**Data Creation Procedure**

\*\*\*\*\*\*\*\*

In creating the combined inserts file, it was discovered that the autogenerated IDs for tables such as influencers do not always start at 1 which means the other insert tables which use those IDs as foreign keys are not able to properly create references.

In order to address this issue, we used the Back-Up tool from pgAdmin to extract the table data exactly as it is in our database. This file was tested and ran as expected.

\*\*\*\*\*\*\*\*\*

We initially wanted our dataset to be completely realistic but eventually realized we did not have enough real-world data to be able to complete the project to our original vision. So we had to creatively generate additional information to create connective elements between elements.

Most of the initial data came from datasets we found on Kaggle.com

The data for each table was generally put into a CSV and a python algorithm was created and ran to generate insert statements by going line by line through the CSV (the python code for this can be provided on request). For the contracts and pays table however, the python code use the company and influencer entries to generate new data and insert statements.

Here is a breakdown of how each table was created:

Country

We used a Kaggle dataset which listed countries with population and other statistical info. However, that dataset contained even the smallest uninhabited islands. We decided that using all those countries was unnecessary, especially if we pretend to be a US based marketing firm. Instead, we picked out 66 countries that are among those with the highest GDP and are the most likely to allow a US based marketing firm to operate.

We then found another website (iban.com/country-codes) which listed all the country codes for each nation.

Influencer table

We found several different data sets on Kaggle for social media users, they had names such as “top 1000 youtubers”. We settled on 3 datasets and took 1000 “top influencers” each from YouTube, Instagram, and TikTok. Some had incomplete country data others had no country data at all. So initially we dropped the country data requirement and tried to normalize the data as best we could.

Eventually we realized that it is difficult to create connections between tables as is and decided to generate a country column for influencers. As this did not have to be realistic, we decided to randomly assign countries to each influencer. However, from a data point of view, a purely random distribution would not be interesting. So, we decided to weigh the probability based on a country’s population size. As a result, most of our influencers are listed as residing in either China or India, which we thought was amusing.

Company table

We were unable to find a good table for company data on Kaggle, so we used the website companiesmarketcap.com instead. We did some data scraping and pulled the top 500 companies by market cap.

Contract and pays table

We realized that we did not have enough connectivity between tables to show transactions or to do queries. So, we generated two tables in order to create stronger associations.  
  
Our algorithm randomly selected a company, randomly selected a date within the past 5 years, and generated a random six digit number. This was used to create the contract ID. For example, an Amazon contract from March 2020, could look like AMA\_MAR20\_123456.

Then a random influencer was selected to serve as the individual the contract was with, and an end date for the contract was picked at random a few months after the start of the contract.

For every entry in the contract table there was a corresponding entry made in the pays table. A new random six-digit number was generated to serve as the pays ID, then we gave the contract ID, a date 30 days after the contract end date was given as the payment deadline. Then an amount was generated between 5 to 6 digits to symbolize the value of the contract. The among was split 70/30 serving as payment/fee. This represents that 70% went as payment to the influencer and 30% went to us as agents/talent scouts.