[Spring-2022] 36-652/752 Course Project

Accept this assignment by accessing GitHub classroom via the following URL:
https://classroom.github.com/a/gU0mcYUV

In this project, you will use FIFA Dataset available on Kaggle
https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset?select=players_20.csv

This dataset contains Soccer player statistics for 2015-2020. In this project, you will need all the dataset across all years to conduct the following tasks:

General Expectations

- Follow coding best practices with well-documented code.
- Add your dataset under data folder
- You may choose 1 peer to work with on this project.
- If you choose to work with peer, write the name of your peer in the Canvas submission. If you fail to do so, your peer will not get the grade.
- We will not handle cases where students forget to submit the name of their peers.
- You don't need to include Career mode player data or female player data.

Task-I: Build and populate necessary tables (40% of course project grade)

- Ingest your data into CSV.
- Create the DB table to store the dataset and populate it with cleaned ingested data.
- Identify constraints as needed and document them in your Readme.md file.
- Your tables should be created in schema with the name "fifa".
- In your ReadMe.md, attach a screenshot of your table infrastructure (e.g. DbVisualizer screenshot).

Task-II: Conduct analytics on your dataset (40% of course project grade)

Develop Python functions that run Spark to answer the following questions (given that x, y and z) will be user-entered parameters. All of the core analysis should be conducted via Spark.

1. List the **x** players who achieved average highest improvement across all skillsets.
2. What are the **y** clubs that have largest number of players with contracts ending in 2021?
    - You may use the 2021 dataset to answer this question. Also, you may assume any other year's dataset to answer it.

3. List the **z** clubs with largest number of players in the dataset where z >= 5.
    - You may assume any year's dataset to answer this question.
    - If all teams have the same number of players, print a message on the screen indicating this.
4. What is the most frequent nation_position and team_position in the dataset? (list the most popular for each)
    - Answer this question across all datasets (2016-2022) – i.e. most popular nation_position and team_position for each year (or across all years)
    - team_position is renamed as club_position in some datasets
5. What is the most popular nationality for the players in the dataset?
    - You may answer this question for each year's dataset or across all datasets (i.e. one answer for each year's dataset or one answer for all datasets).

Task- III Test your Code and Document it (20% of course project grade)

- Document all your functions properly.
- In the Python functions you developed above, make sure to handle errors properly.
- Develop unit tests to cover ALL the functions that were developed above. Your unit tests should cover happy and sad path scenarios.
- In your Readme file, provide a summary on the scenarios you covered in your unit tests.

Task- IV (20% Extra-credit) Deploy your code to the Cloud

- Run the code for three tasks above (including the DB creation process) on the cloud.

Submission Guidelines:

- You MUST use the GitHub classroom URL to create your repository. Post your GitHub repository's URL created via GitHub classroom to Canvas. Use the starter code that is provided above as the starter for your code.
- Your GitHub repository should have a ReadMe.md file that lists the "exact" steps on how to get this application working on a new machine (via Docker). I will follow the steps in your ReadMe file and if I can't get it running on my machine, I will deduct considerable number of points from your project grade.
- You should record a video demonstrating two elements:
    1. Code Walkthrough while you are explaining your code changes.
    2. Demoing the running application while you are navigating through EVERY functionality that is working in your application. I will use this video to help assessing your grade. You may lose points for the functionalities that are not demonstrated in the demo.
- Your video size may be large to be uploaded to GitHub. You may use Box to upload the video and add the URL to your ReadMe.md file in your GitHub repository.
    1. Make sure that your video is publicly shared. Private videos won't be visible to the instructor and TAs and therefore, your project grade will be impacted

Common Penalties:

- Late submissions on Canvas or GitHub: 100% reduction (won't be graded)
- Not submitting the GitHub video (for both code walkthrough and functionality demo): 20% penalty (calculated from maximum project grade).

- Not providing clear details in the ReadMe file on how to run the application (or any variables that need to be updated/replaced): 10% penalty (calculated from maximum project grade)

Refer to Course Syllabus for planned course project checkpoints. The course project submission deadline is April 22nd, 11:59PM EST.

Suggested Project Task Schedule:

Week	Task
End of Week-5	Complete Task-I, Write the SQL needed for
End of Week-7	Complete Task-II
End of Week-10	Complete Task-III
End of Week-11	Finish your ReadMe file and the video recordings
End of Week-12	Complete Task-IV (Optional)