Intermediate python for data science

TTT « PERSONAL WORK »

First Name : Sara

Family Name : El-Metwally

Objectives:

You have to choose two Exploratory Data Analysis (EDA) projects with different levels of difficulty (*Low, Medium or High*). Your goal is to design an EDA so that the students knows:

- **Select a real-world dataset**
- **Ask & answer questions about the data**
- **Perform data preparation & cleaning**
- **Perform exploratory analysis & visualizationn**
- **Interpret their results**
- **Summarize their inferences & write a conclusion**

For each of these two problems, you are asked to give:

1. An abstract
2. A description
3. And your "evaluation" of the problem (see below)
4. You can also give, if necessary, a Jupyter Notebook.

Ps: Don't forget to validate your submissions on Teams. Please use this file for your answers.

Problem 1

1.1 Abstract (minimum 100 characters & maximum 300 characters[1]):

K-means algorithm is one of clustering techniques that used to cluster the data into k clusters. It iteratively assigns the data item to its nearest cluster centroid. By analogy, if you have an image with many colored pixels, you can reduce the number of colors in an image by assigning the pixels with similar colors or closely related ones the same cluster. The number of generated clusters will correspond to the number of reduced colors in the image space.

1.2 Description[2] (minimum 2000 characters & maximum 3000 characters):

Each pixel in the image space has three coordinates which as RED, GREEN, BLUE. K-means clustering algorithm can be used in this color space to find the reduced number of colors an image can be

[1] We will not count blank characters on all answers.

[2] You can add pictures of course.

converted to. The RGB values color space has a range of 0-255 colors and this range could be reduced to i.e. 16 colors channel. The number of reduced colors channel will be corresponding to the total number of resulted clusters. The first step is to choose the image that you would like to play with, loaded with the python IDE and display it before any changes occurred to the image. Then, we need to ask how many pixels already in our image and how many RGB channels are there. This will be done by a shape property from a numpy library. The pixels' normalization step will be occurred as a pre-processing step where the colors values in the range from (0-255) will be converted into the range (0.0-1.0) and the integer values will casted to float ones in order to do math more precisely and simply. The 3D dimensional image space will be reshaped into a 2D dimensional array. The k-means clustering algorithm will be setup and the number of reduced color space will be identified. The new image will be reshaped and plotted next to the original one. We will make a scatter plot in order to visualize the color space before and after the compression step.

1.3 Your Evaluation of this problem (minimum 500 characters & maximum 1000 characters):

1. Level of difficulty of this problem? This problem from my point of view is an intermediate level problem and can explain the practical idea of k-means clustering based on simple pre-processing steps of an image pixels data.
2. Interest of this problem? Very interesting since it is real practical example of image compression techniques.
3. Other Remarks ?
4. References ?
   1- https://github.com/SaraEl-Metwally/TOT-ITI-EPITA/blob/main/SaraEl-Metwally_kmeans.ipynb

Problem 2 :

2.1 Abstract (minimum 100 characters & maximum 300 characters[3]):

Covid-19 is a global pandemic issue and in order to track these pandemic in different counties, we will build a tracker that displays some plots to gain insights about each country status. Usually, tracker projects will collect the publish data, analyze it and make some insightful plots. In this demo, we will use the published data for USA different states in order to make conclusions on different states test cases, deaths, and increasing the number of affected peoples in each state.

2.2 Description (minimum 2000 characters & maximum 3000 characters):

The first step is collecting our data from the website (https://covidtracking.com/), and load it as a data frame in order to start pre-processing it. We will print the data frame information in order to explore our data to know our columns names, data types, and all NAN values will be replaced by -1. The unnecessary columns are dropped and the bar and scatter charts will be plotted in order to explore each USA sate individually. The confirmed cases in each state can be collected by tracking the individual tests. Some ratios will be computed such as fatality ratio, hospitalization ratio, and positive case ratio. Finally, bubble chart can be used to compare different states together.

[3] We will not count blank characters on all answers.

2.3 Your Evaluation of this problem (minimum 500 characters & maximum 1000 characters):

1. Level of difficulty of this problem ? difficult problem.
2. Interest of this problem ? very interesting and it is a real case scenario for data analysis and visualization.
3. Other Remarks ? [...]
4. References ?
   1- https://towardsdatascience.com/data-exploration-with-the-covid-tracking-project-d89ac87342bc
   2- https://github.com/SaraEl-Metwally/TOT-ITI-EPITA/blob/main/Data_Visualization_Task.ipynb

Your Personal Comments

You can write here any personal comments (everything will be confidential) on my slides, on my notebooks, etc.
Any idea, any comment is welcome.