Reduce Rework Percentage in Content Formatting

Santhanam

ROADMAP

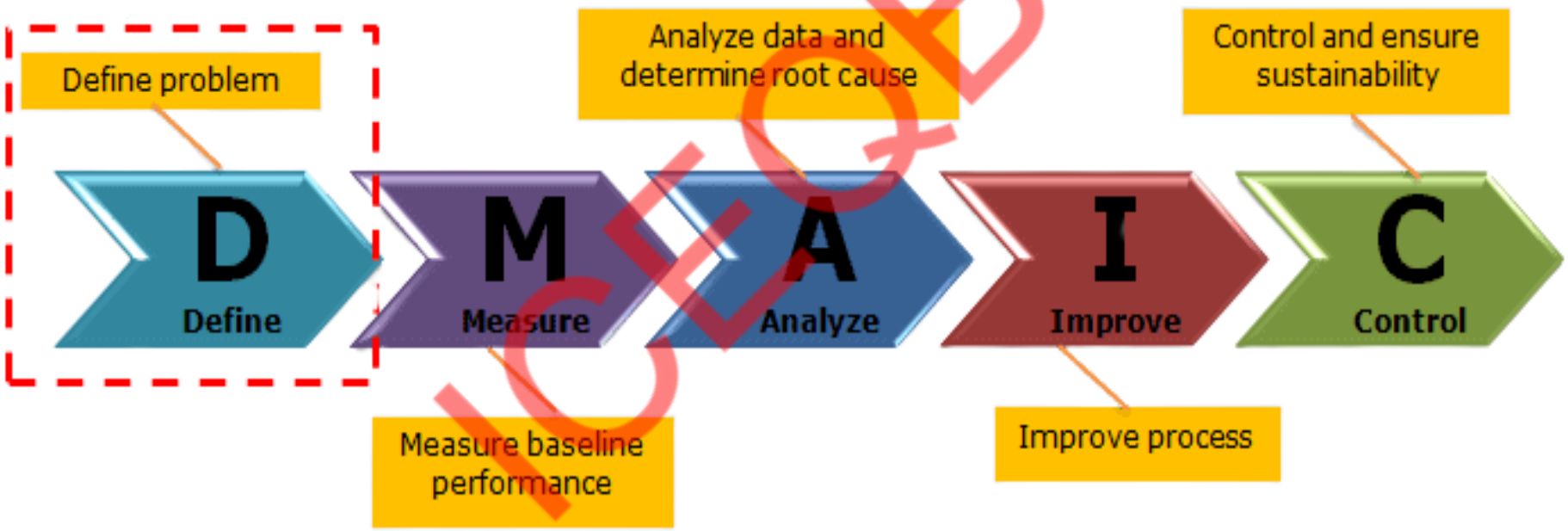
OVERVIEW

Background

The Content Formatting process handles high-volume deliveries where consistency, speed, and quality are critical. Over the past 9 months, the process has shown an average rework rate of 19% (ranging from 13% to 25%), indicating instability and lack of standardization. High rework results in additional QC cycles, increased turnaround time (TAT), operator fatigue, and reduced customer satisfaction.

Currently, 48,000 files are processed monthly, with approximately 9,120 files requiring rework. Each rework consumes an average of 3 minutes of QC effort, leading to nearly 456 extra QC hours per month and an estimated cost impact of ₹4.2 lakhs per month. Reducing rework through a structured Lean Six Sigma approach will stabilize the process, improve first-pass yield, reduce TAT and QC effort, and deliver sustainable cost savings without impacting delivery quality.

DEFINE PHASE

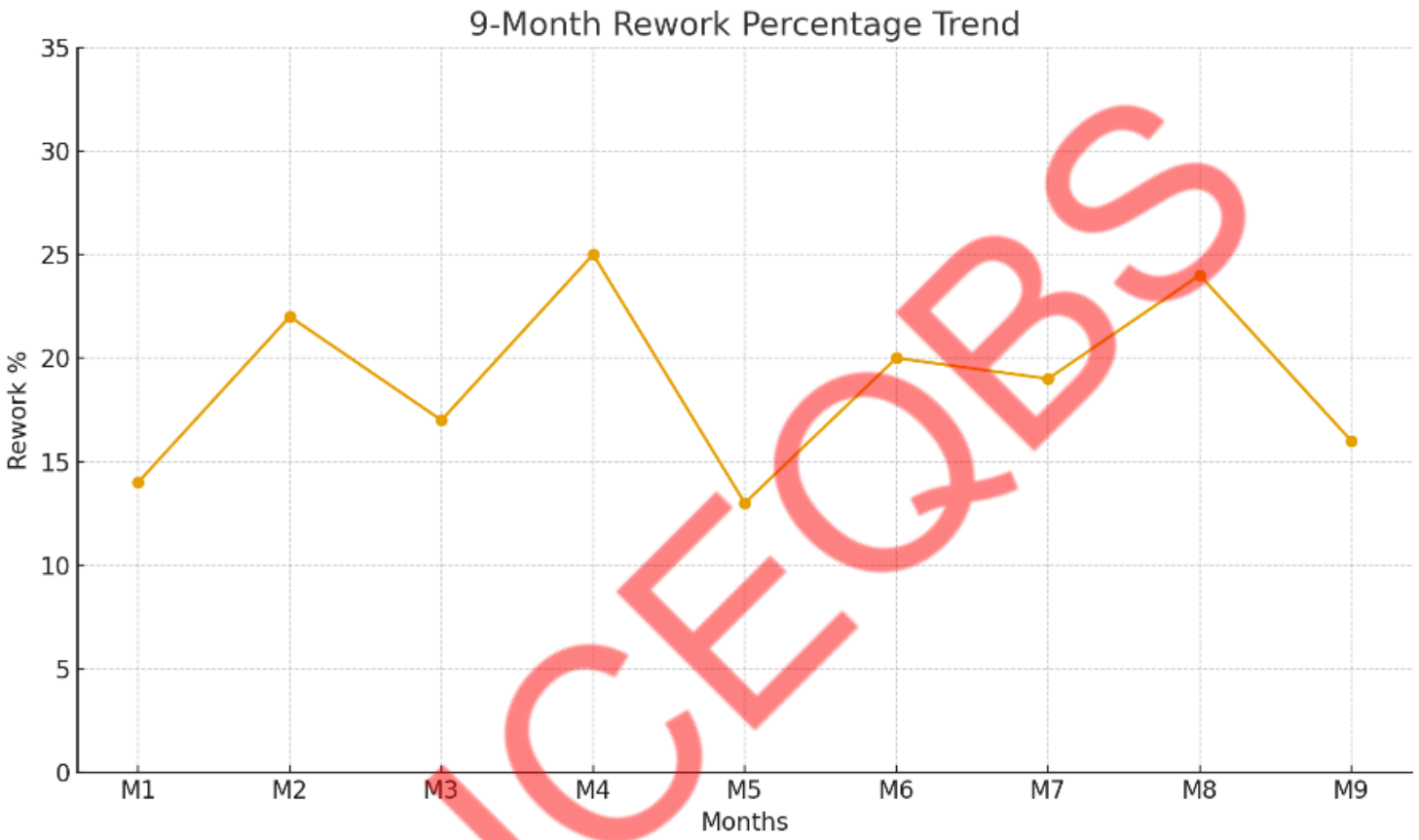


VOC & CTQ

CTQ Tree :

Voice of customer	Critical to X	Primary Metric for improvement
We expect content files to be formatted correctly the first time, with accurate metadata and consistent structure across batches, minimal QC corrections, no repeated errors, and on-time delivery as per SLA to ensure smooth operations and faster turnaround time.	CTQ – Accuracy	**Primary Metric -** Y = 1. Accuracy of article segmentation **Secondary Metric -** **First Pass Yield (FPY)**

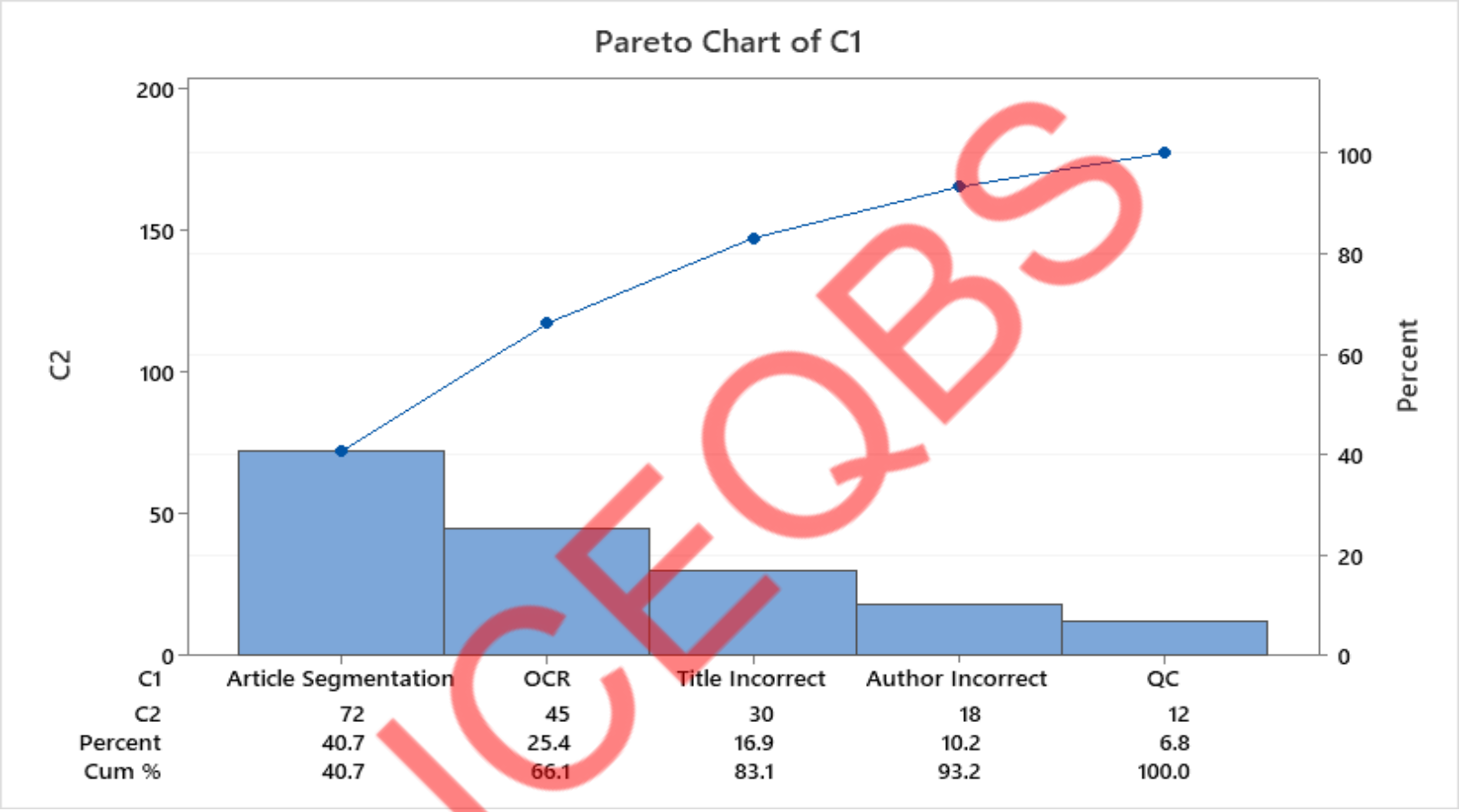
Baseline Performance of Primary Metric (9 months data as Line chart)



Inference :

- Last 9 months data shows a significant variation and hence ideal problem to be taken up as a Six Sigma Project.

Pareto chart



Inference :

- Article segmentation contributes substantially and included in the scope of the project

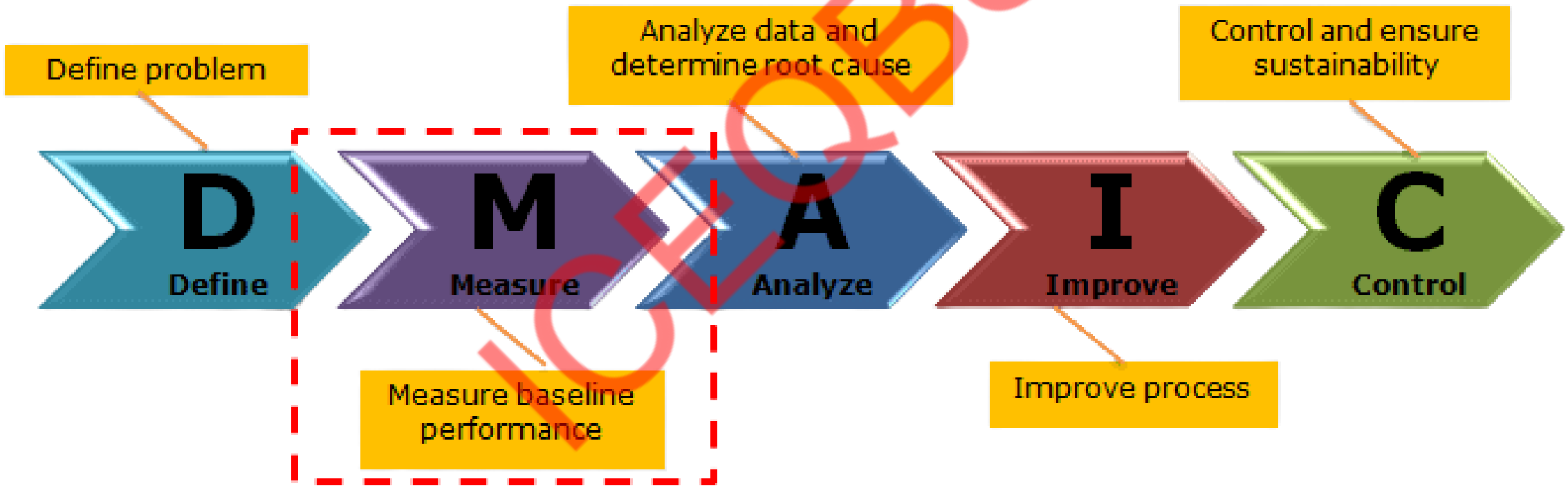
Project Charter

Project Title:	Reduction of Scrap% in Machining process from 3% to
Project Leader	**Project Team Members:**
Santhaam	Sr. QC Analyst – Kavya Sr. Formatter – Arjun MIS Analyst – Varun
Champion/Sponsors:	**Key Stake Holders**
Priya Natarajan (Operations Head)	QC Team Production Team Client QA Team Client Editorial Team
Problem Statement:	**Goal Statement:**
average rework percentage in the Content Formatting process has been 19%, with monthly variation ranging from 13% to 25%,for the last 9 months	To reduce the Rework Percentage from the current baseline of 19% to below 10% within 12 weeks,
Secondary Metric	**Assumptions Made:**
First Pass Yield (FPY)	Stable volume and content complexity No major tool or guideline changes

Project Charter

Tangible and Intangible Benefits:	Risk to Success:
Estimated saving = • ₹4.2 lakhs/month **Other benefits –** • Customer Satisfaction • Accuracy on delivery time	Client guideline changes Non-adherence to standards OCR input quality issues
In Scope:	**Out of Scope:**
Article segmentation accuracy Title formatting and structure Author identification and mapping OCR post-processing checks	Content creation and client edits Tool replacement or system upgrades Downstream publishing activities
Signatories:	**Project Timeline:**
Project Sponsor **Process Owner**	6 months

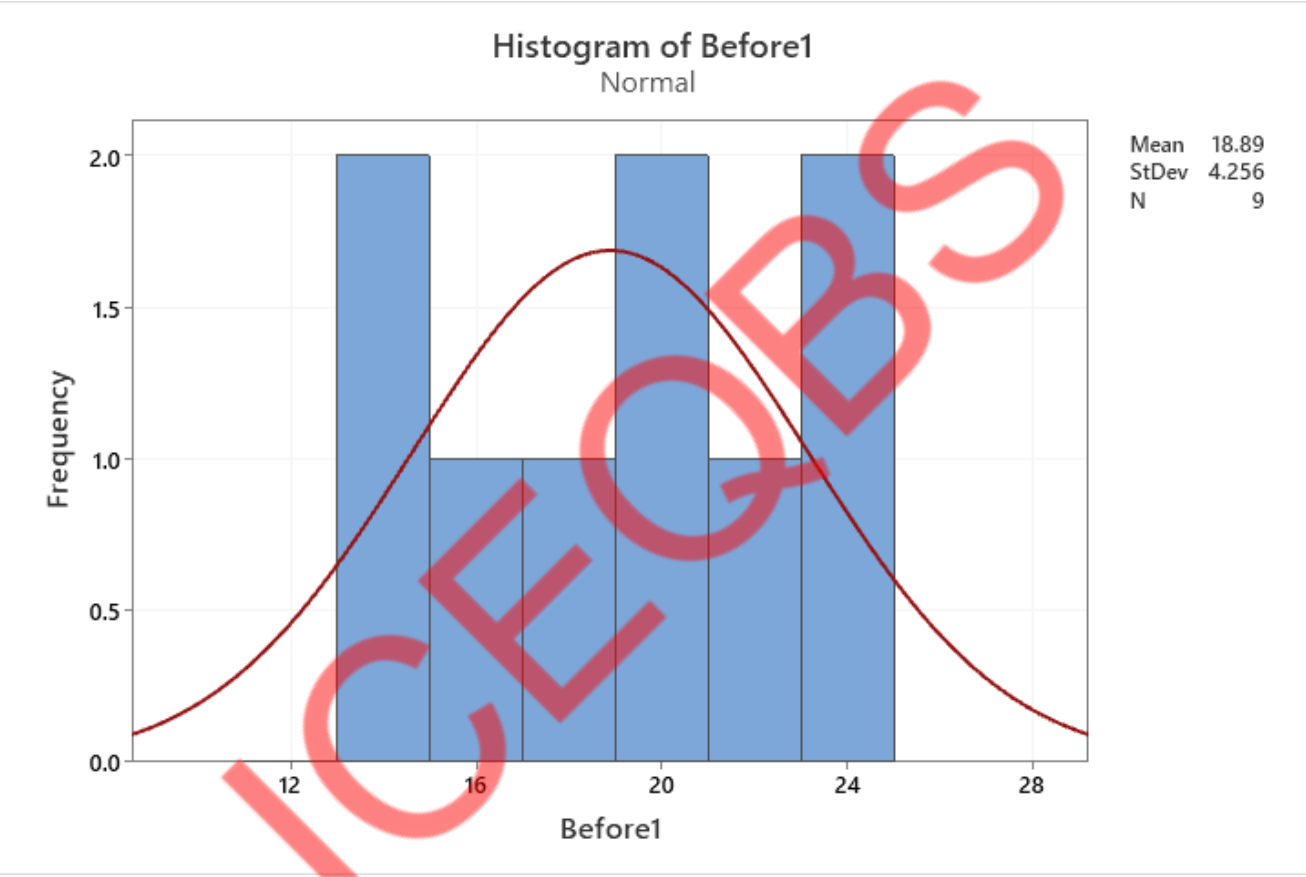
MEASURE PHASE



SIPOC

S – Suppliers	I – Inputs	P – Process Steps	O – Outputs	C – Customers
Client (Content Source Team)	Raw content files (PDF/Word/XML)	Receive input files	Formatted content file	Client QA Team
OCR Tool / OCR Engine	OCR output text	Run OCR & extract text	QC-approved document	Client Editorial Team
Style Guide Owner	Client style guide	Content segmentation	Updated metadata sheet	Client CMS Team
Production Team	Metadata sheets	Format titles & subtitles	Delivery package	External Publishing Team
QC Team	SOP & formatting rules	Author identification	QC feedback summary	Internal QC Team
MIS / Reporting Team	Reference templates	Metadata tagging	Productivity metrics	Production Team
	QC guidelines	Formatting cleanup	Error trend report	Operations Management
		QC review		
		Final packaging & delivery		

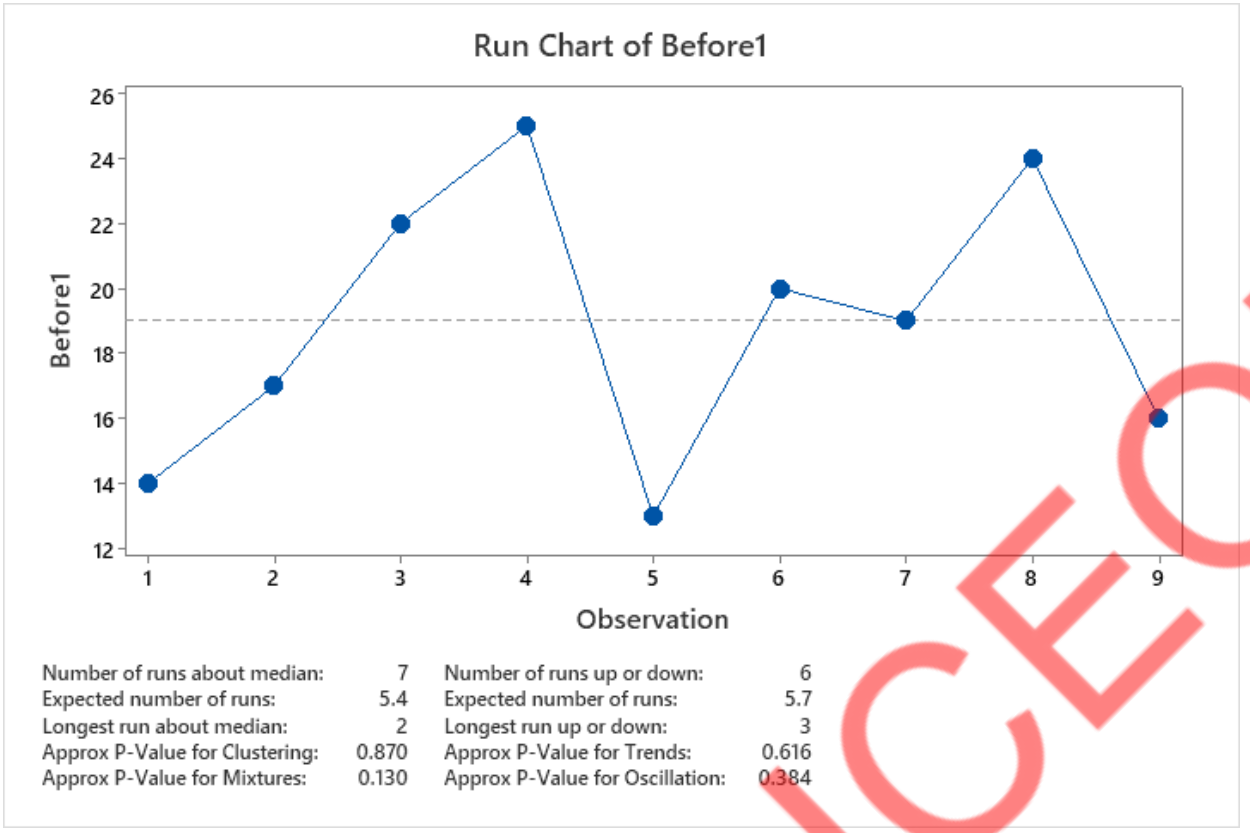
Data collection – Histogram (Before improvement)



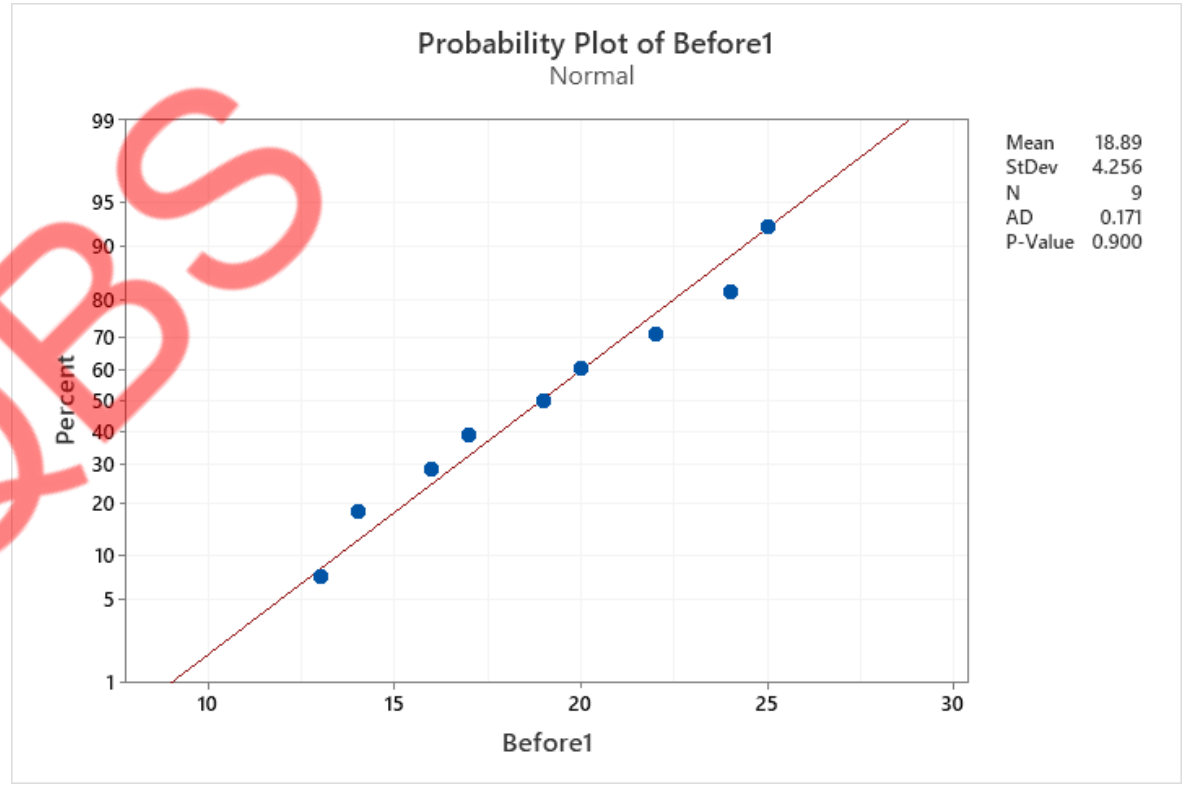
Inference :

- Data is normally distributed over the mean

Data collection – Run Chart (Before improvement)



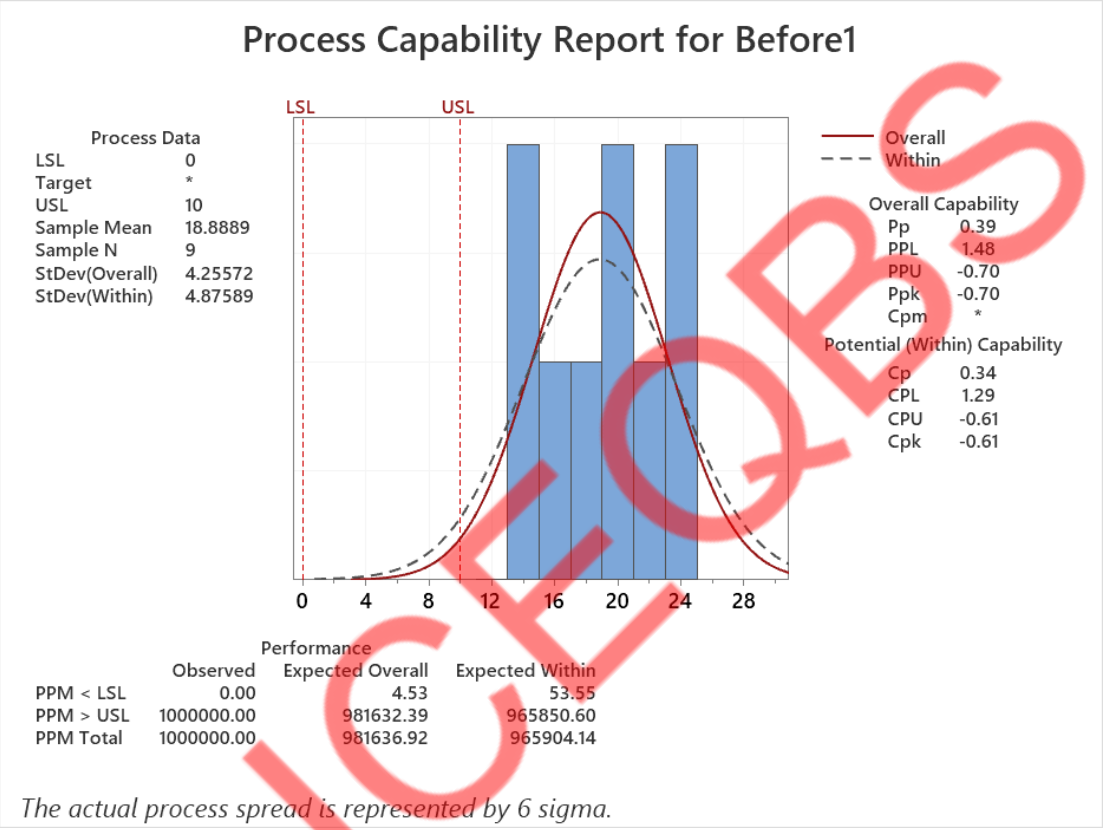
Number of runs about median:	7	Number of runs up or down:	6
Expected number of runs:	5.4	Expected number of runs:	5.7
Longest run about median:	2	Longest run up or down:	3
Approx P-Value for Clustering:	0.870	Approx P-Value for Trends:	0.616
Approx P-Value for Mixtures:	0.130	Approx P-Value for Oscillation:	0.384



Inference :

P > 0.05 – No special causes in the process. Data can be used for further analysis

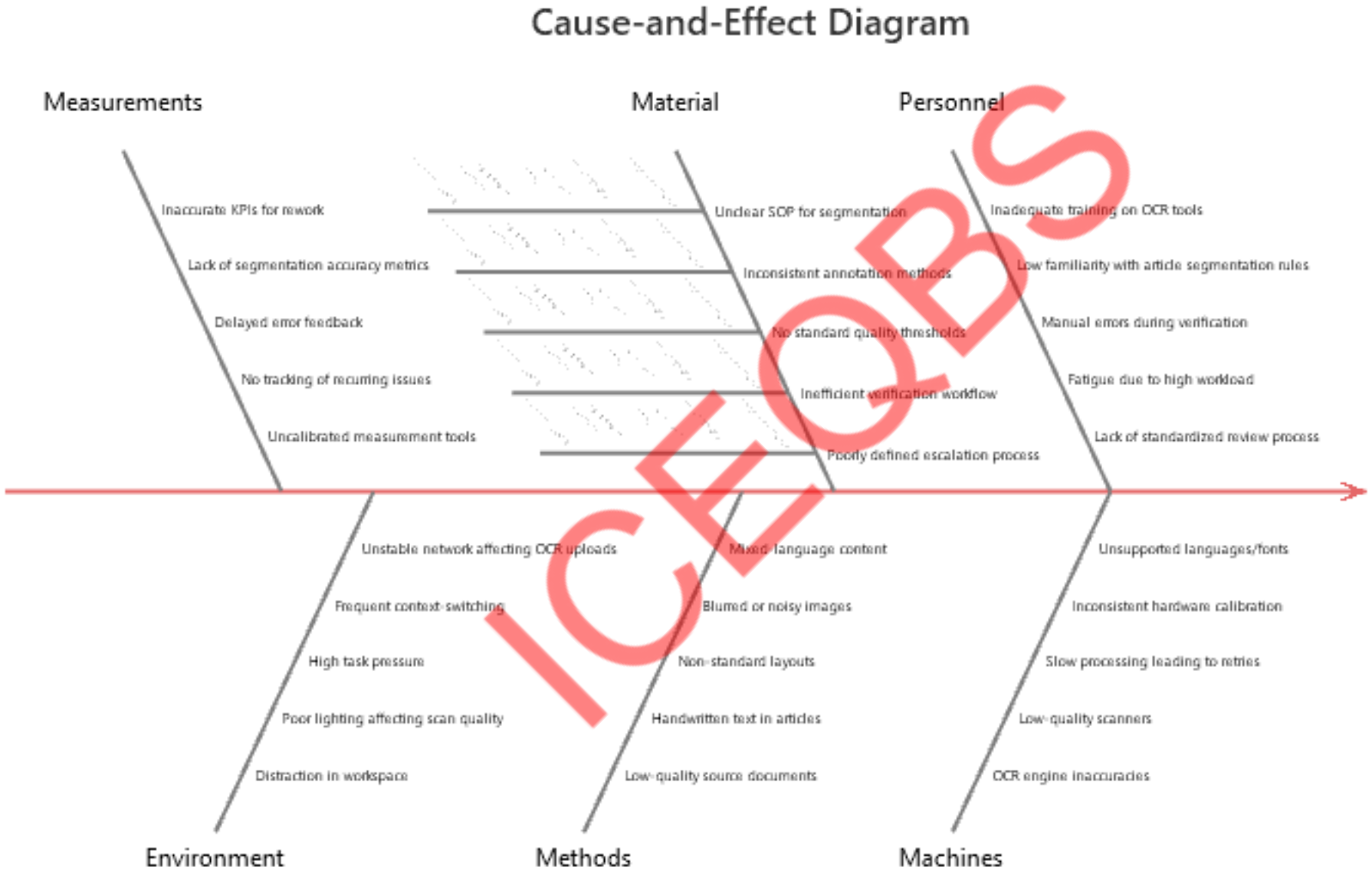
Data collection(Before improvement)



Inference :

- P > 0.05 in all scenarios, thus all the data is normally distributed

Fish Bone Diagram



Common and special causes

Common Causes	Special Causes
Insufficient training	Sudden absence of key staff
Manual fatigue	Unexpected system crash
Lack of standardized work	OCR engine failure
Low experience levels	Emergency process change
Inconsistent review practices	Unexpected file corruption
Low OCR accuracy	Power outage or system downtime
Scanner calibration drift	
Slow processing speed	
Software version mismatch	
System lag	
No standard segmentation guidelines	
Inconsistent annotation method	
Multiple workflow variations	
Complex steps	
Ineffective review procedure	
Inconsistent quality checks	
Manual inspection errors	
No clear accuracy metric	

3M Analysis for Waste

Muda (Waste)	Mura (Unevenness)	Muri (Overburden)
• Rework due to incorrect segmentation or OCR errors • Waiting time when OCR processing is slow or system lags • Overprocessing by performing multiple rounds of manual checking for the same article	• Inconsistent QC feedback leading to variances in how segmentation is done • Variation in document complexity (simple vs. multi-column, tables) causing irregular workload • Different operators producing different quality levels due to variable skill levels	• Operators handling excessive document volume, leading to fatigue and more errors • Expecting staff to manually correct high OCR error rates without automation support • Rushed timelines forcing employees to skip steps or multitask excessively

8 Wastes Analysis

Type of Waste	Example 1	Example 2
Transportation (Unnecessary movement of digital items)	Sending files to multiple reviewers when one is sufficient	Moving documents between multiple shared drives due to unclear storage structure
Inventory (Work piling up)	Queue of OCR outputs waiting for manual verification	Excess unprocessed client inputs stored for future cycles
Motion (Unnecessary movement by people)	Operators repeatedly scrolling back and forth to locate article boundaries	Manually toggling between PDF viewer, OCR tool, segmentation tool
Waiting (Idle time)	Waiting for slow file loading in the segmentation tool	Waiting for clarification from team leads on ambiguous article structures
Overproduction	Producing segmented articles even before client approval of sample output	Running OCR on files that do not require text extraction
Overprocessing	Manually recreating article structures even when automated scripts exist	Adding extra metadata tags not required by the client
Defects	Incorrectly tagging article titles or authors	Missing segments due to OCR dropout zones
Skills (Underutilized talent)	Highly skilled staff doing routine renaming or copy-paste tasks	Not leveraging employee expertise to design templates or automation scripts

Action Plan for Low Hanging Fruits

Gemba Observation (Special Cause)	Lean Tool Used	Action Plan (Low Hanging Fruit)	Expected Benefit
Sudden OCR tool failure	Standard Work	Define restart & fallback SOP	Reduced downtime
Corrupted input files	Poka-Yoke	Input file validation check	Reduced rework
Network/system outage	Visual Management	Daily system health check	Faster escalation

Action Plan for Low Hanging Fruits

Muda (Waste)

Gemba Observation (Special Cause)	Lean Tool Used	Action Plan (Low Hanging Fruit)	Expected Benefit
Sudden OCR tool failure	Standard Work	Define restart & fallback SOP	Reduced downtime
Corrupted input files	Poka-Yoke	Input file validation check	Reduced rework
Network/system outage	Visual Management	Daily system health check	Faster escalation

Mura (Unevenness)

Observed Variation	Lean Tool	Action	Benefit
Different quality by operator	Standardization	Common SOP & samples	Consistent output
Uneven workload across team	Heijunka	Balanced task allocation	Stable flow
QC interpretation differs	Calibration	Weekly QC alignment	Reduced variation

Muri (Overburden)

Observed Overburden	Lean Tool	Action	Benefit
High document load per person	Line Balancing	Redistribute work	Lower fatigue
Manual OCR corrections	Automation	OCR confidence threshold	Reduced strain
Tight delivery pressure	Kaizen	Realistic TAT norms	Sustainable pace

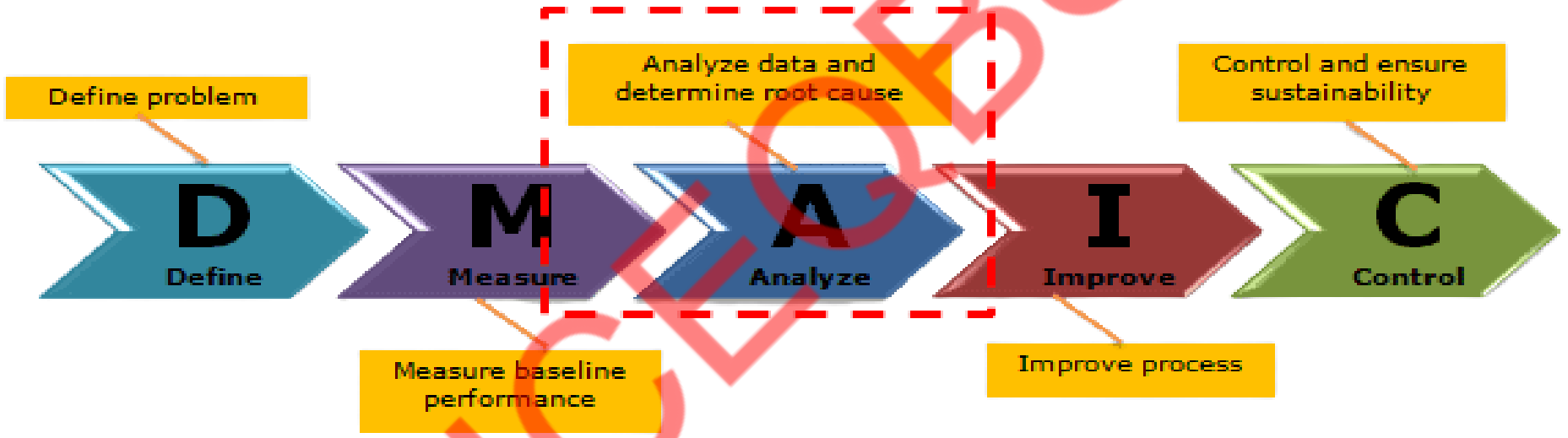
Top Prioritized Root Causes (Based on Net Score)

Input (X – Causes)	Rework Percentage (9)	First Pass Quality (8)	On-Time Delivery (7)	OCR Accuracy (7)	Segmentation Compliance to SOP (9)	Net Score
OCR tool accuracy limitations	9	9	3	9	3	**264**
Insufficient segmentation training	9	9	3	1	9	**262**
High manual intervention	9	9	3	1	9	**262**
Incomplete segmentation SOP	9	9	3	0	9	**255**
QC checklist not followed	9	9	1	0	9	**241**
Complex document formatting	9	3	1	3	9	**214**
Inconsistent metadata rules	9	3	1	1	9	**200**
Poor scan resolution	9	3	1	9	1	**184**

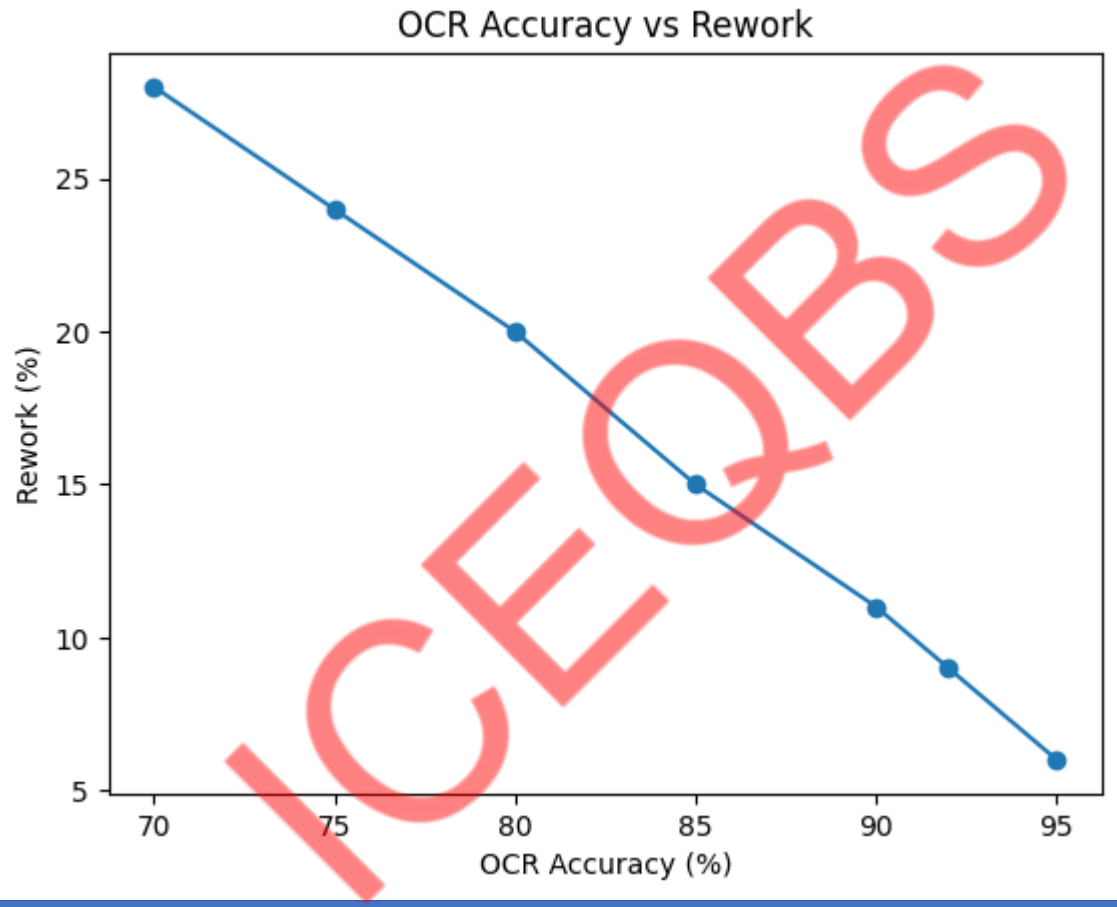
Data Collection Plan

Root Cause (X)	Data to be Collected	Data Type	Source	Sample Size / Period	Collection Frequency	Responsible
OCR tool accuracy limitations	OCR error rate	Continuous	OCR logs	30 files / week	Weekly	IT / QA
Insufficient segmentation training	Training coverage	Discrete	Training records	All operators	Monthly	HR / Ops
High manual intervention	Manual correction count	Continuous	CMS logs	25 files/day	Daily	Production Lead
Incomplete segmentation SOP	SOP gaps	Discrete	SOP review	1 SOP review	One-time	Quality Lead
QC checklist not followed	Checklist adherence	Discrete	QC audit	20 files/day	Daily	QC Supervisor
Complex document formatting	Complexity index	Discrete	Input files	30 files/week	Weekly	Production
Inconsistent metadata rules	Metadata error count	Continuous	QC reports	20 files/day	Daily	Metadata Lead
Poor scan resolution	DPI level	Continuous	Scan properties	30 files/week	Weekly	Client Ops
Excessive workload per shift	Files per FTE	Continuous	MIS reports	All shifts	Daily	Ops Manager
OCR system downtime	Downtime minutes	Continuous	IT logs	Entire period	Daily	IT Support

ANALYSE PHASE

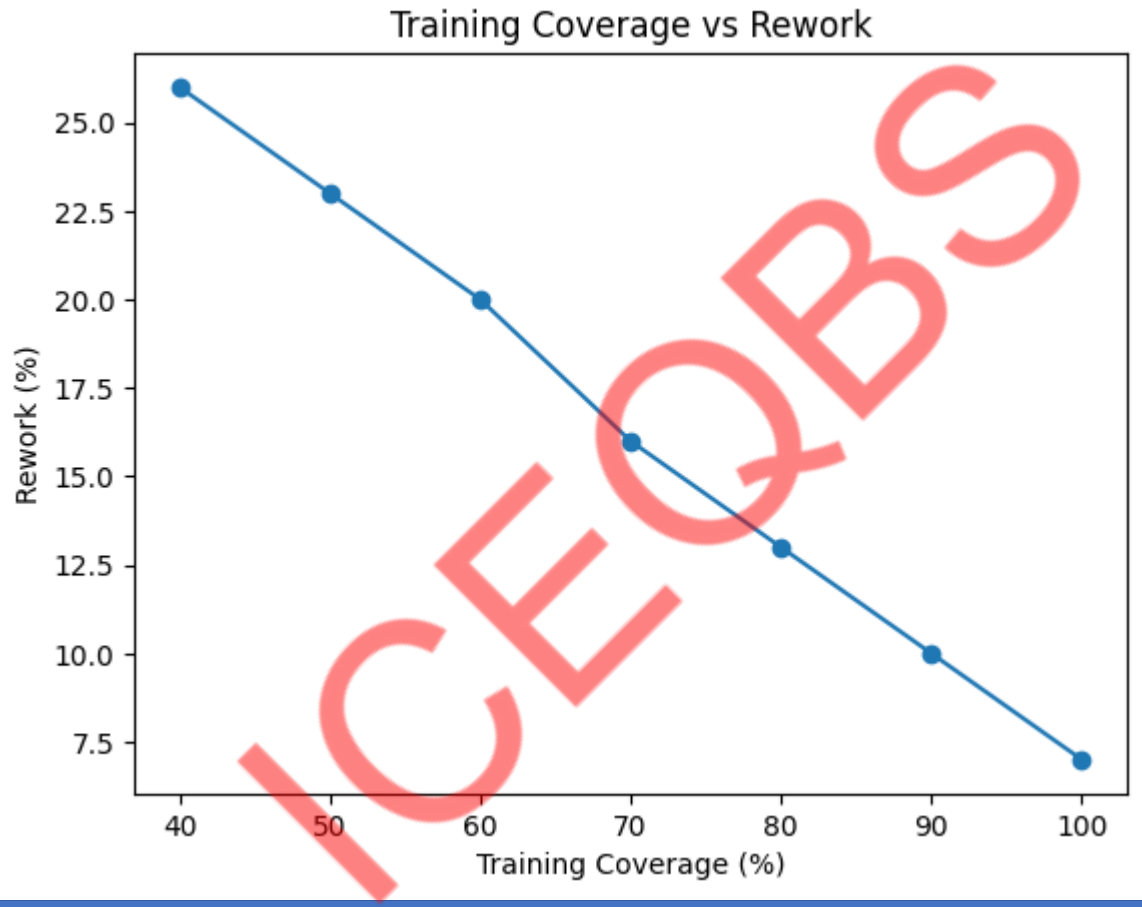


Analyse – Hypothesis testing



Inference :
Higher OCR accuracy leads to significantly lower rework, showing a **strong negative correlation (r ≈ –0.90)**.

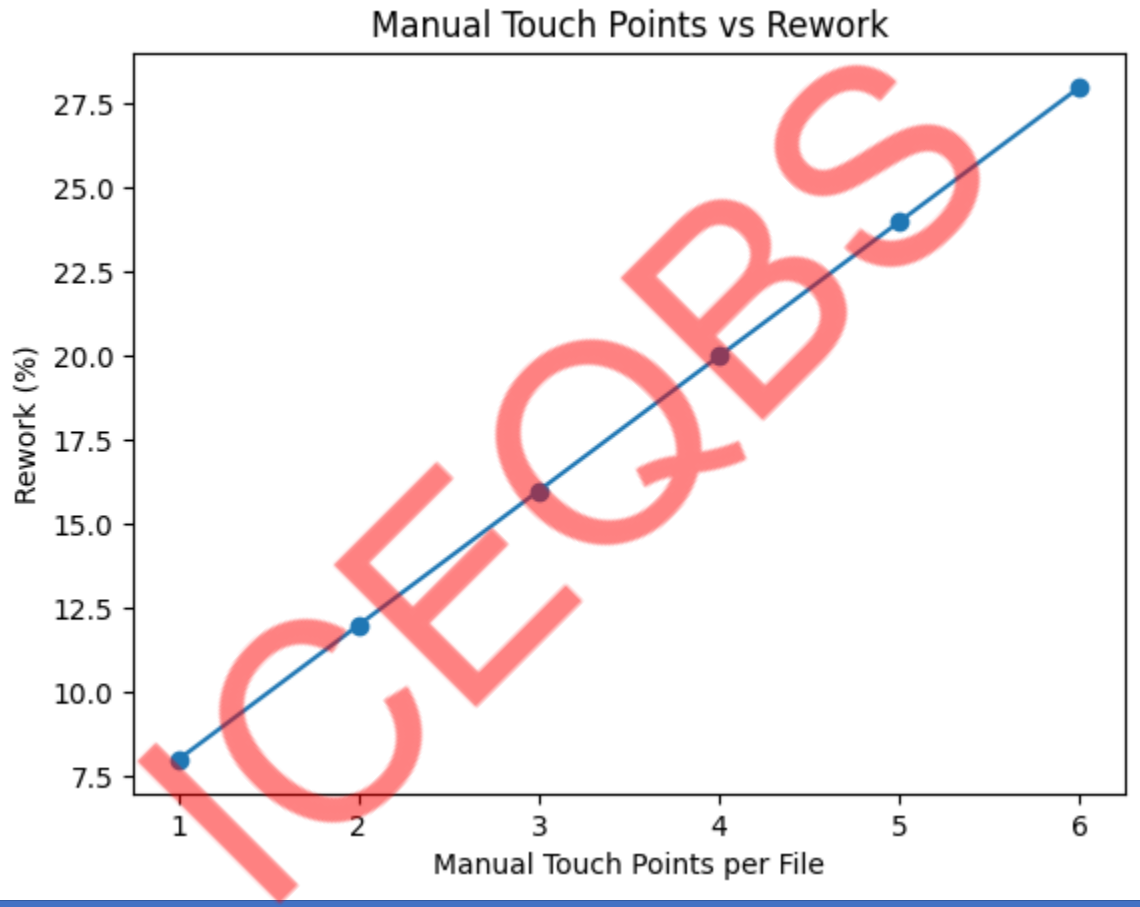
Analyse – Hypothesis testing



Inference :

As training coverage increases, rework percentage consistently decreases, indicating a **strong negative correlation**.

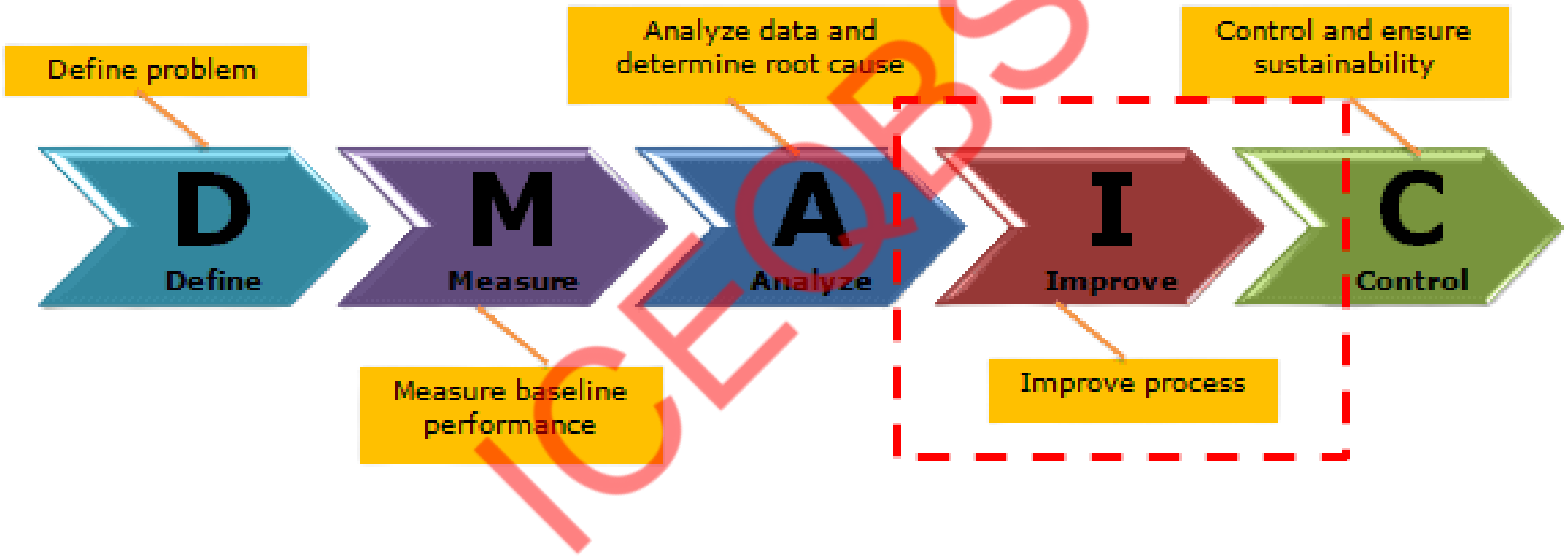
Analyse – Hypothesis testing



Inference :

An increase in manual touch points results in higher rework, confirming a **strong positive correlation**.

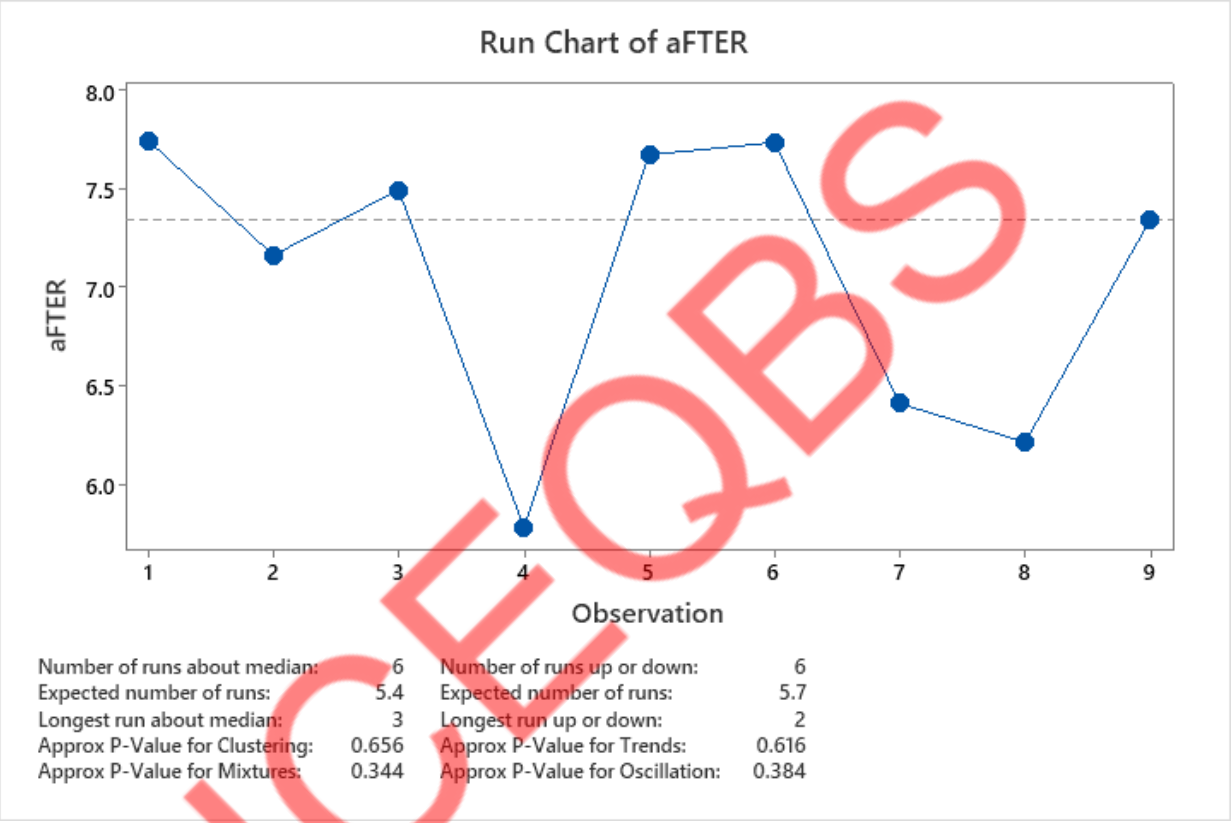
IMPROVE PHASE



Improve

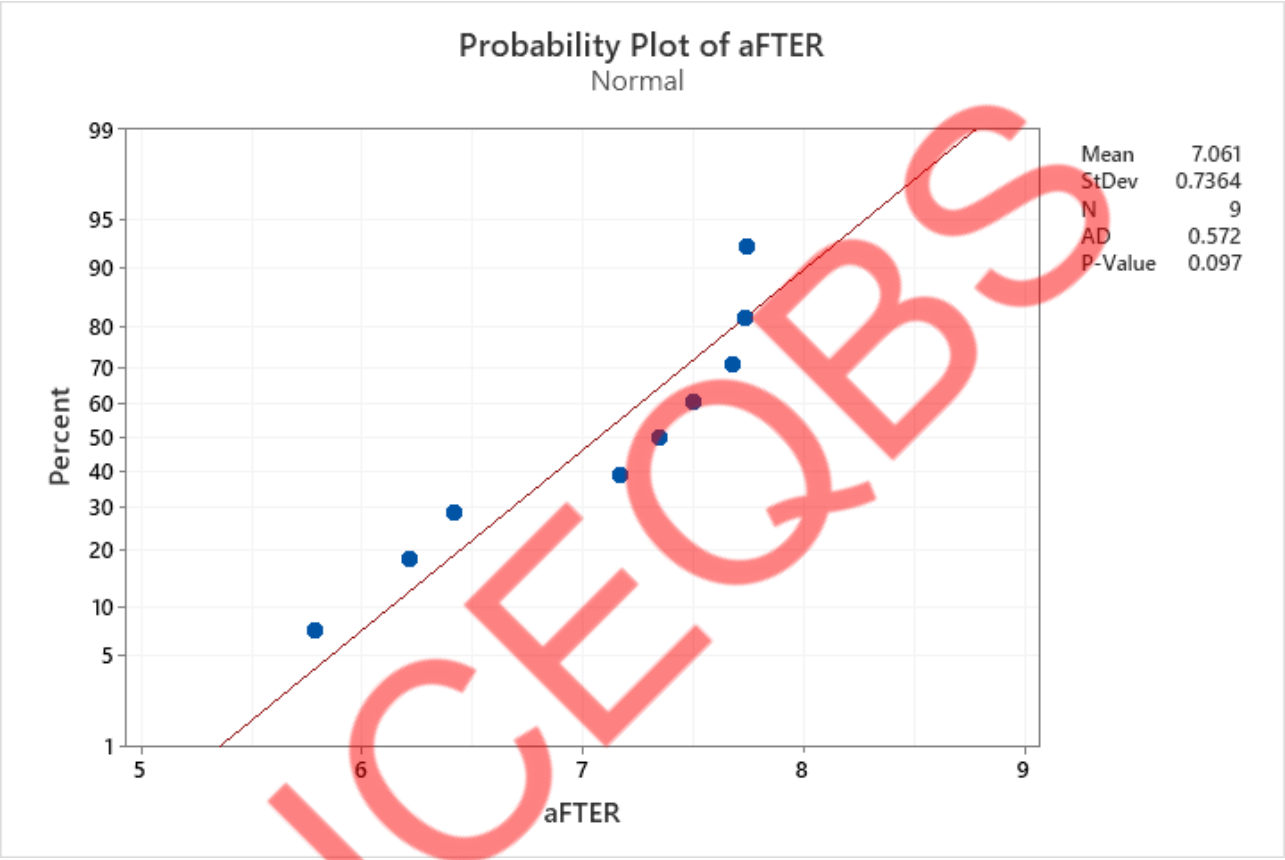
Action No.	Critical Root Cause Addressed	Improvement Action	Key Activities
1	OCR Tool Accuracy Limitations	OCR Model Optimization & Tuning	• Re-train OCR engine using high-error historical documents • Create separate OCR profiles for low-quality scans & complex layouts
2	OCR Tool Accuracy Limitations	Pre-Processing Standardization	• Implement image enhancement (desk, noise removal, contrast correction) • Enforce scan quality standards (DPI, grayscale, alignment)
3	Insufficient Segmentation Training	Mandatory Segmentation Certification Program	• Develop role-based training modules • Certification test with ≥85% pass criteria • Recertification every 6 months
4	High Manual Intervention	Rule-Based & Template-Driven Segmentation	• Create auto-segmentation templates for recurring article formats • Deploy rule engine for title and section identification
5	Insufficient Training & High Manual Intervention	Real-Time Error Feedback Loop	• Embed inline error prompts in tool • Auto-flag recurring errors by operator • Weekly defect review dashboards

Improve



The run chart shows a stable post-improvement process with random variation and no significant trend, shift, or special-cause behavior, indicating sustained control after implementation.

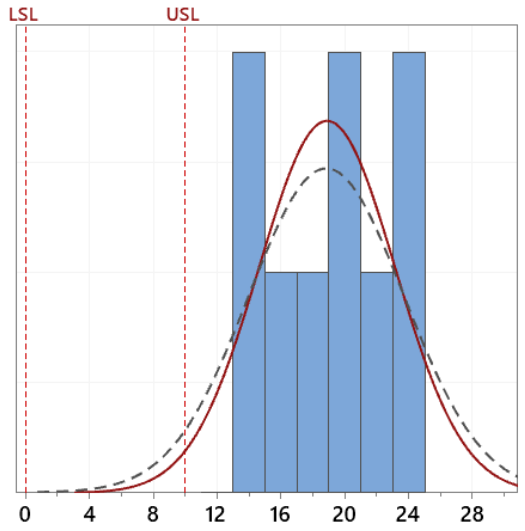
Improve



The probability plot indicates the post-improvement data follows a normal distribution (p-value > 0.05), confirming process stability and suitability for capability analysis.

Improve – Process capability – Before & After Improvement

Process Capability Report for Before1



Process Data	
LSL	0
Target	*
USL	10
Sample Mean	18.8889
Sample N	9
StDev(Overall)	4.25572
StDev(Within)	4.87589

Overall
Within

Overall Capability	
Pp	0.39
PPL	1.48
PPU	-0.70
Ppk	-0.70
Cpm	*

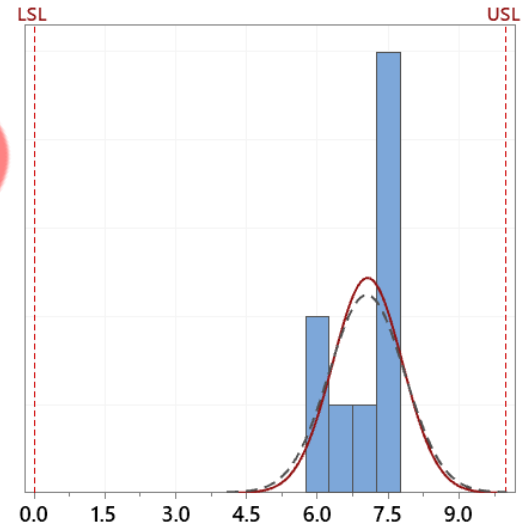
Potential (Within) Capability	
Cp	0.34
CPL	1.29
CPU	-0.61
Cpk	-0.61

Performance

	Observed	Expected Overall	Expected Within
PPM < LSL	0.00	4.53	53.55
PPM > USL	1000000.00	981632.39	965850.60
PPM Total	1000000.00	981636.92	965904.14

The actual process spread is represented by 6 sigma.

Process Capability Report for aFTER



Process Data	
LSL	0
Target	*
USL	10
Sample Mean	7.06138
Sample N	9
StDev(Overall)	0.736397
StDev(Within)	0.800235

Overall
Within

Overall Capability	
Pp	2.26
PPL	3.20
PPU	1.33
Ppk	1.33
Cpm	*

Potential (Within) Capability	
Cp	2.08
CPL	2.94
CPU	1.22
Cpk	1.22

Performance

	Observed	Expected Overall	Expected Within
PPM < LSL	0.00	0.00	0.00
PPM > USL	0.00	32.96	120.24
PPM Total	0.00	32.96	120.24

The actual process spread is represented by 6 sigma.

Inference :

The process capability analysis shows a **clear shift from an incapable process before (negative Cpk and high defects) to a capable and stable process after improvement (Cpk > 1.2)**, with defects reduced to near zero and specifications consistently met.

Improve

Two-Sample T-Test and CI: Before1, aFTER

μ_1: population mean of Before1
μ_2: population mean of aFTER
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Before1	9	18.89	4.26	1.4
aFTER	9	7.061	0.736	0.25

Estimation for Difference

Difference	95% CI for Difference
11.83	(8.51, 15.15)

Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis $H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
8.22	8	0.000

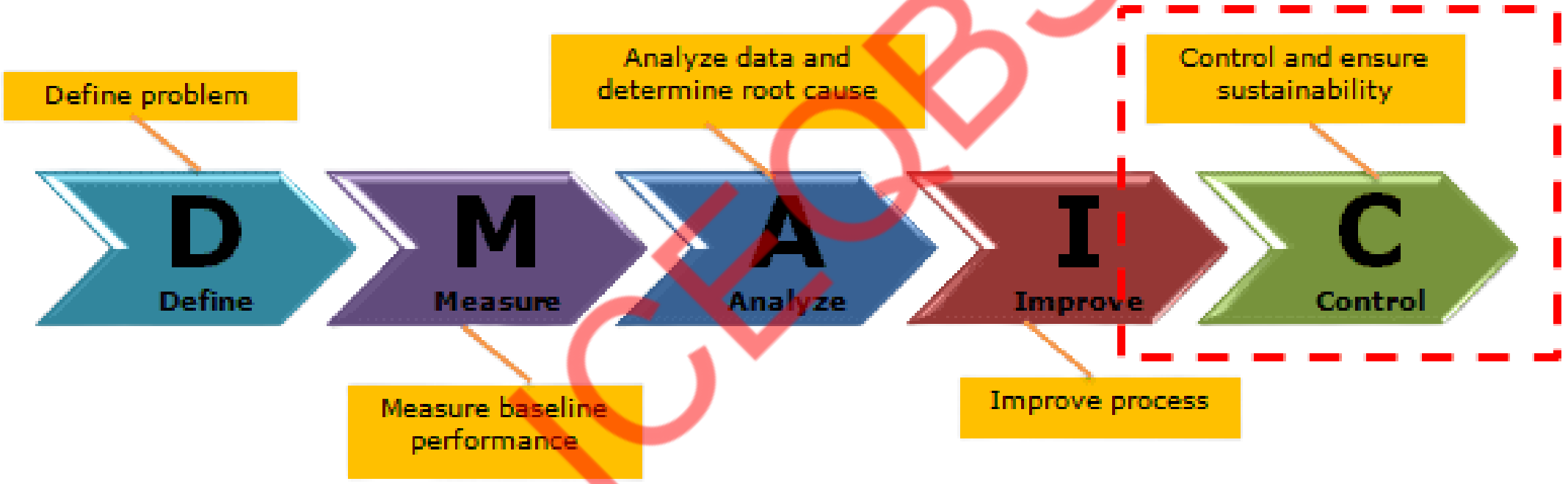
Inference:

The two-sample t-test shows a **statistically significant improvement after implementation**, with the mean reducing from **18.89 to 7.06** ($p < 0.001$), confirming the effectiveness of the improvements.

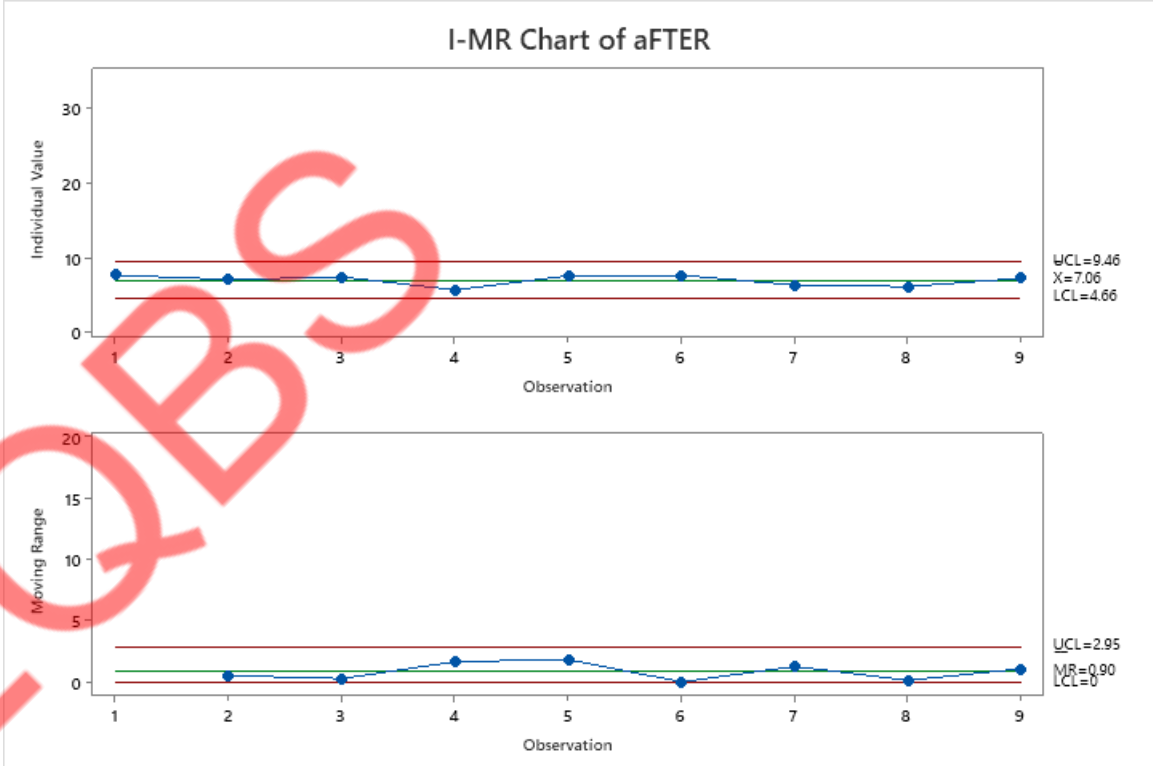
FMEA

Process Step	Potential Failure Mode	Effect of Failure	S	Potential Cause	O	Current Controls	D	RPN	Recommended Proactive Action
OCR model tuning	OCR accuracy does not improve after tuning	Continued OCR-related rework	9	Inadequate training dataset; poor document diversity	6	Pilot testing	6	324	Use stratified OCR training data covering all document types; validate on control set before rollout
Pre-processing standardization	Pre-processing rules not consistently applied	OCR errors reintroduced	8	Manual bypass of pre-processing	5	SOP documentation	7	280	Embed pre-processing as mandatory system step with no override option
Segmentation certification	Operators bypass certification	Inconsistent segmentation & title defects	8	Production pressure	6	Training records	6	288	System access enabled only for certified users (role-based access control)
Template-driven segmentation	Incorrect template selection	Wrong article segmentation	7	Similar-looking document formats	5	Manual selection	6	210	Auto-template suggestion using document pattern recognition with confirmation prompt
QC checklist enforcement	QC checklist filled mechanically	Defects escape to customer	9	Tick-box behavior	4	QC audit	7	252	Randomized mandatory sample review & checklist logic validation

CONTROL PHASE



Improve (Statistical validation for Improvement – I-MR Chart)



Inference:

- The I-MR charts show that process mean and variation have significantly reduced after improvement, with all points within control limits, indicating a stable and controlled process.

Control Plan - 5S and poka yoke mechanism

No.	5S Element	Poka-Yoke Mechanism	How It Works (Error Prevention Logic)	Sustaining Benefit
1	Sort	Document Pre-Qualification Gate	System blocks processing if scan quality (DPI, skew, noise) is below threshold	Prevents poor OCR input → avoids downstream rework
2	Set in Order	Template-Driven Segmentation Lock	Only approved article templates selectable; free-form segmentation disabled	Eliminates incorrect sectioning & title defects
3	Shine	Auto OCR Confidence Flagging	Low-confidence OCR zones auto-highlighted for focused review	Reduces over-editing and missed OCR errors
4	Standardize	Mandatory QC Checklist Hard-Stop	Workflow cannot proceed unless all QC checklist fields are completed	Prevents skipping of critical quality checks
5	Sustain	Operator Error Trend Alerts	System auto-alerts when operator error rate exceeds control limits	Early detection → corrective coaching before defects rise

Control Plan

No.	Critical Process Step	Control Metric (CTQ / X)	Control Method	Frequency	Reaction Plan
1	OCR Processing	OCR Accuracy %	Automated OCR accuracy dashboard with control limits	Daily	If accuracy < control limit → retrain OCR model and review input quality
2	Document Pre-Processing	Pre-processing Compliance %	System-enforced pre-processing log review	Per batch	Non-compliance → block processing and trigger IT review
3	Segmentation Execution	Certified Operator Coverage %	Role-based system access audit	Weekly	Unauthorized access → revoke access and re-certify operator
4	Manual Intervention	Manual Touch Points / Article	Trend chart (I-MR) from workflow logs	Weekly	If trend upward → root cause review and template optimization
5	Quality Assurance	QC Checklist Adherence %	Mandatory QC hard-stop with random sample audits	Daily	Audit failure → retraining and focused quality review

Conclusion

Results after improvement

- This project successfully reduced rework by stabilizing the content formatting process through improved OCR accuracy, targeted training, and minimized manual touchpoints, delivering sustained gains in quality, turnaround time, and cost efficiency.