Individual Assignment (35%)

Big Data Processing – Movie Ratings

Objectives:

1. Understand the process and requirements of loading Big Data.
2. Processing Big Data using Databricks.
3. Create visualizations of Big Data.

Instructions:

1. This is an **individual** assignment.
2. You can only use **PySpark/Pandas DataFrame** for the assignment. The use of **SQL statements** will **not** be accepted.
3. Do note that if you are found to commit **plagiarism**, you will receive **zero** for your assignment and may results in failing the module.
4. Module Assessment Weightage: **35%**
5. **Marking rubrics** for the **conclusions** are found on the last page of this document.

Deliverables:

1. **IT2312 2025 S2 Assignment.docx**
    - **Answers** to the **conclusion question** written in the box provided.
2. **DataBricks DBC (*.dbc) Archive File**
    - **Codes** (**not** SQL statements) that produce the solutions and visualizations for both **Part 1 and Part 2.**
3. **Video recording of you presenting the answers and conclusion for Part 1 and Part 2.**

Background

You are a data scientist looking working for a film production company. You have been tasked to, "Identify new movie genres to target and produce," and "Identify examples and elements of bad movies to avoid replicating." Your data mining goals include:

1. What are the most popular movies among the uncommon genres.
2. What are the worst and best movies based on average rating.
3. What are the top tags that describe the worst movies.

The dataset used in this assignment (ml-25m) describes 5-star rating and free-text tagging activity from MovieLens, a movie recommendation service. It contains 25,000,095 ratings and 1,093,360 tag applications across 62,423 movies. These data were created by 162,541 users between January 09, 1995 and November 21, 2019. This dataset was generated on November 21, 2019.

Users were selected at random for inclusion. All selected users had rated at least 20 movies. No demographic information is included. Each user is represented by an id, and no other information is provided.

The data are contained in the files links.csv, movies.csv, ratings.csv and tags.csv. **Note** that we will be using only **movies.csv**, **tags.csv** and **ratings.csv** files for this assignment. More details about the contents and use of all these files can be found here: https://grouplens.org/datasets/movielens/

Part 1 – Data Ingestion (10 marks)

1. Ingest the 3 files **movies.csv**, **tags.csv** and **ratings.csv** into **DataBricks**.

(5 marks)

2. Print the number of records and the number of columns for each data file.

(5 marks)

Part 2 – Data Exploration (80 marks)

For the data exploration tasks below, please use a **DataBricks Notebook** to document what you did to create the required visualization or tables.

1. Create a DataFrame showing the list of **unique tags** and the **number of occurrences** for each tag for tags that are **NOT** in the following list: **['sci-fi', 'action', 'comedy', 'mystery', 'war', 'politics', 'religion', 'thriller']**. Sort the list in **descending order** by the number of tag occurrences. An example is shown below.

(10 marks)

	tag	cnt
1	atmospheric	6516
2	surreal	5326
3	based on a book	5079
4	twist ending	4820
5	funny	4738
6	visually appealing	4526
7	dystopia	4257
8	dark comedy	4026

2. Create a DataFrame that contains the **movies** with the tags of **'boring'** or **'overrated'**. The view or DataFrame should also show the **title** of the movie and the **average rating** for the movie. Sort the list by **average rating** in **ascending order** and display the **top 10 rows**.

(10 marks)

'boring' or 'overrated'

	movieId	title	avgRating
1	138120	The Expedition	0.5
2	169018	Water Boyy (2015)	0.5
3	61348	Disaster Movie (2008)	1.2055655296229804
4	160012	Andron (2016)	1.4545454545454546
5	186497	When Do We Eat (2005)	1.5
6	182569	Arson Mom (2014)	1.5
7	167616	The Aftermath (1982)	1.625
8	165667	Somnus (2016)	1.6666666666666667

3. Create a DataFrame that contains the movies with the tags of **'great acting'** or **'inspirational'**. The DataFrame should also show the **title** of the movie and the **average rating** for the movie. Sort the list by **average rating** in **descending order** and display the **top 10 rows**.

(10 marks)

'great acting' or 'inspirational'

	movieId	title	avgRating
1	180121	The Life-Changing Magic of Tidying Up (2013)	4.75
2	207640	Vision Portraits (2019)	4.5
3	318	Shawshank Redemption, The (1994)	4.413576004516335
4	858	Godfather, The (1972)	4.324336165187245
5	50	Usual Suspects, The (1995)	4.284353213163313
6	1221	Godfather: Part II, The (1974)	4.2617585117585115
7	1203	12 Angry Men (1957)	4.243014062405697
8	2959	Fight Club (1999)	4.228310618821568

4. Create a DataFrame that aggregates the movie **ratings** (r) into the **ranges** of: **'Below 1', '1 to 2', '2 to 3', '3 to 4'** and **'4 to 5'** into a separate column named **rating_range** where:

Below 1 – r < 1
1 to 2 – r >= 1 and r < 2
2 to 3 – r >= 2 and r < 3
3 to 4 – r >= 3 and r < 4
4 to 5 – r >= 4 and r < 5
5 and more - r >=5

Include the columns **userId**, **movieId**, **rating** and **tag** in the DataFrame.

(10 marks)

Table +

	userId	movieId	rating	tag
1	3	260	4	sci-fi
2	3	260	4	classic
3	4	1732	4.5	great dialogue
4	4	1732	4.5	dark comedy
5	4	7569	3.5	so bad it's good
6	4	44665	5	unreliable narrators
7	4	115569	5	tense
8	4	115713	5	tense
9	4	115713	5	philosophical
10	4	115713	5	artificial intelligence
11	4	148426	2	so bad it's good
12	19	7099	5	post-apocalyptic

5. Create a DataFrame that shows the aggregated movie **rating ranges** and their corresponding **tags** and **count** of the tags. Filter the view or table to show only tag counts that are m**ore than 200**. Sort the view or table by the **rating range** in **ascending order** and the **tag counts** in **descending order**.

(10 marks)

	rating_range	tag	numTag
1	1 to 2	boring	452
2	1 to 2	predictable	275
3	1 to 2	bad acting	228
4	1 to 2	stupid	211
5	2 to 3	boring	612
6	2 to 3	predictable	594
7	2 to 3	sci-fi	510
8	2 to 3	action	425

6. What **conclusions** can you draw from the data exploration you have performed above?

(30 marks)

Part 3 – Data Exploration (10 marks)

1. Record a **5 to 10-minute** presentation video presenting and explaining your answers and your conclusion and submit it on BrightSpace together with this document.

(10 marks)

Marking Rubrics for Conclusions

Inadequate (F)	Acceptable (D)	Satisfactory (C)	Good (B)	Excellent (A)
0 – 14 marks	**15 – 17.5 marks**	**18 – 20.5 marks**	**21 – 23.5 marks**	**24 – 30 marks**
Does not present a conclusion or provides irrelevant statements that do not address the business objective or problem. No logical closure is provided.	Provides a partial or unclear conclusion that only weakly addresses the business objective or problem. Shows limited understanding and lacks depth or meaningful suggestions.	Provides a basic (1 or 2) conclusion that addresses the business objective or problem, but with limited detail or insight. Suggestions may be general, obvious, or lack justification.	Provides a clear conclusion (2 to 3) that addresses the business objective or problem with some insights or recommendations. Shows good understanding with minor gaps in depth, clarity, or justification.	Delivers a clear, concise, and well-synthesised conclusion that directly addresses the business objective or problem. Demonstrates strong critical thinking and provides insightful, actionable recommendations. Shows coherence, strong justification, and strategic depth.

Marking Rubrics for Video Submission

Inadequate (F)	Acceptable (D)	Satisfactory (C)	Good (B)	Excellent (A)
0 – 4 marks	**5 – 5.5 marks**	**6 – 6.5 marks**	**7 – 7.5 marks**	**8 – 10 marks**
No key ideas; presentation shows no effort, organization, elaboration, and use of visuals and graphics. Presenter appears very uncomfortable; displays very poor use of eye contact, posture, gestures, and vocal expressiveness. No effort to engage audience and lack	Key ideas are unclear; presentation shows little effort, organization, elaboration, and use of visuals and graphics. Presenter appears uncomfortable; displays poor use of eye contact, posture, gestures, and vocal	Key ideas lack clarity; presentation shows some effort, organization, elaboration, and use of visuals and graphics. Presenter appears rather uncomfortable; displays limited use of eye contact, posture, gestures, and vocal	Key ideas presented clearly; presentation shows good effort, organization, elaboration, and use of visuals and graphics. Presenter appears comfortable; displays good use of eye contact, posture, gestures, and	Key ideas presented very clearly and convincingly; presentation shows excellent effort, organization, elaboration, and use of visuals and graphics. Presenter appears confident; displays effective use of eye contact, posture,

fluency and good pronunciation. Presented solution that is ambiguous and unorganized, and audience did not understand and could not follow.	expressiveness. Little effort to engage audience and lack fluency and good pronunciation. Presented solution that is not clear and concise, and audience has difficulty understanding and following.	expressiveness. Some effort to engage audience, fluency, and good pronunciation. Presented solution with limited clarity and conciseness; audience sometimes find it challenging to understand and follow.	vocal expressiveness. Good effort to engage audience, fluency, and pronunciation. Presented solution clearly and concisely for audience understanding.	gestures, and vocal expressiveness. Excellent effort to engage audience, fluency, and pronunciation. Presented solution clearly, concisely, and convincingly.