Big Data Emergency Response Project Report

Yifan Wu (wuy56, yfian.wu@Vanderbilt.edu)
Xiaoliang Zhu (zhux19, xiaoliang.zhu@Vanderbilt.edu)

May 1, 2023

1 Overview

The significance of efficient emergency response systems in cities cannot be overstated. They play a crucial role in minimizing loss of life and property during incidents such as accidents, medical emergencies, urban crimes, and fires. Adopting a data-driven, proactive, and principled approach to emergency response management can reduce human and financial losses and inform policy improvements and safety measures.

1.1 Project Overview

In this project, we aimed to develop an interactive interface and analyze data to gain valuable insights into incidents occurring in and around Davidson County, Nashville, TN. By providing unique insights into these datasets, we aimed to enhance future emergency response operations. Our objectives included identifying incident clusters, uncovering spatial-temporal patterns, and evaluating the impact of various factors on emergency response times. This comprehensive analysis enabled us to make informed recommendations for improving emergency response systems in the region.

1.2 Data Overview

We employed five datasets: `incidents` [1], `weather`, `traffic`, `census`, and `demographics`. The `incidents` dataset contains all the incidents recorded between January 2017 to early 2021, with a gap from July 2017 to July 2018. The dataset contains 29770 entries ranging in 13 columns, with notable ones being the `latitude` and `longitude` providing precise location of the incident, `time_utc` giving the Coordinated Universal Time of the incident, `response_time_sec` recording the time it took for the emergency response team to reach the incident, and `xdsegid` which provides the corresponding road segment id in `traffic` dataset if the incident happened on a road.

The `weather` dataset gathers 120,000 weather station hourly reports in Tennessee between January 2010 to February 2012. The `weather` dataset contains similar date and location information recorded in columns `timestamp_utc`, `gps_coordinate_latitude`, and `gps_coordinate_longitude`. Among the 22 fields of weather properties, we picked `humidity`, `rain`, `snow`, `temperature`, and `wind_speed` as our focus because we concluded that properties such as `UV Index`, `solar radiation`, or `dew point` may not have as strong of an impact on the incident occurrence or the emergency response time as the previously mentioned five properties.

The `traffic` dataset contains the traffic sensor data between April 2017 to February 2022 in Davidson country, Nashville, TN, with 11 columns. The data is recorded once every five minutes, with the UTC time stored in `measurement_tstamp`. It also contains the save `xd_id` which is the road segmented ID as the `incidents` dataset. Other notable traffic fields include `speed`, which is the measured speed of the vehicles passing the sensor at the measurement time, `average_speed`, which is the average speed of a particular sensor over a period of time, `reference speed`, which is the typical or speed limit of a particular road segment, `congestion`, which is an estimated index of congestion calculated using the `speed` and `reference_speed`.

The `census` dataset mainly contains the geometry data of all the tracts and county divisions within Tennessee. The data are stored in polygons formed by vertices coded as latitude and longitude coordinates. This data is used to map and visualize the whole system and contains columns such as `countyfp` and `tractce` which are the unique identifier of each division.

Finally, the `demographics` dataset contains demographic information in a 5-year estimate for census tracts in 2015. It contains similar county and tract identifiers as the `census` dataset. We used most columns in

the `demographics` dataset as they all provide some insights, or at least can help rule out some assumptions. Specifically, we picked `total_population`, `median_household_income`, `median_family_income`, and all columns related to gender and race situation in each tract. The difference between household and family is that a household is everyone living in a house, which may contain multiple families, and a family is a group of people related by blood, marriage, or adoption, and can be spread among multiple physical houses. It is also noteworthy to emphasize that these data are collected in the unit of census tract, a geological region, and not by the incident survivors' demographic data.

2 Solution

In this section, we will present in detail our implementation of the solution, including both the design and the specific technologies we used to achieve the result.

2.1 Design

There are two focuses in this project - the census and demographics relationships with the incidents, and the weather and traffic insights about the incidents.

Starting with the census and demographics data, we first need to assign a geographic location to each incident by placing each one in its corresponding census tract polygon, which is achieved by comparing the GPS latitude and longitude of every incident with the geographic range of each census tract. Tracts far from Davidson County and with no incident happening are filtered out. Then, with the geographic location and district identifiers, we are then able to match demographic information including gender, race, and income with each census tract and incident according to the county and tract identifiers.

Then, the weather data is processed with the goal of estimating the weather status at the location and time of each incident. However, the weather data is only recorded at each weather station with a fixed location, and at one-hour intervals. While it is possible to interpolate weather data in between hours for potentially more accurate estimation, we decided that it is not significant as large weather trends do not alter much in the time span of an hour. Therefore, we simply rounded the incident time to the nearest hour and focused our attention on geographical location. We first decided on a 10km by 10km grid spanning the whole of Davidson County. Weather stations within the same grid will have their recordings averaged, and incidents that fall in the same grid will get assigned the same weather data. This grid size is enough such that weather recordings within each grid are consistent. Making smaller grid sizes can make the estimates more accurate; however, the problem which we have already encountered - some grids not getting assigned at least one weather station - would be more severe. Instead of making the grid larger and risking the weather data being less accurate, to solve this problem, we decided to perform a second operation to fill the gaps. We combined all the incidents that were not already assigned weather data from the grid and gathered the nearest neighboring weather station for a weather reading. With this additional operation, every incident can get assigned weather data. With a closer inspection, we found that there are only five anomalies where the weather station is more than 30km away from the incident location, which is an error we are willing to take for the consistency of having no null pairings. After these operations, we fully cover the relationship between incidents and weather.

Finally, we matched each incident that happened on the road with their corresponding traffic data based on road segment and time. Similarly, since traffic data are recorded in five minutes intervals, we rounded the incident times to the nearest 5 minutes. The incidents and traffic datasets are not perfect pairing with some incidents not matched with traffic data. While we can use the road map and geolocation data of the incident to match each incident with the nearest neighboring road segment, unlike weather which is always present, it is meaningless to assign a house fire or park burglary to the average car speed of the nearest road.

After processing all the datasets, we made two interfaces showcasing our findings. The first interface contains four sections - one displaying the temporal trends of incidents during the entire period, one showing the geographical distribution of incidents on the region map, one revealing the temporal distribution of incidents (grouped by month and weekdays), and one reflecting the demographic distribution of each tract, including distributions of ethnicity, gender, and wealth. Detailed incident distribution on month or weekday of a specific

tract will be presented when the cursor hovers over the tract, along with the demographic information of this selected tract. The section on incidents' temporal trends contains a line plot indicating the change in incident counts across the entire period. It allows the user to arbitrarily brush and select sub-periods, while the map will update to reveal the incident data within the selected period. The second interface includes two sections, portraying the impacts of weather and traffic, respectively, on the incident distribution. Each section contains three charts that cover the simple count, frequency, and response time of incidents under a specific weather or traffic condition over the whole geographical and temporal span of the datasets. We did not plot weather and traffic data onto the map, as weather and traffic correlations are more general, and thus is beneficial to adopt the whole large dataset for a more observable and reliable trend.

2.2 Technology

For storage, we used Amazon S3 for its deep integration into the AWS ecosystem and the familiarity from using it throughout the semester. We started by creating one S3 bucket and uploading all the data into this single bucket. We then set IAM permissions for encrypted access from the AWS accounts of both team members. Using a single bucket saved our time for uploading data and also saved storage space. Later during the data-processing stage, the resulting tables are also stored in the same S3 bucket for visualization after completing each processing step.

With Pandas, an SQL join based on the county and tract identifiers was first conducted to match the demographic information to each census tract. We then processed the census data using GeoPandas, one of the first Python libraries we found that can open and manipulate SHP files. With its convenient spatial join, we could achieve our goal of assigning each incident with its corresponding census tract based on coordinates and the polygon.

The weather and traffic datasets are processed using Amazon Athena. It is SQL based and can directly read and write to Amazon S3 buckets. For the weather data, we first performed a left join between the incidents and weather data based on grids, which is achieved by rounding down the longitude and latitude columns to the nearest 0.1 degrees. Weather data with the same time and location after rounding is averaged. The nearest weather station is calculated by using the `ST_Distance` function. For processing speed, we first created a table with only three columns - the incident ID, the nearest weather station ID, and the distance between them. We then used this table as a reference to join the corresponding incidents and weather data. This action is only performed on incidents that have a null value in the `avg_temperature` column, which means it did not get weather data from the previous grid-based pairing. The traffic dataset was simply left joined on road segment ID with the previous result without resolving null columns due to the reasons mentioned in Section 2.1.

Finally, the resulting tables are visualized using Observable. Similar to a Jupyter notebook, an Observable notebook is a JavaScript-based online interactive notebook. We mainly employed D3.js, a JavaScript library for producing dynamic data visualizations in web browsers, and Observable Plot, an open-source, JavaScript library for visualizing tabular data, to implement our visualization interfaces. We constructed our interfaces in two separate notebooks, both of which are stored in a private Observable organization and maintain unlisted public access, so only users obtaining both permission and the notebook URL can access the notebook. The notebooks render the visualization in real-time and require AWS credentials to acquire data from the Amazon S3 bucket. The notebooks also include certain data transformations and manipulations to support interactions and selections from users.

3 Results and Insights

In this section, we will present the insights we gained by navigating through our interface, and try to propose a hypothesis for the cause of any trend we observe.

3.1 Demographic Insights

A full view of the demographics interface can be found in Figure 1. Starting with the second row left, we can set `Box plot time distribution` to `Month`, and set `Map color representation` to `Incident average response time`. Under this setting, we see a trend where the average response time during winter months (September to February) is significantly longer with an average of 394.7 seconds compared to an average of 375.3 seconds during the summer months (March to August). Such a difference is possibly due to responders' need for further preparation on equipment or clothing during winter, as well as people's natural unwillingness to mobilize during cold weather. Moreover, if we set `Box plot time distribution` to `Week`, and set `Map color representation` to `Incident count`, we can see an outlier with significantly more incident counts on Friday (4957) than on other weekdays (< 4500). We suspect that the reason for this outlier may relate to people's anticipation for vacation on weekends. As people are hurried for returning home, it can be more possible for them to cause incidents along the road.

On the second row right, setting `Demographic choice` to `Median family income`, and setting `Map color representation` to `Incident frequency`, we can see that regions with a lower median family income have a higher frequency of incidents. Such a correlation is expected as the frequency of illegal events, like drug overdose and violent crimes, is also relatively high in low-income neighborhoods, while those illegal events are dominant causes of incidents. Interestingly, we also observed that if `Map color representation` is set to `Incident average response time`, there appears a decreasing trend in response time as income decreases. We suspect that this is because the responders put more personnel on hold in low-income neighborhoods due to their more frequent need for emergency responders. We did not find a significant correlation between incident frequency or response time vs. gender or race, and any small amount of trend can be explained better by other factors such as population size or income level. However, the futile search for a trend in gender or race may also be attributed to the fact that the `incidents` dataset does not contain records for these properties on a per-incident basis; thus we have to resort to using the region's demographic data, which would naturally be more correlated by other factors of the region than these columns in question.

By setting `Map color representation` to `Incident count`, we can observe from the map on row three that the downtown area is the peak of incidents, and the farther from the center, the fewer incidents there are. This pattern is obvious as not only is there more traffic and population downtown, but the road is also generally narrower due to the limited real estate causing more potential problems. However, by setting `Map color representation` to `Incident average response time`, we observe the opposite trend: the farther from downtown, the longer the response time is. This is also self-explanatory - since most incidents happened in the downtown area, the responders tend to patrol around the center, making it swifter to respond to incidents around downtown but slower to get to the outer shell of the city.

Finally, there is one interesting fact we found in the last row of the demographics visualization interface. If we set `Time series granularity` to `Month`, we observe a significant dip in incident count around March to May 2020. This coincides with the peak of the COVID pandemic, and consequently, everyone is in quarantine, thus not causing many accidents.

3.2 Weather and Traffic Insights

A full view of the weather and traffic interface can be found in Figure 2. In the weather and traffic interface, the most important charts are the red one, which shows the incident count per 1000 occurrences of a specific weather condition or the count per 100000 occurrences of a specific traffic condition, and the green one, which shows the response time of a specific condition. While the blue plot, the total incident count under a specific condition, can be useful, it generally skews the data and reveals relatively similar patterns to climate or traffic conditions of Nashville than the emergency responders' behavior.

In the weather section, the first notable trend was in `Humidity`. We see a negative relationship between incident frequency and humidity level. This is probably because a lower humidity would result in a higher chance of fire, which significantly increases the frequency of incidents under the climate.

For `Rain` condition, we can utilize the `No Rain` category as a baseline to compare with, as there are significantly more data on not rainy weather than data on rainy weather. We can observe that there are almost

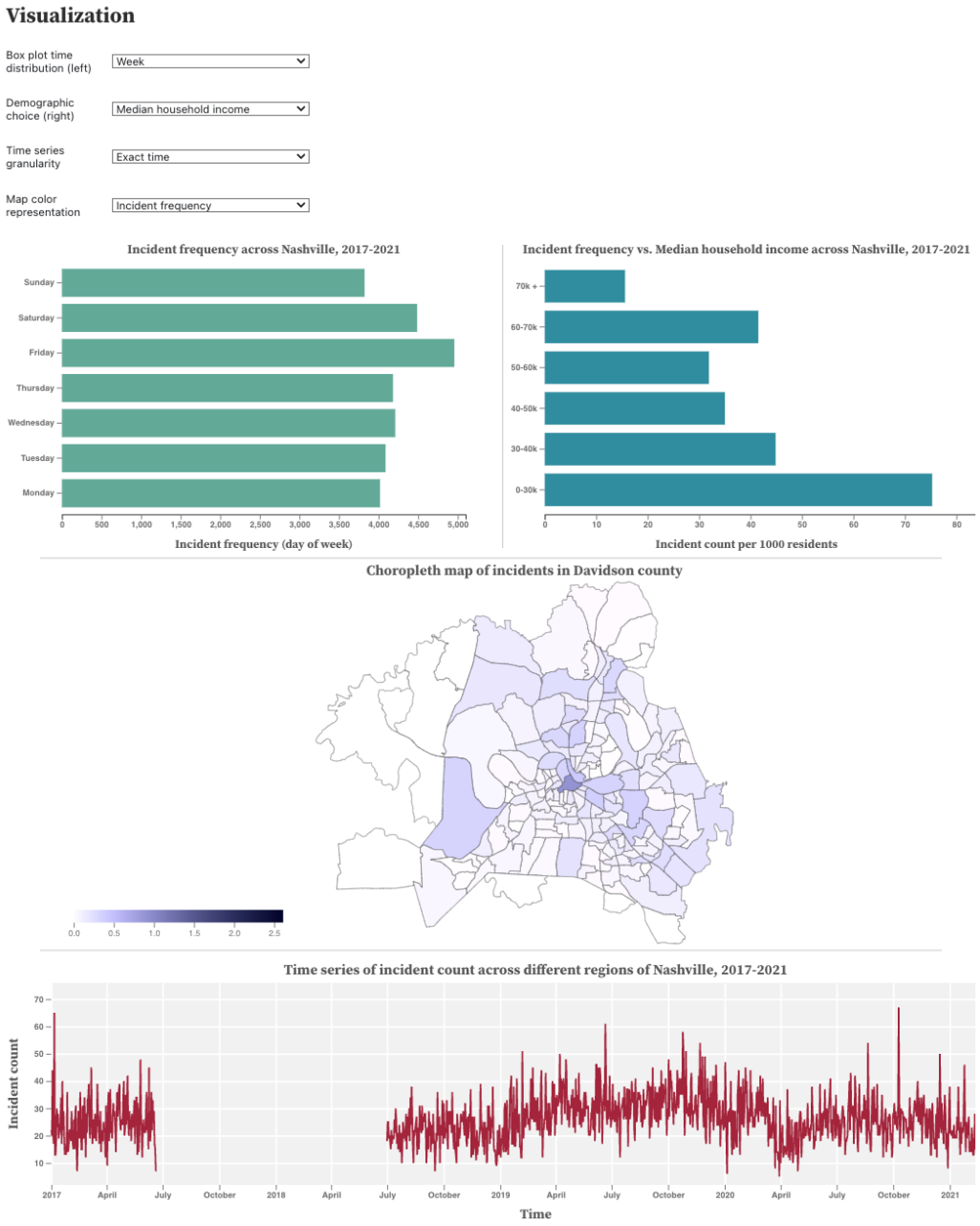

Figure 1: Demographic Interface; **First Row:** Filter Selection Dropdown lists; **Second Row Left:** Incident Count/Frequency/Avg response time in week/month controlled by filters; **Second Row Right:** Incident Count/Frequency/Avg response time vs. Demographic fields controlled by filters; **Third Row:** Map of the census tracts around Davidson county, the color of each region is visualized data controlled by the filters; **Last Row:** Interactable Time vs. Incidents Count graph that allows the user to brush the time window displayed on the map.

universally more frequent incidents when there is some degree of rain compared with no rain. Looking at the response time chart, we can see that the response time on rainy days is also almost universally slower than on no-rain days. Such a pattern indicates that rainfall has a consistent impact on both the frequency of incidents and the speed of emergency response.

We concluded that there is not enough data to form a definitive conclusion on `Snow` condition, as there are only 71 incidents on snowy days among 29770 total incidents count. Therefore, any trend we observe may very likely be from noise than something concrete.

As for `Temperature`, we made two notable observations. First, the incident frequency almost doubled when the temperature reached 30 degrees Celsius. This correlation is similar to the low humidity condition where a high temperature is more likely to cause fire-related incidents. Another `Temperature` related observation is that the response time almost doubled when the temperature dropped below -10 degrees Celcius. While there are only 8 incident cases to support this, we think that it is logical to conclude that with incredibly low temperatures comes a very slow response time.

There is a steady positive correlation between `Wind` and incident frequency. Not only is vehicle control more difficult in stronger wind, but fire also spreads faster, justifying the positive correlation.

In the traffic section, `Average Speed` and `Current Speed` give similar trends that with faster speed comes more frequent incidents. This is self-explanatory as higher speed gives the driver less time to make decisions, leading to more traffic-related accidents. A similar story occurs in `Congestion`. With low to no congestion comes the highest incidents frequency since cars are able to go at a faster speed. Interestingly, there is also an upward trend when there is high congestion (> 0.2). This could be caused by drivers needing to make frequent breaks during extreme congestion, which could lead to mistakes in the process.

4 Visualization Display

In this section, we will exhibit some examples of manipulating the two interfaces and gaining insights from the display.

4.1 Demographics Interface

A typical interaction on this interface is to hover over a specific tract on the map, and the interface will show the incident statistics, demographic information census, and basic information of this tract. The color of each census tract can represent incident count, incident frequency, or average incident response time in this tract, depending on the choice in the drop-down list `Map color representation`. The bar plots on the left-top panel can display the incident distribution on different months or weekdays, depending on the choice of the drop-down list `Box plot time distribution`. When the user hovers over a specific tract on the map, the incident distribution on months or weekdays for this specific tract will also appear in the top-left bar plot, shown in Figure 3. Although the individual bars have a different scale compared to the original scale of the bar plot, it can still reveal insights into the incident distribution of different tracts. The distributions for frequency and average response time are also available. If we set the drop-down list `Demographic choice` to `Ethnicity`, the interface will also display the demographic information for each census tract the user hover over on the map, illustrated in Figure 4.

Another feasible interaction is to brush an arbitrary period on the time-series plot in this interface. Figure 5 shows a selection of the entire year 2019 on the time-series plot of daily incident counts. Upon brushing, the color of each tract in the map above will change simultaneously to reveal the incident distribution within this specific period, along with the numeric value displayed aside. If we set the drop-down list `Time series granularity` from `Exact time` to `Month`, the time-series plot will then display the monthly trend of the incident counts across the entire period, as illustrated in Figure 6. The brushing interaction is also available for this time-series plot, and distributions of incident frequency and average response time can also be displayed on the map when brushing the time-series plot.

4.2 Weather and Traffic Interface

This interface presents 24 bar plots to show the relationships between weather/traffic conditions and incident distributions. The user can choose from five weather conditions from a drop-down list to select one and display three bar plots about the incident counts, frequency, and average response time, as illustrated in

Figure 2: Weather and Traffic Interface; **Top half:** Weather condition; **Bottom half:** Traffic condition; **Blue:** Incident count vs. weather/traffic condition; **Red:** Incident count per 1000 occurrences of weather or per 100000 occurrences of traffic condition vs. weather/traffic condition; **Green:** Response time vs. weather/traffic condition.

Figure 7. Similarly, another drop-down list for traffic conditions is available as well, where the user can pick one condition and visualize impacts from this traffic condition, as illustrated in Figure 8.

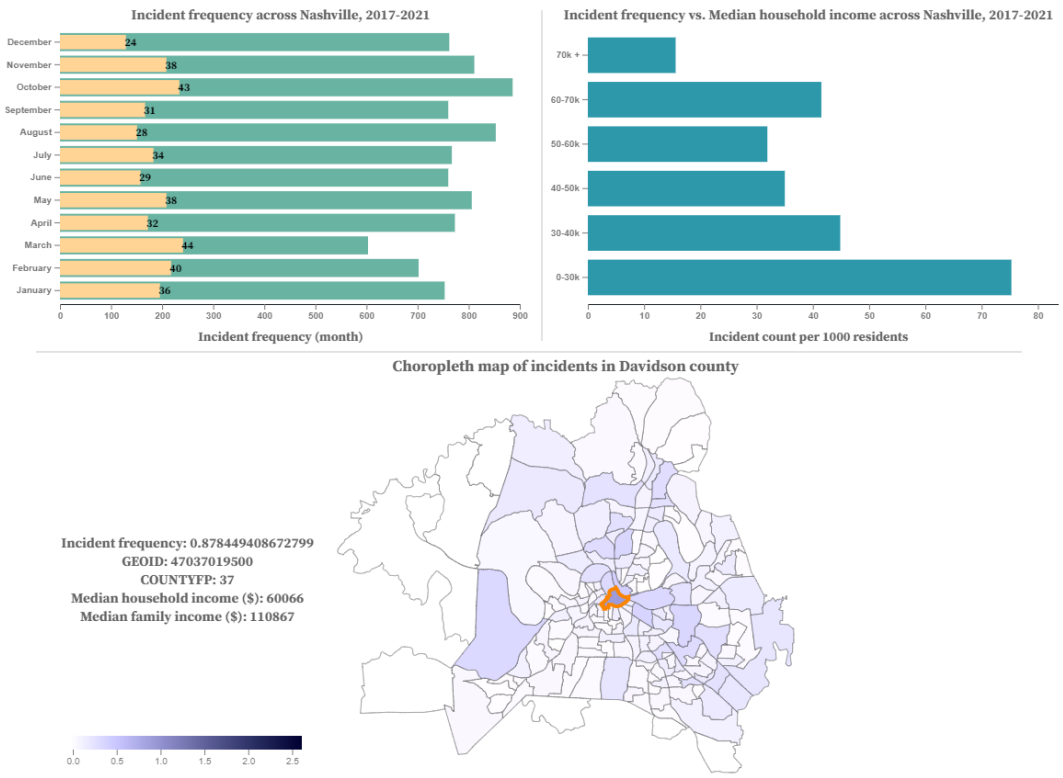

Figure 3: Demographic Interface; Refer to Figure 1 for layout explanation.

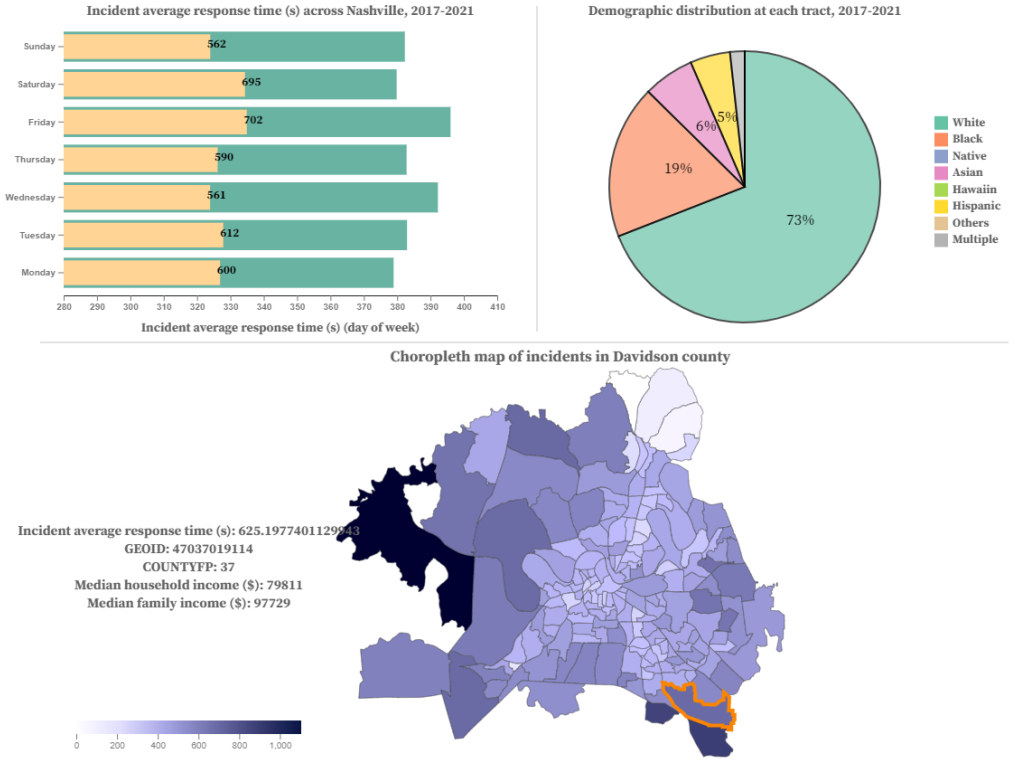

Figure 4: Demographic Interface; Refer to Figure 1 for layout explanation.

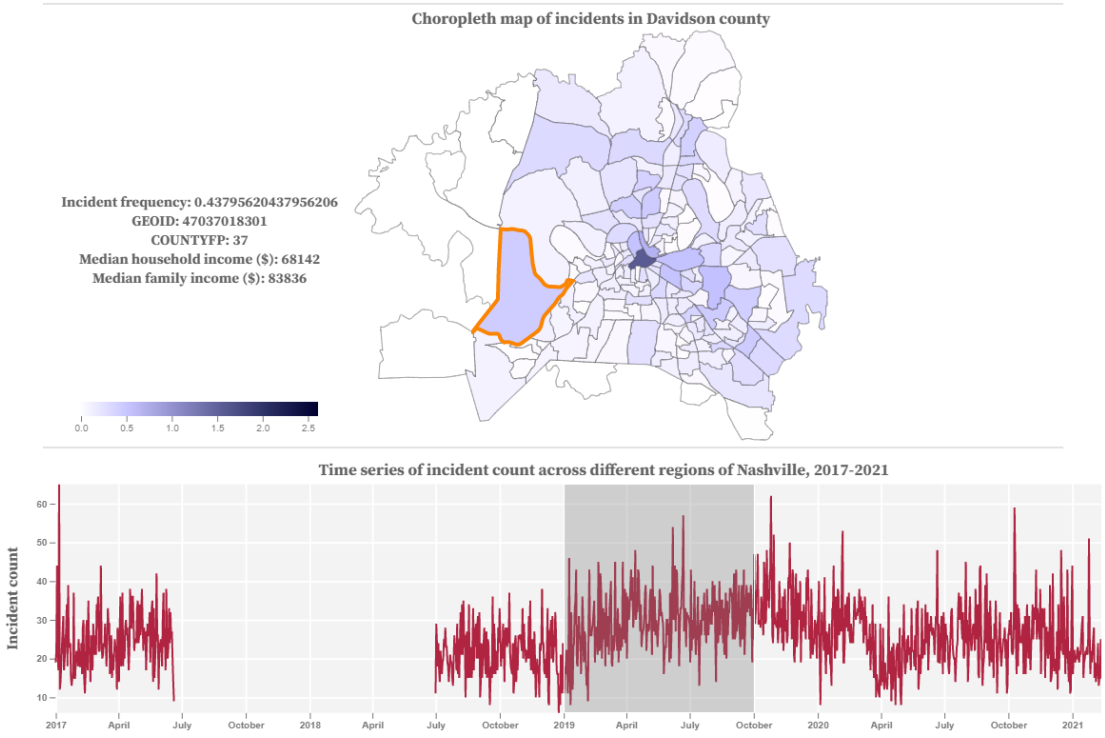

Figure 5: Demographic Interface; Refer to Figure 1 for layout explanation.

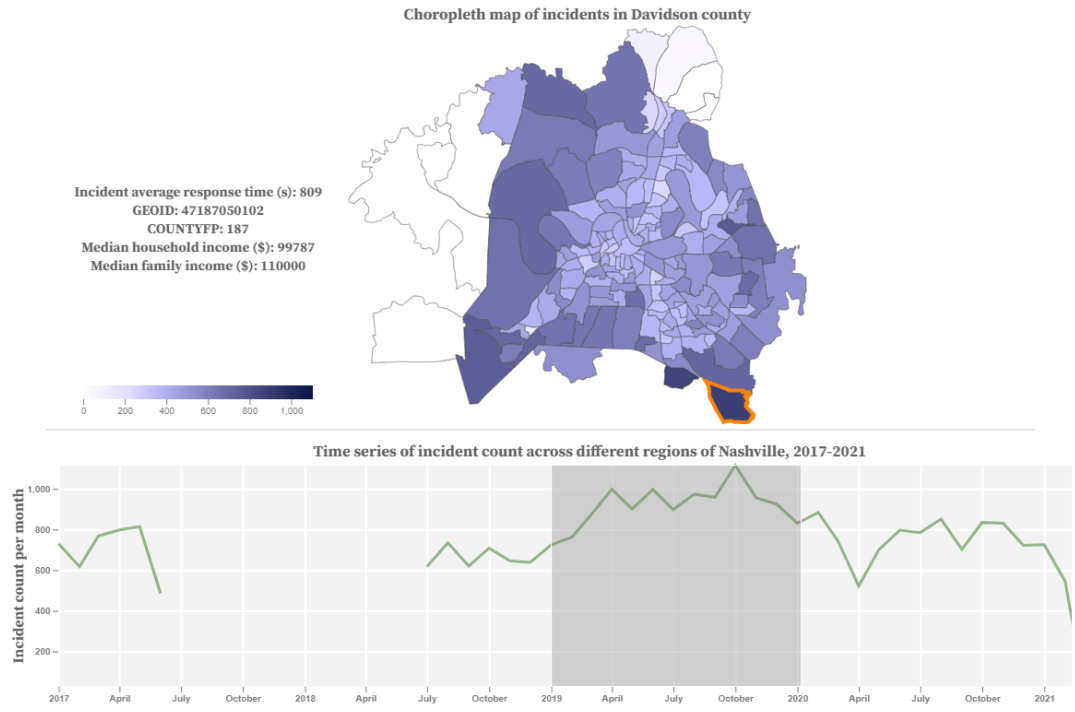

Figure 6: Demographic Interface; Refer to Figure 1 for layout explanation.

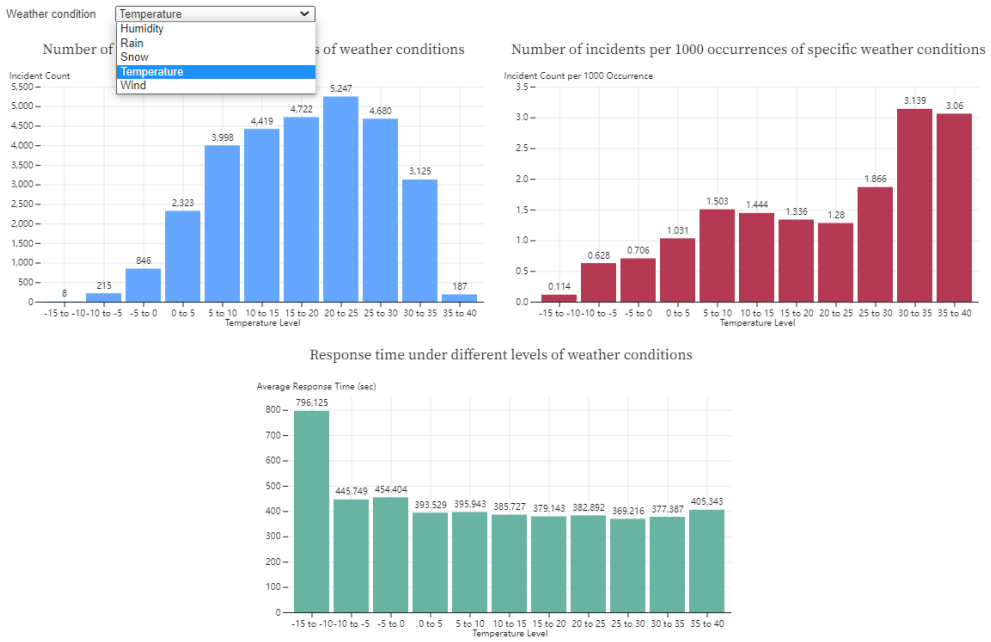

Figure 7: Weather and Traffic Interface; Refer to Figure 2 for layout explanation.

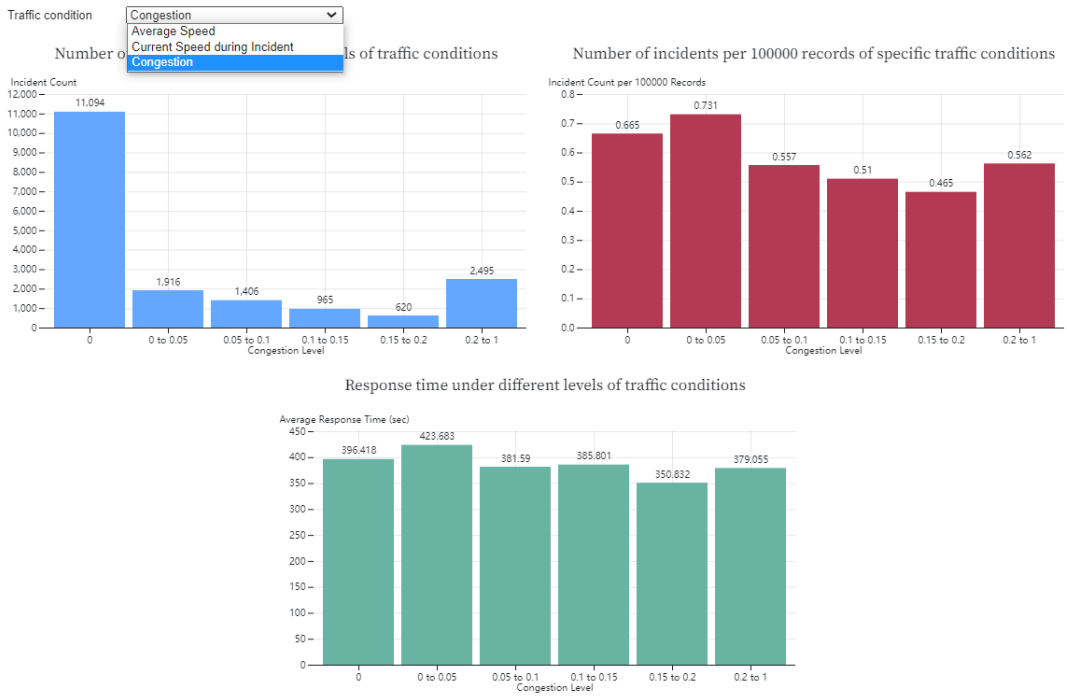

Figure 8: Weather and Traffic Interface; Refer to Figure 2 for layout explanation.

5 Conclusion

In this study, we aimed to understand the relationship between emergency incidents, demographics, weather, and traffic in Davidson County. By developing interactive visualization interfaces, we were able to explore various trends and correlations between these factors. We found that emergency response times are longer during winter months, and incidents are more frequent in lower-income areas. Interestingly, response times seemed to be faster in these lower-income regions, possibly due to a higher presence of emergency personnel.

Weather conditions, such as humidity, rain, and temperature, also play a significant role in the frequency of incidents and emergency response times. We observed that lower humidity and higher temperatures result in a higher frequency of incidents, potentially due to increased fire risks. Rainy days exhibited a higher frequency of incidents and slower response times compared to non-rainy days. Additionally, high wind speeds and traffic congestion were positively correlated with incident frequency.

These insights can help emergency management authorities understand the factors affecting incident occurrence and response times. By identifying patterns and correlations, authorities can prioritize resource allocation, staffing, and equipment distribution to effectively mitigate emergency situations and improve response times. Future work could expand on these findings by incorporating additional variables, refining the granularity of the data, or incorporating real-time data feeds to support more dynamic decision-making.

References

[1] Senarath, Y., Mukhopadhyay, A., Vazirizade, S. M., Purohit, H., Nannapaneni, S., and Dubey, A. Practitioner-centric approach for early incident detection using crowdsourced data for emergency services, 2021.