PREDICT 420. Individual Assignment 4

Consulting Request. It seems that Yahoo! Mail has undergone various redesign efforts in recent years, some intended to make the system look more like Google's gmail. Despite these changes, Yahoo! Mail has experienced a decline in usage, while Google's gmail has been on the rise. See

http://www.google.com/trends/explore?hl=en-US#q=Gmail,+Yahoo+Mail,+Hotmail,+AOL+Mail,+Outlook&cmpt=q

Concerned about usage trends, executives at Yahoo! are hoping that you will be available to serve as an outside consultant to the firm. First, they would like you to provide a summary of what you have learned about corporate e-mail usage by working with the Enron e-mail archive, one of the few e-mail databases in the public domain. How do corporate users utilize e-mail folders? How long do users maintain messages and how large of an e-mail storage area is required for the typical user? How long are to-address, cc-address, and bcc-address lists? What are typical message (body) lengths in kilobytes? It is understood that much of the relevant data may be messy and incomplete. Do your best to characterize corporate e-mail usage. Feel free to indicate places where the Enron data are incomplete or where your analysis should be regarded as preliminary. Let Yahoo! executives know if there are any sampling issues. To what extent can the Enron e-mail usage be used to generalize to e-mail usage at major U.S. corporations?

Use any method that you feel is appropriate for gathering, cleaning, and analyzing relevant data from the Enron e-mail archive. The **`pandas`** package, for example, may be especially useful in generating descriptive statistics for your report. Descriptive statistics are sufficient. No inferential statistics or predictive models are required.

Deliverable:

1. Submit on Canvas your IPython Notebook script.
2. PDF document that has the code snippets and your output answers for the different deliverables