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

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

During this week, we divided our task into two parts, data exploratory analysis and modeling. I was mainly responsible for figure out the important factors in the dataset and performed some visualizations so that it could be much cleared for both the client and us. Also, I would help to build the model. We regarded the feedback in the preliminary draft as a good guidance to make progress. In the past few weeks, we made little progress and the main reason was that we always concentrated on the same problem which would be inefficient. Thus, I suggested that we could work separately according to the preliminary draft and discussed after we had some progress or interesting findings.

Rashi and I mainly focused on the question:

1. when will the client close a deal?
2. Is there an identifiable cluster of timings or the types of deals closing?
3. Determine the effectiveness of each business stage

At first, I merged the historical dataset and the current one because the ‘createddate’ and ‘closedate’ exist in different dataset and I need both columns to work on the first question. After that, I subset the merged set into won and lost since either of them represent a closed deal. I calculated the average and median time for closing a deal and performed visualization based on Mariem’s advice. Also, I found there are two similar columns, both of with might represented the created date for each deal. Mariem gave us some useful advice about how to figure out which one would more likely to be our targeted column. For example, we could focus on the proportion of missing values to get to know which one is more complete. For the third question, I calculated how many deals finally closed in each stage, in which way I could get to know their effectiveness. As a result, we were informed how important each stage is.

For the modeling part, our team worked on the variable selection. Since we didn’t receive the data dictionary before, we had to picked up some variables that we thought are important. Also, there were many columns that we still felt confused and clarification is needed. During the meeting with Mariem, we received many supportive information with regard to our new questions. In terms of the time series method, we would design a sample dataset to present logic in a structured framework as suggested by our professor.