Exploratory Data Analysis

Dataset: Marketing campaign data from Oct 2019 to July 2020.

This dataset contains time-series data of marketing campaign of Facebook and Google campaigns including campaign platforms, audience types, spends, clicks, link clicks, impressions, device, etc.

The dataset contains 15 columns. During the analysis, it was found that three columns contain null values and other columns contains special characters instead of values. The columns with special characters were replaced by nan values. Finally, the columns are as follows:

product : 0

phase : 0

campaign_platform : 0

campaign_type : 0

communication_medium : 0

subchannel : 0

audience_type :15101

creative_type :15101

creative_name : 15101

device : 0

age : 0

spends : 0

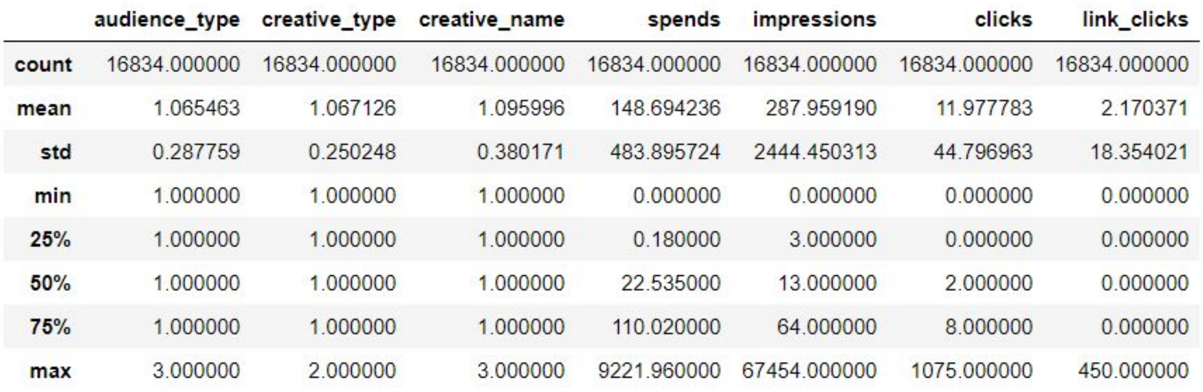
impressions : 0

clicks : 0

link_clicks : 546

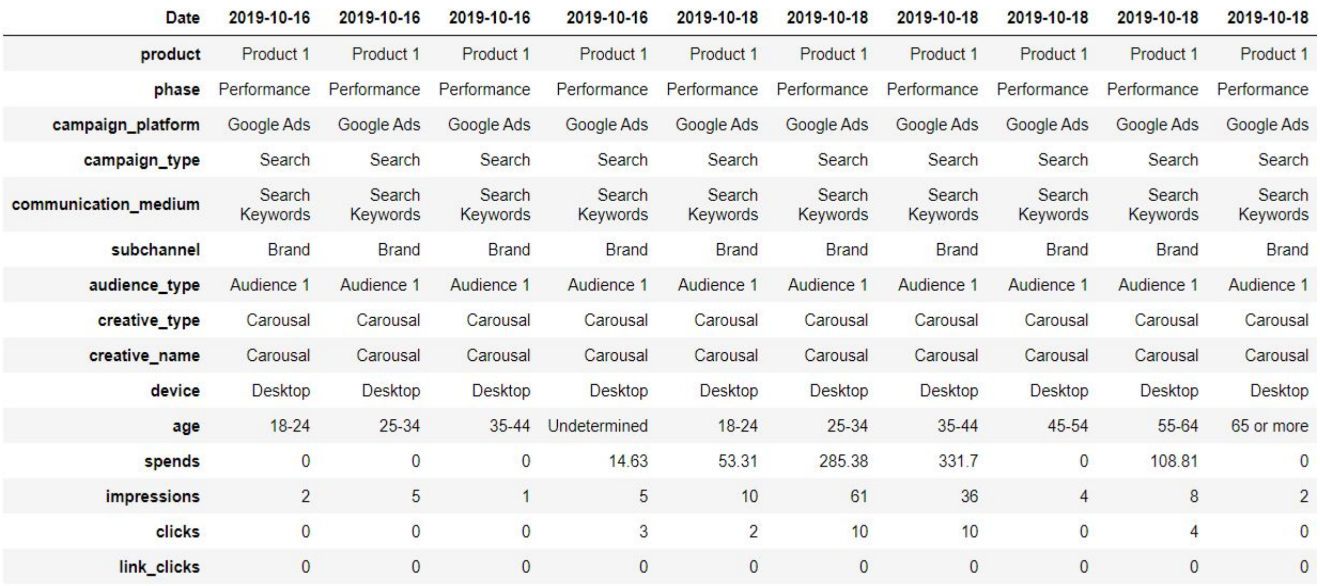
These columns with null values were imputed using KNNImputer from the scikit-learn library.

The statistical analysis of the data frame given by the describe method is:



	audience_type	creative_type	creative_name	spends	impressions	clicks	link_clicks
count	16834.000000	16834.000000	16834.000000	16834.000000	16834.000000	16834.000000	16834.000000
mean	1.065463	1.067126	1.095996	148.694236	287.959190	11.977783	2.170371
std	0.287759	0.250248	0.380171	483.895724	2444.450313	44.796963	18.354021
min	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	1.000000	1.000000	0.180000	3.000000	0.000000	0.000000
50%	1.000000	1.000000	1.000000	22.535000	13.000000	2.000000	0.000000
75%	1.000000	1.000000	1.000000	110.020000	64.000000	8.000000	0.000000
max	3.000000	2.000000	3.000000	9221.960000	67454.000000	1075.000000	450.000000

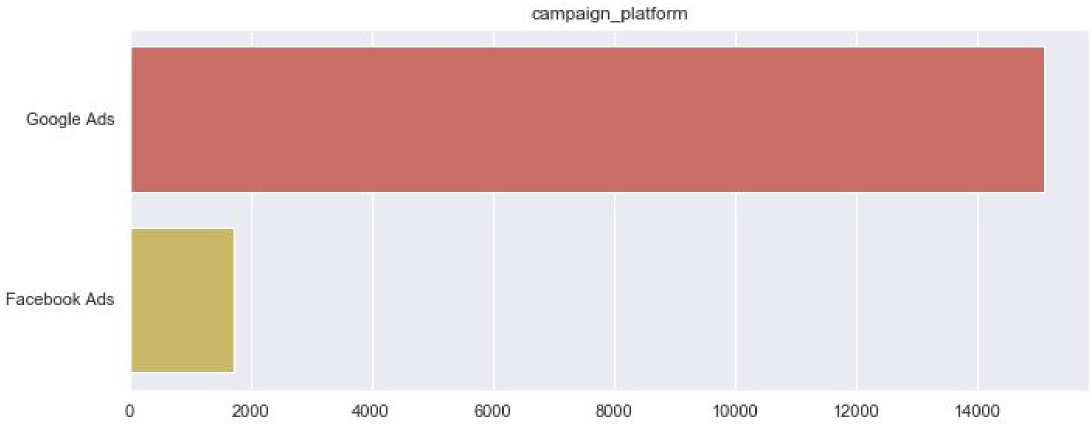
The sample data after filling in null values looks like the following:



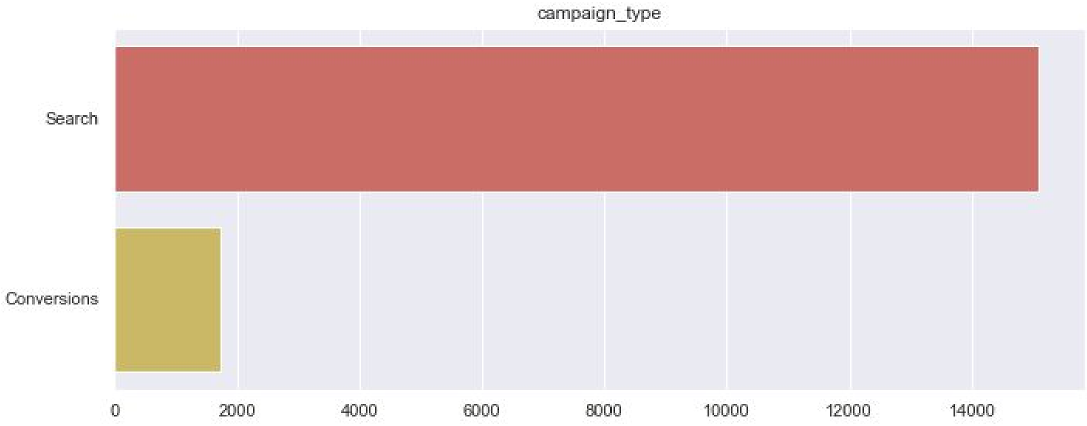
Date	2019-10-16	2019-10-16	2019-10-16	2019-10-16	2019-10-18	2019-10-18	2019-10-18	2019-10-18	2019-10-18	2019-10-18
product	Product 1	Product 1	Product 1	Product 1	Product 1	Product 1	Product 1	Product 1	Product 1	Product 1
phase	Performance	Performance	Performance	Performance	Performance	Performance	Performance	Performance	Performance	Performance
campaign_platform	Google Ads	Google Ads	Google Ads	Google Ads	Google Ads	Google Ads	Google Ads	Google Ads	Google Ads	Google Ads
campaign_type	Search	Search	Search	Search	Search	Search	Search	Search	Search	Search
communication_medium	Search Keywords	Search Keywords	Search Keywords	Search Keywords	Search Keywords	Search Keywords	Search Keywords	Search Keywords	Search Keywords	Search Keywords
subchannel	Brand	Brand	Brand	Brand	Brand	Brand	Brand	Brand	Brand	Brand
audience_type	Audience 1	Audience 1	Audience 1	Audience 1	Audience 1	Audience 1	Audience 1	Audience 1	Audience 1	Audience 1
creative_type	Carousal	Carousal	Carousal	Carousal	Carousal	Carousal	Carousal	Carousal	Carousal	Carousal
creative_name	Carousal	Carousal	Carousal	Carousal	Carousal	Carousal	Carousal	Carousal	Carousal	Carousal
device	Desktop	Desktop	Desktop	Desktop	Desktop	Desktop	Desktop	Desktop	Desktop	Desktop
age	18-24	25-34	35-44	Undetermined	18-24	25-34	35-44	45-54	55-64	65 or more
spends	0	0	0	14.63	53.31	285.38	331.7	0	108.81	0
impressions	2	5	1	5	10	61	36	4	8	2
clicks	0	0	0	3	2	10	10	0	4	0
link_clicks	0	0	0	0	0	0	0	0	0	0

Analysis Findings

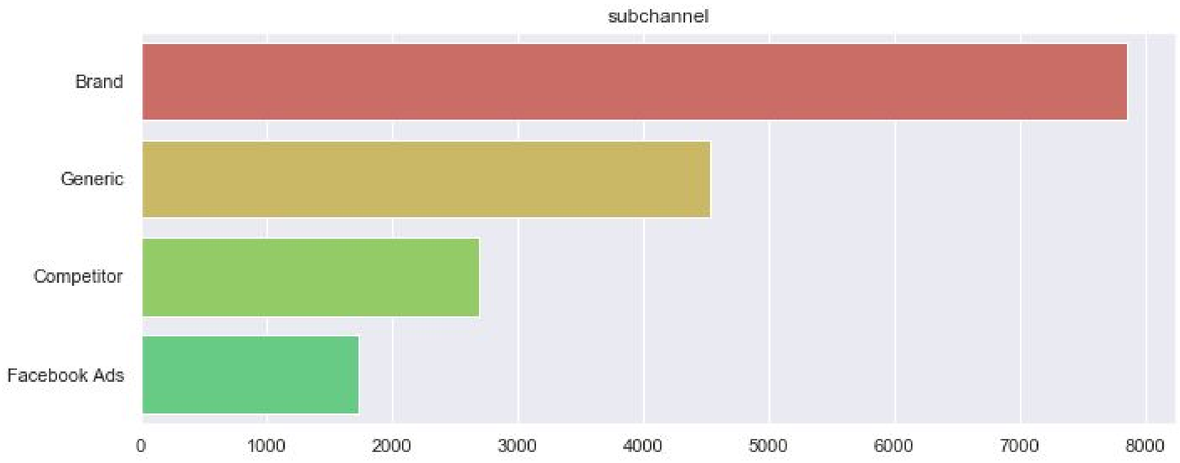
- As mentioned in the task description two campaign platforms were found with Google Ads being the majority than Facebook Ads.



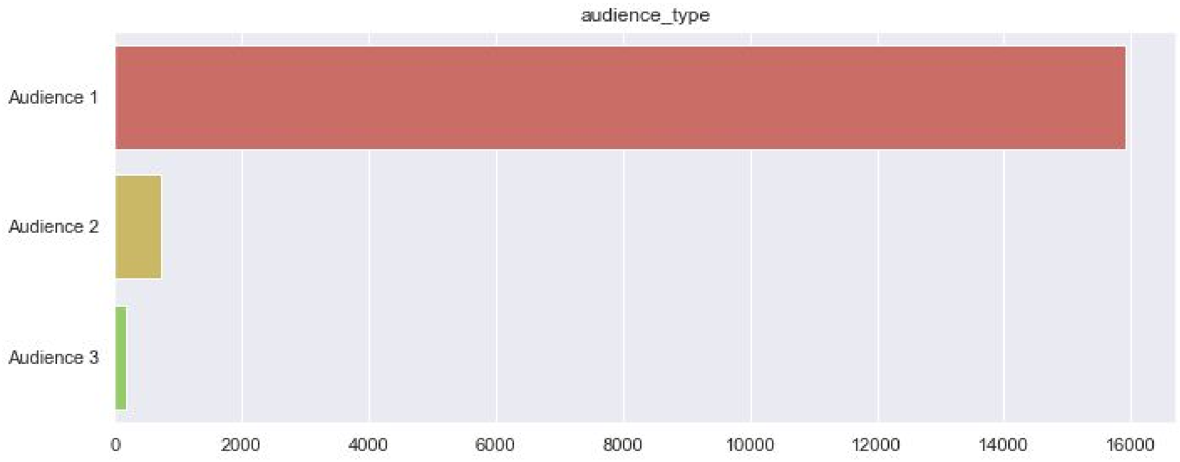
- There were two campaign types namely Search and Conversions with more than 90% Search campaign types.



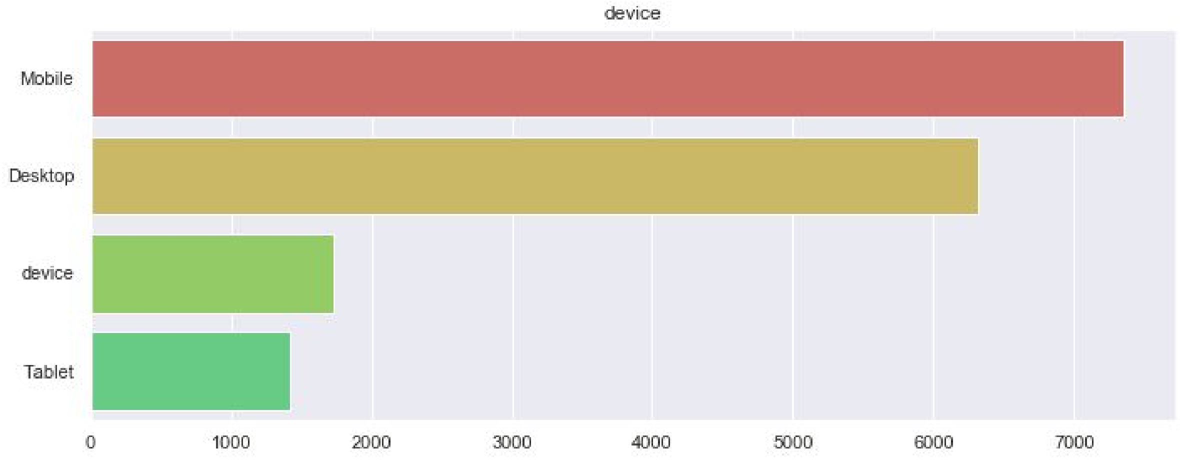
- Most popular subchannel used was found to be “ Brand”(~ 8000) followed by Generic.



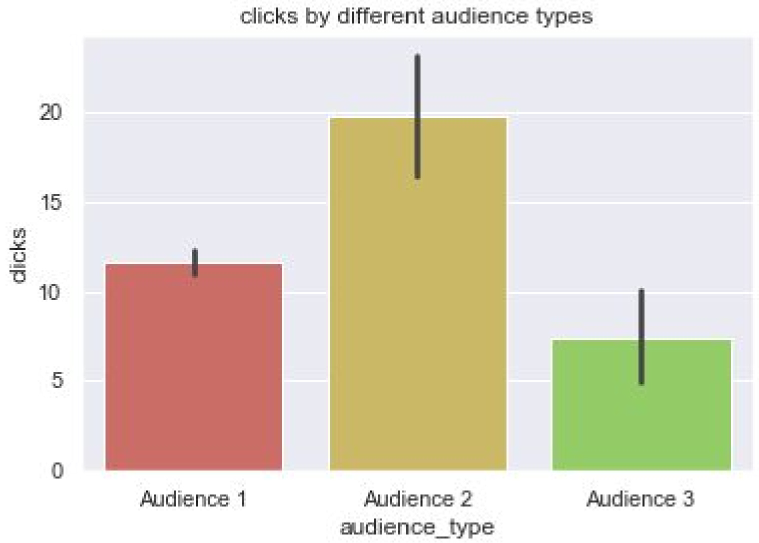
- Most common Audience type was found to be Audience 1 having a huge majority over Audience 2 and Audience 3.



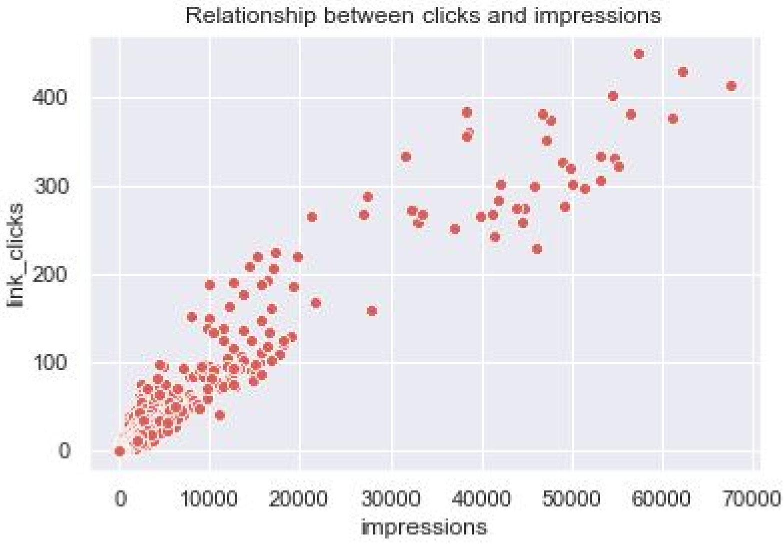
- Most commonly used devices among Mobile, Desktop, device, tablet:



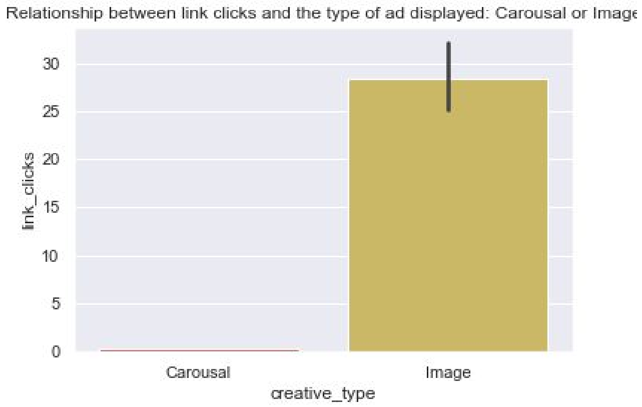
- The audience_type who clicked the most on the ad page(clicks) was found to be Audience 2.
  (Clicks are the number of clicks by a user anywhere on the page.)



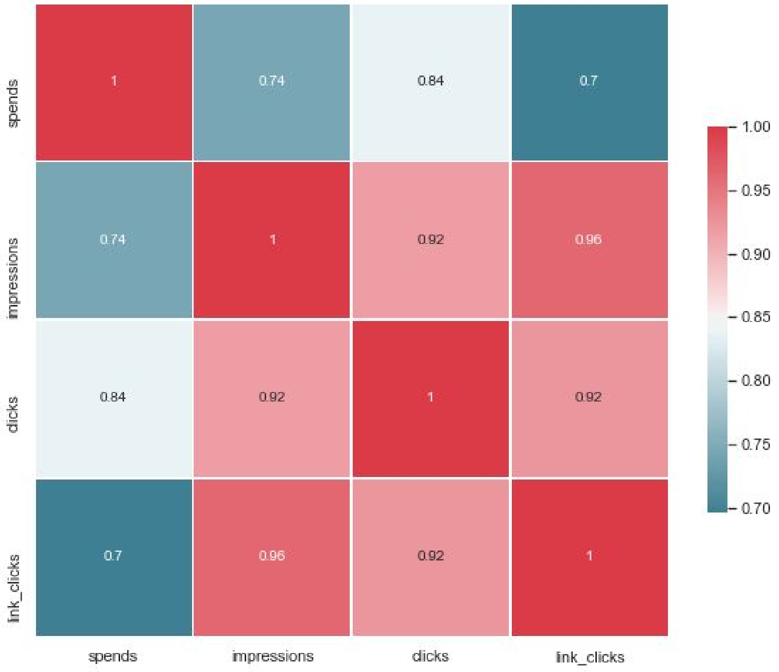
- Relationship between impressions and link_clicks:
  Impressions are the total number of times an ad is displayed on a user's screen and link_clicks are the number of times a user clicks on it.
  This shows that the more times ads were shown to the user the more the links were clicked on by the user.



- The relationship between the type of creative ad shown to the user and the number of times the user clicked on the ad link shows that more users have clicked on ads containing images than carousals.



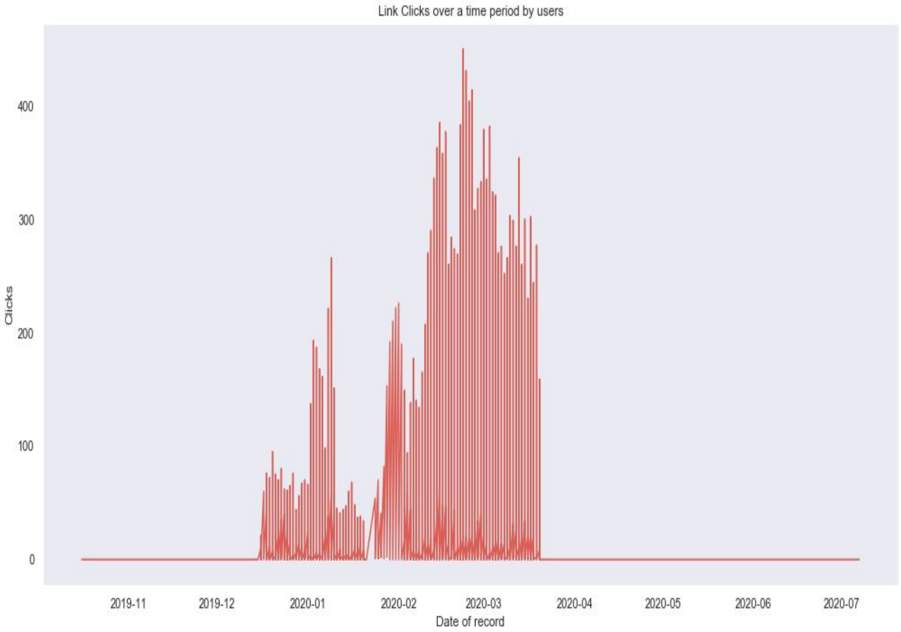
- Correlation between all the columns



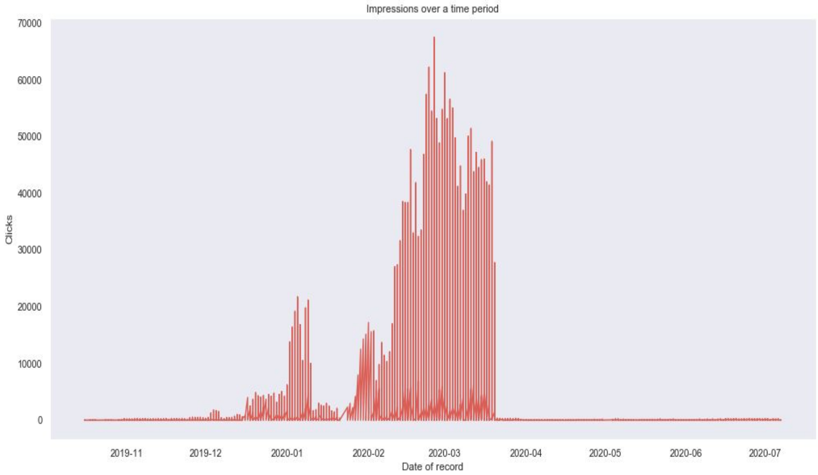
We can see that there is a high correlation between link_clicks and impressions.

Trend Analysis

- Trends across months: According to the time-series data more users clicked on ad links from February to March months of 2020.

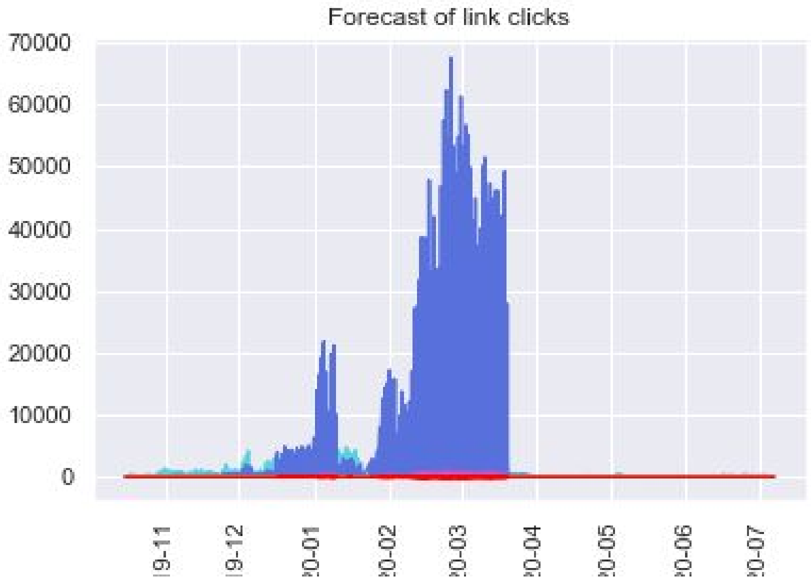


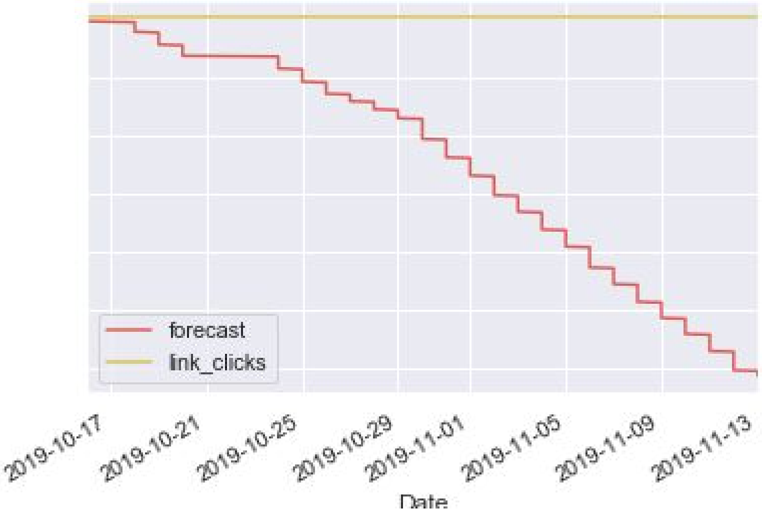
- The graph below shows the number of impressions i.e. the total number of times an ad displayed on a user's screen was high during March months which supports our earlier analysis that more ad links were clicked on during March and February months.
  Lesser number of ads were shown to users in late 2019 and after April 2020 which resulted in lesser number of link_clicks by the users.



Time Series Forecasting

Using the ARIMA model, the link clicks by the users for the future time period are predicted and are as follows:





Link to Jupyter Notebook:
https://github.com/devikathampi/EDA/blob/master/Assignment1_EDA.ipynb