**Week of February 9th-Feb 17st**

**Blog 3: Feb 16th, 2019**

In these weeks, our group focused on the initial data preprocessing. There are many missing values in some columns and we started to clean the dataset. Our mentor Chris Kelly met with us and gave some suggestions about how to crack the problem before working on the dataset, which was quite helpful. We had a deeper understanding about how probabilities are set to different stage and how long it took to move from one stage to another.

After being authorized with the access to the virtual system, we created a Share Drive and made sure that we could work on the project and communicate more effectively. Using Jupyter notebook, we performed some basic descriptive analysis and discussed our methods with Mariem weekly. Initially, we had a lot of problems about the concept for different column name and the relations among those datasets. We sent out emails asking for clarify so that we would not work in a wrong way. Also, we asked whether we could be given a clear data dictionary for us to better understand the datasets. Meanwhile, we kept working on the exploratory data analysis.

From the file ‘APACAMER.csv’, we found some variables that we thought would be important for our further analysis. We tried to figure out the ideal timing for closing a deal, whether it’s won or lost eventually. We learned about the difference between ‘stage’ and ‘status’ columns which always made us feel confused before. We also found that Probabilities will be discrete values which means we could regard it as a categorical value. We tried to figure out whether there is a relation between product group level and our target variable.

From another file ‘Atlas.csv’ which includes historical data about customers, we got to know more detailed information about the change with regard to the expected closed dates, amount for each deal and so on. According to email from Mariem, the primary key for this file is opportunity ID, which we could use to merge with ‘APACAMER.csv’ file. Hopefully we can generate more valuable information after we reached out the merged dataset.