**Week of March 11th-March 17th**

**Blog 5: March 17th, 2019**

During this week after spring break, we continued with our data exploration but moved towards the modeling aspect too. We took the draft as a starting point to pick questions that we would like to incorporate in the modeling process. Jeremy and I focused on the question: ‘when will the client close a deal? Is there an identifiable cluster of timings or the types of deals closing?’. This was a question that was highlighted as a go-ahead from the client on our preliminary draft.

The first step was to merge the current dataset to the historical one. For the purpose of our question, the column names of key interest were: ‘actualcloseddate’, ‘createdate’, ‘opportunityID’. The column name that was merged with was ‘previousopportunityid’. From this, we could get an understanding of the distribution of the different stages, understanding the proportion of won and lost deals and the time range of won and lost deals (ie. number of days taken). There was some confusion regarding similar column names for ‘createdate’. Mariem later clarified this during our call. She also suggested that to better understand which column names to select, or the distribution of it, we could look at the proportion of missing values or NAs to get a sense of which one is the most complete record. Currently, the output shown only had codes, however, taking in the feedback, we would be working on more visualizations so that is much better for the client to understand the variable spread, and the calculations of differing time-ranges for each of the ‘won’ and ‘lost’ deals.

During our call, we got feedback and clear direction on how to include visualizations. For example, for the output of the time-taken for ‘won’ and ‘lost’ deals, it would be better to interpret it in a graph so that the difference can be easily interpreted. It was also suggested to use median to understand the date range. Another possible route that was suggested was to bin time ranges upon our discretion. Such as we can say that ‘more than 6 months’ would be categorized as ‘high’.

Over the weekend, we worked on adding visualization to better support our analysis. For the output of timerange (number of days taken) for the deals ‘won’ and ‘lost’, we created a graph to see which date range does the maximum ‘won’ deals cluster into. A similar visualization was created for the time range of ‘lost’ deals and comparisons were made.

We also tried to understand how we could include ‘modifieddate’ for a time-series model to understand the change of different features over time and what effect they have.

We also progressed in making a sample dataset on Colab to understand and present logic in a manageable framework.