Text modified and adapted from text from statistics by Jim.

Principal Component Analysis Guide & Example

What is Principal Component Analysis?

Principal Component Analysis (PCA) takes a large data set with many variables per observation and reduces them to a smaller set of summary indices. These indices retain most of the information in the original set of variables. Analysts refer to these new values as principal components.

The principal components themselves are a set of new, uncorrelated variables that are linear combinations of the original variables.

Principal component analysis simplifies large data tables. With a vast sea of data, identifying the most important variables and finding patterns can be difficult. PCA's simplification can help you visualize, analyze, and recognize patterns in your data more easily.

This method is particularly beneficial when you have many variables relative to the number of observations or when the variables are highly correlated. Principal component analysis helps resolve both problems by reducing the dataset to a smaller number of independent (i.e., uncorrelated) variables.

Typically, PCA is just one step in an analytical process. For example, you can use it before training a machine learning model, using a clustering algorithm, or creating a visualization.

While PCA provides many benefits, it's crucial to realize that dimension reduction involves a tradeoff between potentially more robust models/improved classification accuracy versus reduced interpretability. Each principal component represents a mixture of the original variables. The process creates a linear combination of variables that squeezes the most explanatory power into each component rather than creating conceptual groupings of variables that make sense to us humans.

Principal Components Properties

In PCA, a component refers to a new, transformed variable that is a linear combination of the original variables. Think of them as indices that summarize the actual variables for each observation.

Each principal component (PC) captures as much information as possible in a single index. The process produces an uncorrelated set of principal components that avoids including redundant information. In other words, each index provides unique information even when the original variables are highly correlated.

The analysis produces a series of components: the first component explains the most variance, the second component explains the second most variance, and so on.

By using only the first several components that explain the greatest amount of variation, you can include most of the original information while reducing the dataset's complexity.

The procedure can produce the same number of principal components as variables in the complete dataset. However, analysts typically use a much smaller subset of components than

variables. Analysts often determine the number of PCs using statistical measures, such as the explained variance the downstream task requires.

Principal components have beneficial properties for statistical and machine learning models. Namely, components maximize the amount of unique information they contain, they are uncorrelated, and there are fewer of them. In other words, principal component analysis produces a small number of high-quality indices for your model.

Each principal component has an eigenvector and an eigenvalue.

Brief Example of Principal Component Analysis

Analysts looking for patterns and trends in stock prices have an extensive dataset containing numerous stocks along with dozens of variables, including closing price, trading volume, earnings per share, market liquidity and volatility, GDP, inflation, company earnings and revenue, dividend yield, international conditions, supply and demand factors, competition, and so on.

That's an ocean of data with a ton of variables. It'll be easy to miss the forest for the trees!

Principal components can take the multitude of variables and reduce them to the most important indices. This method finds a smaller set of values that explain most of the variation in stock prices. Importantly, PCA ranks the components by importance, helping you know which ones to focus on.

From the full set of variables, you might end up with four principal components that explain 90% of the original data. That's much easier to understand!

Factor Analysis vs PCA

As you read this article, you might think that principal component analysis sounds like factor analysis. While there are similarities, there are also stark differences.

Factor Analysis	**PCA**
Identifies latent variables that cause the observed values of outcome variables.	Reduces dimensions and produces components with optimal statistical properties.
Analysts guide the procedure to produce factors that are interpretable and applicable to the real world.	The procedure prioritizes retaining maximal information over the interpretability of the principal components.
Factors maximize conceptual understanding.	Components maximize explained variance.
Model assumes latent causal variables exist.	Model does not assume latent variables exist.
Mathematical: Decomposition of a correlation matrix where the diagonal entries are replaced by 1 — variance of a variable's factor.	Mathematical: Decomposition of the covariance matrix that does not include factors.

Mathematically and conceptually, the two analyses differ. Factor analysis incorporates conceptually understandable latent factors into the analysis, while principal component analysis focuses on producing the most statistically sound indices without considering interpretability.

While the interpretation of principal components can suffer, their statistical soundness can still help you explore, understand, and analyze your data, as you'll see in the next section.

Reasons to Use PCA

Principal component analysis aims to use the fewest components to explain the most variance. But why do you want to do that?

In today's world of big data, analysts frequently have too many variables. There's so much information that it can cause problems. That might be surprising, but it is known as the curse of dimensionality—that sounds like a great Halloween blog post!

In traditional statistical analyses and machine learning, having many dimensions (i.e., many variables or features per observation) can lead you to overfit the model. This phenomenon occurs when you include too many variables in a model. It starts to fit the noise rather than the general trends in the data. When your data contain more variables than observations, you can get particularly strange results that do not apply outside the sample data!

Additionally, if those variables are correlated (i.e., multicollinearity), they increase the model's error. Consequently, your model loses statistical power and its estimates become less precise.

In regression analysis and machine learning, variables providing redundant information can reduce the precision of a model. Too many variables cause clustering algorithms, such as K-Means Clustering, to have difficulties. Similarly, it can be hard to visualize patterns and relationships in your data.

Principal component analysis can solve these problems and prepare your data for exploration or additional statistical or machine learning tasks. Let's look at the top reasons for using PCA!

Dimensionality reduction

Reducing the number of dimensions can increase the dataset's manageability and computational efficiency. By identifying the principal components that explain the most variation in the data, PCA reduces redundant information by creating a set of entirely uncorrelated components.

When you have more features than observations, principal component analysis can reduce the number of variables yet retain most of the information. In this manner, you avoid overfit models.

And, because PCA produces uncorrelated components, it also addresses multicollinearity.

These properties are helpful for classification or regression tasks, as they can lead to faster and more accurate results. For instance, partial least squares regression uses principal component scores to fit the model rather than the original data values.

Data visualization

Creating a lower-dimensional representation of a high-dimensional dataset can help analysts visualize and understand underlying relationships in the data. Towards this end, analysts frequently use the 1st and 2nd principal components as the X and Y axes to graph the data in two dimensions and identify clusters.

Noise reduction

PCA can remove noise or other nuisance variation by identifying the principal components that explain the most variation. Typically, noise reduction occurs by eliminating components that capture only small amounts of the variance.

Outlier detection

Principal component analysis can help you identify outliers. Look for observations that have large residuals after PCA has transformed the data into principal component space.

Feature extraction

Principal component analysis can extract new features from the data that you can use for further analysis, such as classification or clustering.

Analysts use PCA as a feature selection technique by retaining only those most strongly associated with the top principal components. This process can be helpful when you have many features because it identifies those that contribute the greatest amount of unique information.

PCA is a valuable tool for data exploration, visualization, and preprocessing. It can help improve the performance of downstream tasks and make the data more interpretable.

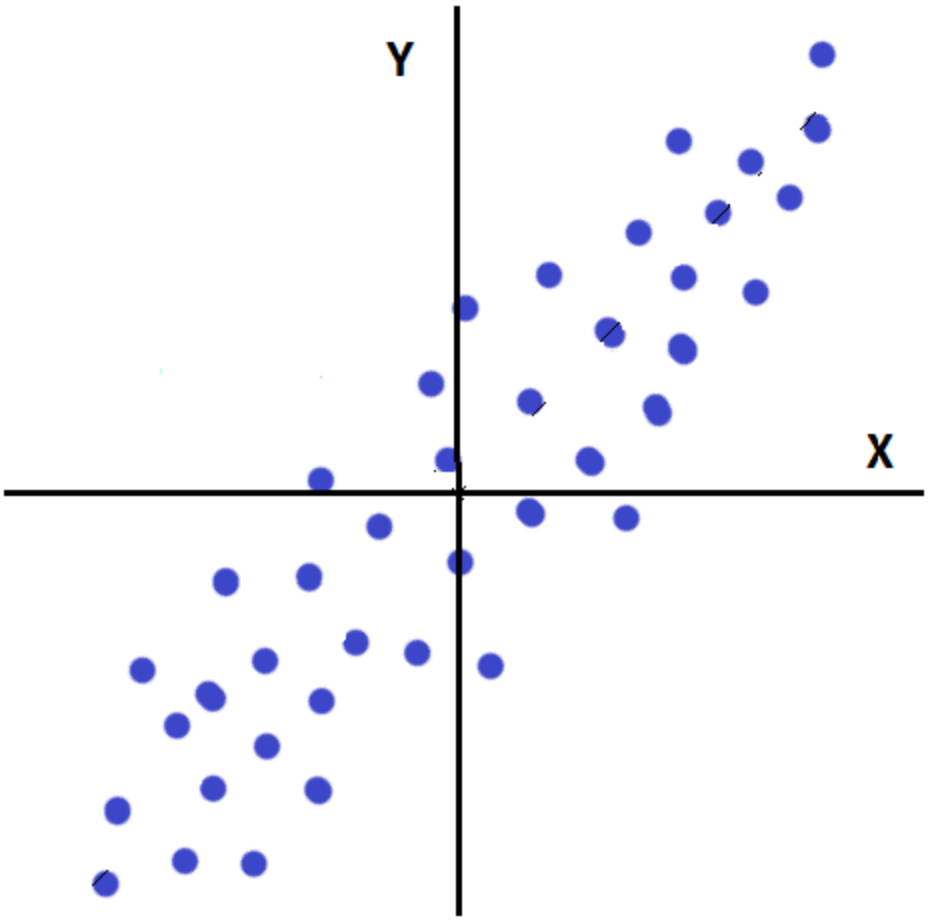
Geometric Explanation of Principal Component Analysis

Principal component analysis works by rotating the axes to produce a new coordinate system. Conceptually, think of the process as changing your vantage point to gain a better view of the data. Given these geometric underpinnings, using graphs can help explain how PCA finds the components.

Throughout this explanation, I link the graphs to characteristics I discuss above. This approach really brings principal components to life!

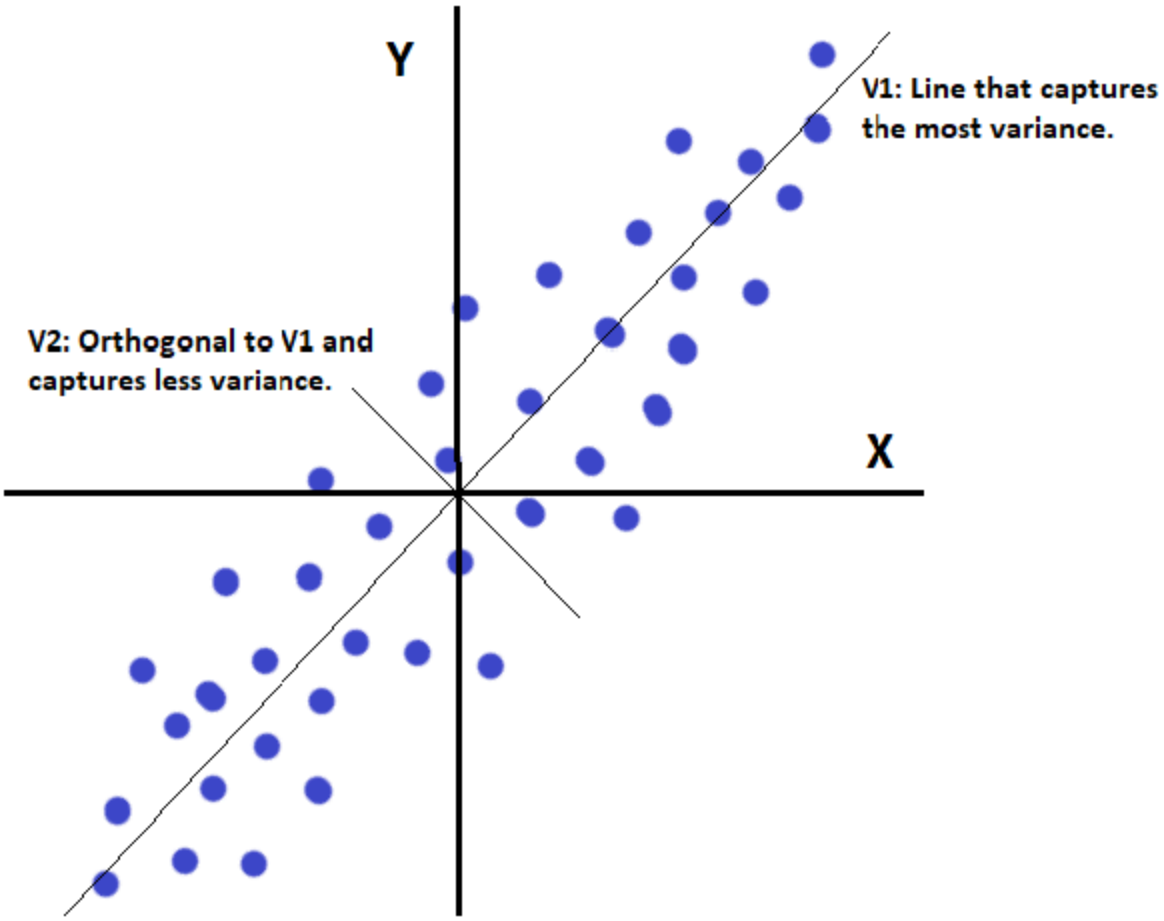
Note that I focus on rotating the axes because this is an introductory article. But PCA involves other crucial steps, such as standardizing all variables, assessing their correlation matrix, and transforming the original data values. Additionally, to keep things simple, we'll use an example where we have two variables, X and Y, and want to reduce them to one component.

Here's our original data. Clearly, the variables are correlated. This graph uses the native axes, where each one represents an original variable.



Finding the New Axes

PCA uses the correlation between variables to find the vectors that explain the most variance.



In this graph, V1 and V2 represent new vectors and are the principal components. The distance that data spread out along each vector represents the amount of variance the component captures. More variance equals more information.

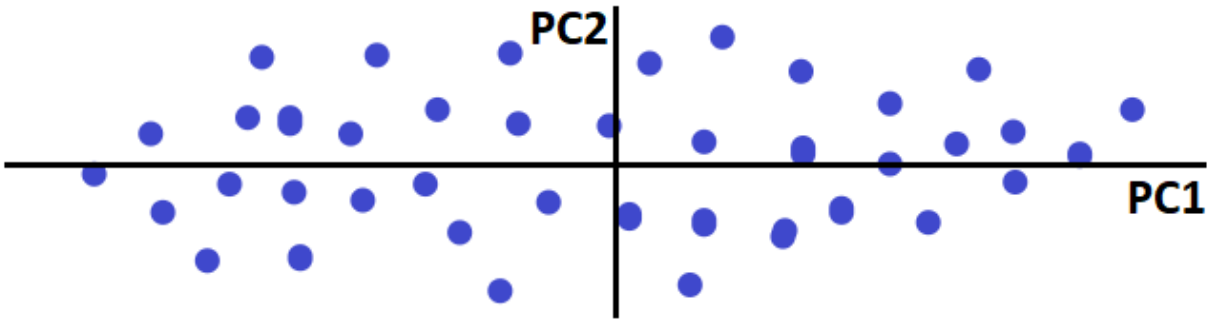
V1 represents the first component. The standard vertical distance between the data points and the new axis is less than their distance to the original X-axis. Technically, the vector minimizes the sum of the squared errors like a least squares regression line.

The length of the second component's line (V2) is shorter than the first. The relative lengths indicate that the first component accounts for significantly more variance than the second.

Finally, the two vectors are perpendicular, making them orthogonal.

Transforming the Coordinate System

Now let's use the vectors to transform the coordinate system.



The data points represent the same observations as before but now use new coordinates based on the orthogonal vectors. As you can see, there's no correlation because the data points have no slope.

Next, we want to simplify the data and use only one principal component. A single principal component score (a value on the new X-axis) represents each data point. That value is a linear combination of both original variables.

As you saw in the rotation process, the new axis uses the relationship between the original variables to combine information from both, but you can't directly interpret the new values. However, you can use the principal component scores in subsequent analyses.

When your dataset contains more dimensions, simply extend the process. Each subsequent axis transformation must satisfy the following:

- It accounts for more variance than any other potential new axis.
- It must be perpendicular/orthogonal to all previous axes.

Eigenvectors and Eigenvalues in Principal Component Analysis

Now that you've got the geometry down, it'll be easier to understand eigenvectors and eigenvalues.

Each principal component has a pair of these values. In the context of the geometric example:

- Eigenvectors signify the orientation of the new axes.
- Eigenvalues represent the line length or the amount of variance/information the new axis explains.

Principal component analysis computes these values from the correlation matrix.

Most statistical software ranks the principal components by their eigenvectors from largest to smallest. This process creates a list of components ordered from explaining the most to least variance.

Worked Example of Principal Component Analysis

Suppose a bank collects the following eight variables for loan applicants: Income, Education, Age, Residency at current address, Years at current employer, Savings, Debt, and the number of credit cards.

Let's use principal component analysis to see if we can reduce the number of dimensions yet retain most of the information.

How Many Components?

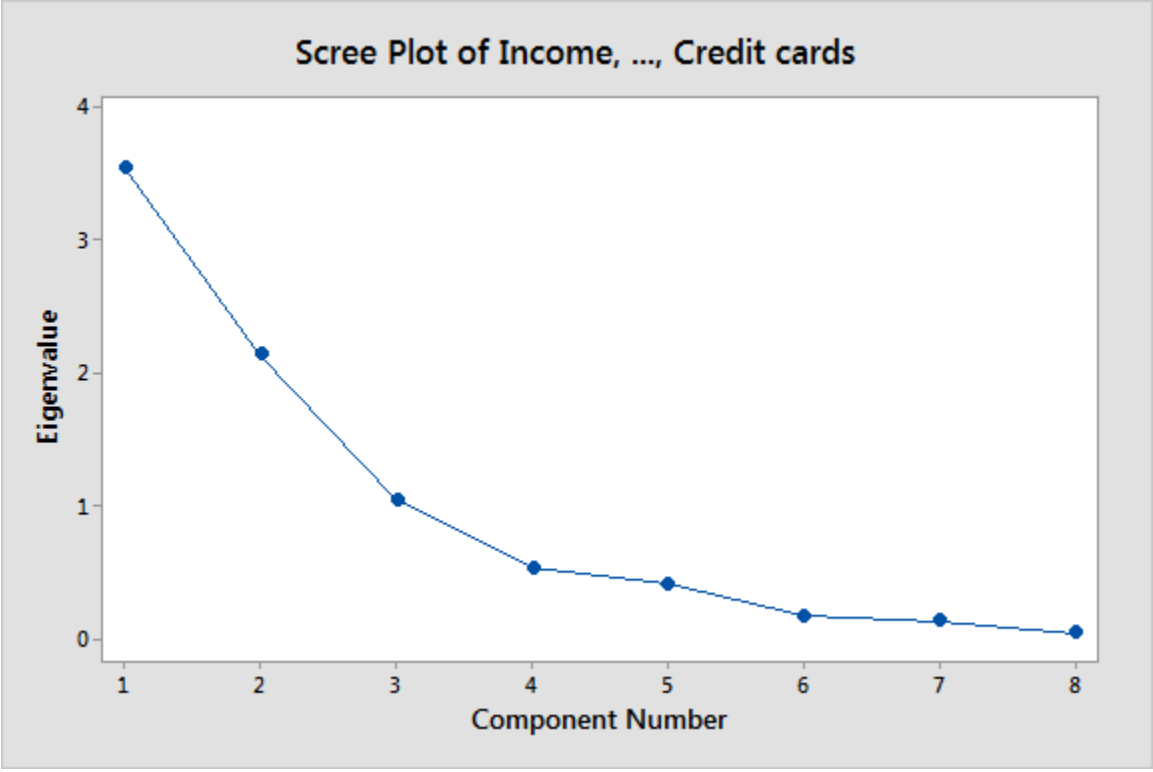
The first question we need to answer is, how many principal components should we use?

Analysts use several standard methods to select the number of components, including the following:

- **Scree plots**: The point before where the curve flattens.
- **Eigenvalues**: Use all components with eigenvalues greater than 1.
- **Subject-area knowledge**: knowing the percentage of variance or the number of components the downstream tasks require.

I don't have any specialized knowledge in this area, so I'll stick with the statistical measures! Fortunately, the scree plot and eigenvalues agree that we should use three principal components.

This scree plot shows that the flat portion of the curve begins at component 4, indicating we should use the first three principal components.



Eigenanalysis

The eigenanalysis below shows that the first three principal components have eigenvalues greater than 1. PC 4 has an eigenvalue of only 0.5315.

The eigenanalysis also indicates that PC1, PC2, and PC3 account for 44.3%, 26.6%, and 13.1% of the variance, respectively. Together, these three principal components account for 84.1%!

In this statistical output below, I've circled the eigenvalues and explained variance for the first three components.

Principal Component Analysis: Income, Education, Age, Residence, Employ, Savings, Debt, Credi

Eigenanalysis of the Correlation Matrix

Eigenvalue	3.5476	2.1320	1.0447	0.5315	0.4112	0.1665	0.1254	0.0411
Proportion	0.443	0.266	0.131	0.066	0.051	0.021	0.016	0.005
Cumulative	0.443	0.710	0.841	0.907	0.958	0.979	0.995	1.000

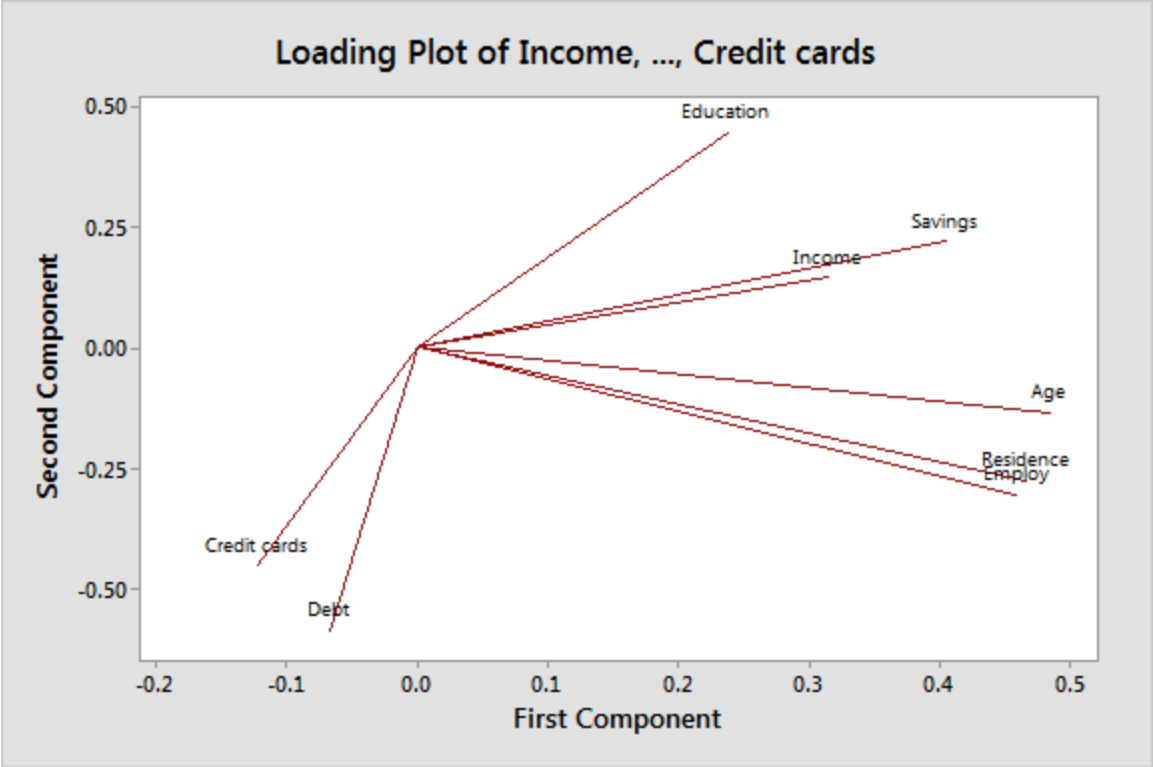
Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Income	0.314	0.145	-0.676	-0.347	-0.241	0.494	0.018	-0.030
Education	0.237	0.444	-0.401	0.240	0.622	-0.357	0.103	0.057
Age	0.484	-0.135	-0.004	-0.212	-0.175	-0.487	-0.657	-0.052
Residence	0.466	-0.277	0.091	0.116	-0.035	-0.085	0.487	-0.662
Employ	0.459	-0.304	0.122	-0.017	-0.014	-0.023	0.368	0.739
Savings	0.404	0.219	0.366	0.436	0.143	0.568	-0.348	-0.017
Debt	-0.067	-0.585	-0.078	-0.281	0.681	0.245	-0.196	-0.075
Credit cards	-0.123	-0.452	-0.468	0.703	-0.195	-0.022	-0.158	0.058

Loadings

The lower section of the output shows the correlations, or loadings, between the variables and the components. Use the loadings to identify the features that most strongly correlate with each component.

For example, the first principal component has high loadings on age (0.484), residence (0.466), employ (0.459), and savings (0.404). Hence PC1 is strongly associated with these variables.

Below, the loading plot displays the correlations graphically using the first and second components for the X and Y axis.



On the plot, age, residence, and employ group together with large, positive loadings for the first component. So, PC1 measures long-term stability—older, more time at current employer, and living longer in the current residence.

Conversely, credit cards and debt have negative correlations with PC2. It's a measure of debt usage.

Keep in mind that principal components are not always interpretable. For this example, we're fortunate we can interpret them to a degree. However, that's not always possible. Principal component analysis prioritizes retaining information over creating interpretable components!

Even if we couldn't interpret the principal components, understanding their loadings can help identify variables closely associated with the most informative components.

Using the transformed coordinate system, the software can calculate the principal component scores for all observations. I've saved the scores for the first three PCs in the dataset. Analysts can use these scores for additional analyses.