EXPLAINING THE CODE (in simple terms)

1. Data Exploration

We started by loading the dataset into Python using this:

df = pd.read_csv('small_business_ghana.csv')

This gives us a table where each row is a small business, and each column is a type of information (like revenue, expenses, region, etc.).

We then asked Python:

- What columns are in the data?
- What kind of data is in each column (numbers or text)?
- Are there any empty spaces (missing values)?

This helps us understand the shape and health of the data before doing any analysis.

2. Dealing with Missing Values

Some businesses didn't report their revenue or expenses (20 of them). We used the median value to fill in the blanks.

Why median?
Because if one or two businesses made unusually high revenue, the average (mean) would be misleading. Median is the "middle number," so it's more stable.

---

3. Feature Engineering

We created new columns from existing ones:

df['profit'] = df['revenue'] - df['expenses']

df['profit_per_employee'] = df['profit'] / df['employee_count']

Why do this?

- Profit tells us how much money they made after paying their expenses.
- Profit per employee shows how efficiently they're using their workers.

4. Scaling the Data

Some columns (like revenue and advertising) have very big numbers. Others (like satisfaction) are small.

To balance the numbers, we used:

StandardScaler()

This made every column more consistent — so no column dominates just because its numbers are bigger.

5. Encoding Categorical Data

Columns like business_type or owner_education are words. But computers work with numbers, not words.

So we:

- Converted education levels to numbers (0 for no education, up to 3 for tertiary).
- Turned categories like "Retail" or "Manufacturing" into separate columns with 1s and 0s.

This lets machine learning models understand the data.

6. Visualizations

We made some graphs:

- **Revenue vs Profit**: Not all businesses with high revenue made big profits.
- **Customer satisfaction**: Some business types have happier customers.
- **Advertising**: Spending a lot doesn’t always mean you’ll get better profit.

EXPLAINING THE REPORT (Layman's Terms)

Introduction

We studied small businesses in Ghana to understand what makes them succeed or struggle — by looking at their money (revenue, expenses, profit), how many people they employ, and how satisfied their customers are.

What We Did

- Some businesses didn’t report all their data, so we filled in the gaps using smart estimates.
- We created new ideas like "profit" and "profit per worker" to better judge performance.
- We made all the numbers balanced and neat so they could be fairly compared.
- We changed words into numbers so computers can understand them.
- We used charts to look for patterns and answer questions.

What We Found

- High revenue doesn’t always mean success — big expenses can wipe out profit.
- Retail and service businesses often had more satisfied customers.
- Advertising alone doesn’t guarantee success — some spent less but still made good profit.

We cleaned the data, created useful insights, and showed patterns. Now, this data can be used for decisions — like helping businesses grow, or helping the government plan support.

Full Report

Project Report: Analyzing Small Business Performance in Ghana

Name: Patrick Ayeh Ayisi & Team

Date: May 8, 2025

1. Introduction

Small businesses in Ghana play a critical role in economic development and employment generation. This analysis explores operational and financial data from 200 small businesses to understand performance factors, improve insight, and inform data-driven decisions. The study focuses on missing value treatment, feature engineering, normalization, encoding, and visualization to uncover trends and relationships.

2. Data Overview and Exploration

The dataset contains 200 records with 12 features, including numerical and categorical variables like revenue, expenses, business_type, region, and owner_education.

- Missing Values:
  - revenue: 20 missing
  - expenses: 20 missing
- Summary Statistics (before scaling):
  - Average revenue: GHS 25,464.13
  - Average expenses: GHS 21,837.69
  - Average employee count: 52
  - Profit margin range: 5% to 49%
- Initial Visual Insights:
  - Most common business type: Manufacturing
  - Most represented region: Greater Accra
  - Customer satisfaction levels are widely varied

3. Missing Value Treatment

To maintain data integrity:

- Median imputation was used for revenue and expenses, as these features showed skewed distributions with potential outliers.

Median is more robust than mean in such cases, minimizing the influence of extreme values.

4. Feature Engineering

Two new features were engineered:

- Profit = Revenue - Expenses
- Profit per Employee = Profit / Employee Count

These features help quantify both raw financial success and operational efficiency.

5. Scaling and Normalization

Numerical features such as revenue, expenses, profit, advertising, sector_growth, and profit_per_employee were scaled using StandardScaler.

- This technique standardized the data to have a mean of 0 and standard deviation of 1, ensuring features contribute equally to models.

6. Encoding Categorical Variables

- Ordinal Encoding:

o owner_education was encoded using the logical order:

 No Formal Education < Primary < Secondary < Tertiary

- One-Hot Encoding:

o Applied to business_type, region, and credit_access to convert categories into numerical form without introducing hierarchy.

7. Visualizations and Insights

1. Revenue vs Profit

- High revenue does not always lead to high profit.
- Some businesses operate on slim profit margins despite high revenues, likely due to high expenses.

2. Customer Satisfaction by Business Type

- Some sectors, such as Retail and Services, showed relatively higher customer satisfaction.
- Manufacturing, though frequent, had mixed satisfaction levels.

3. Advertising vs Profit Margin

- A weak or inconsistent correlation between advertising spend and profit margin.
- Implies that ad effectiveness varies, or other factors (e.g., sector growth or operational cost) influence profit more directly.

8. Conclusion

This analysis provided meaningful insights into the financial and operational dynamics of small businesses in Ghana. Handling missing values, creating performance-based features, scaling numeric data, and encoding categories enabled a cleaner, more model-ready dataset. Visualizations revealed the need for strategic expense management and suggested that advertising alone may not guarantee profit increases.

These insights can guide small business owners and policymakers in making data-informed decisions to enhance efficiency, profitability, and customer satisfaction.

Files Used:

- small_business_ghana.csv (Dataset)