← Go Back to Elective Project

Course Content

# Project FAQ - Data Analysis and Visualization

1. **How to install the scikit-learn-extra library in my working environment?**

   Please execute the below command in the Jupyter Notebook to install scikit-learn-extra library

   ```
   !pip install scikit-learn-extra
   ```

   In case you face any issues with the installation, please make sure the version of Python is greater than or equal to 3.6 and scikit-learn is equal to 0.24. You can check the Python and scikit-learn version using the below codes:

   ```
   import sys
   print(sys.version)
   import sklearn
   print(sklearn.__version__)
   ```

   In case the Python version is lower than 3.6, run the below code in **Anaconda Prompt (or terminal for Mac OS)** to update Python. Run the below command in Anaconda prompt

   ```
   conda update python
   ```

   In case the scikit-learn library version is not 0.24, please update the scikit-learn library to the specific version by using the below command in the **Jupyter Notebook**.

   ```
   !pip install scikit-learn==0.24
   ```

   Once the above conditions are met, please execute the below command in the **Jupyter Notebook**

   ```
   !pip install scikit-learn-extra
   ```

   After making all the above changes, you can **restart** the jupyter notebook and try to import the scikit-learn-extra library.

2. **When I try to load the CSV dataset file using the pandas read() method, I'm getting this error -** ***UnicodeDecodeError: 'utf-8' codec can't decode bytes in position 15-16: invalid continuation byte.*** **How to resolve this error?**

   The error that you are getting is "UnicodeDecodeError". When we are reading a file due to extra spaces, and slashes in the file path, we might get this kind of error. It can be resolved by adding r to the starting of the file path which helps to convert the file path into a raw string that helps to resolve this error. To do the same, please refer to the below line of code.

```
df = pd.read_csv(r'File_name.csv', encoding = 'latin-1') #replace File_name.csv with your filepa
```

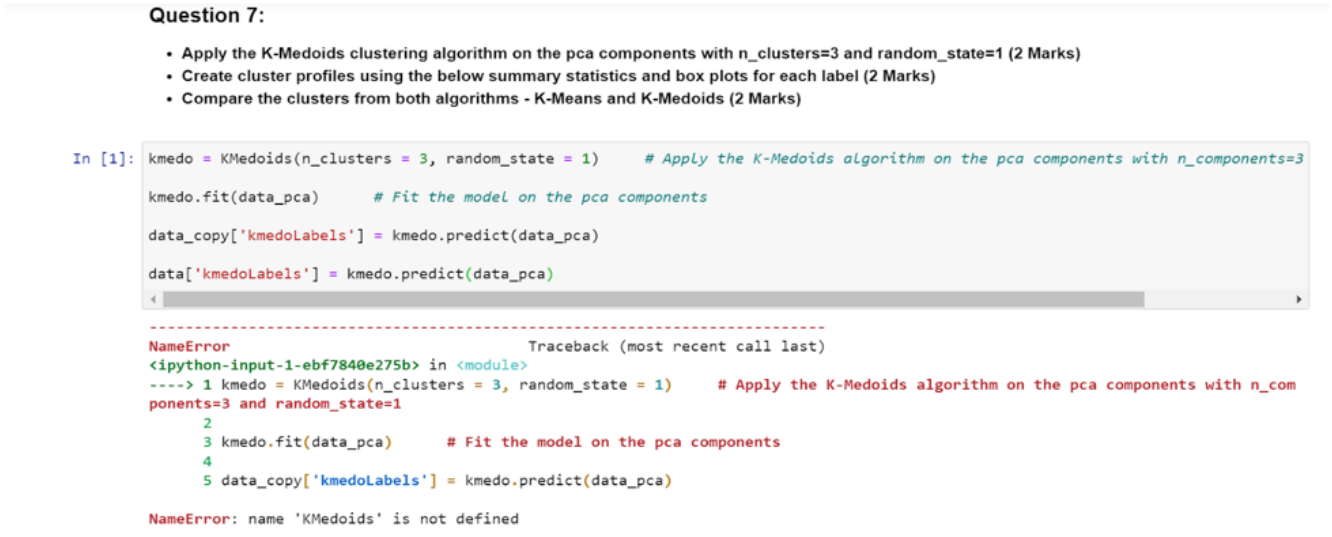
3. **How can we drop the duplicate keys in question 1 of project part 2?**

   To drop the duplicate data from the customer key feature present in the data. First, we need to find the duplicate elements of the Customer Key column present in the data using the duplicated() function that is available in pandas and store the results into a separate variable. Then extract the rows which are not duplicates by using the ~ operator.

4. **How to solve the error "ModuleNotFoundError: No module named 'sklearn_extra"?**

   The error that you are getting is due to the sklearn_extra library not being installed in your working environment. To import any module, it should be installed first to use it. To install the sklearn_extra library, please refer to question 1 of this FAQ page.

5. **I got an error when calling the K-Medoids function, which I think is related to the installation of the package (see image).**




**How can I solve this?**

The error is due to K-Medoids not being imported into your working environment. Please import the K-Medoids from the sklearn_extra library, please refer to the below code:

```
from sklearn_extra.cluster import KMedoids
```

Note: If the scikit learn extra is not installed on your system, please install it. To do that please refer to question 1 of this FAQ page.

6. **When I run the following code in the project, I get an error - *"name 'pca' is not defined".* How to fix it?**

   In your code, pca is not defined to use it. To use any library functions, first, it should be imported into your working environment. Please define or import the PCA library in your notebook as shown below:

   ```
   from sklearn.decomposition import PCA as pca
   ```

7. **I'm trying to do principal component analysis on the dataset, but whenever I want to apply pca.transform from the sklearn.decomposition module I keep getting this error: **AttributeError: 'PCA' object has no attribute 'mean_'*. How to fix it?**

   You should have to fit the PCA on the data before accessing its attributes like mean_, explained_variance_mean, etc. To overcome the issue, please fit the pca on the data after creating an instance for the pca. After fitting the pca on the data, now try to access the attribute values that you desired to access.

8. **How to resolve this error - *AttributeError: 'pca' object has no attribute 'explained_variance_ratio_'* ?**

   The problem is you do not need to pass your parameters through the PCA algorithm. To access the model attributes, we need to access them through the dot(.) operator. It can be resolved by adding .explained_variance_ratio_ to the end of the variable that you assigned the PCA.

   For example:

   ```
   pca = PCA(n_components = 2).fit_transform(df_transform)
   var_exp = pca.explained_variance_ratio_  # Add 'pca' before 'explained_variance_ratio_'
   ```

9. **While creating the scatterplot in question 5 of the DAV project part 1, we are receiving a key error, how can we resolve it?**

   In Pandas, Data Frame will be created with the numbers as the column names when we haven't specified the column names. It is a special case, and when we are accessing the columns from that data frame we don't need to access it like df[0], it should be enough to access it with direct numerical numbers, i.e., 0.

10. **I am receiving an error that just says *KeyError: 'mpg' not found in the axis. H*ow to resolve it?**

    It happens when the accessed key is not present in the data frame. To resolve this error, you need to pass the key that should be present in the data frame. When accessing any column, we should need to pass the column name that should exactly matches the case of the column name that is present in the data frame. Please kindly access the mpg variable from the data as how it is defined in the data frame.

11. **When I remove duplicates, the data shape is (649, 5) but your text indicates it should be (644, 5), What was happening here?**

    Running duplicated method on the entire dataset will search for the duplicate records for the whole dataset. You need to run that on the customer key column, then you can get the duplicate customer keys.

    The below code will be helpful to implement the above suggestion in python.

```
# Getting duplicates for the Customer Key rows
duplicate_keys = data['Customer Key'].duplicated()
# Extracting the duplicate rows using the duplicate_keys variable
data[duplicate_keys]
```

12. **I wonder why we need to install the warning function into the code. Shouldn't we keep the warning so they tell us what we did in the program?**

    Generally, we don't want to see the warnings The warnings in Python are raised when some outdated class, function, keyword, etc., are used. These are not like errors. When an error occurs in a program, the program terminates. If we don't ignore the warnings sometimes they take up a large space of up to a few pages and the coding becomes difficult as we need to scroll up and down. However, you can also feel free to remove the specific piece of code and continue. There is a significant difference between Warnings and Errors. If we are facing an error like an attribute error that needs to be fixed before we progress. The example of missing a parenthesis that needs to be fixed whereas warnings can be ignored and will not stop the execution of the code.

‹ Previous

Next ›

Proprietary content.©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

© 2023 All rights reserved.

Help