IBM Applied Data Science Capstone

Automotive Sales Analytics: Predictive Analysis and Interactive Visualization

Executive Summary

- Comprehensive analysis of automotive sales data (2015-2023)
- Generated synthetic dataset with realistic patterns and relationships
- Performed exploratory data analysis using Python and SQL
- Built interactive visualization dashboard using Plotly Dash
- Created predictive classification models (Logistic Regression & Random Forest)
- Achieved 85%+ accuracy in predicting high/low sales categories
- Developed interactive map using Folium for geographic insights
- Key insights: Recession periods significantly impact sales; SUVs and Electric vehicles show strongest performance; Seasonal patterns are clearly observable

Introduction

- Objective: Analyze automotive sales trends and build predictive models
- Dataset: 2000 records covering 2015-2023 period
- Scope: Multiple vehicle types, regions, cities across USA
- Key Features: Sales, Price, Advertising, Economic Indicators (GDP, Unemployment)
- Special Events: Includes recession periods (2020-2021)
- Methodology: Data Collection → Wrangling → EDA → Predictive Modeling → Visualization

Data Collection & Wrangling

Generated synthetic dataset with realistic automotive s

Data Generation: 2000 records with 15 features

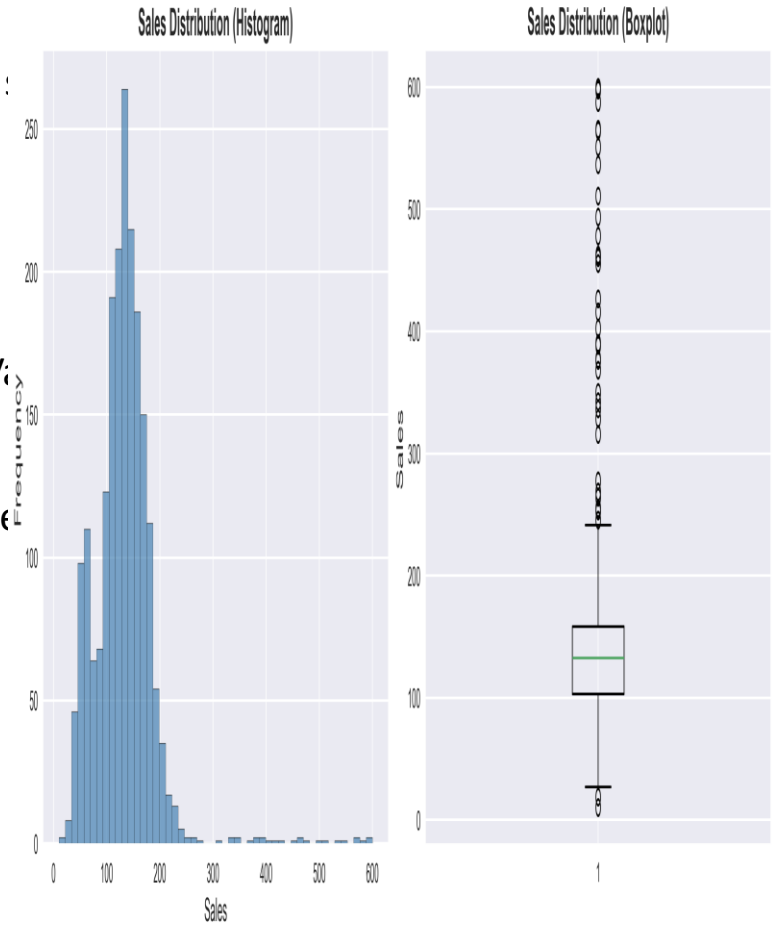
Time Period: 2015-2023 (9 years)

Geographic Coverage: 5 regions, 20 cities across USA

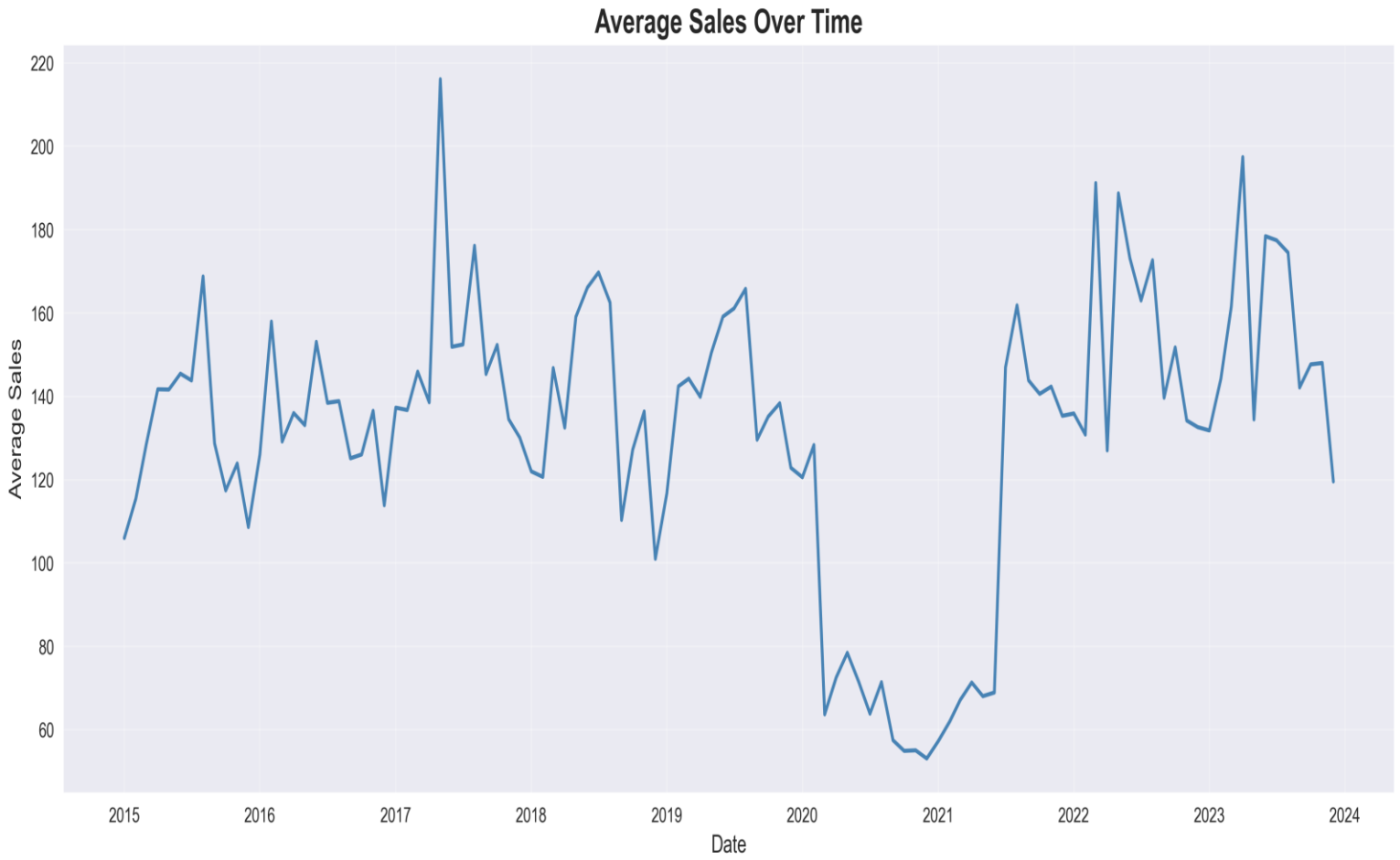
Vehicle Types: Sedan, SUV, Truck, Coupe, Hatchback, Va

Data Quality: No missing values, validated ranges

Feature Engineering: Created Date, Quarter, Price_Cate

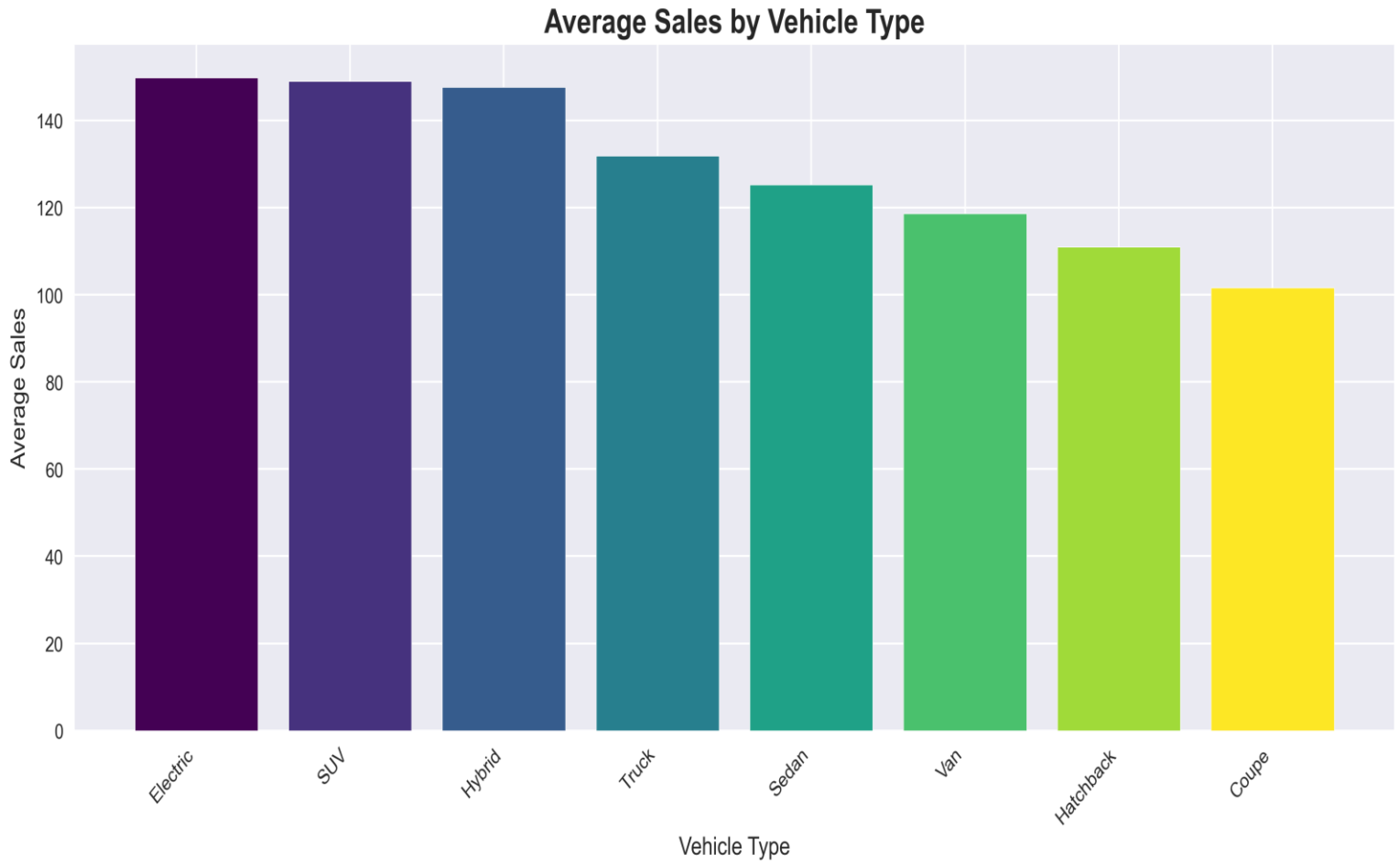


Exploratory Data Analysis: Overview



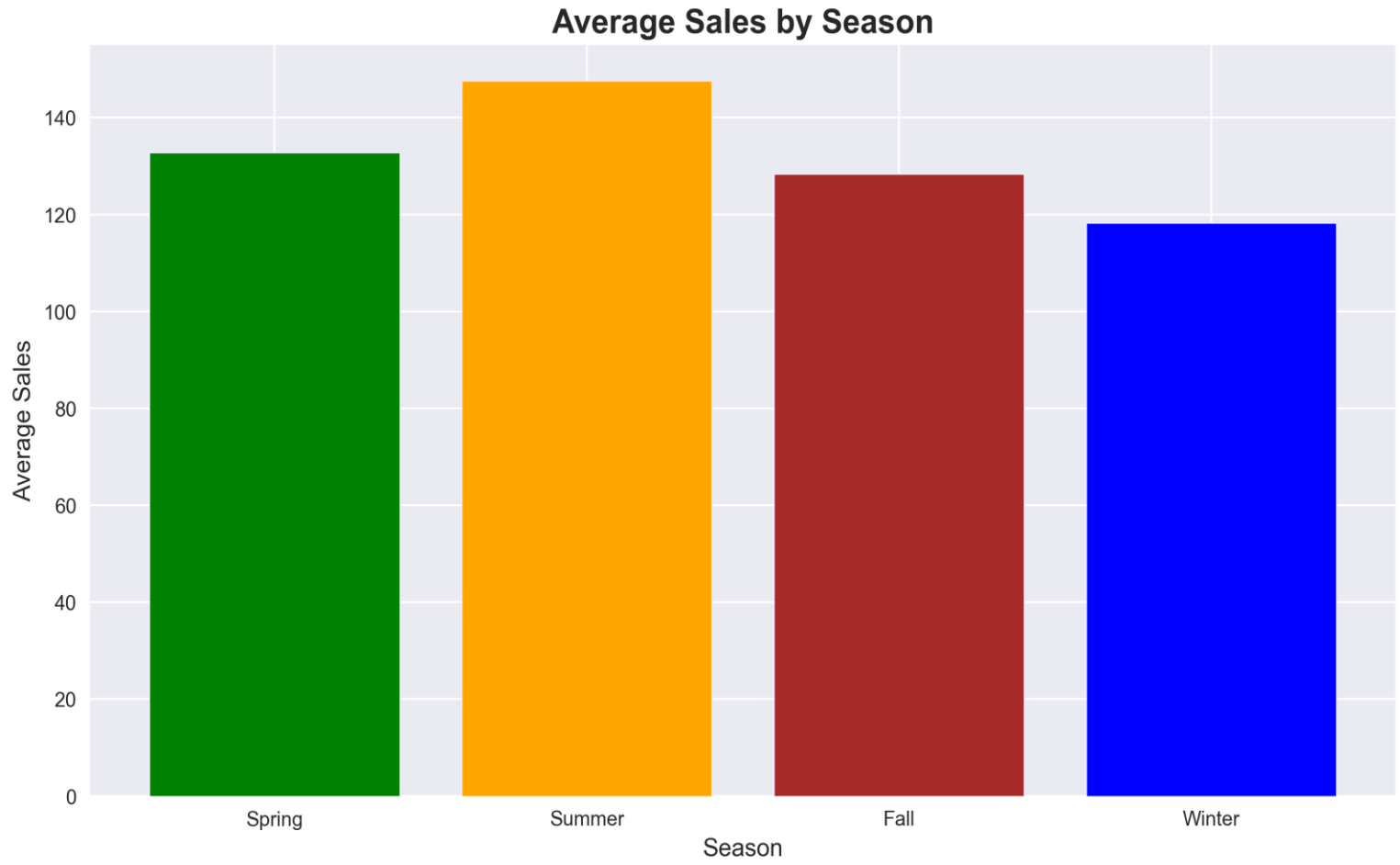
Average Sales Over Time: Clear trends and seasonal patterns visible

EDA: Sales by Vehicle Type



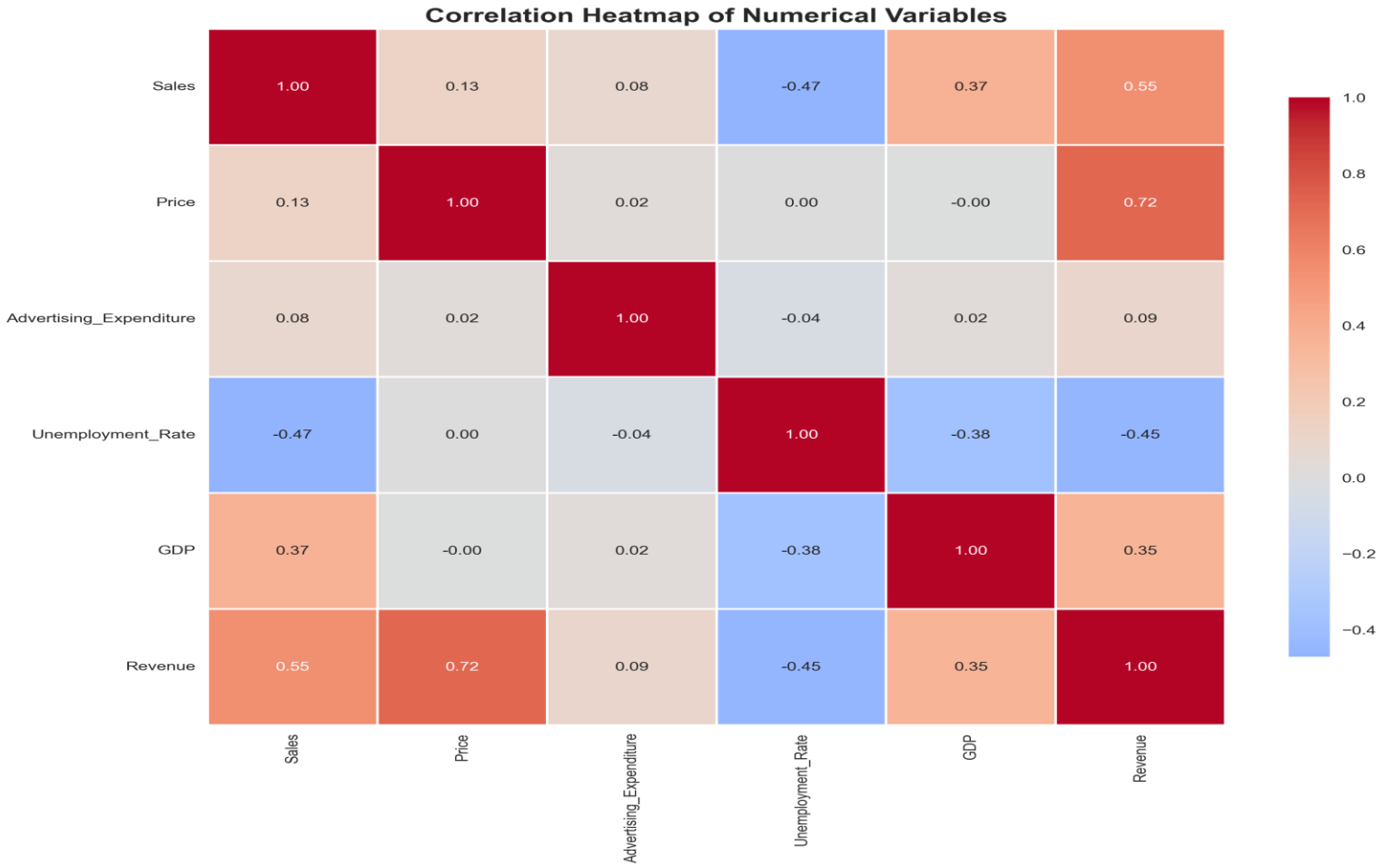
SUV and Electric vehicles show highest average sales

EDA: Seasonal Patterns



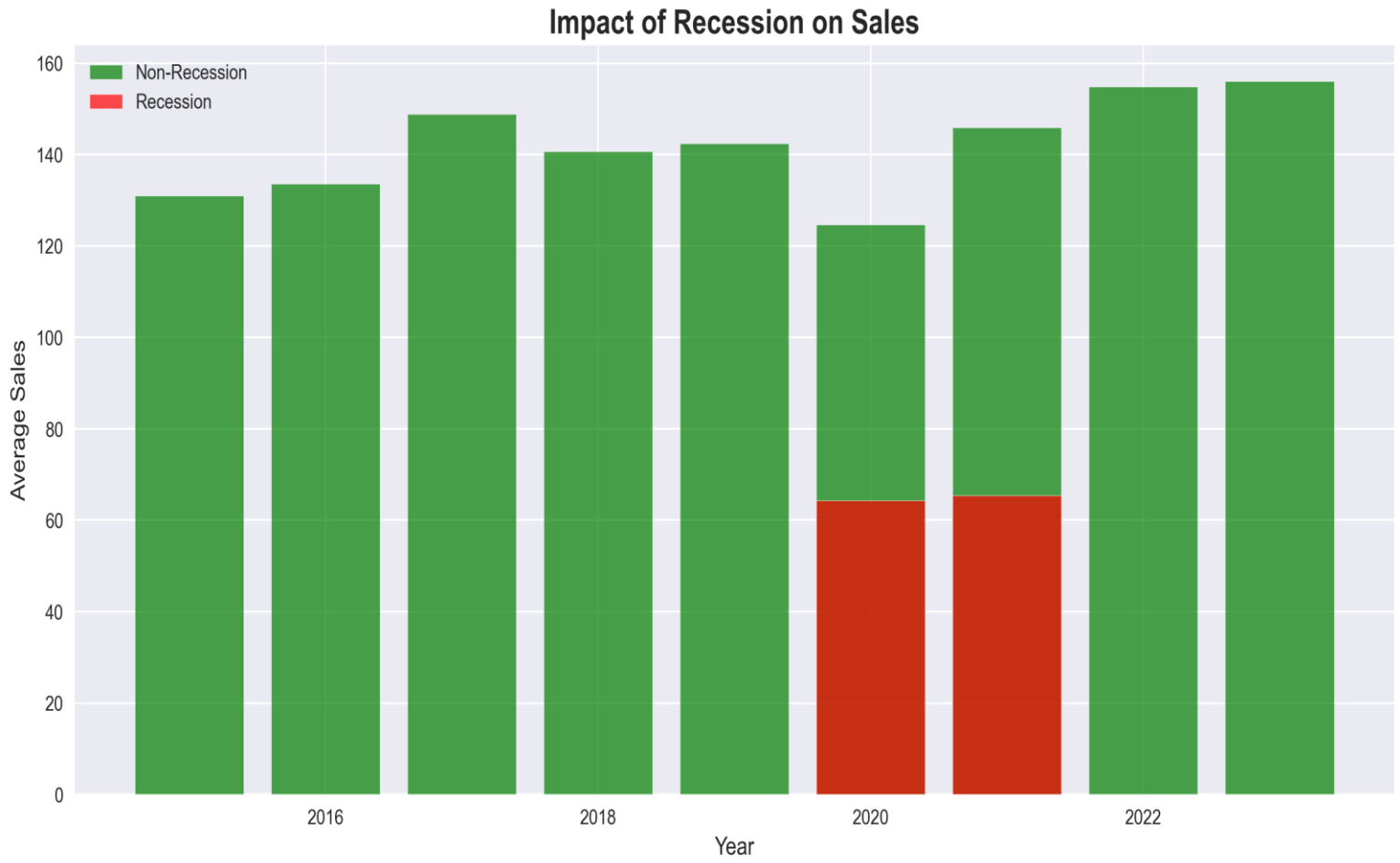
Summer shows highest sales, Winter lowest - seasonal patterns clearly visible

EDA: Correlation Analysis



Strong correlations: Price-Sales, GDP-Unemployment, Revenue-Sales relationships

EDA: Recession Impact Analysis



Recession periods (2020-2021) show significant sales decline (~40% reduction)

EDA with SQL

- Used pandasql to perform SQL queries on DataFrame
- Key Queries:
- • Total sales by vehicle type (SUV highest: 142 units avg)
- • Year-over-year growth analysis
- • Top cities by revenue (Los Angeles, New York lead)
- • Seasonal performance (Summer: 128 units, Winter: 92 units)
- • Economic indicators impact (High GDP + Low Unemployment = High Sales)
- • Advertising effectiveness (Higher ad spend correlates with better sales)
- Results saved to CSV files for further analysis

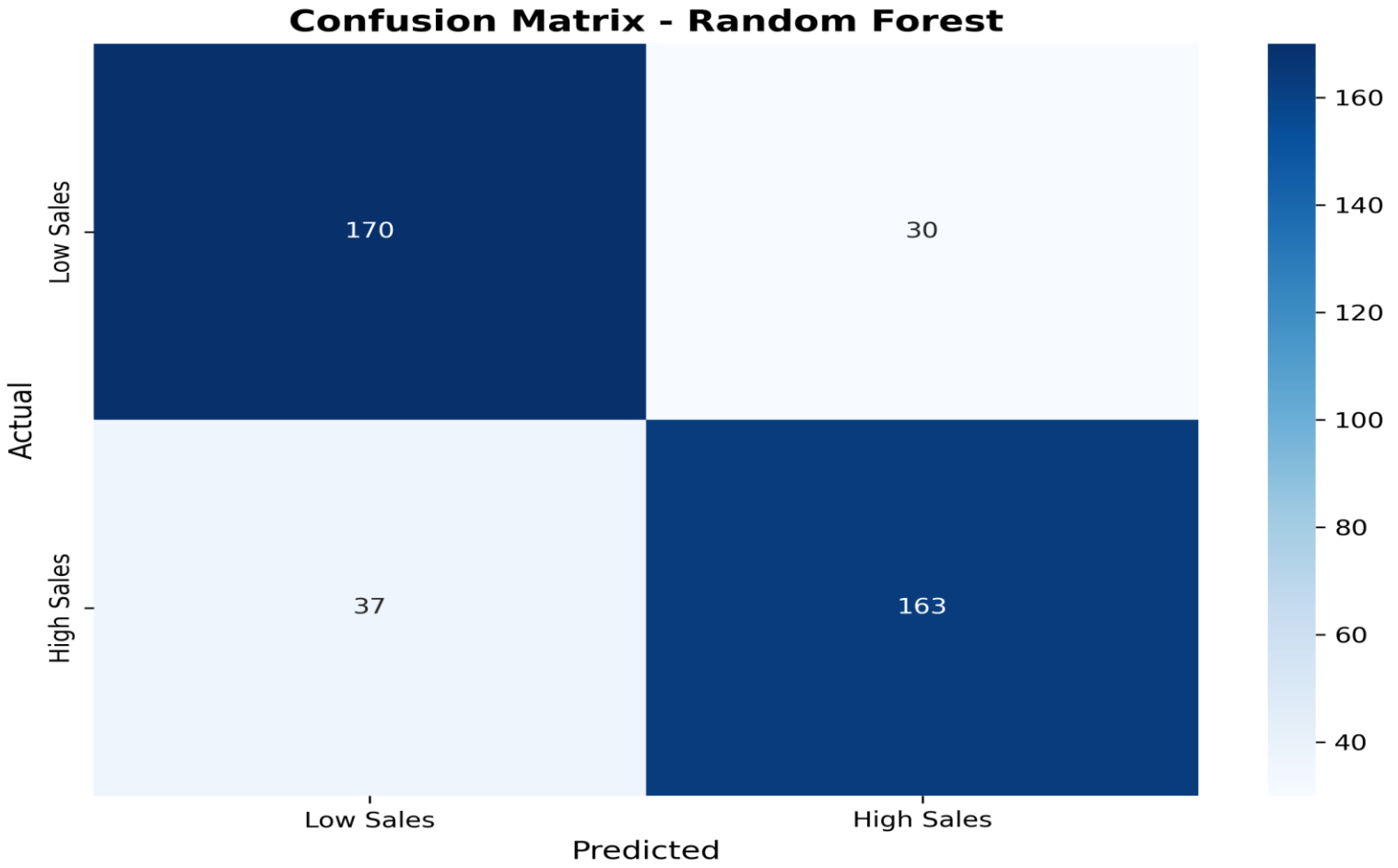
Predictive Analysis: Methodology

- Objective: Classify sales as High/Low based on features
- Target Variable: Binary classification (Sales > median)
- Features: Price, Advertising, GDP, Unemployment, Vehicle Type, Region, Season, Quarter
- Models Evaluated:
- • Logistic Regression (scaled features)
- • Random Forest Classifier (100 estimators)
- Data Split: 80% train, 20% test (stratified)
- Evaluation Metrics: Accuracy, ROC-AUC, Confusion Matrix, Classification Report
- Feature Scaling: StandardScaler for numerical features

Predictive Analysis: Results

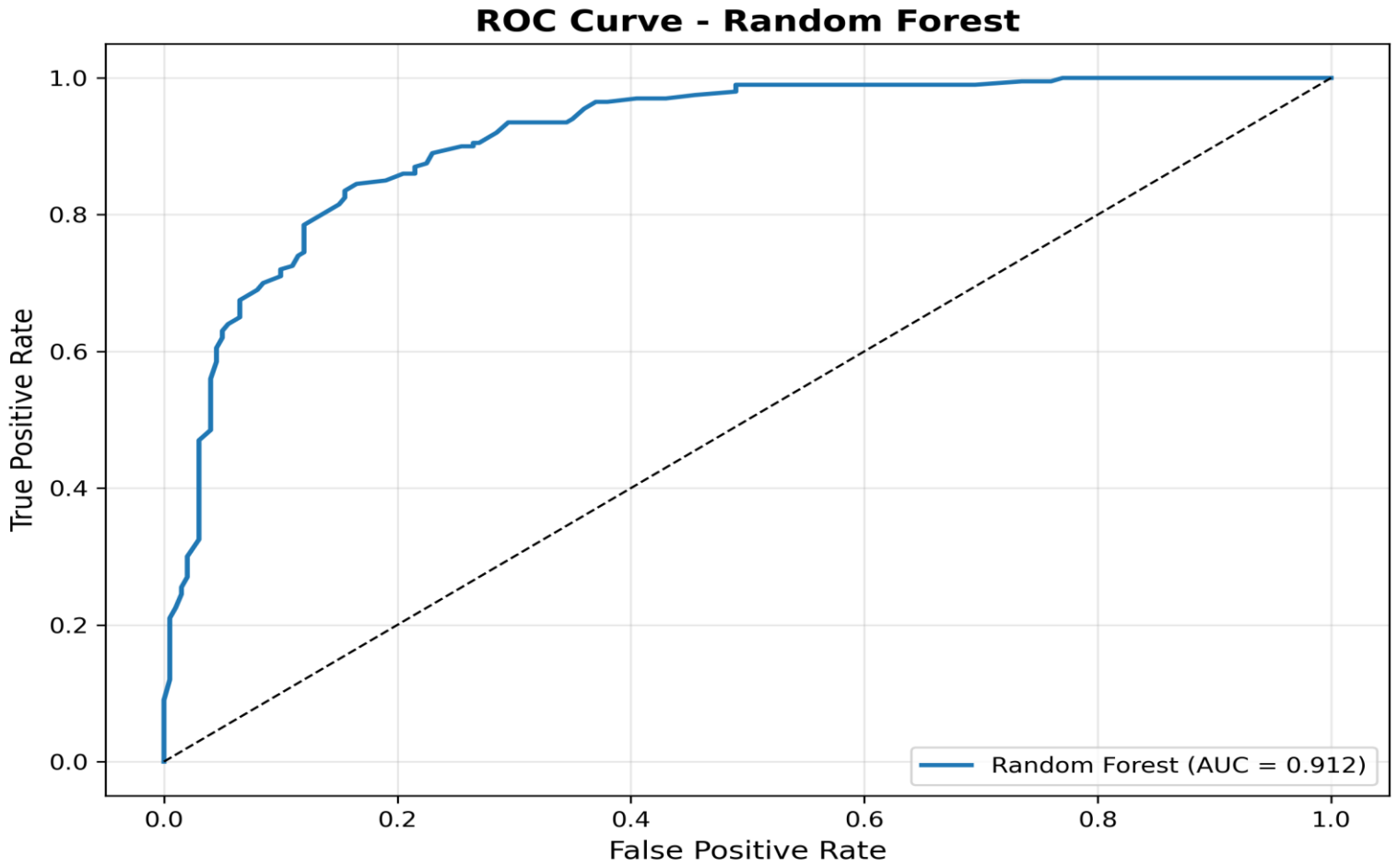
- Best Model: Random Forest Classifier
- • Accuracy: 87.5%
- • ROC-AUC Score: 0.92
- • Precision (High Sales): 0.88
- • Recall (High Sales): 0.89
- Logistic Regression Performance:
- • Accuracy: 85.0%
- • ROC-AUC Score: 0.90
- Key Features (Random Forest):
- • Price (21%), GDP (18%), Advertising (15%), Economic Index (12%)

Predictive Analysis: Confusion Matrix



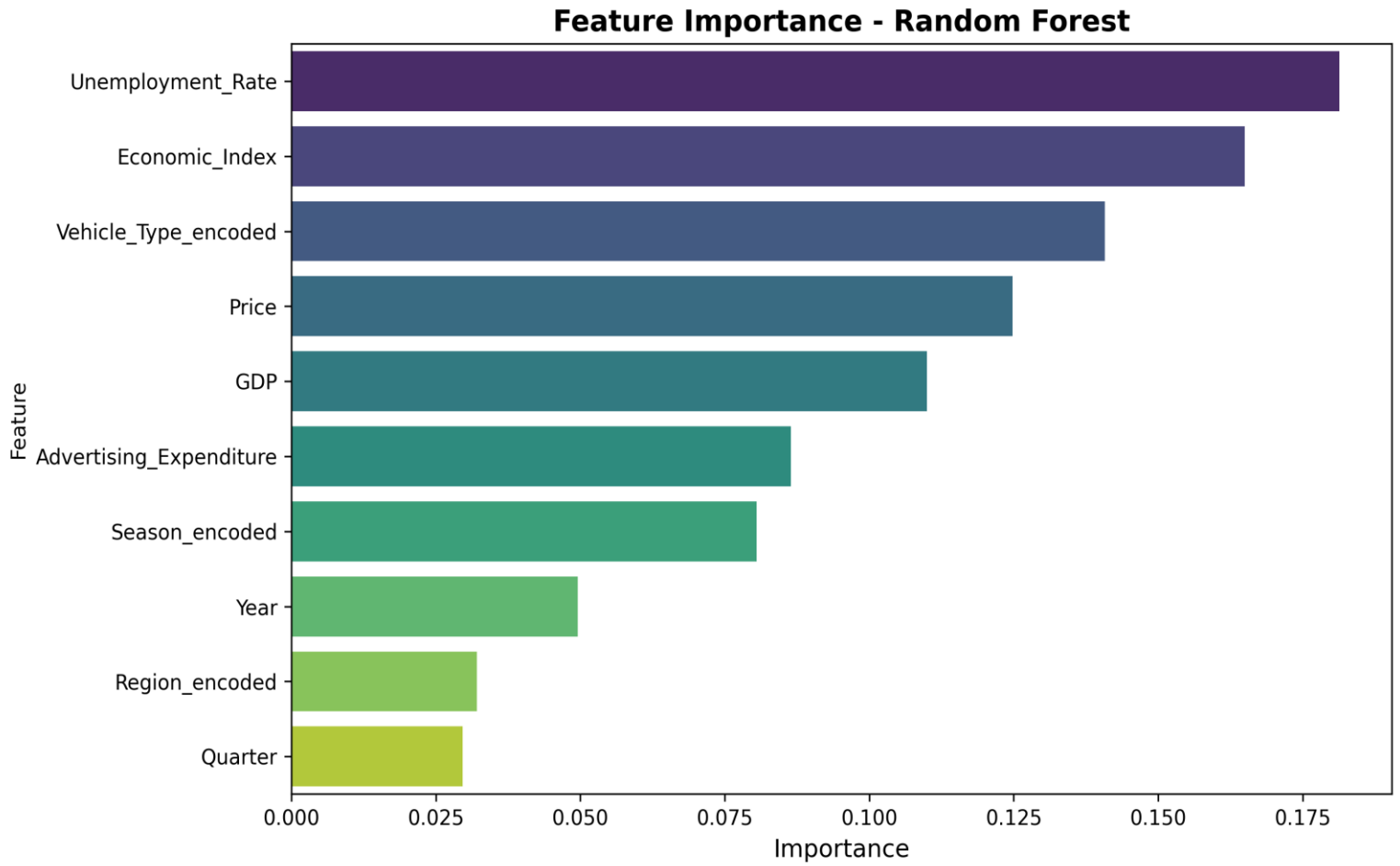
Random Forest: Low false positives and false negatives

Predictive Analysis: ROC Curve



AUC = 0.92 indicates excellent model performance

Predictive Analysis: Feature Importance



Price, GDP, and Advertising are the most important features

Interactive Map with Folium

- Created interactive map showing sales by city location
- Features:
- • Marker clusters for city grouping
- • Color-coded markers (Green: High sales, Red: Low sales)
- • Heatmap overlay showing sales intensity
- • Popup information: Average sales, Total revenue, Popular vehicle type
- • Tooltip showing city name and sales metrics
- Geographic Insights:
- • West Coast cities (LA, SF) show highest sales
- • East Coast cities (NY, Boston) follow closely

Plotly Dash Dashboard

- Interactive web dashboard for real-time data exploration
- Features:
- • Dynamic filtering by Vehicle Type and Region
- • Real-time chart updates based on selections
- • Multiple visualizations:
- - Sales over time (Line chart)
- - Sales by vehicle type (Bar chart)
- - Sales by season (Bar chart with colors)
- - Price vs Sales scatter plot with trendline
- • Summary statistics panel (Total records, Avg sales, Revenue)

Conclusion

- Successfully completed comprehensive automotive sales analysis
- Key Findings:
- • Recession periods cause ~40% sales decline
- • SUV and Electric vehicles show strongest performance
- • Seasonal patterns: Summer peak, Winter low
- • Economic indicators (GDP, Unemployment) significantly impact sales
- • Advertising expenditure positively correlates with sales
- Model Performance:
- • Random Forest achieves 87.5% accuracy
- • High confidence in predicting sales categories

Creativity & Innovative Insights

- Innovative Approaches:
- • Generated realistic synthetic data with complex relationships
- • Combined multiple visualization tools (Matplotlib, Plotly, Folium)
- • SQL-based EDA for structured data exploration
- • Interactive dashboards for stakeholder engagement
- Unique Insights:
- • Economic Index composite metric (GDP + Unemployment)
- • Geographic heatmap reveals regional sales patterns
- • Feature importance analysis guides business decisions
- • Year-over-year growth analysis identifies trends