Final Report

Andrea Pappa, Carlie Azar, Dave Moorman, Jake Geiser

Extract

The purpose of this project was to collect data from outside sources, combine them together in a fashion that would allow for an in-depth, and provide a thorough look at a relational database. We chose datasets formatted into CSVs that collected international cities, specifically where the population exceeded over 100,000 people. Using Kaggle, our group was able to pull CSVs of the city and its respective country, of these cities there was another dataset describing the cost of living for items, including the average cost of a cappuccino, a ticket to the movies(cinema), a glass of wine, a gallon of gasoline, the average rent, and the average disposable income all measured in GBP from 2017. A third CSV included ratings on the overall quality of life for each city, including ratings on crime, health care, pollution, and purchasing power (average cost of living with the average local wage). Alongside our original sources, we also decided to obtain the actual population count of each city that exceeded 500,000 people, the previous years' population count, and the growth rate from another CSV obtained through the World Population Review website. We used SQLAlchemy to generate a relational database of all of our collected CSVs to allow others to also work with this data collection.

Transform

Following the extraction of the CSVs from Kaggle, and the CSV from world population review for population count and growth, we uploaded the datasets into Pandas to begin cleaning and transformations. The CSVs containing city location, quality of life ratings, and cost of living averages in GBP were analysed for any duplicates or any "N/A" values. Finding no duplicates or any non-applicable values, we joined the datasets for location and cost of living by the "city" variable, and created a new dataframe to analyze the relationship between the average cost of rent and the average disposable income. The same was found and done for the CSVs containing city location and the quality of life ratings per city. We desired to look at the overall quality of life rating, so we renamed the columns and sorted to display such values. Once the sort was completed, we merged the dataframes containing the average cost of rent and average disposable income with the dataframe reporting the average quality of life rating per city.

The data pulled from world population review as a CSV file was merged with the quality of life, the cost of living dataset, and the cities dataset. To complete this, the datasets pulled from Kaggle were cleaned by dropping any city with an unknown country, which was indicated by showing up as a duplicate due to the cities.csv having duplicate city names and the other Kaggle files lacking a country column. The population CSV was converted into a pandas dataframe, where the columns were renamed to better suit the data and the population rank was dropped. All

of the clean dataframes were merged, dropping any null or duplicate values, and then moved the “country” variable to the front of the table beside the “city” column.

Load

The newly transformed database was then loaded into PostgreSQL to be made available for others to further manipulate the data and conduct analysis of their own. We chose to use an open-source relational database manager to load our data, in order to more efficiently make the data available for use in various applications. Using an ORM will allow the automation of transferring data stored in this relational database that can be used to speed up application development for projects needing this data.

Guidelines for ETL Project

This document contains guidelines, requirements, and suggestions for Project 1.

Team Effort

Due to the short timeline, teamwork will be crucial to the success of this project! Work closely with your team through all phases of the project to ensure that there are no surprises at the end of the week.

Working in a group enables you to tackle more difficult problems than you'd be able to working alone. In other words, working in a group allows you to ****work smart**** and ****dream big****. Take advantage of it!

Project Proposal

Before you start writing any code, remember that you only have one week to complete this project. View this project as a typical assignment from work. Imagine a bunch of data came in and you and your team are tasked with migrating it to a production data base.

Take advantage of your Instructor and TA support during office hours and class project work time. They are a valuable resource and can help you stay on track.

Finding Data

Your project must use 2 or more sources of data. We recommend the following sites to use as sources of data:

* data.world
* [Kaggle](https://www.kaggle.com/)

You can also use APIs or data scraped from the web. However, get approval from your instructor first. Again, there is only a week to complete this!

Data Cleanup & Analysis

Once you have identified your datasets, perform ETL on the data. Make sure to plan and document the following:

* The sources of data that you will extract from.
* The type of transformation needed for this data (cleaning, joining, filtering, aggregating, etc).
* The type of final production database to load the data into (relational or non-relational).
* The final tables or collections that will be used in the production database.

You will be required to submit a final technical report with the above information and steps required to reproduce your ETL process.

Project Report

At the end of the week, your team will submit a Final Report that describes the following:

* ****E****xtract: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).
* ****T****ransform: what data cleaning or transformation was required.
* ****L****oad: the final database, tables/collections, and why this was chosen.

Please upload the report to Github and submit a link to Bootcampspot.