Principal Component Analysis

Principle Component Analysis: A statistical technique used to examine the interrelations among a set of variables in order to identify the underlying structure of those variables. Also called *factor analysis*.

It is a non-parametric analysis and the answer is unique and independent of any hypothesis about data distribution.

These two properties can be regarded as weaknesses as well as strengths.

Since the technique is non-parametric, no prior knowledge can be incorporated.

PCA data reduction often incurs a loss of information.

The assumptions of PCA:

1. Linearity
   - Assumes the data set to be linear combinations of the variables.

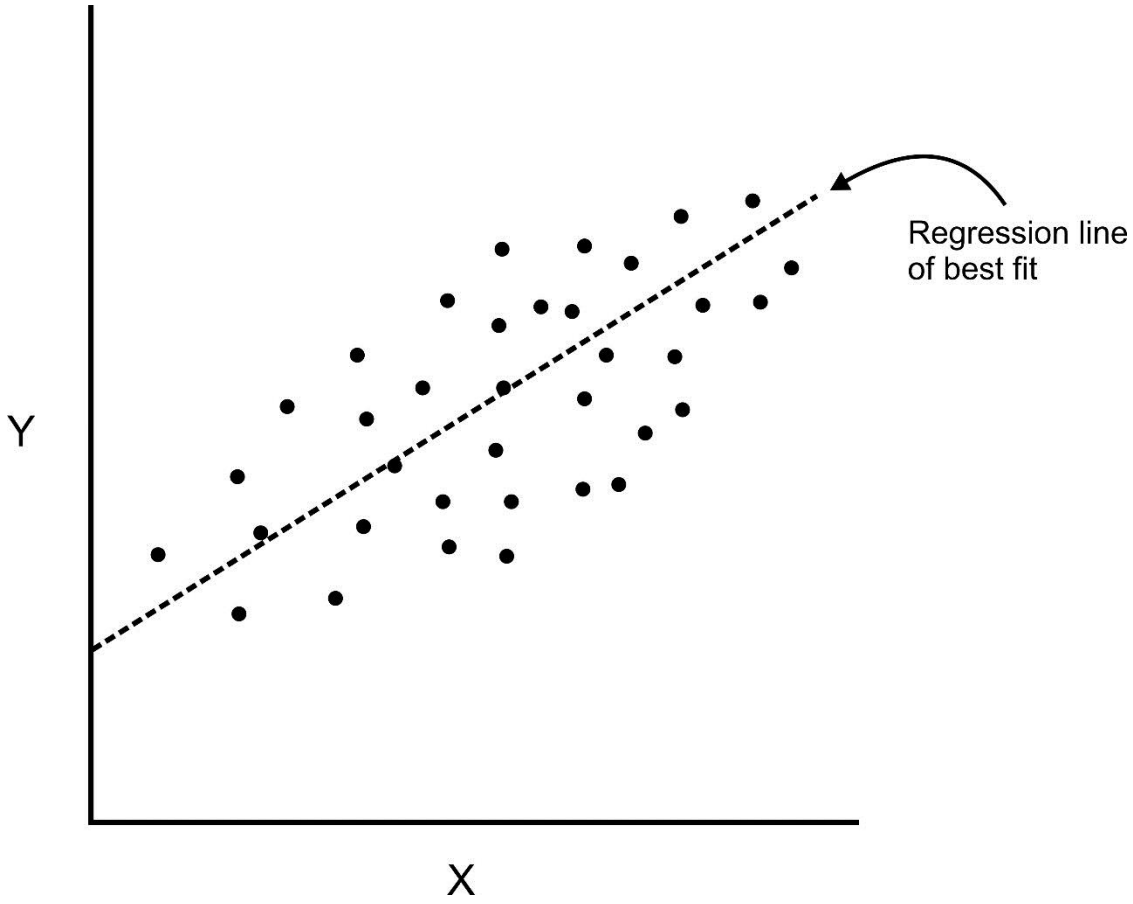
2. The importance of mean and covariance
   - There is no guarantee that the directions of maximum variance will contain good features for discrimination.

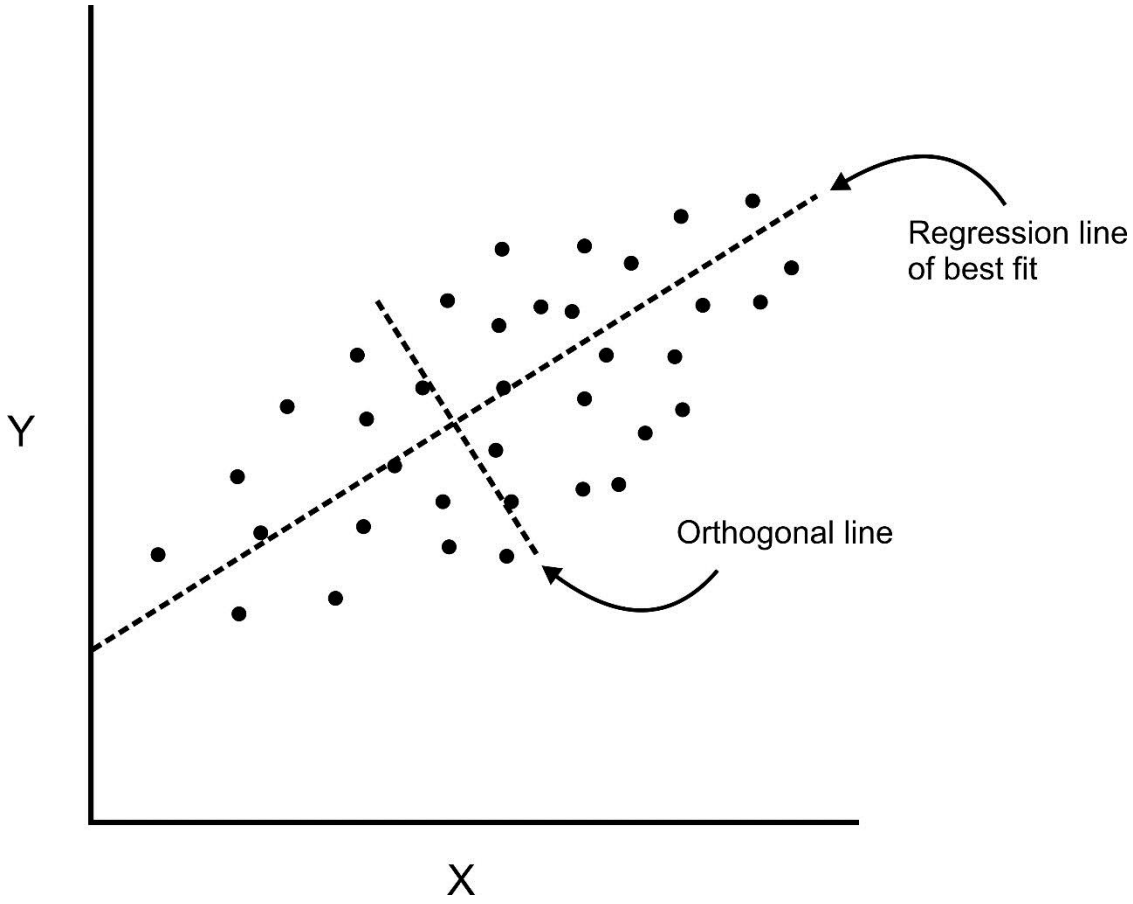
3. That large variances have important dynamics
   - Assumes that components with larger variance correspond to interesting dynamics and lower ones correspond to noise.

Where regression determines a line of best fit to a data set, factor analysis determines several orthogonal lines of best fit to the data set.

Orthogonal: meaning “at right angles”. Actually the lines are perpendicular to each other in *n*-dimensional space.

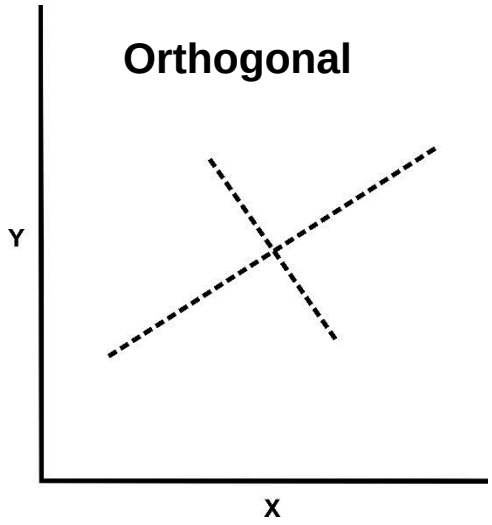
n-Dimensional Space: the variable sample space. There are as many dimensions as there are variables, so in a data set with 4 variables the sample space is 4-dimensional.

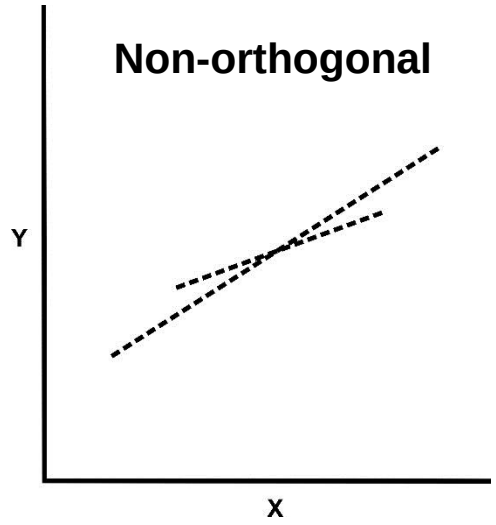

Regression line
of best fit
Y
X

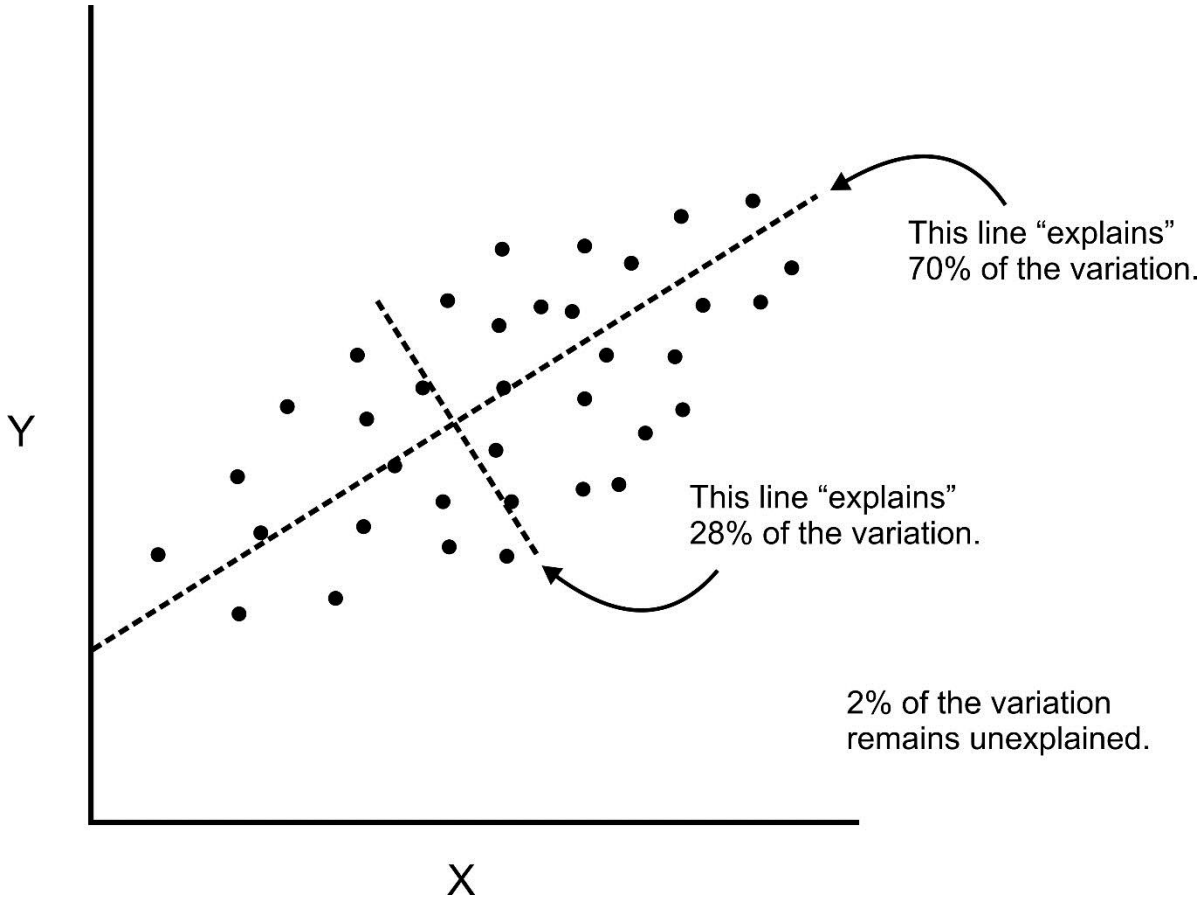

Regression line
of best fit
Orthogonal line
Y
X

Components: a linear transformation that chooses a variable system for the data set such that the greatest variance of the data set comes to lie on the first axis (then called the *principal* component), the second greatest variance on the second axis, and so on ...

Note that components are uncorrelated, since in the sample space they are orthogonal to each other.



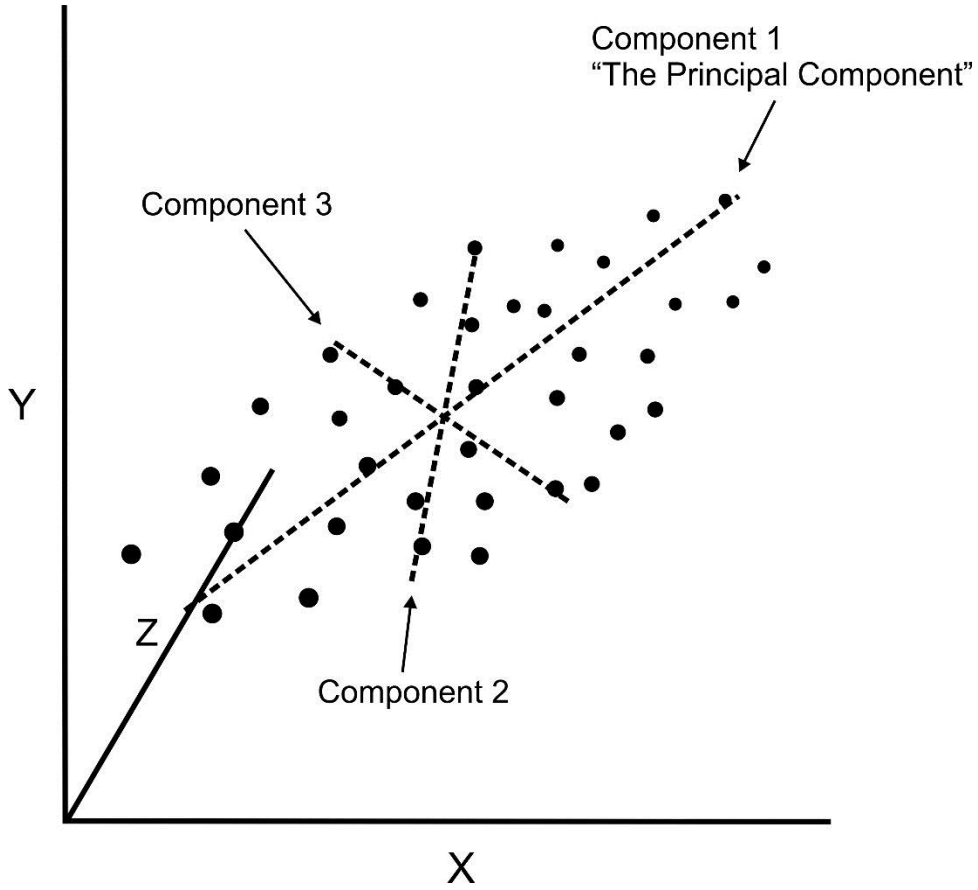



This line “explains”
70% of the variation.
This line “explains”
28% of the variation.
2% of the variation
remains unexplained.
Y
X

Locations along each component (or *eigenvector*) are then associated with values across all variables. This association between the components and the original variables is called the component's *eigenvalue*.

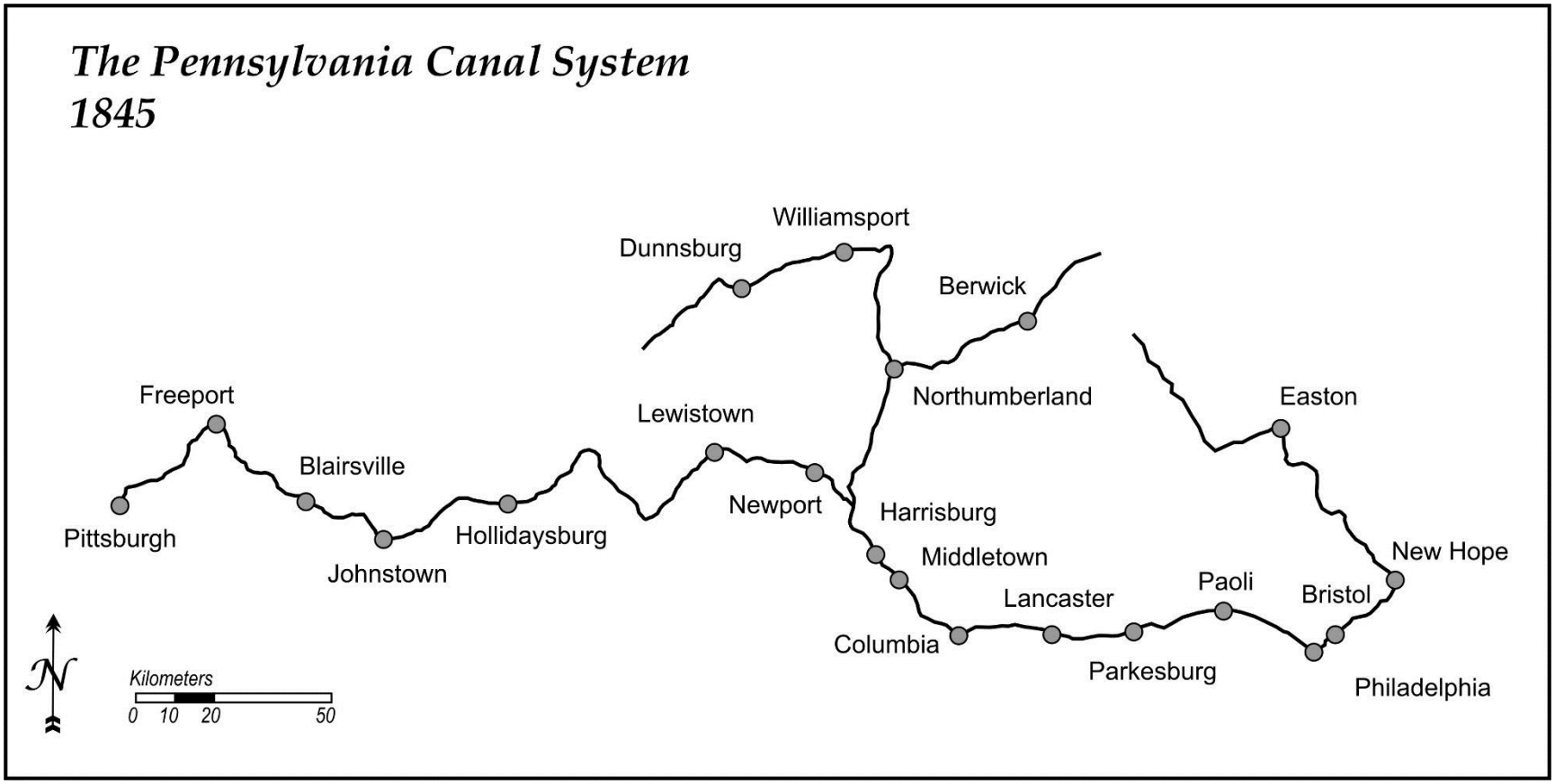
In multivariate (multiple variable) space, the correlation between the component and the original variables is called the *component loadings*.

Component loadings: analogous to correlation coefficients, squaring them give the amount of explained variation. Therefore the component loadings tell us how much of the variation in a variable is explained by the component.


Component 1
“The Principal Component”
Component 3
Y
Z
Component 2
X

If we use this technique on a data set with a large number of variables, we can compress the amount of explained variation to just a few components.

What follows is an example of Principal Component Analysis using canal town commodity production figures (percentage of total production) for 1845.


The Pennsylvania Canal System
1845
Williamsport
Dunnsburg
Berwick
Freeport
Northumberland
Easton
Lewistown
Blairsville
Newport
Harrisburg
Pittsburgh
Hollidaysburg
Middletown
New Hope
Johnstown
Paoli
Lancaster
Bristol
Columbia
Parkesburg
Philadelphia
N
Kilometers
0 10 20 50

Towns

Columbia
Middletown
Harrisburg
Newport
Lewistown
Hollidaysburg
Johnstown
Blairsville
Pittsburgh
Dunnsburg
Williamsport
Northumberland
Berwick
Easton
New Hope
Bristol
Philadelphia
Paoli
Parkesburg
Lancaster

Variables

Corn
Wheat
Flour
Whiskey
Groceries
Dry Goods

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.533	42.211	42.211	2.533	42.211	42.211	1.887	31.452	31.452
2	1.565	26.084	68.295	1.565	26.084	68.295	1.880	31.328	62.780
3	1.504	25.073	93.368	1.504	25.073	93.368	1.835	30.587	93.368
4	.174	2.901	96.269						
5	.119	1.988	98.257						
6	.105	1.743	100.000						

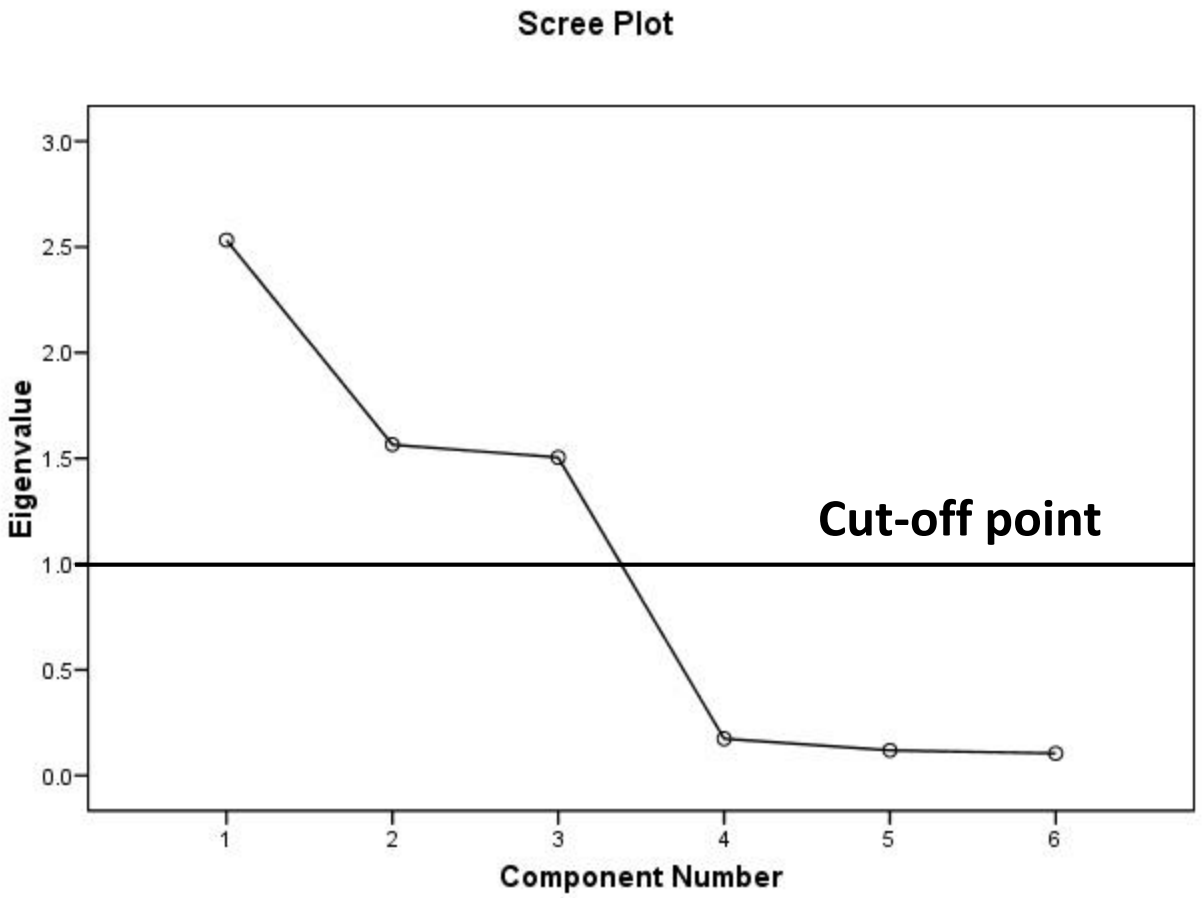
Extraction Method: Principal Component Analysis.

In this case, 3 components contain 93.368% of the variation of the 6 original variables. Note that there are as many components as original input variables.

Component 1 explains 42.211% of the variation, component 2 explains 26.084%, and component 3 explains 25.073%.

The remaining 3 components explain only 6.632%.

A scree plot graphs the amount of variation explained by each component.



Rotated Component Matrix (a)

	Component		
	1	2	3
Corn	-.065	.936	.214
Wheat	-.104	.952	-.057
Groceries	.962	-.092	-.086
Dry Goods	.963	-.074	-.092
Flour	-.126	-.097	.954
Whiskey	-.057	.275	.927

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

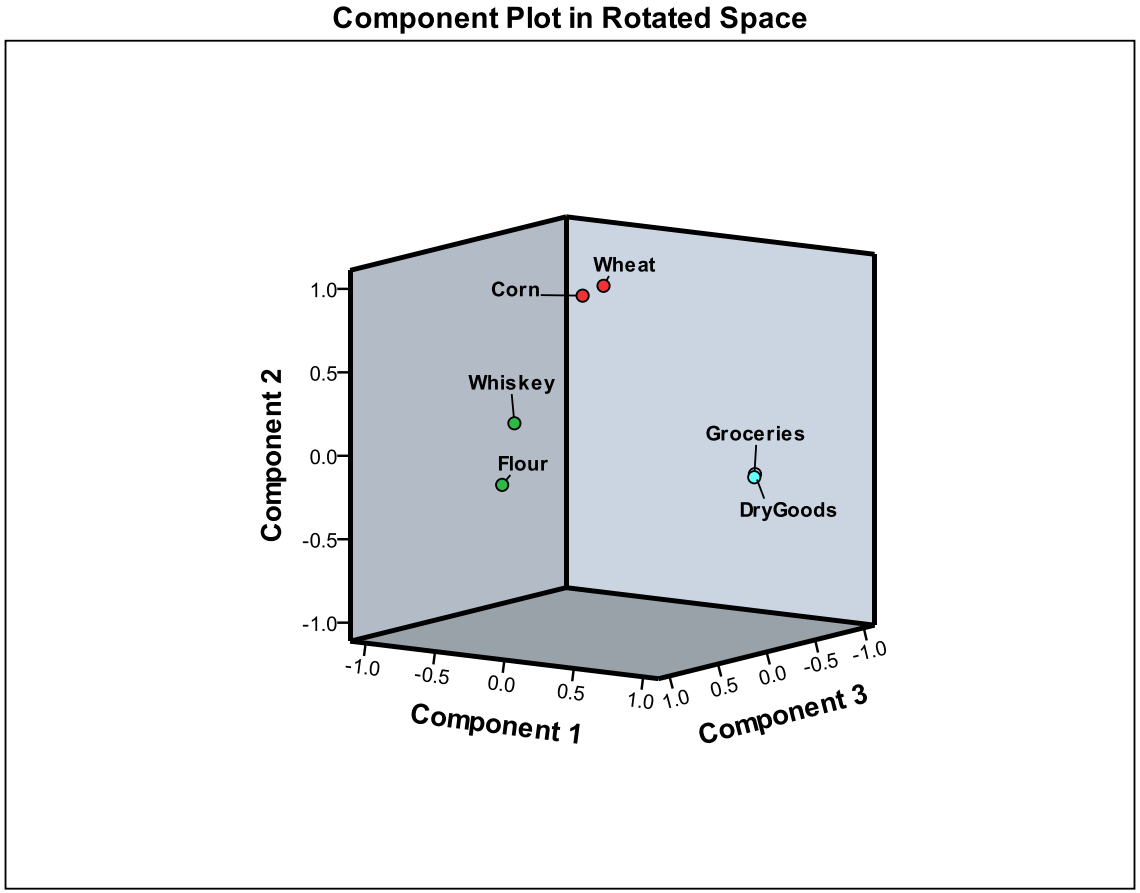
a. Rotation converged in 4 iterations.

Highest Component Loading

Component 1: Groceries and dry goods.

Component 2: Corn and wheat.

Component 3: Flour and whiskey.



Note how the variables that make up each component fall close to each other in the 3-dimensional sample space.

What do these components mean (how do we interpret them)?

- *Component 1 (groceries and dry goods)* – these two items are highly processed and value added.

- *Component 2 (corn and wheat)* – these two items are not processed (raw) and have no value added.

- *Component 3 (flour and whiskey)* – these two items are moderately processed and value added.

It appears that the components are indicators of either the amount of processing or value adding (or both).

The most challenging part of PCA is interpreting the components.

1. The higher the component loadings, the more important that variable is to the component.

2. Combinations of positive and negative loadings are interpreted as ‘mixed’.

3. The specific sign of the is not important.

4. ALWAYS use the ROTATED component matrix!!

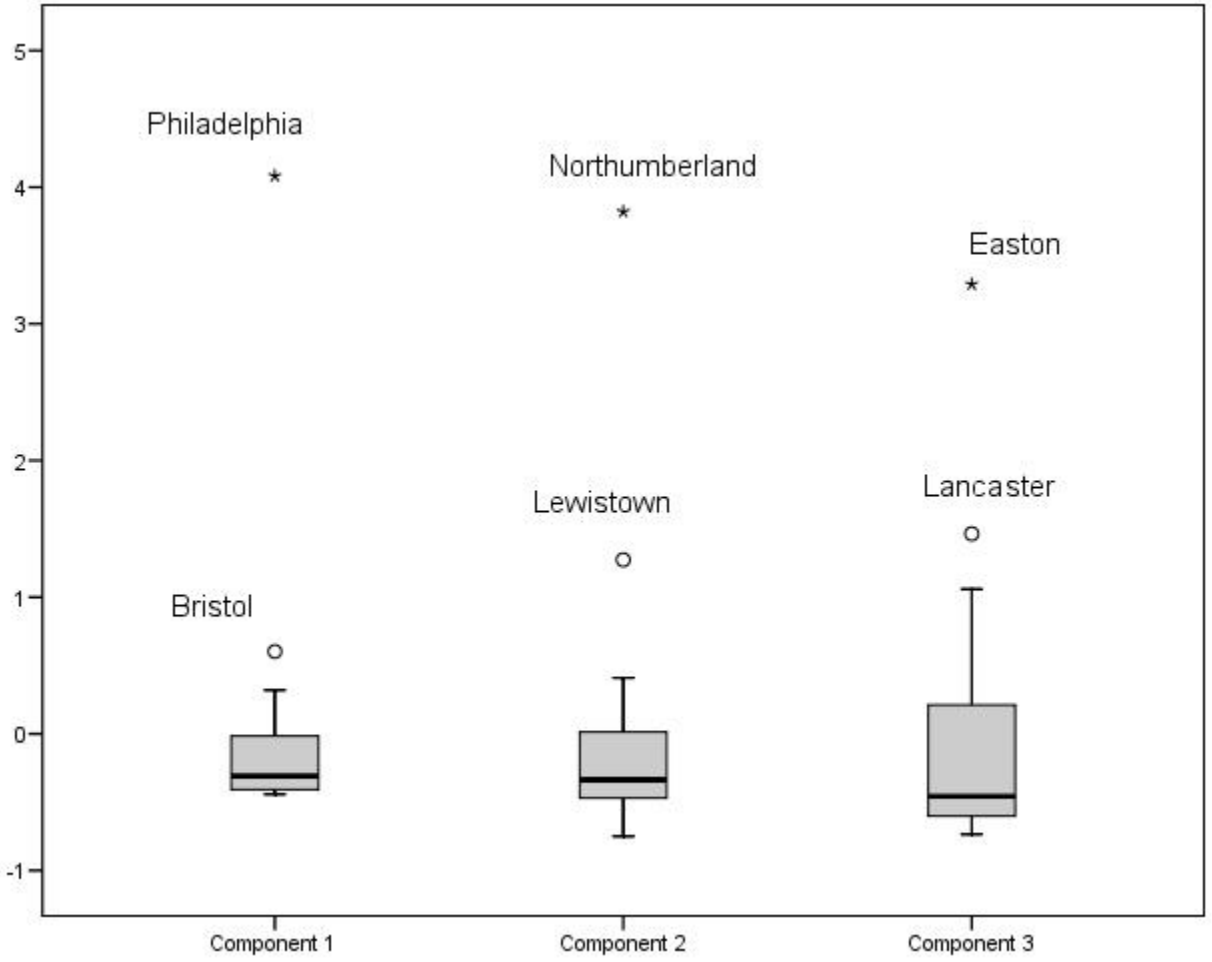
Component score: the new variable value based on the observation's component loading and the standardized value of the original variable, summed over all variables.

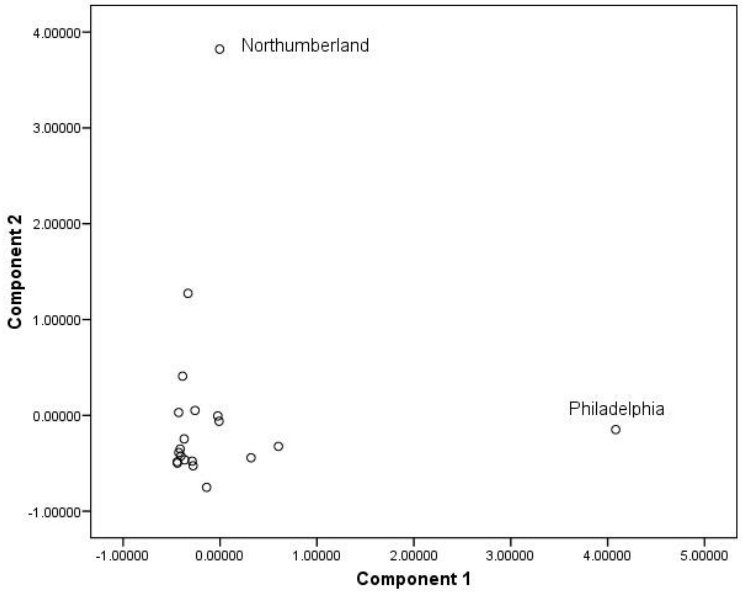
$$Score_{ik} = \sum D_{ij} L_{jk}$$

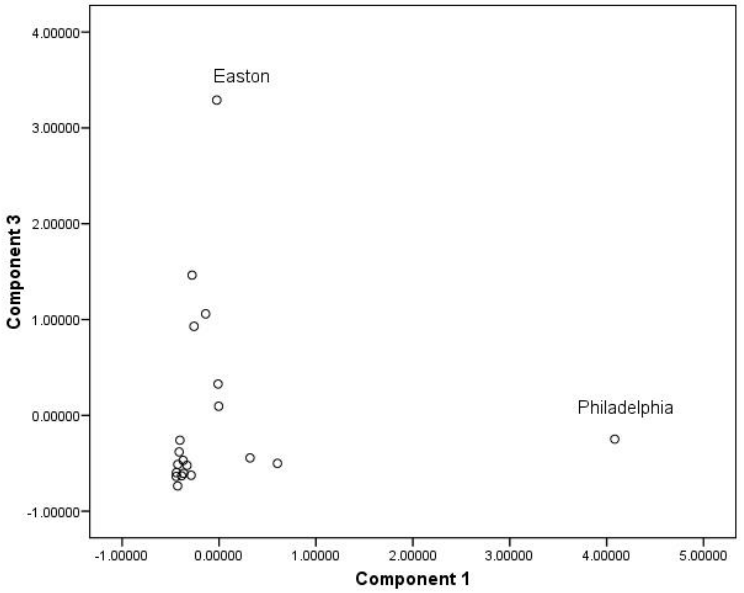
where D_{ij} is the standardized value for observation *i* on variable *j* and L_{jk} is the loading of variable *j* on component *k*.

Examining the component scores for each town may give some clues as to the interpretation of the components.

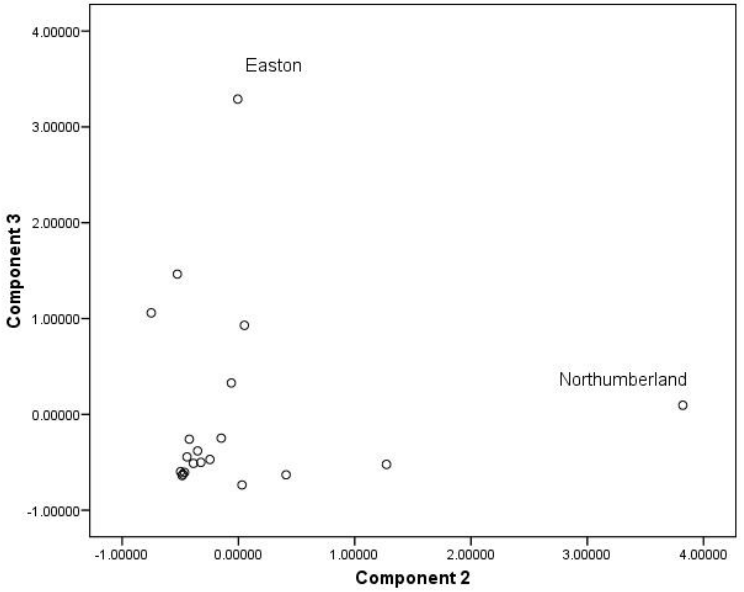
Component Score Box Plot







Easton, Philadelphia, and Northumberland are the only towns that load highly on a single component.



Scoring highly on a single component simply means that the original variable values for these locations are overwhelmingly explained by a single component.

In this case, it means that the variation among ALL of the variables for Philadelphia (for example) is more completely explained by a single component composed of groceries and dry goods.

Rotated Component Matrix[a]

	Component		
	1	2	3
Corn	-.065	.936	.214
Wheat	-.104	.952	-.057
Groceries	.962	-.092	-.086
Dry Goods	.963	-.074	-.092
Flour	-.126	-.097	.954
Whiskey	-.057	.275	.927

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 4 iterations.

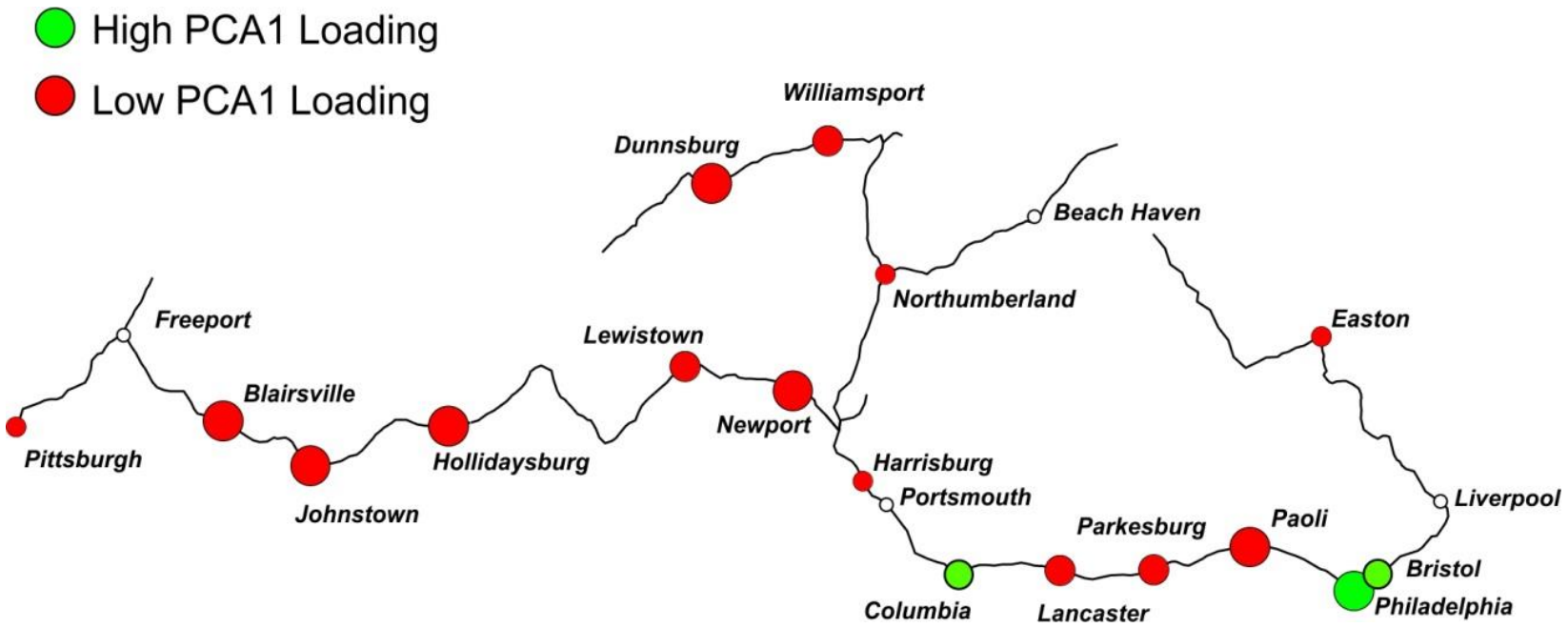
Town Component Scores

Town	Component 1	Component 2	Component 3
Columbia	0.31989	-0.44216	-0.44369
Middletown	-0.37101	-0.24531	-0.47020
Harrisburg	-0.00974	-0.06105	0.32792
Newport	-0.38678	0.40935	-0.62996
Lewistown	-0.33132	1.27318	-0.52170
Hollidaysburg	-0.44018	-0.49770	-0.59722
Johnstown	-0.44188	-0.48447	-0.63736
Blairsville	-0.42552	-0.38759	-0.51107
Pittsburgh	-0.13834	-0.75021	1.05942
Dunnsburg	-0.42728	0.03072	-0.73622
Williamsport	-0.28812	-0.47716	-0.62453
Northumberland	-0.00398	3.82169	0.09538
Berwick	-0.36503	-0.46398	-0.60501
Easton	-0.02349	-0.00587	3.28970
New Hope	-0.40354	-0.42291	-0.25891
Bristol	0.60267	-0.32311	-0.50086
Philadelphia	4.08309	-0.14799	-0.24733
Paoli	-0.41174	-0.35103	-0.38109
Parkesburg	-0.25890	0.05125	0.92910
Lancaster	-0.27880	-0.52566	1.46363

← **Middletown is a 'mixed' town because it loads on all components equally.**

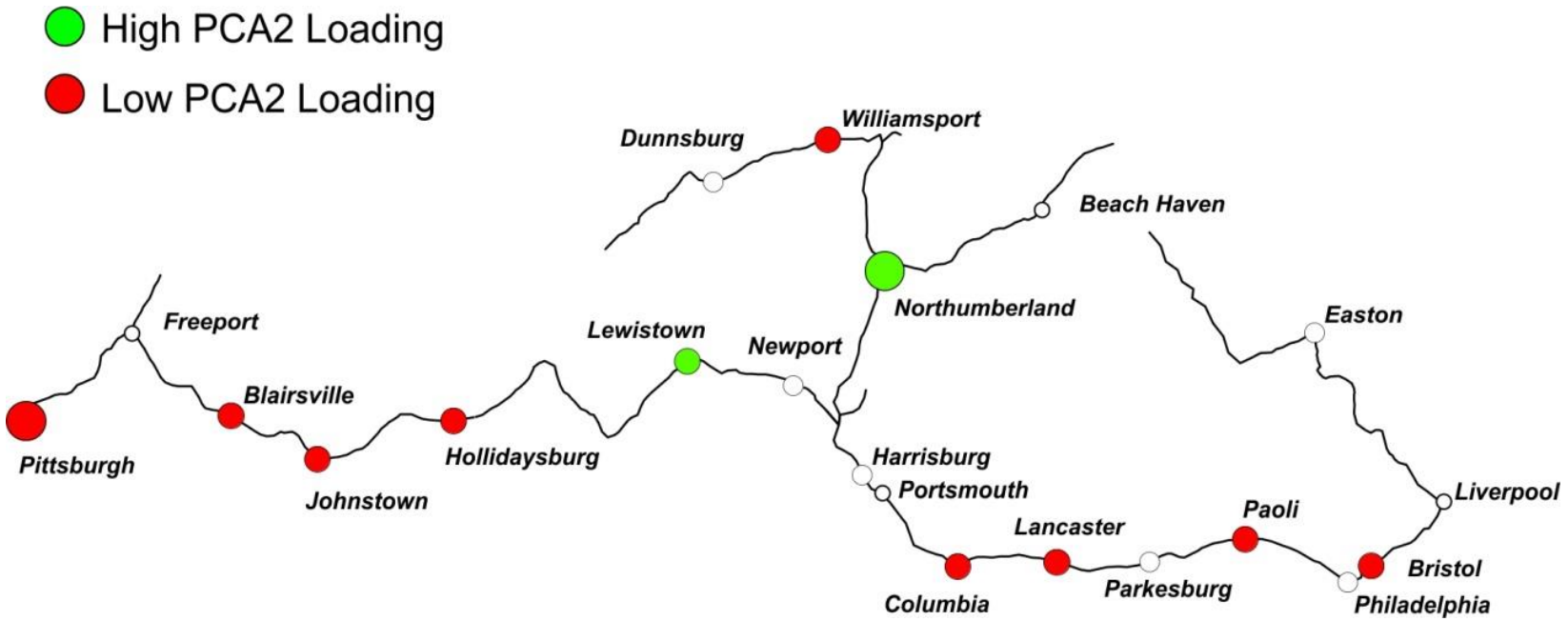
← **Philly is a 'processed goods 'town.**

Component 1: Processed Goods



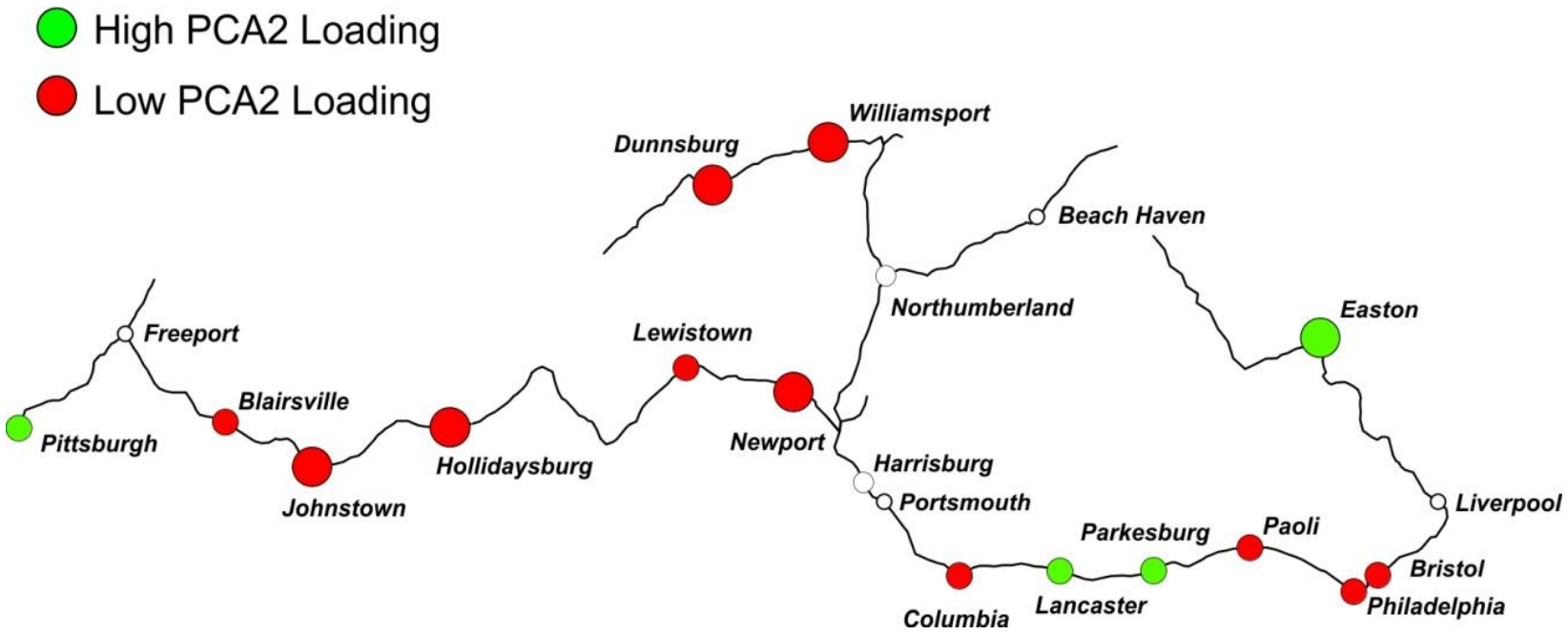
The *green* town were producers of <u>processed goods</u>, while the *red* towns were consumers of those goods.

Component 2: Non-Processed Goods



The *green* town were producers of <u>non-processed</u> goods, while the *red* towns were consumers of those goods.

Component 3: Partially Processed Goods



The *green* town were producers of <u>partially processed</u> goods, while the *red* towns were consumers of those goods.

What information did PCA provide concerning the goods exported by the canal towns?

- The goods fell into recognizable categories (highly processed, moderately processed, not processed).
- A small number of towns were responsible for exporting most of these goods.
- The location of these towns relative to the goods they produced make sense.
    - Industrial towns on the Columbia railroad exported finished goods.
    - Small farming towns on the canal exported produce.
    - Midsize towns exported moderately processed goods.

Without the use of Principal Component Analyses these associations would be difficult to determine.

Principal Component Analyses is also used to remove correlation among independent variables that are to be used in multivariate regression analysis.

Correlation Matrix

		Corn	Wheat	Groceries	DryGoods	Flour	Whiskey
Correlation	Corn	1.000	.812	-.163	-.160	.108	.450
	Wheat	**.812**	1.000	-.183	-.157	-.096	.198
	Groceries	-.163	-.183	1.000	.883	-.191	-.164
	DryGoods	-.160	-.157	**.883**	1.000	-.198	-.163
	Flour	.108	-.096	-.191	-.198	1.000	.806
	Whiskey	.450	.198	-.164	-.163	**.806**	1.000

Correlation

	Dry Goods	Groceries	PCA 2	PCA 3
PCA 1	0.963	0.962	0.000	0.000

Note that PCA1 is highly correlated to dry goods and groceries, but uncorrelated to PCA2 and PCA3.