# 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 and Section which is quite trivial**, when working on that data, therefore we removed it while working on our data.

## Basic statistics of attributes

| | Production TRUNC | degree trun | distance | height |
|---|---|---|---|---|
| **count** | 750.000000 | 750.000000 | 750.000000 | 750.000000 |
| **mean** | 83.209333 | 52.204000 | 110.000000 | 96.666667 |
| **std** | 16.406963 | 19.804781 | 57.483962 | 86.980705 |
| **min** | 43.000000 | 5.000000 | 20.000000 | 0.000000 |
| **25%** | 75.000000 | 36.000000 | 60.000000 | 30.000000 |
| **50%** | 90.000000 | 53.000000 | 110.000000 | 70.000000 |
| **75%** | 96.000000 | 68.000000 | 160.000000 | 150.000000 |
| **max** | 100.000000 | 87.000000 | 200.000000 | 300.000000 |

Table 2: Summary statistics

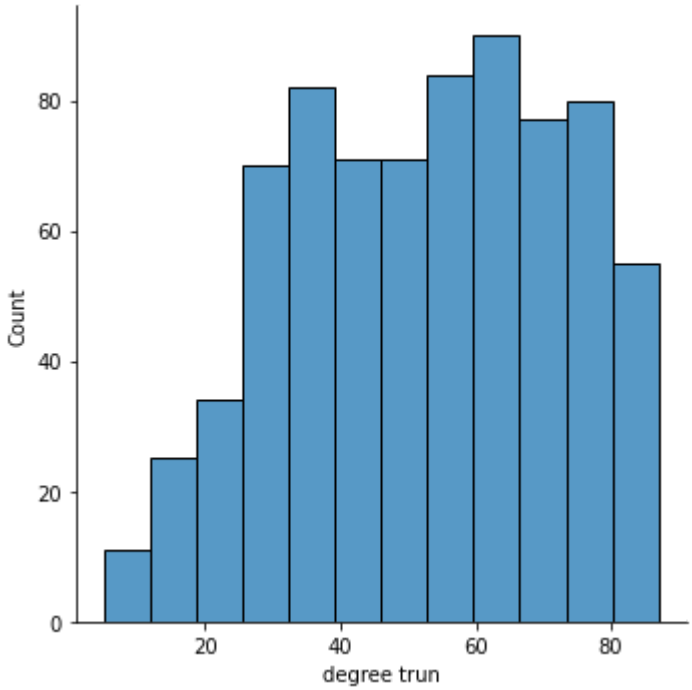
In table 2 we have expressed the general statistics describing the data. All aur features are numerical so we have got the count, mean, standard deviation, minimum, maximum and quartiles. These are the basic statistics which are required for doing data analysis and selecting the right model for prediction.

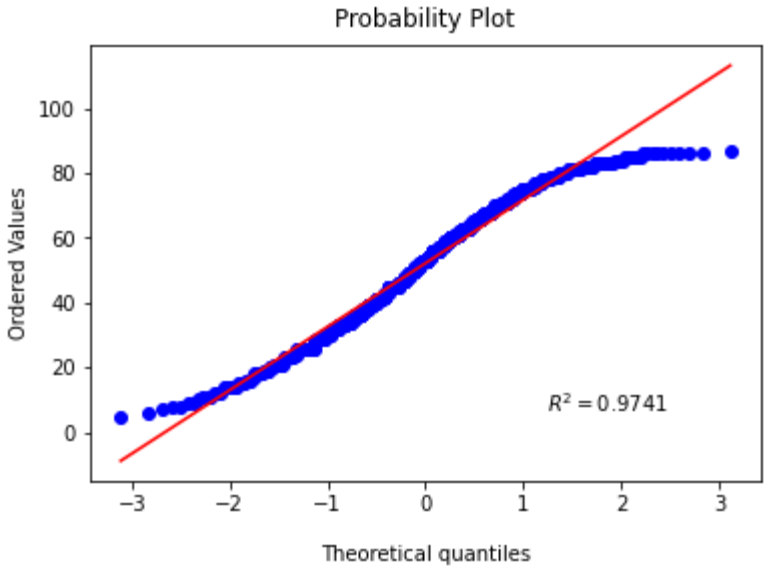
# Data visualization

## The distribution of the attributes

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

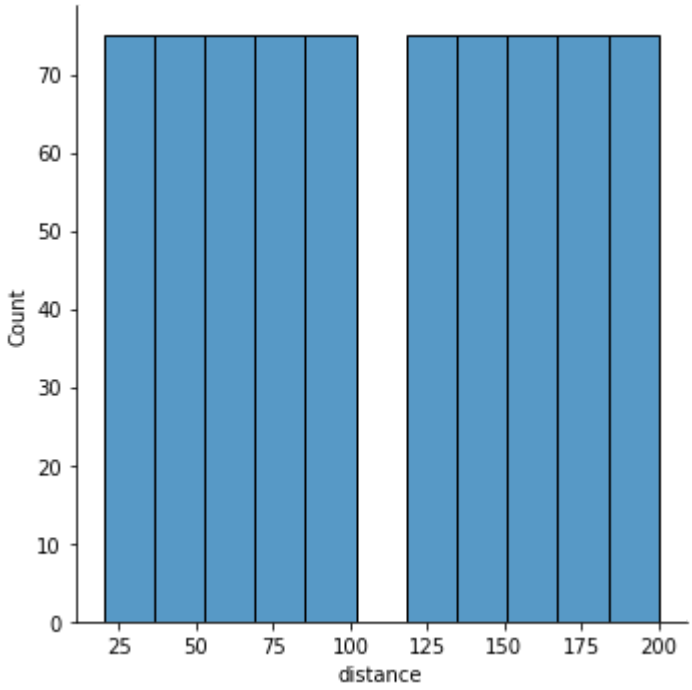
### Histogram and Normal Probability Plot for the attribute Degree Trunc

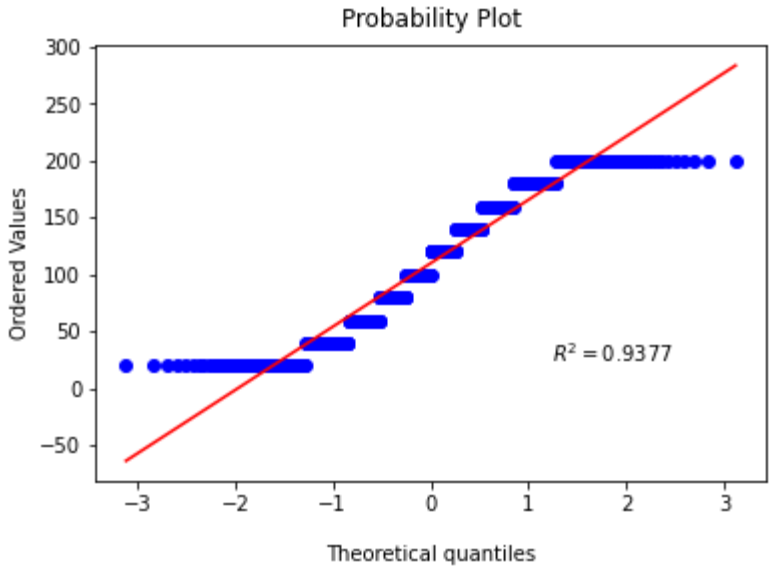



In this case, we see that the data for the attribute **Degree Trunc** resembles a normal distribution a bit, but it skews more to the left .

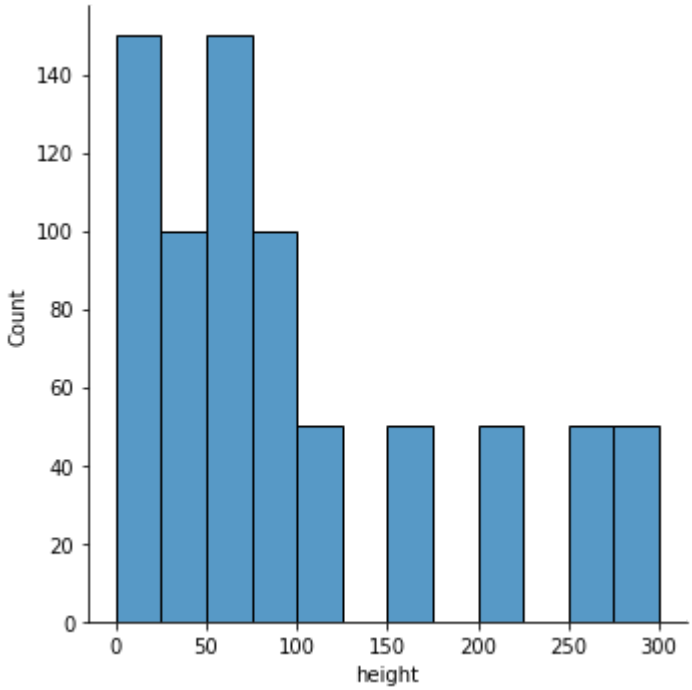
### Histogram and Normal Probability Plot for the attribute distance

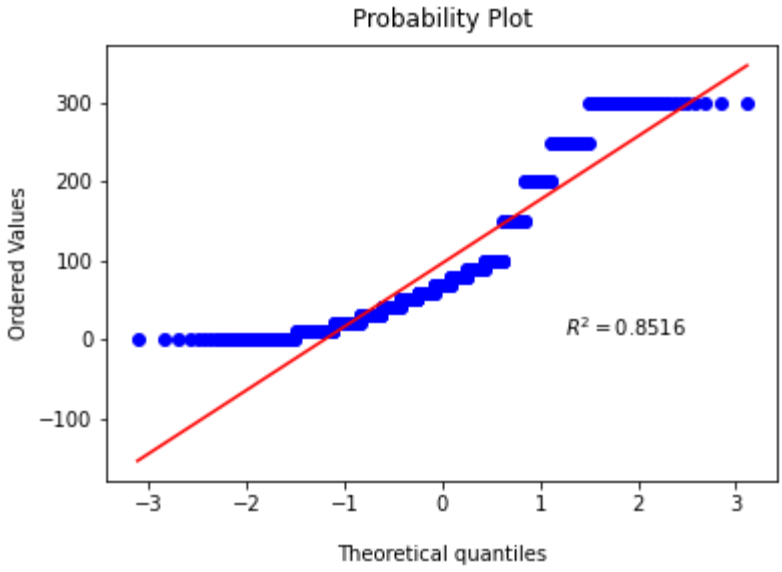



When considering **distance** attributes, the data is not normally distributed.

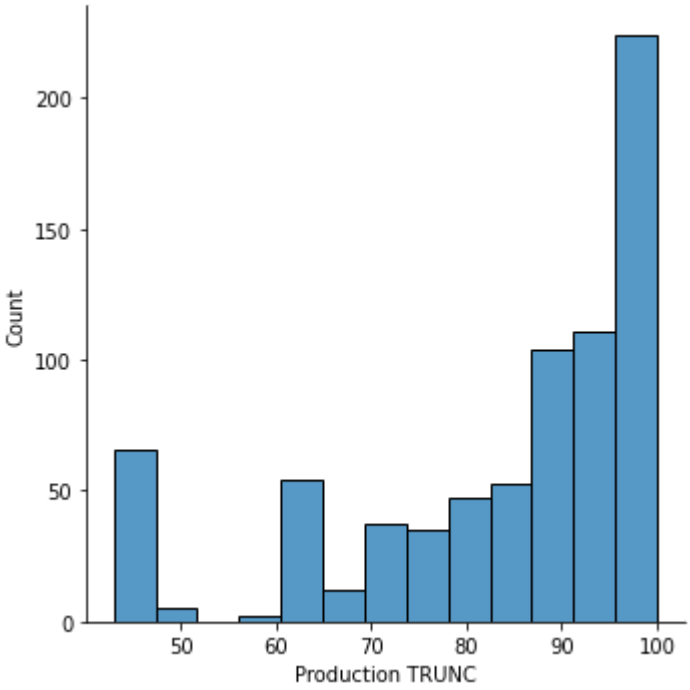
## Histogram and Normal Probability Plot for the attribute height

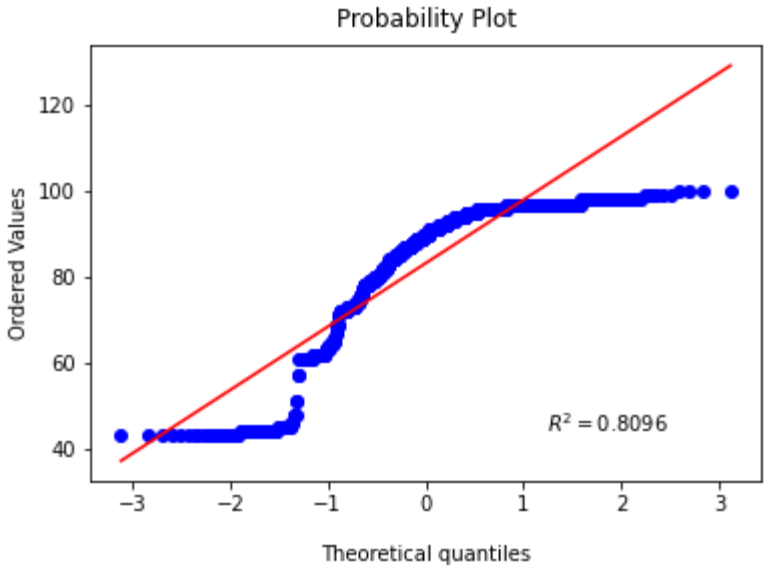



As for **height** , by looking at the histogram and the normal probability plot, we can see that it is skewed towards right and is not completely normally distributed.

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

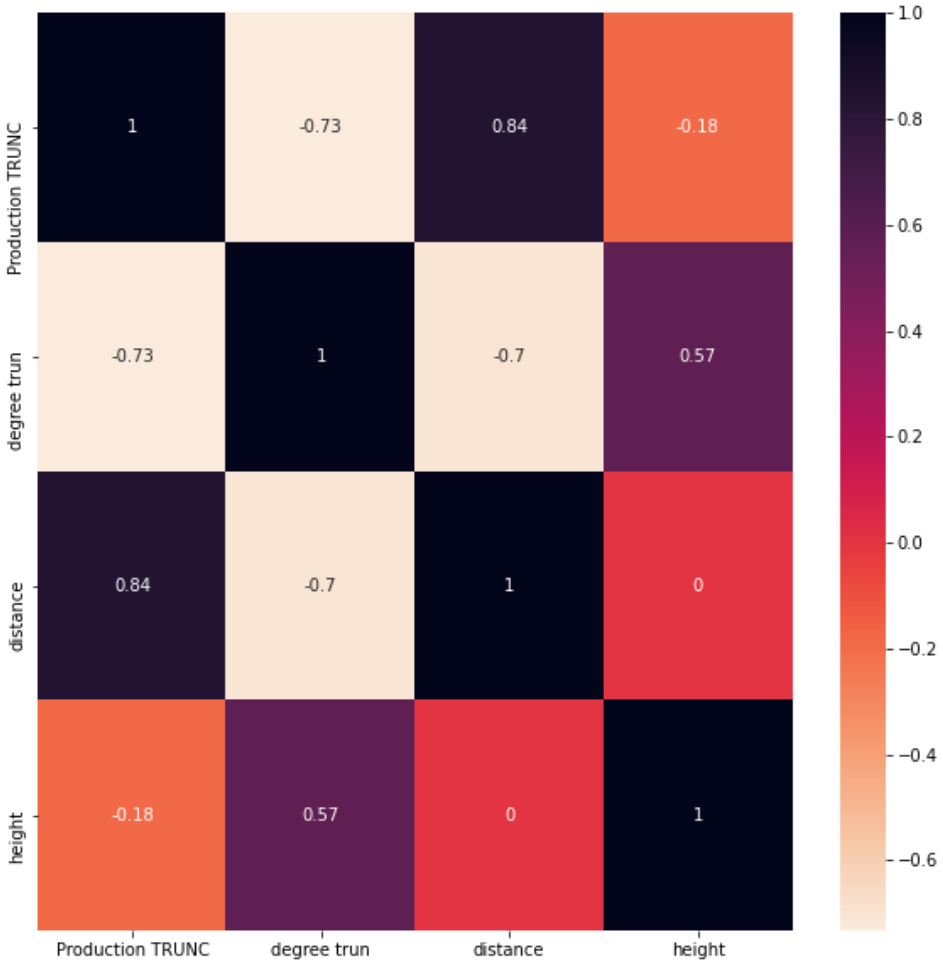



For **Production TRUNC** the graph is skewed towards the left as it is visible from histogram and probability plot.

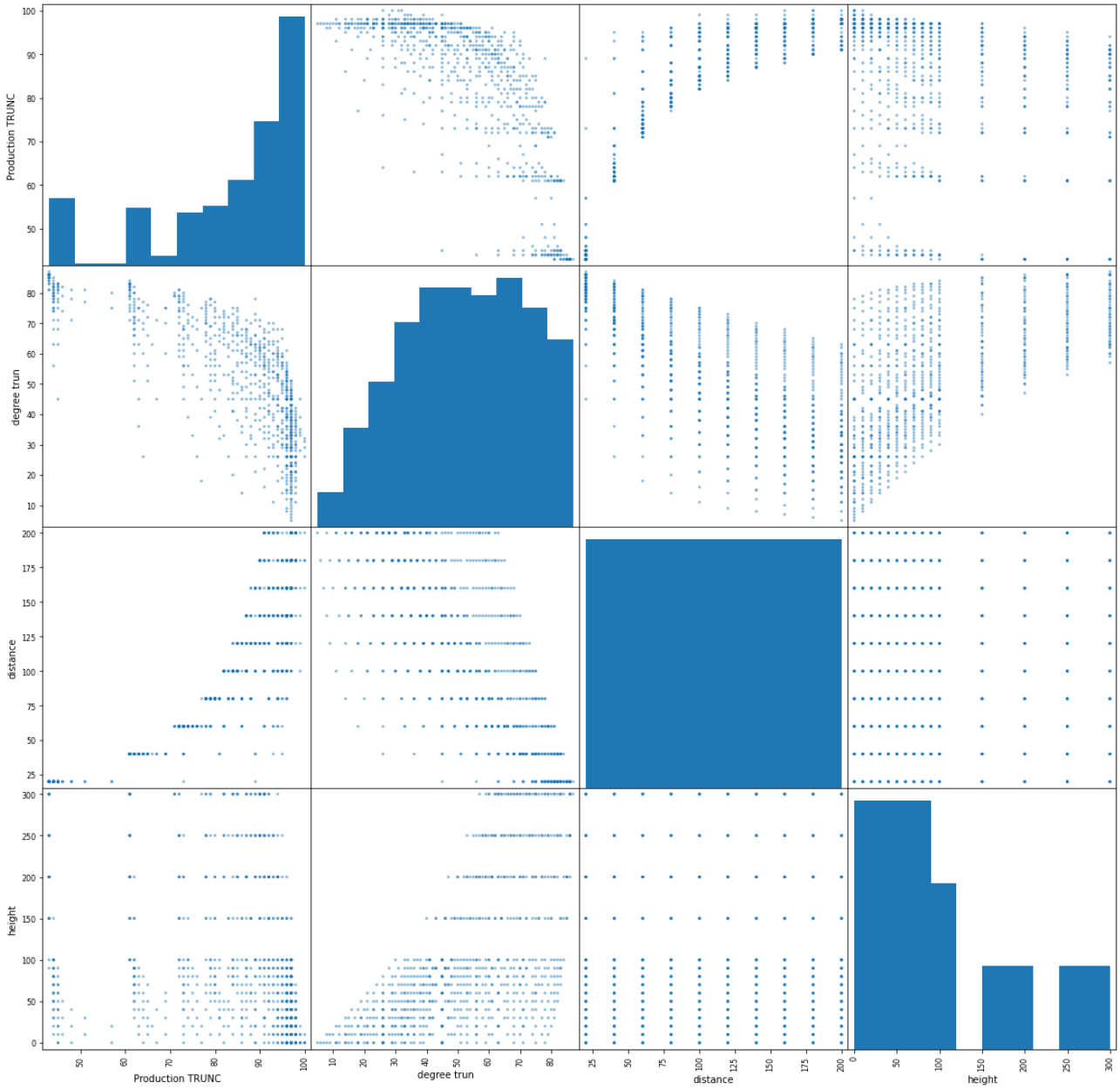
## Correlation between the variables

## HeatMap

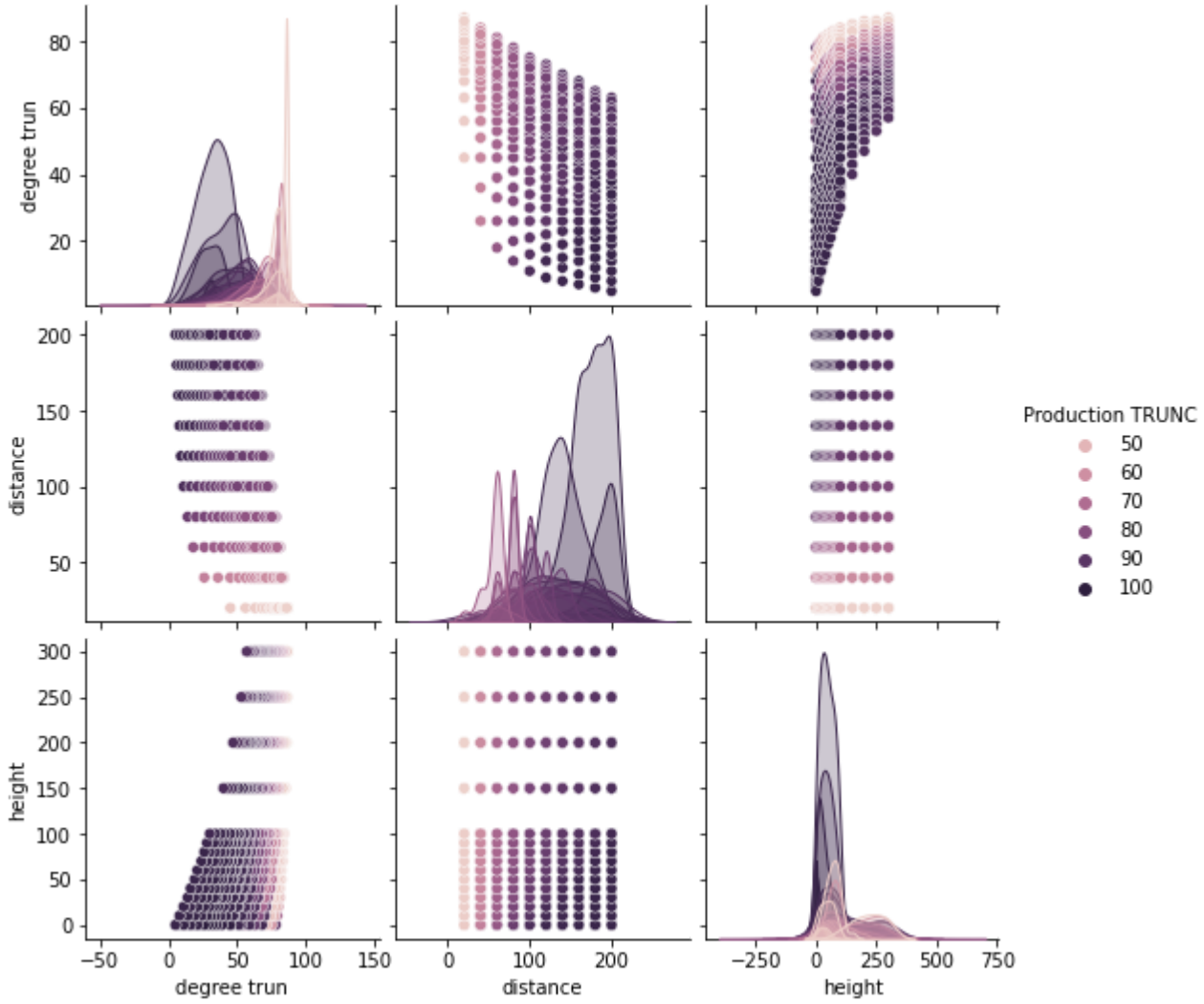



The attributes have both positive and negative correlations. It is interesting to notice that **degree trunc is negatively correlated with the distance and Production Trunc**. On the **other** hand, **distance is strongly correlated with Production Trunc and negatively with height**. Heigth is not very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

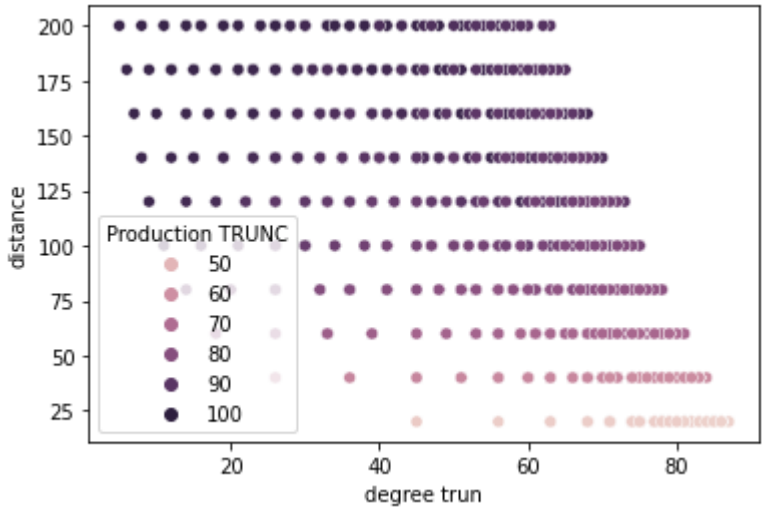
# Scatter Matrix

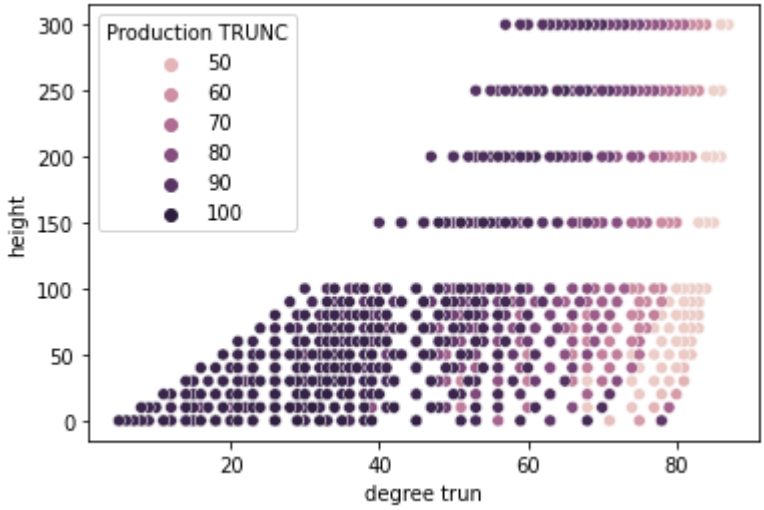


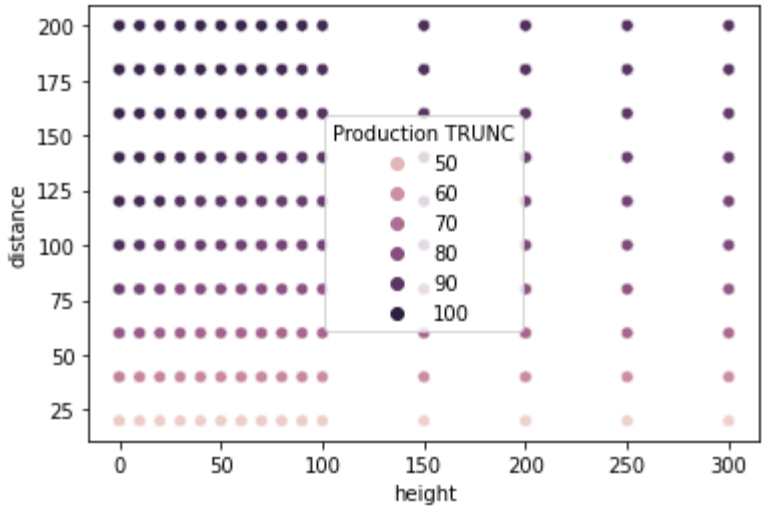
## PairPlot




## Plotted Correlation of Specific Attributes


Production TRUNC
50
60
70
80
90
100
distance
degree trun


Production TRUNC
50
60
70
80
90
100
height
degree trun


Production TRUNC
50
60
70
80
90
100
distance
height

# 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. There appear to be various correlations between the continuous attributes that can be used as indicators.

# Applying Linear Regression:

linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). 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 Trunc are highly positively correlated and Degree trunc is highly negatively correlated with the Production Trunc. Note that production trunc is the output variable for which we need to make prediction.

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

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

## Output:

As we only want to predict Production Trunc so **Production Trunc** is 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 there standard deviation is 1. This technique is called **Standard Scalar.**

Now we are ready to train the model.

## Before Preprocessing

| | degree trun | distance | height |
|---|---|---|---|
| **0** | 78 | 20 | 0 |
| **1** | 79 | 20 | 10 |
| **2** | 80 | 20 | 20 |
| **3** | 81 | 20 | 30 |
| **4** | 81 | 20 | 40 |

## After Preprocessing

| | degree trun | distance | height |
|---|---|---|---|
| **0** | 1.303383 | -1.566699 | -1.112099 |
| **1** | 1.353910 | -1.566699 | -0.997054 |
| **2** | 1.404436 | -1.566699 | -0.882010 |
| **3** | 1.454963 | -1.566699 | -0.766965 |
| **4** | 1.454963 | -1.566699 | -0.651920 |

## Model:

We have split our data **20% and 80% for testing and training** respectively. We applied Linear Regression with Input Provided above and predict Output Production Trunc.

## Testing:

After testing on the testing data we get the following results.

| | OUTPUT | Prediction | Absolute Error |
|---|---|---|---|
| **506** | 88 | 88.309296 | 0.309296 |
| **357** | 82 | 77.282577 | 4.717423 |
| **133** | 61 | 62.117435 | 1.117435 |
| **250** | 80 | 74.905805 | 5.094195 |
| **299** | 77 | 70.532935 | 6.467065 |
| **680** | 98 | 104.525177 | 6.525177 |
| **336** | 89 | 82.161660 | 6.838340 |
| **155** | 82 | 71.308382 | 10.691618 |
| **528** | 97 | 96.362783 | 0.637217 |
| **736** | 97 | 109.851483 | 12.851483 |

These are the Original Output Predicted Output and the absolute error of the testing dataset.

## Error:

So the error values of the model are

```
Mean Absolute Error: 6.8702456312522315
Mean Squared Error: 79.33291892413993
Root Mean Squared Error: 8.906902880583123
```

# 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 Trunc and Production Trunc and a strong positive correlation between Distance and Production Trunc. Height does not have any correlation with Production Trunc but has a weak correlation with Degree Trunc.
2. As there is correlation between attributes and the output is numerical value so it is a good idea to **use Regression.** We have used Simple Linear Regression to fit a line to predict on the testing Data.
3. We used all three attribute to train the Linear Model. Leaving Height attribute was not a good idea as Height is correlated with degree Trunc.
4. The **mean squared error comes out to be 79.33** and **absolute error comes out to be 6.87** which means the for every test observation we had a difference of 6.87 units between original and predicted values.

## Note:

The Dataset was **not adequate** as there were only **750 observations.** From these 750 observations 20% (150 observations) were used as test data. **600 observation** for training are not enough. We should try to **increase the number of observation to get better results.**