Case Study: DATA ENGINEERING

Background

1. Provided to you is set of instructions for web scraping and data 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. Create a Solution and Technical architecture for building the pipeline and storage.
2. Use any Cloud Services of your choice to establish storage & processing components and use key-vaults for connecting all resource. Do not expose credentials in the code.
3. Scrape Wikipedia's Billboard pages from 1992 to 2021.
4. Parse the HTML retrieved to extract ranking, song, and artist information.
5. Construct a Data Frame from parsed data and convert them to the correct data types if needed.
6. Store this Data Frame in ADLS so that you can use it for tasks ahead.
7. Perform Exploratory Data Analysis and answer the following questions:
    a. What has been the trajectory of various genres in the popular zeitgeist?
    b. What are the 30 most popular genres?
    c. How has the popularity of these 30 genres changed with time?
    d. Create a subframe of the ranking and year for each genre.
    e. 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.
    f. Who are the highest quality singers?
    g. Who are the most occurring artists in Billboard's Top 100 list?
    h. 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.
    i. Plot a bar chart of the artists who have occurred at least more than 15 times in the given time frame.
    j. What is the age at which singers achieve their top ranking?
    k. Plot a histogram of the age at which artists reach their top ranking.
    l. At what year since inception do bands reach their top rankings?
    m. Plot a histogram of the years since inception at which bands reach their top ranking.

Please share your output in a google drive and provide the editor level access to the designated folder.