Research Question

In the CalCOFI dataset, which represents over 60 years of oceanographic data with more than 500,000 records, I am interested in exploring the relationship between water salinity and water temperature in the California Current. Specifically, I seek to answer two questions:

1. Is there a statistically significant relationship between water salinity and temperature?
2. Can we predict water temperature based on salinity values, and what models provide the most accurate predictions?

Analysis Plan

To address these questions, I will conduct both exploratory and predictive analyses. First, I will clean and preprocess the dataset to handle missing values, outliers, and inconsistencies. I will then perform an exploratory analysis, visualizing the relationship between salinity and temperature, including scatter plots and correlation analysis, to identify the strength and direction of their relationship.

For the predictive analysis, I will build and compare several regression models to determine the best approach for predicting temperature based on salinity. The models will include:

1. **Linear Regression**: A simple baseline model to assess the linear relationship.
2. **Polynomial Regression**: To capture any non-linear relationship between salinity and temperature.
3. **Decision Tree Regression**: A flexible model to handle potential non-linearities in the data.
4. **Random Forest Regression**: An ensemble method to improve predictive performance over decision trees.

Each model will be evaluated on both training and test sets using metrics such as R^2 and RMSE to compare their performance. Through cross-validation, I will also ensure that the models generalize well to unseen data. Ultimately, I will determine whether salinity is a reliable predictor of water temperature and identify the most effective model for making these predictions.

This dataset's high complexity, large size, and detailed oceanographic parameters make it ideal for this project, allowing for robust modeling and statistical analysis. Specifically, It contains more than 50w data. The CalCOFI dataset’s long-term scope and high frequency of measurements enable a thorough examination of the salinity-temperature relationship over time and across different oceanographic conditions.