1. Introduction

This report presents a comprehensive data analysis of the *Car Sales Dataset*. The main objective is to clean, preprocess, and analyze the dataset to extract actionable business insights and support forecasting and decision-making for sales optimization.

2. Data Cleaning and Preprocessing

Several data quality issues were addressed before the analysis:

- **Handling Missing Values:**
    - Numerical columns were imputed using their median values.
    - Categorical columns were filled using their most frequent category (mode).
- **Removing Duplicates:**
  Duplicate rows were identified and removed to ensure data consistency.
- **Fixing Formatting Issues:**
  Text columns such as *Car Make* were standardized (trimmed spaces, unified case formatting, corrected missing values).
- **Adjusting Data Types:**
  Numerical columns were cast to appropriate types (integers or floats), and date columns were converted to datetime format for time-series analysis.
- **Outlier Treatment:**
  Outliers were detected and capped/removed using the Interquartile Range (IQR) method for key numeric features such as **Cost**, **Profit**, and **Sale Price** to prevent distortion in statistical summaries.
- **Feature Reduction:**
  Unnecessary columns (*Salesperson*, *Customer Name*, *Commission Earned*, *Commission Rate*) were removed as they did not contribute meaningful information to the sales or profitability analysis.

3. Exploratory Data Analysis (EDA)

3.1. Univariate Analysis

- **Customer Gender:**
  Male customers formed the majority of purchases, indicating a stronger buying intent in that demographic.
- **Car Make:**
  Toyota emerged as the top brand, reflecting strong trust and popularity among customers.
- **Car Model:**
  S-Class recorded the highest sales, highlighting a clear preference for premium/luxury vehicles.
- **Sales Region:**
  Idaho and *Pennsylvania* were identified as the strongest performing regions in terms of total revenue.
- **Seasonality:**
  Spring recorded the highest sales and profit, indicating strong seasonal demand.
- **Sale Month and Day:**
  January saw peak monthly sales, and *Monday* recorded the highest daily sales activity, possibly due to early-week promotional effects.

3.2. Bivariate Analysis

- **Profit by Gender:**
  Male customers contributed slightly higher total profit compared to female customers.
- **Customer Age vs Sale Price:**
  The age group **40–50 years** generated the highest concentration of high-value purchases.
- **Top Car Makes by Profit:**
  Toyota, *BMW*, *Audi*, and *Mercedes* dominated profitability, combining volume and luxury strength.
- **Car Year vs Sale Price:**
  Car sale prices dropped from 2018–2020, followed by a slight recovery in 2023–2024, suggesting cyclical market behavior.

- **Profit by Payment Method:**
  Cash transactions generated the highest profits, indicating cost savings from lower processing fees.
- **Discount vs Profit:**
  Optimal discount levels were observed around **10–12%**, beyond which profits declined.

3.3. Correlation Analysis

A correlation heatmap of numerical variables showed:

- Strong positive correlation (**>0.7**) between *Profit* and *Sale Price* — higher sale prices drive profitability.
- Moderate correlation between *Profit* and *Cost* (~0.4).
- Weak correlation between *Profit* and *Quantity* (~0.3).

4. Advanced Exploratory Insights

- **Seasonal Revenue & Profit:**
  Both peaked in *Spring*, suggesting that promotional and marketing efforts are best focused during this season.
- **Regional Performance:**
  Pennsylvania led in total sales revenue, making it a key market for expansion and inventory allocation.
- **Luxury vs Non-Luxury:**
  Luxury vehicles consistently outperformed non-luxury cars in both profit and profit margin percentage.
- **Weekend vs Weekday Revenue:**
  Weekday sales contributed more to total revenue, whereas weekends showed lower customer engagement.
- **Yearly Trends:**
  Overall revenue and profit declined across newer car years, with slight recovery during 2020–2021, hinting at changing consumer preferences or external economic factors.

5. Key Business Insights & Recommendations

1. **Customer Targeting:**
   Focus marketing campaigns on the 40–50 age segment, which represents the most active buyer group.
2. **Brand Strategy:**
   Expand *Toyota*'s market reach while maintaining premium campaigns for *BMW*, *Audi*, and *Mercedes*.
3. **Pricing Strategy:**
   Maintain discounts in the optimal range (10–12%) to maximize profit margins.
4. **Seasonal Planning:**
   Strengthen promotions during spring and replicate February's strong revenue drivers across other months.
5. **Regional Focus:**
   Allocate higher dealership support and inventory to Pennsylvania and Idaho, as they are top-performing markets.
6. **Luxury Segment:**
   Increase inventory and marketing focus on luxury models, which yield higher ROI and profit margins.

Feature Engineering Summary

1. Overview

Feature engineering was performed to enrich the dataset and enhance the predictive capabilities of future forecasting models. Several new attributes were created to capture customer behavior, seasonality, profitability, and product segmentation patterns.

2. Newly Created Features

Feature Name	Description	Type	Expected Impact on Model
Age Group	Segmented customers into 4 groups: 18–30, 31–45, 46–60, 60+.	Categorical	Captures age-related purchasing power and preferences.
Revenue	Calculated as Quantity × Sale Price.	Numeric	Key performance metric; improves forecasting of total sales volume.
Profit Margin %	(Profit / Sale Price) × 10.	Numeric	Represents profitability efficiency; helps identify high-margin sales.
Discount Level	Classified discount values into ‘Low’, ‘Medium’, ‘High’, and ‘Very High’.	Categorical	Helps understand the relationship between discount tiers and profit changes.
Is Luxury	Flag indicating whether the car’s sale price exceeds 50,000.	Binary	Differentiates high-value transactions for profitability modeling.
Is Weekend	Binary variable (1 for Saturday/Sunday).	Binary	Captures weekly sales patterns and temporal seasonality.
YearMonth	Extracted month name from the date for time-series grouping.	Categorical	Facilitates trend and seasonal pattern analysis across months.
Car Model Group	Classified models into “Luxury”, “Midrange”, and “Economy” categories.	Categorical	Improves model interpretability and prediction of profitability patterns.

3. Expected Benefits for Forecast Modeling

- **Enhanced Predictive Power:**
  The engineered features introduce behavioral and temporal dimensions that strengthen sales forecasting and profitability prediction.
- **Improved Segmentation:**
  Grouped features (e.g., Age Group, Car Model Group) enable targeted predictions for different market segments.
- **Seasonality Recognition:**
  Time-based features (e.g., YearMonth, Is Weekend) allow the model to recognize cyclical trends and seasonal peaks.
- **Profitability Insights:**
  Derived features such as *Profit Margin %* and *Discount Level* enable better understanding of pricing strategies and their effects.

4. Conclusion

Through systematic data cleaning, exploration, and feature engineering, the dataset was transformed into a structured, insight-rich foundation ready for modeling and forecasting. The new features are designed to enhance model performance, provide deeper business insights, and support data-driven decision-making in the car sales domain.