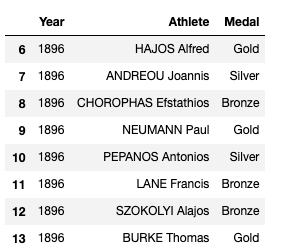
**Assignment #10**

**Course: CPSC 442 (Python For Data Science)**

**Part 1 (Use Olympics Medal Dataset)**

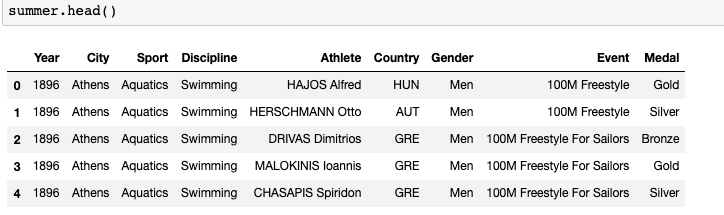
1. Load all three datasets into separate Dataframes. Keep the Dataframe variable name as *summer*, *winter*, *country\_codes.*
2. Display a subset of *Summer* Dataframe with 8 rows (after 6th index) with columns *Year*, *Athlete* and *Medal*

**Expected output:**



1. Replace all commas with space in the *'Athlete'* column of *Summer* and *winter* Dataframe.

**Expected output:**



1. Modify *Summer* Dataframe by merging *Summer* Dataframe with *country\_codes* Dataframe based on the country code.
2. Modify *Winter* dataframe by merging *Winter* Dataframe with *country\_codes* Dataframe based on the country code.
3. Create a function to find the male athlete & female *Athlete* who won the highest number of medals. You should pass the *Summer/Winter* Dataframe object to the function. Also pass the gender to the function. The function should return the name of the male/female *athlete* who won the highest number of medals.

**Function Definition:**

1. Create a function to find the *athlete* who won the highest number of medals in each medal category (Gold, Bronze, Silver).
2. Calculate the number of medals by year and column in a pivot table for Summer and Winter Dataframe. [Use count as aggfunc]
3. Extract a series with total number of medals won by each country in *Summer* and *Winter*.
4. Combine *Summer* and *Winter* Olympics data and create a bar chart showing the top 10 highest medal winning countries.

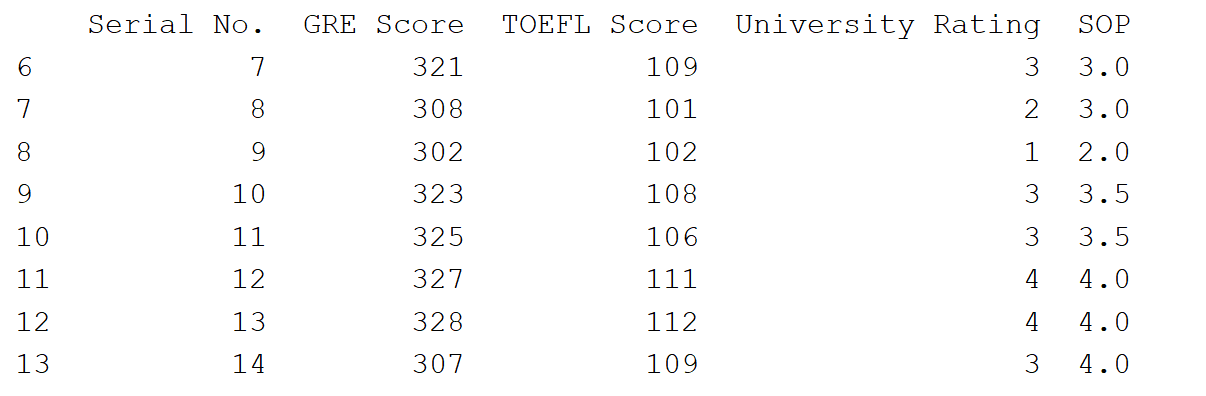
**Part 2 (Use Google Play store dataset):**

1. Import the dataset into data frame and display first and last 10 rows
2. Find out the total number of unique apps in the dataset.
3. Display all the App Categories found in the dataset.
4. Find out if there are any duplicate entries of data. Ie., Find out if any apps have more than one entry in the dataset. List out the apps which has more than one entry in the dataset.[Hint: Use Pandas’ ‘duplicated’ function]
5. Remove the duplicate records in the dataset. Remove the duplicates entries except the first entry [Hint: You can use drop\_duplicates function for removing duplicate entries].
6. Find out if there any NA values in each column. List out the total number of NaN found in each column
7. Drop the rows which has NaN value and display the shape of the dataframe after cleaning the dataset. [Hint: Use Pandas dropna function]

**Part 3(Use Admissions Dataset)**

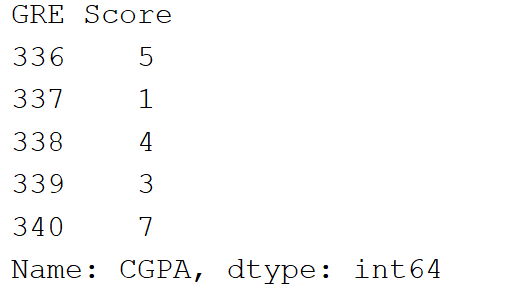
1. From the given dataset, display a subset of data within a certain range. For eg: Display a subset with 8 rows (after the 6th index) and 5 columns.

**Expected Output:**

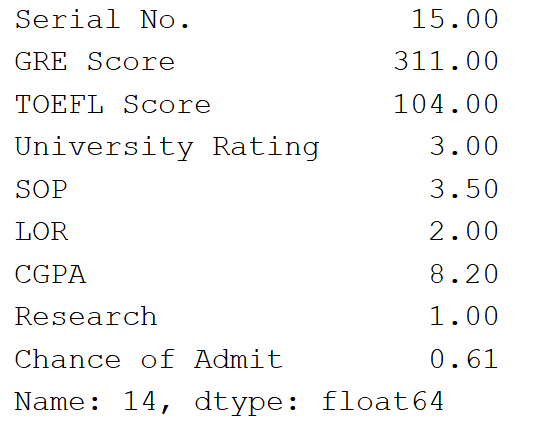


1. Use the given dataset,

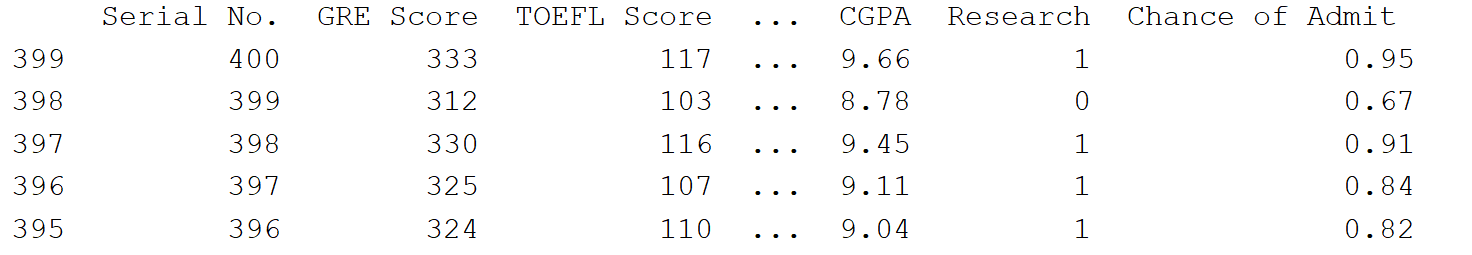
* Find 5 highest unique GRE Score with respect to their respective CGPA.



* Slice the dataframe that has all the admissions data to get the 15th row.



* Find the students with GRE Score above 300, sort them in descending order and then display the data of 5 students with the highest GRE Score.



* Demonstrate the usage of pickle. Create a binary file using pickle and read the stored binary file using pickle.

1. In the given dataset,

* Test for missing values
* If missing values are present in the first 10 rows and 6 columns, fill it with 0.

1. Write a function to find the square of the ‘University Rating’.
2. Use the given dataset,

* To find the serial numbers of top 5 rows with a University Rating = 4.
* Use the numpy mean method to display the mean values of GRE Scores.
* Write a function to find the z-score using the ‘TOEFL Score’ column. Display only the Research field along with the output of the function.
* Use the lambda function to filter the ‘GRE Score’ columns that have a count value = 10.

1. For the given dataset,

* Create 2 clusters by using a random seed of 30 in KMeans algorithm.
* Reduce dimensions using PCA (Use the result of KMeans to find PCA1 and PCA2).

Display the last few data in the clusters with their respective PCA1 and PCA2.

**Part 4 (Use wine dataset):**

1. Use the wine dataset to solve the below mentioned problems using the concept of Univariate plotting

* To find the wines that are from the US along with their points on a histogram.
* To sort the points and represent them on an area graph.

1. Use the wine dataset to solve the below mentioned problems using the concept of Bivariate plotting

* Plot a hexplot graph to display wine data that have the prices below a certain value.
* Plot a scatterplot to represent the wine points for the wines that have their prices below 100$.

1. Use the wine dataset

* To group the data with respect to the ‘points’ column.
* To group the data by ‘points’ in the ascending order of ‘price’.
* To group ‘country’ on the basis of length, minimum and maximum.

1. Use the wine dataset

* To display a wordcloud of the wine description.
* To display a wordcloud of the wine description of the wines from US.

1. Use the wine dataset

* To find points, variety and province of the wines, where points>90. Remove duplicates and ignore the noisy data.
* To obtain a model using knn algorithm.