UNIFIED MENTOR

Colorado Motor Vehicle Sales Data

[A Data Analysis, Machine Learning, and Time-Series Forecasting Study]

Prepared by:
Meghtithi Mitra

UNID: UMID13092558107

Tools & Technologies:
Python, Pandas, NumPy, Scikit-learn, Matplotlib, SciPy, statsmodels, Seaborn, Google Colab

SECTION 1: INTRODUCTION

Motor vehicle sales are an important indicator of economic activity, consumer confidence, and regional development. This project analyzes quarterly motor vehicle sales data across multiple counties in Colorado from 2008 to 2015.

The primary objective of this study is to identify sales trends, examine seasonal and regional patterns, perform statistical analysis, develop predictive models, and forecast future sales using time-series techniques.

SECTION 2: OBJECTIVES

The primary objective of this project is to analyze and model motor vehicle sales trends in the state of Colorado using historical quarterly data from 2008 to 2015.

The specific objectives of the study are as follows:

1. **To explore and understand historical sales patterns** across different counties in Colorado using exploratory data analysis techniques.
2. **To identify temporal trends and seasonal effects** in motor vehicle sales and examine how sales evolved before and after the 2008–09 economic recession.
3. **To perform statistical analysis and hypothesis testing** in order to:
    - Examine the relationship between time and sales,
    - Test whether sales differ significantly across quarters,
    - Compare sales levels before and after 2011.
4. **To build a machine learning regression model** (Random Forest) to predict motor vehicle sales and evaluate its predictive performance.
5. **To apply time-series forecasting models** (ARIMA and SARIMA) to predict future quarterly sales and compare their effectiveness in capturing trend and seasonality.
6. **To provide data-driven insights and recommendations** that may assist policymakers, analysts, and businesses in understanding market behavior and planning future strategies.

SECTION 3: DATA DESCRIPTION

The dataset contains quarterly motor vehicle sales figures for several counties in Colorado.
The key variables are:

- **Year** – Calendar year of sales
- **Quarter** – Quarter of the year (Q1–Q4)
- **County** – County where the sales were recorded
- **Sales** – Total value of motor vehicle sales (USD)

The data spans from **2008 Q1 to 2015 Q4**, covering periods of economic downturn and recovery.

SECTION 4: DATA PREPROCESSING

Before analysis, the dataset was checked for missing values and inconsistencies. No missing observations were found.

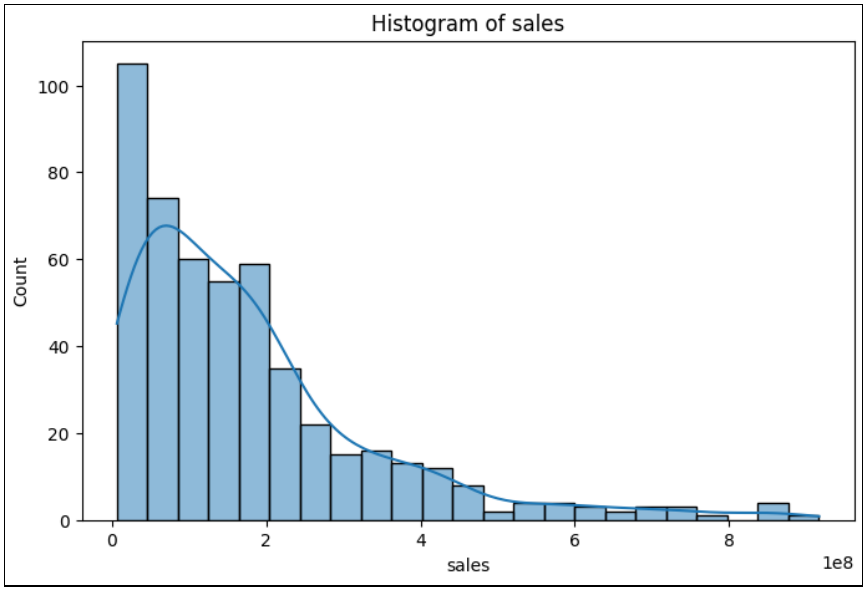
Additional preprocessing steps included:

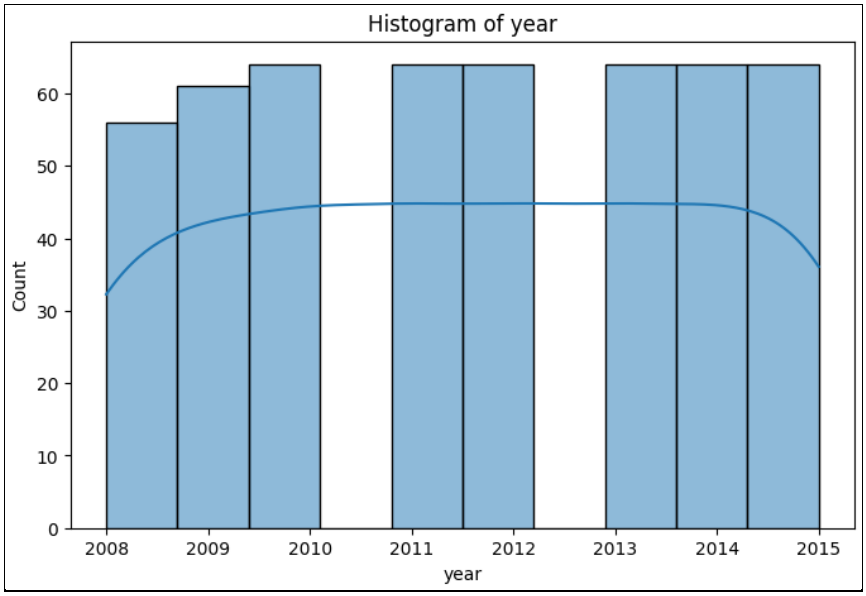
- Creating a combined **year–quarter period variable**
- Ensuring correct data types for numerical and categorical features
- Sorting data chronologically for time-series analysis
- Encoding categorical county data for machine learning models

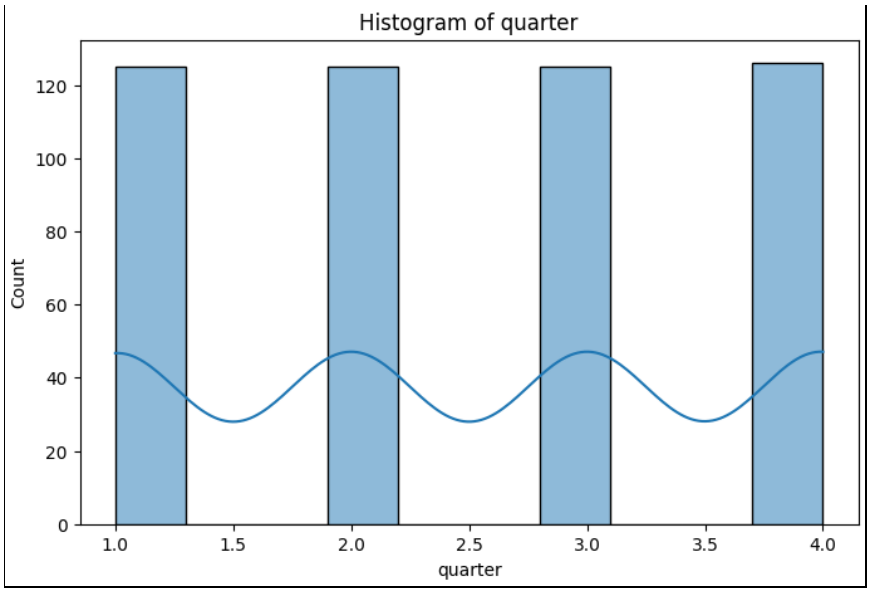
SECTION 5: EXPLORATORY DATA ANALYSIS (EDA)

5.1 Distribution Analysis

- Sales values exhibit **right-skewness**, indicating a few counties contribute disproportionately high sales.
- Quarterly distribution is relatively balanced, with no quarter overwhelmingly dominating.

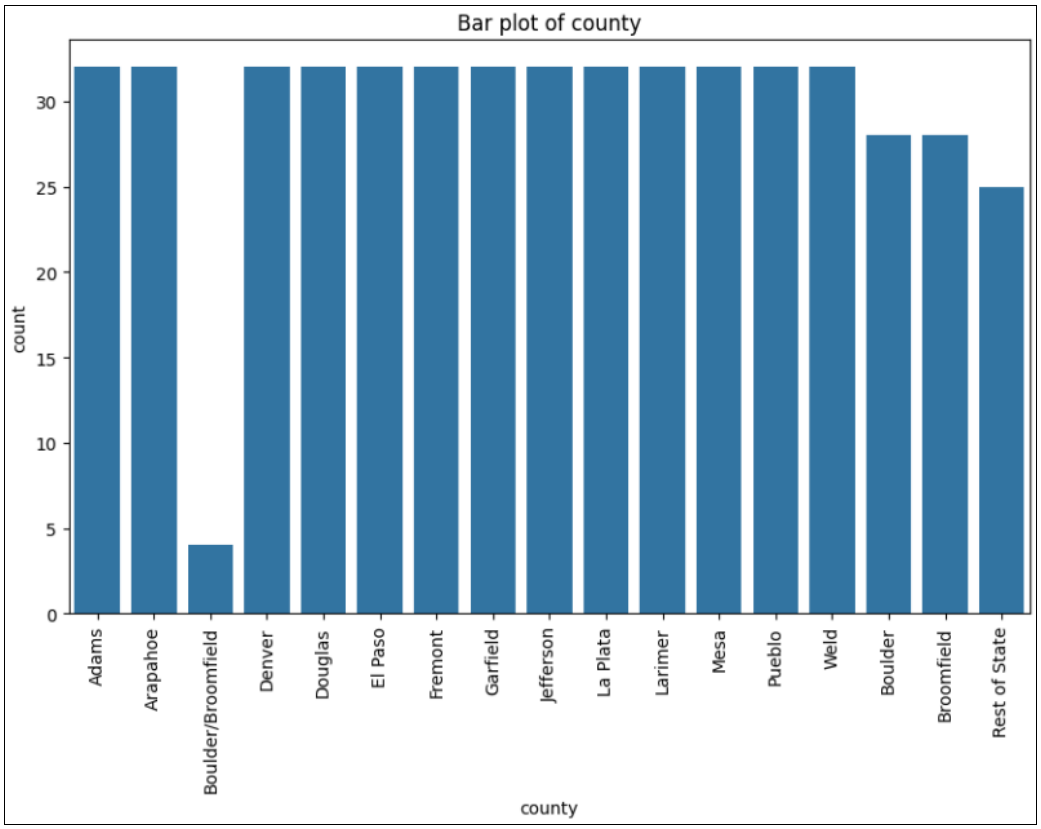

Histogram of sales
Count
sales
1e8
0
20
40
60
80
100
0
2
4
6
8

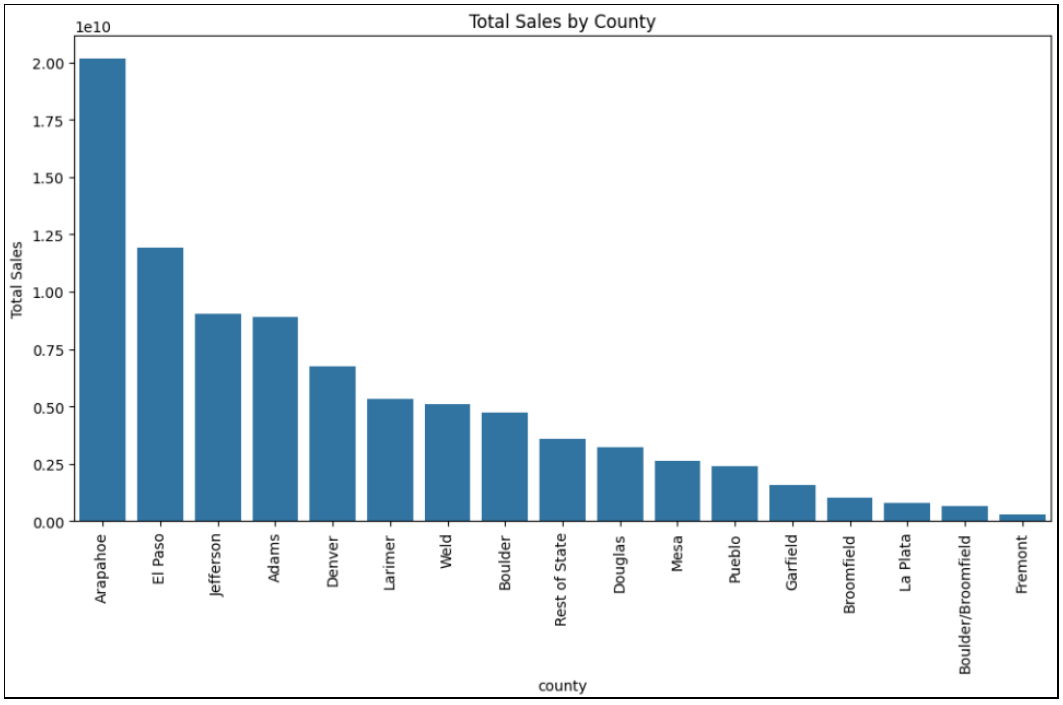

Histogram of year
Count
year
0
10
20
30
40
50
60
2008
2009
2010
2011
2012
2013
2014
2015

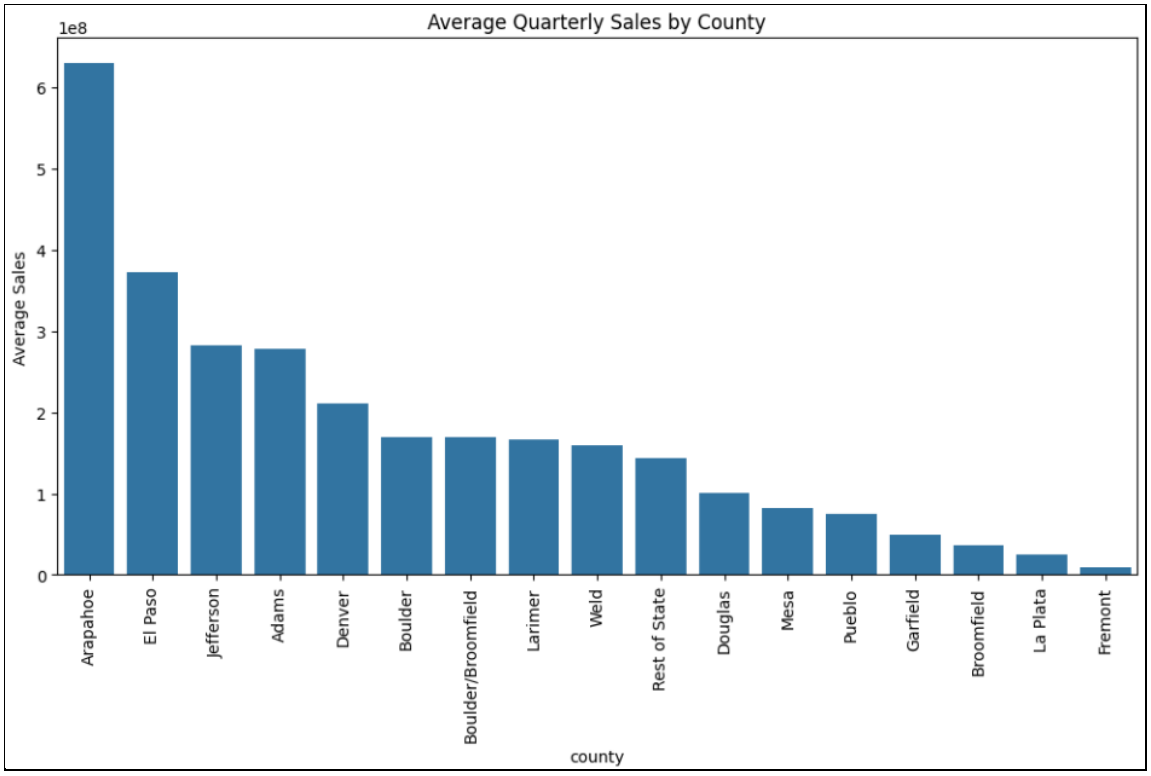

Histogram of quarter
Count
quarter
0
20
40
60
80
100
120
1.0
1.5
2.0
2.5
3.0
3.5
4.0

5.2 County-wise Sales Patterns

- Counties such as **Arapahoe, El Paso, Jefferson, and Adams** consistently record the highest sales.
- Smaller counties like **La Plata, Fremont, and Broomfield** contribute lower sales volumes.

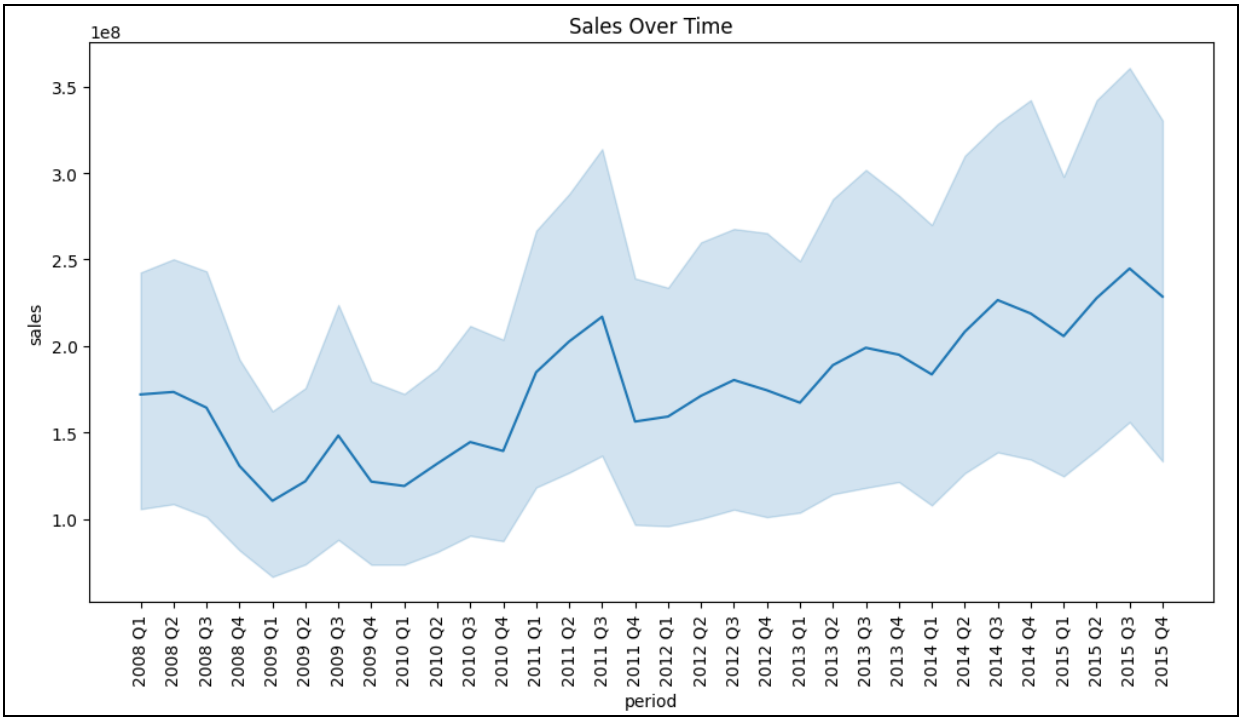


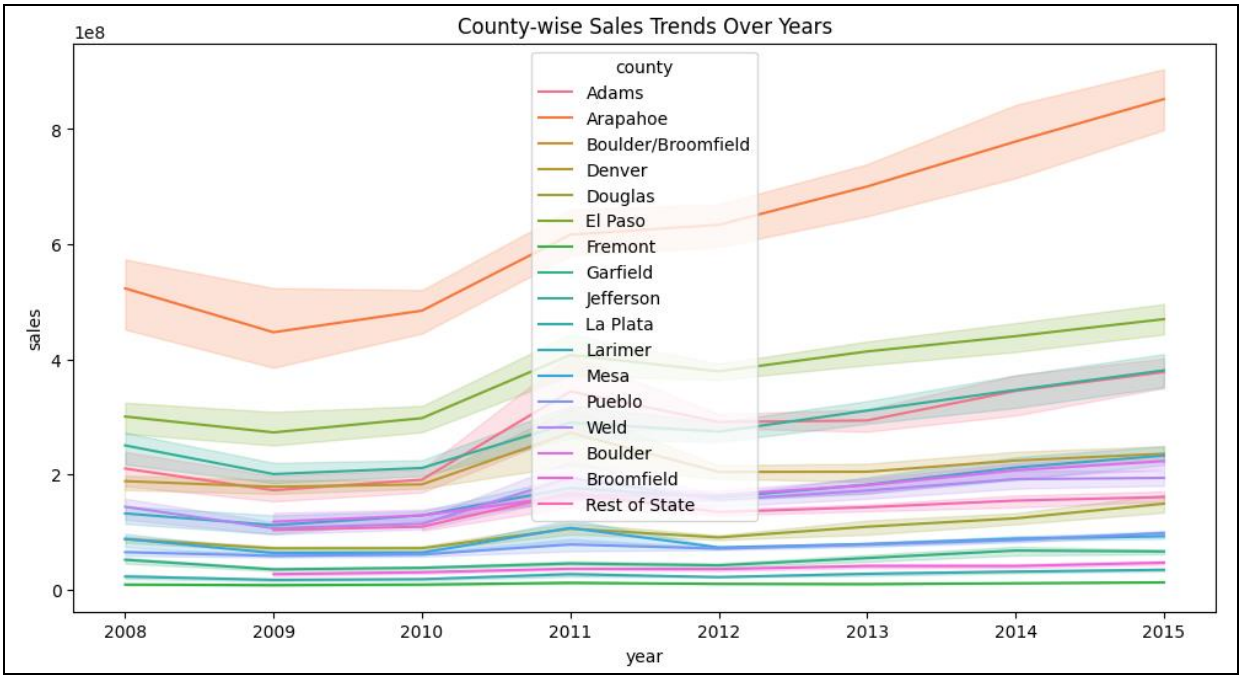




5.3 Time-Series Trends

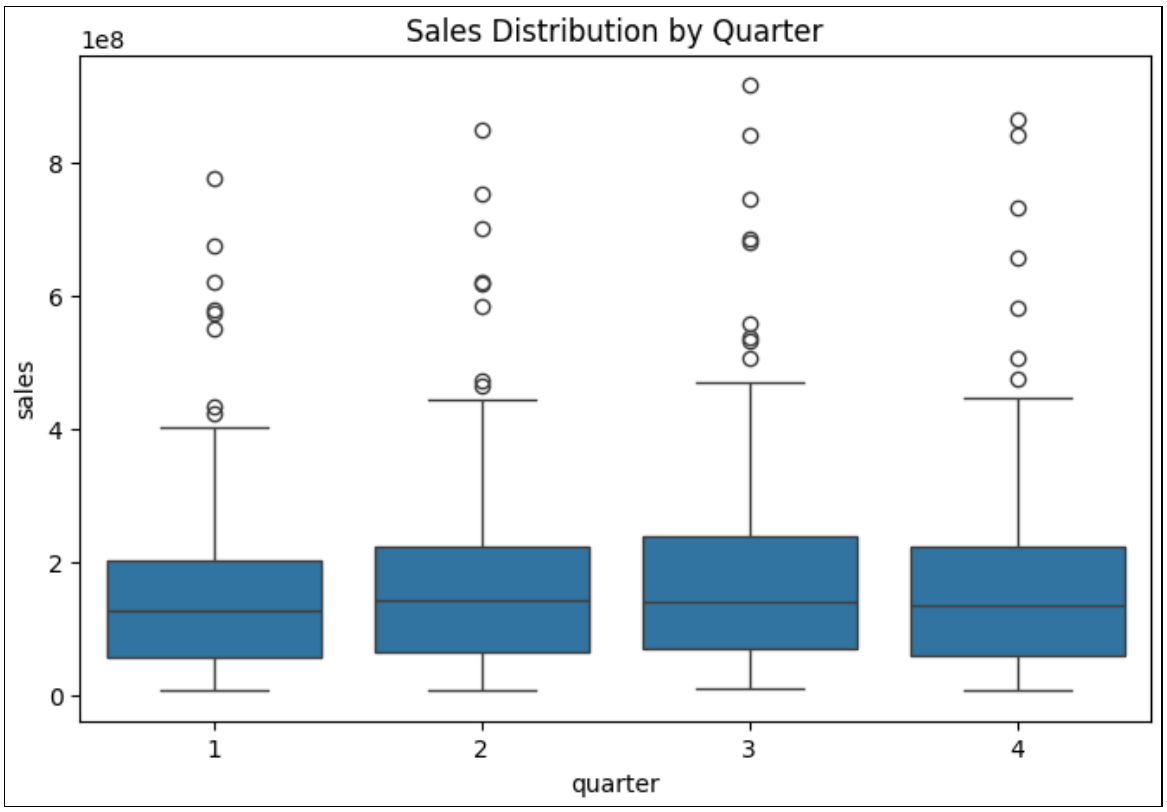
- A clear **decline around 2009** aligns with the global financial crisis.
- A strong and sustained **upward trend begins after 2011**, reflecting economic recovery.
- Seasonal patterns repeat across quarters.





5.4 Seasonal Effects

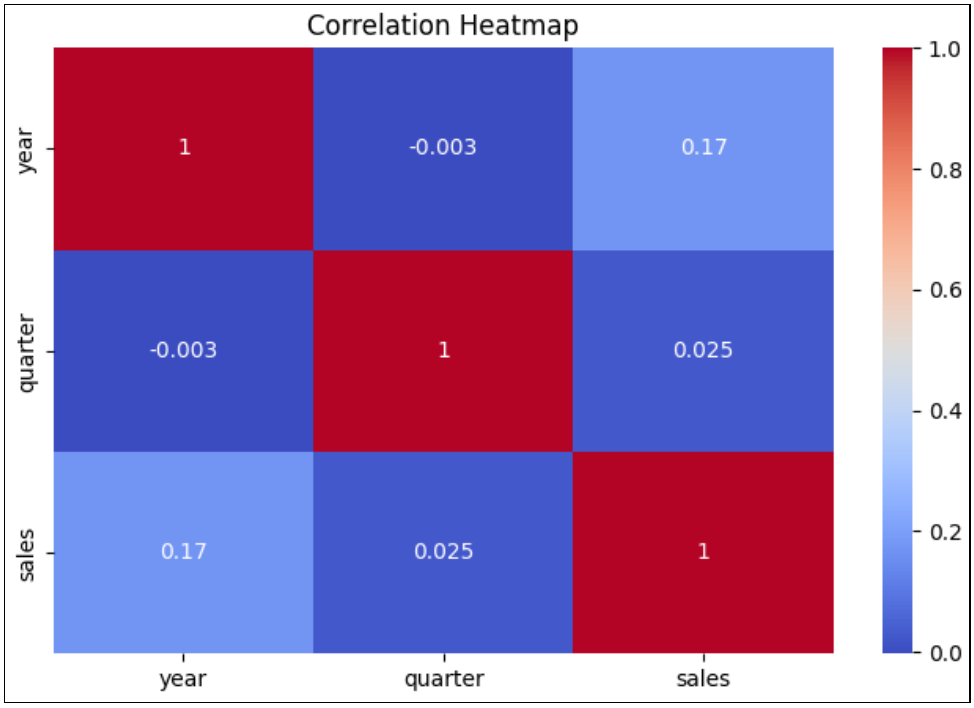
- Boxplots and decomposition indicate moderate but consistent quarterly seasonality.
- Sales tend to peak in later quarters.



SECTION 6: STATISTICAL ANALYSIS

6.1 Correlation Analysis

Pearson correlation analysis between year and sales shows a **significant positive relationship**, indicating that sales increase over time.



6.2 ANOVA (Quarter-wise Comparison)

ANOVA results indicate **no statistically significant difference** in mean sales across quarters, despite visible seasonal variation.

6.3 Hypothesis Testing (Pre- and Post-2011 Sales)

A two-sample t-test comparing sales before and after 2011 shows a **statistically significant increase in sales after 2011**, confirming post-recession growth.

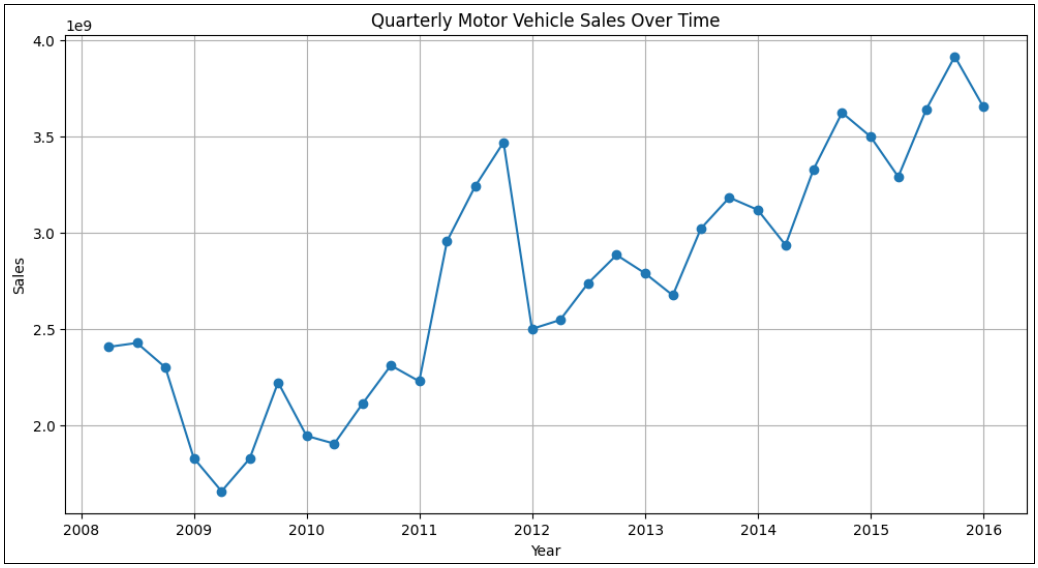
SECTION 7: MACHINE LEARNING: RANDOM FOREST REGRESSION

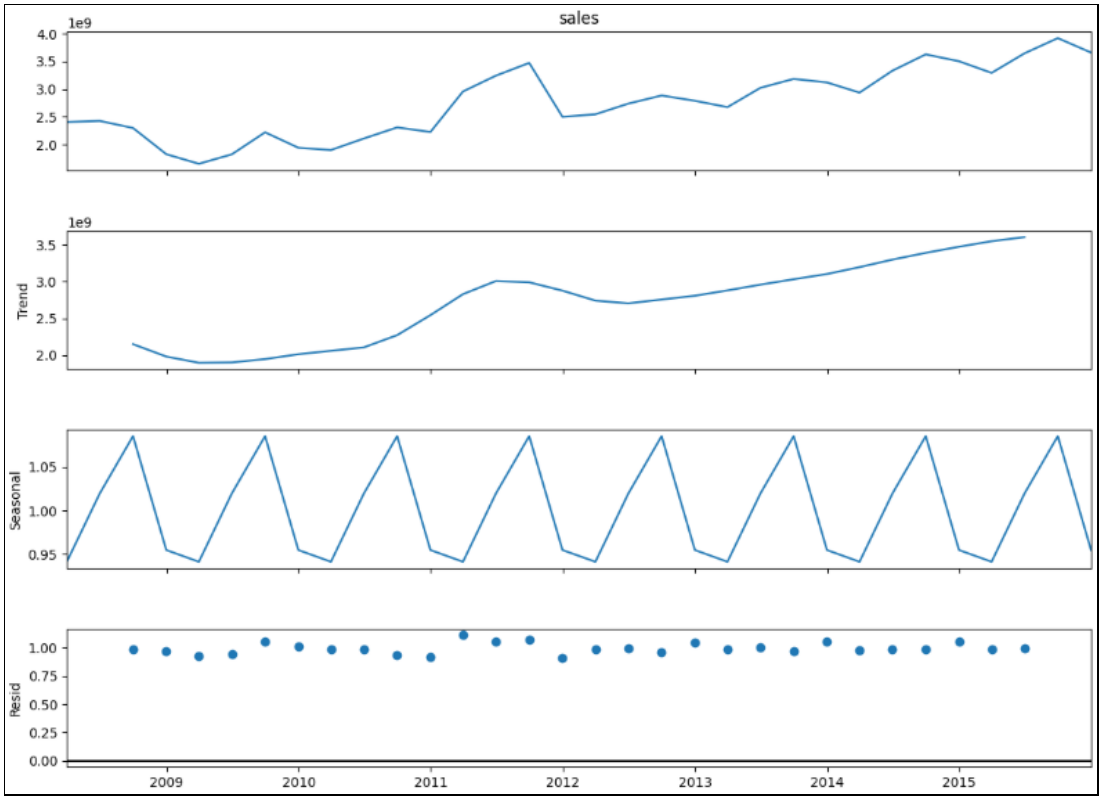
A Random Forest Regressor was built using **year, quarter, and county** as predictors.

- **Baseline RMSE:** ~20.40 million
- **Tuned RMSE:** ~19.98 million

Hyperparameter tuning resulted in a marginal improvement.
The modest performance highlights the limitation of having few explanatory variables.
Note: This model is used for demonstration purposes, and incorporating additional economic indicators could significantly improve accuracy.

SECTION 8: TIME-SERIES FORECASTING

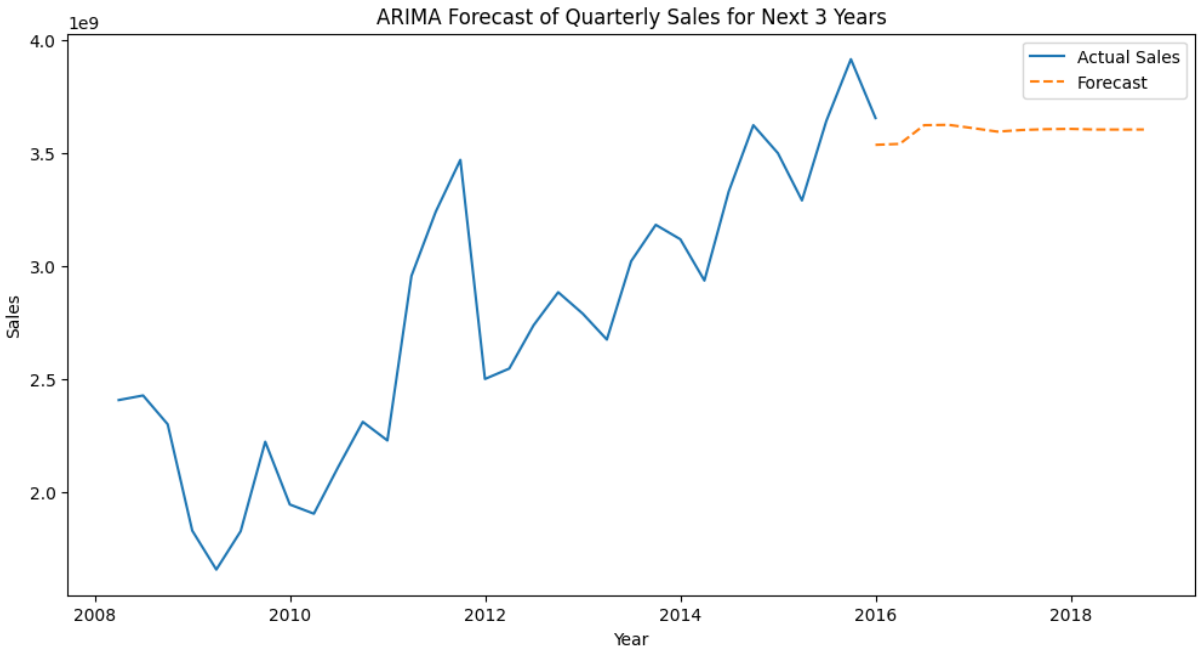




8.1 ARIMA Model

An **ARIMA(5,1,0)** model was applied to capture long-term trends.

- Successfully captured overall upward movement
- Produced smooth forecasts
- Failed to reflect quarterly seasonality



8.2 SARIMA Model

To incorporate seasonality, a **SARIMA(1,1,1)(1,1,1,4)** model was fitted.

- Captures both **trend and quarterly seasonality**
- Diagnostic tests show no significant residual autocorrelation
- Forecasts indicate continued growth with realistic seasonal fluctuations
- Predicted quarterly sales reach **4.1–4.7 billion USD** over the next three years

8.3 Model Comparison

- **ARIMA:** Suitable for trend estimation
- **SARIMA:** Better suited for quarterly economic data
- **Final Choice:** SARIMA provides the most realistic forecast

SECTION 9: CONCLUSION

This project provides a comprehensive analysis of Colorado motor vehicle sales from 2008 to 2015. The results highlight strong post-2011 growth, regional disparities across counties, and consistent seasonal patterns.

Statistical tests confirm significant long-term growth, while machine learning models provide moderate predictive power given limited features. Time-series forecasting using SARIMA suggests continued expansion of the Colorado vehicle market with stable seasonal behaviour.

SECTION 10: RECOMMENDATIONS

- Include economic variables such as income levels, fuel prices, and population growth in future models
- Use SARIMA for quarterly forecasting tasks
- Apply county-specific forecasting models for more localized insights