**Problem Statement:**

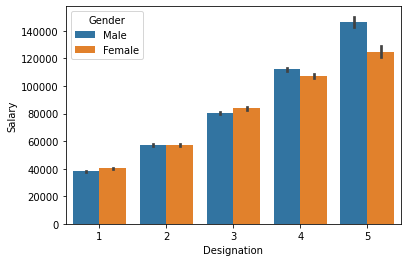
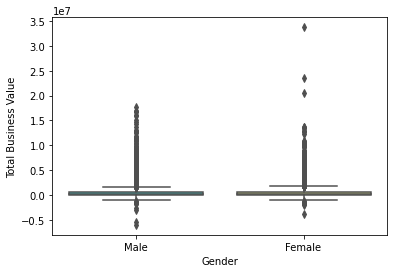
**Predicting Employee Attrition**

In recent years, attention has increasingly been paid to human resources (HR), since worker quality and skills represent a growth factor and a real competitive advantage for companies. After proving its mettle in sales and marketing, artificial intelligence is also becoming central to employee-related decisions within HR management. Organizational growth largely depends on staff retention. Losing employees frequently impacts the morale of the organization and hiring new employees is more expensive than retaining existing ones.

You are working as a data scientist with HR Department of a large insurance company focused on sales team attrition. Insurance sales teams help insurance companies generate new business by contacting potential customers and selling one or more types of insurance. The department generally sees high attrition and thus staffing becomes a crucial aspect.

To aid staffing, you are provided with the monthly information for a segment of employees for 2016 and 2017 and tasked to predict whether a current employee will be leaving the organization in the upcoming two quarters (01 Jan 18 - 01 July 2018) or not.

**Exploratory Data Analysis:**

* Salary,Age Feature are in Normal Distribution.
* In EDA I found that there is no relationship between most of the columns.
* Total Business Value Feature not in Normal distribution so i transformed data to cube root to follow normal distribution.
* There is not much difference in salary of different educational levels.
* Checked the hypothesis of categorical columns to check our hypothesis
* Testing Hypothesis by conducting t-test
* #Null(H0): Average Monthly Salary of Gender are equal
* #Alternate(H1): Average Monthly Salary of Gender are not equal
* Ttest\_indResult(statistic=-1.8406083226609546, pvalue=0.06569577396151297)#p-value is greater than 0.06 there is not enough evidence to reject the null hypothesis at alpha=0.05
* Designation and Salary in linear relationships.
* Found outliers in “Total Business Value” column.

**Data Preprocessing:**

* In our data set actually there are only 2300+ unique employees only but they reported more than once some of them even 24 times so we have 19000+ observations.
* So Grouping and aggregating the features will result in good results because we can’t lose any information about employees.

**Missing Values Treatment:**

* There is only one column Last Working Day having null values
* Actually we can’t fill missing dates here because there is a lot of uncertainty in the column.
* So i converted this column is like a categorical feature whether it is having last working day or not(0 and 1 decoded)

**Feature selection and Feature Engineering:**

* Dropped irrelevant features like reporting date,joining date,Last working date.
* Transformed Age column like categorical column like Young Adult/Senior Adult.
* Created new feature Growth by subtracting joining designation and designation
* Created Avg Rating and Total\_Business\_Value Columns by aggregating the dataset.
* Educational qualification creating ambiguousness in the clusters so i’am dropping that column.

**DummyEncoding and Label Encoding:**

* Dummy encoded the categorical features like city and Label encoded ordinal data like Educational qualification

**Model Selection and Design:**

* Our data lacks a specific dependent variable. I am going with an Unsupervised learning model.
* We know that our data has two specific clusters so k-means clustering is the best method to try in my opinion.

**Model Evaluation:**

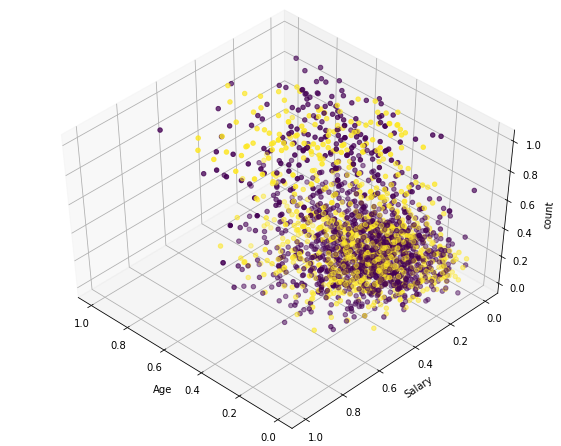
* Even Though i tried my level best to separate two clusters using k-means clustering, I can't get enough accuracy with my model.
* My model got silhouttee\_score of 0.45 not bad with this data

**What not worked:**

* Extracting dependent variable from the data and using supervised learning prediction does not work.
* Using noisy features like city,Educational qualification not giving a good silhouette score.
* Dbscan clustering not working with our data.

**What worked:**

* Scaling the features
* Unsupervised model k-means clustering worked well.

**Final model Output:**