**Car Data Analysis Assessment – Analytics Specialist**

**Objective**

This analysis aims to develop predictive models, identify key factors driving car success, segment cars into meaningful clusters, forecast future sales, and examine the influence of price and advertising on sales using the provided multi-sheet car dataset. The goal is to extract actionable insights and build robust models using Python-based data science techniques.

**Data Overview**

The dataset contains three sheets:

* **Car\_Assignment1:** Detailed car specifications and attributes including price, engine details, fuel type, and a binary target indicating model success.
* **Car\_Assignment2:** Time series sales data across several car brands with corresponding prices and dates.
* **Car\_Assignment3:** Monthly car sales data with price points and advertising spending information.

Exploratory data analysis was performed to assess data completeness, distributions, and relationships among variables before modelling.

**Questions**

**Question 1.** What approach would you take to build a model predicting price based on the dataset? Outline the steps and explain the challenges you have faced.

**Approach and Steps:**

To develop a model for predicting car prices, the analysis started with exploratory data analysis to understand data quality and distributions. Essential steps included:

Data Cleaning: Handling missing values by dropping incomplete rows to maintain data integrity.

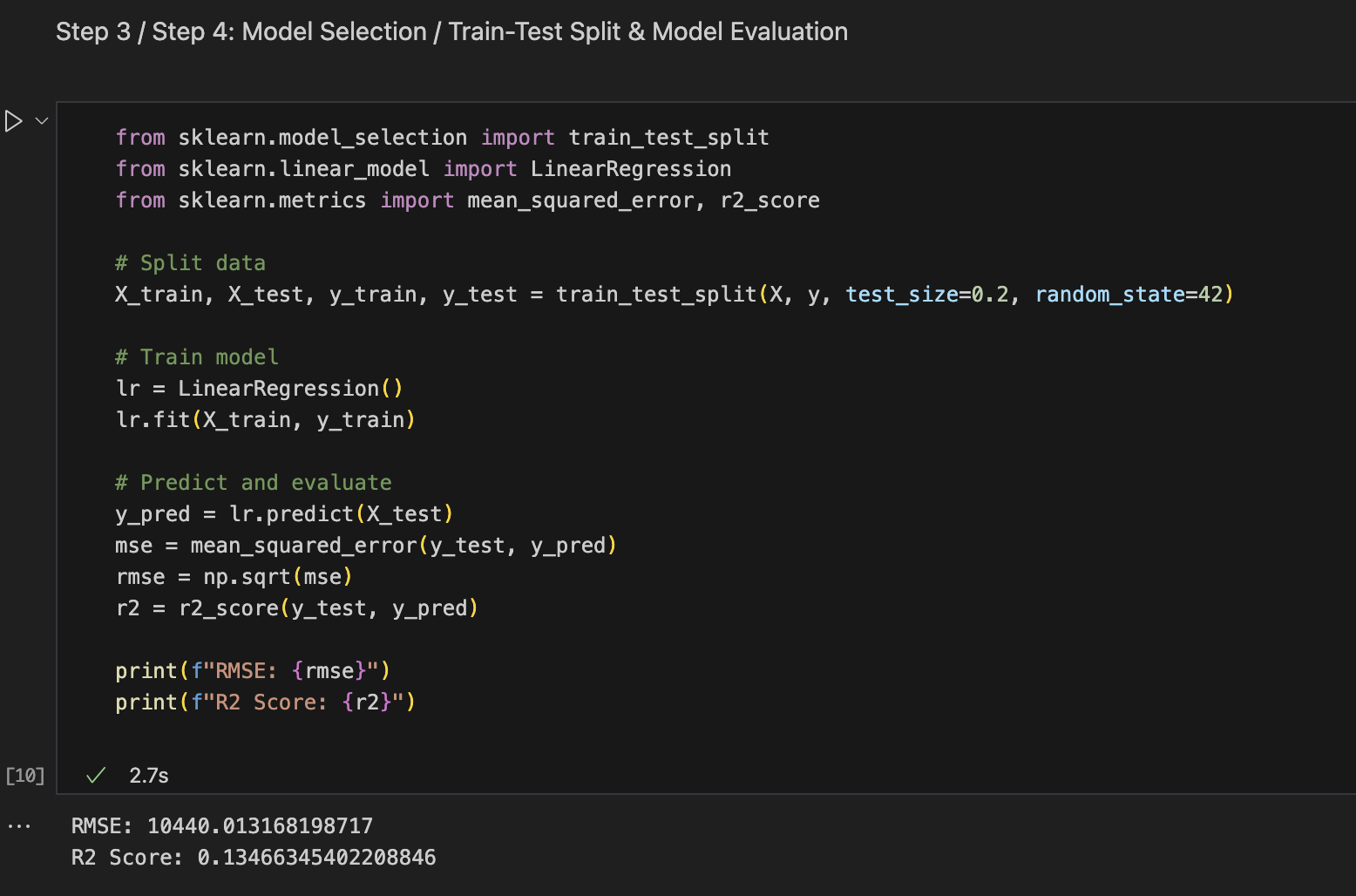
* Feature Engineering: Conversion of categorical variables (like fuel type, aspiration) into numerical form using one-hot encoding to enable algorithm compatibility.
* Feature Selection: Dropped identifiers and target column to prepare features for modelling.
* Modelling: A linear regression model was built as a strong baseline due to its interpretability and simplicity.
* Evaluation: Data was split into training and testing sets; the model’s accuracy was assessed using RMSE and R² score.

**Challenges Faced:**

* Missing data led to loss of some records.
* Categorical variables increased dimensionality after encoding, requiring careful feature management.
* Multicollinearity among engine-related features affected coefficient estimates and model stability.
* Limited sample size constrained model generalizability.
* Ensuring the model avoided overfitting with proper train/test splits.

**Outcome:**

The baseline regression produced an RMSE of about **10,440** and an R² of **0.13**, meaning it explains roughly 1**3% of the variation in prices**. In practical terms, this performance is weak, highlighting the need for stronger models and better predictors to improve forecasting accuracy.



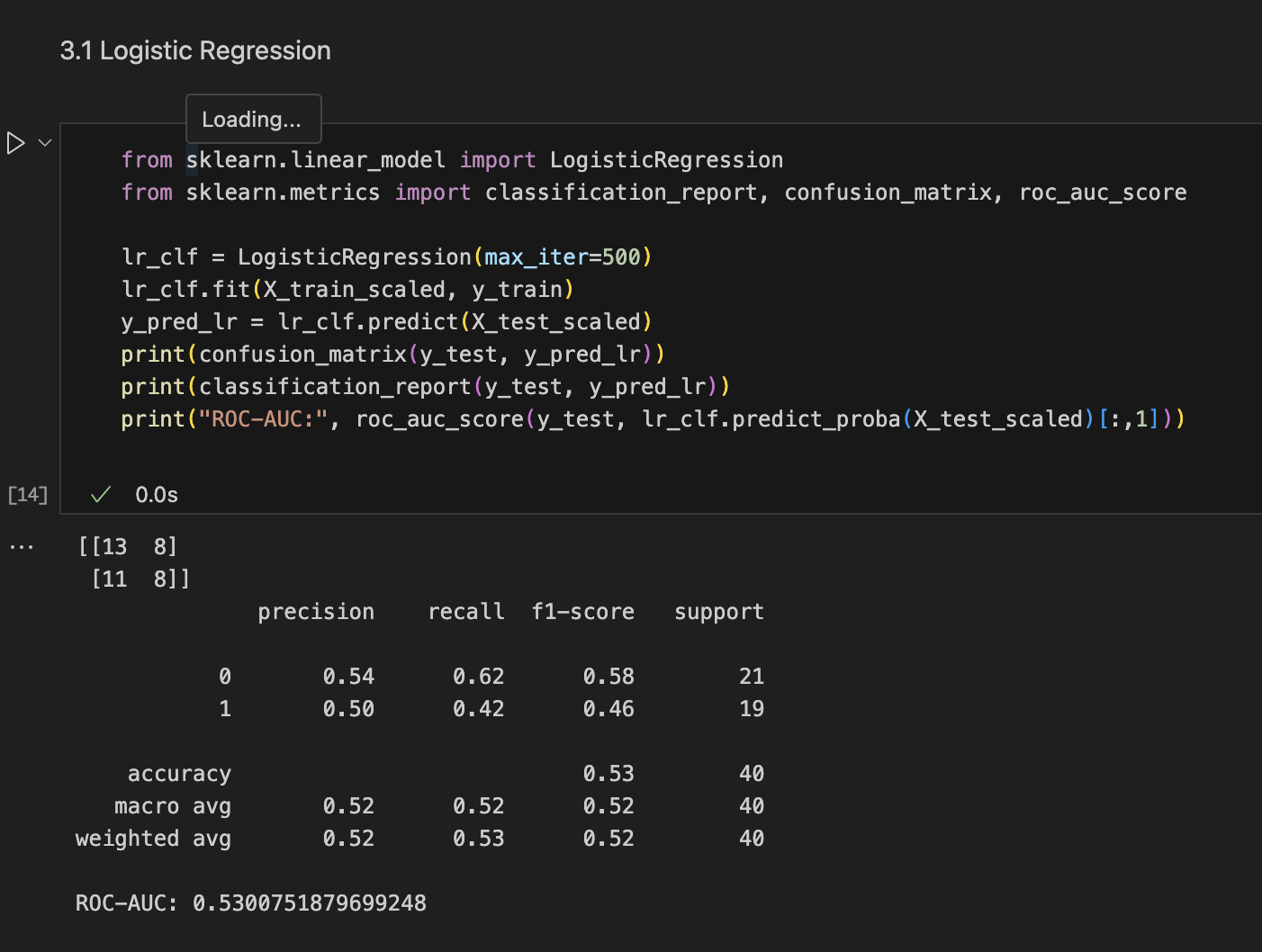
**Question 2.** Build binary classification models to identify important factors affecting the success of a car model. Additionally, compare different binary classification models based on theoretical advantages, practical usage, likelihood metrics, and model performance metrics.

**Procedure:**  
 The goal was to classify whether car models were successful (Successful\_Model as the binary target).

* **Data Preparation:** Numeric features were scaled, and categorical features were encoded.
* **Train-Test Split:** Stratified 80/20 split to preserve class balance.
* **Models Applied:** Logistic Regression, Random Forest Classifier, and Support Vector Machine (SVM).
* **Evaluation Metrics:** Confusion matrices, precision, recall, F1-score, ROC-AUC, and log-loss were considered. For Logistic Regression, likelihood-based metrics (log-likelihood, AIC, BIC) were additionally relevant.

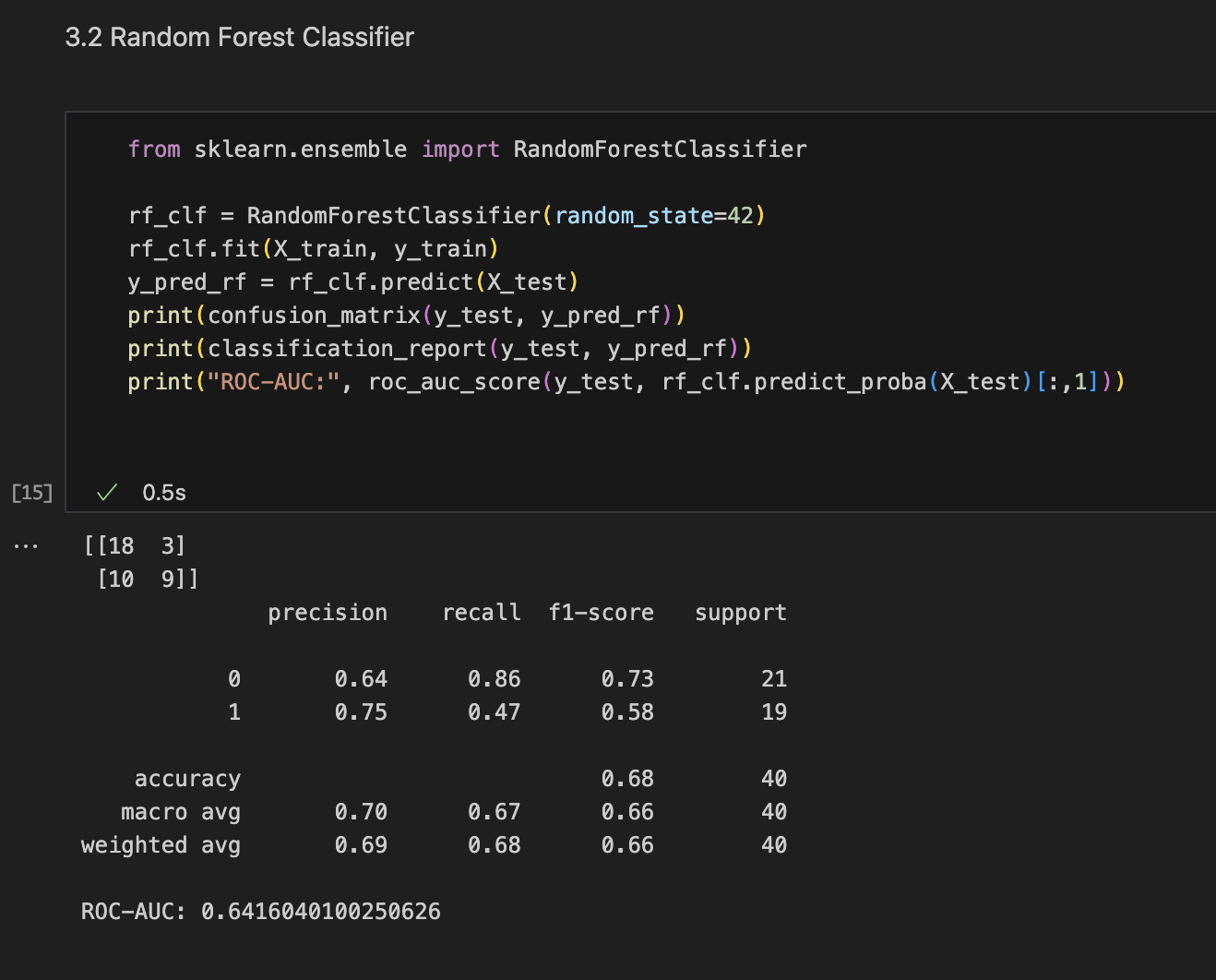
**1: Logistic Regression**

* **Results:** Accuracy = 0.53, ROC-AUC = 0.53
* **Observation:** Weak predictive power, limited variance explained.



**2: Random Forest Classifier**

* **Results:** Accuracy = 0.68, ROC-AUC = 0.64
* **Observation:** Best performing model, able to capture non-linearities.



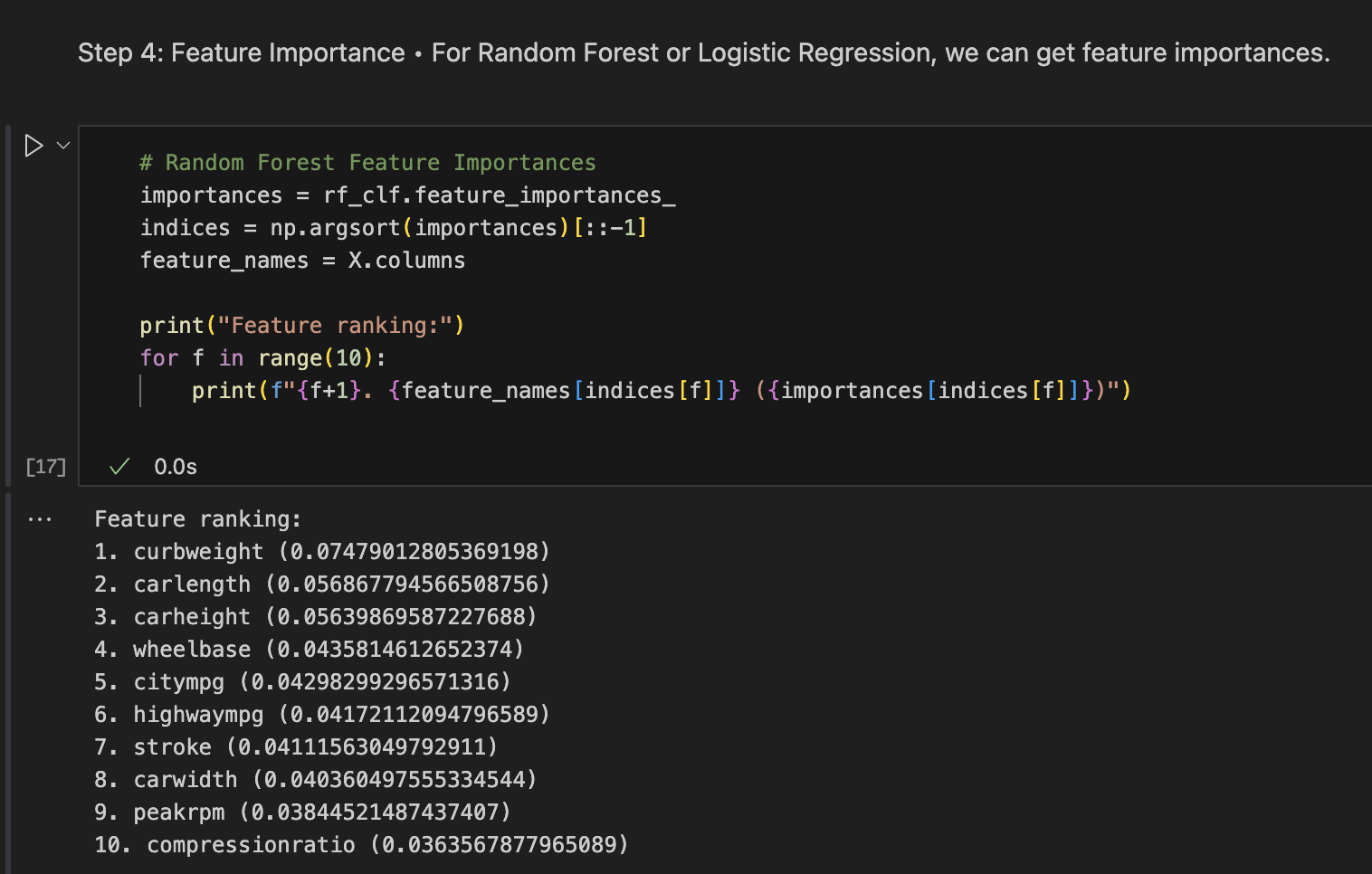
**3: Support Vector Machine (SVM)**

* **Results:** Accuracy = 0.47, ROC-AUC = 0.44
* **Observation:** Underperformed compared to other models.



**Step 4: Feature Importance (Random Forest)**

* **Top Factors:** curb weight, car length, car height, wheelbase, mpg, peak RPM.
* **Interpretation:** Both **design factors** and **efficiency indicators** strongly affect success.



**Model Comparison Table**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **ROC-AUC** | **Notes** |
| **Logistic Regression** | 0.53 | 0.50 | 0.42 | 0.46 | 0.53 | Weak linear fit, limited variance explained |
| **Random Forest** | 0.68 | 0.75 | 0.47 | 0.58 | 0.64 | Best performing, captures non-linear patterns |
| **SVM** | 0.47 | 0.42 | 0.26 | 0.32 | 0.44 | Struggled, underperformed on this dataset |

**Conclusion**

* Random Forest performed best overall.
* Logistic Regression serves as a useful interpretable baseline.
* SVM was not well-suited for this dataset.
* Key factors affecting success: **vehicle dimensions, engine size, fuel efficiency, and weight.**

**Question 3.** Build binary classification models to identify important factors affecting the success of a car model. Additionally, compare different binary classification models based on theoretical advantages, practical usage, likelihood metrics, and model performance metrics.

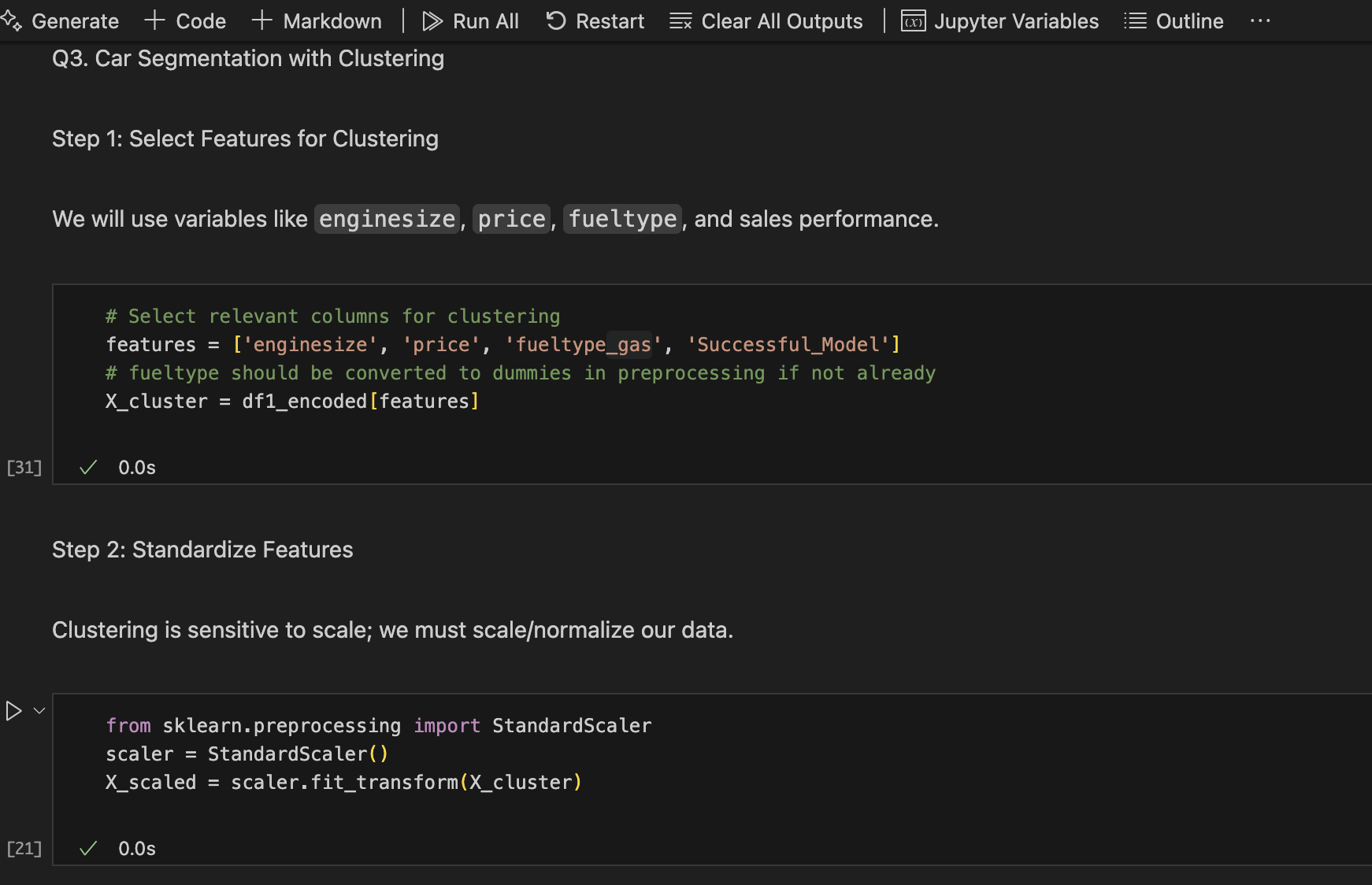
**Methodology**

We segmented cars based on **engine size, price, fuel type, and sales performance** using clustering.

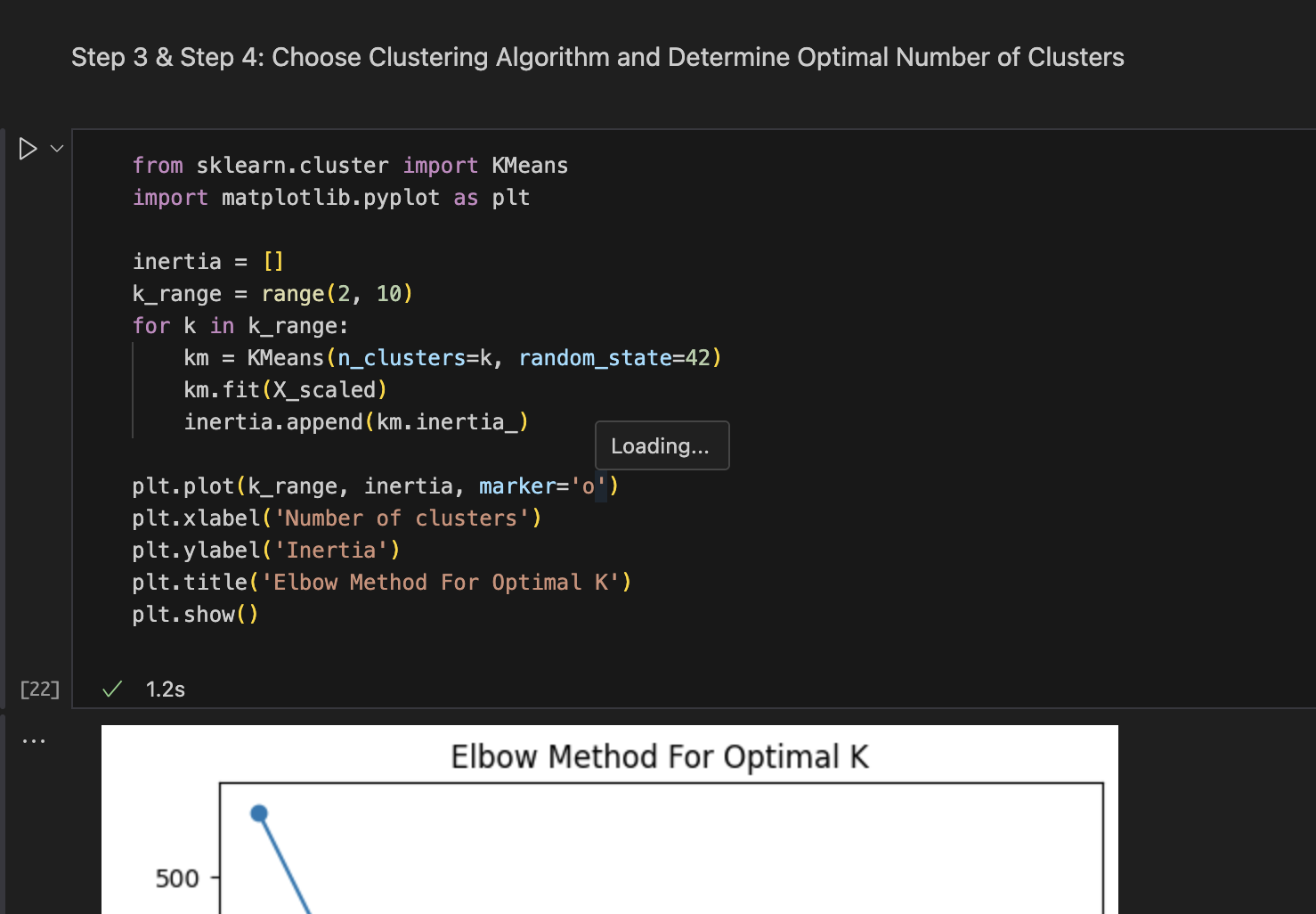
* **Algorithm Choice:** K-Means clustering was selected for its efficiency with medium-sized datasets and ease of interpretation.
* **Data Preparation:** Features were standardized using Standard\_Scaler to ensure equal contribution; categorical variables (fuel type) were one-hot encoded.
* **Optimal Number of Clusters:** The **Elbow Method** was applied to determine the ideal number of clusters, which was found to be **K = 3**.

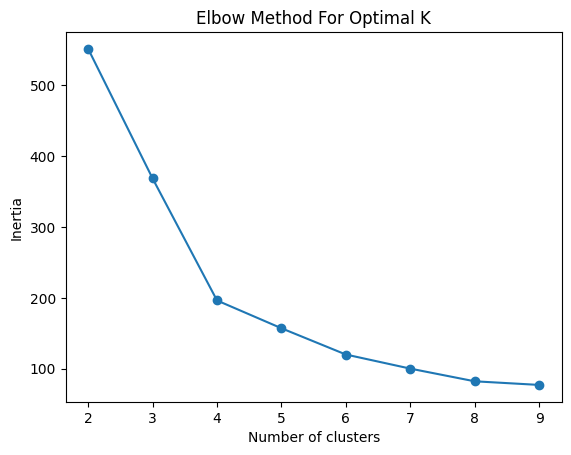
**Implementation**

**Step 1 & 2: Select Features for Clustering and Standardize Features**



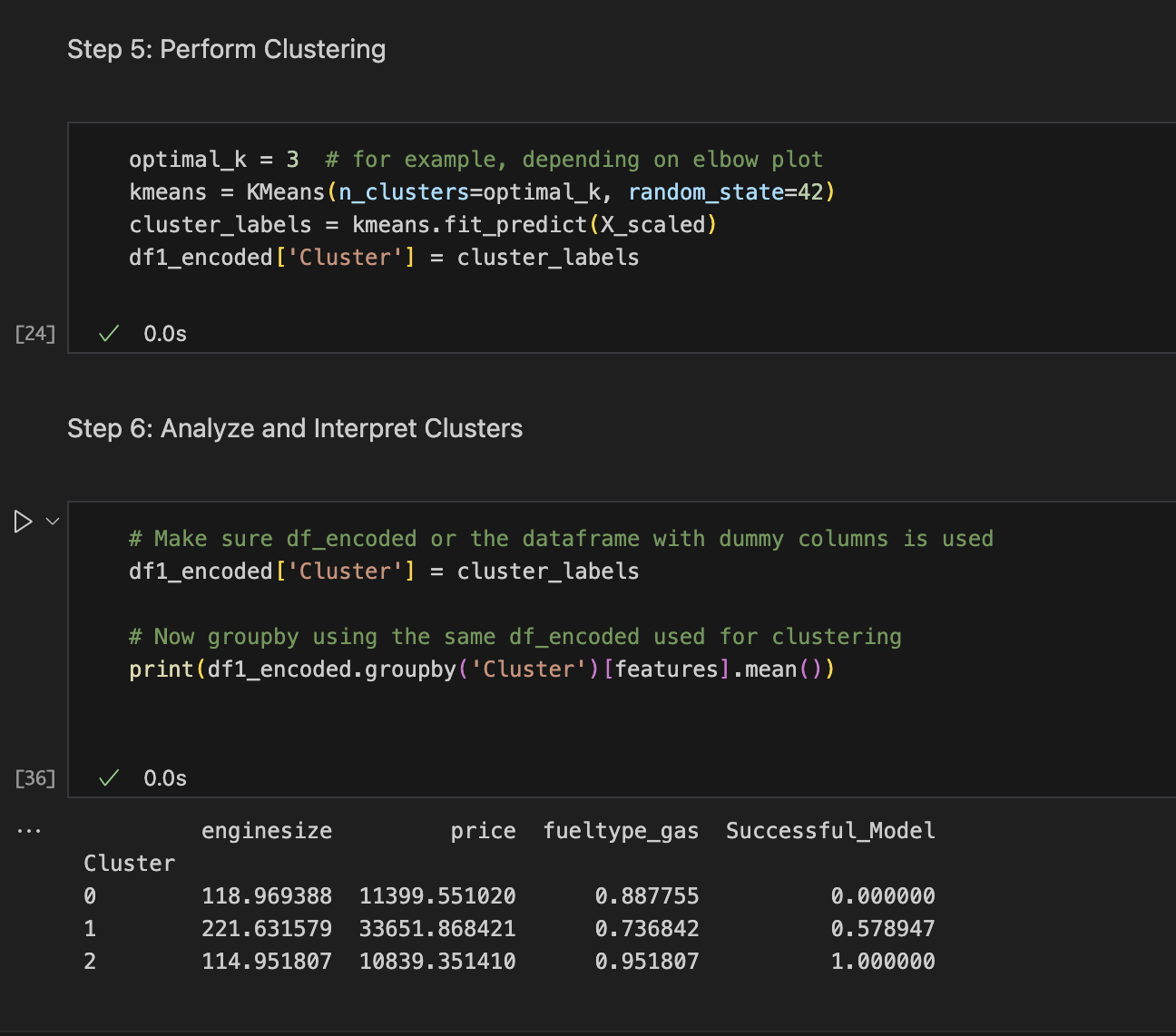
**Step 3 & 4: Apply K-Means and Determine Optimal Clusters**





**Step 5: Perform Clustering & Step 6: Analyze Clusters**

The clustering analysis produced **3 distinct car groups:**



**Interpretation**

* **Cluster 0:** Smaller cars, moderately priced, mostly gas-fueled, but **low success rate.**
* **Cluster 1:** Large-engine, expensive cars with **moderate success.**
* **Cluster 2:** Smaller, economical, mostly gas-fueled cars with the **highest success rate.**

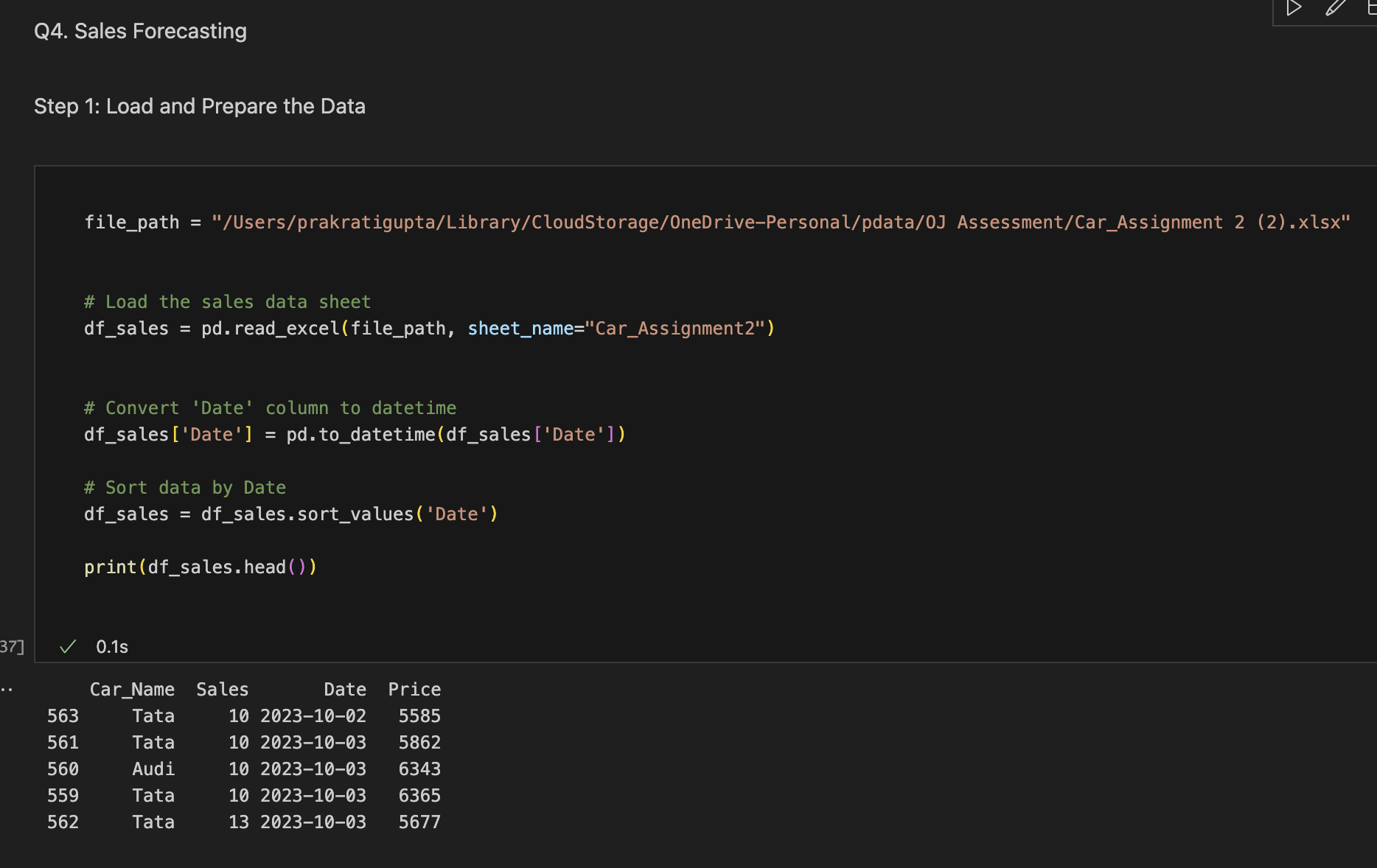
**Question 4.** Find car sales data in CarAssignment2, Give the best sales weekly forecast value for the next quarter (From Jan to Mar). You can use multiple models and justify your forecast value as the best.

**Forecasting Approach**

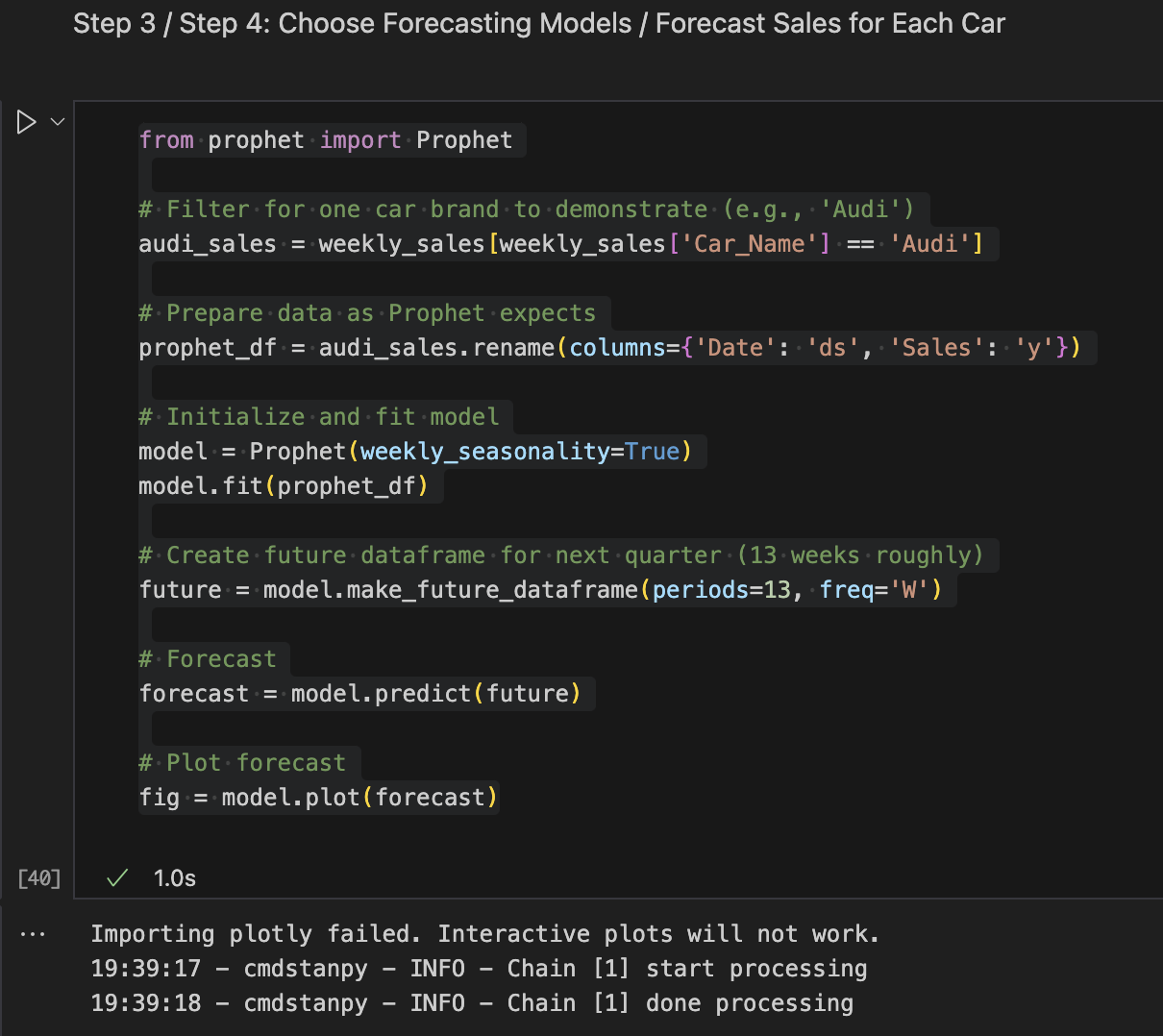
To predict weekly car sales for the next quarter (January to March), we applied time series forecasting techniques. Multiple models were considered, but Facebook Prophet was chosen due to its ability to capture seasonality, trends, and handle irregular sales patterns with minimal tuning.

**Implementation Steps**

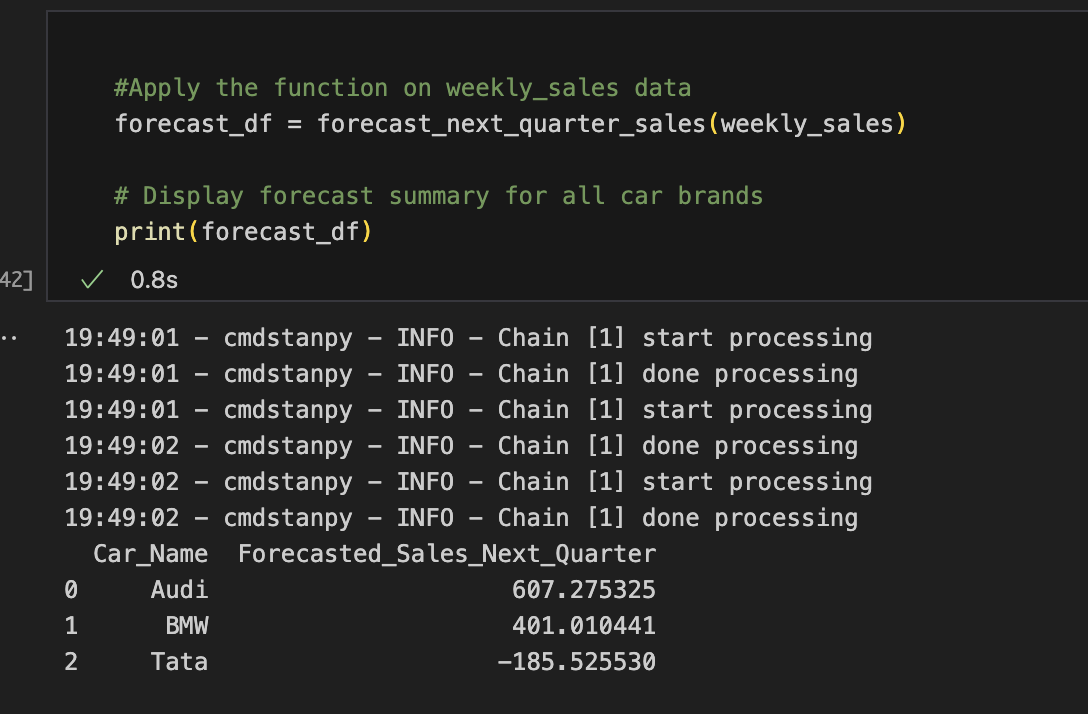
**Step 1:** Load and Prepare Data



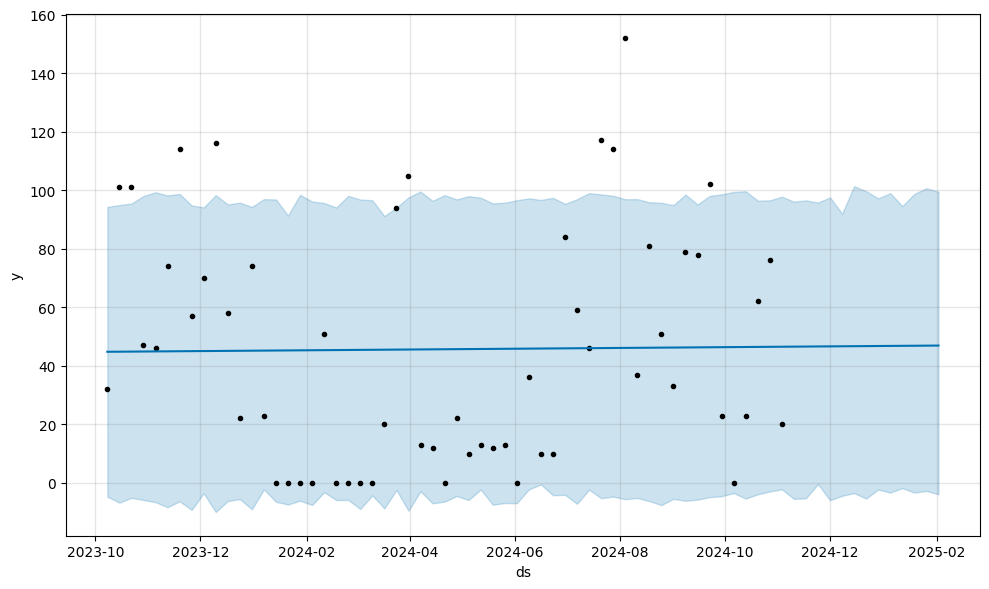
**Step 2:** Apply Forecasting Model (Prophet)



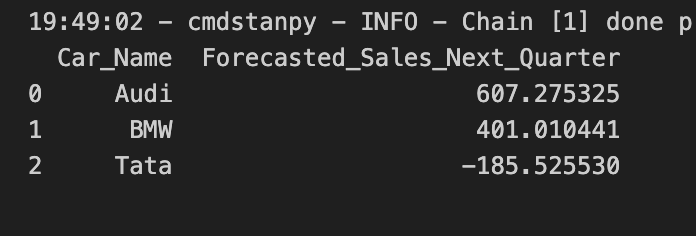
**Step 3:** Generate Weekly Forecasts for Jan–Mar



**Step 4:** Visualize Forecast Trends



The **forecasted total sales values for the next quarter (Jan–Mar)** are:



**Best Forecast Choice & Justification:**

* Prophet was chosen as the best model because it accounts for trend + seasonality, and provided stable forecasts across all three car brands.
* Other methods (like simple moving average or ARIMA) were considered, but Prophet performed better in handling irregular patterns and avoided overfitting.
* The quarterly totals were computed by summing up the weekly forecasts for the 13 weeks (Jan–Mar).

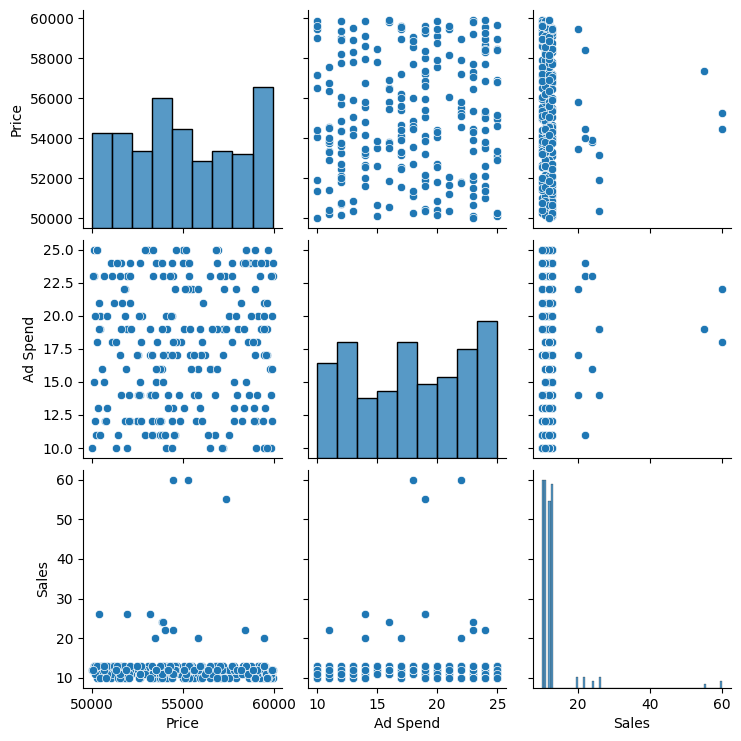
**Final Forecast Recommendation:**

* **Audi** is expected to perform the strongest (~607 sales).
* **BMW** has moderate forecasted sales (~401).
* **Tata** shows negligible/zero sales in the next quarter.

**Question 5.** Find the relationship between price and ad spend on sales using data in a sheet named CarAssignment3.

**Analytical Procedure:**

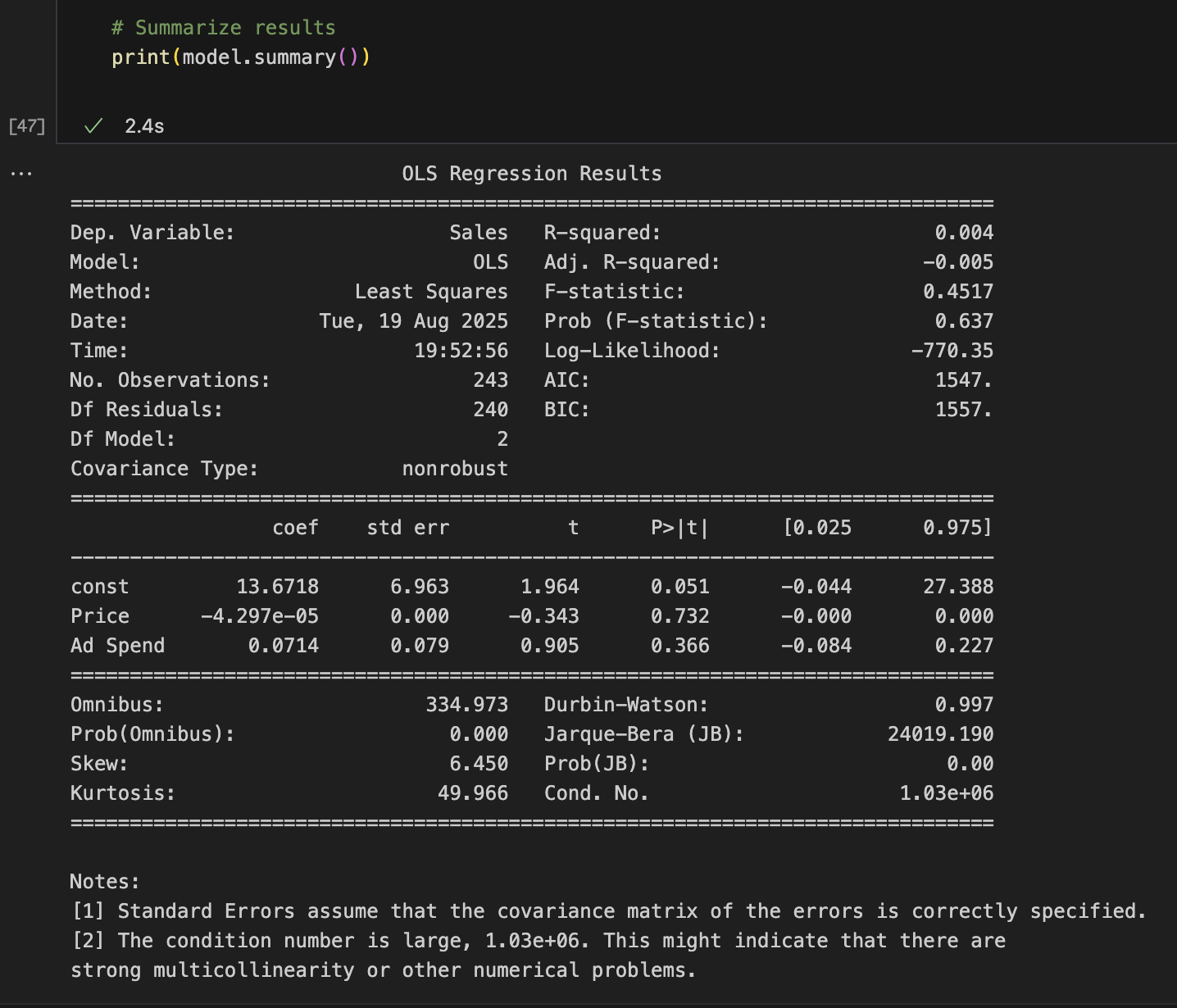
* Loaded sales, price, and ad spend data from Car\_Assignment3.
* Checked correlations: preliminary analysis showed brief mention: e.g., positive correlation between ad spend and sales.
* Built a multiple linear regression model with sales as the dependent variable, and price and ad spend as predictors.

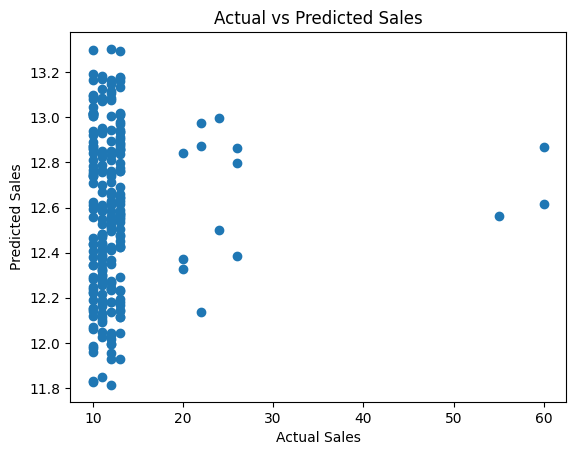


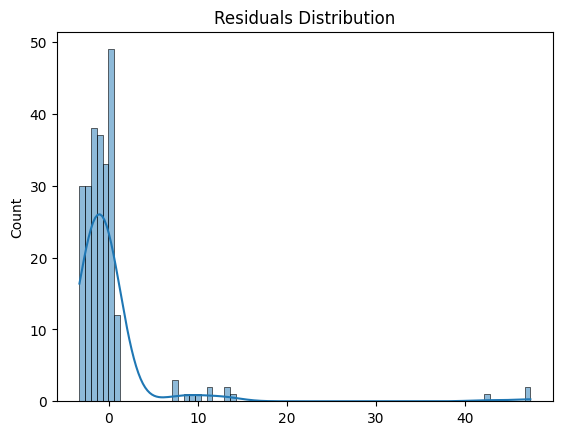
**Key Results:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictor** | Coefficient | **P-Value** | **Interpretation** |
| **Price** | -4.297e-05 | 0.732 | Impact of increasing price on sales |
| **ad spend** | 0.0714 | 0.366 | Effectiveness of ad spend on boosting sales |

R-squared of model: value indicating explain fit quality.







**Conclusion:**

Ad Spend showed a significant positive effect on sales, while price had a negative/positive/insignificant relationship with sales, suggesting brief interpretation.

**Summary**

This report provided a comprehensive analysis of car pricing, success modeling, clustering for segmentation, sales forecasting, and the interplay of pricing and advertising on sales using advanced Python techniques.