**Data Science with Python**

Dept. of Computer Science & Information Systems

Bradley university, Peoria, IL, USA

Submitted to,

Dr. **Christos Nikolopoulos**, Ph.D.

By,

Sai Prasad Ashila

January Interim - 2019

**ACKNOWLEDGEMNET**

It Brings me Great Pleasure for an opportunity to work & submit my Course Report on Introduction to data science with python.

For this I deeply Indebt & sincerely Thankful to **Christos Nikolopoulos** for his help, valuable guidance and elating encouragement throughout the course.

Finally, I am deeply thankful to all the sources my seniors and peers who helped and inspired me in completing this course.

Table of Contents

[1. INTRODUCTION 4](#_Toc535598627)

[1.1 Data Science 4](#_Toc535598628)

[1.2 Python 4](#_Toc535598629)

[1.3 Building a Predictive Model in Python 4](#_Toc535598630)

[2. Installation setup 5](#_Toc535598631)

[2.1 Anaconda Installation 5](#_Toc535598632)

[3. Simple Linear Regression 5](#_Toc535598633)

[3.1 Analysis of the Problem 5](#_Toc535598634)

[3.2 Graphical View of data 6](#_Toc535598635)

[3.3 Importing the data set 7](#_Toc535598636)

[3.4 Splitting the data into training and test Sets 8](#_Toc535598637)

[3.4 Fitting Simple Linear Regression to the Training set 8](#_Toc535598638)

[3.6 Predicting the Test set Results 9](#_Toc535598639)

[3.7 Visualizing the Training set results 10](#_Toc535598640)

# 1. INTRODUCTION

## 1.1 Data Science

Data science is an [interdisciplinary](https://en.wikipedia.org/wiki/Interdisciplinary) field that uses scientific methods, processes, algorithms and systems to extract [knowledge](https://en.wikipedia.org/wiki/Knowledge) and insights from [data](https://en.wikipedia.org/wiki/Data) in various forms, both structured and unstructured, similar to [data mining](https://en.wikipedia.org/wiki/Data_mining). Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the context of [mathematics](https://en.wikipedia.org/wiki/Mathematics), [statistics](https://en.wikipedia.org/wiki/Statistics), [information science](https://en.wikipedia.org/wiki/Information_science), and [computer science](https://en.wikipedia.org/wiki/Computer_science).

## 1.2 Python

Python is an increasingly popular tool for data analysis. In recent years, a number of libraries have reached maturity, allowing R and Stata users to take advantage of the beauty, flexibility, and performance of Python without sacrificing the functionality these older programs have accumulated over the years. Python is a very powerful programming language used for many different applications. Over time, the huge community around this open source language has created quite a few tools to efficiently work with Python. In recent years, a number of tools have been built specifically for data science. As a result, analyzing data with Python has never been easier. Python is a general-purpose programming language that is becoming more and more popular for doing data science. Companies worldwide are using Python to harvest insights from their data and get a competitive edge. Unlike any other Python tutorial, this course focuses on Python specifically for data science.

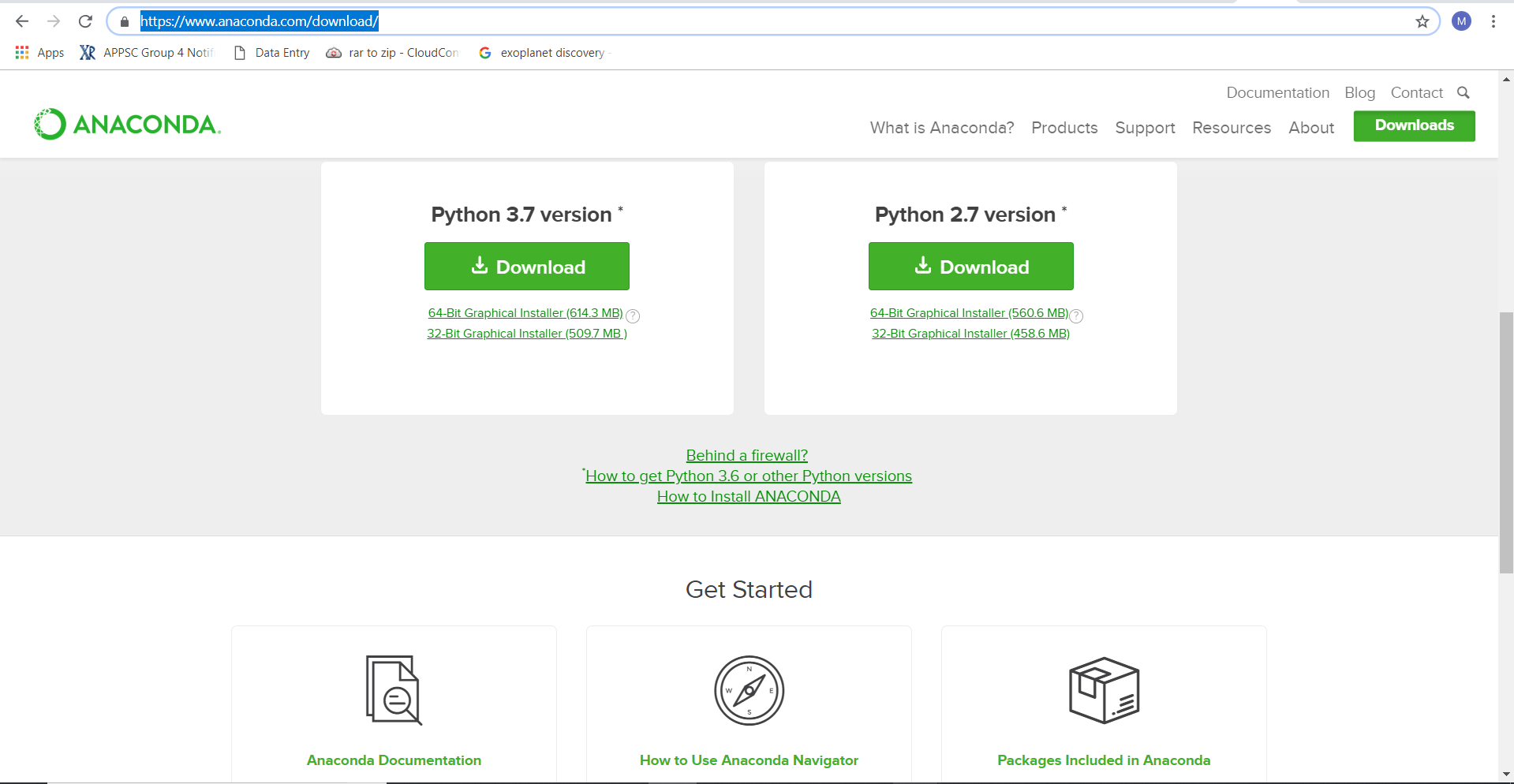
## 1.3 Building a Predictive Model in Python

Few Predicting algorithms are made using python which are the concepts of Machine learning and predictive analytics

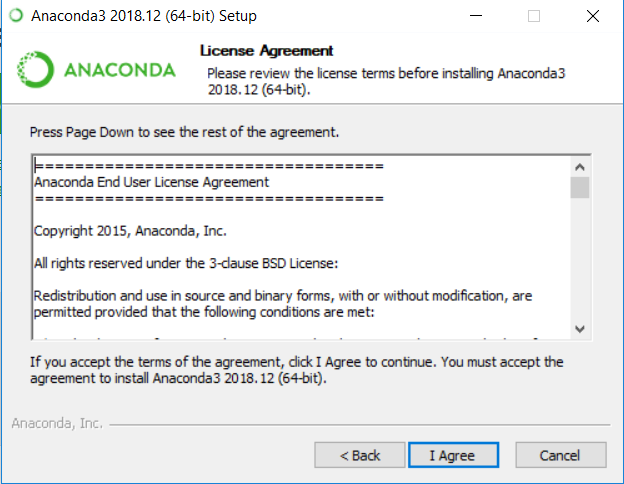
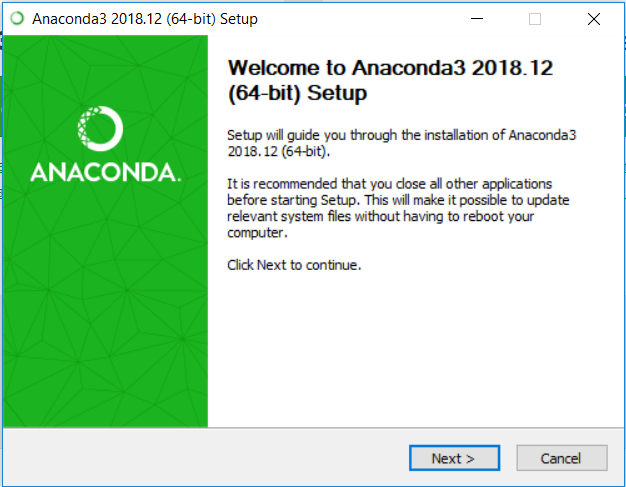
* Simple Linear Regression.
* Multiple linear Regression.
* Polynomial Regression.
* Support Vector Regression.
* Decision Tree Regression.
* Random Forest Regression.

# 2. Installation setup

## 2.1 Anaconda Installation

****

<https://www.anaconda.com/download/>

****

# 3. Simple Linear Regression

## 3.1 Analysis of the Problem

**Problem Statement:** To estimate the salary of the employee for a give particular years of experience based on the trend followed.

**Given:** Data of level of experience and corresponding salaries as shown below.

**Expectation of the regression:**  To estimate the salary for any given experience level.

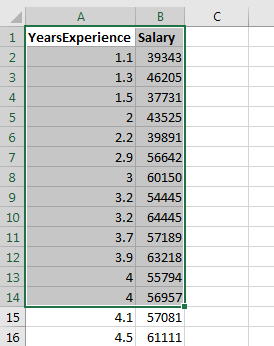
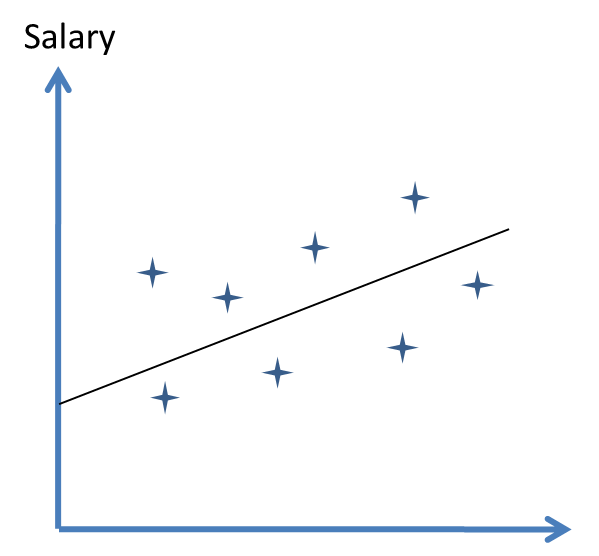
****

Figure 3.1

## 3.2 Graphical View of data



**Figure 3.1.2**

Salary = m(years of experience) + C

m is the regression coefficient

C the intercept indicates the salary of an entry level employee

## 3.3 Importing the data set

Importing the data Set using Pandas Library

1. **import** numpy as np
2. **import** pandas as pd
3. #Importing the data
4. dataset = pd.read\_csv('Salary\_Data.csv')
5. X = dataset.iloc[:,:-1].values
6. Y = dataset.iloc[:,1].values

Above code is used to import the libraries and import the data to the data sets variable and categorizing the dependent and independent variables

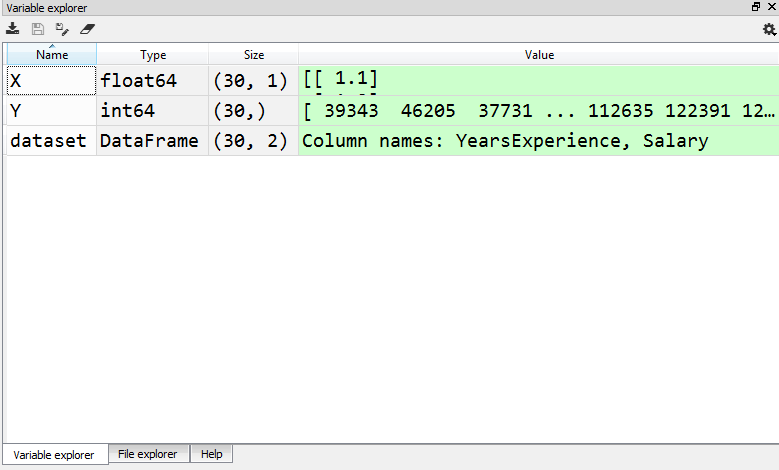
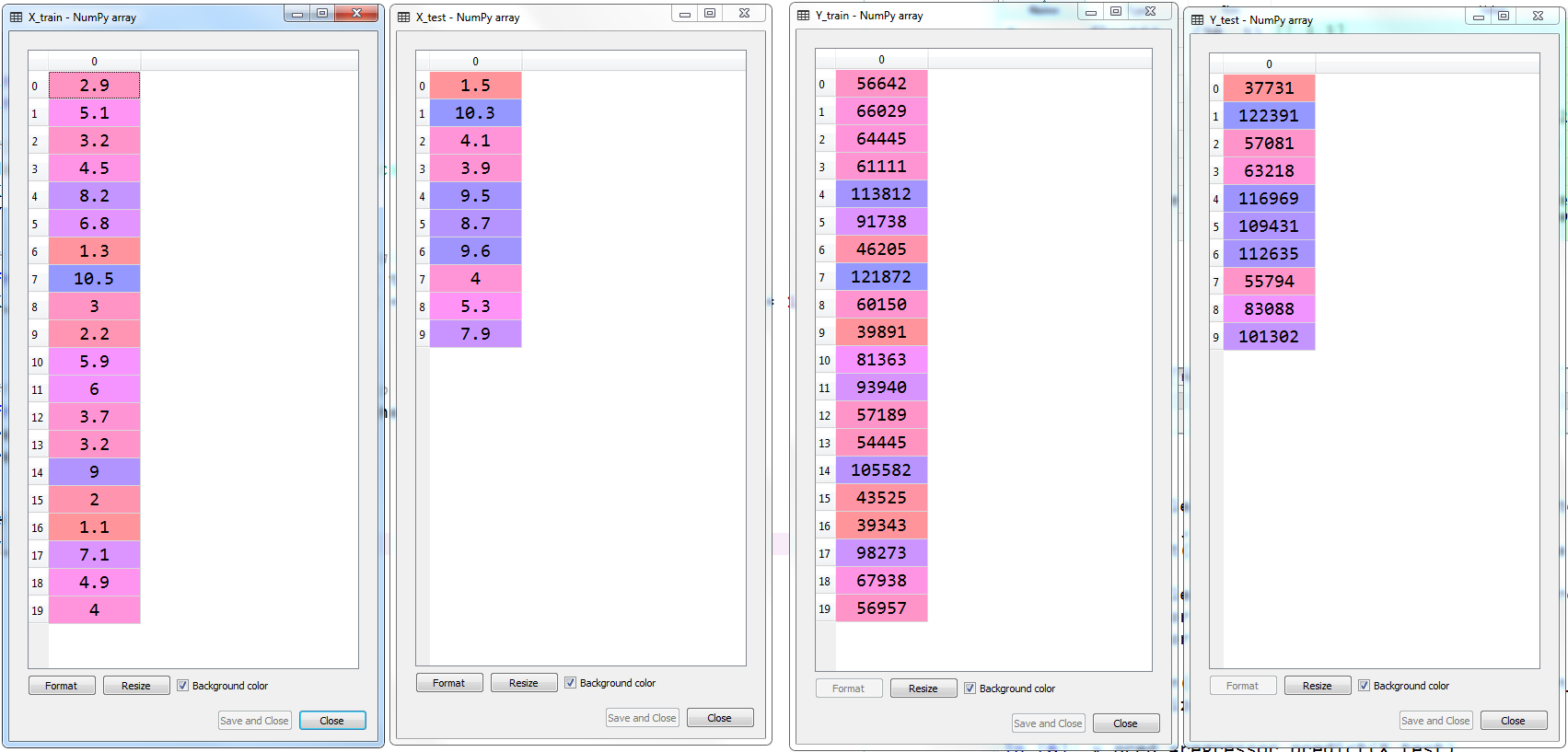


Fig: 3.1.3

## 3.4 Splitting the data into training and test Sets

1. # Splitting the dataset into the training
2. **from** sklearn.model\_selection **import** train\_test\_split
3. X\_train , X\_test , Y\_train , Y\_test =
4. train\_test\_split(X,Y, test\_size = 1/3, random\_state = 0)

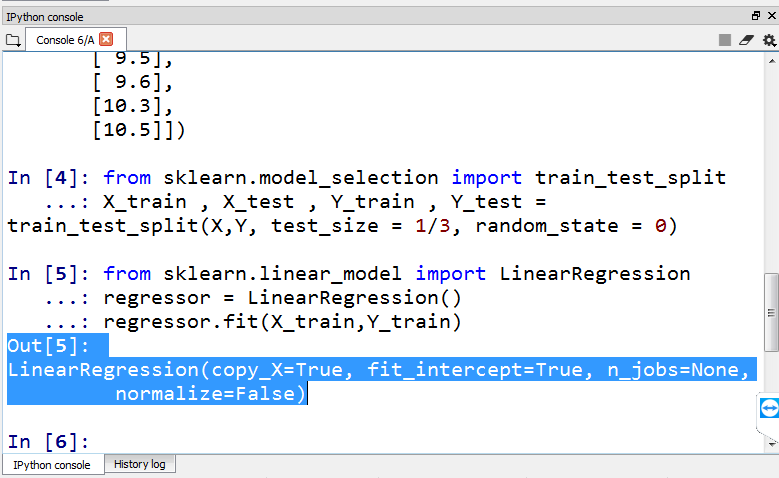
Importing the training set and splitting the given data into training set and test sets in some ratio here (1/3) to test set and 2/3 to training set



## 3.4 Fitting Simple Linear Regression to the Training set

1. #Fitting Simple Linear Regression to the
2. **from** sklearn.linear\_model **import** LinearRegression
3. regressor = LinearRegression()
4. regressor.fit(X\_train,Y\_train)

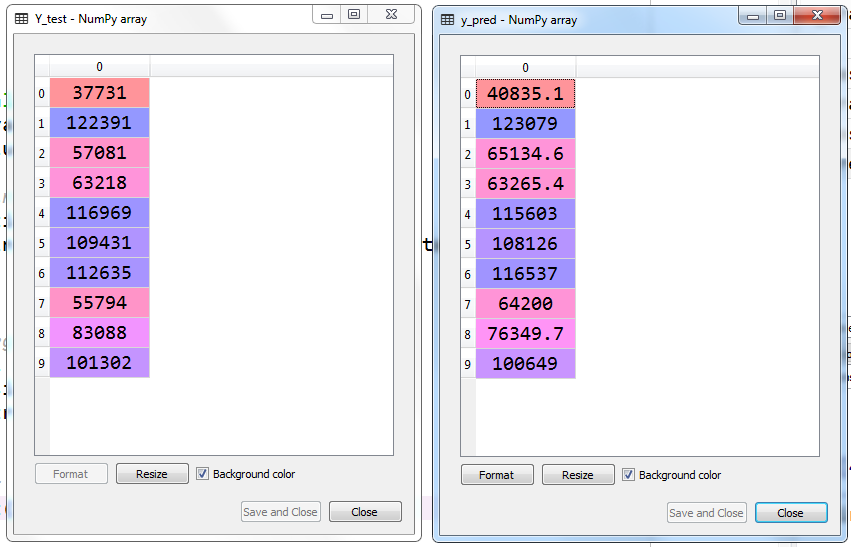
Linear Regress or is set to the training set



## 3.6 Predicting the Test set Results

1. # Predecting the Test set Results
2. y\_pred =regressor.predict(X\_test)

Predict method on the trained regreessor is called and Test set is set as the augment to get the predicted result set and the difference between the actual and predicted values are as below.

****

## 3.7 Visualizing the Training set results

1. #visualising the Training set results
2. plt.scatter(X\_train, Y\_train, color='red')
3. plt.plot(X\_train, regressor.predict(X\_train), color = 'blue')
4. plt.title('Salary vs Experience (Training Set)')
5. plt.xlabel('years of experience')
6. plt.ylabel('Salary')
7. plt.show()

The data processed is shown in the form of a graph using the plot form matplotlib.pyplot as plt libraries. The points shown with the red dots and the line that best fits the points with least possible error.

****