Project 4 Proposal

Dataset: https://www.kaggle.com/datasets/vinicius150987/titanic3

Our Data:
We plan to work with the ship records of the RMS Titanic. We have a data set of 1309 passengers, with information including their class, age, gender, family members aboard, and price paid for their ticket.

Back end (ETL)
Our plan is to start out in python pandas in order to clean and transform our data. We have a lot of variables to work with, including some that are irrelevant to our project. We plan to clean up the data by removing missing data and any columns we don't. Additionally, we will create dummy variables for any of the categorical data such as cabin class (first, second, or third). We can also address any other optimization in this step, such as doing PCA and creating a data set with only principal components, or standardizing variables.

Visualizations:
We will create visualizations in python to help better understand and describe our data. However, our final front end visualizations will be in Tableau and include graphs depicting our model's performance over different optimization iterations and charts on Titanic passenger demographics and their survival rates.

Questions We'll Ask:

1) Who will survive the Titanic sinking according to our model?
2) What feature(s) contributed most to a passenger's survival?
3) What type of machine learning model is most effective in predicting survivors?
4) Can we accurately predict Jack and Rose's fate using our model?