

Part 1: Multivariate Analysis Basics

1. Consider all of the numeric variables (i.e. all of the variables except Brand and Name). Determine the variance/covariance matrix and the correlation matrix of these variables. Interpret briefly what you see.

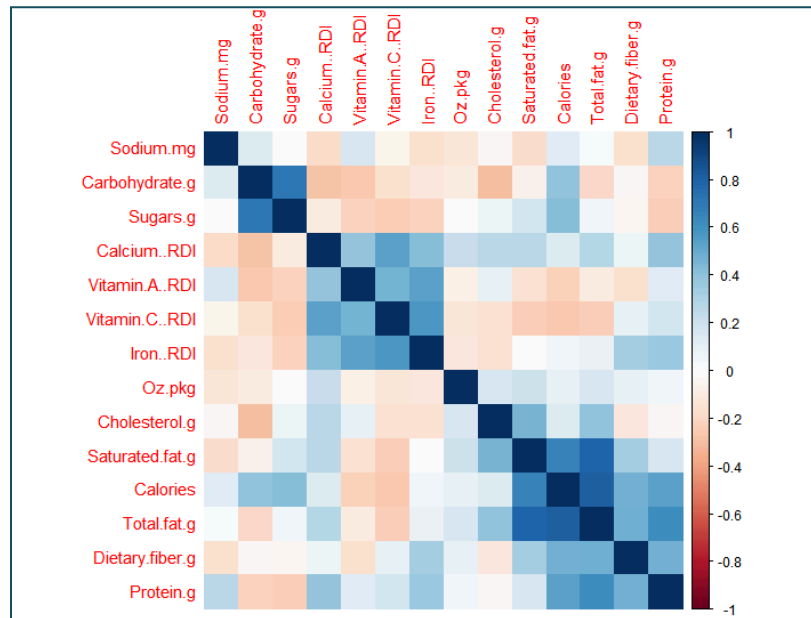
Variance-Covariance Matrix:

The diagonal of the variance-covariance matrix includes the variances for each individual variable. For example, the variance of sugars is 48.946, which means the standard deviation of sugars is approximately 7g. On average, candy bars deviate from the mean amount of sugar by about 7g. The non-diagonals of the matrix are the covariances, describing the relationship between two variables. We can see that all of the positive covariances have a direct relationship between the two variables and the negative covariances have an inverse relationship between the two variables. For example, total fat and carbohydrates have a covariance of -9.549; meaning, as fat increases, carbohydrates tend to decrease and vice versa. On the other hand, total fat and calories have a covariance of 277.532; meaning, as fat increases, calories also increase.

Correlation Matrix:

The correlation matrix is standardized, making interpretations between two variables and comparing these relationships with other relationships easier. While positive and negative correlations still indicate the type of relationship, the value also indicates the strength of the relationship. For example, total fat and calories have a correlation of 0.811, which indicates a strong, positive relationship. On the other hand, sugar and protein have a correlation of -0.236 indicating a relatively weak, negative relationship.

2. Construct a color map on correlations of the variables specified in #1. Comment on what you observe.



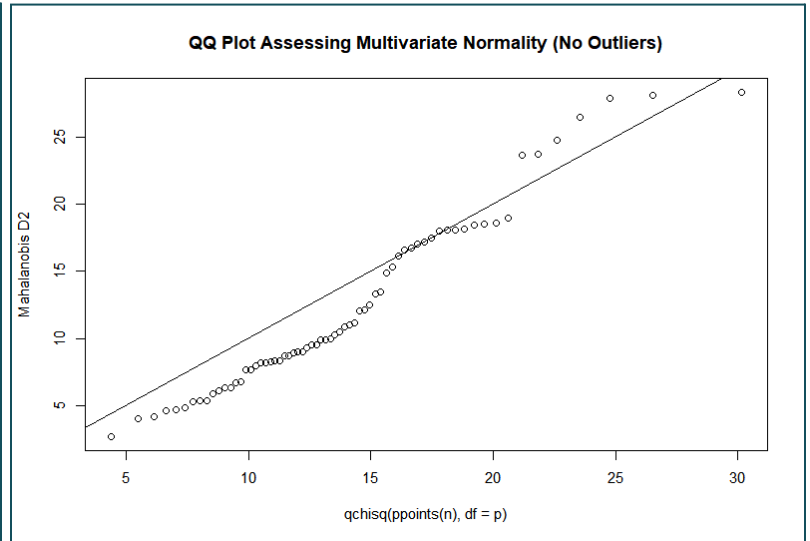
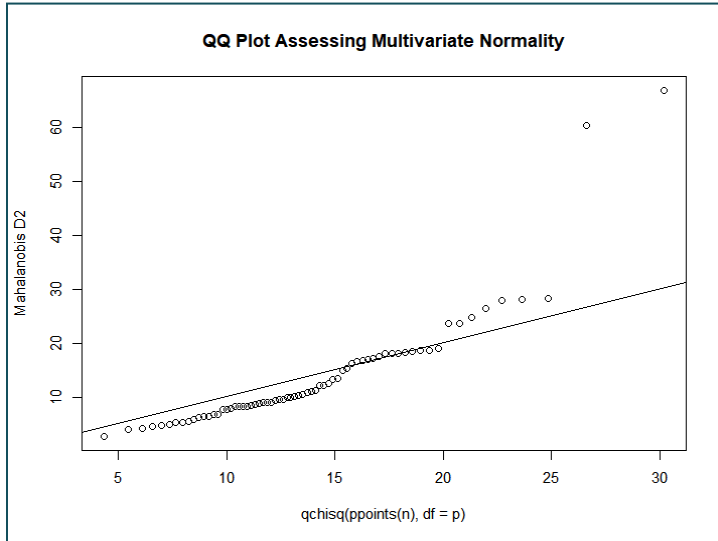
The color map on correlations is a visual representation of the correlation matrix, distinctly indicating positive and negative relationships as well as the severity of the relationship indicated by the deeper the color. In this example we can quickly identify that saturated fat, calories, and total fat all have a strong, positive correlation with one another and may be a potential source of multicollinearity. We can also see that the micronutrients and vitamins such as vitamin A, vitamin C, and iron have a moderately weak, negative correlation.

3. Use univariate probability plots and histograms to assess the normality of the variables specified in #1. Interpret.

Univariate Probability Plots, Histograms

Both probability plots and histograms are methods to assess univariate normality. Histograms of normal data have a symmetrical, bell shape, but may not be the most robust method to determine normality. Probability plots rely less on one's ability to identify a normal curve, but rather compare theoretical data to the observed data. Using both in this example, I would classify total fat, saturated fat, sodium, carbohydrates, sugars, and protein as following an approximately normal distribution. The remaining variables do not follow a normal distribution as the data meanders too far away from the reference line.

4. Use a probability plot to assess the multivariate normality of the variables specified in #1. Interpret.



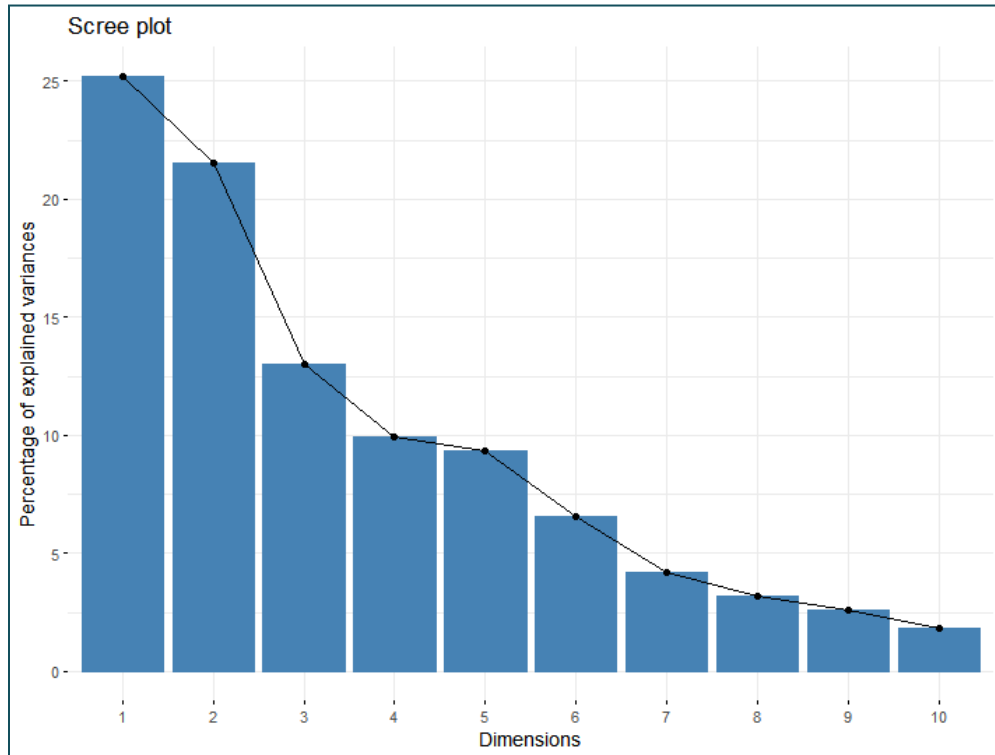
The probability plots above are a method to assess multivariate normality. We are checking to see if the Mahalanobis distances of the data follow a Chi Square distribution. If this is the case, the underlying data follows a multivariate normal distribution, an assumption is preferred for analysis and informs whether covariance and correlation matrices are good measures of similarity. In the example on the left, most of the data follows the reference line except for two obvious outliers. Removing these two outliers, as on the left, causes the remaining data to approximately follow a multivariate normal distribution. When continuing with the analysis, examining these outliers individually and considering the data without them may be beneficial.

Part 2: Principal Component Analysis

1. Find and display the eigenvalues of the correlation matrix. Use these along with a Scree plot (and/or other means) to determine the number of principal components that are to be retained. How much of the variation in the data is explained by your chosen number of principal components? Remember that your goal is dimension reduction...so please be sensible in your choice of number of components.

Eigenvalues

Based on the eigenvalues linked above and the Scree plot, the first five principal components will be retained. By retaining the first five components, 78.97% of the variation can be explained. While the Scree plot does not show a sharp elbow, retaining components beyond the first five does not substantially increase the explained variation, have corresponding eigenvalues less than 1, and continue to add complexity.



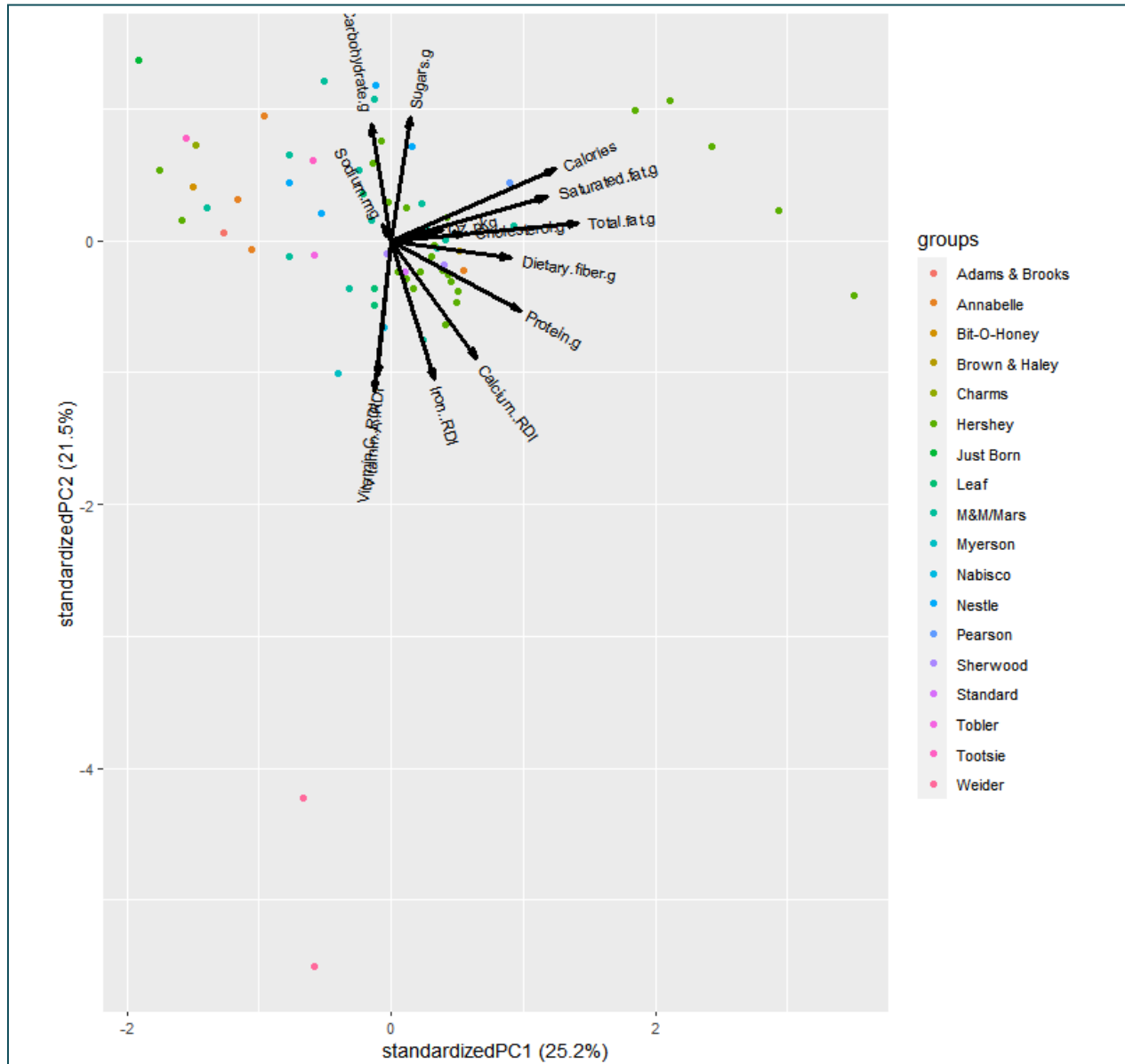
2. Provide the loadings matrix. What do you learn from this matrix? Using this matrix, and as best as possible, interpret the first two principal components (I will expect some interpretation).

Loading Matrix

The loading matrix is a standardized value that indicates the magnitude or influence a variable has in a principle component. Using the loading matrix, we can see the relationships between different variables and the relative influence a variable has in each component. For example, in the first principal component, the strongest loadings (closest to 1 or -1) include fiber, protein, saturated fat, total fat, calories, and perhaps calcium. Each of these loadings are positive and larger than the absolute value of the other variable's loadings. This may indicate that the first principle component is capturing more nutrient dense candy bars that have substantial ingredients like nuts and/or seeds. The variables that correspond with the strongest loadings in the second principle component include sugar, carbohydrates, vitamin A, vitamin C, calcium, and iron. Sugar and calcium have a positive loading and the others have a negative loading, indicating an inverse relationship. This principle component may be capturing the candy bars that are sugar based or have sugary fillings, lacking any other nutrients or substance.

3. Construct a biplot of the loadings and scores for the first two principal components. For this plot, color the observations based on Brand. Are there any natural groupings of the observations? Are there any unusual observations?

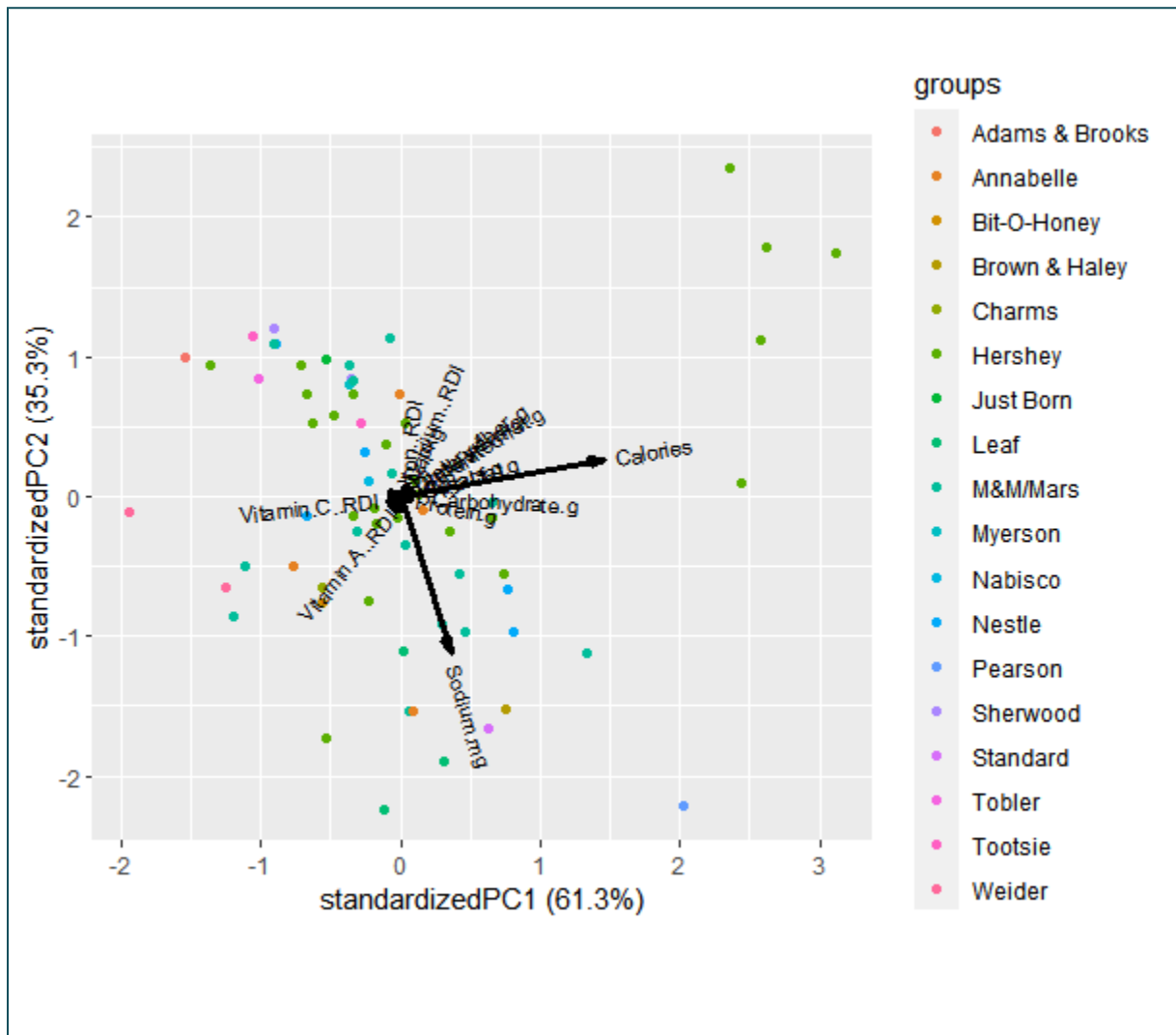
Considering the biplot of the loadings and scores for the first two principal components below, there appears to be a cluster of Hershey brand candy bars on the right. Projecting these five candy bars to the x-axis shows that they are the candy bars that are high in calories, saturated fat, total fat, dietary fiber, and protein. There are a couple pink observations towards the bottom of the plot that stand out from the rest. Because these are in the direction of the calcium arrow (projected on the y-axis), and the opposite direction from the sugars and carbohydrates arrow, we can conclude that these candy bars are perhaps unusually high in calcium and low in sugars and carbohydrates.



4. Show how the results change if PCA were performed on the covariance matrix instead of the correlation matrix. Why did the results change?

Loading Matrix (Covariance)

The loading matrix linked above and the biplot below show the results if the principle component analysis was performed on the covariance matrix instead of the correlation matrix. These are vastly different from the results above because of the different variable's units. The covariance matrix can be used with homogenous data, but because the candy bar data included measurements in oz, grams, milligrams, etc., we need to use a method that standardizes the data (i.e. the correlation matrix).



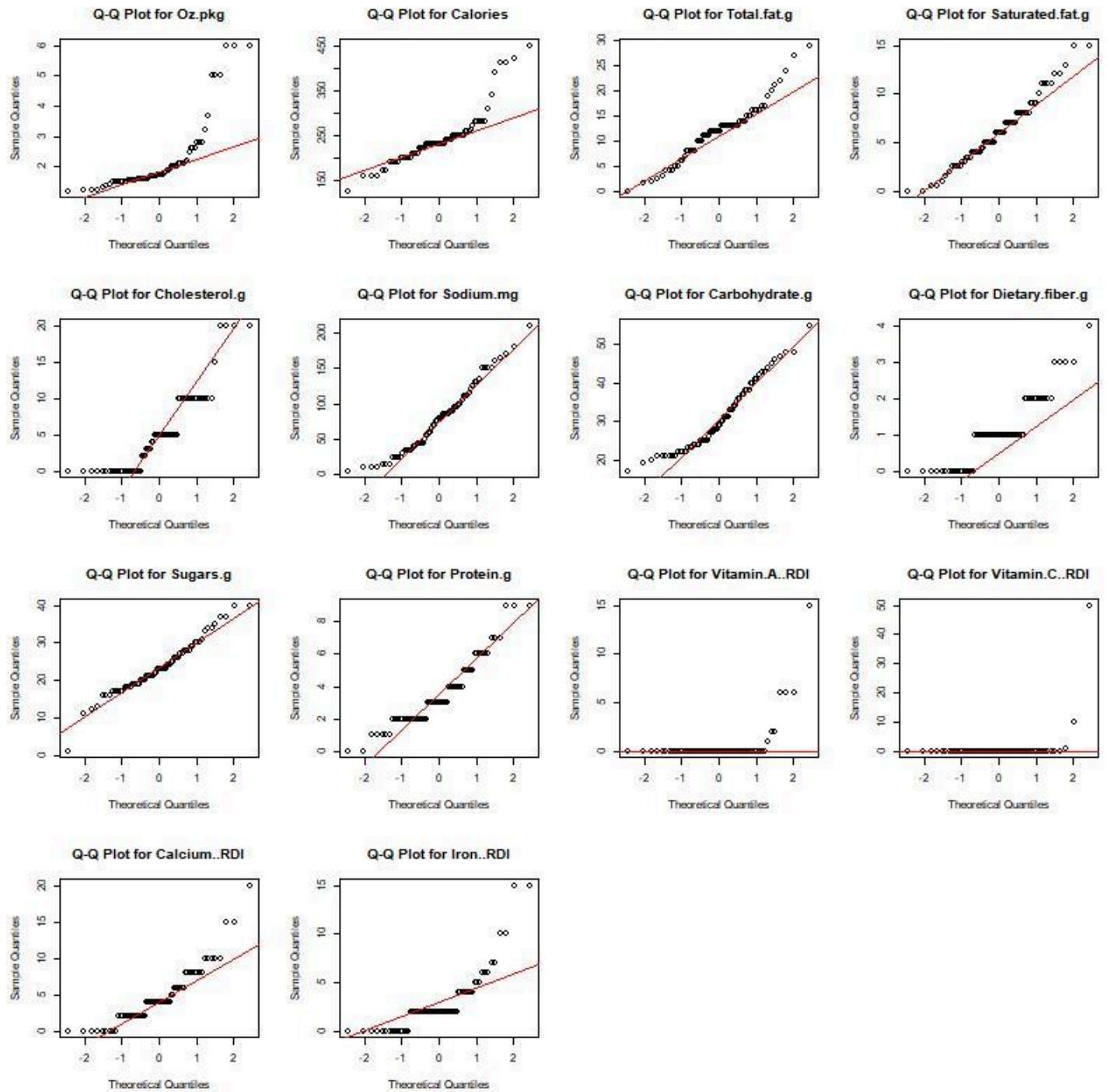
Variance-Covariance Matrix

	Oz.pkg	Calories	Total.fat.g	Saturated.fat.g	Cholesterol.g	Sodium.mg	Carbohydrate.g	Dietary.fiber.g	Sugars.g	Protein.g	Vitamin.A..RDI	Vitamin.C..RDI	Calcium..RDI	Iron..RDI
Oz.pkg	1.312	6.853	0.993	0.803	1.039	-6.881	-1.045	0.109	0.059	0.113	-0.199	-0.856	0.998	-0.412
Calories	6.853	3699.939	277.532	134.768	47.063	375.383	213.615	25.77	179.781	66.889	-29.72	-98.382	31.988	10.868
Total.fat.g	0.993	277.532	31.622	15.082	12.129	8.102	-9.549	2.465	1.88	7.163	-1.339	-8.148	6.232	1.485
Saturated.fat.g	0.803	134.768	15.082	11.611	8.374	-28.5	-2.022	1.004	4.225	1.139	-1.169	-4.924	3.536	-0.181
Cholesterol.g	1.039	47.063	12.129	8.374	29.244	-10.288	-14.526	-0.537	2.725	-0.377	1.153	-4.758	5.532	-2.508
Sodium.mg	-6.881	375.383	8.102	-28.5	-10.288	2299.957	59.658	-6.938	-2.958	26.29	16.633	-16.259	-33.965	-24.244
Carbohydrate.g	-1.045	213.615	-9.549	-2.022	-14.526	59.658	77.399	-0.25	43.757	-3.896	-4.842	-8.789	-9.742	-3.097
Dietary.fiber.g	0.109	25.77	2.465	1.004	-0.537	-6.938	-0.25	0.803	-0.259	0.86	-0.32	0.523	0.253	0.92
Sugars.g	0.059	179.781	1.88	4.225	2.725	-2.958	43.757	-0.259	48.956	-3.37	-3.311	-10.462	-2.906	-4.698
Protein.g	0.113	66.889	7.163	1.139	-0.377	26.29	-3.896	0.86	-3.37	4.165	0.569	2.207	3.068	2.236
Vitamin.A..RDI	-0.199	-29.72	-1.339	-1.169	1.153	16.633	-4.842	-0.32	-3.311	0.569	4.722	6.138	3.211	3.535
Vitamin.C..RDI	-0.856	-98.382	-8.148	-4.924	-4.758	-16.259	-8.789	0.523	-10.462	2.207	6.138	37.457	12.677	10.751
Calcium..RDI	0.998	31.988	6.232	3.536	5.532	-33.965	-9.742	0.253	-2.906	3.068	3.211	12.677	15.012	4.868
Iron..RDI	-0.412	10.868	1.485	-0.181	-2.508	-24.244	-3.097	0.92	-4.698	2.236	3.535	10.751	4.868	9.31

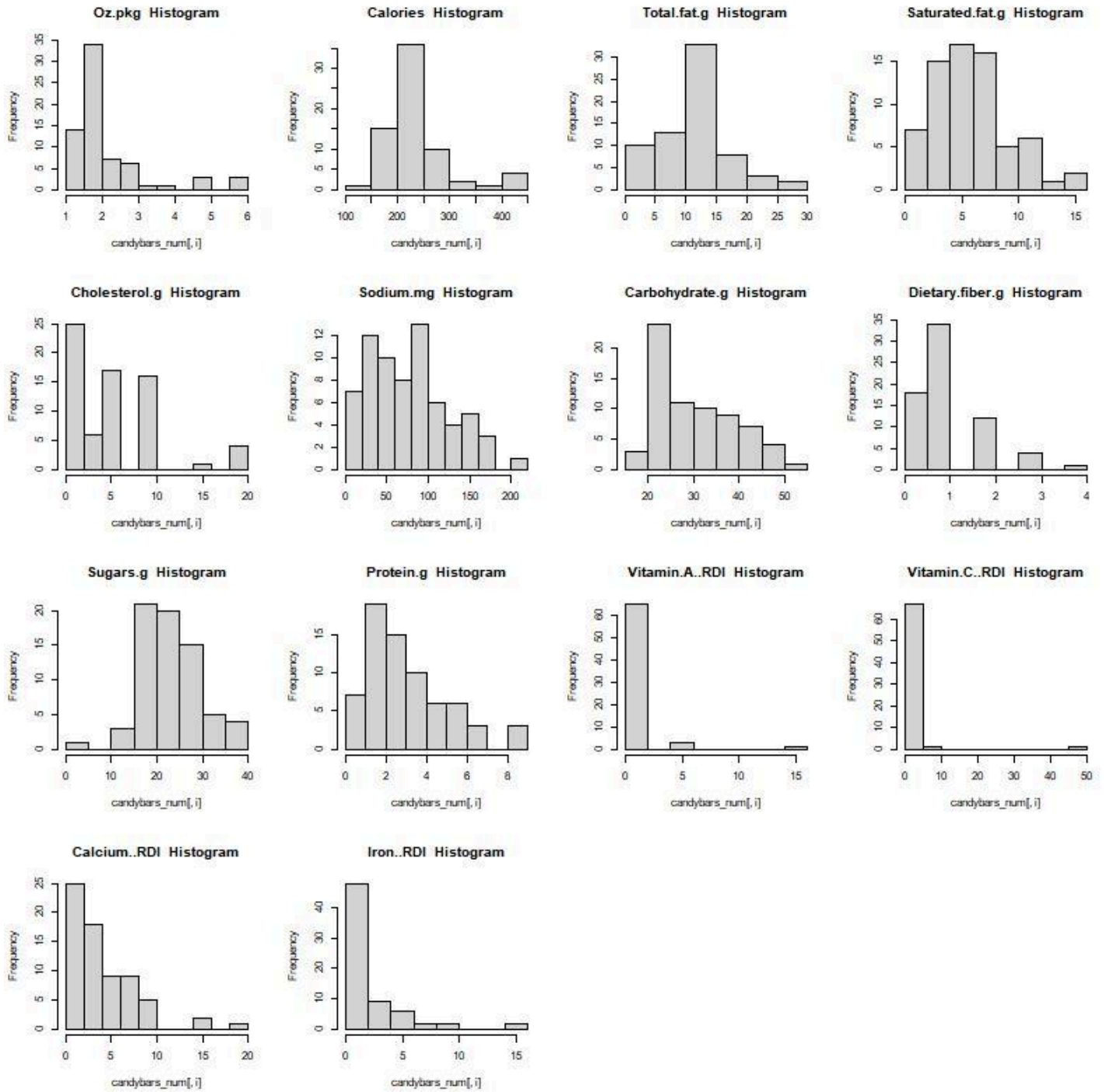
Correlation Matrix

	Oz.pkg	Calories	Total.fat.g	Saturated.fat.g	Cholesterol.g	Sodium.mg	Carbohydrate.g	Dietary.fiber.g	Sugars.g	Protein.g	Vitamin.A..RDI	Vitamin.C..RDI	Calcium..RDI	Iron..RDI
Oz.pkg	1	0.098	0.154	0.206	0.168	-0.125	-0.104	0.106	0.007	0.048	-0.08	-0.122	0.225	-0.118
Calories	0.098	1	0.811	0.65	0.143	0.129	0.399	0.473	0.422	0.539	-0.225	-0.264	0.136	0.059
Total.fat.g	0.154	0.811	1	0.787	0.399	0.03	-0.193	0.489	0.048	0.624	-0.11	-0.237	0.286	0.087
Saturated.fat.g	0.206	0.65	0.787	1	0.454	-0.174	-0.067	0.329	0.177	0.164	-0.158	-0.236	0.268	-0.017
Cholesterol.g	0.168	0.143	0.399	0.454	1	-0.04	-0.305	-0.111	0.072	-0.034	0.098	-0.144	0.264	-0.152
Sodium.mg	-0.125	0.129	0.03	-0.174	-0.04	1	0.141	-0.161	-0.009	0.269	0.16	-0.055	-0.183	-0.166
Carbohydrate.g	-0.104	0.399	-0.193	-0.067	-0.305	0.141	1	-0.032	0.711	-0.217	-0.253	-0.163	-0.286	-0.115
Dietary.fiber.g	0.106	0.473	0.489	0.329	-0.111	-0.161	-0.032	1	-0.041	0.47	-0.164	0.095	0.073	0.336
Sugars.g	0.007	0.422	0.048	0.177	0.072	-0.009	0.711	-0.041	1	-0.236	-0.218	-0.244	-0.107	-0.22
Protein.g	0.048	0.539	0.624	0.164	-0.034	0.269	-0.217	0.47	-0.236	1	0.128	0.177	0.388	0.359
Vitamin.A..RDI	-0.08	-0.225	-0.11	-0.158	0.098	0.16	-0.253	-0.164	-0.218	0.128	1	0.462	0.381	0.533
Vitamin.C..RDI	-0.122	-0.264	-0.237	-0.236	-0.144	-0.055	-0.163	0.095	-0.244	0.177	0.462	1	0.535	0.576
Calcium..RDI	0.225	0.136	0.286	0.268	0.264	-0.183	-0.286	0.073	-0.107	0.388	0.381	0.535	1	0.412
Iron..RDI	-0.118	0.059	0.087	-0.017	-0.152	-0.166	-0.115	0.336	-0.22	0.359	0.533	0.576	0.412	1

Univariate Normal Probability Plots



Histograms



Importance of components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14
Eigenvalues	3.524396	3.014941	1.817274	1.391856	1.306887	0.917613	0.584278	0.44807	0.361236	0.25767	0.169951	0.15218	0.05047	0.003171
Standard Deviation	1.8773	1.7364	1.3481	1.17977	1.14319	0.95792	0.76438	0.66938	0.6010	0.50762	0.41225	0.39010	0.22466	0.05632
Proportion of Variance	0.2517	0.2154	0.1298	0.09942	0.09335	0.06554	0.01173	0.03201	0.0258	0.01841	0.01214	0.01087	0.00361	0.00023
Cumulative Proportion	0.2517	0.4671	0.5969	0.69632	0.78967	0.85521	0.89695	0.92895	0.9547	0.97316	0.98530	0.99617	0.99977	1.0000

Loading Matrix (Correlation Matrix)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14
Oz.pkg	0.25976158	0.05258692	-0.38028943	-0.15745924	-0.1718199952	-0.819068519	-0.226883211	0.027491720	2.583133e-02	0.056480462	0.037842709	-0.049088627	-0.004267099	-8.152703e-04
Calories	0.83872229	0.36688667	0.33843734	-0.03021684	0.1229024988	-0.006251896	0.005319195	-0.092195258	1.789973e-03	0.067851552	-0.080207284	-0.076975179	-0.065744502	3.906503e-02
Total.fat.g	0.95046787	0.08599579	-0.09579008	0.16798628	0.0688191868	0.120571046	-0.005070999	-0.082652487	2.644231e-03	-0.032856738	0.062032686	-0.067346942	-0.123262747	-3.202942e-02
Saturated.fat.g	0.79276399	0.22195442	-0.27501494	-0.18544899	-0.0046951243	0.203615711	-0.051293065	-0.051473169	3.878635e-01	-0.022625433	0.040432048	-0.003702616	0.113473343	-6.263470e-04
Cholesterol.g	0.37243232	0.03582563	-0.65844331	-0.19777830	0.3974611211	0.197875000	-0.030501708	0.319700073	-2.207651e-01	0.182408578	-0.067636182	-0.005858714	0.021433189	1.191338e-04
Sodium.mg	-0.04256518	0.07854961	0.28005937	0.51994342	0.7057427226	-0.218024827	-0.016261487	0.202772487	1.848674e-01	0.027891899	0.080071921	0.121562386	-0.010554327	8.920102e-04
Carbohydrate.g	-0.10217797	0.59118532	0.63665615	-0.37150796	0.1217924858	-0.109445436	-0.016170507	-0.009263486	4.933201e-02	0.164751260	-0.197099244	-0.003263468	0.023176280	-2.405879e-02
Dietary.fiber.g	0.60807957	-0.08759741	0.32936411	0.14143178	-0.5208355377	0.020815213	-0.156914856	0.389527154	-2.902737e-02	-0.153562250	-0.095888124	0.111354879	0.002642215	-5.296122e-04
Sugars.g	0.10148575	0.62917795	0.30330894	-0.56985763	0.2214957773	-0.024105700	0.009046981	0.085662111	-1.796770e-01	-0.204365818	0.213570285	0.005331140	0.018706032	7.743632e-04
Protein.g	0.66227921	-0.35699501	0.31182496	0.41727886	0.0947284944	-0.169339746	0.164561783	-0.114628571	-2.480604e-01	-0.001390448	0.005428494	-0.086608650	0.124943714	-6.107131e-03
Vitamin.A..RDI	-0.06955087	-0.69225412	0.05235639	-0.17558720	0.4714448076	0.039863525	-0.425322286	-0.090056445	9.929815e-05	-0.219113152	-0.128319226	-0.070945790	0.002509861	-5.660114e-04
Vitamin.C..RDI	-0.08750881	-0.77131964	0.27402711	-0.28993160	0.0006426946	-0.047105956	0.282103381	0.289222504	1.684768e-01	0.022666421	0.043504801	-0.203928765	-0.020684097	-1.985643e-04
Calcium..RDI	0.43593769	-0.59947910	-0.12407194	-0.41538179	0.1475000193	-0.192842540	0.357383212	-0.128679523	2.453463e-03	-0.066240198	-0.078096703	0.219954866	-0.023421230	-2.427688e-05
Iron..RDI	0.22233440	-0.70671681	0.40627132	-0.22308735	-0.1153264123	0.146116693	-0.297778366	-0.073453253	-4.150409e-02	0.262361166	0.165755319	0.097692092	-0.001313441	7.366969e-04

Loading Matrix (Covariance Matrix)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14
Oz.pkg	0.002474042	0.006573849	-0.01982236	-0.026471301	-0.01944828	0.00568717	0.111000218	0.055719481	-0.002135683	-0.02826316	0.392771799	0.060704584	-0.0340027715	-0.0064106114
Calories	1.811350695	0.406763167	-0.04984618	0.039615976	0.04147331	-0.01338253	-0.002336156	-0.005231214	-0.022913162	0.01222023	0.005063780	-0.004154563	-0.0015357855	0.0049622674
Total.fat.g	0.133093020	0.046130786	-0.39711636	-0.148951477	-0.04881821	0.17175581	-0.050079809	-0.061966436	0.071197548	-0.01024762	-0.064094048	0.194923411	0.0103997748	-0.0419291149
Saturated.fat.g	0.061093656	0.047288487	-0.14737403	-0.136319253	-0.14214775	-0.02009401	0.115566191	-0.119124556	0.513190848	-0.03639162	0.012717382	-0.155148638	-0.0162922259	-0.0035056371
Cholesterol.g	0.021416219	0.017305552	-0.28066299	-0.414277232	-0.76720700	-0.53411710	-0.112942258	-0.030796939	-0.120801332	-0.03235304	0.005154190	-0.021848564	0.0024744825	-0.0001950406
Sodium.mg	0.444601716	-1.685660154	-0.02344814	-0.002047387	-0.02132594	0.01119717	0.003364846	0.001387596	0.010408985	-0.00884177	0.000609515	0.001305215	0.0008086006	0.0002091439
Carbohydrate.g	0.110397388	-0.006135319	1.01717120	0.186952669	0.01852267	-0.45386331	0.067107074	0.054947054	0.069446197	-0.01301131	-0.022177613	0.026835088	0.0074062234	-0.0210849686
Dietary.fiber.g	0.011432393	0.010107078	-0.02698088	0.036208895	0.04274254	0.02895888	-0.053350932	-0.054747538	-0.011204907	-0.06408972	0.074952383	-0.066823101	0.2132861278	-0.0038574622
Sugars.g	0.086994274	0.036130001	0.64574023	-0.185839817	-0.62980058	0.61715765	-0.103140204	-0.021273115	-0.019653169	-0.02385292	0.006503689	-0.015329255	-0.0029103161	0.0005923367
Protein.g	0.034657429	-0.007564463	-0.14022320	0.122619292	0.07308180	0.09340311	-0.011376732	0.073177764	-0.213984413	-0.02374513	0.006478799	-0.280623093	-0.0418986597	-0.0293071578
Vitamin.A..RDI	-0.012433146	-0.018208735	-0.07486717	0.147708378	-0.13529752	-0.01068016	-0.248927082	0.264944085	0.108571456	0.40979878	0.032682741	-0.018040188	0.0210544114	-0.0029076650
Vitamin.C..RDI	-0.049558190	-0.006037642	-0.14178705	0.914688678	-0.36618915	-0.02228105	0.002791469	-0.346146365	-0.023737064	0.04680192	0.018625941	0.018312294	-0.0105816363	-0.0003037805
Calcium..RDI	0.011003690	0.032196276	-0.22101294	0.321342697	-0.36939326	0.06558805	0.464815659	0.399919066	0.014464398	-0.07440192	-0.060284054	0.027708295	0.0264155213	0.0011514488
Iron..RDI	0.002048084	0.020510980	-0.09535022	0.360025334	-0.01129146	-0.02722441	-0.497169616	0.253347990	0.117841592	-0.27037060	0.010759339	0.018281238	-0.0192057114	0.0026461064