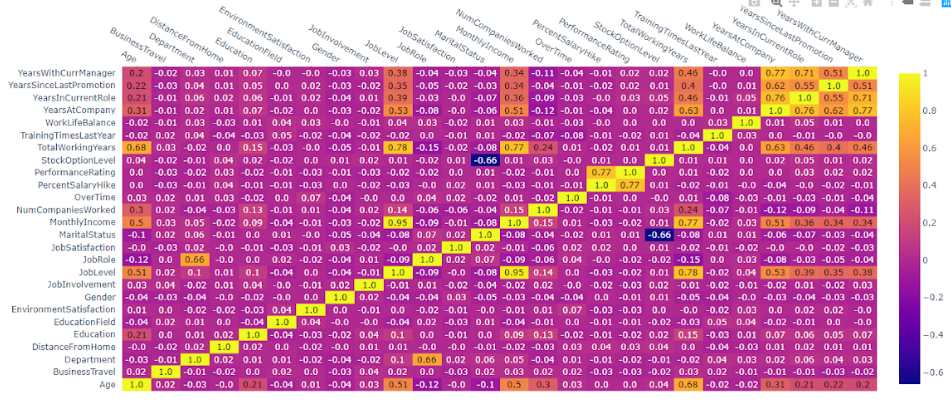
Analysis and Visualization of Employee Data

**Exploring relationships among variables:**

The following columns in the dataset had categorical data – Business Travel, Department, Education Field, Gender, Job Role, Marital Status and OverTime. So the data had to be converted into numerical form using LabelEncoder, in order to use the data to plot a correlation matrix. The screenshot below shows the correlation plot.

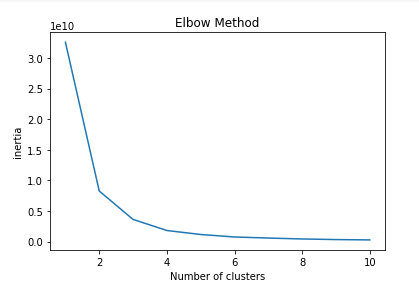


From the correlation plot above, we could find meaningful relationships among some of the data columns.

* The variables Years since last promotion, Years in current role, years at company and years with current manager are highly positively correlated.
* The variables Job Level, Age and Monthly Income are highly positively correlated.
* The variables Performance rating and percent salary hike are positively correlated.
* The variables MaritalStatus and StockOptionLevel have negative correlation.

**Data clustering using k-means technique:**

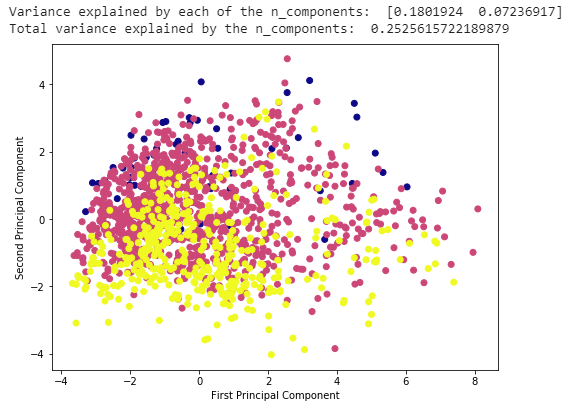
We have used k-means clustering technique to visualize groupings/clusters of employees from the dataset. First we have normalized the features in the dataset using StandardScalar() function. StandardScalar() transforms each column in the data in such a way that its distribution will have a mean value 0 and standard deviation of 1. Once the features are normalized, k-means algorithm is applied to determine the number of clusters in the data using Elbow method.



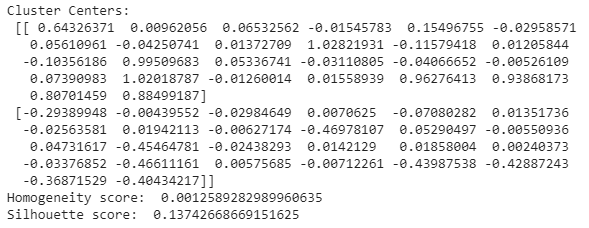
The screenshot above shows the Elbow method which determines the number of clusters based on the location of bend or elbow in the plot. In the above plot, we see that the bend is at 2, hence the number of clusters in our data will be 2.

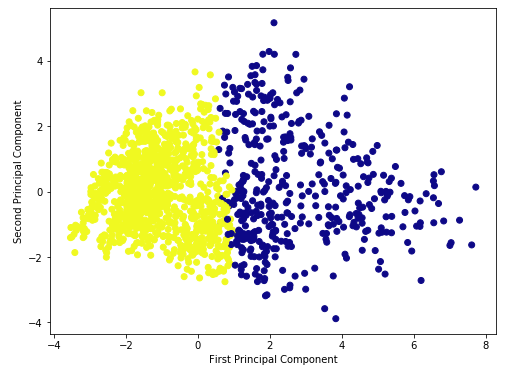
**Dimensionality Reduction using PCA:**

Principal Component Analysis (PCA) is a dimension-reduction technique that converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. It can be used to reduce a large set of variables to a small set that still contains most of the information in the large set. It is used to emphasize variation and bring out strong patterns in the dataset. The screenshot below shows the PCA scatterplot.



We have applied different cluster quality metrics to figure out more about the quality of the clusters, so that we can judge the goodness of fit of the cluster labels to the correct labels.





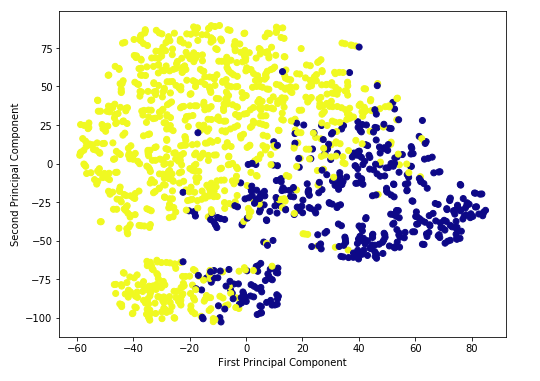
The screenshot above shows the kmeans scatter plot with the Cluster Centers, Homogeneity Score and Silhouette Score metrics.

* The homogeneity score shows to what extent all the clusters contain only data points which are members of a single class.
* The silhouette score measures how similar a data point is to its own cluster compared to other clusters. The silhouette scores usually range from -1 to 1, where a higher value indicates that the object is better matched to its own cluster than the neighbouring clusters.

The scores in the screenshot above don’t seem to be great. We see that the value for silhouette score is 0.15, which is close to 0, which indicates that the sample is on or very close to the decision boundary between two neighbouring clusters. This could mean that the samples could have been assigned to the wrong cluster.

**Dimensionality Reduction using t-SNE:**

Next we have implemented t-SNE (t-Distributed Stochastic Neighbour Embedding) on the data. t-SNE is a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets. The screenshot below shows the t-SNE scatter plot.

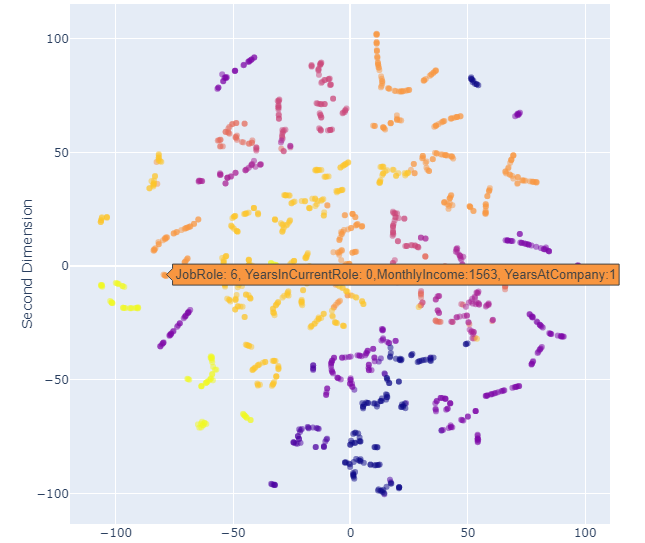


The t-SNE scatter plot shows significant improvement in comparison to the PCA scatter plot by showing clear clusters of data points.

**Exploring relationships among data variables at company level:**

We have used the variables - JobRole,YearsInCurrentRole,MonthlyIncome,YearsAtCompany

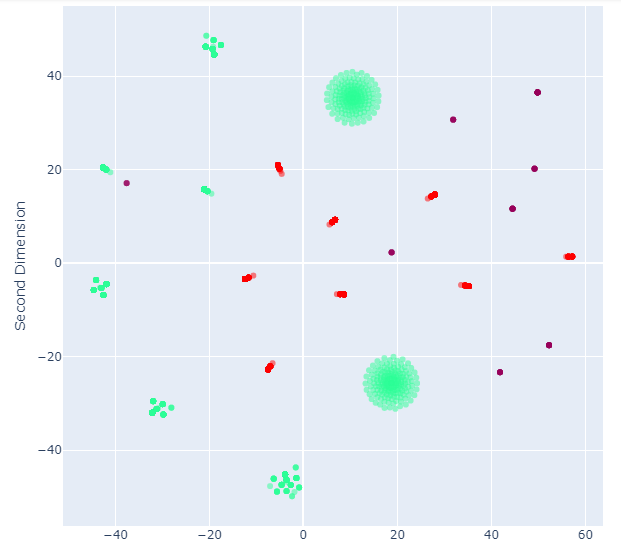
The data is grouped on the basis of JobRoles in the company. We see that the employees with low JobRole are more than higher Job Roles. The employees with low JobRoles have less YearsInCurrentRole,MonthlyIncome and YearsAtCompany compared to employees with higher JobRoles.



**Exploring relationships among data variables at department level:**

The variables used here are Department,Age,JobSatisfaction,Gender

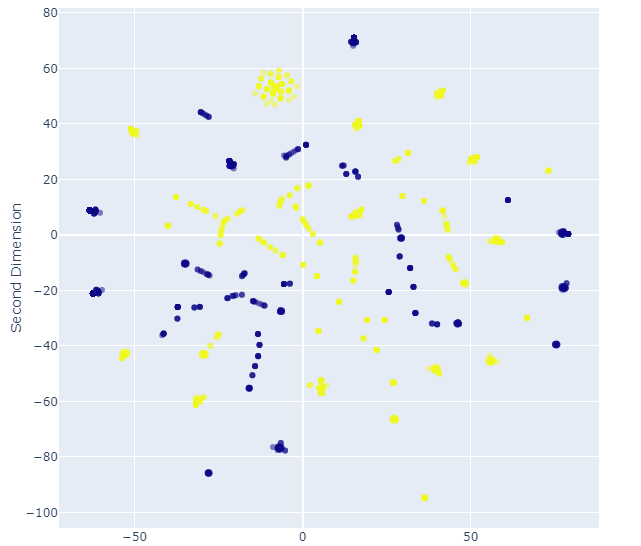
There are clusters observed for employees of Research & Development Department. It is noticed that the employees in this department more age, more job satisfaction and are mostly male employees.



**Exploring relationships among employee data based on Churn:**

The variables used are JobSatisfaction,OverTime,JobLevel,YearsSinceLastPromotion

The employees with high JobSatisfaction are more than low JobSatisfaction. The employees with low JobSatisfaction are due to low Job Level and more YearsSinceLastPromotion even though they have no overtime.



In order to retain the employees the company can do the following things:

1. Providing a platform for employees to speak their mind freely within the organization
2. Improving the work life balance by providing sufficient leaves for sick days, family vacations, etc. Also reducing the level of pressure on employees at all times and allowing them to relax by conducting team building activities or giving short breaks during the day.
3. Showing the employees that they are being trusted by giving them responsibilities that allow them to grow, encouraging them to gain new skills, providing education opportunities and conducting internal hiring wherever possible, and giving promotions based on their performance.