**Fitness Club Analysis**

**1. In addition to the spreadsheet/code or programming output you submit, include a separate written document of 250-500 words that summarizes your process.**

First, I import the needed libraries and the data. After displaying and analyzing the data initially, I found that the data type of “enrolled data” is numeric in the original file. I changed it to “datetime” type to exclude this column from further analysis.

Second, I calculated basic stats of variables. Also, I grouped by several variables such as “payment type”, “use”, and “gender” since they contain countable groups. I also describe the statistics by groups.

Third, I plot each variable using histograms to see their distribution. I also plot the combination of x variables such as payment type, use, and gender and default together to roughly estimate whether there exists any relationship. It seems like “use” has a relatively strong relationship with “default”.

Nest, I use the linear and logistic regression models to see more details about the relationships. Beginning with separating the test and train samples, I used two possible x, which is “use” and “payment type” as x and “default” as y. Then, I used two models, both Random Forest and Logistic Regression to predict the test sample. The result can be found in the heat map of confusion matrix and the accuracy score. According to the graph, I found that the score is not that significant using Logistic Regression but more significant using the random forest method. I also tried the Linear SVC method and KNN method to get the prediction, but it seems that those two methods have the same score as the logistic regression. However, for the linear regression graph, I did not find any linear relationship between “use”, “gender”, and “payment type” and “default” respectively.

**2. Discuss why certain variables you expected to be significant are/are not and any other unexpected insights.**

Because of the noise in the data (for instance, there are ages 0 and 99 in the dataset, which is nearly impossible for the people in those ages go to the fitness club), the model cannot predict the final results correctly. I originally thought that “use” and “payment type” will generate some interesting points, but the data itself has not been clean enough to illustrate my thought. Just by looking at the histogram roughly, I can find some relationships between the variables and churn, but the model methods cannot verify the results.