Emails, emails, emails: spam analysis of a Gmail inbox

By Dennis Francis, Dec 8 2023

Just how much email is spam?

Traditional spam filters don’t entirely work. Despite unsubscribing, blocking, and other tactics, I still find myself with an inbox full of messages that are still junk.

The purpose of this is to answer the question, how much of email is spam?

Dataset: 650 emails from my personal Gmail account accessed via OAuth 2.0 Gmail API (you must create a GCP Google Cloud Platform to access the API)

The data spans from June 22 2023 to December 9th 2023

Why analyze spam?

We’re all familiar with spam. But why is this topic interesting?

Getting acquainted with the Gmail API requires you to become familiar with the Google Cloud Platform GCP. GCP has many products pertinent to Big Data, so it’s well worth getting to know.

Secondly,

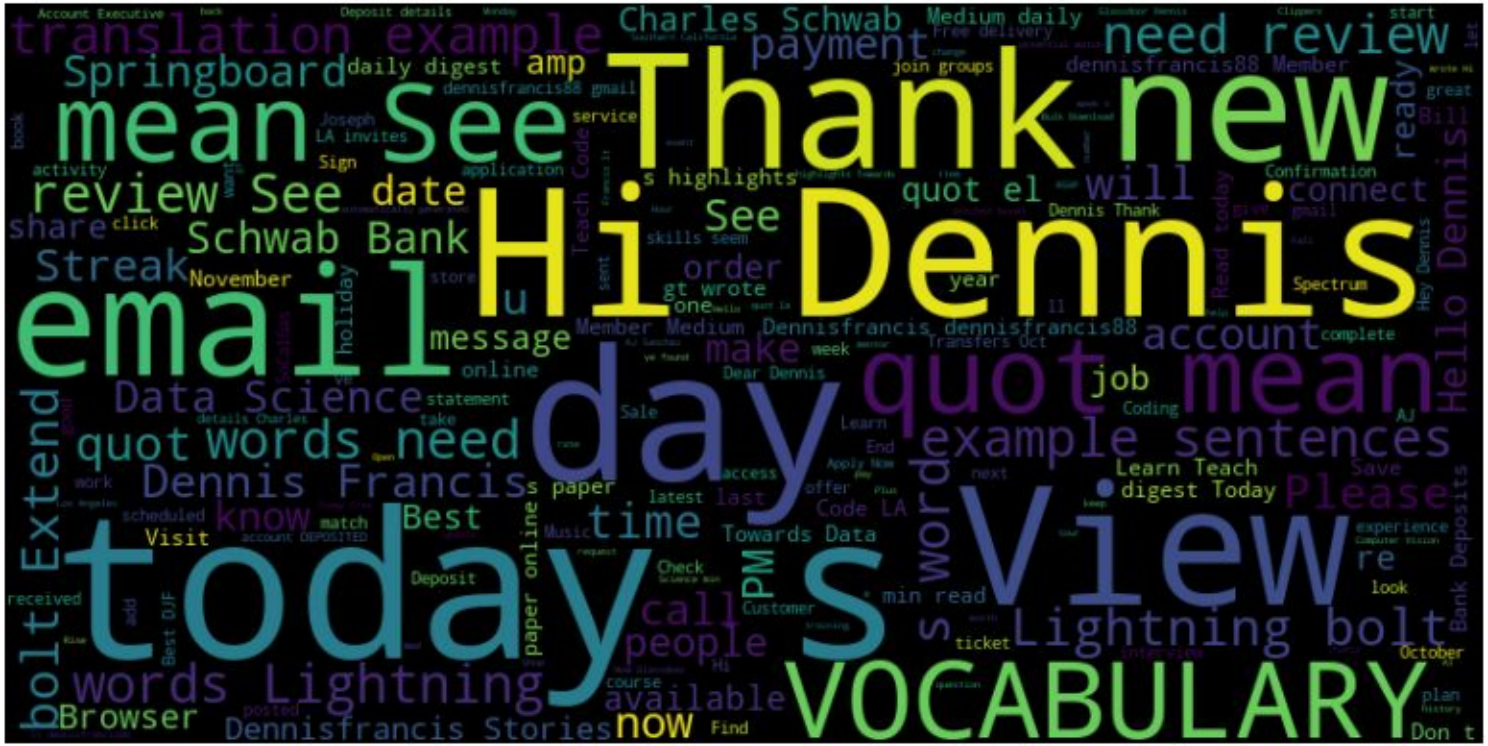
Questions we seek to answer:

What email senders send the most spam

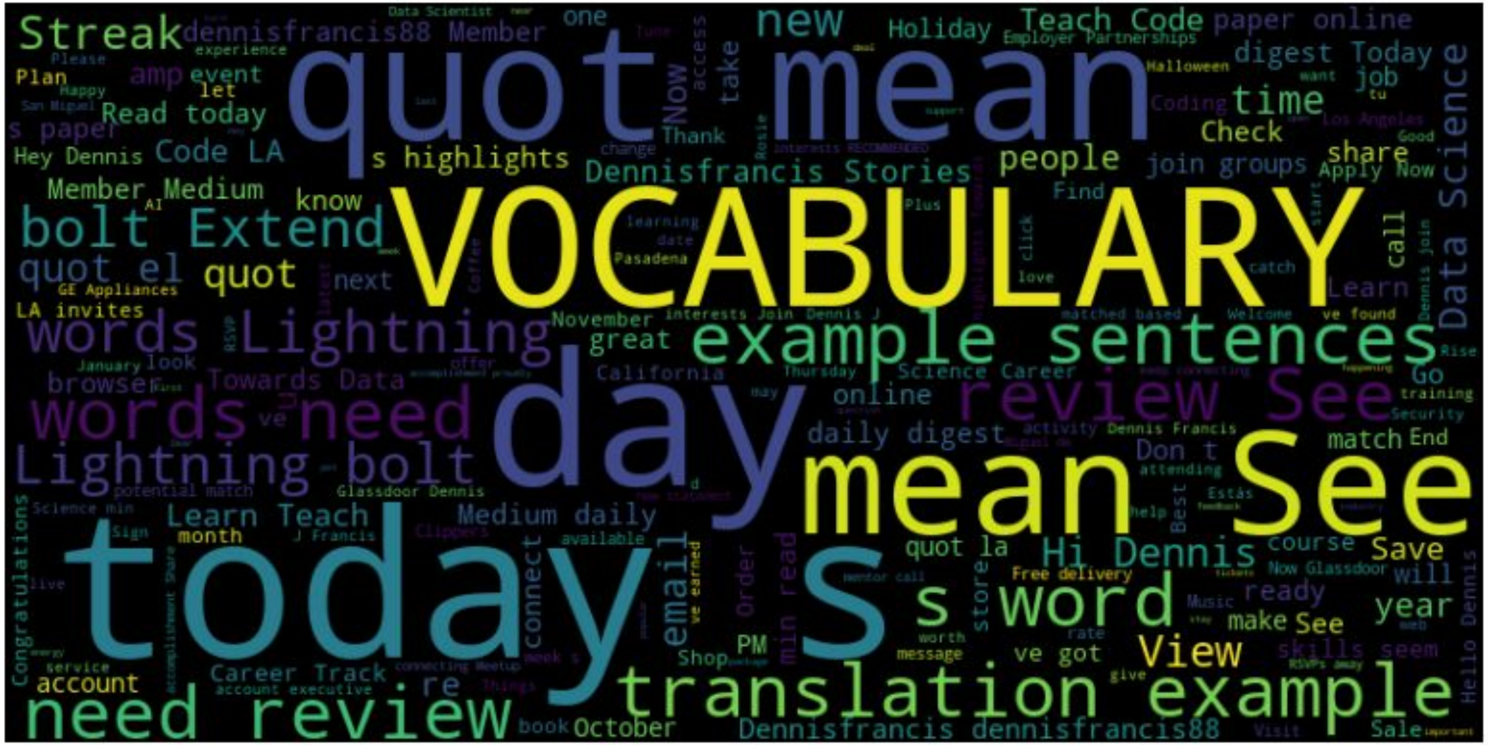
Raw Data

	mssg_id	date	sender	subject	snippet
0	18c4a36027b33bbf	8 Dec 2023 16:15:00 +0000	The Home Depot Consumer Credit Card <homedepot...	Same great benefits. One Account.	Add an Authorized User to Your Account The Hom...
1	18c4a34f38b046cc	Fri, 8 Dec 2023 16:01:58 +0000 (GMT)	SoCalGas <webmaster@socalgas.messages2.com>	SoCalGas Customers: It's Time for Holiday Appl...	Wrap up our energy-efficient appliances from t...
2	18c49ffb07bb84fc	Fri, 08 Dec 2023 15:15:35 +0000 (UTC)	Medium Membership <members@medium.com>	The Edition: Do you innovate or do you problem...	It's that time of year when people tend to...
3	18c49f15bf2e2fc2	Fri, 08 Dec 2023 15:00:00 +0000 (UTC)	Medium Daily Digest <noreply@medium.com>	I Was Trapped In A Room Full Of Business Milli...	Dennisfrancis Stories for Dennisfrancis @denni...
4	18c49e0ec6272460	Fri, 08 Dec 2023 08:41:58 -0600	Quest <promo@e.questdiagnostics.com>	Last chance to save 15% on thyroid tests	Don't miss out on important insights ...
5	18c49da20065e776	Fri, 8 Dec 2023 14:34:40 +0000	Quincy Larson <quincy@freecodecamp.org>	Learn DevOps + Machine Learning with Python [F...	Here are this week's five freeCodeCamp res...
6	18c49bdd1d6988ed	Fri, 08 Dec 2023 14:03:44 +0000	"Chewy.com" <chewy@paws.chewy.com>	Prepare to pounce on up to 50% off	It's a great time to restock their favorit...
7	18c497fa061faa42	Fri, 08 Dec 2023 12:54:52 +0000	"Juan at SpanishDictionary.com" <noreply@spani...	"la vela" — Learn New Words Today with Spanish...	What does "la vela" mean? See the tr...
8	18c49699702abc07	Fri, 08 Dec 2023 12:31:43 +0000	ThriftBooks <hello@shop.thriftbooks.com>	Get the Most Value Out of Your ThriftBooks Sho...	Holiday Hacks! ...
9	18c4922cf6c91369	Fri, 08 Dec 2023 05:10:30 -0600	Los Angeles Times <enotify@email.latimes.com>	Your eNewspaper arrived	Read today's paper online ...
10	18c481e9ce8bb185	Fri, 08 Dec 2023 06:30:14 +0000 (UTC)	Your South Pasadena Central neighbors <no-repl...	Top post: Hi my neighbors,	I am just making sure You are doing great. It ...
11	18c4777436b6c720	Thu, 07 Dec 2023 21:27:14 -0600	"Charles Schwab & Co., Inc." <donotreply@email...	Three financial moves to consider before Decem...	Your top year-end financial tips view on web E...
12	18c46ea506a71f7d	Thu, 07 Dec 2023 18:50:43 -0600	Experian Alerts <support@s.usa.experian.com>	Dennis, you have 1 new credit alert to review	Find out how these changes may affect your cre...
13	18c46bd0612e3a31	Fri, 08 Dec 2023 00:04:01 +0000	Springboard Career Services Team <careerservic...	How was your recent coaching call?	Hi Dennis, Thanks for attending your career co...
14	18c46aaf17b33689	Thu, 07 Dec 2023 23:44:16 +0000	LA Clippers <fanassist@clippers.com>	Season of Gifting, Clips/Nuggets Highlights & ...	Full Court Press The new season is less than a...

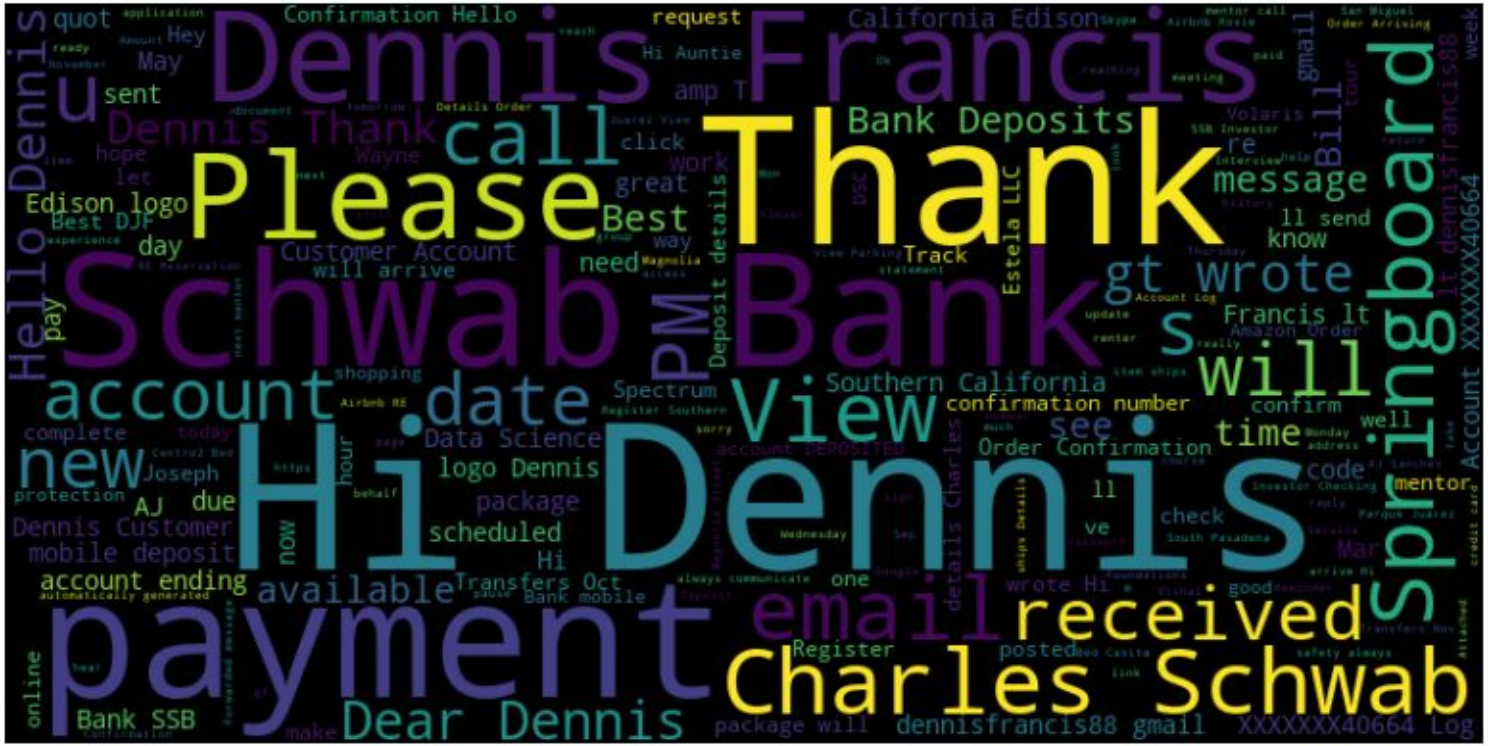
Email Word Cloud



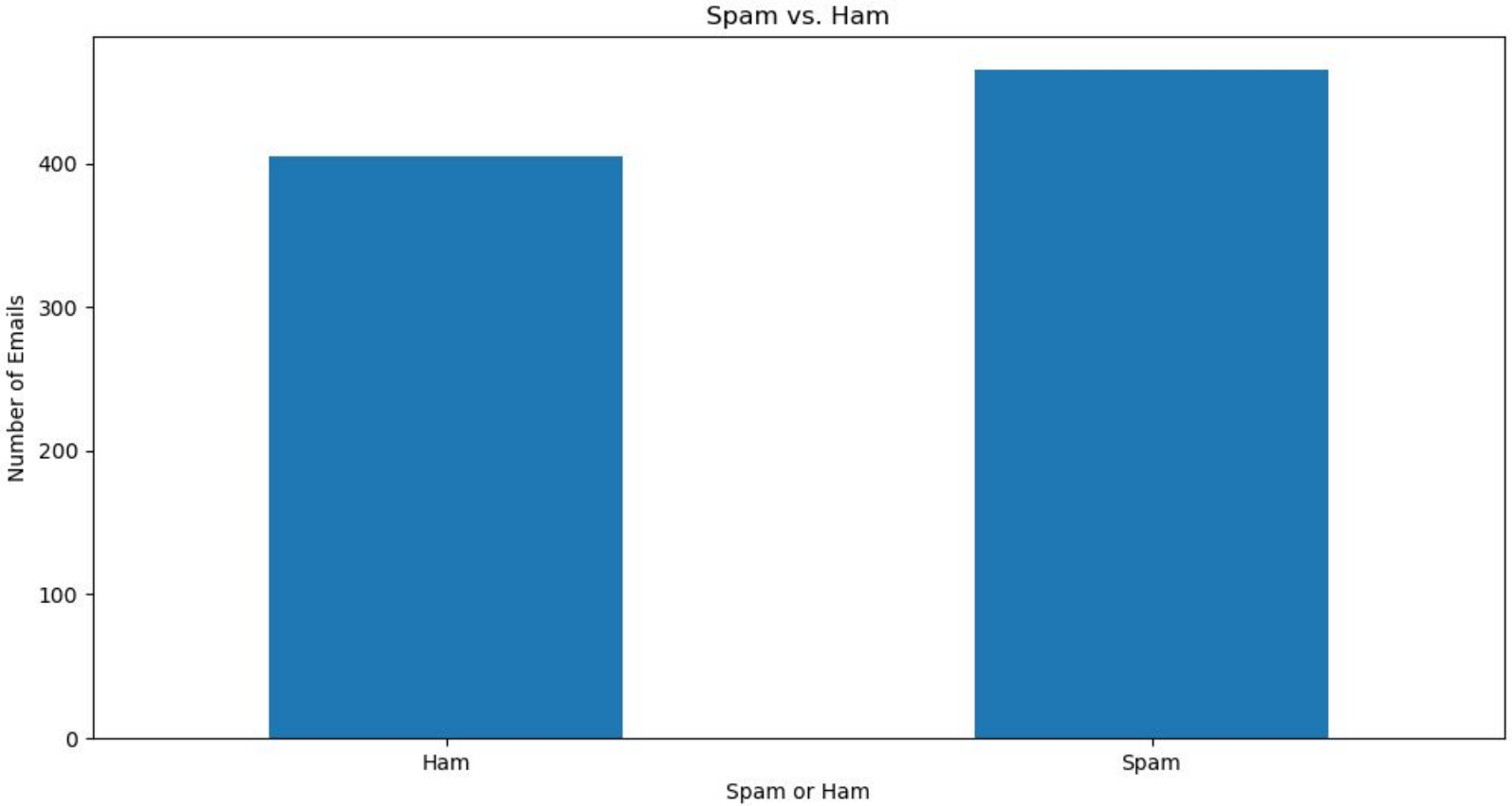
Spam Word Cloud



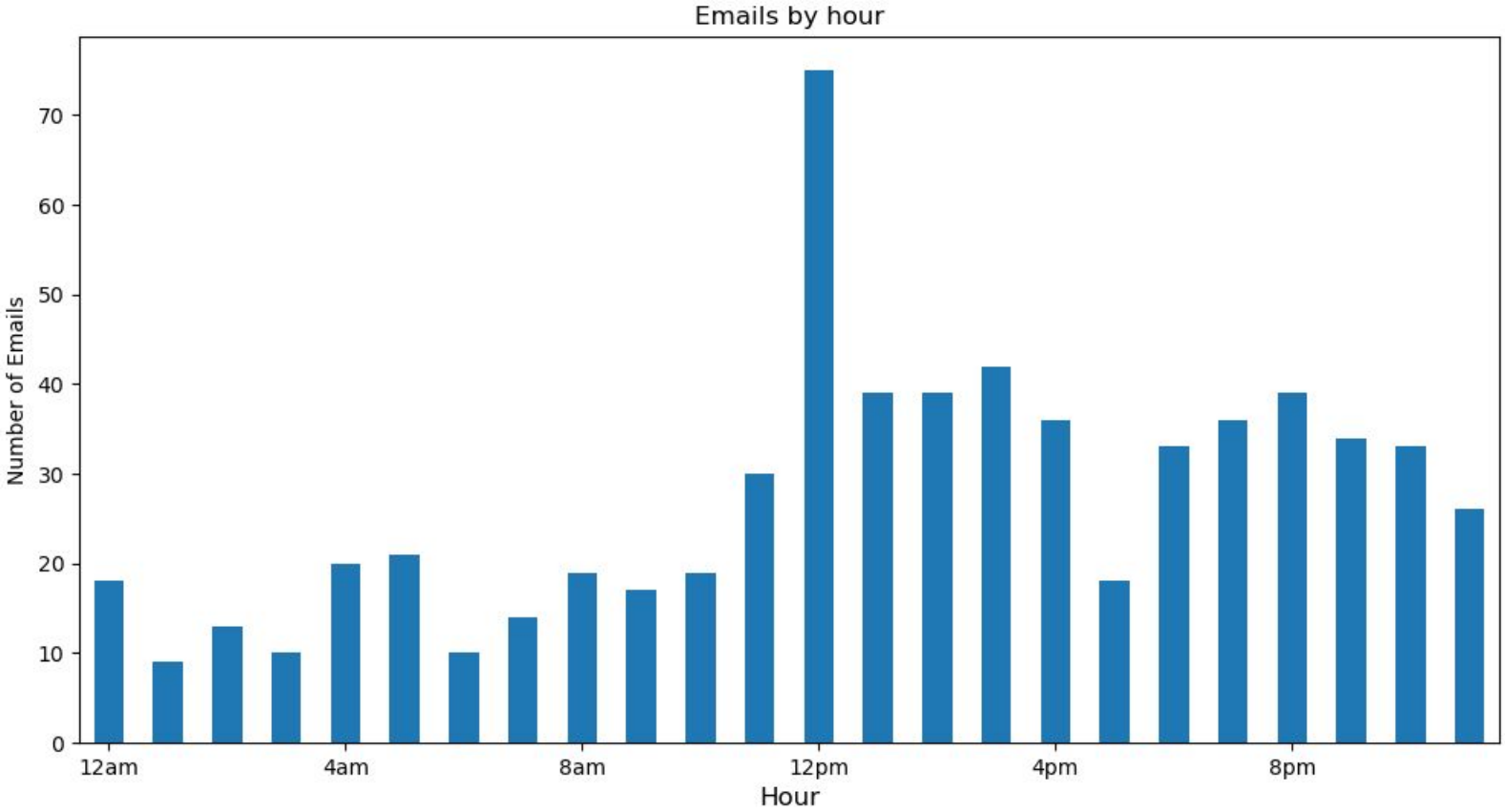
Ham Word Cloud



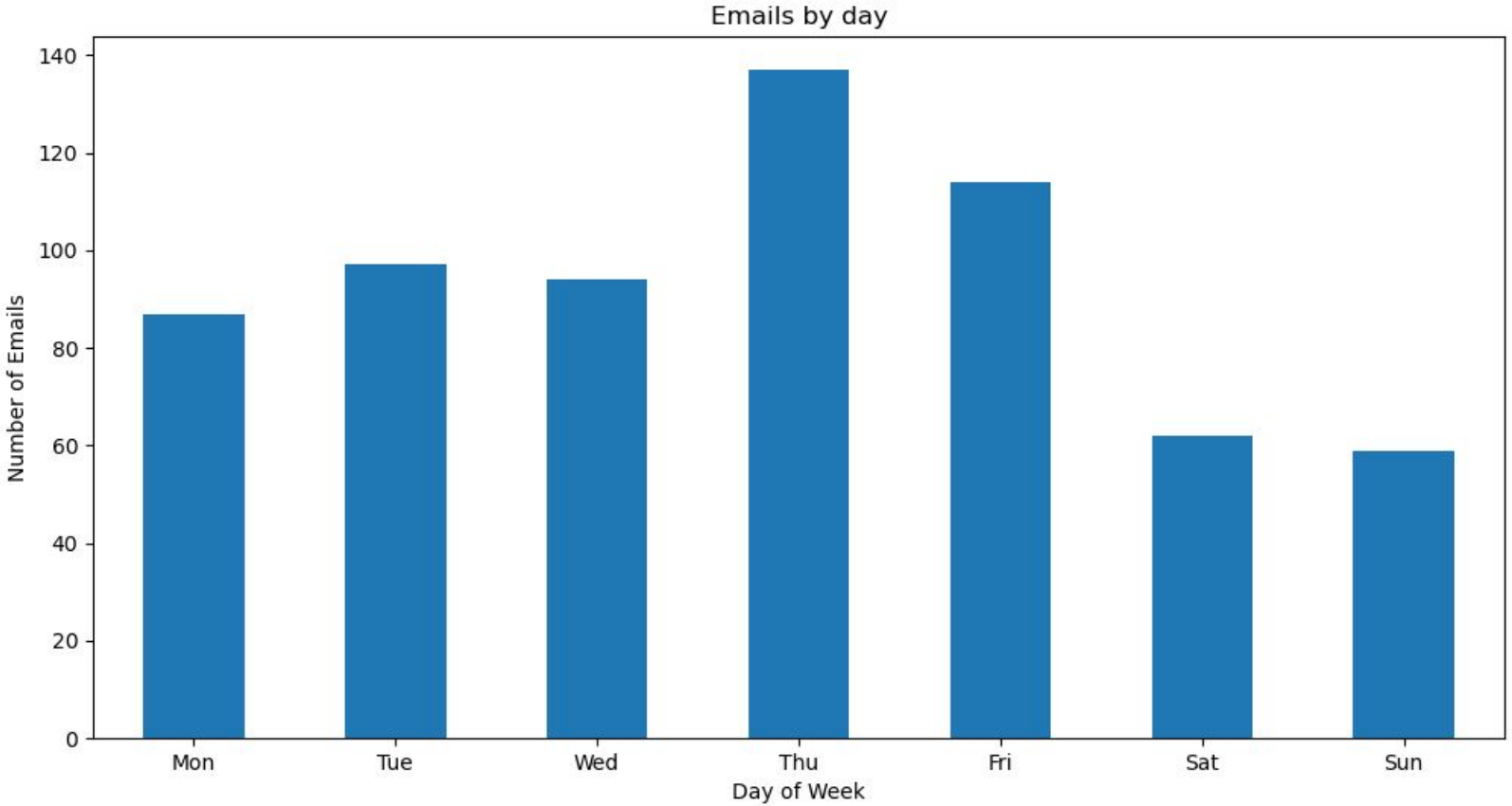
Spam vs. Ham


Spam vs. Ham
Number of Emails
0
100
200
300
400
Ham
Spam
Spam or Ham

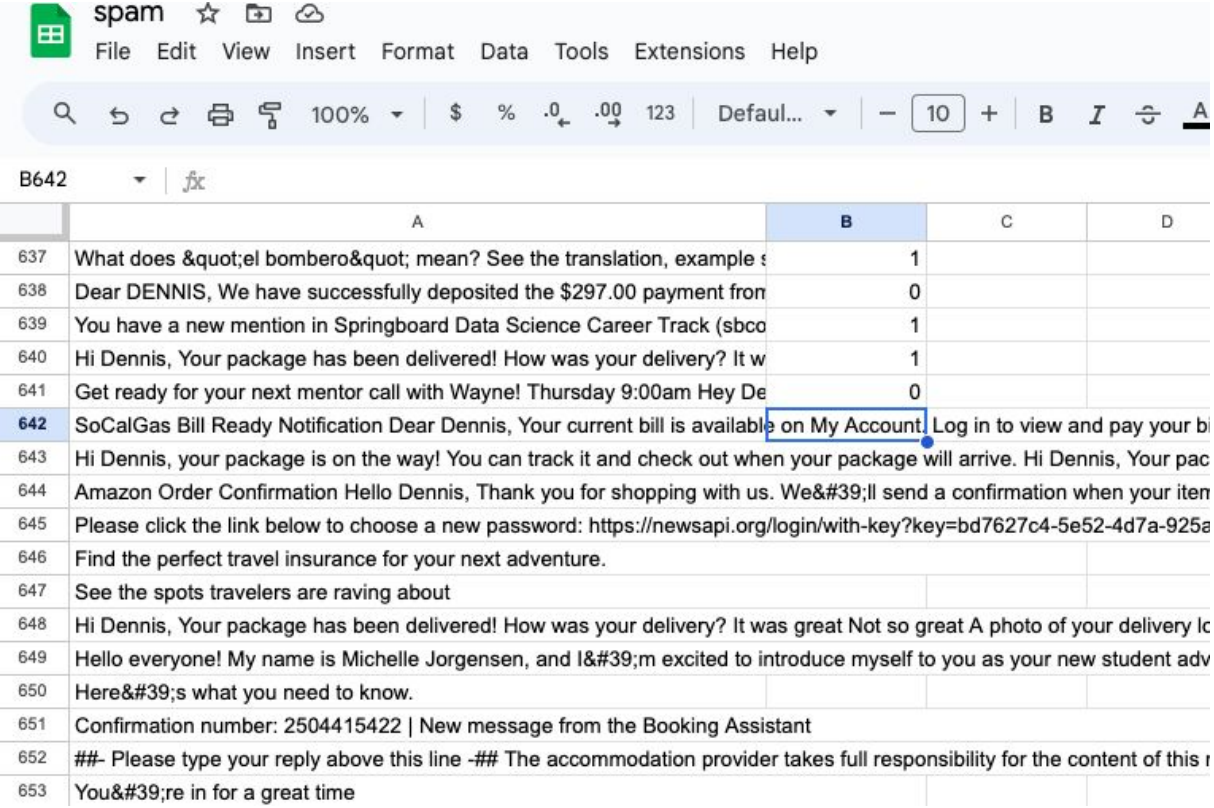
Emails by hour - spike at 12pm noon



Emails by day - Thursday and Friday most frequent



Labeling the Data By Hand:



	A	B	C	D
637	What does "el bombero" mean? See the translation, example s	1		
638	Dear DENNIS, We have successfully deposited the $297.00 payment from	0		
639	You have a new mention in Springboard Data Science Career Track (sbco	1		
640	Hi Dennis, Your package has been delivered! How was your delivery? It w	1		
641	Get ready for your next mentor call with Wayne! Thursday 9:00am Hey De	0		
642	SoCalGas Bill Ready Notification Dear Dennis, Your current bill is available on My Account. Log in to view and pay your bi			
643	Hi Dennis, your package is on the way! You can track it and check out when your package will arrive. Hi Dennis, Your pac			
644	Amazon Order Confirmation Hello Dennis, Thank you for shopping with us. We'll send a confirmation when your item			
645	Please click the link below to choose a new password: https://newsapi.org/login/with-key?key=bd7627c4-5e52-4d7a-925a			
646	Find the perfect travel insurance for your next adventure.			
647	See the spots travelers are raving about			
648	Hi Dennis, Your package has been delivered! How was your delivery? It was great Not so great A photo of your delivery lo			
649	Hello everyone! My name is Michelle Jorgensen, and I'm excited to introduce myself to you as your new student adv			
650	Here's what you need to know.			
651	Confirmation number: 2504415422 | New message from the Booking Assistant			
652	##- Please type your reply above this line -## The accommodation provider takes full responsibility for the content of this r			
653	You're in for a great time			

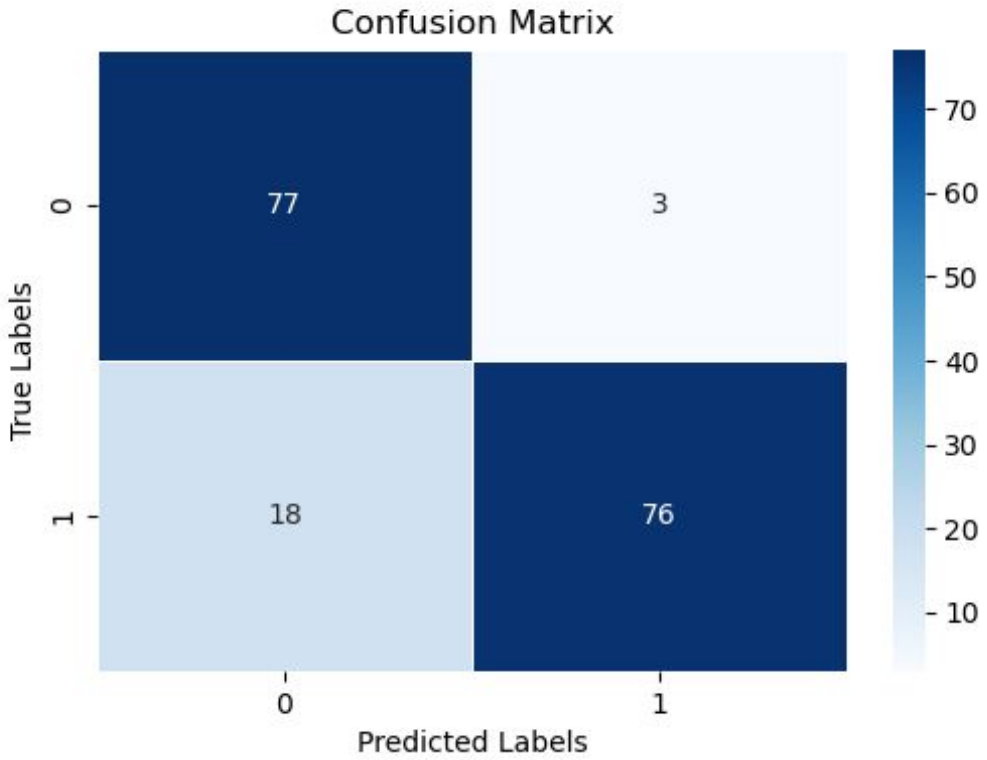
Labor-intensive, but

It needed to be done

Naive Bayes' Performance: 88% Accuracy

Precision: .962

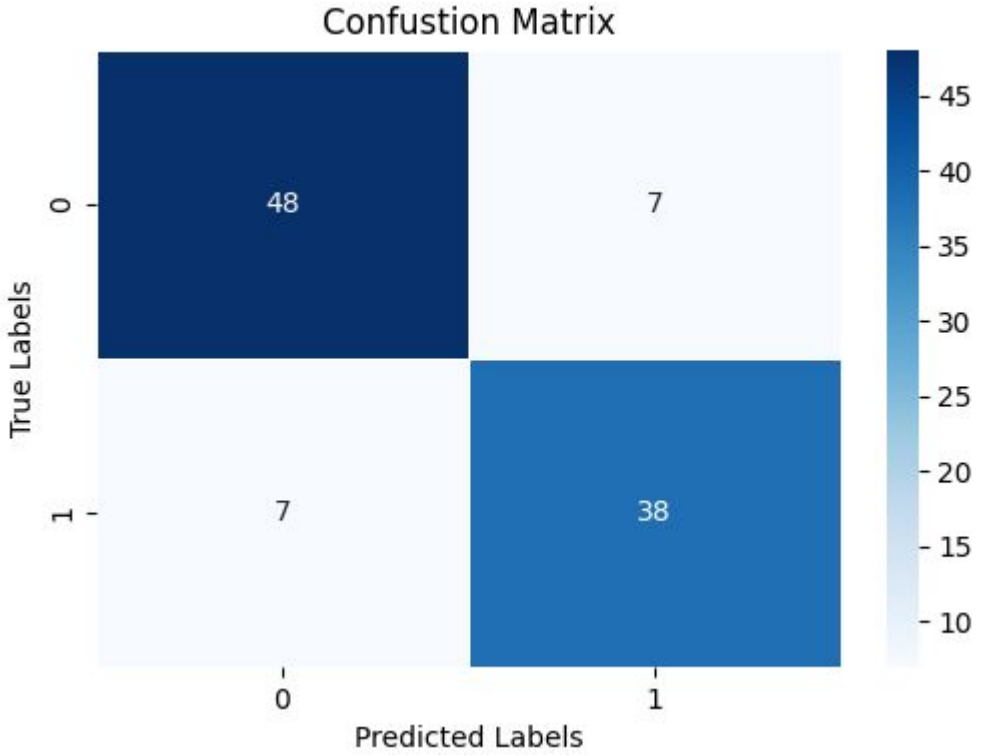
Recall: .809



DistilBERT Base Uncased Model Performance

Precision: .844

Recall: .844



Reasons why Deep Learning Model lost:

Only 650 datapoints, which I subjectively labeled by myself.

Naive Bayes' has historically performed on small sparse datasets and especially spam detection

Notebooks available at: