Case Study: Web Scraping

Background

1. Provided to you is set of instructions for data to be gathered using web scraping and some pre-processing.
2. Git hub repository for version control, code submission and pertaining resources.

Resources

1. Click here to access GIT repository.

Key Deliverables

1. Scrape Wikipedia's Billboard pages from 1992 to 2022 (link provided in the Jupyter notebook). There are also some hints provided in the notebook for optimization and guidance for best practices to be followed.
2. Parse the HTML retrieved via response call to extract ranking, song, and artist information.
3. Construct a Data Frame from parsed data and convert columns to their correct data types if needed.
4. Store this Data Frame in csv or pickle format so that you can use it for tasks ahead. Push this data to a Google Drive (secure, non-public)
5. Perform exploratory data analysis and answer the following questions:
    a. What has been the trajectory (trends) of various genres in the popular zeitgeist? & How has the popularity of these 30 genres changed with time?
    b. What are the 30 most popular genres over the 3 decades?
    c. Create a subframe of the ranking and year for each genre.
    d. Use Group by () function to group by year to create a Data Frame that contains the rankings of every song from that genre each year.
    e. Who are the highest quality singers?
    f. Who are the most occurring artists in Billboard's Top 100 list?
    g. Count the number of times a singer appears in the top 100 over a certain period. Consider an artist appearing twice in a year as two appearances.
    h. Plot a bar chart of the artists who have occurred at least more than 15 times in the given time frame.
    i. What is the age at which singers achieve their top ranking?
    j. Plot a histogram of the age at which artists reach their top ranking.
    k. At what year since inception do bands reach their top rankings?
    l. Plot a histogram of the years since inception at which bands reach their top ranking.
6. Connect this data, *without making it public*, to PowerBI or Tableau and create a basic report using visualization to answer the questions asked above.

Please share your output in a google drive and provide the editor level access to alerts@datachamps.ai .