INF552-HW2
Feb.1st
Jialin Shi
UID # 7948542502

Exploring the data:

(b) - i. How many rows are in this data set? How many columns? What do the rows and columns represent?

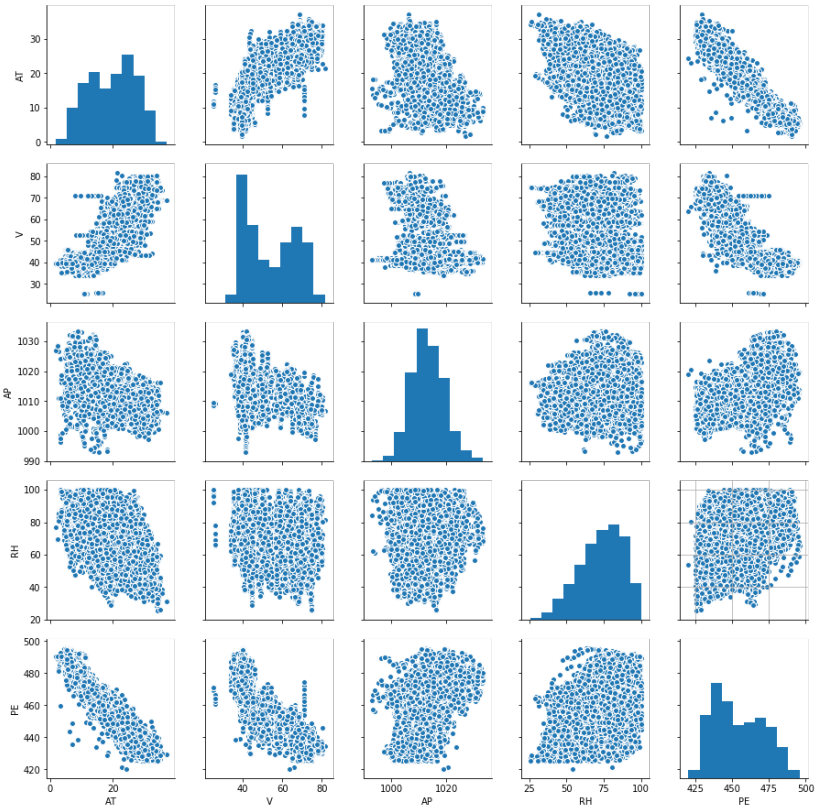
9568 rows and 5 columns.

Each column is a variable.

Each row is an instance(observation).

```
Rows: 9568 entries, 0 to 9567
Data columns (total 5 columns):
  AT    9568 non-null float64
  V     9568 non-null float64
  AP    9568 non-null float64
  RH    9568 non-null float64
  PE    9568 non-null float64
```

(b) - ii. Make pairwise scatterplots of all the varianbles in the data set including the predictors (independent variables) with the dependent variable. Describe your findings.



AT	-0.948128
V	-0.869780
AP	0.518429
RH	0.389794
PE	1.000000

Variable AT and Variable V seem to have strong linear relationship with Independent Variable PE.

RH has weak linear relation with Variable PE.

(b) - iii. What are the mean, the median, range, first and third quartiles, and interquartile ranges of each of the variables in the dataset? Summarize them in a table.

	AT	V	AP	RH	PE
mean	19.6512	54.3058	1013.26	73.309	454.365
median	20.345	52.08	1012.94	74.975	451.55
range	(1.81, 37.11)	(25.36, 81.56)	(992.89, 1033.3)	(25.56, 100.16)	(420.26, 495.76)
25%	13.51	41.74	1009.1	63.3275	439.75
75%	25.72	66.54	1017.26	84.83	468.43
iqr	12.21	24.8	8.16	21.5025	28.68

 # (c) For each predictor, fit a simple linear regression model to predict the response. Describe your results.

AT - PE: R-Squared =0.899, p-value=0.000, statistically significant
V - PE: R-Squared =0.757, p-value=0.000, statistically significant
AP - PE: R-Squared =0.269, p-value=0.000, statistically significant
RH - PE: R-Squared =0.152, p-value=0.000, statistically significant

OLS Regression Results

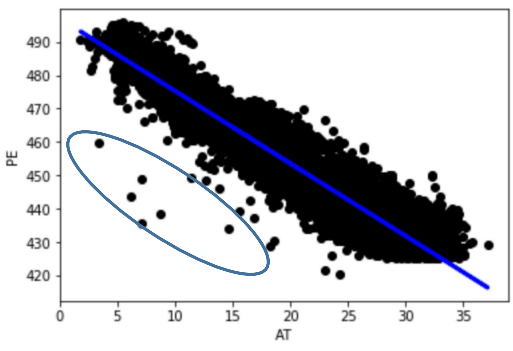
Dep. Variable:	y	R-squared:	0.899
Model:	OLS	Adj. R-squared:	0.899
Method:	Least Squares	F-statistic:	8.510e+04
Date:	Fri, 01 Feb 2019	Prob (F-statistic):	0.00
Time:	21:29:13	Log-Likelihood:	-29756.
No. Observations:	9568	AIC:	5.952e+04
Df Residuals:	9566	BIC:	5.953e+04
Df Model:	1		
Covariance Type:	nonrobust		

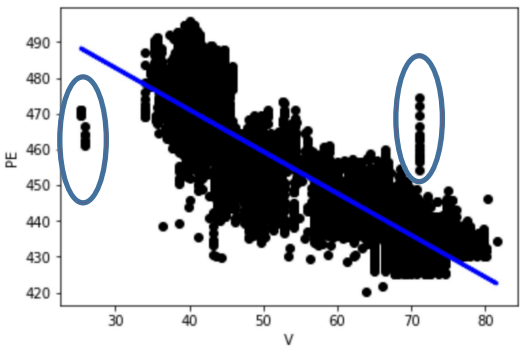
	coef	std err	t	P>|t|	[0.025	0.975]
const	497.0341	0.156	3177.280	0.000	496.727	497.341
AT	-2.1713	0.007	-291.715	0.000	-2.186	-2.157

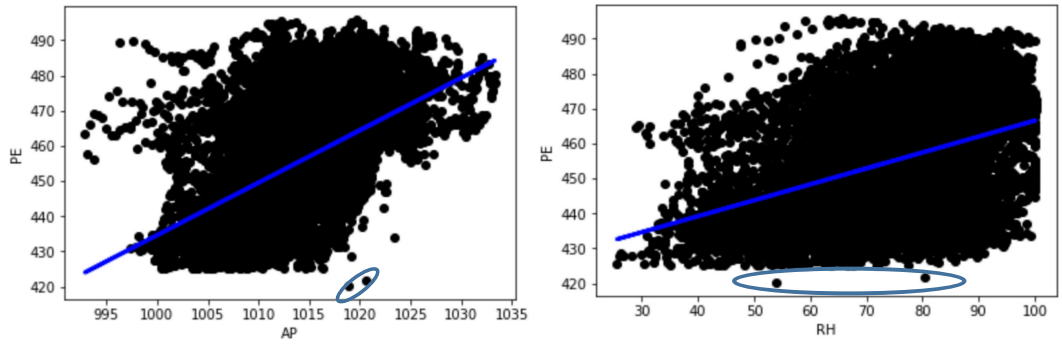
```
==============================================================================
Dep. Variable:                      y   R-squared:                       0.757
Model:                            OLS   Adj. R-squared:                  0.756
Method:                 Least Squares   F-statistic:                 2.972e+04
Date:                Fri, 01 Feb 2019   Prob (F-statistic):               0.00
Time:                        21:29:13   Log-Likelihood:                -33963.
No. Observations:                9568   AIC:                         6.793e+04
Df Residuals:                    9566   BIC:                         6.794e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        517.8015      0.378   1370.218      0.000     517.061     518.542
V             -1.1681      0.007   -172.402      0.000      -1.181      -1.155
==============================================================================
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.269
Model:                            OLS   Adj. R-squared:                  0.269
Method:                 Least Squares   F-statistic:                     3516.
Date:                Fri, 01 Feb 2019   Prob (F-statistic):               0.00
Time:                        21:29:13   Log-Likelihood:                -39224.
No. Observations:                9568   AIC:                         7.845e+04
Df Residuals:                    9566   BIC:                         7.847e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -1055.2610     25.459    -41.449      0.000   -1105.167   -1005.355
AP             1.4899      0.025     59.296      0.000       1.441       1.539
==============================================================================
Omnibus:                      525.438   Durbin-Watson:                   1.996
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              612.290
Skew:                           0.616   Prob(JB):                    1.10e-133
Kurtosis:                       2.859   Cond. No.                     1.73e+05
==============================================================================
```

```
==============================================================================
Dep. Variable:                      y   R-squared:                       0.152
Model:                            OLS   Adj. R-squared:                  0.152
Method:                 Least Squares   F-statistic:                     1714.
Date:                Fri, 01 Feb 2019   Prob (F-statistic):               0.00
Time:                        21:29:13   Log-Likelihood:                -39933.
No. Observations:                9568   AIC:                         7.987e+04
Df Residuals:                    9566   BIC:                         7.988e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        420.9618      0.823    511.676      0.000     419.349     422.574
RH             0.4557      0.011     41.399      0.000       0.434       0.477
==============================================================================
Omnibus:                      772.278   Durbin-Watson:                   1.998
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              319.245
Skew:                           0.231   Prob(JB):                     4.75e-70
Kurtosis:                       2.234   Cond. No.                         383.
==============================================================================
```







. Conclusion: they are all significant.

Are there any outliers that you would like to remove from your data for each of these regression tasks?

. Please see

 # (d) fit a multiple regression model

OLS Regression Results

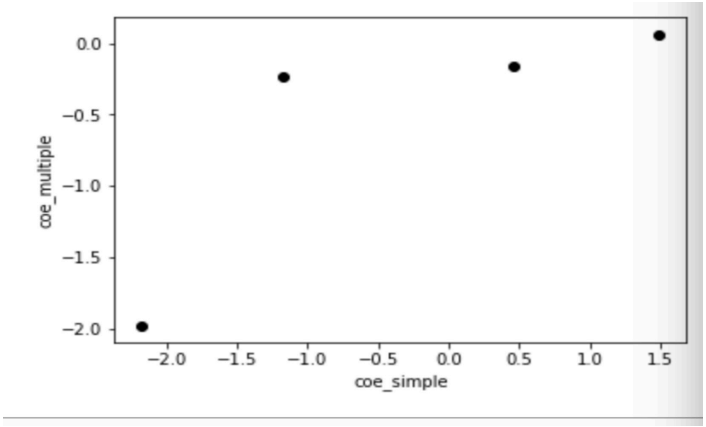
Dep. Variable:	PE	R-squared:	0.929
Model:	OLS	Adj. R-squared:	0.929
Method:	Least Squares	F-statistic:	3.114e+04
Date:	Sat, 02 Feb 2019	Prob (F-statistic):	0.00
Time:	23:19:12	Log-Likelihood:	-28088.
No. Observations:	9568	AIC:	5.619e+04
Df Residuals:	9563	BIC:	5.622e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	454.6093	9.749	46.634	0.000	435.500	473.718
AT	-1.9775	0.015	-129.342	0.000	-2.007	-1.948
V	-0.2339	0.007	-32.122	0.000	-0.248	-0.220
AP	0.0621	0.009	6.564	0.000	0.044	0.081
RH	-0.1581	0.004	-37.918	0.000	-0.166	-0.150

Omnibus:	892.002	Durbin-Watson:	2.033
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4086.777
Skew:	-0.352	Prob(JB):	0.00
Kurtosis:	6.123	Cond. No.	2.13e+05

Description:
r-squared = 0.929
P-values for all predictors are statistically significant.
We could not reject any of them.

 # (e)



 # (f) Is there evidence of nonlinear association between any of the predictors and the response?

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

For "AT", p-values for all coefficients are significant.

OLS Regression Results

Dep. Variable:	PE	R-squared:	0.912
Model:	OLS	Adj. R-squared:	0.912
Method:	Least Squares	F-statistic:	3.299e+04
Date:	Sat, 02 Feb 2019	Prob (F-statistic):	0.00
Time:	23:25:51	Log-Likelihood:	-29101.
No. Observations:	9568	AIC:	5.821e+04
Df Residuals:	9564	BIC:	5.824e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	492.7281	0.673	732.248	0.000	491.409	494.047
x1	-0.6103	0.124	-4.941	0.000	-0.852	-0.368
x2	-0.1251	0.007	-18.199	0.000	-0.139	-0.112
x3	0.0027	0.000	22.594	0.000	0.002	0.003

For "V", p-values for X2 and X3 are NOT significant.

OLS Regression Results

Dep. Variable:	PE	R-squared:	0.775
Model:	OLS	Adj. R-squared:	0.775
Method:	Least Squares	F-statistic:	1.098e+04
Date:	Sat, 02 Feb 2019	Prob (F-statistic):	0.00
Time:	23:25:51	Log-Likelihood:	-33585.
No. Observations:	9568	AIC:	6.718e+04
Df Residuals:	9564	BIC:	6.721e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	554.1468	9.151	60.557	0.000	536.209	572.084
x1	-2.1444	0.509	-4.214	0.000	-3.142	-1.147
x2	-0.0027	0.009	-0.294	0.768	-0.021	0.015
x3	0.0001	5.45e-05	2.465	0.014	2.75e-05	0.000

For “AP”, p-values for all coefficients are significant.

Dep. Variable:	PE	R-squared:	0.275
Model:	OLS	Adj. R-squared:	0.275
Method:	Least Squares	F-statistic:	1813.
Date:	Sat, 02 Feb 2019	Prob (F-statistic):	0.00
Time:	23:25:51	Log-Likelihood:	-39184.
No. Observations:	9568	AIC:	7.837e+04
Df Residuals:	9565	BIC:	7.840e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0747	0.009	8.415	0.000	0.057	0.092
x1	25.2556	3.001	8.415	0.000	19.372	31.139
x2	-0.0500	0.006	-8.439	0.000	-0.062	-0.038
x3	2.514e-05	2.92e-06	8.613	0.000	1.94e-05	3.09e-05

For “RH”, p-values for all coefficients are significant.

OLS Regression Results

Dep. Variable:	PE	R-squared:	0.154
Model:	OLS	Adj. R-squared:	0.153
Method:	Least Squares	F-statistic:	579.2
Date:	Sat, 02 Feb 2019	Prob (F-statistic):	0.00
Time:	23:25:51	Log-Likelihood:	-39923.
No. Observations:	9568	AIC:	7.985e+04
Df Residuals:	9564	BIC:	7.988e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	468.4135	10.545	44.422	0.000	447.744	489.083
x1	-1.7292	0.486	-3.557	0.000	-2.682	-0.776
x2	0.0321	0.007	4.433	0.000	0.018	0.046
x3	-0.0002	3.51e-05	-4.340	0.000	-0.000	-8.34e-05

Conclusion: For “AT, AP, RH”, p-values for all coefficients are significant.
For “V”, p-values for X2 and X3 are NOT significant.

 # g Interaction.

OLS Regression Results

Dep. Variable:	PE	R-squared:	0.936
Model:	OLS	Adj. R-squared:	0.936
Method:	Least Squares	F-statistic:	1.405e+04
Date:	Sat, 02 Feb 2019	Prob (F-statistic):	0.00
Time:	23:45:33	Log-Likelihood:	-27548.
No. Observations:	9568	AIC:	5.512e+04
Df Residuals:	9557	BIC:	5.520e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	685.7825	78.640	8.721	0.000	531.631	839.934
AT	-4.3470	2.373	-1.832	0.067	-8.999	0.305
V	-7.6749	1.351	-5.682	0.000	-10.323	-5.027
AP	-0.1524	0.077	-1.983	0.047	-0.303	-0.002
RH	1.5709	0.773	2.031	0.042	0.055	3.087
AT*V	0.0210	0.001	23.338	0.000	0.019	0.023
AT*AP	0.0018	0.002	0.752	0.452	-0.003	0.006
AT*RH	-0.0052	0.001	-6.444	0.000	-0.007	-0.004
V*AP	0.0068	0.001	5.135	0.000	0.004	0.009
V*RH	0.0008	0.000	1.716	0.086	-0.000	0.002
AP*RH	-0.0016	0.001	-2.125	0.034	-0.003	-0.000

Description:
There is evidence of association of interactions of predictors with the response.
Interaction term “AT*V”, “AT*RH”, “V*AP” statistically significant.

 # h – i
Train the regression model on a randomly selected 70% subset of the data with all predictors.

OLS Regression Results

Dep. Variable:	PE	R-squared:	0.929
Model:	OLS	Adj. R-squared:	0.929
Method:	Least Squares	F-statistic:	2.194e+04
Date:	Sat, 02 Feb 2019	Prob (F-statistic):	0.00
Time:	23:52:08	Log-Likelihood:	-19630.
No. Observations:	6697	AIC:	3.927e+04
Df Residuals:	6692	BIC:	3.930e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	467.8414	11.502	40.673	0.000	445.293	490.390
AT	-2.0044	0.018	-110.340	0.000	-2.040	-1.969
V	-0.2271	0.009	-26.301	0.000	-0.244	-0.210
AP	0.0493	0.011	4.416	0.000	0.027	0.071
RH	-0.1600	0.005	-32.337	0.000	-0.170	-0.150

H – ii

run a regression model involving all possible interaction terms and quadratic nonlinearities, and remove insignificant variables using p-values

First put all 14 features into the regression model, and then use backward selection to remove the feature with the highest p-value. Stop when all p-values show 0.000.
I removed the features in the order of 'V*RH', 'V**2'. 'V*AP', 'AT*AP'.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     PE   R-squared:                       0.938
Model:                            OLS   Adj. R-squared:                  0.938
Method:                 Least Squares   F-statistic:                 1.017e+04
Date:                Sun, 03 Feb 2019   Prob (F-statistic):               0.00
Time:                        00:01:05   Log-Likelihood:                -19166.
No. Observations:                6697   AIC:                          3.835e+04
Df Residuals:                    6686   BIC:                          3.843e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -1.046e+04   1091.512     -9.581      0.000   -1.26e+04   -8317.797
AT            -2.4293      0.100    -24.221      0.000      -2.626      -2.233
V             -0.4542      0.032    -14.187      0.000      -0.517      -0.391
AP            21.1531      2.158      9.804      0.000      16.924      25.383
RH             5.6754      0.755      7.512      0.000       4.194       7.156
AT*V           0.0080      0.001      5.509      0.000       0.005       0.011
AT*RH         -0.0069      0.001     -7.937      0.000      -0.009      -0.005
AP*RH         -0.0053      0.001     -7.244      0.000      -0.007      -0.004
AT**2          0.0170      0.002      7.478      0.000       0.013       0.021
AP**2         -0.0102      0.001     -9.558      0.000      -0.012      -0.008
RH**2         -0.0021      0.000     -7.638      0.000      -0.003      -0.002
==============================================================================
```

h – iii

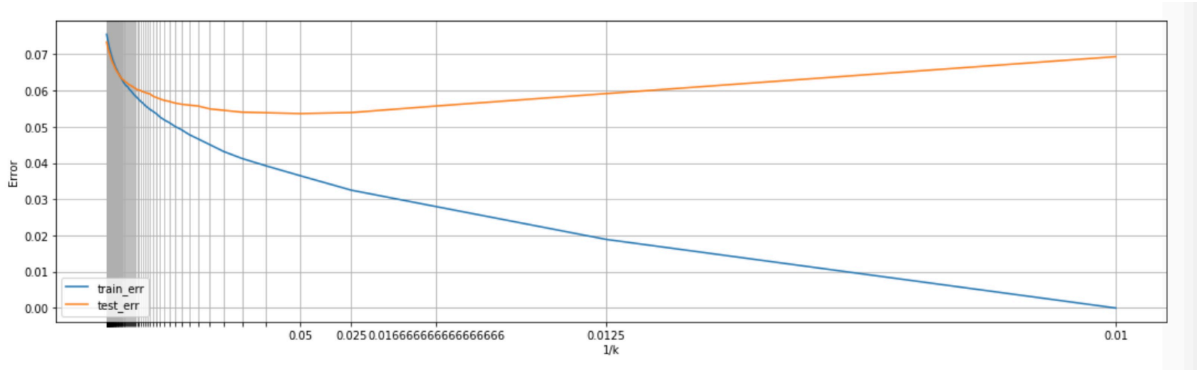
Test both models on the remaining points and report your train and test MSEs.
MSE4 refers to the model with 4 predictors.
MSE10 refers to the model with 10 features.

MSE4test	21.24
MSE4train	20.581
MSE10test	18.694
MSE10train	17.918

 # (i) KNN regression

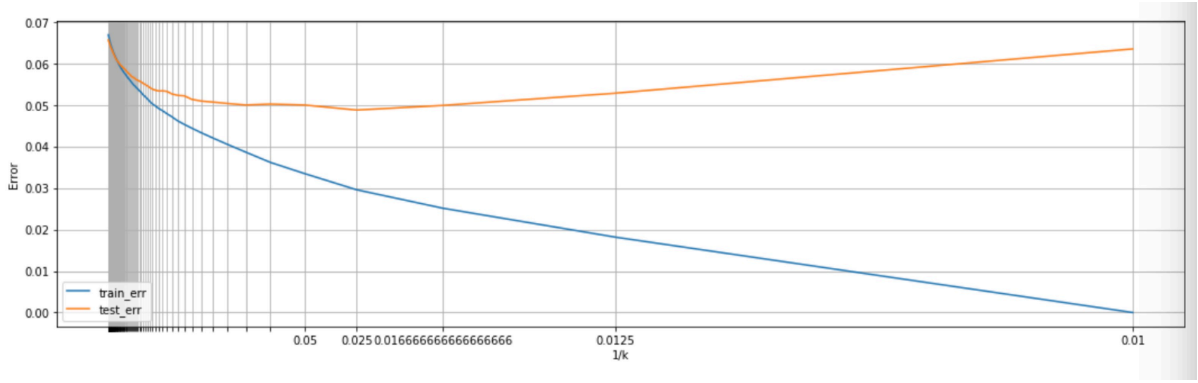
I – raw features

```
When K= 5, test error is minimal: 0.054
```

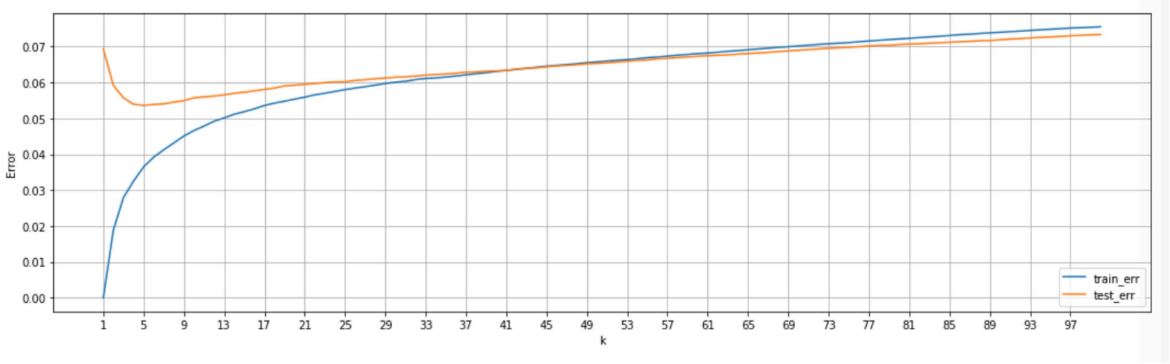


I — SCALED FEATURE

When K= 4, test error is minimal: 0.049



 # (j)



best test err for linear regression is : r**2 = 0.938

when k in {2,3,4....63}, then KNN works better then the best Linear Regression which has r^2=0.938.

ISLR

2.4.1

 # (a)

when sample size n is extremely large, and the number of predictors p is small, we would generally expect the performance of a flexible statistical learning method is

better than an inflexible method.
Since n is large, means we do have a lot of data to test the model and because p is small,
It is less likely to over fit the model.
In this case, we can try to conduct a flexible method.

 # (b)

when number of predictors p is extremely large, and the number of observations n is small, we would generally expect the performance of an inflexible statistical method is better then flexible method.
Since n is small, and p is extremely large, p might bigger than n.
If p is bigger than n, it is hard to test significance of each predictor,
and over fit might happen.
In this case, we expect an inflexible method.

 # (c)

when the relationship between the predictors and response is highly non-linear,
we would expect the flexible method works better.
Since there is non-linear, it means simple linear regression would not be enough.
In this case, we expect a flexible method to perform better.

 # (d)

when the variance of the error terms is extremely high,
it means noise is big.
So we should use an inflexible method.

#2.4.7

(a)

obs=1	distance	3.0	Color:	red
obs=2	distance	2.0	Color:	red
obs=3	distance	3.162	Color:	red
obs=4	distance	2.236	Color:	green
obs=5	distance	1.414	Color:	green
obs=6	distance	1.732	Color:	red

(b) k=1 color?

color = 'green' because when obs = 5, the distance is the shortest.

(c) k=3, color?

color = ‘red’ because closest three obs = 5, 6, 2.
2 out 3 are 'red.'

when	k=1	['green']
when	k=2	['green']
when	k=3	['red']

when	k=4	['green']
when	k=5	['red']

(d) If the Bayes decision boundary in this problem is highly non- linear, then would we expect the best value for K to be large or small? Why?

K should be small. Smaller K would bring more flexibilities since the model is highly non-linear, we need a flexible method.