**Data Wrangling Overview**

Gathering data from my Capstone project diverged a bit from how data scientists normally gather, prepare and tidy data to perform exploratory data analysis.

Traditionally one would scrape APIs or use SQL/DPLYR to connect to some sort of database and gather the data necessary to perform an analysis. In my case I had to reach out to various advertising partners to get data that came in various formats but were restricted to csv files.

Each advertiser measures performance in various ways and scores conversions differently due primarily to the fact that offline and online advertisers impact customers in different patterns.

After receiving all the data I anonymized and factored the values for privacy purposes on my end. I then structured the data outside of the traditional tidy format that relies on key-value pairs and having each column be a variable while each row is an observation.

The reason for my divergence is due to the nature of the problem that I am trying to solve. I am trying to see how money spent per advertiser affects new customer acquisition. While advertisers provide their measurements for customer counts individually one cannot aggregate them because their sum will exceed the actual counts on the companies end!!

This phenomena is known as multi-touch attribution and impacts all marketers who use an onmi-channel approach. There are third party tools in place to help this but come at high costs. With that said there are no true key value pairs when comparing advertiser x spent on day y and how many customers were acquired by that partner on that day.

I discussed this with my mentor after the fact and we decided to use the tidy approach while filtering out some criteria. It was a great mental exercise that generated more ideas on my end.