2 2 2 2 2 2 2 2 1
```

Convert Home Ownership values to numbers.
Renter = 1
Owner = 2

```r
ChurnData$Home_Ownership_Number <-
as.character(ChurnData$Home_Ownership)
ChurnData$Home_Ownership_Number[ChurnData$Home_Ownership_Number ==
"RENTER"] <- "1"
ChurnData$Home_Ownership_Number[ChurnData$Home_Ownership_Number ==
"OWNER"] <- "2"
ChurnData$Home_Ownership_Number <-
as.numeric(ChurnData$Home_Ownership_Number)
ChurnData$Home_Ownership_Number[1:20]
```

```
##  [1] 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2
```

Convert Income to numeric values

I converted the lower number of the bucket to a number
For less than $20,000, I assigned the value of 1

```
ChurnData$HH_Income_Number <- as.character(ChurnData$HH_Income)
ChurnData$HH_Income_Number[ChurnData$HH_Income_Number == "$  30,000 - $39,999"] <- "30"
ChurnData$HH_Income_Number[ChurnData$HH_Income_Number == "$500,000 Plus"] <- "500"
ChurnData$HH_Income_Number[ChurnData$HH_Income_Number == "$100,000 - $124,999"] <- "100"
ChurnData$HH_Income_Number[ChurnData$HH_Income_Number == "$200,000 - $249,999"] <- "200"
ChurnData$HH_Income_Number[ChurnData$HH_Income_Number == "$  50,000 - $59,999"] <- "50"
ChurnData$HH_Income_Number[ChurnData$HH_Income_Number == "$150,000 - $174,999"] <- "150"
ChurnData$HH_Income_Number[ChurnData$HH_Income_Number == "$400,000 - $499,999"] <- "400"
ChurnData$HH_Income_Number[ChurnData$HH_Income_Number == "$175,000 - $199,999"] <- "175"
ChurnData$HH_Income_Number[ChurnData$HH_Income_Number == "$125,000 - $149,999"] <- "125"
ChurnData$HH_Income_Number[ChurnData$HH_Income_Number == "Under $20,000"] <- "1"
ChurnData$HH_Income_Number[ChurnData$HH_Income_Number == "$  80,000 - $89,999"] <- "80"
ChurnData$HH_Income_Number[ChurnData$HH_Income_Number == "$  90,000 - $99,999"] <- "90"
ChurnData$HH_Income_Number[ChurnData$HH_Income_Number == "$300,000 - $399,999"] <- "300"
ChurnData$HH_Income_Number[ChurnData$HH_Income_Number == "$  20,000 - $29,999"] <- "20"
ChurnData$HH_Income_Number[ChurnData$HH_Income_Number == "$  70,000 - $79,999"] <- "70"
ChurnData$HH_Income_Number[ChurnData$HH_Income_Number == "$  60,000 - $69,999"] <- "60"
ChurnData$HH_Income_Number[ChurnData$HH_Income_Number == "$  40,000 -
$49,999"] <- "40"
ChurnData$HH_Income_Number[ChurnData$HH_Income_Number == "$250,000 -
$299,999"] <- "250"
ChurnData$HH_Income_Number <- as.numeric(ChurnData$HH_Income_Number)
ChurnData$HH_Income_Number[1:20]

##  [1]  30 500 100 200  50 500 150 400 175  50 125 500 125 400 150
200   1
## [18]  80 400 200
```

Convert Dummy for Children

Y -> 1
N -> 2

```
ChurnData$dummy_for_Children_Number <-
as.character(ChurnData$dummy_for_Children)
ChurnData$dummy_for_Children_Number[ChurnData$dummy_for_Children_Numbe
r == "Y"] <- "1"
ChurnData$dummy_for_Children_Number[ChurnData$dummy_for_Children_Numbe
r == "N"] <- "2"
ChurnData$dummy_for_Children_Number <-
as.numeric(ChurnData$dummy_for_Children_Number)
ChurnData$dummy_for_Children_Number[1:20]

##  [1] 2 1 1 2 2 1 2 1 2 2 2 2 2 1 2 2 2 2 1 2
```

Verify Year of Residence is a numeric value

```
ChurnData$Year_Of_Residence <- as.numeric(ChurnData$Year_Of_Residence)
ChurnData$Year_Of_Residence[1:20]

##  [1]  1 14  7 23 23 10  4 22  2  6  7 25 23 16 11  1  7  7 12 26
```

Convert Age Range

I converted the lower number of the bucket to a number
For <24, I changed it to 18

```
ChurnData$Age_range_Number <- as.character(ChurnData$Age_range)
ChurnData$Age_range_Number[ChurnData$Age_range_Number == "25-29"] <-
"25"
ChurnData$Age_range_Number[ChurnData$Age_range_Number == "50-54"] <-
"50"
```

```
ChurnData$Age_range_Number[ChurnData$Age_range_Number == "45-49"] <-
"45"
ChurnData$Age_range_Number[ChurnData$Age_range_Number == "55-59"] <-
"55"
ChurnData$Age_range_Number[ChurnData$Age_range_Number == "60-64"] <-
"60"
ChurnData$Age_range_Number[ChurnData$Age_range_Number == "75 years or
more"] <- "75"
ChurnData$Age_range_Number[ChurnData$Age_range_Number == "65-69"] <-
"65"
ChurnData$Age_range_Number[ChurnData$Age_range_Number == "70-74"] <-
"70"
ChurnData$Age_range_Number[ChurnData$Age_range_Number == "40-44"] <-
"40"
ChurnData$Age_range_Number[ChurnData$Age_range_Number == "24 years or
less"] <- "18"
ChurnData$Age_range_Number[ChurnData$Age_range_Number == "35-39"] <-
"35"
ChurnData$Age_range_Number[ChurnData$Age_range_Number == "30-34"] <-
"30"
ChurnData$Age_range_Number <- as.numeric(ChurnData$Age_range_Number)
ChurnData$Age_range_Number[1:20]
```

```
##  [1] 25 50 45 55 60 45 75 45 45 65 50 55 70 50 65 25 25 60 40 70
```

Weekly Fee

For most levels, I converted the lower number of the bucket to a number
For ‘$0’ and ‘$0-0.01’, I made them both 0

```
ChurnData$weekly_fee_number <- as.character(ChurnData$weekly_fee)
ChurnData$weekly_fee_number[ChurnData$weekly_fee_number == "$7.00 -
$7.99"] <- "7"
ChurnData$weekly_fee_number[ChurnData$weekly_fee_number == "$0.01 -
$0.50"] <- "0.01"
ChurnData$weekly_fee_number[ChurnData$weekly_fee_number == "$1.00 -
$1.99"] <- "1"
ChurnData$weekly_fee_number[ChurnData$weekly_fee_number == "$8.00 -
$8.99"] <- "8"
ChurnData$weekly_fee_number[ChurnData$weekly_fee_number == "$0 -
$0.01"] <- "0"
ChurnData$weekly_fee_number[ChurnData$weekly_fee_number == "$2.00 -
$2.99"] <- "2"
ChurnData$weekly_fee_number[ChurnData$weekly_fee_number == "$9.00 -
$9.99"] <- "9"
ChurnData$weekly_fee_number[ChurnData$weekly_fee_number == "$4.00 -
$4.99"] <- "4"
ChurnData$weekly_fee_number[ChurnData$weekly_fee_number == "$0.51 -
$0.99"] <- "0.51"
ChurnData$weekly_fee_number[ChurnData$weekly_fee_number == "$3.00 -
$3.99"] <- "3"
ChurnData$weekly_fee_number[ChurnData$weekly_fee_number == "$5.00 -
$5.99"] <- "5"
ChurnData$weekly_fee_number[ChurnData$weekly_fee_number == "$6.00 -
$6.99"] <- "6"
ChurnData$weekly_fee_number[ChurnData$weekly_fee_number == "$10.00 -
$10.99"] <- "10"
ChurnData$weekly_fee_number[ChurnData$weekly_fee_number == "$0"] <-
"0"
ChurnData$weekly_fee_number <- as.numeric(ChurnData$weekly_fee_number)
ChurnData$weekly_fee_number[1:20]
```

```
##  [1] 7.00 0.01 0.01 1.00 8.00 0.00 2.00 0.01 2.00 1.00 2.00 9.00
9.00 2.00
## [15] 4.00 1.00 2.00 0.51 0.51 0.01
```

Create New Dataframe

I have then taken the columns I need and converted them to a modified dataframe.
I then created variable ChurnData2 that is all data with NA values removed.

```
ChurnDataModified <- ChurnData[, c("Subscriber", "Subscriber_Number",
"Home_Ownership_Number", "HH_Income_Number",
"dummy_for_Children_Number", "Year_Of_Residence", "Age_range_Number",
"weekly_fee_number", "Zip_Code")]
ChurnData_2 <- na.omit(ChurnDataModified)
```

Final Dataset

Here is the structure of my clean dataset.

```
str(ChurnData_2)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    15561 obs. of  9
variables:
##  $ Subscriber               : chr  "NO" "YES" "YES" "NO" ...
##  $ Subscriber_Number        : num  2 1 1 2 1 2 2 2 2 1 ...
##  $ Home_Ownership_Number    : num  1 2 2 2 2 2 2 2 2 2 ...
##  $ HH_Income_Number         : num  30 500 100 200 50 500 150 400
175 50 ...
##  $ dummy_for_Children_Number: num  2 1 1 2 2 1 2 1 2 2 ...
##  $ Year_Of_Residence        : num  1 14 7 23 23 10 4 22 2 6 ...
##  $ Age_range_Number         : num  25 50 45 55 60 45 75 45 45
65 ...
##  $ weekly_fee_number        : num  7 0.01 0.01 1 8 0 2 0.01 2 1 ...
##  $ Zip_Code                 : num  90802 92657 92604 92677
92688 ...
##  - attr(*, "na.action")= 'omit' Named int  31 127 224 294 317 322
386 392 403 481 ...
##   ..- attr(*, "names")= chr  "31" "127" "224" "294" ...
```

What information is not self-evident?

It is not yet clear if any of the independent variables are closely correlated with the dependent variable (subscription status). It is also not clear of any of the independent variables are correlated with the another. For example, does household income correlate with the amount paid for a subscription?

What are different ways you could look at this data?

Ultimately, I want to be able to identify what types of customers are more likely to keep the newspaper subscription. I think the best way to do that is to build a customer profile of multiple variables and see what types of customers are most likely to keep their subscription.

How do you plan to slice and dice the data?

I removed all residence variables except for zipcode. For this project, it seemed cumbersome to be dealing with so many different levels for one particular datapoint that may reveal something about the customer profile.

How could you summarize your data to answer key questions?

Once I am able to identify the variables that have the strongest prediction of churn, I will be able to summarize an easy-to-understand customer type that is most likley to remain a customer and who will not. Furthermore, a summary may be able to predict where a newspaper may be most successful (partcular subscription payments, for example).

What types of plots and tables will help you to illustrate the findings to your questions?

A scatterplot will be the best visual to identify trends and correlations amongst the variables.

When reviewing the independent variables, I will use histograms to understand the distribution of data.

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

Yes. I want to be able to predict the subscription status of the customer. I will create training and test datasets to do this. At this stage, I am not sure if AUC or KNN will be the best method for the analysis but I will investigate this in the next stage of the project.

Part 3

Introduction

Predicting customer churn is important to any subscription business. It makes economic sense to these businesses to invest in keeping the customers that continue to pay for the services. While this encompasses many different types of industries, the dataset I worked with was specific to a California-based newspaper. I set out to explore this dataset to see what characteristics could be found about customers who ended their newspaper subscription.

The problem statement you addressed

Overall, the problem to be solved is predicting customer churn. I set out to investigate if there were any characteristics that would help to predict this so that the customer could be saved. I realized that this could be as simple as one characteristic or a compilation of different characteristics.

How you addressed this problem statement

I focused on 7 different variables to address this problem:
1. Household Income
2. Home Ownership
3. Children in Household
4. Years of Residence in Community
5. Age Range
6. ZipCode
7. Weekly Fee

Additionally, I had one predicting variable which is subscription status.

Analysis

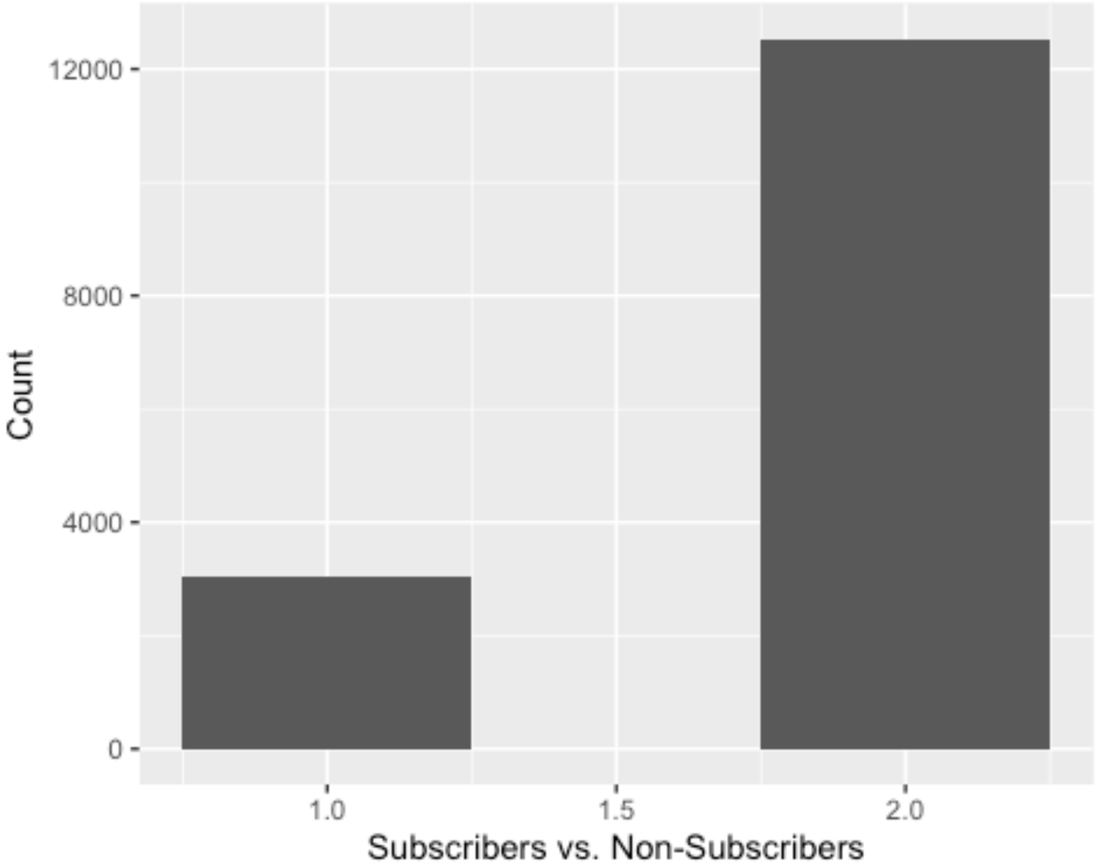
Understanding the variables

Subscribers

The histogram shows that the data contains more information on customers who have ended their subscription than continued with it. This makes sense when understanding the context of the data source which is trying to understand customer churn (we need to understand the customers that ended their subscription).

Subscribers = 1
Non-subscribers = 2

```
ggplot(data = ChurnData_2, aes(Subscriber_Number)) +
  geom_histogram(bins = 2, binwidth = 0.5)+
  xlab("Subscribers vs. Non-Subscribers") +
  ylab("Count")
```

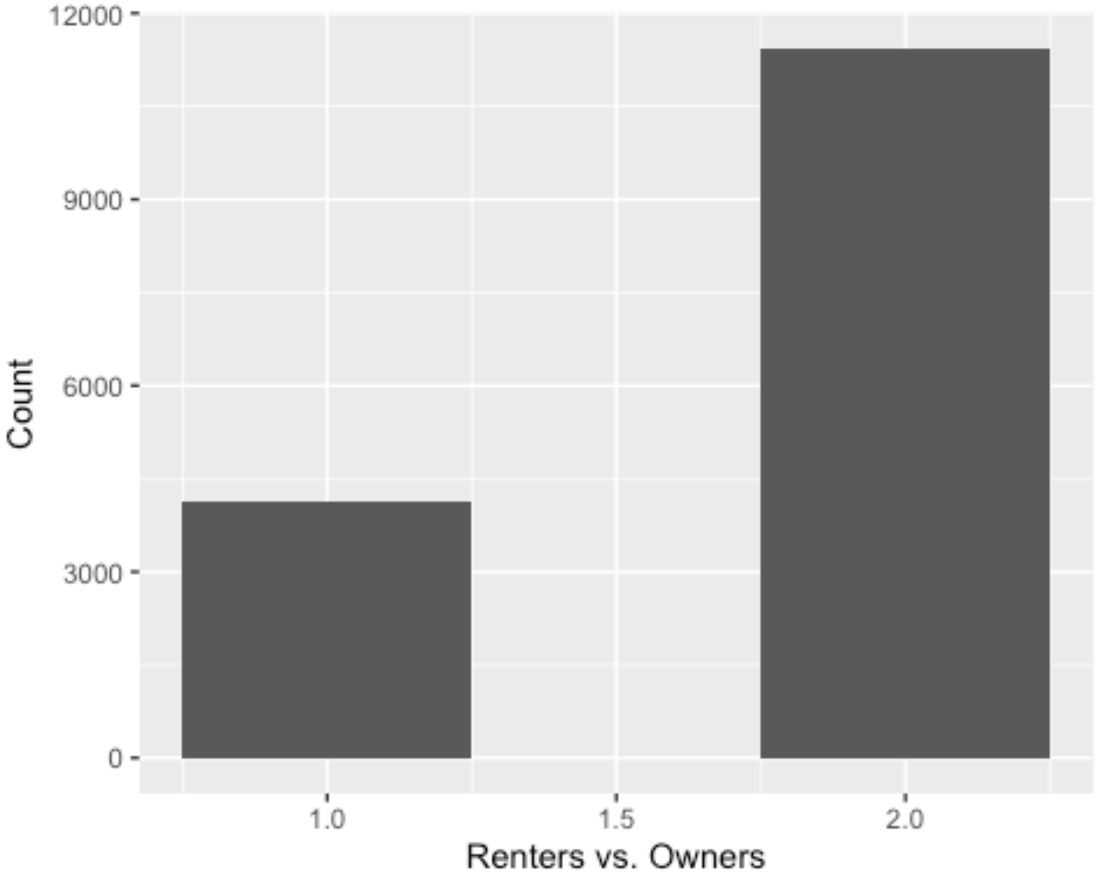


Home Ownership

Renters who unsubscribed had a lower mean of years of residence. The histogram also shows that the data includes far more home owners than renters.

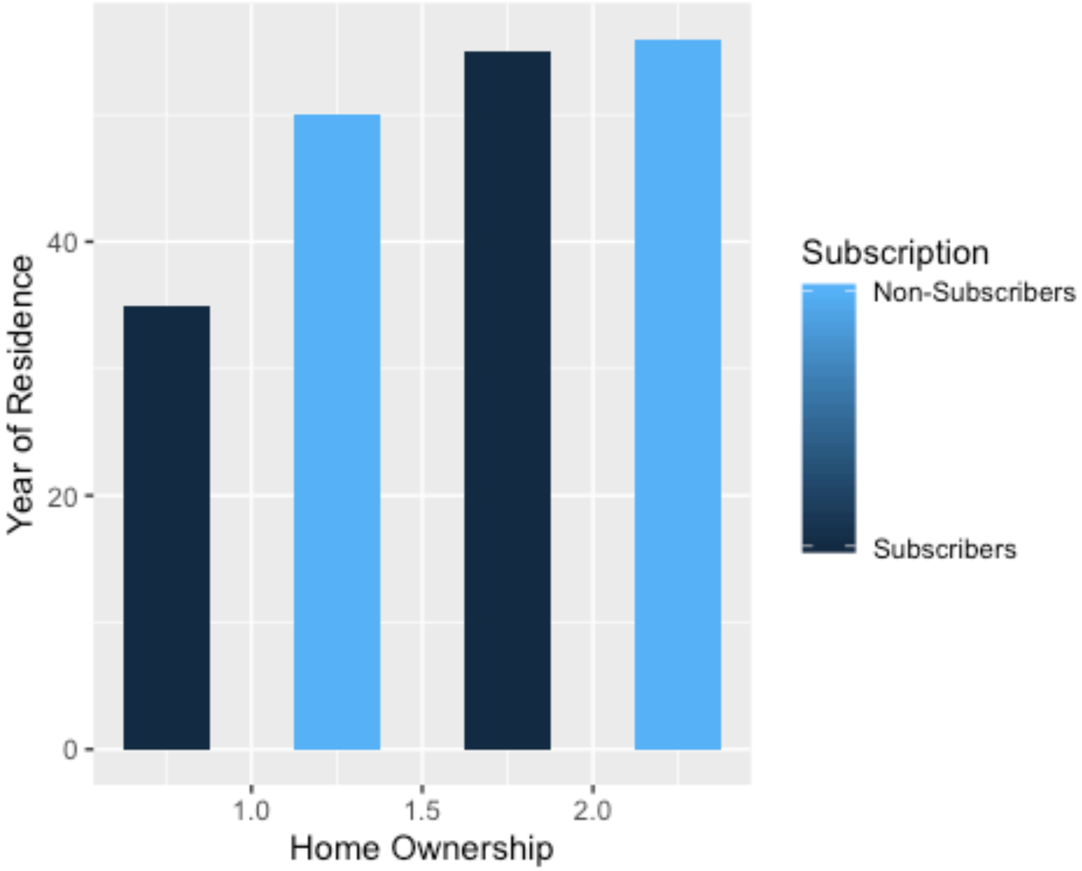
Renter = 1
Owner = 2

```
ggplot(data = ChurnData_2, aes(Home_Ownership_Number)) +
  geom_histogram(bins = 2, binwidth = 0.5)+
  xlab("Renters vs. Owners") +
  ylab("Count")
```



Home Ownership -> Year of Residence

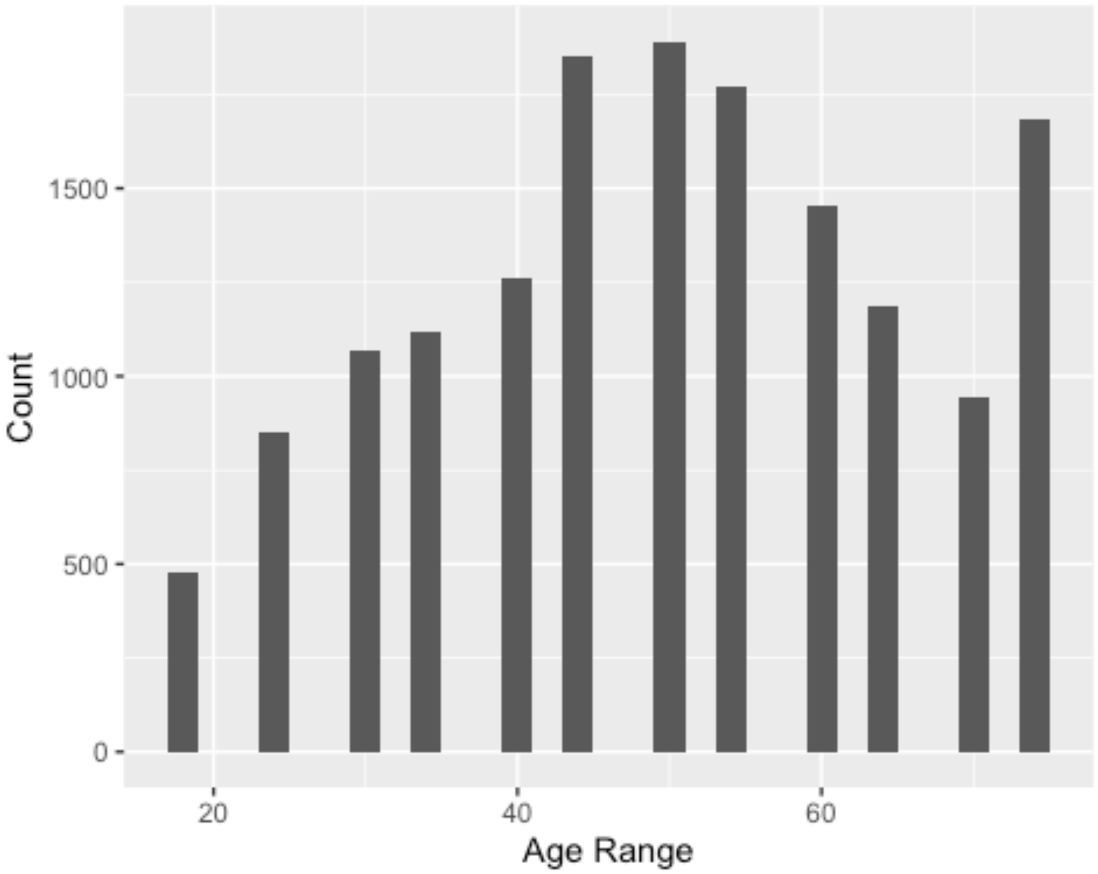
```
bar <- ggplot(data = ChurnData_2, aes(x = Home_Ownership_Number, y =
Year_Of_Residence, group = Subscriber_Number, fill =
Subscriber_Number))
bar + geom_bar(stat = "identity", position = position_dodge(1), width
= 0.5) + scale_fill_continuous(name="Subscription",
                      breaks=c(1, 2),
                      labels=c("Subscribers", "Non-Subscribers"))+
  xlab("Home Ownership")+
  ylab("Year of Residence")
```



Age Range

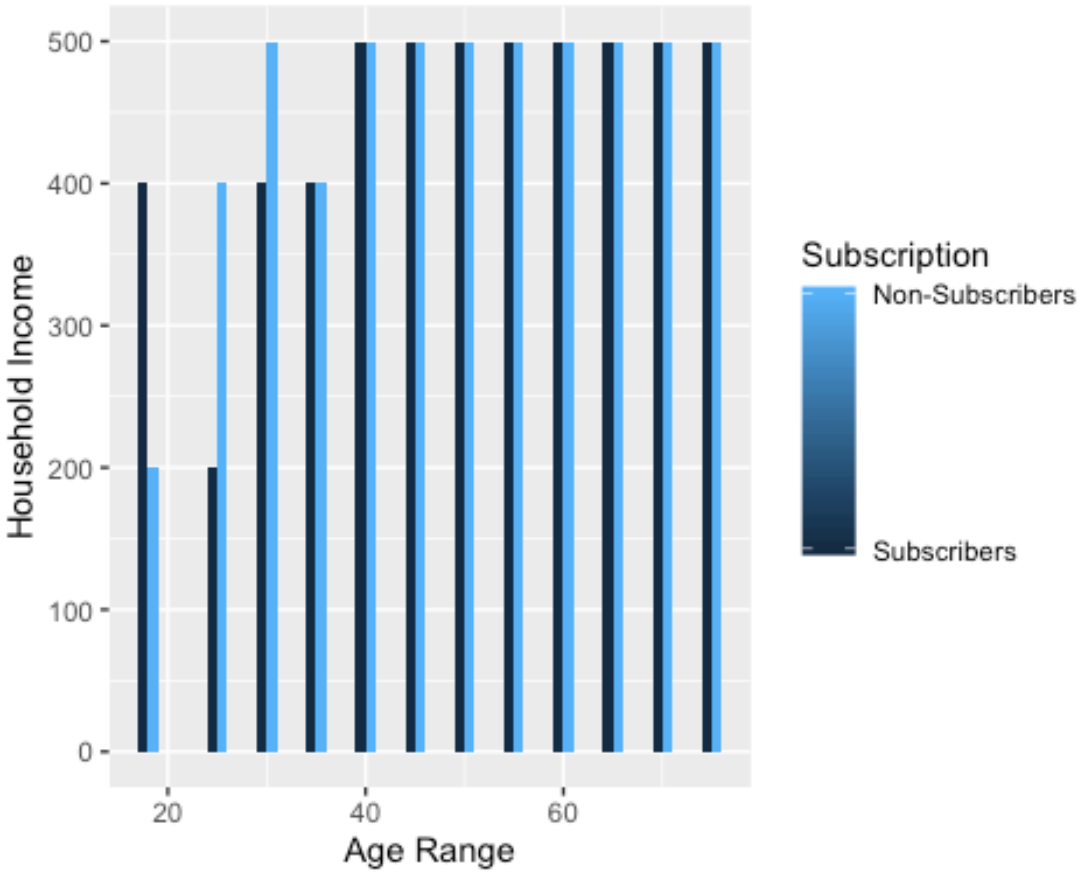
Over 40, household income and subscription status are not distinguishing factors. However, between 20-40, there appears to be a correlation between customers continuing to subscribe having a lower household income. There is also a potential outlier here as subscribers under 20 but with a $400K income are subscribers. Furthermore, when we look at the bar chart for age range against year of residence, we start to see inconsistencies. There are subscribers in the 20-30 age group who have more than 40 years of residency. The histogram shows a normal distribution between 18 and 70. However, there is a spike in data collection for those over 75.

```r
ggplot(data = ChurnData_2, aes(Age_range_Number)) +
  geom_histogram(bins = 6, binwidth = 2)+
  xlab("Age Range") +
  ylab("Count")
```



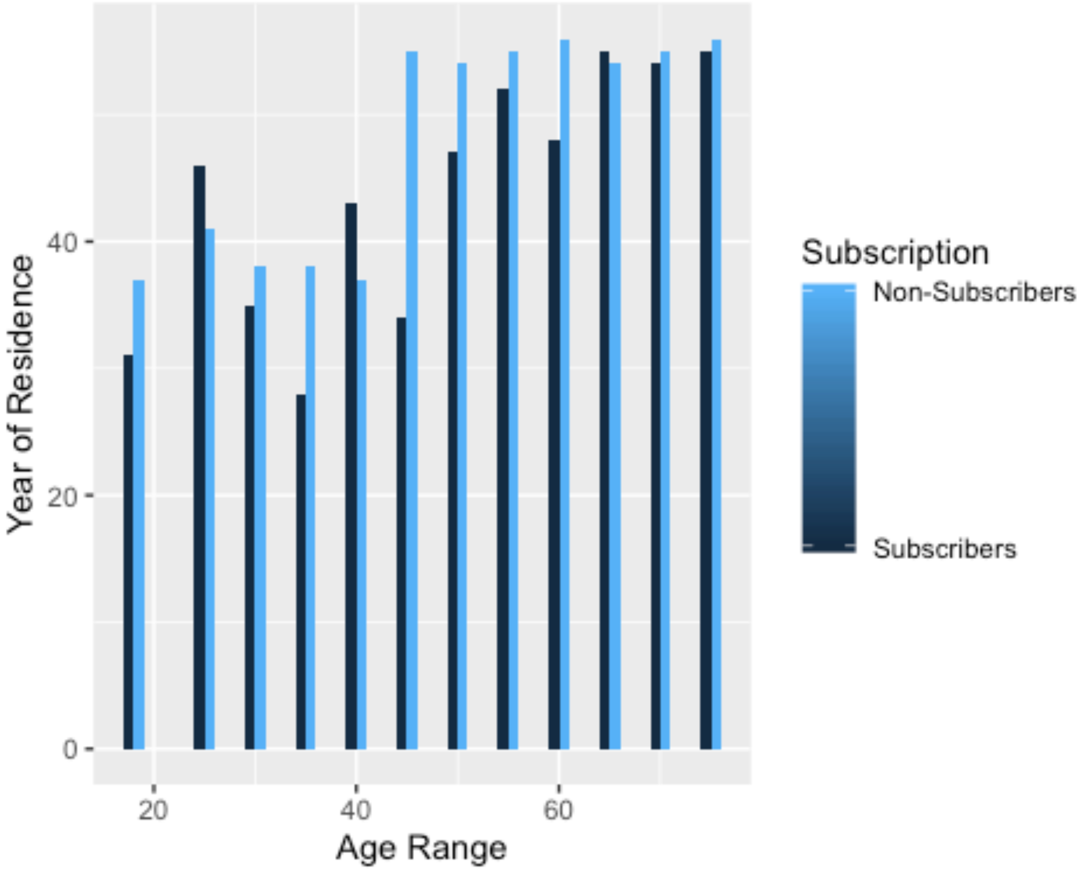
Age Range -> Household Income

```
bar <- ggplot(data = ChurnData_2, aes(x = Age_range_Number, y =
HH_Income_Number, group = Subscriber_Number, fill =
Subscriber_Number))
bar + geom_bar(stat = "identity", position = position_dodge(2), width
= 2) + scale_fill_continuous(name="Subscription",
                      breaks=c(1, 2),
                      labels=c("Subscribers", "Non-Subscribers"))+
  xlab("Age Range")+
  ylab("Household Income")
```



Age Range -> Year of Residence

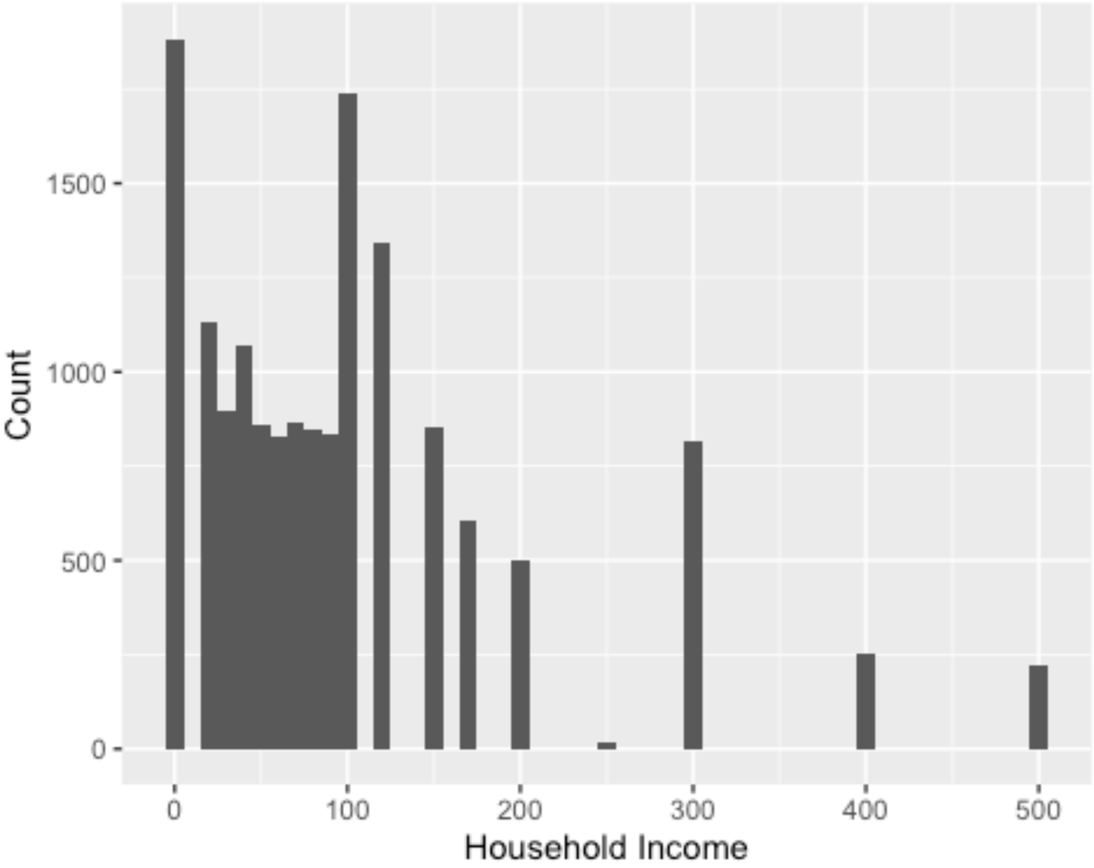
```
bar <- ggplot(data = ChurnData_2, aes(x = Age_range_Number, y =
Year_Of_Residence, group = Subscriber_Number, fill =
Subscriber_Number))
bar + geom_bar(stat = "identity", position = position_dodge(2), width
= 2) + scale_fill_continuous(name="Subscription",
                      breaks=c(1, 2),
                      labels=c("Subscribers", "Non-Subscribers"))+
  xlab("Age Range")+
  ylab("Year of Residence")
```



Household Income

The histogram shows that most of the customers in the data make under $100K.

```
ggplot(data = ChurnData_2, aes(HH_Income_Number)) +
  geom_histogram(bins = 17, binwidth = 10)+
  xlab("Household Income") +
  ylab("Count")
```

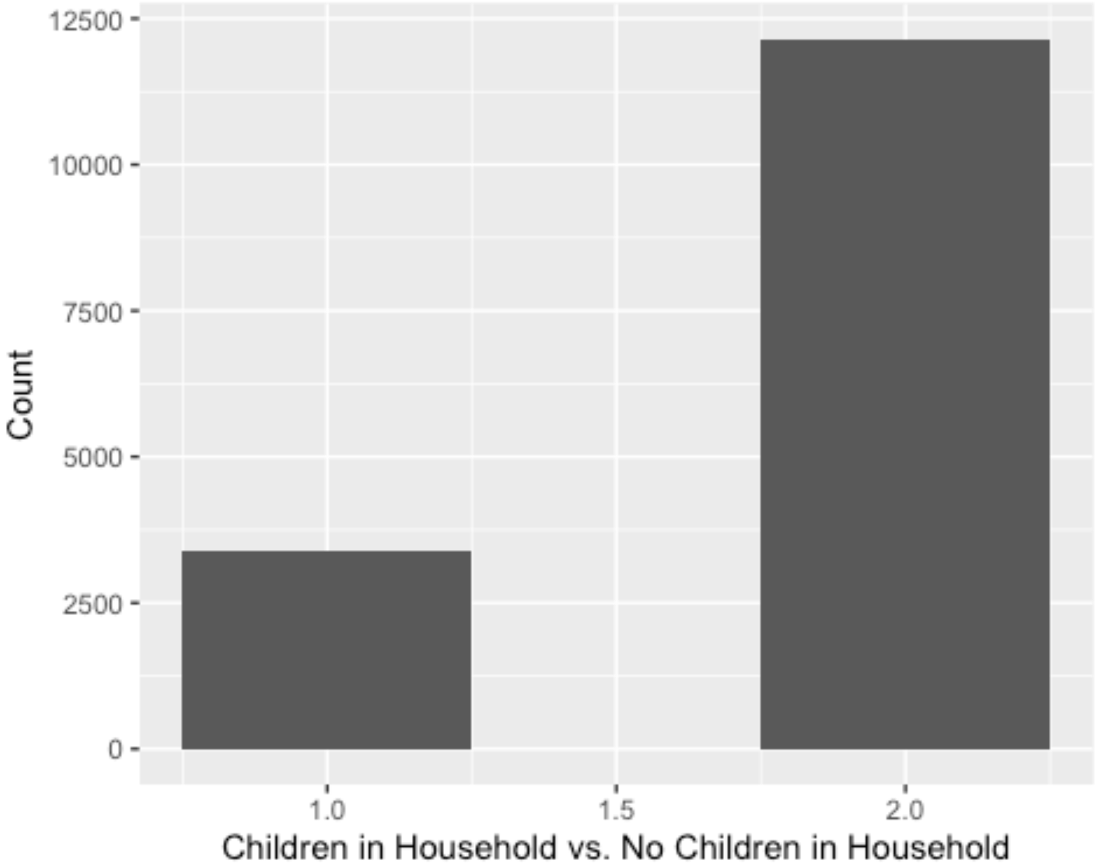


Children in Household

Most of the customers in the data have no children in the household.

1 = Children in household
2 = No children in household

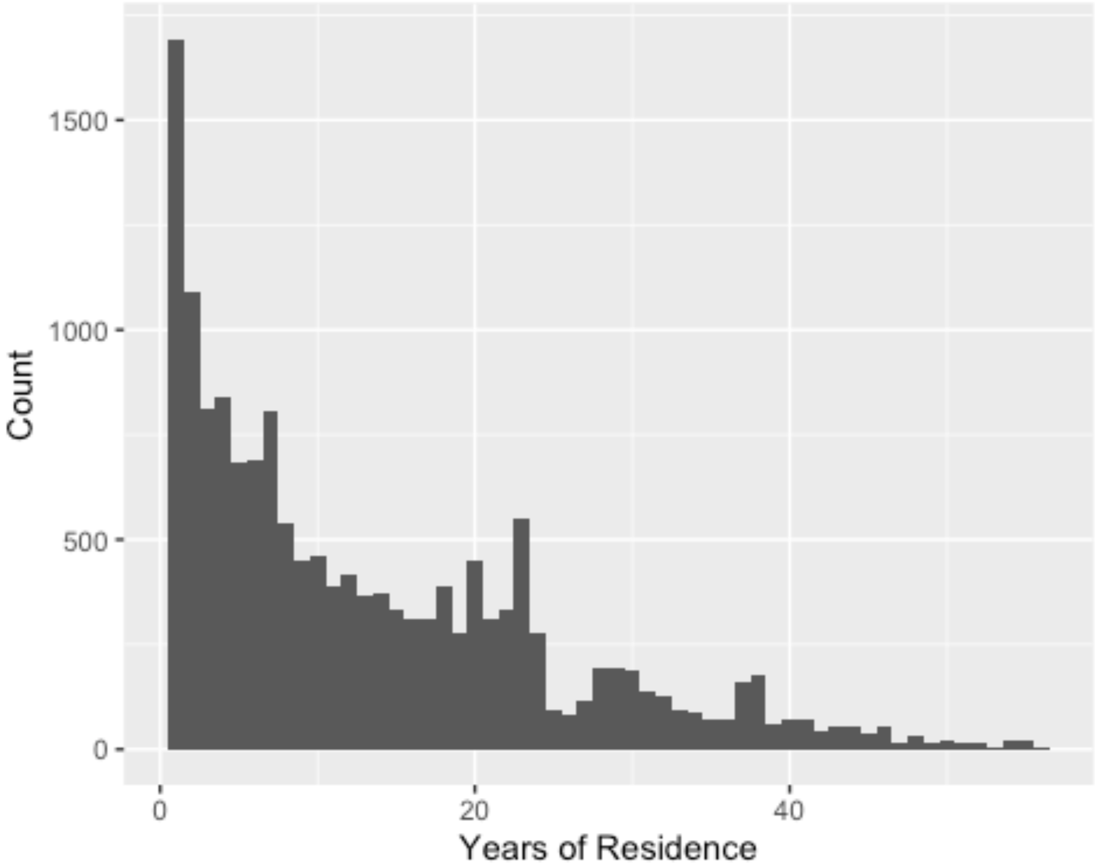
```
ggplot(data = ChurnData_2, aes(dummy_for_Children_Number)) +
  geom_histogram(bins = 2, binwidth = 0.5)+
  xlab("Children in Household vs. No Children in Household") +
  ylab("Count")
```



Years of Residence

The histogram shows that most customers in the data have been residents of the area for less than two years. It makes sense that the number of residents decreases as the number of years increases as this would likely be true for most residential areas.

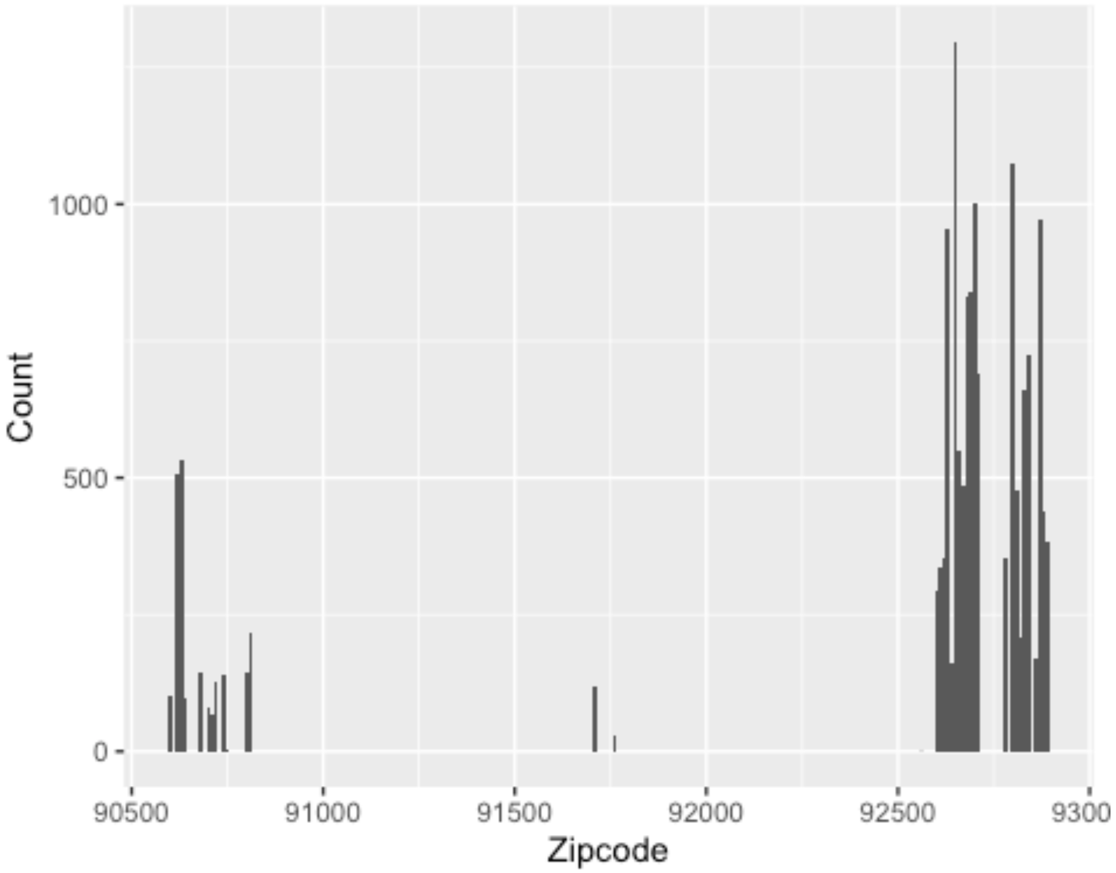
```
ggplot(data = ChurnData_2, aes(Year_Of_Residence)) +
  geom_histogram(bins = 56)+
  xlab("Years of Residence") +
  ylab("Count")
```



Zipcode

The histogram shows that most subscribers fall into a zipcode beginning with 925

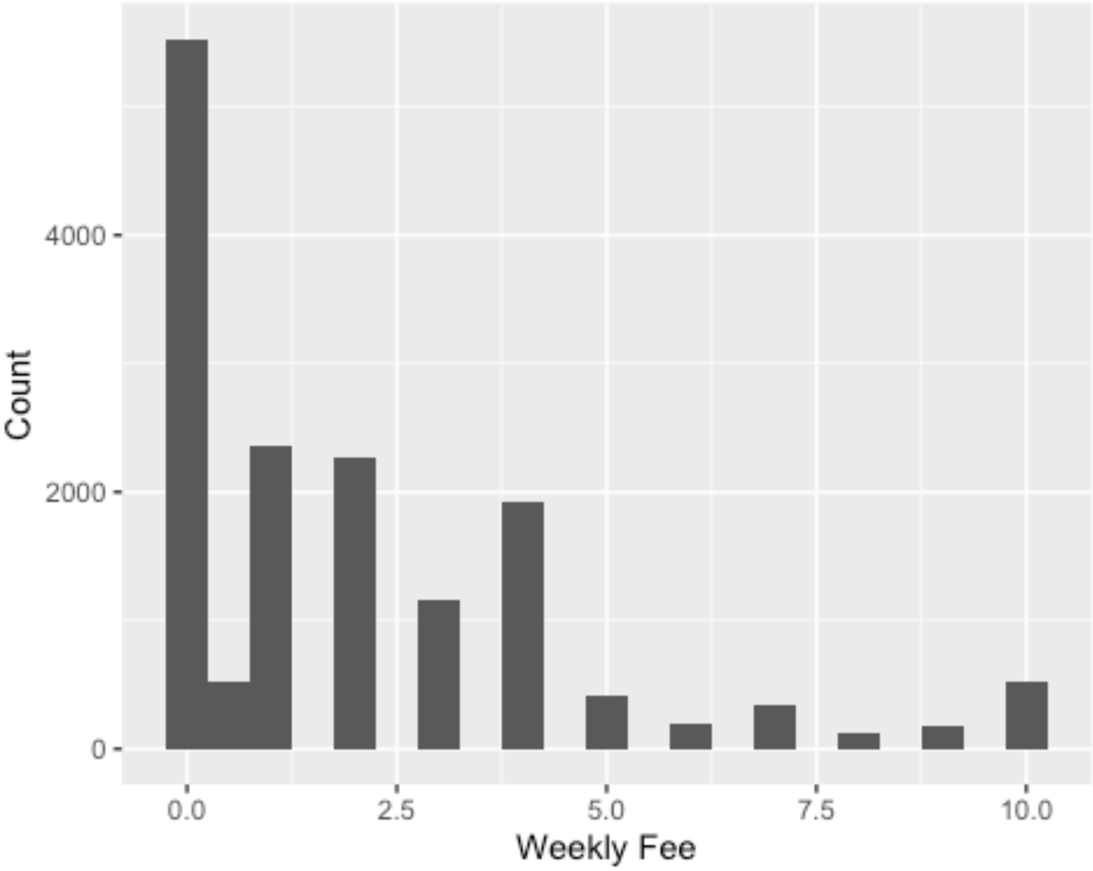
```
ggplot(data = ChurnData_2, aes(Zip_Code)) +
  geom_histogram(bins = 25, binwidth = 10)+
  xlab("Zipcode") +
  ylab("Count")
```



Weekly Fee

This histogram reveals some potential issues with this particular variable. A significant portion pay $0-$0.01 for a subscription (empty values have been removed). I will continue to investigate any potential correlations with other variables but it begs the question of whether to consider a customer lost that paid nothing for a subscription is really a customer that should not churn.

```
ggplot(data = ChurnData_2, aes(weekly_fee_number)) +
  geom_histogram(bins = 13, binwidth = 0.5)+
  xlab("Weekly Fee") +
  ylab("Count")
```



Relationships between variables

I started looking at relationships between one variable and subscriber status. There were no significant positive correlations between any single variable and subscriber status. Here are the calculations for reference.

Household Income

```
cov(ChurnData_2$Subscriber_Number, ChurnData_2$HH_Income_Number)

## [1] -3.969126
```

Home Ownership

```
cov(ChurnData_2$Subscriber_Number, ChurnData_2$Home_Ownership_Number)

## [1] -0.02088387
```

Children

```
cov(ChurnData_2$Subscriber_Number,
ChurnData_2$dummy_for_Children_Number)

## [1] 0.0006507477
```

Year of Residence

```
cov(ChurnData_2$Subscriber_Number, ChurnData_2$Year_Of_Residence)

## [1] -0.9493491
```

Age Range

```
cov(ChurnData_2$Subscriber_Number, ChurnData_2$Age_range_Number, use =
"na.or.complete")

## [1] -1.141553
```

Zipcode

```
cov(ChurnData_2$Subscriber_Number, ChurnData_2$Zip_Code)

## [1] 2.935874
```

Weekly Fee

```
cov(ChurnData_2$Subscriber_Number, ChurnData_2$weekly_fee_number, use
= "na.or.complete")

## [1] -0.1950764
```

I then looked into any potential relationships between any two variables.
Here is a list of the strongest relationships and their value:
1) Zipcode & Weekly Fee - 69.48444
2) Year of Residence & Age Range - 104.295
3) Household Income & Zipcode - 6182.525
4) Household Income & Age Range - 215.9841
5) Household Income & Year of Residence - 216.5007

After plotting the strongest covariance values between two variables and calculating their correlations, there continues to be no significant relationship between two singular variables. Therefore, if there is any relationship, it will have to be combination of multiple variables and I will need to move into a more complex regression analysis. Since I am looking into predicting a categorical variable, I will look into Logistic Regression Analysis.

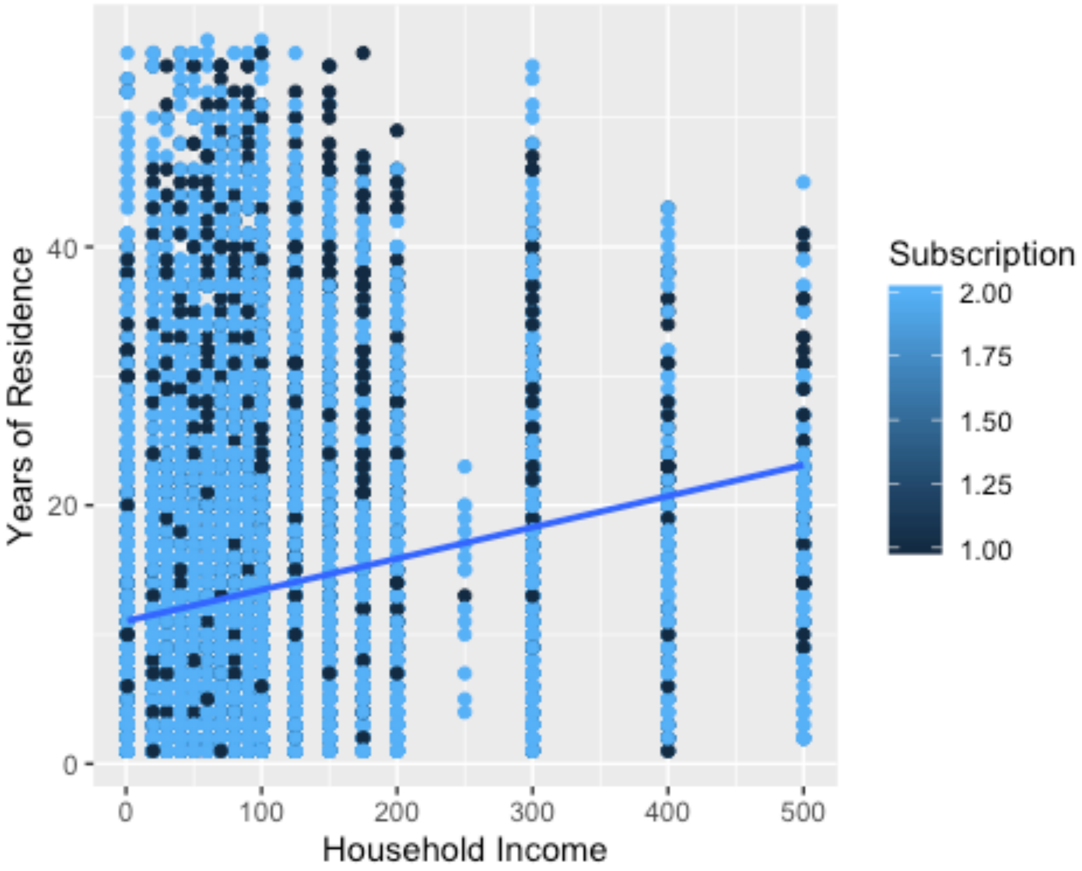
For reference, here are the calculations and plots for the five relationships with the strongest potential.

Household Income - Year of Residence

```
cov(ChurnData_2$HH_Income_Number, ChurnData_2$Year_Of_Residence, use =
"na.or.complete")

## [1] 220.135

ggplot(data = ChurnData_2, aes(x = HH_Income_Number, y=
Year_Of_Residence, color = Subscriber_Number)) +
  geom_point() +
  geom_smooth(method = lm, se=FALSE)+
  xlab("Household Income") +
  ylab("Years of Residence") +
  labs(color = "Subscription")
```

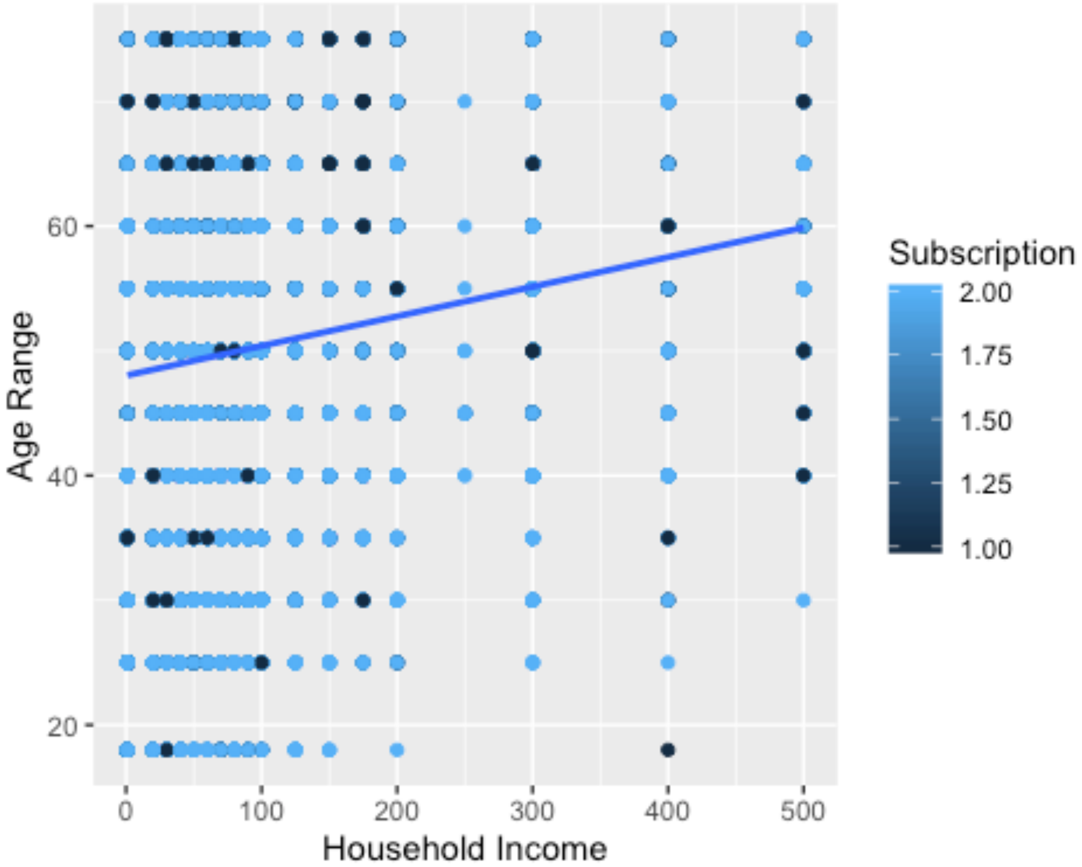


Household Income - Age Range

```
cov(ChurnData_2$HH_Income_Number, ChurnData_2$Age_range_Number, use =
"na.or.complete")

## [1] 216.1004

ggplot(data = ChurnData_2, aes(x = HH_Income_Number, y=
Age_range_Number, color = Subscriber_Number)) +
  geom_point() +
  geom_smooth(method = lm, se=FALSE)+
  xlab("Household Income") +
  ylab("Age Range") +
  labs(color = "Subscription")
```

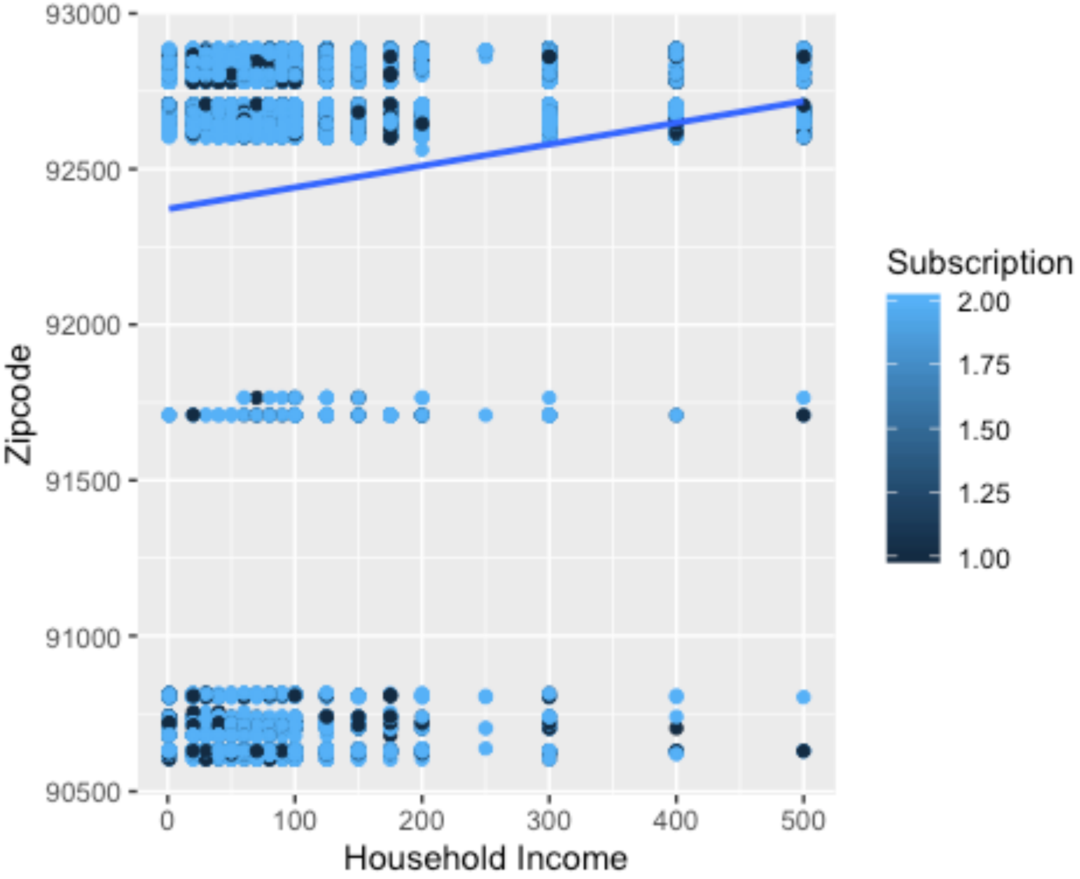


Household Income - Zipcode

```
cov(ChurnData_2$HH_Income_Number, ChurnData_2$Zip_Code, use =
"na.or.complete")

## [1] 6304.287

ggplot(data = ChurnData_2, aes(x = HH_Income_Number, y= Zip_Code,
color = Subscriber_Number)) +
  geom_point() +
  geom_smooth(method = lm, se=FALSE)+
  xlab("Household Income") +
  ylab("Zipcode") +
  labs(color = "Subscription")
```

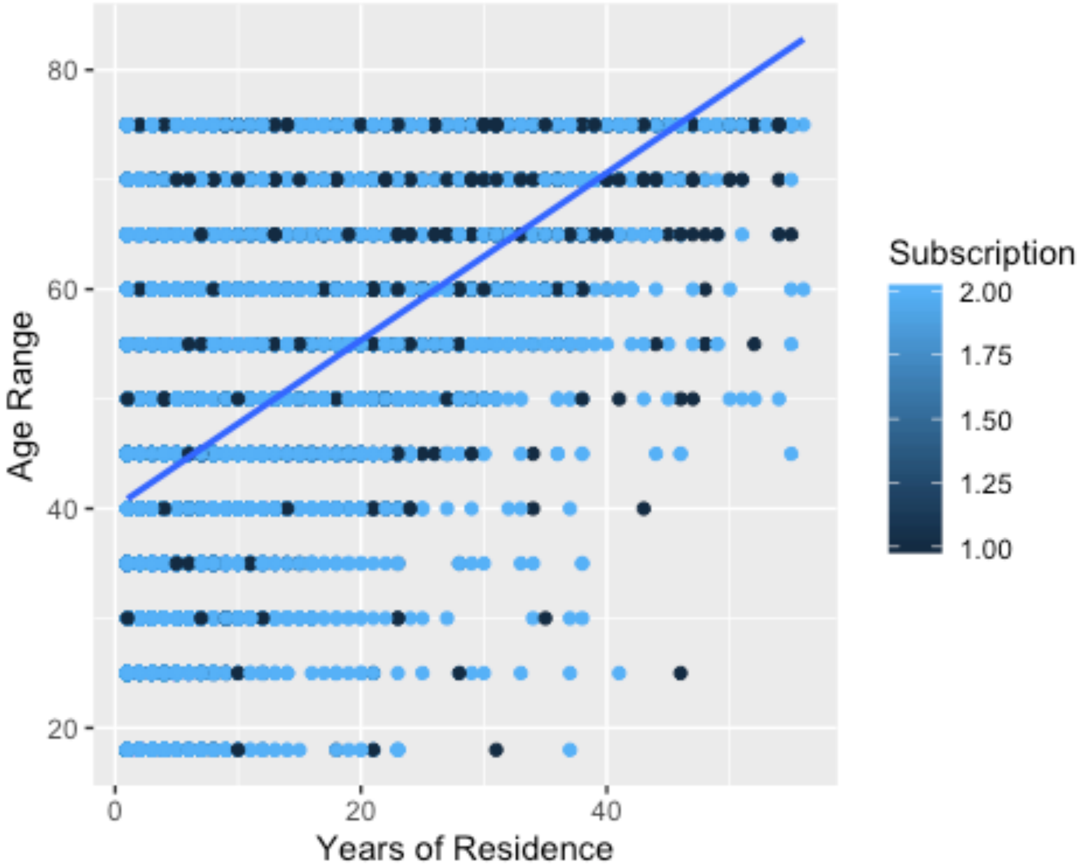


Year of Residence - Age Range

```
cov(ChurnData_2$Year_Of_Residence, ChurnData_2$Age_range_Number, use =
"na.or.complete")

## [1] 103.5607

ggplot(data = ChurnData_2, aes(x = Year_Of_Residence, y=
Age_range_Number, color = Subscriber_Number)) +
  geom_point() +
  geom_smooth(method = lm, se=FALSE)+
  xlab("Years of Residence") +
  ylab("Age Range") +
  labs(color = "Subscription")
```

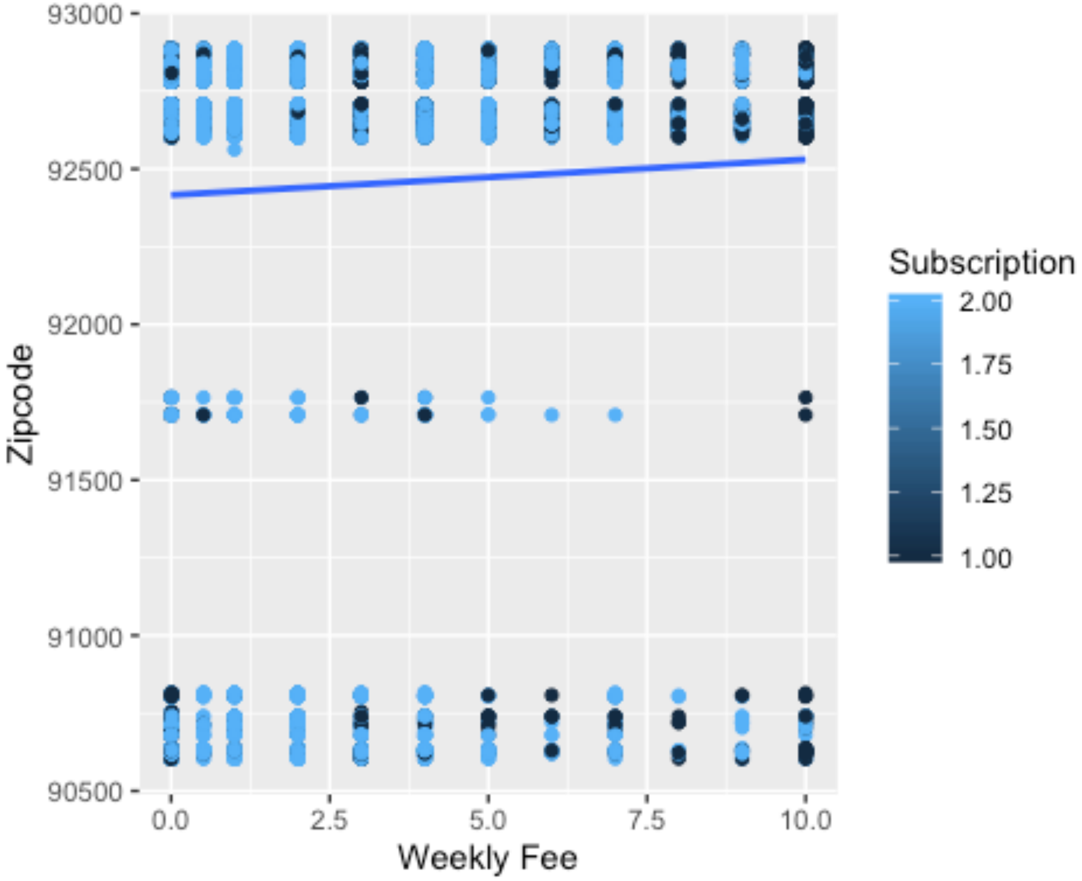


Zipcode - Weekly Fee

```
cov(ChurnData_2$Zip_Code, ChurnData_2$weekly_fee_number, use =
"na.or.complete")

## [1] 71.11318

ggplot(data = ChurnData_2, aes(x = weekly_fee_number, y= Zip_Code,
color = Subscriber_Number)) +
  geom_point() +
  geom_smooth(method = lm, se=FALSE)+
  xlab("Weekly Fee") +
  ylab("Zipcode") +
  labs(color = "Subscription")
```



Logistic Regression Analysis

In order to do logistic regression, I need to correct a portion of my data.

The Subscriber number needs to be binary (0 and 1). Here is the code to create this binary variable.

```
ChurnData_2$Subscriber_Binary <- as.character(ChurnData_2$Subscriber)
ChurnData_2$Subscriber_Binary[ChurnData_2$Subscriber_Binary == "YES"]
<- "1"
ChurnData_2$Subscriber_Binary[ChurnData_2$Subscriber_Binary == "NO"]
<- "0"
ChurnData_2$Subscriber_Binary <-
as.numeric(ChurnData_2$Subscriber_Binary)
```

I created a model with all variables with the Subscription status being the predicted variable.

```
newModel <-
glm(Subscriber_Binary~HH_Income_Number+Year_Of_Residence+Age_range_Num
ber+Home_Ownership_Number+weekly_fee_number+Zip_Code+dummy_for_Childre
n_Number, data = ChurnData_2, family = binomial())
summary(newModel)

##
## Call:
## glm(formula = Subscriber_Binary ~ HH_Income_Number +
Year_Of_Residence +
##     Age_range_Number + Home_Ownership_Number + weekly_fee_number +
##     Zip_Code + dummy_for_Children_Number, family = binomial(),
##     data = ChurnData_2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5486  -0.6644  -0.5388  -0.4113   2.4076
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                1.439e+00  2.700e+00   0.533  0.59404
## HH_Income_Number           1.169e-03  2.217e-04   5.272 1.35e-07
***
## Year_Of_Residence          2.122e-02  2.134e-03   9.946  < 2e-16
```

```
***
## Age_range_Number           1.501e-02  1.794e-03   8.365  < 2e-16
***
## Home_Ownership_Number      1.782e-01  6.332e-02   2.814  0.00489 **
## weekly_fee_number          1.321e-01  7.682e-03  17.201  < 2e-16
***
## Zip_Code                  -5.031e-05  2.922e-05  -1.722  0.08507 .
## dummy_for_Children_Number -4.025e-02  5.274e-02  -0.763  0.44531
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 15320  on 15560  degrees of freedom
## Residual deviance: 14210  on 15553  degrees of freedom
## AIC: 14226
##
## Number of Fisher Scoring iterations: 4
```

The Z scores for each variable are below:
Household Income - 5.272
Year of Residence - 9.946
Age Range - 8.365
Home Ownership - 2.814
Weekly Fee - 17.201
Zipcode - -1.722
Children - -0.763

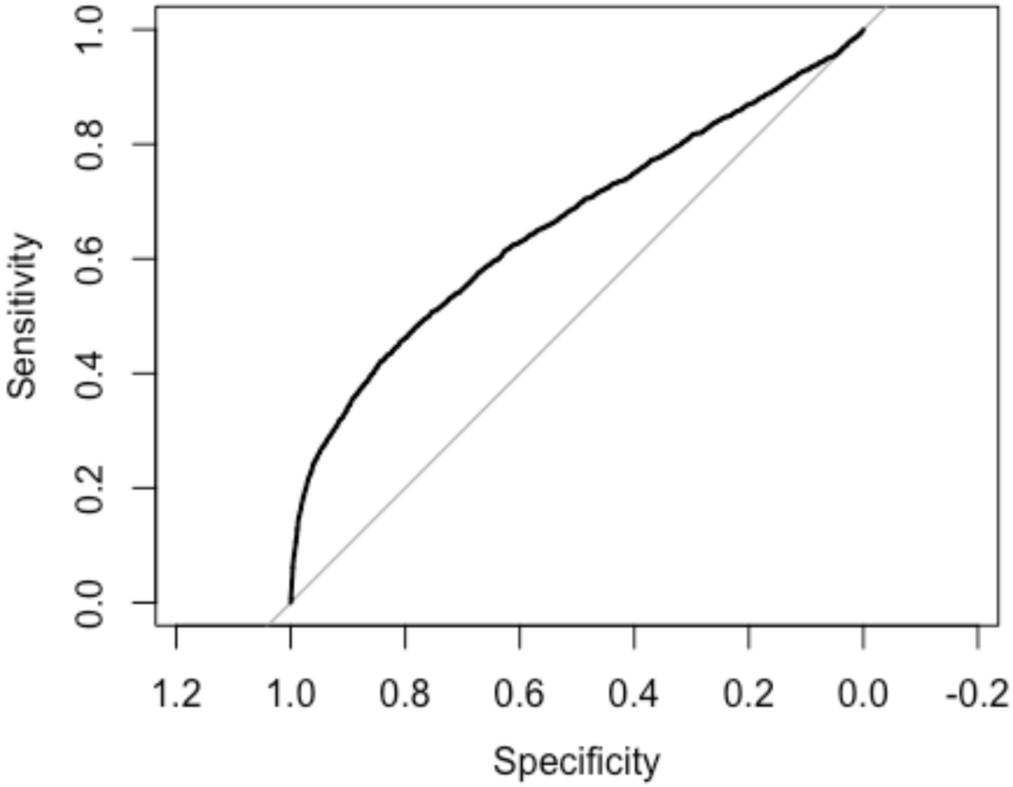
Children and Zipcode are the weakest which I took into account when calculating the Area Under Curve.

First, I calculated the AUC with all variables.

```
roc(ChurnData_2$Subscriber_Binary, newModel$fitted.values, plot=TRUE)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



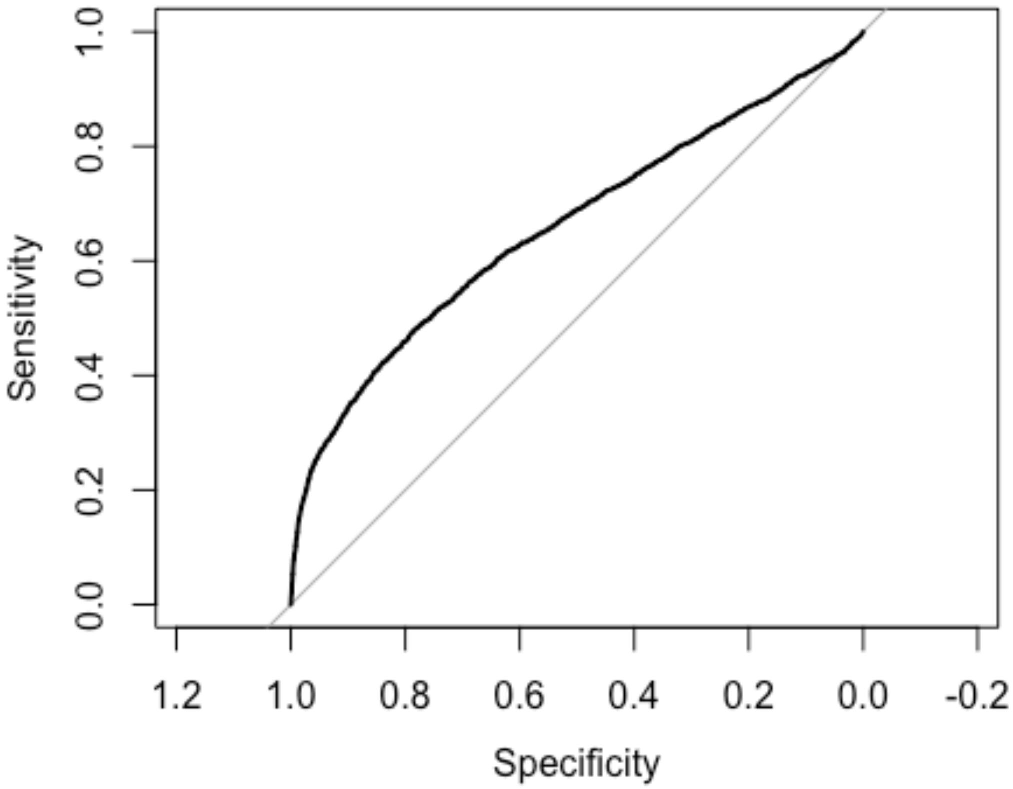
```
##
## Call:
## roc.default(response = ChurnData_2$Subscriber_Binary, predictor =
newModel$fitted.values,     plot = TRUE)
##
## Data: newModel$fitted.values in 12539 controls
(ChurnData_2$Subscriber_Binary 0) < 3022 cases
(ChurnData_2$Subscriber_Binary 1).
## Area under the curve: 0.6605
```

Area under curve is 66% which means there is a 66% accuracy in predicting with all variables. I want to find out if it improves if I remove children and zipcode.

```
newModel_2 <-
glm(Subscriber_Binary~HH_Income_Number+Year_Of_Residence+Age_range_Num
ber+Home_Ownership_Number+weekly_fee_number, data = ChurnData_2,
family = binomial())
roc(ChurnData_2$Subscriber_Binary, newModel_2$fitted.values,
plot=TRUE)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```



```
##
## Call:
## roc.default(response = ChurnData_2$Subscriber_Binary, predictor =
newModel_2$fitted.values,     plot = TRUE)
##
## Data: newModel_2$fitted.values in 12539 controls
(ChurnData_2$Subscriber_Binary 0) < 3022 cases
```

```
(ChurnData_2$Subscriber_Binary 1).
## Area under the curve: 0.6594
```

Accuracy went down <1% so it is not worth removing these variables.

66% is a strong accuracy for prediction so I decided to attempt KNN analysis on all variables. I separated the data into training and testing datasets (70/30) and then calculated the number of rows to assess the best K Value.

```
ChurnData_3 <- ChurnData_2[, c("Subscriber_Number",
"Home_Ownership_Number", "HH_Income_Number",
"dummy_for_Children_Number", "Year_Of_Residence", "Age_range_Number",
"weekly_fee_number", "Zip_Code")]
dat.d <- sample(1:nrow(ChurnData_3),size=nrow(ChurnData_3)*0.7,replace
= FALSE)
train.data_sample <- ChurnData_3[dat.d,]
test.data_sample <- ChurnData_3[-dat.d,]
train.data_sample_labels<-ChurnData_3[dat.d,1, drop = TRUE]
test.data_sample_labels <- ChurnData_3[-dat.d,1, drop = TRUE]
NROW(train.data_sample)

## [1] 10892

NROW(train.data_sample_labels)

## [1] 10892

NROW(test.data_sample)

## [1] 4669

NROW(test.data_sample_labels)

## [1] 4669
```

The number of rows is 10892
The square root of 10892 is 104.36 so I will set the K value at 104 and 105.

```
knn.104<-knn(train=train.data_sample, test=test.data_sample,
cl=train.data_sample_labels, k=104)
knn.105<-knn(train=train.data_sample, test=test.data_sample,
cl=train.data_sample_labels, k=105)
ACC.104 <- 100 * sum(test.data_sample_labels == knn.104)/
NROW(test.data_sample_labels)
```

```
ACC.105 <- 100 * sum(test.data_sample_labels == knn.105)/
NROW(test.data_sample_labels)
ACC.104

## [1] 80.14564

ACC.105

## [1] 80.14564
```

KNN has an accuracy of 80% which is a significant improvement over the 66% from AUC.

Implications

Being able to predict a customer's likelihood to churn at 80% could be significant for any business. The variables ranged from geography to household economics to age.

Limitations

The dataset did not include the reasons why a particular subscriber unsubscribed. This information would have been particularly helpful in looking deeper into the relationship between age range and subscription status. A significant portion of the data was for subscribers 75 and older. This also connects to another limitation which is the data collection on age range. There were inconsistencies identified that cannot be explained.

Another limitation is the data about the Weekly Fee. The histogram showing the spread of this data indicates that many customers are paying $0-$0.01 for the subscription. There needs to be more context as to why there are subscribers with this low of a rate (ex: special offer, rewards program, error in data).

Concluding Remarks

By building a customer profile that includes household income, home ownership, children in household, years of residence in community, age range, zipcode, and weekly fee, I was able to get to a 80% accuracy about if a subscriber was going to churn. By getting access to more detailed reasons why the churn occurred and clarification on some additional variables and context, it is possible that the accuracy would be able to increase.