Subscriber Optimization for Rosetta Stone



**Ishan Supanekar**

**Irene Huang**

**Timothy Tan**

**Brian Kim**

**Jackie Zhao**

**JJ Goh**

**Table of Contents:**

i. Basic Raw Dataset Statistics 2

ii. Sample Observations from Dataset 3

iii. Data Transformation / Cleaning 3

1. Determining The Most Valuable Customer 4-7

2. Understanding Subscriber Segments 8-15

3. Customers Who Can be Sold Additional Services 16-19

4. Customers Who Aren’t Continuing with Usage 20-23

5. Extra Business Analysis / Opportunities 24

6. Challenges and Limitations 25

**i. Basic Raw Dataset Statistics**

Found through Excel Pivot Tables:

* Top Purchased Languages:
  + “ESP” - 24.55%
  + “ALL” - 15.70%
  + “FRA” - 11.25%
* Subscription Types:
  + Lifetime - 16.32%
  + Limited - 83.68%
* Subscription Event Type:
  + Initial Purchase: 73.79%
  + Renewal: 26.21%
* Duration (Subscription End Date - Start Date)
  + Average: 4918 Days
  + Max: 29312 Days
  + Min: 30 days
  + Std. Dev: 10561 days
  + Median : 282 days
* Demo User:
  + No - 65.39%
  + Yes - 34.61%
* Free Trial user:
  + No - 85.45%
  + Yes - 14.55%
* Auto Renew:
  + Off - 64.32%
  + On - 35.68%
* Country:
  + Europe: 12.49%
  + Other: 37.48%
  + US/Canada: 50.02%
* User Type:
  + Consumer: 67.56%
  + Other: 32.44%
* Type of Lead Platform:
  + App: 37.58%
  + Unknown: 31.37%
  + Web: 31.05%
* Email Subscriber:
  + No: 51.48%
  + Yes: 48.51%
* Push Notifications:
  + No: 31.38%
  + Yes: 68.62%
* Send Info count from email:
  + Average: 32.39
  + Max: 4370
  + Min: 1
  + Median: 10
* Open Info count email:
  + Average: 8.23
  + Max: 4365
  + Min: 0
  + Median: 0
* Click Count email:
  + Average: 2.15
  + Max: 4348
  + Min: 0
  + Median: 0
* Unique Click Count:
  + Average: 3.89
  + Max: 196
  + Min: 0
  + Median: 1
* App Session user:
  + 0: 9.96%
  + Android: 9.63%
  + Ios: 38.62%
  + NULL: 35.99%
  + Web: 5.80
* App Activity:
  + App Launch: 23.02%
  + Completed: 9.68%
  + NULL: 35.99%
  + Onboarding: 0%
  + Other: 21.50%
  + Start: 9.80%

**ii. Sample Observations from Dataset**

1. Data are in different currencies
2. There are a lot of values in “Purchase Amounts” that are extremely high
3. There are a lot of NULL values in “Purchase Amounts”column
4. All NULL values in “Purchase Amounts” are from App platform
5. App platform is not released until January 2019
6. Only 30% of limited consumers renewed their subscriptions
7. Almost all App purchase data is inaccurate
8. Purchase amounts are not in sync with the duration
9. You would expect a high duration with a high purchase amount

**iii. Data Transformation / Cleaning**

Data Cleaning and Feature Transformation

1. Converted all currencies to USD using exchange rate on November 16th 2020
2. Removed rows if “Purchase Amounts” is above USD $996 since we found out $999 was the maximum amount offered in Korea
3. Created variable “Duration” based on “Subscription Start Date” and “Subscription Expiration”
4. Merged App Activity data to main Subscriber Information dataset
5. Changed “App Activity Type” to individual columns from row cells to see counts of: “App Launch”,”Start”,”Completed”,”Start”,”Other”
6. Grouped Languages other than “ALL”, “ESP”, “ITA”,”FRA” as Other
7. Deleted free trial dates
8. Removed outliers for “Send Count”, “Click Count”, and “Open Count” above 750
9. Created dummy variables from all categoricals
10. Created “Open Rate” and “Click Rate”, percentage of emails that were opened/clicked from the sent amount from Rosetta Stone. (Value 0-1)
11. Created “Champion Binary”, binary variable whether a customer is high value according to analysis for question #1.

**1. Determining the Most Valuable Subscribers**

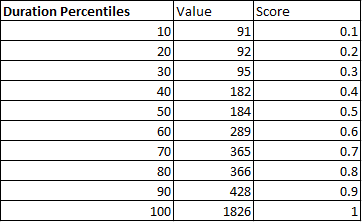
**Goal**

* Identify the top percentile customers and examine the characteristics that most likely predict the shared characteristics.

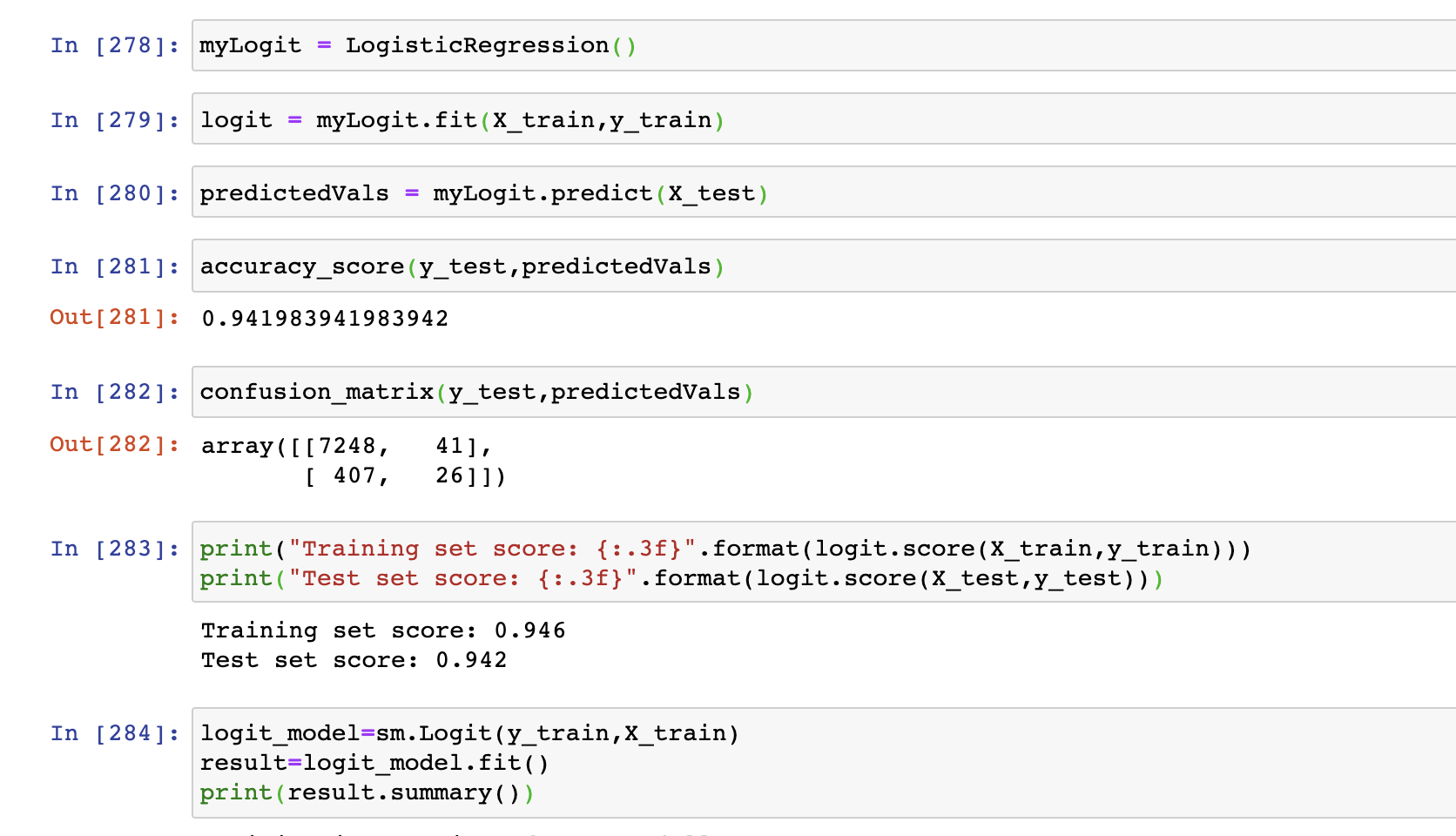
**Metrics**

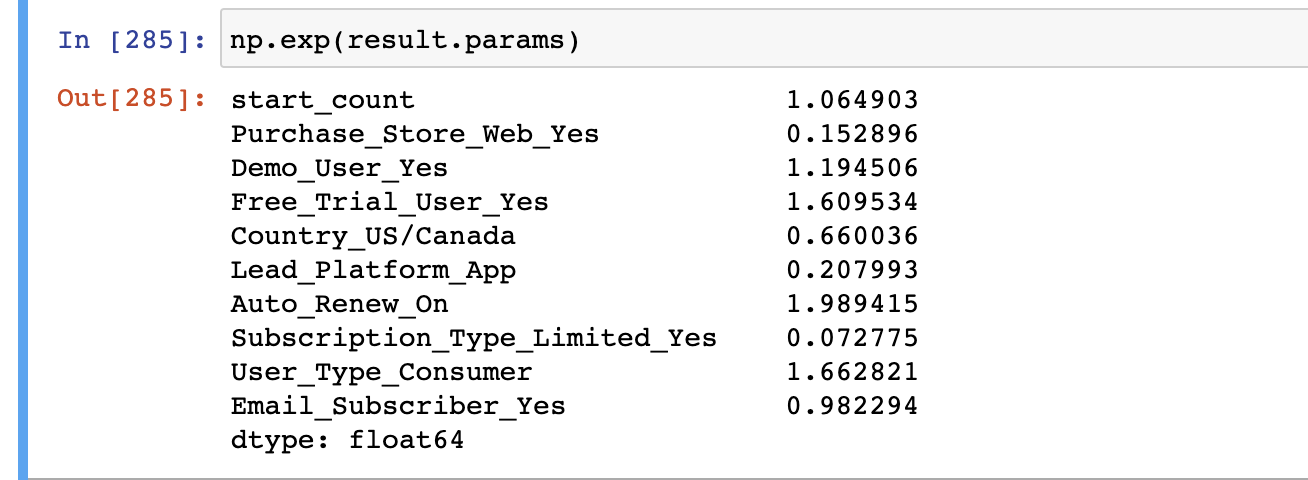
* Duration - A created variable that takes the difference of subscription start date and end date - displays customer history/loyalty with RosettaStone
* Push Notifications - Whether a user keeps notifications on or not
* Open Rate - The rate at which a user opens their emails from Rosetta Stone
* App Launch - Number of times a user opened the Rosetta Application
* Completion - Number of times a user completed a Rosetta stone lesson

**Methodology**

* Divide up our metrics by percentile (0th-100th) and award points to customers based on their percentiles to see the highest performers in the following categories:
  + The Scoring Criteria for Limited and Lifetime:
    - Limited
      * Duration: 0 to 1
      * Push Notifications: 0 or 1
      * Open Rate: 0 to 1 **Total score = 5**
      * App Launch Score: 0 to 1
      * Completion Score: 0 to 1
    - Lifetime
      * Push Notifications: 0 or 1.25
      * Open Rate: 0 to 1.25 **Total score = 5**
      * App Launch Score: 0 to 1.25
      * Completion Score: 0 to 1.25
    - Excel Example: Duration scoring for Limited Customers 
  + Why we had separate scoring for Limited and Lifetime and Limited Users:
    - In terms of Limited users, the start and end date **matters** as this acts as a window of time Rosetta Stone has to offer further marketing and promotions to get limited users to renew service and become Lifetime possibly
    - In terms of Lifetime users, the end date is trivial, as these users will be with the service for an indefinite period. There is no feasible window of time that Rosetta has to act on that
    - This is why Duration was not used as a scoring metric for Lifetime customers and scores for each metric were weighed higher (1.25 instead of 1).
* Assigned scores based on what value for the respective category falls into the scoring rating
* Compile the scores as a Total Score, ranked from 1-5
* In order to see the top performers, we label them as “Champions” who are ranked from 4-5

**Business Insights from our most valuable customers**

* After we graded each customer, took the Champion binary and converted to binary as “1 = Champion Customer” and “0 = Not a Champion Customer”
* We then inputted data into Python for Logistic Regression, predicting the binary value of the Champion Customer
* The reasoning behind this model is to see which variables are most predictive of a champion customer, and if we find customers with these traits that are not a champion yet, we can target them with promotional activities to convert them to a high-value customer
* Here is the code we performed for Logit Regression:
  + 
    - Training score of model: .946
    - Test score: .942
* Here is the output, and the the level of importance our predictors have in terms of determining the outcome of a champion customer



**\***Take note that variables used to determine a champion customer or not were left out in the logistic regression since they might have high interdependencies with the outcome

\*\*Categorical variables were transformed to dummy variables for simple interpretation with the odds ratio

**Interpretation of Logistic Regression**

* The reasoning behind this model is to see which variables are most predictive of a champion customer, and if we find customers with these traits that are not a champion yet, we can target them with promotional activities to convert them to a high-value customer
* According to the output we found that the following most contributed to the outcome of whether a customer is a champion or not in odds ratio terms:

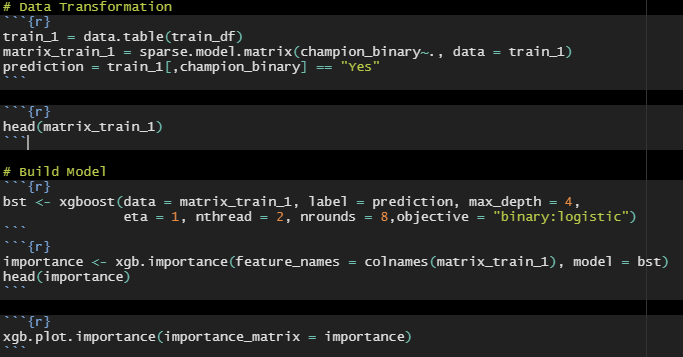
1. Auto Renew On: If a customer sets his/her auto renew function on, it increases the likelihood of them being a champion customer by 98%
2. User Type Consumer: If a customer is a labeled as consumer, it increases the likelihood of them being a champion customer by 66%
3. Free Trial User: If a customer is a free trial user, it increases the likelihood of them becoming a champion customer by 60%

**Business Actions.**

* Now that our team has determined characteristics that our most indicative of a champion customer, we can utilize these traits for special marketing campaigns/promotions
  + For example:
    - Rosetta Stone can send tailored emails for special promotions for added services such as a finite amount of free Live Tutoring sessions, or a discount for a next installment of their renewal.
  + The rationale is for Rosetta Stone to make sure they keep their high valued customer engaged and happy with the overall service. Rosetta Stone should make it a priority not to lose these customers.
* Another segment to target are customers who have these traits and are not champion customers
  + For example:
    - Rosetta Stone can offer these customers promotions to continue their subscription service.
    - The goal is to convert these customers from “non-champions” to “champions”

**Attempt at Gradient Boosting**

* We tried to use gradient boosting to verify our logistic model as it often provides a very high predictive accuracy
* We also thought that the data integrity was very low since we played around with some variables and it gave us a strangely high odds ratio
* However, we were unsuccessful in translating our findings to business insights and it caused more confusion compared to our initial findings
* We suspected that it might be a coding/formatting error
* Here is our attempt at Gradient Boosting:



**2. Understanding the subscriber segments present in the database**

**Goal**

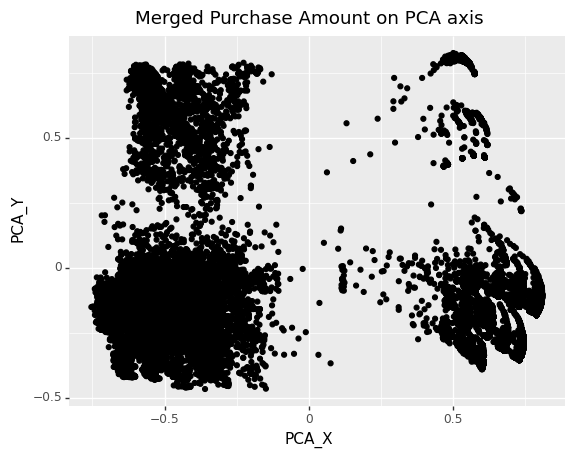
* To separate the whole dataset of customers into clear segments to see which ones are targetable based on our model’s findings

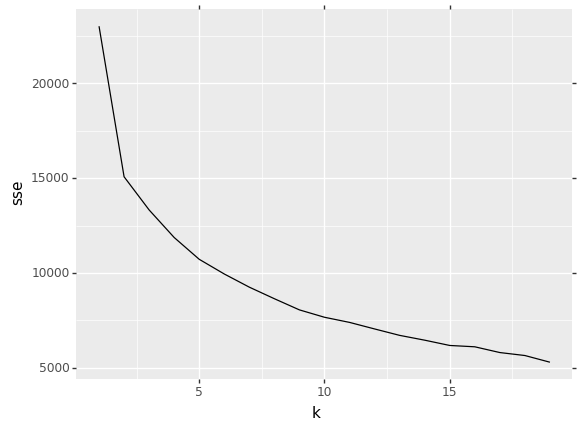
**Metrics**

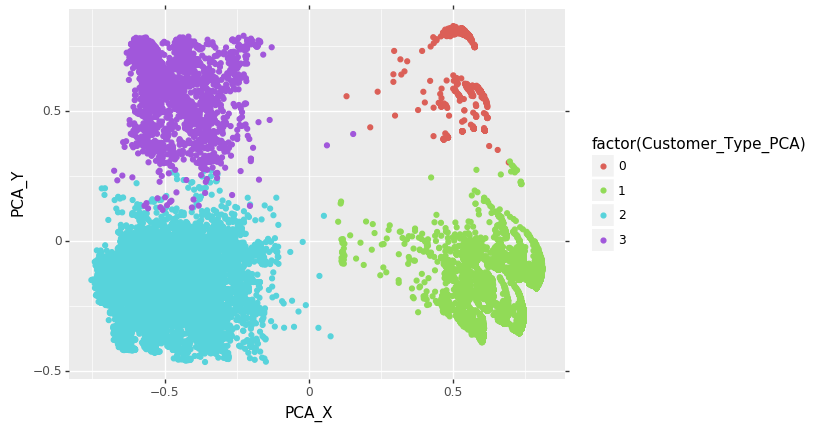
* Able to use all of the metrics the models

**Methodology**

* PCA
  + Used PCA to make graphs visually pleasing and able to show data points that encapsulates every variable
    - Using PCA also standardizes and normalizes each variable so none will be skewed for segmentations
  + Created PCA variables of 2 to use when graphing out the data points
  + Created PCA variables of 13 to use for K-means model



* K-Means Clustering
  + Used K-means to create segmentations of each plot
    - We are able to see that there are 4 defined clusters within our PCA axes
    - The plot below is an elbow plot that shows how many segments we should be separating in correlation to their sum squared error
  + After performing K-Means we are left with a graph that clearly shows the segmentations

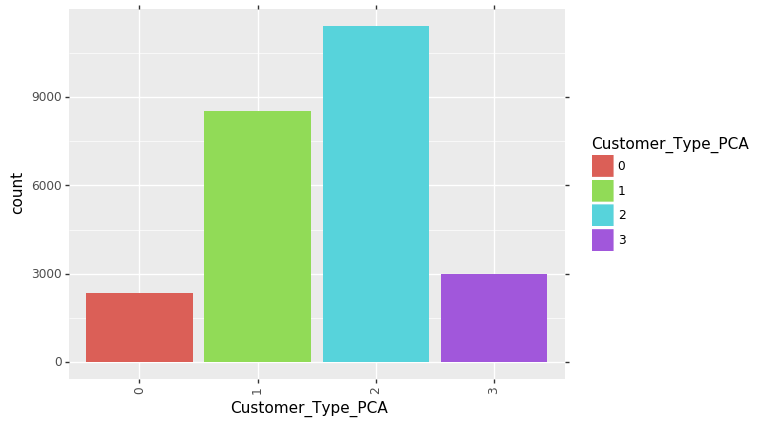


* Graphs
  + After performing the K-means clustering, I graphed out each of the variables and took the average when separated by each segment to characterize each of the segments
  + I use a combination of ggplot and group by functions to plot each variable

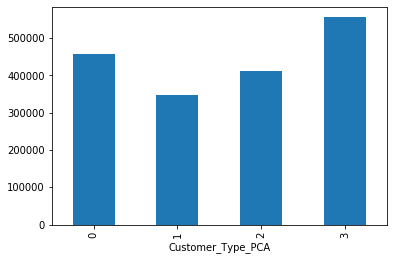
**Summary of Findings**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Segments | 0 | 1 | 2 | 3 |
| Count (Graph 1) | 2358 | 8540 | 11403 | 2990 |
| Total Purchase Amounts (Graph 2) | $457,550.96 | $348,651.29 | $411,739.36 | $555,739.26 |
| Avg. Purchase Amount per Customer (Graph 3) | $194.04 | $40.83 | $36.11 | $185.87 |
| Avg. Duration | 28779 | 257 | 359 | 28596 |
| Avg. App Launch Times | 1 | 1 | 26.85 | 22.39 |
| Avg. Send Count | 5.78 | 8.70 | 33.43 | 66.82 |
| Avg. Open Count | 2.34 | 2.31 | 5.92 | 20.89 |
| Avg. Click Count | .52 | .43 | 1.57 | 6.32 |
| Avg. Unique Open Count | .82 | .95 | 2.81 | 9.99 |
| Avg. Unique Click Count | .11 | .09 | .28 | 1.47 |
| Subscription Type | Lifetime | Limited | Limited | Lifetime |
| Subscription Event Type (Graph 4) | All Initial Purchases | Some renewal, Mostly Initial Purchase | Some renewal, Mostly Initial Purchase | All Initial Purchases |
| Demo User (Graph 5) | None | Some Demo Users | Some Demo Users | None |
| Free Trial User (Graph 6) | None | Some Free Trial Users | None | Some Free Trial Users |
| Country (Graph 7) | Non-Western | None- Western | Western | Western |
| User Type (Graph 8) | Other | Other | Consumer | Consumer |
| Email Subscriber (Graph 9) | None | None | Mostly Subscribers | Mostly Subscribers |
| Push Notifications (Graph 10) | None | None | Yes | Yes |

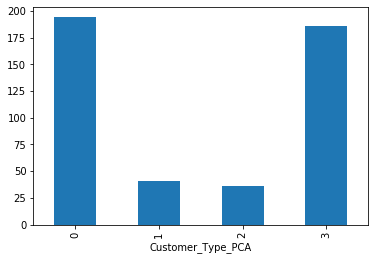
**Count (Graph 1)**



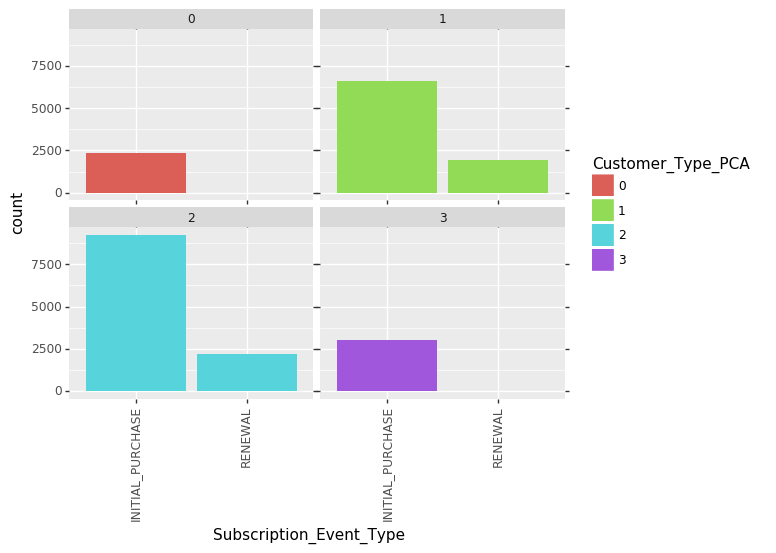
**Total Purchase Amount (Graph 2)**



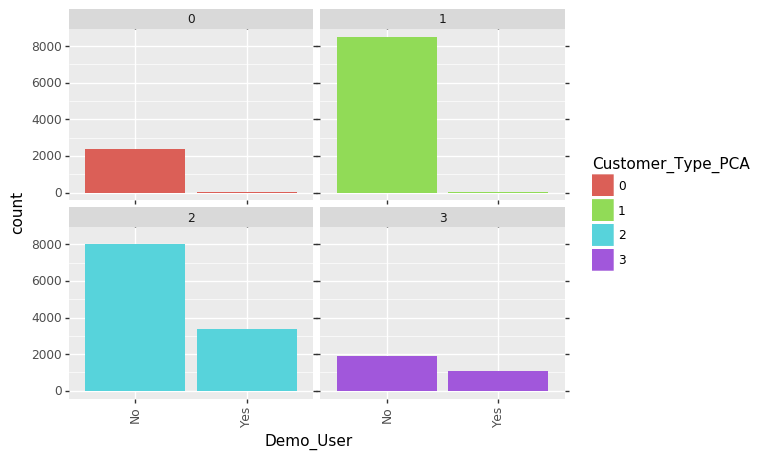
**Average Purchase Amount per Customer (Graph 3)**



**Subscription Event Type (Graph 4)**



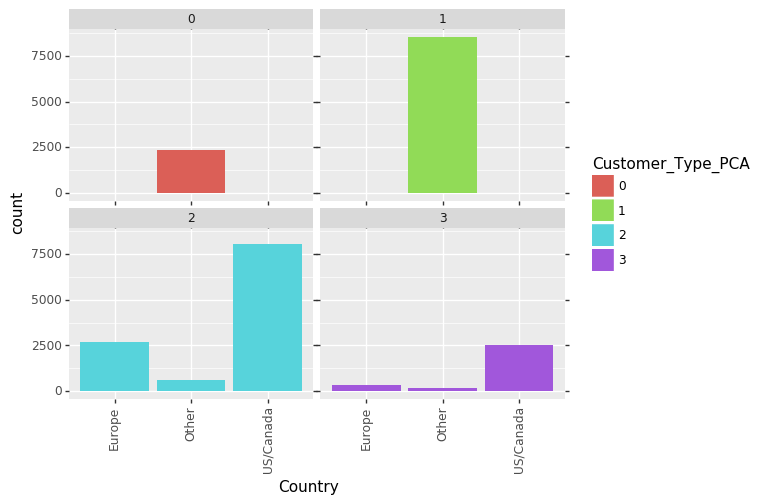
**Demo User (Graph 5)**



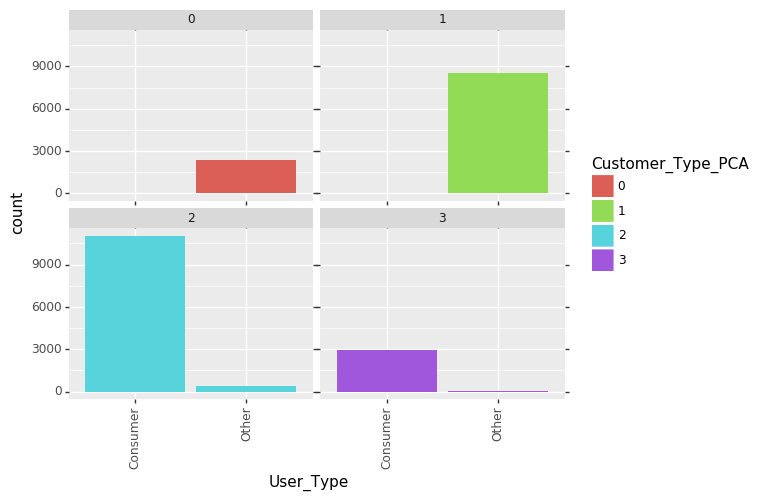
**Free Trial User (Graph 6)**



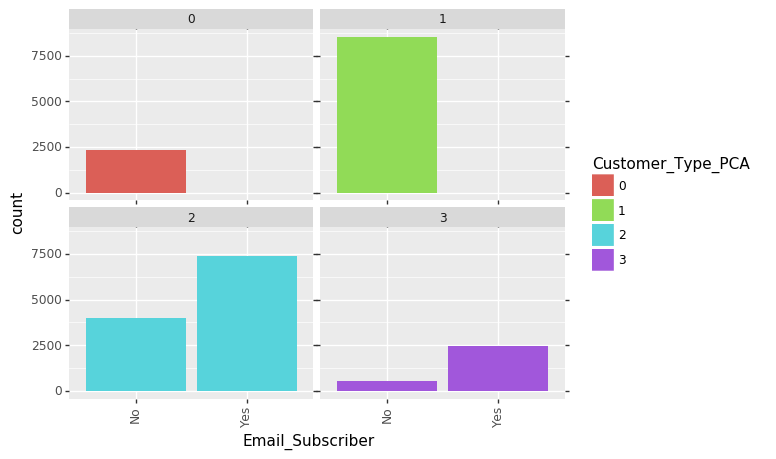
**Country (Graph 7)**



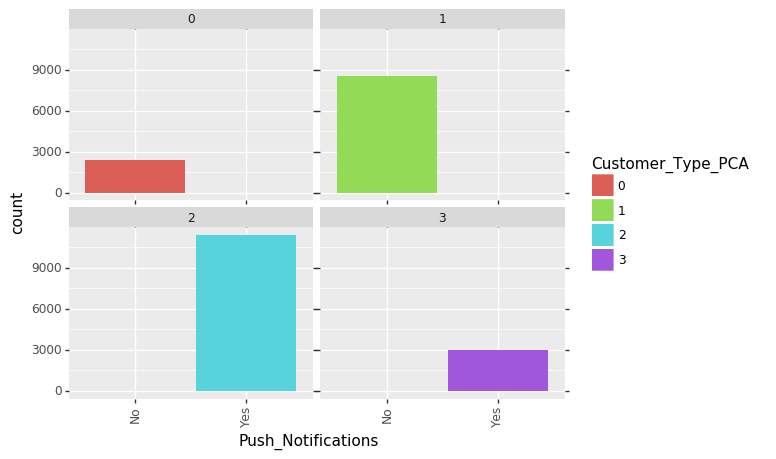
**User Type (Graph 8)**



**Email Subscriber (Graph 9)**



**Push Notifications (Graph 10)**



**Business Insights from Segmentations**

* One of the most important distinctions we have to make is that Segments 0 and 3 are Lifetime subscribers while Segments 1 & 2 are limited subscribers
  + With the lifetime subscription, segments 0 & 3 have access to all of the languages. Segments 1 & 2 mostly pick a single language, mostly Spanish
  + Average Duration shows that segments 0 & 3 have a duration of around 28,600 days that corresponds to being a lifetime subscriber. Segments 1 & 2 have an average duration of under 360.
* Graph 1 shows that a majority (78.85%) of the consumers (Segments 1 &2) are limited
* Graph 2 also shows that our minority (21.15%) is generating more than half of the revenue
* Graph 3 shows that we have a much higher average Purchase Amount per Customer for Segments 0 & 3 which explains why those segments are generating much higher revenue.
* Graph 4 shows that Segments 0 & 3 are all Initial Purchases while Segments 1 & 2 have some consumers that Renewal while the majority are Initial Purchases
* Graph 5 & 6 are very similar in that Segments 0 & 1 consists of consumers that did not use both the Demo and Free Trials. Segments 2 & 3 have some users that were using demo and free trial but a majority of them were not using demos and free trials
* Graph 7 shows that Segments 0 & 1 were separated into Others while Segments 2 &3 were separated into Us/Canada and some Europe. We have identified these segments to be Non-Western for Segments 0 & 1 and Western for Segments 2 & 3.
* Graph 8 shows that Segments 0 & 1 are on non-consumer plans while Segments 2 & 3 are mostly on consumer plans.
* Graph 9 show that Segments 0 & 1 have no consumers that are subscribed to the emails but a majority of Segments 2 & 3 are subscribed to the email
* Graph 10 is very similar to graph 9 in that Segments 0 and 1 are not using push notification while the Segment 2 & 3 are.
* For the email average statistics (Send Count, Open Count, Click Count, Unique Open Count, Unique Click Count) shows that as your ascend from segments 0-4 the averages for each stats are growing.
* Another important stat to see is that the average launch of the apps is much lower for Segments 0 & 1 at 1. Segments 2 & 3 are launching the app at around 25 times at average.

**Interpretations**

* Taking all of this into context we can see some of the defining characteristics of the segments
* Segment 0 are consumers that are Non-Western consumers that have Lifetime subscriptions. They are non-consumer accounts and are not keeping up with Rosetta Stone as seen because they are not using Push Notifications. They have a very high average price which makes up for the lack of numbers in the segments. Their total revenue from the segment was the second highest. It seems like these users were buying the apps and not using the apps at all with an average launch of 1.
* Segment 1 is similar to segment 0 except they have a limited subscription. They are also non-western consumers that do not have push notifications or keep up with the email subscriptions. They have a much lower average price but have a high number of consumers within the segment. Just Like Segment 0, they were not using the app a lot with an average of 1 time launching the app. This segment also consisted of Non-western countries that were not consumer accounts
* Segment 2 are people in the western countries (US, Canada, Europe) and have a limited subscription. These guys have a lot of use from Rosetta Stone which is shown by their higher average of launching the app. They also keep push notifications on and are subscribed to the emails. They are much more susceptible to the emails with the 2nd highest averages of email numerical statistics.
* Segment 3 are people in western countries with a lifetime subscription that are having a lot of use from their app with a high average launch times. They also keep the push notifications on, are subscribed to the emails, and have a high average for all the email statistics. This is the cash cow of the data set because they have the most total revenue from the segments..

**Business Actions**

* Looking at each segment, we can see that the most money is generated by those who have a lifetime subscription which are segments 0 & 3.
* Segments 1 & 2 have the most potential of all our segments due to being able to renew their subscription. If I were to pick a segment between the 2, I would pick segment 2 because they are more open to the emails with a higher average of click and opening of the emails.
* It should also be noted that segment 2 tend to be the consumer subscriptions which means that a single user will be using the account.

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

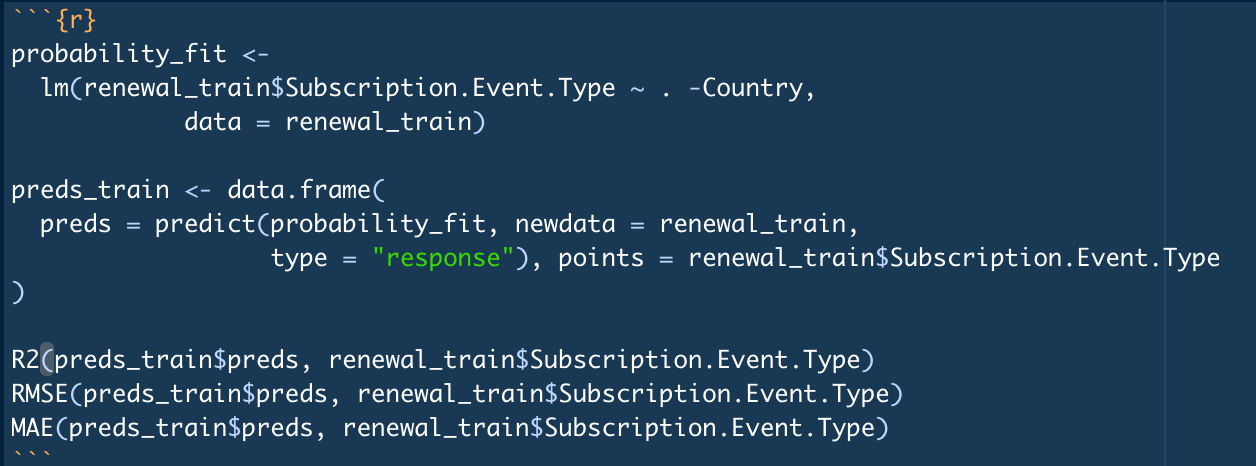
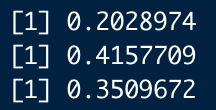
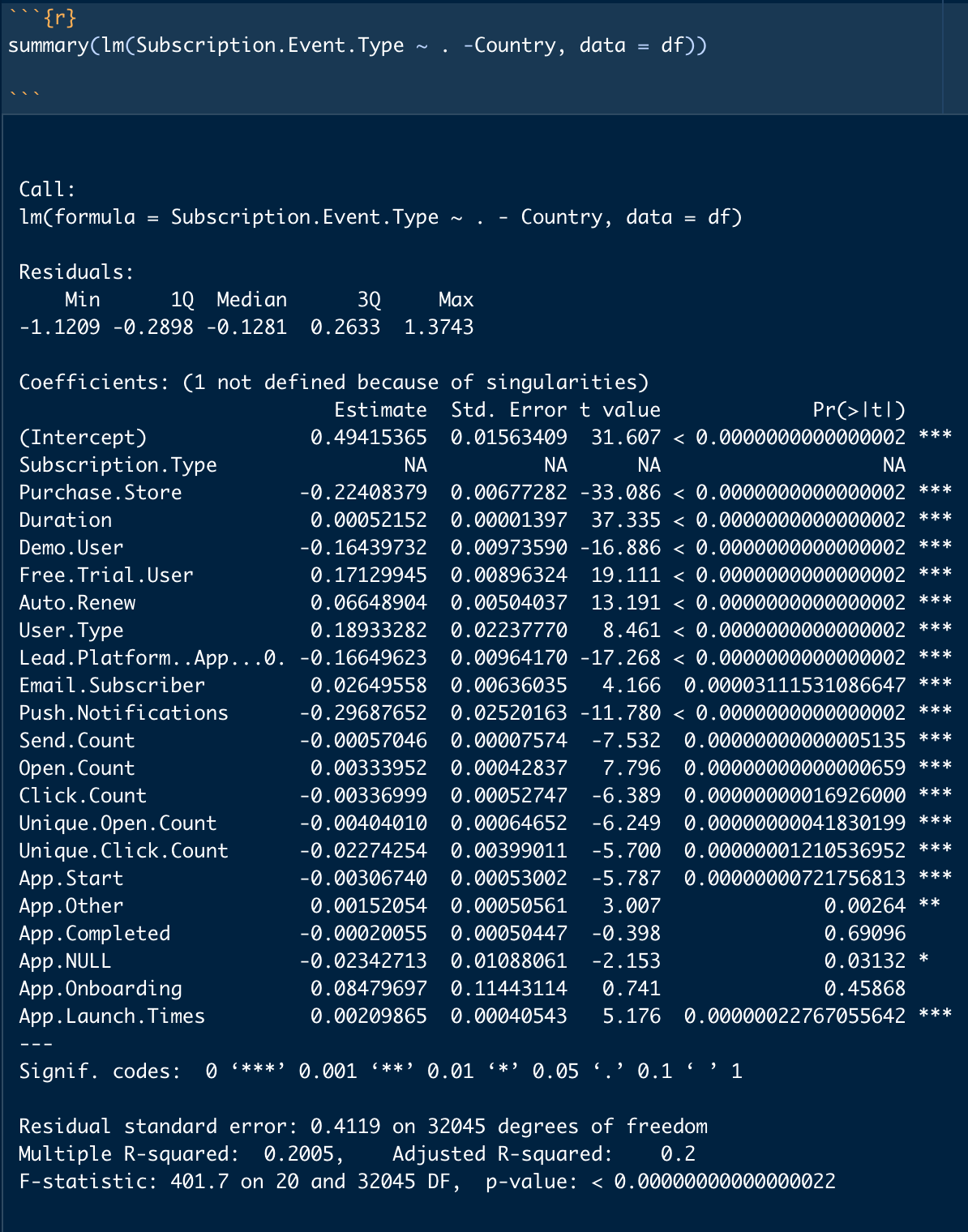
**Goal**

* Identifying the subscribers which are most likely to buy more products/services
* The subscribers who are most likely to buy more products/services are ones who have done so before in the past (we need to find these customers)

**Metrics**

* Subscription Event Type - dependent variable which determines whether or not the subscriber is an initial purchaser or has had renewals in subscription
* Continuous and binary variables as independent variables
  + Took out categorical variables

**Methodology**

* Used data exploration through powerbi to find out which consumers renewed their subscriptions
  + Only 30% of limited consumers resubscribed
  + 0 lifetime consumers resubscribed
* Changed all of the variables that were binary variables to 1s and 0s
  + Assigned different values to 1s and 0s
    - Ex: ios =1 other = 0
* Decided to use linear probability model in order to find out probability of consumers resubscribing dependent on independent binary/continuous variables
* Ran linear probability model with subscription.event.type as the dependent variable and took out country as an independent variable (R code displayed below)
* To check that the results were correct, we used chi-squared test to check and we arrived at a result of this:
  + We were able to test that Demo.User, in the chi-squared test above, had a p-value less than 0.05 at 0.03066 which is also significant in the linear probability model test.
* Got results that most variables were significant with a pvalue less than an alpha of 0.05 but noticed that most app activity didn’t affect the consumer deciding to renew or not
* Ran train/test model to test if variables could predict subscription event type
  + 
  + Achieved results of R2: 0.2029 , RMSE: 0.41577, and MAE: 0.35097
  + 
  + We were able to achieve a low RMSE and MAE which is good but the R2 is too low and is only able to account for 20% of variation within the model.
    - This can be caused by not removing the non-significant app usage data as shown in the linear probability model 

**Business Insights (use linear probability model results above)**

* The top variables in determining whether or not the consumers would/would not resubscribe were:
  + Purchase.Store
  + Demo.User
  + Free.Trial.User
  + User.Type
  + Lead.Platform.App
  + Push Notifications
* The reason why we believed that Linear Probability would be the best model in this case is because we wanted to see if any of the other variables would have an impact upon whether or not consumers would be more likely resubscribe
  + If the consumer is more likely to subscribe due to a specific variable which we can influence, it would important to highlight those variables to incentivize resubscribing or buying other products/services
  + Consumers who resubscribe are more likely to buy other products/services
* The way consumers bought their subscription mattered. If consumers bought their subscription through the Web, they would be less likely to resubscribe.
* There is surprisingly a big difference between being a demo user and a free trial user. As a demo user, there is a less likely chance that the user will resubscribe while being a free trial user increases the probability of the user resubscribing.
* If the User.Type is a consumer then they will most likely resubscribe. We assumed that if the User.Type wasn’t a consumer, then it would be through a business/organization that would only pay for the subscription as a one time fee.
* Lead.Platform coincides with the Purchase store with 0 = App and 1 = Web so the probability ended up being similar.
* Finally, push notifications with 1 = On and 0 = Off led to a result in which if consumers had push notifications on, they would be less likely to resubscribe. This can be due to consumers focusing diligently on their language plans through reminders on push notifications and would learn the language faster. Those who have their push notifications off might not get the daily notification to learn their language and might need to resubscribe again.

**Interpretation of Linear Probability**

* With the top variables mentioned above, these are the increases/decreases in likelihood of resubscribing dependent on the independent variable:

1. Purchase.Store (Web) = -22.1%
2. Demo.User (Yes) = -16.44%
3. Free.Trial.User (Yes) = 17.13%
4. User.Type (Customer) = 18.93%
5. Lead.Platform (Web) = -16.65%
6. Push notifications (On) = -29.69%
7. Duration (Per Hour) = 0.052152%

**Business Actions.**

* To incite more subscription or to sell additional products/services Rosetta Stone can send more marketing advertisements towards mobile users since it shows that mobile users tend to renew their subscriptions.
* The free trials should be advertised more when giving consumers a new account as it leads to a higher probability of users converting to renewing their subscriptions.
* More advertisements should be marketed towards individual customers rather than businesses since individual customers tended to resubscribe rather than businesses as expressed through the User.Type variable.

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

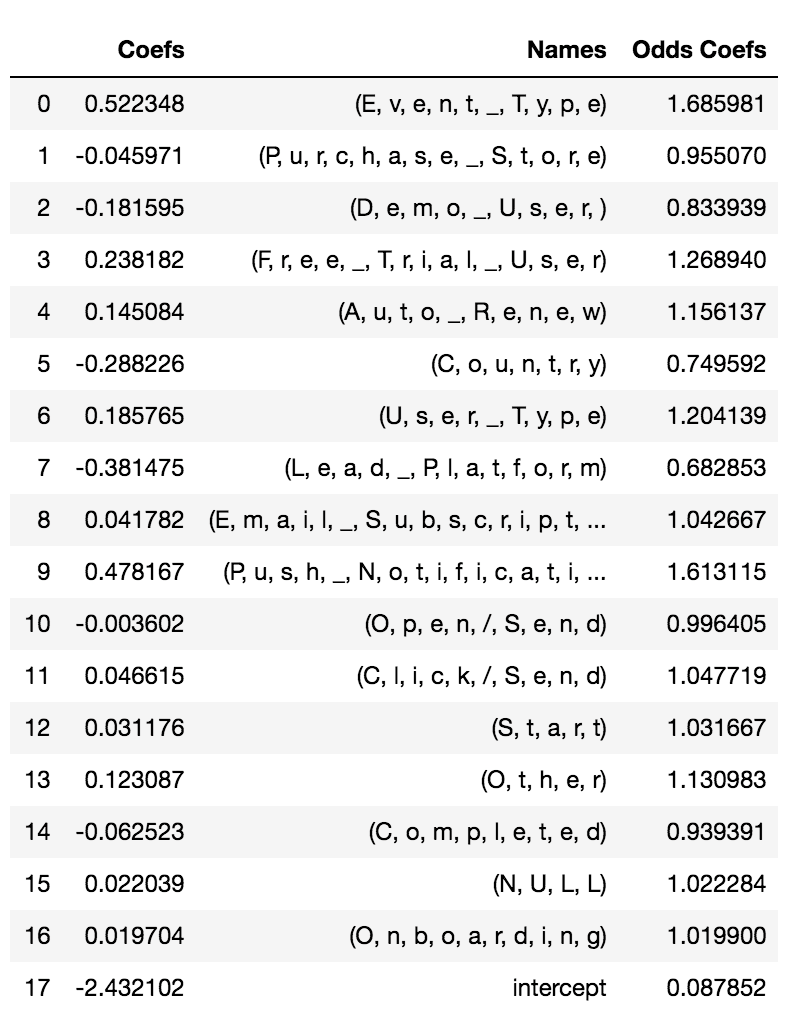
**Goal**

* Identify the subscribers who are most likely to stop using the product and find common characteristics within those subscribers

**Metrics**

* Expiration Date - The date that the subscriber stops his/her subscription
* Duration - A created variable that takes the difference of Subscription Start Date and Subscription Expiration Date
* Quitter - A created variable that defines if a subscriber stops using the product before its expiration date
* Event Type - Whether the purchase action is subscriber renewing the product or initially purchasing the product
* Push Notification - Whether subscriber is opted in to receive push notifications
* Free Triail User - Whether the subscriber registered for a limited time free trial
* User Type - Whether the subscriber is consumer or others (ex. School, business)

**Methodology**

* Defined the variable “Quitter”
  + Filtered Expiration Date and took data before August 2020 because we supposed the latest date the professor extracted the data would be August 2020 and because we would not know if a subscriber actually chooses to stop using the product if the expiration date is in the future.
  + Since a normal subscriber usually subscribes for a month / two months / three months / a year or so, we took out data where its duration is around the multiple of 30 or its duration repeated a lot in the dataset (because it might be due to some kind of promotion).
  + Define the remaining data as 1 in the “Quitter” variable
  + Define the data that was eliminated and its Expiration Date is before August 2020 as 0 in the “Quitter” variable
* Created Logistic Regression Model
  + We used “Quitter” as the y variable to see which variables influence whether a subscriber is a quitter or not.
  + Here is the code we performed for Logistic Regression:
    - 
    - The accuracy score is 90.29%
  + We created a table with variables, its coefficients, and its odds coefficients to show the result. Following is the table:
    - 

**Interpretation of Logistic Regression**

* Looking at the coefficient, the variables that have the highest absolute value of coefficient are Event Type, Push Notifications, and Lead Platform. We choose to interpret the odds coefficient and skip the log odds because odds are generally easier to understand.
* Looking at the odds coefficient, the variables that have the highest absolute value of coefficient are Event Type, Push Notification, Free Trial User, and User Type. To interpret the odds based on each variable:
  + Events Type
    - An increase in one unit of Event Type is associated with 1.69 times increase in the odds.
    - If a subscriber’s event type is renew instead of initial purchase, the more likely the subscriber will become a quitter.
    - We can make an assumption that this result is due to the level of passion towards the product. When a subscriber initially buys the product, he/she is likely to want to take time exploring and learning how to best utilize the product. Whereas subscribers who choose to renew might not renew the product on purpose, but simply because he/she forgot to turn off the auto renew system. Another reason for renew subscription being more likely to lead to the discontinue in product usage is that the subscriber might finish the lesson ahead of time and therefore choose to quit the product.
  + Push Notification
    - An increase in one unit of Push Notification is associated with 1.61 times increase in the odds.
    - If a subscriber has his/her notification on, the more likely he/she will stop using the product.
    - A possibility of why subscribers who have push notifications on are more likely to stop product usage is that subscribers got too annoyed with the notification and therefore stopped the subscription. Another possibility is that since these subscribers get notifications as reminders, they tend to take the lessons more often and thus finish the course material before the expiration date.
  + Free Trial User
    - An increase in one unit of Free Trial User is associated with 1.27 times increase in the odds.
    - If a subscriber uses free trial, the more likely he/she will stop using the product.
  + User Type
    - An increase in one unit of User Type is associated with 1.2 times increase in the odds.
    - If the user type of a subscriber is consumer instead of others, the more likely he/she will become a quitter.

**Business Insights**

* After understanding the characteristics that increase the probability of subscribers stop using the product, we found that most of the findings match up with the results from question 3 (in terms of the least potential customer).
* We highlighted some findings and came up with some ideas that might help Rosetta Stone to keep its customers:
  + From what we found in question 3, we know that subscribers who have notifications on are more likely to finish the courses earlier; those subscribers are also more likely to quit using the product afterwards, which also might be the reason why renew users are more likely to quit than initial users.
    - Rosetta Stone can provide more customized content suggestions for those subscribers who have notification on to keep them interested in the product.
  + From question 1, we know that free trial users are more likely to become a champion user. However, in question 4, we found that free trial users are also more likely to become a quitter. We believe the result is due to the different dependent variables we use in each model and can explain it as free trial users are more likely to become quitters but are also more likely to become champion users once they choose to subscribe to the product (assuming champion user and quitter are two mutually exclusive events).
    - Rosetta Stone to improve the transition process between free trial and an actual subscription to increase the amount of champion users.

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

**Goal**

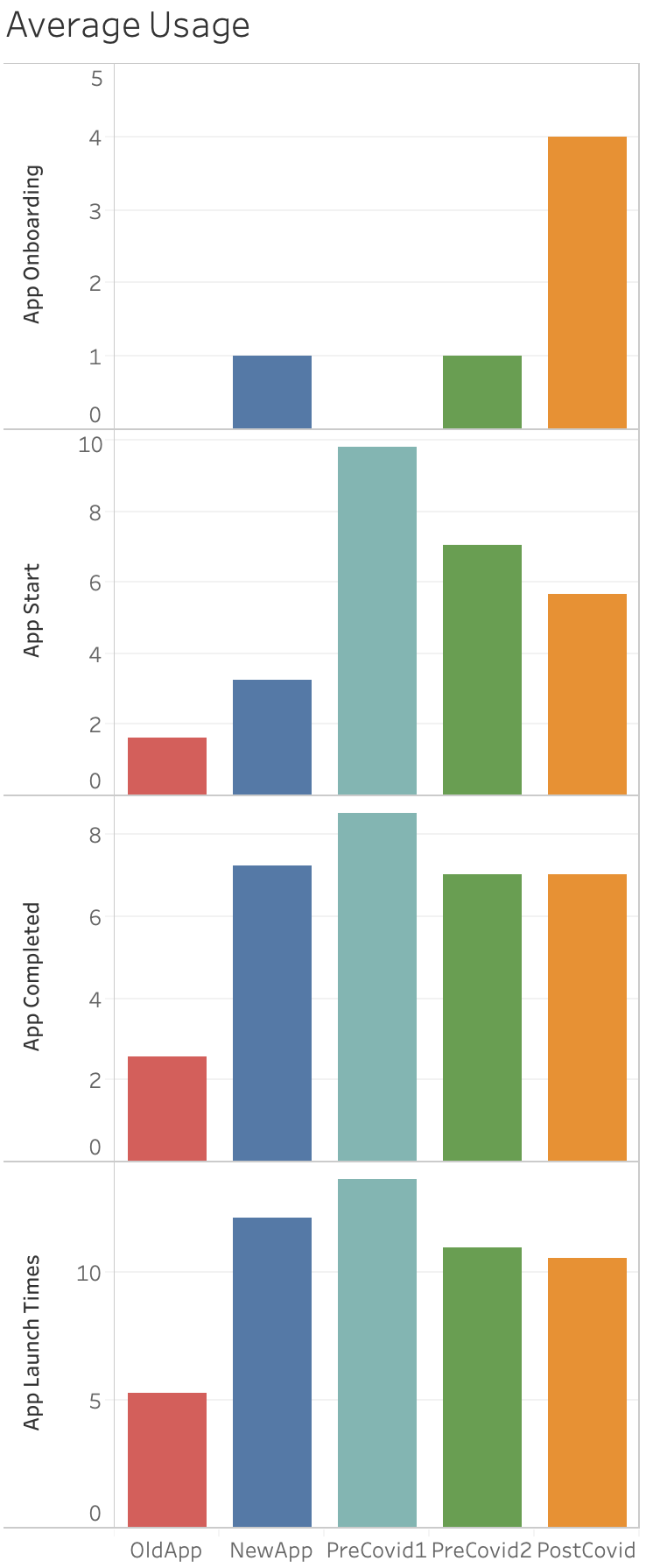
* On January 31, 2019, Rosetta Stone released a new application for iOS (optimized). We want to evaluate its performance, explore, and identify any additional insights for different time periods prior to and after the app’s launch.

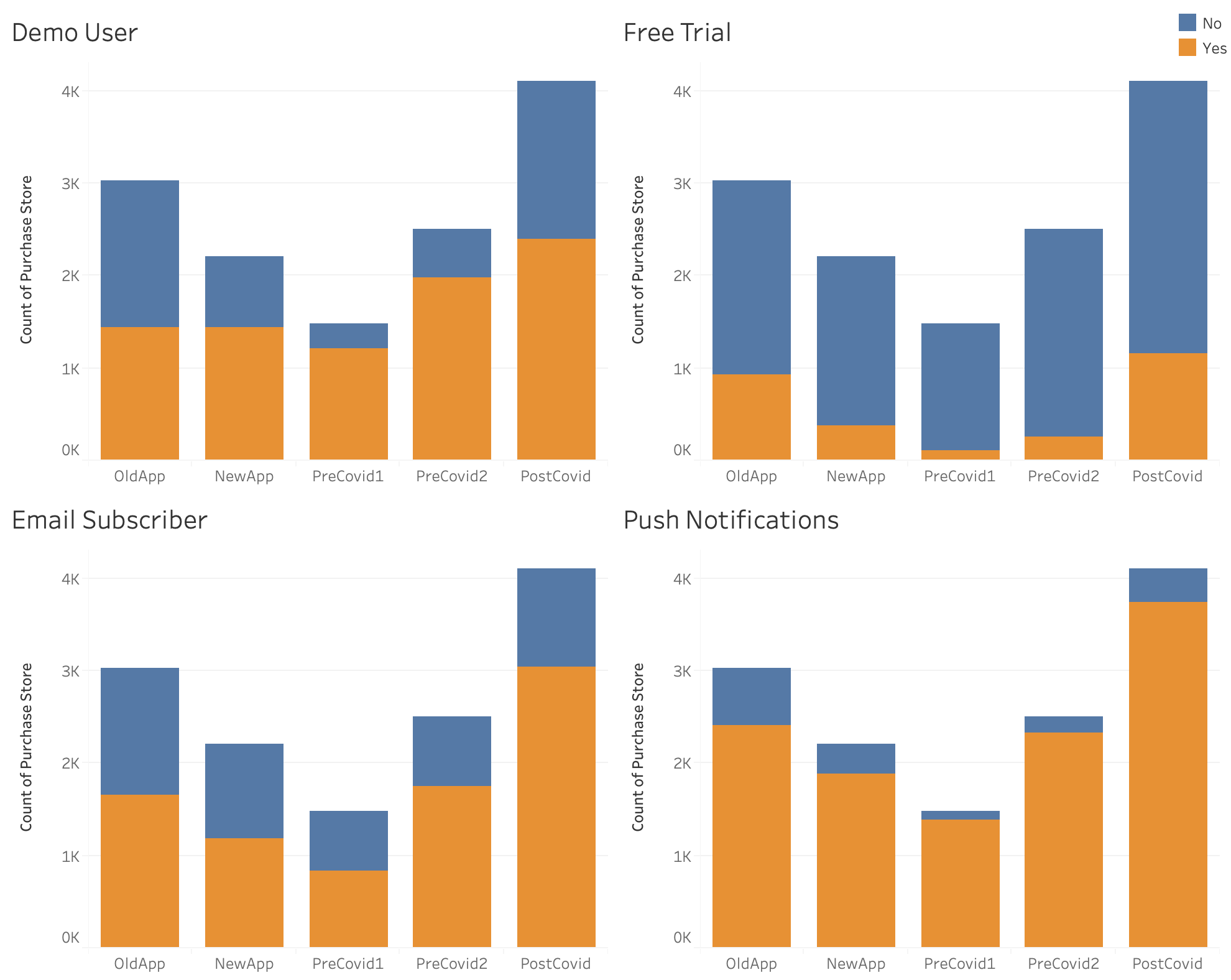
**Metrics**

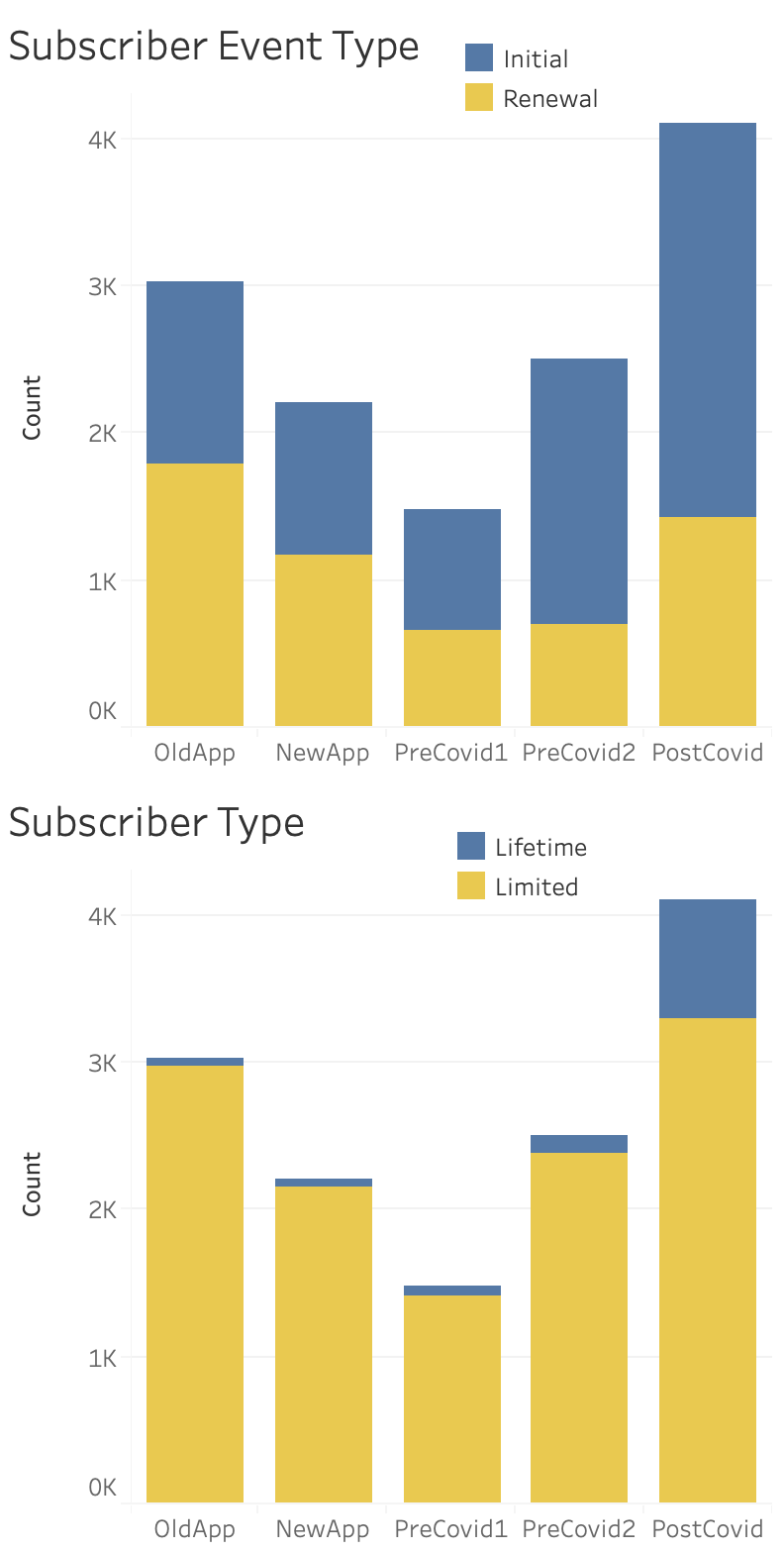
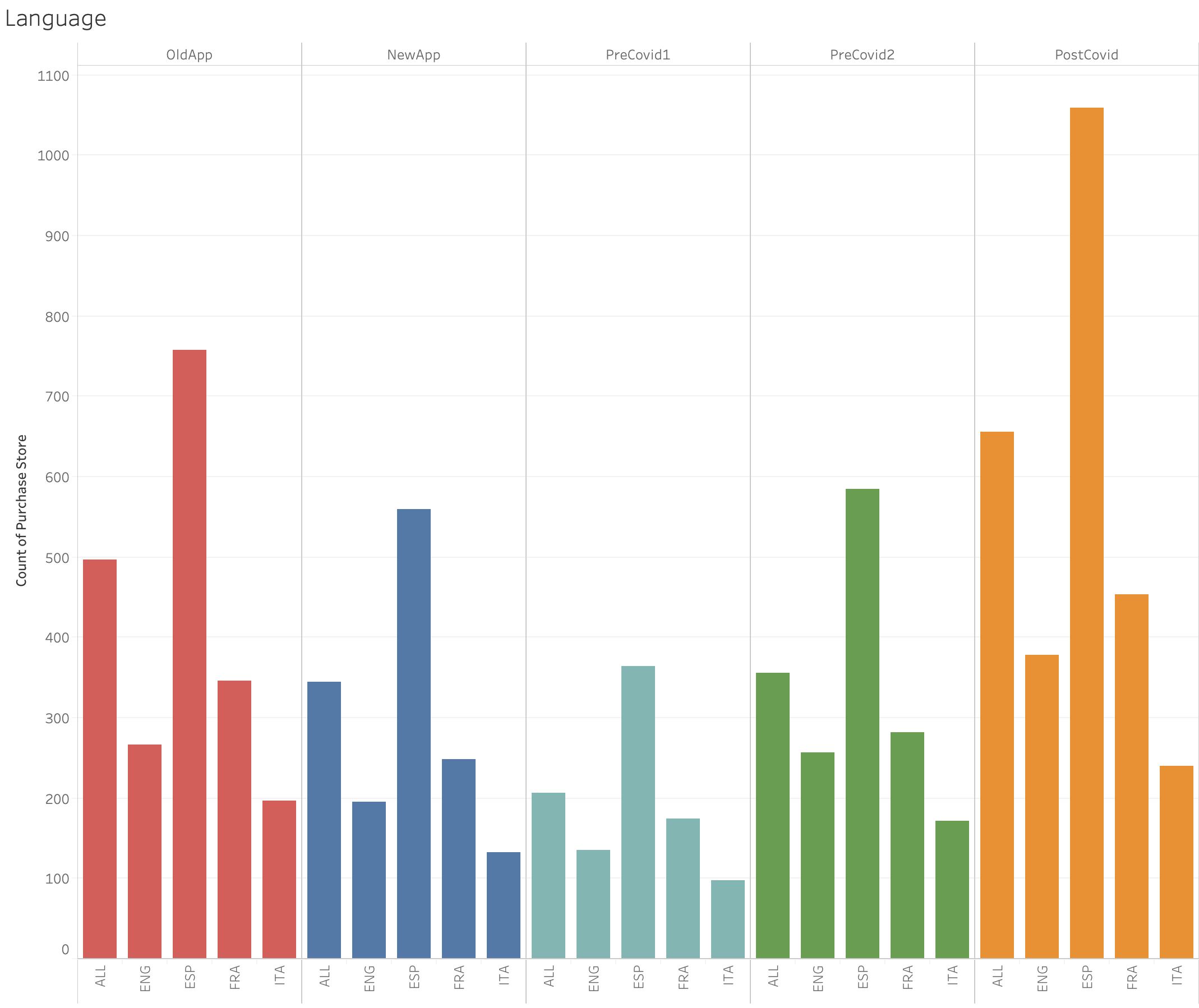
* Average App Usage
  + Onboarding: average consumption of onboarding content
  + Start: average activity related to user starting a lesson or unit
  + Completed: average activity related to user completing a lesson or unit
  + Launch Times: average number of times app was launched
* Subscriber Event Type: initial or renewed purchase
* Subscriber Type
  + Limited: defined expiration date (typically 12 or 24 months)
  + Lifetime: perpetual subscription
* Other Traits
  + Demo: subscriber has used demo content on the app
  + Free Trial: subscriber registered for a limited time free trial (usually 3 days)
  + Email Subscriber: subscriber is opted in to receive emails
  + Push Notifications: subscriber is opted in to receive push notifications
* Language.Top: factor lumped top 5 most studied languages with all else in Other category

**Methodology**

* Divided App activity into 5 main time periods using Subscription Start Date
  + OldApp: Earliest date in dataset; app users before release of optimized iOS app (about 4 months: 10/1/18-1/30/19)
  + NewApp: App users after release of optimized iOS app with OldApp duration as a base (about 4 months: 1/31/19-5/31/19)
  + PreCovid1: App users after release of optimized iOS app between PostAppShort period to PreCovid2 (about 4.5 months: 6/1/19-9/14/20)
  + PreCovid2: App users after release of optimized iOS app between PreCovid1 period until around beginning of lockdown (about 4.5 months: 9/15/20-2/29/20)
  + PostCovid: App users after release of optimized iOS app between PreCovid period until most recent date in dataset (about 1 month: 3/1/20-3/31/20)
* Examine average app usage, subscription patterns, and languages studied using time periods above in Tableau







**Interpretation**

**Average Usage**

* Onboarding: likely did not have onboarding media prior to new App launch, no users consuming content during PreCovid1 period, but spike in PreCovid2 and PostCovid
* Start: more users starting lessons on average after App launch; huge spike in users starting lessons during PreCovid1, but decline in later periods
* Completed: general trend of more users completing lessons for the first three periods, but dip and stagnate in latter 2 periods
* Launch Times: similar to behavior to lessons completed

**Subscription**

* Subscription Event Type: Lower amount of subscribers for periods between OldApp and PostCovid periods, but increasing ratio of initial purchases to renewals, which is good for the company since initial purchases indicate higher customer value
* Subscriber Type: little increase of percentage of lifetime customers in the first four periods, but PostCovid brought in higher proportion of lifetime customers, which is another positive for Rosetta Stone

**Other Traits**

* Demo User: increasing rates of demo users, which is concerning since demo users are more less likely to resubscribe; however, the rate of increase is relatively low
* Free Trial: decline in free trial users until COVID, which makes sense since there are more users in general, so they would want to test the product before buying
* Email Subscriber: similar to free trial, where there was an initial decline in email subscribers, but rates are increasing as time progresses. The rate, however, is more aggressive than free trial
* Push Notifications: over time users are opting in for more and more push notifications, which is both good and bad. It reflects a higher rate of engagement; however, this migh also drive users to complete their lessons and learn the language quicker, which is ultimately a loss.

**Challenges and limitations**

1. We noticed majority of purchase amounts from App users were ‘nulls’ and tried to predict them by looking at their subscription length and subscription type

a. We realized that this would create a high bias for any model

b. This would also lead to wrong predictions and classifications

c. We also found out that Rosetta stone prices in other countries were different from the United States, hence it would be very hard to predict purchase amounts

d. Promotional activity would also affect true cost of prices

1. We were unsure if ‘Purchase Amounts’ are reflected for the customer’s entire duration of their subscription or if it is based on a new billing cycle
   1. E.g. Some customers have a subscription length of over 400 days with a low purchase amount
   2. We thought that these customers might have renewed their subscription at 365 days and purchased another subscription at a lower price
2. Implementing a textbook / traditional CLV was challenging:
   1. In the spreadsheet we noticed there was no possible way to track purchase behavior as each main row was unique. Thus, it was hard to implement purchase frequency. This may be due to the subscription business model.
   2. Because of this challenge, we incorporated a ranking system that was inspired by CLV to group customers against each other in importance
3. We realized Rosetta Stone App was not available before January 2020, which might potentially create bias in our data
4. We know to explain App performance with caution, because the iinitial launch of optimized iOS launch may not be main driver of changes in consumer habits (i.e. COVID increased engagement and new users between certain periods rather than App). We also know that the time periods are varied and not standardized between each other, so that can certainly have an impact that is unaccounted for.