# **PRINCIPAL cOMPONENT aNALYSIS (pca)**

## **What is PCA?**

*Principal Component Analysis (PCA) is a statistical technique used to reduce the dimensionality of a data set while retaining as much of the original information as possible. This is a very popular preprocessing step for other analyses.*

*This is done by linearly transforming the initial data into a new coordinate system where most of the variation in the data can be described by fewer dimensions than the initial data.*

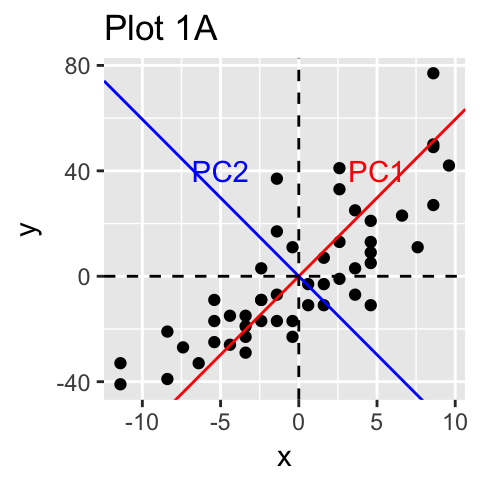
***When to use PCA?***

PCA is used when analyzing data sets with many correlated variables. By reducing the dimensionality, PCA can: 

* Make it easier to visualize and analyze data
* Decrease computation time in code
* Reduce noise and detect outliers in the dataset
* Help mitigate the problem of overfitting

***What are the data requirements and assumptions for PCA?***

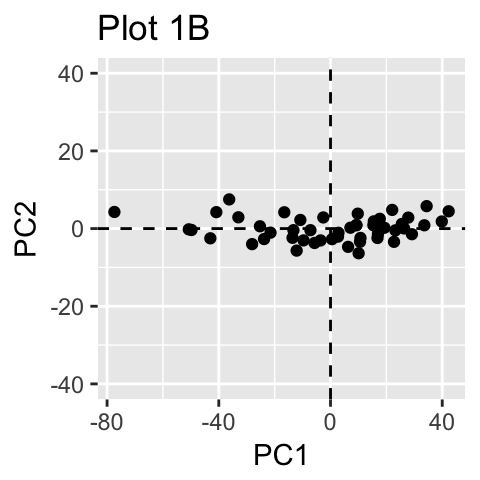
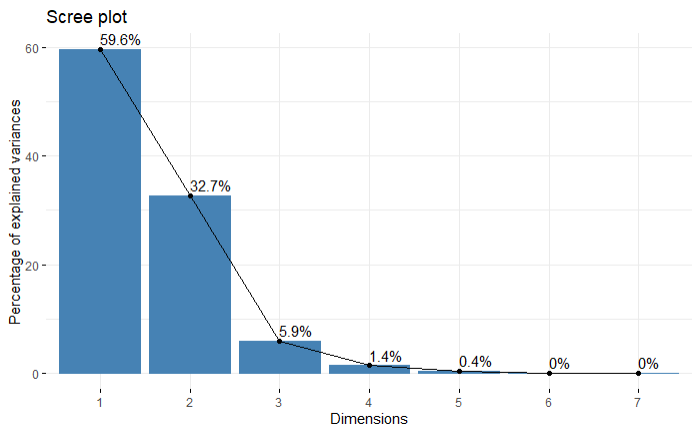
* Data must be numeric
* Data must have at least three features / variables
* Data must be linear (assess visually with pairwise plots or matrix scatter plots)
* Data must be standardized and continuous
* Data with missing values must be removed
* Data set should be highly correlated   
  (assess with Pearson correlation)



**Popular places where PCA is used:** Computer Vision, bioinformatics, machine learning, speech processing, and many more!

***How PCA works:***

***NOTE: This is the Singular Value Decomposition (SVD) method of performing PCA.***

1. **Standardize the data!** Data must be on the same scale.
2. **Find the first principal component (PC1)!** Find the best fit multiple regression line through the data.
3. ****Find the second principal component (PC2)!** Find the best line of fit that is perpendicular to PC1.
4. **Repeat for each variable!** Find the PCs for each variable.
5. **Interpret the results!** Analyze the relationship between variables using the PCs.

*Figure 2: Image of Scree plot to help determine how many PCs are necessary to explain a percentage of variance in the data.  
  
Look for the “elbow” or point where the curve flattens for the optimal number of components to retain*

*Figure 1: Plot 1A is a scatterplot with identified principal components; Plot 1B is reduced data by projecting each sample onto the first PC*

**COMP-4442 | Marina Garceau, Sammantha Firestone, Bradley Robasky**

***Resources:***

[StatQuest: Principal Component Analysis (PCA), Step-by-Step](https://www.youtube.com/watch?v=FgakZw6K1QQ&t=475s)

[Data Camp: Principal Component Analysis in R Tutorial](https://www.datacamp.com/tutorial/pca-analysis-r#introduction-to-pca)

[Toward Data Science: Principal Component Analysis (PCA) 101, using R](https://towardsdatascience.com/principal-component-analysis-pca-101-using-r-361f4c53a9ff)

[BuiltIn: Principal Component Analysis](https://builtin.com/data-science/step-step-explanation-principal-component-analysis)

[UC business Analytics R Programming Guide: Principal Component Analysis](https://uc-r.github.io/pca)

[STHDA: Principal Component Methods in R: Practical Guide](http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/)

[Statology: Principal Component Analysis in R: Step-by-Step Example](https://www.statology.org/principal-components-analysis-in-r/)

[Keboola: A Guide to Principal Component Analysis (PCA) for Machine Learning](https://www.keboola.com/blog/pca-machine-learning)

[Geeks for Geeks: Principal Component Analysis with R Programming](https://www.geeksforgeeks.org/principal-component-analysis-with-r-programming/)

[CRAN: Step-by-Step PCA](https://www.geeksforgeeks.org/principal-component-analysis-with-r-programming/)