User Search Behavior Analysis Report

Bingjie Zhang

Data Set Description

The data set used to analyze user behavior contains three types of user behavior, namely start, results and click. The start actions describe situations when users started their Internet search. Results actions describe the time it took to display users' search result as well as when their searches happened. Click actions describe the time when users clicked a certain search result.

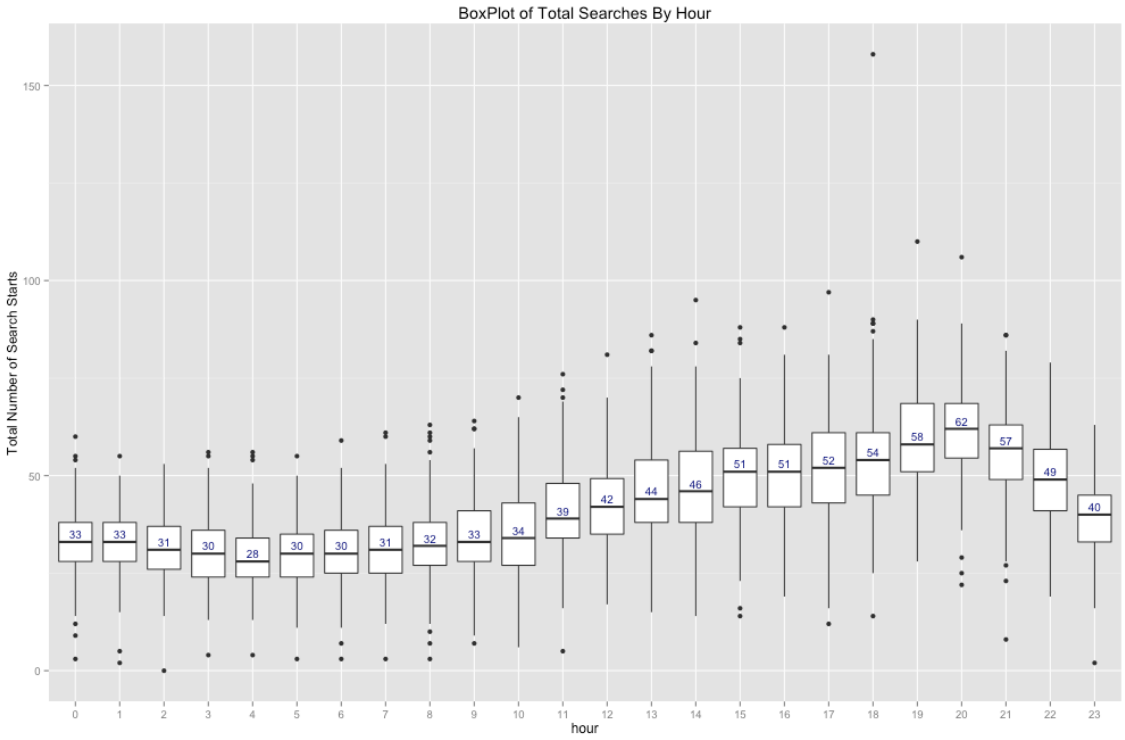
The data set contains 1 million observations in total, which describes user search behavior from November 2014 to May 2015. There are three variables in the data set: timestamp which describes when the action happened, event_action variables indicating the type of user behavior and the time it takes to display results. As there are only 8 observations in November, it does not seem to be very representative of user search behavior. The observations happened in November are eliminated from the dataset for the following analysis.

Data Preprocessing

Five time variables are created from the timestamp. They are year, month, number of days in the month, number of days in the week and hour. Total number of started searches and total number of clicking behavior by hour are calculated. Average display time is also calculated for results actions.

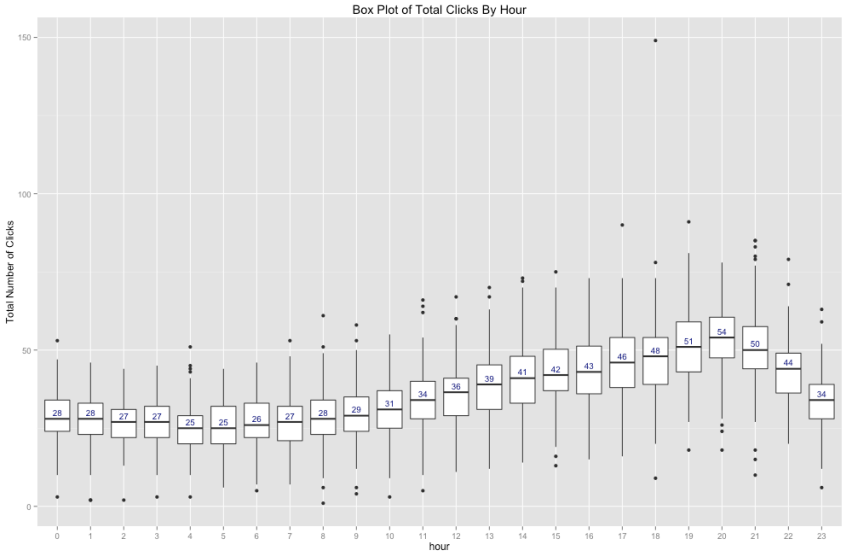
Hourly User Behavior Analysis

Graph 1 shows the boxplot of total number of searches that started in each hour. It shows that the number of searches per hour tend to be stable from midnight to early morning (12:00 am – 10:00 am). Then the number of searches began to grow hour by hour and reached to peak at 8:00 pm. After 8:00 pm, the number began to decline. 7:00 pm to 9:00 pm can be considered as the rush hour of the website.



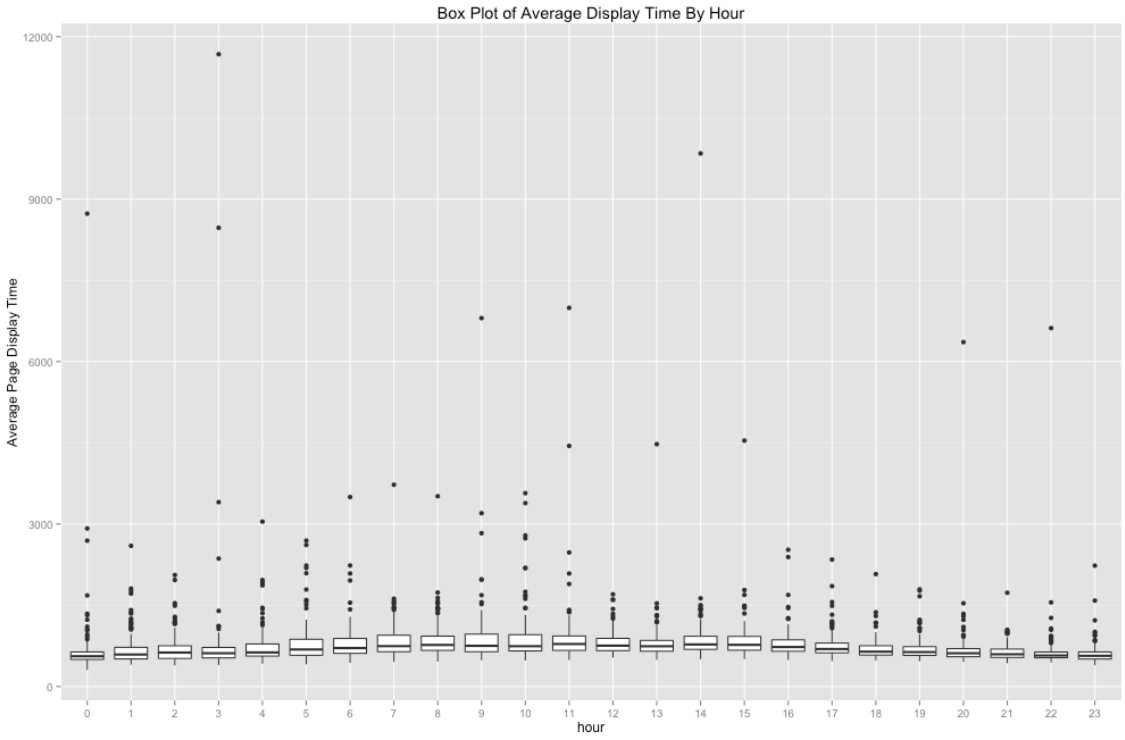
Graph 1

Graph 2 shows the boxplot of total number of clicks by hour. It shows similar trend as graph 1. It is also shows that there are some extreme points in the data set.



Graph 2

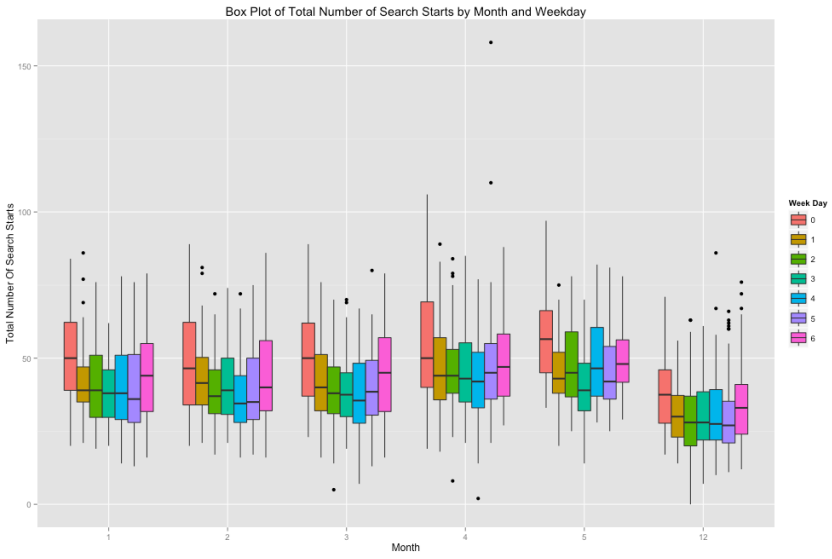
Graph 3 shows box plot of the average time it takes to display users' search result per hour. It shows that the median of average display time do not change dramatically with the time. However, there are many extreme observations and there tends to be more extreme observations in late night and early morning.



Graph 3

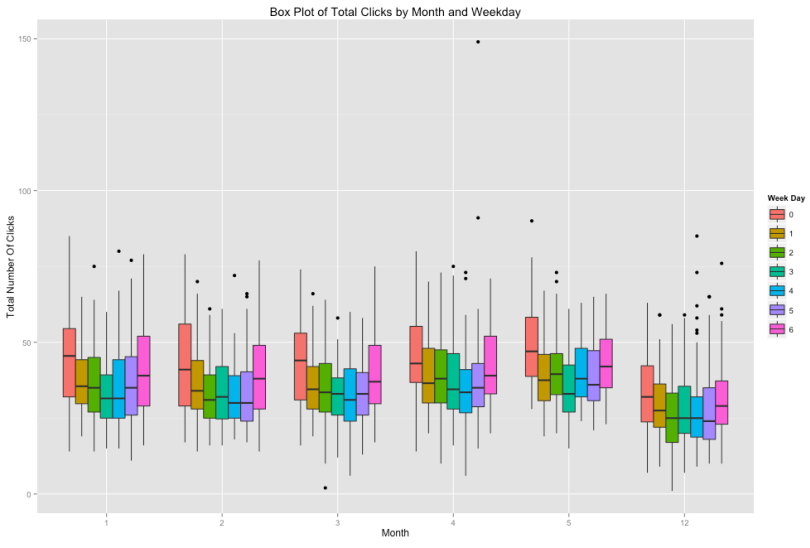
Monthly and Weekly User Behavior Analysis

Graph 4 shows the boxplot of search started per hour by month and weekday. It is easy to observe that there tends to be more searches on weekends and fewer searches in weekdays.



Graph 4

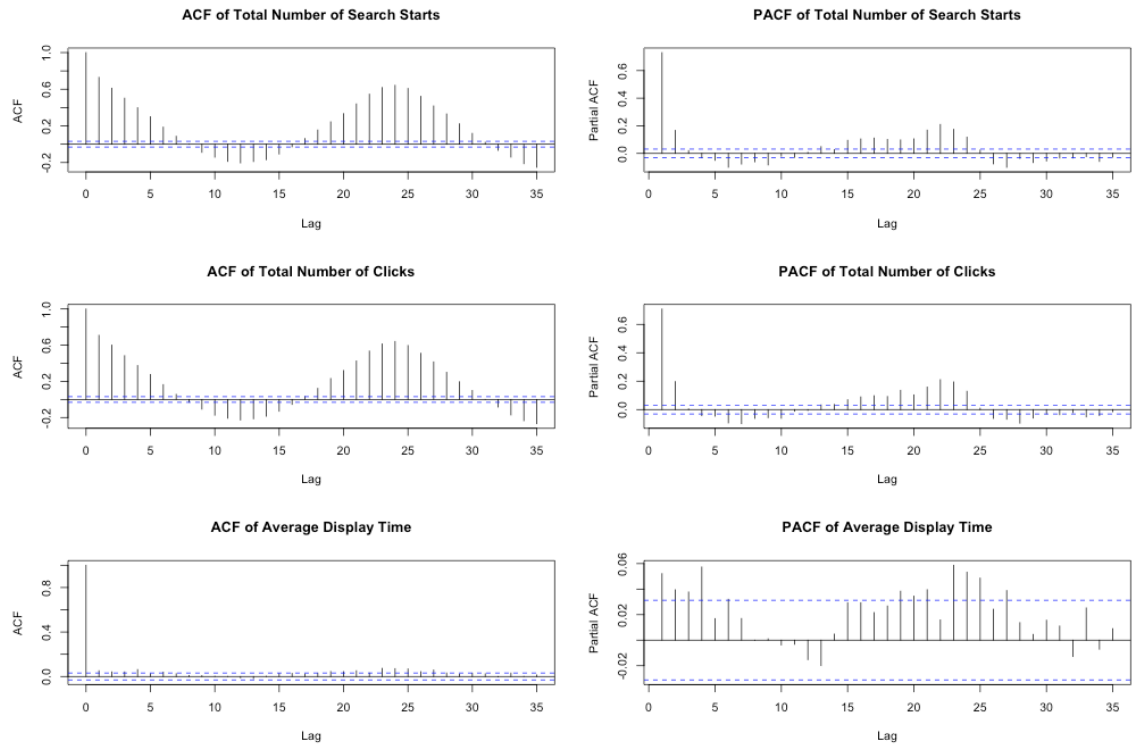
Graph 5 shows the box plot of total number of clicks per hour by month and weekday. There are also more click actions during the weekends and less on weekdays.



Graph 5

Seasonality Detection

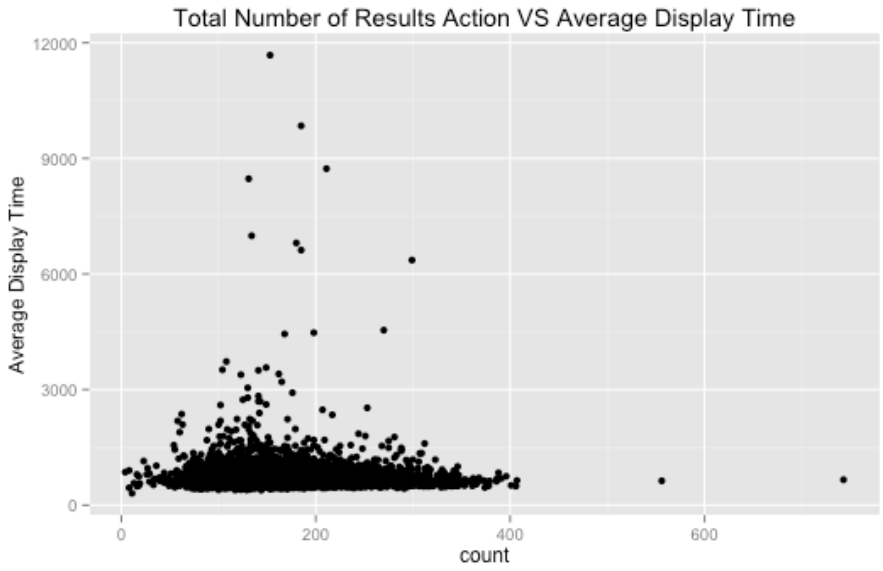
Both sample autocorrelation function (ACF) and partial sample autocorrelation function (PACF) are calculated on total number of searches started per hour, total number of clicks per hour and average display time. Graph 6 shows the results of ACF and PACF. From graph 6 we can see that there exists seasonal pattern of total number of search started and total number of clicks per hour. The PACF shows that lag 22, 23, 24 are significant for both variables. The average display time, however, do not have significant seasonal pattern.



Graph 6

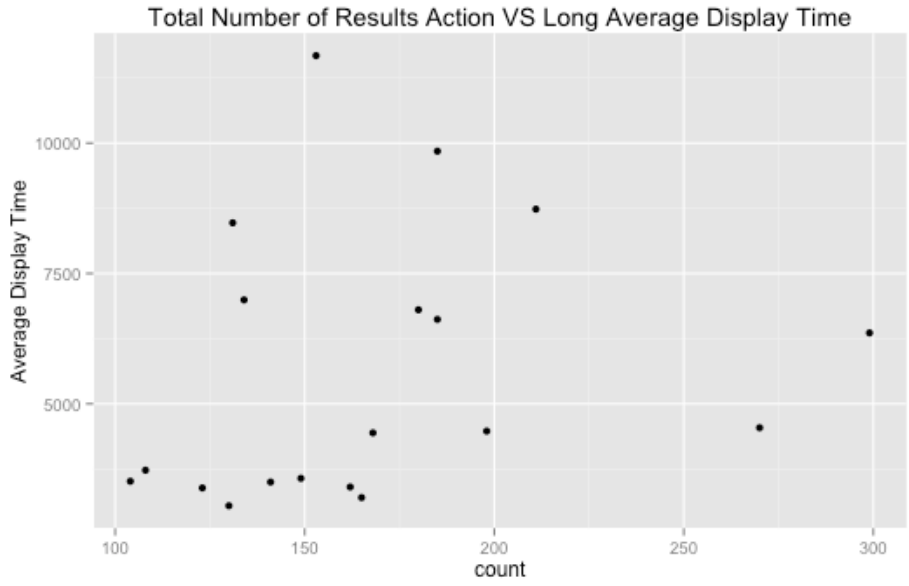
Temporal Pattern and Extreme Points Analysis

Graph 7 shows the scatterplot of average display time and total number of pages displayed by hour. It shows that average display time do not have significant linear relationship with total number of pages displayed.



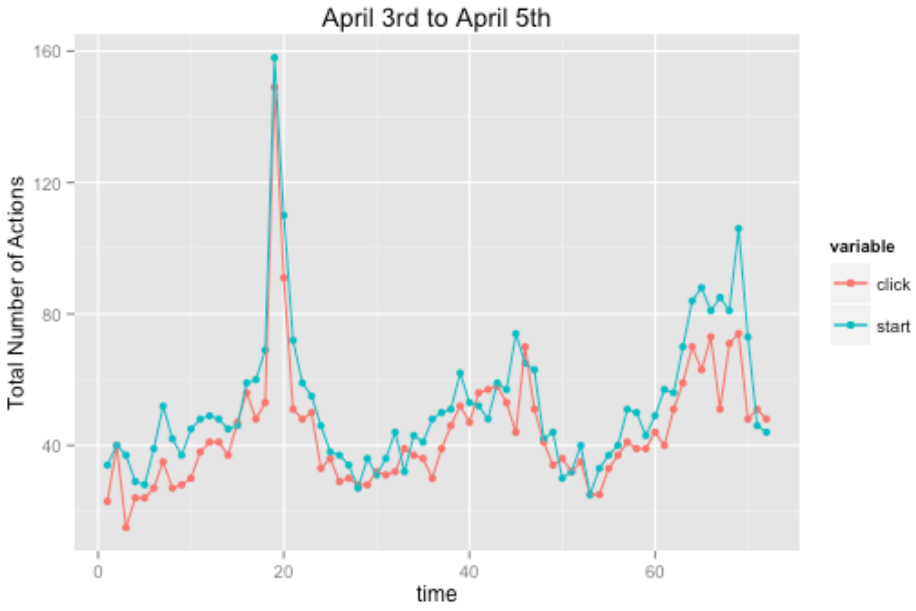
Graph 7

Graph 8 shows the scatter plot of observations with average display time longer than 3000. We can see that all observations with long average display time appeared when there are more than 100 page display requests per hour, which means they all happened during rush hour.

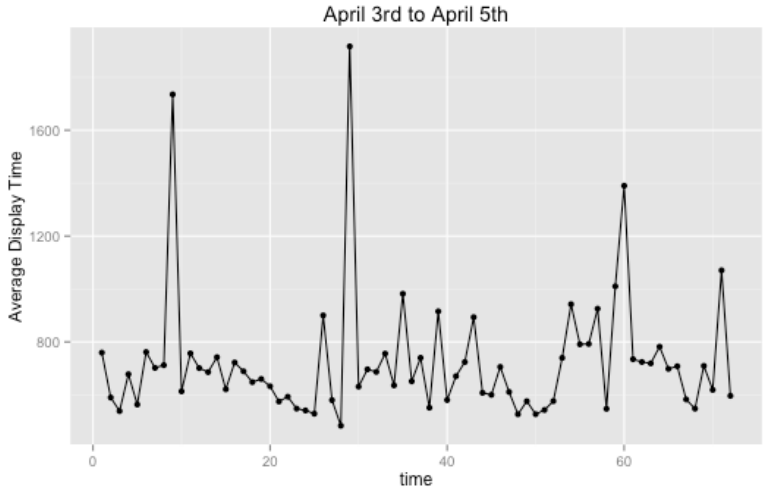


Graph 8

Graph 9 shows the observations before and after the search peak, which happened on April 3rd, 2015. It shows that there are two high peaks during these three days. Graph 10 shows the corresponding average time to display search results. It is easy to see that average page display time peaked correspondingly with the number of searches.



Graph 9



Graph 10