Code ▾

Netflix Userbase Data Preprocessing

Deepthi Suresh & Tathya Grover

08/08/2023

Setup

Hide

```
library(kableExtra)
library(readr)
library(dplyr)
```

Data Description

The data set provides an insight on Netflix user base with a 1 month plan and the monthly revenue generated. The data contains information on 2500 users between the age group of 26 and 51. Each user can be identified by their unique User ID. The data set has the following variables:

- User Id - Unique identifier of the users.
- Subscription Type -The type of subscription plan the user has opted for. Netflix has three types of subscription: Basic, Standard and Premium.
- Monthly Revenue - The monthly revenue generated by their subscription.
- Join Date - The date they joined Netflix.
- Last payment Date - The date of their last payment.
- Country - The country in which they are located in.
- Age - Age of the user base.
- Gender - Gender of the user base.
- Device - device they use to access Netflix (Laptop, Smart TV, Tablet, Smartphone)
- Plan Duration - The Netflix plan duration that the user has opted for.This data set contains users with a 1 month plan.

The data is sourced from Kaggle and has the below URL:

J, Arnav. n.d. Netflix Userbase Dataset.Accessed on 01/08/2023.Available from https://www.kaggle.com/datasets/arnavsmayan/netflix-userbase-dataset?resource=download (https://www.kaggle.com/datasets/arnavsmayan/netflix-userbase-dataset?resource=download).

Read/Import Data

Hide

```
setwd("/Users/deepthisuresh/Library/CloudStorage/OneDrive-RMITUniversity/Data Wrangling-Deepthi’s MacBook Air/Practical Assignment 1")
Userbase<-read_csv("Netflix Userbase.csv")
class(Userbase)
head(Userbase)
```

The working directory is first set to the location in which the data set is located by using ‘setwd’ function.The data set of Netflix user base is downloaded from Kaggle in the CSV format and imported into R using the ‘read_csv’ function from the Readr package.The read function also provides additional information about the data set. The data set has 2500 observations and 10 column specifications.As a default, while importing, the data is stored in the form of data frames as confirmed through the ‘class’ function. There are currently 7 character variables and 3 Numeric or double variables. The ‘head’ function is called to look into the first 6 observations to preview and make sure the data is imported correctly.

Inspect and Understand

Hide

```
dim(Userbase)
names(Userbase)
str(Userbase)

unique(Userbase$`Subscription Type`)
Userbase$`Subscription Type` <- factor(Userbase$`Subscription Type`,levels = c("Basic","Standard","Premium") ,ordered=TRUE)
levels(Userbase$`Subscription Type`)

unique(Userbase$Gender)
Userbase$Gender <- as.factor(Userbase$Gender)
levels(Userbase$Gender)

unique(Userbase$Country)
Userbase$Country <- as.factor(Userbase$Country)
levels(Userbase$Country)

unique(Userbase$Device)
Userbase$Device <- as.factor(Userbase$Device)
levels(Userbase$Device)

str(Userbase)
summary(Userbase)
```

The ‘dim’ function in R returns the dimensions of the data set. There are 2500 rows and 10 columns. The ‘Names’ function returns the variable names - User ID, Subscription Type, Monthly Revenue, Join Date, Last Payment Date, Country, Age, Gender,Device,Plan Duration.The ‘Str’ function shows the structure of the data along with details on the data types. In this data set, User ID, Monthly revenue and Age are numerical variable and rest are Character variable. Some of the character Variables like Subscription Type, Gender,Country,Device were then converted into a factor as they met the requirements of a factor variable. In order to do that, ‘Unique’ functions was used to identify the unique values in each variable and then ‘as.factor’ function was applied on it. When the levels of the above factor variable was checked, it was noticed that the Subscription Type could be a ordered factor variable with the ranking Basic < Standard < Premium. This was done using a combination of ‘as.factor’, ‘levels’ and ‘ordered’ functions

The ‘Str’ function now showed the updated factor/ordered variables, along with other numeric/character variable. The ‘Summary’ functions provides quantitative information such as mean,median, quartiles as well as the total count at each level for the factor variables.

Subsetting

Hide

```
user_sub <- Userbase[1:10,]
str(user_sub)

mat1 <- as.matrix(user_sub)
is.matrix(mat1)
rownames(mat1) <- c("observation1", "observation2", "observation3","observation4","observation5","observation6","observation7","observation8","observation9","observation10")

attributes(mat1)
str(mat1)
```

A subset of 10 observation from the data set was created and named ‘user_sub’ using the subsetting feature. User_sub is a data frame and its variables have the same data types as the parent set thus doesn’t require any data type conversions. The subset was then converted into a matrix using the ‘as.matrix’ function and named ‘mat1’ and then the rows were renamed using ‘rownames’ function. The ‘attribute’ function shows that these changes have been completed and the new row names are displayed. When the structure of the mat1 matrix is checked, it can be noticed that the structure has changed to character data type. This is because, by definition matrix can have only one single datatype, so a mix of factors and characters converts it to a character data type as a character is the most flexible of all data types, leading the entire matrix to become character.

Create a new Data Frame

Hide

```
df1 <- data.frame(CustomerAge = c(35,42,28,50,31,39,47,55,24,36),
                  FeedbackRating = c( 3,    4,  5,  1,  4,  2,  3,  1,  4,  2))

is.data.frame(df1)
str(df1)

df1$CustomerAge<-as.integer(df1$CustomerAge)

unique(df1$FeedbackRating)
df1$FeedbackRating <- factor(df1$FeedbackRating , levels = c(5, 4,  3,  2,  1),
                                labels = c("Excellent","Very Good", "Good","Average","Poor"),
                             ordered = TRUE)
str(df1)
levels(df1$FeedbackRating)

PurchaseFrequencyPerMonth <- c(6,5,8,1,6,5,7,2,3,4)
df1 <- cbind(df1,PurchaseFrequencyPerMonth)
head(df1)
df1$PurchaseFrequencyPerMonth <- as.integer(df1$PurchaseFrequencyPerMonth)
str(df1)
```

A data frame called ‘df1’ was created using a new dataset with the help of ‘data.frame’ function. The data set contains information about Customer Age and their feedback rating for a retail store. The customer age group ranges from 24-55 and the rating is given on a scale from 1-5, where 1 is Poor and 5 is Excellent. The data frame variables were then converted into the appropriate data types i.e Customer Age was converted into integer using the ‘as.integer’ function and Feedback Rating as Ordered Factor, with 5 levels - Excellent, very Good, Good, Average and Poor using the ‘as.factor’ function along with ‘ORDERED = TRUE’. (The ‘unique’ function here helps in identifying the unique values in the observations).

Another variable called ‘purchase frequency per month’ (the number of purchases by the customer in a month) is added to the existing data frame using the ‘cbind’ function, creating a data frame of 10 observations and 3 variables. The purchase frequency per month variable was then converted into an integer using the ‘as.integer’ function for correct data manipulations.

References

J, Arnav. n.d. Netflix Userbase Dataset.Accessed on 01/08/2023.Available from https://www.kaggle.com/datasets/arnavsmayan/netflix-userbase-dataset?resource=download (https://www.kaggle.com/datasets/arnavsmayan/netflix-userbase-dataset?resource=download).

Sona Taheri (2023) Data Wrangling [Module 2 & 3 Demo and Notes], RMIT University,Melbourne