Competency Monitoring System application

CRISP-DM

Jason Tomeij, Baha Kucuk, Fatima Memon, Yara Mousa & Zoë González

Version Management

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

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 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.

Table of Contents

Version Management 2
Introduction 3
Table of Contents 4
Collect Initial Data 5
Describe Data 15
Explore Data 18
Verify Data Quality 20

Collect Initial Data

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.

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.

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.

Data Acquisition Methodology

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.

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.

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.

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.

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.

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.

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.

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.

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.

Our Datasets

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.

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 shown in an Excel file, now there are 7 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.

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

	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 vacaturetekstW									
3	2	(Afstudeer	VacaturegegevensDienstverbandStage Locatie1394 Nederhorst den Berg Volledige vacaturetekstW									
4	3	(Afstudeer	VacaturegegevensDienstverbandStage Locatie1394 Nederhorst den Berg Volledige vacaturetekstW									
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

	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

Quantity of the data

Dataset of HU-curriculum:

- 35 rows
- 3 columns

Dataset of the job scraping from Indeed:

- 8-86 rows
- 3-4 columns

Explore Data

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

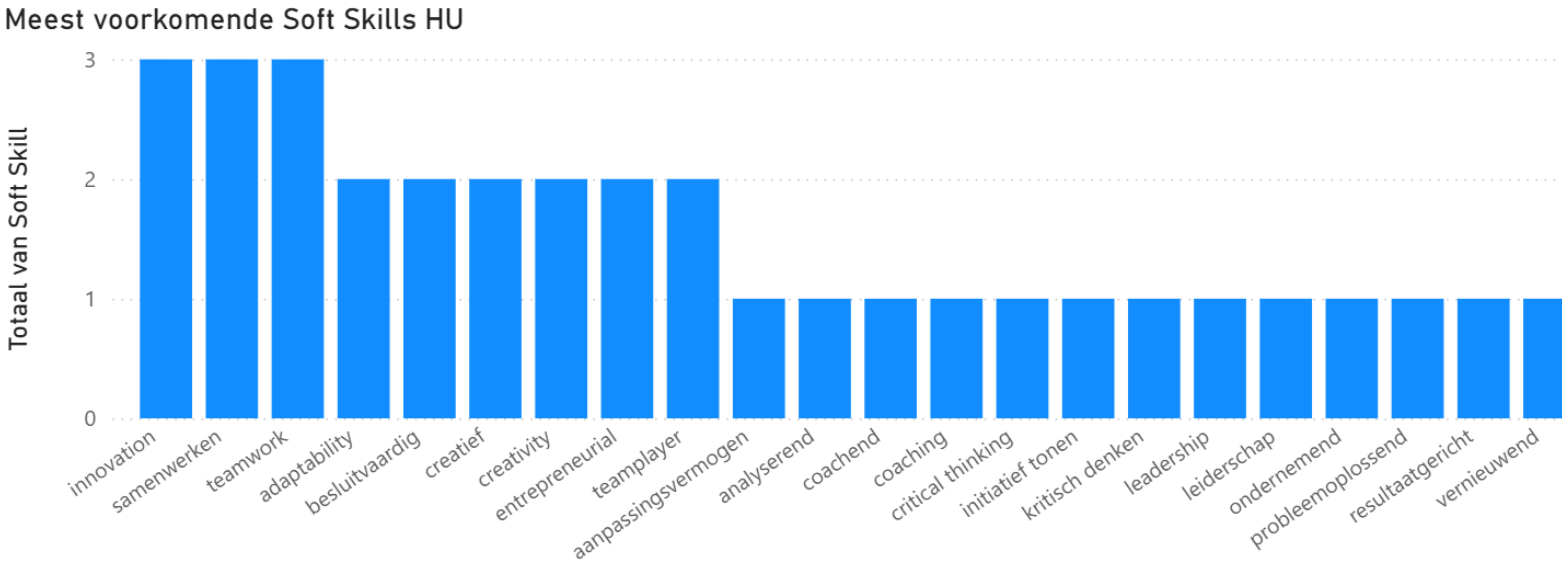
Number of unique soft skills: 13

Source: HU programs dataset

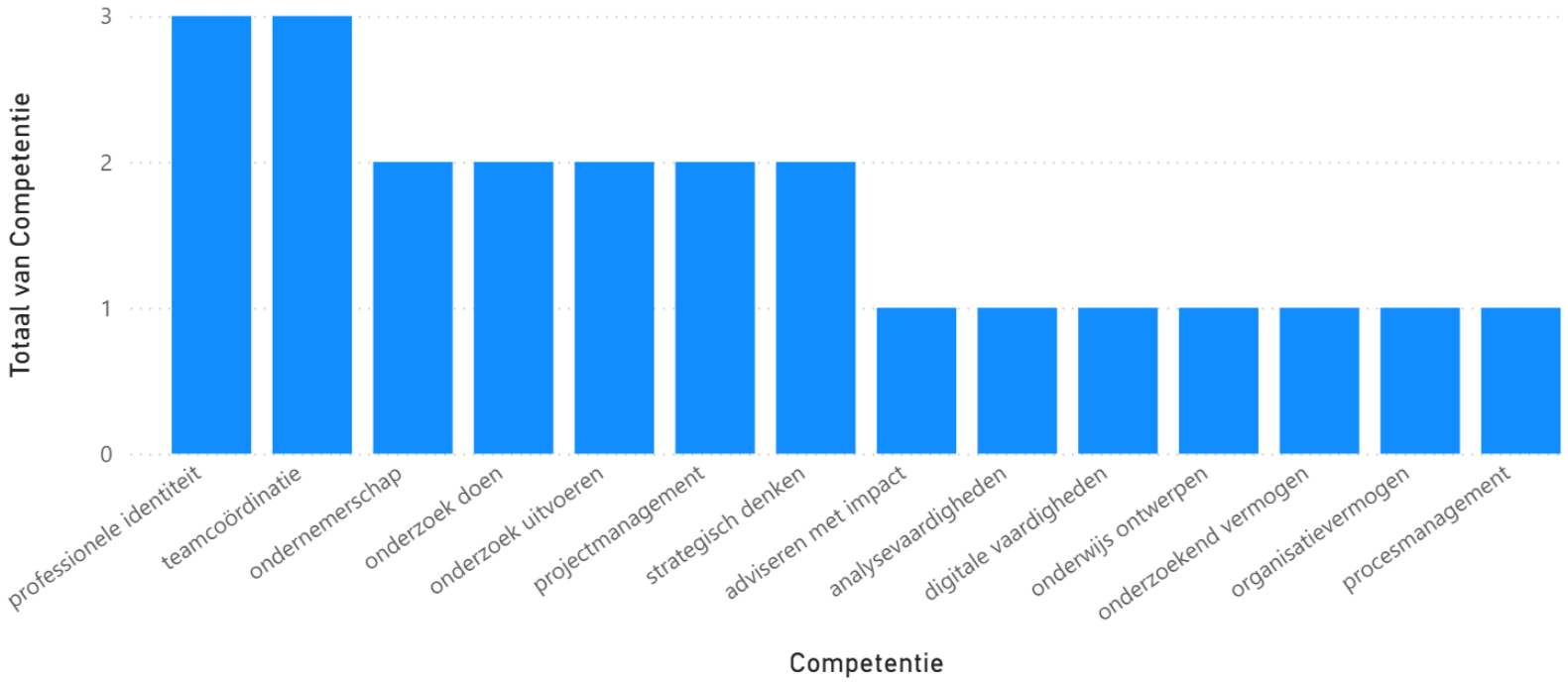
Visualizations

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



Verify Data Quality

Data Quality Metrics

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.

Data Quality Dimensie	Purpose	Method	Metric	Reason
Completeness	Are there empty cells	Count empty cells using Excel function =COUNTBLANK( )	% of non-empty cells	Missing values reduce the reliability of analyses.
Uniqueness	Are there duplicate records?	Use Excel “Remove Duplicates” function	% of unique rows	Duplicates can distort results
Consistency	Are formats and column structures uniform?	Compare columns and check for structural consistency	% of consistent values	Inconsistent columns make it difficult to merge datasets.
Validity	Are the values logical and correct?	Check data types, and expected values	% of valid records	Invalid values can cause errors in analyses.
Timeliness	Is the data recent and representative?	Compare the collection date with the current date met de huidige datum.	< 60 days old	Old data may no longer reflect the current labor market.

Dataset 1 – HU Curriculum

Dimension	Measured Metric	Metric Achieved	Comments
Completeness	11 empty cells out of 34 programs (in the *Competency* column) → 67.6% complete	x	Some *Competency* values are missing, which may affect the completeness of the analysis.
Uniqueness	100% unique	✓	Each combination of *Program – Soft Skill – Competency* appears only once.
Consistency	Structure and column layout are uniform	✓	Some terms are mixed (English/Dutch); this can be standardized later during the Data Preparation phase.
Validity	All values are logical and correctly entered	✓	The remaining values are accurate and usable.
Timeliness	Dataset originates from the HU Curriculum 2024–2025	✓	Data is up to date.