PREDICTIVE ANALYTICS AND SEGMENTATION IN THE AUTOMOTIVE INDUSTRY: PRICE, SALES, SUCCESS, AND CLUSTERING

PRESENTING BY

RAJKUMAR K P

SELF INTRODUCTION

Hi, I'm **Rajkumar**, I completed MSc Biotechnology in 2022, I'm a passionate and detail-oriented **Data Analyst** with almost **2 years of experience**, specializing in **healthcare data analysis, dashboard development**, and **predictive modeling**.

I've worked extensively with **Electronic Health Records (EHRs), claims data**, and **clinical datasets**, using tools like **Python, SQL, Power BI**, and **Oracle Database** to uncover insights that support improved patient outcomes and business decisions. At **Guidehouse India Pvt. Ltd**, I developed multiple analytics solutions, and ensured compliance with **HIPAA, GDPR and CCPA** standards.

I'm also skilled in **machine learning**, and alongside strong data visualization capabilities using **Power BI, Tableau, and Excel**.

I'm always eager to explore new domains. I bring a **self-driven, collaborative mindset** and am excited about opportunities where I can apply my analytical thinking to solve real-world problems.

EXECUTIVE SUMMARY

This assessment comprises five advanced data science problems focusing on regression, classification, clustering, time series forecasting, and mixed-effects modeling using a car dataset. Each question was approached with structured methodology, industry-standard tools, and best modeling practices.

Tools & Techniques Used:

- **Languages:** Python (pandas, scikit-learn, statsmodels, matplotlib, seaborn)
- **Models:** Random Forest, Logistic Regression, ARIMA, Exponential Smoothing, KMeans, Hierarchical Clustering, Mixed Linear Model
- **Techniques:** EDA, preprocessing, scaling, encoding, model tuning, time series analysis, unsupervised learning
- **Metrics Used:** R^2, RMSE, ROC-AUC, PR-AUC, confusion matrix, MAPE

I. Car Price Prediction Model using Random Forest

Objective:

Predict car prices accurately using available car features to support pricing and business decisions.

Methodology:

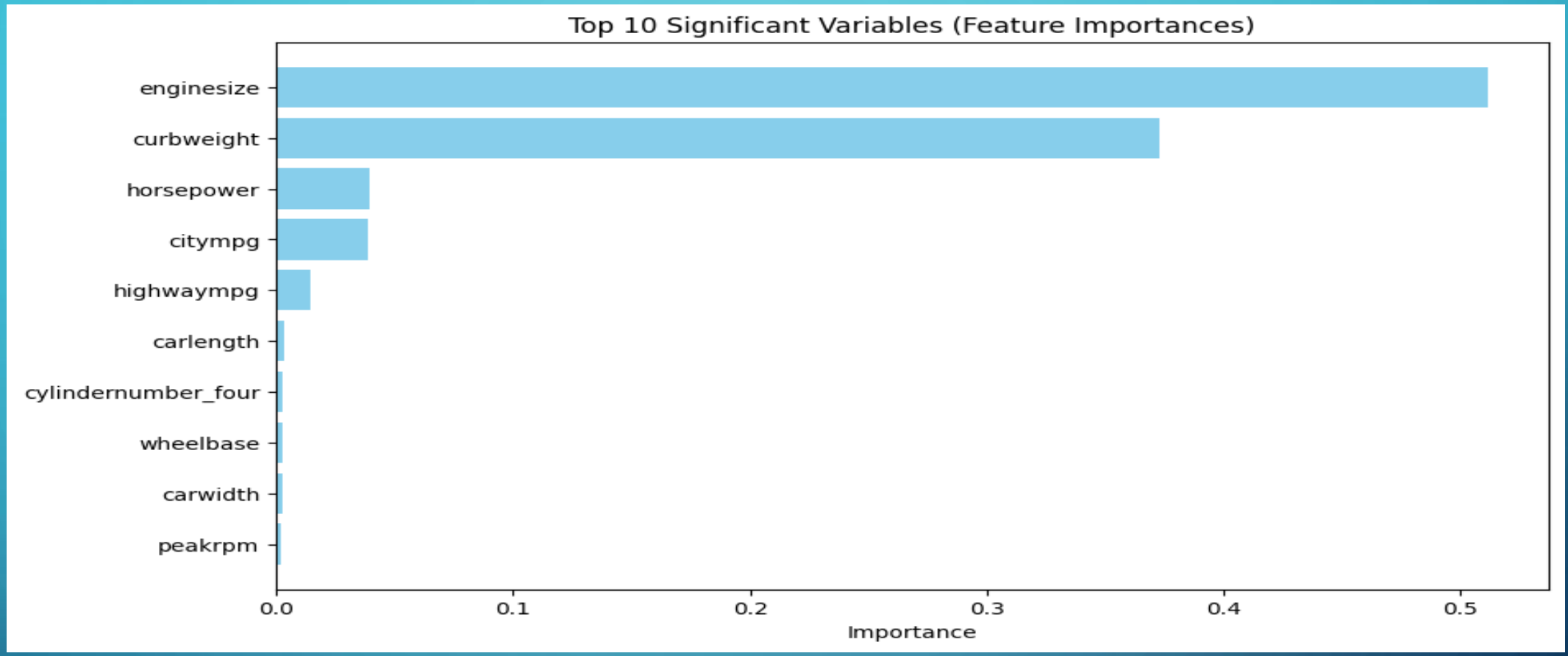
- Loaded and explored the dataset for missing values.
- Pre-processed data: dropped rows with missing target, imputed missing values, removed irrelevant columns, and encoded categorical variables.
- Built an initial Random Forest Regressor model.
- Tuned model hyperparameters (n_estimators, max_depth, min_samples_leaf) to improve generalization and reduce overfitting.

Metrics:

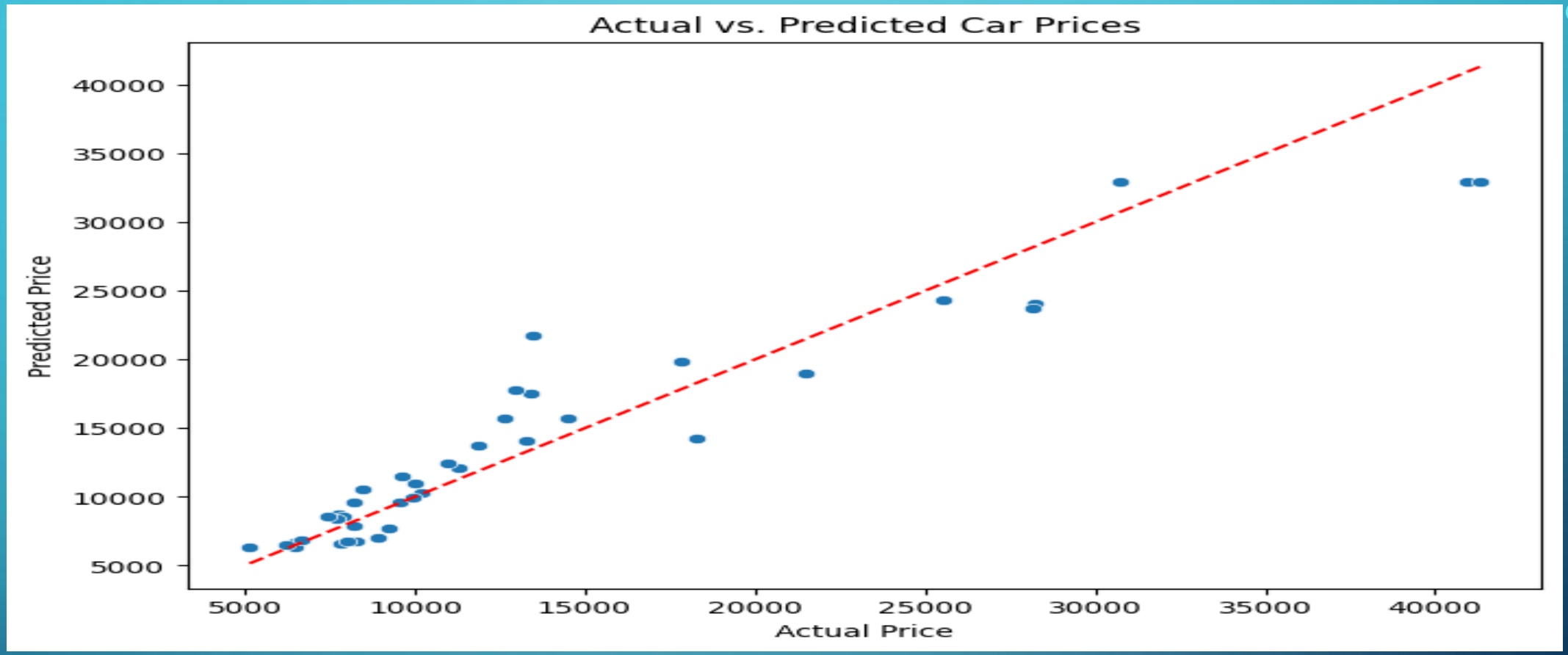
- R^2 (Coefficient of Determination) to measure model accuracy.
- Feature importance to identify key drivers of car price.

Results:

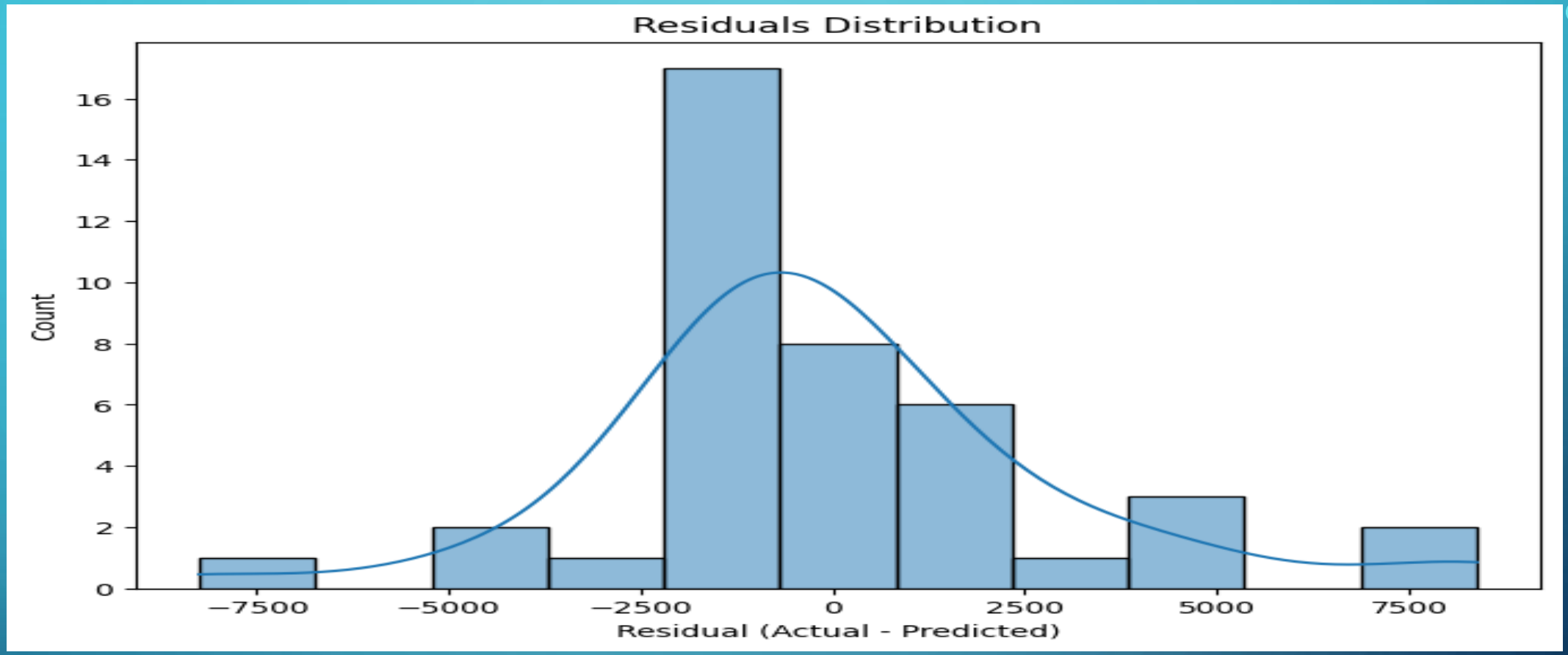
- Initial R^2: 0.95 (indicating possible overfitting).
- Tuned R^2: 0.89 (more realistic and generalizes better).
- Top features: engine size, curb weight, horsepower, citympg.
- **Conclusion:**
  The tuned Random Forest model predicts car prices well, mainly using physical and performance features. Tuning improved model reliability and reduced overfitting, making the results more actionable for business use.



- Engine size and curb weight are the most important features for predicting car price, followed by horsepower and citympg.
- Other features like highwaympg, car length, and cylinder number have much lower importance.
- This means physical and performance attributes of the car are the main drivers of price.



- Most points are close to the red diagonal line, showing good agreement between actual and predicted prices
- Some outliers exist, but overall, the model predicts prices accurately for most cars.
- Indicates the Random Forest model generalizes well after tuning.



- Residuals are mostly centered around zero, with a roughly normal distribution.
- This suggests the model does not have major bias and errors are spread evenly.
- A few large residuals indicate some cars are harder to predict, but most predictions are close.

II. Binary Classification Models for Car Model Success

Objective:

Classify car models as successful or not based on their features, to identify key drivers of success and support strategic decisions.

Methodology:

- Defined success as cars with price above the median.
- Preprocessed data: removed irrelevant columns, encoded categorical variables.
- Built and compared Logistic Regression and Random Forest classification models.
- Evaluated models using confusion matrix, ROC-AUC, and PR-AUC.
- Analyzed feature importances to identify main factors influencing success.

Metrics:

- Confusion Matrix (True/False Positives/Negatives)
- ROC-AUC (Receiver Operating Characteristic Area Under Curve)
- PR-AUC (Precision-Recall Area Under Curve)
- Feature Importance

Results:

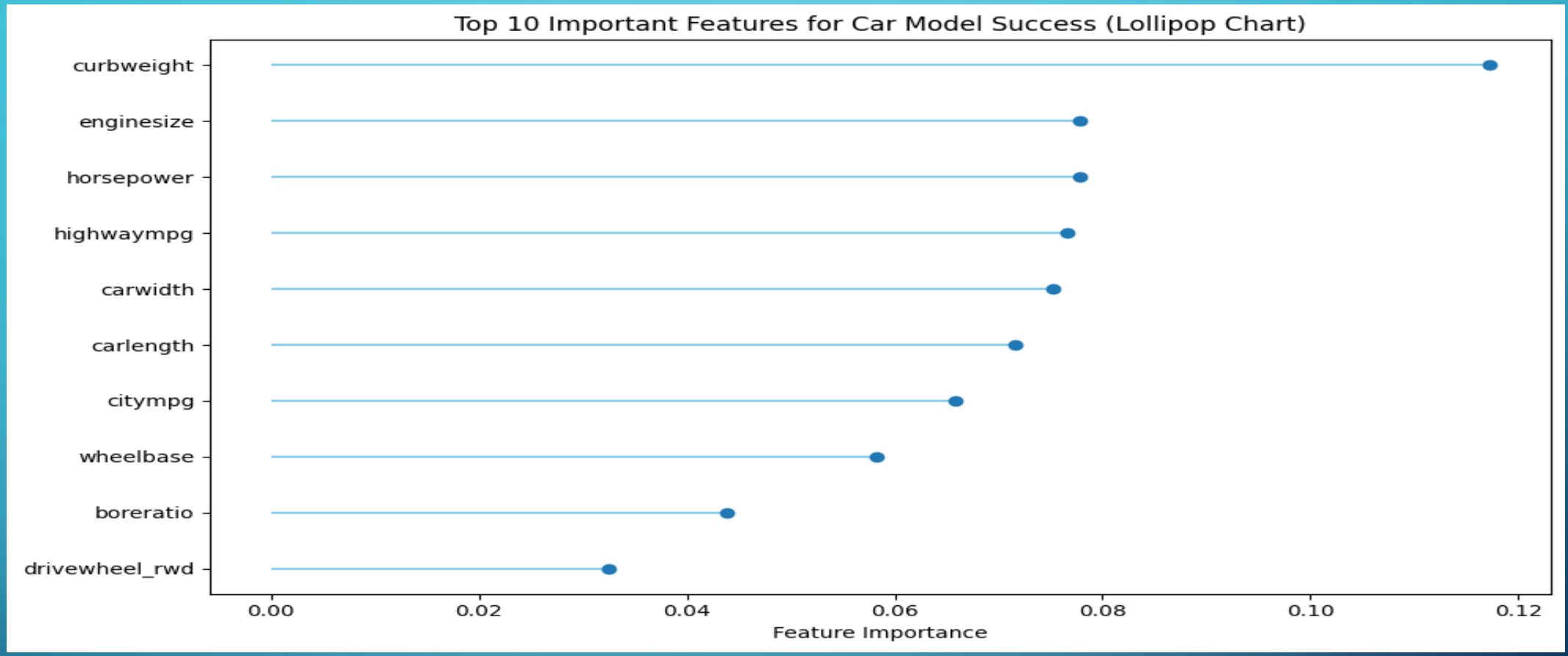
- Both models achieved very high ROC-AUC and PR-AUC scores (close to 1.00).
- Random Forest slightly outperformed Logistic Regression and had zero false negatives.
- Most important features: curb weight, engine size, horsepower, highwaympg.

Conclusion:

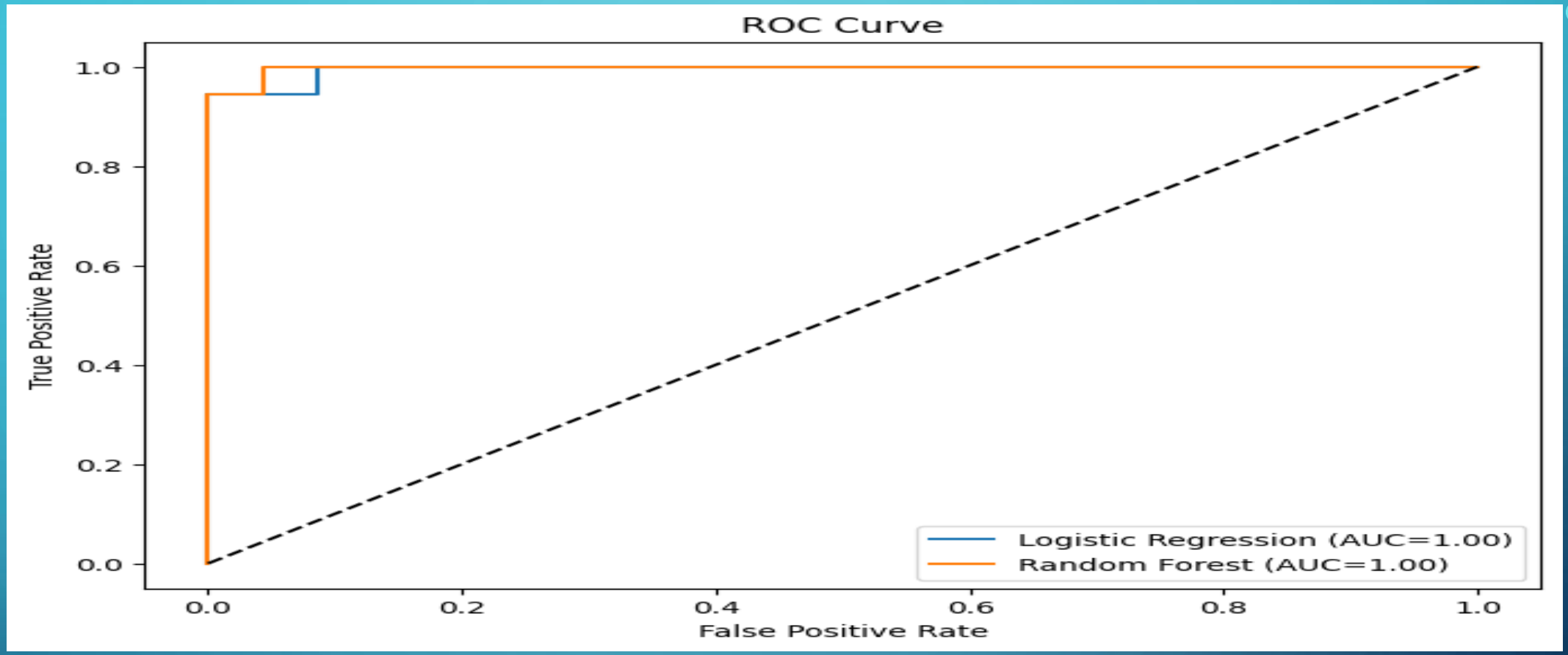
Both models classify car model success with high accuracy. Physical and performance features are the main drivers of success. Random Forest is preferred for its slightly better performance and no false negatives, making it reliable for identifying successful car models.

COMPARISON OF DIFFERENT BINARY CLASSIFICATION MODELS

Model	Theoretical Advantages	Practical Usage	Likelihood Metric	Confusion Matrix	ROC-AUC	PR-AUC	Interpretability
Logistic Regression	Simple, interpretable, outputs probabilities	Good baseline, fast, feature weights	Log-likelihood	[[22, 1], [1, 17]]	0.995	0.994	High
Random Forest	Handles non-linearity, robust, ranks features	Handles complex data, less overfitting	Ensemble likelihood	[[21, 2], [0, 18]]	0.998	0.997	Medium



- Curb weight and engine size are the most important features for predicting car model success
- Other significant features include horsepower, highwaympg, car width, and car length.



- Both Logistic Regression and Random Forest models have ROC-AUC scores very close to 1.00, indicating excellent classification performance.
- The curves are almost at the top left corner, meaning both models can distinguish successful car models from unsuccessful ones with high accuracy.
- Random Forest slightly outperforms Logistic Regression, as seen in the curve and AUC values.

III. Binary Classification Models for Car Model Success

Objective:

Segment cars into distinct groups based on key features to identify market segments and guide business strategies.

Methodology:

- Selected relevant features (engine size, price, fuel type, success).
- Preprocessed data: encoded categorical variables and scaled features.
- Applied KMeans clustering; determined optimal number of clusters using the Elbow Method and dendrogram.
- Visualized clusters and analyzed cluster centers.

Metrics:

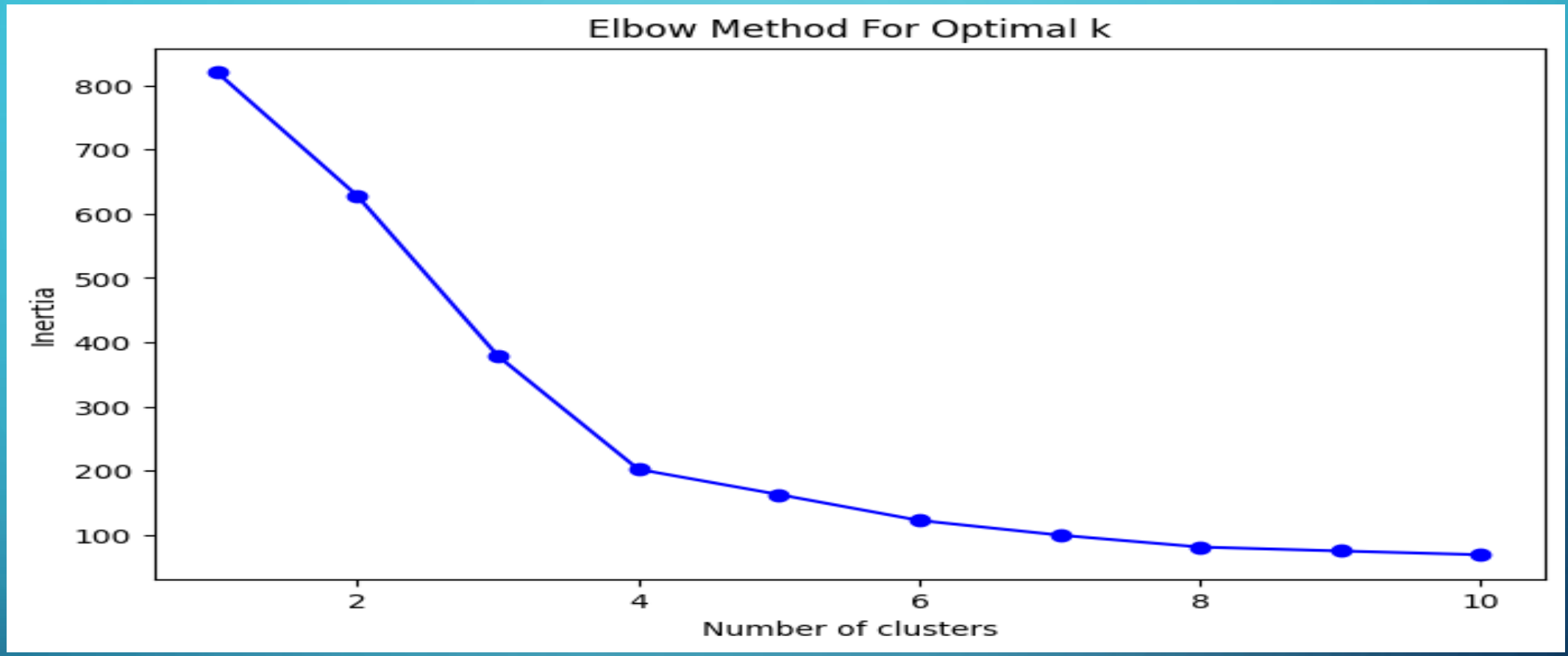
- Inertia (Elbow Method) to select optimal cluster count.
- Cluster centers to interpret segment characteristics.

Results:

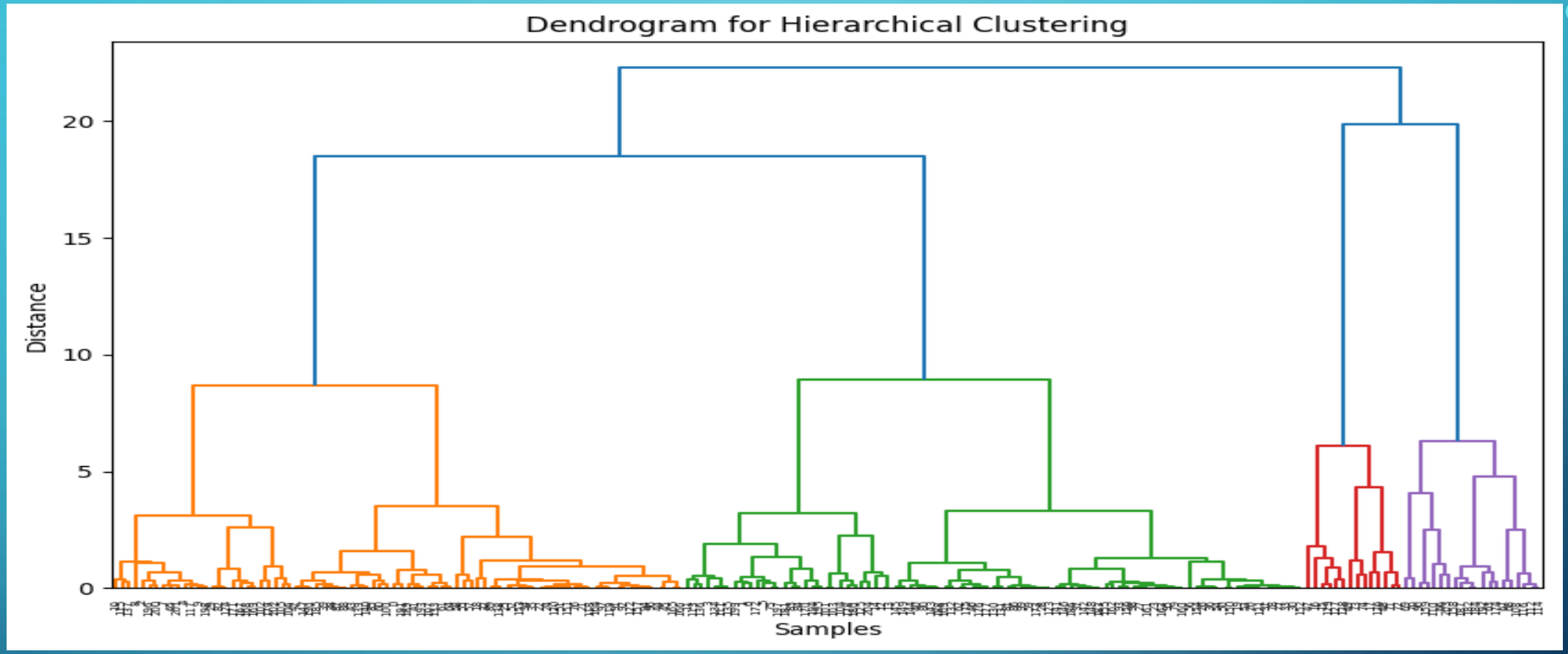
- Identified 4 distinct car segments:
  - Cluster 0: Small engine, low price, all gas, all successful.
  - Cluster 1: Small engine, low price, all diesel, all successful.
  - Cluster 2: Large engine, high price, mixed fuel, all successful.
  - Cluster 3: Medium engine, mid price, mostly diesel, mostly not successful.
- Cluster 2 has the highest average price (premium segment); Clusters 0 and 1 are budget segments.

Conclusion:

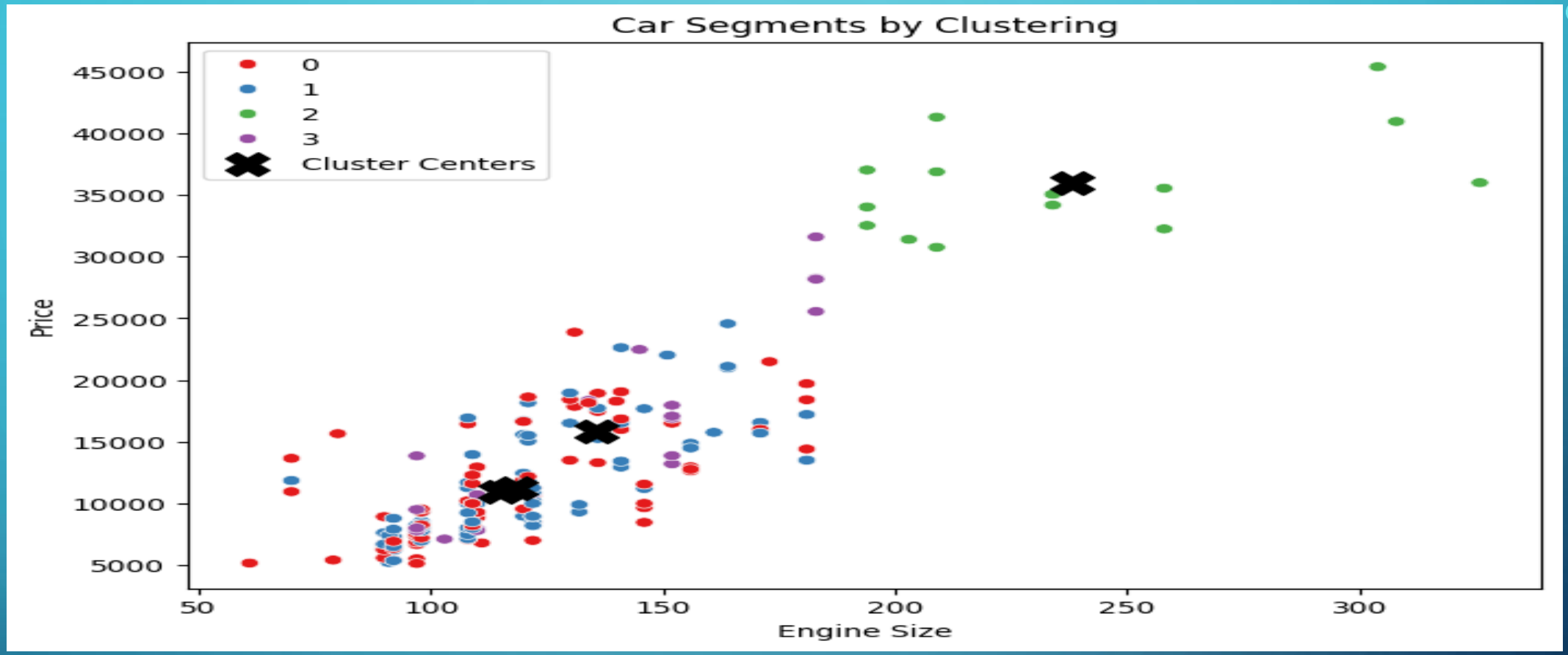
Clustering revealed clear car segments based on engine size, price, fuel type, and success. These insights can be used for targeted marketing, product positioning, and strategic decision-making.



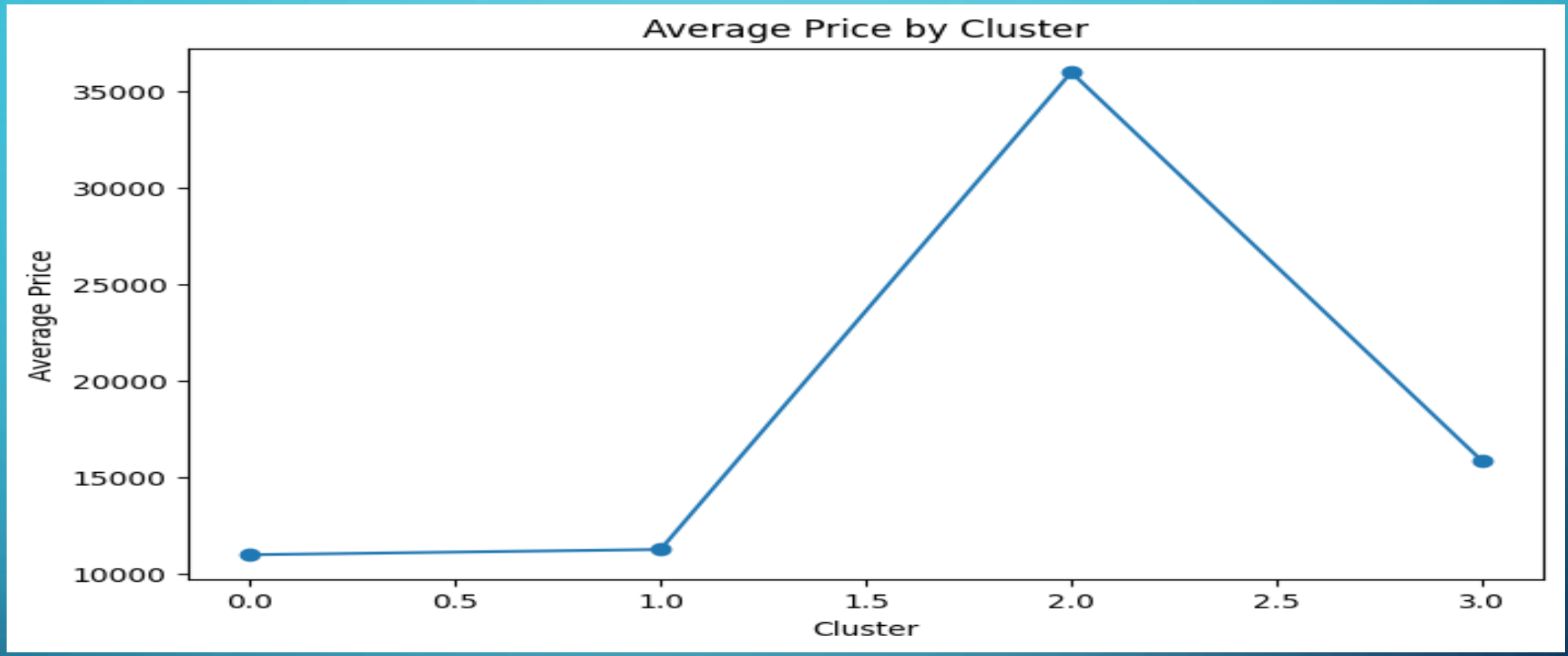
- The Elbow Method visual shows how inertia (within-cluster sum of squares) decreases as the number of clusters (k) increases
- In this plot, the elbow appears around k=4, indicating that segmenting the cars into 4 clusters balances model simplicity and segmentation quality.



- Large vertical jumps (long branches) indicate major splits between clusters, suggesting natural groupings in the data.
- The first few big splits (at higher distances) suggest an optimal cluster count of around 4, matching the elbow method result.
- Cars within the same lower branches are more similar to each other in terms of features (engine size, price, fuel type, success).
- This hierarchical structure helps confirm and visualize the distinct car segments found by KMeans, supporting the choice of 4 clusters for segmentation.



- Cluster 0: Small engine, low price, all gas, all successful.
- Cluster 1: Small engine, low price, all diesel, all successful.
- Cluster 2: Large engine, high price, mixed fuel, all successful.
- Cluster 3: Medium engine, mid price, mostly diesel, mostly not successful.



- Cluster 2 has the highest average price, indicating this group contains premium or luxury cars.
- Clusters 0 and 1 have the lowest average prices, representing budget or economy car segments.
- Cluster 3 is mid-range, possibly representing standard or family cars

IV. Car Sales Weekly Forecast (Time Series Model)

Objective:

Forecast weekly car sales for the next quarter to support inventory planning and business decisions.

Methodology:

- Loaded historical weekly sales data.
- Visualized sales trends and checked autocorrelation.
- Split data into training and test sets (last 12 weeks as test).
- Built and compared ARIMA, Simple Exponential Smoothing, and Holt's Linear Trend models.
- Selected the best model based on error metrics and forecasted future sales.

Metrics:

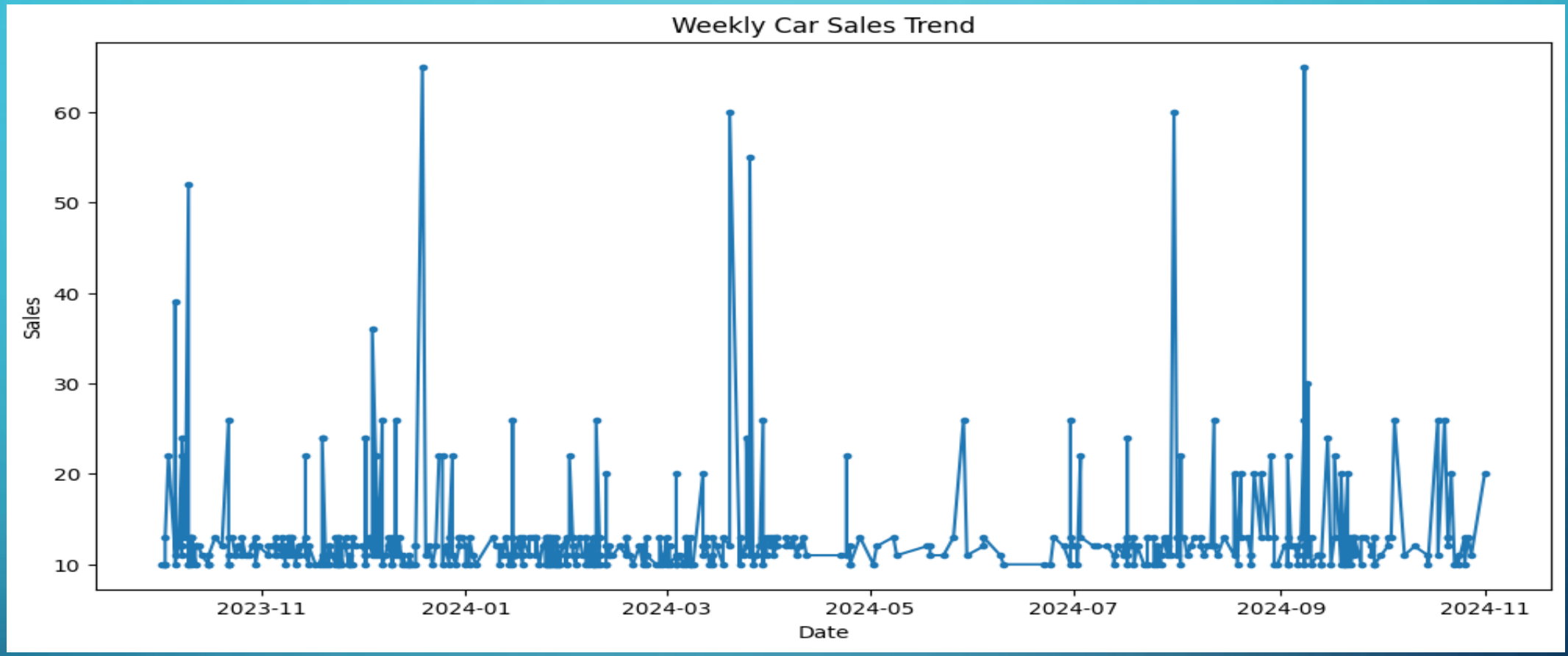
- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)

Results:

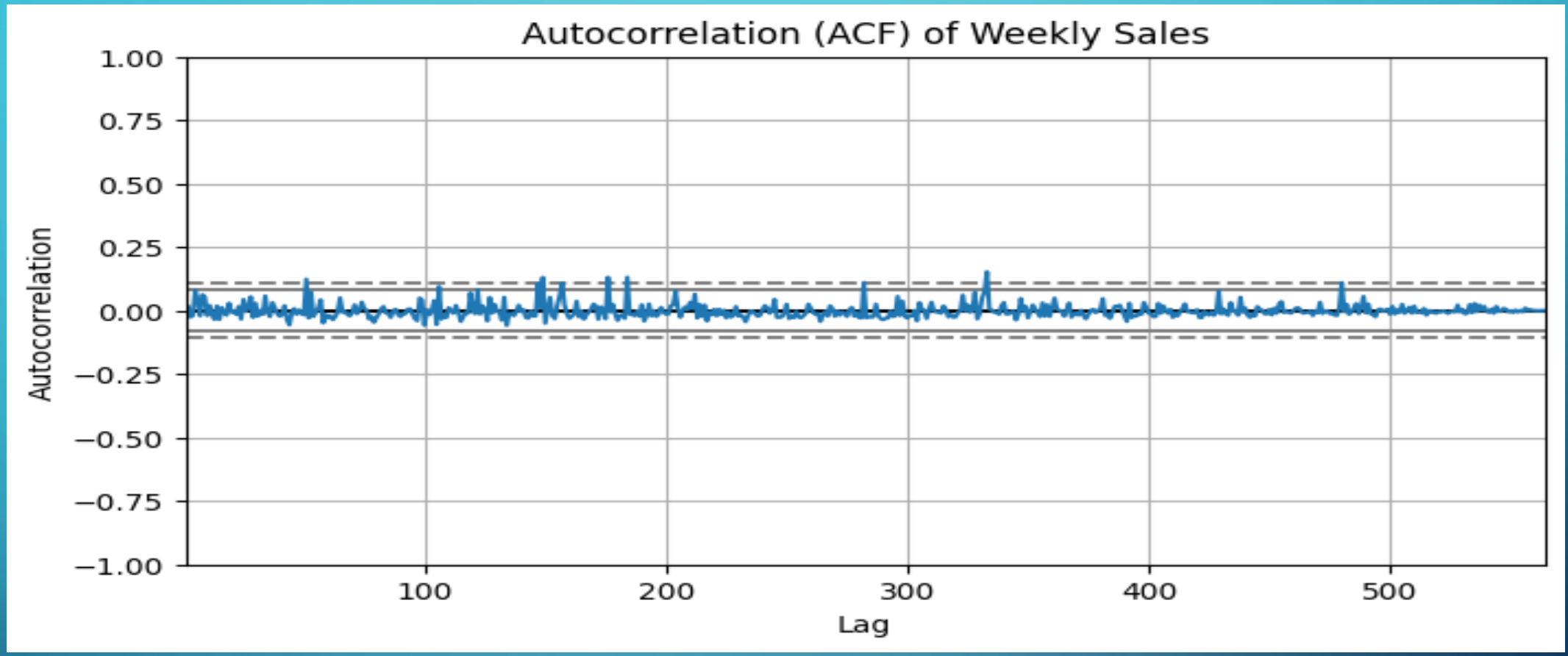
- All models produced smoother forecasts than the actual sales, which were highly volatile.
- ExpSmoothing and ARIMA predicted nearly flat lines, while Holt's Linear showed a steady upward trend.

Conclusion:

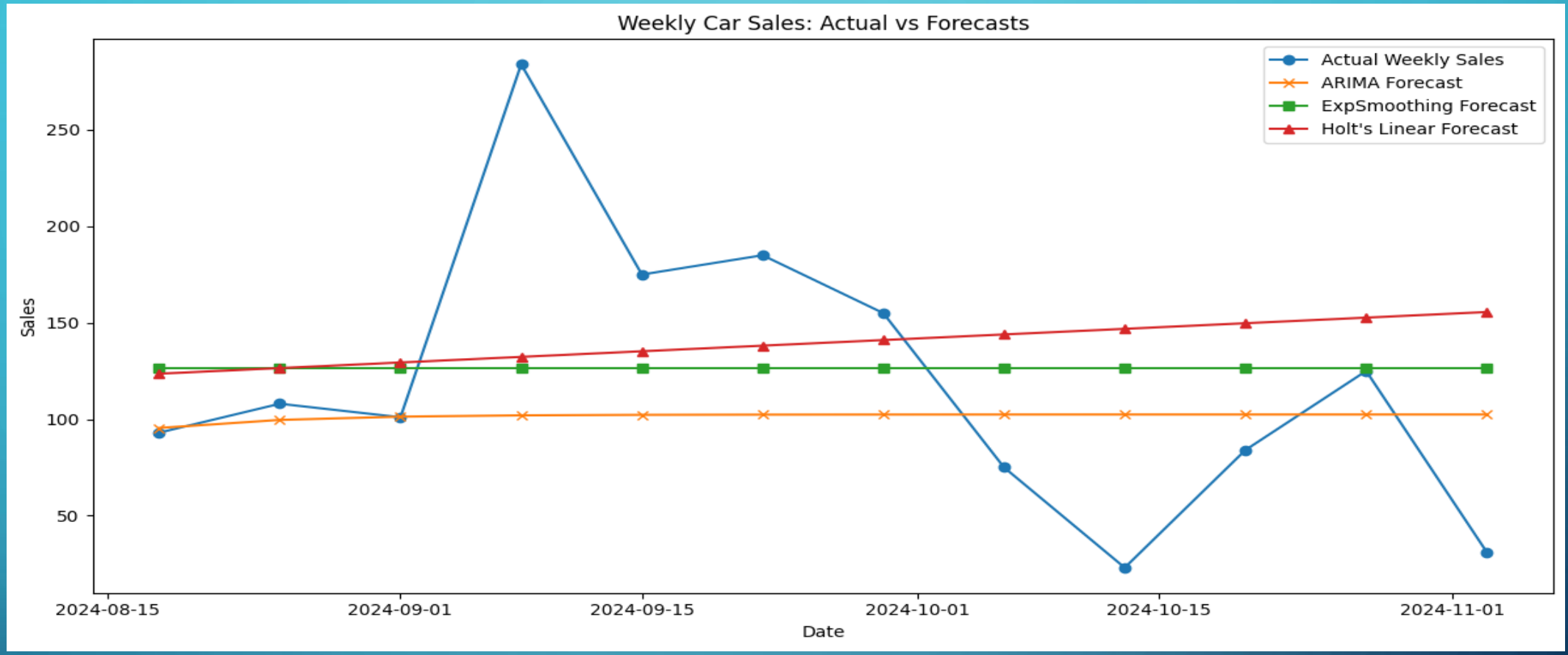
ExpSmoothing achieved the lowest RMSE, indicating the smallest typical error.
ARIMA achieved the lowest MAE, indicating the smallest average error. Sales are expected to remain steady, enabling confident planning for inventory and operations.



- Most weekly sales values cluster between 10 and 20, indicating stable baseline demand.
- There are occasional sharp spikes, with sales jumping above 50 or even 60 units in some weeks—these may be due to promotions, new launches, or seasonal effects.
- No clear upward or downward trend is visible; sales fluctuate around a consistent average.
- The pattern suggests that while most weeks are predictable, the business should be prepared for occasional surges in demand.



- Most autocorrelation values are close to zero across all lags, indicating that weekly car sales do not show strong seasonality or repeating patterns.
- There are no significant spikes above the confidence bounds, so past sales do not strongly predict future sales at any lag
- This suggests sales are mostly random week-to-week, with little persistence or trend.
- Time series models like ARIMA are appropriate, as the data does not have strong autocorrelation or seasonal effects.



- Actual Sales are Highly Volatile: The blue line (actual sales) shows large fluctuations, with some weeks having very high or very low sales.
- All Models Predict Smoother Trends: All three forecasting models (ARIMA, ExpSmoothing, Holt's Linear) produce much smoother lines compared to the actual sales.
- ExpSmoothing and ARIMA are Flat: ExpSmoothing (green) predicts a nearly constant value for all weeks. ARIMA (orange) also predicts a nearly flat line, slightly below ExpSmoothing. This suggests both models expect sales to remain stable, not accounting for sudden changes.

V. Relationship Between Price, Ad Spend, and Sales (Mixed Effects Model)

Objective:

Determine how car price and advertising spend impact sales, while accounting for differences between car models.

Methodology:

- Loaded car sales, price, ad spend, and car model data.
- Visualized relationships with scatterplots.
- Built a mixed effects regression model: Sales ~ Price + Ad Spend +(Random Effect: Car_Id)
- Estimated fixed effects for price and ad spend, and random effects for car models.

Metrics:

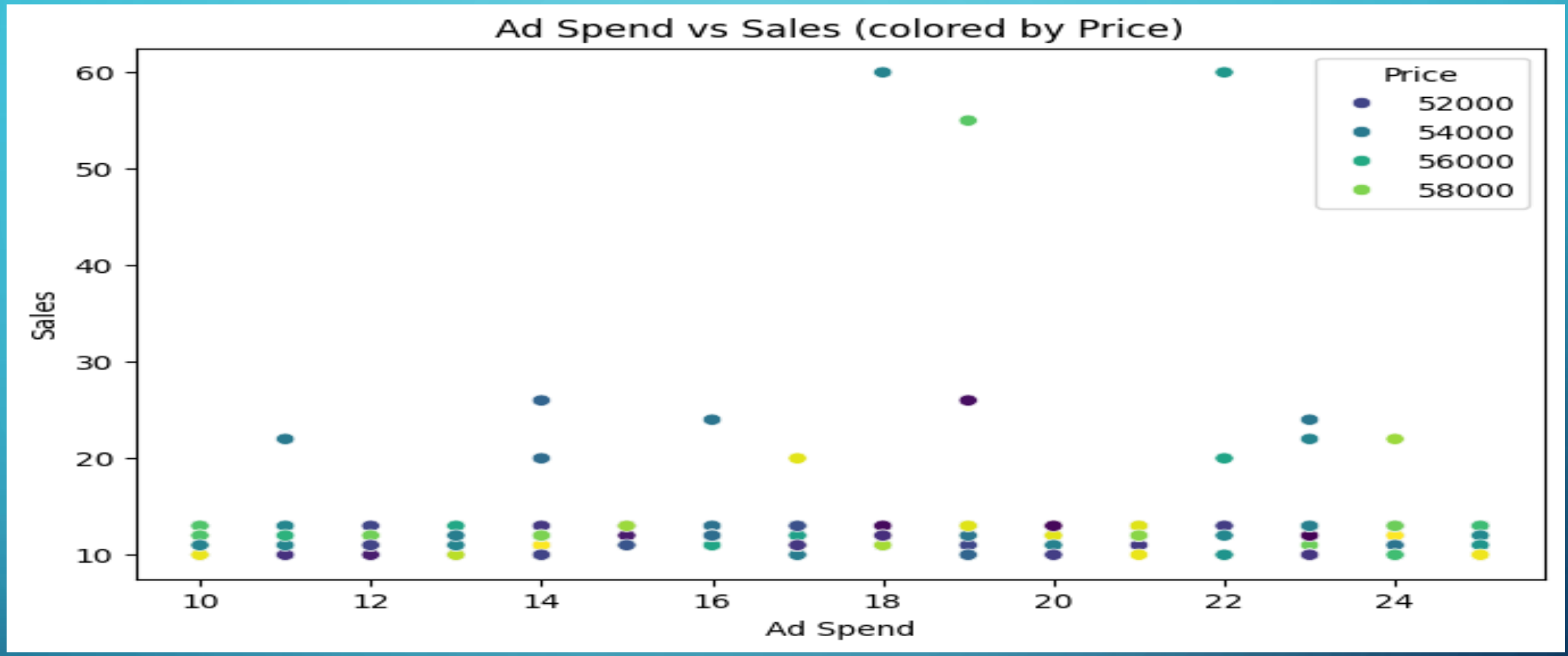
- Coefficient estimates for Price and Ad Spend
- Statistical significance (p-values)
- Model intercept and group variance

Results:

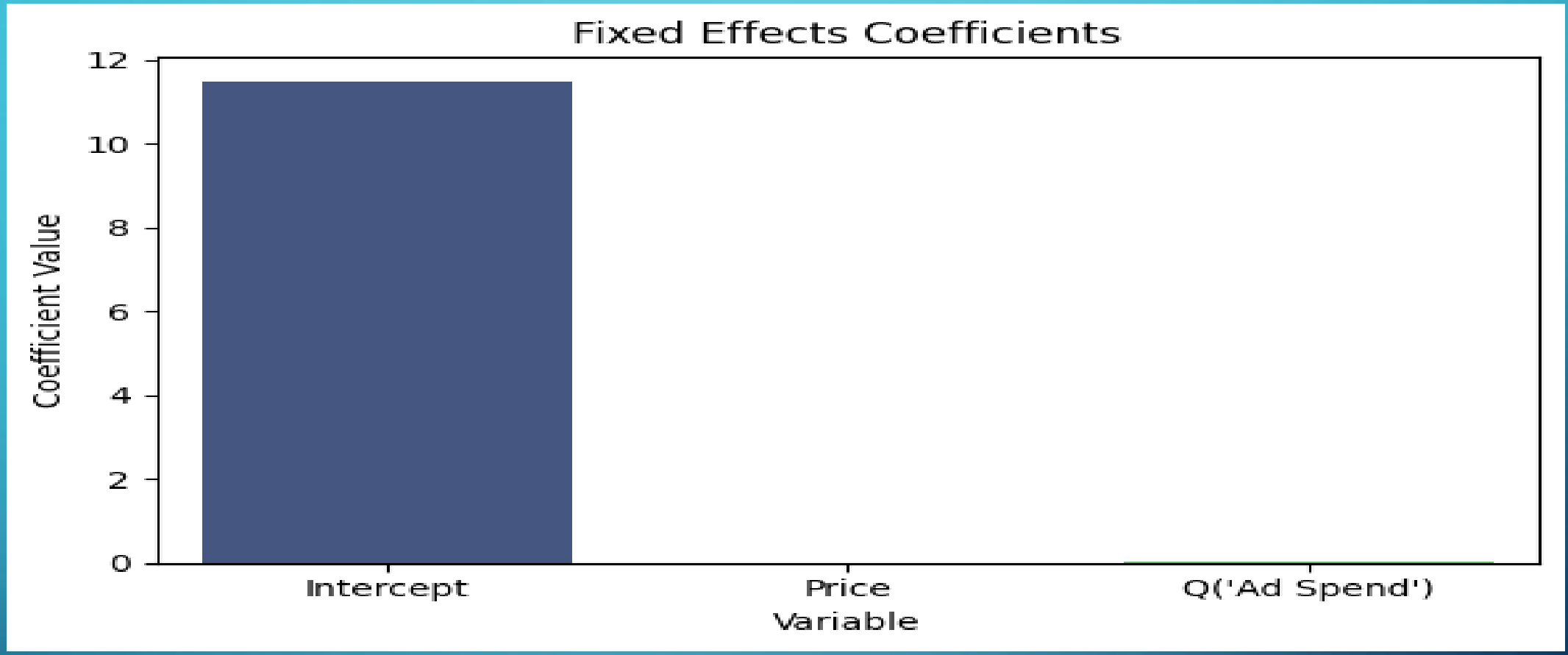
- Price and Ad Spend coefficients are very close to zero and not statistically significant ($p > 0.05$).
- Most variation in sales is explained by differences between car models (high group variance).
- Visualizations confirm no clear relationship between price/ad spend and sales.

Conclusion:

Price and advertising spend do not significantly affect car sales in this dataset. Sales differences are mainly driven by car model characteristics. Business strategies should focus on model-specific factors rather than adjusting price or ad spend.



- Most sales values cluster between 10 and 15, regardless of ad spend or price.
- A few points show much higher sales (above 20, up to 60), but these are rare and spread across different ad spend levels.
- There is no clear upward trend—increasing ad spend does not consistently lead to higher sales.
- Price groups (colors) are spread throughout, showing no strong relationship between price and sales in this visual.



- Intercept is much higher than the coefficients for Price and Ad Spend.
- Both Price and Ad Spend coefficients are very close to zero, indicating they have minimal impact on sales in this model.
- The visual confirms the regression results: neither price nor ad spend significantly influences sales.
- Most of the variation in sales is explained by the baseline (intercept) and differences between car models (random effect), not by changes in price or ad spend.

OVERALL OUTLINE OF STEPS

- **Data Loading & Exploration**
  - Loaded datasets from Excel using pandas.
  - Explored data for missing values, data types, and initial statistics.
- **Data Preprocessing**
  - Dropped rows with missing target values.
  - Imputed missing values in features using the median.
  - Removed irrelevant columns (e.g., car_ID).
  - Encoded categorical variables using one-hot encoding.
- **Feature Selection**
  - Selected relevant features for each model (regression, classification, clustering, time series, mixed effects).
- **Model Building**
  - **Model 1:** Built and tuned a Random Forest Regressor for car price prediction.

- **Model 2:** Built Logistic Regression and Random Forest Classifier for binary classification of car model success.
- **Model 3:** Applied KMeans clustering for car segmentation and visualized clusters.
- **Model 4:** Built ARIMA, Exponential Smoothing, and Holt's Linear models for weekly car sales forecasting.
- **Model 5:** Built a Mixed Effects Model to analyze the relationship between price, ad spend, and sales.

- **Model Evaluation**
  - Used appropriate metrics: R^2 for regression, ROC-AUC/PR-AUC/confusion matrix for classification, inertia for clustering, MAE/RMSE for time series, and coefficient interpretation for mixed effects.
  - Compared models and selected the best-performing ones.
- **Visualization**
  - Visualized feature importances, actual vs. predicted values, cluster assignments, time series forecasts, and model coefficients.

OVERALL CHALLENGES FACED

- **Overfitting:**
  Initial models (especially Random Forest) showed very high R^2, indicating possible overfitting. Required hyperparameter tuning to improve generalization.
- **Missing Data:**
  Needed careful imputation to avoid bias and ensure model robustness.
- **Categorical Encoding:**
  High-cardinality categorical features increased the number of dummy variables, which could lead to overfitting and reduced interpretability.
- **Feature Selection:**
  Too many features after encoding made models complex and harder to interpret.
- **Hyperparameter Tuning:**
  Required several iterations to find the right balance between model complexity and performance.

- **Time Series Volatility:**
  Actual sales data was highly volatile, making it difficult for simple models to capture extreme peaks and troughs.
- **Model Interpretability:**
  Some models (e.g., Random Forest, clustering) were less interpretable, requiring additional visualization and analysis.
- **Integration of Multiple Models:**
  Managing different modeling approaches and ensuring consistent preprocessing across tasks was challenging.

OVERALL CONCLUSION

The analysis shows that car sales and success are primarily influenced by product characteristics (engine size, curb weight, etc.), not by price or advertising spend. Predictive models and clustering provide actionable insights for pricing, segmentation, and inventory planning. Business strategies should prioritize car model features and segment-specific approaches over broad pricing or marketing adjustments.

THANK YOU !