Write-up Question 0:

Collaborators:

- N/A

Links:

- https://linuxconfig.org/how-to-parse-data-from-json-into-python
- https://docs.python.org/3/library/datetime.html
- https://www.geeksforgeeks.org/converting-string-yyyy-mm-dd-into-datetime-in-python/#:~:text=yyyy%2Dmm%2Ddd%20stands%20for,get%20the%20string%20to%20datetime.
- https://docs.python.org/2/library/os.html#os.listdir
- https://stackoverflow.com/questions/3207219/how-do-i-list-all-files-of-a-directory
- URL to Image PIL:
  - https://stackoverflow.com/questions/7391945/how-do-i-read-image-data-from-a-url-in-python
  - https://stackoverflow.com/questions/7391945/how-do-i-read-image-data-from-a-url-in-python
- Retrieving keys in a JSON
  - https://dev.to/bluepaperbirds/get-all-keys-and-values-from-json-object-in-python-1b2d
- MatPlotLib
  - https://matplotlib.org/stable/index.html
- File Opening
  - https://www.w3schools.com/python/python_file_open.asp
- String Replace
  - https://www.adamsmith.haus/python/answers/how-to-strip-a-specific-word-from-a-string-in-python#:~:text=replace()%20to%20strip%20a,in%20old%20to%20remove%20them.
- Heapq
  - https://stackoverflow.com/questions/47548953/find-n-largest-values-from-dictionary
  - https://docs.python.org/3/library/heapq.html
  - https://www.geeksforgeeks.org/python-n-largest-values-in-dictionary/
- Bar Plot
  - https://www.tutorialspoint.com/matplotlib/matplotlib_bar_plot.htm
- Line Graph
  - https://datatofish.com/line-chart-python-matplotlib/

Write-up Question 1 (3 points): What company did you pick, and why did you pick that company's data download?

I picked Spotify for my data download because I recently switched from Apple Music to Spotify at the end of this past December. I also listen to music as much as possible so I thought it would be interesting to analyze what Spotify has learned about me given such a short period of time with (hopefully) sufficient information. I also just love music, so I thought it would be the most interesting to learn about the types of music I tend to gravitate towards and which I don't listen to as much or how my music taste has changed over time. One thing I found interesting about the data download is that it is only available for 14 days after requesting the data, even though (I assume) Spotify holds onto the data long after the data download request has been made and retrieved.

Write-up Question 2 (2 points): What is the high-level format of the data download? (Write just a sentence or two.)

The data is separated into various categories in the form of JSON files as seen in *Figure 1*. Since taking Surveillance Aesthetics and having parsed through JSON files for a similar assignment, I feel more confident in my abilities to traverse through the data with more ease than when first learning to parse through files using Python in the fall.

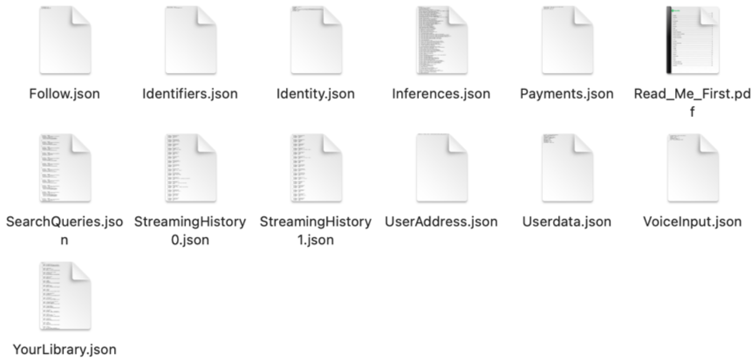
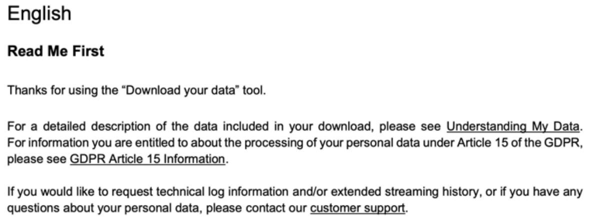


Figure 1


English

Read Me First

Thanks for using the "Download your data" tool.

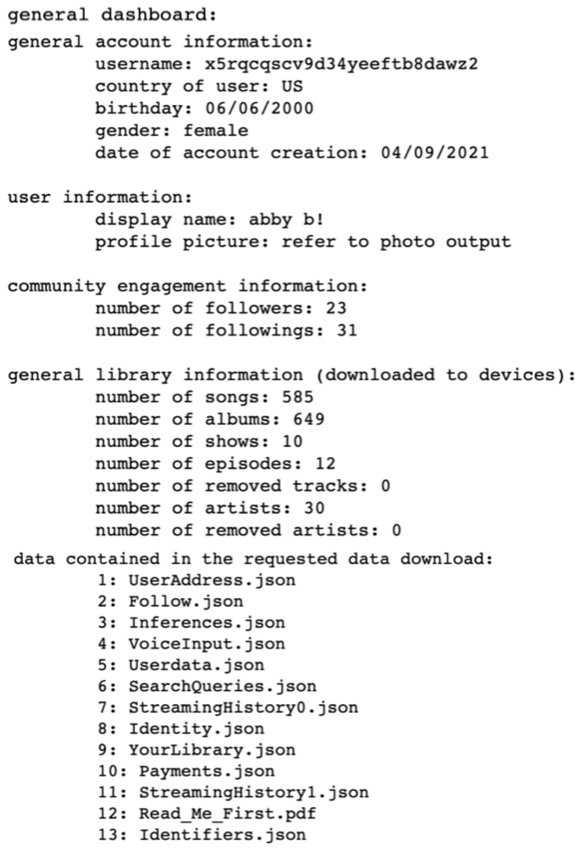
For a detailed description of the data included in your download, please see Understanding My Data. For information you are entitled to about the processing of your personal data under Article 15 of the GDPR, please see GDPR Article 15 Information.

If you would like to request technical log information and/or extended streaming history, or if you have any questions about your personal data, please contact our customer support.

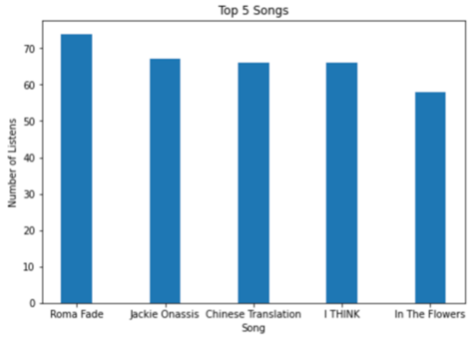
Figure 2

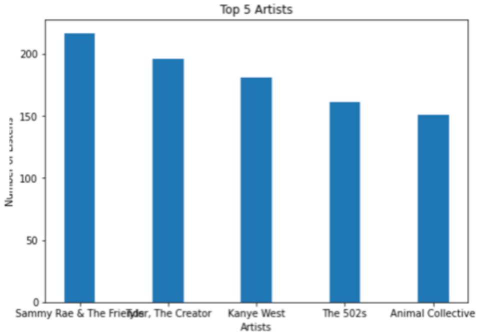
The data download also includes a Read_Me_First.pdf (*Figure 2*) which includes instructions on where to learn more about the collected data. From here, I used the "Understanding My Data" link to better understand the information given in my downloaded files. The files are relatively self-explanatory, but Spotify seems to go in depth with the information available for the user to collect, which will make later tasks presumably less convoluted.

Write-up Question 3 (5 points):



```
general dashboard:
general account information:
        username: x5rqcqscv9d34yeeftb8dawz2
        country of user: US
        birthday: 06/06/2000
        gender: female
        date of account creation: 04/09/2021

user information:
        display name: abby b!
        profile picture: refer to photo output

community engagement information:
        number of followers: 23
        number of followings: 31

general library information (downloaded to devices):
        number of songs: 585
        number of albums: 649
        number of shows: 10
        number of episodes: 12
        number of removed tracks: 0
        number of artists: 30
        number of removed artists: 0
 data contained in the requested data download:
        1: UserAddress.json
        2: Follow.json
        3: Inferences.json
        4: VoiceInput.json
        5: Userdata.json
        6: SearchQueries.json
        7: StreamingHistory0.json
        8: Identity.json
        9: YourLibrary.json
        10: Payments.json
        11: StreamingHistory1.json
        12: Read_Me_First.pdf
        13: Identifiers.json
```





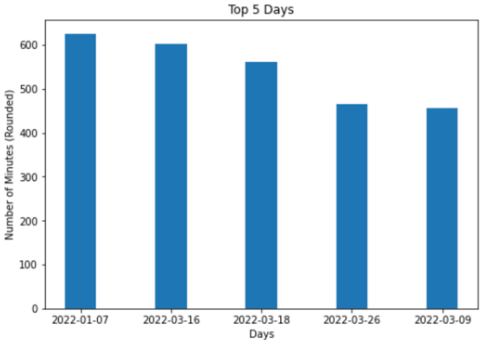
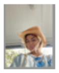
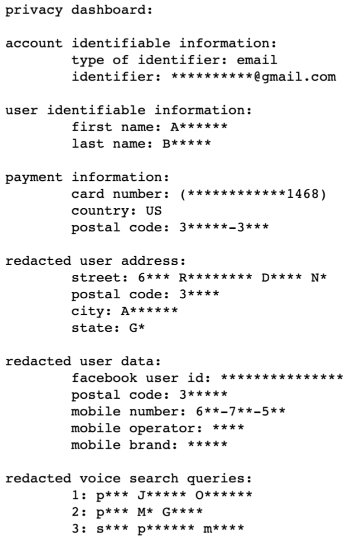


photo output



Write-up Question 4 (5 points):

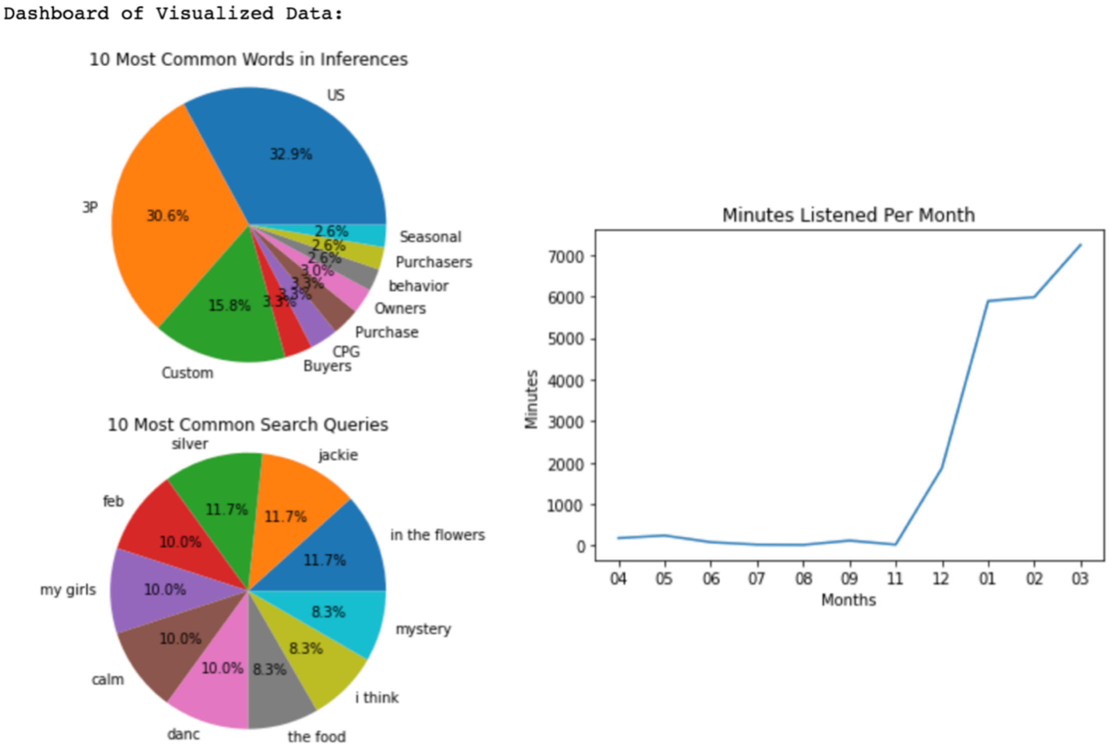


```
privacy dashboard:

account identifiable information:
        type of identifier: email
        identifier: **********@gmail.com

user identifiable information:
        first name: A******
        last name: B*****

payment information:
        card number: (************1468)
        country: US
        postal code: 3*****-3***

redacted user address:
        street: 6*** R******** D**** N*
        postal code: 3****
        city: A******
        state: G*

redacted user data:
        facebook user id: ***************
        postal code: 3*****
        mobile number: 6**-7**-5**
        mobile operator: ****
        mobile brand: *****

redacted voice search queries:
        1: p*** J***** O******
        2: p*** M* G****
        3: s*** p****** m****
```

Write-up Question 5 (3 points):

Going into this section of the assignment, I worked to visualize data that was both interesting regarding my searched in Spotify as well as visualizing data that was not revealing any information about myself that would be considered “private”. Especially when considering the Search Queries data and the Inferences data, I didn’t realize Spotify would take note of my searches (I thought there would be a larger emphasis placed on the songs I ended up listening to instead of the searches themselves), and it was interesting to see the inferences the application was making about me, especially considering I have premium so I don’t receive ads (this makes me wonder how these inferences play into the recommended music I see).

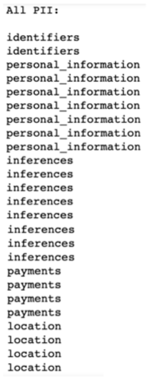
Write-up Question 6 (5 points):



```
Dashboard of Visualized Data:
```

Write-up Question 7 (5 points): In a sentence or two, describe the underlying concept of your visualization, summary, or interaction.

I wanted to show just how much information that could be considered Personal Identifiable Information is contained in the data that Spotify collects on us and could be exposed in a data breach, leaving the user in a compromised position. My visualization incorporates compiling all accounts of information that could be considered sensitive and repeats the information according to the number of times that general type of information appeared in the data that was requested and subsequently downloaded (note that *Figure 3* is not the actual number of repetitions per word but rather a condensed version of the list).



```
All PII:

identifiers
identifiers
personal_information
personal_information
personal_information
personal_information
personal_information
personal_information
personal_information
inferences
inferences
inferences
inferences
inferences
inferences
inferences
inferences
payments
payments
payments
payments
location
location
location
location
```

Figure 3

Write-up Question 8 (5 points): Please refer to *Figure 3*

Task 2: Exploration: Machine Learning (20 points)

1. Google Lens
    a. Who is the manufacturer (or authors) of the product or system?
        i. Google
    b. Include a link to where you read about the product or system.
        i. https://9to5google.com/2022/04/07/google-lens-multisearch/?utm_campaign=etb&utm_medium=newsletter&utm_source=morning_brew
    c. What decision is being made automatically using machine learning?
        i. The scanning and recognition of images combine with sourcing results based on both an image input and words associated from the user.
    d. What algorithm or machine learning approach is being used?
        i. Google Translate's neural machine translation (NMT) algorithms
    e. What data was used to train the algorithm, if applicable?
        i. Google's Knowledge graph
        ii. https://www.youtube.com/watch?v=Wd-yoxlVU4U&t=742s
    f. Briefly, what is one thing you think could go most wrong with using machine learning in this way? This could be related to the training data, the way input data is featurized, how the model's decisions are acted upon, etc.
        i. Giving results that are graphic or inappropriate could be an issue especially with untrained models given the expansive nature of the internet and the difficult that AI may have when scanning and "understand" what is in an image.
2. Tesla Beta Software
    a. Who is the manufacturer (or authors) of the product or system?
        i. Tesla
    b. Include a link to where you read about the product or system.
        i. https://electrek.co/2022/02/02/tesla-full-self-driving-beta-software-update-10-10-remove-rolling-stop-corner-cases/
    c. What decision is being made automatically using machine learning?
        i. The driving of the car on roads (more expansive than previous Tesla software in terms of distance the cameras can cover and read from)
    d. What algorithm or machine learning approach is being used?
        i. Ultrasonic sensors as neural networks internally in the car
    e. What data was used to train the algorithm, if applicable?
        i. Driver's responses (those who have high enough safety scores) to the environment combine with the objects and movements that the ultrasonic sensors in the car detect
    f. Briefly, what is one thing you think could go most wrong with using machine learning in this way? This could be related to the training data, the way input data is featurized, how the model's decisions are acted upon, etc.
        i. The car not properly categorizing an object and hitting someone/something
3. Yelp's Machine Learning Platform
    a. Who is the manufacturer (or authors) of the product or system?

        i. Yelp
    b. Include a link to where you read about the product or system.
        i. https://engineeringblog.yelp.com/2020/07/ML-platform-overview.html
    c. What decision is being made automatically using machine learning?
        i. Data inputs to make recommendations tailored to each user
    d. What algorithm or machine learning approach is being used?
        i. Deep learning analysis
    e. What data was used to train the algorithm, if applicable?
        i. Robust user data from searches and responses
    f. Briefly, what is one thing you think could go most wrong with using machine learning in this way? This could be related to the training data, the way input data is featurized, how the model's decisions are acted upon, etc.
        i. The car not properly categorizing an object and hitting someone/something

4. Spotify Algorithmic Playlists
    a. Who is the manufacturer (or authors) of the product or system?
        i. Spotify
    b. Include a link to where you read about the product or system.
        i. https://bestfriendsclub.ca/spotifys-algorithm-playlists/
    c. What decision is being made automatically using machine learning?
        i. Songs within a given playlist and songs being recommended to a user.
    d. What algorithm or machine learning approach is being used?
        i. Approximate nearest-neighbor algorithm.
    e. What data was used to train the algorithm, if applicable?
        i. Songs that the user listens to and songs that other users enjoy/frequently listen to.
    f. Briefly, what is one thing you think could go most wrong with using machine learning in this way? This could be related to the training data, the way input data is featurized, how the model's decisions are acted upon, etc.
        i. Getting songs that are not to a user’s liking, deterring them from using the given service.

5. Pinterest
    a. Who is the manufacturer (or authors) of the product or system?
        i. Pinterest
    b. Include a link to where you read about the product or system.
        i. https://medium.com/pinterest-engineering/an-update-on-pixie-pinterests-recommendation-system-6f273f737e1b
    c. What decision is being made automatically using machine learning?
        i. Images and content being shown to a given user.
    d. What algorithm or machine learning approach is being used?
        i. Pixie Random Walk Algorithm; graph based recommendation system
    e. What data was used to train the algorithm, if applicable?
        i. Pixie uses user data and data of users interacting with similar content to curate “pins” that are similar to that of the one being interacted with by the given user.

f. Briefly, what is one thing you think could go most wrong with using machine learning in this way? This could be related to the training data, the way input data is featurized, how the model's decisions are acted upon, etc.

i. Recommending pins that are not related to the interested of the user creating disengagement from the platform.

6. Twitter

a. Who is the manufacturer (or authors) of the product or system?

i. Twitter

b. Include a link to where you read about the product or system.

i. https://blog.hootsuite.com/twitter-algorithm/

c. What decision is being made automatically using machine learning?

i. What content to show to the user

d. What algorithm or machine learning approach is being used?

i. Could not find; seems like ranking signals and having a bias towards newer content

e. What data was used to train the algorithm, if applicable?

i. User data including past likes, followed users, followed topics, and previous engagement influence the algorithm.

f. Briefly, what is one thing you think could go most wrong with using machine learning in this way? This could be related to the training data, the way input data is featurized, how the model's decisions are acted upon, etc.

i. From the article, it is know that twitter’s cropping and content shown shows racial biases and political biases respectfully.

7. Google’s Deep Dream Generator

a. Who is the manufacturer (or authors) of the product or system?

i. Google

b. Include a link to where you read about the product or system.

i. https://en.wikipedia.org/wiki/DeepDream

c. What decision is being made automatically using machine learning?

i. How to combine images and how the art is being portrayed and manipulated.

d. What algorithm or machine learning approach is being used?

i. Convolutional neural networks and algorithmic pareidolia

e. What data was used to train the algorithm, if applicable?

i. User text and photographic inputs

f. Briefly, what is one thing you think could go most wrong with using machine learning in this way? This could be related to the training data, the way input data is featurized, how the model's decisions are acted upon, etc.

i. Using images pulled from APIs based on the textual input that were not intended (based on manipulated data/photo inputs)

8. Facebook ChatBot

a. Who is the manufacturer (or authors) of the product or system?

i. Facebook

b. Include a link to where you read about the product or system.

        i. https://www.firstpost.com/tech/news-analysis/facebook-researchers-shut-down-ai-bots-that-started-speaking-in-a-language-unintelligible-to-humans-3876197.html

    c. What decision is being made automatically using machine learning?

        i. What the chatbots are saying to each other to achieve a negotiated goal

    d. What algorithm or machine learning approach is being used?

        i. Neural Language Processing

    e. What data was used to train the algorithm, if applicable?

        i. Common English phrases and researchers inputted goals

    f. Briefly, what is one thing you think could go most wrong with using machine learning in this way? This could be related to the training data, the way input data is featurized, how the model's decisions are acted upon, etc.

        i. From the article, the chatbots created their own, unintelligible language that was seemingly English but nonsensical to the human researchers. This sort of black boxing could lead to (and has led to in the past) unintended, biased and/or discriminator conversations between chatbots.

9. NukkAI Bridge AI

    a. Who is the manufacturer (or authors) of the product or system?

        i. NukkAI

    b. Include a link to where you read about the product or system.

        i. https://www.sciencetimes.com/articles/36874/20220330/artifical-intelligence-system-french-startup-nukkai-defeats-eight-world-champions.htm

    c. What decision is being made automatically using machine learning?

        i. Moves in the game of bridge

    d. What algorithm or machine learning approach is being used?

        i. Could not find – discussion of "white boxing"

    e. What data was used to train the algorithm, if applicable?

        i. Rules of the game bridge and user actions/trial and error over time (practice).

    f. Briefly, what is one thing you think could go most wrong with using machine learning in this way? This could be related to the training data, the way input data is featurized, how the model's decisions are acted upon, etc.

        i. AI pursuing ill-defined or dangerous goals given a lack of user clarity

10. Netflix Recommender System

    a. Who is the manufacturer (or authors) of the product or system?

        i. Netflix

    b. Include a link to where you read about the product or system.

        i. https://towardsdatascience.com/deep-dive-into-netflixs-recommender-system-341806ae3b48

    c. What decision is being made automatically using machine learning?

        i. Media being recommended to the user.

    d. What algorithm or machine learning approach is being used?

        i. Specified as Netflix's "machine learning algorithm", but I could not find anything more specific than that.

    e. What data was used to train the algorithm, if applicable?

    i. User interaction with shows and movies cross-listed with other users engagement with the application.

f. Briefly, what is one thing you think could go most wrong with using machine learning in this way? This could be related to the training data, the way input data is featurized, how the model's decisions are acted upon, etc.

    i. Recommendation inappropriate or irrelevant content leading to a decrease in user engagement.