Bryan Rogers

Bellevue University

DSC630: Predictive Analytics

Course Project

Reducing Customer Churn in a Subscription based service (Netflix)

Problem Statement:

High customer churn in Netflix user accounts. Netflix and other subscription streaming services are at risk for a high turnover in accounts and can sometimes struggle with account retention. With countless options for content and entertainment in this current environment, it seems that streaming services are constantly trying to keep up with each other to retain the most subscribers.

Importance/Interest:

This issue of retaining subscribers and reducing customer churn is important for streaming services in a variety of ways. Losing subscribers can contribute to loss of current and future revenues but can also affect how streaming services long term relationships and referrals for future customers. I am interested in this topic and learning more about customer retention because, I am a current subscriber of a handful of streaming services, and I am considering consolidating and removing a few of them. I am curious to see how

Netflix handles their customer retention and what they are doing to keep and acquire new subscribers compared to how other streaming services accomplish this. I think using predictive analytics are a crucial in this environment, as the streaming service landscape is ever evolving and constantly shifting. Through predictive analytics can find and determine solutions for business growth, a better customer experience, gaining a competitive advantage, and over all having better subscriber retention and growth.

Target Audience:

The target audience of this project would be a few different groups of people within the corporate offices of Netflix. These include executive leadership, who would be interested in revenues, growth, and sustainability. Marketing teams for Netflix and other streaming services would show interest in this report as they are responsible for customer retention and acquisition costs and optimization. Customer support is another target since they deal with customer satisfaction and are also key in retention. Finally, Research & Development is a big target with them being responsible for finding insights and efficiencies for consumer trends and behavior as well as ways to enhance both the product and the subscriber experience.

Data Source:

There are a handful of data sources that I will attempt to find and use. I have a found a Kaggle dataset on Netflix subscription fees that I will be using as well as a Kaggle dataset on Netflix user information dataset. In addition to these sources, I will be looking for

feedback and survey responses as a possible data source as well market research and customer support records.

Data Usefulness:

Using the data that will be acquired will be very helpful in solving the issue of customer churn and retention because it is subject to predictive analytics modeling. Looking at historical customer data as mentioned above can give us a better insight into how to best keep customer retention and potentially have customer growth over time. We can use predictive analysis and forecast modeling to predict future customer churn and retention and be able to identify higher risk cases. Using the data allows to create targeted strategies to develop solutions for at-risk customers. The data provides us with the ability to create informed decisions to and initiate customer retention strategies.

Types of Models:

I will attempt to use various methods of forecasting models to achieve my predictive analysis, including Neural Networks, Random Forrest modeling, and Logistic Regression. Neural Networks will provide deep learning models that can show patterns in data and allow us to target trends. Random Forrest will allow us to understand some of the complex relationships created and be able to identify key performance indicators in our datasets. Logistic Regression will be very helpful to run forecasts and predict future performance allowing us to influencing factors for churn.

Evaluation Metrics:

Accuracy and precision will be the main driver in evaluation of the metrics and dataset. Historic trends will also be used to evaluate and lend to forecasting decisions in the model.

Learning Objectives:

There are a few things that I hope to accomplish and learn in my findings. I think the main thing that I am looking to understand is the key performance indicators and what are the biggest drivers for customer churn and growing retention. I would like to learn how to create optimized strategies based on the key predictors and maximize retention and minimize churn. I would like to understand better how to create an efficient model that runs continuous that delivers productive results.

Risks and Ethical Concerns:

There are a few ethical concerns that we must worry about when dealing with large amounts of customer data. The biggest risk would be privacy to protect the consumer against data intrusion and other things of said nature. We must also account and check for bias to ensure that our data is clean, correct, and provides quality results. I think transparency is another viable concern as predictive analysis must ensure that the consumer is comfortable with their data being looked at and explaining the reasoning and purpose of the analysis.

Contingency Plan:

If for some chance, it looks like my proposal will not go according to plan, or if the data that I am working with doesn't meet the guidelines or requirement, I will have to go in another direction. I will evaluate the quality and quantity of the data that I will be using and continue to research if there are other datasets available. I will see if there are other graphs and models that I could use to better represent the data if my original models don't project an accurate representation of the data. I will be monitoring my timeline and plan, and if it feels like I am not satisfied with the direction of the project, I will make a pivot quick enough to allow enough time to succeed in my contingency plan.

Beginning of Milestone 3:

Will I be able to answer the questions I want to answer with the data I have?

I think being able to answer the questions that I am posing with the data I have been able to collect will come down to three main points or targets. These 3 points would be data adequacy, data quality, and model fit. With data adequacy, it will come down to if the data is able to provide significant coverage of relevant factors in relation to customer churn. If it turns out that I am missing information, I will have to dig around for additional dataset that provide relevant data. After addressing data adequacy, going through data quality is the next important step. This can be done by scouting for missing values, inconsistences, and outliers. Double checking if the data aligns with the proposed business problem is key and finding additional data sources if needed. After looking at data quality, assessing the

models that I am using are a good fit for the dataset and they can provide solutions for the inquiry. Adjusting the model may be needed if the model doesn't align with the dataset.

What visualizations are especially useful for explaining my data?

I think there are many visualizations that I could use to explain my data for showcasing customer churn and how to reduce churn in streaming services. I will create journey maps based on the customer, by showing visualizations to identify potential issues or points of interest that is leading to churn. These customer journey maps will be highlighted by visualizations such as bar and line charts or heatmaps that highlight the most pivotal points affecting churn in streaming services. Using confusion matrices will be key in using visualizations to identify points of interest in churn. Showing true positives and negatives as well as false positives and negatives could show a clear picture of how the model is performing to assess the problem.

Do I need to adjust the data and/or driving questions?

I think it is too early in the EDA process to make critical adjustments or my data selection. It may be beneficial to make small tweaks to some of the driving questions that I have proposed earlier. I think the adjustments are dependent on certain factors, such as, is data performing the way it should, is the data able to be transformed into visualizations to answer the driving questions? I could potentially create new features with the data or modifying current data features to better provide a clearer picture and enhance the patterns and relationships in the model. I think the most important and simplest

adjustment to make would be to search and add additional data sources to the model and adding a more robust data profile to get a better picture of customer behavior. After running the initial EDA and once I start to generate results, that would be the best time to start to consider adding small or large tweaks to the driving questions stated at the beginning of this project. These decisions will be made based off new insights discovered and additional thoughts that come up through the EDA.

Do I need to adjust my model/evaluation choices?

Making the adjustments in model and evaluation choices will also depend on how the EDA process goes and if I feel that my model is not trending the way I think it is supposed to be trending. If needed, I can implore hyperparameter tuning, where I can tweak and adjust the model in a finely tuned manor using the performance metrics and allow the model an opportunity to have optimal accuracy in predicting results. Adding ensemble methods can be a helpful tool to leverage strength of having a model with multiple algorithms if needed. If I must adjust with the evaluation choices, I will consider looking at the metrics that I am using to evaluate customer churn and adjust based on the priorities that Netflix is looking at when trying to prevent customer churn. Having false positives and negatives could show outliers and present results that could lead Netflix down the wrong direction and end up with more churn that originally predicted.

Are my original expectations still reasonable?

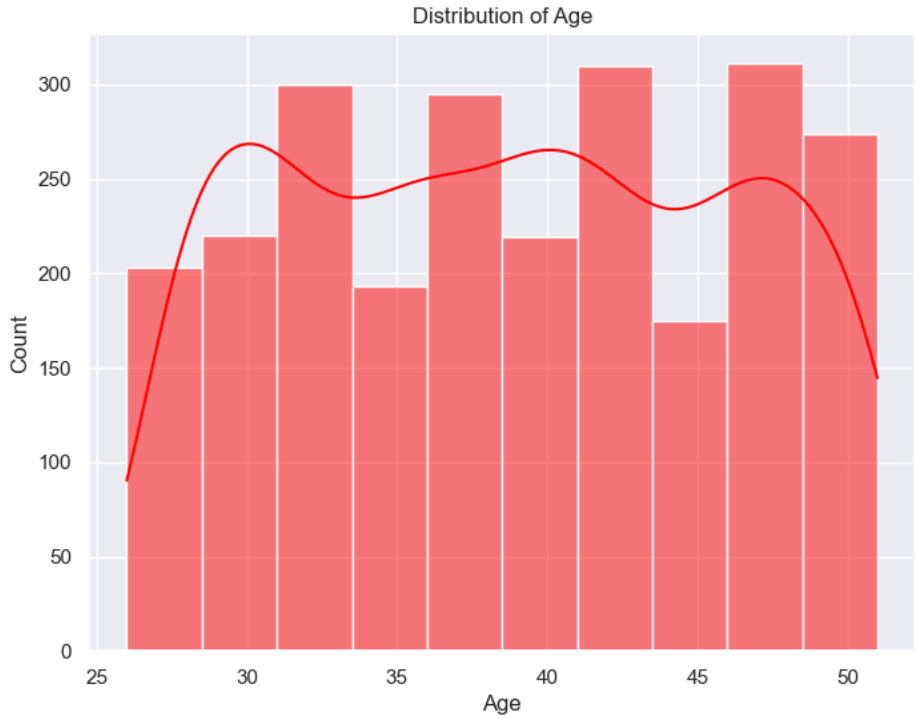
I think after beginning the EDA process, I would say that the original expectations are still reasonable considering that the results of model could go in any direction and there is no true right answer, since it is a predictive analysis study. The data exploration is just going to show an answer and more questions would need to be answered after the initial results are shown. This would require additional research and more detailed thought and modeling. I think, with this being a study on predictive analytics, some iteration would be needed, and I might consider refining the approach of continued EDA and additional ongoing analysis.

Milestone 4:

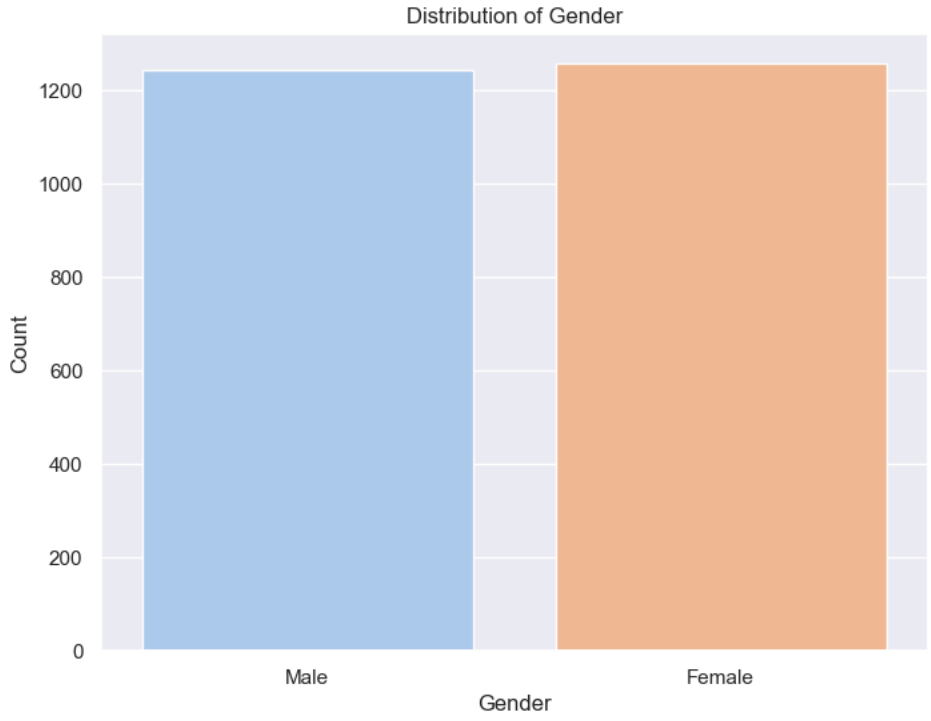
Data

After collecting the dataset required and cleaning and prepping the data for exploration, it is time to look at what metrics we are looking for to prevent customer churn and look at the results of the Exploratory Data Analysis. First, I think it important to establish what are the driving metrics that cause customers to leave and look at some of the demographics of our data. I have created some plot and graphs that show our customer base, based off age, gender, and user location.

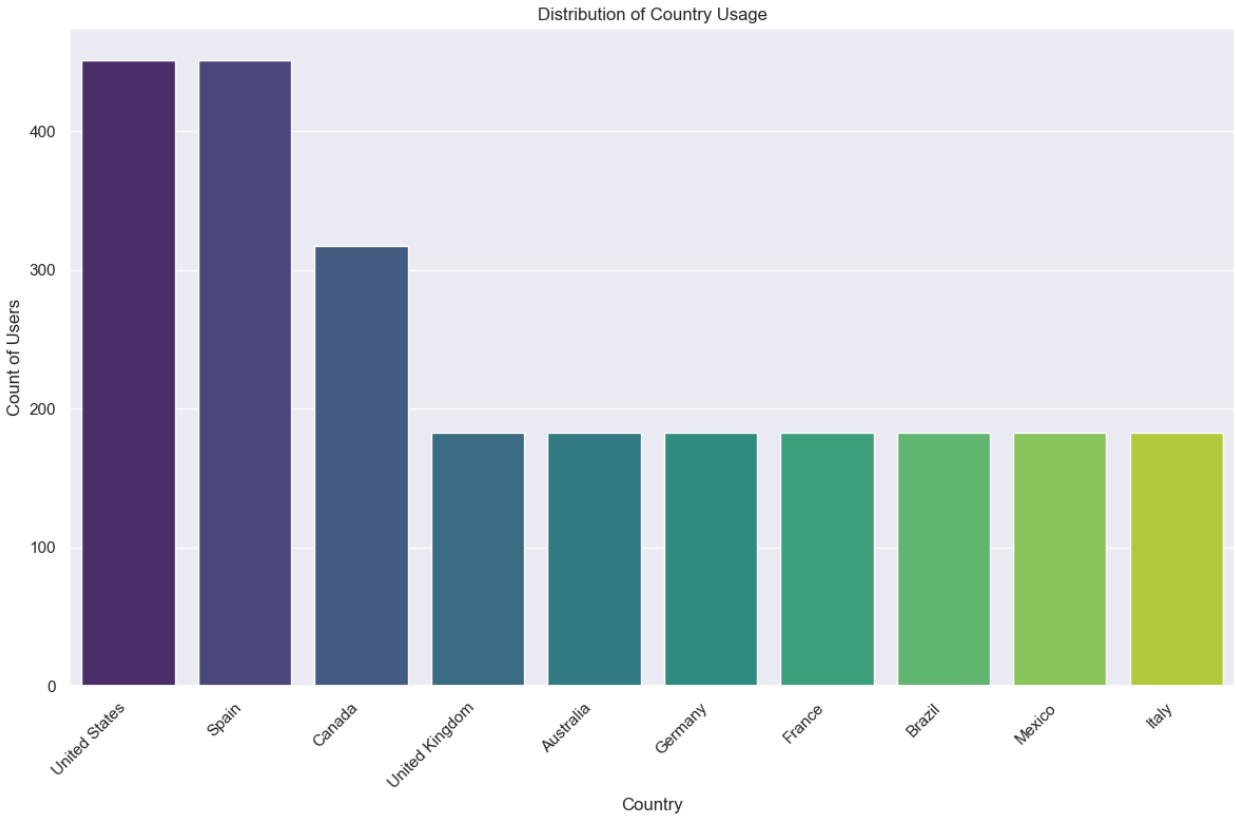
As you can tell from our first plot of age distribution, the average age of the Netflix user is right around the 40 years old mark, with 25–35-year-old and 35-45-year-old age groups making up most of the sample size.



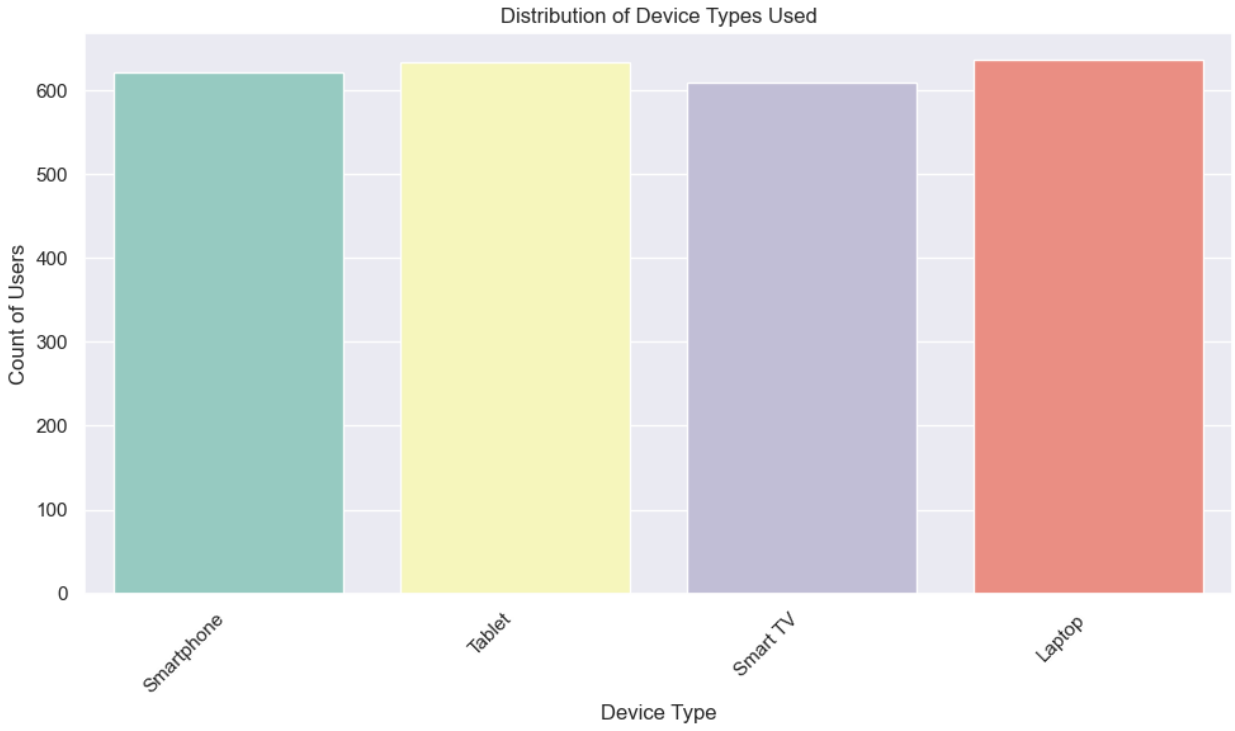
In the second plot, I am using a bar chart to determine gender distribution. It looks like there is an even distribution of male and female users in this grouping, which is not surprising, as there is expected to be with the sample size.



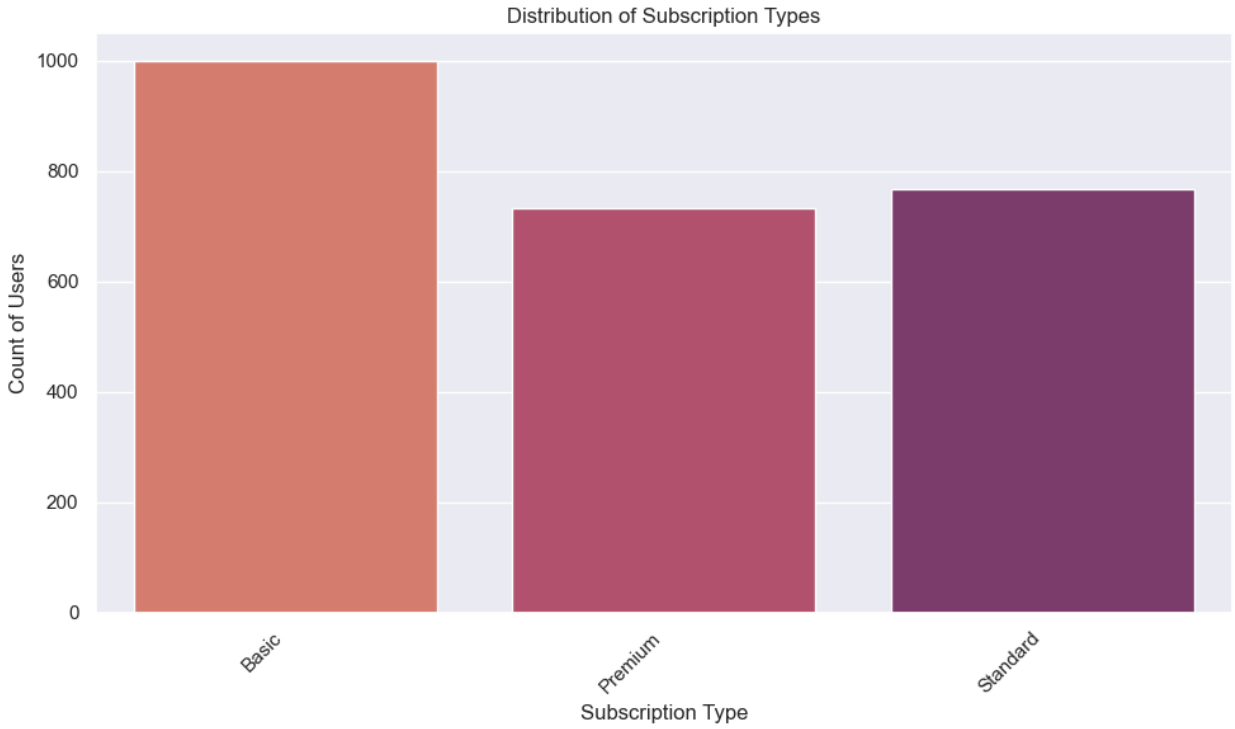
In our third plot, we can see that the largest amount of users comes from the United States, which is also to be expected as Netflix is U.S based and U.S population would dictate that they would have the majority share of users. I am surprised that Spain is the second highest country and is almost equal to the United States.



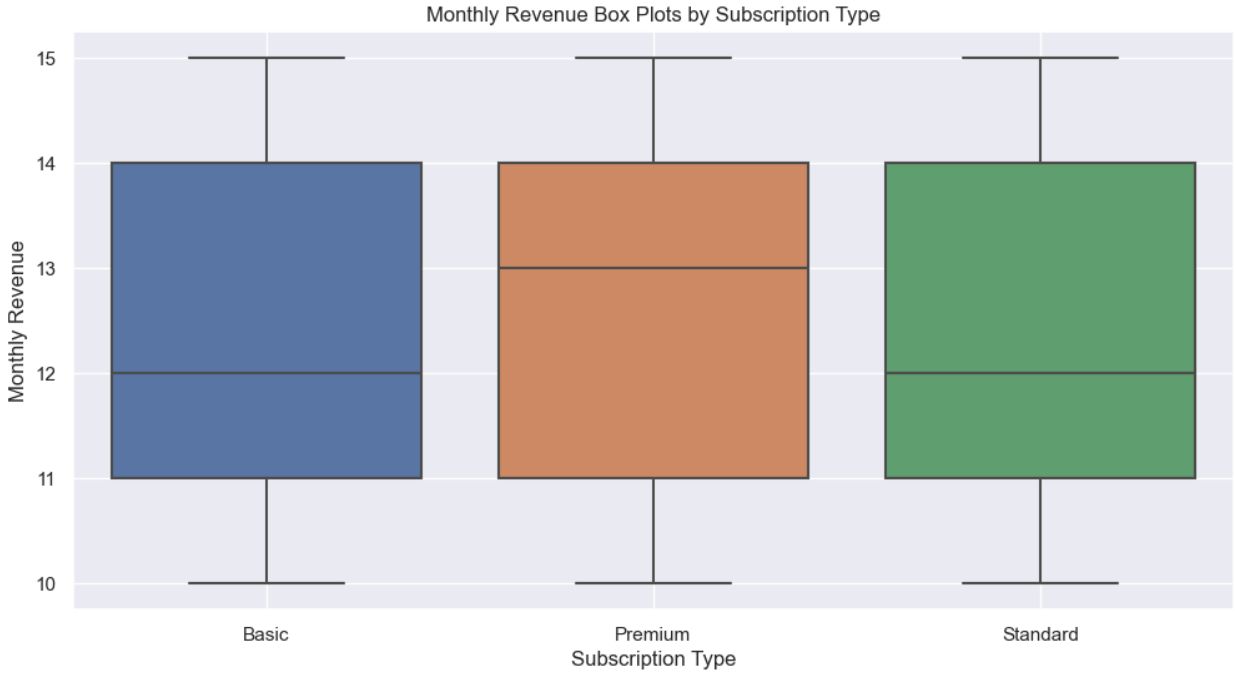
Next, we will look at how our users are using Netflix, by looking at what type of devices that our users are using. From our plot, it looks like there is an even distribution among the four platforms. I am a little shocked that Smart TV is the least used of the four, which most users are streaming their content to mobile devices.

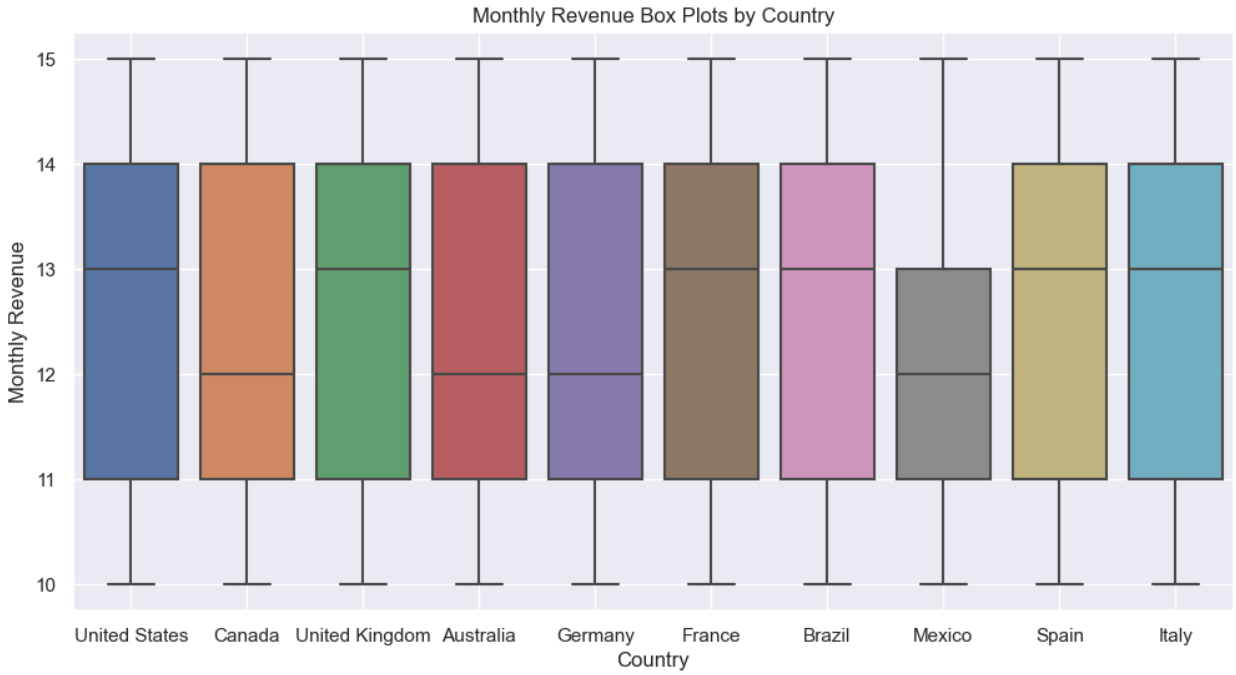


After determining how our users are streaming their content, we will look at what subscription packages are being used, as to get a better picture of users most likely to churn or leave. From the plot based on subscription package, we can see that the Basic plan is the most used, which is a sign of potential churn.



After looking at how users are streaming their content, it is time to look at revenue that Netflix is bringing to determine what the percentage of churn is a possibility.





As you can tell from the above graphs, there is a lot to look at in terms of monthly revenue. First, we will go further into revenue by subscription type, and the graph tells us that the 'Premium' subscription brings in the most revenue, which makes sense since it is the highest costing plan. When looking at revenue based on device, we can see that there is

even split amongst all devices and there is no real impact to monthly revenue. Finally, looking at the box plot of monthly revenue based on country, we can see that there is once again, an even split amongst the countries. This could be because the cost of Netflix plans is relatively uniform across the globe.

After examining all the data, it is time to run a model to predict customer churn. I was unable to come up with a successful model to accuracy predict customer churn. I tried running Random Forests as well as Logistic Regression, and I was hit with countless errors with my dataset not being able to be converted from timedate and I was not able to run any model successfully. I was able to create use the 'Join Date' and 'Last Payment Date' columns to determine what a churn rate would be, and it came out to 0.0016. I don't believe that is an accurate representation of what customer churn should be according to documentation from sources that state the customer churn for Netflix is around 2.5% (Gokul,2023). I would have been able to get more concrete results and made a better prediction had my model ran successfully.

Before concluding, I would need to see more results and some accurate predictive analysis, to make any recommendations to a board as far steps that can be taken to reduce customer churn. Based on the results I was able to come up with, I would say that there is nothing to improve on and not a worry about customers leaving, but that is a false statement.

References:

Arnav. (2023, July 4). *Netflix userbase dataset*. Kaggle. https://www.kaggle.com/datasets/arnavsmayan/netflix-userbase-dataset

Gokul. (2023, December 1). *The secrets of netflix recommendation system revealed - learn how Netflix manages such low churn rate*. The Secrets of Netflix Recommendation System Revealed - Learn how Netflix Manages such Low Churn Rate. https://www.argoid.ai/blog/how-does-netflixs-recommendation-engine-manage-low-churn-rate#:~:text=Netflix%20masters%20customer%20retention%20with%20lowest%20churn%20rate,-OTT%20players%20today&text=With%20more%20than%20247.15%20million,continues%20to%20constantly%20optimize%20content.