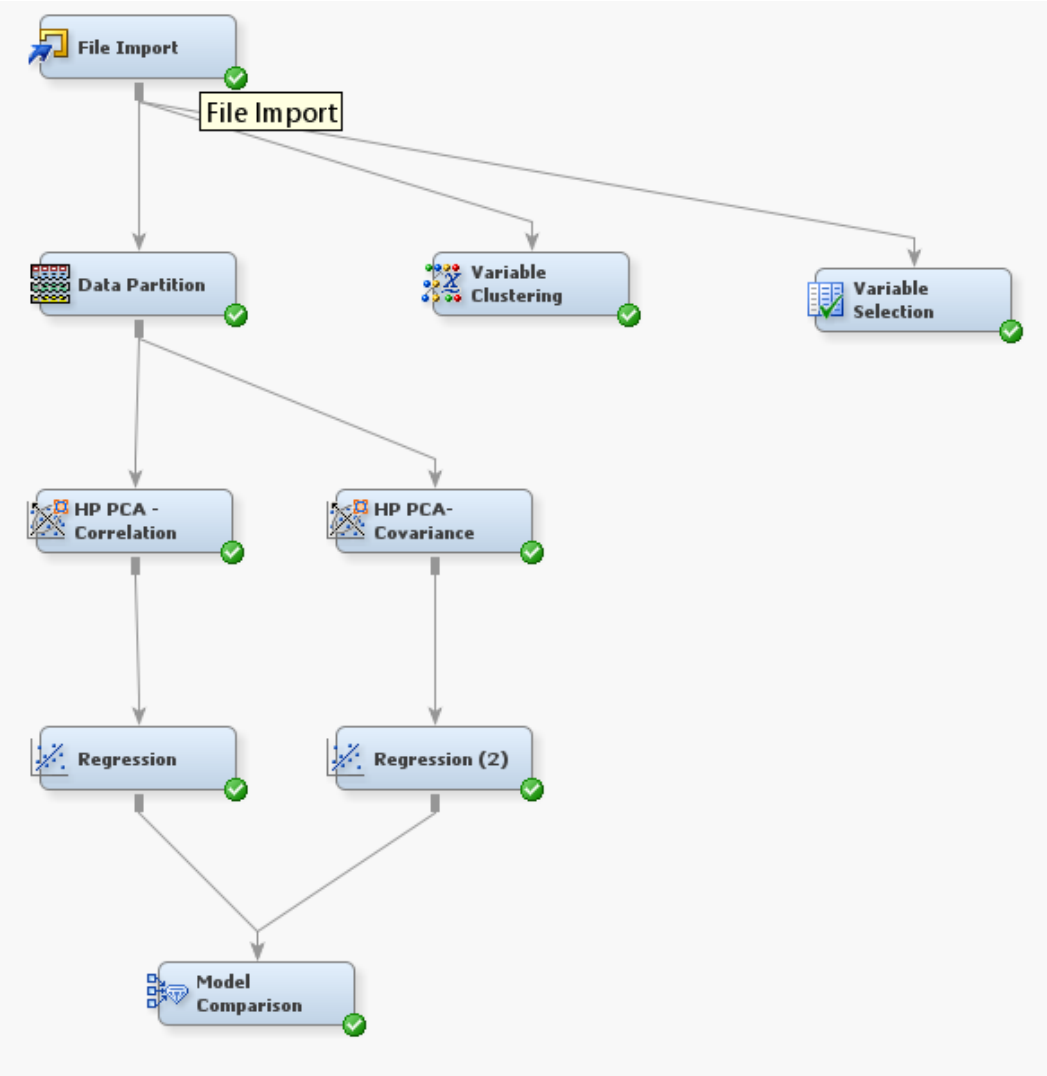
Diane Nguyen
Class Assignment 2
MSBA 645
Due: 3/31/2019

- The flow diagram



- The top 10 principal components for covariance and correlation matrix are:
    - For correlation: calories, carbo, cups, fiber, potass, protein, shelf, sodium, sugars, and vitamins

Variable	Label	Prin1	Prin2	
calories	calories	-0.25633	0.23578	
carbo	carbo	-0.25053	-0.11615	
cups	cups	-0.31120	-0.20482	
fat	fat	0.05505	0.21598	
fiber	fiber	0.41717	0.21749	-
potass	potass	0.40750	0.25205	
protein	protein	0.31530	0.06504	
shelf	shelf	0.22185	0.13317	
sodium	sodium	-0.28046	0.20879	
sugars	sugars	-0.19080	0.23159	
vitamins	vitamins	-0.12062	0.16422	
weight	weight	-0.02597	0.32488	
_DMV1_L0		0.13145	-0.37074	
_DMV1_L1		-0.16306	0.13107	
_DMV1_L2		-0.05857	-0.03406	-
_DMV1_L3		0.25871	0.02023	-
_DMV1_L4		0.04644	0.13806	
_DMV1_L5		0.02852	-0.16176	-
_DMV1_L6		0.02988	0.05880	
_DMV2_L0		-0.13145	0.37074	-
_DMV2_L1		0.13145	-0.37074	

    - For covariance: calories, carbo, fiber, potass, protein, shelf, sodium, sugars, vitamins and cups

Variable	Label	Prin1	Prin2
calories	calories	0.09232	0.02837
carbo	carbo	0.02055	-0.00922
cups	cups	0.00098993	-0.00114
fat	fat	-0.00027765	0.00289
fiber	fiber	-0.01640	0.02903
potass	potass	-0.46896	0.88079
protein	protein	-0.00439	0.00909
shelf	shelf	-0.00371	0.00262
sodium	sodium	0.87289	0.45973
sugars	sugars	0.00818	0.00119
vitamins	vitamins	0.09384	0.10496
weight	weight	0.00027205	0.00132
_DMV1_L0		-0.00047748	-0.00034366
_DMV1_L1		0.00156	0.00060636
_DMV1_L2		0.00089862	-0.00021977
_DMV1_L3		-0.00095764	0.00050861
_DMV1_L4		-0.00043416	0.00047601
_DMV1_L5		-0.00043171	-0.00142
_DMV1_L6		-0.00015319	0.00039607
_DMV2_L0		0.00047748	0.00034366
_DMV2_L1		-0.00047748	-0.00034366

- The difference between the two sets of PCAs in terms of:
    - 1) the number of components needed to explain 90% of variation:

      For correlation, in this case the first Prin1 you can see that fiber and potass have the highest contribution so they dominate here in Prin1.
      In the other case using covariance matrix, you can see that in Prin1 the first one is sodium (.87) and potass is the second.

    - 2) The top two weighted contributions to the first components:

      For correlation matrix the first and second principle components make up the top weighted contributions with 17.89% and 16.64%

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	3.757490	0.263527	0.1789	0.1789
2	3.493963	0.493543	0.1664	0.3453
3	3.000420	0.669955	0.1429	0.4882
4	2.330465	0.812541	0.1110	0.5992
5	1.517924	0.148928	0.0723	0.6714
6	1.368996	0.049928	0.0652	0.7366
7	1.319068	0.422960	0.0628	0.7994
8	0.896108	0.128686	0.0427	0.8421
9	0.767422	0.001467	0.0365	0.8787
10	0.765955	0.115291	0.0365	0.9151
11	0.650664	0.231727	0.0310	0.9461
12	0.418938	0.139252	0.0199	0.9661
13	0.279686	0.067663	0.0133	0.9794
14	0.212023	0.045503	0.0101	0.9895
15	0.166520	0.142097	0.0079	0.9974
16	0.024423	0.002127	0.0012	0.9986
17	0.022296	0.014655	0.0011	0.9996
18	0.007641	0.007641	0.0004	1.0000
19	0	0	0.0000	1.0000
20	0	0	0.0000	1.0000
21	0		0.0000	1.0000

For covariance matrix the first and second principle components make up the top weighted contributions with 54.26% and 38.66%

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	7238.351801	2080.789495	0.5426	0.5426
2	5157.562306	4547.751974	0.3866	0.9292
3	609.810332	297.962220	0.0457	0.9749
4	311.848111	293.448752	0.0234	0.9983
5	18.399360	15.891580	0.0014	0.9996
6	2.507780	1.515599	0.0002	0.9998
7	0.992181	0.505460	0.0001	0.9999
8	0.486721	0.159694	0.0000	0.9999
9	0.327026	0.161793	0.0000	1.0000
10	0.165233	0.053708	0.0000	1.0000
11	0.111525	0.015353	0.0000	1.0000
12	0.096172	0.039727	0.0000	1.0000
13	0.056445	0.015616	0.0000	1.0000
14	0.040829	0.000612	0.0000	1.0000
15	0.040217	0.008383	0.0000	1.0000
16	0.031834	0.010781	0.0000	1.0000
17	0.021053	0.019914	0.0000	1.0000
18	0.001139	0.001139	0.0000	1.0000
19	0	0	0.0000	1.0000
20	0	0	0.0000	1.0000
21	0		0.0000	1.0000

- Explain the reason for the differences you have observed in the previous question and briefly state the general guideline on when to use correlation matrix and when to use covariance matrix in computing PCAs.

Due to the fact that we have a lot of variables that often contain a large variation in values, the PCAs are weighted at nearly 98%, which implies that there is strong correlation. We need to use correlation when we have to perform normalization on the scales. We will use covariance when the scales are similar and you do not need to normalize them. Covariance is affected by the change in scale, while correlation is not. Correlation is really just a specific case of covariance which we can find by standardizing the data. When making a choice, which is a better measure of the relationship between two variables, correlation is better because it remains unaffected by the change in location and scale.

- Provide the RMSs for the two regression models and state whether you should use covariance matrix or correlation matrix in this particular case (i.e., this cereals dataset).
    - The Root Mean Square Error for the Correlation regression is 3.015. The Root Mean Square for the Covariance regression is 8.6031. Here you should use correlation matrix in this particular case.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Valid: Root Mean Square Error
Y	Reg	Reg	Regressi...	rating	rating	3.015944
	Reg2	Reg2	Regressi...	rating	rating	8.603189