**UMD ALUMNI EVENTS ANALYSIS**

**SUBJECT CODE: BUDT704**

**SUBJECT: PYTHON**

**SUBMITTED BY TEAM 14**

Jinping Guo

Siying Li

Sowmya Murukuti

Tanya Singh

**Introduction:**

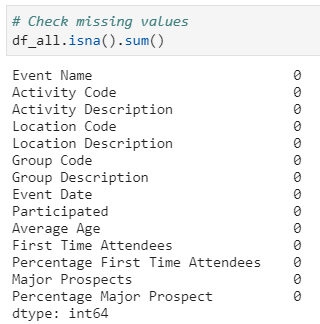
UMD Alumni Events Analysis would like to find variables that correlate with higher event attendance of first-time attendees and prospects for major gifts substantially. With these correlations, we can enhance both new and existing alumni engagement events to serve better to these two groups. As a result, we need to understand which all events attract the most first-time attendees and major gift prospects so that we can use this information in planning future events.

**Methodology :**

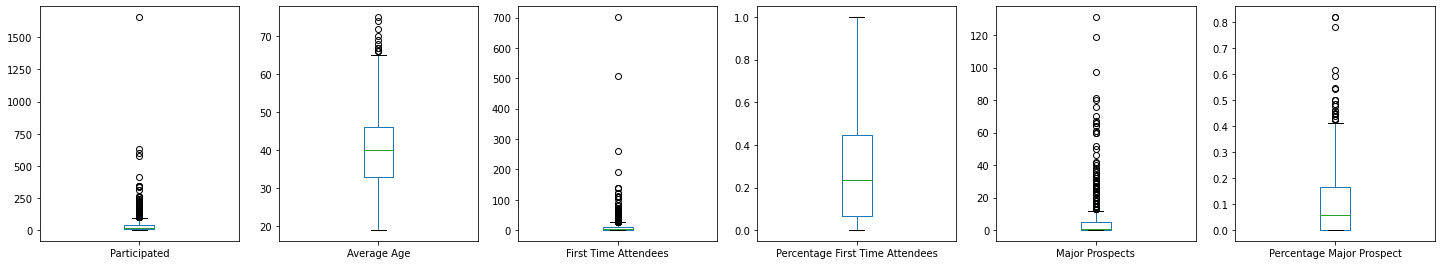
The descriptive analysis [4] method is applied to process our data since more than half of the data are categorical. Initially, we organize and summarize a large amount of raw data information, and then analyze the concentration trend and dispersion trend by describing the inner pattern of the raw data. In our report, the major descriptive methods include the analysis of concentration trends, the analysis of data dispersion, the analysis of frequency distribution of data, and the analysis of variable correlation. Generally, descriptive analysis is the basis for analyzing categorical data.

Firstly, data preprocessing [3] The purpose of data preprocessing is to improve the quality of the raw data by the way of cleaning the abnormal data (abnormal data here refers to data that has no practical significance for data analysis, or has an abnormal format, or is not within the specified range) and finding the data distribution. Finally, after data preprocessing, the raw data will have the characteristics of integrity, uniqueness, consistency, etc., and better show the relationship between them.

In the first step, isnull() function [1] is used for checking missing data, which reduces the negative impact and increases result accuracy. (The result shows in 1.1)



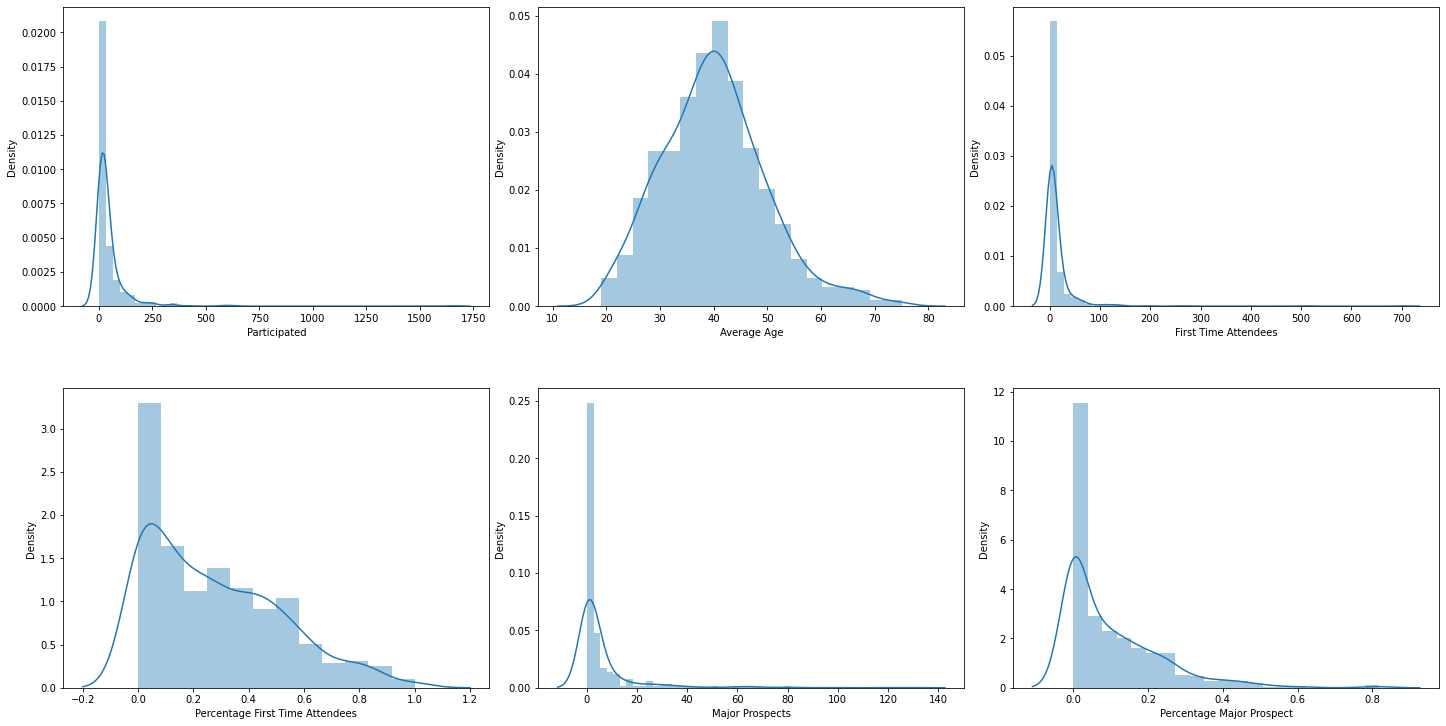
(1.1)

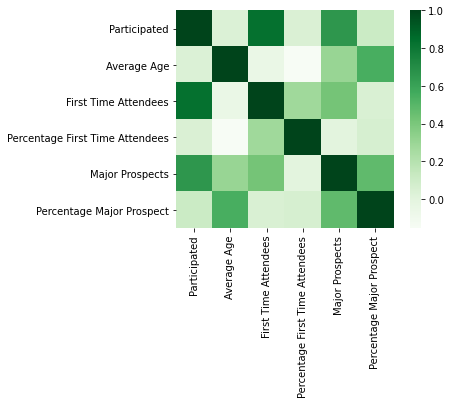
Secondly, to evaluate the outlier. Outlier is the abnormal data that deviates significantly from the rest of them. In our case, the box plot is used for analyzing the variables outliers and the results shown in 1.2. However, it should be noted that the outlier only means that the value is abnormal but not absolutely equal to the wrong value. Therefore, although all variables exist outliers, we still couldn’t simply exclude them. Instead, the treatment of outliers depends on the specific analysis: to determine whether they are reasonable. After analyzing the data frame, we conclude that all the outliers are meaningful. For example, even though most of the events were smaller in scale(no more than 100 people), there were still several events with attendees greater than a thousand. Eventually, after evaluating the outlier, we also keep all of them in our dataframe.

(1.2)

Before doing variable dimensionless normalization, we need to analyze the distribution of the data first. For example, normalization requires that the distribution must obey a normal or skew-normal distribution, and the continuous variables ( such as average age in our project ) also need to determine their unit of frequency based on their distribution. Thus, we plot our variable distribution to analyze their density.

This histogram (1.3) shows that the Average Age is the approximately normal distribution. The rest of the numeric columns are highly right-skewed. In this step, analyzing the distribution of the raw data gives us a general profile of the past activities scales and the age segmentation of the participants, which could give insight to plan the future activities.



(1.3)

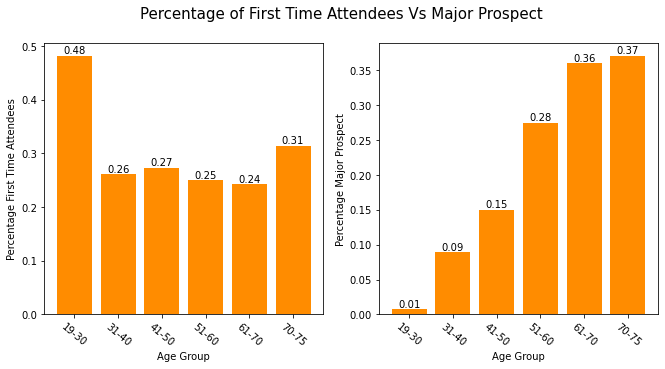
Next, we need to analyze the data correlation, in other words, we need to find variables that are highly correlated with first-time attendees and major gift prospective, separately. The heat map [2] could achieve this goal by color difference. Therefore, the heat map shows (1.4) that the participating people and age group contribute to our results significantly.

The two primary methods of processing data for descriptive analysis are data aggregation and data mining. Based on the correlation analysis, we have an initial idea to measure the dimensionality of the variables, which was to subdivide the participants by age in tens and the activities by time in months. In addition to that, to better find which variables are effective, we need to standardize the metrics - that is, normalization with dimensionless parameters, by comparing their percentage. Since we regrouped the data, we needed to recalculate the average major prospects rate and first-time attendance rate.

**Findings & Results:**

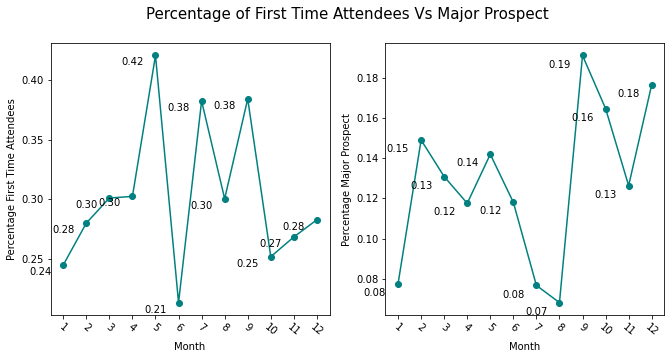
**The percentage of FTA(First Time Attendees) and MP(Major Prospect) for each age group.**

The average age was divided into 6 age groups. For the major prospect, there is an increasing trend of percentage of major prospects, as age goes up. The highest percentage is 37% in the age group 70-75, and the lowest percentage is 1% in the age group 19-30. This means the older alumnus are more likely to donate. However, alumni aged between 19-30 have the highest percentage of first-time attendees is 48%. Based on this finding, we would recommend the Alumni Association target alumni aged from 61 to 75 to get more prospects, increasing the attendance rate of first-time attendees by inviting alumni aged between 19 and 30.

****

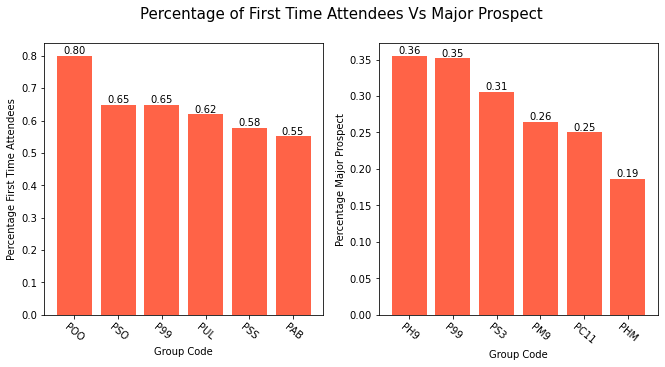
**The percentage of FTA and MP for each month.**

The month with the highest median percentage of first-time attendees is May(42%), July and September also have a high percentage. According to the left graph below, June and January can be regarded as the two months with the lowest median percentage of first-time participants in the whole year, so these two months should not be the first choice. The right graph shows the percentage of major gift prospects for each month. Obviously, September and December attracted the highest percentage of potential customers for major gifts. The lowest percentage of major gift prospects attracted in January, July, and August. September will be a good choice to host more events to attract more first-time attendees and donors as well.

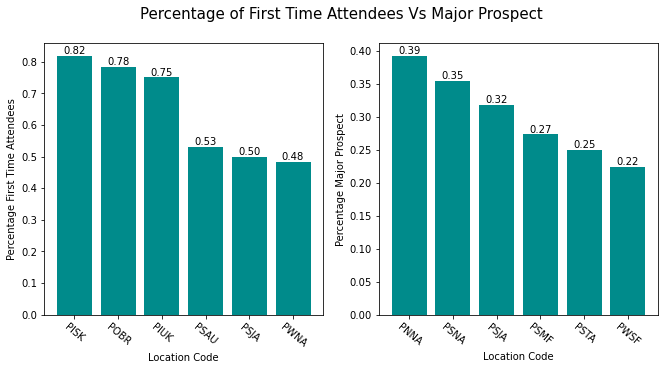
****

**The top 6 groups that have the most percentage of FTA and MP**

The graph POO has the highest percentage of first-time attendees, 80%. In addition, there is not a big difference between every two adjacent groups. Groups like PSO, P99, PUL also have high attendance rates, 65%, 65%, 62% respectively. PH9 has the highest percentage of major gift donations, with 36%, and the P99 group is the second highest. And I notice the P99 group shown in both below the graph. What I will guess is that this group has more new people and has relatively nice financial strength. I will recommend targeting this group to increase more first-time attendees and major prospects.

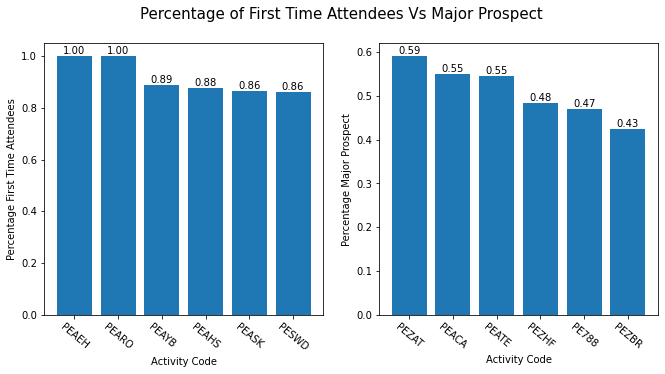
****

**The top 6 Locations that have the most percentage of FTA and MP**

It's clear to see the location: PISK, POBR, and PIUK attracted more than 70% first-time attendees. However, when I check the raw data for those locations, there are only a few locations data of that location. Therefore, if we want to obtain more accurate information, we need to gather more data to predict them. As for the major prospects, the PINNA, PSNA, and PSJA are favorite locations to attract more major prospects. In the future, we could find the meaning behind those group codes, then classify those group codes specifically, which will help us to find more accurate information. 

**The top 6 Activities that have the most percentage of FTA and MP**

Activities like PEAEH, PEARO have a super high percentage of first-time attendees, with 100%. However, I double-checked the data for this, I found those two activities only organized once, so which is not representative. Therefore, we could increase the event PEAYB, PEASK, and PESWD. As for the major prospect, the activity that has the highest percentage of major prospects is PEZAT, 59%. We recommend increasing the PEZAT and PEACA activities, and decreasing the number of PEZBR.

****

**Recommendations**

From our charts and basic analysis, we believe the following recommendations can be considered to reach our mission objectives:

1. **To increase the percentage of first-time attendees:**

* Since the 19-30 years age group has the highest percentage of first-time attendees, we can target market our events in that age group.
* Since more first-time attendees are observed around May and fewer number of them are present in June, it is a good idea to have regular events during the month of may and fewer events during the month of June.
* CP International- South Korea’s location code has the highest percentage of first-time visitors, so more events can be held in this location code.
* Group code POO has the highest number of first-time participants so it is a good idea to implement more marketing around this group to attract more first-time attendees.

1. **To increase major gifting prospects:**

* Since active participation was majorly observed in age groups 70-75, more marketing can be done around those age groups to increase funds for that year.
* More events can be conducted during September to increase the participation of people who contribute to gifting prospects. Fewer events can be organized during the month of August since the number of people donating this month is minimum.
* Gifting prospects are found to be highest in location code CP- NOrtheast- General, so it is a good idea to conduct more events in this location code to collect high donations.
* More marketing can be done for groups PH9 and P99, this way.

**Future Works**

**1. Use Machine Learning for better strategy formation.**

Instead of just using descriptive analytics and forming strategy manually, we can automate the process by using Machine Learning models. This makes the process much more efficient and robust.

**2. Use Cloud Storage and availability of Long - term Data Structure.**

Usage of Cloud-based storage instead of the conventional methods will help make the data more secure and easily accessible. The threat of Data loss/ erasure is removed and we can also provide limited access to people depending on their needs.

**3. Gather More Demographic Details.**

If we have more data, we can be more sure about our analysis and use it to build better models. We can also find out more features that affect our target variables.

Reference:

[1] Perera, P. Prasad. (2017). *The Boston Housing Dataset*. Retrieved October 6, 2021, from <https://www.kaggle.com/prasadperera/the-boston-housing-dataset/notebook>

[2] P, D. (2016). *A Heat Map to Visualize Mean Crowdedness*. Retrieved October 6, 2021, from <https://www.kaggle.com/demetripananos/a-heat-map-to-visualize-mean-crowdedness>

[3] Jain, D. (2021, July 29). *Data Preprocessing in Data Mining*. Retrieved October 12, 2021, from <https://www.geeksforgeeks.org/data-preprocessing-in-data-mining>

[4] Rawat, A. S. (2021, May 31). *An Overview of Descriptive Analysis*. Retrieved September 12, 2021, from <https://www.analyticssteps.com/blogs/overview-descriptive-analysis>