Machine Learning

Assignment 1

Name: Tavleen Bajwa

Roll no.: PGDBD202204017

Part A: Linear Regression

Description: The Garment Industry is one of the key examples of the industrial globalization of this modern era. It is a highly labour-intensive industry with lots of manual processes. Satisfying the huge global demand for garment products is mostly dependent on the production and delivery performance of the employees in the garment manufacturing companies. So, it is highly desirable among the decision makers in the garments industry to track, analyse and predict the productivity performance of the working teams in their factories. This dataset is to be used for regression purpose by predicting the productivity range (0-1).

Preliminary Dataset Analysis:

"garments_worker_productivity.csv" is a regression dataset consists of shape 1153 rows and 12 columns. It has 506 null values in the "wip" column and 0 duplicated rows

Column Description:

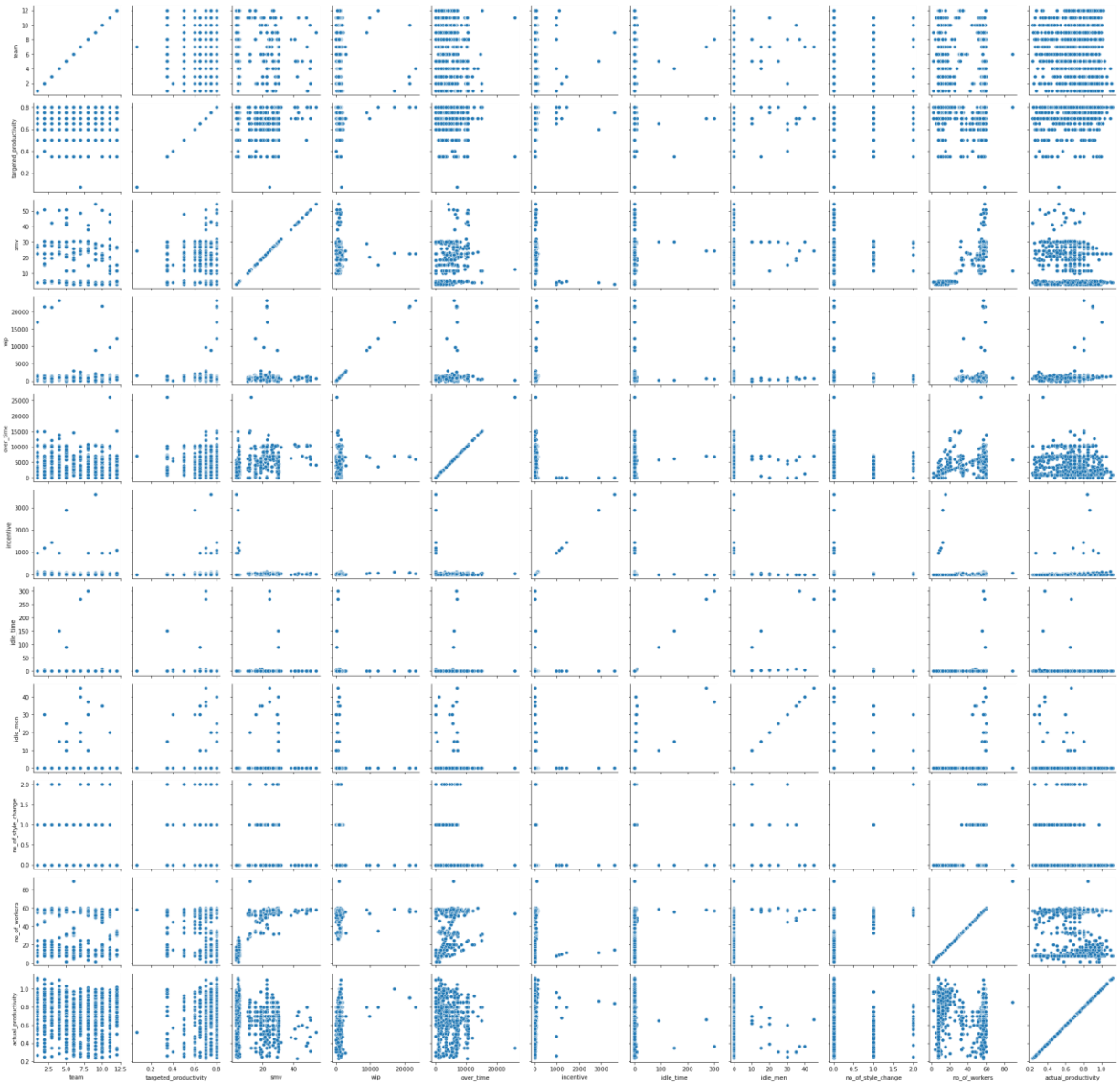
1. date : Date in MM-DD-YYYY
2. day : Day of the Week
3. quarter : A portion of the month. A month was divided into four quarters
4. department : Associated department with the instance
5. team_no : Associated team number with the instance
6. no_of_workers : Number of workers in each team
7. no_of_style_change : Number of changes in the style of a particular product
8. targeted_productivity : Targeted productivity set by the Authority for each team for each day.
9. smv : Standard Minute Value, it is the allocated time for a task
10. wip : Work in progress. Includes the number of unfinished items for products
11. over_time : Represents the amount of overtime by each team in minutes
12. incentive : Represents the amount of financial incentive (in BDT) that enables or motivates a particular course of action.
13. idle_time : The amount of time when the production was interrupted due to several reasons
14. idle_men : The number of workers who were idle due to production interruption
15. actual_productivity : The actual % of productivity that was delivered by the workers. It ranges from 0-1.

Column datatypes:

1. Categorical : date, quarter, department, day

2. Discrete: team, over_time, incentive, idle_men, no_of_style_changes
3. Continuous: targeted_productivity, smv, wip, idle_time, no_of_workers, actual_productivity

Scatter plot of all the columns in raw dataset using Seaborn:



1. Independent variables (x1, x2…x14): date, quarter, department, day, team, over_time, incentive, idle_men, no_of_style_changes, targeted_productivity, smv, wip, idle_time, no_of_workers.
2. Dependent variable (y): actual_productivity

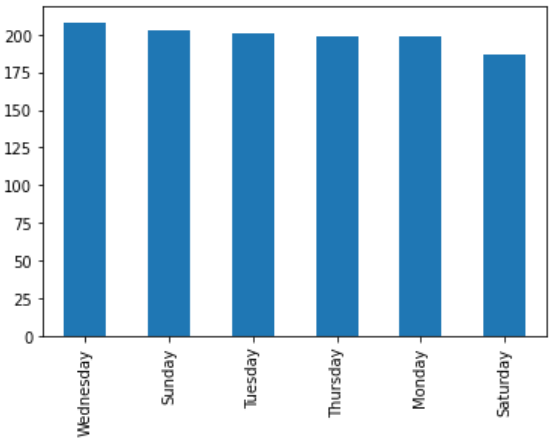
From the above scatterplot it is observable that with w.r.t actual_productivity “target_productivity”, “over_time”, “no_of_workers”, “wip” show linear correlation.

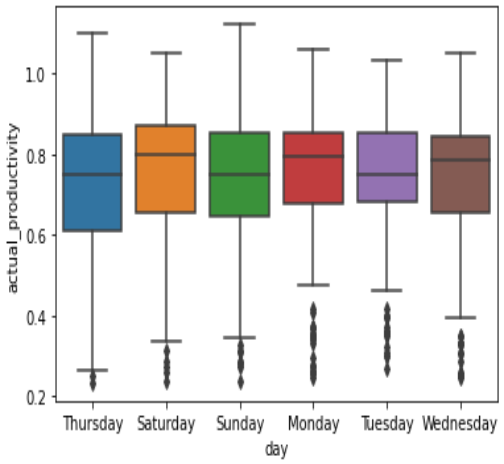
Analysis of Columns of the Dataset:

The correlation between the x independent columns was done w.r.t y (“actual_productivity”) to see if they are significant enough to train the model.

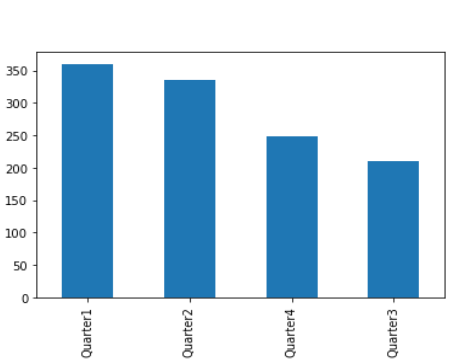
1. **“date” column**
   59 unique dates were observed in the data. No. works initiated on each date were in range 15-24. No much significance w.r.t actual_productivity. So,was dropped.

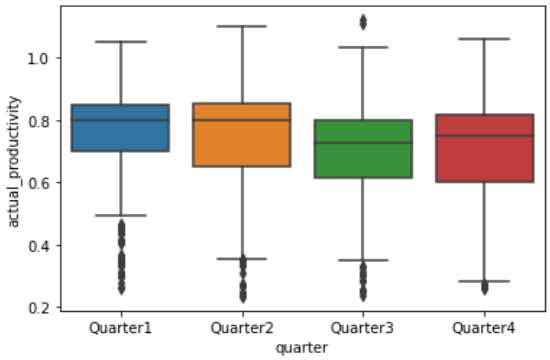
2. **“day” column**
   The value_counts of days of the week in which work was done is shown below. In the barchart and box-whisker shown below, no much variation w.r.t actual productivity was observed. So, can be discarded.





3. **“quarter” column**
   A month is divided into 4 quarters but the value_counts showed 5. Of which 5th column was less in number. So, we dropped the 44 rows of quarter 5. Value_counts of 4 quarters along with box-whisker plot w.r.t productivity is shown below:

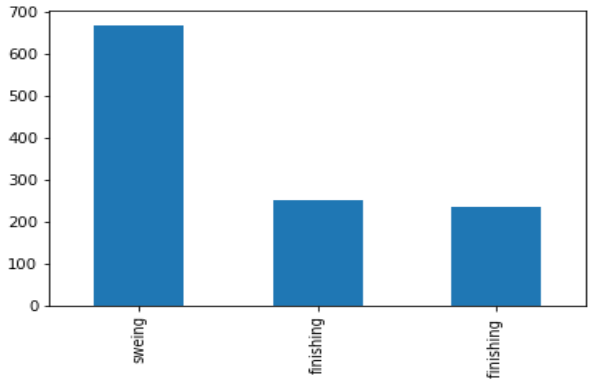


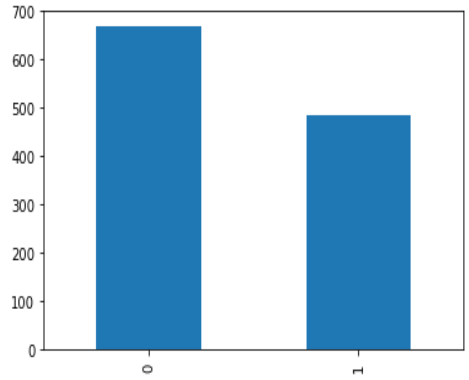


Different quarters show minute differences in actual_productivity. All 4 lie in the range 0.6 to 0.9 which concludes that this feature can be dropped from our dataset.

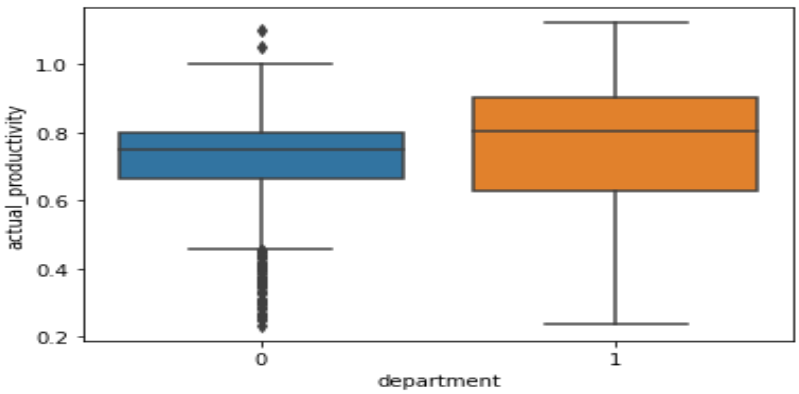
4. **"Department" column**

There were 2 main departments associated with the instance: sweing, finishing. finishing is appearing as 2 seperate columns so will try to replace it using 0 and 1



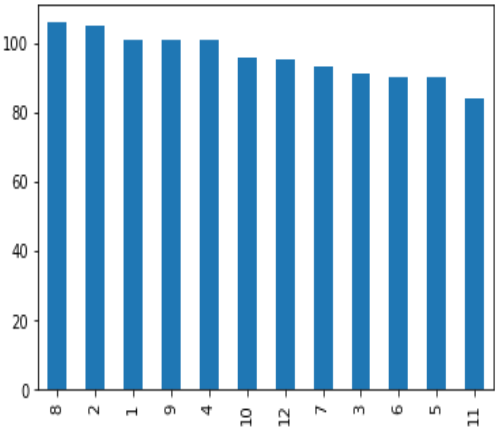


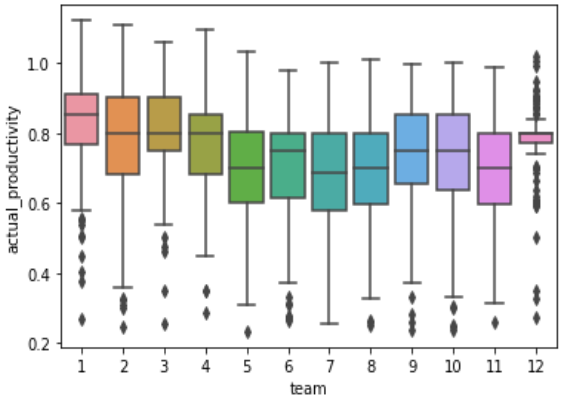
Box-whisker plot between department and actual_productivity showed no significance difference, mostly ranges between 0.6 - 0.9 roughly, also in case of sweing outliers was observed, so can be dropped.



5. **"teams" column**

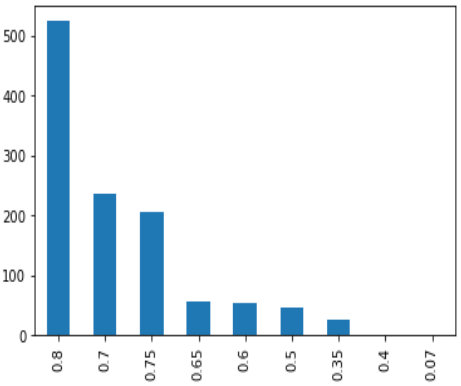
There were 12 unique teams associated with instance. Value_counts of different teams is shown below. Also, different teams show different range of actual_productivity as shown in the Box-Whisker plot.

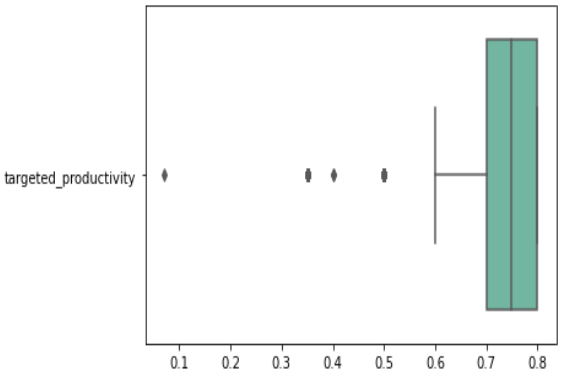




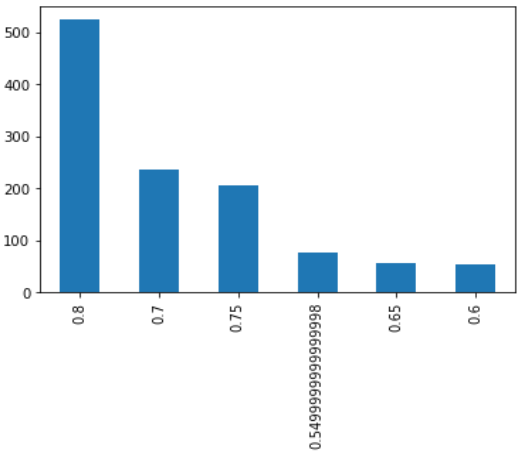
6. **"targeted_productivity" column**

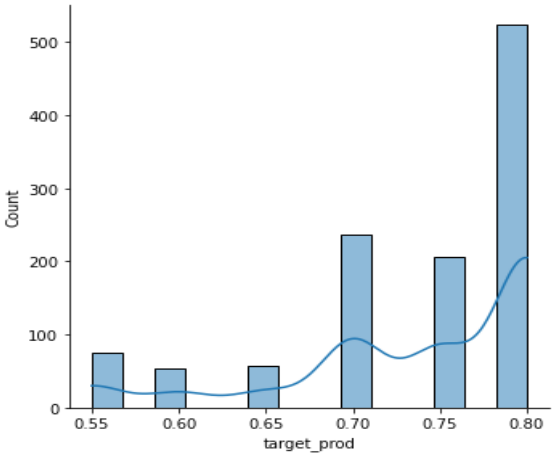
Targeted productivity set by the Authority for each team for each day. The value_counts showed a range from 0.07 to 0.8 and boxplot showed presence of some outliers as shown below

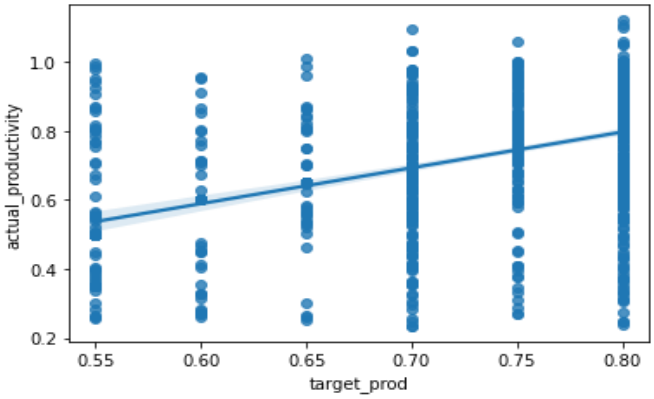




Outliers were removed from the column. Value_counts, distplot and replot post outlier removal are shown below.

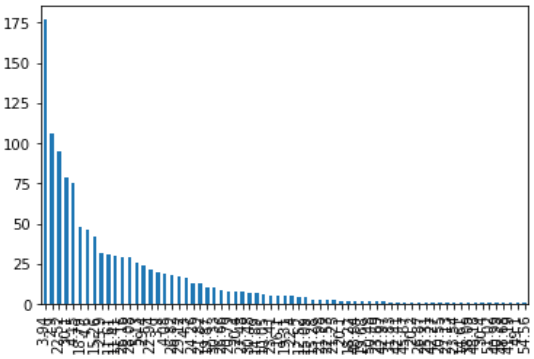


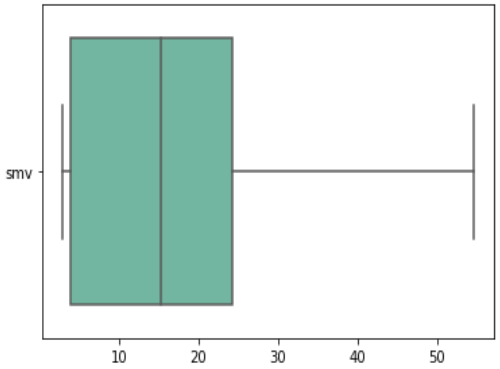


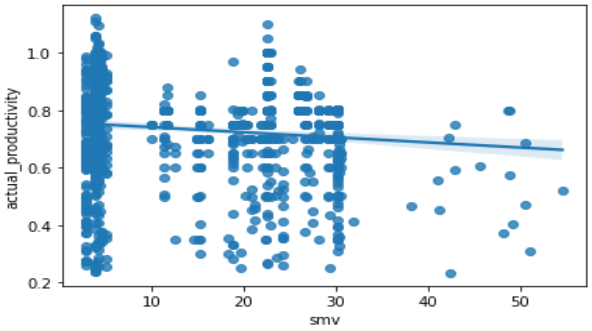


7. "smv" column

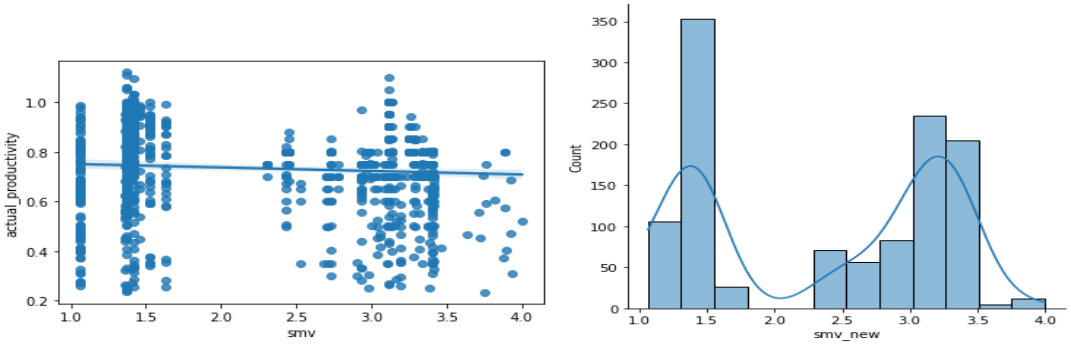
Standard Minute Value is the allocated time for the task. Value_counts shown below showed a lot of skewness. Box-plot shown below showed no outliers in the data. Regplot w.r.t actual productivity showed some linear correlation.





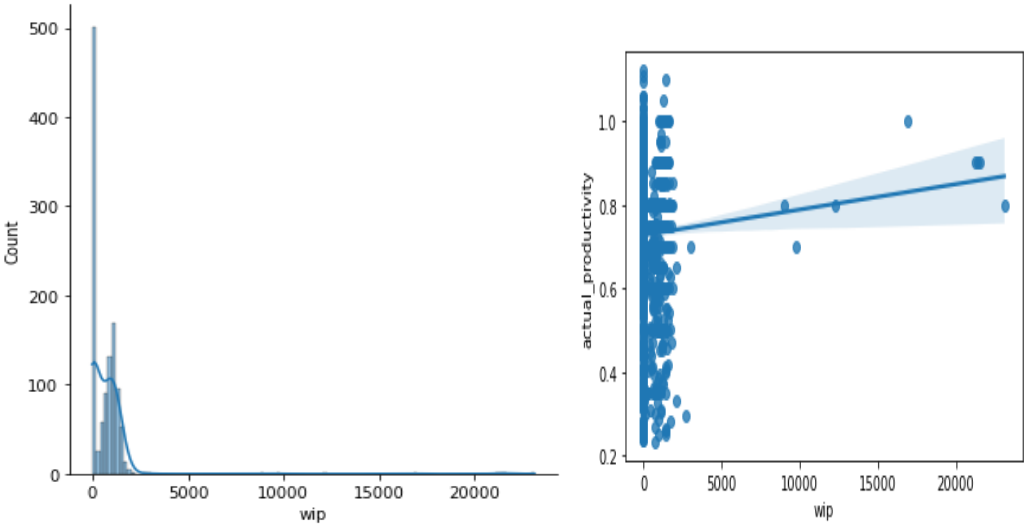


Log transformation was done on column data to reduce skewness. Regplot and distplot post log transformation is shown below.

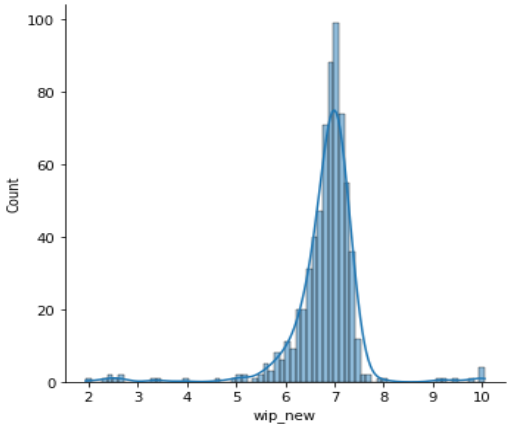


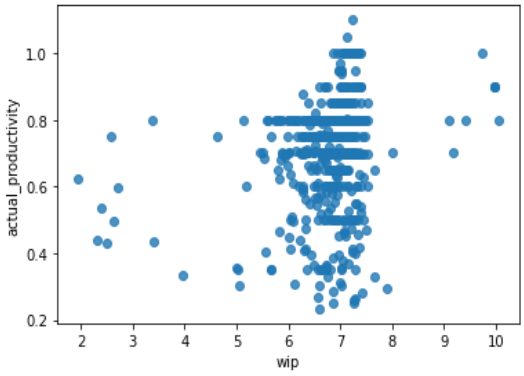
8. "wip" column

Work in progress includes the number of unfinished items for the products. Wip column had null values which were imputed with 0, assuming NaN as no work in progress. Value_counts distplot and regplot before log transformation are shown below.



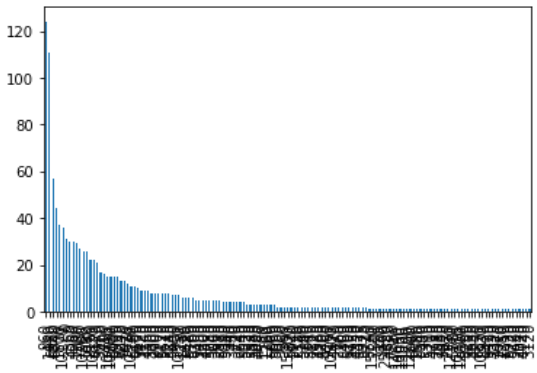
Value_counts distplot and regplot post log transformation is shown below

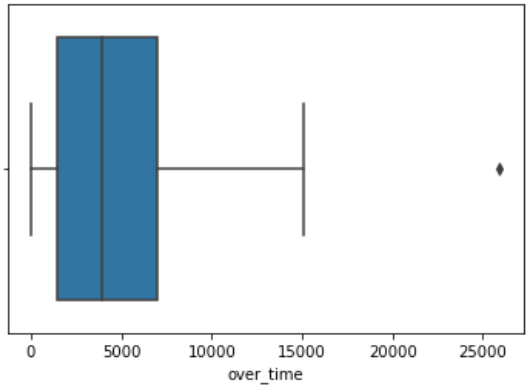


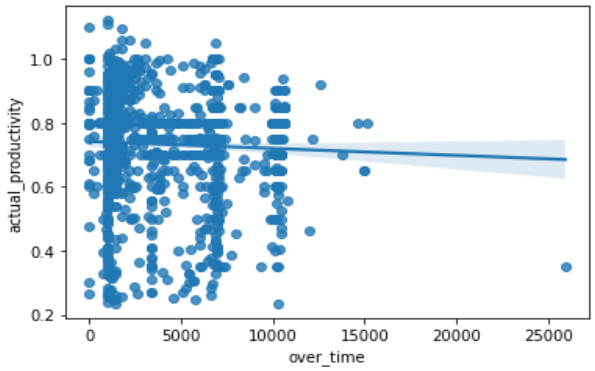


9. "over_time" column

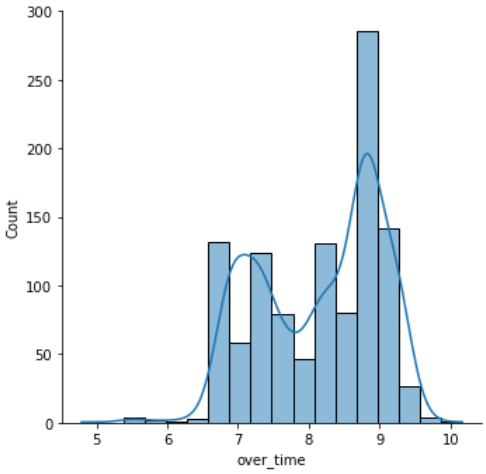
Represents the amount of overtime taken by each team in minutes. Value_counts, regplot wrt actual_productivity and boxplot are shown below.

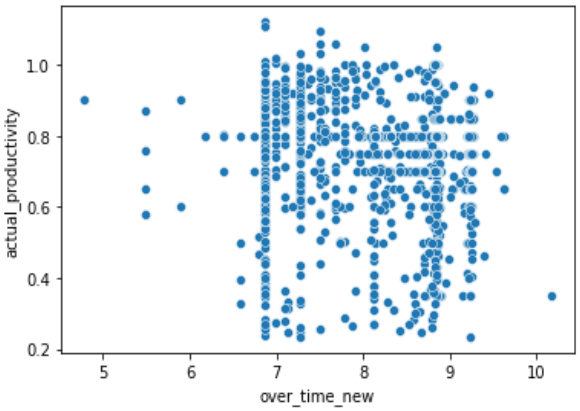






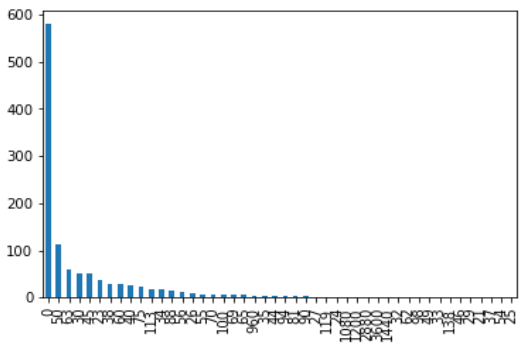
Post log transformation inorder to reduce skewness in the data, Bimodal distribution was observed as shown below in the distplot and regplot.

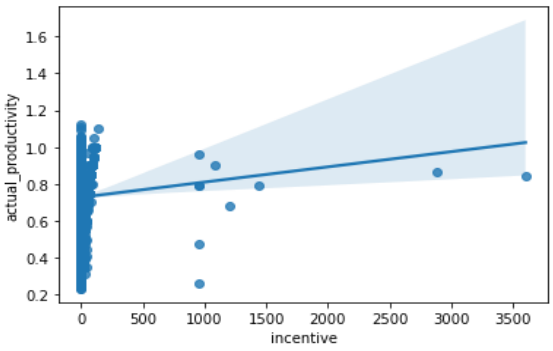




10. “incentive” column

Represents the amount of financial incentive (in BDT) that enables or motivates a particular course of action. value_counts and regplot wrt actual_productivity is shown below

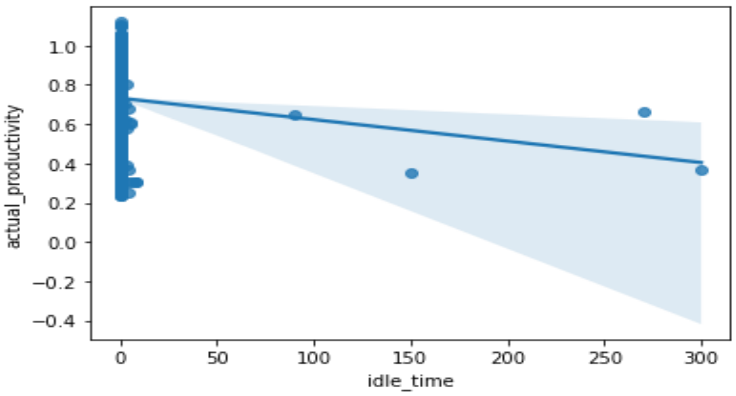




Data in incentive column is highly skewed but it also shows a positive linear correlation with actual_productivity.

11. “idle_time” column

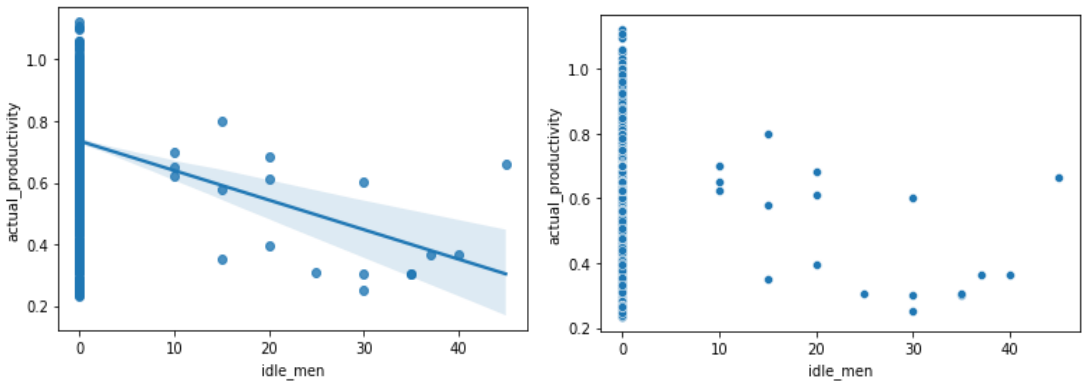
The amount of time when the production was interrupted due to several reasons. Regplot shown below



No significant correlation and most of the values are 0, so can be dropped.

12. "idle_men" column

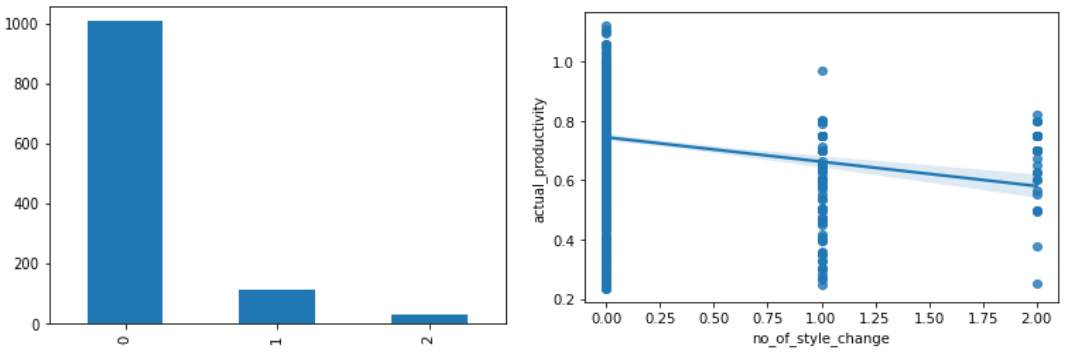
The number of workers who were idle due to production interruption. Scatterplot shown below.

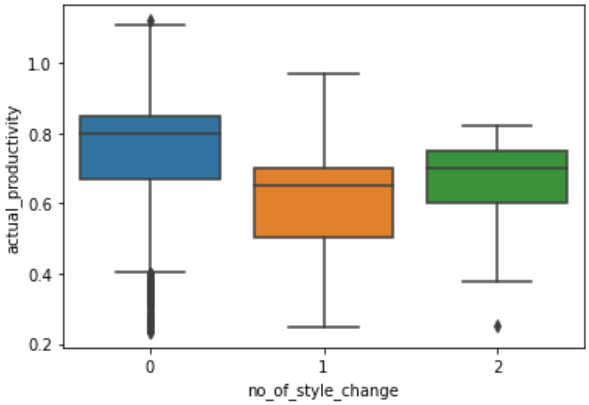


Highly negative correlation with productivity and mostly values are 0 so can be removed. Also, there is some multicollinearity between "idle_time" and "idle_men" which can make the prediction poor therefore we will discard it.

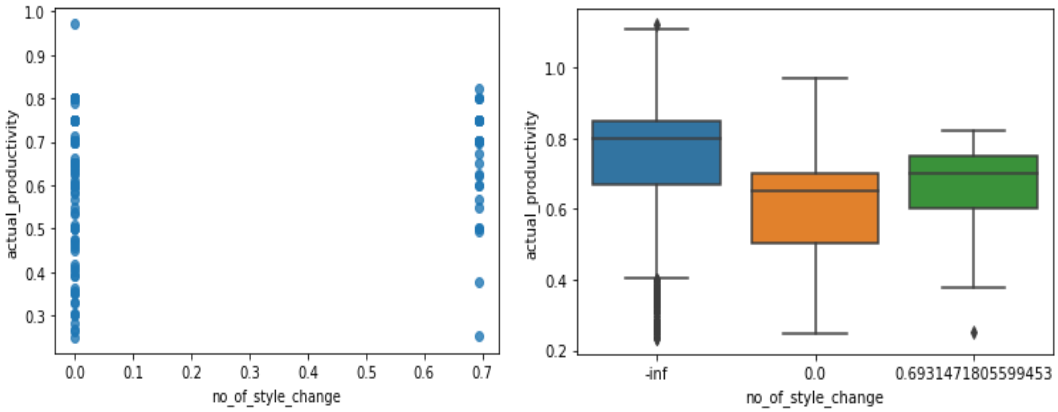
13. "No_of_style_changes" column

Number of changes in the style of a particular product. Value_counts and regplot shown below



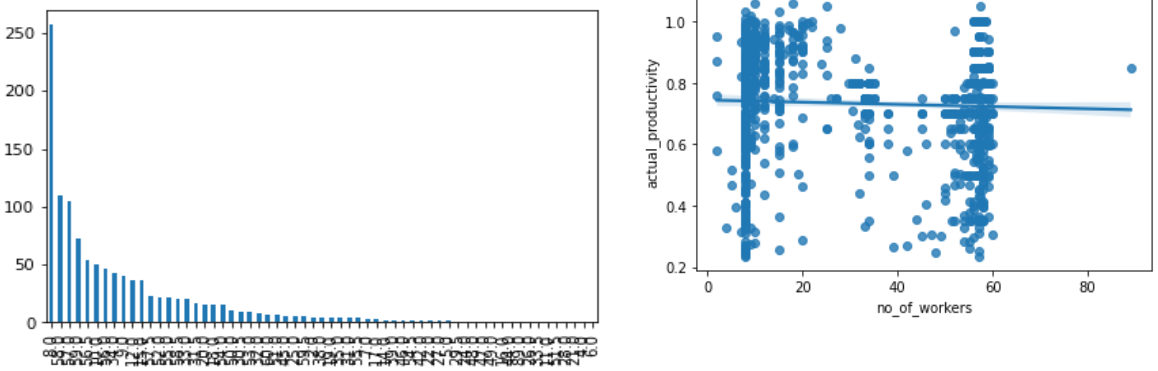


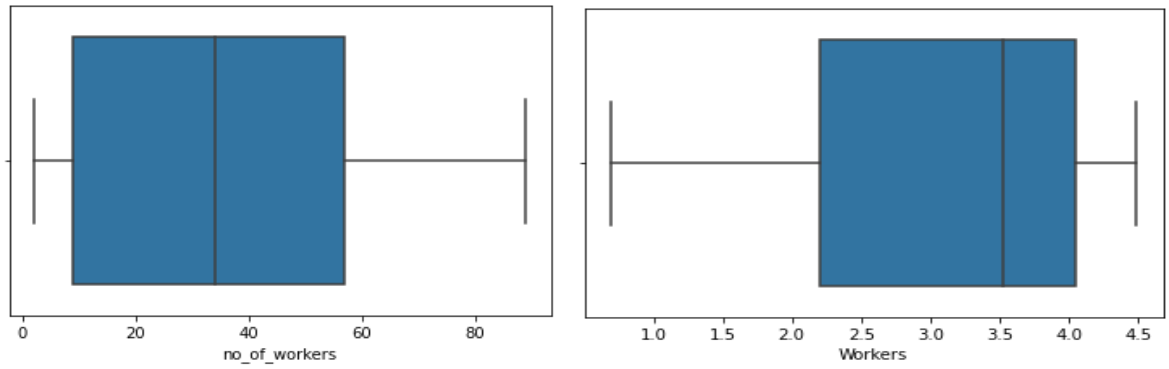
Post log transformation



14. "no_of_workers"

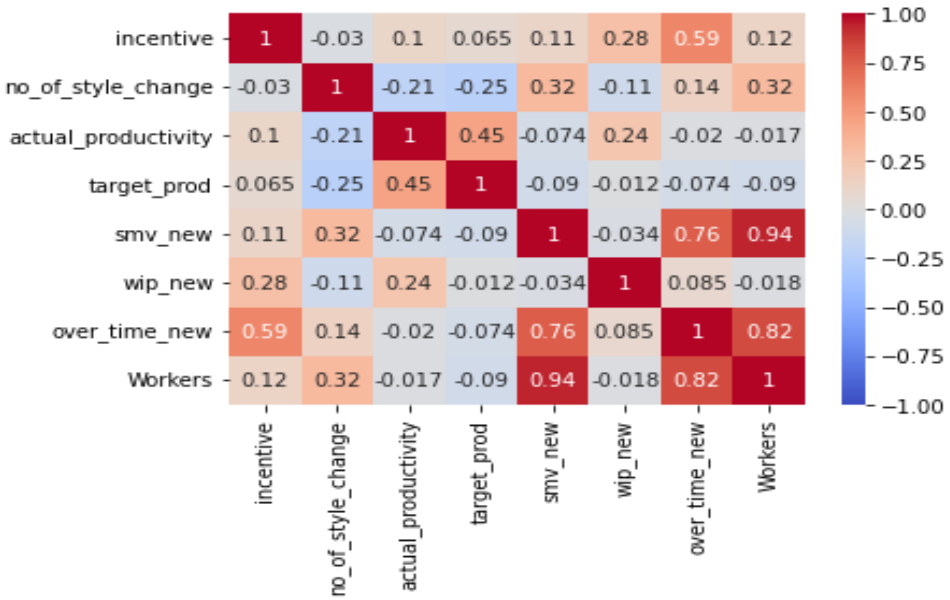
Represents the no. workers in each team. Value_counts and scatterplot are shown below.





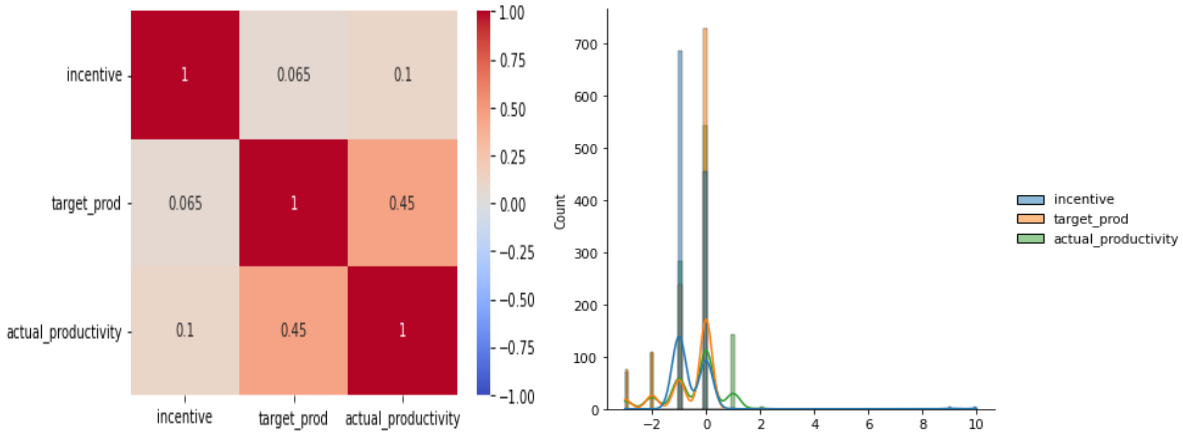
Exploratory Data Analysis and Feature Selection:

All the numerical columns were normalized to make the data scaled so that the Gradient Descent, Closed Form methods so the converging is faster. We further analysed heatmap of numerical columns for feature selection using seaborn.



Wrt to actual_productivity: Columns like target_prod, wip_new, incentive_new show positive correlation.

Post normalization all wip values have become Nan, so we can remove that column



Results:

Multivariate Linear Regression:

The dataset used for Multivariate linear regression had final columns : Incentive, Target_prod, actual_productivity.

All the values were normalised followed by extraction of x: independent (Incentive, Target_prod) features and y:dependent features (actual_productivity)

Followed by Train and Test split

X_train had 923 rows and 2 columns, x_test had 230 rows and 2 columns

Y_train had 923 values and y_test had 230 values.

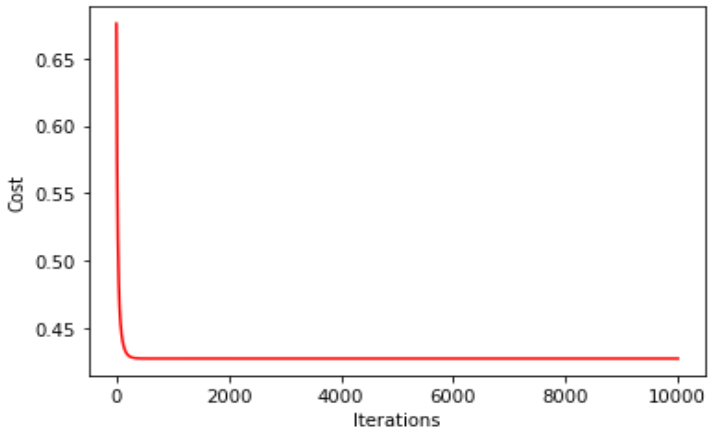
An extra column was added as bias term(Wo). Followed by loss function calculation and optimal value of weights for incentive and target_prod was determined using gradient descent and further Root mean square error was calculated.

Comparison between Weight values, and RMSE using ScikitLearn library and without using ScikitLearn library is given below:

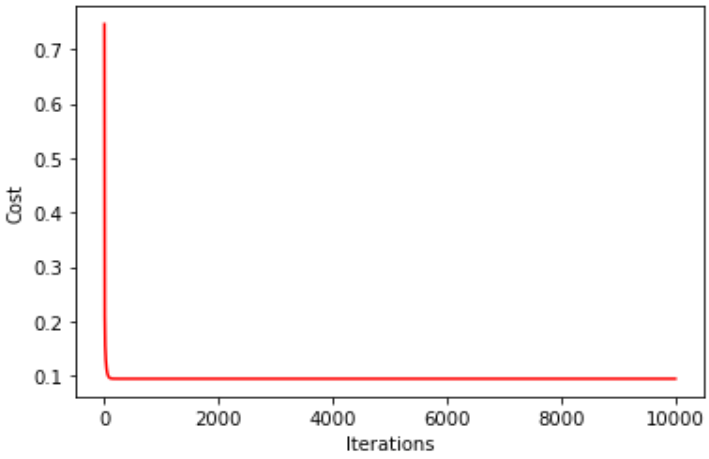
	Train data	**Test data**
Using ScikitLearn	Weights= 0.3935, 0.4654 RMSE = 0.924	Weights = 0.028, 0.425 RMSE = 0.868
W/o SciKitLearn	Weights = 0.393, 0.4654 RMSE = 0.822	Weights = 0.028, 0.425 RMSE = 0.746

Gradient descent plot (Iterations vs Cost)

Train data:



Test data:



Univariate Linear Regression:

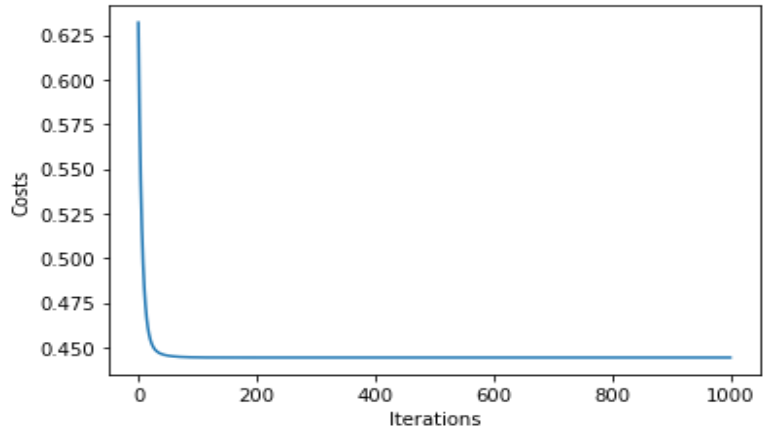
x = target_prod, y = actual_productivity

Closed form:

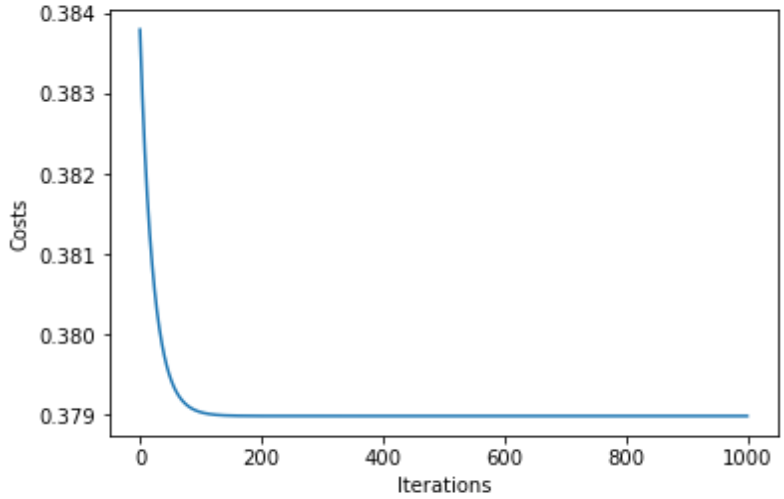
	Train data	**Test data**
Using SciKitLearn	M = 0.528 C = -0.164	M = 0.431 C = -0.276
Without SciKitLearn	M = 0.528 C = -0.164 Train error = 0.888 RMSE value = 0.444	M = 0.431 C = -0.276 Test Error = 0.768 RMSE value = 0.095

Gradient Descent:

Training data:



Test data:



Train data	**Test data**
Cost function = 0.665 M = 0.528 C = -0.16	Cost function = 0.630 M = 0.431 C = -0.276

Part B: Classification Task

Classification Dataset:

Description:

Data were extracted from images that were taken from genuine and forged banknote-like specimens. For digitization, an industrial camera usually used for print inspection was used. The final images have 400x 400 pixels. Due to the object lens and distance to the investigated object gray-scale pictures with a resolution of about 660 dpi were gained. Wavelet Transform tool were used to extract features from images.

Attributes:

1. variance of Wavelet Transformed image (continuous)
2. skewness of Wavelet Transformed image (continuous)
3. curtosis of Wavelet Transformed image (continuous)
4. entropy of image (continuous)
5. class (integer)

Preliminary Data Analysis:

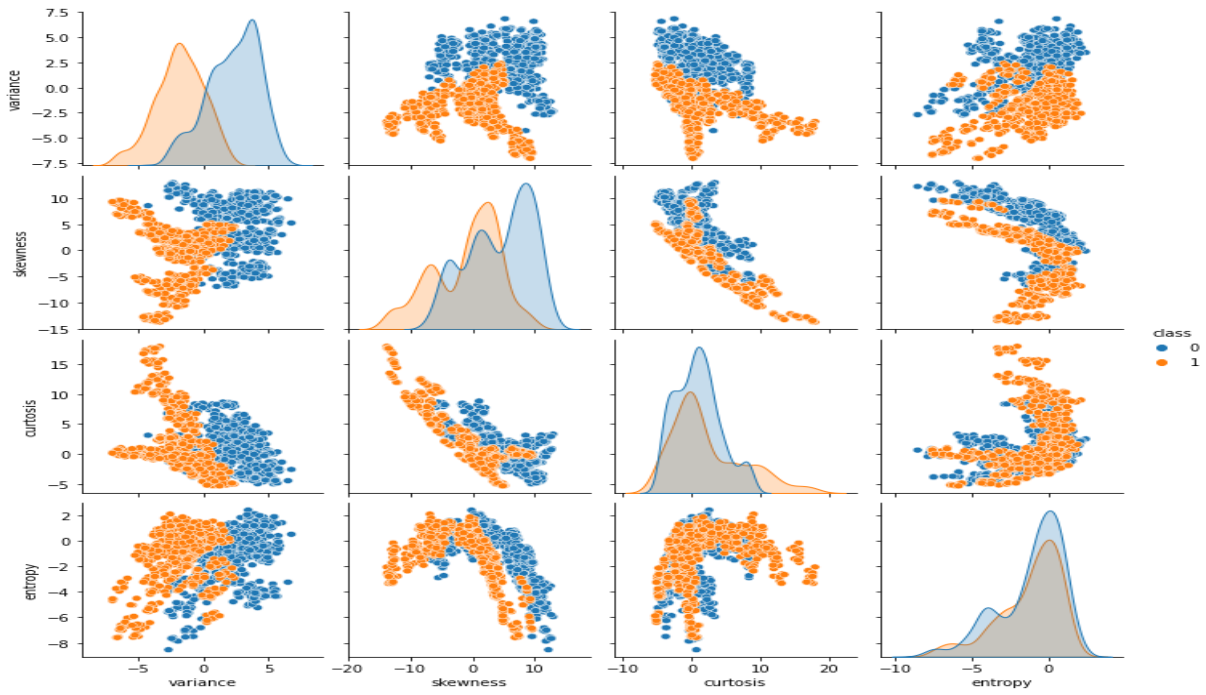
data_banknote_authentication.txt dataset has 1372 rows and 5 columns. There were 0 null values and 24 duplicated rows which were removed. Final dataset has 1348 rows and 5 columns.

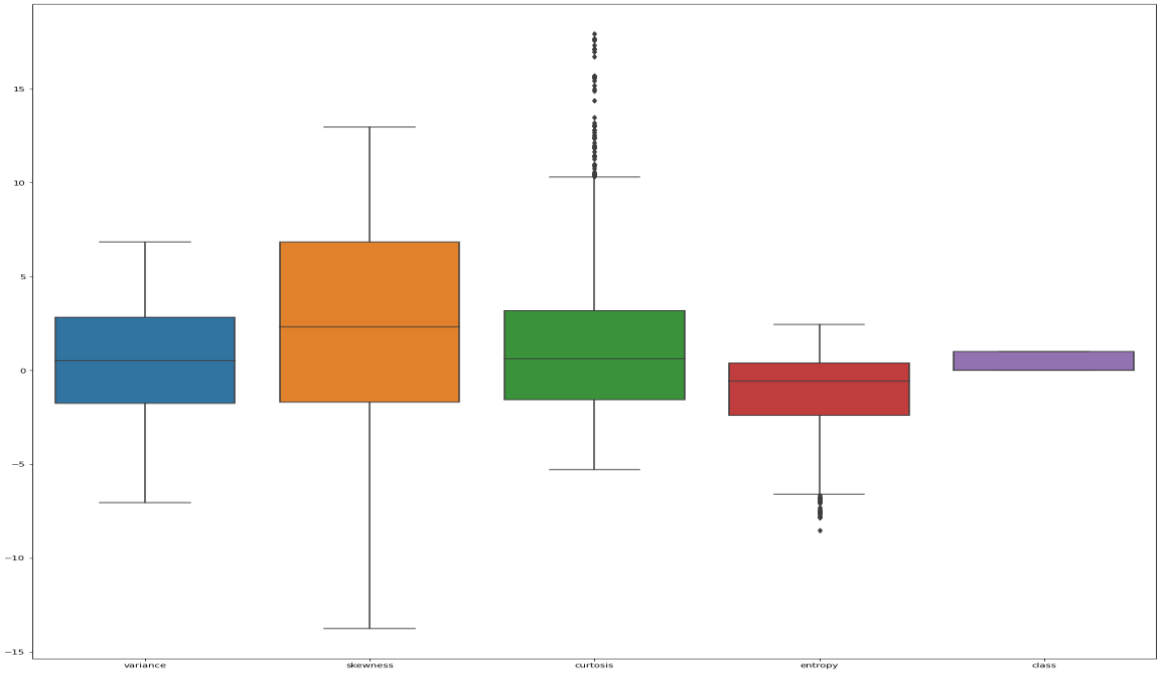
Column wise Data Analysis:

Pairplot of different columns is shown below containing hue which helps in identifying the difference in two set of variables of a dataset and form the most separated clusters to understand the linear separation.

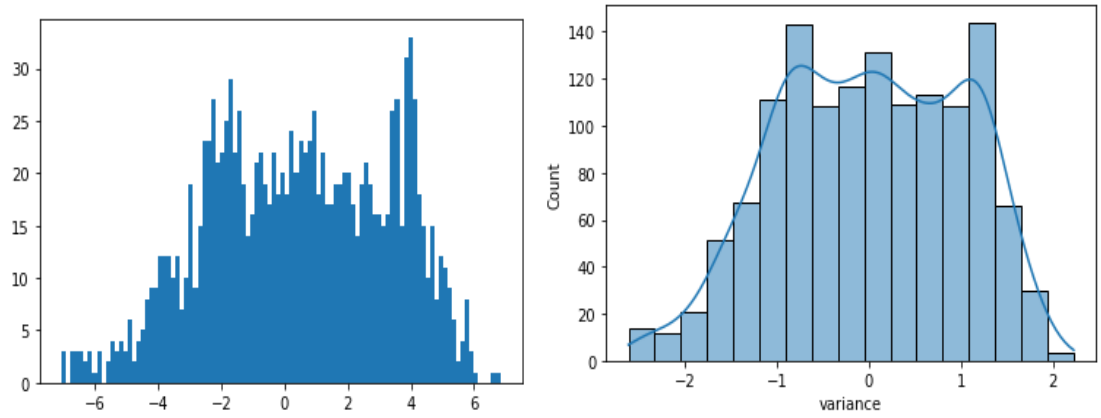
As shown below in the pairplot variance and other 3 variables entropy, curtosis and skewness show best clusters.

Box plot shows that all columns have values in range -10 to +10 (varying range) and curtosis has outliers.

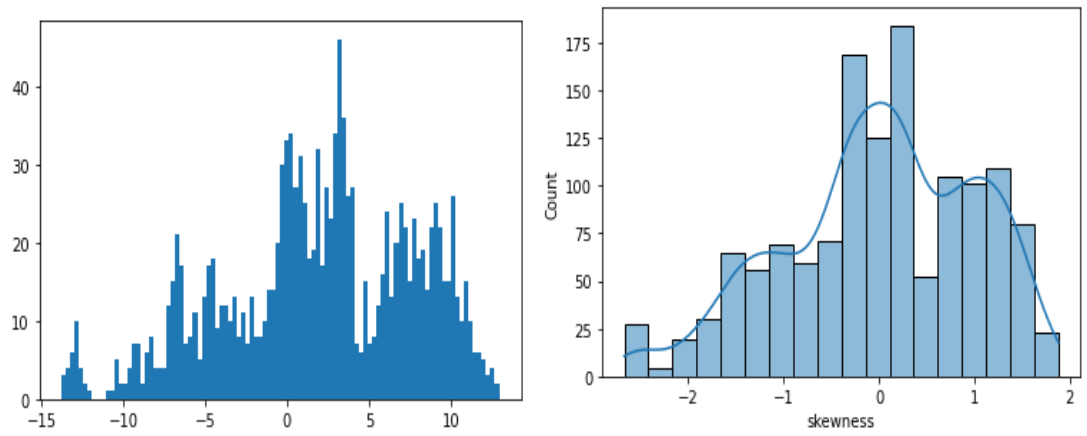




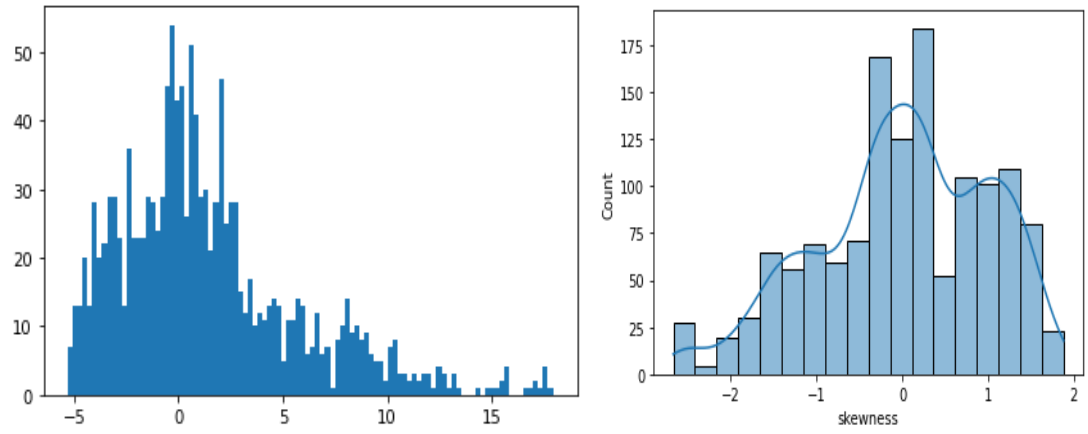
1. **“Variance” column**
   Shows bimodal distribution in the histplot shown below. Distplot post normalization and standardization is shown on right.



2. **“Skewness” column**
   Shows near normal distribution in the histplot shown below. Distplot post normalization and standardization is shown on right.

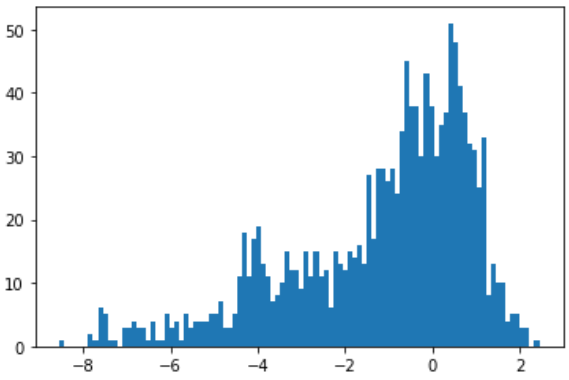


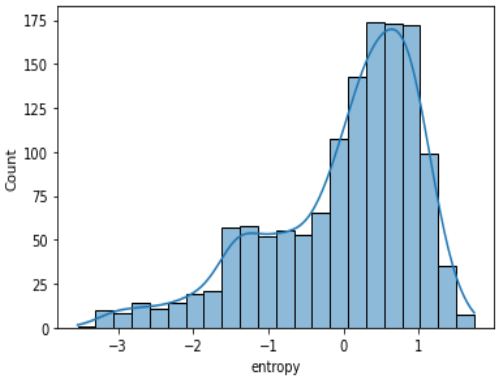
3. **“Curtosis” column**
   Shows near right skewed distribution in the histplot shown below. Distplot post normalization and standardization is shown on right.



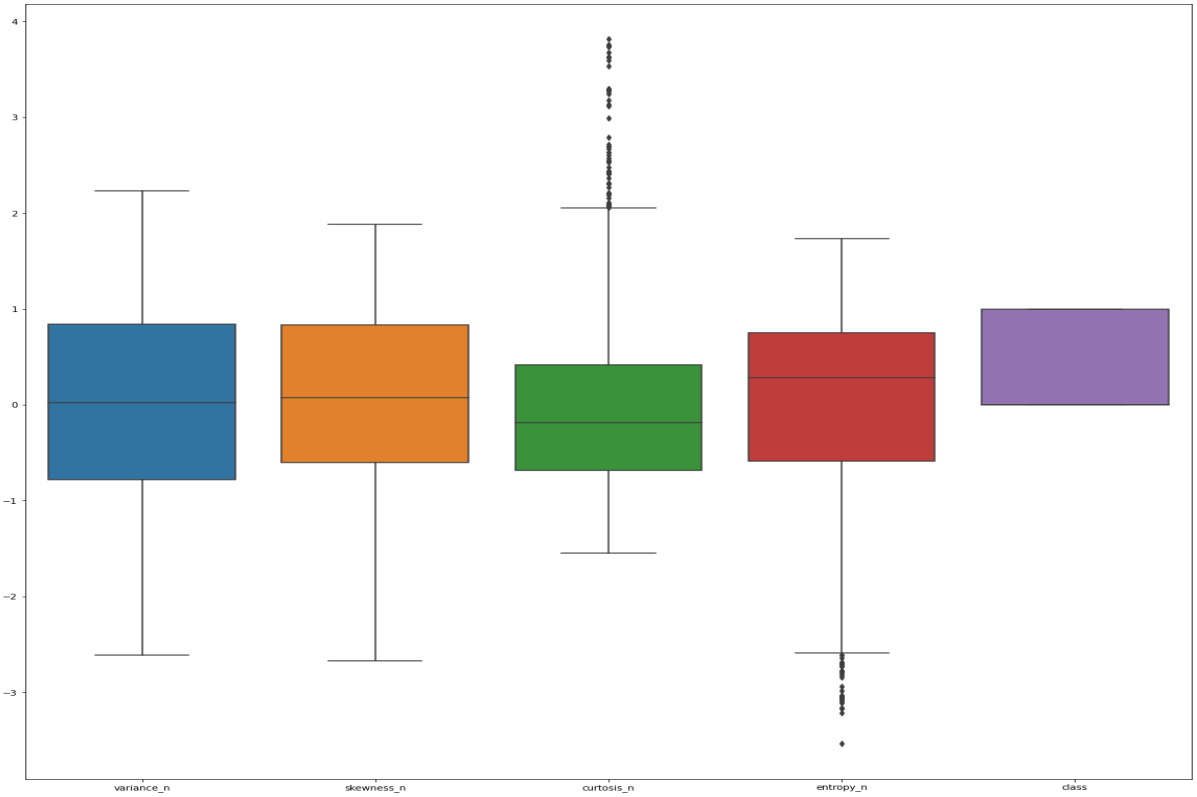
4. **“Entropy” column**

Shows near left skewed distribution in the histplot shown below. Distplot post normalization and standardization is shown on right.





Box plot post Normalization and Standardization:



EDA and Feature selection

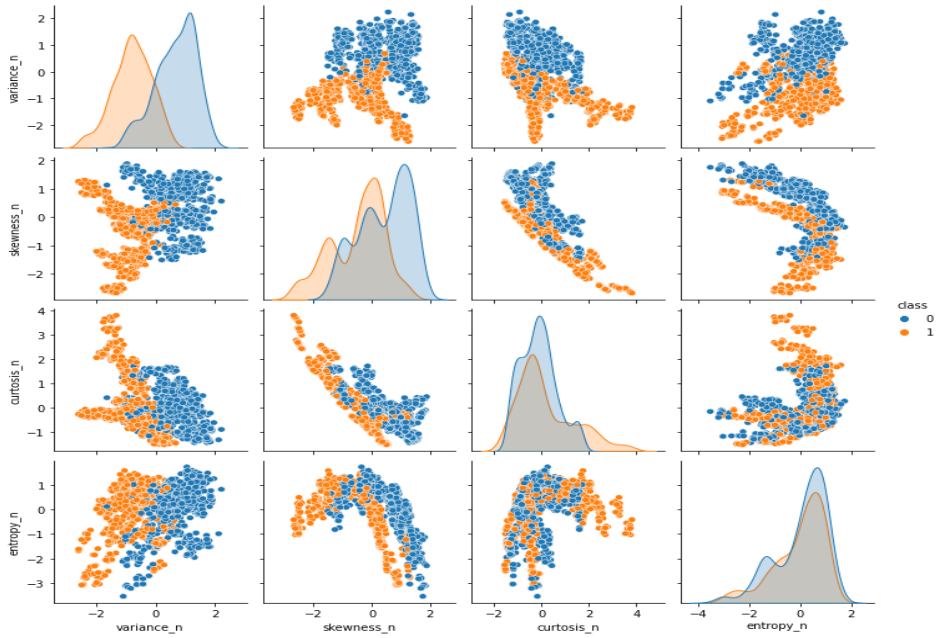
Showed no much variation in case of variance, skewness and entropy. Curtosis has a shorter range but can be considered for initial analysis.

Results:

Naïve Bayes:
Was done on all the columns first and here is the accuracy score calculated.
Dataset had 1348 rows and 5 columns (including “class” column).
x & y were split into train and test data in the ratio 8:2

x_train has 1078 training examples and 4 columns
y_train has 270 test examples and 4 columns

As observed from the pair-plot Variance and skewness were showing better separated clusters so we ran the model on variance and skewness.



Columns	Training data	Test data
Accuracy score (All columns)	0.8441	0.8703
Accuracy score (Variance & Skewness Column only)	0.8738	0.9111

Compared to all columns, the accuracy score was improved a bit when we used Variance and Skewness columns only.

Logistic Regression:

Weights were initialized as vector of ones. Using sigmoid function we calculated predicted values. The values which were below 0.5 were appended as 0 and more than 0.5 were appended as 1 in the final list which was converted into a numpy array.

Using Gradient descent, at the learning rate of 0.001 and 1000 iterations, we calculated the lost function. After each iteration weight new was calculated by subtracting the product of w1(Loss func) and weights (Ws) old calculated at each step.

Columns	**Training data**	**Test data**
All columns	Accuracy: 0.147 Precision: 0.103 Recall: 0.114 Specificity: 0.175 F1 Score: 0.108	Accuracy: 0.147 Precision: 0.103 Recall: 0.114 Specificity: 0.175 F1 Score: 0.108
Variance and Skewness columns only	Accuracy: 0.498 Precision: 0.454 Recall: 0.497 Specificity: 0.498 F1 Score: 0.475	Accuracy: 0.498 Precision: 0.454 Recall: 0.497 Specificity: 0.498 F1 Score: 0.475

Model accuracy was improved a lot when we used Variance and Skewness columns only.