Supply Chain Optimization of Beauty Company (DMAIC Approach)

1st Manan Mishra
40321141
manan7340mishra@gmail.com

2nd Sandip Devrao Misal
40323469
sandipmisal1403@gmail.com

3rd Vatsal Raviya
40324600
1vatsalraviya@gmail.com

https://github.com/MananMishra-7/INSE6210

***Abstract*—In today's evolving and competitive market, it is really important that an enterprise or company's supply chain operates on highest level of efficiency avoiding all possible bottlenecks. This study focuses on using the techniques described in Six Sigma DMAIC in the supply chain data of a Beauty Company. Six Sigma is a comprehensive structured data-driven approach where improving operational efficiency is an important task. Here, DMAIC is applied to the data of a Beauty Company with the goal of reducing the defect rate by 10% and finding possible root causes utilizing techniques from root cause analysis (RCA) for a high defect rate. This study also covers exploratory data analysis, correlation analysis, cause and effect, and Pareto analysis together with the implementation of Random Forest Regressor to derive importance of features.**

***Index Terms*—Six Sigma, DMAIC, exploratory data analysis, correlation analysis, Pareto analysis, Random Forest Regressor.**

I. Introduction

DMAIC approach is a data driven methodology which is primarily used for process improvement in the Six Sigma methodology. It Consists of 5 phases: Define, Measure, Analyze, Improve and Control.Each stage is explained briefly in their respective sections. This framework is designed to systematically identify and eliminate defects, inefficiencies, variations, and improve processes [8]. Another framework used within industry is DMADV, which focuses more on implementation. It has 5 phases Define, Measure, Analyze, Design and Verify [2]. Here DMAIC approach is used which can be implemented in the industries like manufacturing, healthcare,finance, logistics etc.basically wherever any process exists and can be measured using meaningful metrics Six Sigma can be applied. Having a meaningful metric is important to so that improvements can be measured later on. DMAIC framework is a repeatable approach where the organization can apply to various challenges over time.

This paper focuses on a dataset of a Beauty Company [1]. The Data set provides a comprehensive coverage of the complete supply chain of the company showcasing real world relevance. The data set also doesn't consist of any arbitrary or missing values, making it easier to pinpoint opportunity of improvement and problematic areas. The problem addressed in this study is high Defect Rates with the current average value of 2.27% with the goal to reduce it by at least 10%. The reduction in defect rate will overall increase customer satisfaction resulting into revenue growth in future sales.

II. Tools Used

In this study we have used Microsoft Excel for preliminary tasks such as formatting feature names. For later in-depth analysis of the data set, Python programming has been utilized. Some key libraries utilized are as follows:

- Pandas
- NumPy
- Matplotlib
- Seaborn
- scikit-learn

III. Define Phase

Define phase is not only the first step of the process but also the important phase of the DMAIC approach. The main objective is to clearly identify the problem statement, set improvement goals and understand the customer needs, Project scope, defining key deliverables, and identifying stakeholders are the components of this phase [2]. Tools like SIPOC diagram, CTQ analysis are used to ensure the problem is well-structured and aligns with the business needs.

A. Problem Statement

The current supply chain is experiencing a high defect rate, which affects customer satisfaction. These defects also contribute to rework, delays in delivery and increased costs. Hence, defect rate must be reduced for better customer satisfaction. An increased customer satisfaction will also result into customer retention [3] which will will help to maintain as well as increase revenue over time.

B. Goals

The primary goal of this study is to reduce the Defect rates by 10%, by figuring out key features causing high defect rate. Other goals include identifying and addressing root cause of these defects based on the features which are contributing the most. And finally, providing few actionable and practical insights which will help to reduce the Defect Rates.

C. Key Stakeholders

Identifying stakeholders in Define phase is a important step as it helps to aligns project goals with business needs and helps to clarify role and accountability [4]. Key stakeholders in this supply chain process are manufacturing team, marketing team and the suppliers of the raw products. Another key

stakeholders are the customers and the management of the company because they are directly related to the impacts associated with the product defects and quality.

D. CTQ identification

After the preliminary analysis of the dataset, and brainstorming within the group, we were able to identify Crtical-to-Quality (CTQs) present in the Figure 1. CTQs reflects what is most important to the customers and helps in defining measurable meaningful metrics. These CTQ's identified are not specific to this data set but rather generic, but the rank of each CTQ may change based on other data set. Here the ranking is done on the basis of sorting the total score from highest to lowest, where each score was given by each team member.

CTQ (Critical to Quality)	Manan	Vatsal	Sandeep	Total
Defect Rate	10	10	10	30
Customer Satisfaction	10	10	10	30
Product Availability	8	8	10	26
Manufacturing Time	8	5	8	21
Supplier Lead Time	8	8	5	21

Score	Importance Level
0	Not Important
3	Slightly Important
5	Important
8	Very Important
10	Critical

Fig. 1. CTQ's

As shown in the Figure 1, the CTQs are rated out of 10 where, 10 being the critical and 0 being least critical.

E. SIPOC

SIPOC is a high-level process mapping tool which is used to identify all relevant elements of a process [10]. It also helps in communicating with the stakeholders and aligns the team with discussed goals. It helps to identify process components as shown below

- **Supplier** - Component Suppliers, Production department, Warehouses.
- **Inputs** - Transportation Services, Stock and inventory, Production schedule and machinery.
- **Process** - Packaging and Labeling, Dispatch via transportation, Final Inspection before shipping.
- **Outputs** - Delivery ready products, Delivery tracking, transportation costs, Packaged goods.
- **Customer** - Retailers, End Customer, Distributors.

IV. Measure Phase

Measure Phase assists in analyzing and monitoring of the dataset with the help of statistical measures [5]. It helps to capture trends and identify data points which are not within permissible limits. It also helps with establishing baselines metrics and identify variations.

A. Exploratory Data Analysis(EDA)

EDA is an important de facto practice before proceeding further with an sort of analysis because it helps to understand the data in depth by analyzing data types, central measures like mean median mode, distributions, outliers etc using graphical and statistical approaches [6]

It is crucial to understand the features and their definitions for better context and highlight some important information of each. This data set consists of 24 features in total with 100 entries. There is no null value and no missing value in the data set.As this data set is taken from Kaggle it is highly probable some sort of cleaning or interpolation has been already applied on it and it is also possible that this data set may not be real and might be synthetic.

The feature definition and some key facts discussed are as follows:

1) **Product type:** This feature describes the nature of product into three categories 'haircare', 'skincare', or 'cosmetics'. Here Skincare has the highest frequency.
2) **SKU units:** Unit number for that particular product. There are 100 unique entries for each row.
3) **Price:** Price of the product.
4) **Availability:** Number of products available for each SKU.
5) **Number of products sold:** Total units sold.
6) **Revenue generated:** Revenue generated from the each SKU.
7) **Customer demographics:** Categorical values 'Unknown', 'Male', 'Female', or 'Non-binary' are identified here. Based on the orders given by he customers most of the customers don't' share this information hence 'Unknown' has the highest frequency.
8) **Stock levels:** Quantity of stock available to ship if an order is placed.
9) **Business Lead Time:** Renamed from 'Lead time' for clarity, measured in days. Business lead time means total time taken for a business process to be completed from initiation to final outcome.
10) **Order quantities:** Quantity of orders placed.
11) **Shipping times:** Number of days required for shipping.
12) **Shipping carriers:** Categorical values representing different carriers. Their are 3 carriers ' Carrier A', 'Carrier B' and 'Carrier C' with 'Carrier B' with preferred choice over others.
13) **Shipping costs:** Cost incurred in shipping of the products.
14) **Supplier name:** Name of the supplier.Their are 5 suppliers 'Supplier 1', 'Supplier 2', 'Supplier 3', 'Supplier 4' and 'Supplier 5'. Here 'Supplier 1' has the highest frequency followed by 'Supplier 2'.
15) **Location:** Locations include 'Mumbai', 'Delhi', 'Bangalore', 'Kolkata' and 'Chennai'. Here 'Kolkata' is location where maximum number of orders are delivered.
16) **Supplier Lead Time:** Renamed from 'Lead time' to avoid confusion.
17) **Production volumes:** Quantity produced.
18) **Manufacturing Lead Time:** Time required to manufacture products once a request has been made .
19) **Manufacturing costs:** Costs involved in manufacturing.

20) **Inspection results:** Values include 'Fail', 'Pending', and 'Pass'. Here most of the SKUs have 'Pending' Status on them
21) **Defect rates:** Percentage of defective items on the basis of each SKU. This is our target variable which we are trying to reduce. The current average defect rate is 2.27% with the standard deviation of 1.46% with the median value of 2.14%. The Max defect rate is 4.94% and minimum is 0.02%.
22) **Transportation modes:** Categorical values like 'Road', 'Sea', 'Air', and 'Rail'. Here 'Road' and 'Rail' are preferred mode of transportation.
23) **Routes:** There are 3 routes defined for transportation 'Route A', 'Route B' and 'Route C'. ' Route A' seems to be the most preferred route for transportation.
24) **Transportation cost:** Renamed from 'costs' to avoid confusion with other cost features and better understanding .

Figure 2 shows the Average supplier Lead time for each supplier . Based on which it is evident that 'Supplier 3' takes the most time.

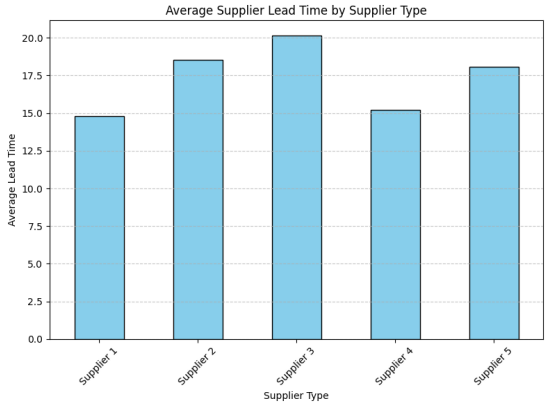

Fig. 2. Average supplier Lead

Figure 4 is an important diagram often mentioned as Correlation Matrix or Heat Map. It's generated using the Python code shown in Figure 3

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

Here:

- r is the correlation coefficient.
- x_i and y_i are the individual data points of variables x and y respectively.
- $\bar{x}$ is the mean of the variable x.
- $\bar{y}$ is the mean of the variable y.
- n is the number of observations.

The correlation coefficient r shows the strength of the linear relationship weather direct or inverse between two variables [11]. Its value ranges between -1 and 1(unlike covariance):

- $r = 1$ implies a perfect direct linear correlation.
- $r = -1$ implies a perfect inverse linear correlation.
- $r = 0$ implies there is no linear correlation.

```
numerical_variables = ['Price', 'Availability', 'Number of products sold',
                       'Revenue generated', 'Stock levels', 'Business lead time', 'Order quantities', 'Shipping times',
                       'Shipping costs', 'Supplier lead time', 'Production volumes', 'Manufacturing lead time',
                       'Manufacturing costs',
                       'Defect rates', 'Transportation costs']
correlation_matrix = df[numerical_variables].corr(method='pearson')


plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)
plt.title('Correlation Matrix')
plt.show()
```

Fig. 3. Code Snippet

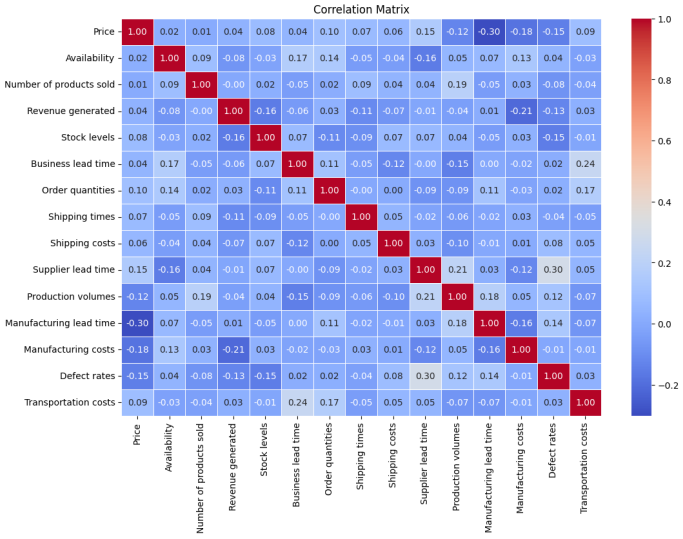

Fig. 4. Heatmap

Figure 4 shows the output of the variables which are highly correlated to each other which helps to draw potential reasons for high defect rate.

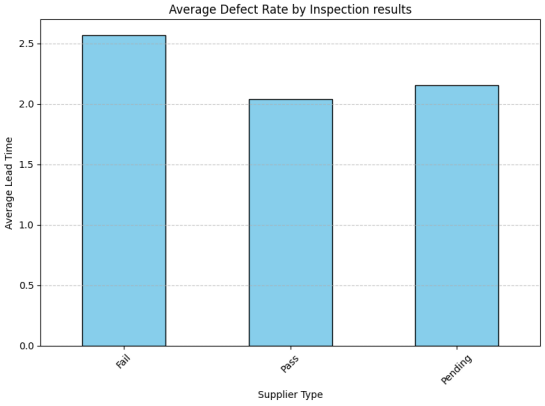

Fig. 5. Average Defect Rate based on Inspection Result

Figure 5 shows that whenever the inspection results fail the defect rate is highest with an value average value above 2.5% In Figure 6 blue bar shows the Average Defect Rate by different transportation modes and respective routes. It displays that (Rail, Route C) has the highest defect rate of of 2.5%. The red line where as displays the corresponding

frequency for each of them. For (Rail, Route C) the defect rate may be the highest but the frequency is the lowest among all.

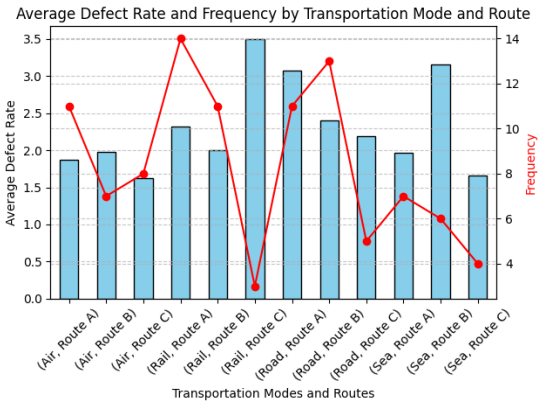

Fig. 6. Average Defect Rate by Transportation Mode and Route

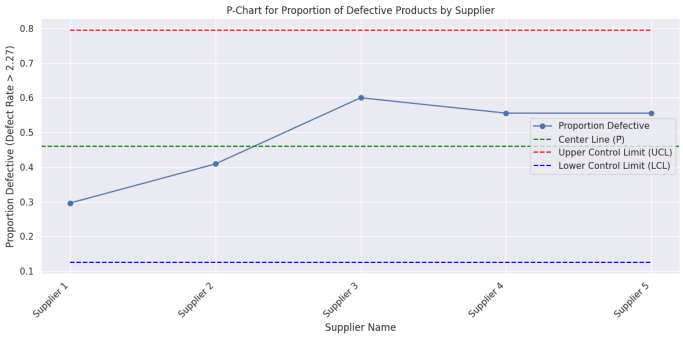

Fig. 7. Supplier Name P-Chart

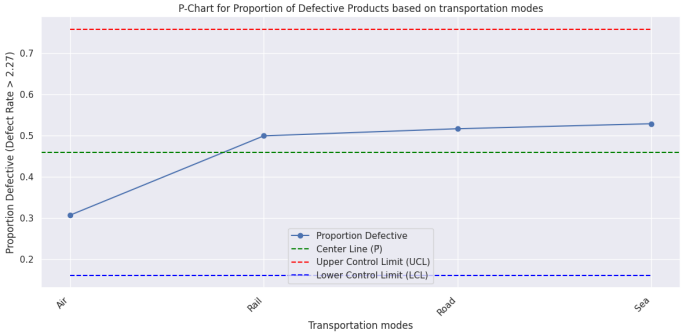

Fig. 8. Transportation Mode P-Chart

Figure 7 Figure 8 Figure 9 Figure 10demonstrates P-Charts where every item which has a defective rate above the average value 2.27% is considered defective. X axis consists of different categories of different features. Here the sample size is average sample size so that the UCL and LCL are uniform across all samples.

Figure 7 Figure 8 Figure 9 Figure 10 shows that the process is may be capable but to conclude that the process first must

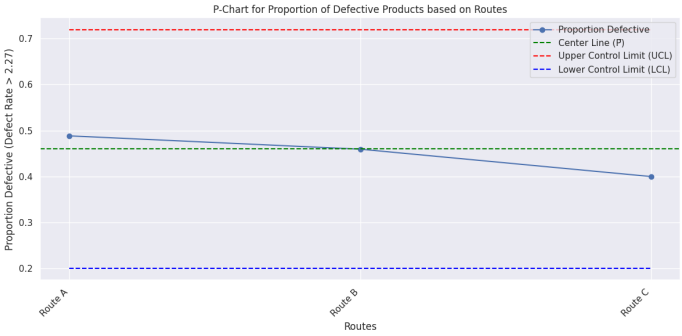

Fig. 9. Routes P-Chart

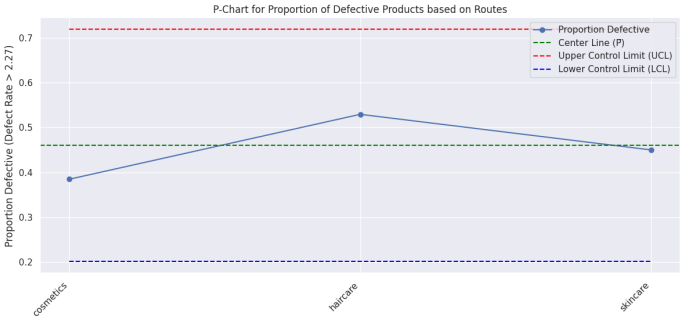

Fig. 10. Product Type P-Chart

be is statistical control. Figure 14 is a run chart with Y axis showing the defect rate and X- Axis showing SKU .Based on this graph it can be concluded that the process is not capable as one SKU to other SKU the variation is too much and at multiple instances Western Electric rules are violated.

V. Analysis Phase

This phase focuses on the identification of the root causes based on the data and insights collected in the Measure phase [12]. These root causes are not based on intuition but are rather derived statistically or logically.

A. Pareto Analysis

Pareto Analysis is a very common illustration which works on 80/20 rule. The rule states that 80% of problems are caused

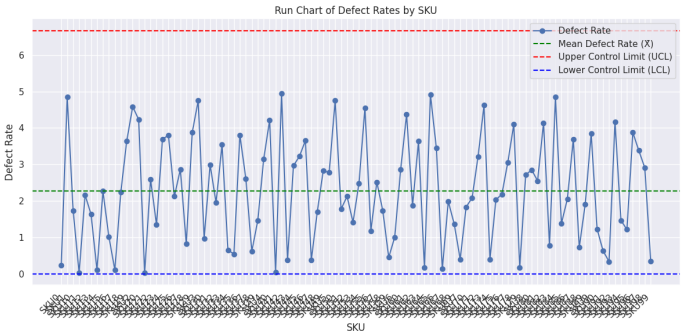

Fig. 11. Defect Rate and SKU

by 20% of causes [13]. In other words it can be said as, high Defect Rate is not because of all possible causes but it is due to some specific features. Hence it is important to figure out the features from the data set which are causing high Defect Rates.

To figure out these few important features, Random Forest (RF) model is used. RF is a popular model which is used for both regression and classification task [14]. It is often considered over Decision Trees as it reduces the chances of overfitting on a small data set [15]. For the extraction of important features causing high defect rates, only numeric values have been used and all the categorical variables are not considered. There are ways to convert the categorical variable to numerical by manually assigning them integer values or using some inbuilt function like `LabelEncoder()`. These ways work best when the categorical variable are ordinal in nature, meaning there is order by which they can be ranked. For example in case of Supplier type there are 5 sub categories: Supplier 1-5, which cannot be ordered like an ordinal data type. Hence the task in hand is a regression one as the selected independent and dependent features are numerical in nature. Scikit-learn library in python provides `RandomForestRegressor` for this task.

RF model is trained on 70% of data and remaining 30% on predictions. Based on the output of RF feature importance are shown in Figure 12. Based on Figure 12 values Figure 13

	Feature	Importance
0	Supplier lead time	0.213478
1	Manufacturing lead time	0.121494
2	Price	0.097101
3	Number of products sold	0.074198
4	Production volumes	0.066282
5	Business lead time	0.059496
6	Transportation costs	0.057378
7	Shipping costs	0.056376
8	Stock levels	0.050610
9	Manufacturing costs	0.043974
10	Order quantities	0.043941
11	Availability	0.042410
12	Revenue generated	0.041760
13	Shipping times	0.031503

Fig. 12. Feature importance

can be derived. Some possible reasons to explain similar importance value between features In Figure 13 are as follows:

- **Genuine low importance:** It's possible that these features as compared to lead times does not impact the Defect Rate significantly
- **Correlation between Features:** Features are correlated to each other as shown in the EDA due to which models like RF often split importance between features.

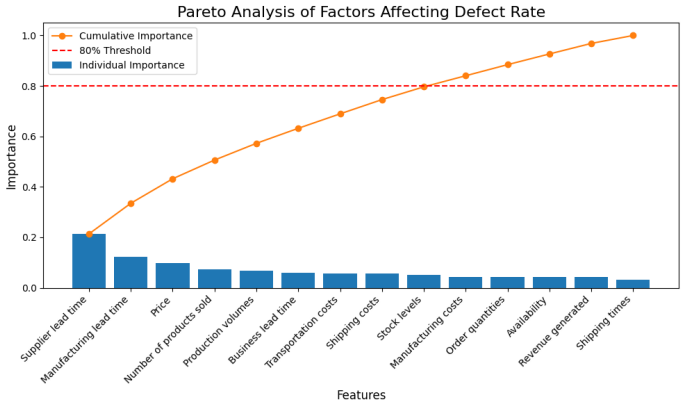

Fig. 13. Pareto Chart

- **Redundant Features:** Their might be redundant features in the data set due to which the values are similar.

B. Fish-bone Diagram

Fish-bone diagram which is also known as the Ishikawa or Cause-and-Effect diagram, is used to systematically identify and organize potential root cause. It explores all the possible factors which lead to an issue, ensuring that the solutions address the real root cause [2]. As seen in the Fig. 14, there are 5 categories - Man, Method, Material, Machine, and Measurement, which can be be considered as a potential root cause. Under the Man category, it highlights the issues like untrained staff and inadequate supplier capacity suggesting limitations in the workforce. Whereas in the Method category, inefficiencies like long manufacturing and supplier lead time are identified. The Material category identifies incorrect packaging and late deliveries as contributors to defects. Machine issues like breakdowns, long setup times, and inappropriate tools point to equipment related inefficiencies. In Measurement, the absence of root cause tracking and lack of trend analysis lead to defects in the process.

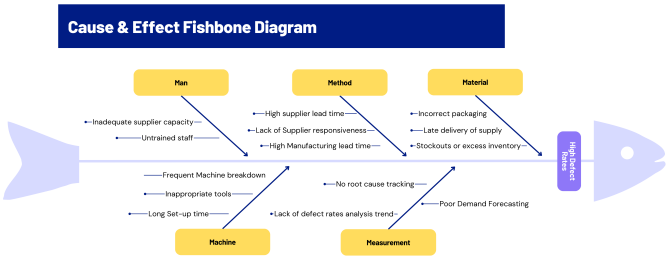

Fig. 14. Fishbone Diagram

C. Root Cause Analysis(RCA)

RCA is an crucial task after figuring out most important factors affecting the Defect rates, but its important to find

the underlying problems rather than symptoms [16]. To find potential root causes in causing high Supplier lead time, high Manufacturing lead time and less Number of products sold Why-Why Analysis can be depicted in Figure15, Figure16 and Figure17

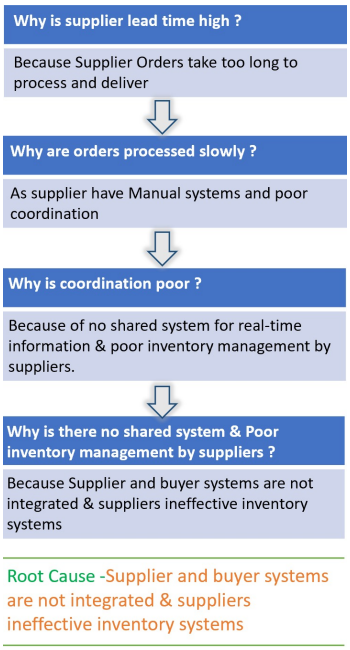

Fig. 15. High Supplier Lead time

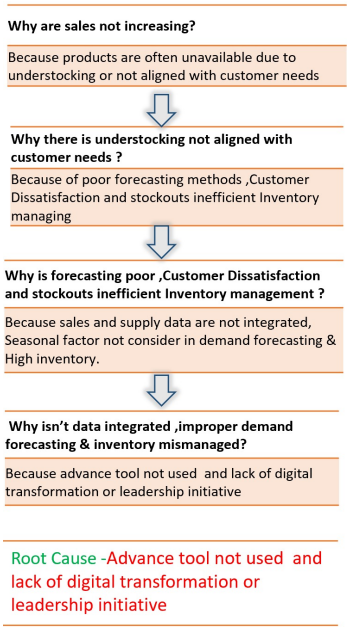

Fig. 16. High Manufacturing Lead time

VI. Improve

The improvement phase focuses on the identification and testing of probable solutions for the possible root causes determined in the analysis phase. These solutions are researched, brainstormed and then implemented to check if they can give expected results.

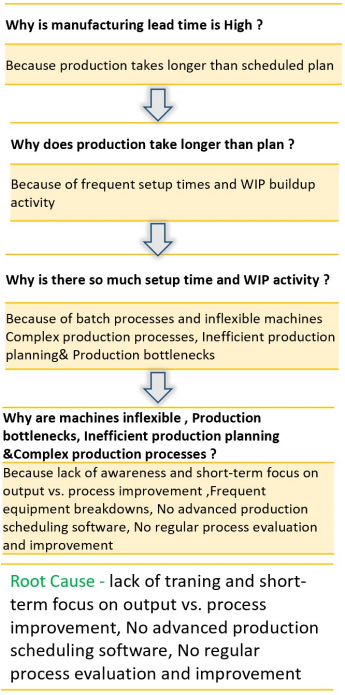

Fig. 17. Less Number of Products sold

A. Feature Modification

The RF model was able to extract the most important features which are causing high defect rate. The top 5 features are shown in TableI.

Feature	Importance
Supplier lead time	0.213478
Manufacturing lead time	0.121494
Price	0.097101
Number of products sold	0.074198
Production volumes	0.066282

TABLE I
Top 5 Features

Out of these 5 features only 3 features will be modified to achieve the goal set at the start. These 3 features are Supplier lead time, Manufacturing lead time and Number of products sold. Reason behind selecting these features are that these features align closely with the identified stakeholders in Define Phase. Also these features does not affect the nature of the products directly whereas fluctuating Price or Production Volume might change the nature of products.
After the selection of the key important features, their values in the test data will be modified and predictions will be made on 2 separate test data one which will have the improved features and the other original test data. The Changes made in the features are as follows:

- Number of products sold (improved) = 1.3*Number of products sold

- Supplier lead time (improved) = 0.6*Supplier lead time
- Manufacturing lead time (improved)= 0.7Manufacturing lead time

The promotion and reduction of features is not done on arbitrary basis, rather correlation analysis is used for it as shown in Figure18.Defect rate has a negative correlation with number of products sold hence it is increased by 30%. Supplier lead time and Manufacturing lead time on the other hand have a positive correlation hence their values are reduced by 40% and 30% respectively. The amount by which the features are modified where determined by hit and trial. It is important to note that for the calculation of correlation values only numerical variables were considered and no categorical was used in for this.

Fig. 18. Defect Rate Correlation

The RF model is now ready to be tested on 2 test data set to show if their are any significant changes in the predicted Defect Rate. Average value is calculated first on the unchanged data and then it is calculated on the modified data. Let:

$$\hat{y}_{\text{improved}} = \text{rf.predict}(X_{\text{test, improved}})$$

$$\hat{y}_{\text{original}} = \text{rf.predict}(X_{\text{test}})$$

Then:

$$\text{Original Average Score} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_{\text{original},i}$$

$$\text{Improved Average Score} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_{\text{improved},i}$$

Metric	Average Score
Original	2.04
Improved	1.77

TABLE II
COMPARISON OF ORIGINAL AND IMPROVED AVERAGE SCORES

Hence based on the numbers obtained:

$$\text{Percentage Change} = \left(\frac{\text{Original} - \text{Improved}}{\text{Original}}\right) \times 100$$

$$= \left(\frac{2.04 - 1.77}{2.04}\right) \times 100$$

$$= \left(\frac{0.27}{2.04}\right) \times 100$$

$$\approx 13.24\%$$

Hence the proposed changes in the feature predicts a reduction of defect rate by 13.24%. These results highlights the fact that changing mere 3 features reduced the predicted defect rate significantly.

B. Proposed Solutions

To address high supplier lead times, a dynamic approach is proposed, which would diversify the suppliers through multi-sourcing and near-shoring. This would include auditing new suppliers and then leveraging advance payments to enhance supplier responsiveness. Also maintaining safety stock, dividing complex products into sub assemblies are some of the strategies which would be helpful in managing supply chain risks. The use of latest AI/ML tools, premium logistics and greater collaboration can further improve supply reliability. Implementing Kanban systems, developing contingency plans, and fostering long-term supplier relationships are also crucial. For High manufacturing lead time, preventive maintenance, adopting advanced production scheduling software, continuous evaluation and streamlining of production processes should be implemented.
To tackle low sales, demand forecasting tools, JIT inventory methods, and improved safety stock policies by each region should be included. Customer service should be enhanced and outsourcing low-demand items should be tailored to distribute network. Sales can be also be improved based on region analysis on the basis of 'Locations'

VII. CONTROL PHASE

Control phase is the final stage of the DMAIC process. It focuses on sustaining the improvement made during the process and ensures that the improvements are maintained [2]. Key activities include developing control plans, standardizing new procedures and implementing control charts. The goal is ensure long-term stability and improvement aligned with customer expectations. Below information and steps reflect the measures that should be taken to control this process.

1) **Process Monitoring System** - A strong monitoring system guarantees that essential supply chain metrics including lead times, inventory levels, and order fulfillment rates should be regularly monitored. Using dashboards, alerts, and real-time data gathering to monitor the process. Implementing systems such as SCADA, ERP, or tailored analytics platforms enhances co-operation among departments and suppliers, encouraging a culture of open-ness and data-informed management [9].
2) **Response Plan for Out-of-Control Situations** - Creating an well structured response strategy for uncontrolled situations like shipping delays, supplier issues, or unexpected demand increases reduces interruptions [9]. These plans detail responsibilities, escalation, and remedial measures to be implemented when any metrics exceed its control limits. The action plan helps to measures & regain stability. It also helps to record every incident for understanding the root causes. Using

this framework enhances organizational flexibility and ensures continuity in the supply chain to avoid uncertainties.

3) **Standardized Operating Procedures (SOPs)** - SOPs offer uniform set of instructions for executing standard supply chain activities, including order processing, warehousing, and supplier engagements. They remove uncertainty, minimize mistakes, and helps to achieve quality standards. By standardizing activities, organizations can ensure process control despite changes in the workforce or scaling initiatives [9]. SOPs also function as training resources for new staff and a standard for audits and process evaluations. In Six Sigma, standardization is essential for maintaining enhancements and guaranteeing that optimized processes are consistently replicated across teams and locations.
4) **Management Review Process** - Regular management review meetings with structured agenda increase the efficiency of control systems, review Key Performance Indicators(KPIs), and make informed choices for ongoing enhancement. These evaluations consist of assessing performance patterns, recognizing risks, and ensuring that control measures continue to align with business objectives. The procedure encourages cross-functional collaboration, and strengthens quality and performance [9]. Recording and monitoring results from these assessments promotes transparency and enables follow-up to identify new improvement opportunities.
5) **Continuous Improvement Framework** - A methodology such as PDCA (Plan-Do-Check-Act) or Kaizen aids in enhancing the supply chain effectiveness and advancement over time. This structure motivates teams to identify inefficiencies at workplace and evaluate outcomes. Mechanism for updating the control plan helps to avoid uncertainties in supply chain. Employees are encouraged to propose improvements, managers assess ideas with data, and outcomes are recorded in knowledge systems [9]. Over time, this encourages operational efficiency and enhances the ability to adjust to evolving market needs or supply chain challenges.
6) **Demand Forecast Validation System** - Predictions guide supply planning, but incorrect estimates can result in inefficiencies. A validation system evaluates predictions against real demand and determines forecast error. Metric such as Mean absolute percentage error (MAPE) or forecast bias can be used to improve Demand Forecasting models [9]. Cooperation among sales, marketing, and supply chain teams reducing the forecast error and improve predictions include market insights. Gradually, this system creates a dependable and flexible planning system capable of adjusting to business trends and seasonality factor.

VIII. Future works

In this whole study categorical features like Supplier Name, Location, Transportation Mode, Shipping Carrier etc are only considered till the EDA and not in the Analysis phase. To incorporate these nominal features also these following techniques can be used:

- **One-Hot Encoding**: In this method, a new column is defined for each category [17]. For example, if *Transportation Mode* is considered for this type of encoding, for each category (Road, Rail, Air, and Sea), a new feature will be defined like `Transportation_Mode_Air` with 1 in the column matching the category and 0 in all other cases.
- **Frequency Encoding**: In this method, each category in a particular column is replaced by its frequency [18].

These changes can help us to incorporate the categorical variables also in our analysis. It's not necessary that only one type of encoding is done on every categorical feature but rather based on use case a mix of encoding techniques can be used.

References

[1] A. Motefaker, "Supply Chain Dataset," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/amirmotefaker/supply-chain-dataset. [Accessed: Apr. 24, 2025].

[2] K. Selvi and R. Majumdar, "Six sigma—overview of DMAIC and DMADV," *Int. J. Innov. Sci. Mod. Eng.*, vol. 2, no. 5, pp. 16–19, 2014.

[3] R. T. Rust and A. J. Zahorik, "Customer satisfaction, customer retention, and market share," *J. Retailing*, vol. 69, no. 2, pp. 193–215, 1993.

[4] A. H. Adepoju, A. Eweje, A. Collins, and O. Hamza, "Developing strategic roadmaps for data-driven organizations: A model for aligning projects with business goals," *Int. J. Multidiscip. Res. Growth Eval.*, vol. 4, no. 6, pp. 1128–1140, 202

[5] Q. P. He and J. Wang, "Statistical process monitoring as a big data analytics tool for smart manufacturing," *J. Process Control*, vol. 67, pp. 35–43, 2018.

[6] E. Camizuli and E. J. Carranza, "Exploratory data analysis (EDA)," *The Encyclopedia of Archaeological Sciences*, pp. 1–7, 2018.

[7] George, M. L., Rowlands, D., Price, M., & Maxey, J. (2005). The Lean Six Sigma Pocket Toolbook: A Quick Reference Guide to Nearly 100 Tools for Improving Quality and Speed. McGraw-Hill Education.

[8] Christopher, M. (2016). Logistics & Supply Chain Management (5th ed.). Pearson Education Limited.

[9] "Kaizen: The Art of Endless Evolution," Leading Business Lmprovement, 2024. [Online]. Available: https://leadingbusinessimprovement.com/kaizen/.

[10] Brown, C. (2019). Why and how to employ the SIPOC model. In *Journal of business continuity & emergency planning*, 12(3), 198–210. Henry Stewart Publications.

[11] Asuero, A. G., Sayago, A. and González, A. G. (2006). The correlation coefficient: An overview. *Critical reviews in analytical chemistry*, 36(1), 41–59. Taylor and Francis.

[12] Christopher, M. (2016). Logistics & Supply Chain Management (5th ed.). Pearson Education Limited.

[13] Pyzdek, T. (2021). Pareto analysis. In *The Lean Healthcare Handbook: A Complete Guide to Creating Healthcare Workplaces* (pp. 157–164). Springer.

[14] GeeksforGeeks. (2021). Random Forest Regression in Python. Retrieved from https://www.geeksforgeeks.org/random-forest-regression-in-python/

[15] GeeksforGeeks. (2021). Overfitting in Decision Tree Models. Retrieved from https://www.geeksforgeeks.org/overfitting-in-decision-tree-models/?ref=asr30

[16] Leszak, M., Perry, D. E., & Stoll, D. (2000). A case study in root cause defect analysis. In *Proceedings of the 22nd International Conference on Software Engineering* (pp. 428–437).

[17] Seger, C. (2018). An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing.

[18] Bolikulov, F., Nasimov, R., Rashidov, A., Akhmedov, F., & Young-Im, C. (2024). Effective methods of categorical data encoding for artificial intelligence algorithms. *Mathematics*, 12(16), 2553. MDPI AG.