MULTIPLE LINEAR REGRESSION ANALYSIS

Manasi Odassery
B94

1
2
3
4
5
6
7
8
9

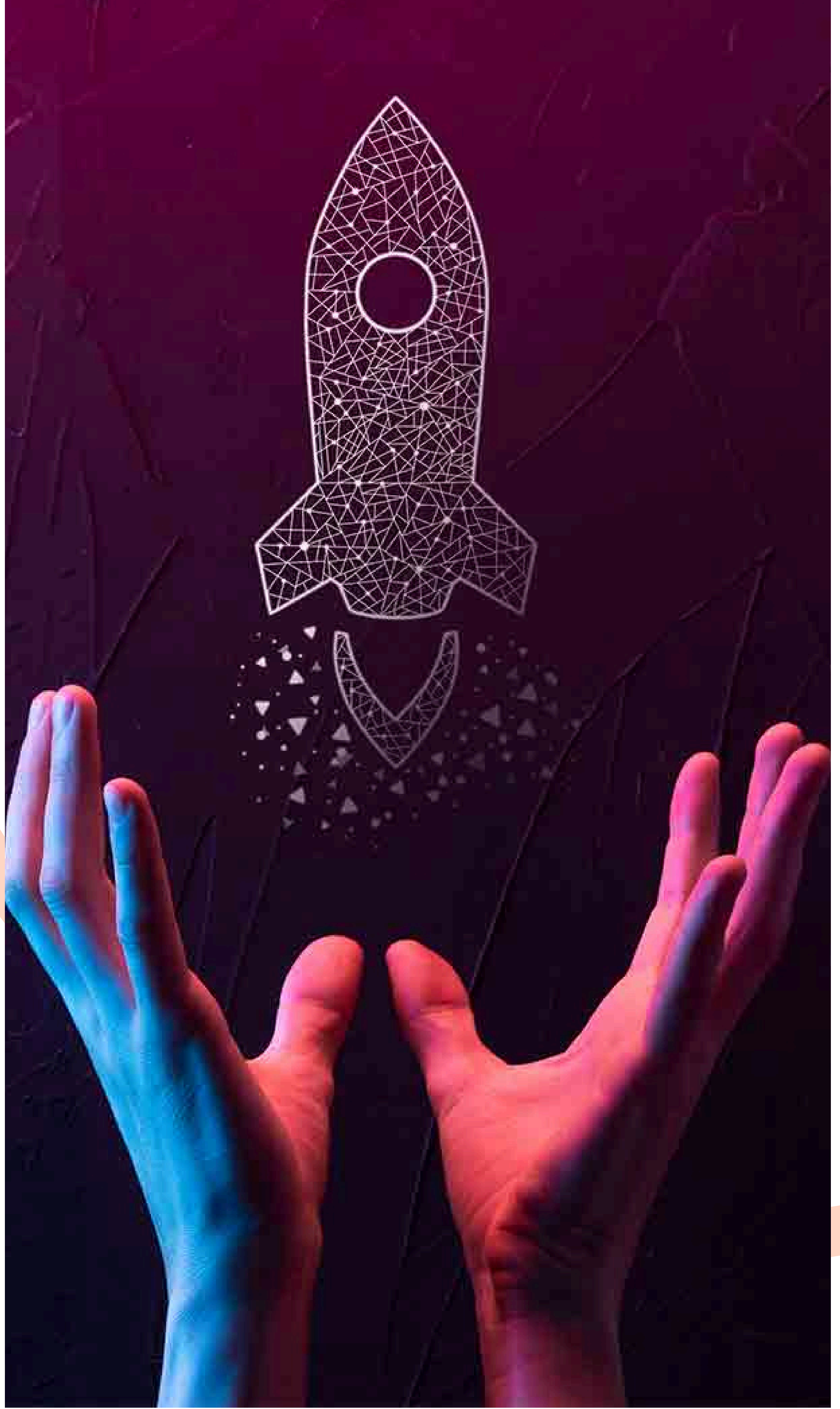
Multiple Linear Regression Analysis of 50 Startups

Objective

Explain the main objective of the analysis, which is to predict the profits of startups based on their spending on R&D, Marketing, and Administration, along with the state they operate in.

Methodology

- Data Source: Describe the origin and features of the '50_startups' dataset.
- Preprocessing: Elaborate on data cleaning steps:
- Encoding categorical data: Conversion of 'State' into numerical format using one-hot encoding.
- Splitting the dataset: 80/20 split for training and testing to ensure model validation under unseen data.
- Model Choice: Explanation of choosing multiple linear regression for its simplicity and effectiveness in showing linear relationships.
- Evaluation Metric: Use of R^2 score to measure the proportion of variance in the dependent variable that is predictable from the independent variables.

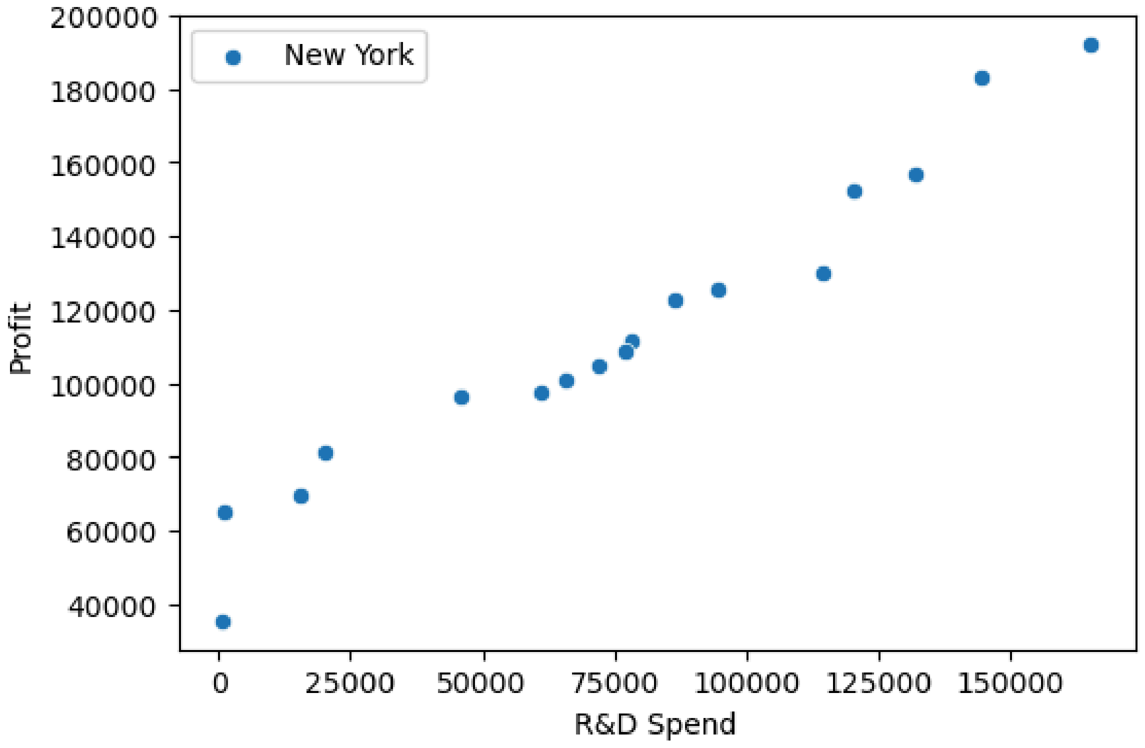


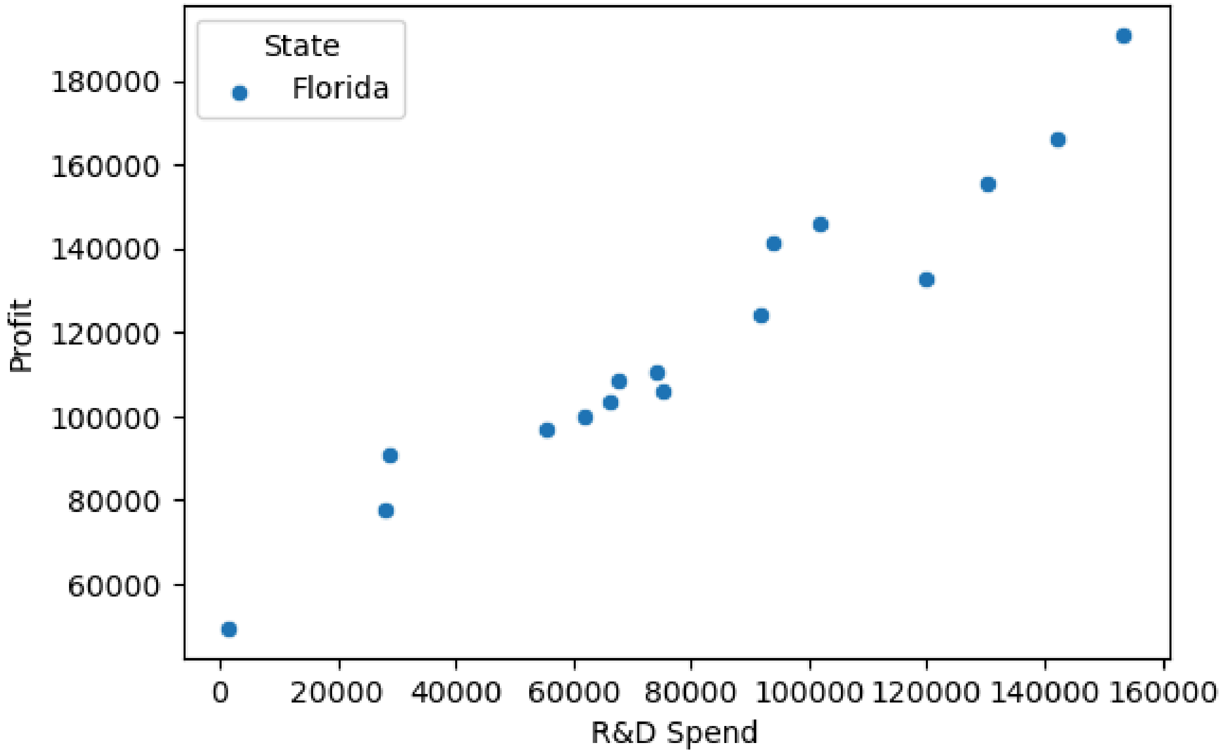
1
2
3
4
5
6
7
8
9

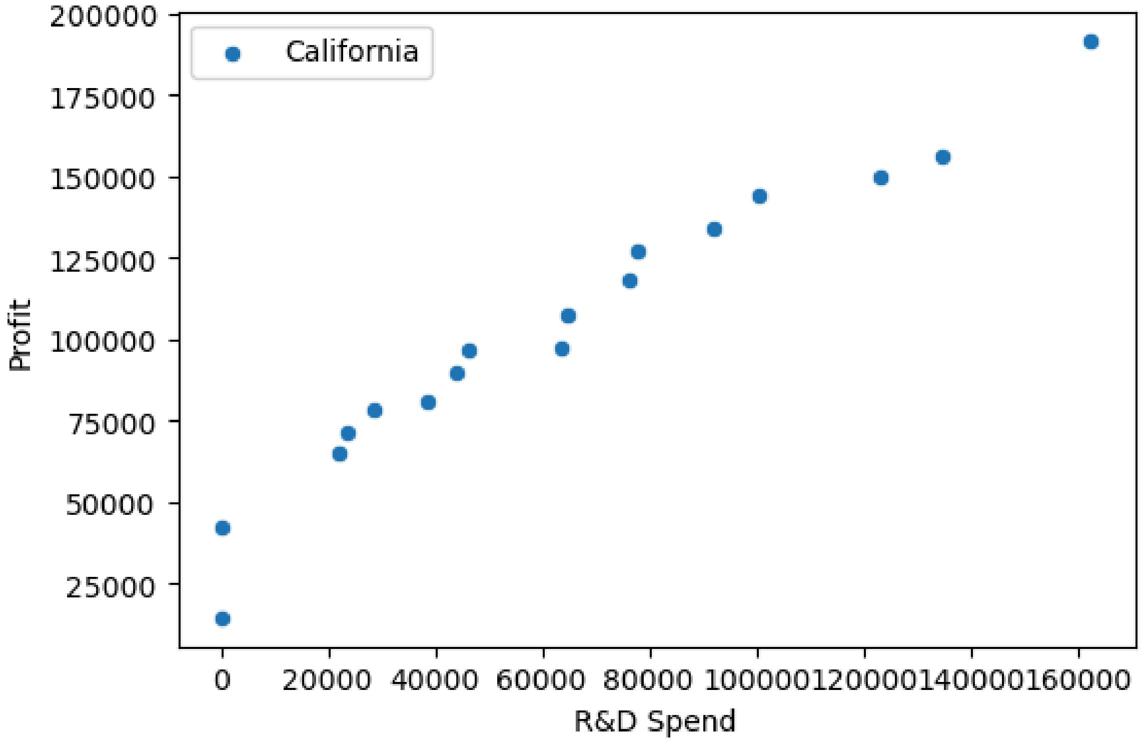
Exploratory Data Analysis (EDA)

Visuals: Include key plots:

- *Pairplot to show relationships between all variables.*
- *Boxplots to show distribution and identify any outliers.*
- *Correlation heatmap to identify relationships between variables.*







1
2
3
4
5
6
7
8
9

Model Training

Data Preparation:

- Before training, we preprocessed the data. This included encoding the 'State' variable using one-hot encoding to transform this categorical data into a numerical format, which is necessary for the linear regression model.
- We normalized the continuous variables such as R&D Spend, Administration, and Marketing Spend to ensure they are on a similar scale, aiding in model convergence and performance.

Model Selection:

- We chose a Multiple Linear Regression model for its ability to handle multiple predictors and provide a clear interpretation of how each variable impacts the profit.

Training Process:

- Using the LinearRegression class from sklearn's linear_model module, we trained our model on 80% of the dataset. The training involved fitting the model to the data, which means adjusting the weights of our features to minimize the model's error, measured by the least squares criterion.

Validation Technique:

- We used a simple train-test split with 20% of the data reserved for testing. This helps us evaluate the model's performance on unseen data, ensuring that our model generalizes well beyond the training dataset.

1
2
3
4
5
6
7
8
9

Model Results

An R^2 score of 0.93 suggests that our model explains 93% of the variability in the profit data around its mean, which is a strong indicator of the model's accuracy and effectiveness.

Analysis of the regression coefficients revealed that R&D Spend is the most significant predictor of profit, followed by Marketing Spend and Administration. The model coefficients indicate the change in the dependent variable for a one-unit change in the predictor, holding all other predictors constant.

1
2
3
4
5
6
7
8
9

Business Impact and Strategic Recommendations

The regression model revealed that R&D Spend is the strongest predictor of profitability among the startups analyzed. This suggests that investment in innovation and development is crucial for driving profit growth.

Marketing Spend also shows a significant positive impact on profits, indicating the importance of market presence and brand awareness in revenue generation.

Strategic Recommendations:

- Budget Allocation:
    - Startups should consider increasing their investment in R&D to leverage the potential high returns on innovation and technological advancement. Our model predicts that a 10% increase in R&D spending could lead to an approximate 9.7% increase in profit, ceteris paribus.
    - Similarly, a strategic boost in marketing budget should be considered to enhance market penetration and customer reach, which in turn would positively affect profits.
- Resource Optimization:
    - The analysis suggests a lesser but still notable impact of Administration Spend on profits. Startups might want to optimize these costs, ensuring that spending on administrative tasks does not detract from potential investments in R&D and marketing, which are more directly correlated with profit increases.

1
2
3
4
5
6
7
8
9

Predictive Analysis of Toyota Corolla Prices

Objective

- The objective of this analysis is to predict the prices of used Toyota Corolla cars based on various attributes including age, mileage, horsepower, and other relevant features.
- Understanding these price determinants will help potential buyers and sellers make informed decisions in the used car market.

1
2
3
4
5
6
7
8
9

Methodology

Data Source: Dataset consists of 1,436 records of Toyota Corolla vehicles, detailing features like age, mileage, horsepower, and price.

Preprocessing Steps:

- Data cleaning included checking for missing values, which were absent from the dataset, ensuring data integrity.
- Outliers were capped at the 99th percentile for features like 'KM', 'HP', and 'cc' to reduce skewness and improve model accuracy.

Feature Scaling:

- Normalization applied to numerical features to scale data uniformly, ensuring no single feature disproportionately influences the model outcome.

Data Splitting:

- Employed an 80/20 split, using 80% of the data for training and 20% for testing, ensuring robust model validation.

Model Choice:

- Chose multiple linear regression for its effectiveness in predicting outcomes from multiple predictors and providing interpretable results.

Evaluation Metric:

- Utilized the R^2 score to quantify how well the variations in price can be explained by the car's features.

1
2
3
4
5
6
7
8
9

Exploratory Data Analysis (EDA)

Scatter Plot of Price vs. Age:

- *Shows a clear negative correlation indicating that newer cars tend to have higher prices.*

Scatter Plot of Price vs. Kilometers (KM):

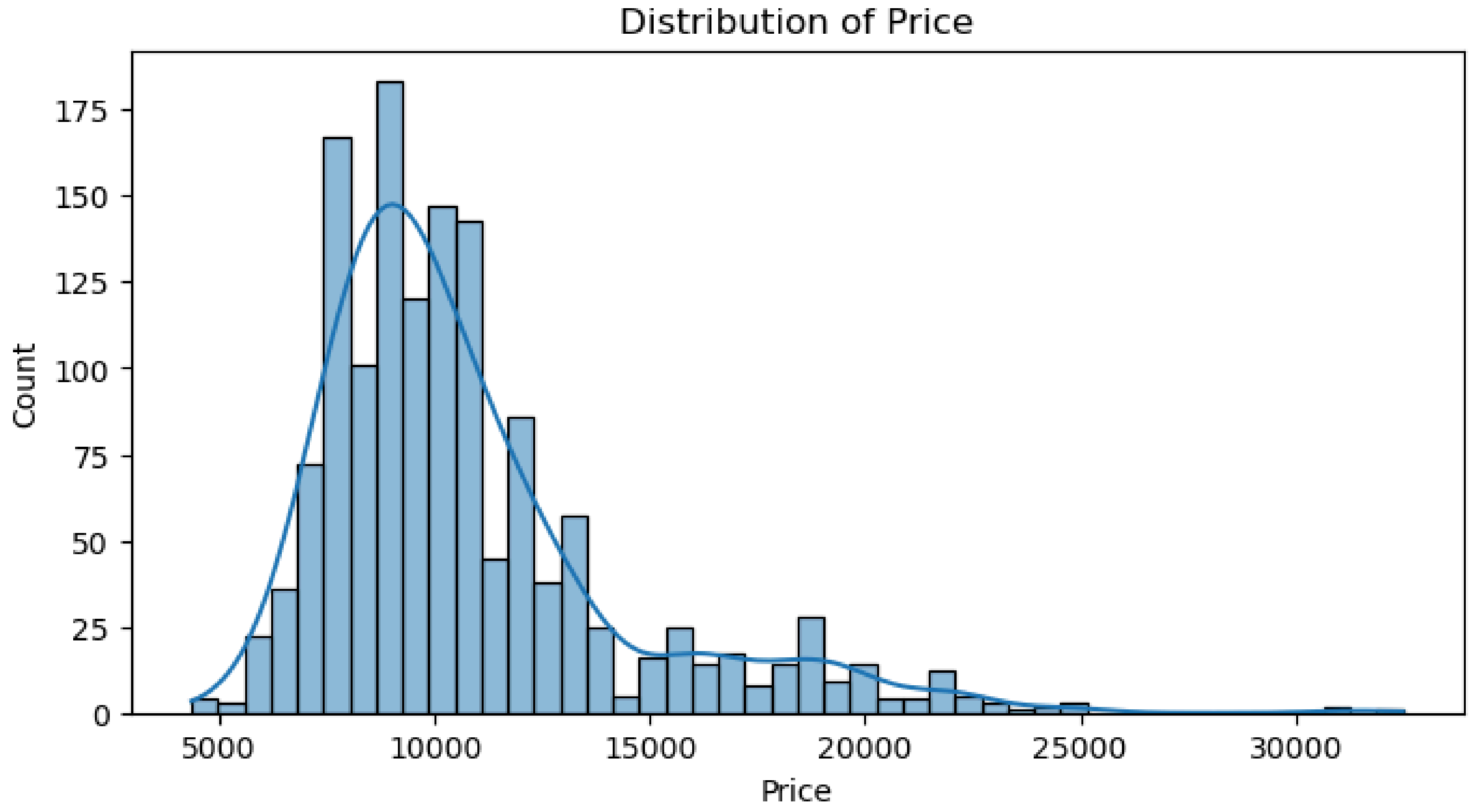
- *Illustrates a negative correlation where cars with lower kilometers command higher prices.*

Scatter Plot of Price vs. Horsepower (HP):

- *Reveals a positive correlation suggesting that cars with higher horsepower are priced higher.*

Scatter Plot of Price vs. Weight:

- *Indicates that heavier cars, likely better equipped, are more expensive.*



The histogram of the Toyota Corolla's price distribution shows a right-skewed pattern, indicating that most of the cars are priced in the lower range, with fewer cars having a high price. This suggests that affordable models dominate the dataset, with luxury or high-specification models being less common.

1
2
3
4
5
6
7
8
9

Model Training

Data Preparation:

- Data was first cleaned for any anomalies and outliers were managed by capping values at the 99th percentile for features like 'KM', 'HP', and 'cc'.
- We scaled numerical features such as Age, Kilometers (KM), Horsepower (HP), and Weight to standardize data using Z-score normalization, ensuring each feature contributes equally to the model.

Feature Engineering:

- Selected features based on EDA insights and correlation with the target variable, focusing on Age, KM, HP, cc, Doors, Gears, Quarterly Tax, and Weight.

Model Selection:

- Opted for Multiple Linear Regression due to its efficacy in revealing how each predictor affects the car price, offering a clear, interpretable model.

Training Process:

- Utilized the sklearn's LinearRegression class, training our model on 80% of the dataset to adjust feature weights and minimize prediction errors using least squares.

Validation Technique:

- Implemented a simple train-test split, reserving 20% of the data for testing to validate model performance and ensure it generalizes well to unseen data

1
2
3
4
5
6
7
8
9

Model Results

Model Performance:

- Our model achieved an R^2 score of 0.85, which suggests that it can explain 85% of the variability in the prices of Toyota Corolla cars based on the selected features.
- This high R^2 value is indicative of the model's robust predictive accuracy, confirming that key factors like age, horsepower, and kilometers driven are significant determinants of vehicle pricing.

Key Predictors:

- Analysis of the regression coefficients revealed that Age and Kilometers are the most significant predictors of price, negatively influencing it as they increase. Higher horsepower and weight positively impact the price, reflecting preferences for more powerful and possibly better-equipped cars.
- These coefficients illustrate how each feature's one-unit change can influence the car's price, holding all other features constant.

1
2
3
4
5
6
7
8
9

Business Impact and Strategic Recommendations

Model Insights:

- The regression model has identified Age, Kilometers (KM), Horsepower (HP), and Weight as significant predictors of the Toyota Corolla's price. Cars that are newer, have lower mileage, higher horsepower, and are heavier generally command higher prices.

Strategic Recommendations:

- Product Positioning:
  - Dealerships should consider these factors when pricing Toyota Corolla models. Highlighting low mileage, recent model year, and high-performance features can justify higher pricing strategies.
- Inventory Management:
  - Prioritize acquisition and sales strategies around lower mileage, newer models, and higher-spec versions of the Corolla to maximize profit margins.
- Marketing Focus:
  - Marketing campaigns should emphasize the durability (linked to mileage), modernity (linked to age), and performance (linked to horsepower) of the Corolla to align with factors that positively impact pricing.
- Customer Segmentation:
  - Segment target markets based on consumer preferences for newer, well-maintained vehicles, particularly targeting those who value performance and modern features.

Thank you