ASSIGNMENT REPORT

2A:

Principal component analysis (PCA): PCA is a statistical procedure that allows you to summarize the information content in large data tables by means of a smaller set of “summary indices” that can be more easily visualized and analyzed.

Principal component regression (PCR): PCR is a regression analysis technique that is based on PCA.

Pearson correlation coefficient (PCC): PCC is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

PCA:

Feature	**Variance captured by each feature**
1	5.48731144e-01
2	1.92813088e-01
3	7.35309249e-02
4	6.31572405e-02
5	4.71656518e-02
6	3.14033504e-02
7	2.23317847e-02
8	5.35698043e-03
9	3.62874021e-03
10	3.49194138e-03
11	1.87469442e-03
12	1.59380795e-03

13	1.25540155e-03
14	1.19839536e-03
15	6.11145629e-04
16	4.28672222e-04
17	3.34703905e-04
18	2.93863064e-04
19	2.27397278e-04
20	1.89879679e-04
21	1.38210786e-04
22	1.00460752e-04
23	6.30531176e-05
24	4.59147186e-05
25	3.35533594e-05
26	2.96377187e-33

Size of Feature set	**Minimum training error**	**Minimum testing error**
1	5475.6543	5052.1854
2	5475.6540	5052.2260
3	5475.3439	5054.8476
4	5475.3214	5055.1062
5	5361.06381	4951.6232
6	5359.0517	4960.1709
7	5353.3342	4950.1681
8	4950.1681	4950.7214
9	5305.6464	4947.1827

10	5230.8101	4865.3772
11	5221.3176	4888.9139
12	5057.8778	4741.3293
13	5057.7243	4743.8077
14	5023.1442	4728.2930
15	5008.1167	4717.1585
16	5006.1033	4735.4546
17	5005.8473	4737.7148
18	5005.3811	4737.9703
19	4897.4472	4661.2507
20	4896.7766	4660.8957
21	4869.7337	4629.2286
22	4769.8110	4530.9834
23	4739.8143	4494.3016
24	4724.6843	4488.7075
25	4722.5256	4489.4170
26	4722.5470	4489.8931

Pearson Correlation:

Size of Feature set	Minimum training error	Minimum testing error
1	5392.0778	4967.93021
2	5388.8171	4955.0967
3	5386.4775	4955.5322
4	5374.6035	4945.5430
5	5372.1422	4937.6952

6	5341.2007	4873.4968
7	5290.4869	4882.05422
8	5250.8987	4841.7667
9	5087.9548	4711.0290
10	5005.5596	4633.7378
11	4985.8061	4626.2657
12	4979.2784	4634.3607
13	4966.7537	4620.2467
14	4953.1673	4612.1162
15	4954.1332	4616.0165
16	4954.3275	4617.0483
17	4918.6459	4649.1929
18	4897.5756	4641.7657
19	4897.2695	4637.3046
20	4881.6924	4633.1586
21	4877.5406	4618.9635
22	4877.0762	4615.4979
23	4875.8047	4613.5309
24	4849.8889	4598.2297
25	4849.7040	4597.4068
26	4849.7030	4597.4009

2B:

The submitted code uses the gradient descent algorithm used previously to find and return minimum testing and training error.
The Algorithm is implemented in the following way:

- At the beginning , we read the data from the provided csv file using the pandas package in python.
- Later , we Normalise and then shuffle the data using appropriate inbuilt python functions.
- Then we perform the requested 80-20 split for training and testing data models.

The algorithms used are described below:

- Forward Selection: Forward selection is an iterative method in which we start with having no feature in the model. In each iteration, we keep adding the feature which best improves our model till an addition of a new variable does not improve the performance of the model.

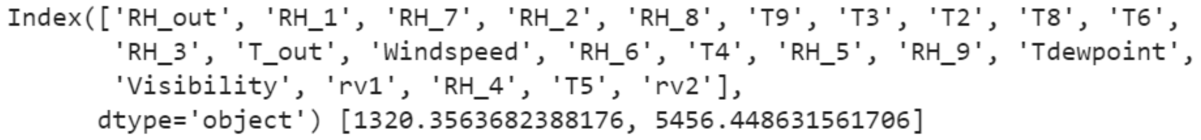
- Backward Elimination: In backward elimination, we start with all the features and removes the least significant feature at each iteration which improves the performance of the model. We repeat this until no improvement is observed on removal of features.

We have currently implement Forward Selection and Backward Elimination

Greedy Forward Selection

Size of Feature set	**Minimum testing error**	**Minimum training error**
1	1514.6902	6182.0988
2	1492.0149	6080.3331
3	1450.0074	5917.5312
4	1423.3719	5819.9712
5	1416.5578	5789.6621
6	1406.4179	5766.3617
7	1368.6529	5592.1788
8	1359.3154	5553.8335
9	1353.2345	5527.0720
10	1345.0807	5508.3836
11	1338.1567	5500.1137

12	1333.0890	5492.9345
13	1329.3477	5480.5061
14	1326.5313	5477.1114
15	1323.5774	5474.1741
16	1322.5808	5468.7408
17	1321.9484	5468.3758
18	1321.5592	5467.2803
19	1321.2264	5465.4820
20	1320.9138	5465.3467
21	1320.6552	5456.8860
22	1320.3563	5456.4486
23	1320.3563	5456.4486
24	1320.3627	5456.1594
25	1320.4248	5456.1520
26	1321.8437	5455.3079

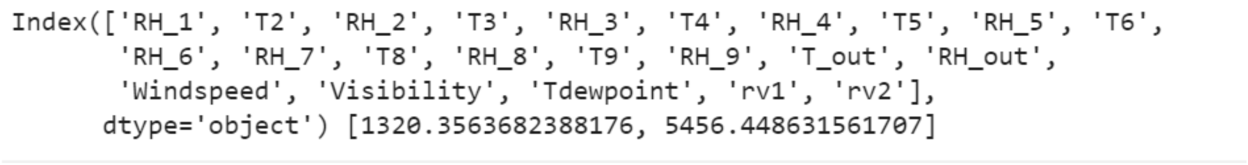


```
Index(['RH_out', 'RH_1', 'RH_7', 'RH_2', 'RH_8', 'T9', 'T3', 'T2', 'T8', 'T6',
       'RH_3', 'T_out', 'Windspeed', 'RH_6', 'T4', 'RH_5', 'RH_9', 'Tdewpoint',
       'Visibility', 'rv1', 'RH_4', 'T5', 'rv2'],
      dtype='object') [1320.3563682388176, 5456.448631561706]
```

Greedy Backward Elimination

Size of Feature set	Minimum testing error	Minimum training error
1	1557.5036	6301.2635
2	1477.8433	6019.4832
3	1424.3228	5822.3518
4	1413.7135	5793.3227

5	1375.7774	5620.8901
6	1362.8160	5563.2537
7	1355.5859	5531.0683
8	1346.5276	5514.5499
9	1341.0281	5504.2492
10	1336.7570	5492.0573
11	1331.2710	5486.4798
12	1328.9742	5484.2771
13	1326.0642	5480.0230
14	1323.3382	5474.3012
15	1322.3194	5468.9446
16	1321.6927	5468.5196
17	1321.5328	5459.7504
18	1321.2216	5459.2360
19	1320.8902	5457.1861
20	1320.6016	5457.0648
21	1320.5662	5457.0626
22	1320.3563	5456.4486
23	1320.3563	5456.4486
24	1320.3627	5456.1594
25	1320.4242	5456.15209
26	1321.8437	5455.3079



```
Index(['RH_1', 'T2', 'RH_2', 'T3', 'RH_3', 'T4', 'RH_4', 'T5', 'RH_5', 'T6',
       'RH_6', 'RH_7', 'T8', 'RH_8', 'T9', 'RH_9', 'T_out', 'RH_out',
       'Windspeed', 'Visibility', 'Tdewpoint', 'rv1', 'rv2'],
      dtype='object') [1320.3563682388176, 5456.448631561707]
```

2C:

From the above data obtained of the using the four feature selection techniques as listed above

1. Pearson Correlation Coefficient
2. Principal Component Analysis
3. Greedy Forward Selection
4. Greedy Backward Elimination

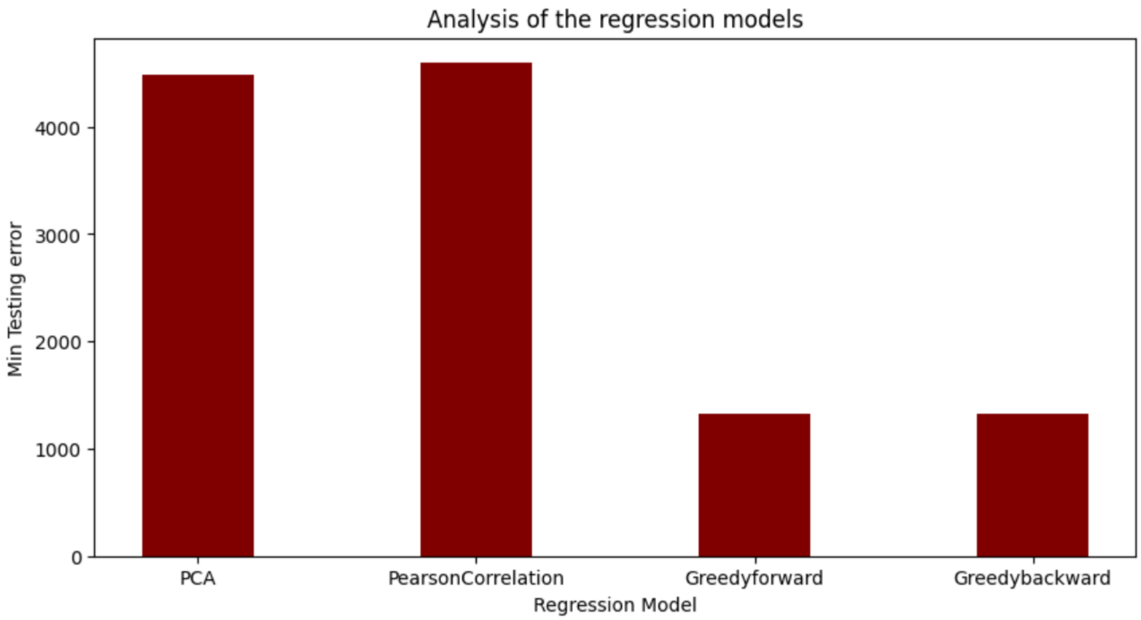
Using Pearson Correlation Coefficient, we get the least testing error of 4597.4009 with 26 features.
Using Principal Component Analysis, we get the least testing error of 4488.7075 with 24 features.
Using Greedy Forward Selection, we get the least testing error of 1320.3563 with 23 features.
Using Greedy Backward Elimination, we get the least testing error of 1320.3563 with 23 features.

Hence, we can conclude that Greedy Forward Selection and Greedy Backward Elimination performan the best.

For the case of choosing all (26) features, we get the testing errors for all four feature selection techniques as follows.

1. Pearson Correlation Coefficient :- 4597.4009
2. Principal Component Analysis :- 4489.8931
3. Greedy Forward Selection :- 1321.8437
4. Greedy Backward Elimination :- 1321.8437

Hence, we can again conclude that Greedy Forward Selection and Greedy Backward Elimination performan the best.


Analysis of the regression models
Min Testing error
0
1000
2000
3000
4000
PCA
PearsonCorrelation
Greedyforward
Greedybackward
Regression Model