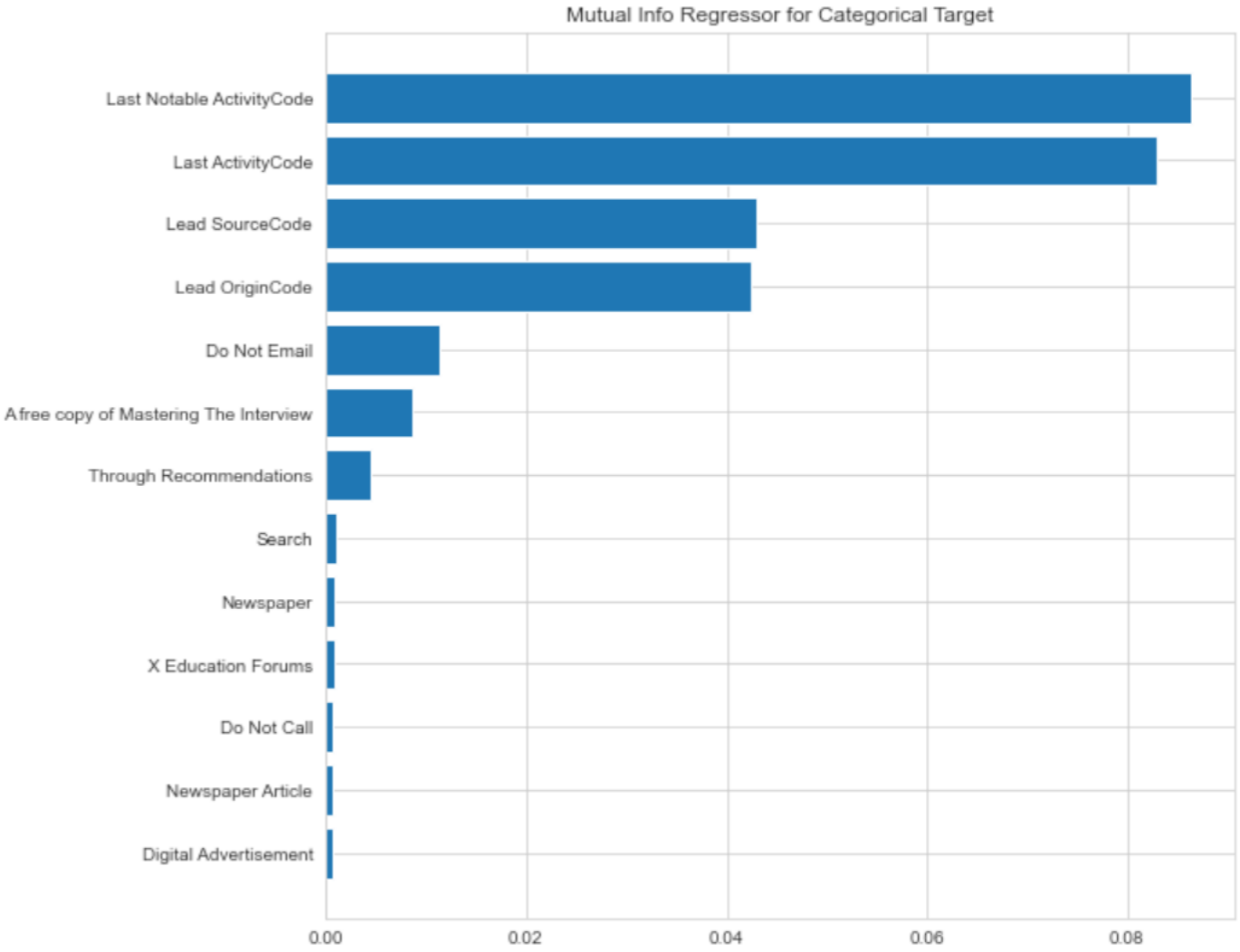
Lead Generation

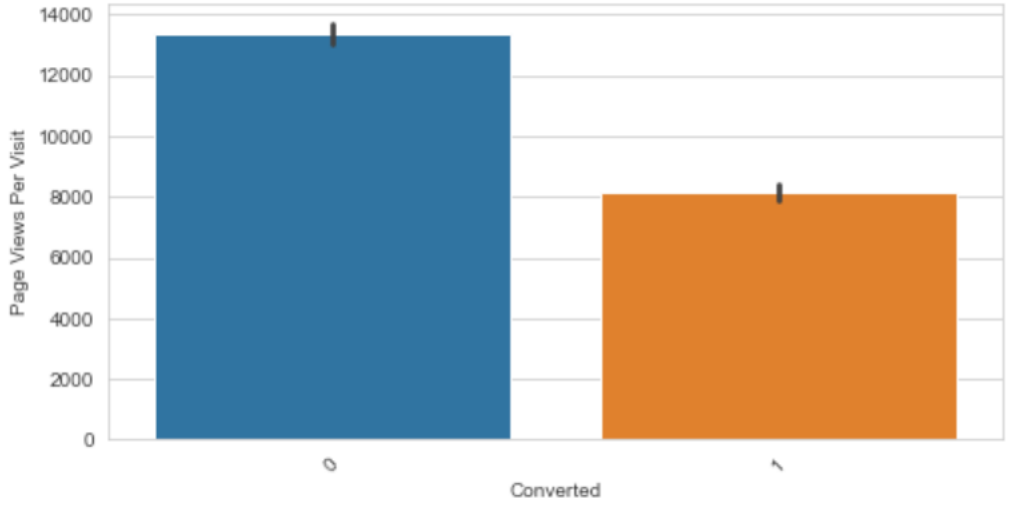
A Case Study

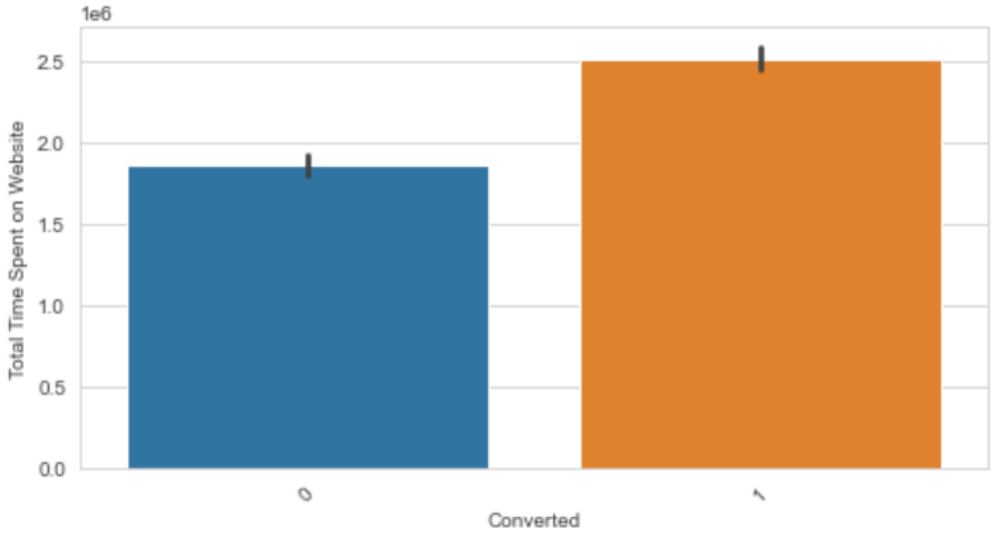
The Idea is to identify maximum number of Leads



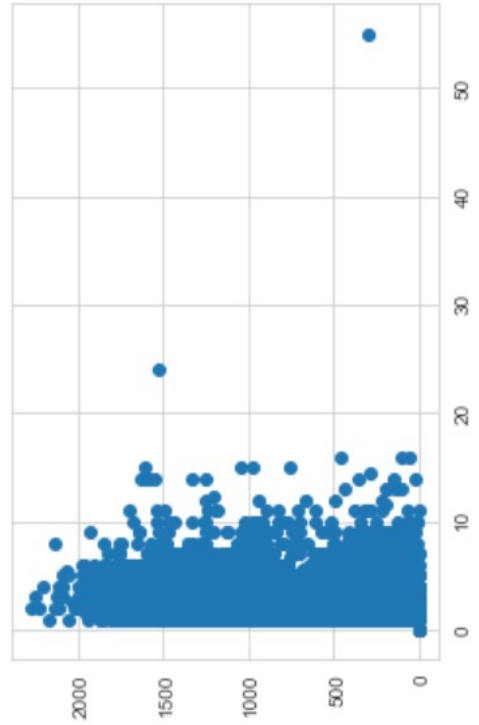
- First we cleaned the data, to retain only the significant information.
- We did not do any missing value imputation, keeping in mind the business problem.
- Next we identified the necessary **categorical** variables that contribute most.

- Based on statistical analysis we selected these to **numeric** variables:
- ['Total Time Spent on Website', 'Page Views Per Visit']



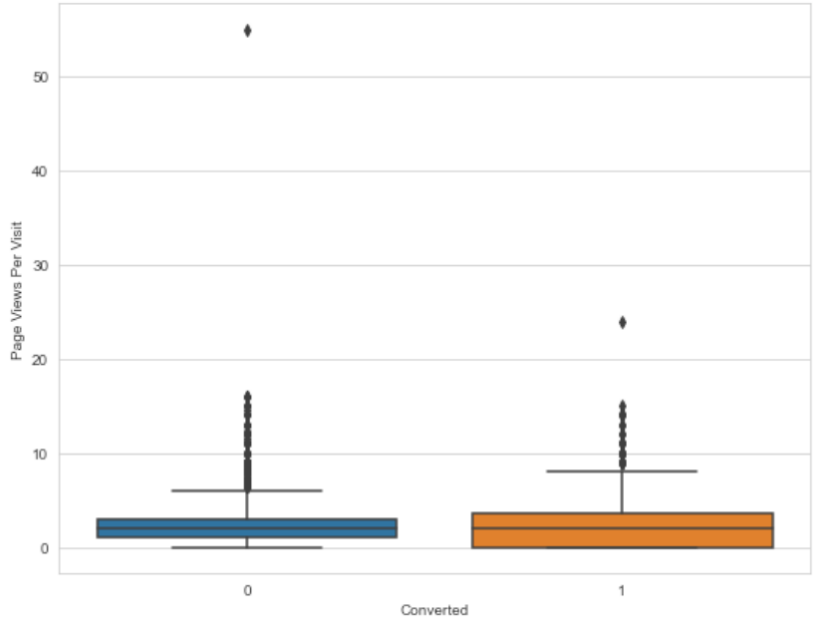


We applied aggregate function on both These variables, and plotted them against the target variable.



X = TIME SPENT
Y = PAGE VIEWS/VISIT

In total, Target people go to fewer number Of Pages per visit, but spend more time on The website

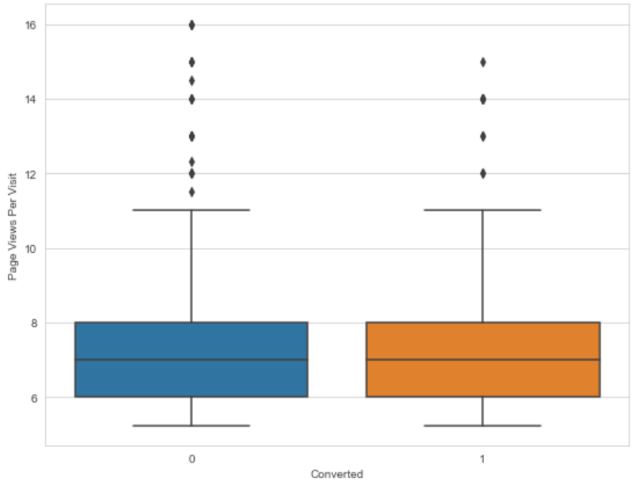


As there were some clear outliers in the Data, we devised them into two groups

1. Who visited more than 5 pages each time, but less than 20 pages each time.
2. Who visited less than 5 pages each time.

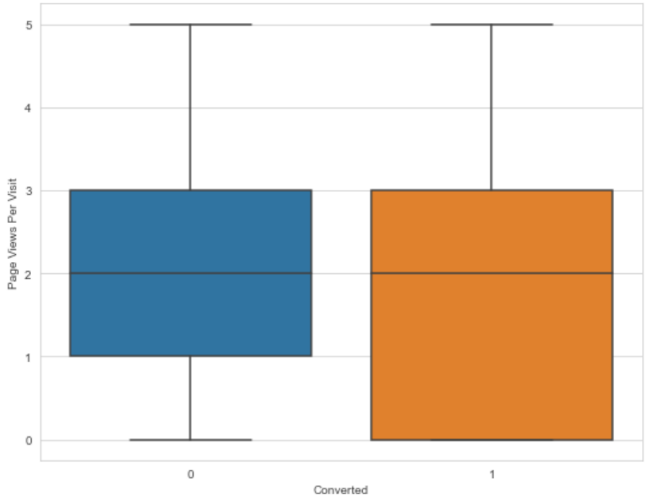
This is how they behave

1. There are a lot of Target people who visit less number of pages



Group 1

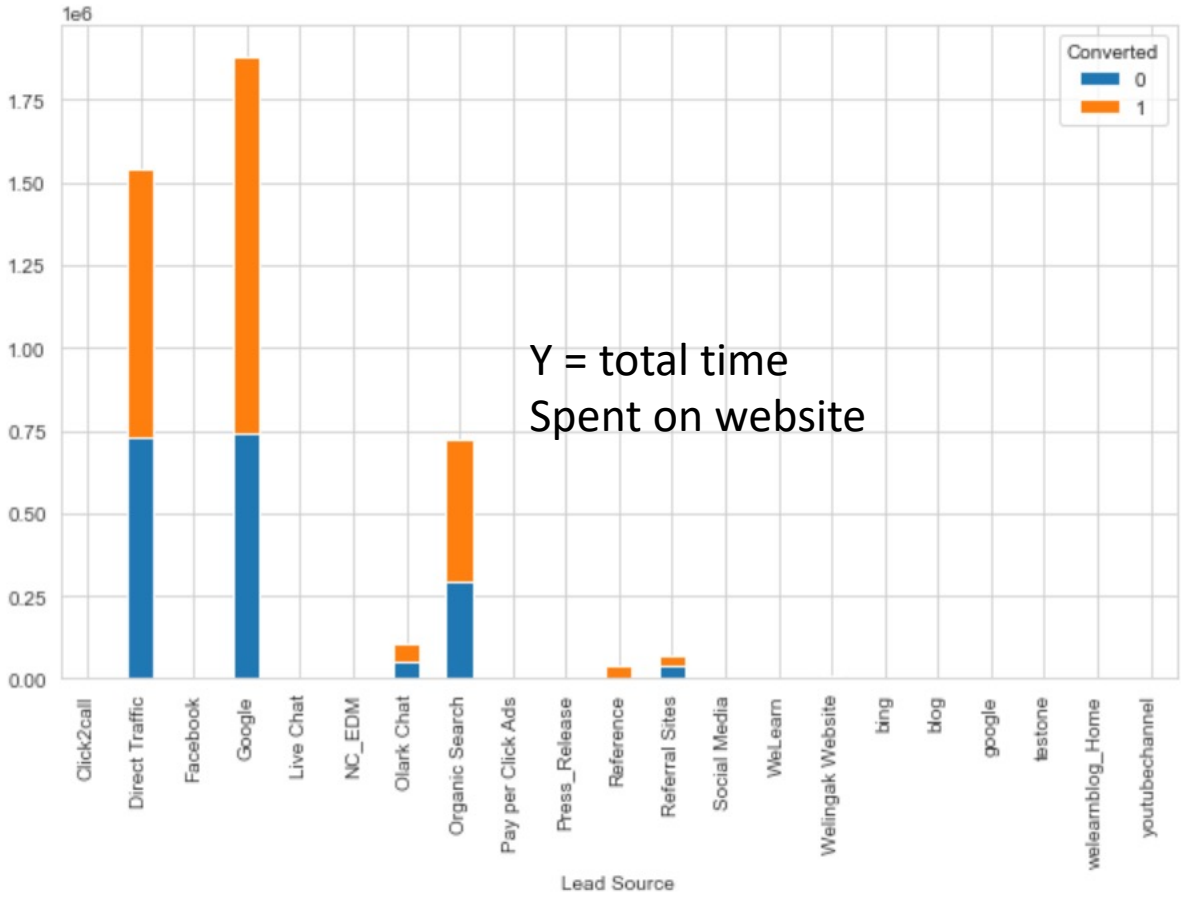
2. There are a group of people who visit lot of pages, but don't opt for any courses



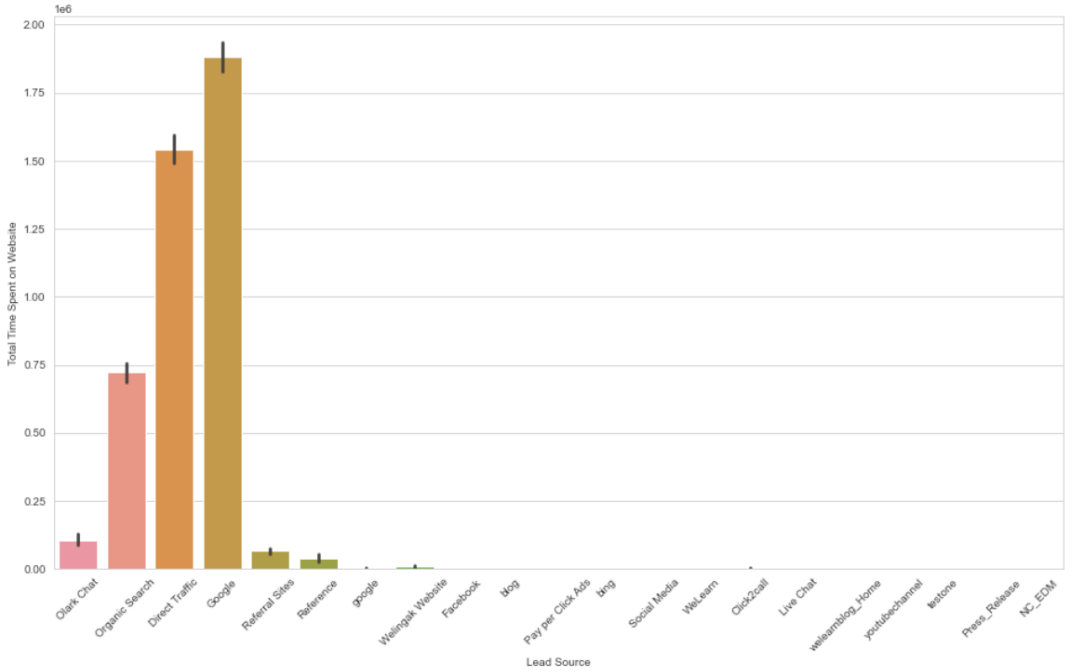
Group 2

Insights:

1. We see the three main **Lead Sources** that bring in the traffic.
2. There are a lot of lead sources that don’t bring any business.
3. The company may discontinue or prioritize those

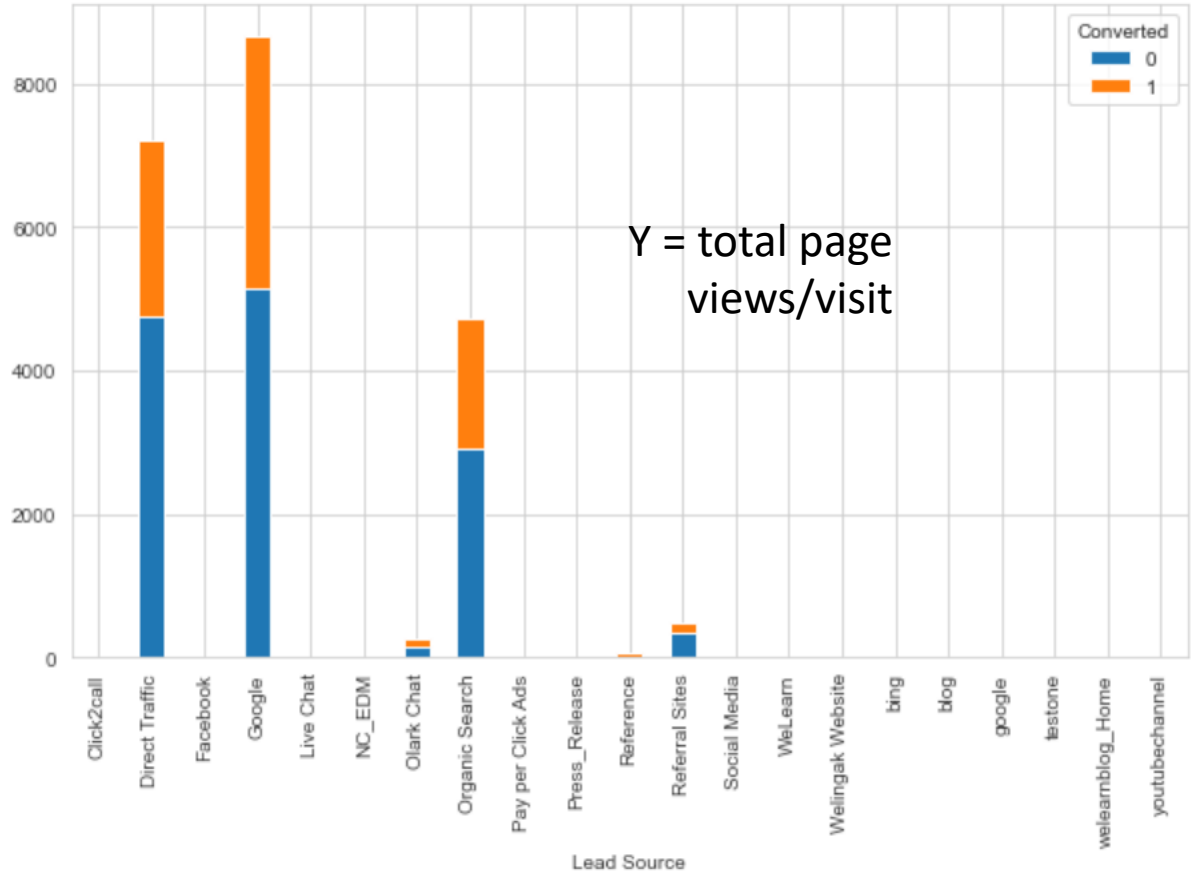


Y = total time
Spent on website



The different Lead Sources according to how much web-traffic they attract in total

We see this trend that people who visit lot of pages each time don’t opt for any courses.

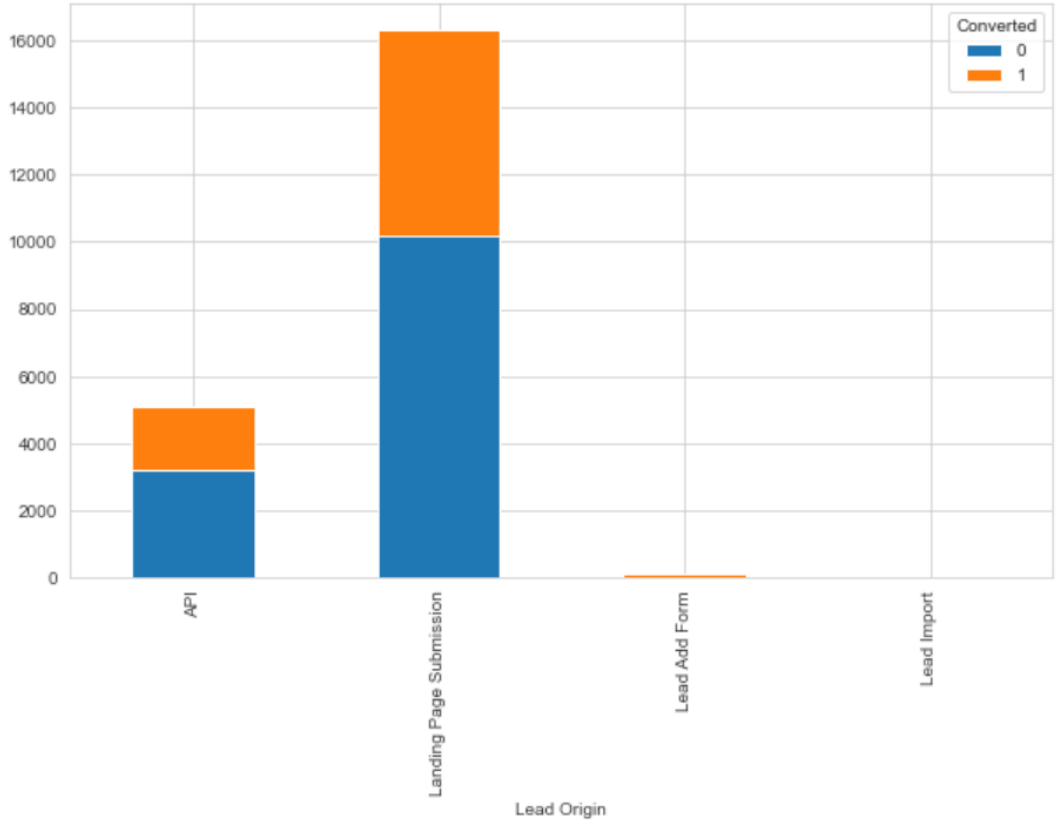


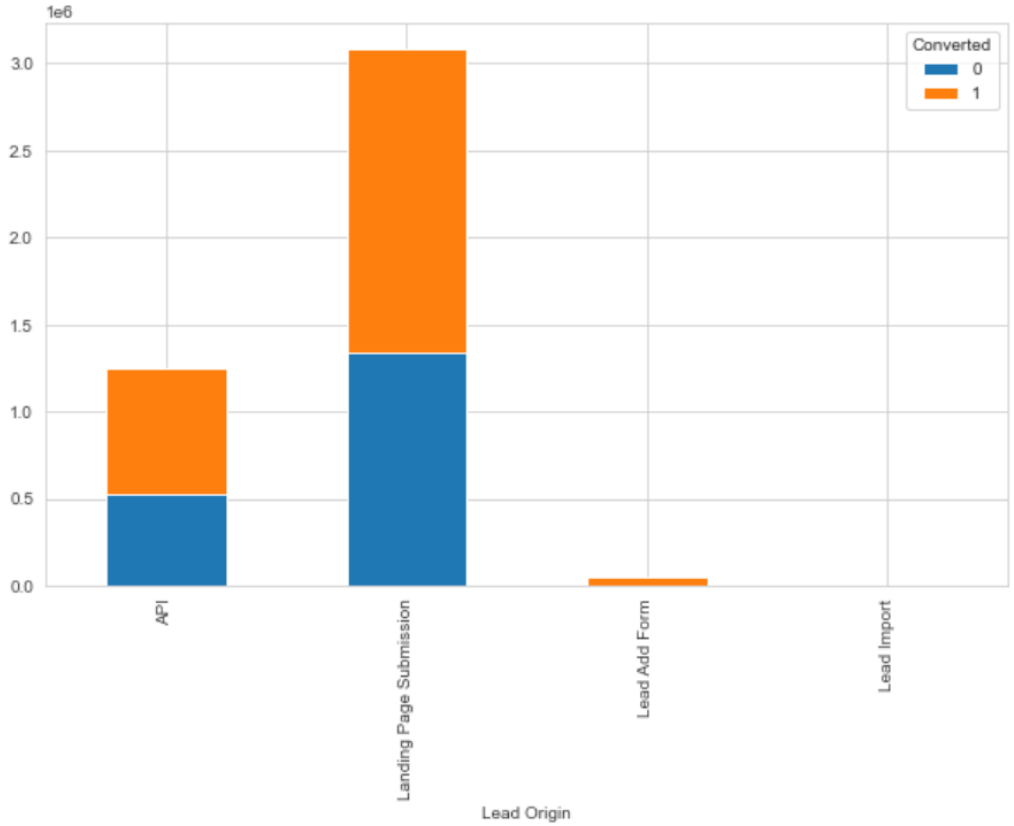
It may mean they are just casual surfers, and they are not serious about the courses.

Y = page views/visit

We plotted the Lead Origin variable against both the numeric types, and stacked them with two possible outcomes

Y = total time Spent on website

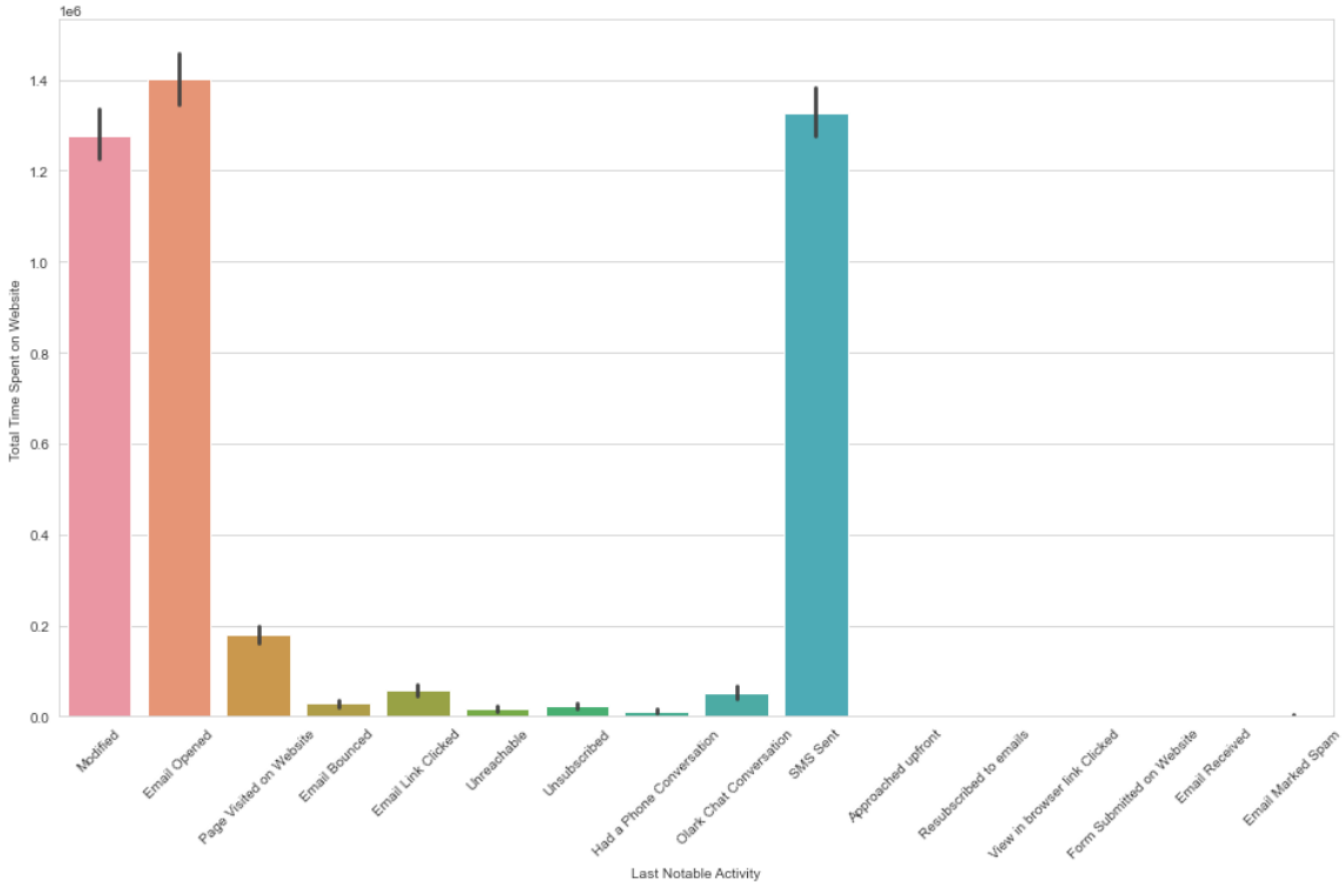


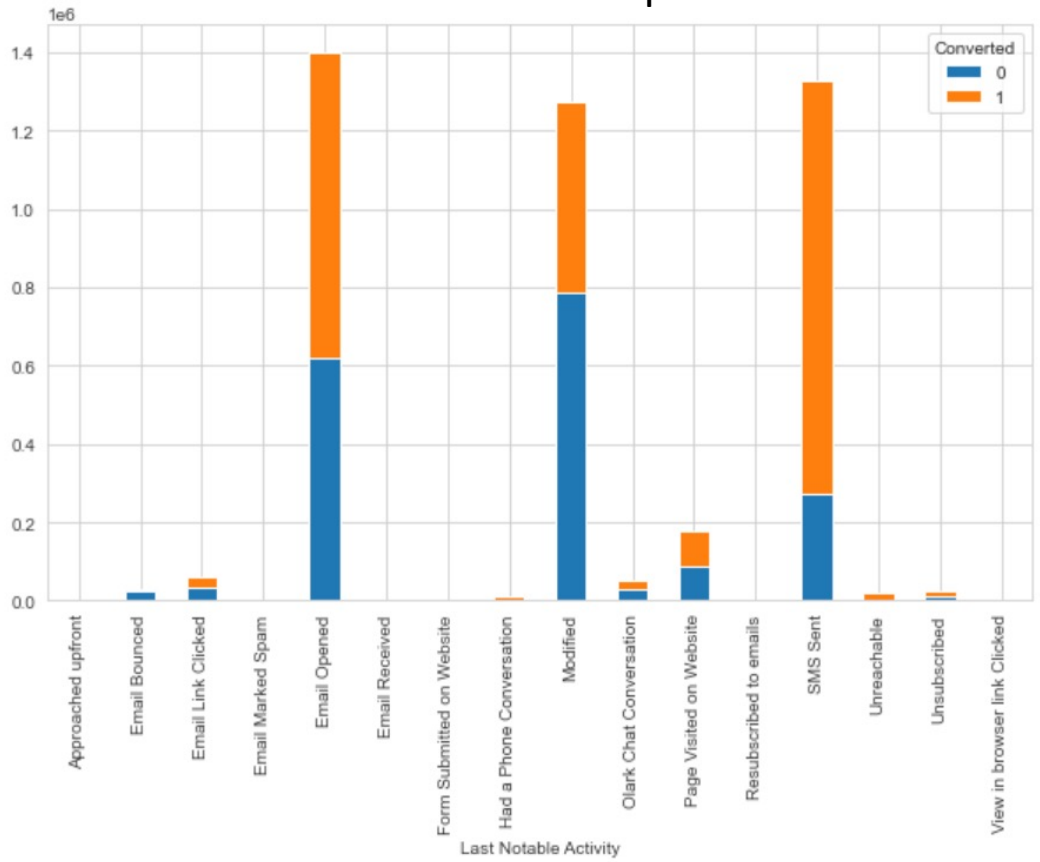


Again this shows, that people visiting lot of pages and not spending enough time on the website don't get converted.
This plot also shows the lead origins of almost all the traffic is concentrated in two levels.

Next, we did the same with '**Last Notable Activity**' variable

Y = total time
Spent on website



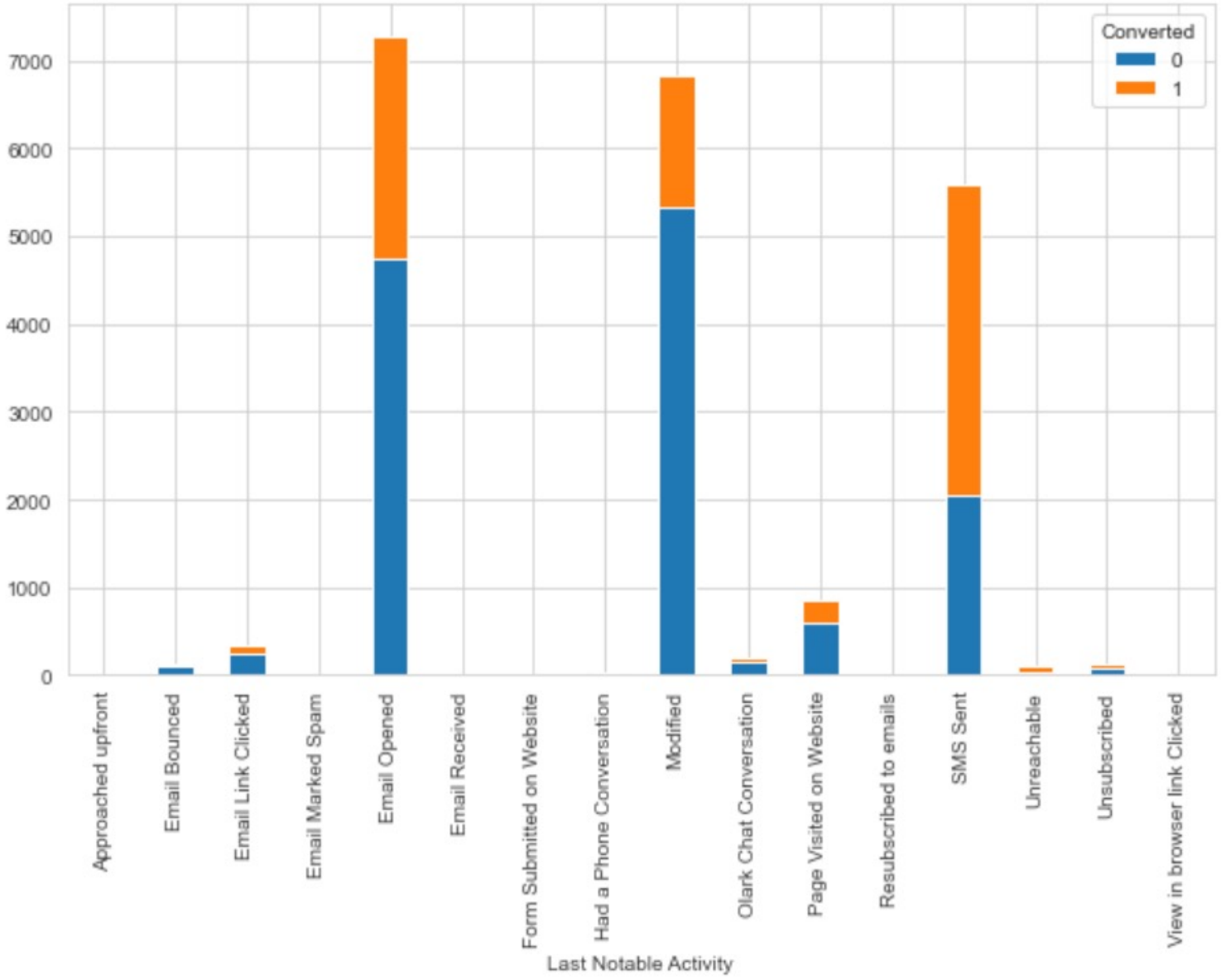


Spending more time on website means higher chances of conversion, as we saw before. We see three distinct notable activities that are associated with higher traffic.

Insights:

1. Those who have modified their profiles and spent a lot of time on website have lower conversion rates.
2. While whose who have sent an SMS and spent a lot of time on the website have a much higher conversion rate.

Next we plotted total **Page Viewed/visit** in similar fashion



Insights:

1. People who have modified their profile or opened their Emails and visited lot of pages have low conversion rate

2. People who have sent an SMS and visited lot of pages have a much higher conversion rate.

We fit a Logistic Regression model based on these variables

- We decided on two cut-offs, they can be interchangeably used based on human resources the company has at different year of the time.
- Both these models have a 78% accuracy, the conservative cutoff has a higher sensitivity.
- The less conservative one has a higher sensitivity.