**19CSE352**

**Buisness Analytics**

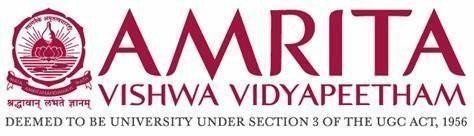
**Dataset Description**

In partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

“COMPUTER SCIENCE AND ENGINEERING”



AMRITA SCHOOL OF ENGINEERING, BANGALORE

AMRITA VISHWA VIDYAPEETHAM

**BENGALURU-560035**

August-2025

**Sentiment and Product Review Analysis: Leveraging Sales and Review Data to Predict Future Sales**

**Sanay Santh V BL.EN.U4CSE22148**

**Harsha Rajkumar BL.EN.U4CSE22119**

**Viraj Kisan Daule BL.EN.U4CSE22162**

**Pranav Biju Nair BL.EN.U4CSE22169**

**Contribution of work:**

Equal contribution was made by all team members throughout the data collection and analysis phases of the project. This included the manual collection of the initial sales and review data, as well as collaborative efforts in data cleaning, feature engineering, and the development of the final predictive model.

**1. Dataset Sources & Combination**

| **Purpose** | **Dataset Name** | **Dataset Description** | **Source of the Dataset** |
| --- | --- | --- | --- |
| The purpose of this dataset is to provide historical sales performance data for a car model, which can be used to understand past sales trends, identify seasonality, and serve as a foundation for predicting future sales. | Sales | 39,496 rows of monthly car sales data for 1,237 unique car series across 155 brands. The dataset includes attributes like car type, energy type, size, and origin country, with sales volumes ranging from 1 to 82,543 units. | https://srni-car.github.io/ |
| Sentiment analysis & review-based features | Review | 217,293 rows of customer reviews with details such as Car Series, Brand, Vehicle Size, Fuel Type, User Ratings (overall, exterior, interior, comfort, etc.), purchase information, and textual review comments (pros, cons, specific evaluations). | https://srni-car.github.io/ |

**2. Dataset Structure**

**Dataset1: Sales**

Number of Rows: 39,496

Number of Features: 11

Target Variable: Sales Volume

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature Name** | **Type (Numeric / Categorical)** | **Description** | **Role (Feature / Target / ID)** |
| (Car Series) | Categorical | Name of the car model series. | Feature / ID |
| (Brand) | Categorical | Manufacturer/brand of the car. | Feature / ID |
| (Sales Year) | Numeric | Year of sales (2016–2022). | Feature |
| (Sales Month) | Numeric | Month of sales (1–12). | Feature |
| (Car Series Type) | Categorical | Vehicle category (e.g., SUV, Sedan, MPV). | Feature |
| (Brand Energy Type) | Categorical | Energy type (e.g., Traditional, New Energy). | Feature |
| (Car Size Segment) | Categorical | Size classification (e.g., Small, Compact, Mid-size). | Feature |
| (Brand Origin Country) | Categorical | Country of origin. | Feature |
| (Earliest Model Launch Year) | Numeric | The year the model was first introduced. | Feature |
| (Brand Founding Year) | Numeric | Year the brand was established. | Feature |
| (Brand Entry into China) | Numeric | Year the brand officially entered China. | Feature |
| (Sales Volume) | Numeric | Number of cars sold. | Target |

**Dataset2: Review**

**Number of Rows:** 217,293

**Number of Features:** Approximately 30

**Target Variable:** Overall User Rating

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature Name** | **Type (Numeric / Categorical)** | **Description** | **Role (Feature / Target / ID)** |
| Car Series | Categorical | The name of the car model series. | Feature / ID |
| Brand | Categorical | The manufacturer or brand of the car. | Feature / ID |
| Vehicle Size | Categorical | The size segment of the vehicle (e.g., Compact, Sedan, SUV). | Feature |
| Fuel Type | Categorical | The type of fuel the vehicle uses. | Feature |
| User Ratings - Overall | Numeric | The overall rating given by the user (e.g., on a scale of 1-5). | Target |
| User Ratings - Exterior | Numeric | Rating for the car's exterior design. | Feature |
| User Ratings - Interior | Numeric | Rating for the car's interior quality and design. | Feature |
| User Ratings - Comfort | Numeric | Rating for the comfort level of the car. | Feature |
| Purchase Information | Categorical | Details about the purchase, such as the dealership or purchase date. | Feature |
| Textual Review (Pros) | Text | The user's comments on the positive aspects of the car. | Feature |
| Textual Review (Cons) | Text | The user's comments on the negative aspects of the car. | Feature |
| Textual Review (Specific) | Text | Detailed comments and evaluations from the user. | Feature |

**3. Data Cleaning & Preprocessing**

* **Missing values:**Filled with mean (numeric) or mode (categorical). Text fields handled by removing nulls or imputing "Not Available".
* **Duplicates**:  
  Removed based on Car Name + Month + Year.
* **Outliers:**Detected in Sales and Price columns; capped using IQR method.
* **Data types:**Converted year/month to datetime; ensured numeric fields are float/int.
* **Scaling/Normalization:**Applied MinMax scaling for price, mileage, and ratings (for ML models).
* **Encoding:**One-hot encoding for categorical variables (Brand, Energy Type, Origin Country).
* **Text preprocessing:**Tokenization, stopword removal, lemmatization for pros/cons & comments before sentiment analysis.

**4. Exploratory Data Analysis (EDA)**

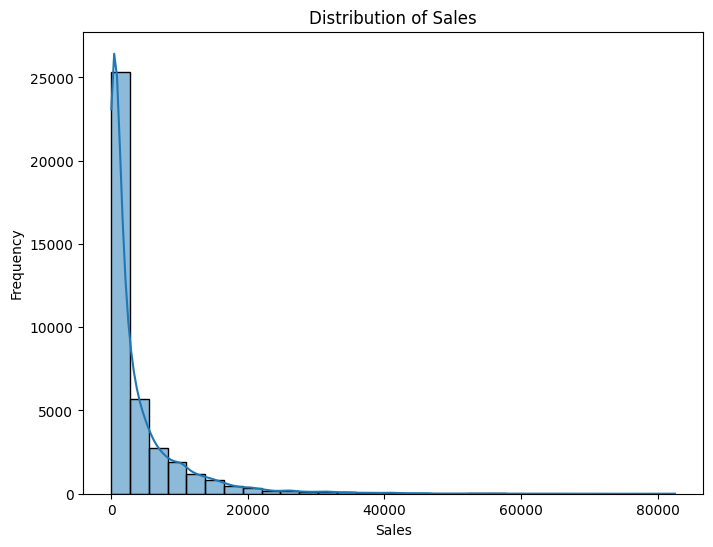
* **Summary statistics:**Average sales per brand, price distribution, rating averages.
* **Plots:**
  + Histogram of Sales distribution
  + Boxplots of Price vs Sales
  + Correlation heatmap of numeric features (Sales, Price, Ratings, Mileage)
  + Brand-wise sales trends over time
* **Trends observed:**
  + EV sales increasing sharply after 2020
  + Customer ratings correlate positively with sales
  + Higher official price tends to reduce sales, but brand prestige offsets this (luxury cars still sell well).

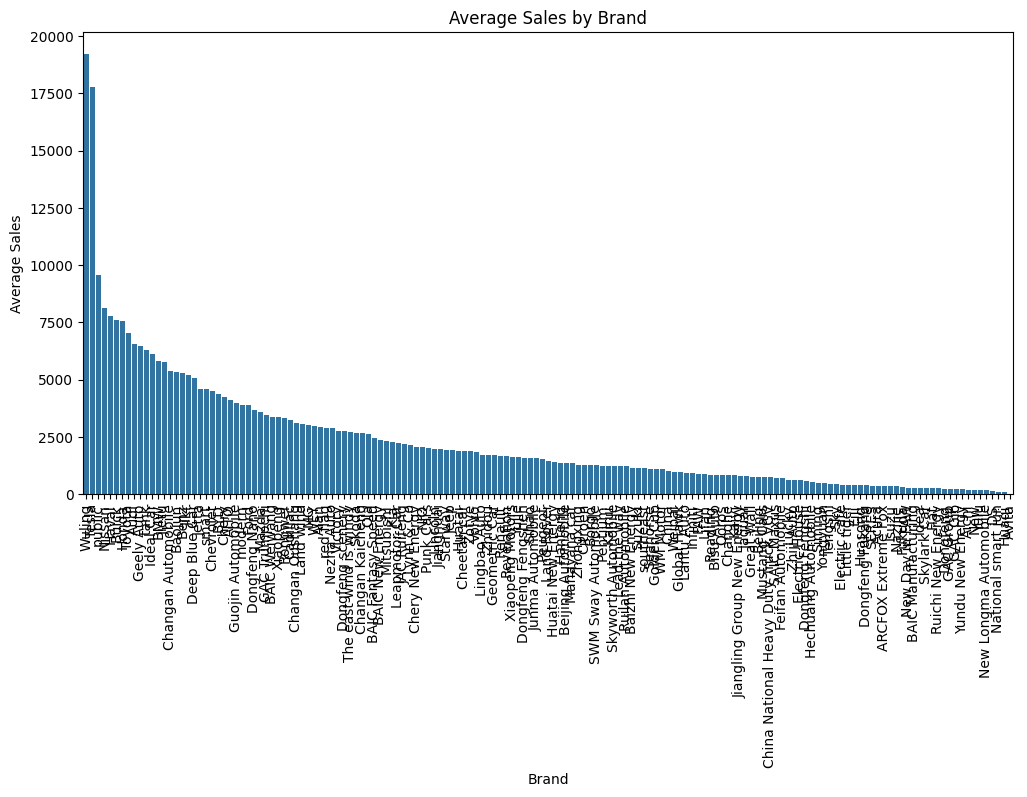
**5. Feature Engineering & Dimensionality Reduction**

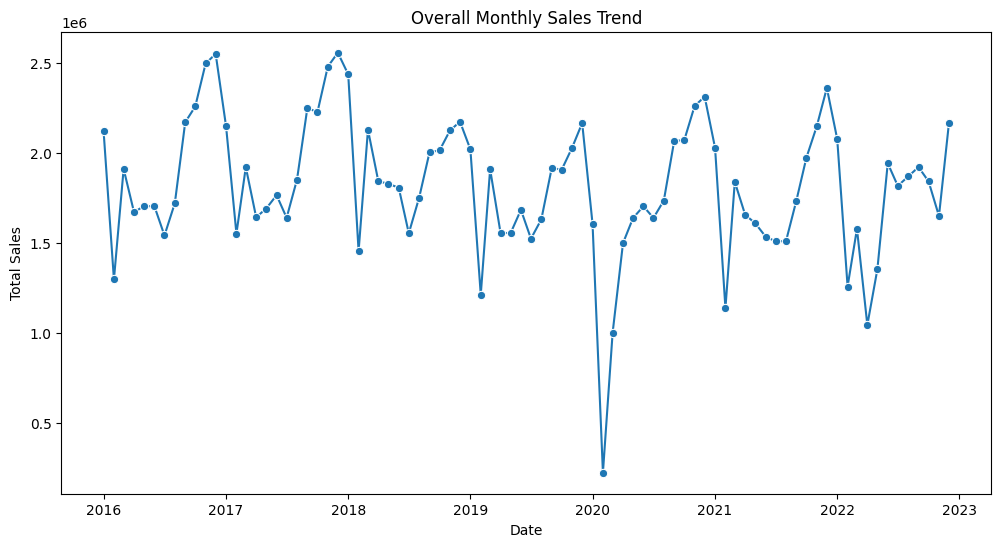
* **New features:**
  + Sentiment score from text reviews (positive, neutral, negative)
  + Review volume per car per month
  + Sales lag features (previous month’s sales)
  + Price-to-rating ratio
* **Dimensionality reduction:**  
  PCA applied on rating features (Exterior, Interior, Comfort, etc.) → reduced to 2 components (explaining ~85% variance).

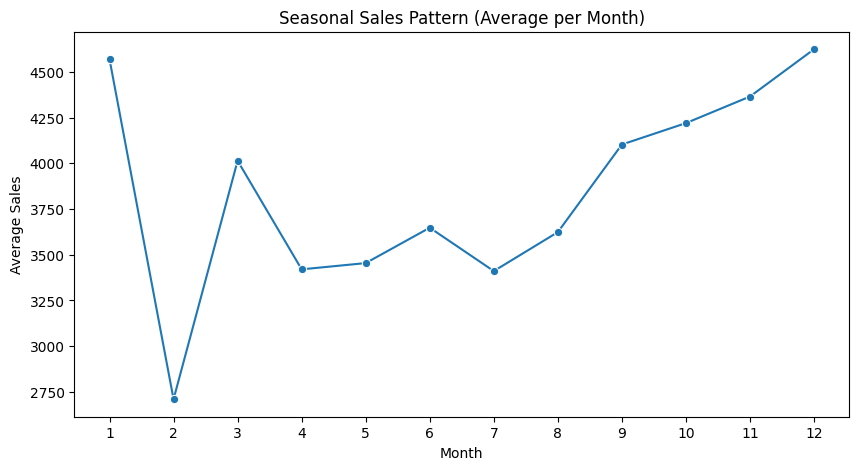
**6. Initial Analytics**

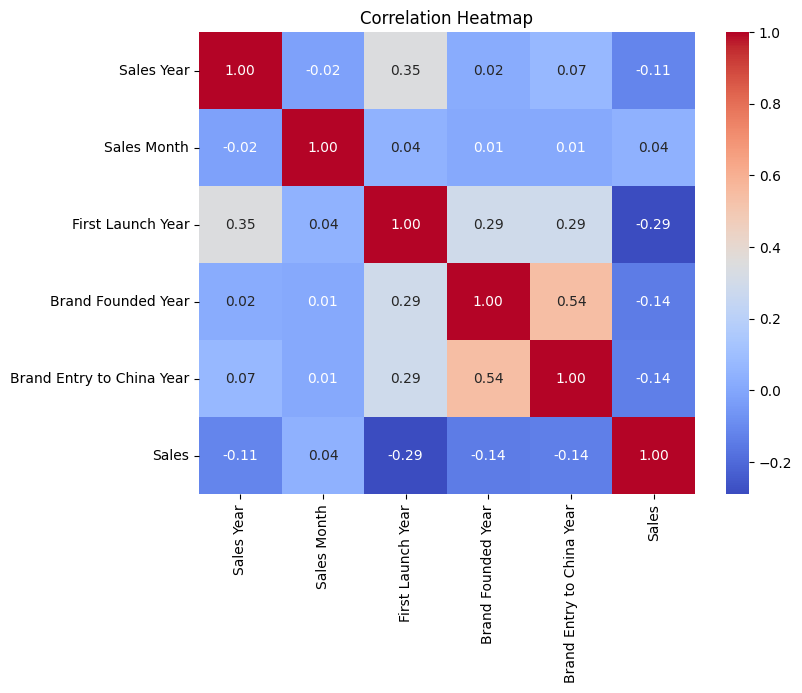
* **Correlation Analysis:** Strong correlation between User Overall Rating and Sales.
* **Clustering:** Cars clustered into Economy, Mid-range, and Luxury segments based on Price & Sales.
* **Sentiment Insights:** Cars with more positive sentiment reviews tend to sustain higher sales.

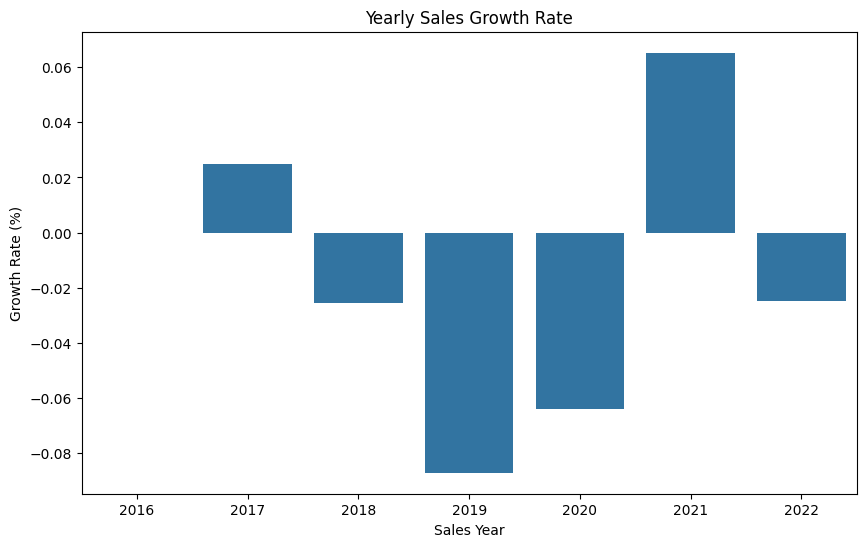
****

****

****

****

****

****