Hyperiondev

TASK

Exploratory Data Analysis on the Movies Data Set

Visit our website

Introduction

Summary of the data set

DATA CLEANING

The following techniques were carried out during data cleaning:

Looking into the data

- In order to get an idea of the nature of the data within the dataset, the dataset was inspected in order to identify the data types for each column and identifying the field that may not be of interest in the data analysis.

```
In [26]: movies_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4803 entries, 0 to 4802
Data columns (total 20 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   budget                4803 non-null   int64
 1   genres                4803 non-null   object
 2   homepage              1712 non-null   object
 3   id                    4803 non-null   int64
 4   keywords              4803 non-null   object
 5   original_language     4803 non-null   object
 6   original_title        4803 non-null   object
 7   overview              4800 non-null   object
 8   popularity            4803 non-null   float64
 9   production_companies  4803 non-null   object
 10  production_countries  4803 non-null   object
 11  release_date          4802 non-null   object
 12  revenue               4803 non-null   int64
 13  runtime               4801 non-null   float64
 14  spoken_languages      4803 non-null   object
 15  status                4803 non-null   object
 16  tagline               3959 non-null   object
 17  title                 4803 non-null   object
 18  vote_average          4803 non-null   float64
 19  vote_count            4803 non-null   int64
dtypes: float64(3), int64(4), object(13)
memory usage: 750.6+ KB
```

Checking for duplicates

- In order to get the best out of the analysis, the continuous and categorical features within the dataset are to be identified, to ensure that an appropriate strategy is implemented on the duplicating observations.

```
In [27]: #checking for duplicates
         movies_df.nunique()
```

```
Out[27]: budget                 436
         genres                1175
         homepage              1691
         id                    4803
         keywords              4222
         original_language       37
         original_title        4801
         overview              4800
         popularity            4802
         production_companies  3697
         production_countries   469
         release_date          3280
         revenue               3297
         runtime                156
         spoken_languages       544
         status                   3
         tagline               3944
         title                 4800
         vote_average            71
         vote_count            1609
         dtype: int64
```

Data reduction

- Some feature can be dropped if they do not add value to the analysis.

```
# code here
del_col_list = ['keywords', 'homepage', 'status', 'tagline', 'original_language', 'homepage', 'overview',
                'production_companies', 'original_title']

movies_df = movies_df.drop(del_col_list, axis=1)
movies_df.head()
```

	budget	genres	id	popularity	production_countries	release_date	revenue	runtime	spoken_languages	title	vote_average	vote
0	237000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	19995	150.437577	[{"iso_3166_1": "US", "name": "United States o...	2009-12-10	2787965087	162.0	[{"iso_639_1": "en", "name": "English"}, {"iso...	Avatar	7.2	
1	300000000	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	285	139.082615	[{"iso_3166_1": "US", "name": "United States o...	2007-05-19	961000000	169.0	[{"iso_639_1": "en", "name": "English"}]	Pirates of the Caribbean: At World's End	6.9	
2	245000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	206647	107.376788	[{"iso_3166_1": "GB", "name": "United Kingdom"...	2015-10-26	880674609	148.0	[{"iso_639_1": "fr", "name": "Fran\u00e7ais"},...	Spectre	6.3	
3	250000000	[{"id": 28, "name": "Action"}, {"id": 80,	49026	112.312950	[{"iso_3166_1": "US", "name": "United States o...	2012-07-16	1084939099	165.0	[{"iso_639_1": "en", "name": "English"}]	The Dark Knight Rises	7.6	

MISSING DATA

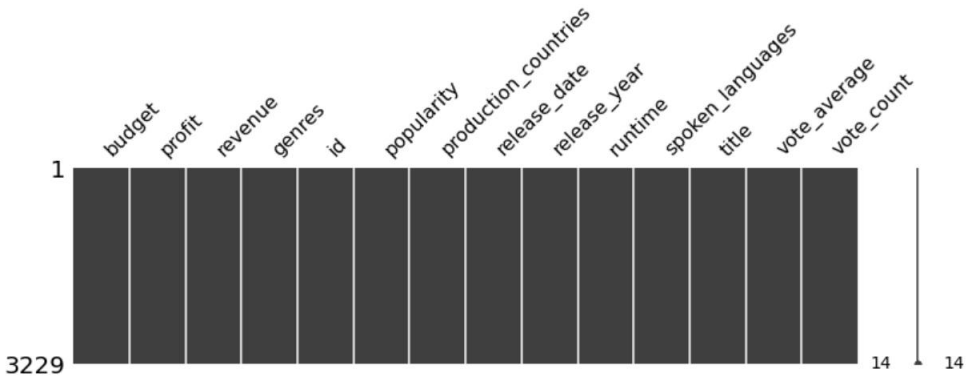
- There are missing observations within the data set.
- In the case of the movies dataset, there are an insignificant number of missing values, therefore it is best to remove the observations with missing values.

```
#Checking for the sum of missing values
movies_df.isnull().sum()
```

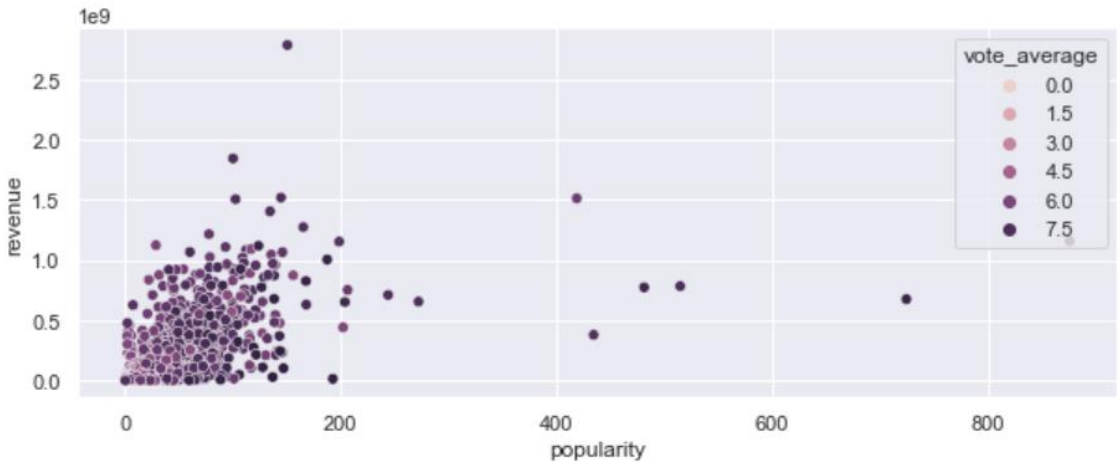
```
budget                  0
genres                  0
id                      0
popularity              0
production_countries    0
release_date            1
revenue                 0
runtime                 2
spoken_languages        0
title                   0
vote_average            0
vote_count              0
dtype: int64
```

```
# Visualise missing data
missingno.matrix(movies_df, figsize = (13,3))
```

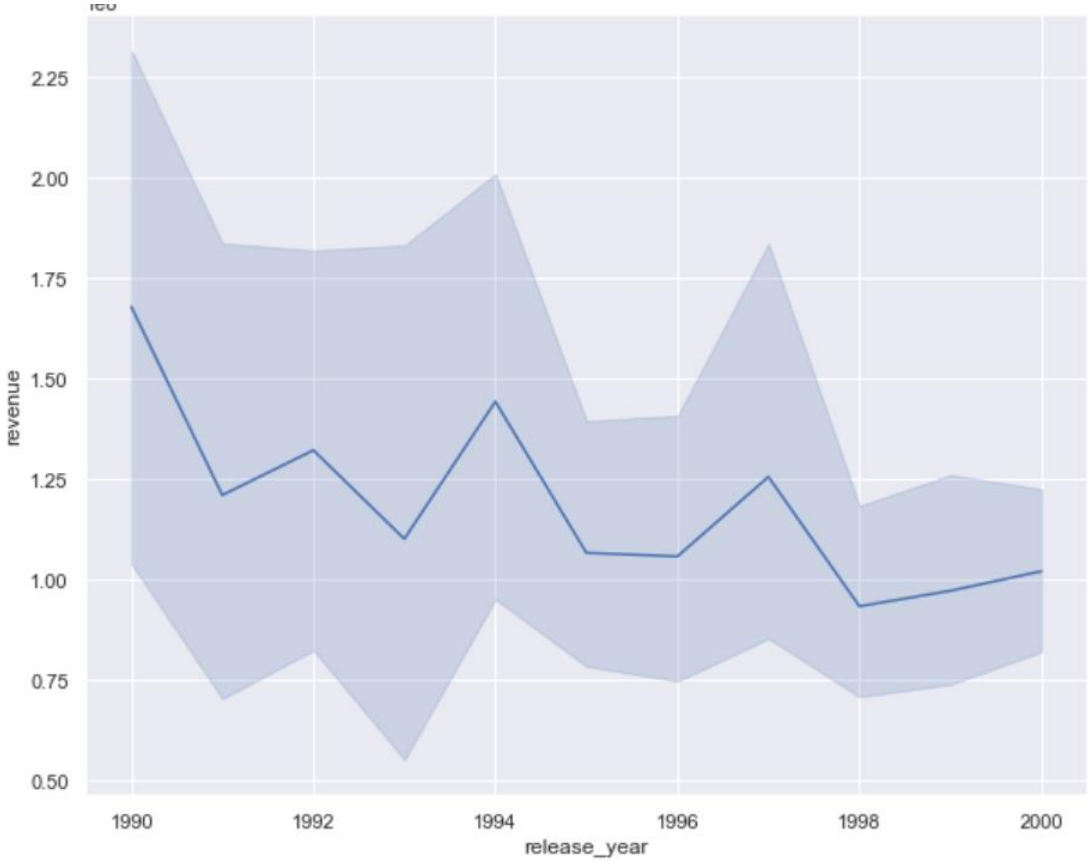
16]: <AxesSubplot:>

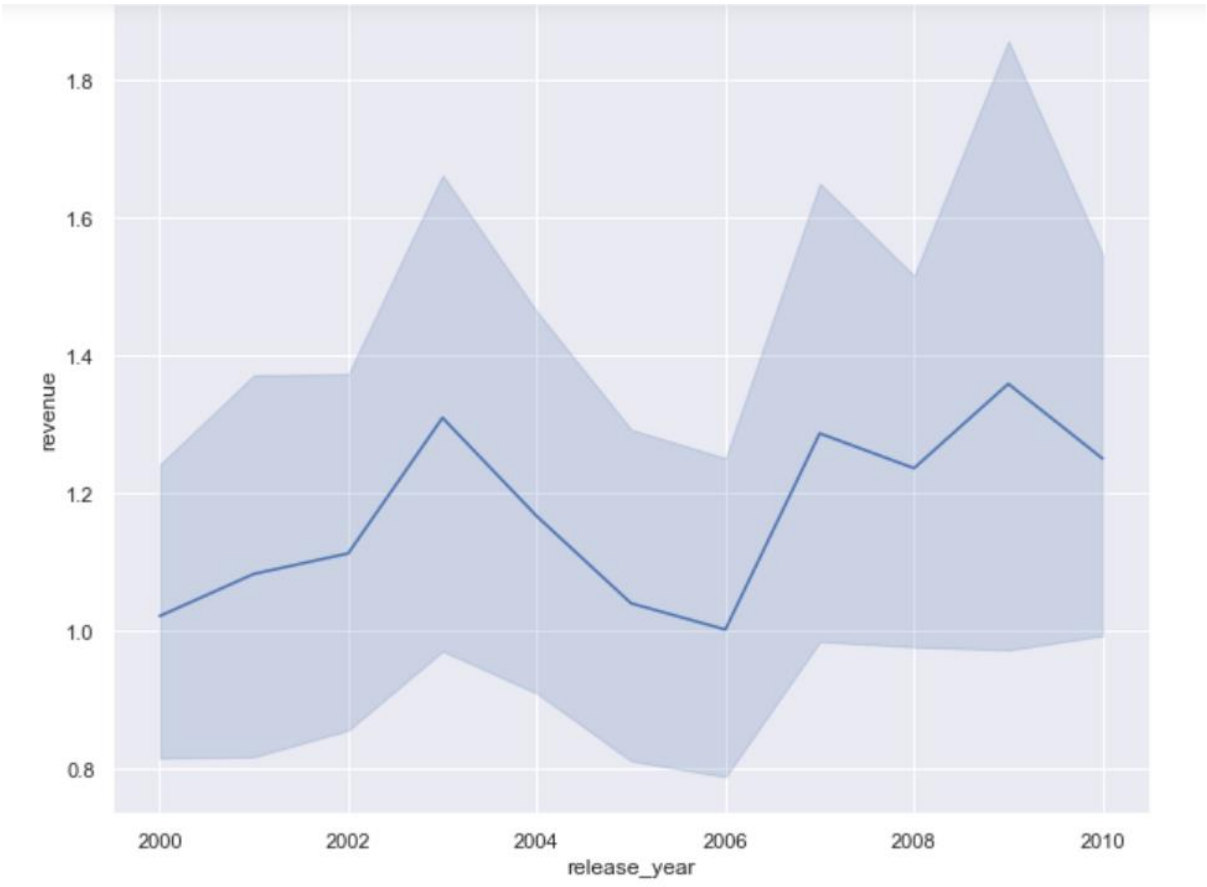


DATA STORIES AND VISUALISATIONS

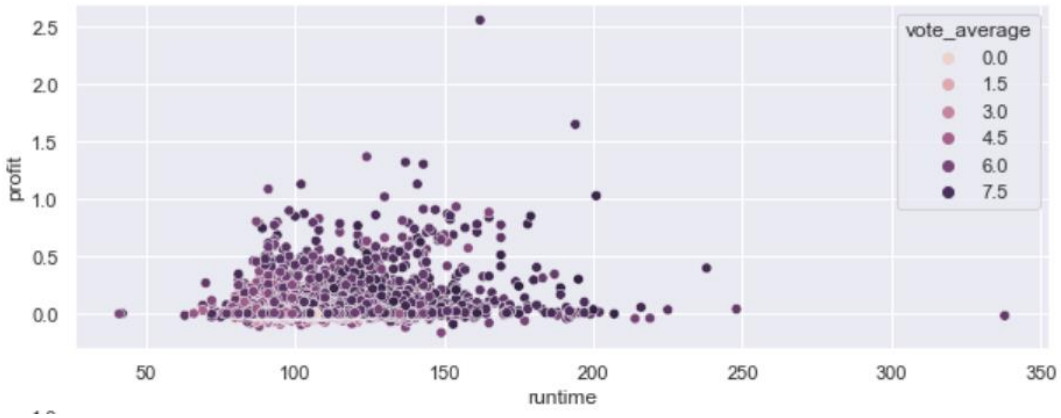


From the scatterplot above, there is a relationship between the popularity of the movie and the revenue generated. This implies that the more popular the movie, the more likely it is going to be sold out, hence an increase in the revenue.





The two line plots above were selected for release year from 1990 to 2010 and the corresponding revenue. For the last decade before the millennium, there has been a steady decline of revenue for the movies released and a steady increase from 2000 to 2010.



There is a correlation between the movie runtime and profit generated. From the scatter plot, the shorter the runtime, the greater the profit and also the greater the average vote.

THIS REPORT WAS WRITTEN BY : Ndatadzeyi B Chiota