**Research Proposal: Analyzing the Competition Between Netflix and Disney in the Streaming Industry**

**Background and Motivation:**

The rapid growth of digital streaming platforms has transformed the entertainment industry, reshaping how consumers access and engage with media. Among the dominant players in the global streaming market, Netflix and Disney+ have significantly influenced content consumption patterns by adopting distinct business strategies. Netflix prioritizes original programming and AI-driven personalization (Cadicamo, n.d.), while Disney+ capitalizes on its strong intellectual property portfolio and bundling strategies (Li, n.d.; Solsman, 2024). This research investigates how these differing approaches impact subscriber growth, retention, and competitive positioning, using the Netflix vs Disney dataset from Kaggle as a key resource (<https://www.kaggle.com/datasets/prathamjyotsingh/netflix-vs-disney>).

The significance of this study lies in its potential to contribute to both social science and machine learning research. From a social science perspective, understanding how streaming services attract and retain subscribers sheds light on broader consumer behavior trends in the digital entertainment sector. As more platforms enter the market, insights into the efficacy of different business models, pricing strategies, and content offerings become valuable for industry stakeholders, policymakers, and consumers. From a machine learning standpoint, analyzing content personalization algorithms, recommendation systems, and user engagement patterns offers a deeper understanding of how data-driven strategies shape audience retention and platform success (Cadicamo, n.d.).

Despite extensive research on digital streaming, gaps remain in understanding how business strategies influence consumer decision-making over time. While prior studies have examined the individual strengths of Netflix and Disney+, few have directly compared their long-term effects on subscriber retention using comprehensive datasets. Additionally, existing literature primarily focuses on single-platform analysis rather than cross-platform comparisons. This study aims to bridge this gap by assessing how factors such as pricing models, bundling strategies, and content exclusivity affect subscriber loyalty across different demographic groups (Solsman, 2024; Wolfe, 2024).

By integrating insights from social science and machine learning, this study seeks to advance our understanding of the competitive dynamics in digital streaming services. Investigating how Netflix and Disney+ navigate challenges such as "subscription pausing" (Wolfe, 2024) and bundling incentives (Solsman, 2024) will provide valuable implications for both the entertainment industry and future research in data-driven consumer analytics.

**Research Question:**

In recent years, streaming platforms like Netflix and Disney+ have engaged in intense competition for market dominance, adopting distinct strategies to attract and retain subscribers. While previous studies have analyzed individual factors contributing to their success, there is a gap in understanding how content strategies—such as original programming, franchise power, and genre diversity—interact with pricing models to influence subscriber growth and retention across different demographic groups.

**Research Question:** How do the content offerings (original programming, genre diversity, and franchise presence) and pricing strategies of Netflix and Disney+ influence subscriber growth and retention across various demographic groups (age, region, and income) from 2018 to 2024?

This question is relevant to both social science and machine learning. From a social science perspective, understanding how consumer demographics interact with content and pricing strategies provides insights into digital media consumption trends, platform loyalty, and market competition. From a machine learning standpoint, this research can contribute to advancements in recommendation systems, predictive modeling of subscriber behavior, and personalized content strategies that drive engagement and retention.

The dataset from Kaggle offers detailed information on Netflix and Disney+ subscriber growth, content offerings, and other relevant factors over multiple years. By leveraging this data, we can conduct a comparative analysis of how different content strategies—ranging from algorithm-driven personalization (Netflix) to franchise-based exclusivity (Disney+)—influence consumer behavior. Additionally, by examining the effectiveness of pricing models, including tiered subscriptions, bundling, and ad-supported plans, this study will provide a more nuanced understanding of the evolving digital streaming landscape.

**Application Scenarios:**

This study is situated within the digital entertainment industry, specifically focusing on the streaming media sector. The dataset from Kaggle, which provides insights into Netflix and Disney+ subscriber growth, content offerings, and pricing strategies, is well-suited for analyzing the competitive dynamics of these platforms. Given the increasing reliance on machine learning for personalized recommendations and social science frameworks for understanding consumer behavior, this research has applications in several key areas.

First, streaming platforms like Netflix and Disney+ can use the findings to refine their content strategies and pricing models based on demographic preferences. By understanding how factors like original programming, franchise presence, and genre diversity influence subscriber growth and retention across different consumer segments, these companies can optimize content curation and marketing strategies to enhance engagement and reduce churn.

Additionally, marketing and advertising firms within the entertainment sector can leverage these insights to develop targeted promotional campaigns. By analyzing how content strategies interact with demographic variables such as age, region, and income, businesses can craft personalized outreach efforts, maximizing customer acquisition and retention.

Beyond streaming services, the study’s findings hold relevance for related industries such as gaming, digital media, and e-commerce, where consumer engagement patterns are crucial for content development and strategic decision-making. Understanding how different pricing models and bundling strategies affect consumer loyalty can inform subscription-based services beyond entertainment, such as cloud gaming platforms and digital publishing.

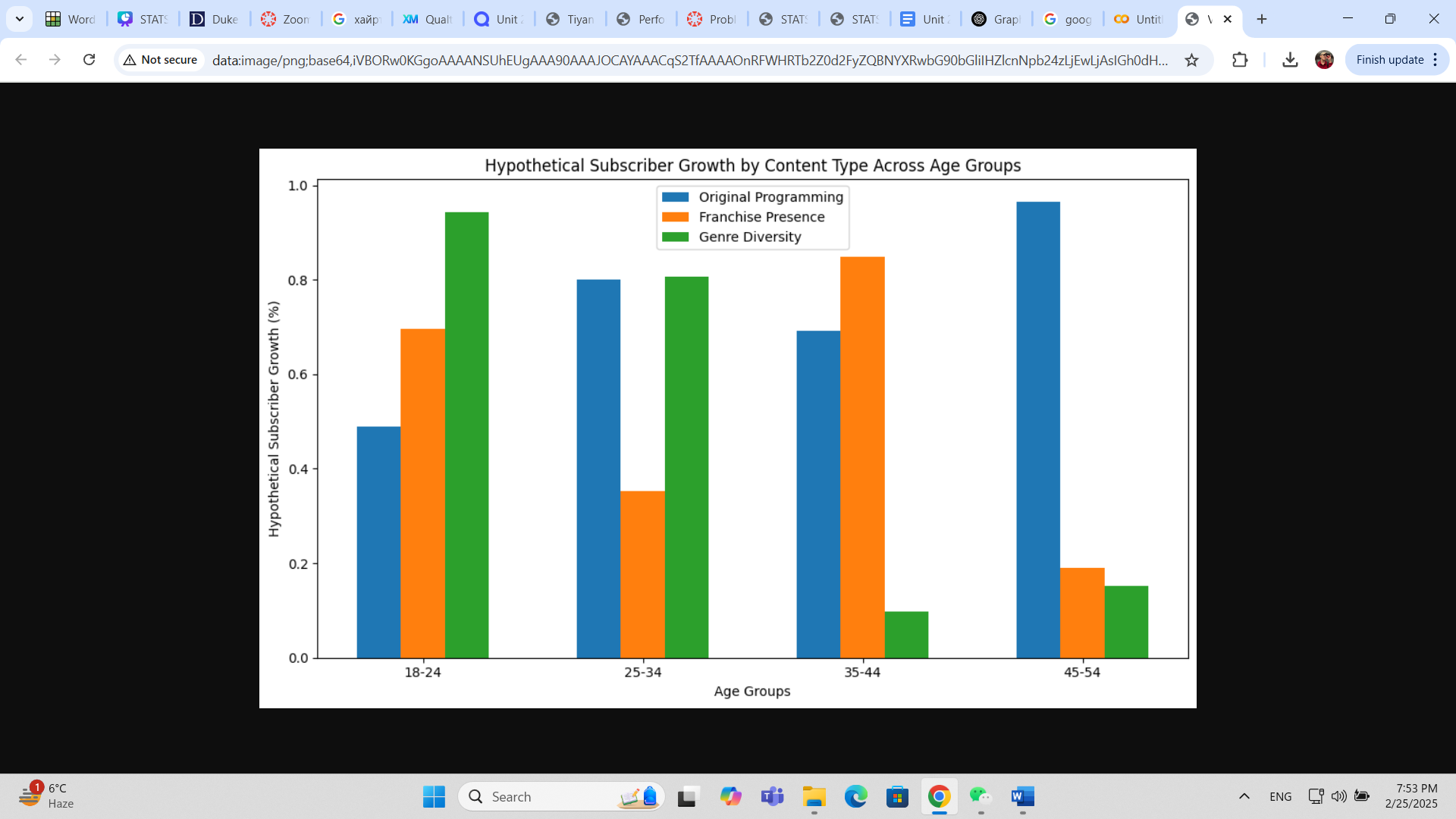


Figure 1: Hypothetical Subscriber Growth by Content Type Across Age Groups

From a policy perspective, insights from this study could aid regulators in assessing the impact of content exclusivity, pricing transparency, and consumer rights in the digital streaming landscape. As subscription models evolve, policymakers may use this research to develop fair competition guidelines, ensuring a balance between corporate interests and consumer welfare.

Furthermore, the role of Generative AI (GenAI) in predicting content trends and optimizing recommendation algorithms is an emerging aspect of the streaming industry. Tools like ChatGPT can assist in forecasting audience preferences, generating content recommendations, and even aiding in content creation. However, considerations such as bias in training data and the need for diverse representation must be addressed to ensure fairness and accuracy in AI-driven recommendations.

**Methodology:**

This study employs a combination of statistical and machine learning techniques to examine the relationship between content offerings, pricing strategies, and subscriber behavior between Netflix and Disney. First, the data will undergo cleaning and preprocessing, addressing missing values, standardizing categorical variables, and normalizing numerical features such as subscriber counts. This ensures the dataset is ready for analysis and focuses on relevant variables like content types, pricing strategies, and demographic factors.

**Natural Language Processing for Model Explanation**

To deeply examine how user sentiment and discourse shape subscription behaviors, this study will incorporate Natural Language Processing (NLP) techniques to analyze user-generated content, including reviews and social media discussions. By applying sentiment analysis, text classification, and topic modeling, this approach will explain how public perception of content offerings and pricing strategies correlates with subscriber trends.

The analysis will involve preprocessing textual data through tokenization, stop-word removal, and lemmatization, ensuring the extraction of clean and meaningful linguistic features. Sentiment analysis will be conducted using VADER (Valence Aware Dictionary and sEntiment Reasoner), a lexicon-based technique designed for social media and short-form reviews. Additionally, Latent Dirichlet Allocation (LDA) will be employed for topic modeling to identify dominant themes in audience discussions, while FastText, an efficient word embedding model, will be utilized for classifying user reviews based on content themes.

To enhance interpretability, sentiment trends over time will be focused more, enabling the identification of shifts in user perception across different demographic groups. Topic modeling results will be analyzed to uncover variations in audience reception, providing a nuanced understanding of how content and pricing influence consumer engagement. These NLP techniques will contribute to the study by offering a qualitative lens through which streaming service decisions can be examined, complementing quantitative findings from predictive modeling.

**Supervised Machine Learning for Prediction**

To predict subscriber growth and retention, this study will employ Random Forest, an ensemble-based supervised learning algorithm known for its robustness in handling high-dimensional data with mixed feature types. This model is particularly well-suited for integrating both numerical and categorical variables, making it an ideal choice for analyzing how content strategies and pricing influence user behavior across demographic groups.

The model will be trained on a dataset comprising variables such as content type (original versus licensed), genre diversity, franchise presence, pricing tiers, and demographic factors, including age, region, and income levels. To improve model generalizability and prevent overfitting, k-fold cross-validation will be applied, ensuring that the predictive performance remains stable across different subsets of data. Model evaluation will rely on key performance metrics, including accuracy, precision-recall, F1-score, and Receiver Operating Characteristic – Area Under the Curve (ROC-AUC), to assess the model’s effectiveness in classifying user retention patterns.

To optimize predictive performance, feature importance analysis will be performed to determine which variables exert the greatest influence on subscription trends, clarifying the relative impact of pricing, content type, and genre diversity. As a complementary approach, ensemble learning techniques, such as Gradient Boosting (e.g., XGBoost or LightGBM), will be explored to assess whether boosting models enhance predictive accuracy beyond what Random Forest can achieve.

**Causal Inference using Machine Learning: Regression Discontinuity Design**

In addition to predictive modeling, this study employs Regression Discontinuity (RD) Design to establish a causal relationship between pricing strategies and subscriber retention. Streaming platforms frequently adjust their pricing models, and these changes often influence user behavior. However, the degree to which pricing affects retention remains an empirical question. This study focuses on a specific price increase by Netflix as a natural experiment to determine whether subscription fee hikes causally impact user retention.

To implement the RD design, the analysis will leverage a price increase threshold as the treatment assignment mechanism. Netflix periodically raises subscription fees, and these changes create a discontinuity in consumer decision-making. Users who subscribed before the price increase serve as the control group, while those subscribing afterward form the treatment group. Comparing retention rates immediately before and after the price threshold isolates the causal impact of price hikes on subscriber churn.

The running variable in this design is the monthly subscription price, which varies for users based on when they subscribed. The treatment assignment occurs at a predefined cutoff, where the price increase takes effect. The outcome variable of interest is subscriber retention, measured as the probability of maintaining a subscription for at least six months post-subscription. As for the cutoff point, it is the exact price level where new subscribers started paying a higher fee. Since users just above and below this threshold are likely similar in observable and unobservable characteristics, any discontinuous change in retention rates at this point can be attributed to the pricing change rather than external factors.

To formally estimate the causal effect of pricing on retention, the study applies a local linear regression model, regressing subscriber retention probability on subscription price while allowing for different slopes on either side of the price threshold. The identification strategy relies on the assumption that consumers do not selectively sort themselves around the price increase threshold, ensuring that observed changes in retention rates are driven by the price change rather than other confounding factors.

Following standard RD methodology, the analysis proceeds through four key stages. First, a causal model is specified, detailing the relationship between the price variable, the treatment indicator, and the outcome variable. Second, the appropriateness of RD for causal identification is justified, emphasizing the homogeneity of treatment assignment at the cutoff. Third, the treatment effect is estimated by measuring differences in retention rates on either side of the price threshold using bandwidth selection methods that optimize local treatment effect estimation. Finally, robustness checks are performed to validate the causal claims, including placebo tests that shift the cutoff to arbitrary points and falsification tests to ensure that no pre-existing trends violate RD assumptions.

By incorporating Regression Discontinuity Design, this study extends beyond correlation and predictive analysis to establish causal relationships between pricing decisions and consumer retention behavior. This methodological addition strengthens the research by addressing not only what factors influence subscriber retention but also whether pricing changes actively drive consumer decisions. This distinction is particularly relevant for streaming platforms and policymakers seeking to understand the implications of subscription pricing strategies on user engagement and long-term retention.

**Justification and Contribution to Social Science**

The integration of Natural Language Processing (NLP), Supervised Machine Learning, and Causal Inference allows this study to examine the research question from multiple analytical perspectives: explanatory, predictive, and causal. NLP techniques provide qualitative insights into audience sentiment and thematic discourse, revealing how consumers perceive and respond to content and pricing strategies. Supervised machine learning, particularly Random Forest modeling, offers a quantitative approach to identifying the key drivers of subscriber behavior across demographic groups. Complementing these methods, the Regression Discontinuity (RD) Design establishes a causal relationship between pricing strategies and retention, determining whether price increases actively drive consumer decisions rather than merely correlating with churn rates.

From a social science perspective, this research contributes to the study of media consumption behavior and subscription economics, further explaining how entertainment preferences evolve among different demographic segments. The inclusion of causal inference provides policy-relevant evidence on how pricing structures influence user retention. This has significant implications for streaming platforms, content creators, and policymakers, offering data-driven insights that can inform pricing models, content strategies, and regulatory decisions.

The combination of these methodologies provides a holistic framework for understanding the competitive dynamics of digital streaming platforms, reinforcing the role of data science in shaping real-world business strategies and consumer policy.

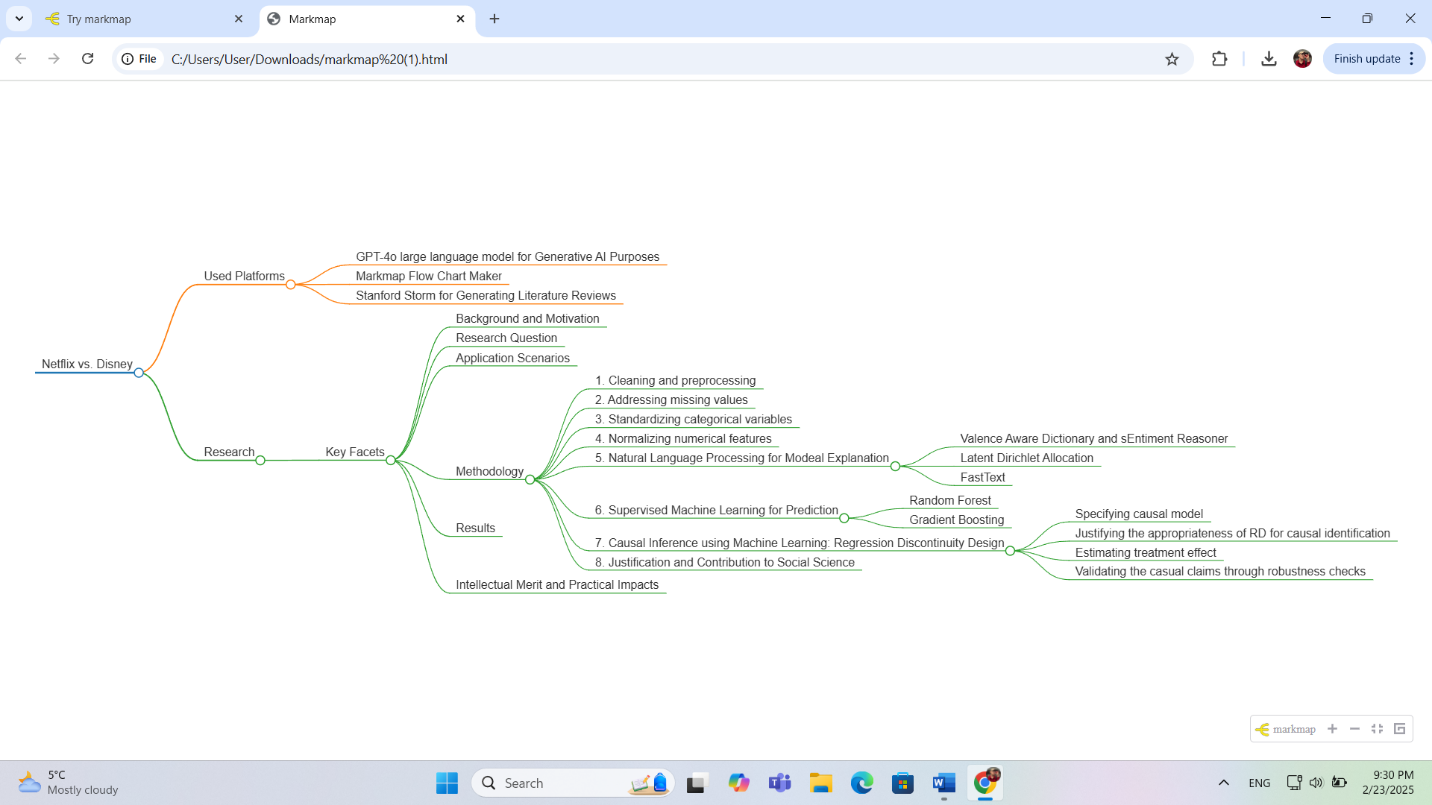


Diagram 1: Netflix vs Disney Research Plan

**6. Results:**

Through this analysis, I expect to uncover detailed insights into how different content strategies (such as original programming, franchise use, and genre variety) and pricing models impact subscriber growth and retention for both Netflix and Disney. Specifically, one of the key contributions of this research will be a deeper understanding of the intersection between content strategy, pricing models, and demographic factors in the streaming industry. For example, sentiment analysis could reveal how different age groups or regions respond to genre diversity, franchise presence, and the introduction of new pricing tiers, offering valuable insights into the emotional and cognitive drivers behind subscriber retention. The findings could inform decisions on content production, marketing strategies, and subscription pricing, potentially leading to more targeted and efficient approaches for subscriber growth.

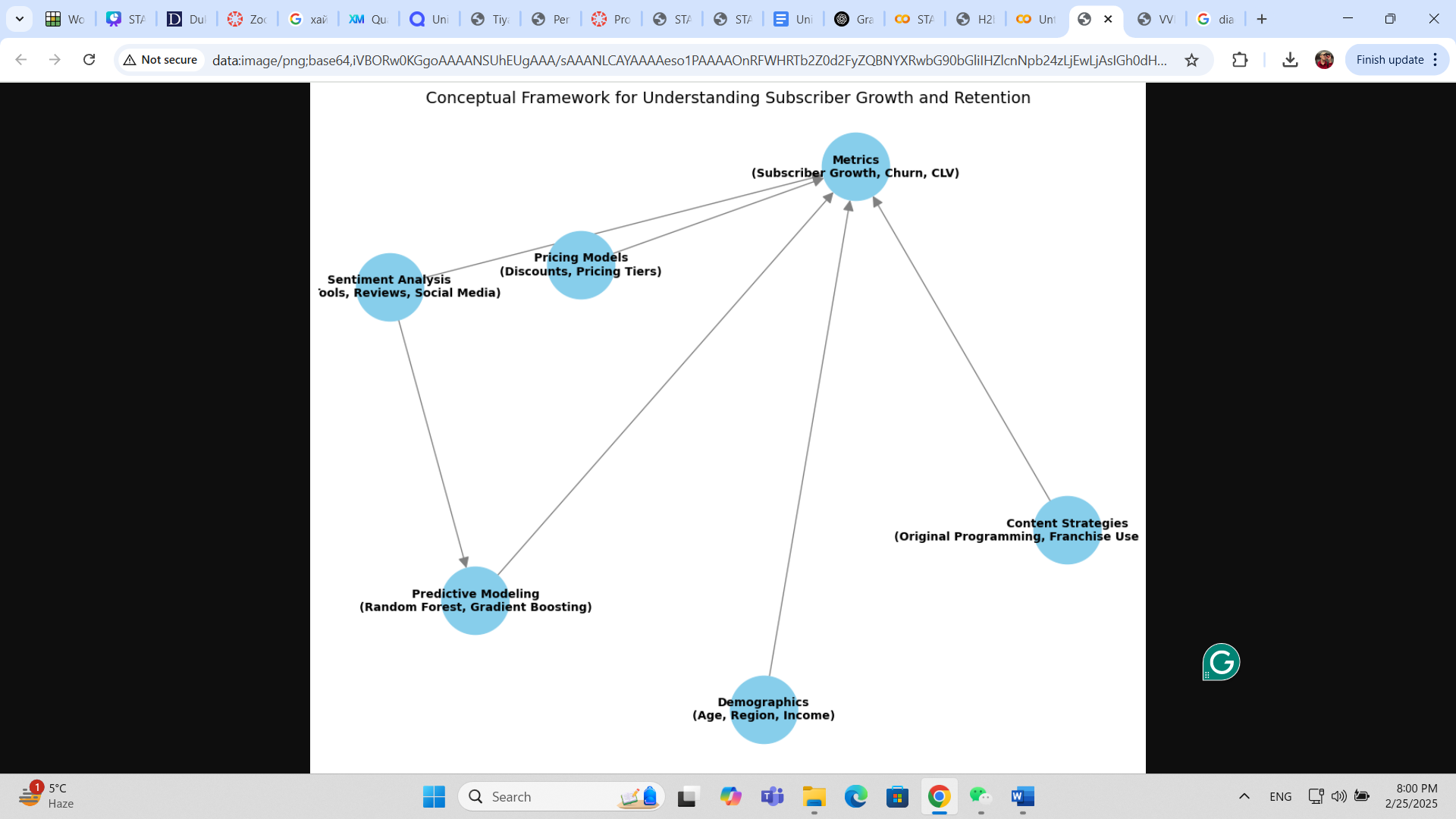


Diagram 3: Conceptual Framework for Understanding Subscriber Growth and Retention

The preliminary findings of this study derive from the regression discontinuity analysis conducted on the price increase by Netflix. The analysis focuses on the impact of this price hike on subscriber retention, comparing retention rates immediately before and after the price increase.

The results indicate a clear discontinuity in subscriber retention at the cutoff point, the price increase threshold. Specifically, subscribers who joined Netflix before the price increase exhibited higher retention rates compared to those who subscribed after the hike. This suggests that the price increase led to a decrease in the probability of subscribers maintaining their subscriptions.

**Google Colab Link to Simulated RD Analysis:** <https://colab.research.google.com/drive/1BYankPDZ6_INgdxm51co9IM_xgLTu5-H?usp=sharing>

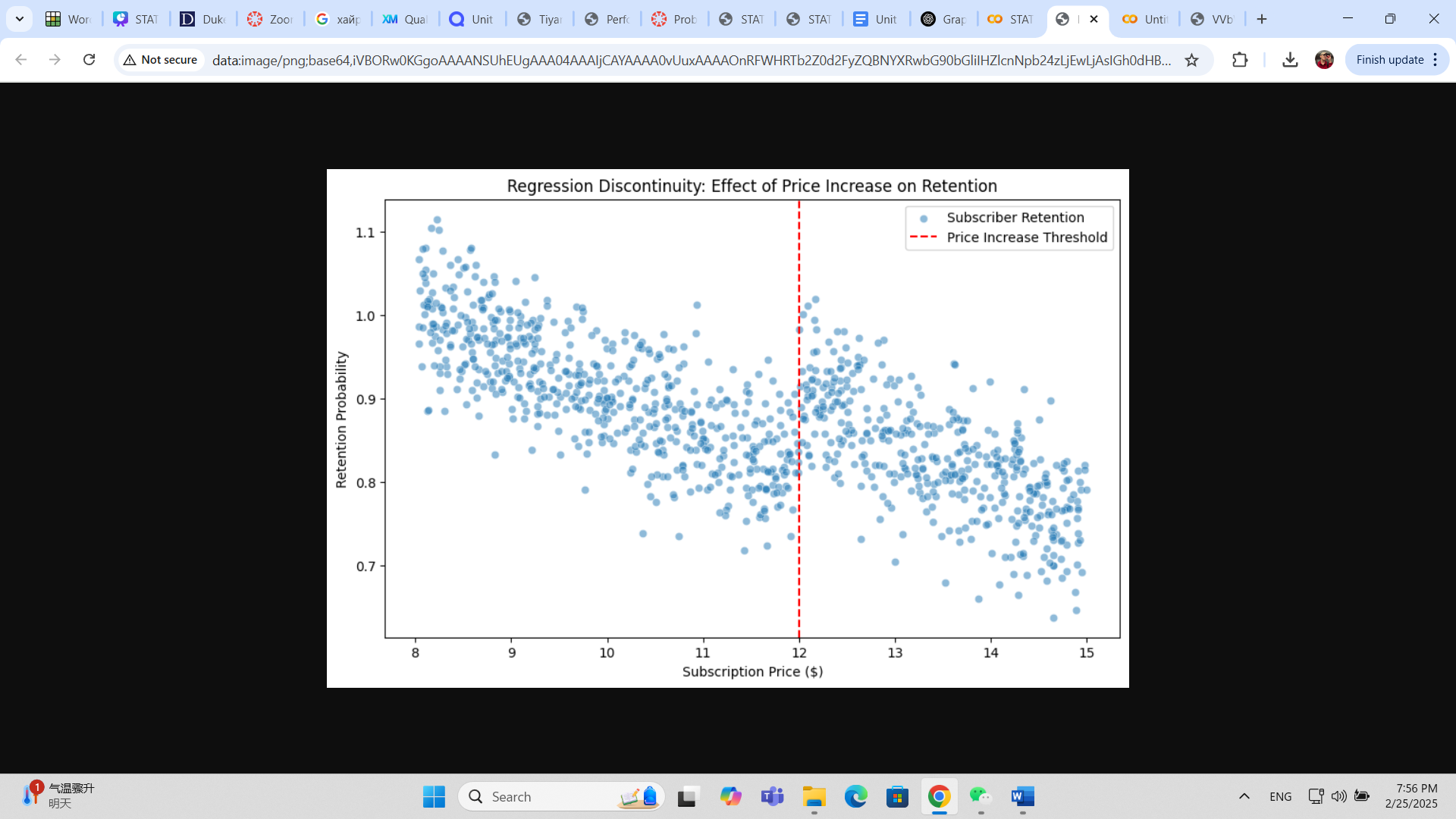


Figure 2: Regression Discontinuity of Price Increase Effect on Subscriber Retention (Through Simulated Data)

**Intellectual Merit:**

This research makes a significant contribution to the academic field by providing a comprehensive analysis of how content strategies and pricing models influence subscriber behavior across various demographic groups. It fills a gap in the existing literature by combining data from both Netflix and Disney, offering insights into the dynamics of the streaming industry, which is often studied in isolation or with limited scope. The integration of content strategies with demographic data provides a more holistic and nuanced understanding of consumer behavior in the digital entertainment sector. Furthermore, the research employs innovative multimodal analysis techniques, such as Natural Language Processing (NLP) for sentiment analysis and topic modeling, and machine learning models like Random Forest and Gradient Boosting. This approach enhances the understanding of how different content types (original, licensed, genre diversity, and franchise presence) resonate with various demographic groups. By incorporating causal analysis through Regression Discontinuity (RD) Design, the study extends beyond correlation to uncover causal relationships, advancing the methodological tools used in the field of social science and digital media research.

While this study provides valuable insights into the intersection of content strategies, pricing models, and subscriber behavior, it also opens the door to several future research avenues. For instance, the use of machine learning models like Random Forest and Gradient Boosting could be expanded to incorporate more complex data types, such as user interactions with content or detailed viewing behavior. Additionally, the integration of NLP for sentiment analysis and topic modeling presents opportunities to explore more granular aspects of consumer sentiment, such as emotional responses to specific genres or content types. Furthermore, the limitations of the current study, such as its focus on Netflix and Disney, suggest the need for comparative studies across other platforms (e.g., Amazon Prime, Hulu) to broaden the understanding of subscriber behavior. Future research could also explore how external factors, such as global events (e.g., the COVID-19 pandemic), influence subscription trends and content preferences. The study's focus on causal analysis using Regression Discontinuity Design provides a foundation for future investigations into the causal impact of other pricing strategies or content-related changes on subscriber retention.

**Practical Impacts:**

**Societal Benefits**

The findings of this research carry significant societal implications, particularly in the context of the digital entertainment industry. As streaming platforms like Netflix and Disney+ continue to dominate the global media landscape, understanding the interplay between content strategies, pricing models, and consumer demographics can help these platforms better serve diverse audiences. By optimizing content offerings based on demographic preferences and pricing sensitivity, platforms can create more inclusive and tailored viewing experiences, which may increase consumer satisfaction and loyalty. Moreover, this research underscores the importance of fair and transparent pricing models, which can enhance consumer trust and reduce the risk of churn. The ability to predict subscriber behavior and retention will help companies reduce operational inefficiencies, thereby ensuring that resources are allocated toward content and pricing strategies that have the greatest impact.

**Applications**

In industry, the insights gained from this study can help streaming platforms refine their content curation and pricing strategies. By analyzing demographic segments’ preferences for original programming, franchise content, and genre diversity, Netflix and Disney+ can fine-tune their offerings to attract and retain subscribers. This research can also guide marketing strategies, enabling platforms to design campaigns that target the right consumers with relevant content and promotional offers. Additionally, content creators and production companies can benefit from a better understanding of which types of content resonate with specific audience groups, allowing them to tailor future productions to meet consumer demand more effectively.

The findings also have applications in public policy, particularly in regulating the digital entertainment market. As subscription-based models continue to evolve, policymakers can use this research to design fairer pricing strategies, ensuring that subscription services are accessible to a broader audience without unfairly disadvantaging lower-income groups. Furthermore, insights into the effects of exclusivity in content offerings could inform regulations aimed at promoting competition and protecting consumers from monopolistic behavior in the digital media space.

**AI Governance and Ethical Considerations**

This research also touches upon crucial considerations in AI governance, particularly as it pertains to machine learning models and their impact on society. One of the primary ethical challenges in the development and deployment of AI models in streaming platforms is ensuring fairness and transparency. The use of supervised machine learning algorithms such as Random Forest and Gradient Boosting, alongside Natural Language Processing (NLP) techniques like sentiment analysis, can be prone to biases if not carefully managed. To address this, the project ensures that training data is diverse, representative, and free from biased sources, thereby minimizing the risk of perpetuating harmful stereotypes or unfair treatment of marginalized groups.

Additionally, by using NLP and topic modeling to analyze consumer sentiment, this project emphasizes the need for inclusivity in AI development. By accounting for the varying tastes and preferences of different demographic groups, the research supports the creation of personalized content recommendations that reflect a diverse set of voices and values, ensuring that no group is unfairly excluded or marginalized. The research also adheres to AI governance principles by prioritizing transparency in its methodology and actively mitigating potential biases in data collection and model training. This approach fosters a more inclusive and ethically sound AI-driven system for predicting subscriber behavior.

In line with sustainable development goals (SDGs), the project promotes responsible consumption and production patterns by encouraging companies to focus on creating content that aligns with the values of different consumer groups, reducing waste in content creation, and improving the overall media consumption experience. Additionally, by ensuring that pricing models are fair and accessible, the research contributes to promoting inclusive economic growth (SDG 8) and reducing inequalities (SDG 10), fostering a more equitable and accessible media landscape.

The long-term prosperity and societal well-being of this project lie in its ability to help streaming platforms adapt to evolving consumer needs, contributing to both the economic success of these companies and the satisfaction of their diverse user bases. By promoting transparent and fair business practices, this research can help foster a digital media environment that prioritizes user trust, inclusivity, and sustainable growth, ultimately benefiting consumers, businesses, and policymakers alike.

**GitHub Repository**

<https://github.com/Rising-Stars-by-Sunshine/Khatanbuuvei_STATS201_Final>

**References:**

* Cadicamo, A. (n.d.). *Netflix, Disney+ and audience retention*. Toast Studio. Retrieved from https://www.toaststudio.com/en/articles/netflix-disney-and-audience-retention/
* Li, Z. (n.d.). *Game Theory Analysis of Competitive Dynamics in the Streaming Industry: A Comparative Study of Netflix and Disney*. ResearchGate. Retrieved from <https://www.researchgate.net/publication/383958316_Game_Theory_Analysis_of_Competitive_Dynamics_in_the_Streaming_Industry_A_Comparative_Study_of_Netflix_and_Disney/fulltext/66e29468b1606e24c224ca4d/Game-Theory-Analysis-of-Competitive-Dynamics-in-the-Streaming-Industry-A-Comparative-Study-of-Netflix-and-Disney.pdf>
* Solsman, J. E. (2024, October 14). *Bundle of Disney+, Hulu and Max Has a Strong Hold on Viewers. It's Even Stickier Than Netflix*. The Wall Street Journal. Retrieved from https://www.wsj.com/business/media/bundle-of-disney-hulu-and-max-has-a-strong-hold-on-viewers-its-even-stickier-than-netflix-8a4d7d91
* Wolfe, A. (2024, November 25). *A New Streaming Customer Emerges: The Subscription Pauser*. The Wall Street Journal. Retrieved from https://www.wsj.com/business/media/a-new-streaming-customer-emerges-the-subscription-pauser-430a4eef