Data Source #1 = https://www.kaggle.com/eliasdabbas/5000-justdoit-tweets-dataset

Data Source #1

First we read in the CSV file into Pandas. Then we show the .head() which is the first 5 rows of the data. We list the column headers, the headers that start with “tweet\_” tweet related and the headers that start with “user\_” are the users who did the tweet. We split the tweet & user data into two datasets.

We clean User data. We then convert user\_id column values from number to a string data type. We move the user\_id column values to the 5th column of the dataset. We start cleaning the dataset. We split the data & time in the “user\_created\_at” column. Created two separate columns for date & time. We run a For Loop in the dataset, run .append() function & concatenate data with “+”, then we drop the column “user\_created\_at” in the dataset. We dedup the data which is dropping duplicate values from the dataset. In SQL, we put the primary key as User ID which is a unique value.

Next, we clean the Tweet data by separating data & time, create two columns for time\_created & date\_created, get tweet\_user\_id, extract tweet\_source\_url, & drop other columns.

Data Source #2

We used BeautifulSoup to webscrape the URL=<http://best-hashtags.com>, for popular just do it hashtags #. We find all related hastags & to make a dataframe & drop #justdoit since Source 1 has restrictions on this tweet & to avoid not getting an incomplete match.