Information Visualization
Checkpoint I

Checkpoint I: Project Proposal

Group: G20
Date: 2025/09/13

Problem Domain

With the growing number of streaming platforms available, each with its own unique and extensive content library, consumers face a challenge: choosing the right service. The problem domain is to help users navigate the complex streaming landscape. The goal is to develop a visual tool that allows confused users to compare platforms based on their personal preferences, whether they value critically acclaimed films, family-friendly content, international films, or specific genres. This work serves as an interactive guide to help users orient themselves and make a more informed decision.

Task Abstraction

1. Abstract Form: Compare the count of items for a filtered categorical attribute X between two values of a categorical attribute Y.

- **Concrete Question:** Between Netflix and Disney+, which one has more family-friendly titles (age_category = toddlers, child)?
- **Task Type:** Comparison, Filtering.
- **Target:** For families with small children wondering what would be the best streaming platform.

2. Abstract Form: How does the distribution of a categorical attribute X compare between two values of a categorical attribute Y?

- **Concrete Question:** What is the geographic distribution of production countries for content on Apple TV+ and Disney+?
- **Task Type:** Distribution, Comparison, Geolocation
- **Target:** For viewers interested in understanding how globally diverse the content catalogs are when comparing streaming platforms.

3. Abstract Form: Which value of a categorical attribute Y has the most items when filtered by a temporal range Z and a categorical attribute X?

- **Concrete Question:** Which platform has more Western movies and TV shows made between 1970 and 1990?
- **Task Type:** Rank, Comparison, Filtering.
- **Target:** For fans of genre-specific content wanting to know what platform has the best access.

4. Abstract Form: Compare the count of items for a filtered quantitative attribute X between two values of a categorical attribute Y.

- **Concrete Question:** Between Amazon and Netflix, which one has more highly-rated shows (imdb_score > 7.5)?
- **Task Type:** Comparison, Filtering.

- **Target:** For viewers who want to choose the platform with the highest number of critically acclaimed shows.

5. Abstract Form: Is there a correlation between a quantitative variable X and a quantitative variable Y?

- **Concrete Question:** Is there a correlation between a platform's monthly subscription price and the number of high-rated (IMDb > 7.5) titles it offers?
- **Task Type:** Correlation, Relationships, Comparison.
- **Target:** For consumers who want to know whether higher subscription fees are justified for high-quality content.

6. Abstract Form: What is the change in a quantitative variable X over a temporal range Z for a given categorical attribute Y?

- **Concrete Question:** By how much has Netflix's subscription price (price) changed since 2020?
- **Task Type:** Comparison, Filtering
- **Target:** For viewers to decide if Netflix remains worth the cost.

7. Abstract Form: Which value of a categorical attribute Y offers a given item X at the best value of a quantitative attribute Z within a temporal range T?

- **Concrete Question:** What was the streaming platform in 2023 where it was possible to watch Tarzan for the best price?
- **Task Type:** Comparison, Filtering.
- **Target:** For movie fans seeking the most affordable option to watch a specific title.

Data Abstraction

Initial Database Collection

Our project uses seven datasets, all obtained from **Kaggle**:

1. **Streaming content datasets**
   a. Contains information about movies and TV shows across multiple platforms (originally split by platform, later merged).
   b. Contains 15 variables and 25938 items (Paramount: 3307 Netflix: 6250 Amazon: 11196 Apple: 171 HBO: 3087 Disney: 1927)
   c. Static table dataset
2. **Pricing dataset**
   a. Contains subscription prices for streaming services over time.
   b. Contains 3 variables and 778 items
   c. Static table dataset

Initial Processing and Cleaning

- **Merging platform datasets**:

  - Originally, each streaming platform had a separate dataset.
  - We added a column (streaming_platform) to identify the source platform and merged all files into a single CSV.

- **Attribute selection**:
  - We removed attributes that are not needed for our analysis: *seasons, imdb_id, imdb_votes, tmdb_popularity, tmdb_score, description*

- **NAN values**:
  - Rows containing NaN values were dropped to ensure consistency.

- **Data standardization**:
  - production_countries: cleaned according to the dataset author's recommendations, which involved leaving only the first country associated with the production
  - converted ISO codes (e.g., US) to full names (e.g., *United States*) for better understanding.
  - genres: cleaned according to the dataset author's recommendations.

- **Pricing dataset alignment**:
  - Filtered out streaming platforms not present in the content dataset.
  - Renamed platforms to match those in the base dataset for consistent merging.
  - Aggregated subscription prices to a yearly level by taking the maximum price recorded per year per platform.

- **Derived measures:**
  - In our project, we have one derived measure called age_category where we divide the age_certifications into 4 categories: toddlers (TV-Y, TV-Y7-FV), child (G, TV-G,TV-Y7, PG, TV-PG), teenager (PG-13, TV-14) and adult(R, NC-17, TV-MA).

- **Exports**:
  - Both cleaned datasets were saved in **CSV format** with new names, streaming_platforms.csv (11 variables and 11275 items) and streaming_prices.csv (3 variables and 52 items), respectively, for later integration in D3 visualizations.

Attribute	**Type**	**Scale / Characteristics**	**Semantics**
title	Nominal	Identifier	Name of the movie/TV show
type	Nominal	Category	Content type (*movie* or *show*)
release_year	Interval	Temporal (linear)	Year the content was released
age_certification	Ordinal	Ordered categories	Age rating (*G, PG, PG-13, R, etc.*)
runtime	Ratio	Quantitative (minutes)	Duration of the content
genres	Nominal	Multiple categories	Main genres of the content
production_countries	Nominal	Category	Country (or countries) where content was produced
imdb_score	Ordinal	Quantitative (0–10 scale)	IMDb rating of the content
streaming_platform	Nominal	Category	Platform hosting the content (*Netflix, Disney+, etc.*)
age_category	Ordinal	Ordered categories	Grouped age suitability (toddler, child, teenager, adult)

File 1 – streaming_platforms.csv

Attribute	**Type**	**Scale / Characteristics**	**Semantics**
streaming_platform	Nominal	Category	Streaming service (aligned with Streaming_platform)
date	Interval	Temporal (linear, instant)	Date of price recording
price	Ratio	Quantitative (USD)	Subscription price of the service

File 2 - streaming_prices.csv

Mapping

Question 1: Between Netflix and Disney+, which one has more family-friendly titles (age_certification = G, PG, PG-13)?

- **Data needed:** streaming_platform, age_certification from streaming_platform.csv.
- **Approach:** Filter rows where age_certification is G, PG, or PG-13. Then count the number of titles per streaming_platform. Compare counts between Netflix and Disney+.

Question 2: What is the geographic distribution of production countries for content on Apple TV+ and Disney+?

- **Data needed:** streaming_platform, production_countries from streaming_platform.csv.
- **Approach:** Filter for Apple TV+ and Disney+. Then group by production_countries and count items per country for each platform.

Question 3: Which platform has more Western movies and TV shows made between 1970 and 1990?

- **Data needed:** streaming_platform, genres, release_year.
- **Approach:** Filter for genres containing “Western” and release_year between 1970 and 1990. Then count items per platform.

Question 4: Between Amazon and Netflix, which one has more highly-rated TV shows (imdb_score > 7.5)?

- **Data needed:** streaming_platform, imdb_score, type.
- **Approach:** Filter rows where imdb_score > 7.5 and type=SHOW. Then count the number of titles per platform.

Question 5: Is there a correlation between a platform's monthly subscription price and the number of high-rated (IMDb > 7.5) titles it offers?

- **Data needed:** streaming_platform, price (streaming_prices), imdb_score (streaming_platforms).
- **Approach:** Filter content with imdb_score > 7.5. Next, count high-rated titles per platform and calculate the correlation between price and the number of high-rated titles.

Question 6: By how much has Netflix’s subscription price (price) changed since 2020?

- **Data needed:** streaming_platform, date, price (streaming_prices).
- **Approach:** Filter for streaming_platform = 'Netflix' and for year >= 2020. Then calculate the difference between the latest price and the 2020 price.

Question 7: What was the streaming platform in 2023 where it was possible to watch Tarzan for the best price?

- **Data needed:** streaming_platform, title, date ,price (streaming_prices).
- **Approach:** Filter streaming_platforms that have the movie Tarzan, the year for 2023 and those platforms on the streaming_prices dataset. Check which one has the lowest price.