Term project II

An analysis on movie successes

18 November 2021

Team Bogota:
Péter Kaiser
Rauhan Nazir
Sára Vargha
Xibei Chen

Agenda

01 Introduction

02 Data sources

03 Knime workflow

04 Analytical questions & findings

05 Q&A

Introduction

Step 1

Data collection

We collected data from Kaggle, GitHub and various APIs (World Bank, OMDb, Eurostat)

Step 2

Data modeling

For creating our data model, we used MySQL to create a DB from our Kaggle dataset and Knime to connect our data sources to each other

Step 3

Analytics & visualization

For analytics and visualization we used various Knime nodes and represented our results with bar charts and a scatter plot

Scope of analysis

Uncover the relationship between the success of different movies - measured by the number of awards won and the box office revenue - and various economic factors, such as the government expenditure on education or the population of cultural employment.

Data sources

Movies data

Kaggle

The dataset provided information on **title, genre, year of production, origin, language, runtime, ratings and availability on streaming platforms of 9,515 films** and was last updated in August 2021

OMDb API

The OMDb API was used to get information on: **(1) box office revenues**, and **(2) awards won by the films**

Economic data

GitHub

To get the **country codes** based on which countries are identified in the World Bank and Eurostat APIs, we used an intermediary data table from GitHub

World Bank API

We extracted **(1) GDP per capita and (2) government spending on education** from the World Bank API.

Eurostat API

For European countries, we retrieved information from the Eurostat API on their **population of cultural employment**

Knime workflow I.

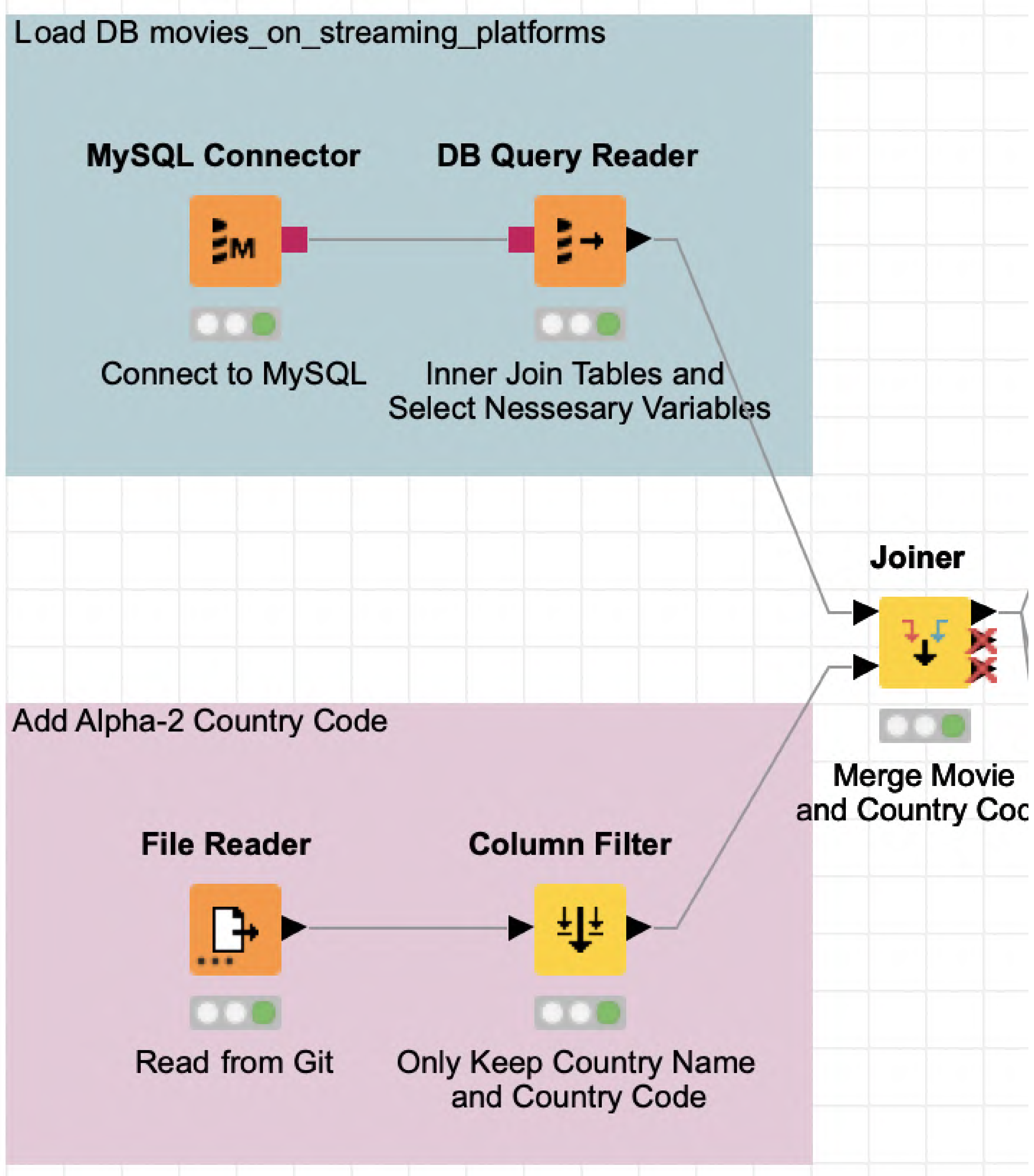
Preparation

Step 1

Load Kaggle dataset into MySQL and create tables to get the database 'movies_on_streaming_platforms'. (using dump for reproducibility)

Step 2

Add Alpha-2 country codes from GitHub data table



Load DB movies_on_streaming_platforms

MySQL Connector

Connect to MySQL

DB Query Reader

Inner Join Tables and Select Nessesary Variables

Joiner

Merge Movie and Country Coc

Add Alpha-2 Country Code

File Reader

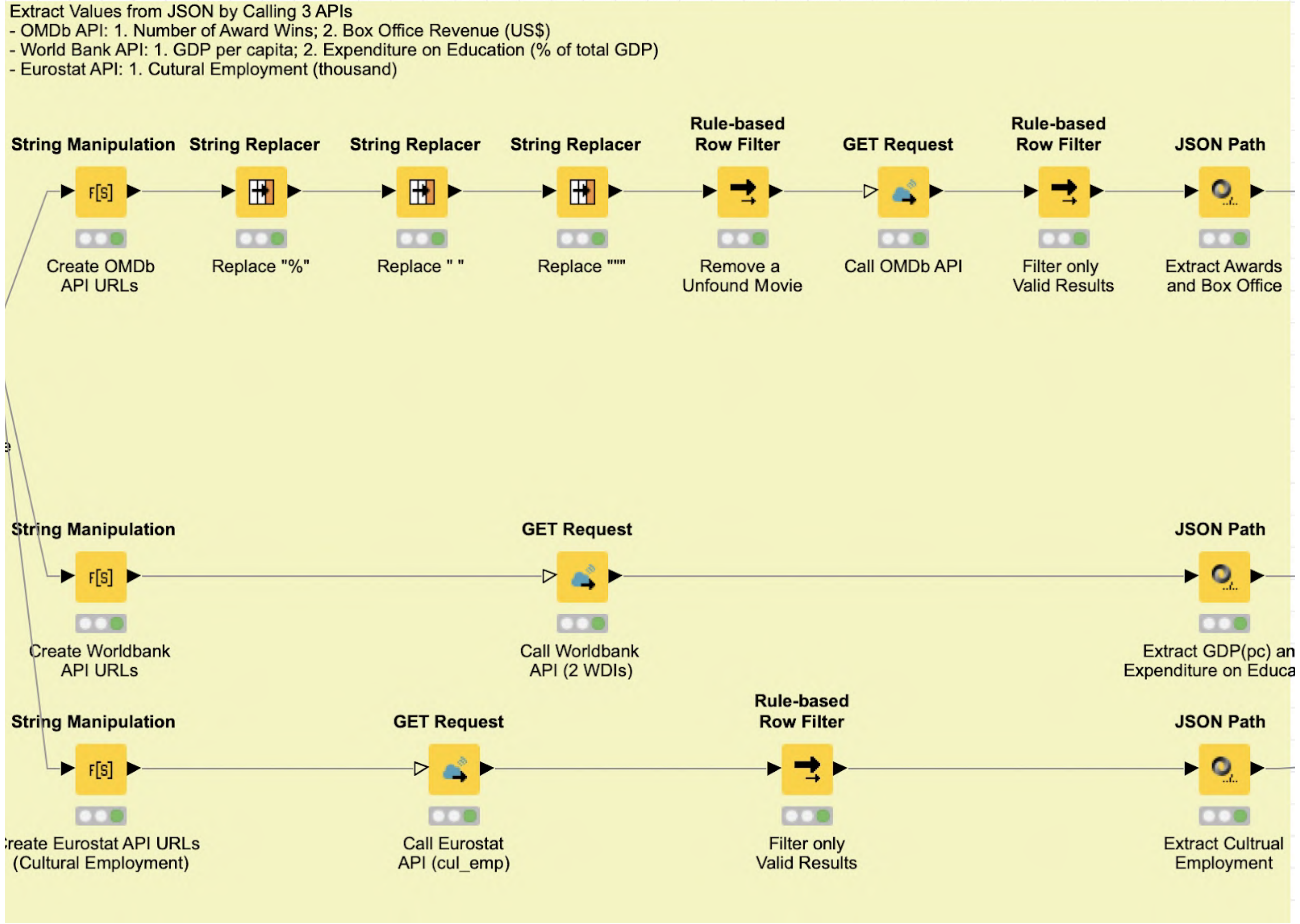
Read from Git

Column Filter

Only Keep Country Name and Country Code

Knime workflow II

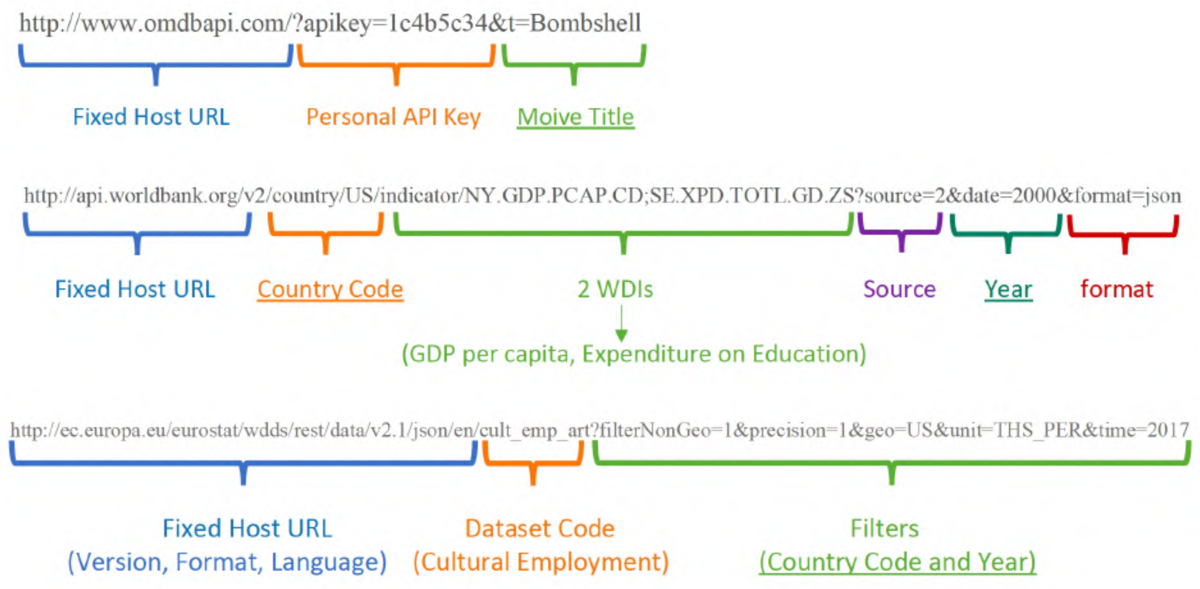
Call APIs and Extract Values



Extract Values from JSON by Calling 3 APIs
- OMDb API: 1. Number of Award Wins; 2. Box Office Revenue (US$)
- World Bank API: 1. GDP per capita; 2. Expenditure on Education (% of total GDP)
- Eurostat API: 1. Cutural Employment (thousand)

Step 1

Manipulate strings to get URLs for APIs



http://www.omdbapi.com/?apikey=1c4b5c34&t=Bombshell

http://api.worldbank.org/v2/country/US/indicator/NY.GDP.PCAP.CD;SE.XPD.TOTL.GD.ZS?source=2&date=2000&format=json

http://ec.europa.eu/eurostat/wdds/rest/data/v2.1/json/en/cult_emp_art?filterNonGeo=1&precision=1&geo=US&unit=THS_PER&time=2017

Step 2

Call the three APIs

Step 3

Filter only valid results

Step 4

Extract the following Values:

OMDb: Awards, Box Office
World Bank: GDP(pc), Expenditure on Education
Eurostat: Cultural Employment

Knime workflow III

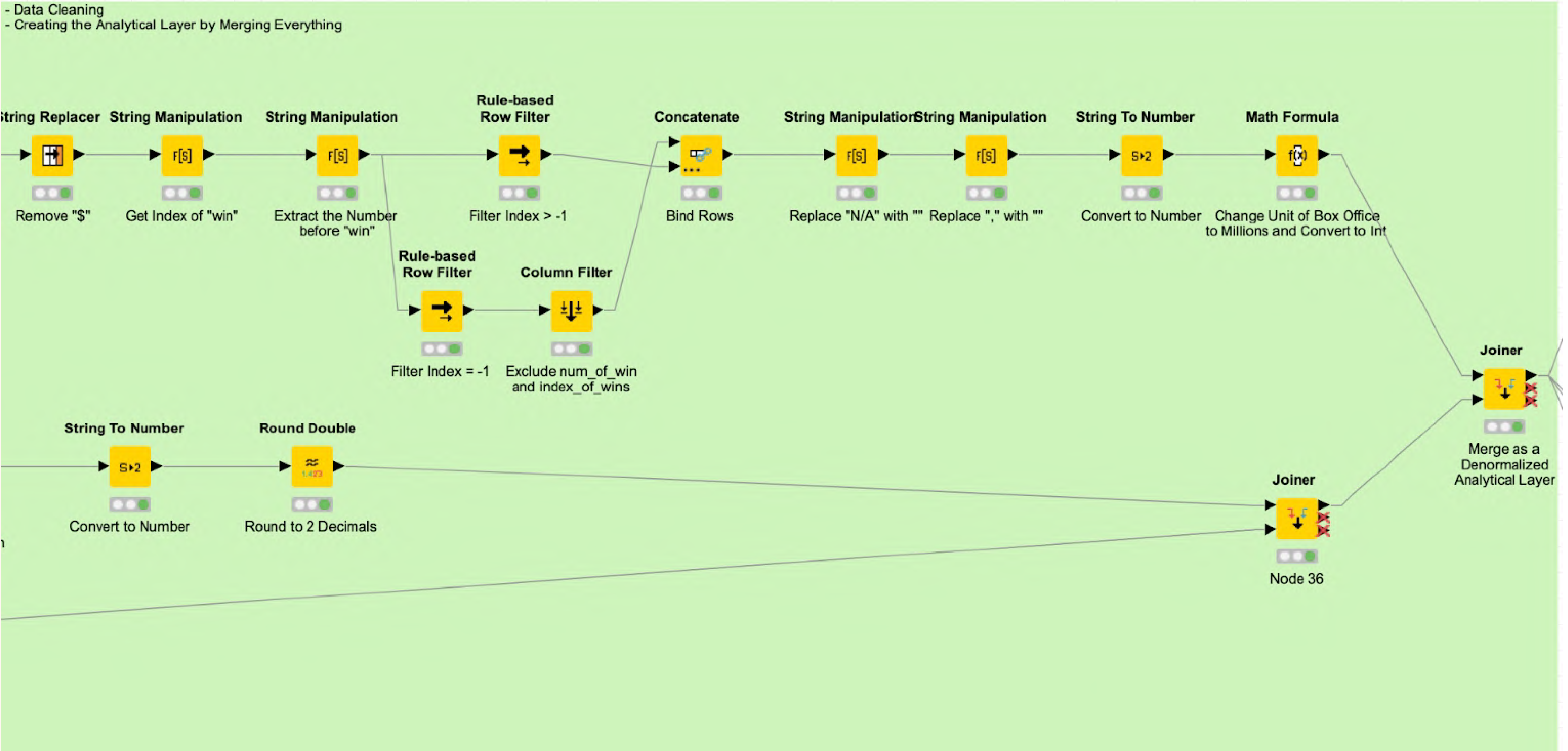
Cleaning and Joining data sources

Step 1

- Covert GDP to number and round;
- Join World Bank and Eurostat values with a full outer join based on the titles of the movies, keeping all data from both data sources

Step 2

- Extract only wins from awards, clean box office to only number without dollar sign;
- Left join the cleaned movies data to the merging result of our World Bank & Eurostat join



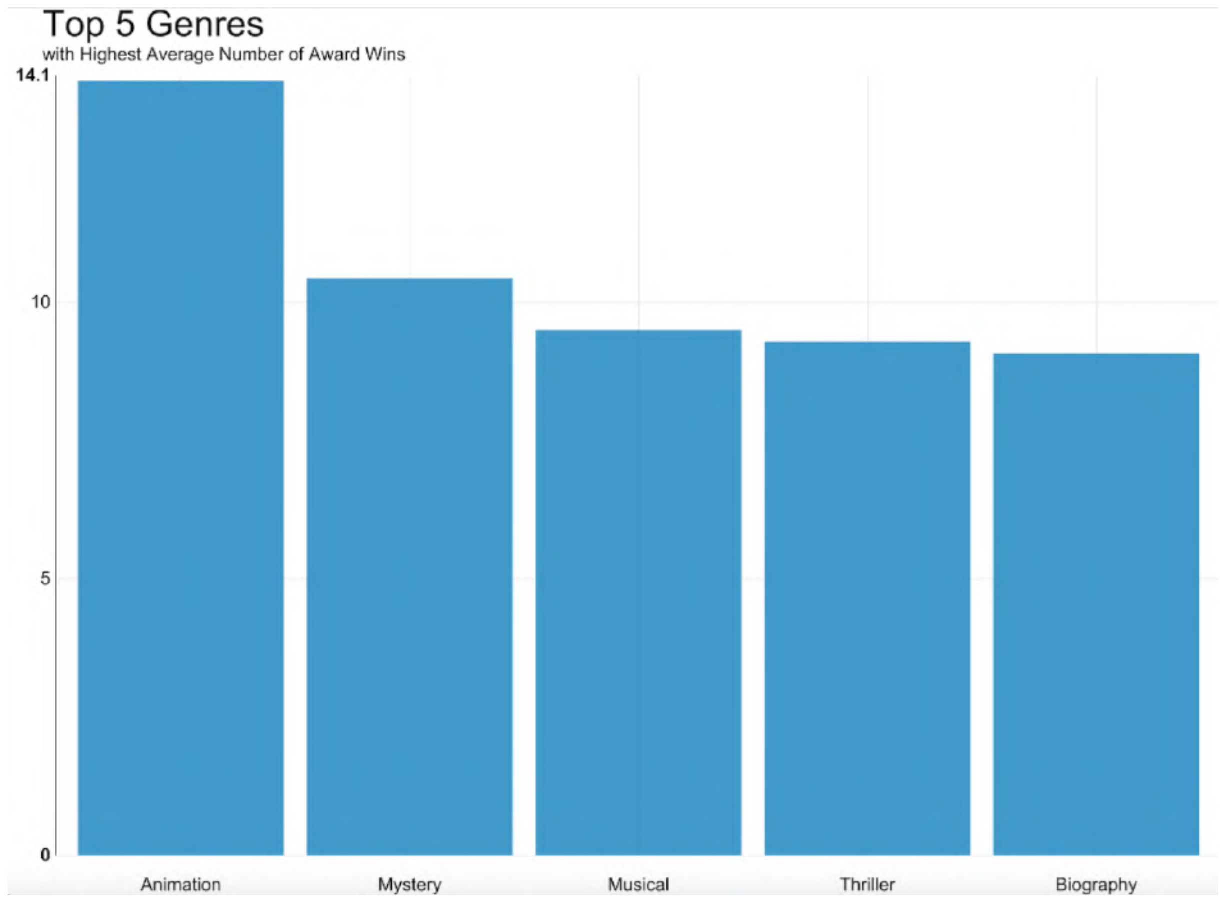
Analytic questions & findings I.

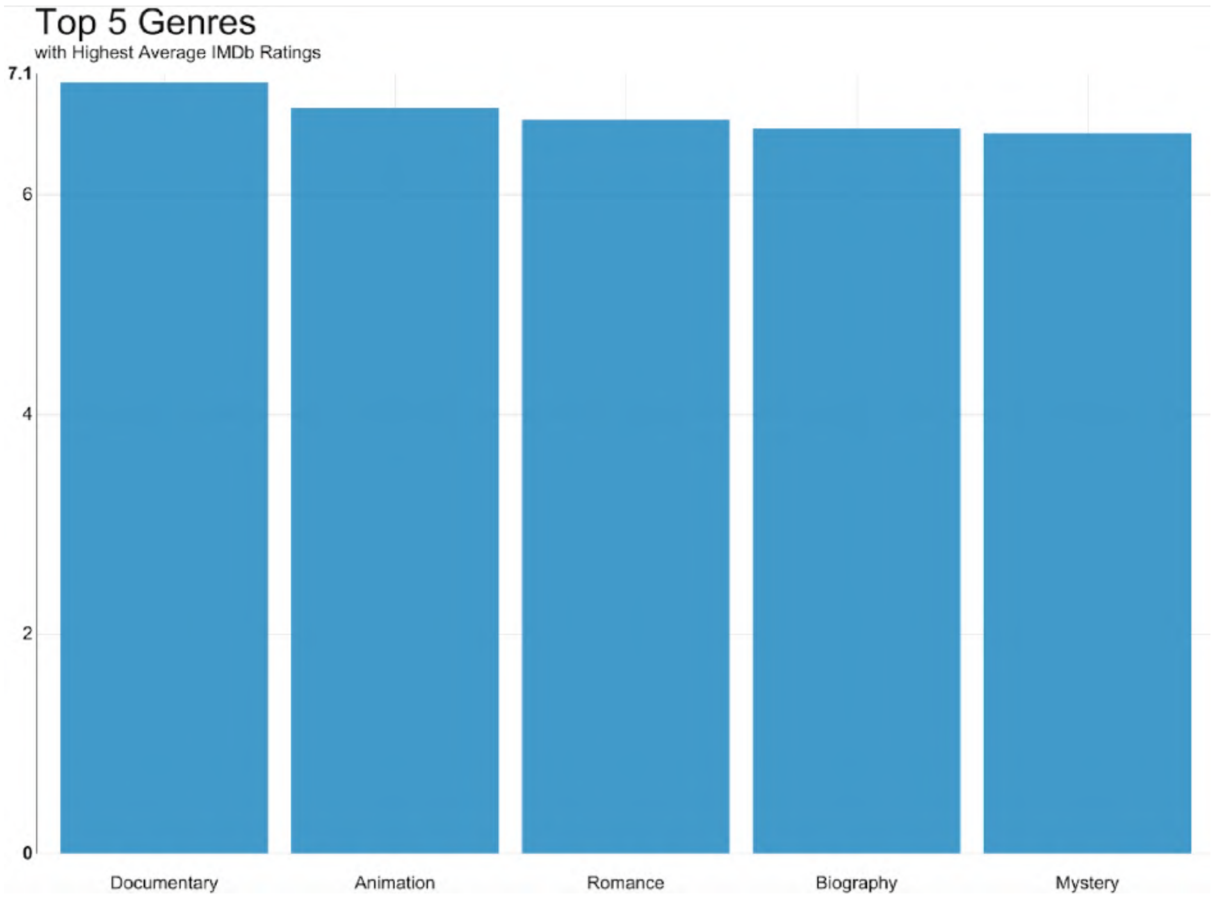
Question 1

Which genres are the top 5 in terms of average number of awards won and IMDb ratings?

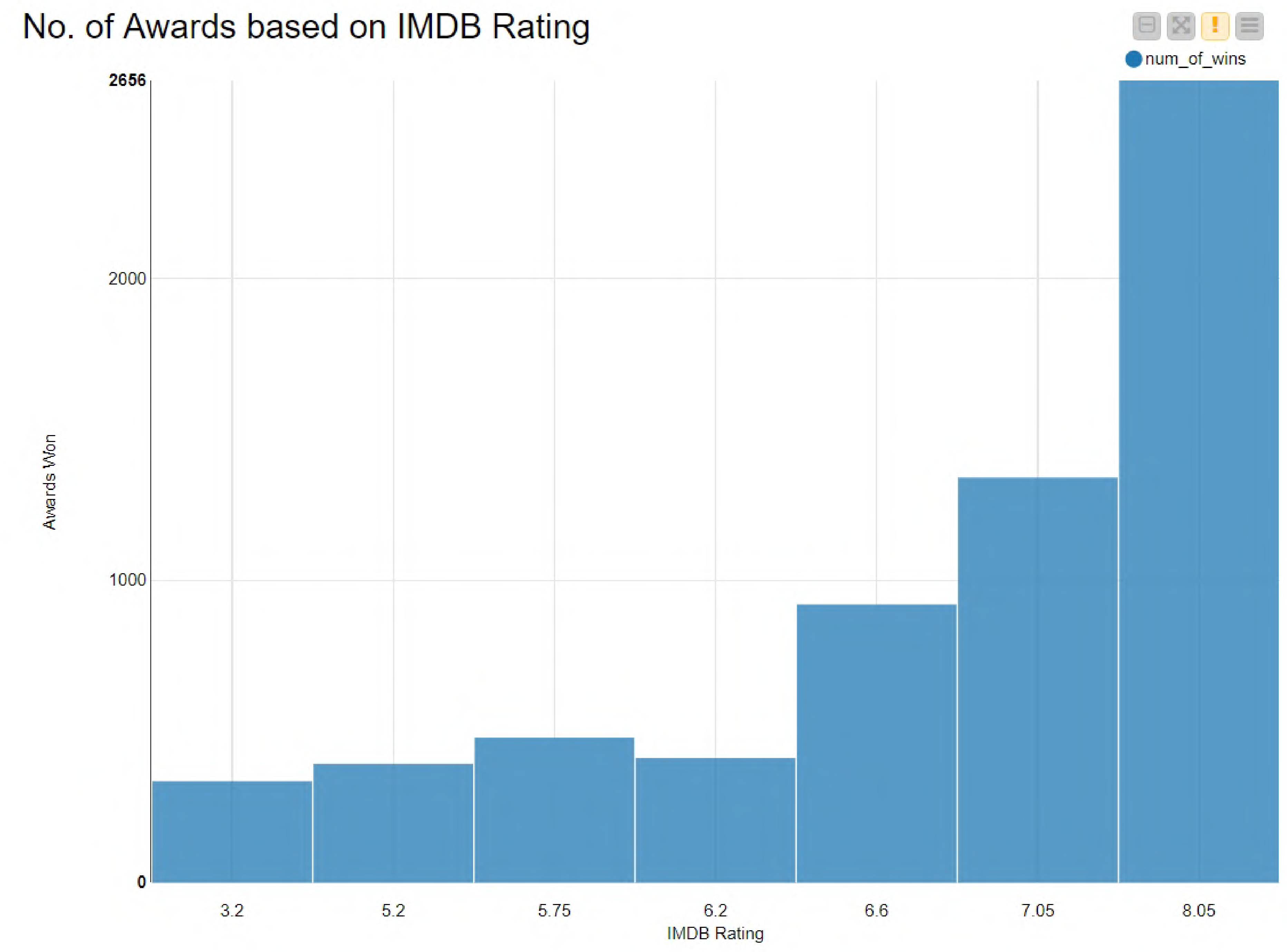
Conclusion

1. Top genres regarding award wins and IMDb ratings are different.
2. Animation wins significantly more awards on average than the 4 other genres.
3. Documentary is the first in terms of average IMDb ratings, however, the difference between the top 5 genres are only slight.





Analytic questions & findings II.



Question 2

What is the relationship between number of awards won and IMDb ratings?

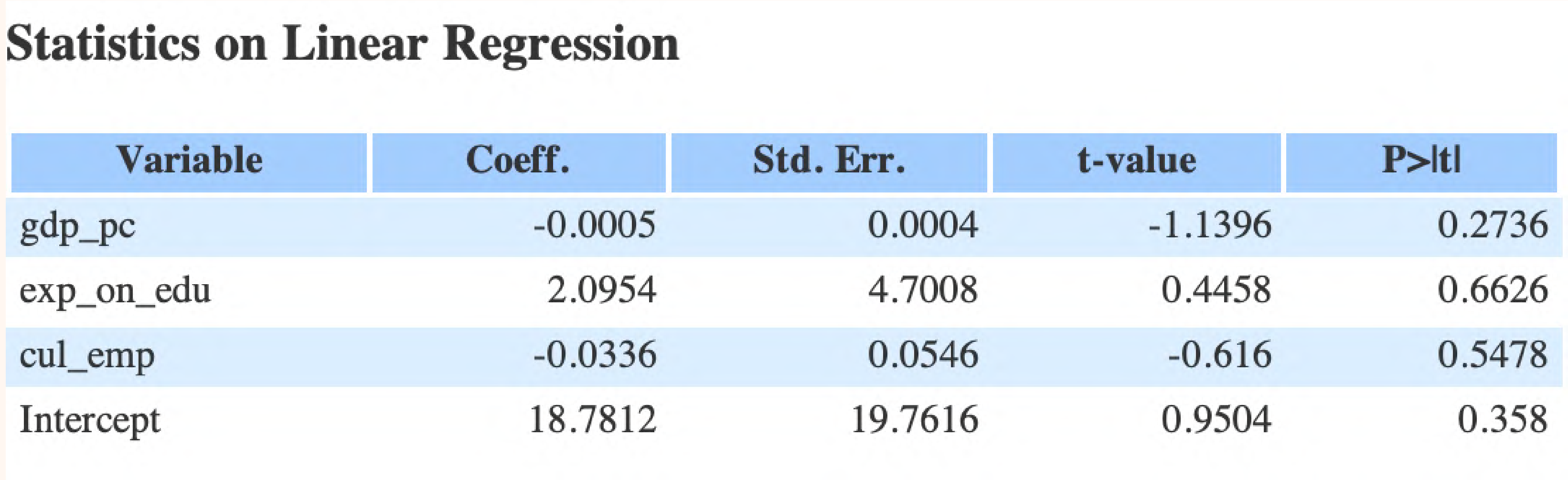
Conclusion

We can say on average, as the rating of the movie increases, so does the performance of the movie in terms of awards won.

Analytical questions & findings III.

Question 3

What is the relationship between the box office revenue of a movie and the GDP per capita, expenditure on education, and population of cultural employment of its country of origin?



Statistics on Linear Regression

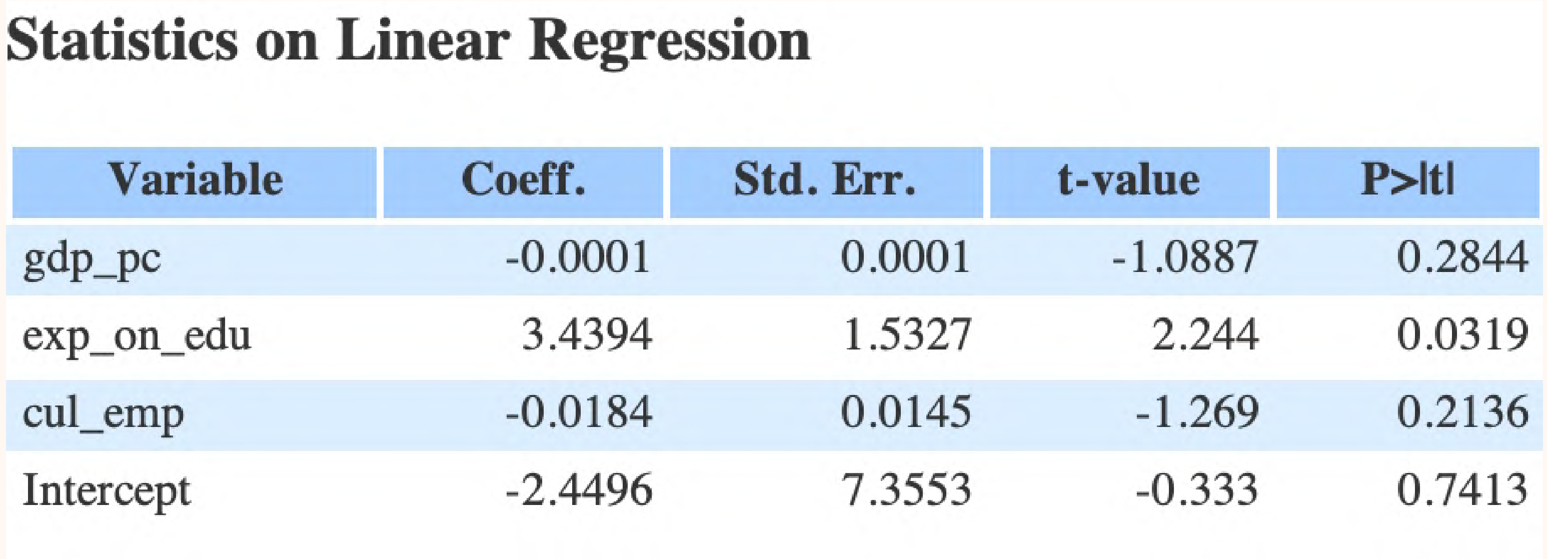
Variable	Coeff.	Std. Err.	t-value	P>|t|
gdp_pc	-0.0005	0.0004	-1.1396	0.2736
exp_on_edu	2.0954	4.7008	0.4458	0.6626
cul_emp	-0.0336	0.0546	-0.616	0.5478
Intercept	18.7812	19.7616	0.9504	0.358

Conclusion

There is no statistically significant correlation between average box office revenue of movies and (1) the country's GDP per capita, (2) expenditure on education or (3) population of cultural employment.

Question 4

What is the relationship between the number of awards won by a movie and the GDP per capita, expenditure on education, and population of cultural employment of its country of origin?



Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
gdp_pc	-0.0001	0.0001	-1.0887	0.2844
exp_on_edu	3.4394	1.5327	2.244	0.0319
cul_emp	-0.0184	0.0145	-1.269	0.2136
Intercept	-2.4496	7.3553	-0.333	0.7413

Conclusion

The number of awards won by a movie is positively related to the country's expenditure on education. However, no statistically significant correlation with (1) the country's GDP per capita or (2) population of cultural employment.

Thank you for your attention!

Q&A