

SAS Assignment 2 – PCA

A study of whether air pollution continues to mortality

JOO YONG YOON

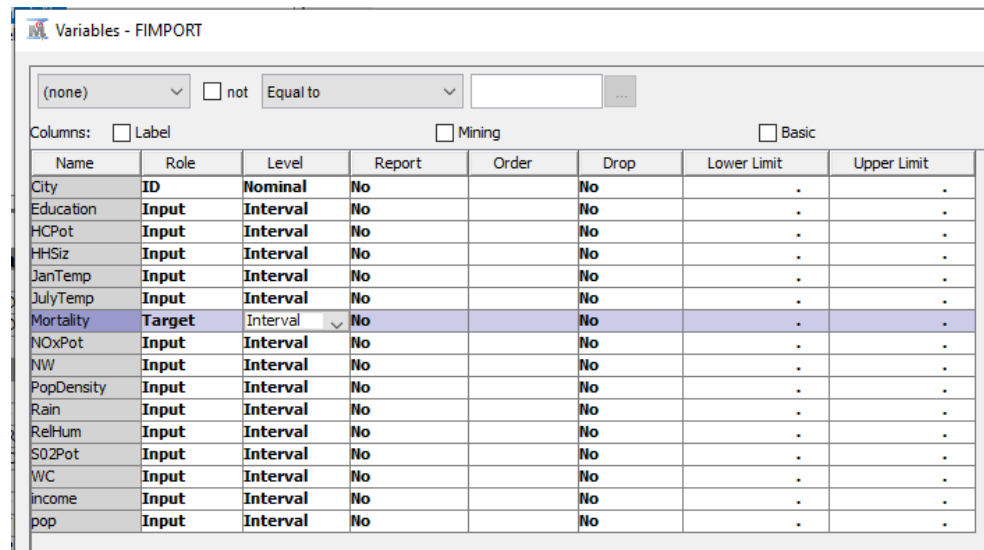
MI 353

Dr. CEVIKPARMAK

22 Feb 2023

Answers to the questions presented (#5~10)

* Please refer to the figure(s) and table(s) provided for the question.



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
City	ID	Nominal	No		No	.	.
Education	Input	Interval	No		No	.	.
HCPot	Input	Interval	No		No	.	.
HHSiz	Input	Interval	No		No	.	.
JanTemp	Input	Interval	No		No	.	.
JulyTemp	Input	Interval	No		No	.	.
Mortality	Target	Interval	No		No	.	.
NOxPot	Input	Interval	No		No	.	.
NW	Input	Interval	No		No	.	.
PopDensity	Input	Interval	No		No	.	.
Rain	Input	Interval	No		No	.	.
RelHum	Input	Interval	No		No	.	.
SO2Pot	Input	Interval	No		No	.	.
WC	Input	Interval	No		No	.	.
income	Input	Interval	No		No	.	.
pop	Input	Interval	No		No	.	.

Figure 1. Variable Settings.

5. Model 1 / Linear regression using all the original variables

a. What are the R2 and adjusted R2 Values?

Model Fit Statistics			
R-Square	0.7619	Adj R-Sq	0.6862
AIC	432.1213	BIC	444.1162
SBC	463.2844	C(p)	15.0000

Table 1. Model 1: Model Fit Statistics

R2 and adjusted R2 values of Model 1 are **0.7619** and **0.6862**, respectively. The closer the adjusted R2 is to 100%, the better the regression model is. Adjusted R2 of 0.6862 means that input variables explain about 69% of the variation in the target variables, which I assume is passable at this point of the project.

b. Which variables (if any) are significant based on the t value statistics and associated probabilities ($Pr > |t|$)

Based on Table 2, it is **not clear** to determine which variable is significant for model 1. All variables but NW (percentage of non-whites) fail to reject the null hypothesis as their p-value is **not less than .05**. This means that all the variables but NW do not have a statistically significant relationship between mortality. Additionally, the absolute t value (5.85) of NW is greater than other variables, and this could be a sign of

significant difference as the t value measures the size of the difference relative to the variation in the sample data (the closer T is to 0, the more likely there isn't a significant difference). I decide to **exclude this possibility** as the t value in the output is calculated from only one sample from the entire population.

Parameter	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	1	1400.0	281.6	4.97	<.0001
Education	1	-11.0467	8.9980	-1.23	0.2261
HCPot	1	-0.6712	0.4556	-1.47	0.1478
HHSiz	1	-38.0416	40.2752	-0.94	0.3501
JanTemp	1	-1.4411	0.7638	-1.89	0.0658
JulyTemp	1	-2.9503	1.9351	-1.52	0.1345
NOxPot	1	1.1785	0.9144	1.29	0.2042
NW	1	5.3027	0.9064	5.85	<.0001
PopDensity	1	0.00472	0.00434	1.09	0.2826
Rain	1	0.9695	0.5851	1.66	0.1046
RelHum	1	0.1360	1.1508	0.12	0.9065
SO2Pot	1	0.0846	0.1357	0.62	0.5362
WC	1	-1.4923	1.2325	-1.21	0.2324
income	1	-0.00043	0.00129	-0.33	0.7435
pop	1	3.402E-6	4.116E-6	0.83	0.4129

Table 2. Model 1: Analysis of Maximum Likelihood Estimates

	Correlation Coefficient
Mort and Rain	.535
Mort and Education	-.581
Mort and NW	.659
Mort and HCPot	-.199
Mort and NOxPot	-.099
Mort and SO2Pot	.418

Table 3. Correlation Coefficient, Excel

According to Table 3 and 4, calculated **correlation coefficient does not** help to determine which variable is significant. Mort VS. Rain, Mort VS. NonWht, and Mort VS. SO2 are positively correlated. Mort and SO2 has weak positive correlation. Mort VS. Rain, and Mort VS. NonWht have moderate positive correlation. The coefficients of the three pairs are not close to +1 which indicates that there is no strong linear relationship. Mort VS. Educ, Mort VS. HC, and Mort VS. NOx are negatively correlated. Mort and Educ has moderate negative correlation. Mort VS. HC, and Mort VS. NOx have weak negative correlation. The coefficients of the three pairs are not close to -1 which indicates that there is no strong linear relationship. For some variables such as HHSiz, JanTemp,

Rain, etc have a similar mean and median value, which suggest a healthier distribution though.

Table 4. Stat Explore

Variable	Role	Mean	Standard Deviation	Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Education	INPUT	10.9661	0.850681	59	0	9	11	12.3	-0.20048	-0.77886
HCPot	INPUT	38.47458	92.63878	59	0	1	15	648	5.552357	34.15214
HHSiz	INPUT	3.24661	0.182923	59	0	2.65	3.27	3.53	-1.68806	3.740761
JanTemp	INPUT	33.79661	10.15191	59	0	12	31	67	1.017078	1.246211
JulyTemp	INPUT	74.40678	4.601794	59	0	63	74	85	0.066507	0.045033
NOxPot	INPUT	22.9661	46.66571	59	0	1	9	319	5.126234	29.7871
NW	INPUT	11.87627	8.997592	59	0	0.8	9.5	38.5	1.120044	0.86769
PopDensity	INPUT	3910.492	1441.69	59	0	1441	3626	9699	1.416453	3.732248
Rain	INPUT	38.50847	11.57341	59	0	10	38	65	-0.18155	0.984186
RelHum	INPUT	57.74576	5.380659	59	0	38	57	73	0.206293	4.342768
S02Pot	INPUT	54.66102	63.55167	59	0	1	32	278	1.89764	3.482461
WC	INPUT	46.38644	5.069702	59	0	33.8	45.5	62.2	0.469013	1.234342
income	INPUT	33246.66	4473.095	59	0	25782	32452	47966	1.277273	2.138679
pop	INPUT	1438037	1541736	59	0	124833	914427	8274961	2.889073	9.55786
Mortality	TARGET	941.1731	62.42133	59	0	790.73	946.19	1113.16	0.066275	0.166972

6. Model 2 (Stepwise Model) / Linear Regression w/ only variable that were significant

* Since I was unable to determine which variable is significant from the regular regression model, I applied **Stepwise Model** in order to choose which variable to use for Model 2.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	1	1037.2	84.5260	12.27	<.0001
Education	1	-13.6778	6.6281	-2.06	0.0440
JanTemp	1	-1.6471	0.5341	-3.08	0.0032
NW	1	4.3214	0.6375	6.78	<.0001
Rain	1	1.1251	0.4928	2.28	0.0264
S02Pot	1	0.2742	0.0804	3.41	0.0012

Table 5. Stepwise Model, Analysis of Maximum Likelihood Estimates

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	F Value	Pr > F
1	NW	1	1	40.94	<.0001
2	Education	1	2	18.67	<.0001
3	JanTemp	1	3	11.04	0.0016
4	S02Pot	1	4	7.43	0.0086
5	Rain	1	5	5.21	0.0264

Table 6. Stepwise Model, Summary of Stepwise Selection

According to Table 5 and 6, the Stepwise model suggests **5 variables to be significant** (NW, Education, JanTemp, SO2Pot, Rain) as the model decides that no (additional) effects met the 0.05 significance level for entry into the model.

Variables - Reg3

(none) ☐ not Equal to

Columns: ☐ Label ☐ Mining

Name	Use	Report	Role	Level
Education	Yes	No	Input	Interval
HCPot	No	No	Input	Interval
HHSiz	No	No	Input	Interval
JanTemp	Yes	No	Input	Interval
JulyTemp	No	No	Input	Interval
Mortality	Yes	No	Target	Interval
NOxPot	No	No	Input	Interval
NW	Yes	No	Input	Interval
PopDensity	No	No	Input	Interval
Rain	Yes	No	Input	Interval
RelHum	No	No	Input	Interval
SO2Pot	Yes	No	Input	Interval
WC	No	No	Input	Interval
income	No	No	Input	Interval
pop	No	No	Input	Interval

Figure 2. Variable Settings, Model 2

I used these 5 variables to build Model 2 regression model and got the same results as Stepwise Model.

a. Why is this an important step when running regression models?

First, it helps determine all of the variables that are related to the outcome, which makes the model complete and accurate. Second, it helps select a model with few variables by eliminating irrelevant variables that decrease the precision and increase the complexity of the model. Ultimately, variable selection provides a balance between simplicity and fit.

b. What are the R2 and adjusted R2 values?

Model Fit Statistics			
R-Square	0.7091	Adj R-Sq	0.6816
AIC	425.9437	BIC	429.2765
SBC	438.4089	C(p)	6.0000

Table 7. Model 2. Model Fit Statistics

R2 and adjusted R2 values of Model 1 is **0.7091** and **0.6816**, respectively.

c. Which variables (if any) are significant based on the t value statistics and associated probabilities ($Pr > |t|$)

All 5 variables (NW, Education, JanTemp, SO2Pot, Rain) selected by Stepwise Model are to be significant as p-value for each variable is less than .05.

7. PCA

a. How many principal components are selected?

Eigenvalues of Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.03031103	1.63667241	0.2879	0.2879
2	2.39363862	0.56509389	0.1710	0.4589
3	1.82854473	0.49964336	0.1306	0.5895
4	1.32890137	0.28416255	0.0949	0.6844
5	1.04473882	0.12574698	0.0746	0.7590
6	0.91899184	0.34753034	0.0656	0.8247
7	0.57146151	0.08226091	0.0408	0.8655
8	0.48920060	0.06969365	0.0349	0.9004
9	0.41950694	0.02753669	0.0300	0.9304
10	0.39197025	0.12106222	0.0280	0.9584
11	0.27090804	0.10251079	0.0194	0.9777
12	0.16839725	0.03081492	0.0120	0.9898
13	0.13758233	0.13173567	0.0098	0.9996
14	0.00584666		0.0004	1.0000

Table 8. PCA, Evaluation of Correlation Matrix

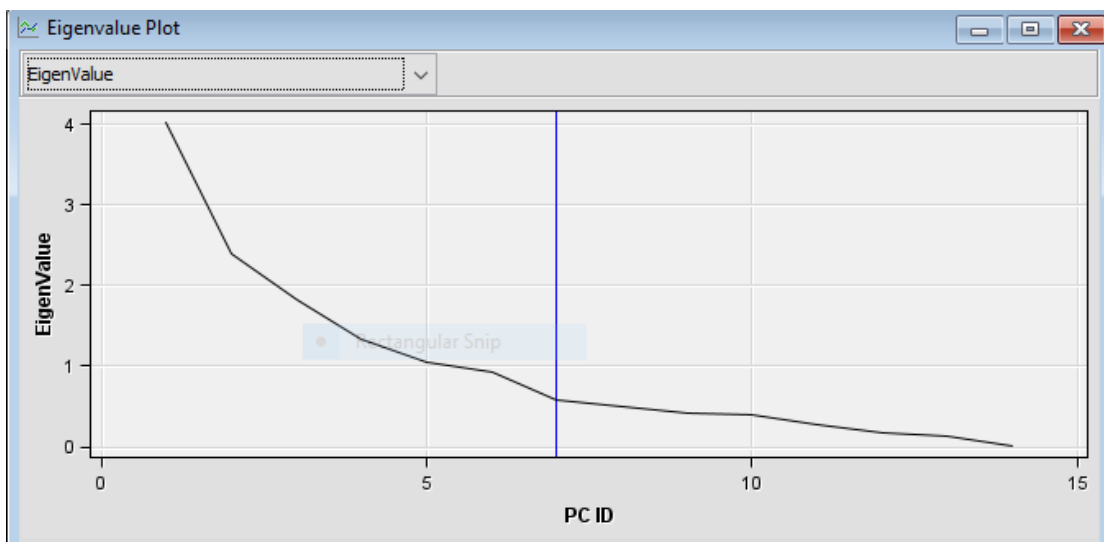


Figure 3. PCA Eigenvalue Plot

For the original PCA model, there are 14 principal components that corresponds to accumulative value of 1.0. With 85% similarity, there are **7 principal components** selected.

b. Does this seem like a reasonable number? Explain your answer.

I think 7 principal components can be a reasonable number. It honestly seems that deciding the number of principal components can be subjective to the goal of the model. Lowering the number of principal components by maintaining low similarity can be a aggressive reduction, but it could also be useful if the goal if to reduce the number of principal components. On the other hand, maintaining the similarity, cumulative variance, in the data set could be a another goal. I believe reducing the number of variables should be the priority for this dataset due to the results of Model 1. Additionally, principal components are not co-related to each other. Regression can be very sensitive to correlation, and reducing correlation among variables could be also effective.

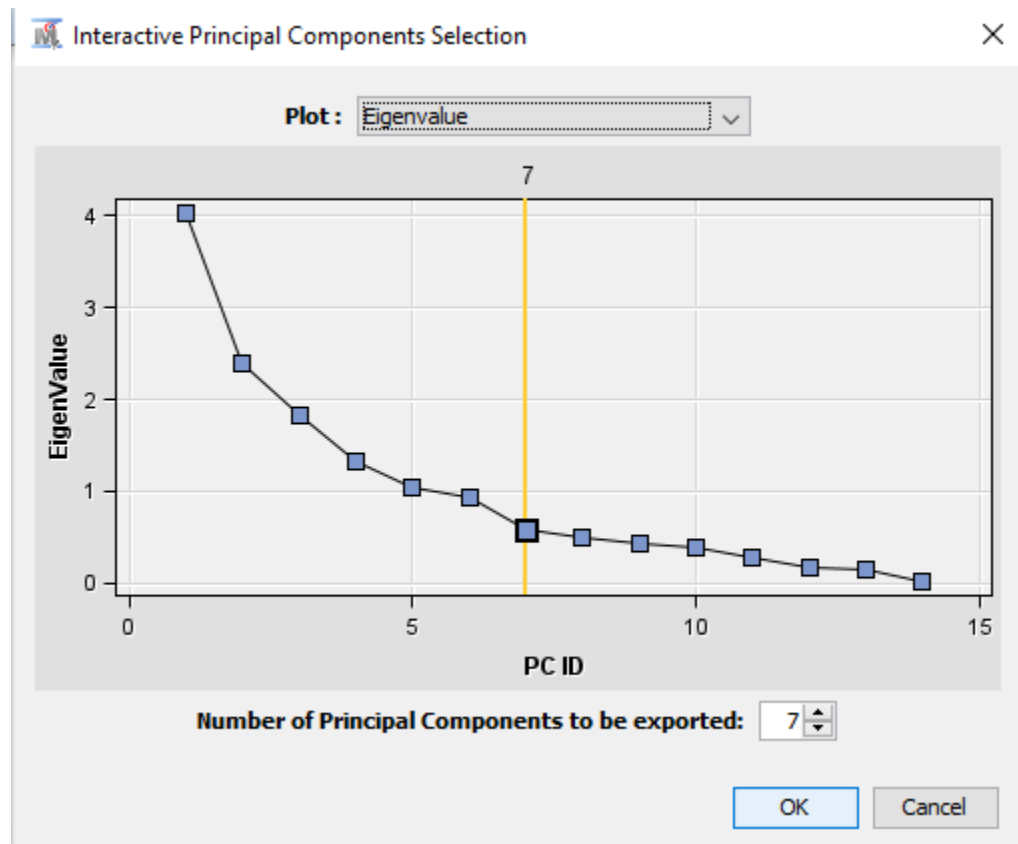


Figure 4. PCA Selection

8. Model 3 / PCA – Regression

a. What are the R2 and adjusted R2 Values?

Model Fit Statistics			
R-Square	0.6390	Adj R-Sq	0.5895
AIC	442.6779	BIC	447.1385
SBC	459.2982	C(p)	8.0000

Table 9. Model 3. Model Fit Statistics

R2 and adjusted R2 values of Model 1 is **0.6390** and **0.5895**, respectively.

b. Which variables (if any) are significant based on the t value statistics and associated probabilities ($Pr > |t|$)

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	1	941.2	5.2070	180.75	<.0001
PC_1	1	-11.3761	2.6159	-4.35	<.0001
PC_2	1	22.9959	3.3944	6.77	<.0001
PC_3	1	-10.1128	3.8837	-2.60	0.0120
PC_4	1	-5.6194	4.5557	-1.23	0.2230
PC_5	1	6.8720	5.1380	1.34	0.1870
PC_6	1	21.4491	5.4783	3.92	0.0003
PC_7	1	1.5443	6.9471	0.22	0.8250

Table 10. Model 3. Analysis of Maximum Likelihood Estimates

Model 3 suggests that PC1, 2, 3, and 6 are significant. The below is my interpretation on each principal models that the regression model found significant.

PC#	Principal Components Coefficient	Interpretation
PC1		In large perspective, PC 1 can be named as pollution level -HCPot (HC pollution potential) and NOxPot (Nitrous Oxide pollution potential). This data shows that HCPot and NOxPot are primary contributors to the largest variances within data. It could also be a valid interpretation that Rain may reduce both HCPot and NoxPot level.

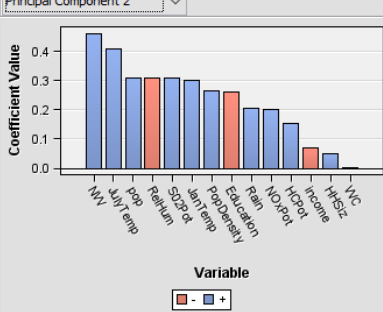
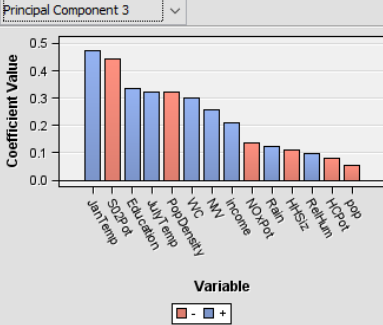
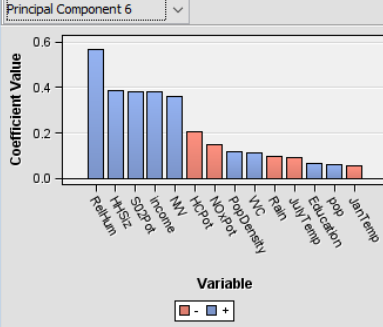

PC2		<p>It is a bit more complicated to name PC2 than PC1 because NW and July Temp cannot be correlated. Not only the pollution level but the racial and ethnic differences can affect the mortality rate. It is also possible that July Temperature can affect the mortality rate (heat wave). Since NW contributed more, I will name PC2 as percentage of non-whites.</p>
PC3		<p>PC3 should be named as temperature. It can be also interpreted that high January temp can affect lower SO2 plot. July temperature also contributed quite some to PC3.</p>
		<p>PC6 should represent Relative Humidity. It is hardly predictable that high relative humidity is related to Household size and percentage of non-whites.</p>

Table 11. Significant PCs, Interpretation

9. Model 4 / PCA – Regression w/ Significant Variables

 Variables - Reg5

(none) ☐ not Equal to

Columns: ☐ Label ☐ Mining

Name	Use	Report	Role	Level
Mortality	Yes	No	Target	Interval
PC_1	Yes	No	Input	Interval
PC_2	Yes	No	Input	Interval
PC_3	Yes	No	Input	Interval
PC_4	No	No	Input	Interval
PC_5	No	No	Input	Interval
PC_6	Yes	No	Input	Interval
PC_7	No	No	Input	Interval

Figure 5. Variable Settings, Model 4

a. What are the R2 and adjusted R2 Values?

Model Fit Statistics			
R-Square	0.6152	Adj R-Sq	0.5867
AIC	440.4420	BIC	443.3508
SBC	450.8297	C(p)	5.0000

Table 12. Model 4. Model Fit Statistics

R2 and adjusted R2 values of Model 1 is **0.6152** and **0.5867**, respectively.

b. Which variables (if any) are significant based on the t value statistics and associated probabilities ($Pr > |t|$)

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	1	941.2	5.2243	180.15	<.0001
PC_1	1	-11.3761	2.6246	-4.33	<.0001
PC_2	1	22.9959	3.4057	6.75	<.0001
PC_3	1	-10.1128	3.8966	-2.60	0.0121
PC_6	1	21.4491	5.4965	3.90	0.0003

Table 13. Model 4. Analysis of Maximum Likelihood Estimates

Based on a t-test, all variables (PC 1, 2, 3, 6) are significant for model 4.

10. Comment on the results of the regression models (Models 1-4). Do you have a preference among them? Explain

Fit Statistics								
Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Train: Average Squared Error	Train: Akaike's Information Criterion	Train: Average Squared Error
Y	Reg	Reg	Model 1	Mortality	Mortality	911.9782	432.1213	911.9
	Reg3	Reg3	Model 2	Mortality	Mortality	1114.314	425.9437	1114
	Reg4	Reg4	Model 3	Mortality	Mortality	1382.749	442.6779	1382
	Reg5	Reg5	Model 4	Mortality	Mortality	1473.841	440.442	1473

Figure 6. Model Comparison

Surprisingly, SAS's model comparison chose Model 1 yet I believe this is not the best choice. It is important, in my opinion, to recall the purpose of this regression model: a study of whether air pollution contributes to mortality. Therefore, a variable such as HC pollution potential should be investigated more than the average household size. That is one of the reasons why many tasks of this assignment focus on dealing with variables, aggressively reducing variables, and using PCA to ignore the correlation between variables and groups.

Even though I do not understand all the information under the result of the model comparison of SAS, I suspect that SAS's model comparison chose Model 1 because it has the highest adjusted R square value.

Model #	Adjusted R2	F
1	.6862	10.06
2	.6816	25.84
3	.5895	12.90
4	.5867	21.59

Table14. Model Comparison

In my opinion, Model 4 is the most preferable because it has a relatively high F value and reasonable R2 value with PCs. For the most part of this assignment, I have focused on the t-test in order to determine the significance of variables. However, it is also important to evaluate the F value as it is a performance measure for each model. If the R2 value of model 4 is significantly less than other models' it would be better to adjust the PC and examine the results.

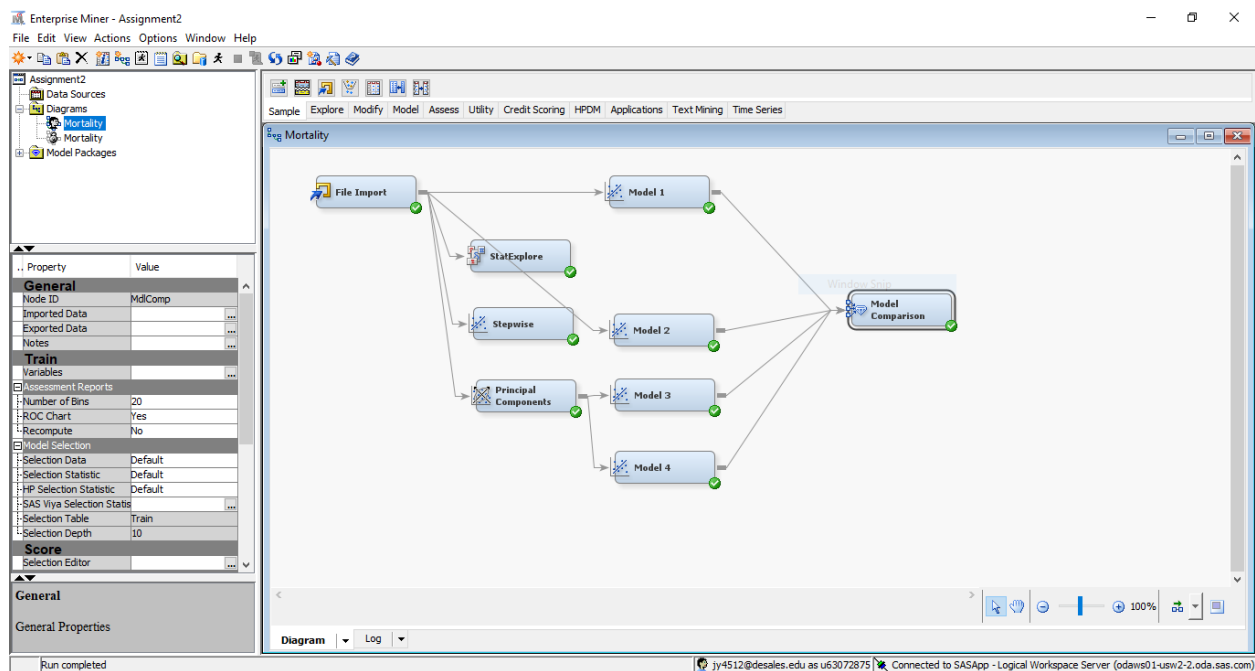


Figure 7. Diagram