**Logo

Description automatically generated**

[**Project Deliverable 2:**](https://uncc.instructure.com/courses/171082/assignments/1503788) **Data Understanding** a) Exploratory Data Analysis b) Dashboard

And **Data Preparation**

Group 4

Members - Dhruv Jani, Dhruvil Patel, Karan Issrani, Mohit Gupta, Yaw Frempong

**Data Understanding**

The Data we have was divided in 4 different files. The link of the data we used is <https://files.grouplens.org/datasets/movielens/ml-latest-small.zip>

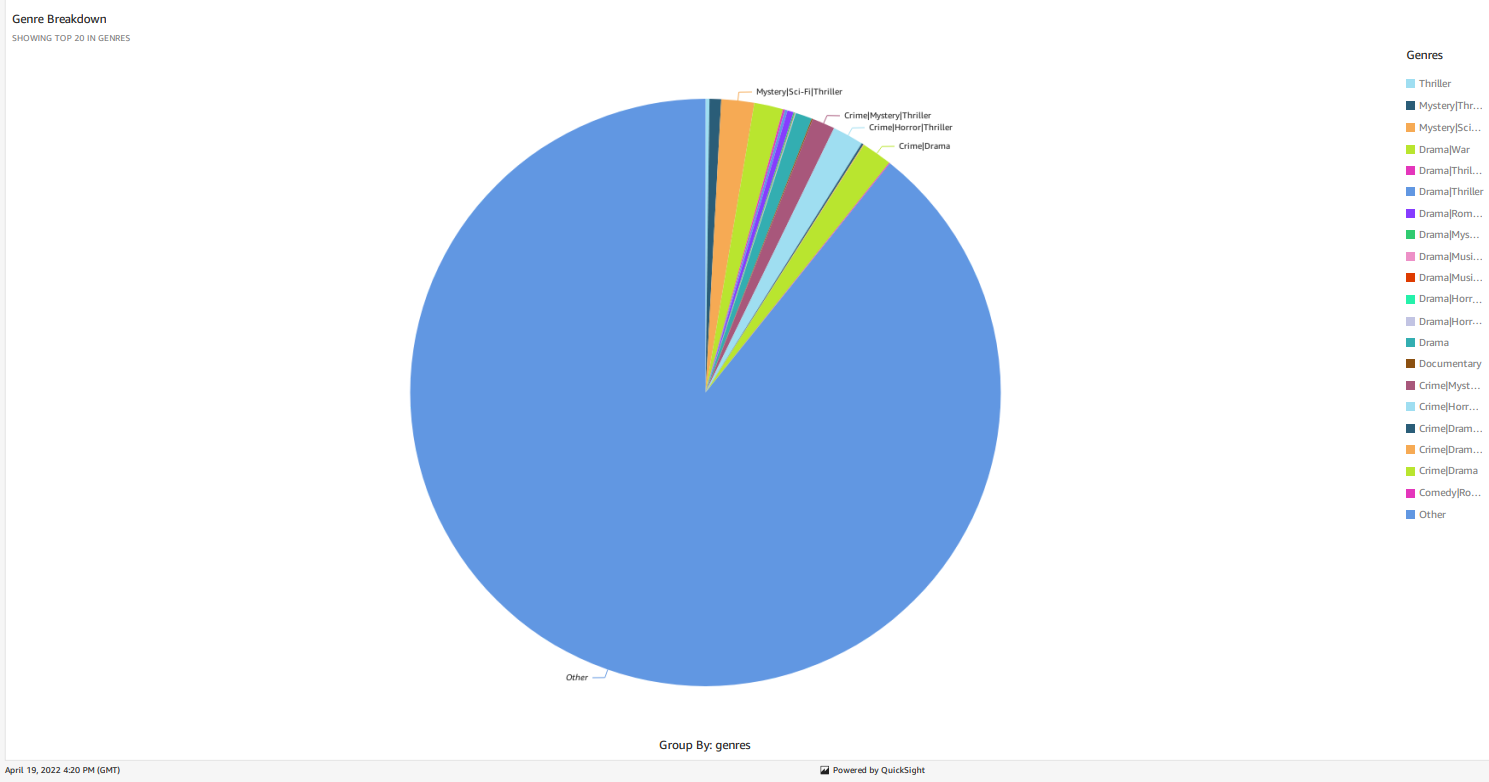
The data is divided into links.csv, movies.csv, tags.csv and ratings.csv

We used the primary key Movie ID to merge these data files and form our data set. Our data set consists of 100000 entries. The dataset merged\_data.csv has all the data in it. Our data has movieId,imdbId,tmdbId,title,genres,rating,userId,tag,timestamp,userId column names respectively.

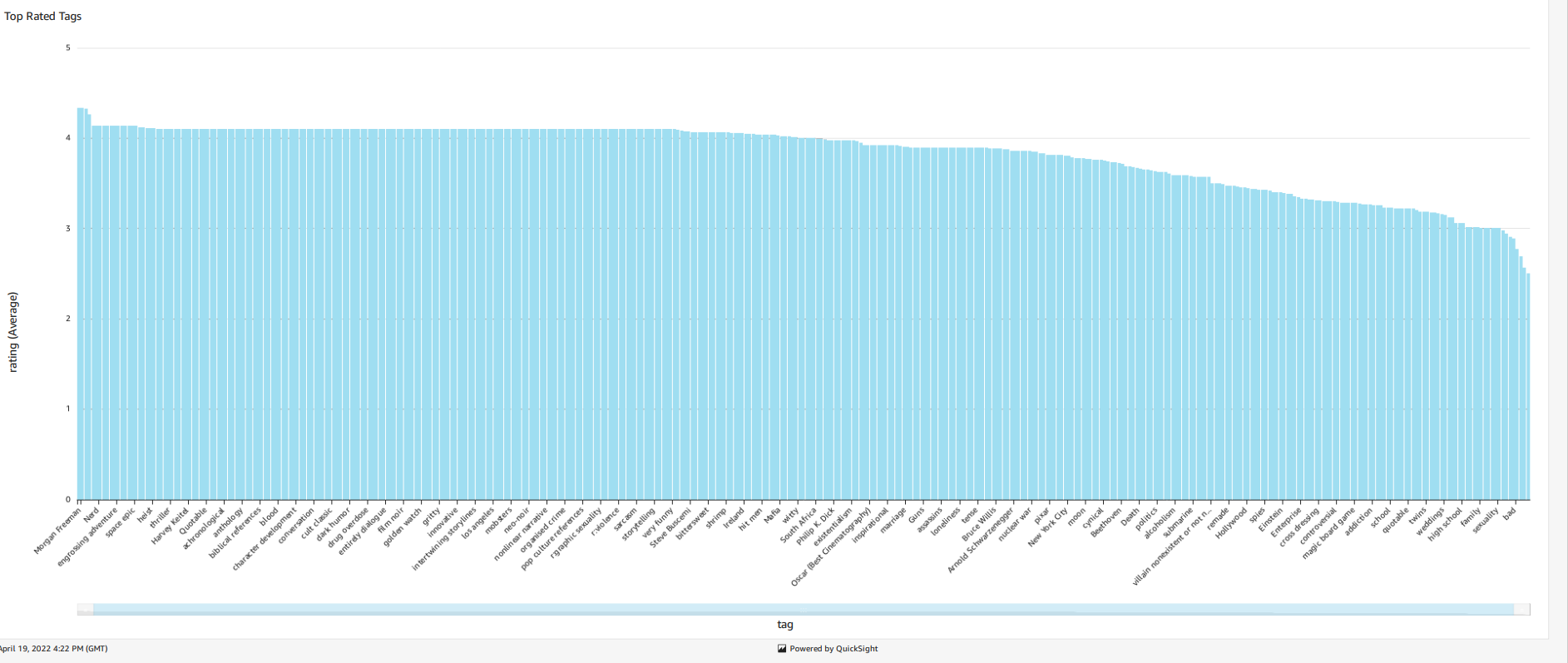
Now we do exploratory data Analysis on our dataset.

a) Exploratory Data Analysis

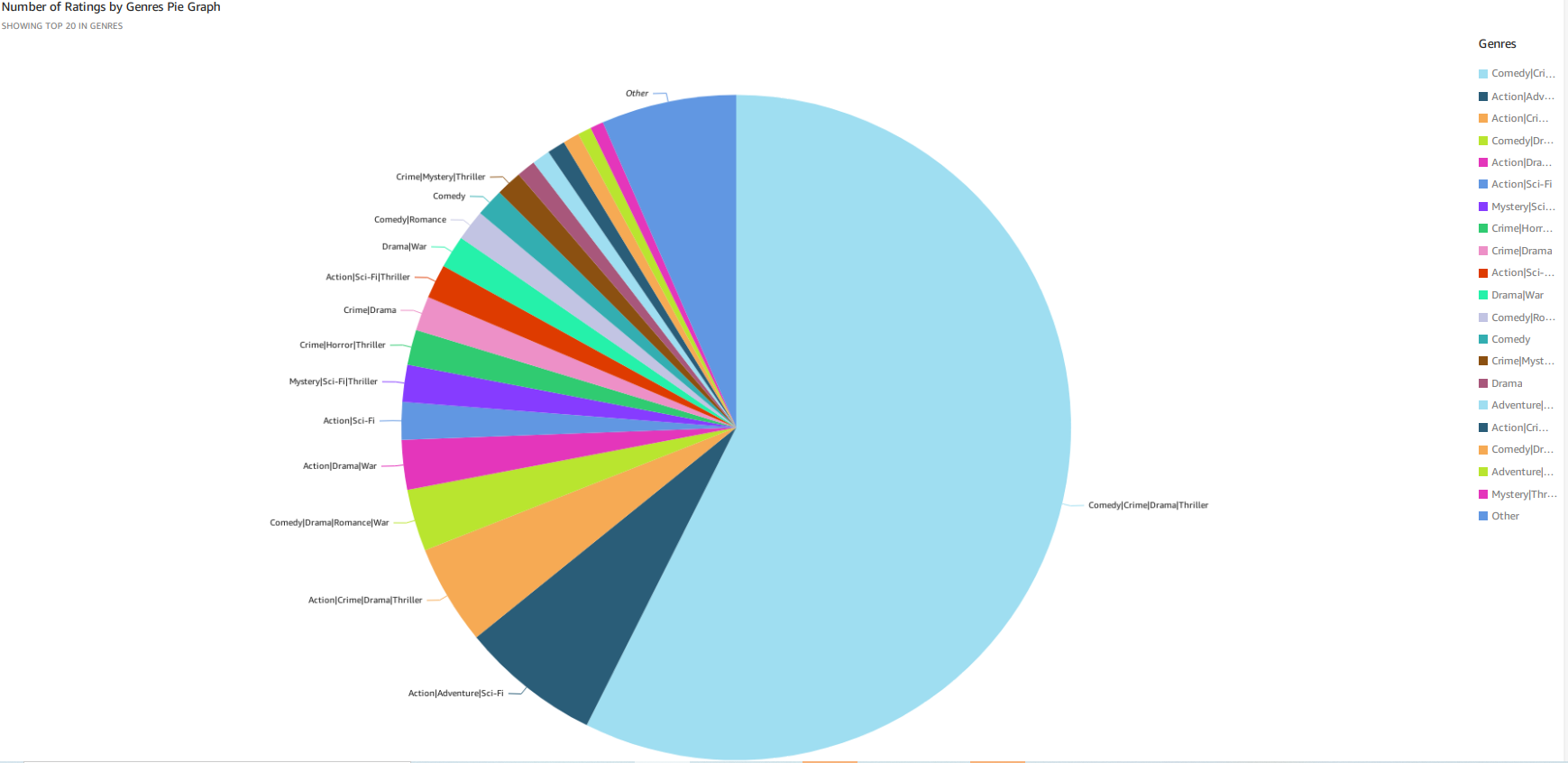
We use AWS Quick sight for our exploratory data analysis.



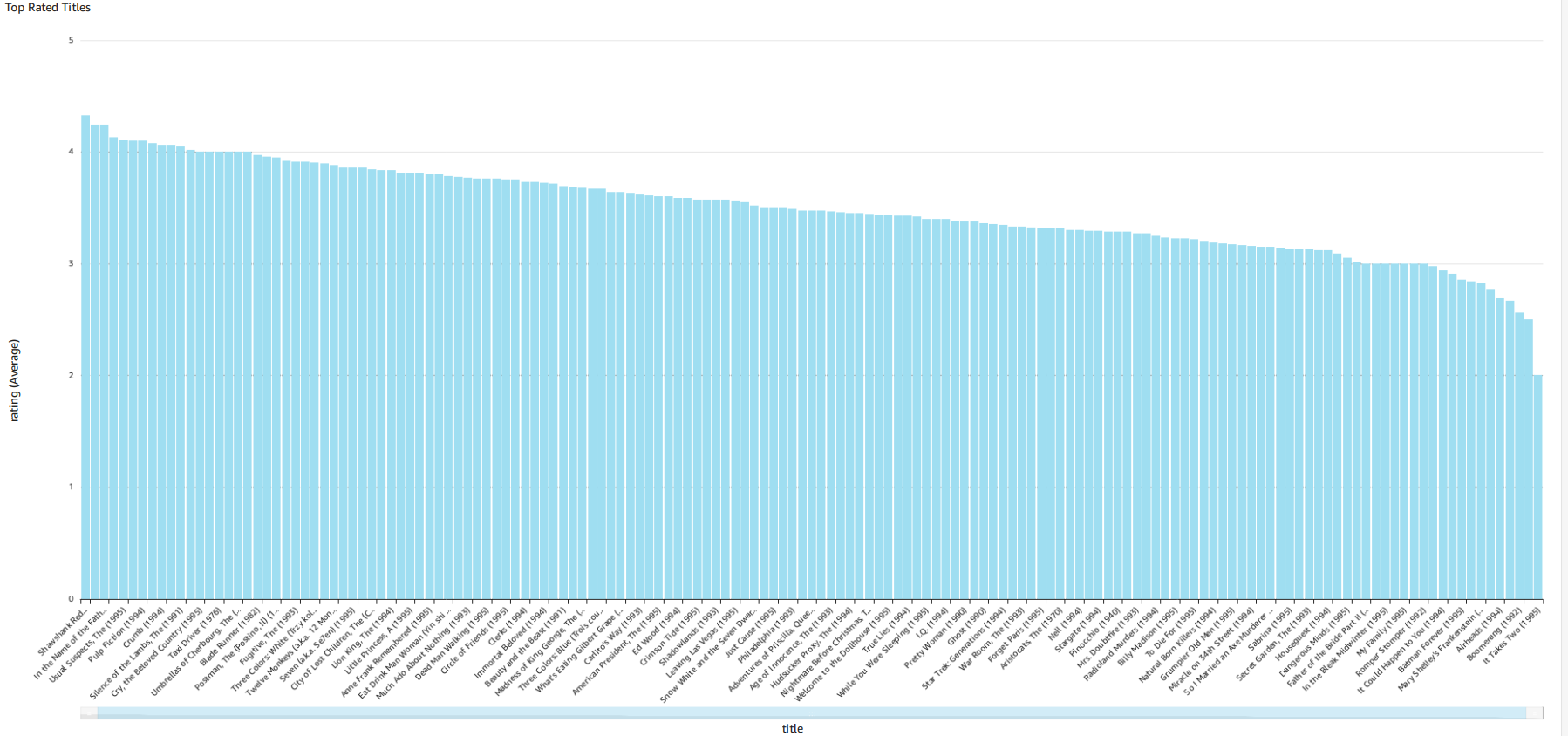
The graph above is the pie chart for representing the movies by genres. Here we can see that a large number of movies dont have one unique genre. But lie in the others category. The others category here represents that a large proportion of movies has more than one genre.



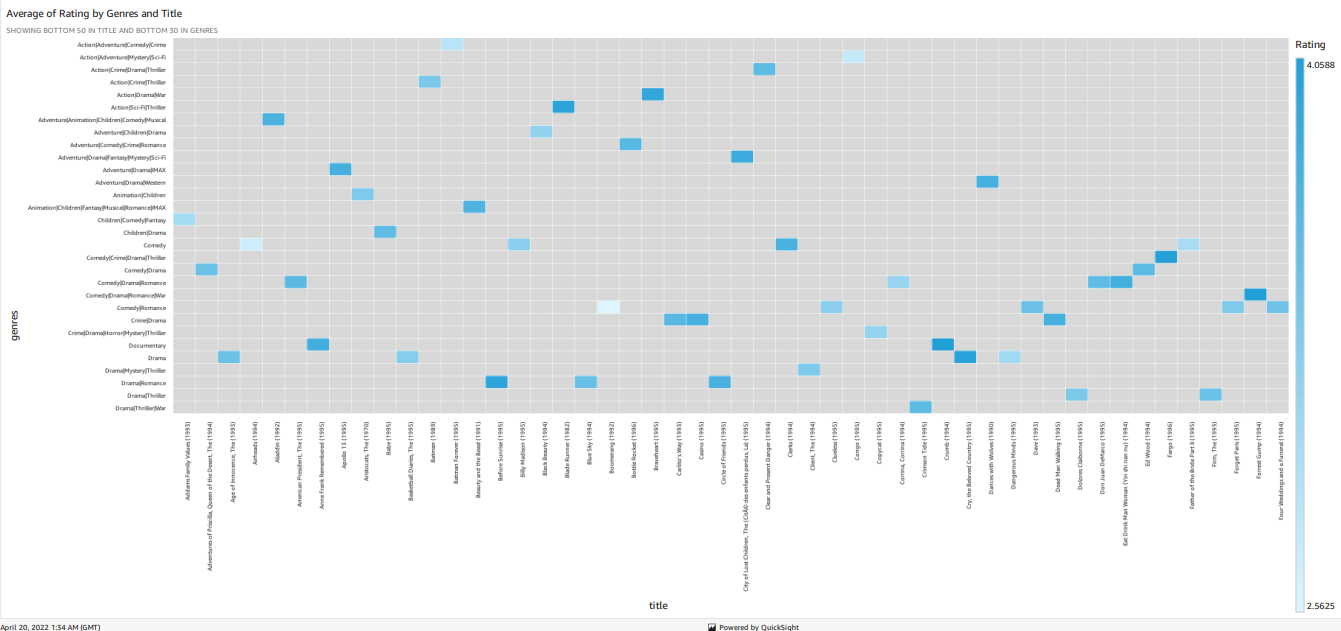
Here we see the graph of ratings vs tags. The artists who have the highest to lowest ratings are given below. This gives us an idea of the peoples choice. We can see that in our data Morgan Freeman has the highest number of ratings. This also gives us an idea that the graph is skewed. Tags can’t be considered as a feature in our model.



Here we plot the pie diagram to represent the number of ratings by genres. This is an interesting plot which shows us that which genres united together gives the best combination and has the highest rating. Our data says that the movies having genres like Comedy, Criminal, Drama and Thriller are the highest rated movies among users. The movies with Action genre precede these choices. The least preferred genre would be a combination of Criminal, Mystery, Thriller and Adventure movies. These features give the recommendation system a clear idea of what the preferences are in similar audience.



The most important relation of all is the title and ratings analysis. This graph shows us the average ratings for a movie out of 5. The favourites and relevant favourites can be mapped through this. Our main aim is to take the users input and based on the input give similar movies that might be the users favourite or are rated similar by the other users.



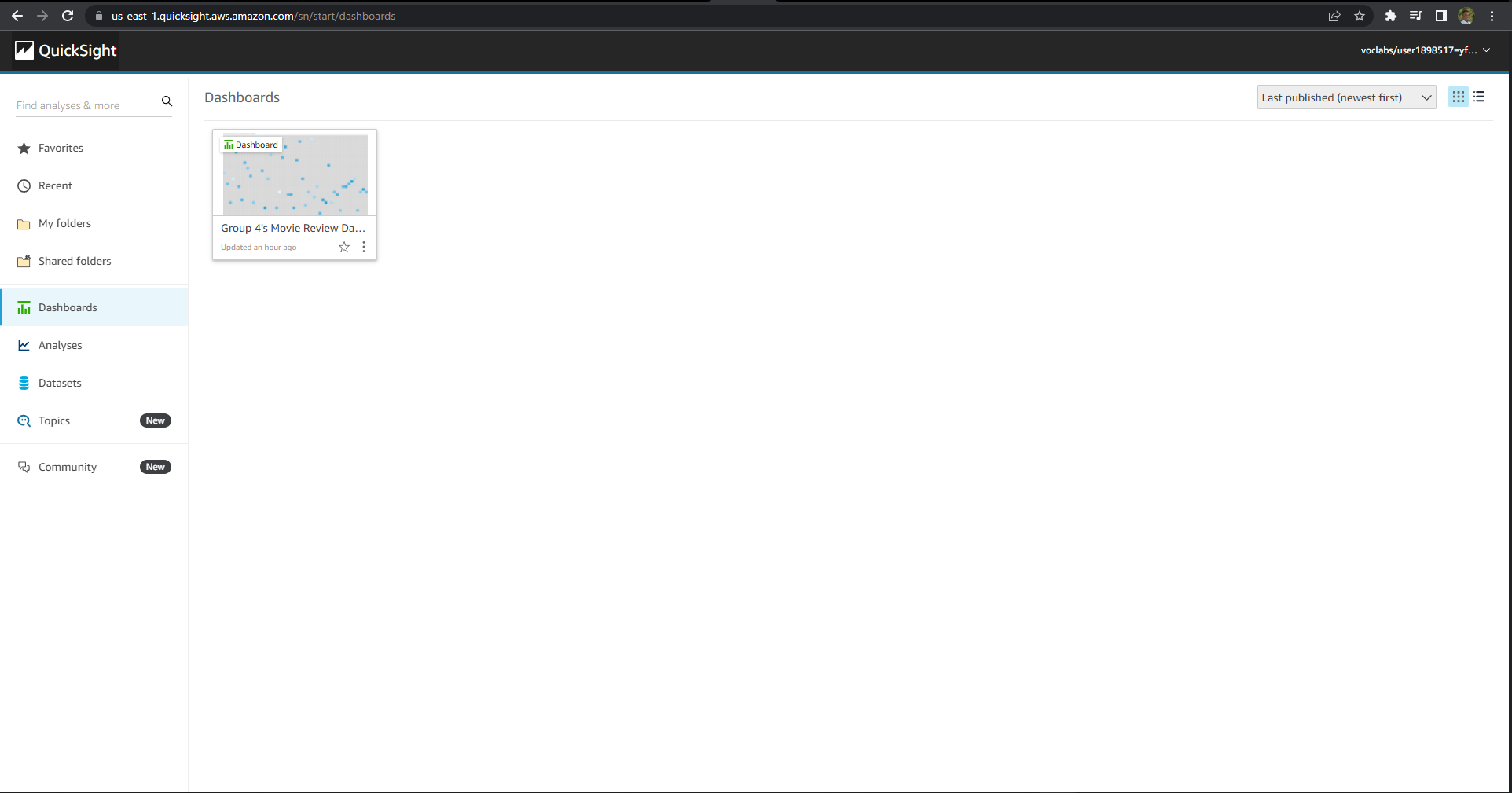
Average of Ratings by Genre and Title this tells us the value of these 3 features in our model. This clearly shows that our data is ready to undergo further process. Each movie can be distinguished as to which movie belongs to which genre and what rating it scores in its genre.

b) Dashboard

The link to the dashboard is <https://us-east-1.quicksight.aws.amazon.com/sn/accounts/530834130991/dashboards/1b211208-2faf-4afd-844a-6dc13bcfe64f?directory_alias=yawfrempong>

Since it is a private instance. It cant be accessed by others.

The screenshot below will show how the dashboard in Quicksight Looks.



**Data Preparation**

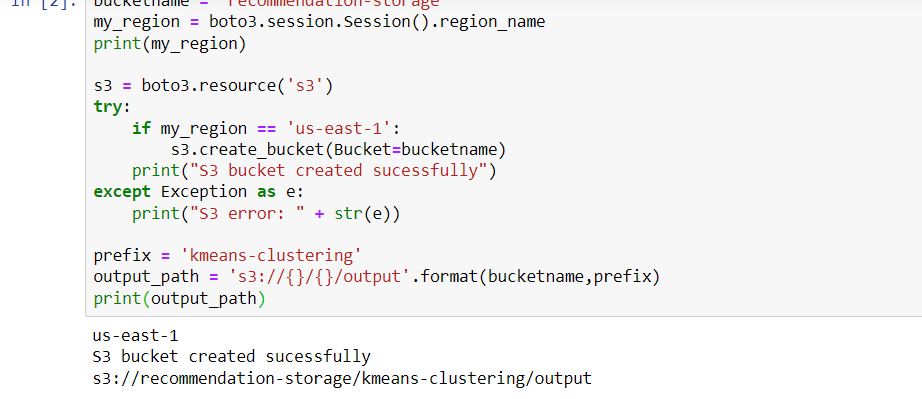
Step 1

We use Sage maker for the data preparation process and data cleaning process. We take Imports that are required for the process.



Step 2

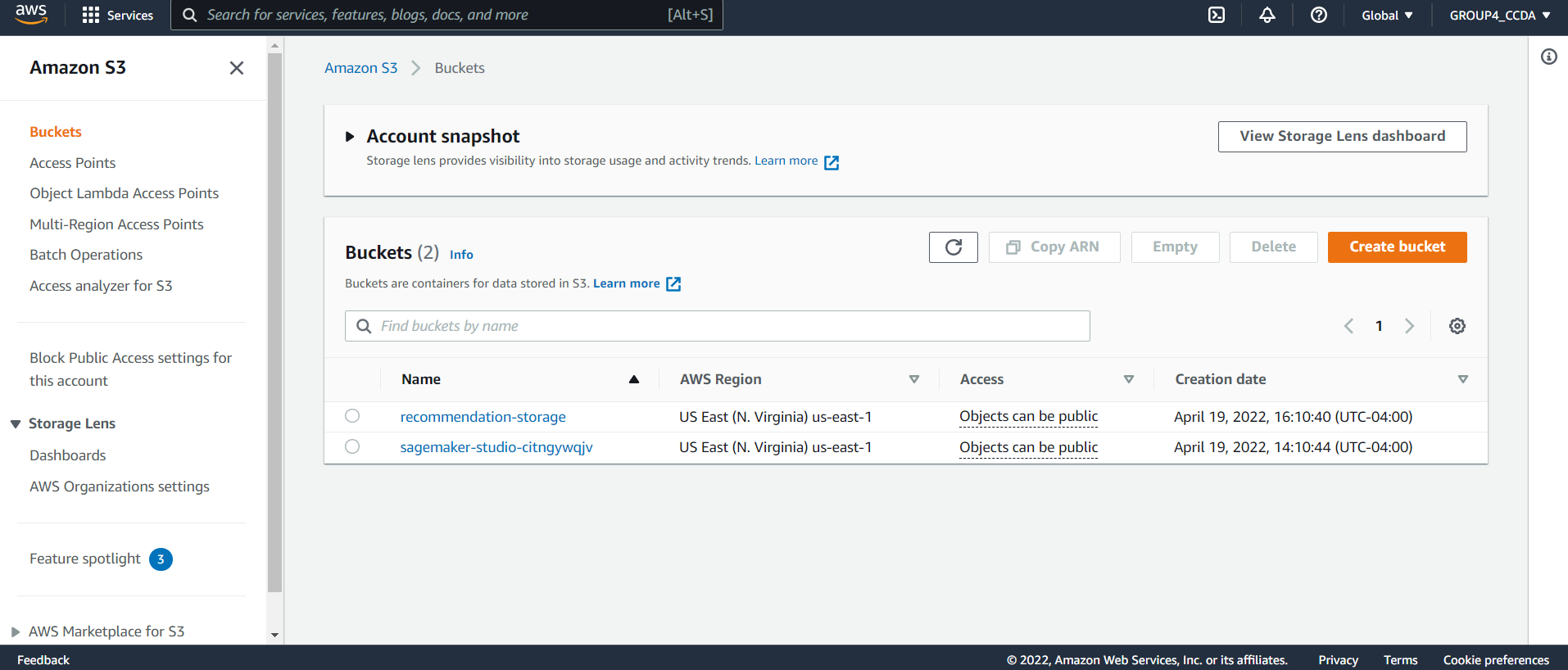
We need to Create an S3 bucket where all our csv files will be stored. We write simple python code to get the csv files. Convert them to dataframes and merge these data frames to get the required data set.



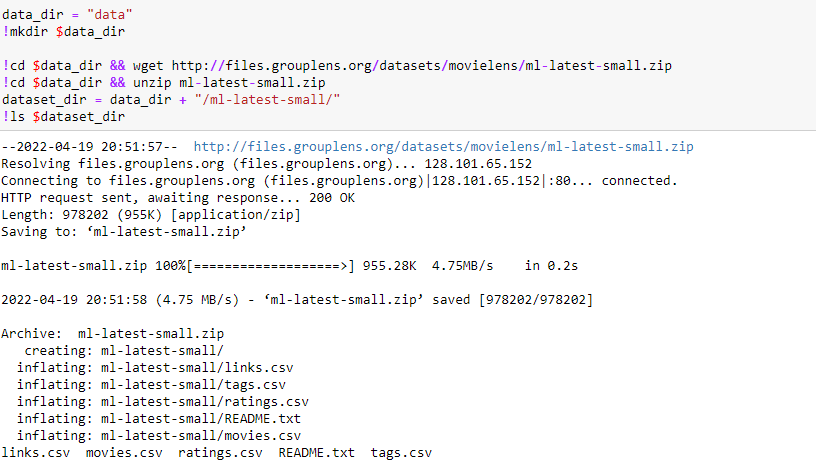
In this code we check if the region is us-east-1 so that our account location matches the S3 bucket location.

We get the file merged\_data after merging the files we need for our analysis. Here MovieId is the primary key for movies.csv and a foreign key in all other csvs.

Step 3

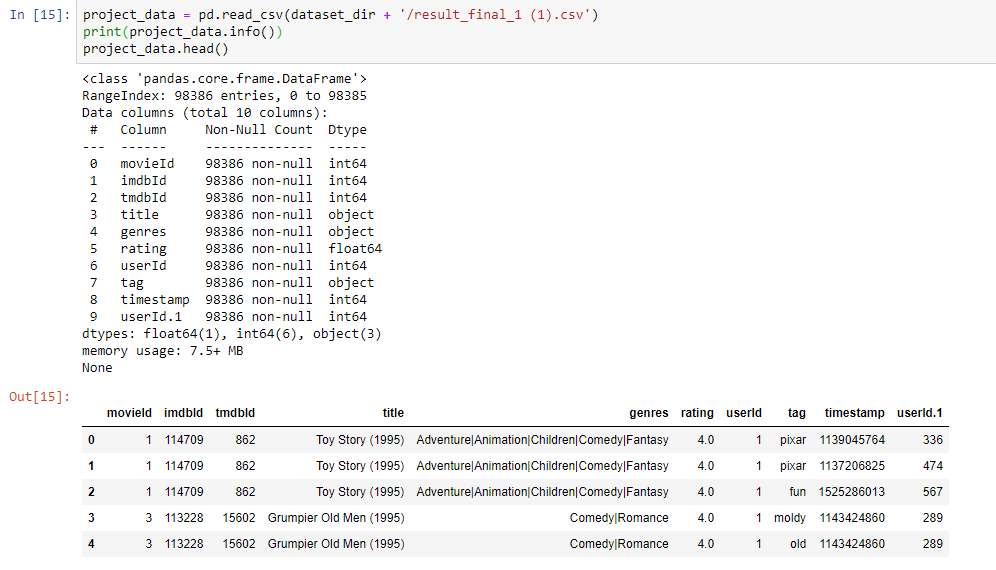
Check our S3 folder has been created. This folder will have our S3 repository and all other data. The folder now also has the newly merged data file.The recommendation-storage is the bucket we need.

Step 4

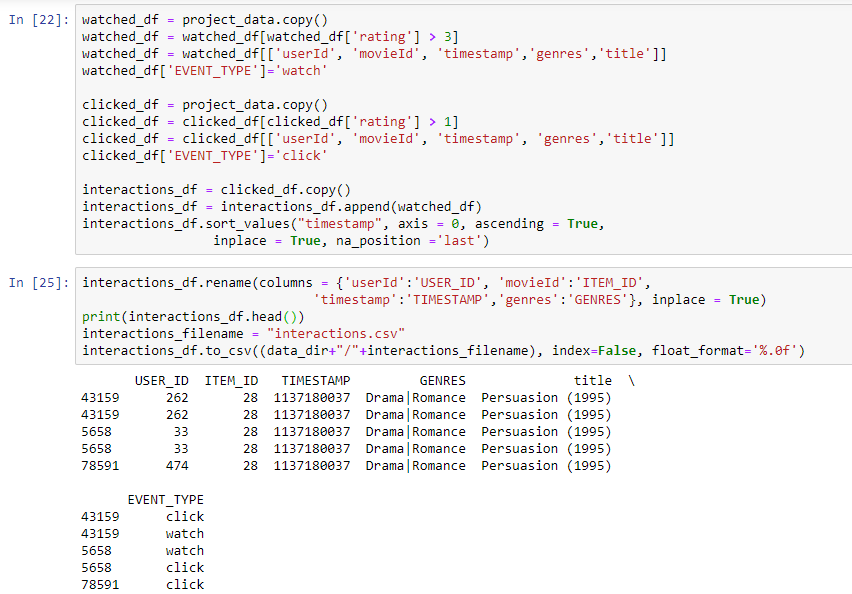
We can write a json code to take our data 

Step 5

Now we read the newly merged csv file to perform further cleaning of data according to our needs.



Step 6



In this step, we define two variables in the dataset to filter out unliked movies and better simulate data gathered by a video-on-demand (VOD) platform.

Since this is an explicit feedback movie rating dataset, it includes movies rated from 1 to 5. For this tutorial, we want to include only moves that were "liked" by the users, and simulate a implicit dataset that is similar to data that is gathered by a video-on-demain (VOD) platform. For that, you will next filter out all interactions below 2 out of 5, and create two EVENT\_TYPE variables: *click* and *watch*. Any movies rated 2 and above are assigned as *click*, and any movies rated 4 and above are assigned as *click* and *watch*.

This is the final step for our data cleaning. We get the data like we want from here.

Further, We will choose a ML algorithm to do further processes.

The AWS features we used are:-

AWS S3

AWS SAGE MAKER

AWS QuickSight

Sources

https://grouplens.org/datasets/movielens/

<https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html>

https://numpy.org/