Module: EGT206 Data Management

Practical 1: Data Cleaning using Python

Data scientists spend a large amount of their time cleaning datasets and getting them down to a form with which they can work. In fact, a lot of data scientists argue that the initial steps of obtaining and cleaning data constitute 80% of the job. It is important to be able to deal with messy data, whether that means missing values, inconsistent formatting, malformed records, or nonsensical outliers.

In this lab, you are expected to achieve the following:

- Counting Missing Values
- Dropping Missing Data
- Identifying Duplicate Records
- Replacing Text in Strings
- Changing String Case

Here are the datasets that we will be using:

- BL-Flickr-Images-Book.csv – A CSV file containing information about books from the British Library

Instructions:

1. Import the required module (pandas)

2. Create a DataFrame out of the csv file 'BL-Flickr-Images-Book.csv' and look at the first five entries.

	Identifier	Edition Statement	Place of Publication	Date of Publication	Publisher	Title	Author	Contributors	Corporate Author	Corporate Contributors	Former owner	Engraver	Issuance type	
0	206	NaN	London	1879 [1878]	S. Tinsley & Co.	Walter Forbes. [A novel.] By A. A	A. A.	FORBES, Walter.	NaN	NaN	NaN	NaN	monographic	http://w
1	216	NaN	London; Virtue & Yorston	1868	Virtue & Co.	All for Greed. [A novel. The dedication signed...	A., A. A.	BLAZE DE BURY, Marie Pauline Rose - Baroness	NaN	NaN	NaN	NaN	monographic	http://w
2	218	NaN	London	1869	Bradbury, Evans & Co.	Love the Avenger. By the author of "All for Gr...	A., A. A.	BLAZE DE BURY, Marie Pauline Rose - Baroness	NaN	NaN	NaN	NaN	monographic	http://w
3	472	NaN	London	1851	James Darling	Welsh Sketches, chiefly ecclesiastical, to the...	A., E. S.	Appleyard, Ernest Silvanus.	NaN	NaN	NaN	NaN	monographic	http://w
4	480	A new edition, revised, etc.	London	1857	Wertheim & Macintosh	[The World in which I live, and my place in it...	A., E. S.	BROOME, John Henry.	NaN	NaN	NaN	NaN	monographic	http://w

3. List the sums of missing values for each column

```
Identifier                 0
Edition Statement       7514
Place of Publication       0
Date of Publication      181
Publisher               4195
Title                      0
Author                  1778
Contributors               0
Corporate Author        8287
Corporate Contributors  8287
Former owner            8286
Engraver                8287
Issuance type              0
Flickr URL                 0
Shelfmarks                 0
dtype: int64
```

4. Drop columns with missing values.

	Identifier	Place of Publication	Title	Contributors	Issuance type	Flickr URL	Shelfmarks
0	206	London	Walter Forbes. [A novel.] By A. A	FORBES, WALTER.	monographic	https://www.flickr.com/photos/britishlibrary/t...	British Library HMNTS 12641.b.30.
1	216	London; Virtue & Yorston	All for Greed. [A novel. The dedication signed...	BLAZE DE BURY, MARIE PAULINE ROSE - BARONESS	monographic	https://www.flickr.com/photos/britishlibrary/t...	British Library HMNTS 12626.cc.2.
2	218	London	Love the Avenger. By the author of "All for Gr...	BLAZE DE BURY, MARIE PAULINE ROSE - BARONESS	monographic	https://www.flickr.com/photos/britishlibrary/t...	British Library HMNTS 12625.dd.1.
3	472	London	Welsh Sketches, chiefly ecclesiastical, to the...	APPLEYARD, ERNEST SILVANUS.	monographic	https://www.flickr.com/photos/britishlibrary/t...	British Library HMNTS 10369.bbb.15.
4	480	London	[The World in which I live, and my place in it...	BROOME, JOHN HENRY.	monographic	https://www.flickr.com/photos/britishlibrary/t...	British Library HMNTS 9007.d.28.

5. Check if there is any duplicated rows in data

0

6. In column "Flickr URL", replace "http" with "https".

	Identifier	Place of Publication	Title	Contributors	Issuance type	Flickr URL	Shelfmarks
0	206	London	Walter Forbes. [A novel.] By A. A	FORBES, Walter.	monographic	https://www.flickr.com/photos/britishlibrary/t...	British Library HMNTS 12641.b.30.
1	216	London; Virtue & Yorston	All for Greed. [A novel. The dedication signed...	BLAZE DE BURY, Marie Pauline Rose - Baroness	monographic	https://www.flickr.com/photos/britishlibrary/t...	British Library HMNTS 12626.cc.2.
2	218	London	Love the Avenger. By the author of "All for Gr...	BLAZE DE BURY, Marie Pauline Rose - Baroness	monographic	https://www.flickr.com/photos/britishlibrary/t...	British Library HMNTS 12625.dd.1.
3	472	London	Welsh Sketches, chiefly ecclesiastical, to the...	Appleyard, Ernest Silvanus.	monographic	https://www.flickr.com/photos/britishlibrary/t...	British Library HMNTS 10369.bbb.15.
4	480	London	[The World in which I live, and my place in it...	BROOME, John Henry.	monographic	https://www.flickr.com/photos/britishlibrary/t...	British Library HMNTS 9007.d.28.

7. Change all values in Contributors to be Uppercase

	Identifier	Place of Publication	Title	Contributors	Issuance type	Flickr URL	Shelfmarks
0	206	London	Walter Forbes. [A novel.] By A. A	FORBES, WALTER.	monographic	https://www.flickr.com/photos/britishlibrary/t...	British Library HMNTS 12641.b.30.
1	216	London; Virtue & Yorston	All for Greed. [A novel. The dedication signed...	BLAZE DE BURY, MARIE PAULINE ROSE - BARONESS	monographic	https://www.flickr.com/photos/britishlibrary/t...	British Library HMNTS 12626.cc.2.
2	218	London	Love the Avenger. By the author of "All for Gr...	BLAZE DE BURY, MARIE PAULINE ROSE - BARONESS	monographic	https://www.flickr.com/photos/britishlibrary/t...	British Library HMNTS 12625.dd.1.
3	472	London	Welsh Sketches, chiefly ecclesiastical, to the...	APPLEYARD, ERNEST SILVANUS.	monographic	https://www.flickr.com/photos/britishlibrary/t...	British Library HMNTS 10369.bbb.15.
4	480	London	[The World in which I live, and my place in it...	BROOME, JOHN HENRY.	monographic	https://www.flickr.com/photos/britishlibrary/t...	British Library HMNTS 9007.d.28.