To: Professor Amini
From: Wanyi (Evelyn) Dong, Matt Andersen, Jenny Shang
Subject: Sales Report
Date: November 13, 2019

Introduction & Executive Summary

This report serves as a summary of our analysis of sales data from a subscription-based game. Our objective is to use the data to provide recommendations to achieve two main goals: optimizing user monetization and increasing engagement. After thoroughly analyzing the data, we propose the following recommendations:

1. Narrow the focus of the game's targeted marketing efforts in order to prioritize high-value users and increase marketing ROI (A-level).
2. Motivate users to log in to the game more frequently by implementing daily rewards or time-gated features, with the goal of extending average user subscription time (A-level).
3. Reallocate investments in user acquisition partnerships to prioritize high-performing partners, such as partners 2, 5, and 6, while minimizing investments in partners that do not outperform players generated without the use of partnerships (B-level).

Assumptions

We made several assumptions to better frame our analysis. First, we assumed that all users started with a free trial without a required payment method. Second, given the fact that we cannot predict the future lifetime value of an active user, our analysis regarding lifetime value only focused on canceled accounts that generated profits for the game. Finally, we assumed that active users with the same characteristics as canceled users would show similar behaviors.

Answers to Prompt Questions

From our data analysis, we determined the free-trial-to-sales conversion of 990 paying users out of 2122 total users, which translates to a conversion rate of 46.7% (we did not include users who refunded their entire purchase as paying users). The lifetime value we have from a paying customer as defined in our assumptions section is $51.02. The average retention duration of a paying customer is 163 days.

To determine the best customer type, we defined the "best customers" as customers who have the highest lifetime value, and we encoded the categorical variables to identify the characteristics of these "best customers." We found that the "best customers" are those who do not often request a refund, who have a higher level of education (i.e. bachelor or graduate level), who come to the game directly rather than through a partner, and who have a long retention duration. By contrast, we defined the "worst customers" as customers who have the lowest lifetime value. These customers often request a refund, have a lower education degree, come from some sort of partner channel, and have short retention durations.

To characterize users who are more likely to request a refund, we looked into the data of canceled accounts who had made a payment (because customers cannot ask for a refund unless they have paid). We elected to filter out active accounts because they still have the potential to

request refunds during their account lifetime. We then ran a decision tree model to find the most important features, such as age, gender, and number of logins, which help predict the likelihood that a user requests a refund. After exploring the distribution of refunds for each category, we found that older women who have a low number of total logins are the most likely demographic to request a refund.

Questions 7 and 8 are addressed in the "Recommendations" section of this report.

Exploratory Data Analysis and Modeling

To find characteristics of the "best customers," we built a linear regression model to find users with the highest lifetime value for all canceled customers. We excluded "Type" and "Last Refund Datetime" since most values in these two columns are missing (see Figure 3 in Appendix). We encoded all categorical data and assigned "10" to all missing values in the category "Partnership #." We also changed the date columns to ordinal data and defined two new features: "differencecancel," which is "DateTime Cancelled" minus "DateTime Created," and "differencelogin," which is "Last Login DateTime" minus "DateTime Created." We then performed initial feature selection by Pearson correlation heatmap to drop highly correlated features (see Figure 4 in Appendix), and performed backward elimination with a significance level of 0.05.

The final result contains five significant features (see Figure 5 in Appendix). There is a positive relationship between customer lifetime value and the following features: "Education_Bachelors Degree," "Education_Graduate Degree," "Partnership #_10.0 (null values)," and the difference between cancellation date and datetime created. As expected, there is also a negative relationship between lifetime value and "Refund Count." The adjusted R-squared is 0.83, which means that the model can explain 83% of the total variability in customer lifetime value. We suggest further investigations to see the meaning of null values in "Partnership #." If a null represents that the source is direct selling, the company should end relationships with low-value partners and allocate more budget to direct selling as users from these partners do not generate as much value.

To find characteristics of users who are more likely to ask for a refund, we focused on all canceled paying customers and ran a decision tree classification model (using 1 for "has ever asked for a refund" and 0 for "has never asked for a refund"), which produced an accuracy rate of 0.8 (see Figure 6 in Appendix). From the decision tree, we found that the most important features are age, gender, and number of logins. From the exploratory analysis (see Figure 7 in Appendix), we found a positive relationship between the probability that a user asks for a refund and age, and a negative relationship between the probability that a user asks for a refund and the number of logins. Also, women are more likely to request refunds; however, this is because women are more likely to make payments in the first place.

Recommendations

Recommendation 1 (A-level):

Our first recommendation is to narrow the focus of the game's targeted marketing efforts in order to prioritize high value users and increase marketing ROI. Since we have identified that valuable users tend to be highly educated, aged 50 or older, and that female users remained subscribed slightly longer than male users, we recommend that the game advertise on websites or within apps that primarily reach that audience.

The most direct way to specifically target this demographic is to seek out mobile games or web-based games that have an established presence among educated users older than 50. Examples may include puzzle games such as word puzzles, and casual games such as Homescapes or Gardenscapes. This way, you are guaranteed to reach an audience that is interested in online gaming, and that is likely to fit your most highly valued cohort of players.

Aside from piggybacking off of existing games, it is also possible to place ads on websites that receive a high amount of traffic from this targeted demographic. According to Blue Fountain Media[1], 70% of internet users older than 50 use the internet to read news, and 71% search for health information online. This could make pharmaceutical websites or news media outlets great resources to bring high-value players into the game.

Recommendation 2 (A-level):

Our group reviewed the login behavior of the game's users, and we noticed that the vast majority of players have only logged in 2-3 times (see Figure 1 below). Even when narrowing the data down to users who have paid for the game, the median logins per capita is still only 3. Since this is a subscription-based game, long-term engagement is very important to keep players active and continuously paying. As expected of subscription-based monetization models, Figure 8 in the Appendix shows that there is a direct correlation between the amount of time that a player is subscribed and the total dollar amount that they have been charged.

We believe that a median login count of 3 is very low for a game that relies on long-term engagement, and that it should be a high priority to incentivize repeat logins. One method to increase the game's login cadence could be to offer "daily rewards," such as in-game currency or items, in order to drive players back to the game across multiple days. Another option is to introduce timed gameplay elements, such as chests that require 12 hours to unlock, or an exclusive game mode that only appears once every 48 hours. Motivating players to spread their game sessions out over a longer time frame rather than concentrating their time into one long session will stretch their activity hours and increase the likelihood that they remain paying subscribers for a longer time.

*Figure 1: How many users logged in fewer than 20 times*

---

[1] https://www.bluefountainmedia.com/blog/reach-older-audience-online
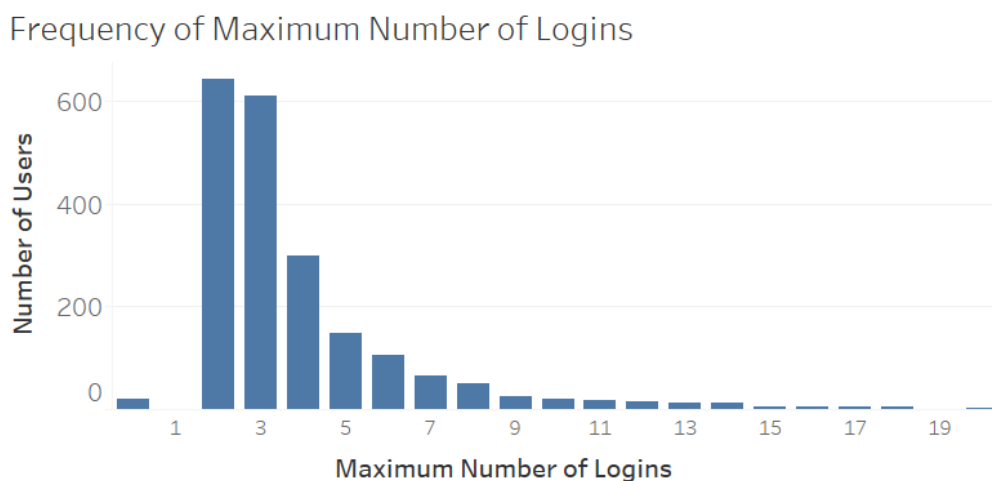
Recommendation 3 (B-level):

        Our final recommendation is to re-evaluate the game's current user acquisition partners, and reallocate resources towards the highest performing channels. We evaluated the ARPU, Conversion Rate, and average length of subscription for users sourced from all nine channels and compared them with users who were acquired without the use of a partner (shown as null values within the data). Figure 2 shows that while some partners outperformed the control group of independently acquired users, the majority of them actually underperformed. Since it likely costs a premium to acquire users through a partner, it is imperative to end relationships with low-value partners in order to prioritize high performing ones.

        For example, partners 3 and 9 were effective at acquiring new users, but the ARPU of these users is more than $10 lower than the group of users who were not sourced by a partner. On the other hand, partners 2, 5, and 6 did not provide many users, but these users turned out to be very valuable. Although the sample size from these partners is very small and may be influenced by outliers, we believe that it is worth expanding these relationships in case there is potential to acquire even more high-value players.

Figure 2: Partner performance

Partnership Statistics
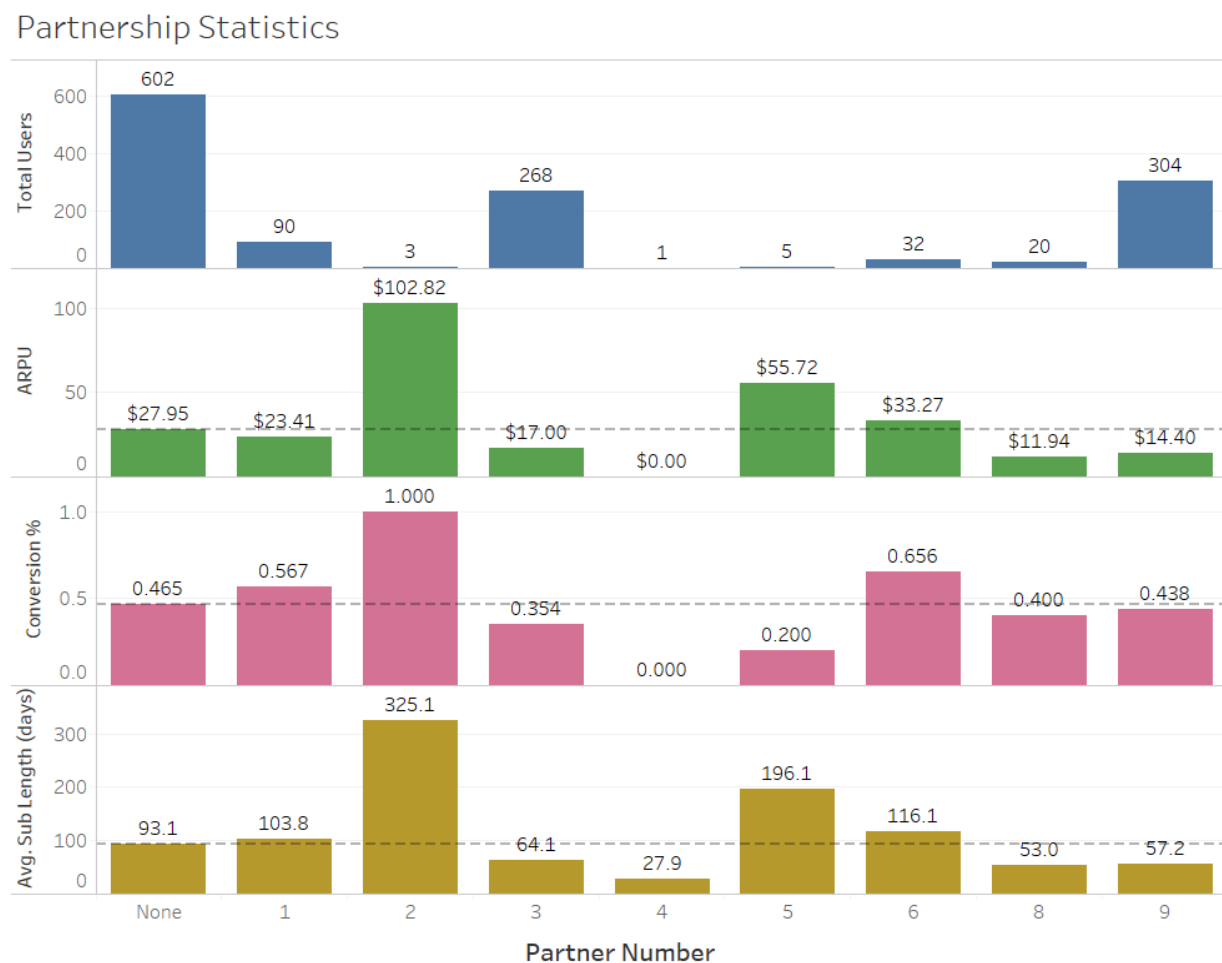
Conclusion

      Using the provided player data, our group was able to determine the demographic that is most likely to provide value for the game, which will allow the game's marketing efforts to be more precisely targeted. We were also able to evaluate partner performance and develop a strategy to drive long-term engagement by increasing login frequency. The fact that this game has been sustaining a base of paying subscribers since 2010 is very impressive; however, with the data analysis techniques mentioned in this report, we believe that the game's performance can be optimized even further.

Appendix

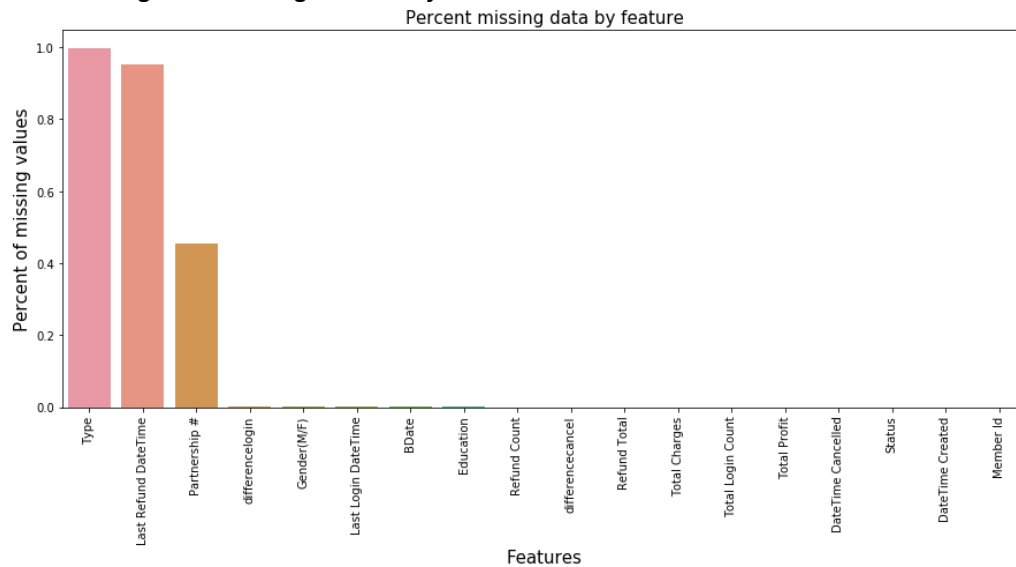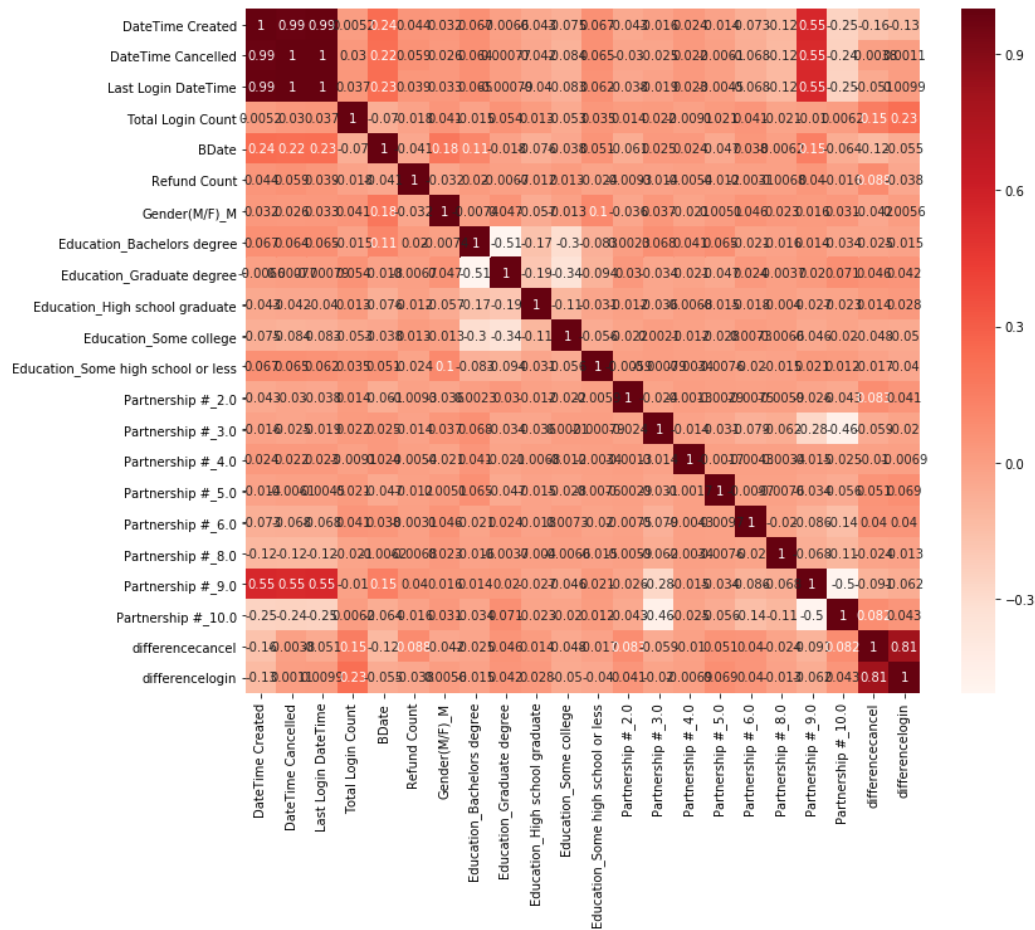*Figure 3: Percentage of missing values by feature*

Percent missing data by feature

Figure 4: Pearson correlation heatmap

Figure 5: Linear regression to find users with highest lifetime values

```
                              OLS Regression Results
================================================================================
Dep. Variable:           Total Profit   R-squared:                      0.835
Model:                            OLS   Adj. R-squared:                 0.834
Method:                 Least Squares   F-statistic:                    1333.
Date:               Tue, 12 Nov 2019   Prob (F-statistic):              0.00
Time:                        08:50:36   Log-Likelihood:               -5806.0
No. Observations:                1323   AIC:                         1.162e+04
Df Residuals:                    1317   BIC:                         1.166e+04
Df Model:                           5
Covariance Type:            nonrobust
================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const                       -5.9366      1.082     -5.488      0.000      -8.059      -3.815
Refund Count               -11.1129      1.515     -7.335      0.000     -14.085      -8.141
Education_Bachelors degree   3.4004      1.347      2.524      0.012       0.757       6.044
Education_Graduate degree    2.8601      1.299      2.202      0.028       0.312       5.408
Partnership #_10.0           2.5433      1.085      2.344      0.019       0.415       4.671
differencecancel             0.3118      0.004     80.949      0.000       0.304       0.319
================================================================================
Omnibus:                     1891.297   Durbin-Watson:                   1.947
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         1199685.848
Skew:                          -7.755   Prob(JB):                        0.00
Kurtosis:                     149.705   Cond. No.                        552.
================================================================================
```
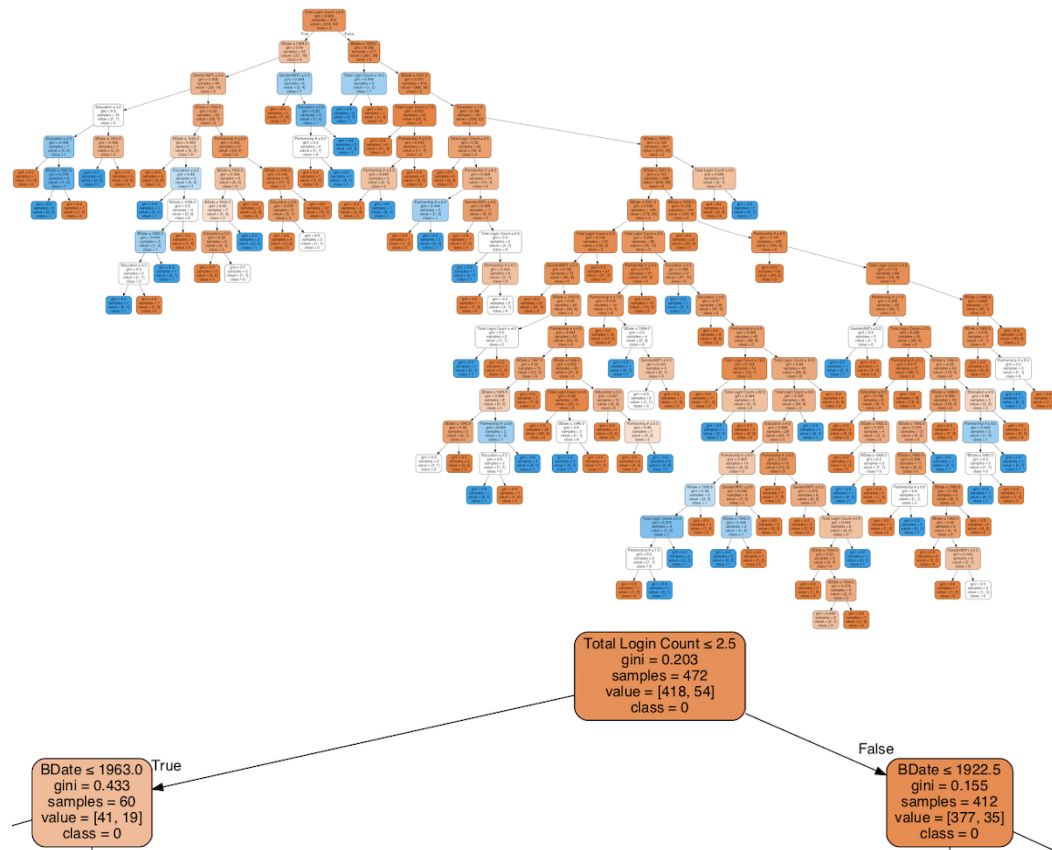
Figure 6: Decision tree to find the most important features

Figure 7: Exploratory analysis on the most important features

Refunds Per Person by Age Group
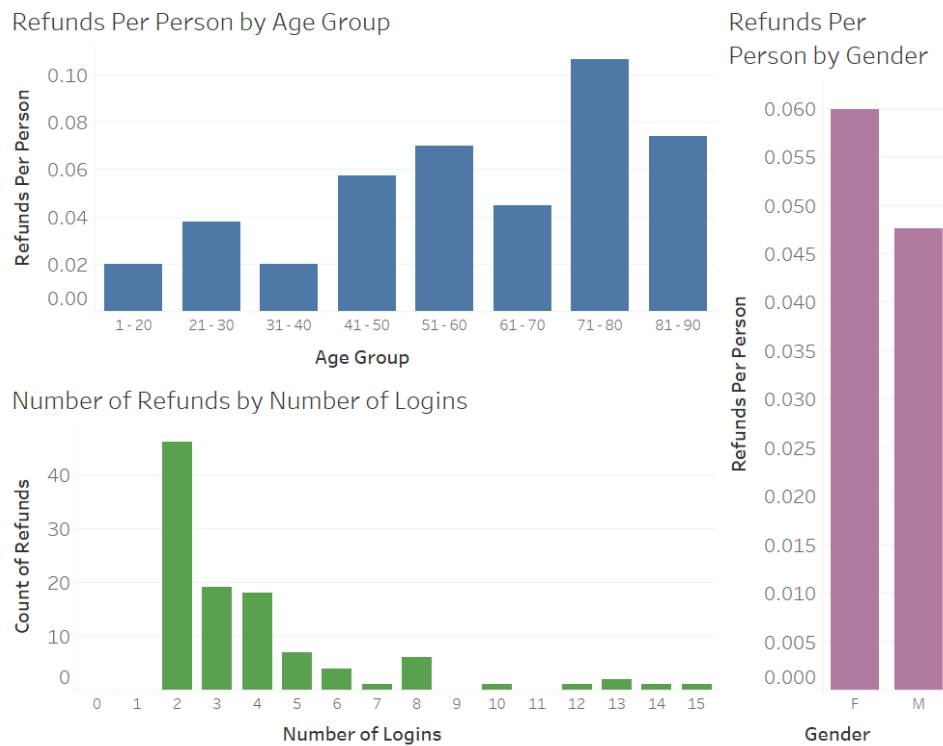
Refunds Per Person by Gender

Number of Refunds by Number of Logins

Figure 8: Relationship between Subscription Length and Total Charged:
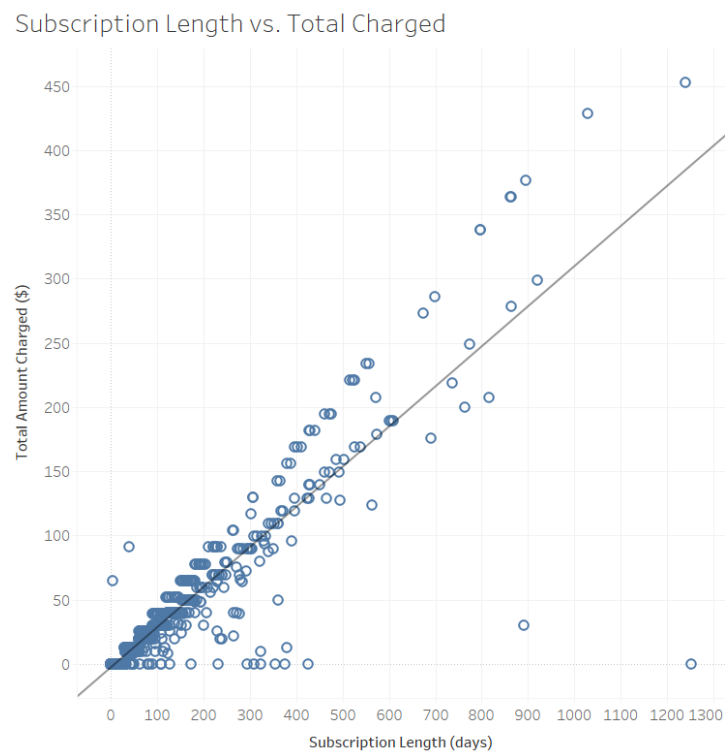

Subscription Length vs. Total Charged

Figure 9: KPIs Organized by Age Group

Various KPIs Organized by Age Range
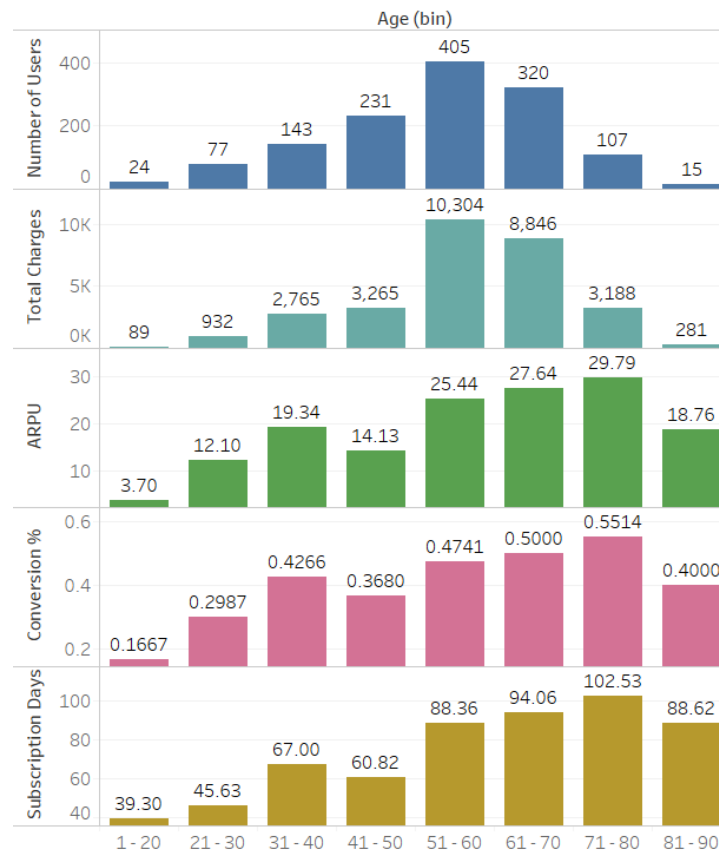
Figure 10: KPIs Organized by Education Level

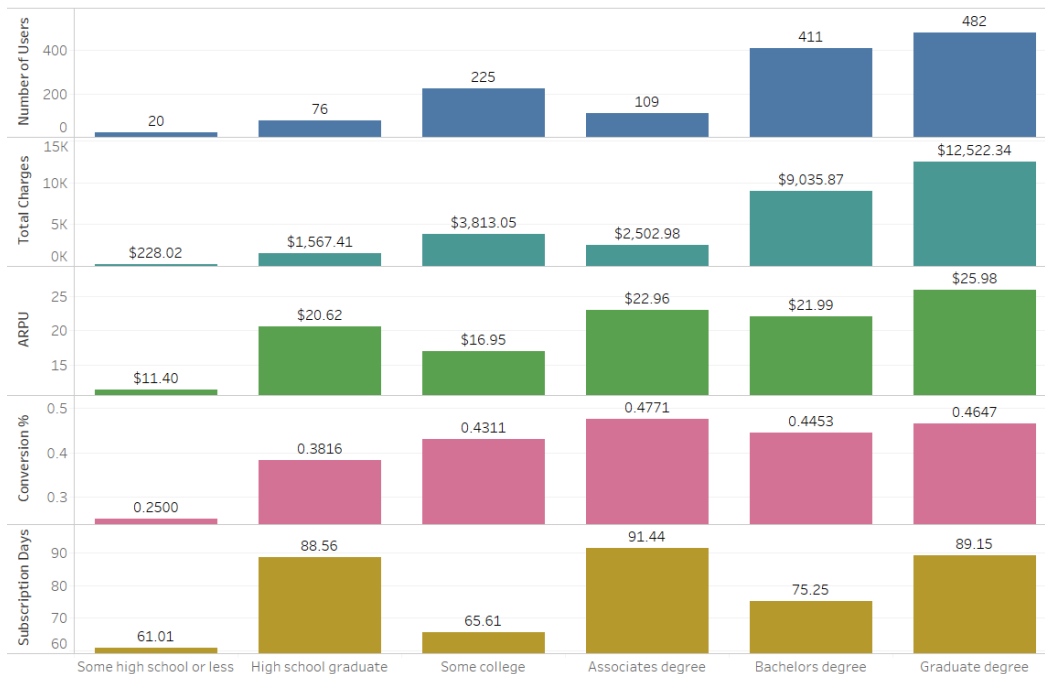

Various KPIs Organized by Education Level

Figure 11: KPIs Organized by Account Creation Date
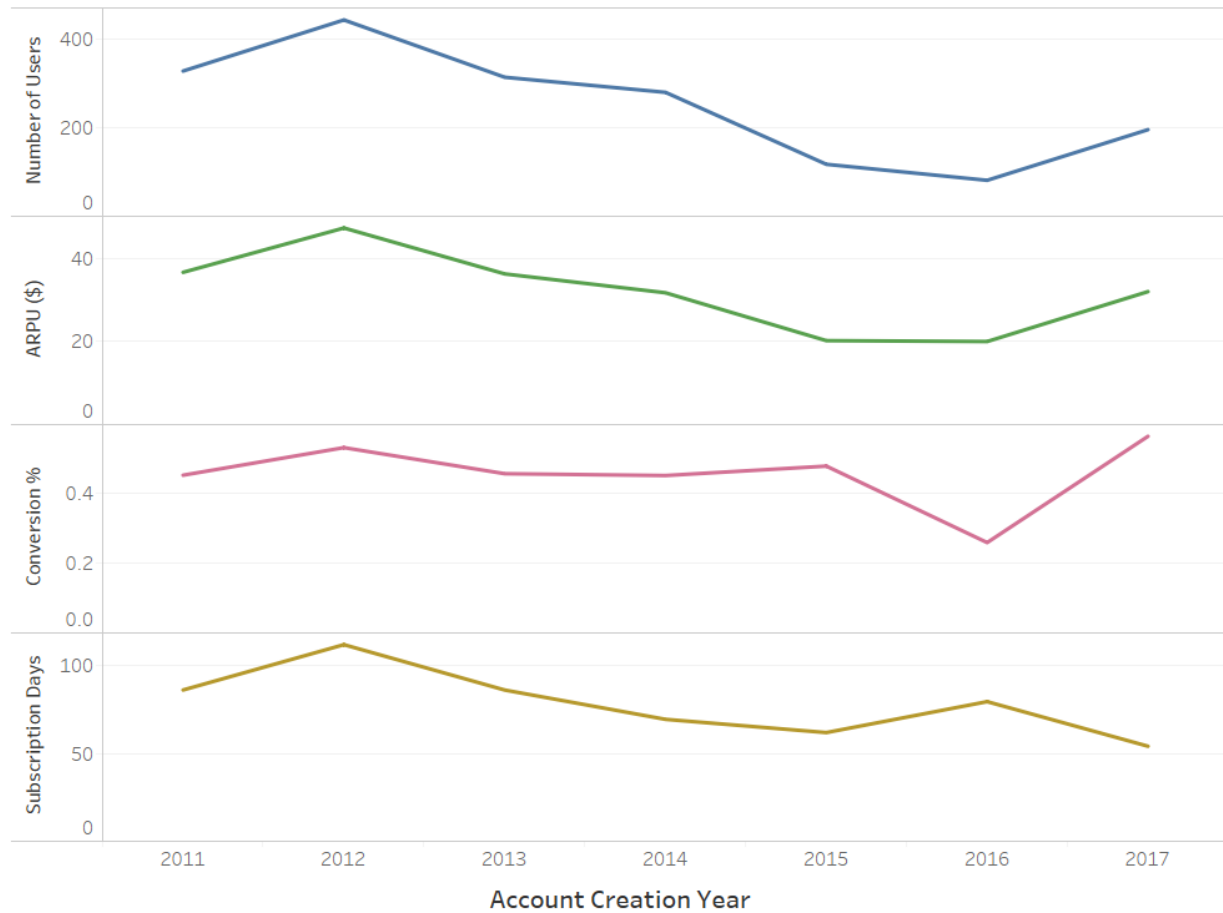

Various KPIs Organized by Account Creation Year

Figure 12: Number of new accounts by month: