Outbrain Click Prediction

Advances in Data Sciences

Final Project

**Under the guidance of Professor Srikanth Krishnamurthy**

**Team 8**

**TABLE OF CONTENT**

1. [About Outbrain](#_About_OutBrain)…………………………. ………………………………………………………………………………. 3
2. [Overall Flow](#_OVERALL_FLOW) ……………………………………………………………………………………………………………….. 4
3. [Data Pre-Processing](#_DATA_PRE-PROCESSING) ………………………………………………………………………………………………….. 5

3.1 [Data Sources](#_DATA_SOURCES) ………………………………………………………………………………………………………. 5

3.2 [Data Scraping](#_DATA_SCRAPING)  ……………………………………………………………………………………………………. 6

1. [Data Exploration](#_DATA_EXPLORATION) …………………………………………………………………………………………………………. 7

4.1 [Analyis on Python](#_Analyis_on_Python)…………………………………………………………………………………………………

4.2 [Analysis in Tableau](#_Analysis_in_Tableau)……………………………………………………………………………………………….

1. [Team Members and Contribution](#_TEAM_MEMBERS_AND) ……………………………………………………………………………….

# About OutBrain

https://us-east-1.online.tableau.com/#/site/ads\_outbrain\_finalproject/workbooks

# OVERALL FLOW

Flowchart of the Steps Performed

# DATA PRE-PROCESSING

Dhasdjbhjbjhbadsbhbshdbjbadsbjasdbasdbhjabsdbhasbdbasbdbjbadsbbasdbbasbdbashbdhbasbdhbasbdhabsdbasdbsadbashdbhsabdhbahsdbhjabshdbashbdhbahsdbhjabshbdhasbdbhjbasdhbashbdhbahsbdhabshdbhsbh

# DATA SOURCES

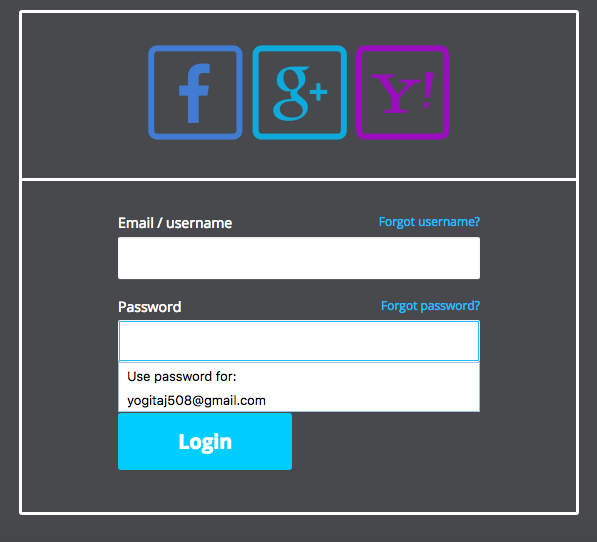
We used requests python library to scrape data from Kaggle.com. Before downloading the data, it requires to log in. Passed the credentials to the script completed the downloading of the data. There are 11 zip files that were downloaded

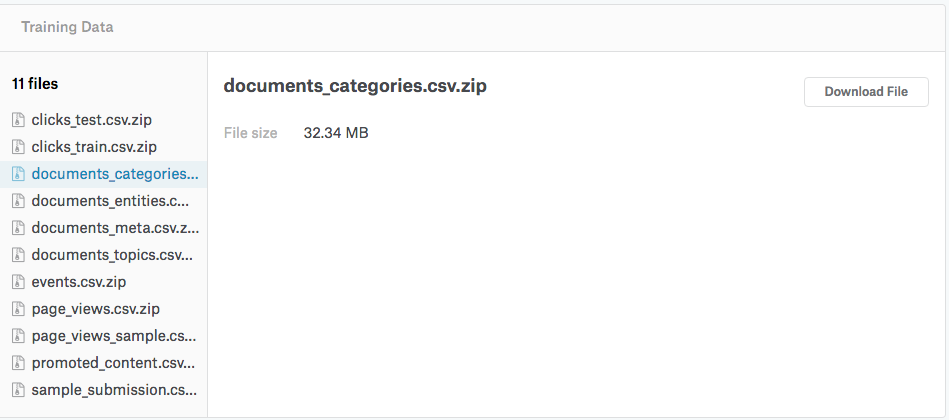
Files:

* Clicks\_test.csv.zip
* Clicks\_train.csv.zip
* Documents\_categories.csv.zip
* Documents\_entities.csv.zip
* Documents\_meta.csv.zip
* Documents\_topics.csv.zip
* Events.csv.zip
* Page\_view.csv.zip
* Page\_views\_samples.csv.zip
* Promoted\_content.csv.zip
* Sample\_submission.csv.zip

# DATA SCRAPING

**Log in Page:** Below is the screenshot of the login page that we will start scraping from and land on data download page as shown below.





PYTHON CODE

The below function in python code will download the files from the kaggle website after successful log in.

****

# DATA EXPLORATION

## Analyis on Python

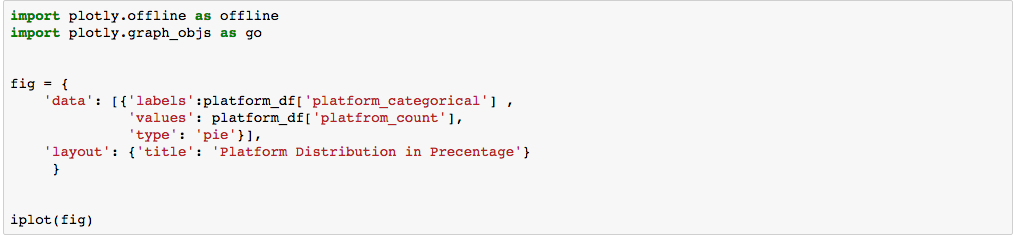
**Following are the analysis that were done on events data to look at specific insights**

1. **Analysis of percentage of clicks across different platform’s in events table:**

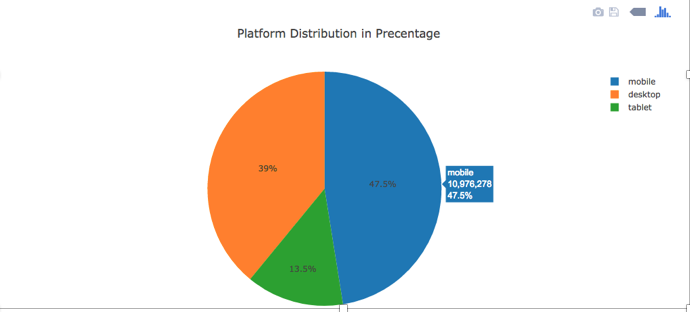
PYTHON CODE

Since platform was numeric, converted it to categorical column as below.

Screen%20Shot%202017-04-27%20at%2010.21.47%20AM.png



GRAPH



OBSERVATIONS

* Maximum percentage of clicks were made through mobile phones, followed by desktop and then tablets. Mainly because:
* App Availability: the app is available that is available on desktop is now present on model too.
* Convenience: Games or social networking apps frequently serve as a way to pass the time while on the subway commuting home or in a cab or surfing net. This directly reflects the increase in uses of mobile devices

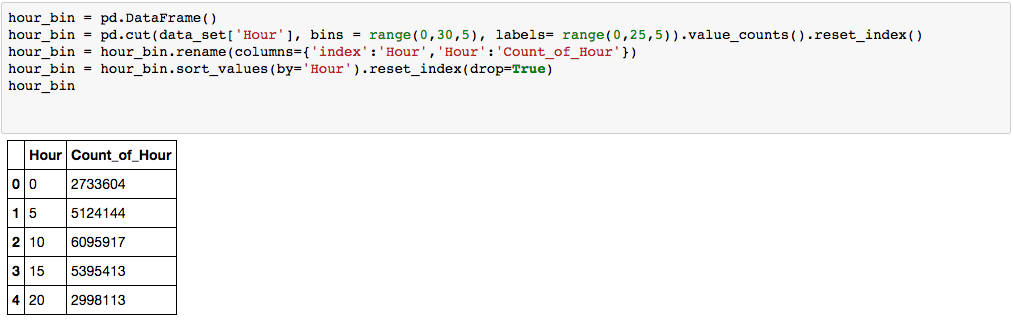
1. **Analysis of click frequency during different hours of a day in events table.**

PYTHON CODE

The events table has a column timestamp, which is a Unix timestamp. We converted that to hours and days using the standard formula and created a new hour column as shown below.

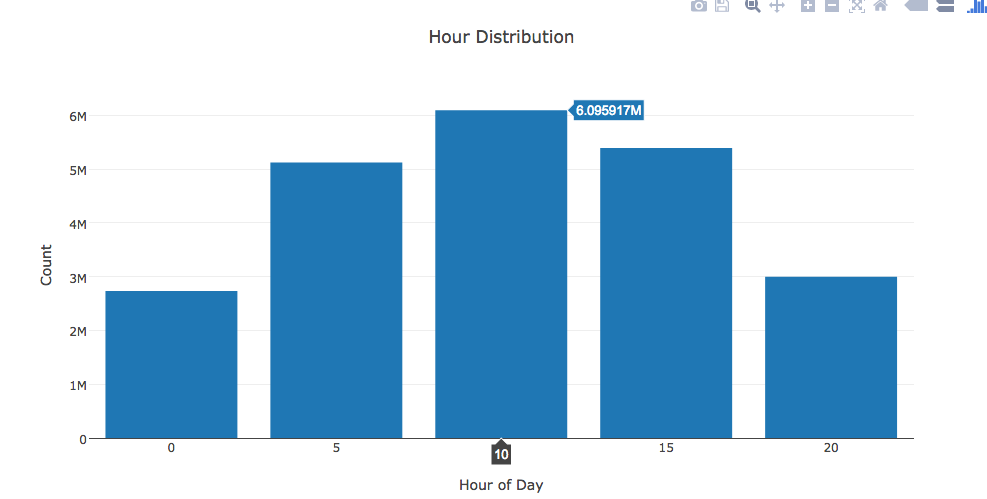
**Screen%20Shot%202017-04-25%20at%2011.41.01%20PM.png**

Next, we divide the hours into 5 bins with a difference of 5 between them and found count in each hour and added to the bin as shown in the below code.

****

****

GRAPH

****

OBSERVATIONS

* We observed that the click frequency is very high between 10-15 hrs. Around 10-11 a.m. time people are usually on their commute to work, school etc. and pass their time surfing net. Around 1-3 p.m. is usually lunch time where people get time to surf internet and that’s the frequency is high.
* The frequency in the bucket 15-20 is again when people are traveling back home and around dinner time when they have time to surf internet.

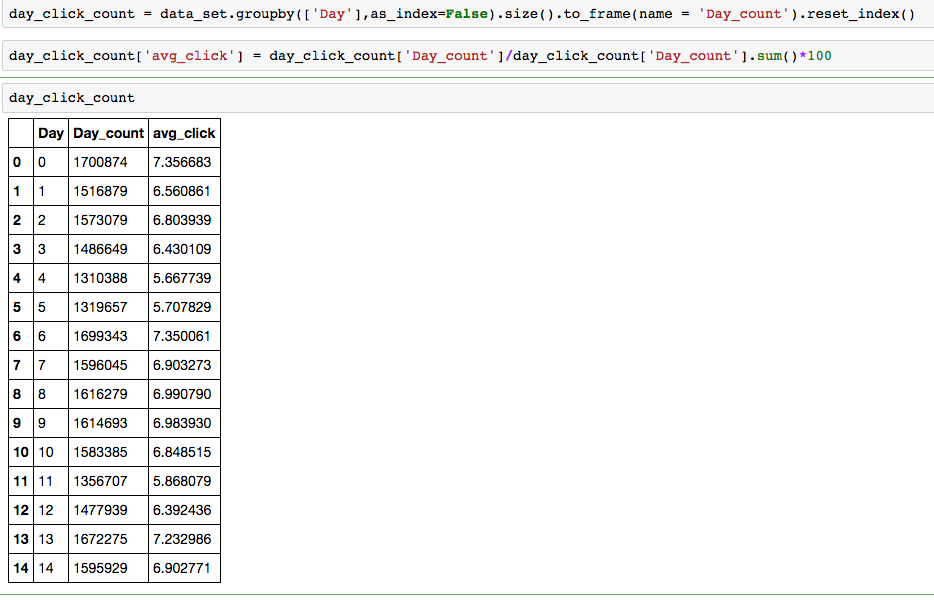
**c) Analysis of click frequency on different days in events table**

PYTHON CODE

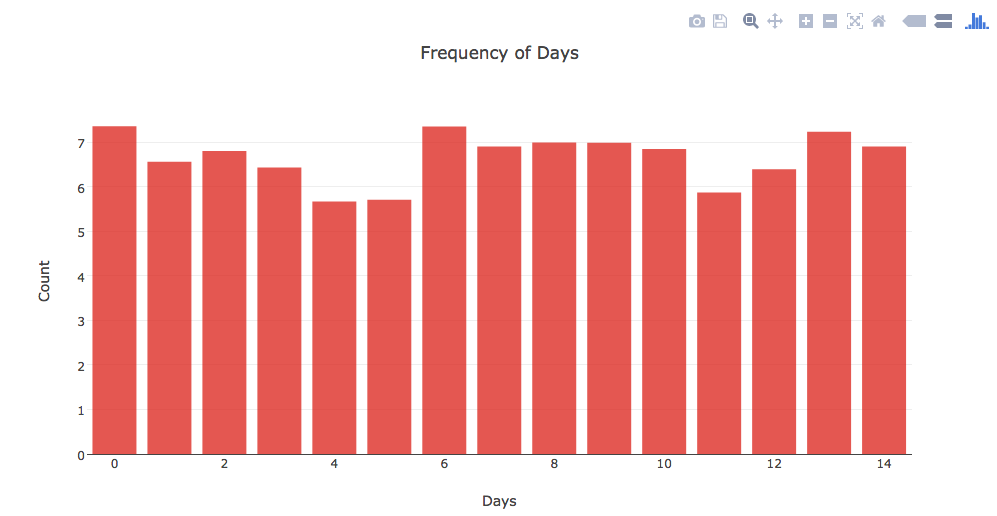
The dataset contains data for only 14 days. Hence we will see the distribution of click frequency on these 14 days. We extracted day from the timestamp column and created a new day column.

Screen%20Shot%202017-04-26%20at%2012.00.25%20AM.png

Next, we calculated the frequency of each day and using that calculated the percentage of occurrence of each day.



GRAPH



OBSERVATIONS

**Following are the analysis that were done on Page views data to look at specific insights**

Page views is basically a log of users visiting the documents.

1. **Analyze the frequency of page views across different platforms**

PYTHON CODE

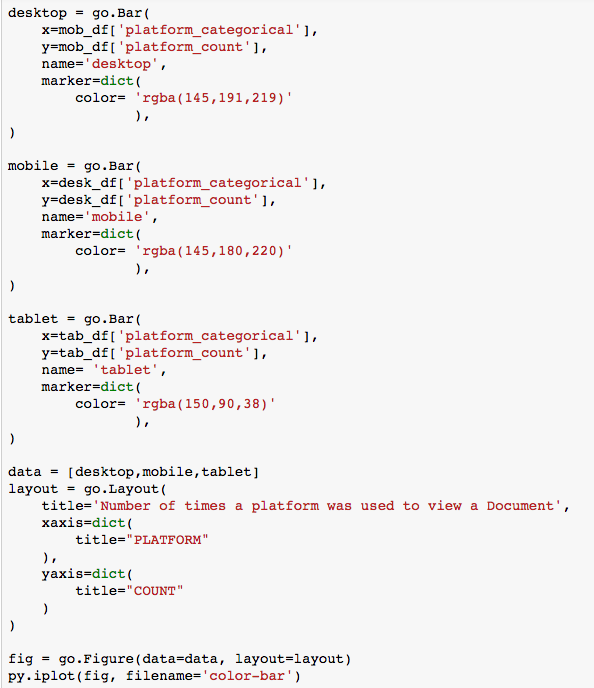
There are three platforms: desktop, mobile and tablet. The dataset contains a platform column which is

numerical so we converted into categorical column by using a map function as below.

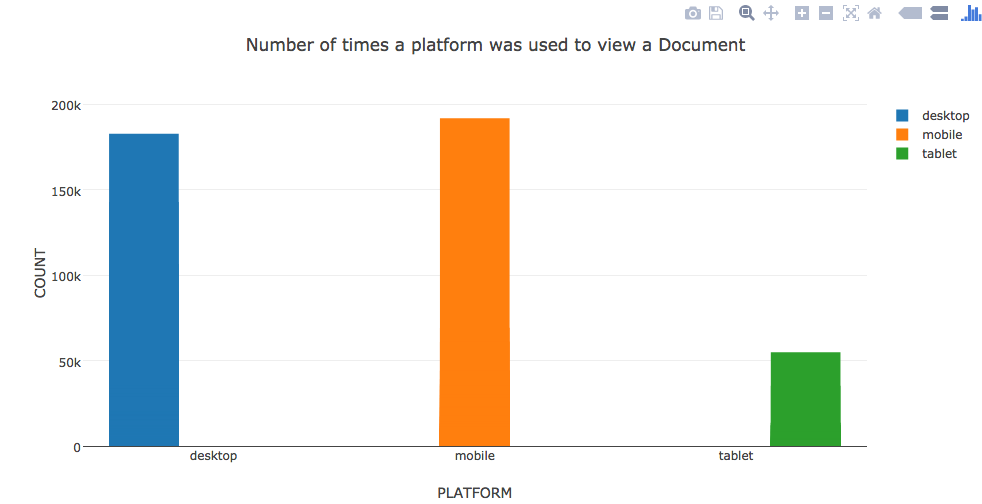
Screen%20Shot%202017-04-26%20at%2012.26.30%20AM.png

Next, we calculated the frequency of each platform group by display\_id and then created three separate data frames for each platform type to draw a plot.





GRAPH



OBSERVATIONS

* The number of views to page or a document online are highest through mobile, followed by desktop and then tablet.
* Again the main reason is people mostly go online when they are travelling and mobile being the most convenient device.

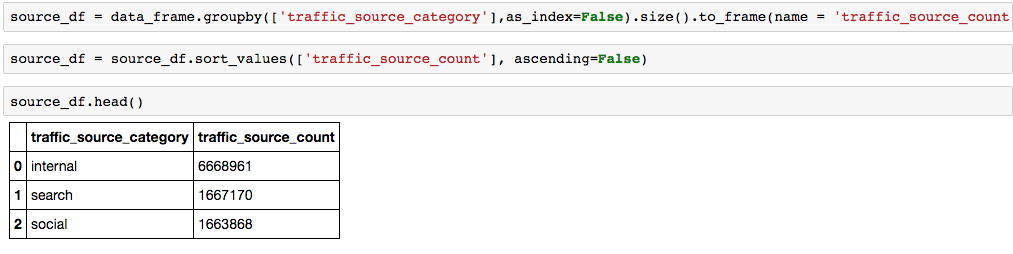
1. **Analyze how traffic sources is related to accessing a page online**.

PYTHON CODE

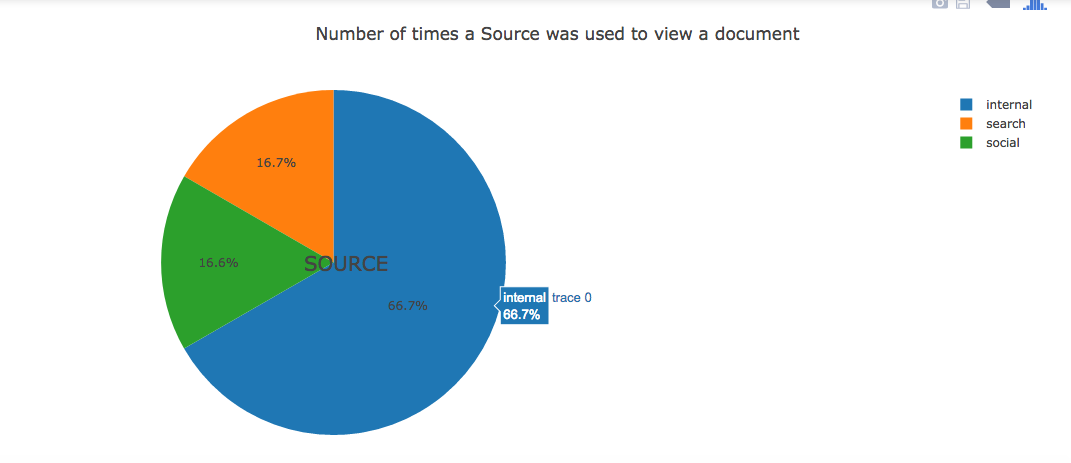
There are three types of traffic sources in the document. Since they were numerical we converted them to categorical



Next, We calculated the frequency of occurrence of each traffic source.



GRAPH



OBSERVATIONS

* Internal traffic source was about 66.7 % which is more than half of the traffic.

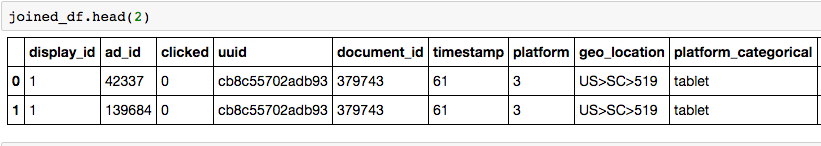
**Following are the analysis done on events and clickstrain data to look at specific insights**

1. **Analysis of number of clicks by hour of day in the events table and clicks\_train table.**

PYTHON CODE

For this analysis we did an inner join on the two datasets: events and clicks\_train on display\_id.

Screen%20Shot%202017-04-26%20at%209.56.49%20AM.png



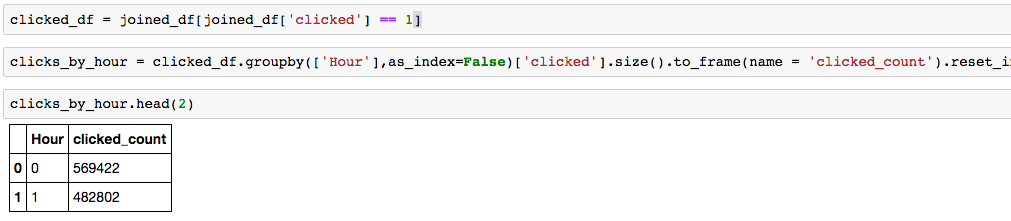
Since the timestamp is in milliseconds, we calculated day and hour using the below formula.



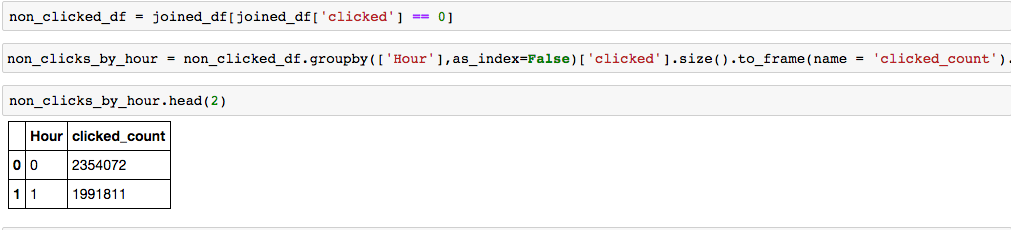
Next, created two separate data frame for click and no-clicks based on the condition:

(joined\_df[‘clicked’] == 0 or joined\_df[‘clicked’] ==1]

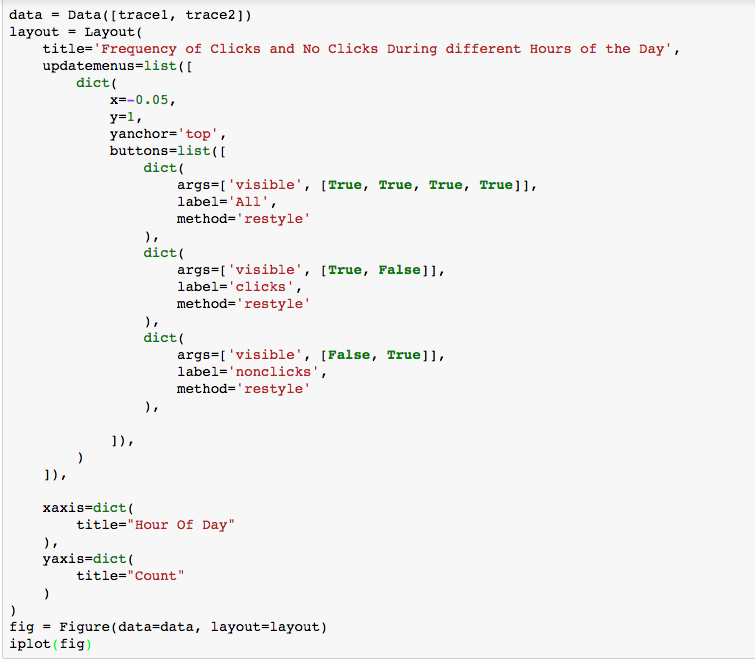
**clicked\_df**



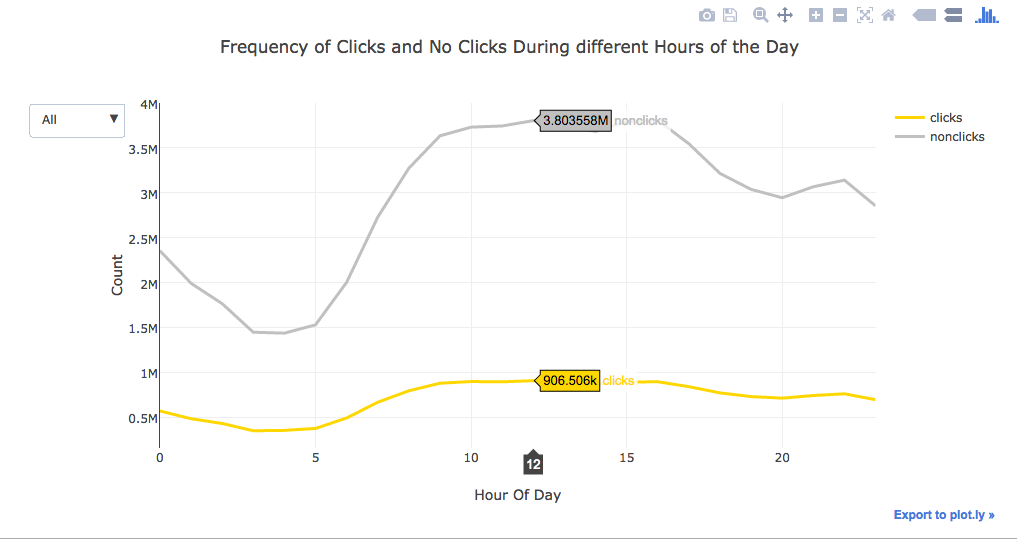
**non-clicked\_df**

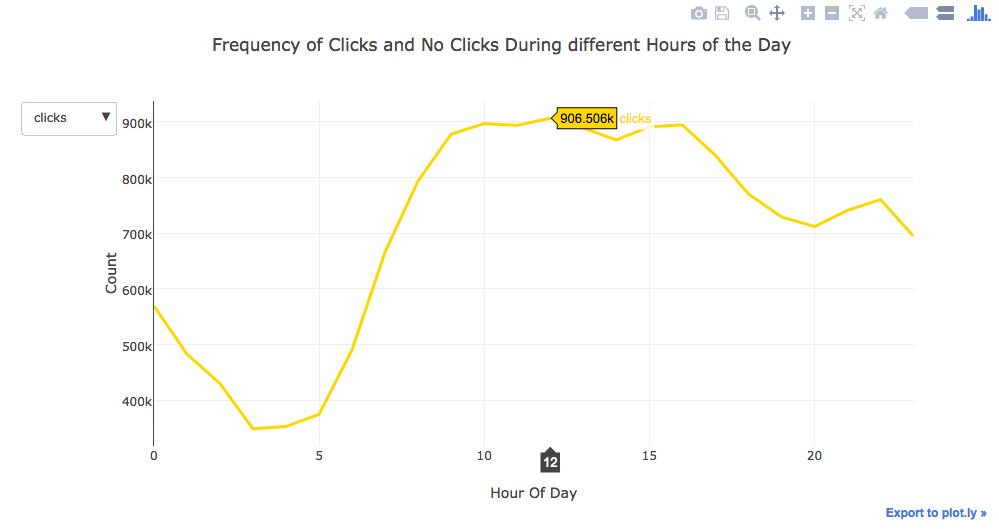


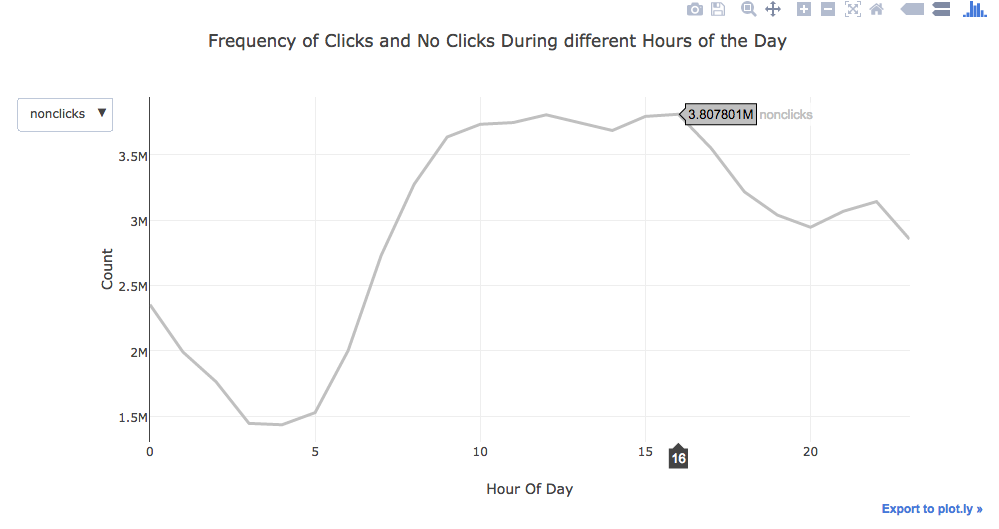




GRAPH







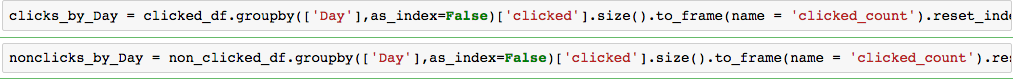
OBSERVATIONS

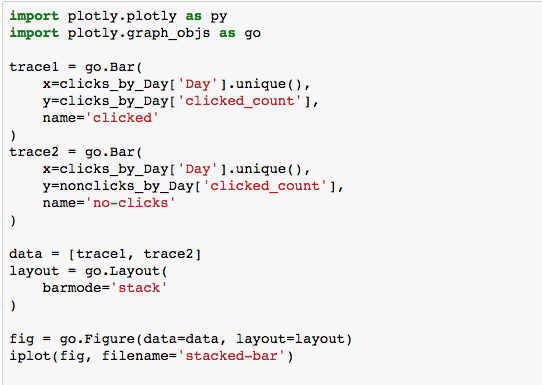
* The number of clicks increase from 9.am. and reaches highest at around 10 a.m. and remains high till 12p.m. and then gradually decreases till it hits 3.p.m. It again rises from 3p.m.
* The least number of clicks is again

1. **Number of Clicks and no-Clicks on Different Days**

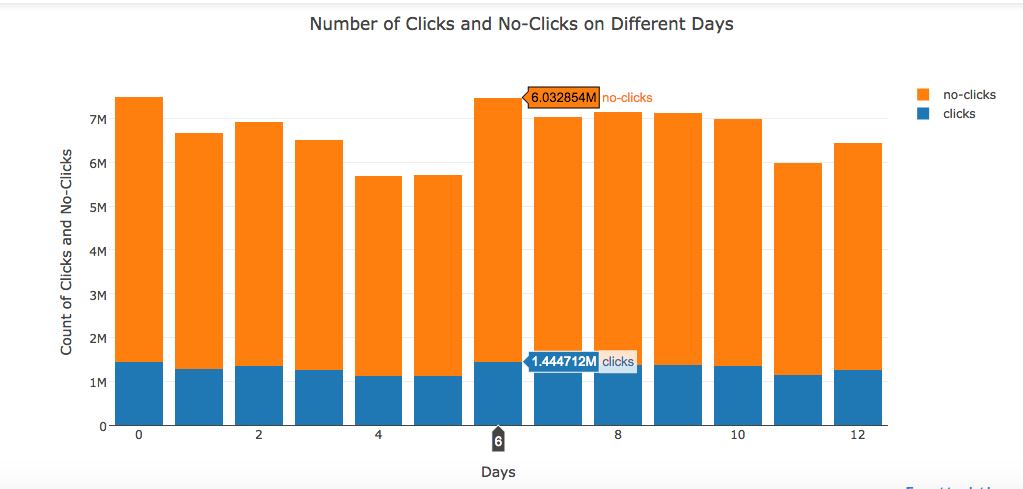
PYTHON CODE

Calculated the number of clicks and no-clicks grouping by day and created two separate data frames as shown below.





GRAPH



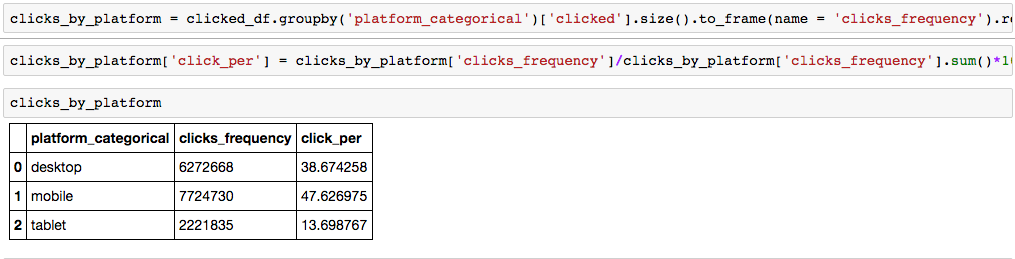
OBSERVATIONS

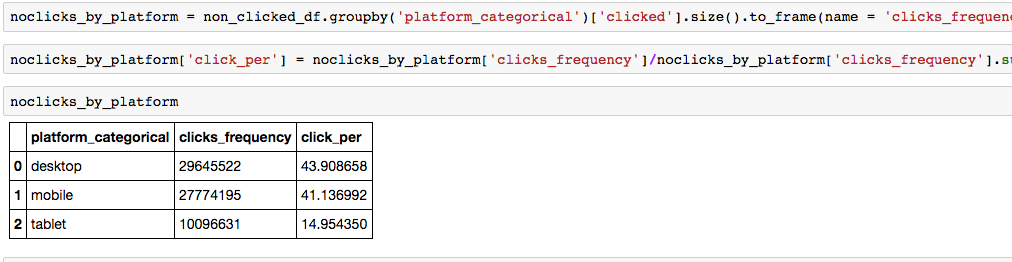
* The number of no clicks on the day 6th ranks the highest.
* The number of clicks remains almost constant during all the 14 days of the data.

1. **Number of Clicks and no-clicks by different Platforms**

PYTHON CODE

In the previous analysis we created two new data frames: clicked\_df and non\_clicked\_df. Using these data frames we calculated number of clicks and no-clicks based on platforms, grouping them platform\_categorical column. Code as below.





BAR CHART PLOT

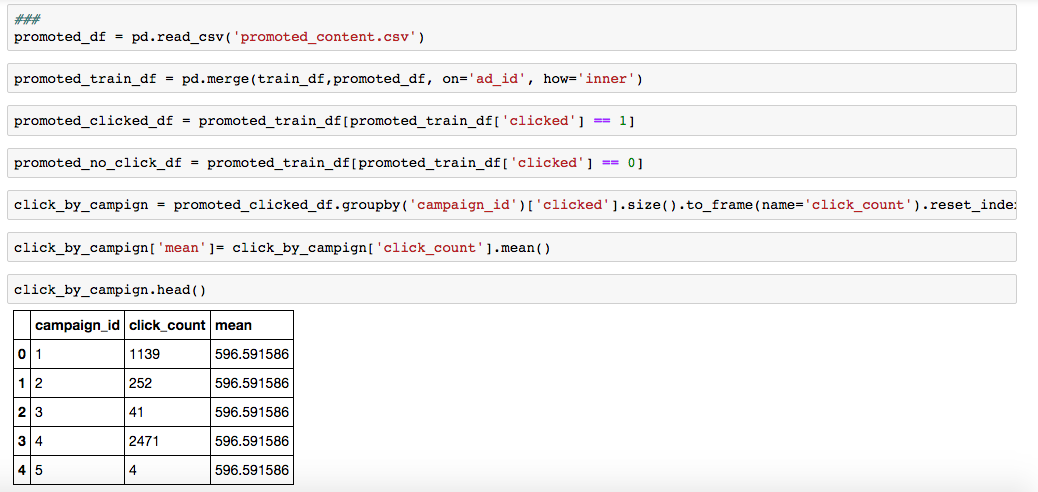
OBSERVATIONS

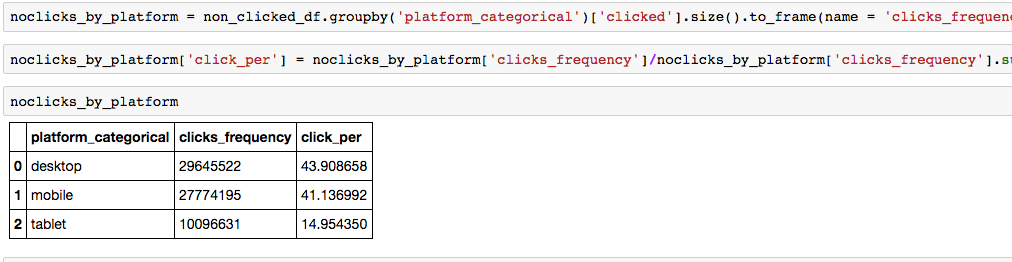
1. **Analyze the Frequency of Clicks for various Ads**

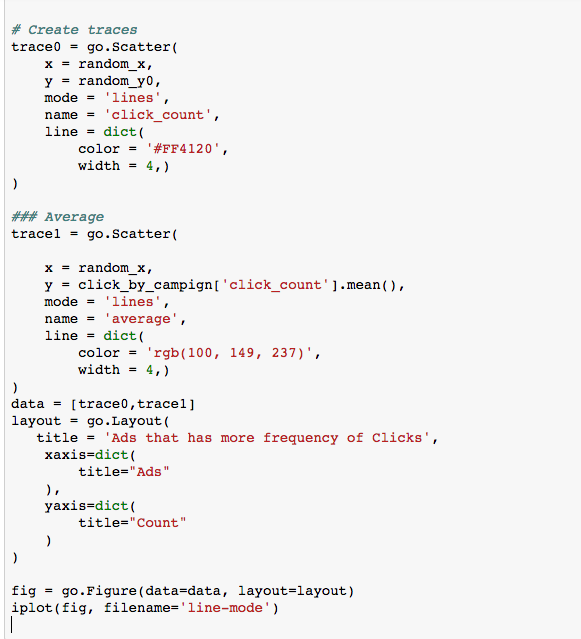
PYTHON CODE

In order to analyze the frequency of clicks across different ads in the dataset we did an inner join promoted\_content and clicks\_train files on ad\_id as below.

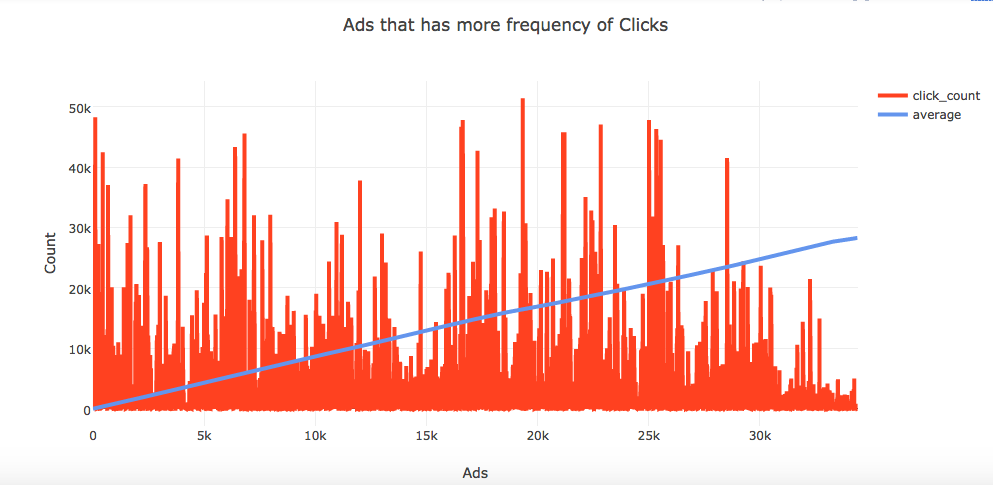
Created two different data frames for promoted content, and one with clicks and one with no clicks.

****

****

****

GRAPH



OBSERVATIONS

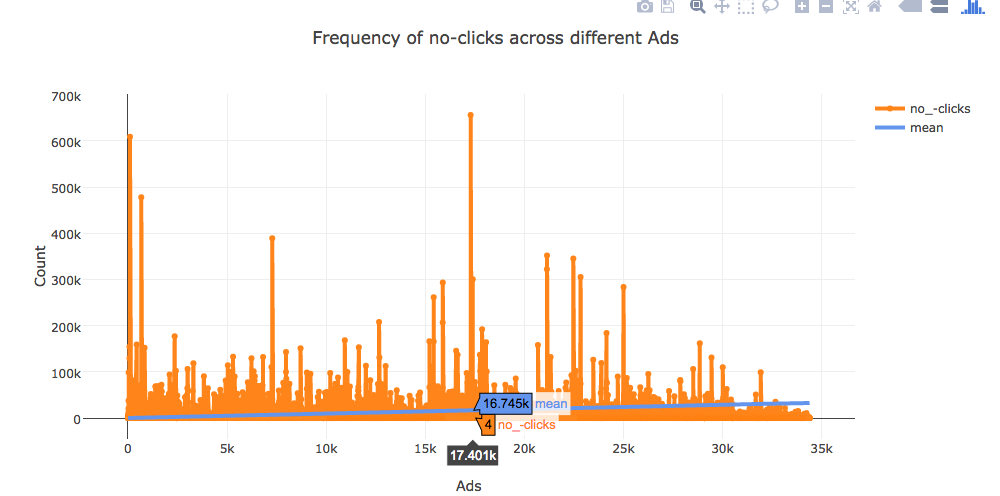
1. **Find number of no-clicks for each campaign id**

PYTHON CODE





BAR CHART PLOT



OBSERVATIONS

## Analysis in Tableau

# TEAM MEMBERS AND CONTRIBUTION

1. Antriksh Sharma

NUID :

Linkedin Profile URL :

1. Sweta Bajaj

NUID :

Linked Profile URL :

1. Yogita Jain

NUID :

Linked Profile URL :

CONTRIBUTION CHART