Challenge 5

Honeypots are mechanisms through which unauthorized use of information systems could be detected by means of one or more computer systems intended to mimic likely targets of cyberattacks.

In this challenge, you are provided with a dataset of 451,581 records. Each record is a cyber attack (or attempt) and its features, collected from March to September 2013 from Amazon Web Services (AWS) honeypots. The features are summarized as below:

Features of **AWS_honeypot_geo** dataset:

- **host**: The host name
- **src**: The source identifier
- **proto**: The communication protocol
- **spt**: The source port
- **dpt**: The destination port
- **srcstr**: The source IP address
- **country**: The source country
- **locale**: Locale (geographical location)
- **postalcode**: Postal code
- **latitude**: Latitude of the source
- **longitude**: Longitude of the source

Prepare a report and submit a PDF file by Tuesday (**11/10/2020 before 6 pm**) considering the below details:

1. Apply the k-means clustering algorithm to the dataset. To obtain the optimal number of clusters (k), rely on silhouette analysis. Thus, calculate and visualize the silhouette coefficients of different clusters for $k \in [2, 6]$. Then, report and discuss what is the optimal value of k. Once you obtained the optimal value, run the k-means clustering algorithm once again and save the result (i.e., cluster labels).

2. Apply the DBSCAN clustering algorithm (eps = 0.4, min_samples = 10) to the same dataset and save the result.
3. Visualize the clustering results of steps 1 and 2 for latitude and longitude features (i.e., only two columns). For this step, you're required to insert two figures in your report that show how data points are clustered in k-means and DBSCAN algorithms.