Enron Email Analysis

Hicham REMLI

2025-05-26

Introduction

The Enron email dataset provides a unique opportunity to study large-scale corporate communication within a major U.S. energy company prior to its collapse in 2001. The dataset includes over 500,000 messages exchanged between employees, executives, and departments, making it a valuable resource for network analysis, behavioral insights, and communication trends.

The objectives of this project are:

- To clean and preprocess the raw Enron dataset for analysis.
- To conduct exploratory data analysis (EDA) to uncover patterns related to senders, recipients, timing, job roles, and email content.
- To build an interactive Shiny app that allows users to dynamically explore these patterns.
- To interpret results and provide scientific commentary supported by statistical visualizations.

Data Description

The dataset used in this project is the **Enron Email Corpus**, a large collection of emails made public during the investigation into Enron Corporation. It contains both the **content** and **metadata** of emails exchanged by Enron employees from 1998 to 2002.

The project utilizes the following key datasets:

- **message**: Contains the core metadata for each email, including `mid` (message ID), `sender`, `date`, and `subject`. It forms the foundation for linking other information.
- **recipientinfo**: Maps each email to one or more recipients, along with their recipient type (`TO`, `CC`, or `BCC`).
- **employeelist**: Contains job-related information about employees, including `Email_id`, `status` (job role), department, and name.
- **referenceinfo**: Tracks message reply relationships using a reference ID field.
- **message_with_roles**: A merged and expanded version of the `message` dataset, enriched with employee roles through join operations.

Additional derived datasets were created for analysis purposes:

- `emails_by_day`: Aggregates emails sent per day to analyze activity over time.
- `emails_by_role_full`: Summarizes the total number of emails sent by different job roles, with missing roles labeled as “Unknown”.
- `recipients_per_message`: Measures the number of recipients per message.
- `subject_words`: Extracted and tokenized subject line words for basic content analysis.
- `top_senders_named`: Joins sender emails with the employee list to show named top senders.

Data Dimensions

- **Total messages**: ~252,000

- **Total recipients**: ~2 million
- **Employees with role info**: 149
- **Time span**: 1998–2002

All datasets were preprocessed to handle `NA` values, merge relevant role data, and standardize date formats for timeline-based filtering.

Exploratory Data Analysis (EDA)

1. Top Email Senders

To identify the most active communicators, we started by analyzing the total number of emails sent by each unique sender. This gives us a basic but important insight into who the key actors are in Enron's internal communication network.

The following steps were applied:

- The `message` dataset was grouped by `sender`.
- Emails were counted using `n()` to get the total sent per sender.
- The top 10 senders were selected.
- A bar chart was used to visualize the results.

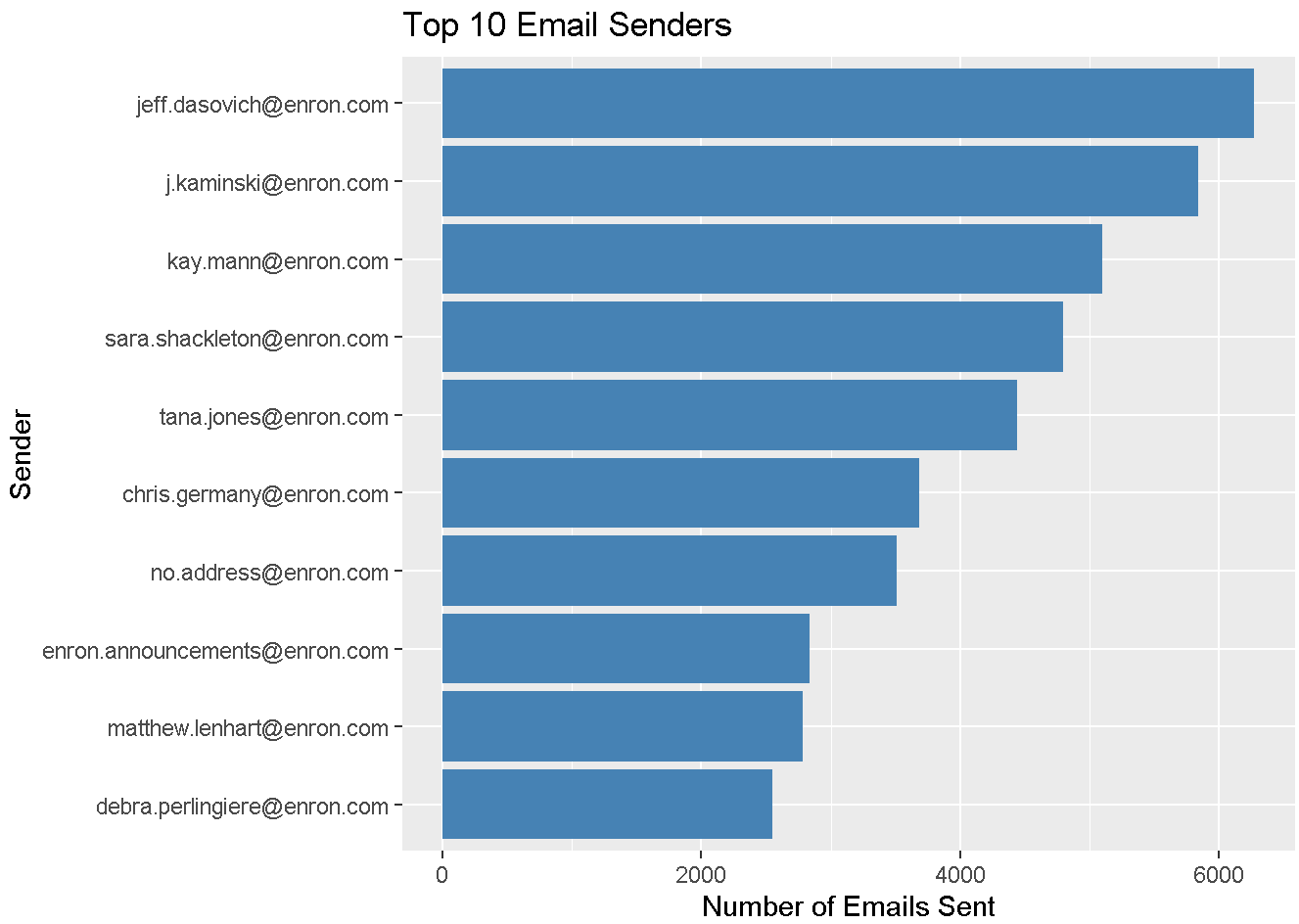
```
## Warning: le package 'dplyr' a été compilé avec la version R 4.4.3
```

```
##
## Attachement du package : 'dplyr'
```

```
## Les objets suivants sont masqués depuis 'package:stats':
##
##     filter, lag
```

```
## Les objets suivants sont masqués depuis 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Warning: le package 'ggplot2' a été compilé avec la version R 4.4.3
```

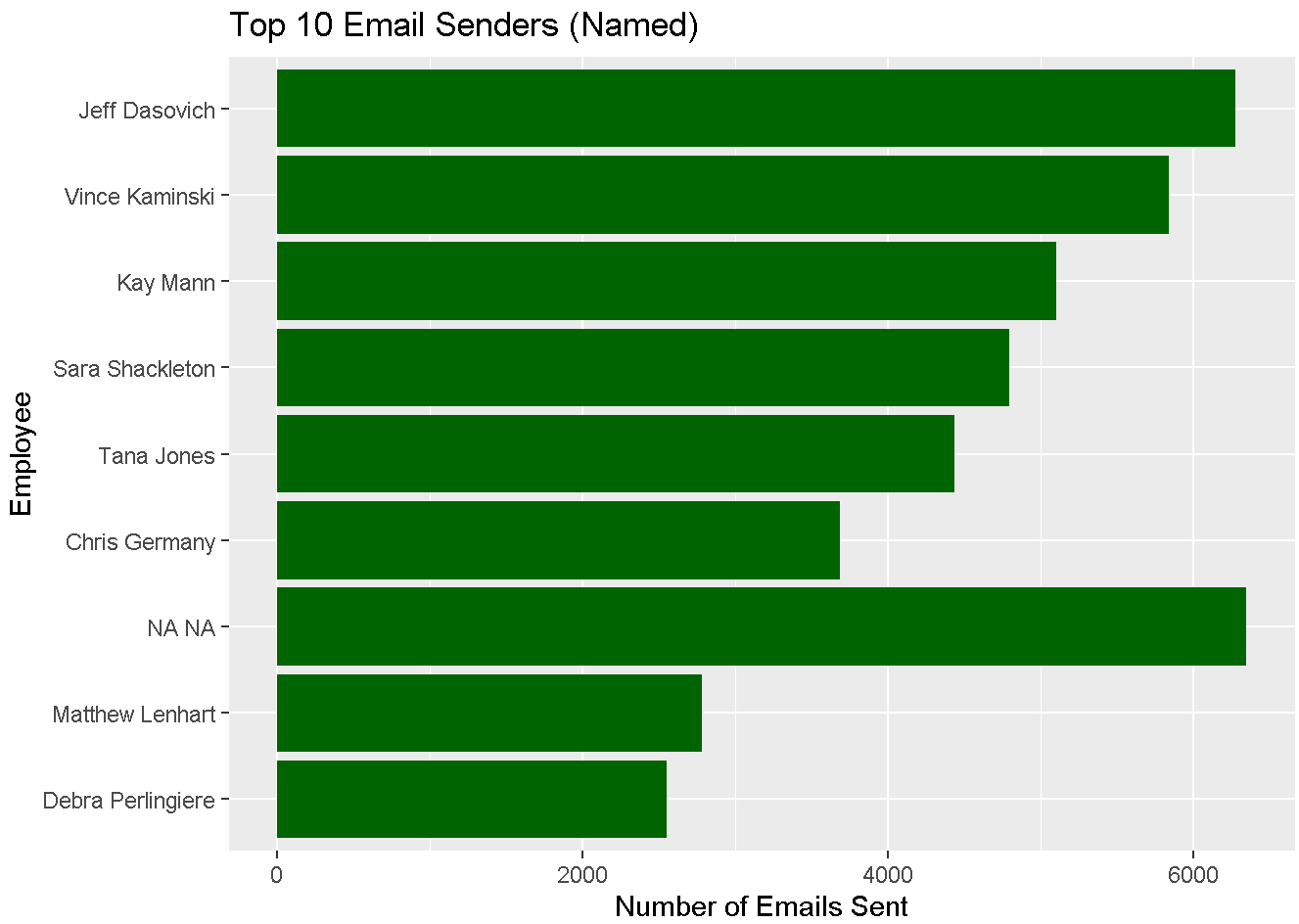


Explanationt:

This analysis shows that email traffic within the organization was **highly concentrated**: a small number of individuals were responsible for a large proportion of communication. This may reflect **management roles** or **departmental responsibilities**, but could also indicate an **over-reliance on a few central communicators**, which may become a risk in terms of transparency or resilience in communication networks.

2. Top Senders Named

To better understand who the most active senders are, we joined the email sender addresses with employee names and job titles from the `employeelist` dataset. This gives us more context beyond email IDs.



Explanation:

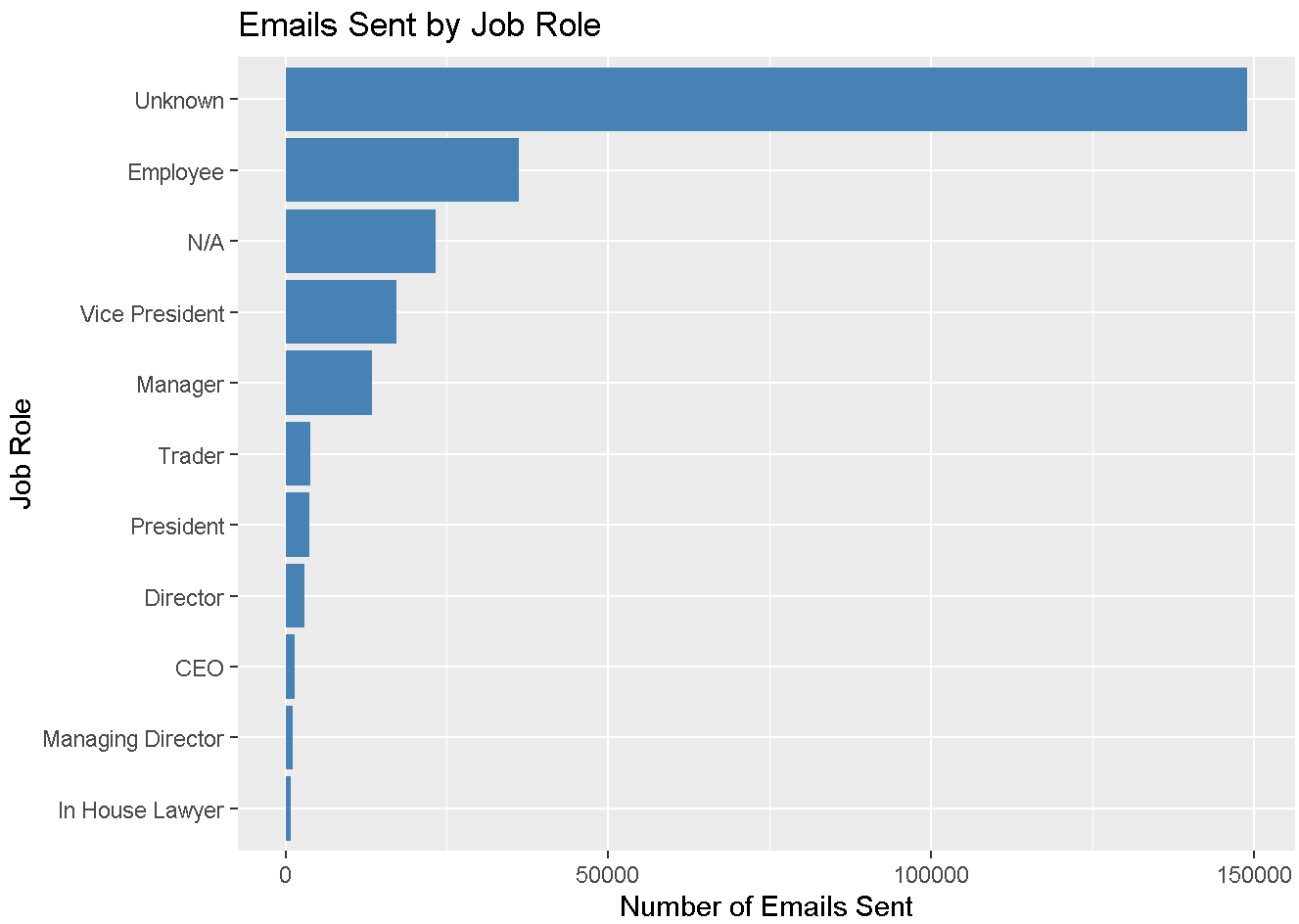
Matching senders to employee data reveals more than just frequency ,it connects email activity to individual **roles** within the company. Notably, several of the top communicators hold positions like Vice President or Manager, suggesting a concentration of communication responsibilities at mid-to-high levels of the organizational hierarchy.

This also allows us to track communication patterns by **person**, not just by anonymous email address, which is essential for understanding how influence and coordination flowed across the company.

3. Job Role Analysis

To analyze communication patterns across different job roles, we joined the message dataset with the employeelist. We then grouped by the employee status (job title) and counted the number of messages sent per role.

Missing or undefined roles were labeled as “Unknown”.



Explanation:

This analysis shows how communication volume is distributed across job roles. Unsurprisingly, roles such as **Vice President** and **Manager** tend to send a large number of emails, reflecting their involvement in daily operations.

Interestingly, some roles show lower-than-expected activity, and a significant number of messages are associated with senders labeled as “Unknown”. This could be due to: - Missing data in the employeelist - External or non-standard email addresses - Employees who are no longer in the database

This suggests that internal communication was partially decentralized or involved actors not fully captured in the metadata.

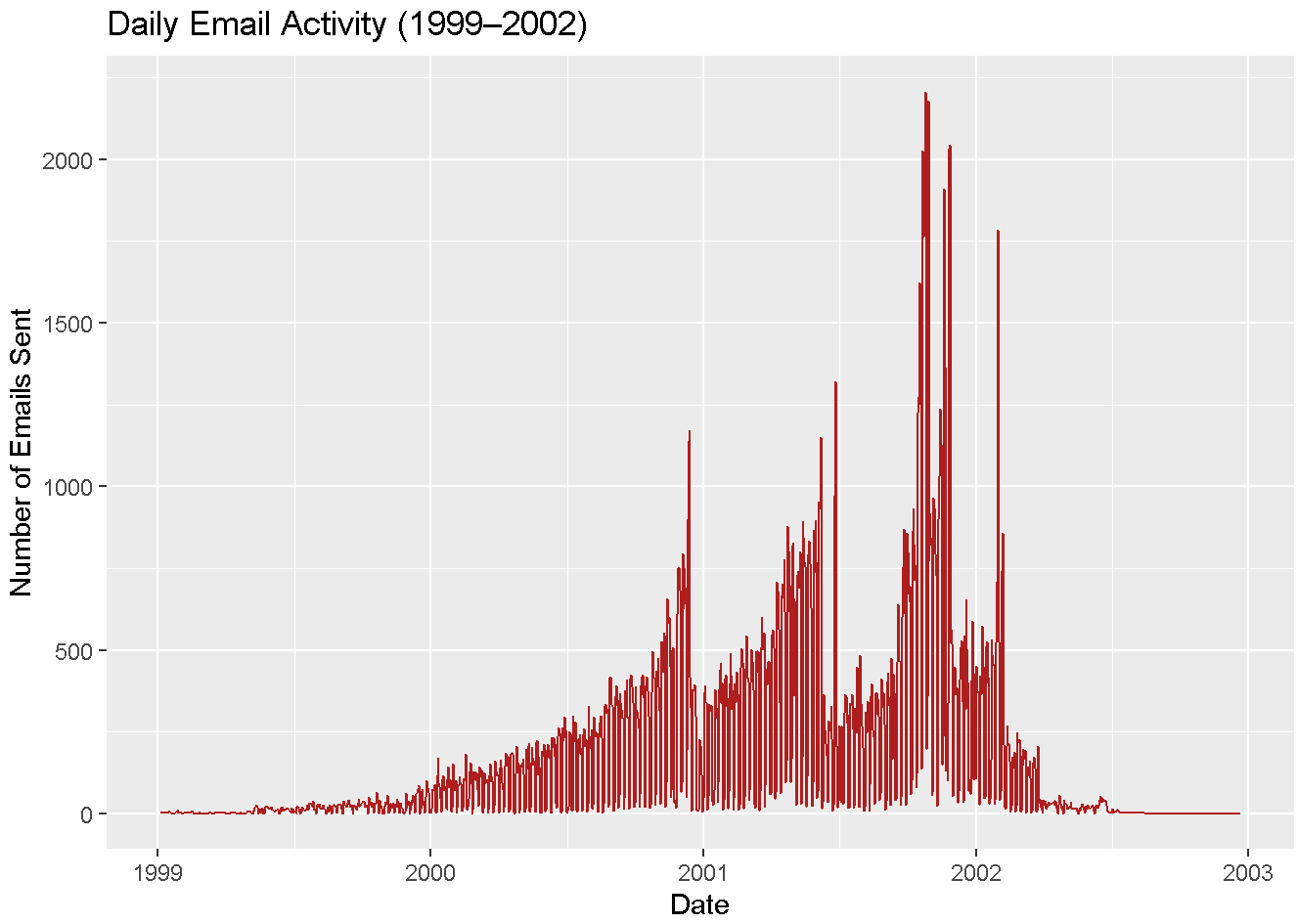
4. Email Activity Timeline

To analyze how email communication evolved over time, I counted the number of messages sent per day. and I applied a date filter to remove clearly invalid values, limiting to the period from January 1, 1999 to December 31, 2002.

```
## Warning: le package 'lubridate' a été compilé avec la version R 4.4.3
```

```
##
## Attachement du package : 'lubridate'
```

```
## Les objets suivants sont masqués depuis 'package:base':
##
##     date, intersect, setdiff, union
```



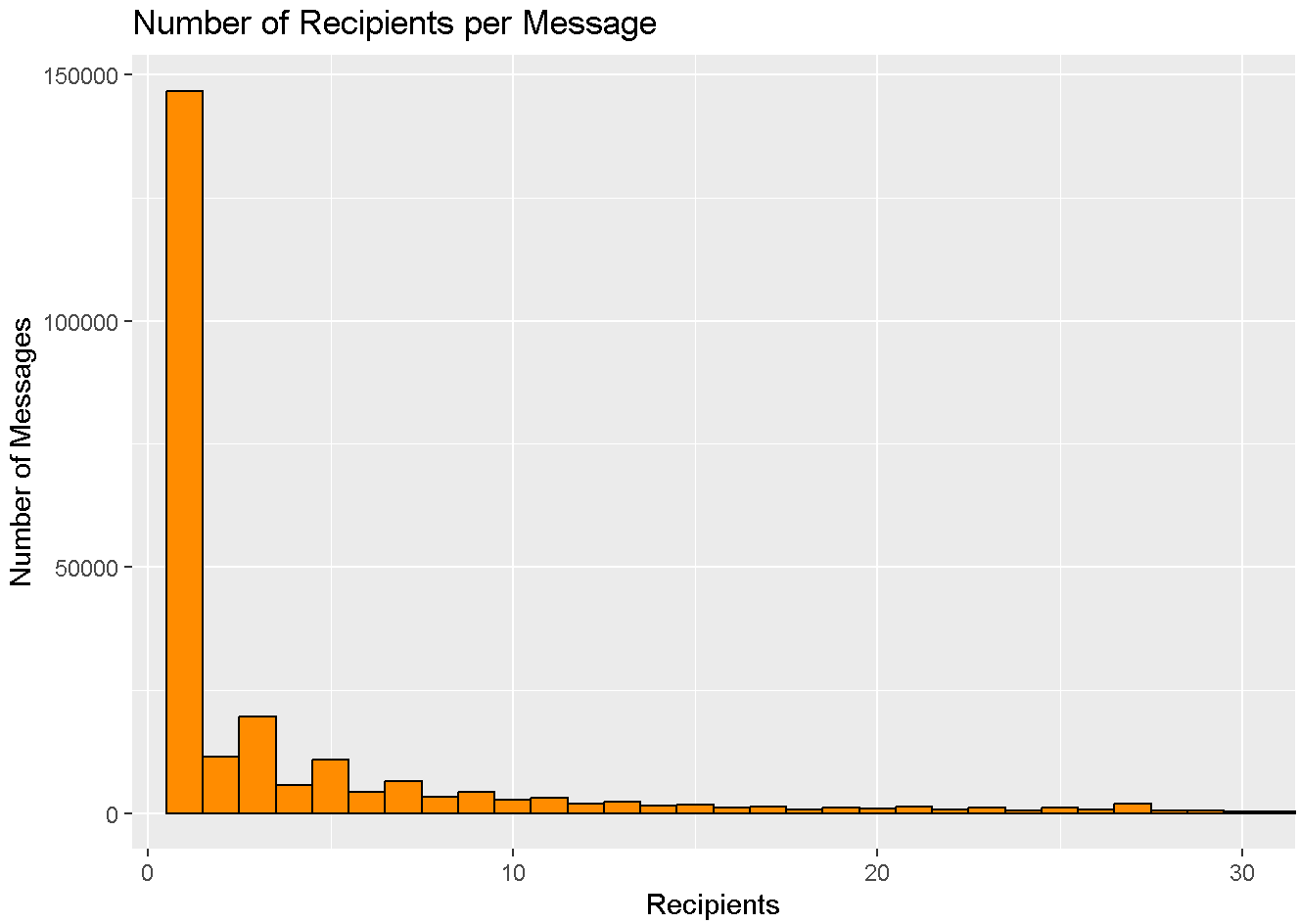
Explanation:

This timeline reveals clear **fluctuations** in email activity over the years. The most intense periods appear around the year **2001**, aligning with the months leading up to Enron's financial collapse in December 2001.

This suggests that internal communication likely intensified during periods of organizational stress, which is consistent with known crisis behaviors in large companies. The decrease in email volume in 2002 could reflect **downsizing, loss of staff**, or reduced operational complexity after the scandal broke.

5. Recipients per Message

We analyzed how many people received each message by counting how often each message ID appeared in the `recipientinfo` dataset. This includes TO, CC, and BCC recipients.



Explanation:

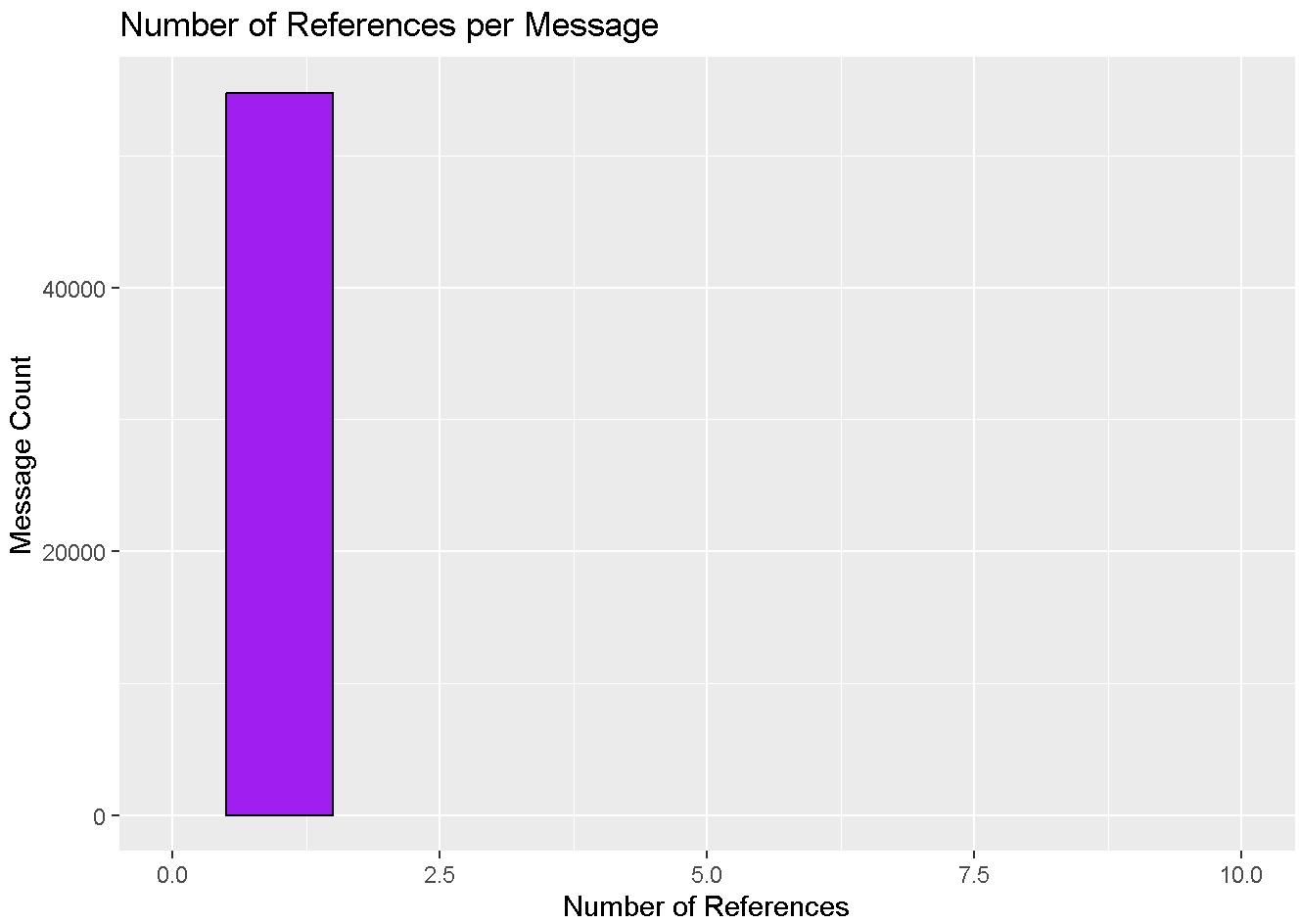
Most emails in the Enron dataset were sent to **1–5 recipients**, indicating that much of the communication was **targeted and personalized**. However, the long tail in the distribution shows that some messages were sent to dozens or even hundreds of people, which may reflect:

- Company-wide announcements
- Project updates
- Broadcast-style communication from management

This distribution shows a blend of private and mass communication behavior, typical in large organizations.

6. References per Message

The `referenceinfo` dataset links messages together using a reference ID. By counting how many messages each message references, we can estimate how often emails were part of ongoing threads or replies.



Explanation:

Most messages have **zero or one reference**, suggesting that many emails were not part of long threads. This implies that a large portion of Enron's email traffic may have consisted of **announcements, stand-alone messages, or short conversations**.

However, the presence of messages with multiple references shows that **some email chains were highly active**, possibly reflecting collaborative discussions or issue resolution threads — particularly within teams or departments.

This distribution highlights how different roles may lead to different patterns of communication: support staff might reply more often, while executives may send more one-directional messages.

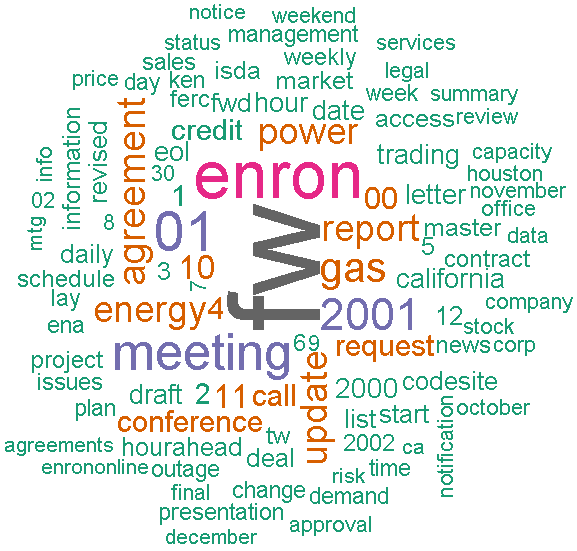
7. Content Analysis – Common Words in Email Subjects

To get an idea of the common themes in email communication, we tokenized the subject lines using the `tidytext` package and visualized the most frequent words using a word cloud.

```
library(tidytext)
library(wordcloud)

subject_words <- message_df %>%
  select(subject) %>%
  filter(!is.na(subject)) %>%
  unnest_tokens(word, subject) %>%
  anti_join(stop_words, by = "word") %>%
  count(word, sort = TRUE) %>%
  filter(n > 50)

wordcloud(words = subject_words$word,
          freq = subject_words$n,
          min.freq = 1,
          max.words = 100,
          random.order = FALSE,
          colors = brewer.pal(8, "Dark2"))
```



Explanation:

The word cloud highlights recurring words in email subject lines, such as “meeting”, “update”, “schedule”, or “project”. These terms reflect Enron’s **corporate routines**, such as coordinating meetings, providing progress reports, and discussing project planning.

This high-level view of email content confirms that much of the communication was related to **daily operations** rather than personal or unrelated messaging. If extended to full email bodies, content analysis could reveal **topics, sentiment**, or even detect **crisis communication patterns**.

Shiny App Summary

As part of this project, we developed an interactive Shiny application that allows users to explore the Enron email dataset visually and dynamically.

The app includes the following features:

- **Email activity filtering by date**: Users can specify a time range to focus on a particular period (e.g., before or after the Enron scandal).
- **Job Role Email Distribution**: A bar chart shows the number of emails sent by employees in different roles, such as Manager, Director, or Vice President.
- **Activity Timeline**: Displays a time-series line plot of daily email volume, helping users spot periods of high or low communication.
- **Top Senders**: Lists the most active email senders, with options to view sender names and roles.
- **Recipients Analysis**: Shows how many recipients are included per message, which helps understand how widely information was shared.
- **Basic Content Analysis**: A word cloud of common terms from email subject lines helps highlight communication themes.

The Shiny app allows users to interact with the dataset without writing any code. It supports **exploratory data analysis**, makes patterns visible, and provides a practical way to engage with complex corporate communication data.

In a real-world business context, this kind of application could help internal audit teams or data analysts investigate communication flows, detect anomalies, or identify key influencers within an organization.

Conclusion

This project provided a hands-on opportunity to explore a real-world communication dataset using R and Shiny. Starting with a large collection of internal Enron emails, we performed a step-by-step analysis covering:

- Identification of top email senders
- Communication patterns by job role
- Temporal analysis of email activity
- Recipient behavior
- Message referencing and reply dynamics
- A basic content analysis of subject lines

We then built an interactive Shiny app that enabled dynamic exploration of this data through visualizations and filters.

Key Insights:

- Communication was **highly concentrated** among a small number of individuals.
- **Managers and Vice Presidents** were responsible for much of the email traffic.
- **Email volume peaked** before the company’s collapse, possibly indicating organizational stress.
- Most emails were sent to **just a few recipients**, though some were shared widely.

- Subject line content showed a strong focus on **project coordination and scheduling**.

These insights reflect how communication flowed within a large corporation during both routine and turbulent times.

Appendix: Code Used in Visualizations

This section collects the R code used to generate the visualizations and tables in this report.

---

1. Load and Prepare the Data

```
load("Enron.RData")
message_df <- message
```

---

2. Top Email Senders

```
top_senders <- message_df %>%
  group_by(sender) %>%
  summarise(total_sent = n()) %>%
  arrange(desc(total_sent)) %>%
  slice_head(n = 10)
```

---

3. Top Senders Named

```
top_senders_named <- message_df %>%
  group_by(sender) %>%
  summarise(total_sent = n()) %>%
  arrange(desc(total_sent)) %>%
  slice_head(n = 10) %>%
  left_join(employeelist, by = c("sender" = "Email_id")) %>%
  mutate(full_name = paste(firstName, lastName))
```

---

4. Emails by Job Role

```
emails_by_role_full <- message_df %>%
  left_join(employeelist, by = c("sender" = "Email_id")) %>%
  mutate(status = ifelse(is.na(status), "Unknown", as.character(status))) %>%
  group_by(status) %>%
  summarise(total_sent = n()) %>%
  arrange(desc(total_sent))
```

---

5. Activity Timeline

```
emails_by_day <- message_df %>%
  filter(date >= as.Date("1999-01-01") & date <= as.Date("2002-12-31")) %>%
  group_by(date) %>%
  summarise(daily_emails = n())
```

6. Recipients per Message

```
recipients_per_message <- recipientinfo %>%
  group_by(mid) %>%
  summarise(num_recipients = n())
```

7. References per Message

```
references_per_message <- referenceinfo %>%
  group_by(mid) %>%
  summarise(ref_count = n())
```

8. Subject Content Analysis

```
subject_words <- message_df %>%
  select(subject) %>%
  filter(!is.na(subject)) %>%
  unnest_tokens(word, subject) %>%
  anti_join(stop_words, by = "word") %>%
  count(word, sort = TRUE) %>%
  filter(n > 50)
```