



Introduction to Data Management

(Project Semester August-December 2021)

Project Report

On

Analyzing MovieLens Website Data

Submitted By

UTKARSH SHAHI

11911032

Bachelor of Technology (CSE)

Section - KM006

Course Code- INT217

Under the Guidance of

Ms. Komal Arora: 17783

Assistant Professor(LPU)

Lovely Professional University

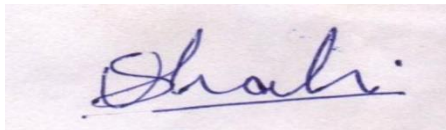
Phagwara, Punjab

Student Declaration

To whomsoever it may concern

I, **UTKARSH SHAHI, 11911032**, hereby declare that the work done by me on “**Analyzing MovieLens Website Data**” from August to December 2021, is a record of original work for the partial fulfilment of the requirements for the award of the degree, BACHELOR OF TECHNOLOGY (CSE).

Utkarsh Shahi (11911032)

A handwritten signature in blue ink, appearing to read 'Shahi', is written over a horizontal line.

Dated: December 17, 2021

Acknowledgement

I would like to express my gratitude towards my University as well as my mentor **Ms. Komal Arora** for providing me the golden opportunity to do this wonderful project regarding Data Science, which also helped me in doing a lot of homework and learning. As a result, I came to know about so many new things. So, I am really thanking to them.

Nevertheless, I express my gratitude toward my families and colleagues for their kind co-operation and encouragement which help us in completion of this project.

Utkarsh Shahi

11911032

Contents

Sr. No.	Topic	Page No.
1	Introduction	5
2	Objectives	6
3	About Dataset	7
4	ETL Process	10
5	Power Pivot	14
6	Analysis	15
7	Dashboard	20
8	Bibliography	21

INTRODUCTION

Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions. The data used for analysis can come from many different sources and presented in various formats.

Data increasingly is seen as a corporate asset that can be used to make more-informed business decisions, improve marketing campaigns, optimize business operations and reduce costs, all with the goal of increasing revenue and profits.

Data analysis is the process of collecting, modeling, and analyzing data to extract insights that support decision-making.

- This Project is based on Data Analysis of MovieLens Website.
- MovieLens is **a web-based recommender system and virtual community that recommends movies for its users to watch**, based on their film preferences using collaborative filtering of members' movie ratings and movie reviews. It contains about 11 million ratings for about 8500 movies. This MovieLens Dataset contains 6 worksheets having 3, 2, 3, 4, 3 and 4 data fields.
- Having 1M+ rows

Objective/Scope of Analysis

On observing the data set, I come to answer the following objectives:

- ✓ Tag wise - Most Viewed Movies
- ✓ Movies with Top Ratings
- ✓ Most Active User (in terms of watching)
- ✓ Customizing movies with respect to Genre
- ✓ Ordering movies on the basic of Relevance score

About Data Set

Source: <https://www.kaggle.com/grouplens/movielens-20m-dataset>

This dataset contain information related to Millions of movies on MovieLens website.

The columns included in the dataset are given below:

❖ movieId: Unique number assigned to every movie

❖ tagId: No. assigned to every tag

❖ relevance: relevance score of movies

❖ tag: tag name for every tagId

❖ imdbId

❖ tmdbId

❖ userId: unique no. for every user

❖ rating: rating for every single movie

- ❖ timestamp: occurrence date and time for the particular movie
- ❖ title: movie title
- ❖ Genres: specific type

Dataset

The screenshot shows an Excel spreadsheet with the following data:

movieid	title	genres
1	Toy Story (1995)	Adventure Animation Children Comedy Fantasy
2	Jumanji (1995)	Adventure Children Fantasy
3	Grumpier Old Men (1995)	Comedy Romance
4	Waiting to Exhale (1995)	Comedy Drama Romance
5	Father of the Bride Part II (1995)	Comedy
6	Heat (1995)	Action Crime Thriller
7	Sabrina (1995)	Comedy Romance
8	Tom and Huck (1995)	Adventure Children
9	Sudden Death (1995)	Action
10	GoldenEye (1995)	Action Adventure Thriller
11	American President, The (1995)	Comedy Drama Romance
12	Dracula: Dead and Loving It (1995)	Comedy Horror
13	Balto (1995)	Adventure Animation Children
14	Nixon (1995)	Drama
15	Cutthroat Island (1995)	Action Adventure Romance
16	Casino (1995)	Crime Drama
17	Sense and Sensibility (1995)	Drama Romance
18	Four Rooms (1995)	Comedy
19	Ace Ventura: When Nature Calls (1995)	Comedy
20	Money Train (1995)	Action Comedy Crime Drama Thriller
21	Get Shorty (1995)	Comedy Crime Thriller
22	Copycat (1995)	Crime Drama Horror Mystery Thriller
23	Assassins (1995)	Action Crime Thriller
24	Powder (1995)	Drama Sci-Fi
25	Leaving Las Vegas (1995)	Drama Romance
26	Othello (1995)	Drama
27	Now and Then (1995)	Children Drama
28	Persuasion (1995)	Drama Romance
29	City of Lost Children, The (1995)	Adventure Drama Fantasy Mystery Sci-Fi
30	Shanghai Triad (Yao a yao yao) (1995)	Crime Drama
31	Dangerous Minds (1995)	Drama

The screenshot shows an Excel spreadsheet with the following data:

tagid	tag
1	1
2	2 007 (series)
3	3 18th century
4	4 1920s
5	5 1930s
6	6 1950s
7	7 1960s
8	8 1970s
9	9 1980s
10	10 19th century
11	11 3d
12	12 70mm
13	13 80s
14	14 09-Nov
15	15 aardman
16	16 aardman studios
17	17 abortion
18	18 absurd
19	19 action
20	20 action packed
21	21 adaptation
22	22 adapted from:book
23	23 adapted from:comic
24	24 adapted from:game
25	25 addiction
26	26 adolescence
27	27 adoption
28	28 adultery
29	29 adventure
30	30 affectionate
31	31 afi 100

ETL Process

ETL is a process that extracts the data from different source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system. Full form of ETL is Extract, Transform and Load.

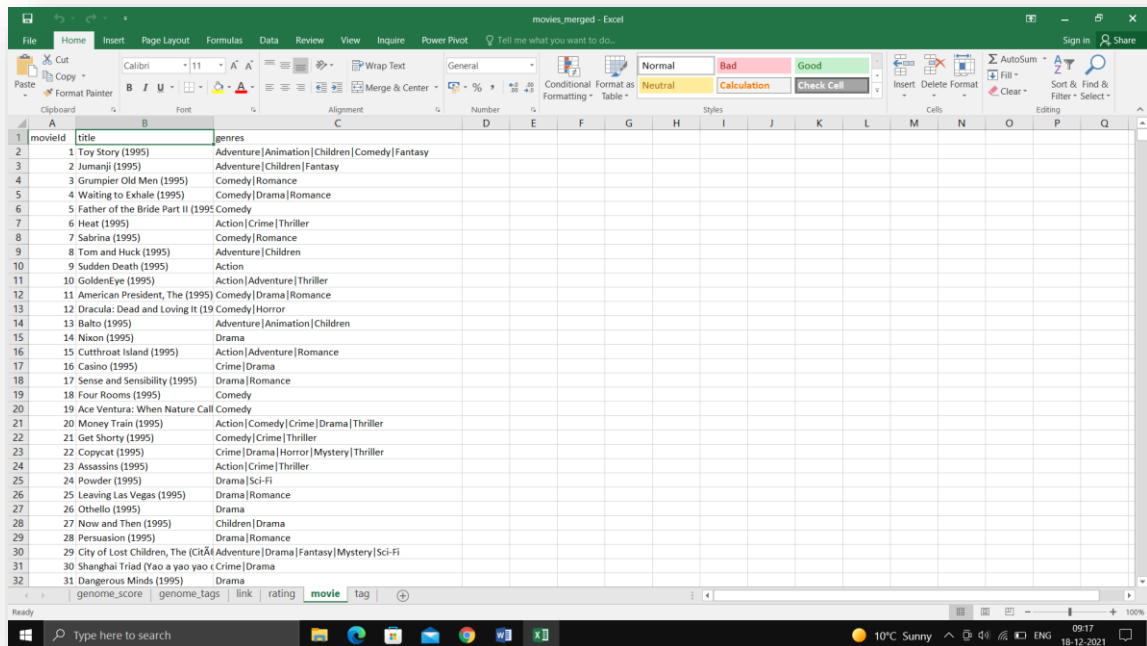
ETL process can also use the pipelining concept i.e. as soon as some data is extracted, it can be transformed and during that period some new data can be extracted. And while the transformed data is being loaded into the data warehouse, the already extracted data can be transformed.

- ❖ Allows sample data comparison between source and target system.
- ❖ Helps to improve productivity as it codifies and reuses without additional technical skills.
- ❖ Allows sample data comparison between source and target system.

The cleansing stage is crucial in a data warehouse technique because it is supposed to improve data quality. The primary data cleansing features found in ETL tools are rectification and homogenization. They use specific dictionaries to rectify typing mistakes and to recognize synonyms, as well as rule-based cleansing to enforce domain-specific rules and defines appropriate associations between values.

Steps taken to clean data through ETL

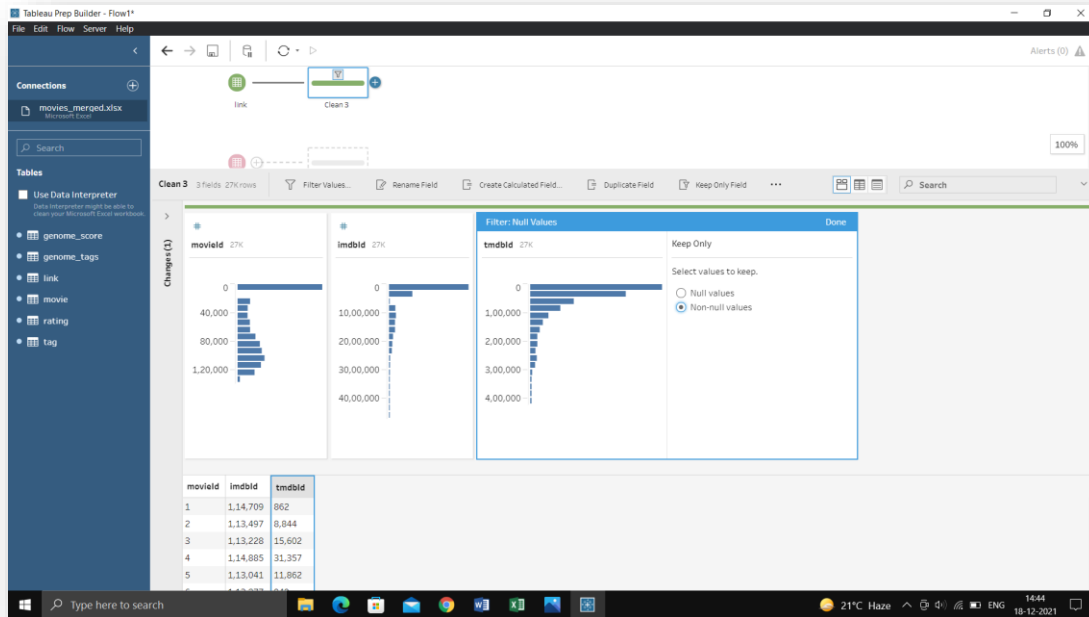
1. First off all open data set and understand carefully.



The screenshot shows an Excel spreadsheet with the following data:

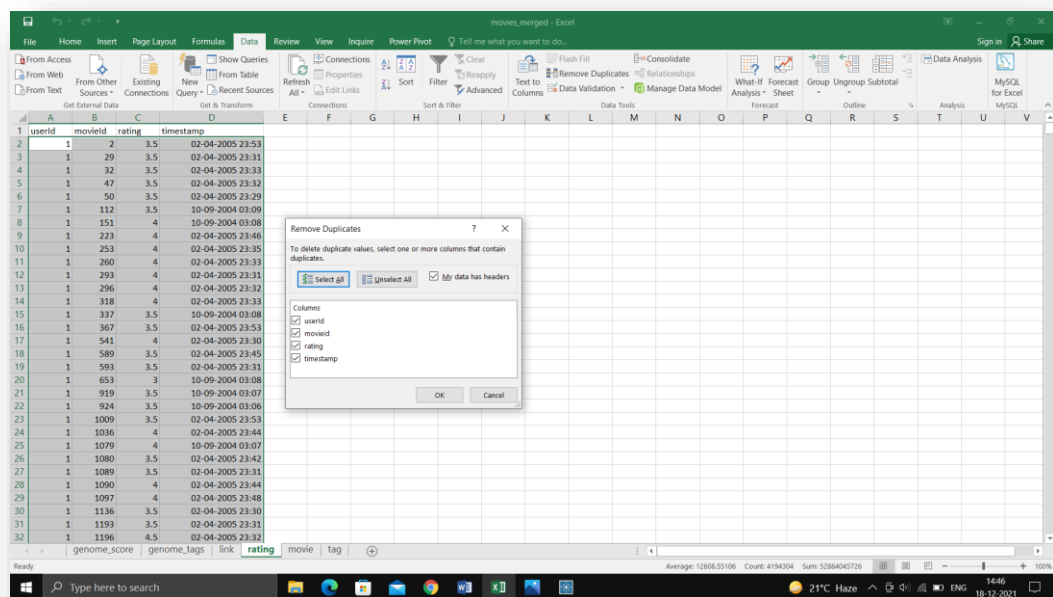
movieid	title	genres
1	Toy Story (1995)	Adventure Animation Children Comedy Fantasy
2	Jumanji (1995)	Adventure Children Fantasy
3	Grumpier Old Men (1995)	Comedy Romance
4	Waiting to Exhale (1995)	Comedy Drama Romance
5	Father of the Bride Part II (1995)	Comedy
6	Heat (1995)	Action Crime Thriller
7	Sabrina (1995)	Comedy Romance
8	Tom and Huck (1995)	Adventure Children
9	Sudden Death (1995)	Action
10	GoldenEye (1995)	Action Adventure Thriller
11	American President, The (1995)	Comedy Drama Romance
12	Dracula: Dead and Loving It (1995)	Comedy Horror
13	Balto (1995)	Adventure Animation Children
14	Nixon (1995)	Drama
15	Cutthroat Island (1995)	Action Adventure Romance
16	Casino (1995)	Crime Drama
17	Sense and Sensibility (1995)	Drama Romance
18	Four Rooms (1995)	Comedy
19	Ace Ventura: When Nature Calls (1995)	Comedy
20	Money Train (1995)	Action Comedy Crime Drama Thriller
21	Get Shorty (1995)	Comedy Crime Thriller
22	Copycat (1995)	Crime Drama Horror Mystery Thriller
23	Assassins (1995)	Action Crime Thriller
24	Powder (1995)	Drama Sci-Fi
25	Leaving Las Vegas (1995)	Drama Romance
26	Othello (1995)	Drama
27	Now and Then (1995)	Children Drama
28	Persuasion (1995)	Drama Romance
29	City of Lost Children, The (1995)	Adventure Drama Fantasy Mystery Sci-Fi
30	Shanghai Triad (Yao a yao yao) (1995)	Crime Drama
31	Dangerous Minds (1995)	Drama

2. Remove that null values completely from dataset
3. Cleaning step removing null values by Tableau Prep



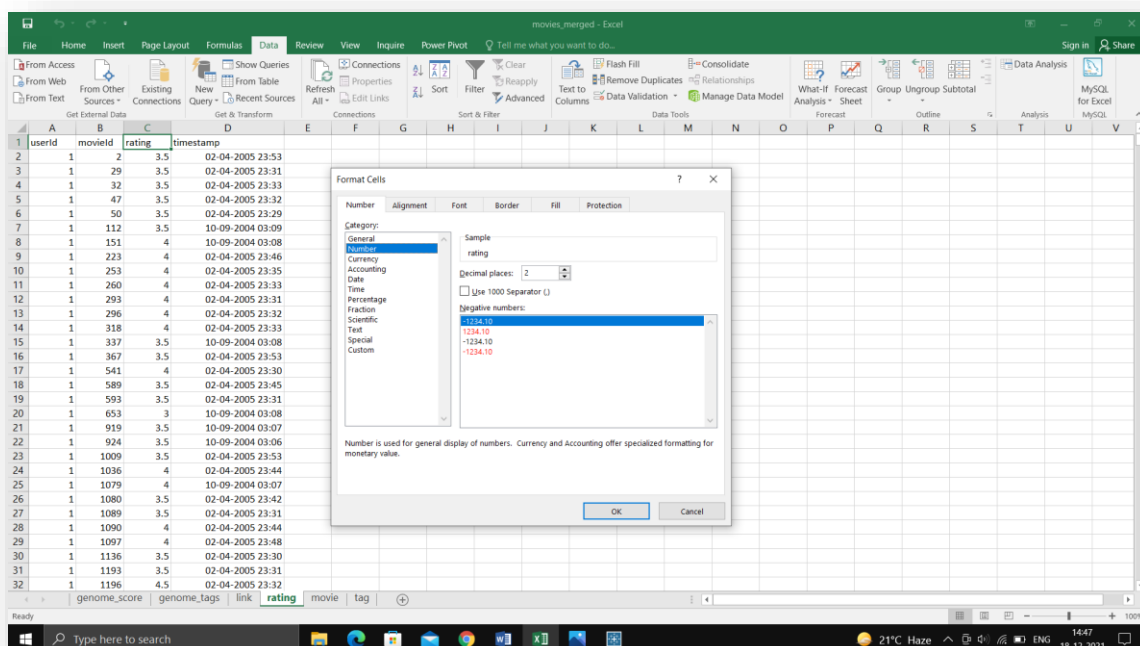
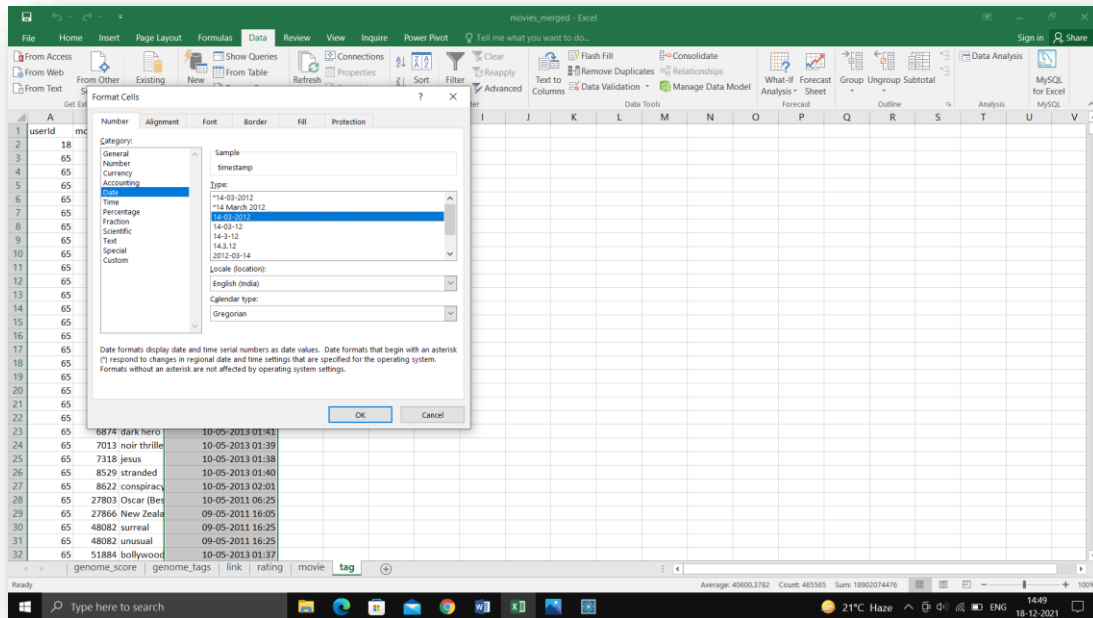
4. Now open excel and delete duplicate values

DATA -> remove duplicates



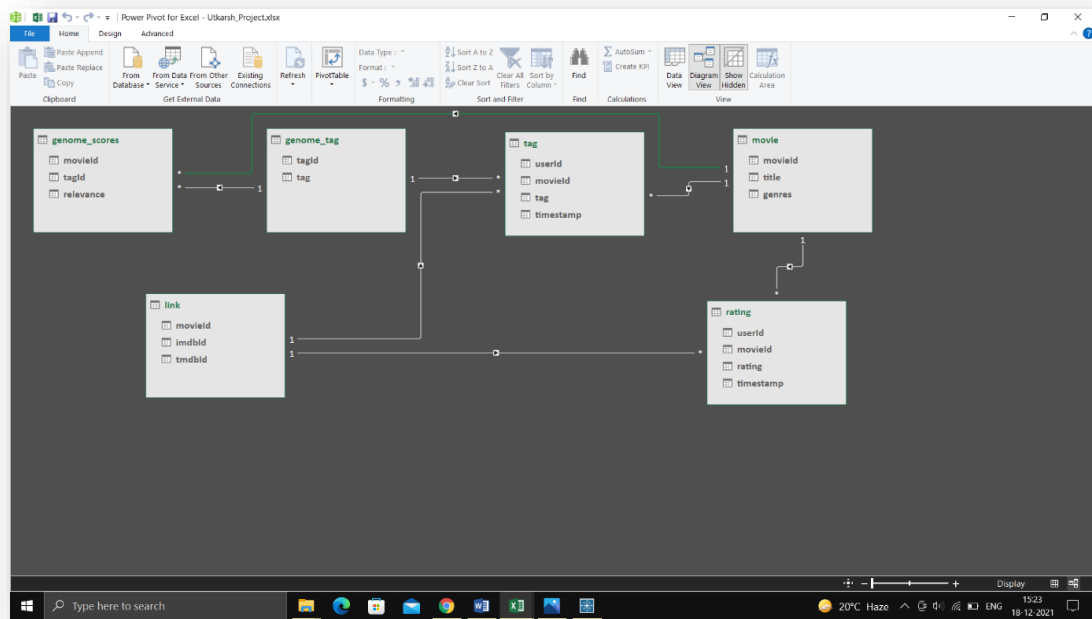
5. Change data type of data fields as required

INSTALLS(TEXT) -> final install(NUMBER), SIZE



POWER PIVOT

With the Help of this Connected all sheets to each other so that relations can be formed



Connections: -

1. tagId(genome_scores) -> tagId(genome_tags)
2. movieId(genome_scores) -> movieId(movie)
3. tag(genome_tags) -> tag(tag)
4. movieId(tag) -> movieId(movie)
5. movieId(movie) -> movieId(rating)
6. movieId(link) -> movieId(rating)
7. movieId(link) -> movieId(tag)

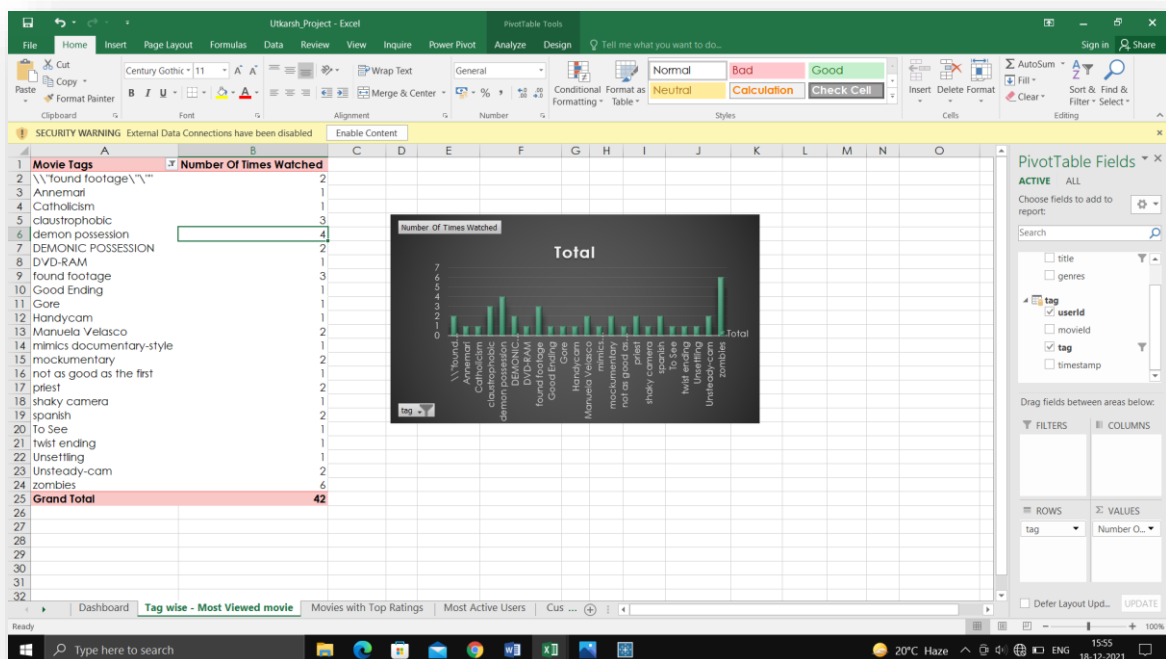
Analysis on Dataset

1. Tag wise – Most Viewed Movies

Description: By using tag along with userId we can analyze Tag Wise – Most Viewed Movies

- ❖ Create pivot table
- ❖ Put tag name in rows and userId in values
- ❖ Apply filter for top 10 apps
- ❖ Insert pivot chart

Sample result:

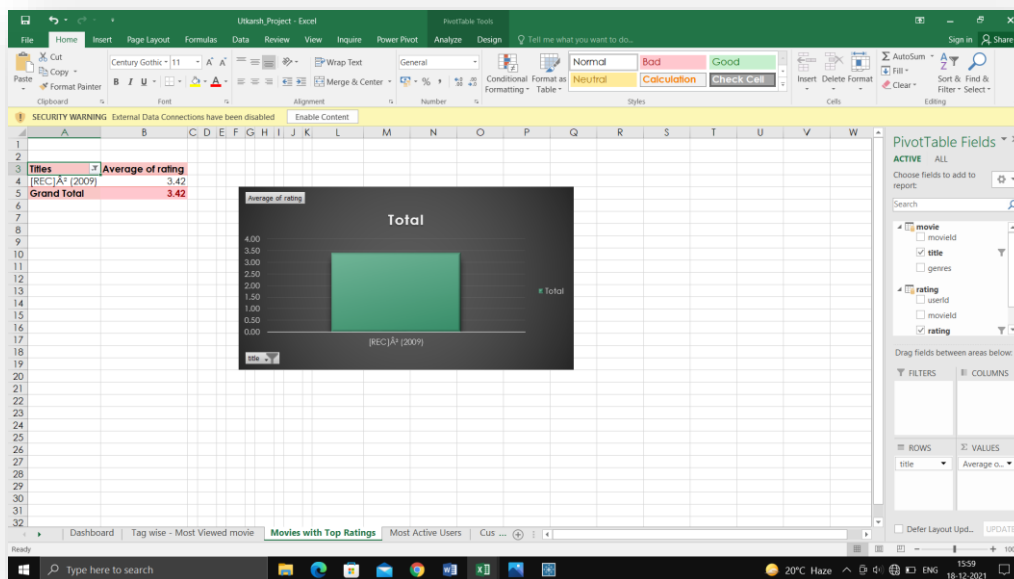


2. Movies with Top Ratings

Description: With the help of title and rating we can find their average rating

- ❖ Create pivot table
- ❖ title in rows and rating in values
- ❖ Value field setting of rating -> Average
- ❖ Insert pivot chart

Sample result:

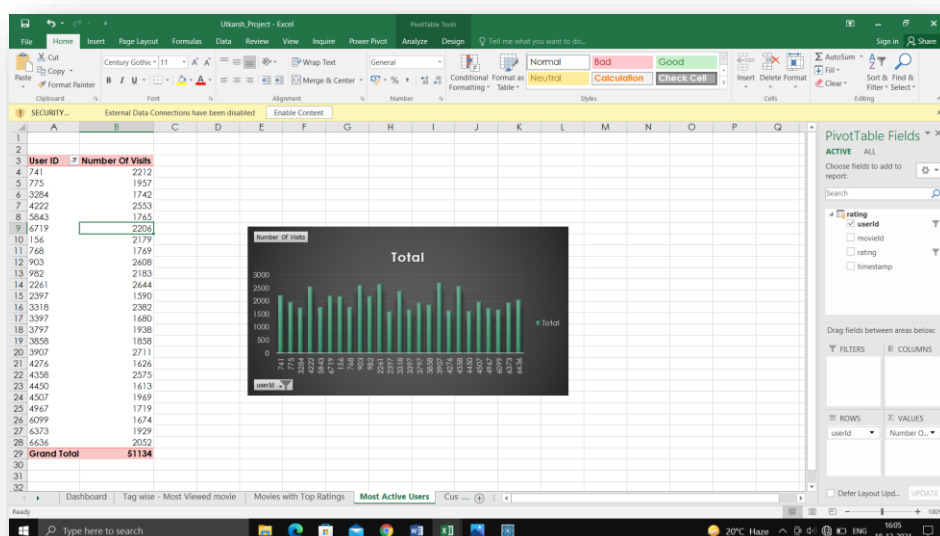


3. Most Active Users

Description: For this analysis, we consider userId in rows and put userId in values also

- ❖ Create pivot table
- ❖ userId in rows and userId in values
- ❖ Value field setting of userId(values) -> Count
- ❖ apply filter for top 25
- ❖ Insert pivot chart

Sample Results:

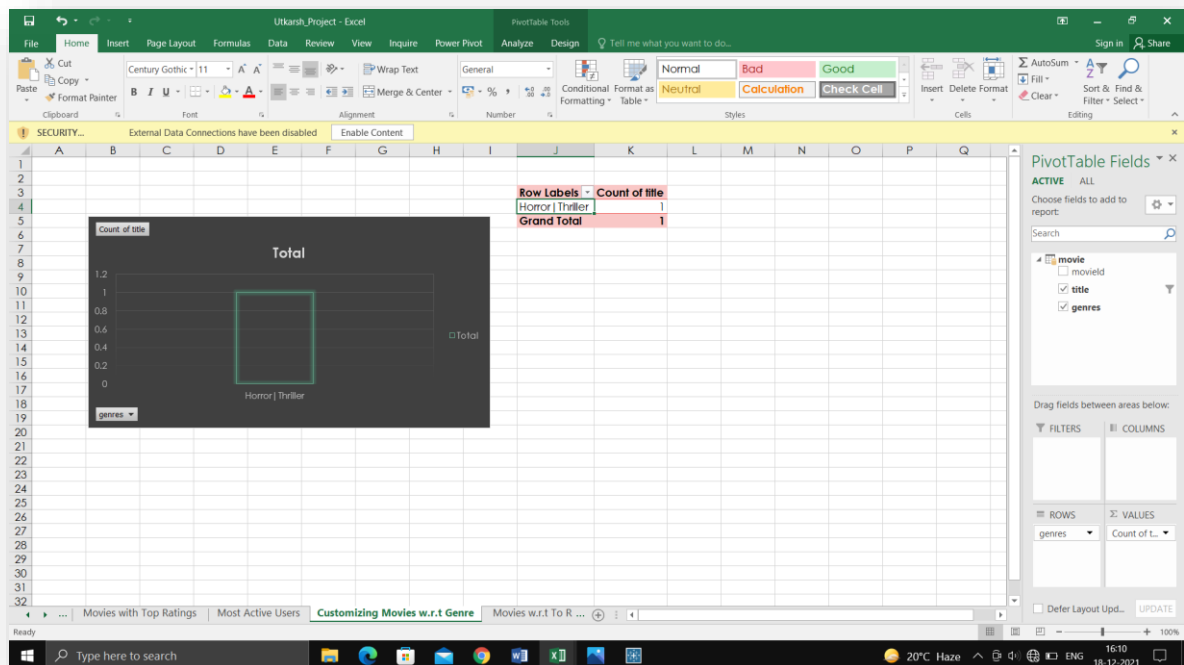


4. Customizing Movies with respect to Genre

Description: Taking help of genre and putting in rows and title in values and applying for Count in Value field setting

- ❖ Create pivot table
- ❖ genre in rows and title in values
- ❖ Value field setting choose Count
- ❖ Insert pivot chart

Sample Results:

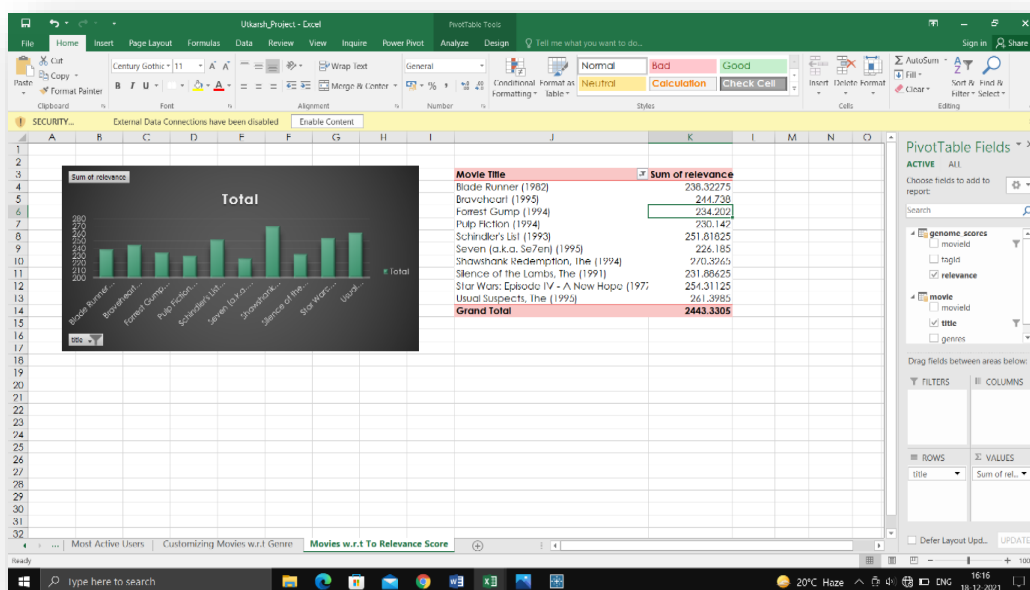


5. Movies with respect to Relevance Score

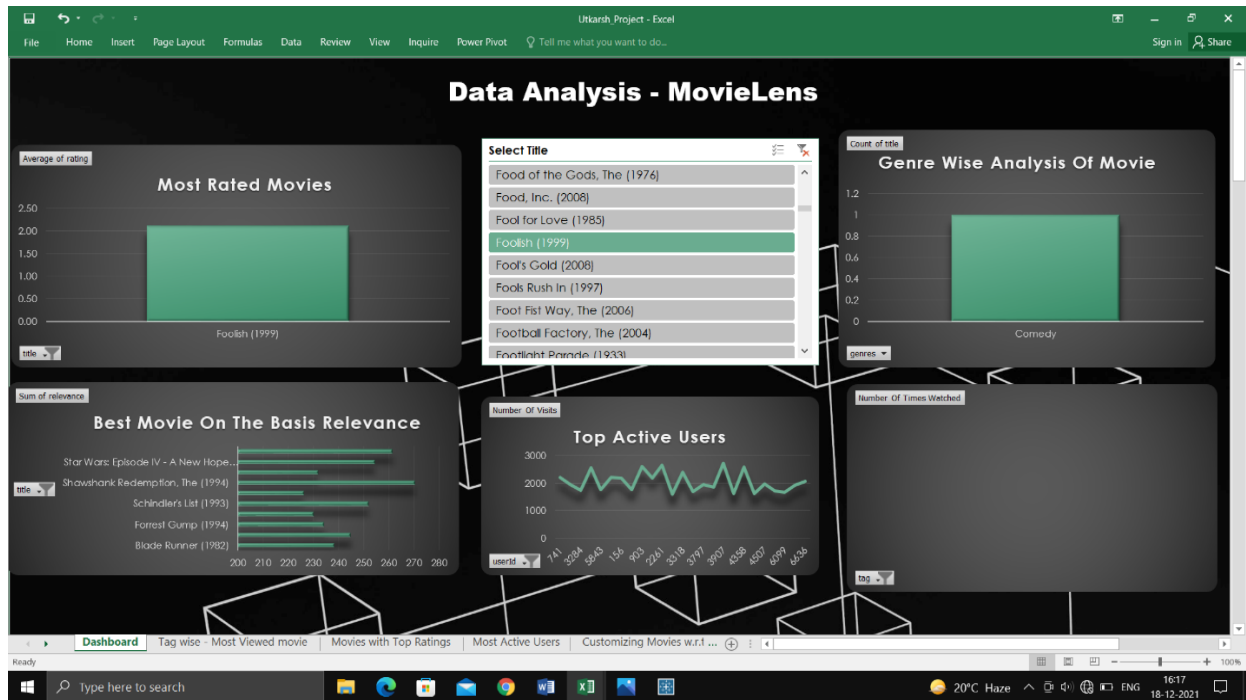
Description: This is relevance score assigned to all movies by analyzing their positive and negative feedback. Here we had organized Movies titles on the basic of their Relevance Score.

- ❖ Create pivot table
- ❖ title in rows and relevance in values
- ❖ apply top 10 filter
- ❖ Value field setting choose Sum
- ❖ Insert pivot chart

Sample Result:



Final Dashboard



Bibliography

❖ Data set Source:

<https://www.kaggle.com/grouplens/movielens-20m-dataset>

❖ Background Image:

Unsplash.com

❖ LPU PPT of INT217

❖ Some online Websites