

Final Project

Internet Prices Around the World

Aretha Kassegnin | Explorations In Data Science | 12-13

# Stage 1

This project is about how internet prices differ around the world. I found out that the country India has free internet for its people, and I started to wonder how that was possible and if it cost a lot to have internet at home. I pay about $100 per month just to have access to the internet. Using the data set I made the target variable: average price of 1 GB and the predictor variables…

1. number of internet plans per country
2. most expensive 1GB
3. internet population per country
4. cheapest 1GB

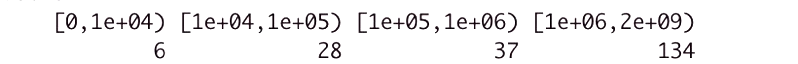
I wanted to see if there’s a relationship between any of these variables and the target.

# Stage 2

To clean up the data I removed all the rows with missing values. I ended up with 205 data points and removed about 35 data points.

Then I binned all the data points and provided below, 5 number summaries from each variable and a 5 number summary of each z-score.

#### Internet Population Per Country



**Figure 0‑1 Internet population binned**

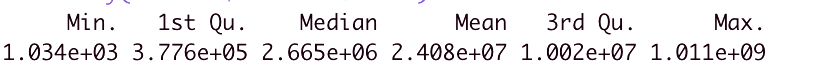


Figure ‑ 5 number summary of internet population

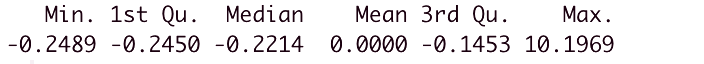


Figure ‑ 5 Number Summary of internet population Z-Score

This variable is for the population that uses the internet as we can see there are high outliers that are 10 standard deviations away from the mean. We can see that from the binning there are 134 counties that has a high population and not that many counties in the lower population.

#### Most Expensive 1GB

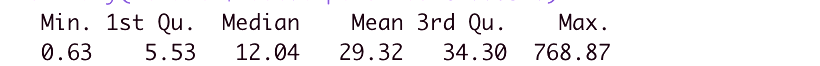


Figure ‑ 5 number summary of most exp. 1gb

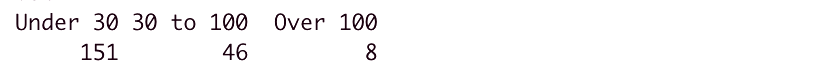


Figure ‑ Binned variable most exp. 1GB

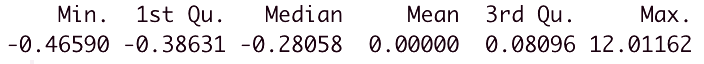


Figure ‑ Most expensive z score 5 Summary

This variable is the most expenses 1GB offered in each country and as we can see there are very high outliers 12 standard deviations away from the mean but I see that most of the data for most expensive is under $30 but the high outliers means that the data is skewed.

#### Cheapest 1GB

A picture containing text

Description automatically generated

Figure ‑ Cheapest 1GB Binned

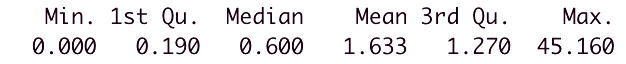
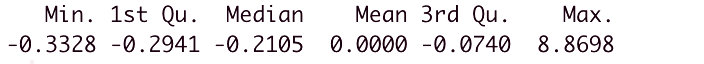


Figure ‑ 5 number summary of Cheapest 1gb



The variable shows the cheapest 1GB offered in each country and this variable also has high outliers and like the most expensive 1GB.

#### Number of Internet Plans per Country

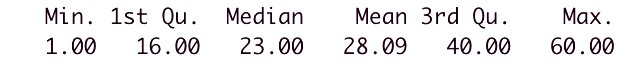
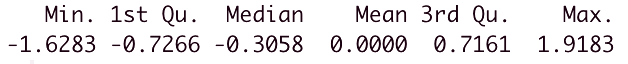


Figure ‑ 5 number summary of # of plans

Text

Description automatically generated with medium confidence

Figure ‑ Number of plans binned



This variable shows the number of plans offered in each country the data here is spread out evenly and we see no outliers and my bins have data is each one of them I think we will see this data in the CART MODEL.

# Stage 3

##### Internet Population

Chart, histogram

Description automatically generatedChart, bar chart, histogram

Description automatically generated

Figure 3.1(normalized) and 3.2(stacked) Average Price vs Number of Plans

##### Most Expensive 1GB

Chart, bar chart, histogram

Description automatically generatedChart

Description automatically generated

Figure 3.3(normalized) and 3.4(stacked) Average Price vs Most Expensive

##### Cheapest 1GB

Chart, bar chart, histogram

Description automatically generatedChart, line chart

Description automatically generated

Figure 3.1(normalized) and 3.2(stacked) Average Price vs Cheapest Plan

##### Number of Internet Plans

Chart, bar chart

Description automatically generated Chart

Description automatically generated

Figure 3.1(normalized) and 3.2(stacked) Average Price vs Internet Users

# Stage 4

### Validate the partition

For this I used the Kruskal test and this is a table of the p- values I got…

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | P-values | Boxplot of Testing vs Training | Thoughts |
| Internet Population | .303 |  | The p-value is greater then 0.05 which means that we can move forward and look at the boxplots we can see that the data for testing and training looks the same so the variable is good to go! |
| Most Expensive | .272 |  | The p-value is greater then 0.05 but the boxplot looks crazy. I wonder its because the data is so weird for this variable we can see from figure 0.6 that the high outliers are about 12 standard deviations away from the mean and we can see here that train got most of the outliers. |
| Cheapest Plan | .670 |  | The p-value is greater then .05 and the boxplots are equal which means this variable is good to go! |
| Number of Internet Plans | .0887 |  | The p-value is greater then .05 and the boxplot looks the same at 1st but the median is not in the same location. I think its important to point this out because it might affect the CART model. |

## Baseline for Model

To establish a baseline for this data I concluded that about 56% of the average price of 1 GB between $1-$4 . This is because based on the figure below…

Text

Description automatically generated

Figure ‑ Training data on Target Variable

# Stage 5/6

Timeline

Description automatically generated

Figure ‑ CART MODEL of DATA

Table

Description automatically generated with medium confidence



Figure ‑ Predictor using training Data

This was the table that was created but I noticed for accuracy that we only worked with two variables when dealing with calculating so I combine the tables but then I realized that if I was to do that my CART model would not work because I divide the data by 3 because it was so skewed in the beginning. So I used the highlighted section to work on my accuracy. Which was 21+71+4/155 = 92% Which means that my CART model is working very well because my baseline was only 56%. For sensitivity I got 71/16+71+1= 72%. Finally for specificity 16/21+16 = 43%

### Appendix of RCode

sum(is.na(newdata))

#DATA PREP

newdata <- na.omit(newdata)

summary(newdata)

average\_price <- as.numeric(as.character(newdata$Average.price.of.1GB..USD.))

summary(average\_price)

average\_price <- cut(x = average\_price , breaks = c(0, 1, 5, 15, 50),

right =FALSE, labels= c("Under $1","$1 to $4","$5 to $15","Over $15"))

#predictors

range(newdata$Internet.users)

users <- cut(x= newdata$Internet.users, breaks = c(0, 10000, 100000, 1000000, 2000000000),

right = FALSE)

summary(newdata$Internet.users)

table(users)

summary(newdata$Most.expensive.1GB..USD.)

most <- cut(x= newdata$Most.expensive.1GB..USD., breaks = c(0, 30,100, 800),

right = FALSE, labels =c("Under 30", "30 to 100", "Over 100"))

table(most)

summary(newdata$NO..OF.Internet.Plans)

amount <- cut(x= newdata$NO..OF.Internet.Plans, breaks =c(0, 15,30 , 61),

right = FALSE, labels= c("Under #15 ", "#15 to 30", "Over #30"))

table(amount)

summary(newdata$Cheapest.1GB.for.30.days..USD.)

least <- cut(x= newdata$Cheapest.1GB.for.30.days..USD., breaks = c(0,1,5,15,50),

right = FALSE, labels= c("Under 1","1 to 4","5 to 15","Over 15"))

table(least)

#Amount of Plans in Each Country

summary(d$NO..OF.Internet.Plans)

boxplot(newdata$NO..OF.Internet.Plans ~ average\_price, horizontal = TRUE)

sd\_mean.no <- sd(newdata$NO..OF.Internet.Plans)

mean\_no <- mean(newdata$NO..OF.Internet.Plans)

z\_score\_no.of.int<- (newdata$NO..OF.Internet.Plans - mean\_no)/sd\_mean.no

summary(z\_score\_no.of.int)

library(ggplot2)

ggplot(newdata, aes(newdata$NO..OF.Internet.Plans))+

geom\_histogram(aes(fill = average\_price),

color = "black", position = "stack")

ggplot(newdata, aes(newdata$NO..OF.Internet.Plans))+

geom\_histogram(aes(fill = average\_price),

color = "black", position = "fill")

#2 COST THE MOST

boxplot(newdata$Most.expensive.1GB..USD.~ average\_price, horizontal = TRUE)

sd\_mean.most <- sd(newdata$Most.expensive.1GB..USD.)

mean\_most <- mean(newdata$Most.expensive.1GB..USD.)

z\_score.most <- (newdata$Most.expensive.1GB..USD.- mean\_most)/sd\_mean.most

summary(z\_score.most)

ggplot(newdata, aes(newdata$Most.expensive.1GB..USD.))+

geom\_histogram(aes(fill = average\_price),

color = "black", position = "stack")

ggplot(newdata, aes(newdata$Most.expensive.1GB..USD.))+

geom\_histogram(aes(fill = average\_price),

color = "black", position = "fill")

#3 users

boxplot(newdata$Internet.users~ average\_price, horizontal = TRUE)

sd\_internet <- sd(newdata$Internet.users)

mean\_internet <- mean(newdata$Internet.users)

z\_score\_users <- (newdata$Internet.users - mean\_internet)/sd\_internet

summary(z\_score\_users)

ggplot(newdata, aes(newdata$Internet.users))+

geom\_histogram(aes(fill = average\_price),

color = "black", position = "stack")

ggplot(newdata, aes(newdata$Internet.users))+

geom\_histogram(aes(fill = average\_price ),

color = "black", position = "fill")

#4 CHEAPEST

boxplot(newdata$Cheapest.1GB.for.30.days..USD. ~ average\_price, horizontal = TRUE)

sd\_cheap <- sd(newdata$Cheapest.1GB.for.30.days..USD.)

mean\_cheap <- mean(newdata$Cheapest.1GB.for.30.days..USD.)

z\_score\_ <- (newdata$Cheapest.1GB.for.30.days..USD. - mean\_cheap)/sd\_cheap

summary(z\_score\_)

ggplot(newdata, aes(newdata$Cheapest.1GB.for.30.days..USD.))+

geom\_histogram(aes(fill = average\_price),

color = "black", position = "stack")

ggplot(newdata, aes(newdata$Cheapest.1GB.for.30.days..USD.))+

geom\_histogram(aes(fill = average\_price),

color = "black", position = "fill")

t1 <- table(average\_price, amount) # amount of Internet plans vs average

t2 <- table(average\_price,least) #cheapest plan for 30 days

t3 <- table(average\_price, users) # internet user population

t4 <- table(average\_price, most)

x <- prop.table(t1)\*100

# checking corr

#CART MODEL

d <- newdata

d <- newdata

sum(is.na(d))

d$Average.price.of.1GB..USD. <- average\_price

d$Internet.users <- users

d$Most.expensive.1GB..USD. <- most

d$Cheapest.1GB.for.30.days..USD. <- least

d$NO..OF.Internet.Plans <- amount

head(d)

library(caret)

set.seed(55)

inTrain <- createDataPartition(

y = d$Average.price.of.1GB..USD., p = .75,

list = FALSE)

d.train <- d[inTrain,]

d.test <- d[-inTrain,]

#check on it

d.train$trainortest <-

rep("train", nrow(d.train))

names(d.train)

d.test$trainortest <-

rep("test", nrow(d.test))

names(d.test)

d.all <- rbind(d.train, d.test)

boxplot(d.all$Cheapest.1GB.for.30.days..USD. ~ (trainortest),

data = d.all)

boxplot(d.all$Most.expensive.1GB..USD. ~ as.factor(trainortest),

data = d.all)

boxplot(d.all$Internet.users ~ (trainortest),

data = d.all)

boxplot(d.all$NO..OF.Internet.Plans~ (trainortest),

data = d.all)

boxplot(d.all$Average.price.of.1GB..USD.~ (trainortest),

data= d.all)

kruskal.test(d.all$Most.expensive.1GB..USD. ~ as.factor(trainortest),

data = d.all)$p.value

#p value = .917

kruskal.test(d.all$NO..OF.Internet.Plans ~ as.factor(trainortest),

data = d.all)$p.value

#p value= .691

kruskal.test(d.all$Cheapest.1GB.for.30.days..USD. ~ (trainortest),

data = d.all)$p.value

#p value = .740

kruskal.test(d.all$Internet.users ~ (trainortest),

data = d.all)$p.value

#p-value =.88

#p-values are > 0.05

########################

#####CART###########

library(rpart)

cart01 <- rpart(Average.price.of.1GB..USD.~ .,

data = d.train,

method = "class")

cart02 <- rpart(Average.price.of.1GB..USD.~.,

data = d.test,

method = "class")

library(rpart.plot)

rpart.plot(cart01, type = 4, extra = 102)

rpart.plot(cart02, type = 4, extra = 102)

library(rattle)

cart01.pred <- predict(object = cart01, newdata = d.train,

type = "class")

t1\_cart.pred <- table(d.train$Average.price.of.1GB..USD., cart01.pred)

prop.table(t1\_cart.pred)\*100

car01.pred\_test <- predict(object= cart01, newdata = d.test,

type = "class")

dim(d.train)[1]

sum(is.na(d.test))

table(d.train$Average.price.of.1GB..USD., pred.cart)

table(d.train$Average.price.of.1GB..USD., pred.cart\_test)

pred.cart

predict(average\_price, newdata=d.train)

#kfold

set.seed(325)

train.control <- trainControl(method = "cv",

number = 100)

d.train.cca <- na.omit(d.train)

model <- train(Average.price.of.1GB..USD. ~ ., data = d.train.cca,

method = "rpart",

trControl = train.control)

plot(model$finalModel)

text(model$finalModel, cex = 0.6)

# rattle package

fancyRpartPlot(model$finalModel,

cex = .6)

# Accuracy of caret CART model

pred.kfold.train <- rpart.predict(object = model, newdata = d.train.cca)

table(d.train.cca$income, pred.kfold.train)

pred.kfold <- rpart.predict(object = model, newdata = d.test)

###########

###########

#C5

#DIDINT WORK

library(C50)

set.seed(325)

library(caret)

mod1 <- C5.0(d.~.,data = d.train)

summary(mod1)

plot(mod1)

table <- table(d.train$Average.price.of.1GB..USD.)

summary(d.train$Average.price.of.1GB..USD.)

prop.table(table)\*100

dim(d.train)[1]

table(d.train)