For this process, I extracted the relevant data from our data warehouse. After discussions with the data scientists at work I understood that there are cases when activities are replicated, i.e the same activity is recorded twice to eliminate this duplication, I coalesced the time-stamp, the anonymous identifier and the referrer url. Looking at the unique values of these coalesced columns removed duplications, in the event that an anonymous id was associated with the same referrer url -which refers to either the ad they may have clicked or landing page they arrived on before moving to a different page then this was the same activity that had been duplicated for some reason.

In this particular case, missing data was not necessarily an issue, there are cases where a user may have an anonymous id but no associated user id, since the anonymous id is assigned when someone visits any of our pages, this often means that the person may not have started the sign-up process - and therefore a user id was never linked to their anonymous id, or for some reason we are missing data to associate that anonymous id to a user id. This happens in very limited circumstances and would not have a substantial effect on my overall project.

I then filtered out all activities associated with portal logins- where a user is visiting our site to login to their dashboard. I specifically wanted to look at the user's journey from session all the way through to them completing a sign-up, therefore logging in was not relevant to my overall analysis.

I lastly went on to write case when statements on SQL to label each activity based on the url path the user was visiting. For the most part, each path reveals where in the signup journey that user may be (session, lead, opportunity, complete).

This is all in the redshift SQL code I used to extract and carry out some data wrangling
GitHub Link:
https://github.com/EmmS21/Springboard-DSC/blob/master/AttributionModel/Code/Data%20Extraction.sql

After getting this data as a CSV extract I then imported it to a pandas dataframe using Python, appending the relevant column name to each column. In addition to this I grouped each ad source cleaning up cases, renaming every utm_source containing the word Facebook as a Facebook referral and double-click or Google as Google Ads, I then saved this cleaned up version of my data extract as a CSV for further analysis

GitHub Link:
https://github.com/EmmS21/Springboard-DSC/blob/master/AttributionModel/Code/Cleaning%20up%20attribution%20data.ipynb