**BUDT704\_0507\_03**

Xinyuan Chen, Sara Kamal, Quikai Chen, Oluwadamilola Yusuf

1. **Constructure of our Python codes**

Our Python codes can be viewed as six parts.

* Part One (Step 1 & 2): Prepare the data
* Part Two (Step 3 & 4): Initial observation for variables
* Part Three (Step 5): Preliminary analysis to explore the factors which influence First Time Attendees
* Part Four (Step 6): Preliminary analysis to explore the factors which influence Major Prospects
* Part Five (Step 7): Further analysis for Part Three and Part Four
* Part Six (Step 8): Other analysis: area comparison

1. **Method used for analysis**

The main methods we used include:

* **Bar chart & Histogram**

We mostly used bar charts and histogram in Part Two, to show many of our initial observations such as number of activities held in each quarter, area, etc., and the distribution of numerical variables such as Participated, Percentage First Time Attendees, etc.

* **Boxplot**

In Part Three and Part Four, most of our analysis were conducted by boxplots, to observe to the mean, median and distribution of Percentage First Time Attendees/Major Prospect or number of First Time Attendees/Major Prospects among different categories within one factor.

* **k-means clustering**

We used 3-means clustering to show relationship between number of Participated alumni and number of First Time Attendees, and relationship between number of Participated alumni and number of Major Prospects in Part Three and Four.

* **Multiple regression**

In Part Five, we conducted multiple regressions to further analyze which factors influenced First Time Attendees and Major Prospects. According to regression output, we picked the ones with high positive coefficient and small p values simultaneously to be the significant factors.

We also used other methods, such as analysis of variance, colormap etc. More detail about how we applied these methods is showed in the next part “*Detailed introduction for each part of our Python codes*”.

1. **Detailed introduction for each part of our Python codes**

*# The charts we put in this part are only the examples that help to explain what we did. If you want to see all the charts or tables in our project, please read .ipynb file.*

* **Part One: Prepare the data**

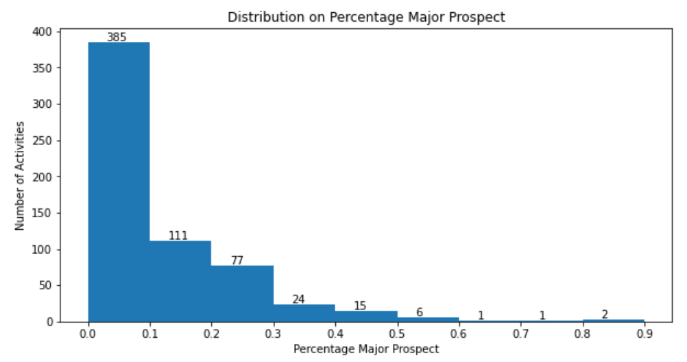
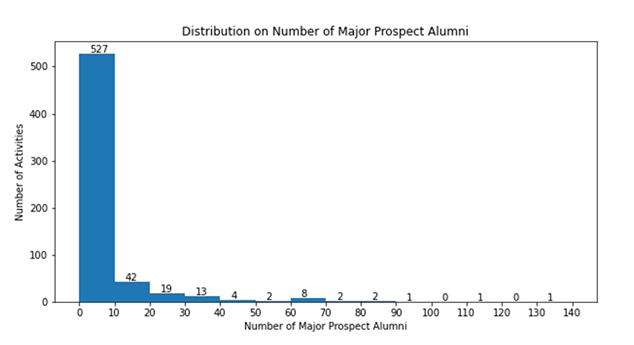
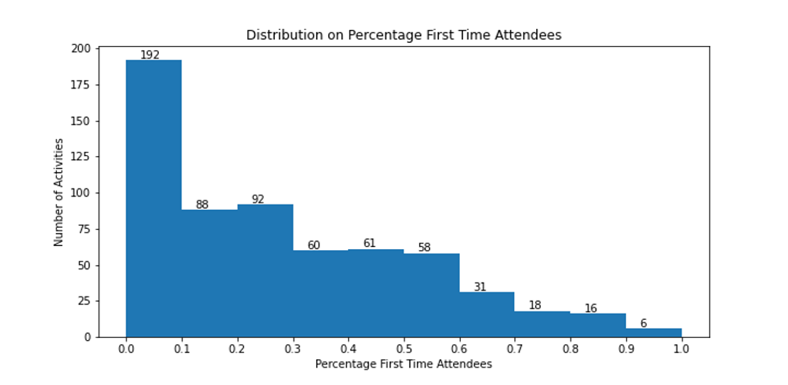
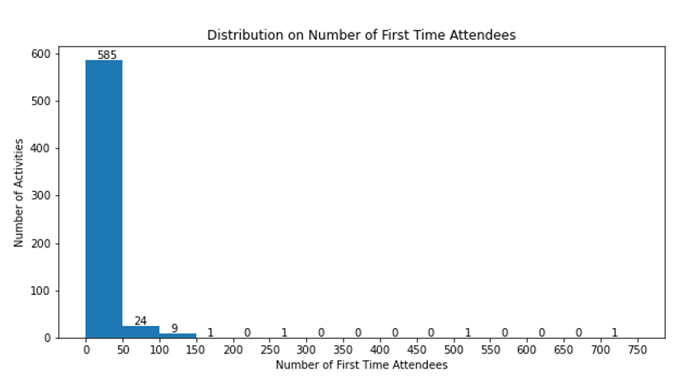
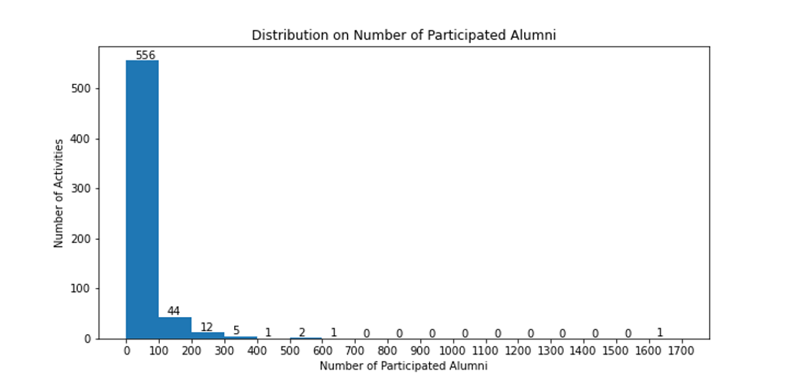
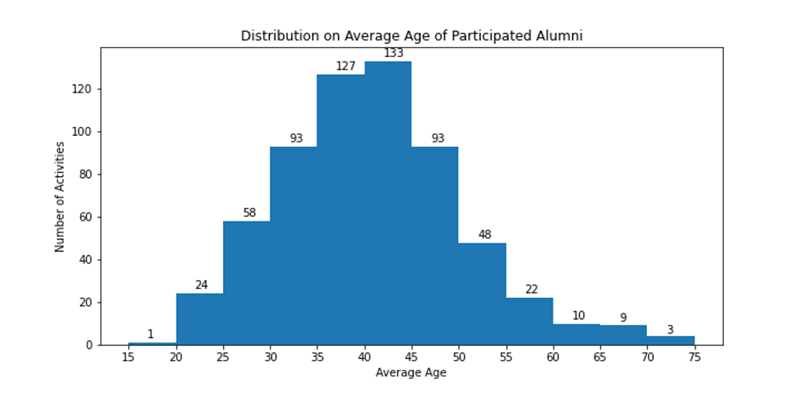
To start the assignment, we tried to prepare the data by concatenating all the years’ data into one big data frame, added a column called “Year” and detected whether there are some missing values.

* **Part Two: Initial observation for variables**

To start our analysis, we did initial observation for each variable first, to help us know what information we had and the overall situation of UMD alumni activities in the past several years. Then, we decided the variables we wanted to analyze.

In this stage, we regarded Location, Group, Event Date and Participated as the factors which may influence First Time Attendees and Major Prospects. We measured First Time Attendees and Major Prospects by both percentage and number of first time attendees/ major prospects.

For numerical variables, we got the descriptive statistics and distribution of them. Then we find that most of activities had less than 100 participants and showed average age from 25 to 55 years old. We also observed that most activities had a percentage of first time attendees/ major prospect less than 0.3, number of first time attendees of less than 50, and number of major prospects of less than 10.



Chart

Description automatically generatedFor Location, by extracting the second letter of Location Code, we got the activity-held area and created a new column named “Area” whose values include Online, DMV Area, Northeast Area, Southeast Area, West Coast Area, International, and USA. When looking at the number of activities held in each area, we can see that over 300% more activities were held in the DMV compared to other areas.

Histogram

Description automatically generatedFor Group, by extracting the second and the third letter of Group Code, we got the intended purpose and intended audience for each row and we created two new columns named “Purpose” and “Audience”. By bar chart to show the number of activities held for each purpose and audience, we found that intended purpose of Social, Athletics, ProDev were most common for activities and general kind of intended audience were most often observed.

A screenshot of a computer

Description automatically generated with low confidenceFor Event Date, we transformed the date to quarter and weekday. Then we looked at how many activities were held in every quarter and on each weekday. And we find that Quarter 4 (Oct-Dec) and Saturday held the highest number of activities.

* **Part Three: Preliminary analysis to explore the factors which influence First Time Attendees**

A picture containing chart

Description automatically generatedThe first method we used was to choose top 10 Percentage First Time Attendees/ number of First Time Attendees in each year, to see their locations, intended purpose etc. For example, we found most of top 10 Percentage First Time Attendees activities were held in DMV area. But only by this we could not draw the conclusion that holding activities in DMV area could attract more first time attendees, because when we choose Bottom 10 ones, most of them were also held in DMV area.

So, we used the second method -- boxplot, to show the difference of Percentage First Time Attendees or number of First Time Attendees among different categories within one factor. We drew the boxplots by plt.boxplot() for Area, Purpose, Audience, Quarter, and Weekday variables, and we found that percentage of first time attendees or number of first time attendees did not change much among different weekdays, which we regarded as the insignificant factor. But within some factors, some categories had obviously higher mean or median of percentage/ number of first time attendees, which may be significant factors. Finally, we concluded that Area, Purpose, Audience, and Quarter may be significant factors for Percentage First Time Attendees, while Purpose and Audience may be significant factors for number of First Time Attendees.

A picture containing box and whisker chart

Description automatically generated

For Activity variable, we could not treat it as Location variable to divide them into several categories, since it had so many kinds of activities. What we did was to preview both top 10 and bottom 10 Percentage First Time Attendees/ number of First Time Attendees, and picked some special events to observe the times they showed among top 10 and bottom 10 activities. Then we found some kinds of events showed much more times among top 10 and less time among bottom 10, such as gala and party, which might be more popular events. In contract, some events showed much more times among bottom 10 and less time among top 10. For example, Game Watch showed 19 times among bottom 10 number of First Time Attendees while only 3 times among top 10. To attract more first time attendees, we will recommend the ones which showed much more times among top 10.

For Participated, we used 3-means clustering to explore the relationship between number of Participated alumni and number of First Time Attendees, then we found the positive relationship between them.

Chart, scatter chart

Description automatically generated

* **Part Four:** **Preliminary analysis to explore the factors which influence Major Prospects**

To do the preliminary analysis on Major Prospects, we did the same thing that was done for First Time Attendees, including (1) looking at the distribution of different factors among top 10 and bottom 10 Percentage Major Prospect and number of Major Prospects, (2) drawing boxplots to see the mean and median of percentage/number of major prospects among different categories within one factor, (3) showing the times of different kinds of activities appearing among top 10 and bottom 10 Percentage Major Prospect and number of Major Prospects, and (4) 3-means clustering to see the relationship between number of Participated alumni and number of Major Prospects.

What we found was that Area, Purpose, Audience, Activity and Participated may be significant factors which influences Major Prospects, while Weekday and Quarter are not significant ones.

* **Part Five: Further analysis for Part Three and Part Four**

In Part Three and Part Four, we mainly explore the factors from charts. In this part, we planned to use quantitative methods -- multiple regression, to do further analysis. Before multiple regression, we did some other analysis.

The first one is to calculate the correlations between numeric variables. By plotting color map of correlations, we can see that the deeper the color it shows, the higher the correlation is. We noticed that correlation between Average Age and Percentage First Time Attendees is the only negative one, thus we can invite more elder people to increase the Percentage First Time Attendees. But on the other hand, increasing the average age would decrease the Percentage Major Prospect. Therefore, how to balance them is a tricky problem. In addition, we found that the correlation between Average Age and Percentage Major Prospect is 0.55, the correlation between Average Age and Major Prospects is 0.31, thus we can include Average Age in our multiple regression model on Percentage Major Prospect or Major Prospects.

Chart, treemap chart

Description automatically generated

Then, we did a simple regression between Average Age and Percentage Major Prospect, and another one between Average Age and Major Prospects to further explore their relationships. According to the regression output, we can know that linear model is better for both, and Percentage Major Prospect will increase by 0.0074 when Average Age increases by 1 and Major Prospects will increase 0.4470 when Average Age increases by 1 if anything else remain the same.

We also conducted an analysis of variance to exclude some insignificant categorical variables before multiple regression. According to the output, we found that different levels of Weekday does not have significant different impact on Percentage First Time Attendees and Major Prospects, different levels of Purpose, Quarter and Weekday do not have significant different impact on First Time Attendees, and different levels of Quarter, Weekday do not have significant different impact on Percentage Major Prospect.

Then, we began to conduct multiple regression on Percentage First Time Attendees, First Time Attendees, Percentage Major Prospect, and Major Prospects. What we know is that International Area, General Purpose, Student and Muslim Audience, Quarter 3 have a statistically significant positive effect on Percentage First Time Attendees while Average Age has a statistically significant negative effect on Percentage First Time Attendees; Student Audience, Participated have a statistically significant positive effect on First Time Attendees while Average Age has a statistically significant negative effect on First Time Attendees; Northeast area, Southeast area, USA, Stewardship Purpose, Mid Career Audience and Average Age have a statistically significant positive effect on Percentage Major Prospect; Northeast area, Membership and Stewardship, Participated and Average Age have a statistically significant positive effect on Major Prospects.

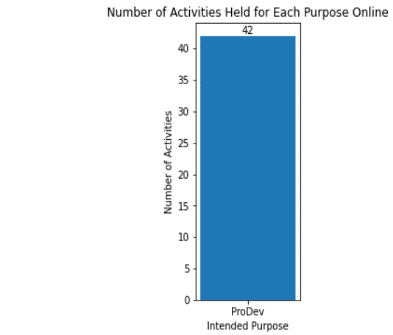
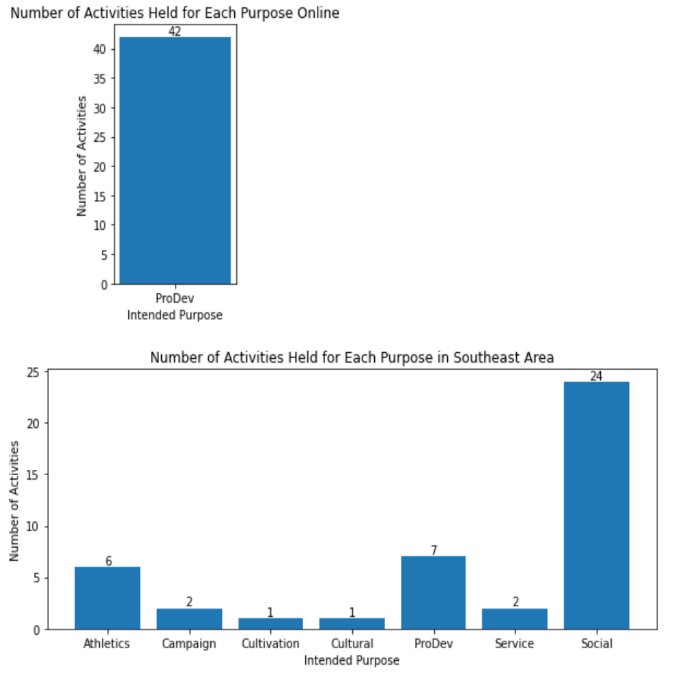
Based on the output of multiple regression and analysis in Part Three and Part Four, we gave our suggestions about how to increase First Time Attendees and Major Prospects to UMD alumni association.

* **Part Six: Other analysis: area comparison**

According to the analysis above, we know that Percentage Major Prospect was significantly influenced by Area. For some area with lower percentage of major prospect, we want to give specific recommendations to it according to its situations. We picked Online, one with relative lower average percentage of major prospect, to analyze. By comparing Online with Southeast Area, which has the highest average percentage of major prospect, we found that Online held activities with only one kind of intended purpose while Southeast Area has various of ones. We think increasing the diversity of activities may be a good method for Online to increase percentage of major prospect.

Chart, bar chart

Description automatically generated



1. **Accomplishment**

* **Identified which factors influence the First Time Attendees and Major Prospects**

Boxplots and k-means clustering gave us the preliminary insight into which factors may have significant effect and which ones may not. For the variables which are difficult to analyze by these methods, such as Activity, we also tried other methods showed above, to explore their influence. Combining with multiple regression analysis, we identified which factors significantly influenced the First Time Attendees and Major Prospects.

* **Gave suggestions to UMD Alumni Association based on our analysis**

Based on our analysis, we gave some suggestions to UMD Alumni Association to help them attract more first time attendees and major gift prospects in the future events.

Since there was a situation where only one person participated the activity and he/she was first time attendee, which made percentage of first time attendees be 100%, we think high percentage of first time attendees does not always mean high number of first time attendees, so does major prospects. So, we gave suggestions according to the different demand of UMD Alumni Association. The suggestions are showed in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Demand | Higher | | | |
| **Percentage First Time Attendees** | **Number of First Time Attendees** | **Percentage Major Prospect** | **Number of Major Prospects** |
| Area | International | / | Southeast, USA general | Northeast |
| Purpose | General | / | Stewardship | Stewardship, Membership |
| Audience | Student, Muslim | Student | Mid Career | / |
| Age | Young | Young | Elder | Elder |
| Quarter | 3 | / | / | / |
| Activity | More about party and gala, less about TPN | Less about TPN and Game Watch | More about meet up and dinner/lunch/breakfast, less about Game Watch | Less about Happy Hour and Game Watch |
| Participated | / | Attract more participants | / | Attract more participants |

1. **How to Test the Project?**

You can test the Python codes in Jupyter Notebook step by step. Then you will see all the work we did.