Machine Learning Report

# **A Description of DataSet**

## **Problem of Interest**

Predict Production Trunc by using other features of the data.

## **Data Source**

The data was provided by **turkyaljehni** to perform data analysis and predict Production Trunc with the help of other features provided.

# **A Detailed explanation of the data attributes**

## **Data issues and description**

Fortunately, this dataset contains no missing values. However it contains column **Scenario which is quite trivial**, when working on that data, therefore we removed it while working on our data.

A new attribute, **Output,** is calculated using Production output.

It is assigned a string value, according to the corresponding Production output value. Following table shows how Production TRUNC is mapped to Output.

| Production Trunc (Value Range) | Output |
| --- | --- |
| 41-50 | 41 to 50 |
| 51-60 | 51 to 60 |
| 61-70 | 61 to 70 |
| 71-80 | 71 to 80 |
| 81-90 | 81 to 90 |
| 91-100 | 91 to 100 |

## **Basic statistics of attributes**

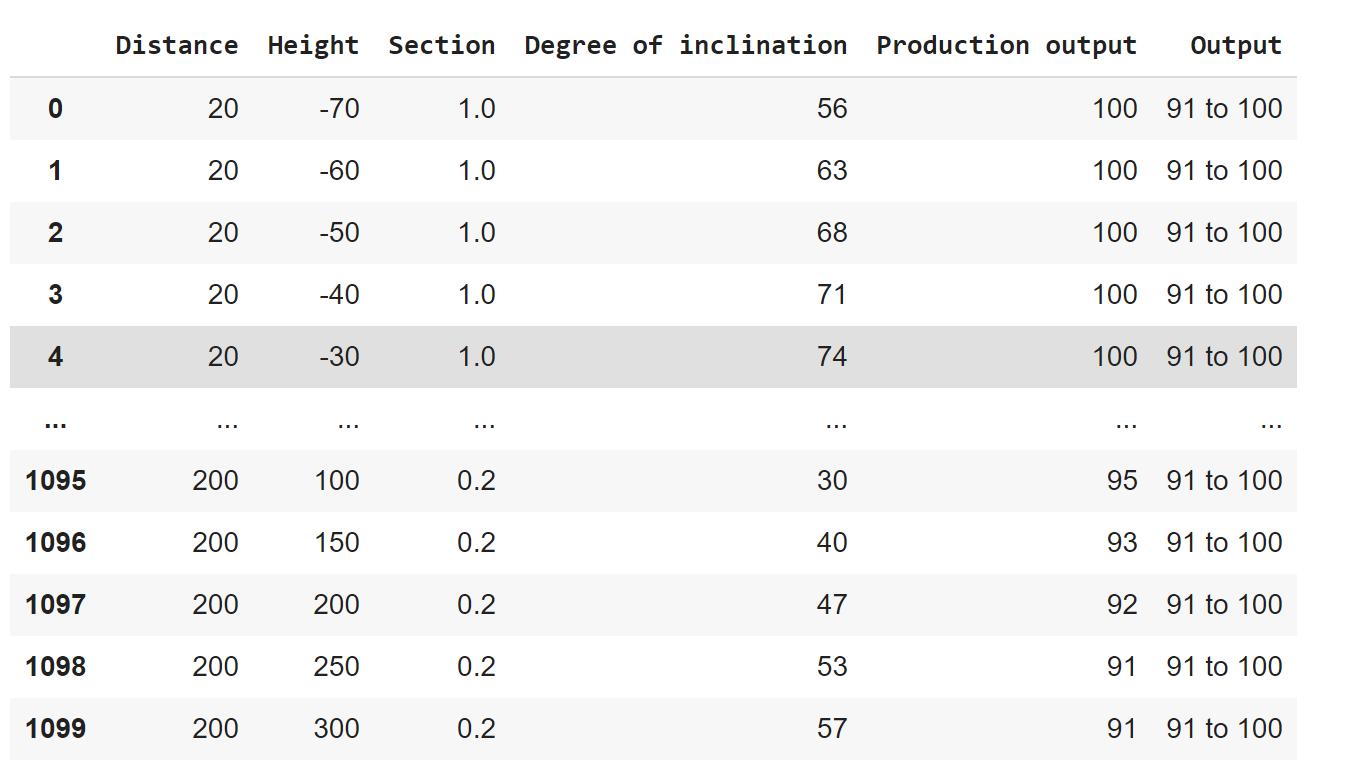


Table 2: Summary statistics

In table 2 we have expressed the general statistics describing the data.

Features concerning numeric data are count, mean, standard deviation, minimum, maximum and quartiles. Numeric data includes distance, height, section, degree of inclination and Production output.

Features concerning non numeric data are count, unique, top and frequency. Non numeric data include Output.

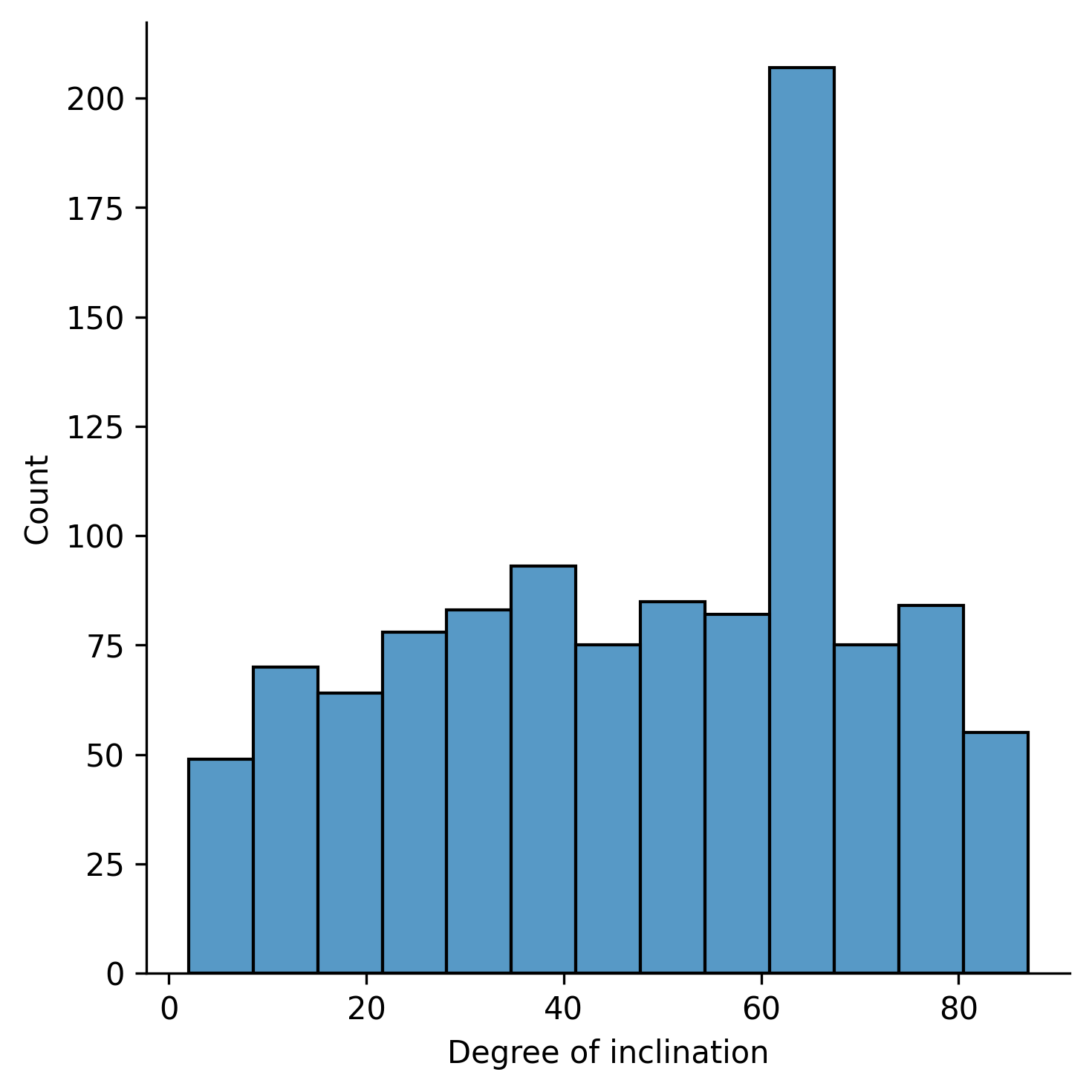
These are the basic statistics which are required for doing data analysis and selecting the right model for prediction.

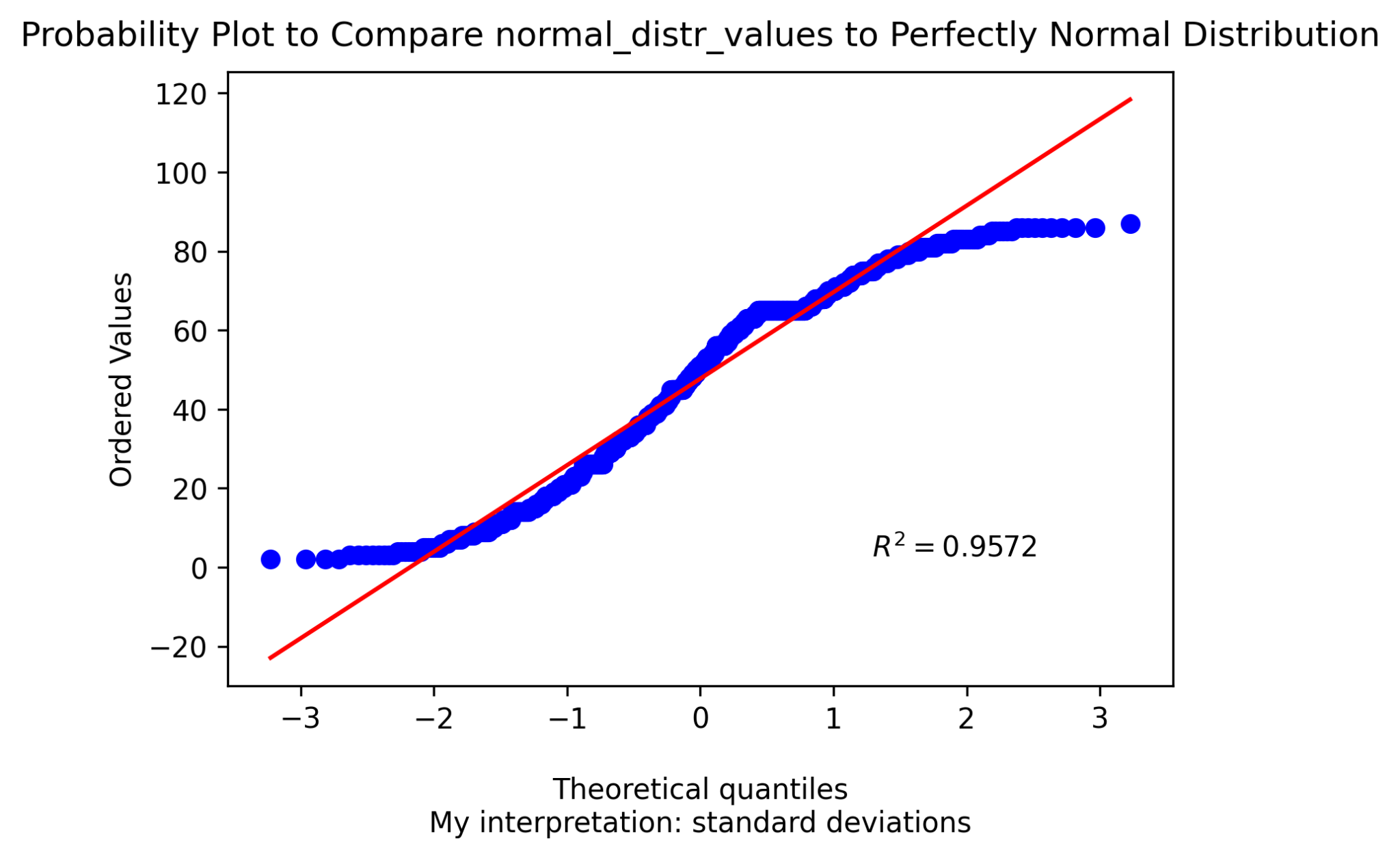
# **Data visualization**

## **The distribution of the attributes**

In this section, an analysis of the distributions of the attributes will be made on the data set.

**Histogram and Normal Probability Plot for the attribute Degree of Inclination**

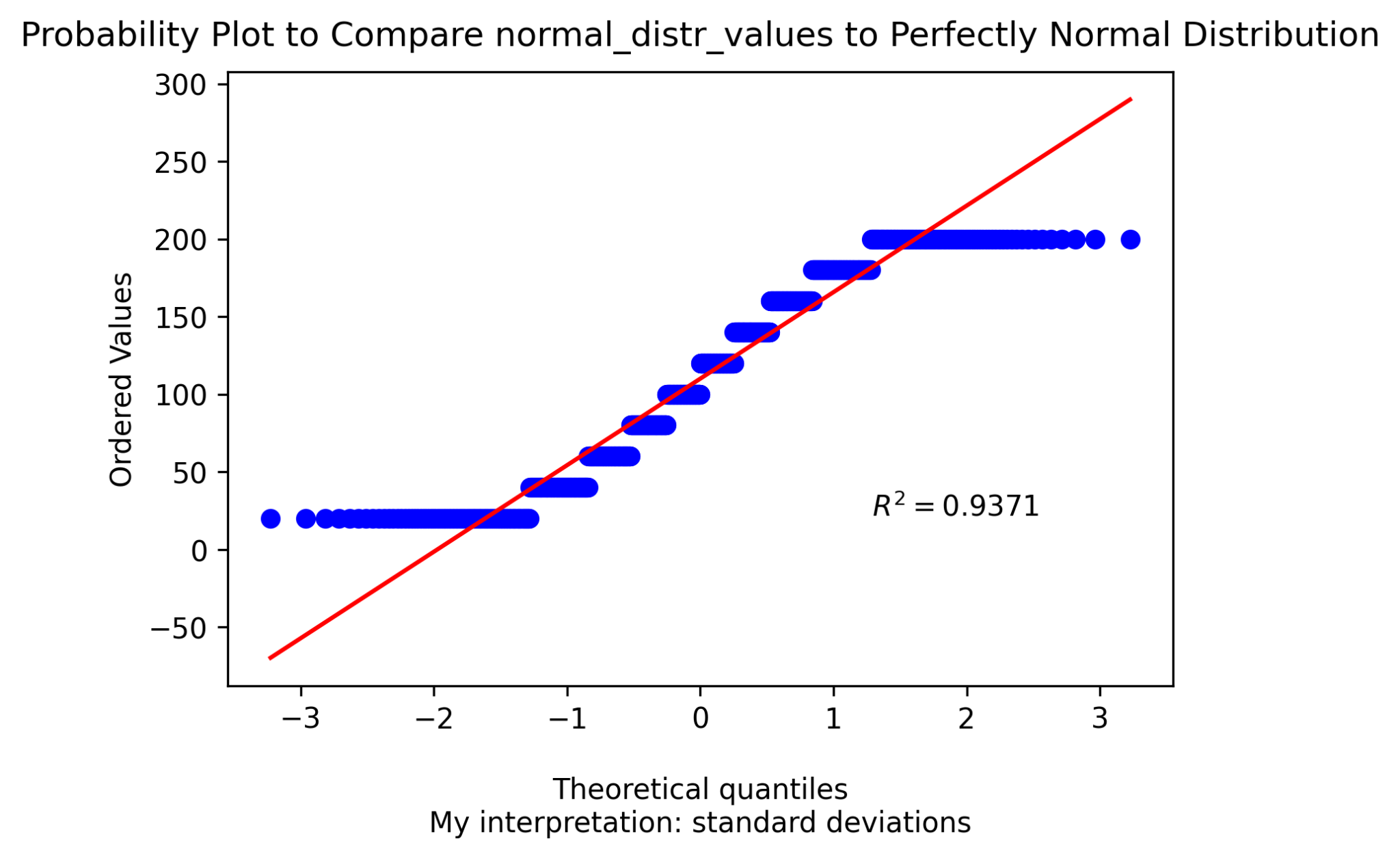
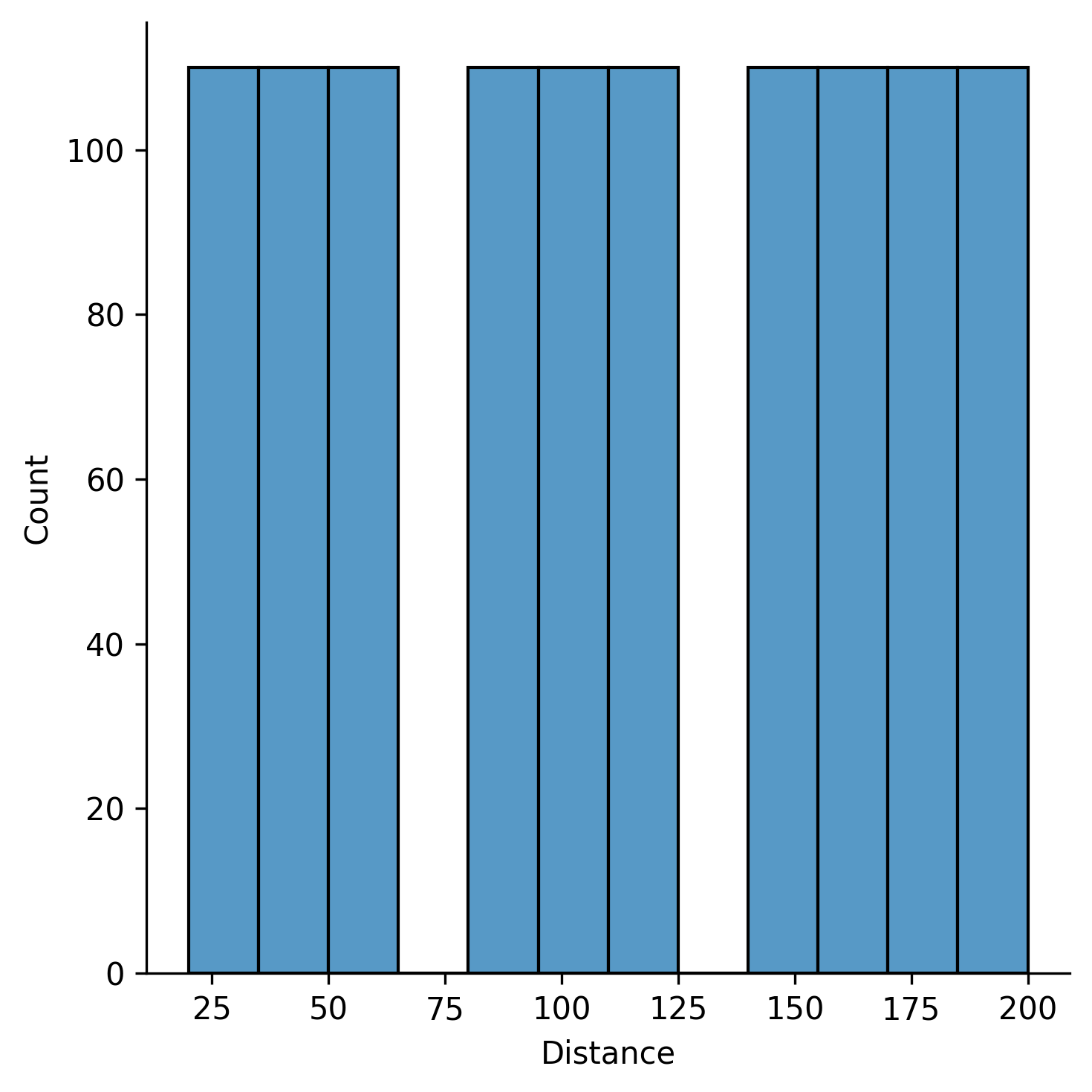




In this case, we see that the data for the attribute **Degree of inclination** resembles a normal distribution a bit, but it skews more to the left .

#### 

#### Histogram and Normal Probability Plot for the attribute distance

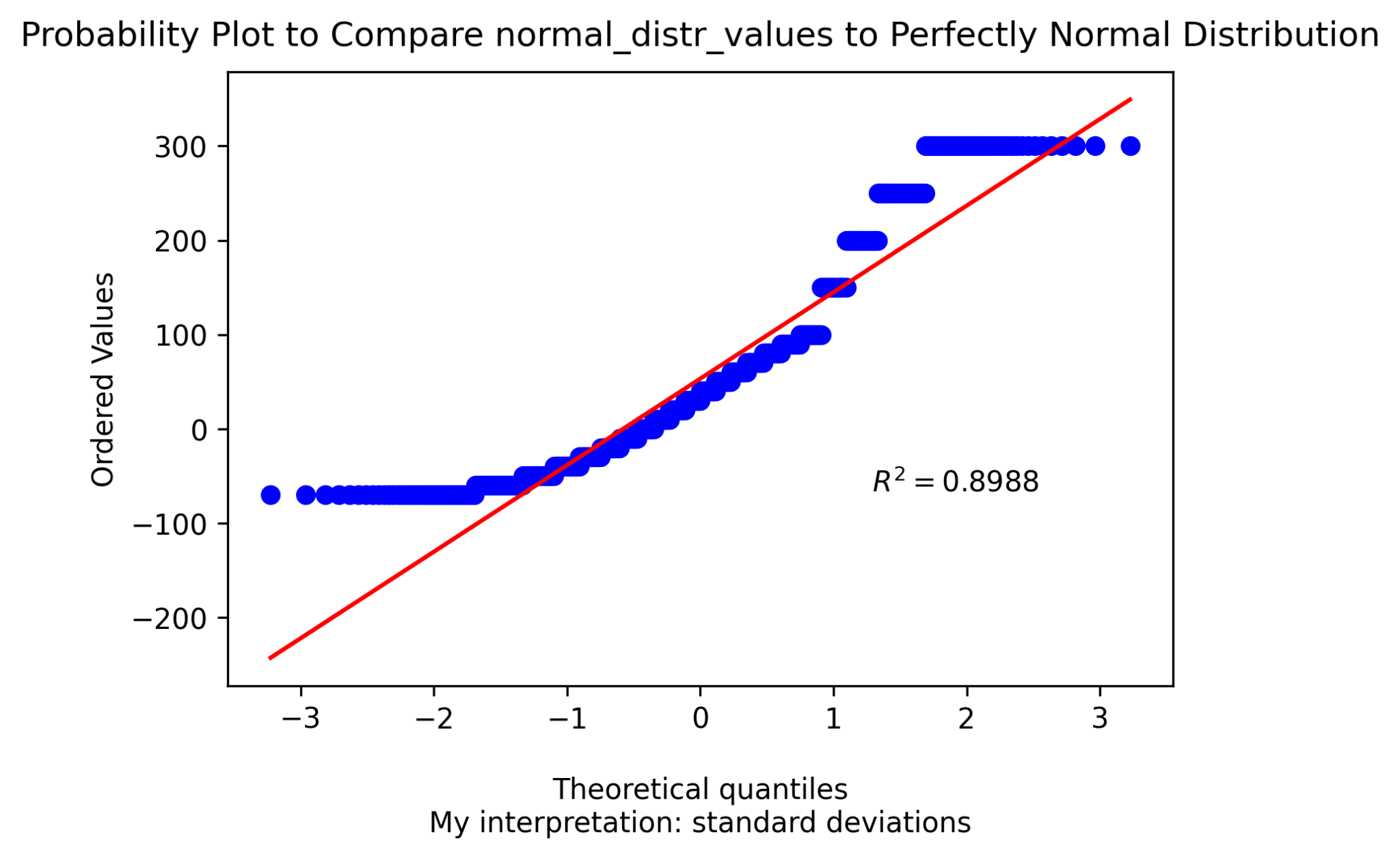
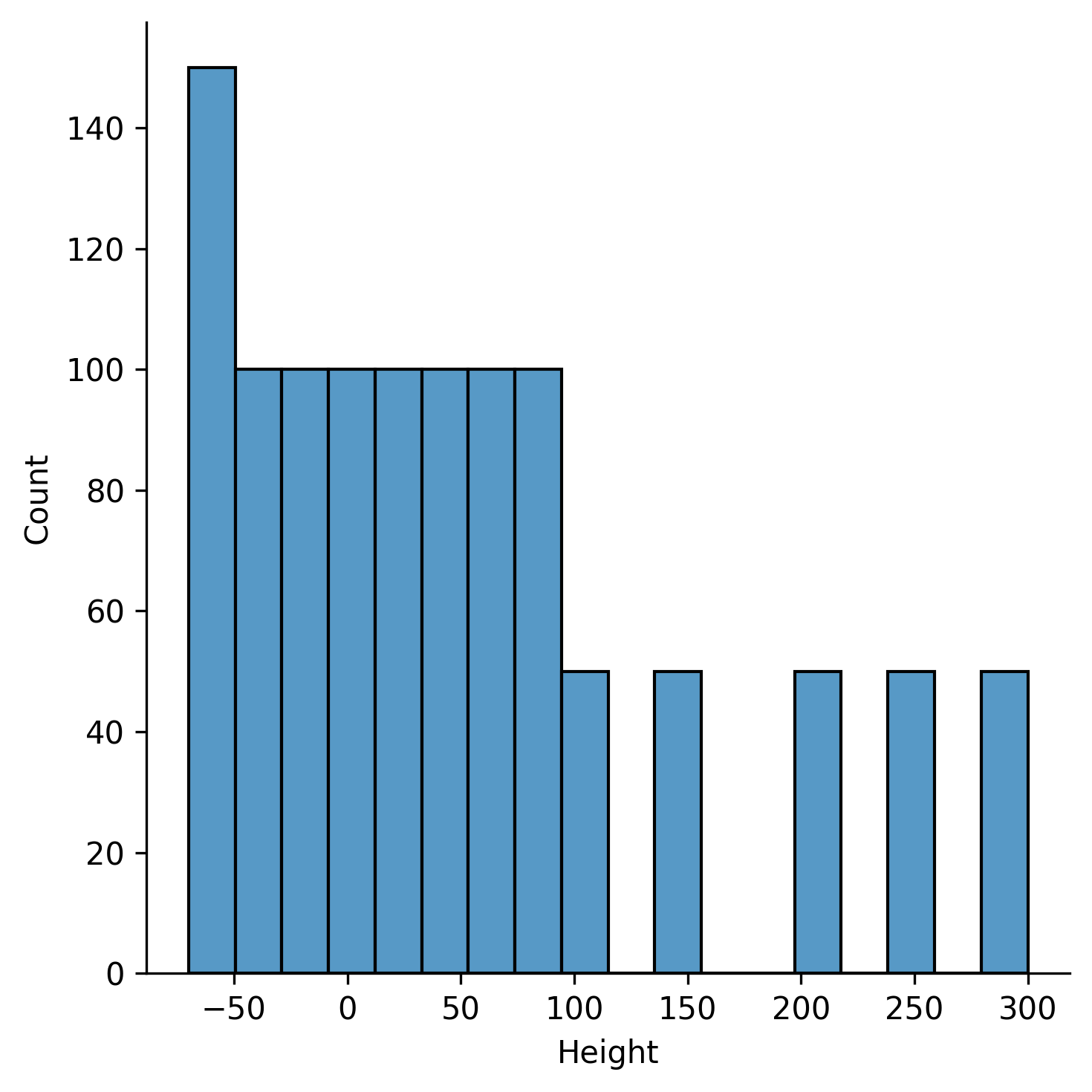


When considering **Distance** attributes, the data is not normally distributed.

#### 

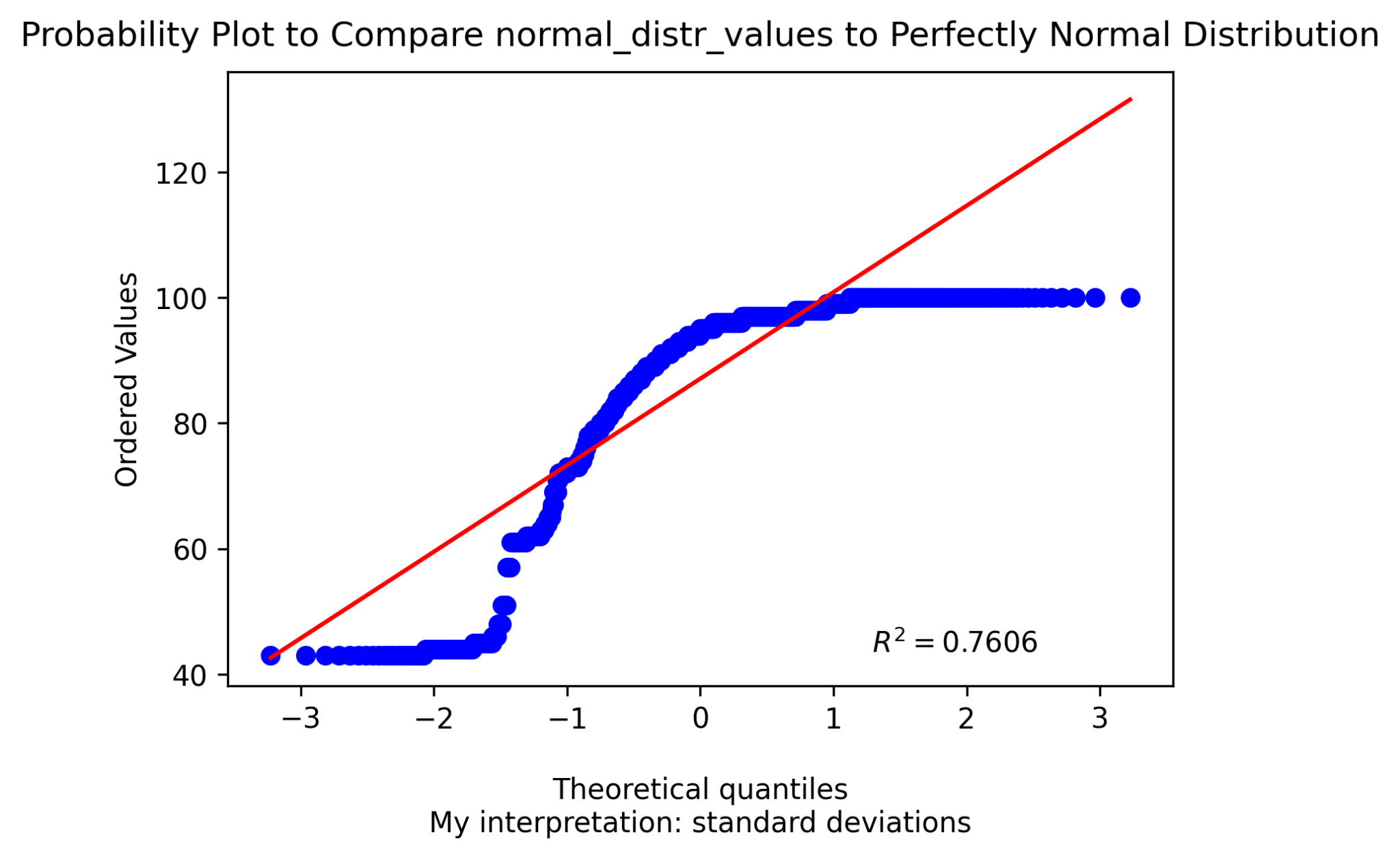
#### 

#### Histogram and Normal Probability Plot for the attribute Height



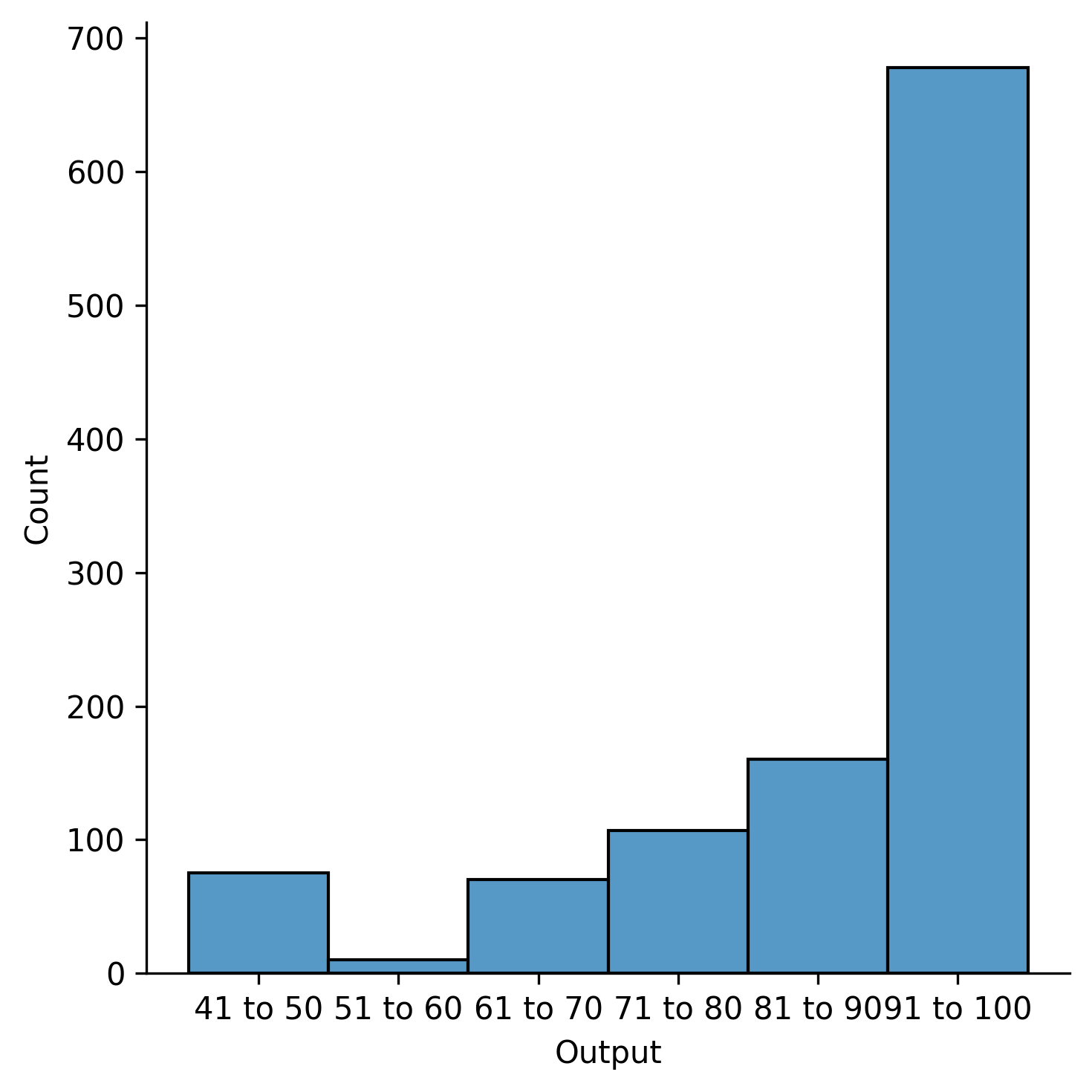
As for **Height** , by looking at the histogram and the normal probability plot, we can see that it is skewed towards right and is not completely normally distributed.

#### Histogram and Normal Probability Plot for the attribute Production Output



For **Production Output** the graph is skewed towards the left as it is visible from histogram and probability plot.

#### Histogram for the attribute Output



This histogram shows that the data is unequally distributed among the six categories. **91 to 100** contains the highest count of data while  **51 to 60** contains the lowest count of data.

## **Correlation between the variables**

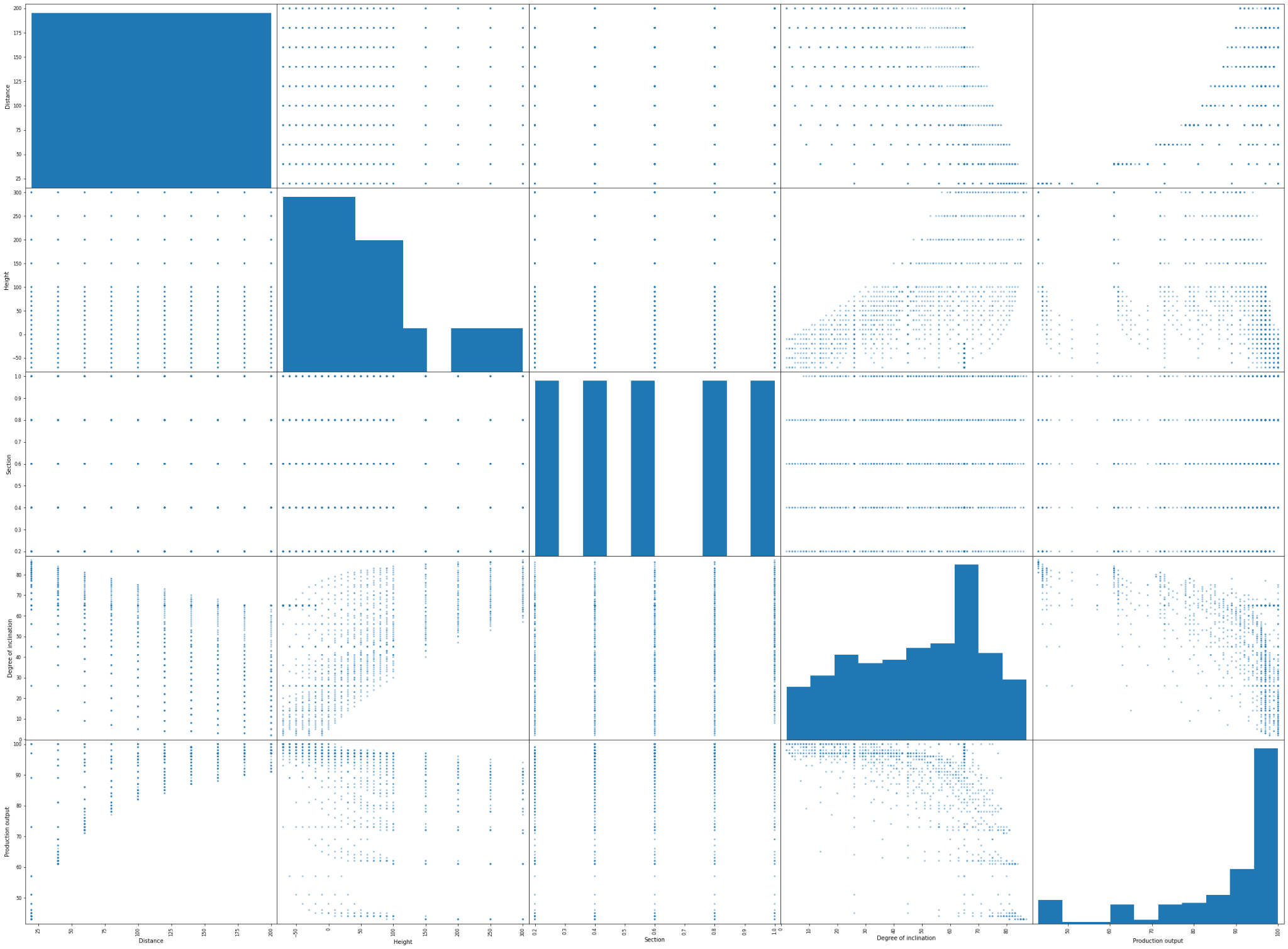
#### HeatMap



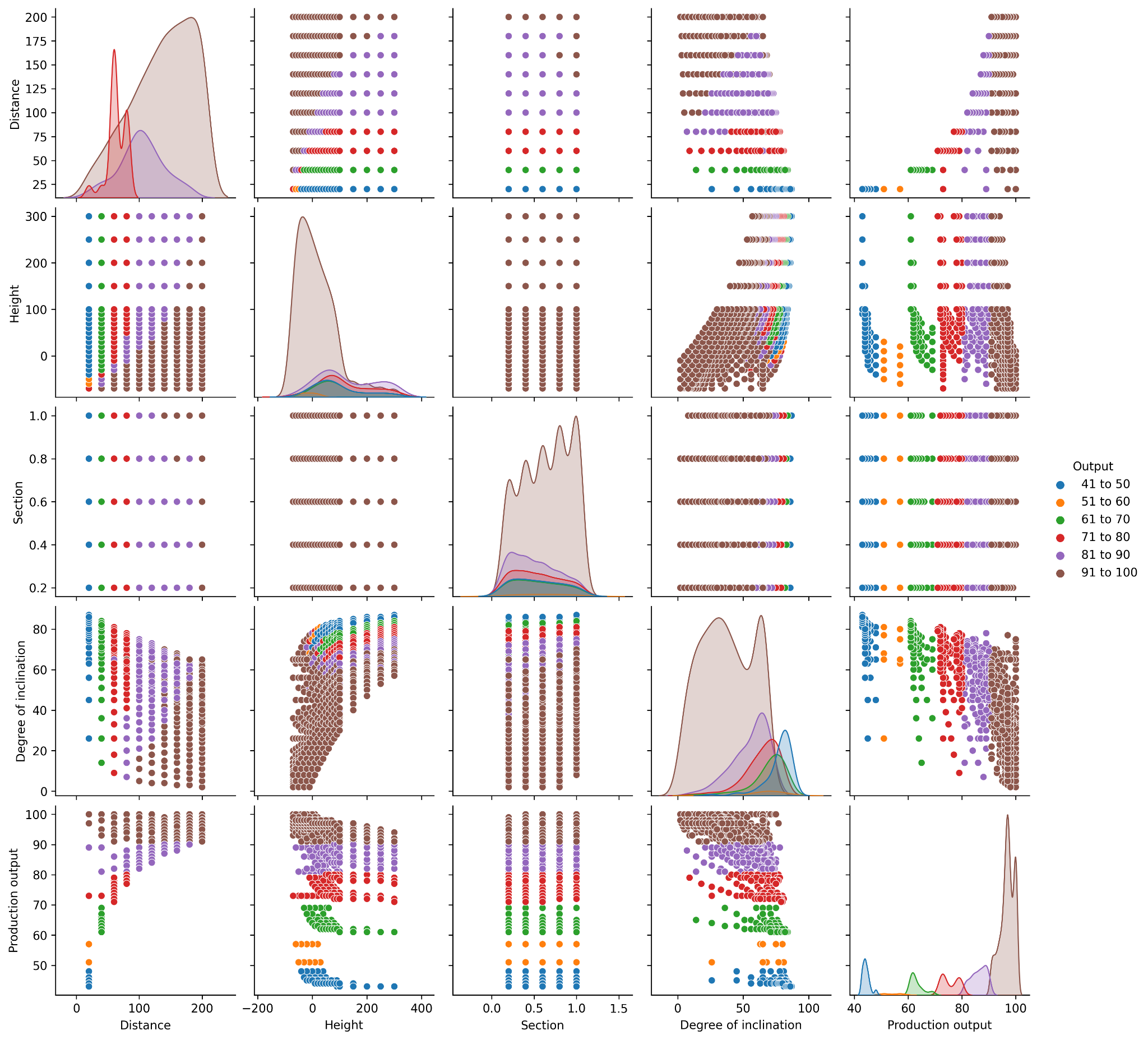
In the heatmap the greater values show strong relationships, zero values show no correlation and small values show weak relationships. Negative value shows that the relation is negative or inverse.

The attributes have both positive and negative correlations. It is interesting to notice that the degree **of inclination is negatively correlated with the distance and Production Output**. On the **other** hand, **Distance is strongly correlated with Production Output.** Heigth is not very highly correlated with Production Output but is correlated with degree of Inclination.

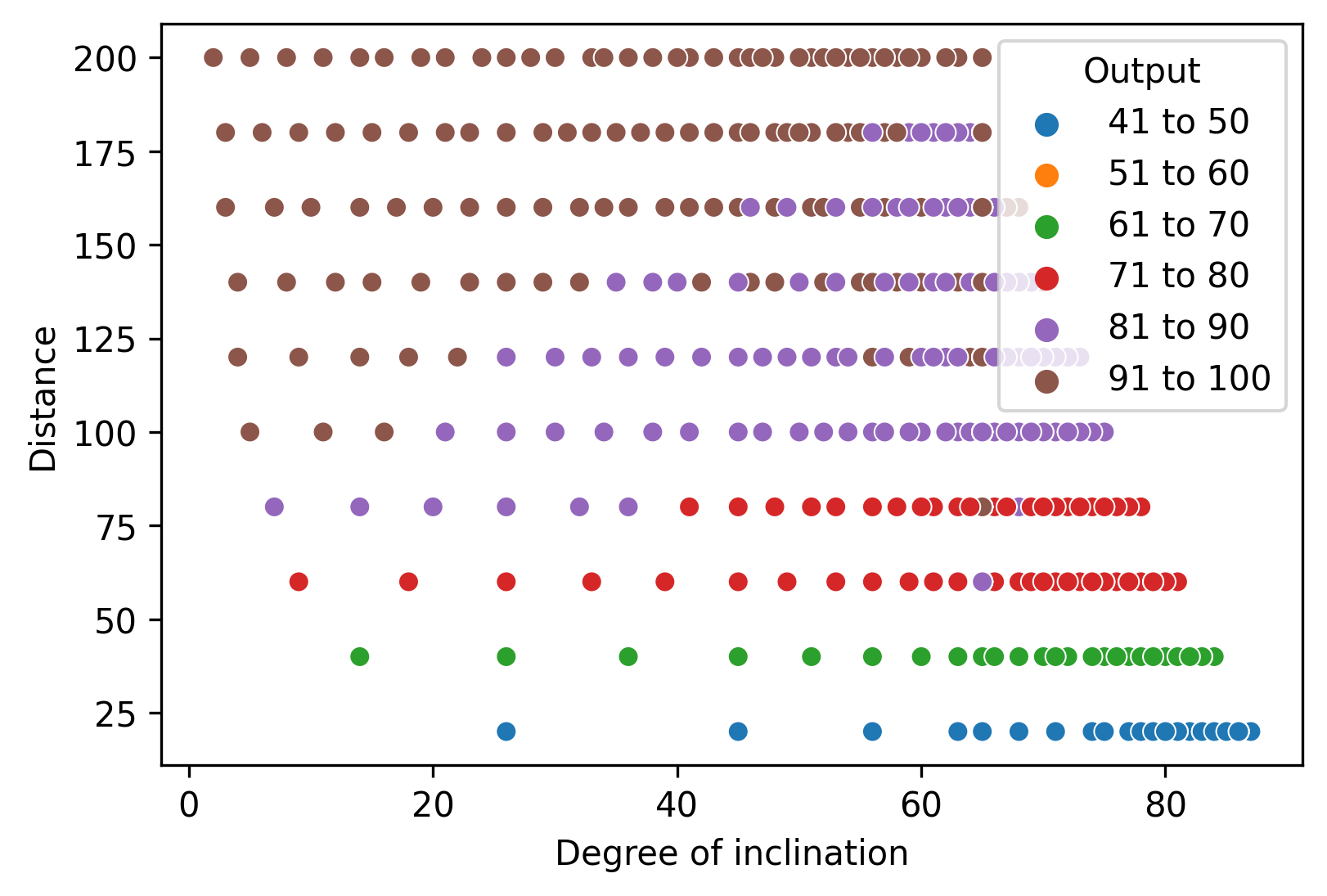
#### Scatter Matrix

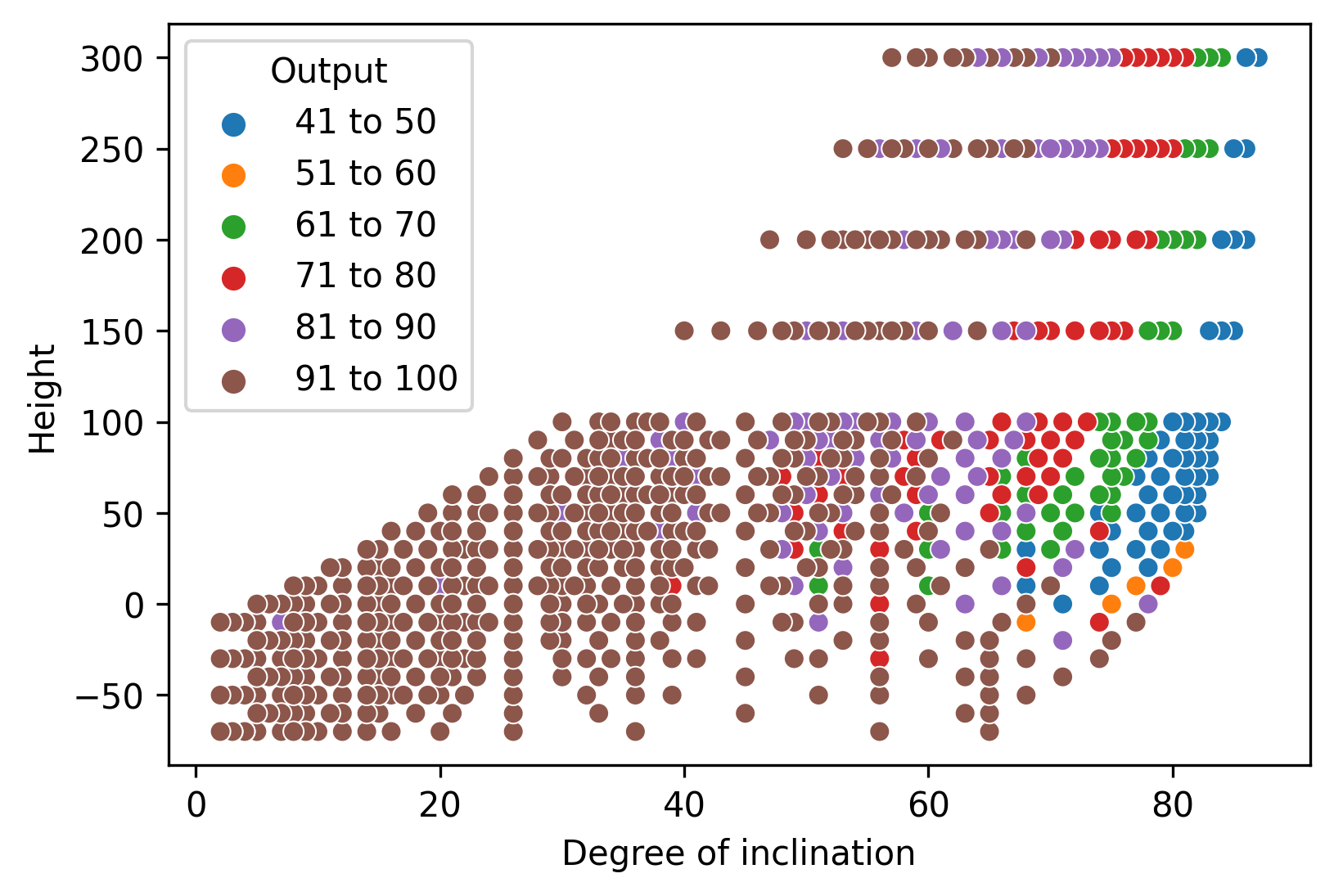


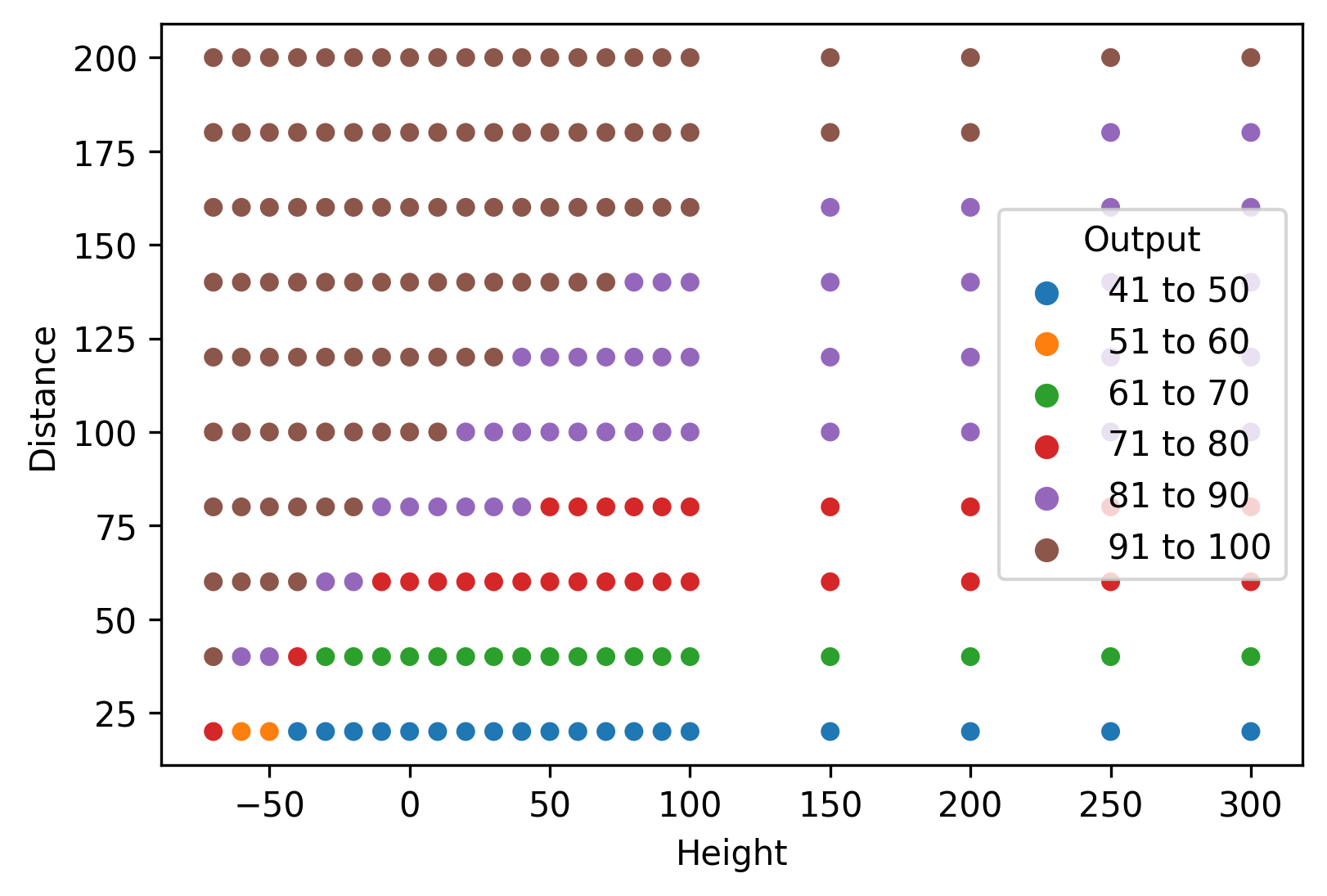
#### PairPlot



#### Plotted Correlation of Specific Attributes







## **Primary machine learning applicability based on visualizations**

Based on the initial visualizations there are good indications that the data set is capable of supporting the intended aim with the respective methods proposed. The data is consistent but **not completely normally distributed** and the continuous attributes hold regressional possibilities. Classification may also be applied using the **Output** attribute as the output of the model. There appear to be various correlations between the continuous attributes that can be used as indicators.

# **Applying Classification:**

In machine learning, classification refers to a **predictive modeling** problem where a **class** label is predicted for a given example of input data. We have to select input and output of the models.

## **Input:**

As we have seen the correlation of the attributes of the dataset. Distance and Production output are highly positively correlated and Degree of inclination is highly negatively correlated with the Production output. Note that the **Output** attribute is the output variable for which we need to make predictions and it is created using **Production output**.

Height is not very strongly correlated with Production Trunc but it is weakly correlated with degree Trunc so we cannot simply ignore it.

We would use **Degree Trunc, Distance,Section and Height** as the input to the classification model.

## 

## **Output:**

As we only want to predict Output so **Output** is the output of our Model.

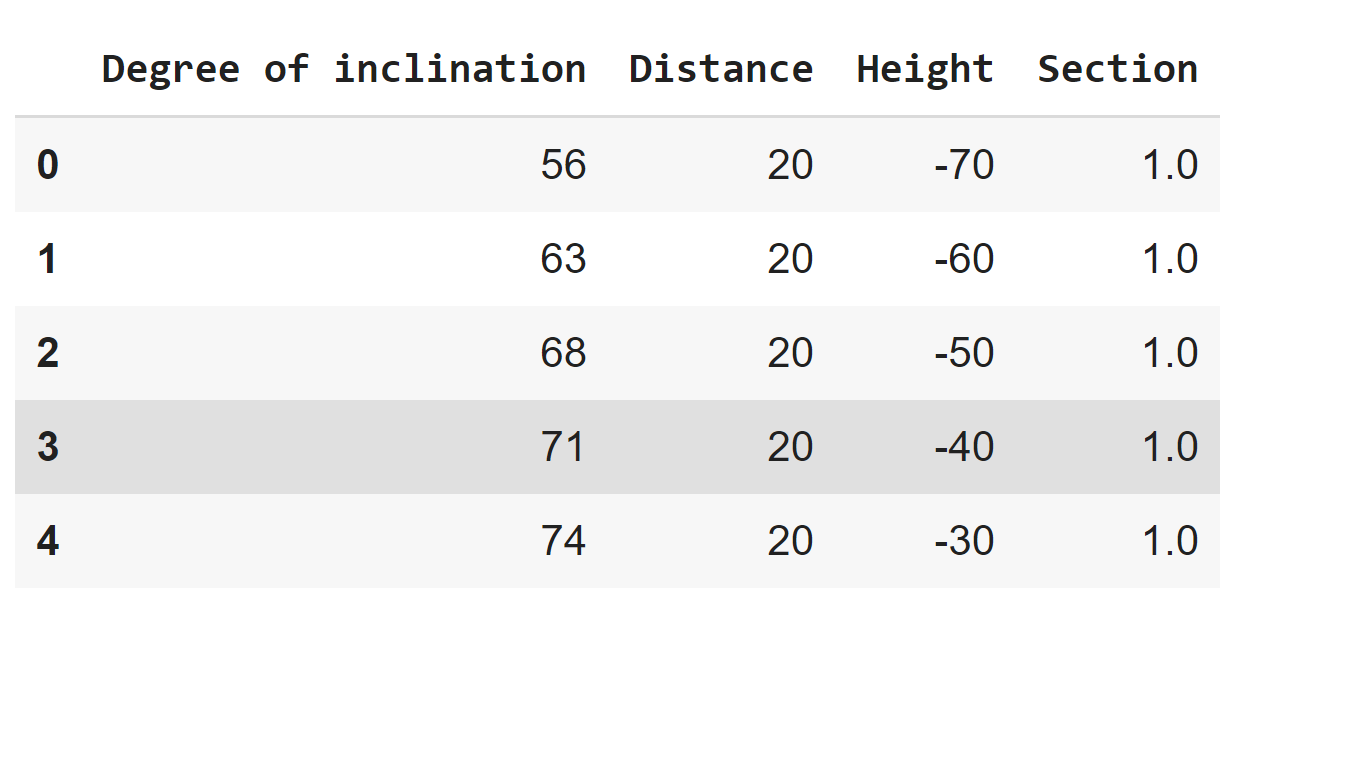
## **Preprocessing:**

Before training our model we would need to preprocess our data as the output accuracy is highly dependent on the data we are entering into the model.

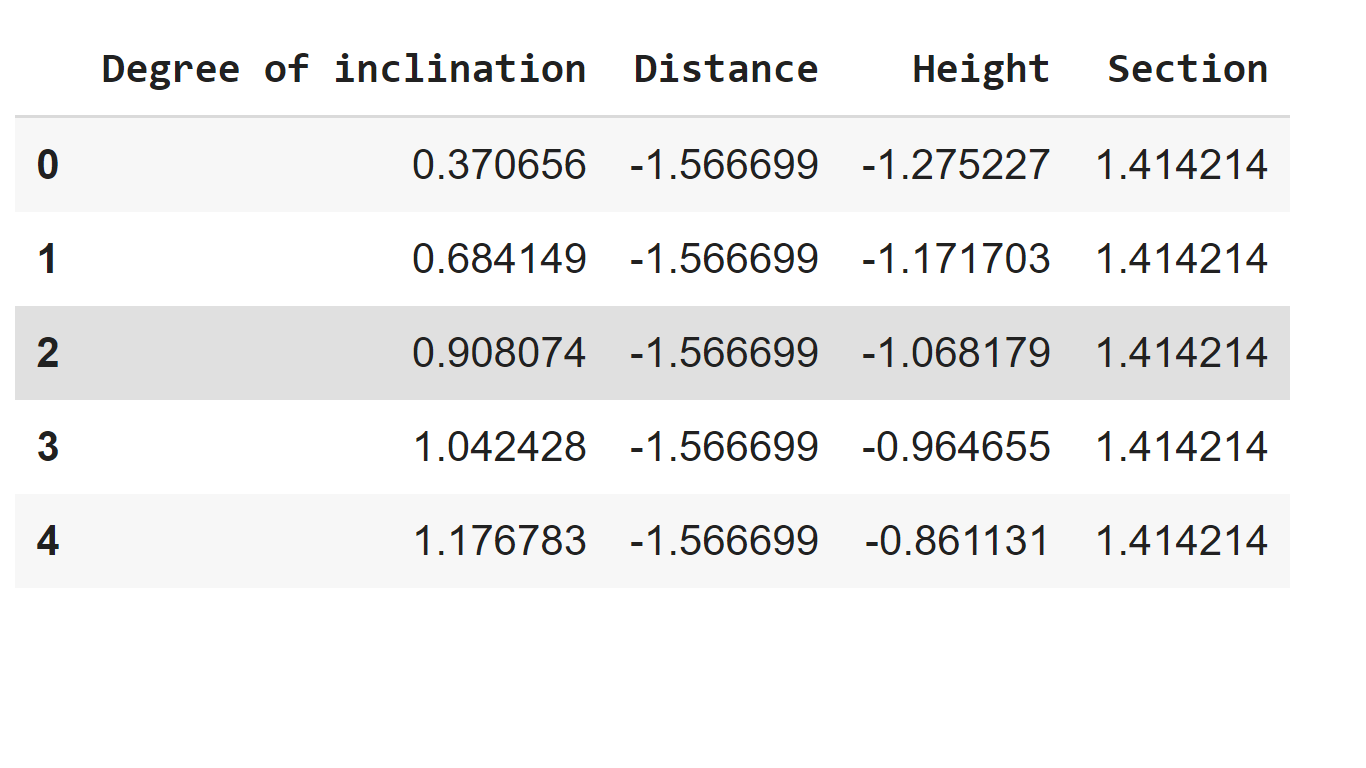
We will convert all our data attributes into standard form so that their standard deviation is 1. This technique is called **Standard Scalar.**

Now we are ready to train the model.

### **Before Preprocessing**



**After Preprocessing**



## **Model:**

We have split our data **20% and 80% for testing and training** respectively. We applied Classification with Input Provided above and predicted Output.

We have used the following classification models:

1. **Logistic Regression**

The logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick.

1. **Support Vector Machine**

SVMs are based on the idea of finding a hyperplane that best divides a dataset into two classes, as shown in the image below.

1. **Random Forest Classifier**

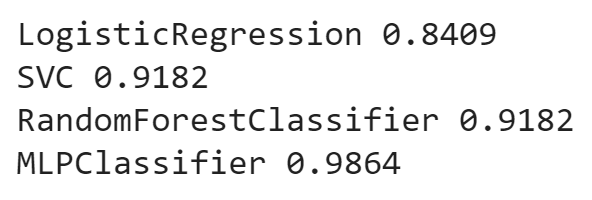
Random forest classifier creates a set of decision trees from a randomly selected subset of the training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

1. **Multilayer Perceptron**

Multilayer perceptrons train on a set of input-output pairs and learn to model the correlation (or dependencies) between those inputs and outputs.

## **Testing:**

Accuracy of these models against the training data is as follows



It is quite evident that multilayer perceptron has the best accuracy out of all these models.

# 

# **Discussion**

## **Summary**

All in all, we can conclude the following points about what we have learned from the data:

1. There was a strong negative correlation between Degree of inclination and Production Output and a strong positive correlation between Distance and Production Output. Height does not have any correlation with Production output but has a weak correlation with Degree of inclination.
2. We used all four attributes to train the Classification Model. Leaving Height attribute was not a good idea as Height is correlated with degree of Inclination.
3. Using MLPClassifier we get an accuracy of 0.9864. Which means that our model predicted 98.7 % of the data correctly which is a huge achievement.
4. This 98.7 % accuracy proves that the attributes given input has a deep relationship with the production output.