exercises-pollinators-datasets-exploration

March 22, 2022

1 Exercises - Pollinators datasets exploration

Exercises with some pollinators datasets.

1.1 Packages import

[26]:
```
import os # operating system functions
import chardet # Universal Character Encoding Detector
import requests # web requests
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib as plt # data visualization
import seaborn as sb # data visualization
import graphviz # grahp visualization
from sklearn.model_selection import StratifiedShuffleSplit # dataset subsetting
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import LabelEncoder # mange categorical data
from sklearn import metrics # results evaluation
```

We probably will download and save more than 1 datase so let's make a funcition for it

[20]:
```
def DatasetDownload(dataset_url, dataset_directory_path, dataset_file_name):
    print("Download started")
    request_dataset = requests.get(dataset_url, allow_redirects=True)
    print("Download completed")
    if request_dataset.status_code != 200:
        print(f"Request status: {request_dataset.status_code}")
    else:
        print("Writing started")
        os.makedirs(dataset_directory_path, exist_ok=True)
        open( dataset_directory_path + dataset_file_name , 'wb').
 ↪write(request_dataset.content)
        print("Writing completed")
    print("End")
    return
```

1.2 Insect Pollinator Initiative - Natural History Museum Data Portal

Graham N Stone; Alfried Vogler; Adam Vanbergen; Jacqueline Mackenzie-Dodds (2017). Dataset: Insect Pollinators Archive. Resource: Insect Pollinator Initiative. Natural History Museum Data Portal (data.nhm.ac.uk). https://doi.org/10.5519/0062900

Retrieved: 16:39 19 Mar 2022 (GMT)

1.2.1 IPI-NHMDP - Data download - (One shoot execution)

Let's use the original website.

Next steps are "one shoot execution", you should execute it only the first time, once did it you can go directly to *Starting points* that youll'find along the code.

[3]:
```
# Dataset url
NHMDP_PI_dataset_url = 'https://data.nhm.ac.uk/dataset/
 ↪46e122c6-7acd-44ec-a354-81a412da419a/resource/
 ↪784d74b6-6b0e-4fd4-b0b5-798ac7b1a11b/download/ipifordataportal.xlsx'

# Desired directory
NHMDP_PI_dataset_directory = 'Datasets/Pollinators/NHMDP/PollinatorsInitiative'

# Desired file name
NHMDP_PI_dataset_name = 'PollinatorsInitiative.xlsx'
```

[12]:
```
# Download and Save
DatasetDownload(NHMDP_PI_dataset_url, NHMDP_PI_dataset_directory,␣
 ↪NHMDP_PI_dataset_name)
```

```
Download started
Download completed
Writing started
Writing completed
End
```

1.2.2 IPI-NHMDP - Data import - Starting point

[4]:
```
IPI_NHMDP_dataset = pd.
 ↪read_excel(NHMDP_PI_dataset_directory+NHMDP_PI_dataset_name,␣
 ↪engine='openpyxl')
```

1.2.3 IPI-NHMDP - Exploration

[14]:
```
IPI_NHMDP_dataset.describe()
```

[14]:

	Specimen No/Barcode
count	1.185400e+04
mean	1.006605e+07

```
std           7.403999e+03
min           1.005246e+07
25%           1.005963e+07
50%           1.006886e+07
75%           1.007182e+07
max           1.007598e+07
```

[5]:
```
IPI_NHMDP_dataset.head()
```

[5]:
```
                              Project Name Specimen No Prefix  \
0  Insect Pollinator Initiative - agriland              NHMUK
1  Insect Pollinator Initiative - agriland              NHMUK
2  Insect Pollinator Initiative - agriland              NHMUK
3  Insect Pollinator Initiative - agriland              NHMUK
4  Insect Pollinator Initiative - agriland              NHMUK

   Specimen No/Barcode Specimen Code         Country Province/State/Territory  \
0             10052460   AL_11_01750  United Kingdom                   England
1             10052461   AL_11_01751  United Kingdom                   England
2             10052462   AL_11_01753  United Kingdom                   England
3             10052463   AL_11_01754  United Kingdom                   England
4             10052464   AL_11_01755  United Kingdom                   England

  District/County/Shire Precise Locality  Coll Date    Method     Collector  \
0       West Yorkshire      Harden Moor 2011-06-27  Pan trap  M. McKerchar
1       West Yorkshire      Harden Moor 2011-06-27  Pan trap  M. McKerchar
2       West Yorkshire      Harden Moor 2011-06-27  Pan trap  M. McKerchar
3       West Yorkshire      Harden Moor 2011-06-27  Pan trap  M. McKerchar
4       West Yorkshire      Harden Moor 2011-06-27  Pan trap  M. McKerchar

    Collector 1 Collector 2     Identifier  \
0  M  McKerchar              S P M Roberts
1  M  McKerchar         NaN  S P M Roberts
2  M  McKerchar         NaN  S P M Roberts
3  M  McKerchar         NaN  S P M Roberts
4  M  McKerchar         NaN  S P M Roberts

                             Determination     SEX Stage
0  Lasioglossum cupromicans (Pérez, J., 1903)  Female   NaN
1  Lasioglossum cupromicans (Pérez, J., 1903)  Female   NaN
2  Lasioglossum cupromicans (Pérez, J., 1903)  Female   NaN
3  Lasioglossum cupromicans (Pérez, J., 1903)  Female   NaN
4       Lasioglossum fratellum (Perez, 1903)  Female   NaN
```

[6]:
```
IPI_NHMDP_dataset.columns
```

[6]:
```
Index(['Project Name', 'Specimen No Prefix', 'Specimen No/Barcode',
       'Specimen Code', 'Country', 'Province/State/Territory',
       'District/County/Shire', 'Precise Locality', 'Coll Date', 'Method',
       'Collector', 'Collector 1', 'Collector 2', 'Identifier',
       'Determination', 'SEX', 'Stage'],
      dtype='object')
```

Mmm I don't see particularly interesting information.

Let's check how many per state differnt specimes have been collected

[14]:
```
IPI_NHMDP_dataset[["Country","Specimen Code"]].groupby("Country").describe()
```

[14]:

	Specimen Code			
	count	unique	top	freq
Country				
United Kingdom	11852	11807	Wi-01-3.13-P10003	2

[15]:
```
IPI_NHMDP_dataset[["Province/State/Territory","Specimen Code"]].
↪groupby("Province/State/Territory").describe()
```

[15]:

	Specimen Code			
	count	unique	top	freq
Province/State/Territory				
England	10028	9996	Ca-05-1.12-P30003	2
Scotland	1824	1811	Ay-15-3.12-P50013	2

[16]:
```
IPI_NHMDP_dataset[["Province/State/Territory","District/County/Shire","Specimen␣
↪Code"]].groupby("District/County/Shire").describe()
```

[16]:

	Province/State/Territory			
	count	unique	top	freq
District/County/Shire				
Bedfordshire	1053	1	England	1053
Cambridgeshire	2356	1	England	2356
Cumbria	113	1	England	113
Dorset	492	1	England	492
Dumfries and Galloway	137	1	Scotland	137
East Ayrshire	523	1	Scotland	523
East Renfrewshire	29	1	Scotland	29
East Riding of Yorkshire	1471	1	England	1471
Highland	651	1	Scotland	651
Kent	173	1	England	173
Lancashire	219	1	England	219
North Lanarkshire	167	1	Scotland	167
North Yorkshire	254	1	England	254
Renfrewshire	14	1	Scotland	14
South Lanarkshire	303	1	Scotland	303

```
Staffordshire                              1359      1   England  1359
West Yorkshire                              895      1   England   895
Wiltshire                                  1643      1   England  1643
```

District/County/Shire	Specimen Code count	unique	top	freq
Bedfordshire	1053	1052	AL_11_03988	2
Cambridgeshire	2356	2340	Ca-01-1.13-P40002	2
Cumbria	113	113	Yo-08-1.12-P30003	1
Dorset	492	492	AL_12_07052	1
Dumfries and Galloway	137	137	Ay-08-3.12-P10001	1
East Ayrshire	523	523	Ay-01-3.12-P20001	1
East Renfrewshire	29	29	Ay-12-3.12-P10001	1
East Riding of Yorkshire	1471	1467	AL_11_02429	2
Highland	651	643	In-04-1.12-P50001	2
Kent	173	173	AL_12_06790	1
Lancashire	219	219	AL_11_02651	1
North Lanarkshire	167	162	Ay-15-3.12-P50009	2
North Yorkshire	254	253	AL_11_06052	2
Renfrewshire	14	14	Ay-09-3.12-P30001	1
South Lanarkshire	303	303	Ay-04-3.12-P10009	1
Staffordshire	1359	1359	St-02-3.12-P10001	1
West Yorkshire	895	894	AL_11_02507	2
Wiltshire	1643	1634	Wi-01-3.13-P40001	2

Could be nice try to represent these data on a geographical map… but it's a bit out of the exercise scope

1.3 Global pollinator database - Boreux & Klein - Figshare Dataset

Boreux, Virginie; Klein, Alexandra-Maria (2019): Global pollinator database. figshare. Dataset. https://doi.org/10.6084/m9.figshare.9980471.v1

1.3.1 GPD-F - Data download - (One shoot execution)

[17]:
```
# Dataset url
GPD_F_dataset_url = 'https://figshare.com/ndownloader/files/18003863'

# Desired directory
GPD_F_dataset_directory = 'Datasets/Pollinators/Figshare/
 ↪GlobalPollinatorDatabase'

# Desired file name
GPD_F_dataset_name = 'GlobalPollinatorDatabase.csv'
```

```
# Description dataset url
GPD_F_description_dataset_url = 'https://figshare.com/ndownloader/files/
 ↪18003860'

# Desired file name
GPD_F_description_dataset_name = 'GlobalPollinatorDatabaseDescription.csv'
```

[21]:
```
# Download and Save
DatasetDownload(GPD_F_dataset_url, GPD_F_dataset_directory, GPD_F_dataset_name)
```

```
Download started
Download completed
Writing started
Writing completed
End
```

[22]:
```
# Download and Save description
DatasetDownload(GPD_F_description_dataset_url, GPD_F_dataset_directory,␣
 ↪GPD_F_description_dataset_name)
```

```
Download started
Download completed
Writing started
Writing completed
End
```

1.3.2 GPD - Data import - Starting point

[30]:
```
GPD_dataset = pd.read_csv(GPD_F_dataset_directory+GPD_F_dataset_name)
```

read_csv on dtaset description rise an error of text decoding: *UnicodeDecodeError: 'utf-8' codec can't decode byte 0x96 in position 292: invalid start byte*

Let's check the encoding

[27]:
```
with open(GPD_F_dataset_directory+GPD_F_description_dataset_name, 'rb') as file:
    print(chardet.detect(file.read()))
```

```
{'encoding': 'Windows-1252', 'confidence': 0.73, 'language': ''}
```

[28]:
```
with open(GPD_F_dataset_directory+GPD_F_dataset_name, 'rb') as file:
    print(chardet.detect(file.read()))
```

```
{'encoding': 'ascii', 'confidence': 1.0, 'language': ''}
```

[29]:
```
GPD_dataset_description = pd.
 ↪read_csv(GPD_F_dataset_directory+GPD_F_description_dataset_name,␣
 ↪encoding='Windows-1252')
```

1.3.3 GPD-F - Exploration

[31]:
```
GPD_dataset.describe()
```

[31]:
```
       Unnamed: 0    diameter      tongue        body
count  796.000000  474.000000  293.000000  633.000000
mean   398.500000   27.781814    7.291297   11.592891
std    229.929699   31.164702    4.009739    3.862993
min      1.000000    2.000000    2.000000    2.000000
25%    199.750000   12.200000    5.000000    9.000000
50%    398.500000   25.000000    5.500000   11.500000
75%    597.250000   25.000000    9.000000   13.500000
max    796.000000  150.000000   26.400000   25.000000
```

So... seems we have to deal with a lot of missing values... yeah! XD

[33]:
```
GPD_dataset.columns
```

[33]:
```
Index(['Unnamed: 0', 'crop', 'type', 'season', 'diameter', 'corolla', 'colour',
       'nectar', 'b.system', 's.pollination', 'inflorescence', 'composite',
       'visitor', 'guild', 'tongue', 'body', 'sociality', 'feeding'],
      dtype='object')
```

[34]:
```
GPD_dataset_description.describe()
```

[34]:
```
       Unnamed: 0
count   15.000000
mean     8.000000
std      4.472136
min      1.000000
25%      4.500000
50%      8.000000
75%     11.500000
max     15.000000
```

[36]:
```
GPD_dataset_description
```

[36]:
```
    Unnamed: 0           Name       Group        Type    Unit  \
0            1           type       Plant    discrete  levels
1            2         season       Plant    discrete  levels
2            3       diameter       Plant  continuous      mm
3            4        corolla       Plant    discrete  levels
4            5         colour       Plant    discrete  levels
5            6         nectar       Plant    discrete  levels
6            7       b.system       Plant    discrete  levels
7            8  s.pollination       Plant    discrete  levels
8            9  inflorescence       Plant    discrete  levels
9           10      composite       Plant    discrete  levels
```

```
10          11          guild  Pollinator    discrete  levels
11          12         tongue  Pollinator  continuous      mm
12          13           body  Pollinator  continuous      mm
13          14      sociality  Pollinator    discrete  levels
14          15        feeding  Pollinator    discrete  levels

                                          Description  \
0                       arboreous or herbaceous plant
1   Flower season: Describes the seasonal range. F…
2                                    Flower diameter
3                                Flower corolla type
4                                      Flower colour
5                      Whether flower contains nectar
6                                Type of bloom system
7                                    Self pollination
8                                Type of inflorescence
9                  Whether flower is composite or not
10                                   Pollinator guild
11                          Pollinator tongue length
12                            Pollinator body length
13             Whether pollinator is sociality or not
14                                 Feeding behaviour

                                             Levels
0                               arboreous, herbaceous
1   sprisum, summer, spriaut, spring, autspri, sum…
2                                               NaN
3                           campanulate open, tubular
4       white, yellow, purple, pink, green, blue, red
5                                           yes, no
6   insects, insects/bats, insects/bats, insects/b…
7                                           yes, no
8    solitary, solitary/clusters, solitary/pairs, yes
9                                           yes, no
10  andrenidae, bumblebees, butterflies, coleopter…
11                                              NaN
12                                              NaN
13                                          yes, no
14                 oligolectic, parasitic, polylectic
```

[37]:
```
GPD_dataset.head()
```

[37]:
```
   Unnamed: 0                 crop        type   season  diameter  \
0           1  Vaccinium_corymbosum   arboreous  sprisum       NaN
1           2  Vaccinium_corymbosum   arboreous  sprisum       NaN
2           3        Brassica_napus  herbaceous   summer      12.5
3           4        Brassica_napus  herbaceous   summer      12.5
```

```
4           5        Brassica_napus  herbaceous   summer      12.5

       corolla  colour nectar      b.system s.pollination inflorescence  \
0  CAMPANULATE   white    yes       insects            no           yes
1  CAMPANULATE   white    yes       insects            no           yes
2         OPEN  yellow    yes  wind/insects            no           yes
3         OPEN  yellow    yes  wind/insects            no           yes
4         OPEN  yellow    yes  wind/insects            no           yes

  composite              visitor       guild  tongue  body sociality  \
0        no     Andrena_wilkella  ANDRENIDAE     NaN  10.5        no
1        no  Andrena_barbilabris  ANDRENIDAE     NaN  10.5        no
2        no    Andrena_cineraria  ANDRENIDAE     NaN  12.0        no
3        no     Andrena_flavipes  ANDRENIDAE     NaN  11.0        no
4        no      Andrena_gravida  ANDRENIDAE     NaN  13.0        no

       feeding
0  oligolectic
1   polylectic
2   polylectic
3   polylectic
4   polylectic
```

Maybe we can try some clusterng tecnique on this dataset to find out some interesting relationship