**Data Science- MDA512**

**Assignment 2 (A business model for efficient operations)**

**lecturer: Dr. Samar Shailendra**

**Group -4**

**Problem Statement and Background**

Retail businesses often rely on data analysis to guide strategy, inventory, and marketing. This report focuses on analysing a retail sales dataset to uncover patterns in customer behaviour, product performance, and seasonal trends. The goal is to gain actionable insights that can inform business decisions. Specifically, we examine how sales vary by product category, customer demographics (age and gender), and time (month), and we explore predictive modelling approaches. The dataset contains individual sales transactions, which include details such as transaction date, product category, customer age and gender, and sales amount. Understanding these factors can help the retailer optimize product offerings and target high-value customers.

**Resources**

The primary resource is the [*retail\_sales\_dataset.csv*](https://www.kaggle.com/datasets/mohammadtalib786/retail-sales-dataset), a synthetic retail sales dataset containing 1,000 transactions from the year 2023. Each record includes fields: Transaction ID, Date, Customer ID, Gender, Age, Product Category (Beauty, Clothing, Electronics), Quantity, Price per Unit, and Total Amount. After initial inspection, no missing values were found, ensuring data completeness [1]. The date range spans January to December 2023. A cleaned version (*retail\_sales\_dataset\_cleaned.csv*) was produced by converting the Date column to a datetime type and extracting the numeric Month and Year as new features for analysis. This preprocessing step facilitates monthly trend analysis. All preprocessing and analysis were performed in Python, using libraries such as pandas, matplotlib, and scikit-learn.

**Business Model**

**Research Questions.** We formulated several key questions to guide the analysis:

1. Which product categories generate the highest total sales?
2. How do sales vary across customer ages and genders?
3. What are the seasonal (monthly) trends in sales, overall and by category?
4. Which customer age groups have the largest average basket size?
5. Can we predict whether a transaction is “high-value” based on customer and product features?
6. How well can we predict sales amounts from features like category, age, gender, and month?

**Challenges.** The analysis faced typical challenges of retail data: mixed data types (numeric, categorical), the need for careful data cleaning, and defining what constitutes a “high value” transaction. We addressed these by converting date strings to datetime, encoding categories numerically for modelling, and choosing a sales threshold (e.g. median Total Amount) to label transactions as high or low value. Another challenge is model evaluation: with limited data (1,000 samples), we used train/test splitting and metrics like RMSE for regression. Class imbalance was modest (roughly balanced high vs low, per the decision tree sample counts), so standard classifiers were applied without special weighting.

**Data Preprocessing.** Using pandas, the raw CSV was loaded and the Date column parsed into datetime. We extracted Month and Year from the date, e.g.:

df['Date'] = pd.to\_datetime(df['Date'])

df['Month'] = df['Date'].dt.month

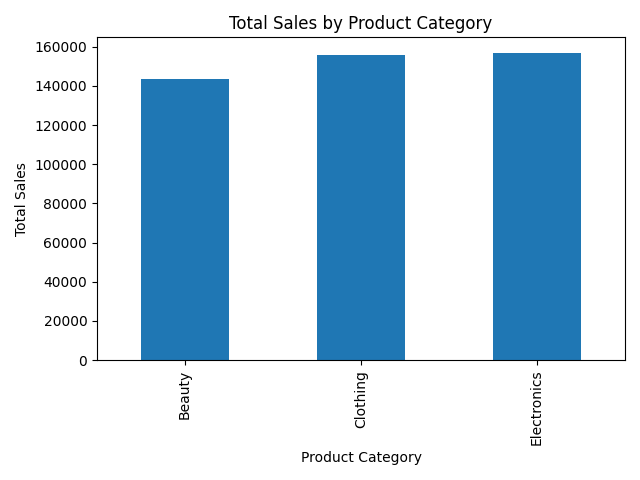
df['Year'] = df['Date'].dt.year

This follows common practice for time features. The additional columns allow grouping by month. No records were dropped as there were no missing entries. Gender and Product Category were encoded (e.g. numerical codes or dummy variables) for modelling.

**Tools and Algorithms.** The analysis used Python with pandas for data manipulation, matplotlib for visualization, and scikit-learn for machine learning. Pandas provided fast, flexible data handling. Matplotlib was used to generate bar charts, line plots, pie charts, and a heatmap to explore trends [2]. For predictive modelling, we applied two algorithms from scikit-learn: a decision tree classifier to predict high-value transactions, and a linear regression model to predict total sales amount. Scikit-learn is a widely used machine learning library featuring regression and classification tools. We split data into training (70%) and testing (30%) sets for model evaluation.

**Data Analysis**

Our analysis proceeded with exploratory data analysis (EDA) to answer the research questions and generate visual insights.



*Figure 1: Product Category Sales*

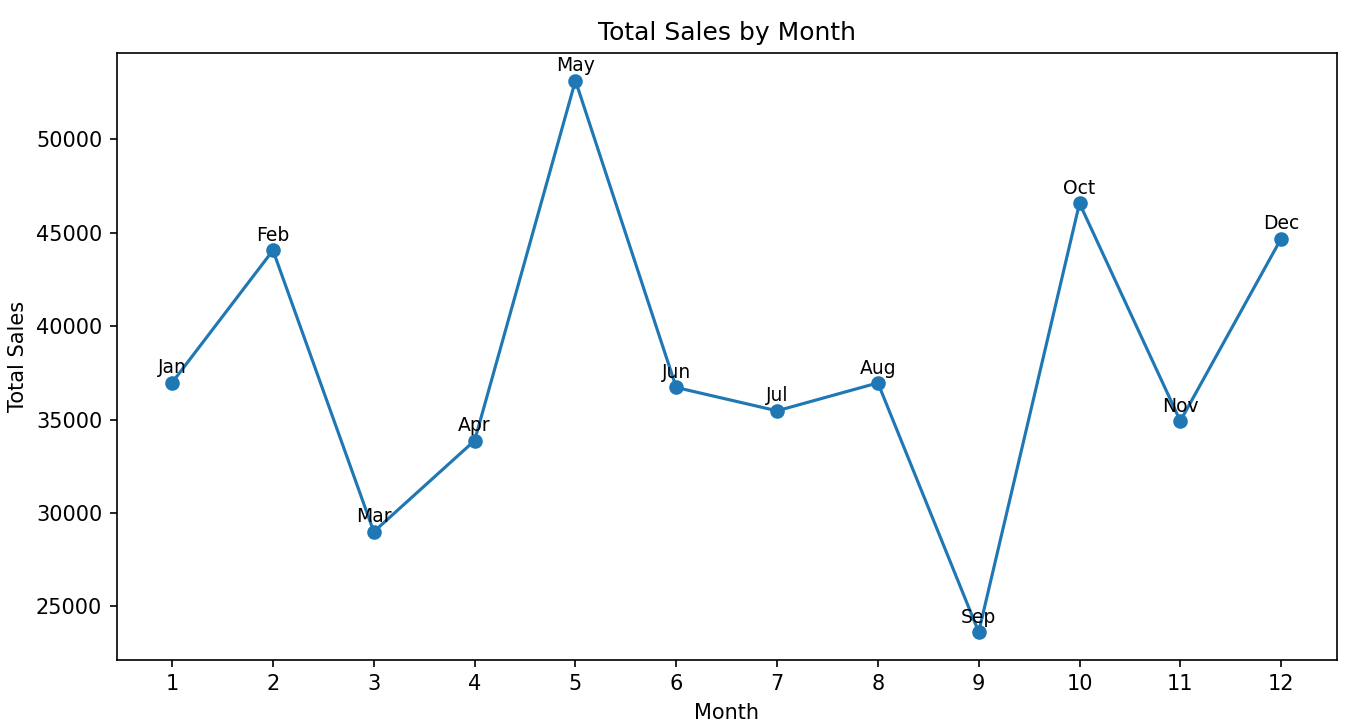
We aggregated total sales by product category. The bar chart (Figure 1) reveals that *Clothing* and *Electronics* generate the highest revenue (around $158k each), while *Beauty* trails slightly (about $143k). This suggests focusing inventory or promotions on Clothing and Electronics could yield greater revenue.

A graph of sales by customer age

AI-generated content may be incorrect.

*Figure 2: Sales by Customer Age*

A bar chart of total sales by customer age (Figure 2) shows which ages contribute most. The highest sales come from customers around age 43, with strong values in the early-40s range. In general, middle-aged groups (30–50 years) dominate sales, indicating these age groups have greater purchasing power or frequency. Younger (18-25) and older (60+) ages contribute less. This insight could guide targeted marketing toward the 30–50 age segment.



*Figure3: Monthly sales Trend*

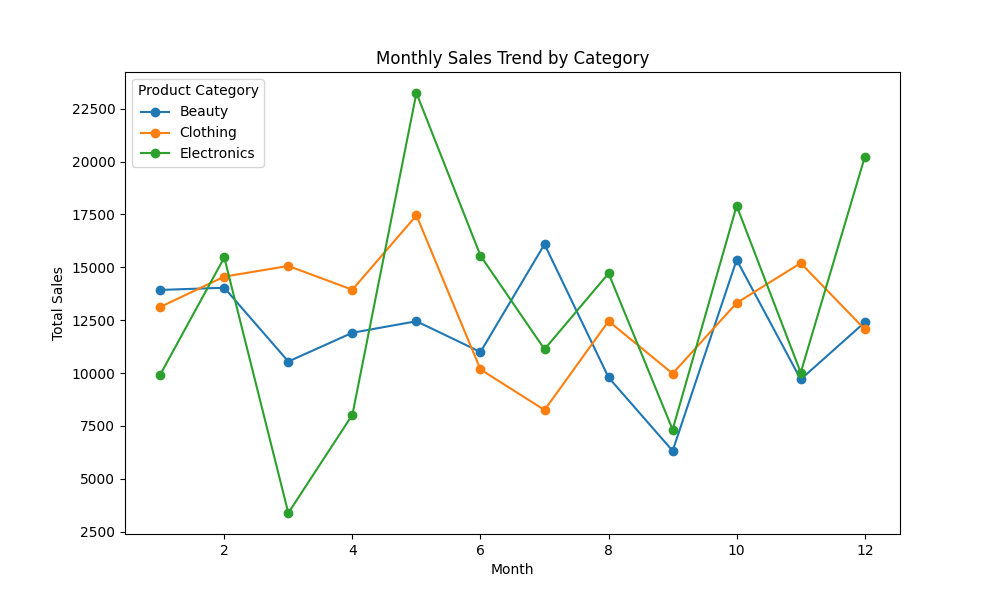
We plotted total sales by month for 2023 (Figure 3). There is clear seasonality: May (month 5) has the peak (~$53k), possibly due to mid-year promotions. There is a sharp dip in September (month 9, $46k and $45k), suggesting holiday shopping spikes. This trend helps the retailer plan for inventory and marketing pushes in peak months.

**A blue and orange pie chart

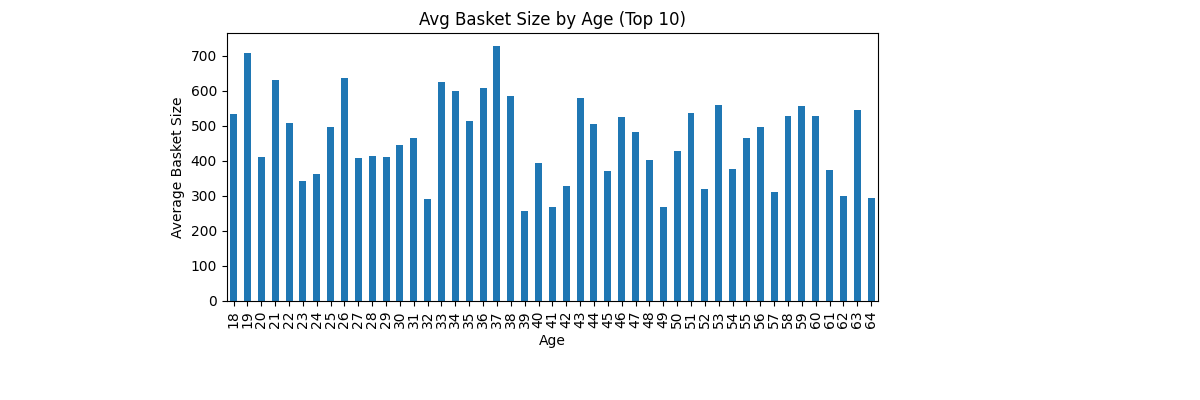
AI-generated content may be incorrect.**

*Figure 4: Sales by Gender*

A pie chart shows sales split roughly evenly between Female (51.1%) and Male (48.9%). This balance indicates no extreme gender bias in overall spending, though a slight female majority exists. It confirms both genders are important customer segments.

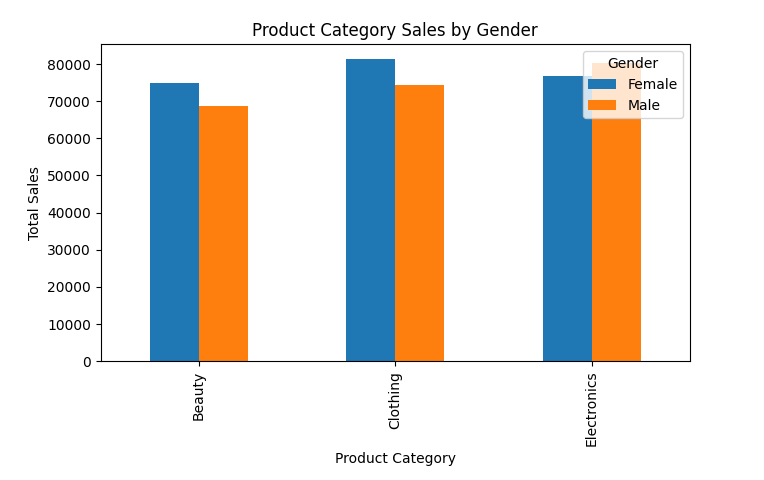
****

*Figure 5: Monthly Trend by Category*

We analysed each category’s monthly sales trend (Figure 5). *Clothing* has relatively consistent high sales with a peak around May, possibly reflecting spring/summer fashion. *Electronics* spikes in May and again strongly in December, aligning with tech releases and holiday season. *Beauty* sales surge in October, perhaps due to autumn promotions or holiday prep. These patterns suggest seasonal marketing: e.g. maximize Electronics promotions in December, Clothing in spring, Beauty around October.

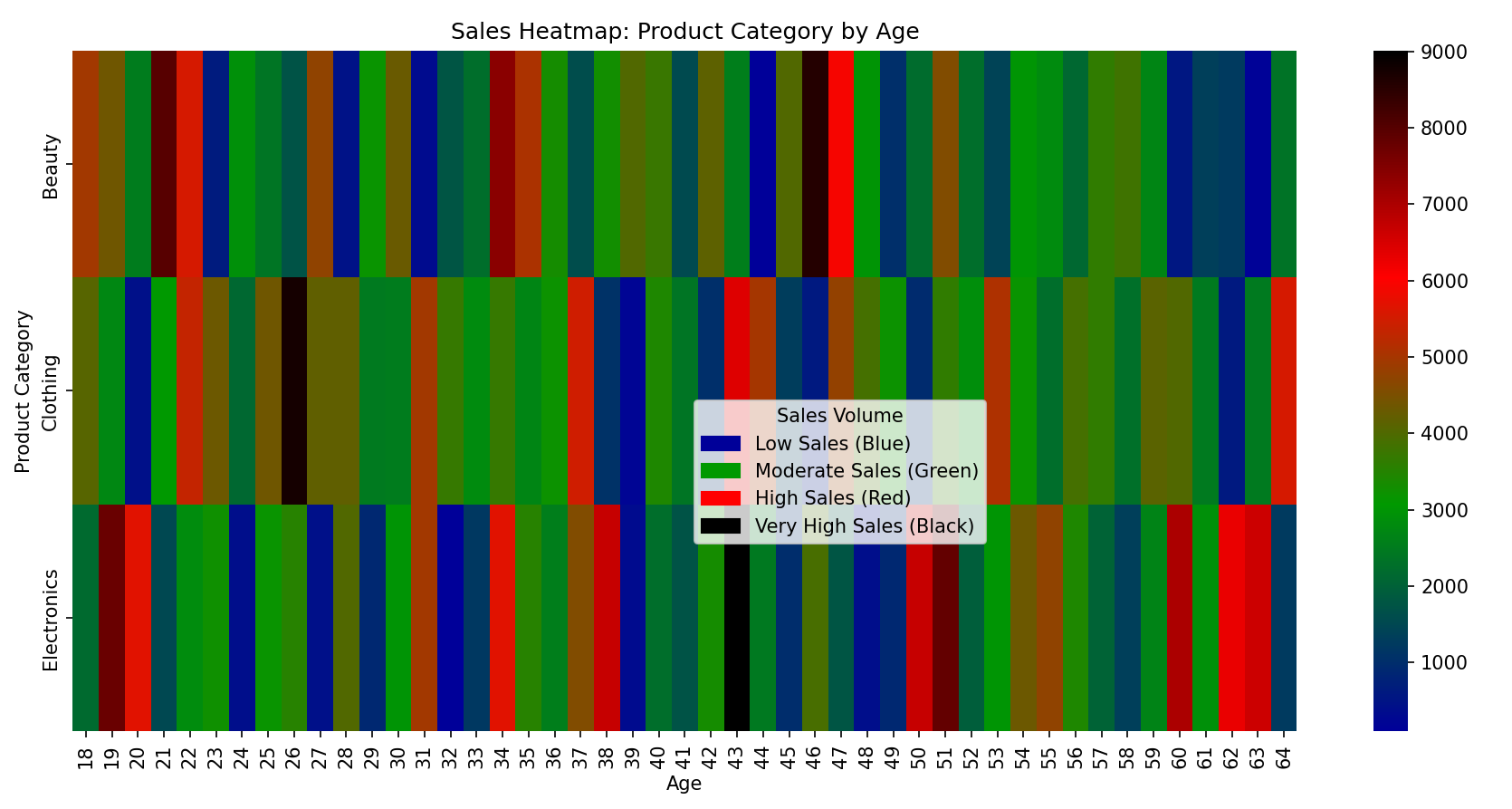
*Figure 6: Average Basket Size by Age*

We computed the average “basket size” (average total amount per transaction) by customer age. Figure 6 (top 10 ages plotted) shows that customers in their mid-30s (around age 35–38) have the largest average basket (over $700). Other ages in late 20s and late 30s also have high baskets (around $500–$600). This implies customers in their 30s not only buy frequently but also spend more per visit. It highlights a key demographic for upselling or loyalty programs.



*Figure 7: Category Sales by Gender*

To break down the gender insight by category, we plotted total sales for each category separated by gender (Figure 7). In all three categories (Beauty, Clothing, Electronics), female customers outspent males. For example, in Clothing females contributed ~$82k vs males $75k; in Beauty about $75k vs $69k. This suggests targeted marketing (e.g. female-focused advertising or products) could further boost sales, especially in categories where female spending already leads.



*Figure 8 : Sales Heatmap: Category by Age*

We created a heatmap of sales for each category by customer age (Fig. 8), Where the blue represents low sales, green indicates moderate sales, red highlights high sales, and black marks very high sales volumes. For instance, high Beauty sales (red/black) appear in the early 20s and mid-30s, high Clothing sales are visible in the late 20s and mid-40s, and Electronics sales concentrate in the mid-30s and late 40s. The heatmap reinforces that different age groups have preferences: younger adults lean toward Beauty and Clothing, while middle-aged groups drive Electronics purchases.

This multidimensional view can guide product-targeted campaigns for specific age segments.

A diagram of a company's flowchart

AI-generated content may be incorrect.

*Figure 9: Predicting High-Value Transactions*

We labelled each transaction as “High” or “Low” value (threshold = median total amount) and trained a decision tree classifier. Figure 9 shows the tree structure. The root splits on *Age ≤ 60.5*. Transactions by older customers (right branch, Age > 60.5) are mostly labelled Low (brown boxes). Younger customers flow left, where further splits use age, product category, and month. For example, on the left branch (Age ≤ 60.5), a split at *Age ≤ 49.5* leads to a large cluster of “High” values when age is above ~42.5 (cyan leaf nodes indicate High class). On the right subtree (Age > 60.5), a split by Product Category shows that low-value purchases dominate for some categories. The tree highlights that customers in mid-age ranges (roughly 25–50) are more likely to make high-value purchases, and that context (month, category) further refines the prediction. This model can help the business identify likely high-value customers (e.g. younger, category purchases) and tailor offers accordingly [3].

**Sales Prediction Using Random Forest Regression**

To further enhance our ability to predict transaction amounts, we applied a Random Forest regression model that incorporates not only basic features (age, gender, product category, and month), but also engineered features like the squared value of age and the interaction between age and product category. The relative importance of each feature in predicting total sales amount is shown in Figure 10.

A graph with blue bars

AI-generated content may be incorrect.

*Figure 1o: Random Forest Feature Importances for Predicting Total Sales Amount*

As shown in Figure 10, the month of purchase stands out as the most influential factor, highlighting the dominant role of seasonality and timing in retail sales. The interaction between age and product category, as well as age itself (including non-linear age effects), also significantly impact sales predictions. In contrast, gender and product category alone have much less influence.

This result suggests that retailers should prioritize their marketing, promotional campaigns, and inventory planning around specific high-performing months and seasons. Additionally, businesses should tailor their product recommendations and marketing campaigns to different age groups and their preferred categories, rather than relying solely on broad category or gender-based strategies. By aligning business operations with the most influential factors revealed by the model, retailers can more effectively target high-value customers and maximize revenue during key periods.

The moderate predictive accuracy of the model also highlights the complexity and variability of retail transactions, suggesting that further gains could be made by collecting more granular data or exploring additional predictive features [4].

**Results**

Our analysis of the retail sales dataset uncovered several clear opportunities for business improvement. Sales are dominated by the Clothing and Electronics categories, highlighting where inventory and promotions should be focused. We found that middle-aged customers, especially those in their 30s to 50s, are the most valuable segment, consistently making larger and more frequent purchases. Females slightly outspent males overall, particularly in Beauty and Clothing, suggesting targeted marketing for these segments could be effective.

Seasonality was a major trend, with sales peaking in May and December and varying by product type throughout the year. This means the business can maximize revenue by aligning promotions and stock levels with these high-demand periods. Our predictive modelling, especially the Random Forest analysis, showed that the timing of purchases and the specific combination of age and product category are the biggest drivers of sales amounts, far more than gender or product category alone.

While the models had moderate predictive accuracy, they reliably highlighted which factors matter most for driving sales. These insights suggest that targeted, data-driven marketing and inventory decisions focusing on key months and age groups will help the business achieve better results.

**Conclusion**

In summary, this analysis has revealed actionable insights that can directly support better business decisions for the retailer. Clothing and Electronics stand out as the top-performing categories, and middle-aged customers particularly those aged 30 to 50 represent the most valuable market segment. Seasonal peaks, especially in May and December, highlight the importance of aligning marketing campaigns and inventory planning with these high-activity periods. Predictive modelling further confirmed that when customers shop and the combination of their age and product preferences are the strongest predictors of sales, while gender and product category alone play lesser roles. By focusing efforts on the most influential factors seasonality and customer segmentation the retailer can optimise sales strategies, better target high-value customers, and ultimately improve overall business performance.

[**GitHub**](https://github.com/hridoyindata/Retail-Data-Analysis)

**References**

[1] M. Talib, “Retail Sales Dataset,” *Kaggle.com*, 2023. https://www.kaggle.com/datasets/mohammadtalib786/retail-sales-dataset (accessed May 18, 2025).

[2] C. to, “comprehensive library for creating static, animated, and interactive visualizations in Python,” *Wikipedia.org*, Oct. 14, 2005. https://en.wikipedia.org/wiki/Matplotlib (accessed May 18, 2025).

[3] GeeksforGeeks, “Decision Tree in Machine Learning,” *GeeksforGeeks*, Dec. 16, 2017. https://www.geeksforgeeks.org/decision-tree-introduction-example/ (accessed May 18, 2025).

‌  
[4] Prasanna Ghattikar, “Using Random Forest For Feature Importance And Feature Selection,” *Medium*, Apr. 19, 2023. https://medium.com/@prasannarghattikar/using-random-forest-for-feature-importance-118462c40189 (accessed May 18, 2025).