[**Executive Summary**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.eflfm09fsrz3)[**2**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.eflfm09fsrz3)

[**Introduction**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.txczudurctoc)[**2**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.txczudurctoc)

[**Problem Formulation**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.r5os4tyouxtf)[**3**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.r5os4tyouxtf)

[**Data Description**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.bo7q4d7fsc5)[**4**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.bo7q4d7fsc5)

[**Model Development**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.obdip7v666i6)[**4**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.obdip7v666i6)

[**Results**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.p3ox5mt2vp46)[**5**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.p3ox5mt2vp46)

[**Recommendations and Implications**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.xk4uimycslm5)[**5**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.xk4uimycslm5)

[**Conclusion**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.vls7u4291fm4)[**6**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.vls7u4291fm4)

[**Reference**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.clrn1ebvx2ki)[**7**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.clrn1ebvx2ki)

[**Appendix**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.dshyy8vjn71p)[**8**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.dshyy8vjn71p)

[**Model Development**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.obdip7v666i6)

1. Has the online community increased user revenue?

To analyze whether the online community has increased user revenue, we divided users into two groups, those who have joined the community and those who haven’t. We did this in order to exclude immeasurable effects (mainly the effect of natural growth of users’ spending across time) from the true effect of them joining the community. We used the Difference in Difference technique to run the regression, and extracted the difference in user revenue caused by their joining decision only.

**Results**

The result shows that the online community has a positive effect on user revenue. If a user joins the community, it will lead to an increase of $29 more spending in the following month The 29 dollars is solely caused by the decision of joining the community.

[**Recommendations and Implications**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.xk4uimycslm5)

From the regression model, we can tell the online community can improve user revenue in the short term. This means when users first enter the community, they are likely to be stimulated by the incentives and spend more on the game. But whether this effect will last longer and produce more user value is still to be discussed.

**Finding the Difference in the Revenue of the Two groups**

The first approach was to analyze the effect of the online community on revenue by comparing and finding the difference in the average change overtime (month) of revenue from those who joined the online community, compared to the average change overtime (month) from those who did not join. The Difference in Difference (DID) model allows us to compute this comparison/difference. *According to Wooldridge DiD is an experimental research in which “outcomes are observed for two groups for two time periods. The second group is not exposed to the treatment during either period. In the case where the same units within a group are observed in each time period, the average gain in the second (control) group is subtracted from the average gain in the first (treatment) group.”* In this case, we assume the two groups have similar spending patterns in the month before the community launch. One of them does join the online community and the other does not. We used the DID model to extract the difference in user revenue to quantify the difference and see what the difference could be.

The results show that gamers who joined the online community increased their spending by $29 for the month that followed (Exhibit 1). The results show that launching the online community will have a positive effect on revenue if gamers decide to join. From this we see that there is a short-term increase in revenue which means when gamers joined the community are more likely to be stimulated to spend. In the following sections we discuss whether this effect last in the long term.

Wooldrige, Imbens. Lecture Notes 10, Summer ‘07. <https://www.nber.org/WNE/lect_10_diffindiffs.pdf>

**(2) Has the online community led to increased retention?**

**Model**

A logistic regression model is used to detect whether joining the online community can lead customers to retain on the game within a three-month timeframe. Besides joining the online community, the ages of customers and the spendings of customers are two potential factors that impact the retention rate. In the logistic regression model, we use whether the customer joined the online com  
munity, the age of the customer, and the customer’s spend to predict the churn rate.

**Result**

The result of the logistic regression model shows that, based on the given data, the age of the customer and the customer spending do not have proved impact on the churn rate. In contrast, the online community does impact on the churn rate. However, the result shows the online community has a negative impact on retaining customers.

**Recommendations and Implications**

With the result, the company should conduct further study on the reason why customers who join the online community are more likely to churn after three months. The company can repeat the model on different data samples to do a robust test. If the model is robust, the word of mouth is one potential factor that needs to be monitored in the online community. The further study would be whether the online community distract customers from playing the game. More other factors that may impact the churn should be discovered. Model 2:

Our first model showed an increase in revenue in the short term. As a follow up we wanted to know if the online community led to increased retention. Maintaining a strong customer base is important for success of a business. According to Kumar (2015), customer retention refers to the act of engaging customers in the continues purchase of a company’s products or services. To find out the impact on the retention we run a logistic regression which allowed us to analyze whether the online community could increase retention. Besides the factor of joining the online community, we noticed that the customer's age and spending are factors that could potentially impact the retention rate. Therefore, we added all three factors to predict the churn rate.

The results showed that the age and spending did not impact the churn rate, but the online community does impact the churn rate. the online community has a negative impact on retaining customers Percentage?. The result showed that gamers who joined the online community where more likely to churn after three months. We recommend further studies to analyze what variables negatively affect retention. Maybe gamers don’t see in value the online community. Kumar (2015) added, “if companies don’t give customers good reason to stay, their competitors will.” One way to avoid customer churn is to do a survey with both groups and understand what they need. There are a limitation found in this study that will be discussed below.

**(3) Has the online community led to an increase in CLV? (*Assume a margin of 50% of customer spend*)**

**Model**

To understand the relationship between the online community and CLV, a linear regression model is conducted here for this question. To compute the value of each customer’s CLV, an equation has been deployed detailed in Figure.X. In the linear regression model, it uses whether the gamer joined the community as an indicator and detect how this factor contributed to the CLV of users.

[**Results**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.p3ox5mt2vp46)

The result of the linear relationship model between the online community and the CLV is shown in Figure.X. As shown in the results, it is statistically significant to conclude that there is a positive relationship between the online community and the CLV. In all, the online community would lead to an increase in CLV. This result is also consistent with another test in Appendix XX, which confirmed the joined group had a higher CLV than unjoined group.

[**Recommendations and Implications**](https://docs.google.com/document/d/1njbd3EDm9jGE-Sh60mlGyGbRr0HFSwVM/edit#heading=h.xk4uimycslm5)

Although the result shows the online community would lead to an increase in CLV, this correlation does not mean causality. There are two factors that would harm (don’t know how to describe….) the result in this case. One factor is the joined customer is already a heavy user, they tend to spend more than the others whether there is an online community or not. The other factor is that even though there is a positive correlation within these two factors, how large is the gap between the revenue generated by a joined customer and a non-joined customer is still questionable. In the real-world case, there are more factors could? to affect CLV. A multi variables with CLV model need to be conducted to clear the question.

**Model 3:**

Ideally when customer retention increases the customers’ lifetime value and revenue also increase. KyngaCell’s data analysis show that their retention rate decreases, so we want to know if the online community led to an increase or decrease in the Customers’ Lifetime Value (CLV). We used the single linear regression for this model. In the linear regression model, we use whether the user joined the community as an indicator and detect how this factor contributed to the CLV of users.. The analysis of the CLV is done by using the sum of the three months divided by the margin cost of 50% of customer spend (Appendix X - results).

The results show that it is statistically significant to conclude that there is a positive relationship between the online community and the CLV. The results also show that the CLV of the gamers who joined the online community was $46 higher than of those who did not join. Overall, this means the online community leads to an increase in CLV. This result is consistent with others results shown in Appendix XX which confirms that those who joined have a higher CLV. Even though, there seems to be a correlation between the online community with the increase in CLV, correlation does not mean causality. For example, it is highly possible that the users who joined the community are the existing hardcore users. They joined the community because they are hardcore users. If there is no community launched, they would also spend more than ordinary users anyway.

**Appendix**

Exhibit 1

Model 1 Difference in difference approach and the Results

In the model, we assumed that users, whether joined the community or not, have parallel trends in spending over a short period of time. We divided users into two groups. Those who have joined the community went to the treatment group, and those who haven’t were put in the control group. Then we compared the difference in spending between two groups using the Difference in Difference approach.

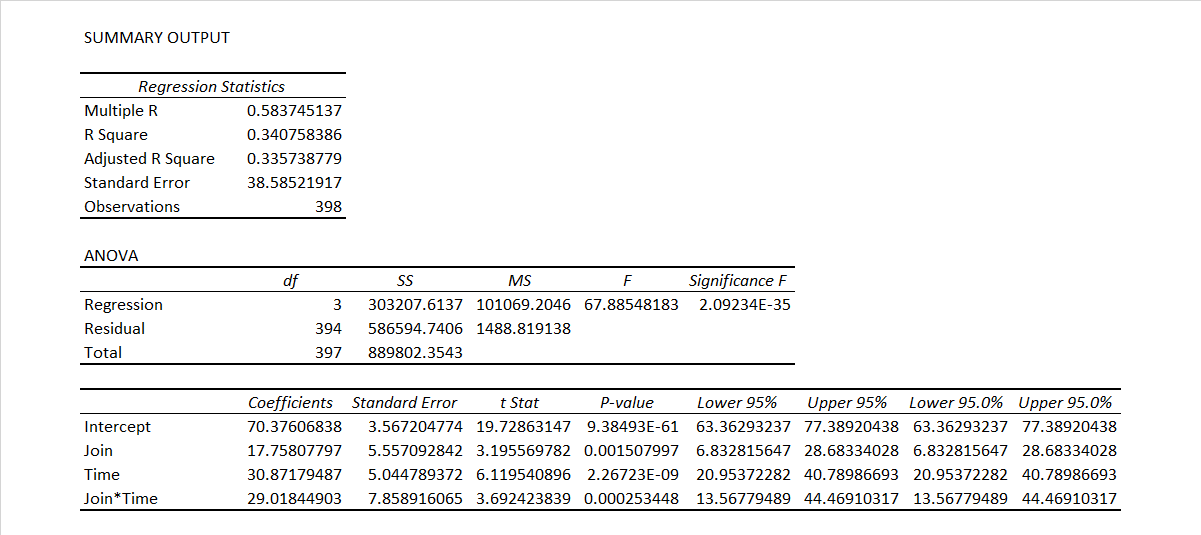


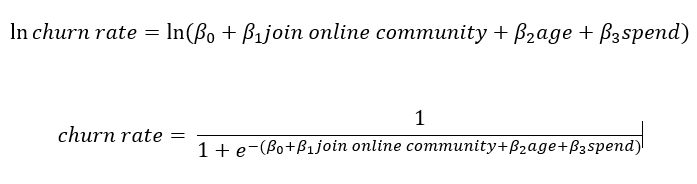
Figure 1. Regression Model using Difference in difference approach

The model we ran is as follows:

Y= β0 + β1\*[Time] + β2\*[Join Online Community] + β3\*[Time\*Join Online Community] +ε

From the result above, the intercept β0 ≈ 70.38 means the baseline average revenue of users before joining the community. The coefficient of Join β1 ≈ 17.76, which means the difference between the two groups (Join or Not Join) before the joining decision happens. The coefficient of Time β2 ≈ 30.87, which means the time trend in control group. The coefficient of Join\*Time β3 ≈ 29.02, which means the difference brought only by joining the community. Here we focused on β3 because it excluded the time effect and revealed the true effect joining the community had on user revenue. The p-value shows it has a significant effect on the revenue. The coefficient is positive, which can be interpreted as joining the community will lead to an average increase of $29 more spending per user in the following month.

Appendix X Model 2 Logistic Regression Model and the Results



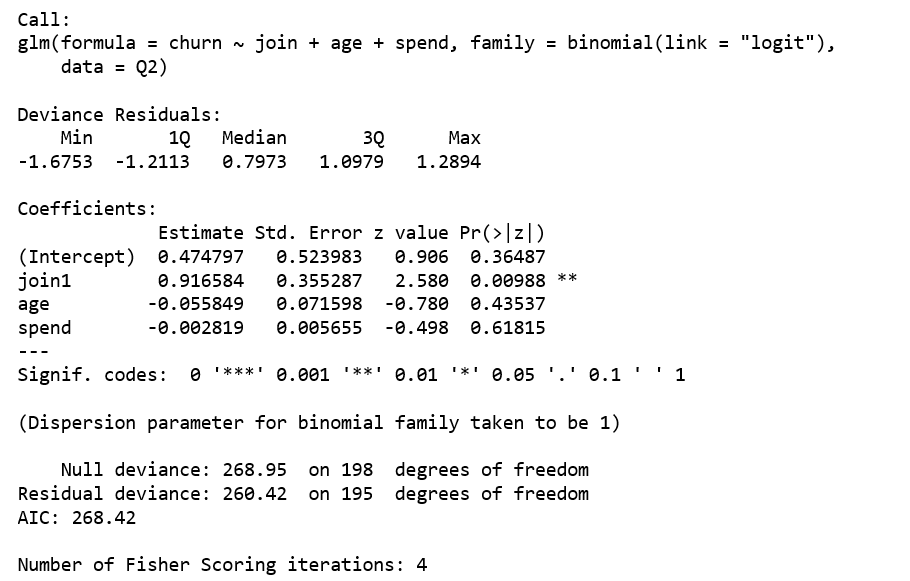


Figure 1. Logistic Regression Model: The model shows that only the online community has significant effect (with p-value 0.00988) on the churn rate. The result of the logistic regression model shows that, based on the given data, the age of the customer and the customer spending do not have proved impact on the churn rate. In contrast, the online community does impact on the churn rate. However, the result shows the online community has a negative impact on retaining customers.

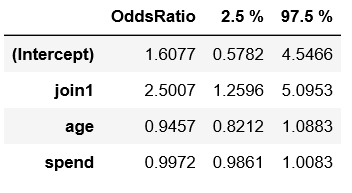


Figure 2. Exponent Coefficients: After transformation the coefficients of variables, we can find the online community’s coefficient with the churn rate is 2.5007, a positive number that means people joined the online community would have 2.5007 times possibility to churn within three months than people who did not join the online community. The word of mouth is one potential factor that needs to be monitored in the online community. The further study would be whether the online community distract customers from playing the game. More other factors that may impact the churn should be discovered.

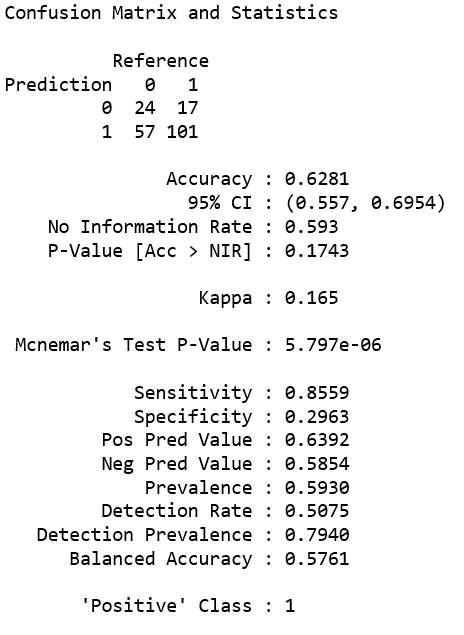


Figure 3. Confusion Matrix and Statistics: From the confusion matrix, the accuracy of this model is 0.6281, which is an acceptable number. The F-1 Score is 0.7319, which means the model is fine to predict the churn rate, and the result of the model is acceptable.

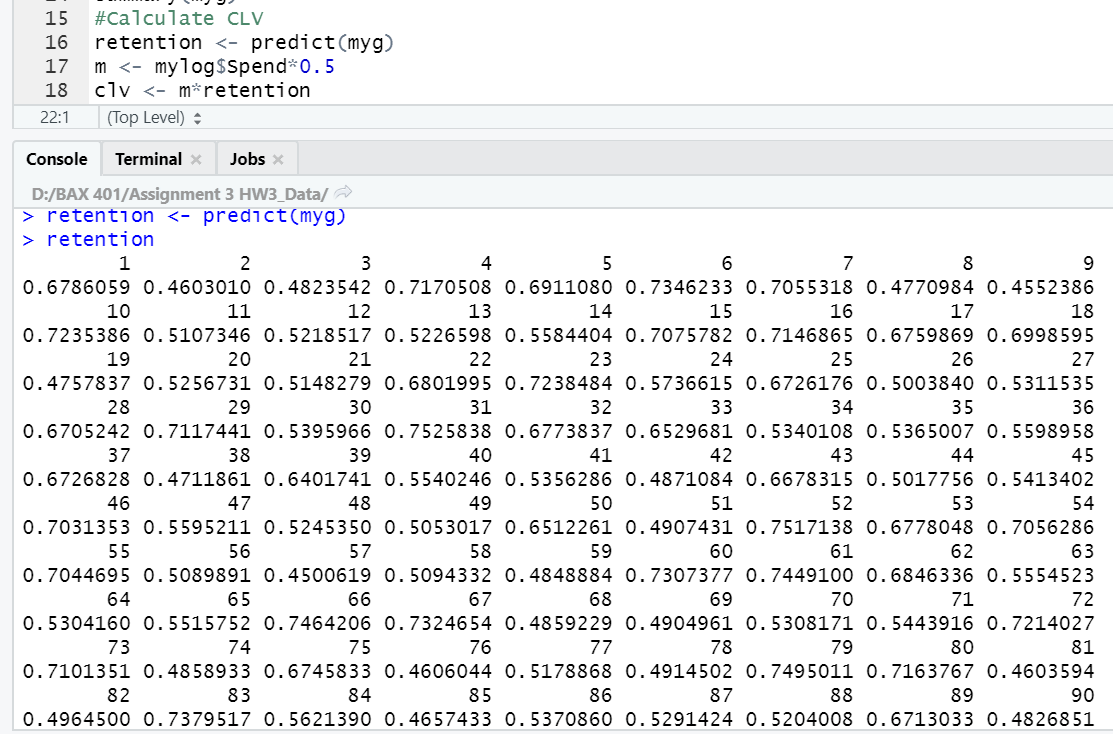


Figure.X Calculation of CLV: The equation of computing CLV is shown below.

CLV = m \* L - AC

Margin is 50% of customer spend and the expected lifetime of customer (L) is the prediction output of logistic regression model between online community and churn rate. Since the linear regression is looking for the relationship between CLV and online community, the interception here (AC) will be offset.

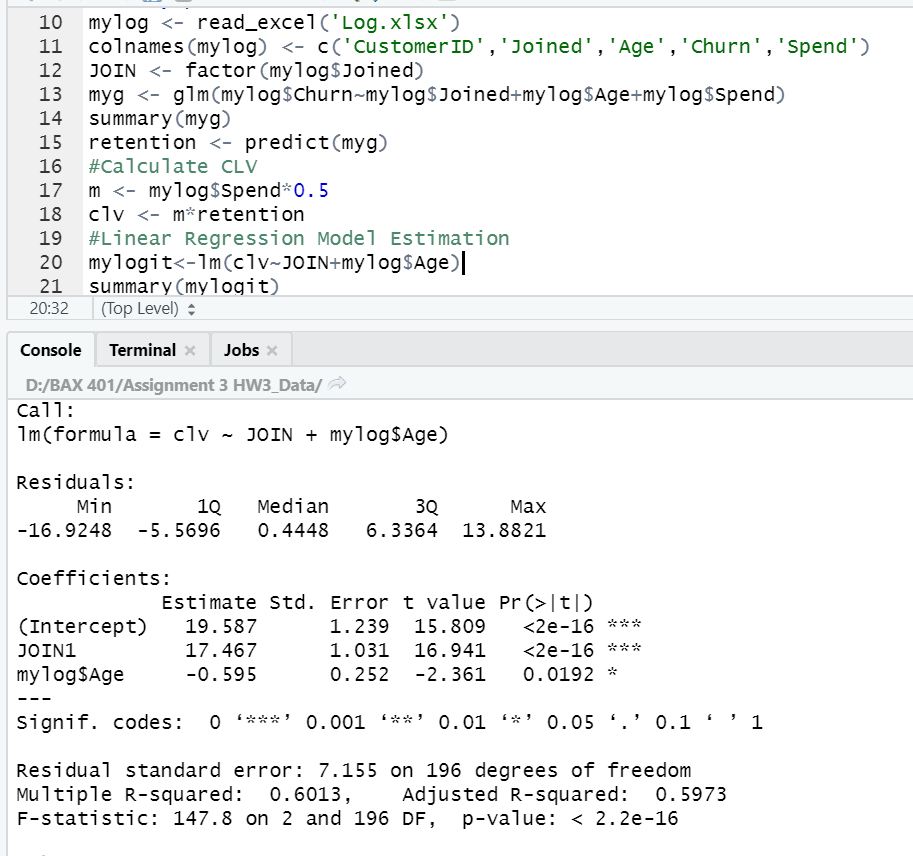
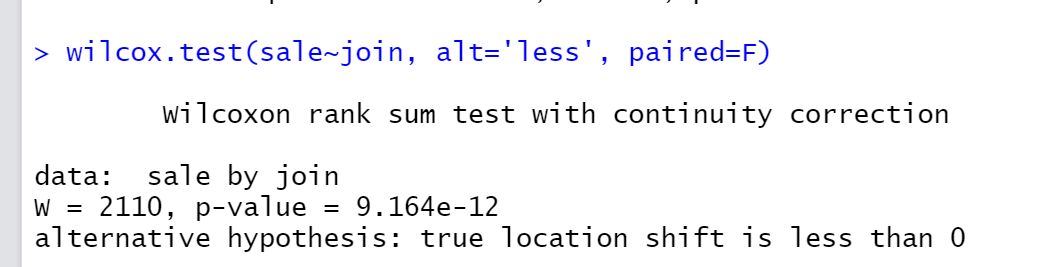


Figure.X Linear Regression Model: This model shows the linear relationship with CLV and the online community. From this model, since estimation of Join is a positive value and p-value is extremely significance, the model concludes that there is a strong positive linear regression between CLV and the online community.

Appendix XX:

Anova Analysis:

It is not normal but has constant variance. So, the wilcox test is used.



The users who joined community spend more than the users who did not join in the 90 days.

**Question 4:**

Marketing campaigns are frequently used by firms to increase brand awareness, attract customers and increase sales. One way to measure how successful the campaign is they to monitor whether the campaigns reached its target audience or not. Kyngacell also run campaigns and keeps the records of the customers they acquired from the campaigns. On the other hand, the users who joined the game not from campaigns are defined as organic users. By comparing organic users with users acquired from campaigns, meaningful insights can be drawn.

**Model development:**

We split users into 2 groups. One is organic user group and the other is campaign user group. Firstly, we calculate the descriptive statistics of the 2 groups and run tests to judge whether their difference is existed if any. Secondly, we try to use this classification of gamers to predict the retention rate. Specifically, we want to know if the users from campaigns are more likely to leave than organic users.

**Results:**

As shown in Appendix (A-D), we can find campaign users and organic users generated the same data. Their average spending, retention rate, percentage of joining the community and the age at the time of joining the community are basically the same. This is also confirmed by the results of formal statistical testing (see Appendix A-D). Also, in Appendix E, adding the information that whether a user is organic or from campaign would not help to enhance the prediction of retention rate. In other words, churn rate is not depended on the type of users. In addition, as illustrated in Appendix F, user type has nothing to do with their priority to join the community.

**Recommendation and limitations:**

As discussed above, we can conclude that the user behavior of organic users and campaign users are the same. This could give us some insights toward the effectiveness of the marketing campaigns. One major objective of a successful marketing campaign is to reach its target customers and convert these potential customers to daily users. The analysis drawn from the data of 199 users clearly demonstrated the users from campaign are actually their target audiences. Because they spend the same amount of money and stayed the same length of time as the organic users. If the acquisition cost of these users are acceptable, the current marketing campaigns can be proved as successful and should continue to perform. At last, the sample size of 199 might be not large enough compared with the huge user base of a game company. Knowing the acquisition cost would also be helpful for further analysis.

Will think more about this section later.

**Main framework for reference:**

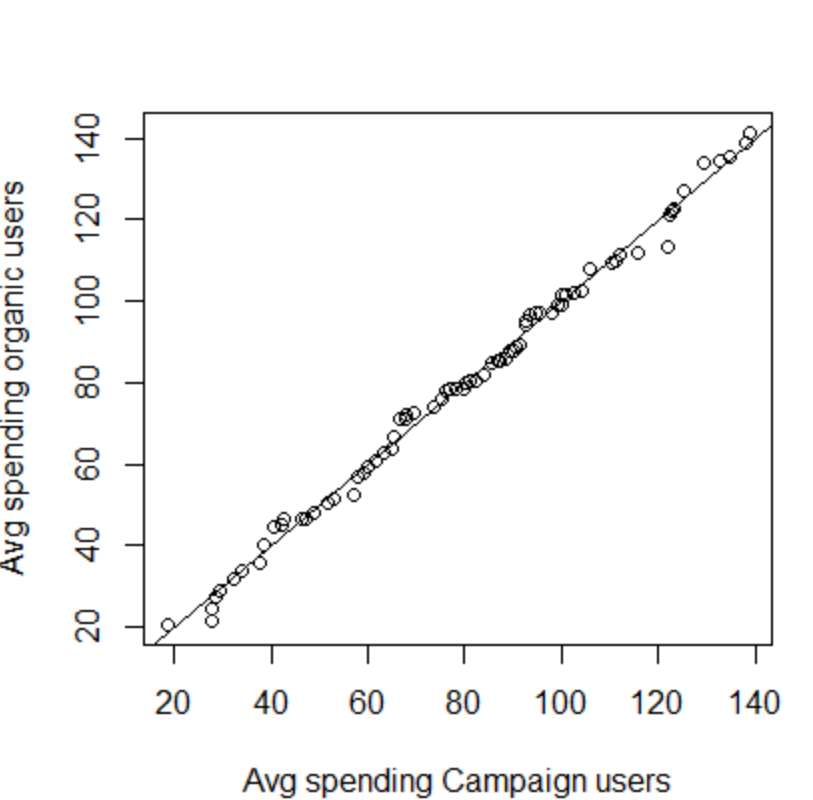
1. I will split users into 2 groups. One is organic users and the other one is campaign users.
2. I will research how different/similar of these 2 group users.
3. I will run lots of tests against different Y variables. Eg, whether the 2 groups spend the same amount of money?
4. At last, no difference found in the 2 groups, which means the behavior of campaign users and organic users are the same
5. This can prove the markeitng campaigns are very successful. Because the campaigns reached its targeting marketing. It attracted the real users just like organic users. Not the users who spend less and quick very soon.

**Appendix:**

1. The mean average spending in last 3 months of campaign users is $80.38 with a standard deviation of 30.25.

The mean average spending in last 3 months of organic users is $80.30 with a standard deviation of 30.76.

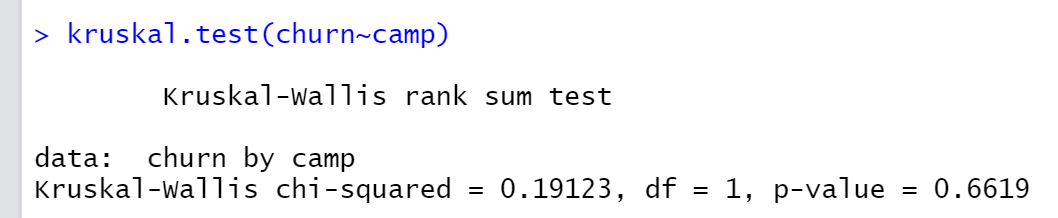
According to the QQ-plot below, we can clearly see there is no difference between the spending of organic users and the users from campaigns.



1. Retention rate of the campaign users is 60.48% with a standard deviation of 0.49

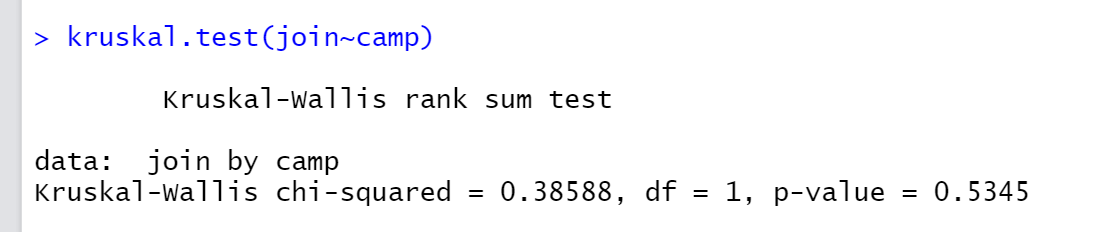
Retention rate of the organic users is 56% with a standard deviation of 0.5

According to the Kruskal test below, we can conclude that there is no statistical difference between the retention rate of organic users and campaign users.



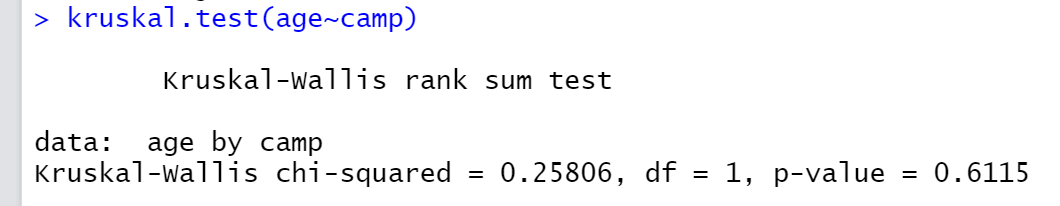
1. In our dataset, 39.52% of the campaign users joined the community and 44% of the organic users joined the community.

However, according to the Kruskal test below, we can conclude that there is no statistical difference between the percentages.

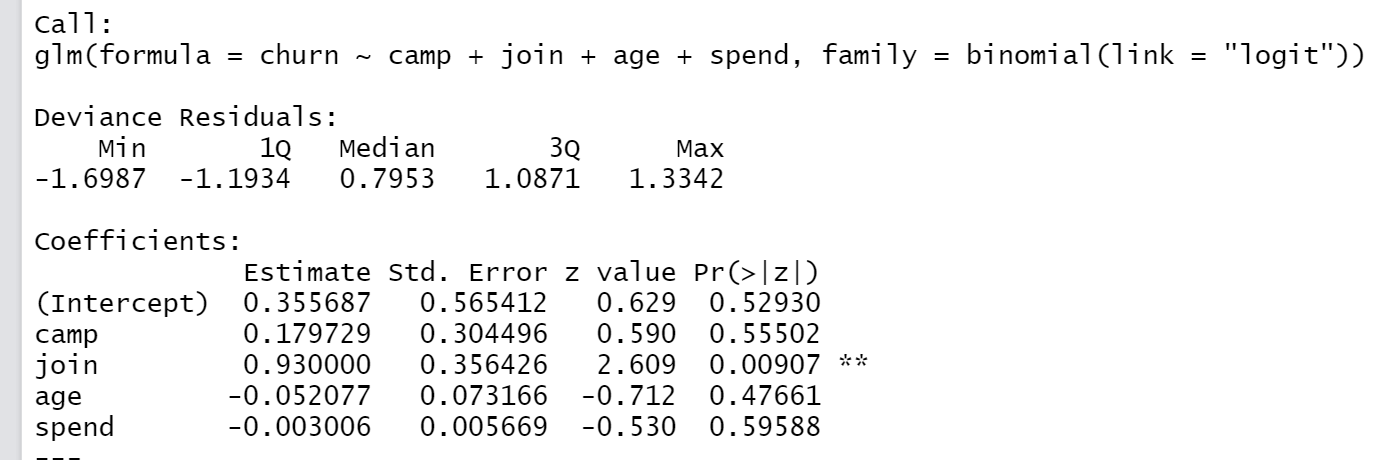


1. The mean customer age at the time of joining the community of campaign users is 4.12 months with a standard deviation of 1.86 and that of the organic users is 4.07 months with a standard deviation of 2.26.

According to the Kruskal test below, we can conclude that there are no differences between these 2 numbers.



1. By adding the variable ‘Campaign’ into our logistic regression model, there is no improvement to the overall model. In other words, the ‘Campaign’ variable did not provide any useful information for depicture the churn rate.



1. Using logistic regression to measure whether users are more likely to join the community based on user types, there is no evidence to say which type of users made a different. In other words, organic users and campaign users have the same chance to join the community.

