Can Time On a Website Predict the Website’s Convenience to Users?

*Analysis of Website Users Trends and Habits*

Blake Conrad

STAT-35000

Professor Tooman

September 8th, 2016

Abstract:

The internet age has ushered in a surge of data directly linked with individual’s time and interest. Computer Scientists work tirelessly to provide structured, convenient, and reliable websites for users all around the world. By analyzing where individuals spend their time, Computer Scientists can more easily draw conclusions on where the users’ interests lie, how to tailor effective websites, and utilize content in a more direct way to their intended audiences. By studying a sample of reliable data found at <https://analytics.usa.gov/defense/data/>, patterns are studied to find quantitative relationships related to user activity, time, and webpages viewed. Mean, median, standard deviation, correlation, and regression are all analyzed in the research on target variables from the provided data set. By looking at relationships form the target variables in the sample space, a strong foundation is laid for inferential statistics to be applied with the findings from the research for a broader population.

In Computer Science, the understanding and utilization of the internet is essential to becoming a successful Computer Scientist. The rapid increase of internet activity in the 1990’s and 2000’s has drastically increased internet traffic and interest. The age of the internet and internet coding has increased drastically as well in recent years, and with this surge of internet activity has come a surge of data. Not just any kind of data, but data directly related to where people spend their time. Real people with unique interests from all over the world choose to spend their time on the internet every day. By analyzing where specific individuals spend their time, one can more easily draw conclusions on where the users’ interests may lie. By understanding the interests of a particular set of users, a good Computer Scientist can more easily program their websites with the user’s interests in mind. So by finding a relationship between users’ time spent on a webpage and how many webpages they viewed, we are indirectly finding the relationship between how effective that website was to its audience of users. Of course, not all data is the same data, however, by analyzing a small sample of reliable data, implications may be able to be drawn with reasonable confidence through inductive reasoning that other samples of data in different areas that may be seemingly unrelated, also hold the same or similar relationship(s). Data comes in all shapes and sizes on the internet. We are most interested in a small sample of reliable data from reliable websites (domains). This will allow for better analysis, relationships, and conclusions. By looking at United States Government and Military websites data from the past 30 days, we can find credible website activity from a credible source.

With data found at <https://analytics.usa.gov/defense/data/>, there exists 66 different government websites including: defense.gov, navy.mil, ms.army.mil, italy.army.mil, health.mil, and many more similar websites related to the U.S Government and U.S Military. These 66 domain observations include their own compressed set of variables which associate directly with the activity in the last 30 days of each domain. The variables associated with each observation are: visits, page views, users, page views per session, average session duration, and exits. So the data set is comprised of 66 website domains as our observations, each with 6 variables to describe each observation’s activity in the past 30 days. The data set contains only quantitative variables, as listed in *Figure 1*.

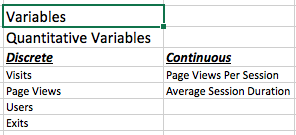
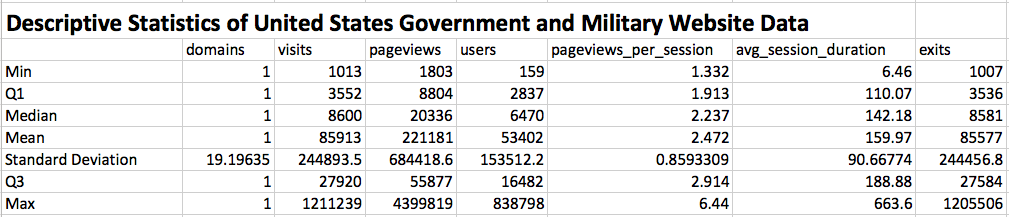
By only having quantitative variables, adding or averaging these values makes sense. So we can also find units of measurement such as the mean, median, mode, and the distributions associated with certain variables against others. As shown in *Figure 2*, there is a list of some descriptive statistics on all of the variables in our data set. One example of a categorical variable that could have been associated with this data set might have been the addition of a variable distinguishing whether the domain observation is government or military. Having a categorical variable to divide the data into two groups would allow for further exploration into our data set by diving it into categories to analyze side by side or against one another.

Figure 1:

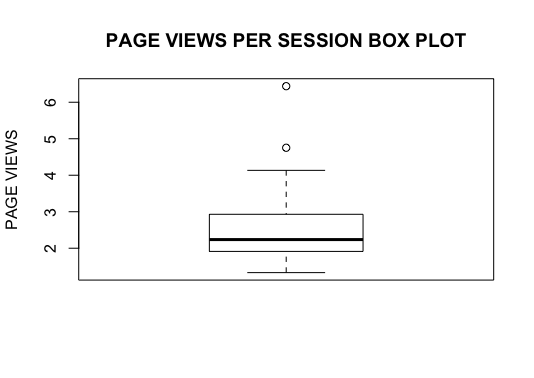
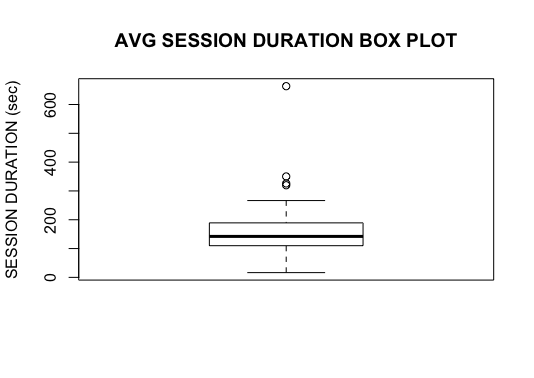
Figure 2:



With all of our variables being quantitative, we can measure any of them against one another and look for relationships. From our initial conjecture, we know that the two variables regarding pages viewed and time spent on a page, will support our intuition of whether or not a page is relevant to a user or not. By viewing these two variables against one another we can more clearly see what relationship the two variables have with our goal. Before diving into the relationships amongst our target variables, we can examine *Figure 2* followed by *Figure 3* to understand our data set better. Off hand, notice how our mean and median are very closely related for page views per session and average session duration, individually in *Figure 2*. We can also note that there may be some positive skewness because of average session duration’s max value. While considering max values, our illustration in *Figured 3* shows the box plots of both of our individual variables with some positively placed outliers in both variables. This connection may lead to some interested conclusions as we continue in our analysis. We can even visually see some outliers in *Figure* 4, as we will discuss more in depth later. *Figure 3* expresses what *Figure 4* illustrates by the points being outside of the fenced areas. With this background information in mind, we can try to begin our attempt for a prediction as pages viewed increases on a website, so does time spent on the page and by nature the relevance of the website to the user’s interests.

Figure 3:

::



# 

As we noted before in *Figure 3,* we can see outliers for both our target variables. *Figure 4* also visualizes these outliers, but not all of them are as explicit. We can see our prediction in action by a scatter plot and linear regression in *Figure 4*. The correlation coefficient of our target variables is 0.74, showing us a medium strength between our target variables. This information shows us that we have a medium, positive, and linear relationship with some outliers identified from *Figure 3*. The regression line has the form: , with a more linguistic representation as: Time on Page (Pages Viewed) = -34.047 + 78.486 \* Pages Viewed. By plugging in a few of our data points into our regression function we can find some residuals. For example if we use our data points of Tricare.mil with 3.632494495 pages, the actual average time spent on the site was 228.8350474. We can use our regression line’s prediction value, 251.0529629, and our actual value 228.8350474. Thus our residual is -22.21791553 (Actual – Predicted). This simply tells us our actual value was -22.21791553 less than the predicted value. We also have an r2 of 0.5533, which tells us the proportion of variation in response from time spent on the page, based on our linear relationship with pages viewed is 55.33%.

Figure 4:

::

# images/scatter_1.png

One must always be aware that correlation ≠ causation. Without further analysis using inferential statistics, we simply cannot draw good or accurate conclusions to a population. However, by view of *Figure 4,* we can certainly see average time a user spends on a military or government webpage has a linear relationship with the pages each user viewed. A relationship between our target variables has certainly been discovered, which is a great discovery. A Computer Scientist can certainly utilize this intelligence to learn and construct even more useful and relevant websites for the intended audience. By observing a very detailed and reliable sample data set we can see patterns in our data. With a 55.33% variability in our prediction, we can’t guarantee the user found all the content on the web pages helpful, but through further exploration in this data set and use of more advanced statistical methods, we might be able to draw probable inferences that other web sites in general might behave in the same way. Additional exploration in our data set could be the analysis of how the two different types of domains, government and military, were used in comparison to one another. Only through further exploration can we hope to gain true answers to our population as a whole and potentially answer our question: *Can time on a website predict the website’s convenience to users?*