Data wrangling 1

1. Import all the required Python Libraries

```
import pandas as pd
```

```
import numpy as np
```

```
import matplotlib.pyplot as plt
```

```
%matplotlib inline
```

```
from sklearn import metrics
```

```
from sklearn import metrics
```

```
from sklearn.neighbors import KNeighborsClassifier
```

```
from sklearn.linear_model import LogisticRegression
```

```
from sklearn.model_selection import train_test_split
```

2. Locate an open source data from the web

....

3. Load the Dataset into Pandas DataFrame.

```
df= pd.read_csv("https://raw.githubusercontent.com/okfn/dataportals.org/master/data/portals.csv")
```

```
print(df)
```

```
                          generator                          api_endpoint  \
0                               NaN                                   NaN
1                               NaN                                   NaN
2                     CKAN: 2.1.3      http://africaopendata.org/api/
3                               NaN                                   NaN
4                               NaN                                   NaN
..                              ...                                   ...
599                            ckan       https://opendatanepal.com/api
600                             NaN                                   NaN
601  Custom/in-house implementation         https://api.dadosjusbr.org/
602  Custom/in-house implementation  https://opendata.bratislava.sk/api
603                      ArcGIS Hub                                   NaN

                           api_type  \
0                               NaN
1                               NaN
2                               NaN
3                               NaN
4                               NaN
..                              ...
599                        CKAN API
600                             NaN
601  Custom/in-house implementation
602  Custom/in-house implementation
603                             NaN

                           full_metadata_download
0                                             NaN
1                                             NaN
2                                             NaN
3                                             NaN
4                                             NaN
..                                            ...
599                                           NaN
600                                           NaN
601                                           NaN
602                                           NaN
603  https://open-data-kosice-mesto.hub.arcgis.com/...

[604 rows x 22 columns]
```

now there is a data processing method

```
df.isnull()
```

	name	title	url	author	publisher	issued	publisher_classification	description	tags
0	False	False	False	False	False	True	True	False	False
1	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	True	True	False	False
3	False	False	False	False	False	True	False	False	False
4	False	False	False	False	False	True	False	False	False
...	...	...	...	...	...	...	...	...	...
599	False	False	False	False	False	False	False	False	False
600	False	False	False	False	False	True	False	False	False
601	False	False	False	False	False	False	False	False	False
602	False	False	False	False	False	True	False	False	False
603	False	False	False	False	False	True	False	False	False

604 rows × 22 columns

describe() : returns the statistical summary of dataframe or series.

size() : count the number of element along given axis.

shape() : gives the number of elements in each dimension of an array.

ndim() : return the number of dimensions of an array.

```
df.describe()
```

	name	title	url	author	publisher	issu
count	604	604	604	539	543	
unique	600	598	600	497	506	
top	state_of_washington	Washington	http://dados.recife.pe.gov.br	African Development Bank Group	African Development Bank Group	27/08/2(
freq	4	4	2	21	22	

4 rows × 22 columns

```
size = df.size
shape = df.shape
df_ndim = df.ndim
series_ndim = df["name"].ndim
print("Size - {}\nShape = {}\nShape[0]x shape[1]= {}".format(size,shape, shape[0]* shape[1]))
```

```
Size - 13288
Shape = (604, 22)
Shape[0]x shape[1]= 13288
```

```
print("ndim of DataFrame = {}\nndim of series ={}".format(df_ndim,series_ndim))
```

```
ndim of DataFrame = 2
ndim of series =1
```

...

5. Data Formatting and Data Normalization: Summarize the types of variables by checking the data types of the variables in the data set. If variables are not in the correct data type, apply proper type conversions.

dtypes: to check the data types of colums in a DataFrame

```
dataTypeSeries = df.dtypes
print("Data type of Each column of DataFrame is:")
```

```
Data type of Each column of DataFrame is:
```

```
print (dataTypeSeries)
```

```
name                        object
title                       object
url                         object
author                      object
publisher                   object
issued                      object
publisher_classification    object
description                 object
tags                        object
license_id                  object
license_url                 object
license_notes               object
place                       object
location                    object
country                     object
language                    object
status                      object
metadatacreated             object
generator                   object
api_endpoint                object
api_type                    object
full_metadata_download      object
dtype: object
```

6. Turn categorical variables into quantitative variables in Python.

get_dummies() : this method will return dummy variable colums.

concat() : to concatenate dummy columns into DataFrames

```
dumies = pd.get_dummies(df.author)
```

```
merged = pd.concat([df,dumies], axis  = 'columns')
merged.drop(['author'],axis = 'columns')
```

	name	title	url	publisher	issued	pu
0	a2gov_org	Ann Arbor, Michigan	http://www.a2gov.org/services/data/Pages/defau...	City of Ann Arbor	NaN	
1	acikveri-sahinbey-bel-tr	Açık Veri Portali - Test Yayını	http://acikveri.sahinbey.bel.tr/dataset	SahinBey Belediyesi	31/01/2015	
2	africa_open_data	Africa Open Data	http://africaopendata.org/	Africa Open Data	NaN	
3	ajuntament-de-tarragona	Open Data Tarragona	http://opendata.tarragona.cat/	Ajuntament de Tarragona	NaN	
4	ajuntament-de-terassa	Open Data Terassa	http://opendata.terrassa.cat/	Ajuntament de Terassa	NaN	
...	...	...	...	...	...	
599	open-data-nepal	Open Data Nepal	https://opendatanepal.com/	Open Knowledge Nepal	2018-03-03	
600	stat-tj	Agency on Statistics under President of the Re...	https://www.stat.tj/en	The Statistical Agency under President of the ...	NaN	
601	dadosjusbr	DadosJusBR	https://dadosjusbr.org/	Instituto Federal de Alagoas	2018-12-31	
602	bratislava-opendata	Bratislava Open Data Portal	https://opendata.bratislava.sk	The city of Bratislava	NaN	
603	sk-kosice-opendata	The city of Košice Open Data Portal	https://opendata.kosice.sk	The city of Košice	NaN	

604 rows × 518 columns

```
print(merged)
```

```
                      name  \
0                  a2gov_org
1    acikveri-sahinbey-bel-tr
2            africa_open_data
3     ajuntament-de-tarragona
4       ajuntament-de-terassa
..                        ...
599           open-data-nepal
600                  stat-tj
```

```
601               dadosjusbr
602       bratislava-opendata
603        sk-kosice-opendata

                                          title  \
0                                Ann Arbor, Michigan
1                     Açık Veri Portali - Test Yayını
2                                  Africa Open Data
3                                Open Data Tarragona
4                                  Open Data Terassa
..                                               ...
599                                Open Data Nepal
600  Agency on Statistics under President of the Re...
601                                        DadosJusBR
602                          Bratislava Open Data Portal
603                The city of Košice Open Data Portal

                                            url  \
0    http://www.a2gov.org/services/data/Pages/defau...
1              http://acikveri.sahinbey.bel.tr/dataset
2                           http://africaopendata.org/
3                        http://opendata.tarragona.cat/
4                         http://opendata.terrassa.cat/
..                                              ...
599                        https://opendatanepal.com/
600                            https://www.stat.tj/en
601                          https://dadosjusbr.org/
602                      https://opendata.bratislava.sk
603                          https://opendata.kosice.sk

                                          author  \
0                                City of Ann Arbor
1                                         pinardag
2                                Africa Open Data
3                           Ajuntament de Tarragona
4                             Ajuntament de Terassa
..                                              ...
599                           Open Knowledge Nepal
600  The Statistical Agency under President of the ...
601  Instituto Federal de Alagoas (IFAL) in partner...
```