Rachel Harvill, Marco Gonzalez, Maha Hamdi, D’Angelo Guiton

Professor Haiyan Cheng

Data Science CS 429

7 February 2018

Homework #1

As an example of exploratory data analysis, thirty-one New York Times data files (nyt1-31.csv) were used to create a set of visual understandings for metrics and distributions over time. The variables immediately available in the nyt data sets included “Age,” “Gender,” “Impressions,” “Clicks,” and whether or not the user was “Signed-In.” To better understand the data, metrics were created to help visualize trends within the statistics. These metrics were defined as CTR (Click-Through-Rate) and ACR (Age-Click-Rate) and allowed for a more focused analysis by limiting the immediately observed variables to “Clicks,” “Age,” and “Impressions.” While further use of other metrics could lead to a deeper understanding of the data trends, it is important to limit the scope of analysis as it is a mistake to presume that every variable contributes in some meaningful way.

Our analysis began first with observations of the immediate variables. Skimming through the data, it was immediately apparent that certain quantities were either incorrect or missing. Examples of this included several user’s ages being set to a value of zero. To correct this issue, Marco solved the issue of data cleaning through the creation of a PHP script which not only removed such values, but compiled the thirty-one individual nyt data sets. With the data cleaned, the CTR metrics was observed by dividing clicks by impression while ACR was defined as Age divided by Clicks. This process was done in R and with the help of the subset function, was divided among age groups with an approximate ten year spread.   
  
  
 Continuing the exploratory data analysis began with visualizing the data and extending the study across all 31 days. Through the use of R’s ggplot function, impressions were categorized by age groups in a blox plot to observe variation while Clicks by Gender were visualized in a histogram. CTR by Age group was later plotted to further the group’s understanding of each user’s engagement with the ads. ACR was visualized through a density plot to help draw conclusions about user tendencies with regards to their age. Finally plots of Clicks and CTR over time were generated through bar and step graphs respectively.

Conclusions drawn from the data and their respective visualizations allowed the group to determine that impressions vary over time but the mean value of said impression does not change. Furthermore, the mean number of clicks have a skewed left distributions, indicating an increase at the end of the day. Similarly, the mean CTR is skewed left in a similar matter to clicks, but the max CTR does not appear to change significantly over time. Finally, based on data from the ACR variable and its respective graphs, indications that younger users click ads more often than older users could be clearly seen.