**PRE-PROCESSING**

* **Understanding data**

1. We looked at the data thoroughly and we searched every column name to understand what it means

* **Removing duplicates**

1. Using Excel, we searched for duplicates using the customer numbers column and we found 401 duplicate customers

* **Changing every Blank, ‘N/A’ and “Unknown” values into a Null value**
  1. We found 3962 “unknown” EX: Educational Level, Gender and Martial Status
  2. We found 100 Blanks values in the Attriation\_Flags
  3. We found 602 “N/A” EX: Gender and Card Category
* **Validating data**
  1. We validated the data using a formula in Excel to calculate the Age by the Date of Birth (“=INT (YEARFRAC ("Cell Value", TODAY ()))")
* **Identifying outliers** 
  1. Based on the IQR for the Age we calculated the lower range to be “22” and the upper “70”
* **Removing outliers**
  1. On the Age column we capped the data depending on the limits we referenced above
* **Filling the NULLs**
  1. We addressed the missing values in the Gender column, the Income\_Category column and the Card\_Category column using mod method
* **Categorization**
  1. We categorized the Age in 5 categories and we named the new column Age\_group
  2. We categorized the Credit\_limit in 5 categories and we named the new column credit\_limit\_category
  3. We categorized the Total\_Trans\_Amt in 5 categories and we named the new column transaction\_level