1/24/2022

**Data Exploration Interview Problem**

For this interview problem, you will be investigating a breast cancer data set from a tumor micro array where each row is a tumor sample from a patient. These tumor samples were then analyzed to segment cells and determine how many of those cells are positive for each biomarker/biomarker combination. These numbers were then normalized to total cells and by specific cell subtypes. Additionally, a spatial analysis was performed on each sample to determine the distance of certain cell types to other cell types. To analyze this data set, you are asked to investigate the data by the stage of the cancer category, and will be asked to:

1. Find the 10 most important features contributing to the variance in the stage of cancer.
2. Run a PCA and find the first 2 principal components and plot them against each other and color code the points by stage.

Bonus points if you can make a confidence ellipse that fits the points for each stage of cancer on the PCA plot.

Notes: The data is organized in way where each column is a measurement, and each row is a sample. The metadata containing the stage information is in the middle of the data sheet.