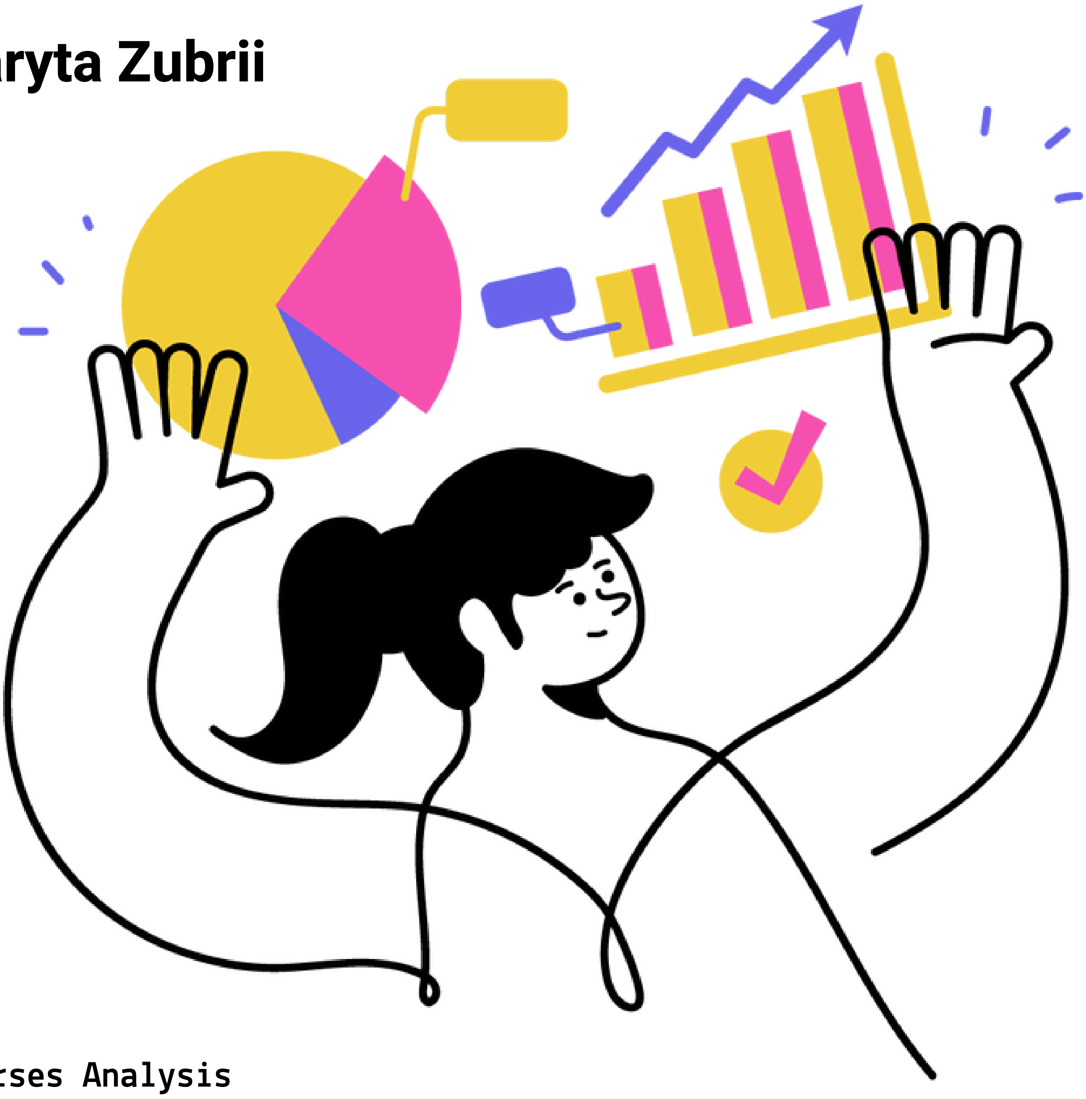
Python projects

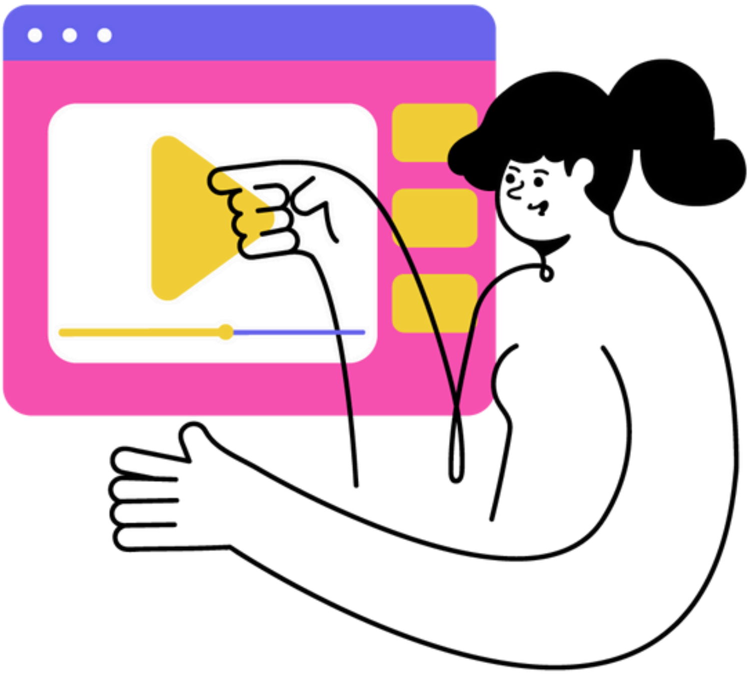
by Margaryta Zubrii



CONTENTS

🎓 COURSERA COURSES ANALYSIS

PROJECT OVERVIEW



Python libraries
numpy, pandas, scipy, seaborn,
matplotlib, plotly.express

Execution Environment
Jupyter Notebook (Google Colab)

Project Steps
Descriptive statistics,
Data visualization,
Exploratory Data Analysis (EDA)

GitHub Link
Coursera EDA project

ABOUT THE COMPANY

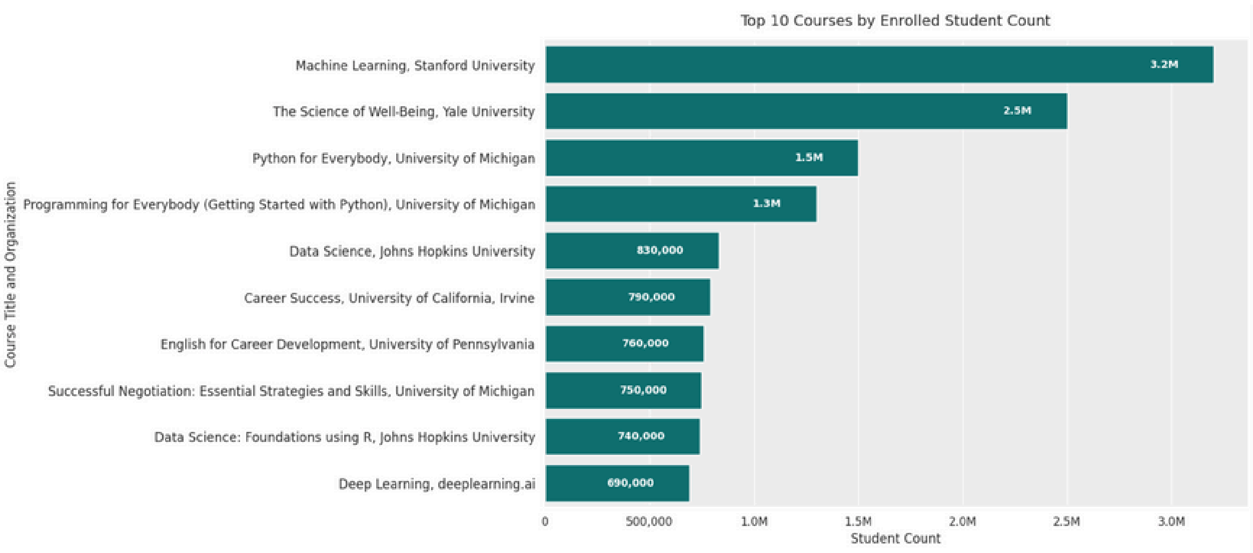
Coursera is an **online educational platform** which partners with more than **350 leading universities and companies** to bring flexible, affordable, job-relevant online learning to individuals and organizations worldwide. **175+ mln** learners joined Coursera to develop their skills since **2012.**

PROJECT DESCRIPTION

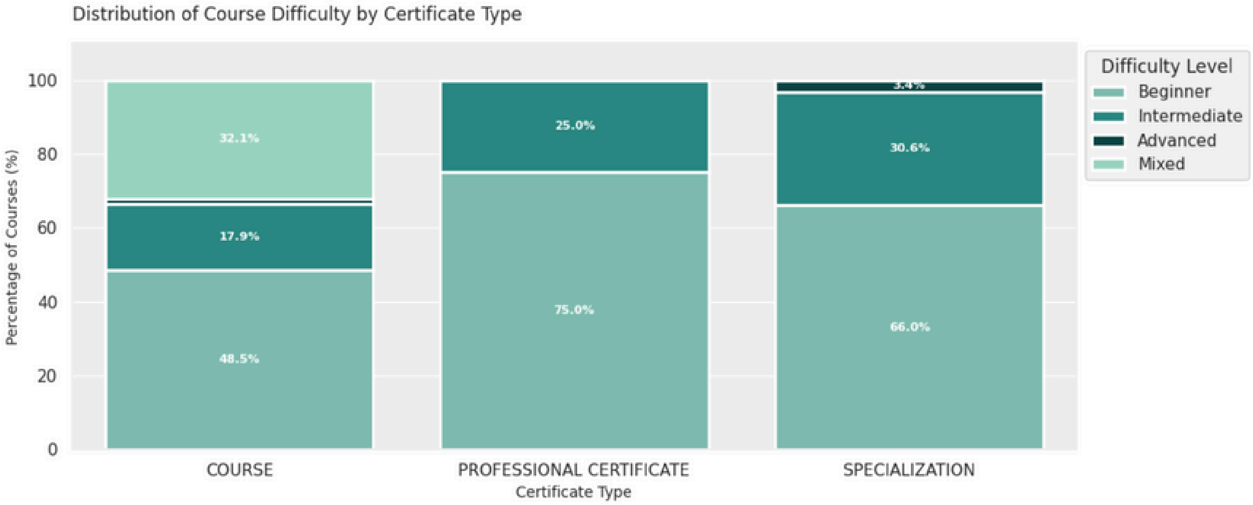
This analysis explores a dataset of **~900 courses** highly rated courses in **data and business fields** from **150+ organizations** on Coursera. It explores key indicators: course ratings, student count, course types and course providers.This analysis is relevant for various stakeholders in online education field.

EXECUTIVE SUMMARY

This EDA revealed that dataset is characterized by exceptionally high and consistent course quality coupled with a highly stratified enrollment structure. The top-enrolled courses serve as highly visible entry points, while the majority of high-quality content remains relatively niche.



The most popular study fields by the number of students include: programming languages (Python, R), machine learning, data science, English language, bussiness and finance, and career development.



The platform is dominated by "Beginner" level courses, aligning with its role in foundational skill acquisition across popular fields.

The most popular course providers by both the number of students enrolled and course count are University of Michigan (7.4M students, 41 courses), University of Pennsylvania (5.5M students, 59 courses), Stanford University (4.9M, 16 courses).

DATA INSIGHTS

Course ratings:

The platform demonstrates high perceived course quality with minimal variation in ratings. Any course with a rating below 4.30 is considered unusually low compared to the majority. There 17 outliers with low course ratings.There are 34 organizations with average course rating equal to or higher than 4.8. Organizations with less courses tend to have higher average ratings than those with multiple courses.

Course popularity:

The vast majority of courses have a relatively small number of enrolled students (up to 222,5k). A small, highly popular set of 78 courses with high enrollment make outliers.

Rating vs. Popularity:

Correlation test (rs = 0.0268 (including outliers), 0.0038 (excluding outliers)) presented that there is no statistically significant monotonic relationship between course rating and the number of students enrolled in this sample. A course's success (high enrollment) does not reliably predict its quality (high rating), and vice versa. Very popular courses don’t have higher perceived quality (rating) than less popular ones.

SUGGESTIONS

Further analysis can be applied to selected course organizations or study fields to find valuable insights on the quality and popularity of particular courses within organizations or areas where new courses can be introduced to fill in the gaps in learning material.

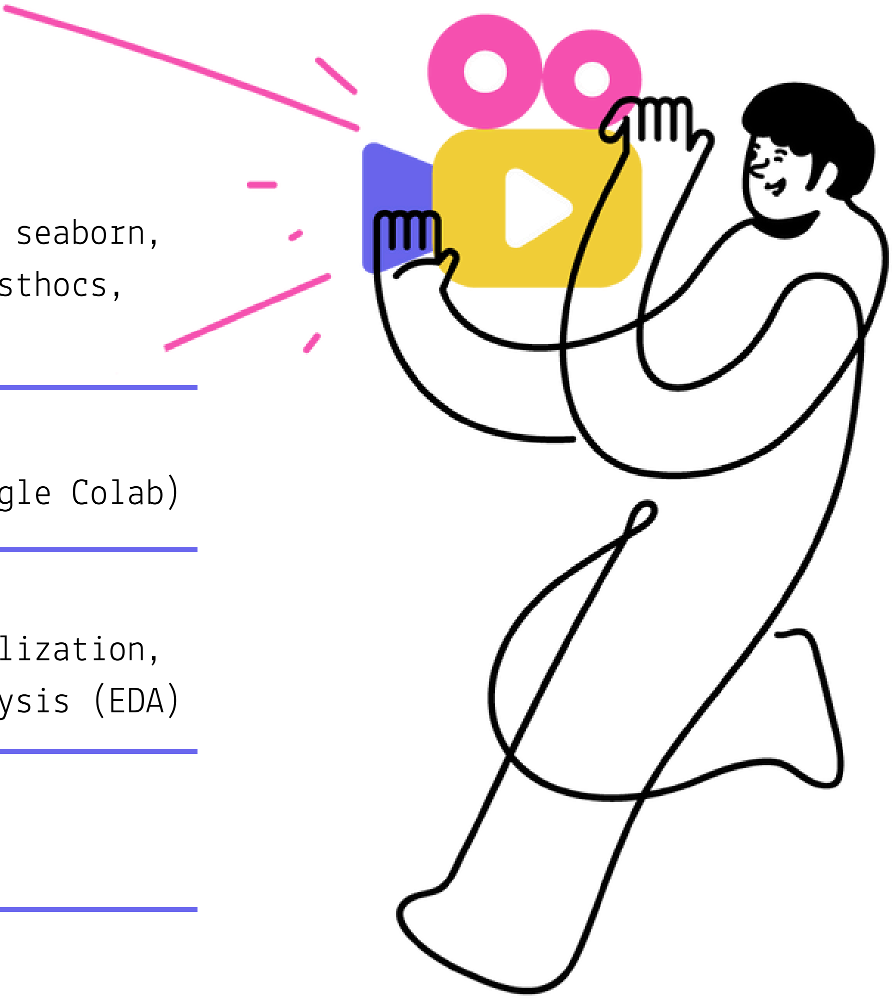
Course organizations could invest into developing more "Intermediate" and "Advanced" level courses to attract more specified audiences or return students from "Beginner" level courses to continue learning with the organization.

Course providers should prioritize content development exclusively within their key domain areas to ensure high-quality delivery, which directly impacts and helps maintain high average course ratings.

Full project in .ipynb file

🎬 IMDB TOP 1000 MOVIES ANALYSIS

PROJECT OVERVIEW



Python libraries
numpy, pandas, scipy, seaborn, matplotlib, skikit_posthocs, plotly.express

Execution Environment
Jupyter Notebook (Google Colab)

Project Steps
Data cleaning & visualization, Exploratory Data Analysis (EDA)

GitHub Link
IMDb EDA project

ABOUT THE COMPANY

IMDb is an **online database** of information related to films, TV series, podcasts, video games, and other streaming content - including cast, production crew and biographies, plot summaries, trivia, ratings, and fan and critical reviews.
Today, it ranks as the **40th most visited website** in the world.

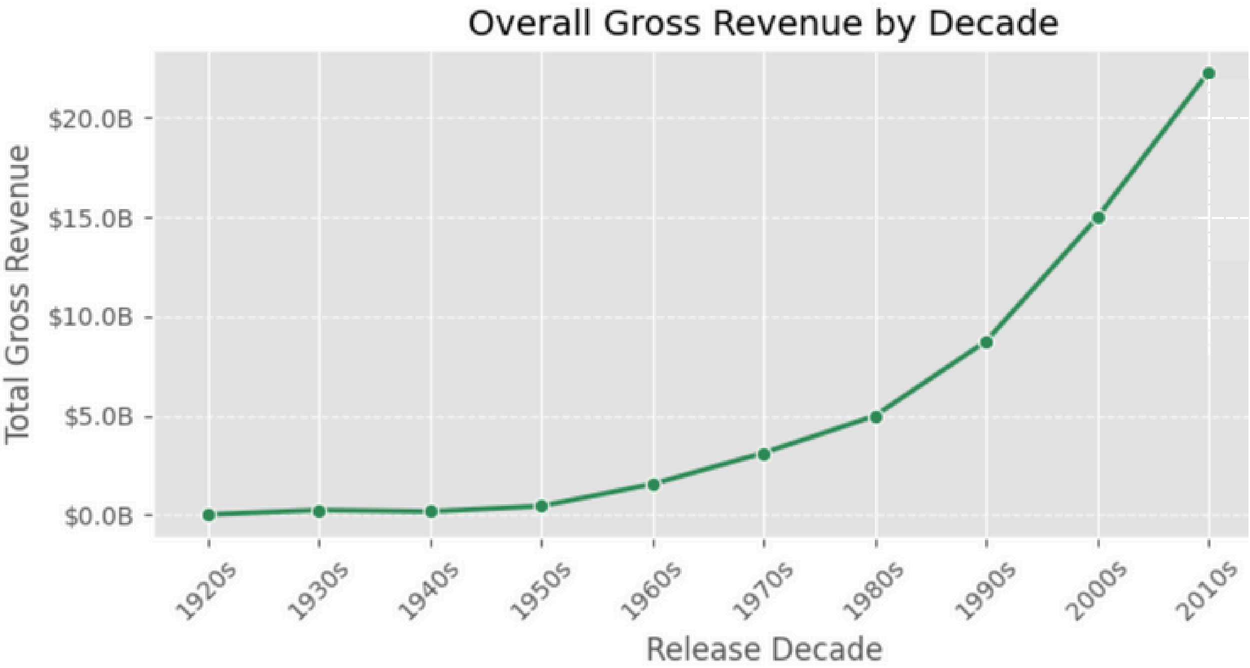
PROJECT DESCRIPTION

This project presents an EDA of a dataset featuring the **Top 1000 movies** on the platform, spanning a century of cinematography **from 1920 to 2020.** It explores key features like movies' gross revenue, director, actor & genre popularity, runtime distribution to answer stakeholders' questions and uncover insights about film industry.

EXECUTIVE SUMMARY

The dataset comprises 1000 movies representing 21 genres and featuring 548 directors. The movie count per decade shows a clear trend of increasing releases over time, peaking after the year 2000.

Raise in movie count also contribute to a raise of overall gross revenue. Data shows a rapid growth of film industry and its market size since the 1960s, and especially since 1990s, peaking sharply in the 2010s.



Key stats:

Genres: the most common movie genres are drama, comedy, crime, adventure, and action. The vast majority of movies are multi-genre.

Key Individuals: Alfred Hitchcock is the most prolific director, while Robert De Niro leads the actor count.

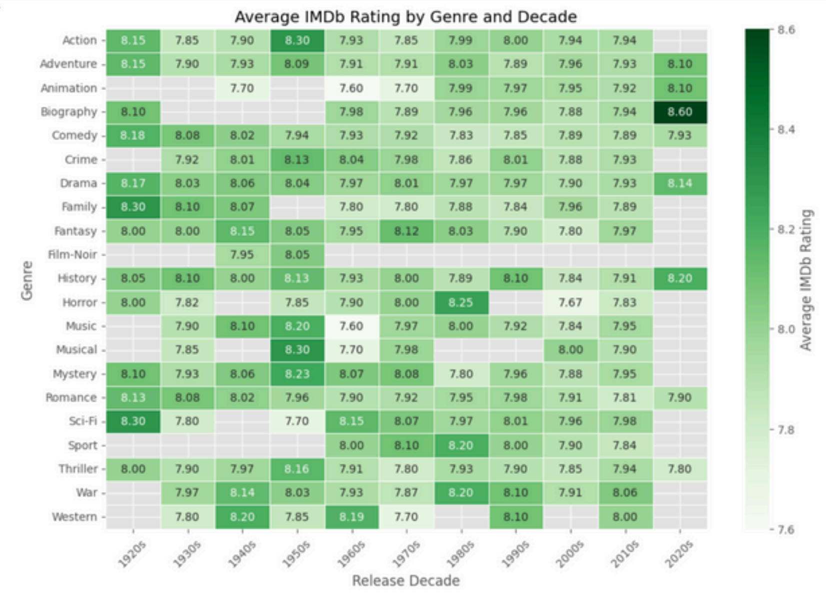
Ratings: the overall average IMDb rating for movies in the dataset is 7.95.

Runtime: the average movie runtime is 122.9 minutes.

DATA INSIGHTS

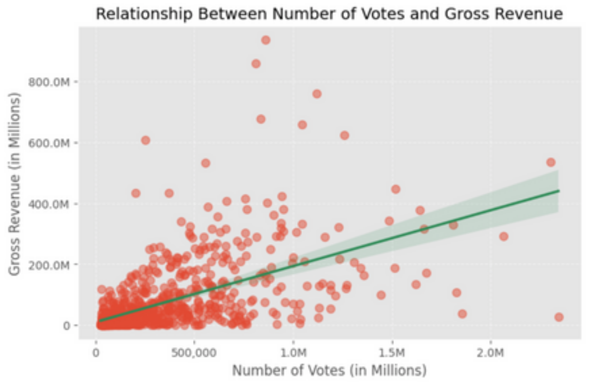
Ratings by genre:
Few genres display high average ratings (>8.0) throughout the studied period: biography, drama, film-noir, history, sport, war, and mystery.

Simultaneously, animation, horror, and thriller genres have the lowest average ratings.

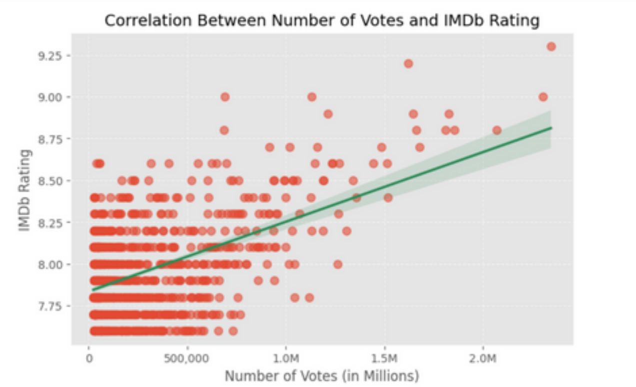

Average IMDb Rating by Genre and Decade

Genre	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s	2010s	2020s
Action	8.15	7.85	7.90	8.30	7.93	7.85	7.99	8.00	7.94	7.94	
Adventure	8.15	7.90	7.93	8.09	7.91	7.91	8.03	7.89	7.96	7.93	8.10
Animation			7.70		7.60	7.70	7.99	7.97	7.95	7.92	8.10
Biography	8.10				7.98	7.89	7.96	7.96	7.88	7.94	8.60
Comedy	8.18	8.08	8.02	7.94	7.93	7.92	7.83	7.85	7.89	7.89	7.93
Crime		7.92	8.01	8.13	8.04	7.98	7.86	8.01	7.88	7.93	
Drama	8.17	8.03	8.06	8.04	7.97	8.01	7.97	7.97	7.90	7.93	8.14
Family	8.30	8.10	8.07		7.80	7.80	7.88	7.84	7.96	7.89	
Fantasy	8.00	8.00	8.15	8.05	7.95	8.12	8.03	7.90	7.80	7.97	
Film-Noir			7.95	8.05							
History	8.05	8.10	8.00	8.13	7.93	8.00	7.89	8.10	7.84	7.91	8.20
Horror	8.00	7.82		7.85	7.90	8.00	8.25		7.67	7.83	
Music		7.90	8.10	8.20	7.60	7.97	8.00	7.92	7.84	7.95	
Musical		7.85		8.30	7.70	7.98			8.00	7.90	
Mystery	8.10	7.93	8.06	8.23	8.07	8.08	7.80	7.96	7.88	7.95	
Romance	8.13	8.08	8.02	7.96	7.90	7.92	7.95	7.98	7.91	7.81	7.90
Sci-Fi	8.30	7.80		7.70	8.15	8.07	7.97	8.01	7.96	7.98	
Sport					8.00	8.10	8.20	8.00	7.90	7.84	
Thriller	8.00	7.90	7.97	8.16	7.91	7.80	7.93	7.90	7.85	7.94	7.80
War		7.97	8.14	8.03	7.93	7.87	8.20	8.10	7.91	8.06	
Western		7.80	8.20	7.85	8.19	7.70		8.10		8.00	

Correlation tests:


Relationship Between Number of Votes and Gross Revenue

The Spearman correlation coefficient between Number of Votes and Gross Revenue is: 0.6991


Correlation Between Number of Votes and IMDb Rating

The Spearman correlation coefficient between Number of Votes and IMDb Rating is: 0.2140

There is a moderately strong, positive correlation (rs = 0.6991) between the vote count and gross revenue. This means that popular films (by revenue) usually get more reviews on the platform.

A weak, positive monotonic relationship between the number of votes and IMDb rating (rs = 0.2140) suggests that movies with higher vote counts tend to have slightly higher ratings, but vote count is not a strong predictor of rating.

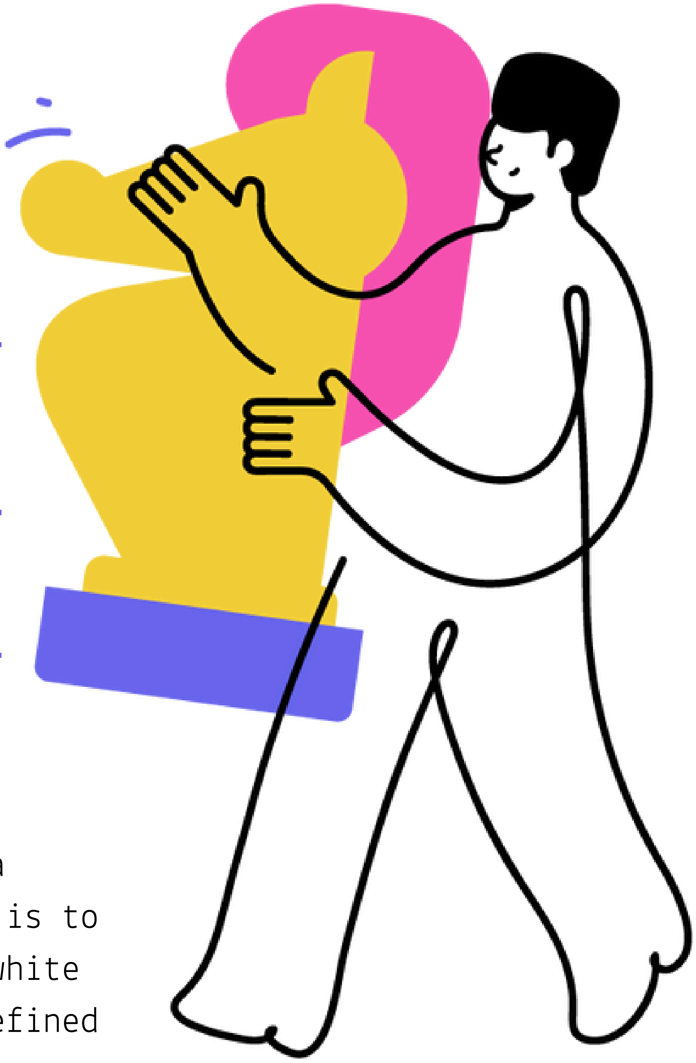
Director has a significant impact on a movie's median gross revenue. However, the significant impact of director's identity on IMDb rating was not confirmed.

Actor selection has influence on median gross revenue of a film as well. Unlike the IMDb rating, in financial terms, hiring an actor from the top tier range is statistically associated with a higher median financial outcome.

Full project in .ipynb file

CHESS PROJECT

PROJECT OVERVIEW



Python libraries
tabulate

Execution Environment
Jupyter Notebook (Google Colab)

GitHub Link
Chess project

PROJECT DESCRIPTION

The project presents a part of a chess game simulation. The goal is to determine which black pieces a white piece can capture from a user-defined board state, using standard chess movement rules.

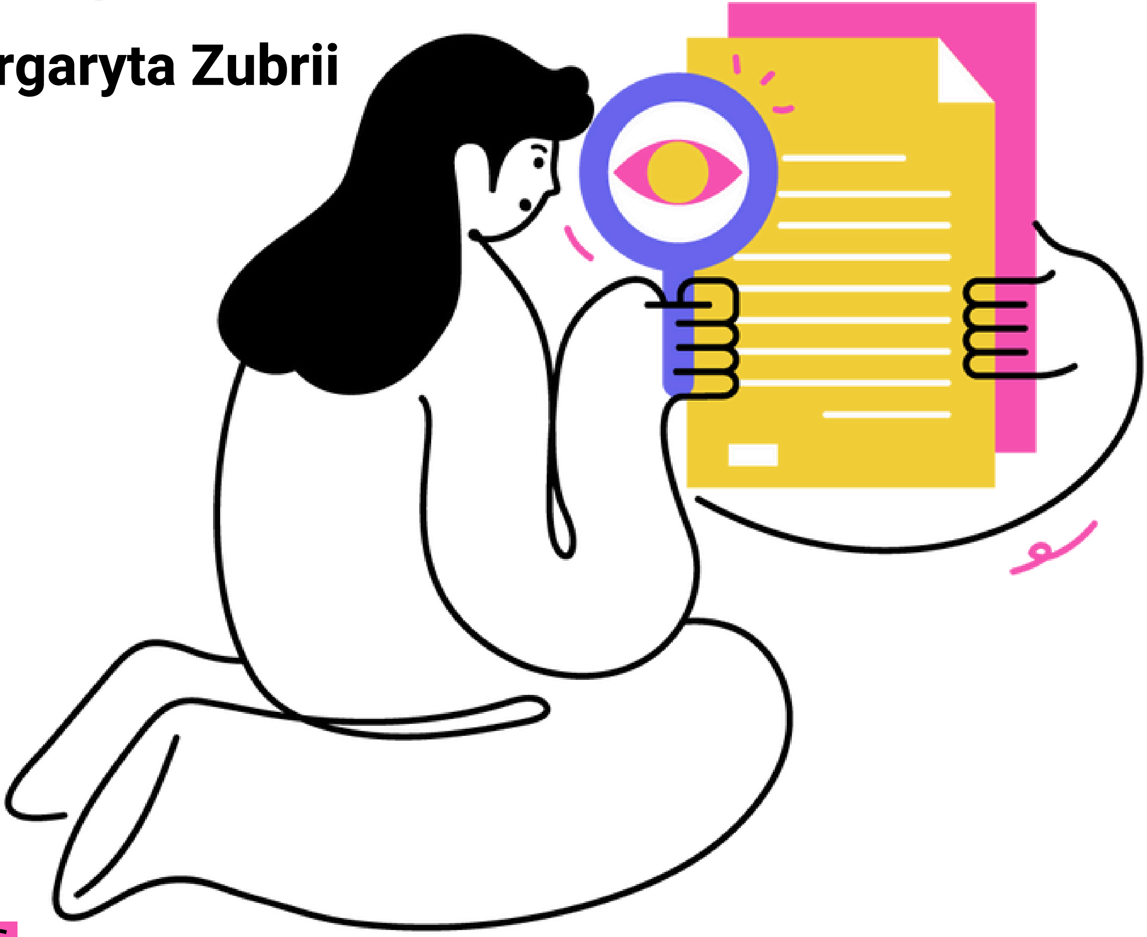
TASK REQUIREMENTS

ChessFormat: position inputs must be valid chess coordinates.
Figure Names: must be valid chess piece names.
White Piece Limit: only 1 white figure is allowed.
Black Piece Limit: accepts 1-16 black figures.
"Done" Command: "done" can be accepted only after at least one black piece has been successfully added to the board.
Final Output: the program must display either a list of all capturable black pieces and their positions, or a message clearly stating that no captures are possible.

Full project in .ipynb file

SQL project

by Margaryta Zubrii

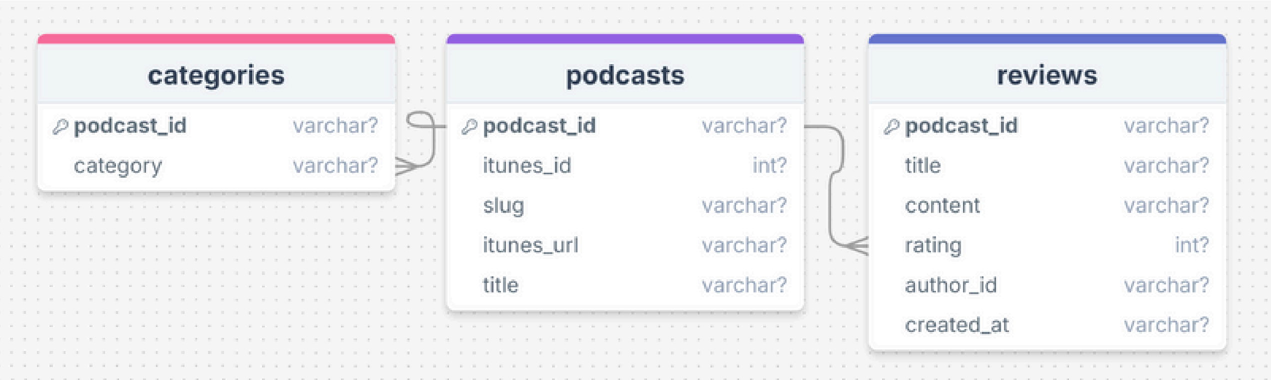


CONTENTS

GitHub Link
Podcasts analysis

DATABASE SCHEMA

This SQL project includes several steps, including creation of database schema, data cleaning, dimensional aggregation and relational data analysis.



categories	
podcast_id	varchar?
category	varchar?

podcasts	
podcast_id	varchar?
itunes_id	int?
slug	varchar?
itunes_url	varchar?
title	varchar?

reviews	
podcast_id	varchar?
title	varchar?
content	varchar?
rating	int?
author_id	varchar?
created_at	varchar?

A database titled ‘podcast_reviews’ has three tables: ‘categories’, ‘podcasts’ and ‘reviews’. Tables can be connected through ‘podcast_id’.

Dataset size:
‘categories’ table has 212,372 rows.
‘podcasts’ table has 110,023 rows.
‘reviews’ table has 2,067,529 rows.

Simple SQL queries used to find number of rows in each table:

```
SELECT COUNT(DISTINCT podcast_id) FROM `podcast_reviews.podcasts`
```

```
SELECT COUNT(*) FROM `podcast_reviews.reviews`
```

```
SELECT COUNT(*) FROM `podcast_reviews.categories`
```

DATA EXPLORATION

Initial exploration presented that a single podcast can belong to multiple categories and have multiple reviews, including review’s ‘author_id’, ‘title’, ‘content’, and ‘rating’ values.

Values in ‘category’ column of ‘categories’ table required cleaning.

RELATIONAL DATA ANALYSIS

Top 5 categories by podcast count:

Category	Podcast count	Reviews count	AVG Rating
Society-culture	19,441	455,191	4,55
Education	13,192	236,485	4,81
Business	13,118	229,577	4,84
Comedy	13,116	360,023	4,63
Spirituality	12,363	149,672	4,83

Popular podcast categories have in between 10k-20k podcasts, their average ratings span from 4,63 to 4,84. Highly rated podcasts belong to niche categories with much lower numbers of podcasts per category.

Top 5 categories by highest ratings:

Category	Podcast count	Reviews count	AVG Rating
Rugby	35	167	4,99
Marketing	1,402	28,869	4,94
Entrepreneurship	3,601	80,242	4,91
Non-profit	506	2,570	4,90
Running	153	6,157	4,89

Digest type podcasts (daily, news, politics) have relatively low avg. ratings.

5 categories with lowest ratings:

Category	Podcast count	Reviews count	AVG Rating
Daily	321	15,030	3,97
True-crime	1,264	162,550	4,16
Politics	1,168	47,888	4,21
Documentary	732	31,338	4,31
News	6,606	190,440	4,31

RELATIONAL DATA ANALYSIS

Top 10 podcasts by review count:

Podcast title	Review count	AVG Rating
Crime Junkie	33,104	4,27
My Favorite Murder with Karen Kilgariff and Georgia Hardstark	10,675	3,61
Wow in the World	9,698	4,78
The Ben Shapiro Show	8,248	3,8
Story Pirates	7,389	4,74
True Crime Obsessed	7,310	4,21
The Daily	5,537	3,34
Crime in Sports	5,099	4,93
Know Your Aura with Mystic Michaela	5,059	4,98
FantasyPros - Fantasy Football Podcast	4,666	4,82

Top 3 podcasts with highest ratings by category (3 most popular categories):

Podcast title	Reviews count	Rating	Category
Viene y Va con Dani G Schulz	317	5,0	Society-culture
Fun and Gains	287	5,0	
Drink in the Movies	267	5,0	
Above Average Podcast with Travis and Jesse	619	5,0	Education
Discover Your Talent-Do What You Love | Build a Career of Success, Satisfaction and Freedom	509	5,0	
Loan Officer Freedom	447	5,0	
THE STEFANIE GASS SHOW - Christian Entrepreneur, Start an Online Business, Work From Home, Get More...	522	5,0	Business
Discover Your Talent-Do What You Love | Build a Career of Success, Satisfaction and Freedom	509	5,0	
Loan Officer Freedom	447	5,0	

INSIGHTS

Podcast popularity vs. quality:

The most popular categories by number of podcasts like society-culture, education, and business all have relatively high podcast counts (13k to 19k) and massive review totals. However, their average ratings cluster tightly between 4.55 and 4.84. This means that there is a mix of high and low quality podcasts that leads to such average values.

In contrast, the categories with the highest ratings like rugby, marketing, and entrepreneurship have smaller podcast counts but their average ratings are much higher, starting at 4.89 and climbing to 4.99. This suggests that the sheer volume of content in major categories leads to an averaging effect, where a high volume of lower-rated podcasts dilutes the score of the top performers.

Digest Content has lowest Rating:

The categories with the lowest ratings are dominated by time-sensitive or high-volume content, indicating audience fatigue or less passionate reviews for general consumption. Content that is produced daily or covers sensitive, polarizing topics like politics or true-crime tends to pull down the overall average rating for its category.

The highest-volume podcast in the 'Daily' category has a very low rating of 3.34, reinforcing the low average rating of the entire Daily category (3.97).

Top Podcasts Dominate the Review Landscape:

The most popular podcast by review count, "Crime Junkie," has 33,104 reviews, which is more than three times the count of the next podcast, "My Favorite Murder" (10,675). This shows that a few dominant market leaders control the vast majority of engagement in their segments.

Cross-Category Performance:

The presence of high-performing podcasts like 'Discover Your Talent...' and 'Loan Officer Freedom' in both the Education and Business categories illustrates that successful podcasts often bridge multiple popular categories, a structure that required ad-hoc data transformation (SPLIT/UNNEST) to analyze accurately.