DSC383W Report

RMSC 1: Ruqin Chang, Valerie Tam, Shawn Wu, Ziyi You, Ruiyu Zhou

April 2021

1 Introduction

This project is sponsored by Rochester Museum and Science Center(RMSC). RMSC is a Rochester local museum that includes three parts: Science Museum, Strasenburgh Planetarium, and Cumming Nature Center. The mission of the RMSC is to inspire a better future for all through curiosity, exploration, and participation in science, culture, and the natural world[1]. We aim to spur museum membership growth, encourage donations from members, and increase overall museum revenue.

In this report, we will first describe the three data sets provided by our sponsor: membership transaction table, earned revenue table, and donation table. We later combined the membership transaction table and earned revenue table together to get the customer complete journey table. We started the exploration of the data sets by doing exploratory data analysis in Tableau and Jupiter Notebook and completed thorough visualizations for each of the four tables. Then we developed three models: multinomial logistic regression classifier, binary logistic regression classifier, and k nearest neighbor classifier to predict the actions of members in four categories: downgrade, upgrade, join, and renew. We also conducted time-series and predictive modeling on the complete journey table. The results and details will be discussed fully in the "Predictive Models" as well as the "Performance and Results" section.

2 Data Description

2.1 Membership Transaction Table

The membership transaction table was provided by the Rochester Museum and Science Center. The data set documented membership transactions each time a member joins, renews, upgrades, or downgrades the membership. There are 64,439 records and 13 features including Constituent ID, Transaction Date, Action, Expiration Date, Is Gift, Membership Program, Membership Level, Amount, Promotion Amount, Promotion Name, Membership Transaction/Revenue Application/Base currency ID, Membership Transaction/Membership Promotion/Base currency ID, and Query Record. We dropped 5 features based on our sponsor's instruction and removed misrecorded and inconsistent membership transactions for data cleaning.

2.2 Earned Revenue Table

The earned revenue table was provided by the Rochester Museum and Science Center. The data set records members' earned revenue transactions. There are 332,066 records and 17 features in-

cluding Business Counter, Order ID, Transaction Date, Order Total, Constituent ID, Member level, Zip, Program Name, Program Date, Program Time, Category, Price type, List Price, Quantity, Net Amount, Discounted, and Sales Meth. We dropped 4 features based on our sponsor's feedback.

2.3 Complete Journey Table

The Complete Journey table was constructed based on the Membership Transaction table. The motivations included: 1) to analyze behaviors of new members who join between 2015-2019; 2) to calculate the number of the total active, new-join, and returning members during certain periods. We removed the duplicated "Join", restricted the years to be within 2015-2019, and removed members without the initial-join record. Then, we also created four new columns: membership duration, join date, final expiration date, and action count. The Complete Journey table has 30328 records, contains 20900 distinctive members, and has 12 features.

2.4 Donation Table

The donation table was provided by the Rochester Museum and Science Center. The data set documented donation transactions from current and past museum members. There are 10k transaction records and 3 features including Constituent ID, Amount, and Date. During the analysis, the donation table was merged with the complete journey table to access more information about the donors.

3 Exploratory Analysis

3.1 Membership Transaction Table

3.1.1 Join and Renewal Count from 2015 to 2019

Figure 1 is the plot of cumulative member count within 2015-2019 by quarter. The y-axis displays the count of distinctive members and the x-axis as the quarter of the transaction date. Each line shows the cumulative number of members of both join and renew action in each year. Each bar indicates the join and renew count for each quarter as well. The line plot indicates that the number of members increases cumulatively, even though sometimes the growth rate is negative. Overall, the cumulative member count increases within 2015-2018, but it decreases a little bit in 2019. Besides, for each year, both the join count and the renewal count are relatively large in the first and fourth quarters.

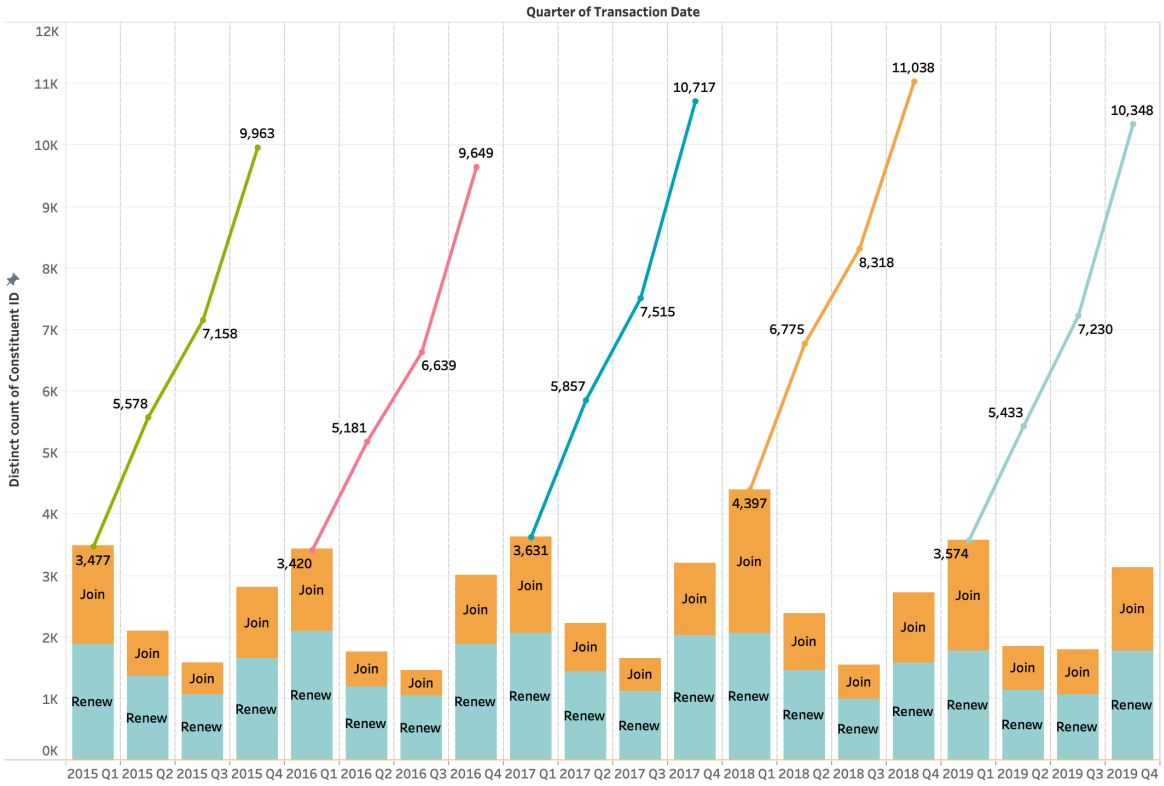

Figure 1: Cumulative member count within 2015-2019 by quarter

Besides, Figure 14 (Appendix) shows the count and growth rate of 5-year renew and join by month. We restricted to the five years from 2015 to 2019, because the data from 2020 are significantly influenced by the pandemic. We grouped all the data points by month. The upper plot shows the number of members who renew and join, and each bar is composed of two colors that display join or renew action. The gray line and the corresponding number denote the sum of join and renew count. The upper plot shows that most of the members tend to join and renew in February and December, which are the spring break and winter break. The lower plot is the growth rate of the sum of join and renew by month. It indicates that October, November, and December, have a relatively large growth rate.

3.1.2 Average Days Between Membership Renewals

Figure 15 (Appendix) shows the number of days between membership renewals. In our analysis, membership renewals are renew, upgrade, and downgrade transactions that happened before and after the previous membership expired. On average, members renew their membership after 477 days. About 51% of renewal transactions happened between 300 and 400 after members' previous membership transactions.

3.2 Earned Revenue Table

3.2.1 Earned Revenue Transactions and Visits

We merged the membership transaction table and earned revenue transaction table to draw insights into members' behavior. Based on our calculation, members made on average 7.63 earned revenue transactions within on membership period. About 49% of members made less than 5 transactions and 25% made between 5 and 10 transactions within on membership period. In addition, members made on average 2 visits within one membership period. A visit is recorded when a member made one or more earned revenue transactions on a specific day. About 30% of members made 1 visit and 20% of members made 2 visits with one period. Furthermore, we investigate the relationship between days between membership renewals and the number of visits made. Figure 16 (Appendix) shows the days between renewals and the number of visits that are color-coded. The percentage distributions of the number of visits are very similar for each 100 days interval. 50% of people visit 1-2 times, 30% visit 3-5 times, and 15% visit 6-10 times.

3.2.2 Activity Participation Rate for Renew Members

The participation rates of museum activities vary for different memberships that have been renewed in the past. In this analysis, museum activities were partitioned into 5 categories based on museum administration: General Admissions, Camps, Cumming Nature Center Admissions, Planetarium, and Others. In Figure 17 (Appendix), it can be clearly seen that Grandparents, Plus Family, and Family are the dominant memberships that have a high participation rate for museum visits. In Figure 18 (Appendix), what stands out is the Single Parent membership has the highest participation rate for camp activities such as summer camps and recess camps. In Figure 19 (Appendix), it is interesting to find out Grandparents and Plus Grandparents have a slightly higher participation rate for the nature center visits. Figure 20 and Figure 21 (Appendix) reveal that Individual membership has the highest participation rate for planetarium visits and also activities in the others category. These findings have great implications for customer behaviors and their activity preferences.

3.3 Complete Journey

We again created a plot of join count and renewal count within 2015-2019, which is shown in Figure 2. The left plot is the number of join and renewal by year, and it shows that both join count and renew count increase during the five years. The right one is the join count and renew count by month, and it shows most members tend to join and renew during February and December.

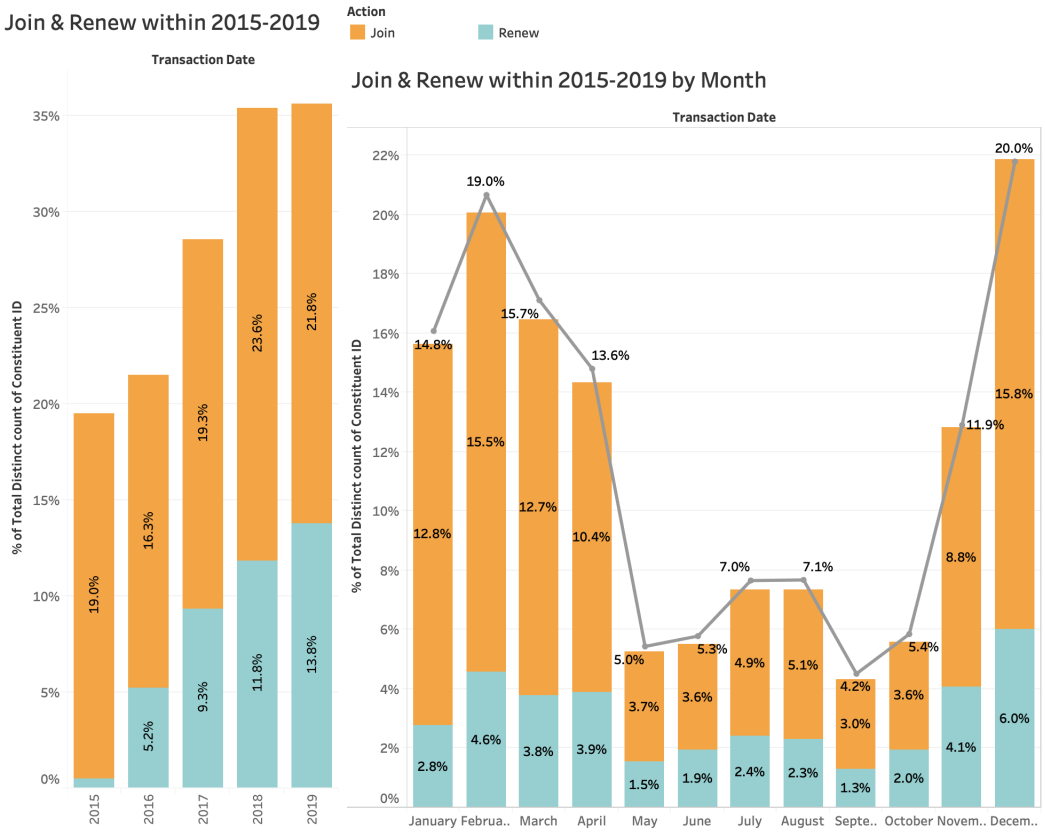

Figure 2: Join Count and Renewal Count of Complete Journey Table

We then investigated the relationship between the number of visits and different times of action. In Figure 22 (Appendix), we only include the data of the top five membership levels, which are Family, Plus Family, Grandparents, Plus Grandparent, and Single Parent. The x-axis shows the types of action, such as the first join, first renew, second renew, and so on. It turns out that the number of visits is largest in members' initial join, and it decreases as people renew more times. Then, we categorize the bar by the top three popular activities, which are general admission, giant screen movie, and star shows. It shows that the number of visits of each of the top three activities is relatively large during the first join and first renew. For most of the membership levels, members are less likely to participate in the top 3 activities after they renew several times.

Figure 23 (Appendix) shows the most popular program for each membership level. For example, for family type, the most popular program is "Animals in the Sky"; for plus family, the most popular program is "Holiday Laser"; for grandparent, it is "The Living Sea".

3.4 Donation Table

The original donation table only contains basic donation information such as donor ID and transaction date. In order to obtain more information about the donors, the donation table was

merged with the complete journey table, which resulted in around 300 transaction records with donor information. In Figure 3, it can be seen that most of the transactions happened when members joined or renewed the membership. In Figure 24 (Appendix), we looked more closely at the join and renew actions and found that a majority of the members donated when they first joined or renewed their membership. In Figure 25 (Appendix), what stands out is that members made the largest average donation amount when they first joined and this amount decreased as they progressed in their membership journey. In Figure 26 (Appendix), it is interesting to see that on average, Grandparent and Plus Grandparent memberships made the largest amount of donation out of all the memberships. These findings can be helpful for the museum to manage relationships with its current and potential donors.

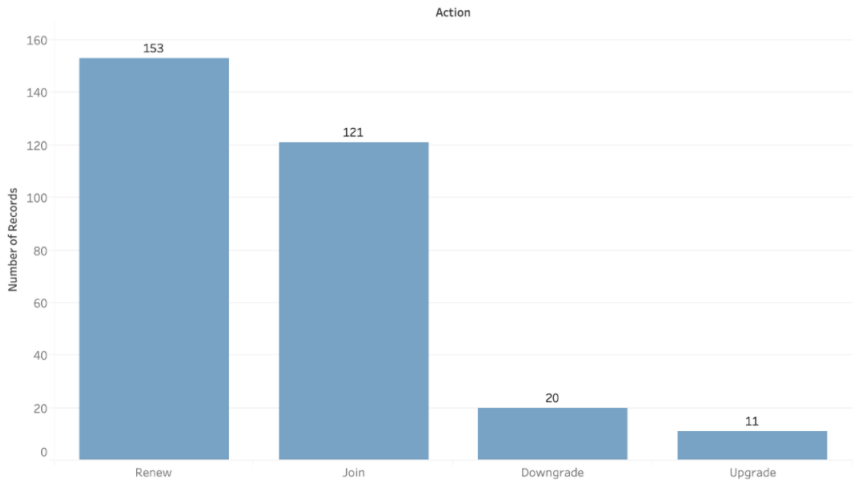

Figure 3: Number of Donations vs. Action

4 Predictive Models

We applied Logistic Regression (LR) to select the important features for the following reasons: LR performs higher accuracy than Random Forest; the sample size is large, and there are less than five classes of the predictor variable; instead of using all features as the input, feature reduction by LR improves the precision and recall by 20%.

We split the data into the training set (70%) and the testing set (30%). We constructed a multinomial LR Classifier and a binary LR classifier for the Complete Journey Table, while we created a K-Nearest Neighbor Classifier (KNN) for a table merged by Complete Journey Table and Revenue Table.

4.1 Multinomial Logistic Regression Classifier

The multinomial logistic regression classifier was created using the Complete Journey table. 70% of the data was used for training and 30% for testing. The goal of this model was to predict four different classes of members' actions, specifically downgrade, join, renew, and upgrade. The

features that were chosen were based on the feature selection plot in Figure 5 below. The top 4 features that were selected include Constituent ID, Transaction Date, join date, and final expiration date.

Figure 4: Importance Score by Logistic Regression Feature Selection

4.2 Binary Logistic Regression Classifier

The binary logistic regression classifier was used to predict two of the members' actions, which were join and renew. The data used to train and test the model was from the Complete Journey table. 70% of the data was used for training and 30% for testing. We used logistic regression feature selection to determine which attributes should be included in the binary logistic regression classifier. Based on Figure 6 below, the top 4 important features were Membership Level, Expiration Date, Membership Duration, and action count. Therefore, we included these 4 features to build the model. The positive scores indicate features predicting class one, while the negative scores indicate features predicting class zero [2].

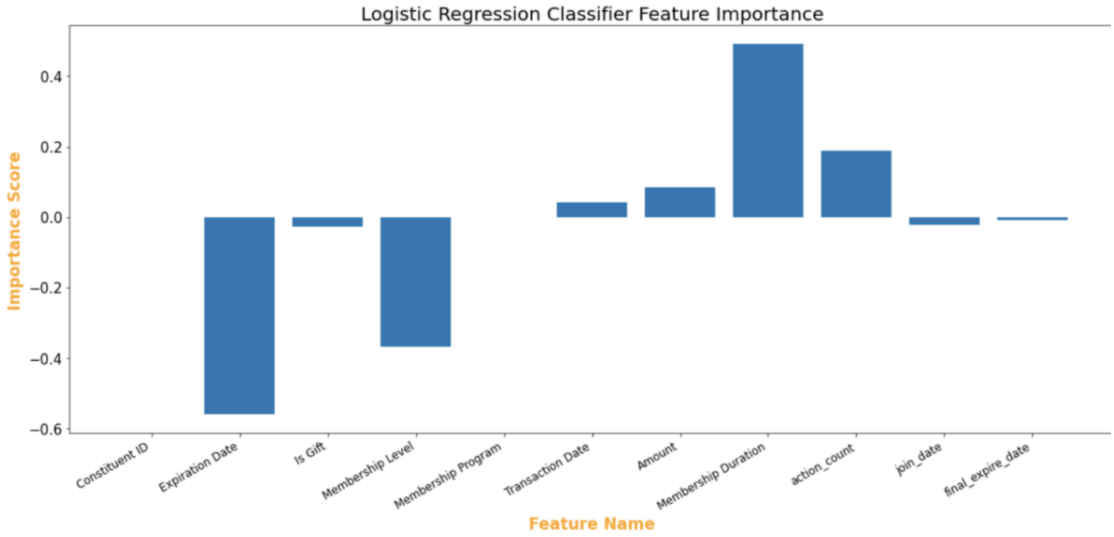

Figure 5: Importance Score for the Binary Logistic Regression Classifier

4.3 K-Nearest Neighbor Classifier

The KNN classifier was created to predict each type of action that might occur for each member, such as 1st renew, 2nd renew, and 1st upgrade. The data that was used to test and train this model was taken from the Complete Journey table merged with the Revenue table. For this model, 70% of the data was used for training and 30% for testing. Before creating the KNN classifier, we completed some data preprocessing. For example, we deleted three classes with a low occurrence ($\leq 0.01\%$). These classes were 2nd downgrade, 5th renew, and 6th renew. After deleting these 3 classes, 7 classes were left. Based on the feature importance plot in Figure 7, the top three important features were Action, Membership Level, and Amount.

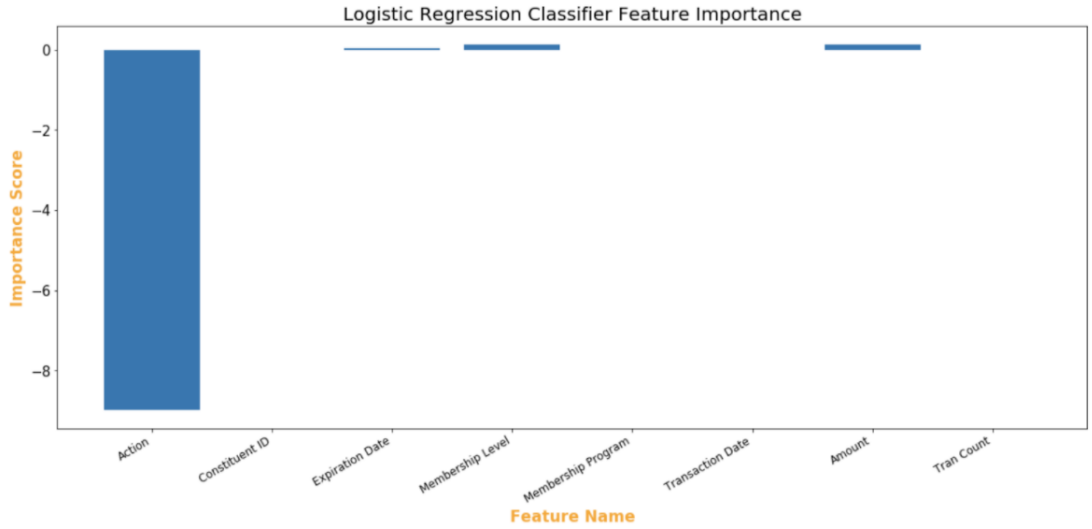

Figure 6: Importance Score for the KNN Classifier

4.4 Time Series Forecasting Models

Interested in finding the number of newly acquired members per period and the number of acquired members who are active per period, we applied cohort analysis to do the calculation [3]. Based on the join date, final expiration date, and cohort date, we are able to calculate the number of total active, new join, and returning members within each period. We define active members as those whose memberships are not expired, new join members as those who first join the membership, and returning members as those who have transactions after the first join. The properties of the time series depend on the cohort period we set. For example, if we calculate the number of total members by day, that is how many members make transactions each day within the five years, then the time series is very noisy, and no explicit seasonality and trend is shown. If we calculate the number of total members by week, that is how many members make transactions weekly within the five years, then the time series is less noisy and seasonality is shown, while the trend is not explicit. If we calculate the number of total members by year, then we find that the number of total active and returning members keeps increasing within the five years, while the number of new join members stays at the same level. Finally, we set the cohort period to be month and practiced predictive models on it.

This is a time series forecasting model of the number of total active members. The x-axis denotes the month within the five years. After decomposition, we found that a multiplicative model suits the time series data well, so we applied a Holt-Winters model and set seasonal to be multiplicative while the trend to be additive. Similarly, for the number of by decomposition, we found that a multiplicative model suits the number of new join members best, so we also applied a Holt-Winters model and both trend and seasonal to be multiplicative. Finally, for the number of returning members, we applied the ARIMA model.

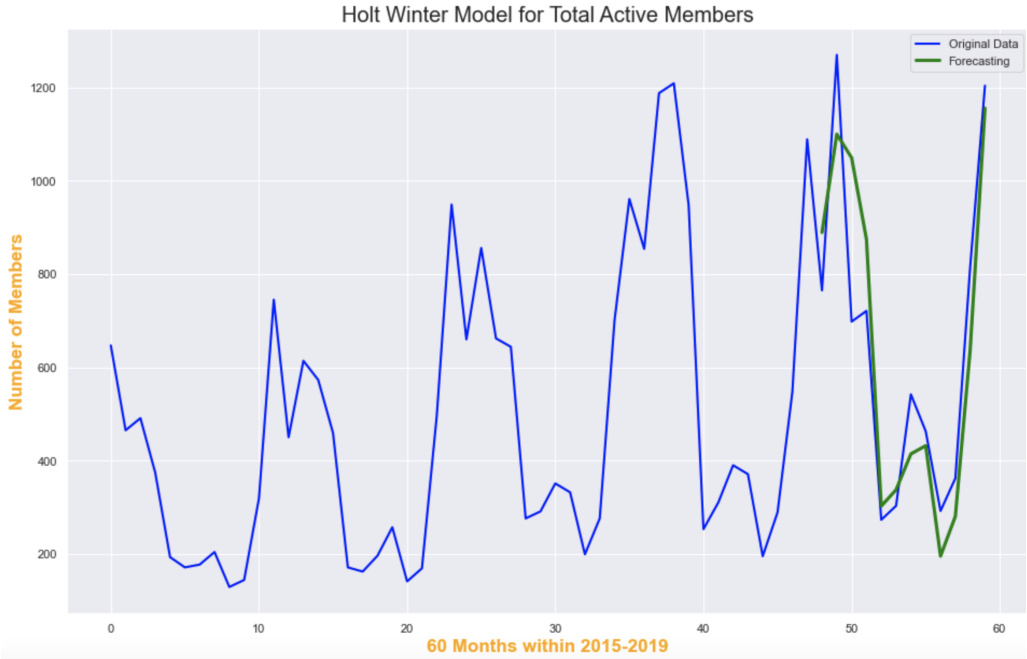

Figure 7: A Multiplicative Model for the Number of Total Active Members

5 Performance and Results

5.1 Multinomial Logistic Regression Classifier

Figure 9 below shows the confusion matrix for the multinomial logistic regression classifier. Based on this plot, we can see that 6,264 data points that were join actions were predicted correctly as join. Similarly, 2,534 renew actions were correctly predicted as renew. The accuracy of this model was 97%. The detailed results of the multinomial logistic regression classifier are shown below in Figure 10. We can see that the testing accuracy is 96.69%. When we take a closer look at the precision scores, the join class has a precision score of 98% and the renew class has a precision score of 93%. However, the classes upgrade and downgrade have a precision score of 0. This is because the number of records with upgrade and downgrade was very low.

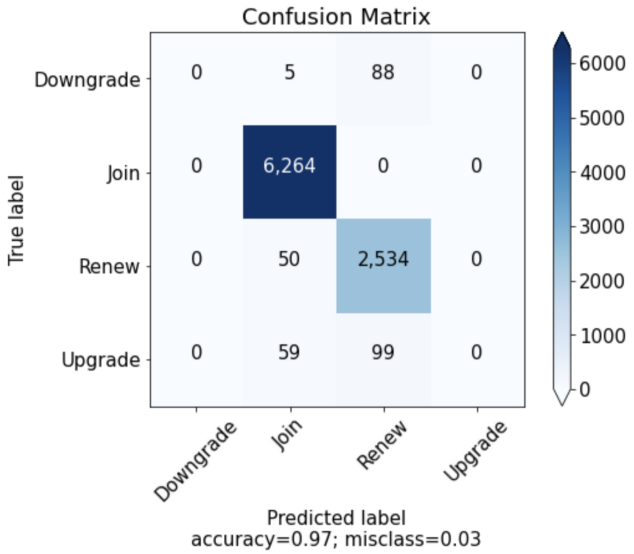

Figure 8: Confusion Matrix of Multinomial Logistic Regression Classifier

```
Training-set accuracy score: 96.75%
Test-set (Model) accuracy score: 96.69%
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        93
           1       0.98      1.00      0.99      6264
           2       0.93      0.98      0.96      2584
           3       0.00      0.00      0.00       158

    accuracy                           0.97      9099
   macro avg       0.48      0.50      0.49      9099
weighted avg       0.94      0.97      0.95      9099
```

Figure 9: Detailed Results of Multinomial Logistic Regression Classifier

5.2 Binary Logistic Regression Classifier

Figure 11 below shows the confusion matrix for the binary logistic regression classifier. Based on this confusion matrix, 6,177 data points that were join actions were predicted correctly as join by the model. 1,659 data points that were renew actions were predicted correctly as renew. The accuracy for this model is also relatively high at 89%. The detailed results of the binary logistic regression classifier are shown in Figure 12. We can see that the accuracy of the test set was 88.51%. We also ran 5-fold cross validation and the average accuracy of both training and testing sets were 89%.

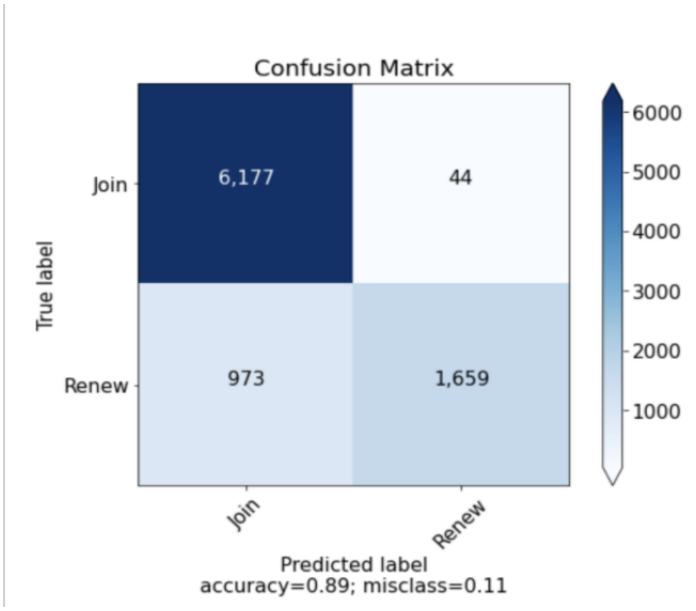

Figure 10: Confusion Matrix of Binary Logistic Regression Classifier

```
Training-set accuracy score: 89.34%
Test-set (Model) accuracy score: 88.51%
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.99      0.92      6221
           1       0.97      0.63      0.77      2632

    accuracy                           0.89      8853
   macro avg       0.92      0.81      0.84      8853
weighted avg       0.90      0.89      0.88      8853
```

Figure 11: Detailed Results of Binary Logistic Regression Classifier

5.3 K-Nearest Neighbor Classifier

Figure 12 below shows the confusion matrix for the KNN classifier. Based on this plot, 1,926 of 1st join data points were predicted correctly as 1st join by the model. 602 data points that were 1st renew actions were predicted correctly as 1st renew. The accuracy for this model is at 87%. The detailed results of the KNN classifier are shown below in Figure 13. We can see that the accuracy of the test set was 86.69%. We also ran 5-fold cross-validation and the average accuracy of the training set was 86% while the testing set accuracy was 85%.

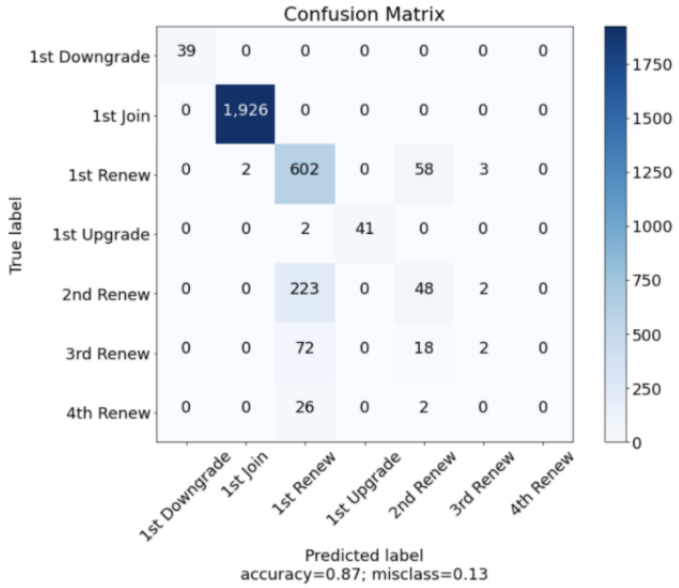

Figure 12: Confusion Matrix of KNN Classifier

```
Training-set accuracy score: 87.82%
Test-set (Model) accuracy score: 86.69%
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        39
           1       1.00      1.00      1.00      1926
           2       0.65      0.91      0.76       665
           3       1.00      0.95      0.98        43
           4       0.38      0.18      0.24       273
           5       0.29      0.02      0.04        92
           6       0.00      0.00      0.00        28

    accuracy                           0.87      3066
   macro avg       0.62      0.58      0.57      3066
weighted avg       0.84      0.87      0.84      3066
```

Figure 13: Detailed Results of KNN Classifier

5.4 Time Series Forecasting Model

For the number of all three types of members, which are the total active, new join, and returning members, we applied the 48-month data during 2015-2018 as the training set to predict the number of members for the next 12 months, which is the year 2019. The root-mean-square error of the

Holt-Winters model of the number of total active members is 147; the root-mean-square error of the Holt-Winters model of the number of new join members is 101; the root-mean-square error of the ARIMA model of the number of returning members is 47.

6 Conclusion and Next Steps

From the membership transaction table, we have the following conclusions. First, the cumulative member count increases within 2015-2018, but it decreases a little bit in 2019. Besides, for each year, both the join count and the renewal count are relatively large in the first and fourth quarters. And members renew their membership on average after 477 days.

From the earned revenue table, we investigated the relationship between days between membership renewals and the number of visits made and we can draw the following conclusions that 50 percent of people visit 1-2 times, 30 percent of people visit 3-5 times, and the rest visit 6-10 times.

From the complete journey table, we can draw the following conclusions. First, the renewal and join count both increase during 2015-2019. People are mostly like to join and renew in February and December, which are around the spring break and winter break. Second, people would like to visit the museum most frequently in their first join and first renew. Third, members of different memberships have varying preferences when choosing activities to participate in RMSC could target specific membership level visitors to their most frequent program.

From the exploratory analysis of the donation table, we can see the following trends. First, people usually make donations at the beginning of their membership journey, for instance, when they first join. Second, the member donation amount goes down as time goes on. Third, the average donation amount of grandparents is the highest among all membership levels.

We encountered some challenges in dealing with the donation table. First, we have very limited donation records of around only 300 records, so we went for visualization graphs instead of predictive modeling. Second, we tried to predict the amount of donation through a linear regression model but it turned out that donation cannot be well predicted by a linear model. Third, there is a significant class imbalance for people who donated versus those who didn't. If we had more time for this project, we would work on finding a better fit model to predict donations from members.

In the next step, our team is going to present the findings to the Rochester Museum and Science Center board committee. We are hoping to gain constructive feedback from the museum administrators to further improve our data analysis and ensure it caters to the analytical needs of the museum. Due to significant class imbalance and limitation on the number of records, we were not able to construct useful predictive models on the donation table. The analysis would have been more interesting if it had included these predictive models and this task could be a potential step for our future analysis.

Appendix A Visualization

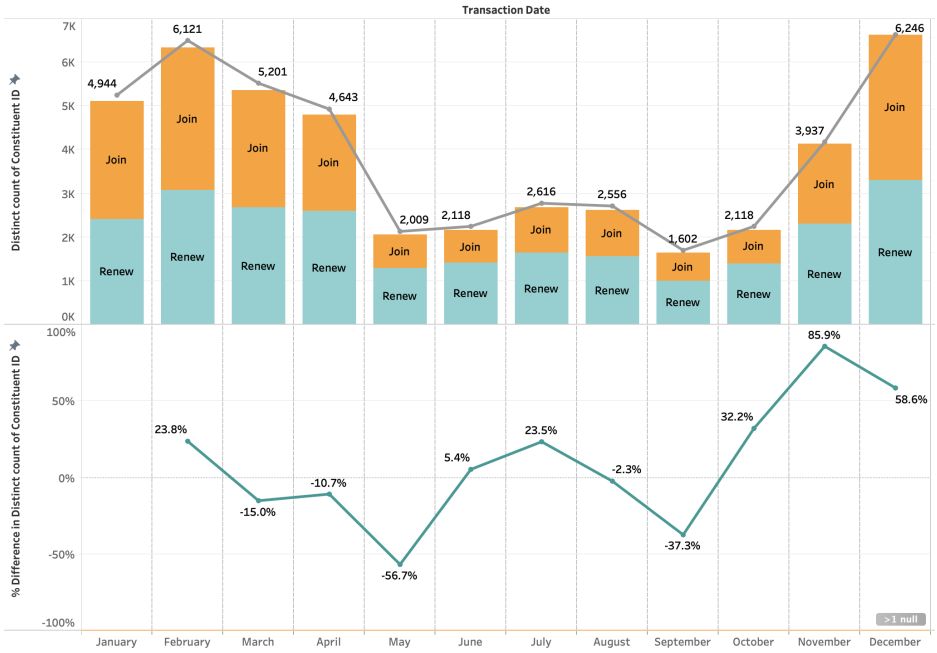

Figure 14: Count and growth rate of 5-year renew and join by month

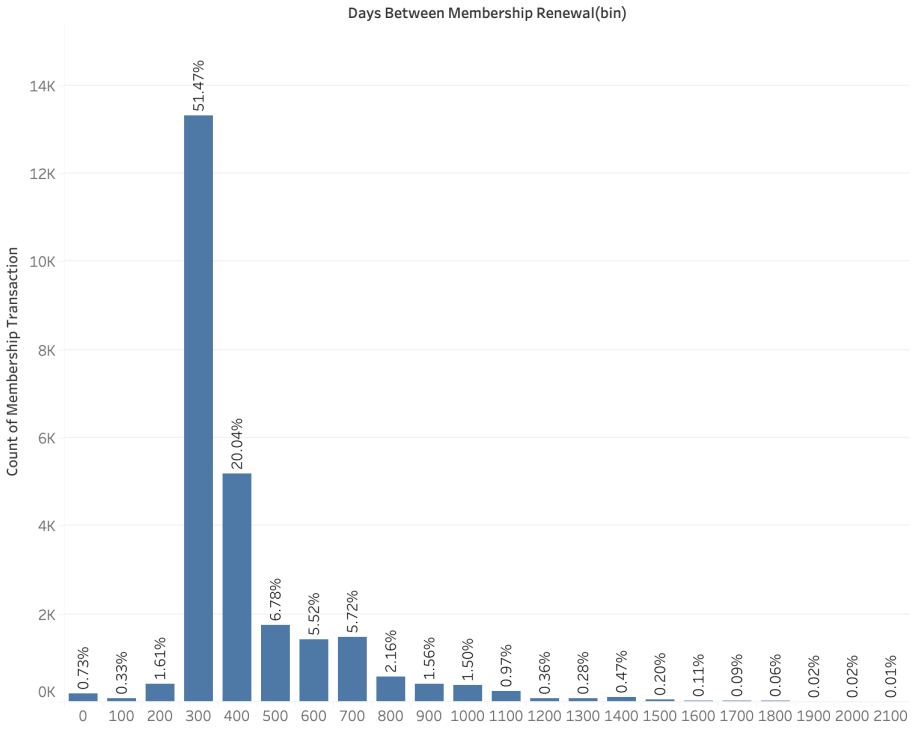

Figure 15: Days Between Membership Renewal

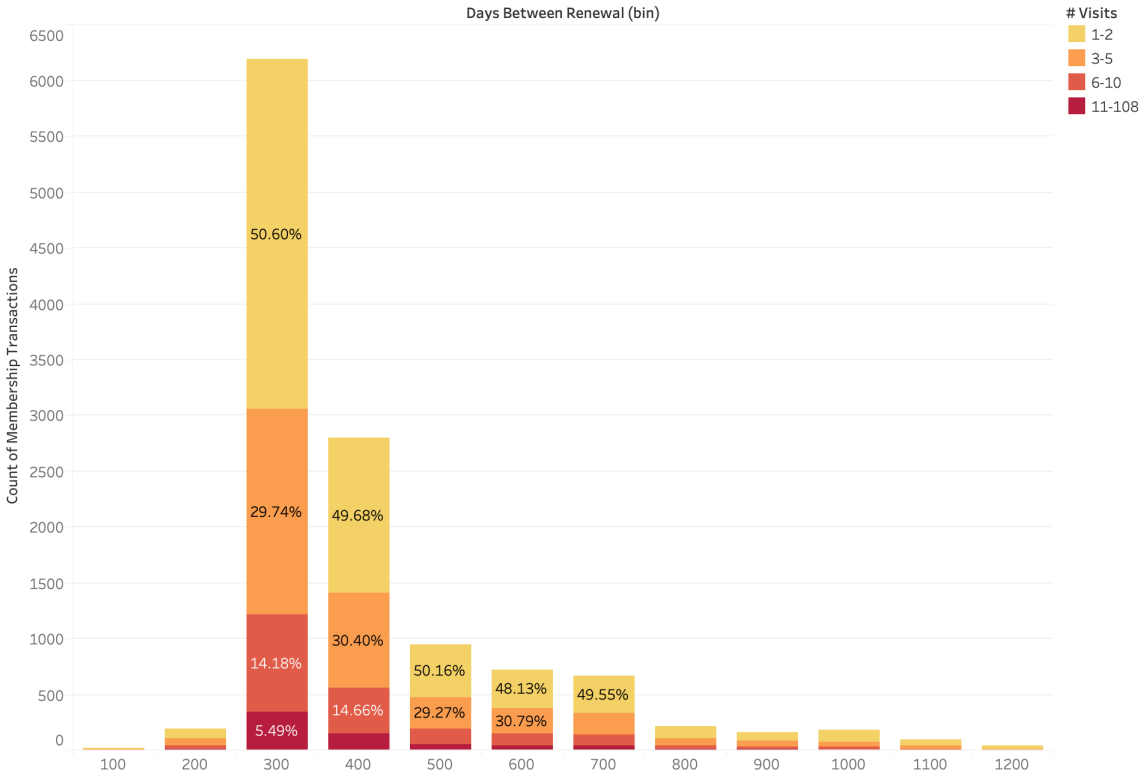

Figure 16: Days Between Renewal and Number of Visits

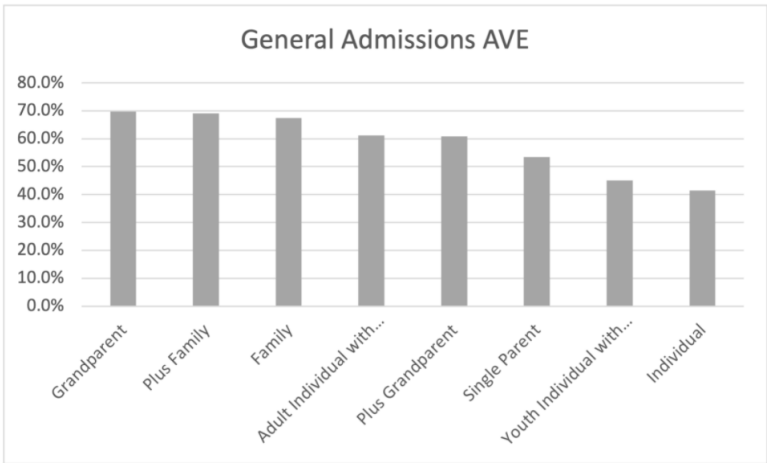

Figure 17: Participation Rate for General Admission

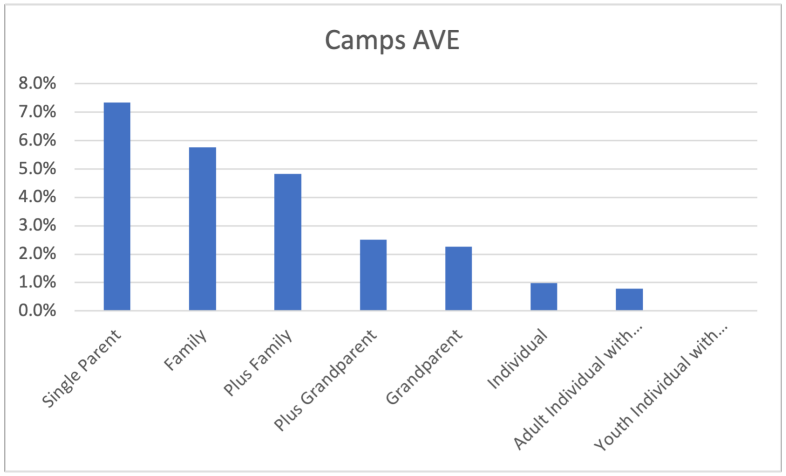

Figure 18: Participation Rate for Camps

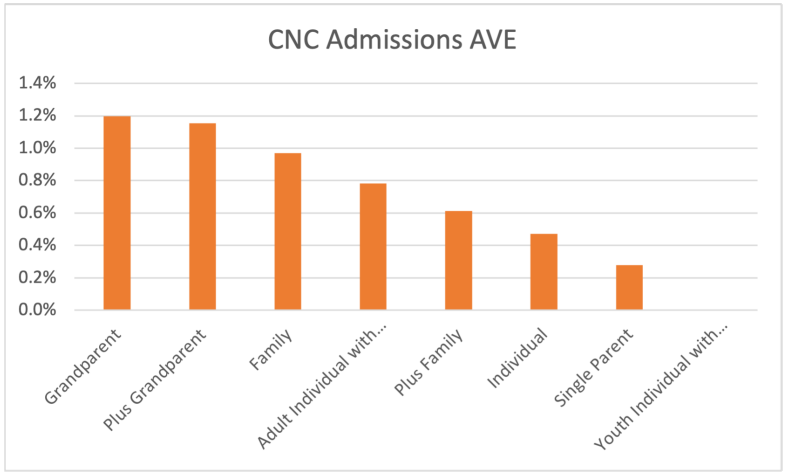

Figure 19: Participation Rate for Cumming Nature Center

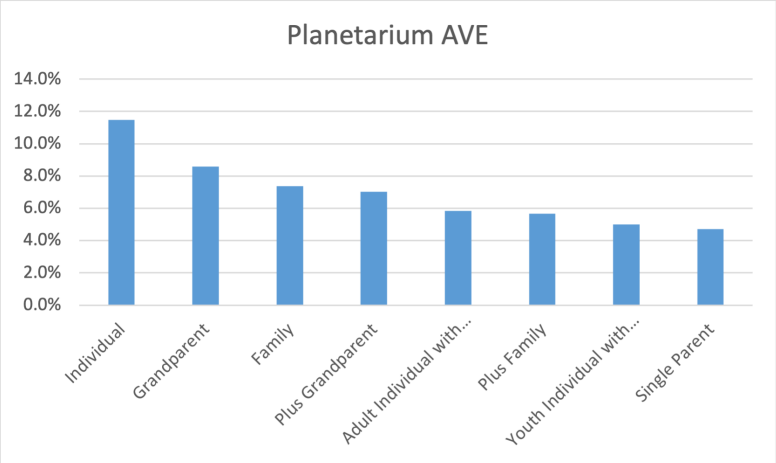

Figure 20: Participation Rate for Planetarium

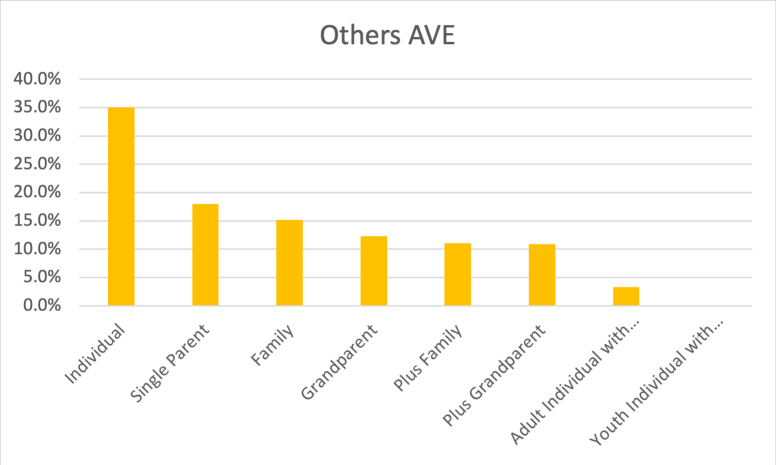

Figure 21: Participation Rate for Others

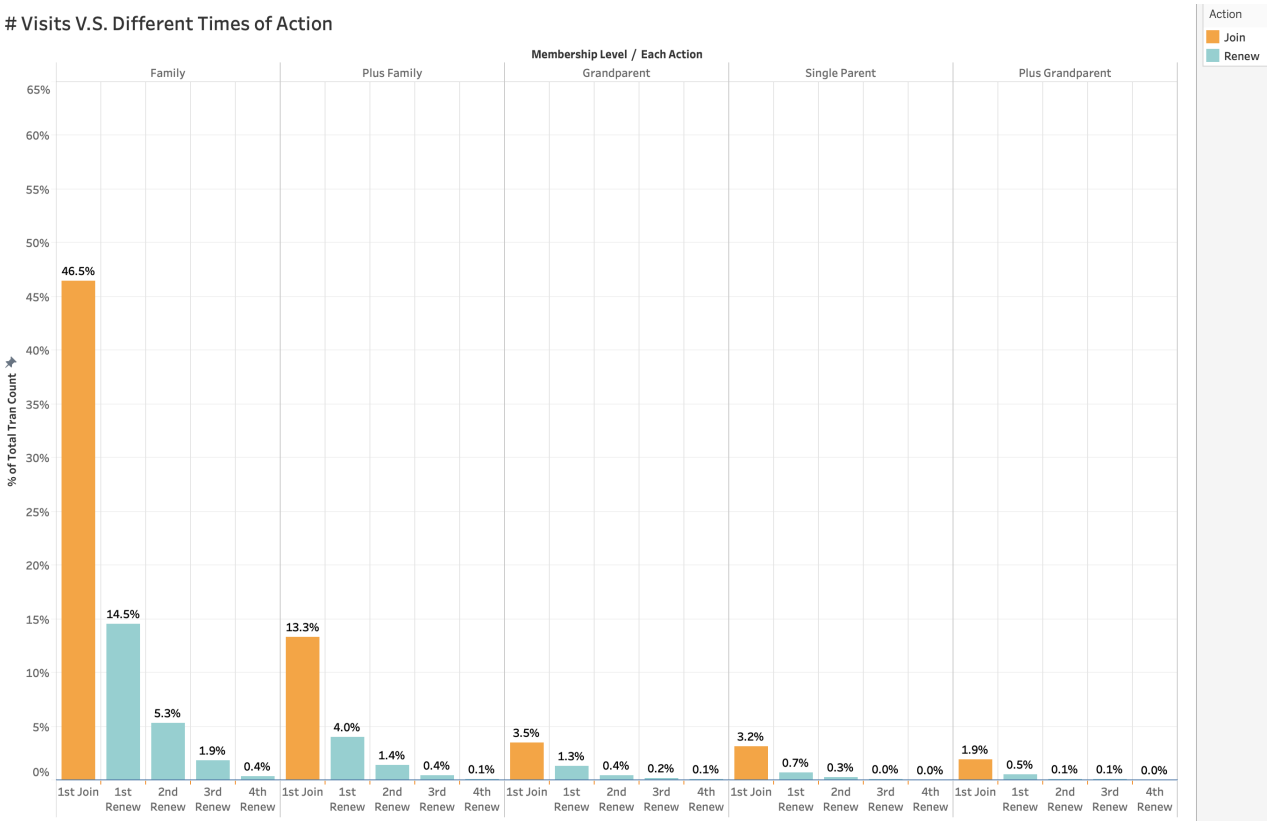

Figure 22: Number of Visit versus Different Times of Action for Top 5 Membership Levels

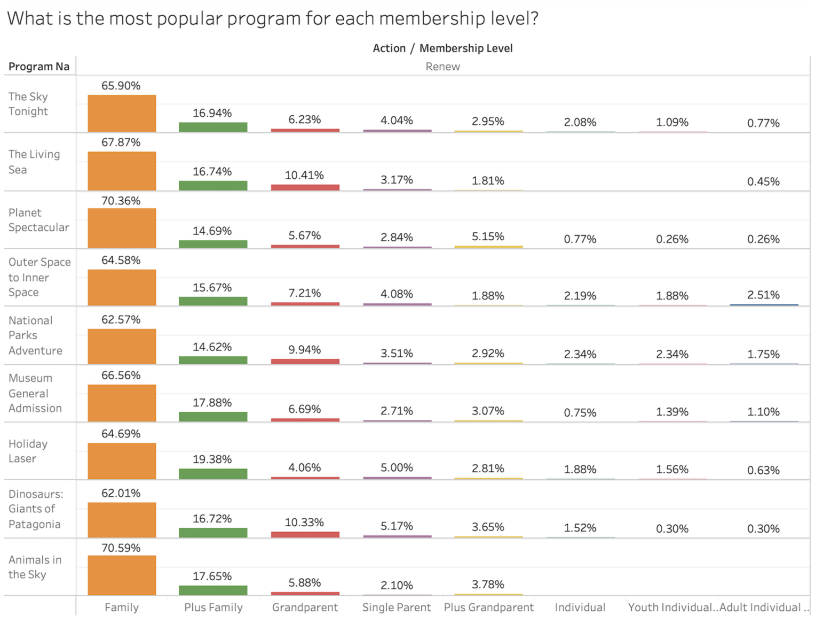

Figure 23: What is the most popular program for each membership level?

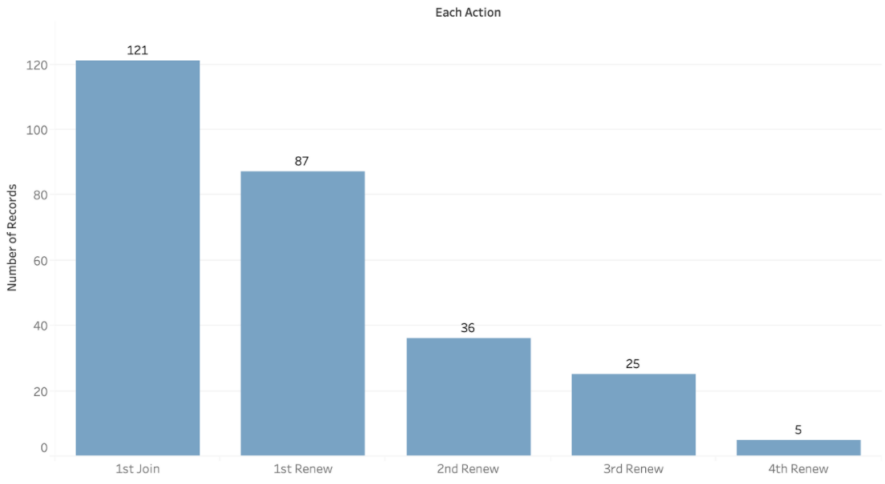

Figure 24: Number of Donations for Renew and Join Actions

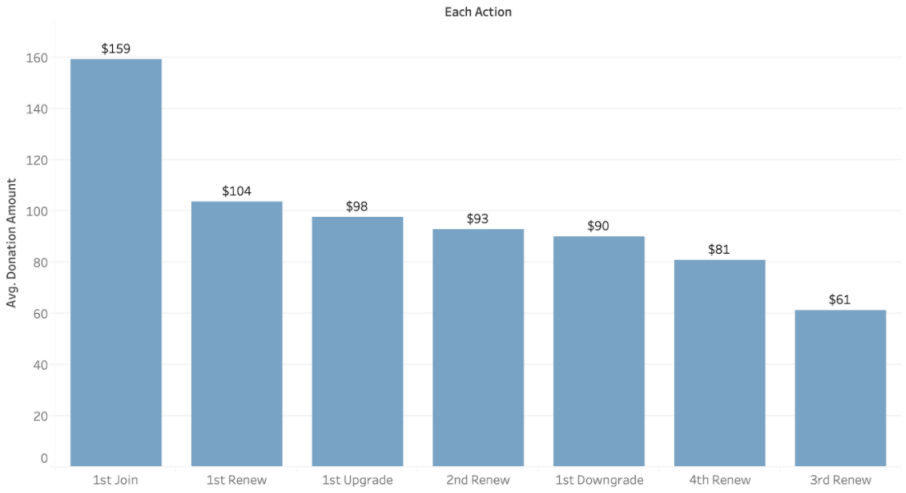

Figure 25: Average Donation Amount for Join and Renew Actions

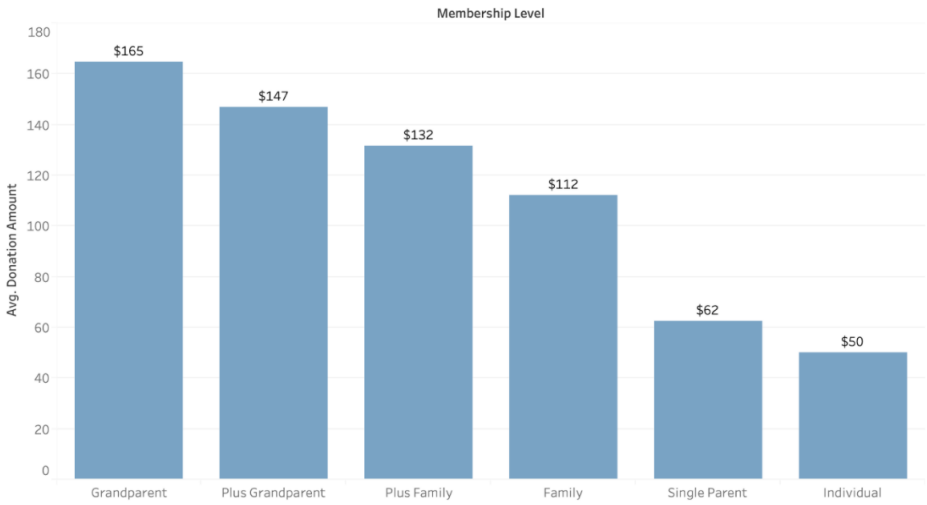

Figure 26: Average Donation Amount for Different Memberships

References

[1] *Rochester Museum and Science Center.* Available at `https://rmsc.org/about/mission-statement`.

[2] *How to Calculate Feature Importance With Python.* Available at `https://machinelearningmastery.com/calculate-feature-importance-with-python/`.

[3] *User Acquisition Stats.* Available at `https://medium.com/multiplyai/user-acquisition-stats-b3fd0928acff`.