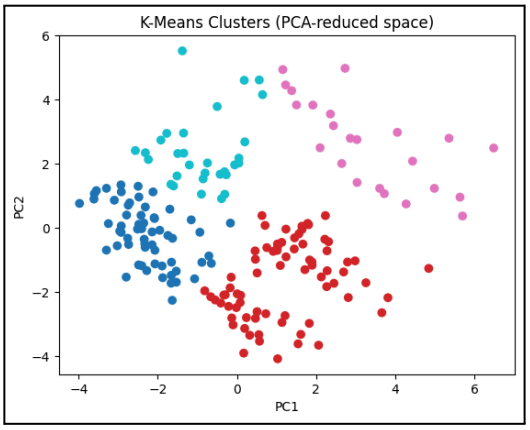
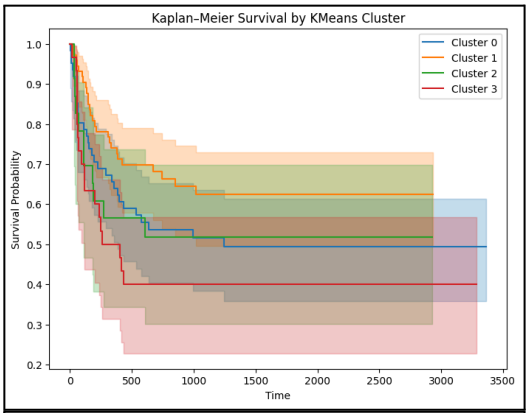
Objective:
The goal of this lab was to apply survival analysis methods to the *Bone Marrow Transplant: Children* dataset from UCI. The dataset contains information about pediatric patients with hematologic diseases who underwent bone marrow transplants.

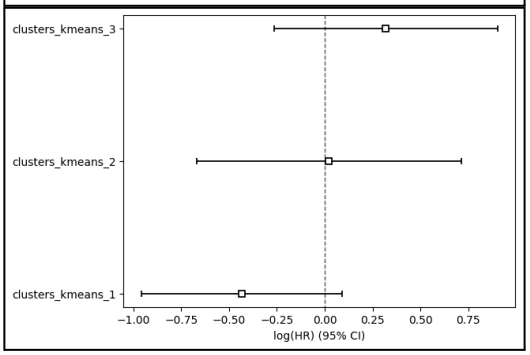
Exploratory Analysis:
Initial inspection revealed that several features contained missing values, although none were missing enough to be dropped. The features that showed strong positive correlation with survival time include aGvHDIIIIV and CD34kgx10d6, while those with strong negative correlation include PLTrecovery and ANCrecovery. A cutoff of $|r| > 0.1$ was used to identify relevant predictors. In terms of feature-feature correlation, high redundant features ($|r| > 0.75$) were grouped and from each correlated group at least one feature was retained to preserve shared information. This resulted in a reduced feature set used for clustering.

Clustering:
K-Means clustering was applied to the selected features to differentiate patients into groups with similar characteristics. Prior to clustering, the data was scaled in order to be standardized, and dimension reduction was done via Principal Component Analysis to visualize the clusters in two dimensions. Several values of *k* were tested, and four clusters were selected as they provided a good balance between separation and interpretation. The resulting clusters showed distinct groupings in PCA-reduced space, suggesting meaningful heterogeneity among patients. The figure shows the K-Means clusters projected onto the first two principal components.



Survival Analysis:
Kaplan-Meier and Cox regression were used to find evidence of survival heterogeneity among the patient subgroups. The Kaplan-Meier curves (as seen to the right) showed clear differences between patient clusters. Some clusters demonstrated consistently higher survival probabilities over time, thus suggesting that the clustering successfully separated patients into groups with distinct prognoses. The Cox regression, as visualized on the right, confirmed survival differences between clusters. The model produced hazard rations indicating relative risk compared to a baseline cluster. For example, one cluster showed an HR > 1 (higher mortality risk) while another had HR < 1 (protective effect).





Morgan Yeazel October 3rd, 2025

Key Takeaways

Exploratory Analysis:

- Several features had missing values, but not enough to be considered significant.
- Both positive and negative correlations with survival time were informative.
    - Features positively correlated with survival suggested longer survival.
    - Features negatively correlated indicated shorter survival.
- Strongly redundant features ($|r| > 0.75$) were grouped; at least one from each group was kept to retain shared information.
- The final feature set was smaller, cleaner, and easier to interpret.

Clustering:

- Four clusters were chosen as the most interpretable grouping.
- PCA projection showed clear separations between clusters in reduced space, confirming that patients were grouped based on meaningful feature patterns.
- Cluster sizes were balanced enough for survival comparison (no cluster dominated the dataset).

Survival Analysis:

- Kaplan-Meier curves showed clear survival differences across clusters.
    - Some clusters exhibited consistently higher survival probabilities, thus confirming that K-Means clustering successfully separated patients into prognostically distinct subgroups.
- Cox regression quantified survival differences.
    - HR > 1 indicated clusters with higher morality risk relative to the baseline.
    - HR < 1 indicated clusters with protective effects and thus lower risk.

Overall, the survival analysis methods provided consistent evidence that patient clusters have distinct survival outcomes.

Google Colab:
https://colab.research.google.com/drive/1DhcOy2Fz3Q9yTpKL7PzgewcJ6N6J6klq?usp=sharing

Github:
https://github.com/mryeazel-729/MLHealth/blob/main/Lab1.ipynb