Technical ETL Report

Step 1

Finding a large dataset on Kaggle that had many data points. Downloaded the files for extraction to a local folder on the Desktop labeled ETL Project.

Step 2

Opening Python and setting dependencies and as well as a file path in Python to the ETL Project Folder to access the csv files. In order to perform extractions we will use a python tool called Pandas.

Step 3

Extract the data from the csv files using the pd.read\_csv.

Step 4

Create dataFrames in pandas.

Step 5

Here is where we transform the data and start to manipulate the dataframes filtering and cleaning only the data we want to use. This can be done many ways but for this project’s purposes, the .loc command was used in selecting the columns desired for the final data set.

Step 6

Once the final data sets were created, pgAdmin(SQL) was used to create a database called ETL\_db.

Step 7

Within the database, created schema to make tables. It’s important to match the column names in your dataframe as well as in your schema so that when you load, it aligns and inserts into your table correctly. Check to make sure tables are correct and empty.

Step 8

Then in Python, create the setup for the load. This involves importing sqlalchemy, and from sqlalchemy import create\_engine, func. These tools allow us the create a connection to the database and use python to load the dataFrames we created into the tables in SQL.

Step 9

Use Pandas to load into the database with the .to\_sql command.

Step 10

Use a query in SQL SELECT\*FROM (table\_name); run this query and verify data has correctly and completely loaded into the database.