1)Data Exploration and Cleaning:

- Upon first look the data seems normal email details with sender and receiver(To,CC,BCC) address and timings.
- Next Step is to check the null values, null values in the BCC and CC is acceptable as it is not mandatory to add receiver list in CC and BCC.
- Null values in Sender, To, Sent , Received fields means data inconsistency as these values cannot be null in proper email representation.
- Once null values are removed, next step is to check for data validity, I,e checking if the fields has proper values.
- To check for validity, each column is filtered with the relevant format. The records which does not match the format are saved in CSV for further analysis.
- Upon analyzing the saved CSV files, (S1→Senders, S2→To, S3→CC, S4→BCC, S5→Sent, S6→Received). There were some records with date and time format in CC and BCC and some emails in Sent and Received.
- After removing the above records the final data was created for further operations.

2) Finding Machine Generated emails:

- Generally, Machine generated emails will have less senders and large receivers as this is the purpose of automating an email.
- To filter the machine generated emails, Sender column was filtered with senders count greater than 250 and To column was filtered with receiver count less than 25.
- 381 unique machine generated emails were found which are displayed in python notebook
- These emails were removed from the data for further analysis

3) Finding Communities/Teams:

- To find the communities, two methods were used
  - First one was to group the emails according to the team/Department names.
  - Second approach was to group the emails according to the domain names.
- 77 unique teams were found, these teams and their respective email addresses are saved in “community.csv”
- 886 unique domains were found after filtering the email addresses with count greater than 50 for each domain. Details are saved in “domaincomm.csv”

3)Finding most trusted email addresses:

- It is assumed that the people mentioned in the BCC are most trusted, as in general setup managers and people with higher responsibilities are included in BCC.
- These emails are filtered for occurrence greater than 500.
- 39 unique email addresses were found to be most trusted. These emails and their frequency counts are displayed in python notebook.