Project-2

Boris Rubel, Dylan Castillo, Melina Heredia, Jacob Trevithick

---

Project Proposal

Motivation:

Our project motivation derived from a marketing strategy to utilize top Twitch streamer data. The strategy utilizes data of top streamers on the interactive livestreaming service for content spanning gaming, entertainment, sports, music, and more. Our team specifically looked at top rank streamers analyzing their views, watch time, followers, and gross income. These insights will ultimately assist our Gaming Clothing & Accessory company to find the best streaming audiences with sufficient buying power and streaming durations to successfully launch our new product line.

Data:

Dataset 1 - https://www.kaggle.com/aayushmihra1512/twitchdata

Dataset 2 - shttps://web.archive.org/web/20211006143529/https://pastebin.com/LjmaPNam

Production:

The data utilized was first extracted from the two data sources cited above. Two csv files were read using the pandas.read_csv command to produce an earnings and user info data frame.

Cleaning:

The two data frames were then cleaned so that only the necessary columns were present. Commands such as ".drop" and ".rename" were used during this process.

To clean the earnings data, dataset 1 was read as the earnings data frame. From this, username, user ID, and gross earnings were used to create the earnings_df.

To clean the user information, dataset 2 was read as the user info data frame. The columns were renamed, specifically highlighting the renaming of column "Channel" as "username." The user info data frame was then merged with the earnings_df on "username" to create user_info_df. This data frame was then used to create user_info_stats_df utilizing columns user_id, watch_time, stream_time, peak_viewers, avg_viewers, followers, followers_gained, and views_gained.

The user information was further cleaned using the merged user information data frame and earnings data frame to create user_info_new_df. This data frame contained columns user_id, username, partnered, mature, and language.

The final step in the cleaning process included merging the earnings data frame and new user information data frame to create earnings_new_df.

Database:

A relational postgreSQL was then used to store the cleaned data. The tables stored within the database include the Twitch streamer's user ID, which can be joined-on to combine the tables.

A user_info table can be used to view user_id, username, if the streamer is partnered, (mature?), and the language they speak while streaming. The data that populates this table originates from the user_info_new_df created during the cleaning process.

An earnings table can be used to view user_id, username, and gross_earnings. The data that populates this table originates from the earnings_new_df created during the cleaning process.

A user_stats table can be used to view user_id, watch_time, stream_time, peak_viewers, avg_viewers, followers, followers_gained, and views_gained. The data that populates this table originates from the user_info_stats_df created during the cleaning process.

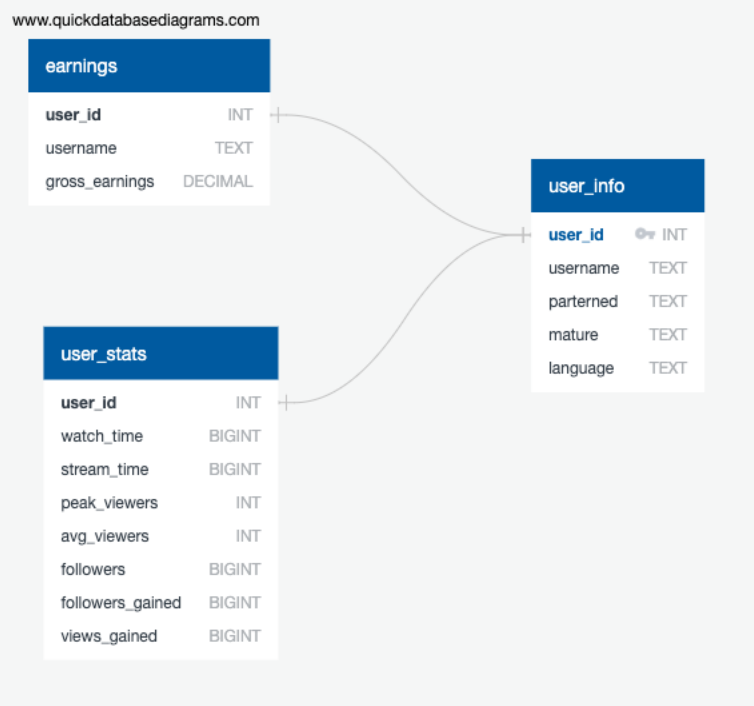

Figure 1: Relational PostgreSQL Database Diagram.

Final Use:

The final database created to store the data cleaned during the project process, including the tables produced, can ultimately be utilized to analyze, and generate insights, to assist in the launch of our Gaming Clothing & Accessory company. The three tables loaded, user_info, earnings, and user_stats, can all be joined in one method or another to view and generate statistics to better make decisions surrounding how to best launch our company. We are now able to view not only top ranked streamers, but a combination of additional data that may influence how to go about marketing our Gaming

Clothing & Accessory company that will be influenced by what streamers may have larger audiences or differing metrics surrounding Twitch streamers.