الجمهورية الجزائرية الديمقراطية الشعبية
PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
وزارة التعليم العالي والبحث العلمي
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
المدرسة الوطنية العليا للتكنولوجيا والهندسة – عنابة
NATIONAL HIGHER SCHOOL OF TECHNOLOGY AND ENGINEERING – ANNABA

Department of Industrial Engineering

In Partial Fulfillment of the Requirements for the Degree of

STATE ENGINEER

Domain: Science and Technology
Field of Study: Industrial Engineering
Specialization: Maintenance and Reliability of Industrial Systems

Presented by
Abderahim REDOUANE
Malak EL MEMI

A Data-Driven Approach for Automated Generation of Interactive Dashboards in Industry 4.0

Supervised by
Dr. Salim KEBIR
NHSTE-Annaba

Examination Board:

Pr. Messaoud DJEGHABA	President	NHSTE-Annaba
Dr. Nadia LACHTAR	Examiner	NHSTE-Annaba
Mr. Ahmed BENDJELLOUL	Guest	SARL ARTEC INT

Academic Year 2025

Abstract

This thesis presents the development of a web-based application "CSViz" that automates the creation of analytical dashboards from CSV files. Users simply upload their datasets, and the system cleans, merges, and visualizes data with minimal intervention. The application aims to democratize data insights by simplifying the dashboard creation process for non-expert users in Industry 4.0 context.

Key words: *Dashboard Automation, Data Visualization, Business Intelligence, AI-Powered Dashboards, Industry 4.0.*

Résumé

Cette thèse présente le développement d'une application web "CSViz" qui automatise la création de tableaux de bord analytiques à partir de fichiers CSV. Les utilisateurs n'ont qu'à télécharger leurs données, et le système nettoie, fusionne et visualise les données avec une intervention minimale. L'application vise à démocratiser l'accès aux perspectives analytiques basés sur les données en simplifiant le processus de création de tableaux de bord pour les utilisateurs non experts dans le contexte de l'Industrie 4.0.

Mots clés : *Automatisation de tableaux de bord, Visualisation de données, Business Intelligence, Tableau de bord basé sur l'IA, Industrie 4.0.*

ملخص

يقدم هذا البحث تطوير تطبيق ويب "CSViz" يقوم بأتمتة إنشاء لوحات البيانات التحليلية انطلاقاًا من بيانات. يقوم المستخدم فقط بتحميل مجموعة البيانات الخاصة به، ويتولى النظام عملية التنظيف، الدمج، والتصوير البياني للبيانات تلقائياًا مع تدخل بسيط. يهدف التطبيق إلى إتاحة الوصول إلى الرؤى البيانية لغير المتخصصين، وتبسيط عملية إنشاء لوحات البيانات في سياق الصناعة 4.0.

كلمات مفتاحية: أتمتة لوحات البيانات، تصوير البيانات، ذكاء الأعمال، لوحات معلومات مدعومة بالذكاء الاصطناعي، الصناعة *4.0*.

Acknowledgments

We would like to express our sincere gratitude to our supervisor, Salim KEBIR, for their invaluable guidance, patience, and continuous support throughout the course of this work. Their expertise and insightful feedback were essential to the development and completion of this thesis.

We also extend our appreciation to our internship hosts at SARL Izdihar Conserverie, whose availability and trust, provided us with a meaningful and enriching experience.

Dedication

To my family,
Your love has always been my safe space. Through the hardest moments, your words, your presence were enough to keep me going. You may not have fully seen the late nights, the doubts, or the pressure, but you've carried the weight with me in your own way. For that, I will always be grateful.

To my parents,
This achievement belongs to you as much as it does to me. Your sacrifices, patience, and unwavering belief in me have been the greatest gift I could ever receive. You raised me with strength, taught me resilience, and supported me in ways that go far beyond material needs. You never asked for recognition, but everything I do is a reflection of your efforts. I hope this makes you proud.

To my little brother,
A promise is a promise, your name belongs here. May this remind you that every goal is worth chasing, no matter how far it seems.

To myself,
For holding on when it felt heavy, and for turning doubt into effort this moment is well earned.

To my thesis partner,
For your steady presence and for the work we carried side by side. I am proud of what we've accomplished together.

Malak

Dedication

"Sic Parvis Magna - Greatness from small beginnings."

This motto always reminded me that even the boldest journeys begin with simple steps.

This thesis is one of those beginnings.

To my future self,
Remember this moment, not for the result, but for the effort and the progress that led you here. Believe in your path; you've only just begun.

To my small family,
To my small but great support system, thank you for constantly believing in me and giving me the space to have my proper journey. I hope this journey, in all its challenges and triumphs, makes you proud.

To my thesis partner,
This work carries both of our names, because it carries both of our efforts. Thank you for the shared ambition, the countless discussions, and the drive to do more.

Abderahim

List of Abbreviations

BI	Business Intelligence
KPI	Key Performance Indicator
LLM	Large Language Model
CSV	Comma-Separated Values
IIoT	Industrial Internet of Things
AI	Artificial Intelligence
ML	Machine Learning
PdM	Predictive Maintenance
ADRS	Automatic Dashboard Recommendation Systems
IQR	Interquartile Range
SLM	Small Language Model

List of Figures

List of Tables

Contents

Introduction

In the era of Industry 4.0, Business Intelligence (BI) and data visualization have become fundamental pillars for enhancing operational efficiency and supporting data-driven decision-making. Organizations are increasingly generating large volumes of data, and the ability to extract meaningful insights from this data is critical to maintaining a competitive edge.

Within this context, dashboards have emerged as essential tools in modern industrial environments. They, thus, serve as a one-stop solution for monitoring operations, visualizing performance metrics, and converting raw data into real-time actionable insights. Dashboards provide real-time tracking, strategic planning, and continuous improvement efforts by aggregating Key Performance Indicators (KPIs) and other pertinent data points.

The structure of this thesis is as follows:

Chapter 1 begins with a background of the internship company, followed by the research problem in Business Intelligence, particularly within modern industrial contexts. It then outlines the motivation for the project, its relevance to Industry 4.0, and its contributions to accessibility, automation, and decision support.

Chapter 2 provides a literature review and introduces the theoretical foundation. It covers key Industry 4.0 concepts, examines existing methods for automated dashboard generation and visualization recommendation. The chapter concludes with a theoretical framework linking data types to suitable visualization techniques.

Chapter 3 presents the methodology and technical contributions of our approach . It details the system architecture, technology stack, and development environment, followed by the data preparation pipeline, both manual and AI-assisted—and the logic behind our metric and visualization recommendation systems.

Chapter 4 showcases the application of the real world data. It demonstrates the dashboard creation process and evaluates the results based on accuracy, usability, and visualization quality. The chapter ends with a critical discussion of the system's strengths, limitations, and directions for future improvement.

Chapter 1

Company description, problem statement, and motivation

In this chapter, we will begin by presenting the background of the internship company to provide context for the industrial environment in which the project was developed. We will then discuss the research problem within the domain of Business Intelligence, focusing on its relevance and challenges in modern industrial settings. Finally, we will outline the motivation behind the project, highlighting its alignment with Industry 4.0 and its intended contributions to accessibility, automation, and decision-making support.

1.1 Company Description

SARL Izdihar Conserverie is an Algerian company operating in the agro-industrial sector, primarily focused on the processing and preserving of vegetable and fruit products. Legally established as a SARL with a share capital of 100,000,000 DZD, it is located in Ain Nechma, Benazouz, Skikda. The company employs about a hundred people and plays a key role in supplying agri-food products to the local market.

Its main product lines consist of harissa, tomato concentrate, ketchup, mayonnaise, tomato sauce, and several fruit-based goods, including jams and compotes. Its core operations involve the canning of fresh produce, supported by specialized facilities for raw material reception, processing, packaging, storage, and distribution. These units are equipped with modern machinery to ensure standardized processes that meet food safety standards.

The production lines cover all transformation stages like washing, cutting, cooking, canning, and sterilization, with operational data collected at each step to enable precise monitoring. This emphasis on traceability and process optimization aligns with the

principles of Industry 4.0. Through its capabilities and quality-driven approach, SARL Izdihar Conserverie strengthens its presence in the Algerian agro-industrial sector while contributing to local food security.

1.2 Problem Statement

In recent years, business intelligence and data visualization tools such as Tableau[1] and Power BI[2] have made significant advances in enabling data-driven decision making across industries. Despite these technological developments, a persistent challenge remains: the substantial learning curve required to develop proficiency in transforming raw data into actionable insights. Engineering professionals in industrial settings must invest considerable time and resources in specialized data analysis training to effectively explore datasets and extract meaningful conclusions.

This skills gap presents a significant barrier to the widespread implementation of data-driven approaches in industrial environments. As organizations transition toward Industry 4.0 frameworks, characterized by interconnected systems and real-time data generation, the need for accessible analytical capabilities becomes increasingly critical. The current data analysis techniques, which requires specialized expertise, creates bottlenecks in decision-making processes and potentially limits the adoption of data-informed practices.

1.3 Motivation

The motivation of this work is to explore an opportunity for innovation in the field of data analysis and visualization through the development of an automatic dashboard generation system tailored to the context of Industry 4.0. Such a tool aims to facilitate broader adoption of data-driven approaches by enabling more agile and informed decision-making, in alignment with the fast-paced and increasingly complex demands of contemporary industrial environments.

In conclusion, this chapter has provided the foundational context for our project by presenting the internship company, outlining the core research problem in Business Intelligence, and explaining the motivation behind our work. It has also highlighted the project's relevance to Industry 4.0 and its intended impact on accessibility, automation, and decision support in industrial settings.

[1] `https://www.tableau.com/`
[2] `https://www.microsoft.com/fr-fr/power-platform/products/power-bi/`

Chapter 2

Literature Review and Theoretical Framework

In this chapter we will present a comprehensive literature review and establish the theoretical foundation for the project. beginning by exploring key concepts of Industry 4.0, followed by an examination of existing methods for automated dashboard generation and visualization recommendation. The chapter concludes with the development of a theoretical framework that connects different data types to the most appropriate visualization techniques.

2.1 Industry 4.0 Concept

2.1.1 Definition and Key Characteristics

According to [1], the industrial sector is currently undergoing a profound transformation, widely recognized as Industry 4.0, which represents the fourth major era of industrial revolution. This latest phase builds upon the advancements of previous revolutions, characterized by a significant convergence of physical and digital technologies. At its core, Industry 4.0 signifies a move towards interconnected and intelligent manufacturing systems, where data plays an increasingly vital role in driving efficiency, enhancing flexibility, and optimizing decision-making processes.

Industry 4.0 is distinguished from the previous industrial eras by a set of characteristics, featured and discussed in [2]:

- **Interoperability** where systems, machines and devices along with people are able to connect and communicate with each other through technologies like Industrial Internet Of Things (IIoT) to share and exchange information seamlessly.

- **Virtualization**, which focuses on data transparency by creating a virtual copy of the physical world using data collected from sensors and simulation models, an example of this is Digital Twin applications.
- **Decentralization**, where it focuses on making systems and smart objects to give decisions autonomously.
- **Real-time capability** signifies the ability to collect and analyze data and provide insights immediately which is essential in applications like predictive maintenance.
- **Service Orientation** highlights the ability of systems to offer services to both internal and external stakeholders.
- **Modularity** refers to flexibility of smart factories to adapt smoothly and quickly to dynamic market requirements.

These characteristics, working synergistically, are fundamental to the transformative potential of Industry 4.0, a departure from traditional towards more adaptable and responsive systems.

2.1.2 Technological Foundations

The ability to collect, process, and analyze vast amounts of data is central to Industry 4.0, making Big Data and Analytics a critical technological foundation. Furthermore, data serves as the central nervous system that connects and drives the functionality of many technologies. For instance, the IIoT generates vast streams of data from industrial assets, which are then analyzed by Artificial Intelligence(AI) and Machine Learning (ML) algorithms, and provide insights from the data, where it gets visualized through dashboards using plots and KPIs.

2.1.3 Data-Driven Industry Trends

Industry 4.0 is fundamentally characterized by a strong emphasis on data, leading to several significant data-driven trends that are reshaping industrial operations. We mention a few trends discussed in [3].

- **Predictive Maintenance (PdM)** stands out as a prominent trend, leveraging data analytics and AI to forecast equipment failures and optimize maintenance schedules.
- **Smart Manufacturing**, which involves the extensive use of data from connected devices and systems to optimize production processes, enhance efficiency, and ensure consistent product quality. This trend encompasses the integration of various technologies to create intelligent and responsive production environments.

- **Supply Chain Optimization**is also heavily reliant on data, with analytics being used to improve forecasting accuracy, manage inventory levels more effectively, and streamline logistics operations across the entire supply network.
- **Dashboards** have been broadly used in business intelligence to help data analysts explore and discover data insights with multiple-view visualizations, which would be helpful in industry 4.0 context.

2.2 Related Works

In this section, we present a literature review and analysis of related works, beginning with studies that address the general principles and methodologies of dashboard generation. We then progressively explore more specialized contributions, including the application of ML techniques and heuristic-based approaches for automated dashboard design.

Praveen Soni et al. in [4] explored the design of Automatic Dashboard Recommendation Systems (ADRS) with a particular focus on how such systems accommodate both novice and expert users. they classified recommendation mechanisms from rule-based to machine learning-driven approaches. Their analysis of 19 systems revealed that ADRS particularly benefits novice users through automation of visualization choices and intuitive interfaces that abstract technical complexity. These systems use heuristics and guided recommendations to help users with limited data literacy gain insights efficiently, democratizing access to business intelligence in industrial contexts.

Qiyue Zhang in [5] studied the impact of interactive data visualization on decision-making in business intelligence and demonstrated its positive impact on decision-making speed and quality by providing immediate data access and enabling scenario-based analysis. The study emphasized how visualization tools can strengthen data-centric organizational culture by making data accessible to non-technical users, which aligns with the proposed system's goal of automating dashboard creation for users with limited data analysis experience.

Dazhen Deng et al. in [6] introduced DashBot, the core of the proposed system is a framework powered by deep reinforcement learning utilizing the Asynchronous Advantage Actor-Critic (A3C) Algorithm for automatic generation of dashboards. This approach models the process as a Markov Decision Process, where an agent learns through a sequence of states of actions, which ensures an iterative and exploratory methodology for dashboard design. The system learns through a process of trail and error, guided by a reward mechanism that evaluates the quality and informativeness of the generated dashboard.

Kevin Hu et al. in [7] propose VizML, an ML-based visualization recommendation system that aims to lower the barrier for exploring basic visualizations by learning from a large-scale corpus of user-generated visualizations on the Plotly platform. The authors extracted over a million dataset-visualization pairs and framed visualization recommendation as predicting many key design choices (chart type, axis assignments, etc.) based on dataset features, using neural networks and baseline models, they achieved high prediction accuracy.

Jock D. Mackinlay, Pat Hanrahan, and Chris Stolte in [8] introduce "Show Me," a feature in Tableau that is designed to guide users toward the creation of visualizations by a set of heuristic rules. These rules intelligently map the specific characteristics of the data attributes selected by the user, with a particular focus on the number and type of dimensions and measures, to a carefully curated selection of appropriate visual encodings.

The reviewed works explore various methods for automating dashboard generation, from heuristic-based systems like Tableau's "Show Me" [8] to advanced machine learning approaches such as VizML [7] and DashBot [6]. ADRS systems [4] focus on aiding both novice and expert users, while interactive visualizations [5] are shown to enhance decision-making and data accessibility. Together, these studies highlight the growing role of AI and user-centered design in democratizing data visualization.

2.3 Theoretical Framework of Visualizations

The most popular visualization types must be defined and categorized in order to support the automated recommendation system. Every chart has a distinct analytical function and works best with specific data structures or relationships. A variety of common visualizations are compiled in the following table, which also describes the kinds of data they represent and their primary purposes. The matching of datasets with suitable graphical representations is based on this theoretical mapping. Additionally, it helps the system choose charts that effectively and clearly communicate insights.

Visualization	**Description**
Line chart	Displays trends and changes in data over time [5]
Bar Chart	Displays data with discrete values and highlights differences between categories [5]
Histogram	Shows the distribution of a numeric variable by grouping values into bins
Box Plot	Shows typical quantiles of the distribution of data, and for single-dimensional measures of central tendency and dispersion of data [9]
Pie Chart	Shows the proportional composition of a whole, with each slice representing a category's relative size [5]
Treemap	Displays hierarchical data as nested rectangles, with the size of each rectangle proportional to the value it represents [5]
Choropleth Map	Displays location shaded according to a variable's value per region
Heatmap	Uses color intensity to represent the magnitude of values in a matrix or grid [5]
Scatter Plot	Displays the relationship between two continuous variables, with each data point represented by a marker [5]

Table 2.1: Visualization Types and Their Applications

In conclusion, Chapter 2 has provided a thorough review of relevant literature and established the theoretical underpinnings of our work. It has covered essential Industry 4.0 concepts and reviewed current methodologies for automating dashboard creation and visualization recommendations. The chapter has also introduced a framework that guides the selection of suitable visualization techniques based on the nature of the data.

Chapter 3

Methodology and Contribution

This chapter presents the methodology and technical contributions of our approach. It begins by outlining the overall system architecture, the chosen technology stack, and the development environment. It then describes the data preparation pipeline, encompassing both manual and AI-assisted techniques. Finally, the chapter delves into the design logic behind our metric selection and visualization recommendation systems.

3.1 System Overview

The application developed during this project was named CSViz, reflecting its core functionality of visualizing and analyzing CSV-based data.

CSViz inputs a CSV file, performs data readiness evaluation and cleaning (manually or AI-assisted), and then automatically generates a dashboard consisting of key metrics and appropriate visualizations.

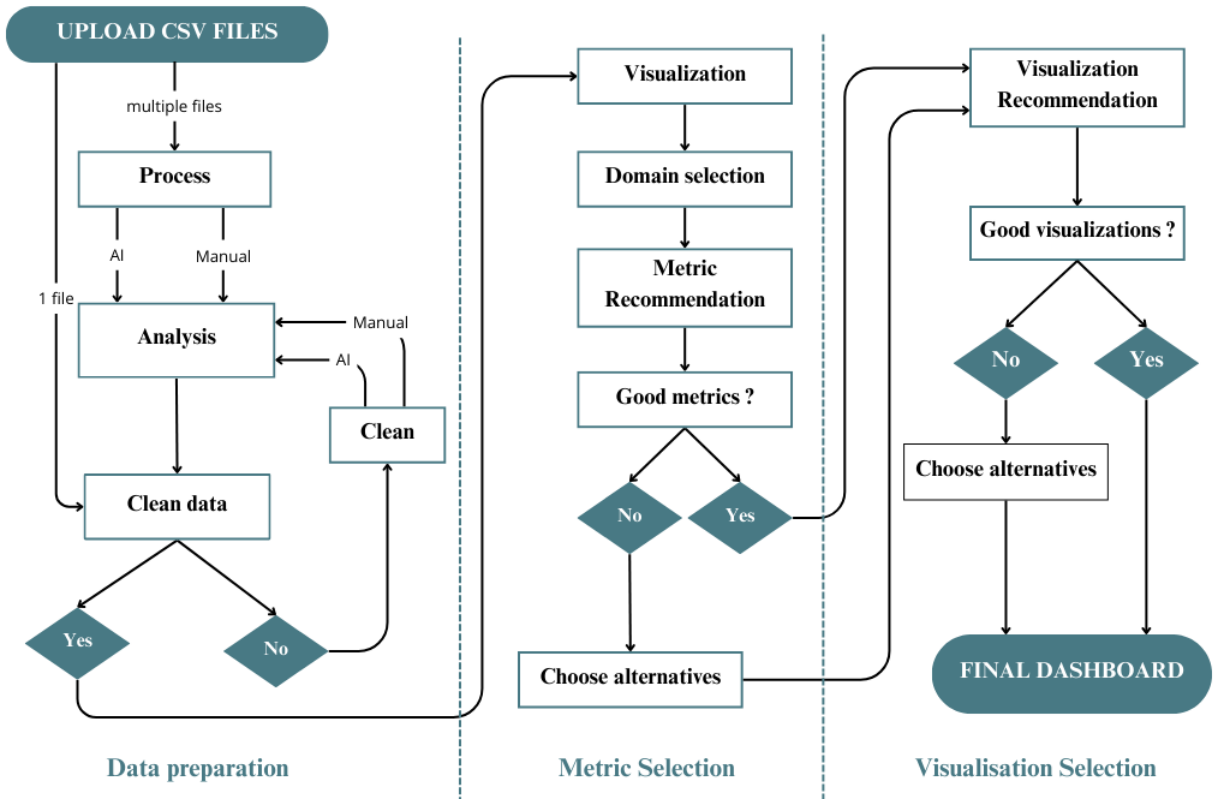

Figure 3.1: Diagram explaining the flow of CSViz

3.2 Tech Stack and Development Environment

The development of the application was done using a Python-based technology stack, consisting of the following tools and libraries:

- **Streamlit**[1] **(≥1.31.0)** is an open-source Python framework for data scientists and AI/ML engineers to deliver dynamic data apps with only a few lines of code.
- **Pandas**[2] **(≥2.0.0)** is an open-source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.
- **NumPy**[3] **(≥1.24.0)** is Python's fundamental package for scientific computing. It is a Python library that provides a multidimensional array object, various derived objects, and an assortment of routines for fast operations on arrays.
- **Scikit-learn**[4] **(≥1.2.0)** A robust machine learning library in Python offering tools for classification, regression, clustering, dimensionality reduction, and model evaluation, used here mainly for statistical scoring and preprocessing logic.

[1] `https://docs.streamlit.io/`
[2] `https://pandas.pydata.org/docs/`
[3] `https://numpy.org/doc/stable/`
[4] `https://scikit-learn.org/`

- **SciPy**[5] **(≥1.10.0)** A scientific computing library that extends NumPy, providing advanced mathematical functions for optimization, integration, interpolation, and statistics.
- **Plotly**[6] **(≥5.14.0)** A graphing library used to create interactive and publication-quality visualizations such as line charts, scatter plots, bar charts, heatmaps, and more. It serves as the primary engine for rendering visual elements of the dashboard.
- **OpenAI**[7] **(≥1.0.0)** The OpenAI Python SDK, used for integrating large language models (LLMs) for optional AI-assisted functionalities such as column renaming, explanation generation, or metric labeling.
- **Word2Number**[8] **(≥1.1.0)** A utility library used to convert written numbers into numeric form, helpful in the data preprocessing phase for standardizing inconsistent formats.
- **VS Code**[9] A lightweight but powerful source-code editor widely adopted by data science professionals for Python development.

3.3 Data Preparation Process

3.3.1 Data Readiness Assessment

According to [10], the establishment of standardized and consistent approaches to data validation is paramount in ensuring data quality. Data quality frameworks serve as the methodological backbone to evaluate and enhance the integrity of data.

In CSViz, we implemented a comprehensive data quality assessment framework that evaluates multiple dimensions of data integrity through a systematic scoring algorithm. The assessment calculates a composite "Data Readiness Score" by analyzing five critical quality dimensions weighted according to their analytical importance.

These dimensions were developed based on a checklist provided by [11]:

Completeness (25% of total score)

Let's define:

- n = total number of cells in the dataset

[5] `https://docs.scipy.org/doc/scipy/`
[6] `https://plotly.com/python/`
[7] `https://platform.openai.com/docs/overview`
[8] `https://w2n.readthedocs.io/en/latest/`
[9] `https://code.visualstudio.com/docs`

- m = number of missing values

The completeness formula is:

$$Completeness = 25\% \left(1 - \frac{m}{n}\right)$$

This rewards datasets with fewer missing values. A dataset with no missing values would receive the full 25% contribution to the final score.

Consistency (20% of total score)

Let's define:

- r = total number of rows
- dr = number of duplicate rows
- c = total number of columns
- dc = number of duplicate column names

The consistency formula is:

$$Consistency = 20\% \left(1 - \frac{dr}{r} - 0.05dc\right)$$

This penalizes both duplicate rows (as a percentage) and duplicate column names (5% penalty per duplicate).

Validity (15% of total score)

Let's define:

- o = number of outlier values
- n = total number of values
- ti = number of text formatting issues
- tt = total number of text values

The validity formula is:

$$Validity = 15\% \left(1 - \frac{o}{n} - \frac{ti}{tt}\right)$$

This accounts for both statistical outliers in numeric columns and formatting inconsistencies in text fields.

Data Type Correctness (15% of total score)

Let's define:

- ci = number of columns with inappropriate data types
- c = total number of columns

The data type correctness formula is:

$$DataCorrectness = 15\% \left(1 - \frac{ci}{c}\right)$$

This penalizes columns where numeric data is stored as text or other inappropriate data type assignments.

Data Format Consistency (25% of total score)

Let's define:

- cm = number of columns with mixed data types
- c = total number of columns
- A column is considered "mixed" if between 20 and 80% of its values are numeric

The data format consistency formula is:

$$DataConsistency = 25\% \left(1 - 0.15cm\right)$$

Each column with mixed data types reduces this component score by 15%.

Final Score Calculation

Let's define:

$S_{base} = Completeness + Consistency + Validity + DataCorrectness + DataConsistency$

h = number of high severity issues (capped at 5)

M = boolean indicator of mixed data presence (1 if mixed data exists, 0 otherwise)

The final score calculation is:

$$S_{final} = \min(\max(S_{base} - 5 * h, 0), M * 75\% + (1 - M) * 100\%)$$

$$S_{final} \in [0, 100]\% \tag{3.1}$$

This applies the high severity penalties (5 points each, up to 25 points) and implements the mixed data cap that limits the maximum score to 75% if any mixed data columns exist.

3.3.2 Data Processing

Manual Processing Techniques

A specialized module was created to enable users to combine and merge several CSV files with little technical work as part of the manual data preparation procedure. In actual industrial settings, where data frequently originates from diverse sources and formats, this preprocessing phase is essential [12].

Concatenation and merging are the two main features of the developed tool [13].

In the concatenation tab as shown in Figure A.1 , depending on the format of the datasets, users can select between vertical or horizontal concatenation when uploading several CSV files. A column mapping system enables users to standardize column names across files for vertical concatenation, guaranteeing structural compatibility. Common columns are automatically identified during horizontal concatenation and can be renamed to prevent conflicts.

In the merging tab as shown in Figure A.1, users can pick any two datasets and configure join operations based on one or more key pairs. There is support for inner, left, right, and exterior joins, among other kinds. To resolve overlapping column names, users can also specify suffixes. Every operation's outcome is shown in real time, along with the opportunity to download the processed dataset straight away.

For non-technical users, this interface offers a low-code preparation layer to get datasets ready for visualization. In addition to increasing accessibility, it guarantees structural integrity and data alignment, both of which are necessary for creating trustworthy and understandable dashboards.

AI-Assisted Preprocessing Methods

The wide use of AI and Large Language Models (LLMs) is fundamentally changing the data science world, making it imperative that we add an AI preprocessing mode where users with limited to no knowledge about data processing would be able to clean, transform,

and prepare their data for analysis with minimal effort [14].

CSViz incorporates a powerful AI preprocessing mode that connects directly with OpenAI's API to handle complex data manipulation tasks automatically. This feature represents a significant advancement in democratizing data preprocessing by removing any technical barriers that traditionally required specialized knowledge.

Behind the scenes, our system constructs a detailed prompt that guides the AI's processing approach, the prompt is shown in 4.6

By incorporating this AI preprocessing capability, we've significantly lowered the technical threshold for meaningful data analysis, making powerful analytical capabilities accessible to business users, researchers, and decision-makers regardless of their technical background [15].

3.3.3 Data Cleaning

Manual Cleaning Techniques

Data cleaning is a crucial step that ensures accuracy, consistency, and usability of datasets before visualization[16]. A specialized data cleaning module in our web application addresses common data quality problems that could compromise dashboard clarity and analytical reliability. The module offers several configurable operations to handle specific data anomalies:[17]:

- **Textual Cleaning and Standardization**: Users can delete special characters, standardize spacing and capitalization across object-type columns, and reduce white spaces to improve consistency in categorical variables, enhancing the visualizations' ability to group or filter data.
- **Column Splitting**: Users can divide compound columns into distinct attributes based on a selected delimiter, with options to keep or remove the original column and use custom values for any gaps.
- **Column Renaming and Dropping**: Users can manually rename columns or drop columns to improve clarity or conform to naming conventions.
- **Data Type Transformation**: Users can convert data types for selected columns, with a helper mechanism facilitating the parsing of numeric values from textual input.
- **Outlier Detection and Treatment** The module supports:
    - Interquartile Range (IQR) clipping, which caps values beyond the 1.5IQR threshold.

    - Winsorization which limits values outside the 5th–95th percentile range.
- **Missing Value Handling**: Strategies include deletion of rows with nulls, imputation using mean/median/mode, and sequential interpolation for numeric time-series data, applied on a column-by-column basis.
- **Duplicate Row Removal**: The module identifies and removes duplicate rows to prevent data overrepresentation.
- **Column Reordering**: The module allows columns to be reordered manually through a validated editor interface to enhance logical structure and readability.

AI-Assisted Cleaning Methods

Data cleaning has traditionally been a labor-intensive task, often requiring specialized knowledge. However, the advent of LLMs has introduced tools that automate this process, making data preparation more accessible. For instance, DANGO utilizes LLMs to generate context-aware models for tabular data cleaning, effectively identifying and rectifying data inconsistencies [18].

In CSViz, we implemented an AI-powered data cleaning mode that operates by securely transmitting datasets to OpenAI's API, accompanied by specific instructions in a prompt detailing the cleaning requirements.

The prompt used with OpenAI's API is shown in 4.6.

CSViz AI cleaning mode significantly enhances the ability of users with minimal technical expertise to efficiently prepare datasets for analysis.

3.4 Dashboard Generation Approach

3.4.1 Metrics Recommendation System

A custom recommendation system was created and incorporated into the dashboard generation pipeline to automate the process of identifying pertinent metrics in industrial datasets. The most statistically and semantically significant numerical columns can be chosen by this system to be highlighted as Key Performance Indicators (KPIs) after tabular data analysis. Rule-based logic, statistical transformations, and domain-specific semantic enrichment are the foundations of the system. Other than the initial dataset upload and domain selection, it is completely automated and doesn't require any human intervention.

Metric Selection Logic

Both data-driven analysis and semantic interpretation are essential components of the metric selection logic. All column names are first cleaned and standardized by the system. This entails formatting tasks like deleting special characters and changing camelCase or snake_case to lowercase with spaces.

This preprocessing step ensures that the semantic matching component functions correctly across datasets with inconsistent naming conventions. Each numerical column is evaluated based on a set of quantifiable characteristics that capture distributional, structural, and informational properties. The table below summarizes the features used for ranking metrics:

Characteristic	**Description**	**Normalized Formula**	**Weight**
Coefficient of Variation (CV)	Measures relative variability of a column; higher values indicate dynamic metrics	$Score_{cv} = \tanh\left(\frac{\sigma}{\mu}\right)$	1.5
Skewness	Indicates the asymmetry of the data distribution	$Score_s = \tanh\left(\frac{\lvert\text{Skewness}\rvert}{10}\right)$	1.2
Kurtosis	Measures the tailedness or extremity of outliers in the distribution	$Score_k = \tanh\left(\frac{\lvert\text{Kurtosis}\rvert}{15}\right)$	1.0
Entropy	Captures information richness based on value frequency	$Score_e = \frac{-\sum p_i \log_2(p_i)}{\log_2(n+1)}$	0.8
Uniqueness Ratio	Measures the proportion of unique values to total entries	$Score_u = \sqrt{\frac{Numberofuniquevalues}{TotalEntries}}$	0.7
Outlier Proportion	Measures the percentage of extreme values (beyond $\pm 3\sigma$)	$Score_o = \frac{NumberofOutrliers}{TotalEntries}$	0.6

σ**:** The standard deviation of the values in the column, μ**:** the mean of the column values, n**:** The number of non-empty bins created when the numerical column is discretized, pi**:** The proportion (probability) of entries falling into the i-th bin of a discretized version of the numerical column.

Table 3.1: Metric Ranking Features

These weights were empirically derived to balance statistical informativeness and stability. In addition to quantitative metrics, the system incorporates semantic keyword matching through a domain-specific scoring model. The user selects a relevant domain (e.g., Sales, Logistics, Production), and the system checks column names for the presence of keywords defined in a domain dictionary. Each keyword has a relevance weight (1.5 or 2.0), which is added to the final column score as a Business Relevance Boost.

The final metric score is the weighted sum of normalized statistics and the semantic boost. Contextual significance and data relevance are both guaranteed by this hybrid approach. The system also includes logic to avoid common pitfalls:

- **ID Column Detection**: Automatic filtering is applied to columns that contain phrases like "id," "code," or "reference."
- **Missing Value Check**: A tiered penalty system is applied based on the proportion of missing values in each column: no penalty for $\leq$5%, moderate penalty for 5–20%, severe penalty for 20–50%, and exclusion or manual review for $>$ 50%.

To synthesize all evaluation components into a single interpretable metric, the final score for each column is computed as follows:

$$\text{Final Score} = 1.5 \cdot \text{Score}_{cv} + 1.2 \cdot \text{Score}_{s} + 1.0 \cdot \text{Score}_{k} + 0.8 \cdot \text{Score}_{e} + 0.7 \cdot \text{Score}_{u} + 0.6 \cdot \text{Score}_{o}$$

Automated Recommendation Process

Following the evaluation and scoring of each number column, the system moves on to the recommendation phase. The outcomes are arranged according to the final score in descending order.

- **Top-3 Metric Selection:**
  - The three columns with the highest scores are chosen as the main KPIs automatically.
  - Each is associated with a recommended aggregation method (mean, sum, or count) using a custom rule-based aggregation function:
    - *Count:* for low-cardinality or ID-like columns
    - *Sum:* for columns with names containing "total", "sum", or mostly positive numeric values
    - *Mean:* for columns containing terms such as "price", "rate", "score"
- **Alternative Metrics Interface:**
  - The next five best-performing columns are displayed as alternatives.

- A dropdown menu allows users to choose which of the top three KPIs to replace.
- Each replacement triggers a recalculation of the selected metric's value using its recommended aggregation method.

This hybrid approach of automation and light interactivity empowers users to trust the default recommendations while still offering flexibility to adapt the dashboard content. It significantly reduces the time and expertise required to identify meaningful quantitative indicators within complex industrial datasets.

All things considered, this metrics recommendation engine marks a significant breakthrough in enabling dashboard creation to be more dependable, accessible, and in line with analytical needs particular to a given domain.

3.4.2 Visualization Recommendation System

Visualization Scoring System

In CSViz, we proposed a Heuristic Data Visualization Scoring System similar to [8] that systematically evaluates data columns and their combinations to identify elements that yield the most meaningful and insightful visualizations. The system assesses individual columns, column pairs, column triples, and GroupBy operations, assigning scores based on specific criteria ranging from 0 to 10 points. These scores are weighted and combined to produce final recommendations.

Our system evaluates data elements at four levels of granularity:

- Individual columns, Column pairs, Column triples and GroupBy operations.

Individual Column Scoring

We assess each column across seven dimensions (0-10 points each):

- **Distribution Characteristics:** Evaluates visualization potential based on:
  - Coefficient of variation, skewness, and kurtosis
  - Multimodality, entropy, and time range (for temporal data)
- **Data Type Scoring:** Rates columns based on their data type:
  - Numerical: Higher scores for wider ranges
  - Categorical: Optimal scores for moderate cardinality ($\sim$10)
  - Temporal: Higher scores for spanning multiple time units
- **Data Quality:** Measures cleanliness based on:
  - Completeness (missing value ratio)

    - Outlier presence
    - Unique value ratio
- **Predictive Power:** Measures relationships with other columns:
    - Correlation for numerical
    - Cramer's V for categorical
- **Semantic Content Analysis:** Detects column purpose:
    - Penalizes ID columns
    - Rewards metric and descriptor columns
- **Dimensional Analysis:** Evaluates the role in dimensional modeling:
    - Fact/Measure columns (higher scores for measurements)
    - Dimension columns (higher for categorical with moderate cardinality)
    - Date dimensions (bonus for temporal data)
- **Variance Information Ratio:** Measures information density

The total score for individual column scoring is calculated with the formula:

$$\begin{aligned}\text{column\_score} = {} & 0.20 \cdot \text{distribution\_score} + 0.15 \cdot \text{data\_type\_score} \\ & + 0.10 \cdot \text{data\_quality\_score} + 0.15 \cdot \text{predictive\_power\_score} \\ & + 0.20 \cdot \text{semantic\_content\_score} + 0.10 \cdot \text{dimensional\_analysis\_score} \\ & + 0.10 \cdot \text{variance\_info\_ratio\_score}\end{aligned}$$

Column Pair Scoring

Column pairs are scored on:

- **Statistical Association:** Strength of relationship
    - Using correlation, Cramer's V, or Eta^2, depending on data types
- **Visualization Complexity:** Evaluates visual clarity
- **Pattern Detection:** Identifies interesting patterns
    - Cluster detection for numerical pairs
    - Trend and seasonality for time series
- **Anomaly Highlighting:** Measures the presence of interesting outliers
- **Information Complementarity:** Evaluates how columns enhance understanding
    - Penalizes ID columns (-7 points)
- **Redundancy Penalization:** Detects when columns show similar information
    - Penalizes highly correlated or dependent columns
- **Practical Utility Score:** Scores visualization of usefulness
    - Rewards high-value combinations (categorical + numerical)

- Penalizes low-value columns (IDs, timestamps)

The total score for pair column scoring is calculated with the formula:

$$\begin{aligned}\text{pair\_score} = {} & 0.15 \cdot \text{statistical\_association} + 0.10 \cdot \text{visualization\_complexity} \\ & + 0.15 \cdot \text{pattern\_detection} + 0.05 \cdot \text{anomaly\_highlighting} \\ & + 0.15 \cdot \text{information\_complementarity} + 0.10 \cdot \text{redundancy\_penalization} \\ & + 0.30 \cdot \text{practical\_utility\_score}\end{aligned}$$

Column Triple and GroupBy Scoring

Column Triples are evaluated on:

- Dimensionality appropriateness for 3D visualization
- Pattern richness
- Complexity balance
- Visual differentiability
- Cognitive load

The total score for column triples scoring is calculated with the formula:

$$\begin{aligned}\text{triple\_score} = \frac{1}{5} \cdot \big(&\text{Dimensionality Appropriateness} + \text{Pattern Richness} \\ &+ \text{Complexity Balance} + \text{Visual Differentiability} \\ &+ \text{Cognitive Load}\big)\end{aligned}$$

GroupBy Operations are scored on:

- GroupBy column suitability (cardinality, distribution)
- Aggregation column quality (numeric properties)
- Pair effectiveness (differentiation, balance)

The total score for GroupBy operations columns scoring is calculated with the formula:

$$\begin{aligned}\text{groupby\_score} = {} & 0.40 \cdot \text{group\_differentiation} + 0.15 \cdot \text{group\_balance} \\ & + 0.25 \cdot \text{aggregation\_meaningfulness} + 0.20 \cdot \text{visualization\_potential}\end{aligned}$$

Final Recommendation Selection

- **Top selected columns:** The system uses weighted averages of dimension scores to select the top 5 visualization recommendations:
  - Top 2 individual columns

- Top 2 column pairs
- Top 1 column triple
- Top 1 groupby-aggregation pair

- **Alternative Visualisations Interface:**
  - The system maintains a secondary repository of the next five highest-performing column visualizations, accessible as alternatives to the primary recommendations.
  - Users can interact with a dropdown menu interface that facilitates the substitution of any of the primary visualizations with these alternatives.

3.4.3 Decision Tree Methodology

Our methodology employs a decision tree algorithm inspired by [19] to optimize visualization recommendations based on data characteristics. The system analyzes column properties to intelligently suggest appropriate visualization types across multiple scenarios: single columns, column pairs, column triples, and grouped data aggregations. This rule-based approach evaluates factors including data types (temporal, numeric, categorical), cardinality (unique value counts), statistical distributions, and relational properties to identify the most informative visual representation.

3.4.4 Dashboard Display

Dynamic Metric Display

Using the Streamlit metric() component, the selected metrics are dynamically shown on the dashboard. This includes a tooltip displaying the column's final score as well as the metric name, value (aggregated), and method.

Visualisations Display

The visualization rendering employs Plotly as the primary visualization library and organizes the visual content in a structured $4 * 4 + 1$ dashboard format. This configuration was specifically designed to optimize information processing capabilities for users. The determination of presenting five plots was informed by [20], which established that humans typically have a working memory capacity of seven items, plus or minus two.

Chapter 4

Results and Evaluation

In this chapter we will present the implementation of the dashboard creation process from real world data. This chapter also provides a comprehensive analysis of the results obtained through the application of the system.

4.1 Use Case Analysis

We evaluated CSViz using a synthetic dataset modeled after production characteristics observed during our internship at a Izidhar factory, generated with Python's pandas and numpy libraries. The dataset contained 520 observations over 31 days in March 2025, including variables of different types such as production date, product type, packaging format, production quantity, unit economics, and quality control parameters as shown in the dataset sample in Figure D.1.

As illustrated in Figure 4.1 ,the CSViz dashboard presents:

- Three KPI cards showing average production quantity, total revenue, and total cost.
- A histogram displaying production volume distribution.
- A scatter plot matrix revealing the relationship between revenue and production.
- Bar charts showing distribution by product type and packaging format.
- A line chart tracking production quantity over time.

These visualizations enable multi-perspective analysis of production processes, supporting operational responsiveness and continuous improvement aligned with Industry 4.0 principles.

Figure 4.1: Generated Dashboard

4.2 Usability and Functionality Evaluation

In order to asses the usability of the CSViz ,we conducted a stuctured survey with closed ended questions with several faculty teachers from diffrent domains: mathematics, computer science, mechanical engineering, and electrical engineering. Participants completed specific tasks and provided anonymous feedback.

The results are summrized in Figure 4.2

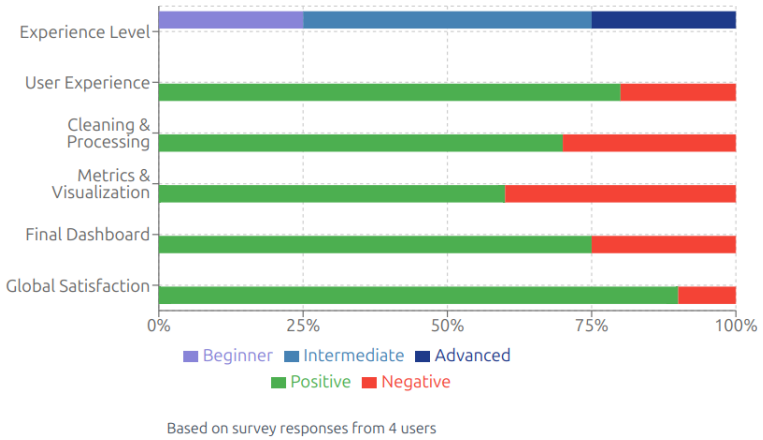

Figure 4.2: CSViz User Assessment Results

0. Preliminary Experience: Respondents had varying levels of data tool experience, with most at intermediate level. All used Excel, while some used tools like R, Python, and Tableau across academic, professional, and educational contexts.

1. User Experience : Respondents rated the application favorably, agreeing it was simple to use with a pleasing interface. The cleaning and visualization process was deemed seamless with acceptable loading times.

2. Cleaning and Processing : The CSV cleanliness detection was considered practical, with well-received color-coding. Users appreciated both AI and manual cleaning options and found the interface user-friendly.

3. Metrics and Visualization Accuracy : This area received more mixed responses. While domain selection was clear and metrics replacement was straightforward, some users requested more flexibility in visualization options.

4. Final Dashboard : The dashboard functionality received positive reviews for its structure, data comprehension support, and decision-making assistance. Users appreciated the professional display and expressed willingness to reuse the tool.

5. Global Satisfaction : Overall satisfaction was high, with users indicating they would recommend CSViz. Suggested improvements included visualizing errors at the data level, adding warning dialogs before corrections, implementing "undo" functionality, and adding export capabilities.

4.3 Discussion

CSViz successfully streamlines dashboard generation for non-expert users in Industry 4.0 environments through its hybrid logic combining statistical profiling, semantic domain modeling, and visualization rules.

Metric Prioritization and Visualization Relevance

The recommendation engine prioritized analytically significant metrics aligned with industrial KPIs, accurately identifying production quantity, revenue, and cost as high-priority metrics. The visualization engine assigned appropriate graph types based on data distributions and context, ensuring dashboards were both statistically sound and actionable.

Usability and Workflow Efficiency

User feedback highlighted the system's accessibility through guided preprocessing, transparent metric substitution, and balanced automation. Participants with limited experience successfully created dashboards without requiring coding or statistical expertise.

Methodological Coherence

The end-to-end pipeline demonstrated transferability across domains through techniques like ID column filtering, missing-value penalty systems, and automated aggregation suggestions. The rule-based components ensured interpretable logic, avoiding "black-box" pitfalls.

CSViz achieves its goal of lowering technical barriers to dashboard generation by integrating data-driven automation with domain-aware adaptability, positioning it as a practical tool for non-expert analytics in smart manufacturing.

4.4 Strengths of the approach

Streamlit has emerged as a powerful open-source Python library specifically engineered for the swift creation and deployment of interactive data applications. This capability revolutionizes how data-driven insights are visualized and interacted with, offering significant advantages over traditional web development frameworks. [21] Furthermore, Streamlit simplifies the deployment process immensely, allowing users to effortlessly share their applications either locally within their environment or globally on the internet with just a few commands, abstracting away the complexities typically associated with web application hosting.

Also CSViz uses **CSV** format, which is known for its remarkable compatibility and portability across diverse technological environments further solidify its position as a preferred data input format. [22].

A significant strength of CSViz lies in its provision of both **AI-powered and manual** options for data processing and cleaning. This dual approach empowers users with flexibility and control. [23].

The integration of **domain context** into the process of selecting relevant metrics for data visualization represents a significant strength of CSViz. [24].

The use of **heuristic scoring** for both metric and visualization selection across various data

combinations, represents a powerful approach to automating the discovery of potentially informative visualizations. [25].

Dashboards are powerful tools for the efficient communication of key findings and insights to a broader audience, including stakeholders who may not be deeply involved in the data analysis process.[6].

4.5 Limitations and constraints

The prototype nature of CSViz showed many limitations and constraints, which will be discussed in this section.

While Streamlit offers significant advantages for prototyping, it also presents several limitations when considering the development of production-ready software [26]. One notable constraint is performance, particularly when dealing with larger datasets. The way Streamlit handles UI rendering can lead to slower loading and running times for applications with extensive data or complex computations.

Beyond the limitations of Streamlit itself, Python, as the underlying programming language, also presents certain challenges in large-scale, production-ready applications. One significant aspect is performance. As an interpreted language, Python is generally slower than compiled languages like C++ or Java [27].

Another drawback lies in the dependence on OpenAI's API for AI cleaning and processing modes. Despite the advantages, relying on external LLM APIs like OpenAI for core functionalities also introduces several potential drawbacks and constraints. [28] This dependence carries inherent risks related to the API's availability, potential changes in pricing structures, and the possibility of service disruptions.

4.6 Potential Improvements

One potential improvement is developing a custom AI model for data cleaning and processing, as opposed to relying solely on an external LLM API, [28] evaluates the shift from proprietary LLM APIs to open-Source models, While acknowledging that proprietary LLMs like GPT-4 offer cutting-edge capabilities and ease of use [29], the research highlights several advantages of SLMs (Small Language Model) that make them viable and, in some cases, more cost-effective alternatives.

Another potential improvement that we received from feedback is implementing the

ability to export dashboards to PDF and Excel formats, which would significantly enhance the usability and collaborativeness of the application [30].

In conclusion, Chapter 4 has illustrated the practical application of the dashboard creation process using real-world data. It has evaluated the system's performance in terms of user experience, cleaning and processing, metrics and visualizations accuracy and the global satisfaction. The chapter also provided a critical discussion of the system's strengths and limitations, offering insights into potential directions for future improvements.

General Conclusion

We present CSViz, a Streamlit application to generate insight-driven dashboards. To develop the app, we combined a rule-based metric selection engine with semantic domain knowledge to recommend relevant KPIs and visualizations, automatically tailored to the user's dataset. Along with a decision tree framework to direct visualization choices based on data types and statistical properties, the system includes both AI-assisted and manual modules for data preprocessing and cleaning. Because of these accessibility-focused features, CSViz is particularly well-suited for users with little technical experience or knowledge of data science. Finally, we demonstrated the effectiveness of our app through a user study where several professors of our school tested the app and gave feedback about their experience. Further research can be conducted to develop a proprietary AI model for data cleaning and processing. Moreover, a dashboard export function could be added to enhance the usability and value of the application.

Bibliography

[1] Lane Thames and Dirk Schaefer. "Industry 4.0: an overview of key benefits, technologies, and challenges". In: *Cybersecurity for Industry 4.0: Analysis for Design and Manufacturing* (2017), pp. 1–33.

[2] Rongyan Zhou and Julie Le Cardinal. "Exploring the impacts of industry 4.0 from a macroscopic perspective". In: *Proceedings of the Design Society: International Conference on Engineering Design*. Vol. 1. 1. Cambridge University Press. 2019, pp. 2111–2120.

[3] Alexandros Bousdekis et al. "A review of data-driven decision-making methods for industry 4.0 maintenance applications". In: *Electronics* 10.7 (2021), p. 828.

[4] Praveen Soni et al. "A survey on automatic dashboard recommendation systems". In: *Visual Informatics* 8.1 (2024), pp. 67–79.

[5] Qiyue Zhang. "The Impact of Interactive Data Visualization on Decision-Making in Business Intelligence". In: *Advances in Economics, Management and Political Sciences* 87 (June 2024), pp. 166–171. DOI: 10.54254/2754-1169/87/20241056.

[6] Dazhen Deng et al. "DashBot: Insight-driven dashboard generation based on deep reinforcement learning". In: *IEEE Transactions on Visualization and Computer Graphics* 29.1 (2022), pp. 690–700.

[7] Kevin Hu et al. "Vizml: A machine learning approach to visualization recommendation". In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019, pp. 1–12.

[8] Jock Mackinlay, Pat Hanrahan, and Chris Stolte. "Show me: Automatic presentation for visual analysis". In: *IEEE transactions on visualization and computer graphics* 13.6 (2007), pp. 1137–1144.

[9] Antonis Protopsaltis et al. "Data Visualization in Internet of Things: Tools, Methodologies, and Challenges". In: Aug. 2020. DOI: 10.1145/3407023.3409228.

[10] Russell Miller et al. "A Framework for Current and New Data Quality Dimensions: An Overview". In: *Data* 9.12 (2024), p. 151.

[11] DAMA UK Working Group. *The Six Primary Dimensions for Data Quality Assessment*. Published by the Washington State Board for Community and Technical Colleges. 2013. URL: https://www.sbctc.edu/resources/documents/colleges-staff/commissions-councils/dgc/data-quality-deminsions.pdf.

[12] Alvaro AA Fernandes et al. "Data preparation: A technological perspective and review". In: *SN Computer Science* 4.4 (2023), p. 425.

[13] Pramod Gupta and Anupam Bagchi. *Essentials of Python for Artificial Intelligence and Machine Learning*. Springer, 2024.

[14] Haochen Zhang et al. "Large language models as data preprocessors". In: *arXiv preprint arXiv:2308.16361* (2023).

[15] Elyas Meguellati et al. "Are Large Language Models Good Data Preprocessors?" In: *arXiv preprint arXiv:2502.16790* (2025).

[16] Hui Yang. "Data preprocessing". In: *Pennsylvania State University: Citeseer* (2018).

[17] Stefanie Molin. *Hands-On Data Analysis with Pandas: A Python data science handbook for data collection, wrangling, analysis, and visualization*. Packt Publishing Ltd, 2021.

[18] Wei-Hao Chen et al. "Dango: A Mixed-Initiative Data Wrangling System using Large Language Model". In: *arXiv preprint arXiv:2503.03154* (2025).

[19] Yan Holtz. *From Data to Viz*. Accessed: 2025-05-06. 2025. URL: https://www.data-to-viz.com/.

[20] George A Miller. "The magical number seven, plus or minus two: Some limits on our capacity for processing information." In: *Psychological review* 63.2 (1956), p. 81.

[21] Alex Merced. *Deep Dive into Data Apps with Streamlit*. Accessed: 2025-05-06. 2024. URL: https://dev.to/alexmercedcoder/deep-dive-into-data-apps-with-streamlit-3e40.

[22] Samuel Shaibu. *CSV vs Excel: Making the Right Choice for Your Data Projects*. Accessed: 2025-05-06. 2024. URL: https://www.datacamp.com/blog/csv-vs-excel.

[23] Ki Hyun Tae et al. "Data cleaning for accurate, fair, and robust models: A big data-AI integration approach". In: *Proceedings of the 3rd international workshop on data management for end-to-end machine learning*. 2019, pp. 1–4.

[24] GeeksforGeeks. *Role of Domain Knowledge in Data Science*. Accessed: 2025-05-06. 2025. URL: https://www.geeksforgeeks.org/role-of-domain-knowledge-in-data-science/.

[25] Emily Wall et al. “A heuristic approach to value-driven evaluation of visualizations”. In: *IEEE transactions on visualization and computer graphics* 25.1 (2018), pp. 491–500.

[26] Restack. *Streamlit Limitations Overview*. Accessed: 2025-05-06. 2024. URL: https://www.restack.io/docs/streamlit-knowledge-streamlit-limitations.

[27] MO Balogun. “Comparative analysis of complexity of C++ and Python programming languages”. In: *Asian J. Soc. Sci. Manag. Technol* 4.2022 (2022), pp. 1–12.

[28] Chandra Irugalbandara et al. “Scaling down to scale up: A cost-benefit analysis of replacing OpenAI’s LLM with open source SLMs in production”. In: *2024 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. IEEE. 2024, pp. 280–291.

[29] OpenAI. *Introducing ChatGPT and Whisper APIs*. Accessed: 2025-05-06. 2024. URL: https://openai.com/index/introducing-chatgpt-and-whisper-apis/.

[30] Mohamed Mohamed, Tarek Aly, and Mervat Gheith. “Python Based End User Computing Framework to Empowering Excel Efficiency”. In: *International Journal for Research in Applied Science and Engineering Technology* 12 (Apr. 2024), pp. 2719–2729. DOI: 10.22214/ijraset.2024.60097.

Appendix A

Processing Page Screenshot

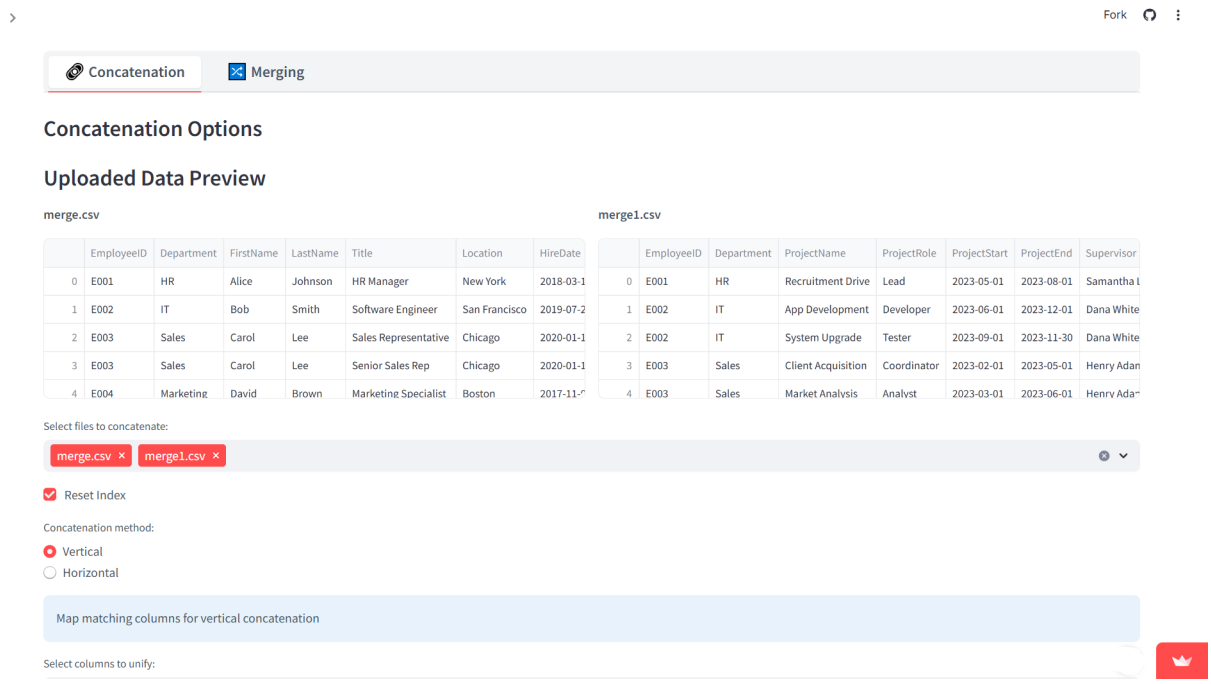

Figure A.1: Manual processing page in the application.

Appendix B

Prompts sent to OpenAI's API

Processing prompt:

> "TASK:
> 1. Analyze these datasets and determine the best way to combine them:
> - Vertical concatenation (stacking datasets with similar columns)
> - Horizontal concatenation (joining datasets side by side)
> - Merging on common keys
> 2. Apply the best combination method intelligently:
> - For vertical concatenation: Map similar columns across datasets
> - For horizontal concatenation: Handle column conflicts
> - For merging: Identify appropriate join keys and join type
> 3. Return ONLY the combined CSV data with no explanations, JSON, or other text."

AI-Cleaning prompt:

> "REQUIREMENTS:
> 1. Clean this dataset thoroughly by:
> - Removing duplicate rows
> - Handling ALL missing/null values (nothing should be left empty)
> - Converting columns to the most appropriate data types
> - Converting text number representations (like "forty", "twenty-five") to actual numbers
> - Standardizing date formats to be consistent
> - Standardizing categorical values (e.g., gender: 'male'/'female' instead of 'm'/'f')

- Fixing any inconsistent text formatting (case, extra spaces, special characters)
- Removing or handling outliers in numeric columns when appropriate
2. Return ONLY the cleaned CSV data in the following format:

column1,column2,column3,...
value1,value2,value3,...
...
3. DO NOT include any explanations, code, or other text - ONLY return the cleaned CSV data.
4. If the dataset is too large, clean the entire dataset using the same logic you would apply to the sample."

Appendix C

Decision tree diagram: Data To viz

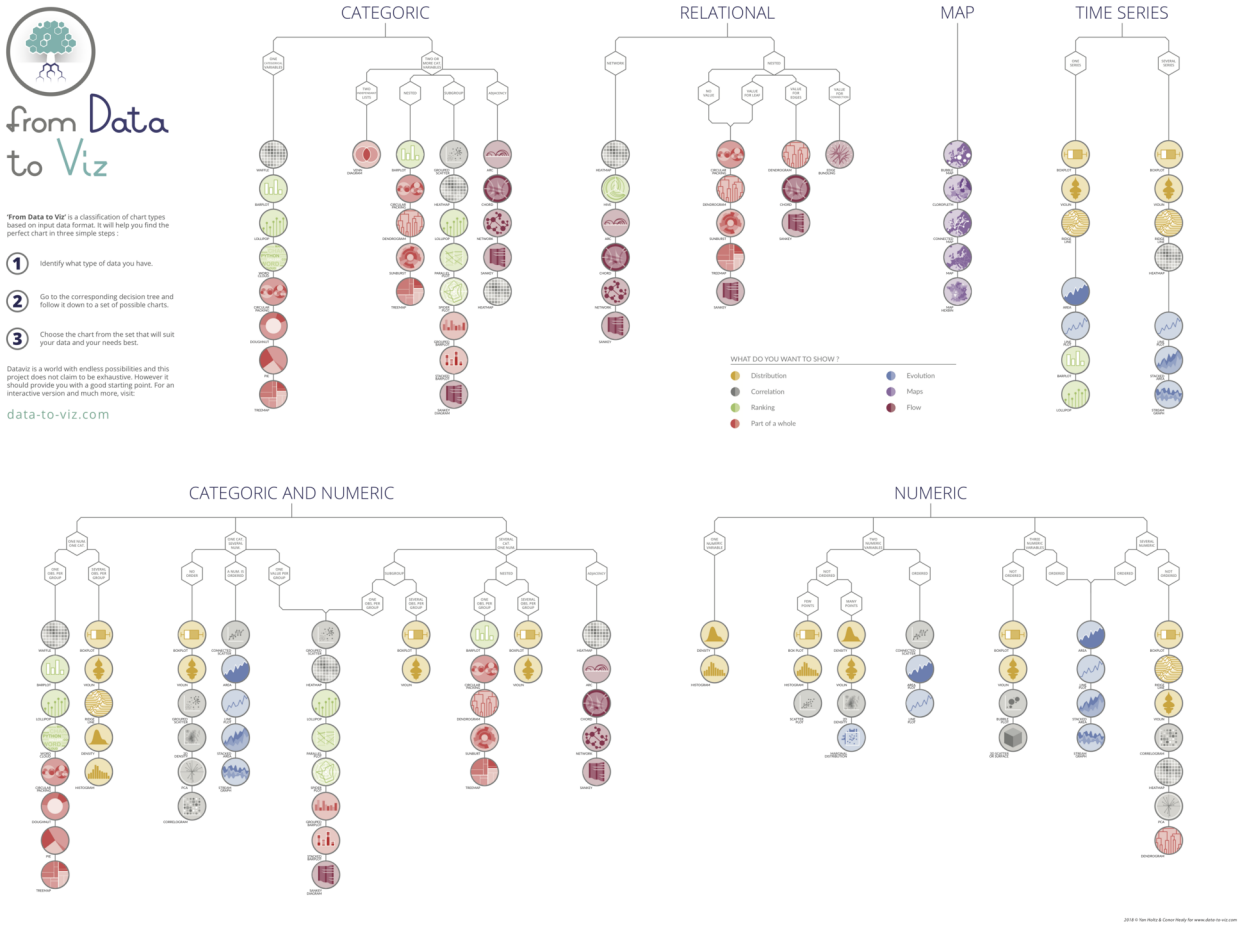

Figure C.1: Decision tree

Appendix D

Dataset Sample

Brand	Model	Year	Engine_Size	Fuel_Type	Transmission	Mileage	Doors	Owner_Count	Price
Kia	Rio	2020	4.2	Diesel	Manual	289944	3	5	8501
Chevrolet	Malibu	2012	2	Hybrid	Automatic	5356	2	3	12092
Mercedes	GLA	2020	4.2	Diesel	Automatic	231440	4	2	11171
Audi	Q5	2023	2	Electric	Manual	160971	2	1	11780
Volkswagen	Golf	2003	2.6	Hybrid	Semi-Automatic	286618	3	3	2867
Toyota	Camry	2007	2.7	Petrol	Automatic	157889	4	4	7242
Honda	Civic	2010	3.4	Electric	Automatic	139584	3	1	11208
Kia	Sportage	2001	4.7	Electric	Semi-Automatic	157495	2	2	7950
Kia	Sportage	2014	2.6	Hybrid	Manual	98700	3	4	9926
Toyota	RAV4	2005	3.1	Petrol	Manual	107724	2	5	6545

Figure D.1: Dataset Sample

Appendix E

CSViz Logo

Figure E.1: CSViz Logo.

Appendix F

CSViz Business Model Canvas

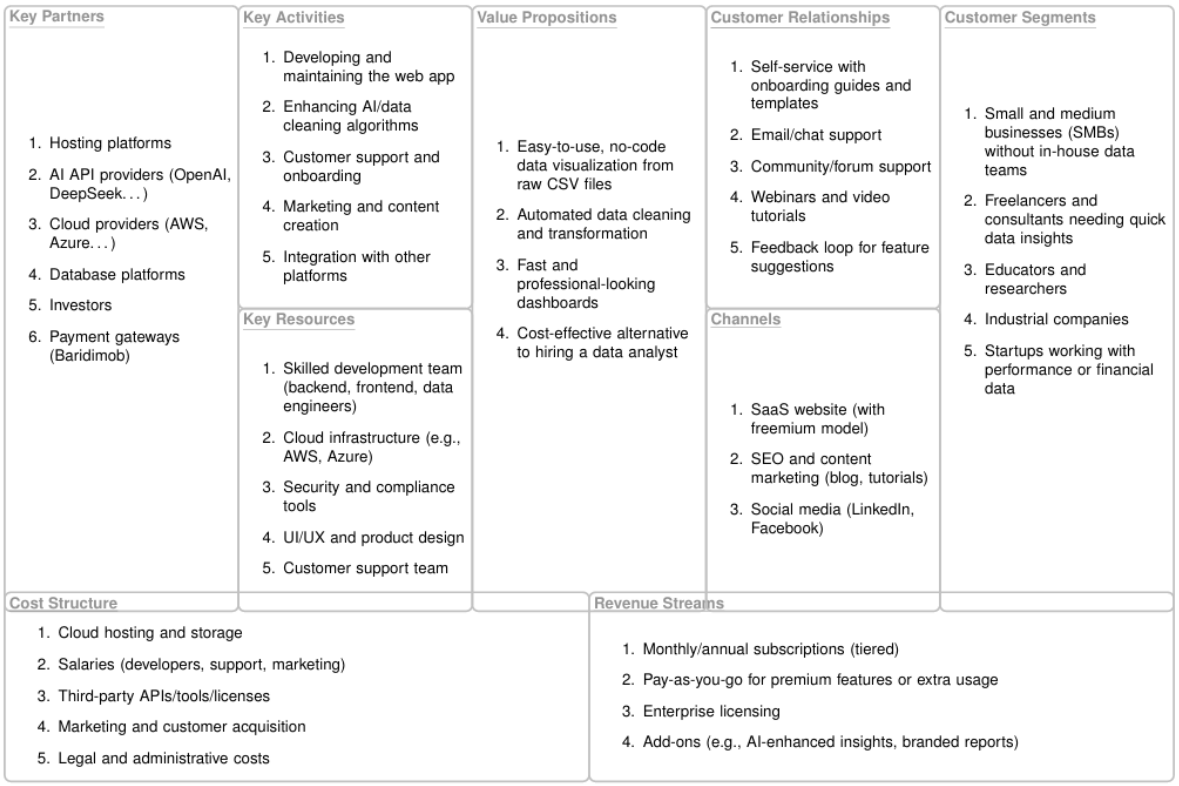

Figure F.1: Business Model Canvas.

NATIONAL HIGHER SCHOOL OF TECHNOLOGY AND ENGINEERING

BUSINESS PLAN

CSViz

2025

Presented by:

Abderahim REDOUANE

Malak EL MEMI

BUSINESS PLAN - CSVIZ

Project Idea

We propose to develop CSViz, a B2B SaaS platform that democratizes data analysis and visualization for Industry 4.0 stakeholders. Our solution fully automates the generation of interactive dashboards from CSV files, eliminating technical barriers that prevent engineers, managers, and SMEs from effectively exploiting their data.

Concretely, CSViz combines an intelligent preprocessing system, a metrics recommendation engine, and a visualization system powered by AI. Users simply upload their CSV file, and our platform automatically generates relevant visualizations, cleans data, recommends metrics based on statistical and semantic relevance, and proposes optimal visualizations through heuristic and AI-based scoring systems.

Our Core Values

- **Accessibility**: Making data analysis accessible to non-experts without requiring advanced data science skills or complex BI tools.
- **Simplicity**: Offering a responsive web interface with no installation required, enabling immediate use from any browser.
- **Intelligence**: Leveraging AI to automate data cleaning, recommend relevant metrics, and optimize visualizations according to business context.
- **Efficiency**: Drastically reducing the time needed to transform raw data into actionable insights, from hours to minutes.
- **Scalability**: Building a cloud-native scalable architecture capable of adapting to the growing needs of industrial enterprises.

Team

Founders: Industrial Engineering students:

- Abderahim REDOUANE
- Malak EL MEMI

Supervisor: Our computer science professor: Dr. Salim KEBIR

Work Organization

- **Development Team**: Responsible for technical architecture, AI algorithm development, user interface, and cloud infrastructure (AWS/Azure).
- **Data Science Team**: In charge of preprocessing algorithm optimization, metrics recommendation system development, and visualization scoring models.
- **Product Team**: Manages user experience, functional specifications, user testing, and interface optimization based on client feedback.
- **Business Team**: Responsible for commercial strategy, client development, industrial partnerships, and international expansion.

First Axis: Project Presentation

Communication and Interaction Methods

- **Weekly team meetings** to synchronize technical, commercial, and product progress, with validation points for developed features.
- **Code management** via GitLab with continuous integration, systematic code reviews, and automated deployment to test and production environments.
- **Structured technical documentation** including API specifications, integration guides, and user documentation with video tutorials.
- **Client communication** via Slack for technical support, integrated CRM for commercial follow-up, and a user feedback platform for continuous improvement.

Project Objectives

- **Short-term Objectives (0-1 year)**

Develop and launch the MVP version of the platform with core functionalities: CSV upload, automatic preprocessing, and basic visualization generation.

Acquire the first 35 paying clients, primarily Algerian industrial SMEs in the manufacturing, agri-food, and logistics sectors.

Achieve client retention rate ≥ 80% and NPS (Net Promoter Score) ≥ 40, demonstrating user satisfaction and product-market fit.

- **Medium-term Objectives (1-3 years)**

Extend the platform with advanced functionalities: multiple data connectors (databases, APIs), real-time collaboration, automatic alerts, and third-party integrations.

Conquer the Maghreb market with 400+ active clients and develop strategic partnerships with local integrators and consultants.

Implement more sophisticated AI models for predictive recommendation and automatic anomaly analysis in industrial data.

- **Long-term Objectives (3-5 years)**

Deploy CSViz internationally, targeting Europe and Sub-Saharan Africa, with multilingual adaptation and local regulatory compliance (GDPR, etc.).

Develop a marketplace of sectoral extensions and templates, allowing partners to create specialized vertical solutions.

Establish CSViz as a regional leader in analytics democratization, with 10,000+ user companies and Series A funding for expansion.

Project Timeline

Our roadmap spans 18 months to achieve a complete commercializable version:

- ➢ **Months 1-3**: Core engine development (upload, CSV parsing, basic preprocessing)
- ➢ **Months 4-6**: User interface and automatic visualization generation
- ➢ **Months 7-9**: Metrics and plots recommendation system and performance optimization
- ➢ **Months 10-12**: User testing, security, production infrastructure
- ➢ **Months 13-15**: Advanced features, integrations, automated onboarding
- ➢ **Months 16-18**: Commercial optimization, customer support, expansion preparation

Nature of Innovations

CSViz stands out through its "zero-code" approach to industrial analytics, combining several technological innovations. Our intelligent preprocessing engine automatically detects data types, cleans anomalies, and structures information according to predefined industrial patterns.

The metrics recommendation system uses hybrid algorithms analyzing both statistical relevance (distribution, correlations) and semantic relevance (business pattern recognition) to suggest the most relevant KPIs and plots according to industry sector.

Innovation Domains

- **Democratized Analytics**: Transforming industrial data analysis from an expert domain to a tool accessible to all operational stakeholders, from technicians to managers.
- **Contextual AI**: Developing algorithms capable of understanding the business context of data to propose visualizations and metrics adapted to each industrial sector.
- **Cloud-first for Industry**: Creating SaaS infrastructure specifically optimized for the security, performance, and scalability needs of industrial enterprises.

Second Axis: Innovative Aspects

Technical Innovation

- **Automated Data Intelligence**: Our preprocessing engine uses machine learning to automatically detect data quality issues, suggest corrections, and optimize data structure for visualization.
- **Context-Aware Visualization**: AI algorithms analyze data patterns and business context to recommend the most effective chart types and dashboard layouts.
- **Natural Language Processing**: Integration of NLP capabilities to understand column headers and data context in multiple languages (French, Arabic, English).

Business Model Innovation

- **Freemium SaaS with AI**: Unlike traditional BI tools requiring extensive configuration, CSViz offers immediate value through AI-powered automation.
- **Industry-Specific Templates**: Pre-built dashboard templates optimized for manufacturing, logistics, quality control, and other industrial processes.
- **Progressive Disclosure**: Interface adapts complexity based on user expertise level, from simple drag-and-drop for beginners to advanced customization for power users.

Third Axis: Strategic Market Analysis

Market Segment

Potential Market

CSViz targets the analytics and business intelligence market for industrial enterprises, a rapidly growing sector with digital transformation. The potential market includes:

- Manufacturing companies (automotive, textile, electronics)
- Agri-food and pharmaceutical industries
- Logistics and supply chain companies
- Engineering consultants and industrial study offices

These stakeholders are primarily located in industrial zones of Algiers, Oran, Constantine, Setif, Skikda, and Annaba. In Algeria, over 2,000 medium and large industrial companies are concerned, plus 8,000+ SMEs with analytics needs.

These organizations seek to:

- Digitize their reporting and analysis processes
- Reduce dependence on rare data science skills
- Improve data-driven decision making
- Optimize operations through real-time insights

Target Market (Segment)

Our solution primarily targets:

- Industrial SMEs (50-500 employees)
- Operational departments of large enterprises
- Specialized consultants and study offices
- Industrial startups in the growth phase

This segment is relevant because these structures have real analytics needs but limited budgets and technical resources, creating a perfect gap for an accessible and affordable solution.

Why We Chose This Target Market

This market combines:

- Urgent need for post-COVID digitization
- Willingness to adopt modern cloud solutions
- Sufficient budgets to justify a SaaS subscription
- Rapid adoption capacity without complex decision processes

By targeting industrial SMEs, we avoid direct competition with traditional BI editors on the enterprise segment while addressing an underserved and rapidly growing market.

Client Contract Possibilities

Our internship company, SARL IZDIHAR, expressed its interest in our service, recognizing its potential to improve operational efficiency and decision-making through data-driven insights. Following our initial discussions and demonstrations, the company highlighted several areas where our solution could be implemented, particularly in predictive maintenance and production monitoring.

Planned national expansion

- Sonelgaz Group (utilities and energy)
- Sonatrach (oil and gas)
- Saidal (pharmaceutical)
- Textile groups
- Agricultural cooperatives

Competition Intensity Measurement

Direct and Indirect Competitors

International Direct Competitors:

- **Tableau** (Salesforce): Global leader in self-service analytics, intuitive interface but high complexity and prohibitive costs for SMEs (>$70/user/month).
- **Power BI** (Microsoft): Solution integrated with the Office ecosystem, cheaper than Tableau but requires significant technical expertise.
- **Qlik Sense**: Associative analytics platform, powerful but high learning curve and significant deployment costs.
- **Looker** (Google Cloud): Modern cloud-native solution, developer-oriented, unsuitable for business users.

Indirect Competitors:

- **Excel/Google Sheets**: Universal tools but limited for advanced analytics, no automation or intelligence.
- **Custom internal solutions**: Expensive in development and maintenance, limited scalability.
- **Traditional BI consultants**: Very expensive custom services, long delays, complex maintenance.

Market Share and Numbers

The Algerian BI/Analytics market remains largely dominated by Excel (>70% of SMEs), with limited adoption of international cloud solutions (<15%) mainly due to costs and complexity.

Large enterprises (Sonatrach, Sonelgaz) mainly use IBM/Oracle/SAP solutions, unsuitable for SMEs.

Strengths and Weaknesses

Competitor	Strengths	Weaknesses
International Solutions (Tableau, Power BI)	● Very complete features ● Mature ecosystem ● International support ● Numerous integrations	● Prohibitive costs for SMEs ● High technical complexity ● Limited local support ● Long deployment time
Excel/Sheets	● Universally known ● Low cost ● Total flexibility ● No training required	● No real-time collaboration ● Scalability limits ● No automation ● Frequent human errors
Custom Solutions	● Perfectly adapted to needs ● Total control ● Native integration	● High development costs ● Complex maintenance ● Limited scalability ● Technical dependency

Our CSViz Positioning:

We fill the gap between Excel (too limited) and enterprise solutions (too complex/expensive) by offering:

- **Simplicity**: Intuitive interface, zero training required, upload-and-go
- **Intelligence**: Integrated AI to automate preprocessing and recommend visualizations
- **Accessibility**: SME-adapted pricing, no minimum users
- **Localization**: French/Arabic support, understanding of local specificities
- **Speed**: Dashboards generated in minutes, not weeks

Marketing Strategy

Strategy Adapted to Our Financial Resources

As a SaaS startup with a limited budget, our marketing strategy prioritizes high-ROI digital channels and a B2B consultative approach, maximizing the value generated.

Content Marketing and SEO

Creating a specialized industrial analytics content hub:

- Technical blog with sectoral use cases
- Practical guides: "How to analyze your production data"
- Monthly webinars on Industry 4.0
- Free dashboard templates by sector

Digital Lead Generation

- Landing pages optimized by persona (quality manager, production manager, etc.)
- LinkedIn Ads campaigns targeting industrial decision-makers
- Google Ads on specialized keywords ("production dashboard", "industrial data analysis")
- Retargeting strategy for prospect nurturing

Strategic Partnerships

- Alliances with digital transformation consultants
- Partnerships with local ERP integrators (reseller role)
- Collaborations with industrial associations

Freemium and Free Trials

- Limited free version for initial hook
- 30-day free trial without commitment
- Guided onboarding with example datasets
- Dedicated success manager for trial→paid conversion

B2B Events and Networking

Organizing proprietary events:

- Monthly "Data Breakfast" with industrialists
- "Analytics 101" workshops in industrial zones
- Sectoral conferences on Industry 4.0

Fourth Axis: Production Plan and Organization

Production Process

As a SaaS solution, our "production" consists of software development, cloud deployment, and client onboarding:

Step 1: Development and Testing

- Sprint development (Agile/Scrum methodology)
- Automated testing and quality validation
- Deployment on staging environments

Step 2: Release and Deployment

- Blue-green deployment on cloud infrastructure
- Load testing and performance monitoring
- Documentation updates and release communication

Step 3: Client Onboarding

- Account configuration and interface customization
- User training and data transfer
- Dashboard setup according to specific needs

Step 4: Support and Evolution

- Usage and client performance monitoring
- Reactive technical support
- Feedback collection for product roadmap

Supply Chain

Purchasing Policy

Our supply strategy focuses on cloud services, software licenses, and human resources:

Cloud services and infrastructure:

- AWS/Azure for hosting and compute (auto-scaling)
- Global CDN for international performance
- Backup and disaster recovery services

Licenses and tools:

- Development tools (JetBrains, VS Code Pro)
- Internal SaaS services (Slack, Notion, GitHub)
- Monitoring solutions (DataDog, Sentry)

Cost optimization policy:

- Reserved instances for stable infrastructure
- Volume-based pricing negotiations
- Continuous cloud cost monitoring with alerts

Most Important Suppliers

Technology Suppliers:

- **Amazon Web Services**: Primary cloud infrastructure, ML/AI services
- **Auth0**: Authentication and user management
- **SendGrid**: Transactional email delivery

Service Suppliers:

- **Algerian Cloud Providers**: Local backup and regulatory compliance
- **Specialized tech law firm**: Intellectual property and contracts

Payment Policy and Deadlines

Cloud supplier payments: Monthly automatic, witha 3-month security provision to avoid service interruptions.

Supply deadlines:

- Cloud services: Instant (auto-provisioning)
- Software licenses: 24-48h average
- Human resources: 4-8 weeks (recruitment + onboarding)

Workforce

Number of Positions

Initial phase (0-12 months): 4-6 people

- 2 full-stack developers
- 1 data scientist
- 1 product manager/UX
- 1 sales/marketing
- 1 founder/CEO

Growth phase (12-36 months): 12-15 people

- Technical team: 6-8 developers, 2 data scientists, 1 DevOps
- Business team: 3 sales, 1 marketing, 1 customer success
- Support: 1 accounting, 1 HR

Required Types and Profiles

- **Full-stack developers**: React/Node.js, B2B SaaS experience, API design
- **Data scientists**: Machine Learning, NLP, industrial experience desirable
- **Product manager**: B2B SaaS background, industrial ecosystem knowledge
- **Sales/Marketing**: B2B sales experience, industrial sector knowledge

Outsourcing Use

- **Design and UX/UI**: Specialized SaaS freelancers for interface and branding
- **Specialized development**: Consultants for complex integrations (ERP, industrial APIs)
- **Digital marketing**: Local agency for paid campaigns and technical SEO

Key Partners

Technology partners:

- Cloud providers (AWS, Azure) for infrastructure
- Industrial solution editors for integrations
- Universities for R&D and talent recruitment

Business partners:

- ERP integrators as resellers/prescribers
- Digital transformation consultants for co-selling
- Industrial associations for credibility and network

Financial partners:

- Banks specialized in tech startups
- Early-stage investment funds
- Innovation support organizations (ANSEJ)

Fifth Axis: Financial Plan

Executive Summary

CSViz is a SaaS platform targeting industrial SMEs in Algeria with AI-powered data analytics. This plan projects conservative growth with realistic market penetration rates and adjusted financial projections.

Revenue Model

SaaS Pricing Structure

- **Starter Plan**: 15,000 DZD/month (SMEs, <5 users)
- **Professional Plan**: 35,000 DZD/month (Medium enterprises, <20 users)
- **Enterprise Plan**: 70,000+ DZD/month (Large enterprises, unlimited users)

Additional Revenue Streams

- Implementation services: 15% of subscription revenue
- Training and certification: 5% of subscription revenue
- Specialized industry modules: 8% of subscription revenue

Financial Projections (5-Year)

Base Scenario Projections

Year	Active Clients	Avg Monthly ARPU (DZD)	Annual Recurring Revenue (DZD)	Services Revenue (DZD)	Total Revenue (DZD)
2026	35	16,500	6,930,000	1,970,000	8,900,000
2027	100	19,200	23,040,000	5,180,000	28,220,000
2028	280	22,800	76,608,000	17,250,000	93,858,000
2029	550	26,400	174,240,000	39,240,000	213,480,000
2030	900	29,500	318,600,000	71,760,000	390,360,000

Cost Structure

Cost Category	2026	2027	2028	2029	2030
Team Salaries	2,400,000	5,000,000	12,000,000	20,000,000	30,000,000
Cloud Infrastructure	450,000	1,500,000	4,500,000	9,000,000	15,000,000
Marketing/Customer Acquisition	1,200,000	2,500,000	7,500,000	12,750,000	19,500,000
R&D/Product Development	600,000	1,250,000	2,800,000	4,500,000	7,000,000
General/Administrative	400,000	700,000	1,500,000	2,500,000	4,000,000
Customer Success/Support	250,000	620,000	1,900,000	3,200,000	5,500,000
Total Operating Costs	**5,300,000**	**11,570,000**	**30,200,000**	**51,950,000**	**81,000,000**

Profitability Analysis

Metric	2026	2027	2028	2029	2030
Total Revenue	8,900,000	28,220,000	93,858,000	213,480,000	390,360,000
Total Costs	5,300,000	11,570,000	30,200,000	51,950,000	81,000,000
EBITDA	3,600,000	16,650,000	63,658,000	161,530,000	309,360,000
Depreciation	800,000	800,000	600,000	400,000	200,000
EBIT	2,800,000	15,850,000	63,058,000	161,130,000	309,160,000
Tax (19%)	532,000	3,011,500	11,981,020	30,614,700	58,740,400
Net Profit	2,268,000	12,838,500	51,076,980	130,515,300	250,419,600
Net Margin	25.5%	45.5%	54.4%	61.1%	64.1%

Scenario Analysis

Optimistic Scenario (+30% client acquisition)

- 2030 Revenue: 507M DZD
- 2030 Net Profit: 325M DZD

Conservative Scenario (-25% client acquisition)

- 2030 Revenue: 293M DZD
- 2030 Net Profit: 188M DZD

Funding Requirements

Phase 1: Seed Funding (2026) - 4,500,000 DZD

- Founder Contribution: 1,500,000 DZD (33%)
- Family & Friends: 800,000 DZD (18%)
- Government Grants (ANSEJ/CNRC): 1,200,000 DZD (27%)
- Accelerator Program: 1,000,000 DZD (22%)

Phase 2: Series A (2027) - 15,000,000 DZD

- Local Investment Funds: 10,000,000 DZD (67%)
- Angel Investors: 3,500,000 DZD (23%)
- Development Bank: 1,500,000 DZD (10%)

Phase 3: Series B (2029) - 40,000,000 DZD

- International VC Funds: 28,000,000 DZD (70%)
- Corporate Investors: 12,000,000 DZD (30%)

Initial Investment Breakdown

Investment Type	Amount (DZD)	Depreciation Period
Technology Infrastructure		
- IT Equipment & Servers	600,000	3 years
- Software Development	1,500,000	3 years
- Licenses & IP	300,000	5 years
Office Setup		
- Office Space & Furniture	400,000	5 years
- Initial Marketing	500,000	1 year
Working Capital	1,200,000	-
Total Initial Investment	**4,500,000**	

Cash Flow Analysis

Monthly Cash Flow (Year 1 - 2026)

Month	Revenue	Expenses	Net Cash Flow	Cumulative Cash
Jan	1,500,000*	500,000	1,000,000	1,000,000
Feb	150,000	450,000	(300,000)	700,000
Mar	250,000	450,000	(200,000)	500,000
Apr	400,000	450,000	(50,000)	450,000
May	550,000	450,000	100,000	550,000
Jun	700,000 + 1,200,000**	450,000	1,450,000	2,000,000
Jul	750,000	450,000	300,000	2,300,000
Aug	800,000	450,000	350,000	2,650,000
Sep	850,000	450,000	400,000	3,050,000
Oct	900,000	450,000	450,000	3,500,000
Nov	950,000	450,000	500,000	4,000,000
Dec	1,000,000	450,000	550,000	4,550,000

*Initial funding
**Government grant

Key Financial Metrics

Investment Returns

- NPV (12% discount rate): 245,000,000 DZD (estimated)
- IRR: 75%
- Payback Period: 2.3 years
- ROI (5-year): 3,900%

Operational Metrics

- Customer Acquisition Cost (CAC): 2,500 DZD (Year 1) → 1,800 DZD (Year 5)
- Customer Lifetime Value (LTV): 350,000 DZD
- LTV/CAC Ratio: 140:1 → 195:1
- Monthly Churn Rate: 3% (Year 1) → 1.5% (Year 5)

Financial Health Indicators

- Self-Financing Ratio: Strong cash generation from Year 2
- Debt Coverage Ratio: Not applicable (equity-financed)
- Cash Conversion Cycle: 18 days (efficient SaaS model)

Break-Even Analysis

Break-Even Points

- Unit Break-Even: 42 clients (achieved Month 8, 2026)
- Cash Flow Break-Even: Month 6, 2026
- Full Profitability: Month 12, 2026

Risk Mitigation

Financial Risks

- Market adoption slower than projected: Conservative client acquisition rates built into base scenario
- Increased competition: 15% marketing budget allocation for competitive positioning
- Currency fluctuation: Pricing adjustments mechanism in contracts

Operational Risks

- Key talent retention: Equity participation program for core team
- Technology scalability: Cloud-first architecture with auto-scaling capabilities
- Customer concentration: No single customer >5% of revenue target

Conclusion

CSViz represents a significant opportunity to democratize data analytics for industrial SMEs in Algeria and the broader MENA region. Our AI-powered, zero-code approach addresses a clear market gap between overly simple tools like Excel and overly complex enterprise BI solutions.

With conservative projections showing profitability by Year 2 and potential for 100M+ DZD annual revenue by Year 5, CSViz offers attractive returns for investors while building a sustainable, scalable business. Our experienced team, validated market need, and innovative technology position us for success in the rapidly growing Industry 4.0 market.

The combination of local market knowledge, international technology standards, and AI-driven automation creates a compelling value proposition that can scale beyond Algeria to serve the broader francophone African market and eventually expand globally.