Final exam for Business Intelligence

Group: BD – 2004

Prepared for: Zuyeva Aigerim

Prepared by:

Zeinep Kuanyshova

Alua Taszhan

Aidana Myrzakassym

Astana IT University

2022

Step 1.

This report presented data on films and TV series of the Netflix platform, which has been popular since 2013. There are 3 datasets connected in our database, which are closely related and complement each other.

The data relate to the **media business secto**r, since we have a film industry.

The main characteristics of columns that were used or somehow involved in this business decision analysis:

1. String columns: type, title, type, director, cast of actors, rating, duration, genre, language, and description.
2. Number columns: release year, added date, runtime, IMDB score.
3. And one geographical column.

Our **main goal** is to build 4 visualizations that will help/give advice to directors / producers how and what should be done to develop the film industry in their understanding and build 6 visualizations that will be auxiliary to key hypothesis.

Step 2.

Key hypothesis:

1) H0: Netflix is more focused on teenage viewers.

Ha: Netflix is focused on all types of viewers.

2) H0: Films featuring popular actors and released in the top years are gaining high IMDB score.

Ha: Films featuring popular actors and released in the top years are not gaining high IMDB score (does not depend on it).

3) H0: African countries gets the lowest grade results in whole release count and Worldwide Netflix IMDB Score

Ha: Some countries in the African continent are above average score, which proves that IMDB does not depend on Continents

By using top5 genre

4) H0: English - language movies / TV series have high ratings for all age ratings according to IMDb.

Ha: There are another language movies/TV shows in this rating.

Secondary hypothesis:

1) H0: Quarantine has increased the number of releases of TV series and films.

Ha: There are fewer releases due to quarantine.

2) H0: With the release of the TV show movies began to lose their relevance

Ha: Movies are also often produced and developed along with TV shows

3) H0: Comedy, Fantasy, Horror, Action and Drama top 5 highly rated genres. (According to RBK experts)

Ha: Comedy, Fantasy, Horror, Action and Drama are not included in the top 5 genres.

4) H0: Usually Documentary films have longest run time in duration among all genres.

Ha: Duration does not depend on genre; each genre has films of different characters and accordingly different Run time.

5) H0: The most common language after English in Netflix Content is Spanish language.

Ha: According to average statistics, the number of movies in Spanish lags behind Hindi, which occupies the respective side of the place after English.

6) H0: The directors of "romantic" films are women.

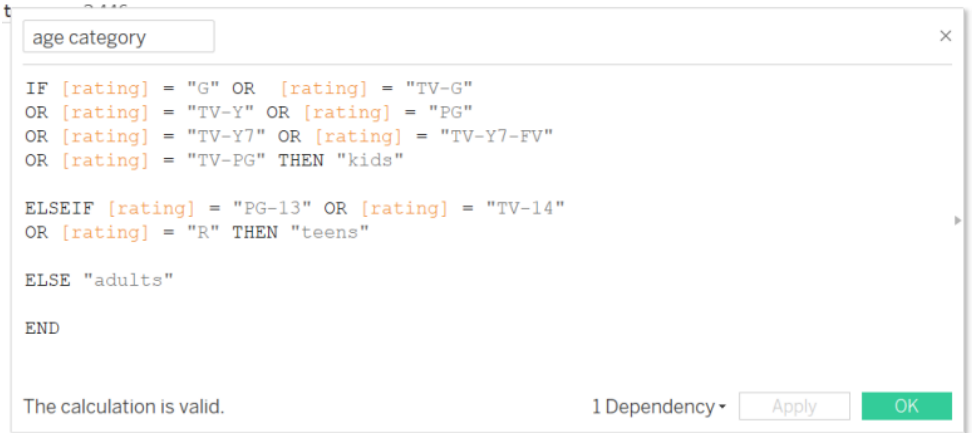
Ha: Films about love are most often made by men.

Step 3.

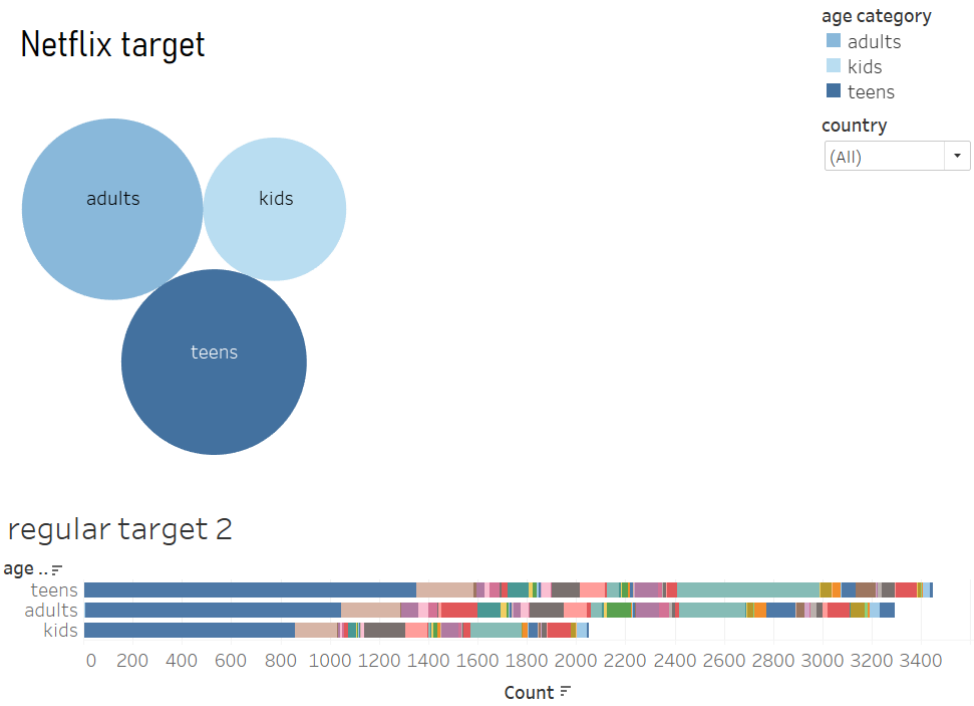
Key hypothesis:

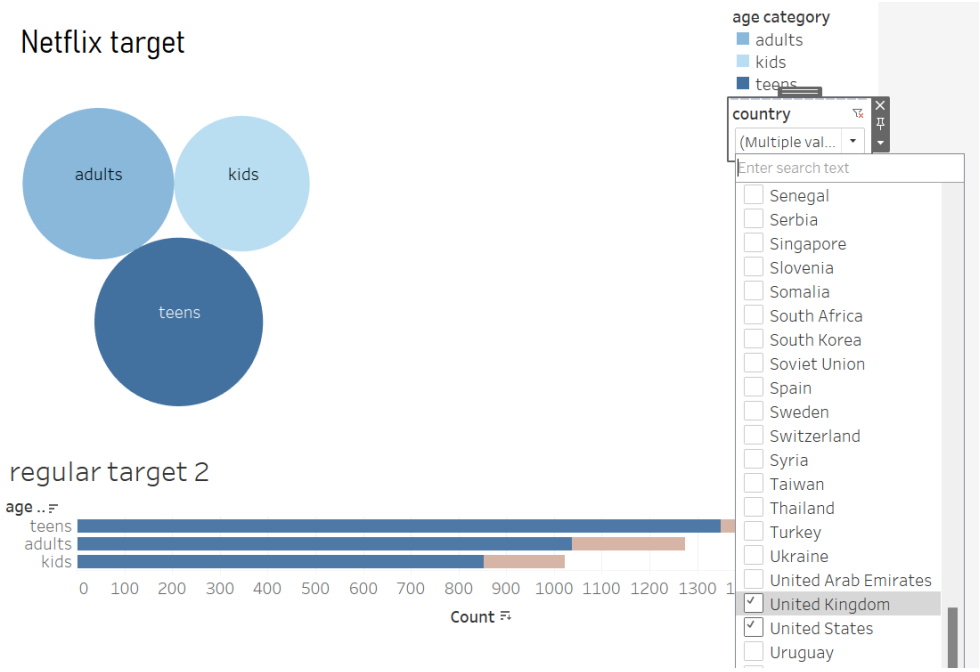
1)

Creating the logical calculation:



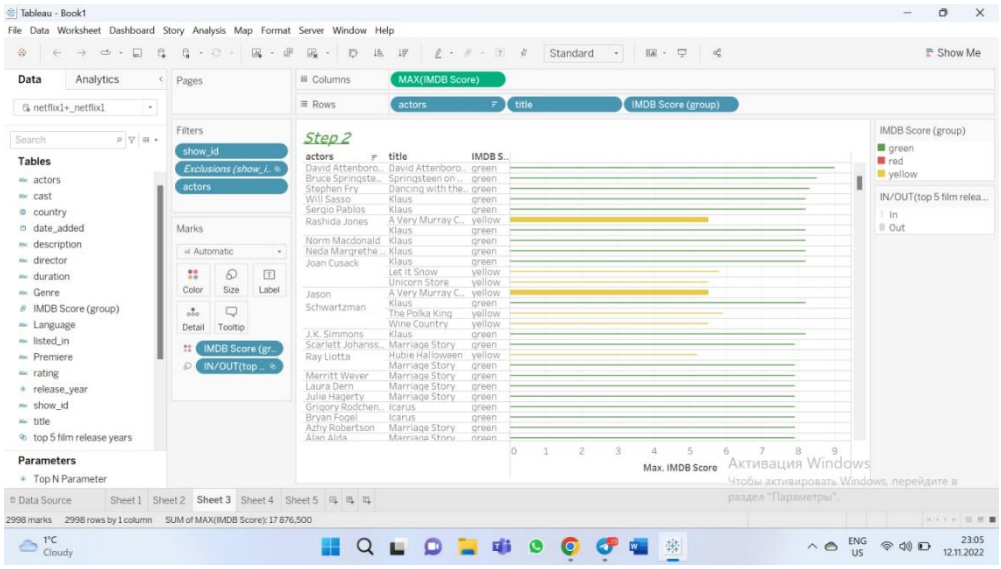
Dashboard: Used “packed bubbles”.



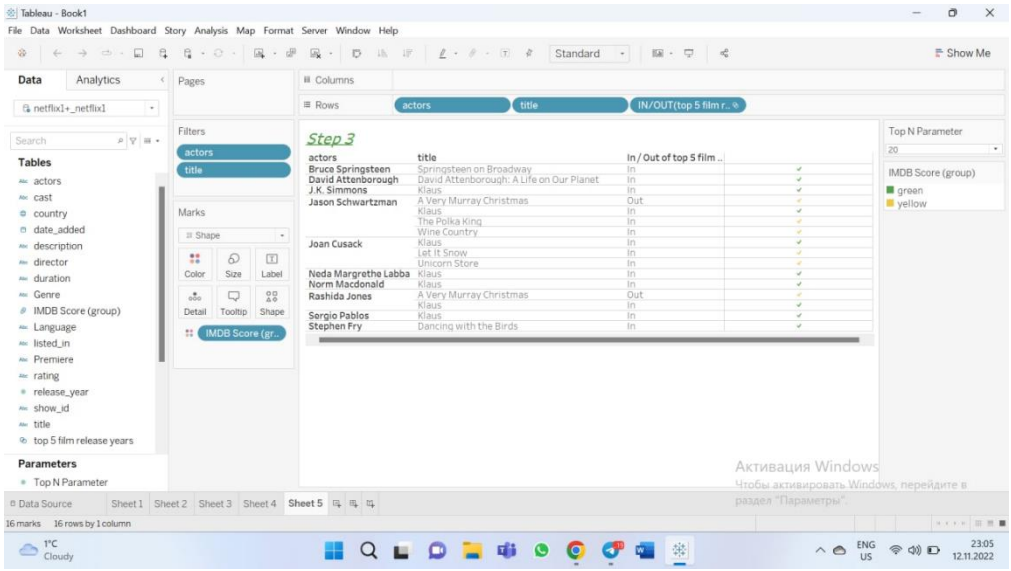


2)

The groups were used to split the IMDB score into the groups, such as green, yellow and red. Green means that the film scored high, yellow average, and red low scores. And sets were used to find the top release years of films. With the use of conventional bar charts, we would not have achieved such clear results as now and it would have taken much more effort and time



Used parameter control to find the top release years of films.

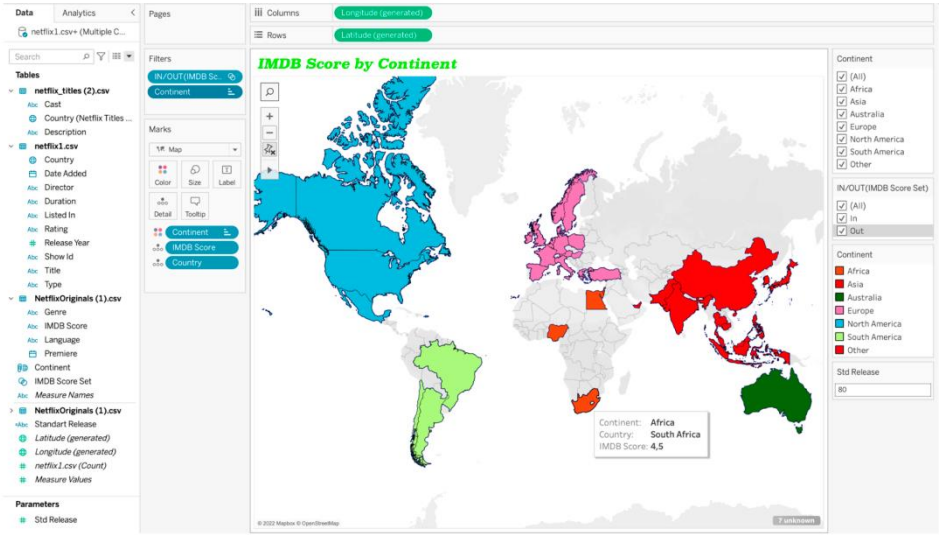


3)

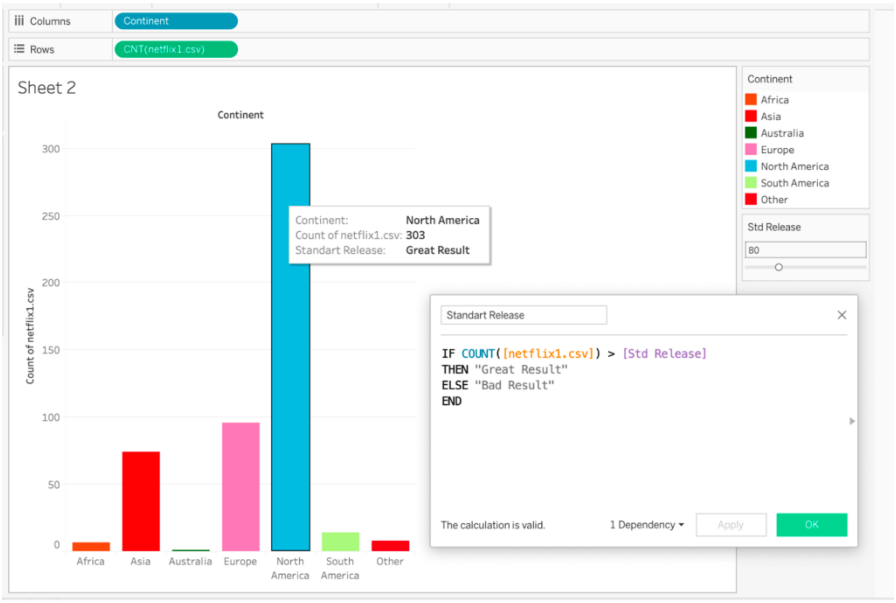
1. Grouping by Continent helps divide countries in such parts to view their total results.

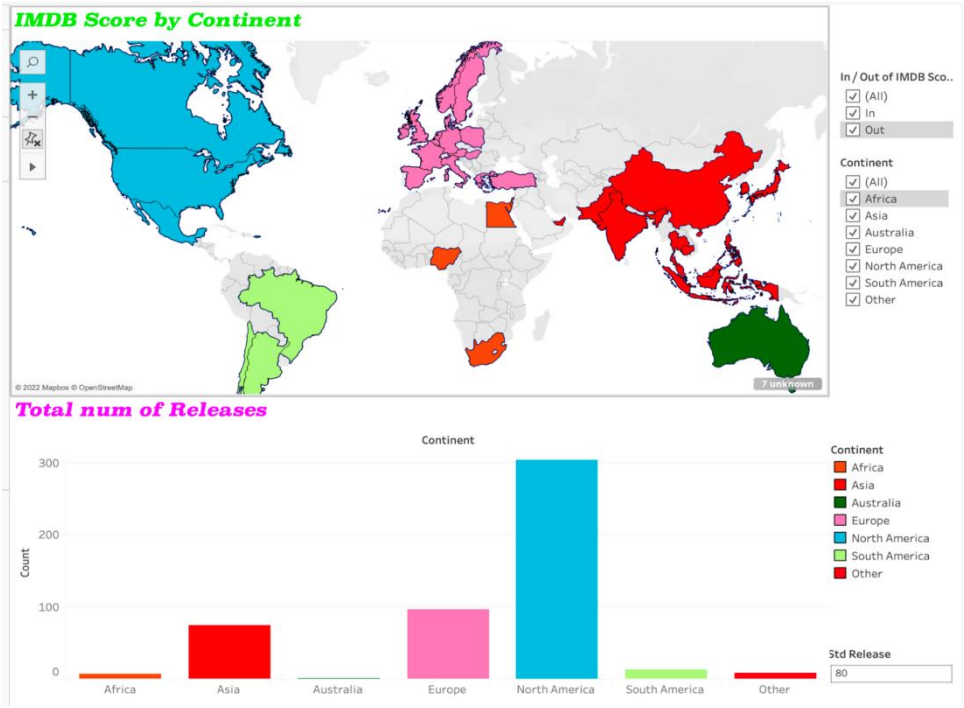
2. Creating Sets by IMDB Score to count Standard high numerical statistics.

3. Add Filters to show compares ratings in Map Graph



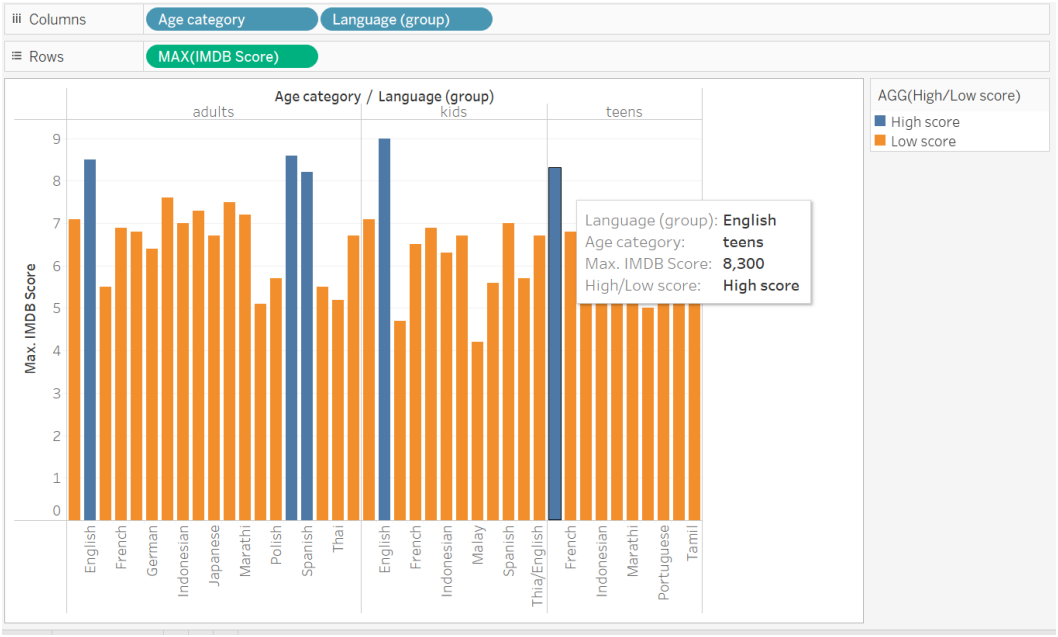
4. Create new Parameter of Standard Release number of Films
5. Compare and identify the results with Calculation Field
6. Africa have not lowest results compared to Australia in the Film Production





4)

Used “group and set”. Group for language and set for identifying high and low IMDB score.

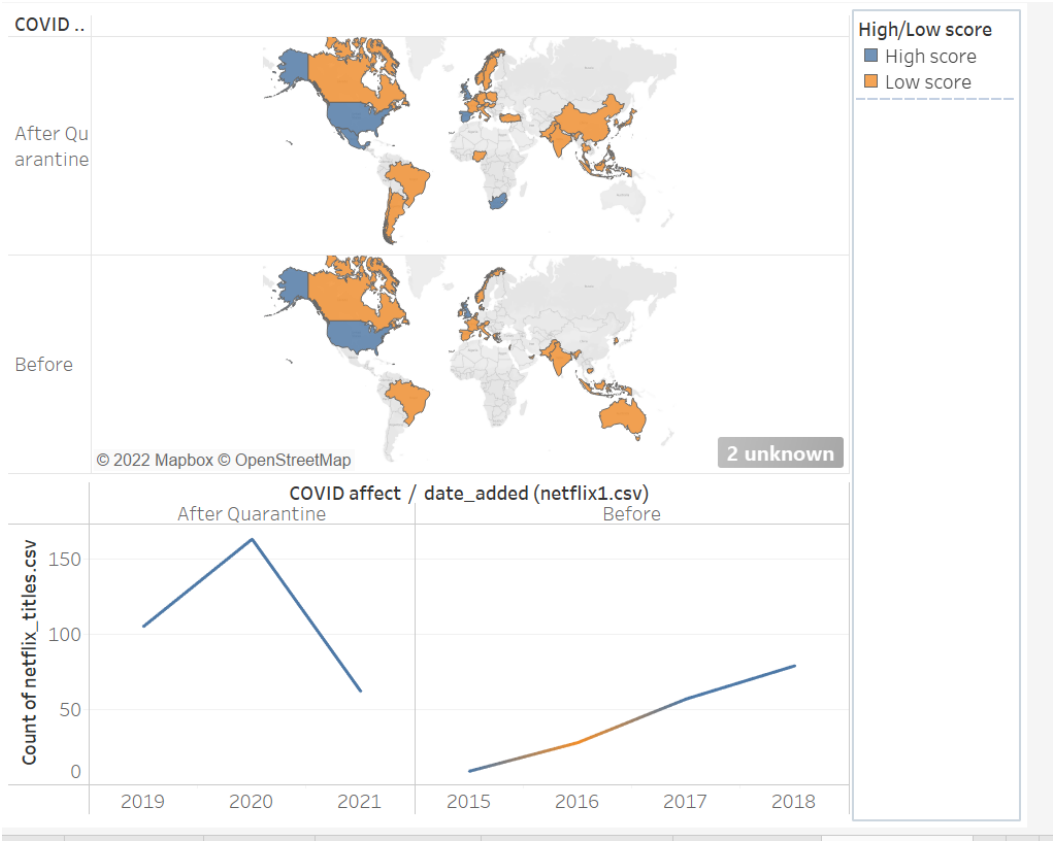


Secondary hypothesis:

1)

Creating the logical calculation:

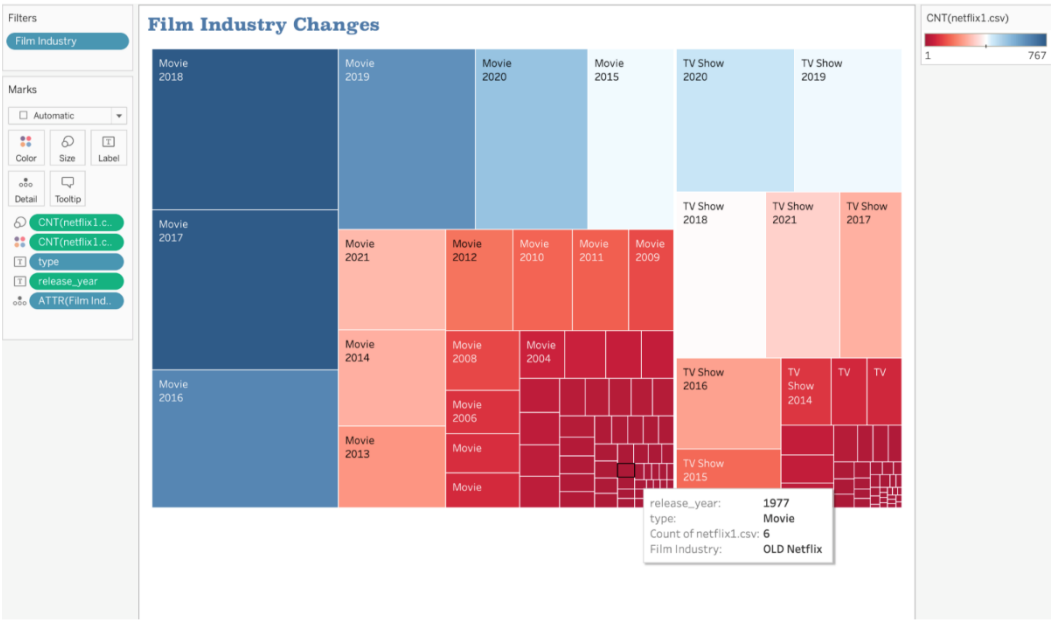
Used “map” and “line”.



After quarantine in many countries, the number of TV series/movies has dropped sharply after continuous increasing.

2)

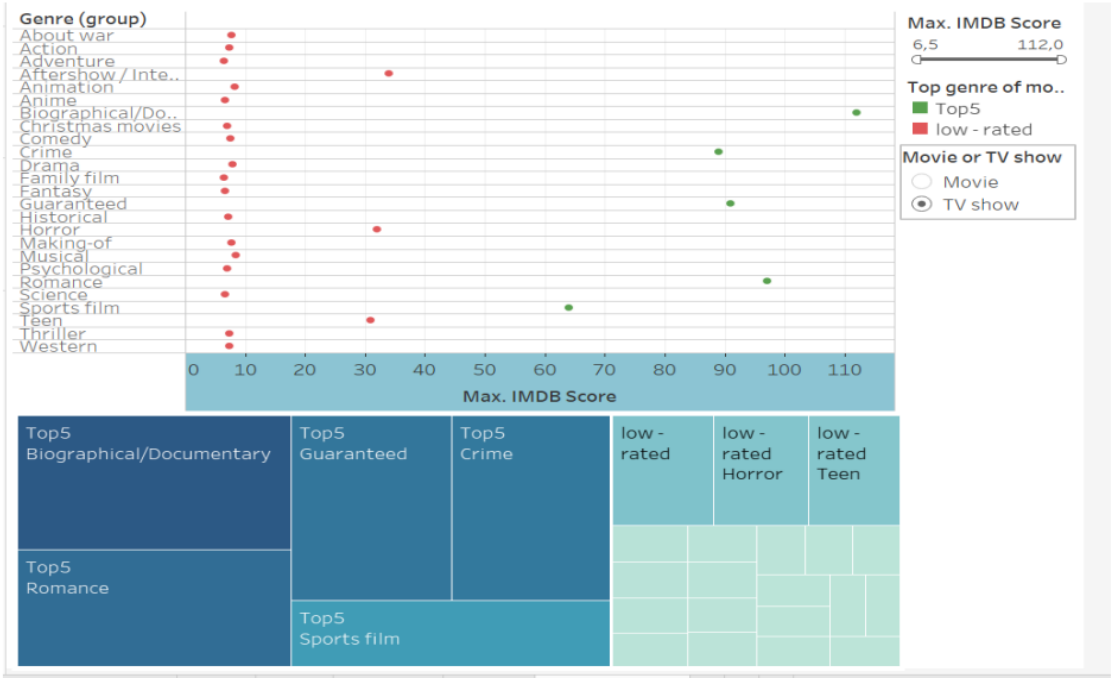
Divided into 2 main types by Netflix Content to show their relevance in different years.

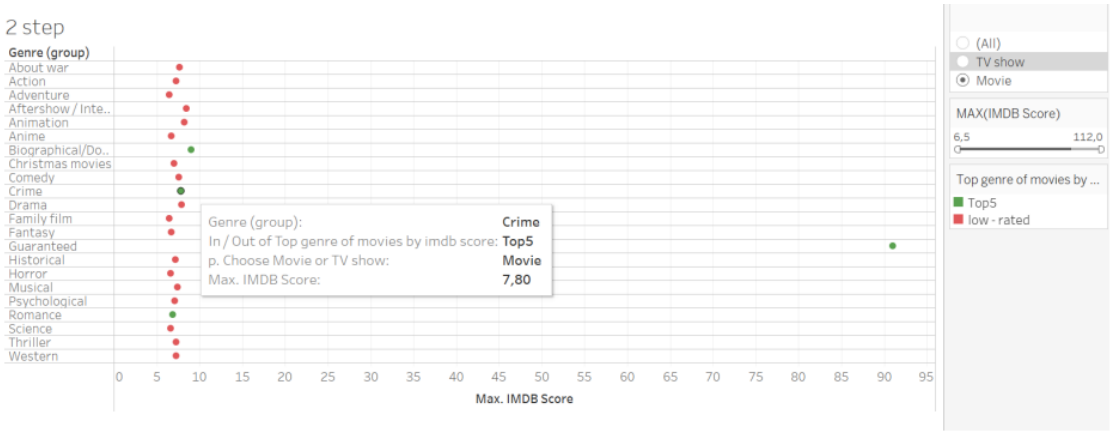


To recognize the urgency of each type by Release Year created the Calculation Filed and use logical function “IF”

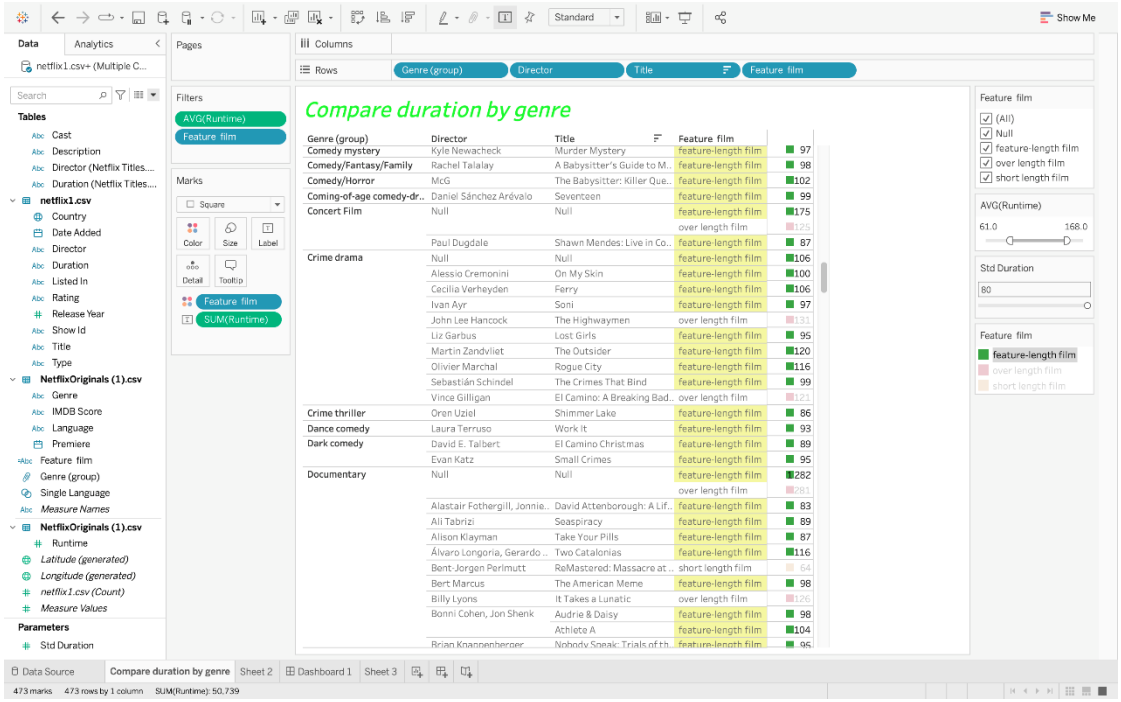
3)

Grouping genre helps me to centralize and choose the right one to reduce the number of many different genres. The set was used in order to collect and mark the top 5 genres by ratings.

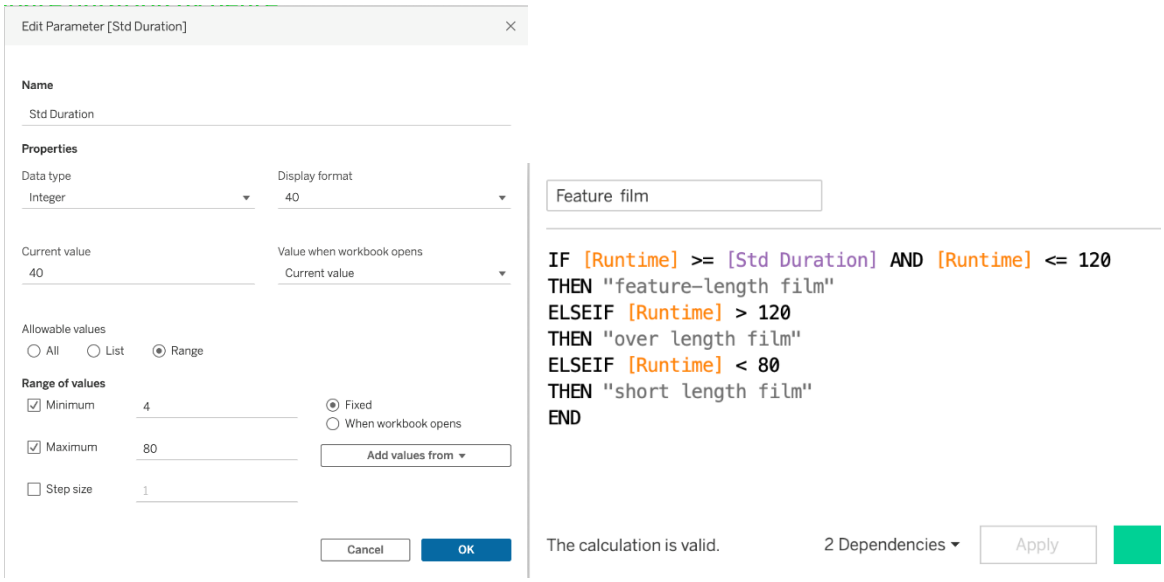

Genre (group)
About war
Action
Adventure
Aftershow / Inte..
Animation
Anime
Biographical/Do..
Christmas movies
Comedy
Crime
Drama
Family film
Fantasy
Guaranteed
Historical
Horror
Making-of
Musical
Psychological
Romance
Science
Sports film
Teen
Thriller
Western
0 10 20 30 40 50 60 70 80 90 100 110
Max. IMDB Score
Max. IMDB Score
6,5 112,0
Top genre of mo..
Top5
low - rated
Movie or TV show
Movie
TV show
Top5 Biographical/Documentary
Top5 Guaranteed
Top5 Crime
low - rated
low - rated Horror
low - rated Teen
Top5 Romance
Top5 Sports film


2 step
Genre (group)
About war
Action
Adventure
Aftershow / Inte..
Animation
Anime
Biographical/Do..
Christmas movies
Comedy
Crime
Drama
Family film
Fantasy
Guaranteed
Historical
Horror
Musical
Psychological
Romance
Science
Thriller
Western
Genre (group): Crime
In / Out of Top genre of movies by imdb score: Top5
p. Choose Movie or TV show: Movie
Max. IMDB Score: 7,80
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
Max. IMDB Score
(All)
TV show
Movie
MAX(IMDB Score)
6,5 112,0
Top genre of movies by ...
Top5
low - rated

4)

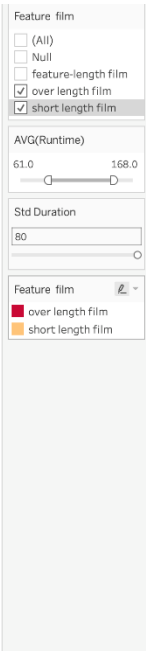


Created a Parameter for Standard Duration, that starts from minimum 40 min. The Screen Actors Guild defines a feature as a minimum of 80 minutes whereas The Academy defines a feature as a minimum of 40 minutes.

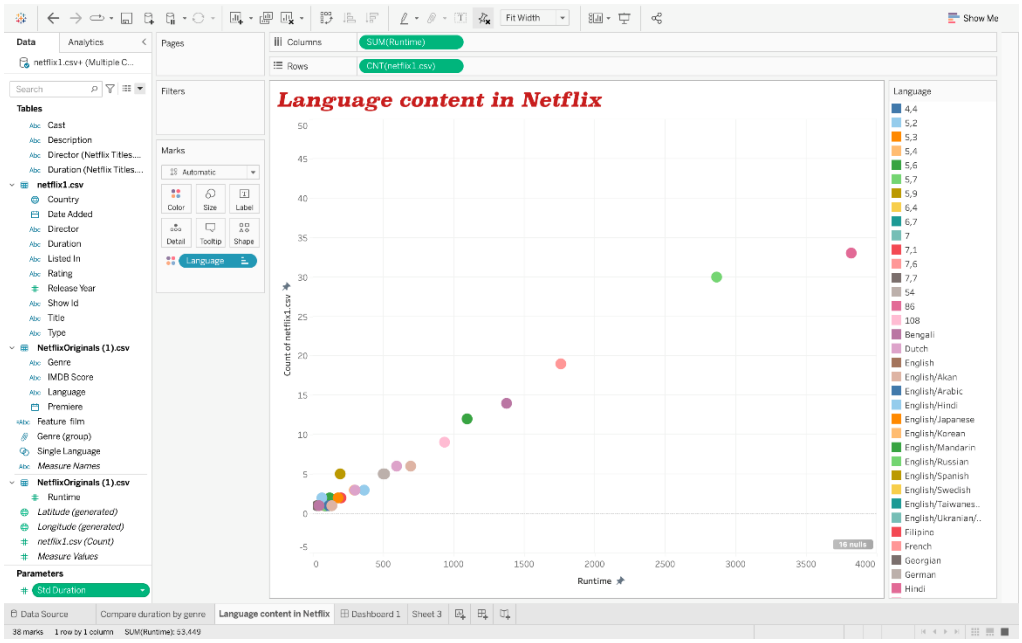


Compare duration by genre

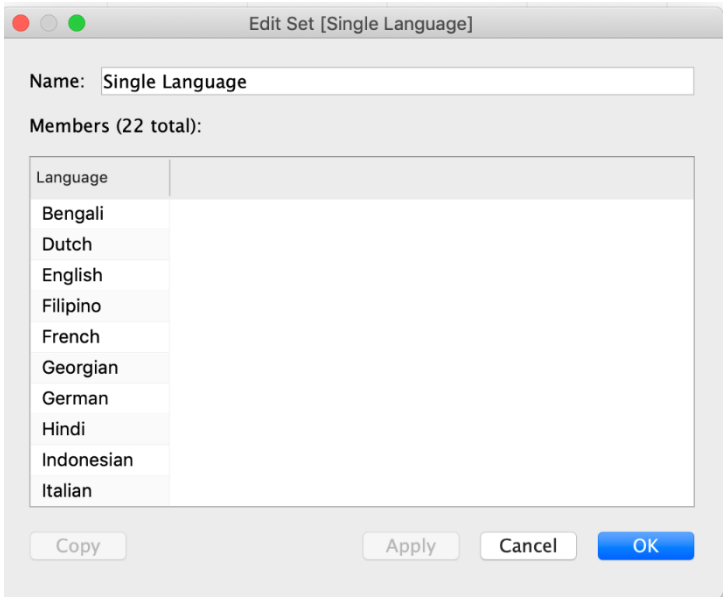
Genre (group)	Director	Title	Feature film	
Action, Action comedy, Action thriller and 3 more	Tarun Mansukhani	Drive	over length film	147
	Timo Tjahjanto	The Night Comes for Us	over length film	121
Animation, Animation / Comedy, Animation / Science Fiction and 4 more	Null	Null	over length film	248
	Al Campbell, Alice Mathias	Death to 2020	short length film	70
	Blair Simmons	Octonauts & the Caves of ..	short length film	72
	Gabriela Tagliavini	Despite Everything	short length film	78
	Kaashvie Nair	Sardar Ka Grandson	over length film	139
	Michael Paul Stephenson	Girlfriend's Day	short length film	70
	Not Given	Invader Zim: Enter the Flo..	short length film	71
	Steven Brill	Sandy Wexler	over length film	131
Anime/Science fiction	Takeru Nakajima, Yoshiyu..	Altered Carbon: Resleeved	short length film	74
Anthology/Dark comedy	Anurag Basu	Ludo	over length film	149
Biopic	Null	Null	over length film	132
Concert Film	Null	Null	over length film	125
Crime drama	John Lee Hancock	The Highwaymen	over length film	131
	Vince Gilligan	El Camino: A Breaking Bad..	over length film	121
Documentary	Null	Null	over length film	281
	Bent-Jorgen Perlmutt	ReMastered: Massacre at ..	short length film	64
	Billy Lyons	It Takes a Lunatic	over length film	126
	Daniel Vernon	Nail Bomber: Manhunt	short length film	72
	David Sington, Heather W..	Mercury 13	short length film	79
	Jacob Kornbluth	Saving Capitalism	short length film	73
	Joe Piscatella	Joshua: Teenager vs. Sup..	short length film	78
	Justin Krook	I'll Sleep When I'm Dead	short length film	79
	Kelly Duane de la Vega	ReMastered: The Two Killi..	short length film	64
	Kevin MacDonald	Sky Ladder: The Art of Cai ..	short length film	79
	Madeleine Gavin	City of Joy	short length film	74
	Petra Costa	The Edge of Democracy	over length film	121
	Rashida Jones, Alan Hicks	Quincy	over length film	124
	Stuart Sender	ReMastered: The Miami S..	short length film	70
	Tom Donahue	Los Tigres del Norte at Fol..	short length film	64
Drama, Drama-Comedy,	Null	Null	over length film	279

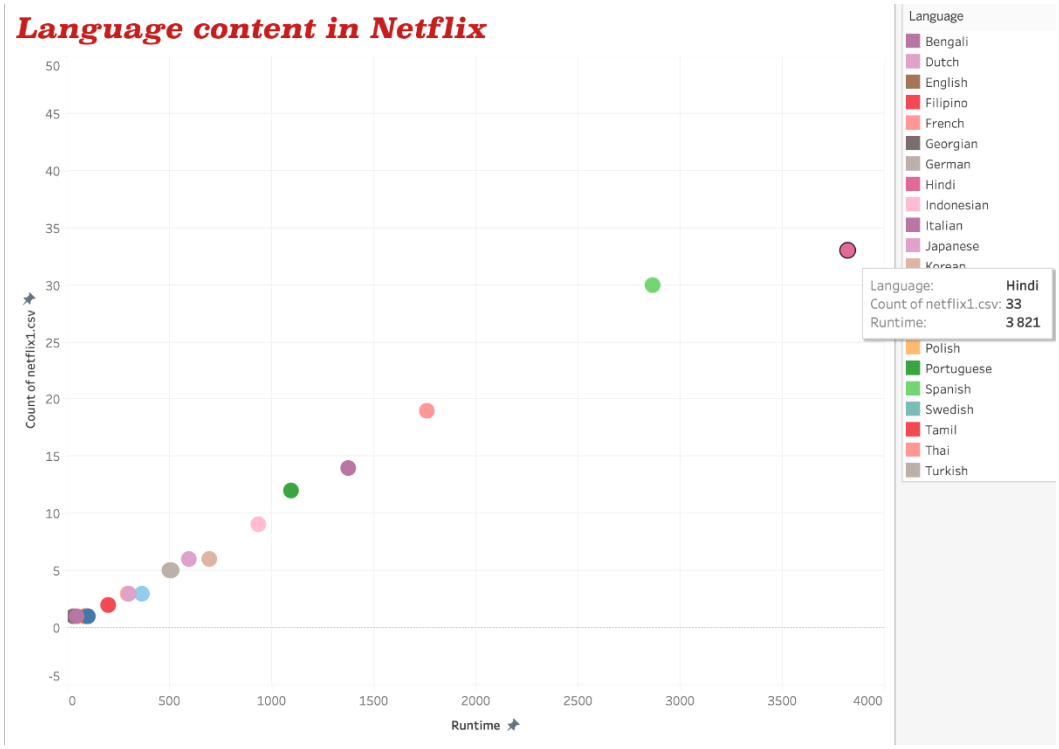


5)



In the full data there are movies that go in two or more mixed languages, and our task is to analyse and compare it is monolingual content. For that firstly we created the set for languages that we need.





6)

Identifying key words from description for using it for better analysing. Used logical functions.

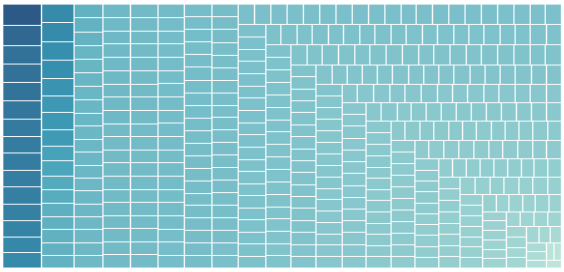
Key words

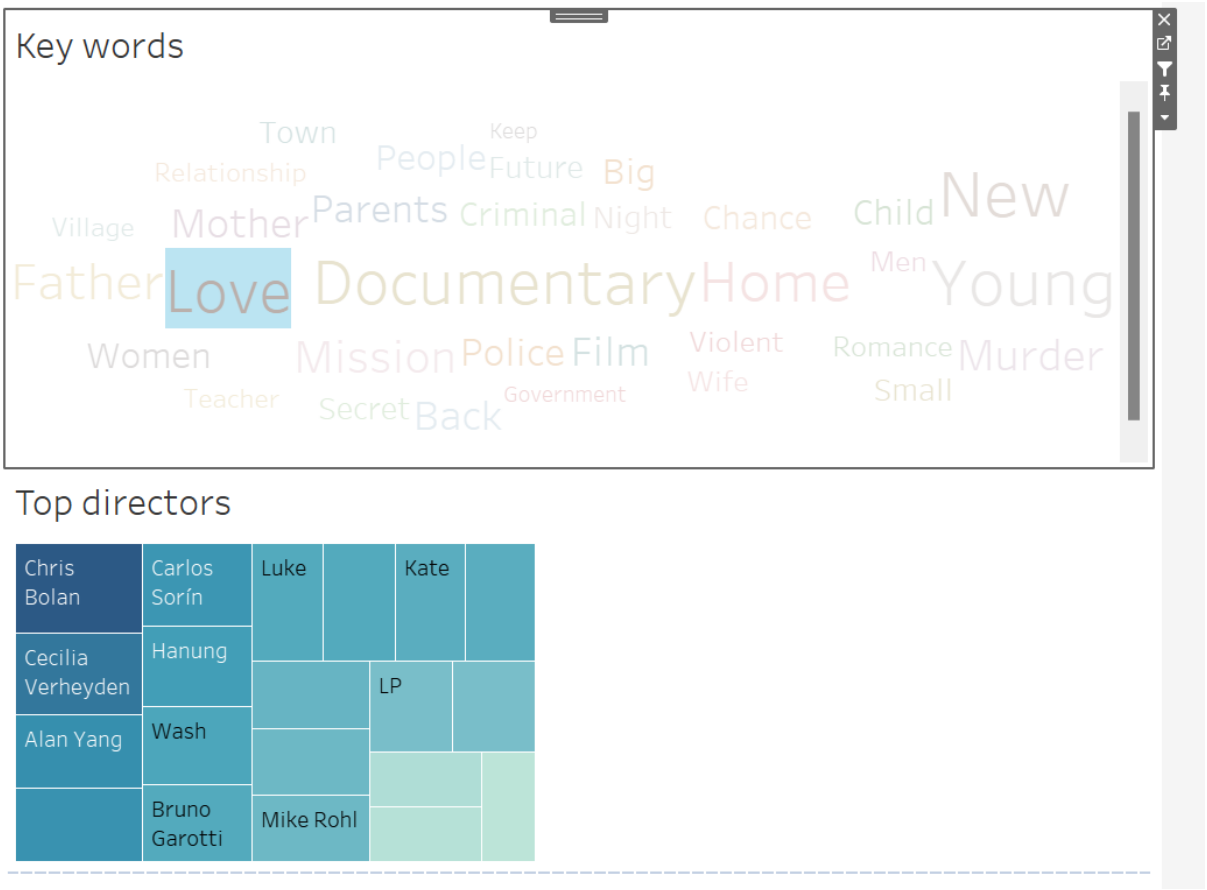
```
IF CONTAINS([description] , " love " ) = TRUE then "Love"
ELSEIF CONTAINS([description] , " criminal " ) = TRUE then "Criminal"
ELSEIF CONTAINS([description] , " home " ) = TRUE then "Home"
ELSEIF CONTAINS([description] , " parents " ) = TRUE then "Parents"
ELSEIF CONTAINS([description] , " murder " ) = TRUE then "Murder"
ELSEIF CONTAINS([description] , " child " ) = TRUE then "Child"
ELSEIF CONTAINS([description] , " future " ) = TRUE then "Future"
ELSEIF CONTAINS([description] , " government " ) = TRUE then "Government"
ELSEIF CONTAINS([description] , " people " ) = TRUE then "People"
ELSEIF CONTAINS([description] , " village " ) = TRUE then "Village"
ELSEIF CONTAINS([description] , " town " ) = TRUE then "Town"
ELSEIF CONTAINS([description] , " wife " ) = TRUE then "Wife"
ELSEIF CONTAINS([description] , " secret " ) = TRUE then "Secret"
ELSEIF CONTAINS([description] , " men " ) = TRUE then "Men"
ELSEIF CONTAINS([description] , " women " ) = TRUE then "Women"
ELSEIF CONTAINS([description] , " mission " ) = TRUE then "Mission"
ELSEIF CONTAINS([description] , " new " ) = TRUE then "New"
ELSEIF CONTAINS([description] , " young " ) = TRUE then "Young"
ELSEIF CONTAINS([description] , " night " ) = TRUE then "Night"
```

The calculation is valid. 1 Dependency Apply OK

Key words

Top directors





Step 4.

Solutions:

Key:

1) Based on the ratings of films and using the Internet, it turned out to determine the categories of each rating. Now looking at it, you can accurately confirm the fact that teenagers are the main target of Netflix rather than other categories. It is also worth noting that most of the material for teenagers is produced by USA and UK. To expand the base of other categories, other countries should also switch to this target.

2) The purpose of this hypothesis is to find out whether films with the participation of a popular actor released in the best years are gaining high IMDB ratings or not. As we can see, the hypothesis is correct, films with the participation of top actors familiar to us were released in the top years, not counting some moments. From this point on, we advise that in order to score a high IMDB score, it is worth hiring these actors, since the cast is one of the main sources of the popularity of the film / series.

3) The results were showed by dividing in Groups and Sets. African continent was in orange colour in Map, and if we choose countries IN Sets with high result, from African c. Map show us only Nigeria which has rating above 5.

Also, by Total number of Releases in different Continents it was compared the Results of each Group. Creating Parameter helps, identify Standard indicator of releases in Range (it may always change). Here Africa scores higher than Australia. We compared it by Logical function (IF).

4) To promote the film, producers should release films in English, since English is an international language, there is more demand for it than in other languages.

Secondary:

1) As statistics show, after covid, the number of releases dropped sharply. This may be due to Quarantine, as quarantine measures required residents not to visit crowded places. And such social isolation can destroy a person's consciousness. From this, we can assume that people are no longer interested in watching movies / TV series at home than spending time on the street.

2) Some results find out that the Movies started to lose relevance and began to less often come out with the advent of TV shows. But statistics provides the opposite, every year the release of Movies and TV shows increases with the same coefficient. However, the Movies take up and show better results.

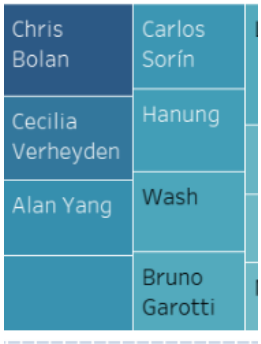
3) With the help of groups and sets, I collected them into certain categories in order to specifically define their genre and find out which genres are in the top 5. As we can see here, the hypothesis that I wrote based on some data was not correct, and the top 5 genre of TV series are documentaries, romance, guaranteed, sports and crime series.

But if we talk about films, then documentaries, crime and romantic films are one of the low-rated.

4) Making a list of all the films we determined the average length of the films. Each of them occupies an individually important duration. And many documentaries have a length of less than the Academy standards. Give them values by Parameter and using the Calculation Field divided it into 3 main Run Time types.

5) According to Release by every year, in Netflix content English is the most common language. But there is question about second famous language in Movie content. Our hypothesis was about prevalence of Spanish films. But Alternatives showed that they are going behind Hindi that demand high results in Movie releases.

6) The supposed hypothesis was incorrect and too stereotypical. As we can see among them only 1 is a woman and the rest are men.



- Our hypotheses are based on various assumptions that are related to Netflix. With their help, we ask specific questions that make us think about the structure and strategy of the film industry in various countries.
- Thanks to our visualizations, we clearly get answers to our questions arising from hypotheses. Visualization makes everything clear for clients who ask for help with questions about how to

improve any area. For example, how to develop the film industry outside of America and the UK so that ratings are as high as in these two countries.

- The main obstacle was that there is no information about 2022 to analyse and compare current results. The Netflix platform's content is exposed and changing every year in a new direction, and it is difficult to raise strong questions and find actual workable solutions that will work according to the statistics of previous years.

Self-reflection - Alua Taszhan:

• What have you learned in each step of the course project?

In this course project, I learned how to correctly put forward hypotheses and answer them. If we analyze it in detail, then at step 1 I identified the business sector, and for this I additionally studied the material about it in order to specifically know what's what. That is, what are usually the business sectors and which of them should be used for analysis. In step 2, I analyzed the previous hypotheses and decided which of them should be left and which should not, and which of them are key and which are secondary. In step 3, I used my knowledge gained throughout this course to create a suitable visualization and apply various methods in addition to them. For example, I used logical functions to divide into categories, parameter control to show "ratings by genre" by type, Groups to centralize and choose the right one to reduce the number of many different genres. The set was used in order to collect and mark the top 5 genres by ratings etc. And finally, there I drew conclusions about how exactly these cases will affect this business sector and the solutions that are provided with the help of my visualizations.

• Where would you like to apply this knowledge?

I would like to apply this knowledge in the future at work. I would like to work as a business analyst, and for this, the knowledge from this course will be very useful.

• Is there anything you would like to learn more about the course?

In this course, I would like to work more on real cases, although it is difficult to arrange. The fact is that we worked with ready-made data from the Internet, where all the characteristics and description of this data are provided. But in reality, this work is much more difficult, due to the fact that most often not such clean and prepared data for analysis are given at work. To do this, we ourselves would have to do additional work for this. Although it sounds a bit burdensome, thanks to this I think we could be 1 step closer to the real case.