ROSETTA STONE: SUBSCRIBER OPTIMIZATION

INTRODUCTION TO THE DATA

This data is taken from Rosetta Stone subscribers who purchased between 10/2018 and 3/2020. Each subscriber is assigned a personal ID used to identify them in the system.

The two datasets:

- Subscriber Information
  - Gives the basic details as to the transactions and retention for subscribers
  - 24 variables and 40,102 observations
- App Activity
  - Shows how and when certain subscribers are active
  - 4 variables and 809,478 observations

ANALYSIS AND CLEANING OF THE DATA

- The datasets give us good insight into subscription optimization and current engagement
- There were a couple problems within the data:
  - Outrageous outliers in purchase prices due to exchange rates to USD
    - Tried converting prices, but ended up using USD data
  - Lifetime subscriptions skew the models when considering the time subscribed
    - Separate dataset for subscriptions under a year
- Throughout our analysis, we will dive further into these issues in order to best answer our objectives

OBJECTIVE

01

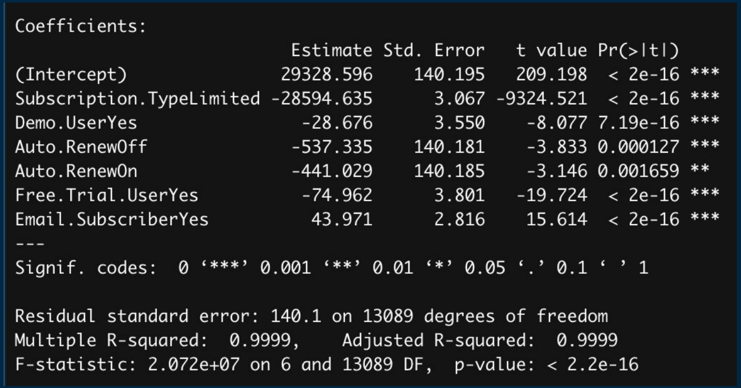
Determine the most valuable subscribers

- Defined “most valuable” as the subscribers that have been consistent for the company, bringing in money along the way
  - Therefore, the most valuable are those with the highest amount of days subscribed and largest amount spent
- We decided that the number of days subscribed (expiration - start date) and total purchase amount allowed us to analyze these individuals further
  - Also looked at Subscription.Type (identify longest days subscribed), Demo.User, Auto.Renew, Free.Trial.User, and Email.Subscriber

THE MOST VALUABLE SUBSCRIBERS

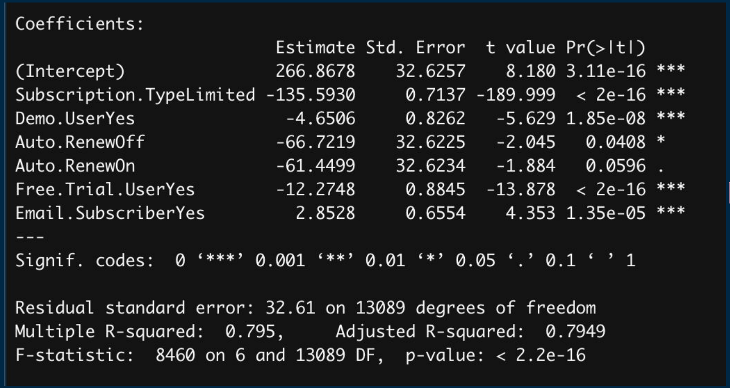
- Created two linear regressions based on this analysis to predict the outcome of the continuous variables: Days.Subscribed and Purchase.Amount (seen below)
- Based on the linear regression models, we can see the relationship between each variables to our continuous variables.
  - In both cases, a limited subscription type is not beneficial, auto renew doesn’t have a large impact, and being an email subscriber helps to being a valuable subscriber

Days.Subscribed



```
Coefficients:
                          Estimate Std. Error   t value Pr(>|t|)    
(Intercept)              29328.596    140.195   209.198  < 2e-16 ***
Subscription.TypeLimited -28594.635     3.067 -9324.521  < 2e-16 ***
Demo.UserYes               -28.676      3.550    -8.077 7.19e-16 ***
Auto.RenewOff             -537.335    140.181    -3.833 0.000127 ***
Auto.RenewOn              -441.029    140.185    -3.146 0.001659 ** 
Free.Trial.UserYes         -74.962      3.801   -19.724  < 2e-16 ***
Email.SubscriberYes         43.971      2.816    15.614  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 140.1 on 13089 degrees of freedom
Multiple R-squared:  0.9999,	Adjusted R-squared:  0.9999 
F-statistic: 2.072e+07 on 6 and 13089 DF,  p-value: < 2.2e-16
```

Purchase.Amount



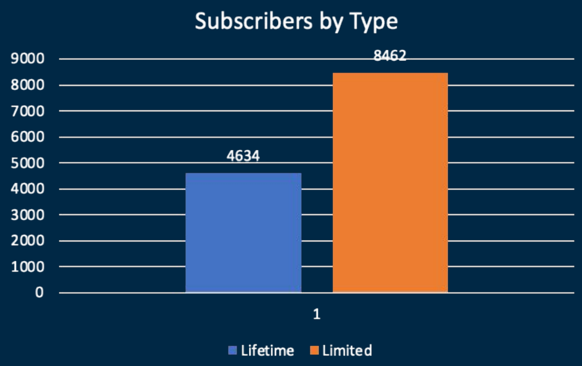
```
Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)    
(Intercept)               266.8678    32.6257    8.180 3.11e-16 ***
Subscription.TypeLimited -135.5930     0.7137 -189.999  < 2e-16 ***
Demo.UserYes               -4.6506     0.8262   -5.629 1.85e-08 ***
Auto.RenewOff             -66.7219    32.6225   -2.045   0.0408 *  
Auto.RenewOn              -61.4499    32.6234   -1.884   0.0596 .  
Free.Trial.UserYes        -12.2748     0.8845  -13.878  < 2e-16 ***
Email.SubscriberYes         2.8528     0.6554    4.353 1.35e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.61 on 13089 degrees of freedom
Multiple R-squared:  0.795,	Adjusted R-squared:  0.7949 
F-statistic:  8460 on 6 and 13089 DF,  p-value: < 2.2e-16
```

OBJECTIVE 02

Understand the subscriber segments present in the database

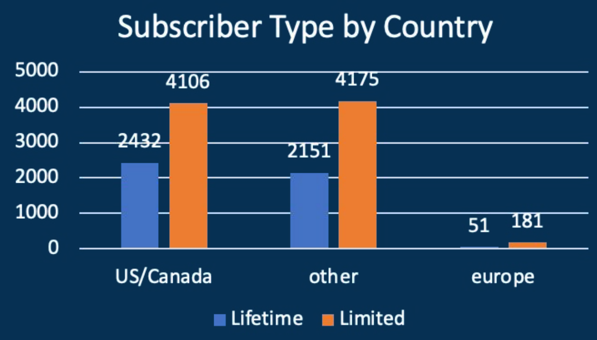
- Based on the datasets provided, we focused on two segmentations:
  - Geographic Segmentation
  - Behavioral Segmentation
- Further segmentation on subscriber type (Lifetime/Limited)
  - Geographic
  - Purchase association
- Evaluated associated purchase and associated email engagement

SUBSCRIBER SEGMENTS



13,906 Total subscribers

- 35% Lifetime subscribers
- 65% Limited subscribers



Subscriber type by country

48% of total = US/Canada (6326)

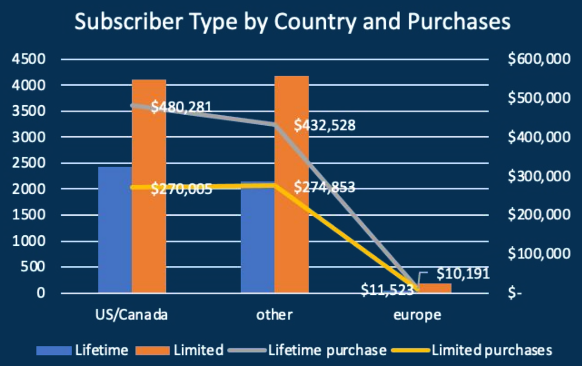
- 37% lifetime
- 63% limited

50% of total = Other (6538)

- 34% lifetime
- 66% limited

2% of total = Europe (232)

- 22% lifetime
- 78% limited



Subscriber type by purchase

US/Canada ($750,286)

- 64% lifetime ($197 AOV)
- 36% limited ($66 AOV)

Other ($707,381)

- 61% lifetime ($201 AOV)
- 39% limited ($66 AOV)

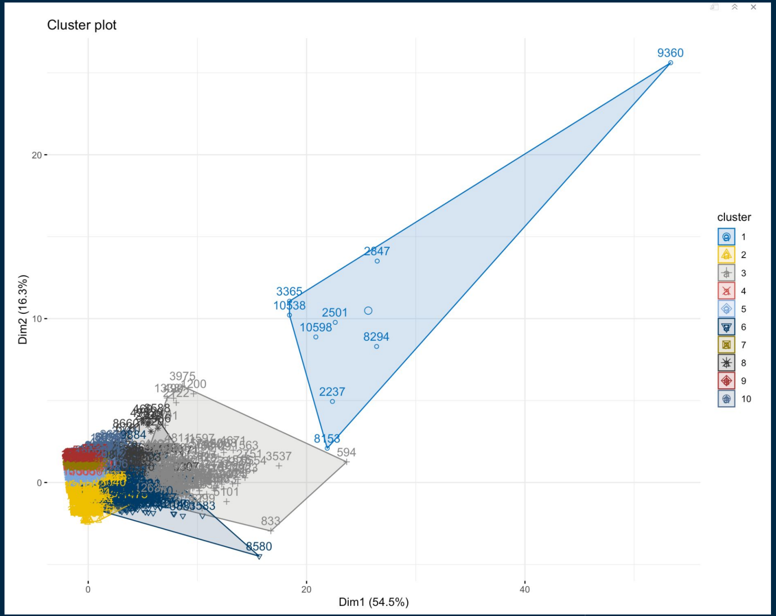
Europe ($21,713)

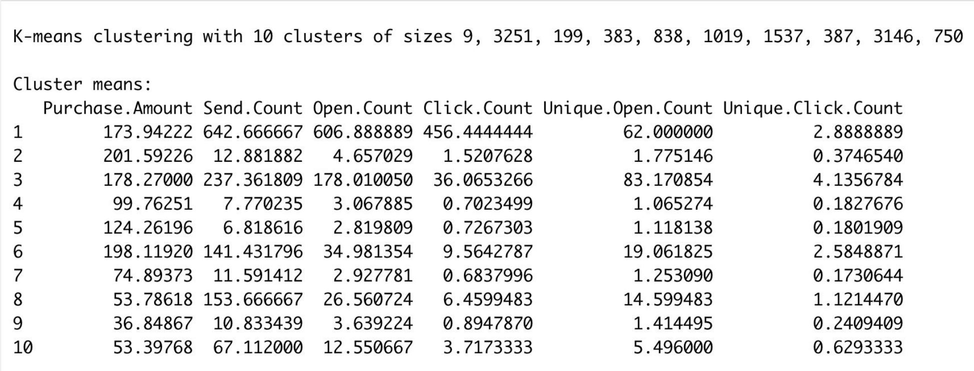
- 54% lifetime ($278 AOV)
- 46% limited ($56 AOV)

CLUSTER MODEL ASSOCIATED OVERALL PURCHASES

- The highest cluster (2) displays a $201 AOV has a 66% email Open rate over emails sent. With a 33% Click to Open rate.
- The second highest cluster (9) has a $37 AOV with a 33% email Open rate over emails sent. 24% Click to Open rate.
- The third highest cluster (7) shows a $74 AOV with a 25% email open rate over emails sent. With a 23% Click to Open rate.

Little trouble/failed modeling with different variables due to lots of categorical variables, as well as other complications; see supporting document





```
K-means clustering with 10 clusters of sizes 9, 3251, 199, 383, 838, 1019, 1537, 387, 3146, 750

Cluster means:
   Purchase.Amount Send.Count Open.Count Click.Count Unique.Open.Count Unique.Click.Count
1        173.94222 642.666667 606.888889 456.4444444         62.000000          2.8888889
2        201.59226  12.881882   4.657029   1.5207628          1.775146          0.3746540
3        178.27000 237.361809 178.010050  36.0653266         83.170854          4.1356784
4         99.76251   7.770235   3.067885   0.7023499          1.065274          0.1827676
5        124.26196   6.818616   2.819809   0.7267303          1.118138          0.1801909
6        198.11920 141.431796  34.981354   9.5642787         19.061825          2.5848871
7         74.89373  11.591412   2.927781   0.6837996          1.253090          0.1730644
8         53.78618 153.666667  26.560724   6.4599483         14.599483          1.1214470
9         36.84867  10.833439   3.639224   0.8947870          1.414495          0.2409409
10        53.39768  67.112000  12.550667   3.7173333          5.496000          0.6293333
```

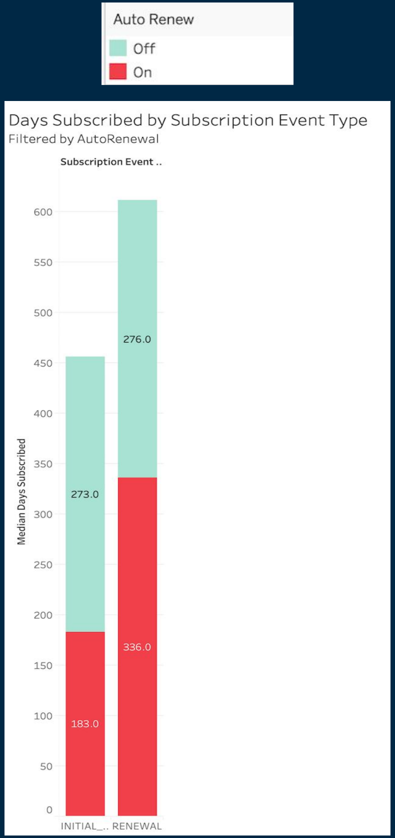
OBJECTIVE 03

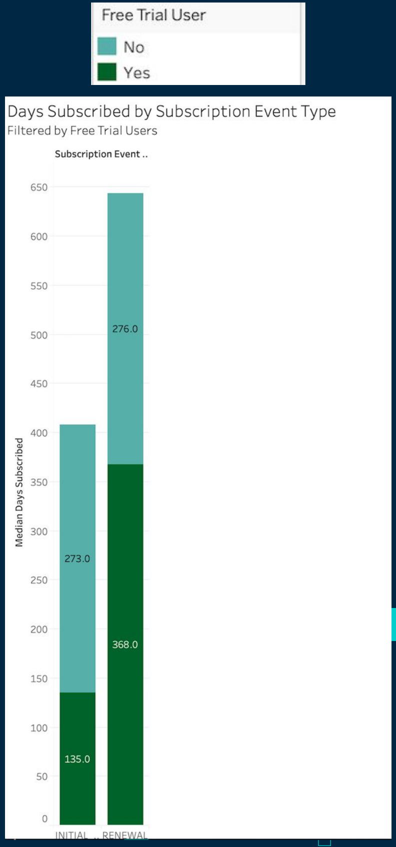
Identify the most likely subscribers who could be sold additional products or services

- Compared days subscribed with users that were on a free trial, auto renewal, or not
- Looked at app session platform to determine which is most active
- Segmented users by country and language learned to see which groups have the highest days subscribed

EXTRA PRODUCTS OR SERVICES

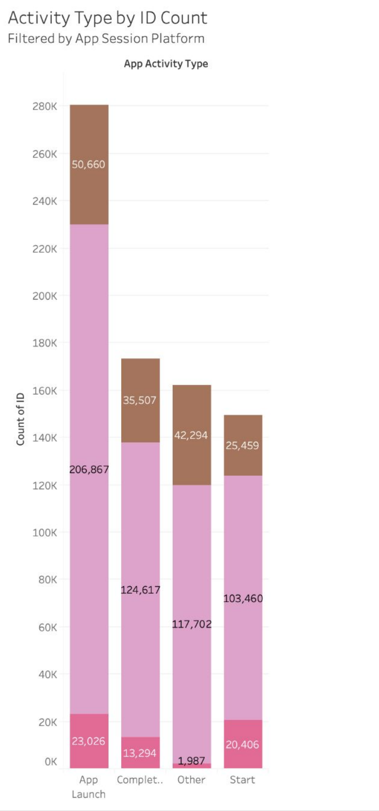
- Subscribers who are renewals instead of initial subscribers have greater days subscribed
- Free trial users are more likely to be renewals instead of initial
- Subscribers with auto renew on and are renewal customers have higher days subscribed than auto renew users who are initial purchasers
- We can conclude that customers who are renewal users are more likely to be sold additional products than initial customers

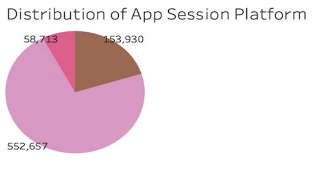




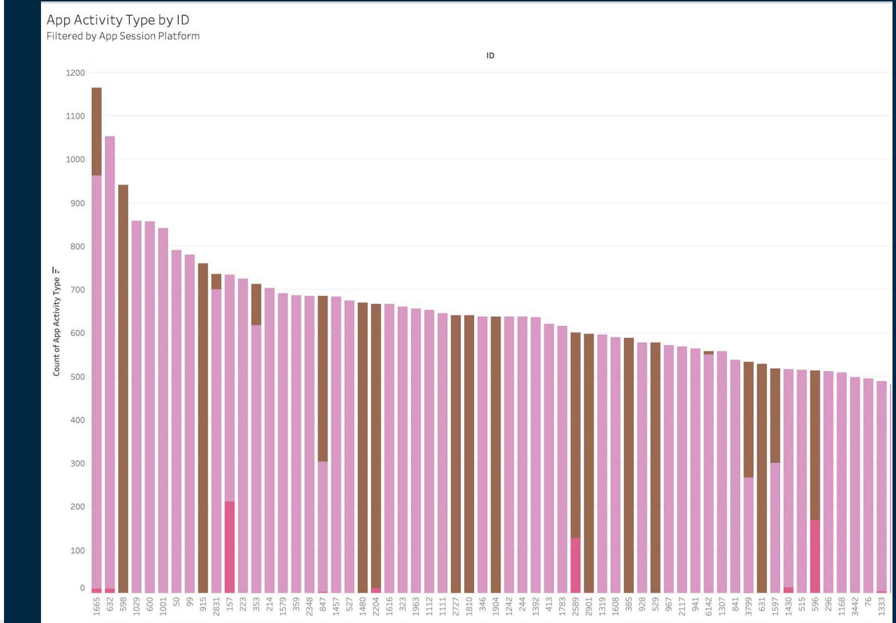
EXTRA PRODUCTS OR SERVICES

- Most activity comes from ios users and the least is from web users
- Ios users complete lessons at a much higher count
- We might say that ios users can be sold additional products/services as these users are the most active subscribers


Activity Type by ID Count
Filtered by App Session Platform

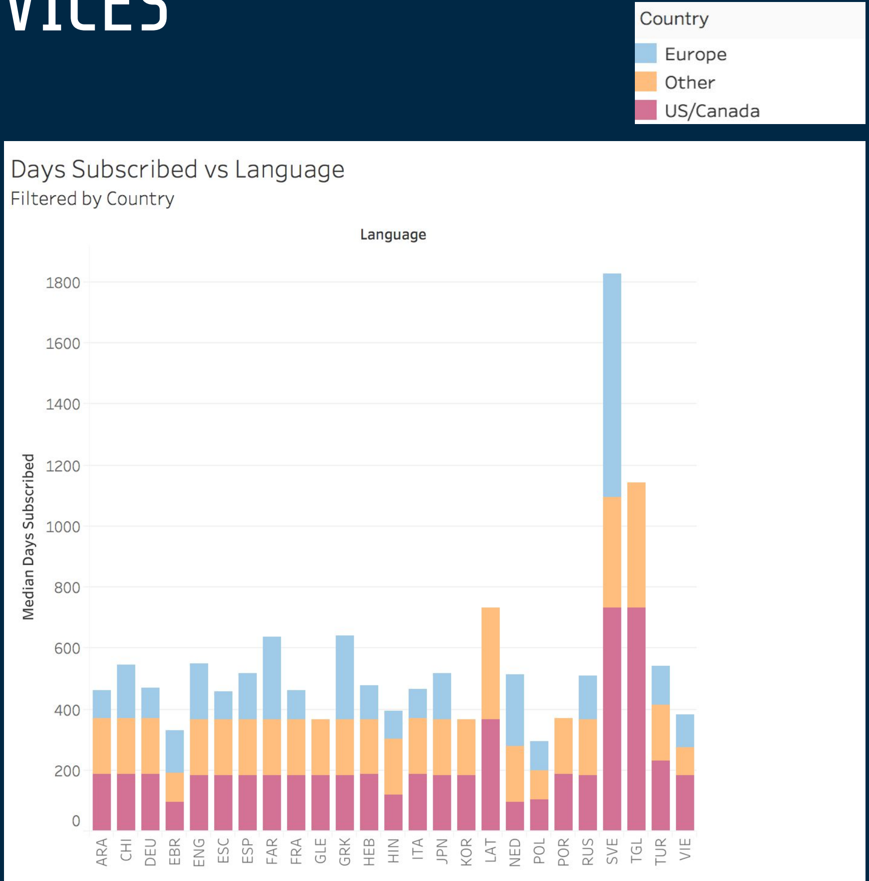

Distribution of App Session Platform

App Session Platform
🟫 android
▨ ios
🟥 web


App Activity Type by ID
Filtered by App Session Platform

EXTRA PRODUCTS OR SERVICES

- US/Canada have the highest number of days subscribed
- SVE (Swedish) and TGL (Tagalog) are the top languages people choose to learn
- Customers located in US/Canada and those learning Swedish or Tagalog can be sold additional products/services as these groups are the most active
- More active → more likely to buy more products/services to enhance their user experience



Days Subscribed vs Language
Filtered by Country

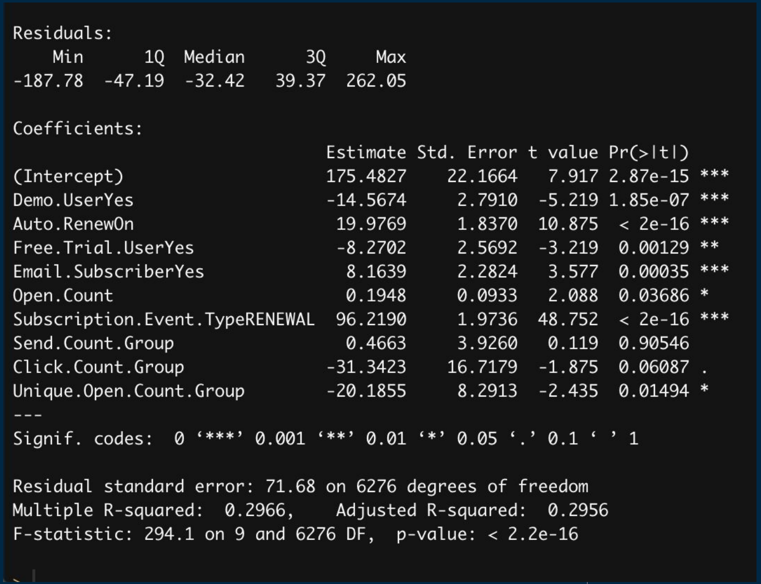
OBJECTIVE

04

Identify the subscriber profile of those not continuing and identify the barriers to deeper subscriber engagement

- We identified those not continuing their subscription by looking at Subscription.Expiration and Days.Subscribed
- Found those with Days.Subscribed under 1 year to be important in identifying why subscribers are not continuing (6286 in total)
  - Also found Demo.User, Auto.Renew, Free.Trial.User, Email.Subscriber, Subscription.Event.Type, and all the counts variables to be important
- Since we wanted to include the counts, we grouped the variables due to the high ranges (except Open.Count)

SUBSCRIBERS NOT CONTINUING



```
Residuals:
    Min      1Q  Median      3Q     Max 
-187.78  -47.19  -32.42   39.37  262.05 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   175.4827    22.1664   7.917 2.87e-15 ***
Demo.UserYes                  -14.5674     2.7910  -5.219 1.85e-07 ***
Auto.RenewOn                   19.9769     1.8370  10.875  < 2e-16 ***
Free.Trial.UserYes             -8.2702     2.5692  -3.219  0.00129 ** 
Email.SubscriberYes             8.1639     2.2824   3.577  0.00035 ***
Open.Count                      0.1948     0.0933   2.088  0.03686 *  
Subscription.Event.TypeRENEWAL  96.2190     1.9736  48.752  < 2e-16 ***
Send.Count.Group                0.4663     3.9260   0.119  0.90546    
Click.Count.Group             -31.3423    16.7179  -1.875  0.06087 .  
Unique.Open.Count.Group       -20.1855     8.2913  -2.435  0.01494 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 71.68 on 6276 degrees of freedom
Multiple R-squared:  0.2966,	Adjusted R-squared:  0.2956 
F-statistic: 294.1 on 9 and 6276 DF,  p-value: < 2.2e-16
```

- We went with a linear regression model to predict the outcome of the continuous variable (Days.Subscribed), trying to see why people are not continuing
- The subscriber profile of those not continuing mainly consists of low days subscribed (<365 days) and a limited subscription type, as identified in our analysis
- Furthermore, the barriers to deeper subscriber engagement stem from the emails (more specifically Click.Count.Group and Unique.Open.Count.Group) and whether auto renew is set up
  - Using this model, we can target these specific groups to increase subscriber retention

OBJECTIVE 05

Outline any other business relevant opportunities present in the data analysis

- Analyzed click count, email data, and activity level to determine any potential business opportunities
- Created a clustering model in order to segment user activity and draw conclusions from it

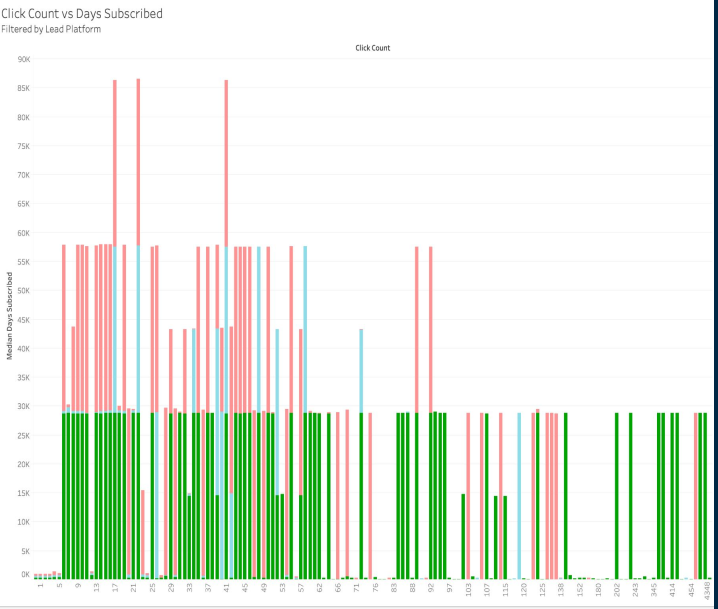
CLICK COUNT ANALYSIS

Lead Platform

🟥 App

🟦 Unknown

🟩 Web


Click Count vs Days Subscribed

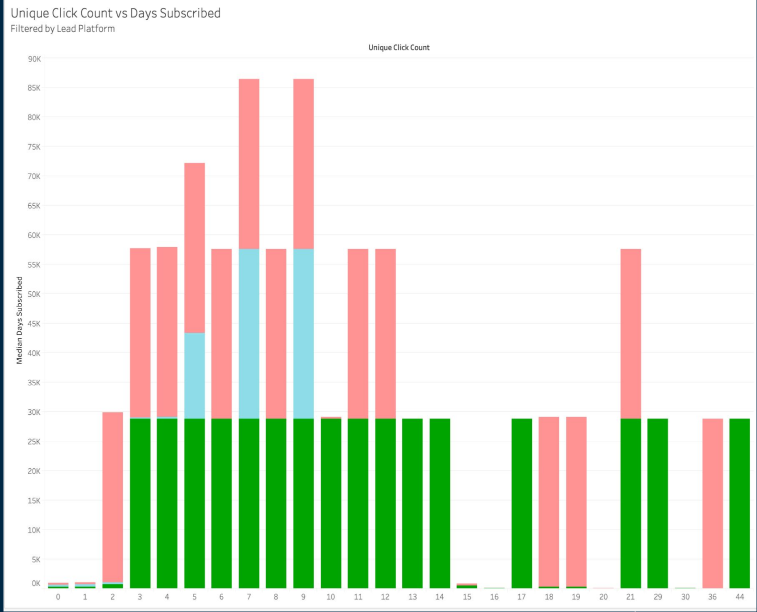
Filtered by Lead Platform

One might assume that higher click count correlates with more days subscribed

However, the data shows us that lower/middle end click counts have higher days subscribed

Web users have the highest click count (unique and non-unique)

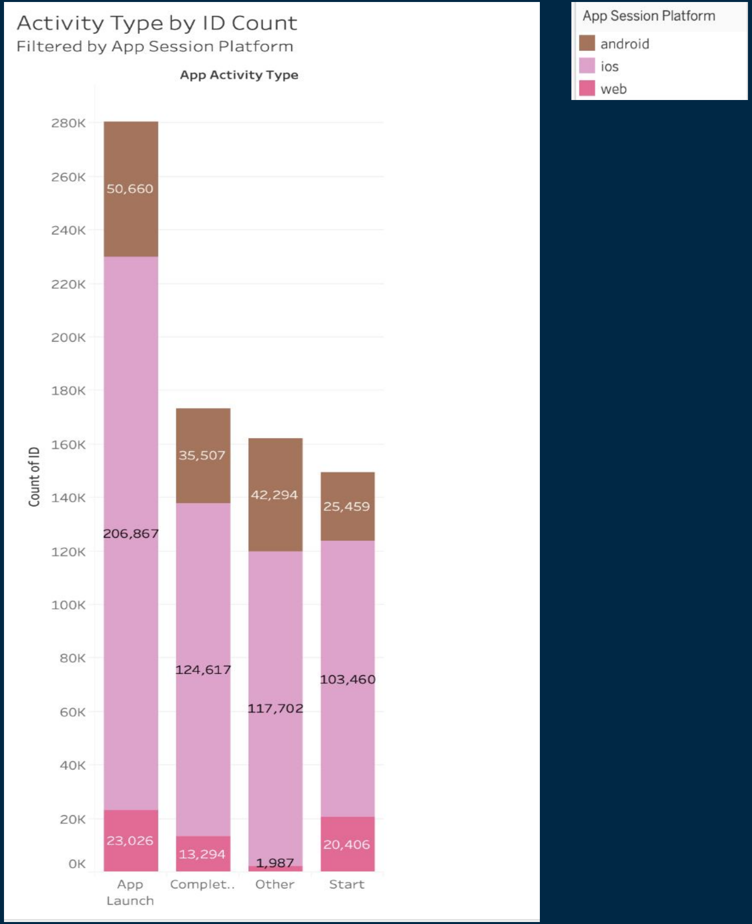
Web users may be more engaged in using Rosetta Stone than app users


Unique Click Count vs Days Subscribed

Filtered by Lead Platform

ACTIVITY ANALYSIS

- Many more users are launching the app compared to actually starting or completing the program
- App launch doesn't mean anything unless the user uses the programs offered (that way, they can give feedback or refer the program to a friend; app launched but not started wastes the consumer’s time)
- Possible business suggestion: having the app automatically start the lesson when the user opens the app
- Eliminates issue of launches with no starting


Activity Type by ID Count

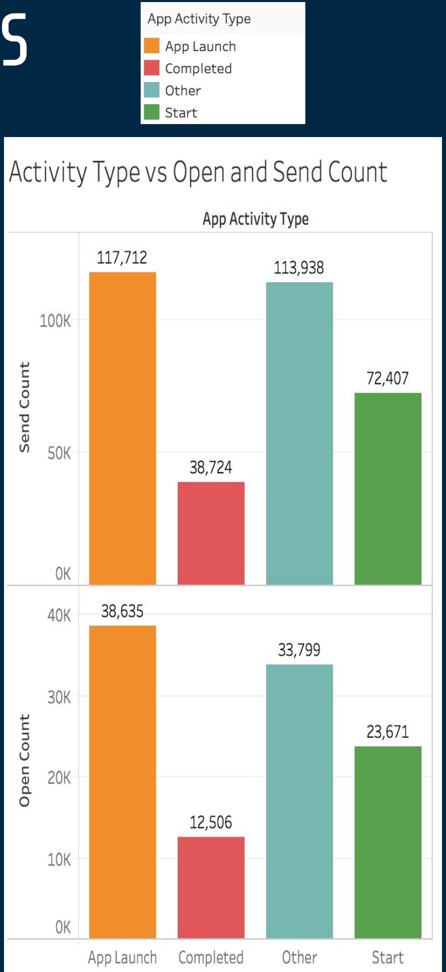
Filtered by App Session Platform

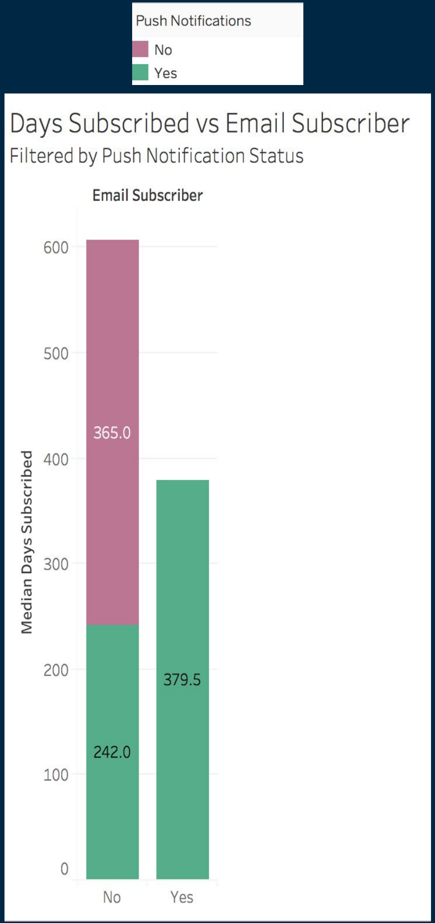
EMAIL ANALYSIS

- Rosetta Stone sends out more emails than their subscribers open
- Those that have the app launched receive and open the most emails
- The send to open ratio among app launch, completed and start is about 3.1
  - Users are not engaging with emails

Possible solutions:

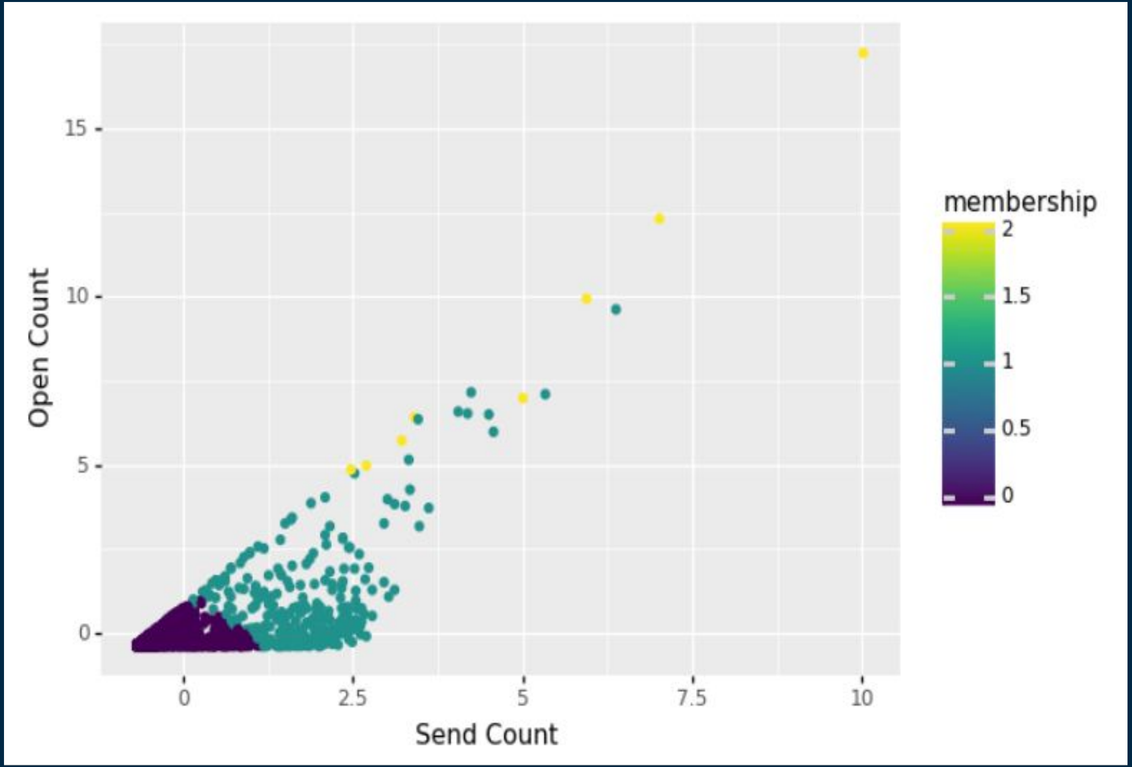
- Rosetta Stone should create more engaging emails
- More visually engaging
- Simpler/more concise email interface





- According to this figure, customers with email subscriptions and push notifications on have the highest days subscribed
- Surprisingly high number of people with emails and notifications off that also have high amount of days subscribed
- All customers with emails on also have push notifications on
- Focus on this segment as they are highly valuable/engaged

CLUSTER ANALYSIS



- The graph compares two variables, Open Count and Send Count, using clustering
- It shows us pretty distinct clusters when we compare the z-scored values of the send count to open count
- The purple cluster refers to those who have completed app activity (low send count to low open count)
- The turquoise cluster refers to those that have started app activity (mid send count to mid open count)
- The yellow cluster refers to those that have the app launched (high send count to high open count)

ANALYTICAL PLAN

Create a timeline

Rosetta Stone needs to set their preferences and create a timeline to help them make changes at a reasonable pace.

Use the models

Using the models that our team created will help Rosetta Stone make beneficial increases in subscriber engagement.

Track the models

As the company changes and makes developments, they need to be continuously searching for new models to incorporate.

Reach their goals

With a heavy focus on these models and for optimizing subscribers, Rosetta Stone will meet its business growth goals.

A more detailed analytical plan is featured in the appendix and continued in the supporting document

THANK YOU!

APPENDIX

ACTION PLAN

01

ANALYZE THE DATA

Analyzing the data allows us to piece together any problems that may arise for the company.

02

IDENTIFY OUR OBJECTIVES

In order to help Rosetta Stone meet their growth goals, we need to identify (5) major objectives.

03

CREATE AN ANALYTICAL PLAN

Based on our analysis and answers to our objectives, we can set an analytical plan in motion.

STEPS TAKEN

1. Downloaded the datasets
2. Individually analyzed the data sets
3. Discussed our personal thoughts and went over the data together
4. Determined how to go about the objectives and overall goal
5. Cleaned the data
6. Created charts and graphs in Excel and Tableau
7. Creating models in R and Python
8. Tied the objectives together to help determine how to have subscriber optimization
9. Designed and communicated an analytical plan
10. Created a report and presentation to present the advised plan and goals

A more detailed list of steps taken is in the supporting document

ANALYSIS OF THE OBJECTIVES

By analyzing these objectives, we are able to see the major strengths and weaknesses of the company in regards to subscribers.

If the company is able to solve these problems, then they are more likely to achieve their business growth goals.

The linear regression and clustering models created will help the company to emphasize these strengths and diminish the weaknesses.

With the business growth goals met, Rosetta Stone will have subscriber optimization and have more active customers.

This analysis on the objectives leads into our analytical plan for Rosetta Stone.

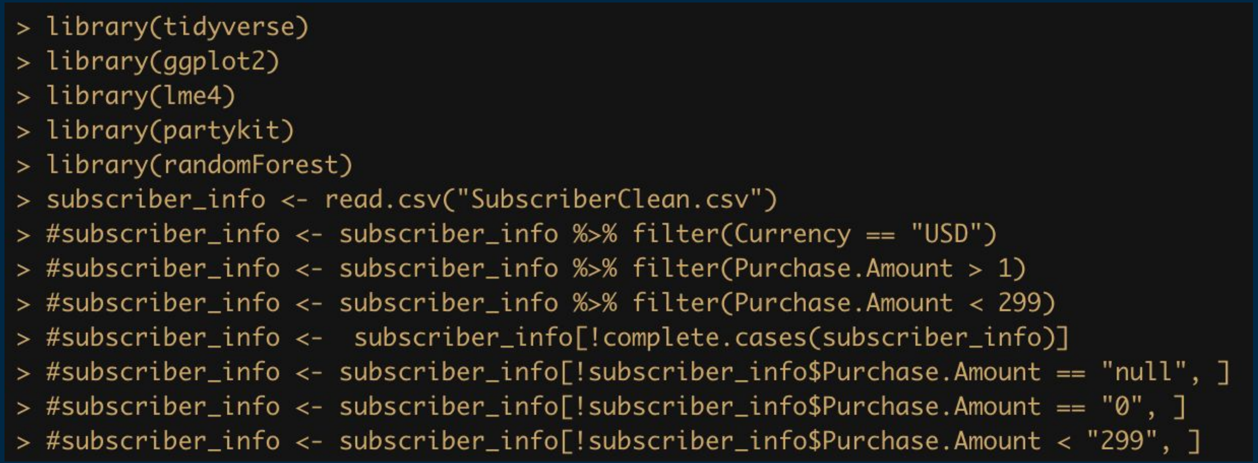
DETAILED ANALYTICAL PLAN

To further expand on the information listed in the presentation and to summarize the information listed in the supporting document, we will discuss the analytical plan in more depth:

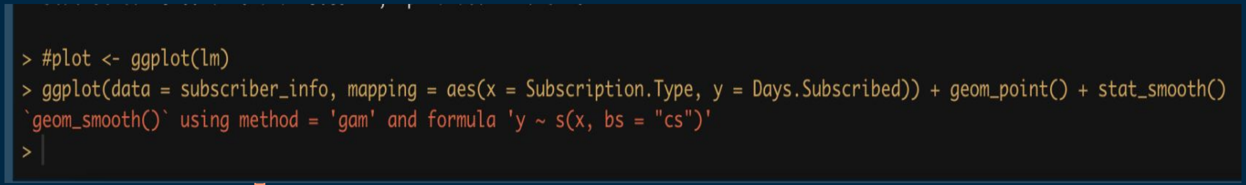
- Rosetta Stone needs to take the analysis of the objectives into account, explaining how each objective tied together
- We will go into the recommendations based on each objective in the supporting document
    - It is important to specify how we came to the conclusions we did
- Inform Rosetta Stone on the best way to go about putting these recommendations into action
    - While it might seem simple, the order of steps taken is important

FAILED CODE/MODELING

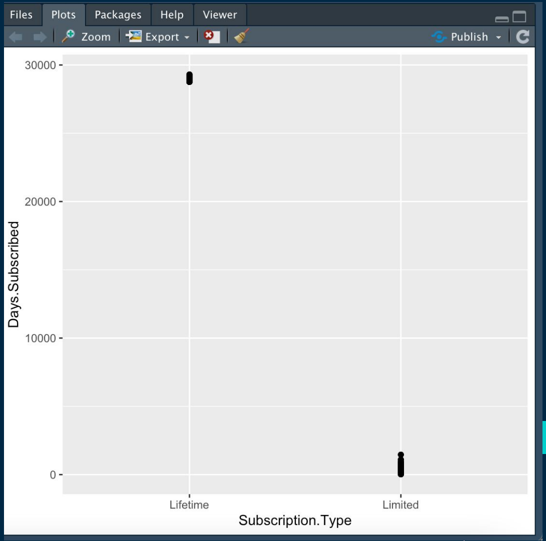
This failed cod/modeling is explained further in the additional document



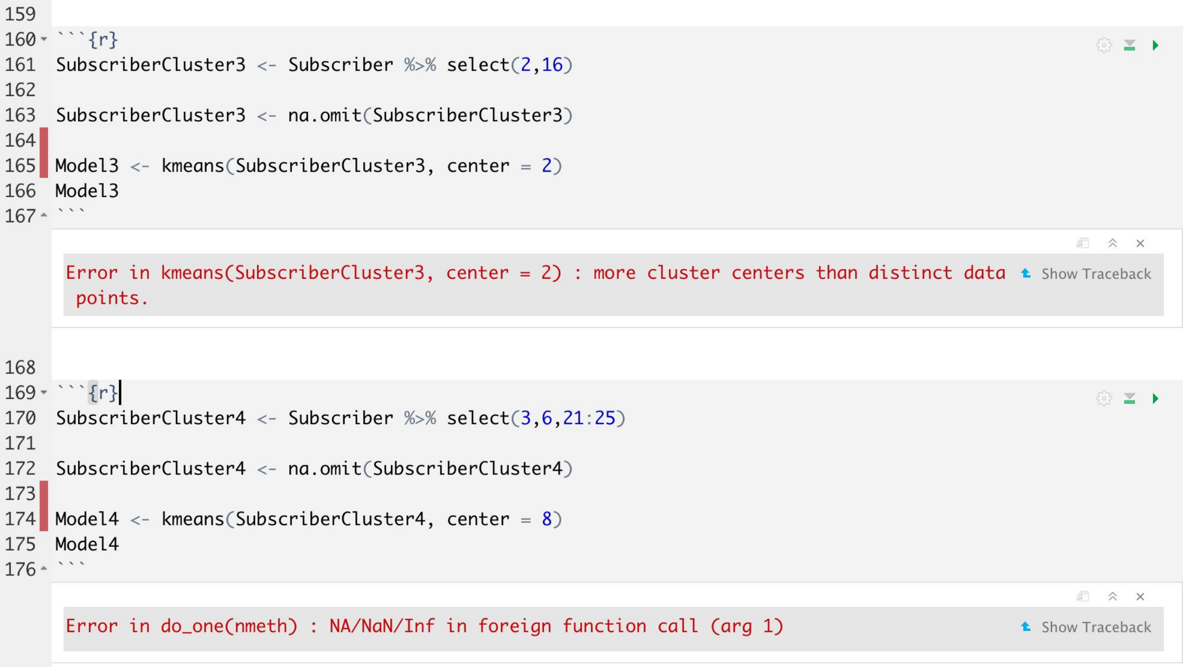
```
> library(tidyverse)
> library(ggplot2)
> library(lme4)
> library(partykit)
> library(randomForest)
> subscriber_info <- read.csv("SubscriberClean.csv")
> #subscriber_info <- subscriber_info %>% filter(Currency == "USD")
> #subscriber_info <- subscriber_info %>% filter(Purchase.Amount > 1)
> #subscriber_info <- subscriber_info %>% filter(Purchase.Amount < 299)
> #subscriber_info <-  subscriber_info[!complete.cases(subscriber_info)]
> #subscriber_info <- subscriber_info[!subscriber_info$Purchase.Amount == "null", ]
> #subscriber_info <- subscriber_info[!subscriber_info$Purchase.Amount == "0", ]
> #subscriber_info <- subscriber_info[!subscriber_info$Purchase.Amount < "299", ]
```



```
> #plot <- ggplot(lm)
> ggplot(data = subscriber_info, mapping = aes(x = Subscription.Type, y = Days.Subscribed)) + geom_point() + stat_smooth()
`geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
> 
```



FAILED CODE/MODELING



~~~
```{r}
SubscriberCluster3 <- Subscriber %>% select(2,16)

SubscriberCluster3 <- na.omit(SubscriberCluster3)

Model3 <- kmeans(SubscriberCluster3, center = 2)
Model3
```
~~~

```
Error in kmeans(SubscriberCluster3, center = 2) : more cluster centers than distinct data
 points.
```

~~~
```{r}
SubscriberCluster4 <- Subscriber %>% select(3,6,21:25)

SubscriberCluster4 <- na.omit(SubscriberCluster4)

Model4 <- kmeans(SubscriberCluster4, center = 8)
Model4
```
~~~

```
Error in do_one(nmeth) : NA/NaN/Inf in foreign function call (arg 1)
```

Failed to run cluster model.

CONCLUSIONS

The data

- Difficult to transform/clean data when you cannot find the readily accessible information to do so
  - Ex. currency exchange rates at a specific point in time for ~40,000 data points
- Does not produce the best models due to the large amount of data involved and outliers present in most of the numeric data
- Pretty clean dataset, so made the general cleaning easier

The project

- Difficult to code as a group online; no quick way to edit code together
- There are so many ways to go about each part of the project, it is hard to decide where to start and have 5 people agree