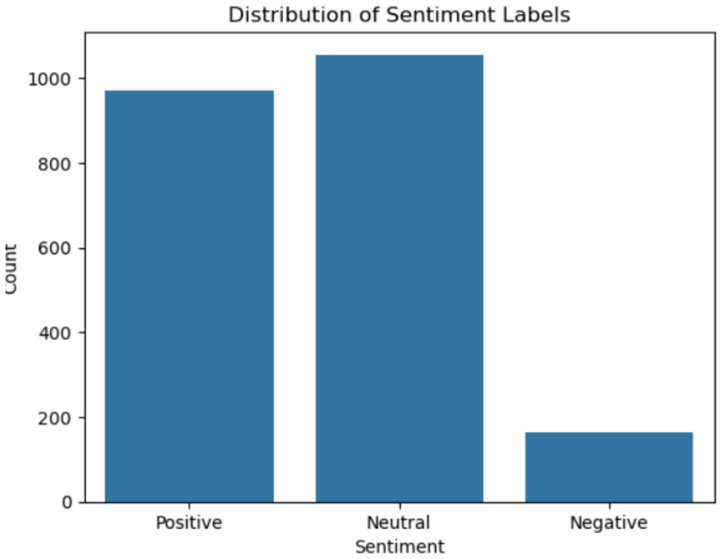
Final Report

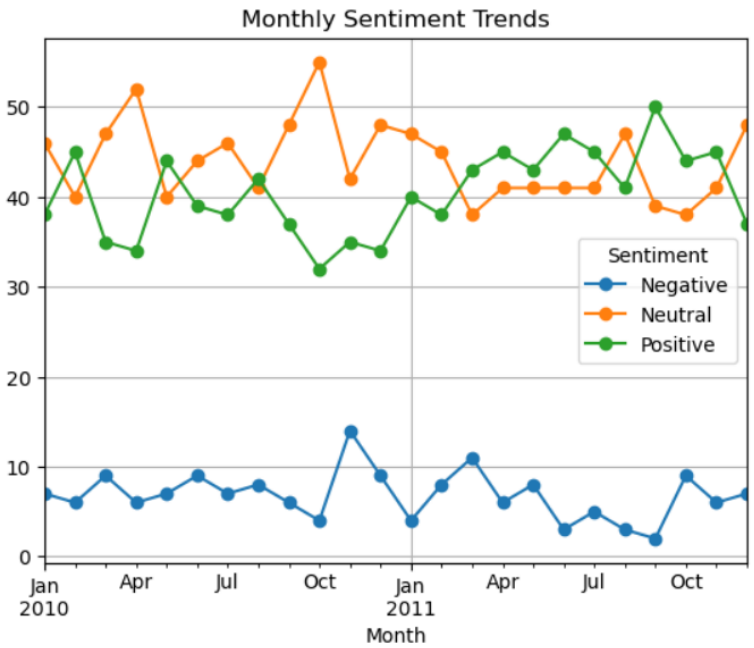
Approach and Methodology

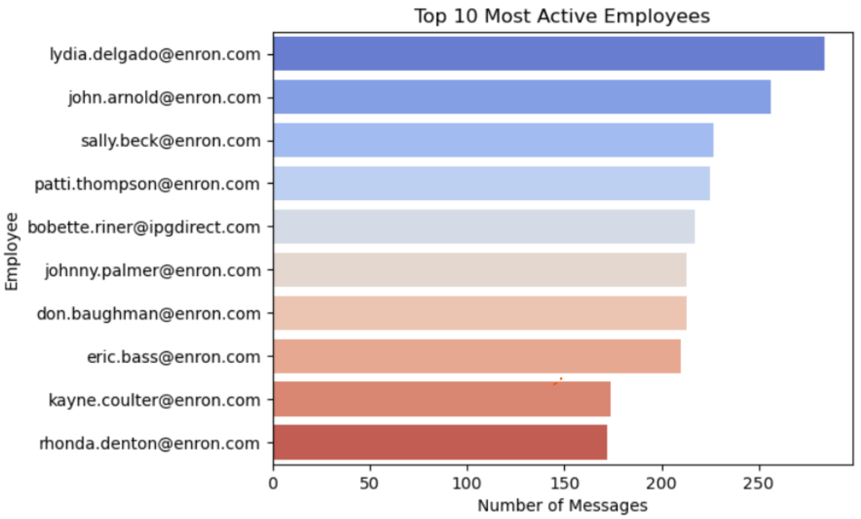
My approach to the Sentiment Labeling task was to use the TextBlob library to calculate the sentiment polarity score for each body of the email. I classified the score based on whether it was greater than 0 (Positive), less than 0 (Negative), or exactly 0 (Neutral). For the EDA, I used .info() to understand the size of the data and how many entries there were. I then selected a few metrics that would help with the tasks ahead, so I looked at the distribution of sentiment classifications, sentiment trends over months, and the most active employees.

Key Findings from EDA

- From the EDA, my key findings were that Positive had the highest number of labels, with Neutral having half as many. Negative labels made up only about one-sixth the number of Positive labels. Another key finding was that the number of messages labeled as Neutral and Positive followed opposite trends each month. When Neutral message counts went down, Positive message counts tended to go up, and vice versa. Negative labels generally followed the Positive trend line but with some inconsistencies. Another insight was gained by ranking employees by their number of messages to identify the most active ones. Lydia Delgado was the most active employee in the dataset.







Explanation of the employee scoring and ranking process

- How I went about the employee scoring process was to create a dictionary containing the respective scores for each sentiment classification. I then created a new column in the DataFrame that mapped these scores to each message. I changed the date column to represent just the year and the month. After these changes were made, it was easy to aggregate all the scores by the same year and month and store that in a column. For ranking, I sorted the monthly sentiment scores per employee in descending order to find the top three positive and top three negative employees for each month. Ties were broken alphabetically by employee name.

Flight Risk Identification Criteria and Outcomes

- To find the employees who are at risk of leaving, the main criteria was to identify any employee who had sent four or more negative emails within any 30-day period. To achieve this, I filtered the dataset to only include entries with negative sentiment and used two nested for loops to check the time difference between messages for each person and to count how many were sent within that time frame. The outcome was that the following employees were identified as flight risks, which included all of the employees in the dataset.

```
['bobette.riner@ipgdirect.com',
 'don.baughman@enron.com',
 'eric.bass@enron.com',
 'john.arnold@enron.com',
 'johnny.palmer@enron.com',
 'kayne.coulter@enron.com',
 'lydia.delgado@enron.com',
 'patti.thompson@enron.com',
 'rhonda.denton@enron.com',
 'sally.beck@enron.com']
```

Overview and Evaluation of Predictive Model

- My predictive model evaluated whether the frequency of messages per month and the specific month influenced sentiment scores. I built the model by creating columns that represented each of these features and merging them into a single DataFrame. I assigned X to include the month and the count, and y to be the score for each month. Then I performed a train and test split. I created a linear regression model, fit it to X_train and y_train, and evaluated the model using appropriate metrics to measure how well these features predicted sentiment scores. My final evaluation showed that there were moderate trends represented through this model. The R^2 value was not extremely high, which is reasonable for a model using only two features. The relatively high variance in the sentiment scores also supports the idea that more complex or additional features would likely improve the model's performance.