**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:** Nicole Haibach

**Student ID:** 001260374

**Capstone Project Name:** Reader Preferences

**Project Topic**: Romance And Action Reader Trends



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

**Research Question:** Between Action and Romance genres, which genre has the highest average rating and number of reviews across the decades?

**Hypothesis**: **Null hypothesis**-There is no statistically significant difference between the mean rating scores (or number of reviews) of books in the Action genre and the Romance genre. **Alternate Hypothesis**- There is a statistically significant difference between the mean rating scores (or number of reviews) of books in the Action genre and the Romance genre.

**Context:** Analyzing which genre, **Action** or **Romance**, has the highest average rating and number of reviews across decades benefits from data analysis as it enables objective and systematic exploration of reader preferences and trends over time. By leveraging metrics like ratings and reviews, we can uncover significant insights about audience engagement, identifying whether one genre consistently outperforms the other or if preferences shift across decades. Statistical methods ensure the differences are meaningful, offering actionable insights for authors, publishers, and marketers. Temporal trends also provide context, reflecting how societal and cultural shifts influence genre popularity, making the analysis both relevant and insightful.

**Data:** *The Top Goodreads Books Collection (1980-2023) dataset contains detailed information about popular books listed on Goodreads over four decades. It includes attributes such as book titles, authors, genres, publishers, languages, and publication dates, alongside key engagement metrics like average rating scores, number of ratings, reviews, current readers, and "want to read" counts. Additional data on book formats, series information, page counts, descriptions, and prices provide further insights. This dataset is rich in quantitative and qualitative data, making it ideal for analyzing genre performance (e.g., Action vs. Romance) and understanding trends in reader preferences, engagement, and popularity across decades.*

The data for the Top Goodreads Books Collection (1980-2023) is sourced from Kaggle, an online platform for data science competitions and projects. Kaggle datasets are typically shared by contributors who aggregate, clean, and format publicly available data or their own curated collections. In this case, the dataset is likely derived from publicly accessible information on Goodreads, an online platform owned by Amazon that provides book-related data like ratings, reviews, and metadata.

**Data Gathering:** *The Top Goodreads Books Collection (1980-2023) dataset will be sourced from Kaggle, where it has been curated from publicly accessible information on Goodreads. The dataset includes attributes like book ratings, reviews, genres, publication details, and reader engagement metrics. After downloading the dataset from Kaggle, the data will be verified for integrity by reviewing its structure, completeness, and relevance to the research question, as well as cross-referencing the documentation for clarity on its compilation. Data cleaning steps, such as addressing missing values and standardizing fields, will be performed to ensure it is suitable for analysis. If additional information is needed, the Goodreads API may be used to supplement the dataset, adhering to ethical guidelines and platform terms of service. This methodology ensures the dataset is ready for accurate and reliable analysis.*

**Data Analytics Tools and Techniques**: *To analyze the Top Goodreads Books Collection (1980-2023) dataset and compare the performance of Action and Romance genres, several data-analysis techniques will be applied. Initially, descriptive statistics will summarize key metrics such as average ratings, number of reviews, and book counts for each genre, providing a clear overview of the dataset. The data will then be grouped by genres (Action and Romance) and publication\_date (organized into decades) to calculate averages and totals for rating\_score and num\_reviews. This aggregation will help identify any significant differences or patterns across time periods.*

*To ensure statistical validity, hypothesis testing will be conducted. A two-sample t-test will compare the mean rating scores and number of reviews between the two genres to determine if the observed differences are significant. If the data does not meet normality assumptions, a non-parametric test like the Mann-Whitney U test will be used. Additionally, trend analysis will explore how ratings and reviews have evolved for both genres over decades, visualized through line graphs, bar charts, and box plots. These techniques will highlight genre performance over time and provide insights into shifting reader preferences. Combining statistical tests with clear visualizations ensures a robust and interpretable analysis.*

**Justification of Tools/Techniques:**

*The chosen data-analysis techniques are appropriate because they align with the structure of the dataset and the research question, ensuring meaningful insights are derived. Descriptive statistics provide a foundational understanding of key metrics such as average ratings and number of reviews, summarizing the dataset's overall trends and helping identify preliminary differences between the genres. By grouping data by genres and publication\_date (organized into decades), the techniques also allow us to capture temporal trends and compare the performance of Action and Romance books over time, which is central to the research question.*

*Hypothesis testing, such as the two-sample t-test, is crucial for determining whether differences in mean ratings and reviews between the genres are statistically significant, ensuring results are not due to random chance. For non-normally distributed data, the Mann-Whitney U test provides a reliable alternative. Additionally, trend analysis and visualizations like bar charts, line graphs, and box plots help highlight patterns and differences in genre performance over decades in an interpretable manner. These techniques together provide a comprehensive, statistically valid, and visually clear analysis tailored to the dataset and research objectives.*

**Project Outcomes**: *The project will deliver a detailed analysis comparing the average ratings and number of reviews for Action and Romance genres, highlighting trends and shifts in reader engagement across decades. Key outcomes include statistically validated insights into which genre performs better, identification of temporal trends in ratings and reviews, and actionable recommendations for authors, publishers, and marketers. Deliverables will include a comprehensive report with findings, visualizations like bar charts and line graphs to illustrate key trends, a concise executive summary, a reproducible codebase for transparency, and a professional presentation summarizing the analysis and its implications. These outputs will provide valuable, data-driven insights into genre performance and reader preferences.*

**Projected Project End Date**: 1/15/25

**Sources**:

Cristaliss. (2023). Ultimate Book Collection: Top 100 Books Up to 2023 [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/cristaliss/ultimate-book-collection-top-100-books-up-to-2023

**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.