# **Solution – Problems2 – Data Analysis**

## **Data**

Here the dataset from data.csv has –

ID – Corresponds to the row Number

Features – Columns (3 – 211) i.e. 208 features

Label – Binary Value –

* 1452
* 1134

From these details, we can deduce that a row has 208 features which might be used for predicting the labels.

## **Data Information**

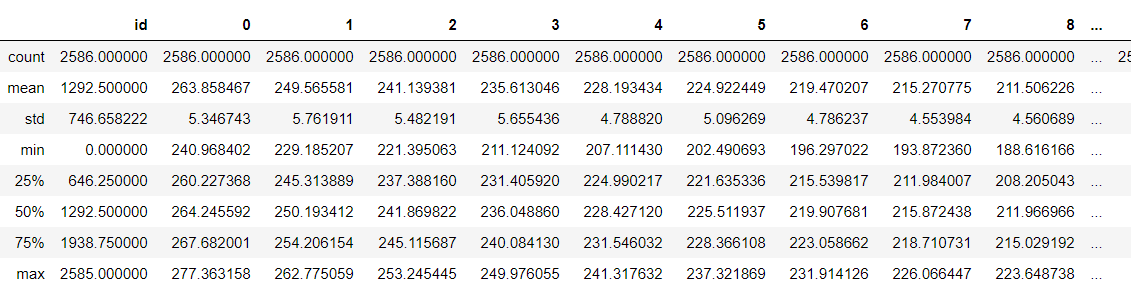
Here is a quick and short information about the data:

RangeIndex: 2586 entries, 0 to 2585

Columns: 211 entries, id to label

## **Data Description**

Here is a quick description(descriptive statistics) about data:



Note – due to large number of features, only first few is shown.

From this description – we have following results:

* Mean of all the features > 200
* Standard Deviation is ~5
* Not Much variation in data i.e. no outliers
* Data seems to be evenly distributed (from the percentiles)

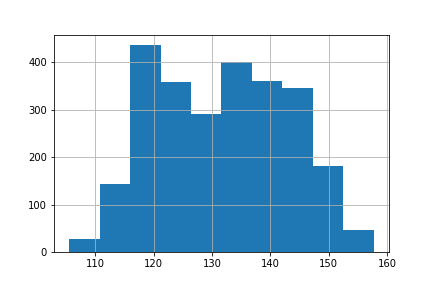
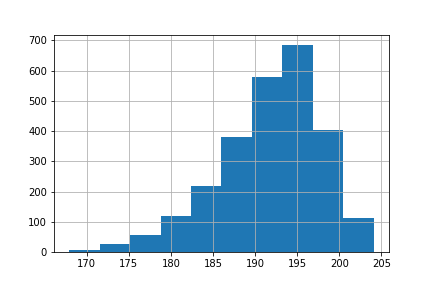
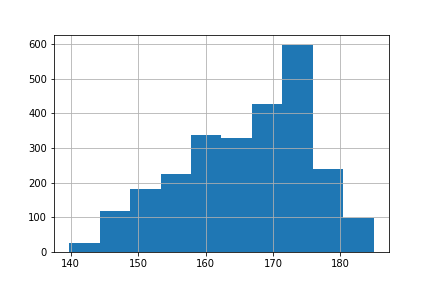
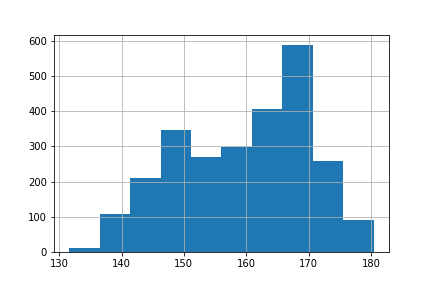
After this if we check the null values in the dataset, it’s none and neither there are any duplicates.

All the features are of Float datatype.

## Visualizations

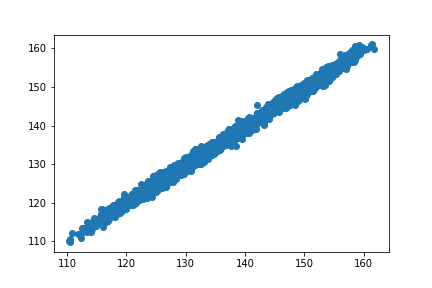
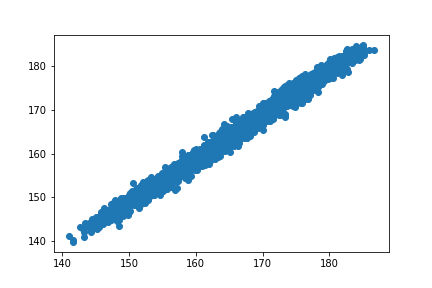
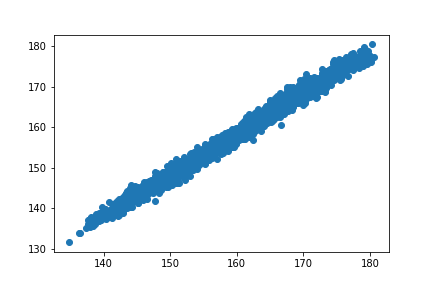
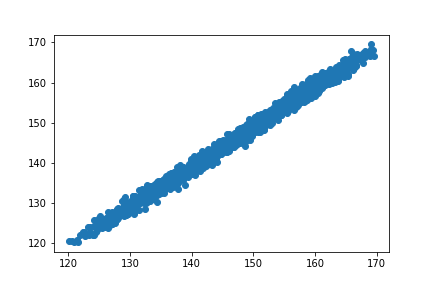
Data shows great extent of correlation and Distribution in some case is normal distribution where as in few other cases it’s skewed.

Here are few: (features ( 8,14,16,34) randomly chosen



To get the correlation plots we have like this for most of the features.

1. 28- 29 feature
2. 33-34 feature
3. 42-43 feature
4. 55-56 feature



Since Correlation is very high, PCA is applied to reduce the dimension of feature:

After Applying PCA, number of features has been reduced to 58, 2-D PCA has been applied.

