Gaming Analytics Project 1

DSO – Fall 2017

Yuzhou Dou

Bangyu Mao

Yu Zhang

Wenzhen Zhao

Shuhan Zhou

Date: Nov, 2017

Insights & Recommendations

- **Play Patterns:** During summer break, users even played game frequently until 2AM; therefore, we also recommend the company allocate more resources during that period of time. However, from July to September, we are losing many of them. From late afternoon(2PM) to midnight(12AM) on Thursday and Saturday, the users are most active. So, we could promote ads during these times to monetize.

- **Frequent Players:** Based on the frequency prediction, we should target older students who want to challenge themselves. Also, keeping the average scores and stars low can also keep users come back more frequently.

- **Churn Analysis:** Players will churn when they don't feel confident in continuing answering the questions right. The analysis shows churners actually are more actively involved in games than non-churners, but their scores and performance are significantly slower. Our recommendations would be: Identify the high-risk churners based on our churn prediction model, and customize (reduce) the game difficulty levels to these players, and send notifications and incentives to re-engage these customers.

- **Activity Outcome:** Students with high average cumulative completion rate, points earned, started rounds, and low average cumulative number of submits and adaptive score are more likely to complete the game. Students' historical playing behavior is more important than the game attributes for the activity in determining activity outcomes.

Statistical Modeling & Results

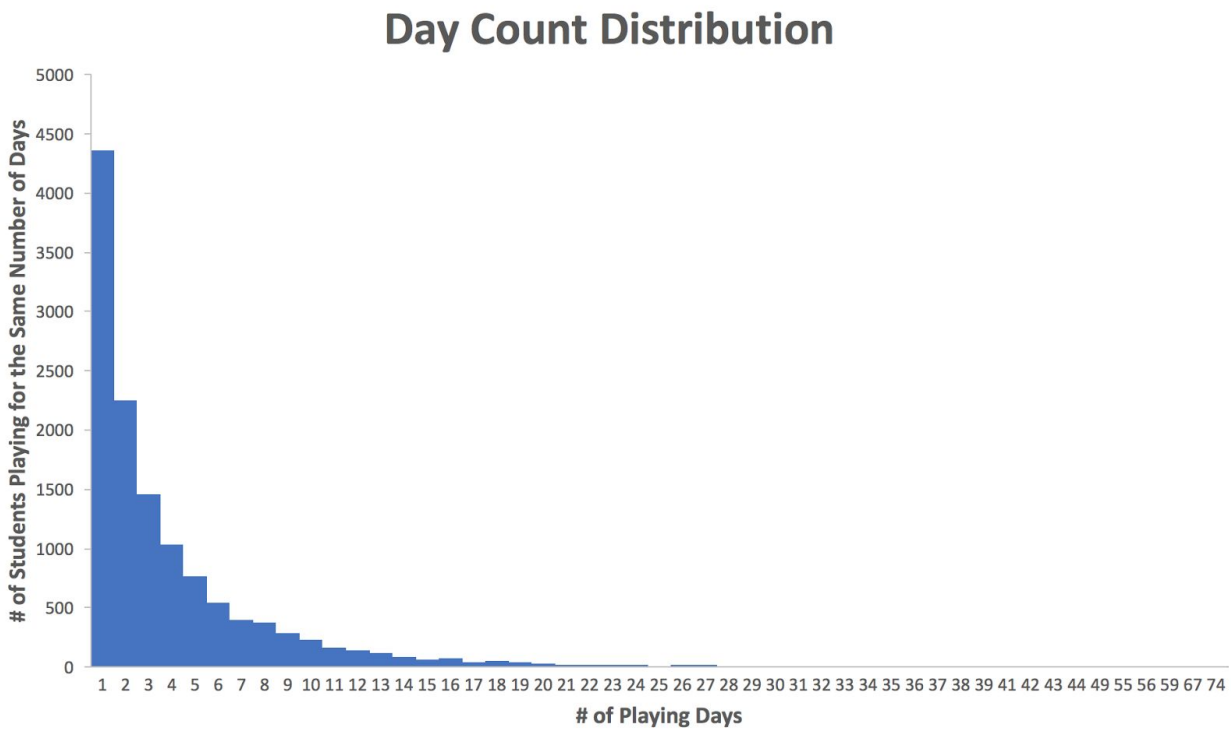
Question 1: Typical Patterns

1. Daily Play Pattern

(1) Business Question

What are the typical play patterns on daily basis?

(2) Data Visualization



(3) Insights

- First of all, among 12696 students, 1531 of them played the game only on the weekends. However, there are many more students, 5289 of them played only on the weekdays, and 5889 of our users played on both. Therefore, it is a game of everyday use.

- 4363 students, played our game for only one day, and over 63% of the students played for only three days. This gives a hint that most of the students are new users and we have a big churn rate potentially.

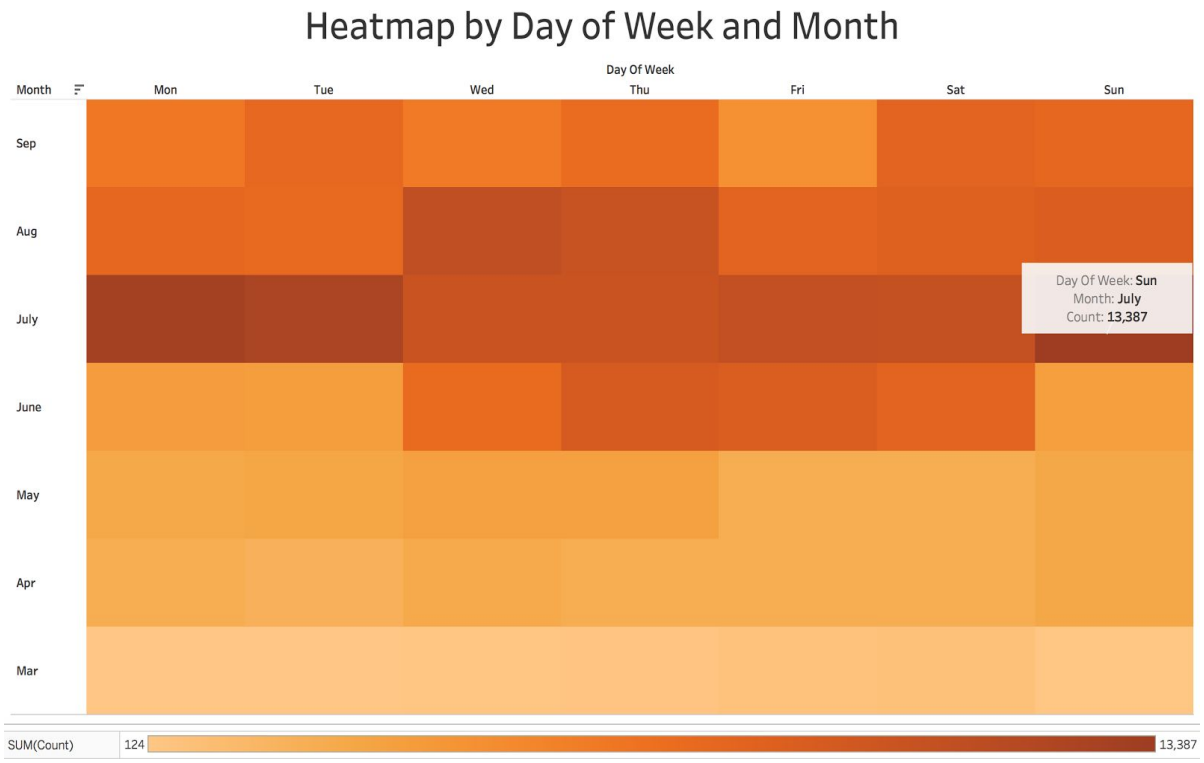
- Since we have a long right tail, we still have many frequently-playing users. As a matter of fact, if we define a loyal user as users that played over 10 time in the given period, we still have 1215 loyal users, accounting for 10%.

2. Weekly and Monthly Play Pattern

(1) Business Question

What are the typical play patterns on week and month basis?

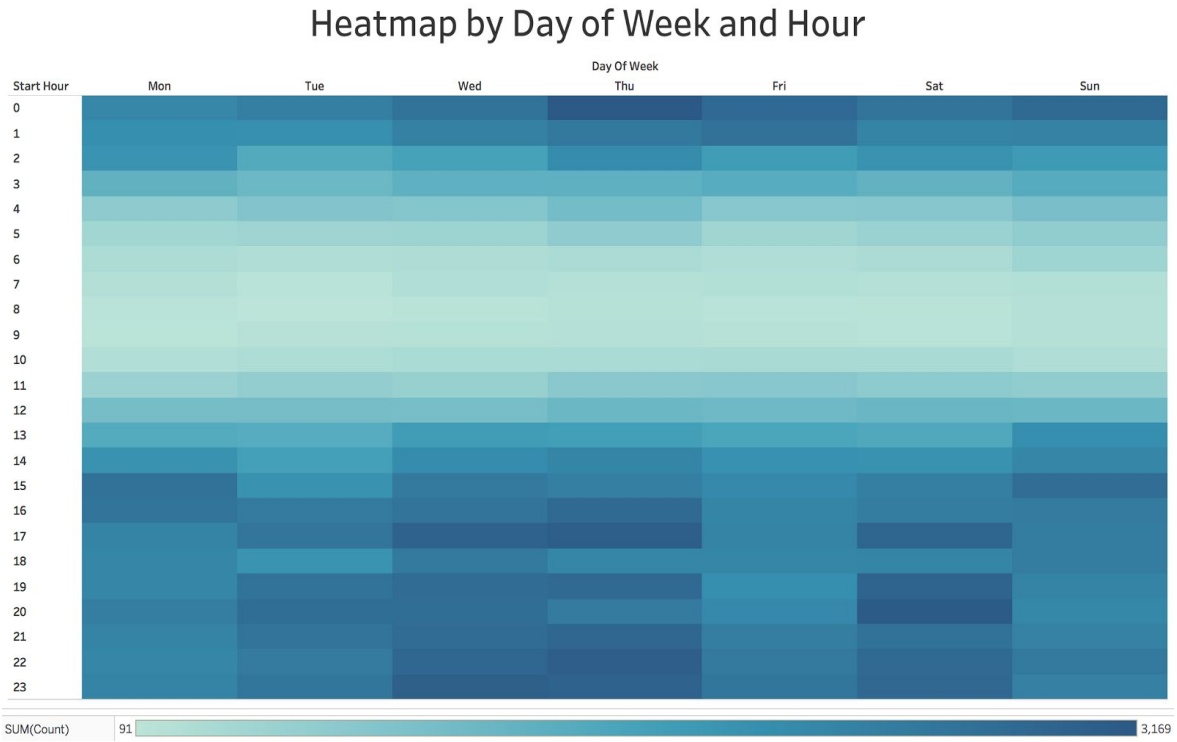
(2) Data Visualization



(3) Insights

- Looking into the play pattern by day of the week and month, we can find that July has the biggest number of activities. Apparently, more students are willing to play during summer vacation, from late June to early September, considering both the quarter and semester year. In addition, although 5289 of the students played only on the weekdays, the busiest time is Sundays in July.

3. User Activity Pattern by Day of Week and Hour of Day



(3) Insights

- If we check the play pattern by day of the week and hour of the day, we can conclude that very few activities happens in the morning, from 4am to 11am. Comparatively, 16pm-17pm, 21pm-22pm are two popular time slots, and Thursday outperforms other days of the week.

- Furthermore, there are some fun facts here. For example, on Friday night, people seldom choose to play our game. Saturday morning and Monday morning seem not a good time for the users to wake up and play.

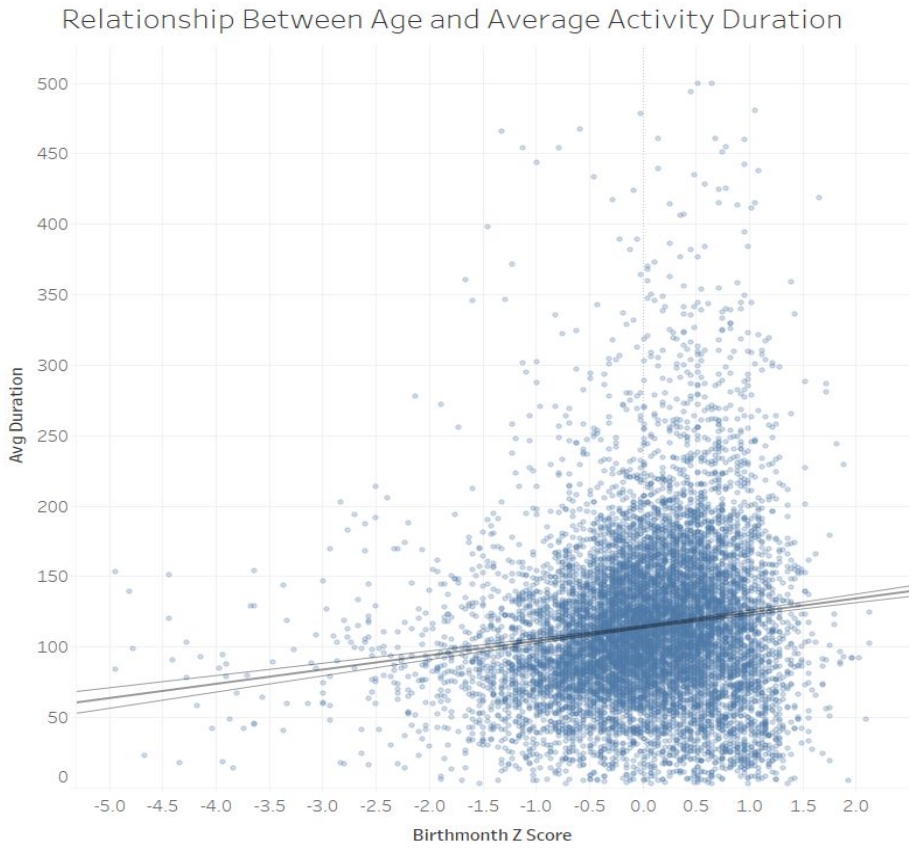
Question 2: Player Segments

1. Relationship Between Age and Average Activity Duration

(1) Business Question

Does the time spent in playing each activity vary by players' age?

(2) Data Visualization



Birthmonth Z Score vs. Avg Duration. The data is filtered on Birthmonth Z Score and Avg Duration. The Birthmonth Z Score filter ranges from -5 to 2.5. The Avg Duration filter ranges from 3 to 500.

(3) Insights

- Most players' birthmonth z scores fall in range [-2.0, 1.5], and their average duration of playing games mostly falls in [3,200].

- Those observations that have average activity duration higher than 200 are mainly plotted near or above birthmonth z score 0.

- From the trendline, we can see there exists a trend that older the age, longer the average activity duration.

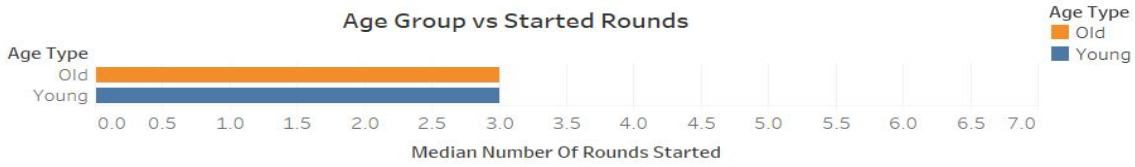
2. Relationship Between Age and Rounds Played

(1) Business Question

Do younger students act differently in terms of starting, completing, or passing activity rounds from older students?

(2) Data Visualization

- For better analyzing, we separate students into two age groups: Young and Old. Those students having birthmonth z score smaller than 0 are assigned to “Young” group, and those students with birthmonth z score larger than 0 are assigned to “Old” group.



(3) Insights

- The median number of rounds started for both Young and Old groups of students is 3.

- The median number of rounds completed for younger students is still 3, but that for older students becomes to 2. Older students tend to complete less rounds than younger students.

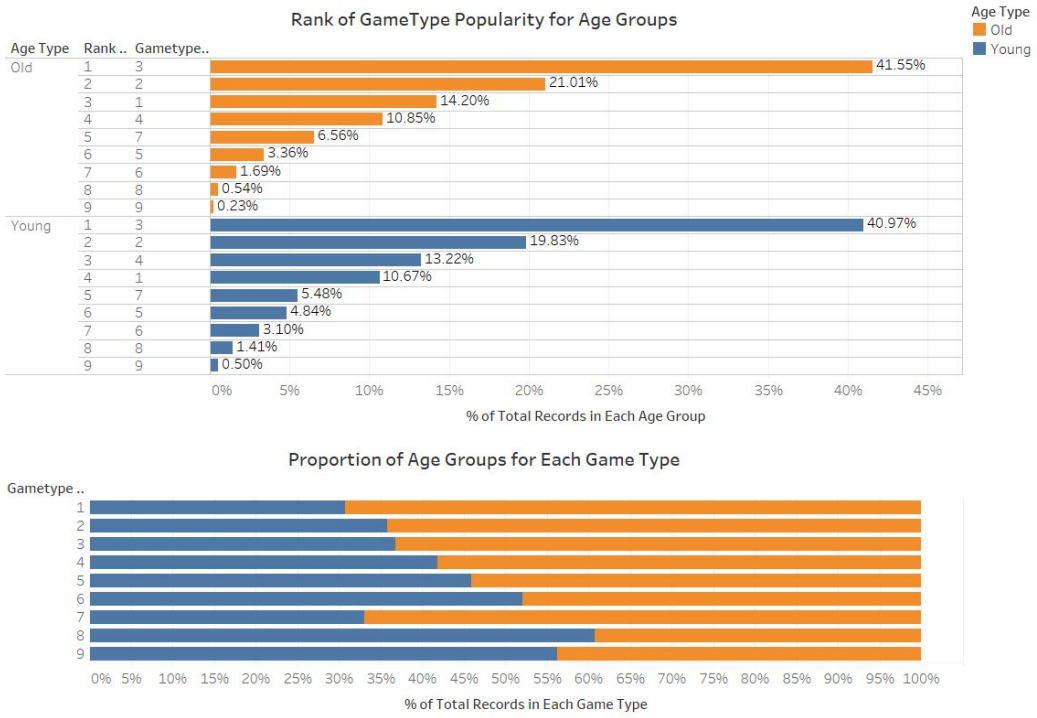
- The median number of passed rounds for younger students is 2, 1 round less than their median number of started and completed rounds. The median number of passed rounds for older students is 1, 1 round less than their median completed rounds and 2 rounds less than their median started rounds. Older students tend to pass less rounds than younger students.

3. Game Type Preference among Age Groups

(1) Business Question

Do younger students prefer different game types from older students?

(2) Data Visualization



(3) Insights

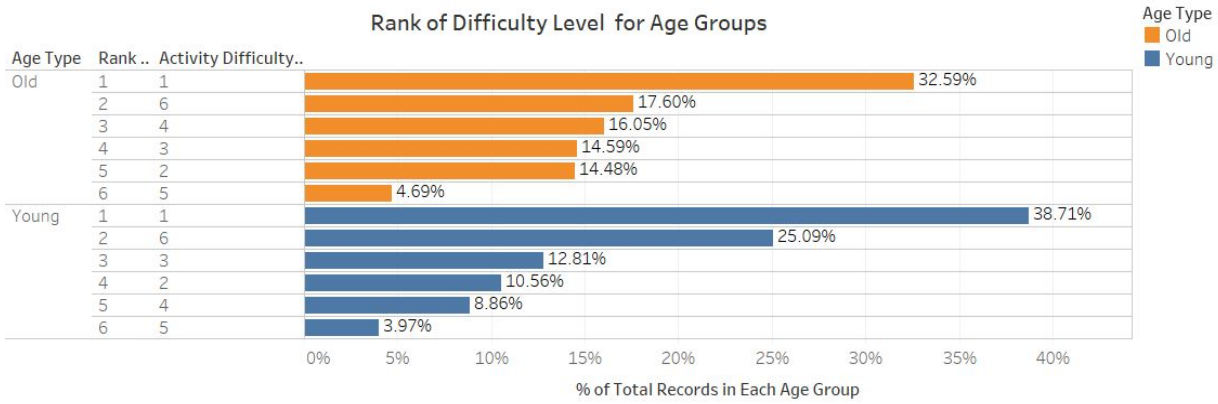
- Game Type 3 and Game Type 2 are the top 2 game types preferred by both younger and older students.
- Younger students prefer Game Type 4 more than Game Type 1, while older students prefer Game Type 1 more than Game Type 4.
- Younger students are major players for Game Type 6, 8 and 9, and older students are major players for all the other game types.

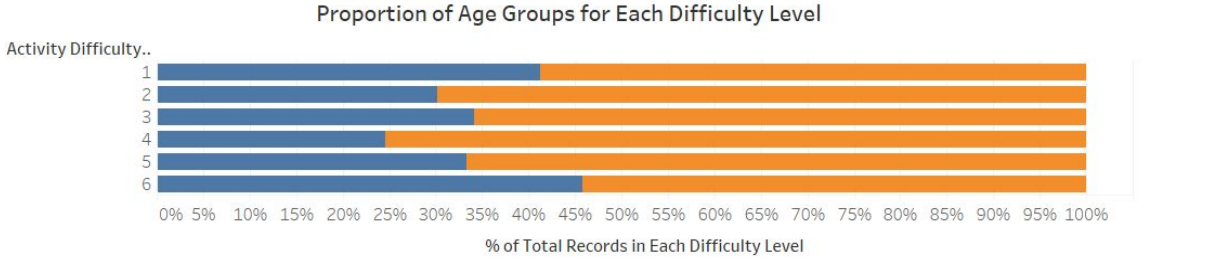
4. Difficulty Level Difference among Age Groups

(1) Business Question:

Do younger students play different difficulty levels from older students?

(2) Data Visualization





(3) Insights

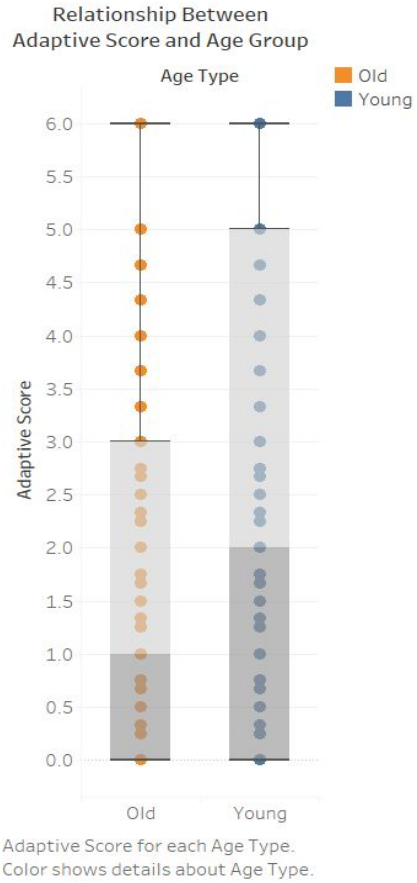
- Difficulty Level 1 and 6 are the top 2 difficulty levels played by both younger and older students. Difficulty Level 5 is the least played difficulty level by both age groups.
- Older students play Difficulty Level 4 more than Difficulty level 3 and 2, while younger students play Difficulty Level 3 more than Difficulty Level 2 and 4.
- For all the difficulty levels, older students are major players.

5. Adaptive Score Difference among Age Groups

(1) Business Question

What's the difference of adaptive scores got by different age groups?

(2) Data Visualization



Average Score for Different Age Groups

Age Type	
Old	1.6751
Young	2.4982

(3) Insights

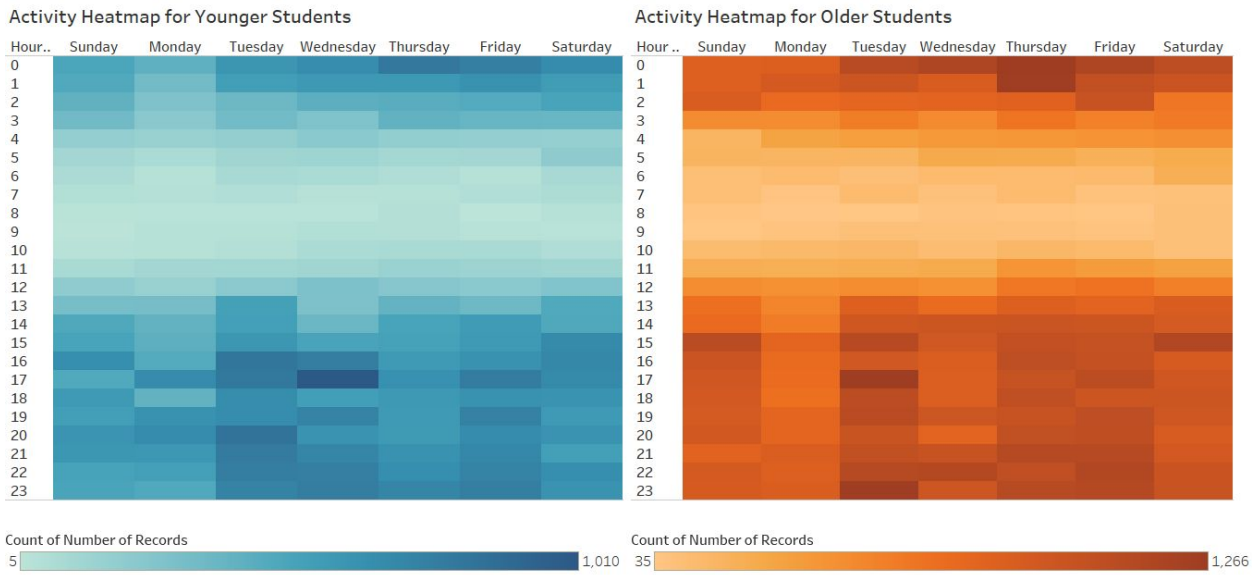
- Although the lowest score is 0 and the highest score is 6 for both age groups, older students have a median score of 1 and upper hinge of 3, while younger students have a median score of 2 and upper hinge of 5. Old group has an average adaptive score of 1.6751 and Young group has an average adaptive score of 2.4982. Younger students tend to get higher scores than older students in general.

6. Activity Time Difference among Age Groups

(1) Business Question

Is there any difference on time of the day and day of week among age groups?

(2) Data Visualization



(3) Insights

- Most activities happened from 1pm to 2pm for both age groups.

- Younger students have most activities on Tuesday and Wednesday and least activities on Monday. Older students have most activities on Tuesday, and least activities on Monday and Wednesday.

- Younger students' activities are more concentrated from 4pm to midnight. Older students' activities are more concentrated from 2pm to 7pm and from 9pm to 2am. In conclusion, older students play more in the midnight.

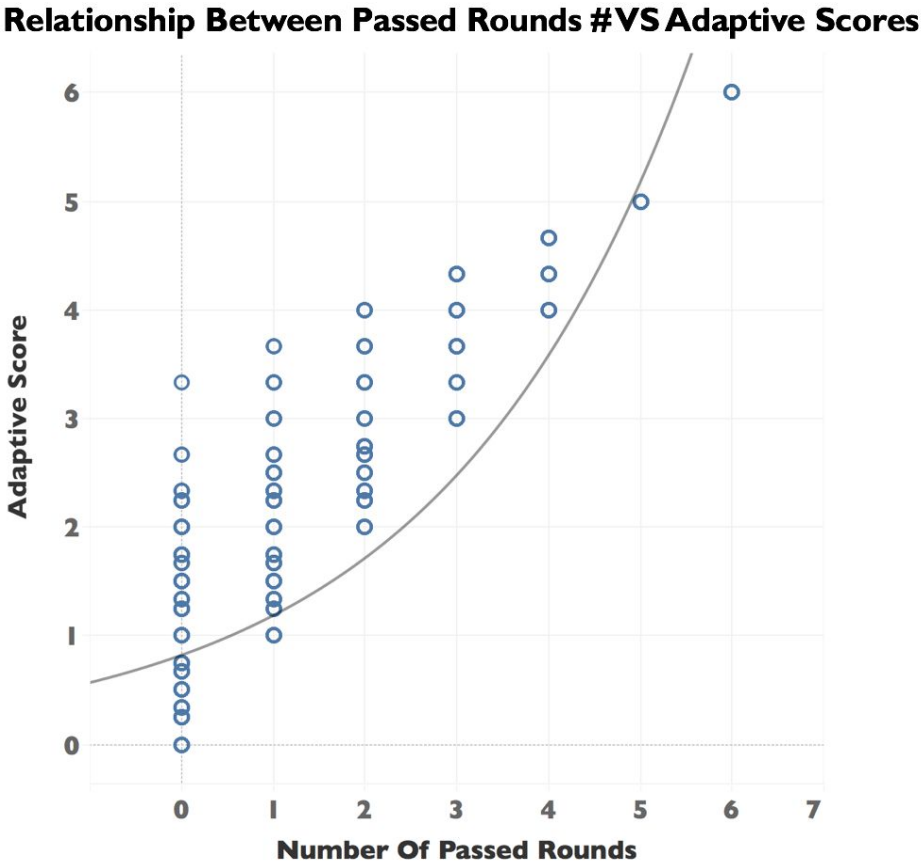
Question 3: Player Ability

1. Business Question

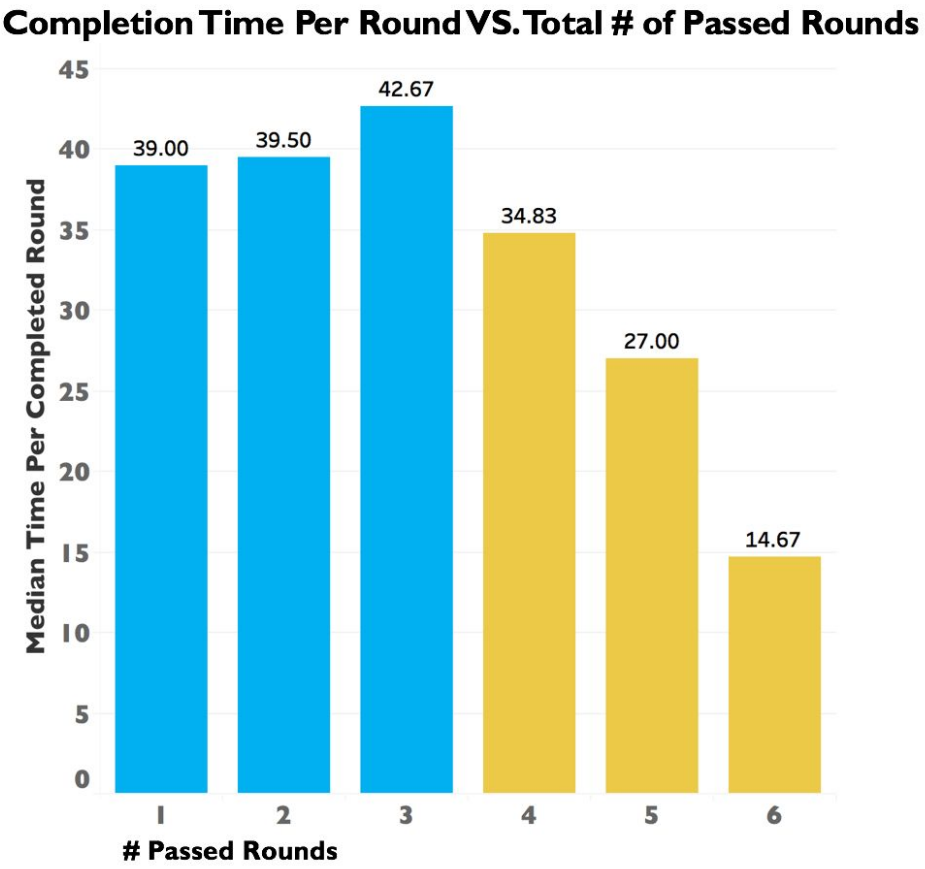
Do students who get more answers right do better than students who do not? Is there a difference between the students who get answers right correctly early versus later?

2. Data Visualization

- We visualize the relationship between quality of answers of students (which is indicated as the number of passed rounds), and the overall performance (which is quantified as adaptive scores).



- In order to find the relationship of students' answering speed with their performance, we calculate the median time interval that a student spends per round of game distributed on the levels of number of rounds that they pass. The bar chart is shown as below:



3. Insights

- The scatterplot illustrates the positive exponential trend between number of passed rounds with adaptive scores. In particular, the people who passed 5 or 6 rounds definitely score higher than rest of the people.

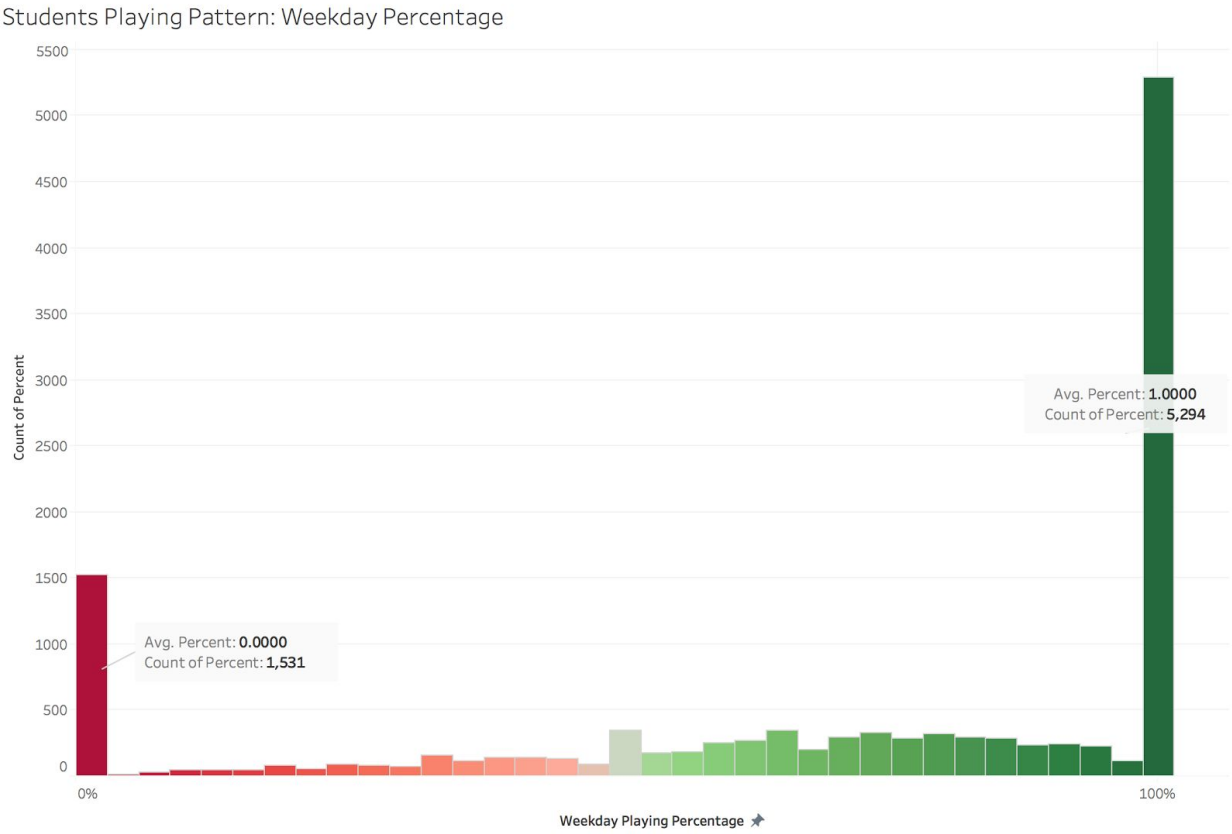
- The bar chart indicates there is a significant difference between whether students submit their answers quickly or slowly in term of getting the answers correctly. Although it seems there is a mild increase of median time spent on each round associated with more rounds passed, time spent per round drops dramatically when observing the pattern of high performing students (when they passed more than 4 rounds). In general, people who get answers quickly usually pass more rounds as well.

Question 4: Typical Play Time

1. Business Question

Are there some students who typically play on weekdays and others who typically play on the weekends?

2. Data Visualization



3. Insights

- We created a label into each activity: if it happened on weekdays, it is 1; if not, it is 0. And then we calculated the percentage of times students played the game on weekdays. 100% means the student played the game all on weekdays, and 0% means the student played the game all on weekends. If equally distributed, the percentage should be 71.43%.

- From the graph, we can see a clear playing pattern that, there are many more students playing the game either on weekdays or on weekends than students playing the game both on weekdays and on weekends.

- Specifically, among 12696 students, 5289 of them only played on weekdays, 1531 of them only played on weekends.

Question 5: Game Difference between in School Year and Not

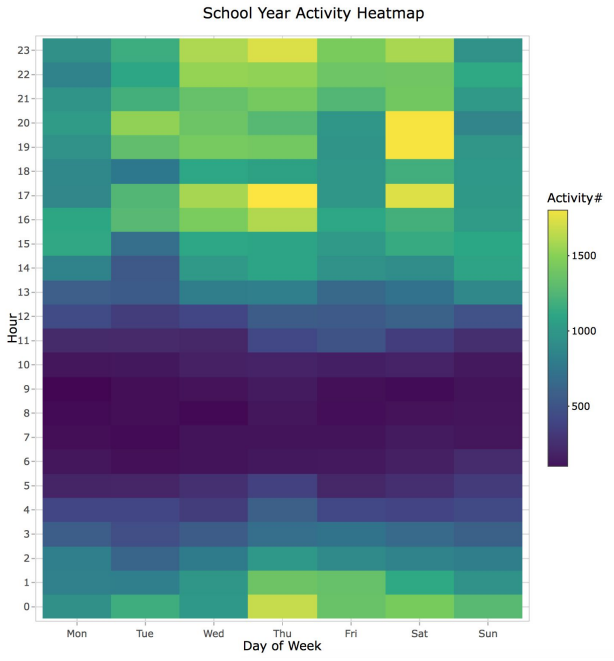
1. Activity Time Difference

(1) Business Question: for games play during and not during the school year, is there a difference on time of the day and day of the weak?

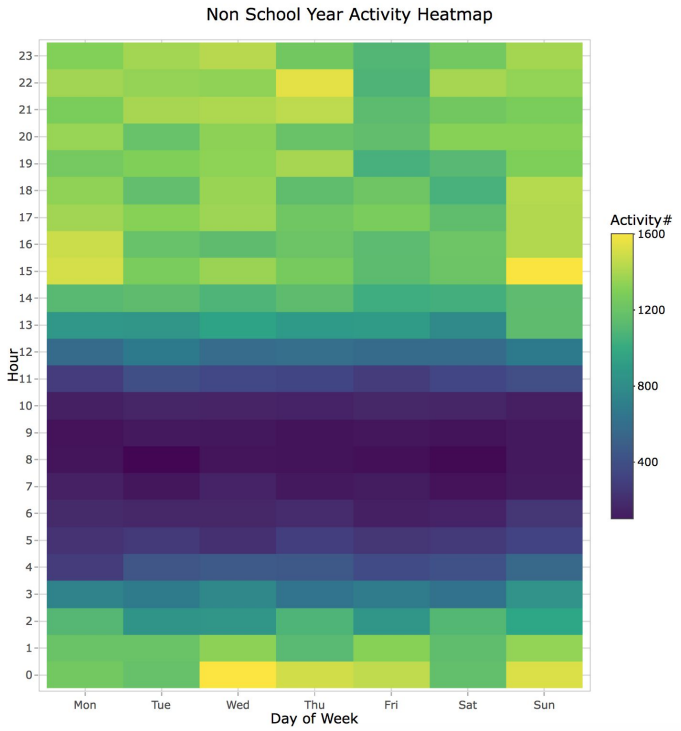
(2) Data Results/Viz

- Different school had different summer break time; for this question, we assume July and August are the non-school months and the rest are school months

- Heatmap - School Year



- Heatmap - Non-School Year



(3) Insights

- No matter it's school year or not, most activities happened from 2PM to 2AM. There's little users played games during the morning(4AM - 10AM).

- In general, the heatmap of activities did not change a lot on school year and non-school years. However, there were still some interesting differences.

- First, users tended to play games later in the midnight during the summer break than school year. During the school year, the activities numbers decreased quickly since 12am; but during the summer break, 12 am was like a peak time of the day and there were still a lot of users play games until 3AM.

- Secondly, during the summer break, there's no significant difference during weekdays and weekends. However, in the school years, Saturday was the apparent peak day of the week, especially during the evening time.

2. Popular Game Type

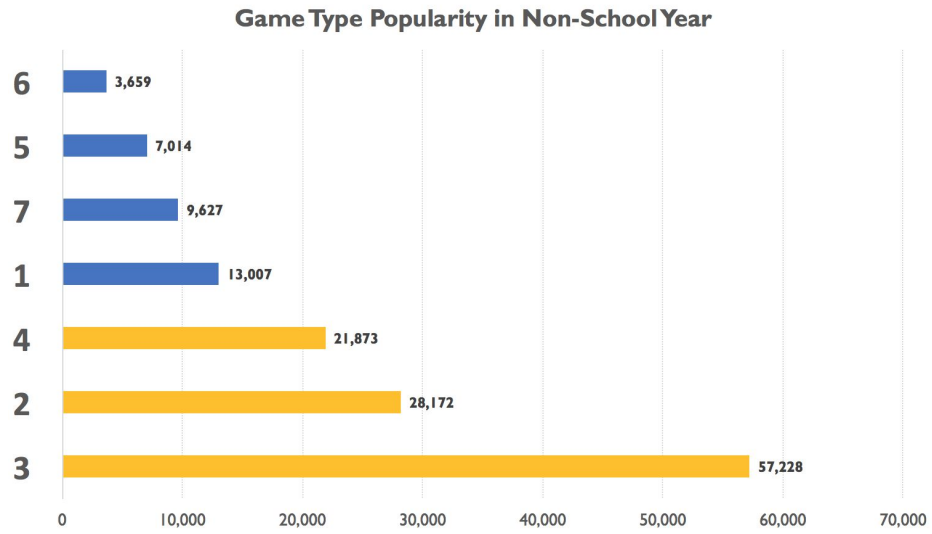
(1) Business Questions

During the school year and non-school, is there any difference on the most popular game type?

(2) Data Results/Viz

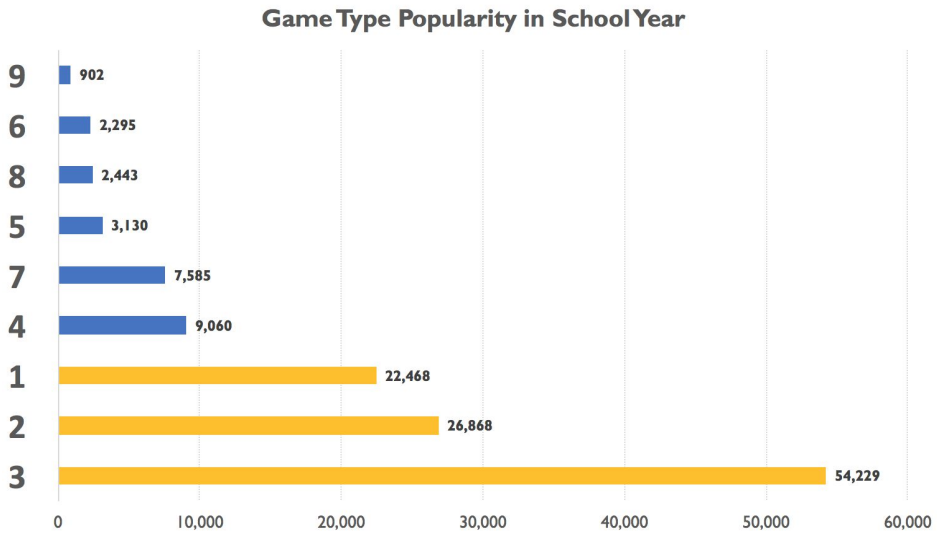
- Non-School Year

Game Type	Total Number
3	57228
2	28172
4	21873
1	13007
7	9627
5	7014
6	3659



- School Year

Game Type	Total Number
3	54229
2	26868
1	22468
4	9060
7	7585
5	3130
8	2443
6	2295
9	902



(3) Insights

- No matter it's school years or not, Type 3 and Type 2 games were the most popular games with highest activity numbers.

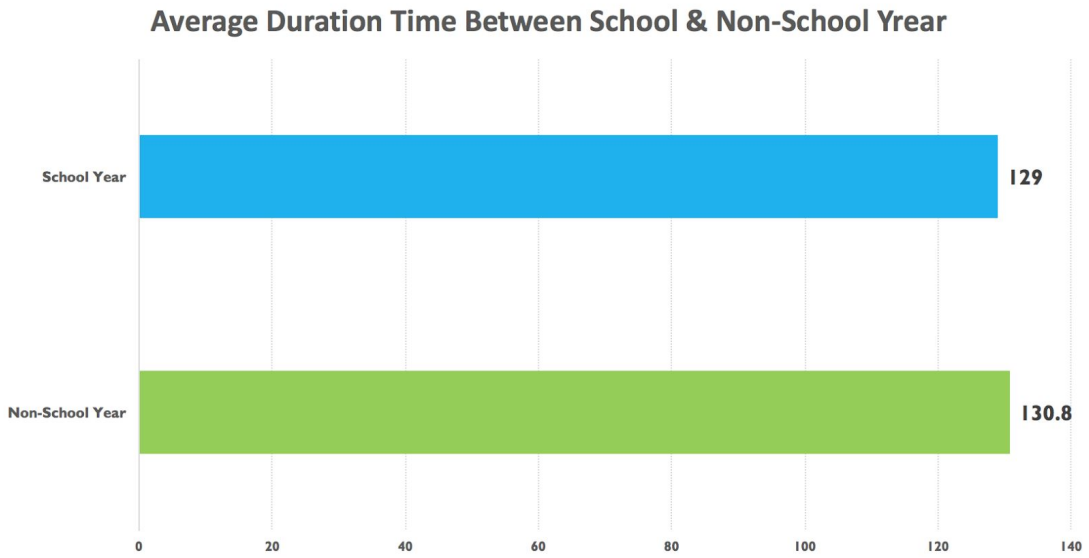
- In School year, Type 1 was the third most popular game type; but during non-school year, type 4 became the third most popular game and the activity number was almost 70% higher than Type 1.

3. Game Related Stats

(1) Business Questions: During the school year and non-school, is there any difference on important metrics like average adaptive score, succeed rate, average difficulty, etc.?

(2) Data Visualization

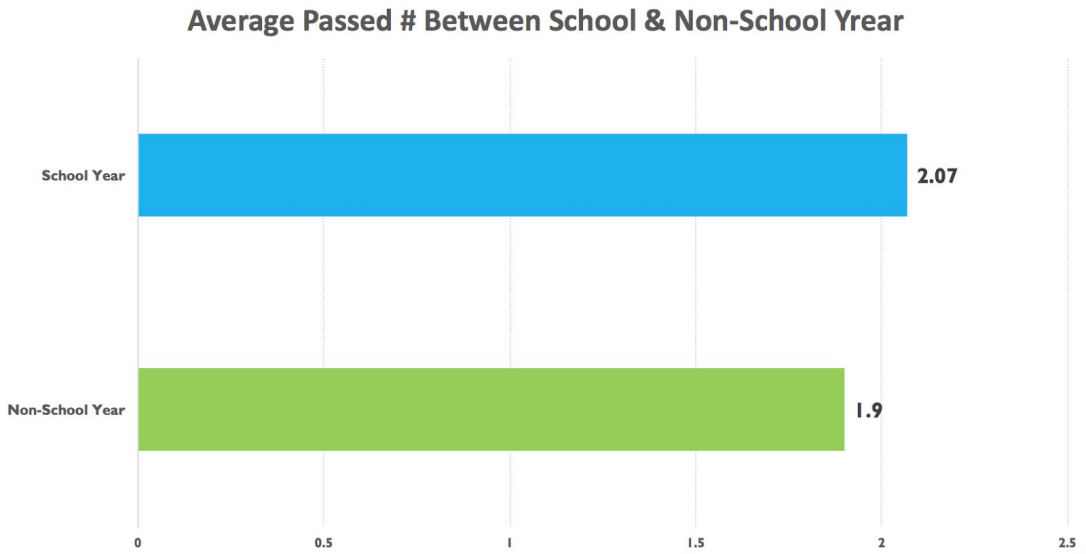
- Average Activity Duration



- T-test

```
t = -3.0454, df = 268140, p-value = 0.002324
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.9163485 -0.6324191
```

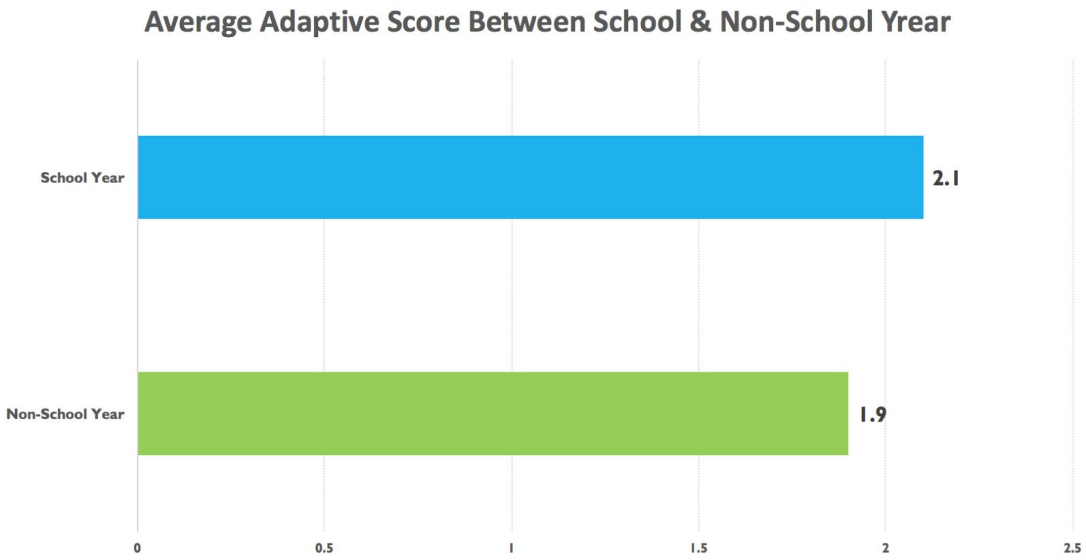
- Average Passed Rounds



- T-test

```
t = 22.851, df = 265240, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1577771 0.1873817
```

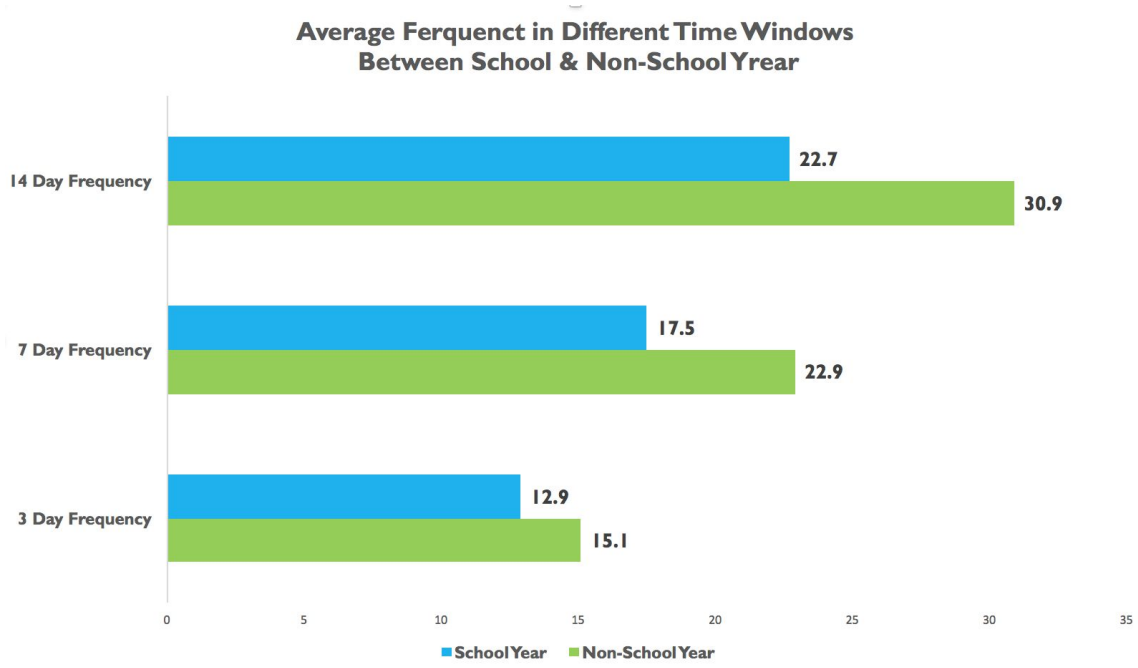
- Average Adaptive Score



- T-test

```
t = 136.48, df = 204920, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.040234 1.070545
```

- Frequency



- 3 days

```
t = -23.196, df = 255670, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.327629 -1.964922
```

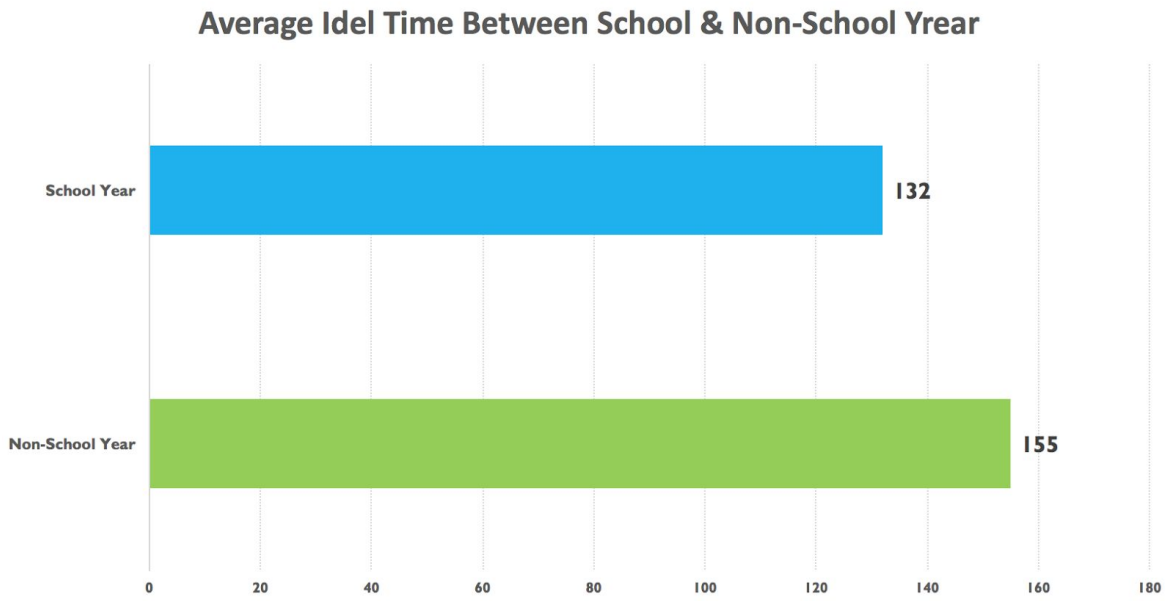
- 7 days

```
t = -35.69, df = 238580, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.681262 -5.089750
```

- 14 days

```
t = -36.399, df = 232710, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.666405 -7.780784
```

- Idle Time

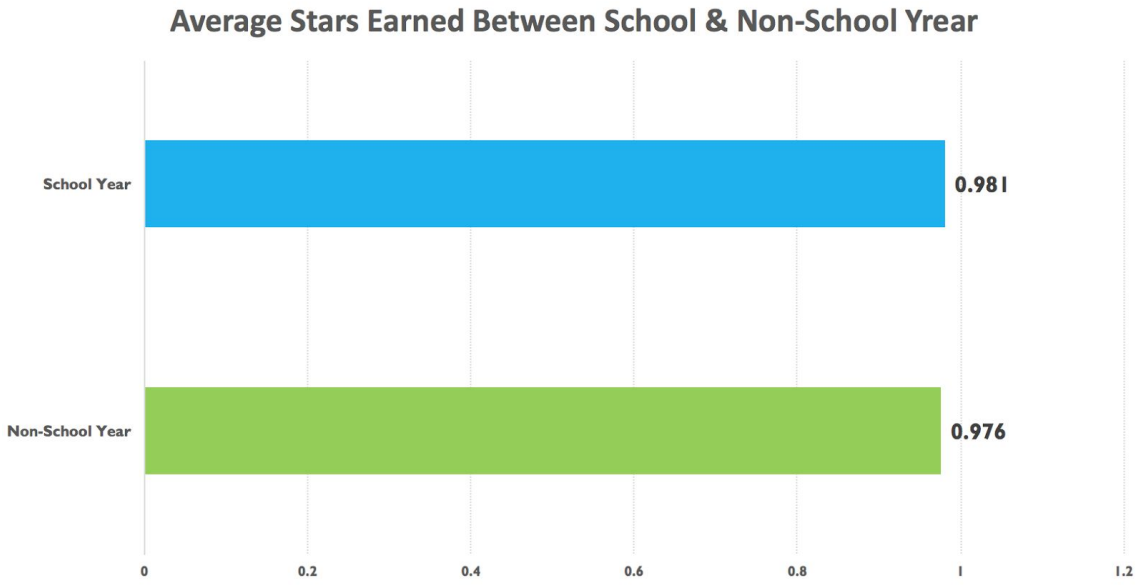


- T-test

```
t = -2.737, df = 268750, p-value = 0.006201
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -39.216003  -6.487395
```

❏ The difference on Idle Time is statistically significant

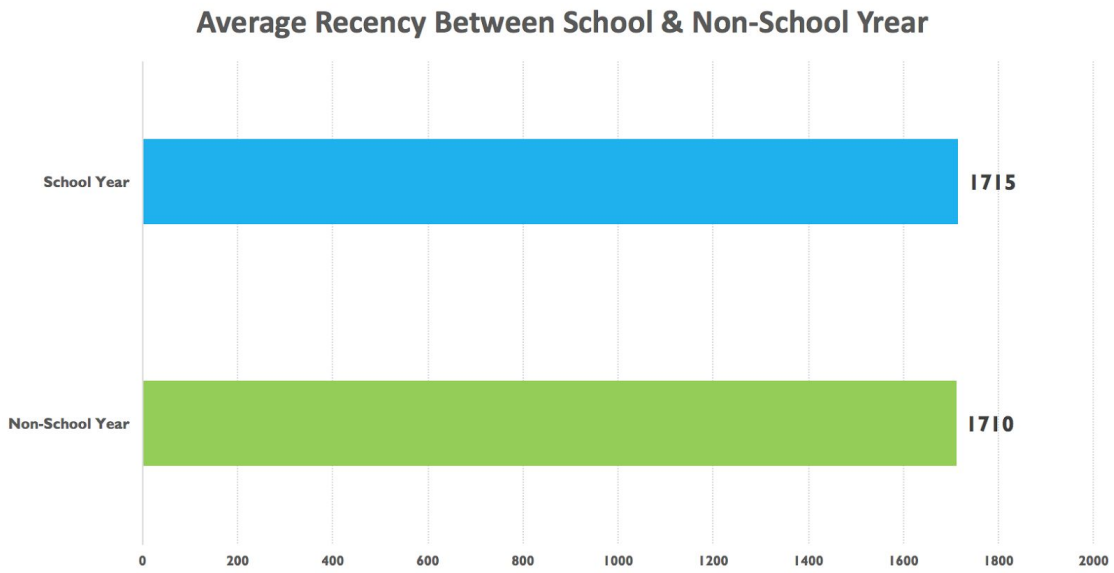
- Average Star Earned



- T-test

```
t = 1.0203, df = 267690, p-value = 0.3076
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.004777078  0.015150353
```

- Average Recency



- T-test

```
t = 0.17484, df = 243050, p-value = 0.8612
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -59.80739  71.52255
```

(3) Insights

- For most most game related metrics, activities in non-school year had significantly better performance than activities in school years. Specifically, activities during the non-school year had higher average duration, higher frequency in all 3,7 and 14 days time windows and also higher idle time. All of these metrics differences were statistically significant. However, for activities in school year, they tended to have higher passed rounds and adaptive scores and they were also statistically significant.

- For Average recency and average stars earned, there were no significant differences between the activities in school year and non-school year.

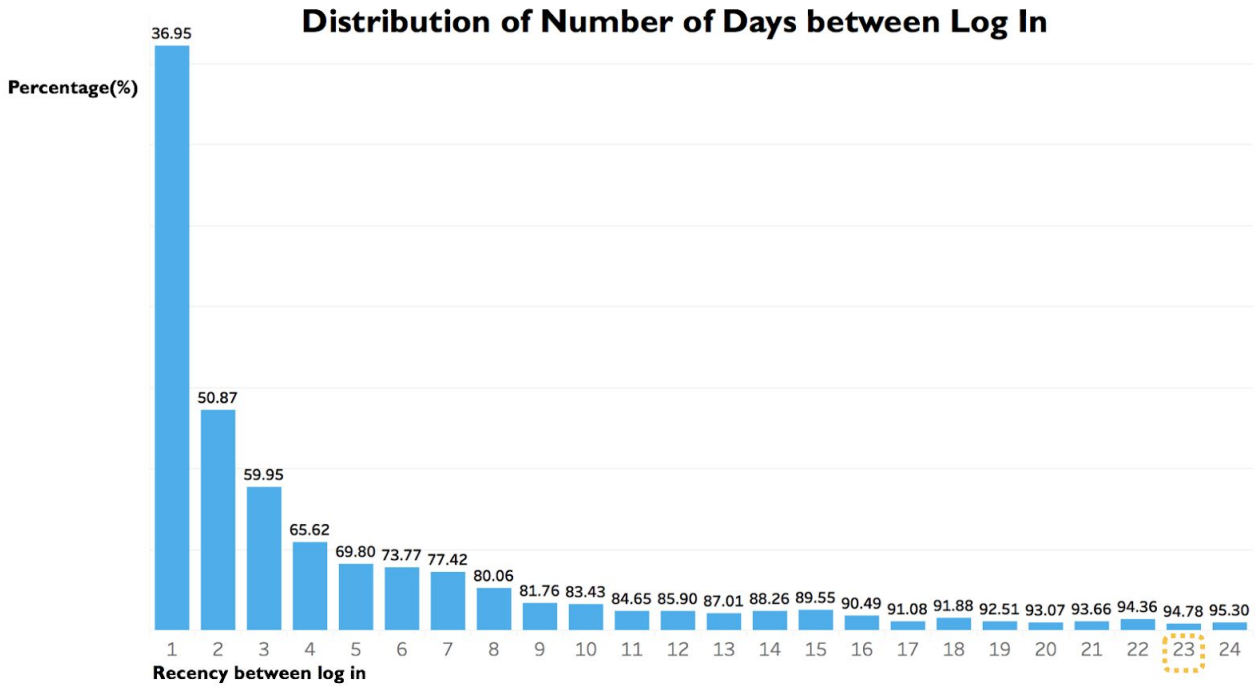
Question 6: Churn Analysis

1. Churn Definition

- Determine the Churn Threshold of players log in recency

- In order to avoid selection bias, we select the observation samples from May 2017 to August 2017, and calculates the log in recency of the active users during this time period (which is defined as if a person has activity on a certain date, then he/she is counted as active and logged in on this specific date).

- We visualize the number of days between consecutive logins for each player as below. The height of the bars represents the percentage of distribution of login intervals, while the text on top of the bars indicating the cumulative percentage since day 0 to day n.



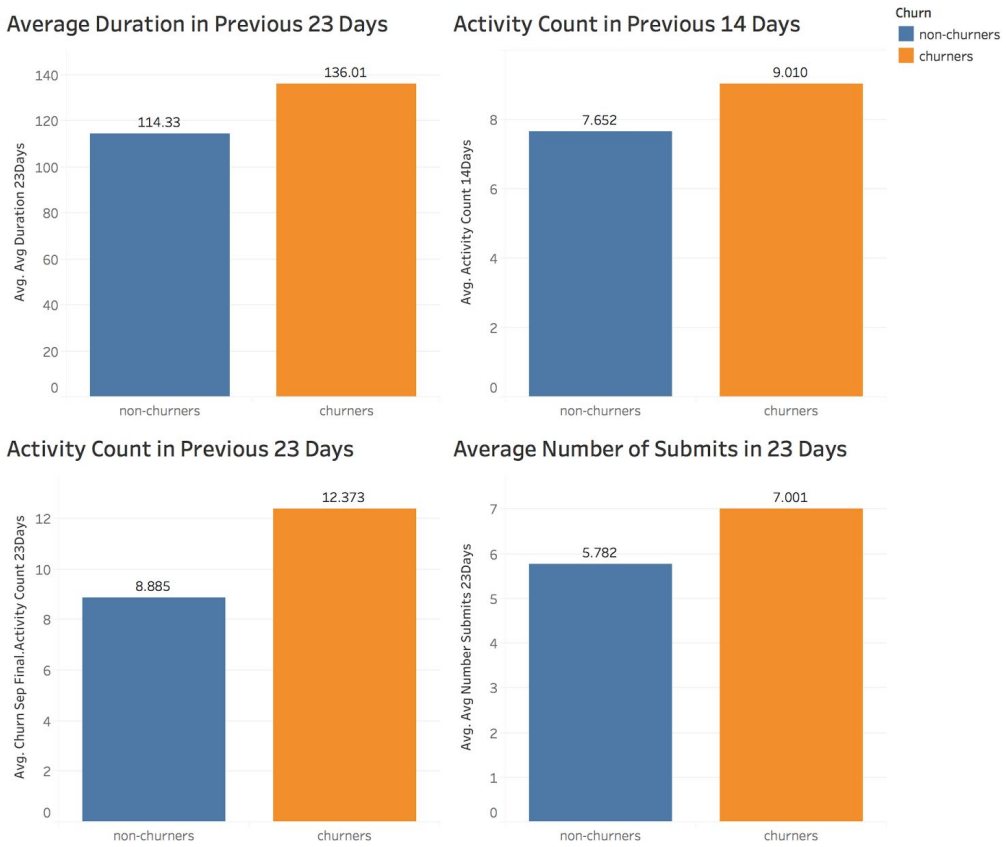
- By looking at the highlighted day 23, we can conclude as : If we observe a user who logged into the game 23 days ago and never come back, the probability that this user would never come back is about 95%.

- So, we cut the threshold on the 95% percentile bin, which is if a user in a certain time window has not logged in in 23 days, then he/she is considered as churn.

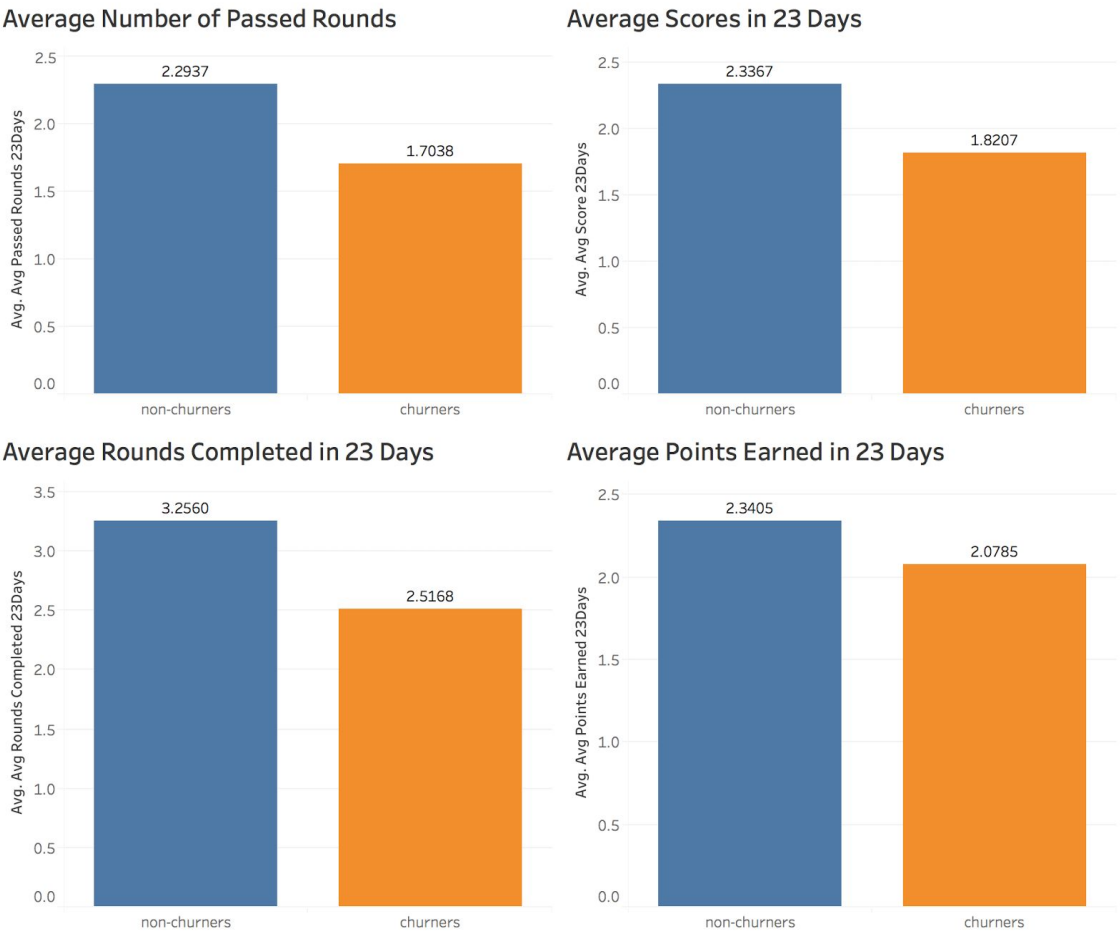
2. Exploratory Analysis on Churners VS. Non-Churners

- We constructed data preparing process in Apache Hive to obtain the aggregated variables for each user given a certain moving time window. And as the churn models in the industry is a dynamic process and get updated nearly real time, the following analysis are based on the dataset from September 2017. Each observation contains the average durations, activity counts, average number of submits, average number of passed rounds, average number of rounds completed, average points earned in previous 23 days from the current active date.

- Observing the contrast of two dashboards below, some interesting insights show up :



- To the contrast of our intuition, the behaviour of churners have been more active than non-churners. The average duration, activity count, and number of submits of churners in the recent 23 day time window are significantly higher than other non-churners.

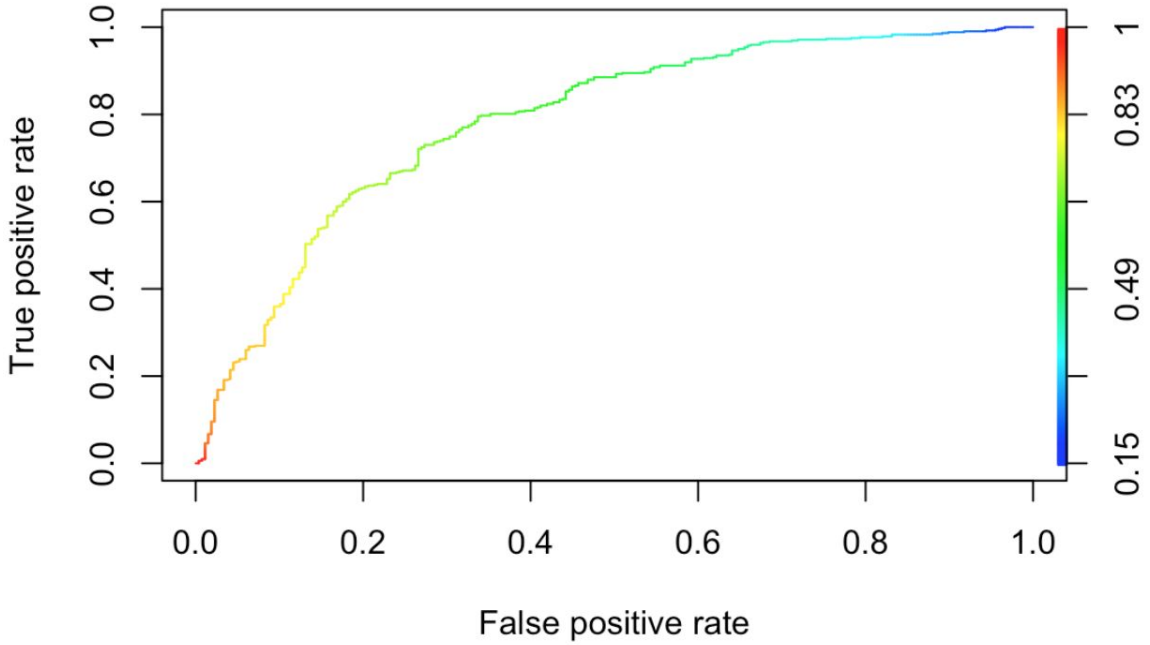


- However, the performance indicators of success during game playing set the two groups apart. Churners had lower scores, and lower number of passed rounds than non-churners. So we can assume in general, the key driver of customer churning is some players lose interests because they fail too much.

3. Churn Model Construction

- The total dataset contains 3950 active users in september 2017, and there are 11 explanatory variables and 1 dependent variable - whether this customer will churn (1 represents churners, 0 represent non-churners)

- Then, randomly sample the data into training(80%) and testing set (20%), and construct a logistic regression model and export the ROC curve, which is indicated as below:



- As the business purpose of the churn model is to capture as many potential churners as possible, so the company can take actions in advance to retain users, so the maximizing True Positive Rate (in this case it would be the ratio of number of actual churners to number of predicted churners) should be our prioritized metrics rather than classification accuracy. As the result, we finalize the probability cutoff threshold on 0.61 as the best candidate, and the ***true positive rate is 0.81 (successfully captures 81% of actual churners)***

- Logistic Regression Output:

```
Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                             1.544913   0.168204   9.185  < 2e-16 ***
activity_count_14days                  -0.042306   0.009531  -4.439 9.06e-06 ***
churn_sep_final.activity_count_23days   0.034417   0.008472   4.062 4.86e-05 ***
avg_duration_23days                     0.005372   0.000752   7.144 9.07e-13 ***
avg_score_23days                        0.001497   0.001521   0.984  0.32507
avg_passed_rounds_23days                4.747842  11.593814   0.410  0.68216
churn_sep_final.avg_failed_rounds_23days 4.810736  11.593689   0.415  0.67818
avg_rounds_started_23days              -1.710949   0.206626  -8.280  < 2e-16 ***
avg_rounds_completed_23days            -3.552323  11.592551  -0.306  0.75928
avg_number_submits_23days               0.029906   0.010720   2.790  0.00528 **
avg_points_earned_23days               -0.254131   0.058047  -4.378 1.20e-05 ***
avg_stars_earned_23days                 0.358899   0.080920   4.435 9.20e-06 ***
```

4. Result Analysis and Business Implications:

- The churn model captures 81% of actual churners, and the top indicators of potential churn are :

-

Negative Indicators on Churn	Positive Indicators on Churn
Activity count in previous 14 days	Average failed rounds in past 23 days
Avg # rounds started within 23 days	Average duration in 23 days
Avg points earned within 23 days	Average stars earned in 23 days

- *Recommendations:*

❖ *1. Identify potential churners from the churn predictive model, and take actions to lower the difficulty levels of the games to these players, and re-engage these high risk users by providing incentives, sending notifications and etc.*

❖ *2. The root cause of churn is players are disheartened by the difficulty in passing rounds. The churners have been even better engaged in the game than non-churners, but they gradually loose passion and ditched the game because they failed to gain confidence in cracking the questions.*

Question 7: Frequent User

1. Business Question

Who are the frequent players?

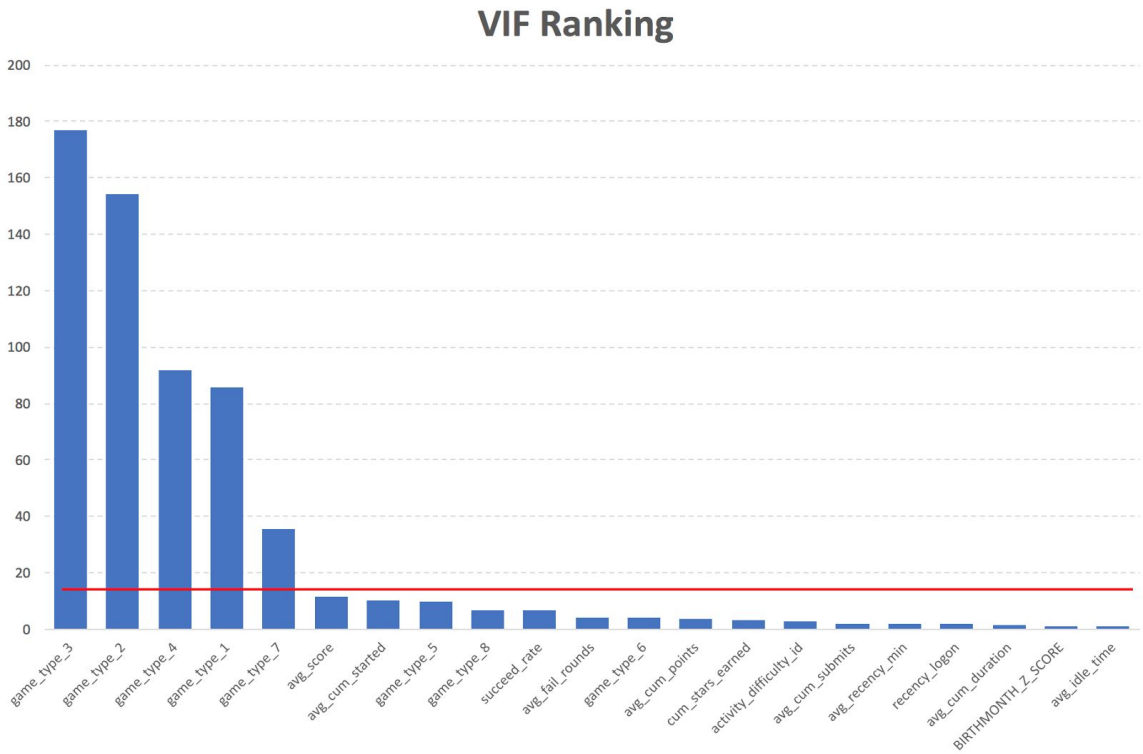
2. Exploratory Data Analysis

- To predict which users will come back most frequently, we define “frequency” as the average number of active days one user has 7 days before certain point. We choose the frequency in 7 days because it has a better distribution, compared to 3 days and 14 days. For the predictors, we aggregate the raw activity data into customer data. For example, we calculate the mean value for most of the variables; besides, for game type, we calculate the weights of the types for each user.

- First, we find two alias in our dataset, that are number of average complete rounds and incompletes rounds, since the number of complete rounds equals to the sum of number of pass rounds and fail rounds.

- Secondly, after deleting them, we further look into the correlation plot. Surprisingly, we find that the correlation of average scores and average pass rounds is 1, which means they are perfectly correlated to each other. There are also some pairs of the predictors have a high correlation. As a result, every time we delete one of the pair to get rid of the unuseful predictors.

- Thirdly ,we want to check VIFs to avoid collinearity. From the plot we can find 5 variables with a VIF larger than 10. Since all of them are from the game type group, all of the weights of other types will give a rough range for the rest one. In order not to eliminate important game type, we keep them for the next step.



3. Linear Model

(1) Multiple Linear Regression

- According to the outcome, if we take a 95% confidence level, we have 16 significant independent variables. They are: success rate, game type 3, game type 5, game type 6, game type 8, idle time, score, fail rounds, activity recency, birthday, log on recency, duration, submits, rounds started, points and difficulty.

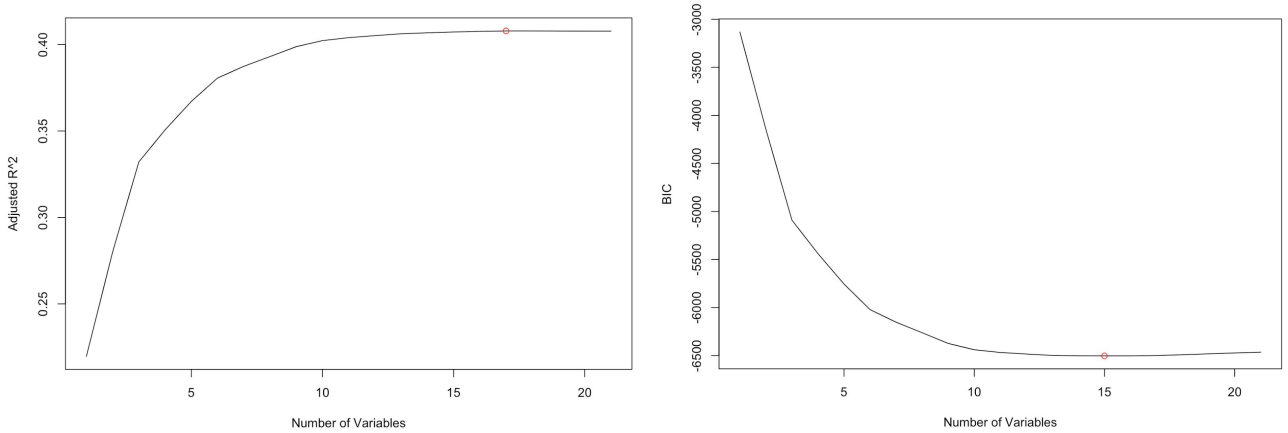
- For the model, we have the adjusted R squared 0.41, which means our model explains 41% of the variance. Additionally, the p value for F test is also very low, making our model more reliable.

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      1.666e+00  2.532e-01   6.582 4.83e-11 ***
succeed_rate                     4.583e-01  4.073e-02  11.254  < 2e-16 ***
game_type_1                     -2.772e-01  2.536e-01  -1.093 0.274275
game_type_2                     -4.156e-01  2.531e-01  -1.642 0.100680
game_type_3                     -6.914e-01  2.532e-01  -2.731 0.006331 **
game_type_4                     -2.742e-01  2.541e-01  -1.079 0.280558
game_type_5                      9.200e-01  2.682e-01   3.430 0.000606 ***
game_type_6                      1.264e+00  3.078e-01   4.108 4.02e-05 ***
game_type_7                     -2.733e-01  2.555e-01  -1.070 0.284763
game_type_8                     -6.967e-01  3.011e-01  -2.314 0.020679 *
avg_idle_time                    1.131e-05  4.610e-06   2.453 0.014176 *
avg_score                       -7.821e-02  1.096e-02  -7.137 1.00e-12 ***
avg_fail_rounds                 -4.767e-02  1.323e-02  -3.603 0.000316 ***
avg_recency_min                  1.758e-05  9.715e-07  18.097  < 2e-16 ***
BIRTHMONTH_Z_SCORE               2.592e-02  5.326e-03   4.867 1.15e-06 ***
recency_logon                   -2.118e-02  5.812e-04 -36.438  < 2e-16 ***
cumulative.avg_cum_duration      1.984e-04  7.455e-05   2.661 0.007797 **
cumulative.avg_cum_number_submits 1.127e-02  2.046e-03   5.510 3.66e-08 ***
cumulative.avg_cum_rounds_started -5.321e-02  1.072e-02  -4.965 6.95e-07 ***
cumulative.avg_cum_points_earned  -7.363e-02  7.461e-03  -9.869  < 2e-16 ***
cumulative.stars_earned           1.624e-02  1.788e-02   0.908 0.363706
activity_difficulty_id            2.838e-01  7.792e-03  36.418  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5296 on 12674 degrees of freedom
Multiple R-squared:  0.4088,    Adjusted R-squared:  0.4078
F-statistic: 417.2 on 21 and 12674 DF,  p-value: < 2.2e-16
```

(2) Best Subset Methodology

- In order to find the best model in terms of the lowest variance, we use best subset to get the best model for this question. Here we use adjusted R squared and BIC as two index to look into the model quality. We have the plots for different number of predictors we set:



- Since we have many predictor, we want to have a bigger penalty on the larger number of predictors. Based on the BIC plot, we choose 15 predictors to get the optimal model. In our model, all the variables are significant, the adjusted R squared is 0.41. We also have some insights in terms of predicting the users who will come back most frequently.

4. Insights

- Demographics: Older students come back more frequently.

- Game Type: It seems from the model that, people playing game type 5 and 6 are willing to come back more frequently. On the other hand, game type 2, 3 and 8 all have negative effect on the user frequency.

- User Behaviors: Number of submits leads to higher frequency, however, number of started leads to lower frequency. This might because if people start too many rounds, they might not want to come very frequently. Besides, people who play on a more difficult level tend to come back more frequently, maybe they still have the dream to pursue.

- Playing Performance: Some performance indicators will also affect user frequency. For instance, success rate has a positive effect, but average scores and points earned have negative effect on frequency. We think that if users have achieved a lot this time, they do not want to come back frequently; if they have a high success rate, equipped with better skills, they want to come back more.

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      1.403e+00  2.434e-02  57.641  < 2e-16 ***
succeed_rate                     4.712e-01  3.826e-02  12.317  < 2e-16 ***
game_type_2                     -1.406e-01  2.621e-02  -5.365 8.22e-08 ***
game_type_3                     -4.364e-01  2.524e-02 -17.289  < 2e-16 ***
game_type_5                      1.214e+00  9.193e-02  13.211  < 2e-16 ***
game_type_6                      1.589e+00  1.677e-01   9.474  < 2e-16 ***
game_type_8                     -3.899e-01  1.185e-01  -3.289 0.001008 **
avg_score                       -8.122e-02  1.080e-02  -7.520 5.83e-14 ***
avg_fail_rounds                 -4.870e-02  1.302e-02  -3.739 0.000185 ***
avg_recency_min                  1.767e-05  9.714e-07  18.192  < 2e-16 ***
BIRTHMONTH_Z_SCORE               2.764e-02  5.289e-03   5.225 1.77e-07 ***
recency_logon                   -2.125e-02  5.811e-04 -36.562  < 2e-16 ***
cumulative.avg_cum_number_submits  1.192e-02  2.026e-03   5.884 4.11e-09 ***
cumulative.avg_cum_rounds_started -5.084e-02  1.041e-02  -4.883 1.06e-06 ***
cumulative.avg_cum_points_earned  -7.271e-02  7.156e-03 -10.161  < 2e-16 ***
activity_difficulty_id           2.892e-01  6.647e-03  43.508  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5299 on 12680 degrees of freedom
Multiple R-squared:  0.4079,    Adjusted R-squared:  0.4072
F-statistic: 582.5 on 15 and 12680 DF,  p-value: < 2.2e-16
```

Question 8: Activity Outcome

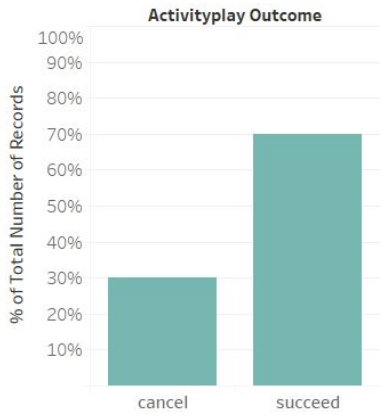
1. Business Question

Can we predict if a student will cancel or complete an activity?

2. Exploratory Data Analysis

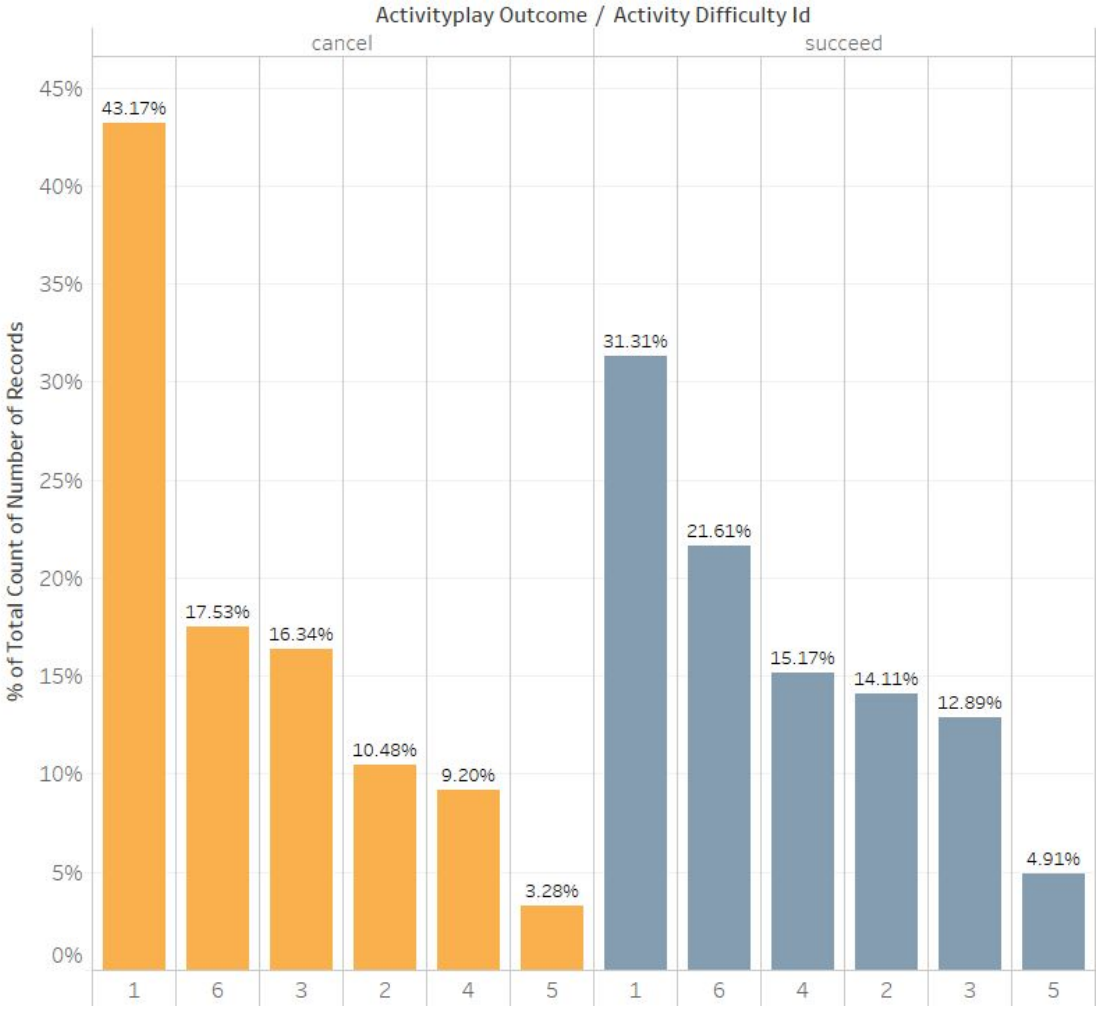
(1) Total Completion Rate

- In our dataset, 70% of total activities completed the game, and 30% failed to complete the game.



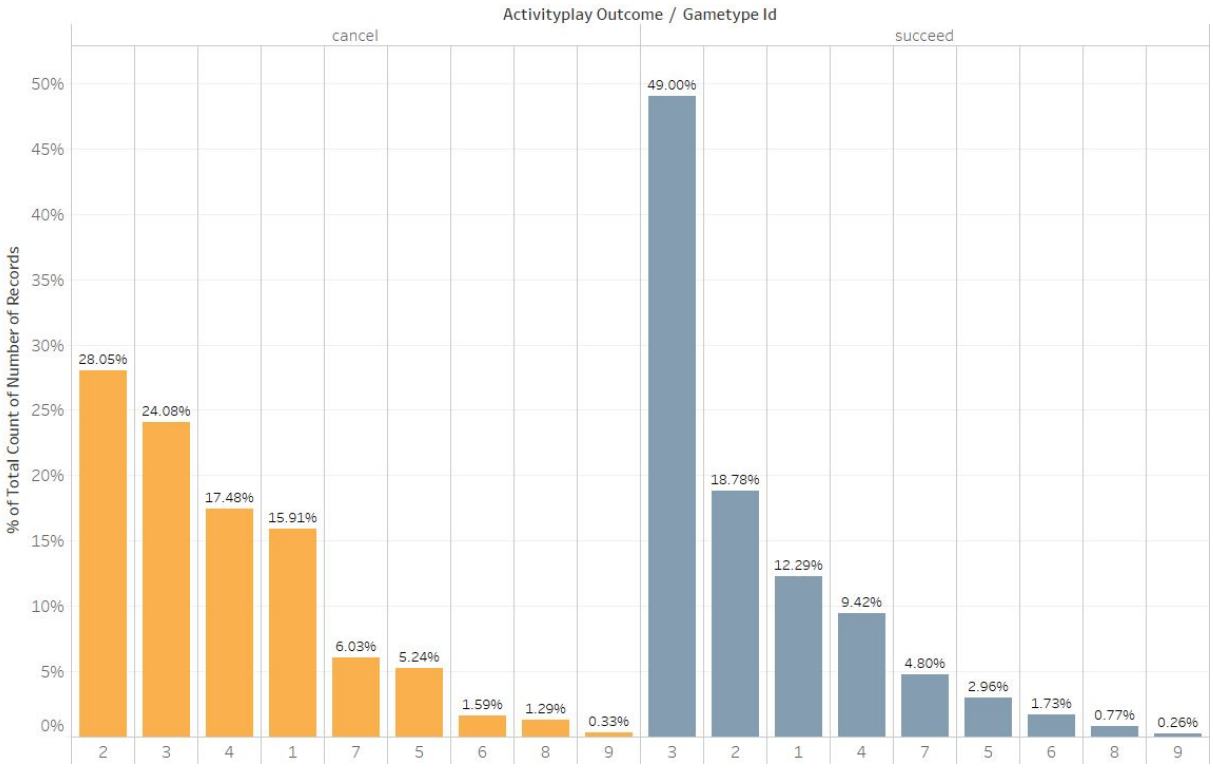
(2) Difficulty Level Distribution within Activity Outcomes

- For both outcomes, Difficulty Level 1 and 6 are the majority and Difficulty Level 5 is the least game type played.



(3) Game Type Distribution within Activity Outcomes

- For both outcome “cancel”, Game Type 2 is the most and Game Type 3 is the second.
- For both outcome “succeed”, Game Type 3 is the most and Game Type 2 is the second.



2. Modeling

(1) Preparation

- In order to predict if a certain student will complete or cancel an activity, we need to predict at activity level using variables that represent the student's playing behavior. Activity outcome (succeed or cancel) is our response variable to solve this question. Original variables like activity duration, score, number of rounds started, number of points earned, number of passed rounds, and so on, can not be generated until a certain game activity has been finished, so that they can not be valid inputs to predict the outcome of an activity. To deal with this problem, we transform these kinds of variables into average cumulative historical data for each specific student. As a result, these variables become good representatives for a student's playing behavior.

$$Avg\ Cumulative\ Variable_i^{student\ A} = \frac{\sum_1^{i-1} Variable^{student\ A}}{1+2+\cdots+i-1}$$

- We also substitute average cumulative number of passed rounds, average cumulative number of failed rounds, average cumulative number of rounds completed and average cumulative number of rounds incompleted with only 2 variables: average cumulative pass rate and average cumulative completion

rate. Relative value is more representative than absolute value in this case. In this way, we remove the effect that different number of rounds completed will have on the number of rounds passed.

- The final 15 predictors are as following:

Variables	Description
Activity_difficulty_id	Difficulty level for the activity; 6 levels
Gametype_id	Game type for the activity; 9 levels
Recency_min	Duration in minutes since the last time the student played the activity
Birthmonth_Z_Score	The distance a student's age from the mean age
Activity_count_3days	Frequency that the student played the activities in last 3 days
Activity_count_7days	Frequency that the student played the activities in last 7 days
Activity_count_14days	Frequency that the student played the activities in last 14 days
Avg_cum_duration	Average cumulative historical activity duration for the student
Avg_cum_score	Average cumulative historical adaptive scores for the student
Avg_cum_rounds_started	Average cumulative historical started rounds for the student
Avg_cum_number_submits	Average cumulative historical number of submits for the student
Avg_cum_points_earned	Average cumulative historical earned points for the student
Avg_cum_stars_earned	Average cumulative historical earned stars for the student
Avg_cum_pass_rate	Average cumulative historical pass rate for the student
Avg_cum_complete_rate	Average cumulative historical completion rate for the student

(2) Logistic Regression

- The first model we built is logistic regression model. We randomly split our data into 70% training data and 30% testing data using set.seed(1), and set a threshold of 50% to determine the potential class of each activity outcome. After running the model, the estimated coefficients for each variable and the model results are shown in the following:

```
Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -8.843e+00  1.408e-01 -62.795  < 2e-16 ***
activity_difficulty_id2         4.574e-01  2.284e-02  20.029  < 2e-16 ***
activity_difficulty_id3         1.640e-01  2.268e-02   7.234 4.70e-13 ***
activity_difficulty_id4         4.080e-01  2.478e-02  16.466  < 2e-16 ***
activity_difficulty_id5        -1.895e-01  3.543e-02  -5.349 8.86e-08 ***
activity_difficulty_id6        -4.295e-01  2.346e-02 -18.309  < 2e-16 ***
gametype_id2                    3.501e-02  2.085e-02   1.679 0.093212 .
gametype_id3                    1.239e+00  2.068e-02  59.917  < 2e-16 ***
gametype_id4                   -3.653e-01  2.501e-02 -14.605  < 2e-16 ***
gametype_id5                   -6.227e-01  3.363e-02 -18.514  < 2e-16 ***
gametype_id6                    7.156e-01  4.803e-02  14.900  < 2e-16 ***
gametype_id7                    1.000e-01  2.873e-02   3.482 0.000498 ***
gametype_id8                    4.812e-02  6.291e-02   0.765 0.444344
gametype_id9                    2.212e-01  1.046e-01   2.114 0.034512 *
recency_min                     4.964e-06  5.397e-07   9.198  < 2e-16 ***
BIRTHMONTH_Z_SCORE              3.157e-02  9.447e-03   3.342 0.000832 ***
activity_count_3days           -3.588e-03  5.793e-04  -6.194 5.87e-10 ***
activity_count_7days            1.155e-03  6.184e-04   1.867 0.061892 .
activity_count_14days           9.655e-04  3.191e-04   3.026 0.002481 **
cumulative.avg_cum_duration     8.144e-04  1.314e-04   6.198 5.70e-10 ***
cumulative.avg_cum_score       -2.498e-04  7.053e-06 -35.414  < 2e-16 ***
cumulative.avg_cum_rounds_started  8.643e-01  2.095e-02  41.261  < 2e-16 ***
cumulative.avg_cum_number_submits -1.076e-02  2.227e-03  -4.833 1.35e-06 ***
cumulative.avg_cum_points_earned   7.424e-01  1.618e-02  45.871  < 2e-16 ***
cumulative.stars_earned            6.043e-01  2.324e-02  26.005  < 2e-16 ***
avg_cum_pass_rate                  1.742e+00  9.512e-02  18.315  < 2e-16 ***
avg_cum_complete_rate              6.800e+00  1.455e-01  46.726  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Logistic Regression

Predict\Outcome	Cancel	Succeed
Cancel	12198	4211
Succeed	11452	52997
Classification Accuracy: 80.63%		
True Positive Rate: 92.64%		

- From the results, it appears that except for game type 2, game type 8, and activity_count_7days, all other variables are statistically significant in predicting the outcome. The classification accuracy for this model is 80.63%, and the true positive rate is 92.64%.

(3) LASSO

- Even though our logistic regression model already performs well, we decide to further conduct variable selection to enhance prediction accuracy. LASSO is chosen to be our method for variable selection.

- For logistic regression, the objective function for the penalized logistic regression uses the negative binomial log-likelihood, and is

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} -\left[\frac{1}{N}\sum_{i=1}^{N} y_i \cdot (\beta_0 + x_i^T\beta) - \log(1+e^{(\beta_0+x_i^T\beta)})\right] + \lambda[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1]$$

- We use 5-fold cross validation to choose the best tuning parameter that minimizes the cross-validation error. In our model, the best parameter turns out to be 0.0002169. However, none of the original coefficient estimates of 15 variables has been shrunk to 0. This means that our variables are all influential in determining the activity outcomes.
- After applying the model on the test data with a probability threshold of 50%, we get our result for LASSO model:

LASSO

Predict\Outcome	Cancel	Succeed
Cancel	12136	4177
Succeed	9168	53123
Classification Accuracy: 80.59%		
True Positive Rate: 92.70%		

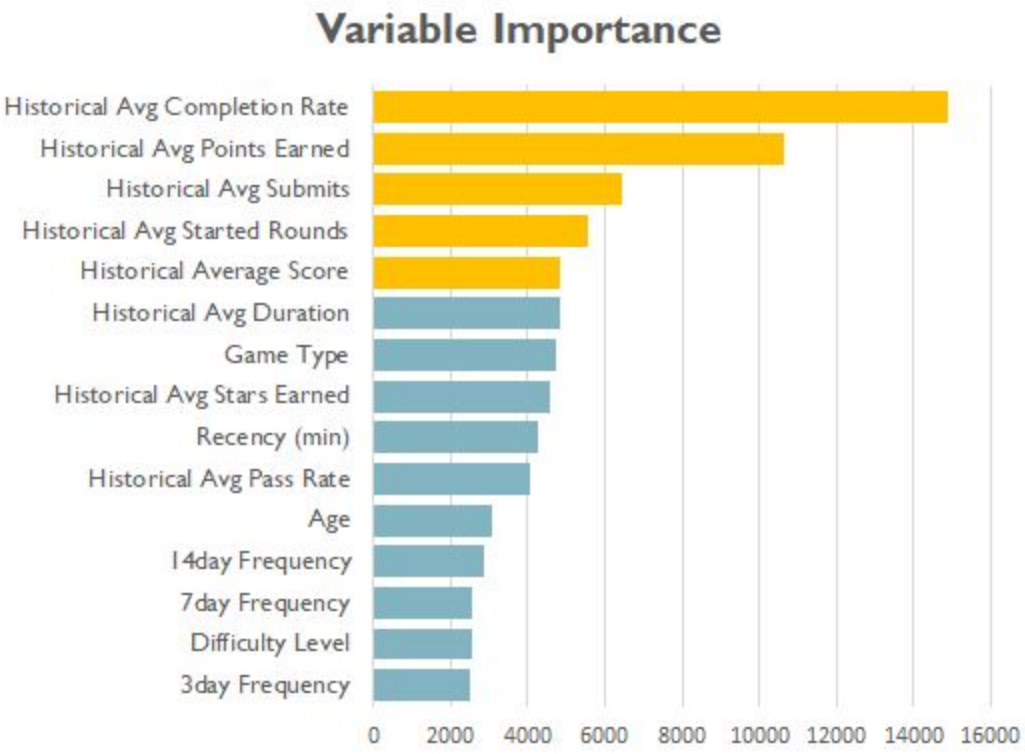
- The classification accuracy for this model is 80.59%, and the true positive rate is 92.70%.

(4) Random Forest

- Lastly, we conduct random forest model majorly to examine the importance of variables in determining the activity outcomes. We set 100 trees to grow and got the following results:

Random Forest

Predict\Outcome	Cancel	Succeed
Cancel	14482	4085
Succeed	9168	53123
Classification Accuracy: 83.61%		
True Positive Rate: 92.86%		



- The performance of Random Forest is even better than the previous two models. The classification accuracy is 83.61%, and the true positive rate is 92.86%. From the Variable Importance plot, we can see that the most important variable is average cumulative historical completion rate. The other top variables are average cumulative points earned, average cumulative submits, average cumulative rounds started, and average cumulative score. All the top variables are students' playing behavior variables that built on their historical playing activities.

Question 9：Passed Rounds Number

1. Business Question

If we knew a user started X rounds during the game, can we predict how many rounds they have passed?

2. Linear Regression Model

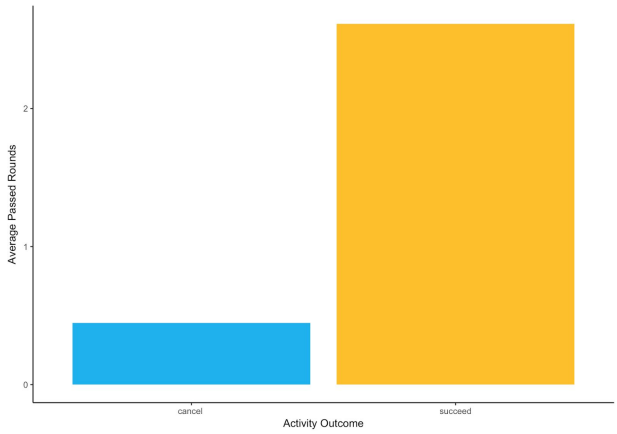
(1) Data Cleaning

- Number of passed round = 0 and activity play outcome is succeeded - 36 rows

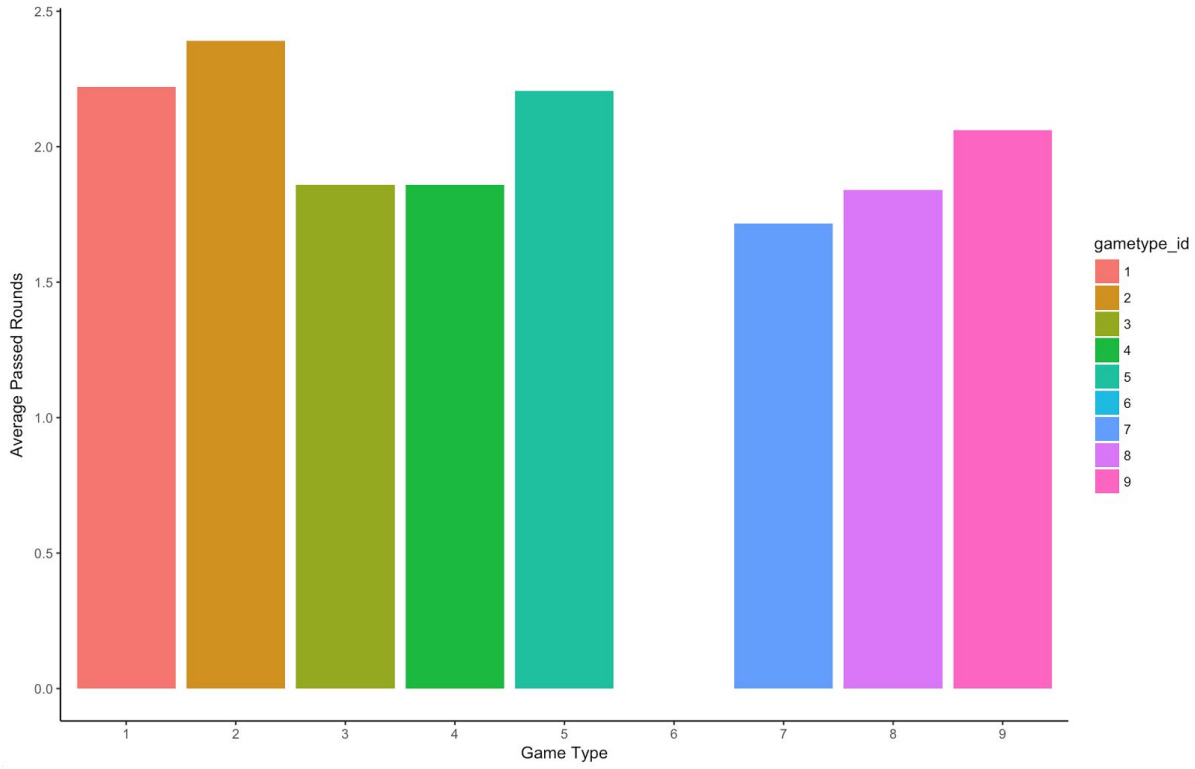
(2) EDA

A. Activity Outcome

- We can clearly see that if the outcome of activity was 'succeed', it would have much higher average passed rounds than 'cancel'.

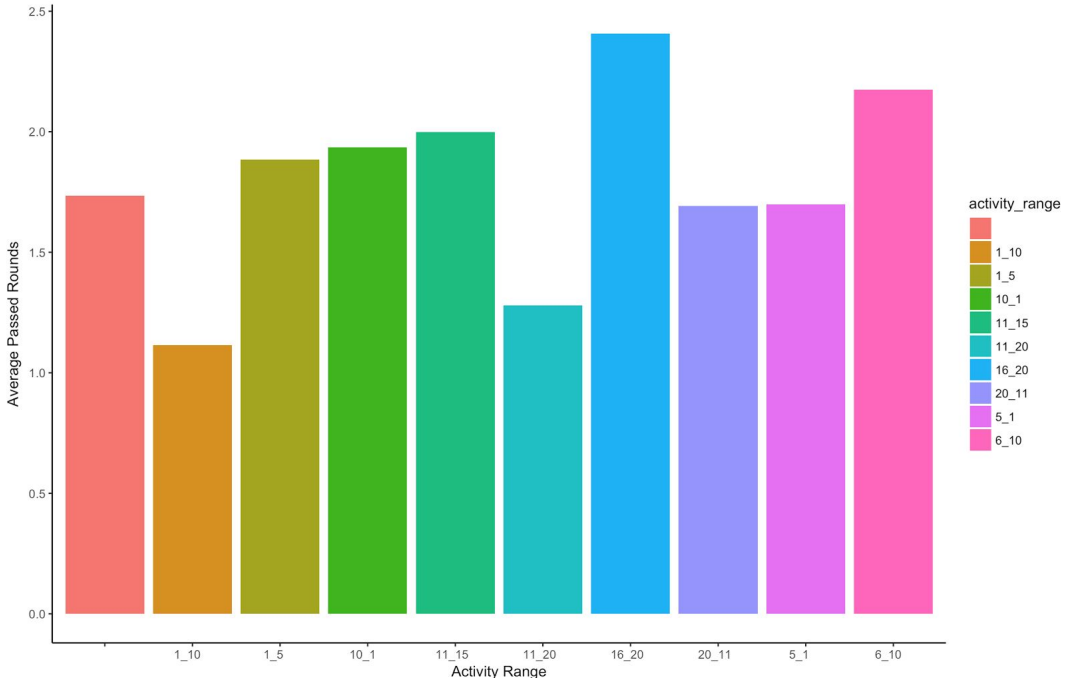


B. Game Type



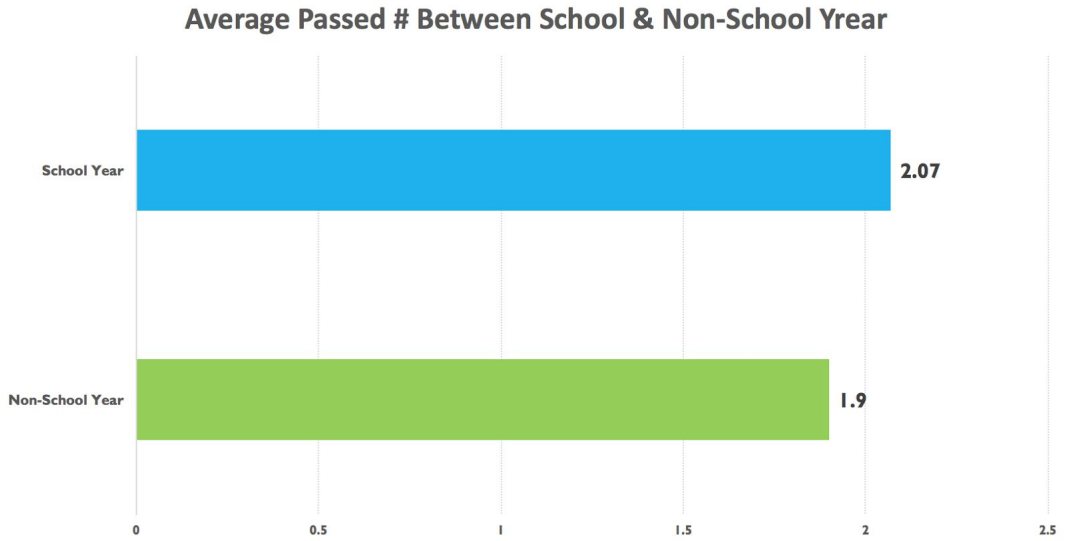
- There’re difference on average numbers of passed rounds among the various game type. It’s also very interesting that for game type 6, the passed rounds are all 0.

C. Activity Range



- Activity range might also be an important categorical factor to predict the passed rounds numbers

D. School Year



- As we mentioned before, games played in and not in school year are also statistically significant on the difference of passed rounds numbers.

(3) Initial Model

- Although we have detailed variables in this dataset, if we want to have a deeper look at the actual situation of predicting the number of passed rounds, we have to remove the independent variables, like number of failed rounds and adaptive scores(only 5778 activities had adaptive scores different from number of passed rounds)

- After the EDA, we decided to put categorical variables, like activity outcome, gametype_id, school and Activity_Range, and also numerical variables, like duration, number of rounds started, number of rounds completed, number of submitted, activity difficulty, frequency_3days, frequency_7days, frequency_14days, idle time, birthmonth Z score and recency, into our initial model. At the same time, when we predict the number of passed rounds, we cannot get information of points earned and stars earned for sure, so we did not to put them in initial model.

- Results:

```
Coefficients:
                              Estimate Std. Error  t value Pr(>|t|)
(Intercept)                 -7.316e-02  1.001e-02   -7.306 2.76e-13 ***
activity_duration           -6.364e-04  1.591e-05  -40.012  < 2e-16 ***
number_of_rounds_started     8.474e-02  8.163e-03   10.381  < 2e-16 ***
number_of_rounds_completed   7.254e-01  8.388e-03   86.481  < 2e-16 ***
number_of_submits           -3.921e-02  2.969e-04 -132.049  < 2e-16 ***
activityplay_outcomesucceed  6.179e-02  8.719e-03    7.087 1.37e-12 ***
activity_difficulty_id      -3.486e-03  1.363e-03   -2.558  0.01053 *
gametype_id2                 2.141e-02  7.384e-03    2.899  0.00374 **
gametype_id3                 2.537e-01  7.276e-03   34.871  < 2e-16 ***
gametype_id4                -2.741e-01  8.635e-03  -31.741  < 2e-16 ***
gametype_id5                 9.337e-02  1.208e-02    7.727 1.11e-14 ***
gametype_id6                -7.126e-01  1.604e-02  -44.429  < 2e-16 ***
gametype_id7                 3.827e-01  1.061e-02   36.079  < 2e-16 ***
gametype_id8                 5.606e-01  2.334e-02   24.022  < 2e-16 ***
gametype_id9                 3.138e-01  3.658e-02    8.580  < 2e-16 ***
activity_count_3days         1.549e-04  1.829e-04    0.847  0.39711
activity_count_7days         2.046e-04  1.920e-04    1.066  0.28656
activity_count_14days       -9.093e-07  9.507e-05   -0.010  0.99237
idle_time                   -3.940e-06  9.490e-07   -4.151 3.30e-05 ***
birthmonth_Zscore           -1.602e-01  2.681e-03  -59.758  < 2e-16 ***
schoolschool                 6.386e-02  4.258e-03   14.997  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.008 on 238353 degrees of freedom
  (31150 observations deleted due to missingness)
Multiple R-squared:  0.7354,    Adjusted R-squared:  0.7354
F-statistic: 3.312e+04 on 20 and 238353 DF,  p-value: < 2.2e-16
```

- As we can see from the results of our initial model, we get Adjusted R^2 at 73.54%. Namely, our model can capture most of variances for predicting the number of passed rounds.

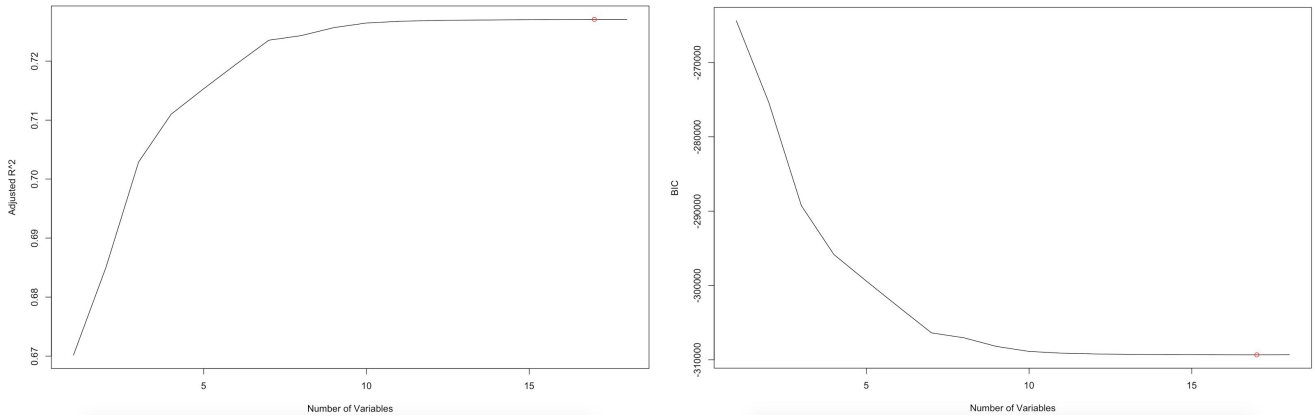
- Among all the predictors, except for frequency in 3 different time windows, all other independent variables are statistically significant for the dependent variables.

- VIF: As we can see the results of VIFs(checking the colinearity of variables), we can see number of started rounds, number of completed rounds and activity_count_7days have VIF larger than 10. This shows we might have collinearity problems in this model. This is because number of started rounds is highly correlated with number of completed rounds and 3 day frequency also is correlated with 7 day frequency. Since activity_count_7day is not statistically significant so we decided to remove that since it's positive correlated with activity_count_3day. Also, number of completed rounds is positively correlated with number of started rounds. In the more realistic scenario, we probably can't get the information like number of completed rounds in order to predict how many rounds the user passed. Therefore, we decided to remove number of completed rounds in our model.

	GVIF	Df	GVIF^(1/(2*Df))
activity_duration	1.364597	1	1.168160
number_of_rounds_started	62.580734	1	7.910799
number_of_rounds_completed	75.483322	1	8.688114
number_of_submits	1.533547	1	1.238365
activityplay_outcome	3.743533	1	1.934821
activity_difficulty_id	1.599359	1	1.264658
gametype_id	2.231487	8	1.051446
activity_count_3days	5.013409	1	2.239064
activity_count_7days	15.127858	1	3.889455
activity_count_14days	8.422768	1	2.902201
idle_time	1.007993	1	1.003989
birthmonth_Zscore	1.044682	1	1.022097
school	1.056674	1	1.027947

(4) Best Subset Selection

- After exploring the relationship between our dependent variable - number of passed rounds and all the independent variables from the initial linear regression model, we decided to use best subset selection model to build a more accurate model and strike a balance between model complexity and prediction accuracy.



- Adjust R^2 and BIC are the metrics that balance the model accuracy and model complexity. From the graphs above, we can see from both those two metrics results, the model contains 17 most important independent variables are the 'best' model that maximize the prediction accuracy but at the same time minimize the model complexity.

Negative Correlated Predictors	
Activity Duration	-0.001
Number of Submits	-0.039
Activity Difficulty ID	-0.015
Game Type 4	-0.28
Game Type 6	-0.791
Activity Frequency 3Days	-0.036

Positive Correlated Predictors	
Number of Started Rounds	0.780
Activity Outcome - Succeed	0.624
Game Type 2	0.034
Game Type 3	0.255
Game Type 5	0.083
Game Type 7	0.370
Game Type 8	0.571
Game Type 9	0.257
School	0.071

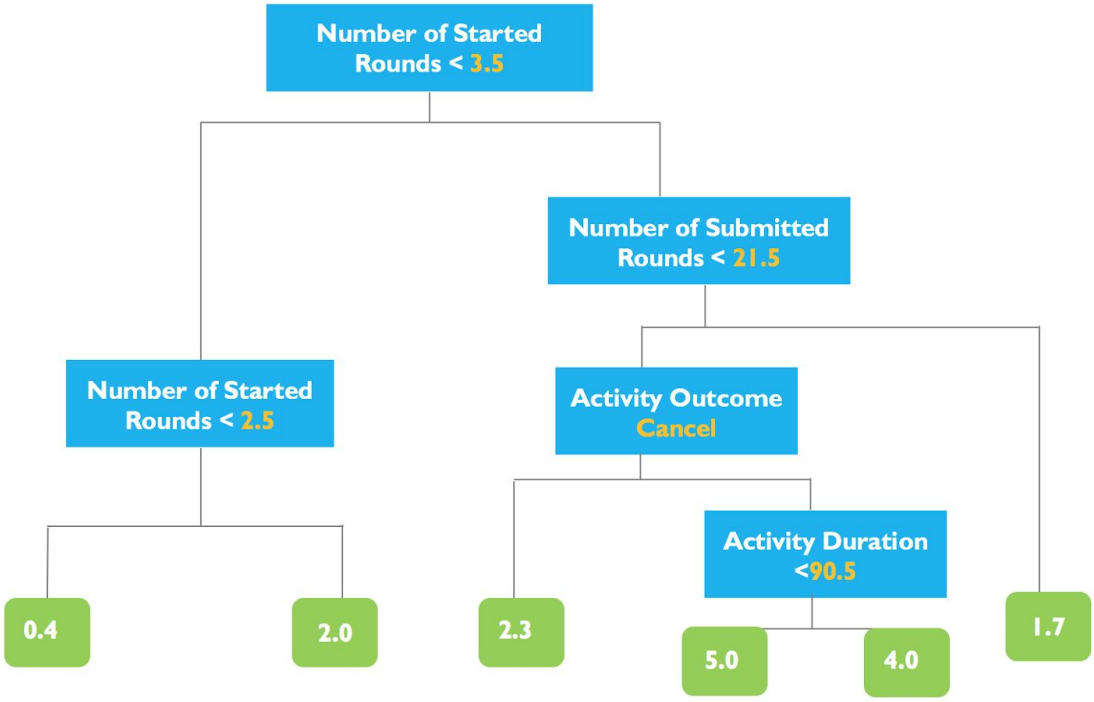
- As we can see the coefficients from best subset selection model. Activity Duration, Number of Submits, Activity Difficulty ID, 3 Day Frequency, Birth Month and Game type 4&6 have the negative relationship with number of passed rounds. This means if user stayed more time on the game and the submit number is higher and play more difficult game type, they are less likely to get a higher passed rounds number. Also, if they are older students and played the game more frequent in the past 3 days, or if if they game type 4 and 6, they are likely to get less passed rounds.

- However, if the user have more started rounds, get a 'succeed' for the final activity outcome, play games during the school year, they are more likely to

get a higher passed rounds number. Meanwhile, if they choose to play game type 2,3,5,7,8,9, they are also more likely to get a higher passed rounds number.

(4) Tree-Based Model

- Decision Tree



- From the results of this pruned decision tree model, we can see that Number of started rounds, number of submitted round, activity duration and activity outcome are the significant factors to predict the passed rounds number.

- Number of started rounds and and Succeed activity had the positive relationship with passed rounds number; Number of submitted rounds and activity duration had negative relationship with number of passed rounds.

- For example, if a user started more than 3.5 rounds but submitted less than 21.5 rounds; at the same time, you got 'Succeed' outcome and spent less than 90.5 seconds on the game. This user, from the tree model, can passed 5 rounds during the game in average.

Appendix

Codes

(1) Recency, Frequency, Cumulative Variables Creation:

In this section, we illustrate some snapshots of the steps of recency, frequency, and aggregated features building.

- Step 1: Import Activity data into **Apache Hive**, and create the table Activity.
- Step 2: Take activity_timestamp_start as time marker, and create aggregate features group by member_id, and arrange the sequence of rows for each member by time.
- Step3: Build several trailing time windows, in order to get the aggregate features. For example, ***activity_count_7days indicates how many activity does user X generate in the recent week on yyyy-mm-dd; historical_avg_score indicates the average score for user X since their first activity until the current date yyyy-mm-dd.***
- Step4: In order better understand churns, we create another table named log_on, and use the similar approach to build recency between two consecutive log in date for each user.
- Below is a screenshot of the code:

```
    > Select member_id, activity_id, activity_play_id, activity_timestamp_start,
    >
    > from_unixtime(unix_timestamp(activity_timestamp_start,'yyyy-MM-dd' 'hh:mm:ss')
,'yyyy-MM-dd') AS start_date,
    >
    > row_number() OVER (partition by member_id order by
    > from_unixtime(unix_timestamp(activity_timestamp_start,'yyyy-MM-dd' 'hh:mm:ss')
,'yyyy-MM-dd') desc ) AS most_recent_record ,
    >
    > count(activity_play_id) over (partition by member_id order by unix_timestamp(a
ctivity_timestamp_start)
    >                                range between 1209600 preceding and current r
ow )AS activity_count_14days,
    >
    > count(activity_play_id) over (partition by member_id order by unix_timestamp(a
ctivity_timestamp_start)
    >                                range between 1987220 preceding and current r
ow )AS activity_count_23days,
    >
    > avg(activity_duration) over (partition by member_id order by unix_timestamp(ac
tivity_timestamp_start)
```

(2) Predictive Modeling- R Codes

- **Customer Churn Prediction Model**

```
set.seed(1)

head(churn_final)
data1=data.frame(churn_final[,c(18,7:17)])
str(data1)
data1$avg_score_23days=as.numeric(data1$avg_score_23days)

data1$churn=as.numeric(data1$churn)

test = sample(1:nrow(data1), nrow(data1)*0.2)
test.data=data1[test,]
train.data=data1[-test,]
head(test)

logistic_model = glm( churn ~ ., data = train.data, family =
"binomial")
summary(logistic_model)
```

```
logistic_probs = predict(logistic_model, test.data, type =
"response")

###roc
library(ROCR)

#ROCRpred<-prediction(pred,obs)
#plot(performance(ROCRpred, measure = 'tpr', x.measure = 'fpr'))

ROCpred=prediction(logistic_probs, test.data$churn)
rocrPERF=performance(ROCpred,'tpr','fpr')
plot(rocrPERF,colorize = T, text.adj=c(-0.2,1.7))

###set threshold
cutoff=data.frame(cut=rocrPERF@alpha.values[[1]],
                  fpr=rocrPERF@x.values[[1]],
                  tpr=rocrPERF@y.values[[1]])

cutoff[cutoff$tpr>0.80 & cutoff$fpr<0.5,]
###cutoff 0.61

logistic_pred_y = rep(0, length(logistic_probs))
logistic_pred_y[logistic_probs >= 0.59] = 1

accuracy(logistic_probs,test.data$churn)
```

- **Frequent Players Prediction**

```
summary(log)
sapply(modeldata, function(x) sum(is.na(x) | x==""))
new = modeldata %>%
  group_by(member_id) %>%
  summarise(recency_min = mean(recency_min),
            activity_count_3days = mean(activity_count_3days),
            activity_count_7days = mean(activity_count_7days),
            activity_count_14days = mean(activity_count_14days),
            cumulative.avg_cum_duration =
mean(cumulative.avg_cum_duration),
            cumulative.avg_cum_number_submits =
mean(cumulative.avg_cum_number_submits),
            cumulative.avg_cum_rounds_started =
mean(cumulative.avg_cum_rounds_started),
            cumulative.avg_cum_points_earned =
mean(cumulative.avg_cum_points_earned),
```

```
            cumulative.stars_earned = mean(cumulative.stars_earned),
            activity_difficulty_id = mean(activity_difficulty_id))

data = cbind(member_aggregate, log, new)

df = data[,c(2, 9:16,18:23,26:28,32,34,39,45:50,37)]
sapply(df, function(x) sum(is.na(x) | x==""))
### birthday/ 2 recency have NAs === 19-21

Med = function(n){
  n[is.na(n)] = median(n ,na.rm = T)
  n
}

df[,c(19:21)]= apply(df[,c(19:21)], 2, Med)

###Check Corr
corrplot.mixed(cor(df), upper = "ellipse")
alias(lm(X7days ~ . , data = df))

df = df[,c(-10, -13, -c(15:18))]

vif = data.frame(vif(lm(X14days ~., data = df)))

lm = lm(X7days ~ ., data = df)
summary(lm)


## Not sig: diff_4_pert, diff_5_pert, avg_idle_time

#### best subset
library(leaps)
regfit.full = regsubsets(X7days ~ .,
                         data=df,nvmax = 21)
regfit.summary=summary(regfit.full)


plot(regfit.summary$adjr2,xlab = "Number of Variables", ylab =
"Adjusted R^2", type = "l")
points(which.max(regfit.summary$adjr2),regfit.summary$adjr2[which.max
(regfit.summary$adjr2)],col="red")

plot(regfit.summary$bic, xlab = "Number of Variables", ylab = "BIC",
type = "l")
```

```
points(which.min(regfit.summary$bic),regfit.summary$bic[which.min(reg
fit.summary$bic)],col="red")


rownames(data.frame(round(coef(regfit.full,15),3)))
lm2 = lm(X7days ~ succeed_rate+game_type_2+game_type_3+game_type_5+

game_type_6+game_type_8+avg_score+avg_fail_rounds+avg_recency_min+

BIRTHMONTH_Z_SCORE+recency_logon+cumulative.avg_cum_number_submits+

cumulative.avg_cum_rounds_started+cumulative.avg_cum_points_earned+
           activity_difficulty_id, data = df)
summary(lm2)

data.frame(lm2$coefficients)
```

- **User Passed Round Number Prediction**

```
### Initial Linear Regression Model
data$gametype_id = as.factor(data$gametype_id)
library(MASS)
lm.fit =
lm(number_of_passed_rounds~activity_duration+number_of_rounds_started
+number_of_rounds_completed+

number_of_submits+activityplay_outcome+activity_difficulty_id+gametyp
e_id+

activity_count_3days+activity_count_7days+activity_count_14days+idle_
time+birthmonth_Zscore+school,
            data=data)

summary(lm.fit)
library(car)
vif(lm.fit)



### Second Model

lm.fit2 =
lm(number_of_passed_rounds~activity_duration+number_of_rounds_started
+
```

```
number_of_submits+activityplay_outcome+activity_difficulty_id+gametyp
e_id+

activity_count_3days+activity_count_14days+idle_time+birthmonth_Zscor
e+school,
             data=data)

summary(lm.fit2)

### Best Subset Selection

library(leaps)

regfit.full =
regsubsets(number_of_passed_rounds~activity_duration+number_of_rounds
_started+

number_of_submits+activityplay_outcome+activity_difficulty_id+gametyp
e_id+

activity_count_3days+activity_count_14days+idle_time+birthmonth_Zscor
e+school,
                          data=data,nvmax = 18)
regfit.summary=summary(regfit.full)


plot(regfit.summary$adjr2,xlab = "Number of Variables", ylab =
"Adjusted R^2", type = "l")
points(which.max(regfit.summary$adjr2),regfit.summary$adjr2[which.max
(regfit.summary$adjr2)],col="red")

plot(regfit.summary$bic, xlab = "Number of Variables", ylab = "BIC",
type = "l")
points(which.min(regfit.summary$bic),regfit.summary$bic[which.min(reg
fit.summary$bic)],col="red")


round(coef(regfit.full,17),3)


### Tree-based Model

train = sample(nrow(data),nrow(data)/2)
library(MASS)
```

```
library(tree)
tree.game =
tree(number_of_passed_rounds~activity_duration+number_of_rounds_start
ed+

number_of_submits+activityplay_outcome+activity_difficulty_id+gametyp
e_id+

activity_count_3days+activity_count_14days+idle_time+birthmonth_Zscor
e+school,data=data)

summary(tree.game)

plot(tree.game)
text(tree.game,pretty = 0)

# Prune the model

cv.game = cv.tree(tree.game)
plot(cv.game$size,cv.game$dev,type = "b")
prune.game = prune.tree(tree.game,best = 6)
summary(prune.game)
plot(prune.game)
text(prune.game,pretty = 0)


## Random Forrest
library(randomForest)

data_rf = na.omit(data)
rf.game =
randomForest(number_of_passed_rounds~activity_duration+number_of_roun
ds_started+

number_of_submits+activityplay_outcome+activity_difficulty_id+gametyp
e_id+

activity_count_3days+activity_count_14days+idle_time+birthmonth_Zscor
e+school,data=data_rf,
                       mtry=4,importance=T)

varImpPlot(rf.game)
```