PRACTICAL EXAMPLE OF THE APPLICATION OF LEAN 6 SIGMA METHODOLOGY IN DATA SCIENCE[1]

Nikola Iliychev Iliev, PhD[2]
D.A. Tsenov Academy of Economics - Svishtov

Abstract: *The paper looks at the application of Lean Six Sigma Methodology in an existing Data Science Example of Data Mining, Data Processing and Report Generation.*

Keywords: *ESG, sustainability, sustainable leadership, triple bottom line, bee sustainable leadership philosophy*[3]

AI> Lean Six Sigma is a business methodology combining the principles of Lean manufacturing and Six Sigma to eliminate waste, reduce defects, and improve overall efficiency. The Lean approach emphasizes elimination of non-value-added process activities, while Six Sigma focuses on reducing variability and improving quality. By combining these methodologies, organizations achieve higher levels of process efficiency, reduce waste and defects, and ultimately improve customer satisfaction.

Data Science is an interdisciplinary field involving statistical, mathematical, and computational techniques to extract insights and knowledge from complex data sets. It encompasses various stages of the data analysis pipeline, including data collection, preprocessing, exploration, modeling, and interpretation. The field also incorporates machine learning and artificial intelligence techniques to build predictive models and automate decision-making processes. Data Science is widely used in various applications such as business analytics, finance, healthcare, and social sciences.

By integrating Lean principles such as eliminating waste and reducing variation with Six Sigma's data-driven approach to quality improvement, organizations can optimize their data processes and deliver better results. Lean Six Sigma (L6σ for short) methodology helps Data Scientists identify non-value-added activities in their processes, streamline workflows, and reduce errors and variability. By adopting L6σ, teams can improve efficiency, reduce costs, and enhance the quality of results, leading to better business outcomes.

The current paper describes a practical example of the application of L6σ methodology over a real Data Science Process. This application is part of the L6σ Certification of employees of Paysafe Group's Global Business Services Department located in Sofia, Bulgaria by InterQuality – a certified provider of trainings in L6σ.

[1] The paper is written as part of scientific study on the project "Artificial Intelligence in the Economic Perspective", No. KP-06-M65/2 dated 12.12.2022., financed by the "Scientific Research" Fund at the Ministry of Education and Culture

[2] The opinions, analysis and conclusions represent those of the author and are not in any way representative for Paysafe Group, Paysafe Bulgaria or InterQuality

[3] The paper was written with the help of artificial intelligence. The same is done in accordance with the theme of the research project, namely "Artificial intelligence in the economic perspective". The texts, for the writing of which artificial intelligence was used, are marked with **AI>** for the beginning and **<AI** for the end

The paper describes the Yellow Belt level certification of this paper's author, which required the completion of a Project by use of L6σ Methodology. **<AI**

In the current case the project can be described in the following Project Charter:

Table 1 – Project Charter

Element	Description
Project name	Improvement of the CAPDEV External Capex calculation
Business case	The CAPDEV External CAPEX calculation and report preparation is a vital part of the monthly process. It consists of the aggregation of multitude sources of non-standardized data, which is transformed to fit an existing report template. This requires a plethora of manual checks, corrections and interactions to standardize data inconsistencies. Later the standardized data is used for the External CAPEX calculation.
Problem statement	These steps take time, require concentrated manual work and cannot safeguard against mistakes due to human error. The final product is a large file, with a multitude of worksheets of data, big part of which aren't needed for the calculation, but only take up space and distract anyone who gets acquainted with its contents.
Project scope	Monthly external CAPEX calculation report & monthly external vendor invoices and timesheets, used in this report preparation.
Goal statement	Objective 1 and KPI 1: To minimize the time required for data input, calculation and report completion, as measured in hours.
	Objective 2 and KPI 2: To reduce the possibility of human error due to manual work, by use of standardization of form and automation of data input and calculation, as measured with number of manual adjustments, interactions and comments needed within the report.
	Objective 3 and KPI 3: To reduce the size of the report file and contents, by minimizing data input to only the required for calculation and report completion, as measured in number of individual data inputs, calculations and results, but also file size.

AI> The definition of a project charter is part of L6σ methodology application, which follows a structured approach to problem-solving known as DMAIC (Define, Measure, Analyze, Improve & Control):

- **Define** the project (charter), establishing the scope, goals, and objectives.
- **Measure** the collected data to define a baseline for current performance and identify sources of variability. KPIs are established to measure progress.
- **Analyze** data to identify root causes of process inefficiencies and defects. Statistical tools and techniques are used to identify trends and patterns in the data and quantify the impact of the root causes on the process performance.
- **Improve** by implementation of solutions to root causes of process inefficiencies or defects. Solutions are tested, and their impact is measured. Once the solutions have been proven effective, they are fully implemented.
- **Control** of the process to ensure an ongoing sustainability in improvements, by measure and assessment of process performance KPIs within established limits. Deviations are identified and corrected promptly to prevent reoccurrence of defects or inefficiencies. **<AI**

For the current project the Define phase includes:

- a project charter, shown in Table 1;

- a SIPOC Diagram explaining key elements and stakeholders, grouped as **S**uppliers, **I**nputs, **P**rocesses, **O**utputs and **C**ustomers[4];
- a stakeholder analysis that juxtaposes each stakeholder's influence on the project against their overall importance;
- a process flowchart, defining the current step-by-step completion of the process, with accent on where the L6σ Methodology will be applied.
- a CTQ[5] Analysis, presented in the following Tree Diagram:

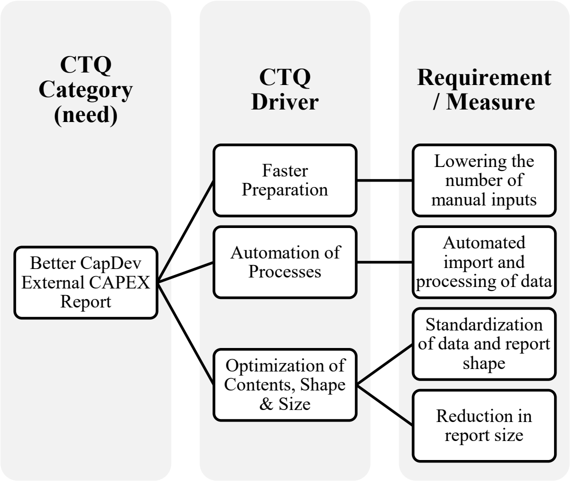

CTQ Category (need)
CTQ Driver
Requirement / Measure
Better CapDev External CAPEX Report
Faster Preparation
Automation of Processes
Optimization of Contents, Shape & Size
Lowering the number of manual inputs
Automated import and processing of data
Standardization of data and report shape
Reduction in report size

Figure 1 -CTQ Tree Diagram

Next is the Measure phase, including a Data Collection Plan and a "Treatment of Time" statement, concerned with the current project specifics. The data can be systemized in the following table:

Table 2 – Initial State of Data

METRIC	INITIAL STATE
Time (work days)	3 Work Days[6]
Size (physical file)	3 MB physical file size
Size (cells of data)	201 993 individual cells of data
Size (data tables)	29 sheets of data tables
# of Manual Inputs	280 manual inputs
# of Manual Calculations	163 manual calculations
# of Automated[7] Inputs	0 automated inputs
# of Automated Calculations	0 automated inputs
# of Outputs	6 109 data outputs
# of Errors	10 data outputs

[4] In our case Suppliers are the people/teams/departments that provide given sources of data, while Customers are the people/teams/departments that use the created report.

[5] CTQ stands for Critical-to-Quality. A CTQ Analysis concerns with the relationship between a CTQ's Category of Need (what do we want), its CTQ Drivers (how do we get that) and their CTQ Requirements (what must be done)

[6] Workdays is used as approximation, as the report completion is conditional on a plethora of variables. To be able to realistically measure time currently spend, and later time saved, we must define a specific rule for time use.

[7] Automatically is assumed to mean physically copy and paste the data provided, instead of enter it manually. Full Automation is only possible when data provided is read by a system with no need for manual Interaction, which is outside of the scope of the current Project.

Time is split into three main categories:

- Data Mining time – related to acquiring reports and sources of data. Cannot be evaded or influenced.
- Manual Labor time – related to manually entering data into the work file. Cannot be evaded but can be reduced by automation.
- Manual Calculation time – related to the need to manually calculate non-general cases, due to data differences and inconsistencies. Can be eliminated by data standardization.

To measure the result of automation and standardization a proper proxy for time must be introduced.

Here the proxy chosen is the number of key-strokes, manual inputs and manual calculations done for the report to be classified as completed. For the project, the current state of time is measured as follows:

Table 3 – Initial State of Data, Treatment of Time

STATUS	INITIAL STATE
Manual Labor time	
New Row Presses	265
Names Key-strokes	2,232
Amounts Key-strokes	881
Saved Key-strokes	0
Total	**3,113**
Manual Calculation time	
Manual Inputs	223
Manual Calculations	287
Automated Inputs	0
Automated Calculations	0
Total	**510**

The next phases are shown simultaneously as to allow focus directly on results.

A Root Cause Analysis allowed for the discovering of inconsistencies between different sources of data of type “Personal Names”. The report preparation requires these names to be cross-referenced, for the purpose of data joining. Currently this is done manually because of differences in data input. Examples are shown in the following table:

NAMES IN INVOICES	NAMES IN INTERNAL DATA
Katharine Mayer	Katharine Mayer
Erika Siguertt Fagundez	Erika Siguertt
Milorad Ničić	Milorad Nicic
Nikolay Uzunov	Nikolai Uzunov
Beniamin Andrei Muntean	Beniamin-Andrei Muntean
Ivaylo Iliev Rashevski	Ivaylo Rashevski
Ivan Ivanov	Ivan V. Ivanov
Rubén Carrillo	Ruben Carillo
Luis Pérez	Luis Perez
Maheshwari Keerthi	Keerthi Maheshwari

Common inconsistencies in names are Differentiating transliteration of Cyrillic Letters, Language specific Letter Generalization, Inclusion or Omission of Surnames, Rotation of First and Last Name and Spelling errors.

To counteract this, a list of "rules" for automated cross-referencing is introduced. Initially, name strings are separated into first name and last name sub-strings. Both are cross-referenced independently. When both names match, we have Rule One – Whole Name. When only first or last names match, we have Rule Two or Two respectively – First Name or Last Name. In Rules where data includes middle name, cross-referencing is done only on border sub-strings, introducing Rule Four – First & Last Name.

In Rules where there are different spellings in names (due to transliteration generalization or spelling errors) a Rule Five rule is introduced, where a custom "anchor" for each sub-string is created, using only first and last letter in each name. Then these custom anchor sub-strings are cross-referenced (e.g., Luis Perez to L--s P---z). Finally, in Rules where names are improperly understood – what is First vs what is Last name, a "crossed" cross reference is done on First against Last and Last against First. Each Rule is only applied after previous haven't return a successful cross-referencing, thus allowing for elimination of general cases toward more specific ones. Possible duplication of First or Last names is minimized by use of introduction of an additional control measure – independent analysis of each group of names by individual value of Vendor or Team affiliation (data for which is available in both cross-referenced sources). Thus, the only cases left for concern of possible duplication is when a given team includes two different people with similar First or Last name and/or similar combination of "anchor" point letters that are also on the same level of names inconsistencies – e.g., Ivan Ivanov will be cross referenced against Ivan M. Ivanov by Rule 2 before a Rule 3 check mistakenly cross-references him against Stoyan Ivanov.

The only existing risk of error would be the existence of a second person with First Name like for example Ivan Petrov who would be mistakenly cross-referenced against Ivan Ivanov. To avoid this a counter of substring duplicates is introduced for each group of Vendor or Team affiliation. This counter returns 0 where this risk is mitigated in advance. For any other value a manual check is needed, but in the existing example such cases are not found.

Other examples of standardization, optimization and automation can be given, but are omitted from the paper due to volume constraints. Still, we can turn our attention towards the results of their application, by looking at the changes in previously mentioned KPIs and data state. We first look at the "after" state of Data, as show in the following table:

Table 4 – Initial and Final State of Data

METRIC	INITIAL STATE	FINAL STATE	CHANGE
Time (work days)	3 Work Days	2 Work Days	33.00% faster
Size (physical file)	3 MB	100 to 200 kb	93.31% smaller
Size (cells of data)	201 993 cells of data	104 643 cells of data	48.19% less cells
Size (data tables)	29 sheets	8 sheets	72.41% less sheets
# of Manual Inputs	280 manual inputs	130 manual inputs	53.57% less inputs
# of Manual Calcs	163 manual calcs.	11 manual calcs.	93.25% less calcs.
# of Automated Inputs	0 automated inputs	148 automated inputs	
# of Automated Calcs	0 automated inputs	278 automated calcs.	
# of Outputs	6 109 data outputs	5 247 data outputs	14.11% less outputs

AI> The data clearly shows that introducing data standardization, calculation automation, and process optimization have had a significant impact on the task, resulting in a smaller file size, fewer manual inputs and calcs, and a reduced number of sheets. The reduction in file size and number of sheets is due to the elimination of redundant or unnecessary data and the optimization of the data structure. The decrease in manual inputs and calcs, along with the introduction of automation, has led to a more efficient and accurate processing of the data, saving time and reducing the likelihood of errors.

The application of L6σ methodology within Data Science is an effective approach to process optimization, as it emphasizes continuous improvement, reducing waste, and improving efficiency. The methodology is well suited to data processing tasks as it involves the use of statistical methods to identify areas for improvement and data-driven decision making. The reduction in the number of manual inputs and calcs and the introduction of automation aligns with the Lean principle of reducing waste, while the optimization of the data structure and standardization of data aligns with the Six Sigma principle of reducing variability. <AI

Additionally, we look at the final state of data in terms of Time, proxied through manual key strokes, already discussed in the paper. The results are, as follows:

Table 5 – Initial and Final State of Data, Treatment of Time

VARIABLE	Initial state	Final state	Change (#)	Change (%)
New Row Presses	265	266	1	0.38%
Names Key-strokes	2,232	1,663	-569	-25.49%
Amounts Key-strokes	881	772	-109	-12.37%
Saved Key-strokes	0	-532	-532	-17.09%
Total Labor Time	**3,113**	**1,903**	**-1,210**	**-54.95%**
Manual Inputs	223	24	-199	-89.24%
Manual Calculations	287	0	-287	-100.00%
Automated Inputs	0	188		
Automated Calculations	0	277		
Total Manual Calculations	**510**	**24**		

AI> The introduction of automation in data input and calculation processes has resulted in significant reduction in the number of manual key-strokes required for data entry. This reduction has led to a corresponding decrease in total labor time required, as shown by the "Total Labor Time" variable. Additionally, the automation of calculations has eliminated the need for manual calculations, resulting in a reduction of the "Manual Calculations" variable to zero. This has also led to a decrease in total manual calculations.

The reduction in manual input and calculations has been achieved without sacrificing data accuracy or quality, as the automated processes have been designed to meet L6σ quality standards. This is evident in the increase of Automated Inputs and Automated Calculations, indicating that the automation has successfully taken over the manual processes. The decrease in manual inputs and calculations has resulted in a decrease in total labor time, which is a key component of L6σ.

The reduction in manual key-strokes and calculations has led to decrease in errors and inconsistencies, as automation eliminates the risk of human error. This is reflected in the negative "Saved Key-strokes" variable, indicating that the automation has prevented mistakes that would have occurred if manual processes had been used. **<AI**

Another result of the application of L6σ methodology relates to data utilization. In the initial process of report creation, a large part of the inputs, calculations and outputs are not required to produce the final result, but are instead a product of repetitive activities, happening due to data inconsistencies. After application of the methodology, we find ourselves in front of a cleaner file, where the number of inputs, calculations and outputs is reduced, without compromise in terms of introduction of deviations, errors or impossibilities in deriving the same results. This could be seen in the following graph:

On the graph we can see different data providers (each shown with two bars of two different variables) and two average levels. For each vendor we see the percentage of data utilization initially. The vendors are sorted from least utilized data to most utilized data. The average data utilization is 69%. After application of L6σ we have an average data utilization of 85%. Individual data utilization of vendors disappears as different data sources are standardized and merged within a single data source "Vendors".

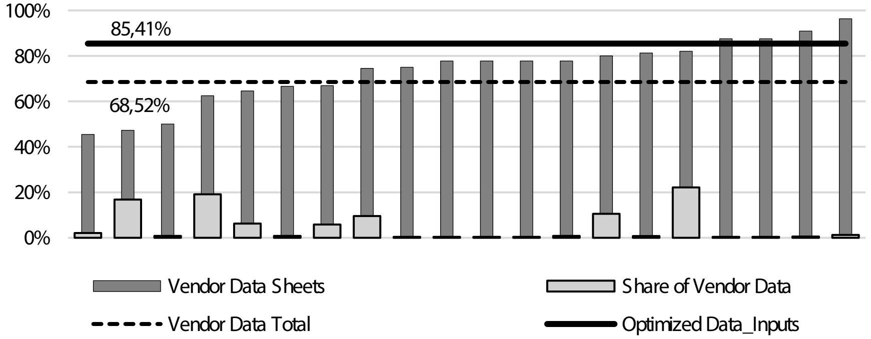
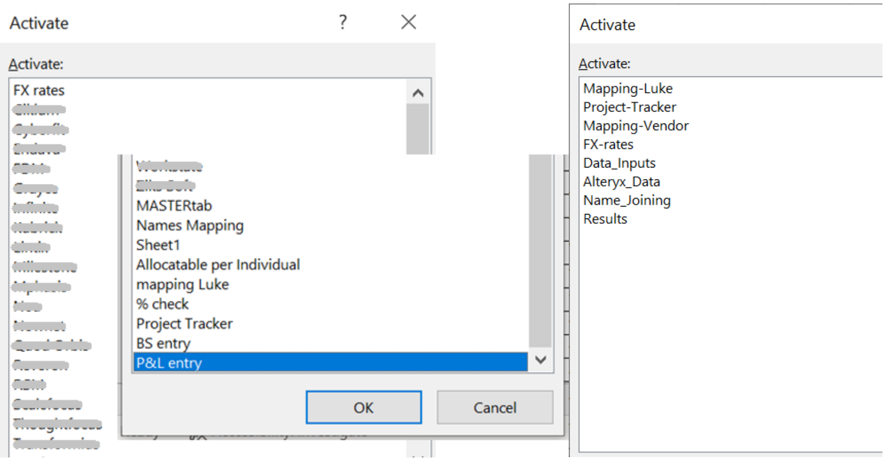


Figure 2 Data Utilization before and after states

Regarding the small number of vendors whose initial level of data utilization was higher, we look at the second wider bar values seen in the bottom. Each represent the share of data of each vendor within the overall size of data source. As it can be seen all vendors whose data is over the optimized average level have a miniscule share in the overall data source. Instead, we turn our attention towards most visible vendors whose shares are between 7 and 21%. The names of all vendors are scrubbed as to not reveal company sensitive information about possible clients or business partners. For mentioned vendors with visible shares in overall data source we can see that all of their utilization has increased after application of L6σ.

Finally, we look a figure, visualizing the number of datasheets of tables in the initial file and in the final file as to compare how much are needed for completion of the calculation and generate the report. The number of rows shows more than anything else that the methodology helps in bettering the work and results and the overall process.



Activate

Activate:

FX rates
[redacted]

[redacted]
MASTERtab
Names Mapping
Sheet1
Allocatable per Individual
mapping Luke
% check
Project Tracker
BS entry
P&L entry

Activate

Activate:

Mapping-Luke
Project-Tracker
Mapping-Vendor
FX-rates
Data_Inputs
Alteryx_Data
Name_Joining
Results

Figure 3 – number of sheets of data tables – before and after state

Having shown this example, the conclusion of this scientific paper is that the practical application of L6σ Methodology for the purposes of Data Science is not only possible, but a very good fit towards bettering the process of data analysis, result generation, and of course time and cost saving.