Tab 1

**BAX372T Final Report: Austin Traffic Incidents**

1. **Project Goals**

The goal of this project is to analyze traffic incident data in Austin, Texas, from May 202 to December 2024 and combine it with local weather data to identify patterns and correlations between location, weather conditions and traffic incidents. By merging our weather and traffic incidents datasets, this analysis aims to identify high-risk areas, specific times, and weather conditions that contribute to a higher number of accidents.

Key objectives include:

1. Identifying traffic incident-prone intersections, neighborhoods, and specific subsections of Austin.
2. Analyzing patterns related to specific days, times, and months that experience higher rates of traffic incidents.
3. Providing insights that can guide city planners and policymakers in directing resources to areas that need them the most.

Ultimately, the main goal of our project is to improve traffic safety in Austin by providing evidence-based recommendations that could lead to more targeted investments in infrastructure, resource allocation, and safety measures, benefiting both drivers and pedestrians across the city.

1. **Data Collection**

The analysis uses two primary datasets: Traffic Incidents and Weather data. The Traffic Incidents Data was sourced from the City of Austin Open Data Portal, containing detailed reports of traffic-related issues from 2018 to 2024. Each record includes fields such as Traffic Report ID, Published Date, Issue Reported (e.g., Collision, Traffic Hazard, Icy Roadway, Boat Accident, Obstruct Highway), Latitude, Longitude, Address, and the Agency responsible for the report. This dataset provides a comprehensive view of traffic incidents across Austin, allowing for geospatial and temporal analysis of incident trends.

The Weather Data was obtained from Visual Crossing Weather, offering daily atmospheric conditions corresponding to the Austin area. The dataset includes key meteorological variables such as datetime, temperature, humidity, precipitation, windspeed, cloudcover, and visibility. These features are essential for exploring the relationship between weather patterns and traffic incidents.

1. **Exploratory Analysis**

After defining our project's goals and sourcing and pre-processing the relevant datasets, we were able to begin performing some exploratory analysis that yielded some interesting patterns and abnormalities.

Before anything else, we explored how incident count distribution looked on average. To achieve this we grouped the entire dataset by Hour of Day and graphed the total number of incidents reported throughout the entire range of dates (01 May 2021 to 31 December 2024). As can be seen in **Figure 1**, most incidents are reported at around 10PM, with high counts happening from noon to 2AM. Interestingly, commuting hours 6AM to 10AM show the least amount of total incidents reported. This does not make much sense, as our instincts told us incidents should be way more common from 6AM to 9AM than from 1AM to 4AM, purely because of the overwhelming higher number of cars on the road. *Perhaps there is some delay in incident reporting, or the dataset reports the time incidents were logged by police instead of when they actually took place.*

Moreover, we wanted to identify any patterns or abnormalities on the amount of daily incidents throughout the date range on the dataset. We looked at various specific incident types and identified that, roughly, incident counts stayed flat on average. I.e, despite a roughly 6% growth in the Austin Metropolitan Area population in the date range, incidents did not increase significantly. Additionally, as can be observed in **Figure 2**, there are some spikes which stand out from the rest of the data. When we took a closer look, *we found that these spikes exactly matched some significant events in the city. For example, the freeze and power crisis in February 2021, the freeze in February 2022, and the ice storm in January-February 2023. Less prominent spikes also take place during each year's SXSW and F1 weekend*. Also somewhat interestingly, there's a drop in incidents around January 2024 for which we cannot find an adequate explanation. *Maybe there was some sort of system failure or police staffing crisis.*

In order to further explore the dataset, we decided that a map visualization of the incidents would be an effective tool (See **Figure 3**). In the resulting heatmap, which roughly delineates Travis County borders with some outliers, we can see the areas where most incidents take place. It is evident that the biggest concentration of incidents takes place in downtown Austin and continues throughout the core of the city. A significant incident trail also follows the path of I-35 and the MoPac Expressway, which also makes a lot of sense, as these roads carry a much higher flow of cars at much higher average speeds. The outskirts and suburbs within Travis County have much lower incident counts, with arterial roads such as Koenig Ln, SH-183, and Parmer Ln being easily visible. Also worth highlighting is the increased incident concentration in the east border of the county, as these areas are sparsely populated and have no major northbound roads, as the heatmap seems to suggest. *A possible explanation is that Travis County Sheriff Deputies misreport incidents that happen in Bastrop County as having taken place in the border of Travis County.*

Additionally, further analysis indicates that there are no interesting trends or abnormalities made apparent by grouping by other variables or factors, like season, year, or month.

1. **Solutions and Insights**

Our solution aims to answer the research question: **Out of all the areas where traffic incidents occurred, which ones can be classified as high-risk and which as low-risk?** This classification allows us to distinguish between consistently critical zones and those that may have experienced a few isolated incidents by chance. We first defined “high-risk” using a data-driven clustering approach. By applying KMeans clustering to the latitude and longitude of each incident, we grouped all incident locations into 50 geographic clusters. Clusters with over 1,000 reported incidents were labeled as high-risk, while those below that threshold were considered low-risk (See **Figure 4**). This approach provides a more objective definition of high-risk based on density rather than arbitrary geographic boundaries.

Next, we engineered a set of features that we believe have explanatory power to the volumes of incidents. Temporal features included hour of the day, day of the week, month, whether the time was during rush hour, and AM/PM indicators. We also incorporated weather data such as temperature, precipitation, windspeed, cloudcover, and visibility. We are mostly leaning into spatial and temporal factors, which makes our model and insights more actionable for city planning, traffic safety, and dispatch optimization.

To classify high-risk areas, we tested several models—Bagging, Random Forest, and Gradient Boosting. We ultimately selected Gradient Boosting, which was configured with 200 decision trees and a maximum depth of 4, because it yielded the best performance. The model offered the highest testing accuracy at 85.5% while gradually reducing prediction error and increasing recall up to 94% for high-risk zones, making it a strong fit for this classification problem. As the false negative error cost is the most significant, recall is more important than precision. Result of high risk being classified as low risk can lead to more accidents and even increased deaths.

We found that some factors were particularly important for the model. Among all the features, visibility and precipitation had high effects on classification. This is expected, as rain and low visibility are typically perceived to make road conditions more dangerous, and thus, higher-risk of incidents. Other engineered features like day of week and rush hour indicator variable were also significant. It follows that higher traffic counts lead to higher rates of incidents, which also makes sense.

Other features had lower importance. For example, temperature, cloud cover, precipitation type, and month, were almost useless. This is not particularly surprising, as either other variables captured most of the effect that they could have potentially exerted, or there was simply no effect, as was the case with month.

Additional insights we gleaned from the datasets can also be found italicized in section III. Exploratory Analysis. Worth highlighting the most is the interesting buildup of incident reports in the low density Travis County-Bastrop County border, which we posit can be explained by incentives to collect reports by the Travis County Sheriff's Office, or incentives to reduce report counts in Bastrop County Sheriff's Office.

In order to address these findings, the City of Austin and the Travis County Sheriff's Office should pay attention to areas classified as high risk, especially those that are not already closely monitored because of highway or arterial road status (think I-35, Congress Ave., Guadalupe St., Parmer Ln, etc.). Examples for these important areas are: 45th St., Cameron Rd., and Oltorf St. Nevertheless, city resources should generally be at least partially redeployed if they are in low-risk areas in order to maximize benefit to the community and increase safety. Moreover, incentives at County Sherrifs’ Offices should be revised to ensure data integrity and transparency.

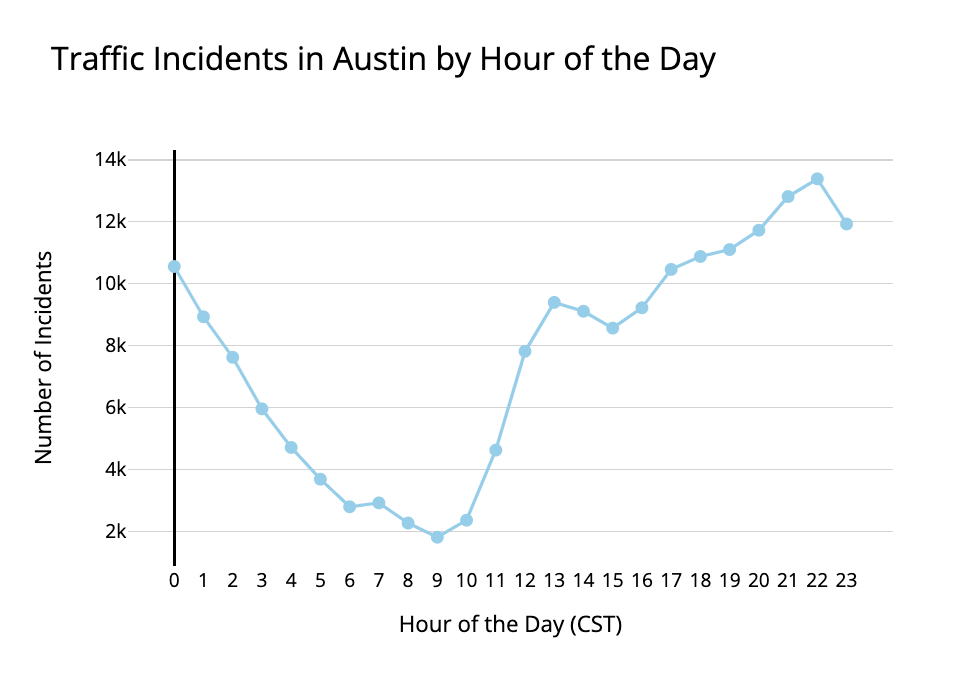
1. **Conclusion**

Our analysis demonstrates that traffic incidents in Austin are not randomly distributed—they cluster in specific geographic areas and time windows, often influenced by environmental factors like weather. From our visualization plots and heatmaps, we observed that the majority of incidents are concentrated in the downtown area and around the UT Austin campus. By further combining temporal and weather data, we built a classification model that identifies high-risk zones. Gradient Boosting proved to be an effective model for this task, helping us move beyond simple counts toward a predictive understanding of incident hotspots.

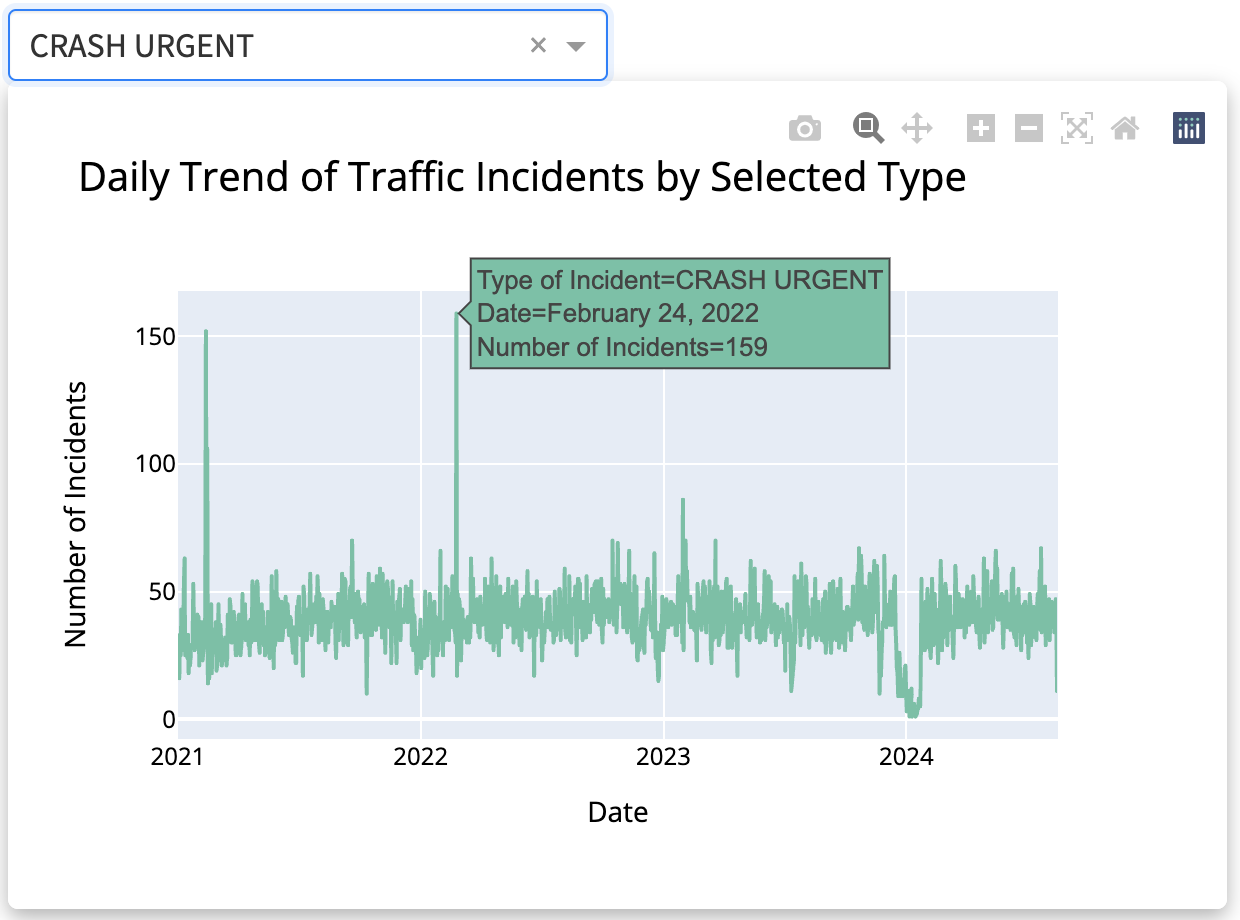
This work lays the foundation for data-informed urban decision-making. The city of Austin could leverage these findings to direct traffic safety resources, optimize emergency response, or even inform infrastructure changes in high-risk zones. Ultimately, we hope that this analysis can contribute to smarter urban planning and a safer community of commuters.

1. **Appendix**

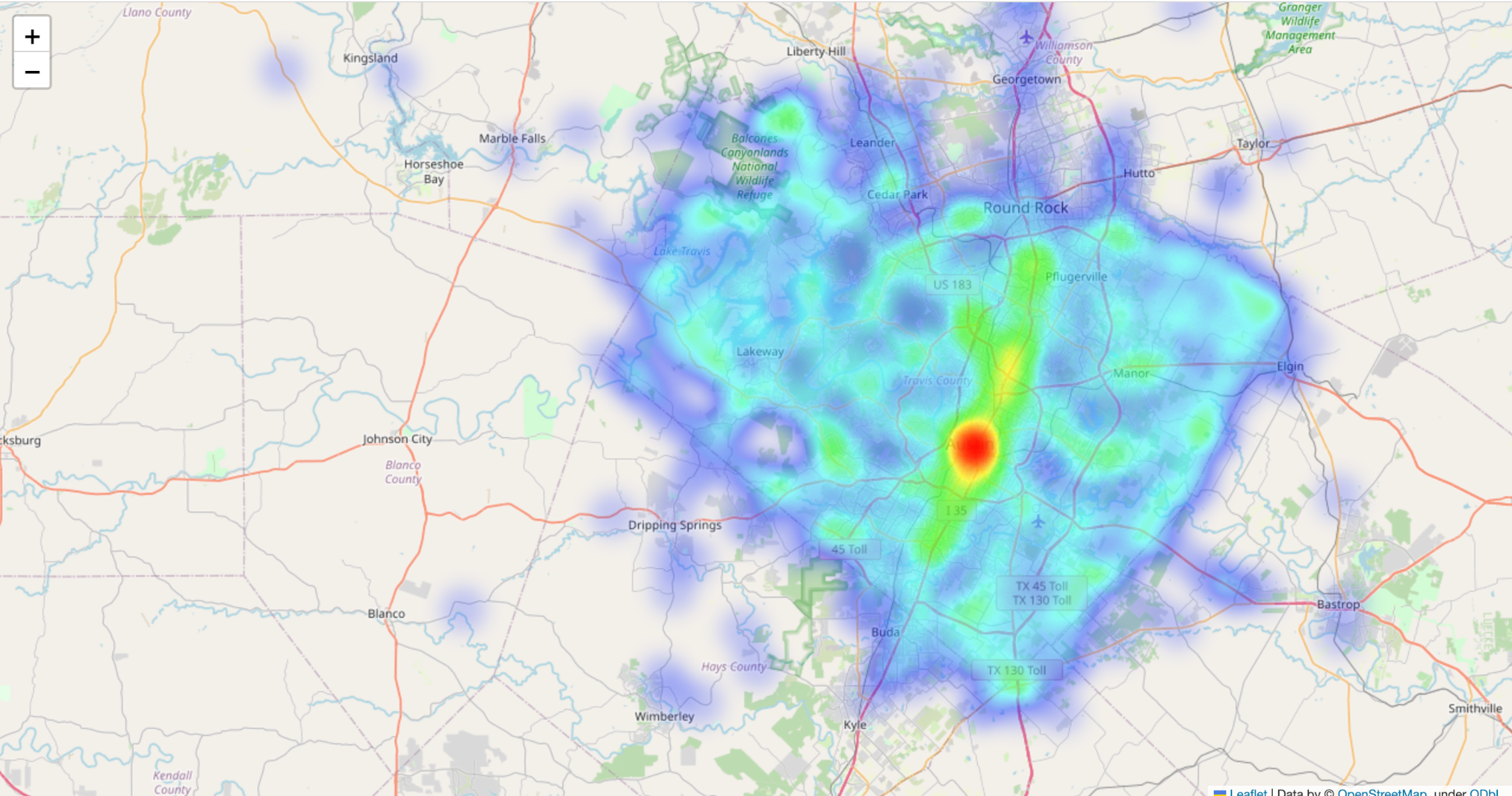
**Figure 1:** Total Traffic Incident Count by Hour



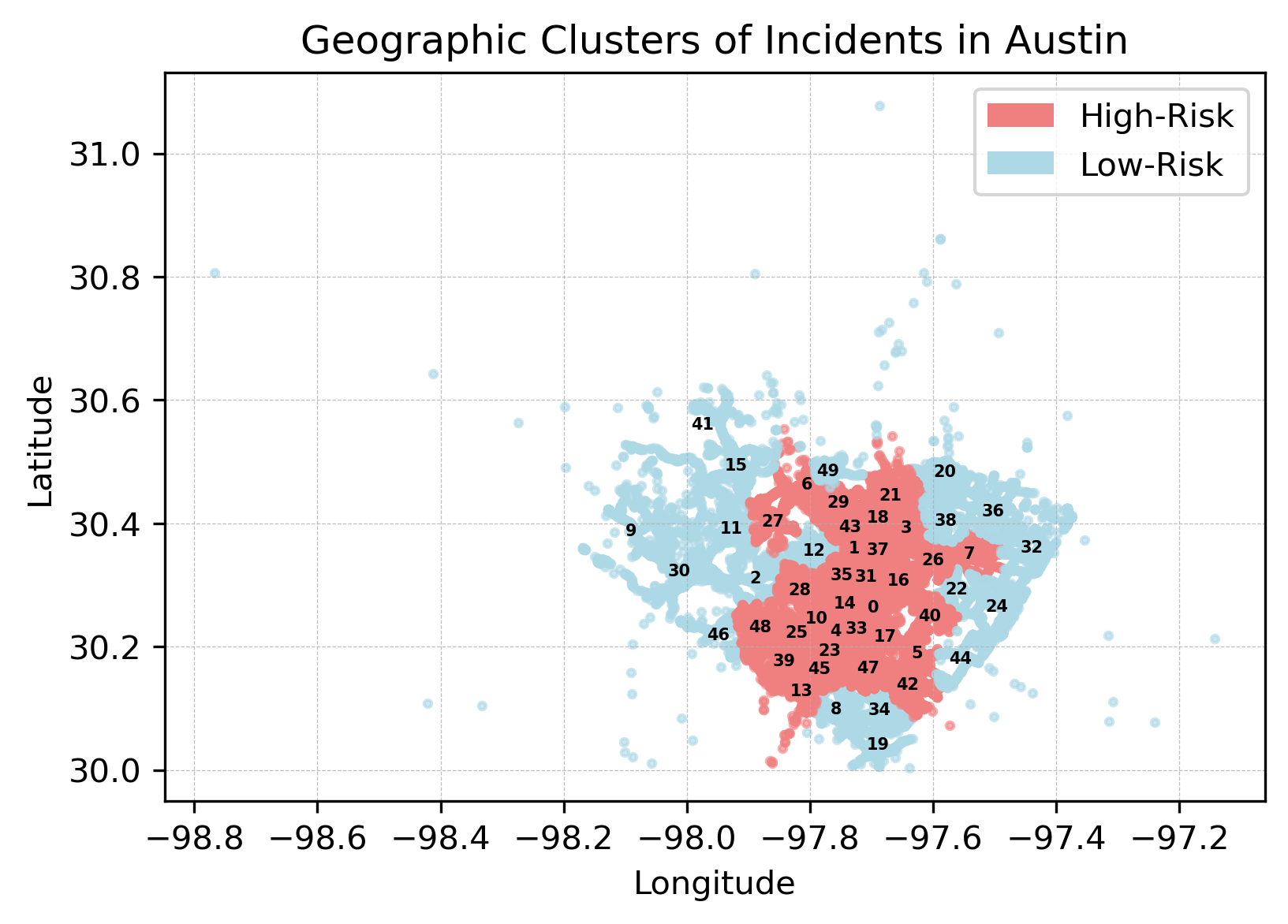
**Figure 2:** Total Daily Crash Count in Date Range



**Figure 3:** Incident Concentration Heatmap



**Figure 4:** KMeans Clustering Risk Classification



1. **References**

“Total Weather Data.” *Visual Crossing*, 2 Apr. 2025, www.visualcrossing.com/weather-data/

City of Austin. “Real-Time Traffic Incident Reports: Open Data: City of Austin Texas.” *Data.AustinTexas.Gov - The Official City of Austin Open Data Portal*, 15 Apr. 2025, data.austintexas.gov/Transportation-and-Mobility/Real-Time-Traffic-Incident-Reports/dx9v-zd7x/about\_data