Predicting CAHOOTS:

Analyzing Temporal and Climatic Patterns in Call Data

By Michael Saccio

Background

This project investigates the relationship between climate variables and the volume and classification of calls received by the Crisis Assistance Helping Out On The Streets (CAHOOTS) program in Eugene, Oregon. A major focus is analyzing the temporal patterns and effects, examining how CAHOOTS call volumes and classifications vary across different time frames such as time of day, day of week, season, and year. Understanding these temporal dynamics, along with the influence of environmental factors like temperature, humidity, air quality, and seasonal weather events, can provide crucial insights for optimizing CAHOOTS' operations and staffing.

The significance of this research lies in its potential to enhance the efficiency and effectiveness of CAHOOTS' mobile crisis intervention services. By identifying environmental influences and temporal patterns impacting call volumes, CAHOOTS can better anticipate community needs and strategically allocate resources to ensure timely crisis assistance. Moreover, integrating an upcoming weather data API enables predictive modeling to categorize future days as neutral or high-risk, refining emergency preparedness and staffing plans. Analyzing these complex relationships is vital for CAHOOTS to proactively respond to crisis situations in the community.

Data

This project utilizes three primary data sources: CAHOOTS call data, climate data from Visual Crossing, and air quality index (AQI) data from the World AQI Project.

CAHOOTS Call Data:

The CAHOOTS call data consists of detailed information about crisis intervention calls received between 2021-01-01 and 2023-12-31. The dataset includes variables such as call date, time, reason for dispatch, client demographics (age, gender, race), and city location. The data source for CAHOOTS call records has been anonymized to protect client privacy.

Climate Data:

Climate data was obtained from the Visual Crossing weather API, a reliable source used by many large corporations. The dataset includes daily measurements of various climate variables, such as maximum, average, and minimum temperatures, dew points, humidity levels, wind speeds, atmospheric pressure, and precipitation amounts for the Eugene area during the study period, as well as any future period required in making predictions. This data is sourced from local weather stations in and around Eugene, ensuring its relevance to the geographic region under study.

Air Quality Index (AQI) Data:

The air quality index (AQI) data, specifically the PM2.5 (particulate matter) values, were collected from The World Air Quality Index Project for the Eugene region, aligning with the timeframe of the CAHOOTS and climate data. Only the PM2.5 value was used from this dataset to incorporate air quality as a potential factor influencing call volumes.

Data Preprocessing:

Several preprocessing steps were undertaken to integrate and prepare the data for analysis, as documented in the project's README file and the following scripts:

1. `weather.ipynb`: This script fetches weather data from the Visual Crossing weather API for the specified location and time period, saving it as 'upcoming_weather.csv' in the 'raw_data' directory.
2. `data_cleaning.ipynb`: This script cleans and preprocesses the raw data obtained from CAHOOTS and weather databases, merging datasets, handling missing values, converting data types, creating dummy variables, and preparing the data for analysis. The cleaned data is saved as CSV files in the 'data' directory.
3. `preprocessing.ipynb`: This script preprocesses the model data for classification, ensuring the data is in the appropriate format and structure required by the classification models.

The data preprocessing steps were documented in Python scripts and Jupyter notebooks, which can be found in the project repository. A README file is included, explaining the purpose and usage of each script, as well as any dependencies or parameters required for execution.

Methods

This project employed a multi-step experimental approach to analyze the relationship between climate variables and CAHOOTS call patterns, as well as develop a predictive model for high-risk days based on upcoming weather data. The analysis was performed using Python and several data science libraries, including pandas, numpy, matplotlib, scipy, statsmodels, and scikit-learn, with coding help from the ChatGPT 2024 and Claude 2024 generative AI models.

Exploratory Data Analysis (EDA):

- The cleaned dataset was explored using visualizations and statistical analyses to understand patterns and relationships between variables.
- Significant correlations between climate variables, call volumes, and call classifications were identified using methods from the scipy and statsmodels libraries.
- The `analysis.ipynb` script was used to conduct EDA and generate relevant plots and correlation matrices.

Data Preprocessing for Modeling:

- The data was preprocessed and formatted appropriately for input into the classification models using the `preprocessing.ipynb` script.
- Feature scaling and encoding techniques were applied as necessary.

Predictive Modeling:

- A Random Forest classifier from scikit-learn was trained on the preprocessed data to predict high-risk days based on call volume.
- The model utilized features such as temperature, humidity, PM2.5, solar energy, day of the week, and day of the year.
- An Ordinary Least Squares (OLS) regression model was used to generate information on the feature importance and coefficients used in the Random Forest classifier.
- The `classifying.ipynb` script implemented the Random Forest classifier and OLS regression model.

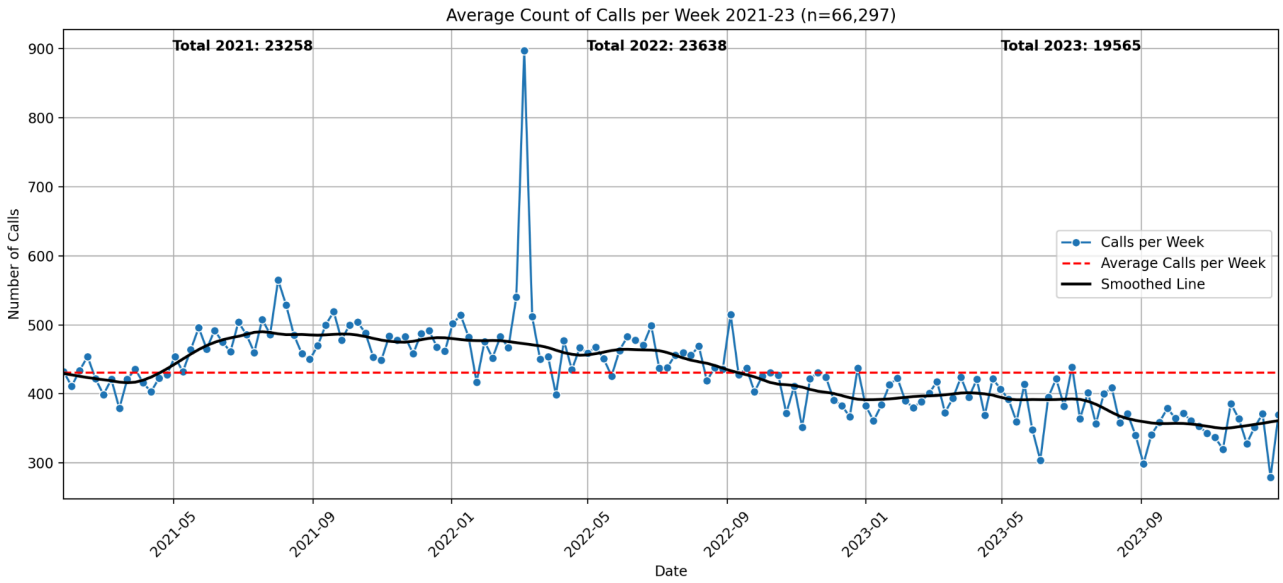
Model Evaluation and Visualization:

- The predictions generated by the Random Forest classifier on the upcoming data were saved as a CSV file (`upcoming_data_with_predictions.csv`) in the data directory.
- The OLS regression weights were stored in a CSV file (`OLS_coefficients.csv`) in the data directory.
- Additional visualizations were created using the `more_plots.ipynb` script to provide further insights into the data and machine learning results.

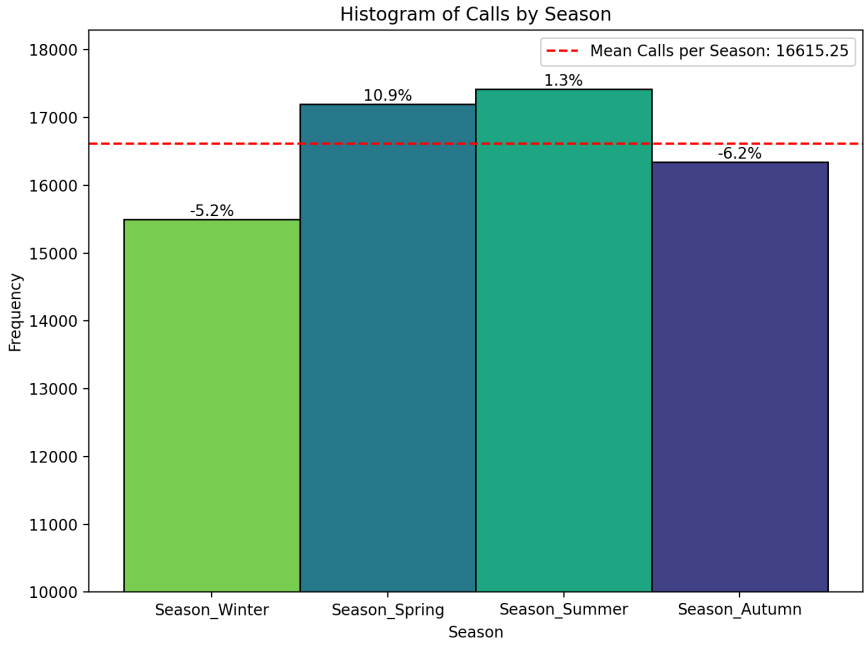
The project repository includes a README file that explains the usage, order of execution, and any dependencies or parameters required for running the scripts.

Results

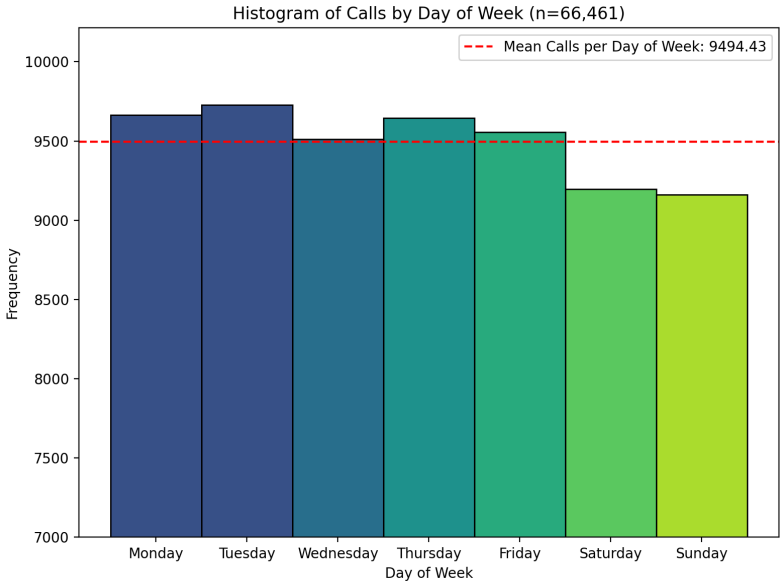
Call Quantity Results:



The line plot above displays the average number of CAHOOTS calls per week and the total number of calls received annually from 2021 to 2023. It shows a consistently high call volume of around 23,000 calls in 2021 and 2022, followed by a noticeable decrease to approximately 19,500 calls in 2023. Interestingly, there is a distinct spike in call volume during late February and early March of 2022, potentially attributable to a surge in COVID-19 cases during that period. This temporary increase in calls lasted for approximately 1-2 weeks before returning to the baseline level.

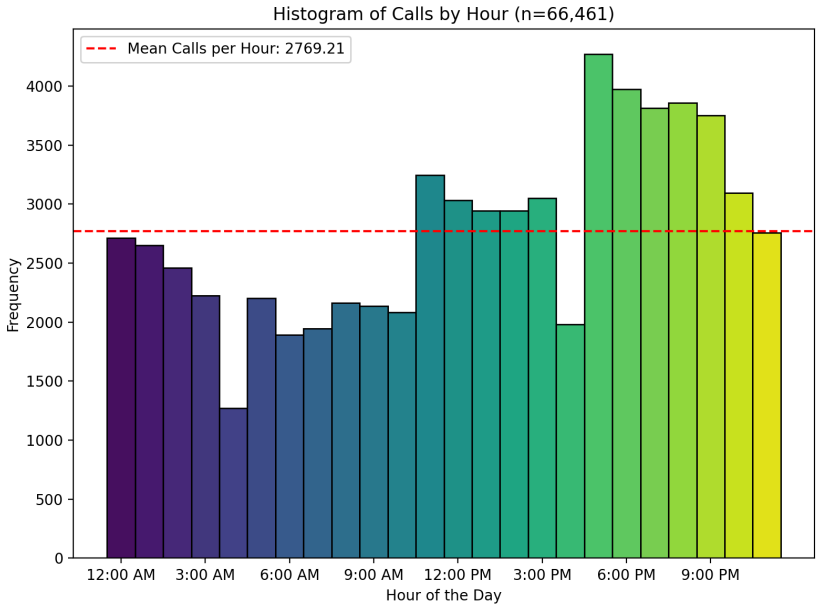


Here's an edited version of the description for the histogram: "The histogram on the right displays the total number of CAHOOTS calls by season, along with the percentage change between seasons. A clear trend emerges, with Spring and Summer seasons receiving the highest call volumes, while Winter and Autumn seasons experience lower call volumes. Notably, the Winter season consistently records the lowest number of calls across the examined period. The percentage changes between seasons highlight the fluctuations in call volumes, with increases observed during the transition from Winter to Spring and from Autumn to Summer, and decreases occurring from Spring to Summer and from Summer to Autumn. Very similar trends were found looking at data by month.



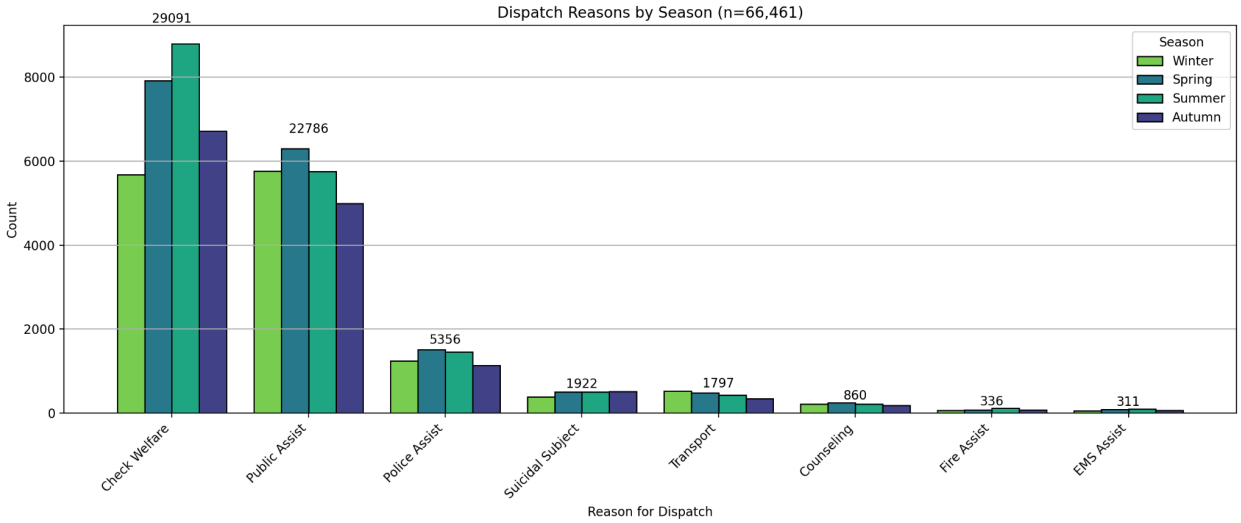
The histogram to the left shows CAHOOTS call counts by day of the week. Weekend days, specifically Saturdays and Sundays, experience significantly lower call volumes compared to weekdays. Conversely, Mondays and Tuesdays consistently record the highest number of calls among all days of the week. This trend suggests that demand for CAHOOTS services is relatively low during the weekend but increases substantially at the start of the workweek.

The histogram on the right depicts the distribution of CAHOOTS call volumes across different hours of the day. Two distinct peaks are observable, with one occurring at 11:00 AM and the other, more pronounced spike occurring at 5:00 PM. The call volume remains elevated during the hours immediately following the 5:00 PM peak, indicating a sustained period of high demand for CAHOOTS services in the late afternoon and evening hours. Conversely, the lowest point in call volume is observed at 4:00 AM, suggesting a diminished need for crisis intervention services during the early morning hours.



Call Classification Results:

The histogram below illustrates the distribution of CAHOOTS call classifications across different seasons, with an annotation representing the total number of calls for each classification. The data reveals that 'Check Welfare,' 'Public Assist,' and 'Police Assist' constitute the majority of calls received by CAHOOTS. For these three predominant classifications, the Spring and Summer seasons exhibit the highest call volumes. Notably, the 'Transport' classification stands out as being most prevalent during the Winter season. This observation aligns with the expectation that transportation needs may increase during winter months due to the presence of ice and snow, which can make mobility more challenging.



The table below presents the correlations between various call classifications and weather features, based on a comprehensive analysis of the cleaned data. The process involved systematically testing the relationships between every possible pair of columns, removing missing values, conducting p-value tests, and calculating the correlations. This rigorous approach identified 1,528 statistically significant relationships, with 381 specifically related to climate variables, all significant at the 1% level.

The most notable trends revealed in the table are the positive correlations between 'Check Welfare' calls and higher temperatures, as well as the negative correlations between 'Check Welfare' and 'Public Assist' calls with colder temperatures and increased snowfall. Interestingly, the 'Transport' classification exhibits a positive correlation with snowfall, once again suggesting a potential increase in transportation needs during snowy conditions.

The table presents a selection of these significant correlations, including the call classification, weather feature, correlation coefficient, p-value, and sample size for each relationship.

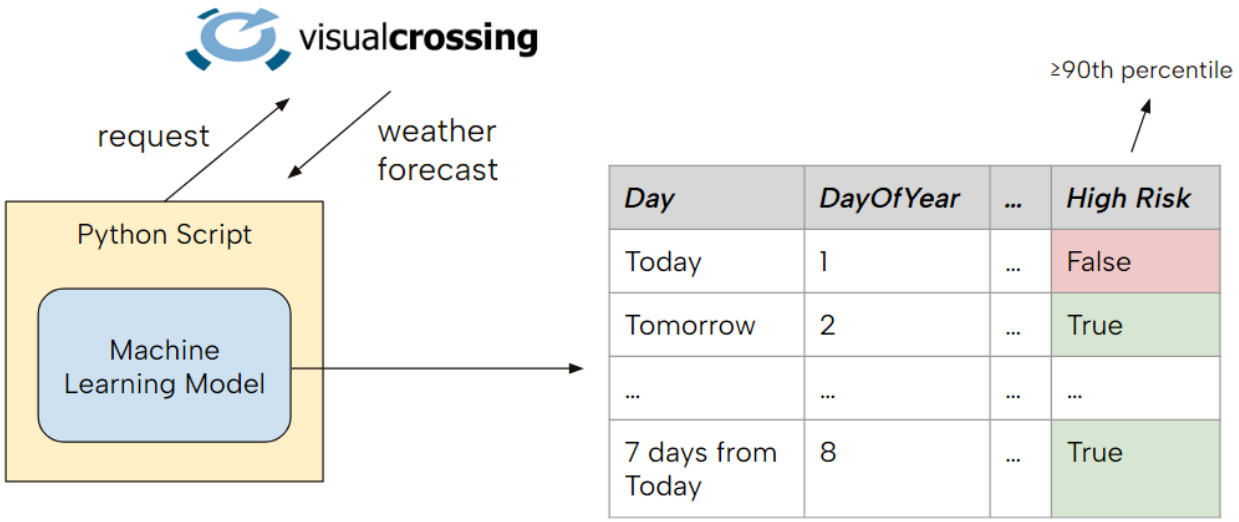
Feature 1	***Feature 2***	***Correlation***	***P-value***	***Sample Size***
Check Welfare	feels_like	0.081884321	2.21E-93	62461
Check Welfare	temp	0.080379998	4.84E-90	62461
Public Assist	feels_like_max	-0.070922517	1.82E-70	62461
Check Welfare	solar_radiation	0.070417611	1.72E-69	62461
Check Welfare	sunset_hour	0.06856535	5.68E-66	62461
Public Assist	temp	-0.068133802	3.64E-65	62461
Transport	snow	0.036399071	9.05E-20	62461

Machine Learning Results:

The flowchart illustrates the process of generating predictions for upcoming days based on weather forecasts and a machine learning model. The initial step involves the weather.ipynb script, which fetches the most recent 7-day weather forecast, including the current day, from a reliable source. Subsequently, the classifying.ipynb script employs a Random Forest classifier to categorize these upcoming days based on various features, such as day of the week, day of the year, and expected climate variables like temperature, humidity, and precipitation.
The Random Forest model classifies each day as either 'High Risk' or 'Normal' based on a predetermined threshold. Specifically, if the predicted call volume for a particular day exceeds or equals the 90th percentile of historical CAHOOTS call volumes, it is classified as 'High Risk.' This categorization allows for proactive planning and resource allocation to address potential increased demand for CAHOOTS services during high-risk periods.

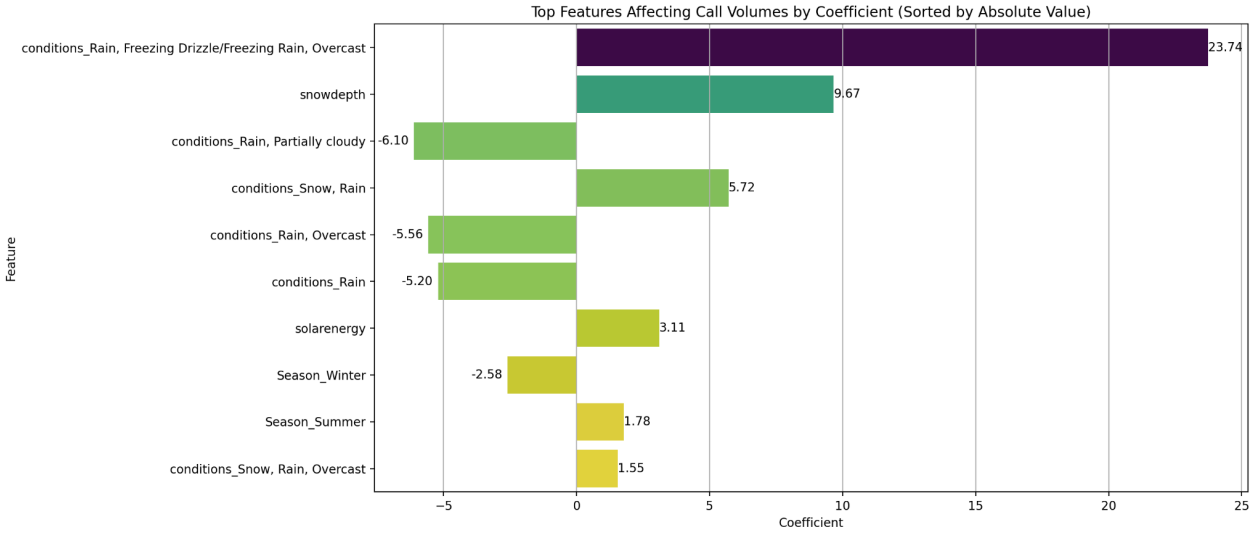
In a professional setting, the modeling and data cleaning processes would be consolidated into a single Python script to enhance usability and streamline the workflow. This integration would provide a more user-friendly and efficient solution for generating predictions and supporting decision-making processes.

In testing, the model correctly categorized 191/217 days as 'High Risk', with an accuracy slightly over 88%.



Day	*DayOfYear*	*...*	*High Risk*
Today	1	...	False
Tomorrow	2	...	True
...	...	...	...
7 days from Today	8	...	True

The plot below highlights the top weather features that affect the number of CAHOOTS calls per day, along with their respective coefficients. Since Random Forest models do not use negative weights, an OLS regression was used to reveal both the positive and negative impacts of each weather feature on daily call volumes. The data shows that weather conditions such as freezing rain and snow significantly increase the number of calls, while milder conditions like rain and overcast skies during fall and winter result in fewer calls. Additionally, spring and summer weather features tend to cause a slight increase in calls on average. All of these coefficients are stored in a CSV file called `OLS_coefficients.csv`.



Discussion

This project aimed to answer several research questions related to how call volumes and classifications change with climate variables, and how these volumes vary by time of day, day of week, season, and year. The main research questions were:

1. How do call volumes and classifications change with climate variables like temperature, humidity, AQI, and more?
2. How does the volume of CAHOOTS calls vary over time of day, day of week, season, and year?
3. How can upcoming weather conditions be used in call quantity prediction?

The results provide valuable insights into these questions. For instance, call volumes were found to decrease in 2023 compared to previous years. Seasonal variations showed an increase in calls during spring, summer, and extreme winter events, while welfare checks were more common in summer and public assists and transports were more frequent in winter. Weekends saw fewer calls and hourly peaks occurred at 11:00 AM and 5:00 PM.

Strengths:

A major strength of this analysis lies in its comprehensive integration of multiple data sources, including CAHOOTS call records, climate data from Visual Crossing, and air quality data from the World AQI Project. The use of an Ordinary Least Squares (OLS) regression model provided a robust methodological approach to identifying significant predictors of call volumes, and the visuals did a good job of showing the most impactful trends in the data.

Weaknesses and Caveats:

However, there are some limitations to this study. The reliance on historical data may not fully capture emerging trends or unexpected future events. Additionally, while the prediction model's accuracy is high, there is still a margin of error that will come from the inaccuracy of the weather forecasts. The data integration process, despite being thorough, could benefit from real-time data updates to improve the model's responsiveness to changing conditions.

Interpretation of Results:

The results largely answer the research questions by identifying clear patterns in call volumes related to specific weather conditions and temporal factors. For example, it was found that extreme weather events such as freezing rain and snow significantly increase the number of calls, while milder conditions like rain and overcast skies during fall and winter result in fewer calls. Additionally, spring and summer weather features tend to cause a slight increase in calls on average.

These findings suggest that call volume and classification analytics are crucial for determining factors like what to bring in the van and how to schedule staff. For instance, knowing that a cold front in January, which can have dangerous side effects, translates into higher call volumes can help CAHOOTS better prepare and allocate resources.

Next Steps:

Next steps to better answer the research questions and improve CAHOOTS operations include:

1. Expanding Data Sources: Incorporating additional data sources, such as socio-economic factors or public health data, could provide a more holistic view of the factors influencing CAHOOTS calls.
2. Qualitative Analysis: Conducting qualitative studies, such as interviews with CAHOOTS responders and community members, could provide deeper insights into the reasons behind call patterns and improve the interpretability of quantitative findings.
3. Model Enhancement: Refining the predictive model to account for more nuanced weather patterns and incorporating advanced machine learning techniques like deep learning could further improve prediction accuracy.