Assignment 4 - Data preprocessing

July 3, 2019

0.1 Assignment 4

0.1.1 Problem

Download the IMDB dataset from https://www.kaggle.com/PromptCloudHQ/imdb-data (or from FormazioneOnLine).

The dataset contains 5,000 popular movies on IMDB from 2006 to 2016. The features included are: Title, Genre, Description, Director, Actors, Year, Runtime, Rating, Votes, Revenue, Metascore.

Perform a preprocessing analysis following the guidelines reported in this blog: http://www.developintelligence.com/blog/2017/08/data-cleaning-pandas-python/.

In other words, reinsert the commands described in the blog and learn how to: * Add a default value for the missing data; * Get rid of (delete) the rows that have missing data; * Get rid of (delete) the columns that have a high incidence of missing data;

0.1.2 Resolution

```
In [1]: import pandas as pd

        data = pd.read_csv('IMDB.csv')

        data.head()
```

Out[1]:

	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)
0	1	Guardians of the Galaxy	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...	2014	121
1	2	Prometheus	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott			
2	3	Split	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan			
3	4	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet			
4	5	Suicide Squad	Action,Adventure,Fantasy	A secret government agency recruits some of th...	David Ayer			

```
1  Noomi Rapace, Logan Marshall-Green, Michael Fa...  2012  124
2  James McAvoy, Anya Taylor-Joy, Haley Lu Richar...  2016  117
3  Matthew McConaughey,Reese Witherspoon, Seth Ma...  2016  108
4  Will Smith, Jared Leto, Margot Robbie, Viola D...  2016  123

   Rating   Votes  Revenue (Millions)  Metascore
0     8.1  757074              333.13       76.0
1     7.0  485820              126.46       65.0
2     7.3  157606              138.12       62.0
3     7.2   60545              270.32       59.0
4     6.2  393727              325.02       40.0
```

Now we must check if there are null values in the dataset:

```
In [2]: data.isnull().sum()

Out[2]: Rank                    0
        Title                   0
        Genre                   0
        Description             0
        Director                0
        Actors                  0
        Year                    0
        Runtime (Minutes)       0
        Rating                  0
        Votes                   0
        Revenue (Millions)    128
        Metascore              64
        dtype: int64
```

There are 128 null values in Revenue column and 64 in Metascore column. We can do 3 different things: * Replace all the null values woth ''; * Delete all the rows with any null values (or all null values); * Delete all the columns with any null values (or all null values).

We start to see the first possibility. Obviously we must reload the data.

```
In [3]: data['Revenue (Millions)'] = data['Revenue (Millions)'].fillna('')
        data['Metascore'] = data['Metascore'].fillna('')
        data.isnull().sum()

Out[3]: Rank                  0
        Title                 0
        Genre                 0
        Description           0
        Director              0
        Actors                0
        Year                  0
        Runtime (Minutes)     0
        Rating                0
        Votes                 0
```

```
Revenue (Millions)    0
Metascore             0
dtype: int64
```

Let's see the second option:

```
In [4]: data = pd.read_csv('IMDB.csv')
        data = data.dropna()
        data.isnull().sum()

Out[4]: Rank                  0
        Title                 0
        Genre                 0
        Description           0
        Director              0
        Actors                0
        Year                  0
        Runtime (Minutes)     0
        Rating                0
        Votes                 0
        Revenue (Millions)    0
        Metascore             0
        dtype: int64
```

Now we can see the third option:

```
In [5]: data = pd.read_csv('IMDB.csv')
        data = data.dropna(axis=1, how='any')
        data.isnull().sum()

Out[5]: Rank                  0
        Title                 0
        Genre                 0
        Description           0
        Director              0
        Actors                0
        Year                  0
        Runtime (Minutes)     0
        Rating                0
        Votes                 0
        dtype: int64
```