Assignment #6
Nitin Gaonkar

Introduction:
The purpose of this assignment is to build regression models for the home sale price, in this assignment I will be using dummy coding of categorical Variables and then build regression models, In the later part of the assignment I will be working on validation frameworks and validations.

Results:

Below is the output of the proc print:

Date	AA	BAC	BHI	CVX	DD	DOW	DPS	GS	HAL	HES	HON	HUN	JPM	KO	MMM	MPC	PEP	SLB	WFC	XOM	VV
03JAN2012	9.23	5.8	51.02	110.37	46.51	29.79	38.34	95.36	34.15	58.4	55.58	9.95	34.98	35.07	83.49	33.41	66.4	70.09	28.43	86	58.18
04JAN2012	9.45	5.81	51.53	110.18	47.02	29.95	38.55	94.74	35.12	59	55.53	9.87	34.95	34.85	84.18	33.76	66.74	69.56	28.56	86.02	58.25
05JAN2012	9.36	6.31	50.82	109.1	46.7	30.14	38.79	94.58	34.56	57.6	55.59	9.82	35.68	34.68	83.8	31.92	66.22	68.07	29.02	85.76	58.44
06JAN2012	9.16	6.18	51.26	108.31	46.04	30.32	38.52	93.42	34.98	56.42	55.18	9.9	35.36	34.46	83.37	31.66	65.39	67.78	28.94	85.12	58.32
09JAN2012	9.42	6.27	51.58	109.49	46.43	30.31	38.52	94.69	35.38	56.9	55.64	9.84	35.3	34.46	83.87	30.96	65.73	68.82	29.3	85.5	58.45
10JAN2012	9.44	6.63	51.95	109.06	47.14	30.98	38.62	98.33	36.33	58.56	56.58	10.2	36.05	34.67	84.3	31.83	65.66	70.75	29.41	85.72	58.99
11JAN2012	9.63	6.87	50.21	107.77	47.3	31.43	38.15	99.76	35.38	57.94	56.46	10.59	36.66	34.03	83.77	32.89	65.01	70.16	29.62	85.08	59.06
12JAN2012	9.93	6.79	48.29	104.97	48.1	32.56	37.96	101.21	34.73	57.01	57.19	11.03	36.85	33.78	84.28	32.64	64.62	69.7	29.61	84.74	59.2
13JAN2012	9.8	6.61	48.02	106.09	48.4	32.02	37.82	98.96	33.94	56.55	56.7	11.21	35.92	33.5	83.6	33.03	64.4	67.99	29.61	84.88	58.95
17JAN2012	9.76	6.48	47.7	106.72	48.54	32.64	37.63	97.68	33.86	57.39	57.16	10.85	34.91	33.68	84.23	33.54	64.65	67.64	29.82	85.69	59.1

Below is the out put for all the returns:

return_AA	return_BAC	response_VV
.	.	.
0.023556	0.001723	0.001202439
-0.009569	0.082555	0.003256494
-0.021599	-0.020817	-.002055499
0.027989	0.014458	0.002226600
0.002121	0.055828	0.009196250
0.019927	0.035559	0.001185938
0.030677	-0.011713	0.002367666
-0.013178	-0.026867	-.004231915
-0.004090	-0.019863	0.002541297

S	return_GS	return_HAL	return_HES	return_HON	return_HUN	return_JPM	return_KO	return_MMM	return_MPC	return_PEP	return_SLB	return_WFC	return_XOM	return_VV
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
2	-0.006523	0.028008	0.010222	-0.000900	-0.008073	-0.000858	-0.006293	0.008230499	0.010421	0.005107	-0.007590	0.004562	0.000232531	0.001202439
6	-0.001690	-0.016074	-0.024015	0.001080	-0.005079	0.020672	-0.004890	-.004524356	-0.056044	-0.007822	-0.021653	0.015978	-.003027130	0.003256494
5	-0.012341	0.012080	-0.020699	-0.007403	0.008114	-0.009009	-0.006364	-.005144475	-0.008179	-0.012613	-0.004269	-0.002761	-.007490672	-.002055499
0	0.013503	0.011370	0.008472	0.008302	-0.006079	-0.001698	0.000000	0.005979449	-0.022358	0.005186	0.015227	0.012363	0.004454350	0.002226600
3	0.037721	0.026497	0.028757	0.016753	0.035932	0.021024	0.006076	0.005113884	0.027713	-0.001066	0.027658	0.003747	0.002569795	0.009196250
5	0.014438	-0.026497	-0.010644	-0.002123	0.037522	0.016779	-0.018632	-.006306917	0.032759	-0.009949	-0.008374	0.007115	-.007494180	0.001185938
3	0.014430	-0.018543	-0.016181	0.012847	0.040709	0.005169	-0.007374	0.006069641	-0.007630	-0.006017	-0.006578	-0.000338	-.004004245	0.002367666
5	-0.022482	-0.023010	-0.008101	-0.008605	0.016187	-0.025561	-0.008323	-.008101069	0.011878	-0.003410	-0.024840	0.000000	0.001650749	-.004231915
6	-0.013019	-0.002360	0.014745	0.008080	-0.032641	-0.028521	0.005359	0.007507632	0.015323	0.003874	-0.005161	0.007067	0.009497638	0.002541297

2.

Below is the correlation between the individual stocks and the market index.

Pearson Correlation Coefficients, N = 501 Prob > |r| under H0: Rho=0													
	return_AA	return_BAC	return_BHI	return_CVX	return_DD	return_DOW	return_DPS	return_GS	return_HAL	return_HES	return_HON	return_HUN	return_JPM
response_VV	0.63241 <.0001	0.65019 <.0001	0.57750 <.0001	0.72090 <.0001	0.68952 <.0001	0.62645 <.0001	0.44350 <.0001	0.71216 <.0001	0.59750 <.0001	0.61080 <.0001	0.76838 <.0001	0.58194 <.0001	0.65785 <.0001

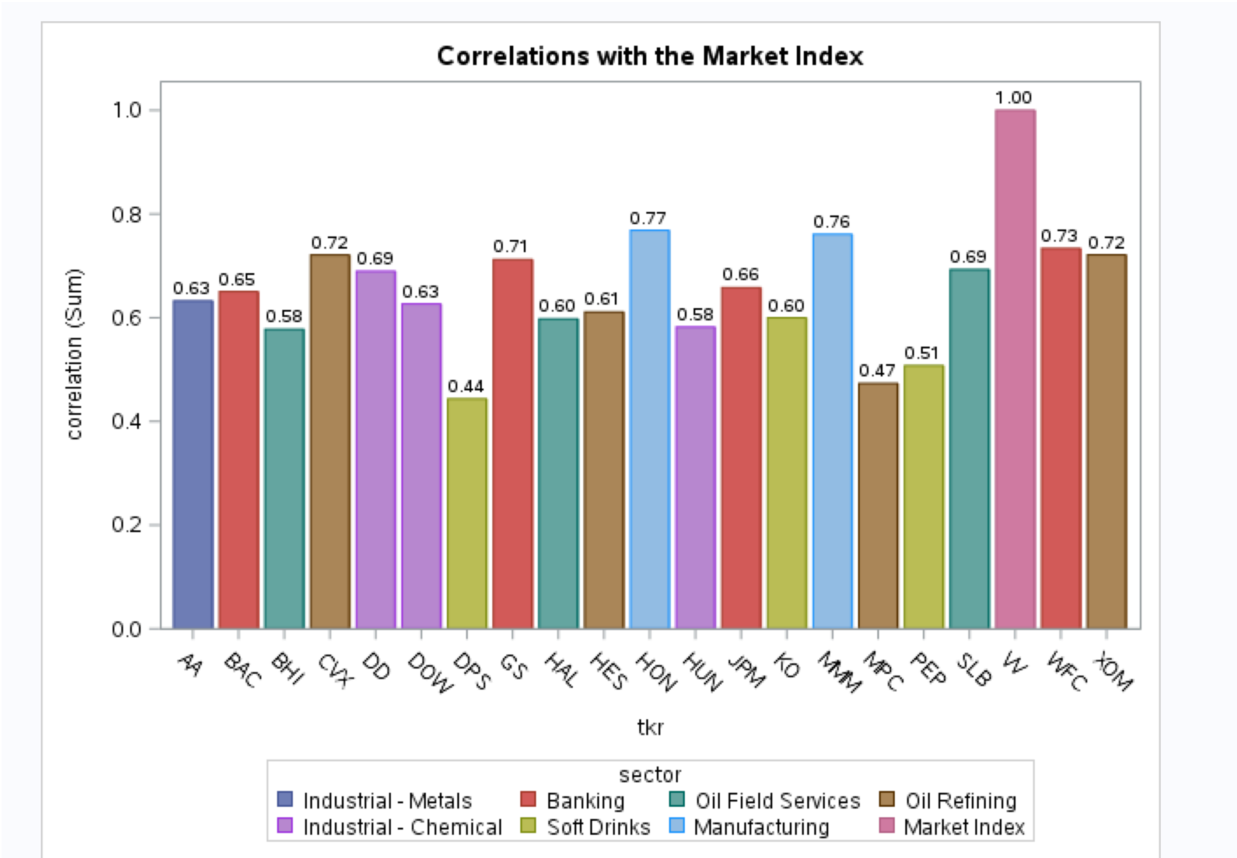
return_KO	return_MMM	return_MPC	return_PEP	return_SLB	return_WFC	return_XOM	return_VV
0.59980 <.0001	0.76085 <.0001	0.47312 <.0001	0.50753 <.0001	0.69285 <.0001	0.73357 <.0001	0.72111 <.0001	1.00000 <.0001

3. Below screen shot gives the observations which are converted from wide format to long format.

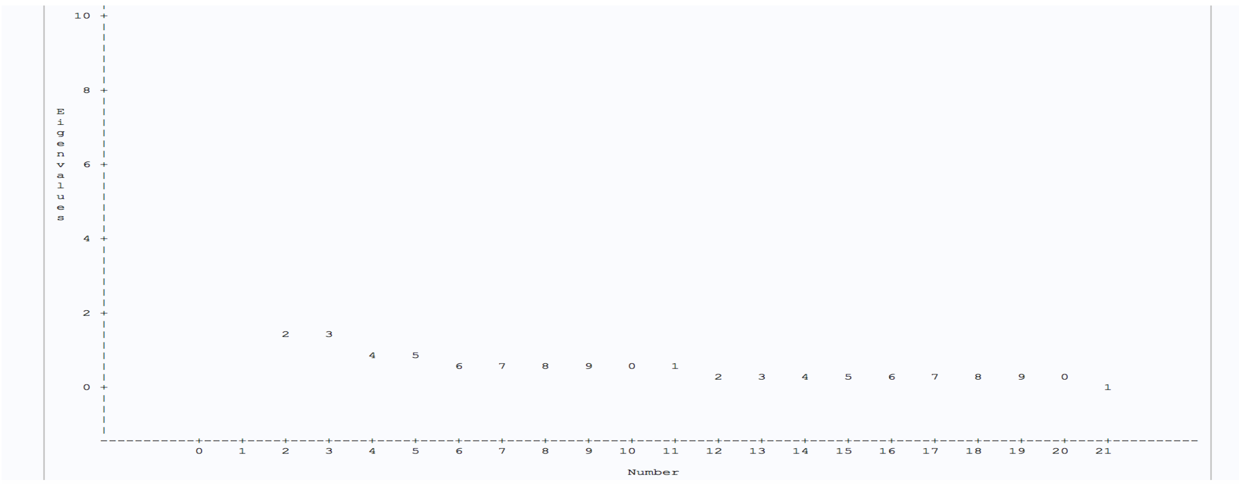
Obs	correlation	tkr
1	0.63241	AA
2	0.65019	BAC
3	0.57750	BHI
4	0.72090	CVX
5	0.68952	DD
6	0.62645	DOW
7	0.44350	DPS
8	0.71216	GS
9	0.59750	HAL
10	0.61080	HES
11	0.76838	HON
12	0.58194	HUN
13	0.65785	JPM
14	0.59980	KO
15	0.76085	MMM
16	0.47312	MPC
17	0.50753	PEP
18	0.69285	SLB
19	0.73357	WFC
20	0.72111	XOM
21	1.00000	VV

4.

Obs	correlation	tkr	sector
1	0.63241	AA	Industrial - Metals
2	0.65019	BAC	Banking
3	0.57750	BHI	Oil Field Services
4	0.72090	CVX	Oil Refining
5	0.68952	DD	Industrial - Chemical
6	0.62645	DOW	Industrial - Chemical
7	0.44350	DPS	Soft Drinks
8	0.71216	GS	Banking
9	0.59750	HAL	Oil Field Services
10	0.61080	HES	Oil Refining
11	0.76838	HON	Manufacturing
12	0.58194	HUN	Industrial - Chemical
13	0.65785	JPM	Banking
14	0.59980	KO	Soft Drinks
15	0.76085	MMM	Manufacturing
16	0.47312	MPC	Oil Refining
17	0.50753	PEP	Soft Drinks
18	0.69285	SLB	Oil Field Services
19	1.00000	VV	Market Index
20	0.73357	WFC	Banking
21	0.72111	XOM	Oil Refining

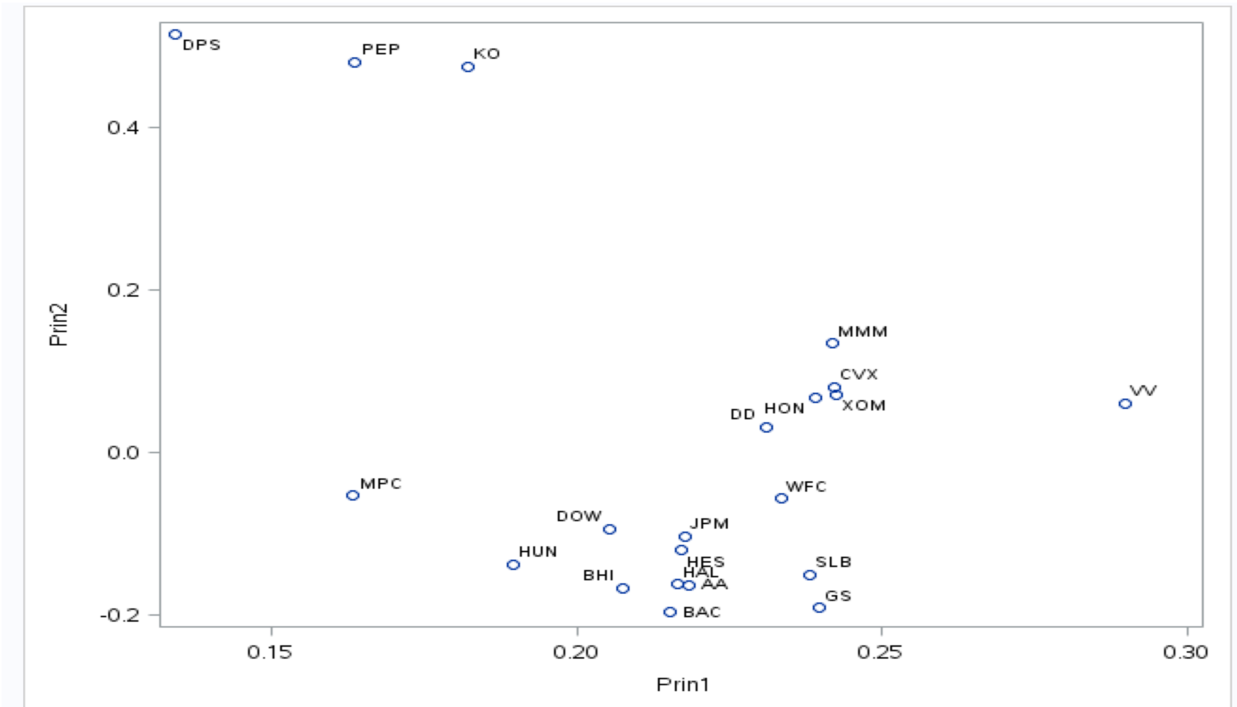

Correlations with the Market Index
correlation (Sum)
1.0
0.8
0.6
0.4
0.2
0.0
0.63
0.65
0.58
0.72
0.69
0.63
0.44
0.71
0.60
0.61
0.77
0.58
0.66
0.60
0.76
0.47
0.51
0.69
1.00
0.73
0.72
AA
BAC
BHI
CVX
DD
DOW
DPS
GS
HAL
HES
HON
HUN
JPM
KO
MMM
MPC
PEP
SLB
W
WFC
XOM
tkr
sector
Industrial - Metals
Banking
Oil Field Services
Oil Refining
Industrial - Chemical
Soft Drinks
Manufacturing
Market Index

5. from the below two plots we can see conclude that we can keep three principle component, as we can see from the eigen value chart we can see that the first three components have value more than one.



Eigenvalues of the Correlation Matrix: Total = 21 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	10.5071587	8.9629547	0.5003	0.5003
2	1.5442040	0.1910305	0.0735	0.5739
3	1.3531735	0.4050880	0.0644	0.6383
4	0.9480855	0.1520940	0.0451	0.6835
5	0.7959915	0.1294957	0.0379	0.7214
6	0.6664958	0.1080561	0.0317	0.7531
7	0.5584397	0.0440820	0.0266	0.7797
8	0.5143577	0.0156109	0.0245	0.8042
9	0.4987468	0.0316307	0.0237	0.8279
10	0.4671161	0.0305970	0.0222	0.8502
11	0.4365191	0.0284333	0.0208	0.8710
12	0.4080858	0.0569741	0.0194	0.8904
13	0.3511117	0.0146788	0.0167	0.9071
14	0.3364328	0.0394074	0.0160	0.9231
15	0.2970254	0.0196849	0.0141	0.9373
16	0.2773405	0.0179758	0.0132	0.9505
17	0.2593647	0.0168429	0.0124	0.9628
18	0.2425218	0.0206015	0.0115	0.9744
19	0.2219203	0.0101847	0.0106	0.9850
20	0.2117355	0.1075627	0.0101	0.9950
21	0.1041729		0.0050	1.0000

Below is the plot of the first two Eigen vectors:
I can see a two groups in the below chart. One with the



Factor Pattern			
	Factor1	Factor2	Factor3
return_AA	0.70767	-0.20292	0.04160
return_BAC	0.69734	-0.24415	-0.39592
return_BHI	0.67292	-0.20705	0.50882
return_CVX	0.78479	0.10023	0.20409
return_DD	0.74868	0.03957	-0.13376
return_DOW	0.66587	-0.11617	-0.19890
return_DPS	0.43475	0.64054	0.04151
return_GS	0.77701	-0.23805	-0.21989
return_HAL	0.70178	-0.20150	0.46725
return_HES	0.70405	-0.14941	0.26417
return_HON	0.77459	0.08484	-0.09359
return_HUN	0.61397	-0.17155	-0.21396
return_JPM	0.70585	-0.12782	-0.39946
return_KO	0.58985	0.59007	-0.01314
return_MMM	0.78390	0.16861	-0.00446
return_MPC	0.52907	-0.06590	0.00812
return_PEP	0.53024	0.59834	0.00278
return_SLB	0.77197	-0.18746	0.37902
return_WFC	0.75677	-0.07034	-0.32691
return_XOM	0.78579	0.08834	0.16749
return_VV	0.93901	0.07571	-0.07446

Variance Explained by Each Factor		
Factor1	Factor2	Factor3
10.507159	1.544204	1.353173

6.

8	Prin9	Prin10	Prin11	Prin12	Prin13	Prin14	Prin15	Prin16	Prin17	Prin18	Prin19	Prin20	Prin21	response_VV	u	train	train_response
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0.75040	0	.
6	-0.58438	0.98252	0.74412	0.35496	0.32373	-1.07575	0.32008	-0.15545	-0.05598	1.21910	0.73499	-0.46395	-0.11466	0.001202439	0.32091	1	0.001202439
2	0.93577	0.12189	-0.73825	0.45978	0.18552	0.93913	-1.39075	0.96007	0.38546	1.64363	-0.69380	-0.57537	-0.29046	0.003256494	0.17839	1	0.003256494
9	0.11672	0.18478	-0.92199	-0.04698	0.38032	1.05000	0.46500	0.25373	-0.26238	0.42693	0.38827	-0.32458	-0.34535	-.002055499	0.90603	0	.
6	-0.23038	-0.26760	0.13941	0.44528	0.37540	-0.55137	-0.37313	-0.21072	-0.19495	-0.64032	0.18602	-0.23415	0.39092	0.002226600	0.35712	1	0.002226600
7	0.68271	-0.12135	0.67252	0.18293	-1.51022	-0.29235	-0.93674	0.27894	-0.35837	-0.19753	-0.76284	-0.32760	0.17658	0.009196250	0.22111	1	0.009196250
0	-0.41563	0.22667	0.39715	0.13389	-0.18985	0.14976	-0.50015	0.12501	-0.45620	-0.37512	0.14085	0.53250	-0.25377	0.001185938	0.78644	0	.
5	-0.88791	0.06860	0.95892	0.80369	-1.17818	-0.35309	1.10458	-0.29630	-1.46978	-0.60197	-0.18047	1.07616	-0.08271	0.002367666	0.39808	1	0.002367666
8	0.20820	-0.39350	-0.47433	-1.41112	1.79423	0.09710	-0.20366	-0.64805	0.21376	0.10883	-0.32231	-0.51949	-0.19579	-.004231915	0.12467	1	-.004231915
3	-0.54238	0.12804	-0.28981	1.21742	0.24114	-0.82071	0.15515	0.50033	-0.45034	0.32789	-0.03422	-0.54516	-0.11285	0.002541297	0.18769	1	0.002541297

7.

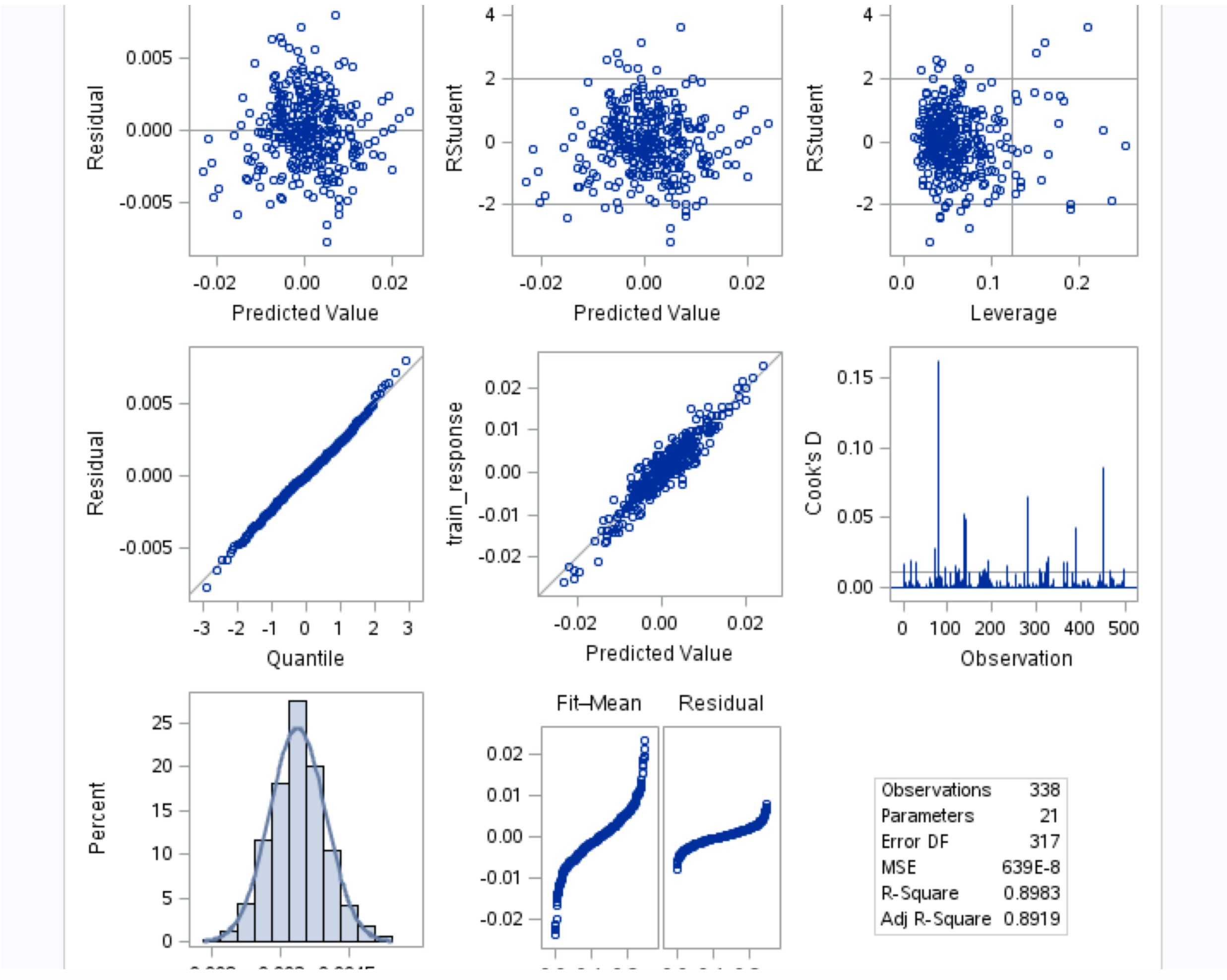
The REG Procedure
Model: MODEL1
Dependent Variable: train_response

Number of Observations Read	502
Number of Observations Used	338
Number of Observations with Missing Values	164

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	20	0.01790	0.00089510	140.04	<.0001
Error	317	0.00203	0.00000639		
Corrected Total	337	0.01993			

Root MSE	0.00253	R-Square	0.8983
Dependent Mean	0.00061635	Adj R-Sq	0.8919
Coeff Var	410.18453		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.00008640	0.00014092	0.61	0.5403	0
return_AA	1	0.01769	0.01317	1.34	0.1802	2.11490
return_BAC	1	0.03198	0.01165	2.75	0.0064	3.10927
return_BHI	1	-0.00111	0.01323	-0.08	0.9333	2.62997
return_CVX	1	0.04907	0.02536	1.93	0.0539	3.07524
return_DD	1	0.04674	0.02037	2.29	0.0224	2.51406
return_DOW	1	0.03642	0.01162	3.14	0.0019	1.88893
return_DPS	1	0.03670	0.01679	2.19	0.0295	1.54768
return_GS	1	0.04849	0.01555	3.12	0.0020	3.10450
return_HAL	1	0.00948	0.01466	0.65	0.5184	3.08758
return_HES	1	0.00359	0.01092	0.33	0.7425	2.10199
return_HON	1	0.12213	0.01924	6.35	<.0001	2.73505
return_HUN	1	0.02712	0.00836	3.24	0.0013	1.79852
return_JPM	1	0.00902	0.01708	0.53	0.5979	3.36439
return_KO	1	0.07903	0.02226	3.55	0.0004	1.93633
return_MMM	1	0.09796	0.02646	3.70	0.0003	2.98277
return_MPC	1	0.01673	0.00809	2.07	0.0394	1.32999
return_PEP	1	0.02911	0.02231	1.30	0.1929	1.68825
return_SLB	1	0.03776	0.01709	2.21	0.0279	3.13690
return_WFC	1	0.07587	0.01848	4.10	<.0001	2.59492
return_XOM	1	0.05467	0.02697	2.03	0.0435	2.98393



From the above ODS outputs we can see that the fit mean is greater than the residual which is good, the SAS curve covers all the point on the histogram, there are few outliers on the cook's D. The QQ plots looks good all the points are on the line and the Square value is 0.8983 which is good.

In our model the VIF value are not high and looks Pretty ok, thus the multicollinearity does not exist.

For training:

MAE calculation run

The MEANS Procedure

Analysis Variable : mae				
N	Mean	Std Dev	Minimum	Maximum
338	0.0019020	0.0015440	6.5094738E-7	0.0080258

MSE: 639 E-8

For test:

MAE calculation

The MEANS Procedure

Analysis Variable : mae				
N	Mean	Std Dev	Minimum	Maximum
163	0.0021449	0.0021758	0.000036770	0.0151721

MSE calculation

The MEANS Procedure

Analysis Variable : MSE				
N	Mean	Std Dev	Minimum	Maximum
163	5.7091237E-8	1.6646024E-7	8.294459E-12	1.41222E-6

8.
Regression with eight principle components using train response as a response variables.

Dependent Variable: train_response

Number of Observations Read	502
Number of Observations Used	338
Number of Observations with Missing Values	164

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	0.01818	0.00227	427.71	<.0001
Error	329	0.00175	0.00000531		
Corrected Total	337	0.01993			

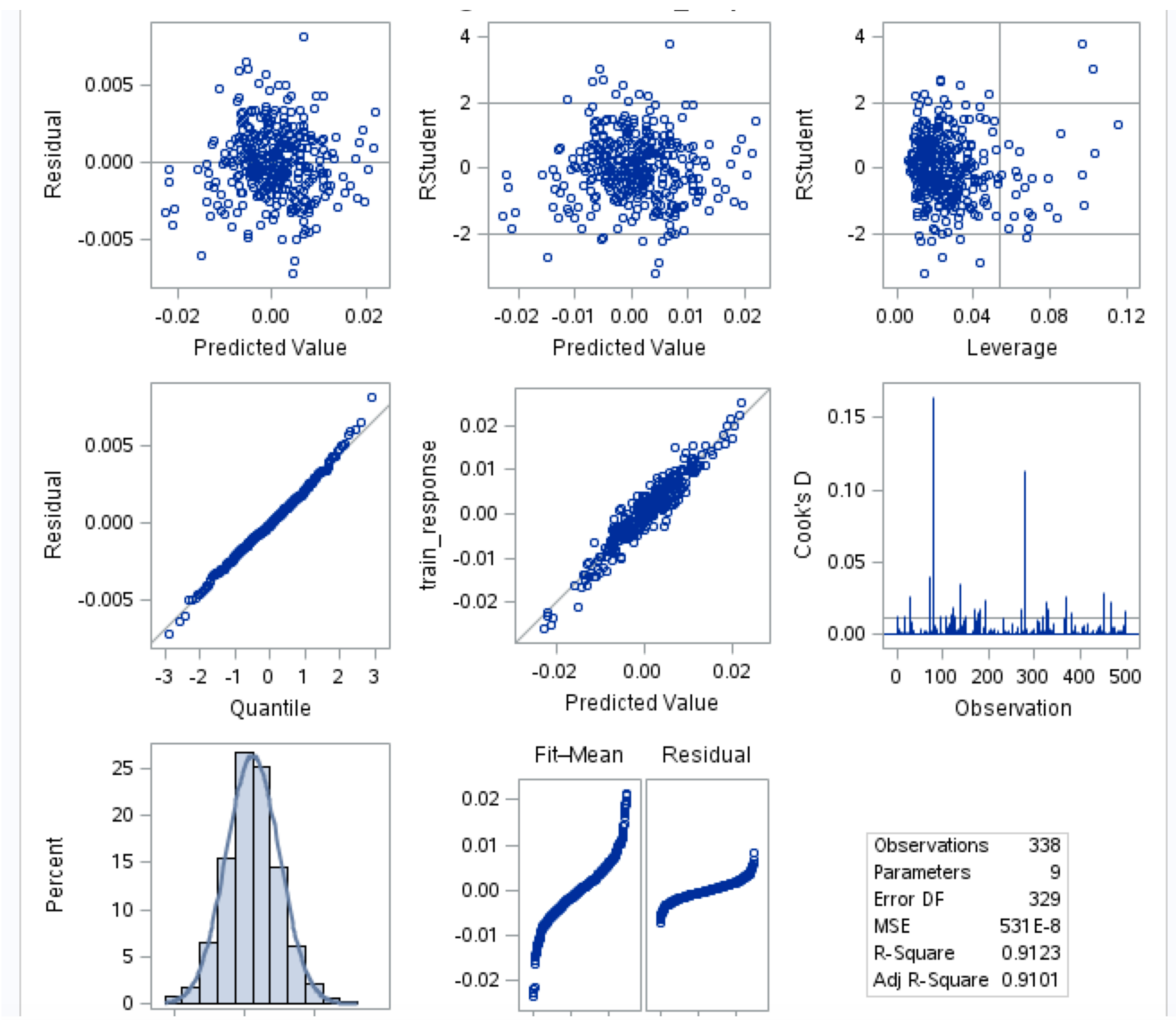
Root MSE	0.00231	R-Square	0.9123
Dependent Mean	0.00061635	Adj R-Sq	0.9101
Coeff Var	373.98445		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.00074905	0.00012615	5.94	<.0001	0
Prin1	1	0.00223	0.00003881	57.47	<.0001	1.00563
Prin2	1	0.00034527	0.00010245	3.37	0.0008	1.00816
Prin3	1	-0.00066667	0.00011041	-6.04	<.0001	1.00938
Prin4	1	0.00029272	0.00013047	2.24	0.0255	1.00632
Prin5	1	-0.00017303	0.00013933	-1.24	0.2152	1.00300
Prin6	1	0.00001816	0.00015367	0.12	0.9060	1.00743
Prin7	1	-0.00012822	0.00016697	-0.77	0.4431	1.02104
Prin8	1	-0.00054701	0.00017987	-3.04	0.0025	1.02030

Below are the ODS output for the model:

From the below ODS outputs we can see that the fit mean is greater than the residual which is good, the SAS curve covers all the point on the histogram, there are few outliers on the cook's D. The QQ plots looks good all the points are on the line and the Square value is 0.9123 which is good.

By observing the VIF values for the model we can say that the Multicollinearity does not exist.



Training set:

Mean square error 531 e-8

MAE calculation

The MEANS Procedure

Analysis Variable : mae				
N	Mean	Std Dev	Minimum	Maximum
338	0.0017738	0.0014252	4.817129E-7	0.0081709

Test set:

MSE calculation

The MEANS Procedure

Analysis Variable : MSE				
N	Mean	Std Dev	Minimum	Maximum
163	4.8608955E-8	1.399492E-7	2.566059E-14	1.1710607E-6

MAE calculation

The MEANS Procedure

Analysis Variable : mae				
N	Mean	Std Dev	Minimum	Maximum
163	0.0019575	0.0020290	2.0451594E-6	0.0138160

By comparing both the models Square value and other ODS outputs and by looking at the variance inflation factors, we can say that the model with the principal components is better than the previous model.

Conclusion:

Overall in this assignment I have tried multi regression using Principal Components Analysis as a method of dimension reduction and as a remedial measure for multicollinearity in Ordinary Least Squares regression.

Code:
Paste your code in at the end.

```
libname mydata "/scs/wtm926/" access=readonly;
proc datasets library=mydata;
run;
 quit;


data temp;
set mydata.stock_portfolio_data;
run;


proc sort data=temp;
by date;
run;
quit;
data temp;
set temp;
* Compute the log-returns - log of the ratio of today's price to yesterday's price;
* Note that the data needs to be sorted in the correct direction in order for us to compute
the correct return; return_AA = log(AA/lag1(AA));
return_BAC = log(BAC/lag1(BAC));
* Continue to compute the log-returns for all of the stocks;
* Name the log-return for VV as the response variable;
```

```
response_VV = log(VV/lag1(VV));

*proc print data=temp(obs=10);
* run;
*quit;




return_AA = log(AA/lag1(AA));
return_BAC = log(BAC/lag1(BAC));
return_BHI = log(BHI/lag1(BHI));
return_CVX = log(CVX/lag1(CVX));
return_DD = log(DD/lag1(DD));
return_DOW = log(DOW/lag1(DOW));
return_DPS = log(DPS/lag1(DPS));
return_GS = log(GS/lag1(GS));
return_HAL = log(HAL/lag1(HAL));
return_HES = log(HES/lag1(HES));
return_HON = log(HON/lag1(HON));
return_HUN = log(HUN/lag1(HUN));
return_JPM = log(JPM/lag1(JPM));
return_KO = log(KO/lag1(KO));
return_MMM = log(MMM/lag1(MMM));
return_MPC = log(MPC/lag1(MPC));
return_PEP = log(PEP/lag1(PEP));
return_SLB= log(SLB/lag1(SLB));
return_WFC = log(WFC/lag1(WFC));
return_XOM = log(XOM/lag1(XOM));
return_VV = log(VV/lag1(VV));
run;
proc print data=temp(obs=10);
run;
quit;
```

2.

```
* We can use ODS TRACE to print out all of the data sets
available to ODS for a particular SAS procedure.; * We can
also look these data sets up in the SAS User's Guide in the
chapter for the selected procedure.;
```

```
*ods trace on;

ods output PearsonCorr=portfolio_correlations; proc corr
data=temp; *var return: with response_VV; var return_:;

with response_VV; run; quit; *ods trace off;

proc print data=portfolio_correlations; run; quit;
```

3.

```
data wide_correlations; set portfolio_correlations
(keep=return_:); run;

* Note that wide_correlations is a 'wide' data set and we
need a 'long' data set; * We can use PROC TRANSPOSE to
convert data from one format to the other;

proc transpose data=wide_correlations out=long_correlations; run; quit;

data long_correlations; set long_correlations; tkr =
substr(_NAME_,8,3); drop _NAME_;

rename COL1=correlation; run;

proc print data=long_correlations; run; quit;
```

4.

```
* Merge on sector id and make a colored bar plot;

data sector; input tkr $ 1-3 sector $ 4-35; datalines; AA
Industrial - Metals BAC Banking BHI Oil Field Services CVX
Oil Refining DD Industrial - Chemical DOW Industrial -
Chemical DPS Soft Drinks GS Banking HAL Oil Field
Services HES Oil Refining HON Manufacturing HUN Industrial -
Chemical JPM Banking KO Soft Drinks MMM Manufacturing MPC
Oil Refining PEP Soft Drinks SLB Oil Field Services WFC
Banking XOM Oil Refining VV Market Index ; run;
```

```
proc print data=sector; run; quit; proc sort data=sector; by tkr; run;

proc sort data=long_correlations; by tkr; run;

data long_correlations; merge long_correlations (in=a) sector
(in=b); by tkr; if (a=1) and (b=1); run;

proc print data=long_correlations; run; quit;

* Make Grouped Bar Plot; * p. 48 Statistical Graphics
Procedures By Example; ods graphics on; title 'Correlations
with the Market Index'; proc sgplot data=long_correlations; format
correlation 3.2; vbar tkr / response=correlation group=sector
groupdisplay=cluster datalabel; run; quit; ods graphics off;
```

5.

```
ods graphics on;
proc princomp data=return_data out=pca_output
outstat=eigenvectors plots=scree(unpackpanel);
run;
quit;
ods graphics off;
* Notice that PROC PRINCOMP produces a lot of output;
* How many principal components should we keep?;
* Do the principal components have any interpretability?;
* Can we display that interpretability using graphics?;
proc print data=pca_output(obs=10); run;
proc print data=eigenvectors(where=(_TYPE_='SCORE')); run;
* Display the two plots and the Eigenvalue table from the output;
* Plot the first two eigenvectors;
data pca2;
set eigenvectors(where=(_NAME_ in ('Prin1','Prin2'))); drop _TYPE_ ;
run;
proc print data=pca2; run;
proc transpose data=pca2 out=long_pca; run; quit;
proc print data=long_pca; run;

data long_pca;
set long_pca;
format tkr $3.;
tkr = substr(_NAME_,8,3);
drop _NAME_;
run;
```

```
proc print data=long_pca; run;

* Plot the first two principal components;
ods graphics on;
proc sgplot data=long_pca;
scatter x=Prin1 y=Prin2 / datalabel=tkr; run; quit;
ods graphics off;


proc factor data=return_data method=prin scree ;
var return_: ;
run;
```

6.

```
data cv_data;
merge pca_output temp(keep=response_VV);   * No BY statement needed here. We are
going to append a column in its current order;   * generate a uniform(0,1) random variable
with seed set to   123; u = uniform(123);   if (u < 0.70) then train = 1;   else train = 0;
if (train=1) then train_response=response_VV; else train_response=.;   run;
proc print data=cv_data(obs=10); run;
```

7.

```
proc reg data=cv_data;
model train_response= return_AA return_BAC return_BHI return_CVX return_DD
return_DOW return_DPS return_GS return_HAL return_HES return_HON return_HUN
return_JPM return_KO return_MMM return_MPC return_PEP return_SLB
return_WFC return_XOM /
vif ;
output out =new_data predicted=yhat;
run;

proc print data=new_data (obs=10);

proc reg data=cv_data;

model train_response= return_AA return_BAC return_BHI return_CVX return_DD
return_DOW return_DPS return_GS return_HAL return_HES return_HON return_HUN
return_JPM return_KO return_MMM return_MPC return_PEP return_SLB
return_WFC return_XOM /
```

```
vif ;
output out =new_data predicted=yhat;
run;


data test1;
set new_data ;
where train=0;
MSE=(1/163) *((yhat-response_vv)* (yhat-response_vv));
run;

proc means data=test1;
var MSE;
title 'MSE calculation';
run;
```

8.

```
proc reg data=cv_data;
model train_response= prin1 prin2 prin3 prin4 prin5 prin6 prin7 prin8 /
vif ;
output out =new_data predicted=yhat;
run;

data test1;
set new_data ;
where train=0;
MSE=(1/163) *((yhat-response_vv)* (yhat-response_vv));
run;

proc means data=test1;
var MSE;
title 'MSE calculation';
run;
```