dsbda-b3-13258-1-1

March 4, 2025

Data Wrangling, Perform the following operations using Python on any open source dataset (e.g., data.csv) 1. Import all the required Python Libraries. 2. Locate an open source data from the web (e.g., https://www.kaggle.com). Provide a clear description of the data and its source (i.e., URL of the web site). 3. Load the Dataset into pandas dataframe. 4. Data Preprocessing: check for missing values in the data using pandas isnull(), describe() function to get some initial statistics. Provide variable descriptions. Types of variables etc. Check the dimensions of the data frame. 5. Data Formatting and Data Normalization: Summarize the types of variables by checking the data types (i.e., character, numeric, integer, factor, and logical) of the variables in the data set. If variables are not in the correct data type, apply proper type conversions. 6. Turn categorical variables into quantitative variables in Python.

[1]:
```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from pandas import DataFrame, Series
```

[2]:
```
import seaborn as sns
```

[5]:
```
sns.get_dataset_names()
```

[5]:
```
['anagrams',
 'anscombe',
 'attention',
 'brain_networks',
 'car_crashes',
 'diamonds',
 'dots',
 'dowjones',
 'exercise',
 'flights',
 'fmri',
 'geyser',
 'glue',
 'healthexp',
 'iris',
 'mpg',
 'penguins',
 'planets',
```

```
 'seaice',
 'taxis',
 'tips',
 'titanic']
```

[6]:
```
data = sns.load_dataset("iris")
```

[8]:
```
print(data)
```

```
     sepal_length  sepal_width  petal_length  petal_width    species
0             5.1          3.5           1.4          0.2     setosa
1             4.9          3.0           1.4          0.2     setosa
2             4.7          3.2           1.3          0.2     setosa
3             4.6          3.1           1.5          0.2     setosa
4             5.0          3.6           1.4          0.2     setosa
..            …            …             …            …        …
145           6.7          3.0           5.2          2.3  virginica
146           6.3          2.5           5.0          1.9  virginica
147           6.5          3.0           5.2          2.0  virginica
148           6.2          3.4           5.4          2.3  virginica
149           5.9          3.0           5.1          1.8  virginica

[150 rows x 5 columns]
```

[9]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   sepal_length  150 non-null    float64
 1   sepal_width   150 non-null    float64
 2   petal_length  150 non-null    float64
 3   petal_width   150 non-null    float64
 4   species       150 non-null    object
dtypes: float64(4), object(1)
memory usage: 6.0+ KB
```

[10]:
```
data.head()
```

[10]:

```
   sepal_length  sepal_width  petal_length  petal_width species
0           5.1          3.5           1.4          0.2  setosa
1           4.9          3.0           1.4          0.2  setosa
2           4.7          3.2           1.3          0.2  setosa
3           4.6          3.1           1.5          0.2  setosa
4           5.0          3.6           1.4          0.2  setosa
```

[11]:
```
data.tail()
```

[11]:

	sepal_length	sepal_width	petal_length	petal_width	species
145	6.7	3.0	5.2	2.3	virginica
146	6.3	2.5	5.0	1.9	virginica
147	6.5	3.0	5.2	2.0	virginica
148	6.2	3.4	5.4	2.3	virginica
149	5.9	3.0	5.1	1.8	virginica

[12]:
```
data.describe()
```

[12]:

	sepal_length	sepal_width	petal_length	petal_width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

[13]:
```
top_left_corner_df = data.iloc[:4, :4]
```

[14]:
```
print(top_left_corner_df)
```

	sepal_length	sepal_width	petal_length	petal_width
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2

[15]:
```
data.to_csv()
```

[15]:
```
',sepal_length,sepal_width,petal_length,petal_width,species\r\n0,5.1,3.5,1.4,0.2
,setosa\r\n1,4.9,3.0,1.4,0.2,setosa\r\n2,4.7,3.2,1.3,0.2,setosa\r\n3,4.6,3.1,1.5
,0.2,setosa\r\n4,5.0,3.6,1.4,0.2,setosa\r\n5,5.4,3.9,1.7,0.4,setosa\r\n6,4.6,3.4
,1.4,0.3,setosa\r\n7,5.0,3.4,1.5,0.2,setosa\r\n8,4.4,2.9,1.4,0.2,setosa\r\n9,4.9
,3.1,1.5,0.1,setosa\r\n10,5.4,3.7,1.5,0.2,setosa\r\n11,4.8,3.4,1.6,0.2,setosa\r\
n12,4.8,3.0,1.4,0.1,setosa\r\n13,4.3,3.0,1.1,0.1,setosa\r\n14,5.8,4.0,1.2,0.2,se
tosa\r\n15,5.7,4.4,1.5,0.4,setosa\r\n16,5.4,3.9,1.3,0.4,setosa\r\n17,5.1,3.5,1.4
,0.3,setosa\r\n18,5.7,3.8,1.7,0.3,setosa\r\n19,5.1,3.8,1.5,0.3,setosa\r\n20,5.4,
3.4,1.7,0.2,setosa\r\n21,5.1,3.7,1.5,0.4,setosa\r\n22,4.6,3.6,1.0,0.2,setosa\r\n
23,5.1,3.3,1.7,0.5,setosa\r\n24,4.8,3.4,1.9,0.2,setosa\r\n25,5.0,3.0,1.6,0.2,set
osa\r\n26,5.0,3.4,1.6,0.4,setosa\r\n27,5.2,3.5,1.5,0.2,setosa\r\n28,5.2,3.4,1.4,
0.2,setosa\r\n29,4.7,3.2,1.6,0.2,setosa\r\n30,4.8,3.1,1.6,0.2,setosa\r\n31,5.4,3
.4,1.5,0.4,setosa\r\n32,5.2,4.1,1.5,0.1,setosa\r\n33,5.5,4.2,1.4,0.2,setosa\r\n3
4,4.9,3.1,1.5,0.2,setosa\r\n35,5.0,3.2,1.2,0.2,setosa\r\n36,5.5,3.5,1.3,0.2,seto
```

sa\r\n37,4.9,3.6,1.4,0.1,setosa\r\n38,4.4,3.0,1.3,0.2,setosa\r\n39,5.1,3.4,1.5,0
.2,setosa\r\n40,5.0,3.5,1.3,0.3,setosa\r\n41,4.5,2.3,1.3,0.3,setosa\r\n42,4.4,3.
2,1.3,0.2,setosa\r\n43,5.0,3.5,1.6,0.6,setosa\r\n44,5.1,3.8,1.9,0.4,setosa\r\n45
,4.8,3.0,1.4,0.3,setosa\r\n46,5.1,3.8,1.6,0.2,setosa\r\n47,4.6,3.2,1.4,0.2,setos
a\r\n48,5.3,3.7,1.5,0.2,setosa\r\n49,5.0,3.3,1.4,0.2,setosa\r\n50,7.0,3.2,4.7,1.
4,versicolor\r\n51,6.4,3.2,4.5,1.5,versicolor\r\n52,6.9,3.1,4.9,1.5,versicolor\r
\n53,5.5,2.3,4.0,1.3,versicolor\r\n54,6.5,2.8,4.6,1.5,versicolor\r\n55,5.7,2.8,4
.5,1.3,versicolor\r\n56,6.3,3.3,4.7,1.6,versicolor\r\n57,4.9,2.4,3.3,1.0,versico
lor\r\n58,6.6,2.9,4.6,1.3,versicolor\r\n59,5.2,2.7,3.9,1.4,versicolor\r\n60,5.0,
2.0,3.5,1.0,versicolor\r\n61,5.9,3.0,4.2,1.5,versicolor\r\n62,6.0,2.2,4.0,1.0,ve
rsicolor\r\n63,6.1,2.9,4.7,1.4,versicolor\r\n64,5.6,2.9,3.6,1.3,versicolor\r\n65
,6.7,3.1,4.4,1.4,versicolor\r\n66,5.6,3.0,4.5,1.5,versicolor\r\n67,5.8,2.7,4.1,1
.0,versicolor\r\n68,6.2,2.2,4.5,1.5,versicolor\r\n69,5.6,2.5,3.9,1.1,versicolor\
r\n70,5.9,3.2,4.8,1.8,versicolor\r\n71,6.1,2.8,4.0,1.3,versicolor\r\n72,6.3,2.5,
4.9,1.5,versicolor\r\n73,6.1,2.8,4.7,1.2,versicolor\r\n74,6.4,2.9,4.3,1.3,versic
olor\r\n75,6.6,3.0,4.4,1.4,versicolor\r\n76,6.8,2.8,4.8,1.4,versicolor\r\n77,6.7
,3.0,5.0,1.7,versicolor\r\n78,6.0,2.9,4.5,1.5,versicolor\r\n79,5.7,2.6,3.5,1.0,v
ersicolor\r\n80,5.5,2.4,3.8,1.1,versicolor\r\n81,5.5,2.4,3.7,1.0,versicolor\r\n8
2,5.8,2.7,3.9,1.2,versicolor\r\n83,6.0,2.7,5.1,1.6,versicolor\r\n84,5.4,3.0,4.5,
1.5,versicolor\r\n85,6.0,3.4,4.5,1.6,versicolor\r\n86,6.7,3.1,4.7,1.5,versicolor
\r\n87,6.3,2.3,4.4,1.3,versicolor\r\n88,5.6,3.0,4.1,1.3,versicolor\r\n89,5.5,2.5
,4.0,1.3,versicolor\r\n90,5.5,2.6,4.4,1.2,versicolor\r\n91,6.1,3.0,4.6,1.4,versi
color\r\n92,5.8,2.6,4.0,1.2,versicolor\r\n93,5.0,2.3,3.3,1.0,versicolor\r\n94,5.
6,2.7,4.2,1.3,versicolor\r\n95,5.7,3.0,4.2,1.2,versicolor\r\n96,5.7,2.9,4.2,1.3,
versicolor\r\n97,6.2,2.9,4.3,1.3,versicolor\r\n98,5.1,2.5,3.0,1.1,versicolor\r\n
99,5.7,2.8,4.1,1.3,versicolor\r\n100,6.3,3.3,6.0,2.5,virginica\r\n101,5.8,2.7,5.
1,1.9,virginica\r\n102,7.1,3.0,5.9,2.1,virginica\r\n103,6.3,2.9,5.6,1.8,virginic
a\r\n104,6.5,3.0,5.8,2.2,virginica\r\n105,7.6,3.0,6.6,2.1,virginica\r\n106,4.9,2
.5,4.5,1.7,virginica\r\n107,7.3,2.9,6.3,1.8,virginica\r\n108,6.7,2.5,5.8,1.8,vir
ginica\r\n109,7.2,3.6,6.1,2.5,virginica\r\n110,6.5,3.2,5.1,2.0,virginica\r\n111,
6.4,2.7,5.3,1.9,virginica\r\n112,6.8,3.0,5.5,2.1,virginica\r\n113,5.7,2.5,5.0,2.
0,virginica\r\n114,5.8,2.8,5.1,2.4,virginica\r\n115,6.4,3.2,5.3,2.3,virginica\r\
n116,6.5,3.0,5.5,1.8,virginica\r\n117,7.7,3.8,6.7,2.2,virginica\r\n118,7.7,2.6,6
.9,2.3,virginica\r\n119,6.0,2.2,5.0,1.5,virginica\r\n120,6.9,3.2,5.7,2.3,virgini
ca\r\n121,5.6,2.8,4.9,2.0,virginica\r\n122,7.7,2.8,6.7,2.0,virginica\r\n123,6.3,
2.7,4.9,1.8,virginica\r\n124,6.7,3.3,5.7,2.1,virginica\r\n125,7.2,3.2,6.0,1.8,vi
rginica\r\n126,6.2,2.8,4.8,1.8,virginica\r\n127,6.1,3.0,4.9,1.8,virginica\r\n128
,6.4,2.8,5.6,2.1,virginica\r\n129,7.2,3.0,5.8,1.6,virginica\r\n130,7.4,2.8,6.1,1
.9,virginica\r\n131,7.9,3.8,6.4,2.0,virginica\r\n132,6.4,2.8,5.6,2.2,virginica\r
\n133,6.3,2.8,5.1,1.5,virginica\r\n134,6.1,2.6,5.6,1.4,virginica\r\n135,7.7,3.0,
6.1,2.3,virginica\r\n136,6.3,3.4,5.6,2.4,virginica\r\n137,6.4,3.1,5.5,1.8,virgin
ica\r\n138,6.0,3.0,4.8,1.8,virginica\r\n139,6.9,3.1,5.4,2.1,virginica\r\n140,6.7
,3.1,5.6,2.4,virginica\r\n141,6.9,3.1,5.1,2.3,virginica\r\n142,5.8,2.7,5.1,1.9,v
irginica\r\n143,6.8,3.2,5.9,2.3,virginica\r\n144,6.7,3.3,5.7,2.5,virginica\r\n14
5,6.7,3.0,5.2,2.3,virginica\r\n146,6.3,2.5,5.0,1.9,virginica\r\n147,6.5,3.0,5.2,
2.0,virginica\r\n148,6.2,3.4,5.4,2.3,virginica\r\n149,5.9,3.0,5.1,1.8,virginica\
r\n'

[16]:
```
ash = data.copy()
```

[17]:
```
print(ash)
```

```
     sepal_length  sepal_width  petal_length  petal_width    species
0             5.1          3.5           1.4          0.2     setosa
1             4.9          3.0           1.4          0.2     setosa
2             4.7          3.2           1.3          0.2     setosa
3             4.6          3.1           1.5          0.2     setosa
4             5.0          3.6           1.4          0.2     setosa
..            …            …             …            …       …
145           6.7          3.0           5.2          2.3  virginica
146           6.3          2.5           5.0          1.9  virginica
147           6.5          3.0           5.2          2.0  virginica
148           6.2          3.4           5.4          2.3  virginica
149           5.9          3.0           5.1          1.8  virginica

[150 rows x 5 columns]
```

[18]:
```
data.count()
```

[18]:
```
sepal_length    150
sepal_width     150
petal_length    150
petal_width     150
species         150
dtype: int64
```

[19]:
```
data.cummax()
```

[19]:
```
     sepal_length  sepal_width  petal_length  petal_width    species
0             5.1          3.5           1.4          0.2     setosa
1             5.1          3.5           1.4          0.2     setosa
2             5.1          3.5           1.4          0.2     setosa
3             5.1          3.5           1.5          0.2     setosa
4             5.1          3.6           1.5          0.2     setosa
..            …            …             …            …       …
145           7.9          4.4           6.9          2.5  virginica
146           7.9          4.4           6.9          2.5  virginica
147           7.9          4.4           6.9          2.5  virginica
148           7.9          4.4           6.9          2.5  virginica
149           7.9          4.4           6.9          2.5  virginica

[150 rows x 5 columns]
```

[20]:
```
data.cummin()
```

[20]:
```
     sepal_length  sepal_width  petal_length  petal_width species
0             5.1          3.5           1.4          0.2  setosa
1             4.9          3.0           1.4          0.2  setosa
2             4.7          3.0           1.3          0.2  setosa
3             4.6          3.0           1.3          0.2  setosa
4             4.6          3.0           1.3          0.2  setosa
..            …            …             …            …      …
145           4.3          2.0           1.0          0.1  setosa
146           4.3          2.0           1.0          0.1  setosa
147           4.3          2.0           1.0          0.1  setosa
148           4.3          2.0           1.0          0.1  setosa
149           4.3          2.0           1.0          0.1  setosa

[150 rows x 5 columns]
```

[21]:
```
data.dropna()
```

[21]:
```
     sepal_length  sepal_width  petal_length  petal_width    species
0             5.1          3.5           1.4          0.2     setosa
1             4.9          3.0           1.4          0.2     setosa
2             4.7          3.2           1.3          0.2     setosa
3             4.6          3.1           1.5          0.2     setosa
4             5.0          3.6           1.4          0.2     setosa
..            …            …             …            …        …
145           6.7          3.0           5.2          2.3  virginica
146           6.3          2.5           5.0          1.9  virginica
147           6.5          3.0           5.2          2.0  virginica
148           6.2          3.4           5.4          2.3  virginica
149           5.9          3.0           5.1          1.8  virginica

[150 rows x 5 columns]
```

[22]:
```
data.any()
```

[22]:
```
sepal_length    True
sepal_width     True
petal_length    True
petal_width     True
species         True
dtype: bool
```

[23]:
```
data.get(40)
```

[25]:
```
ass = data.get(40)
```

[26]:
```
print(ass)
```

```
None
```

[9]:
```
import seaborn as sns
```

[10]:
```
data = sns.load_dataset("iris")
```

[11]:
```
print(data)
```

```
     sepal_length  sepal_width  petal_length  petal_width    species
0             5.1          3.5           1.4          0.2     setosa
1             4.9          3.0           1.4          0.2     setosa
2             4.7          3.2           1.3          0.2     setosa
3             4.6          3.1           1.5          0.2     setosa
4             5.0          3.6           1.4          0.2     setosa
..            …            …             …            …         …
145           6.7          3.0           5.2          2.3  virginica
146           6.3          2.5           5.0          1.9  virginica
147           6.5          3.0           5.2          2.0  virginica
148           6.2          3.4           5.4          2.3  virginica
149           5.9          3.0           5.1          1.8  virginica

[150 rows x 5 columns]
```

[12]:
```
data.iloc[3:5, 0:2]
```

[12]:
```
   sepal_length  sepal_width
3           4.6          3.1
4           5.0          3.6
```

[13]:
```
data.iloc[[1, 2, 4], [0, 2]]
```

[13]:
```
   sepal_length  petal_length
1           4.9           1.4
2           4.7           1.3
4           5.0           1.4
```

[14]:
```
data.iloc[1:3, :]
```

[14]:
```
   sepal_length  sepal_width  petal_length  petal_width species
1           4.9          3.0           1.4          0.2  setosa
2           4.7          3.2           1.3          0.2  setosa
```

[15]:
```
data.iloc[:, 1:3]
```

[15]:
```
     sepal_width  petal_length
0            3.5           1.4
1            3.0           1.4
2            3.2           1.3
```

```
3            3.1           1.5
4            3.6           1.4
..           …             …
145          3.0           5.2
146          2.5           5.0
147          3.0           5.2
148          3.4           5.4
149          3.0           5.1

[150 rows x 2 columns]
```

[16]:
```
data.iloc[1, 1]
```

[16]:
```
3.0
```

[23]:
```
cols_2_4=data.columns[2:4]
data[cols_2_4]
```

[23]:
```
     petal_length  petal_width
0             1.4          0.2
1             1.4          0.2
2             1.3          0.2
3             1.5          0.2
4             1.4          0.2
..            …            …
145           5.2          2.3
146           5.0          1.9
147           5.2          2.0
148           5.4          2.3
149           5.1          1.8

[150 rows x 2 columns]
```

[25]:
```
data[data.columns[2:4]].iloc[5:10]
```

[25]:
```
   petal_length  petal_width
5           1.7          0.4
6           1.4          0.3
7           1.5          0.2
8           1.4          0.2
9           1.5          0.1
```

[30]:
```
data.isnull()
```

[30]:
```
     sepal_length  sepal_width  petal_length  petal_width  species
0           False        False         False        False    False
1           False        False         False        False    False
```

```
2           False        False         False        False    False
3           False        False         False        False    False
4           False        False         False        False    False
..            …            …             …            …        …
145         False        False         False        False    False
146         False        False         False        False    False
147         False        False         False        False    False
148         False        False         False        False    False
149         False        False         False        False    False

[150 rows x 5 columns]
```

[31]:
```
data.isnull().any()
```

[31]:
```
sepal_length    False
sepal_width     False
petal_length    False
petal_width     False
species         False
dtype: bool
```

[32]:
```
data.isnull().sum(axis = 1)
```

[32]:
```
0      0
1      0
2      0
3      0
4      0
      ..
145    0
146    0
147    0
148    0
149    0
Length: 150, dtype: int64
```

[33]:
```
data.isnull().sum()
```

[33]:
```
sepal_length    0
sepal_width     0
petal_length    0
petal_width     0
species         0
dtype: int64
```

[34]:
```
data.isna().sum()
```

[34]:
```
sepal_length    0
sepal_width     0
petal_length    0
petal_width     0
species         0
dtype: int64
```

[51]:
```
data.dtypes
```

[51]:
```
sepal_length    float64
sepal_width     float64
petal_length    float64
petal_width     float64
species          object
dtype: object
```

[ ]:
```
Name= akash pachrne
roll no :13254
```