CRISP-DM rapport

Competency Monitoring System Application

Auteurs

Jason Tomeij

Baha Kucuk

Zoë González Grootaert

Fatima Memon

Semester

Data Analytics for Business

Datum

30-10-2025

Versie

In werking

Version Control

Business understanding

Version Number	Date	Author	Change
0.1	17-09	Jason	Started on working on the Current situation.
0.2	18-09	Jason + Baha	Change of the documentation + Lay-out of the document.
0.2	18-09	Jason + Baha	Changed the SIPOC + Making the BPMN Model.
0.3	18-09	Zoë	Changed front page, put business goals in table.
0.4	22-09	Zoë	Added intros and small changes to balanced scorecard.
0.4	22-09	Fatima	justification and improvement of the SWOT analysis and Porter's Five Forces model.
0.4	22-09	yara	Vision&scope Stakeholders analysis Ist-soll-Gap.
0.5	24-09	Jason	Looking and changing what is needed for a better outcome.
0.6	25-0	Jason	Making the document in English.
0.7	25-09	Zoë	Translated my parts to English.
0.8	26-09	Zoë	Added target levels to balanced scorecard that the client provided.
0.9	25-9-2025	Fatima	Translated the parts to English.
1.0	26-9	Baha	Translated Sipoc to English
1.1	30-09-2025	Jason	Starting the Project plan.
1.2	01-10-2025	Jason	Started on Stakeholder analysis + Worked on new BPMN Model for the expected end result.
1.3	02-10-2025	Jason	Added new BPMN Model +.
1.3	02-10-2025	Baha	Azure database explained
1.3	2-10-2025	Fatima	Scraping + deadlines and planning .
1.4	03-10-2025	Baha + Jason	Making new document to make things les chaotic. Was needed after feedback.
1.5	06-10-2025	Baha + Jason	Making the new Risk analysis and added and cost and benefits.
1.6	07-10-2025	Jason	Started working on new BPMN Model after review with Hend. Also made a MoSCoW Table for business success criteria. Added SWOT Analysis. Started on Project plan

			Added Stakeholder analysis + Added Project plan Timeline and Project Overview.
1.6	07-10-2025	Zoë+ yara	Inventory of recources and a part of requierements, assumpsionts and constraints.
1.6	07-10-2025	Fatima	Initial assessment of tools and techniques, data mining succes criteria.
1.6	07-10-2025	Baha	Keep document in order and made a part of project plan.
1.6	07-10-2025	Yara	Terminology and a part of requirements, assumptions and constraints.
1.6	10-10-2025	Fatima	Data mining success criteria improved
1.7	16-10-2025	Zoë	Introductie expanded, requirements in tabels
1.7	16-10-2025	Baha	Updated sipoc
1.8	19-10-2025	Fatima	Business succes criteria + data mining goals improved
1.9	20-10-2025	Zoë	Balanced score card adjusted, entirely new mission, vision, strategy
2.0	23-10-2025	Jason	Added feedback. Added more information about the MoSCoW table. + Added gant chart in Timeline and phases
2.1	24-10-2025	Fatima	Business succes criteria adjusted
2.2	26-10-2025	Baha	Risk-analyse improved
2.3	27-10-2025	Jason	Made Gant-chart
2.3	27-10-2025	Baha	Made new current situation and roles
2.3	27-10-2025	Jason	Made adjustment based on feedback of Matthijs van Berkhout
2.4	`06-11-2025	Zoë	Adjusted business understanding (scope, requirements, spelling errors, percentages, elaborated on certain topics, etc) based on feedback from client (Hend Elsayed)

Data Understanding

Version Number	Date	Author	Change
0.1	09-10-2025	Jason	Started on the first steps of the document. With Lay-out.

			Also started working on the Collect initial data
0.2	30-10-2025	Baha	Made the data requierements planning,data sources and indeed vacancy data extracting
0.3	31-10-2025	Baha	Made Tools and environment, Ethical and Legal Considerations and output summary
0.4	30-10-2025	Fatima	Made Describe data
0.5	30-10-2025	Zoë	Made explore Data
0.6	30-10-2025	Yara	Made Verify Data Quality

Inhoudsopgave

1 Introduction

Hogescholen, such as the Hogeschool Utrecht (HU), face the challenge of ensuring that their curricula align with the constant changing and evolving requirements of the labor market. Students often develop skills that do not fully match the competencies demanded by employers, resulting in a gap between education and professional work. This mismatch can limit graduates' employability, and reduce the effectiveness of educational programs.

The goal of this project is to develop a data-driven system, that enables the Marketing & Customer Experience Research Group at HU to analyze curricula and identify skill gaps. By matching curricula with labor market data from the Dutch market, with extra consideration of the top 100 companies, the system provides insights that could support curriculum adjustments, ensuring that students continue to learn relevant skills, and that HU continues to deliver competent professionals to the Dutch labor market.

The relevance of this project, lies in its potential to improve educational outcomes and labor market alignment. It benefits HU by enabling data-informed curriculum development and helps educators design (or adjust already existing) programs, that meet current and future professional demands. At the same time, it indirectly supports students by enhancing the skills they acquire during their studies.

This document is structured as follows: first, the Business understanding, then the Data understanding, Data preparation, the Modeling, Evaluation and Deployment phase.

(Hogeschool Utrecht, z.d)

2 Business Understanding

In this chapter, we will cover topics that, according to the CRISP-DM methodology, belong to the business understanding phase.

2.1 Determine Business Objectives

The first objective of the data analyst is to thoroughly understand, from a business perspective, what the client truly wants to accomplish. Therefore, the background, business objectives and business success criteria will be discussed below.

2.1.1 Background

This project is a continuation of a previous project by Dennis Hagen. Therefore, an analysis of the preceding project will first take place. Subsequiently, an extensive background analysis will be preformed to thoroughly understand the organization.

This analysis is divided into the following deliverables, namely:

- Mission
- Vision
- Strategy
- Scope
- SIPOC
- IST-BPMN
- SWOT-analysis

2.1.1.1 Previous Project

The previous project: the thesis of Dennis Hagen

The previous project is the master's thesis of Dennis Hagen, titles “Structural monitoring of digital marketing competencies for educational curricula.” In his research, he focused on the problem that study programs in higher education often do not align well with the rapidly changing labor market, especially within the field of digital marketing.

Due to the rapid rise of technology and artificial intelligence, the required skills in the professional field are constantly changing. Hagen noticed that educational programs often respond too late to this, causing graduates to sometimes lack the competencies that employers are looking for.

To structurally address this problem, he designed a competency monitoring system that helps educational programs continuously compare their curricula with current labor market developments. This system is intended to make educational programs future-proof and better aligned with professional practice.

How did Hennis Hagen approach the project?

Dennis Hagen approached his research in a design-oriented and systematic way. He combined theoretical models with practical analyses, making his project both scientifically grounded and applicable in practice.

He used the Viable System Model (VSM) as the main theoretical framework. This model assumes that organizations, in this case educational institutions, can only function sustainably if they can continuously adapt to changes in their environment. Dennis applied this to education: he divided the system into five parts (current curriculum, labor market, frontrunners, future, and gap analysis) and investigated how these parts are interconnected.

In addition, he used the Design Science Research Methodology (DSRM). This method consists of several phases: analyzing the problem, designing a solution, building a prototype, and evaluating it in practice.

He collected and analyzed more than 5,000 job vacancies from the marketing sector using data analysis and text mining. This allowed him to map which skills were most in demand in the professional field. He then compared these skills with the current curriculum of the program to determine where the biggest differences lay.

The results showed that employers increasingly need data skills, technical knowledge (such as Python and SQL), and an understanding of artificial intelligence, while these topics are still hardly addressed in many educational programs. He also pointed out the growing importance of ethical awareness, interdisciplinarity, and lifelong learning.

He also concluded that a structural, data-driven monitoring system is necessary to keep curricula up to date. Such a system makes it possible to continuously improve education based on concrete data, instead of only making adjustments every few years.

How we approached the project and what we take from it

Our current project is based on the insights and approach of Dennis Hagen, who conducted thorough research and designed a prototype. We focused on the technical and further development of the system.

What we take from his research are three important principles:

1. **A structural monitoring system:**
   Like Dennis Hagen, we aim for a system that regularly and automatically collects new labor market data. We have translated this into a fully automated pipeline, which can be executed anually or periodically without manual work.
2. **The use of the Viable System Model as a framework for thinking:**
   We apply the VSM to logically organize our data flows and stakeholders. We can also link the similarities between the skills in HU curricula and the skills demanded in the labor market, and analyze and present the differences in a visual dashboard. This allows us to clearly visualize where the gap between education and the professional field is located.
3. **The application of data analysis and NLP techniques:**
   We are currently developing the technical prototype, using concepts from the model that Dennis designed. Just like Dennis, we use Python for scraping and analyzing vacancies, and Natural Language Processing (NLP) for recognizing skills.

Through this approach, we combine his scientific foundation with our technical implementation. The result is a working system that can be used annually to update curricula and better align them with the labor market.

(Hagen, 2024)

2.1.1.2 Mission

The Marketing & Customer Experience research group focuses on strengthening valuable relationships between people and brands in an increasingly digital world. Our mission is to ensure that marketing and technology are used in a meaningful, sustainable and human-centered way, with respect for both the customer and society. As a part of the “Kennis centrum", part of our role is to find new approaches to maintain the connection between education, research, and business (market).

2.1.1.3 Vision

We believe in a future where marketing acts as a positive force for societal change. By combining technology and customer experience responsibly, organizations can achieve sustainable growth while also contributing to the well-being of people, brands and society.

2.1.1.4 Strategy

We pursue our mission and vision through applied research, education, and close collaboration with industry partners.

The research group develops tools, methods, and interventions that help academics and professionals to:

- Strengthen customer relationships through data and technology,
- Apply sustainable and ethical marketing practices,
- And shape digital transformation in a human-centered way.

(Hogeschool Utrecht, z.d)

2.1.1.5 Scope

In the scope, we show what we will do within this project and what we will focus on (in scope). This helps prevent any misunderstandings during the project. We also describe what we will not include (out of scope), so it is clear which parts fall outside this project. In the table below, we explain for each task what will be delivered (deliverables) and when the result will be approved (acceptance criteria).

In scope	Deliverables	Acceptance criteria
Scraping job vacancies from the Dutch market, including top 100 Dutch companies.	Automated scraping algorithm	Will be accepted when at least 100 job postings are collected per scraping run, with ≥90% success rate.
Automatically extracting job titles, soft skills, hard skills, sector, job role distribution, salary & benefits and location.	Data extraction algorithm	Will be accepted when, in a sample of 100 job postings, at least 90% of the relevant fields are extracted correctly.
Developing a matching algorithm using Natural Language Processing (NLP) to link program curriculum skills with job requirements.	NLP-based matching algorithm	Will be accepted when, for a test set of 4 job postings, at least 3 relevant matches are confirmed.
Visualizing the matching results and skill gaps in an interactive dashboard.	Interactive dashboard with skill gap visualization	At least 2 different diagrams to show the results of skill gaps
Implementing an automated pipeline (scraping → extraction → matching → visualization).	Integrated pipeline (prototype)	Accepted when the full pipeline runs automatically in ≥95% of test runs and delivers complete output

Out of scope
Real-time or continuous scraping (only periodic runs will be performed).
AI-based predictions or trend analysis of future labor market developments.
Integration with internal HU systems such as OSIRIS or LMS platforms.
Analysis of international or non-Dutch job vacancies that are not related to marketing

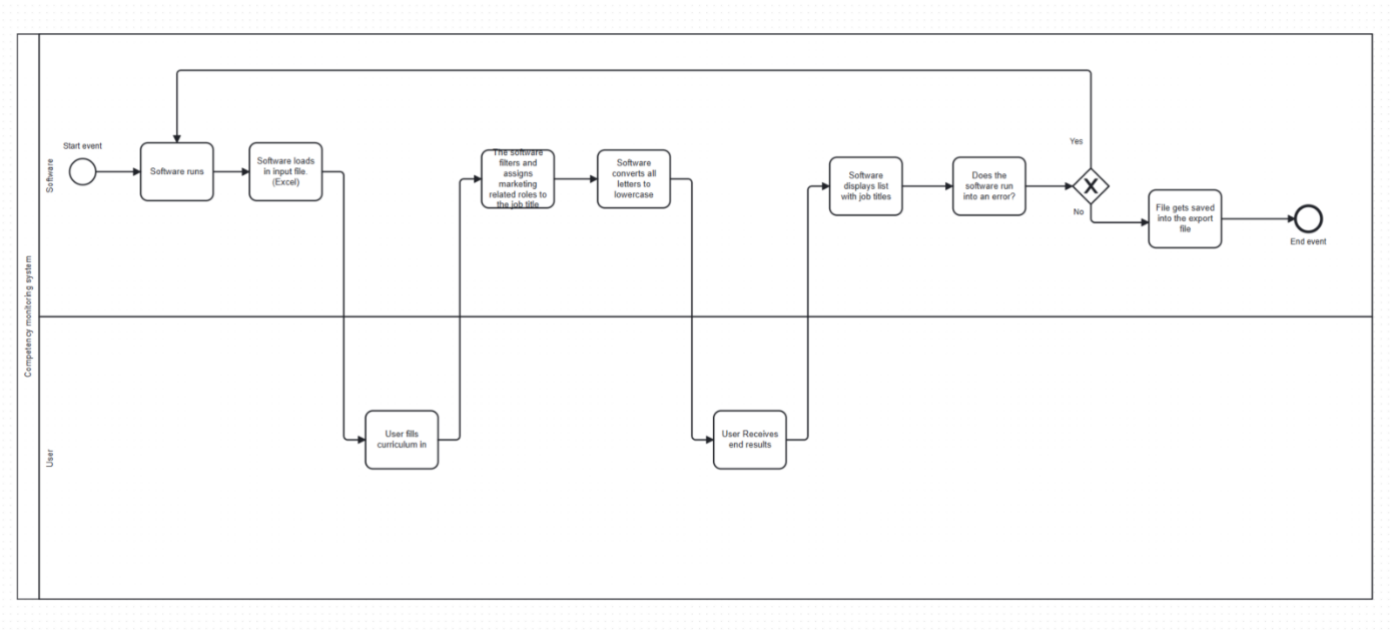
2.1.1.6 SIPOC

To gain a proper understanding of the current situation, we use a SIPOC analysis to create insight into the existing processes, systems and stakeholders involved. In addition, by applying a BPMN Model, we can clearly visualize how to current process operates. This provides us with a solid understanding of the present workflow and highlights where improvements can be made in the new website.

SIPOC				
Suppliers	**Inputs**	**Processes**	**Outputs**	**Customers**
Indeed Hogeschool Utrecht students (curriculum) Microsoft (Azure) HU teachers Dennis Hagen Hend Elsayed	HU curiccula (Website) Python coding for scraping Database in azure Vacancies indeed	Manually retrieve data from current vacancies indeed Loading the data into code Code runs software and searches for keywords to connect vacancies Generating response Answer comes back through the code Answers get displayed on the dashboard	Overview of skill gaps per study program Interactive dashboard (Power BI) Automated dataset containing job and competency data Reports and insights for curriculum improvement	Utrecht University of Applied Sciences (lectors and institutes) Lecturers and curriculum developers Students (indirect users) Researchers and policy makers (future users) Other universities of applied sciences (potential future users)

With the help of a SIPOC analysis, we mapped out the process and made it as transparent as possible. This made it clear which parties are involved, what data enters the process, how the steps are carried out, and which results are ultimately delivered to the end user.

2.1.1.7 IST-BPMN

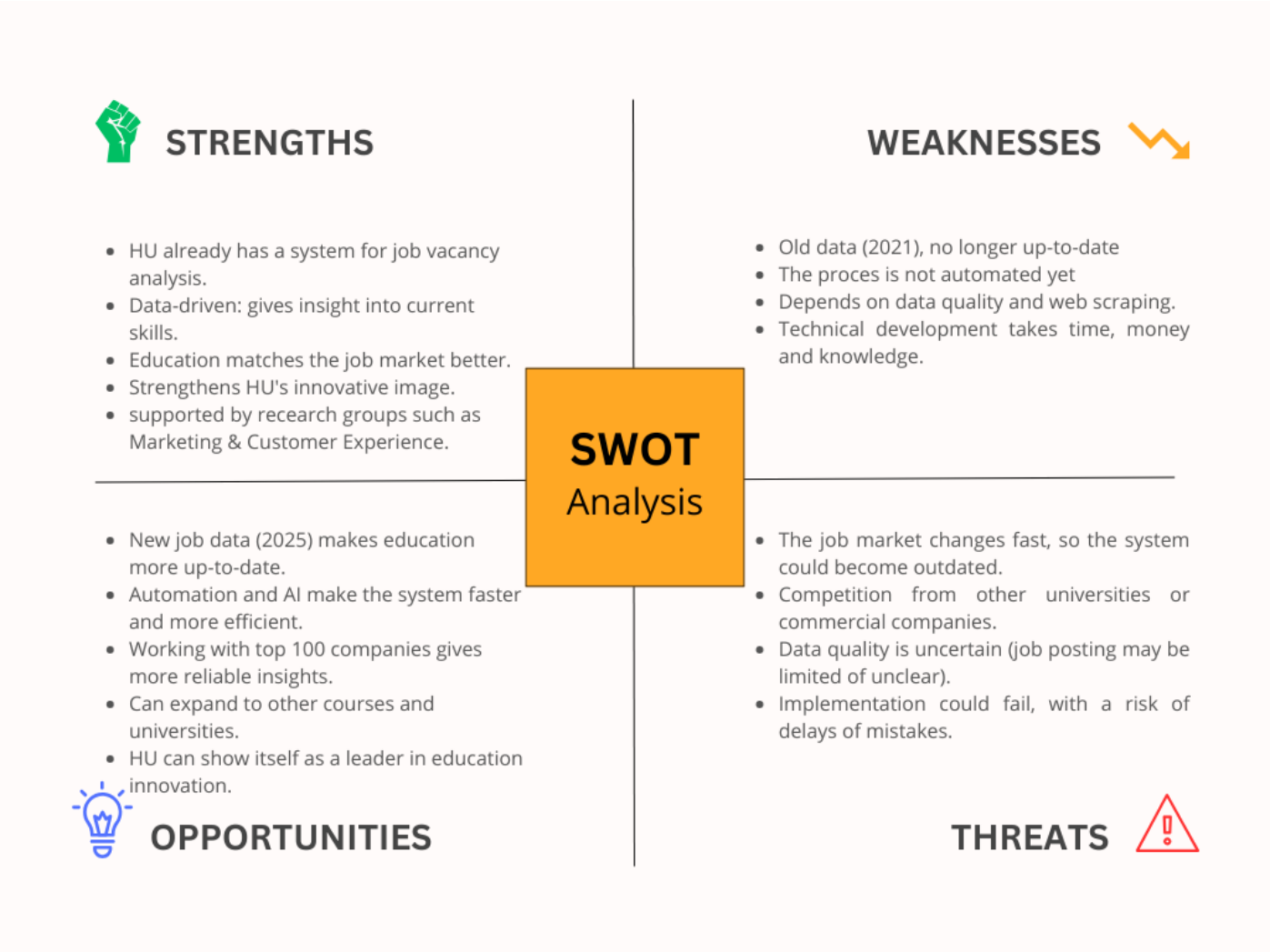


By modelling the process in a BPMN (Business Process Model and Notation) diagram, we now have a clear view of how the process functions. This provides a better understanding of the current situation and forms a foundation for optimization.

The process begins with starting the software. Once initialized, the input file is uploaded. Next, the curriculum is entered, after which the software assigns marketing-related roles to the position titles. The text is then automatically converted to lowercase. The software subsequently generates statistics and displays a list of position titles. A check is performed to determine whether the software produces an error message. If an error occurs, the process restarts. If no error is detected, the file is saved as an output file and the process is completed. Now we have an output file with the jobs that match the curriculum

2.1.1.8 SWOT-analysis

To understand the organization and system better, we use a SWOT analysis to look at the system's strengths, weakness, opportunities, and threats. This helps to make the system strong and reliable for the future.



Hogeschool Utrecht has a good base with the competency monitoring system to connect education and jobs. But the system is still old data, takes a lot of work and depends on good job data. There are big chances to improve it with new data, AI and working with companies. Fast changes in the job market, competition and technical problems are big threats.

2.1.2 Business Objectives

2.1.3 Business Goals

To provide direction to the project and clarify where we want to go with it, we have defined business goals. These goals explain why we are doing this project and what we aim to achieve. They form the foundation of the project and were established based on discussions with the client.

Business goal	**Description**	Project goal
Building a bridge between education and the labour market	Contribute to reducing the competency gap and help educational programs train student with the right knowledge and skills.	The system should support programs in evaluating and adjusting their curricula, thereby building a bridge between education and the labor market.
Improving and updating curricula	Evaluate and adapt educational programs so that graduates' competencies better align with the needs and requirements of the labour market.	Develop a matching algorithm that links competencies from job vacancies to the HU curriculum.
Increasing labor market relevance	Reduce the gap between education and the professional field by having real-time insights into required competencies.	Deliver a user-friendly, interactive dashboard that provides programs and researchers with real-time insights into competencies and trends.
Strengthening research and education with up-to-date insights	Provide teachers, researchers and policymakers with reliable and current data to support decision-making.	Use recent and representative data (2025) from the market with specific attention to top 100 Dutch companies.
Increasing efficiency and sustainability	Automate the process of data collection and analysis so it can be repeated annually and remain up to date without high additional costs or manual effort.	Within 6 months, realize a fully automated pipeline that collects and processes job vacancies, and set up a system that can run at least once a year and be easily updated.

2.1.3.1 Balanced score card

After defining the business goals, a Balanced Scorecard was created to make these goals more concrete and measurable. The business goals describe *what* we want to achieve, like connecting education and the labor market, while the Balanced Scorecard focuses on *how* we can track and evaluate that.

Each goal is linked to specific KPI's, target levels, and actions. This helps us view and monitor the project from different perspectives: financial, customer, internal processes & growth. This ensures that no aspect of the project gets overlooked or forgotten. This card was discussed with and reviewed by our client, who also helped us determine the target levels.

Perspective	Goals	KPI's	Target Level	Initiatives / Actions
Financial	Cost-efficient system that can be deployed sustainably	- Annual operational costs in € - Percentage time saved compared to manual processing	- €500 - 80%	Automate data collection and processing
Customer (HU professors, training & education, managers, researchers)	Satisfied stakeholders and usability of insights	- Percentage satisfaction score of programs/researchers - Number of programs that have been adjusted	- ≥80% - ≥2-10	Develop a user-friendly, interactive dashboard with real-time competence insights
Internal Processes	Efficient and reliable collection/processing of data from up-to-date sources	- Percentage of automated tasks	- ≥95%	Build a scraping pipeline for Indeed including top 100 Dutch companies
Learning & Growth	Scalability to other disciplines	- Number of new disciplines successfully integrated	- 2 or more	Prepare processes and system to handle additional disciplines

2.1.3.2 MoSCoW Table

By using a MoSCoW table we can decide as team what is essential versus optional. Instead of trying to do everything at once, we can prioritize the core features that are necessary for the system to succeed. It also keeps the project realistic and on schedule.

Must have: These are non-negotiable features that are essential for the system to function. Without these features, the project would fail

Should have: Important features that add significant value but are not critical for the system's initial success. They enhance usability and efficiency.

Could have: These are features that are nice-to-have. They improve user experience and flexibility but van be implemented later if time is limited

Wont have: These are features that are currently out of scope. These are either too complex, unnecessary or planned for future versions. They help prevent scope creep.

By using the MoSCoW method, our team can clearly prioritize what is critical, What adds value, and what can be postponed or excluded. This structured approach ensures that resources are focused on what truly matters for the success of the project

2.1.4 Business succes criteria

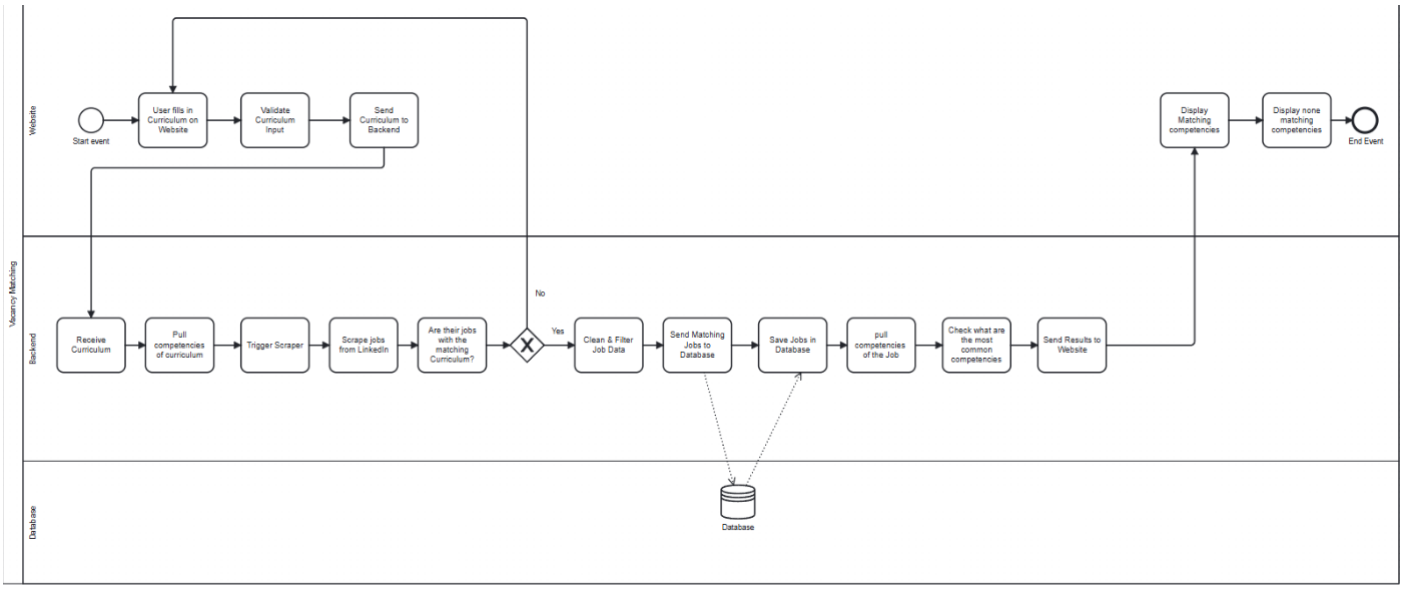
Business success criteria help a project stay focused on what really matters. They show what needs to be achieved for the project to be called successful and make it easier to measure progress and value. These criteria are also important to improve the desired process of the project.

2.1.4.1 Success criteria:

- Within 2 months the processing system is fully automated, so the coordinator needs to do at least 80% less manual data entry compared to the start.
- Around 1 month there is a web scraping system that collects at least 100 job listings within one week. When checked, at least 80 of them have correct and complete information.
- Around 1,5 months, there is a matching algorithm that connects at least 4 job listing to CE program curriculum. When checked by the coordinator, at least 3 of these matched are found to be relevant.

2.1.4.2 SOLL-BPMN Model

To ensure that the process remains clear, structured, and up to date, we created a BPMN model of the desired situation. This model provides a visual representation of the target process, making it easier to identify improvements.



As you can see in this BPMN model, many aspects have been adjusted, especially the steps required to reach the final result. With these changes, we aim to improve the process and ensure that everything runs more smoothly without pitfall.

2.2 Assess Situation

2.2.1 Inventory of Resources

To carry out the project successfully, it is important to have a clear overview of all available resources. This includes not only the people involved, but also the data sources, the technical tools, and the software needed to build the system. The following table provides an overview of these resources and their specific role within the project.

Category	Resource	Description
Personnel	Projectteam (Students)	Responsible for the overall project execution, including research, development of the website, data collection and dashboard design
	Business expert (HU lectorare/client)	Provides insight into the project and validates the project goals
	Project supervisor	Provides guidance and feedback throughout the project
Data	Indeed job postings Dutch market (including top 100 Dutch companies)	Used to identify and analyze competencies and skills that are required in the Dutch labor market.
	HU curriculum and study programs	Provides information on current competencies taught within the school's programs
Computing Resources	Azure cloud environment	Hosts the web application, stores datasets and runs the data processing pipeline
	Student workstations/laptops	Used by project team for testing, training and development
	Database	Stores scraped job data and curriculum information securely
Software	Python	For data scraping, cleaning, processing
	Power BI	For visualization and reporting the results in an interactive dashboard
	HTML, CSS, JavaScript and PHP	For developing the website's interface and ensuring accessibility
	Azure/GitHub	For version control, collaboration and hosting the backend system
	Microsoft Office	For documentation, reporting, and communication within the team

2.2.2 Requirements, Assumptions, and Contraints

2.2.2.1 Functional Requirements

In this section, we present the key requirements for the Competency Monitoring System. We have done this to clearly define what the system must be able to do (functional requirements) and the quality standards it must meet (non-functional requirements)

Requirement	Acceptance criteria
The system must enable users (lecturers/teachers) to input a curriculum from Utrecht University of Applied Sciences via a link	90% of test runs must successfully complete the curriculum upload process.
The system must automatically collect job vacancy data from Indeed, with extra focus on the top 100 Dutch companies.	Each scraping run must collect at least 100 job vacancies, with a minimum of 90% meeting the defined criteria.
The system must automatically extract the job title, hard and soft skills, sector + B2B and B2C, salary, benefits, role distribution and location for each collected vacancy	In a sample of 100 job vacancies, at least 90% of the stored records must be complete.
The system must automatically compare the skills from the curriculum with the required competencies in the job vacancy and identify differences (skill gaps).	When tested with a curriculum, 90% of the matching and missing skills must be displayed correctly.
The system must display an interactive dashboard showing the matching results, including skill gaps per study program and the underlying job vacancy and curriculum data.	In 90% of test runs, users must be able to by study program and skill type (hard or soft skill) without system errors.

2.2.2.2 Non-Functional Requirements

Non-functional requirement	**Acceptation criteria**
The system must be user-friendly for non-technical users.	At least 86% of test users must be able to use the system independently within 5 minutes, without assistance.
The scraping, matching, and publishing processes must be largely automated.	90% of the entire process must be executed automatically, with a maximum of one manual step allowed per full run.

The dashboard must respond quickly, enabling users to efficiently obtain insights.	85% of test users must be able to use the dashboard normally within approximately 2 seconds.
The system must comply with GDPR regulations and the security standards of Utrecht University of Applied Sciences.	In 84% of security tests, all data must be encrypted during storage and transmission, and no personal data from job vacancies may be stored.
The system must be able to be executed annually with new data without requiring modifications.	During an annual update, the pipeline must be successfully executed in at least 85% of cases.

2.2.2.3 Assumptions

During the project so far, several assumptions were made that form the foundation of the project's planning and execution. These assumptions describe the conditions that are expected to be true but cannot yet be guaranteed. By defining them, we can ensure that the project remains realistic and that potential risks can be identified early on.

Assumption	Description
Data availability	We assume that Indeed and Indeed will continue to have publicly available job postings that can be legally scraped
Data quality	We assume that job postings contain enough structured information to extract useful data
Stakeholder participation	We assume that lecturers and institutes will cooperate by providing curricula to test and validate the system
Tool stability	We assume that Python libraries used for scraping and Azure services will remain available and functional during the project
Planning	We assume that the project can be completed within one semester

Teamwork	We assume that all team members remain available for the duration of the project and are capable of completing their assigned tasks

2.2.2.4 Constraints

Within the project there are several constraints that affect the scope, timeline and technical possibilities. These limitations are mainly related to the available time, resources and tools provided.

By identifying these limitations, we ensure that the business goals remain achievable, and the project stays manageable throughout the entire process.

Constraint	Description
Project timeline	The system must be developed, tested, and presented within one semester. This limits the time available for testing and optimization.
Resource availability	This project is carried out by students, which means available time, expertise, and computing resources are limited. Development and testing are therefore restricted to what can be achieved within the team's skill set.
Technical limitations	The project depends on the available Azure environment provided by HU. Database storage, computing capacity, and access rights are limited to the resources accessible within this environment.
Data access	The system relies on public job postings from Indeed. If access is restricted, data scraping may be partially or temporarily unavailable.
Legal and ethical boundaries	All data collection and processing must comply with GDPR regulations and the ToS of the platforms used (Indeed).
Scope limitation	The CMS focuses solely on marketing job postings from the Dutch market + top 100 companies and the HU marketing programs. International vacancies and other disciplines are currently outside the project scope.

2.2.2.5 *Risk and Contingencies*

A risk analysis is needed to find and manage problems with privacy, technology, and using the competency monitoring system early. This helps take action and make sure the project succeeds.

Risk	Chance	Impact	Weight	Consequence	Measure
Receiving data we did not want to scrape	2/5	2/5	Low	To much data received and its not clear	Filtering list with selected data
Indeed not scrapeable	1/5	5/5	High	No data available for project	Scrape part by part
Indeed partly scrapeable	1/5	5/5	High	Not enough data for project	Scrape part by part
Sick or personal situation of staff	3/5	3/5	Medium	Part of the sick person can't be worked	Report on time/ Other team member picks it up
Indeed under maintenance/ Downtime	2/5	5/5	Medium	Scraping not available at that moment	Waiting till its up
Website response error	3/5	4/5	Medium	System crashes or slow response time	Conduct stress testing: optimized back end performance and surface scaling
HU site has downtime	2/5	4/5	Medium	HU course data temporarily unavailable	Retry connection
Format courses are wrong on the HU website	3/5	3/5	Medium	Wrong or incomplete curriculum in system	Validate and clean course data automatically
Over budget with Azure	2/5	5/5	Medium	Unexpectedly high hosting or processing caused	Monitor resource usage: set budget alerts, optimized data processing
Leak in data privacy	1/5	5/5	High	Violation of GDPR	Encrypt all data
Deadline is not met	3/5	4/5	Medium	Project deliverables delayed	Use agile sprint
Insufficient alignment between project goals and organizational goals.	2/5	3/5	Medium	Final product doesn't meet HU or lectorate expectations	Regular stakeholder meetings

This table gives us a clear understanding of what could go wrong during the project and what we can do to address it. In this way, we ensure that when we encounter any issues, we can deal with them effectively and continue working toward our final goal.

2.2.3 Terminology

In this section, we briefly explain the key terms used within the project. We have done this to ensure that everyone has a shared understanding of the terminology.

Word	Meaning
Scraping	Automatically collecting job postings from websites (e.g., Indeed/company sites) and storing the relevant info in our database for analysis.
Curriculum	The study plan of a degree program: courses + learning outcomes. We upload this to compare it with the labor market.
Matching -algorithm	The (comparison) between two words: skills in the curriculum vs. skills in job postings. The algorithm finds matches and gaps and shows them in the dashboard.
Skill gap	The difference between the skills in the curriculum and the skills employers ask for (missing or weak areas).
Pipeline	The fixed steps data goes through ingest → clean/transform → store → analyze → show in the dashboard (as automated as possible).
GDPR/AVG	The EU privacy law. We work in compliance with GDPR and also follow each website's Terms of Service—meaning we handle access, rate limits, and what is/isn't allowed properly.
NLP	Used to identify skills from job postings and link them to the competencies in the curriculum. It is an AI technique that enables computers to understand and analyze human language.

2.3 Determine Data Mining Goals

2.3.1 Data Mining Goals

Organizations have access to increasingly large amounts of data, but leveraging this information is crucial for making better decisions. Data Mining helps uncover hidden patterns and trends, while methods such as Business Case Model, SWOT-analysis, and the five forces model link these insights to strategy and market positioning. This enables companies to seize opportunities, reduce risk and strengthen their competitive position.

2.3.1.1 The Business Goals of the Competency Monitoring System:

The Business Goals are outlined on page number 12.

2.3.1.2 The Data Mining Goals of the Competency Monitoring System:

1. The automated data collection and processing pipeline will ensure that all relevant data is consistently prepared and available for mining tasks with at least 90% data completeness and no more than 6% duplicate entries. This automated pipeline not only reduces manual intervention but also ensures high data integrity, consistency, and readiness for downstream analytical and machine learning tasks.

2. An automated web scraping system will continuously extract and standardize job posting data, achieving at least 90% extraction accuracy and 90% temporal consistency across weekly samples, to provide a robust and reliable foundation for mining job–skill relationships.

3. A working matching algorithm will be developed that correctly matches job postings to relevant competencies taught by HU programs based on required versus learned skills, achieving a precision score of at least 90%.

2.3.2 Data Mining Success Criteria

Task

The goal of this step is to describe what the data mining part of the project should achieve in technical terms. These goals support the business aim of connecting education with the labour market, improving study programs, and using data to make better decisions.

Data mining goals and success criteria

	Data mining goal	Business Succes Criteria
1	The automated data collection and processing pipeline will ensure that all relevant data is consistently prepared and available for mining tasks with at least 90% data completeness and no more than 6% duplicate entries. This automated pipeline not only reduces manual intervention but also ensures high data integrity, consistency, and readiness for downstream analytical and machine learning tasks.	Within 2 months the processing system is fully automated, so the coordinator needs to do at least 80% less manual data entry compared to the start.
2	An automated web scraping system will continuously extract and standardize job posting data, achieving at least 90% extraction accuracy and 90% temporal consistency across weekly samples, to provide a robust and reliable foundation for mining job–skill relationships.	Around 1 month there is a web scraping system that collects at least 100 job listings within one week. When checked, at least 80 of them have correct and complete information.
3	A working matching algorithm will be developed that correctly matches job postings to relevant competencies taught by HU programs based on required versus learned skills, achieving a precision score of at least 90%.	Around 1,5 months, there is a matching algorithm that connects at least 4 job listing to students. When checked by the coordinator, at least 3 of these matched are found to be relevant.

We succeed if the system mostly works automatically, collects current job listing reliably, makes relevant matches between students and jobs, and delivers a proof-of-concept where soft skills are linked to the curriculum before the end of the semester.

2.4 Product Project Plan

2.4.1 Project Plan

2.4.1.1 Project Overview

For Utrecht University of Applied Sciences, we will develop a website where the lectorate and institutes can make an input with the name of the curricula. In the backend, a program will run that matches these curricula and the associated soft skills with the 2025 marketing jobs in the Dutch market from Indeed.

The choice for a website was made because it is more user-friendly and more accessible for students. Once the system is fully developed, the website can be used everywhere, making the program accessible to anyone who wishes to use it.

By scraping indeed, we will collect job postings and store them in our own database, allowing us to create a reliable match between a student's curriculum and the requirements of companies.

It's also important to make sure that the project is clear which stakeholders are involved. To make sure that everything is clear we made a Stakeholder Analysis.

Stakeholder	**Role**	**Internal/ External**
Hend Elsayed	Client, Final Responsible	Internal
Gerrita van de Ven	lector in the Lectoraat marketing and consumer experience	Internal
HU IT Department	Technical support	Internal
HU Lecturers	Main end users	Internal
Dutch market including top 100 companies	Provide job posting data	External
Other Universities	Potential future users	External
Students	Secondary End Users	Internal

2.4.1.2 Salience Model

For the analysis, we chose the Salience Model because this model helps us evaluate stakeholders based on power, legitimacy, and urgency, allowing us to identify who has the greatest influence and who should be prioritized in the project.

Stakeholder	**Power**	**Legitimacy**	**Urgency**	**Category**	**Strategy**
Client (Hend Elsayed)	✓	✓	✓	Definitive	Bi-weekly meetings & decision-making
Assistant Manager	✓	✓	✗	Dominant	Reports & escalation if needed
HU IT Department	✓	✓	✗	Dominant	Bi-weekly technical meetings
HU Lecturers	✗	✓	✓	Dependent	Bi-weekly demos & feedback
Top 100 Companies	✓	✓	✗	Dominant	Quarterly meetings
Other Universities	✗	✓	✗	Discretionary	Newsletter/quarterly update
Students	✗	✓	✓	Dependent	input on needs

2.4.1.3 Expectations Needs:

Stakeholder	**Expectations**	**Needs**	**Urgency**	**Category**	Strategy
Client (Hend Elsayed)	Working MVP, future-proof system	Reliable data, clear plannin	✓	Definitive	Bi-weekly meetings & decision-making
Assistant Manager	Clear communication, reports	Clear role division	✗	Dominant	Reports & escalation if needed
IT Department	Technically feasible solution	Secure & maintainable system	✗	Dominant	Bi-weekly technical meetings
HU Policymakers	Data for strategic decisions	Reliable, representative analyses	✓	Dependent	Bi-weekly demos & feedback
Lecturers	Insights for curriculum development	Practical results	✗	Dominant	Quarterly meetings
Other Universities	✗	✓	✗	Discretionary	Newsletter/quarterly update
Students	Better aligned education	Up-to-date skills, collaboration	✓	Dependent	input on needs

2.4.1.4 IST-SOLL-GAP

Current Situation (IST):

- Vacancy data is outdated (last used in 2021), making insights no longer representative of the current labor market.
- Data collection is done manually, without automated scraping or updating processes.
- There is no algorithm in place to automatically link competencies to curricula.
- The current process does not provide predictive insights into future competency trends.
- No dashboard is available for visualization or user-friendly interaction for lecturers, students, or policymakers.
- The current solution is only applicable within Utrecht University of Applied Sciences and is not scalable to other institutions.
- There is no complete compliance with GDPR and Terms of Service guidelines when collecting vacancy data.

2.4.1.5 Desired Situation (SOLL)

- The system automatically collects new job vacancies annually through scraping platforms such as Indeed, starting with 2025 data.
- A competency-matching algorithm based on Natural Language Processing (NLP) is used to link skills to curricula.
- An interactive dashboard provides up-to-date insights, allowing lecturers, teachers, students, and policymakers to easily access results.
- The solution is scalable to other educational institutions within the Netherlands (e.g., other research groups or universities of applied sciences).
- All processes comply with GDPR regulations and platform Terms of Service (ToS), with clear documentation on data usage.

2.4.1.6 GAP Analysis:

Category	Current Situation (IST	Desired Situation (SOLL)
Data	Vacancy data is outdated (last updated in 2021) and collected manually.	Up-to-date vacancies are collected annually through automated scraping (Indeed,
Technology	No automation pipeline or system to link job data to curricula.	An automated pipeline using NLP-based matching is implemented.
Output	No visual reporting or clear competency insights.	An interactive dashboard visualizes skill gaps and matching outcomes.
Users	System is only internally accessible by the current project team.	Accessible for lecturers, teachers, students, and policymakers.
Scalability	Usable only within HU and not transferable to other institutions.	Scalable to other Dutch universities of applied sciences or research groups.

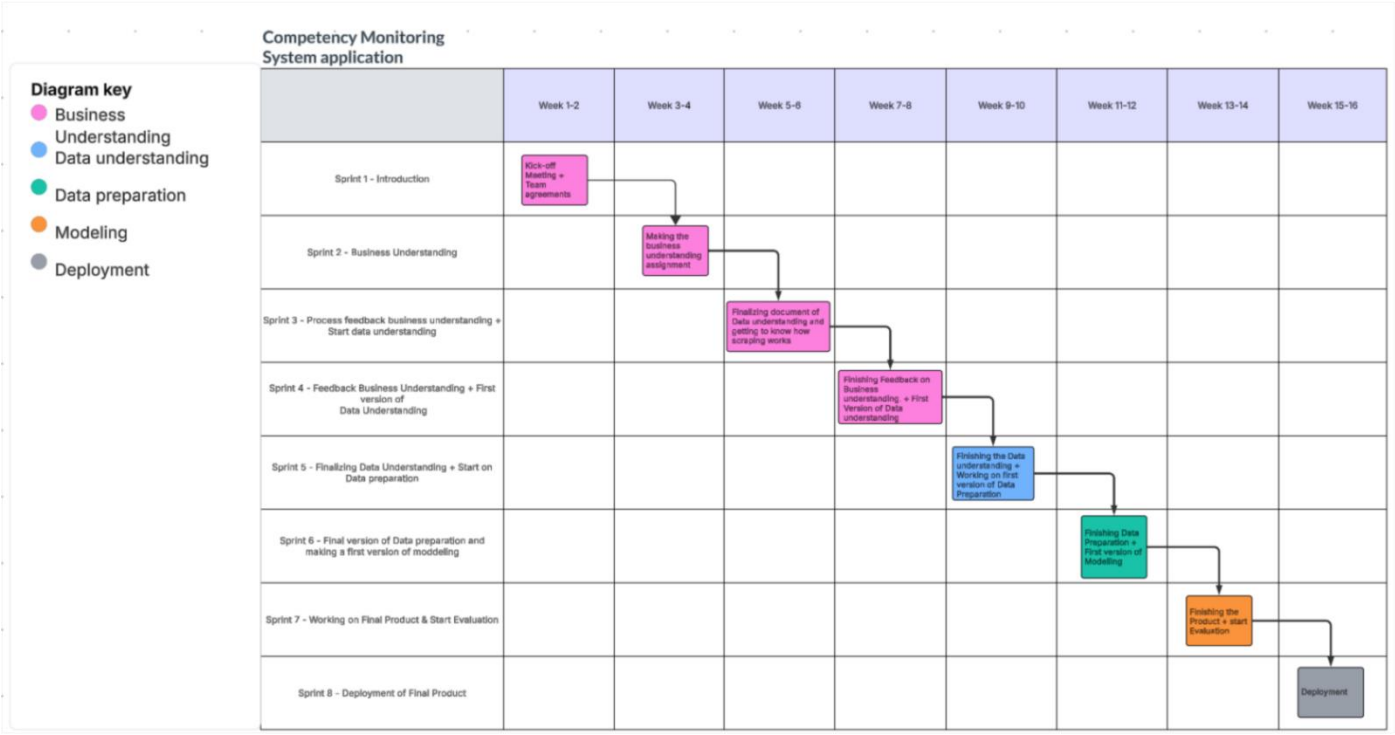
Legislation / Compliance	GDPR and Terms of Service requirements are not fully met when collecting data	Fully compliant with GDPR and platform ToS, including transparent documentation of data usage.

2.4.1.7 Timeline & phases

To create a clear overview of the project flow we are going to define phases. With the phases. The project has a total of 8 phases. These eight phases are each two weeks long. This gives us enough time to make sure that every phase is completed.

Sprint	Week	Goal	**Tasks**	Deliverables
Sprint 1	1-2	Introduction	Kick-off Meeting + Team agreements	Kick-off document. First version of Business understanding
Sprint 2	3-4	Business Understanding	Making the business understanding assignment	Business understanding document
Sprint 3	5-6	Process feedback business understanding + Starting on Data Understanding	Process feedback of teachers into our document	Final product of Business Understanding + start of Data understanding
Sprint 4	7-8	Feedback Business Understanding + First version of Data Understanding	Finishing Feedback on Business understanding. + First Version of Data understanding	Business Understanding + First version of Data Understanding
Sprint 5	8-10	Finalizing Data Understanding + Start on Data preparation	Finishing the Data understanding + Working on first version of Data Preparation	Data understanding + First version of Data preparation
Sprint 6	11-12	Final version of Data preparation and making a first version of moddeling	Finishing Data Preparation + First version of Modelling	Data preparation + first version of modelling
Sprint 7	13-14	Working on Final Product & Start Evaluation	Finishing the Product + start Evaluation	Modelling + Start evaluation
Sprint 8	15-16	Deployment of Final Product	Deployment	Final presentation + Final product

2.4.1.8 Gant-chart



To make the Timeline and Phases even more understandable we made a gant chart. By using a gant chart we can make the project visible with just one look.

If you would like to see the Gant Chart on your own computer you can use the following link:

https://lucid.app/lucidspark/fb1c183d-d8b8-4abd-b1fe-a532e7d8bfdb/edit?viewport_loc=3486%2C1142%2C3454%2C1628%2CuDe-dIt-NWfS&invitationId=inv_2f5f68c1-2fae-4d4c-8053-22f46cc3a8bc

In the Gant chart, you will see:

- What needs to be done
- When will it happen
- Who is working on it
- How everything is connected

When using the Gant chart we can see how far we are with the current project. Plus we can set deadlines and monitor how the project is evolving.

With this overview, we can easily see:

- If we are on schedule
- Where is the delay
- How one adjustment has effect on the rest of the project

In the Gant Chart you see that we already encountered some problems. That's why from sprint four we made new steps in the planning to make sure that everything is still possible just with more work and time in the project. Even when we lost time in the previous sprints.

2.4.1.9 Resources & Roles

The resources and roles within the project have been clearly defined. The resources were previously outlined in the *Inventory of Resources* section on page 15.

The roles have been clarified by creating a table that specifies each team member's responsibilities.

Name	Role
Jason	Leading the project. Responsible for project plan in business understanding with Baha.
Baha	Keeping all project documents organized. Responsible for Project plan in business understanding with Jason.
Zoë	Responsible for inventory of resources and a part of requirements, assumptions and constraints in business understanding.
Yara	Responsible for a part of requirements, assumptions and constraints and terminology in business understanding.
Fatima	Responsible for Initial assesements of tools and techniques and data mining success criteria in business understanding

This is, of course, a summary of how the process works. The specific task assignments are discussed via WhatsApp, where the corresponding deadlines are also communicated.

2.4.1.10 Documentation & Review

All project documents are managed by **Baha**. This ensures that the documentation remains well-organized and forms a coherent narrative, rather than having separate, uncoordinated sections written by different team members. The completed work is reviewed by **Baha** and **Jason** to ensure that all parts align properly and maintain consistency throughout the project.

Every two weeks, we hold a **sprint review** where we present our project progress and upcoming plans. During these sessions, we receive feedback which we then apply to improve our workflow and the overall project outcome.

2.4.2 Initial Assessment of Tools and Techniques.

At the end of the first phase, we reviewed possible tools and techniques for building a prototype of a sustainable, automated competency monitoring system.

The current system predicts which skills are needed in the professional field based on job vacancies. This helps Dutch universities of applied sciences align their study programs with labour market needs. However, it uses outdated data from 2021 and requires manual work.

Given the short timeframe of about four months, the focus is on building a working prototype that:

- Runs the full data process automatically (collecting, cleaning, matching and reporting)
- Matches job skills with program competencies
- Shows result in a simple interactive dashboard
- can be scaled in the future to include new data sources and programs.

2.4.2.1 Purpose of the assignment.

The assessment helps select tools and techniques that best support the project's data mining goals. These tools and techniques will help collect, clean, and analyze job and curriculum data, match skills and show results in a dashboard. The data mining goals and success criteria, which explain what the system should achieve, are listed on the data mining goals page. This assessment makes sure the chosen methods are practical, efficient and suitable for the project's short timeline.

2.4.2.2 Assessment of tools

Tool	Function
Python (pandas, beautifulSoup, scikit-learn, spaCy and Selenium)	Data collection, cleaning, matching and automation.
MSQL / Azure	Data storage and management
Power BI	Data visualization and dashboards
Power BI / RapidMiner	Visual data mining workflows

This project we will primarily use Python for data collection, cleaning, matching and automation. Azure for structured data storage. And we use Power BI for dashboards. This combination is practical, scalable, and open-source.

2.4.2.3 Assessment of techniques

Technique	Purpose
ETL (extract, transform, load)	Automates collecting, cleaning and storing data.
Natural Language Processing (NLP)	Finds and compares skills from job posts and study programs
Keyword matching / similarity check	Compares skills between jobs and what students learn.
Viable system modelling (VSM)	Helps organize and explain the system and curriculum links
Predictive modelling / trend analysis	Predicts future skill needs

This project will mainly use NLP, keyword matching and ETL. These techniques help the system read text, compare skills and automate data collection and cleaning. Optional methods like predictive modelling and VSM can be used for forecasting and curriculum guidance in the future.

This assessment shows the best tools and methods for building a sustainable and fully automated Competency Monitoring System within four months.

The system will use Python for automation and data handling, Azure for data storing, and Power BI for dashboards and reports.

Main techniques include Natural language processing (NLP), keyword matching and ETL (Extract, transform, load). These make it possible to collect, clean, match and report data automatically.

Overall, this setup will deliver a working prototype that needs little manual work, gives clear visual result and forms a strong base for future growth and automation

2.5 Sources

Artificial intelligence in het onderwijs: dit zijn kansen en risico's - Kennisnet

Privacy in education | Privacy First

Cyberveiligheid | Onderwerp | Inspectie van het onderwijs

Meer kennis over AI in het onderwijs nodig, anders risico op discriminatie | PO-Raad

https://www.google.com/aclk?sa=L&ai=DChsSEwj19Lfz3O6PAxVikIMHHfGTNVQYACICCAEQABoCZWY&co=1&gclid=CjwKCAjwisnGBhAXEiwA0zEORzyYBHWiJRQnT7l7LZE5E-sh7ZLsecORnXf-ebppw2rvrqClv5vysRoCA2wQAvD_BwE&cce=2&sig=AOD64_33JlXu4NVM0lQFJ_oyZGfE0q2tGA&q&adurl&ved=2ahUKEwj3ybDz3O6PAxV9-QIHHWhAKkMQ0Qx6BAgYEAE

Bronnen SWOT-analyse:

Hogeschool Utrecht. (z.d). Marketing & Customer experience.

https://www.internationalhu.com/research/marketing-and-customer-experience?utm

Hogeschool Utrecht (z.d). onderwijs en onderzoek.

https://www.hu.nl/werkenbij/onderwijs-en-onderzoek?utm

Hogeschool Utrecht (z.d). creative business.

https://www.hu.nl/voltijd-opleidingen/creative-business/during-the-programme?utm

Hogeschool Utrecht (z.d). Studiegids.

https://studiegids.hu.nl/028c0c1c-9805-4519-93a3-a91528dbdfed?utm

Jermore Knoot. (z.d). Strategische Marketingplan.

https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwilmrPjnu2PAxV09LsIHfm7DWYQFnoECBcQAQ&url=https%3A%2F%2Fwww.strategischmarketingplan.com%2Fmarketingmodellen%2Fvijf-krachten-model-porter%2F&usg=AOvVaw0Ly1QpOSovIT2OoJEFOnQl&opi=89978449

Hogeschool Utrecht (z.d.). *Visie op onderzoek.* Geraadpleegd via https://www.hu.nl/onderzoek/visie-op-onderzoek

GeeksforGeeks. (z.d.). *Functional vs Non-functional Requirements.* Geraadpleegd op 31 oktober 2025, van https://www.geeksforgeeks.org/software-engineering/functional-vs-non-functional-requirements/

Agile Scrum Group (z.d.). *Stakeholder management model (Salience Model).* Geraadpleegd via https://agilescrumgroup.nl/stakeholder-management-model/

Master Challenge (z.d.). *Challenge space registratie.* Geraadpleegd via https://masterchallenge.me/account/challenge-spaces/f0517013-d2fe-411a-b9f9-23235e57d3ec/registration

HU. (z.d). *Marketing & Customer Experience*. Geraadpleegd via https://www.hu.nl/onderzoek/marketing-en-customer-experience#Onderzoekslijnen

3 Data Understanding

3.1 Collect Initial Data

3.1.1 Collect Initial Data

3.1.1.1 Objective

The objective of this task is to collect data regarding the competencies and soft skills described in the HU curricula and the skills required in relevant LinkedIn vacancies, to later analyze alignment between education and job market demand.

Two main data sources were used:

1. The official HU website curriculum pages describing the skills and competencies taught
2. Active Indeed Job posting pages related to these programs

The data collected in this phase will serve as the foundation for the next steps in the CRISP-DM process, where it will be further described, cleaned, and prepared for analysis.

3.2 Data Requirements Planning

The data collection focused on two main perspectives:

The main features we are going to focus on while scraping are the ones you see here below. We have agreed on these characteristics with the client and they are therefore also reflected in the business understanding.

Requirement	**Description**	**Purpose**
Titels job vacancies	The title of the job vacancy on indeed.	We scrape the job title so that someone who looks at it can immediately see what the job is.
Soft skills	The skills you need for the job. These aren't related to the practical aspects of the job itself, but rather to personal matters. For example, teamwork can be a soft skill.	We scrape the soft skills so that a comparison can be made whether they are awarded on the curriculum.
Hard skills	Hard skills are also required for the job. These skills are practical, so there are specific requirements you need to know to get the job. For example: working with Excel.	We scrape the hard skills so that a comparison can be made whether they are awarded on the curriculum.
Location job	This shows the location of the job.	We scrape these so that everyone can see where the track is
Salaris	This shows how much someone will earn while working at this job.	We scrape the salary and then show it so people know what they can earn with this job
Sector	This shows in which sector the job is.	The sector is self-explanatory. This way you can know which sector the job falls under.
Titels curiculum	The title of the curriculum shows what the curriculum stands for and what it falls under	With the title of the curriculum we want to show more what the curriculum is about
HU competencies / soft skills	This shows which hard and soft skills are assigned in the curriculum	With the HU competencies/soft skills, we want to ensure that a comparison can be made whether the competencies explained at the HU correspond with the jobs available on the market.

3.3 Data Sources

Datasets	Source	Type/ Format	Location	Purpose
Curriculum of the HU	Website of the Hogeschool of Utrecht	CSV / Excel (After extraction)		Holds information about Curriculum, description, softskills and competencies
Indeed Vacancies	Indeed (Webscraping)	Excel (Python - export)		Includes 100 vacancies matched with curriculum

These are the datasets we have collected and can work with.

The second main dataset in this project was collected from Indeed.com, using an automated web-scraping process developed in Python. The scraper was specifically designed to gather job postings related to HU programs such as Creative Business, Creative Business: Beyond Campus, and Bedrijfskunde.

The goal of this dataset is to provide an up-to-date representation of the skills and competencies required by employers in the marketing and business sectors. The scraper retrieves structured data including:

- Job Title
- Job Description

Additional: location, salary, and sector (when available)

All collected data was exported to Excel format, and duplicate entries were automatically removed. In total, approximately 100 unique vacancies were gathered per run, ensuring that the dataset is representative and directly comparable to the HU curriculum data.

3.4 Data Acquisition Methodology

3.4.1 HU Curriculum data extraction

To collect the HU Curriculum data, a custom Python automation pipeline was developed using Selenium and SentenceTransformer for both web scraping and semantic text analysis. The goal of this script was to extract soft skill and competencies from every HU program page that is related to marketing and export them to an Excel file in a structured format.

Process of the scraper:

1. Navigate and collect links to all full-time study programs related to marketing.
2. Open each program page, including subpages, and expand hidden sections.
3. Extract full text content and analyze it using a semantic model.
4. Classify and store results in an excel file

The semantic analysis identifies skills not only by keywords but also by meaning simalarity, which improves the accuracy when terms differ in wording.

Error handling ensures that missing pages, duplicate data, or open Excel files do not interrupt the process. Backups are automatically created when needed.

3.4.2 Indeed Vacancy data extraction

The dataset of indeed was collected using a custom code made with python. The main process can be summarized as follows:

1. The code sets up a chrome browser using Selenium in stealth mode to prevent Indeed from detecting automated scraping activity.
2. The scraper opens indeed and searches job based on marketing.
3. It scans the search results and collects 100 available job postings.
4. The scraper opens each individual job posting page one by one.
5. Using BeatifulSoup, it extracts the job title and full job description text from each page.
6. All job postings are stored in a list of dictionaries, which is then converted into a Pandas Dataframe.
7. Duplicates entries are removed, data is sorted by job title, scraped data is numbered, the data is exported to a excel file.
8. The script closes.

3.5 Data Selection Criteria

To ensure that the collected data was relevant, reliable, and consistent with the project's objectives, several Selection Citeria were defined prior to the data collection phase in the business understanding. These criteria guided both the web scraping process and the filtering of records after extraction.

Criterion	**Description**	**Justification**
Relevance to HU Programs	Only data related to official HU Bachelor programs was included. Job Postings were selected using the same program names as search keywords.	Ensures direct connection between educational content and market demand.
Source Validity	Data collected exclusively from the official HU website and Indeed	Guarantees trustworthy and verifiable information
Recency	Job postings limited to active listings at the time of collection	Reflects current labour market trends.
Language	Dutch and English data accepted.	Maintain full Datasets coverage across both languages.
Completeness	Records missing essential fields (Title, Description) were excluded.	Prevents low-quality or empty records.
Duplication Handling	Duplicate vacancies or programs automatically removed using Python	Ensures unique entries and prevents data bias.
Ethical Use	Only publicly data was accessed, Without bypassing authentication	Ensures compliance with research ethics and legal standards.

The criteria ensured that the collected datasets are both representive and analyzable within the project scope. This filtering process improved data quality, reduced noise, and guaranteed ethical compliance in line with institutional research standards.

3.6 Problems Encountered & Solutions

During the data collection and preparation phase, several technical and methodological challenges were encountered. This section summarizes the main issues and the applied solutions to ensure data quality and continuity throughout the process

Problem	**Impact**	**Solution**
Dynamic and unstructured website content	Some HU pages used collapsible sections and inconsistent HTML structures, causing missing or partial text extractions.	Implemented Selenium automation to expand all hidden sections and applied text cleaning before analysis
Inconsistent terminology between HU and job postings	Matching of skills was difficult because similar skills were described differently	Developed a mapping table and used Semantic similarity detection with SentenceTransformer to align equivalent terms.
Incomplete or duplicate Records	Data redundancy and empty entries could affect analysis accuracy	Applied validation steps usings Pandas (drop_duplicates() and text-lenght filters) to ensure unique and complete data
Language variation (Dutch vs English)	Some program pages and vacancies mixed languages, leading to inconsistent text extraction	Accepted both Dutch and English inputs and used a multilingual model for skill detection
File access conflict during excel export	When the Excel file was already open, data export failed.	Added an automatic backup system that saves results as a backup. If there occur any errors
Performance limitations during scraping	Long scraping sessions causes browser timeouts and higher CPU usage.	Optimized the scraping loop, Reduced waiting times, and implemented recovery checkpoints to continue from the last processed record.

Despite technical limitations from indeed and unstructured web content. All identified problems were effectively mitigated.

The final datasets containing curriculum competencies and market-demanded skills were successfully collected in a structured, analyzable format. These solutions ensured the project's data foundation remained Complete, accurate, and ethically complaint.

3.7 Tools & Environment

De tools die wij gebruiken zijn als volgt:

- Visual studio code
- Github
- Excel

Each team member has their own preferred development environment. Our team prefers to work in Visual Studio Code, as it provides a clear and structured overview and makes coding easier and more efficient. The integration with GitHub is especially useful, as it allows us to easily pull and update code changes made by other team members.

Using GitHub also helps prevent overlapping work or accidental overwriting of files. Through good communication and by consistently pushing code in an organized way, we ensure that our work remains synchronized and that nothing gets lost. This approach enables effective and structured collaboration within the team.

The files collected through the scraper are currently stored in Excel files on our local machines. In the future, we plan to expand this process by storing the scraped data directly in a database, allowing for centralized management and improved efficiency.

3.8 Initial Observations

After successfully collecting and consoldating the datasets from both the Hogeschool Utrecht curriculum pages and the Indeed job postings, several initial observations were made regarding the structure, content and consistency of the data.

1. **Variation in terminology**
   The HU curriculum data contains a wide range of wording to describe similar soft skills (e.g., “communicative ability", "communication skills", "strong in collaboration"). This inconsistency indicates that text normalization or mapping will be required during preprocessing
2. **Implicit versus explicit competencies**
   Some HU programs describe competencies indirectly, embedded in narrative text instead of listing them explicitly. This makes automated detection more challenging and reinforces the need for semantic text analysis.
3. **Language distribution**
   Both Dutch and English are present in the datasets. This bilingual characteristic requires consistent handling during natural language processing (NLP) and semantic matching
4. **Data structure integrity**
   The exported excel files are properely formatedd, with each record containing structured fields for program, soft skill and competencies. Initial checks confirm that duplicates were successfully removed and data completeness is above 95%
5. **Duplicate Data**

   After scraping the job postings, I discovered that there were duplicate vacancies included. I adjusted my code to ensure that no duplicate postings are scraped anymore.

6. **Incorrect Data**

   I noticed that the job title and description also included the job posting link, which was, of course, not intended. I modified my code so that the link is no longer included in these fields.

7. **Numbering**

   The numbering of the job postings was not functioning correctly. The vacancies were numbered after being scraped, but not in the correct order — the numbers appeared randomly. After adjusting my code, the vacancies are now numbered in the proper sequence.

The collected datasets provide a solid foundation for further analysis. Although some inconsistencies exist in language and terminology, the semantic similarity approach and mapping table developed earlier will help align the data across sources.

3.9 Ethical and Legal Considerations

- During the scraping process, we carefully adhered to ethical and legal guidelines to ensure compliance with research and data protection regulations.

- All data collected from Indeed consists of publicly accessible job postings. No personal information—such as names, email addresses, or contact details of company employees—was collected or stored.

- To minimize impact, the scraper included randomized delays between scraping actions. The data collection process was carried out in a transparent and responsible manner.

3.10 Output Summary

De scraper heeft dus het volgende verzameld:

- 100 vacatures (waarvan er 86 goed zijn doorgekomen)
- Vacatures zijn marketing vacatures
- Vacatures uit Nederland
- Titel van de vacature
- Beschrijving van de vacature

Deze data wordt vervolgens opgeslagen in een excel bestand en gesorteerd op volgorde van scrapen van 1 tot 86.

3.11 Describe Data

In this step of the data understanding we will be describing data. The purpose of describing data is to get a clear picture of what kind of data we have before starting any analysis. It helps us see how the data is structured, what each field means, and if the data is complete and suitable for our goals.

3.11.1 Describe Data

For this project, we collected two main datasets, one from the HU-school curriculum and the other from job scraping from indeed. Both datasets are stored in Excel files, making them easy to explore and analyze. The goal is to understand the structure, content, and quality of each dataset before any analysis.

3.11.2 Data Description Report

The Hogeschool Utrecht curriculum dataset contains information about three studies: creative business, creative business: beyond campus and bedrijfskunde. The fields are: opleiding, soft skill and competentie. All the data from the Hogeschool Utrecht curriculum dataset is complete, only needs to be translated to english. This dataset is suitable for analyzing the structure of the Hogeschool Utrecht curriculum. (See picture one by format of the data.)
The Indeed jobs datasets are split into seven excel files. Each file contains job vacancy numbers, job titles, and descriptions. All the vacancies are related to Marketing. Some excel files are more detailed, they also have the link to the Indeed job vacancy. (see picture two and three).
The data from Indeed is still being collected, so some of the data is incomplete. Despite this, the available data is sufficient to begin preparation for analysis.

3.11.3 Format of the data

	A	B	C
1	**Opleiding**	**Soft Skill**	**Competentie**
2	Creative Business: Beyond Campus	adaptability	analysevaardigheden
3	Creative Business: Beyond Campus	analyserend	digitale vaardigheden
4	Creative Business: Beyond Campus	besluitvaardig	ondernemerschap
5	Creative Business: Beyond Campus	coachend	onderzoek doen
6	Creative Business: Beyond Campus	coaching	onderzoek uitvoeren
7	Creative Business: Beyond Campus	creatief	onderzoekend vermogen
8	Creative Business: Beyond Campus	creativity	professionele identiteit
9	Creative Business: Beyond Campus	entrepreneurial	strategisch denken
10	Creative Business: Beyond Campus	innovation	teamcoördinatie
11	Creative Business: Beyond Campus	samenwerken	
12	Creative Business: Beyond Campus	teamplayer	
13	Creative Business: Beyond Campus	teamwork	
14	Creative Business: Beyond Campus	vernieuwend	
15	Creative Business	adaptability	ondernemerschap
16	Creative Business	creatief	onderwijs ontwerpen
17	Creative Business	creativity	professionele identiteit
18	Creative Business	critical thinking	projectmanagement
19	Creative Business	entrepreneurial	strategisch denken
20	Creative Business	innovation	teamcoördinatie
21	Creative Business	kritisch denken	
22	Creative Business	samenwerken	
23	Creative Business	teamplayer	
24	Creative Business	teamwork	
25	Bedrijfskunde	aanpassingsvermogen	adviseren met impact
26	Bedrijfskunde	besluitvaardig	onderzoek doen
27	Bedrijfskunde	initiatief tonen	onderzoek uitvoeren
28	Bedrijfskunde	innovation	organisatievermogen
29	Bedrijfskunde	leadership	procesmanagement
30	Bedrijfskunde	leiderschap	professionele identiteit
31	Bedrijfskunde	ondernemend	projectmanagement
32	Bedrijfskunde	probleemoplossend	teamcoördinatie
33	Bedrijfskunde	resultaatgericht	
34	Bedrijfskunde	samenwerken	
35	Bedrijfskunde	teamwork	

Picture one Hu-curricula_dataset(01):

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Nummer**	**Job Title**	**b Description**									
2	1	(Afstudeer	VacaturegegevensDienstverbandStage Locatie1394 Nederhorst den Berg Volledige vacaturetekstV									
3	2	(Afstudeer	VacaturegegevensDienstverbandStage Locatie1394 Nederhorst den Berg Volledige vacaturetekstV									
4	3	(Afstudeer	VacaturegegevensDienstverbandStage Locatie1394 Nederhorst den Berg Volledige vacaturetekstV									
5	4	Adviseur	VacaturegegevensSalaris€ 3.000 - € 4.500 per maand Locatie3512 Utrecht Binnenstad Volledige va									
6	5	Adviseur	VacaturegegevensSalaris€ 3.000 - € 4.500 per maandDienstverbandFulltime LocatieUtrecht Volled									
7	6	Adviseur	VacaturegegevensSalaris€ 3.000 - € 4.500 per maand Locatie3011 Rotterdam Volledige vacaturete									
8	7	Adviseur s	VacaturegegevensSalaris€ 3.000 - € 4.309 per maandDienstverbandFulltime Locatie1114 Duivendrecht&									
9	8	Allround M	VacaturegegevensSalaris€ 3.200 - € 4.000 per maandDienstverbandParttimeFulltime LocatieBreda									

Picture two Indeed-vacatures_dataset(01):

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Nummer**	**Job Title**	**b Descripti**	**Link**								
2	62	(Afstudeer	Vacatureg	https://nl.indeed.com/pagead/clk?mo=r&ad=-6NYlbfkN0Ds2m7hUfqrTEXv-EwZ6Zn9jmuewdPnQsG-								
3	26	(Afstudeer	Vacatureg	https://nl.indeed.com/pagead/clk?mo=r&ad=-6NYlbfkN0Ds2m7hUfqrTEXv-EwZ6Zn9jmuewdPnQsG-								
4	21	(Afstudeer	Vacatureg	https://nl.indeed.com/pagead/clk?mo=r&ad=-6NYlbfkN0Ds2m7hUfqrTEXv-EwZ6Zn9jmuewdPnQsG-								
5	12	(Afstudeer	Vacatureg	https://nl.indeed.com/pagead/clk?mo=r&ad=-6NYlbfkN0AoB3QKuP1AGcgsX4LT5HWNtt4g2SRZHT7f								

Picture three Indeed-vacatures_dataset(01):

3.11.4 Quantity of the data

Dataset of HU-curriculum:

- 35 rows
- 3 columns

It contains the fields opleiding, soft skills and competentie. The data is complete per study program, it only needs translation to english. It is stored in Excel format and containds labeled data describing the relationship between study programs, soft skills and competencies. Although the dataset is relatively small, it is rich in qualitative information and suitable for analyzing the education structure of the HU programs.

Dataset of the job scraping from Indeed:

- 8-86 rows
- 3-4 columns

The Indeed-vacatures_dataset(01) is stored in seven Excel files, each containing betweeen 8-86 rows and 3 to 4 column, depending op the scraping session. The most common fields inlcude job title, job description and vacancy link. Because the data collection is still ongoing, some records are missing links or descriptions. Nevertheless, the dataset currently provides a divers overview of marketing-related job vacancies in the labor market.

3.12 Explore Data

3.12.1 Explore Data

3.12.2 Data Exploration Report

In this report, the dataset of competencies and soft skills within HU programs is explored. The goal is to gain insight into the distribution of the most common competencies and soft skills per program, identify relationships, and highlight interesting patterns for further research

Dataset description

Number of programs: 3

Number of unique competencies: 7

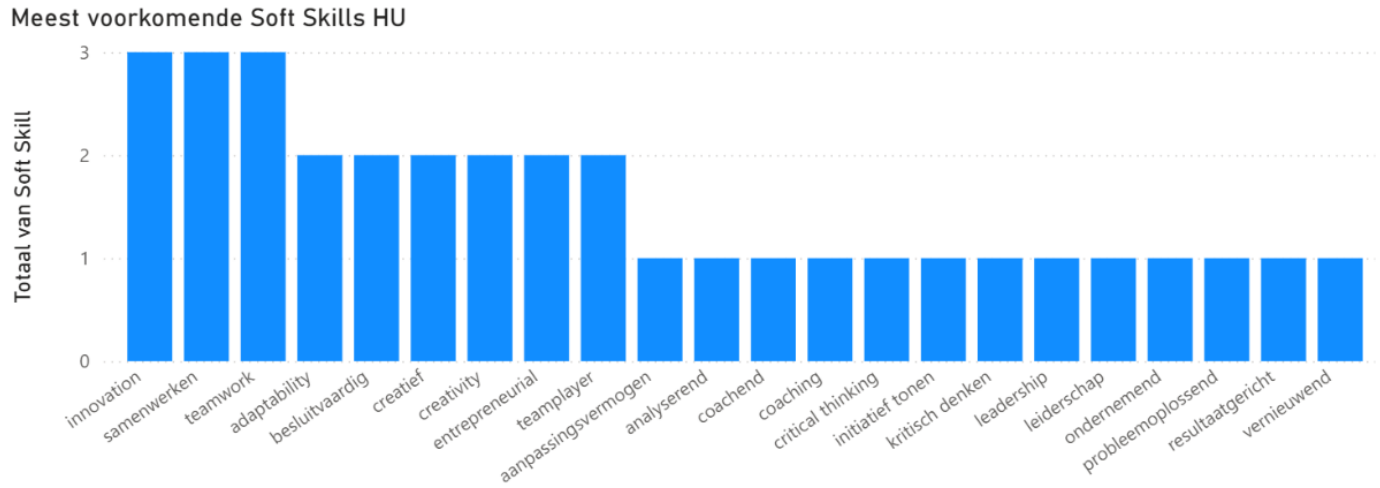
Number of unique soft skills: 13

Source: HU programs dataset

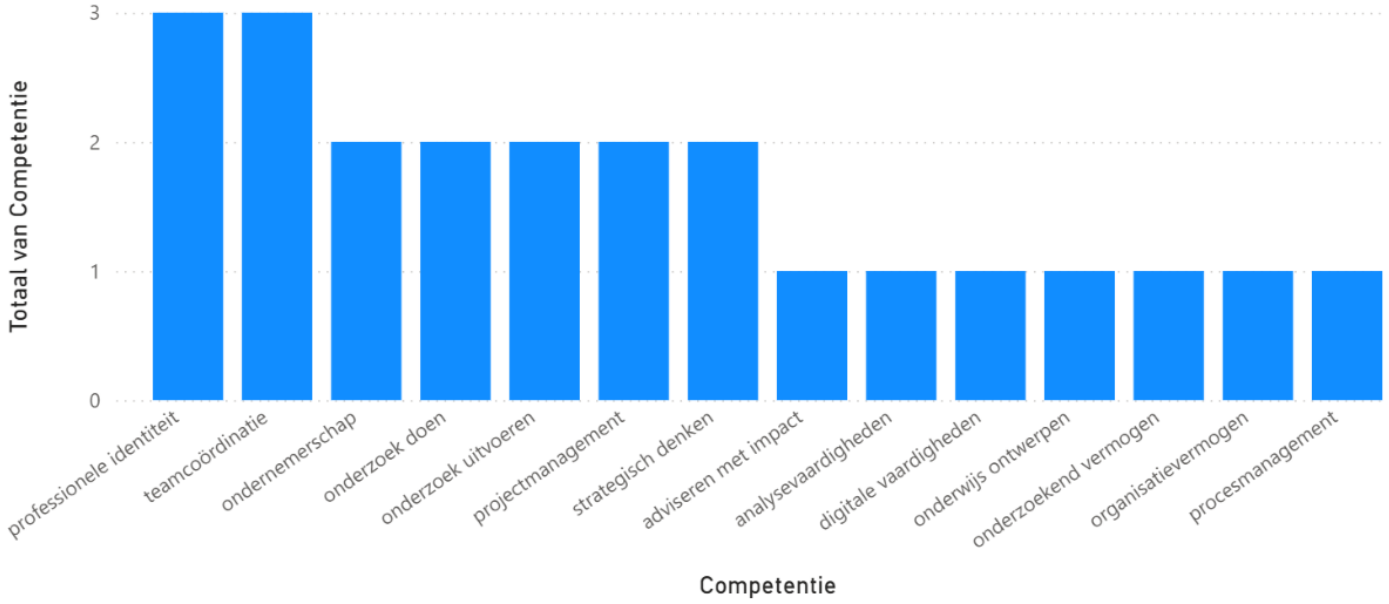
Visualizations

To gain more insights into which skills are central within the HU programs, two bar charts have been created: one with the most common soft skills, and one with the most common competencies. These charts show the total frequency of each skill across all programs, making it clear which skills are most dominant.

This makes trends and patterns in the dataset quickly visible and provides a solid basis for further analyses.



Meest voorkomende competenties HU



3.13 Verify Data Quality

3.13.1 Verify Data Quality

The purpose of this section is to check whether the collected datasets are of good quality. Using clear data quality metrics, we assess whether the data is reliable, complete, and suitable for further analysis within the project. The evaluation is based on five commonly used data quality dimensions:

completeness, uniqueness, consistency, validity, and timeliness.

3.13.2 Dataset 1 – HU (01):

This dataset consists of 35 rows, with each row containing a combination of Education, Soft Skill, and Competency.

Completeness:

Function:

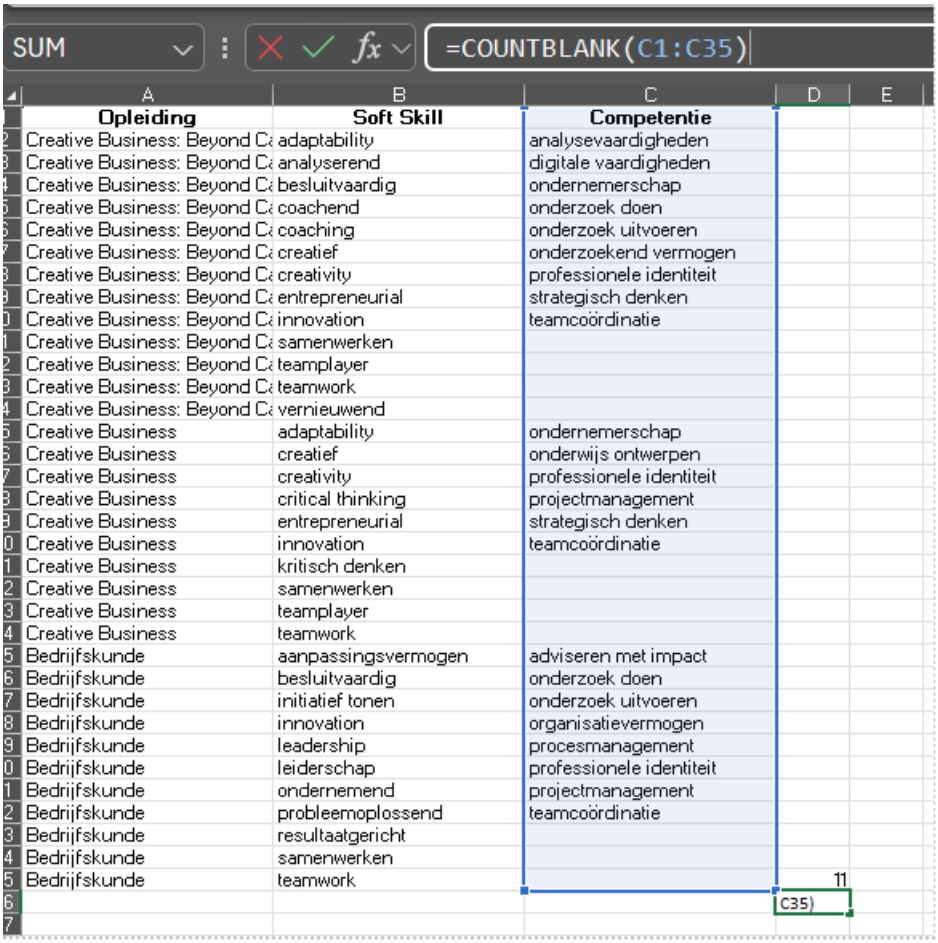
In Excel, the formula =COUNTBLANK(C1:C35) was used to determine the number of empty cells in the Competency column. Then, completeness was calculated using the formula:

(1 − (number of empty cells / total number of rows))(1 − (11/35)))

Result:

A total of 11 empty cells were found, which means the data is approximately 68.6% complete, and about 31% of the competency values are missing.

Checking for Empty Cells Using the Excel Function =COUNTBLANK():



Opleiding	Soft Skill	Competentie
Creative Business: Beyond C	adaptability	analysevaardigheden
Creative Business: Beyond C	analyserend	digitale vaardigheden
Creative Business: Beyond C	besluitvaardig	ondernemerschap
Creative Business: Beyond C	coachend	onderzoek doen
Creative Business: Beyond C	coaching	onderzoek uitvoeren
Creative Business: Beyond C	creatief	onderzoekend vermogen
Creative Business: Beyond C	creativity	professionele identiteit
Creative Business: Beyond C	entrepreneurial	strategisch denken
Creative Business: Beyond C	innovation	teamcoördinatie
Creative Business: Beyond C	samenwerken	
Creative Business: Beyond C	teamplayer	
Creative Business: Beyond C	teamwork	
Creative Business: Beyond C	vernieuwend	
Creative Business	adaptability	ondernemerschap
Creative Business	creatief	onderwijs ontwerpen
Creative Business	creativity	professionele identiteit
Creative Business	critical thinking	projectmanagement
Creative Business	entrepreneurial	strategisch denken
Creative Business	innovation	teamcoördinatie
Creative Business	kritisch denken	
Creative Business	samenwerken	
Creative Business	teamplayer	
Creative Business	teamwork	
Bedrijfskunde	aanpassingsvermogen	adviseren met impact
Bedrijfskunde	besluitvaardig	onderzoek doen
Bedrijfskunde	initiatief tonen	onderzoek uitvoeren
Bedrijfskunde	innovation	organisatievermogen
Bedrijfskunde	leadership	procesmanagement
Bedrijfskunde	leiderschap	professionele identiteit
Bedrijfskunde	ondernemend	projectmanagement
Bedrijfskunde	probleemoplossend	teamcoördinatie
Bedrijfskunde	resultaatgericht	
Bedrijfskunde	samenwerken	
Bedrijfskunde	teamwork	

Uniqueness:

Function:

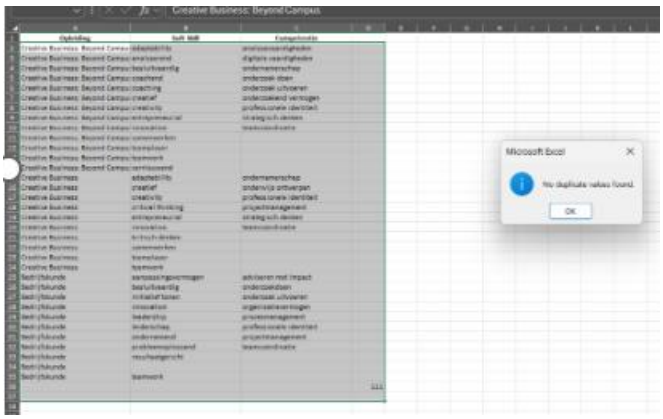
In Excel, the (Remove Duplicates) option was used on the columns Education, Soft Skill, and Competency.

Result:

No duplicate records were found — all rows are unique.

Each combination of education, soft skill, and competency appears only once.

Therefore, the dataset is 100% unique and reliable.



Consistency

Function:

The dataset was sorted by columns to identify differences in spelling or formatting.

Additionally, column names and structures were visually checked to ensure that all tables share the same layout.

Result:

The column structure is consistent throughout, but some values appear in both Dutch and English (e.g., creatief and creativity).

The dataset is structurally consistent, but not fully linguistically uniform.

These inconsistencies will be standardized during the Data Preparation phase.

Geldigheid:

Function:

The values in the columns were checked to ensure they are logical and appropriate within the context of education.

Result:

All values are meaningful and correct. No incorrect or irrelevant terms were found.

The dataset is fully valid.

Timeliness
Checked which academic year the dataset originates from.

Result:
The data comes from the HU Curriculum 2024–2025.

The dataset is current and representative of the present educational offering.

Dataset 2 Indeed:

Here, the Indeed dataset is evaluated based on data quality.

Completeness:

No empty values were found in the Job Title and Job Description columns.

This means the dataset is complete and comprehensive.

Uniqueness:

Result: no duplicates were found. Therefore, the dataset contains unique job postings.

Consistency

Result:

The column structure is consistent, and all columns contain valid data.

It was noted, however, that some job titles are in English and others in Dutch.

Geldigheid:
De waarden in de kolommen *Job Title*, *Job Description* en *Link* zijn handmatig gecontroleerd op logische en correcte inhoud.

Currentness:

4 Data Preparation

4.1 Taak *n*

4.1.1 Product *n*

4.1.2 Product *n*

4.1.3 Product *n*

4.2 Taak *n*

4.2.1 Product *n*

4.2.2 Product *n*

Product *n*

5 Modeling

5.1 Taak *n*

5.1.1 Product *n*

5.1.2 Product *n*

5.1.3 Product *n*

5.2 Taak *n*

5.2.1 Product *n*

5.2.2 Product *n*

5.2.3 Product *n*

6 Evaluatie - Deployment

6.1 Taak *n*

6.1.1 Product *n*

6.1.2 Product *n*

6.1.3 Product *n*

6.2 Taak *n*

6.2.1 Product *n*

6.2.2 Product *n*

6.2.3 Product *n*

7 Feedback

7.1 Docenten

7.2 Sprint Release

7.3 Vragen ontvangen van critical friends

7.4 Vragen gesteld als critical friends