DOT-Data Handling (wrangling??)-Approach

UshaKiran.Kota

March 27, 2016

Coursolve/Need230 : DOT Data Analysis

Part 1 -Introduction

ref://https://www.dotrust.org/about/

Digital Opportunity Trust (DOT) is a leading international social enterprise headquartered in Ottawa, Canada with local operations around the globe. DOT mobilizes youth the talent and energy helping them to develop both an entrepreneurial spirit and technology and business skills that will last a lifetime. Young people are encouraged to become leaders of change as they facilitate technology, business, and entrepreneurial learning experiences to people in their own communities.

DOT's unique youth-led programs empower people living in communities that are developing, in transition, or under stress with the confidence to use technology for entrepreneurial, community, educational, and personal development.

_In September 2015, twenty youth from Kenya, Rwanda, Ethiopia, Uganda, Tanzania, Lebanon, and Indigenous Canada embarked on a project to survey their peers about how they are using technology and social media for work, income, learning, leadership, and employment. 580 youth from urban, peri-urban, and rural areas were surveyed through face-to-face qualitative and quantitative interviews.

The result of this survey is a raw dataset of 580 responses and more than 1600 variables ,

Coursolve/Need230 is a short term project to analyse:

How are youth in Kenya, Rwanda, Ethiopia, Uganda, Tanzania, Lebanon, and Indigenous Canada using technology and social media for work, income, learning, leadership, and employment?

DOT provided the team with a Raw Survey responses.xlsx as an input data for the exploratory Analysis and expects to find answers for some or all of the below questions from the results of Need230

Document Purpose

The purpose of this document is describe one of the several methods of how to handle dataset of large number of categorical variables to reduce the dimensions of the data set for ease of exploration for research questions

Need230/DOT Data Analysis - Research Questions

1.How does access to technology vary by gender?

2.How does use of social media for work, income, learning, leadership, and employment vary by gender?

3.In what ways are youth in these countries using technology and social media to support or engage in formal work? Informal work?

4.In what ways are youth in these countries using technology and social media to supplement income, or to support primary incomes?

5.In what ways are youth in these countries using technology and social media for learning?

6.What barriers and incentives to online learning are there among the surveyed youth?

7.How are young people positioning themselves as leaders in their communities or among peer groups using social media and technology?

8.How are young people gaining and/or participating in formal or informal employment using technology and social media?

Justification

DOT intends to learn more about how our key stakeholder group (youth in the identified countries) are benefitting from technology and social media in the areas of employment, entrepreneurship, learning, and leadership.

DOT will use the results of Need230 to inform program and project design, as well as implementation strategies.

```
filename = "dotData.RDS"


if(file.exists(filename)){
  print("fileexists")
  #create dot survey response data set from the RDS
  dotsr<-readRDS(filename)

} else{
  dotData <- read.xlsx("dotsurveyresp.xlsx", sheetName="Raw survey responses",stringasfactors=F)
  saveRDS(dotData, "dotData.RDS")

}
```

```
## [1] "fileexists"
```

```
#convert dotData to data table
dot.dt<-data.table(dotsr,keep.rownames=TRUE)
```

Part2 - Sample Data Analysis

Research Question : How does access to technology vary by Gender

In this sample analysis part (which is to be elaborated as project progresses), I selected a sample question from the list to apply a top down approach of data analysis , using few Data Wrangling methods. This is to help : 1. bring the data to a compact form, 2. reduced in dimensions : merge redudant variables to meaningful variables 3. drop unwanted data 4. Make ready for explortory analysis

Approach

1. Select a sample question
2. Select subset of observations and variables from DOT survey data , that are relevant to the sample question
3. Clean Data where necessary
4. Transform the data - categorical values
5. Merge the redudant variables
6. Repeat the approach for all variables in the subset data

Data Insights

DOT Raw survey resonponses data has 580 observations over 1600 variables

Each observation(case) has a unique internal ID for the youth who provided the responses to the survey questions asked by DOT interns

The youth who responded are the target audience (male and female respondents) who will be chosen for DOT training programmes further

The survey respondents belonged to 4 categories of Age : 16-19y, 20-24Y,25-30y, and Age >31.

The responses have been mainly categorial.. hence all the response variables are expanded to have logical responses such as 1 or NA.

Here there is a scope for the response variables to be collapsed into independent categorical variables, by generating levels for responses – Such a change in the data reduced the number of response variables to almost 1/5

Hence a few Data wranging techniques like reshape (using melt, dcast) have been applied to 7 (Gender and Age..) variables to reduce them to only 2.

Part3 - Variable names vs.Variable type

To have an idea check how many vairables and type of variables the DOT data has

```
##           var.name       var.type
## 1:      Updated.At POSIXct,POSIXt
## 2:        Location         factor
## 3:        Username         factor
## 4:   GET.Variables        logical
## 5:        Referrer         factor
## 6: Number.of.Saves        numeric
```

```
##                                                                                                 var.name
##  1:                                                                                         Invite.Code
##  2:                                                                                        Invite.Email
##  3:                                                                                         Invite.Name
##  4:                                                                                           Collector
##  5:                                                                                Researcher.email.address
##  6:                                                                                       Your.location
##  7:                                  What.kind.of.data.gathering.exercise.was.this...Face.to.face.interview.
##  8:                                        What.kind.of.data.gathering.exercise.was.this...Phone.interview.
##  9: What.kind.of.data.gathering.exercise.was.this...Skype.or.other.digital.service.interview.
## 10:                                           What.kind.of.data.gathering.exercise.was.this...Focus.group.
## 11:                         What.kind.of.data.gathering.exercise.was.this...Informal.interview.conversation.
##     var.type
##  1:  logical
##  2:  logical
##  3:  logical
##  4:   factor
##  5:   factor
##  6:   factor
##  7:  numeric
##  8:  numeric
##  9:  logical
```

```
## 10:  numeric
## 11:  numeric
```

Part4 - Sample Data cleanup for : How does access to technology vary by gender?

Transform Variables - Combine Variables with Factors

We first subset only those variables that are required for the sample research Q: ***How does access to technology vary by gender?*** We observe that the DOT data variables are of very long length– to many words in it, we can shorten them for ease of computation

*The variables are spread by their response type – which can be categorical : ordinal or non ordinal

*Date and time variables are to be treated in POSIXCt()

*Long variable names have ".." or "...." as a separator, so can be collapsed to form a lengthy name

*Further, The words in the variable names start with questions like "What", "How""/.. the question part of the long names can be replaced with more generic verbs like"access","use""

```
#Get only relevant columns from the dot survey response data set
dot.tech.gender.dt<-subset(dot.dt, select =                                                    g
#check the dimensions of the subset data frame
dim(dot.tech.gender.dt)
```

```
## [1]  580 1016
```

```
#can I drop the rows with Gender..Male and Gender..Female. == NA

dot.tech.gender.dt<-(dot.tech.gender.dt[which(!(is.na(dot.tech.gender.dt$Gender..Male.) & is.na(dot.tech

#make Gender..Female ==0 for all male rows, do away with Gender..Male rename Gender..Female = Gender
dot.tech.gender.dt<-set(dot.tech.gender.dt,which(is.na(dot.tech.gender.dt[["Gender..Female."]])),  "Gend

dot.tech.gender.dt<-rename(dot.tech.gender.dt, replace = c("Gender..Female." = "Gender"))

#drop Gender..Male. col

dot.tech.gender.dt$Gender <-as.factor(dot.tech.gender.dt$Gender)
dot.tech.gender.dt<-dot.tech.gender.dt[, Gender..Male.:=NULL]

#print the transformed variable:
```

Truncate Long Variable Names

```
################################################################################
names(dot.tech.gender.dt) <- gsub("\\.", "", names(dot.tech.gender.dt))

names(dot.tech.gender.dt) <- gsub("\\Agerange", "", names(dot.tech.gender.dt))


names(dot.tech.gender.dt) <- gsub("\\Howoftendoyouactuallyusethesedevices", "use:", names(dot.tech.gende
```

```
names(dot.tech.gender.dt) <- gsub("\\HowoftendoyouusetheInternetforthefollowingactivities", "actuse:", n

names(dot.tech.gender.dt) <- gsub("Whatdevicesdoyouhaveaccessto", "acc:", names(dot.tech.gender.dt))

#HowoftendoyouusetheInternetforlearningVeryoften
names(dot.tech.gender.dt) <- gsub("HowoftendoyouusetheInternetfor", "internetfor:", names(dot.tech.gende

names(dot.tech.gender.dt) <- gsub("Howoftendoyouusethefollowingdigitalservicestoolsforlearning", "digita

names(dot.tech.gender.dt) <- gsub("Howoftendoyouuseanyelectronicdevicesforthefollowingentrepreneurshipre


#HowoftendoyouuseanyelectronicdevicesforthefollowingworkrelatedactivitiesSMStextmessagingforworkVeryoft

names(dot.tech.gender.dt) <- gsub("Howoftendoyouuseanyelectronicdevicesforthefollowingworkrelatedactivit

#HowoftendoyouusethefollowingdigitalservicestoolsforentertainmentPinterestVeryoften

names(dot.tech.gender.dt) <- gsub("Howoftendoyouusethefollowingdigitalservicestoolsforentertainment", "i

#HowoftendoyouuseanyelectronicdevicesforthefollowingactivitiesTotalktomyfamilyNever

names(dot.tech.gender.dt) <- gsub("Howoftendoyouuseanyelectronicdevicesforthefollowingactivities", "misc

#HowoftendoyouactuallyusetheInternetInternetonabasicmobilephoneOnceamonth
names(dot.tech.gender.dt) <- gsub("HowoftendoyouactuallyusetheInternetInternetona", "intdevice:",    nam

#head(dot.tech.gender.dt)
################################################################################
```

Add new columns for categorical data

create tables with labels for categorical variables age, device and frequency of access/use

```
#create factor variables for devices
devices<-matrix(c("Desktop","Laptop","Tablet", "Smart.phone", "Feature.rich.mobile.phone","Basic.mobile
device<-as.table(devices)

#generate levels for Age variables
age.levels <- matrix(c("Age1619","Age.16.19.", "Age2024","Age.20.24.", "Age2530","Age.25.30.", "Age31",

age<-as.table(age.levels)

freq.levels<- matrix(c("Multiple.times..day.", "Once.a.day.",                                    "Every.
                    "Once.a.month.","Less.than.once.a.month.","Never.",
                    "I.don.t.now."),ncol = 1, byrow = T)

freq<-as.table(freq.levels)

#so we need a clean data set with 3 new variables :  Gender, Age, device, access freq. use freq.

#add 4 new variables to dot.tech.gender.dt table
```

```
dot.tech.gender.dt<-dot.tech.gender.dt[,Age:=""]   # no warning

dot.tech.gender.dt<-dot.tech.gender.dt[,device:=""]   # no warning
dot.tech.gender.dt<-dot.tech.gender.dt[,access:=""]   # no warning
dot.tech.gender.dt<-dot.tech.gender.dt[,use:=""]   # no warning

#check the new dimensions again -- since we added new variables

dim(dot.tech.gender.dt)
```

```
## [1]  560 1019
```

Transform all Age~ Variables - to one "Age"

Do away with 5 Age variables – reducing the number variables of Age from 5 to 1

```
#merge columns of same type
#first subset to a smaller table by unique InternalID and Gender

dot.tech.gender.age<-data.table( subset(dot.tech.gender.dt, select =
                                          grepl("Internal|Gender|Age",

#check the dimensions of this subset
dim(dot.tech.gender.age)
```

```
## [1] 560   7
```

```
#http://stackoverflow.com/questions/25144675/how-to-omit-rows-with-na-in-only-two-columns-in-r

#check if there are any observations for which the age has not been recorded
dot.tech.gender.age<-dot.tech.gender.age[!with(dot.tech.gender.age,
                                                 is.na(Age1619) &
                                                 is.na(Age2024) &

#convert the data table to long format so that all Age~~ variables can be collapsed to one and new age
#in the long format all the NA values in Age will be shown as additional records -- we do not want them
dot.tech.gender.age.m = melt(dot.tech.gender.age, id.vars =
                               c("InternalID","Gender","Age"),measure.vars =
                               c("Age1619", "Age2024", "Age2530","Age31")
                               , na.rm=T, variable.name = "Agerange",
                               value.name = "recordedage")

head(dot.tech.gender.age.m)
```

```
##    InternalID Gender Age Agerange recordedage
## 1:   43755427      1      Age1619           1
## 2:   43912947      1      Age1619           1
## 3:   43920051      0      Age1619           1
## 4:   44005998      1      Age1619           1
## 5:   44008123      1      Age1619           1
## 6:   44064013      0      Age1619           1
```

```r
#check the dimensions in the molten data
dim(dot.tech.gender.age.m)
```

```
## [1] 559   5
```

```r
#convert it to data table
dot.tech.gender.age.m<-data.table(dot.tech.gender.age.m)

#set the new "Age" varible to corresponding age of the youth using the labels created in the age.lables

dot.tech.gender.age.m$Age<-age[dot.tech.gender.age.m$Agerange,2][dot.tech.gender.age.m$recordedage == 1]

head(dot.tech.gender.age.m)
```

```
##    InternalID Gender        Age Agerange recordedage
## 1:   43755427      1 Age.16.19.  Age1619           1
## 2:   43912947      1 Age.16.19.  Age1619           1
## 3:   43920051      0 Age.16.19.  Age1619           1
## 4:   44005998      1 Age.16.19.  Age1619           1
## 5:   44008123      1 Age.16.19.  Age1619           1
## 6:   44064013      0 Age.16.19.  Age1619           1
```

```r
dot.tech.gender.c<-data.table(dcast(InternalID+Gender+Age ~Agerange ,
                                    data =dot.tech.gender.age.m,
                                    value.var = "recordedage",
                                    function(x) length(x)))

#show the table after dcast --
head(dot.tech.gender.c)
```

```
##    InternalID Gender        Age Age1619 Age2024 Age2530 Age31
## 1:   43413695      1 Age.25.30.       0       0       1     0
## 2:   43486387      1    Age.31.       0       0       0     1
## 3:   43645338      0 Age.25.30.       0       0       1     0
## 4:   43675088      1 Age.20.24.       0       1       0     0
## 5:   43749321      0 Age.20.24.       0       1       0     0
## 6:   43750712      0 Age.25.30.       0       0       1     0
```

```r
#drop the redundant Age~ columns
dot.tech.gender.c<-dot.tech.gender.c[ ,`:=`(Age1619 = NULL, Age2024 = NULL, Age2530 = NULL, Age31 = NULL

#show that all the Age variables are well organized into single variable -- data set is now reduced by

head(dot.tech.gender.c)
```

```
##    InternalID Gender        Age
## 1:   43413695      1 Age.25.30.
## 2:   43486387      1    Age.31.
## 3:   43645338      0 Age.25.30.
## 4:   43675088      1 Age.20.24.
## 5:   43749321      0 Age.20.24.
## 6:   43750712      0 Age.25.30.
```

Repeat the data wrangling for the other variables in the technology gender subset data to be contd.