# GAS TURBINE - Predicting Turbine Energy Yield (TEY)

*PRESENTED BY :*

**HARSHAD THORAT**

# AGENDA

- INTRODUCTION
- PROBLEM STATEMENT & GOAL
- ATTRIBUTE INFORMATION
- EXPLORATORY DATA ANALYSIS
- FEATURE SELECTION
- MACHINE LEARNING ALGORITHMS
- ENSEMBLE TECHNIQUES
- DECISION ON ALGORITHM
- REPORT
- CONCLUSION

# INTRODUCTION

- Instances - 36,733
- Features - 11
- Measures gathered over one hour , from a gas turbine located in Turkey
- Predicting **TEY** using ambient & process variables as features

# PROBLEM STATEMENT

- Is there a relationship between the process, ambient variables & **Turbine Energy Yield (TEY)**

- Goal is to analyze a dataset containing **ambient, process and emission variables**

- Discover what relationships might exist between **Turbine Yield Energy (TEY)** **&** the other variables

# ATTRIBUTE INFORMATION

**TARGET COLUMN :**

- **TEY: Turbine Energy Yield (MWH)**

**AMBIENT VARIABLES :**

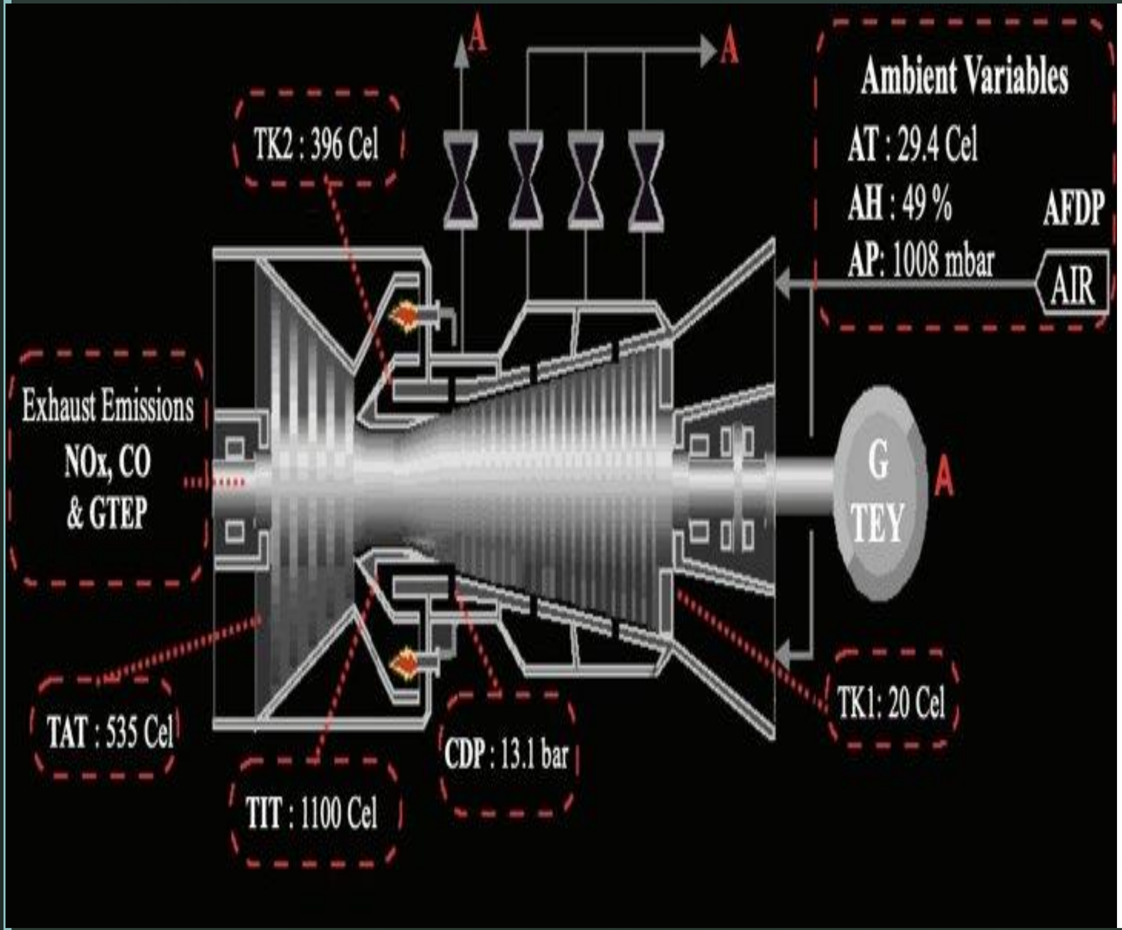
- **AT**: Ambient temperature (C)
- **AP**: Ambient pressure (mbar)
- **AH**: Ambient humidity (%)

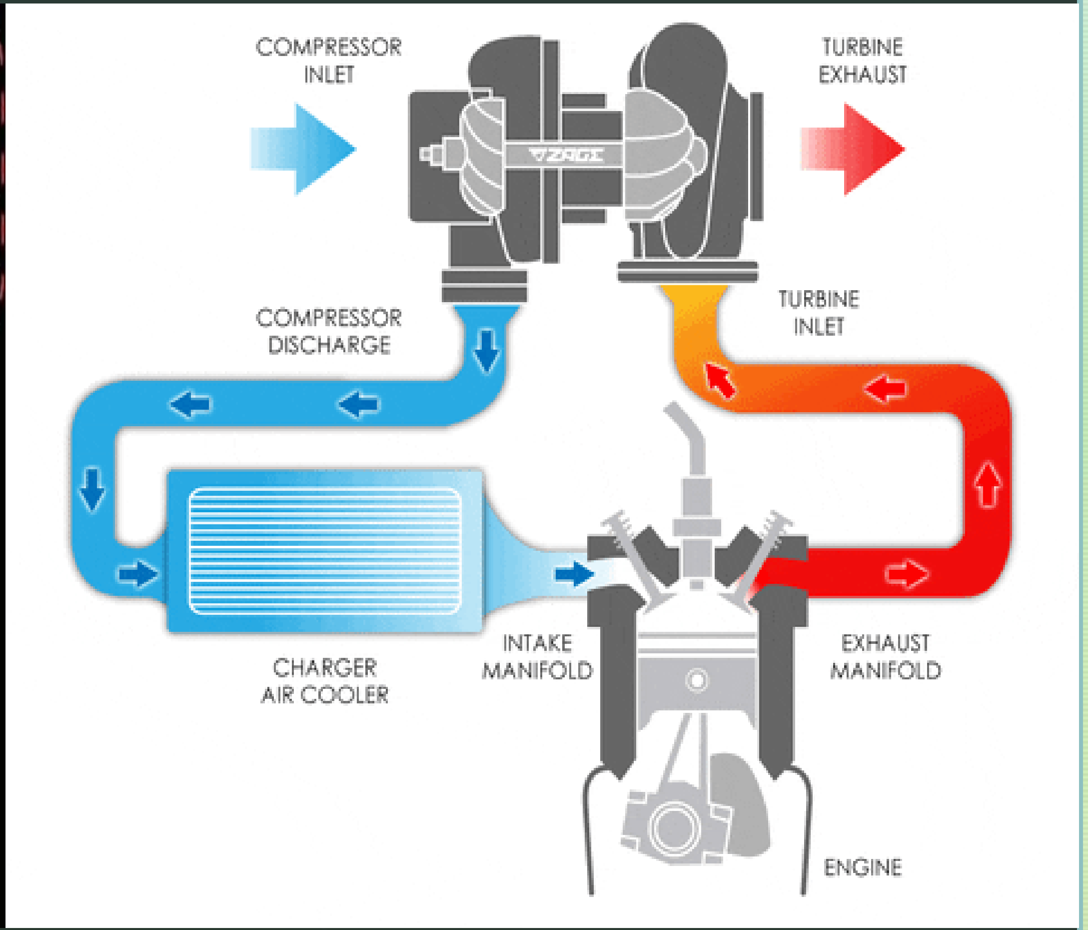
**EMMISION VARIABLES :**

- **CO**: Carbon monoxide (mg/m3)
- **NOX**: Nitrogen oxides (mg/m3)

**PROCESS VARIABLES :**

- **AFDP**: Air filter difference pressure (mbar)
- **GTEP**: Gas turbine exhaust pressure (mbar)
- **TIT**: Turbine inlet temperature (C)
- **TAT**: Turbine after temperature (C)
- **CDP**: Compressor discharge pressure (mbar)




**Pic 1. GAS TURBINE LAYOUT**

**Pic 2. WORKING OF A GAS TURBINE**

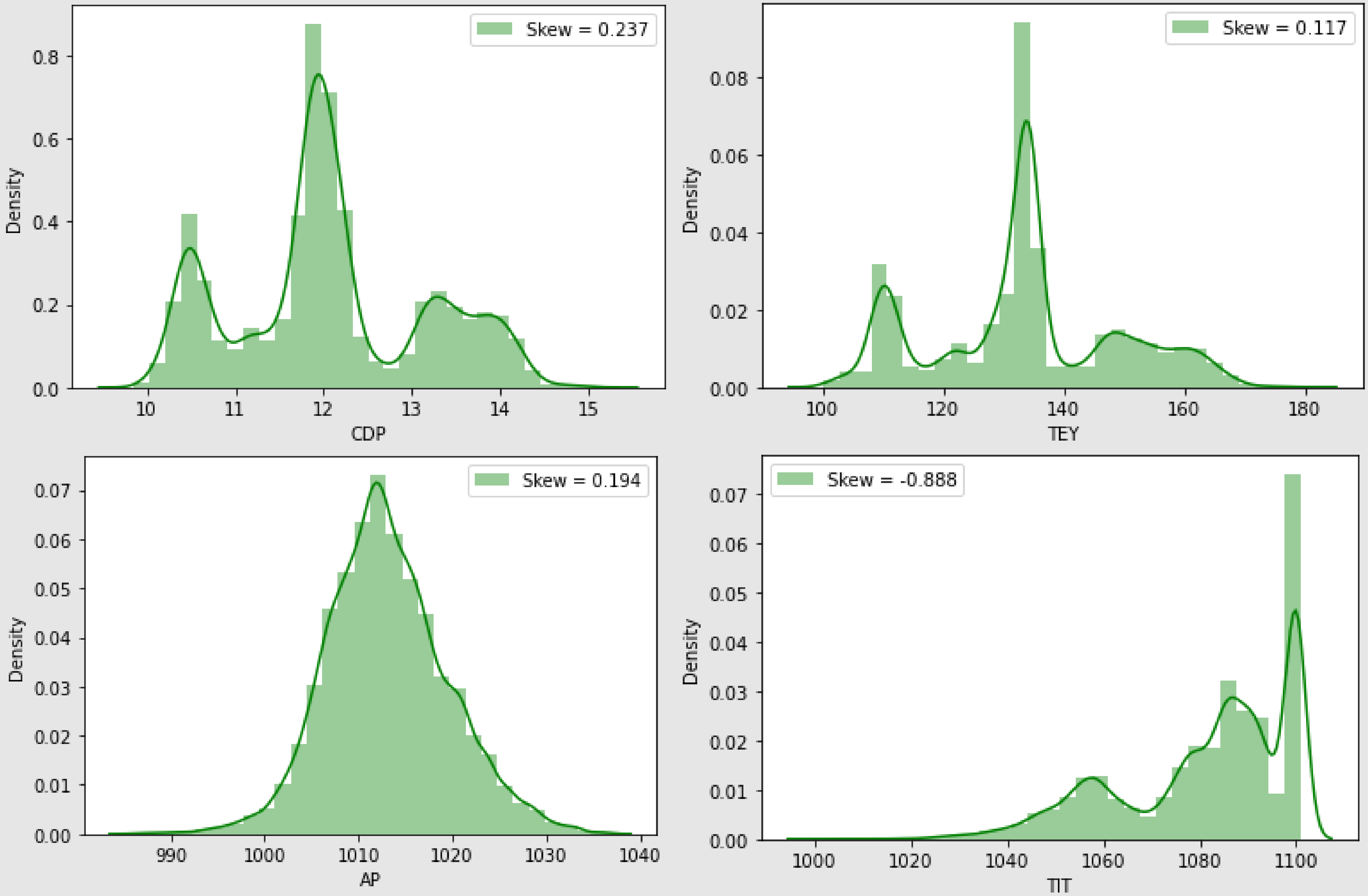
**Source :** Heysem Kaya, Pinar Tufekci and Erdinc Uzun. 'Predicting CO and NOx emissions from gas turbines: novel data and a benchmark PEMS', Turkish Journal of Electrical Engineering & Computer Sciences, vol. 27, 2019, pp. 4783-4796,

Source : www.giphy.com

# EXPLORATORY DATA ANALYSIS

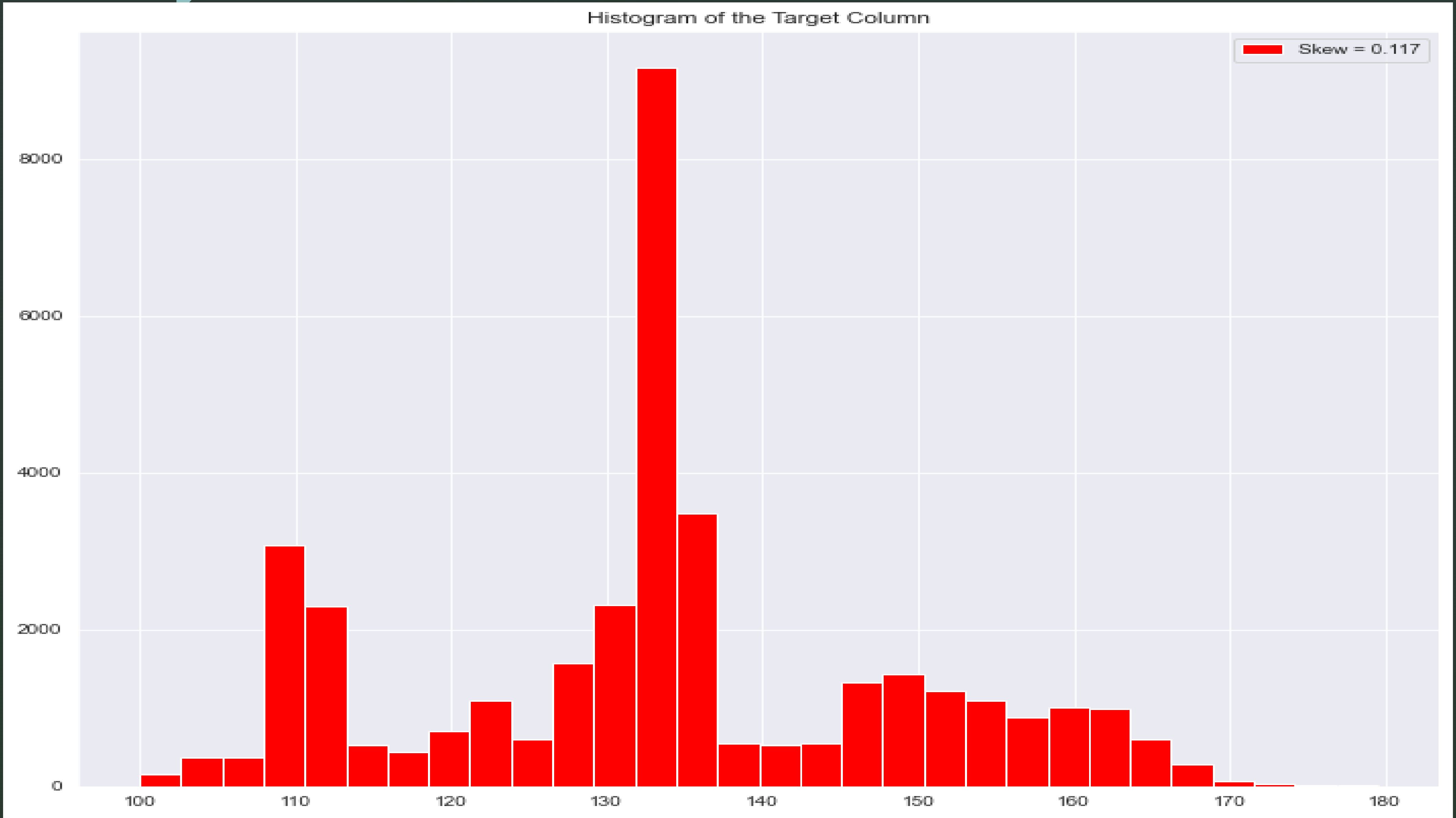
**UNIVARIATE ANALYSIS** :

- Some of the features are normally distributed
- The features AH, CO, TIT and TAT exhibit the highest skew coefficients
- Distribution of CO and TIT and TAT seem to contain many outliers
- Distplots are used to visualize the skewness of the variables

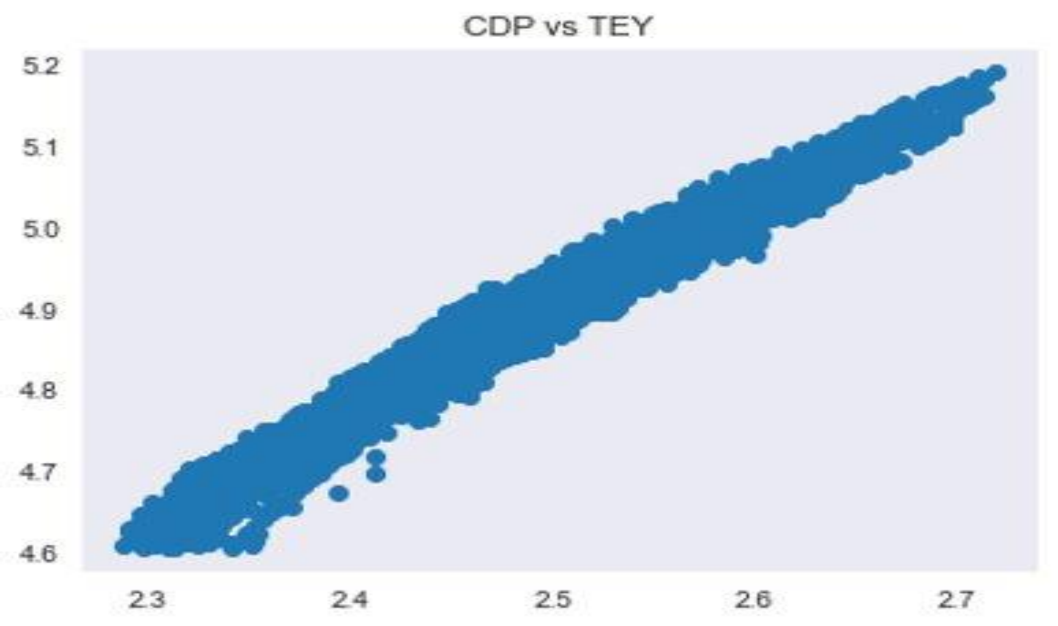


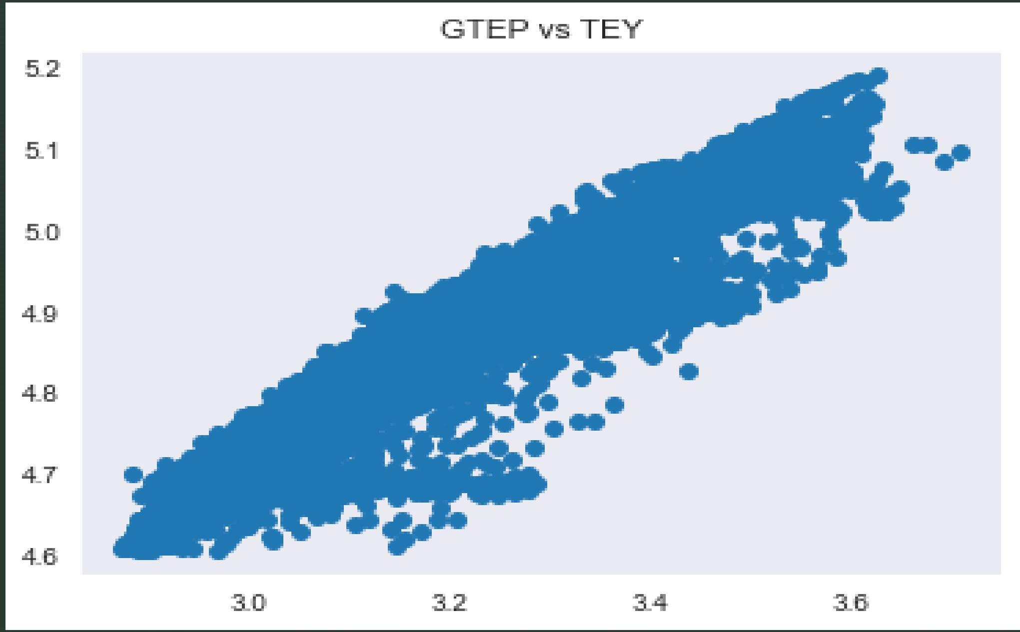
Skew = 0.237
Density
CDP
Skew = 0.117
Density
TEY
Skew = 0.194
Density
AP
Skew = -0.888
Density
TIT

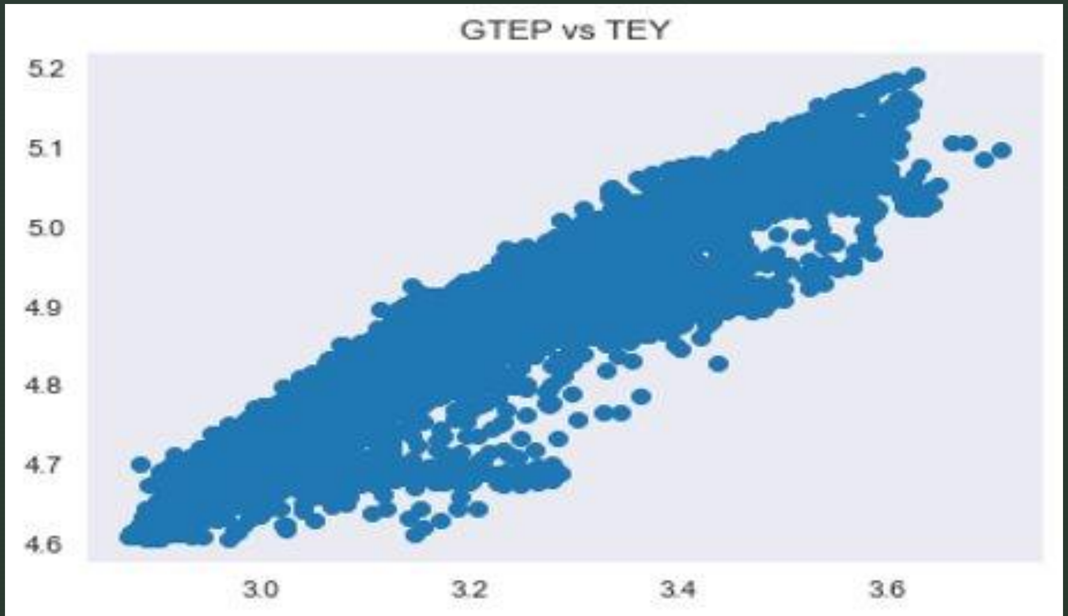
# BIVARIATE ANALYSIS :

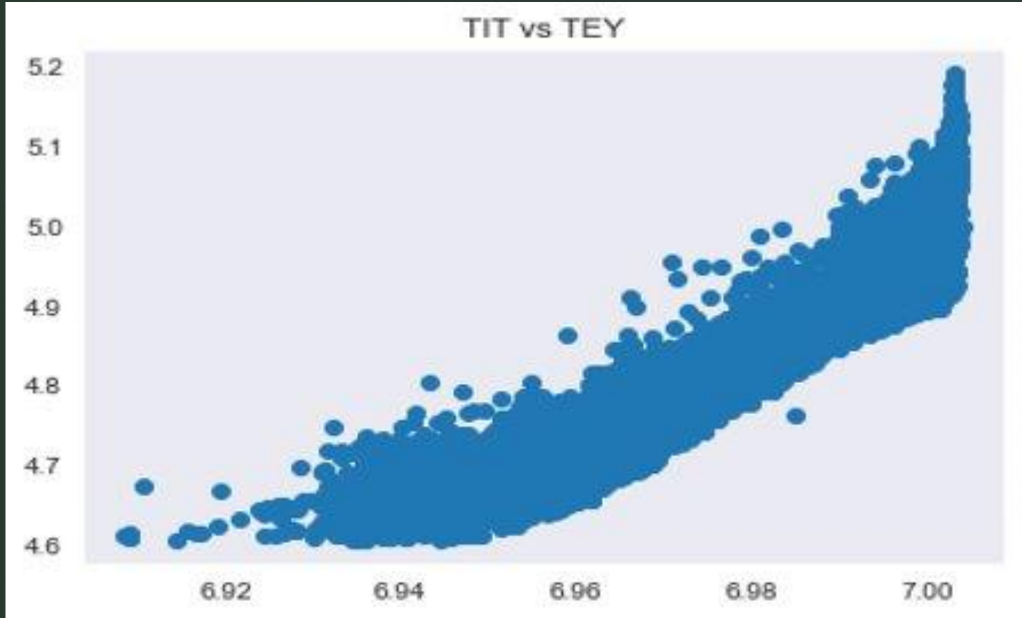


# SCATTER PLOT



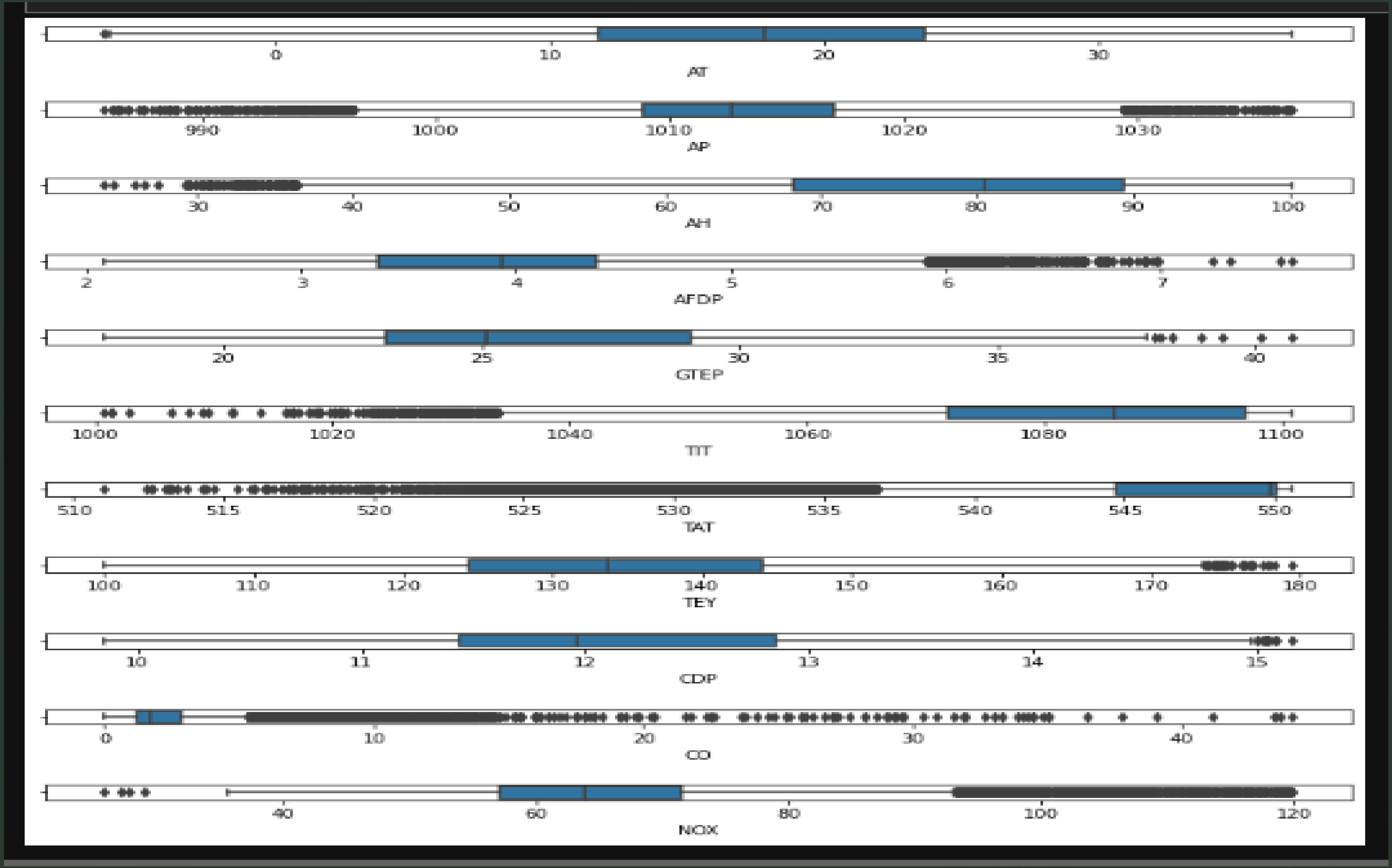



# MULTIVARIATE ANALYSIS :

```
TEY      1.000000
CDP      0.988778
GTEP     0.964127
TIT      0.910297
AFDP     0.665483
AP       0.118224
AT      -0.091152
NOX     -0.116127
AH      -0.137360
CO      -0.569813
TAT     -0.682396
```

# OUTLIERS



0 10 20 30
AT
990 1000 1010 1020 1030
AP
30 40 50 60 70 80 90 100
AH
2 3 4 5 6 7
AFDP
20 25 30 35 40
GTEP
1000 1020 1040 1060 1080 1100
TIT
510 515 520 525 530 535 540 545 550
TAT
100 110 120 130 140 150 160 170 180
TEY
10 11 12 13 14 15
CDP
0 10 20 30 40
CO
40 60 80 100 120
NOX

# FEATURE IMPORTANCE IN DATASET

- As per Univariate and Bivariate analysis *CDP, GTEP, TIT, TAT, AFDP, CO* these variable are very important to our prediction

- In these variable there are many outliers which is directly impact our performance measure

- Values of these features are highly correlated to our target columns, thus all values of features are required to get best accuracy

- So we do not handle the outliers

# STANDARDIZATION USING STANDARD SCALER

- For each feature, the Standard Scaler scales the values such that the mean is 0 and the standard deviation is 1(or the variance)
- x_scaled = x – mean / std_dev
- Standard Scaler assumes that the distribution of the variable is normal
- Thus, in case, the variables are not normally distributed, we either choose a different scaler or first, convert the variables to a normal distribution and then apply this scaler

# MACHINE LEARNING ALGORITHMS

1. SIMPLE REGRESSION
2. MULTIPLE REGRESSION
3. DECISION TREE REGRESSION
4. RANDOM FOREST REGREESION
5. SUPPORT VECTOR REGRESSION
6. K NEAREST NEIGHBOUR
7. BAGGING
8. PASTING
9. ADABOOST
10. GRADIENT BOOST
11. XGBOOST

# TRAIN TEST SPLIT

```
# spliting data the into training and testing
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(std,y,test_size=0.2,random_state=42
```

# SIMPLE REGRESSION

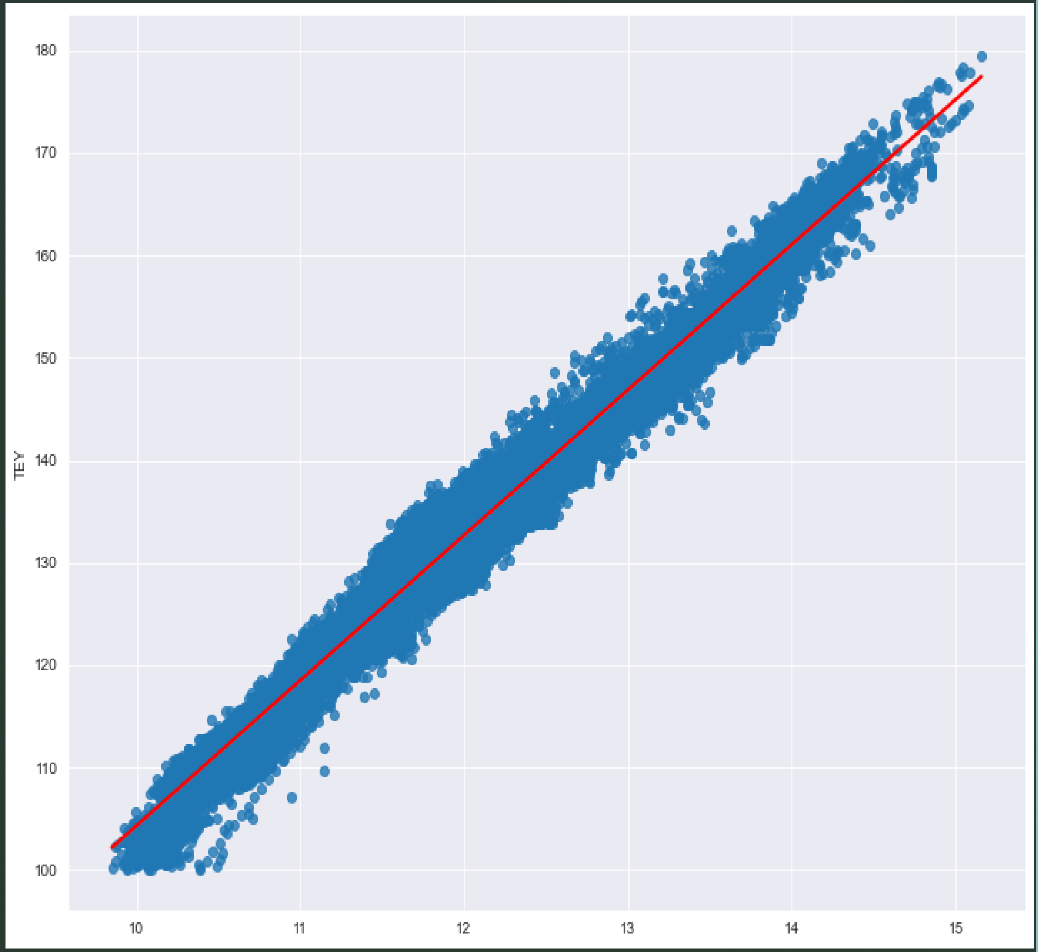
- This Simple Linear Regression Model is applied between input variable CDP (Compressor discharge pressure) & target variable TEY (Turbine Energy Yield)

PERFORMANCE

- Train Result: r2_score: 0.9778
- Test Result: r2_score: 0.9771

*After cross validation*

- Training r2_score: 0.9778
- Testing r2_score: 0.9771




Best Fit Line for CDP & TEY

# MULTIPLE REGRESSION

- In Multiple Regression we use more than on independent variables are used to predict the value of dependent variable.

- Independent Variables : 'CDP', 'GTEP','TIT', 'TAT', 'AFDP', 'CO', 'AT'.

- Dependent variable: ’TEY’

PERFORMANCE

- Train Result: r2_score: 0.995
- Test Result: r2_score: 0.995

*After cross validation*

- Training r2_score: 0.995
- Testing r2_score: 0.995

# DECISION TREE REGRESSION

- Decision Trees a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

- PARAMETER USED:

DTR = DecisionTreeRegressor(max_depth=2)

PERFORMANCE

- Train Result: r2_score: 0.9362
- Test Result: r2_score: 0.9317

*After cross validation*

- Training r2_score: 0.9357
- Testing r2_score: 0.9317

# RANDOM FOREST REGRESSION

- Random Forest contains a number of decision trees on various subsets and takes prediction from each tree and based on the majority votes of predictions it predicts the final output.

- PARAMETER USED:

  RFR = RandomForestRegressor(n_estimators=100)

## PERFORMANCE

- Train Result: r2_score: 0.999
- Test Result: r2_score: 0.998

*After cross validation*

- Training r2_score: 0.998
- Testing r2_score: 0.987

# SUPPORT VECTOR REGRESSION

- Support Vector Machine is a supervised learning algorithm which can be used for regression as well as classification problems.
- The main goal of SVR is to consider the maximum datapoints within the boundary lines and the hyperplane (best-fit line) must contain a maximum number of datapoints.

PERFORMANCE

- Train Result: r2_score: 0.995
- Test Result: r2_score: 0.995

*After cross validation*

- Training r2_score: 0.995
- Testing r2_score: 0.989

# K NEAREST NEIGHBOUR

- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

PERFORMANCE

- Train Result: r2_score: 0.986
- Test Result: r2_score: 0.981

*After cross validation*

- Training r2_score: 0.982
- Testing r2_score: 0.981

# BAGGING

- Bagging, also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset.

- PARAMETERS USED :

```
bag_reg_Bag = BaggingRegressor(DecisionTreeRegressor(),n_estimators=500,bootstrap=True,
                  random_state=42)
```

## PERFORMANCE

- Train Result: r2_score: 0.999
- Test Result: r2_score: 0.998

*After cross validation*

- Training r2_score: 0.998
- Testing r2_score: 0.997

# PASTING

- Pasting creates a dataset by sampling the training set without replacement.

- PARAMETERS USED :

```
bag_reg_past =
BaggingRegressor(DecisionTreeRegressor(),n_estimators=500,bootstrap=False,
                 random_state=42)
```

## PERFORMANCE

- Train Result: r2_score: 0.999
- Test Result: r2_score: 0.998

*After cross validation*

- Training r2_score: 0.998
- Testing r2_score: 0.997

# ADABOOST

- Adaboost is a very popular boosting technique that combines multiple "weak classifiers" into a single "strong classifier".
- PARAMETERS USED :

  Adaboost = AdaBoostRegressor(random_state=42)

## PERFORMANCE

- Train Result: r2_score: 0.984
- Test Result: r2_score: 0.983

*After cross validation*

- Training r2_score: 0.985
- Testing r2_score: 0.983

# GRADIENT BOOSTING

- Gradient Boost helps us to get a predictive model in form of an ensemble of weak prediction models such as decision trees. Whenever a decision tree performs as a weak learner then the resulting algorithm is called gradient-boosted trees.

- PARAMETERS USED:

grad_reg = GradientBoostingRegressor(random_state=40,learning_rate=0.1)

## PERFORMANCE

- Train Result: r2_score: 0.996
- Test Result: r2_score: 0.996

*After cross validation*

- Training r2_score: 0.996
- Testing r2_score: 0.996

# XGBOOST

- XGBOOST is the latest version of gradient boosting which also works very similar to Gradient Boost.
- PARAMETERS USED:

xgb_reg = XGBRegressor(random_state=42,learning_rate=0.1)

## PERFORMANCE

- Train Result: r2_score: 0.998
- Test Result: r2_score: 0.997

*After cross validation*

- Training r2_score: 0.997
- Testing r2_score: 0.997

# REPORT

| Regression | Simple Linear | Multiple | Decision tree | Random Forest | SVM | Bagging | Adaboosta | Gradient Boosting | xgboost Regressor |
|---|---|---|---|---|---|---|---|---|---|
| Training r2_score | 0.9778 | 0.9958 | 0.9357 | 0.9982 | 0.9982 | 0.9964 | 0.9851 | 0.9964 | 0.9977 |
| Test r2_score | 0.9771 | 0.9957 | 0.9317 | 0.9973 | 0.9973 | 0.9949 | 0.9835 | 0.9962 | 0.9971 |

# PREDICTING TARGET USING ORIGINAL & NEW INPUTS

| | Actual value | Predicted value |
|---|---|---|
| **6637** | 154.88 | 154.3841 |
| **4009** | 132.76 | 133.1307 |
| **2951** | 108.59 | 109.0108 |
| **263** | 160.10 | 159.4680 |
| **5568** | 131.03 | 130.6179 |
| **518** | 134.46 | 133.7973 |
| **2320** | 134.68 | 134.2702 |
| **4899** | 129.08 | 129.1317 |
| **315** | 164.34 | 164.6985 |
| **1582** | 148.72 | 151.9823 |

```
# generating predictions for new Data
l=[(11,20,1111,560,3.5,12,4)]
i=np.array(l)
y_pred = RFR.predict(i)
# creating table with test & predicted for test
print('predictions for new Data :',y_pred)
```

```
predictions for new Data : [161.7762]
```

# CONCLUSION

- There is a relationship between the process, ambient variables 'CDP', 'GTEP','TIT', 'TAT', 'AFDP', 'CO', 'AT' and Turbine Energy Yield (TEY) also TEY can be predicted using these variables

- Analyzed the dataset containing ambient, process, and emission variables from the gas turbine and discovered relationships existing between **Turbine Yield Energy (TEY)** and the other variables

THANK YOU!