Team 4:

Kayla Cho
Betsy Heredia
Nikki Heredia
Ben Kahn
Grace Montgomery

Rosetta Stone Analytical Plan

Summary

The group approached this project by starting with meeting, organizing a timeline, creating and outlining drafts of our final documents, as well as delegating tasks to team members. To begin tackling the data, Grace and Ben started with data cleaning to provide a master data sheet with the variables in a usable format. Their initial findings were also used to facilitate the start of researching information on external factors that may have influenced the data by Betsy, Kayla, and Nikki. They also researched Rosetta Stone services and pricing. The team came up with multiple potential methods of analyzing the data in order to reach our goals. Grace and Ben coded the multiples models and discussed results with the team, where executive decisions were on the most applicable and valuable models. While Grace and Ben finalized the models and results, Betsy, Kayla, and Nikki finalized the word document deliverables and implemented all information into the presentation. Both teams within the group communicated throughout the process to ensure tasks were completed and fully understood by all members, and the team was ready to present.

Methodological Approach

First Steps

Initially, our group organized initial meetings to plan our approach to the project. We decided to meet twice a week, every Tuesday and Thursday or Friday. After reviewing the data set together and bouncing ideas off of one another, we delegated tasks based on our strengths. Grace and Ben were going to be in charge of cleaning the data while Nikki, Betsy, and Kayla focused on assembling the analytical plan and presentation. The team also brainstormed different questions about the Rosetta Stone to understand the data in a business context. Then we set up a general timeline to ensure we stay on track to meet the project deadline and keep each other accountable.

Company & Model Research

After looking over the data set, Betsy, Nikki, and Kayla plan to research other potential data sets to add to the data set we were given. Some resources they agreed to check included Kaggle, UCI Repository, and government websites.The girls agreed to research more about the company and answer the questions the team discussed. For example, the group decided we wanted to understand a general timeline of events happening in Rosetta Stone that could explain any patterns in the data. The girls then planned to meet to discuss each business goal and brainstorm possible methods and modeling approaches.

Data Cleaning

Ben and Grace planned to meet to brainstorm different ways to approach the data set. Their goal was to find a way to combine the two data sets and transform it into a mode of usable analysis while searching for any irregularities in the data along the way. After looking at the data set as a team in our first meeting, they also planned on finding a way to convert the currency to USD in order to compare the amount spent and add a row called ‘Email Engagement’ which was ‘open count’ / ‘sent count’.

Data Analysis

The team agreed to each help with modeling the data to address each business goal. Ben and Grace will take the lead in tackling the more complex models. Kayla, Betsy, and Nikki will help with the more simple models.

Business Presentation & Supporting Documents

While Ben and Grace are focused on developing the models, the rest of the team will focus on assembling the presentation and supporting documents. The group working on the presentation will work closely with Ben and Grace to effectively communicate the model information on the slides and continually update the presentation with any additional findings.

Final Steps

After the completion of the analysis and presentation, the team will meet before turning in the presentation to look over the project together and make any necessary changes. As well as ensuring each member of the team fully understands the process and results gathered throughout the project by each member. The team will also discuss who is presenting and what slides they will share with the class. Before the presentation, the group will meet again to rehearse the presentation to ensure it is cohesive.

Expected Outcomes

We anticipate that subscribers who have auto-renew on will be more valuable customers and less necessary to market to. We also expect free trial/demo users to be

our target market since they are testing out the product and haven't purchased it yet. Lastly, we believe that subscribers who have push notifications on and have high email engagement are targets to sell more products and services.

Timeline

4/26: Have cleaned data & researched ideas/methods of determining valuable subscribers
4/28: Complete Analytical Plan
4/30: Have a few models before meeting with Prof. Houldsworth
4/30: Meet with him
5/4: Finalize models
5/5: Presentation draft
5/6: Finish /run through presentation
5/7/21: Submit Project
5/13/21: Rehearse Presentation
5/14/21: Present Project

Action Plan

Methodological Approach

First Steps

After creating our analytical plan in the first meeting and brainstorming questions and ideas about the data set, we each worked on the tasks we were delegated. In addition to our meetings twice a week as a whole group, Ben and Grace met together to work on cleaning the data while Nikki, Betsy, and Kayla met together to clean up the analytical plan and conduct research.

Company & Model Research

While we were unable to find any relevant data sets to add to our project, our team dedicated a lot of time in making connections between what Ben and Grace found putting it in business context. While our original question was when the app was created, it evolved into looking at each peak in app usage and researching potential events that could explain those usage spikes.

Rosetta Stone Timeline

- September/October 2018- Spike of people doing free trials
- January 2019 New & improved iPhone app launch
- March 7, 2019 spike & stock jumped 23%

- May 30, 2019 50% Subscription for Father's Day
- June 20, 2019 HIAS Partnership
- August 22, 2019 Appointed new marketing & communication to board of directors
- December 19, 2019 New advisor
- February 26,2020 Free upgrade for 12-month, 24-month and lifetime new consumer subscriptions
- March 12, 2020 - Arkansas School Partnership
- March 19, 2020 - free subscription for K-12
- September 2020- Cambium Acquiring Rosetta Stone News

By researching the company's history, we learned that one of the largest spikes in app activity in March 2020 could be explained by the free three-month subscription services Rosetta Stone was offering to all K-12 students due to COVID-19. This provided students the opportunity to learn 24 different languages for up to 250 hours. This allowed us to understand why there was such a large spike during that time and possibly why the purchase amounts might not have been what we had expected.

The team also wanted to understand why the purchase amount was high for some customers including some renewals. We found that Rosetta Stone has a B2B partnership program with 20,000 educational institutions and 17,000 private and public sector organizations. However, we were unable to find data around how much each partnership was worth, when exactly they were made, and when the partners started using Rosetta Stone. To gather a better understanding of Rosetta Stone's offerings we researched their subscription rates and services. We learned that first time users on Rosetta Stone's website are offered a three-day free trial. After the trial, another three-months their services cost $11.99 a month or an additional 12 months cost $7.99 a month. Rosetta Stone also offers lifte-time services and unlimited languages that cost $179 (originally $199) or $299 with Rosetta live. Another additional service the company offers include live tutoring sessions where the first session is free and the following sessions are $14 or $19 per session. Another area we tried to research was the difference between a free trial user and a demo user. We concluded that both don't require a credit card and could not differentiate the difference despite trying to speak with their online customer service.

Once we had a better understanding of the business, we focused on brainstorming different ideas and models that address each business goal. For the first goal, "Determine the most valuable subscriber", we noted to do an ElasticNet Models to predict subscription event type, a linear regression model to determine the most important factors in determining the purchase amount, and a logistic regression model to predict the subscription type. To understand the subscriber segments present in the database, we listed out all of the unsupervised clustering models we were familiar with

including K-Means, Expectation Maximization with mixtures of Gaussian, and Hierarchical Agglomerative clustering and potential variables that we could cluster by. For goal three, we brainstormed that an ElasticNet or logistic regression. To identify the subscriber profile of those not continuing with their usage of the product and identify the barriers to deeper subscriber engagement, we thought by creating a column that counts the frequency of ID in App activity we could determine how active each subscriber is on app and find what similarities those who do use the app and those who don't have in common. To address the last goal of the project, we chose to leave it open ended because we knew once we created the models we would find additional insights that we could include.

Data Cleaning

To begin data cleaning, the Coding Team of the group, Grace and Ben, started to look for any sort of irregularities present in either of the data sets, as well as potential ways to combine the two data sets in order to more efficiently analyze the data present. This was done by adjusting the App Activity data in order to merge it with the Subscriber Information. Using the Pivot Table function in Excel, the data could be restructured as sums of various app sessions and app activity per ID number. This method removed the time variable, so we also used the App Activity data set in its raw form to analyze historical trends based on the date variable in Tableau (See data inconsistencies).

After restructuring the App Activity data set, it was very easy to combine it with the Subscriber information data set. However, we realized that the data sets had a different number of rows. This was because the Subscriber information had duplicated ID numbers. After many discussions of how to reconcile this error, our group elected to simply delete the rows, since there were only about 200 of the duplicated ID's, so this would not be removing too much data from the data set of 40,000 data points, and would be less time consuming than any manual cleaning process that we would need to do.

With both data sets together, the Coding team began to look at variables individually in order to clean them up and ensure their usability. This included converting all of the currencies to USD. For the conservation of time and resources, we decided to ignore the change in exchange rates over the years that the data was collected, although ideally this would also be taken into account.

Variables that were deemed unusable were removed from any models, such as Purchase Store. The Purchase Amount variable contained a NULL value if the purchase store was labelled as "App". Purchase Amount was our predicted variable in many of the models, so it was unnecessary to use Purchase Store as a predictor for Purchase Amount since it was always the same. Other variables were transformed to a format that would allow them to analzayed more easily. In some cases this involved

factoring. In some cases, this reformatting occurred solely in Excel. Grace was mostly responsible for the manipulation of the data set to produce a usable form because she is the most proficient in Excel. In some cases, we created new variables. For example, App Session Sum, a new variable, was added in order to quantify the total amount of App Sessions across the Web, App, and NULL values. In addition, we created new variables that were logical data types for all the variables that had 2 possible values. For example, the Lead Platform variable could be either Web or App. We created a variable called “isLeadPlatformWeb” which had a value of TRUE if the Lead Platform was Web and a value of FALSE if the Lead Platform variable was FALSE.

We set up a Github with all the code and remastered data sets so the entire team would have access to a master data file that would be ideal for performing our models and analysis on as well as access to the code for easier collaboration.

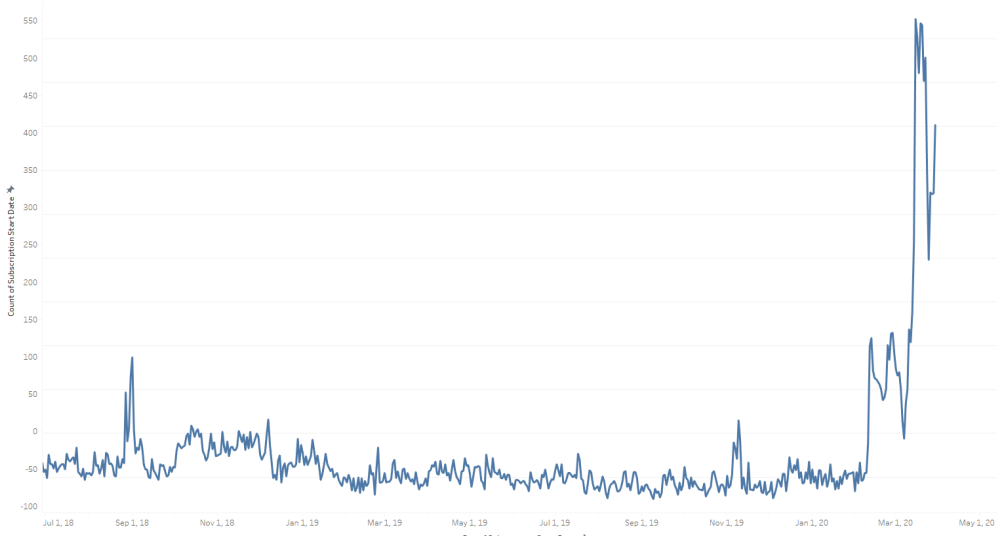
Summary Statistics

From the summary statistics, our team was looking to gain an initial understanding of the data we had at hand. This information was valuable in providing basic counts and statistics that would indicate if the data was skewed in any way. These metrics were vital in identifying the most common features and groups within variables that would act as a starting point for further analysis and research. The counts were performed on the variables: Language, Subscription Type, Subscription Event, Purchase Store, Currency, Demo User, Free Trial User, Auto Renew, Country, User Type, Lead Platform, Email Subscriber, and Push Notifications. This information told the team about our most common data in each column which led to some of the initial hypotheses used when creating our models. With the variables Purchase Amount (USD), Send Count, Open Count, Click Count, Unique Click Count, Unique Open Count, App Session IOS, App Session Android, and all of App Activity, the values produced gave insight into the behaviour of the average customer. Using this information in mind, the team moved forward in analyzing the data to find relationships within these customer behaviors and how they benefit Rosetta Stone in order to provide relevant information regarding our goals.

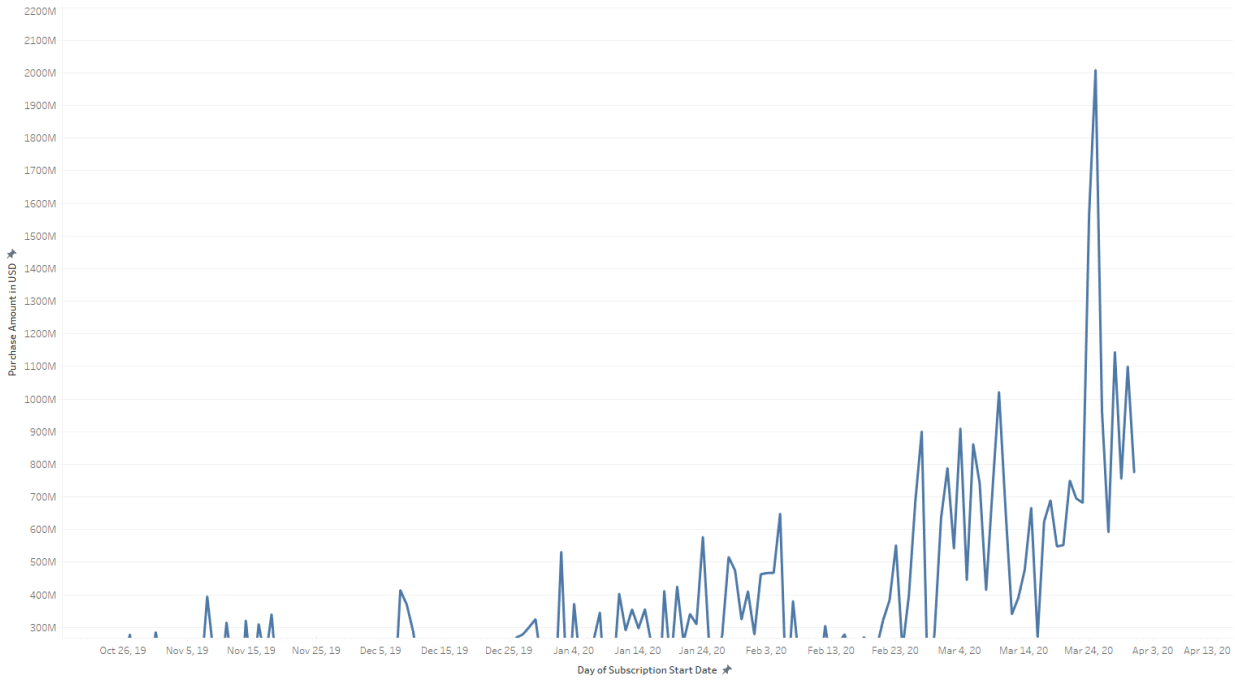
```
   Language    isSubscriptionTypeLifetime isSubscriptionEventTypeRenewal isPurchaseStoreWeb
ESP    :6003   FALSE:21394                FALSE:21988                    FALSE: 1615
ALL    :5184   TRUE : 5411                TRUE : 4817                    TRUE :25190
FRA    :2854
ENG    :1894
ITA    :1493
DEU    :1342
(Other):8035
   Currency     Purchase.Amount.in.USD isDemoUser    isFreeTrialUser isAutoRenew
USD    :20681   Min.   :        0      FALSE:21429   FALSE:23824     FALSE:17234
EUR    : 3248   1st Qu.:        0      TRUE : 5376   TRUE : 2981     TRUE : 9571
GBP    : 2480   Median :       39
BRL    :   79   Mean   :  3182874
CAD    :   76   3rd Qu.:      167
KRW    :   38   Max.   :145187900
(Other):  203
     Country      isUserTypeConsumer isLeadPlatformWeb isEmailSubscriber isPushNotifications
Europe   : 3318   FALSE:11440        FALSE:15530       FALSE:15813       FALSE:11027
Other    :12012   TRUE :15365        TRUE :11275       TRUE :10992       TRUE :15778
US/Canada:11475



  Send.Count        Open.Count       Click.Count       Unique.Open.Count Unique.Click.Count
Min.   :   1.00   Min.   :   0.00   Min.   :   0.000   Min.   :  0.000   Min.   : 0.000
1st Qu.:   4.00   1st Qu.:   0.00   1st Qu.:   0.000   1st Qu.:  0.000   1st Qu.: 0.000
Median :  10.00   Median :   1.00   Median :   0.000   Median :  1.000   Median : 0.000
Mean   :  36.96   Mean   :  10.08   Mean   :   2.859   Mean   :  4.563   Mean   : 0.465
3rd Qu.:  42.00   3rd Qu.:   7.00   3rd Qu.:   0.000   3rd Qu.:  3.000   3rd Qu.: 0.000
Max.   :4370.00   Max.   :4365.00   Max.   :4348.000   Max.   :196.000   Max.   :44.000
NA's   :7406      NA's   :7406      NA's   :7406       NA's   :7406      NA's   :7406
App.Session...android App.Session...ios App.Session...NULL App.Session...web
Min.   :  0.000       Min.   :  0.000   Min.   :0.0000     Min.   :  0.000
1st Qu.:  0.000       1st Qu.:  0.000   1st Qu.:0.0000     1st Qu.:  0.000
Median :  0.000       Median :  0.000   Median :0.0000     Median :  0.000
Mean   :  5.342       Mean   :  7.784   Mean   :0.4474     Mean   :  2.004
3rd Qu.:  0.000       3rd Qu.:  2.000   3rd Qu.:1.0000     3rd Qu.:  0.000
Max.   :938.000       Max.   :952.000   Max.   :1.0000     Max.   :299.000

App.Activity...App.Launch App.Activity...Completed App.Activity...NULL App.Activity...Onboarding
Min.   :  0.000           Min.   :  0.000          Min.   :0.0000      Min.   :0.0000000
1st Qu.:  0.000           1st Qu.:  0.000          1st Qu.:0.0000      1st Qu.:0.0000000
Median :  0.000           Median :  0.000          Median :0.0000      Median :0.0000000
Mean   :  5.516           Mean   :  3.401          Mean   :0.4474      Mean   :0.0005596
3rd Qu.:  5.000           3rd Qu.:  2.000          3rd Qu.:1.0000      3rd Qu.:0.0000000
Max.   :296.000           Max.   :342.000          Max.   :1.0000      Max.   :2.0000000

App.Activity...Other App.Activity...Start
Min.   :  0.000      Min.   :  0.000
1st Qu.:  0.000      1st Qu.:  0.000
Median :  1.000      Median :  0.000
Mean   :  3.999      Mean   :  3.061
3rd Qu.:  4.000      3rd Qu.:  2.000
Max.   :281.000      Max.   :249.000
```

Data Inconsistencies

While running our initial analysis, and creating quick visualizations in Tableau, we realized that there was a large spike in subscriptions around March 20th, 2020. The first graph below shows the large spike in users subscribing to Rosetta Stone. The company went from trends of under 100 subscribers per day to over 550 in one day.



In addition, there was a large purchase amount on that day too. The company showed total subscription revenue of over $2 billion US dollars. This seemed like a very large number, so we did some research into why this might be the case. As discussed in the Company and Model Research section above, we learned that the company offered a free three month subscription to new users. This explains the spike in subscribers during this time, however, it does not explain why there was such a large increase in revenue. Furthermore, we did some research on the company's financial data. We believed that there would be drastic market results for the company if their subscription revenue increased so dramatically.



Below is a screenshot of Rosetta Stone's revenue in 2020. As you can see, the company reported revenue in 2020 of approximately $49 million, which is much less than the $2 billion given in the data set.

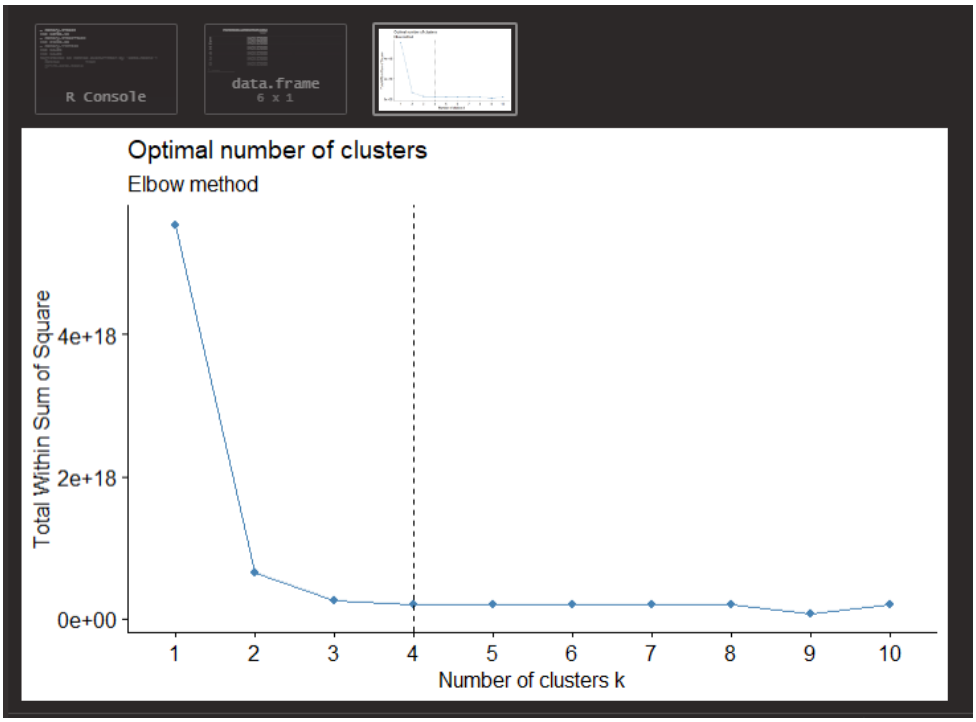
Rosetta Stone's Revenue (in thousands of dollars)

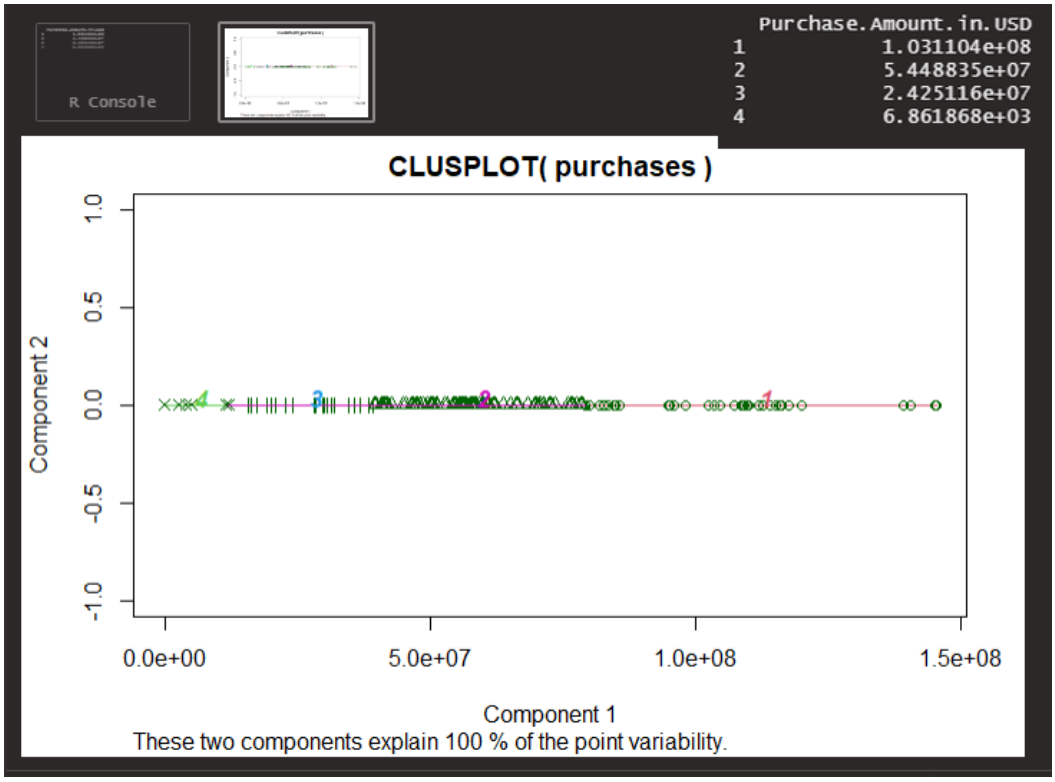
	Three months ended June 30,		Six months ended June 30,	
	2020	2019	2020	2019
Revenue	49,19	45,94	96,37	90,55
	$ 5	$ 2	$ 4	$ 3
Cost of revenue	11,43		22,53	17,28
	6	8,861	7	7
Gross profit	37,75	37,08	73,83	73,26

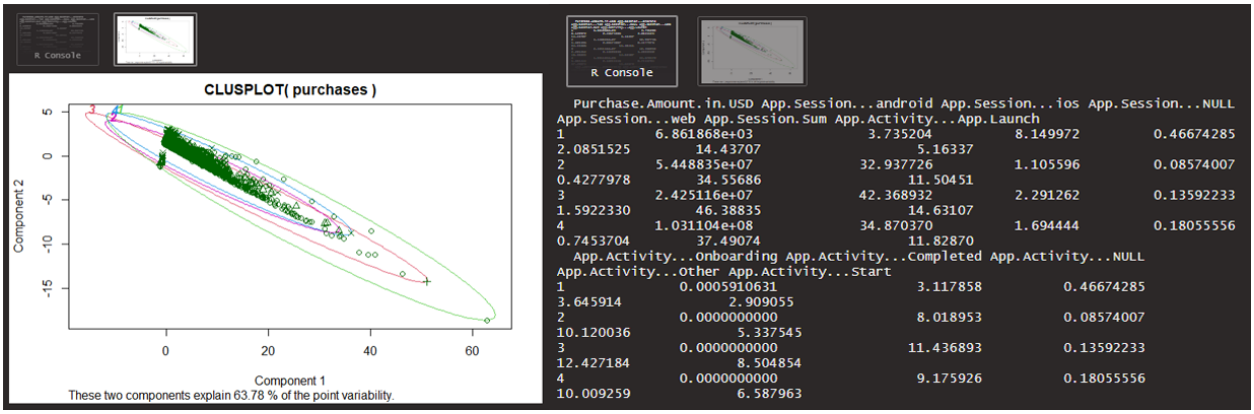
Data Analysis

Goal 1: Understanding the subscriber segments present in the database

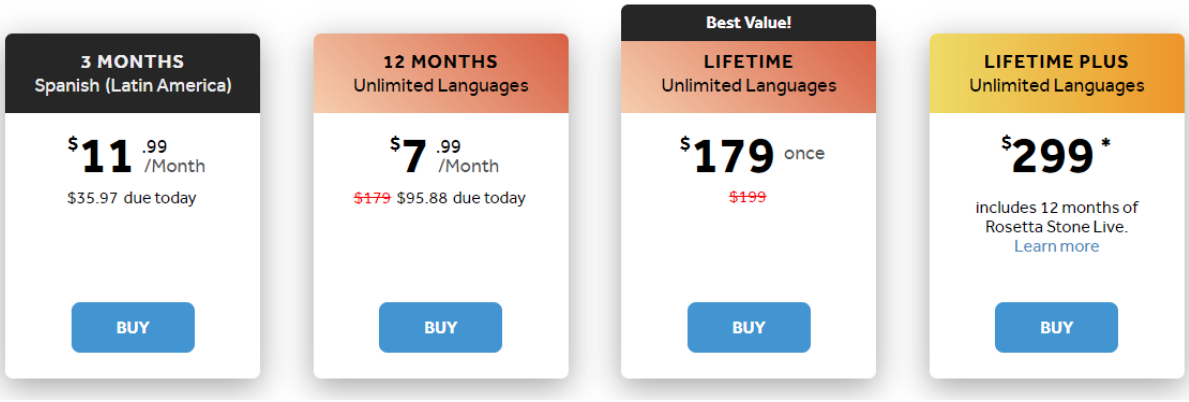
When attempting to determine the segments of subscribers present in the database, the first thing we thought to do was to perform a clustering on the dataset. This proved extremely difficult as clustering requires very specific parameters to run correctly. It requires there to be no NA or NULL values in the dataset in addition to only using numerical values. The first big roadbump was to decide what to do with the large amount of NULL values regarding purchase amount after it had been converted into USD. We decided to ignore them because, despite their large size, they would have made a massive middle cluster had we used the average value for them all and thrown off the rest of our clusters. In addition, we were only limited to using the app activity, app session, and purchase amount data. We found that the clustering, via the elbow method, seemed to be very inefficient as the values in the dataset that had massive values (billions of dollars) completely skewed it. We attempted this with both just the purchase amount and again with all of the numerical values in our dataset. Neither really told us anything regarding actual potential subscriber segments.


R Console
data.frame
6 x 1
Optimal number of clusters
Elbow method
Total Within Sum of Square
4e+18
2e+18
0e+00
1
2
3
4
5
6
7
8
9
10
Number of clusters k


R Console
Purchase.Amount.in.USD
1 1.031104e+08
2 5.448835e+07
3 2.425116e+07
4 6.861868e+03
CLUSPLOT(purchases)
Component 2
1.0
0.5
0.0
-0.5
-1.0
0.0e+00
5.0e+07
1.0e+08
1.5e+08
Component 1
These two components explain 100 % of the point variability.



Instead, we thought about the situation a bit more critically and realized that, for the most part, these purchase amounts likely had to correlate to set plans offered by Rosetta Stone to average consumers for their services. We discovered that they offer four plans to consumers, as can be seen below. Rosetta Stone also offers additional plans to large-scale businesses that we, despite our best efforts (article searching, third party review sites, and contacting Rosetta Stone directly), could not obtain pricing guidelines for. In addition, as mentioned above there were also many users who took advantage of free trials or free subscriptions as well as many data points which were too high to be real given Rosetta Stone's earning reports from this period. We felt we could use these values as they appeared to be the same for every language offered on the site.



We were able to determine five different subscriber segments based on our dataset and the information that we found on the Rosetta Stone website. These segments were manually created based on the outside research, and their frequency was counted. They include the free subscription group, the 3 month (1 language) group, the 12 month (unlimited language) group, the lifetime group, and the lifetime plus group. In addition, for the sake of covering the entire dataset, we also added two other subscriber segments to represent the group with NULL values and the group with the impossibly high values.

NULL	13095
Free Subscription	7010
3 Months	7652
12 Months	5111
Lifetime	730
Lifetime Plus	4825
Too High	1477
Total	39900

As you can see, our groups tend to be pretty similar in terms of users if you combine the lifetime and lifetime plus categories. The reason that the free category is so large is because a massive influx of people signed up for Rosetta Stone during that period in late March 2020 when a free subscription for students was offered due to the COVID-19 pandemic. We also theorize that 3 Months might be as high as it is because we found evidence of many people attempting to sign up for the free subscription being asked for their credit card information. When trying to contact Rosetta Stone directly through their chat feature about this issue and others, we were unable to get clear responses. The reason why the 3 Months might have been smaller otherwise is that if someone is trying to learn only one language in a short period of time, they are much more likely to use a free service such as Duolingo. This might also explain why 12 Months and the Lifetime categories are as high as they are. People who are more earnestly interested in learning new languages on a much larger scale are more likely to pay for a subscription service such as Rosetta Stone over a free application like Duolingo because the quality is bound to be much higher based on the variety of extra resources Rosetta Stone lists on their website.

Goal 2: Determine the most valuable subscribers

To determine the most valuable subscribers, our team decided to use the purchase amount in USD to measure a customers' value. We originally ran a linear regression model to understand the most important variables that affect purchase amount.

Linear Regression Model & Performance

```
mod1 <- lm(Purchase.Amount.in.USD~ .,
           data = df_train)

# print out a summary of the linear model
summary(mod1)
```

```
Residual standard error: 5984000 on 18455 degrees of freedom
  (10413 observations deleted due to missingness)
Multiple R-squared:  0.8524,    Adjusted R-squared:  0.828
F-statistic: 34.91 on 3052 and 18455 DF,  p-value: < 2.2e-16
```

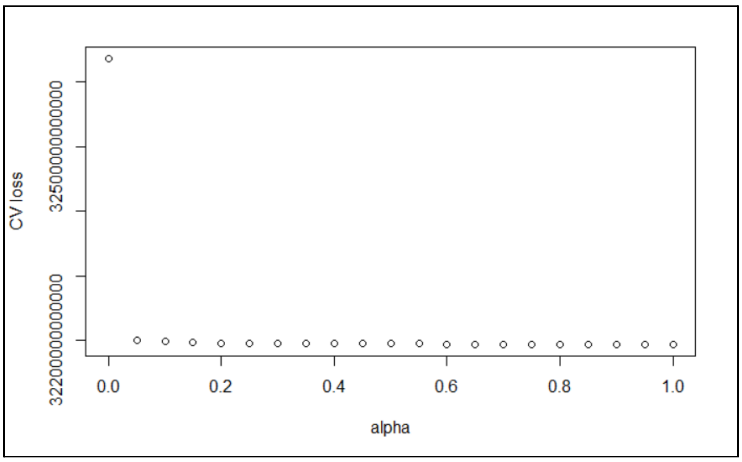
While the model had a high R-squared value, we felt that an ElasticNet model would help us to narrow down the most important variables because the model adds regularization penalties. We ultimately decided to utilize an ElasticNet model because the regularization can help us make inferences on the more accurate impact that each variable has with the purchase amount.

The ElasticNet Model

We included all the variables in the enet model except Purchase Store (see Data Cleaning), with purchase amount as the predictive outcome. In addition, we used a factored version of all the discrete variables, which included currency and language. Going into this model, we knew that there were some significantly large values in Purchase Amount that could potentially skew our results. However, we continued with the same dataset since we couldn't find a justified reason to remove them and it seemed more than just a couple of outliers. About 3.7% of the data appeared to be outliers of this sort. So, we are aware that maintaining these values will affect our results.

We first had to get the best alpha for the enet model that would give the lowest error. We ran an enet model, labeled "enet_model1" that held submodels testing different alphas. We came to the conclusion that our best alpha was 0.85. This means that the model used a mixture of mostly lasso, which converts unimportant coefficients to zero, but still maintains some ridge regression, which decreases variable coefficients to small amounts but not quite zero. This will explain why some variable coefficients disappeared in our model.

```r
enet_mod1 <- cva.glmnet(Purchase.Amount.in.USD ~ .-isPurchaseStore,
                        data = master_data,
                        alpha = seq(0,1, by = 0.05))

plot(enet_mod1)

minlossplot(enet_mod1,
            cv.type = "min")
```



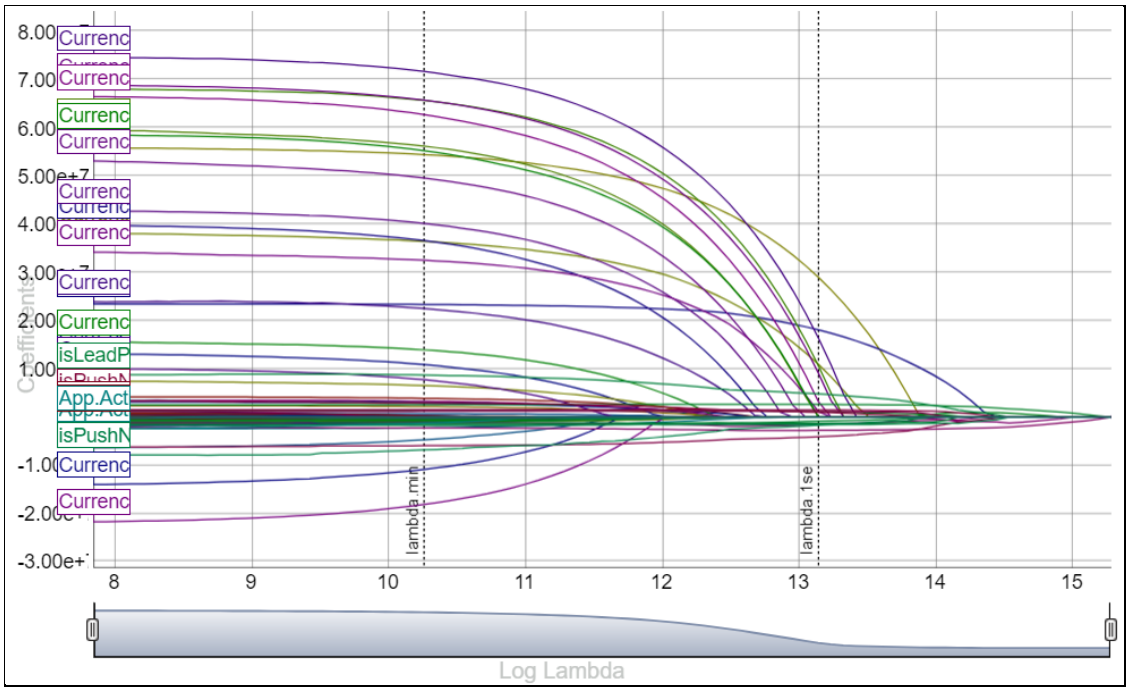
alpha <dbl>	lambdaMin <dbl>	lambdaSE <dbl>
0.85	12152.72	877524.4

Once we knew the parameters for the ideal enet model that reduces error, we coded that model with the specific alpha of 0.85 and labeled it “general_mod”. In addition, we also knew we wanted to observe the coefficients that came from the model with the lambda.1se penalty because it applies a little bit more penalty while still being within one standard deviation away from the penalty with the minimum error.

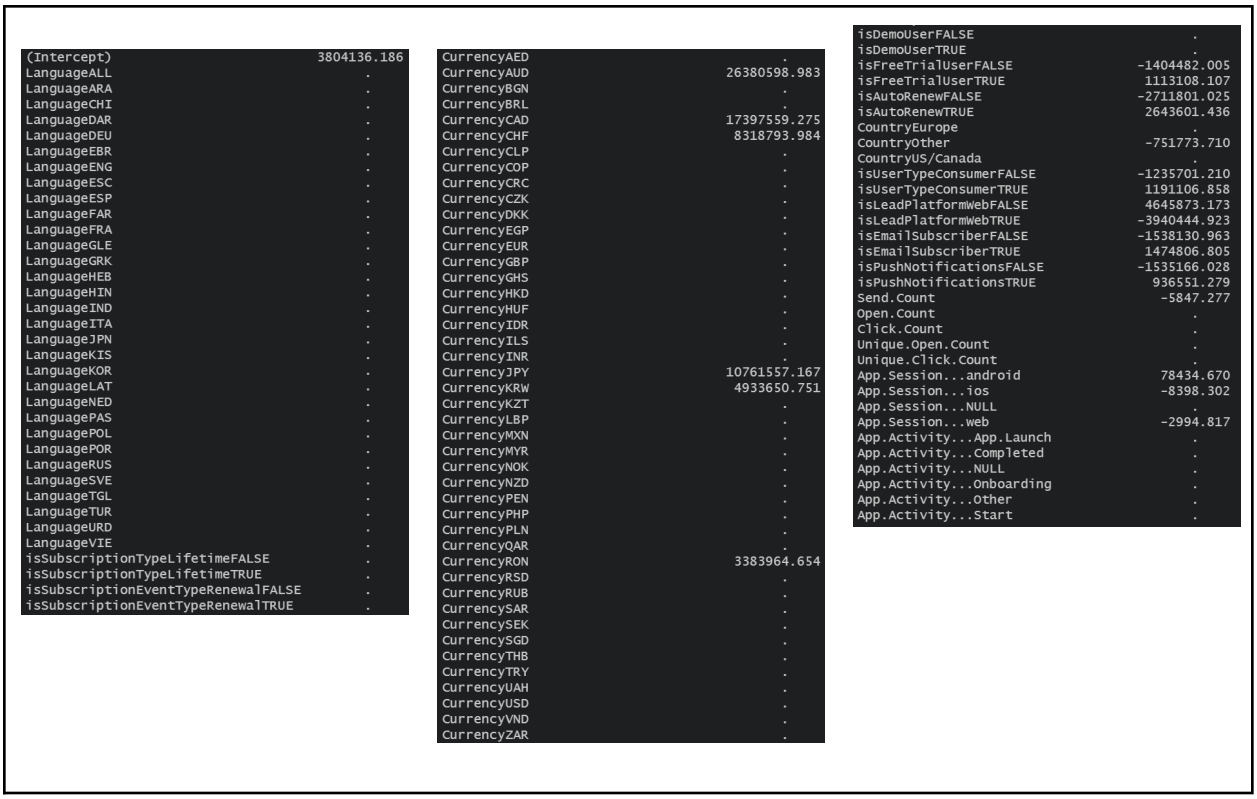
```r
```{r}
general_mod <- cv.glmnet(Purchase.Amount.in.USD ~ .-isPurchaseStoreWeb,
                         data = master_data,
                         alpha = 0.85)

coefpath(general_mod)

```
```



For our purpose of knowing which variables will have a higher impact on the growth of purchase amount, we looked at the coefficients with 1se lambda and got 23 non-zero variables. From those non-zero variable coefficients 13 were positive. We considered the data along with business reasoning and decided to focus on “isLeadPlatformWebFALSE”, “isAutoRenewTRUE”, “isEmailSubscriberTRUE”, “isUserTypeConsumerTRUE”, “isFreeTrialUserTRUE”, and “isPushNotificationsTRUE” as our relevant variables for our goal. Essentially we did not consider language subject, currency used, and android session users. In other words, we concluded that the profile of our most valuable customers in terms of revenue are those that are mainly mobile users, have auto renewal subscription on, subscribe to our emails, started as a free trial user, utilize their push notification, and are regular consumers, rather than classified as the “other” type of user in the data set.



(Intercept)	3804136.186
LanguageALL	.
LanguageARA	.
LanguageCHI	.
LanguageDAR	.
LanguageDEU	.
LanguageEBR	.
LanguageENG	.
LanguageESC	.
LanguageESP	.
LanguageFAR	.
LanguageFRA	.
LanguageGLE	.
LanguageGRK	.
LanguageHEB	.
LanguageHIN	.
LanguageIND	.
LanguageITA	.
LanguageJPN	.
LanguageKIS	.
LanguageKOR	.
LanguageLAT	.
LanguageNED	.
LanguagePAS	.
LanguagePOL	.
LanguagePOR	.
LanguageRUS	.
LanguageSVE	.
LanguageTGL	.
LanguageTUR	.
LanguageURD	.
LanguageVIE	.
isSubscriptionTypeLifetimeFALSE	.
isSubscriptionTypeLifetimeTRUE	.
isSubscriptionEventTypeRenewalFALSE	.
isSubscriptionEventTypeRenewalTRUE	.

CurrencyAED	.
CurrencyAUD	26380598.983
CurrencyBGN	.
CurrencyBRL	.
CurrencyCAD	17397559.275
CurrencyCHF	8318793.984
CurrencyCLP	.
CurrencyCOP	.
CurrencyCRC	.
CurrencyCZK	.
CurrencyDKK	.
CurrencyEGP	.
CurrencyEUR	.
CurrencyGBP	.
CurrencyGHS	.
CurrencyHKD	.
CurrencyHUF	.
CurrencyIDR	.
CurrencyILS	.
CurrencyINR	.
CurrencyJPY	10761557.167
CurrencyKRW	4933650.751
CurrencyKZT	.
CurrencyLBP	.
CurrencyMXN	.
CurrencyMYR	.
CurrencyNOK	.
CurrencyNZD	.
CurrencyPEN	.
CurrencyPHP	.
CurrencyPLN	.
CurrencyQAR	.
CurrencyRON	3383964.654
CurrencyRSD	.
CurrencyRUB	.
CurrencySAR	.
CurrencySEK	.
CurrencySGD	.
CurrencyTHB	.
CurrencyTRY	.
CurrencyUAH	.
CurrencyUSD	.
CurrencyVND	.
CurrencyZAR	.

isDemoUserFALSE	.
isDemoUserTRUE	.
isFreeTrialUserFALSE	-1404482.005
isFreeTrialUserTRUE	1113108.107
isAutoRenewFALSE	-2711801.025
isAutoRenewTRUE	2643601.436
CountryEurope	.
CountryOther	-751773.710
CountryUS/Canada	.
isUserTypeConsumerFALSE	-1235701.210
isUserTypeConsumerTRUE	1191106.858
isLeadPlatformWebFALSE	4645873.173
isLeadPlatformWebTRUE	-3940444.923
isEmailSubscriberFALSE	-1538130.963
isEmailSubscriberTRUE	1474806.805
isPushNotificationsFALSE	-1535166.028
isPushNotificationsTRUE	936551.279
Send.Count	-5847.277
Open.Count	.
Click.Count	.
Unique.Open.Count	.
Unique.Click.Count	.
App.Session...android	78434.670
App.Session...ios	-8398.302
App.Session...NULL	.
App.Session...web	-2994.817
App.Activity...App.Launch	.
App.Activity...Completed	.
App.Activity...NULL	.
App.Activity...Onboarding	.
App.Activity...Other	.
App.Activity...Start	.

Goal 3: Identify the most likely subscribers who could be sold additional products or services.

To identify the most likely subscribers who could be sold additional products or services, our team looked at the Elastic Net we created to find the most valuable subscribers and analyzed the top coefficients that had an inverse relationship with the purchase amount in USD. Since we excluded language and currency as factors we were left with the following variables.

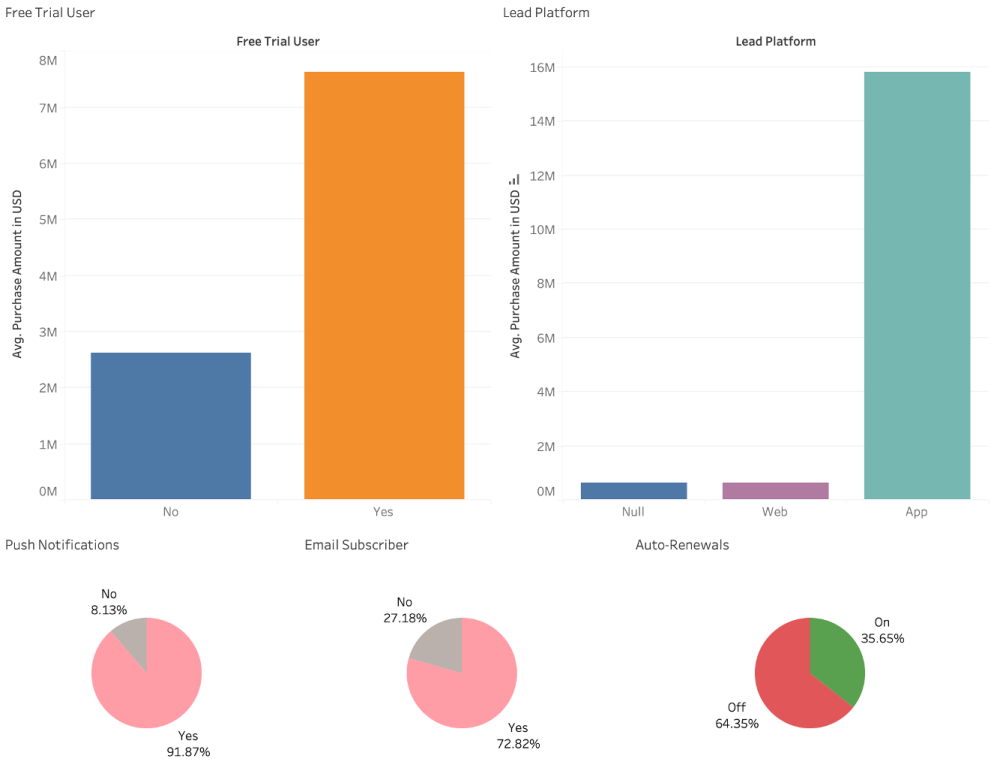
Top 5 Negative Elastic Net Top Coefficients

isLeadPlatformWebTRUE	(4,038,520.22)
isAutoRenewFALSE	(2,726,139.54)
isPushNotificationsFALSE	(1,654,314.55)
isEmailSubscriberFALSE	(1,538,339.85)

isFreeTrialUserFALSE	(1,499,831.93)

This table shows that subscribers who use the web app to access Rosetta Stone, don't have auto-renew or push notifications are less valuable to the company, since they negatively impact purchase amounts in this model. Also users who are not email subscribers or are not free trial users also have a negative relationship with purchase amount. Since these factors negatively impact the value of a subscriber, we concluded that subscribers who do not have their push notifications on, are not subscribed to our emails, and do not have auto-renew on are the subscribers who could potentially be sold additional products or services to increase their value to the company. To look further into what factors drive subscribers to renew their subscription, we created a logistic regression that predicts if a subscriber renews or not.

Top 5 Negatively Value Factors



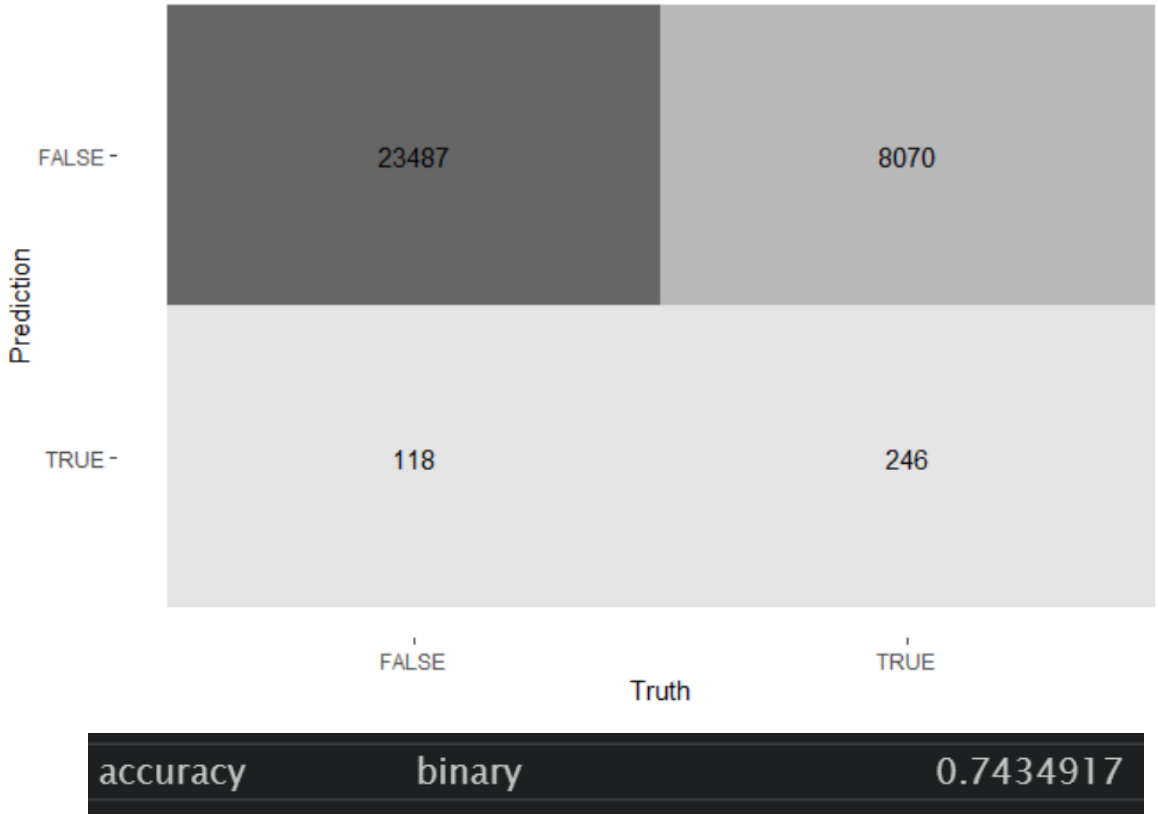
Logistic Regression Model

```r
train_prop <- 0.8
master_split <- initial_split(master_data, prop = train_prop)
master_train <- training(master_split)
master_test <- testing(master_split)
```

```r
logit_fit1 <- glm(isSubscriptionEventTypeRenewal ~ Language + isDemoUser + isFreeTrialUser +
isEmailSubscriber + Country + App.Session...android + App.Session...ios,
                  family = binomial,
                  data = master_train)

#exponentiation
exp(logit_fit1$coefficients)

preds_train <- predict(logit_fit1, newdata = master_train)
preds_test <- predict(logit_fit1, newdata = master_test)


results_train <- data.frame(
  `truth` = as.factor(master_train$isSubscriptionEventTypeRenewal),
  `Class1` =  preds_train,
  `Class2` = 1 - preds_train,
  `predicted` = as.factor(ifelse(preds_train > 0.4,
                                 "TRUE","FALSE")))

results_test <- data.frame(
  `truth` = as.factor(master_test$isSubscriptionEventTypeRenewal),
  `Class1` =  preds_test,
  `Class2` = 1 - preds_test,
  `predicted` = as.factor(ifelse(preds_test > 0.4,
                                 "TRUE","FALSE")))
```

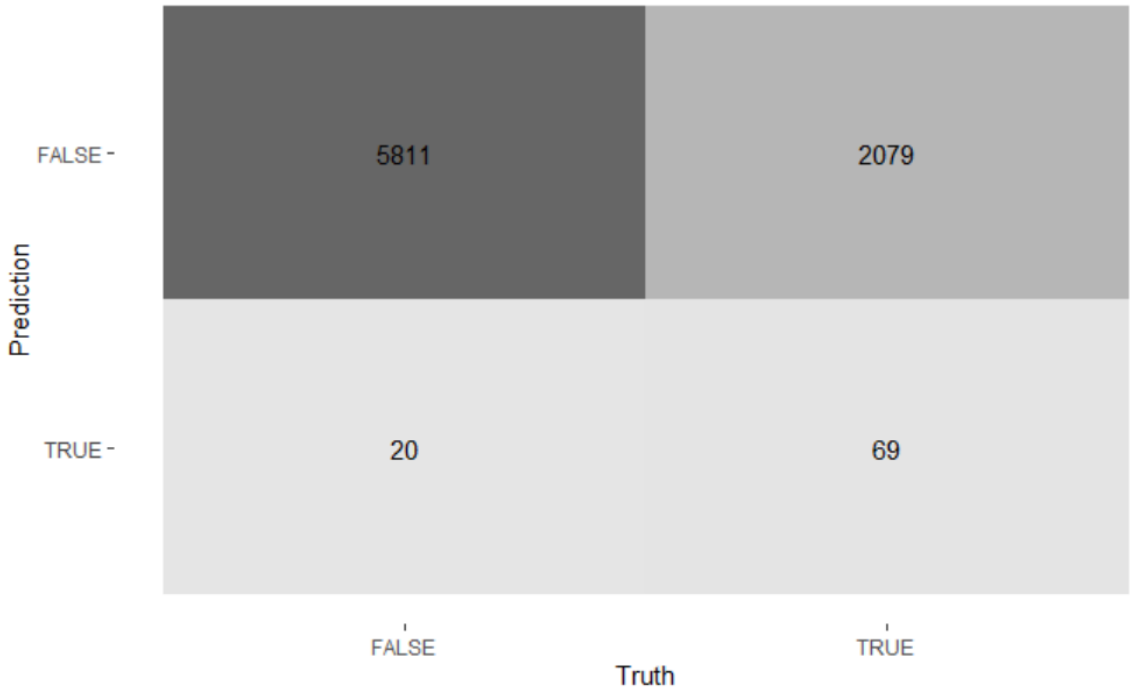
To measure how well our model performs when predicting if a subscriber is renewing or not, we created confusion matrices and measured its accuracy. While we had an accuracy score of approximately 74% for both the train and test set, we noticed our model was producing many false negatives. In other words, our model is better at predicting if the subscriber is doing an initial purchase and often mistakes renewal subscribers as initial purchase subscribers. After looking closer into the data, we realized that almost 75% of the data set is initial purchases and the other 25% is renewals. We also found that 20% of the initial purchases happened in March 2020, around the time period of COVID-19 shut downs in the United States.

Train Model Performance



accuracy	binary	0.7434917

Test Model Performance



accuracy	binary	0.7369345

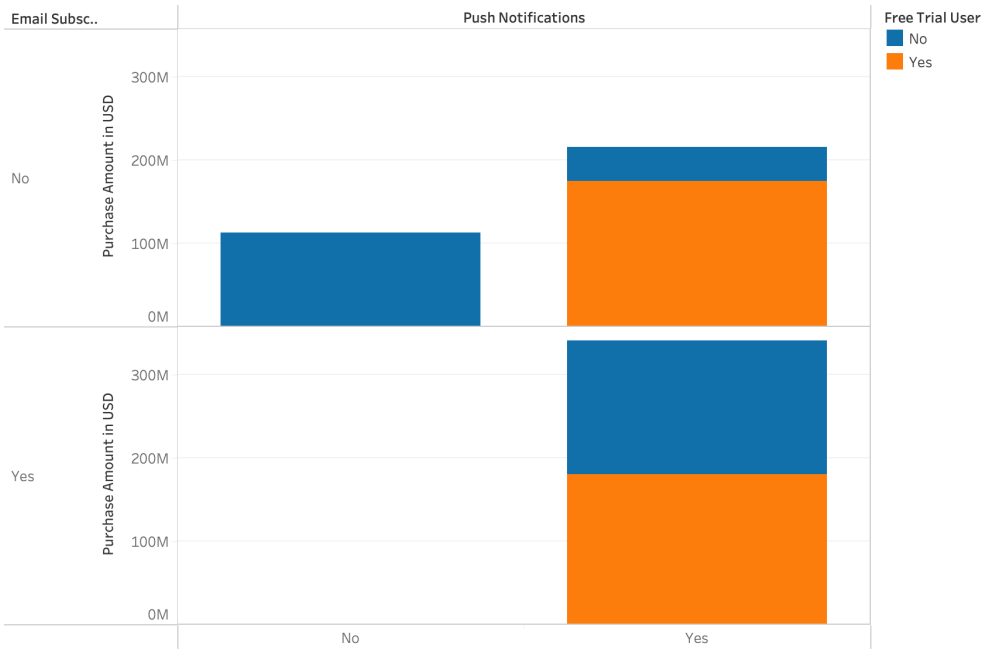
Logistic regression coefficients

(Intercept)	isDemoUserTRUE	isFreeTrialUserTRUE
-1.891874654	0.661312627	1.696568501
isEmailSubscriberTRUE	CountryOther	CountryUS/Canada
0.141420798	0.553574990	0.134063654
App.Session...android	App.Session...ios	isPushNotificationsTRUE
0.001221720	0.003554371	-0.123018505

In conclusion, when looking at the subscriber types, we chose to focus on subscribers who have limited subscriptions and auto-renew off because this factor had the second highest negative relationship in regard to purchase amount. An important thing to note is we did not look at life-time subscribers as those who could be sold additional services or products because the only additional service they could purchase is upgrading to Rosetta Live or purchasing live tutoring sessions — something we didn't have data on. To find out exactly who could be sold additional products or services, we created a visualization that shows, out of those who don't have auto-renew on, how many of them are free subscribers, how many of them are email subscribers, how many have their push notifications on, and how much their purchase amounts were. This graph shows that non-auto-renew users who are doing a free trial have their email subscription on and/or push notifications activated are the most viable users to purchase more products or services. Since Rosetta Stone has an avenue to actively communicate with the

subscriber through push notifications and/or email, they can encourage those users, especially those free trial users, to use auto-renew and try other products and services.

Target Growth Customers



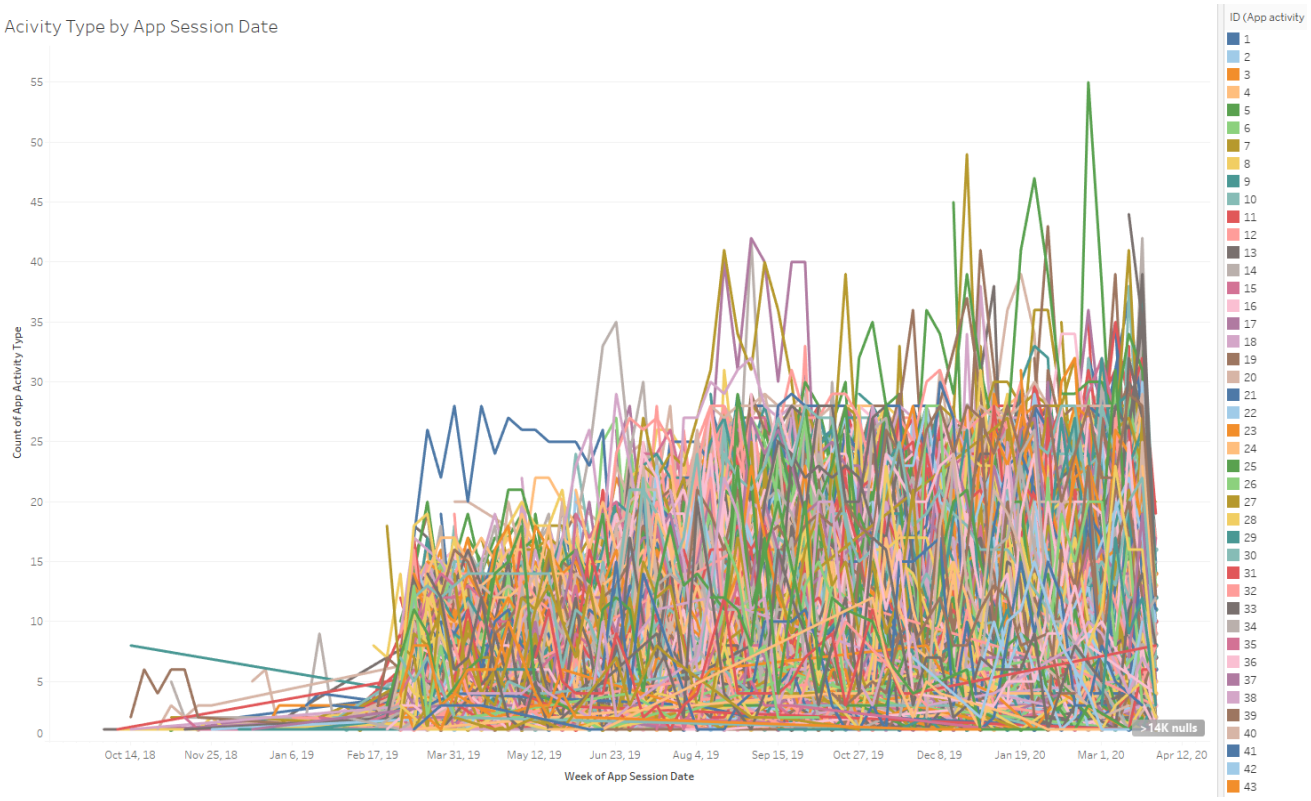
Goal 4: Identify the subscriber profile of those not continuing with their usage of the product and identify the barriers to deeper subscriber engagement where possible.

Originally, as discussed in Goal 3, we believe that subscribers who have push notifications turned on and are subscribed to emails are higher valued customers for Rosetta Stone. As such, any subscribers who are not do not have push notifications turned on and are not subscribed to the company's emails are opportunities for deeper subscriber engagement. However, we did some further analysis and gathered slightly different insights.

Time Series Analysis

Goal 4 contains two parts. The first part asks us to discover the subscriber profile of those who are not continuing with their usage of the product. With the given data set, this would be possible by tracking app activity, and discovering when app activity stops occurring based on ID number. An easier way to do this would be to have historical data of subscriber renewal or purchase history to see where certain ID numbers stopped subscribing.

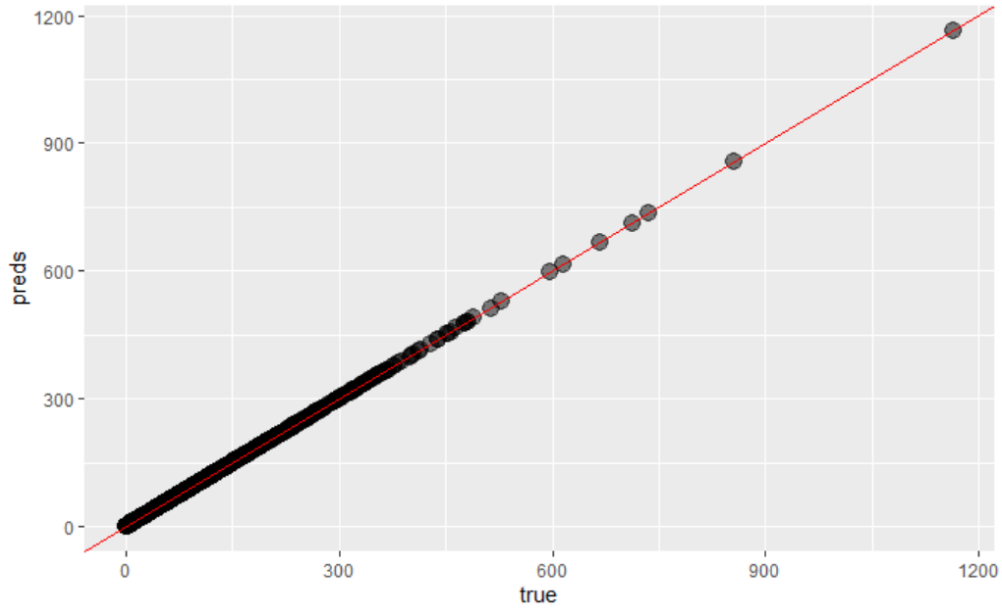
We created a visual in Tableau of the Count of App Activity by App Session Date. We filtered based on ID numbers to try and identify some trends. Below is the resulting graph, which was extremely hard to read and gave us no insights.



Linear Regression

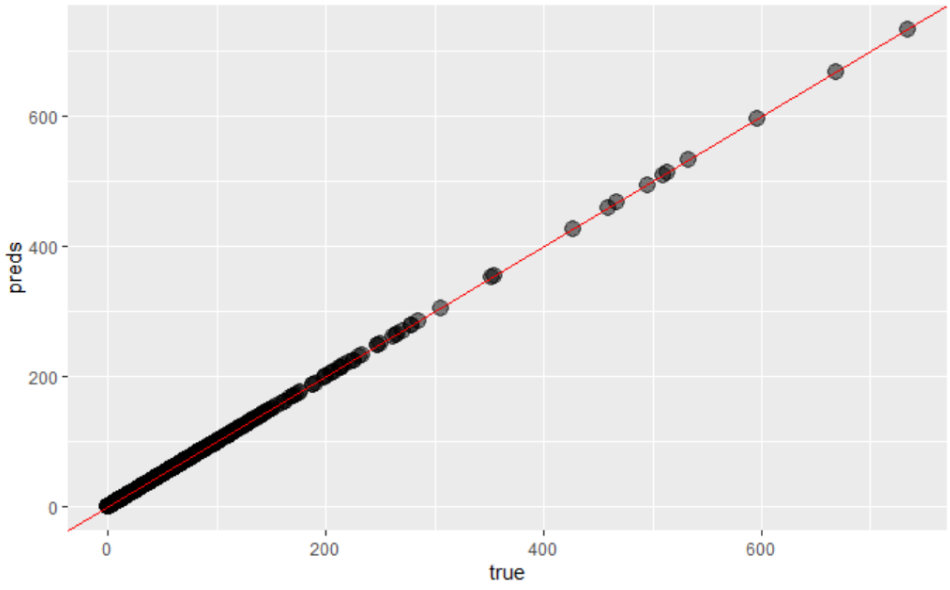
We also attempted to run a linear regression to see what variables were influential to the amount of app activity that an individual engages in. However, we determined that these model results were also unusable. As seen in the below graph, we plotted the model's predicted values against the true values. Surprisingly, this was a straight line. While the R2 value of 1 seems to indicate that the model is very accurate, we realized that this should not be possible. We did some more analysis and we were not able to discover why the accuracy was too high. We are thinking maybe it is because many of the predictor variables are factors.

Training Set Linear Regression Model Predicted vs. True Values



rmse	standard	0.000000000004335235
rsq	standard	1.000000000000000000
mae	standard	0.000000000004094900

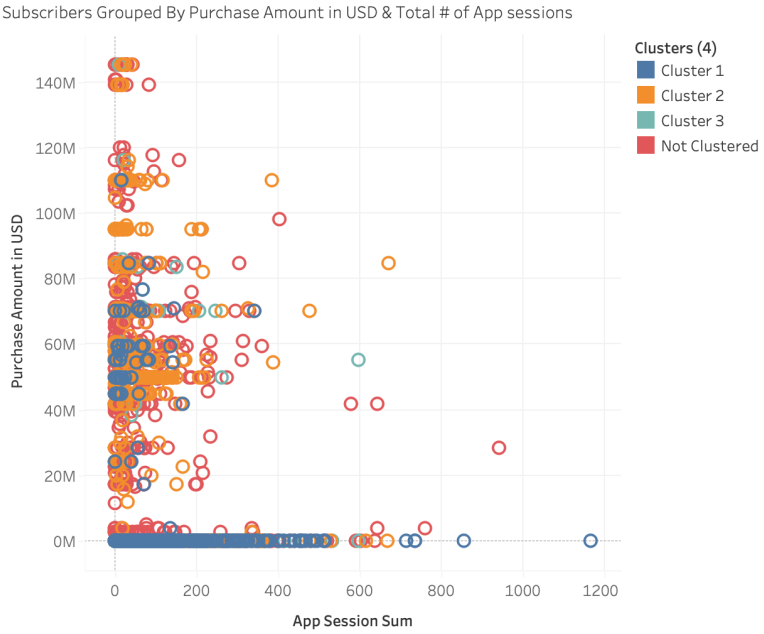
Testing Set Linear Regression Model Predicted vs. True Values



rmse	standard	0.000000000004318397
rsq	standard	1.000000000000000000
mae	standard	0.000000000004110203

Also, from the data we were provided, we used K-Means clustering in Tableau to create subscriber groups based on their total app activity (the total number of times the user goes on the app, web or other), email open rate, and purchase amount in USD to identify what group of subscribers don't actively use Rosetta Stone's products.

K-Means Clustering Model



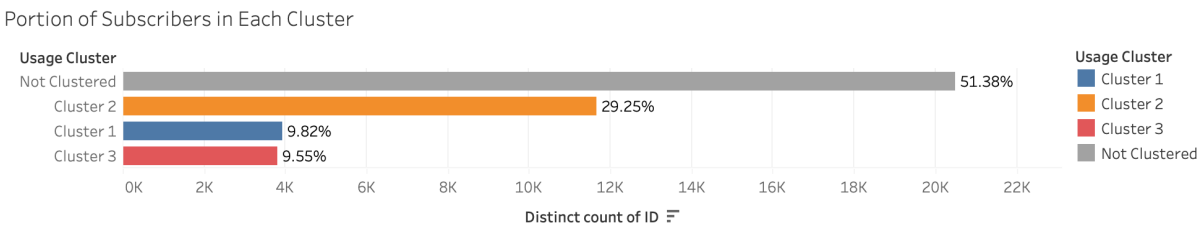
Summary Diagnostics

Number of Clusters:	3
Number of Points:	19399
Between-group Sum of Squares:	2411.4
Within-group Sum of Squares:	338.05
Total Sum of Squares:	2749.5

		Centers		
Clusters	**Number of Items**	**Avg. App Session Sum**	**Avg. Purchase Amount in USD**	**Avg. Email Open Rate**
Cluster 1	3918	23.835	8.1021e+05	0.92014
Cluster 2	11672	12.931	4.009e+06	0.033732
Cluster 3	3809	17.28	1.1857e+06	0.44115
Not Clustered	20501			

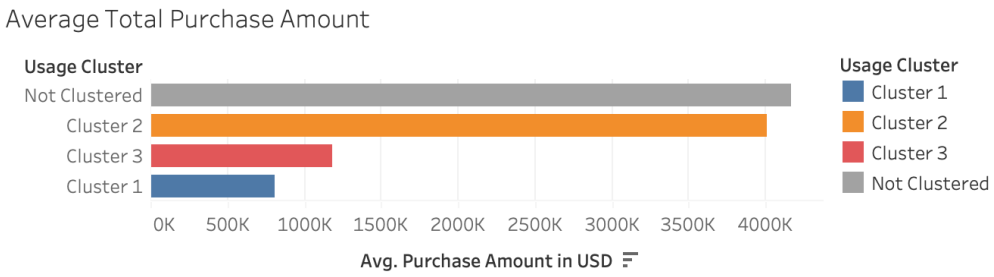
Analysis of Variance:

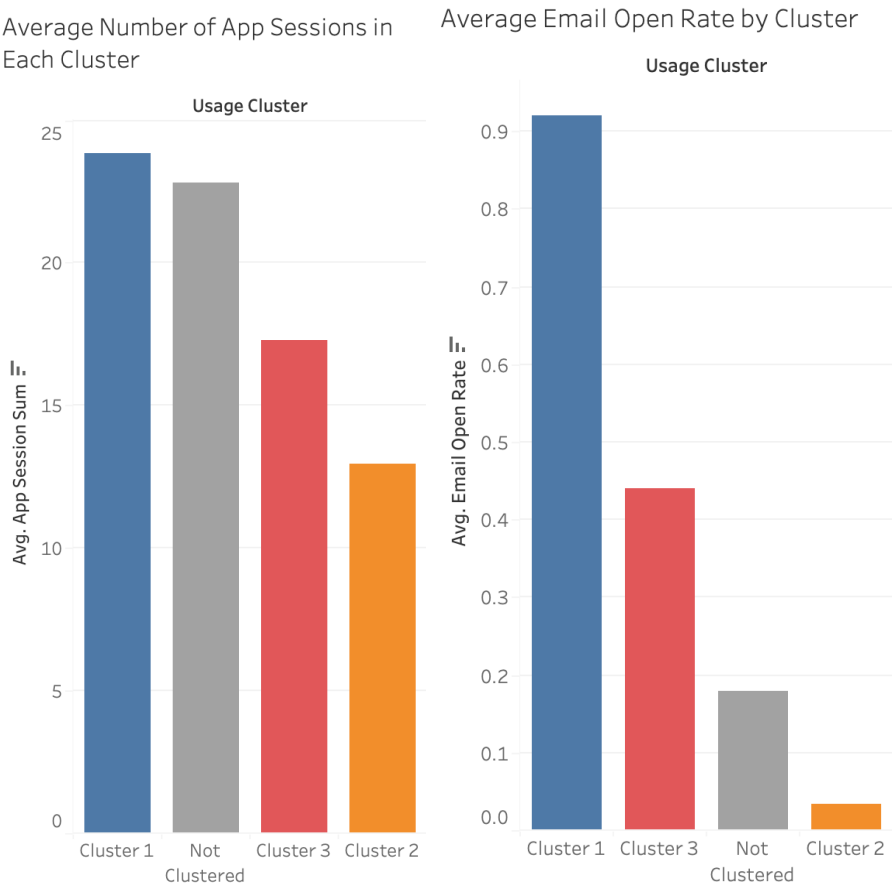
			Model		Error	
Variable	**F-statistic**	**p-value**	**Sum of Squares**	**DF**	**Sum of Squares**	**DF**
Avg. Email Open Rate	9132.0	0.0	2409.0	2	2558.0	19396
Avg. Purchase Amount in USD	116.9	0.0	2.016	2	167.2	19396
Avg. App Session Sum	106.6	0.0	0.2628	2	23.91	19396



Our clustering model performed well in grouping the subscribers because approximately .8770 of the variance is explained by the model. While 51.38% of our subscribers weren't clustered, our focus was on Cluster 2 since they appeared to be the most inactive. We chose to move forward in using this clustering model because the performance was high and Cluster 2 accounted for 29.25% of the subscribers. This model shows Cluster 2's center has the lowest average of total app sessions, purchased the most amount in USD, and had the lowest email open rate. We then looked more in depth about the averages of all the clusters to compare them which aligned with the center's results.

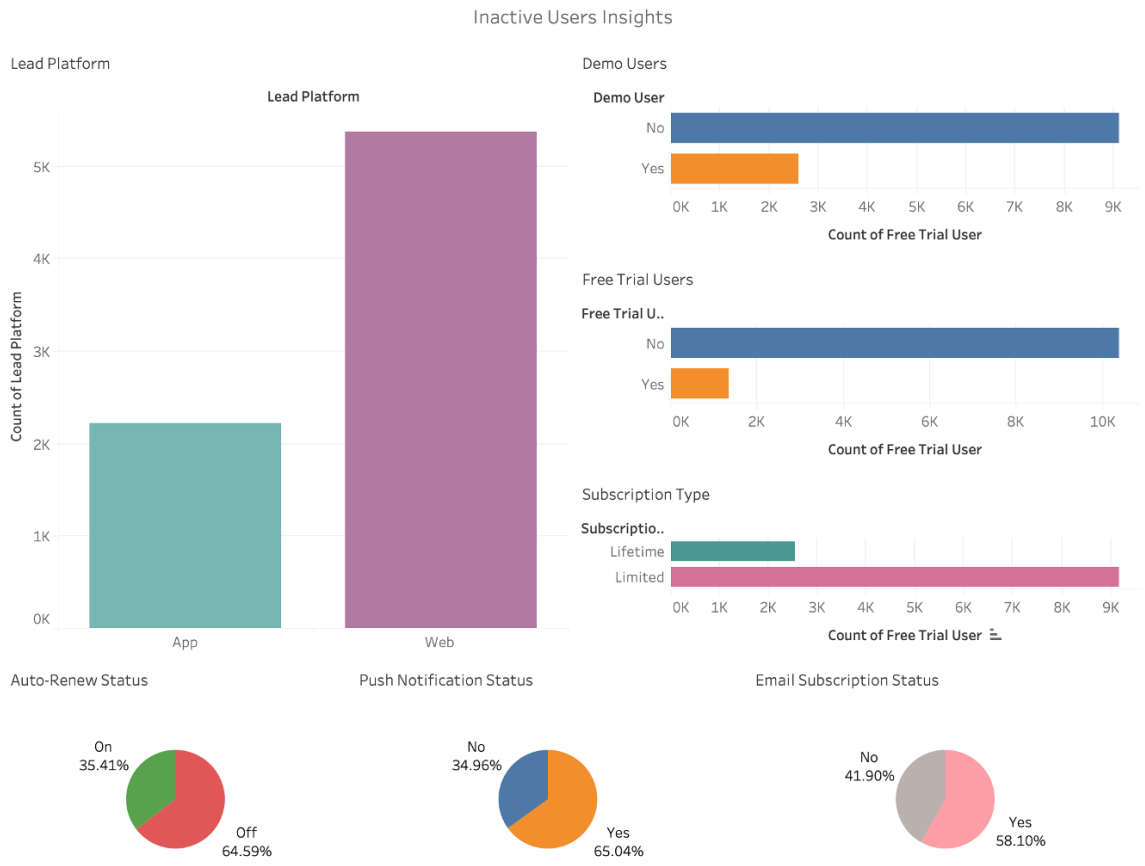
Cluster Average Comparisons





After understanding how the inactive subscriber cluster performed in relation to the other clusters, we looked deeper into understanding the inactive users' profile. We found that the majority of the group was limited subscribers who weren't free or demo users. A large portion of this group also used the web to interact with Rosetta Stone, had their auto-renew off, had push notifications on, and were subscribed to emails.

Inactive User Dashboard



Some barriers we identified that could be hurting our subscriber engagement were auto-renewals, lead platform type, push notifications, and email subscriptions. While the majority of inactive subscribers use the web as their main platform, the iOS app has been updated recently with many unique features and services. Another potential barrier is that inactive users have their auto-renewal setting off. This setting choice could be due to a number of reasons not included in the data set such as they don't want to start a new lesson because they don't intend to continue their subscription. While most inactive subscribers are subscribed to receive emails, the email open rate is low which could be a reflection of these users not being interested in the email content they receive. The same holds true for push notifications. While the majority of the

people have push notifications turned on, this group uses the applications the least. Alternatively, that situation could be a reflection of the push notifications' frequency, type, or verbage.

Goal 5: Outline any business relevant opportunities that are present from your analysis of the data no covered above

Based on our outside research about the company during COVID-19, we believe that there is a business opportunity for Rosetta Stone. Our analysis showed that subscribers who experienced a free trial with Rosetta Stone became more valuable subscribers to the company. As such, we believe that Rosetta Stone would benefit from marketing to those individuals who signed up for a three month free trial during this time period because these individuals are likely to become very valuable subscribers to the company.

Furthermore, Rosetta Stone could benefit from collecting more accurate information from their subscribers. There were many NULL values in the subscriber information data set that made it difficult to run regressions and clustering algorithms. The biggest issue was the missing values from the Purchase Amount variable. For every ID number that made a purchase from an app store, there was no purchase data available. This greatly limited the abilities of our regression since about 25% of the data contained null values. Many of our analysis was based off of the Purchase Amount variable, because we were using it to determine which subscribers were most valuable to the company. Having more information about Purchase Amount would have given us better insights into which subscribers were most valuable, and also offer new potential subscriber segments to analyze.

Final Insights

For Goal 1, after many attempts of clustering, we determined that our qualitative research was the most effective for determining the subscriber segments. These segments were the free subscription group, the 3 month (1 language) group, the 12 month (unlimited language) group, the lifetime group, and the lifetime plus group.

For Goal 2, we were able to determine what types of subscribers have the highest value to the company based on our elastic net model. The results showed us that the most influential variables were the following : "isLeadPlatformWebFALSE", "isAutoRenewTRUE", "isEmailSubscriberTRUE", "isUserTypeConsumerTRUE", "isFreeTrialUserTRUE", and "isPushNotificationsTRUE". In normal terms, subscribers that had the following characteristics were the most valuable:

- Used the app as their lead platform

- Had auto renew turned on
- Were email subscribers
- Had a user type label of “consumer” rather than “other” in the data set
- Had done a free trial in the past
- And are subscribed to push notifications

We used our ElasticNet model to also guide us in determining what subscribers could be sold additional products or services for goal three. Our final result is that subscribers that could be sold additional products have the following characteristics:

- User type: Limited users (free subscription & demo users)
- Had auto renew turned off
- Had push notifications turned on
- Is an email subscriber

To answer goal four, we used K-Means clustering to profile inactive users and pinpoint barriers to deeper subscriber engagement. We developed the following conclusions

- Inactive Subscriber Profile
    - Lowest average of total app sessions, purchased the most amount in USD, and had the lowest email open rate.
    - Weren’t free or demo users.
    - Used the web to interact with Rosetta Stone
    - Had their auto-renew off
    - Had push notifications on
    - Is an email subscriber
- Barriers to Engagement
    - Lead Platform Choice
    - Auto-Renewal Setting
    - Email Content
    - Push Notification Setup

For goal five, we were able to identify a great marketing opportunity for Rosetta Stone. With the three-month free trial they offered to all K-12 students, they are valuable because they could convert to regular subscribers. We also noticed that purchases in the app store were not recorded in the dataset. We believe that combining the purchase data from the app store with the current data set would enable us to understand both platform’s performance and develop more robust models.

Division of Labor

Grace:

I was responsible for doing a lot of cleaning of the data sets and organizing them for use in R. I also created the elastic net model, the logistic regression model, and the linear model for Goal 4. I helped with typing the analytical and action plan. Finally, I will help with the presentation of our slides and answering any technical questions that may arise during the presentation.

Nikki:

I was responsible for working with Kayla and Betsy on outside research and looking for external factors influencing the data, as well as research on Rosetta Stone and their services. I also worked on coming up with potential model ideas for addressing each goal. I worked with Kayla and Betsy on creating and filling out the presentation slides, and filling out the word document including the analytical plan and action plan. I will also be presenting some of the slides during our final presentation.

Betsy:

I worked alongside Kayla and Nikki to create the general timeline and brainstorm methods and models that could be used to answer the goals. After the Enet model was made, I wrote the process and results of the model in the analytical plan to give a definitive answer for goal 2. For the presentation, I also helped create the presentation slides concerning goal 2 and goal 5. I will also help present our slides.

Kayla:

I worked closely with Nikki and Betsy to research Rosetta Stone to put our analysis in a business context and develop model ideas that answer the business goals for Ben and Grace to create. In addition to also creating the general outline of the slides, I worked with Grace to gather and develop insights for goals three and four. I also worked heavily in Tableau to create different visualizations to explain our models in the presentation and paper.

Ben:

I did most of the dataset organization in Excel with regards to currency conversion, NULL removal and/or notification, and creating alternate testing datasets for the sake of better identifying trends from the code. I did this part, as well as the coding portion, together with Grace. I did a lot of the experimental coding in order to discover new trends. I helped to determine how to clean the code as well as creating models based around things like clustering or random forests to help discover more advanced trends.