King Nguyen, Ryan Redjaian, Etai Eilat, Luis Rivas, Paul Zhao, and Todd Hartog
MGSC 410 - Group 6
30 November 2023

Rosetta Stone Project

Rosetta Stone, a leading language-learning software provider, extends its platform to a diverse global clientele, catering to individuals and corporations. In an era of increasing global interconnectedness, the significance of acquiring new languages has experienced exponential growth. Recognizing this trend, Rosetta Stone has strategically prioritized subscriber growth as a linchpin for the success of their business. In this context, a comprehensive understanding of the characteristics and behaviors exhibited by their subscriber base becomes imperative for identifying valuable opportunities for growth and success.

The primary objective of this culminating project was to actively contribute to Rosetta Stone's attainment of its distinct business growth goals. Our strategic approach involved tasks such as discerning the most valuable subscribers, conducting a nuanced analysis of the diverse subscriber segments present in the database, pinpointing subscribers with potential for additional product or service uptake, detailing subscriber profiles, identifying potential barriers hindering deeper engagement, and presenting insightful business opportunities derived from our comprehensive analysis. This concerted effort aimed to provide Rosetta Stone with a roadmap towards meeting and exceeding its business growth expectations.

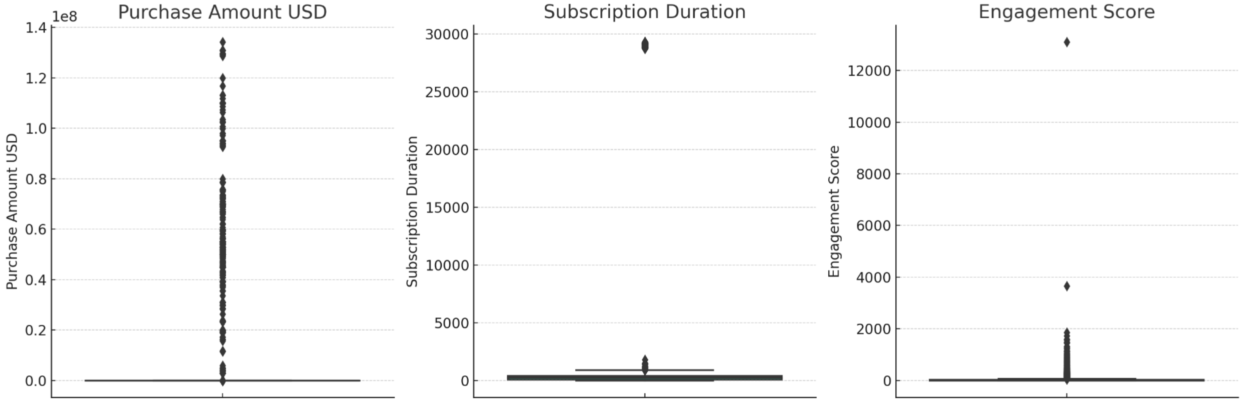
Data Cleaning:

We started by removing all the null values to see how many clean rows the dataset had. After removing all null values, we discovered 2,600 clean entries within the 40,000-row subscriber information dataset. This analysis revealed that eliminating null values could make understanding the data less attainable, thereby exerting a discernible impact on the overall quality of dataset analysis in this context. Notably, a significant concentration of null values is observed at the start and end date of the free trial period. Upon inspecting columns featuring non-null entries for trial start and end dates, it becomes apparent that the trials typically span only three days. Consequently, this particular variable may not have substantial significance in the valuation process due to the variable that says “no” or “yes” in which they have or don't have a free trial where the “no” value of not having labels a null on the start and end date. Therefore, we will keep the null values in the start and end date to keep those rows for building models, but we won't use those variables as they are insignificant.

Some columns needed more information. For example, where the purchase amount was missing, it was assumed it was a free subscription and was marked as $0. For the currency, if it was missing, it was labeled as 'None.' Also, if there were no engagement metrics like how many emails were opened or clicked, it was filled with zeros, assuming no engagement. There was one missing value in the 'Auto Renew' column, which was filled with the most common value in that column.

A few columns needed to be changed to be fit for modeling, and new columns needed to be created. The dates in the dataset were not in the correct format to work with, so they were updated, and a new variable was created called “Subscription Duration.” This variable was calculated by measuring the difference between the start and end date. The dataset also mixed up every purchase amount with the different currencies. Creating a model with each purchase being a different currency wouldn't be possible, so a consistent currency was needed. Our group chose the dollar and converted the rest using approximate exchange rates from April 2023. Many columns dealt with ‘count,’ combined under a new variable called “Engagement Score.” Another new variable was created using the other dataset, App Activity, by running a countif statement on the ID number to see how many times they interacted with the app on the app session date column: this variable was called “App Session Count.” We also created a new variable called Purchase Amount USD which used approximate exchange rates from the time period to convert that amount to a common currency for simpler analysis.

As for outliers, the dataset was filled with them. For purchase amount USD, there was a significant jump between the last values in the $900s, with the next group being in the millions of $. We found that the values greater than 1000 are 1491 rows of data. We decided that these were outliers and decided to remove them as it wouldn’t affect the total data amount by a significant amount.



Analytical Plan

1. Data Understandings
    1. Thoroughly analyze two primary datasets, the app dataset and the subscriber dataset, gaining a comprehensive understanding of their contents, structures, and interdependencies.
2. Dataset Evaluation
    1. Identify strengths and weaknesses within the datasets, assessing data quality, completeness, and potential areas of improvement
3. Feature adjustments
    1. Enhance dataset richness through creating new variables
        1. App session count to count how many sessions a customer used
        2. Engagement score was used to account for all the “count” columns
        3. Standardize currency by introducing a unified currency variable (purchase amount USD)
4. Exploratory Data Analysis (EDA)
    1. Conduct a detailed comparison of different variables within and between datasets, seeking patterns and correlations
    2. Ensure robust aggregation methods to handle NULL values effectively
5. Modeling and Visualization
    1. Explore various analytical techniques, including clustering, regression, churn, and heat maps. Utilize visualizations to enhance insights, especially visualizations generated from models.
        1. Clustering analysis to uncover deeper relationships within the data
        2. Regression models (ridge, logistic, random forests) to understand continuous and binary variables
        3. Heat maps to visualize correlations between different variables
    2. Generate graphs and plots to describe distributions of variables under other categories, especially variables newly made (engagement score, app session count).
6. Insights Generation
    1. Synthesize findings into actionable insights, aligning with Rosetta’s Business Growth Goals

Business Growth Goals

1. Determining the most valuable subscribers
    - Cluster (newly made from Purchase Amount USD): The highest value subscribers are cluster three, with an average Purchase Amount USD of $332.13, more than the other clusters, which have $179.68 and $16.
    - Subscription Duration: Easily identifies high-value subscribers. Clusters show 1618.35, 20800.73, 24271.64, more subscription duration clearly indicates more money spent. Note this shows that lifetime subscriptions are more likely to be high value customers.
    - Currency: All high-value customers pay in either EUR, GBP, or USD. Every other currency falls only in the cheapest cluster.
    - Country: People outside the US/Canada are more likely to be higher-value customers
    - Consumer: Non-consumers are more likely to be higher-value customers

    Models/Graphs

    - In RStudio, we created a k-means model to cluster the cleaned data. Had to further clean data by changing appropriate variables from character to factor types. Clustered based on \`Purchase Amount USD\` to get low, middle, and high spender clusters. This third cluster, the highest spending cluster, is the most valuable subscribers, and further analysis is done to characterize this cluster.
    - Tables included in the presentation show which variables best differentiate the clusters(characterizing subscriber segments): Purchase Amount USD, Subscription Duration, Currencies, Auto Renew, Country, User Type, Lead Platform, Push Notifications, and Open Count. Comparing cluster 3 to 1 and 2 in these tables explains what characteristics high-value customers have.

2. Understanding the subscriber segments present in the data
    - Cluster (newly made from Purchase Amount USD): The three clusters or subscriber segments have average Purchase Amount USD of $16, 179.68, and $332.13.
    - Subscription Duration: Each cluster is well-characterized, with clusters showing 1618.35, 20800.73, and 24271.64; more subscription duration clearly indicates more money spent.
    - Lead Platform: The lowest value customers are usually on the app, while middle and higher spenders usually use the web or other.

    Models/Graphs

    - Using the clusters from part one, we can characterize different subscriber segments present in the data. This allows us to deeply understand what each type of customer is like and what qualities they tend to have based on how much they spend.
    - Tables included in the presentation show which variables best differentiate the clusters(characterizing subscriber segments): Purchase Amount USD, Subscription Duration, Currencies, Auto Renew, Country, User Type, Lead Platform, Push Notifications, and Open Count.

3. Identifying subscribers who could be sold additional products or services
    - Target variable: purchase amount USD

- Language-Based Targeting: Analyzing historical subscription renewal rates across different languages provides valuable insights. Investing in advertising for languages with robust renewal rates is strategic, as it aligns with the potential shared goals and interests within these language-specific user segments.
- Auto-Renewal Engagement: Subscribers who have opted for auto-renewal demonstrate an ongoing commitment to the product, enhancing their likelihood of being well-informed about future services.
- Frequent App Engagement (**app session count**): Customers who frequently utilize Rosetta Stone products through active app sessions are likely to possess a heightened awareness of forthcoming services.
- Engagement score: customers who frequently engage with Rosetta Stone in general, be it through social media, emails, or app sessions are likely to possess more awareness of future services.
- Free trial//Demo User Engagement: Customers currently engaged in a free trial or demo phase present an attractive target audience for future service advertisements, as their initial interest makes them receptive to exploring additional offerings.

Models/Graphs

  - Heat graph of auto renew counter distribution by language
  - Bar graph of subscription duration by language
  - Bar graph of users by region
  - Other graphs unused for presentation in the appendix
- Logistic Regression
  - Explains data from app session count, engagement score, subscription length, and subscription event counter when predicting auto renew as users who auto renew are more likely to purchase future services. Model reported a classification report, accuracy score, confusion matrix, and AUC-ROC score.

4. Identifying the subscriber profile of those not continuing with their usage of the product and identifying the barriers to deeper subscriber engagement where possible
   - Target variable: purchase amount USD
   - Language-Based Targeting: Those pursuing less popular languages may encounter limitations in terms of comprehensive lessons and content, potentially leading to shorter durations of subscription usage.
   - Auto-Renewal Engagement: Customers without auto-renewal in place display a diminished likelihood of maintaining their subscriptions over time.
   - Subscription duration analysis: examining the start and end dates of customer subscriptions for notable trends. Customers opting for short-term subscriptions, for instance, exhibit a higher likelihood of subscription cancellation when compared to those with longer-term commitments.
   - Frequent App Engagement (**app session count**): Customers with a lower count of app sessions demonstrate a reduced probability of continuing their subscriptions.

- Free Trial/Demo User Engagement: Those who have not engaged with the demo or free trial are predicted to be less inclined to make a purchase or maintain their present subscriptions

Models/Graphs

  - Randomforests feature and importance distribution
  - Subscription duration, unique open count distribution, app session count distribution, and unique click count distributions by churn classification
  - Other graphs unused for presentation in the appendix
- Ridge Regression model discusses ridge coefficients from feature variables with MSE & R^2 values for analysis
- Randomforests Regression Model Results

5. Outline any business-relevant opportunities from your analysis of the data not covered above.

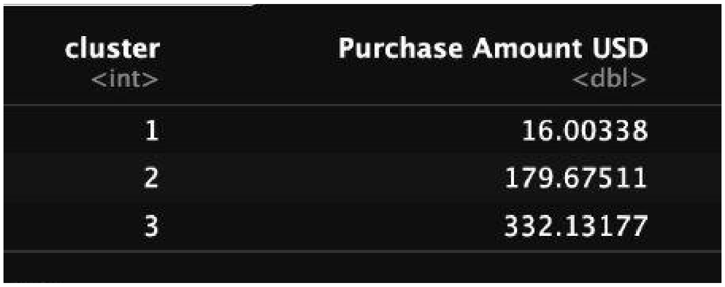
- Tailoring Marketing Strategies: Customizing marketing strategies based on distinct customer groups (low, medium, and high spenders) is essential. This involves offering personalized services or products aligned with their preferences. Focusing on specific demographics or high spenders can notably boost sales and facilitate global expansion beyond the US market.
- Enhancing Email Campaigns: Refining email campaigns based on user engagement is crucial. By crafting more personalized and engaging emails for highly interactive users, businesses can effectively retain subscribers and drive higher sales conversions. This personalized communication approach strengthens the bond between the brand and its audience. This is beneficial because we can see a lot of the interactions coming from the web as seen in the appendix visualizations.
- Product Development through User Analysis: Analyzing user engagement with product features or content is vital. This analysis can help develop products to make subscriptions more appealing and valuable to customers. Identifying and emphasizing features that resonate with users can significantly enhance customer satisfaction and loyalty. This can be seen through comparison of features and the amount of dollars spent in the appendix.

<u>**Data Analysis**</u>

KMeans Clustering:

Using RStudio: changed appropriate variables from character to factor types to analyze more easily. We created a K-means clustering model on `Purchase Amount USD` on the final cleaned dataset.

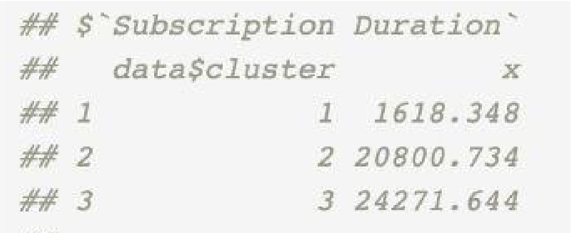
- Cluster 1: low spenders (least valuable customers)
- Cluster 2: middle spenders
- Cluster 3: high spenders (most valuable customers)
- Average purchase amounts in USD by cluster:



```
cluster    Purchase Amount USD
  <int>                  <dbl>
      1               16.00338
      2              179.67511
      3              332.13177
```

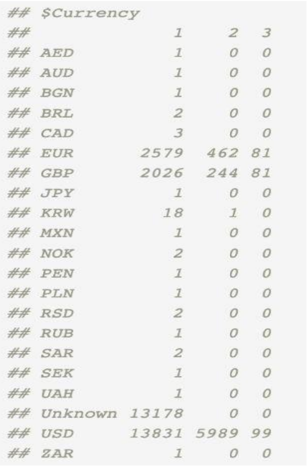
We conducted analyses on the clusters to understand how different variables vary across the different clusters. Below are the variables that best characterize **high-value customers:**

Subscription duration is a significant variable for determining if a customer is in the high, middle, or lower spending cluster.



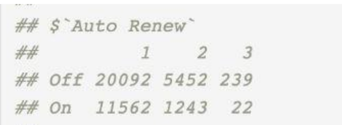
```
## $`Subscription Duration`
##   data$cluster         x
## 1            1  1618.348
## 2            2 20800.734
## 3            3 24271.644
```

All high-value customers pay in either EUR, GBP, or USD. Every other currency falls exclusively in the cheapest cluster.



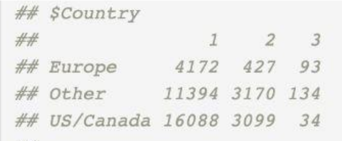
```
## $Currency
##              1    2  3
## AED          1    0  0
## AUD          1    0  0
## BGN          1    0  0
## BRL          2    0  0
## CAD          3    0  0
## EUR       2579  462 81
## GBP       2026  244 81
## JPY          1    0  0
## KRW         18    1  0
## MXN          1    0  0
## NOK          2    0  0
## PEN          1    0  0
## PLN          1    0  0
## RSD          2    0  0
## RUB          1    0  0
## SAR          2    0  0
## SEK          1    0  0
## UAH          1    0  0
## Unknown  13178    0  0
## USD      13831 5989 99
## ZAR          1    0  0
```

High-value customers are more likely to have auto-renew off.



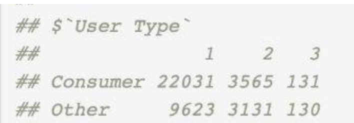
```
## $`Auto Renew`
##         1    2   3
## Off 20092 5452 239
## On  11562 1243  22
```

People outside the US/Canada are more likely to be high-value customers.



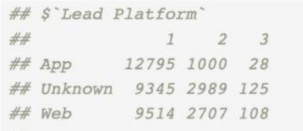
```
## $Country
##               1    2   3
## Europe     4172  427  93
## Other     11394 3170 134
## US/Canada 16088 3099  34
```

Non-consumers are more likely to be higher-value customers.



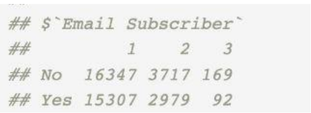
```
## $`User Type`
##              1    2   3
## Consumer 22031 3565 131
## Other     9623 3131 130
```

Non-App users are more likely to be high-value users.



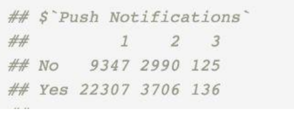
```
## $`Lead Platform`
##             1    2   3
## App     12795 1000  28
## Unknown  9345 2989 125
## Web      9514 2707 108
```

Non-email subscribers are more likely to be high-value customers.



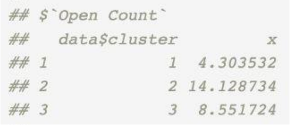
```
## $`Email Subscriber`
##         1    2   3
## No  16347 3717 169
## Yes 15307 2979  92
```

People with push notifications off are more likely to be high-value customers.



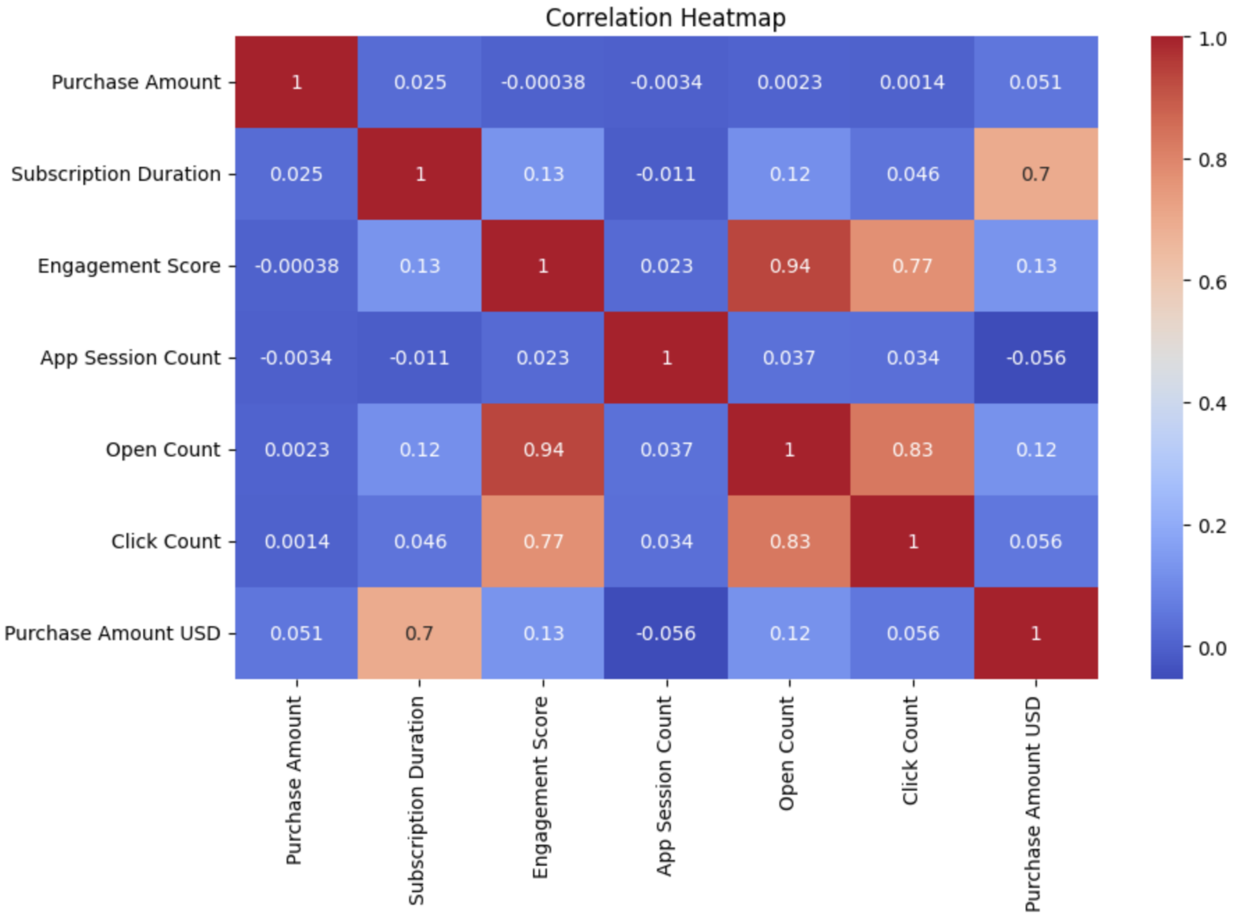
```
## $`Push Notifications`
##         1    2   3
## No   9347 2990 125
## Yes 22307 3706 136
```

High-value customers have higher open counts.



```
## $`Open Count`
##   data$cluster         x
## 1            1  4.303532
## 2            2 14.128734
## 3            3  8.551724
```

Correlation Heat Map and Visualizations:



	Purchase Amount	Subscription Duration	Engagement Score	App Session Count	Open Count	Click Count	Purchase Amount USD
Purchase Amount	1	0.025	-0.00038	-0.0034	0.0023	0.0014	0.051
Subscription Duration	0.025	1	0.13	-0.011	0.12	0.046	0.7
Engagement Score	-0.00038	0.13	1	0.023	0.94	0.77	0.13
App Session Count	-0.0034	-0.011	0.023	1	0.037	0.034	-0.056
Open Count	0.0023	0.12	0.94	0.037	1	0.83	0.12
Click Count	0.0014	0.046	0.77	0.034	0.83	1	0.056
Purchase Amount USD	0.051	0.7	0.13	-0.056	0.12	0.056	1

Engagement and Subscription Duration:

- 'Engagement Score' has a moderate correlation (0.7) with 'Subscription Duration'. This could imply that more engaged users tend to have longer subscription durations. Understanding the drivers of engagement is key to increasing subscription lengths.

Correlation Between Engagement and Purchase Amount:

- There is a correlation between 'Engagement Score' and 'Purchase Amount USD' (0.13), although it's not very strong. It suggests that subscribers with higher engagement scores might spend more, but other factors likely also play a significant role in the purchase amount.

App Usage:

- 'App Session Count' has a negligible correlation with 'Purchase Amount' and 'Purchase Amount USD', indicating that app usage frequency does not directly translate to the amount spent. This might suggest that merely using the app frequently doesn't necessarily mean a subscriber will spend more, and it emphasizes the need to look deeper into what types of engagement or app usage lead to higher revenue.

Potential Revenue Opportunities:

The strong correlation between email engagement metrics and 'Engagement Score' can be leveraged for targeted marketing campaigns. Since 'Open Count' and 'Click Count' are strongly correlated with 'Engagement Score', marketing efforts focused on email campaigns could effectively increase user engagement and potentially subscription duration.

Low Correlation Between App Usage and Email Engagement:

'App Session Count' has a low correlation with both 'Open Count' and 'Click Count', indicating that app engagement and email engagement might represent different user behaviors. It might be beneficial to segment users based on their preferred engagement channels.

These insights suggest several strategies for Rosetta Stone to consider based on these heat maps:

Targeted Email Marketing: Since email engagement correlates with higher engagement scores, personalized and targeted email campaigns could help improve subscriber retention and upsell opportunities.

User Segmentation: Segmenting users by their engagement patterns could enable more personalized experiences or targeted product offerings.

Product Development: Investigate whether features or content that drive higher engagement could be enhanced or developed further to increase the value and appeal of the subscriptions.

Churn Indicator and Visualizations:

In this context, a churn indicator is a binary variable that signifies whether a subscriber has discontinued their usage or has continued using the service. It helps identify subscribers likely to cancel or stop using the product, allowing companies like Rosetta Stone to retain these customers.

Subscription Expiration Date: If the subscription expiration date has passed, it might indicate that the subscriber has churned.

Engagement Metrics: Low engagement metrics, such as fewer app sessions, fewer clicks, or less interaction, could signify potential churn.

Subscription Duration:

The distribution for 'Subscription Duration' shows that users who have not churned ('No') have shorter subscription durations, with a peak at the lower end. A long tail extends towards higher subscription durations, with very few counts, which could include outliers or long-term subscribers. The presence of churned users ('Yes') is significantly less across all subscription durations, indicating that users are not heavily churning after specific subscription periods or that the data for churned users is limited.

For 'App Session Count,' most users, both churned and not churned, have a low count of app sessions, with numbers declining as the session count increases. Most users must be more heavily engaged with the app. The similarity in the distribution for both churned and non-churned users indicates that app session count alone may not be a strong predictor of churn.

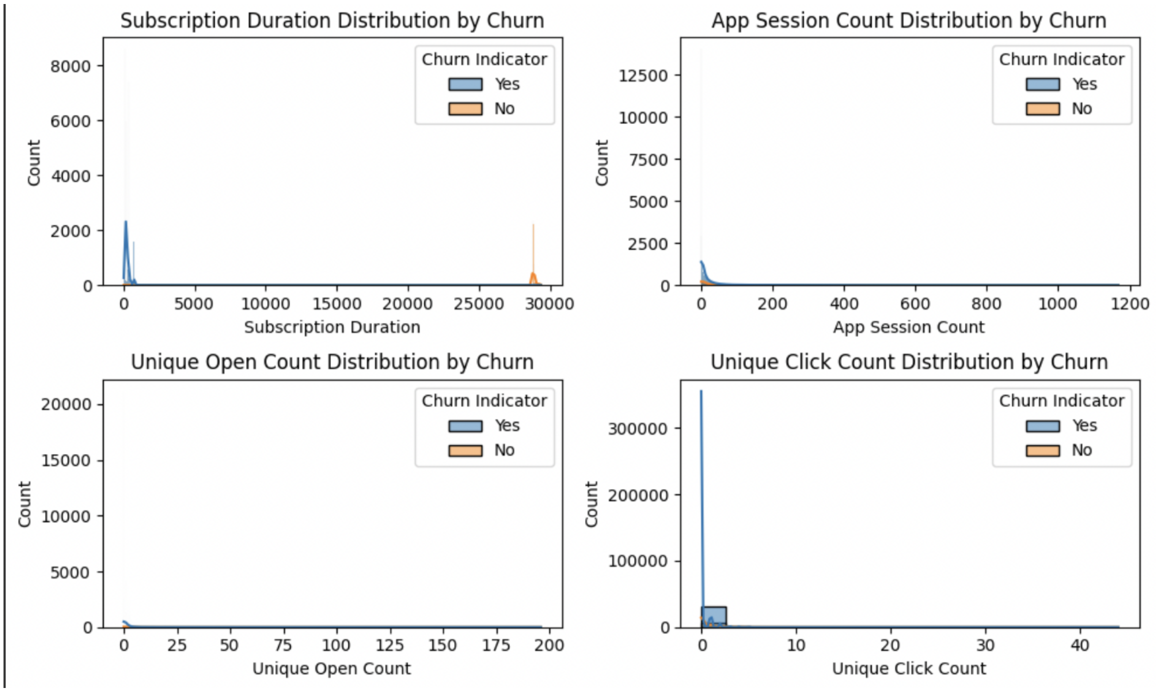
Unique Open Count:

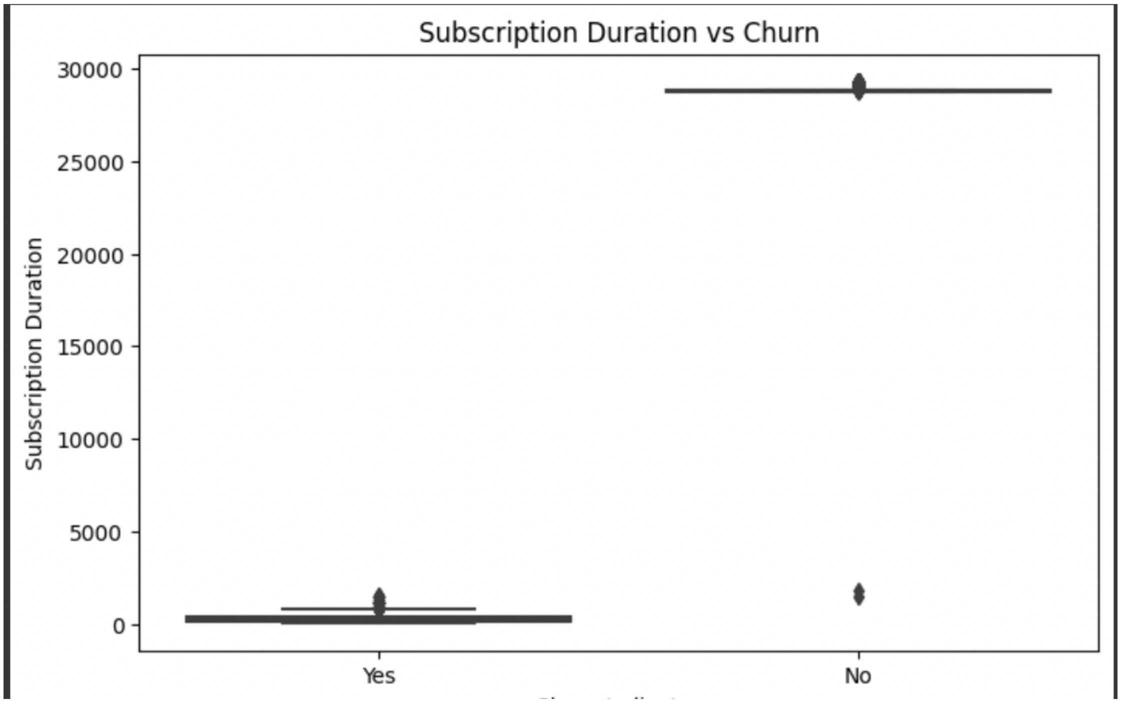
The 'Unique Open Count' histogram shows that most users have a deficient number of unique email opens, with very few reaching higher counts. There is a slight indication that users who have not churned may have a higher open count, suggesting a potential relationship between email engagement and churn.

Unique Click Count:

Similar to the 'Unique Open Count,' the 'Unique Click Count' shows that most users have clicked on emails a few times, with the number of users decreasing sharply as the click count increases. There's a visible but not substantial distinction between churned and non-churned users, suggesting that while email clicks indicate engagement, they may need to be a stronger standalone predictor of churn.

From these insights, we can infer that engagement metrics such as email opens and clicks show some variation between churned and non-churned users may need to be combined with other factors for a more accurate churn prediction. Additionally, the 'Subscription Duration' and 'App Session Count' distributions suggest that a more complex model that considers multiple variables might be necessary to predict churn effectively. The presence of potential outliers in 'Subscription Duration' could be skewing the analysis and might need further investigation


Subscription Duration Distribution by Churn
Count
8000
6000
4000
2000
0
0
5000
10000
15000
20000
25000
30000
Subscription Duration
Churn Indicator
Yes
No
App Session Count Distribution by Churn
Count
12500
10000
7500
5000
2500
0
0
200
400
600
800
1000
1200
App Session Count
Churn Indicator
Yes
No
Unique Open Count Distribution by Churn
Count
20000
15000
10000
5000
0
0
25
50
75
100
125
150
175
200
Unique Open Count
Churn Indicator
Yes
No
Unique Click Count Distribution by Churn
Count
300000
200000
100000
0
0
10
20
30
40
Unique Click Count
Churn Indicator
Yes
No

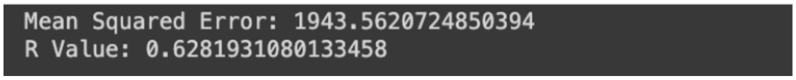

Subscription Duration vs Churn
Subscription Duration
30000
25000
20000
15000
10000
5000
0
Yes
No

Regression Models:

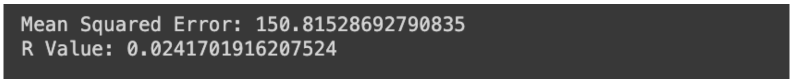
We ran regression models on variables similarly to how we constructed the clustering models. We created plots that visualize the data regarding Purchase Amount USD as scatterplots, boxplots, and histograms, as each graph shows the data differently. All code for models and visualizations can be find in the following google collab file.

https://colab.research.google.com/drive/1RPdPLGGByvgKZ8j9eBvppfUm5AOzNZ31?usp=sharing

1. Linear regression: the linear regression contained feature variables on the subscription type counter (1 is limited), counter subscription event type (1 in initial_purchase), purchase store count (1 is App), auto-renew counter (1 is on), email subscriber count (1 is yes), and push notification count (1 is yes). These variables were used to predict the target variable, Purchase Amount USD, representing every transaction translated back to USD to allow for a simpler overview of financials. The regression model selected the variables, split the data into training and testing sets, were normalized, and z-scored. The test set was transformed using a scaler to maintain consistency; the model was then initialized, fitted, and ready for prediction. The model predicted on the test set returned an R-value of 0.628 and returned an MSE of 1943.56. Since the MSE value was so high, we decided not to consider the results from the linear regression model heavily and attempted the other models below. We attempted another linear regression model that had the same feature variables, but had a target variable of click count. While the MSE was significantly lower (150.815), the R-value was deemed drastically too low for consideration in our business questions (0.628).

    a. 

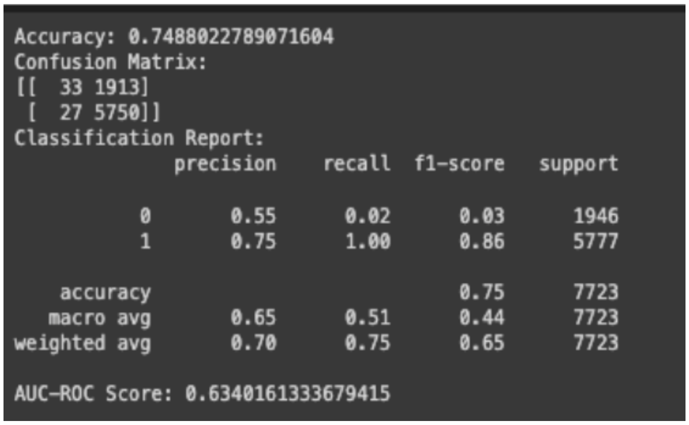
    ```
    Mean Squared Error: 1943.5620724850394
    R Value: 0.6281931080133458
    ```

    b. 

    ```
    Mean Squared Error: 150.81528692790835
    R Value: 0.0241701916207524
    ```

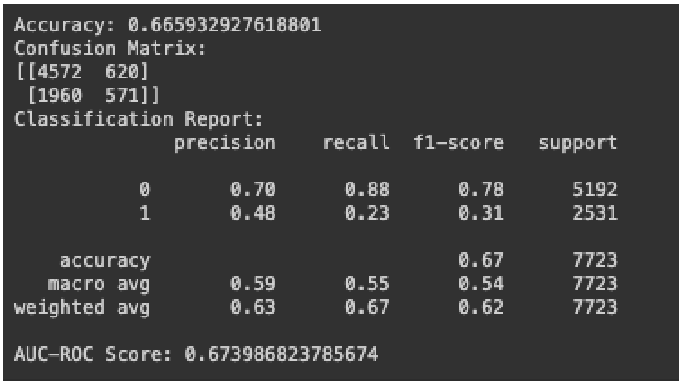
1. Logistic Regression: the logistic regression model was incompatible with Purchase Amount USD as a target variable since logistic regression predicts binary variables, so instead, subscription duration, engagement score, and app session count were utilized as features that would help predict our target variable (Counter Subscription Event Type) which indicated if purchases were isolated events or renewals. The model split the data into training and testing sets, were normalized, and z-scored. The test set was transformed using scaler for consistency and was then intialized to fit the logistic regression model. Finally, the model was evaluated and return an accuracy, confusion matrix, a classification report, and a AUC-ROC score.

    a. An accuracy of 0.7488 means that the mode correctly predicted the target variable 74.88% of the time. The confusion matrix provides the number of times it predicted the target variable (counter subscription event type) true positively, true negatively, false positively, and false negatively. The precision scores explain the ratio of correctly predicted positive observations to the total predicted positives as high precision relates to the low false positive rate. A recall score explains the ratio of correctly predicted positive observations to all observations. High recall relates to a low false negative rate. The F1 score is a weighted average of precision and recall which ranges from 0 to 1. As we can see, the target variable's renewals were correctly predicted 55% of the time and the target variable’s initial purchases were correctly predicted 75% of the time, which is great for how dense the combined dataset came to be. Interestingly, the logistic regression model had a incredibly low recall score for renewals (0.02) and 100% score for initial purchases, which means that the model correctly predicted that were were no false negatives of initial purchases and never mistook them as renewals. A score of 0.02, however, means that the model falsely predicted scenarios where transactions were renewals when they were indeed initial purchases. Lastly, the AUC-ROC represents the model’s ability to distinguish between classes. A score of 0.6340 suggests a moderate ability to distinguish between the class and report an overall moderate performance.

i. 

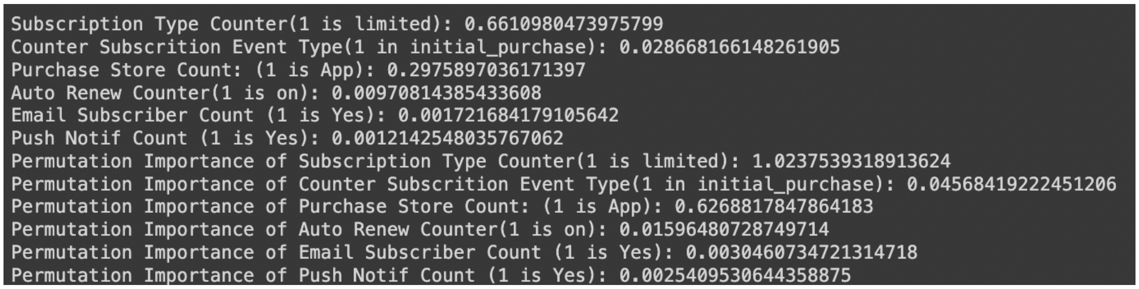
```
Accuracy: 0.7488022789071604
Confusion Matrix:
[[  33 1913]
 [  27 5750]]
Classification Report:
              precision    recall  f1-score   support

           0       0.55      0.02      0.03      1946
           1       0.75      1.00      0.86      5777

    accuracy                           0.75      7723
   macro avg       0.65      0.51      0.44      7723
weighted avg       0.70      0.75      0.65      7723

AUC-ROC Score: 0.6340161333679415
```

b. **This logistic regression model was used in our slides** and had feature variables of: Subscription Duration, engagement score, app session count, and counter subscription event type (1 in initial_purchase). The target variable predicted was auto renew counter (1 is on). The logistic regression model followed the same procedure as the model above and reported the following classification report with an accuracy of 0.6659, meaning the model made correct predictions 70% of the time. As more users had auto renew off, it correctly predicted it 70% of the time and correctly predicted auto renew on 48% of the time. The recall scores were less deviant than the previous logistic regression model as auto renew off had a recall of 0.88, as there were a few false negatives but were overall correctly predicted. A majority of the false negatives were from the auto-renew on group as the recall score was 0.23. The AUC-ROC score was 0.6739, which similarly to the previous model, means that the logistic regression model did a moderate overall performance on distinguishing between classes.

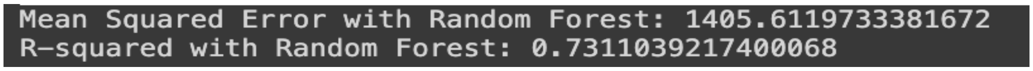
i. 

```
Accuracy: 0.665932927618801
Confusion Matrix:
[[4572  620]
 [1960  571]]
Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.88      0.78      5192
           1       0.48      0.23      0.31      2531

    accuracy                           0.67      7723
   macro avg       0.59      0.55      0.54      7723
weighted avg       0.63      0.67      0.62      7723

AUC-ROC Score: 0.673986823785674
```

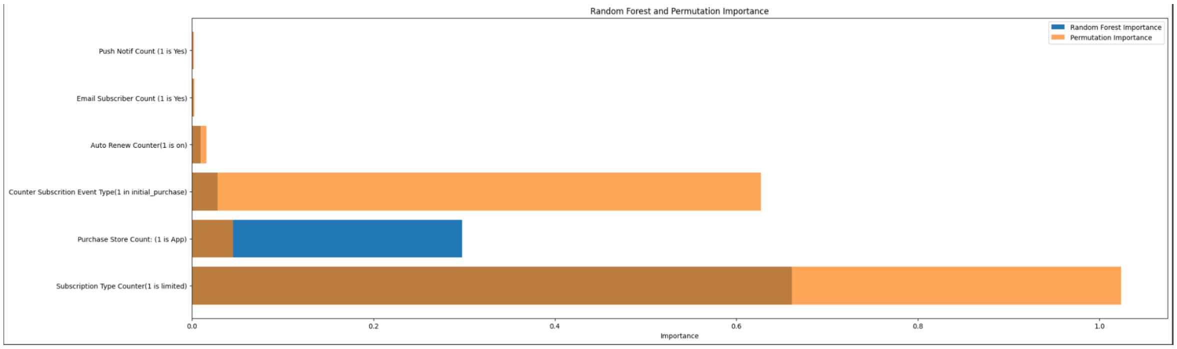
2. Random Forests: The random forests model had feature variables of subscription type counter (1 is limited), counter subscription event type (1 in intial_purchase), purchase store count: (1 is app), auto renew counter (1 is on), email subscriber count (1 is yes) and push notif count (1 is yes). The target variable was Purchase Amount USD. The variables were selected and then split into training and testing sets which was then initialized and fit into the Random Forest regressor. This allowed us to get feature importances from the model and to calculate the permutation importances too. Afterwards, we plotted the feature importance, predicted on the test set, and evaluated the model through MSE and R^2 scores. The report below holds coefficients that contribute to reducing the impurity (MSE) in the decision trees that make up the random forest. Higher values indicate greater importance. As we can see, it looks like subscription type counter is the most important feature. **These images were used in our presentation** and explain their relevance to Purchase Amount USD through permutation importance which shuffles the values of each feature as the model observes its effect on performance. In general, the major contributors to Purchase Amount USD were **Subscription Type Counter (1 is limited), Purchase Store Count (1 is App), and Counter Subscription Event Type (1 in initial_purchase)**. While the R^2 value was healthy and strong, the MSE value was far too high, but was still taken into consideration as it was the second lowest MSE calculated from the regression models and was taken into account by the dense dataset. The MSE score translates to lots of error, but still generated great insights to our business questions.

    a. 

```
Subscription Type Counter(1 is limited): 0.6610980473975799
Counter Subscription Event Type(1 in initial_purchase): 0.028668166148261905
Purchase Store Count: (1 is App): 0.2975897036171397
Auto Renew Counter(1 is on): 0.00970814385433608
Email Subscriber Count (1 is Yes): 0.001721684179105642
Push Notif Count (1 is Yes): 0.0012142548035767062
Permutation Importance of Subscription Type Counter(1 is limited): 1.0237539318913624
Permutation Importance of Counter Subscrition Event Type(1 in initial_purchase): 0.04568419222451206
Permutation Importance of Purchase Store Count: (1 is App): 0.6268817847864183
Permutation Importance of Auto Renew Counter(1 is on): 0.01596480728749714
Permutation Importance of Email Subscriber Count (1 is Yes): 0.0030460734721314718
Permutation Importance of Push Notif Count (1 is Yes): 0.0025409530644358875
```

    b. 

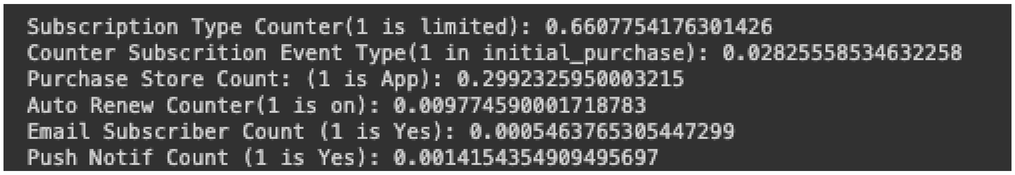
```
Mean Squared Error with Random Forest: 1405.6119733381672
R-squared with Random Forest: 0.7311039217400068
```

    c. 

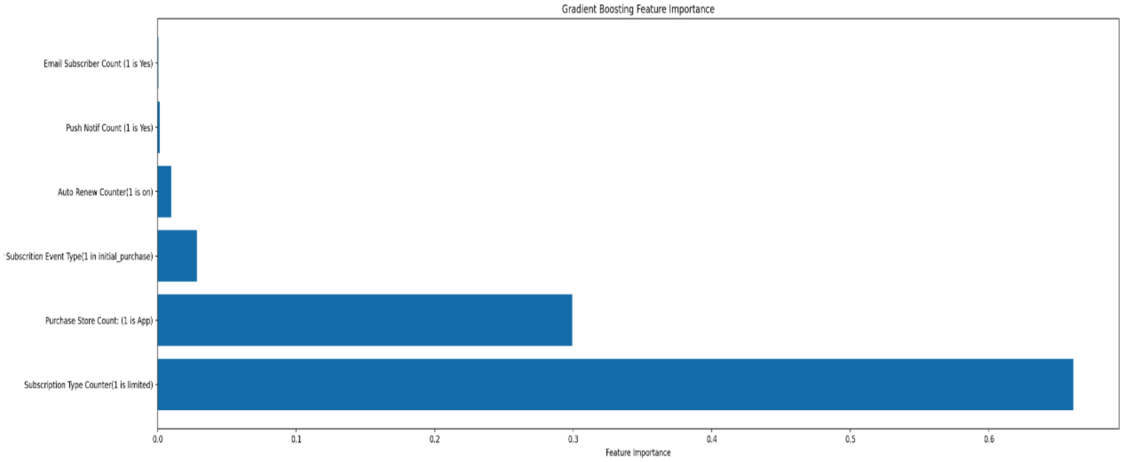
3. Gradient Boosting Regression: We attempted to create a gradient boosting regression model that ensemble machine learning algorithms to create a strong predictive model. We used feature variables subscription type counter (1 is limited), counter subscription event type(1 in intiial_purchase), purchase store count (1 is App), Auto Renew Counter (1 is on), Email Subscriber Count (1 is Yes), and Push Notif Count (1 is Yes). The target variable was again, Purchase Amount USD. the variables were split into training and testing sets, normalized, and

z-scored. The data was then transformed using scaler for consistency and initialized and fitted to the Gradient Boosting Regressor. The model then predicted on the test set and was evaluated using mean squared error. The model also produced feature importance, but returned results nearly identical as the randomforests model. The MSE score was nearly identical as well, leading to the decision to use the randomforests model's data more heavily than the gradient boosting regression model.

a.



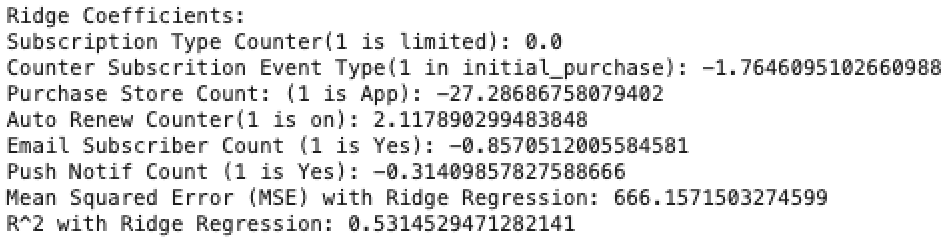
```
Subscription Type Counter(1 is limited): 0.6607754176301426
Counter Subscrition Event Type(1 in initial_purchase): 0.02825558534632258
Purchase Store Count: (1 is App): 0.2992325950003215
Auto Renew Counter(1 is on): 0.009774590001718783
Email Subscriber Count (1 is Yes): 0.0005463765305447299
Push Notif Count (1 is Yes): 0.0014154354909495697
```



b.

4. Ridge Regression: We created a L2 ridge regression model that allowed us to account for overfitting The feature variables were Subscription Type Counter (1 is limited), Counter Subscription Event Type (1 in initial_purchase), Purchase Store Count: (1 is App), Auto Renew Counter (1 is on), Email Subscriber Count (1 is Yes), and Push Notif Count (1 is Yes). The target variable was the same as the rest of the regression models, Purchase Amount USD. THe features were then normalized and split into training and testing sets. Then, the data was initialized and fitted to the Ridge Regression Model. THen model then returned ridge coefficients, predicted on the test set and predicted MSE and R^2 values for analysis. A coefficient of 0.0 for subscription type counter indicates that this feature does not contribute to the prediction. Negative coefficients, such as **Counter Subscription Event Type, Purchase Store Count, Email Subscriber Count, and Push Notif Count,** suggest an inverse relationship where as the feature increases, Purchase Amount USD decreases. Positive coefficients (Auto Renew Counter) suggest a positive relationship as more customers with auto renew on will raise the Purchase Amount USD. The coefficients align with the clustering models as well as high value clusters indicated that the bolded features reversed will target higher value customers, which is discussed more in beginning of the data analysis section and the presentation.

a.



```
Ridge Coefficients:
Subscription Type Counter(1 is limited): 0.0
Counter Subscrition Event Type(1 in initial_purchase): -1.7646095102660988
Purchase Store Count: (1 is App): -27.28686758079402
Auto Renew Counter(1 is on): 2.117890299483848
Email Subscriber Count (1 is Yes): -0.8570512005584581
Push Notif Count (1 is Yes): -0.31409857827588666
Mean Squared Error (MSE) with Ridge Regression: 666.1571503274599
R^2 with Ridge Regression: 0.5314529471282141
```

Future Business Opportunities

Customizing marketing strategies is pivotal for sustained business growth, when targeting diverse customer segments based on their spending patterns—this can be categorized as low, medium, and high spenders like in our analysis.. This approach involves the delivery of personalized services and products crafted to align with the distinct preferences of each group. By focusing on specific demographics and prioritizing high spenders, the Rosetta Stone can not only optimize sales but also position themselves strategically for global expansion beyond the US market.

To fortify customer relationships and drive higher sales conversions, a critical emphasis should be placed on refining email campaigns based on user engagement levels. Concurrently, a deep dive into user engagement with product features and content is imperative for informed product development. By identifying and accentuating features that resonate with users, businesses can elevate customer satisfaction and foster loyalty. This comprehensive strategy aligns with evolving customer expectations and ensures a sustained competitive edge.

In the pursuit of customer retention, a proactive approach involves understanding the reasons behind user disengagement or subscription cancellations. Implementing predictive analysis techniques enables companies to gain insights into potential cancellations based on user activity. This strategic proactive approach is rooted in customer-centricity, leveraging user data to align more closely with customer expectations and preferences. While following these changes and adapting to these strategies, profits and subscriber-base should expect major rises.

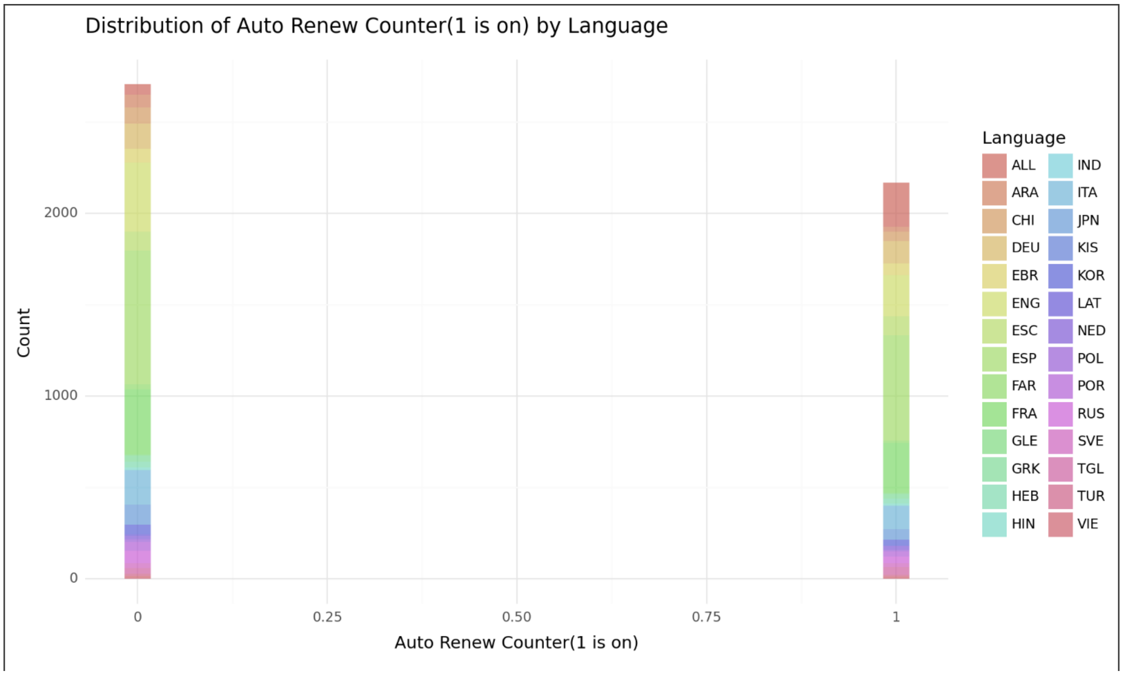
Division of Work

King Nguyen: churn prediction model, churn prediction model section, data cleaning, feature dissection, feature adjustments, Powerpoint slides (19-21), future business opportunities section
Ryan Redjaian: data cleaning, feature adjustments, feature dissection, Powerpoint, Executive Summary (2-4 + 20), presentation
Luis Rivas: clustering/Kmeans analysis, clustering models section, visualizations, Powerpoint slides (5-9), questions 1 and 2
Etai Eilat: data cleaning, feature adjustments, feature dissection, Powerpoint, business opportunities, chart creation, presentation
Paul Zhao: all regression models, regression models section, histograms (below), barplots (below), scatterplots (below), EDA, Powerpoint slides (10-18), analytical plan, overview, business goals section, future business opportunities section
Todd Hartog: Project doc, editing all powerpoint slides, presenting slides, business goals and business opportunities sections, assisting with organizing division of work and planning

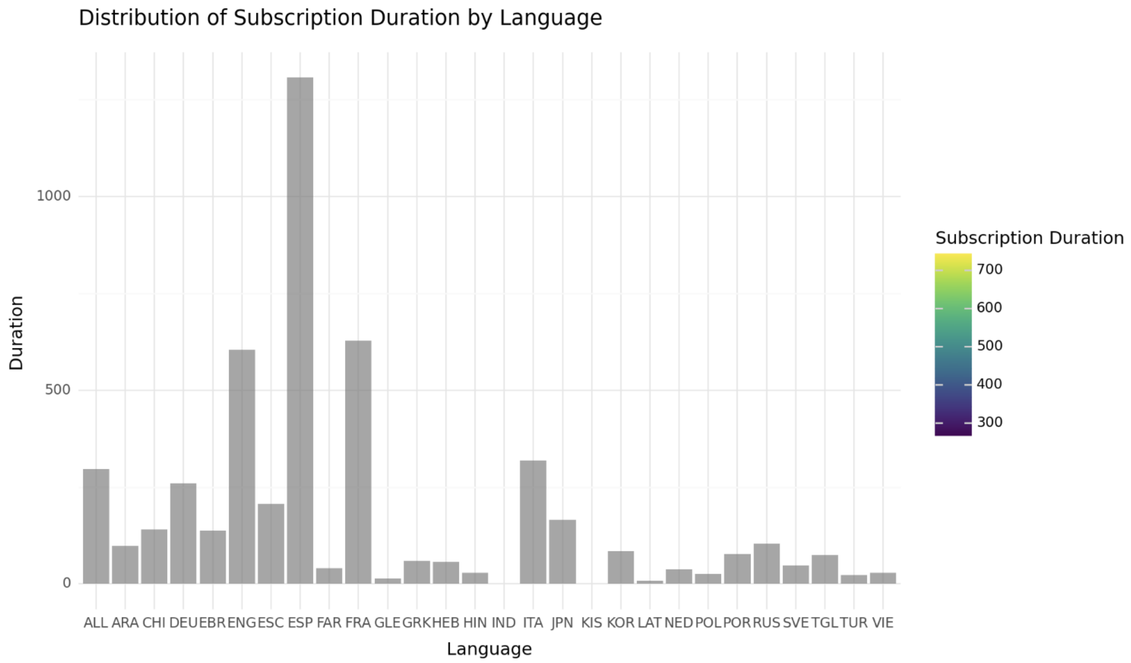
Appendix (unused visualizations)

https://colab.research.google.com/drive/1RPdPLGGByvgKZ8j9eBvppfUm5AOzNZ31?usp=sharing
(has many more histograms, boxplots, and scatterplots of distributions and models)

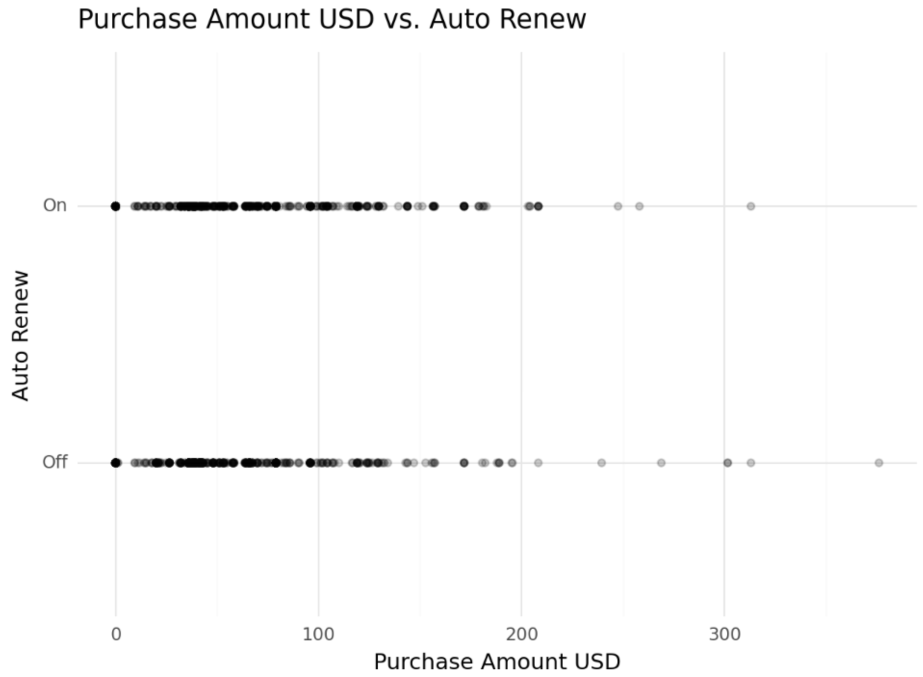
Appendix 1: (boxplot & histogram version in google collab link under title)



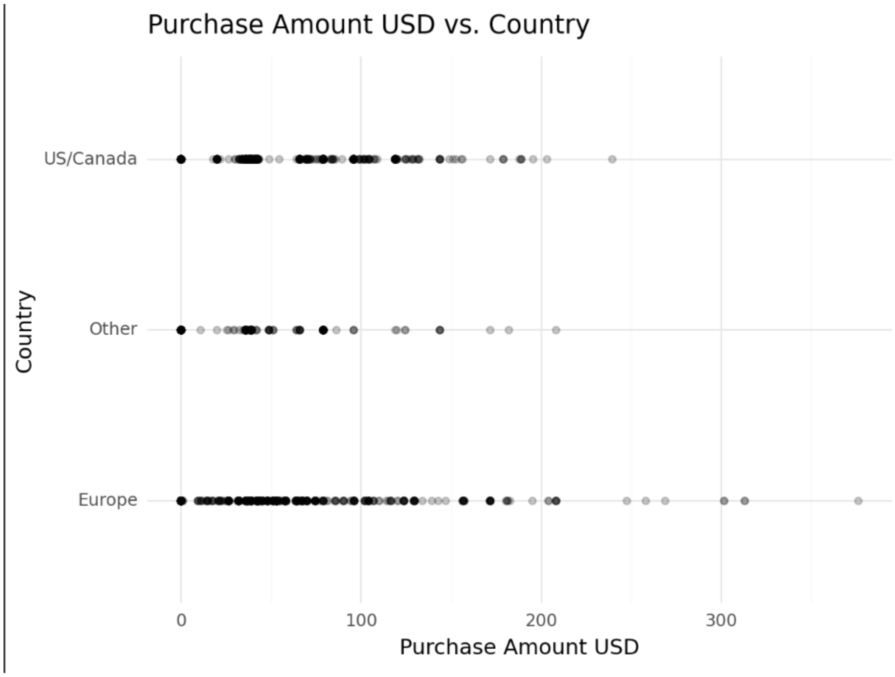
Appendix 2: (boxplot & histogram version in google collab link under title)



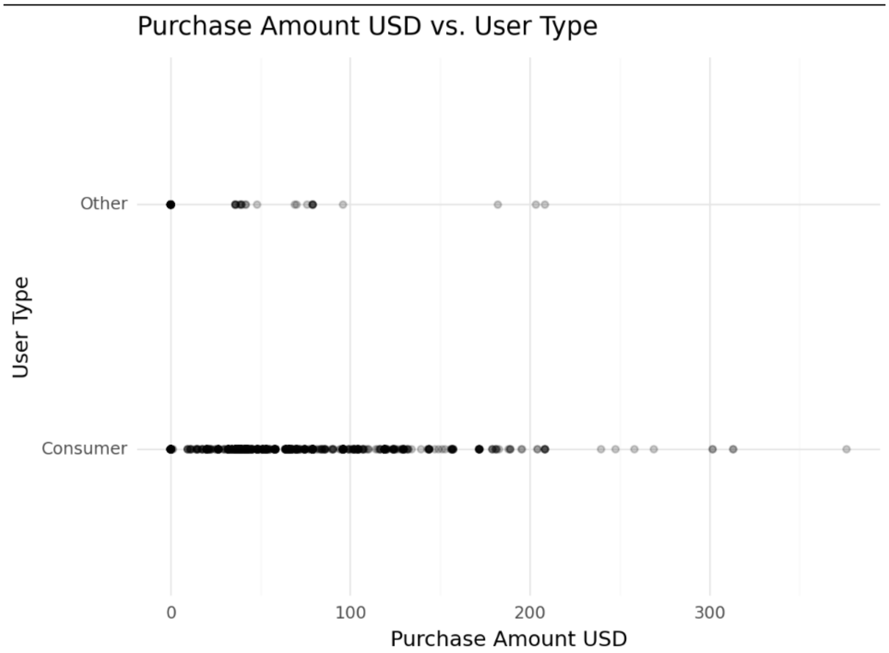
Appendix 3: (boxplot & histogram version in google collab link under title)



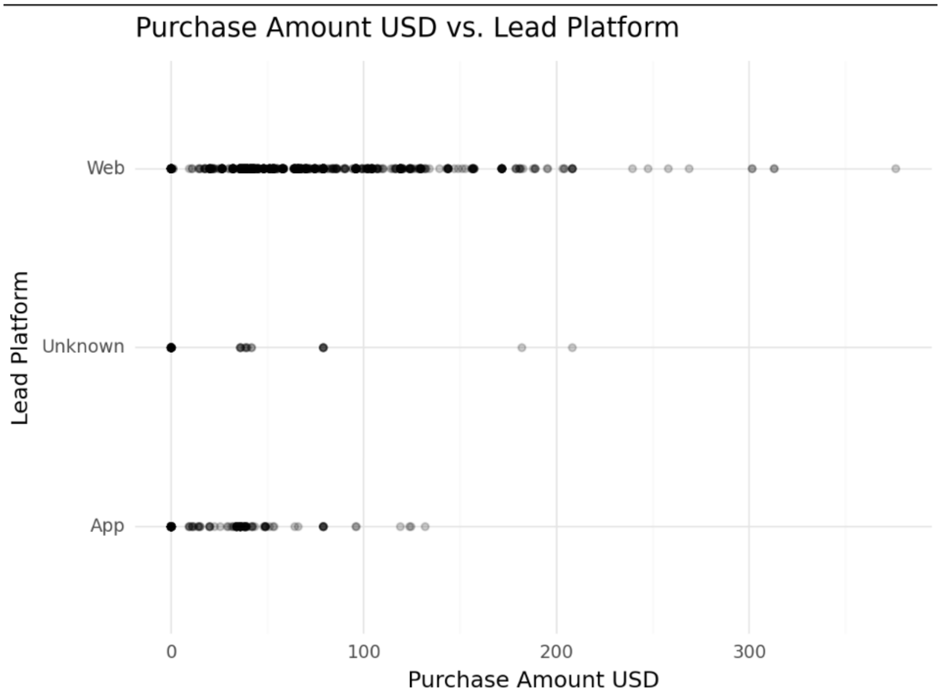
Appendix 4: (boxplot & histogram version in google collab link under title)



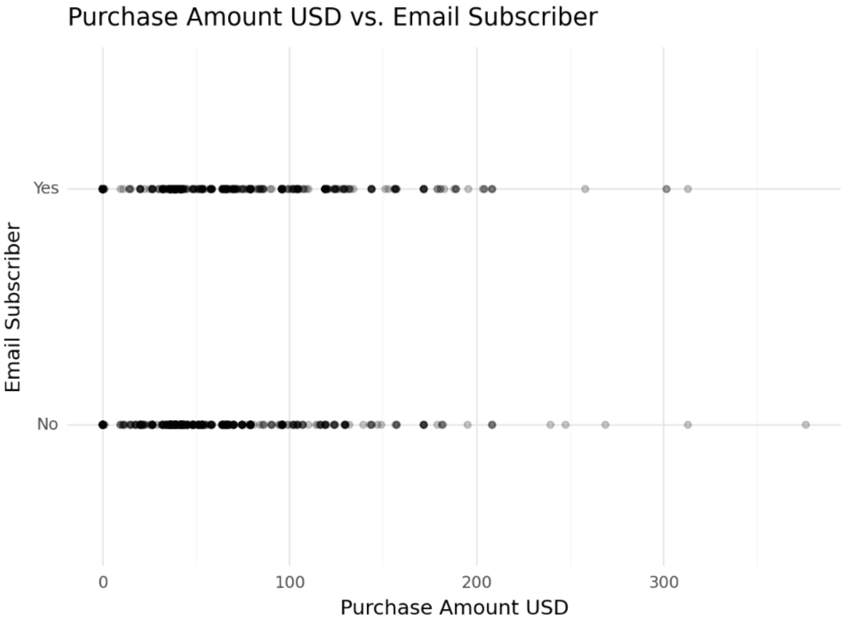
Appendix 5: (boxplot & histogram version in google collab link under title)



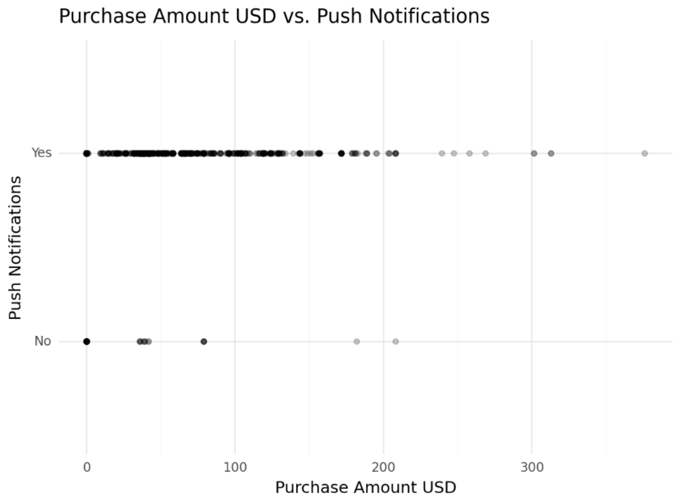
Appendix 6: (boxplot & histogram version in google collab link under title)



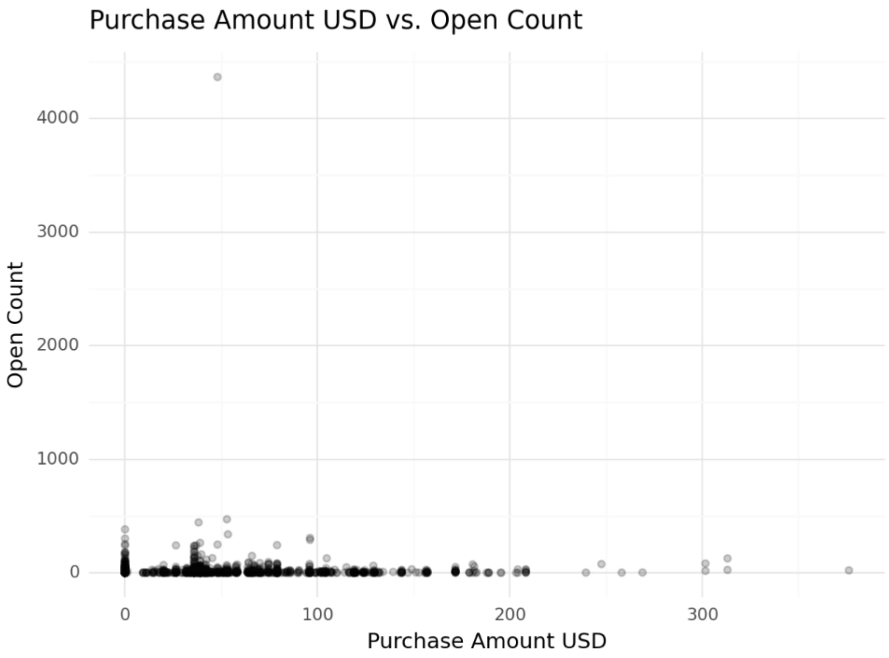
Appendix 7: (boxplot & histogram version in google collab link under title)



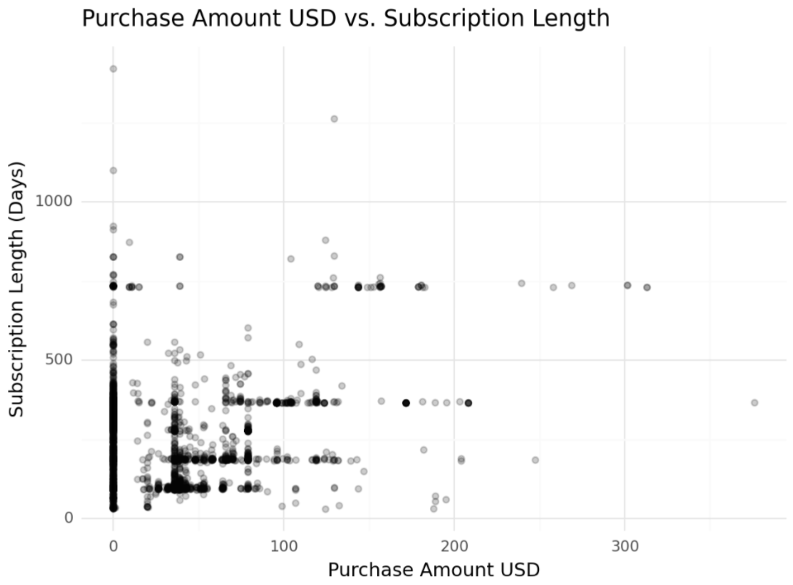
Appendix 8: (boxplot & histogram version in google collab link under title)



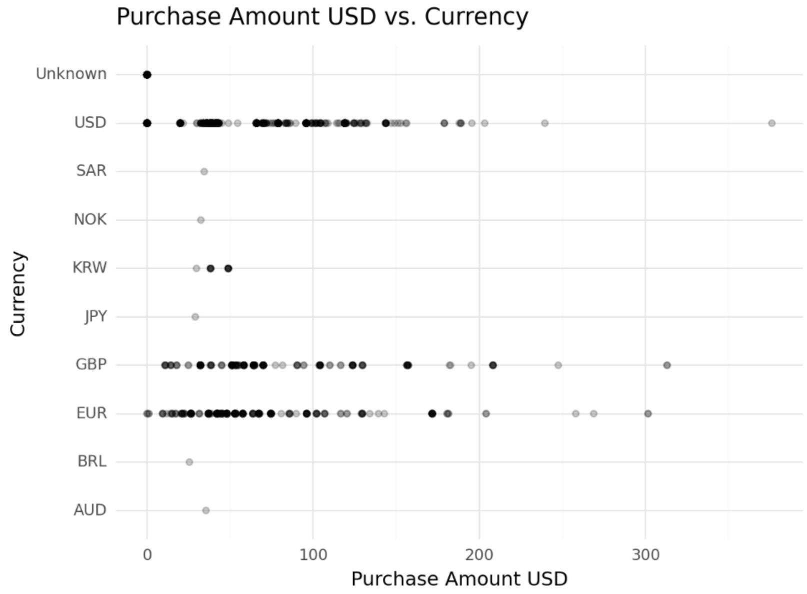
Appendix 9: (boxplot & histogram version in google collab link under title)



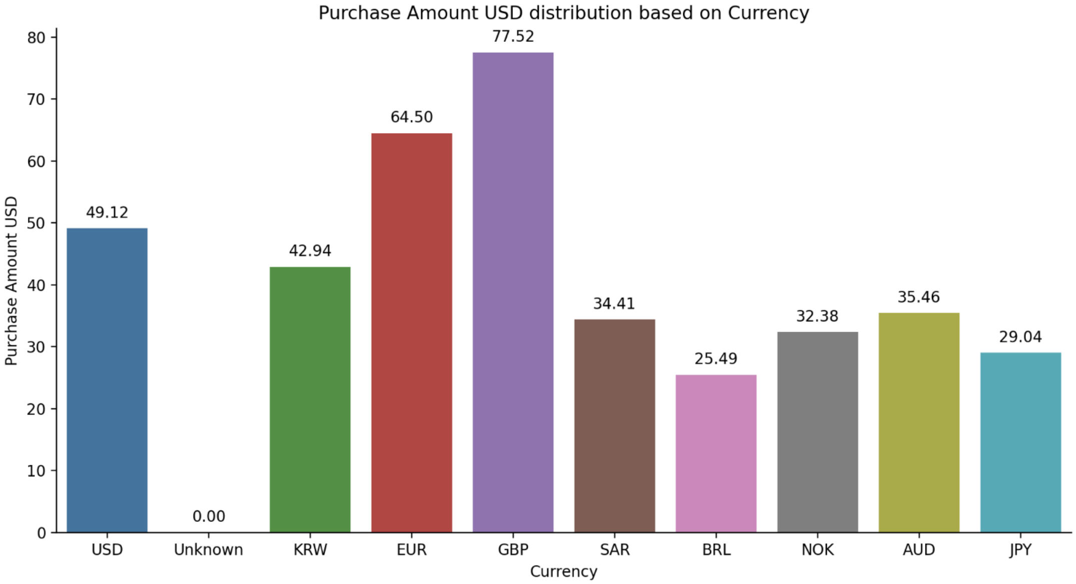
Appendix 10: (boxplot & histogram version in google collab link under title)



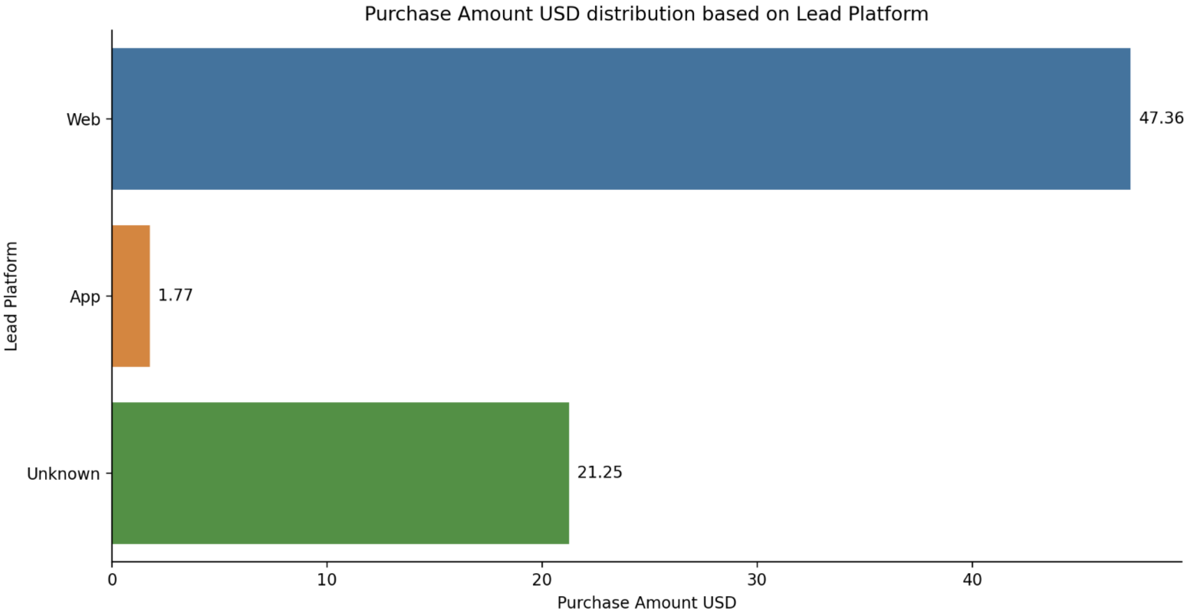
Appendix 11: (boxplot & histogram version in google collab link under title)



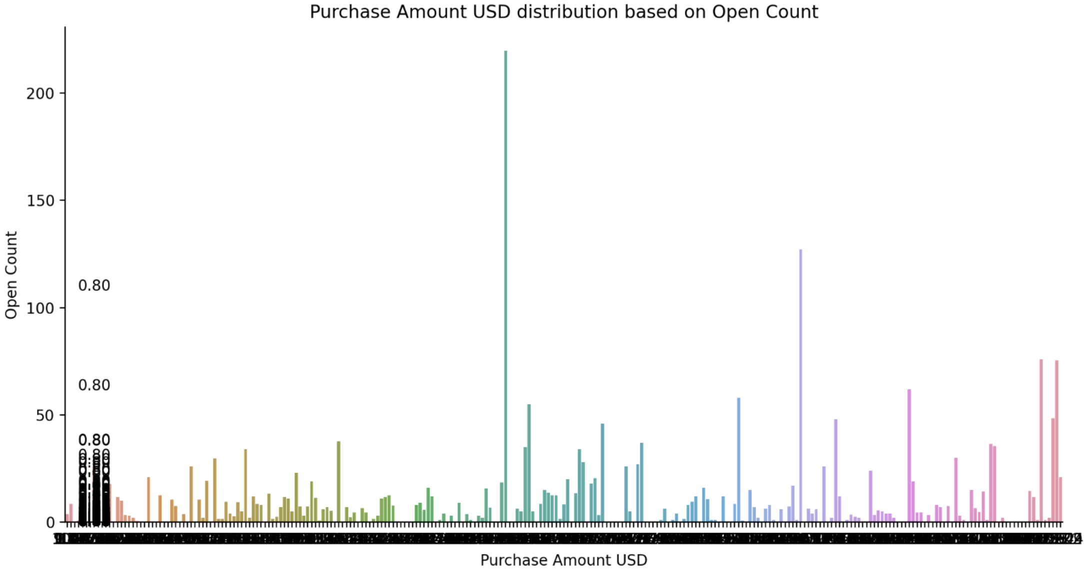
Appendix 12: (boxplot & histogram version in google collab link under title)



Appendix 13: (boxplot & histogram version in google collab link under title)



Appendix 14: (boxplot & histogram version in google collab link under title)


Purchase Amount USD distribution based on Open Count
Open Count
Purchase Amount USD