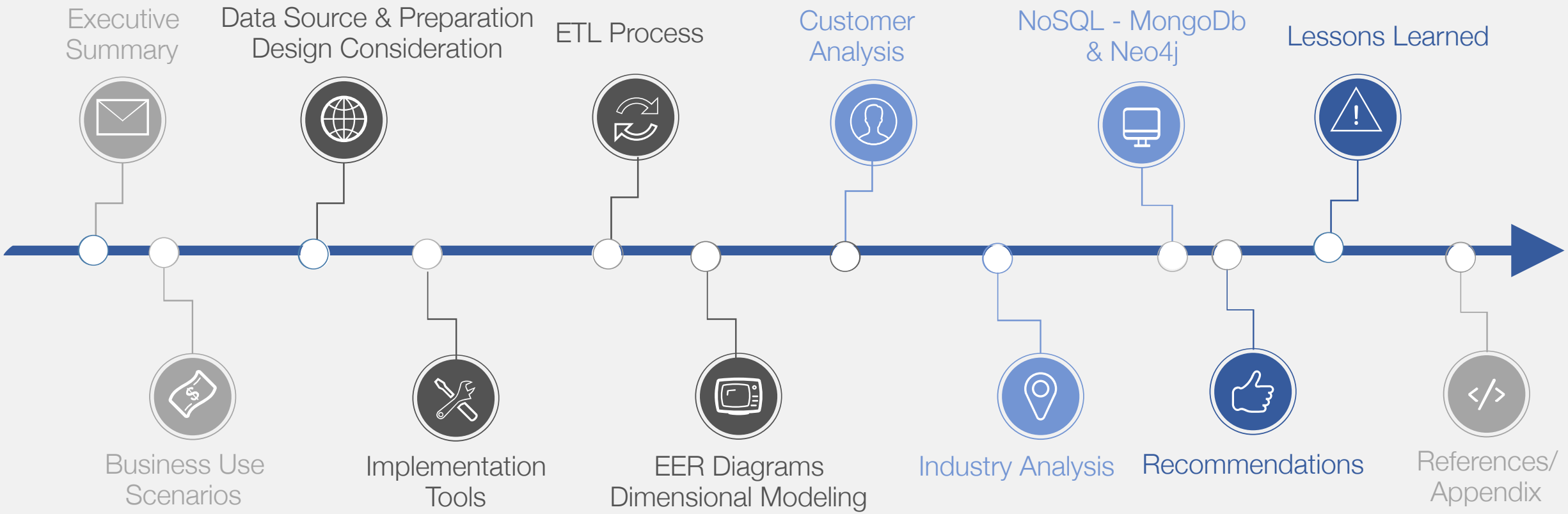
MOVIE INDUSTRY

Statistical Analysis

GROUP 3
Carrie Meijuan Lu, Jenny Zhihan Wang, Kai Li, Ziwei Zhao

AGENDA

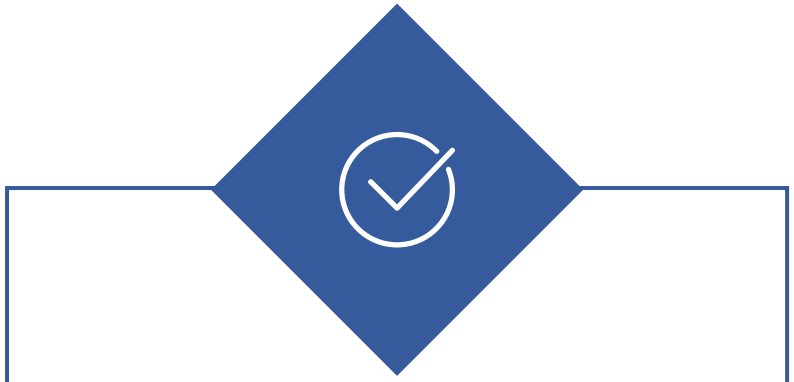

Executive Summary
Data Source & Preparation
Design Consideration
ETL Process
Customer Analysis
NoSQL - MongoDb & Neo4j
Lessons Learned
Business Use Scenarios
Implementation Tools
EER Diagrams
Dimensional Modeling
Industry Analysis
Recommendations
References/ Appendix

01 Executive Summary

Blockbuster Movie always associates with high popularity and financial success. However, investing in movie industry accompanies with high risks.

Our statistical analysis and data visualization will provide insights and recommendations to generate business values on capturing and predicting blockbuster movie success metrics.

02 Business Use Scenarios



BUILD SUCCESS METRICS

•Audience Targeting & Hidden Patterns
•Discovery & Blockbuster Features Prediction

Industry Analysis

MOVIE RECOMMENDATION SYSTEM

Recommend Movies to Users

Customer Analysis

Data Source & Preparation 03

IMDB & TMDB open API

Movie Details: Title, Language, Budget, Revenue, Genres, Keywords, etc.

GroupLens

Ratings

Our DataBase

Roughly 45,000 movies, over 26 million ratings from 270,000 users

Database Sample

Include 7 JSON Columns

budget	genres	id	production_companies	production_countries	spoken_languages	original_language	original_title	overview	poster_path	keywords
30000000	[{'id': 16, 'name': 'Animation'}, {'id': 35, 'name': 'Comedy'}, {'id': 10751, 'name': 'Family'}]	862	[{'name': 'Pixar Animation Studios', 'id': 3}]	[{'iso_3166_1': 'US', 'name': 'United States of America'}]	[{'iso_639_1': 'en', 'name': 'English'}]	en	Toy Story	Led by Woody, Andy's toys live happily in his room until Andy…	/rhIRbceoE9lR4veEXuwCC2wARtG.jpg	[{'id': 931, 'name': 'jealousy'}, {'id': 4290, 'name': 'toy'}, {'id': 5202, 'name': 'boy'}…

release_date	revenue	runtime	staus	tagline	title	video	vote_average	vote_count	movieId	popularity
10/30/95	373554033	81	Released	/	Toy Story	FALSE	7.7	5415	1	21.946943

imdbId	rating	timestamp	userId	cast	crew
114709	1	1425941529	1	[{'cast_id': 14, 'character': 'Woody (voice)', 'credit_id': '52fe4284c3a36847f8024f95', 'gender': 2, 'id': 31, 'name': 'Tom Hanks', 'order': 0, 'profile_path': '/pQFoyx7rp09CJTAb932F2g8Nlho.jpg'}, {'cast_id': 15, 'character': 'Buzz Lightyear (voice)', 'credit_id': '52fe4284c3a36847f8024f99', 'gender': 2, 'id': 12898, 'name': 'Tim Allen', 'order': 1, 'profile_path': '/uX2xVf6pMmPepxnvFWyBtjexzgY.jpg'}…	[{'credit_id': '52fe4284c3a36847f8024f49', 'department': 'Directing', 'gender': 2, 'id': 7879, 'job': 'Director', 'name': 'John Lasseter', 'profile_path': '/7EdqiNbr4FRjIhKHyPPdFfEEEFG.jpg'}, {'credit_id': '52fe4284c3a36847f8024f4f', 'department': 'Writing', 'gender': 2, 'id': 12891, 'job': 'Screenplay', 'name': 'Joss Whedon', 'profile_path': '/dTiVsuaTVTeGmvkhcyJvKp2A5kr.jpg'}…

Design Considerations 03

INTEGER INDEXING

Choose columns (e.g. id, movie_rating_id) with the integer data type (or its variants) for indexing to ensure performance.

OPTIMIZED NORMALIZATION

Implement best practice of normalization to avoid excessive repetition of data or excessive joins across too many tables.

INTEGRITY CHECKS

Set foreign keys and constraints to ensure data integrity with careful design that not to overuse or underuse these integrity checks.

DOCUMENT OF THE MODEL

Communicate the design and make it understandable in the future of what the usage of each object will be.

OLAP SNOWFLAKE SCHEMA FOR THE APPLICATION

Support analysis, reporting, forecasting to fetch and analyze data as fast as possible

Implementation Tools 04

Raw Dataset & Information

IMDb

THE MOVIE DB

GroupLens Research

Data Preparation (Split JSON Columns)

OpenRefine

python

jupyter

Data Normalization ETL Process

python

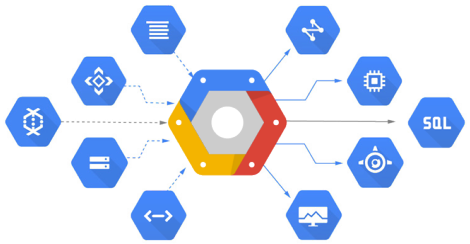
jupyter

MySQL

Database Design & Modeling

MySQL

Data Storage



Google Cloud Platform

Data Analytics

python

jupyter

Data Visualization & Dashboards

tableau SOFTWARE

NoSQL

neo4j

mongoDB

05 ETL Process: Extraction, Transformation & Load

Dataset retrieved from a public source with a combination of data from IMBD & TMBD open API and GroupLens.

Split the JSON columns & Data Normalization

movie_id	cast
862	[{'cast_id': 14, 'character': 'Woody (voice)', 'credit_id': '52fe4284c3a36847f8024f95', 'gender': 2, 'id': 31, 'name': 'Tom Hanks', 'order': 0, 'profile_path': '/pQFoyx7rp09CJTAb932F2g8Nlho.jpg'}, {'cast_id': 15, 'character': 'Buzz Lightyear (voice)', 'credit_id': '52fe4284c3a36847f8024f99', 'gender': 2, 'id': 12898, 'name': 'Tim Allen', 'order': 1, 'profile_path': '/uX2xVf6pMmPepxnvFWyBtjexzgY.jpg'}, {'cast_id': 16, 'character': 'Mr. Potato Head (voice)', 'credit_id': '52fe4284c3a36847f8024f9d', 'gender': 2, 'id': 7167, 'name': 'Don Rickles', 'order': 2, 'profile_path': '/h5BcaDMPRVLHLDzbQavec4xfSdt.jpg'}, {'cast_id': 17, 'character': 'Slinky Dog (voice)', 'credit_id': '52fe4284c3a36847f8024fa1', 'gender': 2, 'id': 12899, 'name': 'Jim Varney', 'order': 3, 'profile_path': '/eIo2jVVXYgjDtaHoF19Ll9vtW7h.jpg'}, {'cast_id': 18, 'character': 'Rex (voice)', 'credit_id': '52fe4284c3a36847f8024fa5', 'gender': 2, 'id': 12900, 'name': 'Wallace Shawn', 'order': 4, 'profile_path': '/oGE6JqPP2xH4tNORKNqxbNPYi7u.jpg'}, {'cast_id': 19, 'character': 'Hamm (voice)', 'credit_id': '52fe4284c3a36847f8024fa9', 'gender': 2, 'id': 7907, 'name': 'John Ratzenberger', 'order': 5, 'profile_path': '/yGechiKWL6TJDfVE2KPSJYqdMsY.jpg'}, {'cast_id': 20, 'character': 'Bo Peep (voice)', 'credit_id': '52fe4284c3a36847f8024fad', 'gender': 1, 'id': 8873, 'name': 'Annie Potts', 'order': 6, 'profile_path': '/eryXT84RL41jHSJcMy4kS3u9y6w.jpg'}, {'cast_id': 26, 'character': 'Andy (voice)', 'credit_id': '52fe4284c3a36847f8024fc1', 'gender': 0, 'id': 1116442, 'name': 'John Morris', 'order': 7, 'profile_path': '/vYGyvK4LzeaUCoNSHtsuqJUY15M.jpg'}, {'cast_id': 22, 'character': 'Sid (voice)', 'credit_id': '52fe4284c3a36847f8024fb1', 'gender': 2, 'id': 12901, 'name': 'Erik von Detten', 'order': 8, 'profile_path': '/twnF1ZaJ1FUNUuo6xLXwcxjayBE.jpg'}, {'cast_id': 23, 'character': 'Mrs. Davis (voice)', 'credit_id': '52fe4284c3a36847f8024fb5', 'gender': 1, 'id': 12133, 'name': 'Laurie Metcalf', 'order': 9, 'profile_path': '/unMMIT60eoBM2sN2nyR7EZ2BvvD.jpg'}, {'cast_id': 24, 'character': 'Sergeant (voice)', 'credit_id': '52fe4284c3a36847f8024fb9', 'gender': 2, 'id': 8655, 'name': 'R. Lee Ermey', 'order': 10, 'profile_path': '/r8GBqFBjypLUP9VVqDqfZ7wYbSs.jpg'}, {'cast_id': 25, 'character': 'Hannah (voice)', 'credit_id': '52fe4284c3a36847f8024fbd', 'gender': 1, 'id': 12903, 'name': 'Sarah Freeman', 'order': 11, 'profile_path': None}, {'cast_id': 27, 'character': 'TV Announcer (voice)', 'credit_id': '52fe4284c3a36847f8024fc5', 'gender': 2, 'id': 37221, 'name': 'Penn Jillette', 'order': 12, 'profile_path': '/zmAaXUdx12NRsssgHbk1T31j2x9.jpg'}]

python

movie_cast_id	movie_id	cast_id	cast_character
0	862	31	Woody (voice)
1	862	12898	Buzz Lightyear (voice)
2	862	7167	Mr. Potato Head (voice)
3	862	12899	Slinky Dog (voice)
4	862	12900	Rex (voice)

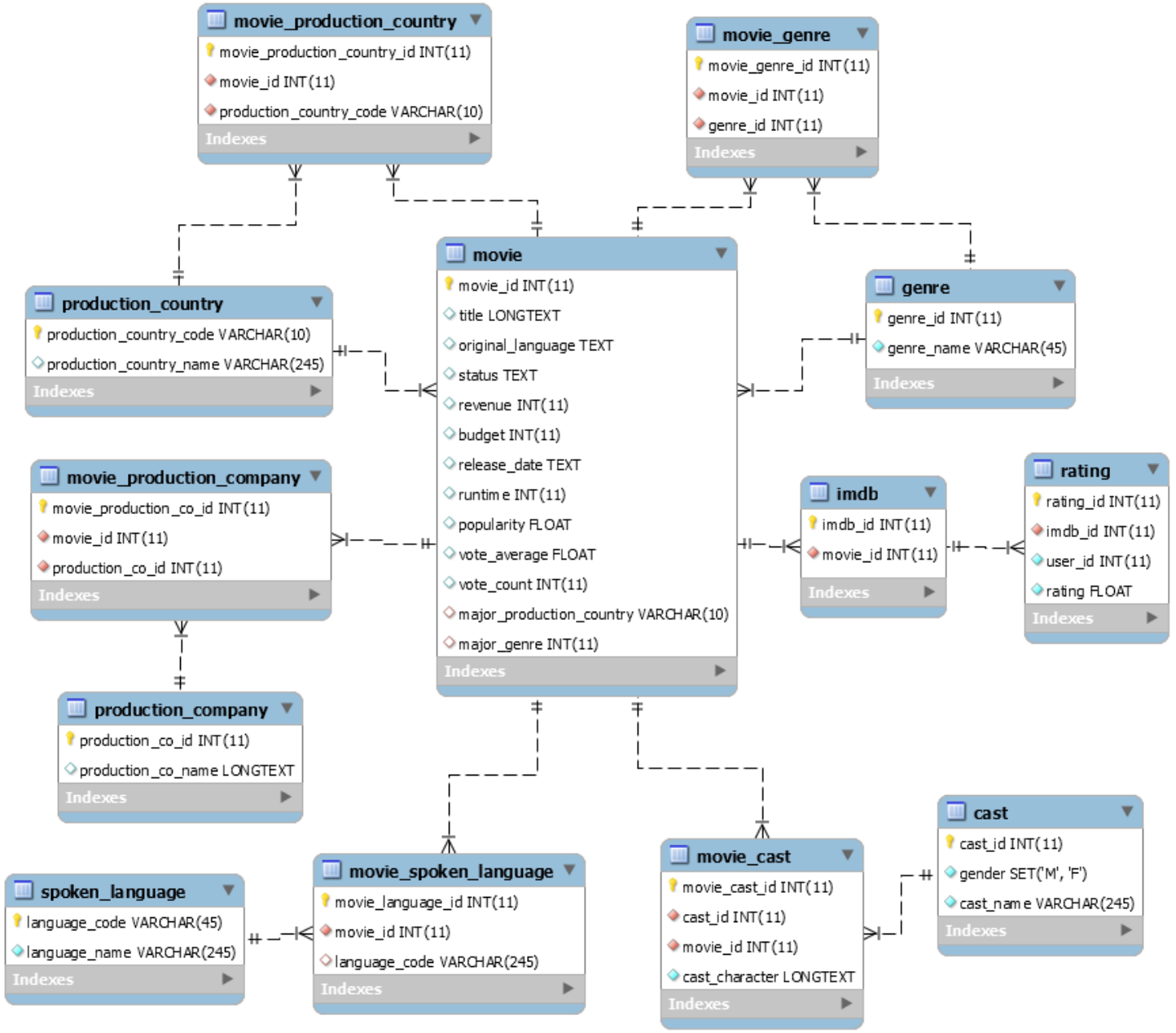
cast_id	gender	cast_name
1	M	George Lucas
2	M	Mark Hamill
3	M	Harrison Ford
4	F	Carrie Fisher
5	M	Peter Cushing

Load data: MySQL

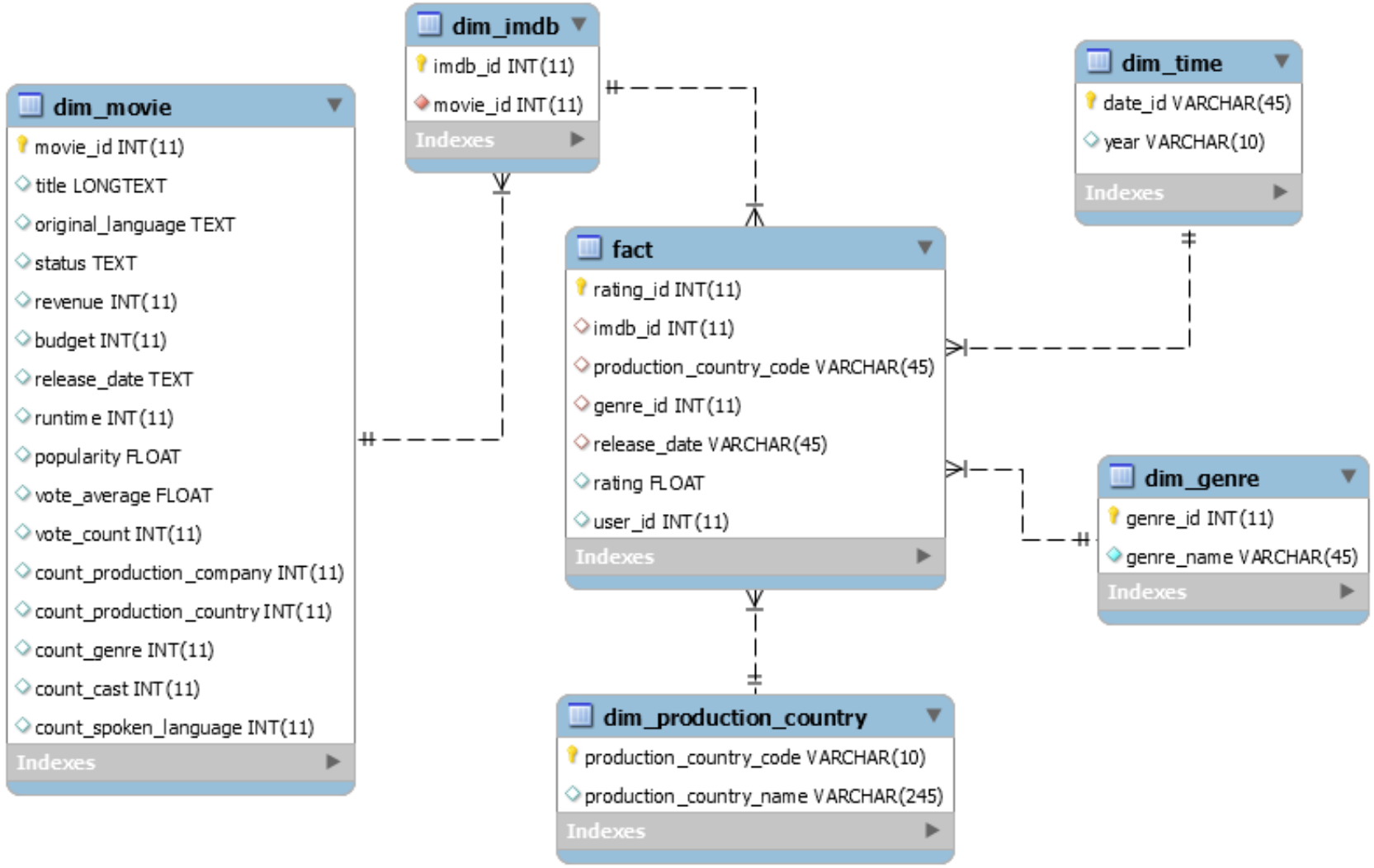
06 EER Diagram

MySQL

Normalization & EER Diagram

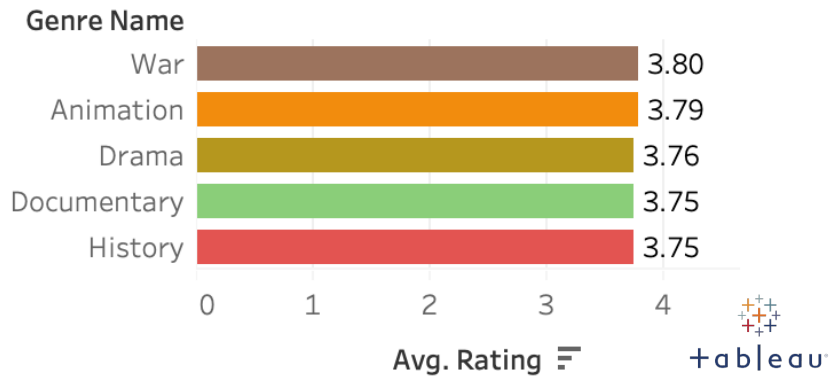


Dimensional Modeling 06

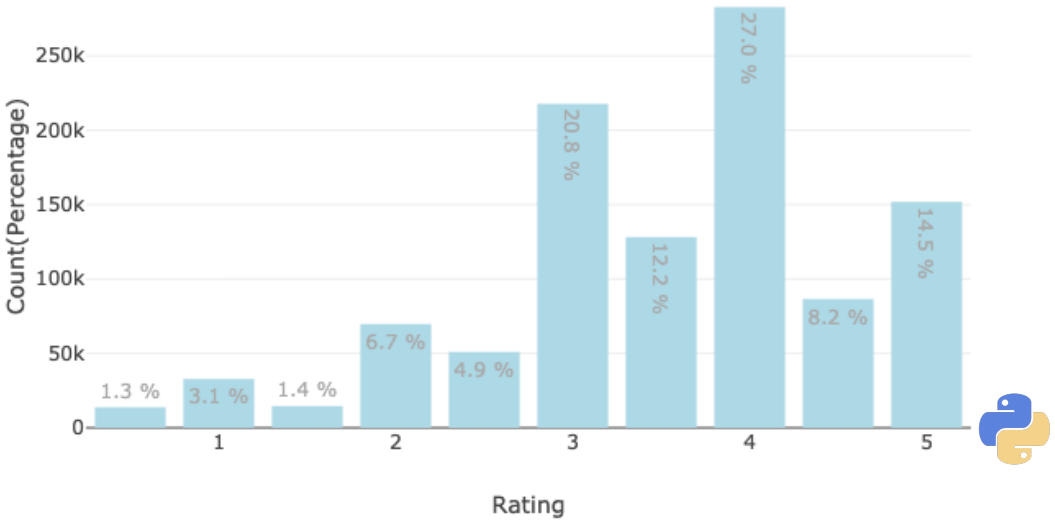


Customer Analysis

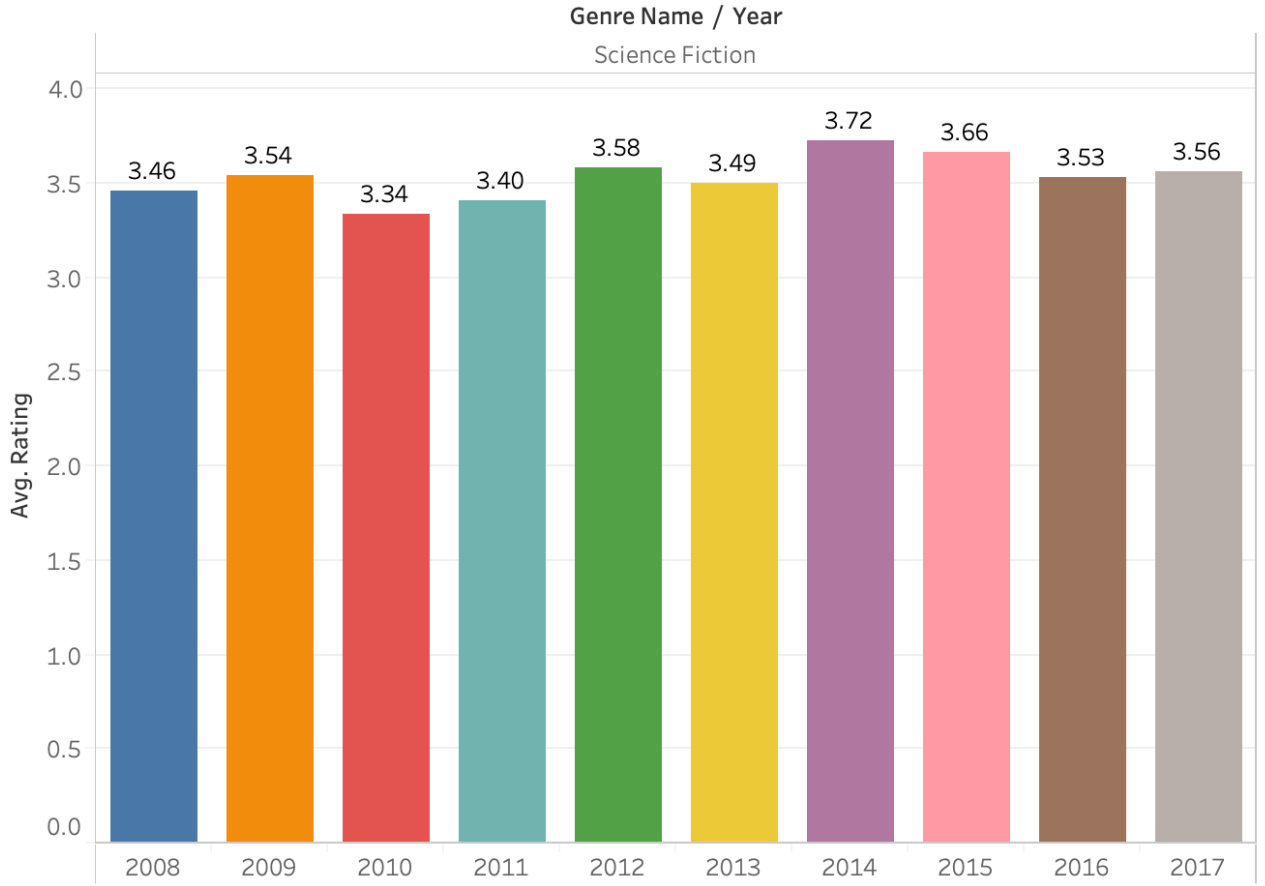
Top 5 Popular Genre Based on Avg Ratings



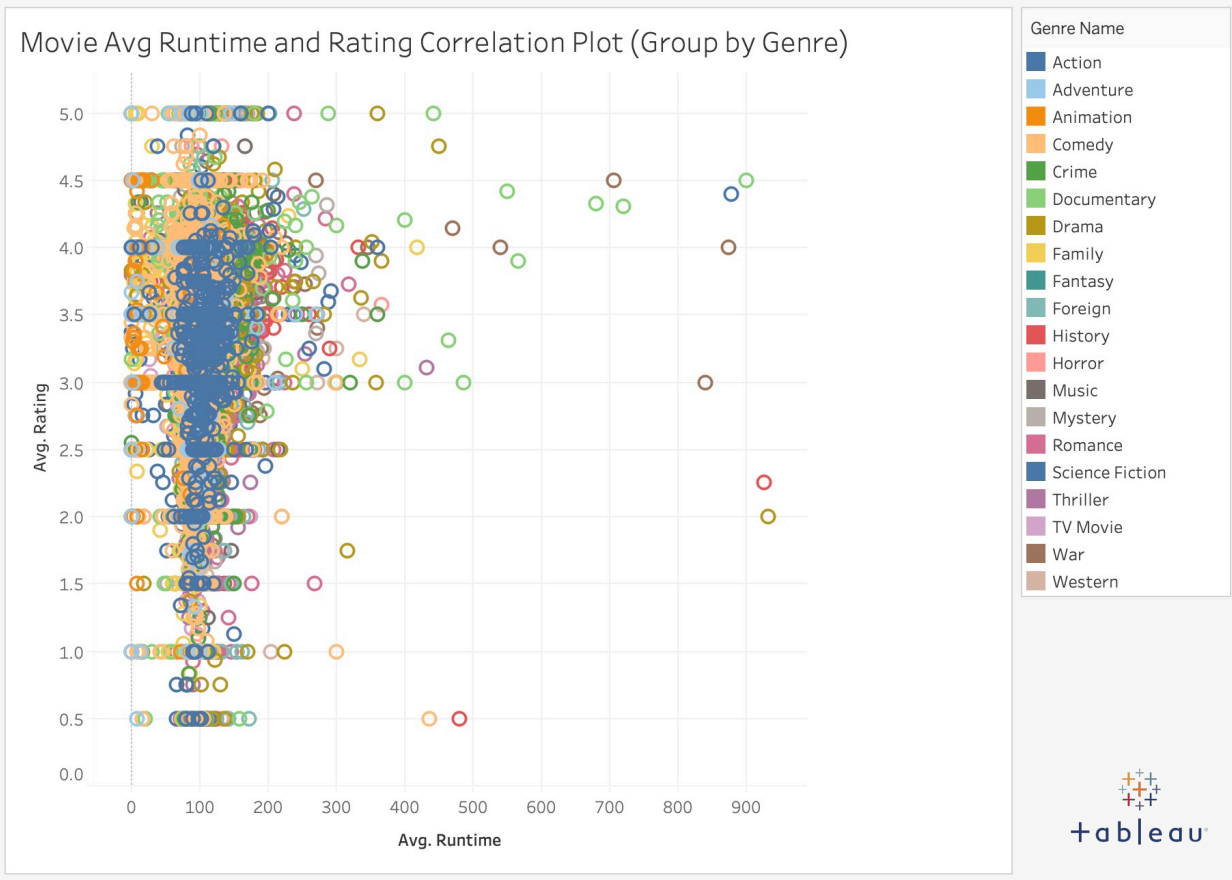
Ratings Distribution

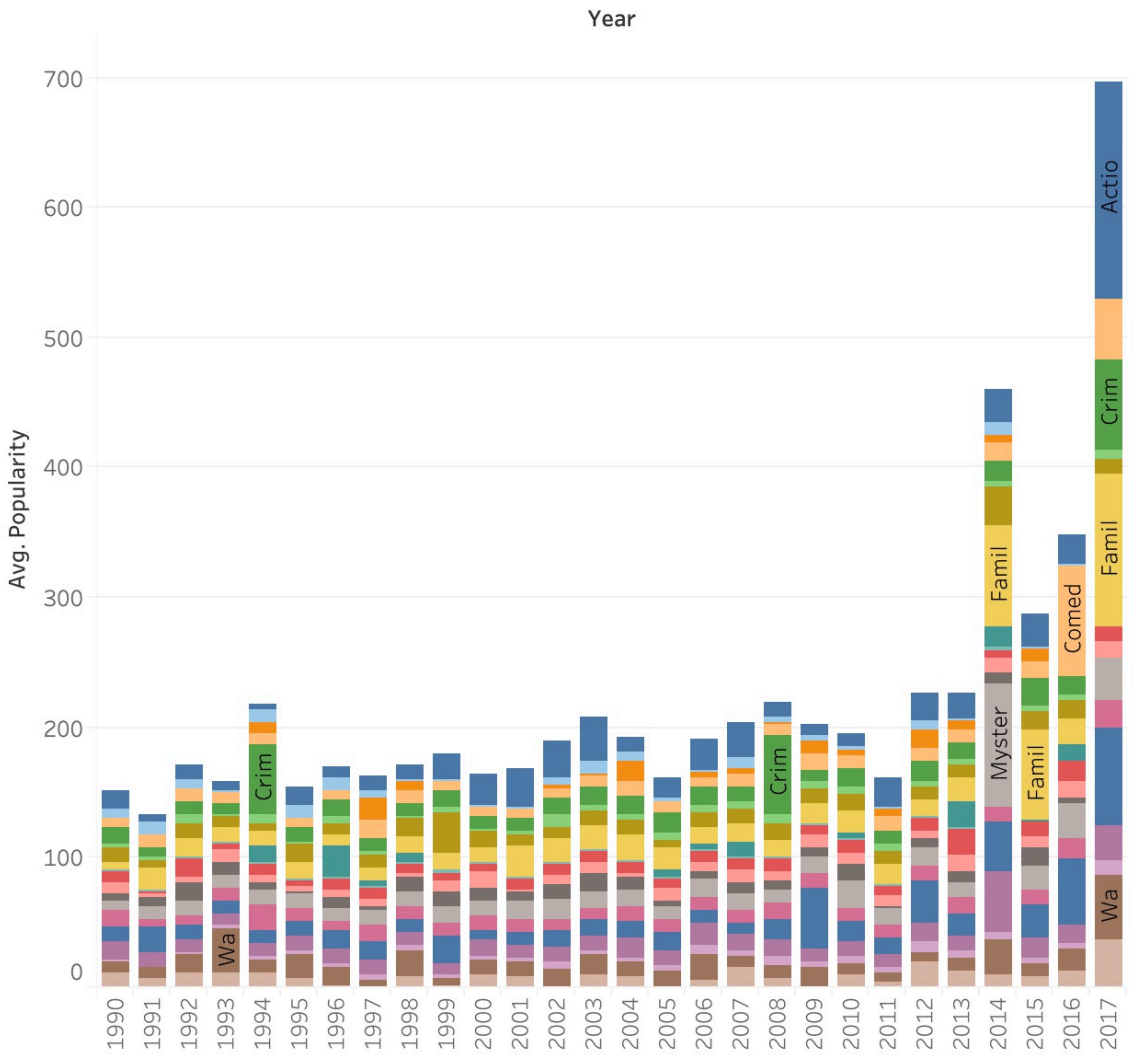


Science Fiction Rating Change in the Last 10 Years



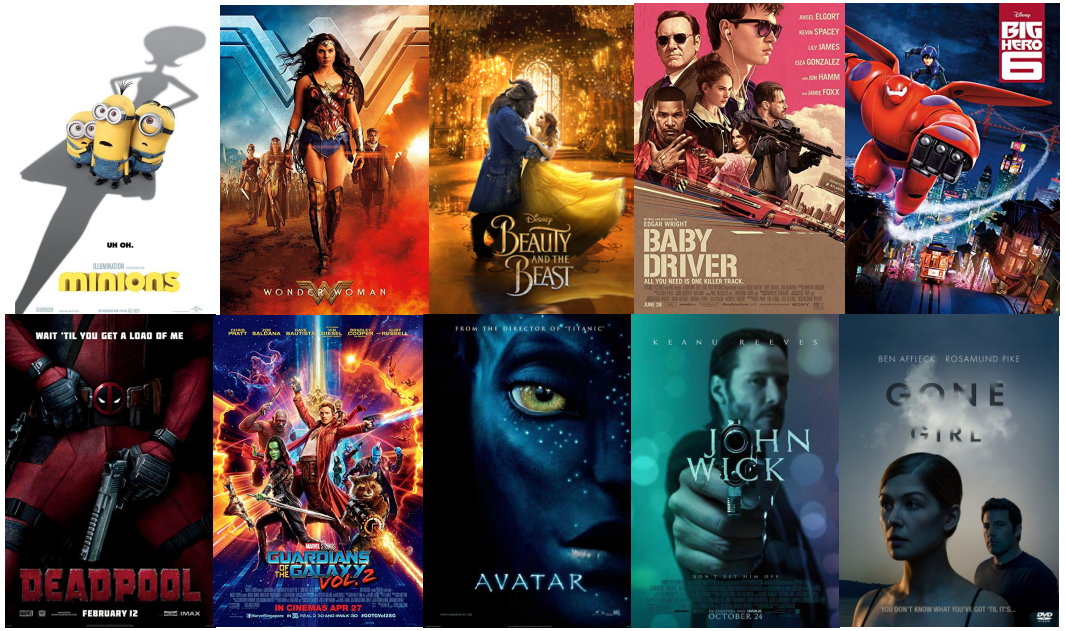
07 Customer Analysis

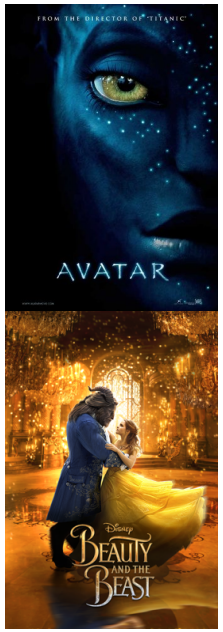


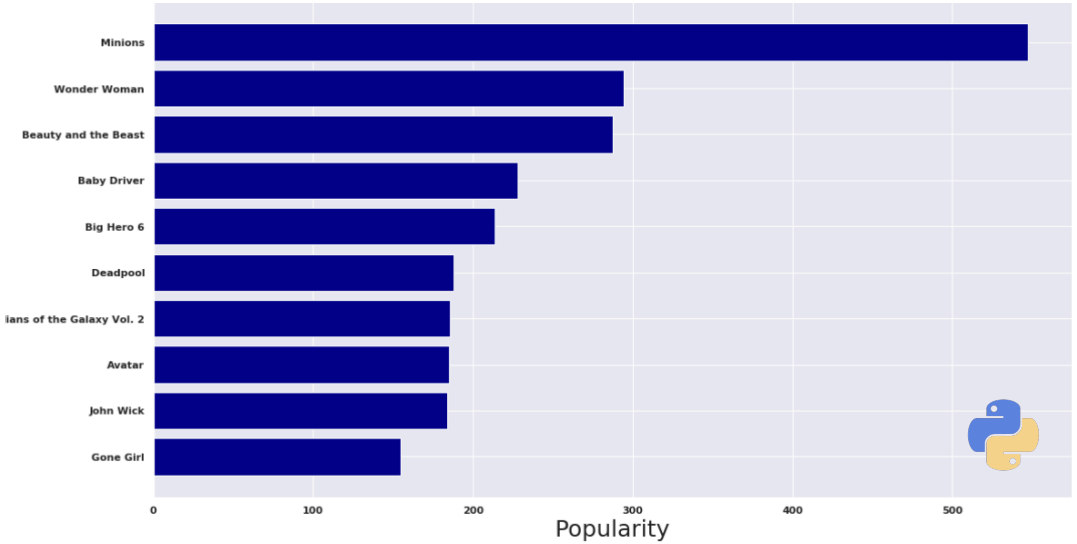


Industry Analysis 08

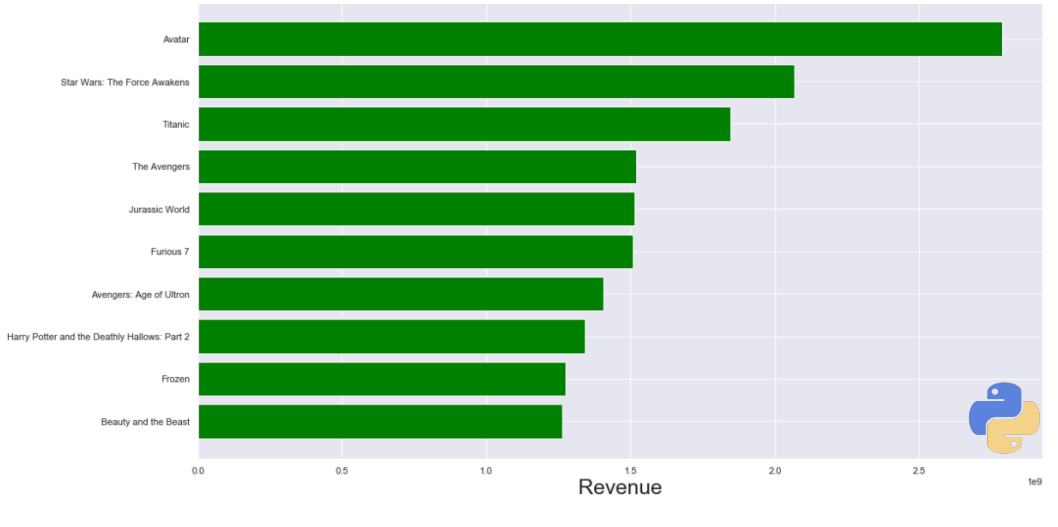
TOP 10 POPULAR MOVIES





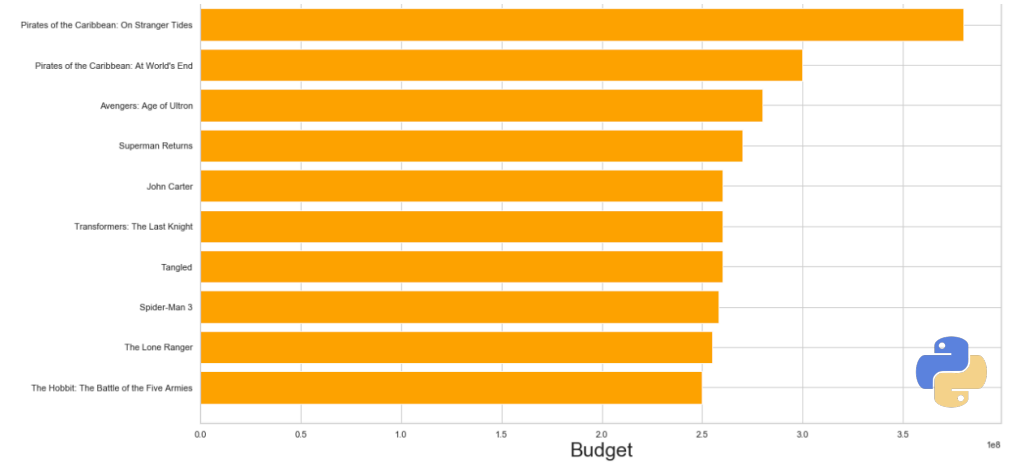


TOP 10 PROFITABLE MOVIES



Industry Analysis 08

TOP 10 EXPENSIVE MOVIES

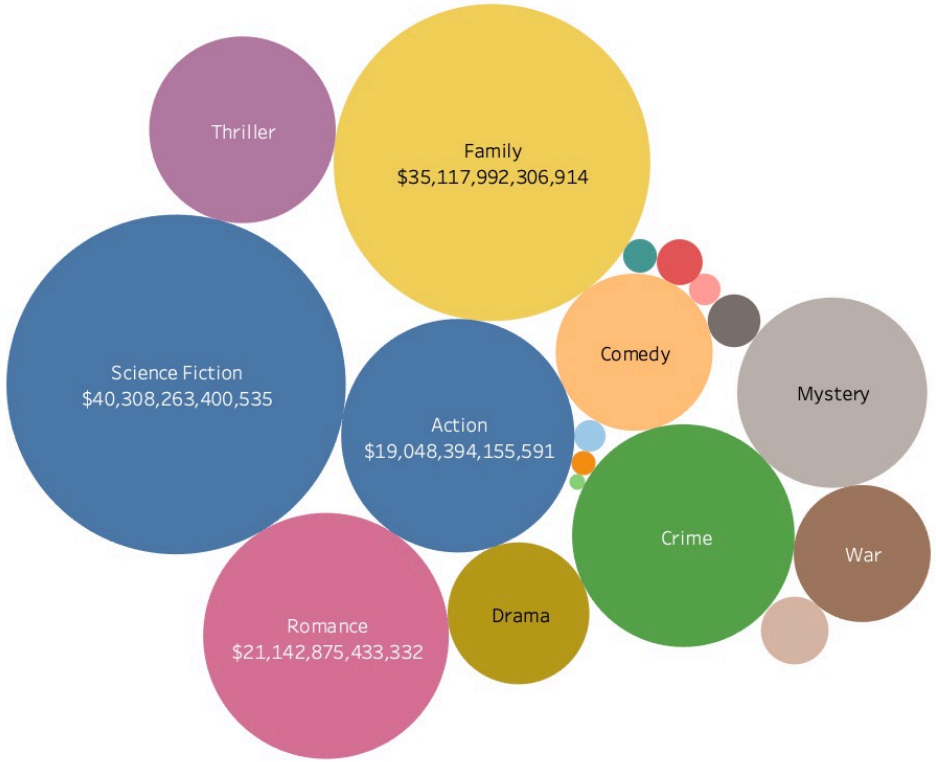


CORRELATION HEATMAP

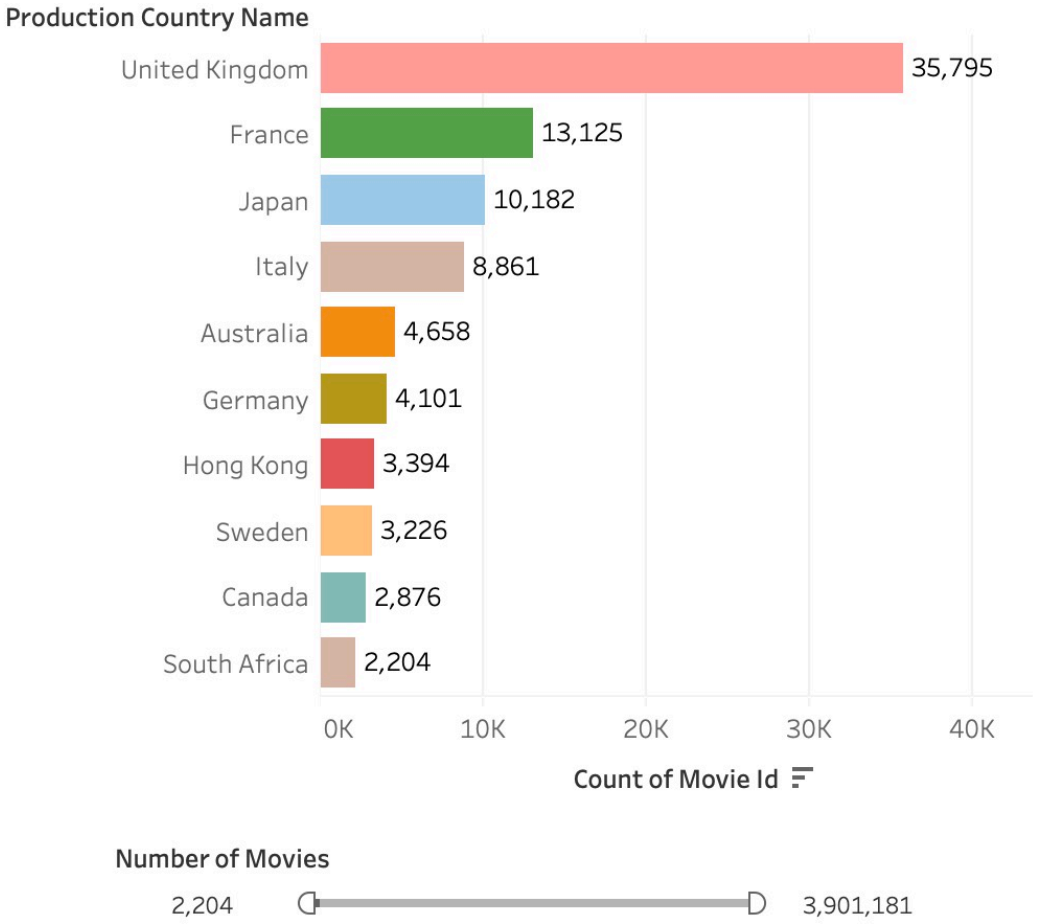
Profitability is highly correlated with budget & popularity, but not runtime

Industry Analysis

Movie Genres that Make the Most Revenue

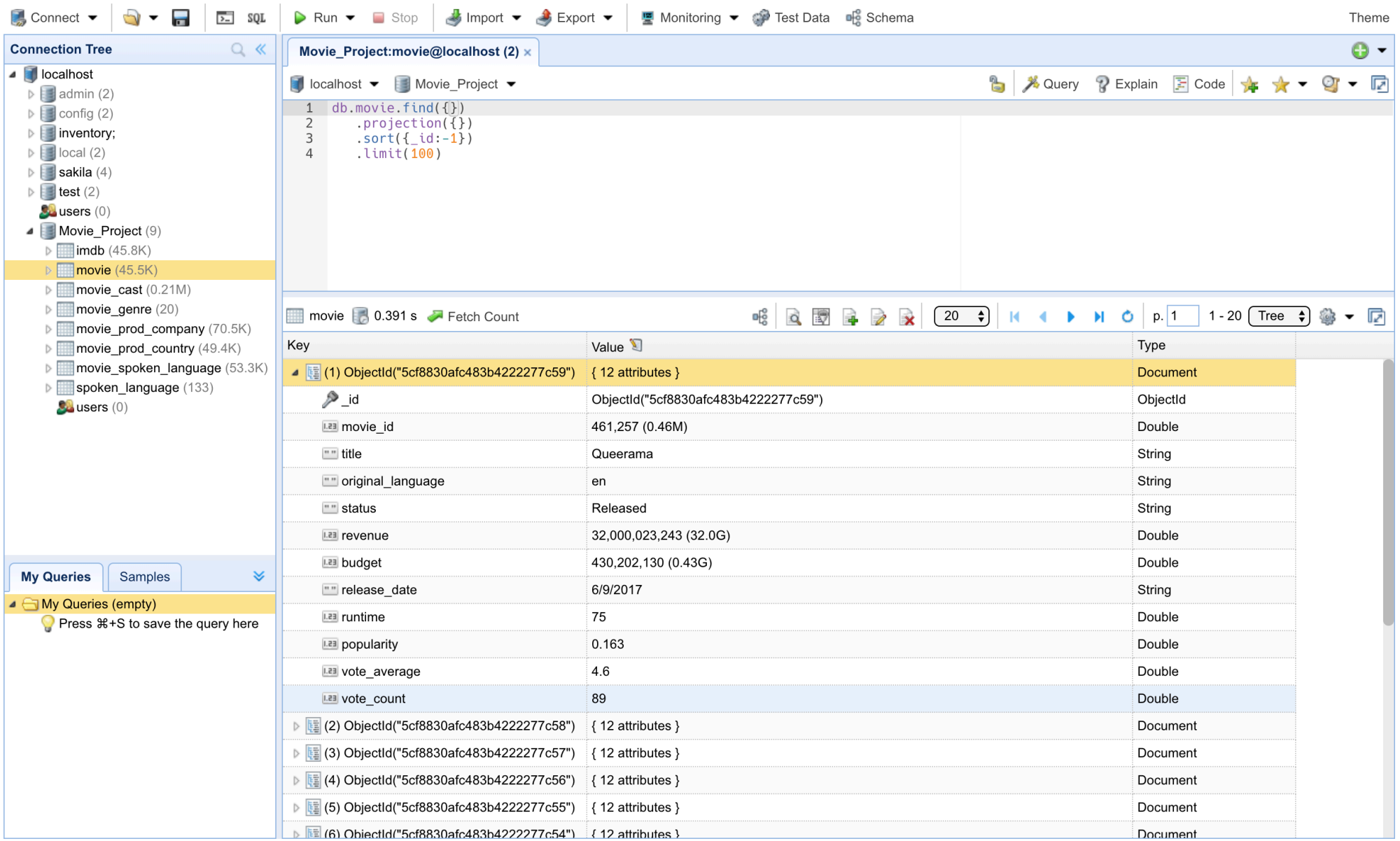


World Top 10 Movie Production Countries (Outside U.S.)
- For US investors

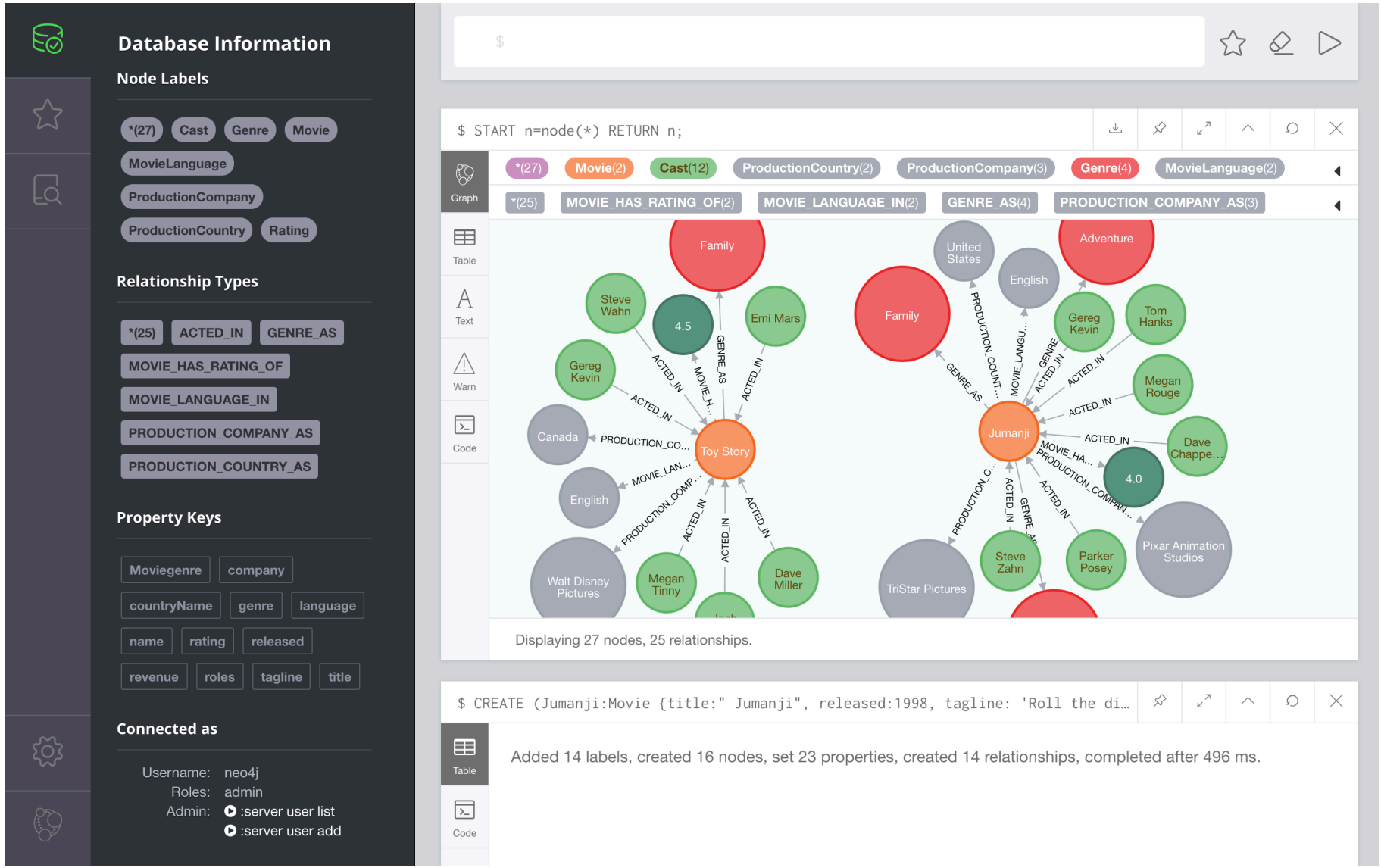


NoSQL - MongoDb

- ✓No need of mapping the application objects to the data objects.
- ✓Secure with the NOSQL database nature and no SQL injection can be made.
- ✓Flex easily to user needs and it's conveniently easy to set up.



NoSQL - Neo4j



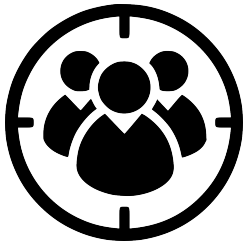
- ✓ Schema-free: Schema-free like other NoSQL databases.
- ✓ Easy representation: Provides a very easy way to represent connected and semi-structured data.
- ✓ Fast Execution and Retrieval: Connected data is very easy to retrieve and navigate, faster other databases comparatively.
- ✓ Performance: Performance remains high even if the amount of data grows significantly.
- ✓ No Join: Doesn't require complex Joins to retrieve connected/related.

10 Recommendations

Industry Analysis

BUILD SUCCESS METRICS

- Audience Targeting & Hidden Patterns
- Discovery & Blockbuster Features Prediction



Customer Analysis

MOVIE RECOMMENDATION SYSTEM

Recommend Movies to Users

USER

Match similar users to make recommendations

CONTENT

Choose from features like genres, cast, language, etc. to make proper recommendations

HYBRID SYSTEM

11 Lessons Learned

DATA PREPARATION

- Find high quality datasource to save time at the project initiation stage
- Arrange more time for data cleaning, ingestion & preparation
- Ensure data accuracy & completeness
- Simplify the process of creating snowflake

ETL PROCESS

- Schedule enough time for large dataset extracting, transforming & loading
- Get rid of the redundant & unnecessary attributes
- Make sure the unique identifiers

DATA ANALYSIS

- Consider the business use cases when analyzing & visualizing data
- Build dashboard and reports based on the stakeholder's perspective

PROJECT MANAGEMENT

- Design the project pipeline & implementation process
- Optimize the procedures & resources allocation

THANK YOU !

QUESTIONS?

GROUP 3
Carrie Meijuan Lu, Jenny Zhihan Wang, Kai Li, Ziwei Zhao

References/Appendix 12

movie_id	A unique identifier for each movie
title	Title of the movie
original_language	The language in which the movie was made
status	"Released" or “Rumored" or “Cancelled”
revenue	The worldwide revenue generated by the movie
budget	The budget in which the movie was made
release_date	The date on which it was released
runtime	The running time of the movie in minutes
popularity	A numeric quantity specifying the movie popularity
vote_average	Average ratings the movie recieved
vote_count	The count of votes recieved
genres_name	The genres of the movie, Action, Comedy ,Thriller etc
production_company	The production house of the movie
production_country	The country in which it was produced
cast	Lead and supporting actors
crew	The name of Director, Editor, Composer, Writer etc
spoken_language	The languages that the movie is translated into

GitHub: https://github.com/KelvinKai/Uchicago_DEP_FinalProject