INFO 7374 : Algorithmic Digital Marketing

Group 6 : Omkar Kalange and Sanghamitra Shanmugam

Claat:
https://codelabs-preview.appspot.com/?file_id=1XOuJ7S7FMm9Nfqhbg6RiSjJA26slJ1vlo1iktMOm1D4#0

Google doc:
https://docs.google.com/document/d/1XOuJ7S7FMm9Nfqhbg6RiSjJA26slJ1vlo1iktMOm1D4/edit?ts=5f774829#

MovieLens 25M : Insights and Analytics

Summary	Uncovering marketing insights and performing analysis on the 25M reviews for movies from 1960 to 2019
Category	Web

About the Dataset

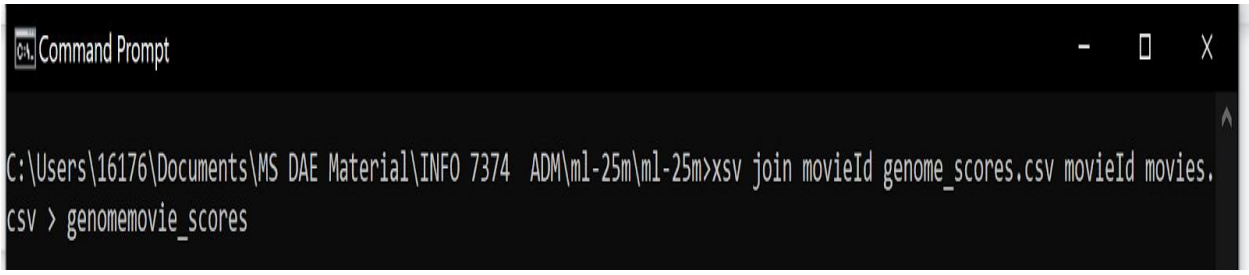
This dataset describes 5-star rating and free-text tagging activity from [MovieLens](http://movielens.org), a movie recommendation service. These ratings were given for the movies during 1995 to 2019. Users were selected at random for inclusion. All selected users had rated at least 20 movies. The dataset also contains the relevance of the tags given by users.

The files contained in dataset were :

- Ratings.csv
- GenomeTags.csv
- Movies.csv
- GenomeScores.csv

Working on Dataset using XSV

1. Using XSV Join command - Performed join operations using the MovieID to get the title for relevance scores of different tags



```
Command Prompt
C:\Users\16176\Documents\MS DAE Material\INFO 7374  ADM\ml-25m\ml-25m>xsv join tagId genomemovie_scores.csv tagId genome
-tags.csv > genome_movies.csv
```

2. Taking samples from genome scores using XSV Sample : For loading into Snowflake Datawarehouse

```
Command Prompt
C:\Users\16176\Documents\MS DAE Material\INFO 7374  ADM\ml-25m\ml-25m>xsv sample 900000 genome-scores.csv > genome_score
s.csv
```

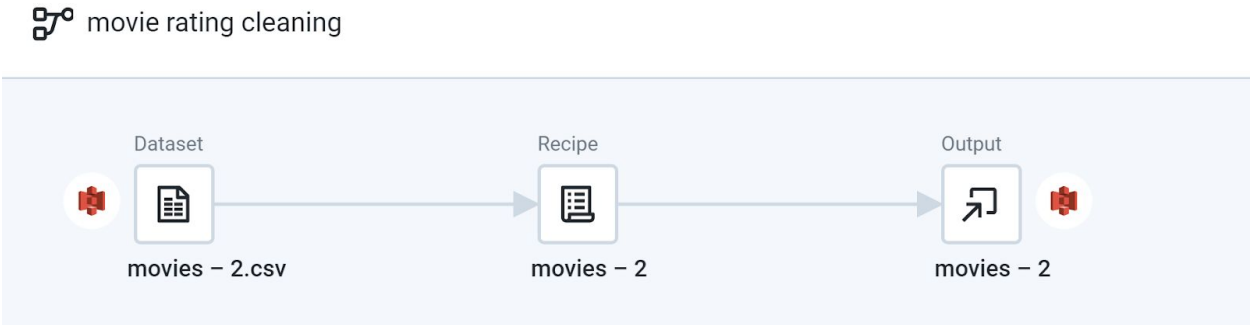
3. Calculated the Stats of the Sampled data for ratings using XSV stats command

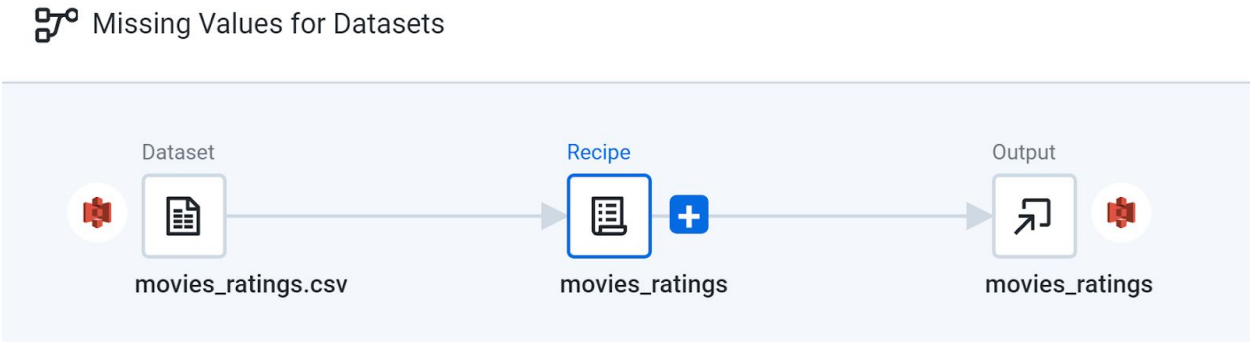
```
C:\Users\16176\Documents\MS DAE Material\INFO 7374  ADM\ml-25m\ml-25m>xsv stats movies_ratings.csv
field,type,sum,min,max,min_length,max_length,mean,stddev
movieId,Integer,3794322266,1,209133,1,6,32586.072363448922,48975.42959183392
title,Unicode,, Days of Summer (2009),Я худею ,1,153,,
release_year,Integer,231945957,1408,2150,4,4,1991.978332188251,20.823757079725702
genres,Unicode,,(no genres listed),Western,3,77,,
,NULL,,,,0,0,,
year_of_rating,Integer,234067971,1996,2019,4,4,2010.2024304362765,6.396004684215687
Avg_rating,Float,384777.19999992487,0.5,5,1,3,3.304510477499171,0.8757024519860734
```

Trifacta Wrangler : Data Cleaning and creating flows

Wrangled the datasets using Trifacta

1. Joined the datasets using XSV and imported them to Trifacta for further analysis





2. Created recipe for regular expressions for deducing new columns

movies – 2

Edit Recipe | Add | •••

Recipe | Data

Steps Preview

```
1  Replace matches of '\"' from title with ''
2  Extract matches of /[\d]{4}/ from title
3  Rename title1 to 'release_year'
4  Delete rows where
   ISMISMATCHED(release_year,
   ['Datetime','yy','yyyy'])
5  Delete rows where ISMISSING([release_year])
6  Replace matches of /[\(][\d]*[\)]/ from title
   with ''
7  Sort rows by title
8  Sort rows by title
9  Split title on delimiters matching '(' into 2
   columns
10 Delete title2
11 Rename title1 to 'title'
```

3. Recipe for Date Time objects

movies_ratings

Edit Recipe | Add | ...

Recipe | Data

Steps Preview

1 **Create** removesymbols_title from REMOVESYMBOLS(title)

2 **Delete** title

3 **Delete rows** where ISMISMATCHED(release_year, ['Datetime','yy','yyyy'])

4 **Change** removesymbols_title **type** to String

5 **Delete rows** where ISMISSING([removesymbols_title])

Steps	5
Updated	Last Tuesday at 11:17 AM
Created	Last Tuesday at 11:14 AM

ratings5l

Edit Recipe | Add | ...

Recipe | Data

Steps Preview

1 **Delete** userId

2 **Change** timestamp **type** to Datetime, mm-dd-yy, mm*dd*yy

3 **Change** timestamp **type** to Datetime, mm-dd-yy, mm*dd*yy

4 **Create** unixtimeformat_timestamp from UNIXTIMEFORMAT(timestamp, 'yyyy-MM-dd')

5 **Delete** unixtimeformat_timestamp

6 **Change** timestamp **type** to Integer

7 **Create** Unixtimestamp from timestamp * 1000

8 **Delete** timestamp

9 **Create** Date from UNIXTIMEFORMAT(Unixtimestamp, 'yyyy-MM-dd')

10 **Delete** Unixtimestamp

Recipe Data

```
8 Delete timestamp
9 Create Date from UNIXTIMEFORMAT(Unixtimestamp, 'yyyy-MM-dd')
10 Delete Unixtimestamp
11 Change Date type to Datetime, yy-mm-dd, yyyy*mm*dd
12 Create year_Date from YEAR(Date)
13 Create month_Date from MONTH(Date)
14 Delete month_Date
15 Create average_rating from AVERAGE(rating) grouped by movieId, year_Date
16 Create Avg_rating from ROUND(average_rating, 1)
17 Delete average_rating
18 Sort rows by year_Date
19 Sort rows by movieId
20 Rename year_Date to 'year'
```

4. Jobs ran successfully eliminating the null values and the data is ready to be staged in SnowFlake

Completed stages

Transform with profile

Completed Last Tuesday at 11:19 AM, started Last Tuesday at 11:17 AM • Ran for 1 min

Environment Spark

100% valid values 0% mismatching values 0% missing values

View steps and dependencies View profile

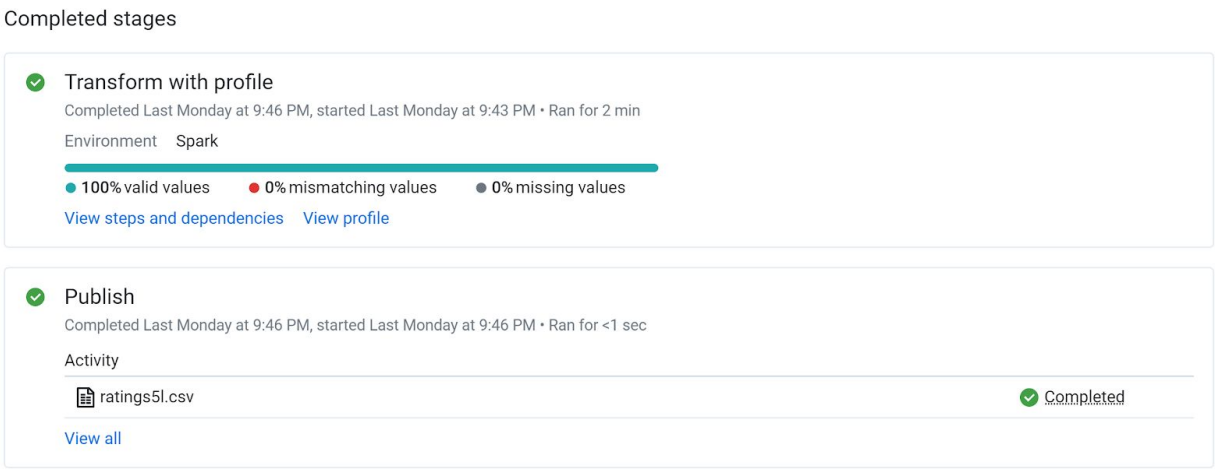
Publish

Completed Last Tuesday at 11:19 AM, started Last Tuesday at 11:19 AM • Ran for <1 sec

Activity

movies_ratings.csv Completed

View all



Snowflake : Staging into Data Warehouse

1. After wrangling data into Trifacta, the data has been loaded into Snowflake. The first thing to be done was to create two tables : Movie Ratings and Genome Relevance

Databases > MOVIES > **MOVIE_RATINGS (PUBLIC)**

Tables | Views | Schemas | Stages | File Formats | Sequences | Pipes

Load Table

Table Name: MOVIE_RATINGS

Column Name	Ordinal ▲	Type	Nullable
MOVIEID	1	NUMBER(38,0)	false
RELEASE_YEAR	2	NUMBER(38,0)	true
GENRES		VARCHAR(200)	true
YEAR_OF_RATING	4	NUMBER(38,0)	true
AVG_RATING	5	FLOAT	true
TITLE	6	VARCHAR(200)	true

Databases > MOVIES > **GENOME_RELEVANCE (PUBLIC)**

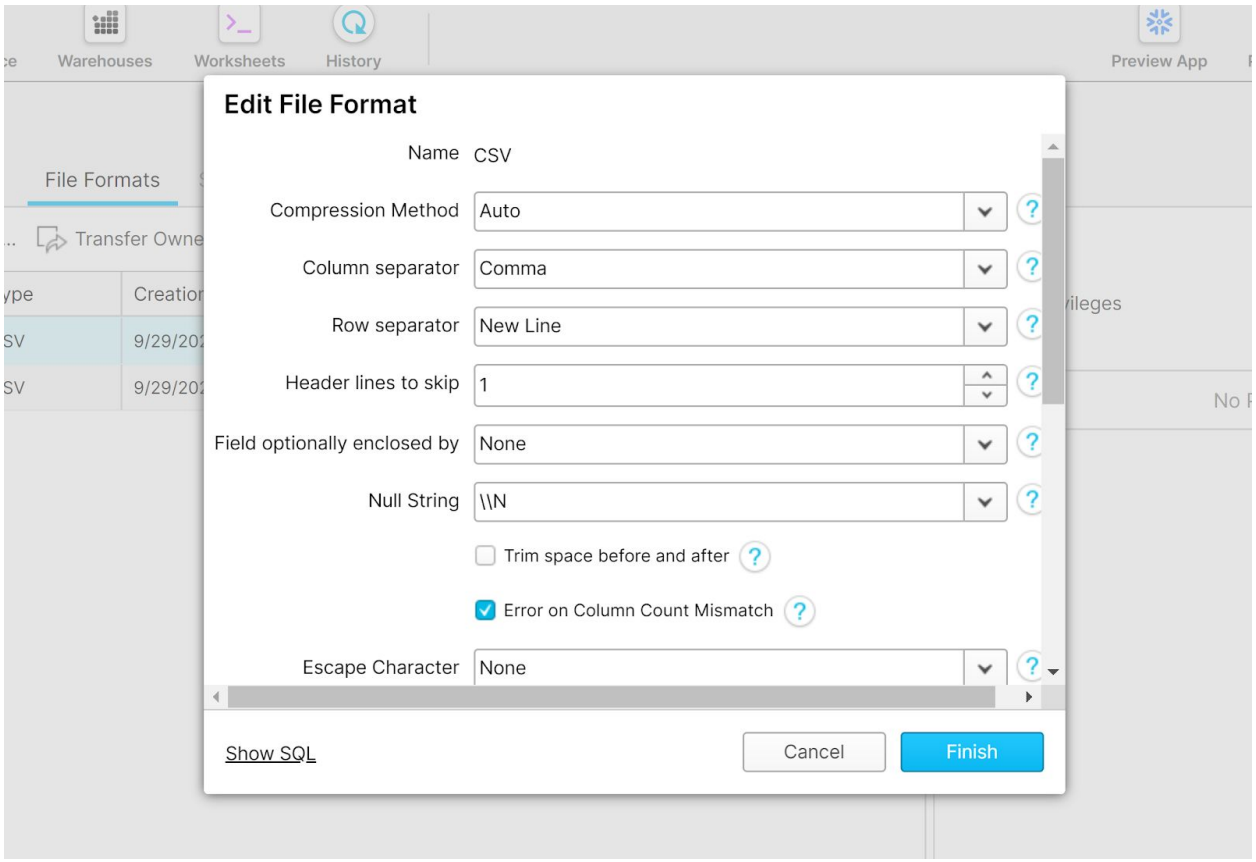
Tables | Views | Schemas | Stages | File Formats | Sequences | Pipes

Load Table

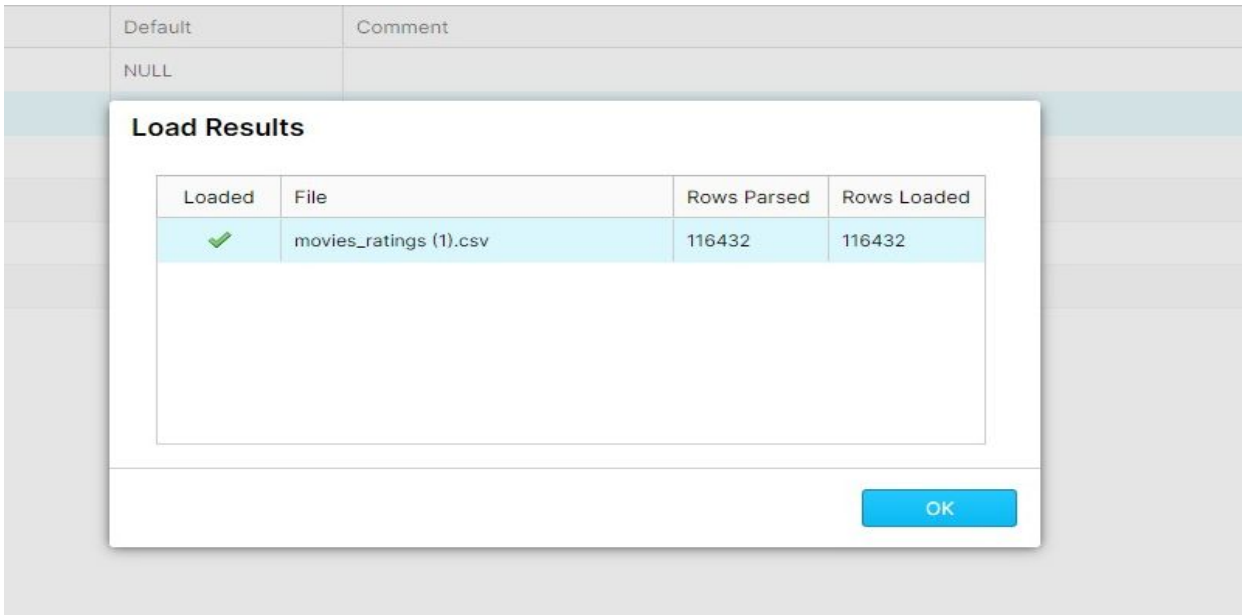
Table Name: GENOME_RELEVANCE

Column Name	Ordinal ▲	Type	Nullable
MOVIEID	1	NUMBER(38,0)	false
TITLE		VARCHAR(100)	false
RELEASE_YEAR	3	NUMBER(38,0)	false
TAG	4	VARCHAR(100)	false
RELEVANCE	5	FLOAT	false
GENRE	6	VARCHAR(100)	true

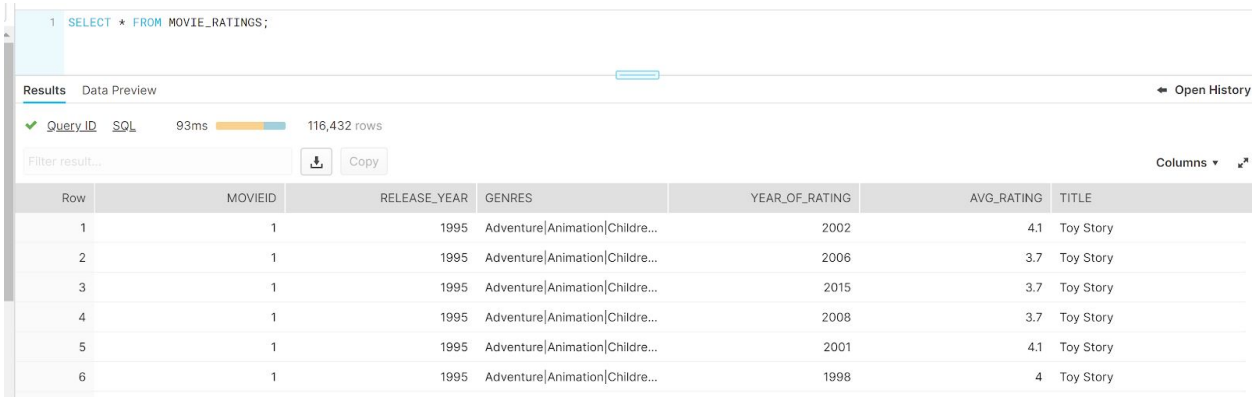
2. Creating file formats for csvs to be loaded into database tables :



3. Loading the data into tables, ensuring the same count of rows and columns from csv to database table :



4. Verifying the data using SQL queries on the data warehouse :



Row	MOVIEID	RELEASE_YEAR	GENRES	YEAR_OF_RATING	AVG_RATING	TITLE
1	1	1995	Adventure|Animation|Childre...	2002	4.1	Toy Story
2	1	1995	Adventure|Animation|Childre...	2006	3.7	Toy Story
3	1	1995	Adventure|Animation|Childre...	2015	3.7	Toy Story
4	1	1995	Adventure|Animation|Childre...	2008	3.7	Toy Story
5	1	1995	Adventure|Animation|Childre...	2001	4.1	Toy Story
6	1	1995	Adventure|Animation|Childre...	1998	4	Toy Story

```
1 SELECT * FROM GENOME_RELEVANCE;
```

Results Data Preview Open History

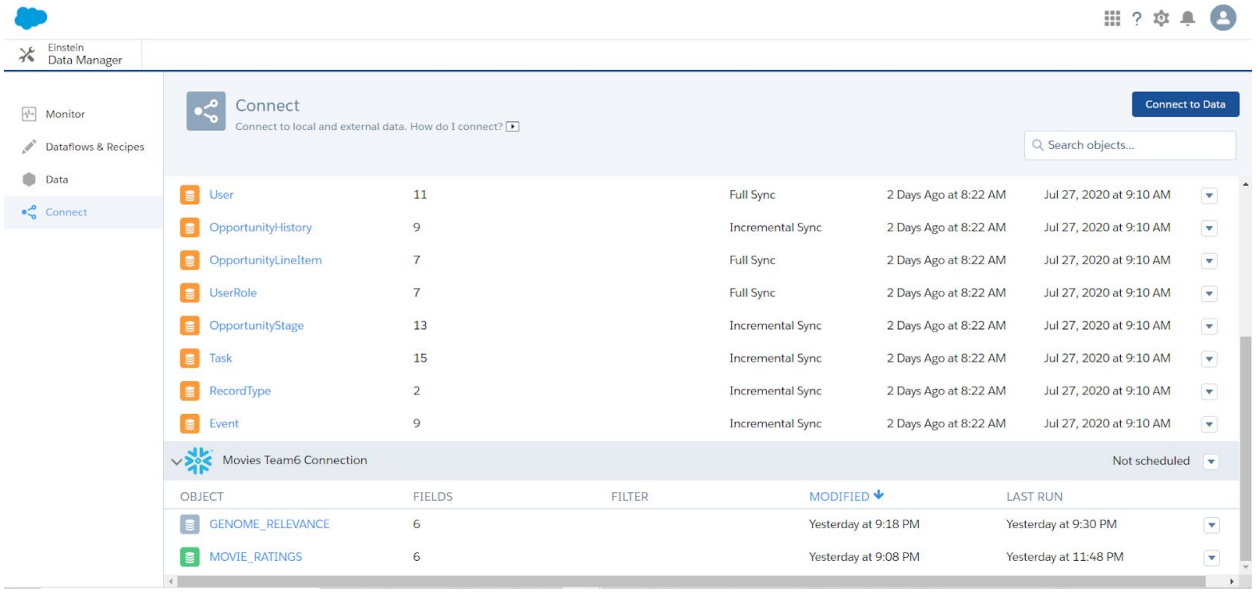
Query ID SQL 1.72s 100,000 rows

Filter result... Copy Columns

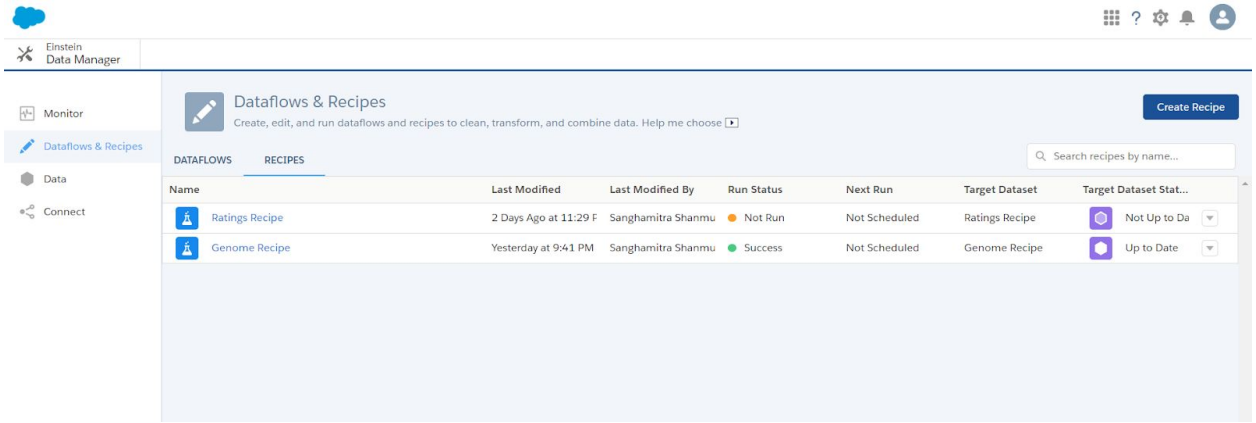
Row	MOVIEID	TITLE	RELEASE_YEAR	TAG	RELEVANCE	GENRE
1	7185	Real Men	1987	unusual plot structure	0.17725	Comedy|Sci-Fi
2	329	Star Trek Generations	1994	computer animation	0.20475	Adventure|Drama|Sci-Fi
3	1702	Flubber	1997	gothic	0.06	Children|Comedy|Fantasy
4	1287	BenHur	1959	entertaining	0.377	Action|Adventure|Drama
5	2606	Idle Hands	1999	harry potter	0.0085	Comedy|Horror
6	74668	District 13 Ultimatum	2009	passionate	0.1145	Action|Sci-Fi
7	89904	The Artist	2011	storytelling	0.9015	Comedy|Drama|Romance

Connecting Snowflake to Salesforce

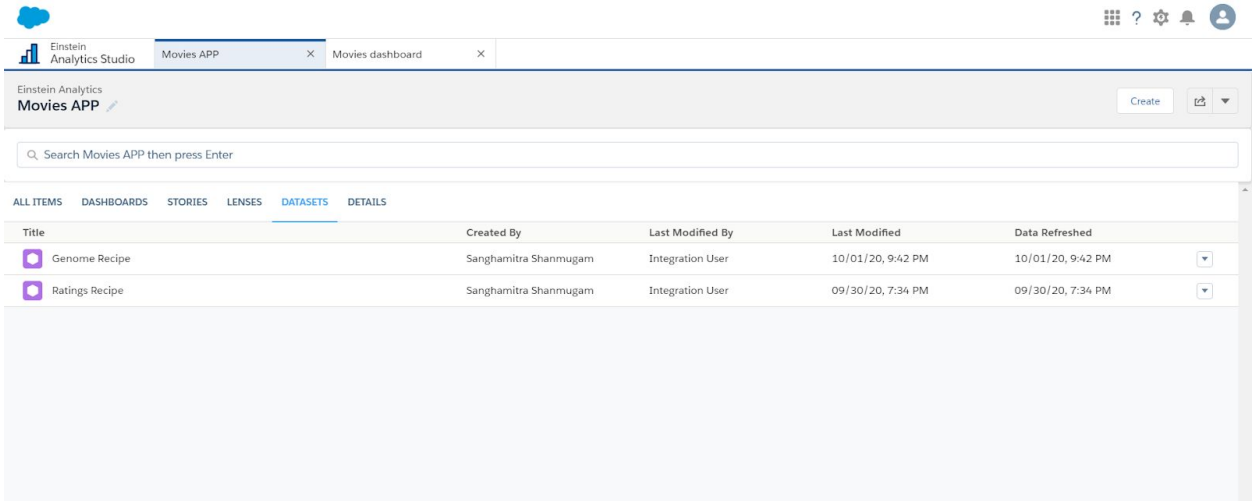
1) Connecting Snowflake to Salesforce and importing data



2) Creating Recipes out of connected data



3) Connecting the App to the recipe



Analytics Dashboard using Salesforce Einstein Analytics

Movie Lens Dashboard

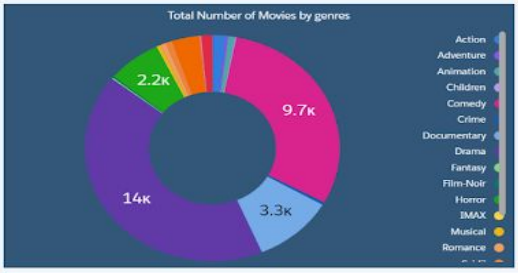
Total Genres and their categories:- 1,279

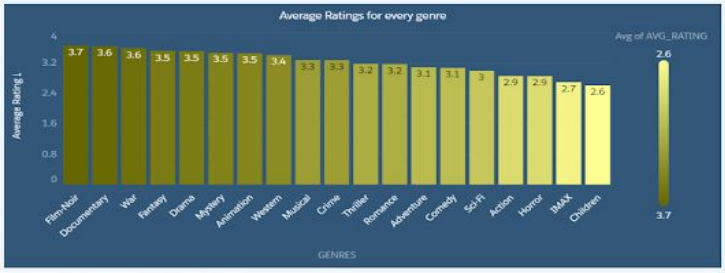
Total Tags:- 1,128

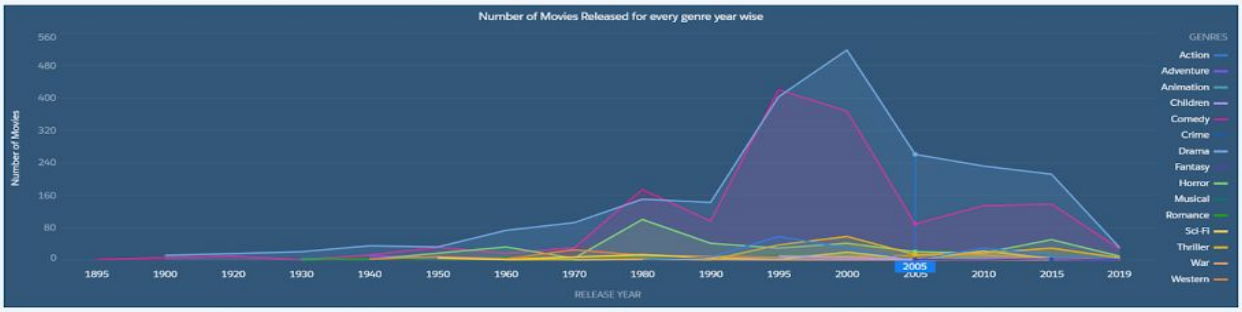
Promotions

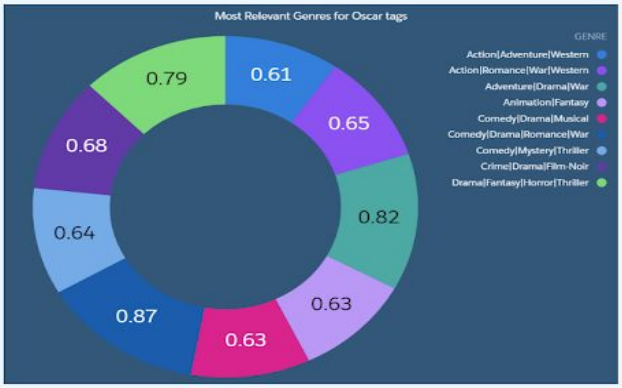
We created these below graphs to get insights-

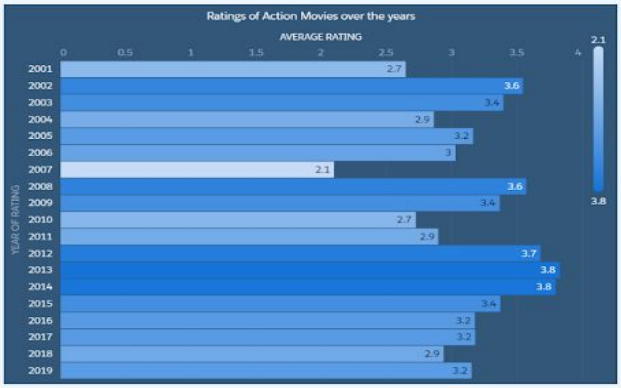
1. **Number of movies in every genre and their ratings**- To get insights on the range of movies released every year and their reach to the public
2. **Track the genre of movies released across all years**- To see the effect of ratings make a significant change in the number of movies released in genres
3. **Analyse users ratings for old movies in recent years** - To see the demand for old movies in the current generation and compare it against the genre of movies released recently
4. **Track the movies with the tag oscar with its ratings and genre:-** To get insights on the most relevant genres with oscar tags

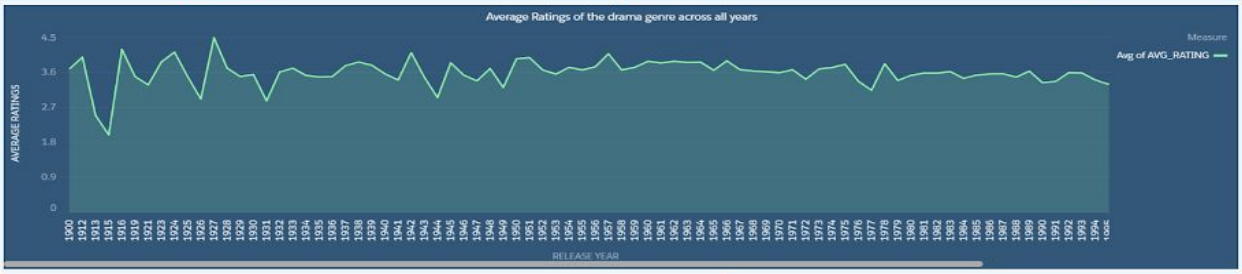

Total Number of Movies by genres
9.7K
3.3K
14K
2.2K
Action
Adventure
Animation
Children
Comedy
Crime
Documentary
Drama
Fantasy
Film-Noir
Horror
IMAX
Musical
Romance

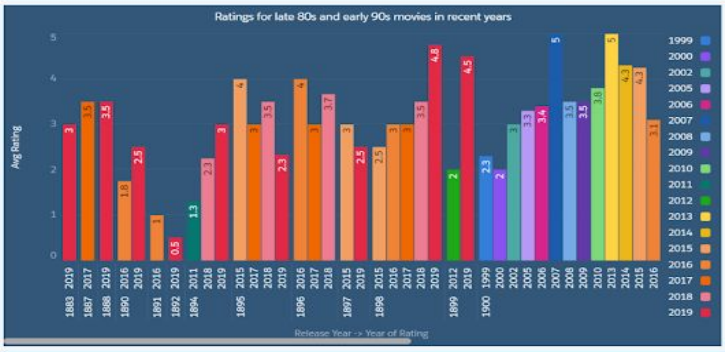

Average Ratings for every genre
Avg of AVG_RATING
Average Rating
GENRES
Film-Noir 3.7
Documentary 3.6
War 3.6
Fantasy 3.5
Drama 3.5
Mystery 3.5
Animation 3.5
Western 3.4
Musical 3.3
Crime 3.3
Thriller 3.2
Romance 3.2
Adventure 3.1
Comedy 3.1
Sci-Fi 3
Action 2.9
Horror 2.9
IMAX 2.7
Children 2.6

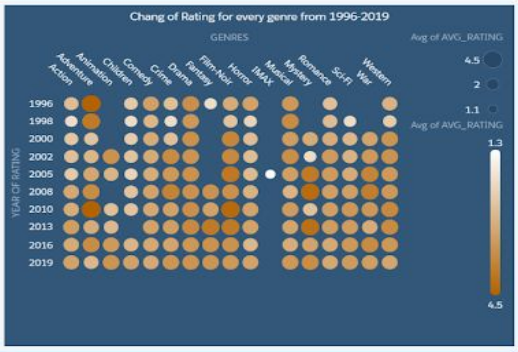

Number of Movies Released for every genre year wise
Number of Movies
RELEASE YEAR
GENRES
Action
Adventure
Animation
Children
Comedy
Crime
Drama
Fantasy
Horror
Musical
Romance
Sci-Fi
Thriller
War
Western

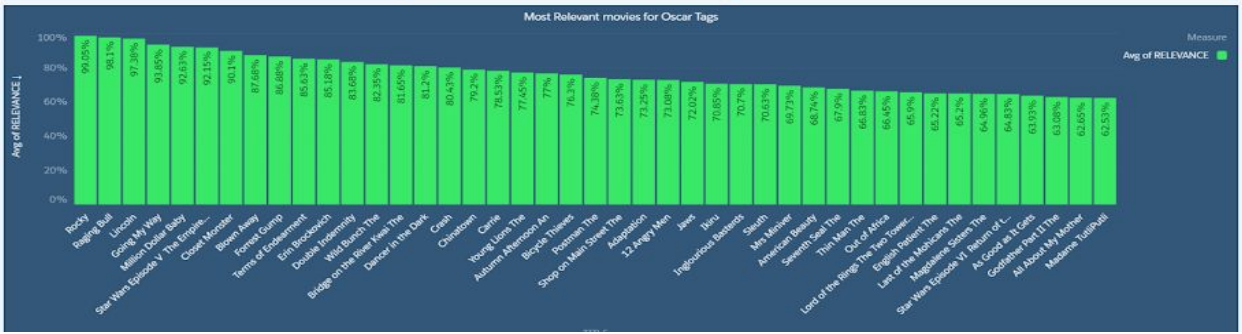

Most Relevant Genres for Oscar tags
GENRE
Action|Adventure|Western 0.61
Action|Romance|War|Western 0.65
Adventure|Drama|War 0.82
Animation|Fantasy 0.63
Comedy|Drama|Musical 0.63
Comedy|Drama|Romance|War 0.87
Comedy|Mystery|Thriller 0.64
Crime|Drama|Film-Noir 0.68
Drama|Fantasy|Horror|Thriller 0.79


Ratings of Action Movies over the years
AVERAGE RATING
YEAR OF RATING
2001 2.7
2002 3.6
2003 3.4
2004 2.9
2005 3.2
2006 3
2007 2.1
2008 3.6
2009 3.4
2010 2.7
2011 2.9
2012 3.7
2013 3.8
2014 3.8
2015 3.4
2016 3.2
2017 3.2
2018 2.9
2019 3.2


Average Ratings of the drama genre across all years
Measure
Avg of AVG_RATING
AVERAGE RATINGS
RELEASE YEAR


Ratings for late 80s and early 90s movies in recent years
Avg Rating
Release Year -> Year of Rating
1999
2000
2002
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019

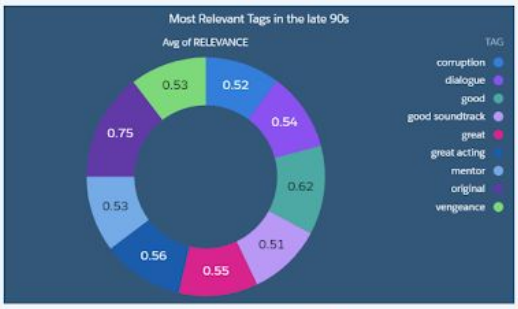

Chang of Rating for every genre from 1996-2019
GENRES
Action
Adventure
Animation
Children
Comedy
Crime
Drama
Fantasy
Film-Noir
Horror
IMAX
Musical
Mystery
Romance
Sci-Fi
War
Western
YEAR OF RATING
1996
1998
2000
2002
2005
2008
2010
2013
2016
2019
Avg of AVG_RATING

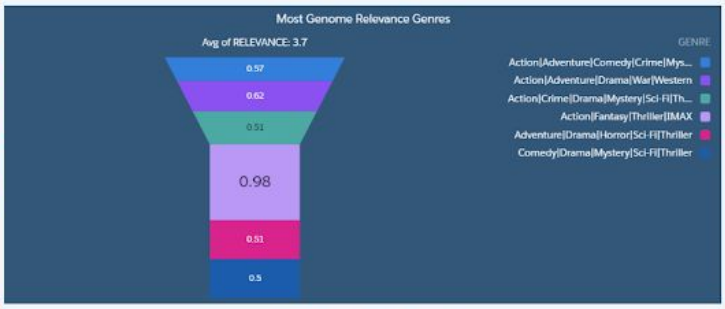

Most Relevant movies for Oscar Tags
Measure
Avg of RELEVANCE
Avg of RELEVANCE
Rocky 99.05%
Raging Bull 98.1%
Lincoln 97.38%
Going My Way 93.85%
Million Dollar Baby 92.63%
Star Wars Episode V: The Empire... 92.15%
Closet Monster 90.1%
Blown Away 87.66%
Forrest Gump 86.88%
Terms of Endearment 85.63%
Eric Brockbosch 85.18%
Double Indemnity 83.68%
Wild Bunch The 82.35%
Bridge on the River Kwai The 81.65%
Dancer in the Dark 81.2%
Crash 80.43%
Chinatown 79.2%
Carrie 78.53%
Young Lions The 77.45%
Autumn Afternoon An 77%
Bicycle Thieves 76.3%
Postman The 74.38%
Shop on Main Street The 73.63%
Adaptation 73.25%
12 Angry Men 73.08%
Jaws 72.02%
Ikiru 70.85%
Inglourious Basterds 70.7%
Sleuth 70.63%
Mrs Miniver 69.73%
American Beauty 68.74%
Seventh Seal The 67.9%
Thin Man The 66.83%
Out of Africa 66.45%
Lord of the Rings The Two Tower... 65.9%
English Patient The 65.22%
Last of the Mohicans The 65.2%
Magdalene Sisters The 64.96%
Star Wars Episode VI: Return of t... 64.83%
As Good as It Gets 63.93%
Godfather Part II The 63.08%
All About My Mother 62.65%
Madame Tutli-Putli 62.53%

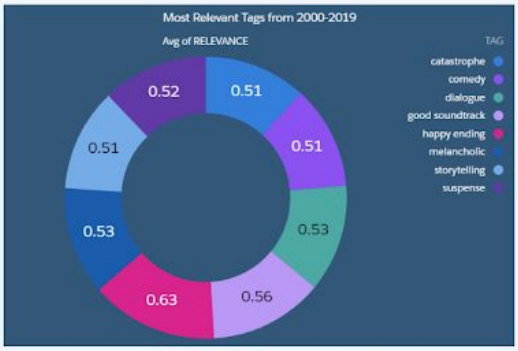
Recommendations

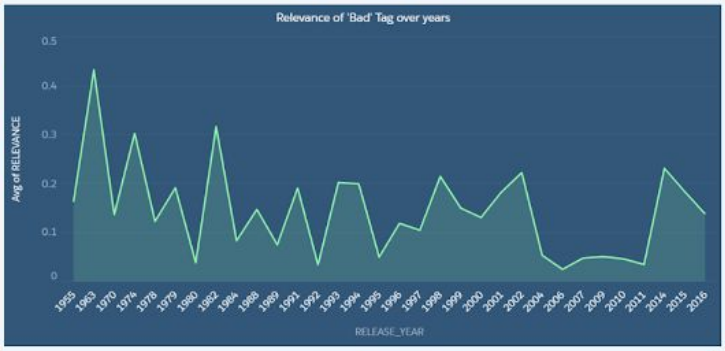
We created these below graphs to get insights-

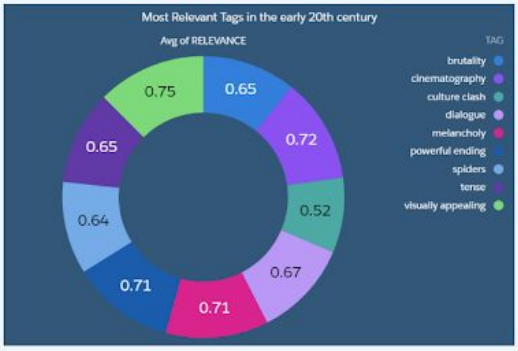
1. **Most Relevant Tags in the 20th century vs 90s vs recent years** - These three graphs combined together can be used to get insights on the type of movies that are being made across all the years
2. **Movie preferences for every genre** - Well rated movies for every genre based on ratings and tags
3. **Study the relevance of 'Bad' Tags over the years -** To get insights on how relevant the tags related to the word 'bad' for all the movies across all years
4. Further we can use collaborative filtering or association rules to recommend movies based on these insights.

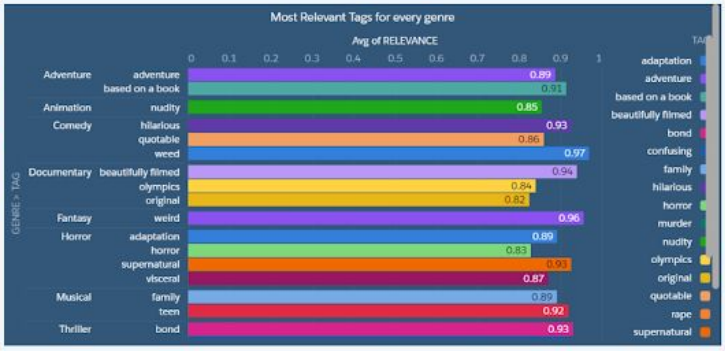

Most Relevant Tags in the late 90s
Avg of RELEVANCE
0.52
0.54
0.62
0.51
0.55
0.56
0.53
0.75
0.53
TAG
corruption
dialogue
good
good soundtrack
great
great acting
mentor
original
vengeance

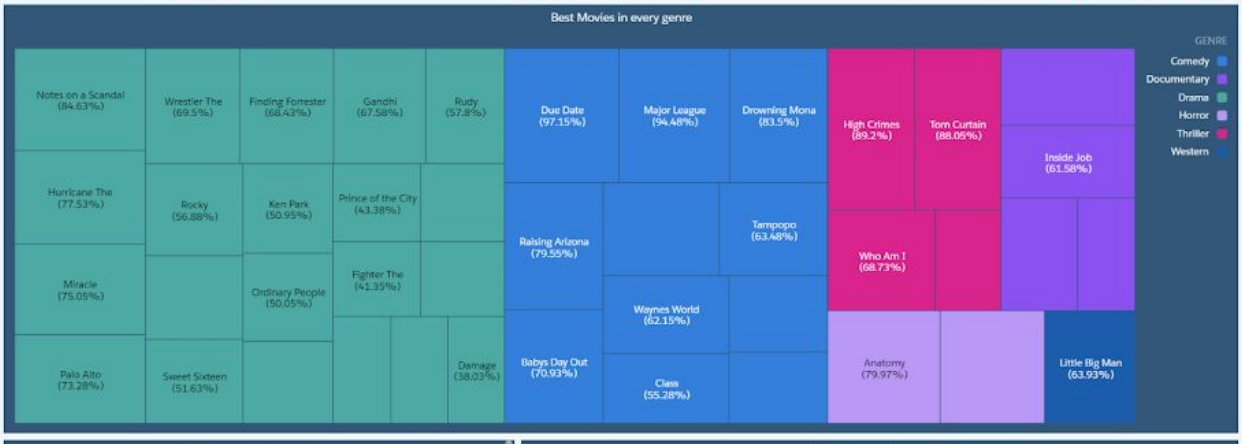

Most Genome Relevance Genres
Avg of RELEVANCE: 3.7
0.57
0.62
0.51
0.98
0.51
0.5
GENRE
Action|Adventure|Comedy|Crime|Mys...
Action|Adventure|Drama|War|Western
Action|Crime|Drama|Mystery|Sci-Fi|Th...
Action|Fantasy|Thriller|IMAX
Adventure|Drama|Horror|Sci-Fi|Thriller
Comedy|Drama|Mystery|Sci-Fi|Thriller

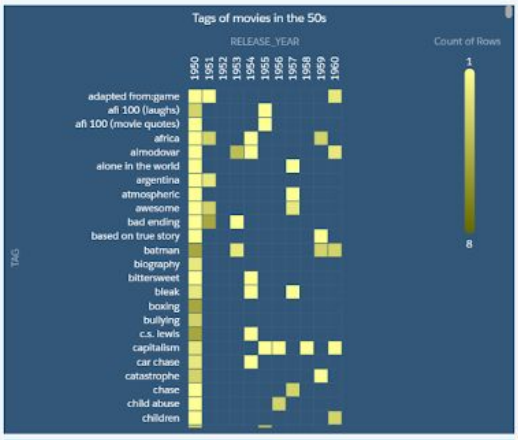

Most Relevant Tags from 2000-2019
Avg of RELEVANCE
0.51
0.51
0.53
0.56
0.63
0.53
0.51
0.52
TAG
catastrophe
comedy
dialogue
good soundtrack
happy ending
melancholic
storytelling
suspense

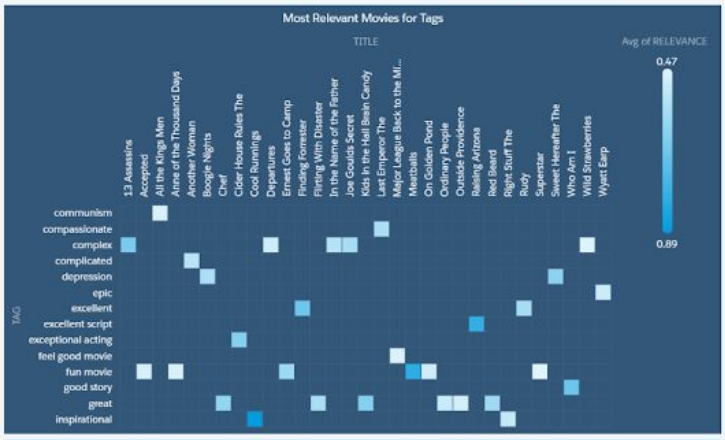

Relevance of 'Bad' Tag over years
Avg of RELEVANCE
0.5
0.4
0.3
0.2
0.1
0
1955 1963 1970 1974 1978 1979 1980 1982 1984 1988 1989 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2004 2006 2007 2009 2010 2011 2014 2015 2016
RELEASE_YEAR

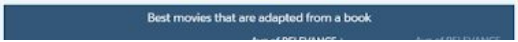

Most Relevant Tags in the early 20th century
Avg of RELEVANCE
0.65
0.72
0.52
0.67
0.71
0.71
0.64
0.65
0.75
TAG
brutality
cinematography
culture clash
dialogue
melancholy
powerful ending
spiders
tense
visually appealing


Most Relevant Tags for every genre
Avg of RELEVANCE
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
GENRE > TAG
Adventure
adventure 0.89
based on a book 0.91
Animation
nudity 0.85
Comedy
hilarious 0.93
quotable 0.86
weed 0.97
Documentary
beautifully filmed 0.94
olympics 0.84
original 0.82
Fantasy
weird 0.96
Horror
adaptation 0.89
horror 0.83
supernatural 0.93
visceral 0.87
Musical
family 0.89
teen 0.92
Thriller
bond 0.93
TAG
adaptation
adventure
based on a book
beautifully filmed
bond
confusing
family
hilarious
horror
murder
nudity
olympics
original
quotable
rape
supernatural


Best Movies in every genre
Notes on a Scandal (84.63%)
Wrestler The (69.5%)
Finding Forrester (68.43%)
Gandhi (67.58%)
Rudy (57.8%)
Due Date (97.15%)
Major League (94.48%)
Drowning Mona (83.5%)
High Crimes (89.2%)
Torn Curtain (88.05%)
Inside Job (61.58%)
Hurricane The (77.53%)
Rocky (56.88%)
Ken Park (50.95%)
Prince of the City (43.38%)
Raising Arizona (79.55%)
Tampopo (63.48%)
Who Am I (68.73%)
Miracle (75.05%)
Ordinary People (50.05%)
Fighter The (41.35%)
Waynes World (62.15%)
Palo Alto (73.28%)
Sweet Sixteen (51.63%)
Damage (38.03%)
Babys Day Out (70.93%)
Class (55.28%)
Anatomy (79.97%)
Little Big Man (63.93%)
GENRE
Comedy
Documentary
Drama
Horror
Thriller
Western


Tags of movies in the 50s
RELEASE_YEAR
1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960
Count of Rows
1
8
TAG
adapted from:game
afi 100 (laughs)
afi 100 (movie quotes)
africa
almodovar
alone in the world
argentina
atmospheric
awesome
bad ending
based on true story
batman
biography
bittersweet
bleak
boxing
bullying
c.s. lewis
capitalism
car chase
catastrophe
chase
child abuse
children


Most Relevant Movies for Tags
TITLE
13 Assassins
Accepted
All the Kings Men
Anne of the Thousand Days
Another Woman
Boogie Nights
Chef
Cider House Rules The
Cool Runnings
Departures
Ernest Goes to Camp
Finding Forrester
Flirting With Disaster
In the Name of the Father
Joe Goulds Secret
Kids in the Hall Brain Candy
Last Emperor The
Major League Back to the Mi...
Meatballs
On Golden Pond
Ordinary People
Outside Providence
Raising Arizona
Red Beard
Right Stuff The
Ruby
Superstar
Sweet Hereafter The
Who Am I
Wild Strawberries
Wyatt Earp
communism
compassionate
complex
complicated
depression
epic
excellent
excellent script
exceptional acting
feel good movie
fun movie
good story
great
inspirational
TAG
Avg of RELEVANCE
0.47
0.89


Best movies that are adapted from a book


Best rated movies in the early 20th century

Insights in all genres, ratings and tags

Metrics:

Movie Rating

Relevance scores for tags

1. **Rating Fluctuations & Effect on movie releases**- Even though the Average Ratings for Film Noir and Documentary is higher compared to other genres, more movies are being released in Drama and Comedy genre compared to the rest of the genres.
2. **Tracking the genres of movies being released-** In the 19th century, Comedy seems to be the most popular genre, however since the 90s there has been a significant surge in the genre Drama followed by comedy which is not far behind
3. **Oscar Tag vs Genre-** Majority of the Oscar tags are associated with the genre which is a mix of Comedy, Drama, Romance and War genre followed by Adventure, Drama and War. Almost all the subcategories of the genres in this category have war associated with them.
4. **Tracking ratings of the genre Drama-** Since the overall drama genre has the maximum number of movies released, we decided to analyse its ratings across all years. The ratings of drama movies has been very consistent throughout though its facing a downfall in the recent years compared to the early days
5. **Ratings of old movies in the recent years-** The movies released in 1900 to 1903 seem to have good average ratings compared to the previous years. However, 80s movies don't have great ratings in recent years.
6. **Change of Ratings for every genre in the recent years-** Even though Adventure movies had good ratings in the late 90s, the ratings for the same movies seem to go down in the recent years which shows that the change in the kind of movies preferred by the public. Likewise, even though Film-Noir was not much appreciated in the late 90s, they seem to have gathered attention in the recent years.
7. **Most Relevant Tags in the early 18th century, 90s and recent years-** In the early 20th century, 'Melancholy' tag seems to the most associated with the movies and in the late 90s, tags like 'original', 'corruption' and 'vengeance' seems to be more related whereas in the years of 2000-2019 'powerful ending', 'brutality', 'visually appealing' have better ratings.
8. **Most Relevant Tags for each Genre**- For the adventures Genre, the most relevant tag is 'based on a book' and for thriller 'bond' seems to have gained popularity the most
9. **Recommending best movies in every popular genre** - Filtering based on the tags focusing on good scripts, good direction and the tags relating to it and the genre, and by also filtering out the movies with the maximum relevance for the above tags, we recommend movies.

10. **Recommending movies based on appreciated tags** - We created a heat map based on the tags that focuses on good script, acting etc and movies that have a high relevance ratings with the above tags
11. **Recommending old movies** - All the best rated movies in the 20th century mainly come from three genres which are Drama, Musical, Comedy and Film-Noir among which Drama takes a huge percentage of the movies. Based on high relevance of tags that mean well, we can recommend movies.

Strength and Weaknesses of Wrangling tools

XSV

Xsv is a program for indexing, slicing, evaluating, separating and joining CSV files on the command line.

- In contrast to other resources, it is much quicker.
- But there is only a small number of operations in terms of Wrangling that can be done as of now as per the source github.
- It is mainly used for handling 10s of GBs of very wide csv files.

Trifacta

For data discovery and self-service data preparation for study, Trifacta develops data wrangling tools.

- Trifacta Wrangler is a connected framework for downstream analytics and visualization to transform knowledge.
- It enables analyst teams in an enterprise to explore and turn data with centralized security, governance and operationalization management using self-service.
- Particularly for individuals who do not like coding and want to prepare their data for forecasts / modeling, it is an easy to use wrangling tool.
- It can handle even some complicated operations such as data imputation, quickly creating new columns from existing columns.
- It also provides you with a facility for using different cloud platforms to export your data.

Answered Questions related to Dashboards

Which columns are dimensions, which columns are measures?

Dimensions -

1. MovieId
2. Title
3. Genre
4. Tag
5. Release Year
6. Year of Rating

Measure

1. Average Rating
2. Relevance

How would you generate new dimensions? What will you do to summarize measures?

1. We can create new columns/dimensions in wrangling tools mentioned above as well as Salesforce.
2. SAQL is used to create new dimensions from existing dimensions

Who would use this dashboard?

1. Business Analyst
2. Data Scientist
3. Data Analyst
4. Business Intelligence
5. Data Scientist - Recommendations

What value would be generated using this dashboard ?

Insights can be generated that could be useful for recommending movies to the user and also the kind of movies that are doing well in the current market. This dashboard can also be used to track how the customers response changes periodically over the years.