Cleaning Netflix Data with R

Shweta Kothawade

Author Note

Northeastern University

Cleaning Netflix Data with R

For cleaning the raw data of the given Netflix excel file that contains **3033** observations and **10** variables, we first save the original excel file as CSV format, then use the **read.csv** command in R to load the data with headers. For import data, we can also use the original xlsx file to R studio which the strings are differently distributed, but it works both ways. After we load the raw data, the empty values in ‘Completed’ is showing ‘NA,’ but the empty ones in ‘Gender’ columns remain as empty, which we used the **equal sign** as a command to set all the cells with empty values to NA for further constructions.

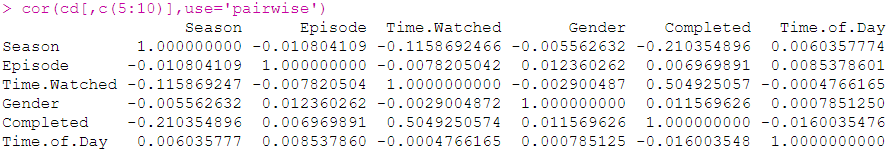
Next, we rename the columns that make confusion about the date. We rename the “Day of Week” to “Date of Week” and “X” as “Day of Week” since the Day of Week has an empty name. Then we had a BOM situation where there are three extra characters before the first column’s name, so we removed it. After we created the table named ‘cd’ as only showing the rows without ‘NA’ using the **drop\_NA** command, then we have **2896** observations and **10** variables in total. As for further establishment, we change the type of Gender into numeric data, male as 1 and female as 0 using the **ifelse** command. As a way to find correlations from the variables, we applied **cor.test** command to apply a hypothesis test between user’s gender and watched time and have the output as **-0.0029**, then used **paris()** to plot the correlation between the two variables

There are more options we can apply regarding data cleaning, but for this assignment, we mainly used basic R functions to clean the missing values. Besides basic functions in R, we can also apply the “tidyverse” package which includes both “dplyr” and “tidyr”. “Piping” would be a great and advanced method to clean and sort the data such as the “rename”, “arrange”, and “group\_by” statements. Piping creates more readable codes to understand, but our group has limited experience with R, so we mainly applied only basic functions from R studio. We will figure out how to work with these packages through future assignments in the course.

Advanced options

The most obvious trend after cleaning the data is that the observation has decreased from 3033 to 2896 since we have 99 values in ‘Gender’ remain NA and 38 in ‘Completed’. This indicates that we have 137 rows (4.5% of the data) that have been removed from the dataset. Here we came up with a question: why is the value in the ‘Completed’ column missing? The result is either 0 (not completed) and 1 (completed), what is the calculation behind this? If the records were automatically saved, in which situations can the data remain missing?

We use the hypothesis test in R to find out the correlation between the gender and watched time of the user in Netflix, the output shows the correlation is -0.0029 which is a significantly weak correlation and almost no correlation between gender and total watched times. We can assume that the time of watched time has no relationship with the gender of users.



Above shows the overall correlations of numeric variables from Season to Time of Day. We can see from the correlation between Time.Watched and Completed returns a positive correlation which indicates that the time of watching on Netflix is correlated to the Completion. There is more to discover from the dataset, in Week 1 we have cleaned and compiled the data.