Step by Step solution for Assignment:

Problem Statement:

In our assignment, we have two datasets. One is **Raw_Skills_Dataset.csv** which contains a lot of jargon mixed in and another one is **Example_Technical_Skills.csv** which contains random examples of technical skills. We have to clean the **Raw_Skills_Dataset.csv** and extract technical skills.

<u>Step→1:</u>

In First step, I imported necessary python libraries and read the Raw_Skills_Dataset.csv file.

Importing necessary modules & dataset

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

Reading the dataset1: Raw_Skills_Dataset

```
[60] data1 = pd.read_csv('/content/drive/MyDrive/data/Raw_Skills_Dataset.csv')
```

Then we take a look at the data.

```
RAW DATA

O What ifs

seniority

familiarity

functionalities

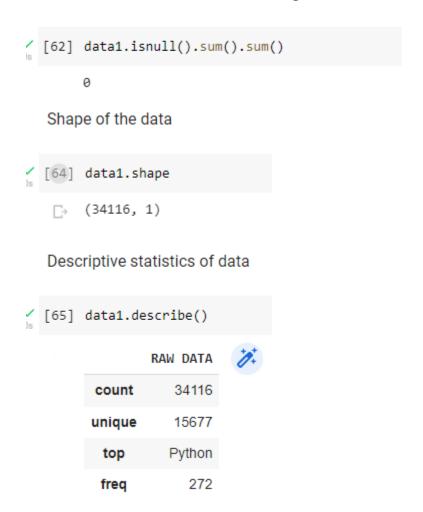
Lambdas
```

In our dataset, there is many unnecessary data those are not technical skills and we have to remove them.

<u>Step→2:</u>

In second step, as a part of preprocessing we check for missing data, shape and descriptive statistics of the data:

Lets check whether there is missing values or not

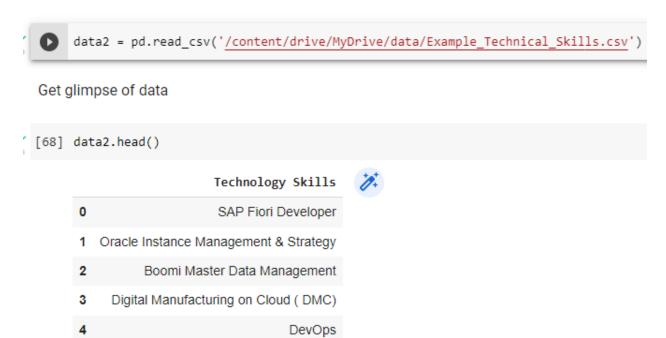


There are no missing values in the data.

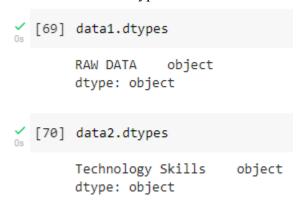
<u>Step→3:</u>

Then we read the second dataset and explore that carefully.

Reading the dataset2: Example_Techical_skills



And we check data types of both dataset's data.



<u>Step→4:</u>

In step→4, we make the skills value of both dataset in lowercase because we have to remove irrelevant data from the first dataset on the basis of the second dataset.

Lowercasing the data

```
[71] data1["RAW DATA"] = data1["RAW DATA"].str.lower()
    data2["Technology Skills"] = data2["Technology Skills"].str.lower()
```

Then we just store those values(skills) in our first dataset that get matched with second datasets skills, otherwise remove them.

Now our Final Task is we have to remove those skills that are not available in our Technology Skills

```
[74] val = data2['Technology Skills'].values

[75] val = list(val)

[76] data1 = data1[data1['RAW DATA'].isin(val)]

[77] data1.head()

RAW DATA

62 mysql

111 github

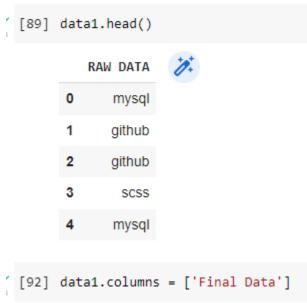
154 github

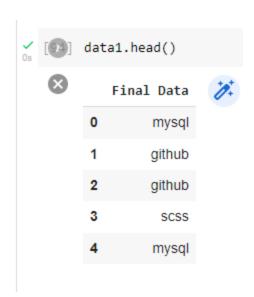
353 scss

354 mysql
```

Then we reset our index and our final dataset becomes:

```
[84] data1 = data1.reset_index()
 (85] data1.head()
            index RAW DATA
         0
               62
                      mysql
         1
              111
                      github
         2
              154
                      github
         3
              353
                       SCSS
              354
                      mysql
(88] data1.drop('index',axis=1,inplace=True)
  So my final dataset is
```





[Note: For more, check out my source code file that I have attached with Email]