Ember Shan, Lizzie Potocsky, Ben Decker

Link To Slidedeck: https://docs.google.com/presentation/d/1518YrtUJBPCOxQE1ziHD7d4Ze6t0S2jbClU-AjWtdVE/edit?usp=sharing

Link To Github Repo: https://github.com/discobaby/finalproject.git

Initial Goals For Project (10 pts)

- Analyze relationship between net worth and avg number of followers/fans of a celebrity.
- Analyze relationship between gender and net worth
- Analyze relationship between gender and number of followers on Spotify and Twitter
- Visualize data corresponding to an artist's average number of Twitter followers pertaining to the artist's genre
  - Answers: Do certain genres lead artists to having more followers on social media?
- Answer the question: As artists gain popularity with their music, do they also gain popularity on subsequent social media platforms?

Goals Achieved (10 pts)

- We achieved goals in seeing if the amount of twitter followers and spotify listeners relate in some way, and if it correlates to a larger net worth.
- We customized a database consisting of a plethora of data related to Twitter/Spotify followers, gender, artist genre, net worth, and average number of fans.
- We utilize our database to create visualizations that illustrate the potential correlations between net worth and average number of followers/fans of a celebrity, between gender and net worth, and between gender and number of followers on Spotify and Twitter.
- We also successfully curated a bar chart showing the average number of followers on Twitter for the artist's corresponding Spotify music genre.

Problems Faced (10 pts)

Our group's original idea was to get the celebrity's number of followers from Tiktok, but TikTok required people to host a website in order to get the data for business uses. Due to this, our group ultimately decided to abandon this idea. We switched to the Twitter API to find the celebrity's number of Twitter followers instead. We also scraped a website to find the top 200 artists and their "fan count".

Another difficulty is that some of the artist names that we found are band names, and the net worth API can only deal with a person. There are also multiple celebrities who have the same name such as Drake. To solve this, we used comparison operators to see if the name returned by the API is exactly the same as the artist name retrieved from the Artist table.

Another difficulty is that the Celebrity API does not have information for some celebrities, so empty fields were represented with a value of -1 or NA.

The Spotify API requires an account that gives a “client id” and “client secret”. I had to include these values in my code and create a variable in order to get an access token that can be used in the functions that follow.

File Containing Calculations (10 pts)

There are two files that contain the calculations: **avg_networth_by_gender.txt** and **avg_twitter_followers_by_genre.txt**.

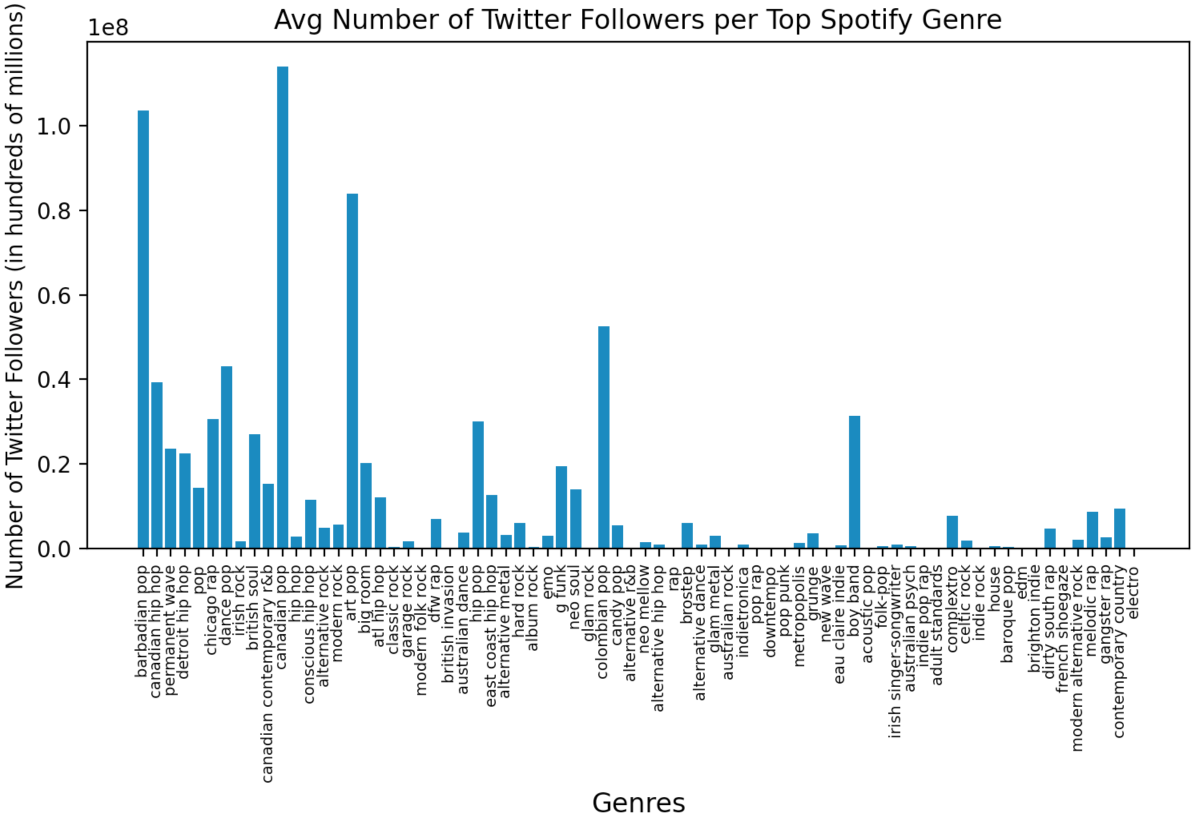
The function **getTwitterFollowersByGenre** in vizualzation.py has the calculations. To do this, we went through each of the genre_id’s that we pulled from spotify. We joined the twitter follower data table and spotify artist table, and I pulled all of the twitter followers by artists in a specific genre, using the genre_ids. I calculated the averages of followers on twitter for each of the genres, and included it in my visualization. So for a hypothetical example, Lady Gaga and Lana Del Rey are considered “art pop.” Gaga has 70 million twitter followers and Lana has 40 million, so the average number of twitter followers for artists who are “artpop” would be 55 million followers ([70m + 40m] / 2 = 55 million). The output file has two columns with genre as the first and average twitter followers as the second column.

The function **getAvgNetworthByGender** in vizualzation.py also contains calculations. It calculates the average net worth of female/male celebrities. The output file is simple which has only two numbers: averages of net worth of male and female celebrities.

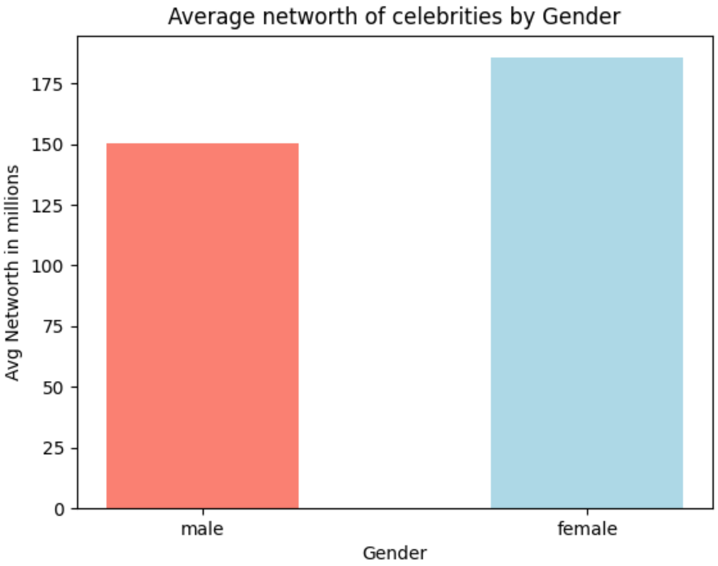
Visualizations- Include screenshots (10 pts)

A total of 4 visualizations

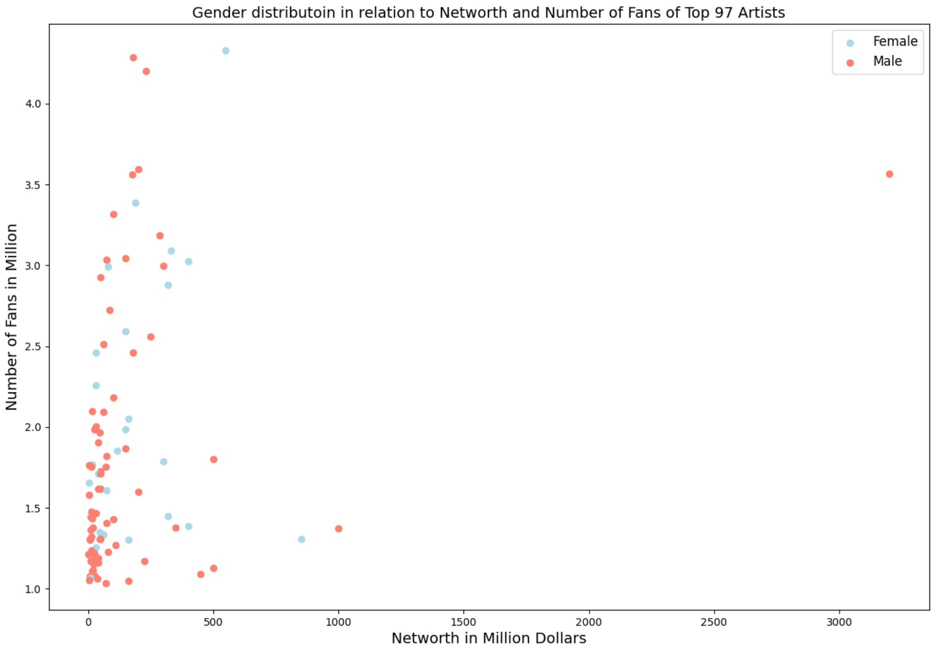
1. Each bar represents a genre and the average number of twitter followers
    a. Finds that artists in “pop” genres have significantly more followers than those of other music types.



2. Bar graph that represents the averages of net worth of female/male celebrities
    a. Average of male celebrity net worth seems to be lower than female, but the data are only collected from 97 artists so it is not representative and there could be bias

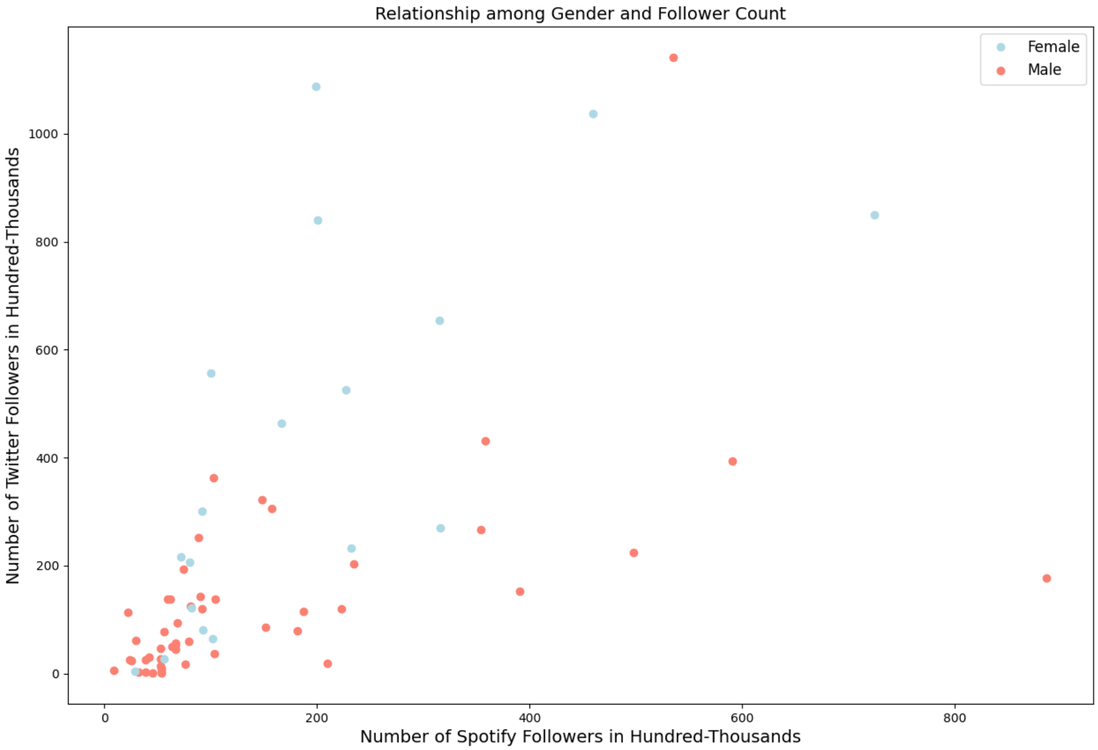


3. Dots are colored based on gender and placed in relation to net worth (x-axis) and Number of fans (y-axis)
    a. It seems like there are more male than female artists among the 97 top artists. But the distributions seem to be pretty similar with few outliers like the red dot on the right side. Additionally, those with higher net worth appear to have a higher number of fans. There does not seem to be a significant difference between male and female artists' data.



4. Relationship among gender and follower count

a. Those with more followers also seem to have more Twitter followers. We seem a positive correlation. It also appears that more female celebrities represent outliers that have higher follower-counts on both social platforms, but it is hard to tell whether there is a correlation between gender and follower-count or if our sample simply had more female celebrities.



Instructions of running the code (10 pts)

Order to run the files:

1. **get_artists_names.py must be run first** to get the list of names which will be used in other python files
2. Then you can run get_net_worth.py, get_spotify_info.py, and get_twitter_data.py, order does not matter
3. visualization.py must be run last as it needs the data processed by other files

All tables and data are stored in the **'finalproject.db'** database. Visualizations are in the **visualizations** folder. And files that store the calculated data are in the **calculation_files** folder.

Documentation for each function; include input & output (20 pts)

get_artist_names.py

createSoup() returns soup based on the url
createSoupFromFile() returns soup based on the html file (only serve as a backup option if the actual website fails)
getLastID(cur) returns the id from the last row of a table
getData(soup, cur, conn) returns nothing but calls the createArtistDatabase function
createArtistDatabase(cur, conn, names, fans, start):

- names and fans are lists, start is id from **getLastID**
- returns nothing and inserts the artist names to the Artists table

main() that calls all the necessary functions to get the 20 names from the website, execute it 10 times to get all the 200 artist names

Get_net_worth.py

getKey() returns the personal key from the API_keys.txt
getLastID(cur) returns the id from the last row of a table
getNameList(cur) returns a list of tuples in the format of [(id, name)]
getNetWorth(names, cur, conn) returns nothing and calls insertIntoDatabase
insertIntoDatabase(networth, cur, conn) returns nothing and inserts data into the NetWorth table
main() needs to be called 10 times to get all the data

Get_twitter_data.py

def getId(cur) gets the api keys
def getNameList(cur) returns a list of twitter names
def getData (name, cur, conn) goes through and adds twitter followers to the names
def main() runs the entire function, returns a database with twitter screen names, artist_ids, and the followers

get_spotify_info.py

getLastID(cur, tablename, idname) creates the SpotifyArtistData table and also the SpotifyGenreData table (has a shared key of genre_id). This function returns the id from the last row of a table.
getListofArtists(cur) returns a list of tuples in the format of [(id, artist name)]
getArtistInfo(names, cur, conn) returns nothing and calls insertIntoDatabase
insertIntoDatabase(artist_info, cur, conn) returns nothing and inserts data into the SpotifyArtistData table and the SpotifyGenreData table
main() needs to be called 10 times to get all the data

Visualization.py

genderScatterPlot(cur, conn) returns the scatterplot illustrating the correlation between an artist's net worth and number of fans, by gender
getAvgNetworthByGender(cur) returns the bar graph illustrating average net worth of celebrities by gender
genderFollowersScatterPlot(cur, conn) returns the scatterplot illustrating the correlation between an artist's Spotify followers and Twitter followers, by gender
createGenrelist(cur, conn) returns a list of genre_ids to iterate through
createGenrelistname(cur, conn)returns a list of genre Id names to later be used on the x-axis
getTwitterFollowersByGenre(cur, conn) returns a bar graph of avg twitter followers by genre
main() takes nothing and returns nothing. establishes the database connection, calls the listed functions, and closes the database connection

Resources

Date	Issue Description	Location of Resource	Result (did it solve the issue?)
11.27	Need to get the id of the last row to determine what next 20 items to scrape/get from website/API	https://stackoverflow.com/questions/4073923/select-last-row-in-mysql	Solved
12.2	Need to remove all comma from the fans_count which is a string	https://www.journaldev.com/23763/python-remove-spaces-from-string	Solved.
12.2	Need to annotate the points in scatter plot with artist names	https://www.geeksforgeeks.org/how-to-annotate-matplotlib-scatter-plots/	Solved
12.6	Cannot pull the files from the remote git repo	https://stackoverflow.com/questions/26281767/run-command-git-branch-set-upstream-to-produce-error/42305655	Need to set upstream first then git pull. Solved.
12.6	Change bar plot to have different colors of bars	https://www.python-graph-gallery.com/3-control-color-of-barplots	Solved