1. **What is your central research question? How will your work contribute to the problem that you have defined?**

Does Age, Rating, Department Name, and Class Name have a significant connection to the Recommended IND column? Do they have a connection with the Positive Feedback Count column?

The work will help us understand the relationship between customer feedback features (Age, Rating, Department Name, Class Name) and the recommendations/feedback on women's clothing in Walmart. Depending on the recommendations and feedback counts, Walmart can choose whether to keep more of a particular item in stock to increase sales.

2. **Explain in short the way you handled the solution. You may want to include something you struggled with and managed to solve.**

The clothing dataset was preprocessed to handle missing values and encode the categorical variables using one-hot encoding, column transformers, and pipelines. I trained classification models (Logistic Regression, kNN, SVC, Random Forest) to predict customer recommendations based on numeric and categorical features. I obtained accuracy, precision, recall, and F1 scores in classification reports from all models. Next, I trained a linear regression model to predict positive feedback counts using the same group of features. I obtained the mean squared error and then conducted an F-test. After doing further research, I realized that I did not need the F-test and that the MSE and RMSE were all that I needed.

3. **Display and explain your results.**

According to the classification reports and confusion matrices, all models had over 90% accuracy in predicting customer recommendations. The most accurate was Logistic Regression at 94.43%. The groups of features had a strong connection with the Recommended IND column due to high performance in all models. *Classification reports are shown in the notebook.*

The linear regression model predicted positive feedback counts with a mean squared error of 30. After conducting an F-test, the initial conclusion was that there was a connection between the features and positive feedback count. I took the square root of MSE to find RMSE, the result was much larger than 1. If the model is greater than 1, then there is no connection. The further the MSE is away from 0, the worse of a connection there will be. Therefore, there is no connection between the group of features and the positive feedback count.