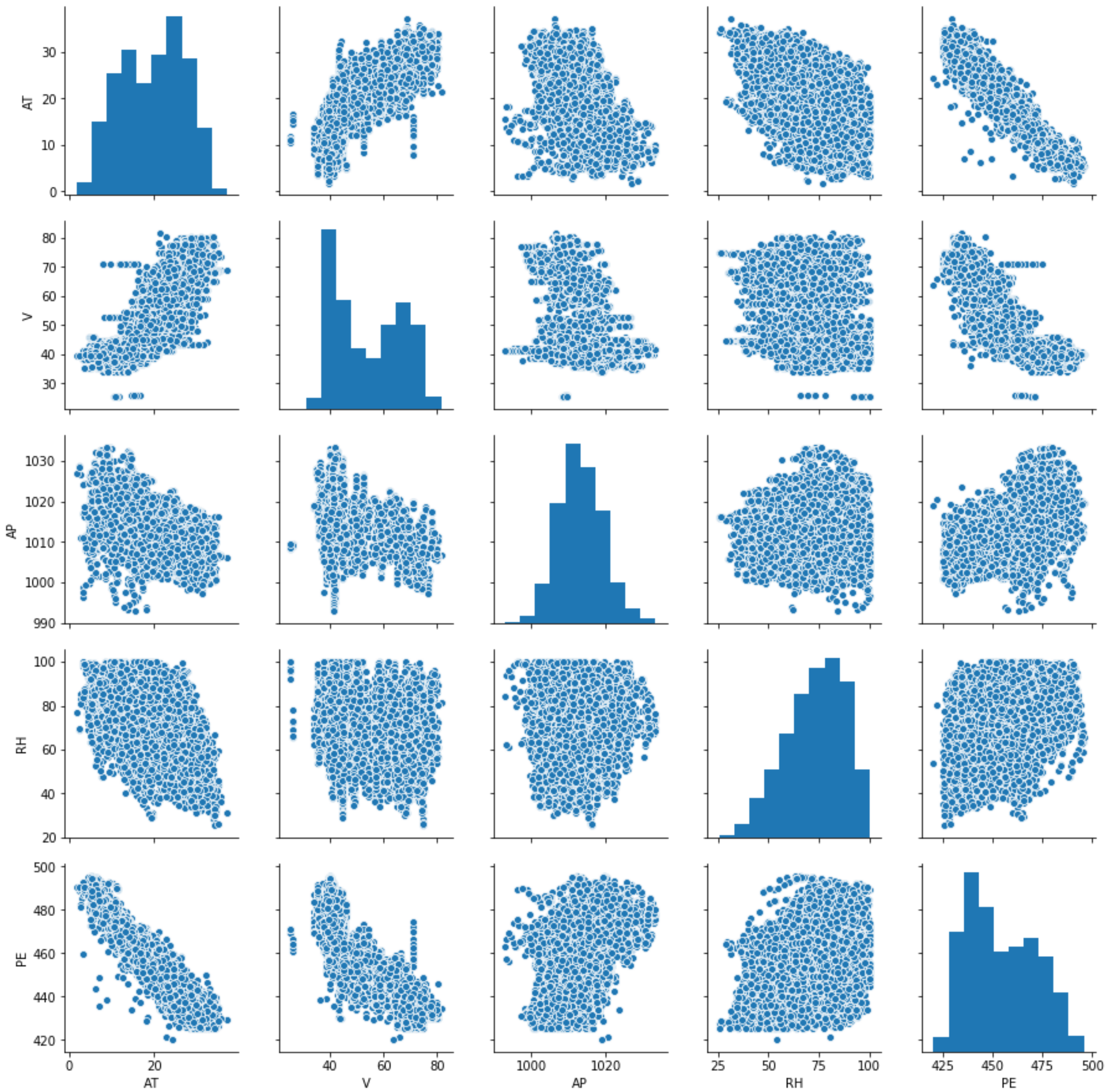
Rubina Iman Kabir
INF 551
Homework 2
678-560-5414

1. (b) (i) Number of rows: 9568
Each row represents the feature values collected for an observation.
Number of columns: 5
The first four column represents values for the Features: hourly average ambient variables Temperature (AT), Ambient Pressure (AP), Relative Humidity (RH) and Exhaust Vacuum (V) to predict the net hourly electrical energy output (EP) of the plant.

(ii)

Figure 1: Pairwise Scatter Plots



(Rough) Distributions of individual predictors (diagonal):

- AT – Normal
- V – None
- AT – Normal
- RH – Skewed right
- PE – Skewed left

Table 1: Correlation Matrix Table

	AT	V	AP	RH	PE
AT	1.000000	0.844107	-0.507549	-0.542535	-0.948128
V	0.844107	1.000000	-0.413502	-0.312187	-0.869780
AP	-0.507549	-0.413502	1.000000	0.099574	0.518429
RH	-0.542535	-0.312187	0.099574	1.000000	0.389794
PE	-0.948128	-0.869780	0.518429	0.389794	1.000000

- Linear: AT-V, **AT-PE**, **V-PE**,
- Moderate: AT-AP, AT-RH, AP-V, **AP-PE**
- Weak: V-RH, **RH-PE**
- None: AP-RH

Since we are implementing a regression analysis on these coefficients, predictors that lie in the linear category may cause problems.

(iii)

Table 2: Variable Statistics Table

	AT	V	AP	RH	PE
MEAN	19.651231	54.305804	1013.259078	73.308978	454.365009
Q1	13.51	41.740	1009.1	63.3275	439.75
Q2(MEDIAN)	20.345	52.08	1012.94	74.975	451.55
Q3	25.720000	66.540000	1017.260000	84.830000	468.430000
RANGE	35.3	56.2	40.40999	74.6	75.5
IQR	12.2099	24.8	8.15999	21.502499	28.68

(c)

Simple Linear Regression

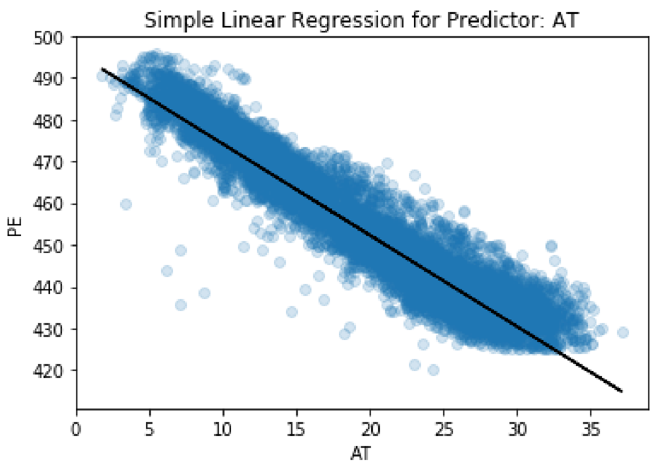


Figure 2: SLR for AT.

Achieves the best R^2 error, implying the approximately 0.9 of the variation in PE can be explained by the relationship to AT.

Also achieves the lowest MSE.

The relationship between AT-PE is nearly linear, so it is not surprising this model performs the best – also this model has the highest correlation value.

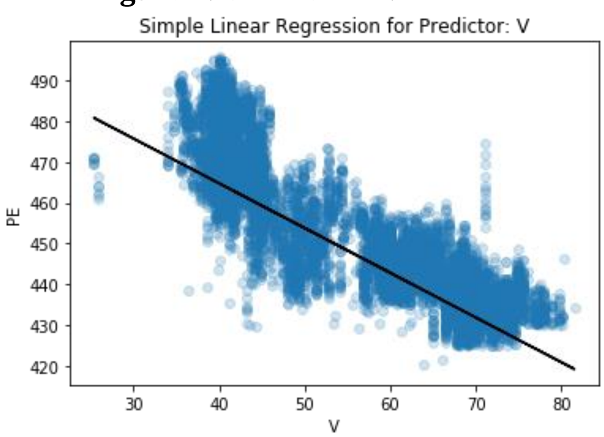


Figure 3: SLR for V.

Ranked second in achieving a high R^2 value, at approximately 0.67

Similarly with it's MSE, it ranks second.

Again, the correlation between V-PE is 0.884, so not surprising that this model is second in performance.

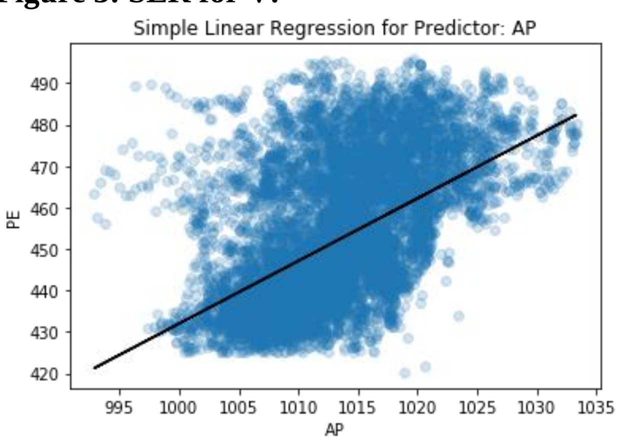


Figure 4: SLR for AP.

Achieves an R^2 of approximately 0.27

This model could be improved by removing values with high marginal error to increase it's correlation, and in turn decreasing it's MSE.

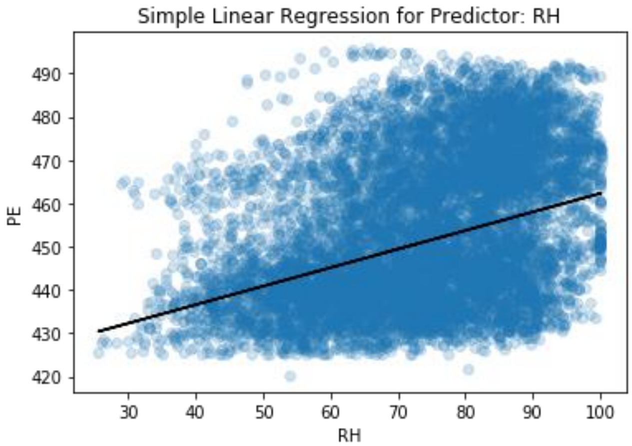


Figure 5: SLR for RH.

Achieves an R^2 value of approximately 0.13; the lowest amongst all predictors.

The model could be improved by removing outliers with high marginal error to increase the correlation between RH-PE, along with it's R^2 value. RH-PE exhibits a weak correlation, so it is not surprising that this model performs the worst with an MSE of 259.61.

Table 3: Summary of each SLR

NOTE: Each row is it's a single model.

	beta0	beta1	R^2	MSE	t-statistic (beta1)	p-value(beta1)
AT	496.055	-2.18753	0.895949	31.1295	-287.002	<.00001
V	508.615	-1.09571	0.667873	99.3645	-138.695	<.00001
AP	-1078.98	1.51098	0.269893	218.431	59.4659	<.00001
RH	419.455	0.428073	0.132252	259.609	38.183	<.00001

For each SLR model, the coefficient on the predictor are all statistically significant. As we move down the rows of the table, the MSE values increase exponentially. Though, this is not surprising based off the correlation values between the individual predictors and the response variable shown in Table 1.

(d)

Table 4: Multiple Linear Regression

	ESTIMATED COEFFICIENT	SE	T-STATISTIC	P-VALUES(ALPHA=0.05)
AT	-1.977513	0.0062255	-317.647	<.00001
V	-0.233916	0.00361171	-64.7661	<.00001
AP	0.062083	0.00783475	7.92405	<.00001
RH	-0.158054	0.00317087	-49.8457	<.00001
INTERCEPT	454.609274	-	-	-

We <u>reject</u> the null hypothesis for all predictors; all estimated coefficients are statistically significant. With this model, we achieve an MSE of 20.767– lower than the best performing simple linear regression model.

(e)

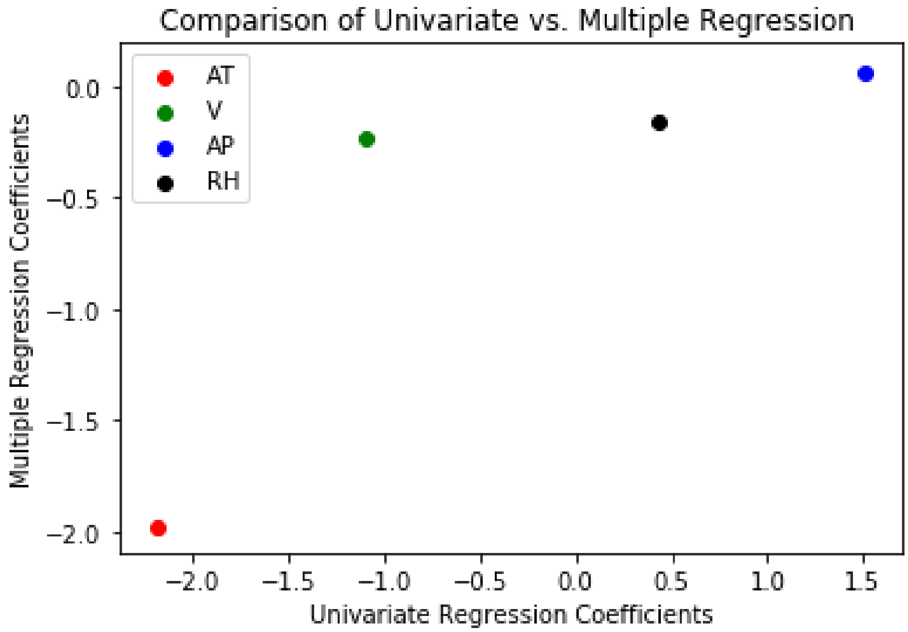


Figure 6: Univariate vs multiple regression coefficients.

AT: The coefficient does not change a lot – about 0.2 decrease.
V: The coefficient decreases by approximately 0.9.
AP: The coefficient decreases by approximately 1.4
RH: The coefficient decreases by approximately 0.5

The large changes in each of the predictor coefficients is related to the issue of multicollinearity since each variable changes approximately 10% or more!

(f)

<u>Table 5: Nonlinear Associations</u>
Note: Each row is an single model.

	X (P-VALUE)	X^2(P-VALUE)	X^3(P-VALUE)	MSE
AT	7.89815e-07	8.83305e-73	3.65218e-110	25.6643
V	2.52659e-05	0.768497	0.0137349	65.5253
AP	4.50274e-17	3.6667e-17	8.26415e-18	211.197
RH	0.000377251	9.39543e-06	1.44028e-05	246.474

AT: All coefficients are significant and this model achieves the lowest MSE value 25.67;
Compared to the simple linear regression on AT, the MSE decreases by 17%.
V: The coefficient of V^2 is statistically insignificant.
Compared to the simple linear regression on V, the MSE decreases by 34%.

AP: All coefficients are significant. Compared to the simple linear regression on AT, the MSE decreases by only 3%.
RH: All coefficients are significant. Compared to the simple linear regression on RH, the MSE decreases by 5%.

Overall, performance increased compared to the simple linear and nonlinear regression models. Amongst all predictors, there is a stronger nonlinear association with the predictor V. Analyzing Figure 3, we can confirm that increasing the degree on the polynomial will result in a function that fits the dataset better than a linear function.

(g)

Table 6: Multiple linear regression with interaction terms

	Coefficient	SE	p-values
intercept	685.782468	78.640060	3.231607e-18
AT	-4.347014	2.373139	6.701873e-02
V	-7.674858	1.350761	1.371251e-08
AT	-0.152355	0.076817	4.735732e-02
RH	1.570907	0.773350	4.225213e-02
AT:V	0.020971	0.000899	3.333358e-117
AT:AP	0.001759	0.002339	4.520509e-01
AT:RH	-0.005230	0.000812	1.216944e-10
V:AP	0.006812	0.001327	2.877026e-07
V:RH	0.000839	0.000489	8.619366e-02
AP:RH	-0.001612	0.000758	3.360557e-02

All coefficients are significant, except AT,V:RH, and AT:AP (when $\alpha = 0.05$.) This model achieves an MSE value of 18.551 – the lowest of all.

(h)

Initial model:

All predictors along with interaction terms and quadratic nonlinearities.

Model found using Backward Selection:

The following features were removed based off their p-values, in order: AP:RH(0.337), V:AP (0.579), AT:RH (0.302), V^2(0.443).

The resulting model is:

$$\hat{y} = \beta_0 AT + \beta_1 V + \beta_2 AP + \beta_3 RH + \beta_4 AT^2 + \beta_5 AT * V + \beta_6 AT * AP + \beta_7 V * RH + \beta_8 AP^2 + \beta_9 RH^2$$

Table7: Multiple linear regression models with test and train datasets.

	TRAIN MSE	TEST MSE
MODEL WITH ALL PREDICTORS	25.731	26.939
MODEL FROM BACKWARD SELECTION	18.435	19.314

The MSE decreases by using backward selection on both the train and test datasets, hence this model would be preferred. The MSE obtain from backward selection is the best MSE obtained in this exercise.

(i)

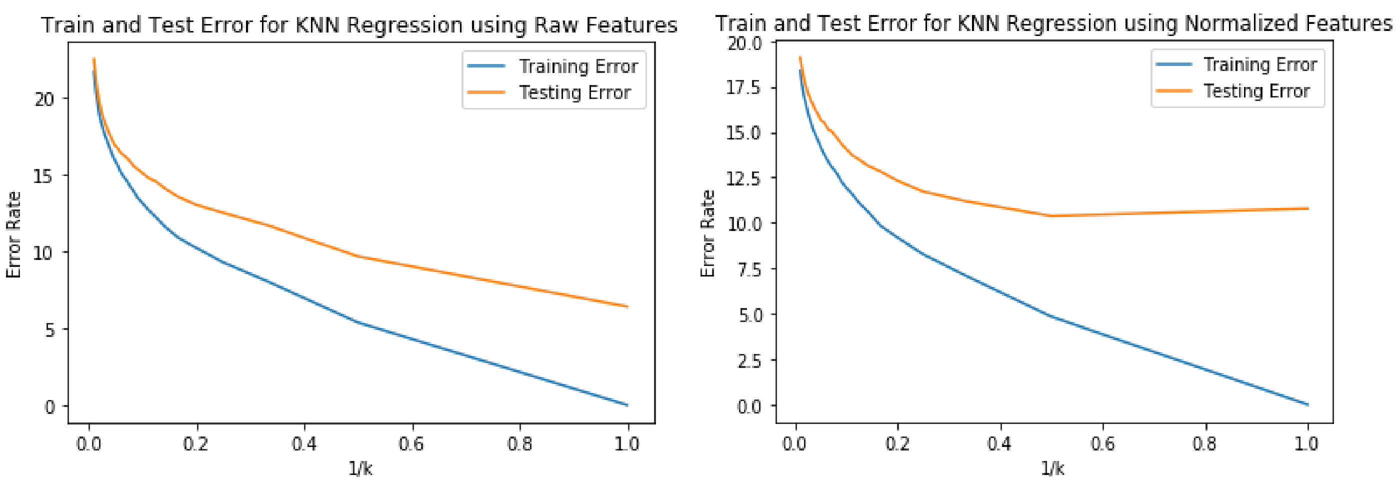


Figure 7: KNN Regression train(70% of data) and test(30% of data) MSE using raw and normalized features.

Based off the figures, for both cases high values of k produce large MSE. Though, using raw features we get lower test MSE as k-decreases compared to using normalized features

Table 8: Minimum Test MSE using KNN Regression

	K	MIN TEST MSE
RAW FEATURES	1 (OR 2)	6.386 (9.639)
NORMALIZED FEATURES	2	10.370

(j)

The smallest MSE achieved from implementing linear regression was found using the interaction terms in part (g), where MSE= 18.551. The KNN regression MSE, for both raw and normalized, is nearly half of the test MSE obtained from linear regression. Since the size of our dataset is very large and the number of predictors is small, KNN performs better than linear regression because KNN is a non-parametric method that does not make any assumptions on our function. Using a flexible non-parametric model, like KNN will provide us more "flexibility" on fitting the data appropriately when training the algorithm versus linear regression which is an inflexible parametric method that has structural assumptions about the data, i.e. linearity.

2. (a) The performance of a flexible model will be **better** than an inflexible model since the flexible model will not assume a type of structure on the relationship between the predictor and the target. Hence, we will have more flexibility to find a function that fits the data well while achieving a low bias.

   (b) The performance of a flexible model will be **worse** than an inflexible model because the flexible model will begin to overfit the data since the number of samples is small – resulting in a large variance.

   (c) The performance of a flexible model will be **better** than an inflexible model since inflexible methods have lower degrees of freedoms and in this cause the relationship between X and Y is highly nonlinear.

   (d) The performance of a flexible model will be **worse** than a inflexible model since increasing the flexibility of the model will start to overfit the data since the variance of the noise is extremely high.

3. (a)

OBSERVATION	DISTANCE
1	3
2	2
3	$\sqrt{10}$
4	$\sqrt{5}$
5	$\sqrt{2}$
6	$\sqrt{3}$

   (b) K=1
   Closest observation is 5: (-1,0,1,Green)
   Thus, Prediction = **Green**

   (c) K= 3
   Closest observations: 2, 5, 6
   P(Y=Green | X=x) = $\frac{1}{3} * 1 = \frac{1}{3}$ P(Y=Red | X=x) = $\frac{1}{3} * 2 = \frac{2}{3}$
   Thus, prediction = **Red**

   (d) We expect **k to be smaller** because as we increase k our model becomes less flexible and produces a decision boundary close to linear.