Data Cleaning

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)

LESSON 7: PRINCIPAL COMPONENT ANALYSIS (PCA)

Lesson 7: Determining Variable Loads

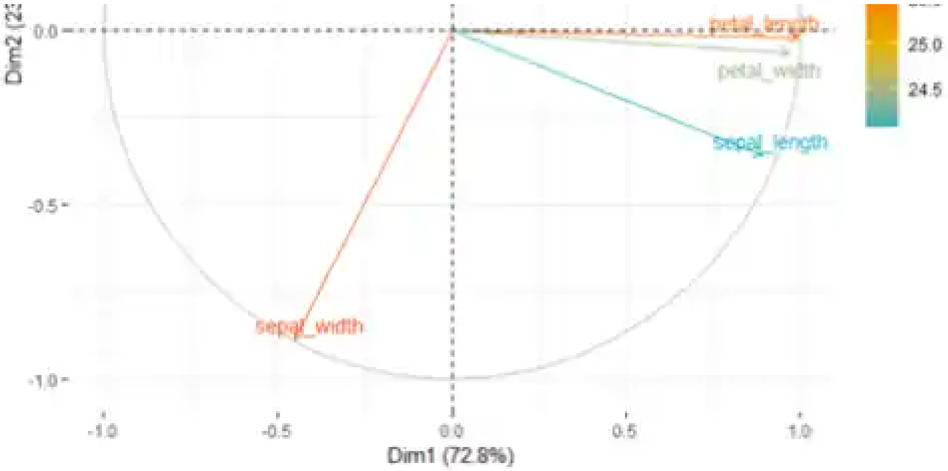
While scree plots help to determine the number of groupings that assist with data reduction, they do not indicate which variables belong to these groupings. Another way to ask this is, What variables "load" into a component. Most PCA analysis has a component loading output that helps you know this information. You may also see a component's load referred to as its rotation.

The larger the magnitude of a given variable's load on a given principal component, the more important that column was in forming that component. One way to visualize this is with a rotation plot:

Data Cleaning

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)

LESSON 7: PRINCIPAL COMPONENT ANALYSIS (PCA)



A rotation plot depicting variable loads.

What does this plot demonstrate? It is a way to visualize how each variable is transformed to create the components in PCA. This visualization can be used for any data set, but for this particular example, there are four total variables (petal_length, petal_width, sepal_length, and sepal_width). The first component (represented by the horizontal axis) captures 72.8% of the variation in the data set, and the second component (vertical axis) captures 23%. The colored lines represent the original variables, and their position on the plot are the values of the variables' respective loads for PC1 and PC2, respectively.

To fully understand this graph, you must consider it one axis at a time. Consider the Dim1 vertical axis to determine which variables are further away. In this case, sepal_width is close to this axis, whereas the other three are going in the opposite direction. Thus, petal_length, petal_width, and sepal_length are loading on Dim1 (the first principal component). You can examine the Dim 2 horizontal axis to identify

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)

LESSON 7: PRINCIPAL COMPONENT ANALYSIS (PCA)

bottom-right of the plot. Variables that did not have a high load factor for either of the first components would be found close to the graph center. You might think of such variables as less important in forming the PCA because they have low load values in the components that explain most of the variance in the data. The process of interpreting load and rotation is always subjective, however. It is your responsibility as a data analyst to continually understand and interpret what it means.

Notice that this graph only shows the load factors of the first two components. This is a typical way to analyze load and variable importance, but not the only way. This same analysis can be extended to any two given components, and you could even imagine a 3D plot being used to analyze the rotation of three components, or a pivot table to analyze all components at once non-visually. At times, you may also be presented a table with values that give "weights" of each variable with the components. Usually, a variable's biggest weighted value indicates which component it is primarily loading onto. You will consider an example of this in the next example.