Nick Bamont, Samara Othmani, Austin Simpson, Christian Watson,

Henry Holt & Karla Carmona

Duncan Houldsworth

MGSC 410-01

12 May 2023

Final Group Project

**Action Plan & Division of Work**

We conducted a thorough analysis of the team's strengths and weaknesses during our initial meeting, with a particular emphasis on each individual's analytical capabilities. Following that, we discussed the task of cleaning the datasets,and determining the best team members for the task in hand. We decided that Austin and Karla would be in charge of cleaning the datasets, while Henry would concentrate on converting the currencies into a standardized substance. Furthermore, we discussed each team member's ability to answer questions using code and visualization, ultimately deciding that Austin would handle questions 1 and 2, while Karla and Christian would collaborate on question 3. Christian would handle question 5, while Henry and Nick would work together on question 4. Samara would be in charge of creating Tableau visualizations and consolidating all of the information on the presentation slides and supporting documents. We acknowledged the importance of consistent communication and collaboration after finalizing our work division to ensure the timely and accurate completion of our project. As a result, we planned regular meetings to go over our progress and share our findings, enabling us to make informed choices and address any potential issues.We hope that by strategically allocating roles and responsibilities, we will be able to optimize our workflow and improve our analytical capabilities, allowing us to provide high-quality outcomes.

Austin Simpson

* Initial Data Cleaning
* Questions 1 & 2

Nick Bamont

* Question 4
* Summary Statistic & Visualization

Henry Holt

* Standardizing Currencies
* Question 4

Christian Watson

* Question 3 & 5
* Outside Research

Karla Carmona

* Data Cleaning
* Question 3
* Slide Deck/Presentation

Samara Othmani

* Data Visualization
* Slide Deck/Presentation
* Word Document

**Analytical Plan: Rosetta Stone**

1. **Research Questions**
   1. Who are the most valuable subscribers?
   2. What subscriber segments exist in the data?
   3. Who are the most likely subscribers who could be sold additional products or services?
   4. What is the subscriber profile of those not continuing with their usage of the product and which barriers can deepen subscriber engagement?
   5. What business relevant opportunities are present from your analysis of the data?
2. **Data Cleaning**

The first step in our data cleaning procedure included a comprehensive visual assessment of the dataset to identify any abnormalities that needed to be corrected. The App Activity dataset was our first focus, and we proceeded to drop all null values. The initial stage for the Subscriber Information dataset was to import the data and execute relevant type conversions, such as Boolean, Category, Int/Float, and Datetime conversions. Furthermore, after extensive investigation, we have taken careful measures to account for severe outliers in the 'Purchase Amounts' variable by editing them properly. Moreover, we used one-hot encoding/dummying techniques to convert categorical data into numerical representations, allowing for more effective analysis. Furthermore, we used one-hot encoding/dummying techniques to convert categorical data into numerical representations. Finally, we used Z-scoring to normalize variables, allowing for more accurate comparison and understanding of the information.

1. **Feature engineering**

In regards to feature engineering, we transformed raw data to new data features that would help with our data analysis. First, for the sake of consistency and comparability, we standardized the 'Currency' variable to USD, which unifies the monetary values across the dataset. Furthermore, we changed the 'Country' variable from self-reported to a more precise currency-based location. In addition, we created a new variable, 'Subscription Length' , in days by subtracting the 'Subscription Start Date' data from the 'Subscription Expiration Date' value. Lastly, to better understand which users have stopped providing value we created a new binary variable, ‘Churn.’ To create this variable we looked at the user ‘Subscription Expiration,’ ‘Auto Renew,’ and ‘App Activity Type.’ ‘App Activity Type’ was aggregated by user. If the user’s subscription had expired, based on last subscription start date, or if their auto renewal was turned off and their app activity was under 5 times over the timeframe of the data set they were considered churned.

1. **Tools**

We carefully analyzed which analytical tools would be most appropriate to our project's objectives in order to enhance our analytical capabilities and streamline our process. We found that GitHub would be the best platform for storing all of the datasets in one place, while Google Colab would be the best site for collaborative coding work. In addition, we used Visual Studio as our major coding tools and Python as our primary programming language. We used Rguroo's skills to create summary statistical data visualizations, while Tableau supplied us with advanced data visualization tools. Other applications used included Microsoft Excel and Google Drive, which were useful in creating the PowerPoint deck and accompanying documents. Finally, we used the internet to obtain a better understanding of Rosetta Stone, allowing us to improve our analysis and provide complete insights. We aim to optimize our analytical capabilities and deliver high-quality results by utilizing a diverse range of analytical tools and resources. Tools are listed below:

* GitHub
  + <https://github.com/Austin-Simpson/MGSC-410>
* Google Colab
* Visual Studio
* Python
  + Pyplot (visualization), Ggplot, Pandas, Sklearn, etc.
* Rguroo
* Tableau
* Microsoft Excel
* Google Drive
* Internet

1. **Model Techniques & Metrics**

* Rather than perform specific modeling tasks to address the question, the conclusions of objective 1 were informed from the subscriber analysis conducted throughout the entirety of the project.
* The need to distinguish subscriber segments in Question 2 led to the use of clustering models to identify which user groups exist within the data. Three methods of clustering were tested with various input parameters. First, a K-Means clustering method was ran utilizing all of the features in the dataset. As expected when modeling with so many variables, this model performed poorly overall and led to the exploration of further dimensionality reduction methods. Thus, a Principal Component Analysis was performed to extract the most important features that explained the top percent of variance in the data. With the results of a successful PCA, the identified principal components were then fed into future iterations of clustering models. A hierarchical agglomerative and spectral clustering were both run with these principal components, and both performed similarly in that there was uneven distribution of users among the clusters. Commonly, one cluster was much more populated than the rest, and while this ended up also being true for the final clustering model, both hierarchical and spectral clustering resulted in this to an extreme degree. Besides being more imbalanced, the final clustering method used to segment users ended up being the K-Means clustering done with PCA because the previous two methods were much more computationally intensive (resource consumption was especially high in regard to RAM.). Fine tuning the model and the creation of a distortion score chart allowed the use of the elbow method to determine how many clusters should be made in the final implementation of K-Means.
* In regards to business objective 3, we planned to utilize the results of two models to inform our conclusions. In order to find out which subscribers could be influential in the future, we wanted to look at variables such as whether the user is an email subscriber, whether they have push notifications on, and whether they have auto renew on. Auto Renew is a variable that we would like to predict as it shows the consumer commitment to the company, as well as the likeness of the products/services offered. First, a logistic model was created with 13 different predictors ranging from boolean values to continuous values. Of these predictors, being an email subscriber, having push notifications on, and being a free trial user explains the majority of the variance within the auto renew variable. The coefficients of free trial user and push notifications are both around 2.3, showing that these two variables have the strongest effect on whether a customer has auto renew on. Additionally, a decision tree model utilizing the features extracted from the Principal Component Analysis was created to further supplement these conclusions. This tree model performed well overall, and utilizing the feature importance function we extracted the same top parameters as resulted from the logistic model. Being an Email Subscriber, Free Trial User, and having Push Notifications on continued to be top features in predicting the Auto Renew variable, as supported by the two well-performing models.
* Question 4 - To identify the subscriber profile of those not continuing with their usage of the product we created the 'Churn' variable (See Feature Engineering). The churn variable was used as the dependent variable in a logistic regression with independent variables:

'Purchase Amount,' 'Send Count,' 'Open Count,' 'Purchase Store', 'Demo User', 'Free Trial User', 'Email Subscriber', 'Push Notifications', 'App Session Platform.'

Independent Variables were turned into dummy variables where appropriate. We evaluated the model using an ROC curve, which can be seen in Figure 7. The coefficients are their values can be seen in Figure 16. The variables with positive coefficients, indicating more likely to churn, are those that purchased on the web, those that are email subscribers, those with higher email send counts, and those that spent more purchasing the product. The variables with negative coefficients, indicating less likely to churn, are those with a higher click and open count, those that use the app on IOS or the web, and those that were demo or free trial users. Metrics used to judge the accuracy model can be seen in Figure 17.

* As we looked more closely at Question 5, it became clear that merely looking at the dataset would not be sufficient to provide a complete understanding of the situation facing the company. It was essential to look into the business' external environment and market trends in order to produce insightful findings that really address the issues at hand. In order to achieve this, we suggested a comprehensive research initiative to assess the company's strengths and weaknesses and to analyze the competitive landscape. Data on customer retention rates, marketing initiatives, and levels of customer satisfaction are some of the key performance indicators that we plan to gather. Additionally, we realize the significance of not only examining the company's current operations but also identifying possible opportunities for development and improvement. We plan to identify new market prospects and prospective areas for product development by researching industry trends and customer behavior, which will help the company obtain a competitive advantage. With this information, we plan to create a business strategy that solves the company's difficulties.

**Executive Summary: Conclusion**

* 1. ***​​Determine the most valuable subscribers.***
  2. ***Understanding the subscriber segments present in the database.***

Using the assigned clusters from the model, specific customer segments could be visualized when exploring these clusters among various columns. Figure 13 exemplifies the most insightful subscriber segments found in these experimentations. We see that one cluster is much more populated that the rest (chart A), but despite this fact, median usage is highest among a differing cluster (chart B). Additionally, churn is quite high for all clusters except one (chart C), and email engagement is low for all clusters except two (chart D). With the knowledge of these clusters, multiple segments could be identified according to the business objectives and users that would like to be targeted.

* 1. ***Identify the most likely subscribers who could be sold additional products or services.***

In trying to understand the consumer base of Rosetta Stone, it is important to understand that about 79% of Rosetta Stone users are located in the United States. Because of this, it is hard to break down which subscribers could be sold additional products or services by the respective country. Therefore, an analytical approach utilizing the results of the modeling techniques described above informed our business strategy conclusions to address this objective.

While Rosetta Stone’s platform is not free, the company relies heavily on free subscribers. This gives the consumer an opportunity to test the product/service, and decide whether they want to subscribe for longer. A tactic that Rosetta Stone uses, along with plenty of other subscriptions based companies, is the default auto-renew for every free trial subscriber. Despite the number of users who forget to cancel their subscription, there is still a large sum of lifetime subscribers that have their auto renew on. Whether or not someone is an email subscriber can be a good indicator of if they are receiving emails concerning promotions, existing subscriptions, and updates to the Rosetta Stone platform. Visibility and persistence is key when trying to further your business and or promote your products/services. The higher your visibility and outreach, the higher you can expect your conversion rate to be. Another key indicator of subscribers that could potentially be sold more products would be those users who have their push notifications on. This allows for real-time updates to keep the user engaged with the platform and the language/s they wish to learn. For most, push notifications can be annoying and unnecessary if it is coming from an app that the user does not engage with often. These notifications keep the already committed subscribers engaged, and can influence low usage subscribers to further their language. While Rosetta Stone can’t make all subscribers commit to their platform for life, they can influence the decisions made by their subscribers with more outreach and heavier engagement. In order to do so, they first have to understand their markets, as well as their target audience.

* 1. ***Identify the subscriber profile of those not continuing with their usage of the product and identify the barriers to deeper subscriber engagement where possible.***

Based on the coefficients from the logistic regression model, we can infer some characteristics of subscribers who are not continuing with the usage of the product. Users who made their purchase from the web are more likely to discontinue their usage. This may suggest that there are barriers with the web purchasing experience that could be causing higher dissatisfaction with the product when compared to those purchasing on the app. Those users that started on a demo and/or free trial are less likely to discontinue the use of the product. This shows that these users had a chance to understand the product during their trial period and make a more informed decision on whether to purchase the product. Users that have subscribed to emails are more likely to discontinue their use of the product. This could be due to many factors including irrelevant content, frequency of emails, or privacy concerns. Additionally, the more emails that they send to these users the more likely they are to discontinue the use of the product. However, users that are receiving emails and interacting with the emails through open and click count are less likely to discontinue the use of the product. Subscribers that are using the app on the Android platform are more likely to discontinue use, followed by the IOS platform, and finally web users are least likely to discontinue their use of the product. Finally, a higher purchase amount leads to a higher likelihood to discontinue use of the product. Although this is the least impactful continuous variable, it could show the perceived value of the product.

To summarize, barriers to deeper subscriber engagement may include issues with the web purchasing experience, the effectiveness of email communication, and the difference in user experience on different platforms. A more personalized and relevant engagement approach could be beneficial with a push for users to start a demo or free trial phase to communicate the value of the product.

* 1. ***Outline any business relevant opportunities that are present from your analysis of the data not covered above.***

Rosetta Stone possesses certain strengths that need to be built upon, and weaknesses that need to be addressed. The language learning market is a very competitive landscape with each company having their own growth strategies. These strategies could include a focus on Artificial Intelligence, creation of new global learning platforms, implementation of cloud based language learning, etc. Rosetta Stone’s major selling point is their speech recognition software accompanying the philosophy of allowing users to learn as if they were a child.

Based on the data and outside research, we can see that over 75% of Rosetta Stone’s users are based in the U.S. While they still have somewhat of a strong presence in Europe and the UK, these foreign markets are being occupied by different language learning competitors. The global outreach aspect is an area that could be improved as there is so much opportunity outside the US. As mentioned earlier, all of these language learning companies have their own niche, and or specialization in their product.

Duolingo, for instance, is the biggest competitor in this language learning landscape. Like Rosetta Stone, they are not a complete language program, however, they have managed to capture over ⅔ of the global market share. They drive profits through expansion into other countries, by offering free usage of the platform, and offering a large list of languages, taught by native speakers. There is also a heavy interactive aspect with DuoLingo that Rosetta Stone misses out on. The game-like interface of DuoLingo allows for a more fun, interpretive way of learning.

In order for Rosetta Stone to compete in this every changing market, they need to adapt, and do some benchmarking of competitors. Rosetta Stone has seen their revenue growth plummet, despite a slight uptick in 2019 due to COVID. They need to leverage their long-standing place in the market, change features within their platform, and expand globally. Education management, higher education, and primary/secondary education are the major industries that use the Rosetta Stone platform. These are global industries that can be infiltrated by Rosetta Stone. By catering to the needs of the education sector in different countries, they can capture more market share, better understand these foreign markets, and ultimately make profit within these markets.

**Appendix**

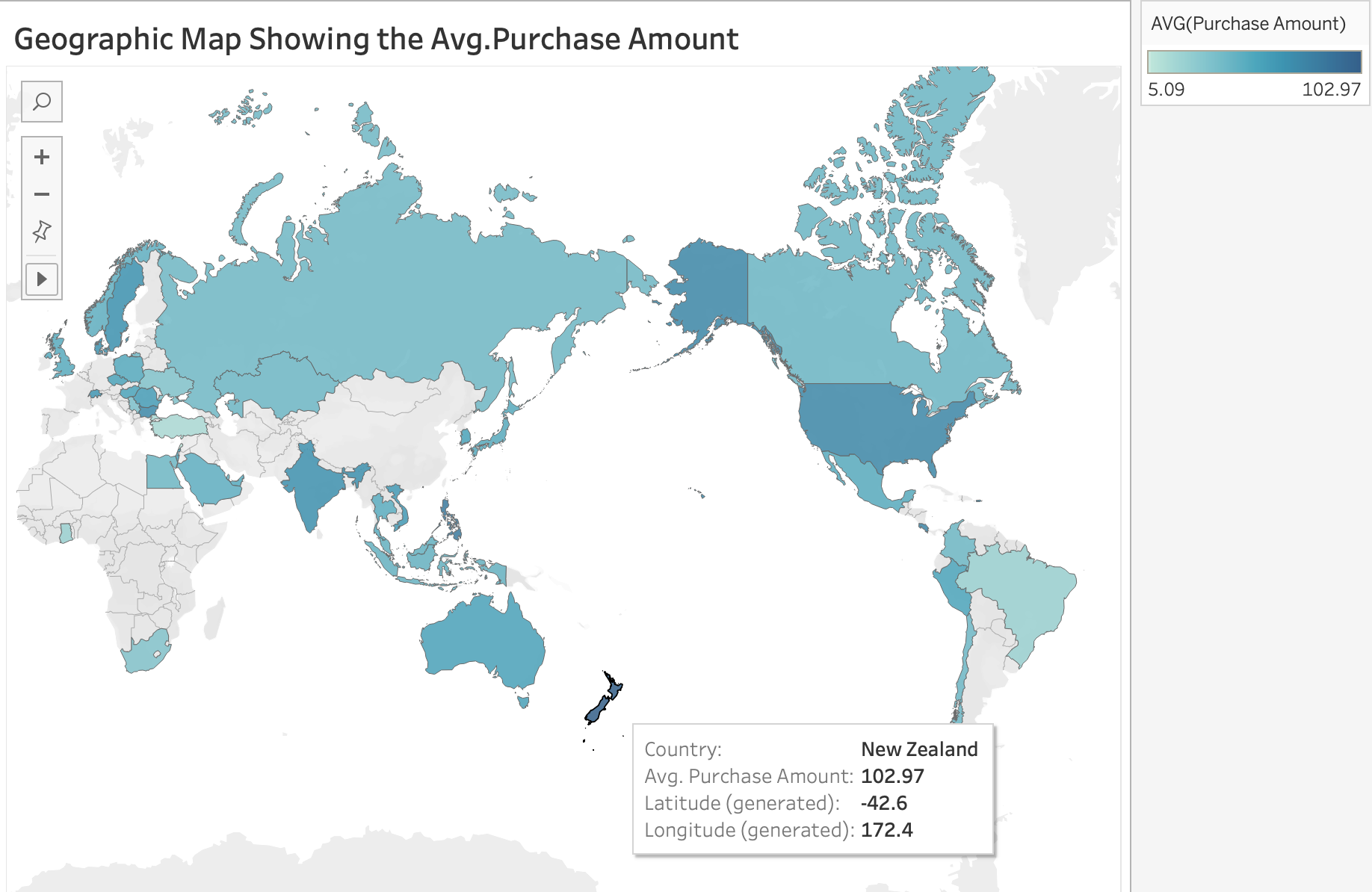
****

Figure 1. Geographic Map showing the average purchase amount for each Country.

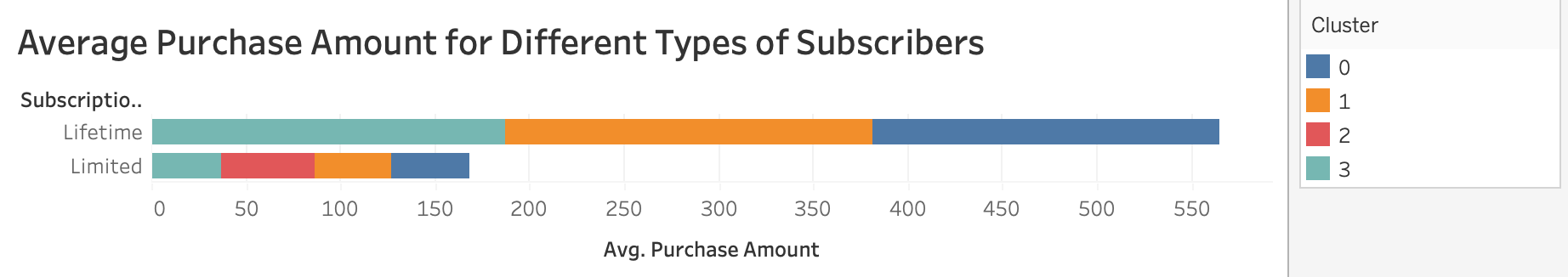


Figure 2. Horizontal bar graph showing the average purchase amount for the different types of subscribers (limited or lifetime).



Figure 3. Bar graph showing the average click amount for free trial users vs. non free trial users.

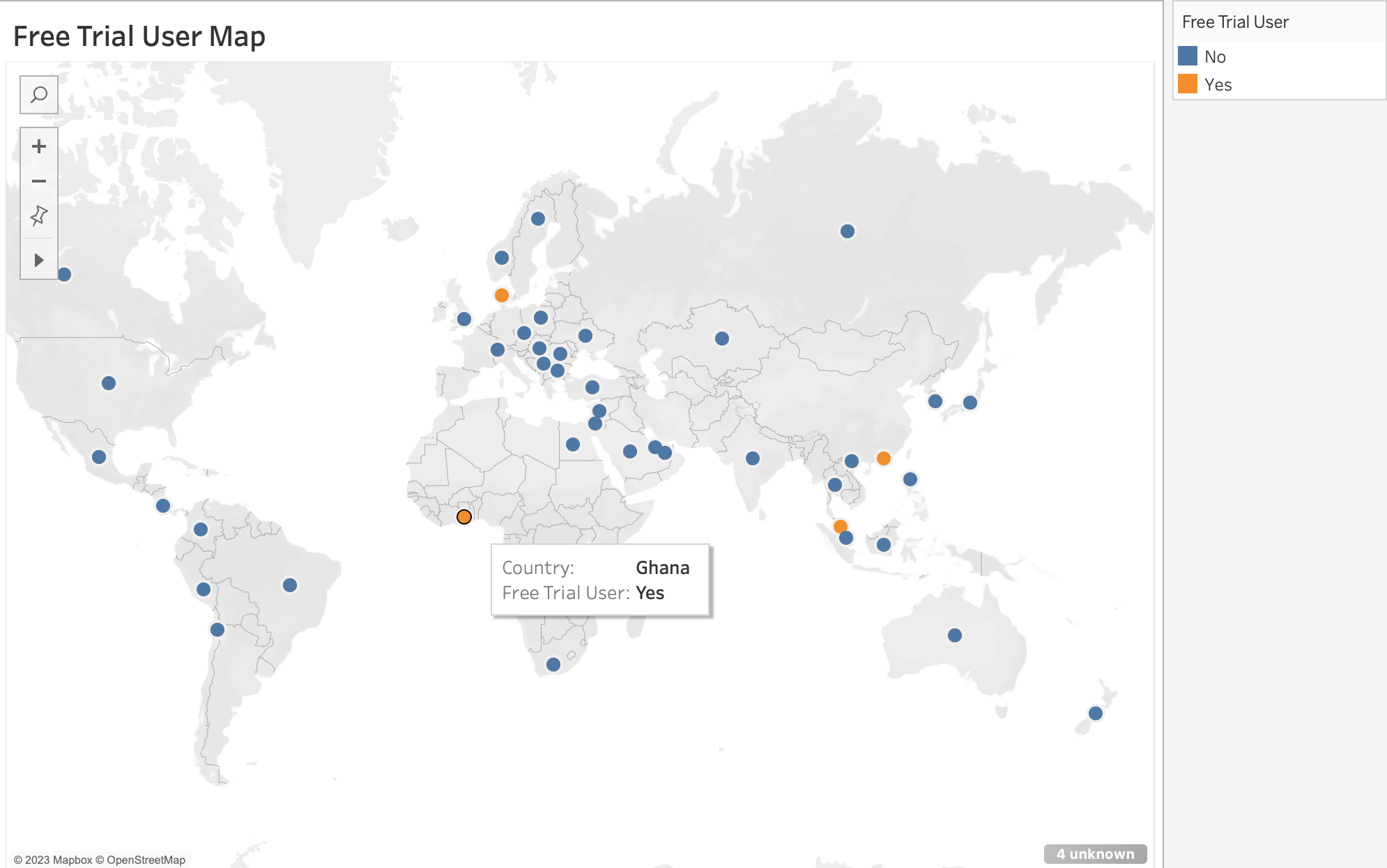


Figure 4. Geographic map showing the location of free trial users across the globe.



Figure 5. Bar graph showing the average click count for the different types of subscriptions.

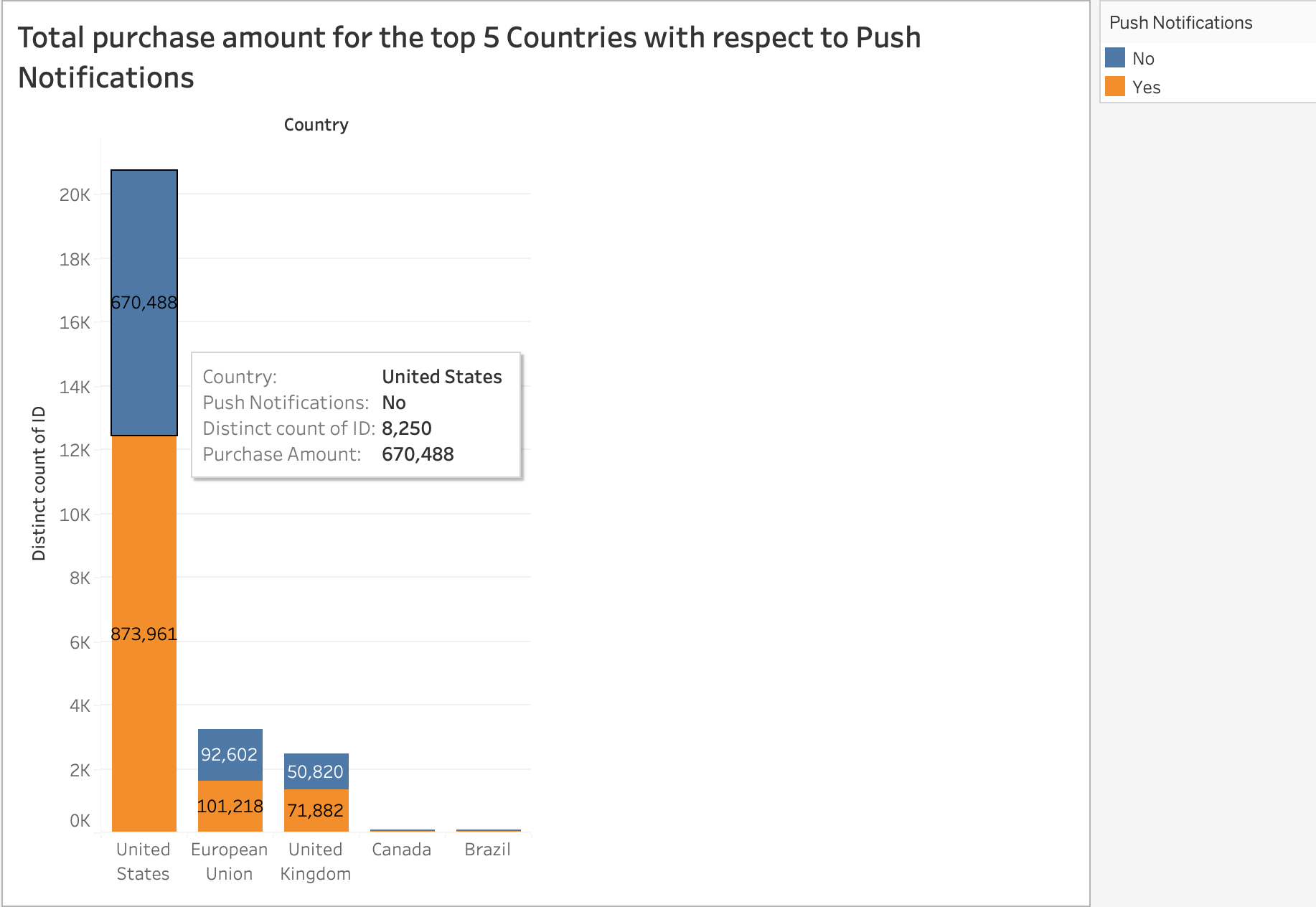


Figure 6. Bar graph showing the total purchase amount for the top 5 most active countries with respect to whether or not users have push notifications on.

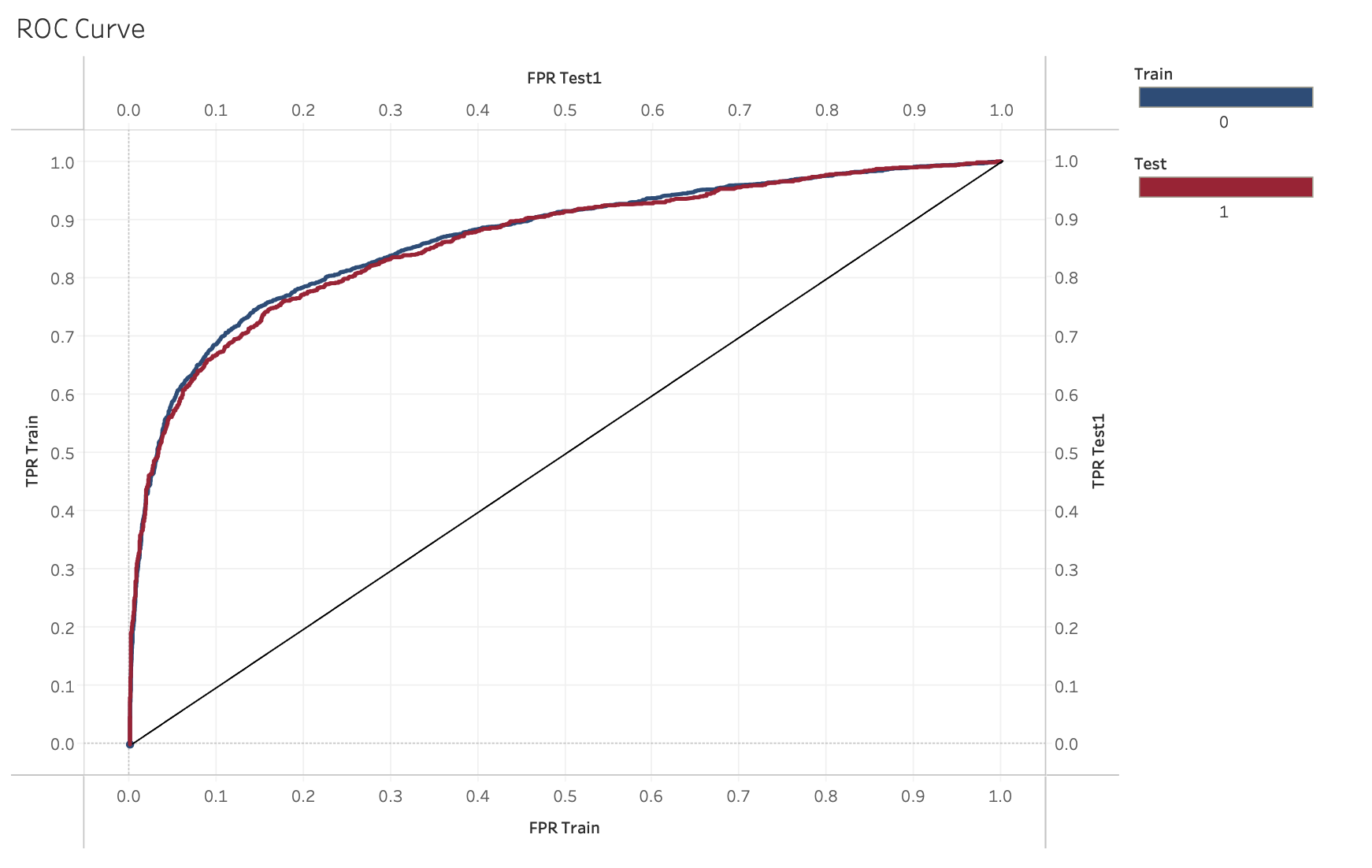


Figure 7. ROC Curve for results from logistic regression from Churn dataset.

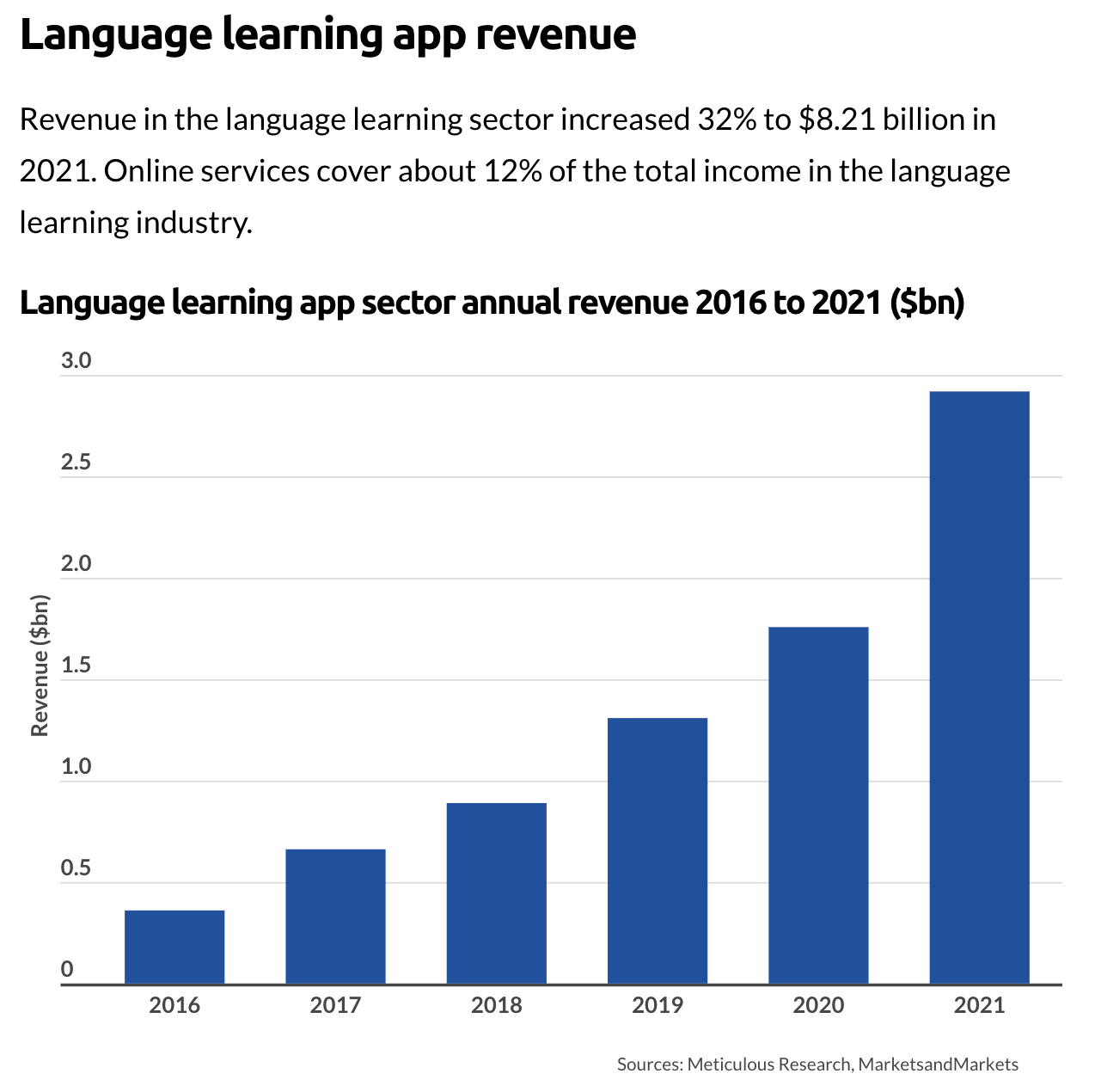


Figure 8. Bar Chart showing the growth in annual revenue for the language learning app sector

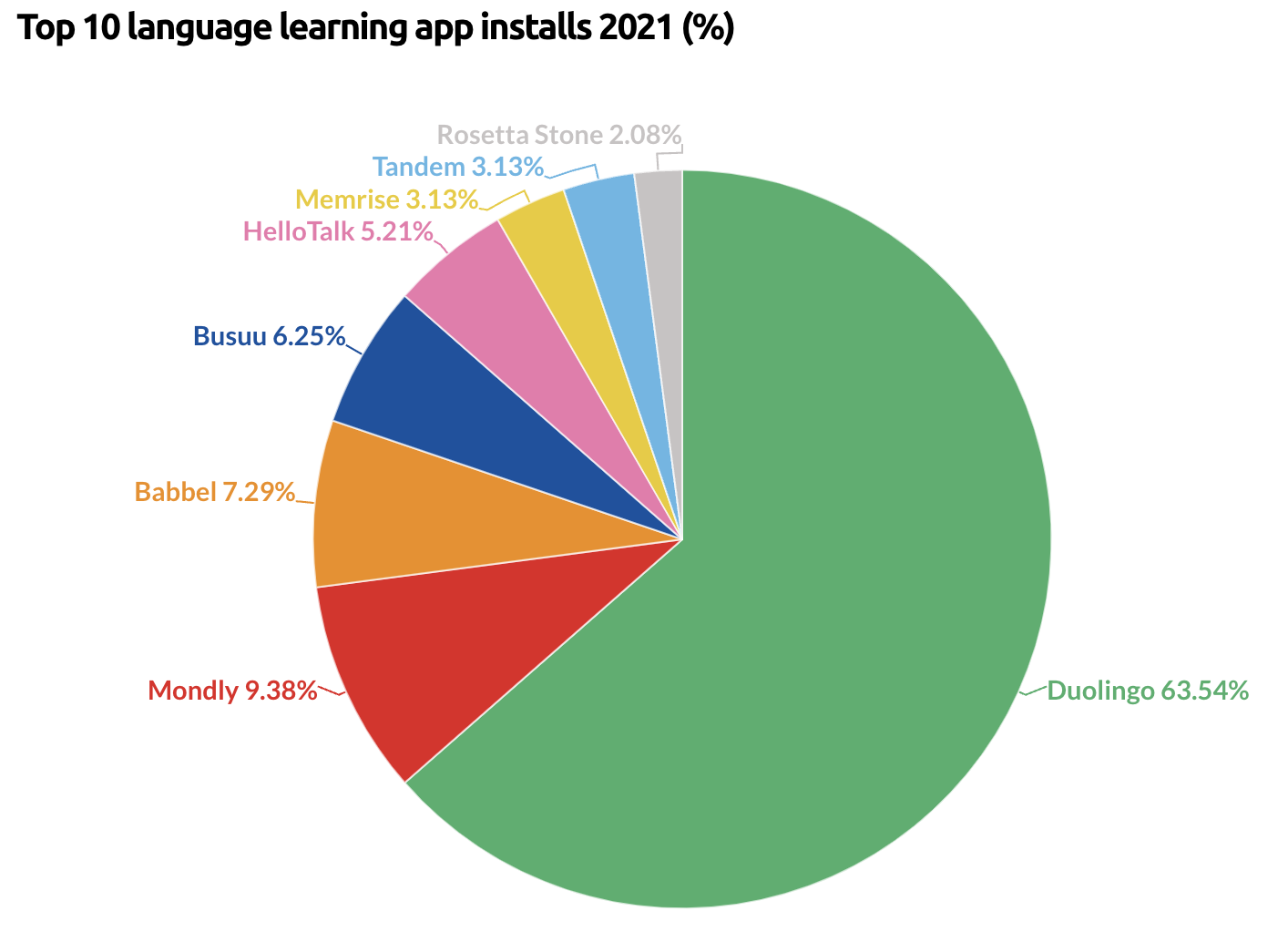


Figure 9. Pie chart showing the dispersion of app installation across different language learning platforms

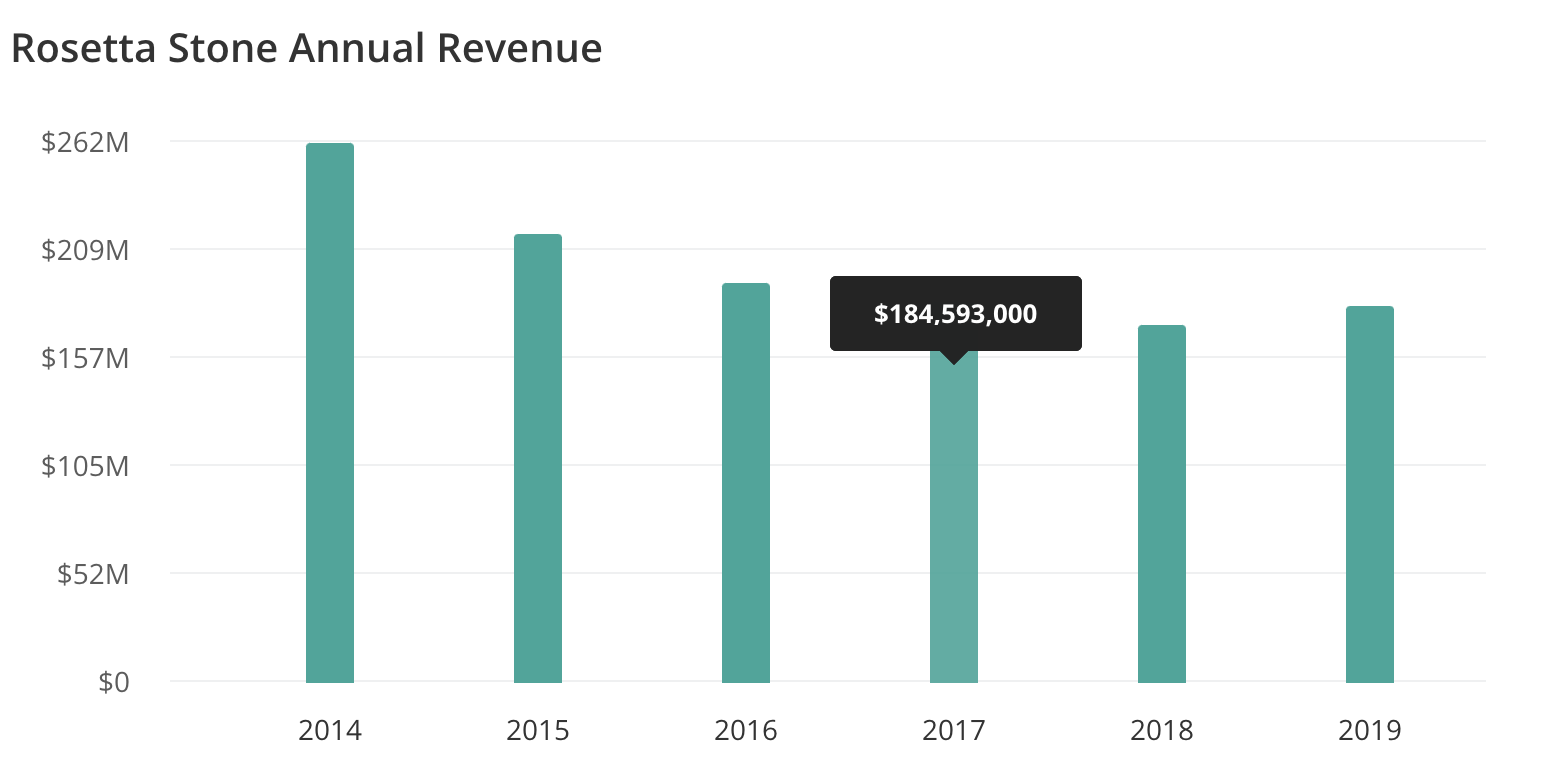


Figure 10. Rosetta Stone annual revenue by year from the years 2014 to 2019.



Figure 11. Decision Tree predictions Confusion Matrix.

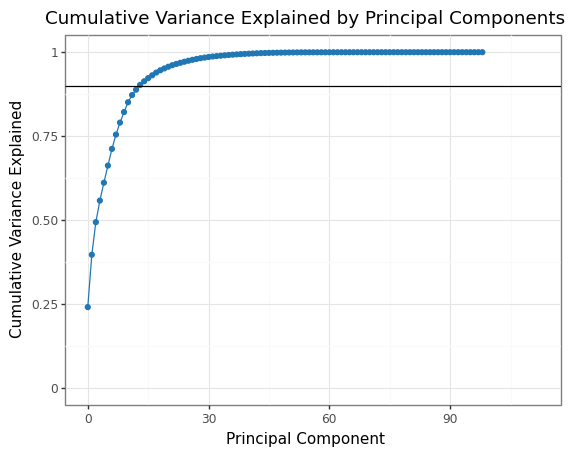
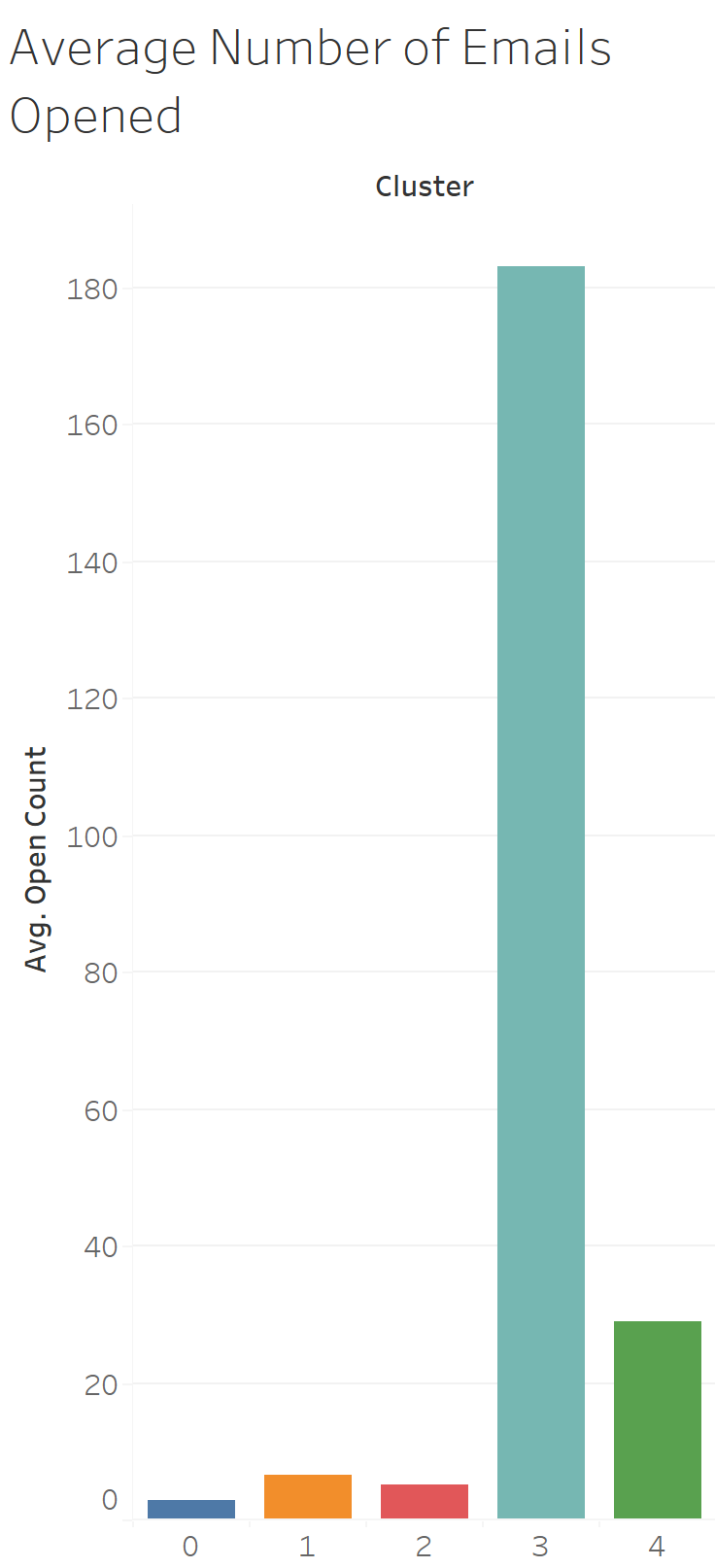
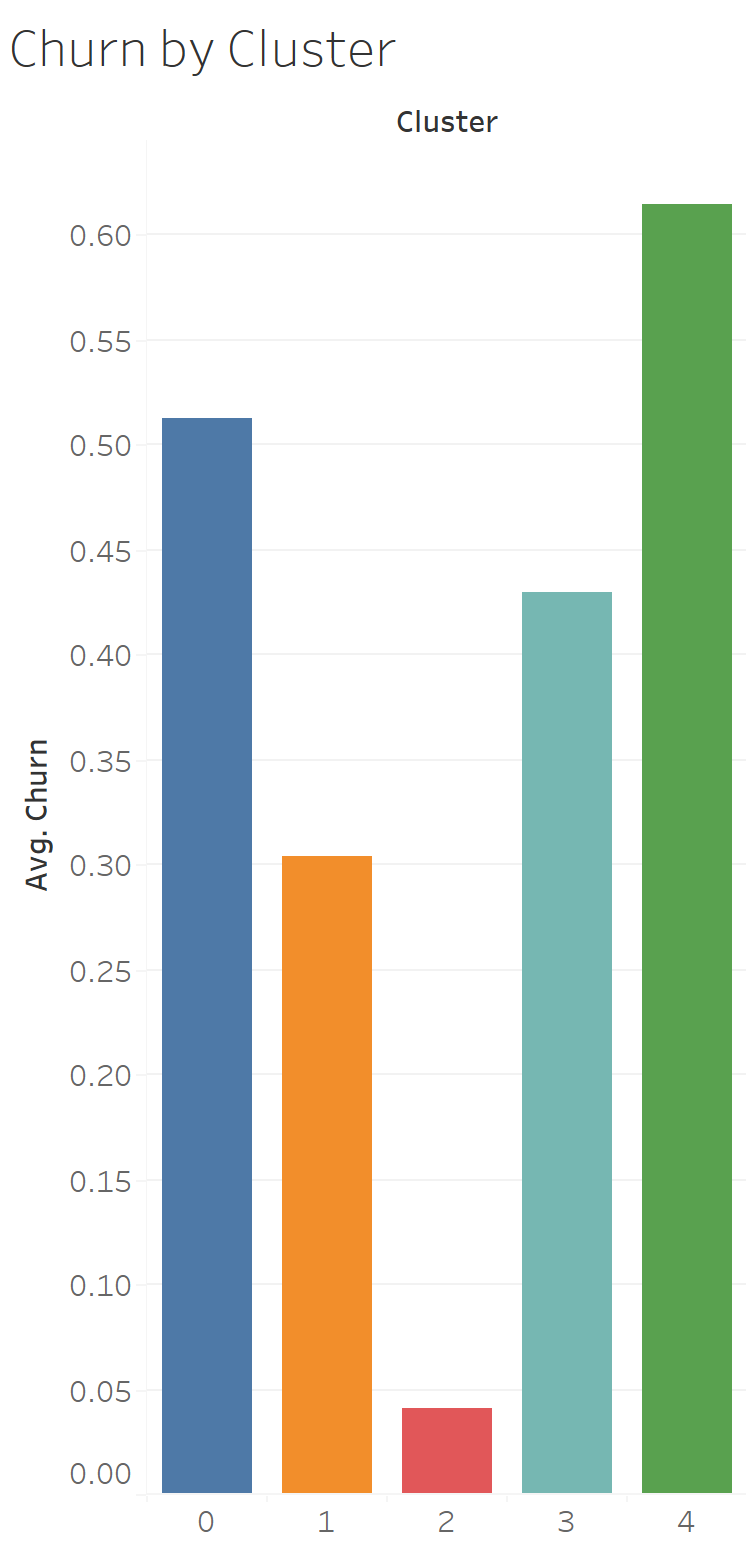
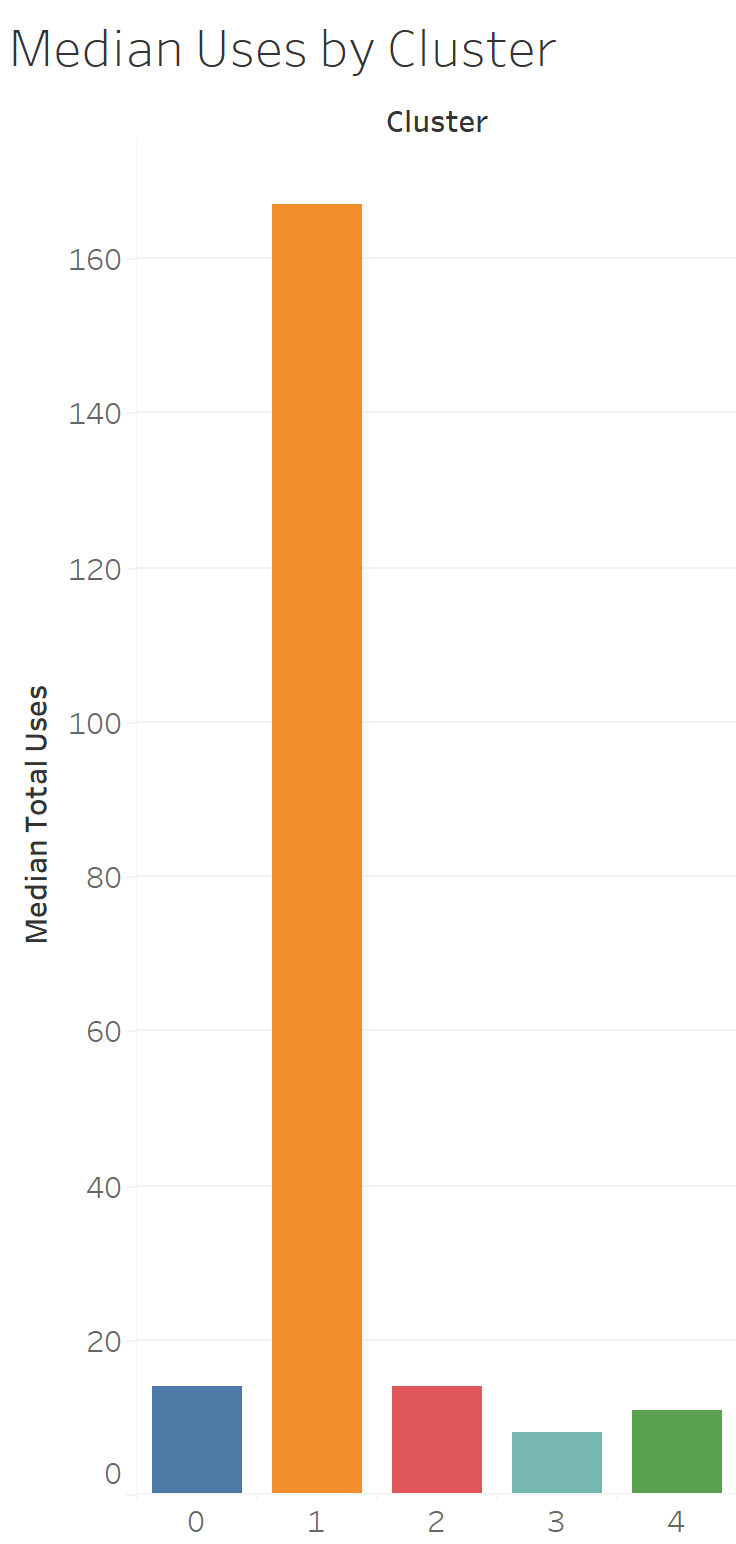
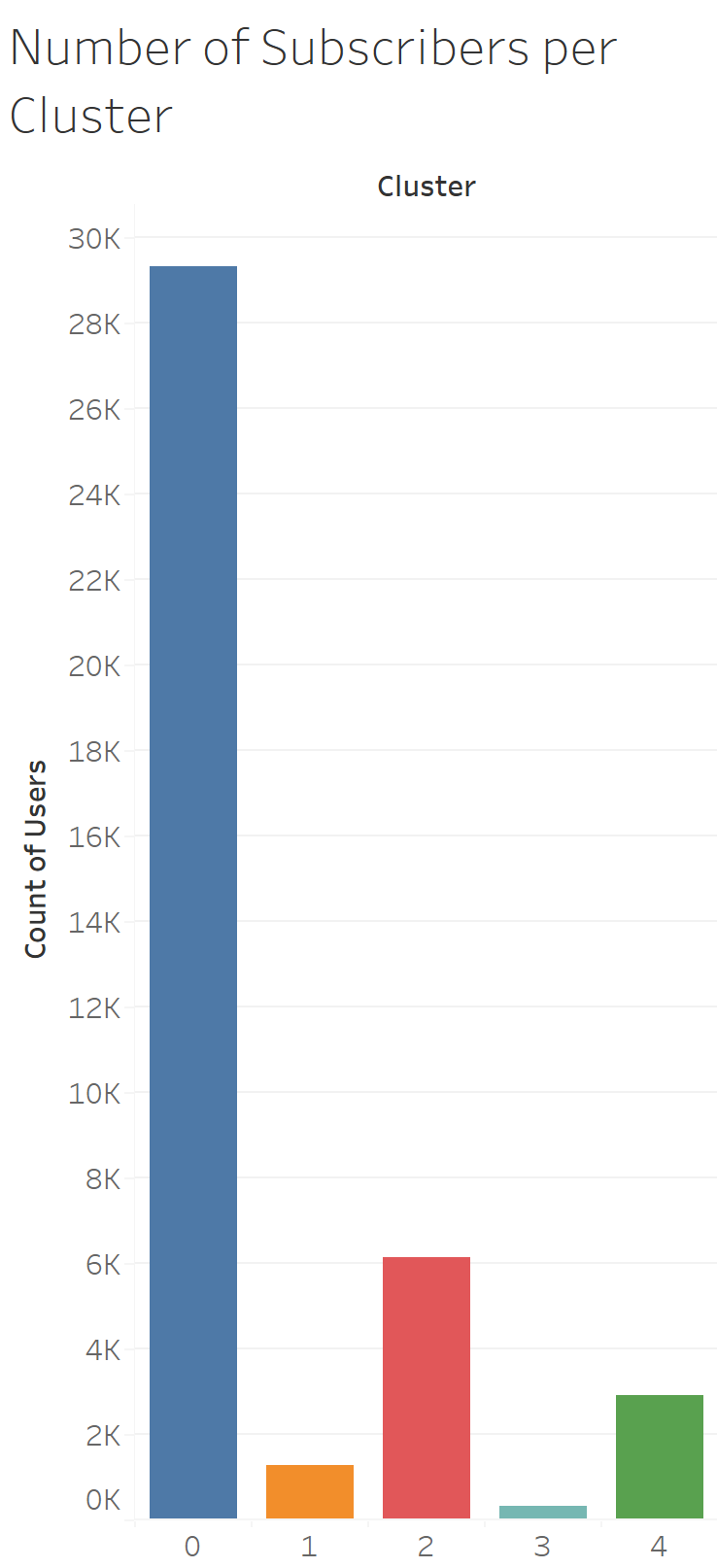


Figure 12. Explanation of variance graph resulting from Principal Component Analysis.



(a) (b) (c) (d)

Figure 13. Subscriber characteristic summarization by cluster

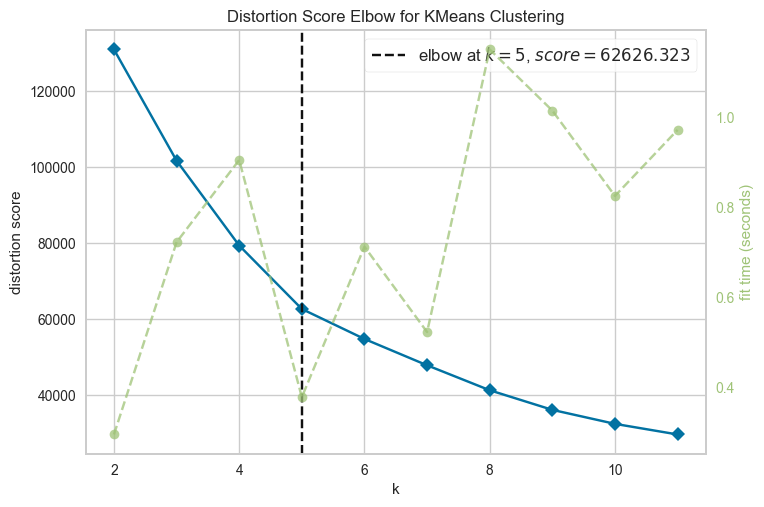


Figure 14. Distortion score elbow chart for Objective 1 K-Means clustering

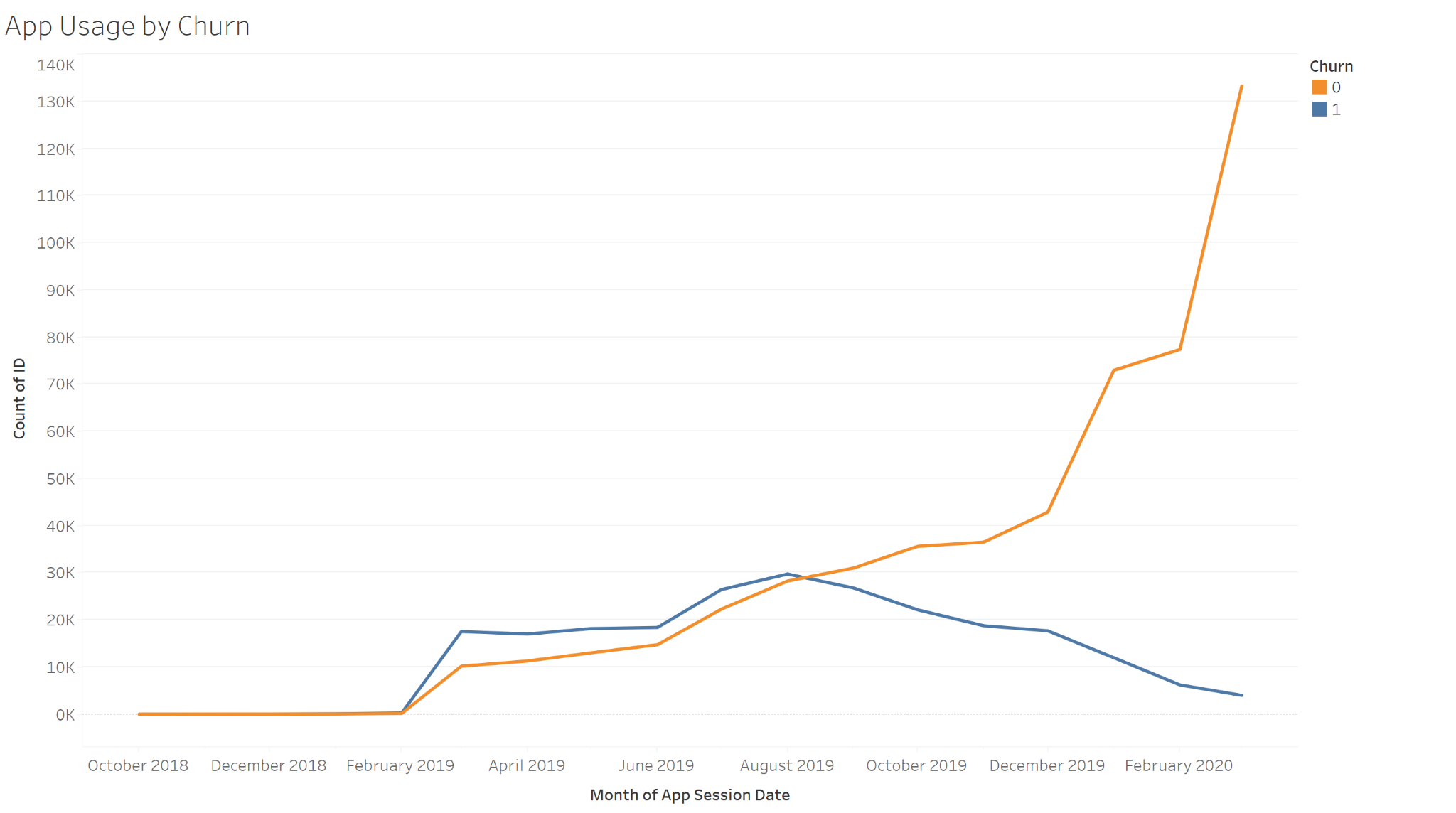


Figure 15. App Usage by Month according to Churn

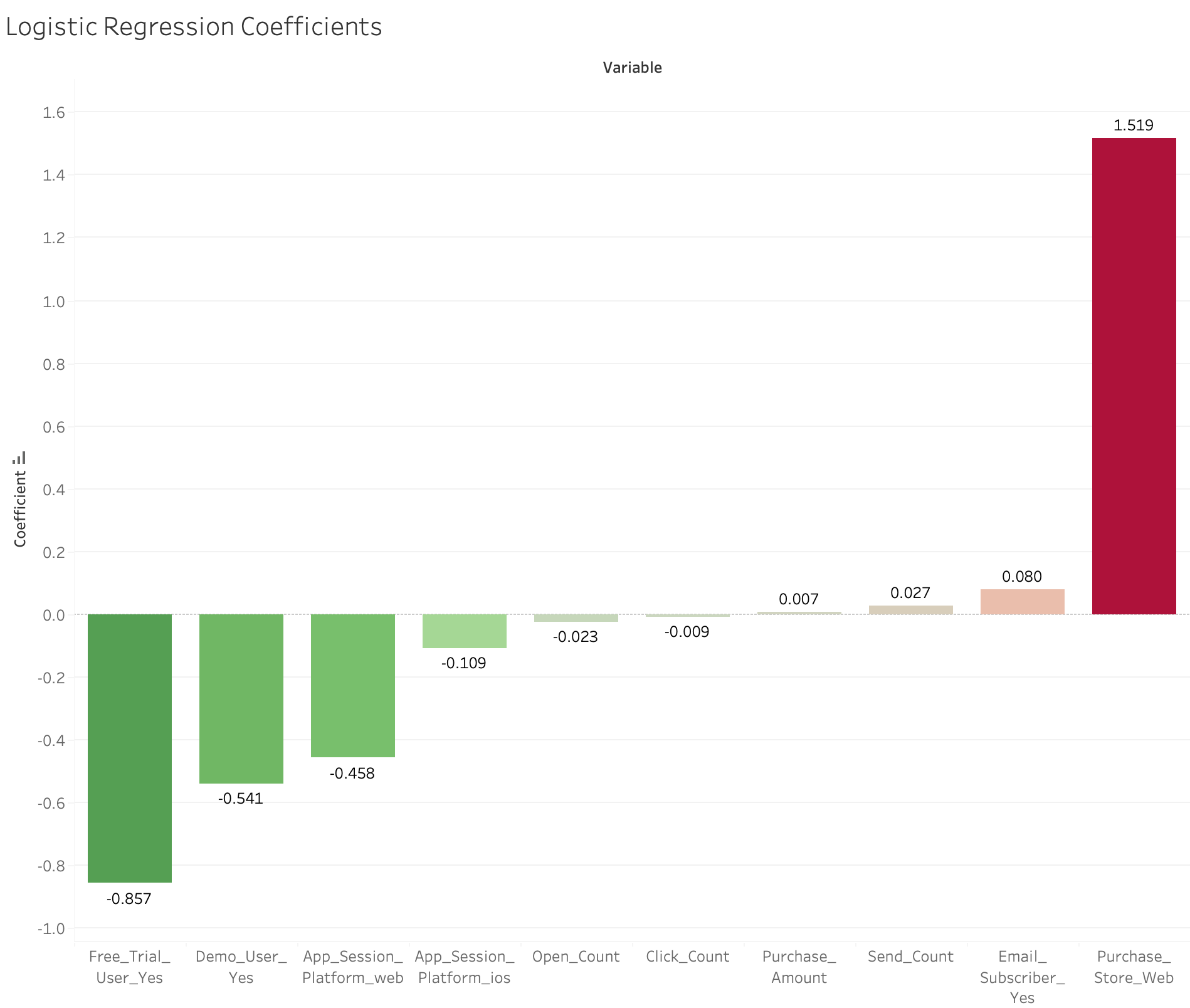


Figure 16. Coefficients from Logistic Regression.

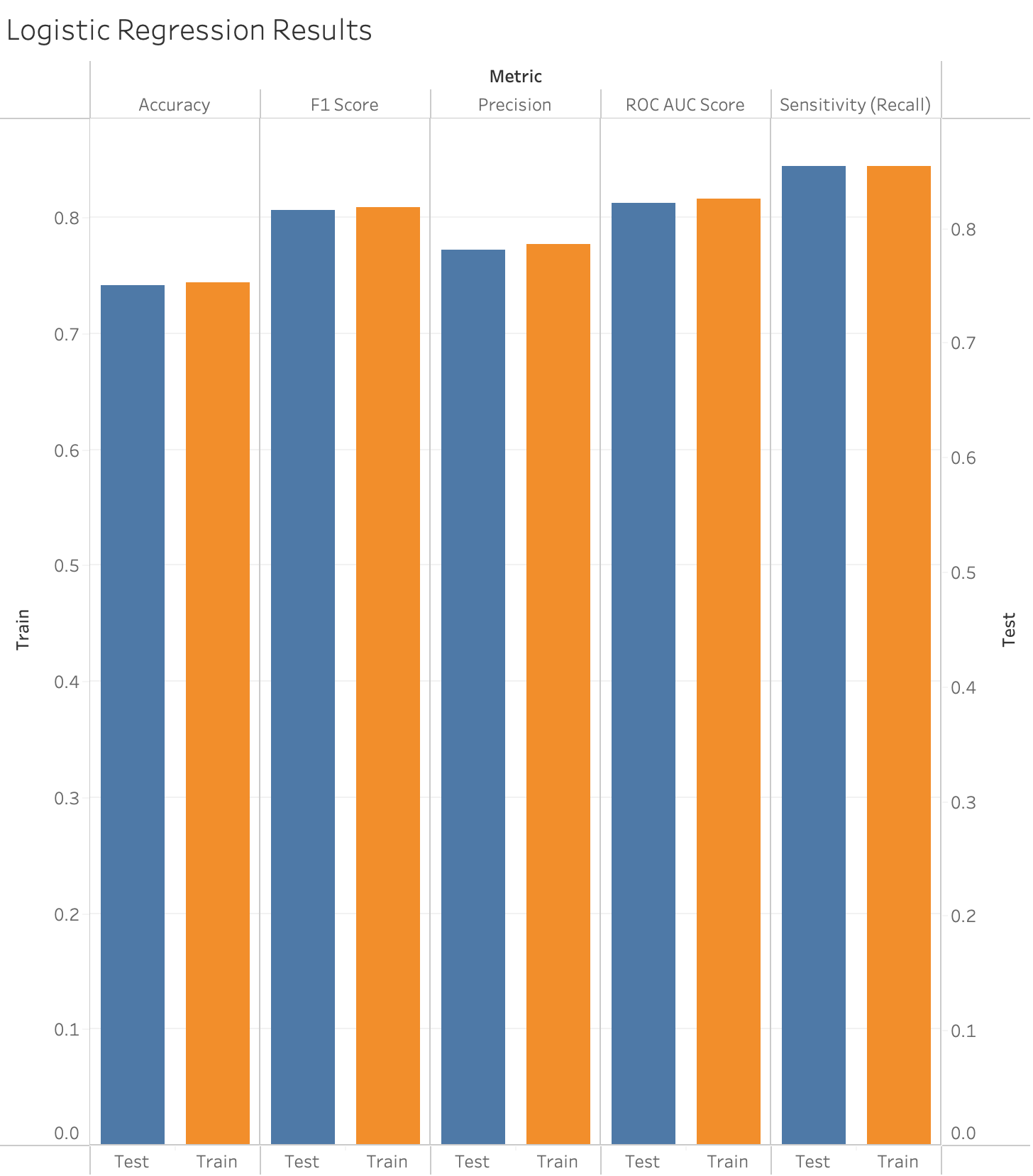


Figure 17. Logistic Regression Metrics.

**References**

Business of Apps. "Language Learning App Market." Business of Apps, 2021, <https://www.businessofapps.com/data/language-learning-app-market/#:~:text=Duolingo%20was%20the%20dominant%20language,most%20popular%20apps%20in%202021>.

Digital Journal. "Cloud-Based Language Learning Market Future Scenario, Growth Rate, Market Segmentation, and Industrial Opportunities to 2028 | Voxy, Inc., Rosetta Stone Inc., Culture Alley." Digital Journal, 2021, <https://www.digitaljournal.com/pr/cloud-based-language-learning-market-future-scenario-growth-rate-market-segmentation-and-industrial-opportunities-to-2028-voxy-inc-rosetta-stone-inc-culture-alley>.

Enlyft. "Rosetta Stone." Enlyft, <https://enlyft.com/tech/products/rosetta-stone>.

Langoly. "Rosetta Stone Review." Langoly, 2021, <https://www.langoly.com/rosetta-stone-review/>.

LinguaLift. "The 15 Best Language Learning Apps of 2021 That Actually Work." LinguaLift, 2021, <https://www.lingualift.com/blog/best-language-learning-apps/>.

Reddit. "Why is Rosetta Stone so bad?" Reddit, 2021, <https://www.reddit.com/r/languagelearning/comments/pq4tjp/why_is_rosettastone_so_bad/>.

Zippia. "Rosetta Stone Revenue, Number of Employees, Annual Growth and Compounded Annual Growth Rate." Zippia, 2021, <https://www.zippia.com/rosetta-stone-careers-9940/revenue/>.