Tastes on Netflix

What kinds of content gain more popularity?

By Ivan Kompaniets

Index

1. Context and audience
2. Metadata Description
3. Exploratory data analysis
4. Insights and recommendations
5. Model selection
6. Dataset preprocessing. Variable selection
7. Hyperparameter selection
8. Model performance
9. Final model
10. Analysis of clusters
11. Conclusion
12. Content suggestions
13. Future lines

Context and audience

Context

Netflix is an online streaming platform that offers series, movies, and other types of content to its users in exchange for a monthly payment.

To have a stable income, the platform needs to retain its existing customers by offering them interesting content and also attract new users.

Therefore, it is essential to understand which content grabs the users' attention the most and what characteristics determine its success.

Audience

This work could be valuable to Netflix's management to improve its content offerings as well as to other streaming platforms looking to increase their market share.

Limitations

The main limitation of this work is that it is based on 2021 data. In the past two years, users' preferences may have changed, so the study would need to be updated with new data.

Questions of interest

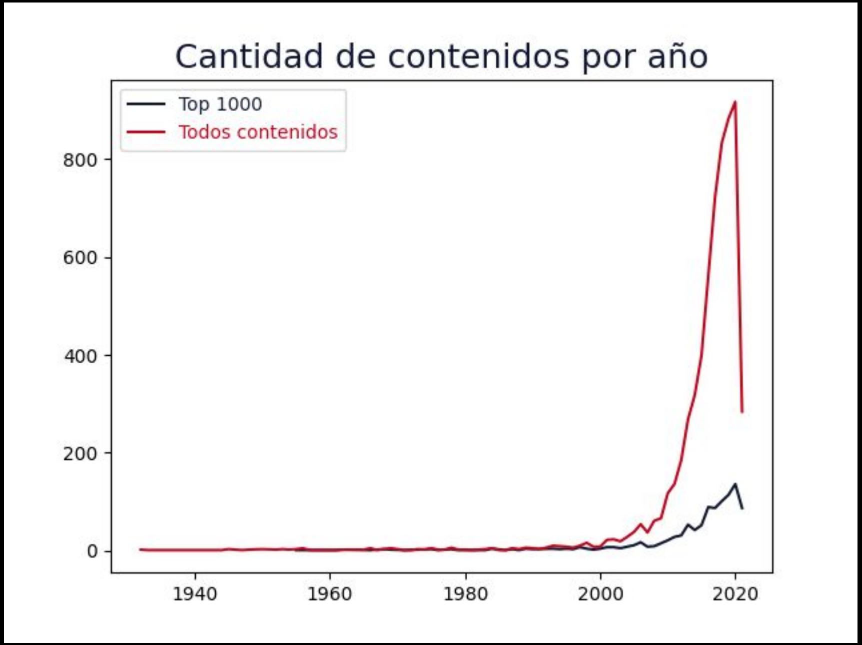
Our main hypothesis is that content with certain characteristics attracts more attention from users than other content. To confirm or refute the hypothesis, we will answer the following questions:

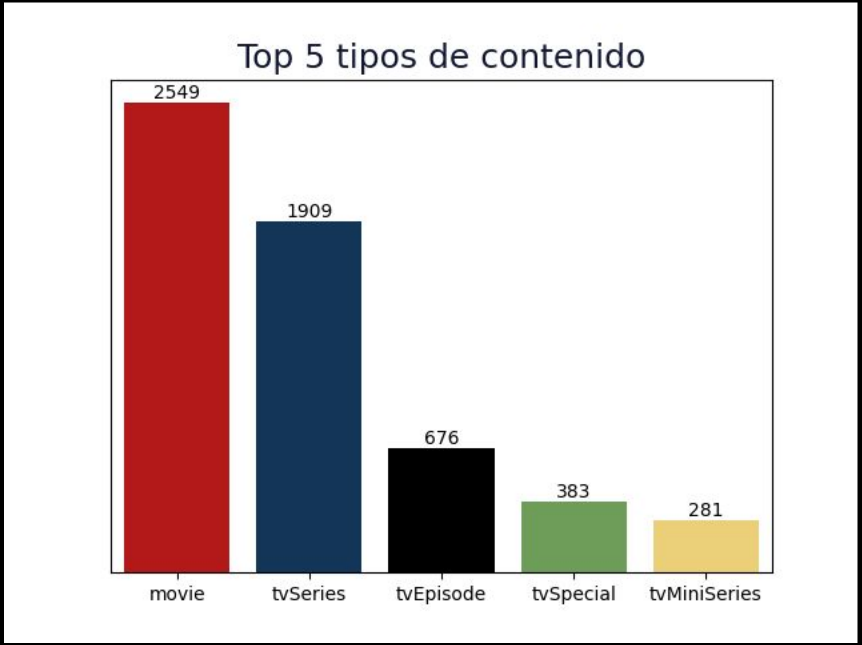
- Is there a relationship between the popularity and quality of content?
- What type of content is more popular?
- What type of content is of better quality?
- What are the most popular genres?
- What are the items of content with the highest rating?

Metadata

Number of content items:
6167

Production country and language
with the most content:





Metadata

As we can see on the graphs, the two most prominent types of content on the platform are series and movies. It is also worth noting the miniseries type, which is similar to series but with fewer episodes. We can combine the two types when processing the data.

Regarding the line graph, it shows that the amount of content increases over the years before experiencing a decline in 2021. This could be explained by the fact that the dataset was compiled that year and also by the impact of the COVID pandemic, which led to a decrease in movie production.

It is also noticeable that the number of content items in the top 1000 increases along with the overall quantity, but not as steeply. Therefore, we can assume that many of the new content items did not manage to become very popular.

The next two slides will present the variables contained in the dataset.

Variables

Variable name	Description	Data type
imdb_id	IMDB Id	string
title	Content title	string
popular_rank	The smaller the rank, the higher the popularity	int
certificate	Suitability for minors (cannot be used as there are many null values)	string
startYear	Year of release	date
endYear	End year (can be different from the year of release for series)	date
episodes	Number of episodes (1 for movies, other values for series)	int
runtime	Runtime (total for a movie or of one episode for series)	int
type	If it’s a movie, TV show or another type	string

Variables

Nombre de variable	Descripción	Tipo de data
orign_country	Production country	string
language	Language	string
plot	Plotline	string
summary	Summary	string
rating	IMDB score	float64
numVotes	Number of votes on IMDB	float64
genres	Genres (one content can belong to several genres)	string
isAdult	If adult content (column deleted, as all rows had the same values)	int64
cast	List of actors	string
image_url	Link to the picture (will not be used)	string

Exploratory Data Analysis

In the following slides, I will present the results of my exploratory data analysis.

Is there a relationship between popularity and content quality?

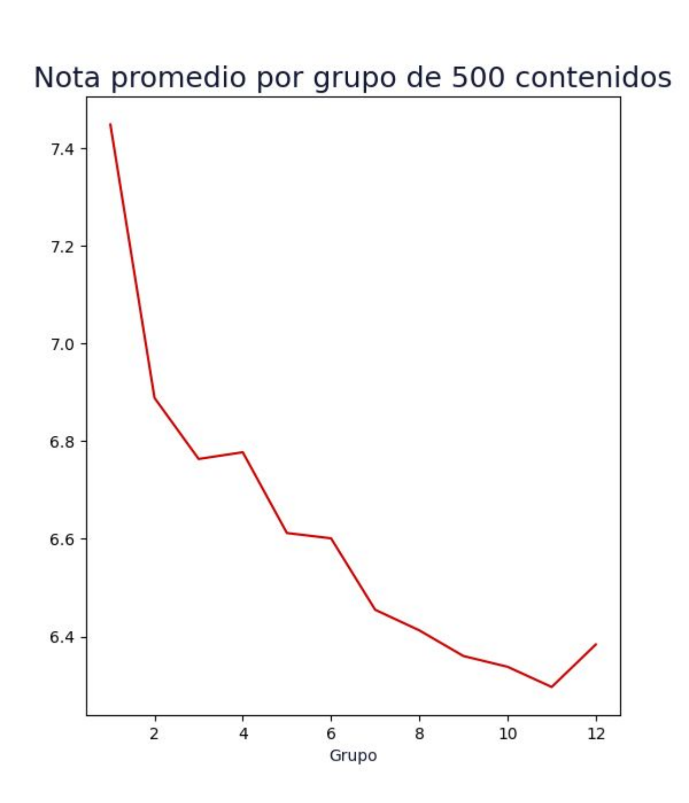
The quality of a movie is often well reflected by the rating it receives on the IMDB page.

To see if a movie's popularity depends on its average rating, we divided the data into 12 clusters and calculated the average rating per group.

The clusters with the highest popularity are on the left and the less popular clusters are on the right.

We can see that there is a relationship between popularity and quality. The average rating decreases as popularity decreases.

The only exception is group 12, which has a slightly higher rating than the previous group.



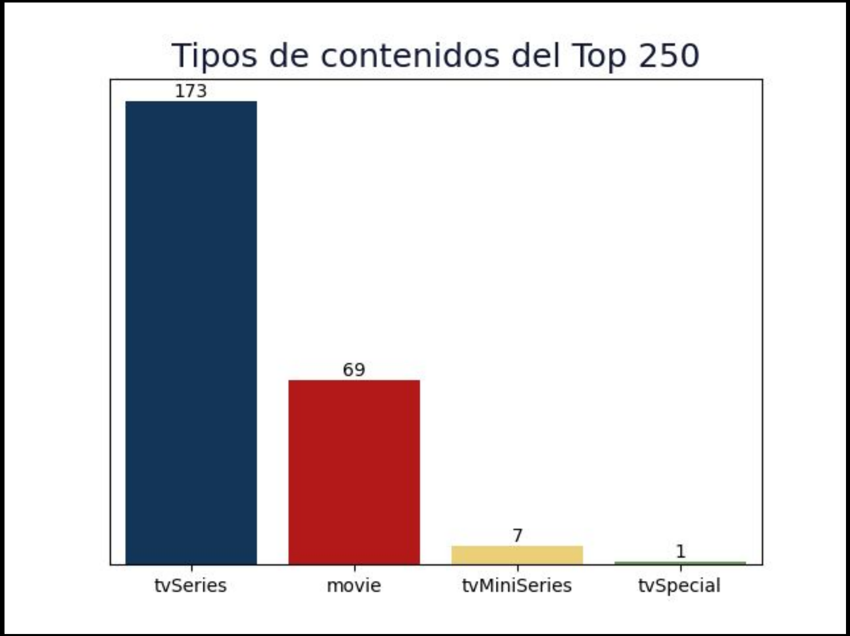
What types of content are the most popular?

To identify the most popular types of content, we focused on the top 250 popular items.

We can observe that the most popular type of content on the graph is series.

This is an interesting result because movies are the most common content type in the dataset overall.

It means that the greater popularity of series is not a mere coincidence but there is a trend behind it.



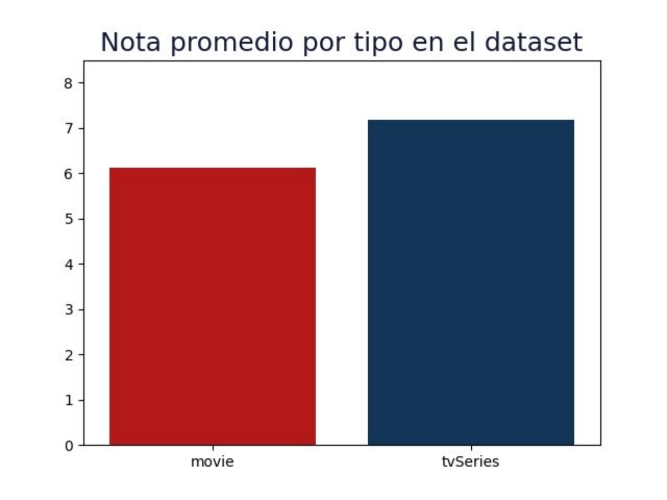
What type of content is of better quality?

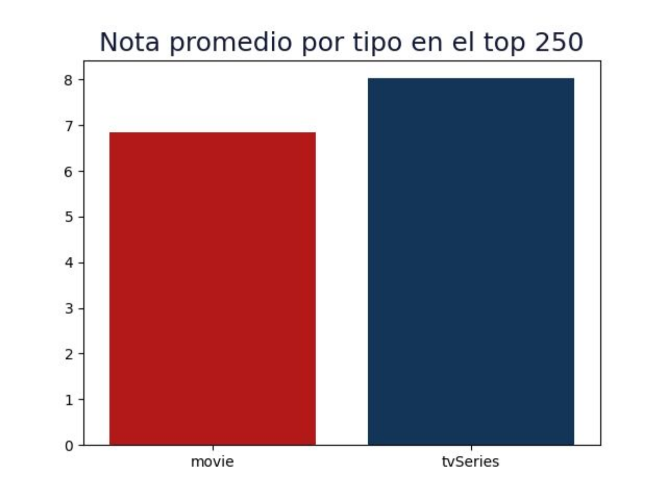
As we saw that movies are less present among the top 250 popular content items, we wondered if there is also a qualitative difference between the two types of content.

As we can observe, the answer is yes!

We can see on the two charts that the quality of series is superior in the entire dataset and in the top 250 items.

Furthermore, if we look at the values, the average ratings in the top 250 are higher than in the entire dataset, demonstrating once again the relationship between quality and popularity.



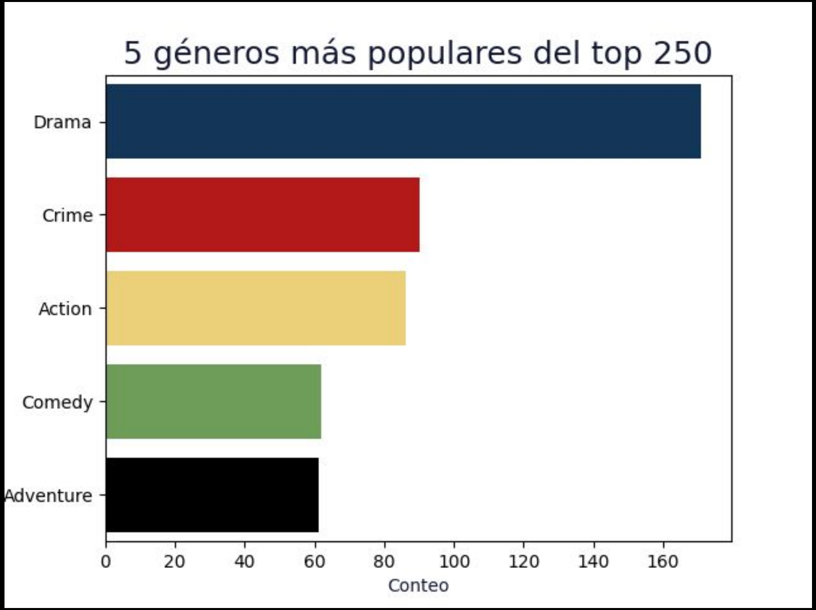


What are the most popular genres?

Once again we looked at the most popular genres in the top 250 and produced the following chart.

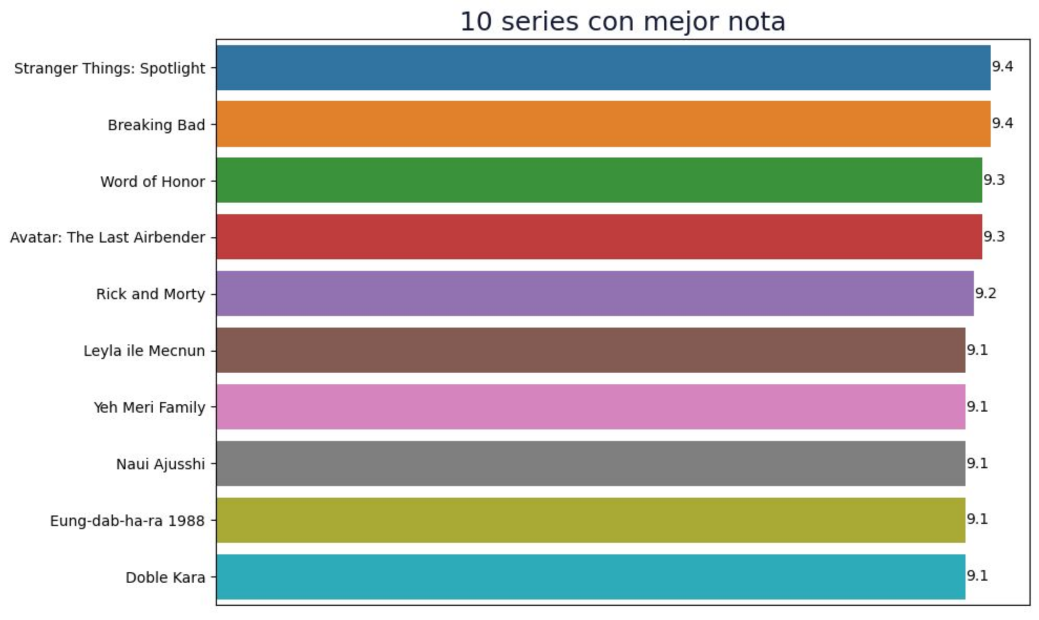
An interesting detail is that a content can belong to several genres at the same time and the genres we see in the chart can be compatible. For instance, we can group the genres “drama” and “crime” as well as “comedy” and “adventure” as these genres often go hand in hand.

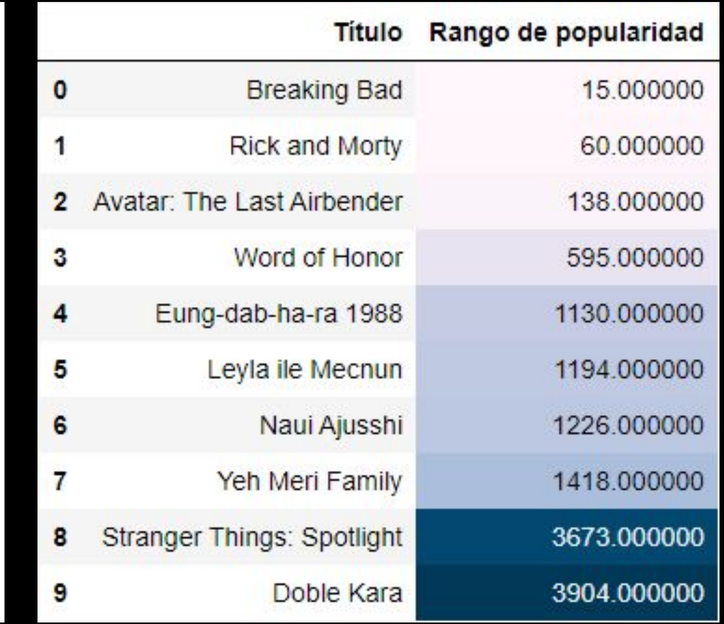
These are the two main genre categories of content most searched on Netflix.



What are the series with the best ratings?

On the chart below, we can see the series with the highest ratings, but as we see in the table, they do not always have the highest popularity index.
Therefore, quality is not the only criteria, but other characteristics such as genre are also important.





	Título	Rango de popularidad
0	Breaking Bad	15.000000
1	Rick and Morty	60.000000
2	Avatar: The Last Airbender	138.000000
3	Word of Honor	595.000000
4	Eung-dab-ha-ra 1988	1130.000000
5	Leyla ile Mecnun	1194.000000
6	Naui Ajusshi	1226.000000
7	Yeh Meri Family	1418.000000
8	Stranger Things: Spotlight	3673.000000
9	Doble Kara	3904.000000

Insights and recommendations

Insights

- We have observed that the popularity of Netflix content depends on several characteristics.
- One important feature is the IMDB rating, which reflects the quality of the content. However, it is not the only factor.
- Another important factor is the genre of the content. For example, Netflix users seem to be particularly interested in crime dramas and comedy adventure movies.

Recommendations

- To improve its content offering, Netflix should continue offering new content (especially series) in users' preferred genres.
- Another strategy could be to improve the quality of the movies available on the platform to attract other clusters of users.
- Finally, it could be useful to offer series and movies in less popular genres to attract users with more diverse interests.

Model selection

The goal of this work is to develop a clustering model that highlights various types of content within the dataset and provides content suggestions to users.

For the clustering task, we will create KMeans, Agglomerative Clustering, Birch, HDBSCAN and OPTICS models and compare their performance based on clustering metrics. We will then select the best suited algorithm to carry out a deeper analysis of the dataset and identify relationships between entities.

Next, we will try to find the most similar contents in order to make suggestions based on user preferences. To do so we will build a NearestNeighbors model that allows searching for the records closest to a chosen one.

Finally, we will write a function to extract the id, title, year, and genre of each suggestion from the initial dataset.

Data set preprocessing.
Variable selection

First, we will select the variables that will be used for building the model.

We will include the following variables:

- startYear: release year. We will modify the variable to represent periods rather than exact years
- type: type of content (series, movie, or other)
- origin_country: the variables origin_country and language contain similar information. However, the variable "country" is more precise as content in Spanish or English can be produced by many different countries.
- rating: the score a content received on IMDB. It's important because it reflects its quality. On the other hand, the variable "numVotes" which indicates the number of ratings received will not be included because it does not tell us anything about the intrinsic value of content.
- genres: an essential variable because users often choose movies based on genre.
- imdb_id: reflects the popularity of the content.

Variable Selection

The following variables will be excluded from the analysis:

- title: the title of the content cannot be used in the analysis.
- episodes, runtime: both describe the duration of content. We will use them to create a total duration variable.
- plot, summary: textual variables. They can be used but require processing.
- cast: a difficult variable to process that contains lists of actors. As we have seen in the exploratory data analysis, specific actors do not necessarily determine the success of a series or movie. Furthermore, using the variable would significantly increase the data dimensionality.
- image_url: is a link to the content's cover. We will not process visual data.

Preprocessing of variables

The genres variable is essential for our analysis as it determines content choices by users. However, in our dataset, each record can correspond to multiple genres. To get around this problem, we used the One Hot Encoding algorithm to encode each genre. If a content belongs to a genre, it receives a 1 in the corresponding column, and if not, it receives 0.

As expected, content can belong to one, two, or three genres.

We also encoded the textual variables type and origin_country. Unlike genres, each record has a unique value in these columns, so we used the Label Encoding algorithm that a number to each textual value.

Creation of new variables

Based on the existing variables, we added two new variables total_rt (total duration) and period.

The total duration is a numerical variable obtained as the sum of the variables episode and runtime. As a result, series consisting of multiple episodes have a longer total duration than movies and other single-episode content items.

As for period, it is a categorical variable obtained from the variable startYear. Viewers often do not decide whether they are interested in a content based on its exact production year but rather on whether it is new or old. Therefore, we highlighted 4 clusters of contents which are:

Group 0: old content from before 2000

Group 1: content produced between 2000 and 2010

Group 2: content produced between 2010 and 2020

Group 3: newest content released after 2020

After creating the new variables, we applied the Standard Scaler algorithm to rescale the numerical variables rating and total_rt.

Model selection

To perform data clustering, we selected the following algorithms:

- Agglomerative Clustering,
- Birch,
- HDBSCAN,
- OPTICS,
- KMeans,

The main difficulty of our dataset is that it contains data on very diverse types of content, so it was expected to obtain many clusters through clustering.

Another risk is that when using an algorithm with a fixed number of clusters, such as KMeans, most of the data may end up in a single group because of not being classified. This would significantly decrease the model's utility.

To evaluate the models, we used the Silhouette Score, Davies-Bouldin score, and Calinski-Harabasz score metrics. We will also calculate the average cluster size and the standard deviation of the cluster sizes.

Hyperparameter Selection

Agglomerative Clustering

The first algorithm we used was Agglomerative Clustering because it allows grouping data by distance without determining a number of clusters when instantiating the model.

We chose the Ward clustering method and set the distance threshold to 200. A high distance value would reduce the number of clusters but could produce some very large clusters that would be less useful for analysis.

Conversely, a very low distance value would result in a large number of very small clusters that would be hard to analyze.

The threshold of 200 allows for higher clustering metrics without creating disproportionately sized clusters.

Hyperparameter selection

HDBSCAN

It is a clustering algorithm that can classify certain records as noise. Then, similar to Agglomerative Clustering, it generates an unlimited amount of clusters.

We set the maximum size of a group to 500 as a hyperparameter to prevent the algorithm from classifying most of the data into the same group.

OPTICS

It is another algorithm that removes noise from the data. The results can be compared to HDBSCAN.

We define the parameter min_samples as 5, which corresponds to the minimum number of records to create a group, as a hyperparameter.

When evaluating these two models, we exclude the data that was classified as noise to focus on the rest of the clusters.

Hyperparameter selection

KMeans

It is an algorithm that clusters the data into a predetermined number of clusters. To choose the best number of clusters we used the elbow method which showed that the optimal number is 15.

Birch

Birch is a hierarchical clustering algorithm that we use with the hyperparameter n_clusters to generate a given number of clusters. We set it to 15, the same number of clusters we use for KMeans. It will allow us to compare the two models.

It is important to see if the clusters are well balanced or if there is one cluster larger than the rest. The latter can happen if a part of the data that does not correspond to the already formed clusters is thrown into a default cluster. In this case, the larger group would be similar to the noise group formed by the OPTICS and HDBSCAN algorithms.

Model performance

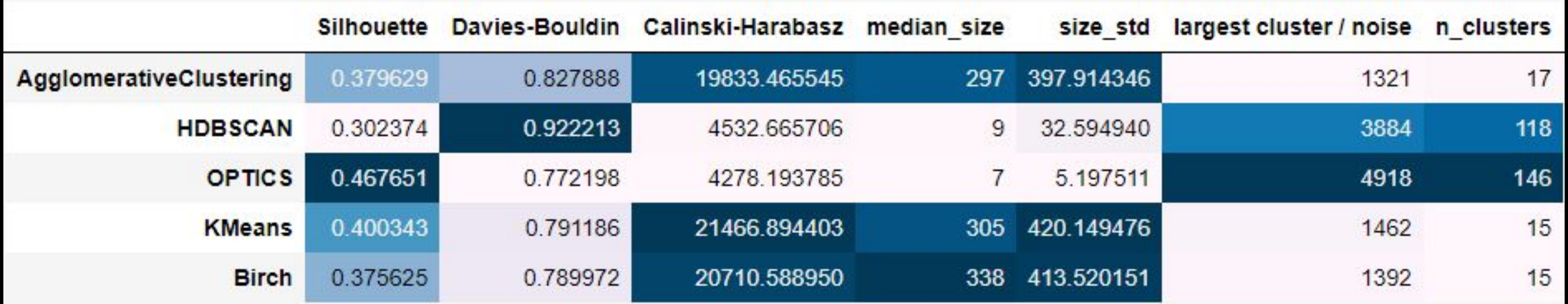
To evaluate the algorithms, we used the following metrics:

- **Silhouette score:** shows whether the formed clusters are well separated from each other. Values can be between -1 and 1. The higher the metric, the better the clustering.
- **Davies-Bouldin score:** assesses the similarity of the clusters. A low value means that the clusters are well delimited and easy to distinguish.
- **Calinski-Harabasz score:** assesses the cohesion within each cluster. When comparing models, the one with the highest value should be chosen.
- **Median size:** the median size of the clusters formed was calculated. Together with the standard deviation, it allows us to assess whether the clusters have similar sizes or whether there is an imbalance.
- **Size of the largest cluster:** this is useful for models with a limited number of clusters. In the case of the HDBSCAN and OPTICS models, the largest cluster is the noise cluster.
- **Standard deviation of the sizes:** we compare the metric with the median size to know if the clusters are of similar size or the clusters are either very large and very small.
- **Number of clusters:** this is useful for Agglomerative Clustering, HDBSCAN and OPTICS algorithms that create an unlimited number of clusters.

Table of results

We create a function to evaluate the initialised models and add the metrics to a table.

It can be found below:



	Silhouette	Davies-Bouldin	Calinski-Harabasz	median_size	size_std	largest cluster / noise	n_clusters
AgglomerativeClustering	0.379629	0.827888	19833.465545	297	397.914346	1321	17
HDBSCAN	0.302374	0.922213	4532.665706	9	32.594940	3884	118
OPTICS	0.467651	0.772198	4278.193785	7	5.197511	4918	146
KMeans	0.400343	0.791186	21466.894403	305	420.149476	1462	15
Birch	0.375625	0.789972	20710.588950	338	413.520151	1392	15

Analysis of the results

By comparing the models, we can notice that the OPTICS algorithm has the best Silhouette score and the best Davies-Bouldin score. However, both OPTICS and HDBSCAN classified more than half of the data as noise: 4918 and 3884 rows respectively. We can also see that they separated the rest of the data into many small clusters. On the one hand, that explains the low value of the Calinski-Harabasz score of both models and on the other hand denotes the great diversity of the data.

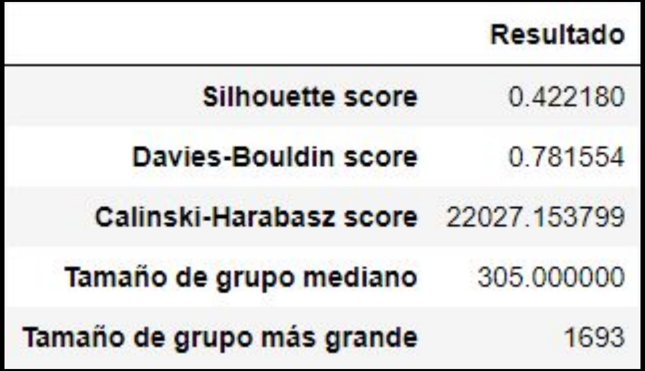
As for the Agglomerative Clustering algorithm, which did not have a predetermined number of clusters, it divided the data into 17 clusters of very different sizes, which can be deduced from the large difference between the largest and the medium-sized group, as well as from the significant standard deviation of the sizes.

Finally, the KMeans and BIRCH algorithms obtained rather similar metrics. The KMeans model has a slightly higher Silhouette score and Calinski-Harabasz score, while the BIRCH model got a better value for the Davies-Bouldin score.

Based on the metrics, we will choose KMeans as our algorithm. First, it formed a smaller number of clusters than the OPTICS and HDBSCAN algorithms, allowing us to highlight more general trends in the data. Second, it obtained better metrics than the hierarchical clustering models, i.e. BIRCH and AgglomerativeClustering.

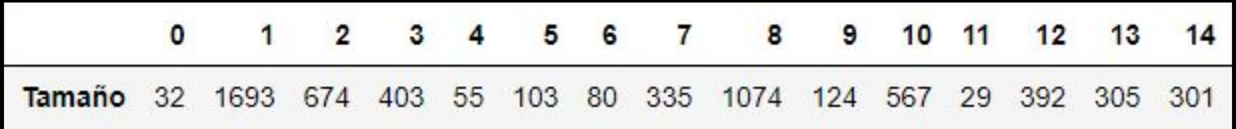
Final model

After re-running the model, we obtained the following metrics:



	Resultado
Silhouette score	0.422180
Davies-Bouldin score	0.781554
Calinski-Harabasz score	22027.153799
Tamaño de grupo mediano	305.000000
Tamaño de grupo más grande	1693

The cluster sizes can be seen below:



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Tamaño	32	1693	674	403	55	103	80	335	1074	124	567	29	392	305	301

We can see that we have two large clusters of more than a thousand units and 4 small clusters of less than 100 units. We will now proceed to the analysis of the clusters obtained.

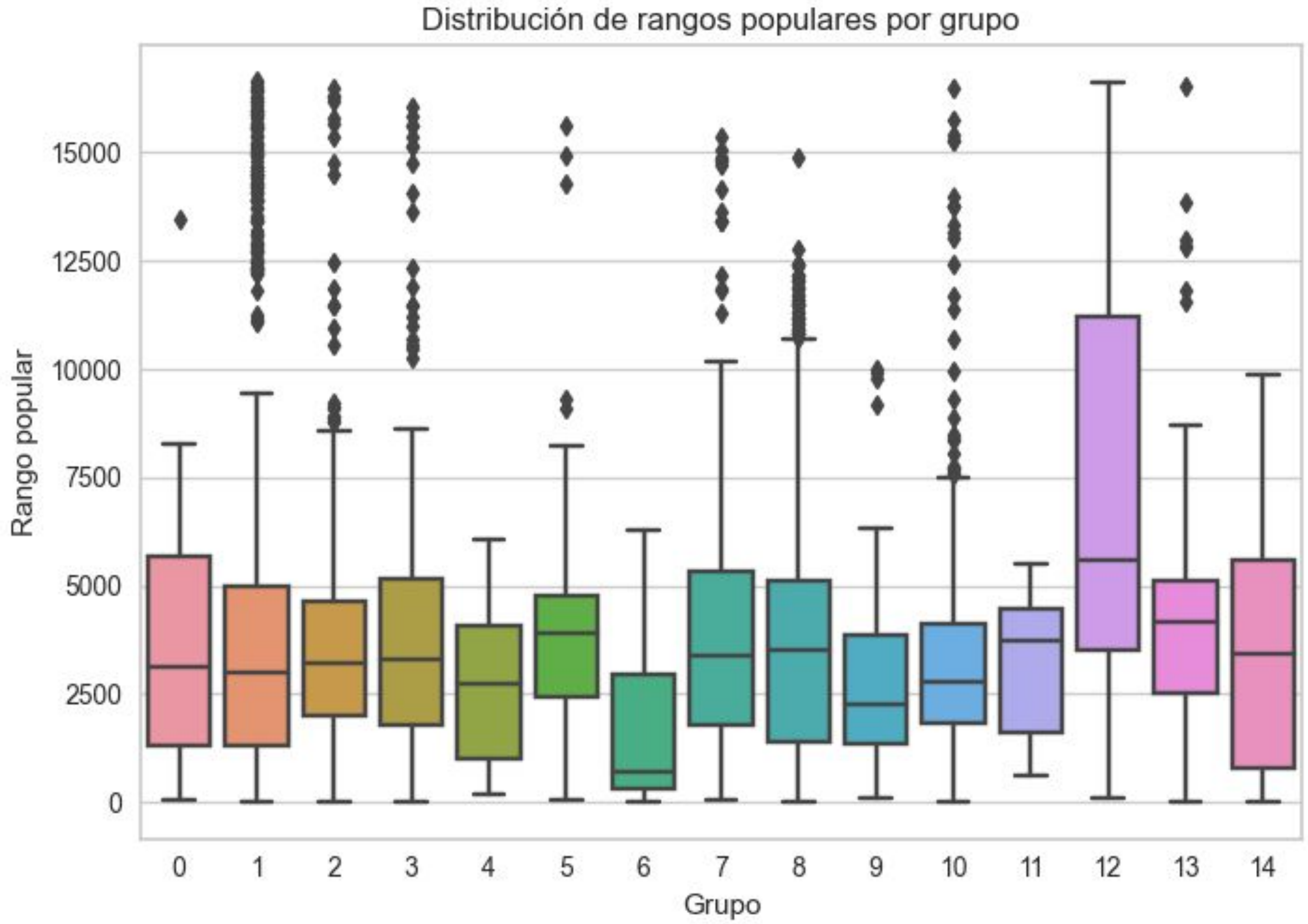
Analysis of the clusters

To carry out an analysis of the clusters, we added the group labels to the initial dataset containing the information about the popular ranks of the content. We then generated a box plot to highlight the clusters with a higher popular rank. The obtained graph can be seen in the next slide.

In the graph, cluster number 6 stands out immediately as having a better popular rank than the rest of the clusters (remember that lower ranks denote higher popularity).

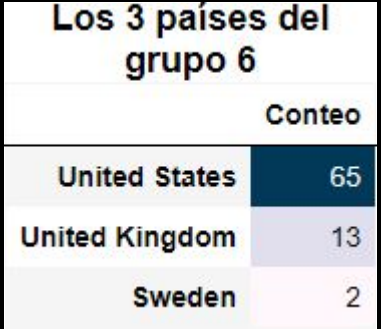
The two other clusters with a high average popularity rank are 4 and 9. Using the size table we can see that these are relatively small clusters. Cluster 6 contains only 80 contents. Cluster 4 contains 55 and cluster 9 contains 124.

We will analyse each of the clusters in detail to identify their particular characteristics.


Distribución de rangos populares por grupo
Rango popular
0
2500
5000
7500
10000
12500
15000
Grupo
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14

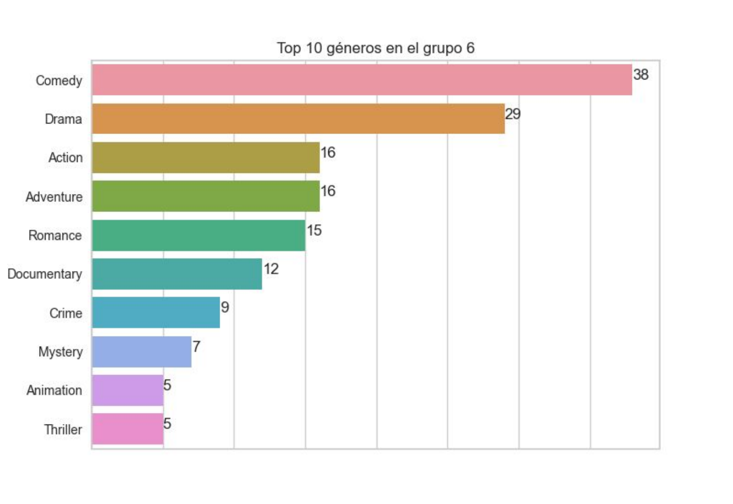
Analysis of cluster 6

We conducted an analysis of the contents of cluster 6 by country, type and genre. We can see that the cluster contains both films and series. However, the content is mostly produced in two countries: the UK and the US. Certain genres also stand out in the genre chart. We can see a predominance of comedy and dramas. Other important genres are action and adventure. This seems to confirm our findings from the analysis of the data on the most popular genres.

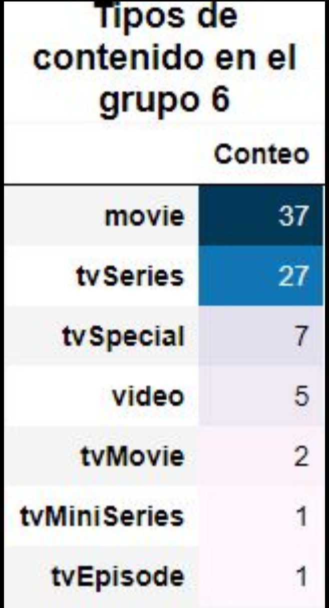


Los 3 países del grupo 6

	Conteo
United States	65
United Kingdom	13
Sweden	2



Top 10 géneros en el grupo 6



Tipos de contenido en el grupo 6

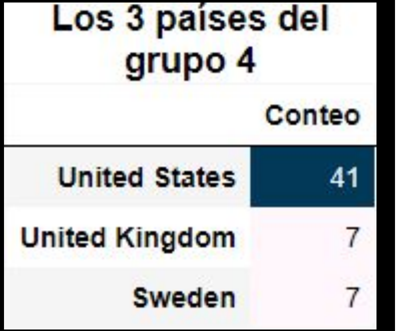
	Conteo
movie	37
tvSeries	27
tvSpecial	7
video	5
tvMovie	2
tvMiniSeries	1
tvEpisode	1

Analysis of cluster 4

This group has one outstanding content type which is drama genre films produced in the United States. The second most spread genre is crime, so we can assume that these are crime dramas.

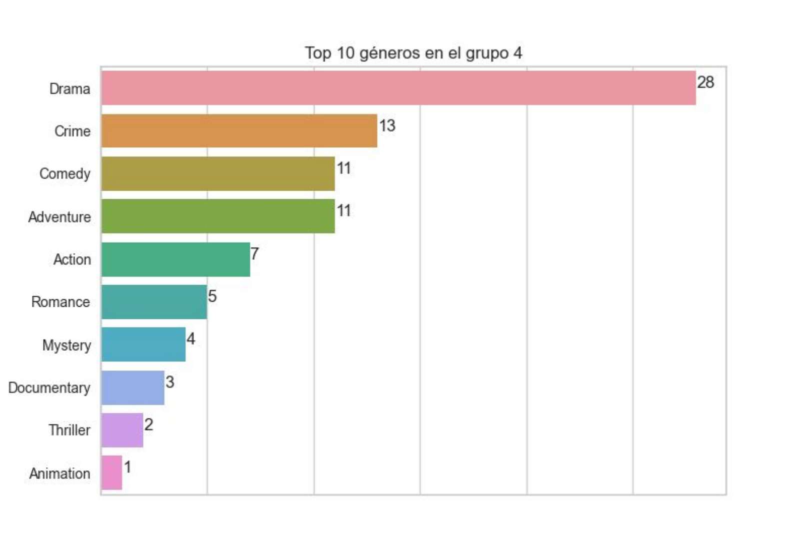
On the other hand, it should be noted that comedies and adventures are also present in the cluster.

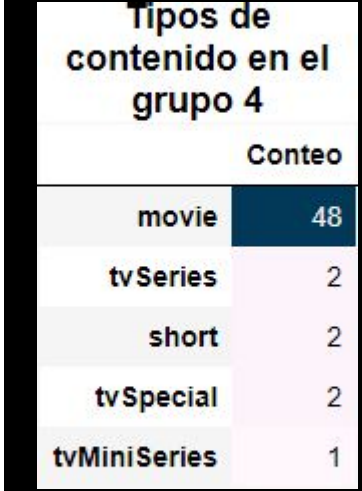
The two other countries included are the UK and Sweden which were also present in the previous group.



Los 3 países del grupo 4

	Conteo
United States	41
United Kingdom	7
Sweden	7





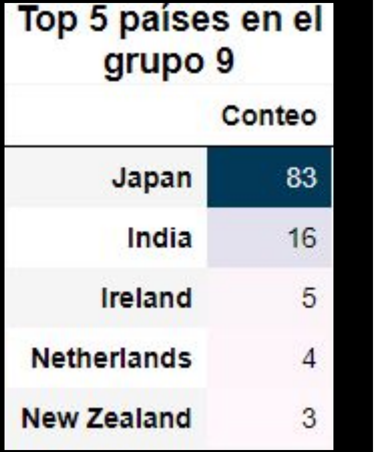
Tipos de contenido en el grupo 4

	Conteo
movie	48
tvSeries	2
short	2
tvSpecial	2
tvMiniSeries	1

Analysis of cluster 9

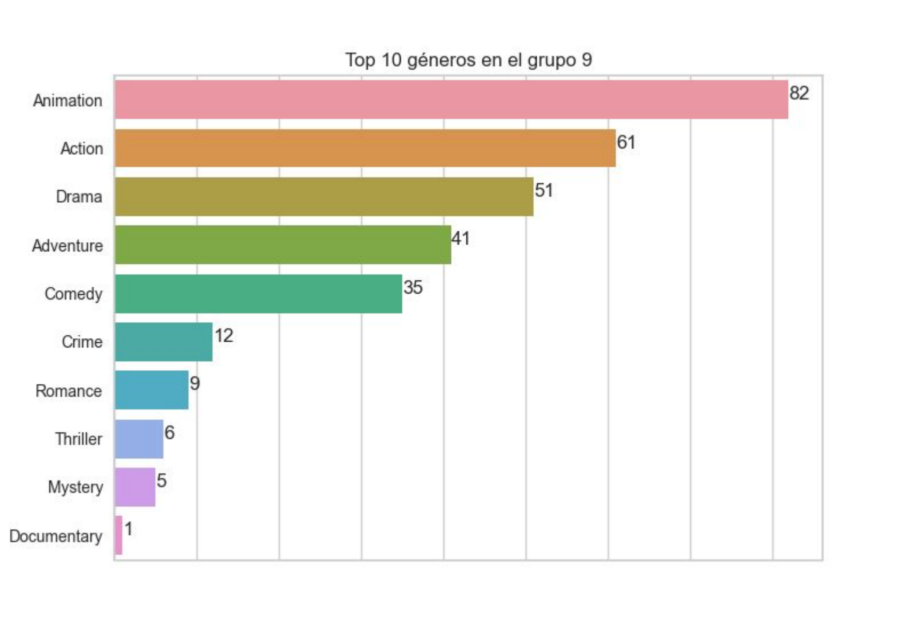
A particular type of content that stands out in this cluster is Japanese animated films and series, i.e. anime, which can belong the the genres action, adventure or drama.

This is an important new category of content that we did not identify in the exploratory data analysis.

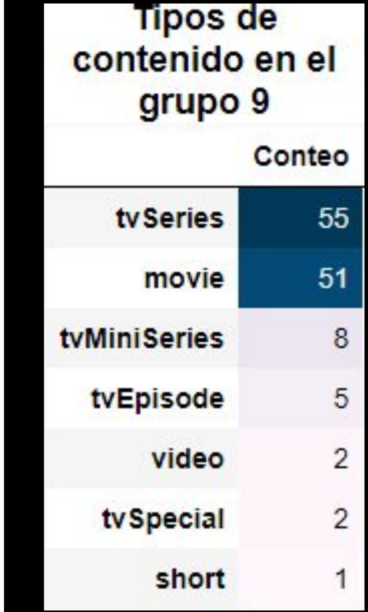


Top 5 países en el grupo 9

	Conteo
Japan	83
India	16
Ireland	5
Netherlands	4
New Zealand	3



Top 10 géneros en el grupo 9



Tipos de contenido en el grupo 9

	Conteo
tvSeries	55
movie	51
tvMiniSeries	8
tvEpisode	5
video	2
tvSpecial	2
short	1

Conclusion

The data analysis and clustering model allowed us to identify the most popular content categories on Netflix. One extremely important characteristic is genre. Drama content is the most popular type. It can be of two types: crime dramas and action dramas. The second most popular genre category is comedies which can be adventure or romantic comedies. Finally, one last outstanding category that we were able to identify using the data clustering algorithm is Japanese animated films, i.e. anime.

Then, another significant characteristic was identified, which is the production country. Content items produced in the United States and the United Kingdom enjoy the highest popularity. It can be explained by the fact that Netflix is a US platform and local content attracts the largest audience. That said, two other producer countries with good potential were highlighted, namely Japan, a major anime producer, and Sweden.

Finally, the most common content types on the platform are movies and series and they are more popular than the other types. However, the clustering algorithm did not make a clear separation between these two types, as only one cluster out of the three is mostly composed of films. So we can conclude that genre and country of production are more relevant selection criteria for the user than content being a series or a movie.

Conclusion

Based on the analysis we can make the following suggestions for the company:

- Films are as important a type of content as series, however they have a lower average quality than series. By investing in good quality films, new user groups could be attracted.
- The genres favoured by users can be divided into two groups: dramas, which can be of two types: crime or action, as well as romantic comedies and adventure comedies. Offering more of this type of content could increase user satisfaction.
- Another category with good potential for the platform is Japanese anime. A wider selection of this type of content would make the platform more attractive to users interested in that genre.
- Finally, the most popular production countries on the platform are the United States and the United Kingdom. It is the content from these countries that gets the most attention from users. However, two other countries with high potential are Japan and Sweden. The inclusion of a larger amount of content produced there would be a good strategy to diversify the content offer on the platform.

Content suggestions

Finally, we created a function to make content suggestions based on liked content.

First, we modified the dataset, leaving only the columns related to type, language, note, genre and period, in order to eliminate noise.

Then we built a Nearest Neighbors model to select the most similar content to a given one and extracted from the initial database the information about the ID, title, type, country and genre of each item suggested by the model.

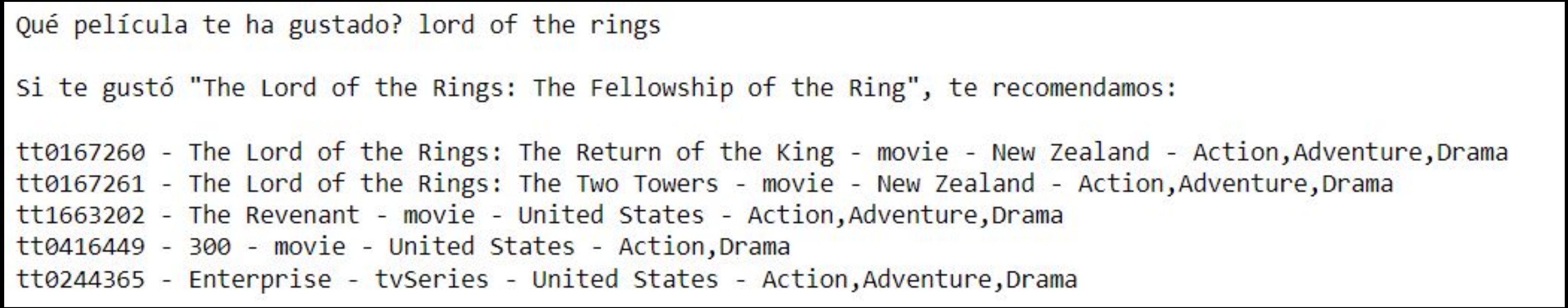
When we asked for suggestions for someone who liked the Lord of the Rings movie and the South Korean series Seutateueob, we obtained the results that can be seen in the following slide.

It can be noticed that for Lord of the Rings the model found the other films of the saga, while for Seutateueob it suggested other series produced in the same country.

Based on this test, we can conclude that the model succeeds in making relevant suggestions.

Suggestion examples

Lord of the Rings:

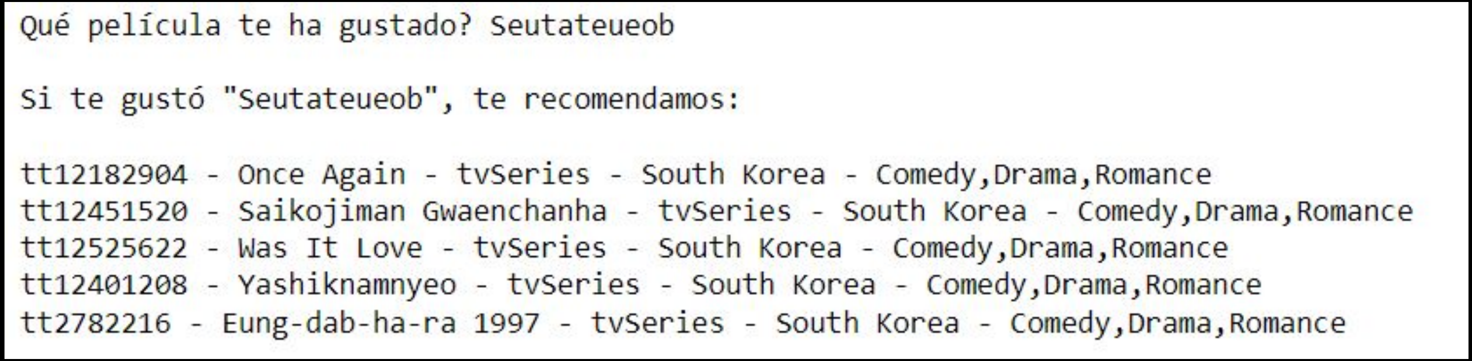


```
Qué película te ha gustado? lord of the rings

Si te gustó "The Lord of the Rings: The Fellowship of the Ring", te recomendamos:

tt0167260 - The Lord of the Rings: The Return of the King - movie - New Zealand - Action,Adventure,Drama
tt0167261 - The Lord of the Rings: The Two Towers - movie - New Zealand - Action,Adventure,Drama
tt1663202 - The Revenant - movie - United States - Action,Adventure,Drama
tt0416449 - 300 - movie - United States - Action,Drama
tt0244365 - Enterprise - tvSeries - United States - Action,Adventure,Drama
```

Seutateueob:



```
Qué película te ha gustado? Seutateueob

Si te gustó "Seutateueob", te recomendamos:

tt12182904 - Once Again - tvSeries - South Korea - Comedy,Drama,Romance
tt12451520 - Saikojiman Gwaenchanha - tvSeries - South Korea - Comedy,Drama,Romance
tt12525622 - Was It Love - tvSeries - South Korea - Comedy,Drama,Romance
tt12401208 - Yashiknamnyeo - tvSeries - South Korea - Comedy,Drama,Romance
tt2782216 - Eung-dab-ha-ra 1997 - tvSeries - South Korea - Comedy,Drama,Romance
```

Further ideas

The analysis of the dataset and the development of the clustering model allowed us to identify the most promising content categories and to create a function to propose content suggestions to the user. However, the present analysis can be extended in the following ways:

- Textual data from the ‘plot’ column could be used.
- More complete data could be extracted for other variables such as ‘certificate’, which would require the use of the IMDB API (not available for free).
- The dataset could be updated to cover more recent Netflix content.
- The suggestion function could be further improved to suggest content based on several user preferences instead of just one.
- A model could be developed to identify promising content in larger datasets containing data on movies that are not currently available on Netflix.