ETL Project

Christopher Turner

Jordan Myer

Ryan Carnes

**Extract**

To begin the project the team discussed which site we would search for the data and what type of source we would use. After a brief discussion we decided to use Kaggle and Data.World for our 2 separate data sources. Which both are CVS file types.

Our team decided to gather data on Spotify songs downloading the Top hits of 2017. The secondly popular holiday music.

Before downloading the CSV files, the team created a new GitHub repository. After creating the repository, the team uploaded the CSV files. Once the CSV files where uploaded the team opened up Jupyter Notebook and began to transform the data.

**Transform**

Once in the Jupyter Notebook the team imported the necessary dependencies: **import pandas to PD and from sqlalchemy import create\_engine**.

From there the team pulled the data from the CSV file and assigned the data a variable called DF for the Holiday\_Music and Music\_DF for the top hits of 2017.

After creating the variables, we created a new variable to return a data frame with only selected columns. The selected columns chosen were Artist, Song Name, and Danceability.

Once the new data frame with selected columns was created, we went into PG Admin and created a new data base. Once the date base was created, we created the tables. The columns names matched the columns selected in the data frame.

**Load**

Once the new data base was created, we connected to our local postgres database in pandas using the following code:

**rds\_connection\_string = “postgres:postgres@localhost:5432/Holiday Music”**

**engine = create\_engine(f’postgresql://{rds\_connection\_string}’)**

Once the connection was complete we checked to confirm the tables were entered using the following code: **engine.table\_name()**

From there we transferred the new data frame into the data base.

Finally, we performed a series of queries to ensure the data properly loaded into PostgreSQL database. The following query was used to confirm: **pd.read\_sql\_query(‘select\*from Table Name’, con=engine).head()**

We used postgres as the final production database. Which is a relational database. Which we used because we are more familiar with that process.