Chapter 4 Assignment - Python Insight Generation

## Instructions

1. You can take help from the lecture notes to revise the concepts that we have covered
2. You have been provided a google sheet named “Top 2000 Universities of the World”, this is your dataset for this assignment.
3. For these questions, you need to work on the Google Colaboratory which is prepared according to your questions. And you have to write a short summary of answers in this word document.
4. To get started with the assignment, you need to make a copy of the Google Colaboratory at [https://colab.research.google.com/drive/1pViSIc4Y9J20IpsCP63gWyfo4XnTC2B5](https://colab.research.google.com/drive/1pViSIc4Y9J20IpsCP63gWyfo4XnTC2B5?authuser=1#scrollTo=cEqxXzTd3u5n)
5. Each question’s answer cell should have the right formulas and calculations. Your python code are also graded.
6. Please submit the assignment through TalentLabs Learning System. You will need to submit this word document and Google Colaboratory link. **Make sure that the notebook is already executed and with the expected answers/results printed**. In order for your mentors to grade it properly, please make sure that you setup view access for the public.

**Please Insert the Link to your Google Colaboratory Notebook here (Make sure the link is accessible)**

|  |
| --- |
| <https://colab.research.google.com/drive/1-8Jrh0zdC_3P6w-TqJK0LgLbvnF8bZq3?usp=sharing> |

**Question 1 (2 points):**

If we recall the flowchart of data analysis, we know that data analysis requires data preprocessing. Although we have covered data pre-processing in detail previously, it is always a good practice to verify your data preprocessing before you start your analysis.

You have been provided the dataset of “Top 2000 Universities of the World”, after having a look at the dataset, verify if you have any null or NA values on your dataset and drop the rows accordingly to make your data ready for analysis.

|  |
| --- |
| Count of Null / NA values:0  (Note: below is how we check the existence of missing values in the data set.) |

**Question 2 (6 points):**

When we jump on working with the data sets in Data Analysis, after verifying the data for missing values and formulating our research questions. We always observe our dataset, describe it and check the dimensions of the dataset.

Data description provides us a quick view of the data columns and the recorded values in our dataset. We need to make a quick view of our dataset to see how many rows, columns we have in our dataset, and how the data is distributed.

Please do a quick study of the data using the describe function in pandas and answer the following:

|  |
| --- |
| **Size / Shape of Dataset:**  Number of Rows: 2000  Number of Columns: 9  Number of Categorical Columns: 2  Number of Numerical Columns: 7  Attach a description of your dataset (the output from the describe() function) : |

**Question 3 (12 points):**

As a data analyst, after having the initial description of the dataset, you need to understand the columns in your dataset so that you can choose your data aggregation and data summary strategies accordingly for generating insights from your dataset.

In this question, you are required do some analysis for 4 of the data columns, and put down

1. type (Numerical, Categorical) of each column,
2. the description of each column, and
3. values in each column (list out 3 sample values for categorical column, and the range of value for numerical values)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Type | Numerical / Categorical | Description | Values |
| Country | String | Categorical | Country of university location | "USA", "United Kingdom", "Brazil" |
| Institution | String | Categorical | University Institution name | "Harvard University", "Stanford University", "University of Cambridge" |
| World Rank | Integer | Numerical | University's World Rank | 1, 2, 3 |
| Score | Float | Numerical | University score | 100, 96.7, 95.1 |

**Question 4 (4 points):**

Imagine you are a student who wants to get into the best university. This dataset provides you information about the Universities and their world ranking of different areas (e.g. Employment, Research etc.).

Let’s say you want to find a university that has good alumni employment opportunities. From the dataset given,

1. List out the top 10 universities based on employment ability.
2. Also, list out the countries where these universities come from.

(You can take help from the data aggregation strategies such as sorting, filtering etc)

|  |
| --- |
| Top 10 Universities:    Distinct Countries Count: 5  Distinct Countries name; 'USA', 'France', 'Japan', 'China', 'Switzerland' |

**Question 5 (5 points):**

While performing Data Analysis, we might not need to consider all the rows and columns. For example, if we want to have a high level overview of the universities in the USA based on their world rank and overall scores, we can make a subset of data having World Rank, Score and Institution only. We can also further filter out universities that are not in the USA.

To achieve that, please

1. Create a subset data frame using filters on above mentioned columns
2. And from this subset, filter out the Universities which belong to the **USA (the output data frame should only contain universities from the USA)**.

Then, answer the following:

|  |
| --- |
| Mention the name of top 5 and bottom 5 universities of USA:  Top 5:    Bottom 5: |

**Question 6 (10 points):**

Next, we would like to analyse which countries are having more high ranking and high quality universities and which countries have less.

In order to generate these insights, you are required to make a summarised data frame with all the countries and their average score of the universities. (Hint: consider using the “groupby” function)

After creating the data frame, answer the following:

|  |
| --- |
| The average score of   1. Ireland: 72.02 2. United Kingdom: 73.63 3. Pakistan: 68.38 4. Germany: 74.47   Two lowest performing countries:   1. Zambia: 67.40 2. Zimbabwe: 66.70 |

**Question 7 (12 points):**

We have studied the Ranges, Quartiles and Interquartile range in detail. We can use it to help us in finding the outliers in the university dataset (i.e. those extremely good or bad universities).

In order to perform this analysis, we will use the 1.5IQR method that we covered in class. If you forgot the method, please refer to the video.

|  |
| --- |
| 1. Identification of Column of interest (1 point):  Using “Score” column for outlier detection.  2. Minimum and Maximum of column used in Analysis: (1 point)  Minimum value of “Score” column: 65.7  Maximum value of “Score” column: 100    3. Q1, Q3 and IQR of the column (2 points):  Q1 = 67.7  Q3 = 74.1  IQR = 6.40  4. Write down the lower and upper expected minimum and maximum of IQR (3 point):  Lower expected min of IQR: 58.10  Upper expected max of IQR: 83.70  5. Number of outliers identified (1 point)  63 rows of outliers identified.  6. Any insights that you can conclude from this analysis (1 point)  Using the 1.5IQR method for outlier detection, we have identified 63 rows as outliers. These universities have "Score" values that fall below 58.10 or above 83.70.  These outliers may represent universities that are either exceptionally good or exceptionally bad compared to the majority of universities in the dataset, based on their "Score" values.  (extra 3 points for correct python code for solving this problem) |