Tunisian Republic
Ministry of Higher Education and Scientific Research

Tunis Business School

BI and DBMS Project Report

Netflix BI Analysis

Professor: Dr. Manel Abdelkader

Prepared by

Gisele Aydi, Adem Maatallah

Junior IT/BA

Academic Year: 2024-2025

Contents

General Introduction

0.1 Business Needs

In an era where streaming services dominate the entertainment industry, data-driven insights are invaluable for maintaining a competitive edge.
This project focuses on analyzing Netflix-related data to uncover trends, user behaviors, and financial patterns.
By understanding this information, the Netflix business can better address user needs, refine its marketing strategies, and improve user experience.

0.2 Project Goals

The primary goals of this analysis are:

- To categorize and analyze Netflix's content catalog based on show types, genres, release years, and ratings.
- To gain insights into Netflix's userbase, including demographic data such as location, age, gender, devices used, and subscription types.
- To evaluate subscription fees across various regions worldwide and identify pricing strategies.
- To assess reviews of the Netflix app on Google Play, segmented by app versions, for insights into user satisfaction and app performance.

0.3 Key Questions to Explore

Through this analysis, we aim to answer several critical questions, such as:

- What are the most popular genres and types of shows on Netflix?
- How do user demographics influence subscription choices and device usage?
- What regional pricing strategies could improve Netflix's global market share?
- How do app reviews and version updates reflect user satisfaction and technical issues?

0.4 Deliverables

This project will produce the following key deliverables:

- An optimized database tailored for data warehousing and OLAP tasks, ensuring efficient data retrieval and analysis.
- Interactive and visually engaging dashboards built with Power BI, designed to make insights accessible and actionable for stakeholders.
- A comprehensive document summarizing conclusions, the challenges encountered during the analysis, and actionable decisions to support Netflix's future strategies.

This report provides a detailed account of the project phases, from data collection, cleaning and modeling to visualization, highlighting the challenges encountered and opportunities for further improvements.

Technical Implementation

0.1 Technologies and Tools Used

This project utilized a variety of technologies and tools to ensure efficient and accurate analysis. Below is a summary of the key tools and their purposes:

- **Kaggle**: Used for collecting diverse datasets related to Netflix.

- **Python (Pandas)**: Employed for data cleaning and preprocessing.

- **Python**: Used for loading and initial exploration of the datasets.

- **MySQL Workbench**: Utilized for data warehousing and managing the database schema.

- **Power BI**: Applied for OLAP operations, data modeling, and visualization to derive actionable insights.

0.2 Project Phases

0.2.1 Data Gathering

The first phase of the project involved collecting and preparing data for analysis. Kaggle served as the primary source, offering various datasets related to Netflix. From these datasets, we extracted only the relevant data necessary to address the key questions identified in the project.

The data included details such as:

- Netflix shows, their genres, types, release years, and ratings.
- User demographics, including age, gender, and devices used.
- Subscription fee details from different regions worldwide.
- Reviews and version updates of the Netflix app on Google Play.

After gathering the data, irrelevant or redundant datasets were removed to ensure a focused and streamlined analysis. The remained datasets were then prepared for the ETL process.

0.2.2 Data Preparation (ETL: Extract, Transform, Load)

The data preparation phase, commonly referred to as ETL, is crucial for ensuring that the data is cleaned, transformed, and loaded into the MySQL data warehouse in a usable format. It involves three main operations: extraction, transformation, and loading.

The extraction phase was already assured during the previous phase of data gathering. The transformation phase focused on cleaning and structuring the data, while the final loading phase involved inserting the cleaned data into the MySQL data warehouse for further analysis.

This process ensures that data quality is maintained, and the database can support accurate analysis.

Data Cleaning

Data cleaning is a critical part of the transformation phase, aimed at ensuring the dataset's integrity and consistency. For the general cleaning, duplicates were removed and missing data was handled. In addition, categorical data was transformed to binary or appropriate values, which helped ensure consistency across the datasets. The following operations were implemented:

- **Handling Duplicates:** Duplicates, especially in identifiers such as 'reviewer IDs' and 'show names', were removed to ensure accurate analysis.
- **Handling Missing Values:** Missing values were either filled with placeholders or rows were removed, depending on the importance of the data.

- **Column Transformation:** Certain columns, such as the 'Netflix original' column, were converted into binary values (0 and 1) for consistency and ease of analysis.

The following Python code snippet illustrates how duplicates and missing values were handled for one of the datasets:

```
# Remove duplicates
df_reviewers.drop_duplicates(inplace=True)
# Remove missing values
df.dropna(inplace=True)
# Handling missing values
df_cleaned["Mapped_Genre_ID"].fillna("Other", inplace=True)
```

Furthermore, specific transformations were applied to categorical columns. For example, the `Netflix original` column was transformed into binary values, making it easier to work with during analysis:

Listing 1: Movies_and_ratings.ipynb

```
# Transform categorical data to binary
data['Netflix_original'] = data['Netflix_original'].map({'
    True': 1, 'False': 0})
```

Data Loading

After the cleaning and transformation steps, the data was loaded into the MySQL data warehouse. Using Python and MySQL connectors, the cleaned data was inserted into the appropriate tables in the database. This process ensured that the data was structured and ready for analysis in Power BI.

The dbconnection.py file established the connection between the Python file and the MySQL server's appropriate schema.

Listing 2: dbconnection.py

```
from sqlalchemy import create_engine
engine=create_engine('mysql+pymysql://root:123456ja@localhost
    /netflix_db')
```

The main.py file is the one responsible for directly loading the data into its appropriate tables into the data warehouse schema when being run.

Listing 3: main.py

```
import pandas as pd
from db_connection import engine
df = pd.read_csv('C:\.Fichiers\.Revision\.Junior\BI&DBM\
    BIproject\Datasets\Cleaned\
    cleaned_Netflix_SubscriptionFees.csv')
df.to_sql('subscription_fees', con=engine, if_exists='replace
    ', index=False)
```

This example shows us how the `cleaned_Netflix_SubscriptionFees` csv file was loaded into the data warehouse, and how the duplicate values were handled by replacing them `if_exists='replace'`.

This process ensured that the data was accurately loaded into the data warehouse for further analysis and visualization.

0.2.3 Data Warehouse Creation and Data Modeling

Data Warehouse Creation

In this phase, the cleaned data was properly loaded into the MySQL data warehouse. The data was organized into fact and dimension tables, ensuring clarity and efficient querying. The primary fact tables identified were:

- **Shows and Ratings**
- **Subscription Fees**
- **Userbase**
- **App Reviews**

For each of these fact tables, appropriate dimension tables were created by extracting specific attributes into standalone tables. This not only reduced redundancy but also ensured that each dimension could be independently managed and referenced. The following dimension tables were created:

- **Shows and Ratings:** Genres, Types, and Netflix Produced dimensions.
- **Subscription Fees:** Subscription Types and Countries dimensions.
- **Userbase:** Countries and Subscription Types dimensions.
- **App Reviews:** Reviewers Data dimension.

Here is an example SQL snippet for creating a dimension table:

```sql
-- Creating the 'subscription_types' dimension table
CREATE TABLE subscription_types
SELECT DISTINCT Subscription_ID, Subscription_Type
FROM userbase
ORDER BY Subscription_ID ASC;
```

Foreign keys were then mapped appropriately between fact and dimension tables to establish relationships. For instance, the `mapped_genre_id` in the `shows_and_ratings` table was linked to the `genres` dimension, ensuring that each show's genre could be accurately referenced.

Data Modeling

After loading the data into the MySQL schema, we moved on to data modeling using Power BI. The imported data was organized based on the fact constellation schema, which links multiple fact tables through shared dimension tables. This schema ensures that the data model is flexible and scalable, making it easier to perform complex analyses and generate insights.

Based on this schema, the OLAP process applied was ROLAP (Relational OLAP). ROLAP involves querying relational databases to retrieve multidimensional views of the data, allowing us to execute complex queries and generate insights without pre-aggregating the data. This will be implemented further during the analysis phase.

Measures were created for each fact table to provide meaningful metrics. For instance, in the `App Reviews` table, the following measures were calculated:

- **Count of Rating 1**
- **Count of Rating 2**
- **Count of Rating 3**
- **Count of Rating 4**
- **Count of Rating 5**
- **Total Reviews**

Here is a brief DAX formula for calculating the count of each rating in Power BI:

```
-- DAX measure for counting rating 1
Count of Rating 1 = COUNTROWS(FILTER('Fact_app_reviews','
    Fact_app_reviews'[Score] = 1))
```

The full data model schema was carefully designed to ensure optimal performance and easy querying. Below is a diagram of the fact constellation schema used for this project:

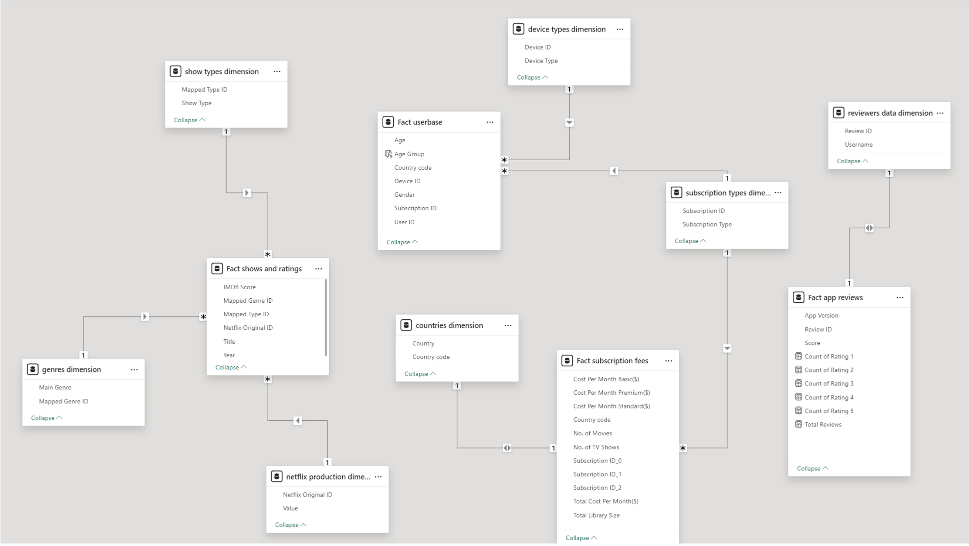


Figure 1: Fact Constellation Schema

This schema allows for efficient analysis of Netflix-related data, providing insights into shows, subscriptions, user behavior, and app reviews, all linked through well-defined dimensions.

0.2.4 Data Analysis Phase

In this phase, we leveraged a Business Intelligence tool, Power BI, to derive and present actionable insights from the Netflix dataset. The dashboards created during this stage were based on the implemented data model and imported data. The primary objectives were to understand user behaviors, analyze content offerings, and evaluate application performance.

Dashboard Insights

The dashboards designed employed various chart types to effectively visualize the data. Below is a summary of the dashboards and the visualizations used:

Netflix App Reviews Dashboard This dashboard summarizes user reviews for the Netflix mobile application on Google Play. It includes:

- **Stacked Bar Chart:** User ratings distribution (1 to 5 stars).
- **Star Visualization:** Average rating in a form of stars.
- **Card:** Average rating (2.82 stars) from over 102,000 reviews.
- **Card:** Total number of ratings (102,000 reviews).
- **Vertical Slicer:** Filters to explore reviews based on app versions.

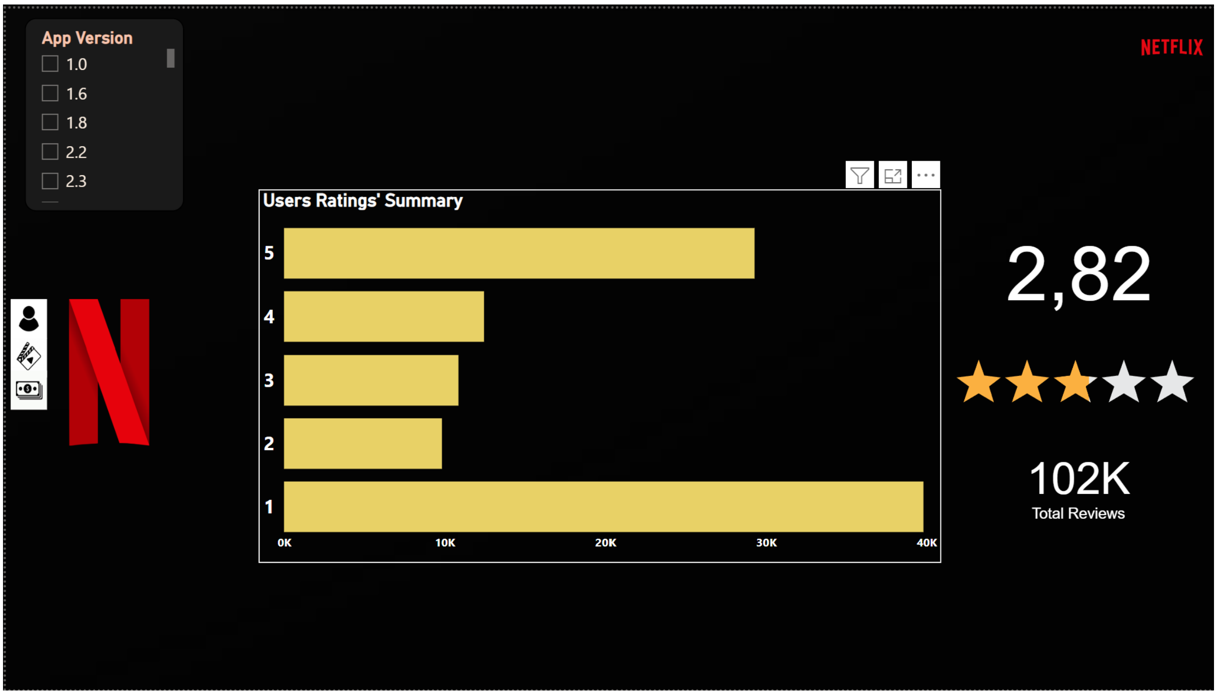


Figure 2: Netflix App Reviews Dashboard

Shows Analysis Dashboard This dashboard focuses on Netflix's content library and its attributes. Key visualizations include:

- **Tile Slicer:** Filters to examine shows solely produced by netflix or not.
- **Dropdown Slicer:** Filters to explore shows based on exact years.
- **Vertical Slicer:** Filters to view shows based on specific genres.
- **Pie Chart:** Distribution of shows by type (movies vs. TV shows).
- **Donut Chart:** Netflix-produced content as a percentage of the catalog.
- **Clustered Bar Chart:** Top genres by total number of shows.
- **Line Chart:** Temporal trends in show releases over the years.
- **Card:** Total number of shows, total genres, and average IMDb score.

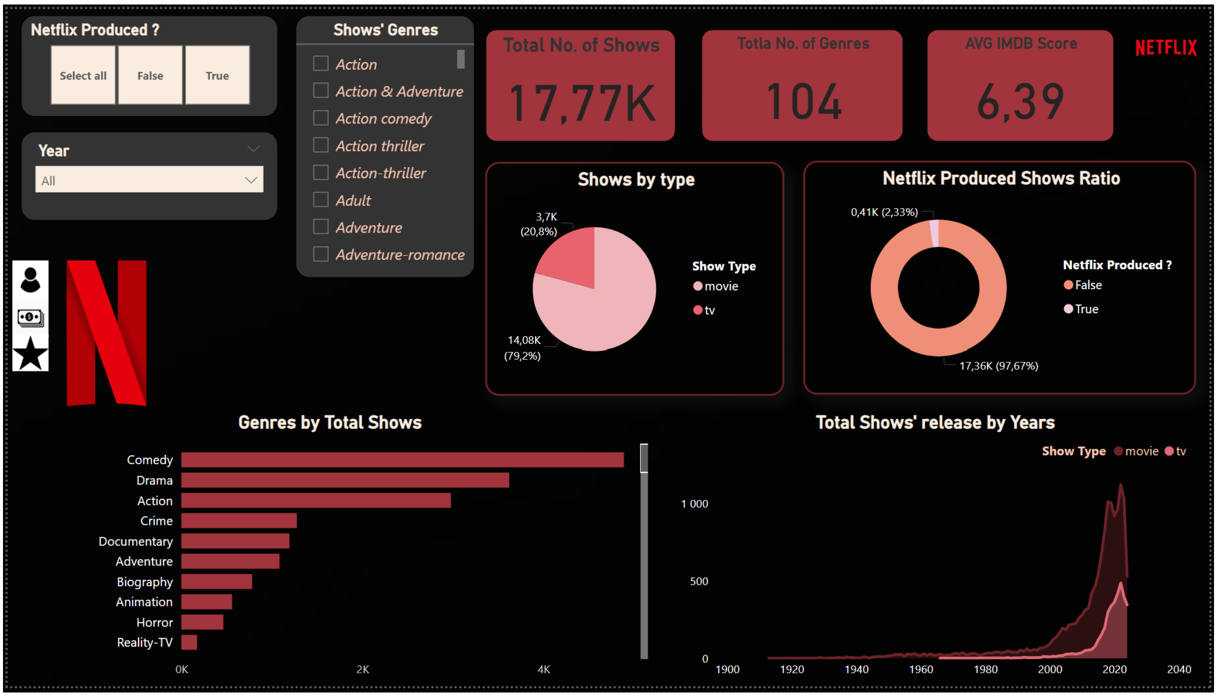


Figure 3: Shows Analysis Dashboard

Userbase Analysis Dashboard This dashboard highlights Netflix's user demographics and preferences. Key visualizations include:

- **Between Slicer:** Filter to examine users' data in an exact range on ages.
- **Map:** Global user distribution showing varying concentrations of users.
- **Infographic Chart:** Gender distribution of users.
- **Infographic Chart:** Device usage (laptops, tablets, smartphones and smart TVs).

- **Clustered Column Chart:** Users segmented by age groups.
- **Donut Chart:** Subscription types distribution of users (Standard, Basic, Premium).

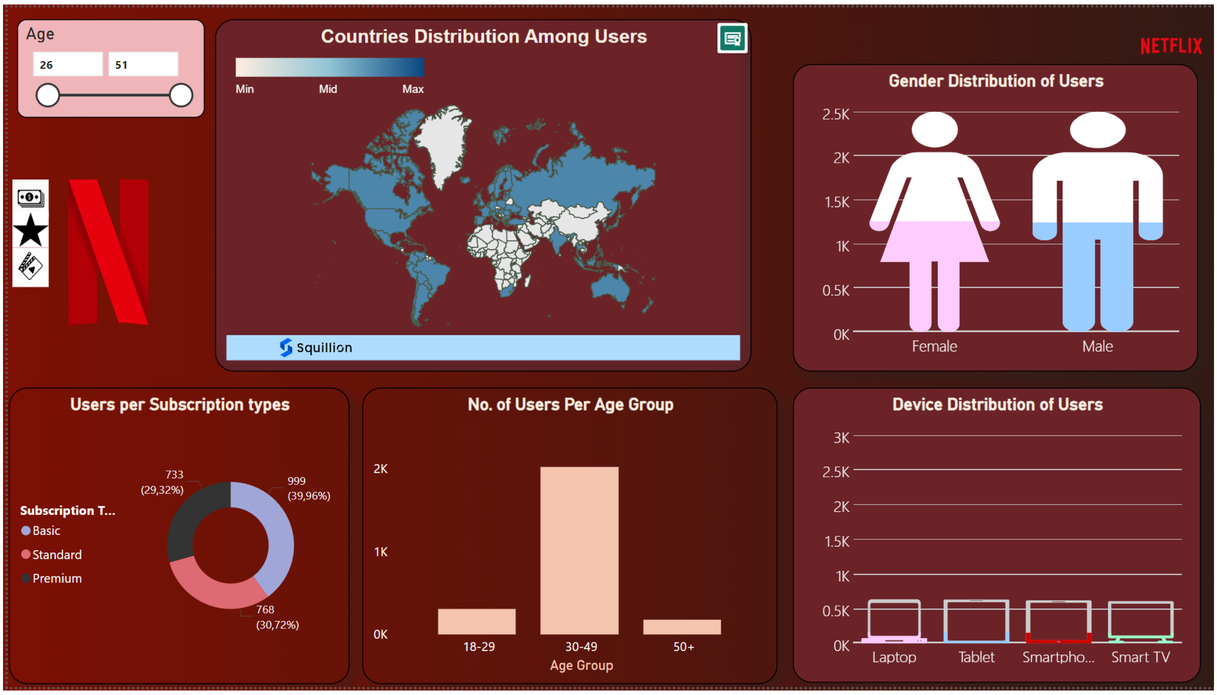


Figure 4: Userbase Analysis Dashboard

Subscription Fees Dashboard This dashboard presents Netflix's subscription fees and library composition across countries. Visualizations include:

- **Map:** Countries where Netflix is available.
- **Pie Chart:** Composition of the library (TV shows vs. movies).
- **Card:** Average monthly cost for Basic, Standard, and Premium subscriptions, and their total .

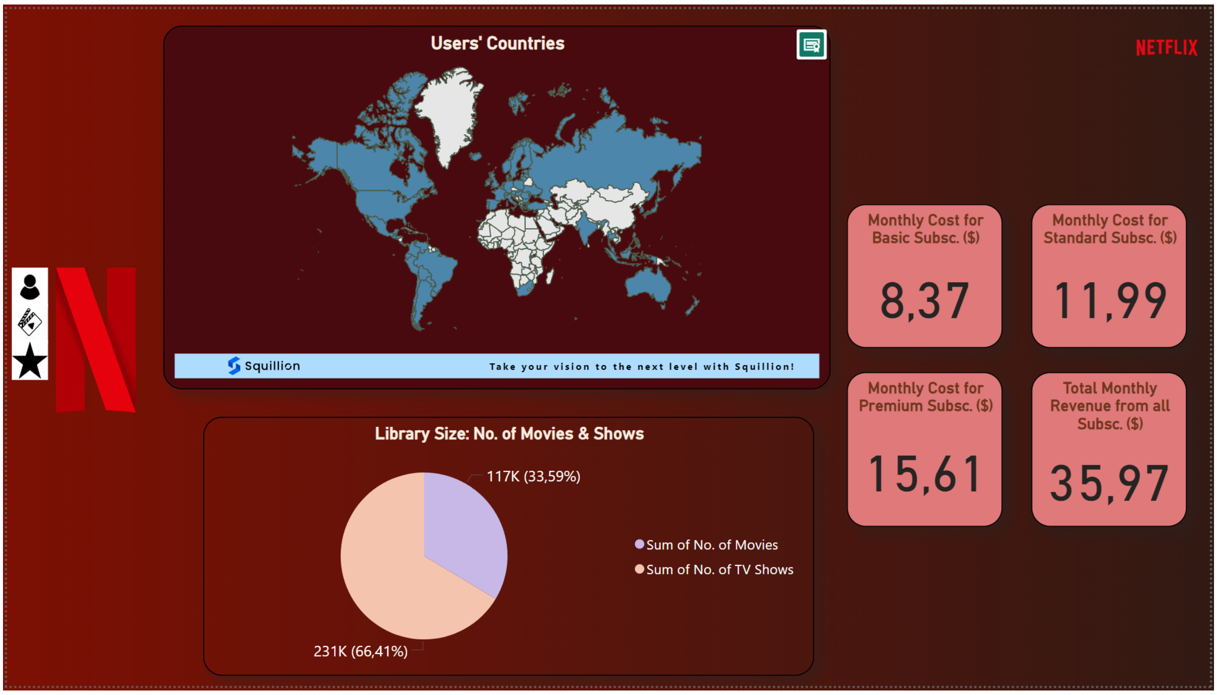

Users' Countries
Squillion
Take your vision to the next level with Squillion!
NETFLIX
Monthly Cost for Basic Subsc. ($)
8,37
Monthly Cost for Standard Subsc. ($)
11,99
Monthly Cost for Premium Subsc. ($)
15,61
Total Monthly Revenue from all Subsc. ($)
35,97
Library Size: No. of Movies & Shows
117K (33,59%)
231K (66,41%)
Sum of No. of Movies
Sum of No. of TV Shows

Figure 5: Netflix App Reviews Dashboard

ROLAP

ROLAP was employed to implement slicing and dicing as well as drilling functionalities:

- **Slicing**:
  - **Shows Analysis Dashboard:** Data was filtered by year, genre, and Netflix production status.

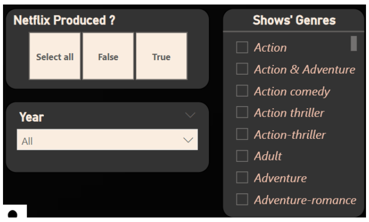


Figure 6: Slicing shows' data

  - **Userbase Analysis Dashboard:** Data was filtered by age groups.

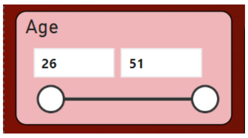
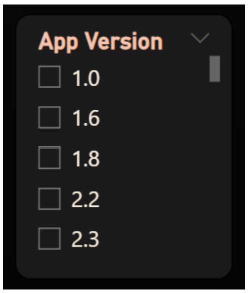


Figure 7: Slicing Data by users' ages

  - **App Reviews Dashboard:** Data was filtered by app version.



App Version

□ 1.0

□ 1.6

□ 1.8

□ 2.2

□ 2.3

Figure 8: Slicing data by app versions

- **Drilling:** Drilling was applied to graphs such as:
  - Number of users per age group, enabling deeper exploration by devices and subscription types.

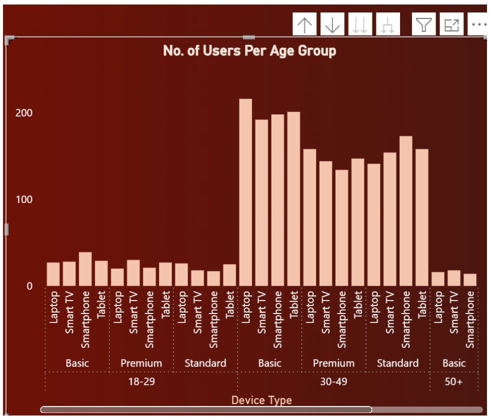


Figure 9: Drilling in Number of Users per Age Group Chart

  - Genres by total shows, adding dimensions like show type and Netflix production status.

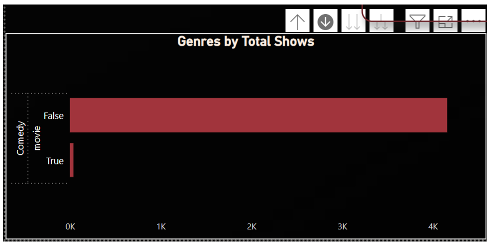


Figure 10: Drilling in Genres by Total Shows

Interactivity

The dashboards were designed with interactive features to enhance user experience:

- Page navigation was enabled through custom icons, allowing seamless transitions between dashboards.

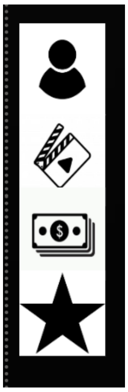


Figure 11: Page Navigation Icons

The dashboards effectively demonstrate the diverse aspects of Netflix's offerings and user interactions, providing a comprehensive overview of its business landscape.

Conclusion

0.3 Key Findings

From the dashboards and data analysis, several insights were derived:

- **Most Popular Genres and Show Types:** Comedy, Drama, and Action dominate Netflix's content offerings. Movies constitute the majority of the catalog (79.2%), while TV shows make up 20.8%. Netflix-original content remains limited, at only 2.33% of the total catalog.
- **User Demographics and Subscription Choices:** Female users slightly outnumber male users, with the 30–49 age group being the largest demographic. Standard and Basic plans are the most popular subscription tiers, and laptops are the most commonly used devices.
- **Regional Pricing Strategies:** Global pricing variations were visualized, highlighting regions where pricing might be a barrier. For example, high subscription costs in low-income countries could hinder user acquisition.
- **App Reviews and Updates:** The app has an average rating of 2.82 stars, with a significant skew towards 1-star reviews. Feedback revealed persistent technical issues across versions, emphasizing the need for improved user experience and feature updates.

0.4 Answers to the Research Questions

The following answers address the questions posed in the introduction:

- **What are the most popular genres and types of shows on Netflix?** Popular genres include Comedy, Drama, and Action. Movies are the predominant show type, but there is room to expand TV show offerings.
- **How do user demographics influence subscription choices and device usage?** The 30–49 age group is the primary subscriber base, favoring the Standard and Basic plans. Device usage trends show laptops as the top choice, followed by tablets and smartphones.
- **What regional pricing strategies could improve Netflix's global market share?** Adjusting subscription fees to align with regional income levels could enhance affordability in emerging markets. Offering bundled subscriptions with telecom providers could also improve adoption.

- **How do app reviews and version updates reflect user satisfaction and technical issues?** Low app ratings and frequent complaints about stability and performance highlight a critical need for technical enhancements and regular updates addressing user feedback.

0.5 Business Decisions

Based on the findings, the following strategic decisions are recommended for Netflix:

- **Content Strategy:**
  - Increase Netflix-original productions, particularly in popular genres like Comedy and Drama, to reduce dependency on third-party content.
  - Expand the catalog of TV shows to cater to evolving viewer preferences.
- **Subscription Models:**
  - Introduce region-specific pricing models to improve accessibility in underserved markets.
  - Consider a new, ultra-low-cost tier with limited content for price-sensitive regions.
- **App and Technical Improvements:**
  - Invest in resolving stability and performance issues to boost app ratings and user satisfaction.
  - Roll out app updates more frequently, ensuring that user feedback is promptly addressed.
- **User Engagement:**
  - Enhance recommendation algorithms to personalize content suggestions, increasing engagement and retention.
  - Develop interactive features, such as polls or quizzes, to boost app interactivity.

0.6 Challenges and Enhancements

Throughout the project, several challenges were encountered:

- **Data Quality:** The datasets required significant cleaning to handle missing or inconsistent entries.
- **Regional Analysis Limitations:** Certain regional user data was either incomplete or unavailable, limiting the depth of the analysis.

For future enhancements:

- **Data Enrichment:** Include more granular user and market data for improved regional insights.
- **Sentiment Analysis:** Apply advanced natural language processing to user reviews for deeper insights into user sentiment.
- **Real-Time Analytics:** Implement real-time data tracking to keep dashboards continuously updated.

0.7 Final General Conclusion

This analysis provided a comprehensive view of Netflix's content offerings, user behaviors, app performance, and pricing strategies. By leveraging dashboards and advanced data visualization, actionable insights were generated to address key business questions and strategic priorities. While challenges were encountered, the findings present significant opportunities for Netflix to improve its market position. Strategic content investments, user-focused pricing, and enhanced app functionality could position Netflix for sustained global success in an increasingly competitive streaming industry.