Data Wrangling/EDA for Final Project

By: Lily Thomson and Alysha Akhtar

Our dataset is from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database. This is a Canada-UK project that looks at a cohort of 1905 women for targeted sequencing data (1). The dataset contains clinical attributes of individuals with breast cancer. Some examples of this are age_at_diagnosis, cellularity, chemotherapy, and her2_status. It also contains genetic attributes of these patients including m-RNA level z-scores for 331 genes and mutations for 175 genes. mRNA or messenger RNA acts as a code for DNA to be transcribed off of. Based on the documentation of this dataset, $\left(\frac{expression\ in\ tumor\ sample - mean\ expression\ in\ normal\ sample}{standard\ deviation\ of\ expression\ in\ normal\ sample} = mRNA\ z\ score\ (3)\right)$ which means that the data is already normalized. Essentially, these scores tell us if a gene is upregulated or downregulated relative to the expression of normal tissue.

According to the American Cancer Society, breast cancer accounts for 1 in 3 of all new female cancers each year. Additionally, incidence rates have increased by 0.6% per year (2). Although genetic links in breast cancer have been thoroughly studied, we want to further understand correlations between clinical variables and expression of various genes. Further study in this area can aid in early detection, risk assessment, and treatment selection. This is a very robust dataset that we have confidence in to be able to draw meaningful conclusions.

Given that this is a very large dataset though, we need to be able to cast a wide net to find the correlation between expression of certain genes and clinical variables. We want to use our methods to validate already established correlations, while also discovering new, less studied ones. Additionally, given how complex of a disease breast cancer is, it is difficult to completely isolate variables without taking others into account. Our plan is to look at combinations of variables side by side in order to perform effective analysis and draw conclusions on our hypotheses.

We looked at both clinical variables and gene expression for our EDA. The distribution of patient age at diagnosis, overall survival in months, tumor size, and mutation counts were looked at by creating a histogram for each variable and looking at the descriptive statistics tables. This was helpful to better understand patient characteristics in this dataset. Additionally, two histograms were created of BRCA2 expression and CDH1 expression, two genes that play an important role in developing breast cancer (4). The relationship between these two gene expressions and survival in months was studied using scatter plots. Finally, a bar plot of positive or negative status of ER, PR, and HER2 was made. Understanding these variables are important because they play a large role in what kind of therapies a patient is eligible for to treat their type of breast cancer. **Plots are shown in the attached python notebook.**

References:

1. “Breast Cancer Gene Expression Profiles (METABRIC).” *Kaggle*, www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric. Accessed 30 Sept. 2024.
2. “Breast Cancer Statistics: How Common Is Breast Cancer?” *Breast Cancer Statistics | How Common Is Breast Cancer? | American Cancer Society*, www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html. Accessed 30 Sept. 2024.
3. By, et al. “Z-Score: Definition, Formula, Calculation & Interpretation.” *Simply Psychology*, 6 Oct. 2023, www.simplypsychology.org/z-score.html.
4. “Inherited Cancer Risk: BRCA Mutation.” *Johns Hopkins Medicine*, 10 Oct. 2023, www.hopkinsmedicine.org/health/conditions-and-diseases/breast-cancer/inherited-cancer-risk-brca-mutation#:~:text=BRCA1%20and%20BRCA2%20account%20for,%2C%20PTEN%2C%20TP53%20and%20NF1.