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

LESSON 7: PRINCIPAL COMPONENT ANALYSIS (PCA)

Lesson 7: Scree Plots

A scree is an accumulation of loose stones lying at the base of a hill or cliff (as indicated by Figure 4 below). Thus, a scree plot helps a data scientist differentiate between the cliff and the rubble of the grouped data in each scenario. More specifically, in Figure 3, the cliff of the data is centered around 31 components, and the rubble data is located on the right side of the graph.

Some data scientists might also interpret a cliff in the scree plot in Figure 3 around seven components and consider most of the dots to the right as rubble. If the original data contained 44 individual variables, this scree plot suggests that the 44 variables might be reduced to either 31 or possibly as few as seven groupings of variables and still provide the same information. How can you be sure that it is providing the same information? This is where the idea of variability comes in. Variability is basically a measure of consistency (or lack of) in the data. More specifically, it is the average amount a specific data point deviates from the average value

Imagine if you went to a doctor because you felt sick, and after hours of testing, the doctor told you "I think I have figured out 5% of what is wrong with you." That probably would not make you feel confident in your doctor's ability to identify what is causing your illness. Instead, most people would rather hear, "I think I have figured out at least 95% of what is wrong with you." The more the doctor can explain about your illness, the more confidence you will have that you can proceed to a prescribed solution. Similarly, example, consider how unlikely it would be that you would be able to place all 100 items into one grocery bag.

The second component in the scree plot explains about 7.5%. Therefore, the first two components explain nearly 20% of the variability in the original data. If you keep accumulating the components, roughly 45% of the variability in the original data is explained in the first seven components. Therefore, using seven groupings of the original 44 variables explains nearly 50% of the original variation. If you consider the data through the 31st component (i.e., the second cliff shown in the figure), roughly 98% of variance is explained. Another way to say that is that 31 components give approximately the same information as 44 individual variables. If the data scientist is not specifically concerned with each of these 44 individual items, it would be more efficient to analyze 31 components rather than the original 44 individual variables. Even more helpful, if the data scientist is not worried about a loss of some information, analyzing seven components should be much more efficient than analyzing all 44 individual variables, with only some loss in information.

One might wonder why the loss of information is considered okay. Again, if the researcher is interested in the individual 44 items, that would not be acceptable. But if the researcher is mostly interested in the essence of that data, some loss of information might be acceptable. Imagine taking a trip and getting directions from a GPS. Most people would not be as concerned with "turn right on Baker Street" as much as "turn right at the next light." Sometimes the details are not needed because you will still end up in the same location.