Play Around with Movie Data

- Wrangling Movie Data

Group Members

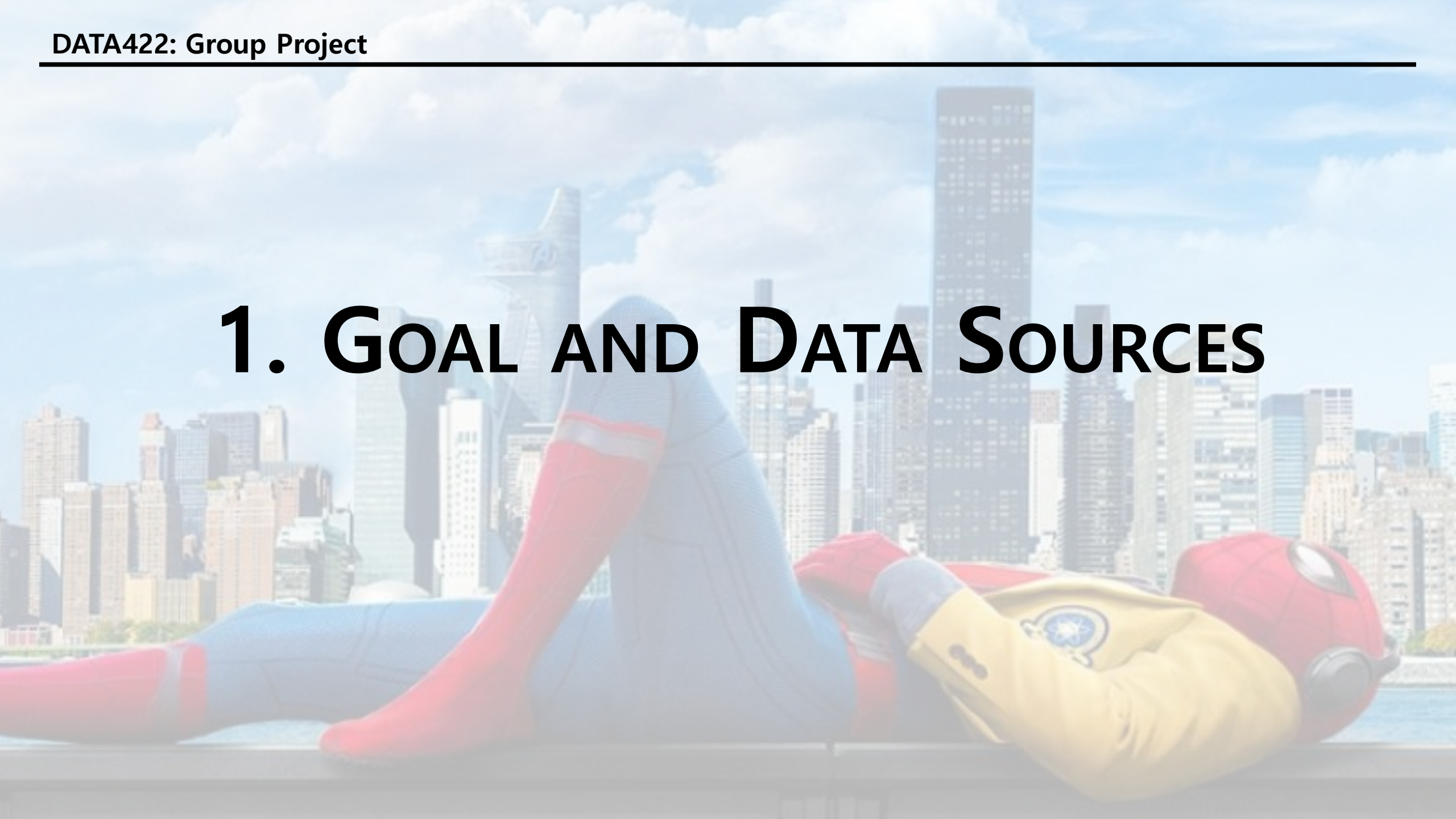
Changmin Moon

Jia Zhiying

Joohyun Ahn

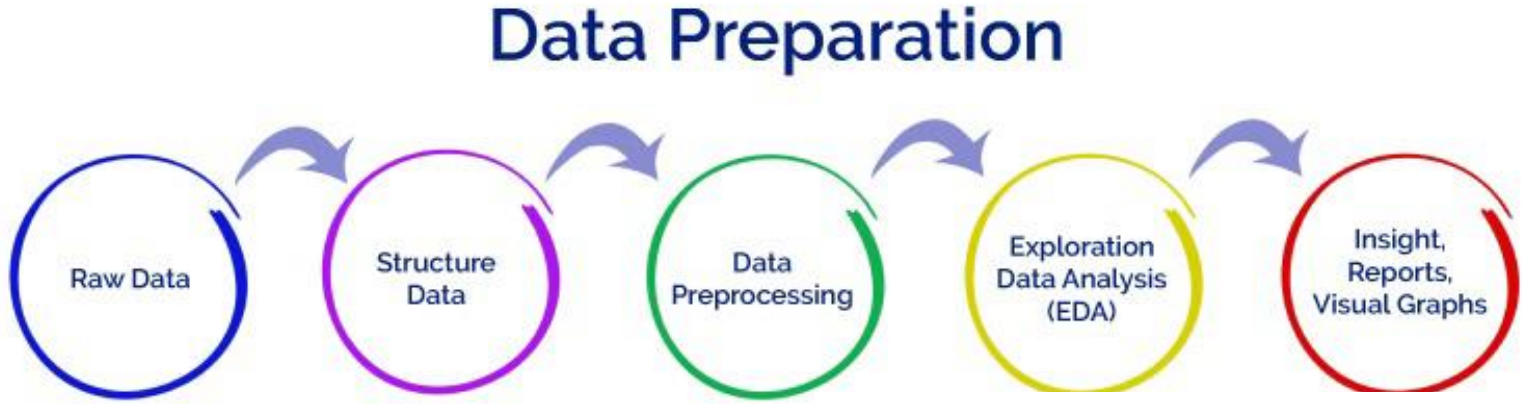
Komlev Viacheslav

INDEX



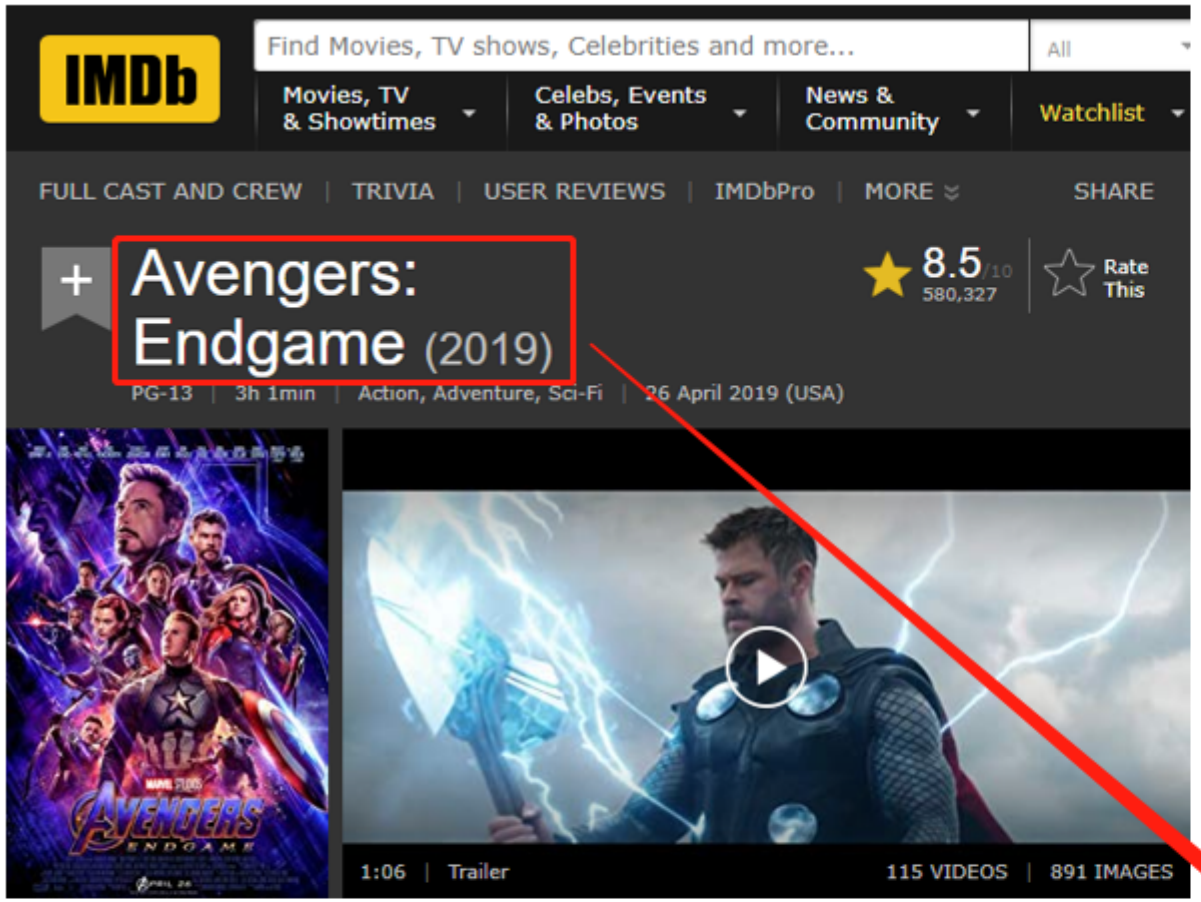
1. Goal and Data Sources

1. Practice and demonstrate the skills we learned in this course



2. Analyze and visualize interesting patterns in movie data

Source: support.ibisos.com/article/381-business-intelligence-solutions



THE NUMBERS®

Where Data and the Movie Business Meet

News | Box Office | Home Video | Movies | People | Research Tools | Our Services | Mobile

Tweet | Like 14 | Share

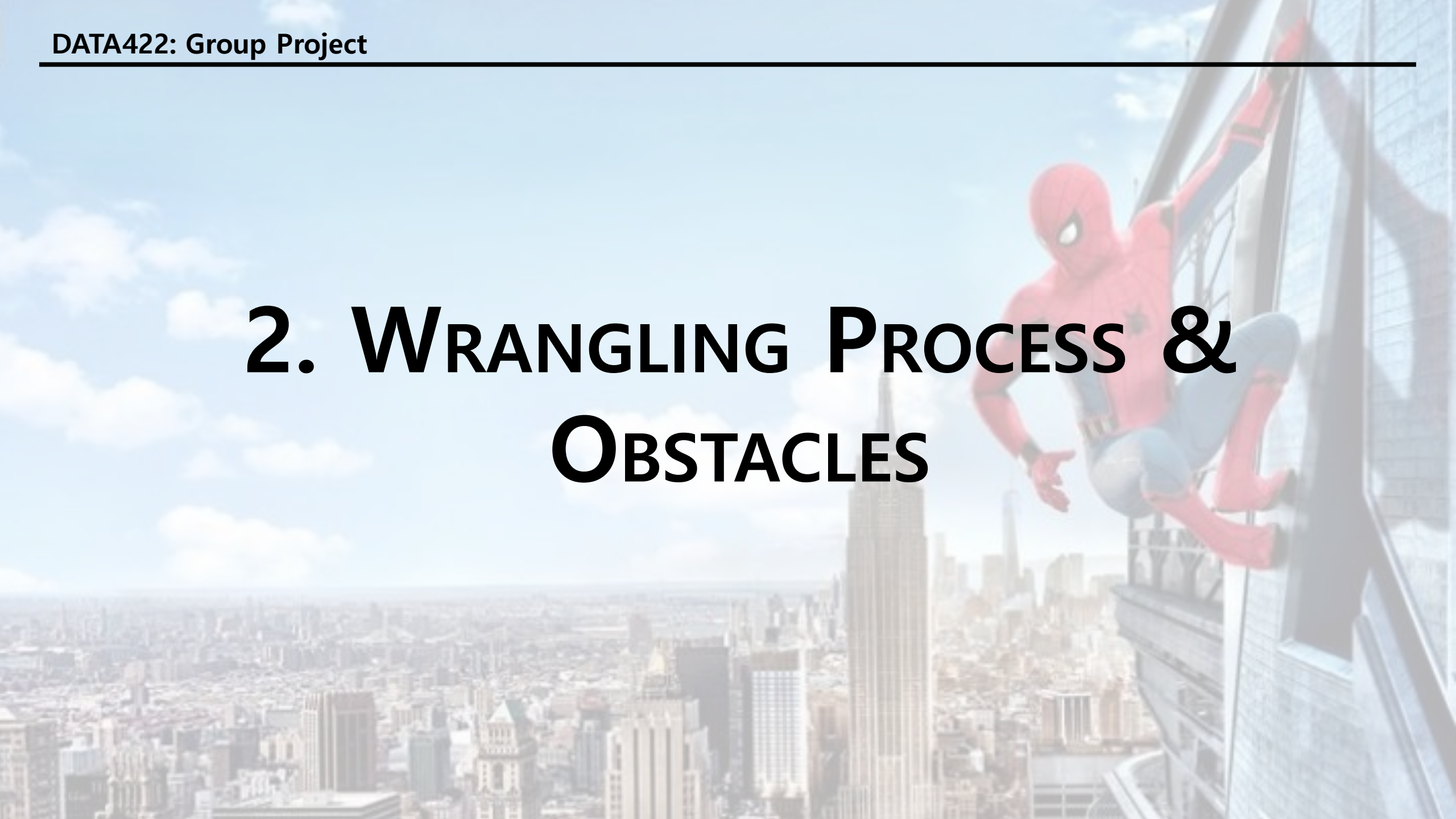
Movie Budgets

Note: Budget numbers for movies can be both difficult to find and unreliable. Studios and film-makers often try to keep the information secret and will use accounting tricks to inflate or reduce announced budgets.

This chart shows the budget of every film in our database, where we have it. The data we have is, to the be of our knowledge, accurate but there are gaps and disputed figures. If you have additional information or corrections, please let us know at corrections@the-numbers.com.

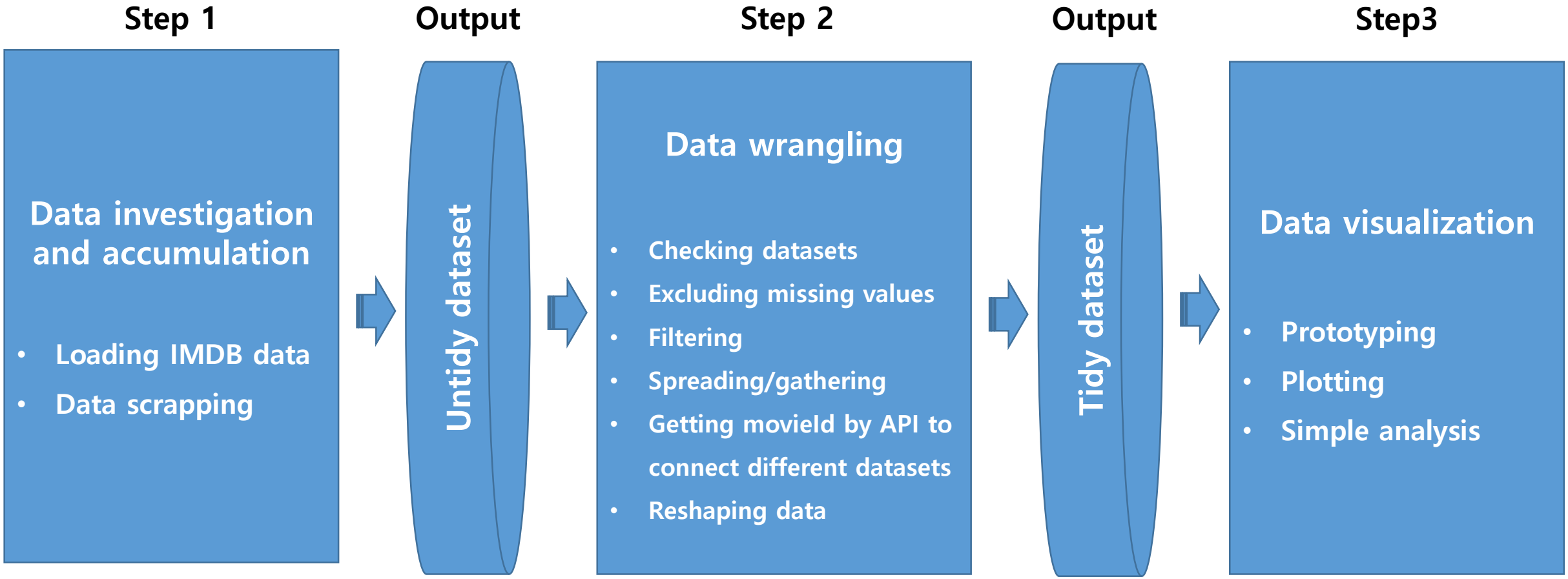
Our movie profit and loss records, based on this budget information, can be found here.

	Release Date	Movie	Production Budget	Domestic Gross	Worldwide Gross
1	Dec 17, 2009	Avatar	$425,000,000	$760,507,625	$2,789,705,275
2	May 20, 2011	Pirates of the Caribbean: On Stranger Tides	$410,600,000	$241,063,875	$1,045,663,875
3	Apr 23, 2019	Avengers: Endgame	$400,000,000	$858,373,000	$2,795,473,000
4	Apr 22, 2015	Avengers: Age of Ultron	$330,600,000	$459,005,868	$1,403,013,963
5	Dec 13, 2017	Star Wars Ep. VIII: The Last Jedi	$317,000,000	$620,181,382	$1,316,721,747
6	Dec 16, 2015	Star Wars Ep. VII: The Force Awakens	$306,000,000	$936,662,225	$2,053,311,220
7	Apr 25, 2018	Avengers: Infinity War	$300,000,000	$678,815,482	$2,048,134,200
8	May 24, 2007	Pirates of the Caribbean: At World's End	$300,000,000	$309,420,425	$963,420,425
9	Nov 13, 2017	Justice League	$300,000,000	$229,024,295	$655,945,209
10	Oct 6, 2015	Spectre	$300,000,000	$200,074,175	$879,620,923
11	Jul 19, 2012	The Dark Knight Rises	$275,000,000	$448,139,099	$1,084,439,099
12	May 23, 2018	Solo: A Star Wars Story	$275,000,000	$213,767,512	$393,151,347
13	Jul 2, 2013	The Lone Ranger	$275,000,000	$89,302,115	$260,002,115



2. Wrangling Process & Obstacles

Project steps



Initial data

IMDB data

Title_principals
- tconst
- ordering
- nconst
- category
- job

Title_basics
- tconst
- primaryName
- titleType
- primaryTitle
- originalTitle
- isAdult
- startYear
- endYear
- runtimeMinutes
- genres

Title_ratings
- tconst
- averageRating
- numVotes

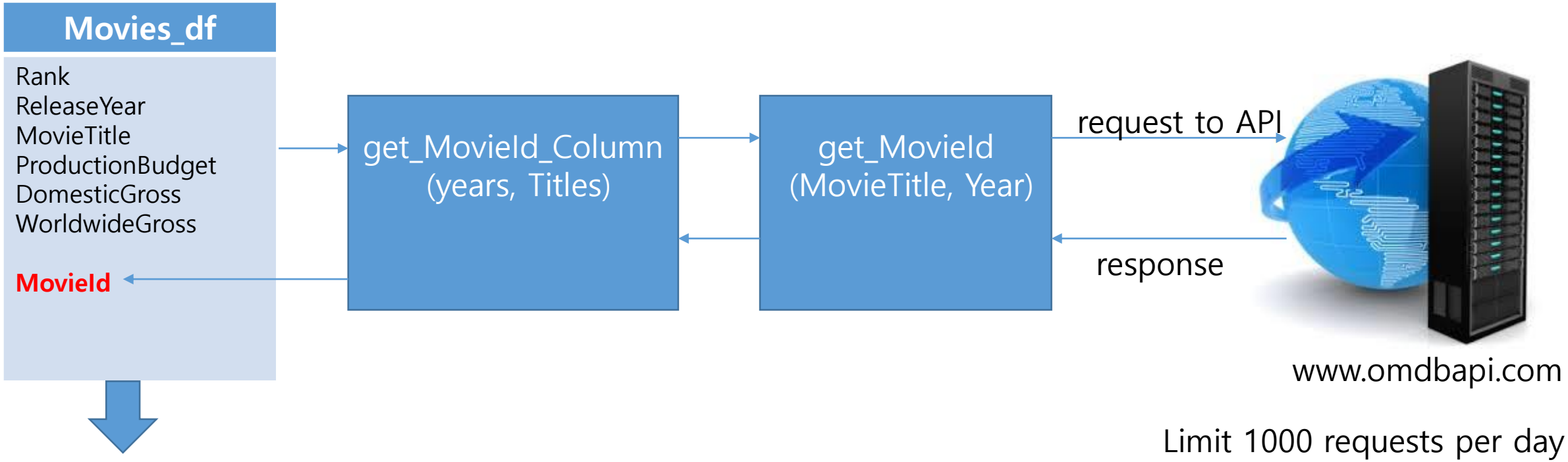
Name_basics
- nconst
- primaryName
- birthyear
- deathYear
- primaryProfession
- knownForTitles

Title_akas
- titleId
- Ordering
- Title
- Region
- Language
- isOriginalTitle

Financial data scrapped from www.the-numbers.com

Movies_df
- Rank
- ReleaseYear
- MovieTitle
- ProductionBudget
- DomesticGross
- WorldwideGross

MovieId retrieval



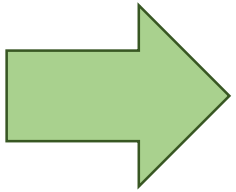
No more than 1000 rows processed per 1 request.
7 new datasets were created with movieId column.

Limit 1000 requests per day
with one API key

Data wrangling techniques

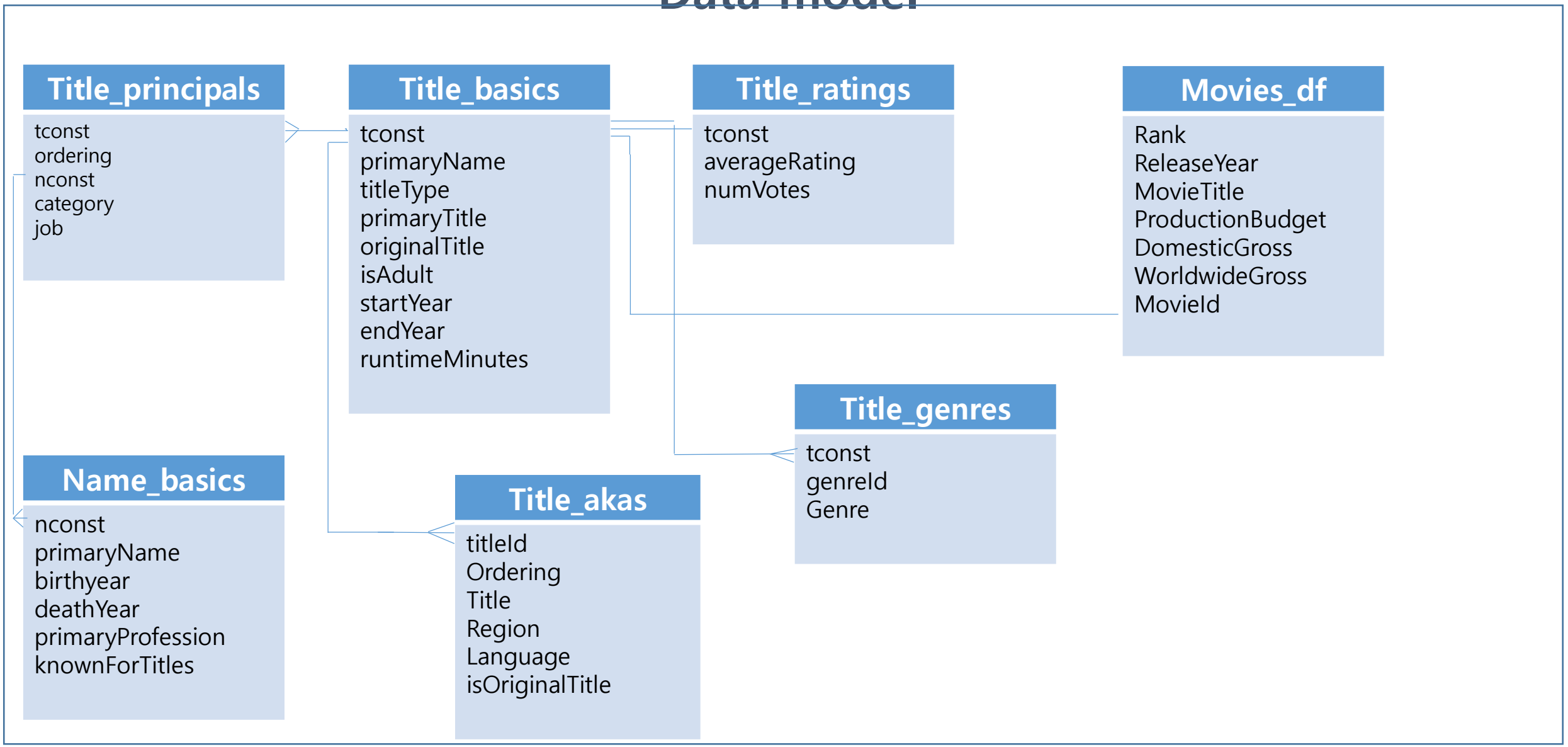
- Data was checked by using R functions glimpse, problems, kable
- Missing data and NA values was excluded so that it could not affect the main idea
- Data was scrapped by using httr library and its functions, also API function was implemented
- Data was partly filtered by years
- Reshaping was implemented (gbind was use to unite several dataframes)
- Mutate was used for creating new columns
- Columns with several values was separated (separate) and gathered(gather) into new tables. For in stance, column Genres.

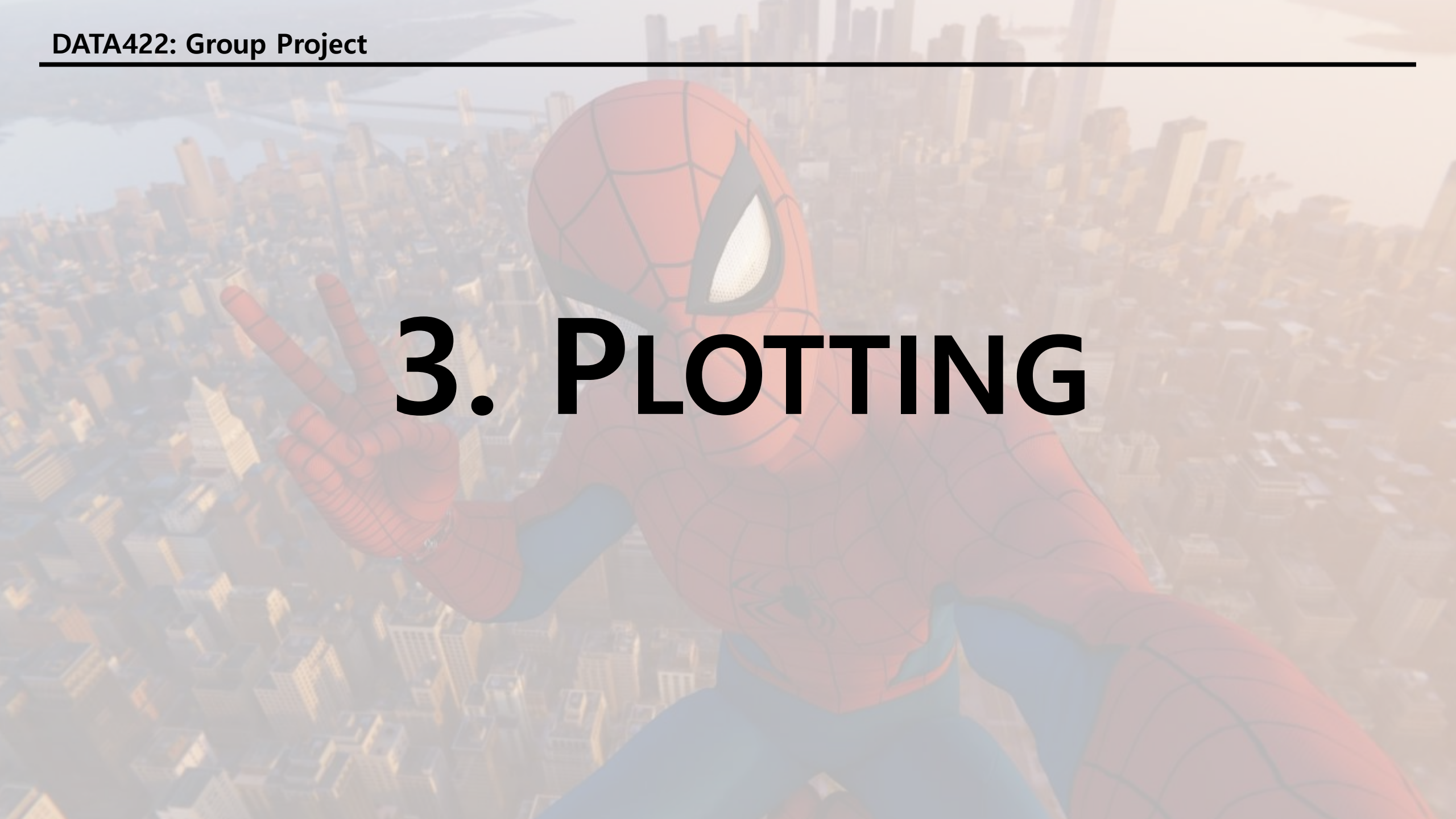
tconst	1	2	3
<chr>	<chr>	<chr>	<chr>
tt0015724	Drama	Mystery	Romance
tt0023331	Documentary	NA	NA
tt0035423	Comedy	Fantasy	Romance



tconst	genreId	genre
<chr>	<chr>	<chr>
tt0015724	1	Drama
tt0015724	2	Mystery
tt0015724	3	Romance

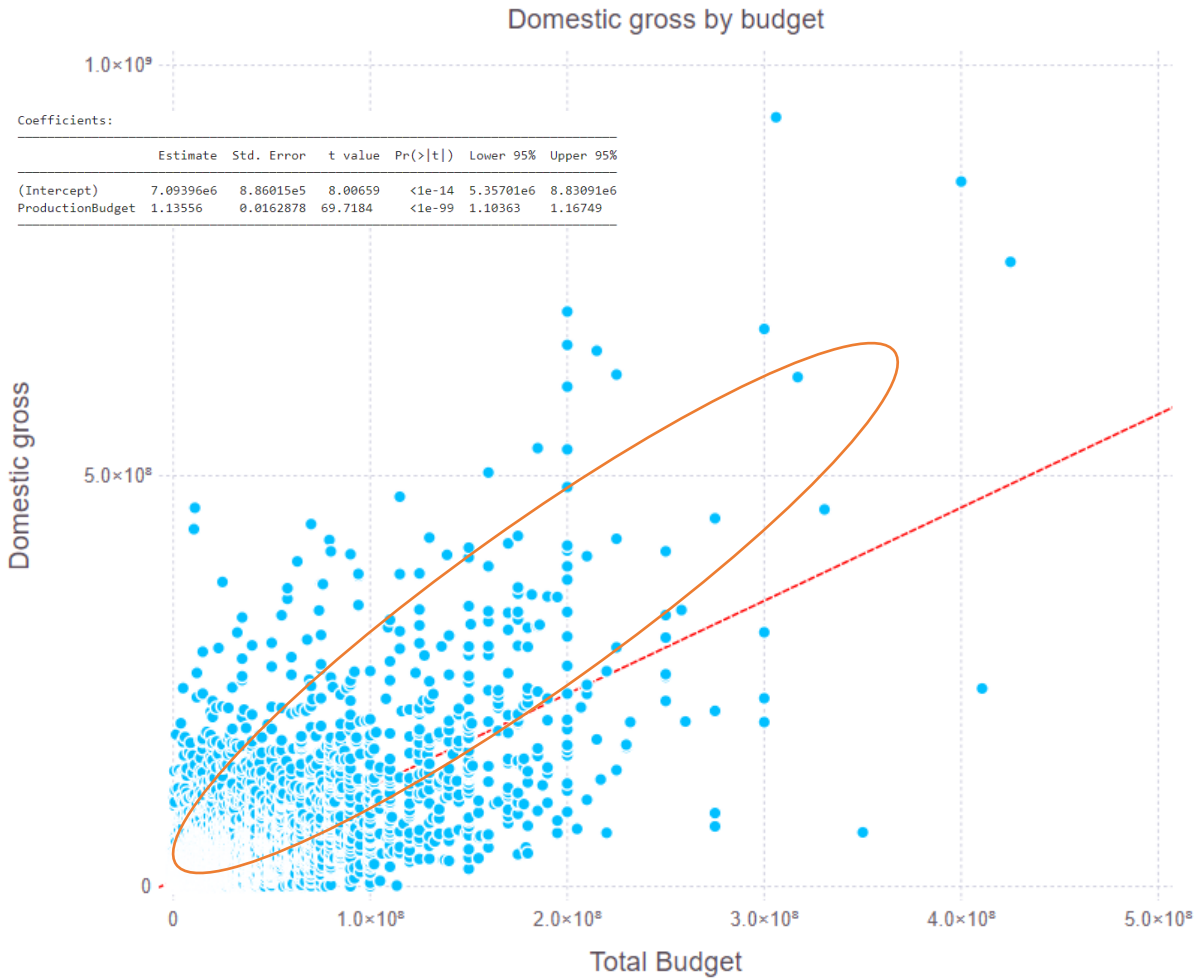
Data model

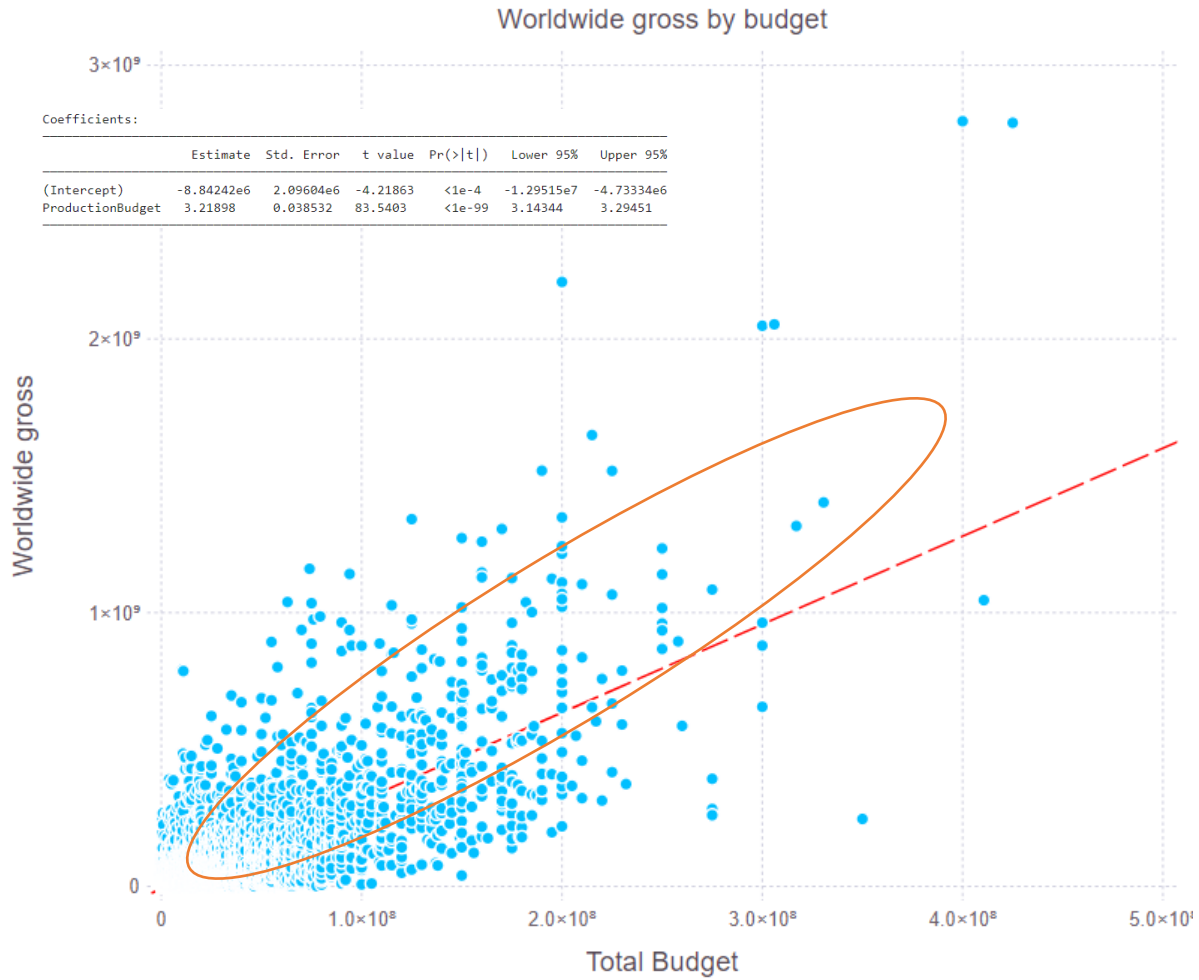

Title_principals
tconst
ordering
nconst
category
job
Title_basics
tconst
primaryName
titleType
primaryTitle
originalTitle
isAdult
startYear
endYear
runtimeMinutes
Title_ratings
tconst
averageRating
numVotes
Movies_df
Rank
ReleaseYear
MovieTitle
ProductionBudget
DomesticGross
WorldwideGross
MovieId
Title_genres
tconst
genreId
Genre
Name_basics
nconst
primaryName
birthyear
deathYear
primaryProfession
knownForTitles
Title_akas
titleId
Ordering
Title
Region
Language
isOriginalTitle



3. Plotting

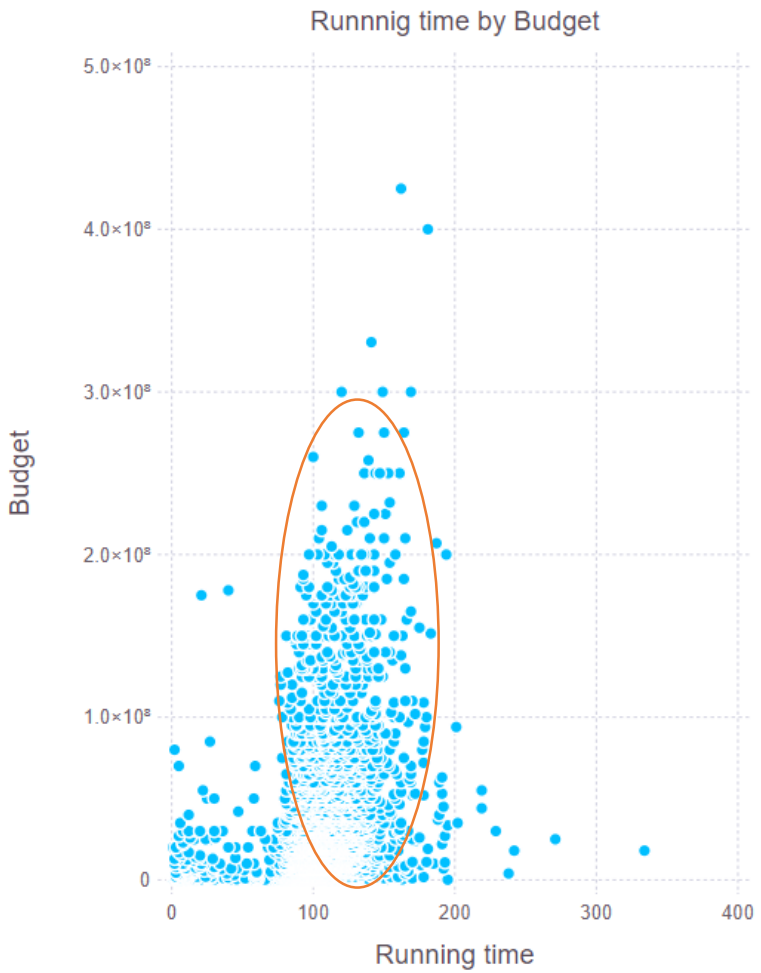
1. Is there any relationship between the marketing budget and successful movies? Yes, it is, it seems to be relevant to some extent.

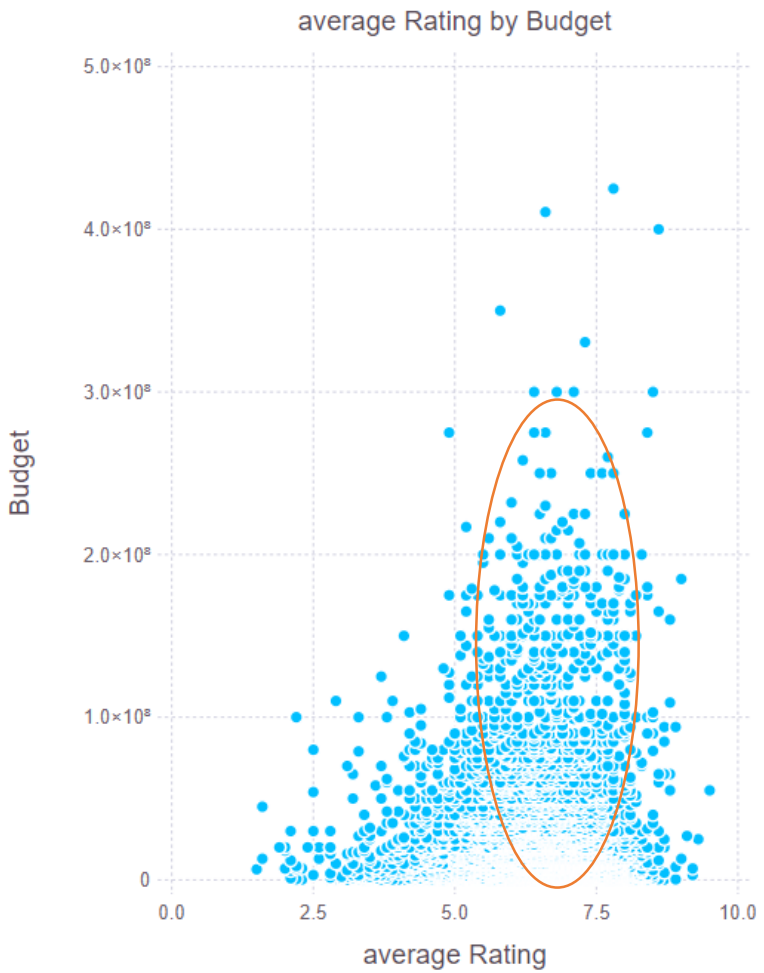


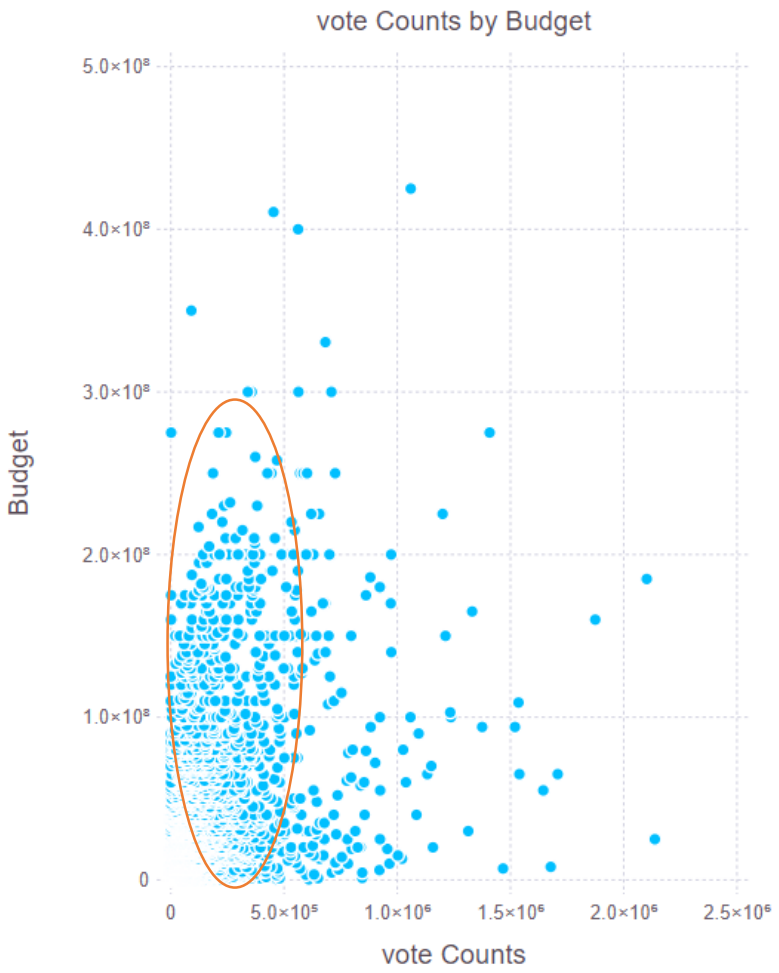


1. So what films tend to attract a lot of marketing budgets?

The running time(100 min ~ 170 min), The rating(6 ~ 8), vote counts seem meaningless

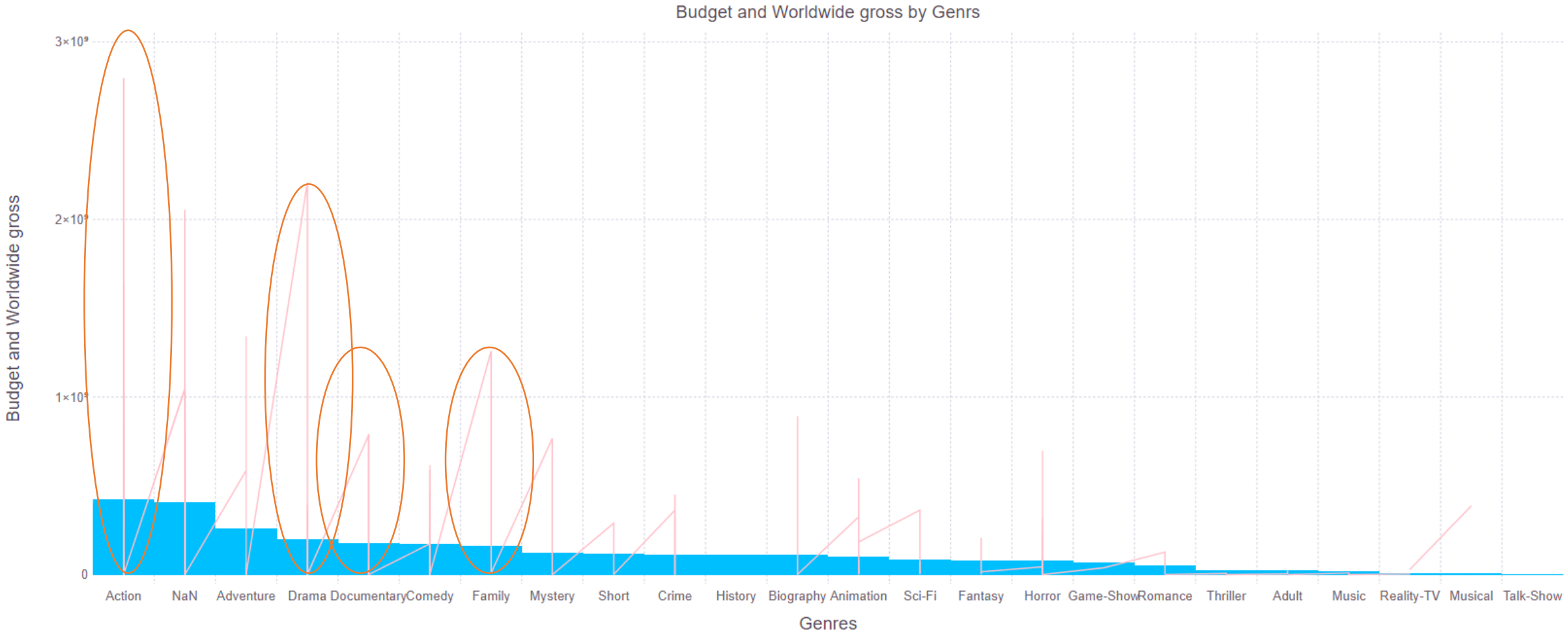






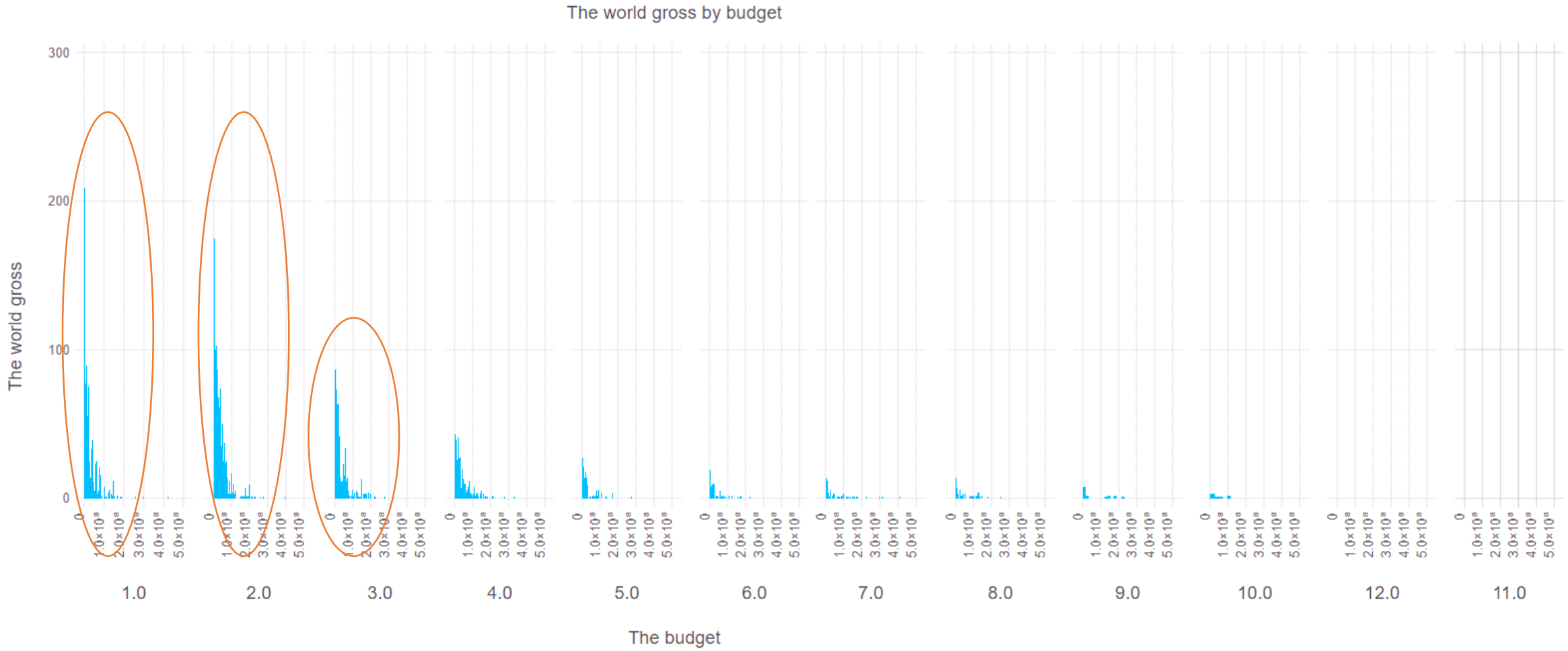
1. What about the relationship with the marketing budget and gross by the genre?

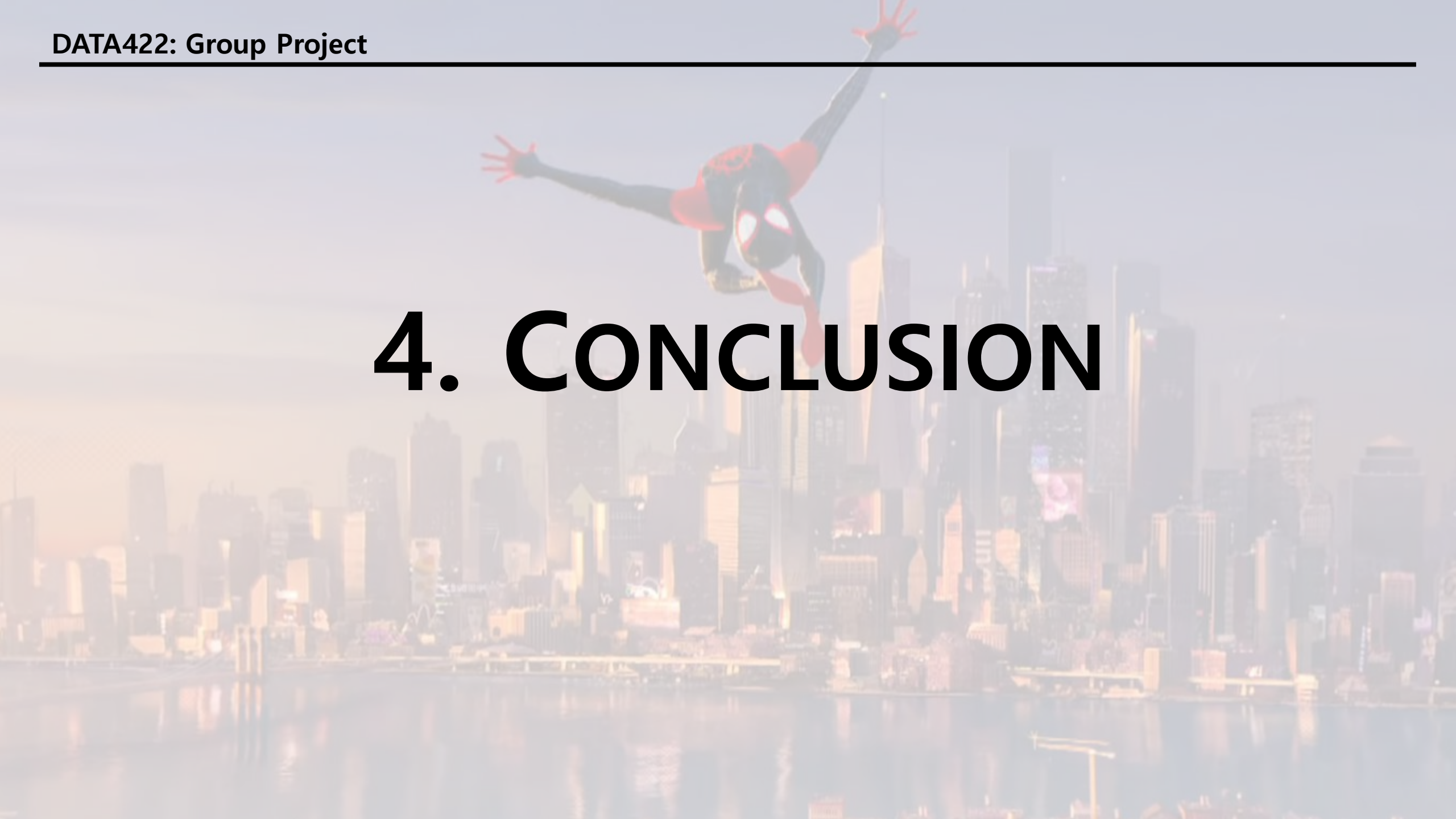
Action, dramas and family films are enjoying relatively good success in preparation for marketing budgets, followed by documentaries.



1. Is there any relationship between the marketing budget and world gross by words of the movie title to success?

Usually, successful movies tend to lie on between 2 words and 3 words





4. Conclusion

1. Do invest in the blockbuster movies
2. Do invest in movies with no more than three hours running-time
3. Do invest in movies with at least 6 rating scores
4. Do not invest unless Action, Adventure, Drama, Family Or Documentary
5. However, if the title of the movie beyond 3 words, give it up!

The insights of the project.

To analyse properly the data wrangling work has to be prepared perfectly. Or not following analysis works can be suffered by untidy dataset every time.

To cooperate with colleagues require some core skills such as communication, organising roles, time management to complete. Due to the complexity of the real world, that situation is very common.

However, two colleagues are better than one, and three are better than 2. Because we could not deal with some problems such as using API, web scrapping without other colleagues.

This movie data were very big and untidy.

CAST

Awesome guy	KOMLEV VIACHESLAV
Smart guy	JOOHYUN AHN
Brilliant guy	JIA ZHIYING
Just korean guy	CHANGMIN MOON
Amazing lecturer	GIULIO VALENTINO DALLA RIVA
Great lecturer	THOMAS LI

THANK YOU FOR LISTENING