Netflix Movies and TV shows

APAN 5310: SQL & Relational Databases

April 30, 2020

Submitted by: Team 6
Hang Yin, Siyuan Huang,
Yi Lin, Zengyi Xu,
Yubei Tang, Harshitha Tummalapalli

Table of Contents

1.1 Scenario

Our client, Netflix, Inc. is an American media-services provider and production company. It is famous for its high-quality original TV series and movies. Each year, there will be hundreds of new series and movies released on Netflix. They have their own rating systems that people can rate for these TV series and movies and search for what they want to watch based on these ratings. Meanwhile, Netflix also gives customers recommendations on TV series and movies based on these ratings. The problem was that people don't always want to watch five-star videos: a member might give "The Godfather" five stars and "Friends" three stars but still be more interested in watching "Friends.", maybe because they prefer the genre “sitcom”, or they just prefer one of the cast. In order to solve this problem and give customers better and more personalized recommendations, They need to improve their database system: A well-designed database system where new information can easily be updated and properly and safely stored; it needs to be easier for analysts to find the trend in different regions or genres so that Netflix can use this system to make better decisions on future TV series and Movie production. What's more, by analyzing the key factors such as the content types, the rating for different content, content's growth rate, content categories, and distribution of movie or TV show's duration through this dataset, Netflix can understand its strengths and weaknesses and be able to improve its content quality and management system.

Our team was hired to develop this database system to meet these requirements. We were given some unstructured data and asked to organize them for easy access and analysis. we will create a movie/series profile database system with normalized tables, and create different authorization levels for departments to allow them to have quick access to only the information they need.

1.2 Proposal

The objective of the project is to create a movie and series profile database system that safely stores data in a structured format, is user friendly, and provides departments with different authorization levels allowing them quick access to the information they need. Analysts will be given the authorization to write and execute SQL code and access the database through Python/R, whilst Managers/C-executives will be given authorization for high-level overview through visualizations and interactive dashboards that automatically update when new data is stored in the database.

The new database system will provide numerous benefits as detailed below:

1. Analysis of the relationships between different variables such as ratings could help Netflix polish its recommendation system by providing specific on-demand recommendations for customers based on their preferences for certain factors.

2. Analytical trends provide an indication of the target audience and their preference thereby guiding what content Netflix should produce using its limited resources i.e. what genre? more movies or TV shows? or should they invest in producing originals or buy rights to classics?
3. Analytical insights can also be used to improve Netflix marketing strategies i.e. Are certain customers trending toward specific types of covers? If so, should personalized recommendations automatically change? Which title colors appeal to which customers? Is there an ideal cover for an original series? Or should different colors be used for different audiences?

In effect to the above, we obtained a dataset from Kaggle that was compiled from a third-party Netflix search engine, consisting of TV shows and movies available on Netflix as of 2019. It includes 6,235 TV shows and movies on Netflix and 11 attributes to describe each one, such as title, director, cast, genre, release year, and TV rating, among others. We further supplemented this dataset with an additional dataset from IMDb. IMDb is a website devoted to collecting movie data supplied by studios and fans. It claims to be the biggest movie database on the web and is run by Amazon. The dataset provides information on customer ratings and votes for Movies & TV shows.

Below are the links to the datasets and a screenshot of the sample data:
Netflix movies & TV shows - https://www.kaggle.com/shivamb/netflix-shows
IMDb ratings - https://www.imdb.com/interfaces/

show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
81145628	Movie	Norm of the North: King Siz	Richard Finn, Tim Maltby	Alan Marriott, Andrev	United State	9-Sep-19	2019	TV-PG	90 min	Children & F	Before planning an awesome wed
80117401	Movie	Jandino: Whatever it Takes		Jandino Asporaat	United King	9-Sep-16	2016	TV-MA	94 min	Stand-Up C	Jandino Asporaat riffs on the chal
70234439	TV Show	Transformers Prime		Peter Cullen, Sumalee	United State	8-Sep-18	2013	TV-Y7-FV	1 Season	Kids' TV	With the help of three human allie
80058654	TV Show	Transformers: Robots in Disguise		Will Friedle, Darren Ci	United State	8-Sep-18	2016	TV-Y7	1 Season	Kids' TV	When a prison ship crash unleash
80125979	Movie	#realityhigh	Fernando Lebrija	Nesta Cooper, Kate W	United State	8-Sep-17	2017	TV-14	99 min	Comedies	When nerdy high schooler Dani fi
80163890	TV Show	Apaches		Alberto Ammann, Elo	Spain	8-Sep-17	2016	TV-MA	1 Season	Crime TV Sh	A young journalist is forced into a
70304989	Movie	Automata	Gabe Ib谩帽ez	Antonio Banderas, Dy	Bulgaria, Ur	8-Sep-17	2014	R	110 min	Internationa	In a dystopian future, an insuranc
80164077	Movie	Fabrizio Copano: Solo piens	Rodrigo Toro, Francisco Schultz	Fabrizio Copano	Chile	8-Sep-17	2017	TV-MA	60 min	Stand-Up C	Fabrizio Copano takes audience p
80117902	TV Show	Fire Chasers			United State	8-Sep-17	2017	TV-MA	1 Season	Docuseries,	As California's 2016 fire season ra
70304990	Movie	Good People	Henrik Ruben Genz	James Franco, Kate Hi	United State	8-Sep-17	2014	R	90 min	Action & Ac	A struggling couple can't believe t
80169755	Movie	Joaqu铆n Reyes: Una y no n	Jos茅 Miguel Contreras	Joaqu铆n Reyes		8-Sep-17	2017	TV-MA	78 min	Stand-Up C	Comedian and celebrity imperson
70299204	Movie	Kidnapping Mr. Heineken	Daniel Alfredson	Jim Sturgess, Sam Wc	Netherlands	8-Sep-17	2015	R	95 min	Action & Ac	When beer magnate Alfred "Fredc
80182480	Movie	Krish Trish and Baltiboy		Damandeep Singh Baggan, Smita		8-Sep-17	2009	TV-Y7	58 min	Children & F	A team of minstrels, including a m
80182483	Movie	Krish Trish and Baltiboy: Bat	Munjal Shroff, Tilak Shetty	Damandeep Singh Baggan, Smita		8-Sep-17	2013	TV-Y7	62 min	Children & F	An artisan is cheated of his payme
80182596	Movie	Krish Trish and Baltiboy: Bes	Munjal Shroff, Tilak Shetty	Damandeep Singh Baggan, Smita		8-Sep-17	2016	TV-Y	65 min	Children & F	A cat, monkey and donkey team u
80182482	Movie	Krish Trish and Baltiboy: Co	Tilak Shetty	Damandeep Singh Baggan, Smita		8-Sep-17	2012	TV-Y7	61 min	Children & F	In three comic-strip-style tales, a
80182597	Movie	Krish Trish and Baltiboy: Ov	Tilak Shetty	Rishi Gambhir, Smita Malhotra, De		8-Sep-17	2017	TV-Y7	65 min	Children & F	A cat, monkey and donkey learn t
80182481	Movie	Krish Trish and Baltiboy: Part II		Damandeep Singh Baggan, Smita		8-Sep-17	2010	TV-Y7	58 min	Children & F	Animal minstrels narrate stories ab
80182621	Movie	Krish Trish and Baltiboy: The	Munjal Shroff, Tilak Shetty	Damandeep Singh Baggan, Smita		8-Sep-17	2013	TV-Y7	60 min	Children & F	The consequences of trickery are e
80057969	Movie	Love	Gaspar No茅	Karl Glusman, Klara K	France, Belg	8-Sep-17	2015	NR	135 min	Cult Movies	A man in an unsatisfying marriage
80060297	Movie	Manhattan Romance	Tom O'Brien	Tom O'Brien, Katherir	United State	8-Sep-17	2014	TV-14	98 min	Comedies, I	A filmmaker working on a docume
80046728	Movie	Moonwalkers	Antoine Bardou-Jacquet	Ron Perlman, Rupert	France, Belg	8-Sep-17	2015	R	96 min	Action & Ac	A brain-addled war vet, a failing b
80046727	Movie	Rolling Papers	Mitch Dickman		United State	8-Sep-17	2015	TV-MA	79 min	Documenta	As the newspaper industry takes a
70304988	Movie	Stonehearst Asylum	Brad Anderson	Kate Beckinsale, Jim S	United State	8-Sep-17	2014	PG-13	113 min	Horror Movi	In 1899, a young doctor arrives at
80057700	Movie	The Runner	Austin Stark	Nicolas Cage, Sarah P	United State	8-Sep-17	2015	R	90 min	Dramas, Ind	A New Orleans politician finds his
80045922	Movie	6 Years	Hannah Fidell	Taissa Farmiga, Ben R	United State	8-Sep-15	2015	NR	80 min	Dramas, Ind	As a volatile young couple who ha
80244601	TV Show	Castle of Stars		Chaivapol Pupart, Jintanutda Lum		7-Sep-18	2015	TV-14	1 Season	Internationa	As four couples with different lifes
80203094	Movie	City of Joy	Madeleine Gavin		United State	7-Sep-18	2018	TV-MA	77 min	Documenta	Women who've been sexually bru
80190843	TV Show	First and Last				7-Sep-18	2018	TV-MA	1 Season	Docuseries	Take an intimate look at the emot
70241607	Movie	Laddaland	Sopon Sukdapisit	Saharat Sangkapreech	Thailand	7-Sep-18	2011	TV-MA	112 min	Horror Movi	When a family moves into an ups
80988892	Movie	Next Gen	Kevin R. Adams, Joe Ksander	John Krasinski, Charly	China, Cana	7-Sep-18	2018	TV-PG	106 min	Children & F	When lonely Mai forms an unlikel
80239639	Movie	Sierra Burgess Is A Loser	Ian Samuels	Shannon Purser, Krist	United State	7-Sep-18	2018	PG-13	106 min	Comedies, F	A wrong-number text sparks a vir

1.3 Database Schema and Normalization Plan

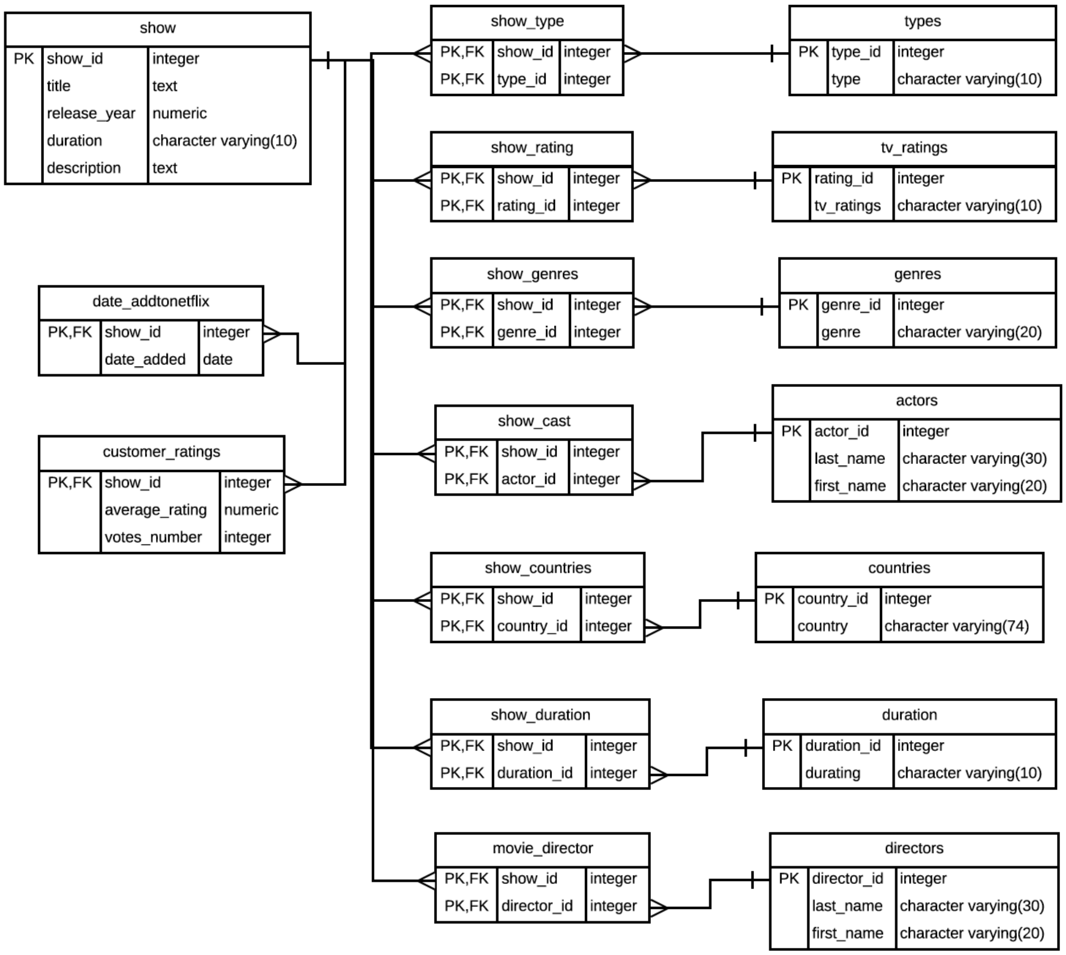
According to the scenario and analytics application, we create 17 tables in the relational database. All the tables we created are in third normal form(3NF). Here are the seventeen tables we created：

- Show
- date_addtoNetflix
- Customer_ratings
- Show_type
- Show_rating
- Show_genres
- Show_cast
- Show_countries
- Show_duration
- Movie_director
- Types
- TV_ratings
- Genres
- Actors
- Countries
- Duration
- Directors

We first create the table show with the attributes to have an overall view of the dataset. Show_id is created to be the primary key. For table Types, TV_ratings, Genres, Actors, Countries, Duration, Directors, we create schemas for these tables in this logic: We set unique ID as Primary key to types, ratings, genres, actors, countries, duration, directors and their corresponding specific content. For table Show_type, Show_rating, Show_genres, Show_cast, Show_countries, Show_duration, Movie_director,date_addtoNetflix, Customer_ratings, we connect the tables we created in the early stage with the show_id. We make the id created in the early stage as the foreign key to these tables whose primary key is show_id. Based on the third normal form (3NF) tables we created, it makes our database more efficient and logical, storing data according to relevance, removing data redundancy, and increasing data quality. The code of creating these tables can be seen at https://github.com/siyuan2016/APAN5310_Project, called Group 6__Create 3NF tables.

The Entity-Relationship Diagram (ER diagram) we created can be seen as follows. This diagram is a graphical representation of the flow of data and information of Netflix Tv shows and movies.

Visualizing entities and relationship sets from ER diagrams allows us to understand the dataset better. This will also help us in interacting with the database and make analytical applications in an efficient way. The link to this diagram is the following:
https://www.lucidchart.com/documents/view/bf9a4440-8282-43c9-8719-8a729060815b/0_0



1.4 ETL Process

Based on the normalization plan in the former stage, we create the database tables in pgAdmin through R studio. With the database and all tables created (3NF), then we extract, transform and load (ETL) the dataset into the database to prepare the dataset for analysis. The detailed process of each table can be seen as following:

- Show:

We check whether there are any duplicate rows in the dataset to ensure the uniqueness of the table. Then we load the data into the show table to have an overall view of the show information.

- Types:

Within the dataset there are repeating movie types so we cannot simply add a column with incrementing integer numbers for the primary key of movies as this would lead to movies with multiple primary keys. There are only 2 types: Movies and TV shows. Therefore, we first use the unique function to make “type” unique. Then, we push the data into the database.

- TV_ratings:

Within the dataset, there are repeating movie ratings so we cannot simply add a column with incrementing integer numbers for the primary key of movies as this would lead to movies with multiple primary keys. There are only 14 types of ratings. Therefore, we first use the unique function to make “tv_ratings” unique. Then, we push the data into the database.

- Genres:

Firstly, we check the uniqueness of genres. Then we use the function 1:nrow() to define the genres_id with incrementing integer numbers.

- Actors:

First, we splitted the cast using function 1:nrow(). We did not get rid of the comma in the cast because one cast just has too many names under it if we delete the comma. This was not helpful for our analysis. Second, we created a subset of df corresponding to the traders database table. Last, we pushed the data into the database.

- Countries:

First, we used the function “unique” for the variable “country” because there are many repeating countries under one country. Then we split the country using the function “1:nrow()”. Lastly, we pushed the data into the database.

- Duration:

First, we used the function “unique” for the variable “duration” because there are many repeating durations. Then we splitted the duration using the function “1:nrow()”. This was not helpful for our analysis. Lastly, we pushed the data into the database.

- Directors:

First, we split the variable directors, getting rid of the comma using the function “strsplit()” because there is more than one person in the directors variable separated by a comma “,”. Then we created a new list called “director_id_list” and inserted the new list into the “df”. Lastly, we pushed the data into the database.

- Date_added

First, we can easily create a subset of df corresponding to the data_added database table, including “show_id” and “date_added”. Then, we pushed the data into the database.

- Customer_ratings

First, we can easily create a subset of df corresponding to the cutomer_ratings database table, including “show_id” and “rating”. Then we renamed the second column of customer_rating table into “rating level”. Lastly, we pushed the data into the database.

- Show_type

This table connects table types and the show_id. First,we use show_type_list to create a list called show_type_list mapping show_id with each type and then inserting this list to the dataset as a new column.

- Show_rating

As the tv_ratings table we have already created to filter the unique rating levels. And then, we create a rating framework to create a list mapping show_id with each show in the dataset and then inserting this list to the dataset as a new column.

- Show_genres

Since we have generated a genres table with unique genres, defining them with incrementing numbers, we create a show_genres_list framework to create a list mapping show_id with each show in the dataset and then inserting this list to the dataset as a new column.

- Show_cast

According to the actor table we created before, we just combine the actor_id and show_id correspondingly to create a show_cast table.

- Show_countries

As the countries table we have already created to filter the unique country names. And then, we create a show_ country framework to create a list mapping show_id with each show in the dataset and then inserting this list to the dataset as a new column.

- Show_duration

This table is related to the table duration. As the duration table we have already created to filter the unique duration levels. And then, we create a `duration_list` framework to create a list mapping show_id with each show in the dataset and then inserting this list to the dataset as a new column.

1.5 Analytics Applications

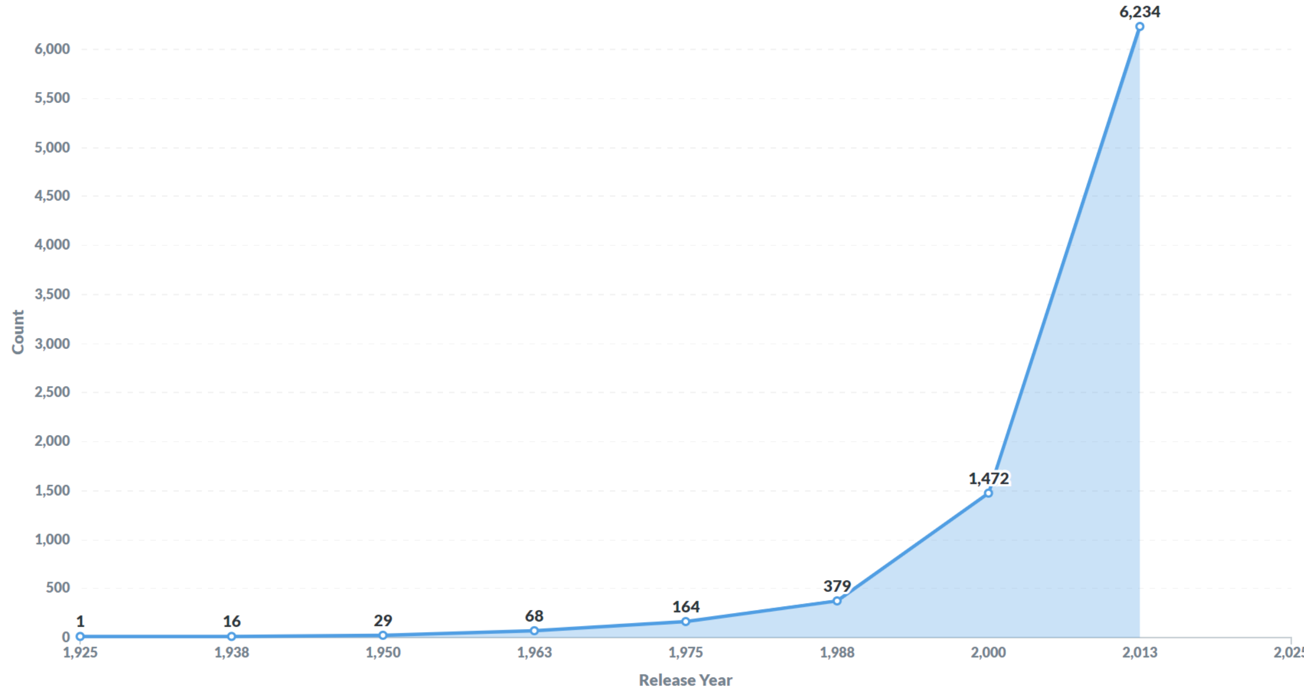
After building the database, we come up with two main applications: first is to build/update the contents management system and the recommendation system for Netflix, and second is to capture the trend of customers' preference and help the management team decide what movie or TV shows to purchase in the future.

General Trends

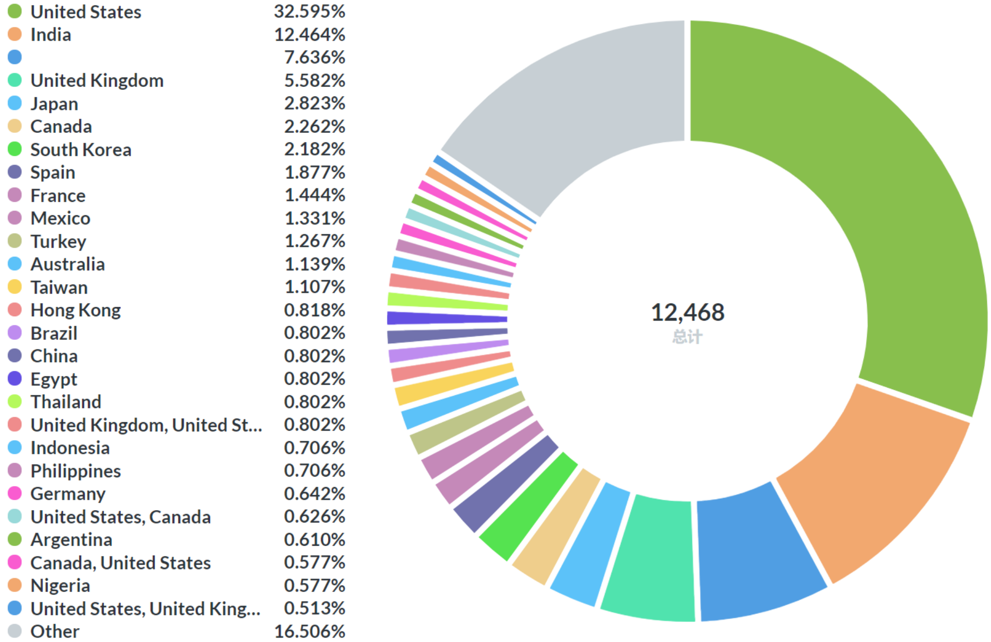
We use Metabase and Tableau to generate business insights, and we also use R and python to make some complicated computations.

Metabase: http://team06-metabase.herokuapp.com

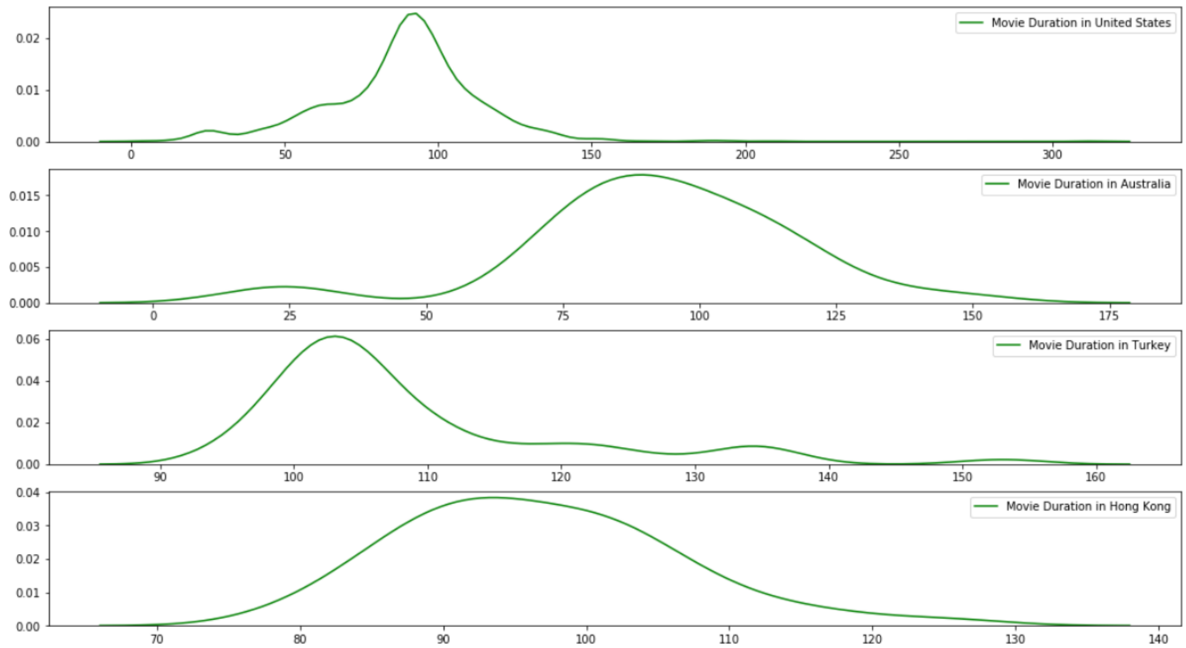
1) Content over the years



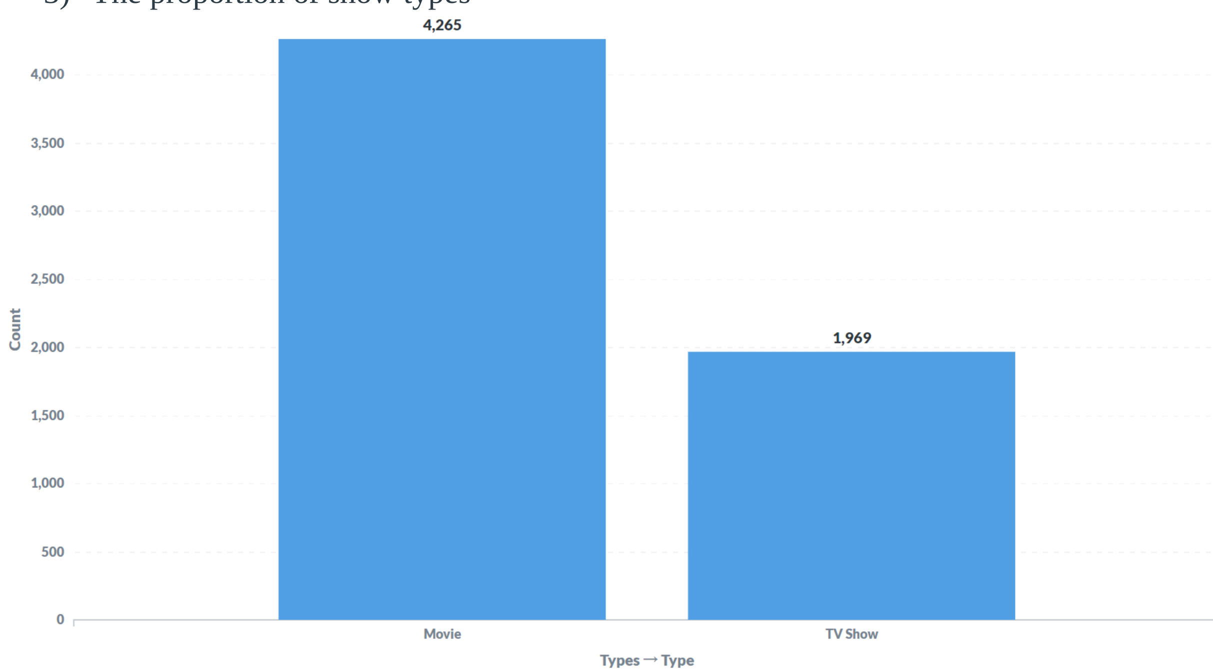
2) Distribution over countries



In North America like the United States and Mexico and European countries like Turkey and the UK, people like movies/shows between 50 minutes and 110 minutes. But in Australia there are more movies/shows between 75 minutes and 125 minutes.

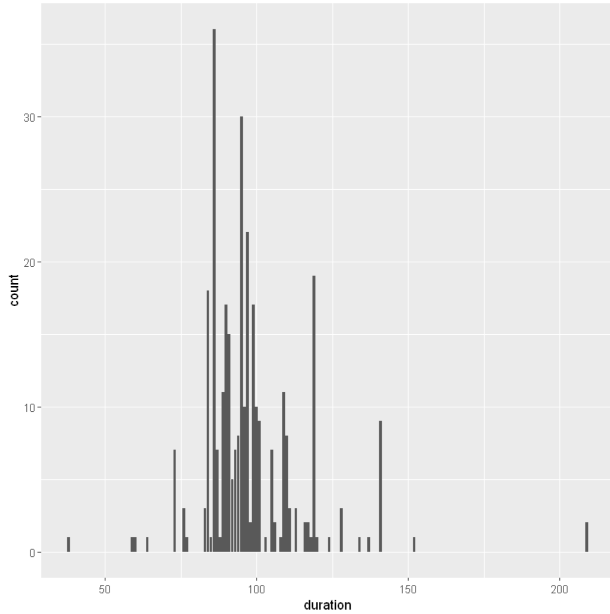


3) The proportion of show types



In total, there are more movies than shows. But there is a difference between different regions. For example, both the US and India produce more movies than shows. In other countries like Japan and the UK produce more shows than movies.

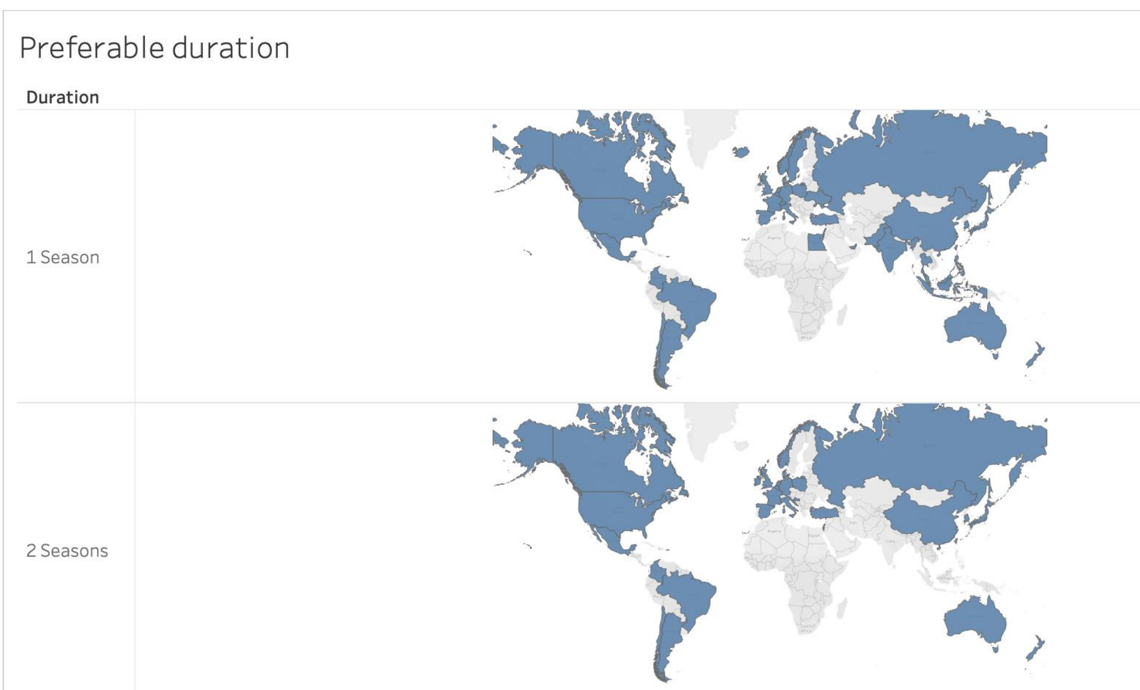
4) Durations distribution of movies

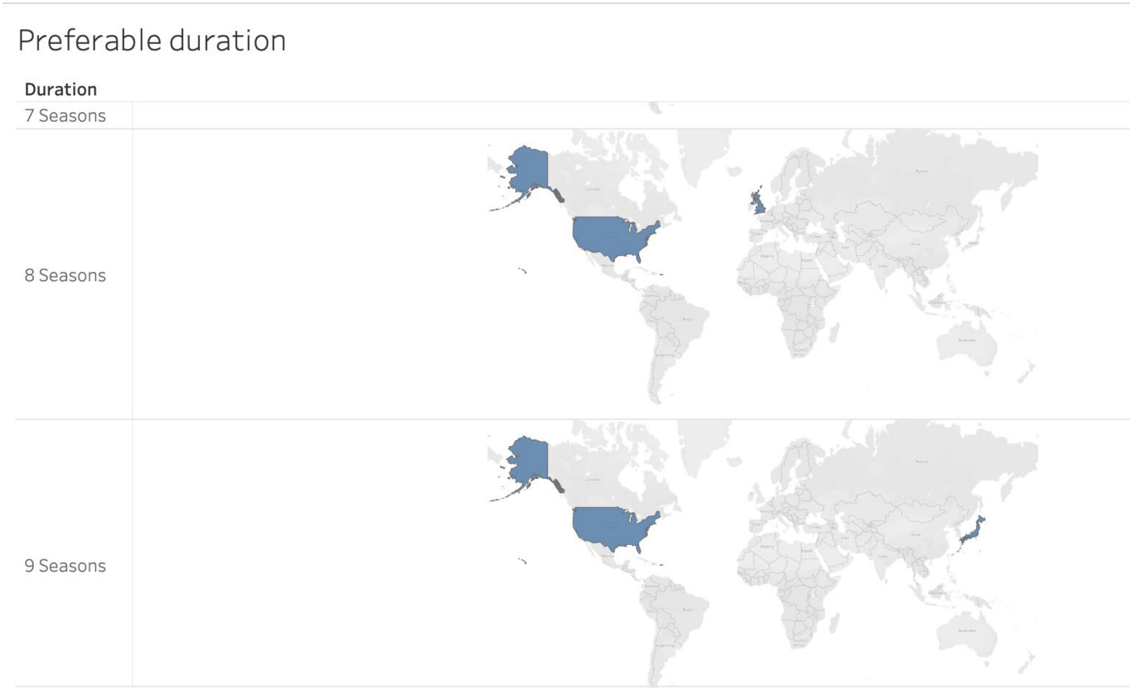


duration	count
86 min	36
95 min	30
97 min	22
119 min	19
84 min	18
90 min	17
99 min	17
91 min	15
109 min	11
89 min	11
100 min	10
96 min	10

As we can see, for movies, a duration of around 80 to 100min is the most popular one.

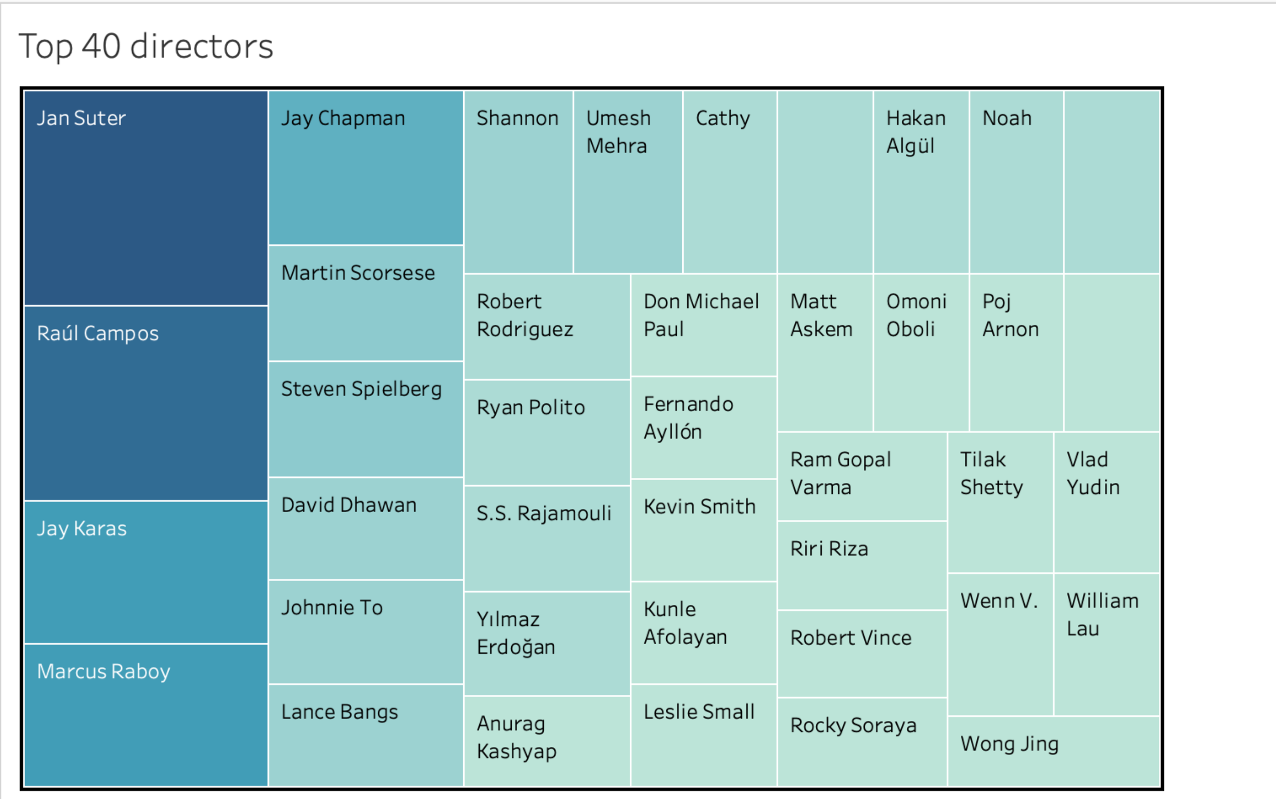
5) Preferable TV show duration





Depending on different regions and countries, there is a visible preference for the show's duration. Most countries prefer having 1 to 3 seasons for a show. However, only the UK, USA, and Japan have the shows which are up to 8 or 9 seasons.

6) Top directors



Here are the top 40 directors ranked by the number of the shows/movies they produced.

Improve Content Management System

All of these charts and tables are made by analytical tools like Tableau, Metabase, R and Python. Our dataset is designed to be 3NF, which is more structured and user friendly. Since Tableau and Python can easily connect to our database, analysts are able to query and extract the relative information very easily and quickly. Moreover, there is no worries about disorganizing the original dataset.

Update Recommendation System

In our database, we do not have useful information. In this case, we cannot make personalized recommendations to each user. Instead, we can build non-personalized recommendations by finding out what kinds of shows are most popular.

7) Linear regression model (the relationship between genre, TV ratings and review ratings)

We joined the review rating table from IMDb with the Netflix show table by matching the movie title, and built a linear regression model using genre and TV ratings to predict review ratings. The result shows that TV ratings are not very related with ratings, but some genres are very correlated with the review ratings. For example: Action+Adventure+Animation, Action+Adventure+Crime, Action+Adventure+Drama, Action+Adventure+Fantasy, Action+Crime+Drama, Action+Drama+Fantasy, Action+Drama+Mystery, Action+Drama+War, Action+Thriller.

```
Call:
lm(formula = V2 ~ rating + genres, data = finaltable)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3916 -0.5638  0.0810  0.6917  3.6956

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                           6.247524   0.211836  29.492  < 2e-16 ***
ratingNR                              0.018665   0.201268   0.093 0.926114
ratingPG                             -0.193578   0.199073  -0.972 0.330892
ratingPG-13                          -0.070123   0.192540  -0.364 0.715725
ratingR                               0.048406   0.189877   0.255 0.798781
ratingTV-14                          -0.035363   0.188597  -0.188 0.851269
ratingTV-G                            0.237382   0.251827   0.943 0.345904
ratingTV-MA                          -0.004705   0.187607  -0.025 0.979993
ratingTV-PG                          -0.020997   0.191324  -0.110 0.912614
ratingTV-Y                           -0.586373   0.382364  -1.534 0.125195
ratingTV-Y7                          -0.275689   0.260347  -1.059 0.289677
ratingTV-Y7-FV                       -0.286005   0.279832  -1.022 0.306794
ratingUR                             -0.372201   0.503995  -0.739 0.460240
genresAction                         -0.467339   0.228742  -2.043 0.041088 *
genresAction,Adult,Crime              1.087840   1.134173   0.959 0.337524
genresAction,Adventure                0.449691   0.257538   1.746 0.080842 .
genresAction,Adventure,Animation      1.014207   0.155239   6.533 6.98e-11 ***
genresAction,Adventure,Biography      0.326583   0.660302   0.495 0.620903
genresAction,Adventure,Comedy         0.232362   0.182769   1.271 0.203657
genresAction,Adventure,Crime          0.790095   0.165921   4.762 1.96e-06 ***
genresAction,Adventure,Documentary    1.273474   1.134854   1.122 0.261845
genresAction,Adventure,Drama          0.934667   0.157373   5.939 3.03e-09 ***
genresAction,Adventure,Family         0.686939   0.287388   2.390 0.016866 *
genresAction,Adventure,Fantasy        0.664688   0.196418   3.384 0.000719 ***
genresAction,Adventure,History        1.857181   0.805372   2.306 0.021146 *
genresAction,Adventure,Horror        -0.765933   0.660186  -1.160 0.246024
genresAction,Adventure,Mystery        0.728726   0.355953   2.047 0.040677 *
genresAction,Adventure,Romance        0.222540   0.660312   0.337 0.736112
genresAction,Adventure,Sci-Fi         0.664664   0.218862   3.037 0.002401 **
```

8) Non-personalized recommendation systems

Since we don't have user information, based on our system we will recommend movies and TV shows with the highest ratings. The top 10 shows we got are:

	title	type	V2
4537	The Chosen	Movie	9.8
5463	The Phantom of the Opera	Movie	9.8
2068	Hush	Movie	9.7
383	Big Time	Movie	9.6
1196	Dog Eat Dog	Movie	9.6
4585	The Code	TV Show	9.6
4594	The Code	TV Show	9.6
5509	The Promise	Movie	9.6
626	Breaking Bad	TV Show	9.5
2212	Jail	Movie	9.5

We also look at the most popular TV shows and movies in 2019 to look for latest trend:

	title	release_year	type	V2
3358	Pegasus	2019	Movie	9.1
5140	The Last Laugh	2019	Movie	9.1
1222	Domino	2019	Movie	9.0
1997	Homeland	2019	Movie	9.0
5138	The Last Laugh	2019	Movie	8.9
1548	For Love or Money	2019	Movie	8.8
1601	Fractured	2019	Movie	8.8
2276	Joy	2019	Movie	8.8
2450	Let's Dance	2019	Movie	8.8
5122	The King	2019	Movie	8.7

	title	release_year	type	V2
4827	The Gift	2019	TV Show	9.5
6261	What If?	2019	TV Show	9.5
293	Bad Blood	2019	TV Show	9.4
1837	Haunted	2019	TV Show	9.4
905	Chosen	2019	TV Show	9.3
6094	Undercover	2019	TV Show	9.2
5524	The Protector	2019	TV Show	9.1
6294	White Gold	2019	TV Show	9.1
5726	The Spy	2019	TV Show	9.0
187	Arrow	2019	TV Show	8.9

1.6 Conclusion

According to our database schema design, we build 17 tables and our database is satisfied with 3NF, which means every non-prime attribute of R is non-transitively dependent on every key of R. We try to make it normalized and decrease redundancy and the requirement for storage, for example, compromise with a higher need of storage for a separate type table and an additional type column in the shows table and create separate tables for actor and director.

We are using PostgreSQL to store tables and write queries as our coding tool. In addition, we would use Metabase and Tableau as our main visualization tools. To sum up, R and PostgreSQL is our coding tool, and Metabase and Tableau is our visualization tools.

Our database can help analysts to better query, extract and analyze data, and it's more convenient to link to the analytical and visualization tools. Here are some finding we discovered: Analytical procedures lead us to find insights that some directors are very popular and have a high ranking. In addition, genres like adventure, drama and animation are very highly correlated with the rating. This concludes our final recommendation for Netflix: Investing more in the top 40 directors and TV/Movies of popular genres, like Action+Adventure+Animation, in the future would improve the overall ratings and polish the recommendation system. Meanwhile, we would also like to continue discovering the relationship between ratings and user satisfaction rate/ratings to design a more personalized recommendation system. In such fierce competitions in the online streaming industry, offering a high-end and personalized watching experience is the main competitive edge for companies like Netflix to compete with traditional movie theatres as well as its direct competitors.

Appendix

- ER Diagram is generated as following link:
  https://www.lucidchart.com/documents/view/bf9a4440-8282-43c9-8719-8a729060815b/0_0

- All code of NETFLIX project is accessible a GitHub:
  https://github.com/siyuan2016/APAN5310_Project

- Metabase: http://team06-metabase.herokuapp.com

- Our dataset comes from Kaggle Project: https://www.kaggle.com/shivamb/netflix-shows

- IMDb_ratings （title.basics.tsv.gz ; title.ratings.tsv.gz) https://www.imdb.com/interfaces/

ER Diagram from Lucidchart