SMU Data Science Bootcamp September 23rd, 2021



**A Case Study of Extract, Transform and Load: Spotify**

Steven Green

Greyson Moore

Saatvi Rajgopal

Lucas da Silva

**Background and Inspiration:**

Spotify is a cloud-based music streaming service that allows users to stream millions of songs to

their phone, computer or internet-connected device. Spotify is the world’s most popular music

streaming service with 248 million total users. This type of music streaming has become the

new normal for music listeners. Because the music is hosted by the same platform, Spotify is

able to present top music charts by genre and location based on streaming numbers. It also has

charts to showcase new and potentially popular songs. The “Viral 50” is determined by

compiling streaming data along with analytics from social media sites and blogs.

**Proposal Topic:**

We will be creating a database on Google Cloud to store song information from Spotify,

specifically, the “Top 200” and “Viral 50” songs. We will be looking at this data from each

individual chart by the most populated countries; United States, Thailand, India, Brazil, Mexico,

Japan, Philippines, Egypt, Vietnam, and Turkey. These are not the most populated countries in

the world, but the most populated countries that Spotify provides the data for. In addition, we will

look at the data globally to get an overall perspective. We will be looking at the data by week,

and the time frame we are using will be from July 29, 2021, to September 1, 2021.

**Process:**

In order to create the database, the group will be web scraping song information from the Spotify

charts website. The website contains a week-to-week list of the “Top 200” streamed songs that

displays the rank, song title, artist and streaming numbers. It also contains a week-to-week list

of the top 50 viral songs, called “Viral 50”. Using the Python library “BeautifulSoup”, the group

will gather the information from each song over the selected time frame and organize it into data

frames using the Python library “pandas.” The data within the data frames will be cleaned and

organized and uploaded to a PostgreSQL database. The group will upload the database to the

Google Cloud.

**ERD:**

**Data Engineering**

**Web Scraping:**

**Data Cleaning:**

**Conclusion:**

**Sources:**

<https://spotifycharts.com/regional/global/weekly/latest>

<https://jennifer-franklin.medium.com/how-to-scrape-the-most-popular-songs-on-spotify-using-python-8a8979fa6b06>

<https://www.worldometers.info/world-population/population-by-country/>