# Challenges and Techniques used:

* 1.There were many missing values for both numerical and categorical variables
  + 1.1- Categorical variables were **binned** together because of **high cardinality** and then **mode** was taken to impute missing values
  + 1.2-Numerical variables were imputed using **model based imputation**,which means that if a particular column has a missing value for a particular row,then the value for missing column is determined ny other columns for the same row,which I feel is better.
* 2.There was a format inconsistency with string value for numerical variable that had to be converted to numerical after splitting up the string.
* 3.Feature Engineering:
  + Some variables weren’t useful as such,such as other skills which was converted to relative\_count of skills with respect to other candidates in order to get a quantifiable metric that would be useful for ranking candidates
  + Variable such as current year of graduation had to be transformed into work experience through logic.
* Normalization
  + Before using k-Means clustering,had to normalize the data in order to avoid overweight of certain variables
* Value of K was chosen as 2 since we want two categories-one high performing and another -low performing.
* Had to use value of centroid to determine which among the clusters was the high performing one.
* Faced issue with data corruption,which was caused due to some transformations used in code.Had to track back,but unfortunately due to time constraint,could not actually get to the root of the data lineage tree.

# Ways to improve:

* Could have looked more into optimum no of clusters and other model parameters
* Could have used more functions to reduce repeated code
* Looking forward to suggestions from you