University Email Analysis Project Report

By: Mohammed Sateh Salem

Introduction

As a university student, I often find myself overwhelmed by the volume of emails I receive daily. These emails range from academic updates and assignment deadlines to administrative notices. I noticed that the constant influx of emails, especially during peak academic periods, was becoming a source of stress for me. This realization motivated me to analyse my own email patterns to better understand when and why these emails are sent and how I could manage them more effectively. By delving into this project, I hoped to uncover actionable insights that could reduce the stress of constantly monitoring my inbox and help me organize my communication more strategically.

Objectives

The objectives of this project were deeply personal and aimed at addressing my specific challenges with email management:

1. **Identifying the Most Frequent Email Senders**: To determine which individuals, departments, or systems dominate my inbox, helping me prioritize responses and understand which contacts require my attention.
2. **Determining the Best Times to Check Emails**: To analyse my email activity patterns and identify optimal times to check my inbox, reducing unnecessary distractions.
3. **Analysing Monthly and Weekly Trends**: To identify peak periods of email activity, such as midterms and finals, and prepare myself better for these high-volume times.
4. **Categorizing Emails by Topic and Sender Type**: To understand how my emails are distributed across academic, administrative, and extracurricular topics, helping me focus on the most important messages.
5. **Visualizing Hourly Distribution**: To pinpoint the busiest hours of the day for receiving emails and organize my schedule accordingly.
6. **Exploring Course-Specific Trends**: To determine which courses generate the most email traffic and understand why some courses, like **CS204** and **MATH203**, appear to send more emails—likely because I took these courses multiple times.

Dataset Description

The dataset for this project was a highly personal collection of emails exported from my university account via Google Takeout and Thunderbird, covering the period from **Monday, 29 August 2022, to Wednesday, 20 November 2024**. This dataset included:

- **Email Timestamps**: To track when emails were sent or received.

- **Sender Details**: To identify frequent senders and key contacts.
- **Subject Lines**: To categorize and understand the content of my emails.
- **Day, Month, and Year Metadata**: To analyse trends over time.
- **Weekday Information**: To uncover weekly patterns in email traffic.
- **Time Details**: To pinpoint specific hours of high activity.

Additionally, some courses, such as **CS204** and **MATH203**, had higher email volumes because I took these courses more than once during the analysed period.

Methodology

1. Data Extraction

- I extracted the email data from .mbox files using Google Takeout and Thunderbird.
- Relevant fields such as timestamps, sender details, and subject lines were parsed for analysis.

2. Data Cleaning

- The data was pre-processed to ensure it was accurate and usable:
    - Timestamps were parsed and formatted to extract specific components like day, month, and year.
    - Sender details were cleaned using regular expressions to separate names from email addresses.
    - Missing or invalid data was handled by assigning placeholders or excluding incomplete records.

3. Data Analysis

- I used Python libraries like Pandas to analyse:
    - The frequency of emails by time of day and day of the week.
    - Patterns across weeks and months in the semester.
    - Trends in emails sent by specific courses and departments.

4. Data Visualization

- Visualizations were created using Matplotlib, Seaborn, and WordCloud to represent the data in a meaningful way:
    - Bar charts for frequent senders and hourly distributions.
    - Line graphs for monthly trends.
    - Heatmaps for weekday and hourly email activity.
    - Word clouds to highlight common keywords in email subject lines.

Hypothesis

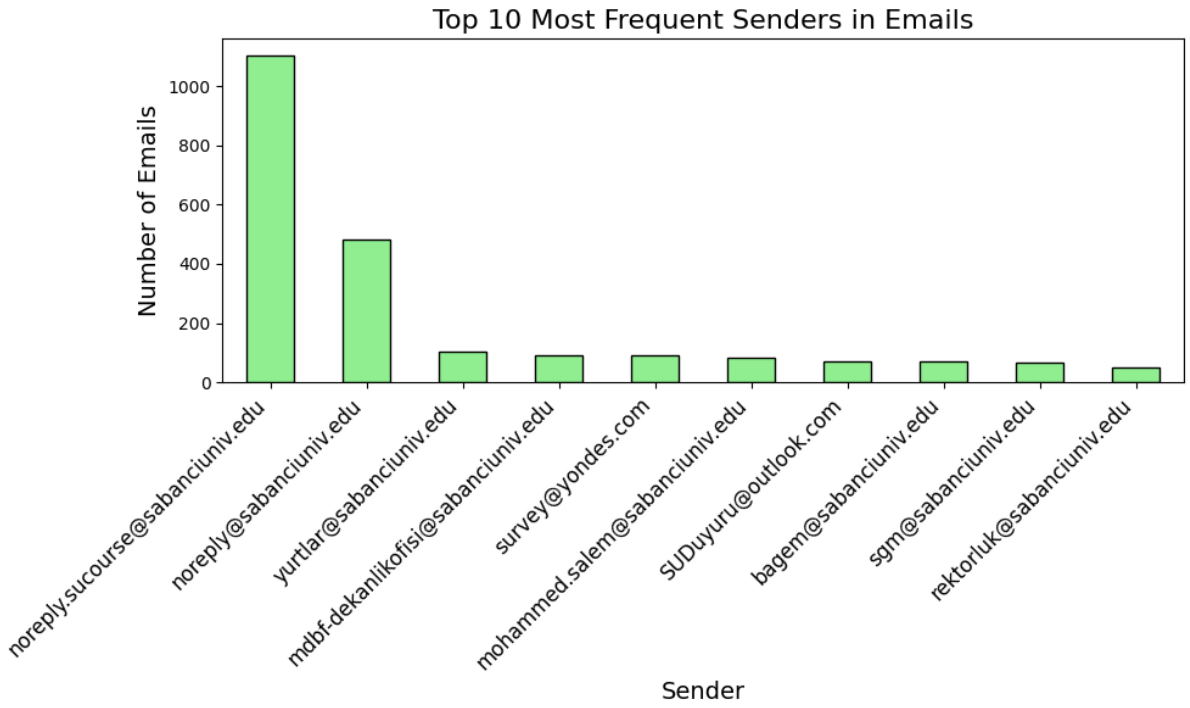
Based on my experiences with university email patterns, I formulated the following hypothesis:

1. **Peak Email Times**: Most emails will be received during the morning hours (e.g., 9:00 AM to 11:00 AM), coinciding with academic and administrative working hours.
2. **Frequent Senders**: Automated systems (e.g., noreply addresses) will dominate the inbox, followed by emails academic departments.
3. **Course-Specific Trends**: Emails related to Computer Science courses (e.g., CS204) will be more frequent due to the volume of assignments, deadlines, and project updates.
4. **Weekly Patterns**: Email traffic will peak midweek (e.g., Tuesdays and Thursdays) and drop significantly on weekends.
5. **Subject Keywords**: Keywords like "submission," "exam," and "assignment" will dominate email subject lines, reflecting academic priorities.

Findings

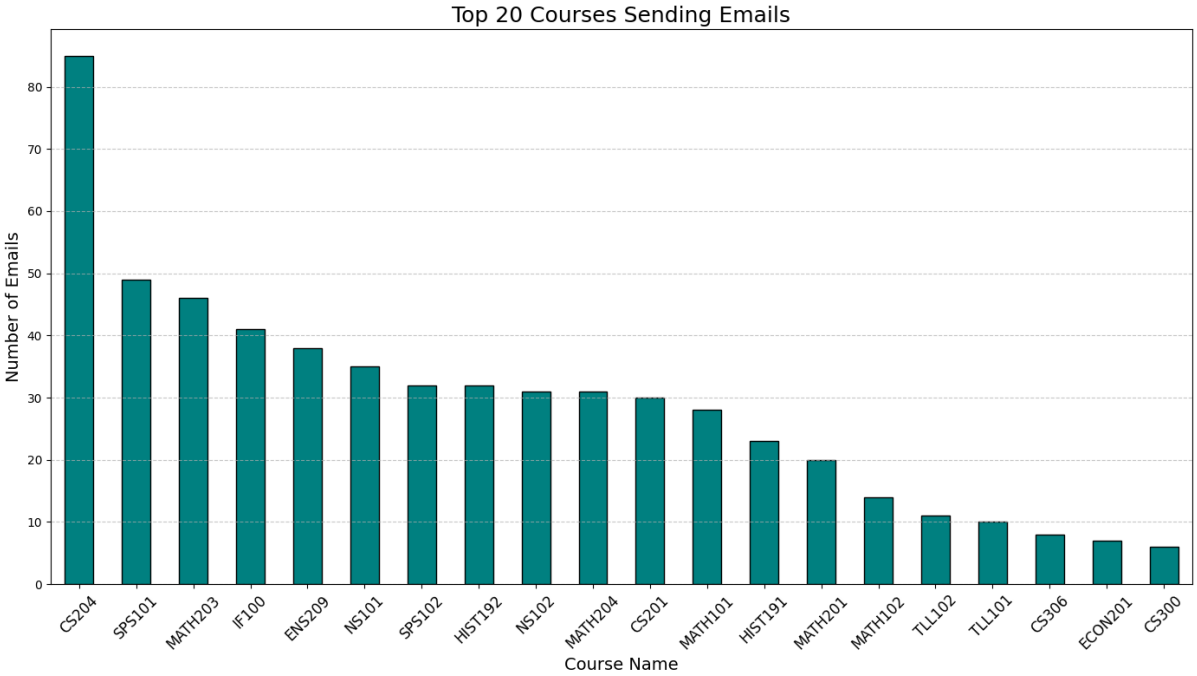
1. Most Frequent Email Senders

- **Observation**: Automated systems like **noreply.sucourse@sabanciuniv.edu** and **noreply@sabanciuniv.edu** dominated my inbox.
- **Key Insight**:
    - The **bar chart** showed the top 10 email senders. The **x-axis** represents the senders, and the **y-axis** indicates the total number of emails received.
    - On closer inspection, **noreply.sucourse@sabanciuniv.edu** was identified as the primary source of course-related announcements sent through SUCourse by professors.
- **Implication**: These findings connect frequent sender data directly to course-specific trends, helping me understand the academic focus of my email traffic.



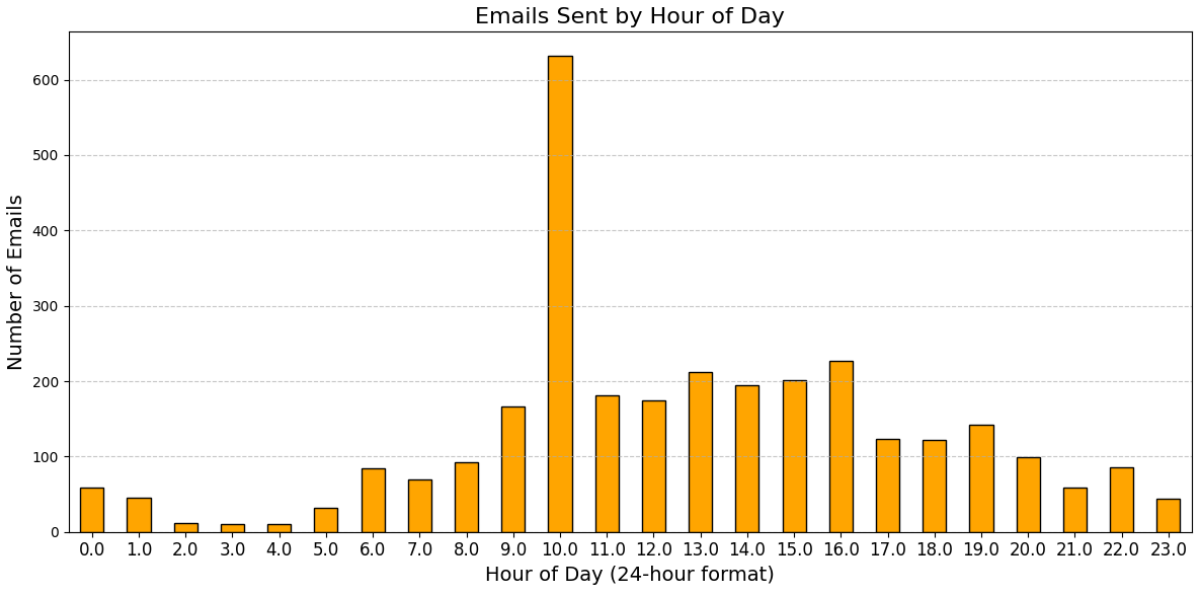
2. Course-Specific Trends

- **Observation**: Courses like **CS204** and **SPS101** generated the highest email volumes.
- **Key Insight**:
    - The **bar chart** of email frequency by course revealed the significant contribution of SUCourse announcements. The **x-axis** represents the course codes, and the **y-axis** shows the total number of emails.
    - Courses like **CS204** and **MATH203** stood out because I took these courses multiple times, inflating the email counts.
- **Implication**: Recognizing course-specific trends allows for prioritizing emails related to these high-traffic courses, especially during key academic periods.



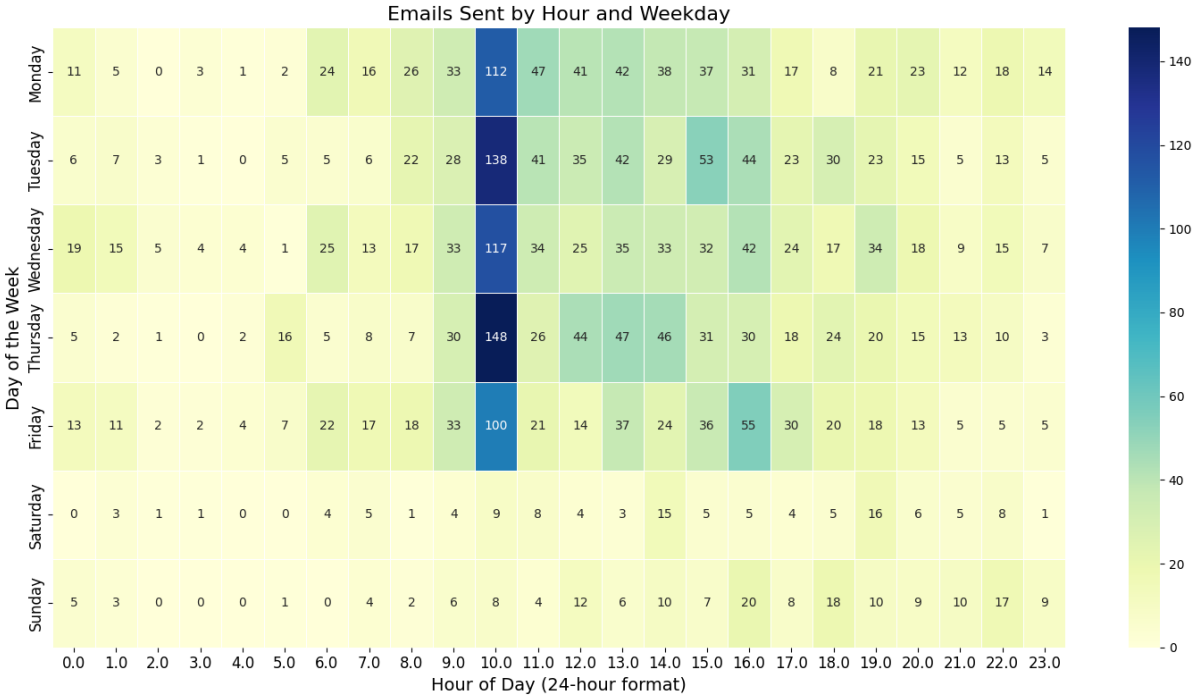
3. Peak Email Times

- **Observation**: The busiest hour for receiving emails was **10:00 AM**.
- **Key Insight**:
    - The **bar chart** of hourly email activity highlights when emails peak during the day. The **x-axis** represents the hours (in 24-hour format), and the **y-axis** shows the total number of emails.
    - The 10:00 AM peak was heavily influenced by daily **GazeteSU** announcements sent at 10:30 AM.
- **Implication**: Understanding peak hours helps optimize inbox management, such as scheduling email checks during high-traffic times.



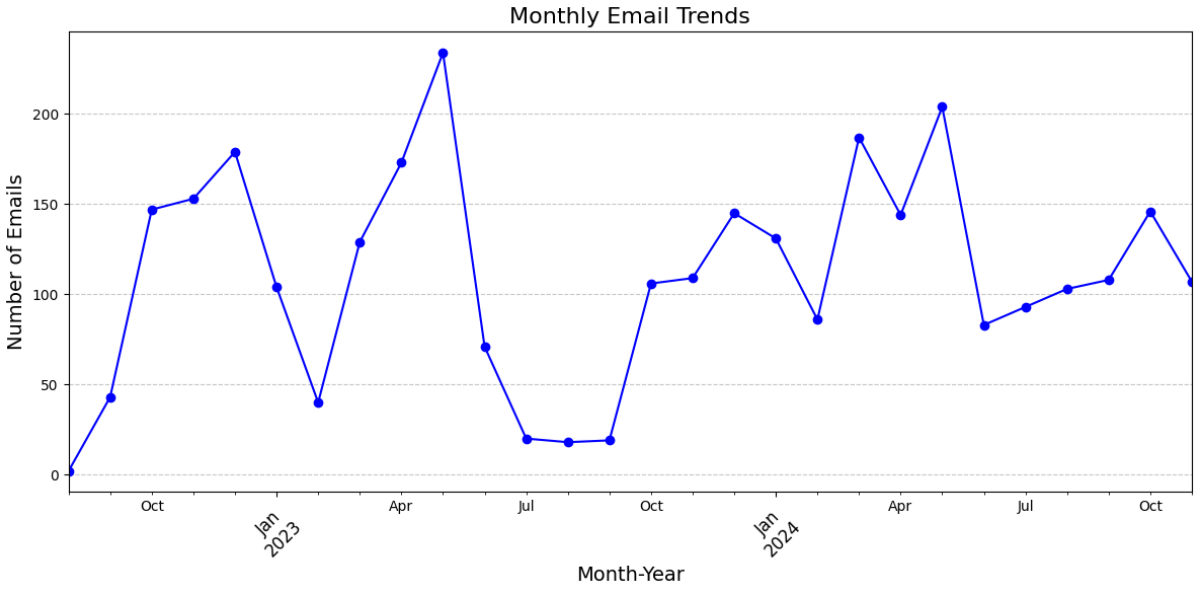
4. Weekly Trends

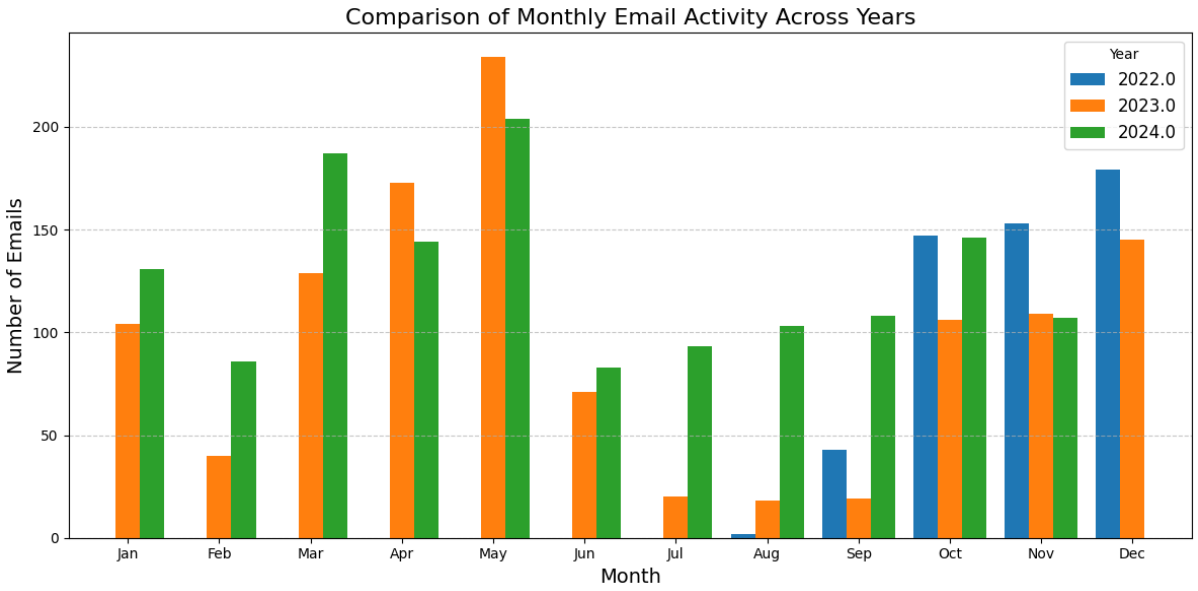
- **Observation**: Email traffic peaked on **Tuesdays** and **Thursdays**, with significantly lower activity on weekends.
- **Key Insight**:
    - The **heatmap** visualized email activity across days and hours. The **x-axis** represents the hours, the **y-axis** represents the days, and the colour intensity indicates email volume.
    - These peaks corresponded to typical academic and administrative schedules.
- **Implication**: Recognizing weekly patterns allows for better time management, such as focusing on important updates during midweek.



5. Monthly and Yearly Trends

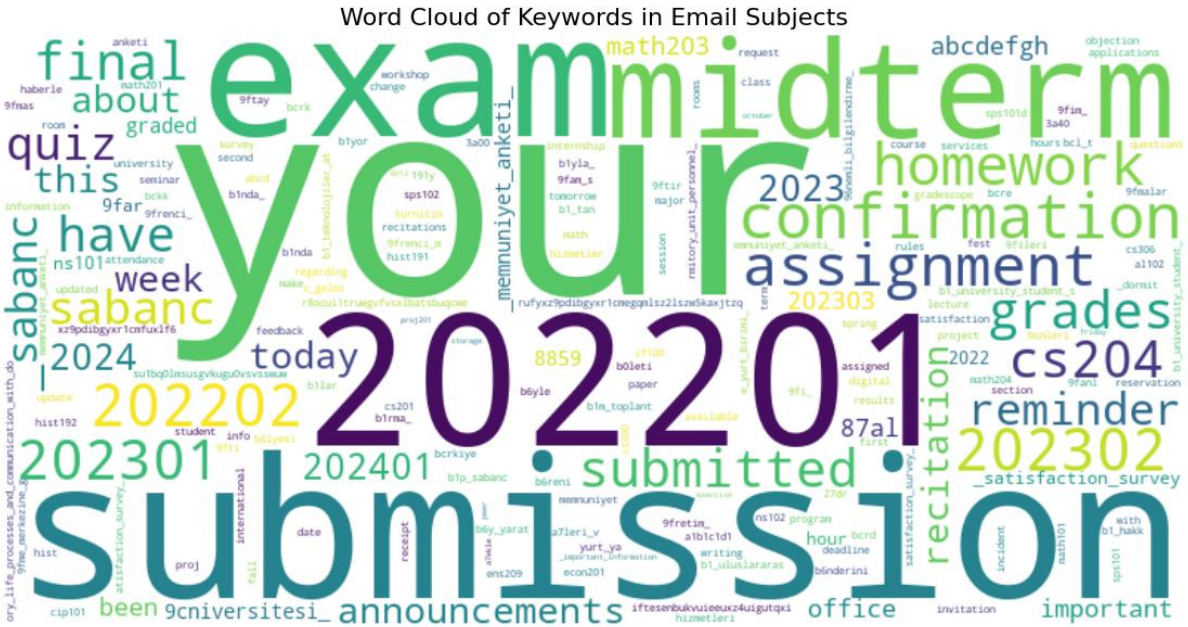
- **Observation**: Email activity spiked in **May** and **December**.
- **Key Insight**:
    - The **line graph** for monthly trends revealed seasonal peaks. The **x-axis** represents the months, and the **y-axis** shows the total number of emails.
    - The **Comparison of Monthly Email Activity Across Years** graph highlighted the consistency of these trends across 2022, 2023, and 2024.
    - These spikes align with midterms, finals, and end-of-semester deadlines.
- **Implication**: Anticipating high-volume months helps prepare for academic and administrative workload.





6. Keywords in Subject Lines

- **Observation**: Keywords like **"submission," "exam," "grades,"** and **"assignment"** were prevalent.
- **Key Insight**:
    - The **word cloud** visualized common terms in email subject lines, emphasizing the academic nature of most emails.
    - These keywords reflected the time-sensitive and task-oriented focus of the emails.
- **Implication**: Recognizing dominant themes in subject lines helps categorize emails and prioritize tasks effectively.



Graph Stories: How Each Graph Builds on the Previous

Bar Chart of Frequent Senders

The analysis began by identifying the top 10 most frequent senders. This gave a clear picture of which automated systems, professors, or departments contributed most to my inbox. Diving deeper into the top sender **noreply.sucourse@sabanciuniv.edu**, I discovered these emails were announcements made by professors through SUCourse, connecting sender activity to course-specific trends.

Bar Chart of Course-Specific Trends

With broader trends established, I drilled down into course-specific email traffic. Courses like CS204 and MATH203 stood out, with higher volumes due to being taken multiple times. This graph helped contextualize academic email traffic within my inbox and tied directly to SUCourse announcements by professors.

Bar Chart of Hourly Email Activity

After identifying key senders, I wanted to understand when emails were sent throughout the day. The hourly distribution showed peak activity at 10:00 AM, influenced heavily by automated daily emails like those from GazeteSU. This insight connected sender behaviour to temporal patterns.

Heatmap of Hourly and Weekly Activity

Building on the hourly distribution, I added a weekly dimension to explore how email activity varied by day. The heatmap revealed that Tuesdays and Thursdays were the busiest days, complementing the earlier insights on hourly peaks.

Line Graph of Monthly Trends

To understand long-term patterns, I analysed monthly email activity. This graph showed clear spikes in May and December, aligning with midterms, finals, and semester-end deadlines. These trends built on the weekly and hourly patterns by highlighting seasonal variations.

Word Cloud of Subject Line Keywords

Finally, the word cloud synthesized the content of emails, highlighting academic priorities like "submission," "exam," and "assignment." This visualization tied together the sender, temporal, and course-specific analyses by focusing on content themes.

Implications

1. **Optimized Inbox Management**:
   Knowing that email traffic peaks around 10:00 AM, I can schedule email checks during this window to stay updated without distractions. Avoiding late-night inbox checks can help create better boundaries between academic and personal time.
2. **Course Prioritization**:
   Identifying courses like CS204 and SPS101 as high-traffic sources helps me focus on these emails for better academic organization. Recognizing that many course-related emails come from SUCourse allows me to filter and prioritize announcements effectively.
3. **Stress Reduction**:
   Anticipating peak periods, such as May and December, helps me prepare for high email traffic and manage my academic workload more efficiently. Understanding weekly and hourly patterns allows me to create a more structured approach to handling my inbox, reducing the anxiety of unexpected emails.
4. **Improved Time Management**:
   By focusing on high-traffic times, such as Tuesday mornings, I can allocate specific slots in my schedule to process emails without interrupting other tasks.
5. **University Recommendations**:
   These insights could guide universities to consolidate emails into fewer, more structured announcements, reducing email overload for students. Universities might also consider scheduling important emails outside of peak times to ensure better visibility and engagement.

Limitations

- **Dataset Bias**: This analysis is based solely on my personal dataset, and trends may not generalize to other students.
- **Incomplete Data**: Some emails lacked metadata or had inconsistent subject lines, limiting the depth of the analysis.
- **Keyword Limitations**: Subject line categorization was challenging due to inconsistent phrasing and missing course codes.

Future Work

1. **Expanded Dataset**: Including anonymized datasets from other students for broader insights.
2. **Automated Categorization**: Using Natural Language Processing (NLP) to classify emails by content and urgency.
3. **Response Time Analysis**: Investigating how quickly I respond to emails to identify inefficiencies.

Conclusion

This project provided a personal lens into my email habits, helping me understand and manage my inbox more effectively. By identifying trends in sender activity, peak times, and course-specific communications, I was able to reduce email-related stress and improve time management. Expanding this analysis in the future could offer even more valuable insights and applications for students and university administrations alike.