Sunday, December 29, 2013

What's in My Inbox? Data Analysis of Outlook

Introduction
Email is the bane of our modern existence.

Who of us hasn't had a long, convoluted, back-and-forth email thread going on for days (if not weeks) in order to settle an issue which could have been resolved with a simple 5 minute conversation?

With some colleagues of mine, email has become so overwhelming (or their attempts to organize it so futile) that it brings to my mind Orwell's workers at the Ministry of Truth in 1984 and their pneumatic tubes and memory holes - if the message you want is not in the top 1% (or 0.01%) of your inbox and you don't know how to use search effectively, then for all intents and purposes it might as well be gone (see also: **Snapchat**).

Much has been written on the subject of why exactly we send and receive so much of it, how to best organize it, and whether or not it is, in fact, even an effective method of communication.

At one time even **Gmail** and the **concept of labels** was revolutionary - and it has done some good in organizing the ever-increasing deluge that *is* email for the majority of people. Other attempts have sprung up to tame the beast and make sense of such a flood of communication - most notably in my mind **Inbox Zero**, the simply-titled smartphone app **Mailbox**, and MIT's recent data visualization project **Immersion**.

But email, with all its systemic flaws, misuse, and annoyances, is definitely here for good, no

question. What a world we live in.

But I digress.

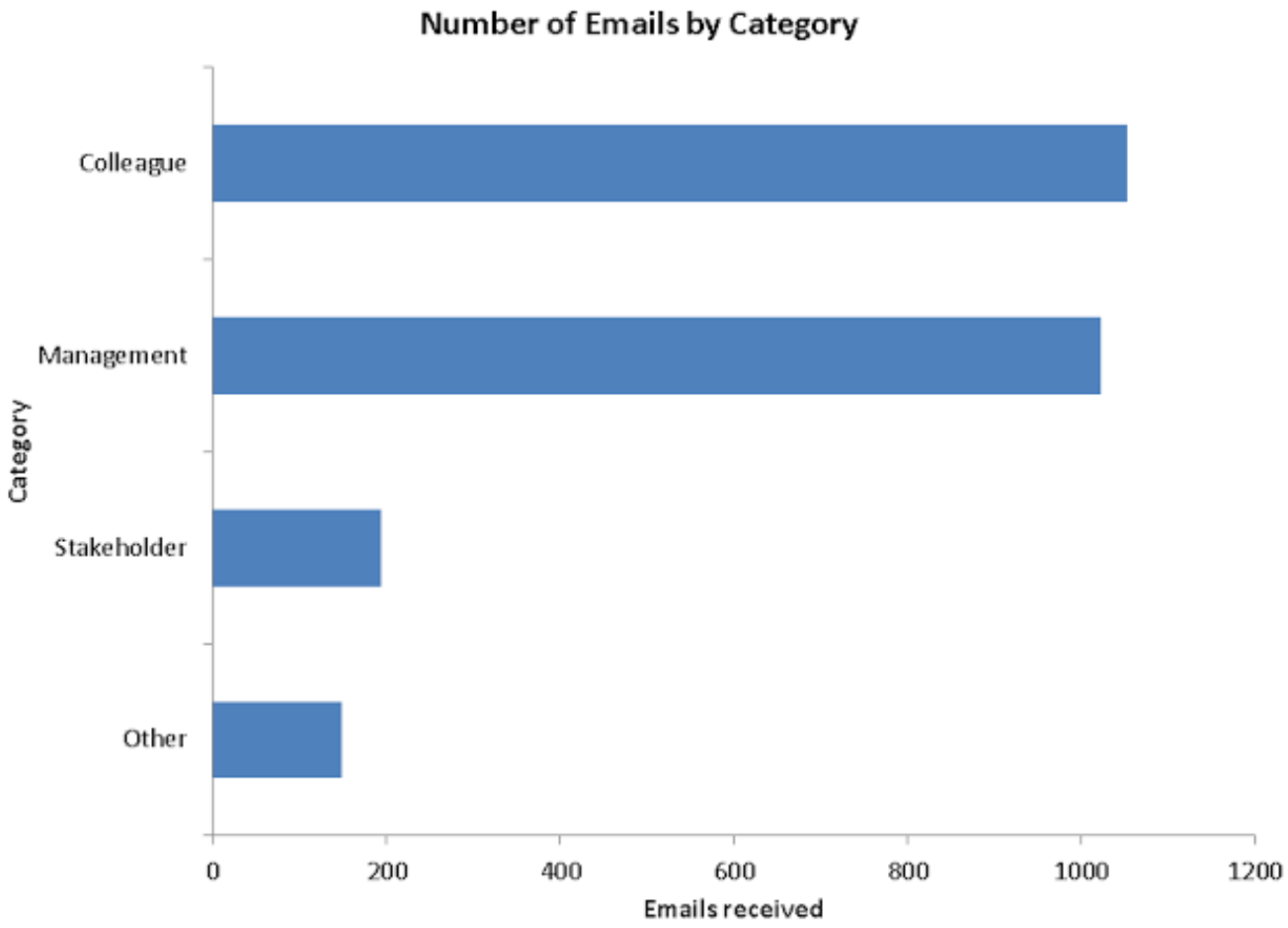
Background
I had originally hoped to export everything from Gmail and do a very thorough analysis of all my personal email. Though **this is now a lot easier than it used to be**, I got frustrated at the time trying to write a Python script and moved on to other projects.

But then I thought, hey, why not do the same thing for my work email? I recently discovered that it's quite easy to export email from Outlook (as I detailed **last time**) so that brings us to this post.

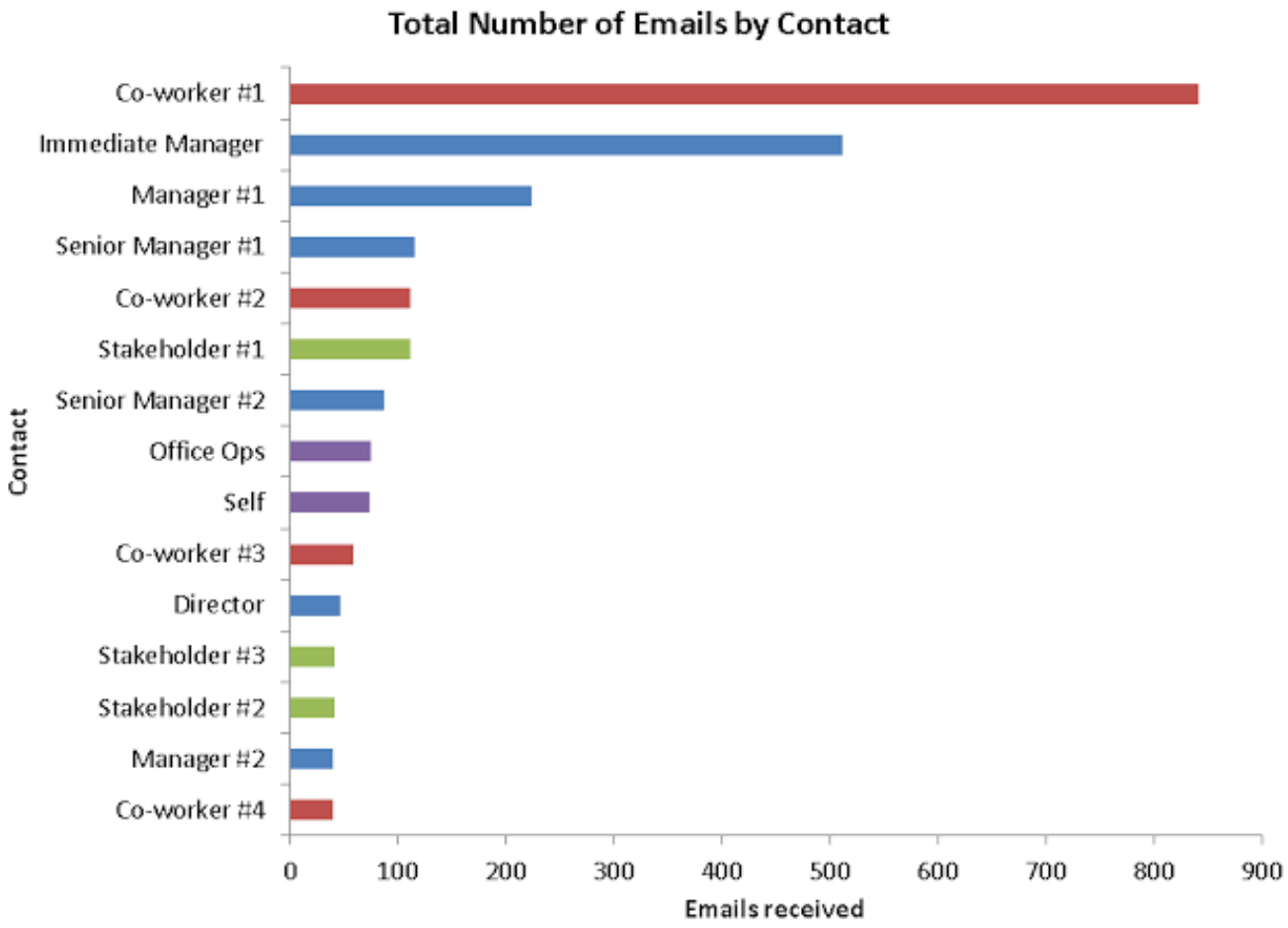
I was somewhat disappointed that Outlook can only export a folder at a time (which does not include special folders such as search folders or 'All Mail') - I organize my mail into folders and wanted an export of all of it.

That being said, the bulk probably does remain in my inbox (4,217 items in my inbox resulted in a CSV that was ~15 MB) and we can still get a rough look using what's available The data cover the period from February 27th, 2013 to Nov 16th, 2013.

Email by Contact
First let's look at the top 15 contacts by total number of emails. Here are some pretty simple graphs summarizing that data, first by category of contact:



In the top 15, split between co-workers/colleagues and management is pretty even. I received about 5 times as much email from coworkers and managers as from stakeholders (but then again a lot of the latter ended up sorted into folders, so this count is probably higher). Still, I don't directly interact with stakeholders as much as some others, and tend to work with teams or my immediate manager. Also, calls are usually better.



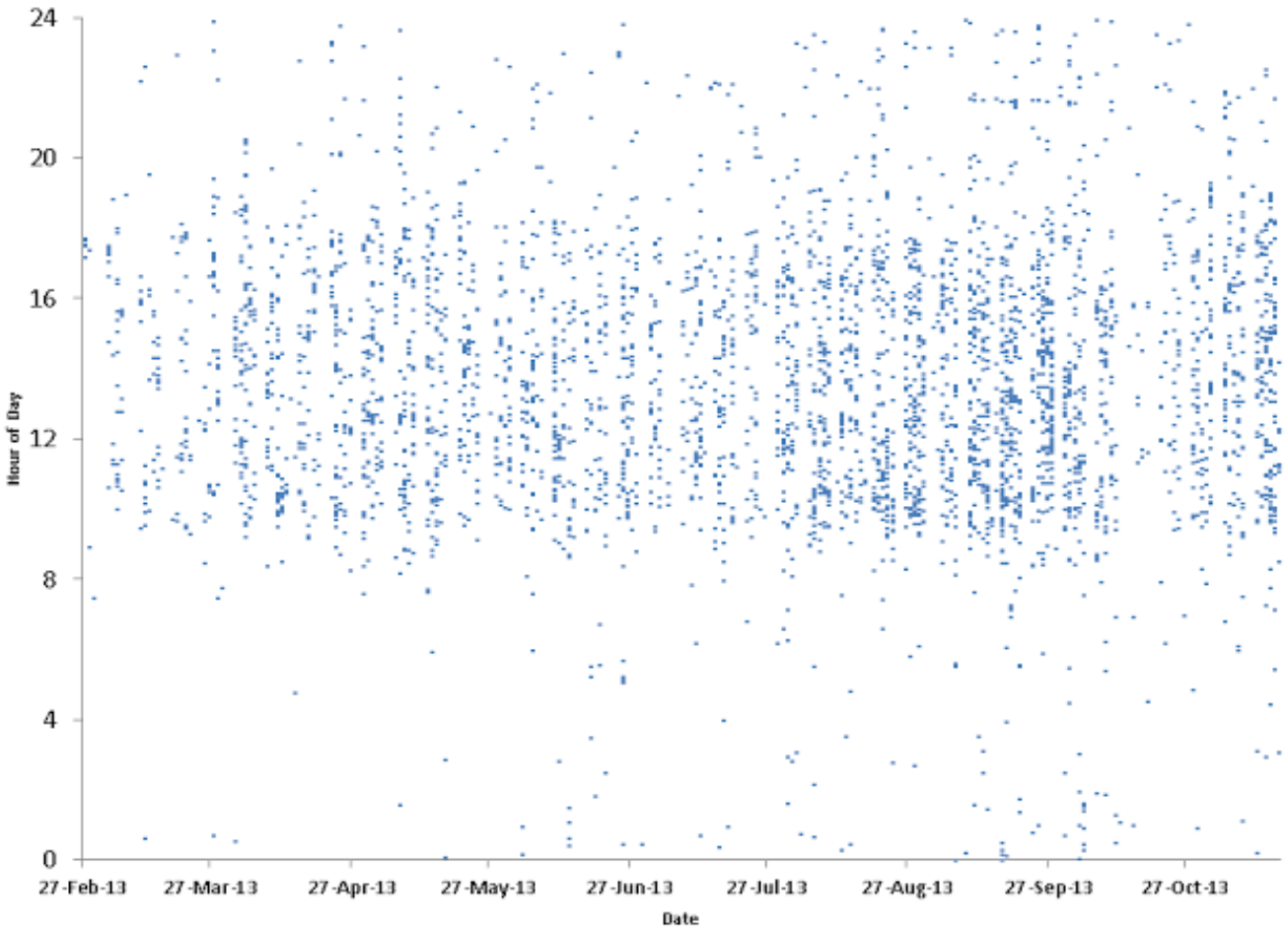
Here you can see that I interacted primarily with my immediate colleague and manager the most, then other management, and the remainder further down the line are a mix which includes email to myself and from office operations. Also of note - I don't actually receive ***that*** much email (I'm more of a "in the weeds" type of guy) or, as I said, much has gone into the appropriate folders.

Time-Series Analysis

The above graphs show a very simplistic and high level view of what proportion of email I was receiving from who (with a suitable level of anonymity, I hope). More interesting is a quick and simple analysis of patterns in time of the volume of email I received - and I'm pretty sure you already have an idea of what some of these patterns might be.

When doing data analysis, I always feel it is important to first visualize as much of the data as practically possible - in order to get "a feel" for the data and avoid making erroneous conclusions without having this overall familiarity (as I noted in an **earlier post**). If a picture is worth thousand words then a good data visualization is worth a thousand keystrokes and mouse clicks.

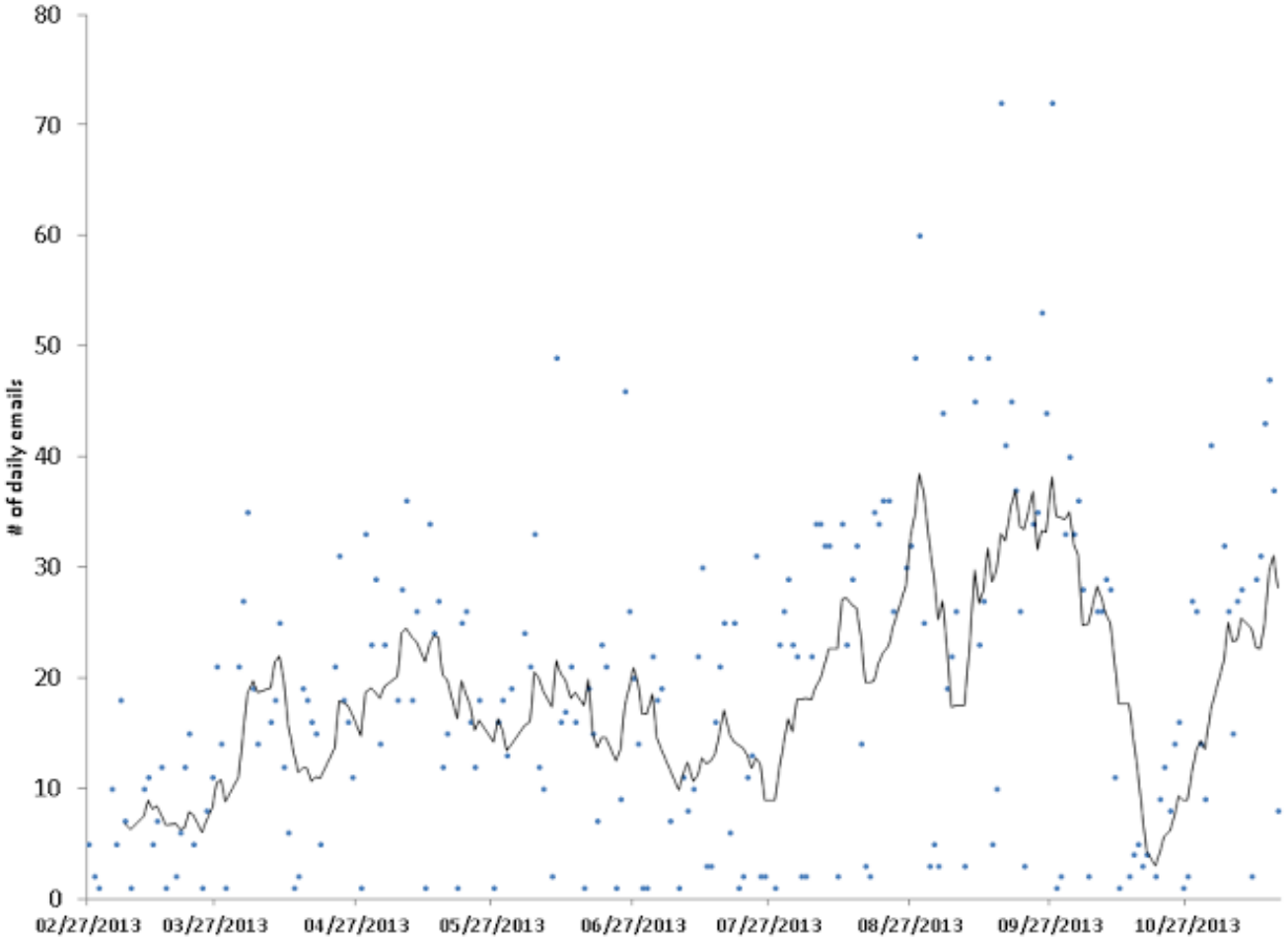
Below is a simple scatter plot all the emails received by day, with the time of day on the y-axis:



This scatterpolot is perhaps not immediately particulary illuminating, however it already shows us a few things worth noting:

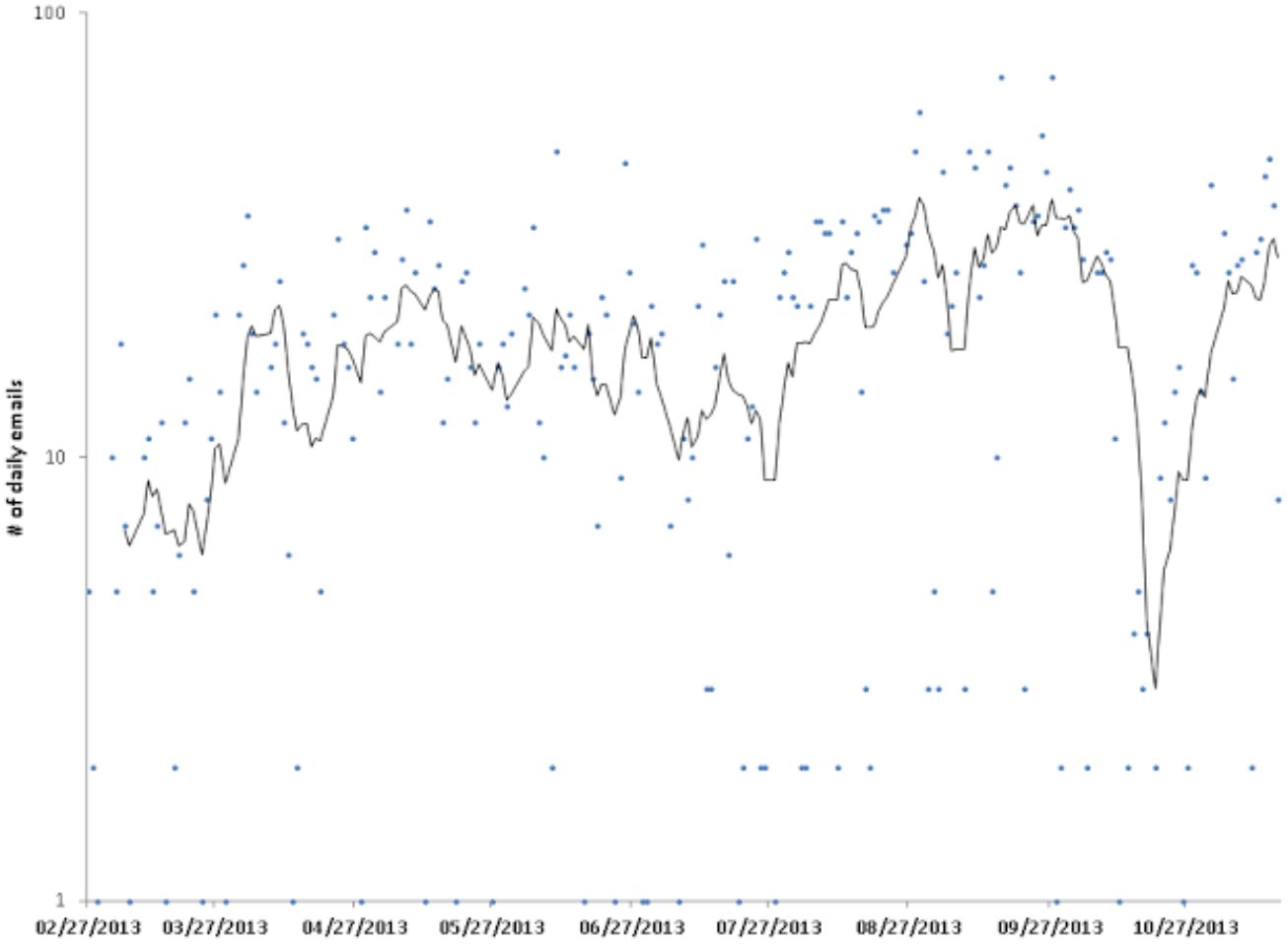
- the majority of emails appear in a band approximately between 8 AM and 5 PM
- there is increased density of email in the period between the end of July and early October, after which there is a sparse interval until mid-month / early November
- there appears to be some kind of periodic nature to the volume of daily emails, giving a "strip-like" appearance (three guesses what that periodic nature is...)

We can look into this further by considering the daily volume of emails, as below. The black line is a 7 day moving average:

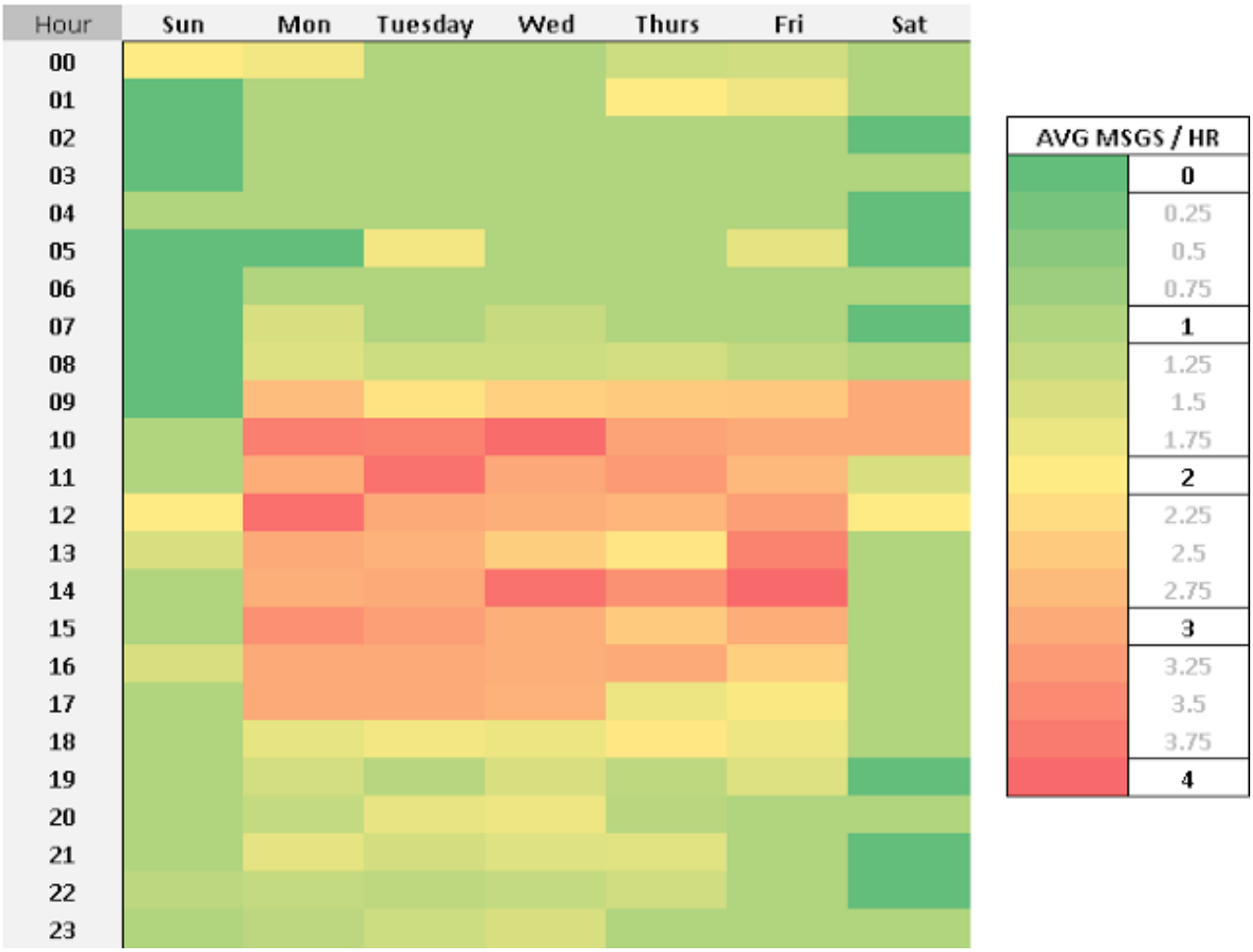


We can see the patterns noted above - the increase in daily volume after 7/27 and the marked decrease mid-October. Though I wracked my brain and looked thoroughly, I couldn't find a *specific* reason why there was an increase over the summer - this was just a busy time for projects (and probably not for myself sorting email). The marked decrease in October corresponds to a period of **bench time**, which you can see was rather short-lived.

As I **noted previously in analyzing communications data**, the distribution of this type of information is exponential in nature and usually follows a **log-normal distribution**. As such, a moving average is not the greatest measure of central tendency - but a decent approximation for our purposes. Still, I find the graph a little more digestible when depicted with a logarithmic y-axis, as below:



Lastly we consider the periodic nature of the emails which is noted in the initial scatterplot. We can look for patterns by making a standard heatmap with the weekday as the column and hour of day as the row, as below:



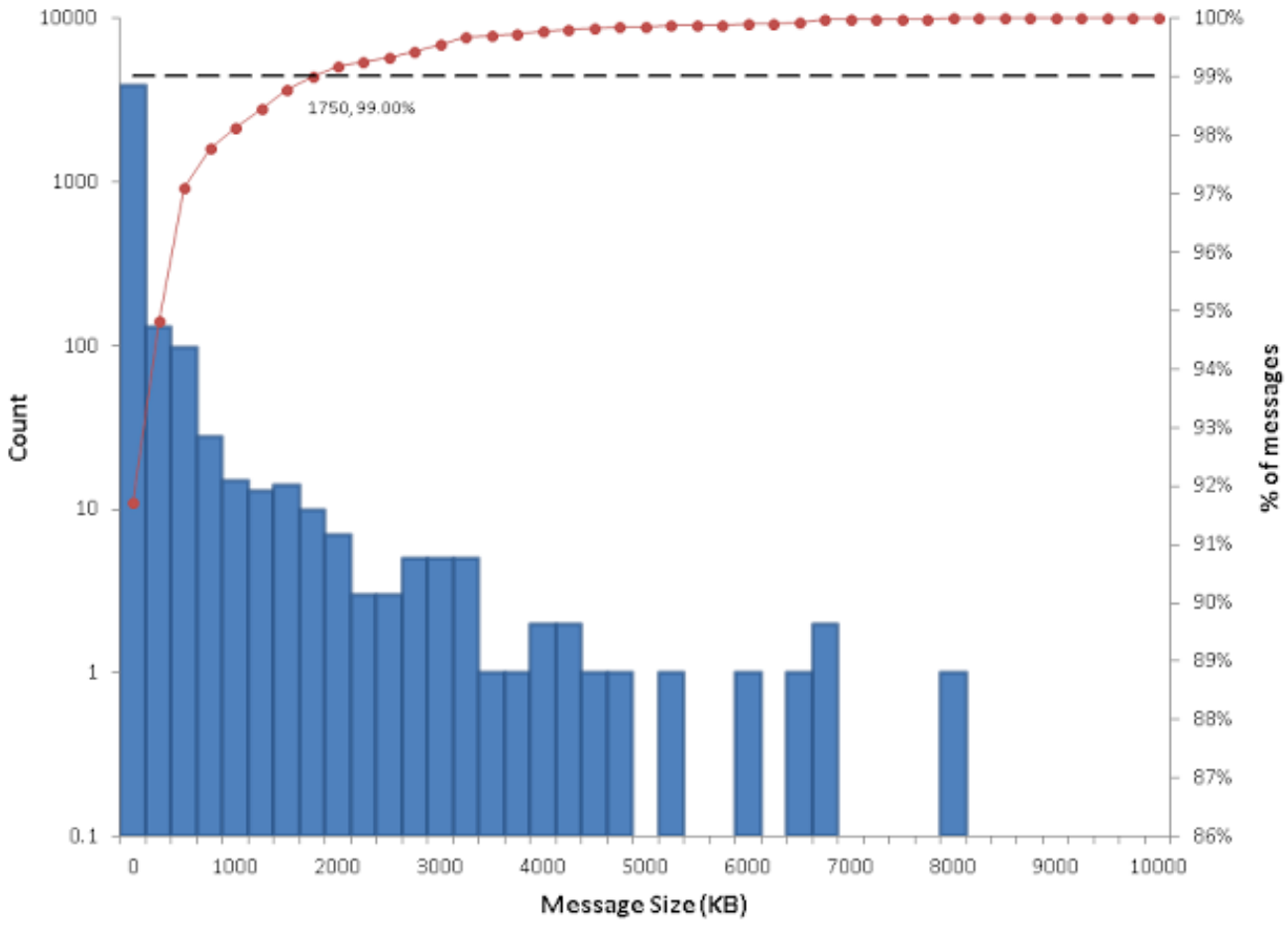
You can clearly see that the the majority of work email occurs between the hours of 9 to 5 (shocking!). However some other interesting points of note are the bulk of email in the mornings at the begiinning of the week, fall-off after 5 PM at the end of the week (Thursday & Friday) and the messages received Saturday morning. Again, I don't really receive that much email, or have spirited a lot of it away into folders as I noted at the beginning of the article (this analysis does not include things like automated emails and reports, etc.)

Email Size & Attachments

Looking at file attachments, I believe the data are more skewed than the rest, as the clean-up of large emails is a semi-regular task for the office worker (as not many have the luxury of an unlimited email inbox capacity - even executives) so I would expect that values on the high end to have largely been removed. Nevertheless it still provides a rough approximation of how email sizes are distributed and what proportion have attachments included.

First we look at the overall proportion of email left in my inbox which has attachments - of the 4,217 emails, 2914 did not have an attachment (69.1%) and 1303 did (30.9%).

Examining the size of emails (which includes the attachments) in a histogram, we see a familiar looking distribution, which here I have further expanded by making it into a **Pareto chart**. (note that the scale on the left y-axis is logarithmic):



Here we can see that of what was left in my inbox, all messages were about 8 MB in size or less, with the vast majority being 250K or less. In fact 99% of the email was less than 1750KB, and 99.9% less than 6MB.

Conclusion

This was a very quick analysis of what was in my inbox, however we saw some interesting points of note, some of which confirm what one would expect - in particular:

- vast majority of email is received between the hours of 9-5 Monday to Friday
- majority of email I received was between the two managers & colleagues I work closest with
- approximately 3 out of 10 emails I received had attachments

- the distribution of email sizes is logarithmic in nature

If I wanted to take this analysis further, we could also look at the trending by contact and also do some content analysis (the latter not being done here for obvious reasons, of course).

This was an interesting exercise because it made me mindful again of what everyday analytics is all about - analyzing rich data sets we are producing all the time, but of which we are not always aware.

References and Resources

Inbox Zero
http://inboxzero.com/

Mailbox
http://www.mailboxapp.com/

Immersion
https://immersion.media.mit.edu/

Data Mining Email to Discover Organizational Networks and Emergent Communities in Work Flows
http://www.orgnet.com/email.html

at **5:19 PM**

Labels: **bar charts**, **dataviz**, **email**, **self-tracking**

No comments:

Post a Comment

Newer Post Home Older Post

Subscribe to: Post Comments (Atom)

Previous

- **2016** (3)
- **2015** (8)
- **2014** (17)
- **2013** (16)
  - **December** (1)
    What's in My Inbox? Data Analysis of Outlook
  - **November** (2)
  - **October** (1)
  - **September** (1)
  - **August** (1)
  - **June** (1)
  - **May** (2)
  - **April** (2)
  - **March** (1)
  - **February** (2)
  - **January** (2)
- **2012** (26)

About Me