**Data Management and Data Analytics Capstone Topic Approval Form**

The purpose of this document is to help you clearly explain your capstone topic, project scope, and timeline. Identify each of the following areas so you will have a complete and realistic overview of your project. Your course instructor cannot approve your project topic without this information*.*

**Student Name: Kimberly Valentine**

**Student ID: 001547930**

**Capstone Project Name:** Streaming Platform Analysis

**Project Topic**:

Analyzing Consumer Value Across Streaming Platforms: A Comparison of Netflix, Hulu, Amazon Prime, Apple TV+, and HBO Max

**Research Question:**

Which streaming platform provides the best value for consumers, considering the percentage of high-rated media (IMDb ≥ 7.5), exclusive high-rated content, the average rating of unique content, content overlap, and subscription price?

**Hypothesis:**

Null Hypothesis (H₀): There is no significant difference between streaming platforms in terms of value when considering high-rated media, exclusive content, and subscription price.

Alternative Hypothesis (H₁): Streaming platforms differ significantly in value, with at least one platform providing the best balance of high-rated content, exclusivity, and affordability.

**Context:**

This analysis is relevant to consumers looking to optimize their streaming subscriptions, given the increasing number of platforms and costs. By identifying high-quality content, exclusivity, and overlap, this project will help determine which platform(s) provide the most value for the money.

**Data:**

* **Streaming Data**:
  + Datasets from platforms such as Netflix, Hulu, Amazon Prime, HBO Max, and Apple TV.
  + Parameters that we will use include content type, IMDb ID, and the countries the titles are available in.
* **IMDb Data**
  + Information about titles on IMDb.
  + Parameters we are interested in are IMDB ID, Promotional Title, the Genres, the Average User Rating, and the Number of Votes.
* **Cancellation Rates Data**:
  + Cancellation rates of TV shows on streaming platforms between 2020 and August 2023.
  + Data collected from Statista: [Cancellation rates of streaming and linear TV shows in the United States by platform](https://www.statista.com/statistics/1410363/tv-show-cancellation-rate-us-by-platform/).
  + Logged into streaming\_costs.csv
* **Pricing Data**:
  + Manual data collection from each platform’s website regarding subscription prices.

**Data Gathering:**

* The Streaming Dataset were collected from Kaggle
  + <https://www.kaggle.com/datasets/octopusteam/full-hulu-dataset>
  + <https://www.kaggle.com/datasets/octopusteam/full-apple-tv-dataset>
  + <https://www.kaggle.com/datasets/octopusteam/full-amazon-prime-dataset>
  + <https://www.kaggle.com/datasets/octopusteam/full-netflix-dataset/data>
  + <https://www.kaggle.com/datasets/octopusteam/full-hbo-max-dataset>
* The IMDb Dataset was collected from the IMDb website.
  + <https://developer.imdb.com/non-commercial-datasets/>
* The Cancellation Rate dataset was collected from Statista.
  + <https://www.statista.com/statistics/1410363/tv-show-cancellation-rate-us-by-platform/>
  + This was manually entered into a csv file and saved as cancellation\_rates.csv
* The Pricing Data was manually collected from the each platform website and manually logged into streaming\_costs.csv

**Data Analytics Tools and Techniques**:

**Tools:**

* Jupyter Lab: for organizing data analysis
* Python: For Analysis and Preprocessing
* Tableau and Matplotlib: For visualizations of overlap percentages, unique content ratings, and affordability comparisons.

**Techniques:**

* Statistical comparisons (e.g., two-sample t-tests to compare average ratings).
* Correlation tests (e.g., overlap percentages vs. platform pricing).
* Descriptive analysis (e.g., high-rated vs. low-rated percentages across platforms).
* Regression Models: To estimate how pricing, content availability, and exclusivity influence consumer value.
* Clustering (e.g., K-Means Clustering to group platforms based on features)

**Justification of Tools/Techniques:**

* Python is chosen for its powerful libraries (Pandas for data manipulation, Matplotlib/Seaborn for visualizations, and statistical testing).
* Regression models and hypothesis testing will allow for testing the relationship between platform characteristics (content exclusivity, pricing) and consumer value, thus supporting the research question with statistical significance.

**Application Type, if applicable (select one):**

**☐** Mobile

**☐** Web

**X** Stand-alone

**Programming/Development Language(s), if applicable:**

* Python, Juypter Lab, Tableau

**Operating System(s)/Platform(s), if applicable:**

* Cross-platform (Windows, macOS)

**Database Management System, if applicable:**

* Not applicable (Data stored in CSV files)

**Project Outcomes:**

* A comprehensive analysis comparing streaming platforms based on content availability, exclusivity, and pricing.
* Recommendations on the best value platforms for consumers.
* A final report with insights and visualizations.
* The potential development of a dashboard summarizing key findings.

**Projected Project End Date:**

December 29, 2024

**Sources:**

* **Streaming Platform Datasets**:
  + [Hulu Dataset (Kaggle)](https://www.kaggle.com/datasets/octopusteam/full-hulu-dataset)
  + [Apple TV+ Dataset (Kaggle)](https://www.kaggle.com/datasets/octopusteam/full-apple-tv-dataset)
  + [Amazon Prime Dataset (Kaggle)](https://www.kaggle.com/datasets/octopusteam/full-amazon-prime-dataset)
  + [Netflix Dataset (Kaggle)](https://www.kaggle.com/datasets/octopusteam/full-netflix-dataset/data)
  + [HBO Max Dataset (Kaggle)](https://www.kaggle.com/datasets/octopusteam/full-hbo-max-dataset)
* IMDb dataset (converted from TSV to CSV)
* Statista cancellation rates data: [Statista](https://www.statista.com/statistics/1410363/tv-show-cancellation-rate-us-by-platform/)
* Platform pricing data (gathered manually from each platform's website)
* **Public Research:** Relevant articles and papers that provide insights into consumer behavior regarding streaming subscriptions, pricing, platform comparison, and content preferences (to be used as secondary research or references).

**Human Subjects or Proprietary Information**

Does your project involve the potential use of human subjects? (Y/N):

Does your project involve the potential use of proprietary company information? (Y/N):

**STUDENT SIGNATURE**

**Kimberly Valentine**

**By signing and submitting this form, you acknowledge** that any cost associated with the development and execution of your data analytics solution will be your (the student) responsibility.

**TO BE FILLED BY A COURSE INSTRUCTOR**

**The capstone topic is approved by a course instructor.**

**COURSE INSTRUCTOR’S NAME AND SIGNATURE: **

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**COURSE INSTRUCTOR APPROVAL DATE: 11/24/2024**

**Project Compliance with IRB (Y/N): Y**