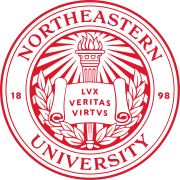
****

**XN Project Draft**

**Event Analytics and Forecasting**

**Background**

The "Event" Event & Location Analytics project is a comprehensive initiative aimed at enhancing event management and planning through the integration of geolocation data and predictive analytics. The project focuses on Toronto's event landscape, leveraging data sourced from Eventbrite and Meetup to analyze event characteristics and attendee behavior. By employing tools like BeautifulSoup for data extraction and advanced analytics techniques, the project delivers insights into attendee preferences and event patterns, offering solutions that improve overall efficiency and participant experience.

**Statement of Purpose**

The project aims to transform the event management process by offering data-driven insights and solutions to common challenges. By employing predictive modeling, the project aspires to forecast event attendance accurately, thereby enabling organizers to make informed decisions. Furthermore, it seeks to integrate real-time data visualization tools, such as heatmaps and clustering, to enhance the effectiveness of event management strategies.

**Project Scope**

The scope of the project is extensive, beginning with the collection of event data from Eventbrite and Luma, specifically focusing on events within Toronto. The data undergoes rigorous cleaning and preparation, which involves removing duplicates, standardizing formats, and engineering new features such as event duration categories and geolocation attributes. The project encompasses a detailed analysis to uncover trends in event timing, locations, attendee counts, and preferences. These insights are presented through geolocation-based heatmaps, enabling stakeholders to visualize event density and plan accordingly. Additionally, actionable recommendations are provided to optimize event scheduling, improve marketing strategies, and manage crowds effectively. Future research possibilities include incorporating external factors like weather conditions, social media trends, and exploring enhanced predictive models to further refine the project outcomes.

**Literature Review**

Trinh, T., & Vuongthi, N. (2022). A predictive paradigm for event popularity in event-based social networks. *IEEE Access*, *10*, 125421–125434. <https://doi.org/10.1109/access.2022.3225734>

The paper by Trinh and VuongThi (2022) proposes a predictive paradigm for estimating event popularity within Event-Based Social Networks (EBSNs), such as Meetup. The study focuses on predicting participant numbers as a metric for event success by leveraging regression methods like Random Forest, Support Vector Machine, and Decision Tree. The authors develop a feature generation framework based on venue, time, and content factors. Experiments conducted on datasets from Sydney, London, and San Francisco demonstrate that Random Forest achieves the best prediction accuracy. Interestingly, the study identifies content as the most significant predictor in city-wide contexts, while time becomes crucial within specific group contexts. The framework is also suggested as a potential online tool for event organizers to plan effectively and reduce costs.

This research aligns closely with the goals of the Event project by highlighting how diverse factors influence event attendance, a crucial element for accurate forecasting. The predictive paradigm offers valuable insights into feature engineering techniques, such as leveraging venue proximity and content relevance, which could be integrated into Event’s forecasting model. Additionally, the study's findings about the varying importance of time and content factors in different contexts provide a nuanced understanding that can enhance the heatmap and predictive functionalities of the Event system. By incorporating these strategies, the Event project could improve its ability to forecast attendance and offer actionable insights to event organizers.

Zhang, X., Zhao, J., & Cao, G. (2015). Who will attend? -- predicting event attendance in event-based Social Network. *2015 16th IEEE International Conference on Mobile Data Management*, 74–83. <https://doi.org/10.1109/mdm.2015.23>

Zhang, Zhao, and Cao (2015) focus on predicting event attendance within Event-Based Social Networks (EBSNs) by analyzing user behavior patterns. The authors propose a model using semantic, temporal, and spatial features to forecast attendance at future events. Semantic features were found to be the most critical, capturing user preferences based on past event participation and topic relevance. Temporal features utilized weekly and daily attendance patterns, while spatial features analyzed user location preferences. The study applied supervised learning methods, including Logistic Regression, J48 Decision Tree, and Naïve Bayes, to evaluate the features. Results revealed that combining all feature sets significantly enhanced predictive accuracy, with semantic features demonstrating the highest impact on event attendance predictions.

This research is highly relevant to the Event project, as it highlights the importance of feature engineering in attendance forecasting. Specifically, the emphasis on semantic features aligns with Event’s goal to analyze user activity data for accurate predictions. The integration of temporal and spatial features offers a broader perspective for refining Event’s models. Additionally, leveraging techniques from this study, such as the J48 Decision Tree, could improve the Event system's capability to forecast attendance and visualize geolocation-based insights through real-time heatmaps. The findings provide actionable strategies for enhancing the project’s predictive and analytical frameworks.

Zhou, D., Li, H., Liu, S., Song, B., & Hu, T. (2017). A map-based visual analysis method for patterns discovery of mobile learning in education with Big Data. *2017 IEEE International Conference on Big Data (Big Data)*, 3482–3491. <https://doi.org/10.1109/bigdata.2017.8258337>

The study introduces a map-based visual analysis method for identifying patterns in mobile learning using educational big data. The authors detail a layered visualization approach, enabling the exploration of geo-tagged data to identify both individual and group learning behaviors. Key techniques include hierarchical information tagging and spatial analytics for visualizing time-series patterns, such as learning activities over days and weeks. The approach leverages map overlays and topological operations to provide meaningful insights into the spatial and temporal dimensions of learning. This method demonstrates the capability to process large-scale educational data while offering practical tools for improving management and decision-making in learning environments.

This work is relevant to the Event project because it highlights the utility of map-based visualizations for uncovering spatial patterns in user behavior, which is integral to forecasting event attendance and visualizing real-time crowd dynamics. The study's emphasis on layered visualization techniques and geo-tagged data aligns with Event's goal of creating dynamic heatmaps and analyzing mobility patterns. Incorporating similar methods could enhance Event’s ability to segment and predict crowd activities more effectively, offering actionable insights to event organizers for planning and safety optimization.

**Design and Data Collection**

The project’s data collection relied on Eventbrite and Meetup as the primary sources, focusing on events in Toronto. Data was extracted using BeautifulSoup, resulting in a dataset of 1,000 records across nine variables, including event name, geolocation, descriptions, and timing details. Rigorous data preparation was a critical aspect of the project. Duplicates were removed, datetime formats were standardized, and missing information, such as geolocation data, was handled by dropping incomplete records.

Feature engineering played a vital role, with additional attributes such as event duration categories and geolocation handling being derived to enhance the analytical capabilities of the dataset. This prepared data formed the foundation for subsequent analyses and modeling. For attendance forecasting, Meetup data is used which has ~1200 records and 11 variables. The advantage of using Meetup data is that it includes the number of attendees for each event which is used to predict the number of attendees.

**Implementation, Methodology, and Strategies**

The implementation phase combined geospatial analysis, temporal analysis, predictive modeling, and visualization techniques to address the project’s objectives. Geospatial analysis focused on identifying event density through heatmaps, revealing that Downtown Toronto is a hub for events. Temporal analysis uncovered trends such as peak activity occurring at 7:00 PM and Saturdays being the most popular days for events.

Predictive modeling was employed to forecast attendance based on historical trends and geolocation data, utilizing techniques like regression analysis and clustering to identify significant patterns. Visualization tools such as Tableau were used to create interactive dashboards, allowing users to explore event trends and geolocation-based insights dynamically. Together, these methods provided a comprehensive understanding of event characteristics and attendee behavior.

**Exploratory Data Analysis (EDA)**

**Event Duration Categories**

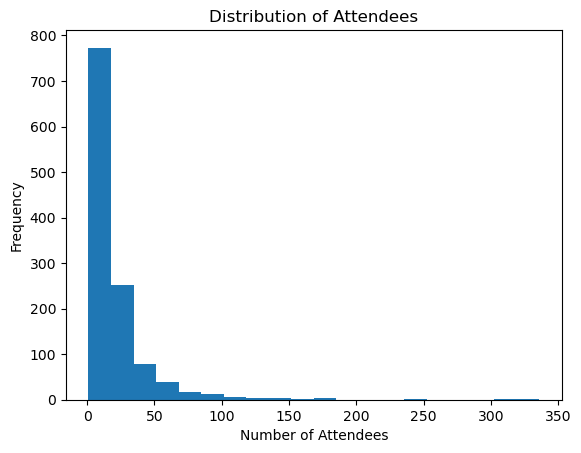
The analysis of event durations shows a significant concentration of events lasting only a day, as highlighted in the bar chart. A smaller proportion of events span 2–7 days, and virtually no events exceed a week. This distribution indicates a preference for shorter-duration events, possibly due to logistical simplicity and higher attendee availability for one-day commitments.

A graph with a bar and a bar

Description automatically generated with medium confidence

**Distribution of Attendees**

The histogram for attendee distribution reveals a highly skewed pattern, with the majority of events having fewer than 50 attendees. There are very few events with larger attendance, indicating that most events are designed for smaller, focused audiences. This could reflect targeted themes or limited venue capacity, catering to niche interests or exclusive groups.



**Frequency of Each Event Type**

The bar chart comparing event types indicates that in-person events dominate the dataset, followed closely by online events. Hybrid events, which combine in-person and virtual elements, represent a negligible proportion. The trend suggests a strong inclination toward traditional in-person gatherings, with online events likely being a secondary option for accessibility or pandemic-related adjustments.

A graph of a bar chart

Description automatically generated with medium confidence

**Events by Day of the Week**

Saturday emerges as the most popular day for hosting events, with significantly higher occurrences compared to other days. Thursday and Wednesday also see a moderate number of events, suggesting a mid-week preference for certain activities. Conversely, Monday and Friday are the least popular, reflecting typical workweek schedules and possible travel constraints.

A graph of events by day

Description automatically generated

**Number of Events Over Time (Daily)**

The line chart detailing events over time highlights fluctuations in the number of events held daily. Peaks correspond to likely scheduled clusters, while dips indicate quieter periods. These trends may align with holidays, weekends, or specific industry schedules, making them essential for planning high-impact activities.

A graph of events over time

Description automatically generated

**Number of Events Over Time (Hourly)**

An hourly analysis reveals that early morning, particularly around 8:00 AM, sees the highest event activity, with smaller peaks during other hours. This trend aligns with the typical workday start or conference schedules, which often begin early to maximize participation.

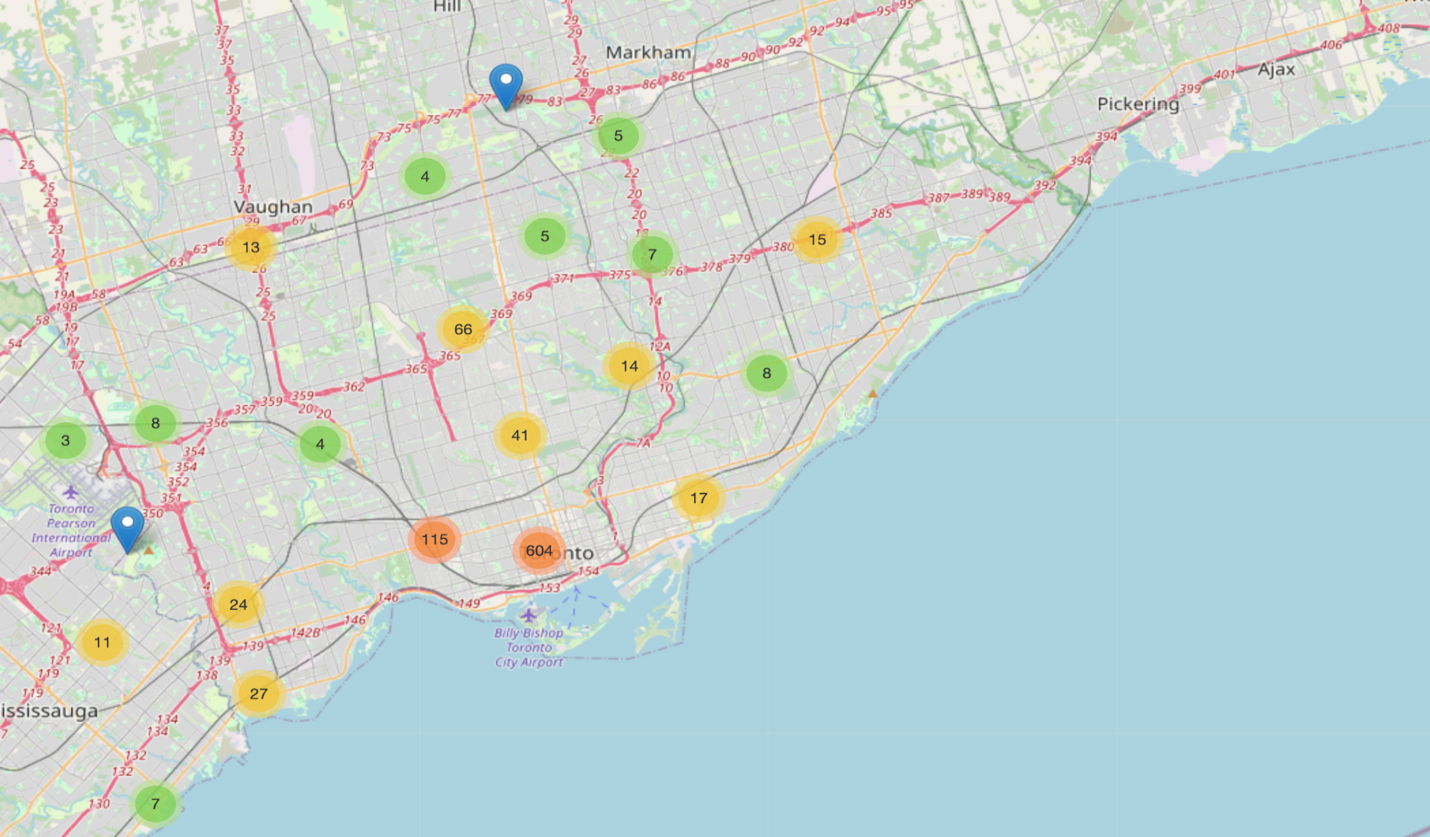
**A graph with blue lines

Description automatically generated**

**Geolocation Analysis**

**Heatmap of Event Locations**

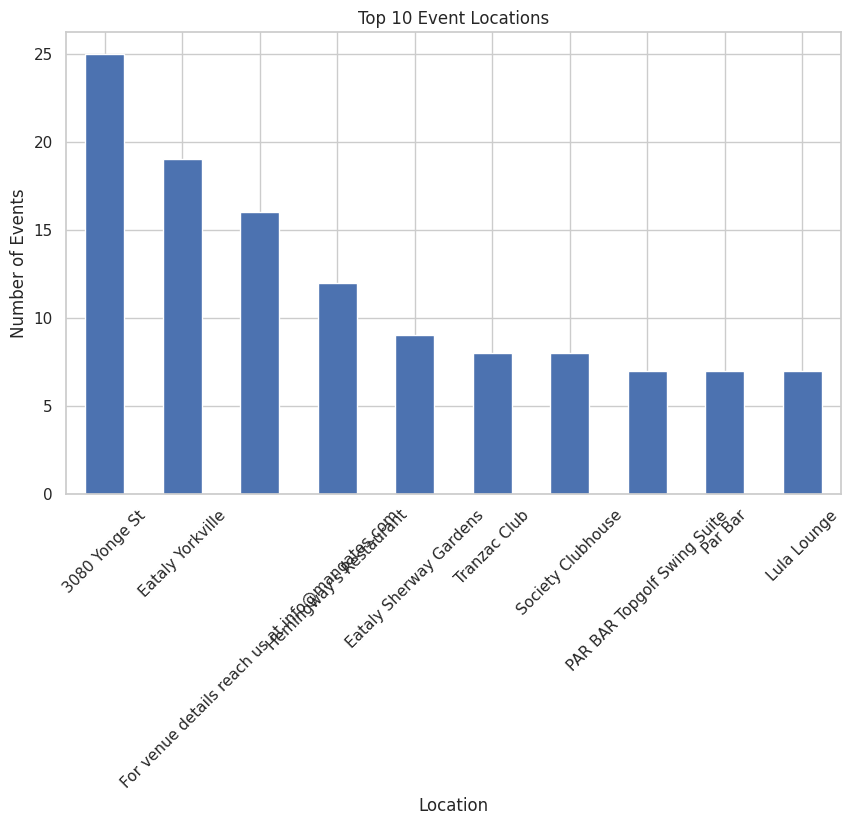
The geolocation heatmap reveals clusters of events across Toronto, with the downtown core being the most densely populated. The central area, including high-density regions such as the Entertainment District and adjacent neighborhoods, serves as a hub for most events. Suburban areas, such as Vaughan and Markham, exhibit lower activity levels. This clustering indicates a preference for central, well-connected venues, likely due to accessibility, infrastructure, and proximity to potential attendees.



The spatial distribution also suggests strategic planning by organizers to cater to high footfall areas and ensure better event turnout. Regions with sparse activity could present opportunities for expansion or new event outreach initiatives to tap into underserved communities.

**Top 10 Event Locations**

The bar chart highlights the most popular event venues, with 3080 Yonge St being the top location, hosting the maximum number of events. Eataly Yorkville and similar premium venues also rank high, reflecting a preference for well-known, upscale settings. These locations are likely chosen for their appeal to specific demographic segments, logistical conveniences, and ability to enhance the event experience.



The diversity in the top locations—from high-end restaurants to casual bars and lounges—demonstrates a wide range of event types, catering to different preferences and purposes. This information can guide future event planning to identify strategic venues that align with target audience interests and logistical requirements.

**Attendance Prediction**

The Attendance Prediction section highlights the application of multiple predictive modeling techniques to estimate event attendance ranges. Using a combination of supervised learning models like XGBoost, Random Forest, and Decision Tree alongside a Deep Learning approach, the analysis explored the impact of various features, including event type and timing, on attendance. The results, as summarized in Table 1, demonstrate the effectiveness of these models in addressing data imbalances and capturing attendance patterns, with XGBoost emerging as the top-performing model in terms of accuracy and consistency.

Table 1: Results of the Predictive Models Performed on the Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy (%) | Precision (Weighted Avg) | Recall (Weighted Avg) | F1-Score (Weighted Avg) |
| XGBoost | 65.27 | 62 | 65 | 62 |
| Random Forest | 62.34 | 56 | 46 | 49 |
| Decision Tree | 61.51 | 60 | 62 | 60 |
| Deep Learning | 59 | 56 | 59 | 57 |

**Supervised Learning**

The analysis utilized supervised learning techniques, including XGBoost, Random Forest, and Decision Tree classifiers, to predict the attendance range for events. The features included categorical and numerical variables such as event type, day of the week, month, hour, and timezone. The target variable, attendance range, was encoded using LabelEncoder. Class weights were computed to handle imbalances in the dataset.

After hyperparameter tuning, XGBoost achieved the highest accuracy of 65.27%, followed by Decision Tree (61.51%) and Random Forest (62.34%). The classification reports highlighted the challenges in predicting smaller classes like "100+" and "50-100," where precision and recall were low, likely due to data imbalance. XGBoost performed best in terms of weighted averages, indicating its suitability for attendance prediction. These models offer a baseline for improving prediction accuracy through feature engineering and advanced techniques.

**Deep Learning**

A deep learning model using a neural network architecture was also implemented for attendance prediction. The model consisted of multiple dense layers with dropout to prevent overfitting and used categorical cross-entropy as the loss function. With early stopping and learning rate reduction, the model achieved an accuracy of 59% on the test set. While its performance was slightly below traditional machine learning models, it captured non-linear patterns in the data. The learning curves showed stable training and validation loss, suggesting a well-generalized model.

The combination of traditional machine learning and deep learning approaches provides a comprehensive framework for attendance prediction, with opportunities to enhance accuracy through data augmentation and additional features.

**Time series forecasting for number of attendees:**

For time series forecasting for number of attendees, the code performs a series of data processing and machine learning tasks aimed at classifying, forecasting, and analyzing event data based on attendance and cluster predictions. The workflow starts by importing the necessary libraries and loading datasets related to event descriptions and their associated characteristics, such as attendance.

The first part of the code focuses on text classification. It uses a TF-IDF vectorizer to transform event descriptions into numerical features, then applies a Logistic Regression model to classify events into clusters based on their descriptions. After training, the model is used to predict clusters for new events. The predicted cluster labels are added to the dataset, and the results are outputted to a CSV file for further analysis. Cluster 0 is for Spiritual and Personal Growth events, Cluster 1 is for Social and Creative Activities and Cluster 2 is for Science and Professional Development.

Next, the code processes the event attendance data, aggregating it by date and predicted cluster, and uses time series forecasting techniques to predict future attendance. It prepares the time series data by creating lag features (i.e., previous attendance values) for each cluster, then splits the data into training and testing sets. An ensemble forecasting method, using a RandomForestRegressor, is applied to make predictions for the next 60 days. The model is trained on historical attendance data, and future predictions are made using the most recent available data, rolling the predictions forward to ensure continuity.

The accuracy of the forecasts is evaluated using various metrics, such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). These metrics are used to assess how well the model has captured the patterns in the data. The forecast results, which include predictions for 60 days into the future, are visualized in plots alongside the training and testing data for each cluster, and the average predicted attendance and range of attendees for the forecast period are also computed.

**Recommendations and Findings**

The analysis revealed several key findings that informed actionable recommendations. It was observed that Saturdays, particularly in November and December, see the highest concentration of events, making these optimal times for scheduling key events. Marketing efforts were advised to focus on high-density areas like Downtown Toronto to maximize reach and engagement. The use of real-time heatmaps was recommended to manage crowd flow and allocate resources efficiently, addressing both logistical and safety concerns. To further improve attendance forecasting, the project suggested incorporating external factors such as weather conditions and exploring advanced predictive techniques like deep learning. These recommendations highlight practical strategies for optimizing event management processes and improving participant experiences.

**Results of time series forecasting for number of attendees:**

The results indicate that the number of events varies greatly across clusters. Cluster 0 has only 27 events, while Cluster 1 has 1010 events, and Cluster 2 has 182 events. These differences in the number of events could suggest that Cluster 1 represents a larger, more diverse set of events, while the other clusters may represent more niche or specialized events.

A graph of a graph

Description automatically generated

In terms of accuracy, the model's performance varies between clusters. For Cluster 0, the Mean Absolute Error (MAE) is relatively low at 21.97, indicating that the model's predictions are fairly close to the actual attendance. However, the predicted range of attendance is narrow, from 1 to 5 attendees, and the average predicted attendance is only 2. This could suggest that the events in this cluster are smaller or more specialized.

A graph with different colored lines

Description automatically generated

Cluster 1, which has the highest number of events, shows a much higher MAE of 357.20, indicating that the predictions for this cluster are less accurate. Despite the larger error, the predicted range of attendance for Cluster 1 is between 441 and 794 attendees, with an average of 540 attendees. This suggests that the events in this cluster tend to have a larger, more variable audience.

A graph of a graph

Description automatically generated

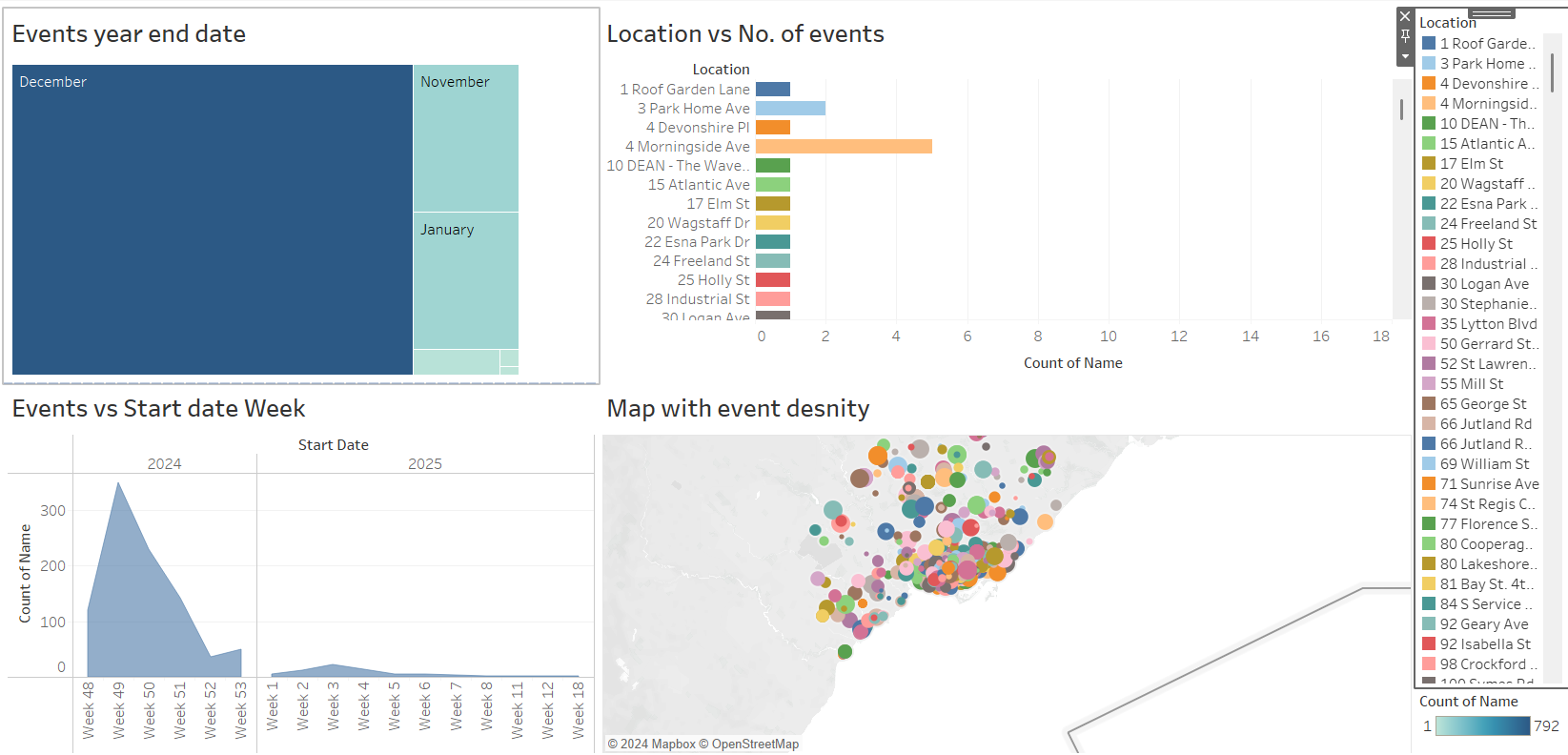
For Cluster 2, the MAE is 114.81, which is moderate compared to the other clusters. The predicted range for this cluster is between 45 and 142 attendees, with an average of 65 attendees. This indicates that the events in this cluster tend to have moderate attendance compared to Cluster 1 but higher than Cluster 0.

**Tableau Dashboard Analysis & Report Guide**

**Link for the dashboard:**

[**https://public.tableau.com/app/profile/nagendra.hegde1985/viz/TorontoEventbritedata/Dashboard1?publish=yes**](https://public.tableau.com/app/profile/nagendra.hegde1985/viz/TorontoEventbritedata/Dashboard1?publish=yes)

This report analyzes the event data using various visuals to uncover trends, key locations, and time-based patterns. Below is the dashboard and the interpretations and key findings from each visual:

****

1. **Events Year-End Date (Tree Map):** The tree map highlights the distribution of events based on their end dates by month:
   * How to Read: Larger blocks represent months with a higher count of events.
   * Insights: December accounts for the majority of event closures, indicating a peak in event activity during the year-end. November follows as the second busiest month.
2. **Location vs. Number of Events (Bar Chart):** This bar chart displays the count of events at each location:
   * How to Read: The length of each bar represents the number of events held at a particular location.
   * Insights: "4 Morningside Ave" is the most active location, hosting the highest number of events, while other locations have relatively fewer events.
3. **Events vs. Start Date Week (Area Chart):** This area chart depicts the weekly distribution of events based on their start dates:
   * How to Read: The height of the chart indicates the number of events starting in each week.
   * Insights: Event activity peaks in Week 49, suggesting a significant rise in events during late November to early December. Activity declines sharply after Week 52, indicating a slowdown at the start of the new year.
4. **Map with Event Density (Geospatial Map):** This map visualizes event density across different geographical areas:
   * How to Read: Each circle represents a location, with the size of the circle indicating the number of events at that location.
   * Insights: Events are clustered in specific regions, with some areas showing significantly higher event density. These high-density regions may warrant focused resource allocation or marketing efforts.

**Overall Insights**

* **Temporal Patterns:** Event activity is highest toward the end of the year (December), with a sharp decline at the beginning of the following year. Weekly trends show a peak in late November (Week 49).
* **Location Analysis:** Certain locations, such as "4 Morningside Ave," stand out as hotspots for events, while other areas have fewer events.
* **Geographical Distribution:** Events are geographically concentrated, with several high-density areas requiring strategic attention.

This dashboard provides a comprehensive overview of event trends, supporting data-driven decisions for event management, resource allocation, and targeted planning.

**Conclusion**

The "Event" project demonstrates how integrating geolocation data and predictive analytics can transform event management by addressing key challenges such as resource allocation and crowd safety. This analysis provides actionable insights for event planners to optimize scheduling, marketing efforts, and crowd management strategies. By identifying clear trends in event distribution and attendance patterns, the project underscores the value of data-driven, targeted planning.

Future enhancements, such as incorporating external factors like weather or social trends and adopting advanced predictive models, will further improve accuracy and applicability. Using clustering and time series analysis, the project categorized events into groups based on descriptions and successfully forecasted attendance. While predictions for larger, more diverse events (Cluster 1) showed greater variability, smaller, more consistent events (Cluster 0) had more accurate forecasts.

These findings are valuable for event planners aiming to allocate resources more effectively and ensure smooth operations. However, refining prediction models for large or unpredictable events will be a key focus moving forward, ensuring that this approach remains practical and adaptable for all stakeholders, whether in business or event management.

**References:**

Trinh, T., & Vuongthi, N. (2022). A predictive paradigm for event popularity in event-based social networks. *IEEE Access*, *10*, 125421–125434. <https://doi.org/10.1109/access.2022.3225734>

Zhang, X., Zhao, J., & Cao, G. (2015). Who will attend? -- predicting event attendance in event-based Social Network. *2015 16th IEEE International Conference on Mobile Data Management*, 74–83. <https://doi.org/10.1109/mdm.2015.23>

Zhou, D., Li, H., Liu, S., Song, B., & Hu, T. (2017). A map-based visual analysis method for patterns discovery of mobile learning in education with Big Data. *2017 IEEE International Conference on Big Data (Big Data)*, 3482–3491. <https://doi.org/10.1109/bigdata.2017.8258337>

Allgaier, R. L., Shaafi-Kabiri, N., Romney, C. A., Wallis, L. A., Burke, J. J., Bhangu, J., & Thomas, K. C. (2019). Use of predictive modeling to plan for special event medical care during mass gathering events. *Disaster Medicine and Public Health Preparedness*, *13*(5–6), 874–879. <https://doi.org/10.1017/dmp.2019.1>

Milusheva, S., Marty, R., Bedoya, G., Williams, S., Resor, E., & Legovini, A. (2021). Applying machine learning and geolocation techniques to social media data (Twitter) to develop a resource for urban planning. *PLOS ONE*, *16*(2). <https://doi.org/10.1371/journal.pone.0244317>

Pennefather, S., & Irwin, B. (2014). An exploration of geolocation and traffic visualisation using network flows. *2014 Information Security for South Africa*, 1–6. <https://doi.org/10.1109/issa.2014.6950507>

Wu, X., Dong, Y., Shi, B., Swami, A., & Chawla, N. V. (2018). Who will attend this event together? event attendance prediction via deep LSTM Networks. *Proceedings of the 2018 SIAM International Conference on Data Mining*, 180–188. <https://doi.org/10.1137/1.9781611975321.21>