***HR Analytics Project- Understanding the Attrition in HR*  
  
1.Problem Definition.**Human resource analytics is an area in the field of analytics that refers to applying analytic processes to provide insight into each process by gathering data and then using it to make relevant decisions about to identify how attrition affect companies and how HR Analytics help in analyzing attrition.  
  
Thus, every year employees leave the company and need to be replaced with the talent pool available in the job market. The management believes that this level of attrition (employees leaving, either on their own or because they got fired) is bad for the company, because of the following reasons -

1. The former employees’ projects get delayed, which makes it difficult to meet timelines, resulting in a reputation loss among consumers and partners
2. A sizeable department has to be maintained, for the purposes of recruiting new talent
3. More often than not, the new employees have to be trained for the job and/or given time to acclimatise themselves to the company

Hence, the management has contracted an HR analytics firm to understand what factors they should focus on, in order to curb attrition. In other words, they want to know what changes they should make to their workplace, in order to get most of their employees to stay. Also, they want to know which of these variables is most important and needs to be addressed right away.  
which aims to identify important factors that might be influential in determining which employee might leave the firm and who may not. This article provides in-depth analysis as well as predictive modelling to understand important factors and make accurate predictions.  
  
  
 **2. Data Analysis.**  
The dataset describes the attributes in HR attrition analytics. It has 1470 observations with 35 variables. Out of the 35 variables, there exists one target variable *Attrition*with possible outcomes *Yes*and *No*. The other 34 variables are independent variables but one, that was, *Employee Number*which denotes the employee number or the identification number.  
Following are the variables:  
Age : Defines the age   
Attrition : Yes/No parameter  
Business Travel :Frequency of travel  
Daily Rate : daily rate   
Department : Company department   
Distance From Home: Commute distance   
Education : No of Courses   
Education Field : Type of education  
Employee Count:; Count of Employees   
Employee Number: ID of employee  
Relationship Satisfaction : Numeric values   
Standard Hours: Numeric values  
Stock Option Level: Numeric values   
Total Working Years: Numeric values   
Training Times Last Year: Numeric values   
Work Life Balance: Numeric values   
Years At Company: Numeric values   
Years In Current Role: Numeric values   
Years Since Last Promotion: Numeric values   
Years With Curr Manager: Numeric values

Another feature namely *BelowAverageIncome* was created based on the department the employee was working in, the average income of that department and *PercentSalaryHike* of the employee. If the monthly income of an employee was less than the average income of that department and the percentage salary hike was less than 16, the employee was given a grade 1 else 0, indicating that the employee was most likely to leave

**3. EDA Concluding Remark**  
Exploratory data analysis employs a variety of techniques (mostly statistical graphics) before making inferences from data. It is essential to examine all variables in the dataset to:

* Catch mistakes
* Generate hypotheses
* See patterns in the data
* Extract important variables
* Detect outliers and anomalies
* Gain deep familiarity with the dataset
* Refine selection of features that will be used to build the machine learning models  
    
  Find patterns in data through data visualization. Reveal hidden secrets of the data through graphs, analysis and charts.
  + Univariate analysis
    - Continous variables : Histograms, boxplots. This gives us understanding about the central tendency and spread
    - Categorical variable : Bar chart showing frequency in each category
  + Bivariate analysis
    - Continous&Continous : Scatter plots to know how continous variables interact with each other
    - Categorical & categorical : Stacked column chart to show how the frequencies are spread between two
    - categorical variables
    - Categorical &Continous : Boxplots, Swamplots or even bar charts
* Detect outliers

Feature engineering  
  
  
**4. Pre-Processing Pipeline.**  
.  
Before starting the process, its important to answer if it's clear what kind of problem we are dealing with, because in many cases isn't so simple to identy it. A good understanding of the problem will help to choose the right data mining and machine learning techniques to make the right predictions. Thus, the first step, is preprocessing the data to look for missing, incomplete or noise values, because, in real word, the raw datas can be collect from many sources like sensors, websites, public data and many others.

To start the step of preprossing the dataset is neccessary to import some useful Python libraries.

* Numpy: Is a fundamental package to use linear algebra and random number capabilities. See: www.numpy.org/
* Pandas: Is a package to work with relacional data as tables.
* importnumpyasnp *# linear algebra*
* import pandas aspd *# data processing*
* Importing necessary modules  
  importmatplotlib.pyplotasplt
* frommatplotlib.pyplotimport  
    
  Other useful method is **info** that shows a summary of the dataset, like number of observations, columns, variable type and the total memory usage. The dataset have 14999 observations, 10 columns and with no null values. The data types of the variables are divided in 2 float, 6 integer and 2 object.data.info()

MatplotMatplotlib: is a plotting lybrary, usefull to plot statistical graphics. See: www.matplotlib.org

Seaborn: is a library based on matplotlib that can draw attractive statistical graphics. See: seaborn.pydata.org/index.html lib: is a plotting library, useful to plot statistical graphics.   
Seaborn: is a library based on matplotlib that can draw attractive statistical graphics. See: seaborn.pydata.org/index.html

The correlation is a very useful statitiscal analysis that describes the degree of relationship between two variables. Let´s see the table below and the heat map to see what relationship are in the data.

In the heat map is possible to see:

* Negative correlation of (-0.39) between satisfaction\_level and the employees that left the company.
* The highest positive correlation is between number of projects and average monthly hours (0.42).
* Last\_evaluation is high correlated to number\_project(0.35)and average\_monthly\_hours(0.34).
  + Work\_accident have a low negative correlation(-0.15)and salary (-0.16) with employees that left.  
      
    It is a relatively young company, on average, employees have 3 or 4 years in the company and the oldest employees are working 10 years.
* The biggest difference in the salary from who stayed and those who left, was found in the managemnet department, in the others departments although the salaries of who stayed be higher in average, it is not a big difference.
* The number of employees that had a work accident is about 14%, of which only 169 employees left the company, so don't seem to have a correlation with the employees leaving.
* In five years only 2% of the employees were promoted. Is possible that many employees get unmotivated and start planning to leave.
* Employees with 7 or longer in the company didn't left. Employees with 5 years have more chances to leaving.
* There are 2 distincts groups of employees performance that left. A group with poor performance with 2 projects and others with high performance with 5 or more projects. It is not necessary retain all the employees, the focus is on keeping employees with high performance.
* The employees with 4 years in the company have the lowest average satisfaction level of all the company with (0.47).
* The satisfaction drops when the employees are working in 5 or more projects. A number of 3 or 4 projects seems to be ideal independent of the time spend in the company.
* The employees with 5 or more projects that left also worked at least 20% more hours than the average of the company.
* The satisfaction level of the employees that left is grouped in totally disappointed, below the average satisfaction and satisfied.

**5. Building Machine Learning Modules.**

The 7 Key Steps To Build Your Machine Learning Model :

Step 1: Collect Data

Step 2: Prepare the data

Step 3: Choose the model

Step 4 Train your machine model

Step 5: Evaluation

Step 6: Parameter Tuning  
Step 7: Prediction  
  
For the Hr attrition analytics , we have used multiple modules as following.  
  
**1. HR Analytics Attrition Dataset Analysis  
  
2. Importing necessary modules**

# **3.Reading the input to from a dataframe**

# **4.Check Attrition=Yes**

# **5.Comparative Analysis of Age and Number of persons on that Particular age group**

# **6.Filteration according to Department**

# **7.Pie Representation on filteration**

# **8.Research Members of the Company in Research and Development**

**6.Conclusion**:  
  
The key to successful HR analytics relies on the understanding that the size of the measured data isn't the key to success, but rather, the **impact the data can have on decision-making** in the organization. The biggest value that HR analytics and metrics can help you with is that it **offers tangible data** that can help you shape the overall strategy of your business in a data-oriented, people-centric way. Having analytics tools that help you better measure your metrics is the way forward in HR efforts.  
Based on the analysis, one can clearly notice a pattern. The employees are more concerned with the materialistic objects that they get directly in hand. Then comes the psychological variables that determines if an employee might leave the organization. Hence, the HR can focus on such aspects and understand from the viewpoint of the employees. This can immensely help the organization.  
  
The HR Department can focus on the important variables that contribute significantly in determining if an employee is going to leave an organization. Such variables are:

· Stock Option Level

· Monthly Income

· Job Satisfaction

· Job Involvement

· Work Life Balance

· Environment Satisfaction

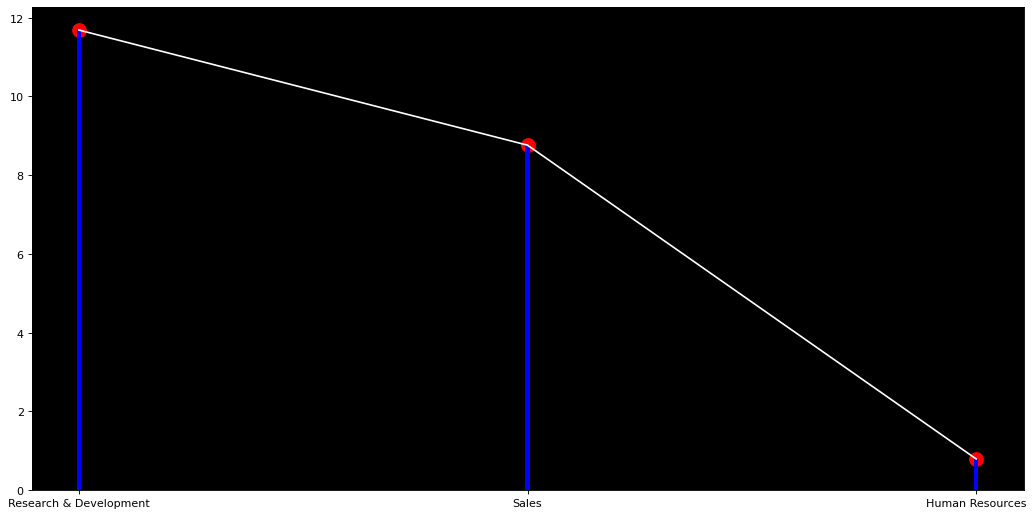
We can conclude that StockOptionLevel plays a very important role in deciding the attrition of the employee. Apart from that, MonthlyIncome, JobSatisfaction, JobInvolvement also are among the top contributors. On the other hand, factors such as PerformanceRating, Gender,Department tend not to contribute as significantly.  
  
The Average age of the Research Members of the Company in the Research and Development department is: 33

The minimum age of the Research Members of the Company in the Research and Development department is: 18

The maximum age of the Research Members of the Company in the Research and Development department is: 58  
  
***>*** *Age*, *DailyRate*, *DistanceFromHome*, *HourlyRate*, *MonthlyRate*, *PercentSalaryHike* tend not to have any outliers.

***>****NumCompaniesWorked*, *TrainingTimesLastYear*, *YearsWithCurrManager*, *YearsInCurrentRole* have a moderate number of outliers.

***>*** *MonthlyIncome*, *TotalWorkingYears*, *YearsAtCompany*, *YearsSinceLastPromotion*have large number of outliers.

  
**1. From the pictorial representation it is clear that the employees of Research and Development department stays with the company for the long time.**

**2. The second position with the sales department and then last it comes with Human Resource department**

**Finally the Attrition rate will be higher in Research and Development sector**