



Data Driven Modeling WS24/25

11.10.2024

Lukas Bader

Data Mining SS24

Seite 1

Fakultät für Informatik und Ingenieurwissenschaften
Institut für Data Science, Engineering, and Analytics

Technology
Arts Sciences
TH Köln

Agenda

- Company August Rüggeberg
- Topic area US economy
- Classification Code-Systems
- Data sources

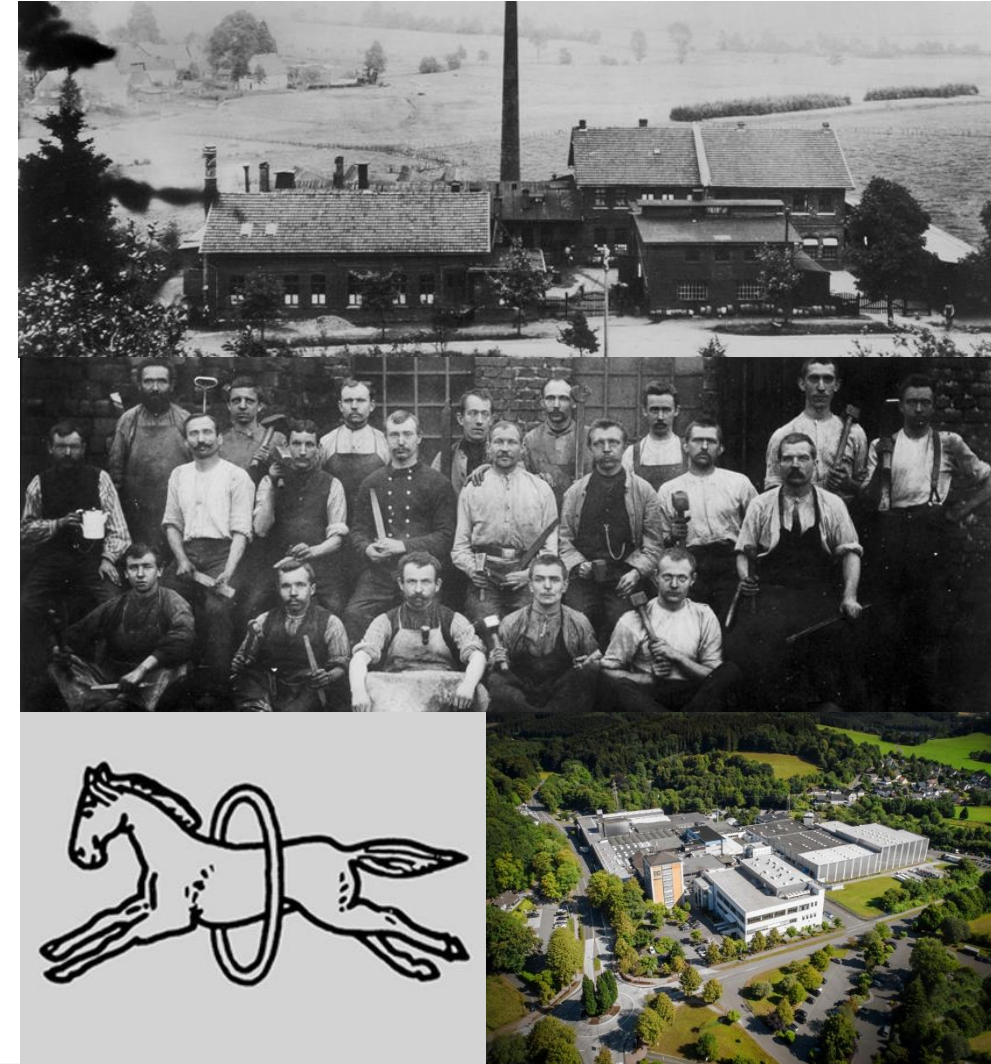
August Rüggeberg

Company description

August Rüggeberg (PFERD) is leading in the

- development
- production
- support
- and distribution

of tool solutions for work on surfaces and material cutting.





Topic area US economy

Topic area US economy

- The focus of this project will be on the US economy
- Various indicators are available to analyze the us economy
- The focus will be on the manufacturing industry with a view to the industries that process metal
- **Identification of regions with high potential for product sales with Clustering**



Foundry



Aerospace



Shipbuilding



Classification Code-Systems

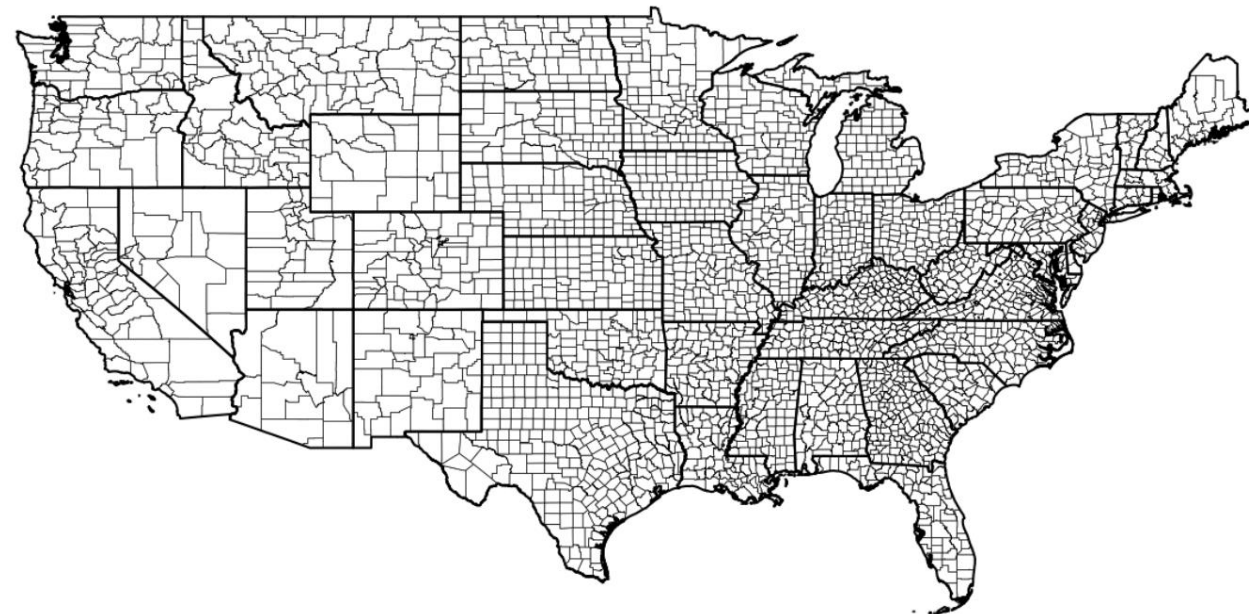
Classification

Classification

Geographical classification

The USA is divided geographically and statistically into different regions. The economic indicators we will use are classified according to the Federal Information Processing Standard Codes ([FIPS Codes](#))

- FIPS 01001:
 - 01003 State 01 = Alabama
 - 01003 County 003 in Alabama = Baldwin County
 - 01999 County 999 in Alabama = unknown
- 56 States
- 3233 Counties



Classification

Classification

Industry classification

The North American Industry Classification System ([NAICS](#)) is the standard used by Federal statistical agencies in classifying business establishments

- Sector: 2-digit code
 - Subsector: 3-digit code
 - Industry Group: 4-digit code
 - NAICS Industry: 5-digit code
 - National Industry: 6-digit code

Level	NAICS Code	Title
Sector	44-45	Retail Trade
Subsector	441	Motor Vehicle and Parts Dealer
Industry Group	4412	Other Motor Vehicle Dealers
NAICS Industry	44122	Motorcycle, Boat, and Other Motor Vehicle Dealers
National Industry	441222	Boat Dealers

Focus on:

21- Mining, Quarrying, and Oil and Gas Extraction	23- Construction	31-33- Manufacturing
---	------------------	----------------------

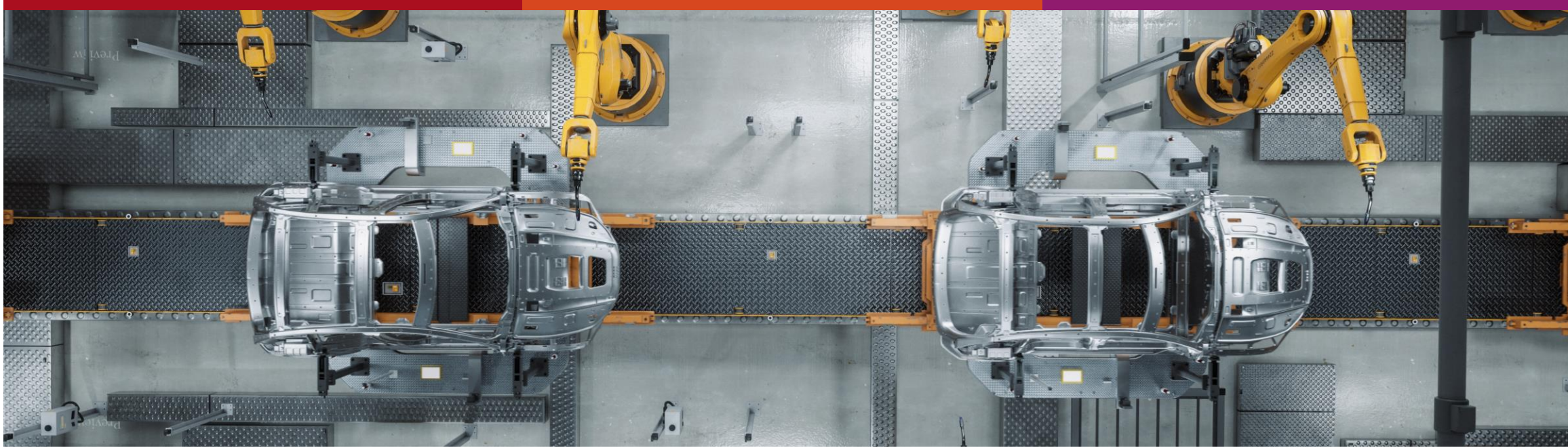
Classification

Classification

Occupation classification

The 2018 Standard Occupational Classification ([SOC](#)) system is a federal statistical standard used by federal agencies to classify workers into occupational categories.

- Consisting of six digits with a hierarchical structure
- Focus on 28 occupations that occur in the metalworking industry, e.g:
 - 47-2221 [Structural Iron and Steel Workers](#)
 - 49-3011 Aircraft Mechanics and Service Technicians
 - 49-9041 Industrial Machinery Mechanics
 - 51-4121 Welders, Cutters, Solderers, and Brazers



Data Sources

Data sources

GDP by industry and region

Granularity		
Industry (NAICS)	Regional	Time
2 digit	County	Year (2022)
Source: U.S. Bureau of Economic Analysis		

- Indicator is the GDP for various industries (2-digit NAICS codes) for the year 2022 on a county level
- Provides information on the value of goods produced in specific industries



	FIPS	GeoName	Region	TableName	LineCode	IndustryClassification	Description	Unit	2022
	0	0	United States	CAGDP9	1	...	All industry total	Thousands of chained 2017 dollars	2.182204e+10
	1	0	United States	CAGDP9	2	...	Private industries	Thousands of chained 2017 dollars	1.928312e+10
	2	0	United States	CAGDP9	3	11	Agriculture, forestry, fishing and hunting	Thousands of chained 2017 dollars	1.700580e+08
	3	0	United States	CAGDP9	6	21	Mining, quarrying, and oil and gas extraction	Thousands of chained 2017 dollars	2.452900e+08
	4	0	United States	CAGDP9	10	22	Utilities	Thousands of chained 2017 dollars	3.183920e+08

108047	98000	Far West	8	CAGDP9	88	42,44-45	Trade	Thousands of chained 2017 dollars	4.292295e+08
108048	98000	Far West	8	CAGDP9	89	22,48-49	Transportation and utilities	Thousands of chained 2017 dollars	1.903768e+08
108049	98000	Far West	8	CAGDP9	90	31-33,51	Manufacturing and information	Thousands of chained 2017 dollars	1.010181e+09
108050	98000	Far West	8	CAGDP9	91	...	Private goods-producing industries 2/	Thousands of chained 2017 dollars	6.650949e+08
108051	98000	Far West	8	CAGDP9	92	...	Private services-providing industries 3/	Thousands of chained 2017 dollars	3.244339e+09

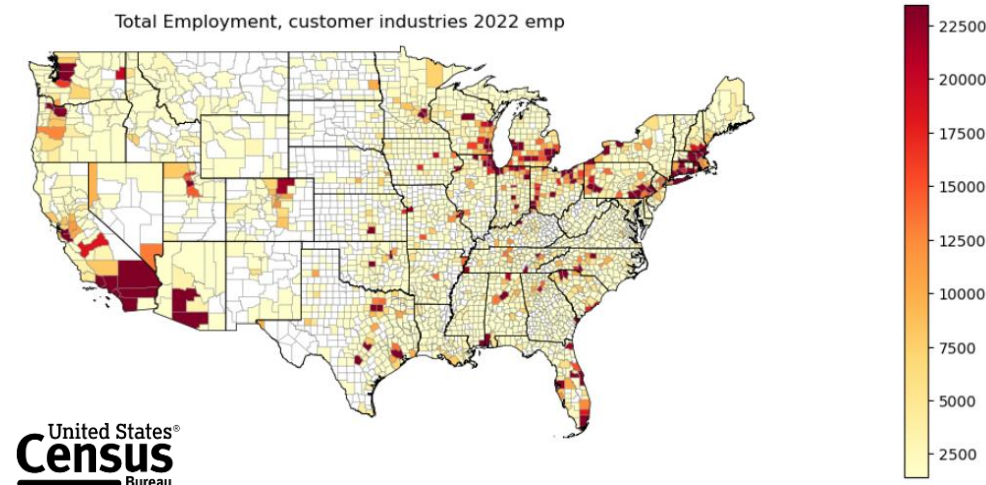
108052 rows × 9 columns

Data sources

Number of companies, employees + wages by industry and region

Granularity		
Industry (NAICS)	Regional	Time
4 digit	County	Year (2022)
Source: U.S. United States Census Bureau		

- Indicators are:
 - Number of employees
 - Number of establishments
 - Company size
 - Wages
 - Noise for each indicator



- Provides deeper insights into the distribution of industries in the USA

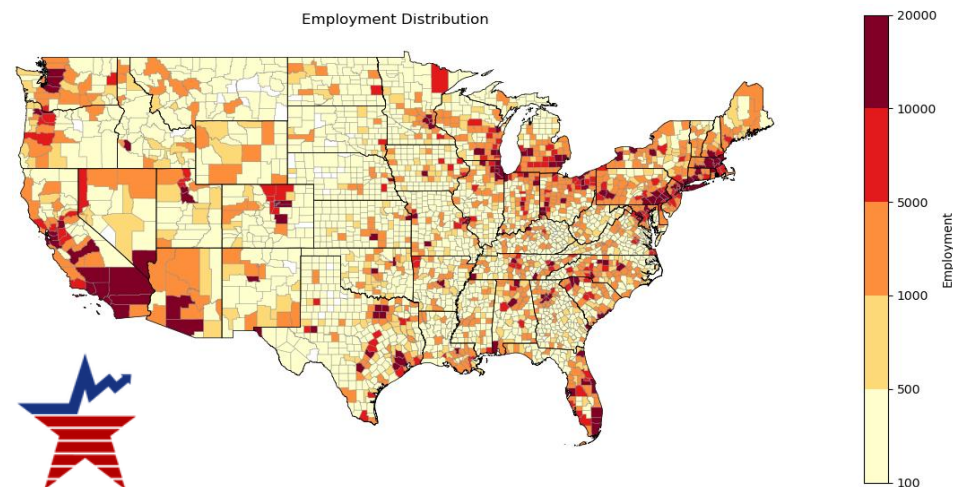
	State_GEOID	County_GEOID	FIPS	naics_2	naics	DESCRIPTION	emp_nf	emp	qp1_nf	qp1	...	n20_49	n50_99	n100_249	n250_499	n500_999	n1000	n1000_1	n1000_2	n1000_3
0	01	001	1001	11	1133	Logging	G	68	G	1213	...	N	N	N	N	N	N	N	N	N
1	01	001	1001	21	2123	Nonmetallic Mineral Mining and Quarrying	G	87	G	1224	...	N	N	N	N	N	N	N	N	N
2	01	001	1001	22	2211	Electric Power Generation, Transmission and Di...	G	129	G	4418	...	N	N	N	N	N	N	N	N	N
3	01	001	1001	23	2361	Residential Building Construction	J	23	G	166	...	N	N	N	N	N	N	N	N	N
4	01	001	1001	23	2362	Nonresidential Building Construction	H	69	H	685	...	N	N	N	N	N	N	N	N	N

Data sources

Number of employees in specific occupations and industries

Granularity		
Industry (NAICS)	Regional	Occupation (SOC-Code)
4 digit	County	6 digit
Source: U.S. United States Census Bureau		

- Combination of NAICS data and occupation data
- Indicator is the number of employees for specific combinations of occupations by SOC code and industries by NAICS code at county level
- Data is pre-filtered for prioritized occupations from the metalworking industry



	FIPS	State_GEOID	naics	NAICS_TITLE	emp_total_county_naics	OCC_CODE	OCC_TITLE	emp_occupation
0	13073	13	3211	Sawmills and Wood Preservation	116	45-4021	Fallers	0.064409
1	13073	13	5617	Services to Buildings and Dwellings	945	45-4021	Fallers	0.106189
2	13075	13	1133	Logging	14	45-4021	Fallers	0.149708
3	13075	13	5617	Services to Buildings and Dwellings	14	45-4021	Fallers	0.001741
4	13077	13	5617	Services to Buildings and Dwellings	799	45-4021	Fallers	0.088781
...
658567	55141	55	4240A3	Miscellaneous Nondurable Goods Merchant Wholes...	147	51-9031	Cutters and Trimmers, Hand	0.038153
658568	55141	55	5613	Employment Services	405	51-9031	Cutters and Trimmers, Hand	0.114460
658569	55999	55	4240A3	Miscellaneous Nondurable Goods Merchant Wholes...	293	51-9031	Cutters and Trimmers, Hand	0.114460
658570	55999	55	4243	Apparel, Piece Goods, and Notions Merchant Who...	4	51-9031	Cutters and Trimmers, Hand	0.038153
658571	55999	55	5613	Employment Services	37194	51-9031	Cutters and Trimmers, Hand	10.415872

658572 rows x 8 columns

Data sources

General information

- The datasets will be published in a github repo with further information
- The datasets are stored in the pickle file format, which can be easily loaded into a python dataframe
- The three data sets can be linked to each other via the FIPS code and the NAICS code
- The metadata is provided with the data sets in a text file