**Blog 3: February 18th, 2019 (Week of Feb 10th - Feb 17th)**

During this week, Chris Kelly, our capstone mentor, had a meeting with us and suggested us to think of some questions ahead before we explore the dataset. It helped point out the data exploration target and avoided aimless and repetitive exploration. He also got us in touch with a professional sales person to answer our questions about the timing of sales pipeline. It gave us another view on how sales people assign probabilities to clients at every stage and how long it takes to advance to the next stage. We really learnt a lot from this chat.

During the weekly meeting with Mariem, we got familiar with the csv data files we had. One of the csv file doesn’t include column heads and we can hardly utilize it. We communicated with Mariem about this situation and continued working on the other 2 files. **We used Google Drive to write down some important things about csv files.**

For the **APAC.csv** file:

* Account ID: Always a number
* If it is the same account, itl will be the same ID
* Lot of fields are null (check for data cleaning before to replace)
* Client ID
* Stage
* Probability: there are few fields added due to system migration (they will be labeled with an underscore)   sth. Starts or ends with amd
* Probabiliies will be distinct values- not continuous- could see them as categorical
* Amount could be different currencies: default would be USD
* Status: values other than won or lose are outliers
* Opportunity for an account or a product;
* Coverage: it could be a coverage 1 person or coverage 2 person : what matters the most is coverage 1(main coverage)- person you will go to for the opportunity basically - there are people rarely for coverage 2- more for backup! GO TO COVERAGE 1 (maybe coverage 2 could be interesting) **Coverage 2 can be ignored / EDA may need it**
* **Modified: system or someone else edit it**
* Products; hierarchy: product group level structure (PGL)=> PGL A,PGL  B, PGL C >> Product (product is more detailed), A is a parent company (Alphabet would be A and Google might be product)

For the **Atlas opportunity file Audit history** file:

* More about opportunity
* Equivalent
* This is one of the archive files
* Nice and Clearer
* Amount yield
* mapping : opportunity ID: you can join different files through this
* NOT complete
* Tie back an opportunity : winning or losing
* Status/stage: is something 0 as opposed to 1: which of the features affected it!!!

We also divide our problems into several parts. My job is to analyze: What are the factors influencing the succes of a deal? Since it is a supervised learning problem, I will utilize logistic regression, decision tree, random forest and other classification methods to deal with this problem. And I will try to figure out the theoretical basis and Application range of different methods.