**Data Analytics Capstone Topic Approval Form**

*The guidelines for the Data Analytics Graduate Capstone course require a student to demonstrate the "application of academic and professional abilities developed as a graduate student. The capstone challenges students to integrate skills and knowledge from several program domains into one project." As a result, it is highly recommended that your topic* ***should resolve a current or perceived business problem.*** *As stated, you want to exemplify what you have learned in your MSDA program to showcase your skills. Remember, your research topic exemplifies scholarship and research at the highest level and is significant and helpful to potential employers in identifying your abilities. Your research will always be something you can look back on with pride. Finally, it is recommended for students to use publicly available data sets for transparency and external validity.*

*The purpose of this document is to help you clearly state the research question you will be exploring in your capstone project, your project’s scope, and your timeline in order to ensure that these align with your degree emphasis. Without clearly defining each of these areas, you will not have a complete and realistic overview of your project, and it cannot be accurately assessed whether your project will be acceptable for the purposes of these courses. Of course, if this a project that you have already completed at work or elsewhere, this should be easy to fill in! Many students do use a project they have already completed in the past year or two. In that case, you will write the proposal as if the project had not been done yet, and when you report on your project, you will use your complete after-implementation report. If you have not yet done your project, this document can help make sure the scope is within the acceptable range for this capstone. A course instructor will approve this form before submitting this task for evaluation. The task will not be evaluated without a course instructor’s signature. The course instructor may ask for additional information before approving the form.*

*Before submitting this form for approval, please remove all italicized directions in the form.*

***Please only submit a Topic Approval Form that has been signed by a course instructor for evaluation.***

**Student Name:** Laurie Narcisse

**Student ID:** 011392441

**Capstone Project Name:** Multiple Linear Regression on Netflix Content

**Project Topic**: Analyzing the Influence of Attributes on the Popularity of Movies and TV Shows on Netflix

**This project does not involve human subjects research and is exempt from WGU IRB review.**

**Research Question:** How does the presence of certain attributes, such as cast, director, genre, and release year, influence the popularity or viewership of movies and TV shows on Netflix?

**Hypothesis**: **Null hypothesis**-. There is no significant relationship between the presence of specific attributes (cast, director, genre, release year) and the popularity or viewership of movies and TV shows on Netflix. **Alternate Hypothesis**-. The presence of certain attributes (cast, director, genre, release year) significantly influences the popularity or viewership of movies and TV shows on Netflix.

**Context:***.*Understanding the factors that contribute to the popularity or viewership of movies and TV shows on Netflix is essential for content creators, producers, and the platform itself. With the vast library of content available on Netflix, identifying key attributes that attract viewers can inform decision-making processes related to content production, acquisition, and recommendation algorithms. By analyzing data on attributes such as cast, director, genre, and release year, we can gain insights into audience preferences and optimize content selection and promotion strategies.

**Data:** The dataset consists of information on movies and TV shows accessible on Netflix, including attributes such as show ID, type (movie or TV show), title, director, cast, country, date added, release year, rating, duration, listed genres, and description. Each entry in the dataset represents a unique movie or TV show available on Netflix, allowing for comprehensive analysis of various attributes and their impact on popularity or viewership (Netflix Movies & Shows Dataset, 2023).The dataset is located at <https://www.kaggle.com/datasets/ashfakyeafi/netflix-movies-and-shows-dataset?resource=download>

**Data Gathering:** The data will be collected from the provided dataset, which contains detailed information on movies and TV shows available on Netflix. No additional data collection is required as the dataset covers a wide range of attributes relevant to the research question.

**Data Analytics Tools and Techniques**: The appropriate data-analysis technique for this project would be regression analysis, specifically multiple linear regression. Multiple linear regression allows us to examine the relationship between multiple independent variables (such as cast, director, genre, release year) and a continuous dependent variable (popularity or viewership) by estimating the coefficients of the regression equation.

**Justification of Tools/Techniques:** Multiple linear regression is suitable for this analysis because it can identify the relative importance of each attribute in influencing the popularity or viewership of movies and TV shows on Netflix. By estimating the coefficients for each independent variable, we can quantify their impact and determine which attributes have a significant effect on viewership.

The planned tools to use include Python and Jupyter Notebooks.

**Project Outcomes**: Key anticipated outcomes and deliverables include:

* Identification of attributes that significantly influence the popularity or viewership of movies and TV shows on Netflix.
* Insights into audience preferences based on attributes such as cast, director, genre, and release year.
* Development of predictive models to forecast the potential viewership of new content based on its attributes.
* Recommendations for content creators, producers, and Netflix to optimize content selection and promotion strategies.
* A final report summarizing findings, conclusions, and actionable insights for stakeholders in the entertainment industry.

**Projected Project End Date**: 3/31/2024

**Sources**:

Netflix Movies & Shows Dataset. (2023, December 12). Kaggle. <https://www.kaggle.com/datasets/ashfakyeafi/netflix-movies-and-shows-dataset?resource=download>

Blokhin, A. (2023, August 13). Linear vs. Multiple Regression: What's the Difference? Investopedia. <https://www.investopedia.com/ask/answers/060315/what-difference-between-linear-regression-and-multiple-regression.asp>

Statistics Solutions. (2021, August 19). What is Multiple Linear Regression? - Statistics Solutions. <https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-multiple-linear-regression/>

GeeksforGeeks. (2023, December 20). R vs Python. <https://www.geeksforgeeks.org/r-vs-python/>

Kumar, V. (2021, December 11). Python vs R: What’s best for Machine Learning - towards data science. Medium. <https://towardsdatascience.com/python-vs-r-whats-best-for-machine-learning-93432084b480> Mikoliuk, E. (2023, December 7). How to compare SAS vs Python for data analysis. MarketSplash. <https://marketsplash.com/tutorials/python/sas-vs-python/>

**Course Instructor Signature/Date:**

The research is exempt from an IRB Review.

An IRB approval is in place (provide proof in appendix B).

Course Instructor’s Approval Status: Approved

Date: Click here to enter a date.

Reviewed by:

Comments: Click here to enter text.