Joyce Shi
Retention Analysis Synthesis

## Discussion of the Dataset

**Exploratory Data Analysis:** A quick examination of the dataset reveals data type inconsistencies (e.g. some "date" columns are dates while others are timestamps) and value inconsistencies (e.g. country of origin values are not standardized). To address this, we first check for unexpected values in our boolean columns and non-datetime-convertible values in our date columns. We also find one duplicate data entry which, upon further inspection of missing values in each column, turns out to be a result of there being 2 entirely empty data entries.

**Data Cleaning:** To fix the issues revealed in our EDA, we standardize all date columns to datetime format, standardize country names by mapping all alternatives to a unified name, and drop any empty rows. This leaves us with missing values in only the conversion_date, cancellation_date, and personal_person_geo_country columns:

- **conversion_date:** We check whether the customer converted to a paid plan at all and find that all customers missing conversion dates simply never converted. We also confirm that only customers who converted have conversion dates. We leave NaT in the empty entries for now.
- **cancellation_date:** We check whether those with a cancellation date have a plan status of "canceled" (i.e. is_canceled==True) and those without a cancellation date do not. We find that there are 87 customers for whom this does not hold—these may indicate resurrected customers or canceled ongoing plans that will still be active until a certain date, depending on how the data was recorded.
- **personal_person_geo_country:** We use "unspecified" as a filler for missing country data.

With these initial changes, we have a cleaned dataset for our retention analysis.

## Discussion of Retention Characteristics

We can conduct our retention analysis along 4 general areas of interest: customer characteristics, customer behavior, churn/resurrection, and revenue trends.

**Segmented Analysis:** We can segment customers by cohort, geography, and provider. In doing so, we find the following trends:

- Cohorts: Defining cohorts by signup date, we find a pattern that roughly resembles a shallow U (see Fig. 1). New cohorts at the monthly level have the highest rates of retention (which is expected since we are measuring active users relative to a reference date of January 16, 2023). At the weekly level, newer (but not the most recent) cohorts have the highest rates of retention, indicating that the most recent cohorts may be experiencing low conversion rates. Older cohorts have moderate rates of retention, which is also to be expected as there are likely to be early adopters who are particularly enthusiastic about the product or some users returning to the product after initially churning. There is a clear dip in retention for cohorts that signed up in late 2021 and early 2022—this trend is worth digging into to understand whether it is a function of the product, seasonality, or other factors.

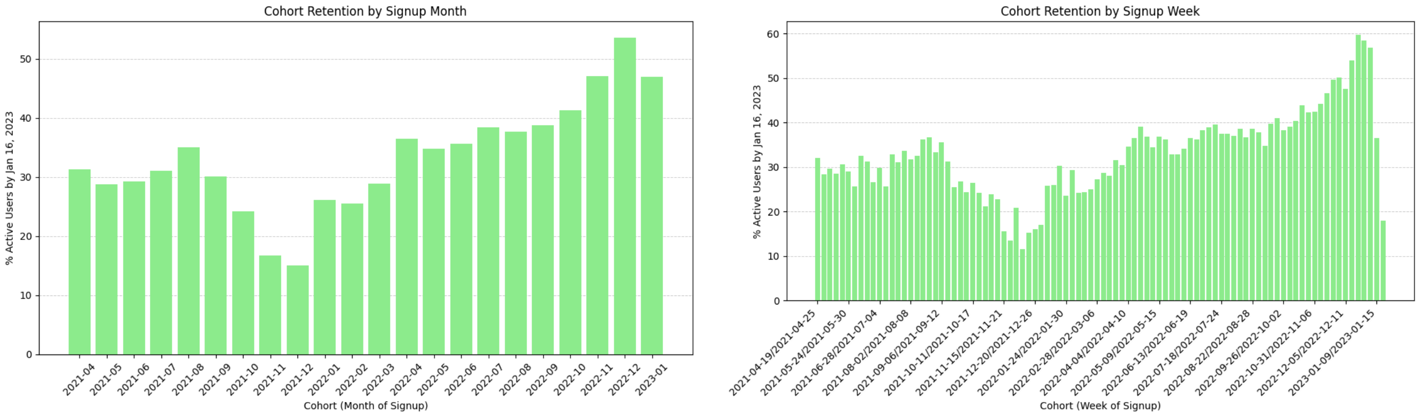

*Figure 1. Cohort Retention by Signup Time*

The dip in retention rates is muted when considering cohort retention filtered to only the customers that have converted to a paid plan (see Fig. 2), implying that the lower retention rates are impacted somewhat significantly by customers who never convert in the first place. It should be a priority for the company to figure out its “magical moment” that pushes customers to convert.

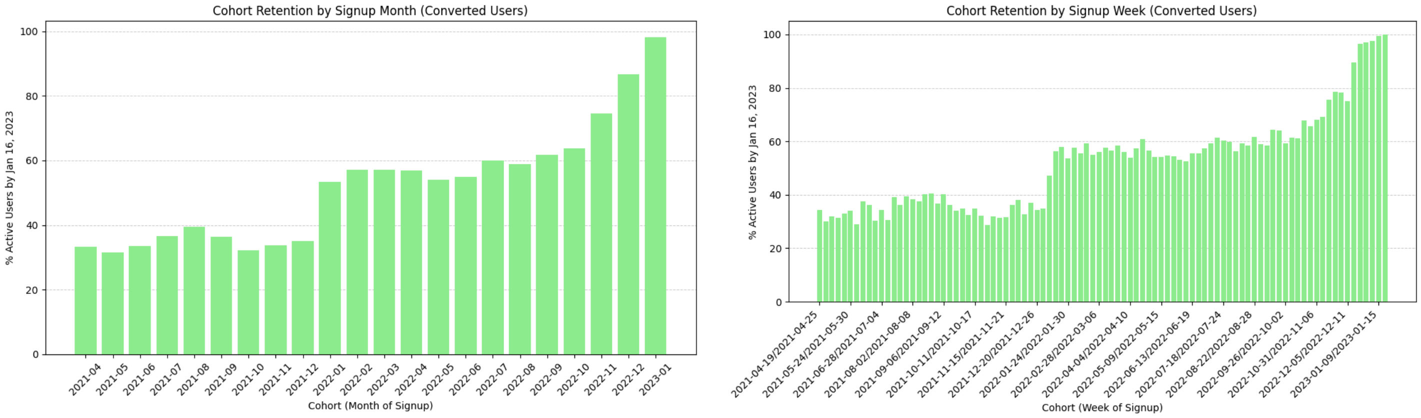

*Figure 2. Cohort Retention by Signup Time (Converted Users Only)*

- Geography: Naive analysis of retention by country appears to show clear differences in retention, with certain countries having incredibly loyal customers (see Fig. 3). This is misleading, however, since we have imbalanced sample sizes across countries. When we compare the number of total users signed up and the number of customers still active on January 16, 2023, the retention rate results are heavily biased for countries that only have a few users.

  To address this imbalance (to some extent), we can group countries into larger regions (roughly by continent) to get a better sense of any actual geographical differences. The results show that certain regions (e.g. North America) have notably higher retention rates than others (e.g. Eurasia), which may indicate that the product is better tailored to certain regional needs than others (see Fig. 4). Product feature diversification may be an important forward strategy for regional expansion. Note, however, that the United States has the most customers at 19,248, with the second highest number of customers being in the United Kingdom at 3,829.

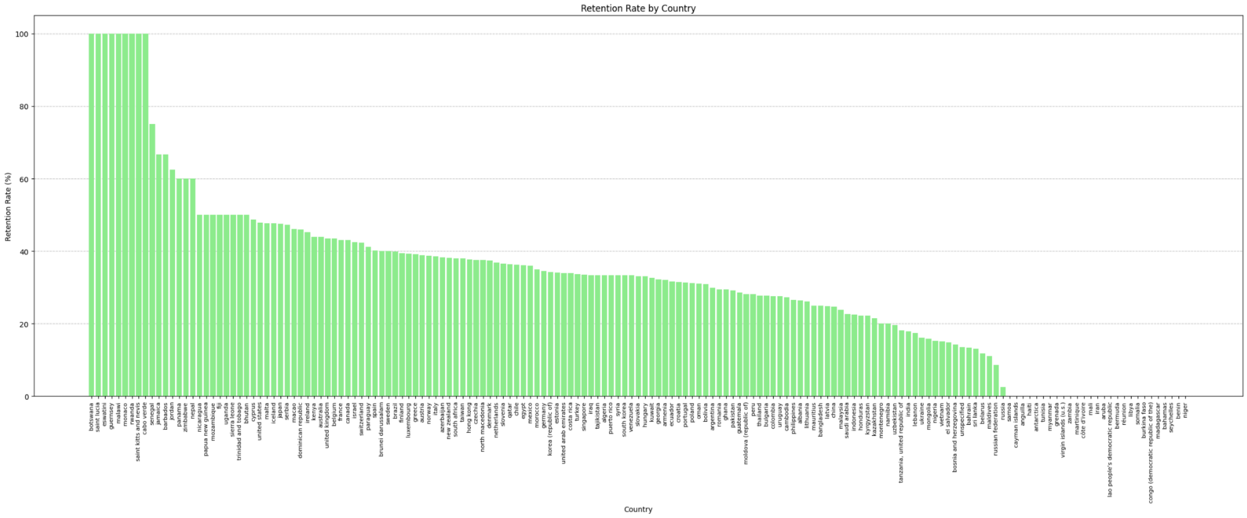

*Figure 3. Retention Rate by Country*

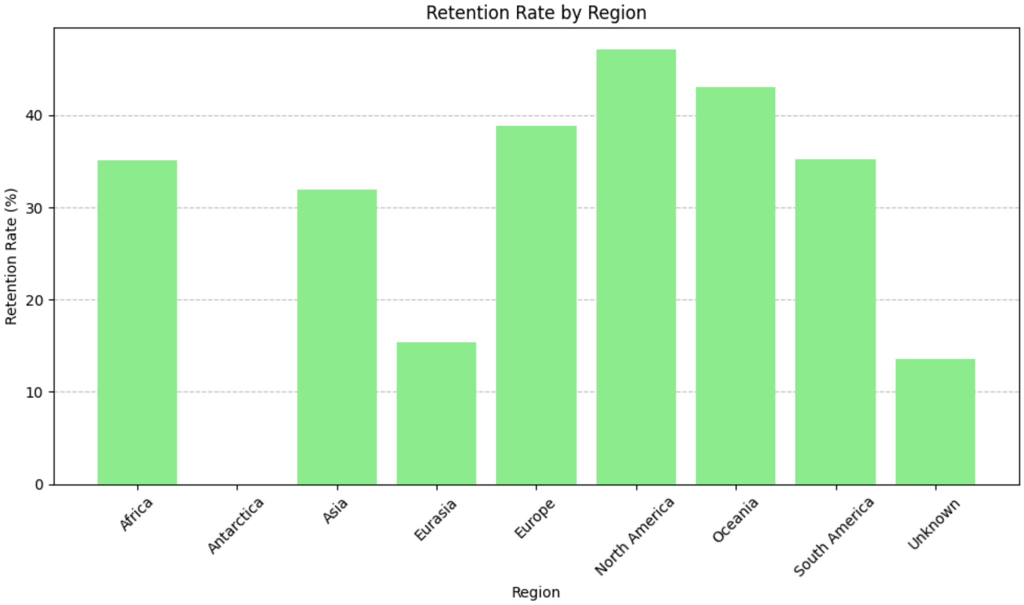

*Figure 4. Retention Rate by Region*

For the top 5 countries with the most customers, we see roughly similar retention trends by signup cohort (see Fig. 5).

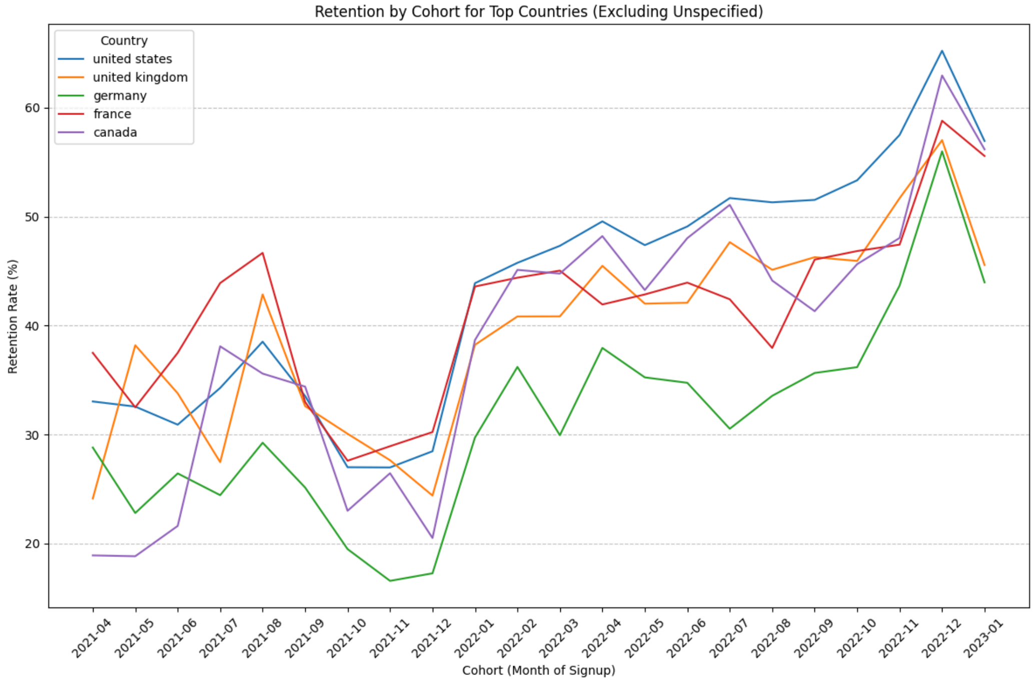

*Figure 5. Retention Rate by Cohort for Top Countries*

- Provider: Some providers are more popular than others among customers. For instance, Apple is used by 98,026 total users (39,906 of whom remain active) while Google is only used by 4,476 total users (2,326 of whom remain active). Though Stripe is used by a sizeable number of customers (32,108), it has the lowest retention rate at only 15.1%, with Apple having retention of 40.7% and Google having retention of 52.0% (see Fig. 6). This difference in retention by provider is interesting — it may be correlated with regional differences for provider preference, or providers may have tangibly different levels of seamless integration with the subscription business, making the experience better for some customers on a provider-dependent basis.

  Looking at retention by cohort across all three providers, we can note the noticeable rise in retention for Google payment users, while Stripe user retention has dropped significantly from the end of 2022 to January 2023 (see Fig. 7). We can also see that the three providers were perhaps introduced as payment platform options at staggered dates, corresponding to the differences in cumulative customer usage across the three (e.g. Apple was integrated first and acted as the only provider for 7 months, which likely explains its dominance of customer share).

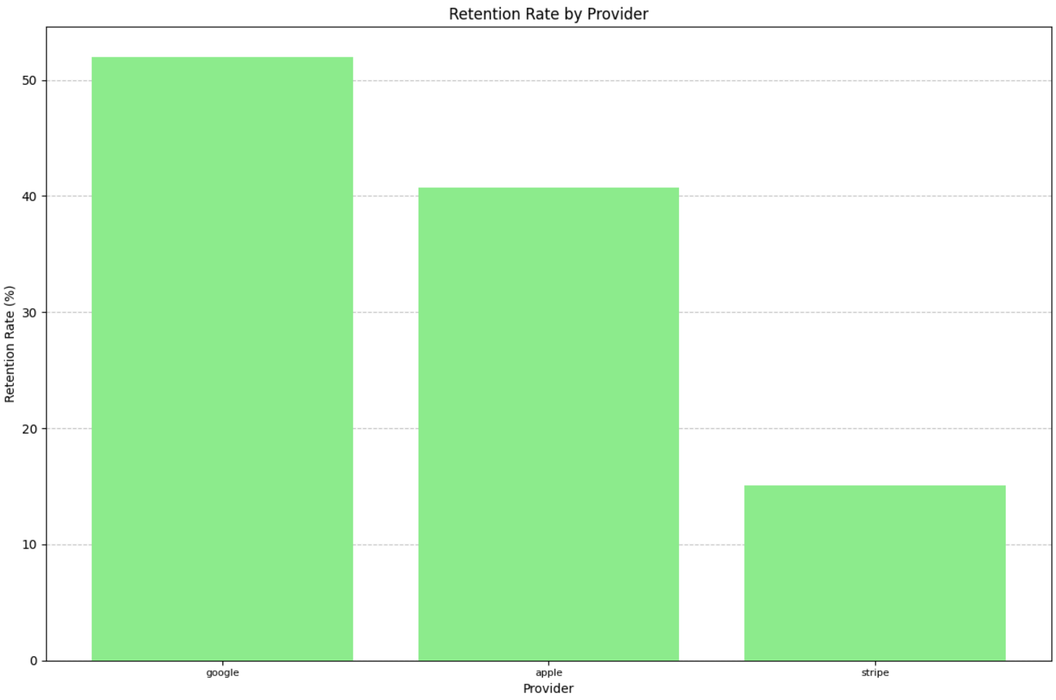

*Figure 6. Retention Rate by Payment Provider*

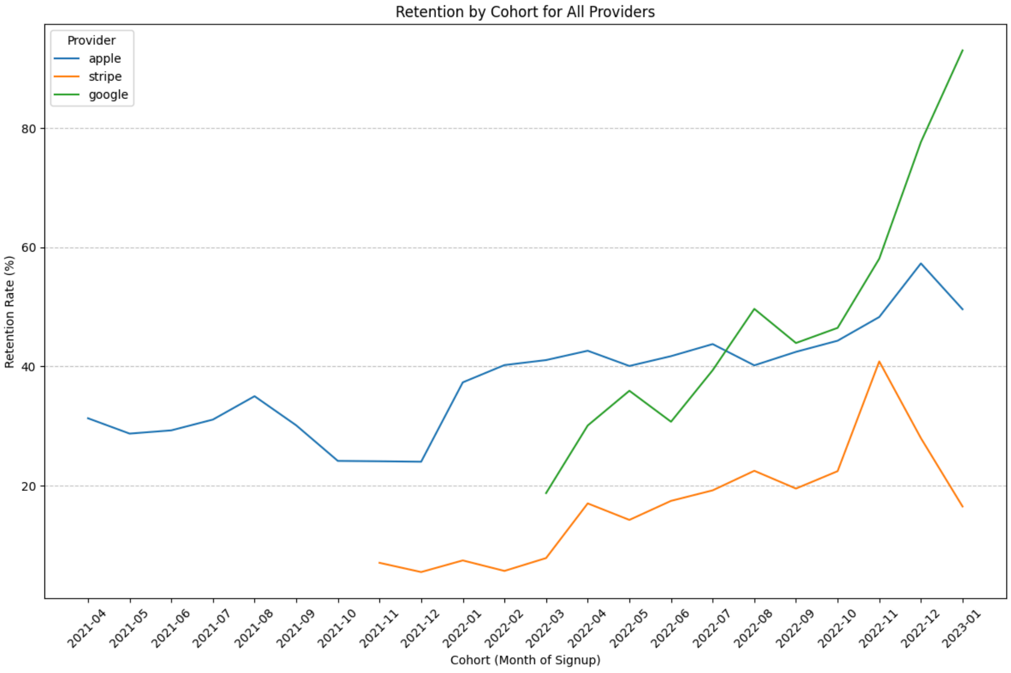

*Figure 7. Retention Rate by Cohort for Providers*

**Customer Behavior Analysis:** We also want to analyze lifetime value (LTV), conversion rate, cancellation rate, and delinquency rate information.

- Lifetime Value: The average LTV of a customer is $24.86. This jumps to $41.26 if we are only considering customers who converted to paid plans. Examining average LTV by geography and provider, we find trends that match our expectations given the retention rate patterns for these segments (see Fig. 8), which suggests that there exists little/no pricing differentiation by geography/provider.

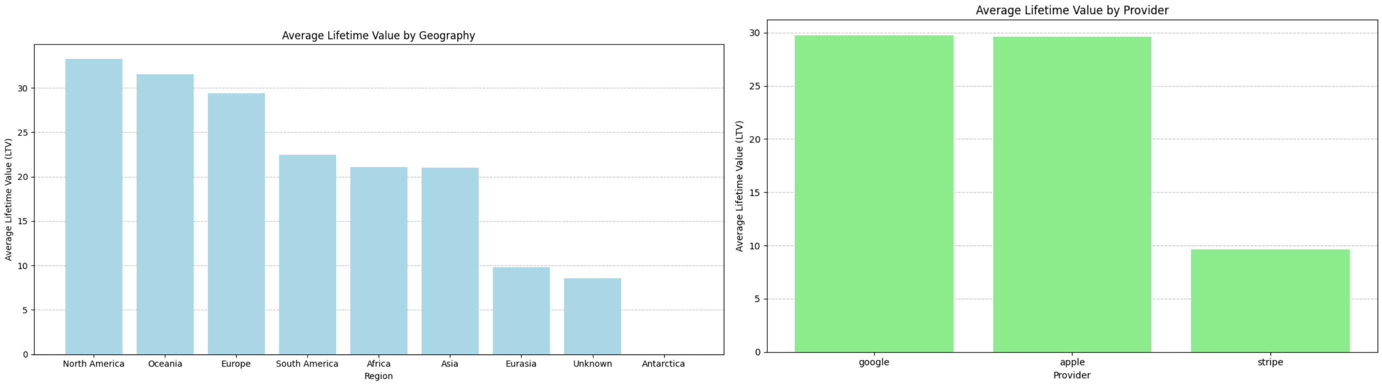

*Figure 8. Average LTV by Geography and Provider*

- Conversion: Tracking the total conversion rate to paid plans over time, we see that there is a steep drop in conversion, after which the overall rate remains low (see Fig. 9). Since this is a cumulative measure of (converted users)/(total signed up users), the drop in conversion indicates a rise in signups where conversion did not keep pace. This sudden decline in conversion should be analyzed further to identify potential reasons (for instance, there could have been a period of time where the product was not being delivered correctly, pushing many customers away from conversion).

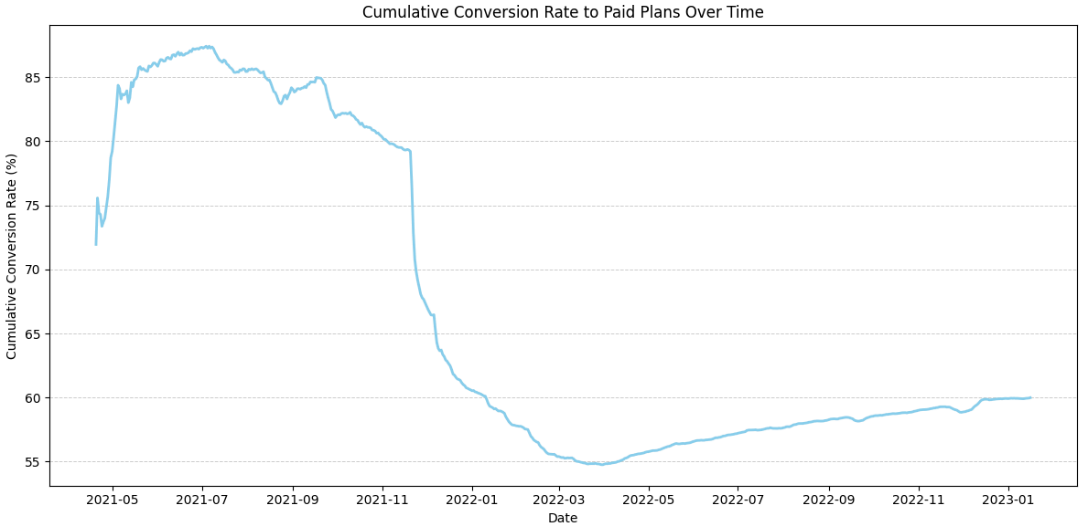

*Figure 9. Cumulative Conversion Rate to Paid Plans Over Time*

We also find that, on average, customers who convert to paid plans do so after ~5 days (see Fig. 10). The first 5 days of product use may thus be a critical period, and it may be important for the company to focus on engagement during this time.

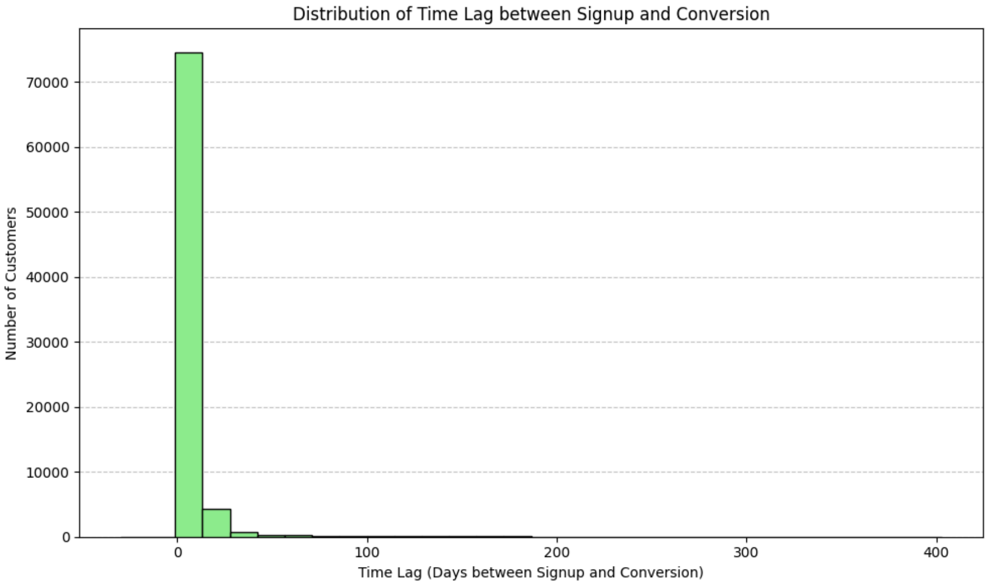

*Figure 10. Time from Signup to Conversion (for Converted Users)*

- Cancellation: Similarly, tracking the cancellation of paid plans, we find that, on average, customers who convert to paid plans and ultimately cancel them spend ~115 days using the subscription before canceling (see Fig. 11). The right-skewed nature of the cancellation distribution, however, indicates that half of these customers cancel by the 61-day mark (with another peak at ~1 year of paid usage). Understanding why customers cancel is crucial, and improving the subscription experience in the first two months may be critical for long-term retention.

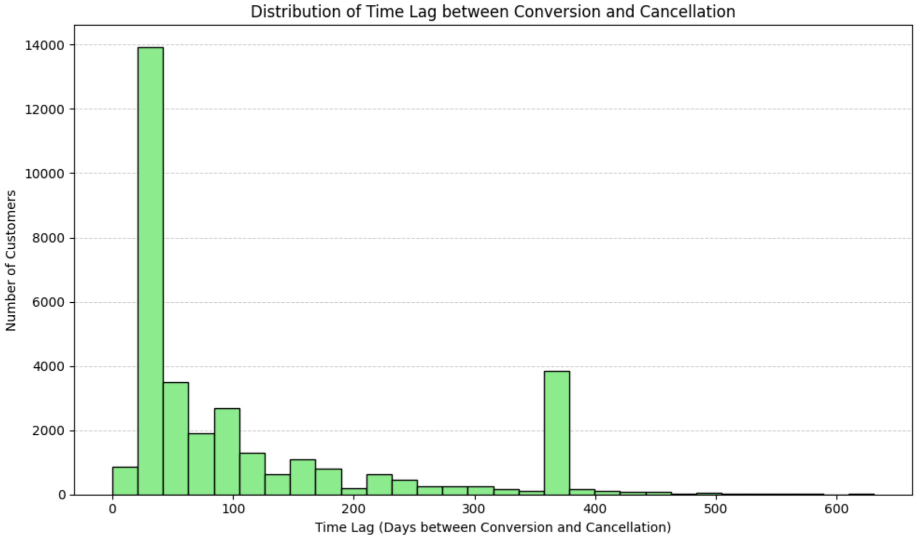

*Figure 11. Time from Conversion to Cancellation*

- Delinquency: Looking at delinquent customers, we find that they are a very small fraction of overall customers, with only 90 delinquent users in our dataset. None of these users have churned as of January 16, 2023, though 7 of them have non-null cancellation dates, indicating that they may have canceled at some point and then reactivated their accounts.

  Delinquency rates peak ~1 year after initial signup, which makes sense (customers probably convert to paid plans when they are able to pay on time, but their situations may change as time passes; see Fig. 12).

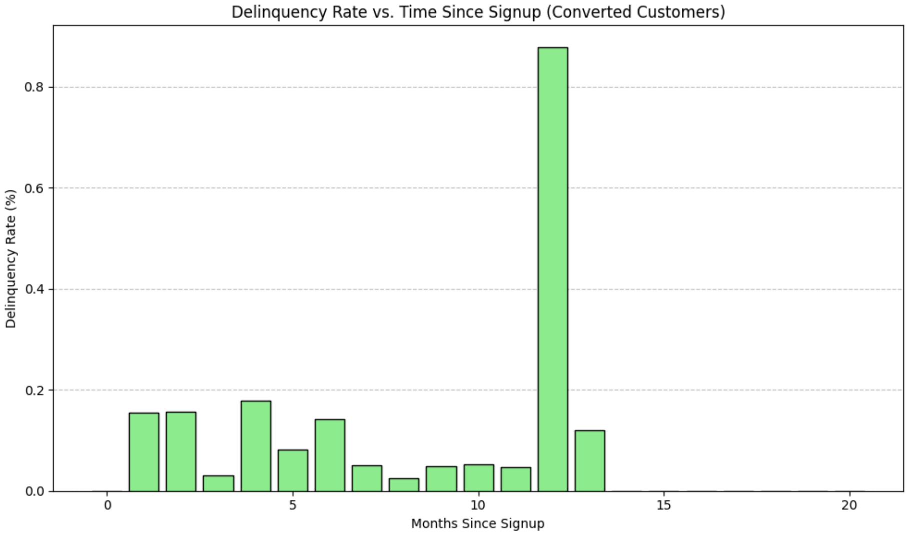

*Figure 12. Delinquency Rate vs. Months Since Signup for Converted Customers*

**Churn and Resurrection:** Churn and resurrection are key metrics for tracking company health and product-market fit. Although the dataset does not include reactivation date data, the fact that there are customers who have a valid cancellation date and an active paid plan at a point after that cancellation date implies that there are resurrected customers (who have reactivated their paid plans under the same customer ID). Under these assumptions, the product has a resurrection rate of 0.26% among users that churn.

Tracking cancellation dates alongside signups and conversions, we see seasonal trends (i.e. peaks near the end of each year) for signups, which trickle into a similar corresponding trend for conversions and, to a much lesser extent, cancellations (see Fig. 13).

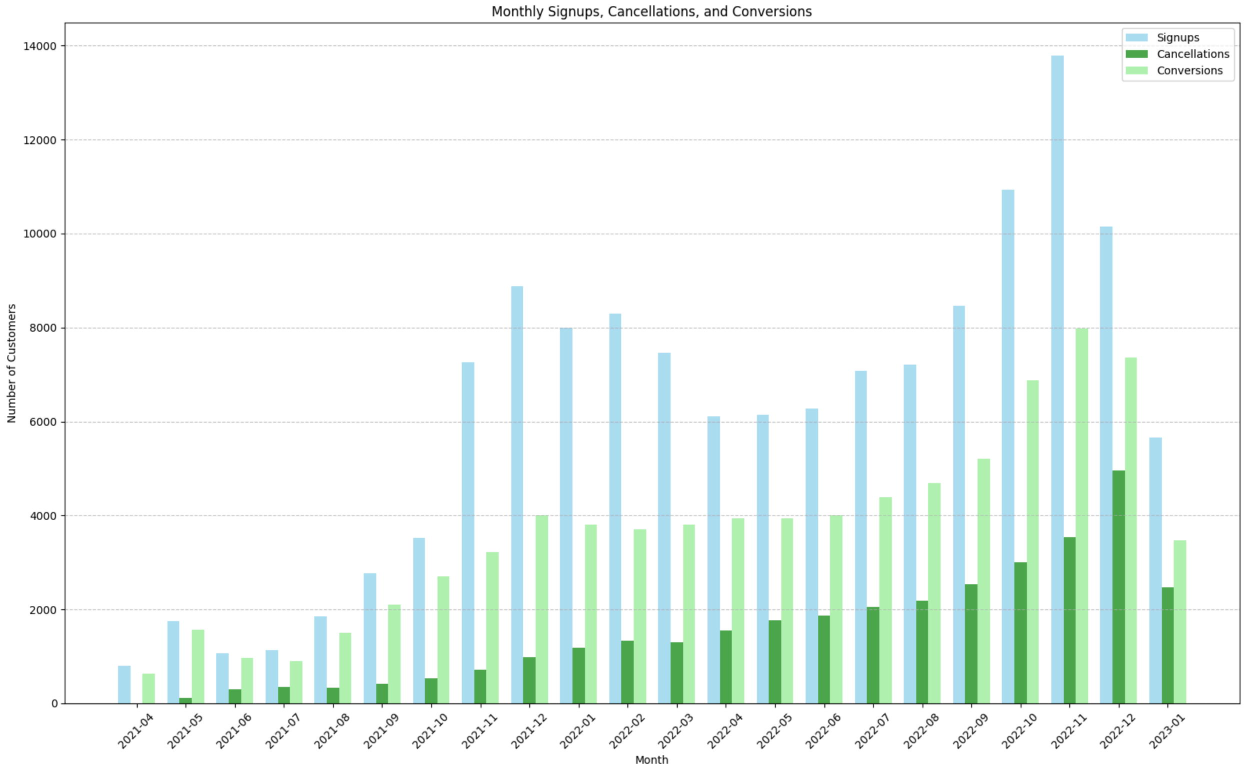

*Figure 13. Signups, Conversions, and Cancellations by Month*

**Revenue Analysis:** Beyond customer retention, we also want to understand revenue retention and customer spending trends. Since we only have current MRR, we can use (total charges) / (months since signup) as a (likely imperfect) proxy for customer MRR at conversion. We can then analyze the amount of current revenue generated by each signup cohort in comparison to the initial amount of revenue generated by that cohort when its customers first converted. Doing so reveals that revenue retention is higher than customer retention for all cohorts until the September 2022 cohort, after which revenue retention is below customer retention (see Fig. 14). For cohorts where the revenue retention rate exceeds the customer retention rate, retained customers contribute more revenue over time (perhaps due to upgraded subscription tiers, add-on purchases, price increases, etc.), so it makes sense that older cohorts would have undergone more of these pricing changes. The opposite is true for cohorts where the revenue retention rate falls below the customer retention rate. There is quite a bit of fluctuation in revenue retention, which could be worth examining further.

Based on these trends, we can infer that there is both revenue expansion and contraction, depending on the customer. Since we only have the current MRR for each customer, we can roughly track the evolution of MRR at the customer level by comparing the number of months a customer has been paying and their total-charge-to-current-MRR ratio (which shows the number of months they *should* have paid for if they were paying at their current MRR for their entire customer lifetime thus far). From this analysis, we see that 43,992 customers have experienced revenue expansion, 2,599 customers have experienced revenue contraction, and 478 customers have experienced no change in MRR. Thus, the vast majority of converted, currently-active customers have experienced revenue expansion—a potential indicator of successful upselling/cross-selling or price increases. See Fig. 15.

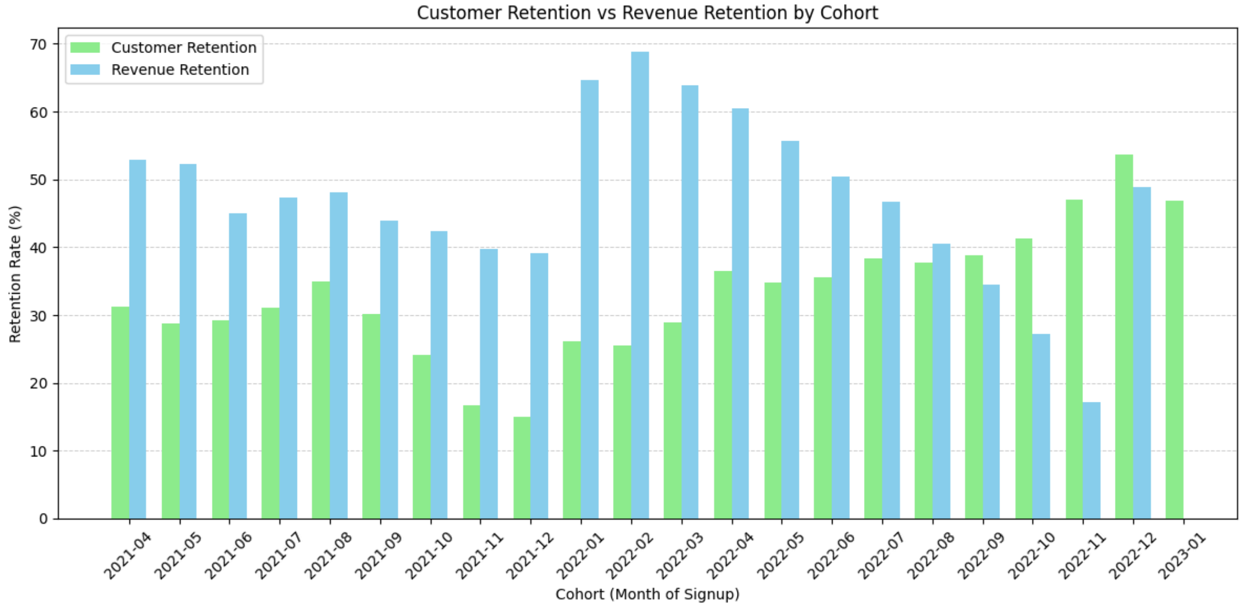

*Figure 14. Customer Retention vs. Revenue Retention by Cohort*

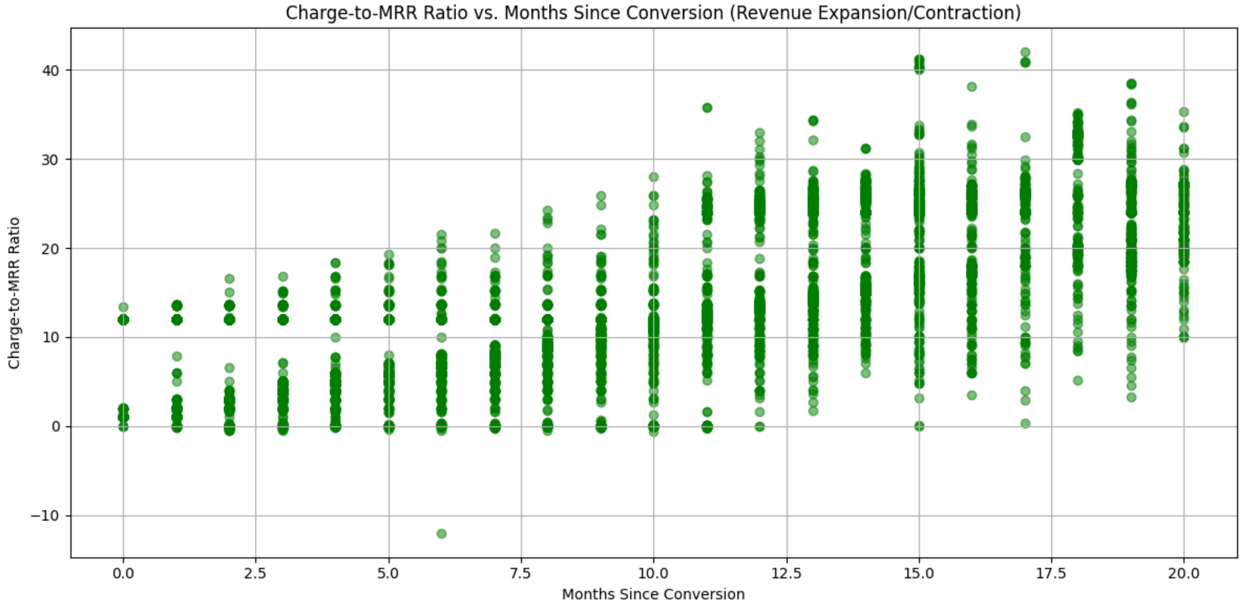

*Figure 15. Charge-to-MRR Ratio vs Months Since Conversion*

Finally, examining average current MRR across starting cohorts gives us some insight into how upselling/cross-selling or pricing might be more effective for some active users depending on when they first became engaged with the product. The overall average current MRR is $5.95 across all cohorts, with the January 2023 cohort having the highest average current MRR at $7.47 and the July 2021 cohort having the lowest average current MRR at $3.85. We have filtered for only converted, active customers to remove deflation from cancelled/unpaid accounts. Generally, newer cohorts have higher average current MRR, which may be due to a variety of reasons (e.g. discounted pricing for early users, special subscription packages for new users, etc.). See Fig. 16.

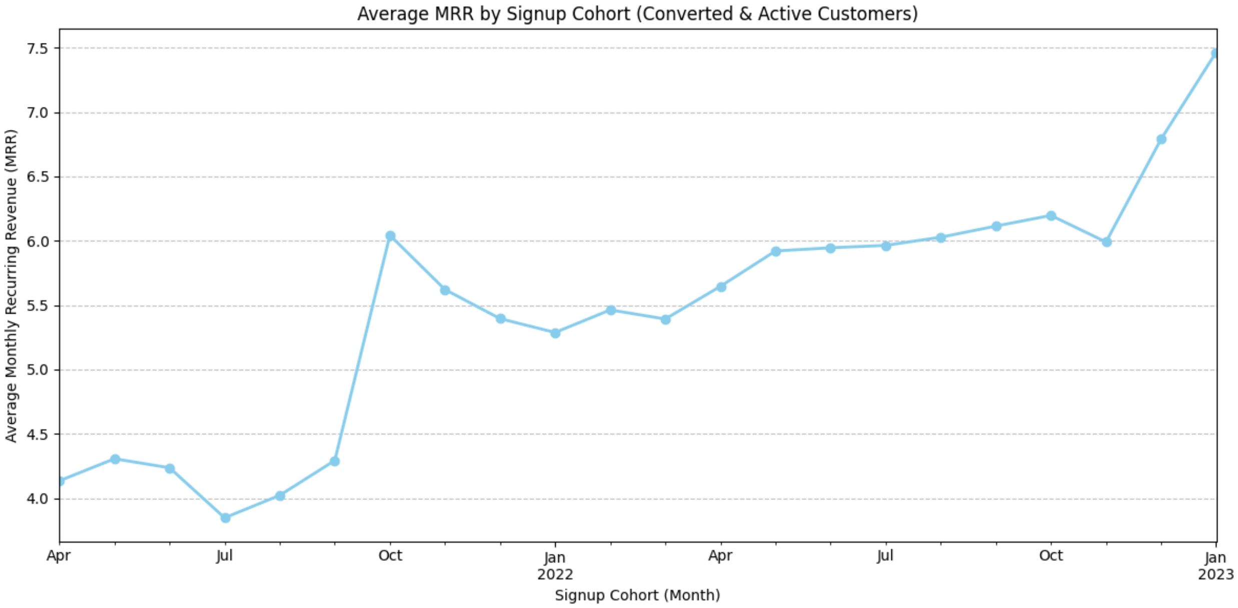

*Figure 16. Average Current MRR by Signup Cohort*

**Directions for Further Analysis**

This is a preliminary analysis of retention characteristics for this product. If we had more data (e.g. MRR data for each customer over time, reactivation dates), we could conduct a more thorough analysis of revenue retention, customer resurrection, and so on.