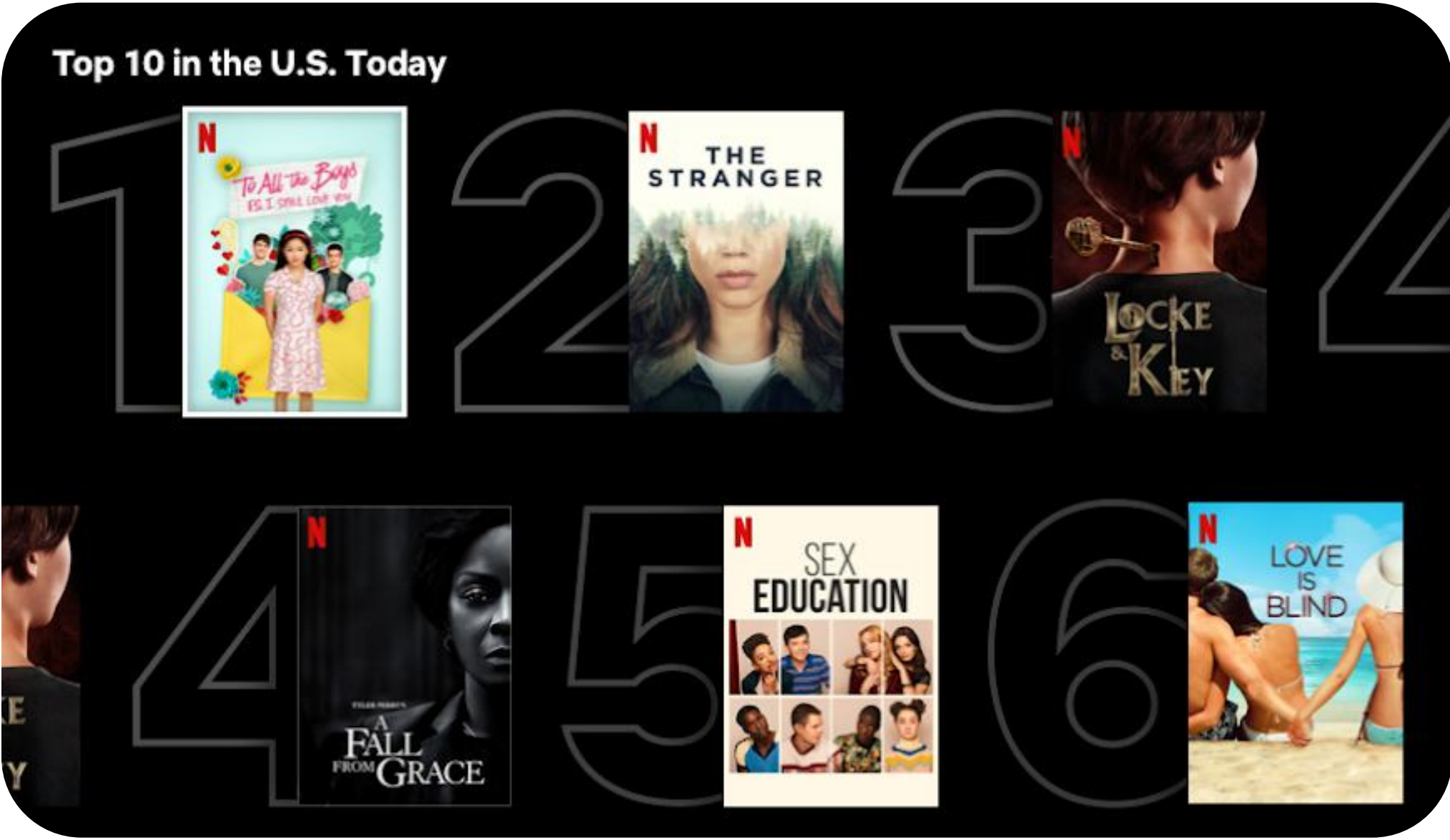# Examining the Netflix Top 10 Feature
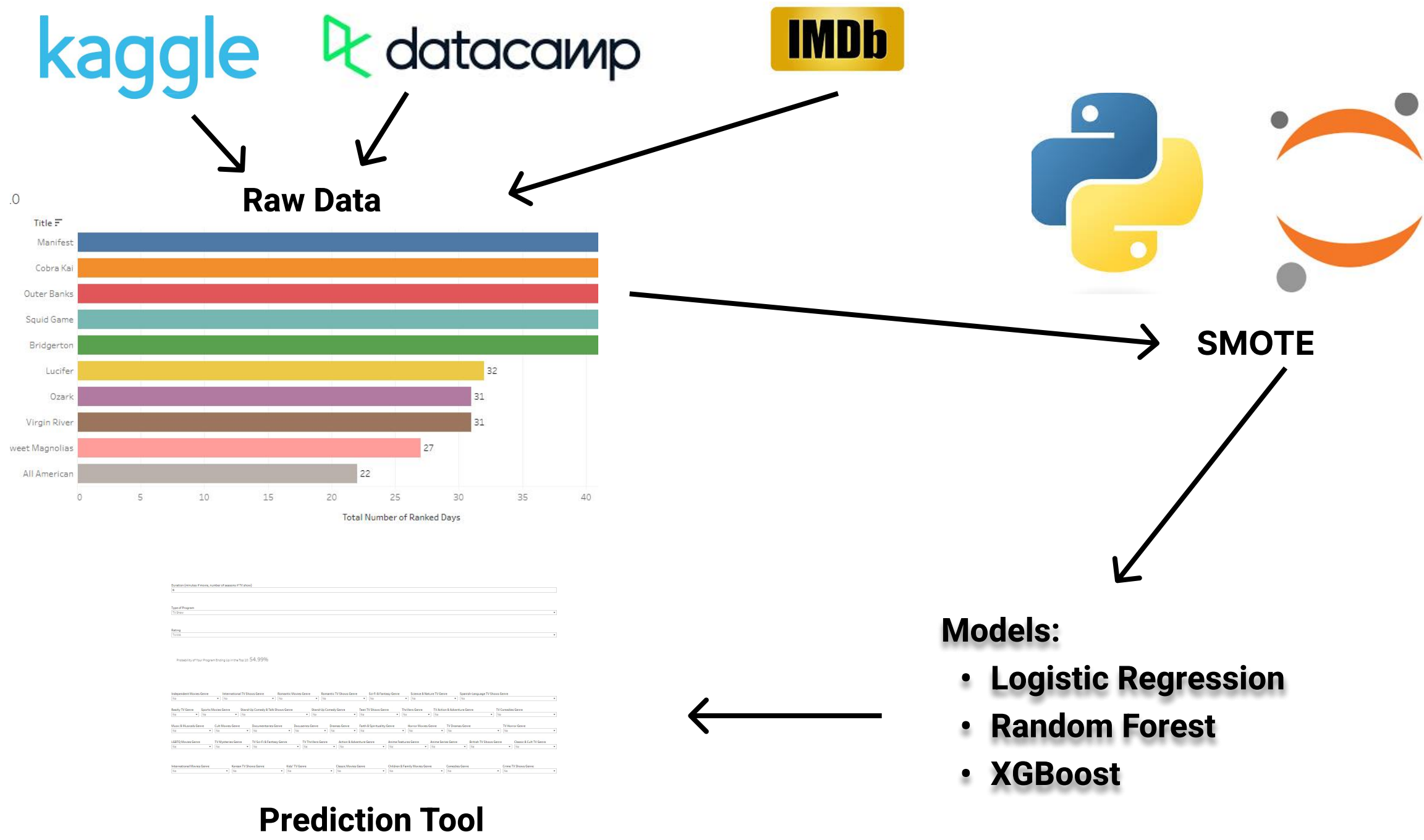
Motivation/Introduction

Welcome to our research project, aimed at dissecting the intricacies of Netflix's Top 10 rankings. While the streaming giant regularly updates its list, providing viewers with popular options, it lacks deeper insights into what truly drives a movie to the coveted Top 1 position. Recognizing the need for more informed consumption choices, we set out to delve into the key factors influencing these rankings.

Our project goes beyond surface-level popularity, focusing on deciphering the intricate workings of recommendation systems. By harnessing data analytics and visualization concepts, we've created a dynamic Tableau dashboard. This innovative tool empowers viewers to make informed decisions about their next watch, while also providing invaluable insights for business executives in the entertainment industry.

Whether you're a viewer seeking the perfect movie or a creator/business looking to optimize production strategies, our project offers a comprehensive solution to navigate the complex world of streaming content.

Top 10 in the U.S. Today

Process Flow

Raw Data

SMOTE

Models:
- **Logistic Regression**
- **Random Forest**
- **XGBoost**

Prediction Tool

Our Approach

Embark on a journey through our data-driven exploration into the dynamics of Netflix's Top 10 rankings. Leveraging a **classification machine learning model**, we endeavored to forecast the likelihood of a show securing a coveted spot in this elite list.

Our methodology encompassed meticulous preprocessing steps, including the **utilization of regex** to disentangle nuanced details such as duration and genre from the dataset. To address the inherent imbalances within the data, we adopted the **SMOTE technique**, ensuring a robust analysis. Through one-hot encoding and the application of **advanced algorithms like random forest, SVM, and XGBoost**, our model exhibited remarkable predictive accuracy. In our initial logistic regression model, validated through cross-validation, we achieved an impressive accuracy of 0.85.

This pioneering approach not only offers insights into viewer preferences but also **provides a predictive framework** beneficial for producers and filmmakers seeking to understand audience tastes. Through rigorous validation, including comparisons with alternative algorithms, we've established the efficacy of our logistic regression approach, paving the way for nuanced insights into the ever-evolving landscape of streaming content consumption.
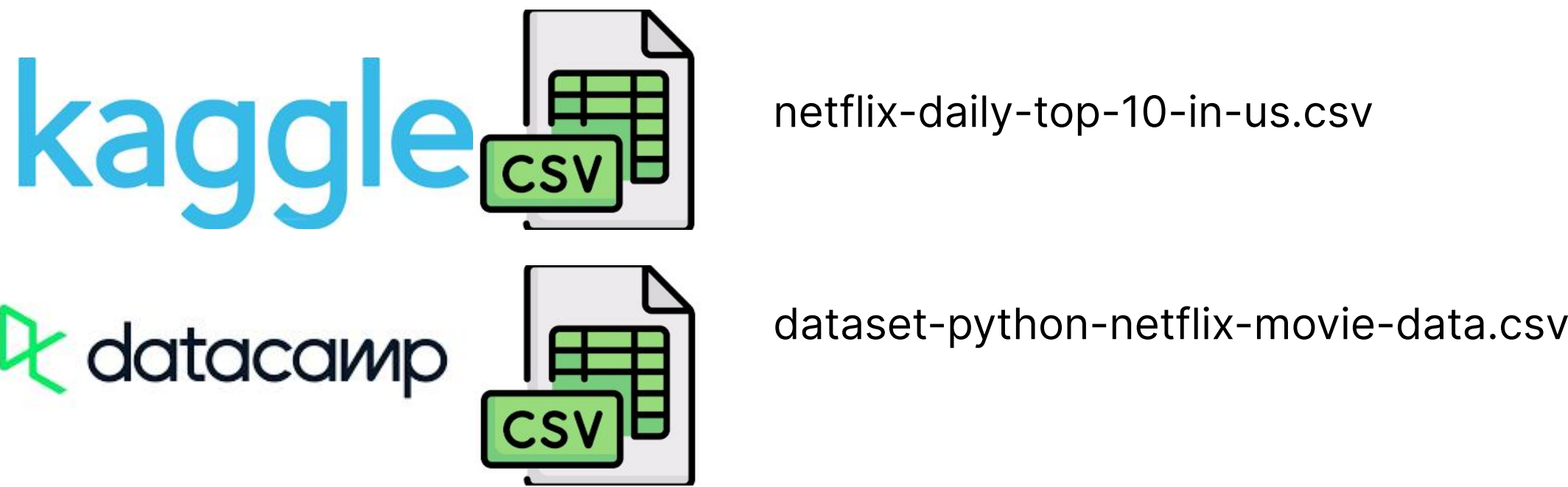
Data

The team acquired the data in the following ways:
1. Netflix Top 10 Dataset: This dataset was publicly available (Kaggle).
2. Additional Netflix Dataset: The team used two additional Netflix datasets to enhance the original top 10 dataset. (Kaggle)

Data Characteristics
1. Size on Disk: The size of the first dataset is a csv file of approximately 500 KB, containing top ten data. The second is a csv file of 3.4 MB containing additional characteristics per program.
2. Number of Records: The exact number of records in the first dataset is 7,100 rows. The second is 8,808 rows.
3. Temporal: The dataset is time-series as they are related to the Netflix top 10 feature, which changes over time.
4. Data Quality Issues: The team encountered several data quality issues, such as:
   - The genre field had all genres listed in a single column, separated by commas.
   - There was a many-to-many relationship between titles and genres and countries (of production).
   - The duration field had different value types depending on whether the content was a TV show or a movie.
   - There were null values that required handling.
5. Data Enrichment: The team tried to use IMDB data to further enhance the Netflix top 10 dataset, but found it to be too "dirty" to incorporate at all.

Data Sources

netflix-daily-top-10-in-us.csv

dataset-python-netflix-movie-data.csv

Data Pre-processing
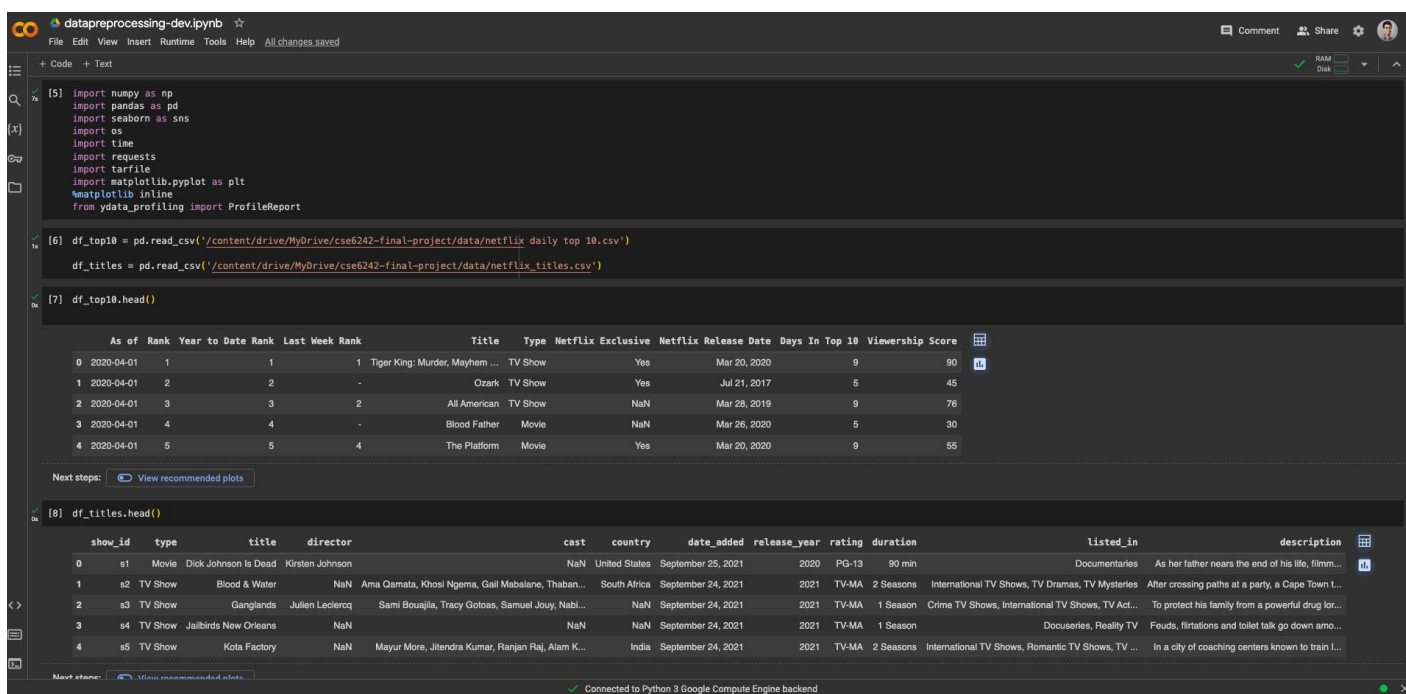
Pre-processing Methods
- Data Joining
- Data munging
- One-Hot Encoding Countries and Genres, respectively

Majority Top 10 Programs are TV Shows

% of Top 10 by Attribute

Legend
- Stand-Up Comedy
- Movie
- TV Show

Most Frequent Programs in the Top 10

Program	Count
Cocomelon	428
Ozark	85
Cobra Kai	81
Manifest	80
Outer Banks	72
Squid Game	66
All American	58
Bridgerton	58
Lucifer	56
Virgin River	55

Most Frequent Programs in the Top 3

Title	Count
Manifest	47
Cobra Kai	46
Outer Banks	46
Squid Game	44
Bridgerton	42
Lucifer	32
Ozark	31
Virgin River	31
Sweet Magnolias	27
All American	22

Total Number of Ranked Days

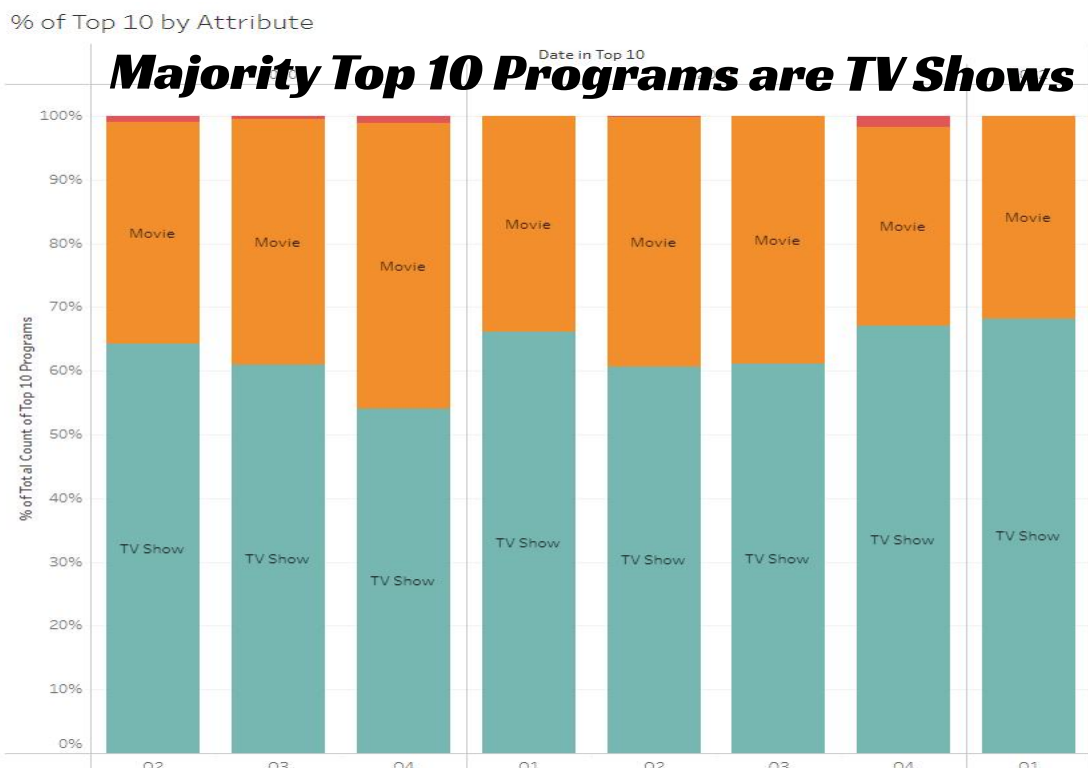## Experiments and Results

Findings:
- The following attributes have the greatest correlation with success of landing in the Netflix Top 10:
  - A TV Show is more likely than a Movie:
    - If the program is a TV Show, then a Miniseries or a TV Show with a smaller number of seasons is more likely to reach Top 10.
    - If the program is a Movie, then having a longer runtime correlates with success
  - Has one of the following genres:
    - Crime, Reality, Drama, Action, Mystery, Sci-Fi, Fantasy

- Manual cross-validation was conducted on the 10 programs on the visualization on the left.
  - Our tool predicts high chances of all the programs on the left ending up in the Top 10 with the exceptions of Cocomelon (3%) and Squid Game (19%).
    - Cocomelon and Squid Game are simply unicorns within their genres.
    - Our tool's output should not be seen as a guarantee of success or failure, but rather an educational guide.
    - Cocomelon does not appear in the most frequent programs in the Top 3 (bottom left visualization)
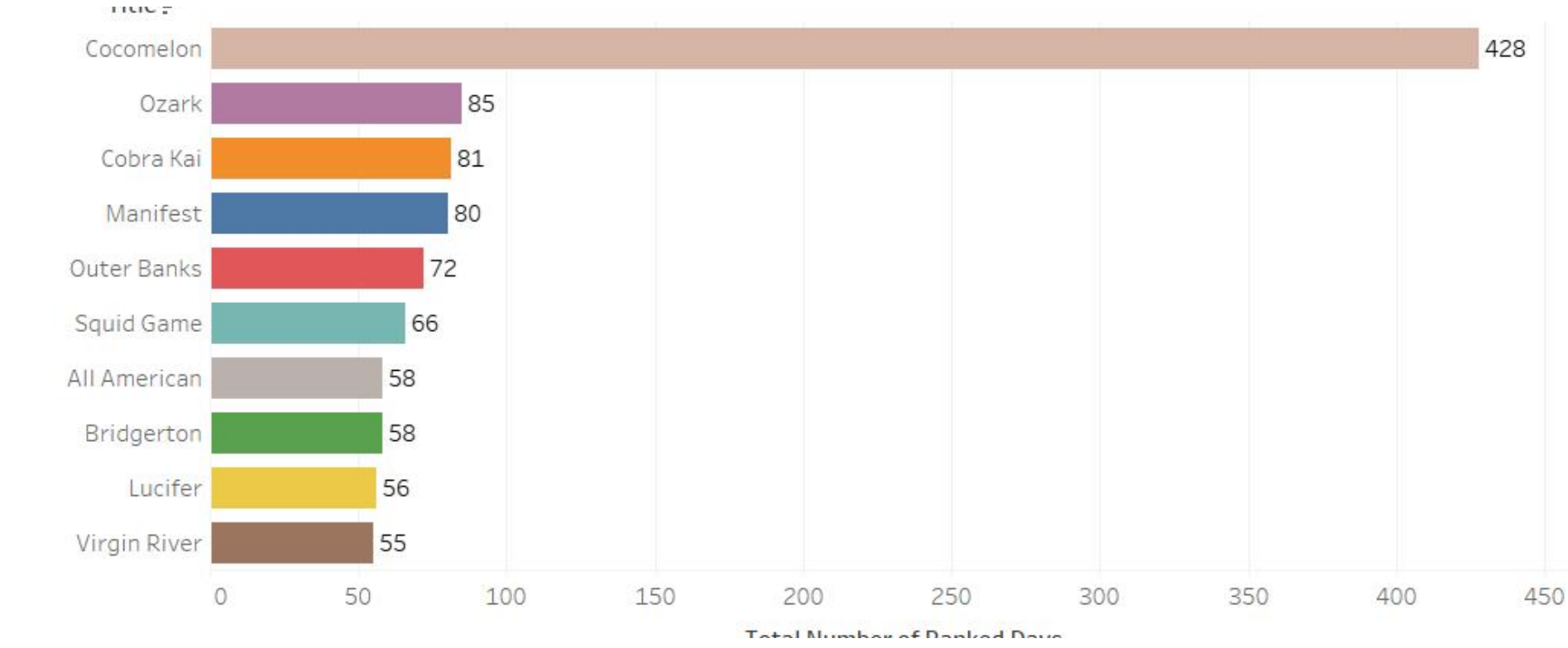
Drawbacks/Future Development:
- Our tool allows for unrealistic scenarios such as:
  - Having no genre at all can often lead to 50% chance of success.
  - Having a unrealistically large number of genres will sometimes lead to 90+% chance of success.
  - Creating a thousand-minute movie will lead to 100% chance of success.

Conclusion:
- For realistic scenarios, our tool does an excellent job of predicting a program's chance of entering the Netflix Top 10.
- Our tool should be used in parallel with the normal creative process for a new program and should not be used as a process replacement.