Technical Report

Team Eta

Leslie Trejo, Julie Tang, Alireza Shams, Ershad Ziaei, Manxuan Zhang

This project allowed us to understand how to extract data from csv files, transform the data in jupyter notebook, and load in mongoDB compass. We obtained both datasets from kaggle. We chose data sets regarding movies/tv shows from Amazon and Netflix to see what movies both streaming services had to offer. Once transferred to jupyter notebook, we collaboratively decided which columns to drop, how we wanted to change the column names, and which column we wanted to drop null values in.

Amazon Data ETL

We used the drop function in order to drop 5 columns; “show_id”, “cast”, ”'date_added”, ”director”, ”country”. The “rename” function was used in order to change the first letter of the headers (of the columns we kept) to uppercase, and we changed the name of “listed_in”, as “genre”. The ratings column had many null values so we used “drop.na” in order to clean our data. We then decided to only keep all of the movies and drop all of the tv-shows. From there, we sorted all the movies in ascending order by the release year. This was then loaded to mongoDB compass by converting the data frame to a list of dictionaries, then we created the database/collections and named them according to their streaming services.

Netflix Data ETL

Similar to our last process, we used the drop function in order to drop 5 columns; “show_id”, “cast”, ”'date_added”, ”director”, ”country”. The “rename” function was used in order to change the first letter of the headers (of the columns we kept) to uppercase, and we changed the name of “listed_in”, as “genre”. In order to check for null values in the columns we used “drop.na” which showed no null values existed. We then decided to only keep all of the movies and drop all of the tv-shows. From there, we sorted all the movies in ascending order by the release year. This was then loaded to mongoDB compass by converting the data frame to a list of dictionaries, then we created the database/collections and named them according to their streaming services.