Data Mining Process and Analysis

During this term we learned a lot of data mining methods using python. In this we are supposed to do one project of our choice. I selected Height and Weight dataset to work using learned method in class. I found it very interesting. It is a small dataset. Now I will explain my methodology step by step.

**Step 1. Data preparation:** In this step, first I import some important libraries, csv file that contains my dataset and then create the data frame using Pandas.

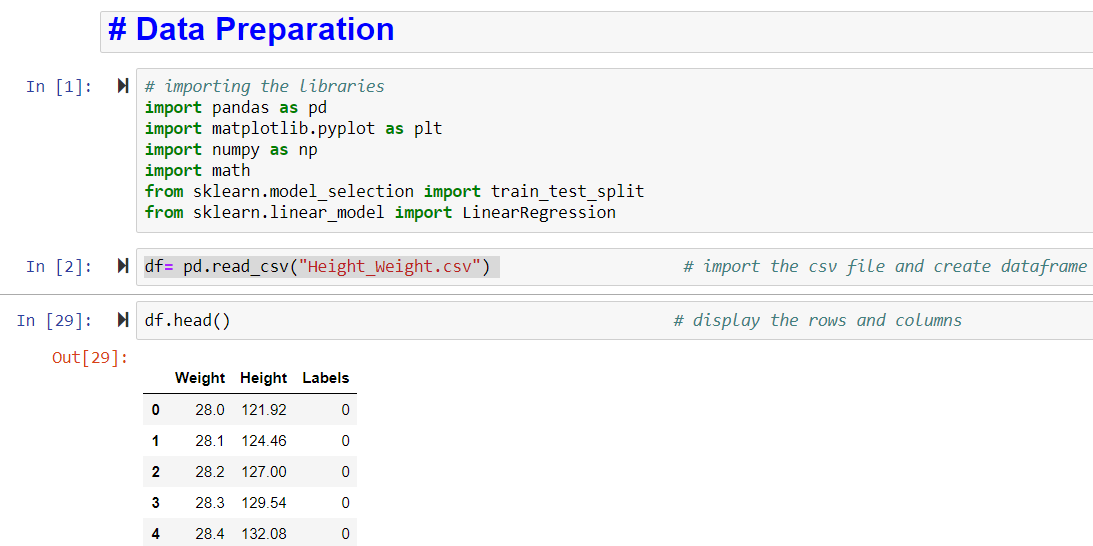
****

Image 1. Data preparation

**Step. 2 Data Exploration:** Basically, data exploration step is all about presenting the deep knowledge of dataset. In this step, I used df.info () function to get the summary of data frame such as how many columns and rows, index dtype (data type), column dtype, and memory usage. To get statistical details of data frame, I used df.describe() function. This method gave me the count, mean, std, min, max, etc. I also found the sum of null values in data frame.

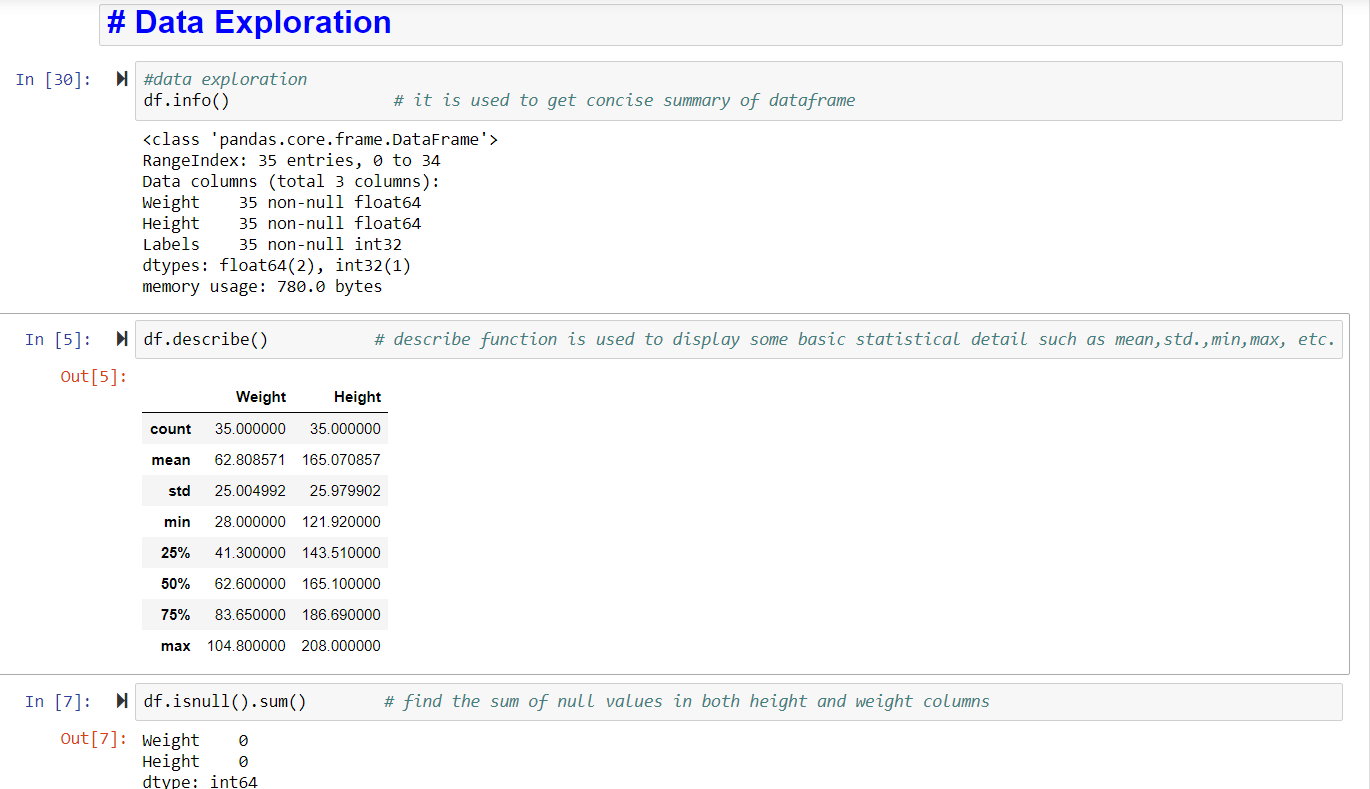
****

Image 2. Data Exploration

**Step-3 Plot the relationship between Height and Weight**

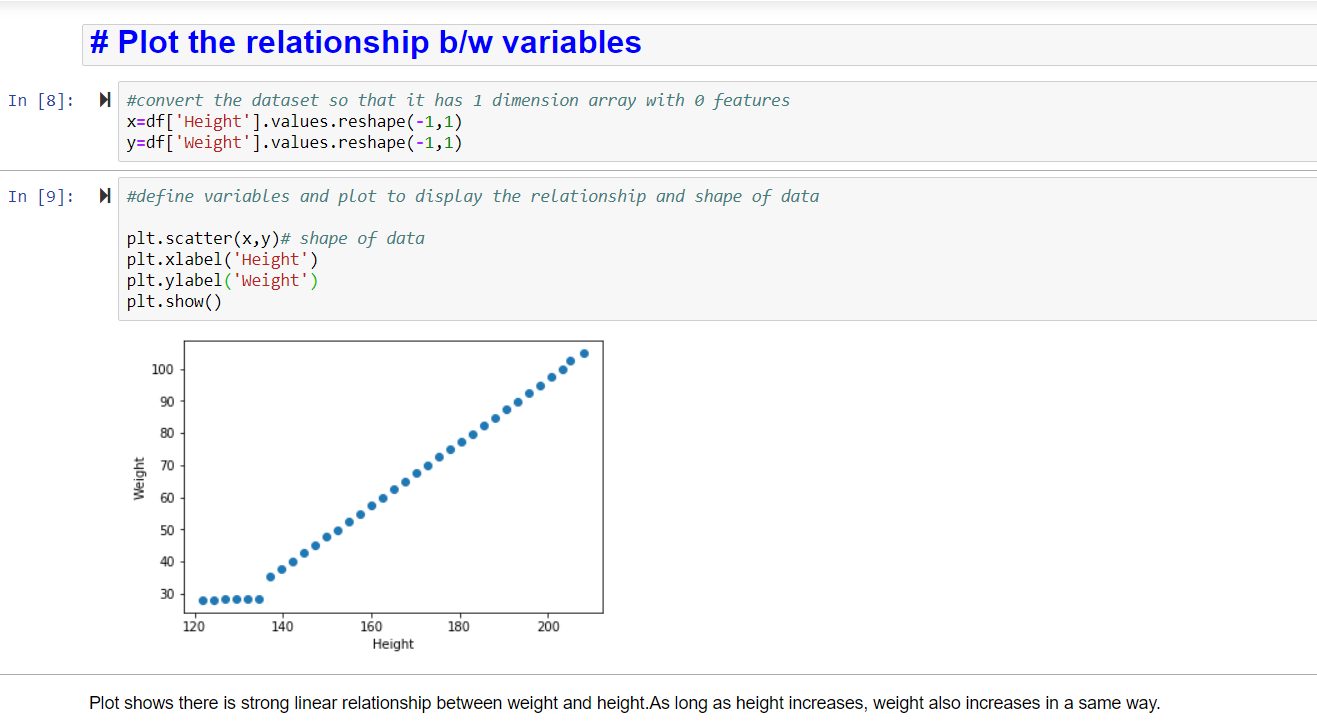


Image 3. Plot shows the linear relationship b/w variables

I plotted the scatter plot between height and weight columns to check their relationship. I found out strong linear relationship between them. As a person’s height increases weight also increases in a same way.

After this I import seaborn library to create high-level visualizations. When dealing with a set of data, the first thing that comes in your mind is how the variables are distributed and I plotted distribution plot to check the data distribution of height and column. This can be a beneficial tool for plotting the shape of a distribution.

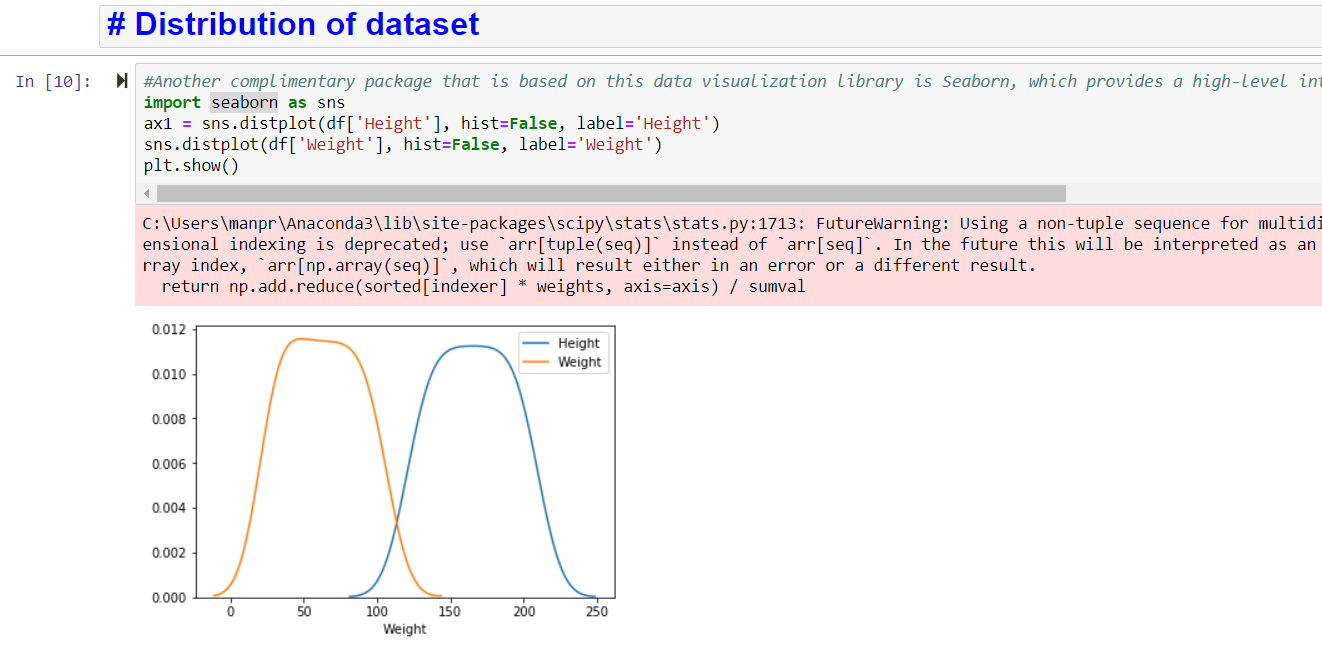


Image 4. Distribution of Dataset

**Step 3: Data pre-processing using Standard Scaler**

The idea behind Standard Scaler is that it will transform your data such that its distribution will have a mean value 0 and standard deviation of 1. After the feature scaling process all the features will be one the same scale and relative to each other.

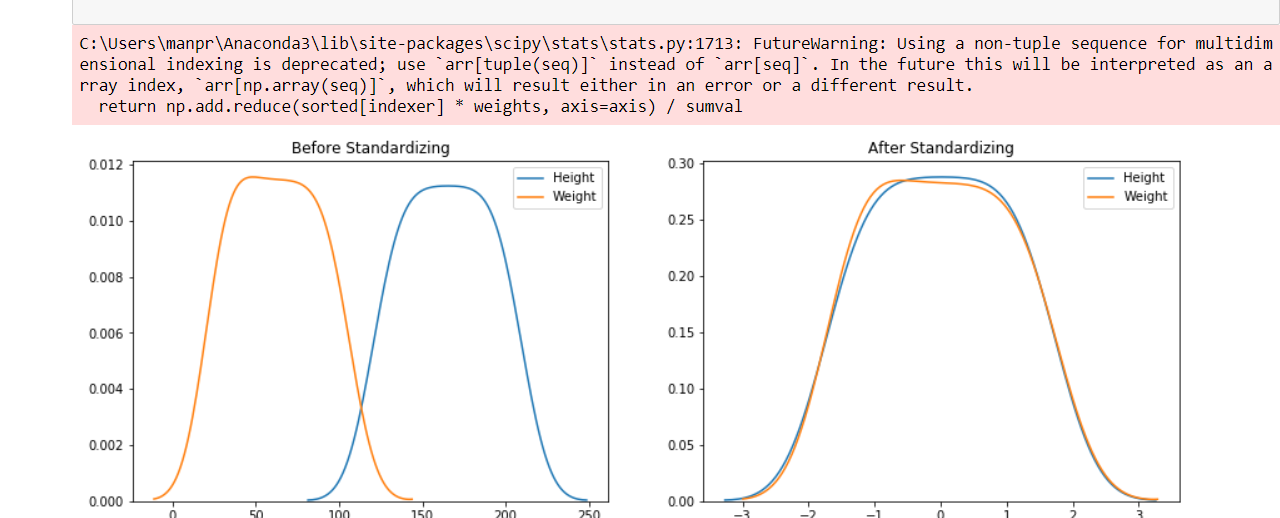
****

Image 5. Data distribution before and after Standardization

**K-means clustering:**

K-means clustering is an algorithm to start exploring an unlabeled dataset. The K in K-Means denotes the number of clusters. This algorithm has 4 steps:

1. Initialize Cluster Centroids
2. Assign datapoints to Clusters
3. Update Cluster centroids
4. Repeat step 2–3 until the stopping condition is met.

I used Elbow method to find out the best value for k mean how many clusters I should take for further analysis and produced some visualization showing the best k’s.

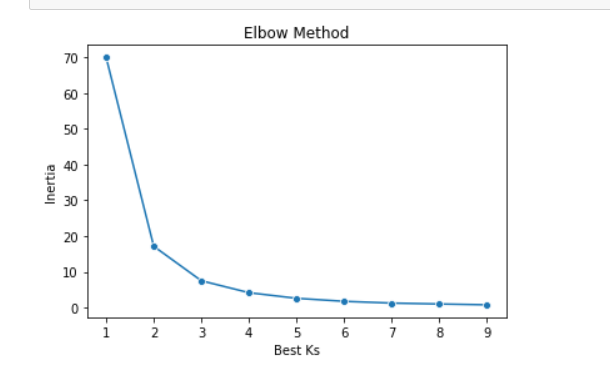


Image 5. Elbow method to find the best value of k

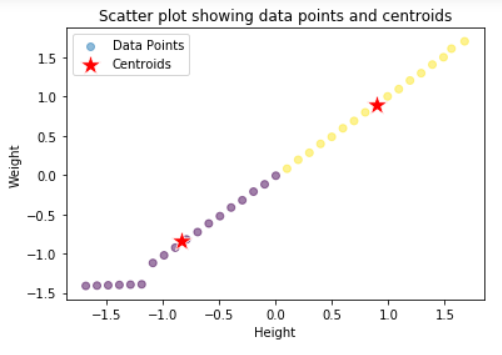
Above visualization shows, the best value of k will be 2. For building the model, I used k = 2 (no. of clusters). Then I found out the centroids which is the center point for each segment or cluster and then found out the data points near the centroids.

Image 6. Scatter Plot showing data points and centroids

**Analysis & Conclusions**

I also used SkLearn to build the model. I used linear regression method to find out the correlation and intercept in my dataset. It is used to measure at what extent there is a linear relationship between two variables (height and weight). In this we predict the value of dependent variable (weight) based on one or more independent variable (height).

After executing the linear regression, I just come to know that there is strong correlation between person’s height and weight. Weight depends upon the person’s height. It increases with the height. I calculated the value of r square and regression coefficient. Based on regression results, It proved the data is fitted to the model strongly.