Mahbubul Hasan (mhasan1)

Elizabeth James (enjames)

Omer Khan (oakhan)

Jennifer Tichenor (jtchnor2)

# Predicting Cable Subscriptions

Using information regarding an individual’s age, gender, income, home ownership, class segment, and number of children, we plan to predict whether or not the individual will subscribe to cable.

<https://www.kaggle.com/amansaxena/cabletv-subscriber-data>

# Report:

This data contains information on people who have subscribed to a certain kind of Cable TV service. The data contains just 300 observations and 7 variables.

The 7 variables are:   
**Age** - the age of the TV subscriber   
**Gender** - the gender of the TV subscriber   
**Income** - the income of the TV subscriber   
**kids** - the number of kids the TV subscriber has   
**ownHome** - if the TV subscriber owns the home or not   
**subscribe** - if they have subscribed to the TV services or not   
**segment** - the segment of the TV subscriber's subscription

Source: Udemy course on Data Analytics

# Preliminary Analysis of the Data

The data is based on 300 observations of people subscribed to cable TV. Cable subscribers range in age from 19 to 80. The median age is 39, while the average age is 41, so the ages skew just a little higher. Income has a minimum of $-5,183 and a maximum of $114,278. We may need to dig into why some incomes report as being negative to understand whether the data is valid or not. The median income is $52,014 while the average income is $50,937. The number of children range from 0 to 7 per subscriber, with the median being 1, while the mean is 1.27. A majority of subscribers have 1 child or no children at all. Additional observations include:

* Slightly more subscribers are women (52%).
* 87% of people polled have not subscribed to cable.
* 53% of people say they do not own their home.
* Given the options to categorize their lifestyle as suburb, urban, travelers, or transitioning,
  + 26% surveyed identified themselves as travelers,
  + 17% selected urban,
  + 23% chose transitioning or “moving up,” and
  + 33% selected suburb.

# R code:

rm(list=ls())

library(readr)

#read/import csv file into R

CableTVSubscribersData <- read\_csv("CableTVSubscribersData.csv")

#determine number of rows

nrow(CableTVSubscribersData)

#nrow < 2500; no volume restriction needed

summary(CableTVSubscribersData)

#summary shows no NA values

#double check by running na.omit and recounting rows

CableTVSubscriber2 <- na.omit(CableTVSubscribersData)

nrow(CableTVSubscriber2)

#new data is same as original data

#no NA values confirmed

#use original CableTVSubscriberData for further calculations

#percentages for character types

#check factors

factor(CableTVSubscribersData$gender)

#returns levels Male Female

#omitted from output report below

factor(CableTVSubscribersData$ownHome)

#returns ownNo ownYes

#omitted from output report below

factor(CableTVSubscribersData$subscribe)

#returns subNo subYes

#subNo subYes will be our on/off end goal for categories

#omitted from output report below

factor(CableTVSubscribersData$Segment)

#returns "Moving up" "Suburb mix" "Travelers" "Urban hip"

#omitted from output report below

library(expss)

#run percentages

Percent\_Male <- count\_if("Male", CableTVSubscribersData$gender)/nrow(CableTVSubscribersData)

Percent\_Male

Percent\_Female <- 1 - Percent\_Male

#only 2 factors makes “1 – “ work

Percent\_Female

Percent\_Own <- count\_if("ownYes", CableTVSubscribersData$ownHome)/nrow(CableTVSubscribersData)

Percent\_Own

Percent\_OwnNo <- 1 - Percent\_Own

#only 2 factors makes “1 – “ work

Percent\_OwnNo

Percent\_Sub <- count\_if("subYes", CableTVSubscribersData$subscribe)/nrow(CableTVSubscribersData)

Percent\_Sub

Percent\_SubNo <- 1 - Percent\_Sub

#only 2 factors makes “1 – “ work

Percent\_SubNo

Percent\_SegMovUp <- count\_if("Moving up", CableTVSubscribersData$Segment)/nrow(CableTVSubscribersData)

Percent\_SegSuburb <- count\_if("Suburb mix", CableTVSubscribersData$Segment)/nrow(CableTVSubscribersData)

Percent\_SegTrav <- count\_if("Travelers", CableTVSubscribersData$Segment)/nrow(CableTVSubscribersData)

Percent\_SegUrban <- count\_if("Urban hip", CableTVSubscribersData$Segment)/nrow(CableTVSubscribersData)

#more than two factors; reason not “1 – “

Percent\_SegMovUp

Percent\_SegSuburb

Percent\_SegTrav

Percent\_SegUrban

# Output:

> rm(list=ls())

> library(readr)

> CableTVSubscribersData <- read\_csv("CableTVSubscribersData.csv")

Parsed with column specification:

cols(

age = col\_double(),

gender = col\_character(),

income = col\_double(),

kids = col\_double(),

ownHome = col\_character(),

subscribe = col\_character(),

Segment = col\_character()

)

> nrow(CableTVSubscribersData)

[1] 300

> summary(CableTVSubscribersData)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | | gender | | income | | kids | | ownHome | |
| Min.: | 19.26 | Length: | 300 | Min.: | -5183 | Min.: | 0 | Length: | 300 |
| 1stQu.: | 33.01 | Class: | character | 1stQu.: | 39656 | 1stQu.: | 0 | Class: | character |
| Median: | 39.49 | Mode: | character | Median: | 52014 | Median: | 1 | Mode: | character |
| Mean: | 41.2 |  |  | Mean: | 50937 | Mean: | 1.27 |  |  |
| 3rdQu.: | 47.9 |  |  | 3rdQu.: | 61403 | 3rdQu.: | 2 |  |  |
| Max.: | 80.49 |  |  | Max.: | 114278 | Max.: | 7 |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| subscribe | | Segment | |
| Length: | 300 | Length: | 300 |
| Class: | character | Class: | character |
| Mode: | character | Mode: | character |

|  |
| --- |
| > CableTVSubscriber2 <- na.omit(CableTVSubscribersData)  > nrow(CableTVSubscriber2)  [1] 300 |
| |  | | --- | | > | |

|  |
| --- |
| > library(expss)  > #run percentages  > Percent\_Male <- count\_if("Male", CableTVSubscribersData$gender)/nrow(CableTVSubscribersData)  > Percent\_Male  [1] 0.4766667  > Percent\_Female <- 1 - Percent\_Male  > Percent\_Female  [1] 0.5233333  >  > Percent\_Own <- count\_if("ownYes", CableTVSubscribersData$ownHome)/nrow(CableTVSubscribersData)  > Percent\_Own  [1] 0.47  > Percent\_OwnNo <- 1 - Percent\_Own  > Percent\_OwnNo  [1] 0.53  >  > Percent\_Sub <- count\_if("subYes", CableTVSubscribersData$subscribe)/nrow(CableTVSubscribersData)  > Percent\_Sub  [1] 0.1333333  > Percent\_SubNo <- 1 - Percent\_Sub  > Percent\_SubNo  [1] 0.8666667  >  > Percent\_SegMovUp <- count\_if("Moving up", CableTVSubscribersData$Segment)/nrow(CableTVSubscribersData)  > Percent\_SegSuburb <- count\_if("Suburb mix", CableTVSubscribersData$Segment)/nrow(CableTVSubscribersData)  > Percent\_SegTrav <- count\_if("Travelers", CableTVSubscribersData$Segment)/nrow(CableTVSubscribersData)  > Percent\_SegUrban <- count\_if("Urban hip", CableTVSubscribersData$Segment)/nrow(CableTVSubscribersData)  >  > Percent\_SegMovUp  [1] 0.2333333  > Percent\_SegSuburb  [1] 0.3333333  > Percent\_SegTrav  [1] 0.2666667  > Percent\_SegUrban  [1] 0.1666667 |
| |  | | --- | | > | |