BT3103: Application Systems Development for Business Analytics
Semester 2 AY 19/20

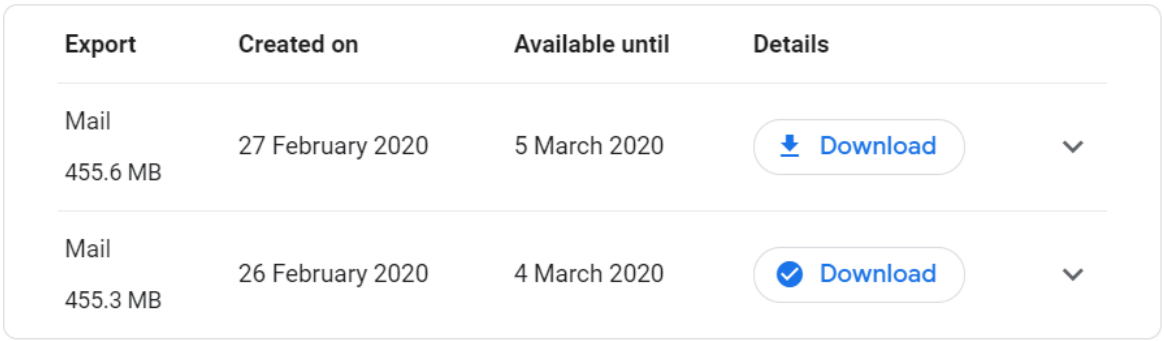
Mid Sem Assignment 1: Email Analysis
Assignment Report

Lee Jun Hui, Sean (A0189893B)

Contents

1. Data Retrieval

The mailbox data has been retrieved by downloading all the information in my personal email account using Google takeout.



Export	Created on	Available until	Details
Mail 455.6 MB	27 February 2020	5 March 2020	Download
Mail 455.3 MB	26 February 2020	4 March 2020	Download

This action exports all the information into a. mbox extension. Afterwards, I parsed the .mbox file using a csv writer in python. I extracted out the To, From, Date, Gmail Labels and the message subject using the above method.

```
writer = csv.writer(open("Extracted_mbox.csv", "w", encoding = 'utf-8')) #w = write binary

for message in mailbox.mbox('new_mail.mbox'):
    writer.writerow([message['To'], message['From'],message['Date'], message['X-Gmail-Labels'], message['subject']]) #
```

2. Data Cleaning & Pre-processing

After parsing the .mbox file into a more workable format (csv), I imported it as a pandas dataframe and started cleaning the dataset first.

Firstly, I dropped all the rows with any null values in them as I noticed that the csv writer has resulted in empty rows on alternate rows.

```
In [333]: df2 = df1.dropna()
```

Next, I reset the index of the dataframe as the dataframe index consist of only values in alternate rows

```
In [334]: new_index = []
          count = 0
          for i in df2.index:
              new_index.append(count)
              count +=1

          df2.index = new_index
```

Next, I tried to understand the dimensions of the dataframe and the data types of the columns that I am working with

```
df2.shape
```

```
(6065, 5)
```

Further understanding of the datatypes of the columns and to check whether there are any null values

```
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 6065 entries, 0 to 6064
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   To            6065 non-null   object
 1   From          6065 non-null   object
 2   Date          6065 non-null   object
 3   Gmail Labels  6065 non-null   object
 4   Subject       6065 non-null   object
dtypes: object(5)
```

There is a need to clean the To, From, and Date columns as they are currently in very inconsistent formats. This will affect the accuracy of the aggregation functions later if we do not clean them into one format.

```
df2.Date.unique()[:20]
```

```
array(['Thu, 20 Feb 2020 15:27:28 -0800',
       'Fri, 21 Feb 2020 02:07:58 -0600',
       'Tue, 4 Feb 2020 05:05:01 +0000 (UTC)',
       'Wed, 19 Feb 2020 17:33:03 +0000 (UTC)',
       'Wed, 12 Feb 2020 15:29:46 +0000 (UTC)',
       'Sun, 26 Jan 2020 20:38:45 +0000 (UTC)',
       'Sat, 15 Feb 2020 02:03:40 +0000',
       'Thu, 16 Jan 2020 16:47:16 +0000',
       'Fri, 14 Feb 2020 07:25:11 +0000',
       'Tue, 25 Feb 2020 16:00:59 +0800 (+08)',
       '1 Jan 2020 12:04:05 +0800',
       'Mon, 03 Feb 2020 00:00:00 +0000 (UTC)',
       'Tue, 11 Feb 2020 11:48:19 +0000',
       'Mon, 17 Feb 2020 13:38:10 +0530',
       'Sat, 22 Feb 2020 01:43:03 -0800',
       'Sun, 22 Dec 2019 06:30:55 +0000 (UTC)',
       'Sat, 15 Feb 2020 23:24:56 -0800',
       'Sat, 15 Feb 2020 06:52:55 -0800',
       'Sat, 18 Jan 2020 03:10:59 +0000 (UTC)',
       'Sat, 18 Jan 2020 01:46:00 +0000 (UTC)'], dtype=object)
```

```
df2.To.unique()[:20]
```

```
array(['<seansljh@gmail.com>', '"Lee Jun Hui, Sean" <seansljh@gmail.com>',
       'Sean Lee <seansljh@gmail.com>', 'seansljh@gmail.com',
       '"Sean Lee" <Seansljh@gmail.com>',
       '"seansljh@gmail.com" <seansljh@gmail.com>',
       'Sean Lee <seansljh@gmail.com>, Sean Lee <seansljh@gmail.com>',
       'SEANSLJH@GMAIL.COM', 'NTU-PACE-ENEWSLETTER@MLIST.NTU.EDU.SG',
       '<SEANSLJH@GMAIL.COM>',
       '"seansljh@gmail.com" <seansljh@gmail.com>,\r\n        "E0325477@U.NUS.E
DU"\r\n\t<E0325477@U.NUS.EDU>',
       'seansljh <seansljh@gmail.com>',
       '=?utf-8?Q??= <seansljh@gmail.com>',
       '"Lee Jun Hui Sean" <seansljh@gmail.com>',
       'Undisclosed recipients:;',
       '"Teh,Joanna,SINGAPORE,Human Resources" <Joanna.Teh@sg.nestle.com>',
       '"Por,Hui Fang,SINGAPORE,Human Resources" <HuiFang.Por@sg.nestle.com>',
       '"seansljh@gmail.com" <seansljh@gmail.com>,\r\n        "E0325477@U.NUS.E
DU"\r\n\t<E0325477@U.NUS.EDU>,\r\n        "seansljh@gmail.com" <seansljh@gmail.
com>',
       '"seansljh@gmail.com" <seansljh@gmail.com>,\r\n        "seansljh@gmail.c
om"\r\n\t<seansljh@gmail.com>,\r\n        "E0325477@U.NUS.EDU" <E0325477@U.NUS.
EDU>',
       '"\'Lee Jun Hui, Sean\'" <seansljh@gmail.com>'], dtype=object)
```

```
df2.From.unique()[:20]
```

```
array(['=?UTF-8?Q?Persona=20Nutrition?= <personanutrition@news.personanutritio
n.com>',
       '"gradsingapore" <mail@gradsingapore.com>',
       'Nicholas Teh via LinkedIn <invitations@linkedin.com>',
       'LinkedIn <jobs-listings@linkedin.com>',
       'LinkedIn Job Alerts <jobalerts-noreply@linkedin.com>',
       'SmileTutor <contactus@smiletutor.sg>',
       '=?utf-8?Q?Slidebean=20Templates=20Blog?= <templates@slidebean.co>',
       'Quora Digest <digest-noreply@quora.com>',
       '"BrightSparks" <brightsparks@jobscentral.com.sg>',
       '"KOI" <singapore@koicafe.com>',
       '"Medium Daily Digest" <noreply@medium.com>',
       'Supercell <noreply@id.supercell.com>',
       'Standard Chartered Singapore <sc.singapore@sc.com>',
       'Google <noreply-utos@google.com>',
       '"Airbnb" <invitation@airbnb.com>',
       'Google Docs <comments-noreply@docs.google.com>',
       '"Tobias Yap (Google Docs)" <comments-noreply@docs.google.com>',
       'Anamika Suresh via LinkedIn <messaging-digest-noreply@linkedin.com>',
       'DataCamp <team@datacamp.com>',
       "Prince George's Park Residences <askpgpr@nus.edu.sg>"],
      dtype=object)
```

Cleaning Date Column

1. I wanted to convert the string representations to datetime objects as it will be easier to reference and filter later (as we can simply use datetime.year, etc to reference what we need)
2. I also wanted to convert all the different string representations of datetime to datetime objects to GMT +8 format so we can make more valid comparisons later in addressing the question requirements

Cleaning To & From Columns

1. Inspecting the dataframe, I can also see that the to and from formats of the email are vastly different (e.g. my email, seansljh@gmail.com can be in various formats such as <seansljh@gmail.com> , SEANSLJH@GMAIL.COM, etc. Hence, there is a need to process the strings to show the emails all in lower cases only
2. For 'To', I noticed that sometimes an email may not just be sent to me only. Hence, I created a new feature known as 'toCount' to show the number of people the email has been addressed to

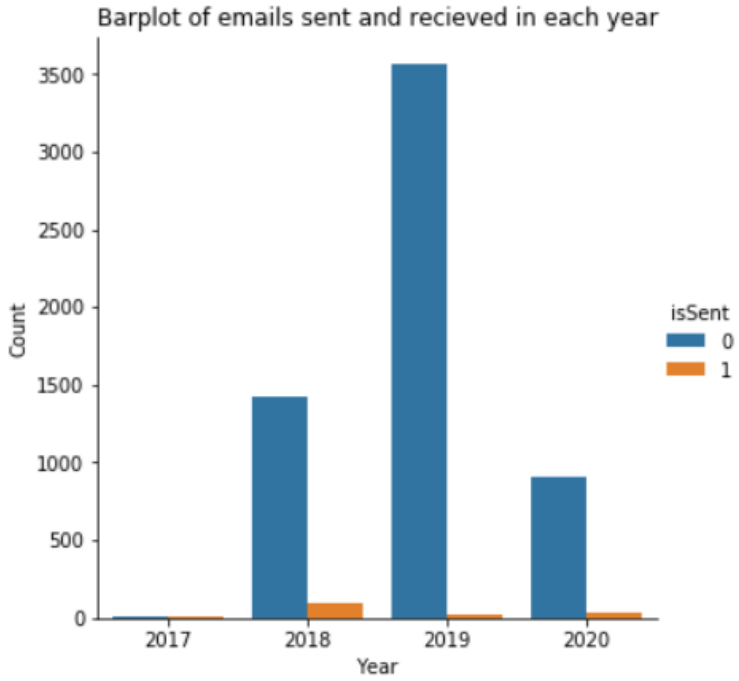
3. Insights to Question Requirements

Question 1

- Plot a barchart of the number of emails sent and received in each year
- Plot a breakdown of the number of emails received between June 2019 - August 2019 by From field(You can use fictitious names for the From sender to protect data privacy)

For the first barchart

1. I first filtered out the records that were from the past 3 years
2. Aggregated the data by the 'year' portion of the datetime object and applied a count() function to get the year and count of mail in a dataframe
3. Plotted it using sns.catplot(), with kind = "bar". I also set "hue" to 'isSent' in order to differentiate the bars by whether it is the count for emails I sent or received

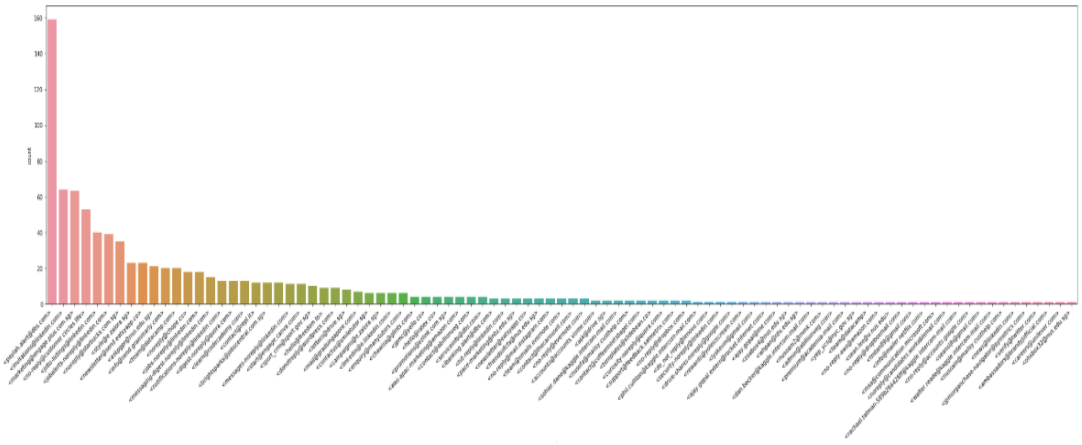


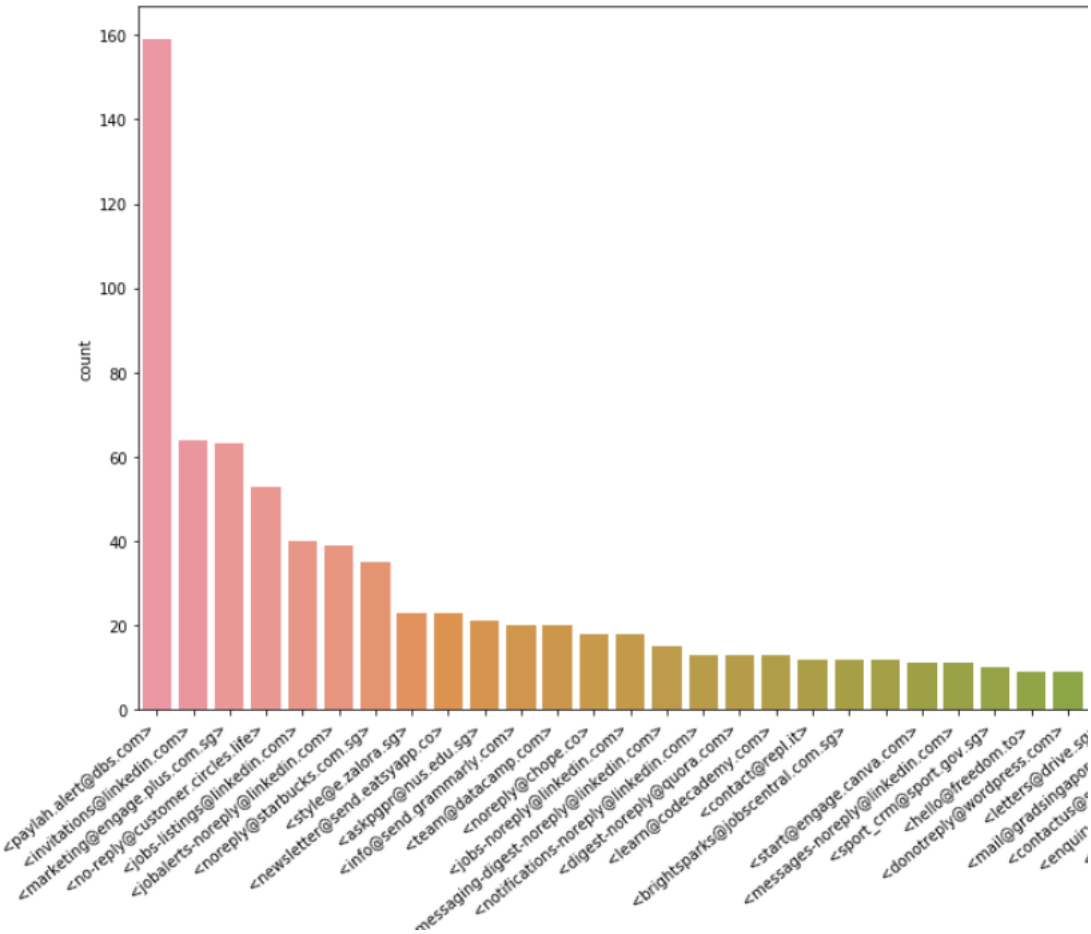
Insights

1. The number of emails I have sent out is much lesser than the emails that I have received (Lesser than 10% of the total number of emails sent and received in all the years)
2. There is a general increasing trend in the number of mails that are being sent and received in my email over the years, which is true as I am currently trying to port all the emails of my various accounts to this new email that I created in 2017.

For the second part of the question

1. I filtered out all the records with date between June 2019 & August 2019 AND also isSent = 0
2. Did a groupby the 'From' column
3. Used a countplot to show the counts of the different categorical groups





(Zoomed in view)

Insights

1. Most of my emails that I received during this period were from LinkedIn and other subscription services that I have opted in (E.g. Zalora)
2. The paylah alert counts were also the highest, which is valid as during this portion of time I was spending a lot of time in school due to Co-Curricular Activities
3. Interestingly, one of the top email addresses I received emails from also includes askpgpr@nus.edu.sg. This could be because I assumed my position as a Resident Assistant during this period (From June 2019 onwards).

Question 2

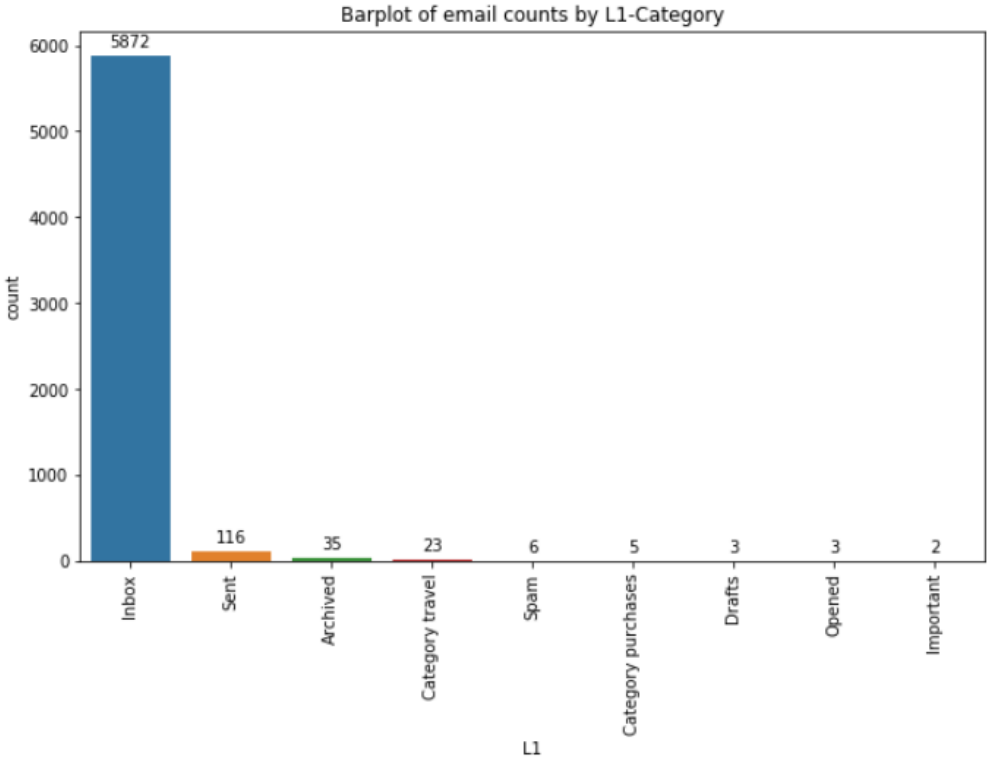
- Categorize your emails based on labels and plot them.
  - Google has a few categories . Use these.

Categorisation

1. For this question, I first went to look at the various unique gmail labels that my mails were tagged as by Google
2. Afterward, recognising that they tend to be broken down in terms of various levels, I decided to split the Gmail-Labels into various categories and worked from there

> Can do create new columns which show the L1 - L3 Email Category, which refers to the order ir which Gmail has categorised the emails into;
>
> Eg. for 'Inbox, Opened, Category Promotions, Starred';
> L1 = Inbox, L2 = Opened, L3 = Category Promotions

Barplot based on L1 Category

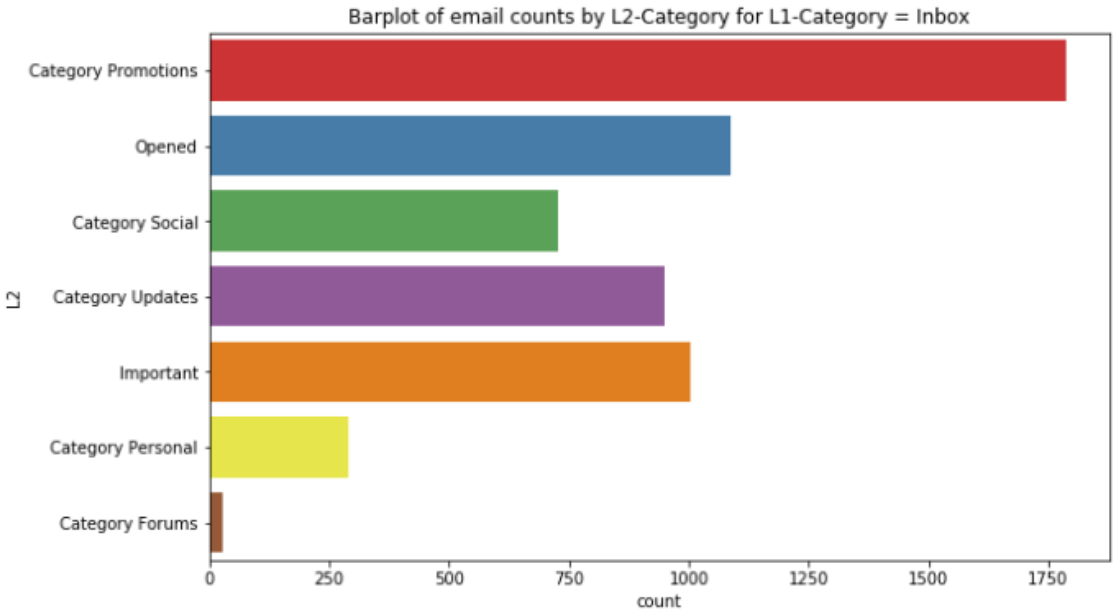


Insights

1. Bulk of my mails go into my Inbox, i.e. I receive way more emails than I send out emails
2. Rest of the emails in other L1 Categories occur in very low amounts (Especially Drafts, Opened, etc)!

Looking at this skewed distribution, I went to drill down on the L2 Categories of those mails tagged with L1 Category as 'Inbox'

L2 Category Barplot for those with L1 Category as ‘Inbox’



Insights

1. Bulk of the mails I receive are categorised as promotions, social media, and updates
2. A small amount refers to forums – which is true as I began reading Quora recently

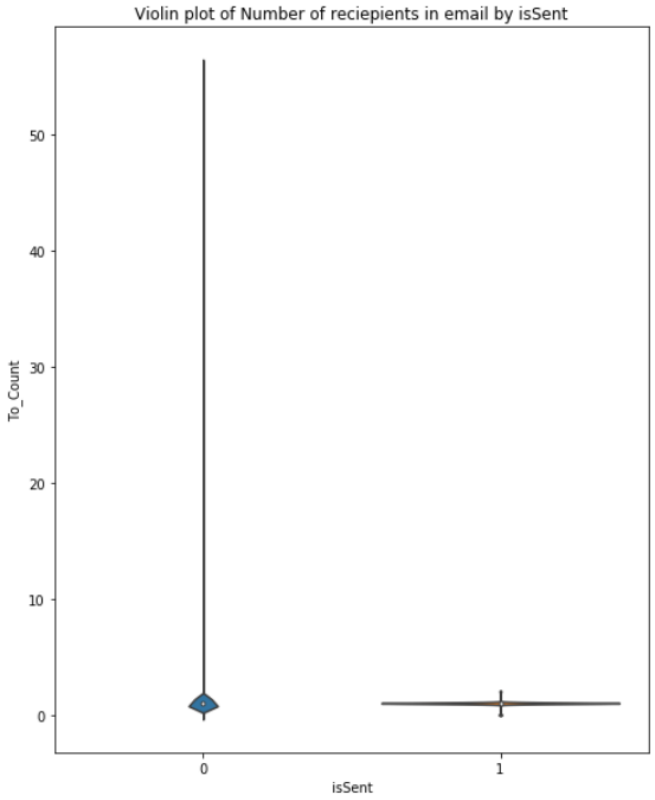
Question 3

Some ways that I would like to analyse the data would include the following

a) Whether the To_Count (number of people an email is addressed to) is affected by whether I was the one sending the email
b) The most popular words that appear in the subject title of my emails
c) Time Series Analysis of my emails (Whether there is a seasonal trend, etc)

Analysis a)

1. I did a violin plot as I would like to analyse the distribution of the continuous variable To_Count



Insights

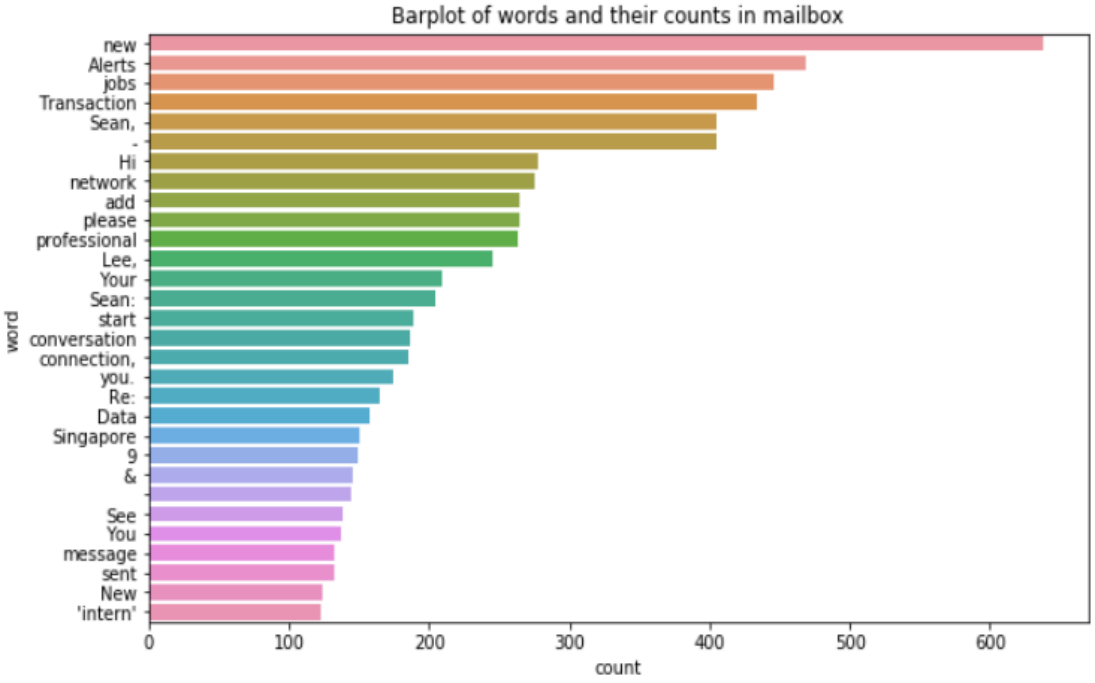
1. From the above, we can see that bulk of the messages that I have sent out tend to have 1 recepient only, while for the messages I recieved, even though bulk of them are sent explicitly to me only, there is a larger portion that is sent to multiple people too (seen by the long tail of the plot for isSent = 0)

Analysis b)

1. For this portion, I took the strings in the 'Subject' column and splitted them by spaces. While not a perfect strategy, I thought this would be one of the best ways to go about with the analysis I wanted to perform
2. I also used the nltk package (A natural language processing package commonly used for text processing purposes) to retrieve a list of common stop words in English. This

is to allow the analysis to be more insightful as stop words tend to be used more commonly in all email subjects.

3. I took out the top 30 words only as I did not want to overwhelm the plot

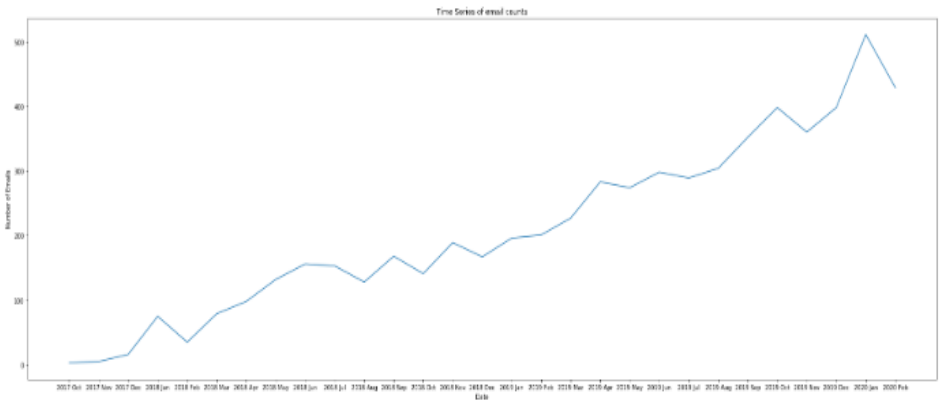


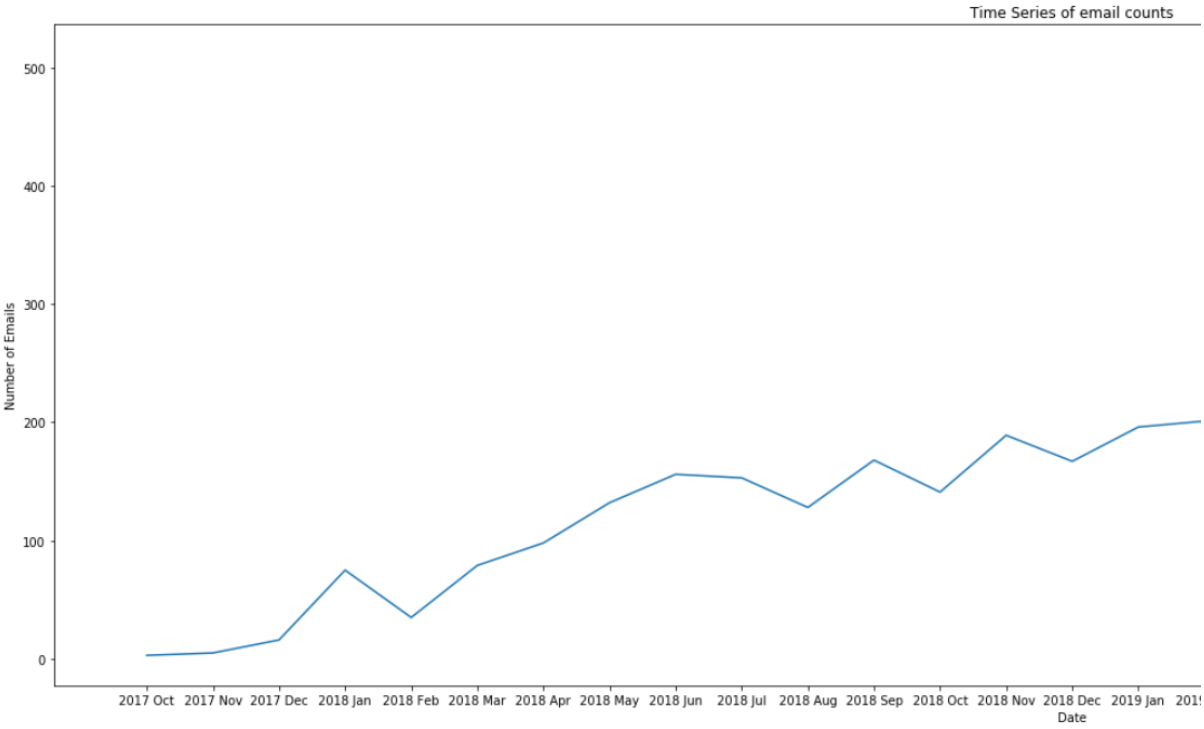
Insights

1. While there is the presence of numerous words that are normal connectors used (e.g. to, at, etc); there is also the presence of words that are used normally by LinkedIn when a new connection is added to my network (network, add, professional, start, connection, conversation, etc)
2. Supports the analysis in Question 1b that I am receiving bulk of my emails from LinkedIn

Analysis c)

1. While an analysis was done on the yearly emails received and sent in Question 1a), I wanted to look at the time series data of the number of emails received and sent across all the months to look at patterns for any seasonality





(Zoomed in version)

Insights

1. There is no apparent seasonal trend in the number of emails that I received, only a gradual steady increase in the number of monthly emails received

4. Conclusion

In conclusion, the analysis conducted in this project has made me more aware of the type of emails that I am receiving and sending out. Most importantly, I believe that it reflects the current state of my life well as it does show that I am using LinkedIn quite aggressively as a social media platform for networking and internship searching purposes.

As further improvements, I believe it would be interesting to compare the data from this email that I have set up for professional purposes to the email account that I set up when I was younger for gaming and social purposes.