

|  |
| --- |
| **EMPLOYEE TURNOVER HR DATA ANALYSIS USING PYTHON** |
| [Final Project] |
| Project Submitted By--------  Kishor Sarkar  Suvadip Mandal  Samiul Aziz  Jaydip Sarkar  Sanjib Mallick |
|  |



**Acknowledgements**

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. We would like to extend our sincere thanks to all of them.

We are highly indebted to Ms. Tania Chakrabortyfor their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

We would like to express my gratitude towards our parents & member of NIVT Skills for their kind co-operation and encouragement which help me in completion of this project.

We would like to express our special gratitude and thanks to industry persons for giving me such attention and time.

## *What is Employee Turnover?*

Employee turnover, or employee turnover rate, is the measurement of the number of employees who leave an organization during a specified time period, typically one year. While an organization usually measures the total number of employees who leave, turnover can also apply to subcategories within an organization like individual departments or demographic groups.

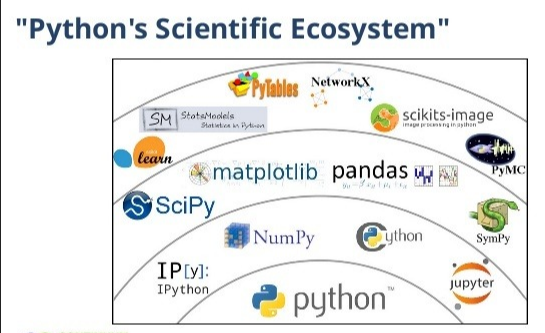
Businesses have long understood that it’s easier (and more cost-effective) to keep your current customers or clients happy than it is to acquire new customers. That’s why there are so many customer loyalty programs, designed to keep customers happy and coming back again and again.

The same holds true for employees. It’s not enough for a company to recruit the best employees – it also has to find a way to retain those employees for as long as possible in order to get the most value out of its recruitment efforts.

This isn’t news to most organizations (and especially not to HR professionals). And yet, holding on to top talent is something virtually every employer, regardless of industry, region or size, struggles to do.

**Why we use the Python Language to Employee Turnover Analytics?**

Python's expansive library of open source data analysis tools, web frameworks, and testing instruments make its ecosystem one of the largest out of any programming community.

Python is an accessible language for new programmers because the community provides many introductory resources Programming languages have unique ecosystems, cultures and philosophies built around them. You will find friction with a community and difficulty in learning if your approach to programming varies from the philosophy of the programming language you've selected. 

Python's culture values open source software, community involvement with local, national and international events and teaching to new programmers. If those values are also important to you and/or your organization then Python may be a good fit.

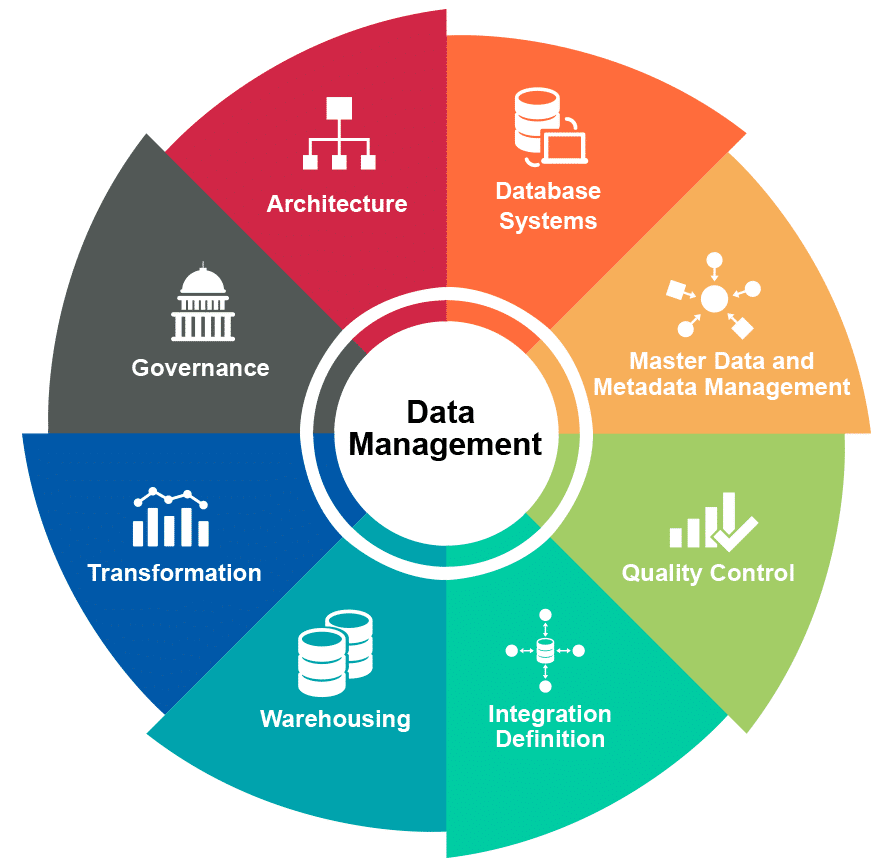
What is Data?

In computing, data is information that has been translated into a form that is efficient for movement or processing. Relative to today's computers and transmission media, data is information converted into binary digital form. It is acceptable for data to be used as a singular subject or a plural subject. Raw data is a term used to describe data in its most basic digital format.

**What Does Data Science Mean?**

****Data science is a multidisciplinary blend of datainference, algorithmm development, and technology in order to solve analytically complex problems. This aspect of data science is all about uncovering findings from data. Diving in at a granular level to mine and understand complex behaviors, trends, and inferences. It's about surfacing hidden insight that can help enable companies to make smarter business decisions.

**DATA MANAGEMENT**

Data management is the practice of organizing and maintaining data processes to meet ongoing information lifecycle needs. Emphasis on data management began with the electronics era of data processing, but data management methods have roots in accounting, statistics, logistical planning and other disciplines that predate the emergence of corporate computing in the mid-20th century.

There are mainly three kinds of Data

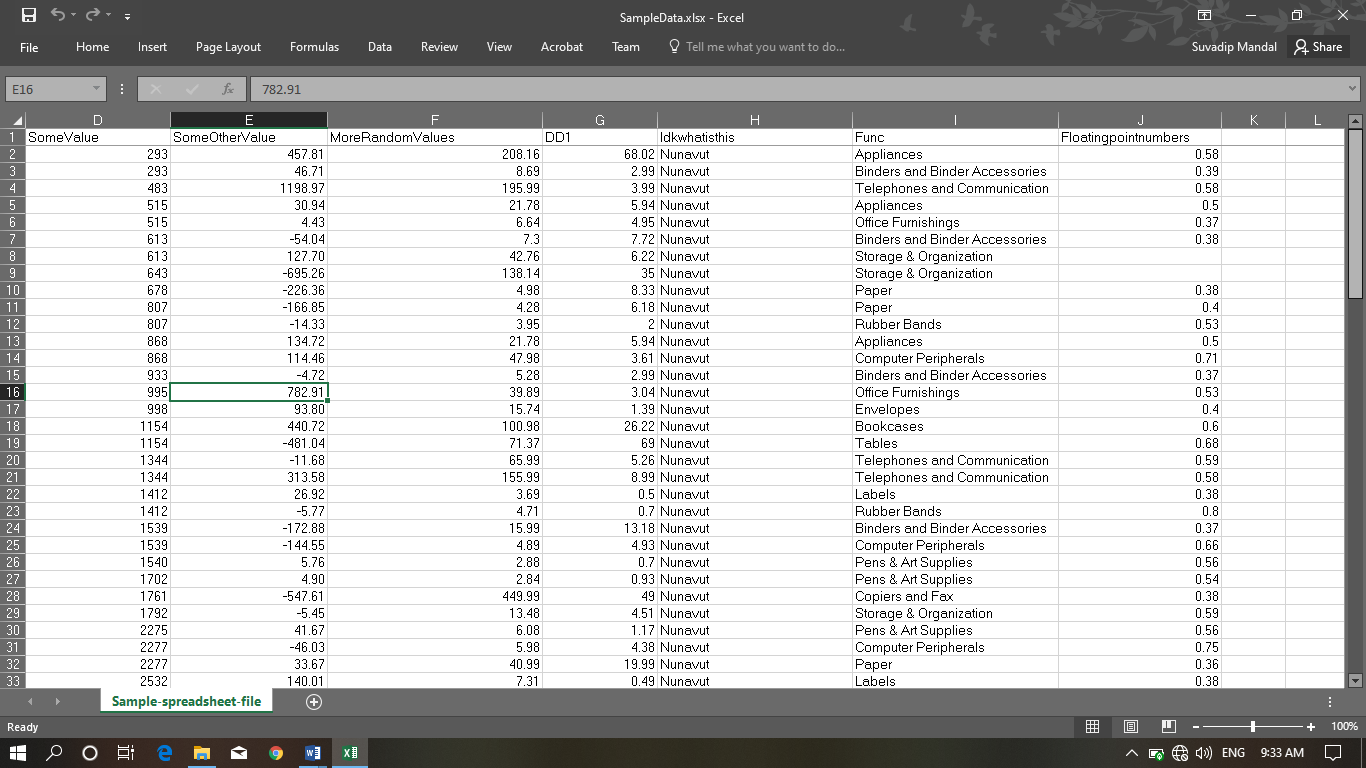
1) Structured Data

2) Unstructured Data

3) Semi Structured Data

**1)Structured Data:**

Structure Data is data that has been organized into aformatted repository, typically a database, so that itselements can be made addressable for more effective processing and analysis.

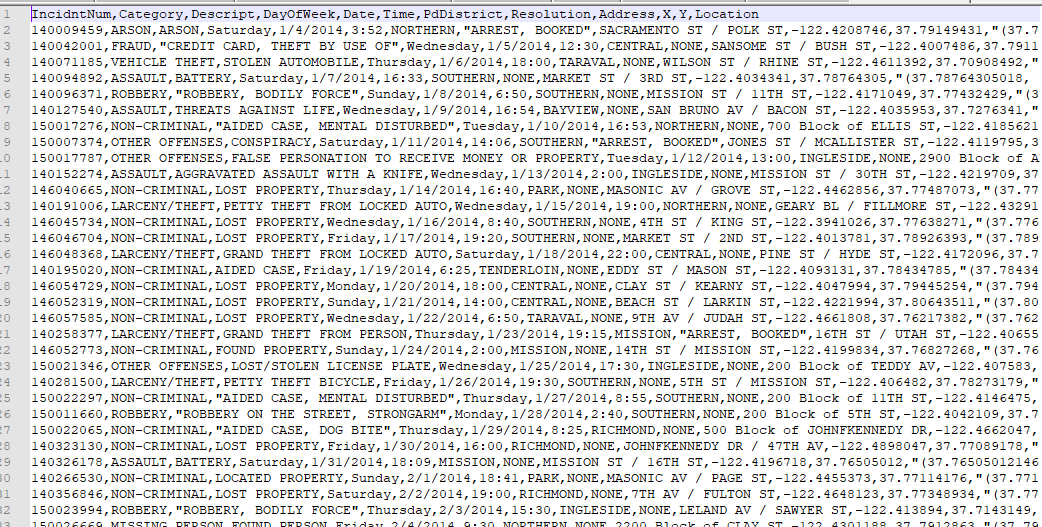


**2)Unstructured Data:**

The phrase Unstructured Data usually refers toinformation that doesn’t reside in a traditional row-column database.Unstructured Data files often include text and multimediacontent.Examples include e-mail messages, word processing, documents, videos, photos, audio files, presentations, textcorpus etc.



**3)Semi-Structured Data:**

Semi-structured data is data that is neither raw data, nor typed data in a conventional database system. Eg-CSV(comma Separated Value), XML, JSON (JavaScript Object Notation).

**Data Management Tools**

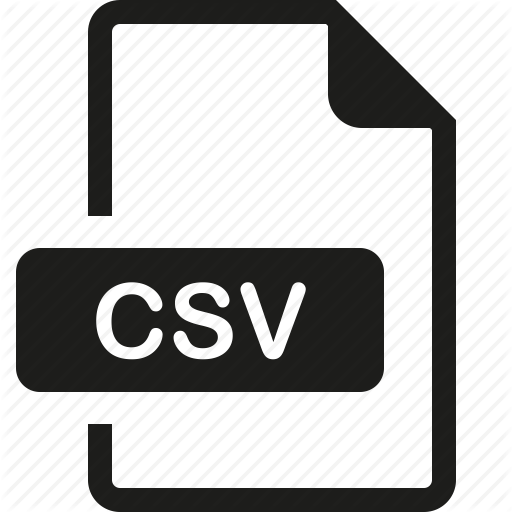
Good data management enables the location, sharing, and reuse of data, and reduces the redundancy of data. These attributes of good data management reduce costs in terms of time and money. Data management tools and best practices can also help to meet the open data requirements of making data more accessible and discoverable.

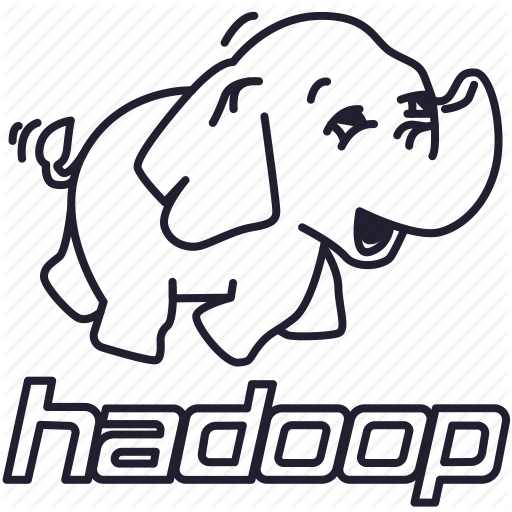
When there is a structured and reasonable amount ofdata, the data tools are used to manage that data arecommonly known as database.

*Like Oracle*, *MySQL, Sql, CSV/Excel*.

**Oracle:** Oracle Corporation is an American multinational computer technology corporation headquartered in Redwood Shores, California. The company sells database software and technology, cloud engineered systems, and enterprise software products—particularly its own brands of database management systems.

******MySQL:** MySQL is an open-source relational database management system. Its name is a combination of "My", the name of co-founder Michael Widenius's daughter, and "SQL", the abbreviation for Structured Query Language.

**CSV: CSV** is a simple file format used to store tabular data, such as a spreadsheet or database. Files in the **CSV** format can be imported to and exported from programs that store data in tables, such as Microsoft Excel or OpenOffice Calc. **CSV** stands for "comma-separated values"

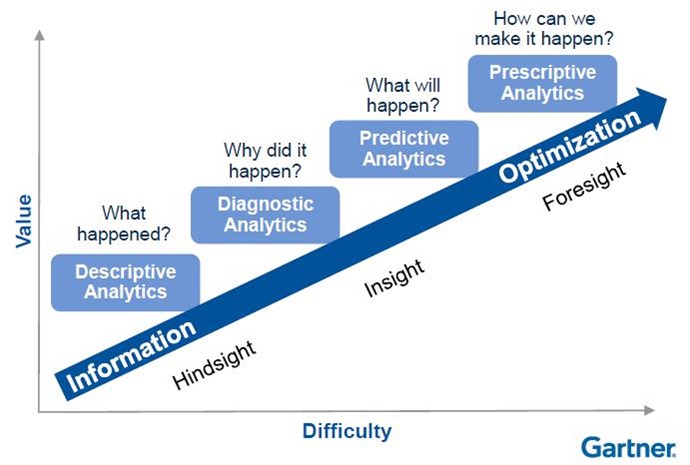
**Hadoop:** Apache Hadoop is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model.

**What is Data Analytics?**

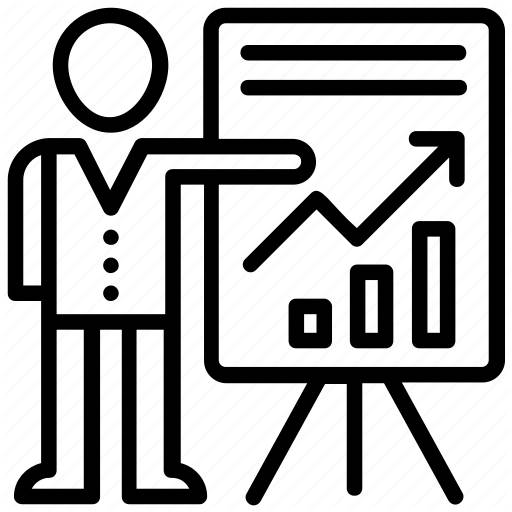
Data analytics (DA) is the process of examining [data](https://searchdatamanagement.techtarget.com/definition/data) sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software. Data analytics technologies and techniques are widely used in commercial industries to enable organizations to make more-informed business decisions and by scientists and researchers to verify or disprove scientific models, theories and hypotheses.

There are four types of data analytics,

1. *Descriptive Analytics.*
2. *Diagonostic Analytics.*
3. *Predictive Analytics.*
4. *Prescriptive Analytics.*

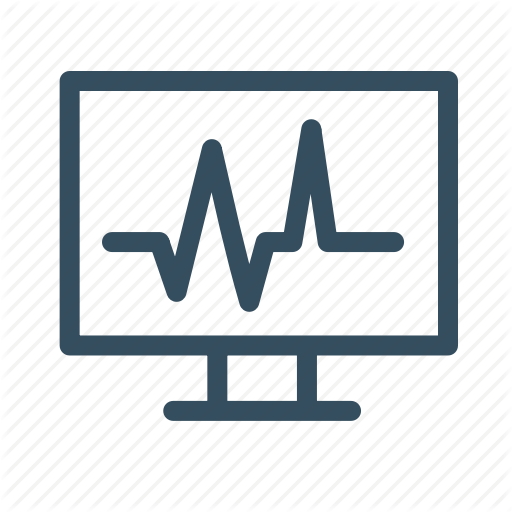
**

**1)Descriptive Analytics:**

******Descriptive analytics is a preliminary stage of data processing that creates a summary of historical data to yield useful information and possibly prepare the data for further analysis.

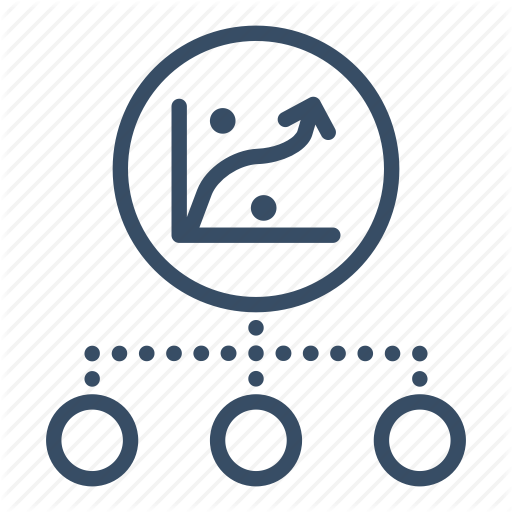
Descriptive analytics is sometimes said to provide information about happened.

**2)Diagnsostic Analytics:**

******Diagnostic Analytics is a form of advanced analytics which examines data or content to answer the question “Why did it happen?”, and is characterized by techniques such as drill-down, data discovery, data mining and correlations.

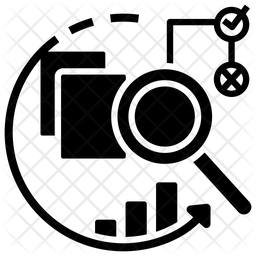
Diagnostic analytics is a deeper look at data to attempt to understand the causes of events and behaviors.

***3)Predictive Analytics.***

Predictive analytics is the practice of extracting information from existing data sets in order to determine patterns and predict future outcomes and trends.

Predictive analytics, which is used to identify future probabilities and trends, is said provide information about what might happen in the future.

***4) Analytics.***

Prescriptive analytics is the area of business analytics (BA) dedicated to finding the best course of action for a given situation. Prescriptive analytics is related to both descriptive and predictive analytics.

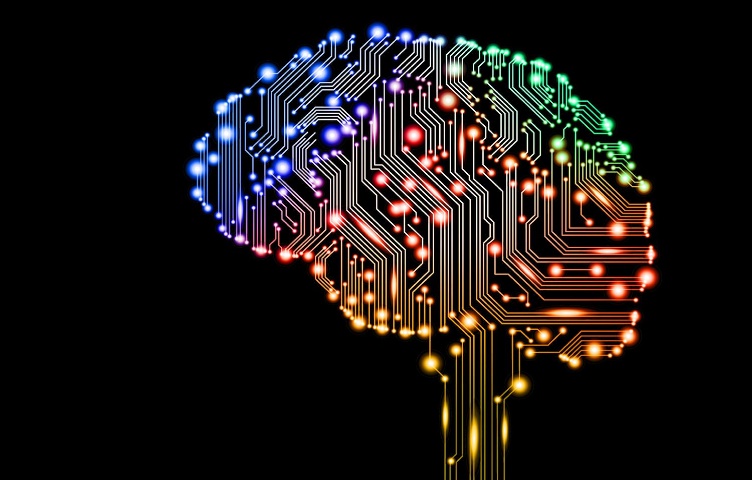
Prescriptive analytics is applied to try to identify the best outcome to events, given the parameters, and suggest decision options to best take advantage of a future opportunity or mitigate a future risk.

**Data Pre-processing**

Data processing occurs when data is collected and translated into usable information. Usually performed by a data scientist or team of data scientists, it is important for data processing to be done correctly as not to negatively affect the end product or data output.

Data processing starts with data in its raw form and converts it into a more readable format (graphs, documents, etc.), giving it the form and context necessary to be interpreted by computers and utilized by employees throughout an organization.

**ARTIFICIAL INTELIGENCE**

**Artificial intelligence (AI) makes it possible for machines to learn from experience, adjust to new inputs and perform human-like tasks. Most AI examples that you hear about today – from chess-playing computers to self-driving cars – rely heavily on deep learning and natural language processing. Using these technologies, computers can be trained to accomplish specific tasks by processing large amounts of data and recognizing patterns in the data.

**WHY WE ARE IMPLEMENTING THE CONCEPT OF DATA SCIENCE IN EMPLOYEE TURNOVER ANALYSIS?**

Data analysis *discovering useful information, informing conclusions, and supporting decision-making* is a process of *inspecting, cleansing, transforming, and modeling data* with the goal of. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains. *In today's business, data analysis is playing a role in making decisions more scientific and helping the business achieve effective operation.*

Human Resourcedata is such a big data with varse number of entries like *15,000 employee observations and 10 features, and* HR analytics demonstrates the causal relationship between the activities exacted by an HR department and the business outcomes that result from this activity. Once this causal relationship is discovered, HR analytics professionals will devise and implement a strategic plan to assist the business in attaining better outcomes, to reach their goal.

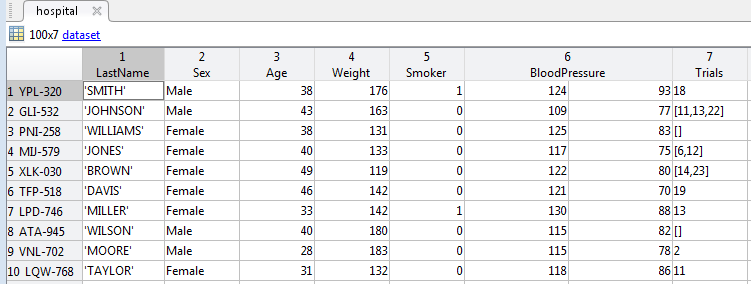
So, this is how we implement Data Science in our HR Analytics domain.

**Dataset**

The Dataset here used, was provided to us by our Mentor Tania Chakraborty of the NIVT training institute. The data is primarily a raw data of human resource records, we analyses data using python.

This data set is said to be original data entries from much known American multinational information technology company, International Business Machines Corporation (IBM). Who have dealt operations in over 170 countries?

In this data analysis report we are *using python as our data analysis tool.* Because, python is a high-level programming language designed to be easy to read and simple to implement. It is open source, which means it is free to use, even for commercial applications. Python can run on Mac, Windows, and UNIX systems and has also been ported to Java and .NET virtual machines.



**Jupyter Notebook:**

****The **Jupyter Notebook** is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more

**Installing Jupyter Notebook**

[Installing Jupyter using Anaconda and conda](https://jupyter.readthedocs.io/en/latest/install.html#id3)

For new users, we **highly recommend** [installing Anaconda](https://www.anaconda.com/download). Anaconda conveniently installs Python, the Jupyter Notebook, and other commonly used packages for scientific computing and data science.

Use the following installation steps:

1. Download [Anaconda](https://www.anaconda.com/download). We recommend downloading Anaconda’s latest Python 3 version (currently Python 3.5).
2. Install the version of Anaconda which you downloaded, following the instructions on the download page.
3. Congratulations, you have installed Jupyter Notebook. To run the notebook:

jupyter notebook

[*Alternative for experienced Python users:* Installing Jupyter with pip](https://jupyter.readthedocs.io/en/latest/install.html#id4)

**Important**

Jupyter installation requires Python 3.3 or greater, or Python 2.7. IPython 1.x, which included the parts that later became Jupyter, was the last version to support Python 3.2 and 2.6.

As an existing Python user, you may wish to install Jupyter using Python’s package manager, [pip](https://jupyter.readthedocs.io/en/latest/glossary.html#term-pip), instead of Anaconda.

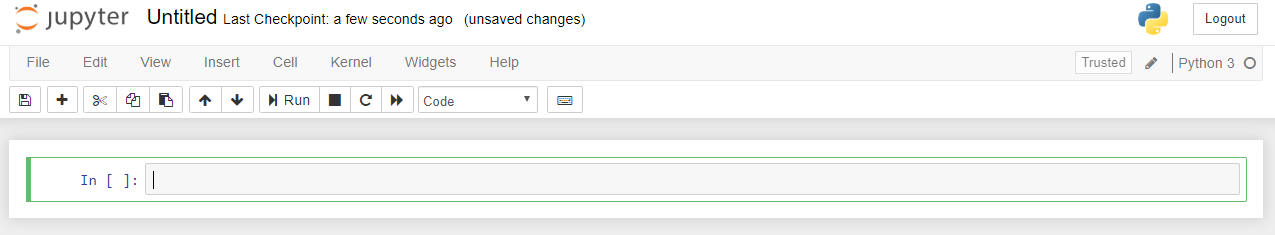
First, ensure that you have the latest pip; older versions may have trouble with some dependencies:

pip3 install --upgrade pip

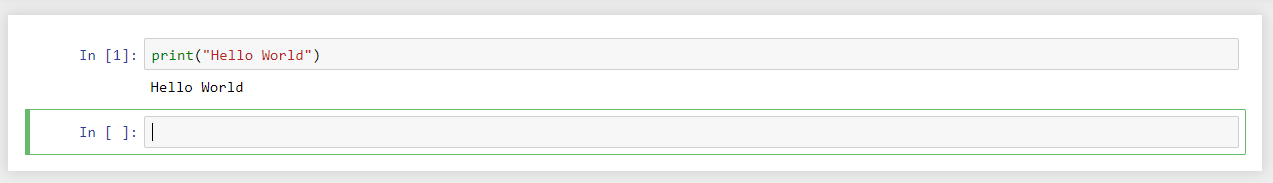
Then install the Jupyter Notebook using:

pip3 install jupyter

Now to implement the codes needed for analysis we need a platform that is *Jupyter Notebook* here.



My First Notebook

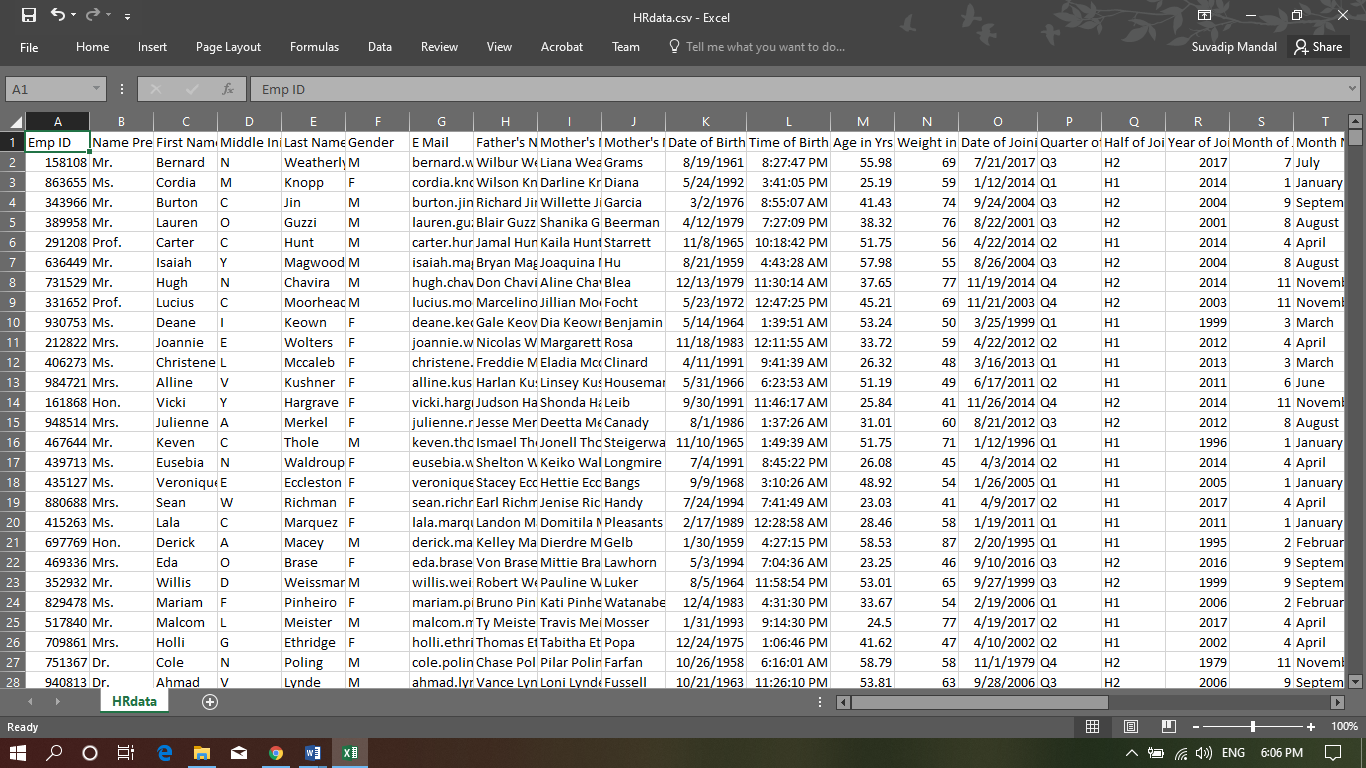


**The Setting**

Now let’s say IBM has hired you as a data scientist to help them make a business decision.

Here’s the question they want answered:

1. Find out the Data Description.
2. Find out the Statistical Description of the data.
3. Find out the correlation matrix of all the variable present in the dataset and visualize the matrix.
4. Plot counts the number of employee who left the company and who didn’t.
5. Count plot to check the number of employees with respect to Salary and Department feature and grouping together
6. Plot a diagram to check the number of employees left with respect to Salary and Department feature and grouping together.
7. How we can predict that an employee’s is salary with respect to satisfaction level, years at company, project count.

To help you make this decision, the company has provided you with HRdata.csv file that contains statistics on how employees are behaving in such manner.

**Implementation:**

Firstly We Import Python Libraries which one we want to use for in this data Analysis so libraries are ‘Numpy’ for Data Procession, ‘Pandas’ for Linear Algebra, ‘Matplotlib.Pyplot’ for Data Visualization , ‘Seaborn’ for Statistical Plotting Creation and for ignore Warnings *we have imported ‘*warnings’*.* Also there are some more Libraries which we imported. Otherwise we can’t do any Statistics Operation in this dataset they are.

‘Sklearn.model\_selection’,’sklearn.linear\_model’,’train\_test\_split’,’LogisticRegression’,’Warning’, we import form sklearn.ensemble’ ‘RandomForestClassifier’, ’sklearn.metrics’ use for  several loss, score, and utility functions to measure classification performance. Some **metrics** might require probability estimates of the positive class, confidence values, or binary decisions values, and supported vector machine as SVM

**Pandas**

Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with 'relationa' or 'labeled' data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python.

**Pandas is well suited for many different kinds of data:**

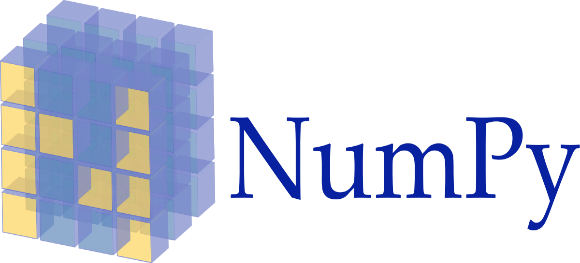
* Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet
* Ordered and unordered (not necessarily fixed-frequency) time series data.
* Arbitrary matrix data with row and column labels
* Any other form of observational / statistical data sets.

**Pandas Basic commands:**

|  |
| --- |
| import pandas as pd |

**Numpy**

“A NumPy tutorial for beginners in which you'll learn how to create a NumPy array, use broadcasting, access values, manipulate arrays, and much more.”

NumPy is, just like SciPy, Scikit-Learn, Pandas, etc. one of the packages that you just can’t miss when you’re learning data science, mainly because this library provides you with an array data structure that holds some benefits over Python lists, such as: being more compact, faster access in

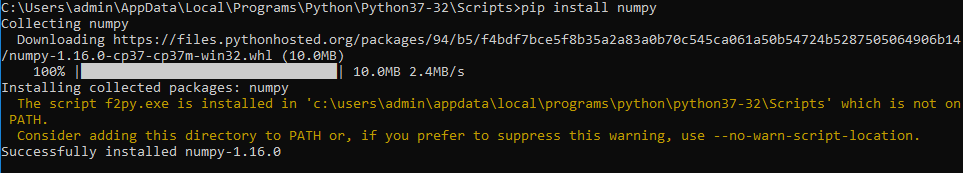
Reading and writing items, being more convenient and more efficient.

NumPy is the fundamental package for scientific computing with Python. It contains among other things:

* a powerful N-dimensional array object
* sophisticated (broadcasting) functions
* tools for integrating C/C++ and Fortran code
* useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

**Getting Started**

To install NumPy, we strongly recommend using a scientific Python distribution. See [Installing the SciPy Stack](https://www.scipy.org/install.html) for details.

Many high quality online tutorials, courses, and books are available to get started with NumPy. For a quick introduction to NumPy we provide the [NumPy Tutorial](https://www.numpy.org/devdocs/user/quickstart.html). We also recommend the [SciPy Lecture Notes](https://scipy-lectures.org/) for a broader introduction to the scientific Python ecosystem.

## Support NumPy

Institutional Partners are organizations that support the project by employing NumPy contributors, with contributing to the project as part of their official duties. Current Institutional Partners include:



# **Matplotlib**

Matplotlib is a python library used to create 2D graphs and plots by using python scripts. It has a module named pyplot which makes things easy for plotting by providing feature to control line styles, font properties, formatting axes etc. It supports a very wide variety of graphs and plots namely - histogram, bar charts, power spectra, error charts etc. It is used along with NumPy to provide an environment that is an effective open source alternative for MatLab. It can also be used with graphics toolkits like PyQt and wxPython.

## Matplotlib Example:-

import numpy as np

import matplotlib.pyplot as plt

# Compute the x and y coordinates for points on a sine curve

x = np.arange(0, 3 \* np.pi, 0.1)

y = np.sin(x)

plt.title("sine wave form")

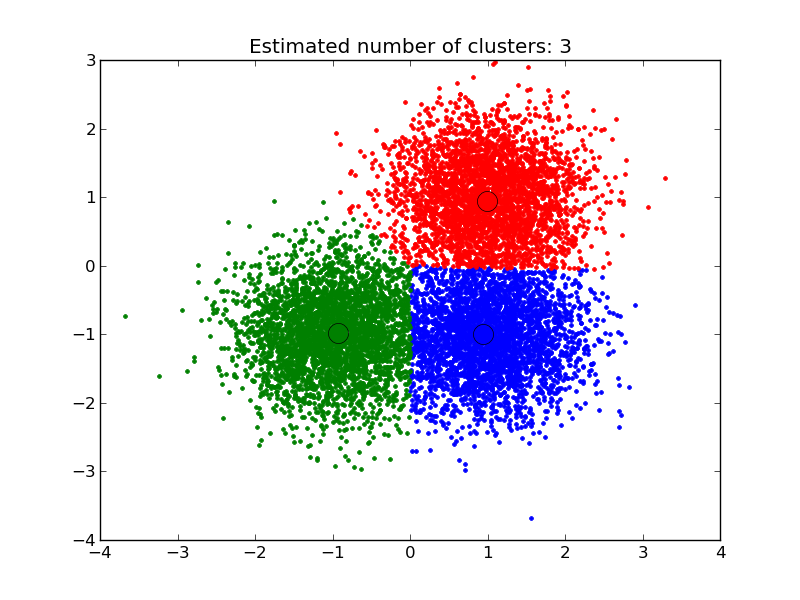
# Plot the points using matplotlib

plt.plot(x, y)

plt.show()

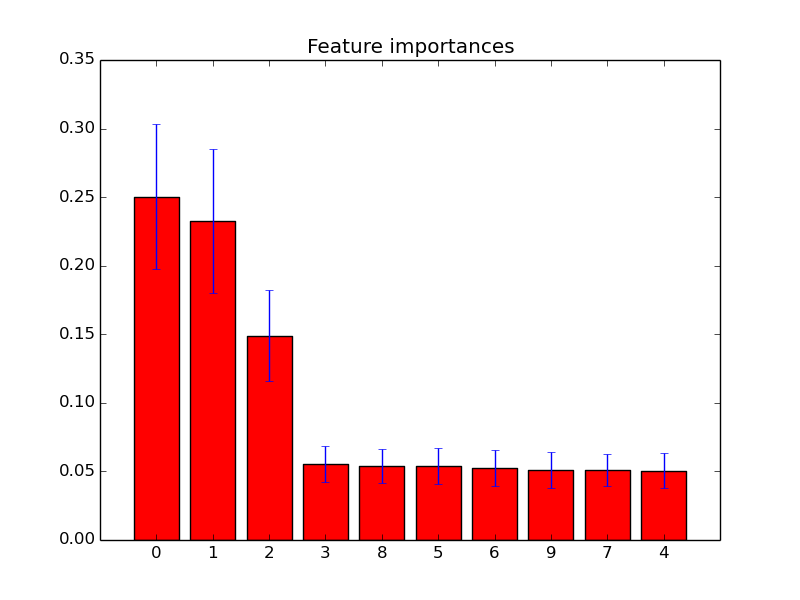
## Sine WaveIts output is as follows –

# **Scikit-learn**

Scikit-learn is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbours, and it also supports Python numerical and scientific libraries like NumPy and SciPy.

To do this, we’ll be using the [Sales\_Win\_Loss data](https://www.ibm.com/communities/analytics/watson-analytics-blog/sales-win-loss-sample-dataset/) set from IBM’s Watson repository. We will import the data set using pandas, explore the data using pandas methods like head(), tail(), dtypes(), and then try our hand at using plotting techniques from Seaborn to visualize our data.

**Sklearn\_Feature Selection**

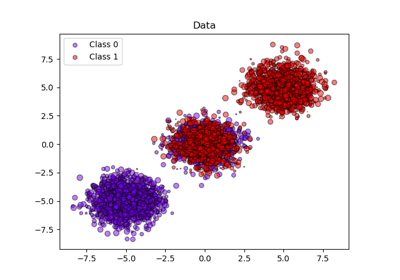
Feature selection is a process where you automatically select those features in your data that contribute most to the prediction variable or output in which you are interested.

Having irrelevant features in your data can decrease the accuracy of many models, especially linear algorithms like linear and logistic regression.

Three benefits of performing feature selection before modeling your data are:

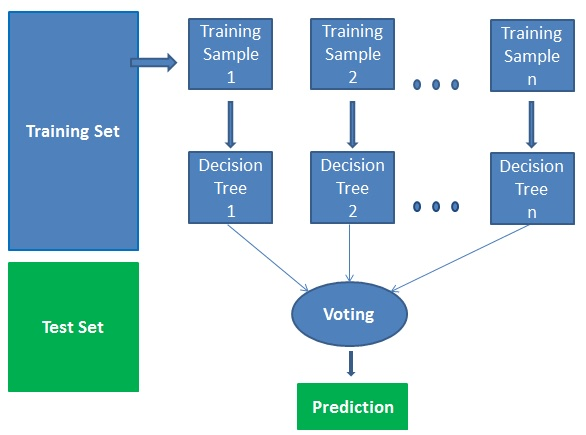
* **Reduces Overfitting**: Less redundant data means less opportunity to make decisions based on noise.
* **Improves Accuracy**: Less misleading data means modeling accuracy improves.
* **Reduces Training Time**: Less data means that algorithms train faster.

[**sklearn.model\_selection**](https://scikit-learn.org/stable/modules/classes.html#module-sklearn.model_selection)**.train\_test\_split**

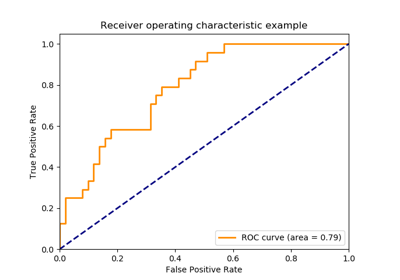
Split arrays or matrices into random train and test subsets.Quick utility that wraps input validation and next(ShuffleSplit().split(X, y)) and application to input data into a single call for splitting (and optionally subsampling) data in a oneliner.

sklearn.model\_selection.train\_test\_split

# **Random Forests Classifiers**

Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.

Random forests has a variety of applications, such as recommendation engines, image classification and feature selection. It can be used to classify loyal loan applicants, identify fraudulent activity and predict diseases. It lies at the base of the Boruta algorithm, which selects important features in a dataset.

**Sklearn metrics**

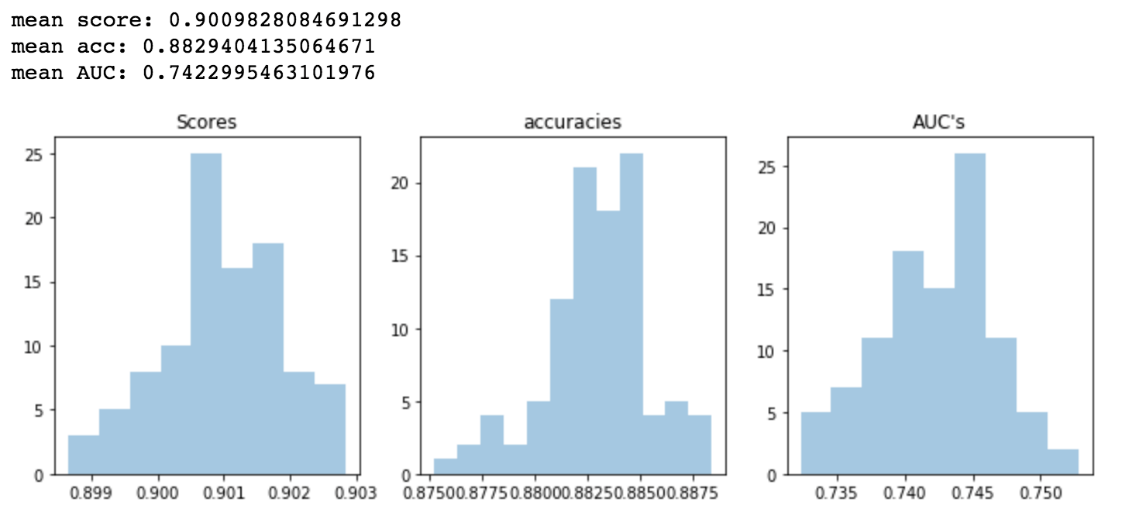
The module [**sklearn.metrics**](https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics) also exposes a set of simple functions measuring a prediction error given ground truth and prediction:

Functions ending with \_score return a value to maximize, the higher the better.

functions ending with \_error or \_loss return a value to minimize, the lower the better. When converting into a scorer object using [**make\_scorer**](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.make_scorer.html#sklearn.metrics.make_scorer), set the greater\_is\_better parameter to False.

Many metrics are not given names to be used as scoring values, sometimes because they require additional parameters, such as [**fbeta\_score**](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.fbeta_score.html#sklearn.metrics.fbeta_score). In such cases, you need to generate an appropriate scoring object. The simplest way to generate a callable object for scoring is by using [**make\_scorer**](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.make_scorer.html#sklearn.metrics.make_scorer). That function converts metrics into callables that can be used for model evaluation.

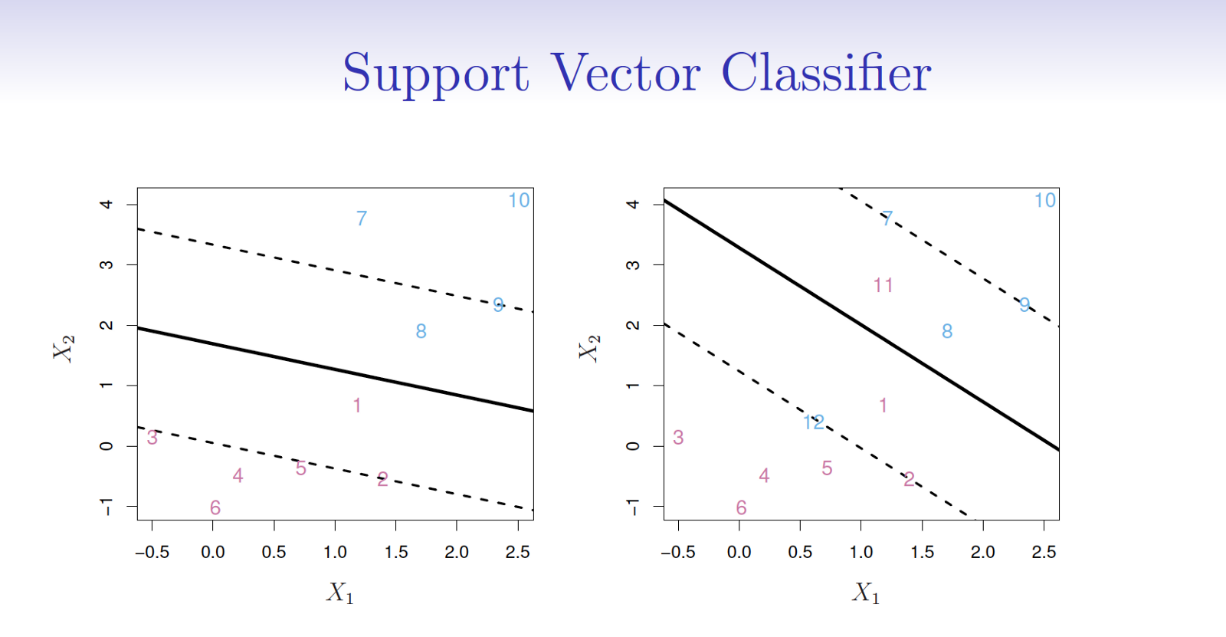
**Sklearn.metrics accuracy\_score**

The [**accuracy\_score**](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html#sklearn.metrics.accuracy_score) function computes the [accuracy](https://en.wikipedia.org/wiki/Accuracy_and_precision), either the fraction (default) or the count (normalize=False) of correct predictions.

In multilabel classification, the function returns the subset accuracy. If the entire set of predicted labels for a sample strictly match with the true set of labels, then the subset accuracy is 1.0; otherwise it is 0.0.

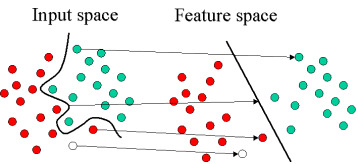
If y^ is the predicted value of the i-th sample and yi is the corresponding true value, then the fraction of correct predictions over nsamples is defined as

**SVC (Support Vector Classifier)**

****The most applicable machine learning algorithm for our problem is [Linear SVC](http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html). Before hopping into Linear SVC with our data, we're going to show a very simple example that should help solidify your understanding of working with Linear SVC.

The objective of a Linear SVC (Support Vector Classifier) is to fit to the data you provide, returning a "best fit" hyperplane that divides, or categorizes, your data. From there, after getting the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is. This makes this specific algorithm rather suitable for our uses, though you can use this for many situations.

**SVM (Support vector machines)**

It is a set of supervised learning methods used for [classification](https://scikit-learn.org/stable/modules/svm.html#svm-classification), [regression](https://scikit-learn.org/stable/modules/svm.html#svm-regression) and [outliers detection](https://scikit-learn.org/stable/modules/svm.html#svm-outlier-detection).

The advantages of support vector machines are:

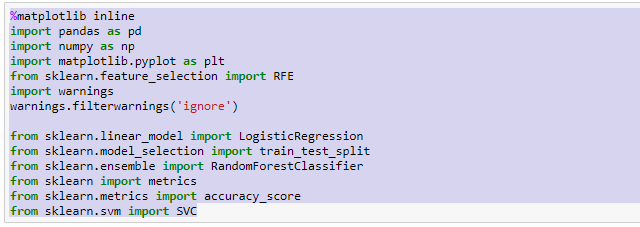
* Effective in high dimensional spaces.
* Still effective in cases where number of dimensions is greater than the number of samples.
* Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
* Versatile: different [Kernel functions](https://scikit-learn.org/stable/modules/svm.html#svm-kernels) can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector machines include:

* If the number of features is much greater than the number of samples, avoid over-fitting in choosing [Kernel functions](https://scikit-learn.org/stable/modules/svm.html#svm-kernels) and regularization term is crucial.
* SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation (see [Scores and probabilities](https://scikit-learn.org/stable/modules/svm.html#scores-probabilities), below).

**Project**

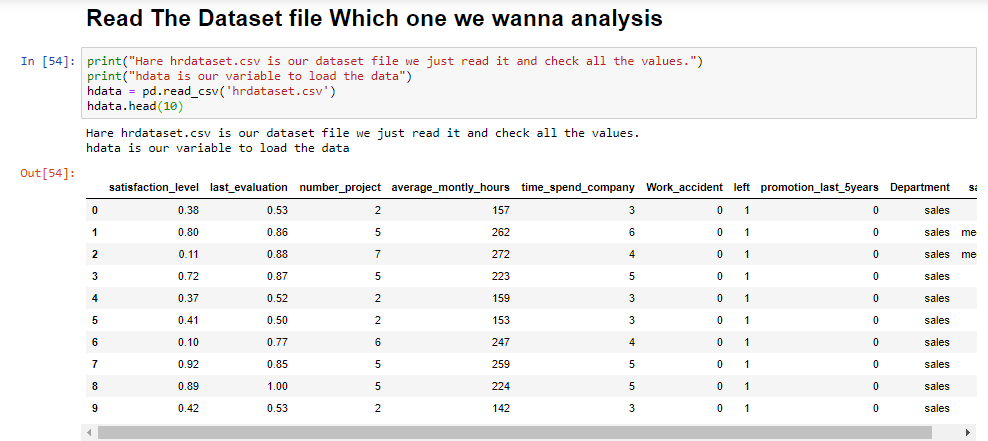
**Employee Turnover Data Analysis with Python**



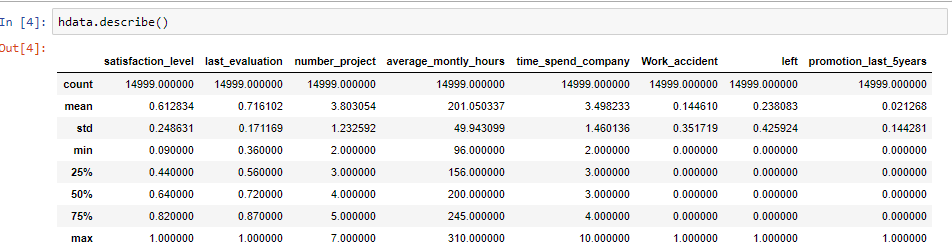
**Load the Dataset file**

After Imported all necessary Libraries we import the dataset CSV file ‘**hdataset.csv’** using the **.read\_csv ()** method from pandas ‘pd’ and store it in a dataFrame Name ‘hdata’

Because it’s a huge dataset file for this we use head() function to check some of data from this dataset.



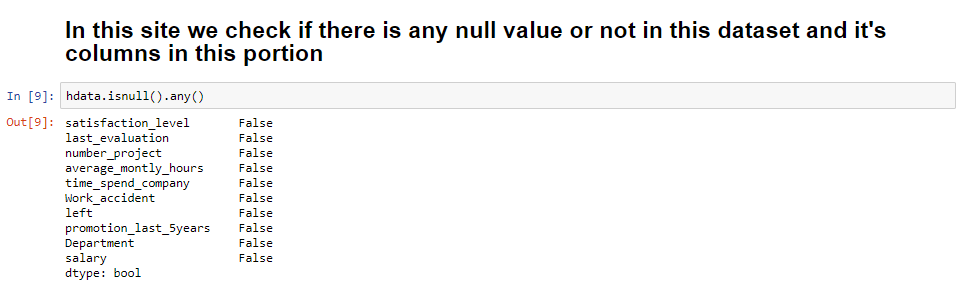
# Useing describe() method will give us some basic summary statistics about various fields of the dataset. so in this section we Run this describe() fucntion in this dataset.



# 

# 

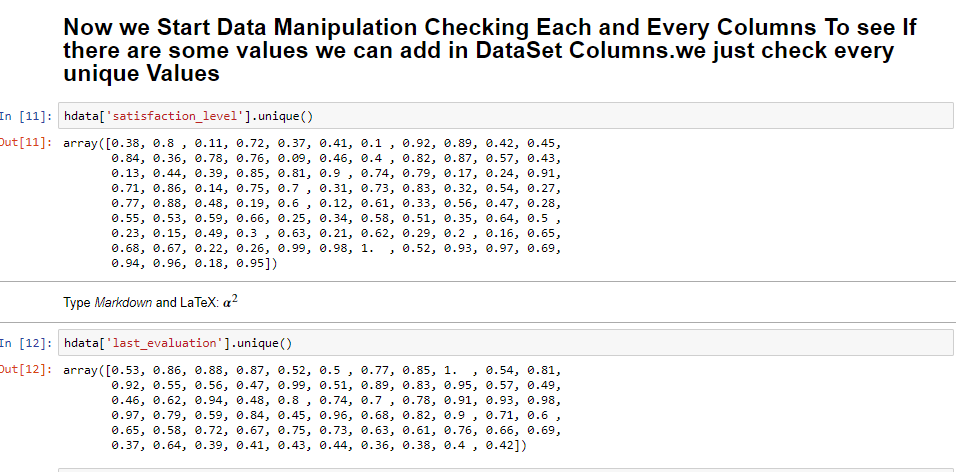
# 

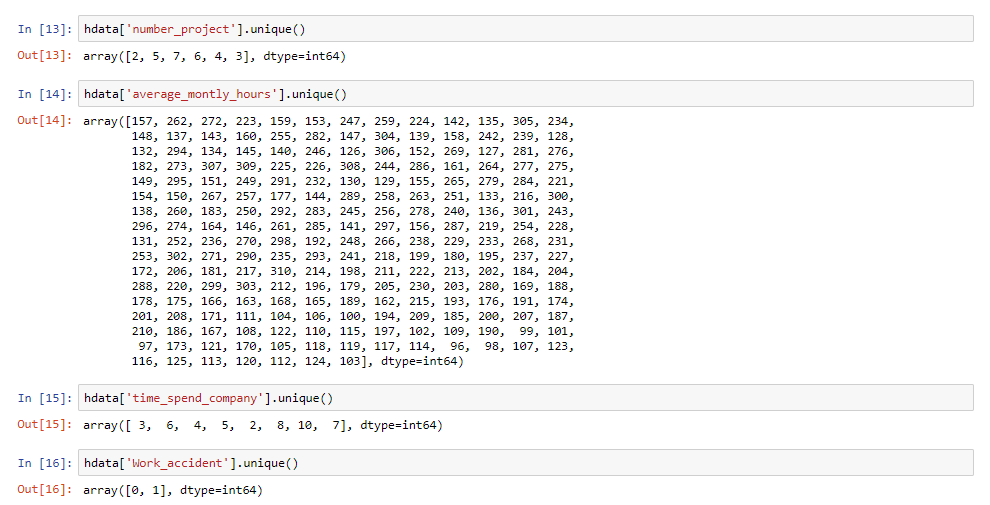


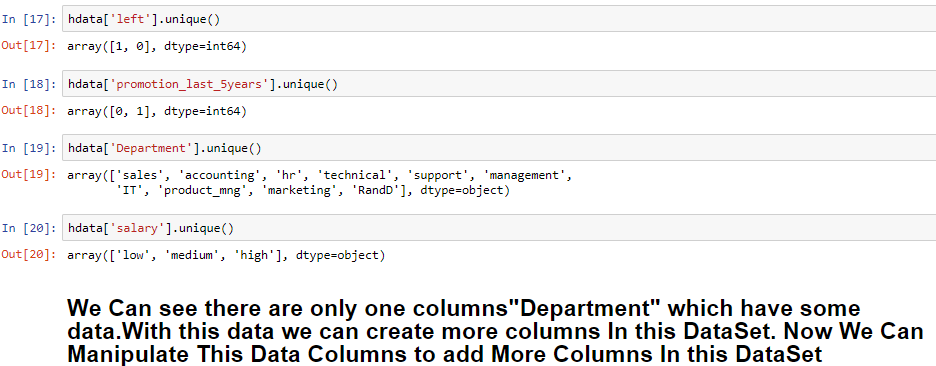
# Check dataset shape using shape method . After using shape you can see there are 15000 columns and 10 rows

# 

# Now we Start Data Manipulation Checking Each and Every Columns to see If there are some values we can add in Dataset Columns. We just check every unique () Values





**

You can see there are two columns that have unique values they are “Department”,”Salary” then we manupulate this two columns.

# There we must need a python library which Name is Numpy. We previously import Numpy as np at the first portion of this project

We check for the unique values for “Department Columns” Which are



There are RandD","accounting","hr","management","marketing", "product\_mng","sales","Technical" total eight(8) Data Unique Values

That we can manipulate

