Data Science Specialization - 4 Months

SALES ANALYSIS FOR ABG MOTORS TO ENTER THE INDIAN MARKET

Internship Project Report

Submitted To

Mayur Dev Sewak
Head, Internships & Trainings
Eisystems Services

Mallika Srivastava
Head, Training Delivery
Eisystems Services

Submitted By

Somanjan Chakraborty

Kalyani Government Engineering College

Registration No.:- 005471 of 2019-20

Roll No.:- 10200319056

Table of Contents

1. Project Summary

This internship project focuses on analyzing historical vehicle sales data to evaluate the feasibility of ABG Motors entering the Indian automobile market. Using Python for data analysis and Tableau for visualization, sales trends from Japan were studied and used as a benchmark to predict potential demand in India. The project identifies key customer segments, sales patterns across age groups, and provides data-driven recommendations for market entry.

---

2. Introduction

ABG Motors aims to expand its operations into new international markets. Before entering the Indian market, it is essential to understand customer behavior, expected demand, and target demographics. This project applies data analytics techniques to support strategic decision-making for market expansion.

---

3. Business Problem Statement

To determine whether ABG Motors should enter the Indian automobile market by estimating potential sales and identifying target customer segments using historical sales data from Japan.

---

4. Dataset Description

The dataset contains vehicle sales-related information including customer age, gender, car age, and country-level sales figures. Japan sales data is used as historical input, while predictive modeling estimates sales for India.

```python
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

```python
Indian_data = pd.read_excel('IN_Data.xlsx')
japanese_data = pd.read_excel('JPN_Data.xlsx')
```

```python
Indian_data.head()
```

	ID	CURR_AGE	GENDER	ANN_INCOME	DT_MAINT
0	20710B05XL	54	M	1425390	2018-04-20
1	89602T51HX	47	M	1678954	2018-06-08
2	70190Z52IP	60	M	931624	2017-07-31
3	25623V15MU	55	F	1106320	2017-07-31
4	36230I68CE	32	F	748465	2019-01-27

```python
japanese_data.head()
```

	ID	CURR_AGE	GENDER	ANN_INCOME	AGE_CAR	PURCHASE
0	00001Q15YJ	50	M	445344.000000	439	0
1	00003I71CQ	35	M	107634.000000	283	0
2	00003N47FS	59	F	502786.666667	390	1
3	00005H41DE	43	M	585664.000000	475	0
4	00007E17UM	39	F	705722.666667	497	1

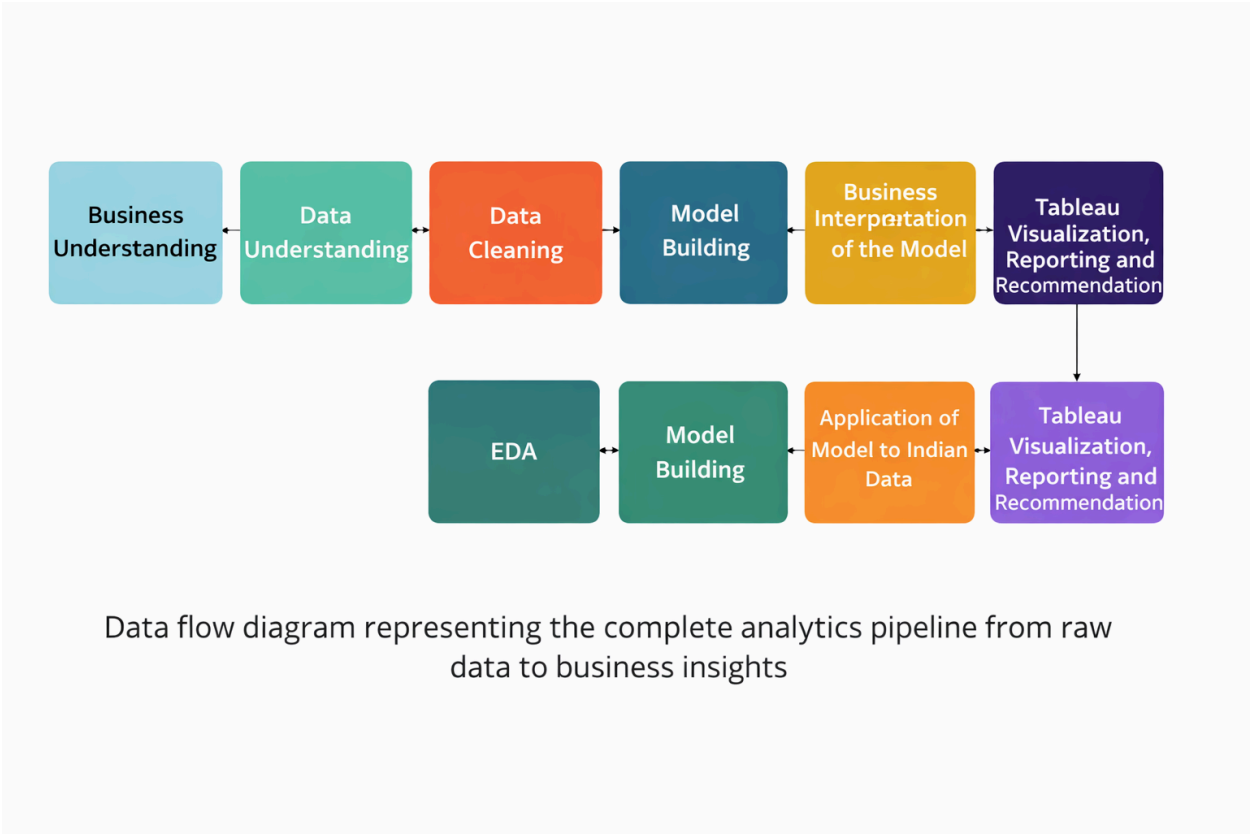
---

5. Methodology / Project Process

This project follows a structured, end-to-end data analytics and machine learning workflow to evaluate ABG Motors' potential entry into the Indian automobile market. The complete process is explained step by step below.

Data flow diagram representing the complete analytics pipeline from raw data to business insights:

> Business Understanding →Data Understanding → Data Cleaning → EDA → Feature Engineering → Model Building →Model Validation and Performance Evaluation → Business Interpretation of the Model → Application of Model to Indian Data →Tableau Visualization → Reporting and Recommendation



Data flow diagram representing the complete analytics pipeline from raw data to business insights

Step 5.1: Business Understanding

The primary objective of this project is to support ABG Motors' strategic decision-making regarding entry into the Indian market. Japan's historical car sales data is used as a benchmark market to understand customer behavior and purchasing patterns. The business target is to assess whether the predicted sales volume in India can exceed a benchmark threshold and justify market entry.

Step 5.2: Data Understanding and Data Cleaning

The datasets for Japan and India were loaded into Python using the pandas library. Initial checks were performed to understand column names, data types, and feature meanings. Missing values, duplicate records, and inconsistencies were identified and handled appropriately. Summary statistics such as mean, median, minimum, and maximum values were analyzed to understand overall data distribution.

df.info() method in the pandas library prints a concise summary of a DataFrame.

df.isnull().sum() counts the number of missing (NaN) values in each column of the DataFrame **df** and returns the result as a pandas Series.

Check for missing values

```
Indian_data.isnull().sum()
```

```
ID            0
CURR_AGE      0
GENDER        0
ANN_INCOME    0
DT_MAINT      0
dtype: int64
```

```
japanese_data.isnull().sum()
```

```
ID            0
CURR_AGE      0
GENDER        0
ANN_INCOME    0
AGE_CAR       0
PURCHASE      0
dtype: int64
```

Step 5.3: Exploratory Data Analysis (EDA)

EDA was conducted to identify patterns and relationships between customer attributes and vehicle purchases. Visualizations were created to study trends across age groups, gender, and other key variables. Correlation analysis helped identify which features had a stronger influence on sales. Insights from Japan were compared with patterns observed in Indian data to validate assumptions.

Step 5.4: Feature Engineering

Relevant features were prepared for analysis and modeling. Categorical variables were encoded into numerical form, numerical features were scaled where required, and age and car-age variables were segmented into meaningful bins. The target variable was clearly defined to represent the likelihood of purchasing a vehicle.

Age binning (VERY IMPORTANT)

```
bins = [20, 30, 40, 50, 60, 70]

labels = ['25', '30', '35', '40', '45', '50']
```

```
df['Curr Age (bin)'] = pd.cut(df['Age'], bins=bins,
labels=labels)
```

Car age segmentation

```
df['AGE_CAR_SEGMENT'] = pd.cut(
    df['Car_Age'],
    bins=[0, 3, 6, 9, 12],
    labels=[1, 2, 3, 4]
)
```

Step 5.5: Model Building

A classification-based modeling approach was adopted, as the business problem involves predicting whether a customer is likely to purchase a car or not. Models such as Logistic Regression were considered due to their interpretability and suitability for binary outcomes. The Japanese dataset was split into training and testing sets to build and evaluate the model.

Step 5.6: Model Validation and Performance Evaluation

Model performance was evaluated using standard metrics such as accuracy, precision, recall, F1-score, and confusion matrix. These metrics ensured that the model was reliable and suitable for business application. The evaluation results confirmed that the model captured meaningful patterns in customer purchase behavior.

Step 5.7: Business Interpretation of the Model

Model outputs were interpreted in business terms. Key factors influencing purchase decisions were identified and translated into actionable insights for ABG Motors. This step ensured that technical results were understandable and useful for non-technical stakeholders.

Step 5.8: Application of Model to Indian Data

The trained model was applied to the Indian dataset after performing the same preprocessing steps. Predictions were generated to estimate the number of potential buyers in India. The predicted sales volume was then compared against the business benchmark to assess market feasibility.

Step 5.9: Tableau Visualization

The final results were visualized using Tableau dashboards. Visualizations included predicted sales in India, historical sales in Japan, age-wise sales trends, and comparative insights between the two markets. These dashboards enable interactive exploration and executive-level understanding of the results.

Step 5.10: Reporting and Recommendation

All findings, insights, and visualizations were compiled into this internship project report. Based on the analysis and predictions, clear recommendations were provided regarding ABG Motors' entry into the Indian market.

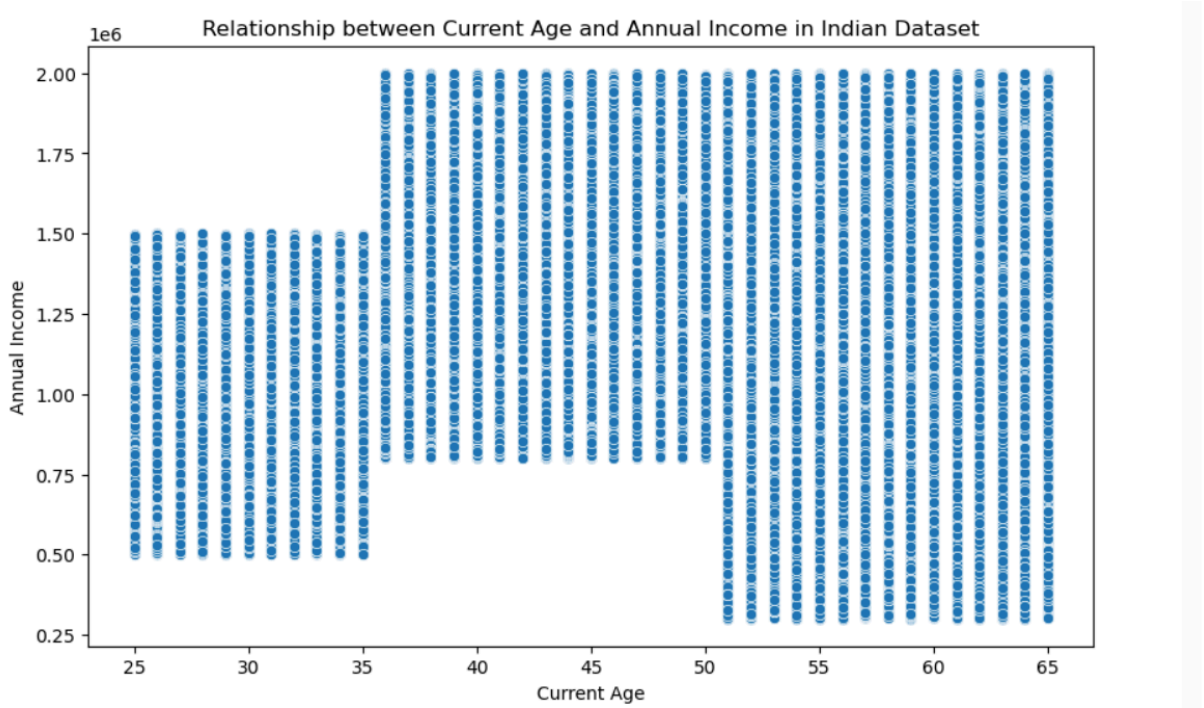
---

6. Exploratory Data Analysis (EDA)

EDA was performed to identify sales distribution across age groups and genders. The analysis revealed that customers aged between 30 and 50 contribute the highest sales volume.

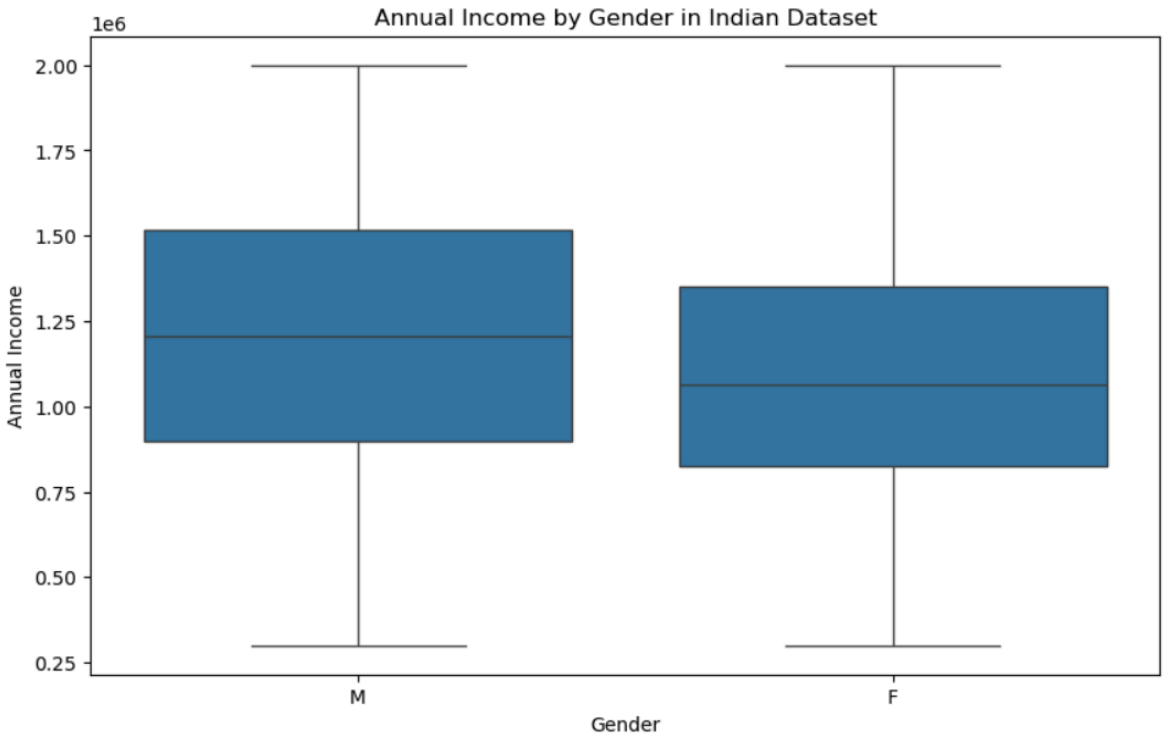
Age vs Sales (Line / Bar)

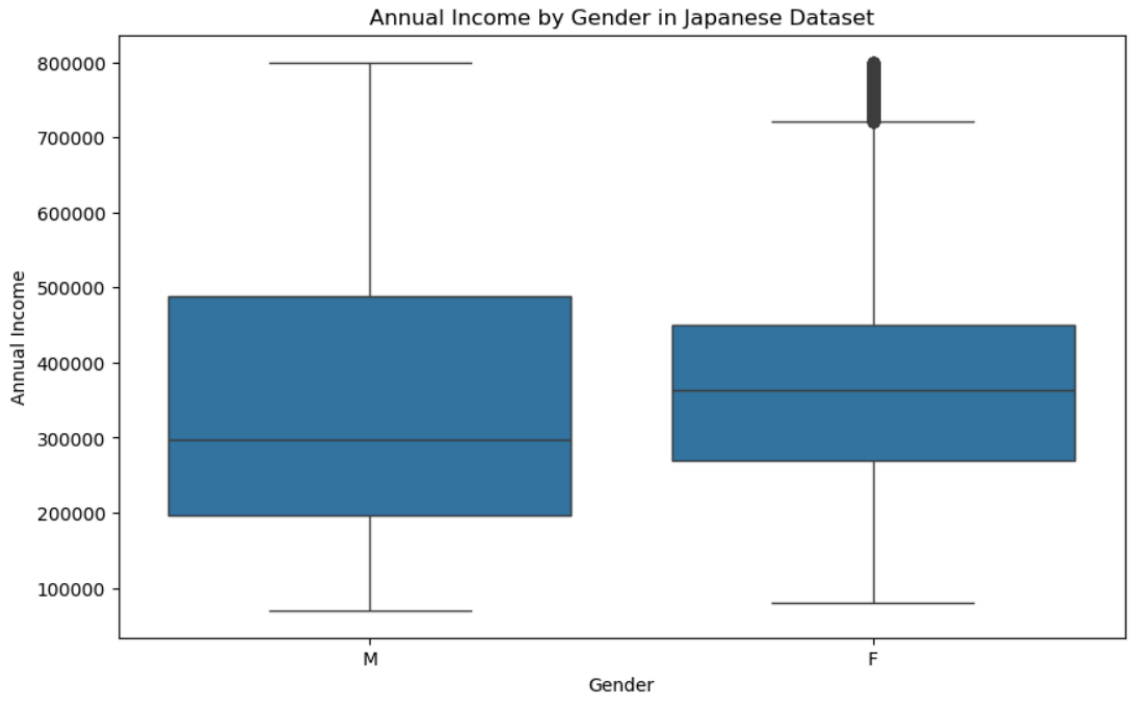
```
age_sales = df.groupby('Curr Age (bin)')['Sales'].sum()

age_sales.plot(kind='line', marker='o')
```



Gender-wise sales

```
gender_sales = df.groupby('Gender')['Sales'].sum()
gender_sales.plot(kind='bar')
```





7. Predictive Analysis

Using observed trends from Japan, sales predictions were generated for the Indian market. The total predicted sales for India were approximately 29,735 units, indicating strong market potential.

Logistic Regression

In [40]:
```
imputer = SimpleImputer(strategy='mean')
X_train = imputer.fit_transform(X_train)
X_test = imputer.transform(X_test)
```

In [41]:
```
log_reg = LogisticRegression()
log_reg.fit(X_train, y_train)
y_pred_log_reg = log_reg.predict(X_test)
print("Logistic Regression Accuracy:", accuracy_score(y_test, y_pred_log_reg))
print("Classification Report:\n", classification_report(y_test, y_pred_log_reg))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred_log_reg))
```

```
Logistic Regression Accuracy: 0.685625
Classification Report:
              precision    recall  f1-score   support

           0       0.65      0.55      0.60      3349
           1       0.71      0.78      0.74      4651

    accuracy                           0.69      8000
   macro avg       0.68      0.67      0.67      8000
weighted avg       0.68      0.69      0.68      8000

Confusion Matrix:
 [[1852 1497]
 [1018 3633]]
```

Decision Tree

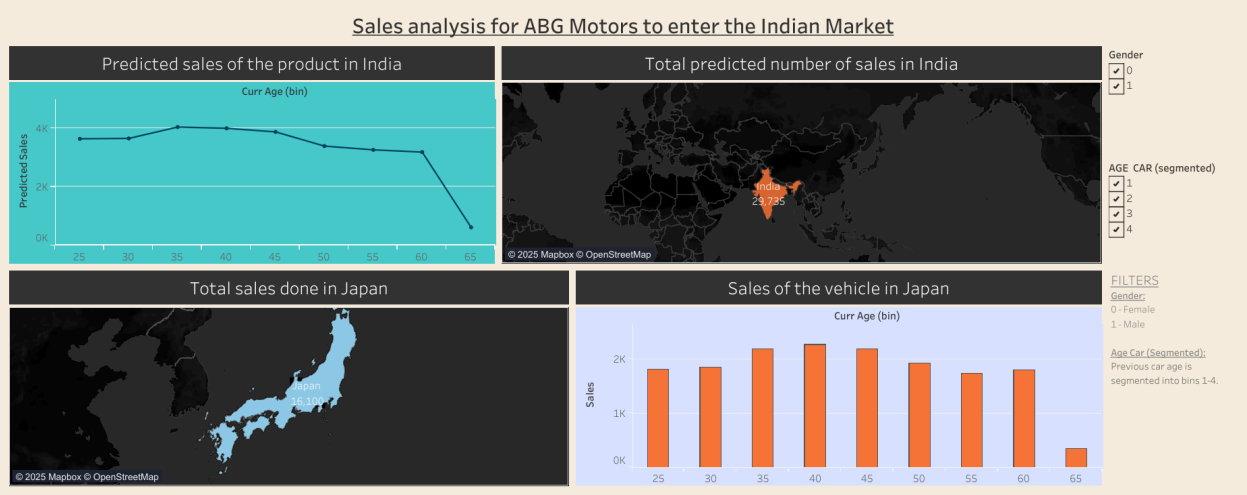
In [42]:
```
dt = DecisionTreeClassifier()
dt.fit(X_train, y_train)
y_pred_dt = dt.predict(X_test)
print("Decision Tree Accuracy:", accuracy_score(y_test, y_pred_dt))
print("Classification Report:\n", classification_report(y_test, y_pred_dt))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred_dt))
```

```
Decision Tree Accuracy: 0.62875
Classification Report:
              precision    recall  f1-score   support

           0       0.56      0.56      0.56      3349
           1       0.68      0.67      0.68      4651

    accuracy                           0.63      8000
   macro avg       0.62      0.62      0.62      8000
weighted avg       0.63      0.63      0.63      8000

Confusion Matrix:
 [[1892 1457]
 [1513 3138]]
```

8. Dashboard & Visualization

An interactive Tableau dashboard was created to visualize: - Predicted sales in India by age group - Total predicted sales in India (map view) - Actual sales in Japan - Age-wise

sales distribution in Japan



---

9. Key Insights & Business Recommendations

- Highest demand observed in the 30–50 age group
- Predicted Indian sales exceed Japan's historical sales
- India presents a high-growth opportunity for ABG Motors
- Focus marketing and product positioning on mid-age professionals

---

10. Conclusion

Based on data analysis and predictive insights, entering the Indian market appears to be a viable and profitable opportunity for ABG Motors. Strategic focus on the identified customer segments can maximize success.

---

11. References

- Eisystems Services Training Materials
- Dataset downloaded from Internet
- Tableau Public Documentation
- Python Libraries: Pandas, NumPy, Matplotlib, Seaborn

  GitHub Link:
  https://github.com/somanjan056/SALES-ANALYSIS-FOR-ABG-MOTORS-TO-ENTER-THE-INDIAN-MARKET.git