My Name (shixuef2)
IE598 MLF F19
Module 5 Homework (Dimensionality Reduction)

Part 1 Exploratory Data Analysis

The Shape, Head and Tail:

```
(8071, 31)
   SVENF01  SVENF02  SVENF03  SVENF04  ...  SVENF28  SVENF29  SVENF30  Adj_Close
0   2.1224   2.0266   2.1023   2.2377  ...   3.6471   3.6970   3.7458  10.130177
1   2.1239   2.0317   2.1096   2.2468  ...   3.6660   3.7153   3.7636  10.130177
2   2.0874   1.9956   2.0844   2.2289  ...   3.6421   3.6847   3.7257  10.150118
3   2.1319   2.0559   2.1451   2.2856  ...   3.7132   3.7630   3.8113  10.130177
4   2.1051   2.0234   2.1180   2.2632  ...   3.6655   3.7098   3.7525  10.130177

[5 rows x 31 columns]
      SVENF01  SVENF02  SVENF03  SVENF04  ...  SVENF28  SVENF29  SVENF30  Adj_Close
8066   6.1632   6.6192   6.9560   7.2403  ...   8.4805   8.4806   8.4807   2.942279
8067   6.2091   6.6589   6.9843   7.2634  ...   8.5768   8.5769   8.5770   2.942279
8068   6.2195   6.6790   7.0240   7.3172  ...   8.5965   8.5966   8.5966   2.942279
8069   6.2215   6.6978   7.0637   7.3688  ...   8.5524   8.5525   8.5525   2.942279
```

The Summary:

```
           SVENF01      SVENF02  ...      SVENF30    Adj_Close
count  8071.000000  8071.000000  ...  8071.000000  8071.000000
mean      3.785311     4.258972  ...     5.167371     5.509793
std       2.648060     2.498137  ...     1.847834     2.491110
min       0.072700     0.327300  ...     0.411100     2.801050
25%       1.144050     1.865600  ...     3.831350     3.130587
50%       3.986500     4.393300  ...     4.669000     4.956219
75%       5.901500     6.221250  ...     6.421850     8.051437
max       9.813800     9.887800  ...    10.535100    10.150118
```

The Summary Statistics for each Feature/Target Column (First six features):

Features:

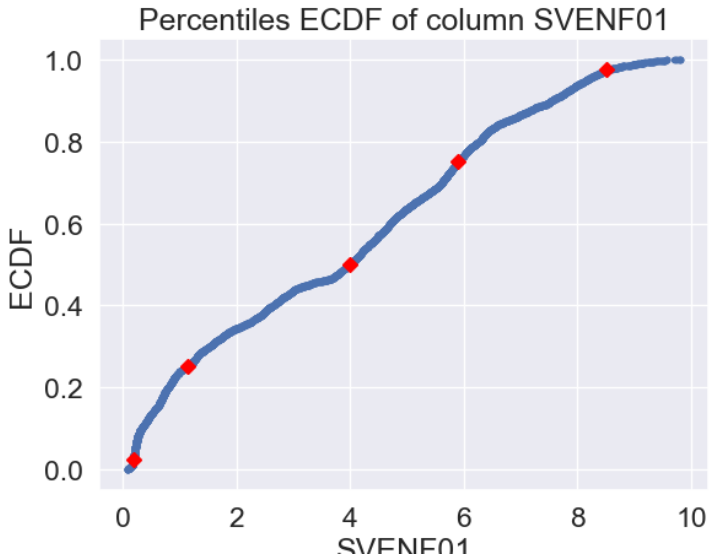
```
The summary statistics of SVENF01
Mean =  3.7853113740552593      Standard Deviation =     2.647895705006851

Boundaries for 4 Equal Percentiles
[0.1887  1.14405 3.9865  5.9015  8.5091 ]

The summary statistics of SVENF02
Mean =  4.258972085243462       Standard Deviation =     2.497981991651637

Boundaries for 4 Equal Percentiles
[0.4735   1.8656   4.3933   6.22125  8.679775]

The summary statistics of SVENF03
Mean =  4.669362829884772       Standard Deviation =     2.341202680682377

Boundaries for 4 Equal Percentiles
[0.909475 2.53655  4.5055   6.4613   8.8504  ]

The summary statistics of SVENF04
Mean =  5.02242974848223        Standard Deviation =     2.221494421670246

Boundaries for 4 Equal Percentiles
[1.4234  3.02305 4.7189  6.6266  9.0034 ]

The summary statistics of SVENF05
Mean =  5.318492888117949       Standard Deviation =     2.1376689073756525

Boundaries for 4 Equal Percentiles
[1.826725 3.5447   5.0513   6.77955  9.160025]

The summary statistics of SVENF06
Mean =  5.559644467847845       Standard Deviation =     2.0802764669332507

Boundaries for 4 Equal Percentiles
[2.182425 4.0633   5.3946   6.90805  9.302225]
```

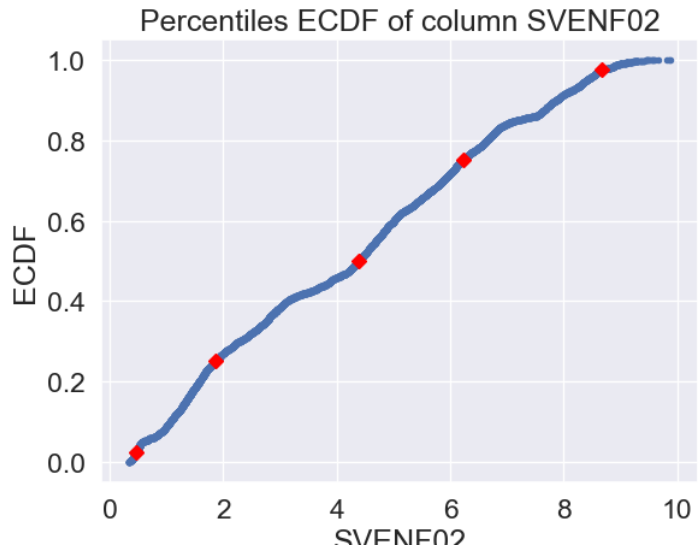
Target:

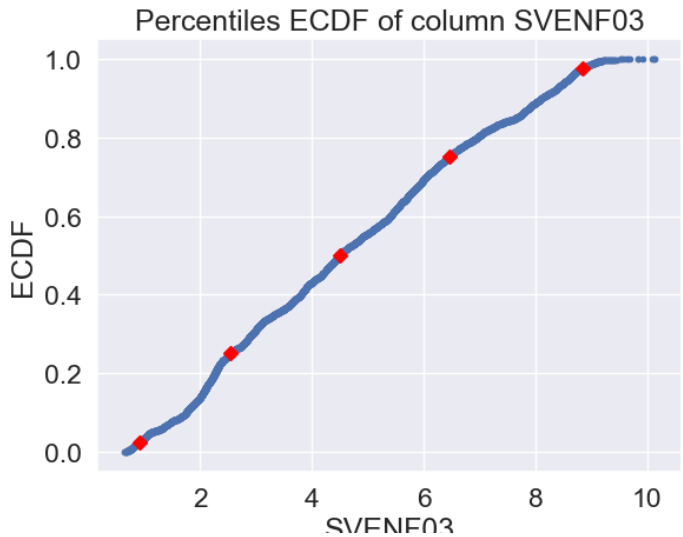
```
The summary statistics of 'Adj_Close'
Mean =  5.509793467352371       Standard Deviation =     2.490956074581464

Boundaries for 4 Equal Percentiles
[2.871666 3.130587 4.956219 8.051437 9.794909]
```

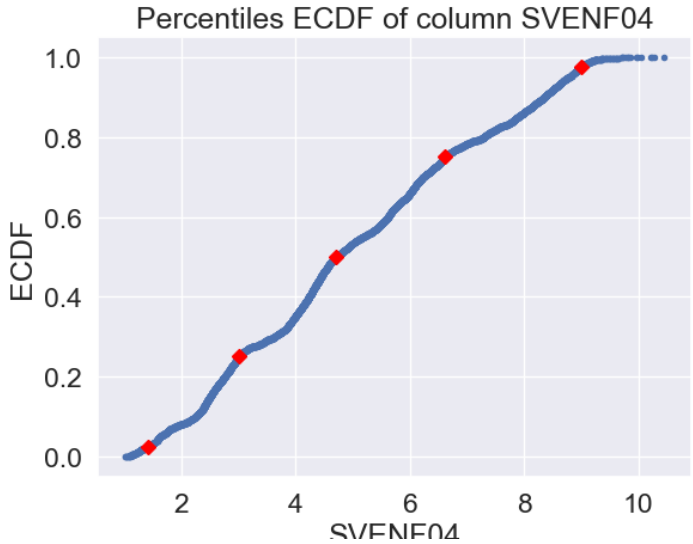
The Percentile ECDF for each Feature/Target Column (First six features):

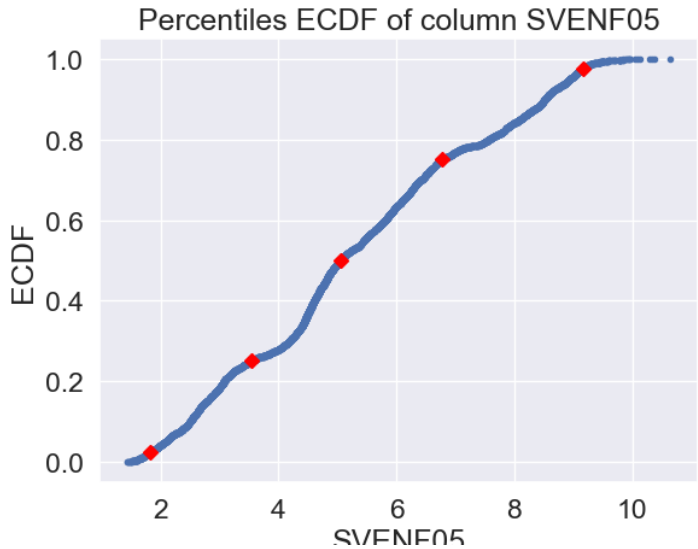
Features:

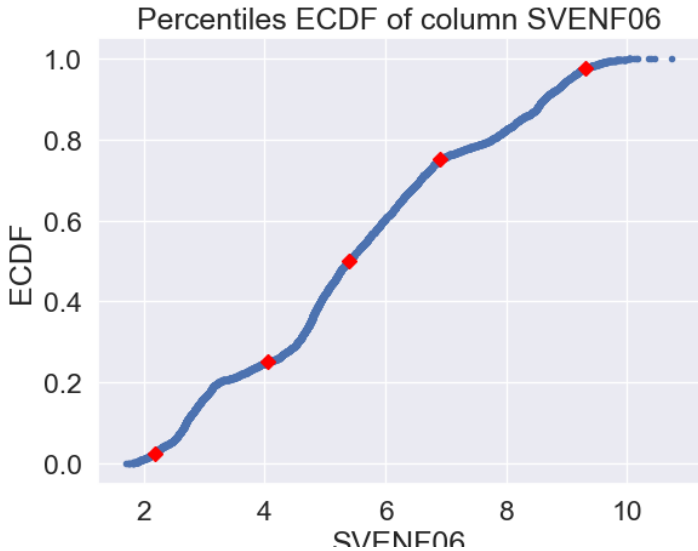




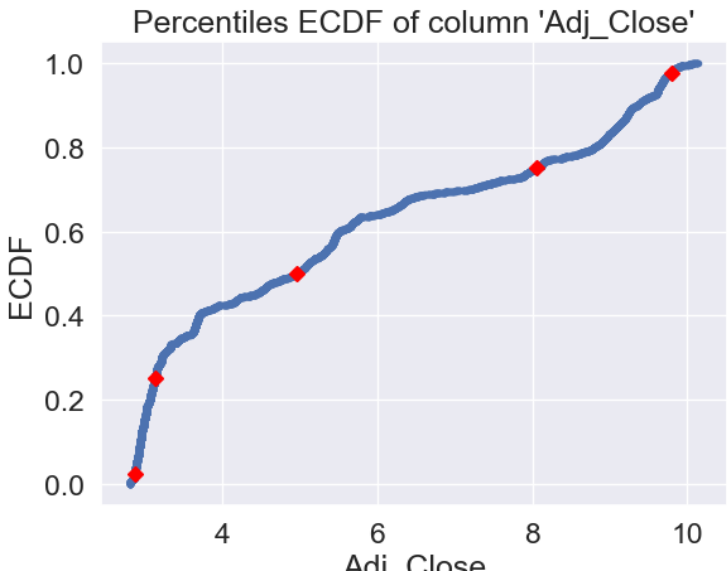



Percentiles ECDF of column SVENF04
ECDF
1.0
0.8
0.6
0.4
0.2
0.0
2
4
6
8
10
SVENF04


Percentiles ECDF of column SVENF05
ECDF
1.0
0.8
0.6
0.4
0.2
0.0
2
4
6
8
10
SVENF05

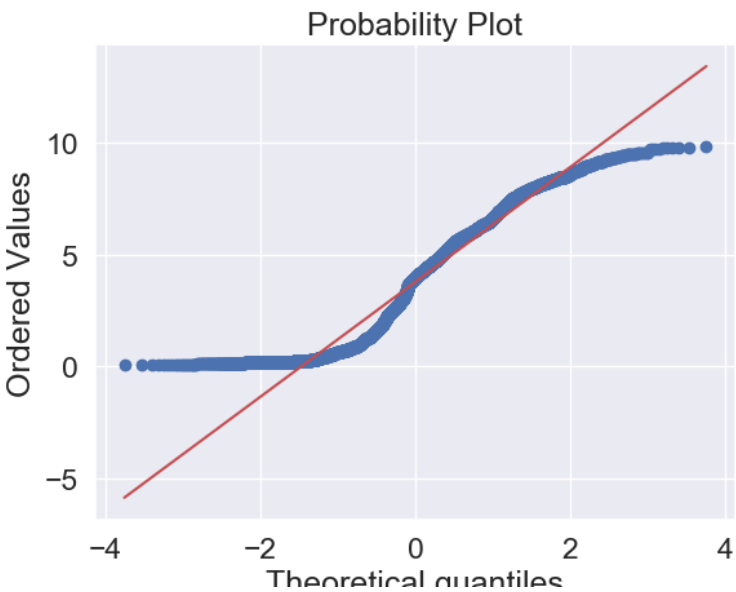

Percentiles ECDF of column SVENF06
ECDF
1.0
0.8
0.6
0.4
0.2
0.0
2
4
6
8
10
SVENF06

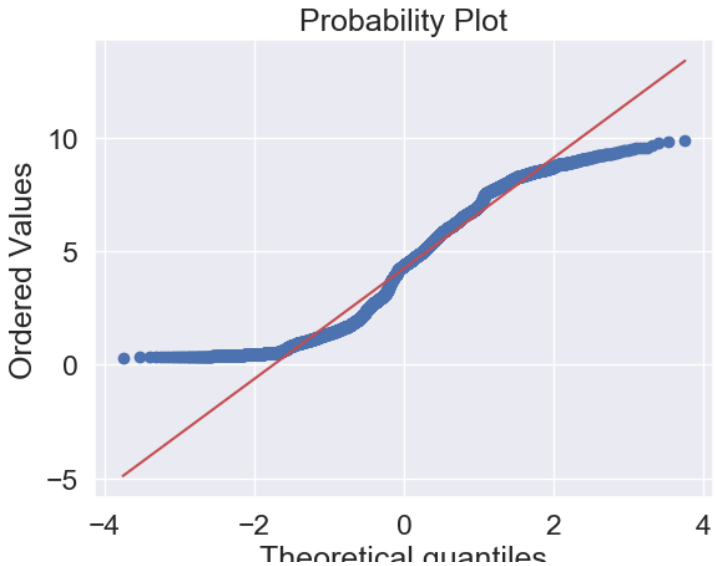
Target:

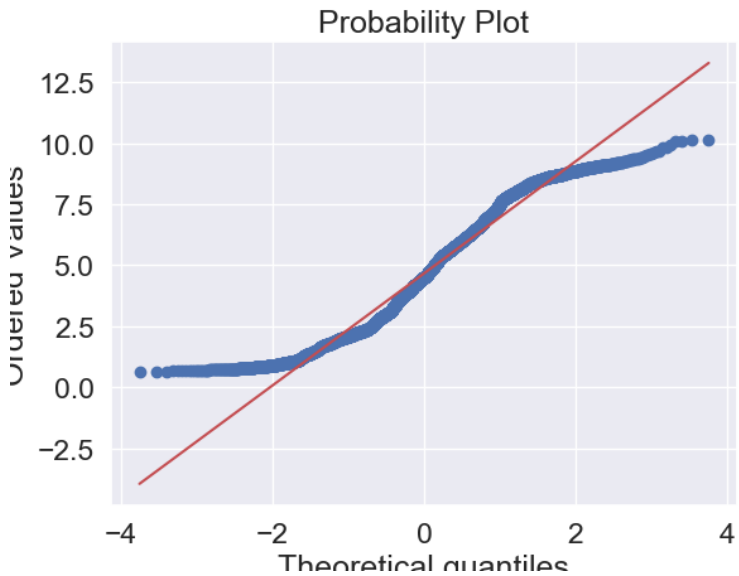


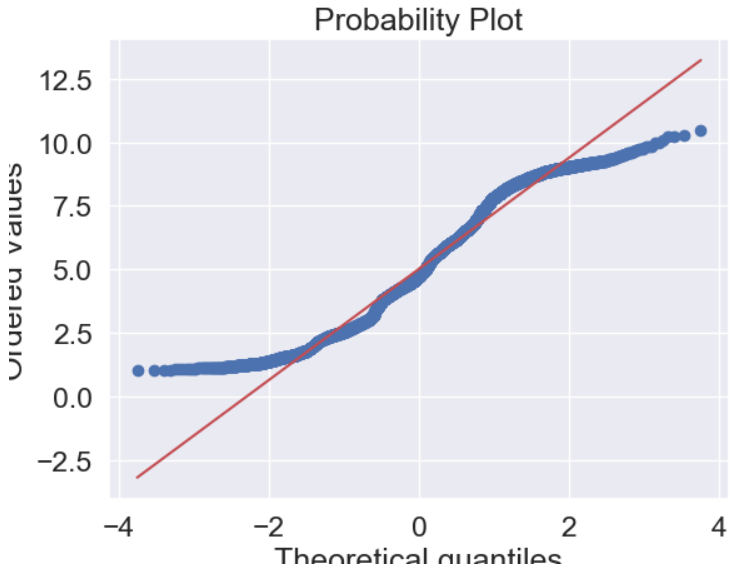
The Percentile ECDF for each Feature/Target Column (First six features):

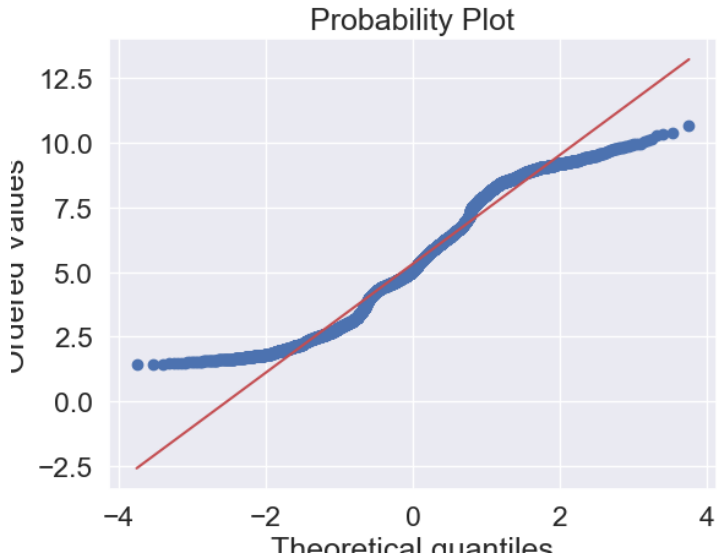
Features:

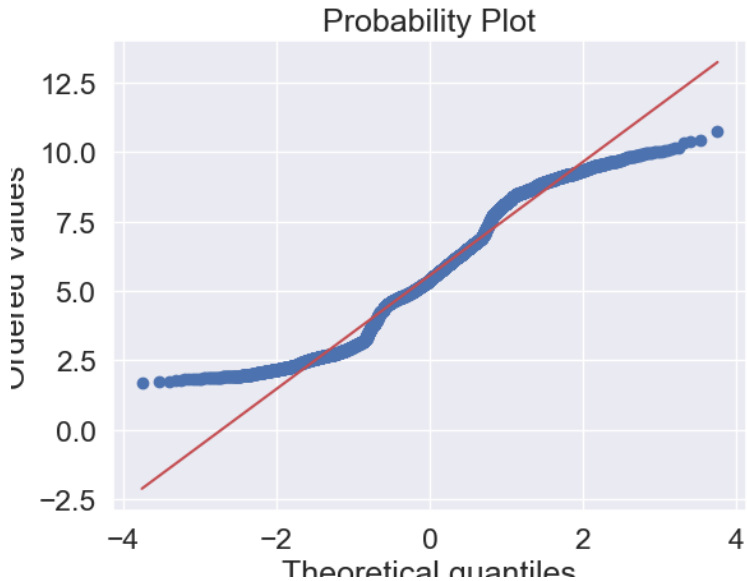



Probability Plot
Ordered Values
10
5
0
−5
−4
−2
0
2
4
Theoretical quantiles

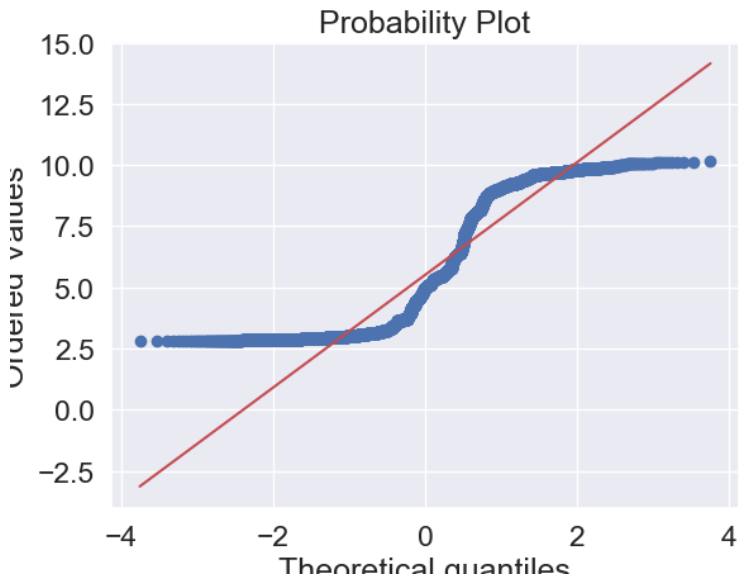

Probability Plot
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−4
−2
0
2
4
Theoretical quantiles


Probability Plot
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−4
−2
0
2
4
Theoretical quantiles


Probability Plot
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−4
−2
0
2
4

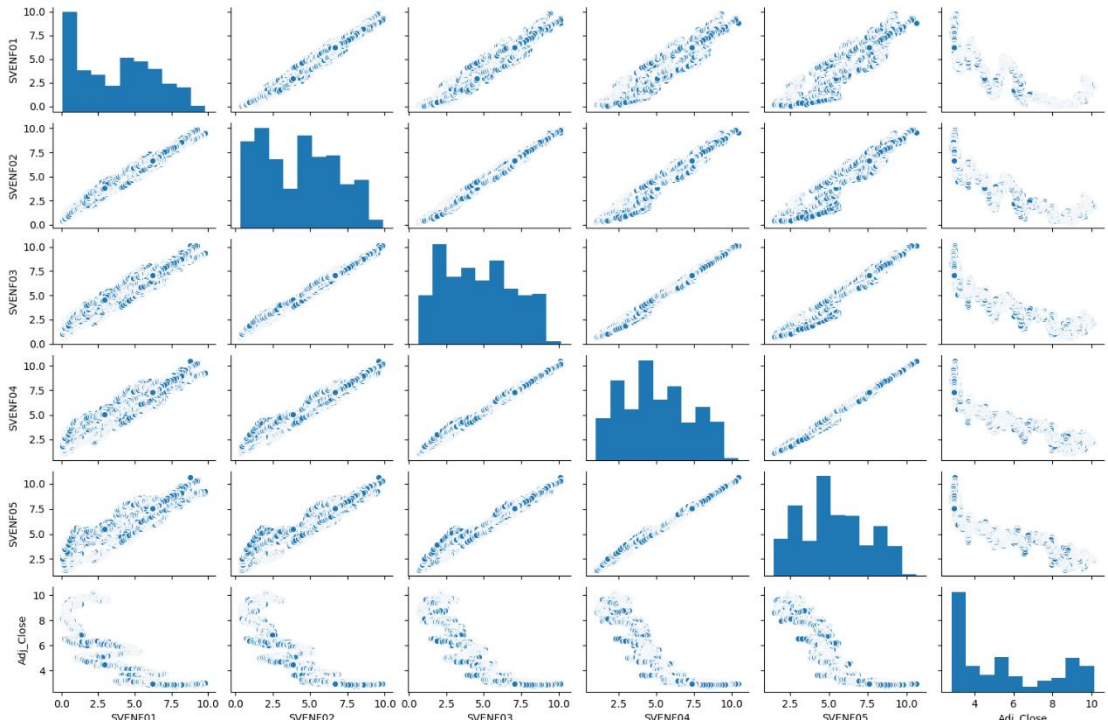

Probability Plot
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−4
−2
0
2
4

Target:

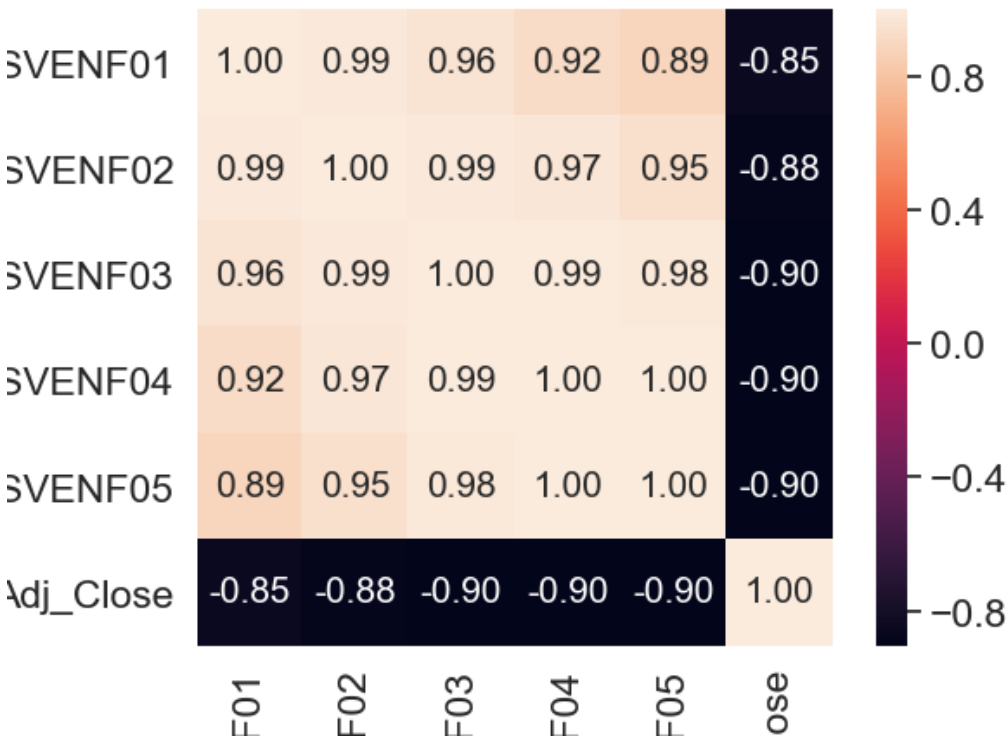

Probability Plot
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−4
−2
0
2
4

The Graphical Summary of the Relationships:

Scatter Plot with first five features and target:



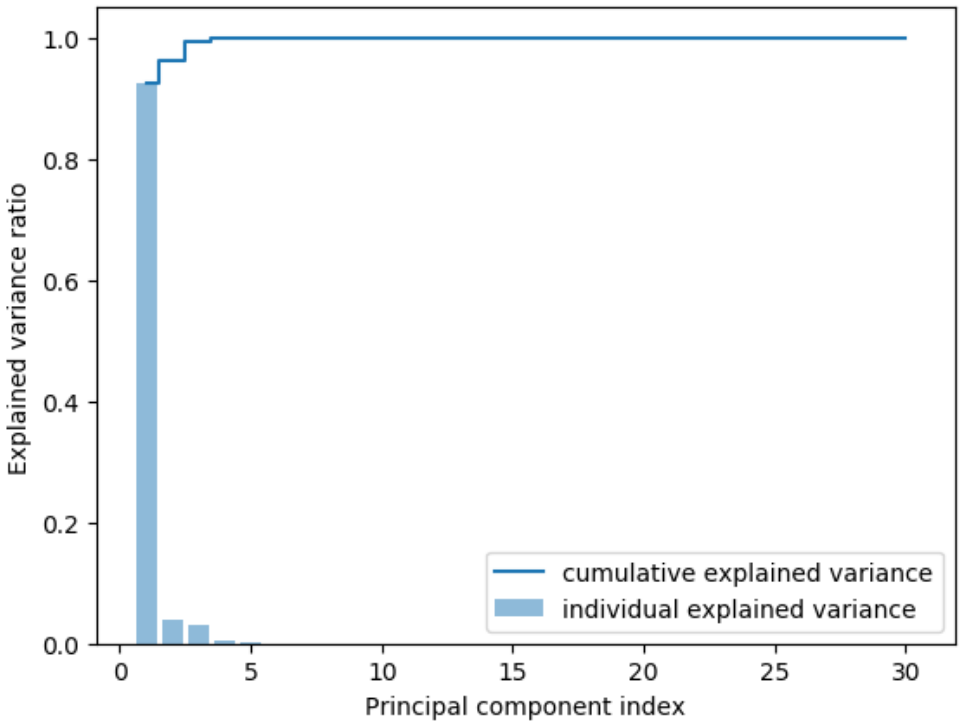
Heat Map:



Part 2 Perform a PCA on the Treasury Yield dataset

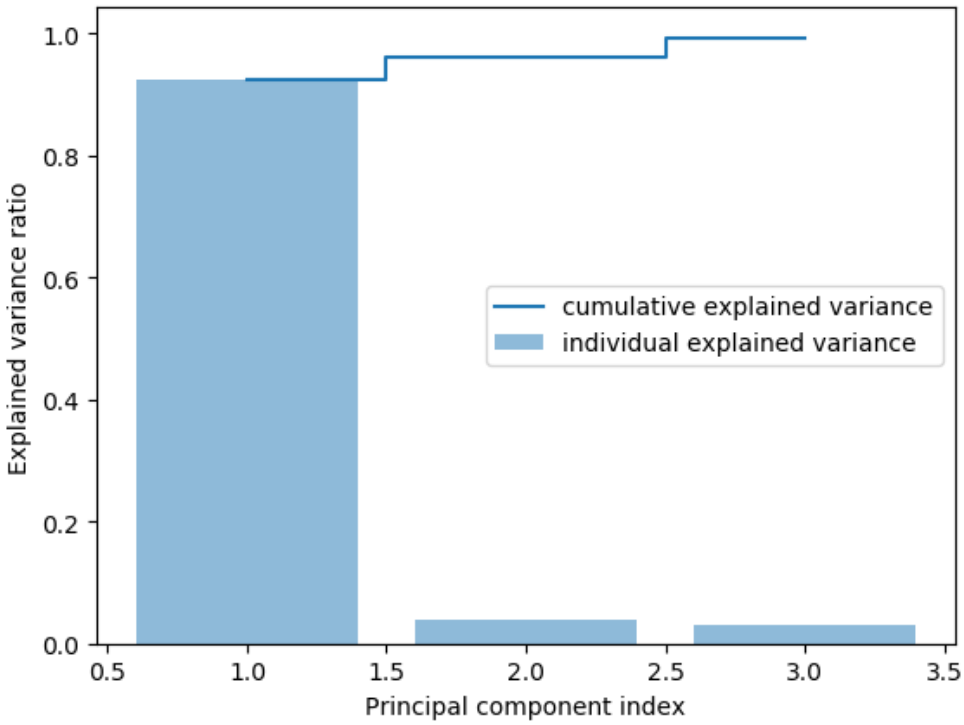
All components:

```
Explained variance ratio:  [9.25027254e-01 3.77198563e-02 3.11962115e-02 5.11829721e-03
 8.45006479e-04 8.14071111e-05 1.06386900e-05 1.23073879e-06
 8.99497477e-08 7.14094977e-09 4.89071592e-10 3.83422436e-11
 8.63162713e-12 7.54060102e-12 7.44722038e-12 7.41409677e-12
 7.37633844e-12 7.36922042e-12 7.21033060e-12 7.16011018e-12
 7.08499808e-12 7.01615861e-12 6.97953948e-12 6.83297854e-12
 6.78790385e-12 6.76011093e-12 6.68796631e-12 6.63106214e-12
 6.57322725e-12 6.42225375e-12]
Explained variance:  [1.07902835e+02 4.39995625e+00 3.63898432e+00 5.97040551e-01
 9.85685500e-02 9.49599926e-03 1.24098486e-03 1.43563560e-04
 1.04924831e-05 8.32979488e-07 5.70493586e-08 4.47255665e-09
 1.00686443e-09 8.79598117e-10 8.68705426e-10 8.64841615e-10
 8.60437170e-10 8.59606866e-10 8.41072642e-10 8.35214516e-10
 8.26452819e-10 8.18422814e-10 8.14151256e-10 7.97055174e-10
 7.91797289e-10 7.88555294e-10 7.80139748e-10 7.73501973e-10
 7.66755634e-10 7.49144835e-10]
```



components=3:

```
Explained variance ratio of the 3-component version:  [0.92502725 0.03771986 0.03119621]
Explained variance of the 3-component version:  [107.90283548   4.39995625   3.63898432]
Cumulative explained variance ratio of the 3-component version:  0.9936054704751927
Cumulative explained variance of the 3-component version:  115.94177605261855
```



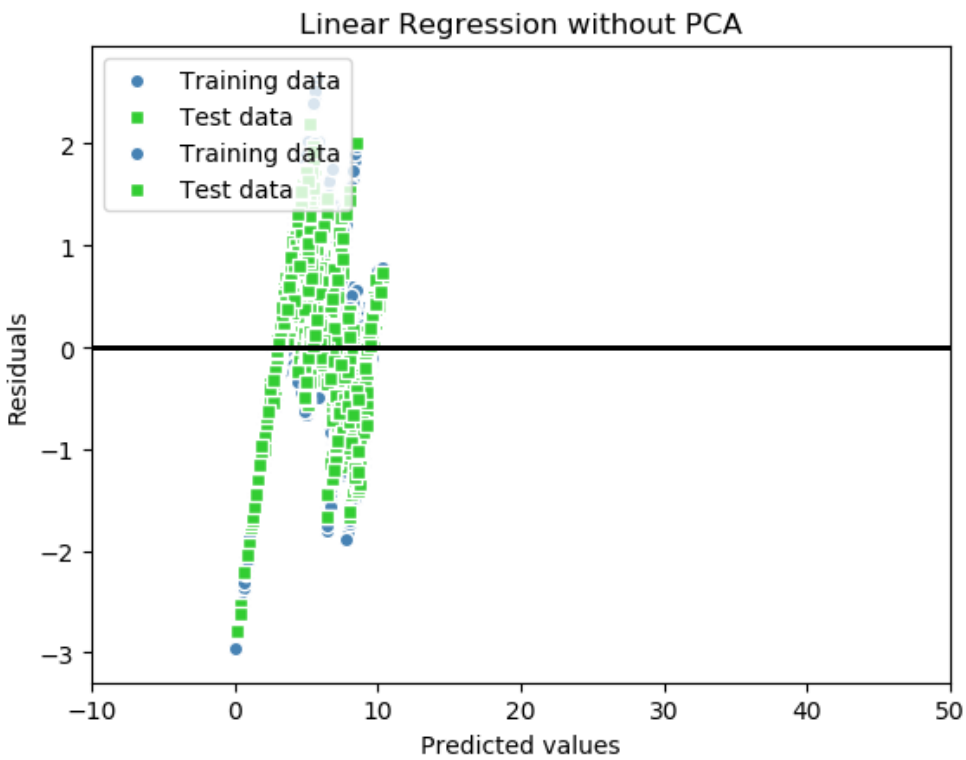
Part 3 Logistic regression classifier v. SVM classifier – baseline

Linear Regression without PCA:

Basic Information:

```
Accuracy R2 score on training set: 0.9022730353400459
Accuracy R2 score on testing set: 0.9041309535336478
Average accuracy R2 score on training set using 5-fold cross-validation: 0.9010357624460449
Average accuracy R2 score on testing set using 5-fold cross-validation: 0.8999735972645933
Root Mean Squared Error of Train Set: 0.776653304036978
Root Mean Squared Error of Test Set: 0.7823695855060767
Slope:
[  -4.83843828   53.15886154 -249.77609515  590.39765971 -686.96356431
  228.09887769  289.2937007  -302.62642323  -44.31624559  320.69207747
 -288.36927381  200.16366115   -0.89256855  -86.9401334   -96.64031266
   -7.50513023 -302.47703104  216.50764238  136.90241245  133.63875552
  562.97736489 -387.63320904  176.17955175 -418.55197044 -795.41172645
  238.76730551  102.69781344  839.17533861  -80.32795403 -336.50681657]
Intercept: 11.807
```

Residual Errors Plot:

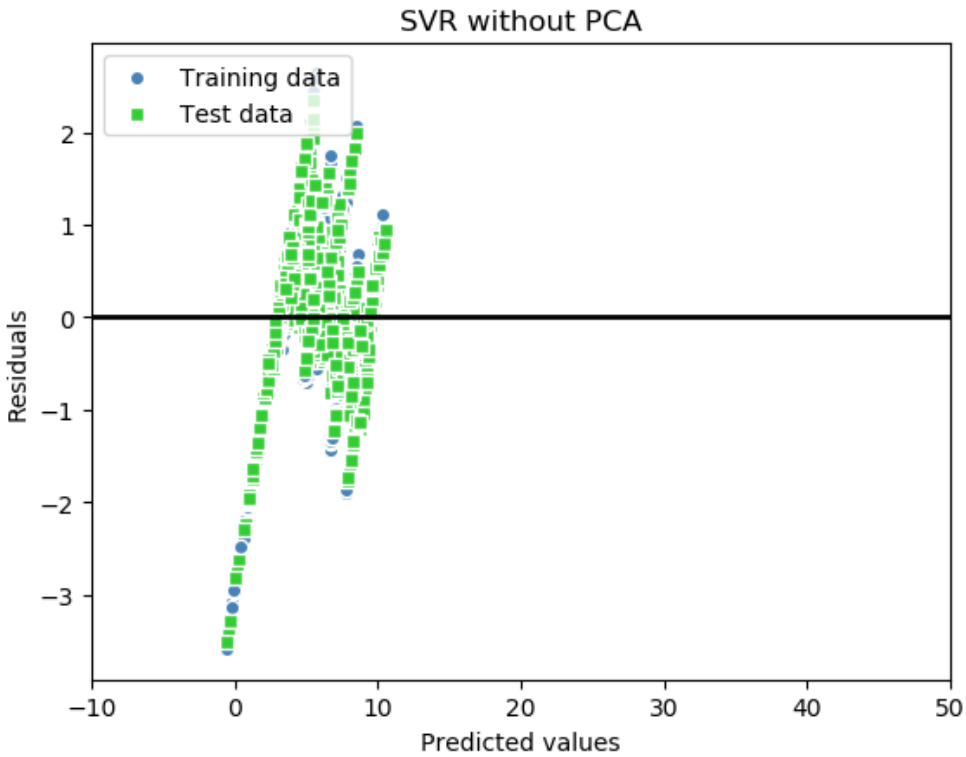


SVR without PCA

Basic Information:

```
Accuracy R2 score on training set: 0.8933230878134524
Accuracy R2 score on testing set: 0.8944429691333482
Average accuracy R2 score on training set using 5-fold cross-validation: 0.8925818002331299
Average accuracy R2 score on testing set using 5-fold cross-validation: 0.8881862799306814
Root Mean Squared Error of Train Set: 0.8114377482436396
Root Mean Squared Error of Test Set: 0.820949302035238
Slope:
[[ 0.28502261  0.17012401 -3.18652086  0.24086713  2.351913    2.18005974
   0.83353371 -0.62989206 -1.6161057  -1.95568616 -1.71667616 -1.07577227
  -0.26184565  0.53300508  1.15724106  1.52993139  1.60312998  1.4044737
   0.97230776  0.38588103 -0.26667849 -0.90004092 -1.41725446 -1.75247579
  -1.82352097 -1.57814458 -0.9889056  -0.01337928  1.34245905  3.09584137]]
Intercept: 11.712
```

Residual Errors Plot:

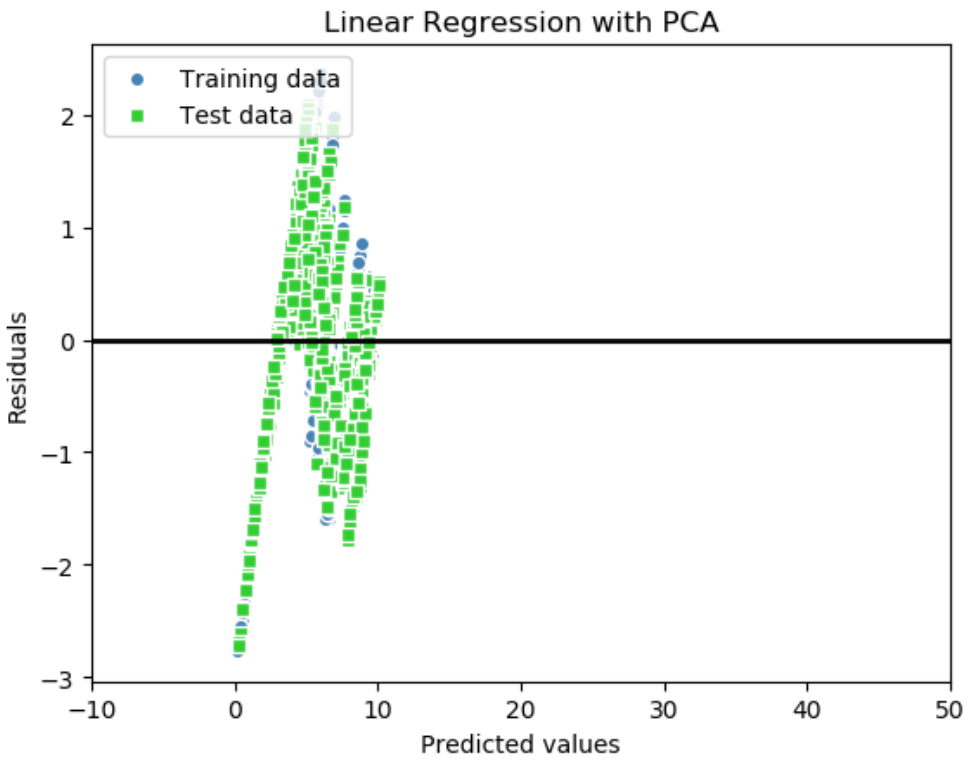


Linear Regression with PCA:

Basic Information:

```
Accuracy R2 score on training set: 0.8673885521430046
Accuracy R2 score on testing set: 0.8663783970490899
Average accuracy R2 score on training set using 5-fold cross-validation: 0.8669518367262727
Average accuracy R2 score on testing set using 5-fold cross-validation: 0.8649701882097635
Root Mean Squared Error of Train Set: 0.904712305616479
Root Mean Squared Error of Test Set: 0.9236577822901371
Slope:
[-0.2175607  -0.24624459  0.01563734]
Intercept: 5.510
```

Residual Errors Plot:

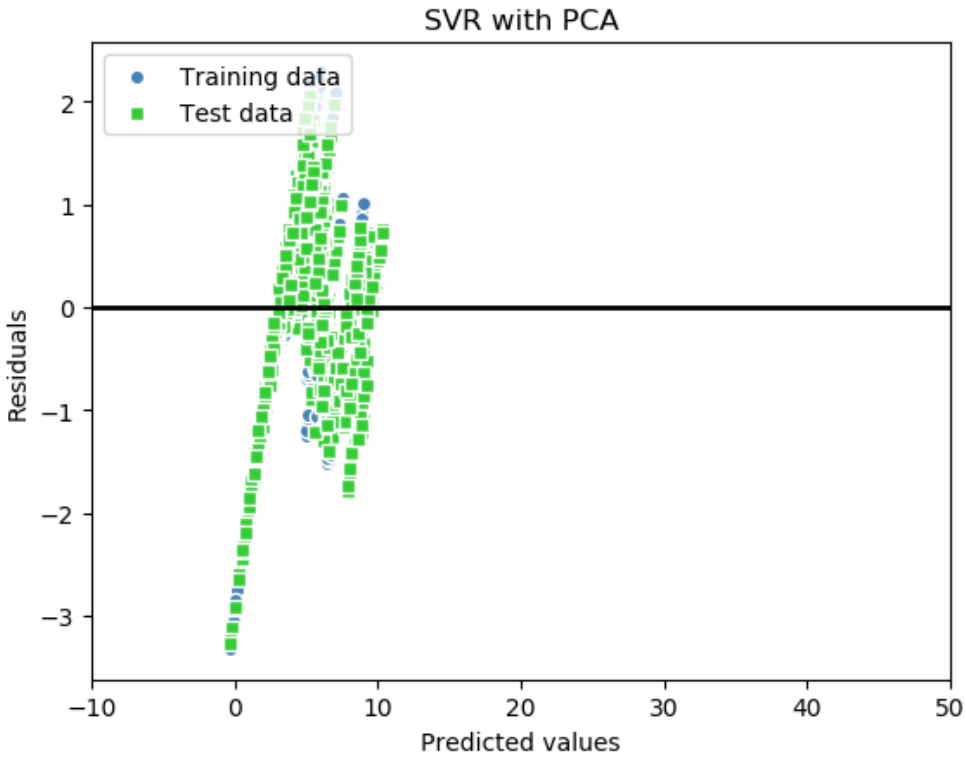


SVR with PCA

Basic Information:

```
Accuracy R2 score on training set: 0.8624757890247121
Accuracy R2 score on testing set: 0.8612014291991967
Average accuracy R2 score on training set using 5-fold cross-validation: 0.8620008317274213
Average accuracy R2 score on testing set using 5-fold cross-validation: 0.8584698835104285
Root Mean Squared Error of Train Set: 0.921318029573932
Root Mean Squared Error of Test Set: 0.9413806175273707
Slope:
[[-0.22866087 -0.29653499  0.03559914]]
Intercept: 5.442
```

Residual Errors Plot:



Part 4 Conclusions

Experiment 1 (Treasury Yields)		
	Linear	SVM
Baseline (all attributes)	Train Acc: 0.902273	Train Acc: 0.893323
	Test Acc: 0.904131	Test Acc: 0.894443
PCA transform (3 PCs)	Train Acc: 0.867389	Train Acc: 0.862476
	Test Acc: 0.866378	Test Acc: 0.861201

	Linear	SVM
Baseline (all attributes)	Train RMSE: 0.776653	Train RMSE: 0.811438
	Test RMSE: 0.782370	Test RMSE: 0.820949
PCA transform (3 PCs)	Train RMSE: 0.904712	Train RMSE: 0.921318
	Test RMSE: 0.923658	Test RMSE: 0.941381

According to the form, we can get two conclusions:

1. Linear Regression better explained the data than SVM with larger R^2 and smaller RMSE.
2. Baseline's regression better explained the data than PCA's regression. Maybe it is because some information will lose after the PCA transformation.

Part 5 Appendix

Link to my code:
https://github.com/fengzixue96/IE598_F19_HW5/blob/master/IE598_F19_HW5.py

The screenshot: