Objective: Each group writes a 2 to 4-page reflection paper on the project experience.
Tasks:
• - Reflect on the challenges faced during data selection, ETL setup and implementation, analysis, and cloud storage.
- Discuss lessons learned, particularly focusing on technical challenges, team coordination, and any improvements for future projects.
- Summarize skills gained and areas for further development.

DS 2002 Final Project
Reflection Paper

Carson Smith, Emma Mills, Neha Dacherla

Data Selection

During the data selection process, when deciding on which data we wanted to use for our project, we identified shared interests and themes that were relevant to our everyday lives. Once we landed on streaming media, we chose the Netflix and Hulu sites to analyze because, as consumers, we were most curious about trends and viewing habits across their worldwide audiences. We didn't face many challenges during this stage as the process of downloading the CSVs from Kaggle and loading them into Colab was something we had practiced numerous times over the course of the semester.

ETL Setup and Implementation

The extraction and loading component of this project was straightforward. During the ETL setup and implementation stages, the transformation stage was something that we had to revisit as we worked through our data visualizations later in the process. We quickly realized that clean and consistent data was a critical foundation for our data analysis, so re-grouping observations, creating new columns, removing unnecessary data points and missing values, and re-formatting data was an essential first step. This was less a challenge and more a key learning point for us as we quickly understood the importance of the cleaning, filtering, and structuring in the transformation stage. We also considered data and cloud storage, given that our existing datasets were relatively small, we built our ETL such that potential future data growth and scalability would be easily incorporated without causing problems in our code.

Data Analysis

In reviewing the respective datasets for Hulu and Netflix, we decided some of our most important columns would be "available_countries", "imdbaveragerating", "genres", and media "type" in order to establish trends and understand consumer behavior on a more nuanced level. We first decided to look at the crossover amongst titles on Netflix and Hulu, to understand how much access to content users had across both platforms. We found that in total, the two services offered a total of 24,472 titles, 2,148 of which overlap. Upon further analysis, we came to understand this number spans across 86 countries, each of which have different licensing agreements, maturity ratings, and local regulations. We found, after segmenting by country, in the combined Hulu and Netflix dataset that Japan (25.3%), the Andorra (22.9%) and the U.S. (10.10%) had the largest access to titles. Yet, when we began to break this datum down to understand voter habits by country, we noticed that there was one key difference

between the separated Hulu and Netflix datasets. While the combined dataset represented 86 countries, when creating visualizations for the number of votes by country for Hulu, we realized that the only available countries that were represented in the dataset were the U.S. and Japan. In cleaning the Hulu dataset some more, we found that Japan actually had more access to the titles featured on Hulu, at 6205 titles, while the U.S. only had access to 2483 titles. In this discovery, we wanted to understand the voting differences between the respective countries and used a plot chart to come to the conclusion that Japan simultaneously outvoted the U.S., while also rating titles at higher levels. When we attempted this same analysis for the Netflix dataset, there were too many plot points as there are 86 countries represented in the data, while there are only two in the Hulu data. Further, we wanted to understand how access to titles in Japan and the U.S. compared between Netflix and the U.S. , and how Hulu's exclusivity to Japan and the U.S. impacted the combined dataset. In cleaning this data, we found that Japan, Andora, and the U.S. still ranked amongst the highest access to titles, yet at different proportions. Japan ranked first with access to 6938 titles, Andorra second with 6277 titles, and the U.S. third with 2780. When interpreting the accompanying world clouds that represent the combined and Netflix data sets, you can see the proportional differences the Hulu datasets create.

We also wanted to explore the differences in genres the two datasets featured, and started by using a histogram to visualize the number of titles in the combined dataset, grouped by genre. We found that the most represented genres were Comedy, Drama, and Action. Further, we wanted to see how these genres were rated on IMDB, hypothesizing that Comedy, Drama, and action would be rated highest, yet our hypothesis was disproved. We found that amongst the 27 genres, average ratings were relatively uniform, with Western films ranking lowest, and War films rated highest. We wanted to explore the actual rating system more by comparing production year differences, and initially, created a line chart visualization that we concluded was likely irreflective of the quality (or consumer taste) for film over time. Our line chart initially indicated that media in the mid to early 1900s was rated significantly higher than media produced any time in the 2000s. Yet, upon creating a scatterplot of the average rating of the media in the combined dataset vs the year the media was produced, it became clear that it was a numbers game; there were significantly more films represented on IMDb from the 2000s than the mid to early 1900s.

Cloud Storage

Connecting to Google Cloud Storage (GCS) via Google Colab was a difficult yet rewarding learning experience, as it was our first time utilizing this program. We set up our project dashboard on GCS and created a storage bucket to house our cleaned data. In order to connect our Python script on Colab to the new bucket in GCS, we used the 'google-cloud-storage' library in Python to establish a connection, interact with the cloud environment, and upload data efficiently. It was rewarding to be able to run this code and instantly see our cleaned data appear in the GCS bucket, highlighting the value of GCS in project efficiency. Though our project for this class was relatively small-scale, interacting with GCS demonstrated the power of these tools to handle much larger projects as well, which we may encounter in our future work with data systems. Overall, though initially intimidating, interacting with GCS for this project increased our confidence in using cloud storage to handle data.

Conclusion

Over the course of this project, we developed several key takeaways about working with data. First, we found that visualizations can be misleading without context. For example, we made two data visualizations to observe iMDB ratings over time- a line graph showing how ratings have changed over time and a scatter plot showing the volume of ratings by the years and how they compared over time. In the first line graph we noticed that ratings had generally decreased over time. However, this was due to two key factors: iMDB was founded in 1990 so there was more data following that year and the "nostalgia factor" was leading critics to rank older movies higher because they were "classics". Without this context, we would have misinterpreted our data visualizations and made a claim that inherently wasn't accurate.

Some insights we drew on our data were that there are more titles available on Hulu than Netflix overall even though Netflix has more subscribers. This led us to believe that Netflix's popularity with viewing audiences is a result of the platform's reputation and popularity with a broader range of demographics. We also found that there was a low overlap of media between the streaming platforms which as a result encourages viewers to subscribe to multiple streaming platforms to gain full access to their streaming needs.

Team coordination was a significant challenge we faced with working on the Colab simultaneously. We quickly realized that we had to alternate who was working on the script and rotate responsibilities amongst team members. For future projects, a more effective approach would involve each team member individually working on various data visualizations prior to collaboration. If we reconvened after individually working on the visualizations, we could quickly identify which ones were the most impactful and integrate them more efficiently.