Investigate_a_Dataset

August 14, 2019

1 Project: Investigate a Dataset (IMDB Database)

1.1 Table of Contents

Introduction

What's the relation between movies and it's directors?: I choosed the movies database to check who are the most directors in number of movies directed and the top rated movies director. **How does movie production improved in the last ten years?** It's also intersting to see how movies production is improving in the last ten years in type of movie per year.

```
In [1]: # importing lib
        import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        %matplotlib inline

        #read CSV
        df = pd.read_csv('tmdb-movies.csv')

        #Check it
        df.shape
```

```
Out[1]: (10866, 21)
```

Data Wrangling

Is the related data clean?: I'll check for nulls and duplicated values in related data only:

1.1.1 General Properties

```
In [2]: # info
        df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10866 entries, 0 to 10865
Data columns (total 21 columns):
id                      10866 non-null int64
imdb_id                 10856 non-null object
popularity              10866 non-null float64
budget                  10866 non-null int64
revenue                 10866 non-null int64
original_title          10866 non-null object
cast                    10790 non-null object
homepage                2936 non-null object
director                10822 non-null object
tagline                 8042 non-null object
keywords                9373 non-null object
overview                10862 non-null object
runtime                 10866 non-null int64
genres                  10843 non-null object
production_companies    9836 non-null object
release_date            10866 non-null object
vote_count              10866 non-null int64
vote_average            10866 non-null float64
release_year            10866 non-null int64
budget_adj              10866 non-null float64
revenue_adj             10866 non-null float64
dtypes: float64(4), int64(6), object(11)
memory usage: 1.7+ MB
```

```
In [3]: df.shape
```

```
Out[3]: (10866, 21)
```

```
In [4]: sum(df.duplicated())
```

```
Out[4]: 1
```

1.1.2 Data Cleaning (Drop unnecessary data!)

```
In [5]: #Drop nnecessary data columns
        df.drop(columns=['homepage','tagline','keywords','cast','keywords','overview','runtime',
        df.shape
```

```
Out[5]: (10866, 14)
```

```
In [6]: # find duplicated entries after removing columns.
        sum(df.duplicated())
```

```
Out[6]: 1

In [7]: #Remove duplicated entries
        df.drop_duplicates(inplace=True)

In [8]: #Check for removal
        df.shape

Out[8]: (10865, 14)

In [9]: #Remove any entries without director name
        df.director.dropna(inplace=True)

In [10]: #Recheck
         df.shape

Out[10]: (10865, 14)
```

Exploratory Data Analysis.

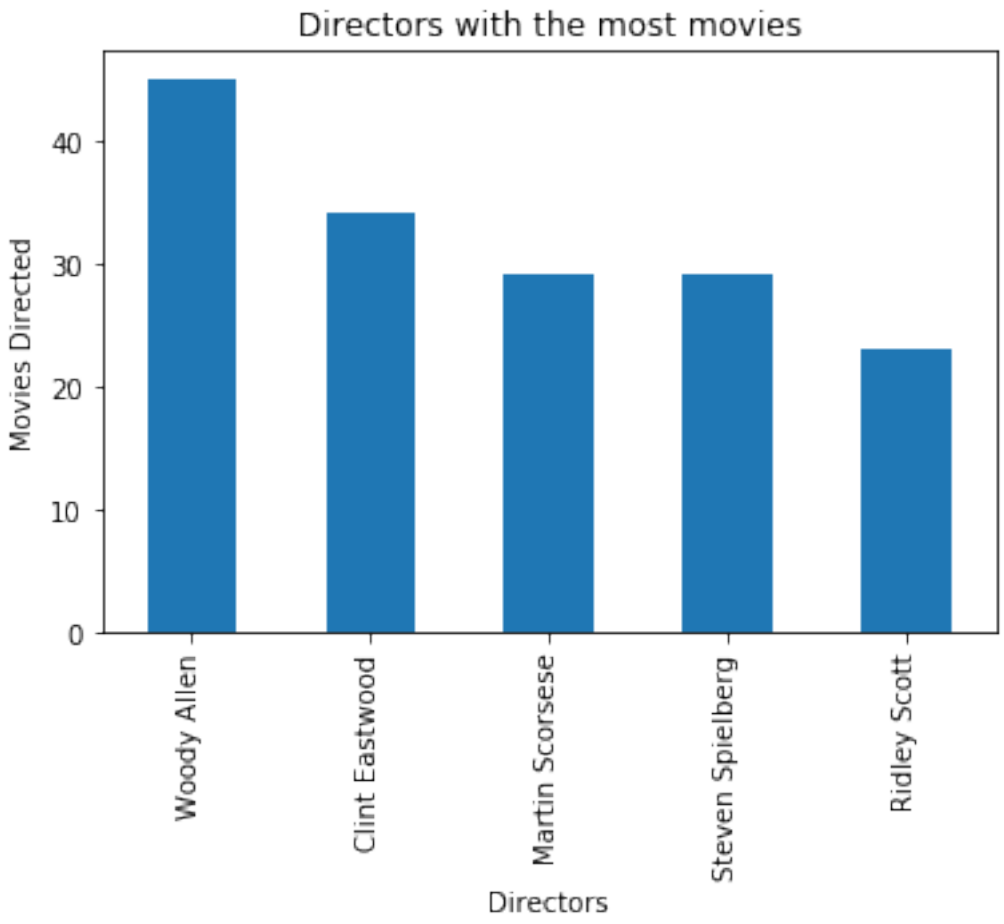
1.1.3 What's the relation between movies and it's directors?

```
In [11]: #Directors who have directed most movies:
         df['director'].value_counts().head()

Out[11]: Woody Allen         45
         Clint Eastwood      34
         Martin Scorsese     29
         Steven Spielberg    29
         Ridley Scott        23
         Name: director, dtype: int64

In [12]: df['director'].value_counts().head().plot(kind='bar')
         plt.title("Directors with the most movies")
         plt.xlabel("Directors")
         plt.ylabel("Movies Directed")

Out[12]: Text(0, 0.5, 'Movies Directed')
```




```
In [13]: #Top Rated Directors:
         top_dir = df['director'].groupby(df.vote_average).unique().tail()
         top_dir
```

```
Out[13]: vote_average
         8.5    [Curt Morgan, James Payne, Martin Scorsese|Mic...
         8.7                                          [David Mallet]
         8.8                        [Carl Tibbetts, Derek Frankowski]
         8.9                                 [Jennifer Siebel Newsom]
         9.2                                          [Mark Cousins]
         Name: director, dtype: object
```

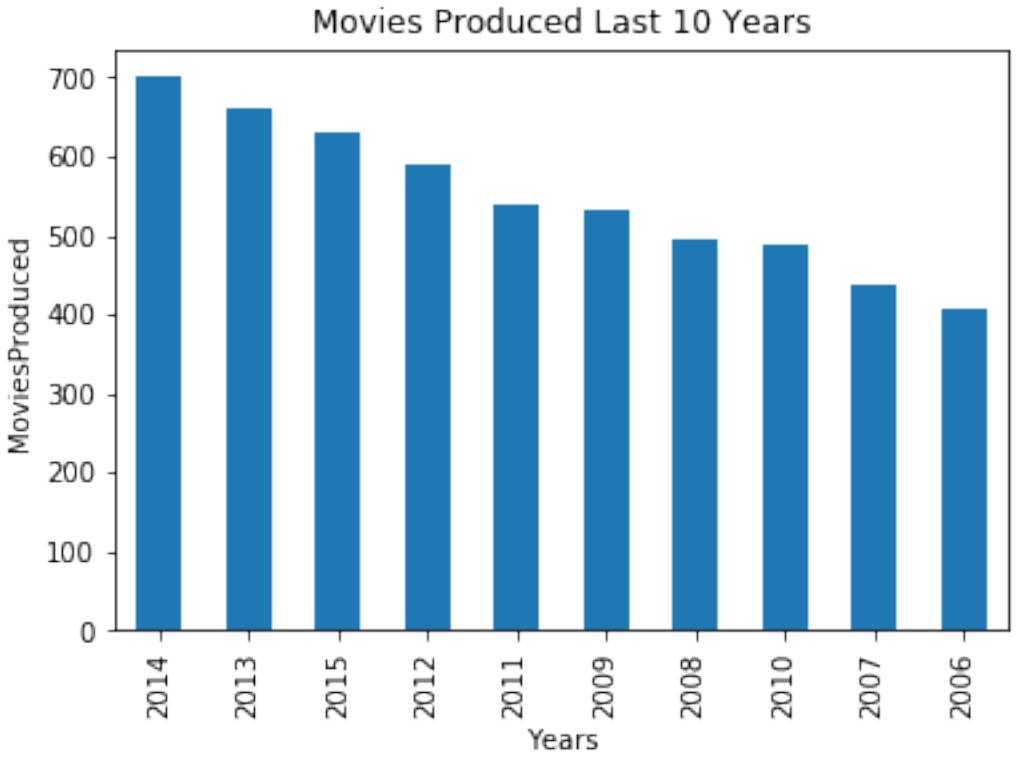
1.1.4 How does movie production improved in the last ten years?

```
In [14]: # How many movies produced per year in the last 10 years:

         df['release_year'].value_counts().head(10).plot(kind='bar')
         plt.title("Movies Produced Last 10 Years")
```

```
        plt.xlabel("Years")
        plt.ylabel("MoviesProduced")
```

```
Out[14]: Text(0, 0.5, 'MoviesProduced')
```



Conclusions

Summary: 1. The most movies were directed by "Woody Allen" with a high score of 45 movies but, unfortunately, none of them was the best-rated, The best-rated movies belong to Mark Cousins with the high score of 9.2 out of 10. 2. Movie production is improving over the years as we can see the number of movies making progress over years although it sometimes drops but, it still improves.