

**PROJECT 03**

Shape

Description automatically generated with low confidence

Send me E-mail

Shape

Description automatically generated with low confidence

LinkedIn Profile

Shape

Description automatically generated with low confidence

YouTube Channel

Shape

Description automatically generated with low confidence

GitHUb

**By:**

**Mehrdad Mansourdehghan**

1. **Data Preprocessing**
2. **Exploratory Data Analysis (EDA)**
3. **Correlation Analysis**
4. **Data Visualization**

**IMDB MOVIES**

* 1. **Project Goals**

In this project, I’m going to preprocess and clean data, run exploratory data analysis (EDA) and find descriptive statistical measurements, finding the correlation among variables as well as analysis data with making an interactive dashboard for a dataset which contains nearly 7000 movies’ data.

* 1. **Language, libraries, tools:**

**Language**: Python, DAX

**Libraries**: Pandas, NumPy, Seaborn, Matplotlib, Regular Expression, StatsModels, SweetViz

**IDE**: Jupyter Notebook

**Application**: Microsoft Excel, Microsoft PowerBI

* 1. **Data**

There are 6820 movies in the dataset (220 movies per year, 1986-2016). Each movie has the following attributes:

**• budget:** the budget of a movie. Some movies don't have this, so it appears as 0

**• company:** the production company

**• country:** country of origin

**• director:** the director

**• genre:** main genre of the movie.

**• gross:** revenue of the movie

**• rating:** rating of the movie (R, PG, etc.)

**• released:** release date (YYYY-MM-DD)

**• runtime:** duration of the movie

**• score:** IMDb user rating

**• votes:** number of user votes

**• star:** main actor/actress

**• writer:** writer of the movie

**• year:** year of release

Here you can follow all steps that were taken in this project. Moreover, the codes in Jupyter notebook are exactly based in these processes.

* 1. **Import Packages**

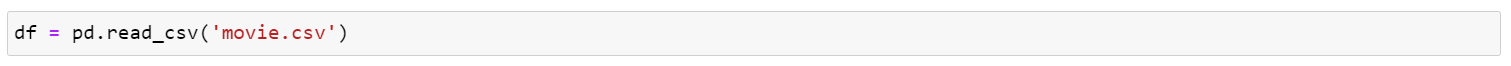
First, I import needed packages. I use Pandas because the structure of the dataset is in tabular format. Also, I use NumPy to have this opportunity to run numerical analysis much easier. Finally, I use Seaborn and Matplotlib for visualization during EDA process. I set the size of all figures and visualizes in this project, at the beginning. Also, the style of visualization is “ggplot” based.

Shape, rectangle

Description automatically generated

* 1. **load dataset:**

Then I load the dataset using with pandas



* 1. **look at data:**

just to make sure the data is imported correctly, let’s see its 3 first rows:

Table

Description automatically generated

* 1. **handle missing values:**

I create a for loop to iterate among all columns to realize whether they have null values or not. I’m looking for the number of null values in every single column as well as the percentage of null values.

Graphical user interface, text, application, email

Description automatically generated

The result shows we must have a different approach to handling null values. For some columns that the percentage of null values are less than 5%, we can drop the records, and for those have more than 5%, we should impute. Let’s drop the null values in “rating, released, score, votes, writer, star, country, gross, company and runtime”.

Chart, scatter chart

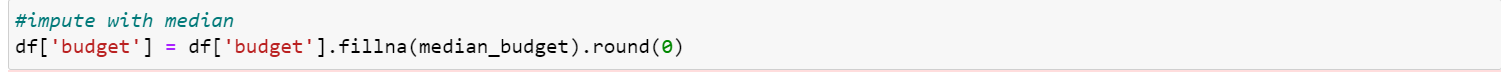
Description automatically generated

And now we should impute null values for “budget”. But before doing this, we must make share, about distribution shape of this column to see whether it’s right-skewed or left-skewed. It can be helpful when we want to decide choosing mean or median for imputing.

Graphical user interface, text

Description automatically generated with medium confidence

So, we choose median:



Finally, we check the null values again:

Graphical user interface, text, application, email

Description automatically generated

* 1. **Sanity checks on “Year”:**

Since we have two columns for the year, we must check the sanity (correctness) to realizer whether the year column is based on the release date or not. By the way, we want to replace the new year column (that is extracted from release date) with the old year column.

Graphical user interface, text, application

Description automatically generated

Now we can drop the old year column.



* 1. **Handle duplicate rows:**

Now we should handle duplicate rows. Since all values might be same, we just we need to check whether there are two rows that all values in all columns are the same or not.

Rectangle

Description automatically generated with low confidence

The result shows fortunately we don’t have duplicate rows.

In this section we do descriptive statistical analysis to know data much more. By doing this, we can realize their distribution, central tendency measurement, the dispersion measurement and shape of the data. First, we import SweetViz package that is very helpful in descriptive statistical analysis. Moreover, we write some functions for Matplotlib and Seaborn to add more information about the statistical analysis by drawing distribution plots (quantitative variables) and bar charts (qualitative variables).

* 1. **Qualitative Variables**

Here we analyze qualitative variables, and we see each label in every single variable account for the highest frequency. Then, we can see the statistical interpretation for categorical variables:

|  |  |  |
| --- | --- | --- |
| rating | Table  Description automatically generated | Chart, bar chart  Description automatically generated |
| The “R” , “PG-13” and “PG” account for the most frequent rating. | | |
| genre | Table  Description automatically generated | Chart, bar chart  Description automatically generated |
| The “Comedy”, “Action” and “Drama” were the types of genres that directors made. | | |
| country | Table  Description automatically generated | Chart  Description automatically generated |
| The most movies are made in “US” and “United Kingdom”. | | |
| year | Table  Description automatically generated | Table  Description automatically generated |
| The in “2015”, “2011” and “2007” companies made more movies compared to other years. | | |
| runtime | Table  Description automatically generated | Chart, funnel chart  Description automatically generated |
| The most movies’ durations long between “91 to 120” minutes. | | |
| director | A picture containing graphical user interface  Description automatically generated | |
| The “Woody Allen”, “Clint Eastwood” and “Steven Spielberg” made many more than other. | | |
| writer | A picture containing graphical user interface  Description automatically generated | |
| The most movies are made with the scenarios that are written by “Woody Allen”, “Stephen King”, “Luc Besson” and “John Hughes”. | | |
| star | Table  Description automatically generated with medium confidence | |
| The “Nicolas Cage”, “Tom Hanks” and” “Robert De Niro” played in the most movies compared to other actresses or actors. | | |
| companu | Table  Description automatically generated with medium confidence | |
| The most movies are made by “Universal Pictures”, “Warner Bros” and “Columbia Pictures”. | | |

* 1. **Quantitative Variables**

Now, let’s analyze descriptive statics for quantitative variables. In this section we can find central tendency, dispersion, and shape measurements. Then we draw the distribution plot to compare with normal distribution. Also, we can see how those variables are related to other variables.

Text

Description automatically generated

Now we can see the statistical interpretation for quantitative variables:

|  |  |
| --- | --- |
| *score* | |
| Table  Description automatically generated | Chart, line chart, histogram  Description automatically generated |
| *The average score of movies is 6.40.*  *Half of the movies scored less than 6.5*  *The difference between the maximum and minimum score is 7.4.*  *25% of the movies are scored less than or equal to 5.8*  *75% of the movies are scored less than or equal to 7.1*  *50% of the movies are scored between 5.8 and 7.1*  *The distribution is almost normal.* | |
| *votes* | |
| Graphical user interface, text, application, table  Description automatically generated | A picture containing chart  Description automatically generated |
| *The average votes of movies is 0.1M*  *Half of the movies voted less than 0.034M*  *The difference between the maximum and minimum votes is 2.4M.*  *25% of the movies are voted less than or equal to 0.010M*  *75% of the movies are scored less than or equal to 0.096M*  *50% of the movies are scored between 0.010M and 0.096M*  *The distribution is considerably right skewed (big outlier).* | |
| *budget* | |
| Graphical user interface, text, application, table  Description automatically generated | Chart  Description automatically generated |
| *The average budget for producing movies is 32.2M.*  *Half of the movies are made with the amount of budget less than 21.8M*  *The difference between the maximum and minimum budget is 356M.*  *25% of the movies are made with budget less than or equal to 14M*  *75% of the movies are made with budget less than or equal to 33M*  *50% of the movies are made with budget between 14M and 33M*  *The distribution is considerably right skewed (big outlier).* | |
| *gross* | |
| Graphical user interface, text, application  Description automatically generated | Chart  Description automatically generated |
| *The average gross of movies is 0.1B*  *Half of the movies could make money less than 0.020B*  *The difference between the maximum and minimum gross is 2.8B*  *25% of the movies could bring benefit less than or equal to 0.0046B*  *75% of the movies could bring benefit less than or equal to 0.0764B*  *50% of the movies could bring benefit between 0.0046B and 0.0764B*  *The distribution is considerably right skewed (big outlier).* | |
| *runtime* | |
| Graphical user interface, table  Description automatically generated | Chart  Description automatically generated with medium confidence |
| *The average duration of movies is 107 minutes*  *Half of the movies last less than 104 minutes*  *The difference between the maximum and minimum duration is 303 minutes.*  *25% of the movies last less than or equal to 95 minutes*  *75% of the movies last less than or equal to 116 minutes*  *50% of the movies last between 95 minutes and 116 minutes*  *The distribution is considerably right skewed (big outlier).* | |

* 1. **Maximum & Minimum Values:**

We can also, see the minimum and maximum values for each quantitative variable:

Text

Description automatically generated

Text, email

Description automatically generated

Table

Description automatically generated

Graphical user interface, text, table

Description automatically generated

In this section, I’m going to see whether there is relationship between variables with gross. Since the gross is very crucial for each producer, it makes sense that we see how many other factors are correlated to gross. For doing this, we should consider two different ways. One way is correlation between numeric variables with gross, and the second way is correlation between categorical variables with gross. For the first one, we use Pearson correlation and for the second one, we use ANOVA test.

* 1. **Pearson Correlation**

First, based on OLS method, we analyze the correlation and regression between variables on plot.

Graphical user interface, text, application, email

Description automatically generated

Chart, scatter chart

Description automatically generated

For all numeric variables, there is a positive relationship between them and gross. However, this relationship between gross with votes and budget are strong, but with scores are weak. For knowing the exact correlation, we can use Pearson function:

A picture containing table

Description automatically generated

As we can see, we can consider the budget as the most numeric variable for having more gross, and if we budget more, we will have more gross. Even though having a higher score and runtime of a movie can increase the gross, it is not very considerable.

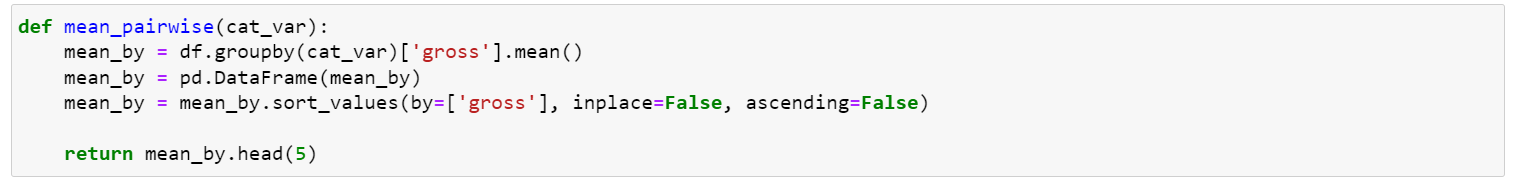
* 1. **ANOVA Test**

For categorical variables, first we should see whether a particular categorical variable impacts gross or not (it is significant or not). For doing this, we run ANOVA test to compare the meaningful difference between means. After that, for the variables that are significant, we run pairwise descriptive analysis for all labels in the particular categorical variable and then compare their impacts on the gross.

Graphical user interface, text

Description automatically generated

According to results, we can come up that name is not significant variables to explain how a movie can make gross, because the p-value is more than 0.05 and we cannot reject null hypothesis. So, in the next step, we want to see for each label in the above categorical variables, which of them has the most impact on gross. Thus, I make s function to calculate the mean for each label in every single categorical variable, and then shows just first top positive influencer:



|  |  |  |  |
| --- | --- | --- | --- |
| Table  Description automatically generated | Table  Description automatically generated | Table  Description automatically generated | Table  Description automatically generated |
| *The most gross can be gained by rate “G”* | *The most gross can be gained by genre “Animation”* | *The most gross can be gained by director “A. Russo”* | *The most gross can be gained by writer “Ch. Markus”* |
| Table  Description automatically generated | Table  Description automatically generated | Table  Description automatically generated | Table  Description automatically generated |
| *The most gross can be gained by star “D. Glover”* | *The most gross can be gained by country “Malta”* | *The most gross can be gained by company “Marvel Studios”* | *The most gross can be gained by year “2020”* |

In this section we want to analyze data based on visualization on “data\_cleaned” file, because I believe the best way to analyze the data is in visualized way. So, by doing this we can answer some ad-hoc questions that might be asked in daily-basis business. So far, we have a good image of the data, and we can help managers or users who are willing to have insight about the data. I use Microsoft Power BI to create an interactive dashboard.

* 1. **Dashboard**

The dashboard contains so many elements that indicate information about the data. In the top left, we have two carousel slicers we can use to filter data based on year and movie’s length. Next to them, there are some cards where we can find some statistical information about movie(s), and on the top right, we have a filled map to have better view about the geographical distribution of movies all around the world. In the middle of the report, we have a trend chart that indicates the total gross for every year and you can use the small carousel (bottom of the chart) to filter the years. Also, the table shows the performance of the top-5 movies in terms of how much they can benefit and earn money per minute. Finally, at the bottom of the dashboard, we can see some bar charts that show top-5 movie, active director, stars, and the best genre (in terms of number of produced movies). Having said that, if you hover mouse on the top-5 movies, you can see their more information there.

A screenshot of a computer

Description automatically generated with medium confidence

* 1. **Movie Table**

Having a whole table is always good practice to give this chance to users for iterative among data. In the next page, there is a table that show list of all movies with some information about them. Moreover, there are some filters that can help you to get closer to your target.

Graphical user interface

Description automatically generated

**- EN­D -**