**Assignment: Linear regression analysis of combined cycle power plant data**

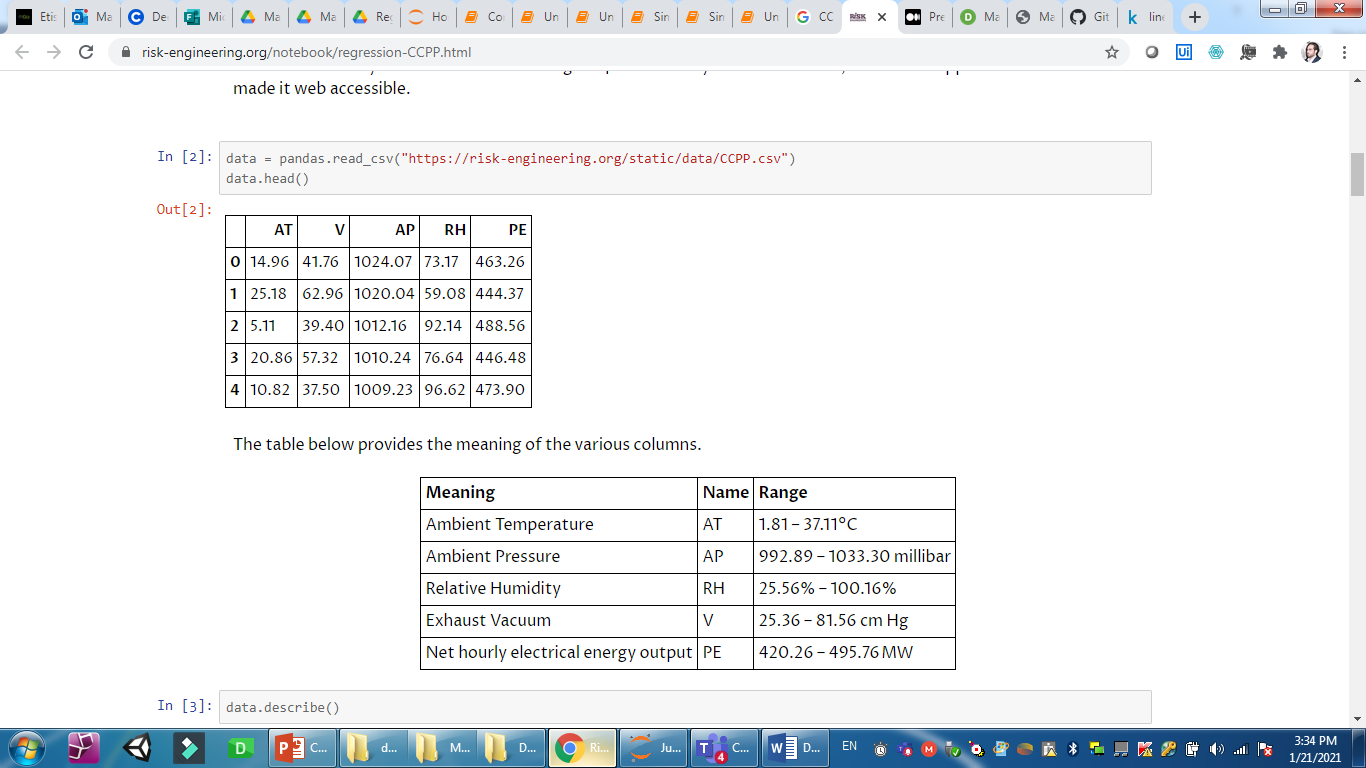
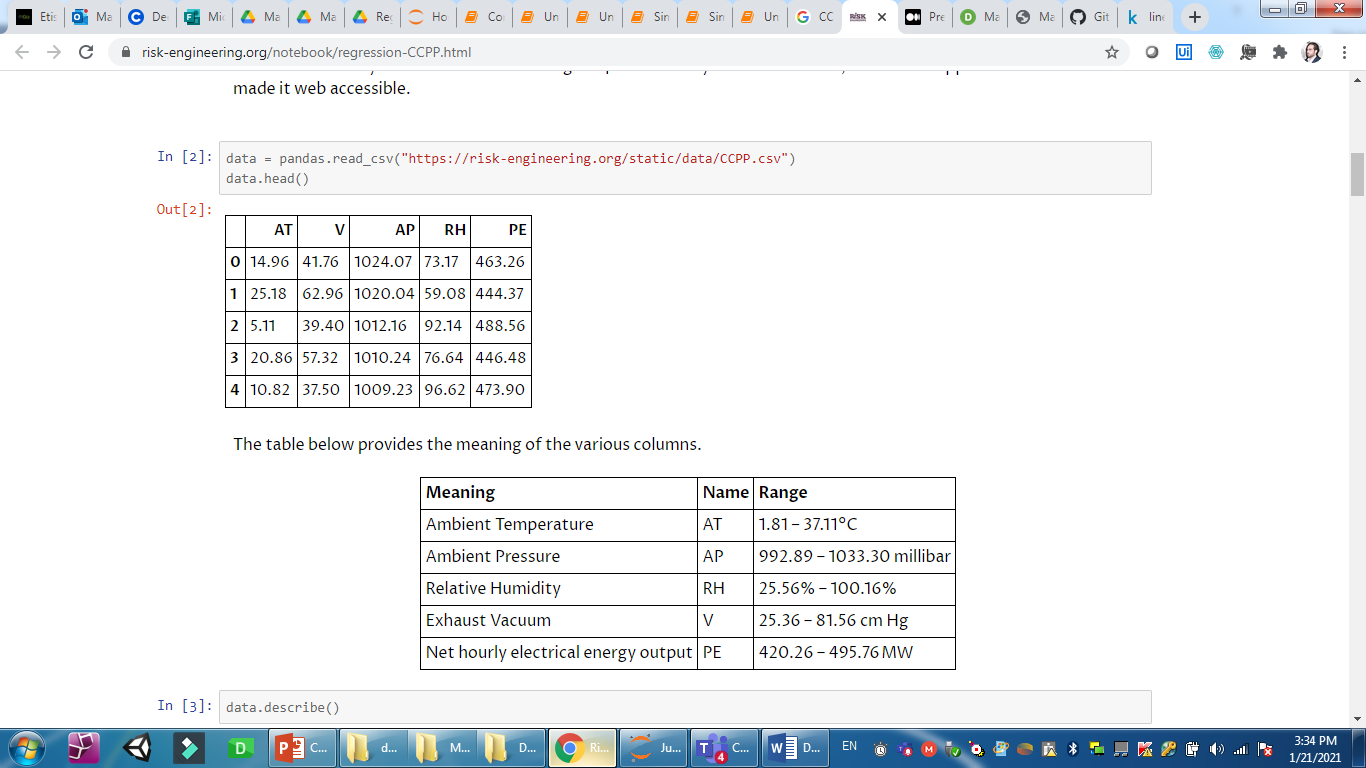
**Task:** Analyze data from a combined cycle power plant (CCPP) to attempt to build a predictive model for output power.

**Subject:** Attached dataset contains 9568 data points collected from a combined cycle power plant over 6 years. A combined cycle power plant is composed of gas turbines, steam turbines and heat recovery steam generators. Electricity is generated by gas & steam turbines. Three ambient variables affect the performance of the gas turbine, and exhaust vacuum affects the performance of the steam turbine.

Data consists of hourly averages taken from various sensors located around the plant that record the ambient variables every second.

# Objective: To develop a predictive model to predict the power output.

**Below are the table columns:** **Below are the meaning of the columns:**

**Tip:** Since the goal is to predict the output power based on some parameters, this is a regression problem. Regression aims to establish a relationship between predictors (variables that help us make a prediction) and the target (the value we want to predict).

# Workflow to solve the tasks:

1. Data Cleaning/mining and transformation
2. Feature Extraction and visualization
3. Develop the model
4. Evaluate the model

# Data Cleaning/Data Mining

* Import python libraries.
* Load the attached dataset.
* Explore the data and check number of (rows & columns).
* Check if there is any missing values and handle it.
* Check the datatype of each column
* Check if there is any duplicate values and remove it using (df.drop\_duplicates(inplace=True))

**Feature Extraction and visualization**

* Check the distribution of the variables using (df.describe) and check the standard deviation
* Determine which features will help to predict the output power using Correlation coefficient (r), Use the following code to check the correlation coefficient (r).

import seaborn as sns  
plt.figure(figsize = (7, 5))  
sns.heatmap(df.corr(), annot = True)

**Tip:** Correlation is measured on a scale of (-1 to 1). -1 means complete negative correlation and 1 means complete positive correlation. 0 means no correlation at all.

**From the above which variable have strong or weak correlation with the target variable?**

* Strong Correlation variable: V& AT
* Weak Correlation variable

**Tip:** We can visualize the distribution of the dataset (which shows how each feature correlates to each other and the PE) using the following code:

import seaborn as sns

sns.set(style=”ticks”)  
sns.pairplot(df, diag\_kind = ‘hist’)

**Caution:** You will notice some independent variables(x) are highly correlated with each other. This is usually not a good thing as our features should be independent of each other. This problem is called **multicollinearity,** in this case we need to remove one of these variables which have less correlation with the target variable.

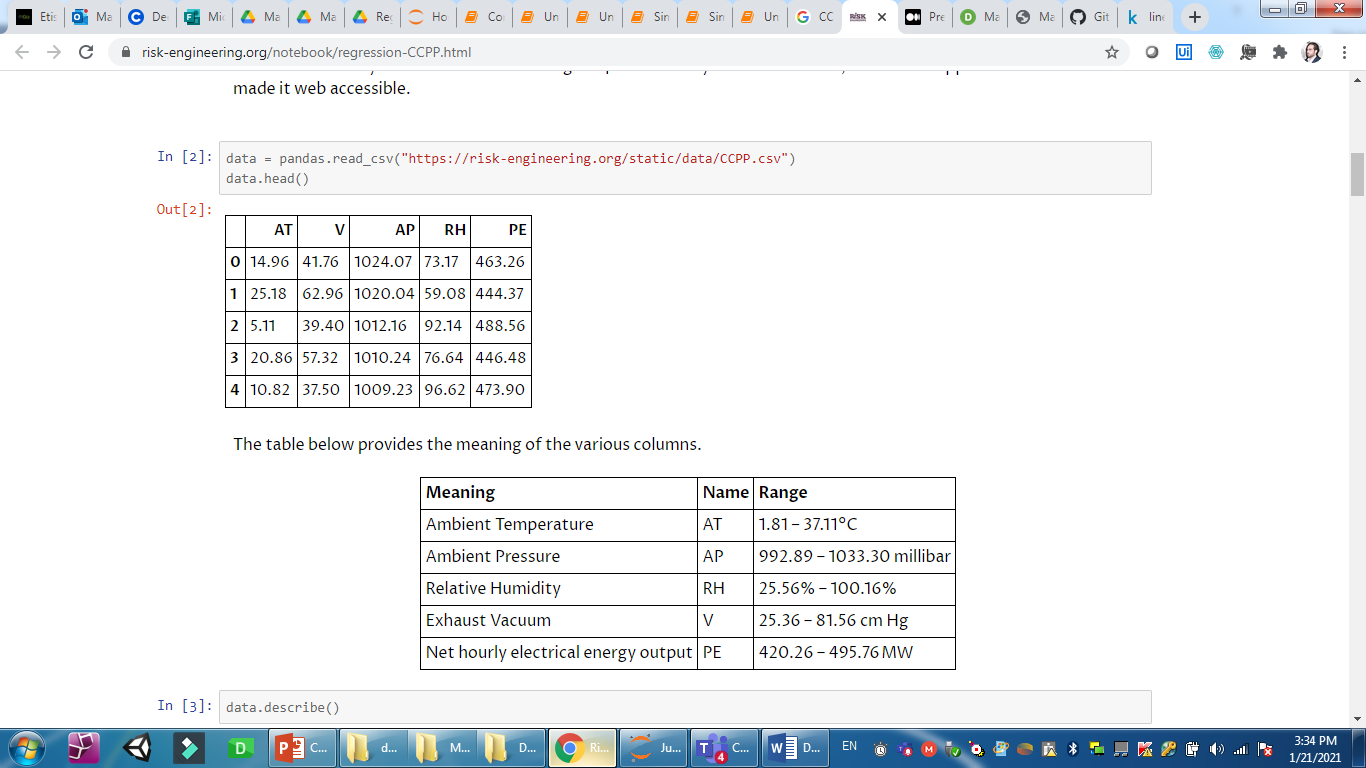
Which variables are correlated with each other’s? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

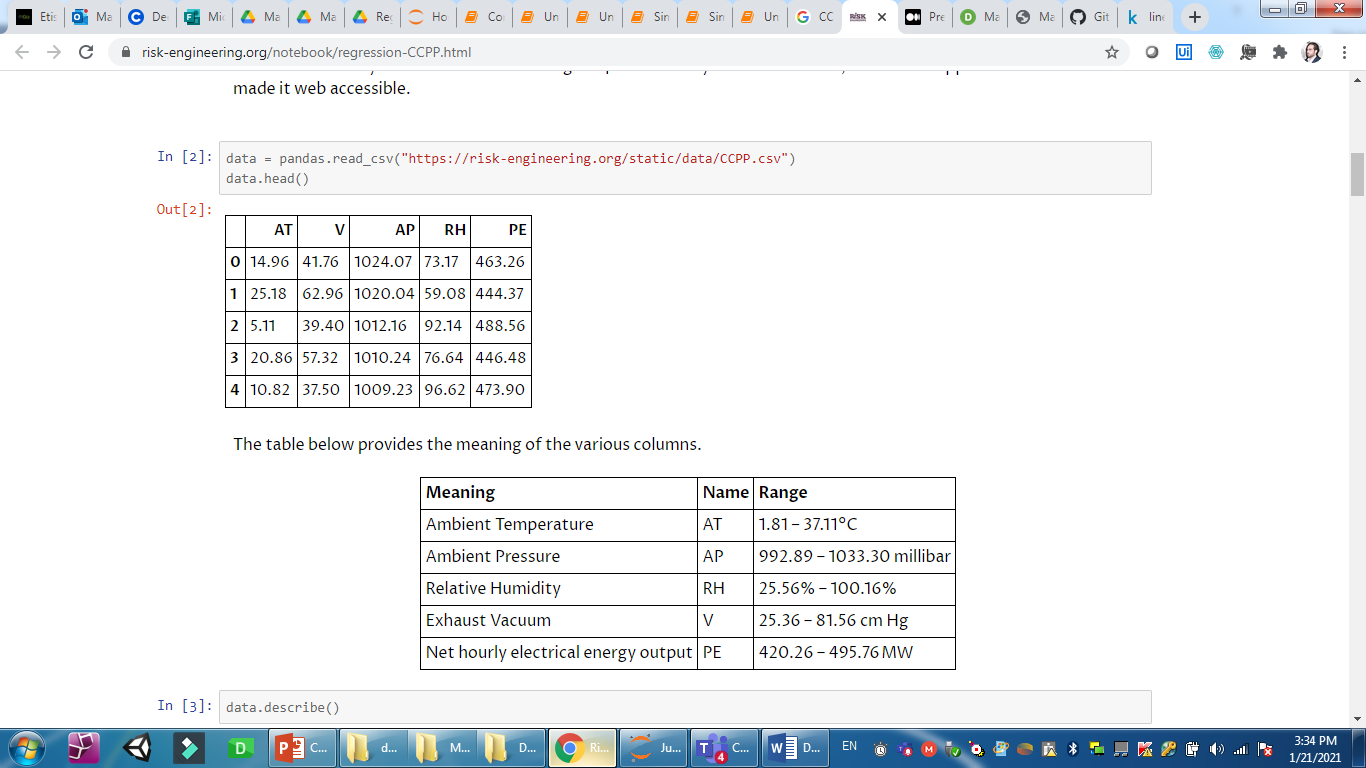
**Develop the model**

* Divide our data into (x) and (y) using .iloc[row,column]
* Split the data set into training and testing using (train\_test\_split) with 30 as a ratio
* Create Multiple linear regression model using fit() function

**Evaluate the model**

* Calculate the score (R-Squared) using .score() function
* Check and get the slop and the y-intercept for the model (y=b+m1x1+m2x2+m3x3+m4x4)
* Create the table to show the predicted value (y-pred) and (y-actual) as the followings :





**Y-Pred**

**Y-actual**

**Try**: To predict the power for the followings independent data input row ([22.1,71.29,1008.2,75.38]) , what is the predicted power?

**Finally, Save the code as a HTML file and submitted it with mail subject as (Assignment 4 – Linear Regression Analysis)**

**Good Luck ☺ !!**