PACE
UNIVERSITY

Fall 2021 - CS 673: Scalable Databases

Midterm Project (Team Submission) Requirements

Total Points : 100

Due Date : 10/29/2021

Task	Points	Self-assessment
Problem Definition: What is the problem you are working on? Which dataset are you using?	10	
PowerPoint Presentation : Your mid-term presentation must have Introduction, Problem definition or hypothesis, dataset used, your EDA backed by appropriate visualizations, results, and conclusion.	30	
Python Notebook: Use of python / SQL concepts taught in the class such as functions, creating own packages, file Handling, exception Handling, use of third-party libraries. Use of different file formats like json, xml with CSV or Excel are encouraged, to demonstrate your knowledge on handling structured and unstructured data. (If worked on both structured and unstructured data you will receive extra credit : 10 points) **Essential key steps to demonstrate in your Python Notebook** 1. Loading data in to DataFrames. Integration of SQL and Python 2. Check the Data Types of your data columns. 3. Drop any NULL, missing values or unwanted columns. 4. Drop duplicate values. 5. Check for outliers using a box plot or histogram. 6. Plot features against each other using a pair plot. 7. Use a HeatMap for finding the correlation between the features(Feature to Feature). 8. Use a scatter plot to show the relationship between 2 variables. 9. Merging two Data Frames. 10. Slicing Data of a particular column value (like year, month, filter values depending on the categorical data) 11. Representing data in matrix form. 12. Upload data to Numerical Python (NumPy) 13. Select a slice or part of the data and display.	50	

14. Use conditions and segregate the data based on the condition (like show data of a feature(column) >,<,= a number) 15. Use mathematical and statistical functions using libraries. 16. Select data based on a category(categorical data based). 17. Libraries expected to try(minimum 4 required): Pandas, Numpy, Seaborn, Matplotlib . 18. Write your own functions and handle exceptions in the functions. 19. Use of *arg and **kwargs. 20. Use of data functions.		
Teamwork: Effective communication and participation with your teammates.	10	

Submission :

- Submit the files with code, data set, and a word document or PPT with the explanation.
- You are required to do a class presentation, with all team members participating. If presenting on zoom, switch on the camera.
- Late submission up to one week, 20% deduction of total points earned.
- Submit the self-assessment along with the above-mentioned files.

Important : No plagiarism, please implement your own idea and submit your own work. Your work will be checked for plagiarism.

NOTE : Extra credit for the class participation (Q&A) – 5 points.