CS24200: Homework 5

Due date: Friday December 6, 11:59 pm EST

Use python to complete this assignment. Submit a PDF with both the code that you used for analysis and your answers to the questions below. Your homework must be typed.

Download the "Chinook" music database here. The Chinook data model represents a digital media store, including tables for artists, albums, media tracks, invoices and customers.

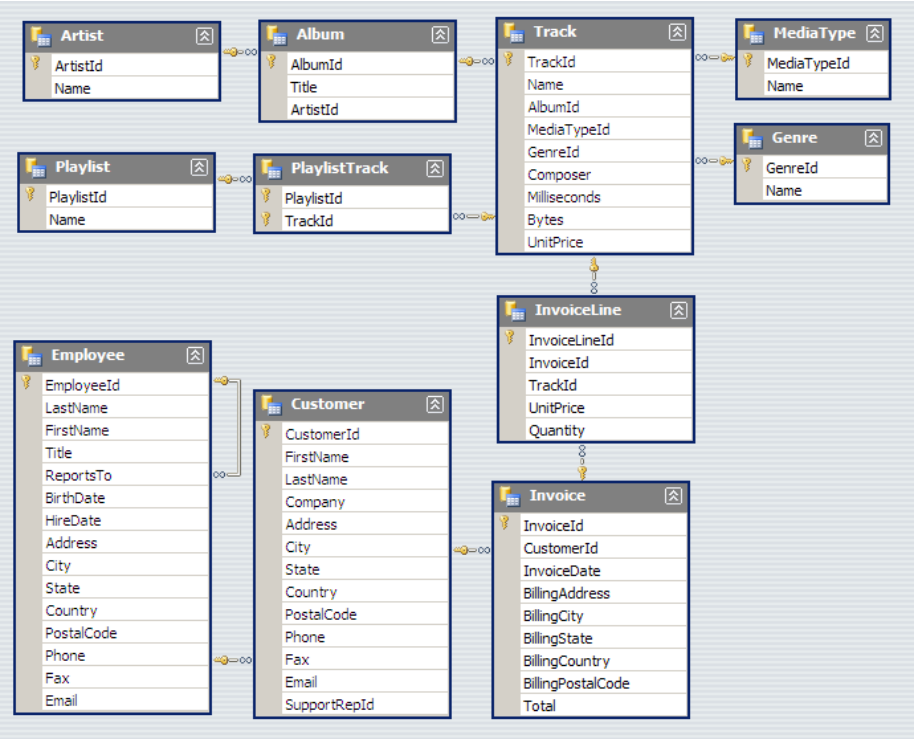

Figure 1: Schema of relations in Chinook database.

In this assignment, you will use sql to select data from the database to parse, transform, and analyze. You will evaluate and visualize the results.

Start by connecting to the `Chinook_Sqlite.sqlite` database file in a python notebook.

1 Basic SQL Queries (20 pts)

Write SQL queries to find the following:

(a) The InvoiceId and CustomerId of the top 15 invoices i.e. with the highest total.

(b) The names of all the employees whose first names match with some customer's first name.

(c) The names of all the employees whose last names match with some customer's last name

(d) The most popular genre (i.e., the genre with most tracks). Return the genre name.

(e) The customer who bought the highest amount of tracks. Return the customer name.

(f) Find all artists that have more than five albums, return the name and the count of their albums.

(g) For the artist Iron Maiden, find all their genres. Return the artist and genre names.

(h) Find the albums that appear on more than three playlists. Return the album name and number of playlists they appear on.

2 Clustering Artists (20 pts)

In this question, you will gather information on the artists and their tracks, and then cluster the artists into groups.

(a) Write an SQL query to gather information about the artists, their albums, tracks, genres, and playlists. Import this information into a pandas data frame. (Note: you can do this in one big join.)

(b) Select all the artists that have more than one album for the analysis below.

(c) Construct a set of ten features for each artist:

- Genre: Create a numerical feature for each of the top 7 genres that records how many songs the artist has from that genre. (Determine the top 7 genres in the same ways as Q1c.)
- Number of albums: Count of how many albums the artist has in the data (note this should be ≥ 1 based on filter above).
- Number of tracks: Count of how many tracks the artist has in the data.
- Number of playlists: Count of how many playlists that include any track of the artist.

(d) Apply k-means clustering to cluster the artists based on the features above. (*Note: you should have one row of features per artist.*) Consider values of $k = [2, 4, 6, 8, 10]$ and choose an appropriate value of k based on `inertia` scores. Include a description, discussion, or plot to support your choice of k.

(e) For your chosen value of k, for each cluster, find and report the three artists that are closest to the cluster centroid. (*Note: these should be different artists for each cluster.*) Include their names and feature values (calculated above). Discuss whether you see any patterns in the discovered clusters (e.g., do they cluster by genre?).