In HW4 we will be working with a dataset called "Dry_Bean" Which will be uploaded in our Telegram Channel, so make sure to download it and have it ready for this homework.

Our objective is to load this dataset and perform simple statistical analysis on it, as well as apply a well-known clustering algorithm called Lloyd's algorithm. We will also conduct some basics of exploratory data analysis (EDA) procedure on this dataset. Throughout this project, we have several tasks to accomplish. Let's go through them:

Task 1: Calculate and display the descriptive statistics of the given data set. (Note that your output should be a dataframe.)

Note: by descriptive statistics we mean the following:

- Mean
- Standard deviation
- Minimum
- 25th percentile (Q1)
- Median (50th percentile)
- 75th percentile (Q3)
- Maximum
- Count (number of non-missing values)
- Sum

Task 2: Calculate the correlation matrix between features in the given dataset and identify the features with a correlation coefficient greater than 0.8 then display the selected features in a data frame where each of its entries is the correlation coefficient between those pairs features.

Task 3: Use the Seaborn module to plot a heatmap of correlation matrix for the given dataset.

Hint: Create a correlation matrix dataframe according to Task 2, then create the heatmap plot of the correlation matrix.

Task 4: Transform the categorical column (last column) of the given dataset to numerical values and fit the Lloyd algorithm with k clusters using a random seed of 42 and then display the plot of final clusters with different arbitrary colors.

Instructions:

1. Identify the categorical columns that need to be transformed into numerical values.
2. For each categorical column, use the appropriate encoding technique (e.g., One-Hot Encoding) to convert them to numerical values.
3. Set the random seed to 42 using the numpy library's proper function. This ensures the reproducibility of results as the Lloyd algorithm is random-based.
4. Fit the Lloyd algorithm to the transformed dataset and Initialize an instance of the Lloyd algorithm with the desired number of clusters (k).
5. plot your results.

Note: The Lloyd algorithm is a variation of the K-means clustering. It is a randomized algorithm, meaning that different initializations can lead to different clustering results. By setting a random seed, we can ensure that the algorithm behaves consistently across runs for easier comparison and reproducibility.

Hint: One-hot encoding is a technique used to convert categorical variables into a binary vector format. It creates new binary columns for each unique category in the original variable, representing the presence or absence of that category. This technique is useful when dealing with categorical data that cannot be directly used in mathematical models.

To use one-hot encoding, you can Look for functions or classes specifically designed for one-hot encodings, such as get_dummies () in the pandas or OneHotEncoder () in the scikit-learn. These functions can take categorical columns as input and transform them into numerical columns using one-hot encoding.

Good Luck