**Main progress**

* Explore the dataset
* Variable exploration

**1. Explore the dataset**

After we get the data, we need to dig into the variables. Since we don’t have the data documentation or the data dictionary, the only way we can do first is to take a guess.

The big audit history dataset doesn’t have column name. Although we received the head file from the Credit Suisse that have the column name, we find the column names from the head file can’t match with the dataset.

Thus I choose first explore the APAC file(the other dataset we have), since this dataset is more well structured. And I will use the code and model built on this dataset in the audit history file after Credit Suisse or we find a way to us it.

**2. Variable exploration**

Because our main aim is to gauge the probabilities of closing a deal and determine what the factor are that influence the probability of closing a deal, we need first find what is the final status(close or not) variable. I set amount, product level, produc\_tid, product\_level, coverage\_id, and coverage\_region as independent variables. I am pretty sure that amount is the amount for the deal; product\_id is the id of the product, product\_level is which segment in Credit Suisse provide that product. I think coverage might mean the customer information. I choose this variables for two reason:

1. This variables don’t have many missing values

2. The variable is categorical variable(except amount), because I’d like to explore the closing deal rates in different sub groups