Author: Aditi Nair (asnair09@gmail.com)
Date: January 24th 2016

Code Challenge Submission for Bitly Data Science Internship

Methodology:
I tried to answer some natural questions I had about the dataset mostly by showing different plots that gave some insight into the answers. I used two different datasets which both initially had size one million.

I was also interested in entries with labeled 'ckw' values, since users who create accounts specifically choose these values and create their URLs, so I also generated plots for only data where 'ckw' is defined.

Assumptions:
Since the data provided was only for one hour's worth of data, and I only analyzed a subset of it, I did not do any temporal analyses because I assumed it was within too short a time period to observe any significant patterns.

Also I assumed 'ckw' corresponds to the custom link suffixes you can provide when you have your own account.

Data Description:
Both the general dataset and the 'ckw' dataset initially contained one million entries. This was partly so that the code would run more quickly and partly so that some of the plots were easier to visualize.

After dropping duplicates the general dataset contained 937,805 entries and the 'ckw' dataset only contained 9,132 entries.

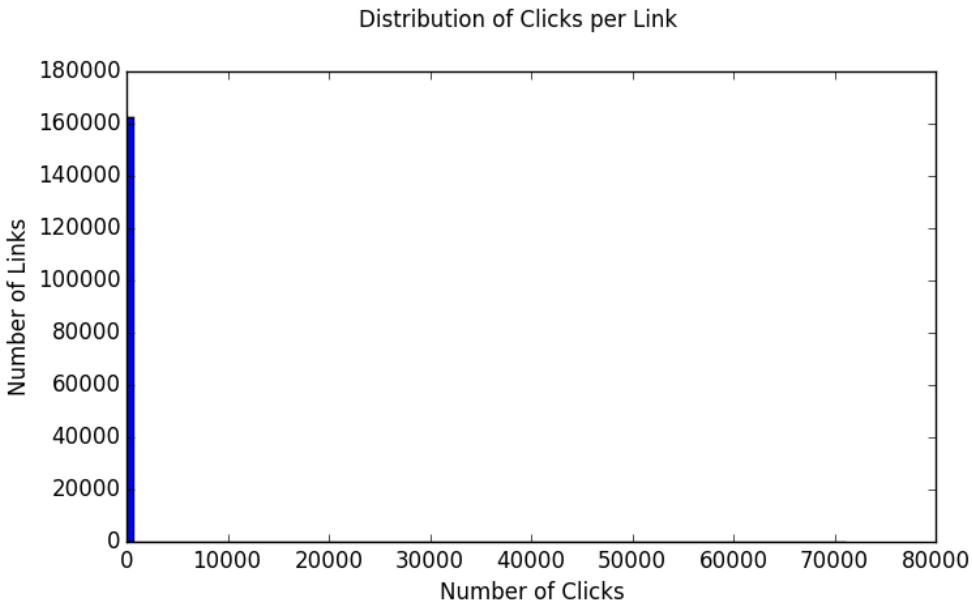
Only about 0.31% of the general dataset contained values in the 'ckw' column. All of the entries in the 'ckw' dataset contained values in the 'ckw' column.

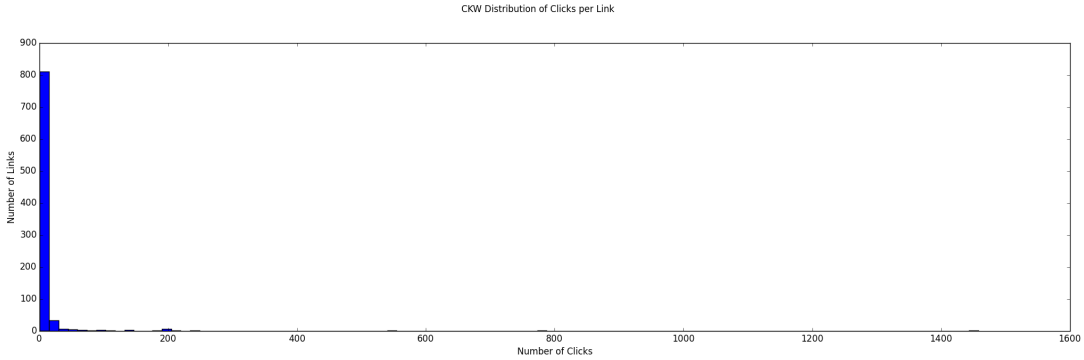
Notes:
You can see that actual png files (easier to look at) in the plots/previously_generated directory.

I noticed that some of the plots – the ones involving 'group by' on the 'u', 'h', or 'u' and 'h' columns looked very similar. I was concerned that I had misunderstood the data fields. However, after creating a quick function to verify that there was in fact a difference between these different group by actions (please see 'group_by_analysis.py'), I realized that the similarity of the plots simply reflected the nature of the data.

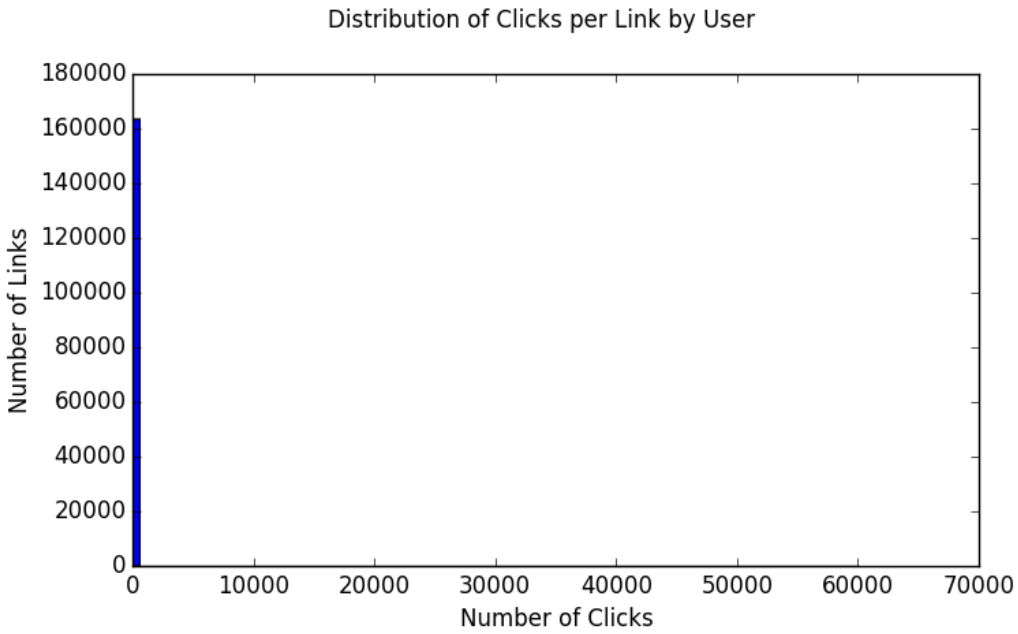
Plots:

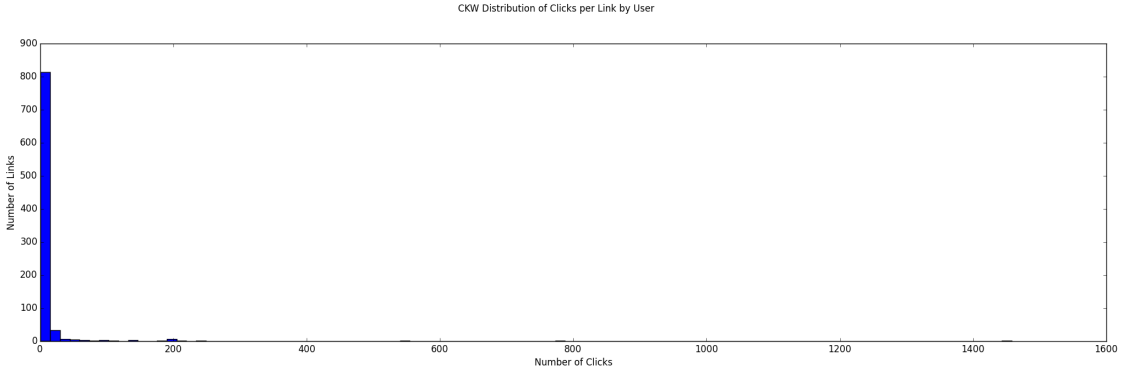
Motivation: How many clicks do links tend to get?



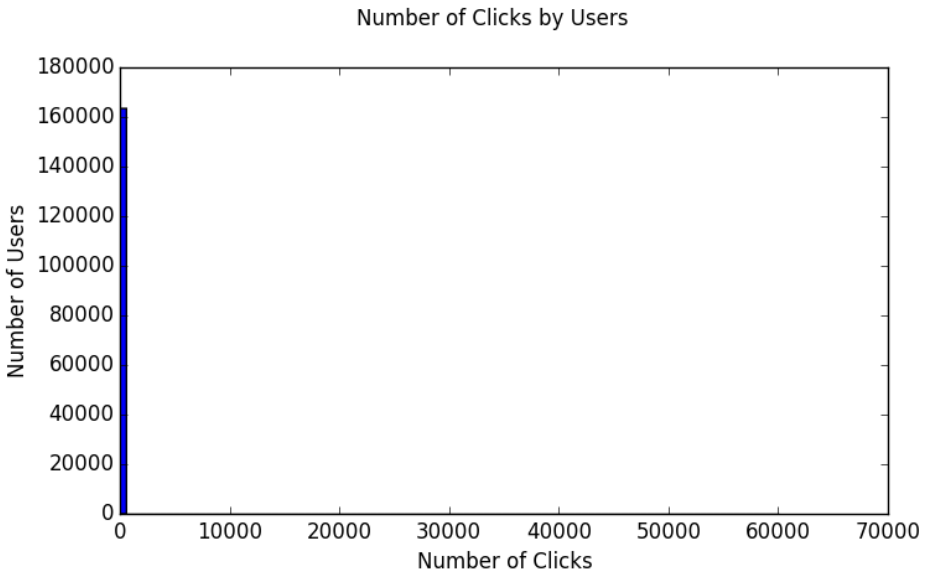


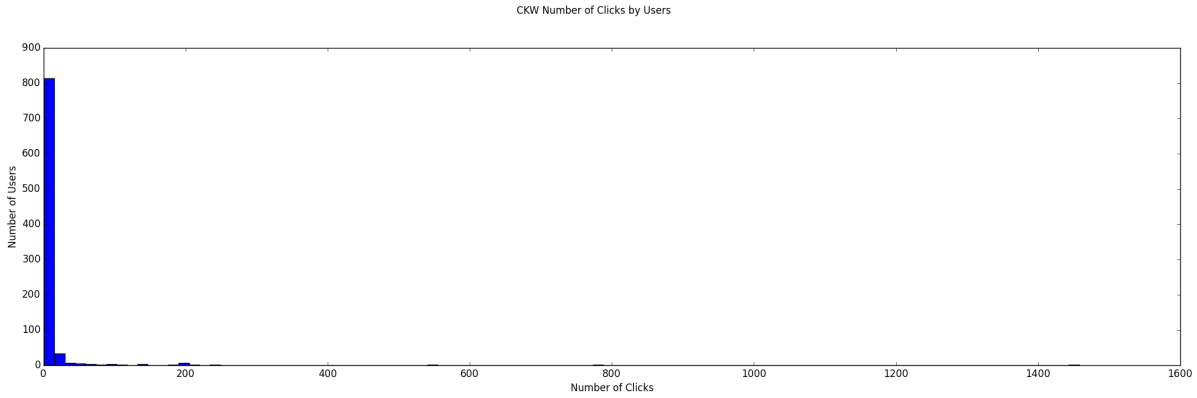
Motivation: How many times does a user click a single link?



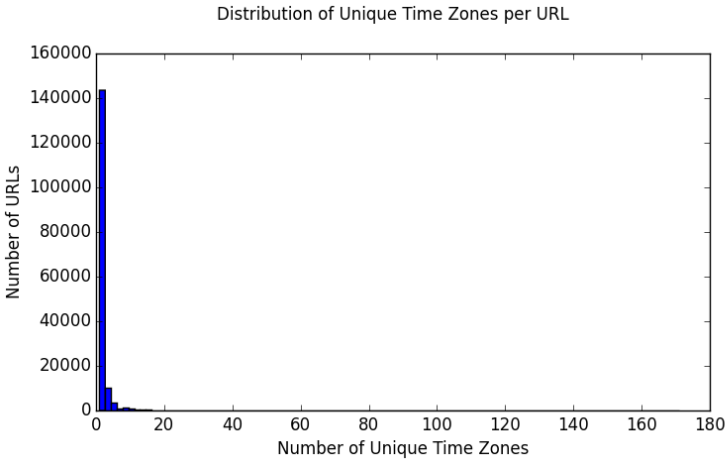


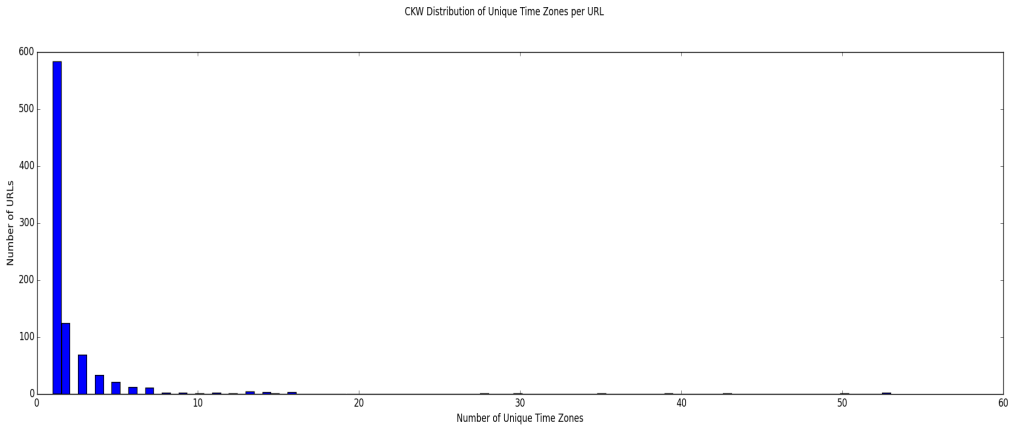
Motivation: How many times does a user click any link?



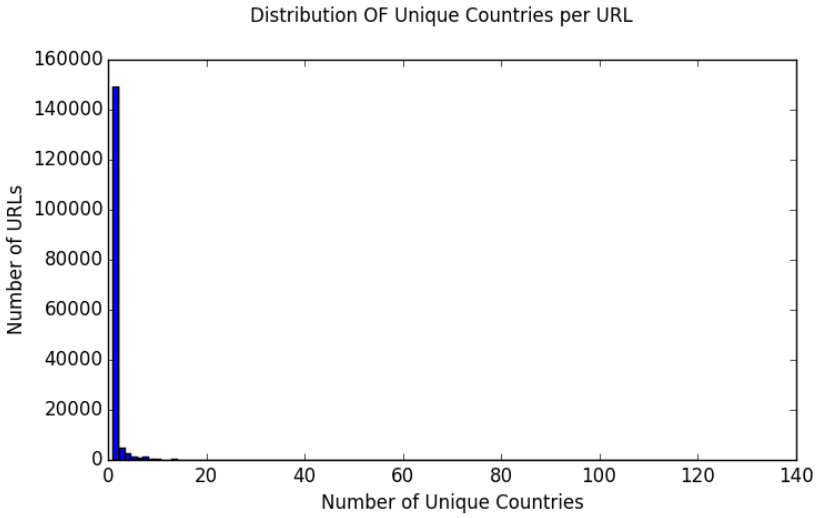


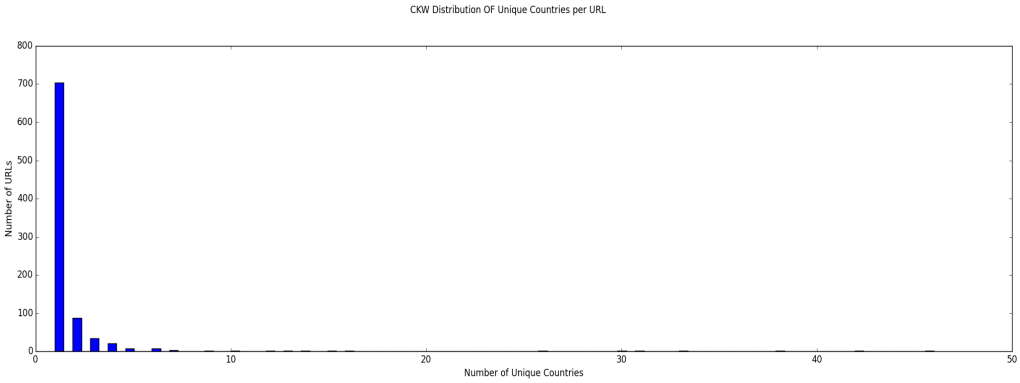
How many different time zones do links tend to be clicked in?





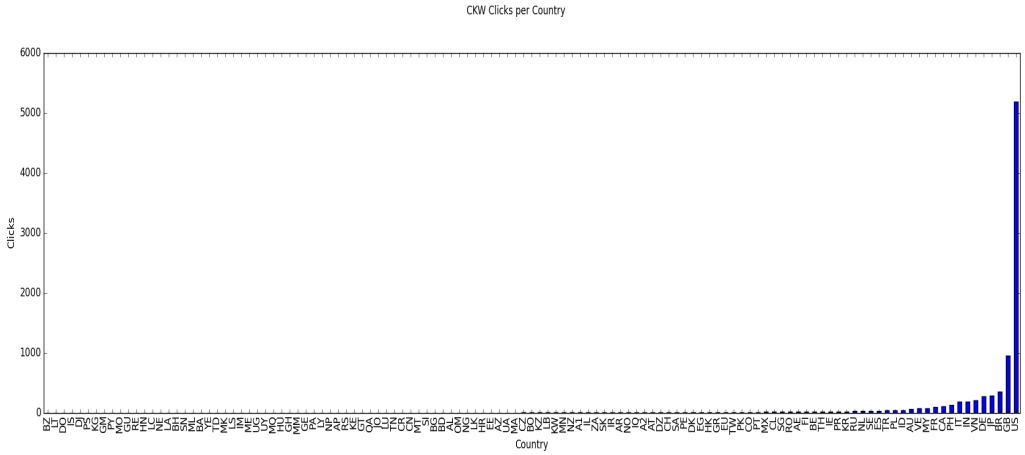
How many different countries do links tend to be clicked in?





What countries have the most clicks?

(I left out the general dataset here since there were so countries associated with it.)



What time zones have the most clicks?

(I left out the general dataset here since there were so many time zones associated with it.)