Business Analytics Consulting for Alumni Association for Higher Attendance

Background

Analyzing data from the UMD Alumni Association and developing meaningful, data-driven insights to increase first-time attendees and major prospects at events. This was done by identifying variables that correlate with higher event attendance of first-time attendees and major gift prospects, understanding what types of events attracted the target groups and optimizing new and existing events to improve engagement of the target groups.

Tools

A variety of Python packages were employed to complete this analysis. They are:

NumPy: array support and manipulation

Pandas: data analysis and manipulation

SciPy: ANOVA test

Matplotlib and Seaborn: statistical plotting and data visualization

Scikit Learn: predictive modeling using linear regression

Tableau: visualization

Data Cleansing

Data cleansing tasks included:

- Merging individual excel worksheets into one dataframe
- Renaming attributes for usability
- Temporal decomposition: Extracting day, month, year, and quarter from event date attribute
- Binning age: To analyze distribution of attendees by age groups
- Missing data: checking for missing data and finding substitutes (no missing data found)

	Event_Name	Activity_Code	Activity_Description	Location_Code	Location_Description	Group_Code	Group_Description	Event_Date	Participated	Average_Age	First_Time_Attendees	Percentage_First_Time_Attendees	Major_Prospects	Percentage_Major_Prospect	age_bins	year	month
0	History of Civil Rights in Baltimore Walking Tour	PEAC4	CP AA-Civil Rights in Bmore Walking Tour	PDBA	CP DMV-Baltimore	PU9	CP Cultural-General	2019-10-13	20	45	1	0.050000	0	0.000000	(39, 49]	2019	10
1	Baltimore in the 21st Century: The Future of t...	PEAB9	CP AA-Baltimore in the 21st Century	PDBA	CP DMV-Baltimore	PC9	CP ProDev-General	2019-10-24	18	41	2	0.111111	1	0.055556	(39, 49]	2019	10
2	Baltimore Mornings with Maryland	PEABM	CP AA-Baltimore Mornings with Maryland	PDBA	CP DMV-Baltimore	PC4	CP ProDev-Mid Career	2019-10-24	22	46	0	0.000000	2	0.090909	(39, 49]	2019	10
3	UMD Does Light City	PEAL1	CP AA-UMD Does Light City	PDBA	CP DMV-Baltimore	PS9	CP Social-General	2019-11-07	100	48	3	0.030000	5	0.050000	(39, 49]	2019	11
4	Baltimore Terps Career Panel	PEABC	CP AA-Baltimore Terps Career Panel	PDBA	CP DMV-Baltimore	PC9	CP ProDev-General	2019-11-12	16	30	1	0.062500	0	0.000000	(29, 39]	2019	11

Preliminary Data Analysis

To understand the data, its structure and what kind of insights could be derived from it, exploratory data analysis was performed to review the following:

- Checking data types of attributes
  All categorical variables were object data types and age bins were category data types. All numerical variables were int64 and there was one datetime attribute that can be decomposed temporally.

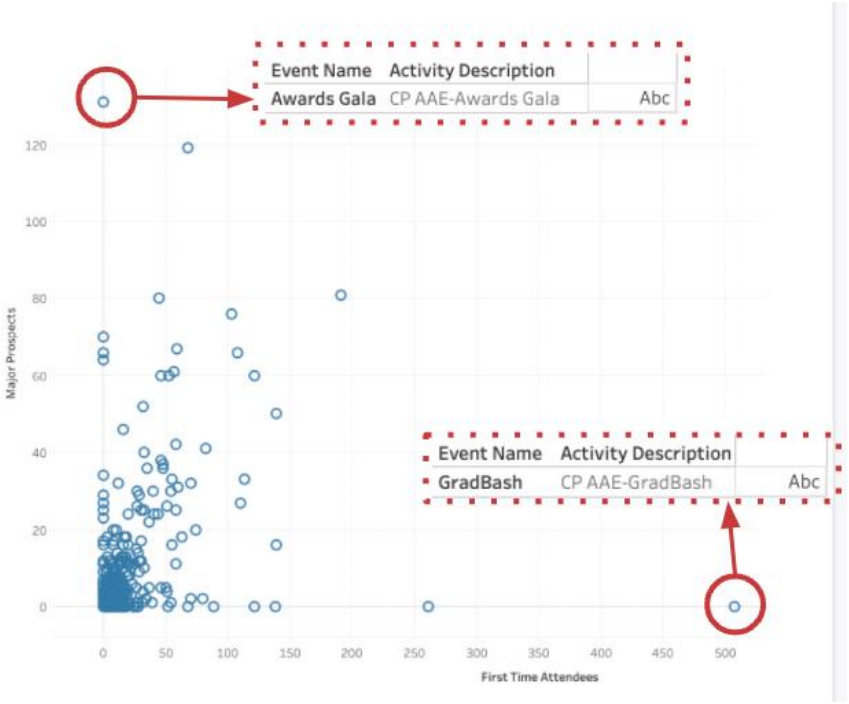
```
Event_Name                          object
Activity_Code                       object
Activity_Description                object
Location_Code                       object
Location_Description                object
Group_Code                          object
Group_Description                   object
Event_Date                  datetime64[ns]
Participated                         int64
Average_Age                          int64
First_Time_Attendees                 int64
Percentage_First_Time_Attendees    float64
Major_Prospects                      int64
Percentage_Major_Prospect          float64
age_bins                          category
year                                 int64
month                                int64
dtype: object
```

- Descriptive statistics and outlier detection
  This showed the percentiles and other summary statistics which helped identify attributes that may possess outliers.

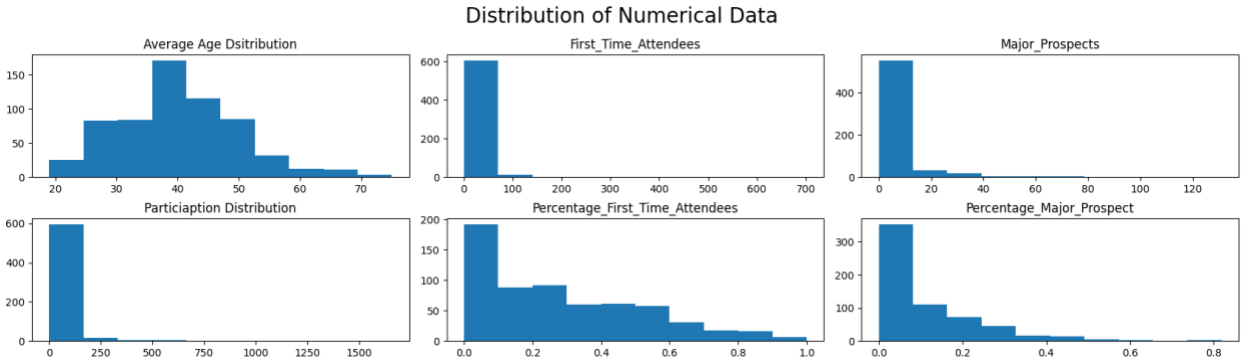
	Participated	Average_Age	First_Time_Attendees	Percentage_First_Time_Attendees	Major_Prospects	Percentage_Major_Prospect	year	month
count	622.000000	622.000000	622.000000	622.000000	622.000000	622.000000	622.000000	622.000000
mean	44.803859	40.117363	13.456592	0.276282	5.966238	0.102214	2017.271704	6.397106
std	93.165049	9.741459	41.103936	0.242273	14.123466	0.131444	1.703970	3.547904
min	1.000000	19.000000	0.000000	0.000000	0.000000	0.000000	2013.000000	1.000000
25%	10.000000	33.000000	1.000000	0.068523	0.000000	0.000000	2016.000000	3.000000
50%	20.000000	40.000000	4.000000	0.237327	1.000000	0.058824	2018.000000	6.000000
75%	44.750000	46.000000	11.000000	0.444444	5.000000	0.166667	2019.000000	10.000000
max	1657.000000	75.000000	702.000000	1.000000	131.000000	0.818182	2019.000000	12.000000

- Cross reference using Tableau
  According to the Tableau, not only it shows the relationship between major prospects and first time attendees, it also shows the specific event under the outlier:
  Awards Gala, which has quite high major prospects even though no first time attendees at all, encouraging us to explore the deep analysis later on. The same for the event:

GradBash, the reason why the quantity for first time attendees is huge but no prospects at all deserves to be analyzed.



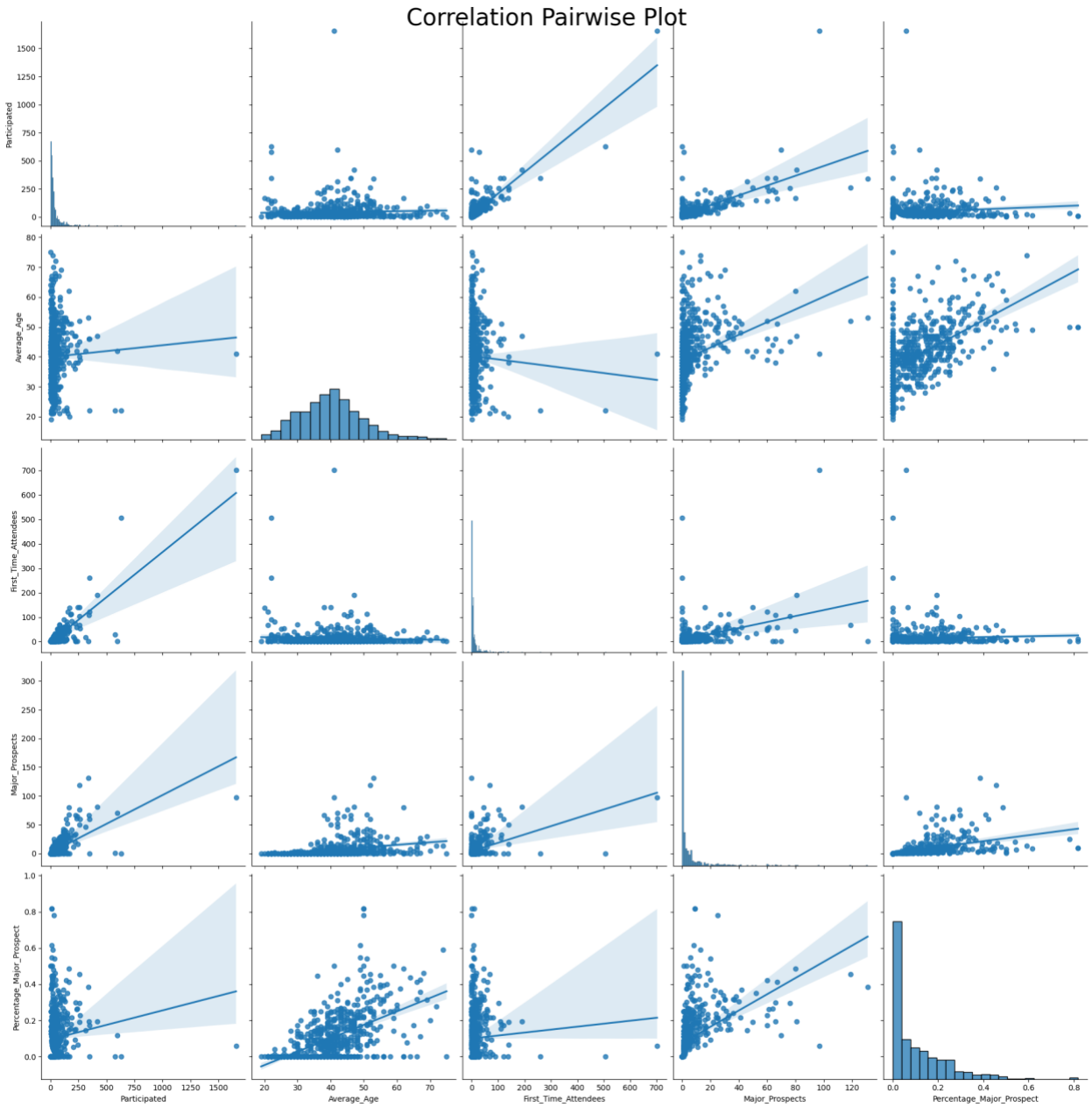
- Checking data distribution



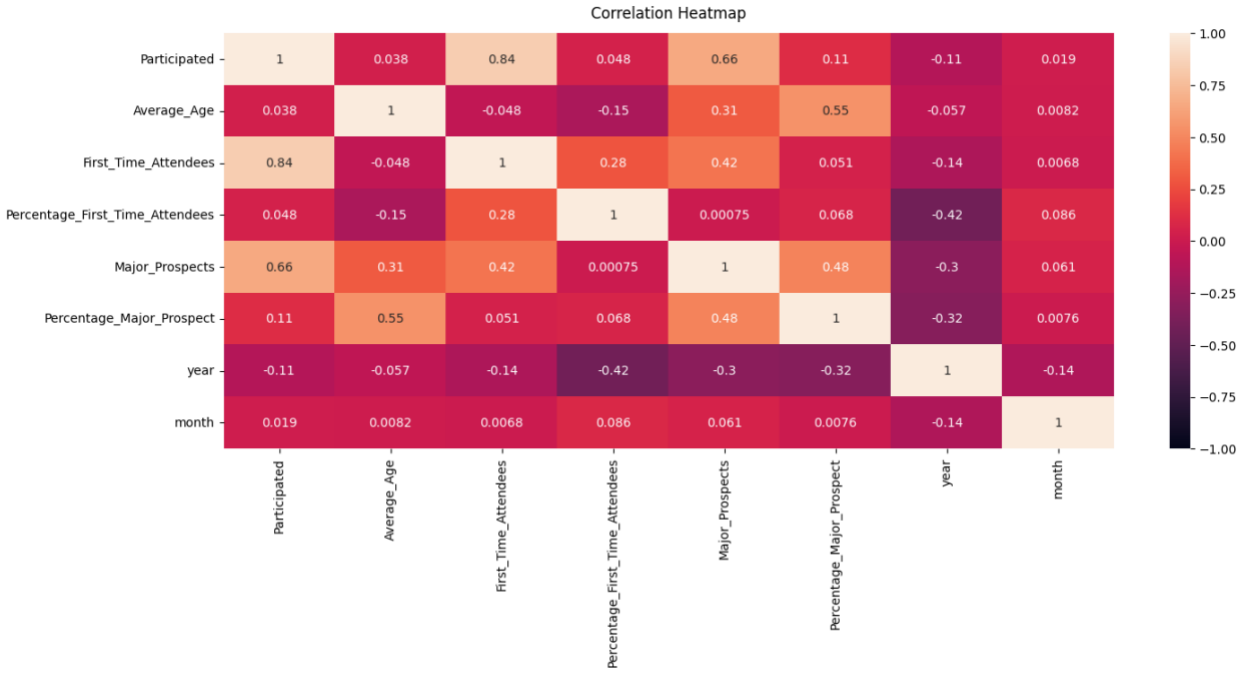
Average age is the only variable with a normal distribution. All others are right-skewed and outliers within them make their shape look so extreme. For further analysis, this data can be normalized by taking log of the variables or exploring removing those outliers.

- Analyzing correlation between variables Pairwise:

This showed the distribution of data via a scatter plot with the confidence intervals of the attribute relationships. This was useful in understanding correlation between attributes and which would be useful to analyse together



The correlation heatmap was an alternate way of visualizing relationships between attributes and viewing the numerical correlation numbers. It was observed that the highest correlation was between participated with first time attendees (0.84) followed by participated with major prospects (0.66). These relationships were further analysed in the predictive modeling stages.



How to read the notebook and document

Please note that beyond this point, the recommendations and insights are organized by attribute (activity, location description, group description, age bin, etc.)

The Python notebook beyond the data cleansing section is organized by the target group (participation, first time attendees and major prospects).

The python notebook was designed for efficiency and was hence grouped by target group. This file is designed to leverage those insights and organize them by attributes so we can compare the groups by attributes and analyse differences.

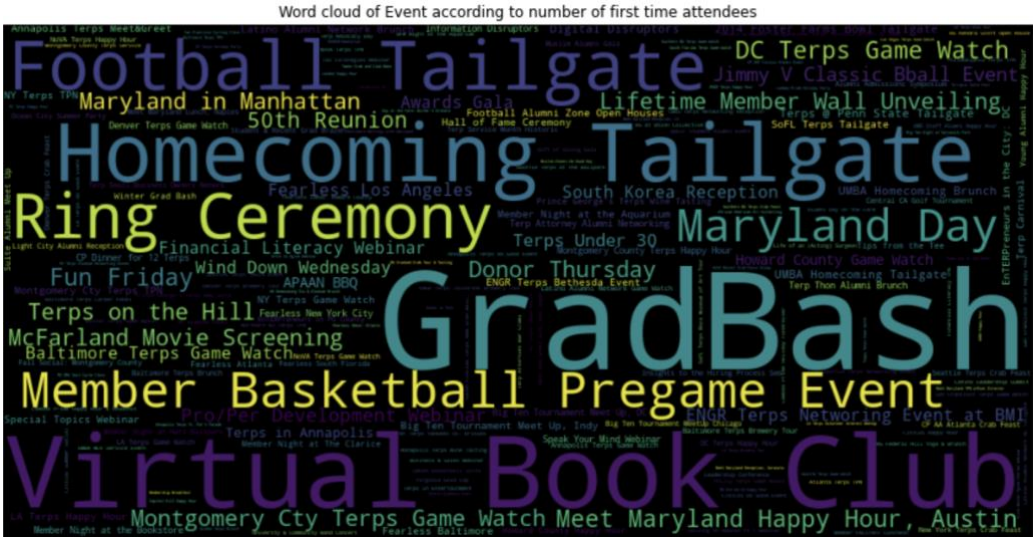
Activity Analysis

By participation

Analysing the event by participation, it was observed that homecoming provided a key opportunity to get alumni involved on campus. Additionally, gradbash and virtual book club also showed promise as being events that could be leveraged to drive alumni engagement. Lastly, Maryland in Manhattan was an example of an event that garnered tremendous attention for out-of-state events. Other out-of-state events should be modeled after this to increase engagement for alumni in other parts of the country.

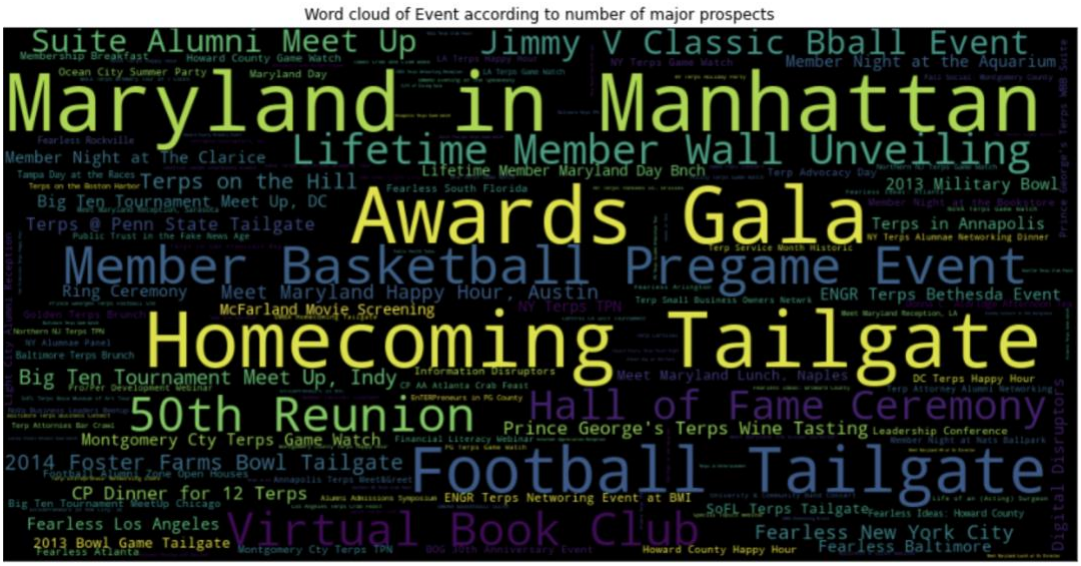
By first time attendees

We see gradbash as a premier event to get first time attendees. This may be because this is one of the first events people attend as alumni. In line with overall participation, we see football tailgating as a good source of first time attendees. Lastly, virtual events like the virtual book club seem to appeal to this group, potentially due to its convenience factor.



By major prospects

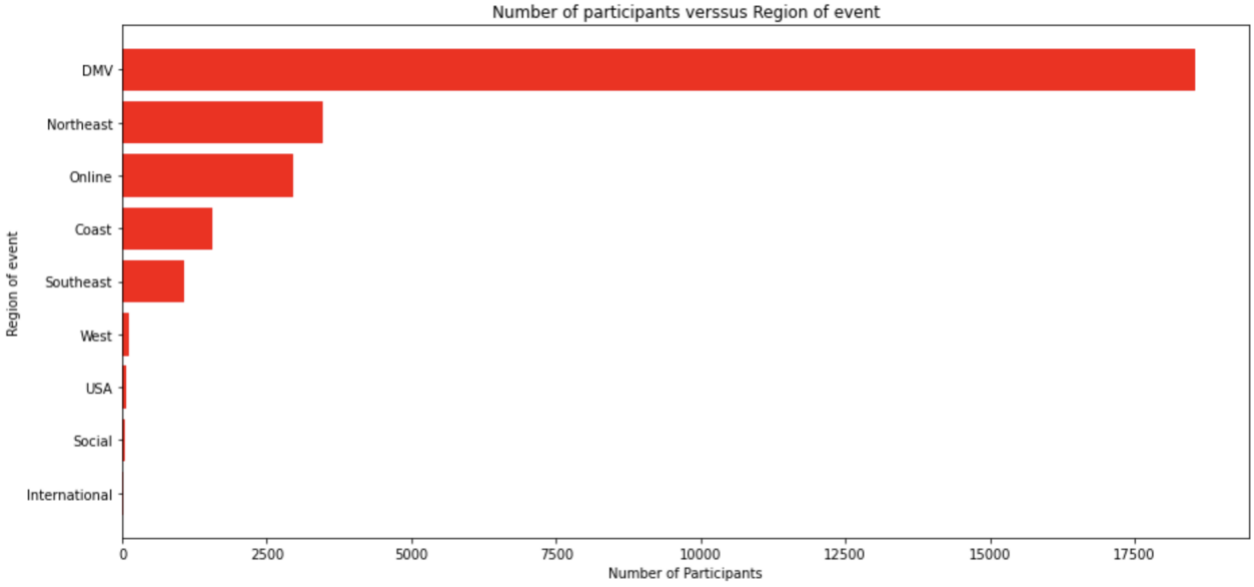
Major prospects tend to attend events that appeal to status, like the awards gala and Maryland in manhattan. These types of ‘status’ events should be marketed to increase engagement for this segment. Lastly, athletic events that have overall strong participation also appeal to this group.



Location Description Analysis- Region

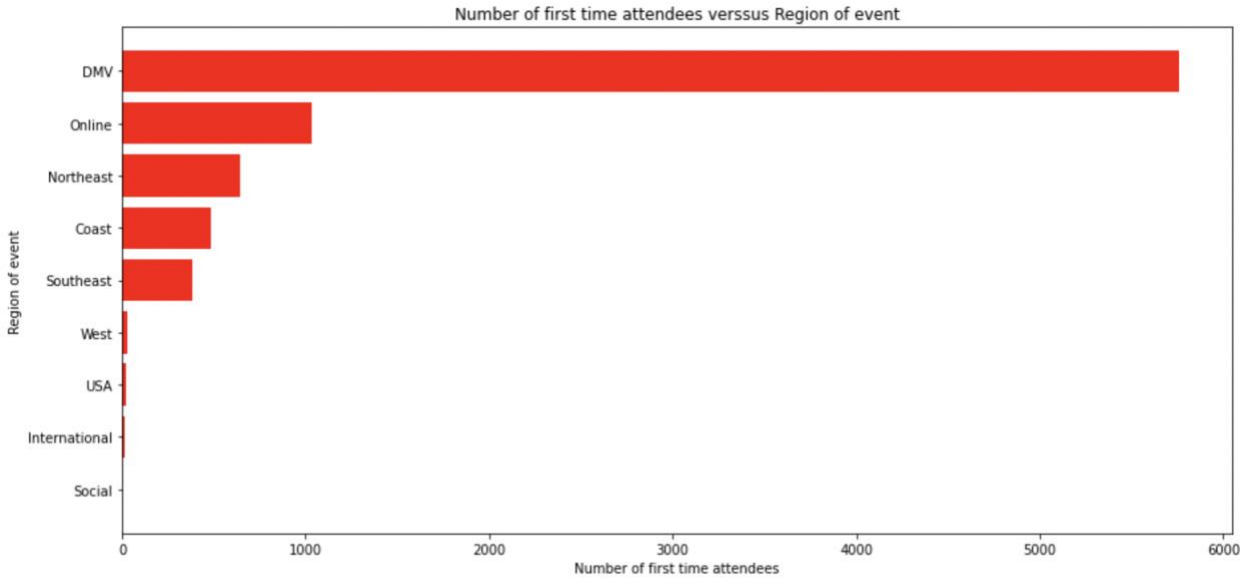
By Participation

It was observed that participants primarily attended events in the DMV area and online. There are continued discrepancies with generalized broad categories like USA, DC and Social that do not align well with location segmentation. It is recommended to reassign these to more specific, existing categories for stronger segmentation.



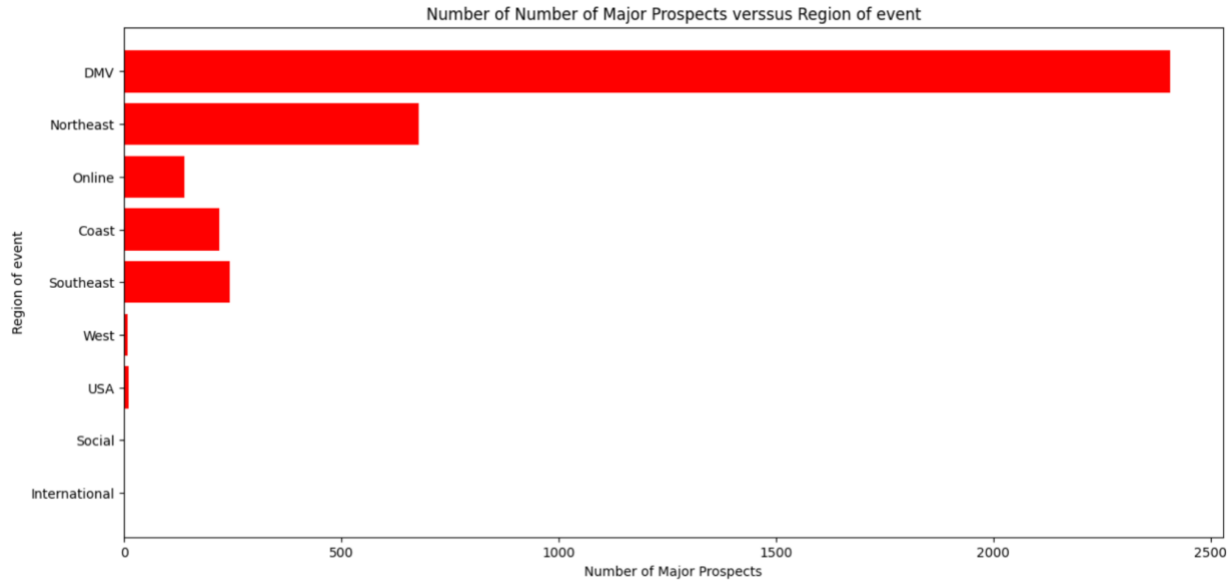
By first time attendees

First time attendees show preference to the DMV area and then online events, relatively similar to overall participation. It is recommended to leverage online events as it has shown high performance in recent years and adds a convenience factor that first time attendees would utilize since they do not need to go out of their way to attend events.



By major prospects

Major prospects seem to favor the Northeast after the DMV area ulike first time attendees who favor online events. The concentration of high income earners in the Northeast make it a hotbed to target for major prospect engagement.

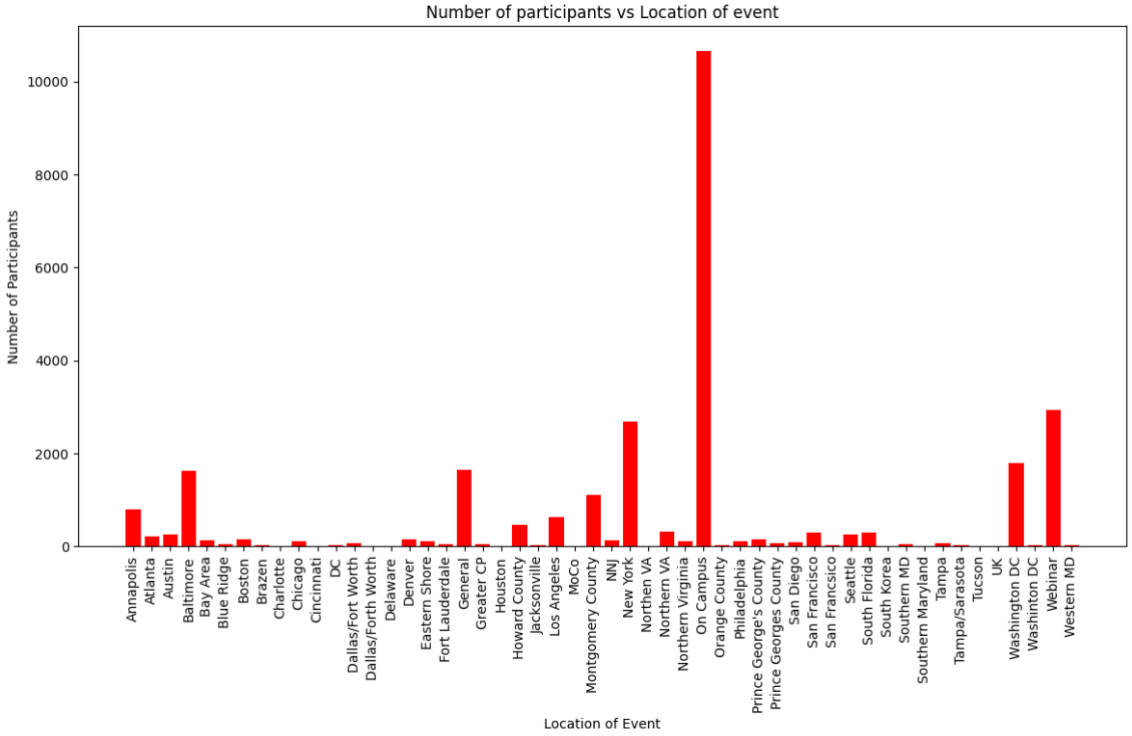


Location Description Analysis- Location

By participation

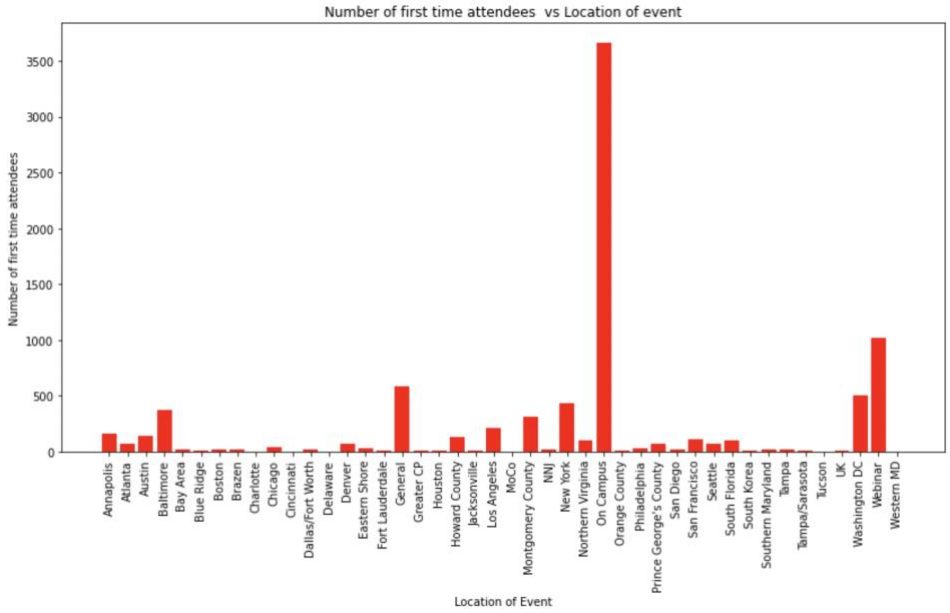
Most participants attended events on campus followed by webinars (online events). There is also relatively high participation in New York. Inconsistencies in this segmentation are also prominent. It is recommended to re-segment by either city or region. Current segmentation includes cities, regions, countries, counties and duplicates due to data entry errors.

The prominence of first time attendees in the DMV area can be attributed to high participation on-campus. This could be treated as an outlier kewing the data.



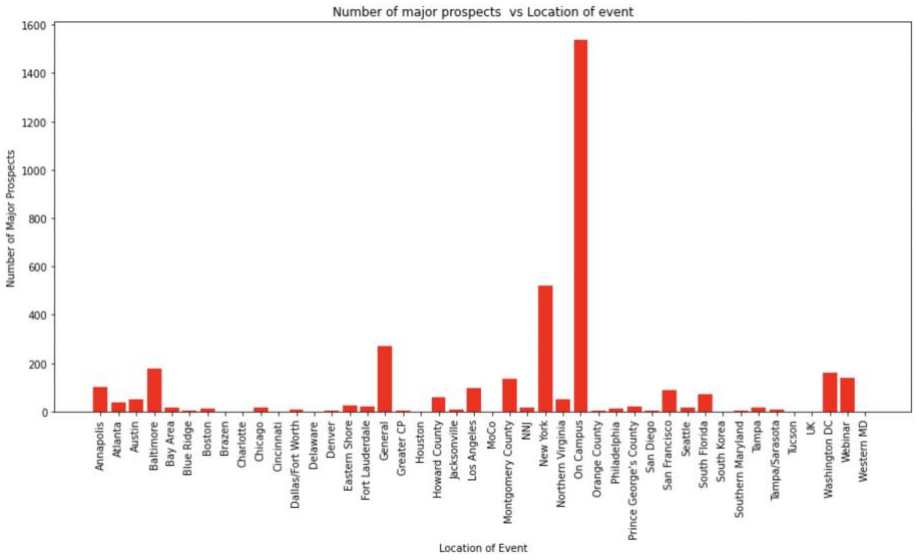
By first time attendees

A similar trend is observed in location preferences for first time attendees with on campus and webinars being the highest. Segmentation in location should be specified better to develop better insights and for future use of the alumni foundation.



By major prospects

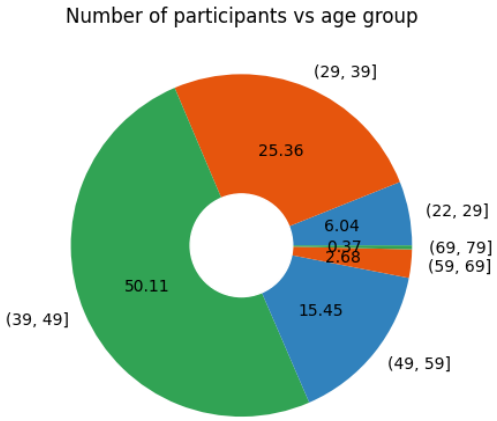
We see a shift in trend with major prospects with most of them preferring New York after on campus. This group does not show any preference for online events.There is a need for stronger segmentation to remove outliers (counties, and other countries)



Age group analysis

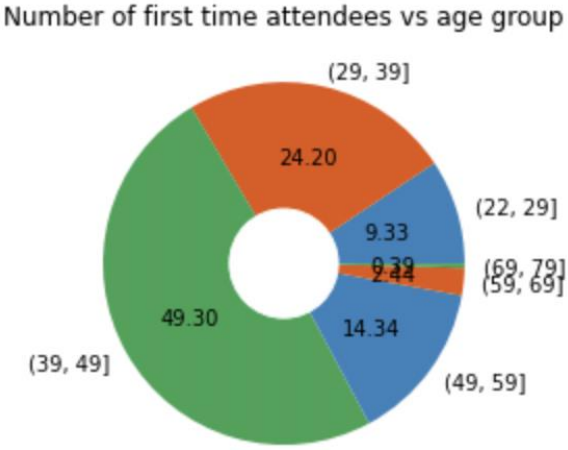
By Participation

50.11% of participants were between 39-49 years of age. The next group was 29-39 year olds. Overall, It seems like participants seem to drop off post-graduation and then get re-engaged with the university closer to mid-career.



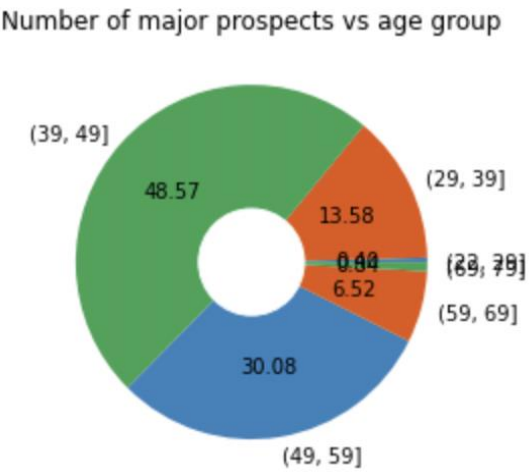
By first time attendees

There is a similar distribution for first time attendees as is for overall participants.



By major prospects

There is a significant increase in the 49-59 age bin (30%) compared to overall participants and first time attendees. This bin is almost double the size of the other groups showing that mid-career alumni are key target groups to be leveraged for major prospect donation.

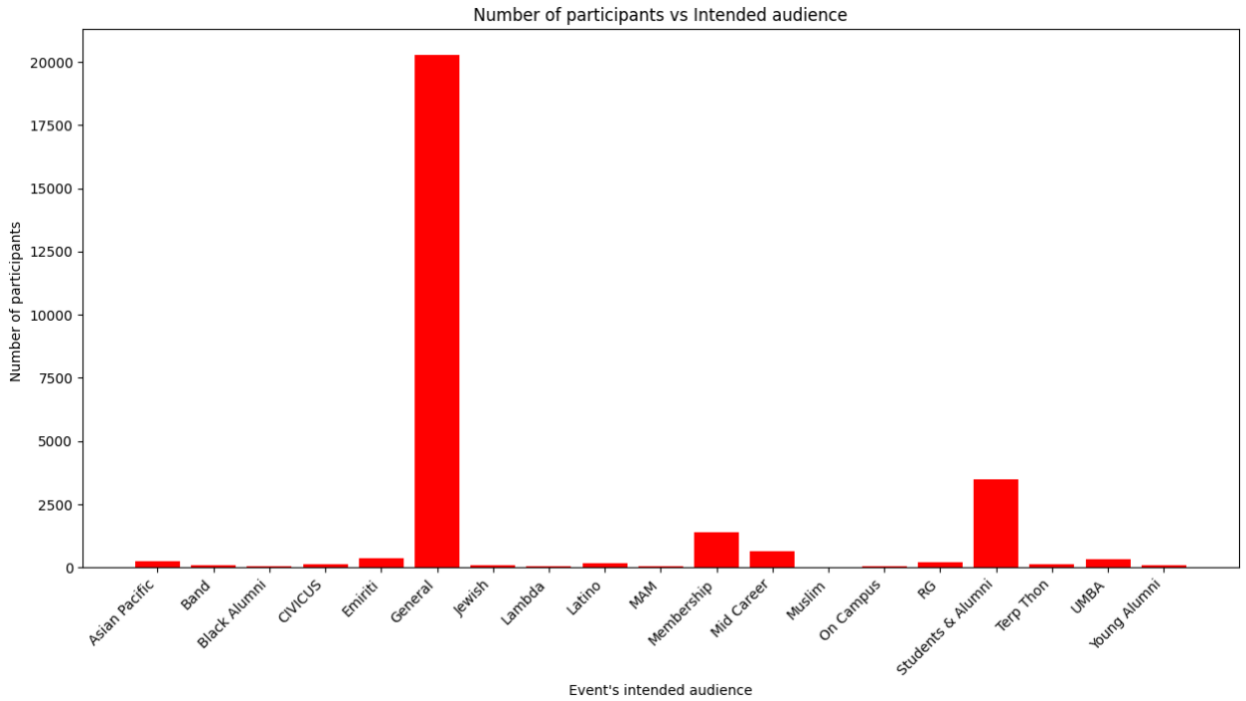


Group Description Analysis- Intended Audience

By participation

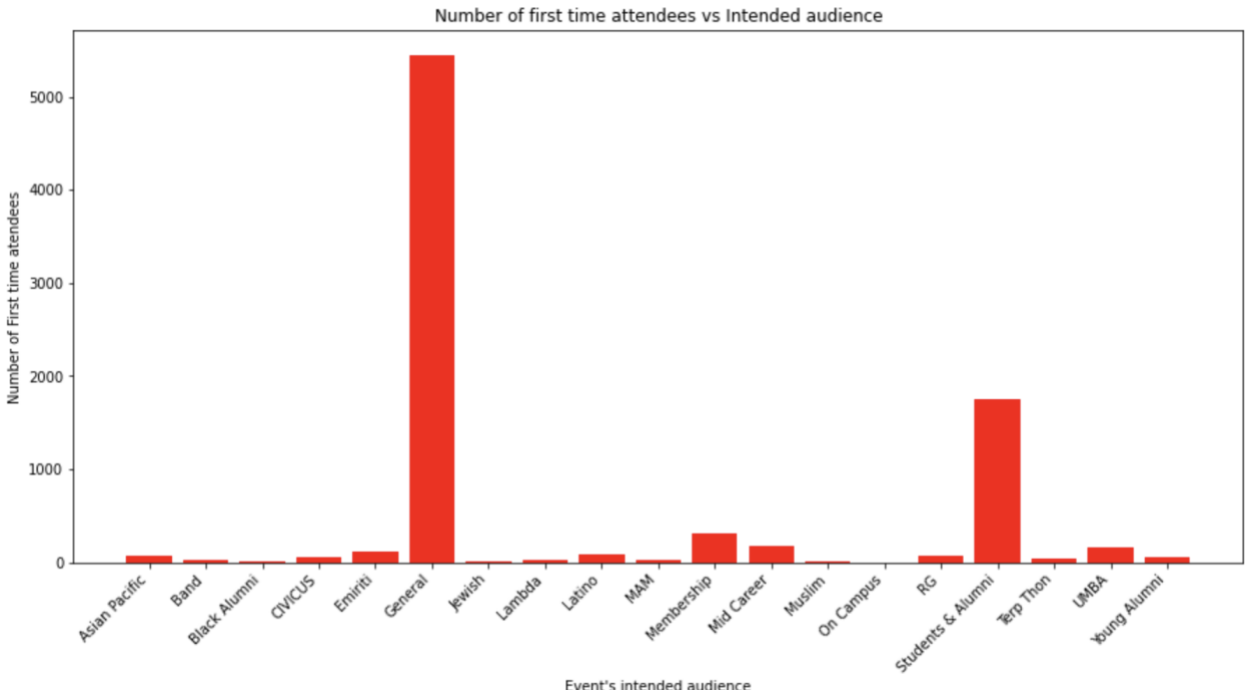
Generalization in the segmentation makes this data hard to interpret. It is recommended to dissect the “General” category into more specific, meaningful segments. Much of the data in this category can be distributed to relevant existing categories to provide a more balanced picture of the intended audience. There are also redundant categories like “students and alumni” that need to be revisited.

Overall, trends even on the first time attendees and segment levels (shown below) do not have distinguishable traits due to the data distribution and skewness due to the general category.



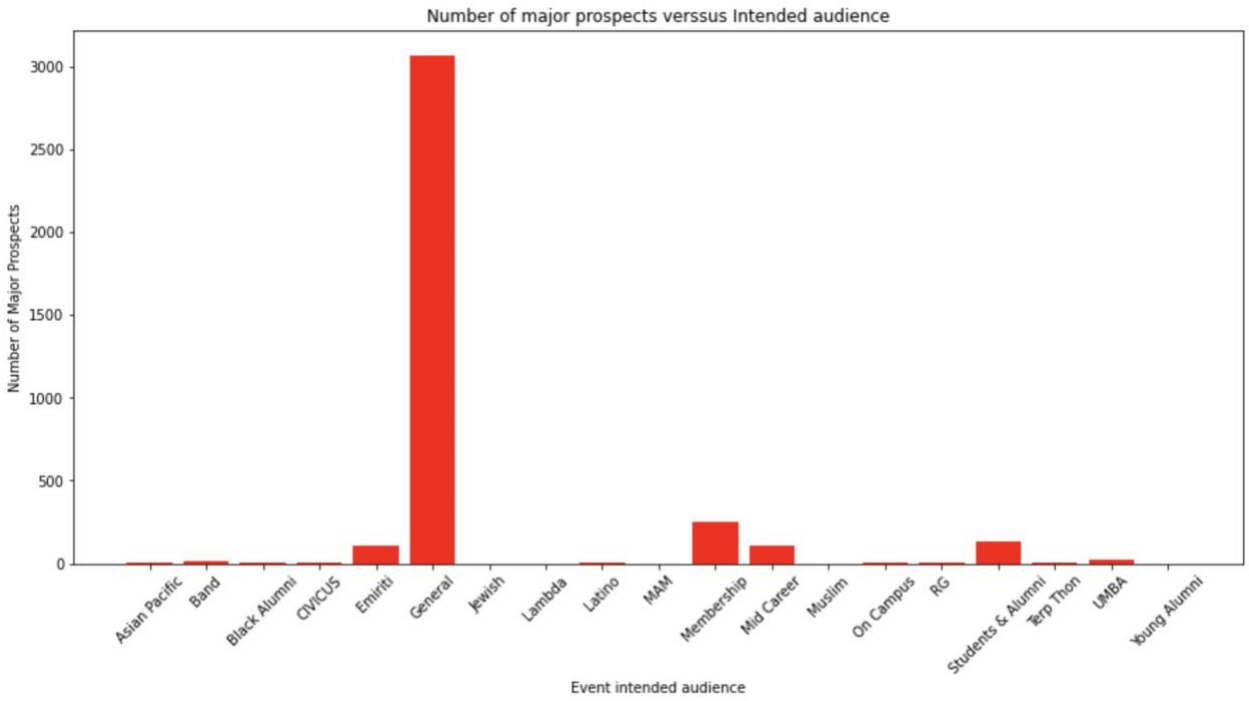
By first time attendees

First time attendees follows an identical trend to overall participation



By major prospects

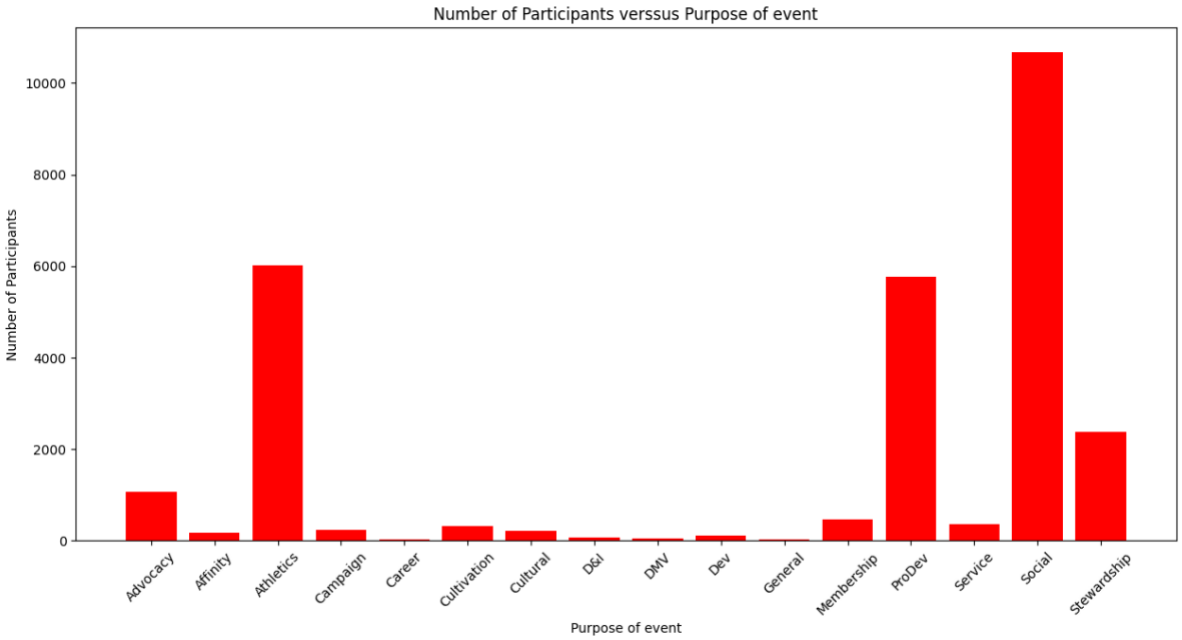
No strong trends are shown for major prospects.



Group Description Analysis- Event Purpose

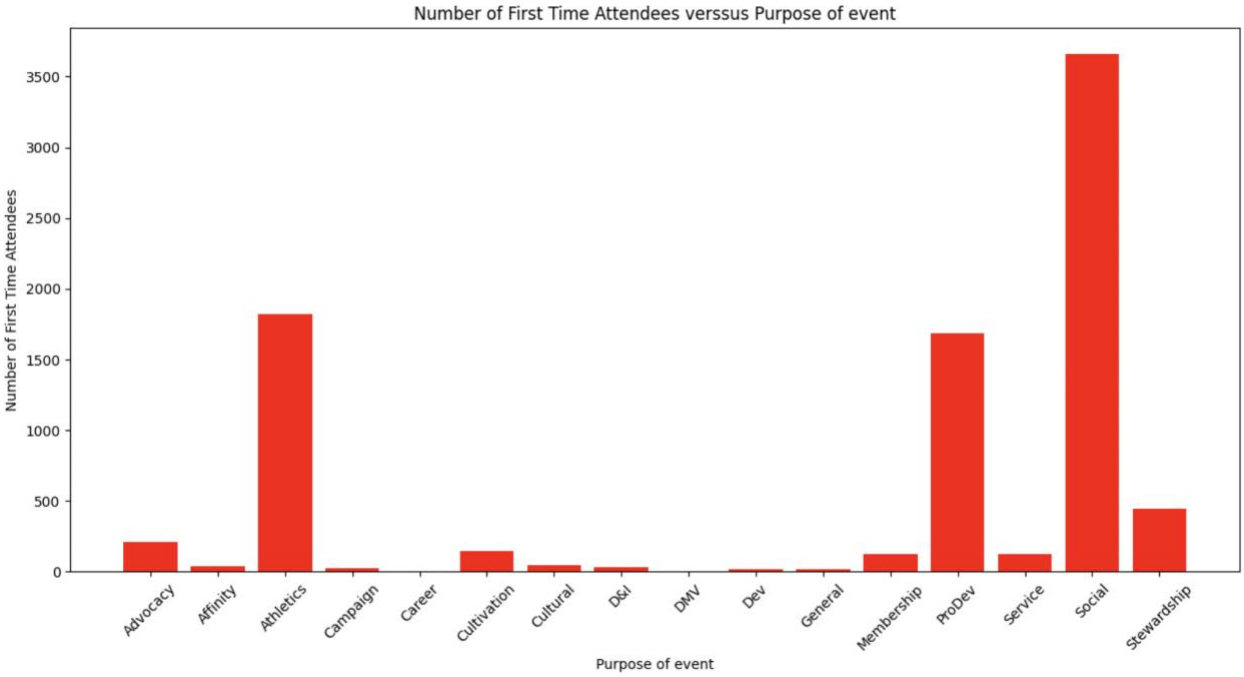
By participation

In terms of intended purpose, most popular events are social and athletic. This category can be leveraged to keep overall participants engaged with the university. The more-specific categories can be used to target specific groups, like stewardship for major prospects and professional development for first time attendees.



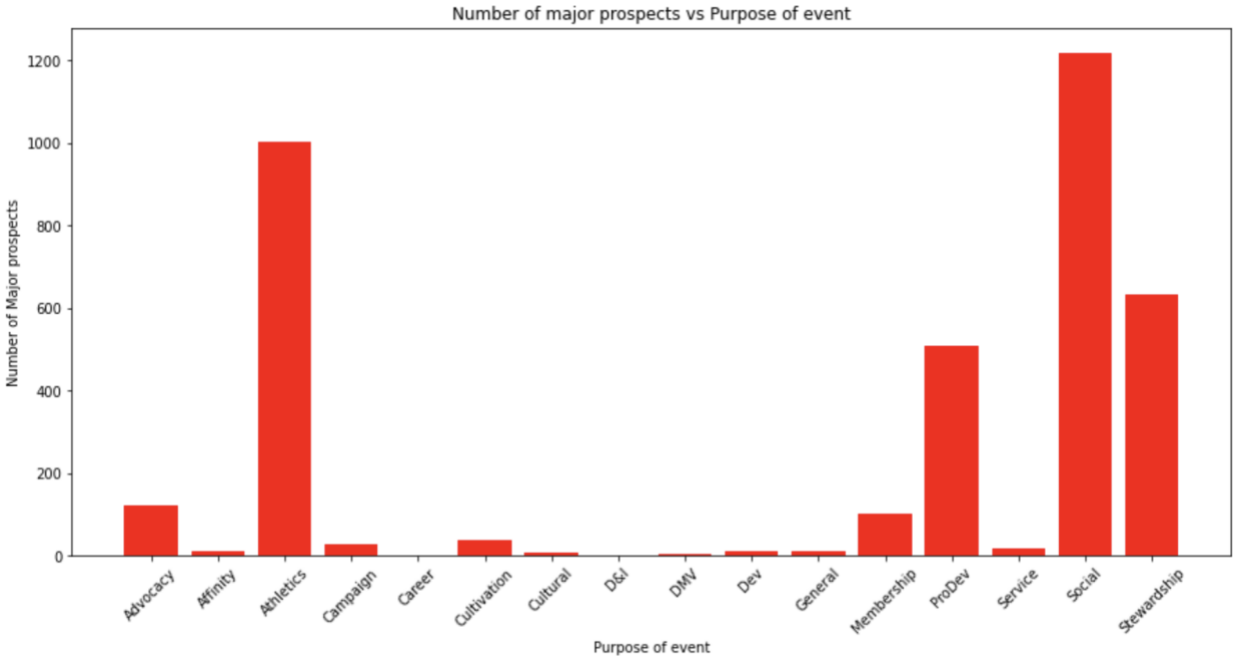
By first time attendees

Other than social and athletics events, there is a strong trend for professional development engagement for first time attendees. It seems like they are eager to network with fellow alumni and reconnect in that aspect.



By major prospects

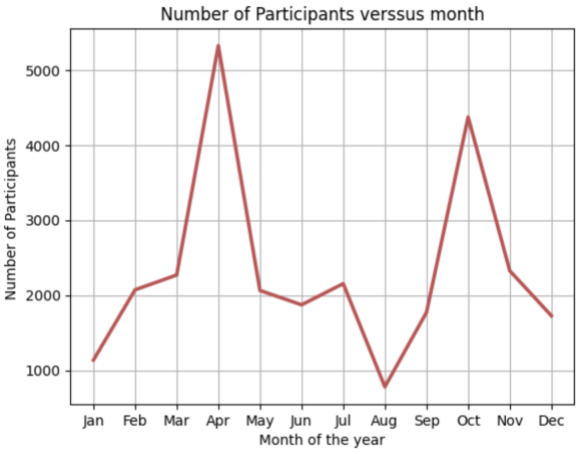
Major prospects show a significant desire for stewardship events. This is different from first time attendees who do not show this desire. It is recommended to focus on stewardship events for major prospects to be engaged and increase the likelihood of donations.



Temporal Analysis- by month

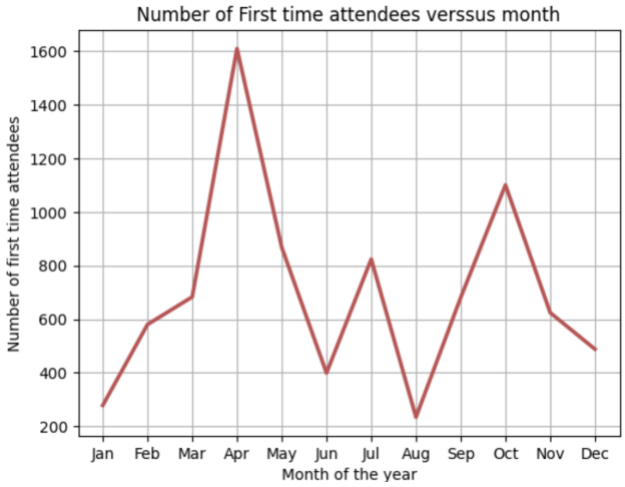
By participation

Overall participation shows peaks in April and October and notably low engagement in August.



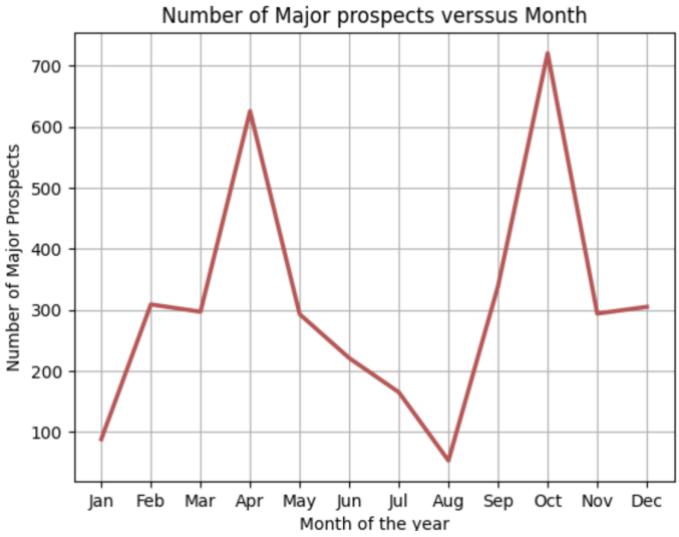
By first time attendees

We see a similar trend in April and October for high involvement of first time attendees.



By major prospects

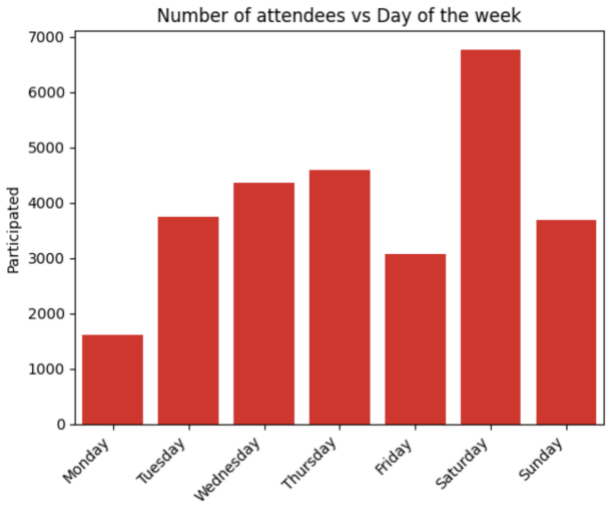
Major prospects show a remarkable trend of engagement in October, surpassing engagement in April, which is different from the other groups. It is recommended to prioritize events towards the end of the year for major prospects



Temporal Analysis- by day of week

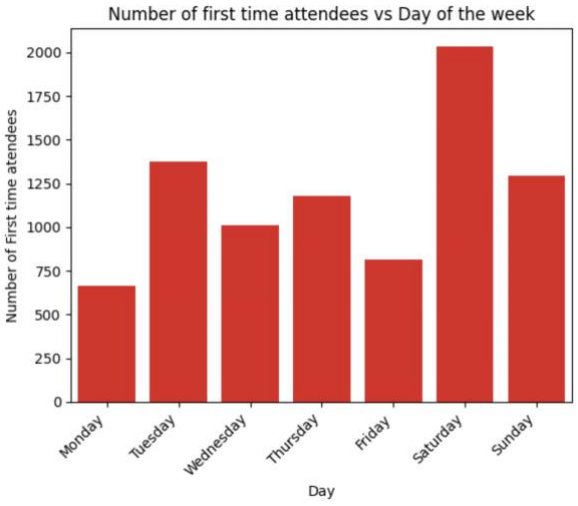
By participation

Overall, participation is strong on weekends, especially saturday. There is also strong engagement on Thursdays as well.



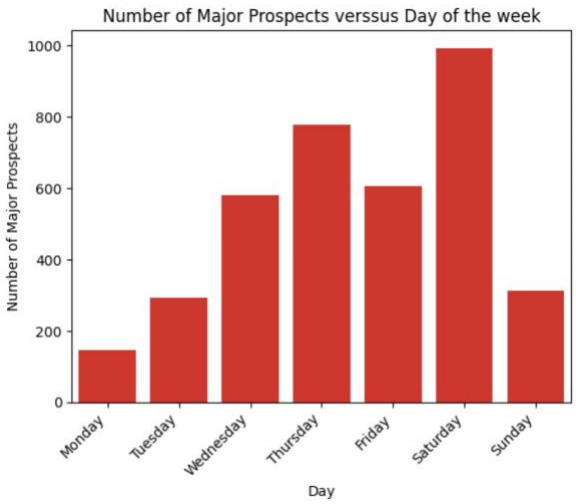
By first time attendees

For this group, Saturdays are a good day for engagements and surprisingly Tuesdays. These trends in attendance clearly indicate this.



By major prospects

Major prospects show similar trends to overall participants.



ANOVA Test

Compare two independent(unrelated) variable groups using F-distribution, to produce P value at the end.

The null hypothesis for the test is that the two means are equal. Therefore, a significant result means that the two means are unequal. Assumption(Ho) is that Major_Prospects and First_time_attendees are not correlated.

In terms of First_time_attendence, P value of quarter is 0.138, higher than 5%, which means lower significant level. Thus, other variables are more relatively deserved to explore and highly correlated with First_time_attendence.

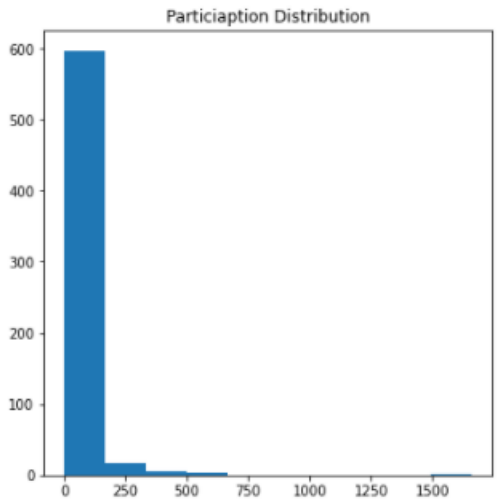
In terms of Major_prospects, P value of quarter is 0.0956, higher than 5%, which means lower significant level. Thus, other variables are more relatively deserved to explore and highly correlated with Major_prospects.

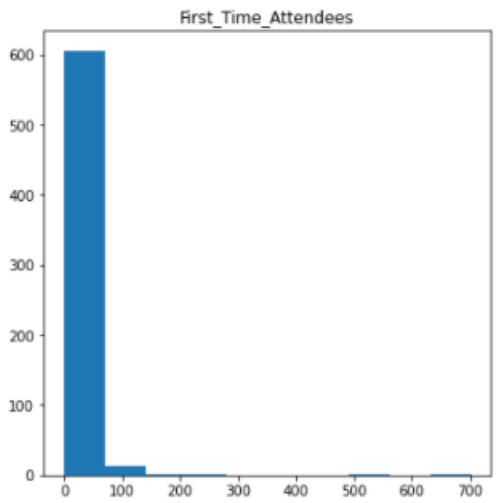
Predictive Modeling

First Time Attendees

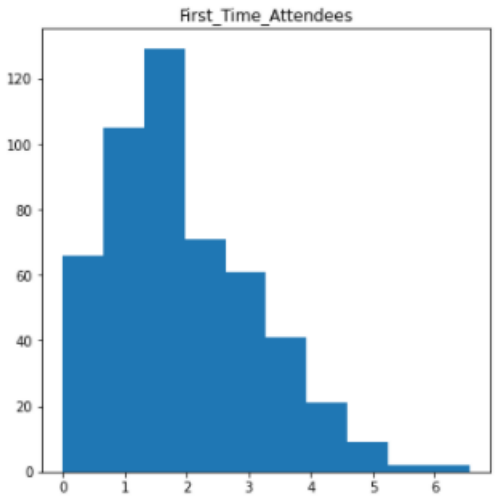
Three regression models were developed predicting first time attendee participation. The attribute with the strongest relationship (correlation) to first time attendees was participation. This was used as the independent variables to build a simple linear model that served as the baseline model. Data from both first time attendees and participation were normalized to create normal distribution.

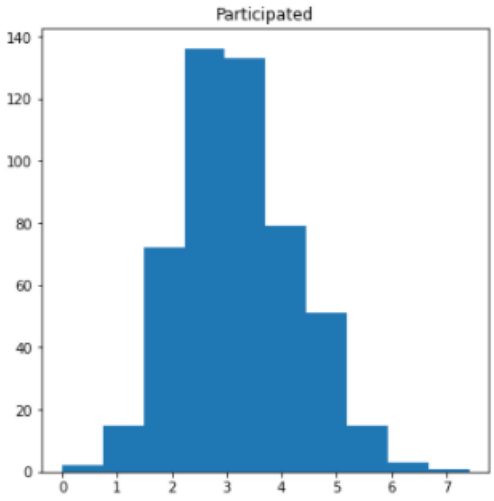
Before normalization:



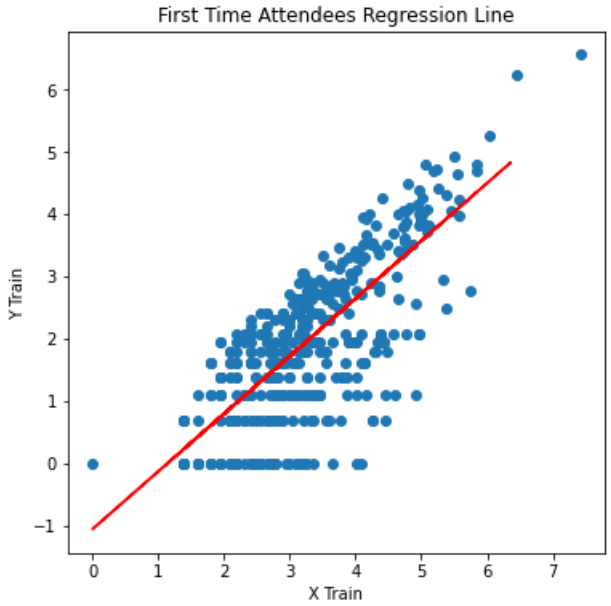


After normalization:





The data was then split into training (80%) and test (20%) sets. The R-square was 62.1% . MSE was 0.618 and RMSE was 0.786. The regression line was plotted. There seems to be outliers in the data that may require additional treatment to improve the fit of the model.



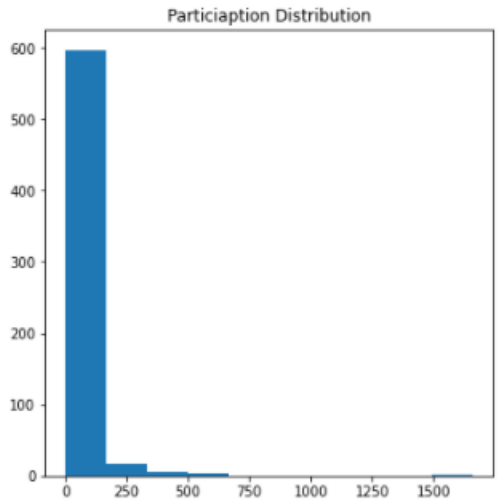
The next linear model included major prospects. This attribute had the next strongest relationship with first time attendees. Major prospects also had to be normalized. The R-square of this model was lower at 55.08% . The MSE and RMSE values also increased to 0.691 and0.831 respectively. This showed that major prospects may not be a reliable variable to help predict first time attendees.

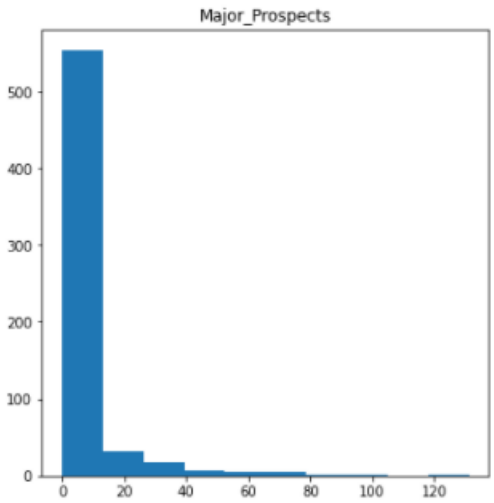
The final model included categorical variables- activity code and location code. These variables were encoded to dummy variables. The final model R-square was 54% with significantly higher MAE and RMSE values than the previous models. This model is likely suffering from overfitting. Outlier detection and deeper analysis of categorical variables may help in identifying data that should be dropped to improve the predictive ability.

Major Prospects

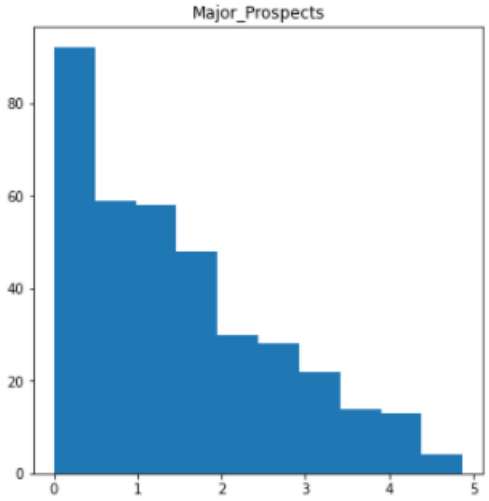
A simple linear model was built as a baseline to evaluate predicting major prospects. Participation was used as the independent variable since it had the strongest relationship with major prospects. Both participation and major prospects had to be normalized.

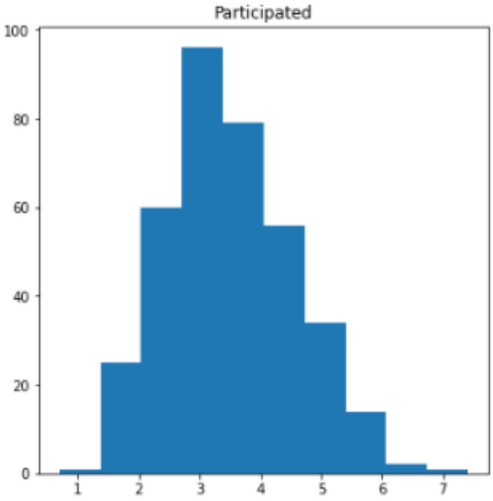
Before normalization:



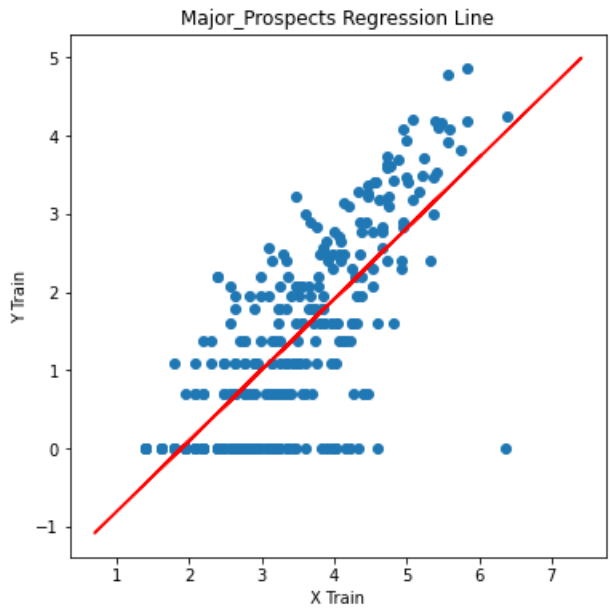


After normalization:





The model yielded a R-square of 58.9% with MSE and RMSE values of 0.624 and 0.79 respectively. The model plot showed several outliers that reduced the performance of the model.



The second model added the next independent variable with strongest relationship to major prospects. This was first time attendees. The model results improved slightly from the baseline with a R-square of 61.6% and MSE and RMSE values of 0.559 and 0.748. The increased R-square and lower MSE and RMSE values showed that the model performance got better when adding first time attendees to the equation.

The final model included categorical variables for location code and activity code. The R square reduced to 25.2% with significantly higher MAE and RMSE values. The categorical variables did not help model performance and reduced the prediction accuracy. It may be beneficial to reduce the number of categorical variables and see the changes to model performance. Determining which categorical variables to drop will need additional analysis.

Models for both first time attendees and major prospects did not provide very strong accuracy or model performance. Since the regression modeling was used in an investigative approach to see how first time attendees and major prospects could be predicted, the scores provided some useful insight. For first time attendees, only participation provided a strong enough relationship to provide any prediction. On the other hand, for major prospects, both participation and first time attendees provided reasonable model performance. Lastly, categorical variables did not add to model performance.

If the predictive modeling was taken as a primary approach to answering the mission objective, more could be done to build a robust model. Outlier within participation, major prospects and first time attendees need to be evaluated. Removing them may likely improve model performance. Categorical variables could be added to the model one at a time to see which one helps improve model performance. Lastly, some of the categories should be removed after conducting a separate analysis of the categories.

Future Work

There are many opportunities to further the analysis of this dataset. More detailed analysis of activity code, location code and group code could help define some trends in event participation. Using the analysis of categorical variables can help better define the independent variables for predictive modeling that was explored in the analysis. Doing time series forecasting may be useful to anticipate growth in attendance based on events and current trends and may be an alternate way to predict first time attendance and major prospects to the linear regression model.

To better define our target groups, clustering is another option that should be explored. This could allow for better segmenting first time attendees and the participants in general. Lastly, classification modeling using logistic regression may be a possibility to see likelihood of major prospect giving a donation. In the initial analysis, it seemed that the current data may not be able to support a reliable classification model but it is still worth exploring.

Overall, the description data needs to be evaluated and made to be more specific for further analysis. Broad categories like “General” and locations like “USA” and using counties as descriptions make the data distribution skewed and hard to interpret. Thorough data cleansing will allow for better insights.

References

The team referenced documentation for matplotlib, pandas, seaborn and several other packages. Additionally, external resources were used to troubleshoot code or for enhancement purposes throughout the project. Links to various references are provided in the python notebook.

Concatenating sheet into single dataframe:

https://stackoverflow.com/questions/25486438/how-to-dynamically-refer-to-dataframes-in-a-for-loop-in-python

Renaming attribute:

https://www.listendata.com/2020/09/How-to-rename-columns-in-Pandas.html

Adding quarter attribute:

https://stackoverflow.com/questions/1406131/is-there-a-python-function-to-determine-which-quarter-of-the-year-a-date-is-in

Extracting month and year:

https://stackoverflow.com/questions/25146121/extracting-just-month-and-year-separately-from-pandas-datetime-column

Adding age bin:

https://www.geeksforgeeks.org/pandas-cut-method-in-python/

Treating inconsistency in Activity description:

https://www.geeksforgeeks.org/python-string-strip-2/

Word cloud:

https://amueller.github.io/word_cloud/

Treating inconsistency:

https://www.journaldev.com/37898/python-numpy-where

Plotting number of Participants each month:

https://www.geeksforgeeks.org/seaborn-lineplot-method-in-python/

Extracting day of the event Date:

https://www.geeksforgeeks.org/python-pandas-to_datetime/

Number of first time attendees:

https://medium.com/@kvnamipara/a-better-visualisation-of-pie-charts-by-matplotlib-935b7667d77f

Word cloud of Event according to number of major prospects:

https://stackoverflow.com/questions/43145199/create-wordcloud-from-dictionary-values

Infinite values:

https://stackoverflow.com/questions/17477979/dropping-infinite-values-from-dataframes-in-pandas

R-Square for training set:

https://www.datacamp.com/community/tutorials/essentials-linear-regression-python

One hot encoder:

https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd