Analyzing Fire Incidents in San Francisco

Ananya Varma Mudunuri
Masters in Data Analytics,
San Jose State University

Nikhil Sarma Gudur
Masters in Data Analytics,
San Jose State University

Sai Sahithi Bethapudi
Masters in Data Analytics,
San Jose State University

Sri Mounika Jammalamadaka
Masters in Data Analytics,
San Jose State University

***Abstract*— This project focuses on predicting and analyzing fire incidents in San Francisco to enhance public safety and optimize resource deployment. Using historical and live non-medical fire incident data from the San Francisco Fire Department, we aim to identify patterns and trends that can improve response strategies. The dataset includes comprehensive details such as incident type, city, property use, and responder actions, offering insights into high-risk areas and times for fire occurrences. Our approach involves data preprocessing, feature engineering, and predictive modeling to inform resource allocation and enable proactive fire prevention measures. By leveraging data-driven insights, this project aspires to contribute to more effective fire response and public safety initiatives. The goal is to cluster those areas by the types of incidents each commonly experience, and also know when-at what hours of the day-these incidents are most likely to occur. This will let you know about regions with similar incident patterns and their peak times, and that is very valuable for resource allocations, prevention, and response planning.**

***Keywords*— Fire Incident Prediction, Public Safety, Resource Allocation, Data-Driven Insights, Predictive Analytics, Fire Risk Analysis, San Francisco Fire Department, Feature Engineering, Proactive Fire Prevention, Emergency Response**

I. Introduction

Fire outbreaks and their effects take a considerable toll on public safety and resource management. For instance, the San Francisco Fire Department is always stretched in responding quickly and efficiently to all kinds of non-medical fire outbreaks, from structural fires to wildfires, which may be of great cost in properties as well as extremely hazardous. This project is aimed at historic fire incident analysis relative to fire incidents handled by the San Francisco Fire Department, in search of patterns and trends, and has been filtered down to non-medical events.

Incident location, fire kind, property loss, and response actions will be analyzed in this project using a dataset of over 678,000 records. Advanced data analysis with predictive modeling is used, searching for actionable insights in order to achieve better resource utilization, response times. The predictive model developed in the course of this project is intended to support the San Francisco Fire Department in proactive strategic decisions on fire prevention toward a safer community.

II. Related Work

Some of the research papers use machine learning models for the purpose of prediction and classification of fire risks; these have their own strengths. For example, Ahn et al.[1] presented a fire risk prediction model using the stacking ensemble methodology that utilized machine learning algorithms such as decision trees, neural networks, and support vector machines. In this respect, one model developed by a dataset of the characteristics of buildings, lands, and demographic structure succeeded in categorizing fire risks into five categories and found that buildings in the highest category of risk-while only 22% of the sample-comprised 54% of the cases of fire, which is a very good example of practical value for fire prevention and assessment of insurance.

The researched work in this area includes that by Coughlan et al. [2], in which the authors developed a fire ignition prediction due to lightning in West Australia using some machine learning models, such as decision trees, AdaBoost, Random Forest, among others; these gave an accuracy of 78%. This will be something worth looking into, basing on how it applies to the fire

management that links forecasted lightning with the probability of ignition, especially in areas with a high probability of wildfires.

For instance, Seo et al. [3] did some fire hazard analyses in terms of property loss in Seoul, South Korea. They then adapted the use of machine learning algorithms through the Random Forest algorithm. The results were an accuracy of 83% in forecasting fire damage. The identification of major risk factors that might have contributed to the outbreak of fire included those conditions associated with residential apartment facilities. It thus served the model's usefulness in preventing urban fires.

Choi and Jun [4] developed fire risk assessment models using logistic regression, deep neural networks, and optimized risk indexing. They further extended these predictions in terms of building and factory fires and provided more enlightening fire insurance risk assessments, by incorporating fire data from the Fire Protection Association in Korea.

Another urban fire prediction study by Shi et al.[5] applied XGBoost with both historical and real-time environmental and structural data to achieve the best model performance in terms of accuracy. The model is also integrated with the City in a Box™ platform for Shenzhen to extend safety to citizens through possible early warnings for urban fire disasters in highly risky regions.

Yin et al. [6], considering some specific factors such as building age and population density, adapted the Bayesian network model in fire risks assessment on campuses. The paper used fault tree analysis showing causal relationship in fire occurrence and presents an effectiveness of this model in planning campus fire prevention.

Seetharaman et al.[7] proposed a real-time fire detection system with temperature, smoke, and flame sensors whose signals will be processed through a Raspberry Pi. This kind of system, on detecting fire, would trigger an alert to nearby emergency responders via GSM for immediate action to reduce further damages.

III. Data Preprocessing

The original dataset is extracted from the San Francisco government website, which has about 682,000 records , one for each incident across 66 columns of data that describe each incident in great detail. Dates and times of fire incidents, address locations of incidents, response times, and fire incident types are just a few of the variables captured in this dataset. More specific data that can be found within this dataset includes, but is not limited to, neighborhoods, zip codes, and fire department response zones.

A. *Dropping of Irrelevant columns*

This broad dataset has great value in analyzing the trend of public safety and fire responses, finding patterns related to frequency and distribution, and studying how factors such as location and time affect efficiency in responding to emergencies.Working at this scale and level of detail, there are quite a few challenges with this dataset when it comes to extensive data cleaning.Initially we check for missing values in each column and dropped the columns that had more than 80% of missing values. Columns like Estimated property loss , action taken secondary , etc have been dropped. For the remaining columns like mutual aid , estimated content loss, etc which are irrelevant to the project and had a lot of ‘unknown’ values are dropped as well.

B. *Handling missing values*

Subsequently , for numeric columns, missing values are often filled with the median of that column.For categorical columns, the mode (most frequent value) is used to fill in missing entries. For example, missing values in the point column are filled with the most common point within each neighborhood_district. Missing City and neighborhood_district values are filled based on a zipcode-to-city and zipcode-to-neighborhood mapping. Missing Supervisor District entries are populated using a similar zipcode mapping.Some columns are filled with “Unknown” where missing values indicate that no reliable replacement can be

derived (e.g., First Unit On Scene or Property Use).For critical columns like timestamps (Alarm DtTm, Arrival DtTm, and Close DtTm), rows with missing values are ultimately dropped.

C. *Feature Engineering*

New columns were created to provide critical metrics and enhance the consistency of the dataset, allowing for more effective analysis of emergency response patterns. For instance, a ResponseDelayedinMins column was added to calculate the delay in response time by measuring the difference between the Alarm DtTm and Arrival DtTm timestamps.There is a 'Property Use' column which was a combination of code and property description which is divided into Code and Description column , subsequently for columns like 'Primary situation' which have similar input structure are dealt similarly. To find if there is any correlation between the neighbourhood population and the number of incidents , different dataset like the population of the neighbourhood is combined with the main dataset using the neighbourhood column and the year to map.
Moreover , new columns for year and month are created by extracting from the Incident Date column.

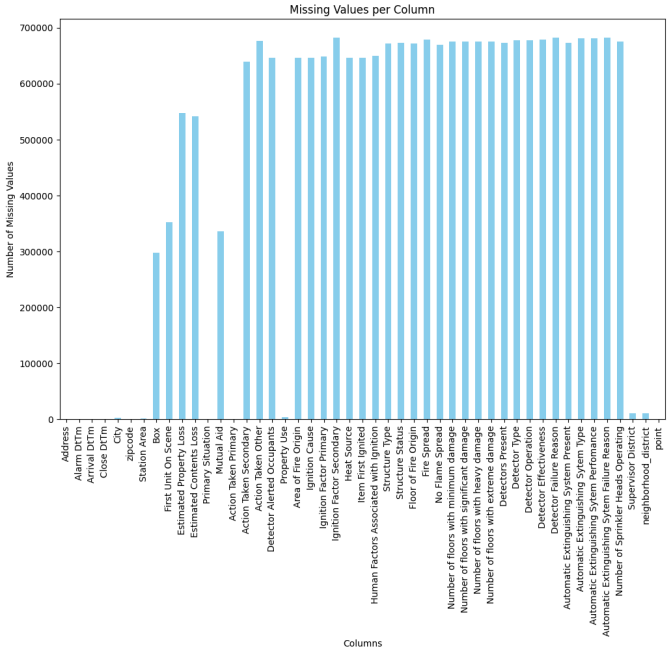

Fig. 1: *Missing Values per Column*

IV. Data Analysis

Data analysis involves a series of visualizations that depict patterns, trends, and other useful insights from the fire incident data. Graphs take on a different perspective depending on various other aspects of the dataset; hence, they are useful in understanding incident frequency, property type distribution, and temporal trends.

A. General Trends

The general trend to perceive from the annual incident data on fire incidents from 2003 up to 2024 would be that starting from a high level in 2003, it gradually went down and reached its lowest ever in 2006. Starting with this period, incidents gradually began to rise, starting in 2007 until 2012, peaking to a point that strongly suggests either a time period of high fire activities or, alternatively, increased reporting.

From 2014, a sharp drop ensues in incidents, after which incident numbers once again began to rise until they peaked in 2018 and fell again with a notable decline in 2021. Starting from the past couple of years, there has been a significant spike in 2023, showing an unusually high number of fire events within that time frame. Immediately after that, however, is an apparent sudden decrease in 2024, suggesting improvements during that time period due to better means of control or other mitigants, which would limit the number of active fires.

These trends reveal fire incident rates that have fluctuated with time; these may be because of changes in environmental factors, changes in policies on fire prevention, or changes in population density and infrastructure. As a matter of fact, the data justifies continued cause analysis to determine what is affecting these fluctuations and to devise strategic policies on future fire prevention and responses.

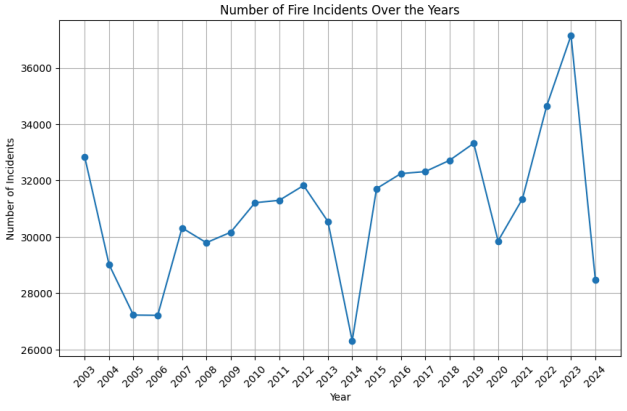

Fig. 2: Number of Fire Incidents Over the Years

Fig 3 graph depicts the trends in firefighter and civilian deaths and injuries from 2003 through 2024. The civilian casualties continue to be, by a long shot, the highest among the counts depicted there in red, with noticeable spikes in variability around the years 2005 and 2013, which represent years that experienced unusually high incidents of civilian injury. Fire injuries, shown in orange, have some fluctuation, and the number really spiked in the years 2009 and 2012. The green shows civilian fatalities, and in blue shows fire fatalities that are relatively low and stable throughout; this reflects the fact that fatalities occur quite rarely compared to injuries. This data shows that civilian casualties from fire incidents are rare, while civilian injuries recur, hence the need for safety and a response mechanism to make sure civilians are protected during fire incidents.

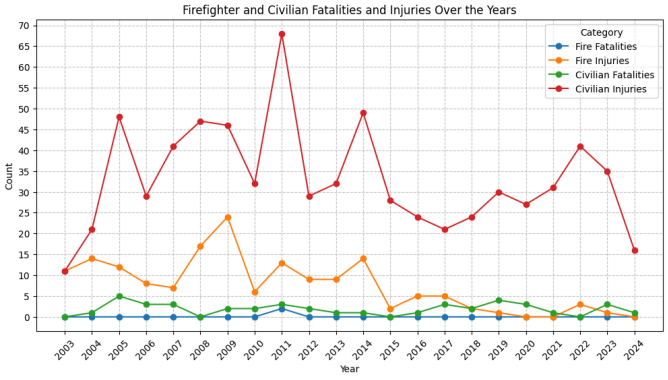

Fig. 3: *Firefighter and Civilian Fatalities and Injuries Over the Years*

The heatmap in figure 4 represents the relationships of fire and civilian fatalities, suppression units, and fire and civilian injuries. Intensities of color and values apply to strengths of these correlations. There would appear to be a moderate positive correlation of 0.38 between fire injuries and civilian injuries; that is, incidents that have fire injuries often have civilian injuries. It is also found that the correlation between fire fatalities and civilian fatalities is 0.27, which was a good indicator suggesting incidents involving fatalities would involve both firefighters and civilians.

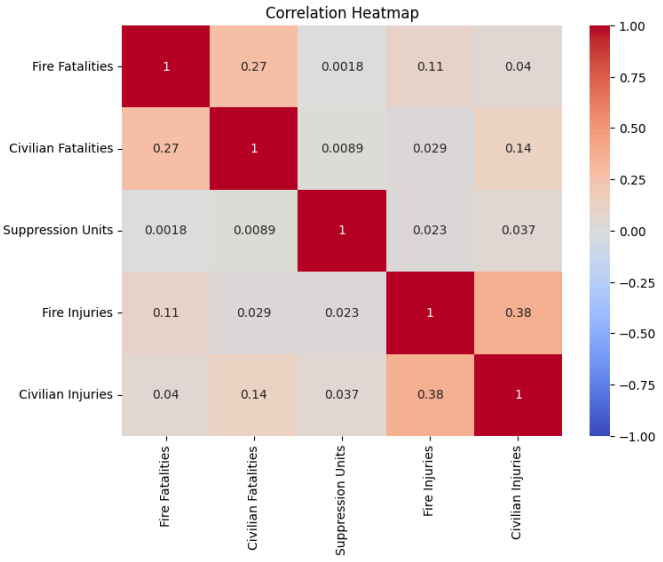

Fig. 4: Relationships of fire and civilian fatalities

B. Unique Trend Analysis

Distribution underlines that residential areas, multifamily dwellings, are within the highest priority of incident rates, probably because of a high population density factor and the presence of different flammable materials. The remaining categories, with a relation to hotels, hospitals, and schools, show lower incident numbers, but highly relevant due to their exposure for specific preventive measures:. This gives a relative standpoint into the different kinds of property targets that are of priority to concentrate fire prevention and safety policy measures, as residential premises and street categories require most attention.

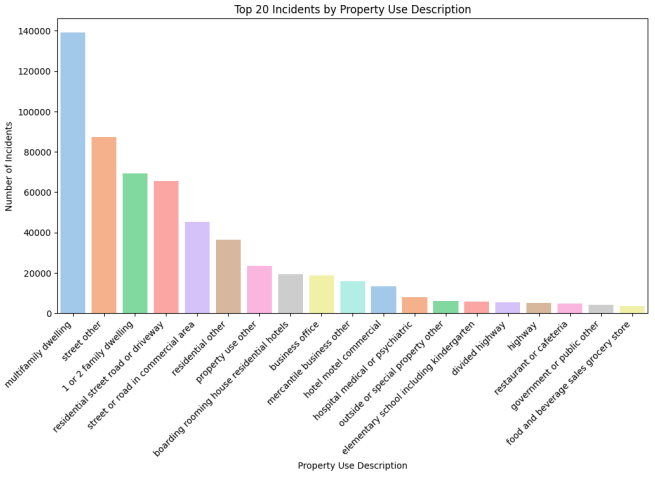

Fig. 5: *Top 20 Incidents by property Use Description*

The Figure 6 bar chart shows the top 10 property use descriptions with the highest average response delay in minutes. The properties described as "aircraft loading area" have an extremely high and multiples higher response delay, averaging above 35 minutes, compared to the remainder of property use descriptions. Other locations from these datasets that include "airport passenger terminal," "bridge trestle," and "memorial structure including monuments," each have very much lower but high average response delays around the 10-minute mark. The above distribution does seem to hint that incidents taking place in places that are either specialized or remote, like airports, bridges, and monuments, are subjected to prolonged response times. Maybe it can be reasoned out that access to these places is highly restricted and requires some special way of responding to them. This was quite an interesting insight, to see where the logistical challenges in emergency response may lie and thus provide possible opportunities for access improvement or response planning in those high delay places.

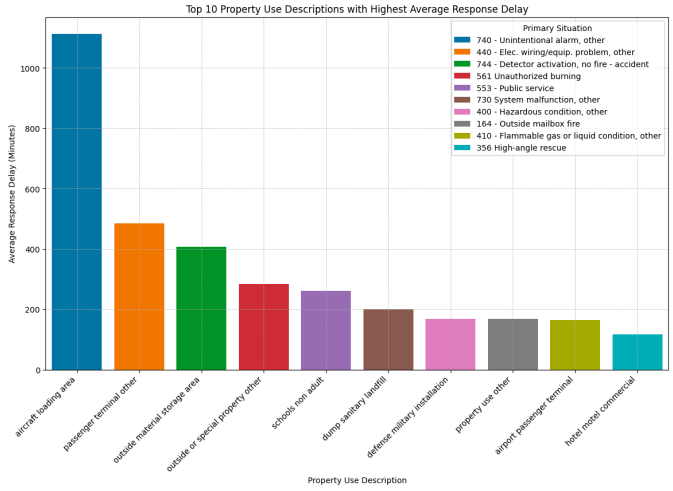

Fig. 6: *Top ten property Use with highest average response delay*

The bar chart below represents the ten types of incidents with the highest severity based on the primary situation, fire fatalities and injuries, civilian fatalities, and civilian injuries. The "Building Fire" stands out for having the most outstanding number, in particular for civilian injuries, because it is represented using the color red. On the other hand, other incident types such as "Passenger Vehicle Fire," "Fire, Other," and "Special Outside Fire, Other" apply relatively low counts to all severity metrics, with just very negligible civilian and firefighter injuries and fatalities. This trend is indicative that building fires are highly hazardous to the safety of civilians compared with incidents of other types, hence specific safety measures and preventive actions indoors are called for.

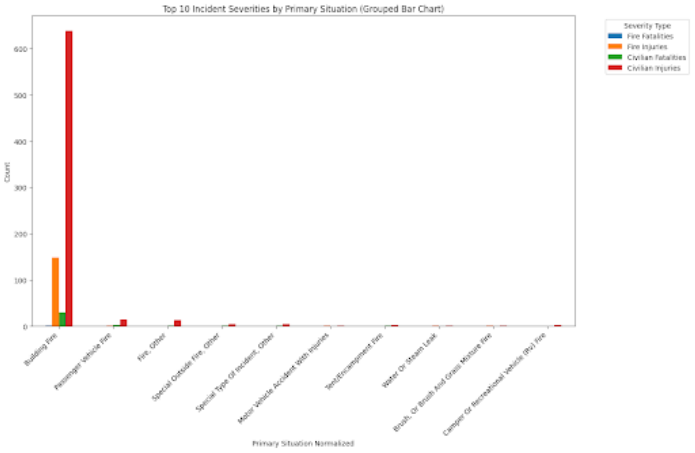

Fig. 7: Top ten incident severities by primary situation

Fire incident distribution by day of the week bar chart illustrates fire incidents, which occurred during the days of the week. Although the number remained quite steady, there is a slight difference on the weekdays. Indeed, incidents occur rather evenly from Monday through Saturday. The most number of incidents fell on Friday, while on Sunday, the count fell a bit below compared to every other day, denoting a minor decline of fire-related incidents on Sundays. This uniformity of representation across days serves as a strong indication that incidents of fire do not depend heavily on the day of the week and that other factors surrounding the time of day, property type, or an environmental condition may be considered as much as, if not more than, the exact day of the week when it comes to incident frequencies.

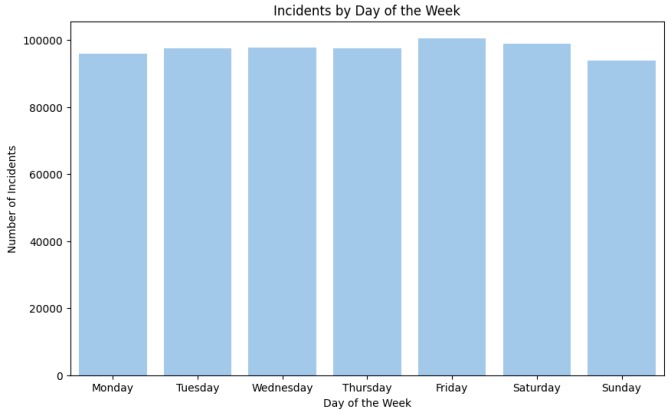

Fig. 8: *Fire Incident Distribution by Day of the Week*

The bar chart in Figure 9 shows the distribution of fire incidents across each hour of the day. Incident counts are generally low during early morning hours, with the lowest count coming at approximately 4:00 AM. Incident frequency begins to sharply increase from 7:00 AM onwards and peaks between 12:00 PM and 5:00 PM, during which the counts keep consistently high. Due to this trend, one can note that daytime incidents of fire are more frequent, probably because of better human exposure and operation of various facilities. It gradually decreases after 6:00 PM into late evening hours, with the least counts past 10:00 PM. This pattern can be seen to show that most fire incidents occur during high activity hours, implying that daytime and early evening are the most incident-prone of the times of day.

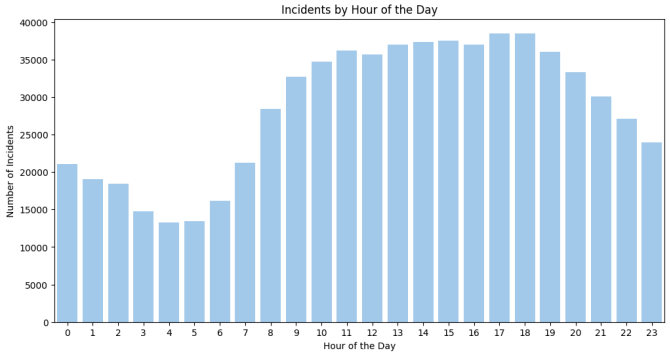

Fig. 9: *Fire Incident Distribution by hour of the day*

Figure 10 depicts the incident count of fires across time in a day of the week. All days hold their peak incident counts in "Early Morning" and hence are the most active incident times. Incident counts are consistently falling to mid-range values throughout the week in both "Evening" and "Morning", with very negligible fluctuation across days. This would be in contrast to the selections of "Mid Night", a period of time when incident counts are at their lowest, reflecting the relatively low incidence of fire incidents in late nightly hours. The daily distribution here across the intervals furthers the fact that time of day has a more increasing effect on incident frequency than days of the week and points out main periods for resource allocation in emergency response efforts.

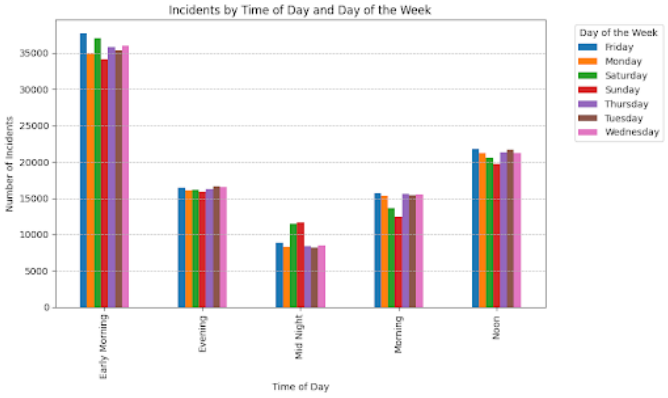

Fig. 10: *Incident by time of day and day of the week*

This map represents the number of populations in neighborhoods and the number of fire incidents taking place within each area. The neighborhoods were shaded in various shades of green to represent the range, light green for low populated neighborhoods within 0 to 1,000 people to dark green for populated neighborhoods with 20,000 to 50,000 people. The orange circles show the relative size of the number of fire incidents: large radiuses imply high incident counts.

Indeed, from this map alone it certainly becomes obvious that incidents are not strictly proportional to the population, since a number of low-population areas represent high incidents-most probably due to infrastructure, density of high-risk properties, or commercial activity. On the other hand, a number of highly populated neighbourhoods have only a moderate incident rate. This spatial distribution probably defines high-risk neighbourhoods in which targeted fire prevention should be identified at levels irrespective of population density.

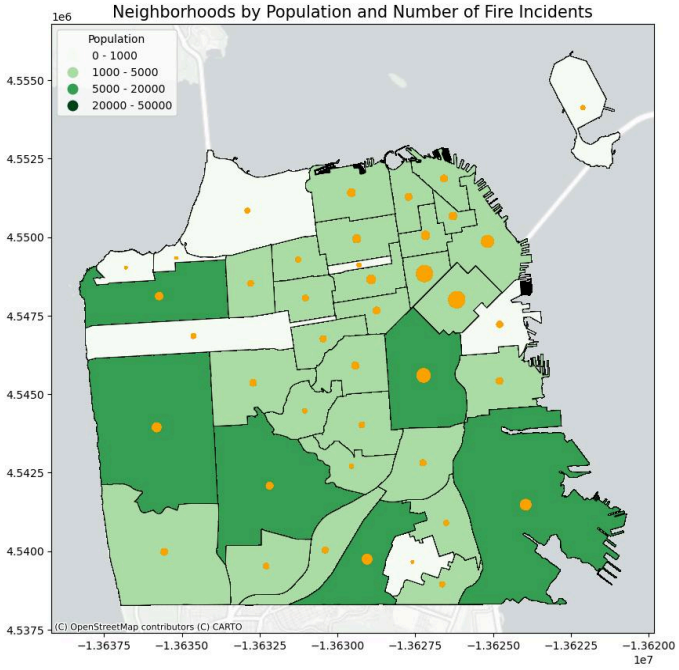

Fig. 11: *Neighborhoods by population and number of fire incidents*

Figure 12 shows the top ten neighborhood districts by fire incident counts. Ranked first, nearly 80,000 incidents were counted across this district ; in order, second and third place are South of Market and Mission; all of these districts reported a high count of incidents. Also, districts such as Financial District/South Beach, Bayview Hunters Point, Sunset/Parkside, reported sizable numbers of incidents falling between 30,000 to approximately 50,000. Relatively fewer incidents occur in neighborhoods like Excelsior, Western Addition, Nob Hill, and Pacific Heights, yet still make the top 10, showing significant fire activity. This distribution underlines the fact that incidents would be more frequent in urban, heavily populated neighborhoods, possibly because of higher residential density, commercial activities, or socio-economic factors, and that attention would have to be paid to fire prevention and response strategies in these areas.

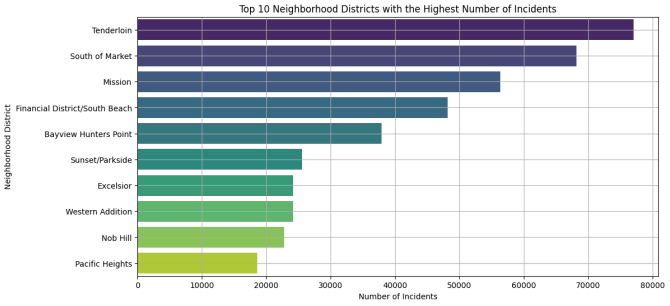

Fig. 12: *Top 10 neighborhood districts sorted by fire incident counts*

V. Community Contribution

In the project, this provides value to the community on safety issues regarding fire management and prevention through actionable insight. Identification of risky location and time identification enables the stakeholders to take proper preventive action against the occurrence of fire incidents effectively. Through the analysis of the deep historic record of fires and high-fire danger periods, the predictions provided by the developed analytical models will help in ascertaining where emergency resources are most likely to be deployed in protecting the public at large and ensuring response efforts are effective. Such findings will help planners within the city, the fire department's heads, and policy makers to identify areas of the highest rates of incidents coupled with response time and lay down strategies for resource allocation. This evidence-based strategy aligns with the public safety work in San Francisco, bringing in a safer community which will help build long-term resilient urbanism.

VI. Conclusion

The project effectively makes use of data from past fire incidents to predict and provide strategic recommendations for the management of fire risks in San Francisco. Using machine learning models on holistic datasets, we mapped the insights in patterns of fire incidents, located the hotspots, and identified various critical factors affecting response times. The predictive capabilities developed under this project introduce proactive overtones to public safety by using data-driven decisions in resource allocation,

preparedness, and a risk-reduction strategy for the San Francisco Fire Department and other stakeholders.

This project lays a scientific foundation by way of data preprocessing, feature engineering, and analysis on which future fire risk assessment and mitigation strategies can be effectuated. This would greatly reduce property damage and delayed response to fire outbreaks and thus engender community resiliency. Consequently, this project will avail the necessary information to city planners and policy makers, pinpointing specific locations and times of high fire risk that inform targeted preventive efforts and enhanced emergency response strategies. This is part of a bigger urban agenda for safety, sustainability, and efficiency in the use of resources, underlining the rise of data analytics in creating safer and more resilient communities.

Reference

[1] S. Ahn, J. Won, J. Lee, and C. Choi, "Comprehensive Building Fire Risk Prediction Using Machine Learning and Stacking Ensemble Methods," *Fire*, vol. 7, pp. 336–346, Sep. 2024, doi: 10.3390/fire7100336.

[2] R. Coughlan, F. Di Giuseppe, C. Vitolo, C. Barnard, P. Lopez, and M. Drusch, "Using Machine Learning to Predict Fire-Ignition Occurrences from Lightning Forecasts," *Meteorological Applications*, vol. 28, 2021, Art. no. e1973, doi: 10.1002/met.1973.

[3] M. S. Seo, E. E. Castillo-Osorio, and H. H. Yoo, "Fire Risk Prediction Analysis Using Machine Learning Techniques," *Sensors and Materials*, vol. 35, no. 1, pp. 1–12, 2023, doi: 10.18494/SAM.2023.3949.

[4] M.-Y. Choi and S. Jun, "Fire Risk Assessment Models Using Statistical Machine Learning and Optimized Risk Indexing," *Applied Sciences*, vol. 10, pp. 4199–4211, Jun. 2020, doi: 10.3390/app10124199.

[5] X. Shi, Q. Li, Y. Qi, T. Huang, and J. Li, "An Accident Prediction Approach Based on XGBoost," in *Proc. 12th Int. Conf. Intell. Syst. Knowl. Eng.*, Nanjing, China, Dec. 2017, pp. 1829–1834, doi: 10.1109/ISKE.2017.8247128.

[6] K. Yin, Y. Niu, and Y. Jiao, "Analysis of Campus Fire Based on Bayes Network in the Background of Big Data," in *Proc. 2021 Int. Conf. E-Commerce and E-Management (ICECEM)*, Yantai, China, Sep. 2021, pp. 74–81, doi: 10.1109/ICECEM54757.2021.00074.

[7] R. Seetharaman, N. Nivetha, R. R. Sreeja, S. Gowsigan, S. V. Dakshin, and M. Barath, "Analysis of a Real-Time Fire Detection and Intimation System," in *Proc. 5th Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Chennai, India, 2023, pp. 1738–1745, doi: 10.1109/ICSSIT55814.2023.10061106.