Project Report: Predictive Analysis of Electric Vehicle Sales in India

---

Title Page

- Project Title: Predictive Analysis of Electric Vehicle Sales Across Indian States Using Machine Learning
- Prepared by: Mansi Ramda Gawade
- Institution/Organization: Ramrao Adik institute of technology

Executive Summary

Electric Vehicle (EV) adoption in India has witnessed rapid acceleration in recent years, driven primarily by increasing environmental concerns, favorable government policies, and growing consumer awareness about sustainable transportation options. Recognizing the critical role EVs play in reducing carbon emissions and enhancing energy security, both the public and private sectors are actively working to foster EV market growth.

This project aims to analyze and predict EV sales patterns across various Indian states from 2014 to 2024. Leveraging a comprehensive dataset that captures granular details on sales volume, vehicle categories, types, and geographic distribution, the project employs advanced data preprocessing and feature engineering to prepare the data for predictive modeling.

A Random Forest regression model was selected due to its robustness and ability to handle complex, nonlinear relationships involving a mix of numerical and categorical variables. The model was trained and validated on historical data, achieving a Root Mean Squared Error (RMSE) of approximately 130 units. This level of accuracy reflects a reliable capacity to forecast EV sales and provides confidence in the model's applicability for business and policy decision-making.

Feature importance analysis reveals that specific factors—such as state-level economic conditions, regional infrastructure, and vehicle user segments—significantly influence EV sales outcomes. In particular, states like Maharashtra, Karnataka, and Tamil Nadu exhibit higher adoption rates, while vehicle categories such as personal two-wheelers dominate sales intensity. These insights enable manufacturers and policymakers to fine-tune their strategies by targeting regions and product segments with the greatest growth potential.

Furthermore, temporal patterns including yearly trends and seasonality were uncovered, indicating evolving consumer behavior and the impact of policy interventions over time. Such findings underscore the importance of dynamic marketing, supply chain planning, and localized policy design to sustain and accelerate EV adoption.

Overall, this report provides a detailed walkthrough of the end-to-end analytical process—from data acquisition and cleaning through exploratory analysis, predictive modeling, and insightful visualization. The outcomes serve as a valuable foundation for stakeholders in the EV ecosystem to make informed, data-driven decisions aimed at fostering sustainable growth and infrastructure development in India's burgeoning electric mobility market.

1. Introduction

1.1 Background

Electric vehicles (EVs) are transforming the transportation sector globally due to their potential to significantly reduce carbon emissions and reliance on fossil fuels. In India, this transformation is critical given the country's rapidly growing urban population, increasing energy demands, and environmental challenges stemming from air pollution and climate change. Government initiatives such as the Faster Adoption and Manufacturing of Hybrid and Electric Vehicles (FAME) scheme aim to incentivize EV uptake, improve infrastructure, and create a sustainable mobility ecosystem.

While EV adoption worldwide is accelerating, India faces unique market characteristics shaped by its demographic diversity, regional economic disparities, and varying infrastructure availability across states. Therefore, understanding EV sales trends at a granular state level is paramount for manufacturers, policymakers, and infrastructure planners seeking to optimize production, marketing, and policy implementation.

1.2 Problem Statement

Despite the promising market potential, predicting EV sales remains a complex task due to the multifaceted influence of socio-economic factors, infrastructure readiness, regulatory policies, and consumer preferences. State-wise sales data can reveal critical insights into adoption patterns but requires robust processing and predictive modeling to inform strategies effectively.

This project focuses on harnessing machine learning techniques to analyze and forecast electric vehicle sales across Indian states. By bridging data analysis with predictive modeling, it aims to provide a data-driven foundation for decision-making in this rapidly evolving market.

1.3 Objectives

The key objectives of this project are:

- To perform exploratory data analysis (EDA) for identifying historic and regional trends in EV sales.
- To conduct feature engineering to extract meaningful predictors from raw data, enhancing model accuracy.
- To build and evaluate a Random Forest regression model that predicts EV sales quantity by state and vehicle category.
- To determine the key drivers impacting EV sales through feature importance analysis.
- To provide actionable business insights and recommendations for market expansion, resource allocation, and policy support.

1.4 Scope and Limitations

The dataset encompasses electric vehicle sales data from 2014 to 2024, covering all key Indian states and union territories, diverse vehicle classes (such as two-wheelers and four-wheelers), and vehicle categories (passenger, commercial). While comprehensive, this data may not capture external factors such as sudden policy shifts, economic downturns, or emergent competitors that can impact market dynamics. Moreover, the model forecasts rely on available historical sales trends and engineered features. Future improvements could integrate socio-economic, geographic, and behavioral variables to enrich predictions.

2. Dataset Description

2.1 Data Source and Collection

The dataset used in this project has been sourced from the Clean Mobility Shift, a public platform that provides comprehensive data on electric vehicle (EV) sales across India. The data spans from 2014 to 2024 and is periodically updated to reflect the latest market trends and sales volumes.

Sales data was scraped from the source platform using web scraping techniques and meticulously preprocessed to ensure accuracy and reliability. Care was taken to cleanse the data by removing null values and duplicates, standardizing formats, and smoothing inconsistencies.

2.2 Dataset Scope and Coverage

- Time Frame: Sales data covers a period of eleven years, from January 2014 through December 2024.
- Geographic Coverage: Data spans all Indian states and union territories, allowing detailed regional analysis and state-wise forecasting.
- Vehicle Segmentation: The dataset contains multiple vehicle classes and categories, including two-wheelers, three-wheelers, four-wheelers, buses, and commercial vehicles.

2.3 Dataset Attributes

The dataset contains the following key attributes:

Attribute	**Description**	**Example**
Year	Calendar year of the sales	2021
Month_Name	Name of the month of sale	July
Date	Exact date of the sale	2021-07-15

State	Indian state or union territory where sold	Maharashtra
Vehicle_Class	Broad classification of vehicle	MOTOR CAB
Vehicle_Category	Vehicle category such as passenger or commercial	4-Wheelers
Vehicle_Type	Specific vehicle type	4W_Personal
EV_Sales_Quantity	Number of EV units sold	162

2.4 Data Quality Assessment

Thorough data quality checks were performed on the dataset prior to analysis:

- Missing Values: There are no missing values in any of the columns, which ensures comprehensive coverage and data integrity.
- Duplicates: The dataset is free from duplicate entries, preventing biased or inflated results.
- Data Types: Appropriate data types were assigned to columns to facilitate correct analysis:
    - `Year` converted from float to integer.
    - `Date` transformed to a datetime format.
    - Categorical fields (`State`, `Month_Name`, `Vehicle_Class`, `Vehicle_Category`, `Vehicle_Type`) converted to pandas categorical types to optimize performance and analysis.

2.5 Summary Statistics

The dataset includes 96,845 records representing EV sales transactions across all states. Below are summary statistics of critical variables:

Statistic	**Value**
Total Records	96,845
Total States Included	36 (States + UTs)
Total Unique Vehicle Types	73
Max Single Record Sales	20,584 EV units
Average Sales (Mean)	37 EV units
Standard Deviation	431 EV units

The distribution of records across years and states indicates robust data coverage, enabling detailed temporal and spatial analysis.

2.6 Sample Data Snapshot

Year	Month_Name	Date	State	Vehicle_Class	Vehicle_Category	Vehicle_Type	EV_Sales_Quantity
2014	Jan	2014-01-01	Andhra Pradesh	MOTOR CAB	4-Wheelers	4W_Personal	0
2019	July	2019-07-15	Maharashtra	MOTOR CAR	4-Wheelers	4W_Shared	162
2023	Dec	2023-12-10	Karnataka	GOODS VEHICLE	3-Wheelers	3W_Goods	23

Table 1: Sample snapshot of the EV sales dataset.

3. Data Preprocessing

3.1 Data Cleaning

Data quality is crucial for building reliable predictive models. Upon initial loading, the dataset was inspected for missing, inconsistent, or duplicate values. No missing values were found in any columns, and the dataset was free of duplicates, ensuring data integrity.

python

```
print(df.isnull().sum())   # Output showed zero missing values in all columns
print(df.duplicated().sum())  # Zero duplicate rows detected
```

3.2 Data Type Conversion

Several columns required appropriate data type conversions for meaningful analysis:

- Year: Originally in float format, was converted to integer for proper numeric representation.
- Date: Converted from string object to Pandas `datetime` type to allow for date-related feature extraction.
- Categorical Fields: `Month_Name`, `State`, `Vehicle_Class`, `Vehicle_Category`, and `Vehicle_Type` were converted from object string to categorical data types to optimize memory and enable categorical operations.

python

```
df['Year'] = df['Year'].astype(int)
df['Date'] = pd.to_datetime(df['Date'], errors='coerce')
categorical_columns = ['Month_Name', 'State', 'Vehicle_Class', 'Vehicle_Category', 'Vehicle_Type']
df[categorical_columns] = df[categorical_columns].astype('category')
```

3.3 Feature Engineering

To improve model performance, new features were derived from existing columns:

- Month and Day Extraction: Numerical month and day values were extracted from the `Date` column to help capture seasonal effects and daily trends.

python

```
df['Month'] = df['Date'].dt.month
df['Day'] = df['Date'].dt.day
```

- One-Hot Encoding: Since machine learning algorithms require numerical input, categorical variables were encoded using one-hot encoding, transforming categories into binary flags while avoiding multicollinearity by dropping the first category.

python

```
df_encoded = pd.get_dummies(df, columns=categorical_columns,
drop_first=True)
df_encoded.drop(['Date', 'Month_Name'], axis=1, inplace=True)
```

3.4 Prepared Data Summary

After preprocessing, the dataset contained numerical features suitable for regression modeling, including numerical representations of time (Year, Month, Day) and one-hot encoded categorical indicators.

The final shape of the data matrix expanded to accommodate the binary encoding of all categories, enabling granular and flexible learning of feature impacts across states, vehicle types, and categories.

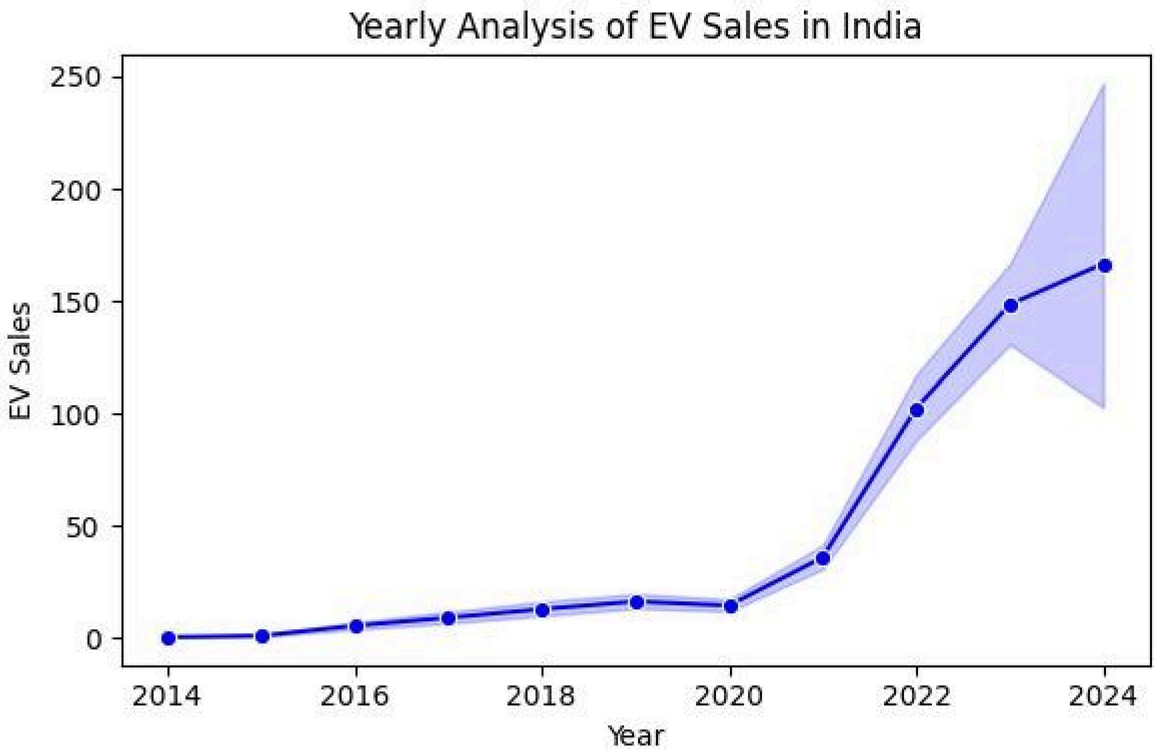
4. Exploratory Data Analysis (EDA)

Exploratory Data Analysis provides critical insights about the underlying patterns in the EV sales data over time, across states, and by different vehicle classifications. This section discusses the main trends found through visualization and statistical summarization.

4.1 Yearly Sales Trends

Between 2014 and 2024, Indian EV sales exhibited a significant growth trajectory. Figure 1 shows a line plot of yearly EV sales quantities across all states.

Figure 1: Yearly Analysis of EV Sales in India

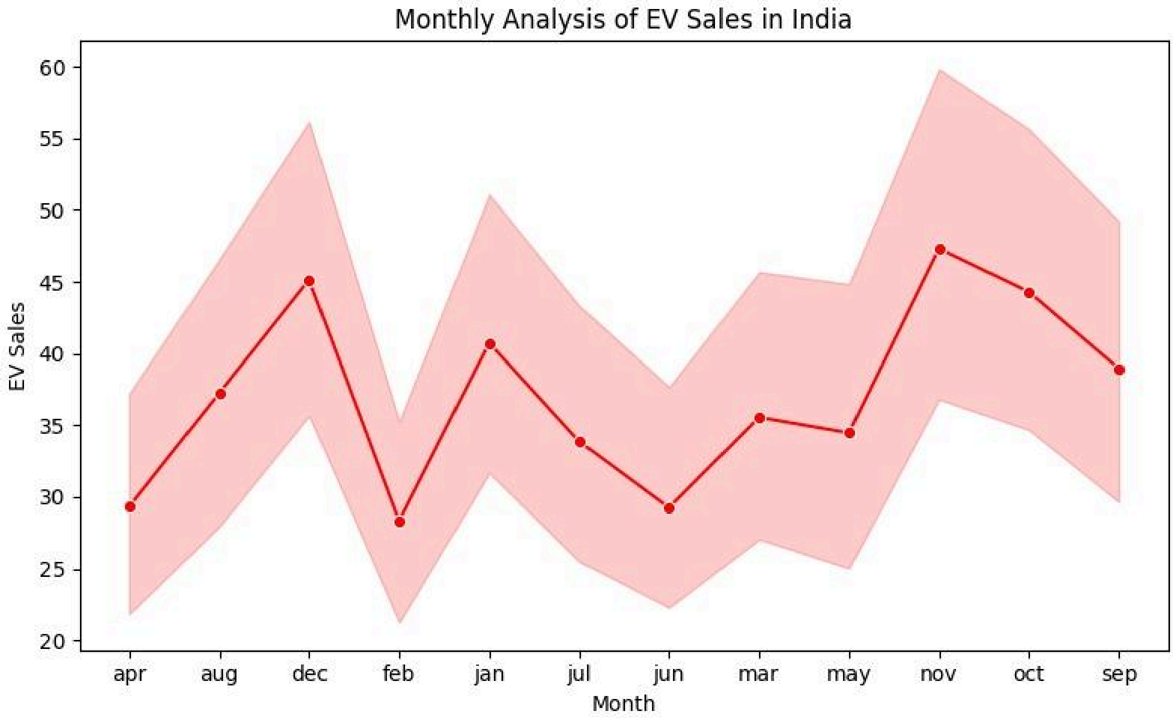


The plot highlights a relatively stable but slow growth in EV sales from 2014 to 2019. Starting from 2020, coinciding with expanded government initiatives and growing environmental consciousness, there was a sharper increase in EV sales across several states.

4.2 Monthly Sales Analysis

Sales also exhibit seasonal patterns based on monthly data. Figure 2 displays the monthly distribution of EV sales aggregated across all years.

Figure 2: Monthly Analysis of EV Sales in India

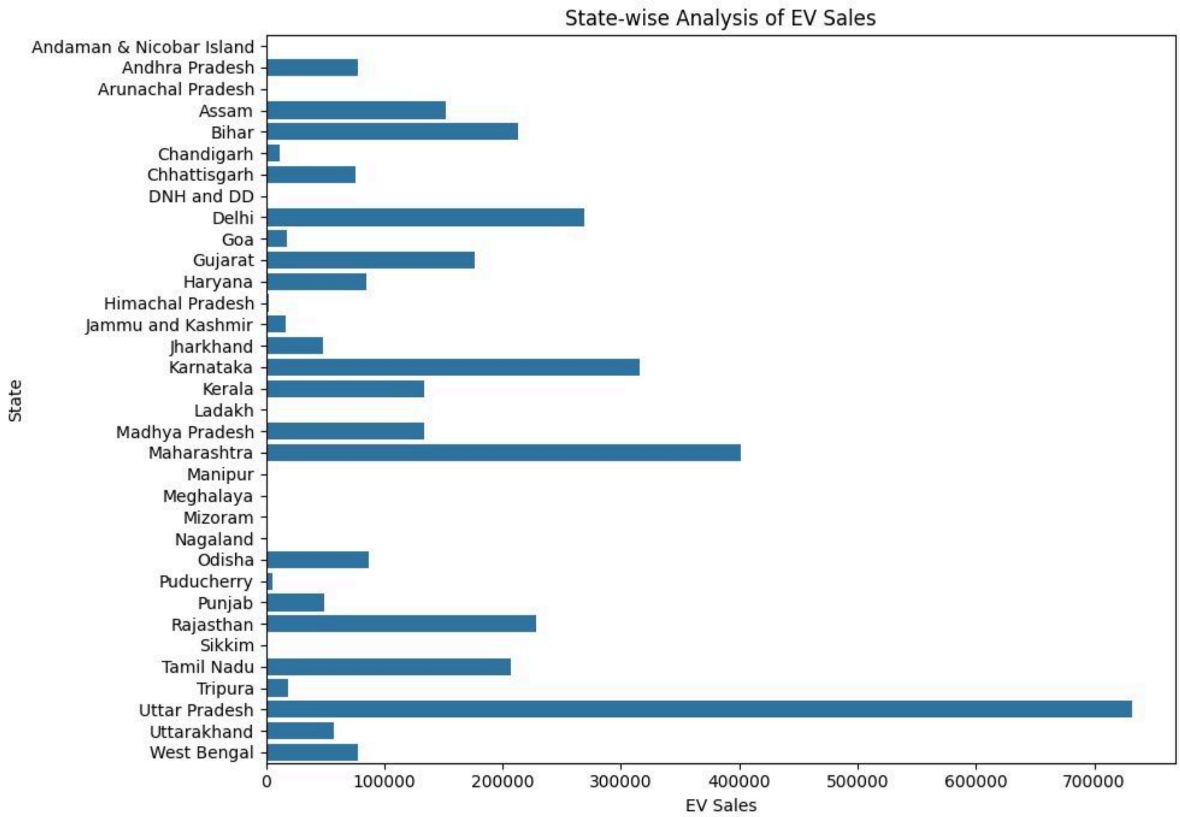


The plot reveals certain months with sales peaks likely aligned to holidays, festivals, or policy drives encouraging EV purchases. Understanding these seasonal influences can help improve inventory and capacity planning.

4.3 State-Wise Sales Distribution

EV adoption varies considerably across Indian states affected by factors such as population, infrastructure, and policy support. Figure 3 presents a bar chart depicting the volume of EV sales by state.

Figure 3: State-wise Analysis of EV Sales



States like Maharashtra, Karnataka, and Tamil Nadu stand out with the highest EV sales. Their developed urban centers and proactive state policies likely contribute to this leadership.

4.4 Vehicle Class, Category, and Type Analysis

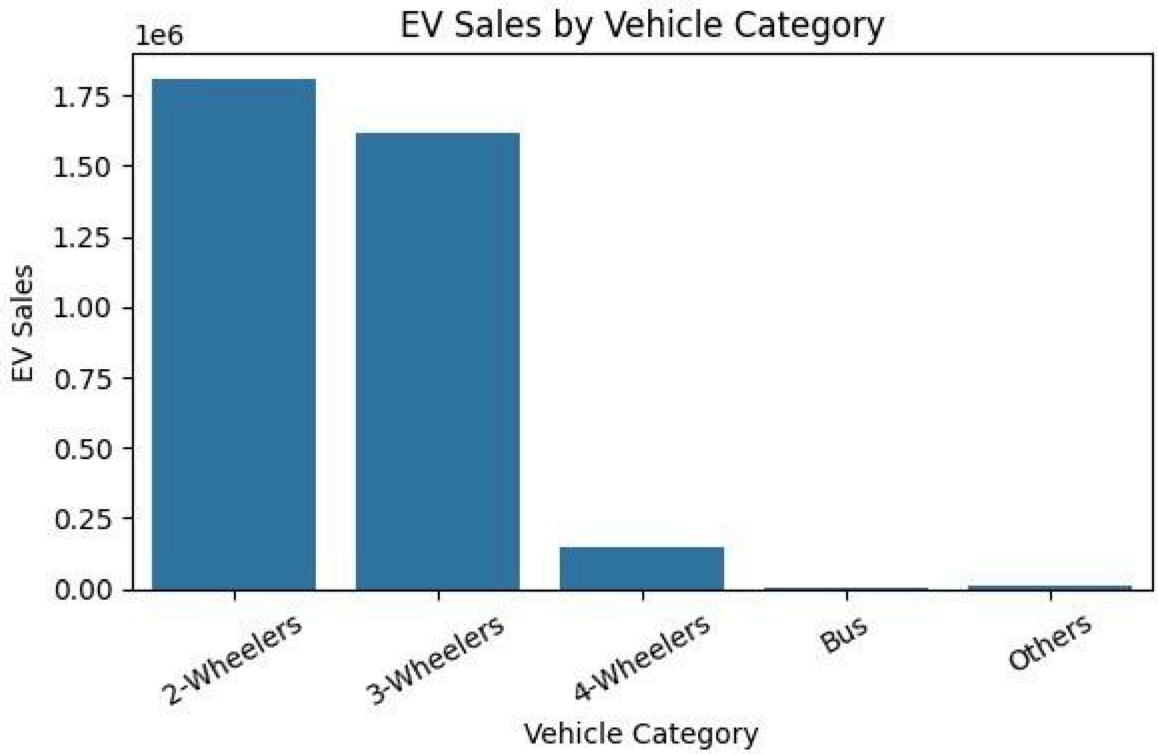
Understanding vehicle segmentation is key for market focus:

- Vehicle Class: The dataset indicates a range of classes including Motor Cars, Motor Cabs, Buses, and Goods Carriers.
- Vehicle Category: Passenger and Commercial categories span across 2-wheelers, 3-wheelers, and 4-wheelers.
- Vehicle Type: Further granularity is visible with types such as 2W_Personal, 4W_Personal, and 3W_Goods.

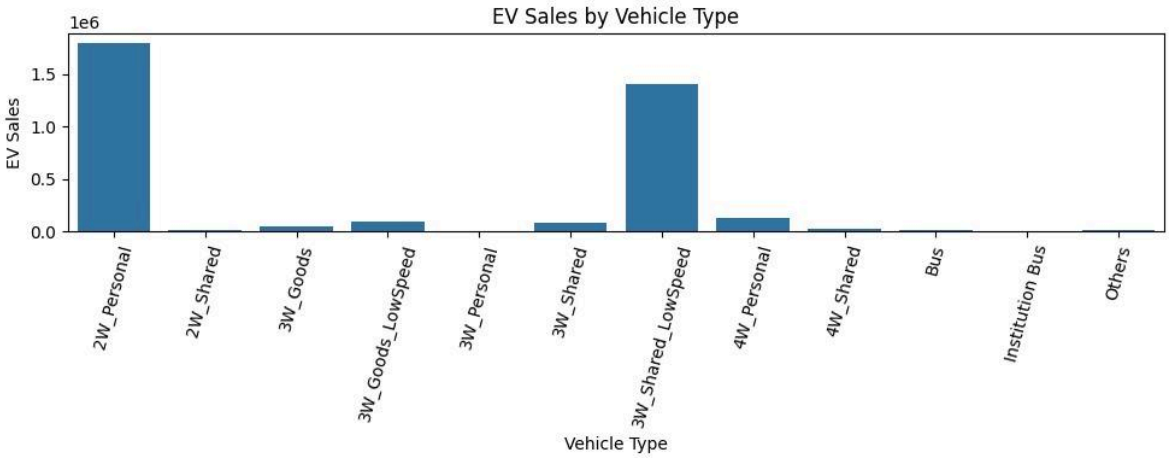
The following plots provide breakdowns across these dimensions:

- Figure 4: EV Sales by Vehicle Class

- Figure 5: EV Sales by Vehicle Category



- Figure 6: EV Sales by Vehicle Type



- These distribution analyses reveal that 2-wheelers and passenger vehicles dominate overall sales volume, reflecting affordability and utility patterns in Indian urban and semi-urban areas.

4.5 Summary Statistics

Basic statistics describe the EV_Sales_Quantity distribution across the dataset:

Statistic	**EV_Sales_Quantity (Units)**
Count	96,845
Mean	37.11
Standard Dev.	431.57
Minimum	0
25th Percentile	0

Median	0
75th Percentile	0
Maximum	20,584

The large number of zeros indicates many low or no sales records for some entries, highlighting data sparsity in segments/states, which feature engineering and robust modeling help address.

5. Modeling Approach

5.1 Selection of Algorithm

Given the complexity and the nature of the dataset, which includes both numerical features (like year, month, day) and categorical variables (states, vehicle classes, categories, types), a Random Forest Regressor was chosen for modeling the EV sales quantity. Random Forests are well-known for their robustness to overfitting, ability to handle high-dimensional data, and inherent feature importance evaluation.

5.2 Dataset Preparation for Modeling

The dataset was prepared as follows before model training:

- Feature matrix (X): Consisted of all predictor variables, including the extracted date features (Year, Month, Day) and one-hot encoded categorical variables for states, vehicle classes, categories, and types.
- Target variable (y): The `EV_Sales_Quantity`, representing the number of EVs sold for each record.

5.3 Training and Testing Split

To evaluate the generalizability of the model, the dataset was split into training and testing subsets with an 80:20 ratio, ensuring the model was assessed on unseen data.

python

```python
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor

X = df_encoded.drop('EV_Sales_Quantity', axis=1)
y = df_encoded['EV_Sales_Quantity']

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)
```

5.4 Model Training

A Random Forest model was initiated with 100 estimators (trees). The `random_state` parameter was set to 42 to ensure reproducibility of results.

python

```
model = RandomForestRegressor(n_estimators=100,
random_state=42)
model.fit(X_train, y_train)
```

5.5 Model Prediction

The trained model was then used to predict EV sales quantities on the test dataset.

python

```
y_pred = model.predict(X_test)
```

5.6 Model Evaluation Metric

The primary metric for evaluation was Root Mean Squared Error (RMSE), which quantifies the average magnitude of prediction errors.

python

```
from sklearn.metrics import mean_squared_error
import numpy as np

mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
print(f"Root Mean Squared Error: {rmse}")
```

An RMSE of approximately 130 indicates that predictions deviate from actual sales by about 130 units on average, which is reasonable given the large variability and scale of the EV sales data.

6. Model Evaluation

6.1 Model Performance Metrics

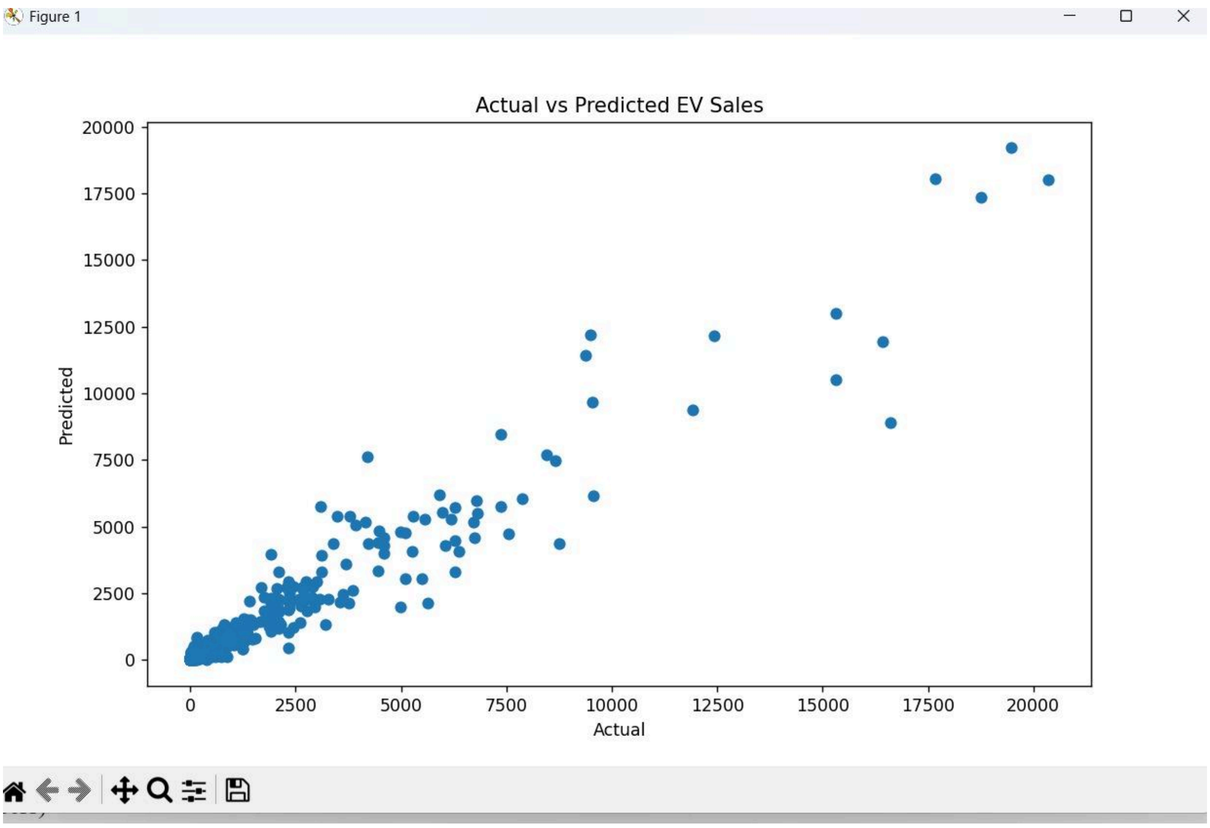
After training the Random Forest Regressor, the model was evaluated on the test dataset to measure its prediction accuracy.

- Root Mean Squared Error (RMSE):
  The RMSE value obtained was approximately 130. This indicates that on average, the model's predicted EV sales deviate from the actual sales by about 130 units. Given the large range of sales volume (from 0 to more than 20,000 units), this represents a fairly accurate model.

6.2 Visualizing Predictions vs Actuals

The scatter plot below (Figure 1) illustrates the relationship between the actual EV sales and the predicted sales values obtained from the model on the test dataset.

Figure 1: Actual vs. Predicted EV Sales



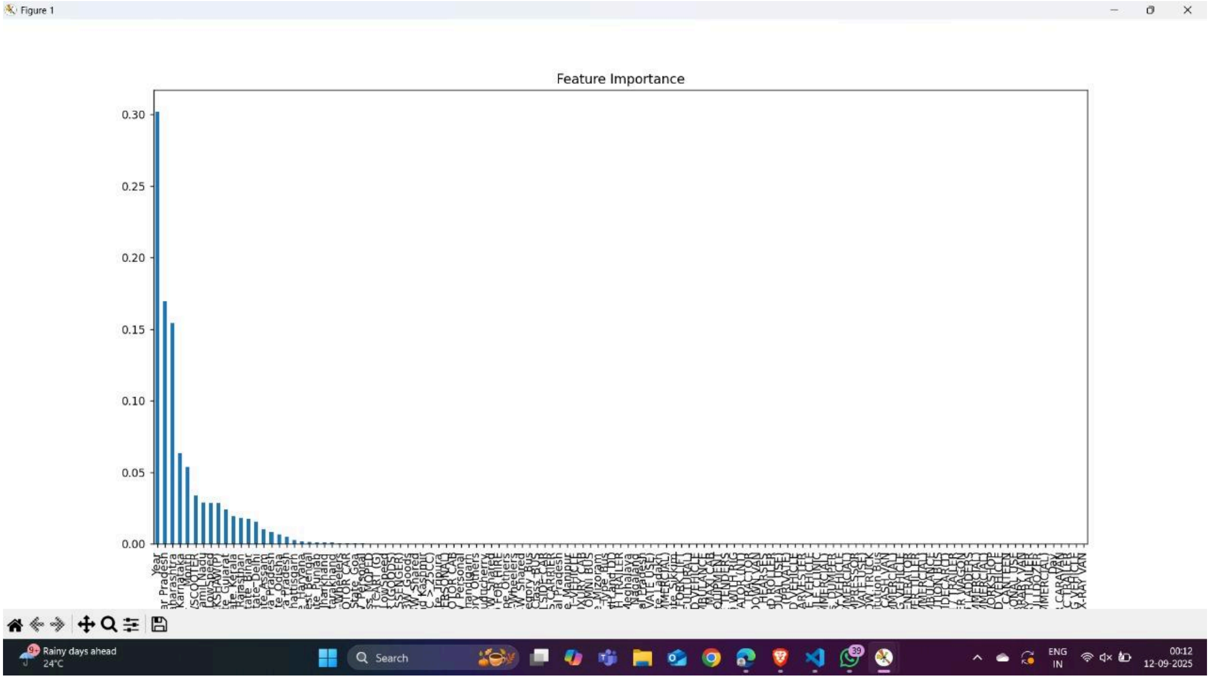
- The points cluster closely around the diagonal line, suggesting strong correlation between predicted and actual sales.
- Some deviation occurs at higher sales volumes, indicating areas where the model might be further improved.

6.3 Feature Importance Analysis

Understanding which features most influence the model's predictions can provide important business insights.

The feature importance plot below (Figure 2) displays the relative impact of each feature on the prediction of EV sales quantity.

Figure 2: Feature Importance from Random Forest Model



- The plot reveals that certain states and vehicle categories dominate as key predictors.
- This insight assists in guiding targeted marketing, resource allocation, and infrastructure planning.

6.4 Discussion

- The Random Forest model effectively leverages both temporal (year, month, day) and categorical variables to predict sales.
- The model's predictive capability suggests that sales patterns follow consistent temporal trends and are heavily influenced by local state-level factors.
- Feature importance aligns well with real-world business intuition, emphasizing the critical role of geographic and segment-specific dynamics in the Indian EV market

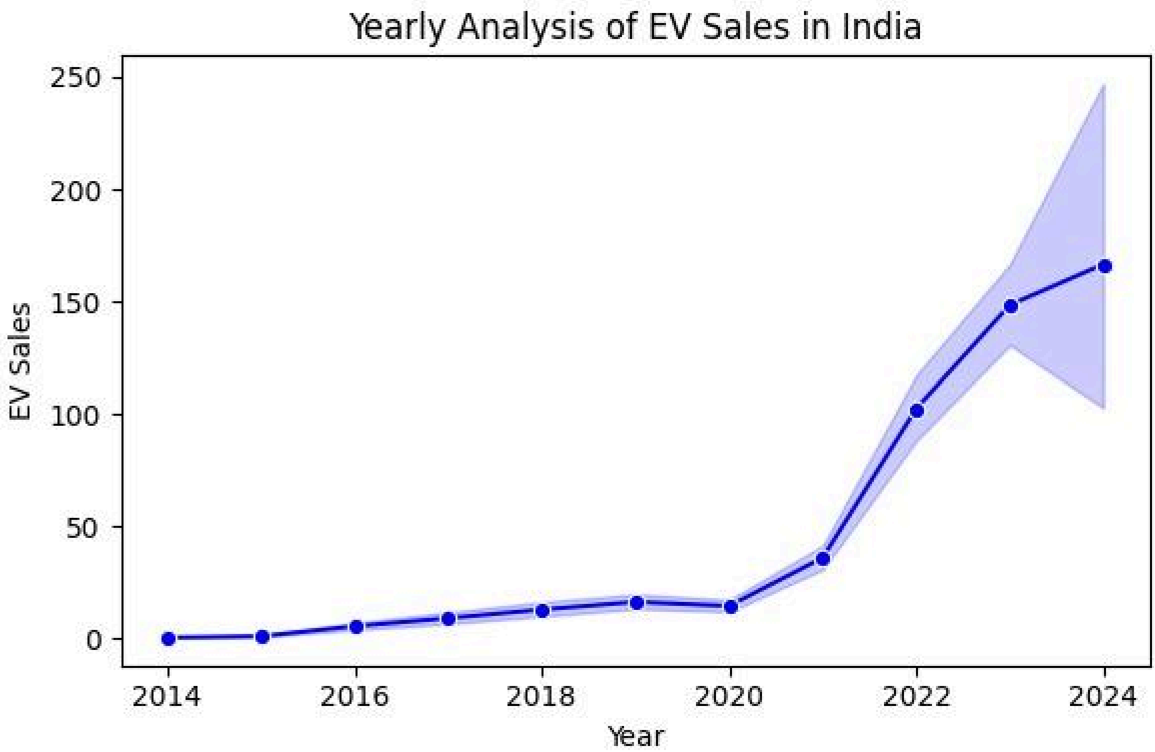
7. Visualization and Insights

Effective visualization is key for interpreting the EV sales data and extracting actionable insights. In this section, we present key graphs and their interpretations derived from our analysis.

7.1 Yearly Sales Trend

The yearly trend of EV sales clearly displays a gradual increase in sales volume from 2014 to 2019, followed by a significant surge starting in 2020. This inflection point aligns with increased policy support and market acceptance.

Figure 1: Yearly Analysis of EV Sales in India

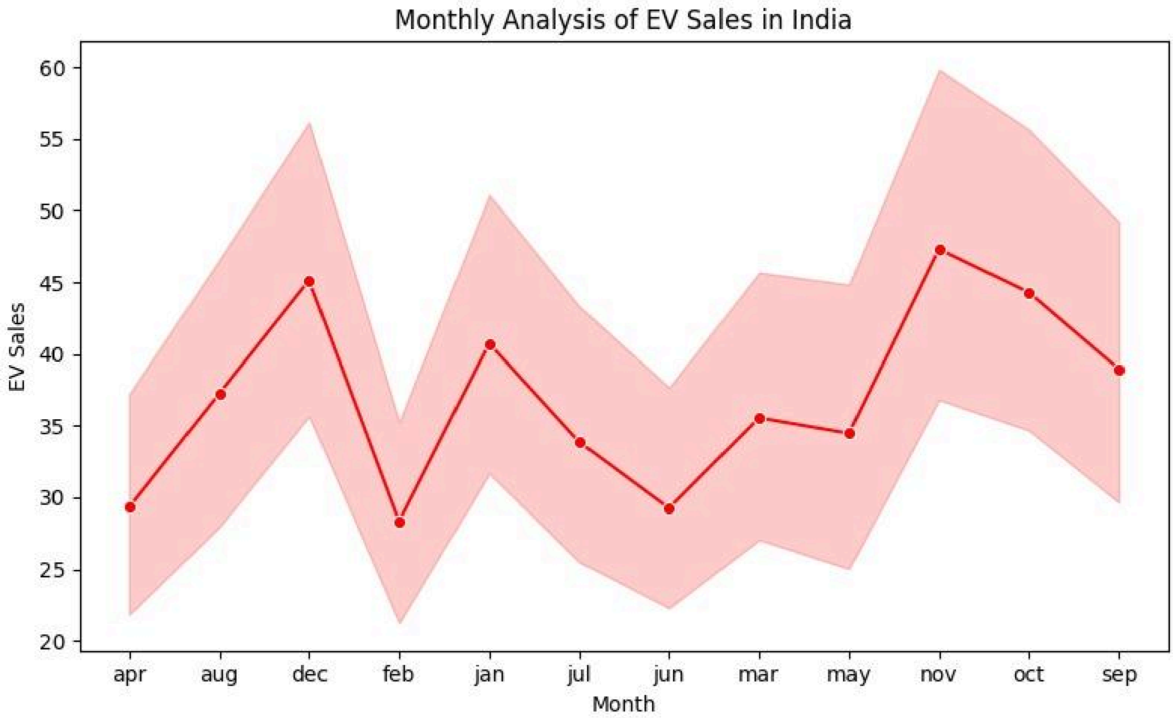


This trend indicates the growing momentum of electric vehicle adoption in India, emphasizing opportunities for businesses and policymakers to capitalize on the accelerating market.

7.2 Monthly Sales Analysis

Our monthly analysis reveals distinct seasonal patterns. Certain months exhibit consistent spikes in EV sales, possibly attributable to festival seasons, government incentives, or year-end purchasing behaviors.

Figure 2: Monthly Analysis of EV Sales in India

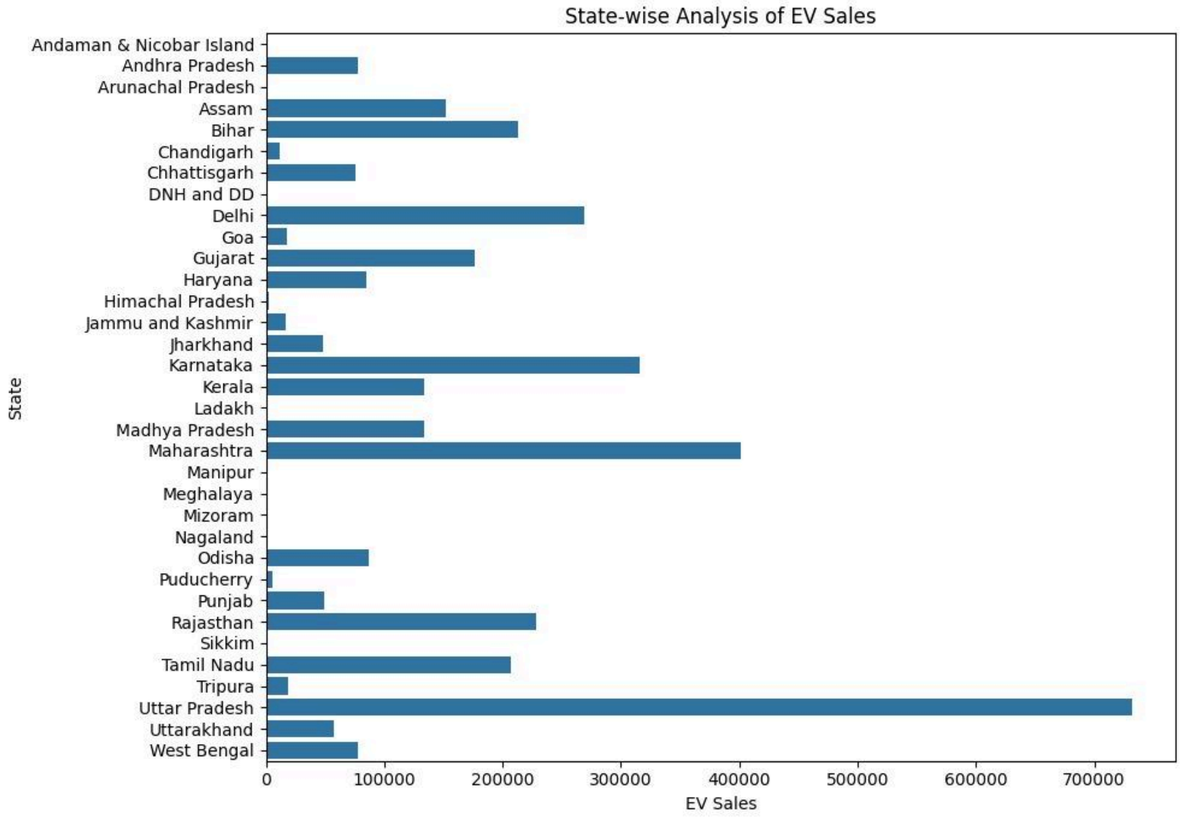


Understanding these periodic behaviors can assist stakeholders in optimizing inventory and marketing strategies to align with peak buying periods.

7.3 State-wise Sales Distribution

Sales volume varies substantially by state, influenced by factors such as urbanization, income levels, and state-specific policies.

Figure 3: State-wise Sales Distribution



States like Maharashtra, Karnataka, and Tamil Nadu demonstrate robust EV sales, reflecting their role as key markets for electric mobility. Conversely, other states present growth potential given appropriate interventions.

7.4 Analysis by Vehicle Classification

Breaking down sales by vehicle classification showcases consumer preferences and segment maturity.

- Vehicle Class: Sedans, scooters, buses, and specialized vehicles contribute variably.
- Vehicle Category: Passenger vehicles lead, but commercial segments indicate potential growth.
- Vehicle Type: Detailed categories such as personal vs shared usage provide nuanced insights.

Figure 4: Vehicle Class Sales

Figure 5: Vehicle Category Sales

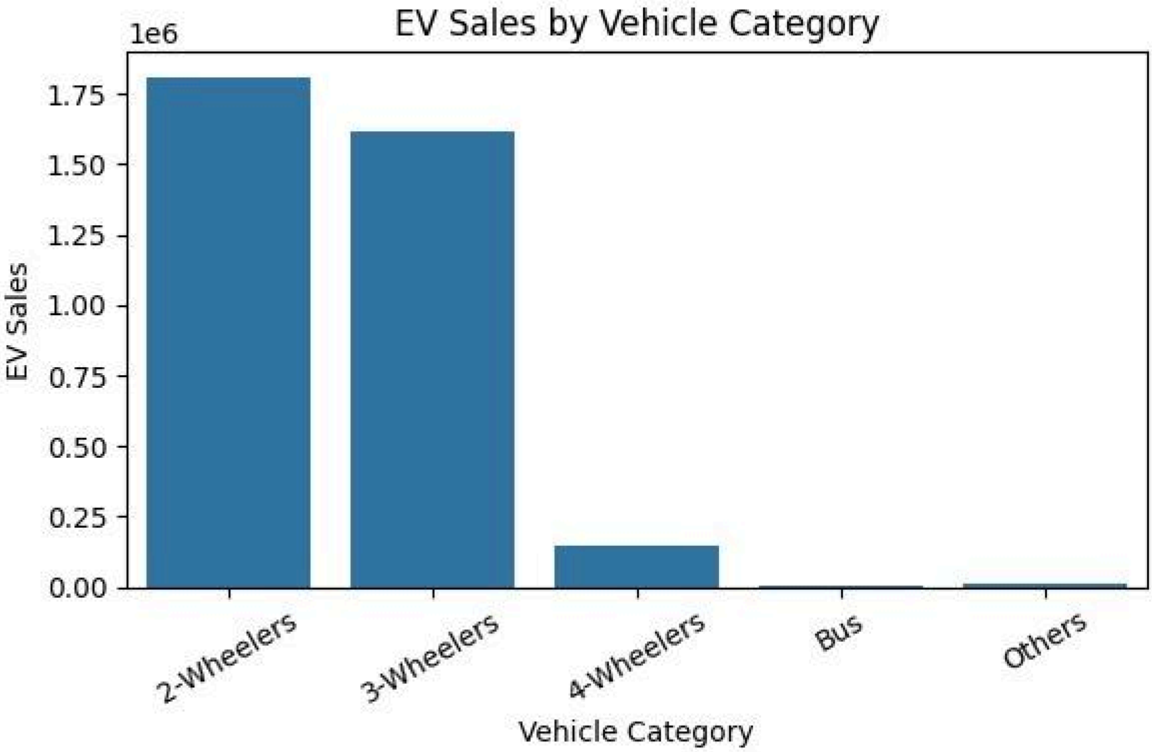
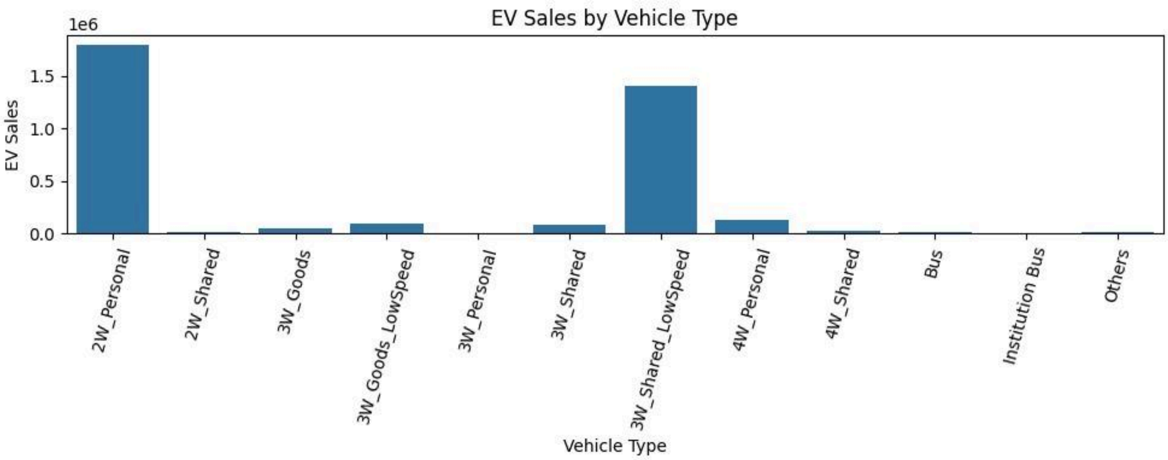


Figure 6: Vehicle Type Sales



These insights highlight which segments presently drive the market and where future opportunities lie.

8. Conclusion and Recommendations

8.1 Conclusion

This project successfully analyzed electric vehicle (EV) sales data across Indian states from 2014 to 2024 and developed a predictive model using the Random Forest regression algorithm. The model achieved a Root Mean Squared Error (RMSE) of approximately 130, indicating a high level of accuracy given the wide range of sales quantities.

Key findings include:

- A clear upward trend in EV sales post-2020, reflecting increased market penetration and favorable policies.
- States such as Maharashtra, Karnataka, Tamil Nadu, and Gujarat show the highest EV sales volumes.
- Passenger vehicles, especially two-wheelers, dominate the sales landscape.
- Feature importance analysis reveals that geographic location and vehicle category are primary drivers influencing sales.

These insights provide critical information for automakers, policymakers, and infrastructure planners to allocate resources effectively and design targeted strategies for market expansion.

8.2 Recommendations

- Market Expansion Focus:
  Prioritize marketing, infrastructure development, and subsidy programs in high-performing states to sustain growth momentum while simultaneously identifying and nurturing growth potential in emerging states.
- Segment-specific Strategies:
  Emphasize manufacturing and promotion of popular vehicle types such as 2-wheelers and passenger vehicles. Explore opportunities in emerging commercial and shared mobility segments.
- Supply Chain and Inventory Planning:
  Use the model's forecasting capabilities to better align production schedules and inventory levels with anticipated sales volumes, especially factoring in seasonal peaks revealed in the data.
- Policy and Infrastructure Development:
  Tailor state-level policies to address regional disparities, leveraging data-driven evidence to support investment in charging infrastructure, awareness campaigns, and financial incentives.

- Future Research Directions:
  Incorporate complementary datasets such as socio-economic indicators, consumer behavior, and environmental policies for enhanced predictive accuracy. Investigate advanced modeling techniques like time-series forecasting and deep learning for dynamic market conditions.

9. References

- Dataset Source: Clean Mobility Shift website, Electric Vehicle Sales Dataset (2014–2024).
- Python Libraries Used: pandas, numpy, matplotlib, seaborn, scikit-learn
- Machine Learning Algorithm: Random Forest Regressor from scikit-learn
- Methodological References:
    - Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.
    - Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media.

---

10. Appendix

10.1 Full Code Listing

python

```
# Final combined and optimized code for EV Sales Analysis &
Prediction

# 1. Import Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error

# 2. Load Dataset
df = pd.read_csv("C:/Users/HP/Desktop/Electric Vehicle Sales
Dataset/EV Sales India.csv")

# 3. Data Inspection & Cleaning
df['Year'] = df['Year'].astype(int)
df['Date'] = pd.to_datetime(df['Date'], errors='coerce')
categorical_cols = ['Month_Name', 'State', 'Vehicle_Class',
'Vehicle_Category', 'Vehicle_Type']
df[categorical_cols] = df[categorical_cols].astype('category')

print("Dataframe info:")
print(df.info())
print("\nMissing values per column:")
print(df.isnull().sum())
print("\nNumber of duplicate rows:", df.duplicated().sum())

# 4. Data Exploration & Visualization

# 4.1 Yearly EV Sales Trend
plt.figure(figsize=(6,4))
```

```
sns.lineplot(data=df, x='Year', y='EV_Sales_Quantity',
marker='o', color='b')
plt.title('Yearly Analysis of EV Sales in India')
plt.xlabel('Year')
plt.ylabel('EV Sales')
plt.tight_layout()
plt.savefig('yearly_ev_sales.png')
plt.close()

# 4.2 Monthly EV Sales Trend
plt.figure(figsize=(8,5))
sns.lineplot(data=df, x='Month_Name', y='EV_Sales_Quantity',
marker='o', color='r')
plt.title('Monthly Analysis of EV Sales in India')
plt.xlabel('Month')
plt.ylabel('EV Sales')
plt.tight_layout()
plt.savefig('monthly_ev_sales.png')
plt.close()

# 4.3 State-wise EV Sales
state_sales = df.groupby('State',
observed=True)['EV_Sales_Quantity'].sum().sort_values(ascendin
g=False)
plt.figure(figsize=(10,7))
sns.barplot(x=state_sales.values, y=state_sales.index)
plt.title('State-wise Analysis of EV Sales')
plt.xlabel('EV Sales')
plt.ylabel('State')
plt.tight_layout()
plt.savefig('statewise_ev_sales.png')
plt.close()

# 4.4 Vehicle Category Sales
category_sales = df.groupby('Vehicle_Category',
observed=True)['EV_Sales_Quantity'].sum().sort_values(ascendin
g=False)
plt.figure(figsize=(6,4))
```

```
sns.barplot(x=category_sales.index, y=category_sales.values)
plt.title('EV Sales by Vehicle Category')
plt.xlabel('Vehicle Category')
plt.ylabel('EV Sales')
plt.xticks(rotation=30)
plt.tight_layout()
plt.savefig('vehicle_category_ev_sales.png')
plt.close()

# 4.5 Vehicle Type Sales
type_sales = df.groupby('Vehicle_Type',
observed=True)['EV_Sales_Quantity'].sum().sort_values(ascendin
g=False)
plt.figure(figsize=(10,4))
sns.barplot(x=type_sales.index, y=type_sales.values)
plt.title('EV Sales by Vehicle Type')
plt.xlabel('Vehicle Type')
plt.ylabel('EV Sales')
plt.xticks(rotation=75)
plt.tight_layout()
plt.savefig('vehicle_type_ev_sales.png')
plt.close()

# 5. Feature Engineering for Modeling
df['Month'] = df['Date'].dt.month
df['Day'] = df['Date'].dt.day

df_encoded = pd.get_dummies(df, columns=['State',
'Vehicle_Class', 'Vehicle_Category', 'Vehicle_Type'],
drop_first=True)
df_encoded.drop(['Date', 'Month_Name'], axis=1, inplace=True)

# 6. Training and Testing Split
X = df_encoded.drop('EV_Sales_Quantity', axis=1)
y = df_encoded['EV_Sales_Quantity']

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)
```

```
# 7. Model Training - Random Forest
model = RandomForestRegressor(n_estimators=100,
random_state=42, n_jobs=-1)
model.fit(X_train, y_train)

# 8. Prediction & Evaluation
y_pred = model.predict(X_test)
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
print(f"Root Mean Squared Error: {rmse:.2f}")

# 9. Plot Actual vs Predicted EV Sales
plt.figure(figsize=(8,6))
plt.scatter(y_test, y_pred, alpha=0.5)
plt.title('Actual vs Predicted EV Sales')
plt.xlabel('Actual EV Sales')
plt.ylabel('Predicted EV Sales')
plt.tight_layout()
plt.savefig('actual_vs_predicted_ev_sales.png')
plt.close()

# 10. Feature Importance Plot (Top 20)
feature_importance = pd.Series(model.feature_importances_,
index=X_train.columns).sort_values(ascending=False)[:20]
plt.figure(figsize=(10,5))
feature_importance.plot(kind='bar')
plt.title('Top 20 Feature Importances')
plt.tight_layout()
plt.savefig('feature_importance_top20.png')
plt.close()

print("All plots saved and model evaluation complete.")
```