موسسه آموزش عالی آزاد توسعه

برگزار کننده دوره‌های تخصصی علم داده

Homework 2: Fall 2019
Due date: ۲۱ آذز
Please email your HWs to y.zerehsaz@gmail.com
Please append all your codes to your response

*****Please hand in your HWs as a word file with your email as the document's name. For instance, I would name my word file as y.zerehsaz@gmail.com.docx.**
*****Make sure to copy and paste the codes that you used for each question. I need to see your plots, results and conclusions but not the long output of your codes.**
*****When asked, please explain your results.**

a) Please read the googleplay3.csv file and place it in a data frame called df.

df<-read.csv('/googleplaystore3.csv', header=TRUE, stringsAsFactors = F)

The highlighted section should be replaced **by your address.**

b) Get the summary and structure of df.
c) Are there any missing values in the data? Which variables?
d) Change the levels of variable "Category" to 1 to 33.
Hint: use the functions **levels** and **seq**.
e) We can use the "unique" function to extract the unique values used in a vector. For instance, the unique values of vector x=c(1,2,2,3,3,3,1) are 1, 2 and 3. Apply this function to the variable "Type" and comment on its unique values.
Hint: unique(df$Type)
What do you get? How many unique values does this variable have? Any missing values in this vector? What do you think these empty values are?
f) Replace the empty values with NA. You can use the same method which we used to replace zeros with NAs in the PIMA dataset. Please keep in mind that these are not **zeroes** but they are

empty values. You need to find these empty values and then replace them with NAs. The empty values are denoted by " " in R.

g) Get the levels of variable "Type" using the **levels** function. What do you get? Get the summary of df and check the NA's for the variable "Type".

h) The "Last.Updated" column should be in "Date" format. Is it? If not, we need to coerce it to be a date variable so that R knows these are dates and not characters. But, the format might be different for each computer. This means that the "Last.Updated" column in your csv file might have a format, like 2018-08-01, 01-08-2018, 2018/08/01, 01/08/2018. Or, the month and day elements might have changed their locations on your computer. So, we need to try different formats when changing the structure of this variable to "date" format. Use the following function.

df$Last.Updated<-as.Date(df$Last.Updated, tryFormats = c("%m/%d/%Y","%Y/%d/%m","%Y-%d-%m","%m-%d-%Y","%d/%m/%Y","%Y/%m/%d","%Y-%m-%d","%d-%m-%Y"))

The argument "tryFormats" indeed tries different date formats to see which one fits the format of your data in df$Last.Updated.

If this does not work for you, just check the format on your excel file, and then add that format in the "tryFormats" argument.

i) After successfully changing the format of the variable df$Last.Updated, we need to compute a new variable called "Updates". The new variable should give the number of days since the last update of the application. So, compute the difference between your system date and the column df$Last.Updated, and place this vector in df$Updates. This will generate a new column in your data frame called "Updates". Now, you just need to coerce this column to be a numeric variable. Do as follows

df$Updates<-as.numeric(df$Updates)

Get the str(df) and check the structures.

j) Get the first five values of the variable "Installs" in the dataset. As you can see, there is a "+" sign at the end of the number of installs. This changes the structure of this variable to character while it should be a numeric variable. We need to remove the "+" sign from the end of the number of installs and change the whole column into a numeric variable.

Consider the first element of this variable which is "10,000+":

Since the "+" sign is at the end of variable, we can access it by first splitting it then removing the last element. These are the steps:

1- Split this value using strsplit("10,000+","") function and put them in a variable called "spinst". You can use df$Installs[1] instead of the exact value in the first argument. Recall that the second argument in the function ("") means that we would like to split the term by its characters.

2- Change this list to a vector as spinst=unlist(spinst)

3- Now, you have a vector with 7 elements (only for this example). Remove the last element from the vector. You need to figure a way to find the last element of vector and remove it.

4- Before rejoining the characters in spinst, note that some of the values in the variable "Installs" contain one or more "," characters. Having such characters in the variables is problematic, for we

cannot force the variables to be numeric. Hence, we need to get rid of them. I will simply give you the code but try to understand it.
spinst=spinst[spinst!=”,”]
5- We can use the paste function with its collapse argument to rejoin all the elements. Use the paste function as paste(spinst,collapse=?) and replace ? with the appropriate argument.
6- As you have seen, we performed all these operations on only one element. In practice, we need to do these steps for all elements. You do not need to do this but the way that I address this problem is to write a function performing all these steps and then use the “apply” function to repeat the whole procedure for all the elements of the variable “Installs”.
k) Use the function mice with m = 2 to impute the missing values. Before that, please remove the first, second, seventh and eighth columns from the dataset. That is, your dataset should not have the columns “X”, “App”, “Last.Updated” and “Installs”. Use the following function to perform the imputation:
imput<-mice(df,m=2,meth=c("polyreg","sample","pmm","logreg","pmm"))
Use the command ?mice to see what this function does exactly and explain it.
l) Use the margin and aggregations plots to interpret the missing values patterns.
m) Use the xyplot, stripplot and densityplot to comment on the performance of imputation.
n) Perform the imputation using the **complete** function and get the summary of the resulting data frames.