Hi Analyst!!, here is your stage 1 task.

Data Analytics Track – Stage 1 Task: Feature Engineering & Exploration.

Task explainer video

Objective

In this stage you will learn how to clean, explore, and perform basic feature engineering on a real-world dataset and generate insights that can later support recommendation systems or analytics.

Dataset: MovieLens Dataset

The dataset is a small portion of the MovieLens dataset (a dataset of movie ratings and metadata collected by the GroupLens research group) chosen for this task.

The main files in the zip file are the ratings.csv and movies.csv.

ratings.csv

userId → unique identifier for each user

movieId → unique identifier for each movie

rating → rating given by the user (usually 0.5 to 5.0 in 0.5 increments)

timestamp → when the rating was made

movies.csv

movieId → links to ratings

title → movie title (often includes release year, e.g., Toy Story (1995))

genres → one or more genres separated by | (e.g., Action|Adventure|Sci-Fi)

tags.csv

Description: User-generated tags for movies.

Columns:

userId: ID of the user who tagged the movie.

movieId: ID of the tagged movie.

tag: The tag text.

timestamp: When the tag was added.

links.csv

Description: Links movie IDs to external databases.

Columns:

movieId: MovieLens ID.

imdbId: IMDb ID.

tmdbId: The Movie Database (TMDb) ID.

Task Instructions:

Step 1: Data Preparation

Load the datasets and merge them on movieId.

Check for duplicates and missing values.

Convert timestamp into a proper datetime format.

Step 2: Feature Engineering

Create at least 6 new features. Some ideas include:

Extract the release year from the movie title (e.g., Toy Story (1995) → 1995).

Count the number of genres a movie has (e.g., Action|Comedy - 2).

These are just examples, you should add your own creative features too, and explain why they might be useful.

Step 3: Exploratory Data Analysis (EDA)

Using the original and new features, carry out exploratory analysis to better understand the dataset. Think along the lines of:

What do ratings look like overall? (e.g., do people tend to rate movies high?)

How do movies or genres differ from each other?

Highlight 6 meaningful insights. Simple plots (histograms, bar charts, scatter plots, etc.) are encouraged but not strictly prescribed.

Tool Requirement:

Make use of Python or R.

Deliverables:

A cleaned dataset with your new features.

A Jupyter Notebook or R Notebook Downloaded as PDF. (NOTE: Make sure you run all cells before downloading your notebook and take care of all errors).

A short report in PDF (6 pages max) explaining:

The features you created and why.

The key insights you discovered.

How these could support building a recommendation system in the future.

A few visualizations/tables that support your findings

Expected Outcome:

By the end of this task, you should:

Gain practical experience in data cleaning and wrangling.

Learn how to design useful features from raw data.

Practice exploratory analysis without depending on step-by-step instructions.

Present insights clearly with both visuals and explanations.

Submission: Submit (Click on submit to access the submission form)

Deadline: 22nd October, 2025 by 11:59 PM WAT

TikTok | HNG Internship

HNG Internship

Data Analysis Stage 1 Task - HNGi13

TikTok | HNG Internship

HNG Internship

Data Analysis Stage 1 Task - HNGi13