BUDT704

Data Processing and Analysis in Python

TEAM MEMBERS:

Bharath Kumar Routhu

Hsin-Yu Tsai

Yingnan Mu

Ying-Ting Lin

Project Introduction

Our project goal is to increase event attendance of first time attendees and major prospect attendees. And there are three mission statements for our project, which are:

- To identify the correlated variables (Time, Location, Group Code) to our objective
- To understand what type of events attract the objective the most
- To figure out how to optimize the current and future event

Project Method

We have separated the project analysis into two different parts. The first part being the 'Preliminary Analysis' part, which is focusing on the participation rate, the number of events held in each location, and the number of events held for each group. We used several pie charts, heat maps, bar charts, histograms, and line charts to figure out the distribution and relationship of different features.

Then, the second part is 'Further Analysis', which is to understand how variables are related to first time attendees, as well as how variables related to the major prospects. We used scatter plots, histograms, tree maps to do so.

Additionally, we also do the clustering and build regression model, try to predict the future first time attendees and major prospects for the future work.

Analysis & Finding

- **Preliminary Analysis**

  In this part, we focus on the participation rate, the number of events held in each location, and the number of events held for each group to see whether we can find any obvious relationship between each variable. Followings are our findings:

  1. Number of events and participants over time (Figure 1):

     The average number of participants decreases by time, however, the number of events increases. Namely, holding more events does not seem to attract more participants. Consequently, holding more events does not really increase the first time attendees and major prospects.

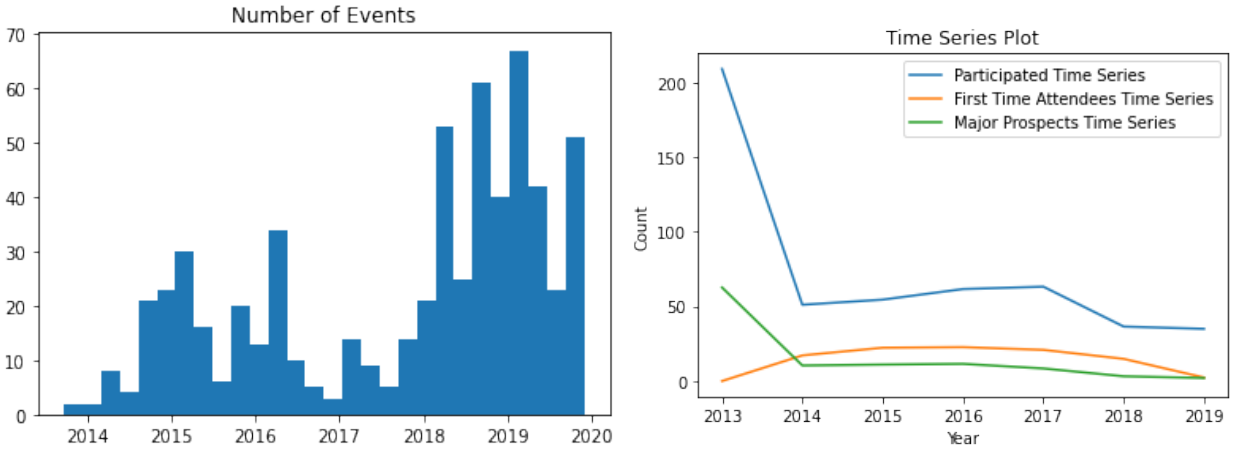

Figure 1: Number of events and participants over time

2. Number of participants over weekdays and weekends (Figure 2):
   a. Holding events during weekends does not intuitively attract more participants.
   b. Holding events on Friday seems to attract more participants, whereas on Saturday might have a lower participation rate.

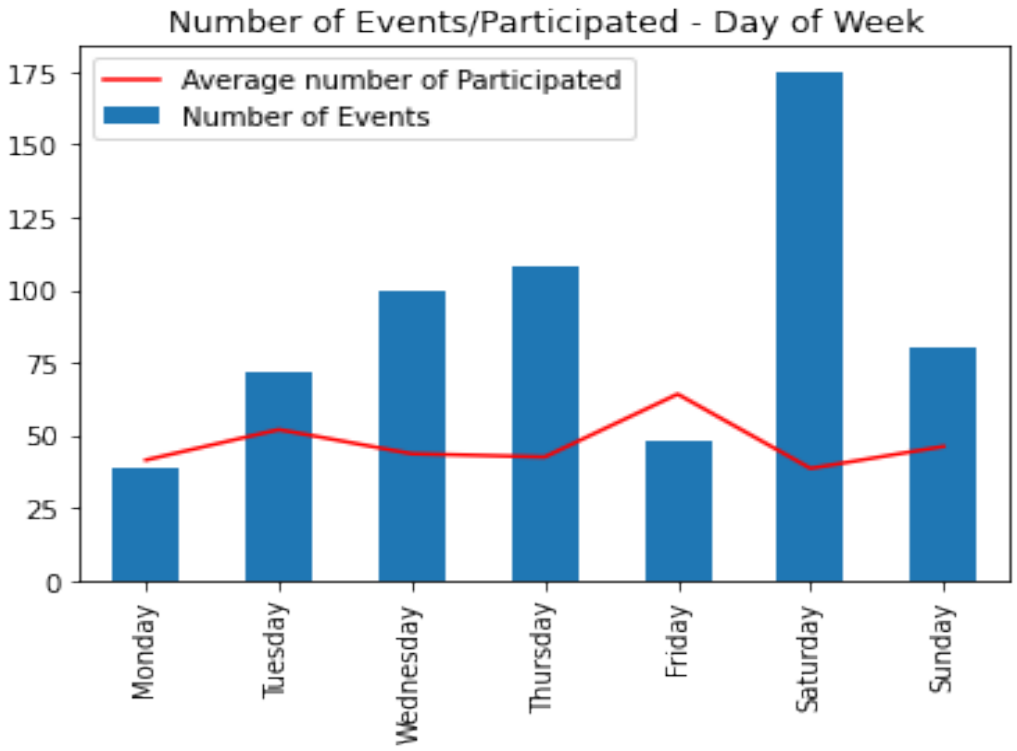

Figure 2: Number of participants over weekdays and weekends

3. Correlations of other variables to objectives (first time attendees & major prospects) (Figure 3):
    a. High correlation between *Participated & First time Attendees* (0.84 - strong correlation); And *Participated & Major Prospects* (0.66 - strong correlation), indicating that for attracting more first time attendees and major prospects, UMD Alumni can hold the event that can accommodate more participants.
    b. The correlation of *Weekend & First time attendees*, as well as correlation of *Weekend & Major prospects* are quite low, meaning that holding events during weekend cannot effectively attract more first time attendees and major prospects.
    c. The *Age* seems to only have a relatively high correlation with *Percentage of major prospects*. Namely, the increase of age of participants seems to increase only the percentage of prospects

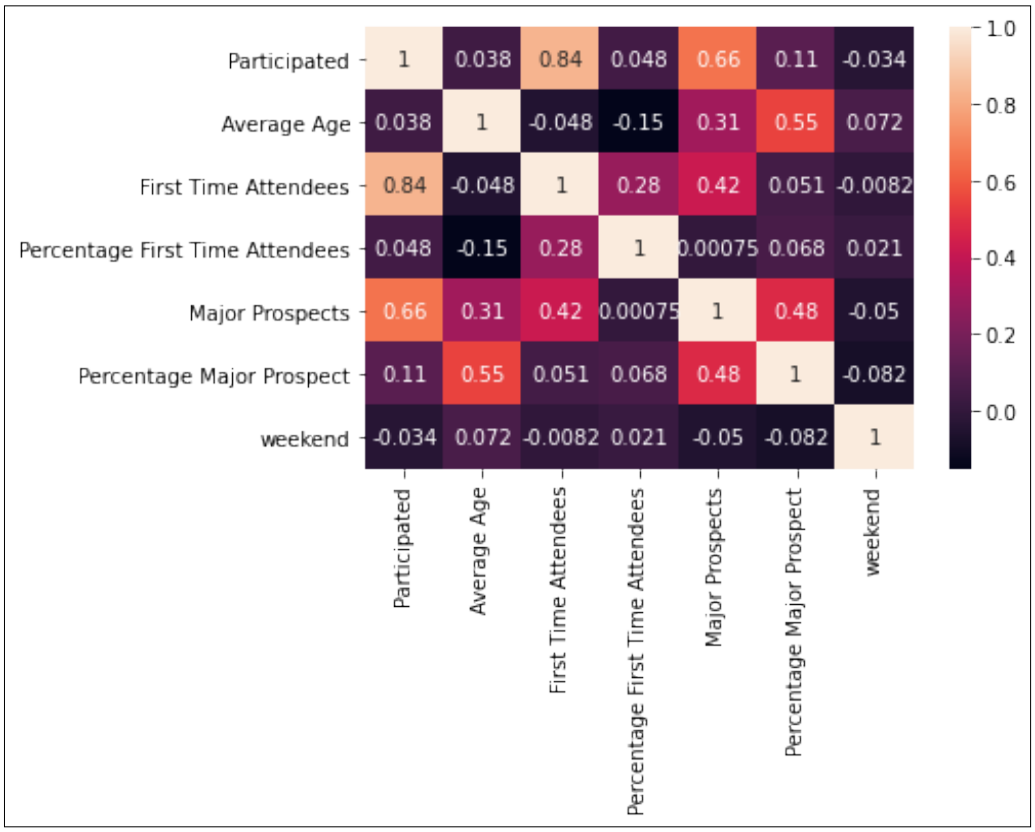

Figure 3: Correlation matrix

4. Percentage of events held in each location and each group (Figure 4):
    a. For *Location*, we can see the number of events held at location code PDON (on-campus - DMV area) is the most, with 18.6% of the total events.

b. For *Group*, the number of events held for group code PS9 (social-general type) is the most, with 28.3% of the total events.

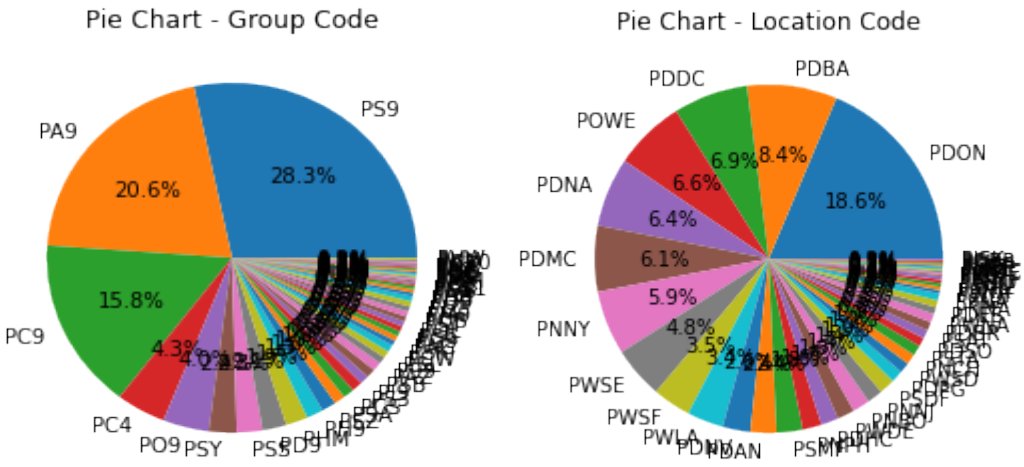

Figure 4: Pie charts of percentage of events held in each location and each group

- **Further Analysis**

  Our further analysis focuses on understanding how variables (Age, Location, Group) relate to first time attendees, as well as how variables relate to the major prospects.

  1. Impacts of Age on the First time attendees (Figure 5):

     From the scatter plots, we can see that the age of participants seems to not have conspicuous impacts on the number and percentage of the first time attendees.

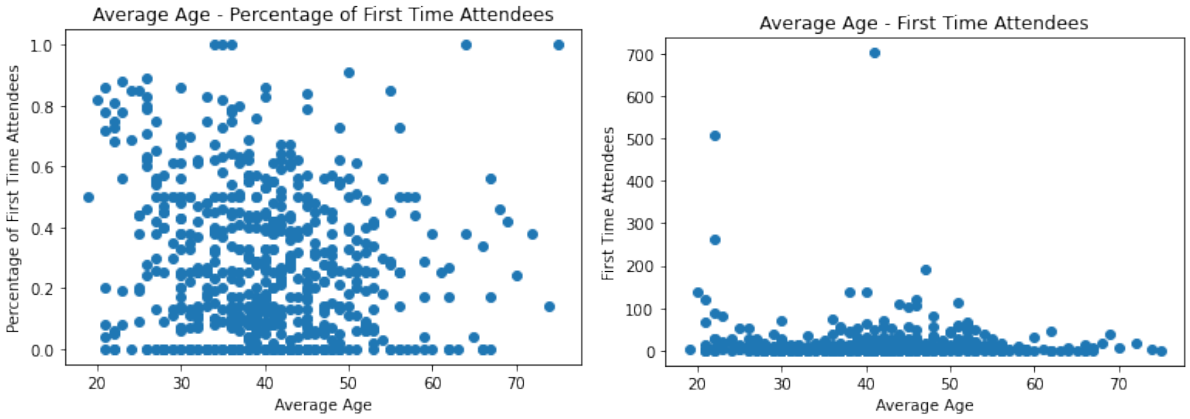

Figure 5: Scatter plots of Age on the first time attendees

  2. Age on the Major prospects (Figure 6):

     The age of participants seems to not have conspicuous impacts on the number of major prospects, yet there seems to have positive relationship between average age and percentage of major prospects (i.e., correlation coefficient = 0.55), so a possible strategy that UMD alumni association can take is to target the participants with higher age.

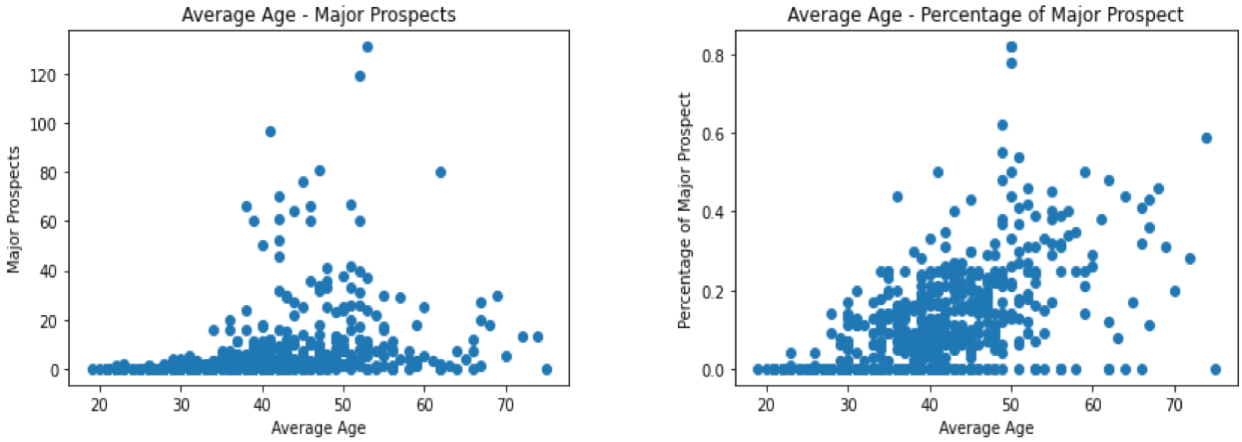

Figure 6: Scatter plots of Age on the Major Prospects

3. Location on the First time attendees (Figure 7):

   From the histogram below, we can see that PSAU (Southeast- Austin) contributes to the greatest number of the first time attendees; However, PSJA (Southeast-Jacksonville) has the highest average of percentage of the first time attendees.

   To Conclude, If we disregard the cost of holding an event, then holding an event at PSAU is possible to attract the most number of first time attendees. However, if we consider the cost, holding events at PSJA might have the highest ROI.

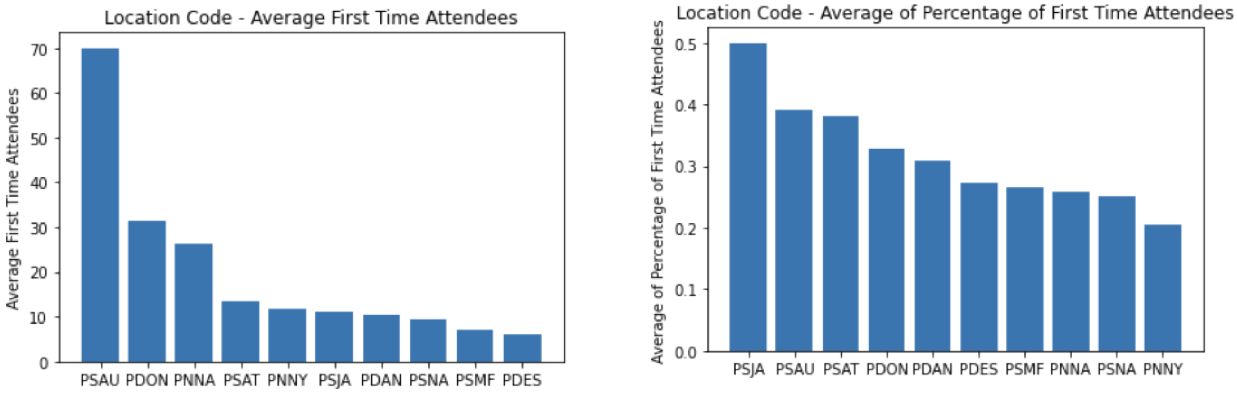

Figure 7: Histograms of the location on the first time attendees

4. Location on the Major Prospects (Figure 8):

   As shown in the left-hand side histogram below, PNNA (Northeast-General) and PSAU (Southeast- Austin), contribute to the greatest number of the major prospects.

   However, PSJA (Southeast-Jacksonville), PSNA (Southeast-General), and PNNA (Northeast-General), have the highest average of percentage of major prospects.

   With the similar concept, we can conclude that regardless of the cost of holding an event, holding events at Location PNNA is possible to attract the greatest number of major prospects. However, if we consider the cost, holding events at Location PSJA might return higher ROI.

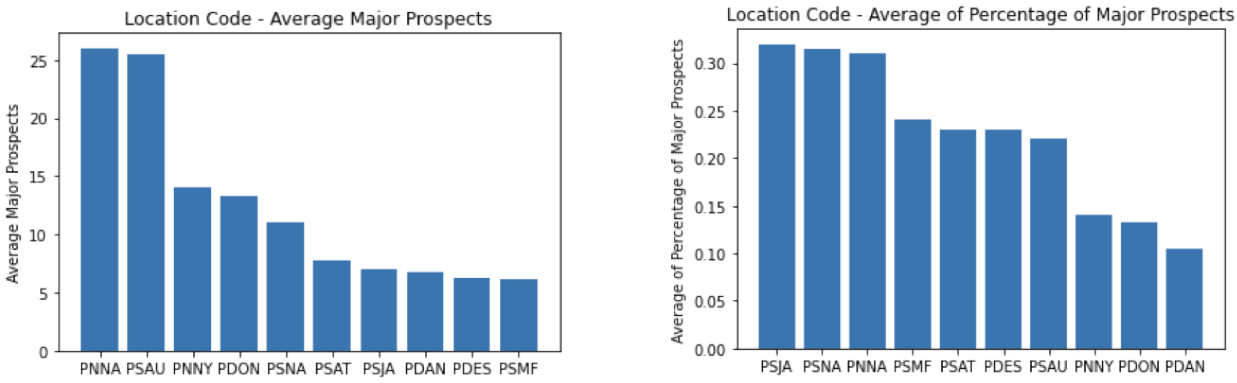

Figure 8: Histograms of the location on the major prospects

5. Group code on the First time attendees (Figure 9):

   For the impacts of group code on the first time attendees, as indicated in the histogram, PM9 contributes to the most number of the first time attendees, and P99 has the highest average of percentage of the first time attendees than other groups.

   Similarly, interpreting from the perspective of ROI, if we are not considering the cost, then holding events for PM9 would be more effective, if we are to consider the cost, then holding events for P99 would probably be more valuable.

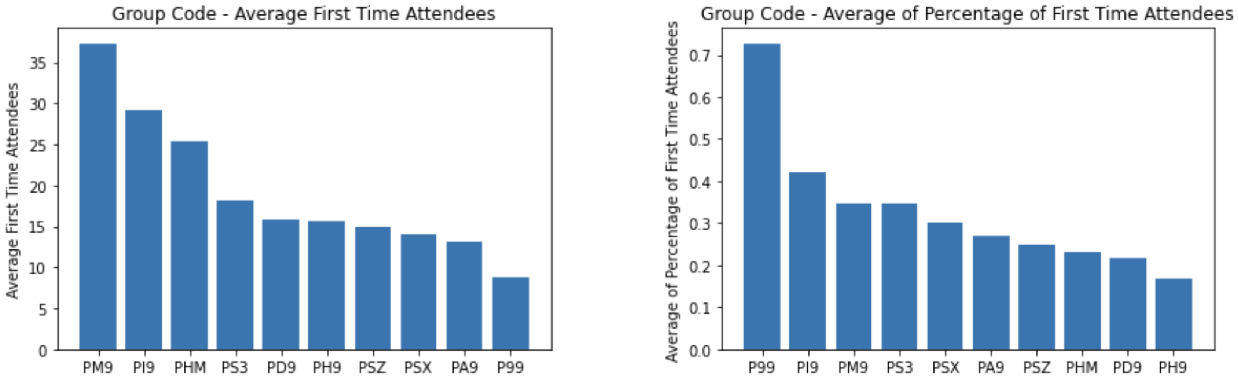

Figure 9: Histograms of the group code on the first time attendees

6. Group code on the Major Prospects (Figure 10):

   Lastly, we examine Group Code versus Major Prospects. The two histograms below present the same result, in which P99 contributes to the greatest number of first time attendees and accounts for the most of the percentage of first time attendees.

   Again, we can conclude that with or without considering the cost of holding an event, holding events for Group PH9 might have the highest ROI and is possible to attract the greatest number of major prospects.

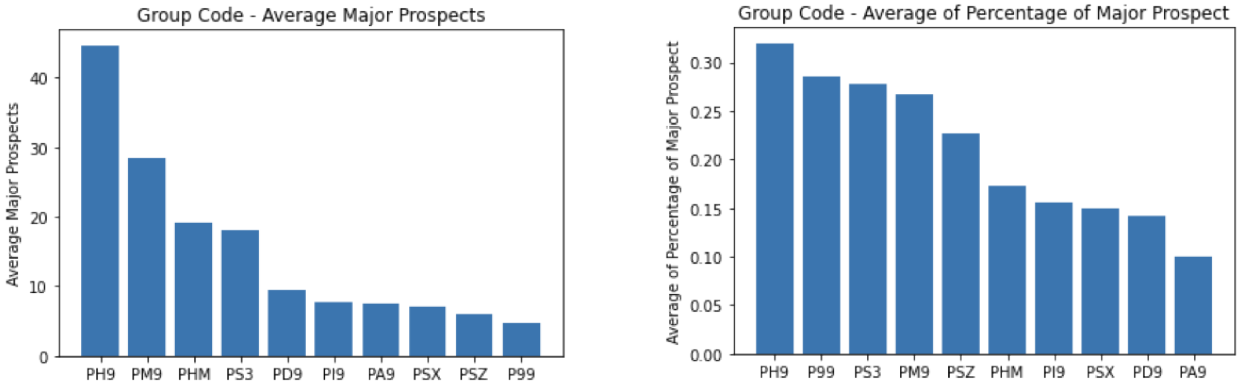

Figure 10: Histograms of the group code on the major prospects

7. Integration of Location and Group Code – First time attendees (Figure 11)

   Considering both locations and groups, the events held at Location PDON(DMV-On Campus) for Group PSS (Social-Students) is possible to attract the most number of first time attendees.

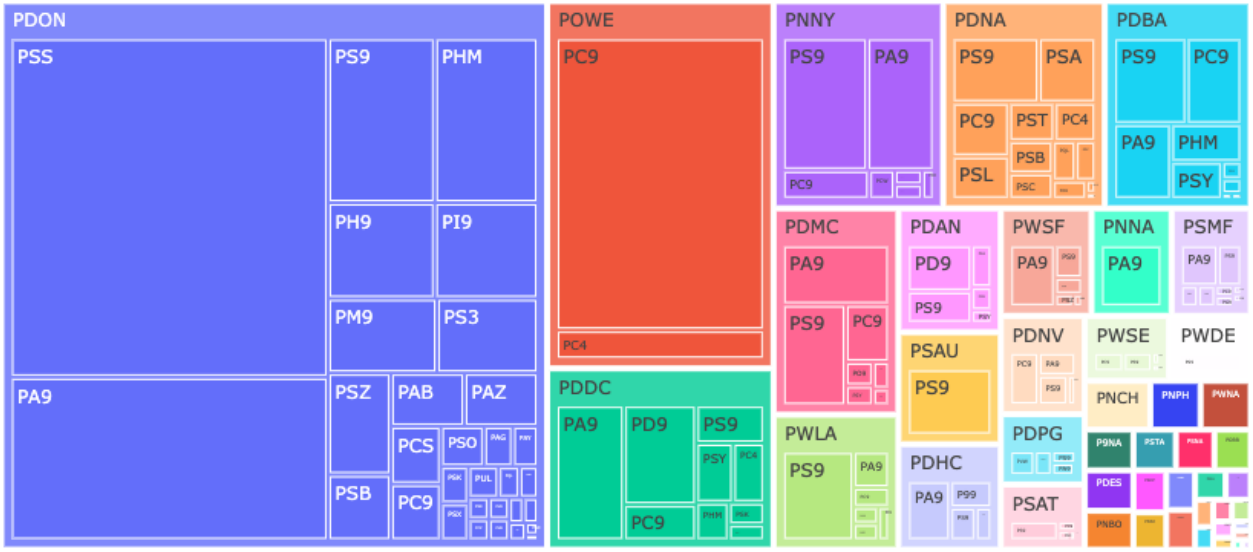

Figure 11: Treemap of Location and Group Code on the first time attendees

8. Integration of Location and Group Code – Major Prospects (Figure 12)

   Considering both locations and groups, the events held at Location PDON(DMV-On Campus) for Group PA9 (Athletics-General) and Group PH9 (Stewardship-General) is possible to attract the most number of major prospects.

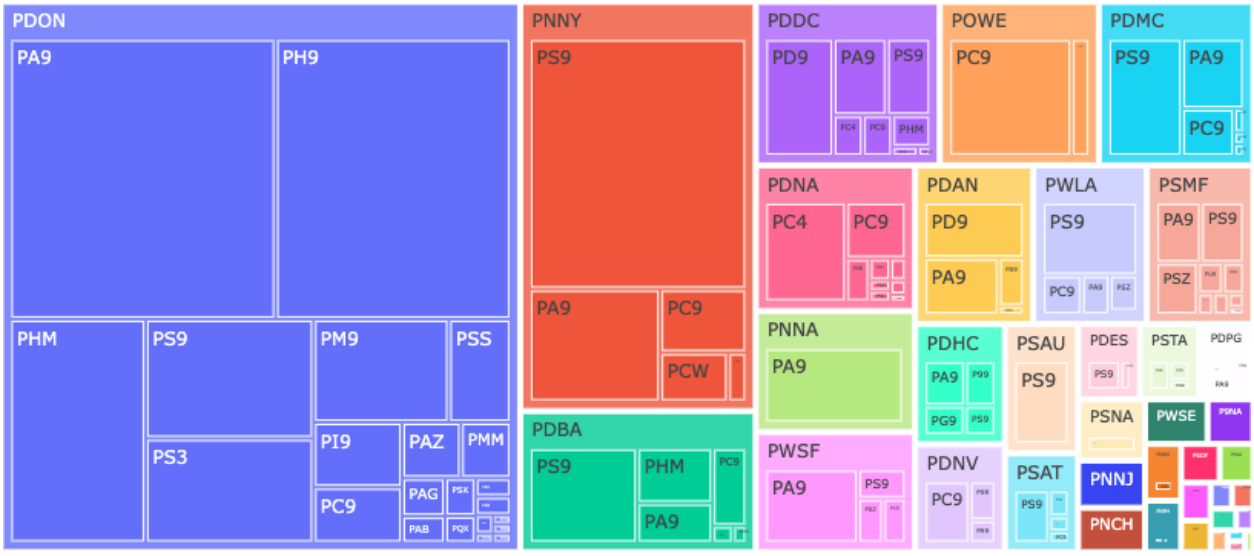

Figure 12: Treemap of Location and Group Code on the major prospects

9. Average age of each group – Percentage of the first time attendee (Figure 13)

   According to the data and pictures, we can see that:

   - The young are more likely to attend Career events. And their First Time Attendees is relatively low.
   - For older alumni, the General event attracts them most. And their First Time Attendees is the highest.
   - For middle-aged alumni, they are more interested in Cultivation, D&I and Service events. And their First Time Attendees is relatively high.

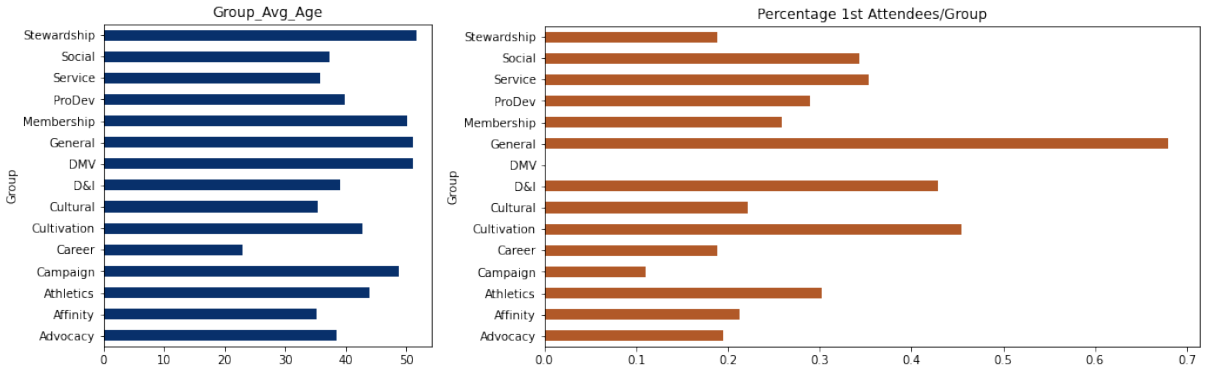

Figure 13: Average age of each group – Percentage of the first time attendee

10. Sublocation – Number of major prospects (Figure 14 and Figure 15)

    Besides, sub-locations were further specified by using the column “Location”. The total number of events held in each sub-locations and the average major prospects in each sublocation are calculated. According to Figure 14, we can see that:

    - Average Major Prospects of some cities are not positively related to numbers of events they held, such as On Campus, with 115 of the total number of events, but only 13.357 of the average number of major prospects.
    - Some cities' Average Major Prospects are obviously higher than others, such as Austin and Fort Lauderdale; on the other hand, some are significantly lower, such as Brazen and Western MD.
    - Finally, we can hold an event in a specific city, such as New York in the Northeastern U.S., Austin and Fort Lauderdale in the Southern U.S., San Francisco in Western U.S., which may be helpful to increase the Major Prospects for the UMD alumni association.

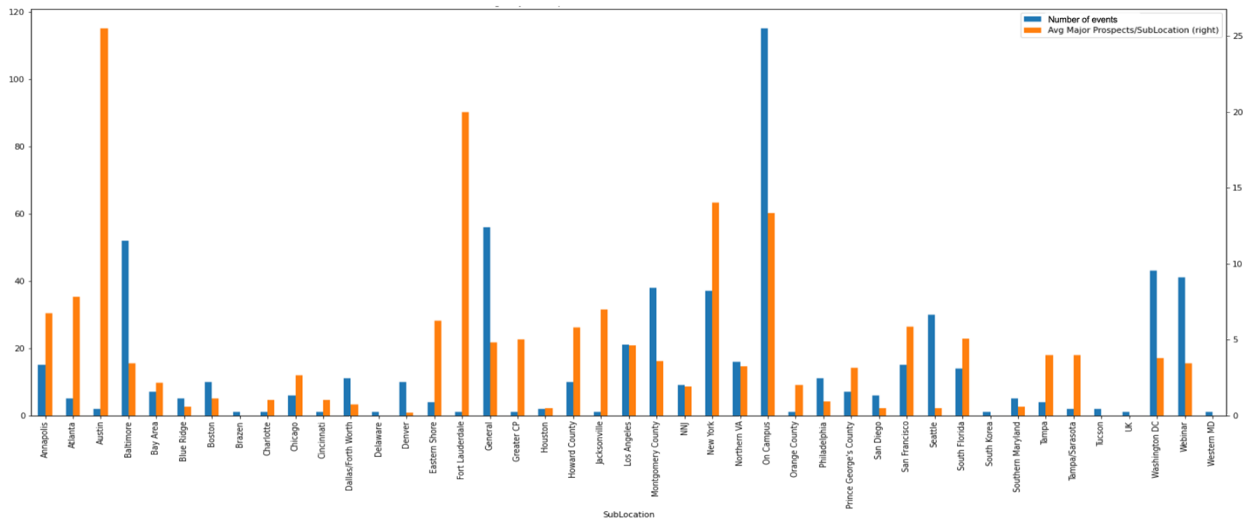

Figure 14: Sublocation – Number of major prospects

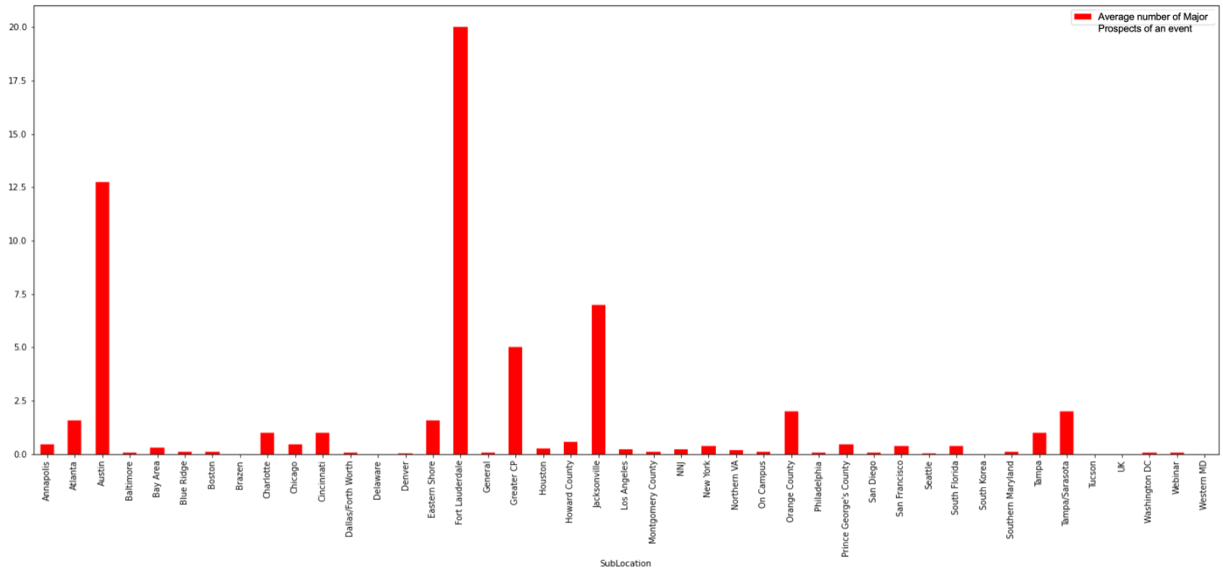

Figure 15: Average number of major prospects for each event in each sub-location

Recommendation

Based on our analysis, we can conclude two main recommendations to achieve our project goal of increasing event attendance of the first time attendees and major prospect attendees. On campus, we could hold more events that attract social general groups since those events usually have more first time attendees. Also, holding more events that attract Athletic and Stewardship are helpful for increasing the major prospects. Off campus, we could hold events that attract Cultural General Groups at SouthEast Jacksonville and Austin, which might help to increase most first time attendees. Also, holding events that attract Stewardship Group at NorthEast Region and SouthEast Austin would have more influence on increasing the major prospects. Besides, we should try to hold more events on Friday, since Friday has the highest average number of participants.

ON CAMPUS	
First Time Attendees	Major Prospects
Events that attract Social General Groups	Events that Attract Athletic and Stewardship Events
OFF CAMPUS	
First Time Attendees	Major Prospects
Location: SouthEast Jacksonville and Austin Group: Events that attract Cultural General Groups	Location: NorthEast Region and SouthEast Austin Group: Events that attract Stewardship Group

Future Work

From analyzing the data, we could come up with conclusions on different parameters such as the day of the week, Location, Group composition, and Age to host the next event to increase the First Time attendees and Major Prospects.

However, the recommendations are limited to pooling the maximum number but do not necessarily give an insight on how much each factor affects the output statistically. In order to make this possible, we came up with two regression models one to predict/estimate the First Time Attendees and the other to estimate the turn up of Major Prospects.

The Data is first clustered with its highest correlated parameter to remove the outliers, and this potentially helps with deciphering any structure or pattern in data that could be clogged with outliers. The number of clusters are evaluated based on the sample data considered. Now let's Understand each Model:

- **First Time Attendees Prediction Model:**

  In the correlation heat map, we can see high correlation between Participated and Number of First time Attendees. Hence, we have clustered the data plotted across Participated and Number of First time attendees, as presented in Figure 16. We have simulated it by increasing the number of clusters to get a point where the outlier data cluster has less than 25 data points. This would imply that the regression algorithm is run on more than 95% (612 in this scenario) of the data set, leading to a much more reliable result.

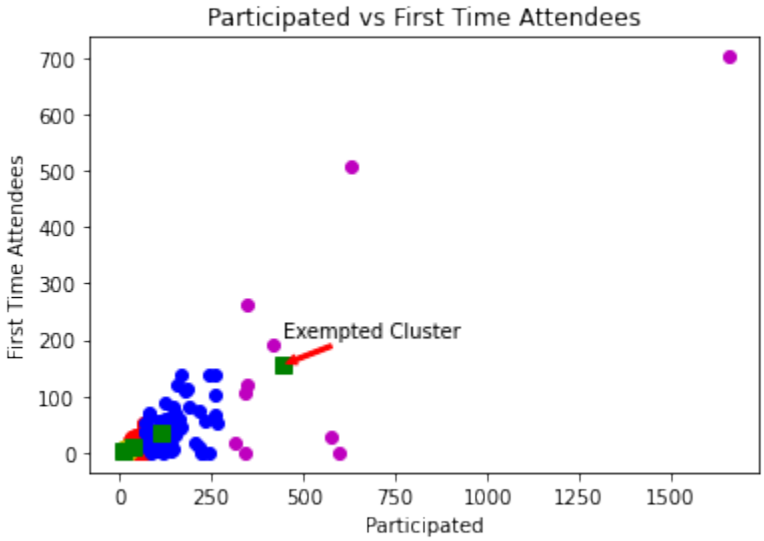

Figure 16: Clustering analysis with respect to participated and the first time attendees

For the regression analysis, we have considered First Time Attendees as the dependent variables, and Participated, Average Age, and Estimated number of Major Prospects as the independent variables. Note that the Group Code, Activity Code, and Location Code have multiple subcategories which add very little significance to the result and hence omitted any categorical variable to the independent variable list.

In consequence, we have considered the number of participants, average age and the estimated number of major prospects as independent variables. Additionally, we also observed that there is high correlation between Average age and Major Prospects, thus the interaction term has been considered to be included in the regression analysis. The Clustered data produced the below regression equation:

$$Interaction = Average\ Age \times MajorProspects$$

$$First\ Time\ Attendees\ (FTA) = 6.647 + 0.333 \times Participated - 0.194 \times Average - 0.797 \times Major\ Prospects + 0.013 \times Interaction$$

We can clearly see from the equation that the FTA are positively correlated to Participated and negatively correlated to Average age and Major Prospects, the same can be inferred from the Correlation Heat Map as well. Hence, we can say that on an Average with an increase in 3 participants the number of the first time attendees increases by 1 when other factors remain unchanged. Similarly, when the Average age and the Major Prospects increase would lead to a decrease in the number of FTA.

The Regression model when tested on the sample data produced the following results, with predicted values and the actual values plotting in Figure 17:

R^2 score is 0.585

Mean Squared Error is 136.849

Mean Absolute Error is 6.905 %

Root Mean Squared Error is 11.698

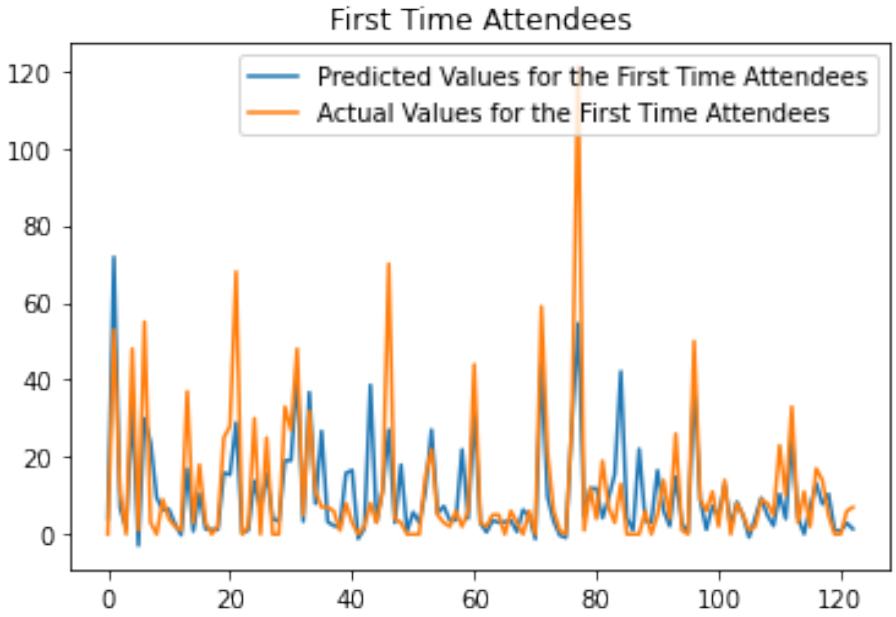

Figure 17: Predicted values for the number of first time attendees – actual values for the number of first time attendees

- **Major Prospects Prediction Model:**

  Similar to that of Number of first Time attendees, we can see from the correlation heat map that there is a high correlation between Participated and Number of Major Prospects. Hence, the Clustering, as shown in Figure 18, was performed with the scatterplot of Major Prospects across Number of Participants. The same simulation technique was followed to determine the number of clusters (i.e by increasing the number of clusters to get a point where the outlier data cluster has less than 25 data points). This would imply that the regression algorithm is run on more than 95% (612 in this scenario) of the data set, leading to a much more reliable result.

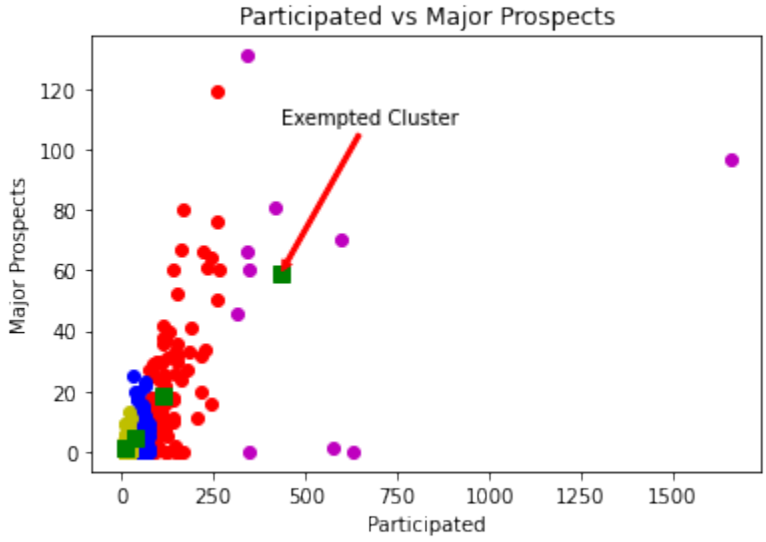

Figure 18: Clustering analysis with respect to participated and the major prospects

We have considered Number of Major Prospects as the dependent variables, and Participated, Average Age, and the estimated number of First Time Attendees as the independent variables. The Group Code, Activity Code and Location Code have multiple subcategories which add very little significance to the result and hence omitted any categorical variable to the independent variable list. In consequence, the Clustered data produced the below regression equation:

$$Major\ Prospects = -13.243 + 0.199 \times Participated + 0.281 \times Average\ Age - 0.021 \times First\ Time\ Attendees$$

We can clearly see from the equation that the Major Prospects are positively correlated to Participated, Average Age and negatively correlated to First Time Attendees, the same can be inferred from the Correlation Heat Map as well. Hence, we can say that on an Average with an increase in 5 participants the number of Major Prospects increases by 1 when other factors remain unchanged. Similarly, when the Average age increases by 4, the number of Major Prospects increases by 1 when other factors remain unchanged. Furthermore, the First Time Attendees increase would lead to a decrease in the number of major prospects.

The Regression model when tested on the sample data produced the following results, with predicted values and the actual values plotting in Figure 19:

R^2 score is 0.664

Mean Squared Error is 41.798

Mean Absolute Error is 4.042 %

Root Mean Squared Error is 6.465

Comparatively, it appears to be a much efficient prediction model when compared to the previous model, with Higher R^2 and lower error values.

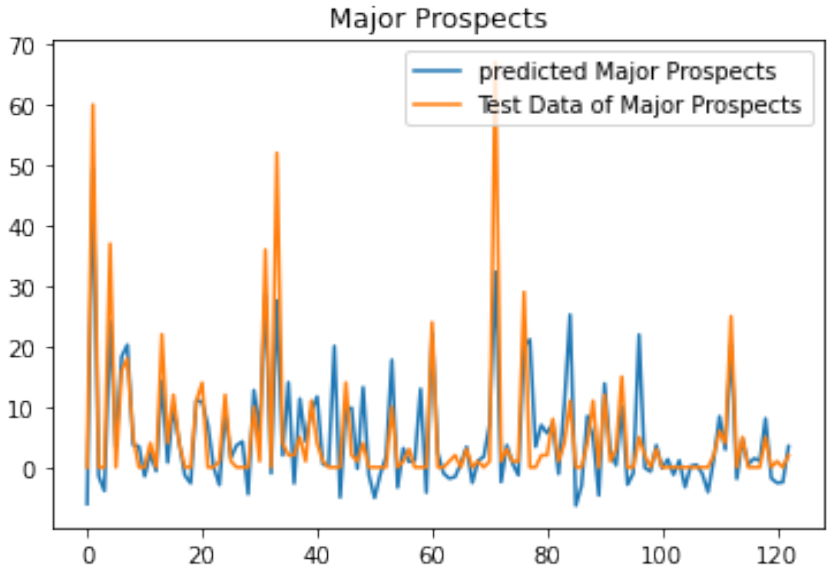

Figure 19: Predicted values for the number of major prospects– actual values for the number of major prospects