109550198 卜銳凱

HW2
Report

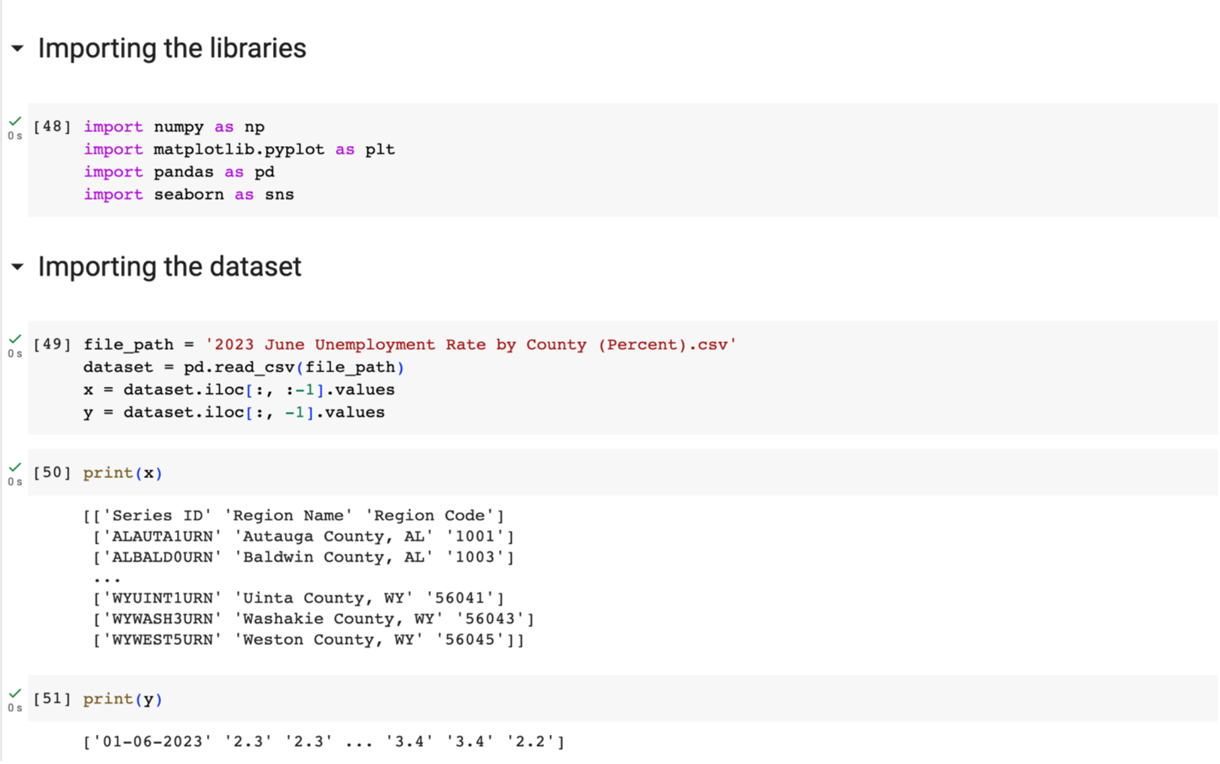
This report describes the analysis of the dataset from the CSV file "2023 June Unemployment Rate by County (Percent).csv." The basic goal of the code is to load the data, do preliminary data exploration, and illustrate the presence of missing values.

Code Summary:
The provided code accomplishes the following tasks:

Data Loading: The code reads the dataset from the CSV file and saves it in a Pandas Data Frame with the name dataset.

Data Splitting:

It divides the dataset into two arrays, x and y, where x contains all except the last column and y contains the last column. This is commonly done to differentiate between features (independent variables) and the desired variable (dependent variable).



Importing the libraries

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

Importing the dataset

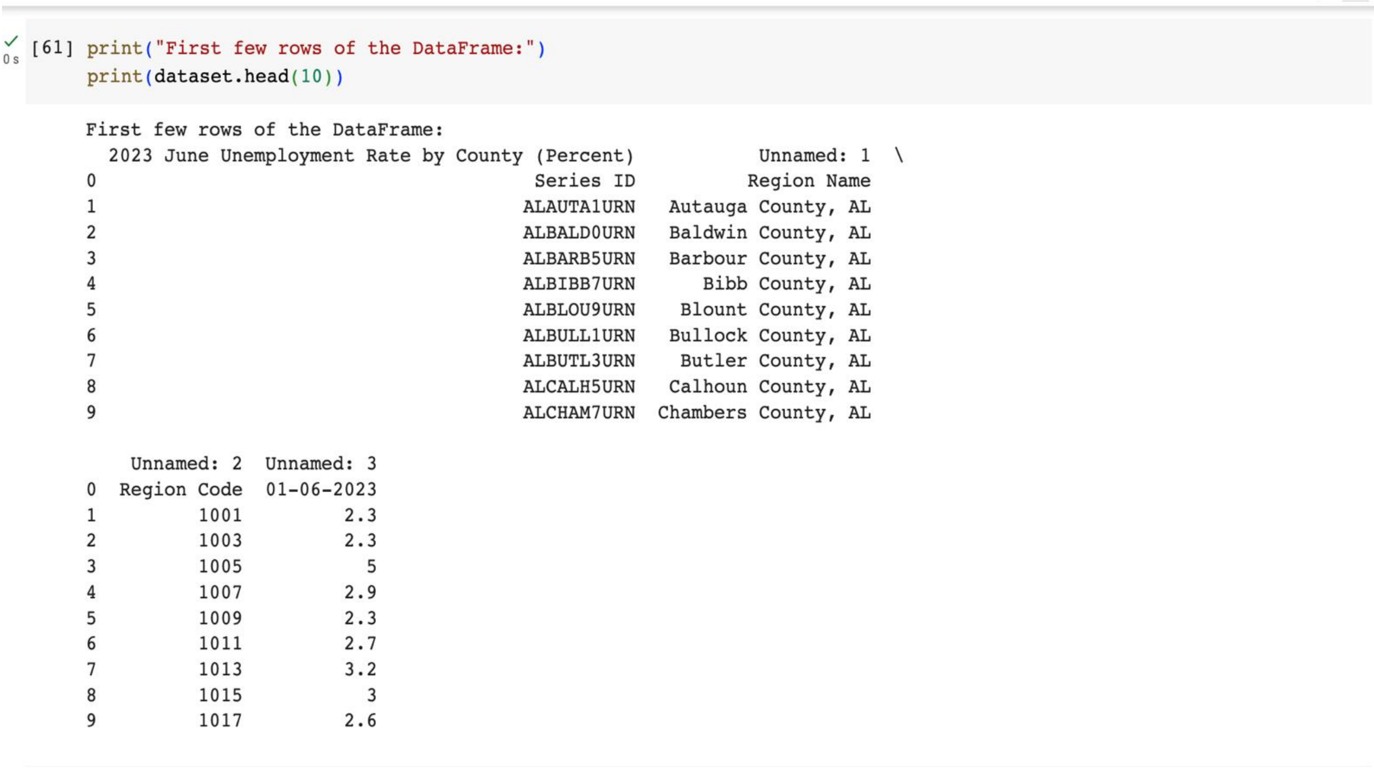
```
[49] file_path = '2023 June Unemployment Rate by County (Percent).csv'
     dataset = pd.read_csv(file_path)
     x = dataset.iloc[:, :-1].values
     y = dataset.iloc[:, -1].values
```

```
[50] print(x)

     [['Series ID' 'Region Name' 'Region Code']
      ['ALAUTA1URN' 'Autauga County, AL' '1001']
      ['ALBALD0URN' 'Baldwin County, AL' '1003']
      ...
      ['WYUINT1URN' 'Uinta County, WY' '56041']
      ['WYWASH3URN' 'Washakie County, WY' '56043']
      ['WYWEST5URN' 'Weston County, WY' '56045']]
```

```
[51] print(y)

     ['01-06-2023' '2.3' '2.3' ... '3.4' '3.4' '2.2']
```

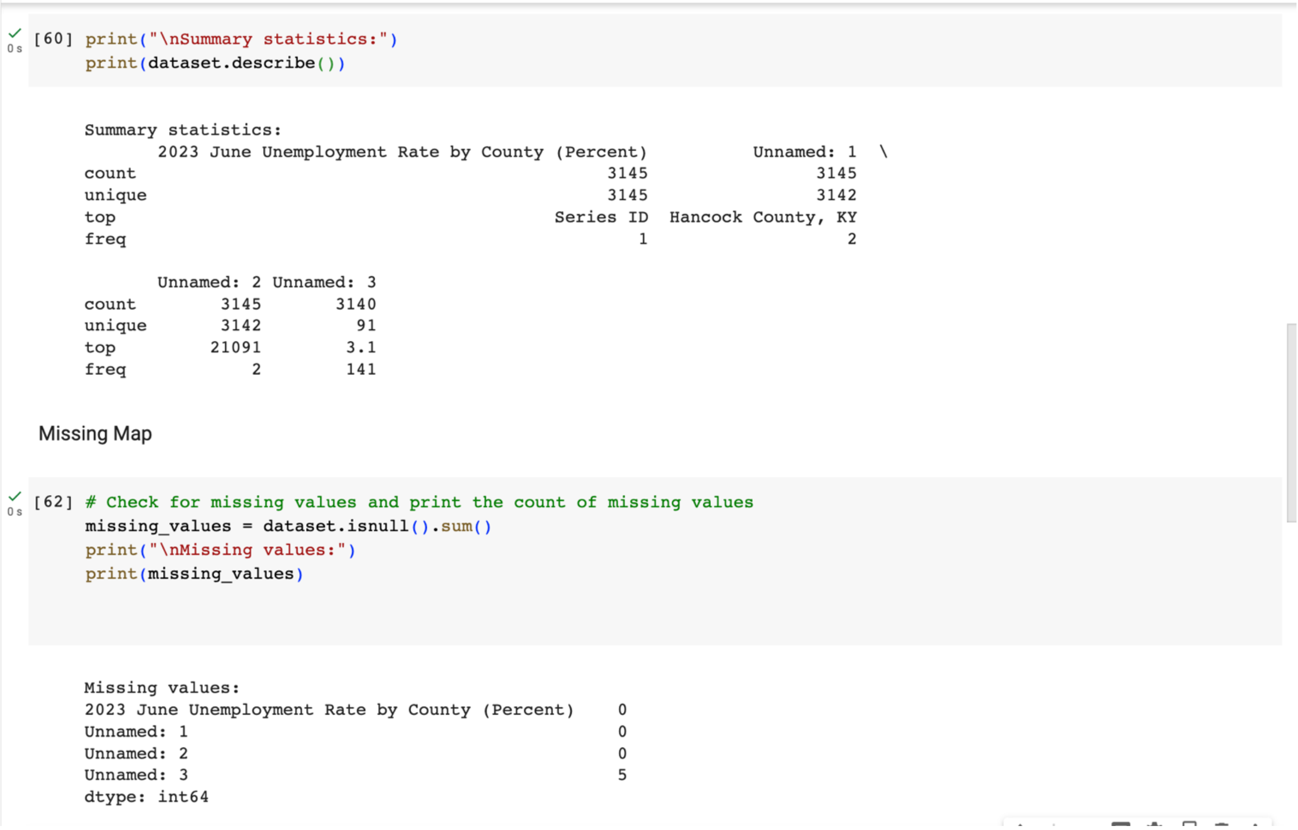
Data Exploration:

It shows the first ten rows of the Data Frame, providing an overview of the data's structure and content.
It computes and displays summary statistics, providing insights into the central patterns and spreads of numeric columns.
It discovers and counts missing values in the dataset and visualizes them using a heatmap.



```
[61] print("First few rows of the DataFrame:")
     print(dataset.head(10))
```

```
First few rows of the DataFrame:
  2023 June Unemployment Rate by County (Percent)           Unnamed: 1  \
0                                       Series ID          Region Name   
1                                      ALAUTA1URN   Autauga County, AL   
2                                      ALBALD0URN   Baldwin County, AL   
3                                      ALBARB5URN   Barbour County, AL   
4                                      ALBIBB7URN      Bibb County, AL   
5                                      ALBLOU9URN    Blount County, AL   
6                                      ALBULL1URN   Bullock County, AL   
7                                      ALBUTL3URN    Butler County, AL   
8                                      ALCALH5URN   Calhoun County, AL   
9                                      ALCHAM7URN  Chambers County, AL   

    Unnamed: 2  Unnamed: 3  
0  Region Code  01-06-2023  
1         1001         2.3  
2         1003         2.3  
3         1005           5  
4         1007         2.9  
5         1009         2.3  
6         1011         2.7  
7         1013         3.2  
8         1015           3  
9         1017         2.6  
```



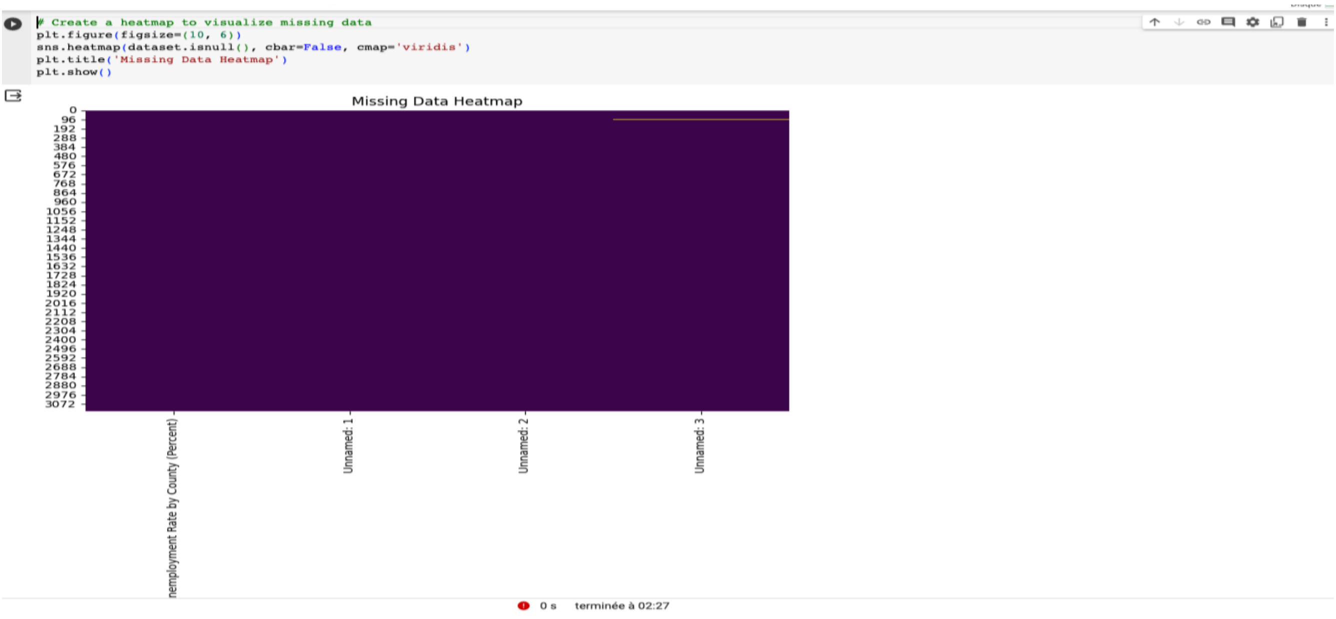
```
[60] print("\nSummary statistics:")
     print(dataset.describe())
```

```
Summary statistics:
       2023 June Unemployment Rate by County (Percent)          Unnamed: 1  \
count                                             3145                3145
unique                                            3145                3142
top                                          Series ID  Hancock County, KY
freq                                                 1                   2

       Unnamed: 2 Unnamed: 3
count        3145       3140
unique       3142         91
top         21091        3.1
freq            2        141
```

Missing Map

```
[62] # Check for missing values and print the count of missing values
     missing_values = dataset.isnull().sum()
     print("\nMissing values:")
     print(missing_values)
```

```
Missing values:
2023 June Unemployment Rate by County (Percent)    0
Unnamed: 1                                         0
Unnamed: 2                                         0
Unnamed: 3                                         5
dtype: int64
```



```
# Create a heatmap to visualize missing data
plt.figure(figsize=(10, 6))
sns.heatmap(dataset.isnull(), cbar=False, cmap='viridis')
plt.title('Missing Data Heatmap')
plt.show()
```

0 s terminée à 02:27

Usage: The code can be used to get a sense of the structure and quality of the provided dataset. It lets users to inspect the data, look for missing values, and depict the distribution of such values.

Discuss possible problems you plan to investigate for future studies.

For future studies, I plan to investigate:

Data Cleaning: t is critical to handle missing values effectively. Consider imputation or removal of rows or columns with missing values, depending on the amount of the missing data.

Exploratory Data Analysis (EDA): To acquire deeper insights, try undertaking more complete EDA, such as investigating connections between variables, displaying distributions, and detecting outliers.

Data Visualization: Extend the visualization capabilities to incorporate different sorts of plots, such as histograms, box plots, scatter plots, or time series plots, to expose new insights in the data.

The provided code is intended to be used as a first step in evaluating the "2023 June Unemployment Rate by County" dataset. It provides data loading, basic exploration, and missing value visualization. To extract useful information and enable informed decision-making, more data cleaning and in-depth exploratory analysis are recommended.