TMDB MOVIES

Ahmed said

EG FWD PROFESSIONAL DATA ANALYSIS

Contents

Abstract

An analysis of a dataset containing information about 10,000 movies collected from the Movie Database (TMDb), including user ratings and revenue.

This report is for people who hadn't a programming background.

Ask questions

Q1: Which genres are most popular from year to year?

Q2: What kinds of properties are associated with movies that have high revenues?

Q3: what are more popular long movies or short movies?

Q4: who is the most successful director 1960-2015 highest revenue_adj (average)?

Q5: who (director) has the most popular movies from 2001-2015?

Q6: who (director) has the highest vote average from 2001-2015?

Q7: who (director) has the highest vote average from 2001-2015 for directors has participated in more than 3 movies?

Cleaning data

After exploring data, I found that ‘home page’ column has only 2936 value, so many missing values in this column so I decided to drop it from dataset.

“production_companies”, “keywords”, “tagline” and “director” have some missing values I dropped only null rows.

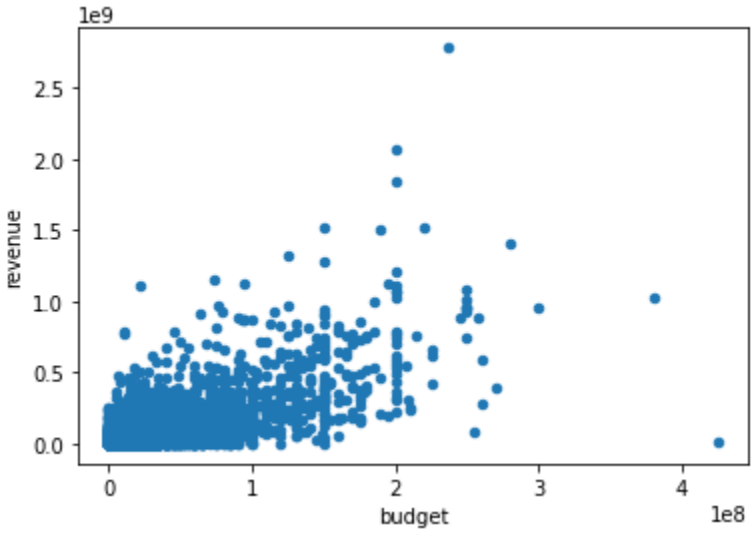
This was the final data set that I worked with:

#	Column	Non-Null Count	Dtype
0	id	7030 non-null	int64
1	imdb_id	7030 non-null	object
2	popularity	7030 non-null	float64
3	budget	7030 non-null	int64
4	revenue	7030 non-null	int64
5	original_title	7030 non-null	object
6	cast	7030 non-null	object
7	director	7030 non-null	object
8	tagline	7030 non-null	object
9	keywords	7030 non-null	object
10	overview	7030 non-null	object
11	runtime	7030 non-null	int64
12	genres	7030 non-null	object
13	production_companies	7030 non-null	object
14	release_date	7030 non-null	object
15	vote_count	7030 non-null	int64
16	vote_average	7030 non-null	float64
17	release_year	7030 non-null	int64
18	budget_adj	7030 non-null	float64
19	revenue_adj	7030 non-null	float64

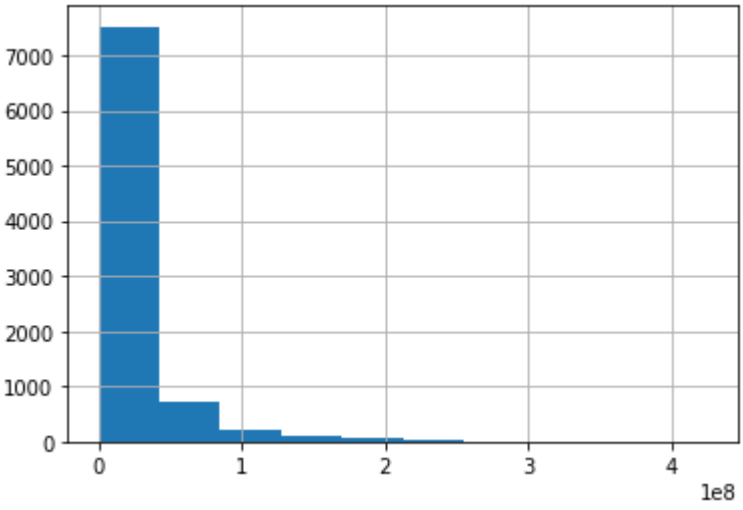
Figure 1:clean dataset

Exploratory Data Analysis

At first I noticed a lot of zero values in revenue and budget



The budget distribution shows that more than half of the values are zero



To answer the first question Which genres are most popular from year to year?

I try to reduce the unique values of genres because genres have more than 1000 unique genre

Because of old format example: action|drama|music|…etc.

I used the first genre of each one example: action, that helped me to reduce unique values to 20 value.

Then I plot the next two plots to show the answer

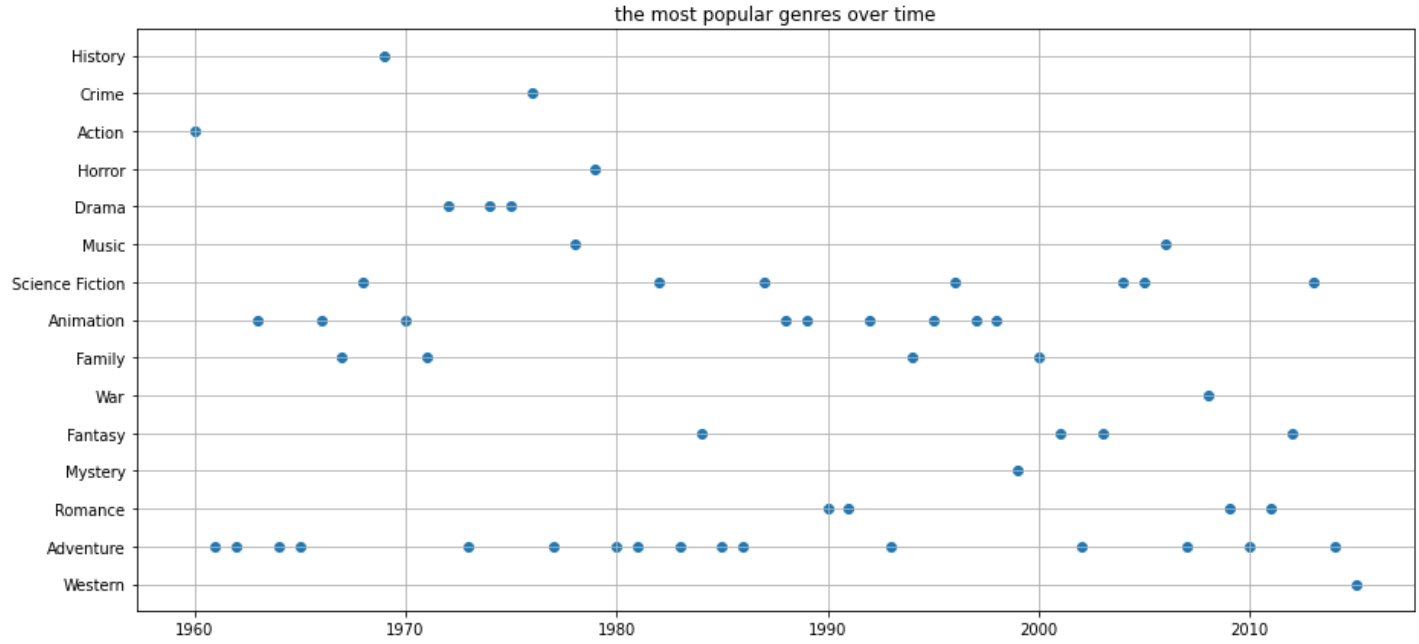


Figure 2:Answer q1 scatter plot

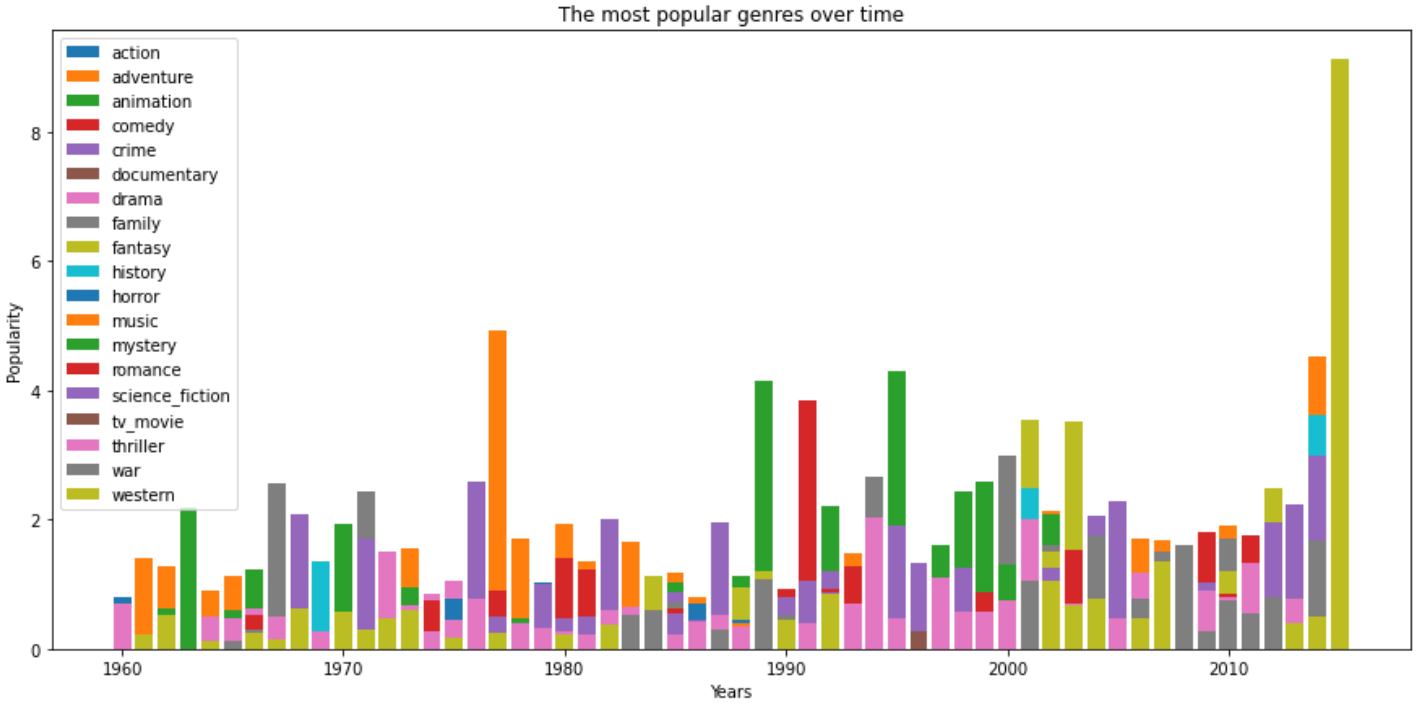
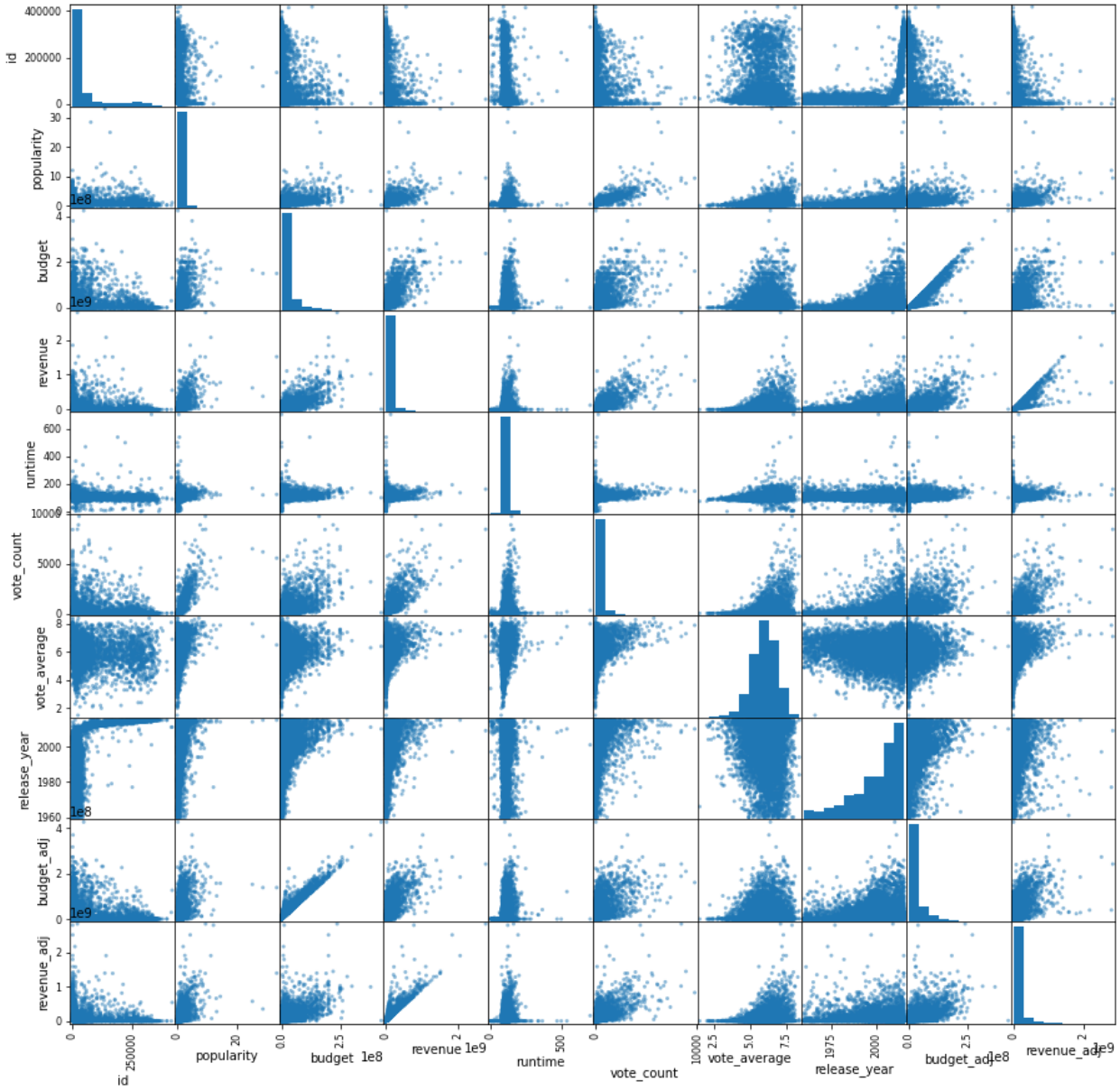


Figure 3:Answer q1 bar chart

Q2: answer

Next plot shows matrix of scatter plot to all attributes which answer the second que stion

I think that budget and popularity have a positive correlation with revenue.



Q3: answer

To answer the third question, I calculated the mean popularity for long movies (mo vies have runtime greater than the median runtime) and for short movies

Long movies average popularity = `1.054`

Short movies average popularity = `0.62`

Long movies are more popular than short, after calculating the evaluation in the sa me way, the same result came out.
Long movies average rating = `6.284`
Short movies average rating= `5.761`

Q4: answer

A simple answer to an easy question I calculated average revenue_adj for each director and selected maximum value
It turns out that “Clyde Geronimi|Hamilton Luske|Wolfgang Reitherman" are the most successful director 1960-2015 highest revenue_adj (average)
With 157,481,473,9 $
Note:
the three directors together in the first place ~~not the first three places~~

Q5: answer

Colin Trevorrow (director) has the most popular (average) movies from 2001-2015

With popularity average: 16.7

Q6: answer

Damien Chazelle (director) has the highest vote average from 2001-2015

With average 8.2

But! He has only one movie and Other directors have participated in more than 10

So I filtered the directors who participate more than 3 movies.

Q7: answer

Christopher Nolan (director) has the highest vote average from 2001-2015 for directors has participated in more than 3 movies

With average 7.6

He turned out to be the highest of all time

With average 7.6375