Data Science Assignment

INSTRUCTION: Read the following questions carefully and attempt them. NOTE: Though you may discuss with your group members, but the assignment should be done independently without copying one another's work. In the case where two or more students provided exactly/closely related answers, the scores of the students involved will be affected. Submission Deadline: 18th of November, 2021.

1. By using a case-study, describe the stages in a Data Science process highlighting the core importance of each stage in the process. List two top Data Science programming languages and discuss their advantages and disadvantages possibly using real-life scenarios.

2. Data acquisition/collection represent one of the stages in of a Data Science process. Discuss at least four forms of gathering dataset when carrying out a Data Science project. Also, support your description with real-life examples, and describe any three data format that a dataset can assume per time.

3. Data cleaning/cleansing is one of the stages of Data Science process in which issues such as missing values, unit conversion, misspellings, duplicate roles, inconsistent format, and unspecified units are resolved to put the dataset in a good shape for the subsequent process. With this understand and what we have learned in class, you are to work with the "Crime Incidence Report" dataset which I handed over to you several weeks ago, performing four different data cleaning operations/tasks. Afterwards, you are expected to clearly report your results (screenshots of the outcome of the cleaning tasks and codes used). The implementation can be done using tools/programming languages that we used in class (Python, Jupiter Notebook) or other tools of your choice.

4. Briefly discuss why Data visualization is an important stage of Data Science process. Using the New York "Airbnb" dataset which I shared with you a few weeks ago, perform at least four different kind of Data visualization tasks using functions/methods from the *Matplotlib* and *Seaborn* libraries. Also, clearly report your results with screenshots of the outcome of the data cleaning tasks and their respective codes.

5. Exploratory Data Analysis (EDA) constitute an impart aspect of a Data Science project. You are required to give a concise but informative description of the "FIFA World Cup" dataset handed to you in the previous lecture and carryout the following EDA tasks: (a) Find the summary statistics of the data and provide a brief explanation of the statistics for any (single) variable in the dataset, (b) Find the mean, median, and mode of the variable "Qualified teams", (c) Generate an histogram plot for the variable "Winner" to understand the number of times a country won the world cup between 1930 to 2014, (d) Find and visualize the correlation relationship among the numeric variables (Year, GoalsScored, QualifiedTeams, MatchesPlayed) in the data using "heatmap" plot, (e) Plot the pairplots of 'GoalsScored', 'QualifiedTeams', 'MatchesPlayed' variables and summarize the information in the plot. Note: Clearly report your results with screenshots of the outcome of the EDA tasks and the respective codes.

Note: Assignment report submission deadline: 18th of November, 2021.

Dr. SAMUEL O.W. (samuel@siat.ac.cn & timitex92@gmail.com)--Course: Data Science
Fall Semester-2021/2022 Academic Year
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China.