**Rosetta Stone Analytical Plan & Analysis**

Group 8
MGSC 410
Chapman University

**Table of Contents:**

**Action Plan**

Our primary goals include analyzing Rosetta Stone's user base to identify specific customer groups to perform in depth analysis on. Once we identify groups of consumers through data modeling we intend to label them based on their attributes including overall app usage, user types, contract duration, and amount paid. Secondly, our analysis pertains to understanding consumer behavior and recognizing barriers that constrain consumers from further investing and engaging with Rosetta Stone's platform. Additionally, we aim to find data supported opportunities that are mutually beneficial to both consumers and Rosetta Stone by providing tangible business solutions. Group wise, we are tackling questions based on interest and vision all while implementing measurable goals that allow us to iteratively progress throughout our project duration. The remainder of this document outlines our methodologies for data management, modeling approaches to answer the specific questions, and our storyline for the presentation.

**Executive Summary (Introduction)**

Valuable customers: Lifetime & Loyal Users

Redflag customers: Inactive & Free Trial Users

Barriers:

1. No Bang for your Buck
2. Short Retention Rate
3. Weak Marketing Techniques

Business opportunities:

- Free Trials don't improve long term subscribers or renewal subscriptions.
- New Email techniques could encourage customers to remain with Rosetta Stone.

**The Data - Connor**

The data gathered is internal data from Rosetta Stone transactions

| | id | language | subscription_type | subscription_event_type | purchase_store | purchase_amount_raw |
|---|---|---|---|---|---|---|
| 0 | 1 | POR | Limited | INITIAL_PURCHASE | App | NaN |
| 1 | 2 | EBR | Limited | INITIAL_PURCHASE | Web | 39.00 |
| 2 | 3 | ESP | Limited | INITIAL_PURCHASE | Web | 0.00 |
| 3 | 4 | KOR | Limited | INITIAL_PURCHASE | App | NaN |
| 4 | 5 | ENG | Limited | INITIAL_PURCHASE | App | NaN |

```
id
language
subscription_type
subscription_event_type
purchase_store
purchase_amount_raw
currency
subscription_start_date
subscription_expiration
demo_user
free_trial_user
free_trial_start_date
free_trial_expiration
auto_renew
country
user_type
lead_platform
email_subscriber
push_notifications
send_count
open_count
click_count
unique_open_count
unique_click_count
```

and application activity. Attached are some screenshots of the two initial datasets. The most notable attributes of this dataset are subscription_type and subscription_event_type. subscription_type describes if a customer who made the purchase is a lifetime or limited customer. A limited subscriber has to keep renewing the subscription to retain the service. scubscription_event_type describes if a transaction is a renewal or an initial purchase. These have significant effects on revenue and will be discussed in the main sections below.

| | id | app_session_platform | app_activity_type | app_session_date |
|---|---|---|---|---|
| 0 | 1 | ios | App Launch | 3/20/2019 |
| 1 | 2 | android | App Launch | 12/3/2019 |
| 2 | 3 | ios | App Launch | 5/2/2019 |
| 3 | 4 | ios | App Launch | 2/6/2020 |

**Data Engineering - Connor**

Next, the data needed to be transformed into a useful format for the analysts and the models they use.

The data initially had many categorical variables. For many models and interpretations this can be troublesome. We turned these variables into one-hot. This means that each category in a column is transformed into multiple mutually exclusive columns. A transaction can be an initial purchase *or* a renewal.

The next and most intensive step of the data engineering process was working with the raw purchase prices. There were two problems: one, that the purchase prices were in different currencies and that the numbers were unrealistic. An example being that Rosetta Stone did a $1.6 trillion sale. That's wrong. To convert the currencies to USD, the model first looked through the purchases and assumed that because most of the sales were US sales, any missing ones would likely be American ones. Then the model used the subscription start date and the two currency rates to convert all transactions to USD. Yes, this methodology is heuristic, but it is a solution that will likely solve >95% of the problem cases.

To deal with these high numbers for sales we decided to take a statistical and automated approach. The high prices in the data *did* have a pattern, *but* these differences in the data were not confirmed. If we spent our time fixing an error that would change in its type, then the time would be wasted. The other option would be to look through the data individually and manually

offset data that is off. Instead we modeled the data on a normal distribution with a max of 5000. This represented a reasonable sale made by Rosetta Stone. Sales over 1000, are extremely unlikely. This approach then removed values that lied 4 standard distributions away from the mean. The standard distribution was $77.84 and the new mean from this reduction was $70.50. The new maximum purchase was $412.97. A little over 14,000 values were removed to make a more reasonable dataset.

Modeling for these missing values is the most logical way to remedy them. We used these variables:

```
predictors = ['language', 'subscription_event_type','purchase_store', 'demo_user',
              'free_trial_user','auto_renew', 'country', 'user_type', 'email_subscriber',
              'push_notifications', 'unique_open_count', 'unique_click_count',
              'subscription_length_days', 'total_app_interactions', 'launch_app_interactions',
              'subscription_type_lifetime', 'subscription_type_limited']
```

The outcome was purcahse_amount_usd. We then replaced these values and had an R^2 of 0.75. Usually R^2 is not a good way to find causality, but in this situation all we needed to do is replace the missing values as close to the real value as possible. We then predicted on the missing values and any negative predictions were replaced with 0. This imputed data wasn't used for modeling directly on price and purchasing decisions, but was primarily for visualization and gathering other general insights about the data.

Simple transformations were done to extract some more information from the attributes. A subscription length variable was created by finding the difference between expiration end and start. As our analyses found, time had no significant effect on usage or sales. We treat records as equal and look at the differences in length between them. Then we looked into subscription types and found that ~6500 subscriptions were renewal subscriptions and ~33500 subscriptions were lifetime.

On the export of the data, id and subscription_expiration were dropped because they were not needed for any more data transformations or would be useful in the models we made. Next, we start the modeling on this transformed data.

**Methodologies for questions**

- **Question 1 - Liz and Nora**

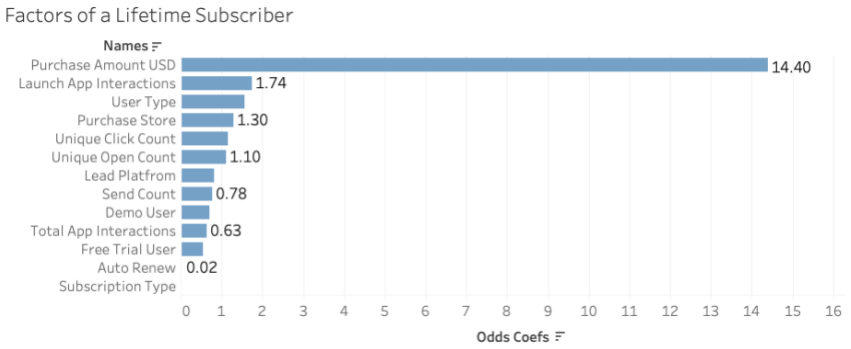
Initially we defined what a valuable subscriber meant to us. In determining value, we looked at, on average, the group of individuals who paid the most per subscription and who had their auto renewals on. In the figure below, the data indicates that limited type subscribers (0) on average pay ~ $40.21 per subscription while lifetime subscriptions pay ~ $183.66. Intuitively that makes sense as lifetime subscribers invest more in the company due to the educational benefits that they obtain.

| subscription_type_lifetime | purchase_amount_usd_imputed |
|---|---|
| 0 | 40.209295 |
| 1 | 183.655961 |

Running a logistic regression on lifetime subscriber and auto renewal provided insight on the attributes that contribute to the likelihood of each event happening (i.e. renewals turned on or they are a lifetime subscriber).

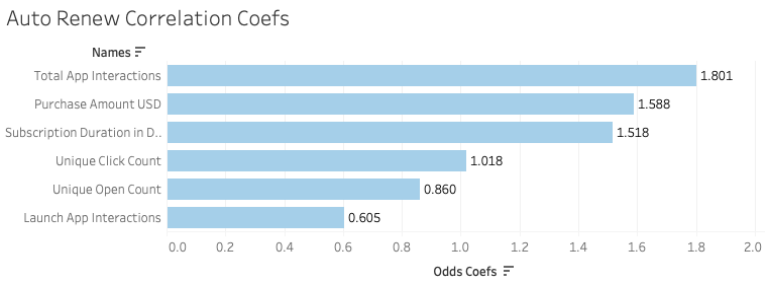
Displayed below are the correlation coefficients contributing to the likelihood of being a lifetime subscriber. We see that variables including purchase amount paid, launch app interactions, user type, and purchase store contribute to such likelihood.

| | Coefs | Names | Odds Coefs |
|---|---|---|---|
| 0 | -8.285867 | subscription_event_type_bin | 0.000252 |
| 1 | 0.245138 | purchase_store_binary | 1.277797 |
| 2 | 0.455682 | user_type_binary | 1.577249 |
| 3 | -0.202858 | lead_platform_type | 0.816394 |
| 4 | -0.369988 | demo_user | 0.690742 |
| 5 | -0.595565 | free_trial_user | 0.551251 |
| 6 | -4.328631 | auto_renew | 0.013186 |
| 7 | -0.284085 | send_count | 0.752702 |
| 8 | 0.110759 | unique_open_count | 1.117126 |
| 9 | 0.164990 | unique_click_count | 1.179381 |
| 10 | -0.442023 | total_app_interactions | 0.642735 |
| 11 | 0.548219 | launch_app_interactions | 1.730169 |
| 12 | 2.645495 | purchase_amount_usd_imputed | 14.090424 |
| 13 | -9.967977 | intercept | 0.000047 |

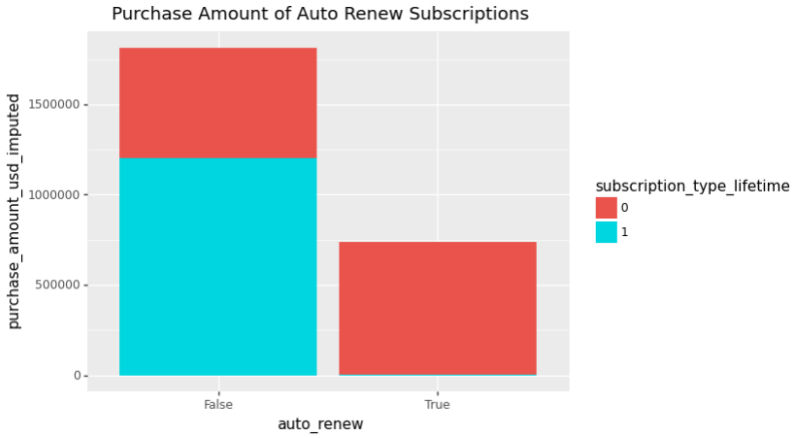



Displayed below are the correlation coefficients contributing to the likelihood of being an auto renewal customer. We see that variables including purchase amount paid, total app interactions, user type, and days subscribed contribute to such likelihood.

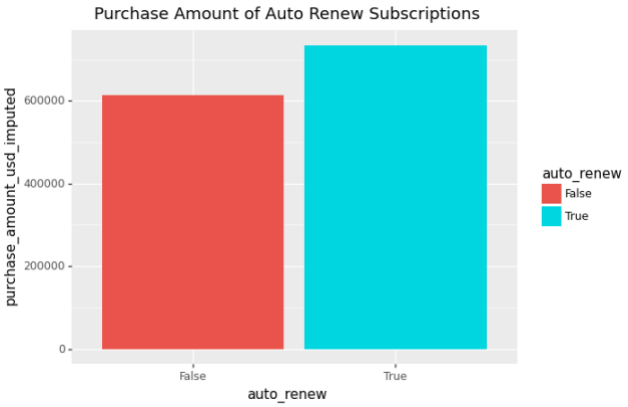
| | Coefs | Names | Odds Coefs |
|---|---|---|---|
| 0 | 0.502942 | total_app_interactions | 1.653578 |
| 1 | -0.427463 | launch_app_interactions | 0.652162 |
| 2 | 0.415348 | subscription_length_days | 1.514898 |
| 3 | 0.460195 | purchase_amount_usd_imputed | 1.584383 |
| 4 | -0.165330 | unique_open_count | 0.847614 |
| 5 | 0.018057 | unique_click_count | 1.018221 |
| 6 | -0.305530 | intercept | 0.736733 |




Since we know that lifetime subscriptions pay the most on average we also know that since they are lifetime, they do not have their auto renews on which is visually displayed below.




In order to get a more accurate depiction of how much the remaining population of auto renewal customers are paying, we created a separate date frame that excluded lifetime subscriptions to then output the following ggplot:




Similarities between lifetime and auto renew subscribers include amount paid and app interactions. Additionally we used the clusters found in question 2 to support our claim that short duration subscriptions paired with high application activity are valuable. Overall, this regression and question 2's cluster analysis supports our conclusion that valuable customers are those who exhibit monetary investment, retention, and high application activity regardless of duration.

- **Question 2 - Aviv**

Utilizing K-Means clustering to categorize users according to their similarities. Patterns among users are helpful to identify users who are valuable, valueless, more likely to purchase additional products, and other non-obvious trends in the data.

Determining customer segments is a helpful first step in answering the specific questions we've been tasked with. Customer clusters illustrate user similarities, which provide intuitive insights. K-Means is chosen because it is computationally effective and statistically rigorous.

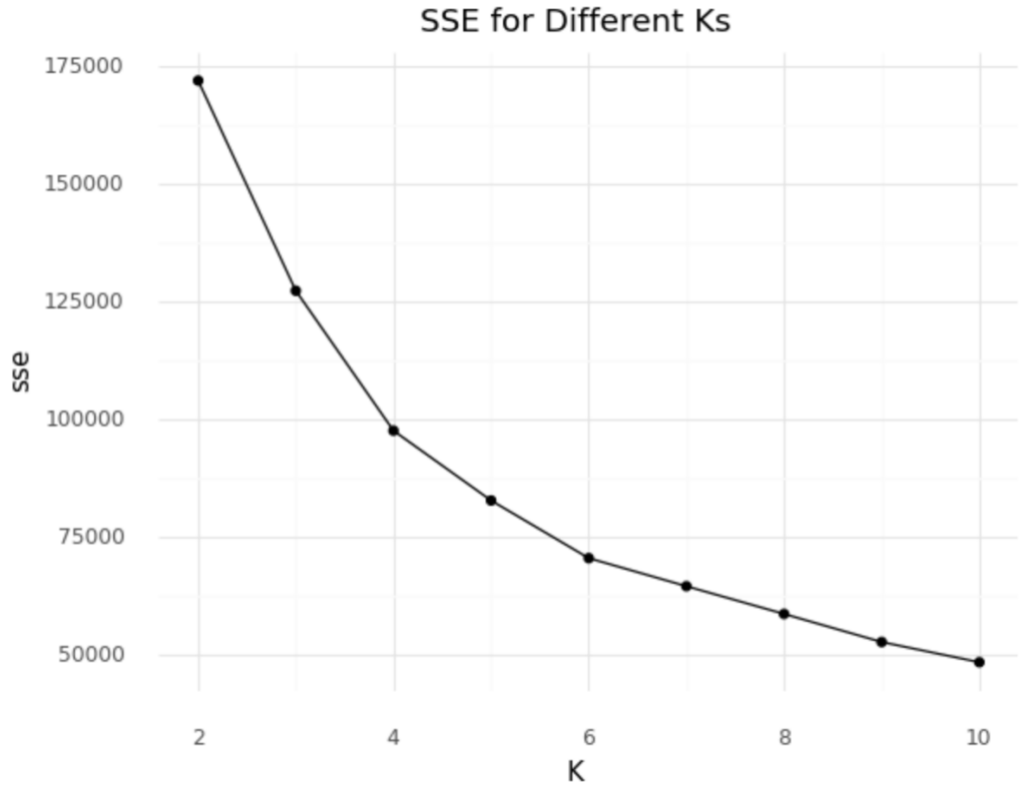
The first step is determining the key features the model is built on. Here, total_app_interactions, launch_app_interactions, subscription_length_days, purchase_amount_usd_imputed, unique_open_count, and unique_click_count are utilized. These

were chosen because they are continuous and intuitively significant, that is it makes sense these should be included for subscriber optimization.

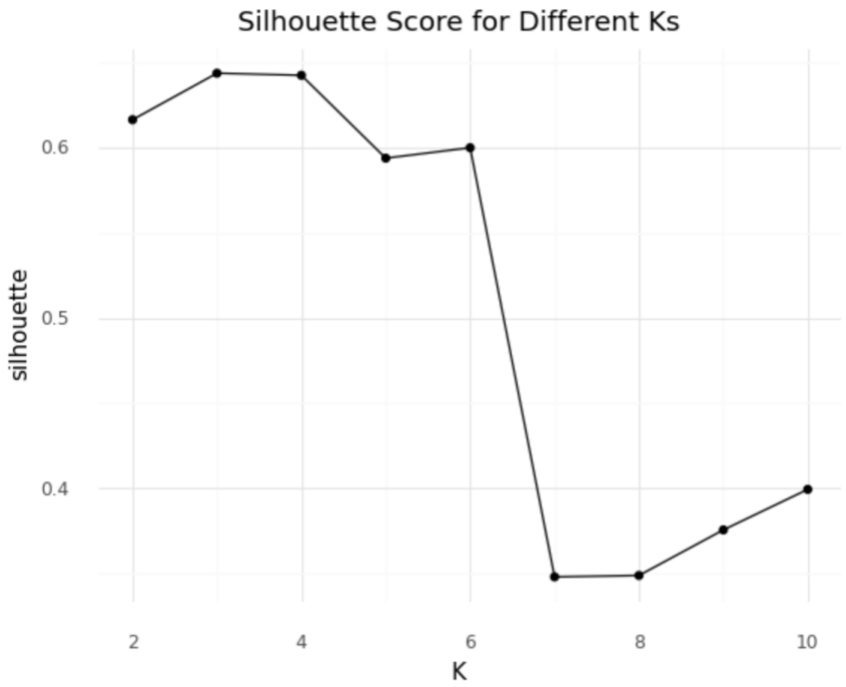
Choosing the quantity of clusters, K, in the model is determined by looping through many options and calculating the silhouette and sum of squared errors at each. This is illustrated in the two images below.

```
# Choosing the perfect K
ks = [2,3,4,5,6,7,8,9,10]

sse = [] # sum of squared errors
sils = [] # silhouette score

for k in ks:
    km = KMeans(n_clusters = k)
    km.fit(X)

    sse.append(km.inertia_) # average distance between point and its center
    sils.append(silhouette_score(X, km.predict(X)))

sse_df = pd.DataFrame({"K": ks,
                       "sse": sse,
                       "silhouette": sils})
```

⇒



⇒



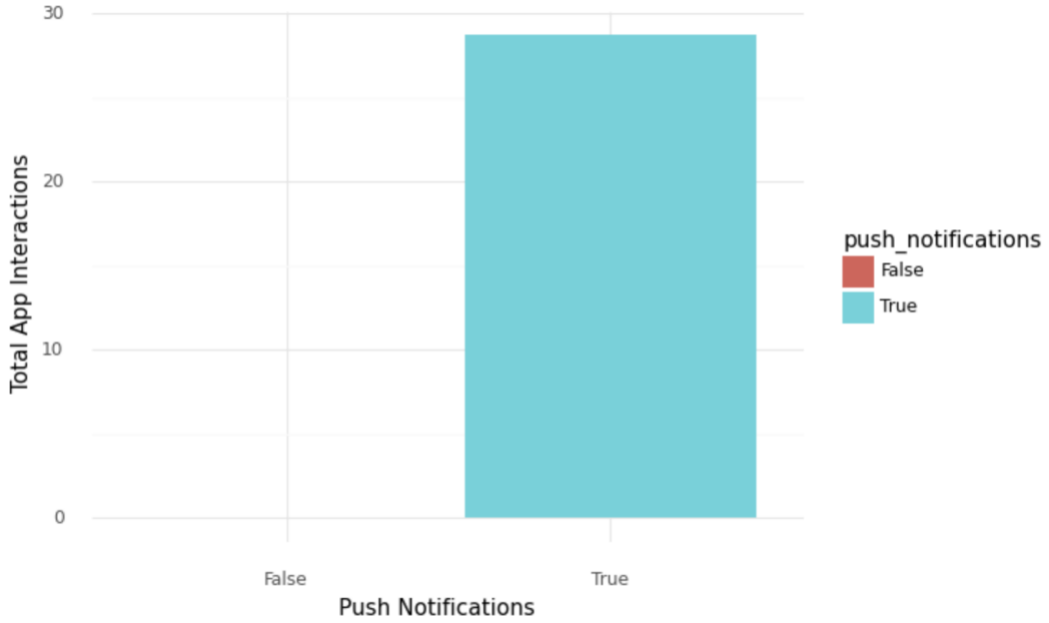

A value at a low sum of squared errors and a high silhouette score is desired. Hence, K = 4 is chosen. The clusters were analyzed thereafter using the graphs at the bottom of this document.

- **Question 3 - Aviv**

Using basic plotting techniques to identify relationships between variables to distinguish users who are more likely to purchase additional items.

Intuitively, users who 1) enable push notifications, 2) are on the email list, 3) have a renewal subscription plan, and 4) have a relatively high total application interaction are more likely to purchase additional items. The next step was to analyze these features to validate the intuition.

I began by scrutinizing users with push notifications enabled because the very first plot I graphed indicated their importance, as shown here:




This graph illustrates that users with push notifications are virtually the only ones using the application. Immediately, therefore, push notifications became a noteworthy trait. The rest of the analysis follows this pattern - identifying relationships between users with push notifications enabled and other features, like whether they had email notifications or whether they were long term subscribers. In short, push notifications is a key characteristic of users who could be sold additional products. Despite convincing evidence, this one feature alone isn't powerful enough. So, upon further analysis, long term subscribers were also categorized as those who could be sold additional products. Thus, the coupling between these two features - push notifications and long term subscription - pinpointed a group of users who are almost certainly willing to purchase additional products.

- **Question 4: Identify the subscriber profile of those not continuing with their usage of the product and identify the barriers to deeper subscriber engagement where possible.**

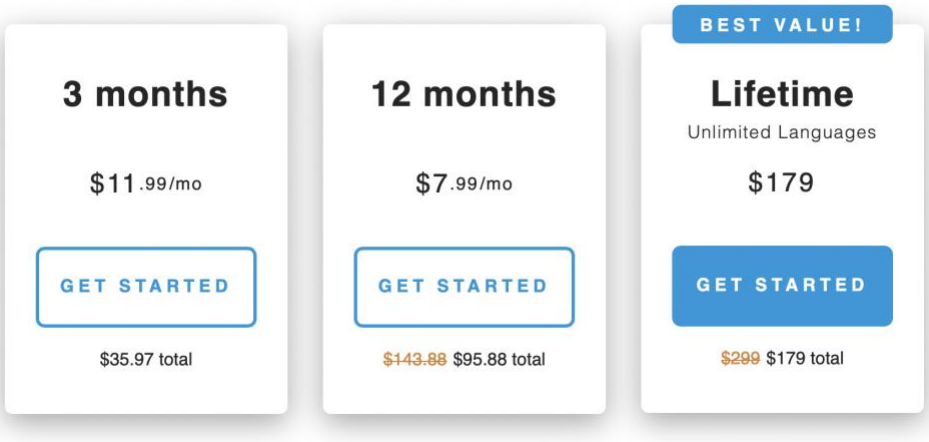
To initially answer this question, I broke it down into two parts: "identifying subscriber profiles of those not continuing" and "identifying barriers to deeper subscriber engagement".

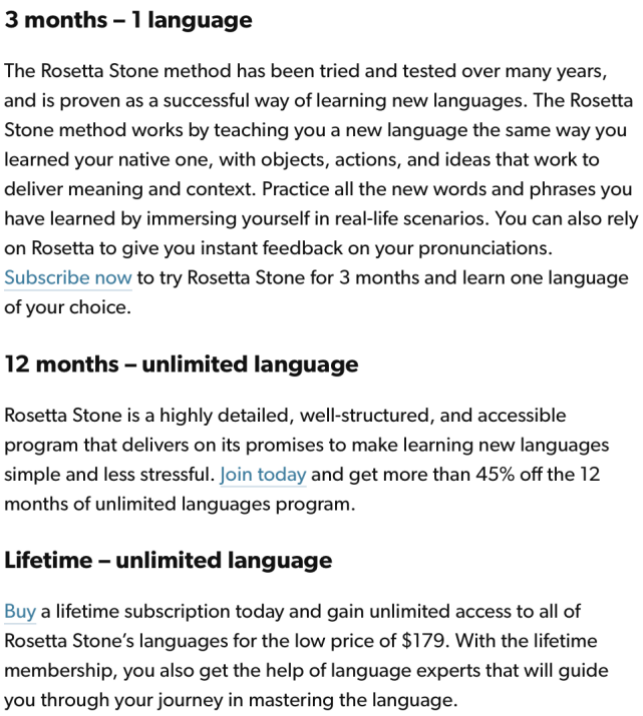
To identify subscriber profiles, I looked into modeling the subscriber profile using a K-Means clustering model. However, the model was inconclusive and could not identify conclusive groups to determine barriers on. Only two clusters were produced, cluster information was not conclusive to determine strong results. However, we should note that the highest correlation to subscription length was purchase_amount_usd_imputed and language_ALL. Taking that into consideration our next approach takes on these assumptions.

The next approach is to take a pure visualization step and observe trends that may lead to possible barriers.
I made the following assumptions to observe the trends:

- Subscription Duration: A subscriber duration period determines how long a consumer stays with the company.
    - Using this, we can determine what factors at certain points of subscriptions that affect the end of the subscription period by looking at the duration.
    - Subscriptions are tiered, but by looking at the number of subscribers in each tier we can see a decline of numbers at higher tiers. Tiers are represented at Month 3, and Month 6, and Month 12, meaning users stay with the company for 3 months, 6 months, and 12 months. The question we essentially want to answer is why do subscribers not subscribe for longer.
    - Based on the two above-mentioned variables, I made the assumption that customers are price sensitive and the more access to languages the customer has, the longer they are willing to stay.

I also did some research on Rosetta Stone's tier to better understand the tier trends I was seeing. This will serve as strong background information to determine each tier.





**3 months – 1 language**

The Rosetta Stone method has been tried and tested over many years, and is proven as a successful way of learning new languages. The Rosetta Stone method works by teaching you a new language the same way you learned your native one, with objects, actions, and ideas that work to deliver meaning and context. Practice all the new words and phrases you have learned by immersing yourself in real-life scenarios. You can also rely on Rosetta to give you instant feedback on your pronunciations. Subscribe now to try Rosetta Stone for 3 months and learn one language of your choice.

**12 months – unlimited language**

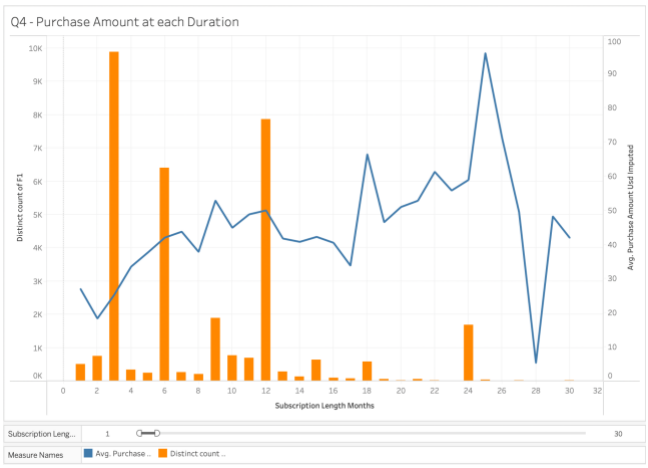
Rosetta Stone is a highly detailed, well-structured, and accessible program that delivers on its promises to make learning new languages simple and less stressful. Join today and get more than 45% off the 12 months of unlimited languages program.

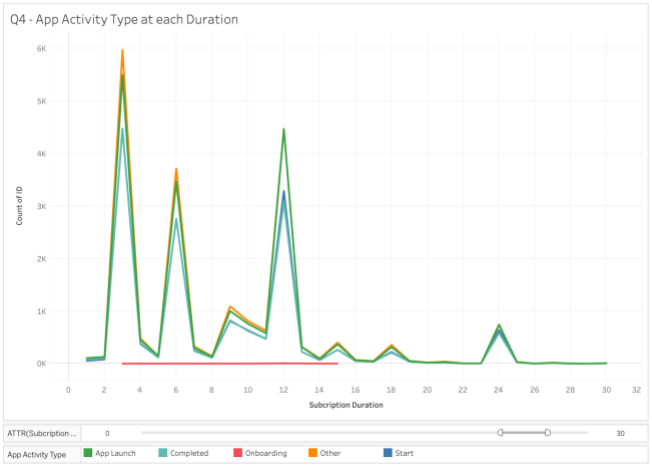
**Lifetime – unlimited language**

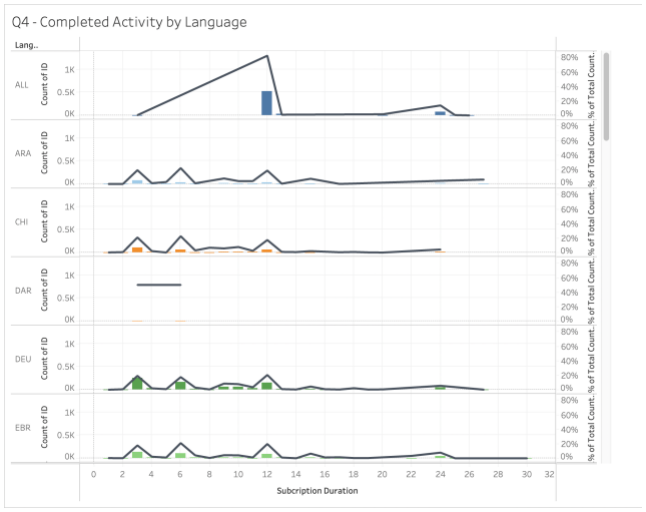
Buy a lifetime subscription today and gain unlimited access to all of Rosetta Stone's languages for the low price of $179. With the lifetime membership, you also get the help of language experts that will guide you through your journey in mastering the language.

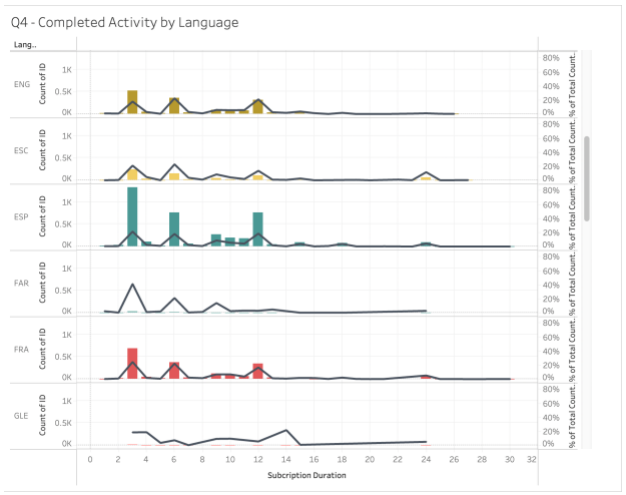
From these the visualization analysis, I saw the following:

- Barrier 1 & 2: Price Sensitive Customers & Language Education Length
    - Observations: Looking at the visualization "Purchase Amount at Each Duration", we see a drop off in customers from Month 3 and 6. This is because the plans here offer only one language. Subscribers who complete a language quickly will not need to renew the plan to finish their language. Most consumers are price sensitive, so they will want to finish early and not renew their subscription for another 3 months, thus explaining the drop in customers ending their subscription at 6 months. This can be visualized as activity drops sharply at the end of these tier plans, instead of continuing on with more subscription time on "Completed Activity by Language" and "App Activity by Duration". We also see a large increase in customers from 6 months to 12 months subscription. This is largely due to the fact that 12 months subscribers get access to all the languages, making the subscription tier more favorable and attract more customers to purchase the tier.
- Barrier 3: Lack of Marketing Effort to Retain Customers
    - In the visualization "Marketing Response at Each Duration ", we can see that marketing at each subscription duration is quite equal to each other. However, email marketing should significantly increase at the end of tiers in order to retain customers for longer subscriptions. It also matters what these emails are offering -- are they offering something or just informational?




*Business Opportunity to Fix Barrier: Entice subscribers through e-mail marketing to stay between Month 3 and Month 6 with a sample of a second language at the end of their subscription. A sample of another language after completing their first language may bring the subscriber to renew for a few more months to complete the new language. This can bring potentially 3 more months of revenue per customer that renews to complete another language.*

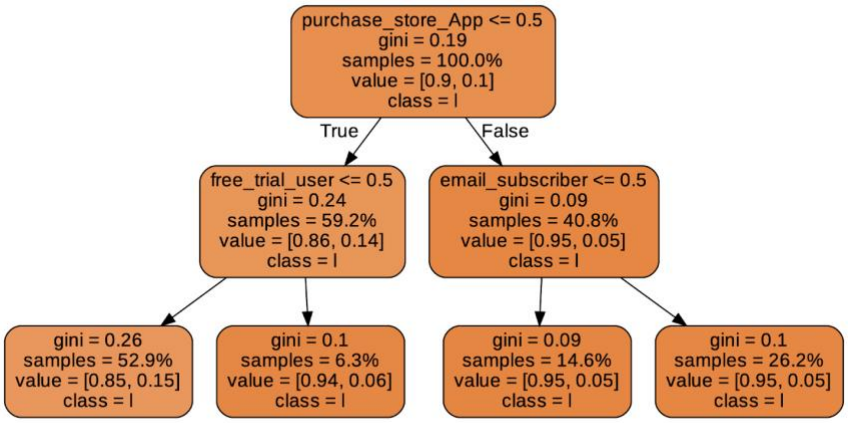
**Question 5 - Connor**

The previous analyses answered a lot of the initial questions that were presented in the data. Initially it was difficult to decide on additional analyses. Then there were a couple of analyses to do. The first one to do was to analyze, by proportion, to find what the most common languages being used on the app were. This is important because it can guide management about which languages to maintain or make content for. If a growing number of users are learning Japanese, then it is a good choice to scale, not linearly, the amount of content for that language. Scaling at a logistic rate means there is a trade-off between a linear scale and a no scale

approach. Linear scaling will cost Rosetta Stone a lot and not return as quickly. No scaling can cause users to become dissatisfied with content and hurt the growth in that language.

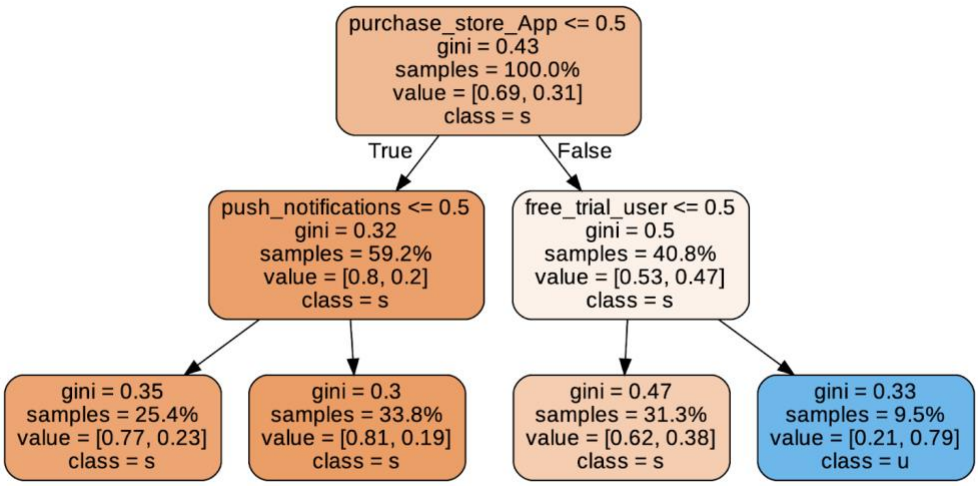
The next analysis dealt with metrics for management to aim for when signing a new subscriber. While, if every subscriber would fall under these categories, they wouldn't necessarily fall into the same groups, but they would *in aggregate* move towards more favorable groups. An example being that if the optimal subscriber purchased on the web and didn't have a free trial and then every subscriber had this profile, there would be benefits, but not everyone would convert into this group.

Decision trees were used with only Yes/No variables to predict a Yes/No Variable. As previously discussed, because periodical subscriptions are valuable only they were used. Rosetta stone should aim for these renewing subscriptions. The first decision tree evaluated if a customer was long term (>420 day subscription length). The best subscriber is someone who purchased on the website and did not have a free trial. This group made up 52.9% of the data and had a distribution of 15% long term subscribers. This may seem low, but the rate of long term subscribers if ~7% and 15% is significantly higher. The left node on the bottom is of interest.



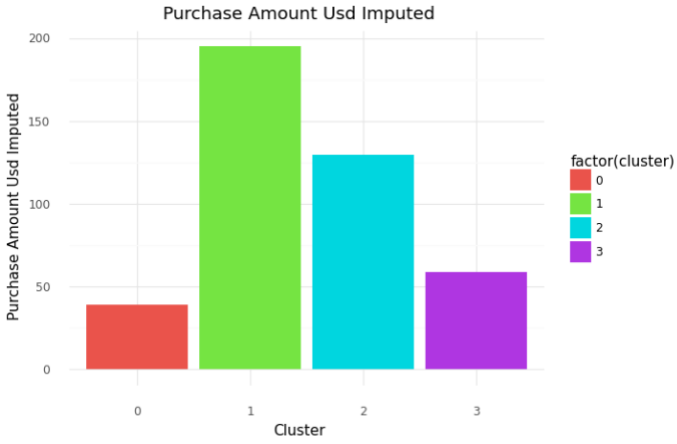

The second decision tree was classifying about what made a renewal transaction. The subscribers who make renewals are contributing to the importance and growth of a periodic subscription based service. The group that made the most renewal subscriptions purchased it on the app store and also was never a free trial user. This group made up 9.5% of the data and was made up of 79% renewal subscriptions. The free trial doesn't necessarily bring people to keep

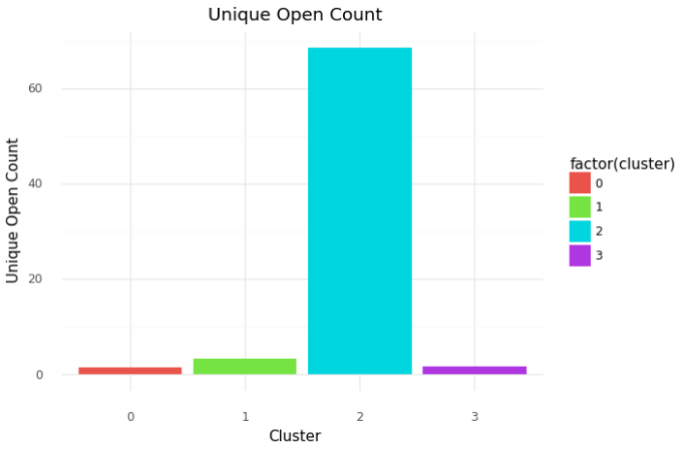
coming back. In the image below, the blue node is the one of interest.




**Extra Visualizations:**

**Clusters from Question 2:**




**What languages are being studied most (Nora)**


**Presentation Methodology - Nora, Aviv, Alyssa**

In order to properly convey what Rosetta Stone aims to provide their target audience with on a global scale, we chose to lead our presentation by introducing the mission statement. We believe that reflecting on our mission statement will allow us to understand where the customers are coming from and what they need so we can better serve them. We first take a look at our broad audience. Even though each of our customers are different, we can gather information from behavioral patterns of how our customer base is segmented. We recognize 4 types of customers: inactive, lifetime, avid and loyal. Differentiating our customers allows us to see how to restrategize our marketing efforts. We acknowledge that lifetime and loyal users are valuable

in their own way. Even though they have different subscription durations, we respect that they will go at their own pace and look forward to building a relationship with them. We also have readjusted our targeting for those who could buy additional products by finding our highest engagers. We shift to analyzing Rosetta Stone's barriers to overcome with first time users, completers and its email list. We offer new strategies each step of the way to improve Rosetta Stone's business model.

**Conclusion**

Rosetta Stone is a remarkable tool utilized globally. But remarkability isn't timeless. To create a product that establishes itself as a necessity, users must be compelled to it. Classifying Rosetta Stone's users is a first step in achieving this aim. The second step is determining the barriers to optimal user engagement, and the last step is to correct these limitations. With our analytical insight, Rosetta Stone will unquestionably expand communication and understanding across the globe.

Additional Links:

Visualizations:
https://public.tableau.com/views/MGSC410FinalGroupProject/BarriersDraft?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link
Currency Exchange: https://docs.google.com/spreadsheets/d/1T2Frs7T-k5W3e38Ryu_QjJ9uYRF3eWhfACU3VL45AQY/edit?usp=sharing