Assignment #7
Nitin Gaonkar

Introduction:
The purpose of this assignment is use factor analysis to identify the sector in the stock market. I will be using factor analysis and factor rotation in this assignment.

Results:

Below is the output of the proc print:

Obs	Date	BAC	BHI	CVX	DD	DOW	HAL	HES	HUN	JPM	SLB	WFC	XOM	VV
1	03JAN2012	5.8	51.02	110.37	46.51	29.79	34.15	58.4	9.95	34.98	70.09	28.43	86	58.18
2	04JAN2012	5.81	51.53	110.18	47.02	29.95	35.12	59	9.87	34.95	69.56	28.56	86.02	58.25
3	05JAN2012	6.31	50.82	109.1	46.7	30.14	34.56	57.6	9.82	35.68	68.07	29.02	85.76	58.44
4	06JAN2012	6.18	51.26	108.31	46.04	30.32	34.98	56.42	9.9	35.36	67.78	28.94	85.12	58.32
5	09JAN2012	6.27	51.58	109.49	46.43	30.31	35.38	56.9	9.84	35.3	68.82	29.3	85.5	58.45
6	10JAN2012	6.63	51.95	109.06	47.14	30.98	36.33	58.56	10.2	36.05	70.75	29.41	85.72	58.99
7	11JAN2012	6.87	50.21	107.77	47.3	31.43	35.38	57.94	10.59	36.66	70.16	29.62	85.08	59.06
8	12JAN2012	6.79	48.29	104.97	48.1	32.56	34.73	57.01	11.03	36.85	69.7	29.61	84.74	59.2
9	13JAN2012	6.61	48.02	106.09	48.4	32.02	33.94	56.55	11.21	35.92	67.99	29.61	84.88	58.95
10	17JAN2012	6.48	47.7	106.72	48.54	32.64	33.86	57.39	10.85	34.91	67.64	29.82	85.69	59.1

Below is the out put for all the returns after using the keep option:

return_BAC	return_BHI	return_CVX	return_DD	return_DOW	return_HAL	return_HES	return_HUN	return_JPM	return_SLB	return_WFC	return_XOM	response_VV
.	.	.	.	.	.	.	.	.	.	.	.	.
0.001723	0.009946	-0.001723	0.010906	0.005357	0.028008	0.010222	-0.008073	-0.000858	-0.007590	0.004562	0.000232531	0.001202439
0.082555	-0.013874	-0.009850	-0.006829	0.006324	-0.016074	-0.024015	-0.005079	0.020672	-0.021653	0.015978	-.003027130	0.003256494
-0.020817	0.008621	-0.007267	-0.014234	0.005954	0.012080	-0.020699	0.008114	-0.009009	-0.004269	-0.002761	-.007490672	-.002055499
0.014458	0.006223	0.010836	0.008435	-0.000330	0.011370	0.008472	-0.006079	-0.001698	0.015227	0.012363	0.004454350	0.002226600
0.055828	0.007148	-0.003935	0.015176	0.021864	0.026497	0.028757	0.035932	0.021024	0.027658	0.003747	0.002569795	0.009196250
0.035559	-0.034068	-0.011899	0.003388	0.014421	-0.026497	-0.010644	0.037522	0.016779	-0.008374	0.007115	-.007494180	0.001185938
-0.011713	-0.038990	-0.026325	0.016772	0.035322	-0.018543	-0.016181	0.040709	0.005169	-0.006578	-0.000338	-.004004245	0.002367666
-0.026867	-0.005607	0.010613	0.006218	-0.016724	-0.023010	-0.008101	0.016187	-0.025561	-0.024840	0.000000	0.001650749	-.004231915
-0.019863	-0.006686	0.005921	0.002888	0.019178	-0.002360	0.014745	-0.032641	-0.028521	-0.005161	0.007067	0.009497638	0.002541297

2.

Here we are doing the factor analysis without the rotation:

- Two factors have been retained by SAS.
- SAS used proportion criterion to select the number of factors to retain.
- In the scree plot we can see that the variance is explained and from the output we can see that the factor 1 is greater than the factor 2.
- Our initial hypothesis was about 3 to 4 factors, but the result that we got is 2 factors.
  Below are the two groups based on the sign of the second factor, but the factor 1 are more highly related to the variables.

  The first appears to combination of both the Banking and industrial sectors (BAC, DD, DOW, HUN, JPM, WFC), and the second appears to be related to Oil refining and field services sectors (BHI, CVX, HAL, HES, SLB, XOM).

  Below are the eigenvalues of the reduced correlation matrix:

The FACTOR Procedure

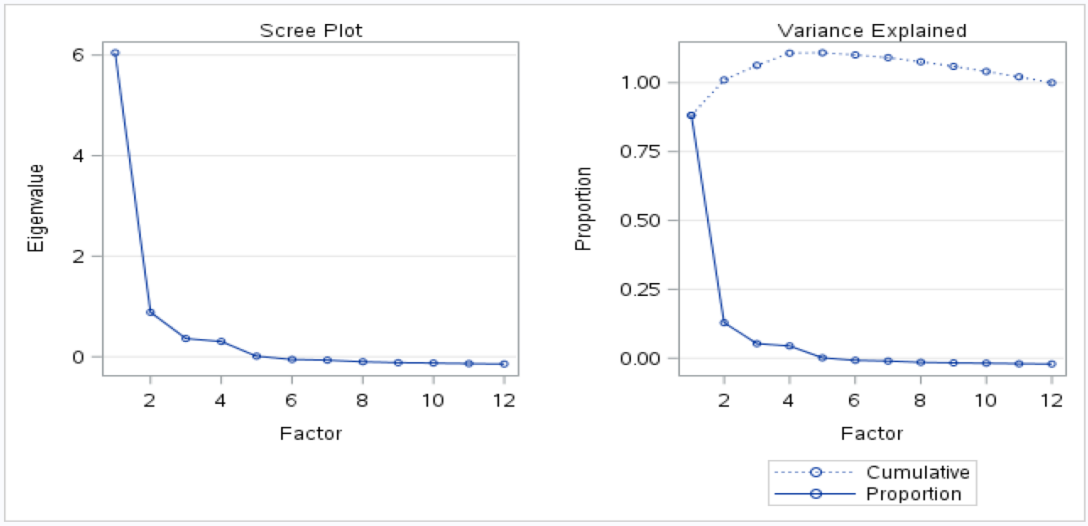
Input Data Type	Raw Data
Number of Records Read	502
Number of Records Used	501
N for Significance Tests	501

The FACTOR Procedure
Initial Factor Method: Principal Factors

Prior Communality Estimates: SMC											
return_BAC	return_BHI	return_CVX	return_DD	return_DOW	return_HAL	return_HES	return_HUN	return_JPM	return_SLB	return_WFC	return_XOM
0.58577906	0.61046627	0.64179539	0.54681402	0.47197670	0.64986770	0.49976057	0.39225125	0.58034671	0.68269067	0.57372531	0.62696933

Eigenvalues of the Reduced Correlation Matrix: Total = 6.86244298 Average = 0.57187025				
	Eigenvalue	Difference	Proportion	Cumulative
1	6.04732583	5.16261770	0.8812	0.8812
2	0.88470813	0.52262870	0.1289	1.0101
3	0.36207942	0.05735386	0.0528	1.0629
4	0.30472556	0.29429115	0.0444	1.1073
5	0.01043441	0.06365245	0.0015	1.1088
6	-.05321803	0.01517115	-0.0078	1.1011
7	-.06838918	0.03291807	-0.0100	1.0911
8	-.10130725	0.01600696	-0.0148	1.0763
9	-.11731422	0.00866270	-0.0171	1.0593
10	-.12597692	0.01040221	-0.0184	1.0409
11	-.13637913	0.00786652	-0.0199	1.0210
12	-.14424565		-0.0210	1.0000

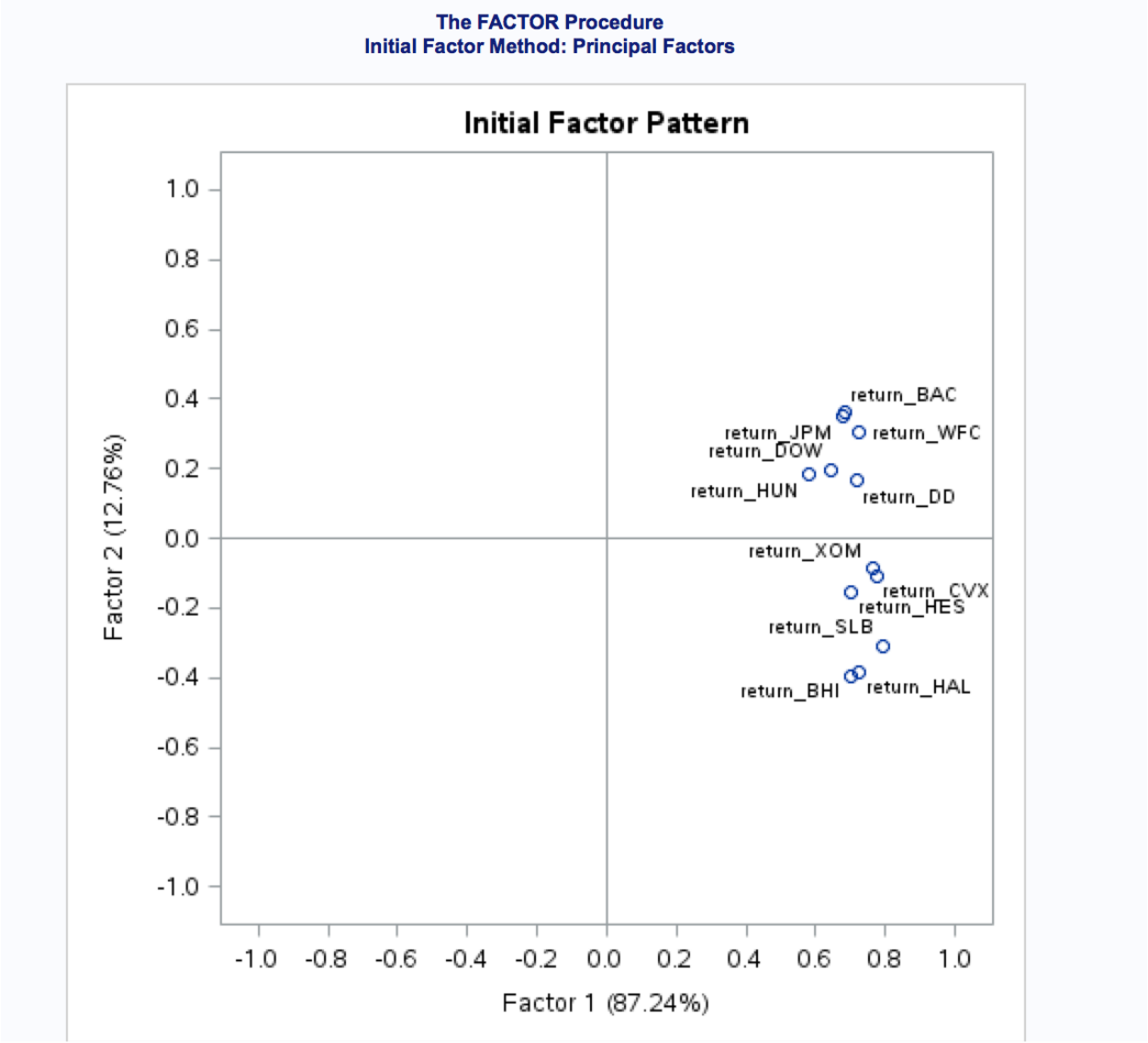
2 factors will be retained by the PROPORTION criterion.



Factor Pattern		
	Factor1	Factor2
return_BAC	0.68475	0.36021
return_BHI	0.69984	-0.39498
return_CVX	0.77402	-0.10833
return_DD	0.71605	0.16703
return_DOW	0.64548	0.19801
return_HAL	0.72630	-0.38221
return_HES	0.70361	-0.15709
return_HUN	0.58030	0.18186
return_JPM	0.67874	0.34813
return_SLB	0.79382	-0.30815
return_WFC	0.72445	0.30517
return_XOM	0.76500	-0.08361

Variance Explained by Each Factor	
Factor1	Factor2
6.0473258	0.8847081

Below we can see factor 1 and factor 2 plot:
We can see that the points are grouped into two parts, as we can see from the below graph the factor 1 has variables all above the loading threshold whereas the factor 2 do not have the all the variables above the threshold. Below graph does give us enough interpretability.



3.

Now we perform factor analysis with rotation, the rotation is meant to improve the interpretability of the model. By default, the unrotated output maximizes the variance accounted for by the first and subsequent factors, and forces the factors to be orthogonal. [WIKI]

- Yes SAS retained same no of factors.
- from the below out put we can see that the factor loadings for both factor one and factor two have changed for the variables after rotation.
- Yes we did obtain a simple structure and the interpretability has increased after the rotation, we can see the below pattern:

Factor 1 is comprised of BAC, DD, DOW, HUN, JPM, WFC.

Factor 2 is comprised of BHI, CVX, HAL, HES, SLB, XOM.

Factor 1 is comprised of Banking and industrial sectors, where as
Factor 2 is comprised of Oil refining and field services sectors.

-And in the rotated factor pattern we can see that all the values are positive and fall under one quadrant.

The FACTOR Procedure
Rotation Method: Varimax

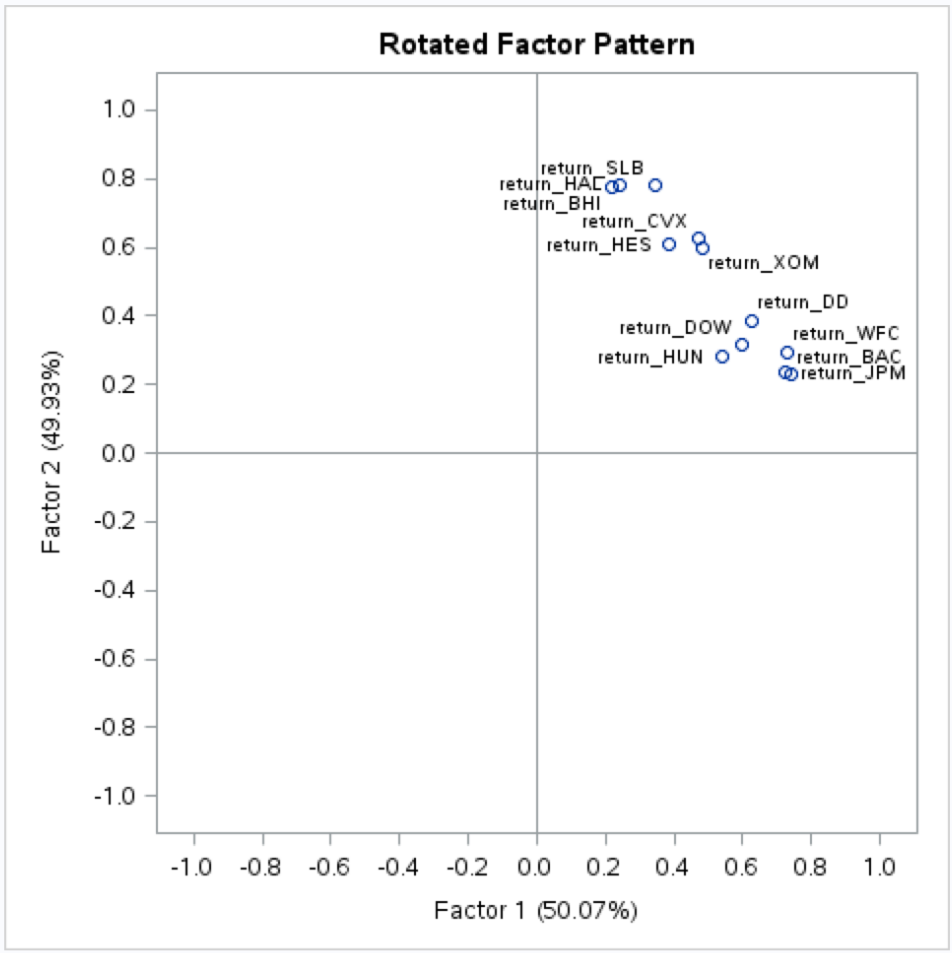
Orthogonal Transformation Matrix		
	1	2
1	0.70781	0.70640
2	0.70640	-0.70781

Rotated Factor Pattern		
	Factor1	Factor2
return_BAC	0.73912	0.22875
return_BHI	0.21634	0.77394
return_CVX	0.47133	0.62344
return_DD	0.62482	0.38759
return_DOW	0.59675	0.31582
return_HAL	0.24408	0.78359
return_HES	0.38705	0.60822
return_HUN	0.53921	0.28120
return_JPM	0.72634	0.23305
return_SLB	0.34419	0.77886
return_WFC	0.72835	0.29575
return_XOM	0.48241	0.59958

Variance Explained by Each Factor	
Factor1	Factor2
3.4711423	3.4608916

Final Communality Estimates: Total = 6.932034											
return_BAC	return_BHI	return_CVX	return_DD	return_DOW	return_HAL	return_HES	return_HUN	return_JPM	return_SLB	return_WFC	return_XOM
0.59863104	0.64577915	0.61083713	0.54062043	0.45584934	0.67359085	0.51974549	0.36982204	0.58188382	0.72509913	0.61795857	0.59221697

The FACTOR Procedure
Rotation Method: Varimax



4.

In this part we will be performing a maximum likelihood estimation to the estimate the common factors with a varimax rotation.

Two factors are retained by the ML process.

We observe that the method computes an initial set of eigenvalues to assess the convergence criterion. Below output displays the results of the analysis with 2 factor. The solution is so close to the optimum that PROC FACTOR cannot find a better solution, hence we receive this message:

```
Convergence criterion satisfied.
```

Below we observe that we see that there is two separate statistical hypothesis test with the null hypothesis stated as 'no common factors' and '2 factors are sufficient' respectively and both test allow us to accept the null hypothesis.

ML produces more generalizable and reproducible results, as it does not inflate the variance estimates when compared to PCA, also in maximum likelihood factor analysis is the goodness-of-fit criteria

The FACTOR Procedure
Initial Factor Method: Maximum Likelihood

Prior Communality Estimates: SMC											
return_BAC	return_BHI	return_CVX	return_DD	return_DOW	return_HAL	return_HES	return_HUN	return_JPM	return_SLB	return_WFC	return_XOM
0.58577906	0.61046627	0.64179539	0.54681402	0.47197670	0.64986770	0.49976057	0.39225125	0.58034671	0.68269067	0.57372531	0.62696933

Preliminary Eigenvalues: Total = 16.9350893 Average = 1.41125745				
	Eigenvalue	Difference	Proportion	Cumulative
1	14.9446192	12.7338755	0.8825	0.8825
2	2.2107436	1.3691513	0.1305	1.0130
3	0.8415924	0.1303533	0.0497	1.0627
4	0.7112391	0.6898126	0.0420	1.1047
5	0.0214265	0.1469738	0.0013	1.1060
6	-0.1255473	0.0709849	-0.0074	1.0986
7	-0.1965322	0.0239638	-0.0116	1.0869
8	-0.2204960	0.0345938	-0.0130	1.0739
9	-0.2550898	0.0286557	-0.0151	1.0589
10	-0.2837455	0.0587253	-0.0168	1.0421
11	-0.3424708	0.0281790	-0.0202	1.0219
12	-0.3706498		-0.0219	1.0000

2 factors will be retained by the PROPORTION criterion.

Iteration	Criterion	Ridge	Change	Communalities										
1	0.6510876	0.0000	0.0766	0.61329	0.67722	0.59574	0.52016	0.43403	0.68819	0.52149	0.34962	0.59394	0.74783	0.65036	0.58326
2	0.6468795	0.0000	0.0233	0.62141	0.68257	0.58380	0.50908	0.42409	0.71149	0.51403	0.34494	0.60820	0.75240	0.65204	0.57178
3	0.6466363	0.0000	0.0044	0.62579	0.68570	0.57998	0.50612	0.42109	0.71292	0.51264	0.34336	0.61064	0.75401	0.65312	0.56870
4	0.6466159	0.0000	0.0017	0.62697	0.68582	0.57927	0.50513	0.42017	0.71351	0.51229	0.34293	0.61230	0.75408	0.65322	0.56796
5	0.6466133	0.0000	0.0006	0.62759	0.68586	0.57907	0.50478	0.41980	0.71345	0.51224	0.34275	0.61278	0.75412	0.65329	0.56779

Convergence criterion satisfied.

Significance Tests Based on 501 Observations			
Test	DF	Chi-Square	Pr > ChiSq
H0: No common factors	66	3656.2617	<.0001
HA: At least one common factor			
H0: 2 Factors are sufficient	43	319.3192	<.0001
HA: More factors are needed			

Chi-Square without Bartlett's Correction	323.30664
Akaike's Information Criterion	237.30664
Schwarz's Bayesian Criterion	55.99257
Tucker and Lewis's Reliability Coefficient	0.88187

Squared Canonical Correlations	
Factor1	Factor2
0.94176593	0.73146692

Below we can see that the factor 1 loading are more related to the variables.

Rotated Factor Pattern		
	Factor1	Factor2
return_BAC	0.76122	0.21969
return_BHI	0.21664	0.79932
return_CVX	0.49806	0.57530
return_DD	0.59542	0.38748
return_DOW	0.56395	0.31884
return_HAL	0.24256	0.80907
return_HES	0.40289	0.59153
return_HUN	0.50588	0.29457
return_JPM	0.75054	0.22277
return_SLB	0.35223	0.79376
return_WFC	0.75994	0.27534
return_XOM	0.51113	0.55362

Variance Explained by Each Factor		
Factor	Weighted	Unweighted
Factor1	8.7156851	3.55022275
Factor2	10.1803287	3.42320994

5.
Maximum Likelihood Factor Analysis, with Rotation and Max Priors

5 factor have been suggested by ML factor analysis

Rotated Factor Pattern					
	Factor1	Factor2	Factor3	Factor4	Factor5
return_BAC	0.19300	0.75425	0.26803	0.17215	0.09285
return_BHI	0.75597	0.14970	0.18684	0.24628	-0.01722
return_CVX	0.37688	0.25354	0.26440	0.70383	0.02658
return_DD	0.24372	0.27524	0.66859	0.31138	-0.13337
return_DOW	0.19396	0.25931	0.64481	0.23505	-0.00701
return_HAL	0.82071	0.18978	0.20801	0.16916	-0.00609
return_HES	0.47834	0.23976	0.25785	0.40900	0.24903
return_HUN	0.22592	0.26677	0.60996	0.06709	0.16770
return_JPM	0.20547	0.77151	0.22874	0.17842	-0.03102
return_SLB	0.72537	0.25575	0.24707	0.30301	0.05701
return_WFC	0.20847	0.61032	0.35934	0.29285	-0.00631
return_XOM	0.37166	0.29603	0.24083	0.66560	-0.02404

Variance Explained by Each Factor		
Factor	Weighted	Unweighted
Factor1	9.48177257	2.55119512
Factor2	6.95572063	2.08400430
Factor3	5.26449075	1.82173920
Factor4	5.80237050	1.59069819
Factor5	0.31984016	0.12246466

The above output suggests that the factor selection is highly dependent on the prior estimates of communalities. By looking at the out put of the factor loading we can group the factor loadings into below groups.

```
Factor 1: BHI, HAL, SLB
Factor 2: BAC, JPM, WFC
Factor 3: DD, DOW, HUN
Factor 4: CVX, XOM
Factor 5: No selections based on loading criterion
```

From the above analysis we can say that the Factor 1 strongly indicates Oil Field services, Factor 2 strongly indicates Banking, Factor 3 strongly indicates Industrial - Chemical, Factor 4 strongly indicates Oil Refining and HES which tends to lean heavily on oil field services factor than factor4 and for factor 5 we don't have any single variable.

We can conclude that the method chosen for prior calculation and communalities is significant over the chosen factors from the model.

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;

data temp;
set mydata.stock_portfolio_data;
* Let's drop some variables to get better factor analysis results;
drop AA HON MMM DPS KO PEP MPC GS ;
run;
proc sort data=temp; by date; run; quit;

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

data temp;
set temp;
* Compute the log-returs;
* Note that the data needs to be sorted in the correct
direction in order for us to compute the correct return;
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_HAL = log(HAL/lag1(HAL));
return_HES = log(HES/lag1(HES));
return_HUN = log(HUN/lag1(HUN));
return_JPM = log(JPM/lag1(JPM));
return_SLB = log(SLB/lag1(SLB));
return_WFC = log(WFC/lag1(WFC));
return_XOM = log(XOM/lag1(XOM));
* Compute the remainder of the log-returs;
response_VV = log(VV/lag1(VV));
run;
proc print data=temp(obs=10); run; quit;
```

Assignment #7: Factor Analysis
PREDICT 410

```
data return_data;
set temp (keep= return_:);
* What happens when I put this keep statement in the set statement?;
* Look it up in The Little SAS Book;
run;
proc print data=return_data(obs=10); run;


ods graphics on;
proc factor data=return_data method=principal priors=smc rotate=none
plots=(all); run; quit;
ods graphics off;

ods graphics on;
proc factor data=return_data method=principal priors=smc rotate=varimax
plots=(all); run; quit;
ods graphics off;

ods graphics on;
proc factor data=return_data method=ML priors=smc rotate=varimax
plots=(loadings); run; quit;
ods graphics off;




ods graphics on;   proc factor data=return_data method=ML priors=max rotate=varimax
  plots=(loadings); run; quit;
ods graphics off;
```