Sales Analytics Report

Submitted By

Sr. No.	Team Members Name	Roll No.
1	Abhilasha Nirmal	SCFU124007
2	Ajinkya Konda	SCFU124059
3	Gaurav Dudam	SCFU124009
4	Shubham Marta	SCFU124079

School of Computing / B.Tech CSE

Academic Year: 2024–2025

MIT Vishwaprayag University

Abstract

This project presents an exploratory data analysis (EDA) of a comprehensive transaction dataset to uncover meaningful patterns in product sales, pricing behavior, and market dynamics. Through data cleaning, transformation, and analytical visualization techniques, key insights were derived across multiple dimensions including size distribution, correlation structures, product performance, and geographic variation. Statistical analysis revealed that mid-range pipe sizes dominate sales volume, while extreme sizes move slowly in the market. Correlation metrics confirmed strong linear relationships among amount-based variables and a moderate connection between size and unit rate. Dimensionality reduction using Principal Component Analysis (PCA) demonstrated that the dataset is highly one-dimensional, with the first principal component explaining nearly all variance. PCA projections and K-Means clustering further clarified transaction patterns, showing consistent behavior across locations and product categories, with clear segmentation between bulk orders and regular sales. Overall, the study provides a data-driven understanding of purchasing trends and item performance, offering valuable insights for pricing strategy, inventory management, and operational planning.

Contents

Chapter 1

Introduction

In today's data-driven environment, businesses increasingly rely on analytical insights to understand sales patterns, optimize inventory, and enhance decision-making. This project focuses on performing an in-depth Exploratory Data Analysis (EDA) on a transactional dataset containing detailed information on product sizes, pricing, quantities sold, locations, and associated revenue. The primary objective of the study is to uncover hidden trends, identify relationships among key variables, and interpret market behavior through statistical and visual techniques.

The analysis begins with systematic data cleaning, transformation, and preprocessing, ensuring consistency in categorical fields such as product type, color, and location. Numerical features are examined using distribution plots, correlation matrices, and scatter visuals to understand pricing logic, size behavior, and sales fluctuations. Through dimensionality reduction using Principal Component Analysis (PCA), the study evaluates the underlying structure of the data, revealing a primarily one-dimensional variance pattern. Additionally, clustering techniques like K-Means are applied to segment transactions into meaningful groups based on attributes such as size, quantity, and rate.

The EDA highlights significant business insights, such as the dominance of mid-range pipe sizes in sales volume, the moderate relationship between product size and price, and the consistent purchasing behavior across locations. These findings provide a foundation for better decision-making in areas such as inventory planning, pricing strategy, and product demand forecasting. Overall, the project demonstrates the importance of leveraging analytical tools to convert raw data into actionable knowledge that supports operational and strategic business goals.

Chapter 2

Data Analysis

The analysis of the EDADV dataset provides a detailed understanding of product performance, pricing behavior, customer demand, and geographical transaction patterns. The dataset contains key attributes such as product size, quantity sold, rate per unit, total amount, GST, colour, and transaction location. Through graphical exploration and statistical interpretation, several important trends and patterns were identified.

The line plot of daily total sales revealed strong fluctuations in purchasing activity over time, indicating periodic demand cycles rather than uniform sales. This reflects variations in market requirements, seasonal trends, or bulk procurement at irregular intervals. The dataset also includes bar charts showing total quantities sold per product, which clearly highlight that PVC Pipe and PVC Elbow dominate demand, while smaller specialised components contribute less overall quantity.

Further analysis using box plots for revenue distribution by product type shows that certain products, especially PVC Pipes and Column Pipes, produce significantly higher and more variable revenue, including many high-value outliers. This indicates frequent large-volume orders or pricing differences among product variants.

The pie chart analysis of colour distribution demonstrates an overwhelming preference for Grey products, which account for over 80% of the dataset. Other colours such as White, Black, Milky White, and Silver appear in much smaller proportions, suggesting limited production or targeted use cases.

Scatter plot visualizations, including Size vs Rate and Rate vs Size grouped by Colour, show a positive but non-linear relationship between size and price. Larger pipe sizes tend to command higher rates and exhibit greater price variability. Grey products display the widest spread in pricing, indicating their presence across both low- and high-end product segments.

In addition, PCA (Principal Component Analysis) was used to identify underlying structure within the dataset. The scree plot shows that most of the variance is captured by the first few principal components, confirming that dimensionality reduction is effective for summarizing the data. PCA 2D projections by product type and by location reveal clear clusters, indicating that similar product categories or locations share common purchasing and pricing characteristics.

A heatmap of correlations highlights strong relationships between Basic Amount, GST, and Total Amount, validating the consistency of financial calculations. Moderate correlations among quantity, size, and revenue provide further insight into how product characteristics influence sales outcomes.

Finally, the average transaction value by location shows considerable variation across regions. Some cities, such as Guntur and Mumbai, exhibit significantly higher transaction values, while others operate at smaller scales. This suggests differing market strengths, purchasing capabilities, or demand levels across locations.

Overall, the dataset demonstrates distinct patterns in customer preference, pricing sensitivity, demand distribution, and geographical purchasing behavior. These insights collectively support effective decision-making in product planning, inventory control, pricing strategies, and regional marketing.

Chapter 3

Graphs, Charts, Images and Analysis

3.1 Daily Sales Trend (Line Plot)

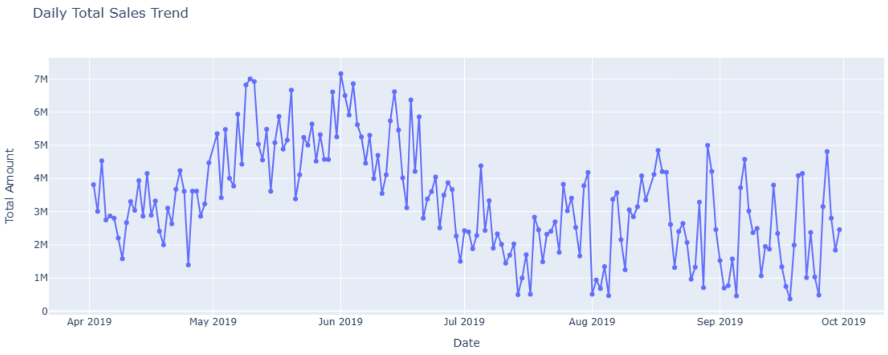

Figure 3.1: Daily Total Sales Trend

Description: The line graph represents the daily total sales trend from April 2019 to October 2019. The sales values fluctuate significantly throughout the period, indicating unstable and varying daily demand. A strong upward movement can be observed between late April and June, where sales frequently reach their peak levels of around 6–7 million, suggesting a high-demand phase. After June, the trend shows a noticeable decline, with July recording some of the lowest sales values, indicating a temporary drop in market activity. From August to September, sales begin to rise again, but the pattern remains irregular, highlighting continued variability in purchasing behaviour. Overall, despite the day-to-day fluctuations, the general level of sales remains relatively consistent across the months, showing no strong long-term upward or downward trend but rather a cyclical pattern of peaks and troughs in customer demand.

3.2 Total Quantity Sold Per Product (Bar Plot)

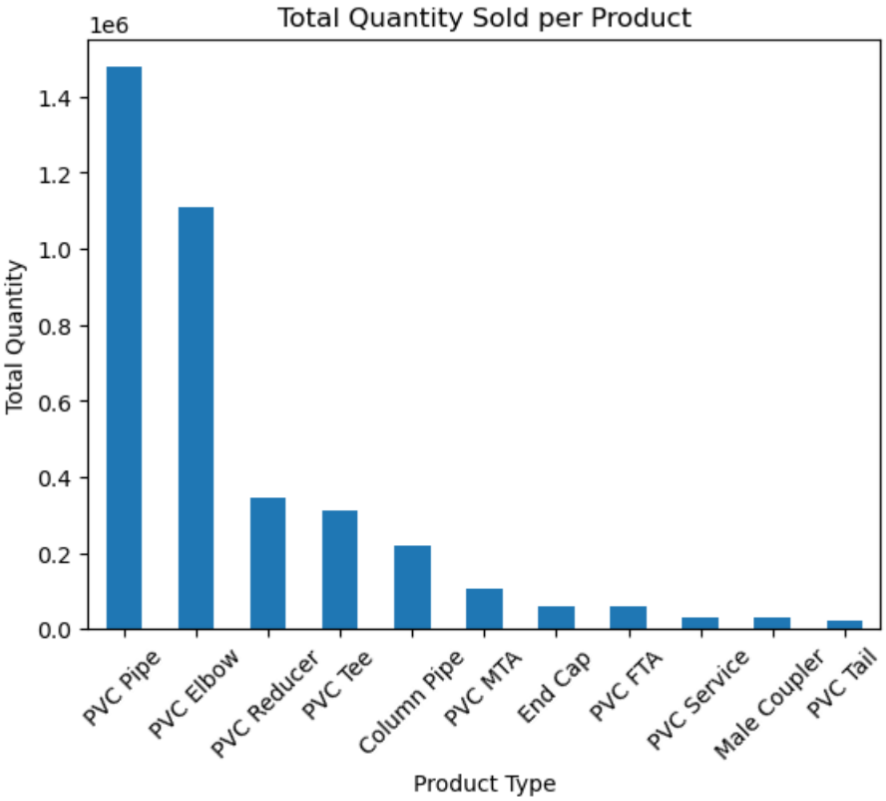

Figure 3.2: Total Quantity Sold by Product Type

Description: The bar chart illustrates the total quantity sold for each product type. Among all items, PVC Pipe is the highest-selling product by a large margin, recording well over 1.4 million units, indicating its dominant demand in the market. This is followed by PVC Elbow, which also shows strong sales at above 1.1 million units, reflecting its frequent usage alongside main pipe installations. Other products such as PVC Reducer, PVC Tee, and Column Pipe show moderate sales volumes, suggesting they are required regularly but in smaller quantities compared to primary pipe components. Items like PVC MTA, End Cap, PVC FTA, PVC Service, Male Coupler, and PVC Tail have relatively lower sales, indicating they are either niche components or required less frequently in typical plumbing or piping projects. Overall, the chart highlights a clear skew in demand, with major pipe products driving the bulk of total sales, while smaller fittings and accessories contribute relatively less.

3.3 Total Revenue Per Product (Log Scale)

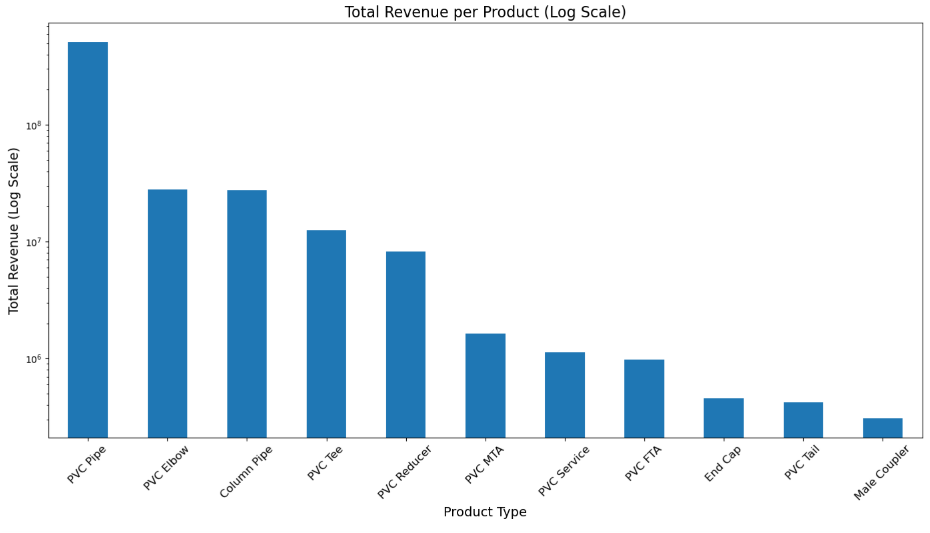

Figure 3.3: Total Revenue per Product (Logarithmic Scale)

Description: The bar chart presents the total revenue generated by each product type on a logarithmic scale, allowing clearer comparison across wide revenue differences. The highest revenue is generated by PVC Pipe, which far exceeds all other product categories, indicating its dominant role in overall earnings. This is followed by PVC Elbow and Column Pipe, both of which contribute significantly to revenue, reflecting their frequent use and relatively higher unit prices. Mid-range contributors include PVC Tee, PVC Reducer, and PVC MTA, which show moderate revenue levels due to their consistent but lower-volume sales. On the lower end, products such as PVC Service, PVC FTA, End Cap, PVC Tail, and Male Coupler generate comparatively smaller revenue, implying they are either low-demand items or low-value components. Overall, the chart highlights a highly skewed revenue distribution, where a few major products account for the majority of earnings, while the remaining items contribute marginally to total revenue.

3.4 Average Transaction Value by Location (Bar Plot)

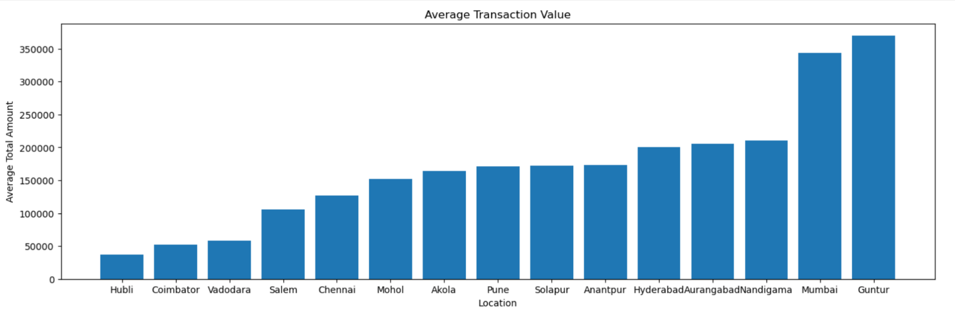

Figure 3.4: Average Transaction Value by Location

Description: The bar chart displays the average transaction value across different locations, revealing significant variation in customer spending patterns. Locations such as Guntur and Mumbai exhibit the highest average transaction amounts, exceeding 350,000, indicating either higher-volume purchases or a stronger market for premium products in these regions. Mid-range cities including Hyderabad, Aurangabad, Nandigama, Anantpur, and Solapur show moderately high average transaction values, ranging between 170,000 and 210,000, suggesting steady commercial activity and balanced consumer demand. Locations such as Chennai, Salem, Akola, and Pune fall in the mid-lower range, reflecting moderate but consistent purchase behavior. On the lower end, cities like Hubli, Coimbatore, and Vadodara show the smallest average transaction values, indicating either smaller purchase volumes or lower-priced product demand in these regions. Overall, the chart highlights clear regional differences in buying patterns, with certain markets contributing substantially more to revenue through higher transaction values.

3.5 Revenue Distribution by Product Type (Box Plot)

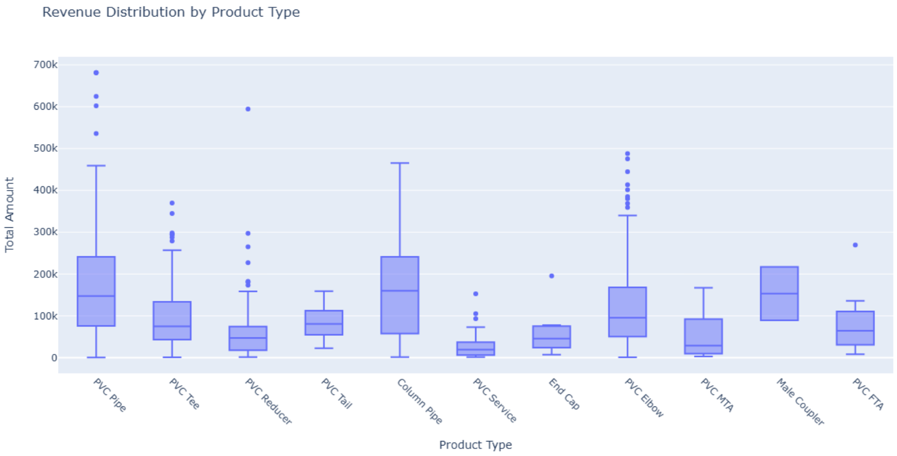

Figure 3.5: Revenue Distribution Across Product Types

Description: The box plot illustrating revenue distribution across different product types highlights substantial variation in earning patterns within the dataset. Products such as PVC Pipe, Column Pipe, PVC Elbow, and Male Coupler show wide interquartile ranges and numerous high-value outliers, suggesting that these items frequently generate large revenue contributions and are subject to high-value transactions. In contrast, items like PVC Reducer, PVC MTA, End Cap, and PVC Service exhibit lower median revenues and tighter distributions, indicating more consistent but comparatively smaller sales amounts. The presence of several extreme outliers—especially for PVC Pipe and Column Pipe—reflects occasional bulk purchases or high-cost orders that significantly boost total revenue. Overall, the plot reveals that a few core product categories dominate revenue generation, while others contribute smaller but more stable amounts, highlighting important differences in product performance and demand intensity.

3.6 Colour Distribution (Pie Chart)

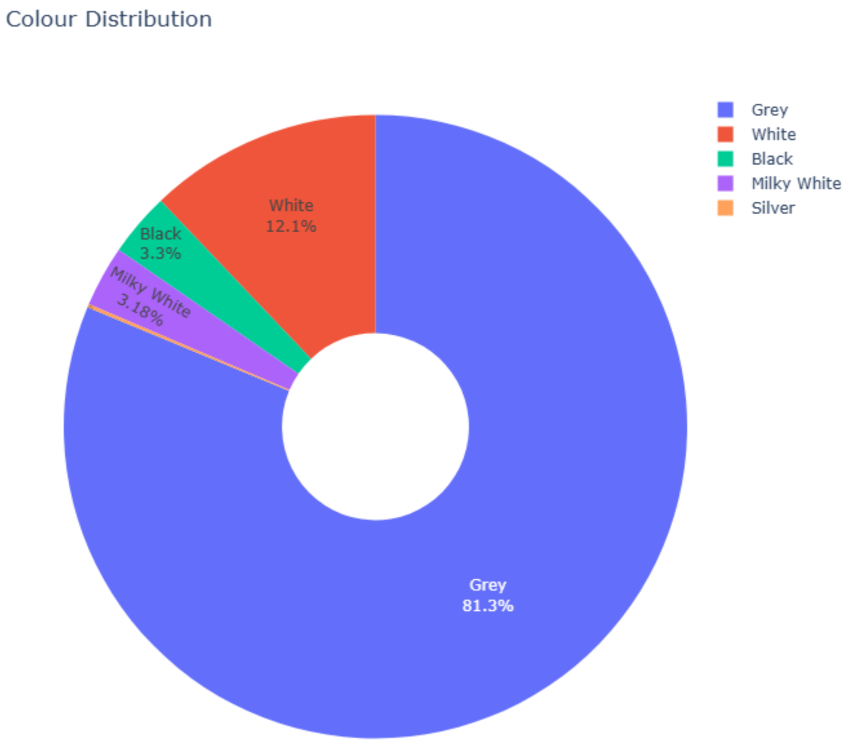

Figure 3.6: Colour Distribution of Products

Description: The pie chart depicting the colour distribution of products shows a highly skewed pattern, with Grey dominating overwhelmingly at 81.3%, indicating that the vast majority of items in the dataset are manufactured or sold in this colour. This suggests that Grey is the standard or most widely demanded colour across product categories. White follows distantly at 12.1%, representing a moderate share of the distribution, likely catering to specific aesthetic or functional requirements. The remaining colours—Black (3.3%), Milky White (3.18%), and Silver (0.1%)—constitute only a small portion of the overall distribution, reflecting limited demand or niche usage. Overall, the chart highlights a strong market preference for Grey products, with other colours contributing minimally in comparison.

3.7 Scatter Plot: Size vs Rate

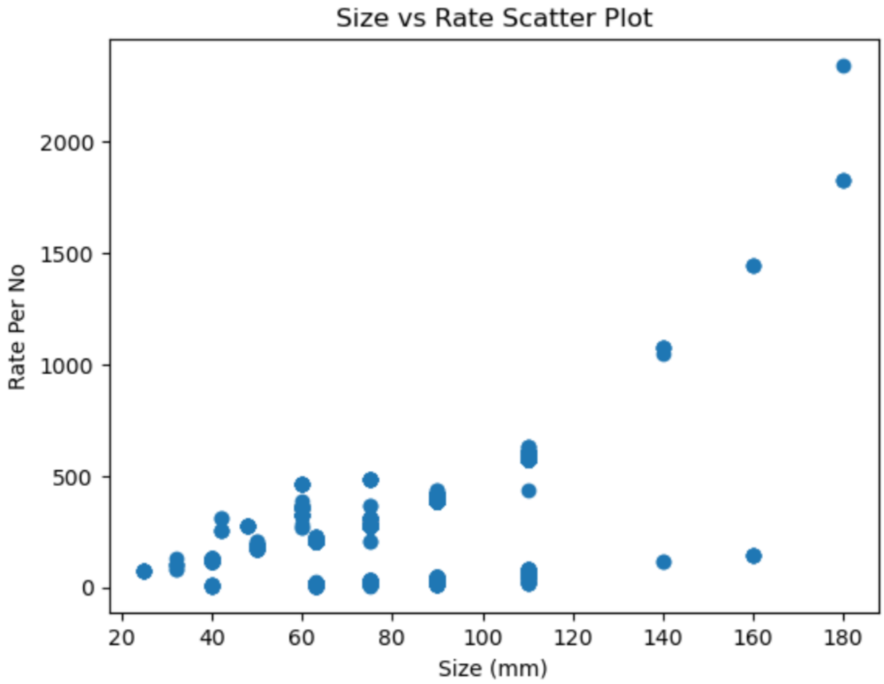

Figure 3.7: Size vs Rate Scatter Plot

Description: The scatter plot illustrates the relationship between pipe size and rate per unit, revealing a generally upward trend as size increases. Smaller pipe sizes are associated with lower and more tightly clustered rate values, indicating consistent pricing within these ranges. As the size increases beyond approximately 100 mm, the rate values become more widely dispersed and significantly higher, suggesting that larger diameters command a premium and exhibit greater price variability. A few noticeable high-value points at larger sizes indicate occasional premium-priced items or specialized products. Overall, the plot highlights a positive but non-linear relationship between size and rate, where large-diameter pipes tend to be priced substantially higher than smaller ones.

3.8 Price Sensitivity: Rate vs Size (Grouped by Colour)

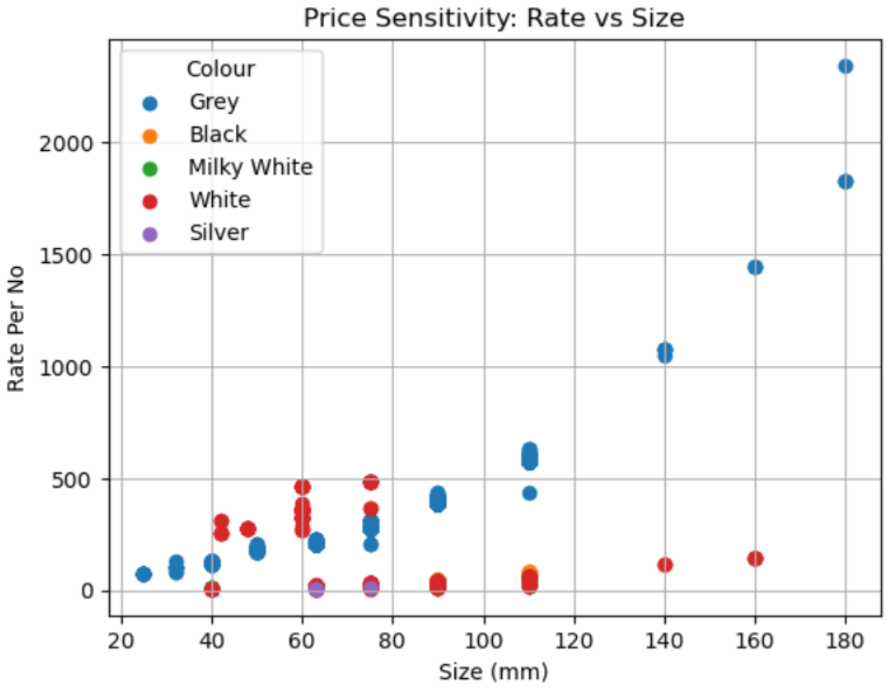

Figure 3.8: Rate vs Size (Grouped by Colour)

Description: The scatter plot showing price sensitivity across different pipe sizes and colours reveals important variations in pricing behavior. While the rate generally increases with size, the distribution differs significantly among colour categories. Grey-coloured products display the widest range of rates, including the highest-priced items, indicating that Grey pipes are not only the most common but also span both low- and high-value segments. White products exhibit moderate price levels with several mid-range clusters, suggesting stable pricing for this colour category. In contrast, Black, Milky White, and Silver products remain clustered at lower rate values, indicating that these colours are associated with smaller or lower-priced items. The spread of points also highlights that price variability increases as size increases, particularly for Grey and White items. Overall, the plot suggests that both size and colour influence pricing, with Grey products showing the highest price sensitivity and variability across larger size ranges.

3.9 3D K-Means Clustering

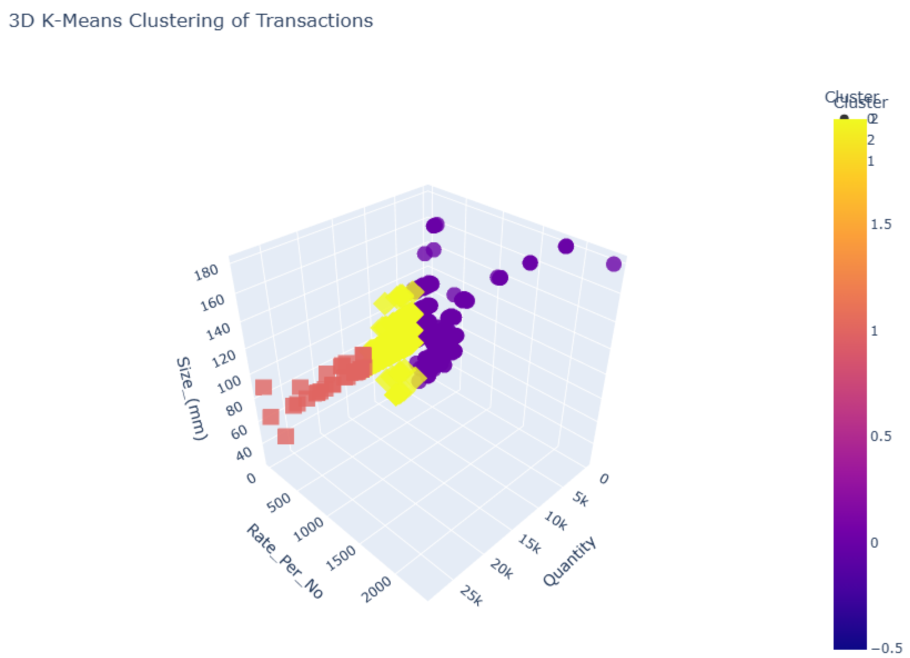

Figure 3.9: 3D K-Means Clustering of Transactions

Description: The 3D scatter plot illustrates the results of a K-Means clustering analysis performed on transaction data using three key variables: Quantity, Rate per Unit, and Size (mm). Each point represents a transaction and is color-coded according to its assigned cluster, revealing distinct behavioral groupings within the dataset. One cluster appears to consist of transactions with low quantities but relatively higher rates and larger sizes, suggesting premium or specialized items. Another cluster contains mid-range quantities with moderate rates and sizes, indicating standard or frequently ordered products. A third cluster shows very high quantities paired with lower rates and mid-sized items, characteristic of bulk or wholesale transactions. The 3D visualization helps highlight separations and overlaps between these groups, making it easier to understand how different transaction types naturally segment based on purchasing patterns.

3.10 PCA Scree Plot

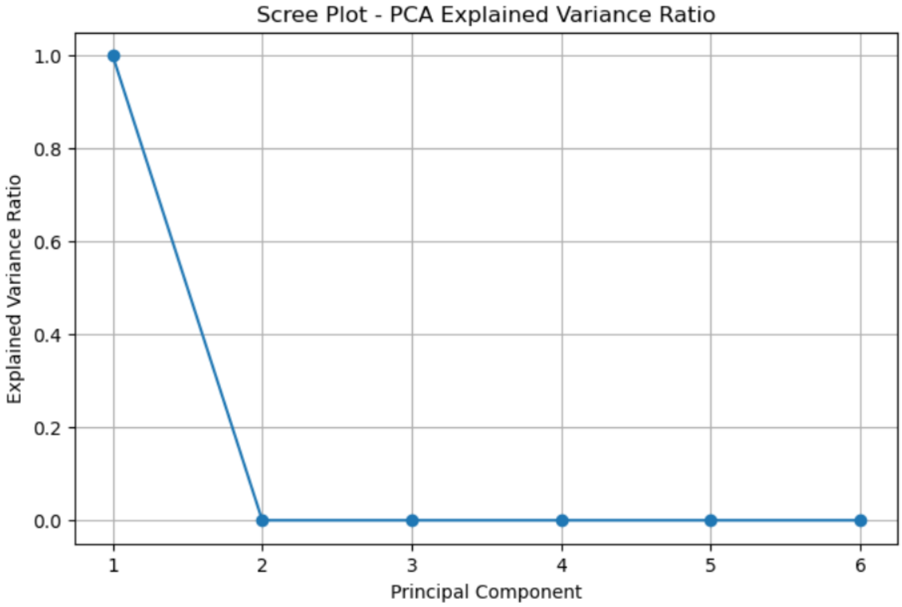

Figure 3.10: PCA Scree Plot

Description: The scree plot displays the explained variance ratio for each principal component in a PCA analysis. The chart shows a dramatic drop after the first component, with the first principal component alone accounting for nearly 100% of the variance in the dataset. All subsequent components contribute virtually no additional explanatory power, as reflected by their near-zero variance ratios. This indicates that the dataset is highly one-dimensional in terms of information content, meaning that almost all meaningful variability can be captured using just a single principal component. In practical terms, dimensionality reduction to one component would preserve nearly all the data's structure, greatly simplifying analysis without significant loss of information.

Conclusion: Most variance is captured in the first few components.

3.11 PCA 2D Projection by Product Type

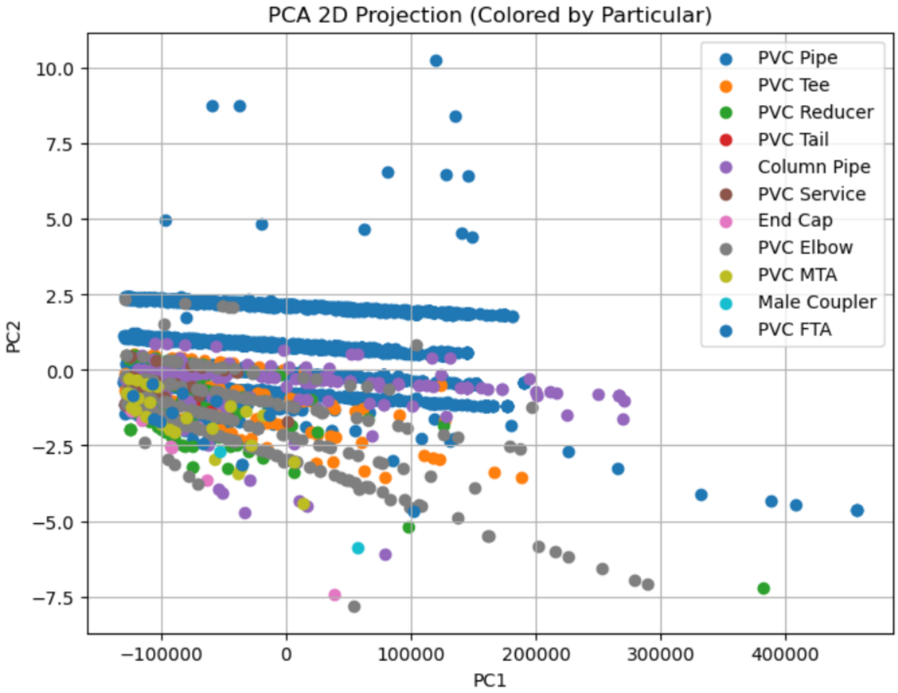

Figure 3.11: PCA 2D Clustering by Product Type

Description: The PCA 2D projection plot visualizes how different product categories distribute across the first two principal components, PC1 and PC2. Each point represents a transaction, colored according to its product type. The wide horizontal spread along PC1 indicates that this component captures most of the variability—consistent with the earlier scree plot showing PC1 explains nearly all variance. Despite all product types largely overlapping due to the dataset's one-dimensional structure, some subtle patterns emerge: PVC Pipe items (blue) form a dense, elongated band, suggesting consistent behavior across transactions. Other categories such as Column Pipe, PVC Elbow, and End Cap show more scatter, indicating greater variability in their attributes. A few outliers positioned far from the main cluster—particularly for PVC Pipe and PVC Elbow—suggest unusual or extreme transaction values. Overall, the plot demonstrates that while different items share similar underlying structure, PC1 primarily differentiates them, with PC2 expressing only minor variation.

3.12 PCA 2D Projection by Location

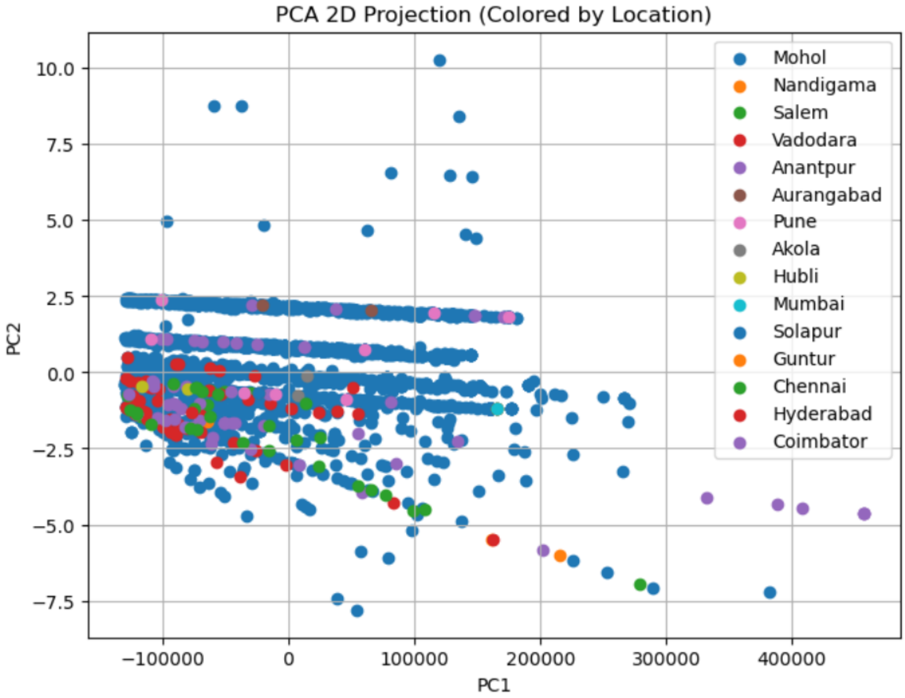

Figure 3.12: PCA 2D Clustering by Location

Description: The PCA 2D projection plotted by location shows how transactions from different geographic centers distribute across the first two principal components, PC1 and PC2. Similar to the earlier PCA visualization, most of the variability lies along PC1, causing all locations to cluster tightly in a long horizontal band. Many cities—such as Solapur, Hyderabad, Pune, and Chennai—exhibit heavy overlap, indicating that their transaction patterns are statistically similar when reduced to principal components. However, a few distinct outliers emerge: locations such as Anantpur and Solapur have points extending far to the right on PC1, suggesting exceptionally high values for one or more original features (e.g., unusually large quantities or rates). Meanwhile, a scattered mix of cities in negative PC1 ranges suggests lower-value or more typical transactions. Overall, the plot shows that geographic location does not strongly separate the data in PCA space, implying that transaction behavior is broadly consistent across regions, with only a few isolated anomalies.

3.13 Pearson Correlation Heatmap

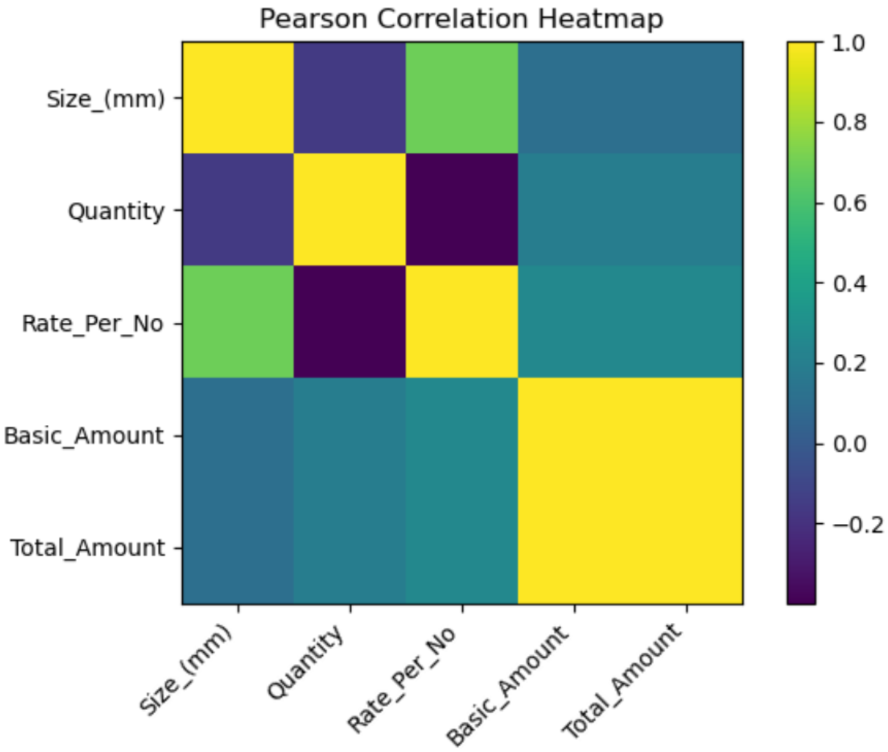

Figure 3.13: Correlation Heatmap of Numerical Variables

Description: The Pearson correlation heatmap highlights the strength and direction of relationships among the key numerical features in the dataset. As expected, Basic Amount and Total Amount show a perfect positive correlation since Total Amount is directly derived from Basic Amount. A strong positive correlation also exists between Rate Per No and Size (mm), suggesting that larger-sized items tend to have higher per-unit rates. In contrast, Quantity shows weak or slightly negative correlations with most variables, indicating that the number of units ordered does not strongly influence item size, rate, or amount within the dataset. The relationships between amount-related variables and Size or Rate are moderate but not dominant, reflecting that variations in order amounts arise from multiple interacting factors. Overall, the heatmap reveals a few clear linear relationships while showing that many variables operate independently of each other.

3.14 Distribution of Size (mm)

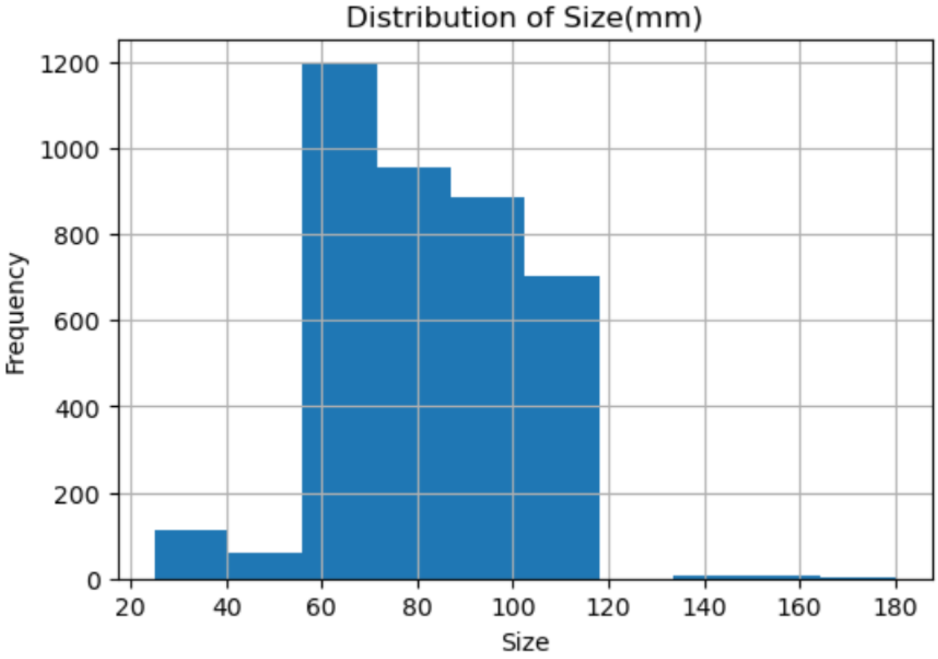

Figure 3.14: Distribution of Pipe Size (mm)

Description: The histogram of Size (mm) shows that the majority of items fall within a relatively narrow size range, with a pronounced concentration between approximately 55 mm and 110 mm. Sizes around 60–80 mm appear most common, indicating that medium-sized items dominate the dataset. A few observations exist below 50 mm and above 120 mm, but these are rare, suggesting they represent specialized or less frequently ordered product types. The distribution is moderately right-skewed, with a long tail extending toward larger sizes, including a small number of extreme values around 140–160 mm, which can be considered outliers. Overall, the distribution implies that the business primarily deals with mid-range product sizes, with only occasional transactions involving unusually small or large dimensions.

3.15 Fast vs Slow Moving Sizes

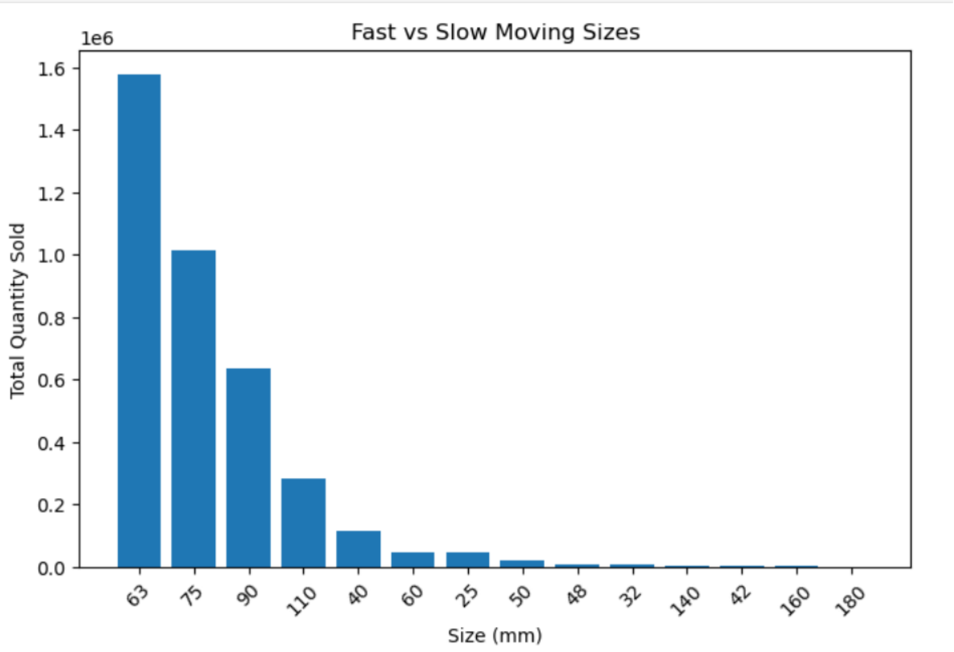

Figure 3.15: Fast vs Slow Moving Sizes

Description: The bar chart illustrates total quantity sold across different product sizes, clearly distinguishing fast-moving sizes from slow-moving ones. The distribution is highly skewed, with a few mid-range sizes dominating sales volume. 63 mm emerges as the fastest-moving size by a wide margin, exceeding 1.5 million units sold, followed by 75 mm and 90 mm, which also show strong demand. As sizes increase beyond 110 mm or decrease below 40 mm, total sales drop sharply, indicating that both very large and very small sizes are slow-moving. Sizes such as 140 mm, 160 mm, and 180 mm show negligible movement, highlighting their limited market demand. Overall, the chart reveals that the majority of transactions concentrate around a core set of popular mid-range sizes, providing insight into inventory planning and demand forecasting.

Chapter 4

Conclusion

The exploratory data analysis reveals several strong patterns within the transaction dataset, offering valuable insights into product behavior, pricing, and sales dynamics. The dataset is dominated by mid-range pipe sizes, with 63 mm, 75 mm, and 90 mm emerging as the fastest-moving items, accounting for the highest sales volumes. Larger and very small sizes show significantly lower movement, identifying them as slow-moving inventory. Correlation analysis highlighted expected relationships, such as the strong link between Basic Amount and Total Amount, as well as a moderately positive relationship between Size and Rate, indicating that larger pipe sizes tend to be priced higher.

Dimensionality reduction using PCA showed that nearly all variance in the dataset is captured by a single component, implying that the core transactional structure is highly linear and influenced mainly by a few dominant numerical variables. The 2D PCA projections revealed substantial overlap between product types and locations, confirming that variations in purchasing behavior are relatively uniform across regions and categories. The K-Means clustering further supported this observation by grouping transactions into distinguishable clusters based on quantity, rate, and size, helping to identify bulk orders, premium items, and standard sales patterns.

Overall, the analysis provides a clear understanding of sales trends, price dynamics, and product performance. These insights can support better inventory planning, pricing strategy refinement, and targeted business decisions. The findings highlight that while the dataset is complex in volume, its underlying structure is predictable, enabling effective data-driven optimization across operations.

Chapter 5

References

[1] Kothari Agritech Pvt. Ltd, “EDADV Dataset,” Provided for academic and analytical use, 2025.

[2] Abhilasha Nirmal, Ajinkya Konda, Gaurav Dudam, Shubham Marta, “EDADV-Analysis-Code,” GitHub. Available: `https://github.com/Ajinkya106/EDADV-Analysis-Code.git`