Email Log Analysis – Documentation

1. Introduction

In this project, we analyzed an email log dataset to understand email activity and communication patterns. The dataset contains:

- Sender – Who sent the email
- Subject – The subject line of the email
- Body – The full content of the email
- Sent Date – When the email was sent

The main idea is to see patterns in email traffic, find out who the most active senders are, what topics come up most often, and identify trends over time.

2. Aim

Our goal for this project was simple:

1. Clean the dataset so we can trust our results.
2. Analyze patterns like top senders, frequent subjects, and common keywords.
3. Visualize email trends over time to make the results easier to understand.
4. Summarize insights that could help someone manage or monitor email activity effectively.

3. Tools We Used

- Python 3 – Our main programming language.
- pandas – For reading, cleaning, and analyzing the data.
- matplotlib & seaborn – To create graphs and visualizations.
- collections.Counter – To find the most common words in email bodies.
- wordcloud – To make a visual representation of the most frequent words.

4.Our Approach

Step 1: Load the Data

We started by loading the email CSV file into Python using pandas. This allowed us to work with the data like a table.

Why: So we could manipulate, analyze, and visualize the data easily.

Step 2: Clean the Data

- Removed duplicates so the same email wouldn't be counted twice.
- Fixed character encoding issues (like â€™ becoming ') so text was readable.
- Cleaned subject lines by removing extra spaces.

Why: Clean data ensures our analysis is accurate and meaningful.

Step 3: Analyze the Data

- Top Senders: Counted which email addresses sent the most emails.
- Top Subjects: Counted which subjects appeared most often.
- Common Keywords: Analyzed email bodies to find the most frequently used words.

Why: This helps us understand who the most active users are and what topics or issues appear most often.

Step 4: Visualizations

We created several graphs to make the analysis easier to understand:

1. Emails per Day (Line Chart) – Shows daily email traffic.
2. Top Senders (Bar Chart) – Highlights the most active senders.
3. Top Subjects (Bar Chart) – Shows which subjects appear most often.
4. Emails per Sender (Pie Chart) – Shows the proportion of emails sent by top senders.
5. Emails per Month (Line Chart) – Shows trends over time.
6. Word Cloud – Highlights common words used in email bodies.
7. Emails by Day of the Week (Bar Chart) – Shows which weekdays are busiest.
8. Email Activity Heatmap (Day vs Hour) – Shows peak email hours during the week.

Why: Visualizations make it easy to spot patterns and understand the data quickly.

5. Results

- We identified the most active senders and the most frequent subjects.
- The common keywords gave insight into the main topics discussed.
- Daily, weekly, and monthly trends showed when email activity is highest.
- The heatmap highlighted peak hours and busiest days.

6. Conclusion

Through this analysis, we learned:

- Who sends the most emails.
- What topics are commonly discussed.
- When email traffic is highest during the day and week.

These insights can help organizations monitor email activity, prioritize responses, and understand communication patterns better.

7. Future Work

- Perform sentiment analysis to see if emails are positive or negative.
- Group emails into categories based on topics.
- Create an interactive dashboard for real-time monitoring.