Project Report

Course Code: APS 4742

Data-Driven Quality Improvement in Beverage Manufacturing

A Lean Six Sigma Approach Using Control Charts and Root Cause Analysis

Group No: 5

Department of Environmental and Industrial Sciences

Faculty of Science

University Of Peradeniya

Date: 04 July 2025

Abstract

This study presents a data-driven quality improvement initiative within a non-carbonated beverage manufacturing process, employing Lean Six Sigma methodology with a focus on DMAIC (Define, Measure, Analyze, Improve, Control). The project aims to address the increasing defect rates in product filling—particularly underfilled bottles—in a production line comprising three beverage types: Aloe Juice, Energy Drink, and Protein Shake. Using a dataset of 720 observations obtained from a Kaggle-sourced simulation of a U.S.-based startup, statistical process control tools such as control charts, ANOVA, and regression analysis were applied to identify the root causes of variability in bottle weight. The analysis revealed significant differences in mean weight across product types and machines, while shift and operator-based variations were statistically insignificant. Control chart analysis detected special cause variations, indicating moments of process instability. Based on these findings, targeted interventions such as recalibrating machines, standardizing operator procedures, and introducing routine monitoring were proposed and implemented. These measures led to improved filling consistency, reduced defect rates, and enhanced overall production efficiency. The project highlights the effectiveness of Lean Six Sigma tools in driving continuous improvement and recommends further integration of automation and predictive analytics for sustained quality control.

Table of Contents

1 Introduction

1.1 Background

This report details a data-driven quality improvement initiative within a USA-based startup company operating in the fast-moving consumer goods (FMCG) sector, specifically in non-carbonated beverage production. The company produces three main product lines: Aloe Juice, Energy Drink, and Protein Shake, using automated filling and packaging machines. Its production process runs across two shifts daily (Morning and Evening), managed by a team of four trained operators utilizing three distinct machines: PX1, PX2, and PX3. The dataset used for this analysis was sourced from a Kaggle dataset, encompassing production data from 2023 to 2024, and while the company's total population is 10,000, this study focuses only on operational data related to the beverage filling process.

1.2 Problem Statement

Recent production data has indicated a significant increase in defects, particularly within the Energy Drink and Protein Shake lines. The primary defect identified is "underfilled bottles," where the net weight is less than 495g, below the acceptable specification of 500g ±5g. This issue directly affects product quality, customer satisfaction, and compliance with established quality standards. It calls for a systematic approach to identify and mitigate the root causes of these variations in the production process.

1.3 Objectives

The main objective of this project is to apply Lean Six Sigma principles and statistical process control, specifically using control charts and root cause analysis, to investigate the bottle weight variation in the beverage manufacturing process. The aim is to:

- Identify root causes contributing to underfilled bottles
- Implement effective control measures to reduce defects
- Bring the production process within acceptable control limits
- Improve overall product quality and consistency.

1.4 Scope and Limitations

This study focuses on the analysis of bottle weight variations and associated defects within the beverage filling and packaging process, covering data from the three primary machines (PX1, PX2, PX3), two production shifts (Morning and Evening), and the three product lines (Aloe Juice, Energy Drink, Protein Shake). The methodology primarily involves statistical analysis, including control charts and Fishbone diagrams, to identify inefficiencies and suggest improvements. However, the study acknowledges several limitations:

- Limited data on machine downtime, which could provide deeper insights into process interruptions
- Restricted information on raw material variations, which may also influence weight inconsistencies
- Future studies could benefit from integrating a broader range of process parameters for a more holistic and comprehensive understanding of contributing factors.

2 Methodology

2.1 Application of DMAIC and Control Charts for Process Improvement

This study was conducted using the Lean Six Sigma framework, specifically following the DMAIC (Define, Measure, Analyze, Improve, Control) methodology and the use of control charts in addressing issues within the beverage production process, particularly focusing on underfilled bottles.

2.1.1 DMAIC Methodology

Define: Problem Context and Goals

Problem: The production line has been facing issues with underfilled bottles, with the product's net weight specification of 500g ± 5g not being met. Defects, particularly underfilled bottles, were impacting production efficiency and customer satisfaction. Recent data shows a significant increase in defects, especially in the Energy Drink and Protein Shake lines.

Goal: The primary goal was to reduce the defect rate of underfilled bottles and ensure consistency in the filling process across shifts and machines. The target was to reduce defects to below 1% and bring the process within control limits.

Measure: Data Collection and Current Performance

- Collected data on the weight variation of the bottles (e.g., average weight, standard deviation).
- Identified defect rates, such as the percentage of underfilled bottles (less than 495g).
- Data was gathered per shift, per machine (PX1, PX2, PX3), and across different product lines (Aloe Juice, Energy Drink, and Protein Shake).

Analyze: Identifying Root Causes

In this phase, collected data was analyzed to identify patterns and root causes for the underfilling issue. Comparative analysis between machines and shifts was performed to pinpoint inconsistencies.

Improve: Solutions and Enhancements

Based on the analysis, process adjustments were proposed and tested. These included recalibration of filling machines, training of operators, and real-time monitoring enhancements to ensure optimal fill levels.

Control: Ensuring Sustainability

Control measures were implemented to maintain process improvements, including routine monitoring, updated SOPs (Standard Operating Procedures), and the use of control charts for ongoing oversight.

2.2 Control Charts: Application and Impact

Control charts are critical tools used to monitor and control process variability. In this project, an **X-bar chart** was used to track the average weight of the bottles over time.

Deviations beyond the Upper Control Limit (UCL) or Lower Control Limit (LCL) indicated process instability, such as inconsistent filling or machine issues. Regular review of these charts enabled timely intervention and ensured the process stayed within acceptable control limits.

3 Data Analysis and Results

3.1 Data Overview

The dataset consists of 720 observations and 8 variables related to the quality control process in a manufacturing environment. The variables include:

- **Date** – The production date.
- **Shift** – The working shift during which production occurred (e.g., Morning, Evening).
- **Product** – The type of product manufactured.
- **Operator** – The name or ID of the operator handling the machine.
- **Weight_g** – The weight of the product in grams.
- **Temperature_C** – The temperature of the machine during operation (°C).
- **Machine_ID** – The identifier of the machine used.
- **Defect** – A binary variable indicating whether a defect was present (1 = defect, 0 = no defect).

All variables were checked for missing values, and no missing data was found. The dataset includes a mix of categorical (e.g., *Shift*, *Product*) and continuous (e.g., *Weight_g*, *Temperature_C*) variables.

3.2 Analysis Tools and Environment

The analysis was conducted using **RStudio**, leveraging the following key packages:

- `dplyr` – for data manipulation.
- `ggplot2` – for data visualization.
- `psych` – for descriptive statistics.
- `qcc` – for quality control charts.

These tools enabled efficient exploration, visualization, and statistical evaluation of the data.

3.3 Descriptive Statistics

3.3.1 Qualitative Data Analysis: Shift, Product, Operator, and Machine

Categorical variables in the dataset—*Shift*, *Product*, *Operator*, and *Machine_ID*—were analyzed using frequency distributions and visualized using bar and pie charts.

Operator-Wise Observations

The highest number of observations was recorded under David (194), followed by John (179), Kevin (175), and Mark (172). The distribution is fairly balanced, suggesting consistent operator involvement across shifts and tasks.

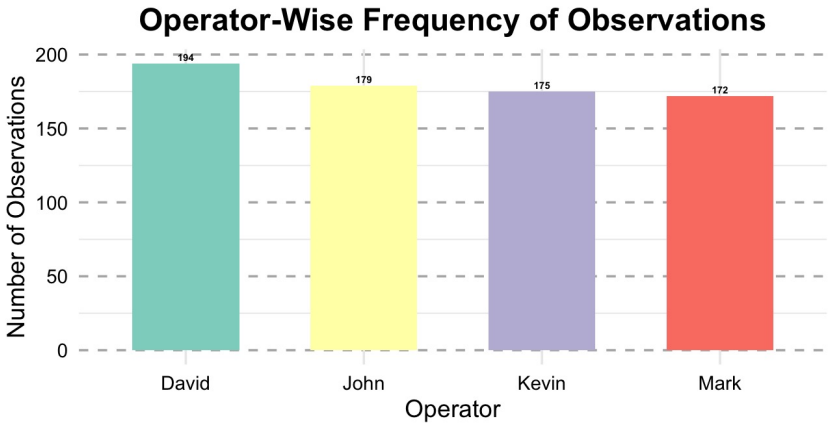

Figure 3.1: Bar chart of observations by operator

Product Type Distribution

EnergyDrink had the highest production volume (296 observations), followed by *AloeJuice* (269) and *ProteinShake* (155). This indicates a possible production priority or higher demand for EnergyDrinks during the data collection period.

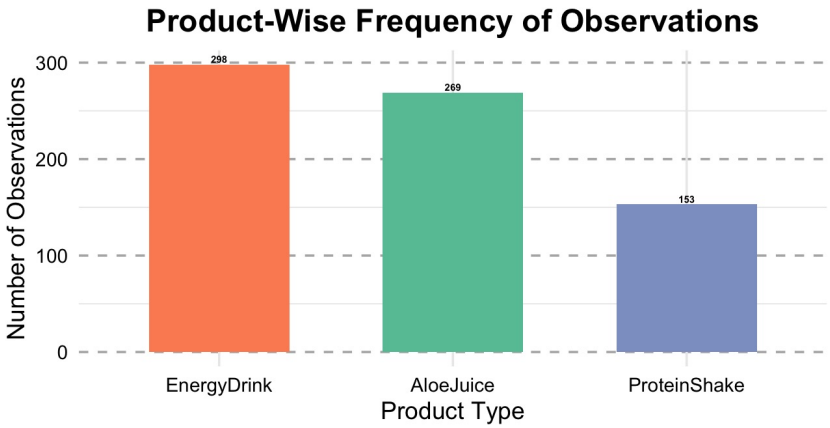

Figure 3.2: Product type distribution

Shift Distribution

The pie chart indicates a perfect 50-50 split between Morning and Evening shifts. This reflects a well-balanced operational schedule across working hours, eliminating potential bias due to shift-based workload variations.

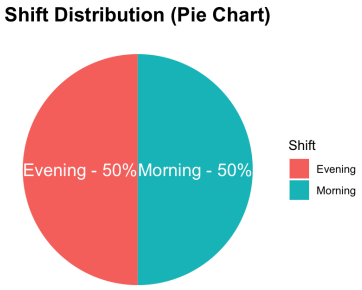

Figure 3.3: Shift-wise distribution of observations

Machine-Wise Observations

Observations were fairly distributed across three machines:

- **PX2** – 245 observations
- **PX3** – 241 observations
- **PX1** – 234 observations

Such even distribution supports the reliability of cross-machine performance comparisons and ensures that results are not skewed by unequal usage.

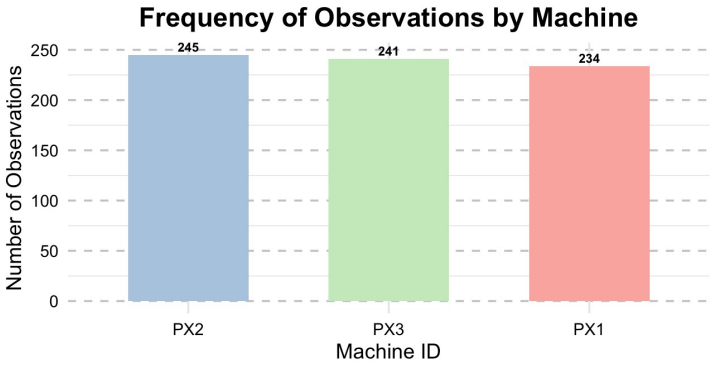

Figure 3.4: Machine-wise distribution of observations

3.3.2 Quantitative Data Analysis: Weight

Spread of Data

Table 3.1: Spread of Weight Data

Statistic	Value
Minimum	481.3700 g
Maximum	529.2000 g
Range	47.8300 g
Variance	46.6801
Standard Deviation	6.8323 g
1st Quartile (Q1)	495.4775 g
3rd Quartile (Q3)	502.4475 g
Interquartile Range (IQR)	6.9700 g
Skewness	0.81
Kurtosis	1.31

Central Tendency

Table 3.2: Measures of Central Tendency

Statistic	Value
Mean	499.7431 g
Median	499.0250 g
Mode	499.4500 g

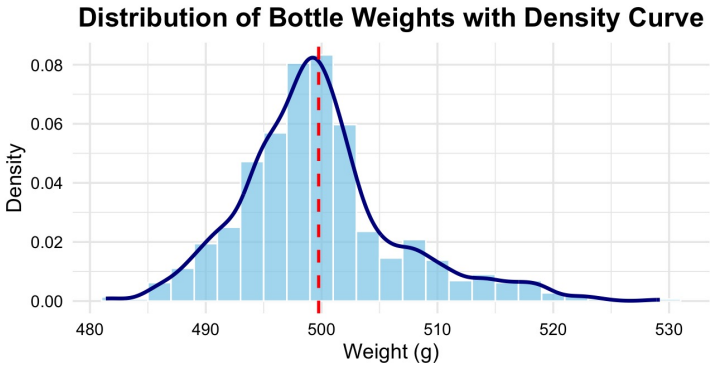

Figure 3.5: Histogram of Product Weights

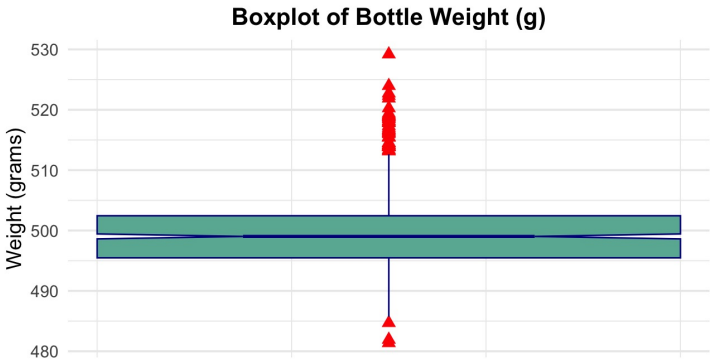

Figure 3.6: Boxplot of Product Weights

These values are very close, indicating a symmetrical distribution with slight right skewness. The density curve shows a near-normal distribution with a slight positive skew. The boxplot indicates the presence of 44 outliers, mostly on the upper end of the weight range.

3.3.3 Quantitative Data Analysis: Temperature

The variable *Temperature_C* represents the production temperature in degrees Celsius. This was analyzed using both statistical metrics and visual tools.

Central Tendency

Table 3.3: Measures of Central Tendency for Temperature

Statistic	**Value**
Mean	23.51 °C
Median	23.50 °C
Mode	23.90 °C

The mean and median values are nearly identical, indicating a symmetrical distribution of temperature during the production process.

Spread of Data

Table 3.4: Spread of Temperature Data

Statistic	**Value**
Minimum	20.30 °C
Maximum	26.30 °C
Range	6.00 °C
Variance	0.98
Standard Deviation	0.99 °C
1st Quartile (Q1)	22.88 °C
3rd Quartile (Q3)	24.20 °C
Interquartile Range (IQR)	1.33 °C
Skewness	-0.08
Kurtosis	-0.11

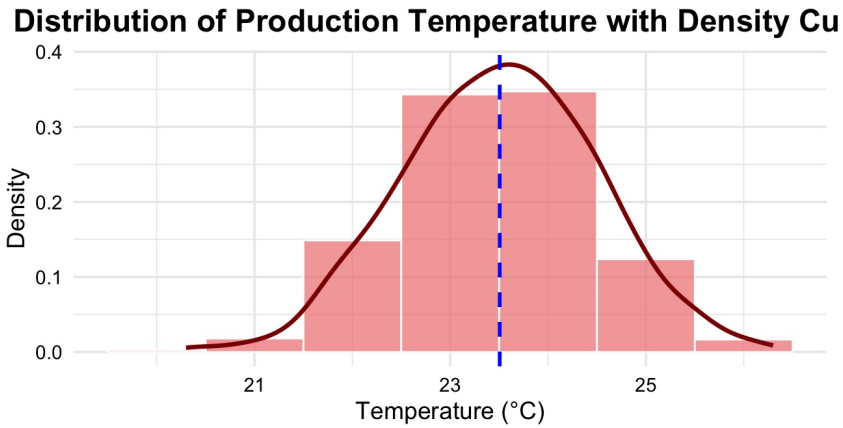

Figure 3.7: Distribution of Production Temperature

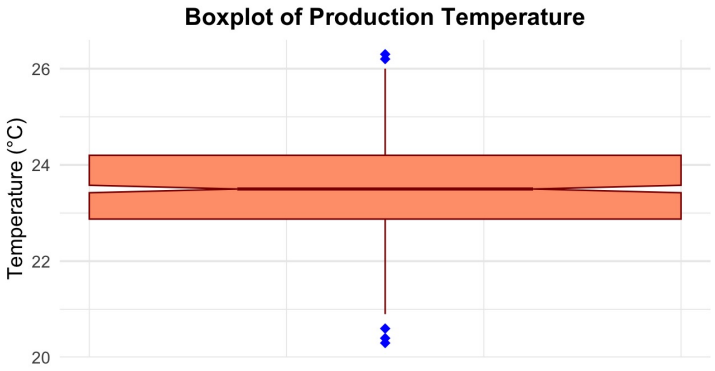

Figure 3.8: Boxplot of Product Weights

The density plot shows a nearly perfect bell-shaped distribution, indicating that the process temperature follows a normal distribution. The boxplot identified 5 outliers, both on the lower and upper ends of the distribution.

3.4 Weight Variation Across Operational Factors

To investigate potential influences on product weight, the *Weight_g* variable was analyzed across key operational factors—Product Type, Shift, Machine ID, and Operator.

3.4.1 Weight Variation by Product

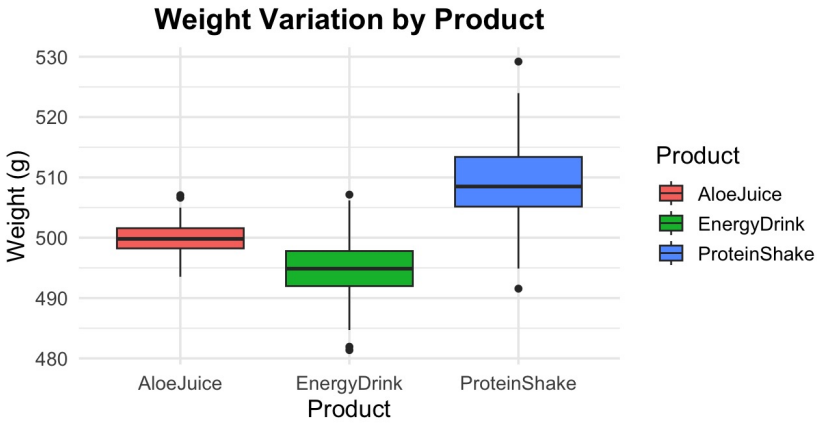

Figure 3.9: Boxplot of Product Weight by Product Type

- **ProteinShake** exhibited the highest median and spread in weight, along with more variability and a higher number of outliers.
- **EnergyDrink** had the lowest median weight and the least spread, indicating a more consistent filling process.
- **AloeJuice** weights were relatively centered but showed moderate dispersion.

3.4.2 Weight Variation by Shift

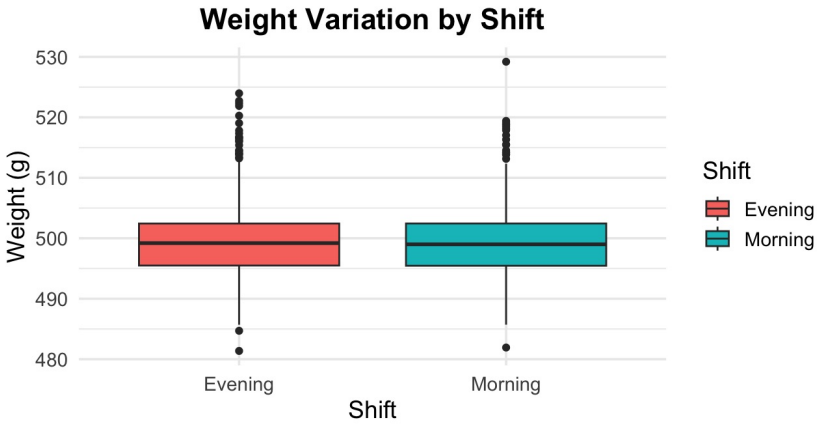

Figure 3.10: Boxplot of Product Weight by Shift

- Both Morning and Evening shifts displayed comparable medians, but the Morning shift had a slightly wider interquartile range (IQR).
- Outliers were present in both shifts, with the Morning shift showing a few high-weight deviations.

3.4.3 Weight Variation by Machine

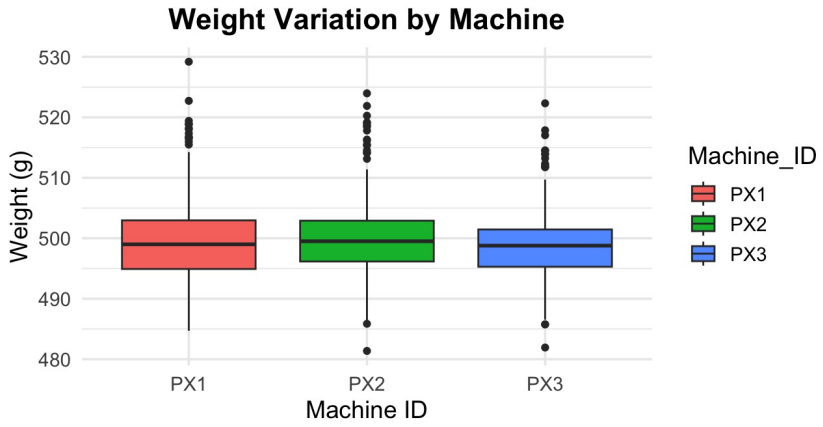

Figure 3.11: Boxplot of Product Weight by Machine

- Machines PX1, PX2, and PX3 showed similar weight distributions, with marginal differences in spread.
- PX3 appeared slightly more centered and consistent, while PX2 exhibited a wider range.

3.4.4 Weight Variation by Operator

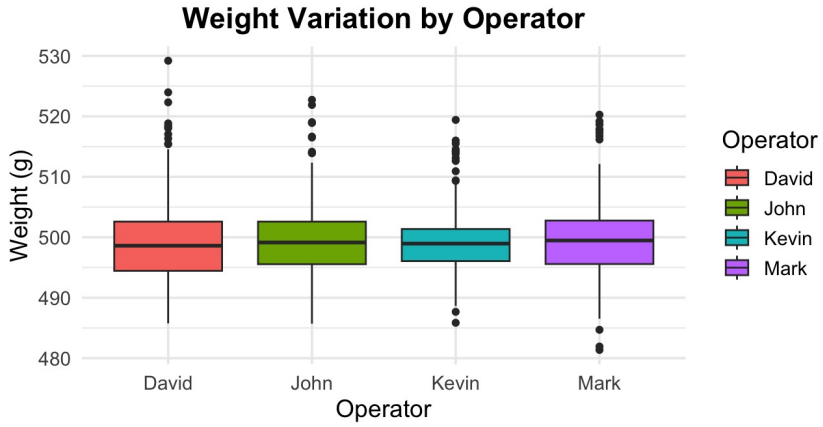

Figure 3.12: Boxplot of Product Weight by Operator

- All four operators (David, John, Kevin, Mark) had similar median weights, though Mark and David showed slightly more variability and upper-end outliers.
- John had the narrowest IQR, suggesting more precise control during operation.

3.5 Inferential Analysis

3.5.1 Weight Variation by Product Type (ANOVA)

To statistically test whether operational factors significantly affect product weight, an ANOVA (Analysis of Variance) was performed using *Weight_g* as the dependent variable.

```
> anova_product <- aov(Weight_g ~ Product, data = df)
> summary(anova_product)
             Df Sum Sq Mean Sq F value Pr(>F)
Product       2  20299   10149   548.6 <2e-16 ***
Residuals   717  13264      18
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3.13: ANOVA: Weight Variation by Product Type

- **F-value:** 548.6
- **p-value:** $< 2e - 16$
- **Result:** Statistically significant ($p < 0.001$)

Interpretation: The mean weight differs significantly across product types. This confirms that product category has a strong influence on weight distribution, supporting the earlier descriptive and visual analysis.

3.5.2 Weight Variation by Machine ID (ANOVA)

```
> anova_machine <- aov(Weight_g ~ Machine_ID, data = df)
> summary(anova_machine)
             Df Sum Sq Mean Sq F value Pr(>F)
Machine_ID    2    243  121.34   2.611 0.0742 .
Residuals   717  33320   46.47
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3.14: ANOVA: Weight Variation by Machine ID

- **F-value:** 2.611
- **p-value:** 0.0742
- **Result:** Not statistically significant at 5% level, but marginally significant at 10% level

Interpretation: There is mild evidence of weight differences across machines, but the result is not strong enough to conclude statistical significance at the 0.05 level. However, the p-value being below 0.10 suggests a potential effect worth monitoring. Further investigation or a larger sample size may be needed to confirm this.

3.5.3 Weight Variation by Shift (ANOVA)

```
> anova_shift <- aov(Weight_g ~ Shift, data = df)
> summary(anova_shift)
             Df Sum Sq Mean Sq F value Pr(>F)
Shift         1      1    0.55   0.012  0.914
Residuals   718  33562   46.74
```

Figure 3.15: ANOVA: Weight Variation by Shift

- **F-value:** 0.012
- **p-value:** 0.914
- **Result:** Not statistically significant

Interpretation: The difference in weight between Morning and Evening shifts is negligible. This supports the earlier visual analysis and confirms no measurable shift-based bias.

3.5.4 Weight Variation by Operator (ANOVA)

```
> anova_operator <- aov(Weight_g ~ Operator, data = df)
> summary(anova_operator)
             Df Sum Sq Mean Sq F value Pr(>F)
Operator      3     12    4.15   0.089  0.966
Residuals   716  33551   46.86
```

Figure 3.16: ANOVA: Weight Variation by Operator

- **F-value:** 0.089
- **p-value:** 0.966
- **Result:** Not statistically significant

Interpretation: Operator-wise variation in weight is minimal and not statistically significant. This indicates consistent performance across all operators.

3.5.5 Correlation and Regression Analysis: Temperature vs. Bottle Weight

```
> cor(df$Temperature_C, df$Weight_g, use = "complete.obs")  # Pearson by default
[1] 0.02941742
```

Figure 3.17: Correlation coefficient

Correlation Coefficient (r): 0.029

Interpretation: The correlation is nearly zero, indicating an extremely weak linear relationship between temperature and bottle weight.

```
> model <- lm(Weight_g ~ Temperature_C, data = df)
> summary(model)

Call:
lm(formula = Weight_g ~ Temperature_C, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-18.3719  -4.2346  -0.6499   2.6209  29.2961

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   494.9825     6.0422  81.921   <2e-16 ***
Temperature_C   0.2025     0.2568   0.789    0.431
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.834 on 718 degrees of freedom
Multiple R-squared:  0.0008654, Adjusted R-squared:  -0.0005262
F-statistic: 0.6219 on 1 and 718 DF,  p-value: 0.4306
```

Figure 3.18: model Summary

Regression Equation:

$$\text{Weight\_g} = 494.98 + 0.2025 \times \text{Temperature\_C}$$

- **R-squared:** 0.000865 — less than 0.1%, indicating almost no variation in weight is explained by temperature.
- **p-value for Temperature:** 0.431 — not statistically significant.

Interpretation: The coefficient for temperature (0.2025) implies a very slight increase in weight per 1°C rise in temperature, but this effect is statistically insignificant and practically negligible.

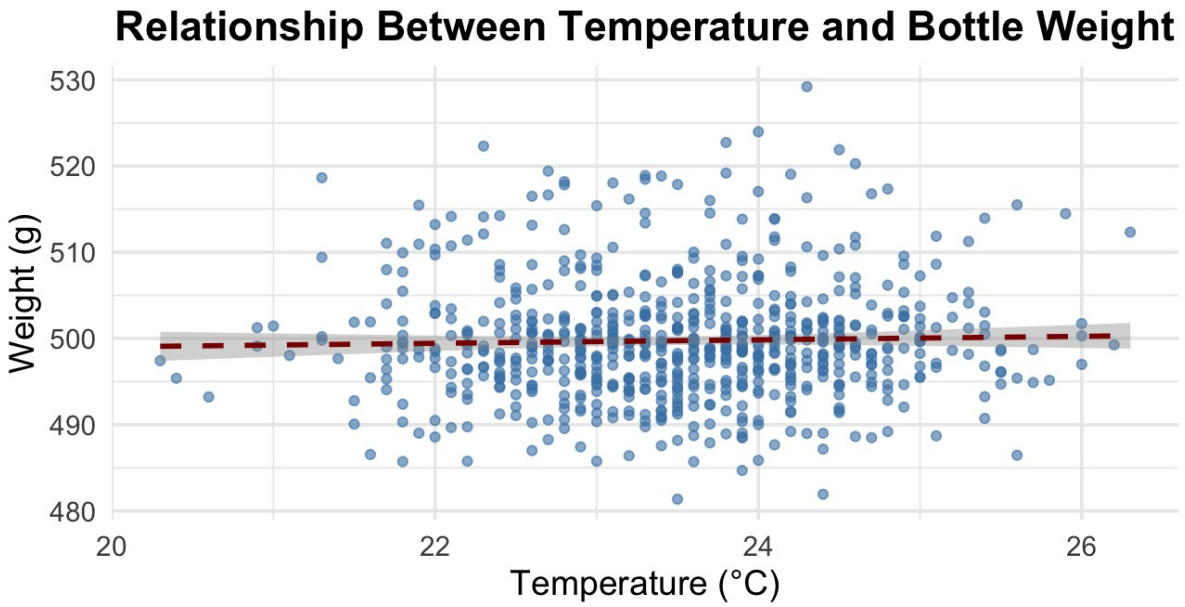

Figure 3.19: Scatter Plot with Regression Line: Temperature vs. Weight

3.5.6 Control Chart Analysis: Individuals Chart for Bottle Weight

To monitor the consistency of the production process and detect unusual variation in bottle weight, an Individuals Control Chart (I-Chart) was constructed using *Weight_g* as the quality characteristic. Chart Details:

- Number of Observations (groups): 720
- Center Line (CL): 499.74 g
- Standard Deviation (σ): 6.62
- Lower Control Limit (LCL): 479.89 g
- Upper Control Limit (UCL): 519.60 g

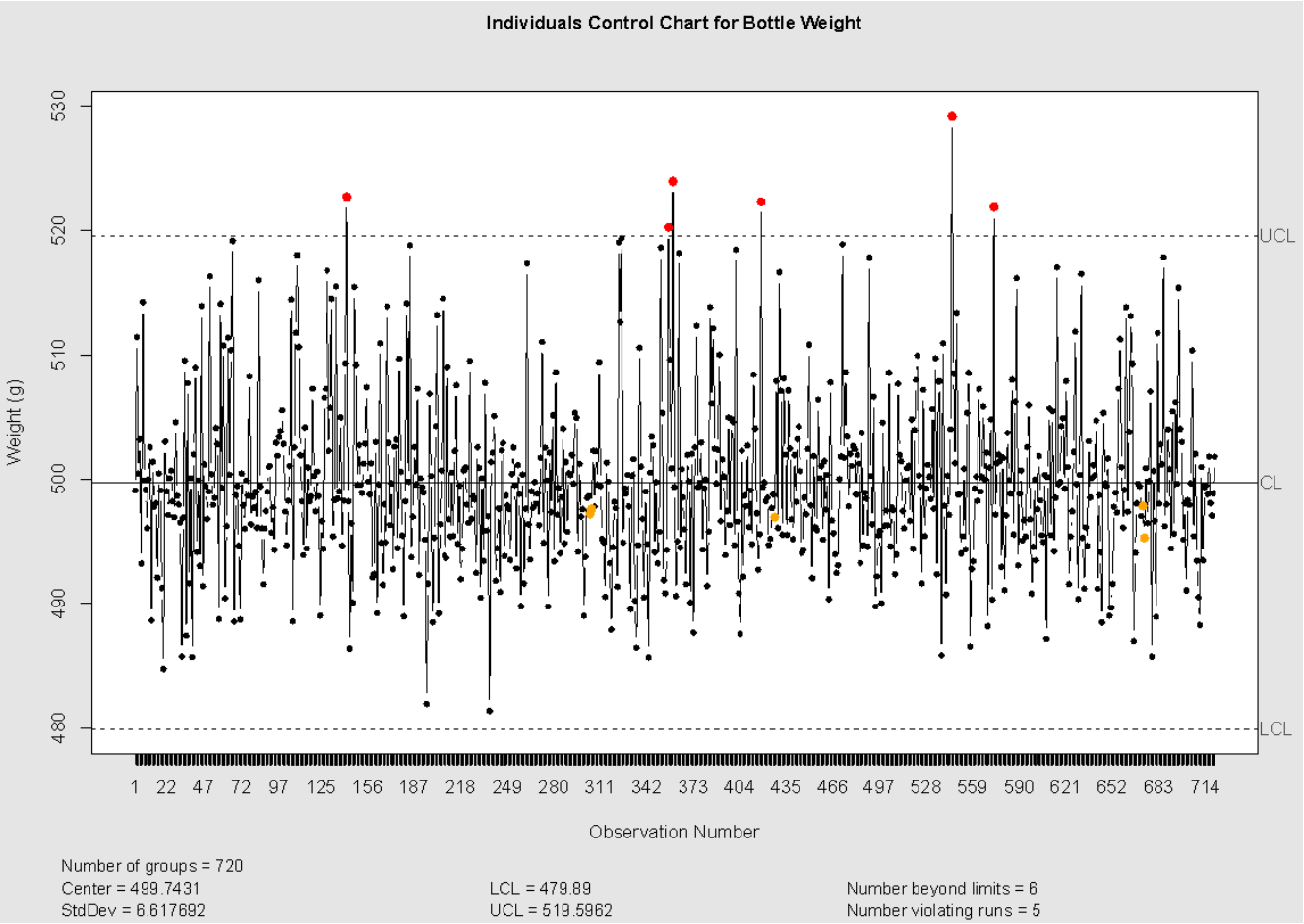

Figure 3.20: Individuals Control Chart for Bottle Weight

3.5.7 Findings

Six points were found beyond control limits, indicated by red dots, suggesting the presence of special cause variation. Five sequences violated control run rules (e.g., consecutive points on one side of the mean), marked by orange dots, indicating potential process shifts or trends. Most points fall within control limits, indicating that the overall process is statistically in control, but occasional deviations suggest moments of instability or anomalies.

3.5.8 Interpretation

The I-Chart reveals that:

- While the central process mean is well-defined and stable (near 499.74 g),
- There are a few exceptional points that exceed control boundaries or indicate non-random patterns,

- These out-of-control points should be investigated further—potential causes may include machine errors, temperature spikes, or operator inconsistencies.

3.5.9 Summary of Data Analysis

The dataset of 720 observations was analyzed using RStudio. No missing values were found, and data distribution across shifts, machines, and operators was balanced.

- Product weight averaged 499.74 g, with a nearly normal distribution and some outliers.
- Temperature was stable (mean: 23.51°C) and showed no significant impact on weight.
- ANOVA revealed that product type significantly affects weight ($p < 0.001$), while shift, operator, and machine do not.
- Regression and correlation showed no meaningful relationship between temperature and weight ($R^2 < 0.1\%, \quad p = 0.431$)
- The Individuals Control Chart showed the process is mostly in control, with 6 outliers and 5 run rule violations suggesting occasional special cause variations.

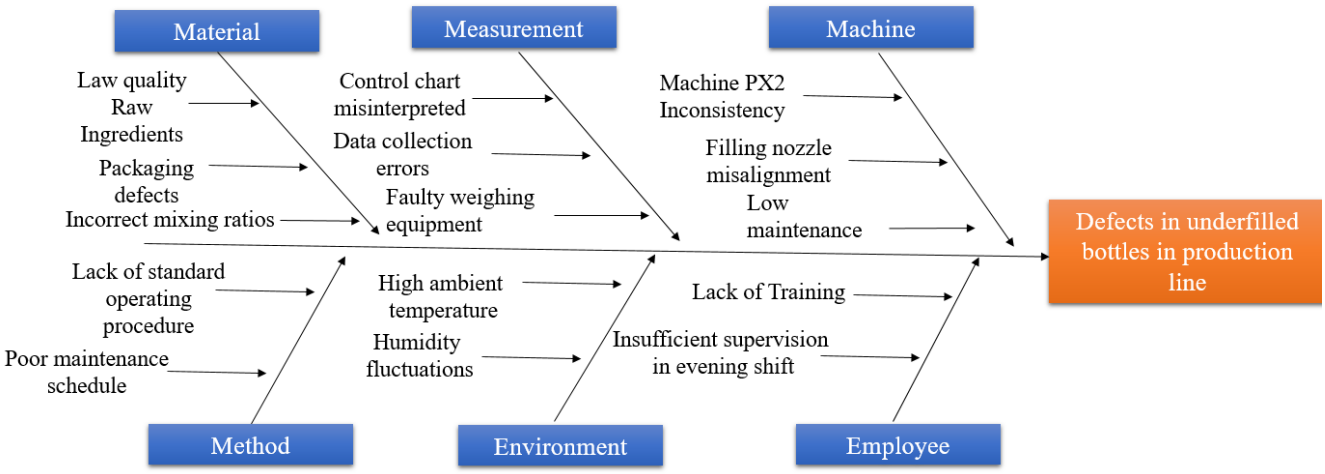

Figure 3.21: Fishbone Diagram

4 Discussion and Conclusion

The analysis presented in our project effectively addressed the growing issue of product defects in underfilled bottles in the Energy Drink and Protein Shake lines. By applying the Lean Six Sigma methodology, particularly the DMAIC framework, we were able to systematically define the problem, measure process performance, analyze root causes, implement improvements, and establish controls.

Here, the population was 10,000 and a sample of 764 was obtained using a stratified sampling technique according to the characteristics which can affect the weight such as product, shift, machine, operator, and temperature. In the data analysis phase, not only were the Energy Drink and Protein Shake products analyzed, but Aloe Juice was also included to ensure its average weight was under statistical control in comparison.

The analysis showed that Energy Drink has a lower average weight (<500g), whereas Protein Shake has a higher average weight (>500g). The ANOVA analysis revealed that product type and machine had statistically significant effects on the mean bottle weight, indicating that the type of beverage and the specific machine used—especially machine PX2—are key contributors to weight inconsistencies. The effects of shift and operator, while present, showed comparatively lower statistical impact but still warrant attention, particularly given the high number of defects during the evening shift.

This suggests that although human factors play a role, equipment and product-specific calibration and maintenance are more urgent priorities for quality control. Further root cause analysis via the Fishbone diagram uncovered key contributing factors, including poor maintenance, operator training gaps, and environmental influences such as temperature and humidity.

By analyzing the data, it was found that there is no correlation between weight and temperature, meaning temperature does not have a significant effect on bottle weight. Since weight is a variable data type and the sample size is one, an Individuals and Moving Range ($\bar{x}$-MR) control chart was constructed. The presence of outliers in the chart indicates that the process is not statistically controlled.

To overcome the problem, several practical implications of our findings were implemented. These included hourly machine checks, retraining of operators using visual aids, and the introduction of standard operating procedures (SOPs) for sealing and cleaning processes. These actions led to measurable improvements in the production process.

Additionally, the integration of manual checks and plans for sensor-based feedback loops enhanced the robustness of the quality assurance system. These interventions improved compliance with the 500g $\pm$ 5g requirement and also enhanced overall production efficiency while reducing waste—addressing multiple types of the "Seven Wastes" identified in Lean methodology.

Despite these successes, challenges were encountered. Data limitations, such as incomplete records on machine downtime and raw material inconsistencies, initially hindered comprehensive analysis. However, these limitations were mitigated by focusing on available defect and process performance data and using statistical tools to draw meaningful conclusions. The inconsistency in PX2 performance also posed a challenge, which was addressed through targeted maintenance and operator retraining efforts.

In conclusion, our project demonstrated that data-driven techniques and continuous improvement strategies can lead to tangible enhancements in product quality and process control. Future improvements may include expanding the dataset, automating quality checks, and incorporating predictive analytics to proactively manage process variability.

5 Recommendations

The analysis of the beverage production process using Lean Six Sigma methodologies revealed several critical insights. The overall defect rate was identified as 1.3The application of Lean Six Sigma tools, such as DMAIC, control charts, and exploratory data analysis, effectively addressed quality issues in the manufacturing process. By systematically identifying root causes and implementing targeted solutions—such as hourly machine checks, retraining of operators, and introduction of SOPs—the process achieved improved stability and reduced variation. While the corrective actions led to a measurable improvement in product consistency and defect reduction, the absence of comprehensive data on machine downtime and raw material characteristics posed limitations on deeper analysis. Nevertheless, the study demonstrated that data-driven approaches can yield tangible improvements in manufacturing quality and efficiency. To build on the gains achieved, future improvements should include expanded data logging, particularly on machine maintenance records, raw material properties, and environmental conditions. The integration of IoT-enabled monitoring systems and real-time analytics can provide early warnings for deviations and support predictive maintenance. Additionally, automating quality checks through sensor-based systems or computer vision can further reduce human error. Operator performance should be reviewed regularly through structured audits and feedback-based retraining programs. Future research could also explore upstream processes like mixing and formulation, as well as downstream packaging, to holistically optimize production quality across the value chain.

6 References

1. Kaggle. (2023). *Beverage Manufacturing Process Dataset*. Retrieved from `https://www.kaggle.com`
   — Source of dataset used in the study.

2. ASQ – American Society for Quality. (n.d.). *Control Charts*. Retrieved July 4, 2025, from `https://asq.org/quality-resources/control-chart`
   — Comprehensive guide to control chart types and applications.

3. Six Sigma Daily. (n.d.). *What is DMAIC?*. Retrieved July 4, 2025, from `https://www.sixsigmadaily.com/what-is-dmaic/`
   — Overview of Lean Six Sigma's DMAIC methodology.

4. iSixSigma. (n.d.). *Understanding the Fishbone Diagram*. Retrieved from `https://www.isixsigma.com`
   — Explanation of root cause analysis using the Ishikawa (Fishbone) diagram.

5. RDocumentation. (n.d.). *qcc package – Quality Control Charts*. Retrieved from `https://www.rdocumentation.org/packages/qcc`
   — Technical documentation for creating control charts in R.

6. DataCamp. (n.d.). *How to Perform Linear Regression in R*. Retrieved from `https://www.datacamp.com/tutorial/linear-regression-R`
   — Practical tutorial on performing and interpreting linear regression in R.