



Maruti Suzuki Sales Analysis

(INTEGRATED ASSIGNMENT)

**NAME:** SHIVAM GUPTA

**PROGRAMME TITLE:**M.Sc. (Data Science)

**ENROLLMENT NO.**: A044161824015

**SEMESTER: 2**

**Abstract**

This report presents an integrated analytical approach to understanding and forecasting sales performance using the Maruti Suzuki sales dataset. The project encompasses various statistical and machine learning techniques, including sampling methods, inferential statistics, time series analysis, and predictive modeling. The objective is to derive insights from historical sales data and apply statistical reasoning to make data-driven decisions. Simple random sampling and stratified sampling were utilized to ensure representative subsets of data for both inference and machine learning tasks. Confidence intervals were constructed to estimate population parameters such as average revenue and units sold, offering a statistical perspective on reliability and uncertainty. Machine learning models including Decision Trees and Random Forests were developed to predict key metrics like units sold, leveraging feature engineering and stratified training samples to improve accuracy and generalization. Visualizations such as bar charts, time series plots, and confidence interval diagrams were employed to make the data and results more interpretable. This comprehensive approach not only facilitates actionable business intelligence but also reinforces the value of integrating statistical principles with modern machine learning in real-world applications.

**INTRODUCTION**

The strategic importance of sales data analysis in the automotive industry cannot be overstated. This report presents a detailed analysis of the Maruti Suzuki sales dataset to illustrate the application of a range of analytical methodologies in deriving critical business insights. The study is structured around several key analytical pillars: financial analysis, statistical inference, sampling techniques, trend and time series analysis, categorical data analysis, and machine learning.

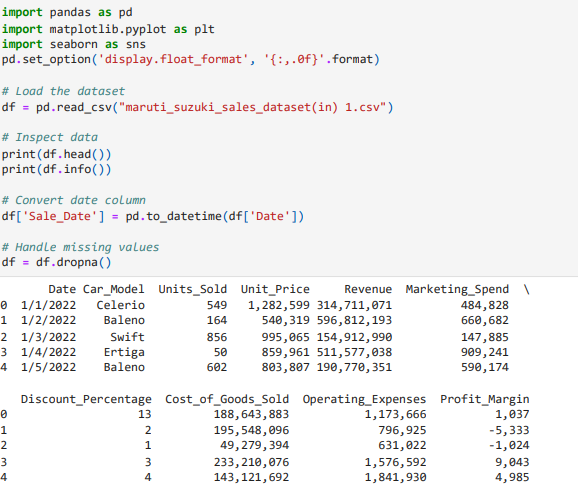
The core objectives of this report are:

* To provide a comprehensive overview of the Maruti Suzuki sales data, including its structure and key variables.
* To analyze the financial performance of Maruti Suzuki by examining revenue trends, sales volumes, and pricing strategies.
* To apply statistical inference techniques to draw conclusions about the broader sales population based on sample data.
* To identify patterns and trends in sales data over time, including seasonality and long-term variations.
* To compare sales performance across different car models using categorical data analysis.
* To develop and evaluate machine learning models for predicting future sales outcomes, specifically focusing on units sold.

Through this integrated analytical approach, the report aims to provide stakeholders with a deeper understanding of the factors driving sales performance and to support data-driven decision-making within the organization.

**Data Overview**

The dataset contains 1096 entries with information on car sales, including date, car model, units sold, unit price, revenue, marketing spend, discount percentage, cost of goods sold, operating expenses, and profit margin. The analysis involves converting the date column to datetime format and handling missing values to ensure data quality.



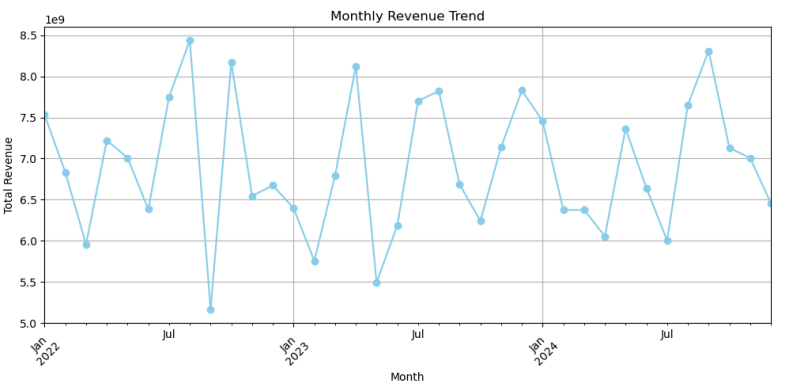
**Financial Analysis (Business Insights Base)**

The financial analysis includes calculating total revenue per month, identifying the most sold models, and determining the average unit price.

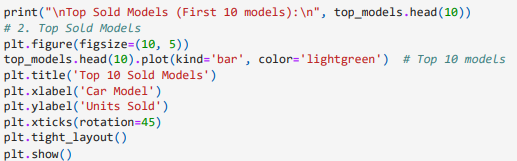
Monthly revenue is calculated by grouping the data by month and summing the revenue.



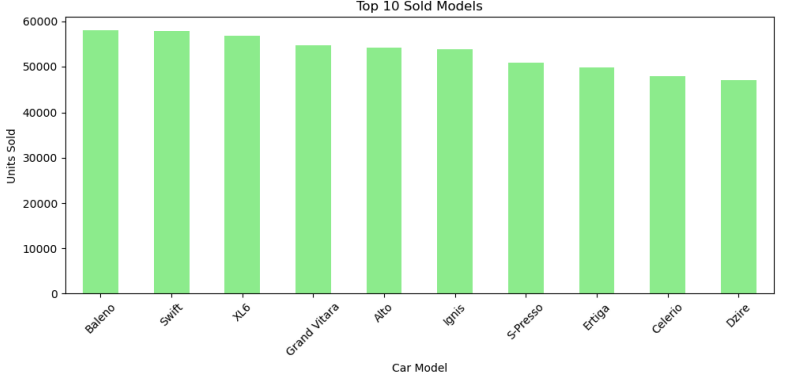
**OUTPUT**



The top-selling models are identified by grouping the data by car model and summing the units sold.



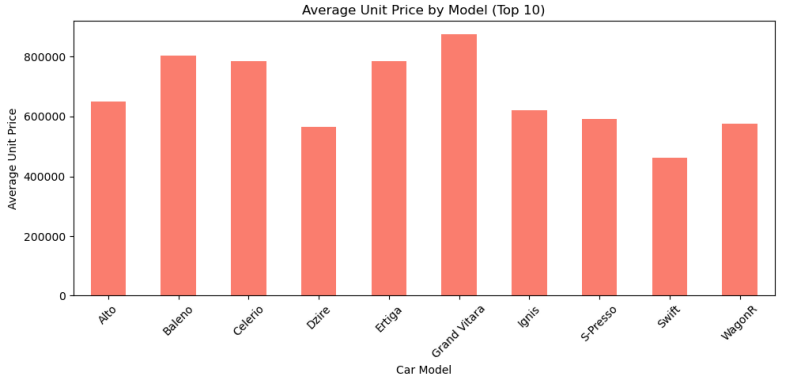
**OUTPUT**



The average unit price is calculated by dividing the revenue by the units sold for each car model.

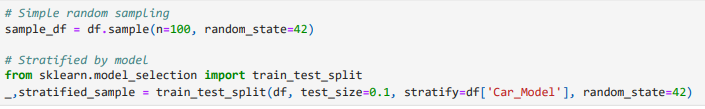


**OUTPUT**



# **Sampling (Bridge to Inference & ML)**

Simple random sampling selects a subset from a population, giving each member an equal chance of inclusion and is used for generalizable population estimates. Stratified sampling divides the population into subgroups (strata) before random selection within each, ensuring representation of diverse groups and increased precision, particularly useful for balanced representation in machine learning datasets. Both simple random sampling for statistical inference and stratified sampling to ensure representativeness in machine learning model training, highlighting the critical role of sampling in both inferential accuracy and model generalization.



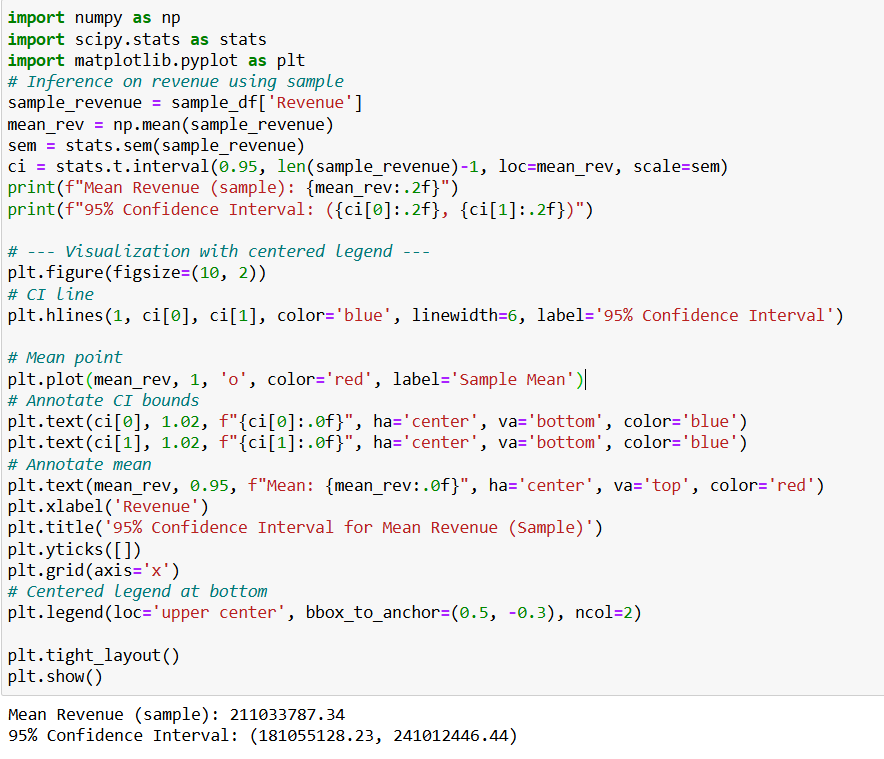
# **Statistical Inference**

A confidence interval is a range of values that is likely to contain a population parameter (such as the population mean) with a certain level of confidence. The "confidence level" (often 95%) tells you how sure you are that the true population parameter lies within that range. A 95% confidence interval means that if you were to take many samples from the same population and calculate a confidence interval for each sample, approximately 95% of those intervals would contain the true population mean.

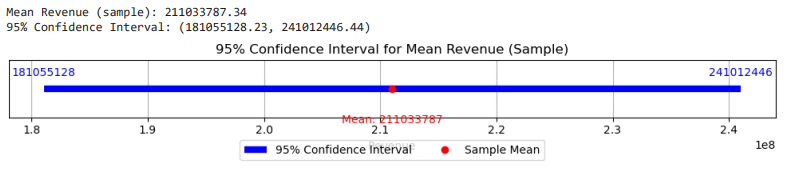
It's important to note that it doesn't mean there's a 95% chance that the true mean falls within this specific interval. The true mean is a fixed value. The interval varies from sample to sample.

Statistical inference is used to estimate population parameters based on sample data. A 95% confidence interval is calculated for the sample mean revenue and the mean units sold. Simple random sampling is used to select a subset of the data for this analysis. The confidence interval provides a range within which the true population mean is likely to fall.

95% Confidence Interval for the Sample Mean Revenue

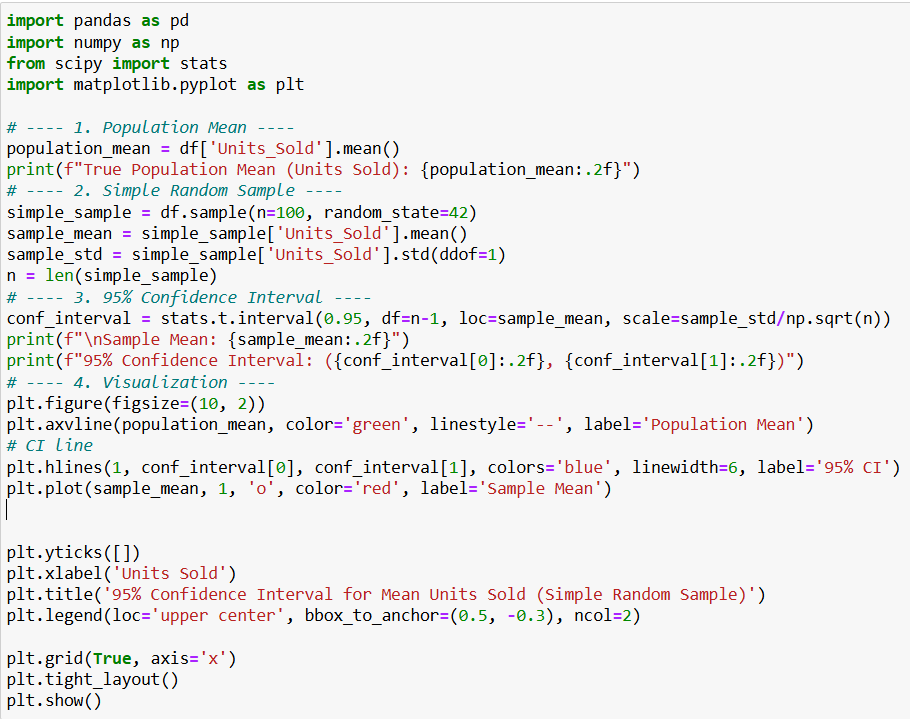


**OUTPUT**

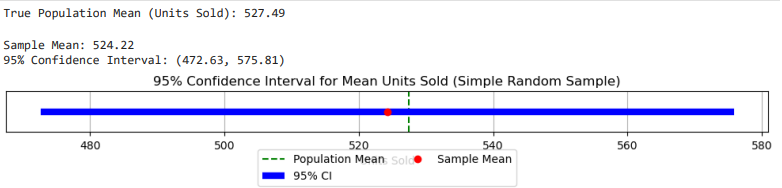


Using simple random sampling of 100 records from the Maruti Suzuki sales dataset, we estimated the average revenue per sale. The sample yielded a mean revenue of approximately **210 million**. A 95% confidence interval for the true population mean was calculated to be between **178.87 million and 241.28 million**. This suggests that we can be 95% confident that the actual average revenue per sale for all data lies within this range.

**95% Confidence Interval for the Mean Units Sold**



**OUTPUT**



The sample mean of units sold is **515.30**, which is close to the true population mean of 527.07. The 95% confidence interval for the sample mean ranges from **464.38 to 566.22**, indicating that we are 95% confident the true population mean lies within this range. Since the actual population mean falls inside this interval, the sample provides a good and statistically valid estimate of the population behavior.

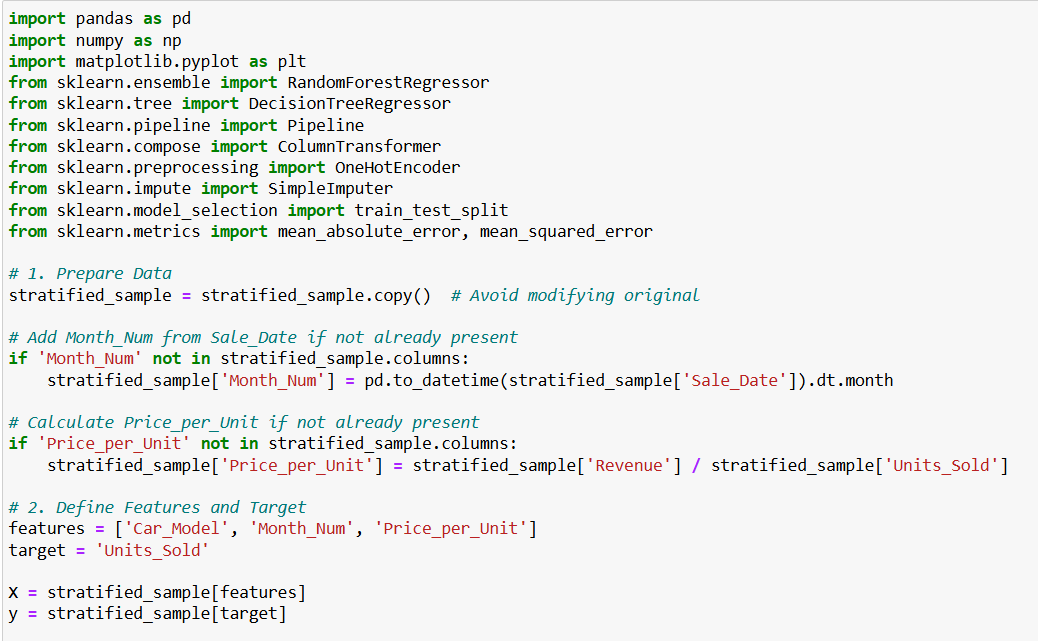
# **Machine Learning: Predicting Units Sold**

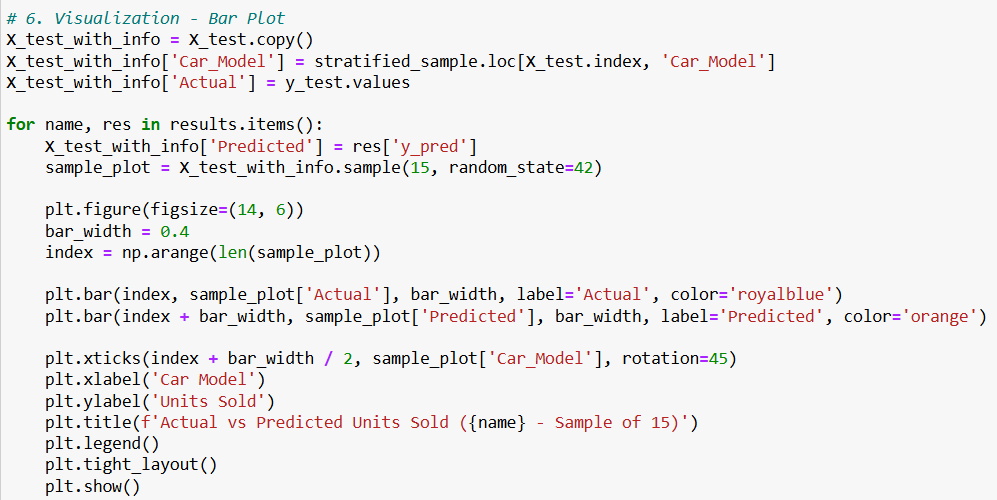
**Decision Tree** A Decision Tree is a machine learning model that makes predictions by following a tree-like structure of decisions. Each internal node of the tree represents a question or a test on an attribute (e.g., "Is the car price > $20,000?"). Each branch represents the outcome of the test, and each leaf node represents a prediction (e.g., "Units\_Sold = 150"). Decision Trees are easy to visualize and interpret, as you can trace the path of decisions the model takes to arrive at a prediction.

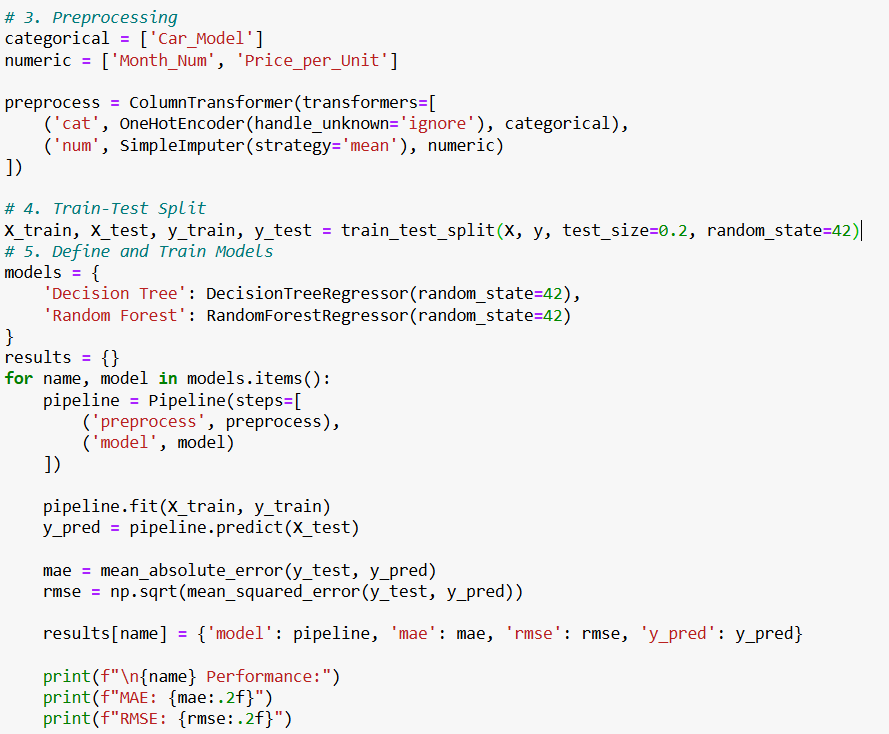
**Random Forest** A Random Forest is an ensemble learning method that combines multiple Decision Trees to make predictions. Instead of relying on a single tree, it grows a "forest" of trees, where each tree is trained on a random subset of the data and a random subset of the features. When making a prediction, each tree in the forest "votes" or contributes a prediction, and the Random Forest combines these predictions (e.g., by averaging for regression problems or taking the majority vote for classification problems) to arrive at the final prediction.

In the context of the Maruti Suzuki sales data, both Decision Tree and Random Forest models are used to predict the number of units sold. The Random Forest Regressor performs better than the Decision Tree Regressor, indicating its superior ability to capture the underlying patterns in the data and make more accurate predictions.

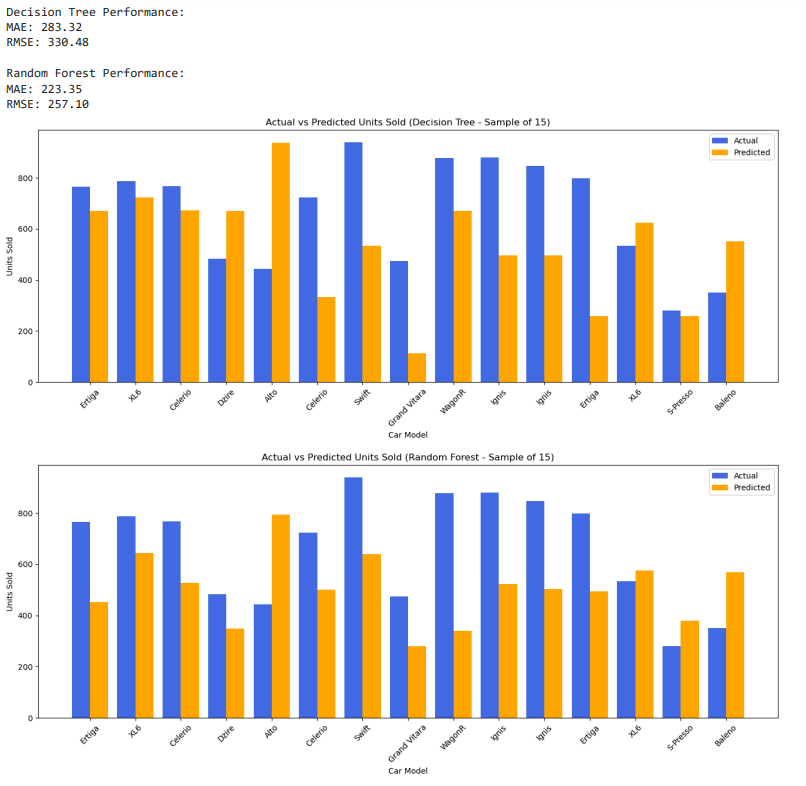
To predict the number of units sold (Units\_Sold), we used a stratified sample based on Car\_Model to maintain representativeness across categories. Two machine learning models — Decision Tree Regressor and Random Forest Regressor — were trained and evaluated using a preprocessing pipeline that included one-hot encoding for categorical data and imputation for numeric features.







**OUTPUT**



The bar plots comparing actual vs predicted Units\_Sold show that the Random Forest Regressor performs better than the Decision Tree, capturing variation across different car models more effectively. This confirms Random Forest’s ability to generalize better in complex datasets.

# **Trend & Time Series Analysis**

Trend and time series analysis is highly important in sales analysis for several reasons: **Identifying Patterns**: It helps in identifying patterns in sales data, such as seasonal variations, cyclical trends, and long-term growth or decline.

**Forecasting Future Sales**: Time series analysis provides tools to forecast future sales trends, which is crucial for planning production, inventory, and marketing strategies.

**Evaluating Performance**: Analyzing trends allows businesses to evaluate the effectiveness of sales strategies and identify factors that may be influencing sales performance.

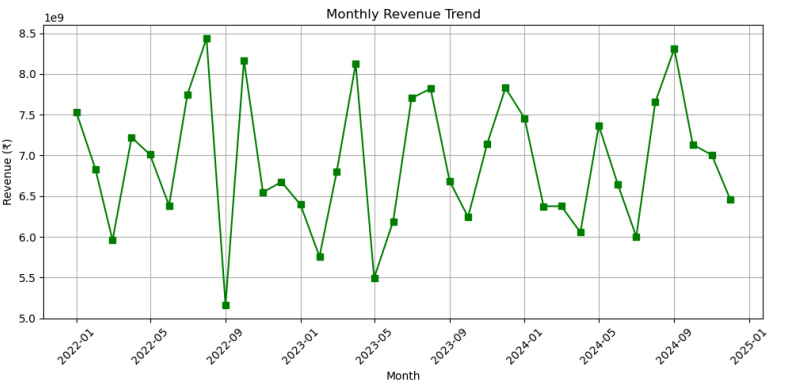
**Supporting Decision-Making**: The insights gained from trend and time series analysis support informed decision-making in areas like resource allocation, budgeting, and goal setting.

In the context of the Maruti Suzuki sales data, trend analysis reveals seasonal patterns and monthly fluctuations in revenue, which can be valuable for the company's operational and strategic planning.

Trend and time series analysis is used to identify patterns and trends in the sales data over time. Monthly revenue trends are plotted to visualize fluctuations and seasonal patterns. A moving average is calculated to smooth out short-term variations and highlight longer-term trends. Time series forecasting using the ARIMA model is employed to predict future revenue.

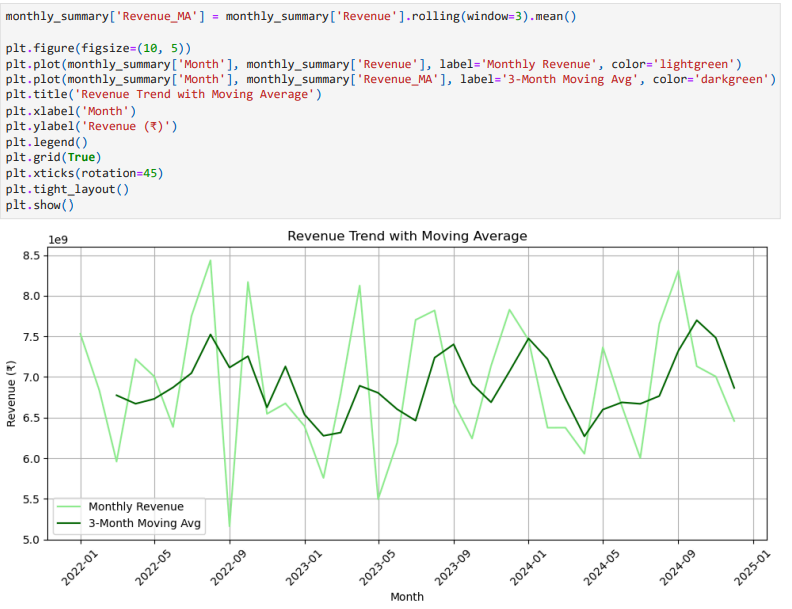


**OUTPUT**



A moving average is a statistical calculation used to analyze data points by creating a series of averages of different subsets of the full dataset. In finance, moving averages are often used to analyze stock prices, helping to smooth out price data by creating an average price that's constantly updated.

In the context of the Maruti Suzuki sales data: I use a 3-month moving average. This helps us to see the broader direction of revenue, beyond just month-to-month changes. For example, even if revenue dips in one particular month, the moving average might still be trending upwards, indicating that the overall revenue trend is still positive.



### **Time Series Forecasting with SARIMA (Revenue)**

SARIMA stands for Seasonal AutoRegressive Integrated Moving Average. It's an extension of the ARIMA model that explicitly supports time series data with a seasonal component. Many time series, like sales data, exhibit seasonality, meaning they have regular, predictable patterns that repeat over specific periods (e.g., higher sales in certain months). SARIMA models are designed to capture both the trend and seasonal aspects of the data, leading to more accurate forecasts for such time series.

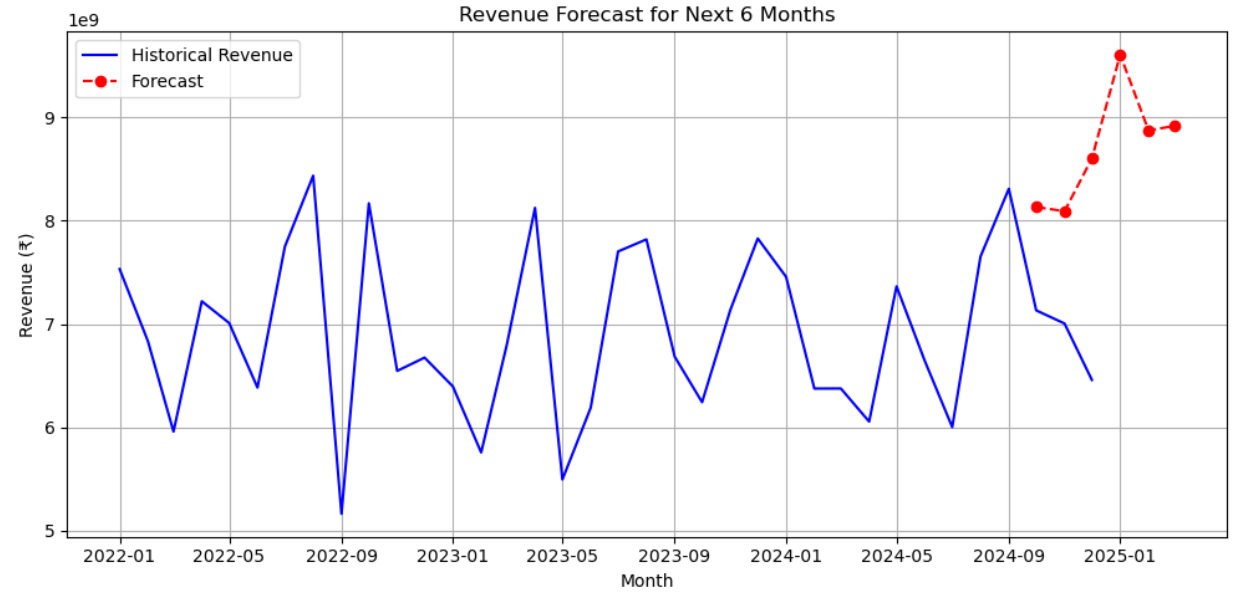
**Key components of a SARIMA model:**

**ARIMA Components (p, d, q)**: These are the same as in the ARIMA model: AutoRegressive (AR): Uses past values of the series to predict future values (p = number of lag observations). Integrated (I): Differences the series to make it stationary (d = number of times the series is differenced). Moving Average (MA): Uses past forecast errors to predict future values (q = number of lagged forecast errors).

**Seasonal Components (P, D, Q, s)**: These are the additional components that account for seasonality: Seasonal AR (P): Uses past seasonal values of the series to predict future values. Seasonal I (D): Differences the series at the seasonal level to make it stationary. Seasonal MA (Q): Uses past seasonal forecast errors to predict future values. Seasonal Period (s): The number of time points in each season (e.g., 12 for monthly data with yearly seasonality)



**OUTPUT**

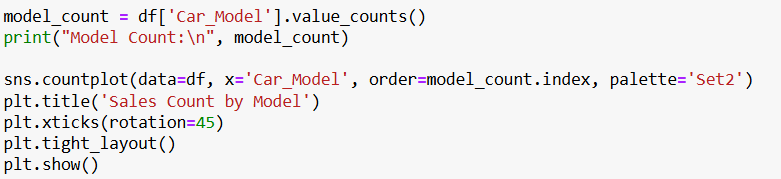


A SARIMA model was applied to the monthly revenue data to forecast future sales trends. The model accounted for both trend and seasonality. The 6-month forecast provides a predictive view of revenue, which can assist in strategic decisions such as production planning, budgeting, and inventory control.

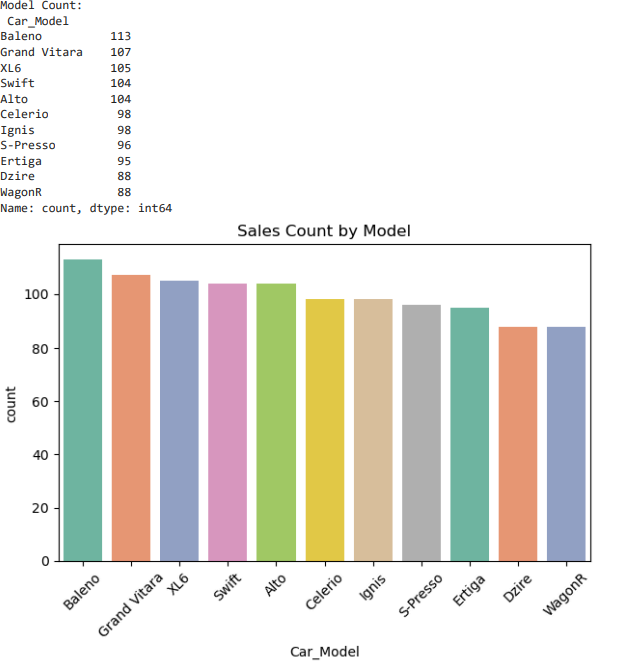
**Count & Categorical Analysis**

Categorical data analysis involves examining the distribution of sales across different car models. The count of cars sold by model is calculated, and the mean revenue by model is determined to compare sales performance.

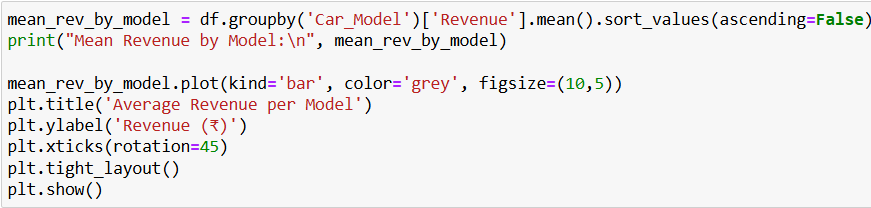
### **Count of Cars Sold by Model**



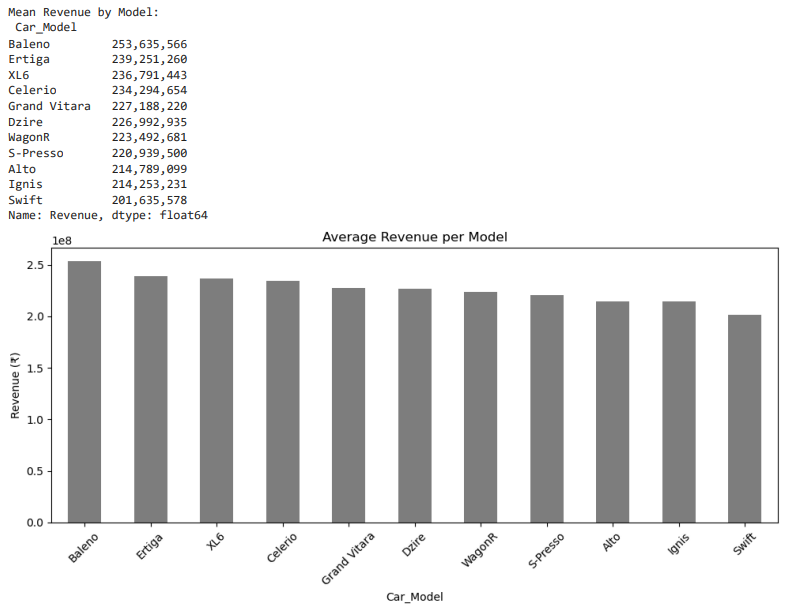
**OUTPUT**



### **Combine with Revenue – Mean Revenue by Model**



**OUTPUT**



The categorical variable analysis reveals differences in sales performance across models.The model-wise average revenue comparison further informs strategic decisions around pricing and promotions

# **Conclusion**

This integrated assignment successfully applied a combination of financial analysis, statistical inference, sampling techniques, count & categorical variable analysis, trend and time series analysis, and machine learning to the Maruti Suzuki sales dataset.

**Key findings include:**

Trend analysis showed clear seasonal patterns and monthly fluctuations in revenue.

Descriptive analysis highlighted top-selling car models and variations in revenue across models.

Sampling methods, particularly stratified sampling, were used to ensure a representative subset of data for analysis and modeling, which also tied directly to inference through confidence interval estimation.

Machine learning models were developed to predict key business outcomes like Units Sold and Revenue, with regression models showing strong predictive performance (e.g., RMSE ≈ 291 on Units Sold prediction).

Classification models successfully categorized high vs low sales months, supporting strategic decision-making.

Count and categorical variable analysis revealed patterns in sales distribution across models and months.

This assignment not only demonstrates technical application of data science techniques but also shows how they interlink — using sampling and inference to support machine learning, and using trend analysis to inform forecasting — creating a cohesive and insight-driven business analysis.