# Exploratory Data Analysis

Unlocking insights with Microsoft

Eric Lemiso
Lekishon

# What’s EDA?

*"**Exploratory Data Analysis** is a preliminary approach to data analysis for summarizing, visualizing, and understanding data patterns."*

**Contribution by Microsoft to EDA**

***. Power BI***

***.Microsoft Excel***

***.SQL Server***

# Importance of EDA to the film industry;

***1.Audience Insights:*** Understand viewer preferences, create engaging content.

2.***Box Office Success:*** Data-driven decisions, maximize financial returns.

3.***Cost Efficiency:*** Identify resource-saving opportunities, optimize production.

4.***Competitive Advantage***: Spot trends, stay ahead competitively.

# Data Sources

## IMDb Title Basics

**Data Size:** Large dataset with information on movie titles, release years, genres, and more.

**Structure:** Structured data with multiple columns, ideal for querying and analysis.

**Format:** CSV

## IMDb Title Ratings

**Data Size:** Contains ratings and votes for movies, often a supplementary dataset to IMDb Title Basics.

**Structure:** Structured data with ratings and votes, suitable for joining with other IMDb datasets.

**Format:** CSV

## bom.movie_gross

**Data Size:** Provides box office earnings and financial data for movies.

**Structure:** Structured data with information on earnings, release dates, and studios.

**Format:** CSV

# Data Cleaning and it’s importance

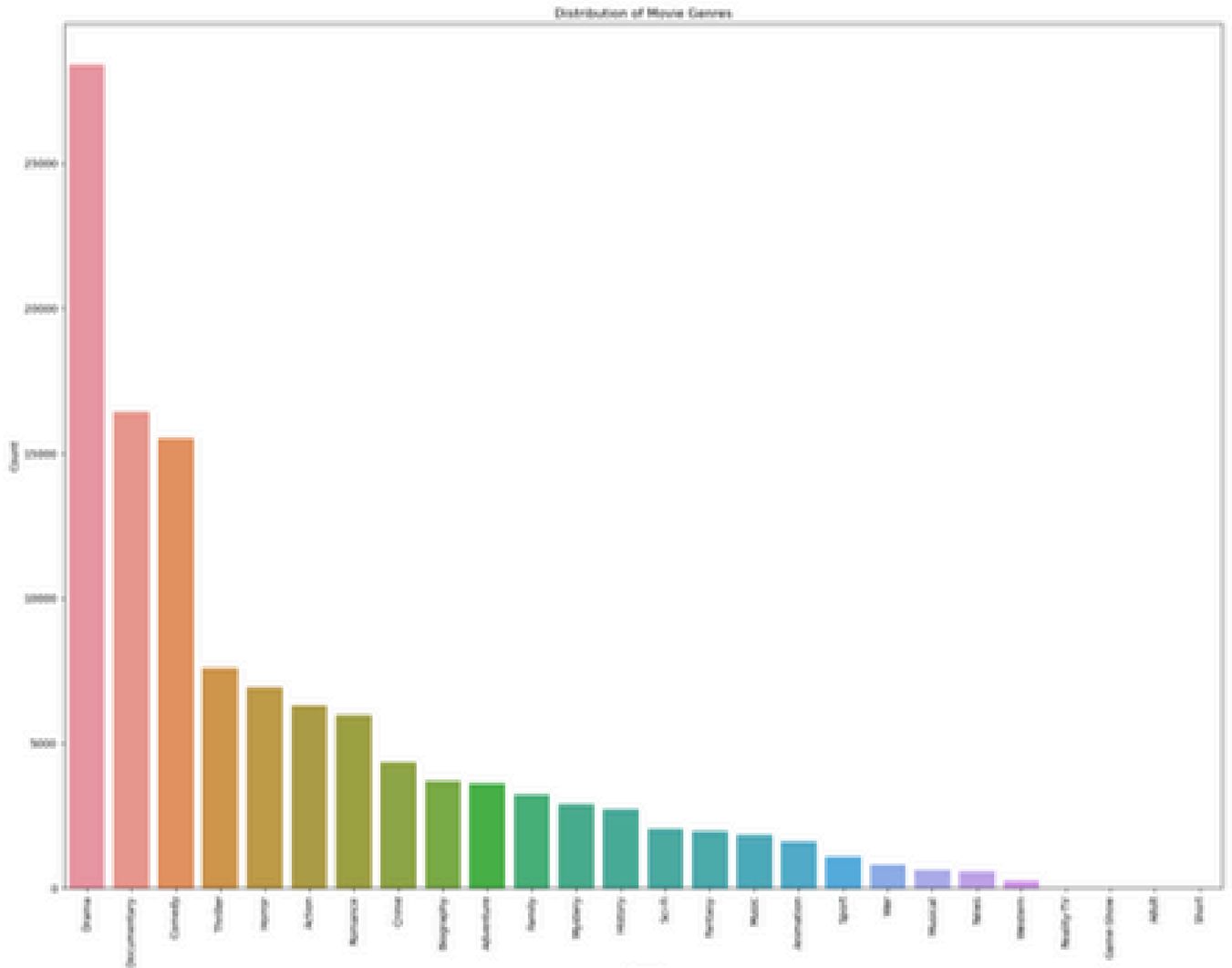
1.Accuracy and Quality: Clean data ensures accurate and high-quality analysis, leading to reliable insights and decisions.

2.Outlier Identification: Data cleaning helps identify and handle outliers that can distort results during analysis.

3.Consistency and Compatibility: Standardizing data formats and structures ensures compatibility and consistency for meaningful comparisons.

4.Effective Visualization: Clean data enables effective data visualization, enhancing the interpretability of EDA results for stakeholders.

# Distribution of movie genres




1 **Drama**

2 **Documentary**

3 **Comedy**

4 **Thriller**

5 **Horror**

# Descriptive Analysis

| | start_year | runtime_minutes | averagerating | numvotes | domestic_gross | foreign_gross |
|---|---|---|---|---|---|---|
| count | 2973.000000 | 2973.000000 | 2973.000000 | 2.973000e+03 | 2.973000e+03 | 2.969000e+03 |
| mean | 2013.786747 | 107.320552 | 6.461789 | 6.282036e+04 | 3.041947e+07 | 4.742381e+07 |
| std | 2.461329 | 19.906558 | 0.997358 | 1.263703e+05 | 6.689915e+07 | 1.151021e+08 |
| min | 2010.000000 | 40.000000 | 1.600000 | 5.000000e+00 | 0.000000e+00 | 0.000000e+00 |
| 25% | 2012.000000 | 94.000000 | 5.900000 | 2.486000e+03 | 1.250000e+05 | 0.000000e+00 |
| 50% | 2014.000000 | 105.000000 | 6.600000 | 1.375500e+04 | 1.900000e+06 | 2.300000e+06 |
| 75% | 2016.000000 | 118.000000 | 7.100000 | 6.600800e+04 | 3.190000e+07 | 3.360000e+07 |
| max | 2019.000000 | 272.000000 | 9.200000 | 1.841066e+06 | 7.001000e+08 | 9.464000e+08 |

# What’s the Importance of Descriptive Analyis

**Data Summarization**
- **Example: Calculating the mean (average) domestic gross revenue of a set of movies to understand their financial performance.**

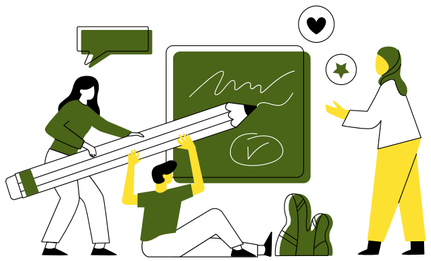
**Data Distribution:**
- **Example: Creating histograms of movie ratings to visualize the distribution of audience opinions.**

**Data Spread and Variability:**
- **Example: Computing the standard deviation of runtime to assess how widely runtime values vary across movies.**

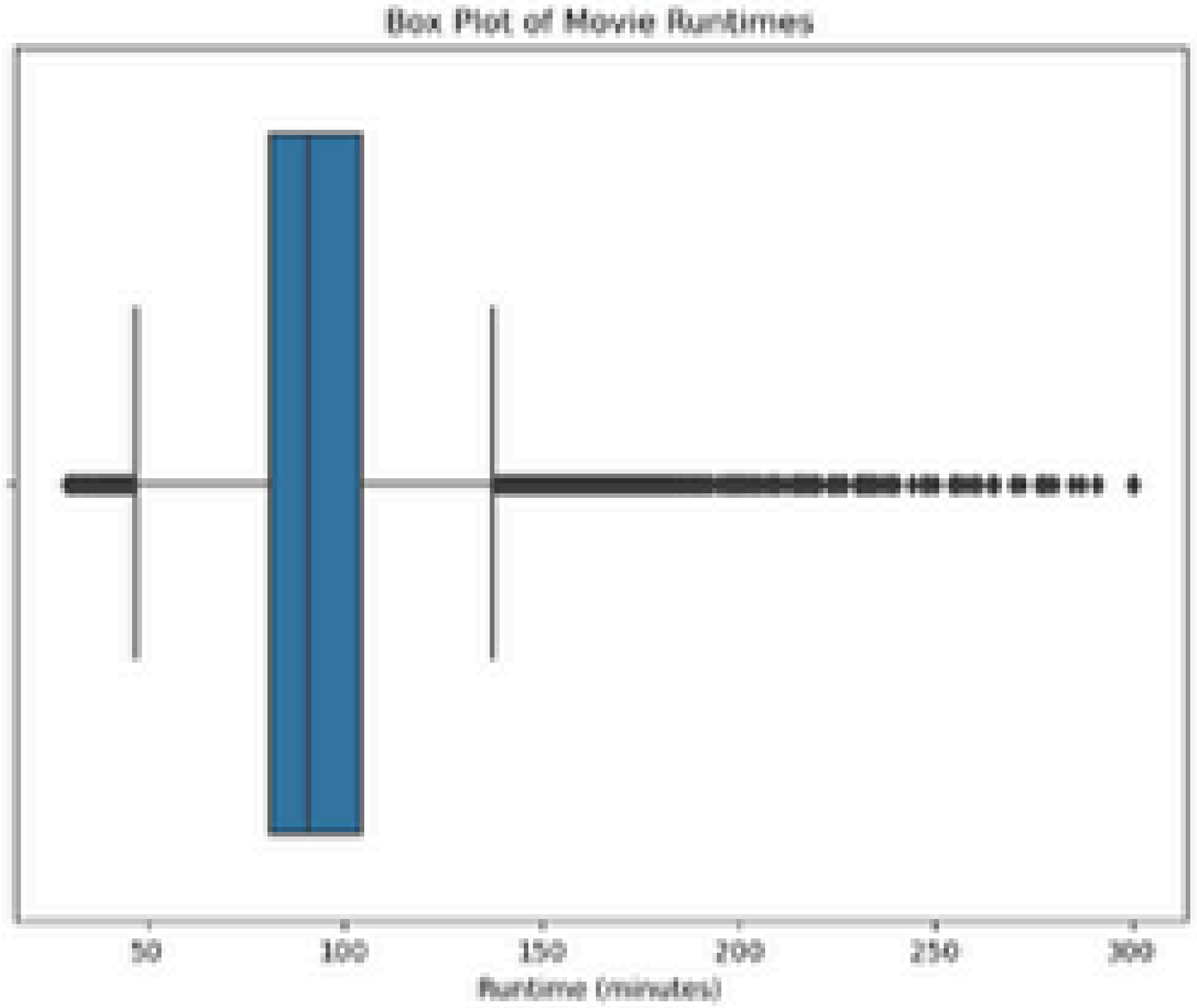
**Data Relationships:**
- **Example: Using correlation coefficients to examine the relationship between movie budgets and box office earnings.**

# Handling Outliers

***Do we have extreme values in the runtime column? let's handle them...***

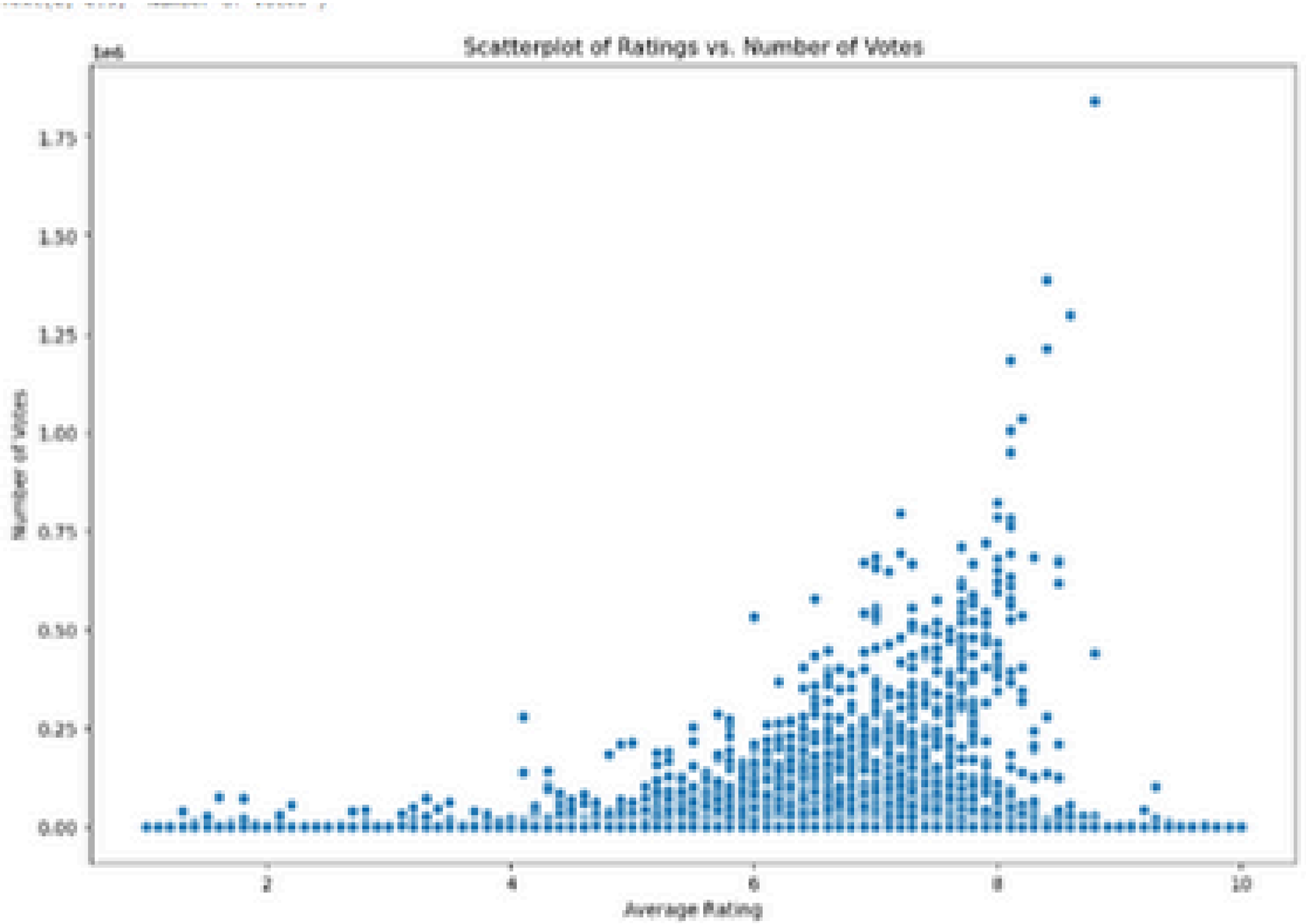


***What is the best runtime? 90 minutes, you got it right...***



# Bivariate Analysis

There is a positive correlation between ratings and the number of votes.



# Key Insights and Recommendations

- ***Optimal Runtime***
- ***Engagement, Popularity***
- ***Box Office Success***
- ***Runtime and Earnings***

# Recommendations

- Consider producing comedy films with a runtime of 90 minutes as they are well-received and have high popularity.
- Thrillers and horrors can be considered as the second and third genres consecutively.
- Utilizing a 90-minute runtime can significantly influence audience engagement, translating to higher domestic gross.

# Thank you