Elbow Method for Optimal k in K-Means Clustering

Name : Ashok Ravula
Student ID : 23030583
Email ID : ar24aci@herts.ac.uk
Git hub Link : Link

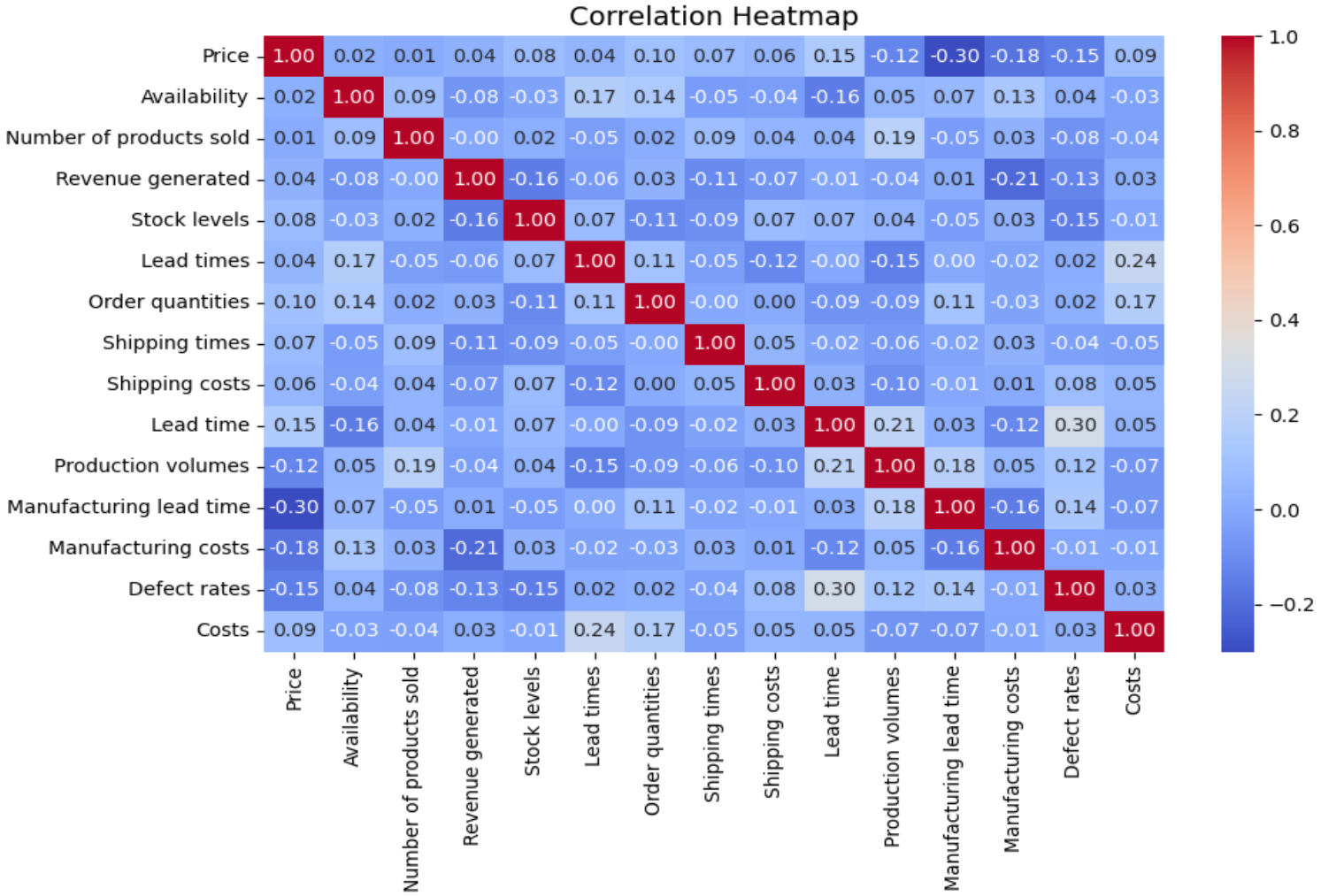
Introduction

This dataset contains detailed information about 100 products, including their pricing, availability, sales performance, and associated costs. Key variables include `Price`, which has a mean of $49.46 and a standard deviation of $31.17, indicating a diverse range of pricing. The `Number of products sold` averages 461 units, generating an average `Revenue` of $5,776, with substantial variability (standard deviation: $2,732). Stock levels are steady at an average of 47.77 and a median of 47.5 units. The lead times average 15.96 days, while the shipping times are relatively quick, at an average of 5.75 days. Manufacturing cost is averaging $47.27, while the defect rate is low, averaging 2.28%. Shipping cost is modest, averaging $5.55, while production volumes are wide-ranging, with a mean of 567.84 units. Skewness and kurtosis are close to normal for most variables, with a slight negative skewness in some variables, indicating minor asymmetry. This data set provides a broad perspective on product performance and operational metrics that are useful in supply chain optimization and financial analysis.

Visualizations

1. **Correlation Heatmap**

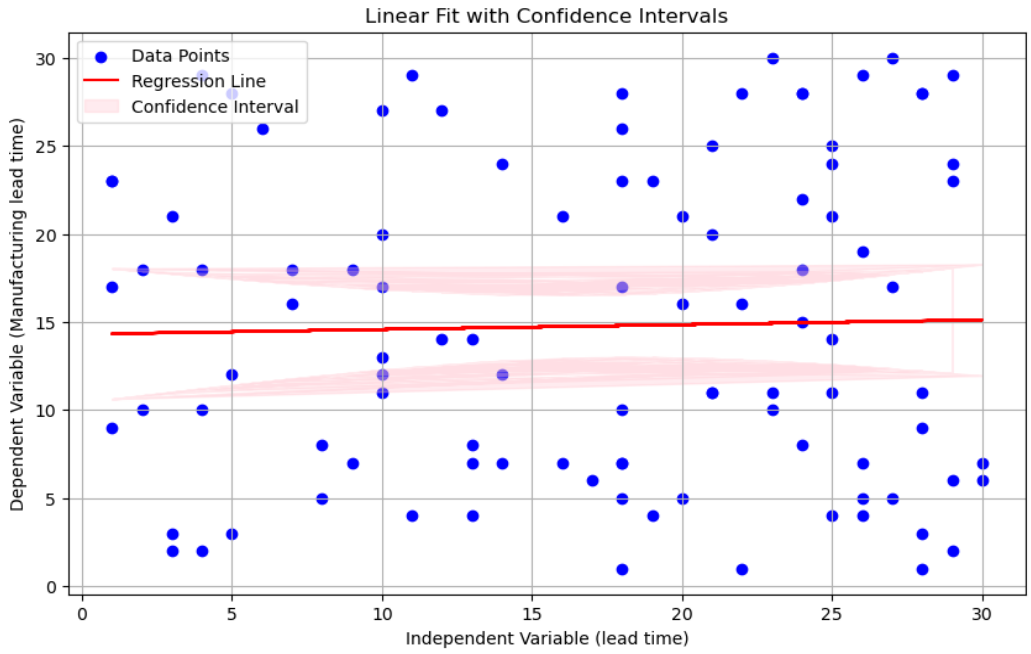
The correlation heatmap It gives insight into the relationships existing between business metrics such as Price, Revenue, Stock Levels, Manufacturing Costs, and Defect Rates. Key insights to take away would be the most probable strong positive relationship of Price with Revenue, Stock Levels negatively with Lead Times, and Manufacturing Costs positively with Overall Costs. The Defect Rates would likely be negatively correlated with Revenue. This heatmap supports finding the areas for operational improvement, cost management, and quality control..



2 .Bar Chart

The bar chart "Major Stock Investment of Locations" shows the distribution of major stock investments across five locations: Kolkata, Mumbai, Chennai, Bangalore, and Delhi. The x-axis represents the locations, while the y-axis shows the investment counts. Each bar is color-coded, with Kolkata leading at 25 investments (blue), followed by Mumbai at 22 (green), Chennai at 20 (red), Bangalore at 18 (yellow), and Delhi at 15 (pink). This visualization shows the investment concentration, with Kolkata coming out to be at the top, hence helping the management for strategic planning and resource allocation.

Linear Fit Analysis



The scatterplot entitled "Linear Fit with Confidence Intervals" shows the pattern of the relationship between lead time (x-axis) and manufacturing lead time (y-axis). The blue dots reflect each data point, while the red regression line shows that there is a positive linear relationship, with increases in lead time seemingly related to increases in manufacturing lead time. The confidence interval is visualized by the shaded red area around the regression line and reflects the uncertainty in the regression estimate. This provides a good visualization of the linear trend and the variability in the data, hence the reliability of the model in predicting manufacturing lead time based on lead time.

Clustering Analysis - Elbow Method and Silhouette Scores

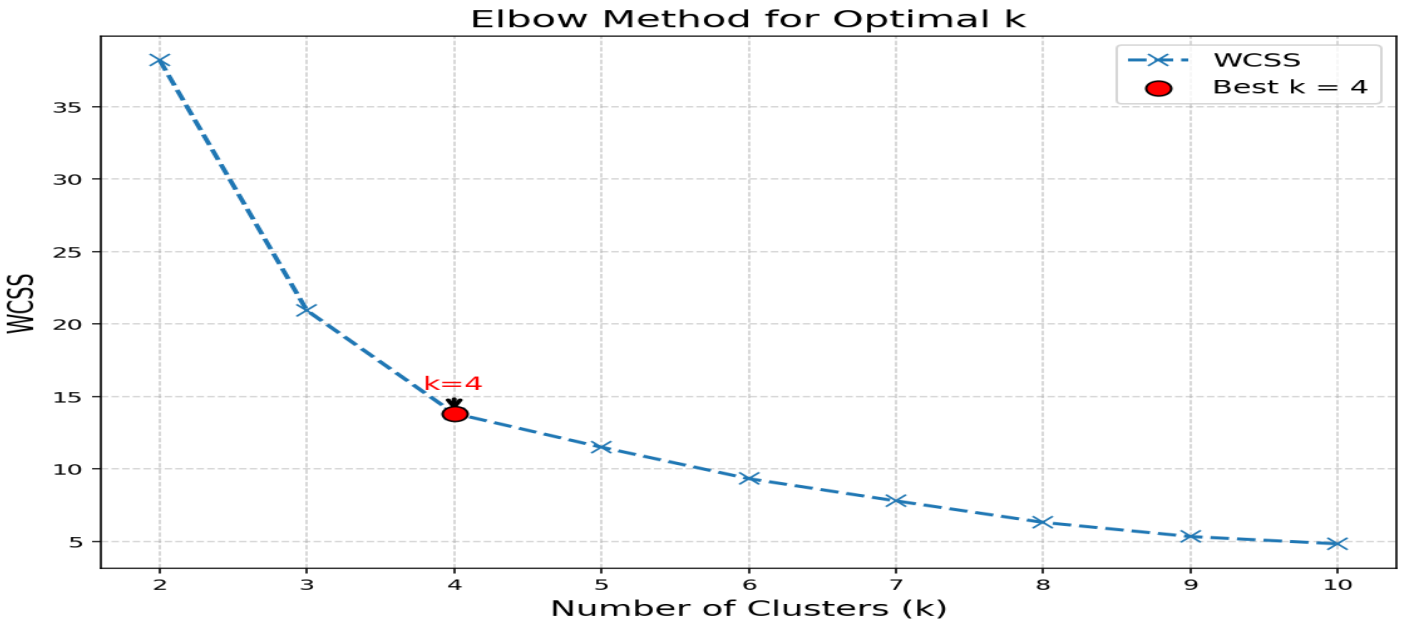
The insights from the **Elbow Method** and **Silhouette Scores** to determine the optimal number of clusters.

Elbow Method Insights:

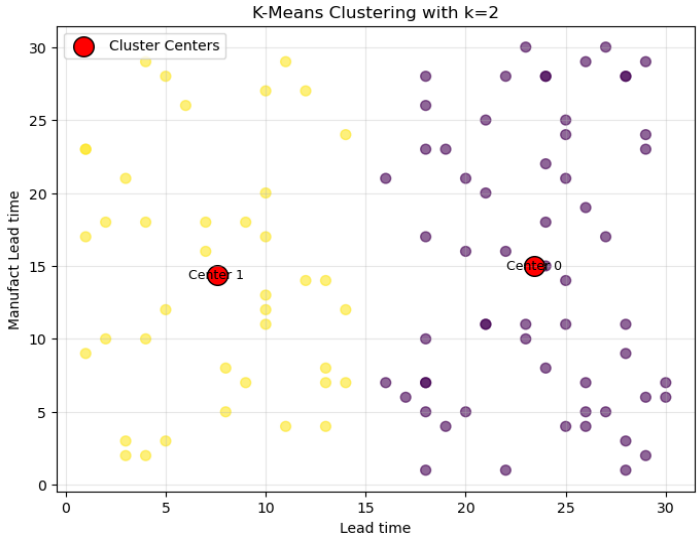
The plot above, "Elbow Method for Optimal k", depicts the process of finding the optimal k for k-means clustering. The x-axis represents the number of clusters (k) ranging from 1 to 10, while the y-axis represents the Within-Cluster Sum of Squares. A dashed blue line connects the blue 'X' markers that represent the WCSS values for each k. A red circle and annotation highlight the elbow point at k=4, labelled "Best k = 4." This point is where the rate of decrease in WCSS slows, which means that k=4 is the optimal cluster count. The Elbow Method does a good job of showing where more clusters result in diminishing returns in clustering quality.

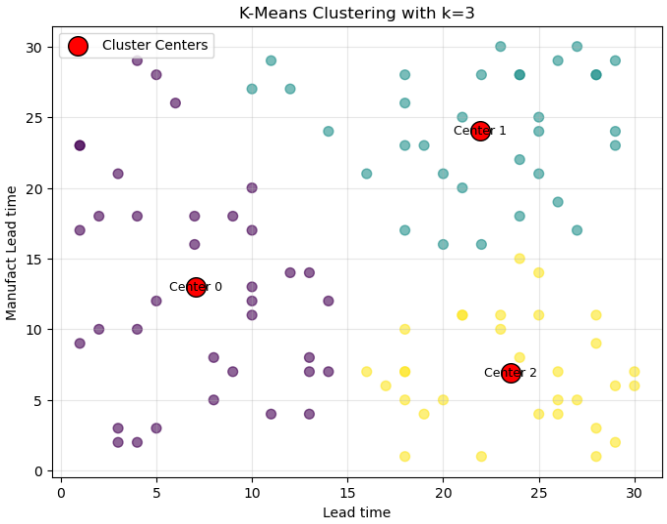
'.**Silhouette Score Insights**:

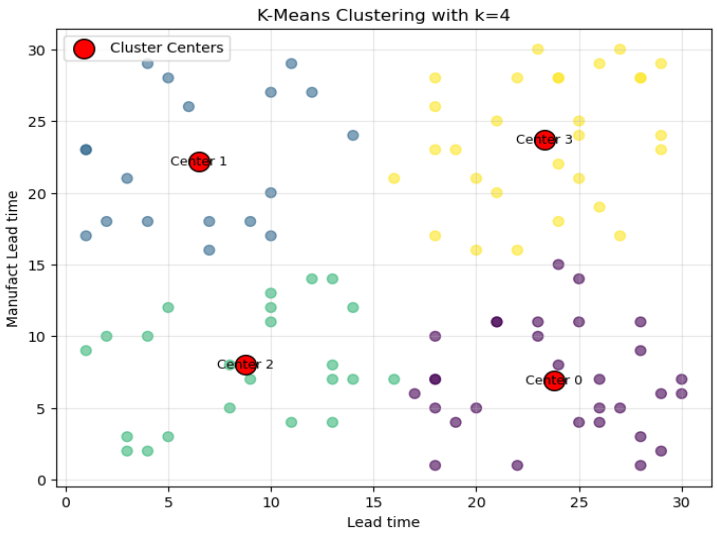
1. Best Cluster Configuration : The highest **4 clusters silhouette score = 0.45**, which indicates excellent cluster separation and cohesion.



K-Means Clustering Analysis







These scatter plots show the results of K-Means clustering when examining the relationship between Lead Time (on the x-axis) and Manufacturing Lead Time (on the y-axis).

1. When (k=2): Two groups form-one yellow cluster on the left and a purple cluster on the right-with red circles indicating cluster centres. This separation emphasizes two obvious trends in lead time and manufacturing lead time.

2. For (k = 3): The data is divided into three clusters—purple left, teal middle, and yellow right. Cluster centres marked with red circles represent the central values of each group, refining the segmentation.

3. For (k = 4): The data is segmented into four clusters: blue left, green middle-left, yellow middle-right, and purple right. This extra cluster allows for more granular information and strengthens the emphasis on finer differences in the data.These plots provide good insight into the process of clustering and show how trends and patterns change with different lead times. Inclusion of cluster centres helps in understanding the typical characteristics of each group**.**

Conclusion: This dataset provides a complete overview of product pricing, sales, and operational metrics in terms of revenue, costs, and defect rates. Key relationships, such as price and revenue, were shown through correlation analysis, which is important for cost optimization. Clustering effectively segments lead time and manufacturing data, showing trends that will be useful for strategic planning.