

Project: Capstone Project 1: Data Wrangling

Project Name: Books Recommendation System.

DataSet.

1. **BX-Books:-**Contains books details like Title and book id(ISBN).
2. **BX-Users:-** Contains Users details with name and their ID.
3. **BX-Book-Ratings:-**Contains details of userid with ratings given to Books.

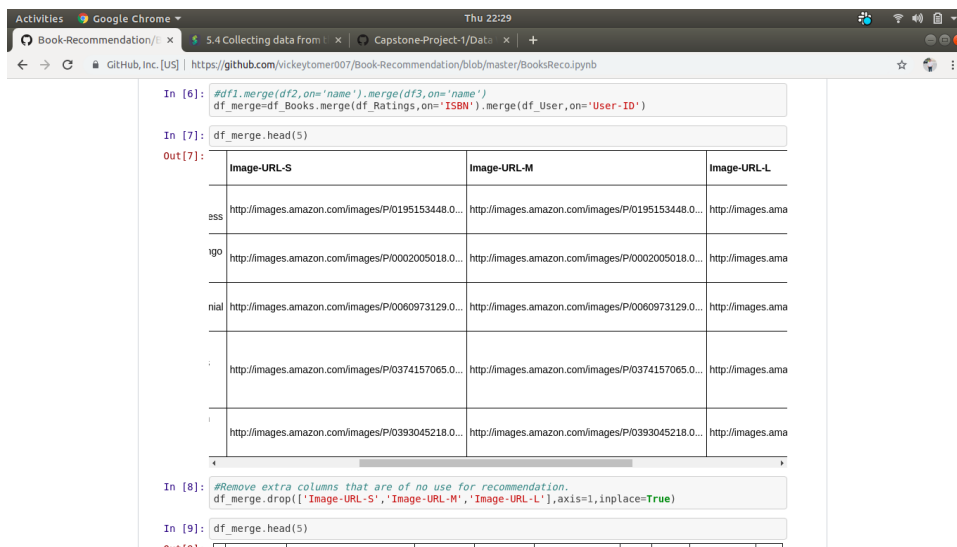
1.How did you deal with missing values, if any?

There are Two main steps for Data Cleaning are:-

- a. Handling missing values or null(NaN) values.
- b. Finding and Removing Outliers.

Here in my data set i have some of the columns that are of no use in Recommendation like:-

'Image-URL-S','Image-URL-M','Image-URL-L' as in below image.



```
In [6]: #df1.merge(df2,on='name').merge(df3,on='name')
df_merge=df_Books.merge(df_Ratings,on='ISBN').merge(df_User,on='User-ID')

In [7]: df_merge.head(5)
Out[7]:
```

	Image-URL-S	Image-URL-M	Image-URL-L
ss	http://images.amazon.com/images/P/0195153448.0...	http://images.amazon.com/images/P/0195153448.0...	http://images.ama
tp	http://images.amazon.com/images/P/0002005018.0...	http://images.amazon.com/images/P/0002005018.0...	http://images.ama
nial	http://images.amazon.com/images/P/0060973129.0...	http://images.amazon.com/images/P/0060973129.0...	http://images.ama
	http://images.amazon.com/images/P/0374157065.0...	http://images.amazon.com/images/P/0374157065.0...	http://images.ama
	http://images.amazon.com/images/P/0393045218.0...	http://images.amazon.com/images/P/0393045218.0...	http://images.ama

```
In [8]: #Remove extra columns that are of no use for recommendation.
df_merge.drop(['Image-URL-S','Image-URL-M','Image-URL-L'],axis=1,inplace=True)

In [9]: df_merge.head(5)
Out[9]:
```

I have removed all the above columns for more dataSet clarity.

```

V. W. Norton
&
company
http://images.amazon.com/images/P/0393045218.0...
http://images.amazon.com/images/P/0393045218.0...
http://

In [8]: #Remove extra columns that are of no use for recommendation.
df_merge.drop(['Image-URL-S', 'Image-URL-M', 'Image-URL-L'], axis=1, inplace=True)

In [9]: df_merge.head(5)
Out[9]:
   ISBN  Book-Title  Book-Author  Year-Of-Publication  Publisher  User-ID  Book-Rating  Location  Age
0  0195153448  Classical Mythology  Mark P. O. Morford      2002  Oxford University Press      2      0  stockton, california, usa  18.0
1  0002005018  Clara Callan      Richard Bruce Wright      2001  HarperFlamingo Canada      8      5  timmins, ontario, canada  NaN
2  0060973129  Decision in Normandy  Carlo D'Este      1991  HarperPerennial      8      0  timmins, ontario, canada  NaN
3  0374157065  Flu: The Story of the Great Influenza Pandemic...  Gina Bari Kolata      1999  Farrar Straus Giroux      8      0  timmins, ontario, canada  NaN
4  0393045218  The Mummies of Urunchi  E. J. W. Barber      1999  W. W. Norton & Company      8      0  timmins, ontario, canada  NaN

In [10]: #Grouped data on the basis of title ratings.
df_merge.groupby('Book-Title')['Book-Rating'].mean().sort_values(ascending=False).head()

```

2. While pivoting the data there are number of columns that has no ratings are coming as NaN for them i have replaced them with the 0 so that i easily can calculated correlation matrix.

```

In [94]: df_merge.head()
Out[94]:
   Book-Title  User-ID  Rating
0  A Light in the Storm: The Civil War Diary of Amelia Martin, Fenwick Island, Delaware 1861 (Dear America)  2  0.0
1  Earth Prayers From around the World: 365 Prayers, Poems, and Invocations for Honoring the Earth  8  0.0
2  Final Fantasy Anthology: Official Strategy Guide (Brady Games)  9  0.0
3  Good Wives: Image and Reality in the Lives of Women in Northern New England, 1650-1750  243  0.0
4  Goosebumps Monster Edition 1: Welcome to Dead House, Stay Out of the Basement, and Say Cheese and Die!  388  0.0
5  It Takes Two  2  0.0
6  LA Gallineta Roja/the Little Red Hen  8  0.0
7  Murder of a Sleeping Beauty (Scumble River Mysteries (Paperback))  9  0.0
8  Q-Space (Star Trek The Next Generation, Book 47)  243  0.0
9  Q-Zone (Star Trek The Next Generation, Book 48)  388  0.0

5 rows x 52830 columns

```

3. Dealing with outliers basically the book having less user ratings are outlier here in my case but this is not the right way may be the book published is very new but here i have selected only the

books having more the 100 user's ratings.

Activities Google Chrome Thu 22:39

SpringBoard/MovieReco: x 5.4 Collecting data from: x Capstone-Project-1/Dat... x Downloads/Books Reco/ x BooksReco x

← → local host:8888/notebooks/Downloads/Books%20Reco/BooksReco.jupyter

jupyter BooksReco Last checkpoint: an hour ago (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

A Light in the Storm: The Civil War Diary of Amelia Martin, Fenwick Island, Delaware, 1861 (Dear America)	NaN	4
Earth Prayers From around the World: 365 Prayers, Poems, and Invocations for Honoring the Earth	-0.002994	10
Final Fantasy Anthology: Official Strategy Guide (Brady Games)	NaN	4
Good Wives: Image and Reality in the Lives of Women in Northern New England, 1650-1750	-0.002994	10
Goosebumps Monster Edition 1: Welcome to Dead House, Stay Out of the Basement, and Say Cheese and Die!	NaN	9

```
In [103]: #corr_Euros.sort_values(['Correlation','num of ratings'],ascending=False).head()
corr_Euros[corr_Euros['num of ratings']>100].sort_values('Correlation',ascending=False).head()
```

Out[103]:

	Correlation	num of ratings
Book-Title		
Ender's Game (Ender Wiggins Saga (Paperback))	1.0	249
Paradise	1.0	172
Ishmael: An Adventure of the Mind and Spirit	1.0	162
Winter Moon	1.0	144
Possession : A Romance	1.0	141

In []: