Data Mining SS25

Agenda

- Company August Rüggeberg
- Topic of the semester project
- Data sources
- CRISP-DM
- General information

Company

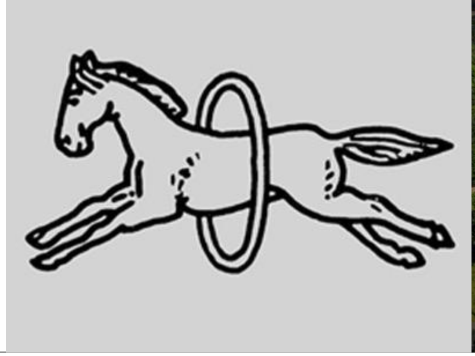
August Rüggeberg

Company description

August Rüggeberg (PFERD) is leading in the

- development
- production
- support
- and distribution

of tool solutions for work on surfaces and material cutting.



PFERD
Germany
Safety code
80 m/s max.
12 200 RPM
CC-GRIND-SOLID
PFERD
ALUMASTER
SD-F 115/125 ALUMASTER
Ne convient pas pour le tronçonnage!
¡No usar para corte!
EH 4/1
PFERD
Germany
NOX®-PNER
H 7506-6 A grob
H 7506-6 A coarse
hart - hard
PFERD
www.pferd.com
Rundbürste
Made in Germany
PFERD
Germany
Safety code
EN 12 413
80 m/s max.
8 500 RPM
PFERD
Safety code
EN 13 743
80 m/s max.
12 200 RPM
125 x 20 x 22,23 mm
POLIFAN
PFC 125 Z40
PFERD
Germany
PFERD
Germany
www.pferd.com
PFERD
n. óptim. 6300
80
PFERD

August Rüggeberg

Customers

- Metalworking is the primary focus among our top customers
- Key focus areas include aerospace, shipbuilding, and automotive, though the tools are also used in many other industrial sectors.

Foundry

Aerospace

Shipbuilding

Topic of the semester project

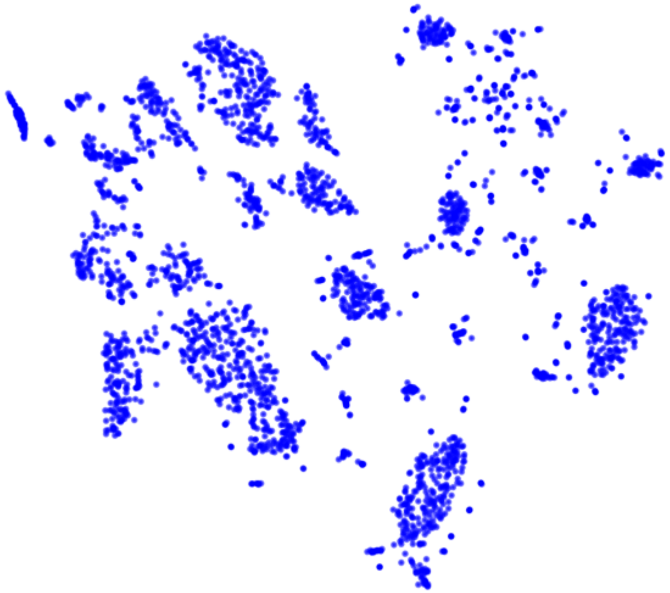
Topic of the semester project

Topic area world economy

Approach: Comprehensive Market Analysis using clustering methods and tailored feature selection

Indicators:

- **Economic Indicators** (GDP, economic growth)
- **Demographic Indicators** (population size, population growth)
- **Trade Data** (import/export volumes of related products)
- **Business-Relevant Metrics** (Steel consumption (use-case reference))
- **Soft Factors** (Corruption Index, Economic Freedom Index)



Goal: Group countries into comparable market clusters to uncover high-potential markets.

Data sources

Data sources

The Global Ecomomy (Link)

theGlobalEconomy.com

Learning resources and data on the world economy

Platform providing reliable economic data for over 200 countries (since 1960)

- Includes 500+ indicators from:
  - Central banks,
  - National statistics offices,
  - International organizations

- Goal: Save time and effort through well-documented, downloadable data
- Free access for: Educators & students from low-income countries
- Founded in 2012 at Georgia State University (USA)

- Selection of approx. 800 annual key features are provided in a python df

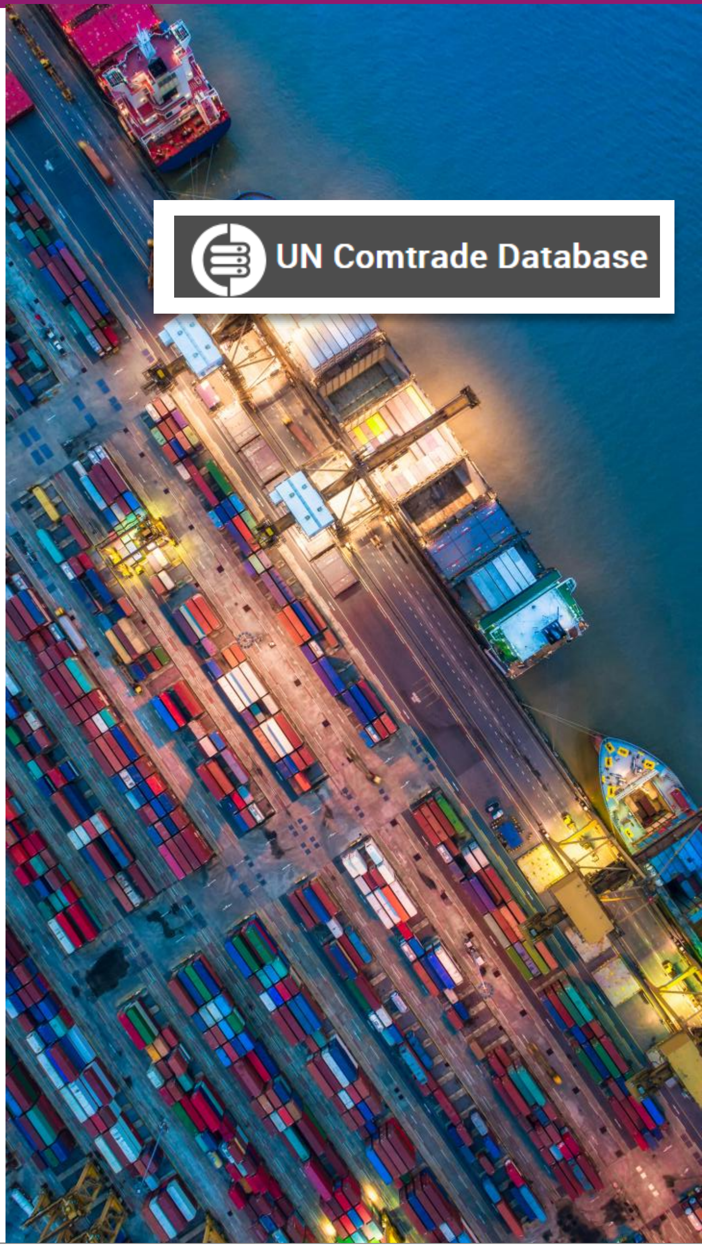
Data sources

UN Comtrade (Link)

- Global trade database maintained by the United Nations
- Contains detailed import and export statistics from over 170 reporting countries
- Covers more than 5,000 product categories, classified by:
  - HS (Harmonized System), SITC, BEC, and other classification systems

- Data reported by national customs authorities and standardized for international comparison
- Available at annual and monthly frequency, depending on country
- Includes trade values and quantities, partner countries, and trade flows

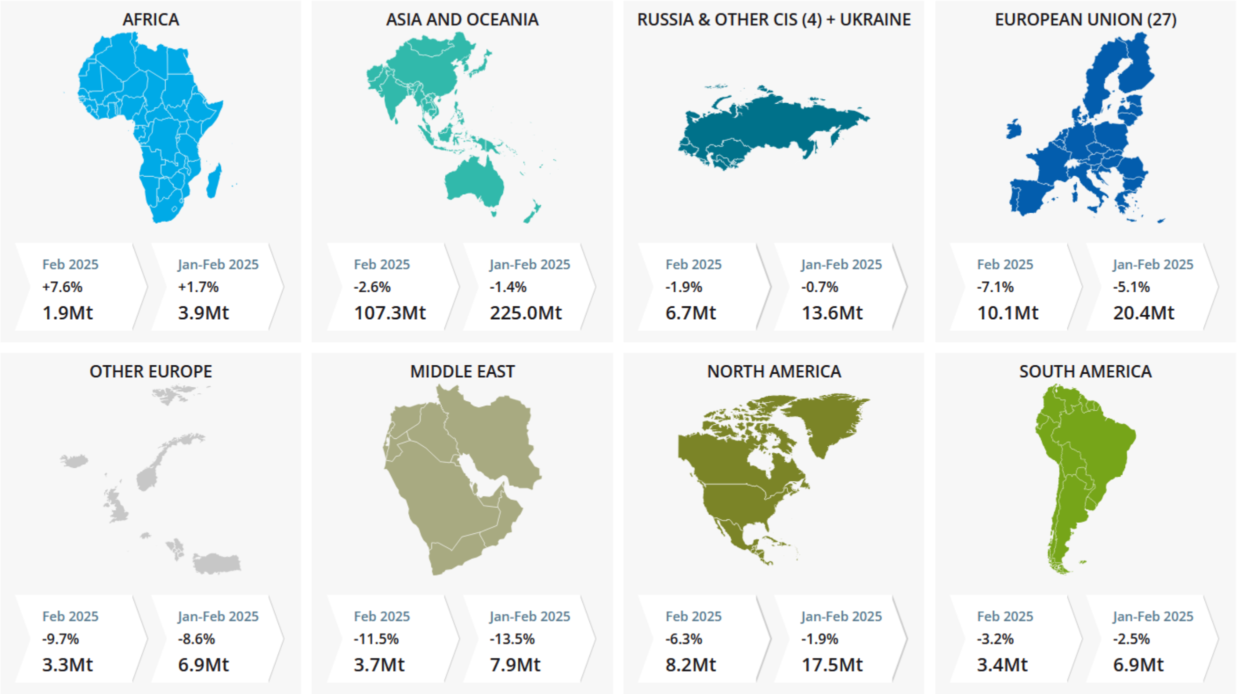
- Data must be downloaded and pre-processed independently



Data sources

World Steel Association (Link)

- One of the largest industry associations in the world
- Represents steel producers, national/regional associations, and research institutes
- Covers around 85% of global steel production
- Provides comprehensive data and statistics on:
  - Crude steel production (monthly & annual)
  - Steel use by country and sector
  - Steel trade flows and demand forecasts



AFRICA
Feb 2025 +7.6% 1.9Mt
Jan-Feb 2025 +1.7% 3.9Mt

ASIA AND OCEANIA
Feb 2025 -2.6% 107.3Mt
Jan-Feb 2025 -1.4% 225.0Mt

RUSSIA & OTHER CIS (4) + UKRAINE
Feb 2025 -1.9% 6.7Mt
Jan-Feb 2025 -0.7% 13.6Mt

EUROPEAN UNION (27)
Feb 2025 -7.1% 10.1Mt
Jan-Feb 2025 -5.1% 20.4Mt

OTHER EUROPE
Feb 2025 -9.7% 3.3Mt
Jan-Feb 2025 -8.6% 6.9Mt

MIDDLE EAST
Feb 2025 -11.5% 3.7Mt
Jan-Feb 2025 -13.5% 7.9Mt

NORTH AMERICA
Feb 2025 -6.3% 8.2Mt
Jan-Feb 2025 -1.9% 17.5Mt

SOUTH AMERICA
Feb 2025 -3.2% 3.4Mt
Jan-Feb 2025 -2.5% 6.9Mt

CRISP-DM cycle with regard to the project

CRISP-DM cycle

Model planning 50%-70%

Business Understanding

- What is the business use case?
- What is the aim of the project?
- What are the stakeholders?
- When is a market interesting for PFERD?
- Which features could be relevant?

Data Understanding

- Have I chosen the right features?
- Are there redundant features?
- Is the data available for many markets?
- How should the various data sets be prepared?
- How do I combine the data sets into a df?
- How must the data be provided for clustering?

Data Preparation

- Is the data suitable for modeling?
- Feature engineering?
- How is the data distributed?
- How to deal with outliers?
- How to deal with missing values?
- Does the data need to be scaled?
- Which method do I use for scaling?
- Must all columns of the df be scaled?

CRISP-DM cycle

Model development 10%-20% & Model deployment / presentation 20%-40%

Modeling

- Which clustering methods are suitable?
- How to find the best model parameters?
- How to save model parameters?
- Do the data need to be adjusted again?

Evaluation

- Does clustering make sense in a visual inspection?
- Which metrics can I use for evaluation?
- What are the best clusters, and is that plausible?
- Are there differences between the methods?
- Which features have the greatest impact?
- Are the results relevant to the use case?

Model (deployment) / presentation

- What are the most important results?
- How can modeling results be a strategic recommendation regarding the use case?
- How do I present my results?
- How do I create trust among stakeholders?

General information

General information

Data:

- Datasets will partly be published in a GitHub repo with further information
- Datasets are stored in the pickle file format, which can be easily loaded into a python dataframe

Procedure:

- After the group assignment, you will be assigned to a GitHub repo
- The project documentation takes place in the repo
- If you have any questions, you can contact me (lukas.bader@th-koeln.de)

Interesting links:

- Country similarity index (use of a distance matrix), Link