Regression Analysis of Seasonal Patterns in Vehicle Theft Data*

Kumari Shyama

October 8, 2025

Every year, a significant number of vehicle theft reports are captured by the Police Department in California, USA. In the year 2024, 7,250 vehicle thefts were reported in San Francisco. The project aims to explore seasonal patterns of vehicle thefts in San Francisco throughout the year using a Simple Linear Regression Model to understand if there is any significant impact of summer, winter, or holiday seasons on the number of thefts occurring. The analysis results suggest that there is an overall decrease in thefts in the year 2024.Though, the regression model only explains approximately 33% variation in the data. Patterns suggesting increase in theft in summer and a decrease in winter are visible. However,a simple linear regression model might not be the best fit for the research question asked as there can be various factors involved affecting the decline and the mid year spike in the number of vehicle theft reports.

1 Introduction

In the United States, the issue of vehicle theft has significantly impacted residents for many years. Incidents of stolen vehicles, break-ins, and parts theft have been increasing rapidly. However, in 2024, San Francisco, California, saw its first significant decrease in the number of such cases (Fang 2025).

Figure 1 displays the sudden decline in the number of motor vehicle theft incidents in San Francisco for the year 2024, compared to previous years starting from 2018. Many factors could have contributed to the resource management and allocation for the San Francisco Police Department (SFPD) for the same. While researchers often emphasize the overall crime trend and spatial data for coming up with theft prevention measures, there has been less emphasis on the seasonal trend of vehicle thefts. Our initial assumption states that during summer,

*Project repository available at: https://github.com/shyamaku/MATH261A-project.

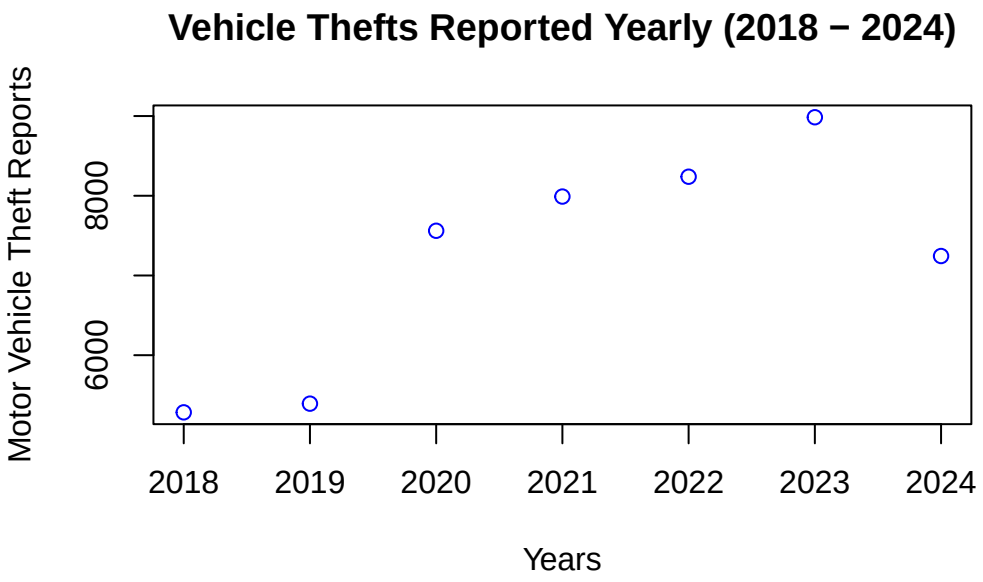

Figure 1

there is a significant increase in people going out, thus increasing the number of vehicles on the street. However, during peak winters, there would be a decrease in vehicles on the road. This project aims to study whether there is a significant impact on thefts based on seasonal variations and holiday periods in the year 2024.

The project analyses trends using scatter plots and applies simple linear regression to interpret the correlation between seasons and vehicle theft. It also examines the limitations of the linear regression model using the residual vs fitted plot. Through this study, the expectation is to provide significant results with the aim of assisting the SFPD with better resource allocations during those periods to further reduce the crime rates.

The remainder of this paper is structured as follows: Section 2 introduces the data used in this analysis. Section 3 describes the simple linear regression model used for the purpose of answering the research question. It also explains the significance of using ANOVA and residuals in assessing the fit of the model. Section 4 explores in depth the results of the analysis and examines if the assumptions of the model being a good fit hold true. Section 5 discusses the interpretations based on the research question and whether the model is efficient in analyzing the seasonal patterns, if any. Additionally, it discusses the limitations of the model for the given data and scope of future research.

2 Data

The dataset used in the project is obtained from DataSF, which is an open data portal (DataSF 2025). It is considered a reliable source of data as the incidents are recorded through official police reporting systems. However, it may contain some wrong classifications due to human errors at the time of reporting, and there might be cases where the thefts have not been reported. Additionally, it does not contain any confidential information related to the incidents. The dataset comprises 969326 records of various incidents (e.g., Fraud, Assault, Motor Vehicle Theft) captured from 2018 to the present. Each row describes the date, time, year of an incident, along with incident description, ID, category, police district, analysis neighborhood, resolution, and other incident area-related fields.

For the scope of this analysis, the data has been filtered to include only records of "Motor Vehicle Theft" for the year 2024. The key variables used in our regression analysis are mentioned below:

Key Variables	Description
week (numerical)	Week of the year. Extracted from the Incident Date field.
num_theft (numerical)	Aggregated number of vehicle theft reports for each week of the year.

The data has been processed in a weekly format, as the number of theft reports daily might not significantly express the trend with the change in seasons, while monthly analysis might miss any granular-level trend in the data. Since there has been a recent decline in vehicle theft incidents in 2024, this project aims to explore in depth trends based on the lower number of crimes for the given year.

3 Methods

This project adopts the simple linear regression model with week of the year as the predictor X_i and number of vehicle thefts as the response Y_i, where $i = 1, 2, ..., n$ and n is the number of records. The model can be represented with the following equation:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

In this model, β_1 represents the slope, which describes the change in the number of vehicle thefts as the week increases. β_0 represents the intercept, which describes the number of vehicle thefts at week 0, and ϵ represents the random error, which describes the unexplained variation of data.

Next, after fitting the model as per the above equation, analysis of variance is performed using ANOVA. Additionally, the project analyses the model's fit using the residual values. Residuals are the difference between observed and predicted values, written as,

$$e_i = Y_i - \hat{Y}$$

The analysis has been implemented using the R programming language (R Core Team 2024) and packages including (Wickham et al. 2023),(Wickham 2016)

4 Results

This section presents the result of the analysis conducted to examine whether motor vehicle thefts follow a seasonal pattern. A scatterplot is used to visualize the relationship between weeks and the number of motor vehicle theft reports. Figure 2 indicates a negative relation between the two key variables

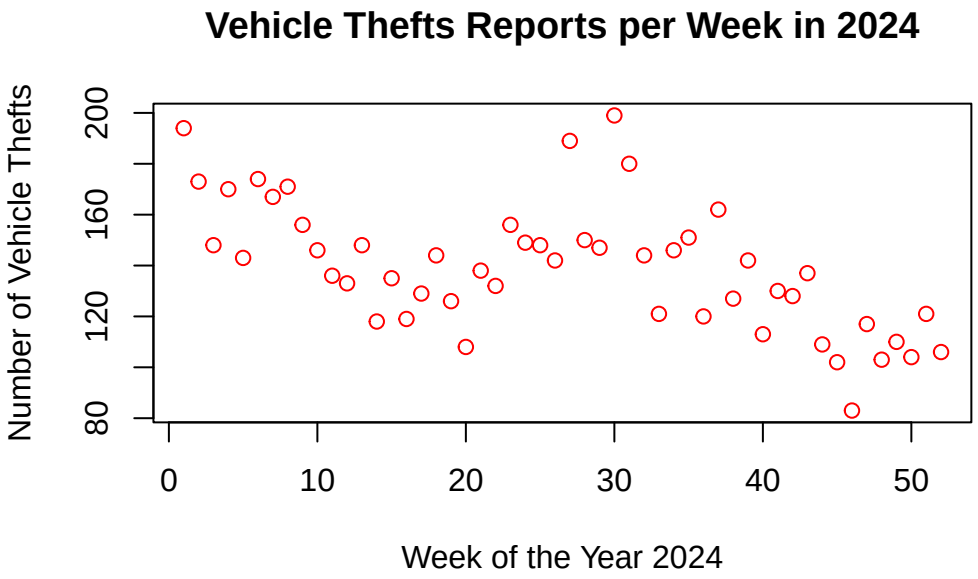

Figure 2: Scatter plot of week in x-axis and number of vehicle theft reports in y-axis.

The summary after fitting the linear regression model with outcome as the number of motor vehicle theft reports and predictor as the week of the year is described below:

- The estimated slope parameter is b_1 = -0.953. In other words, as the weeks progress, there is an average decrease in the number of vehicle theft reports by ≈ 1.

- The estimated intercept is $b_0 = 164.572$, which explains the expected number of thefts at week 0.
- The t value = -4.9389423 and p-value = 9.1749221×10^{-6} show that week is a statistically significant predictor for the estimated decrease in the number of vehicle theft reports. However, this cannot be interpreted as such because it was already established that in 2024, the number of vehicle thefts had decreased significantly, and the decrease does not directly depend on the increase in weeks.
- The R-squared value: 0.3279 explains $\approx 33\%$ of the variation in the theft report numbers per week. This does not cover a significant amount of variation in the data.

To assess the fit of the linear regression model, ANOVA is applied to the fitted model. The null hypothesis $\beta_1 = 0$ (no linear relationship between the two variables) is compared with the two-sided alternative hypothesis $\beta_1 \neq 0$. In order to reject the null hypothesis, the ratio between the mean squared regression (MSR) and the mean squared error (MSE) should be significantly larger than 1. In the above result, MSR = 1.0646452×10^4 » MSE = 436.453, hence the ratio is as desired. On the basis of this, we can reject the null hypothesis, interpreting that there is some amount of linear relationship between the two variables.

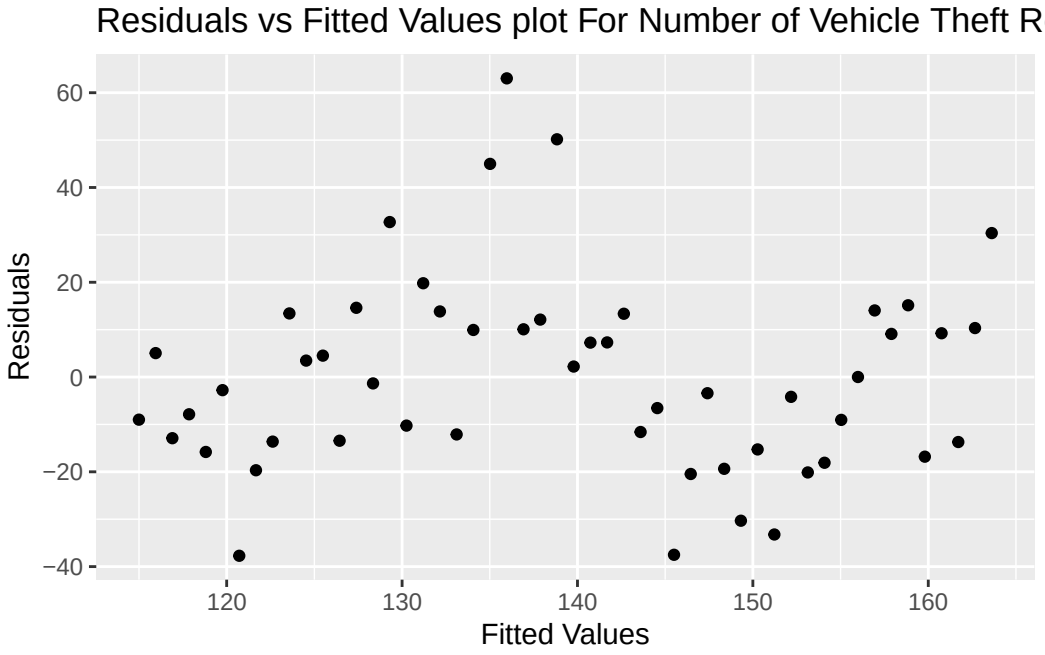

Figure 3: Scatter plot of fitted values (x-axis) and residuals (y-axis)

Another way to assess the fit of our linear regression model is to examine the residual vs fitted value plot and assess if the assumptions of a linear regression model hold for our analysis. Figure 3 examines the residuals to assess model fit. From the scatter plot, the following can be interpreted:

- There is a somewhat linear relationship between the number of vehicle theft reports and weeks, as there are no visible arcs or curves in the plot, but due to some points forming waves, it is difficult to interpret the same with certainty.
- The error terms have equal variance, as there is no major spread in the data points with the increase in weeks.
- The error terms are not independent, as we can see an increase and decrease in the residual points forming waves with the increase in the fitted values.
- The error terms might not be normally distributed, but to verify that a quantile-quantile plot would be needed.

5 Discussion

The analysis of vehicle thefts and the week of the year 2024 explains a few seasonal patterns in the data, according to the research objective. Figure 2 indicates that the initial assumptions hold value. Between week 23-37 (summer, 4th of July), there can be seen a spike in the number of thefts, which can be due to the increase in outdoor activities in the season, leading to higher usage of vehicles. In contrast, week 45-55 (winter, Thanksgiving, Christmas) shows the lowest number of thefts in San Francisco might be due to the colder season and decrease in vehicles on the streets, which falls under the earlier assumption of this project.

Even so, these interpretations cannot be made with certainty, as after assessing the residual vs fitted model plot, it can be said that the linear regression model doesn't necessarily fit the data. As the errors are not independent, there might be non-linear patterns that the linear regression model is not fit to expand. Thus, the increasing and decreasing patterns of the number of vehicle thefts can be influenced by other factors such as an increase in police patrols in the middle to latter half of the year and implementation of automated drones and cameras to increase surveillance.

For further inspection of the assumptions made initially, the analysis should be done with multiple years of data to assess if the seasonal trend is valid for each year. Analysis with the addition of a few other key predictors, like police districts or incident time (day or night), can be used to analyze the seasonal trends better with multiple regression models. Time series models should be explored for a better analysis of the data to explore the non-linear patterns. This could be helpful in exploring different trends of the number of vehicle thefts across weeks, which would prove significant towards the real goal of assisting the SFPD with enhanced planning and reducing the theft rates further.

References

DataSF. 2025. "Police Department Incident Reports: 2018 to Present." Police Department. /url%7Bhttps://data.sfgov.org/d/wg3w-h783%7D.

Fang, Tim. 2025. "Car Thefts in California down 13% Statewide, CHP Says." CBS News. https://www.cbsnews.com/sanfrancisco/news/california-car-thefts-down-13pct-statewide-23-24-chp-says/.

R Core Team. 2024. "R: A Language and Environment for Statistical Computing." Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. "Dplyr: A Grammar of Data Manipulation." https://CRAN.R-project.org/package=dplyr.