Central European University
Data Engineering 1: SQL and Different Shapes of Data
Fall 2024/25

Term project 2

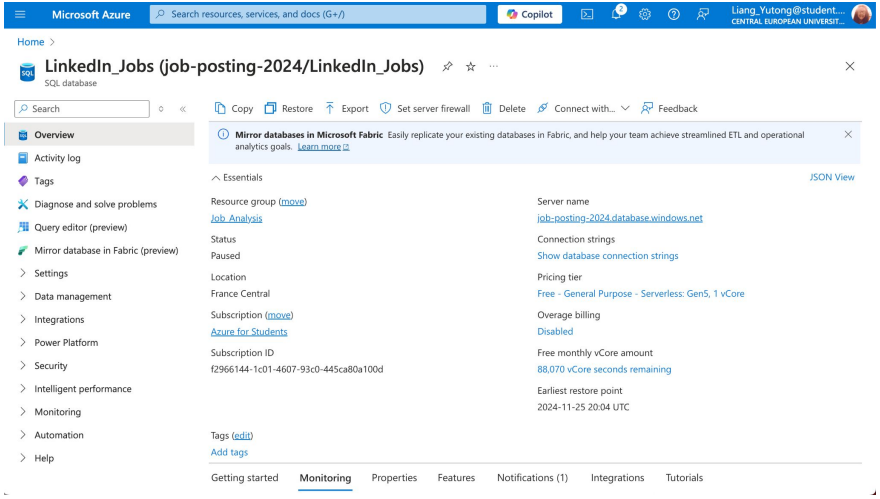
Jobs, Skills, and Salaries:
A Data Engineering Dive into LinkedIn Insights

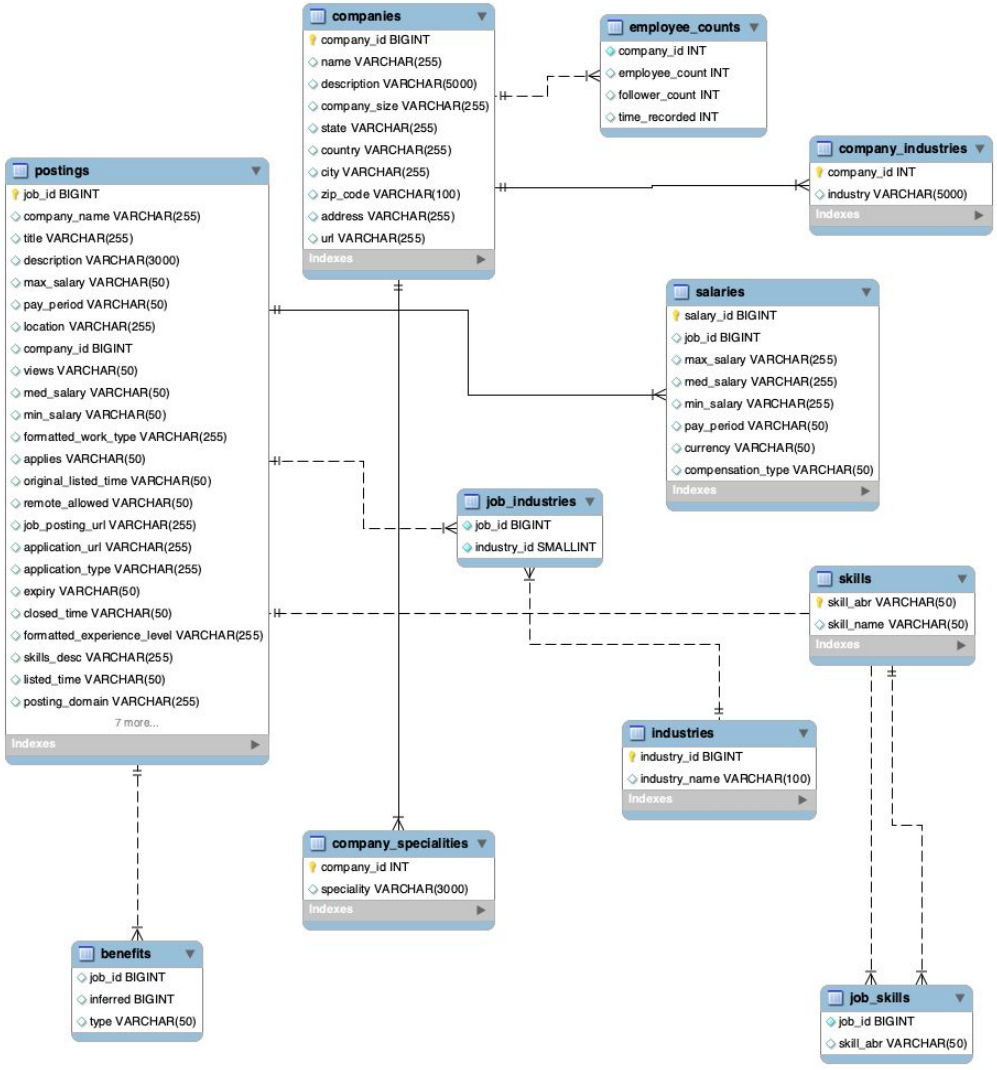
Team #5:
Azizbek Ussenov
Guillermo Leal
Tatyana Yakushina
Yutong Liang

Supervisor: Laszlo Sallo

Azure

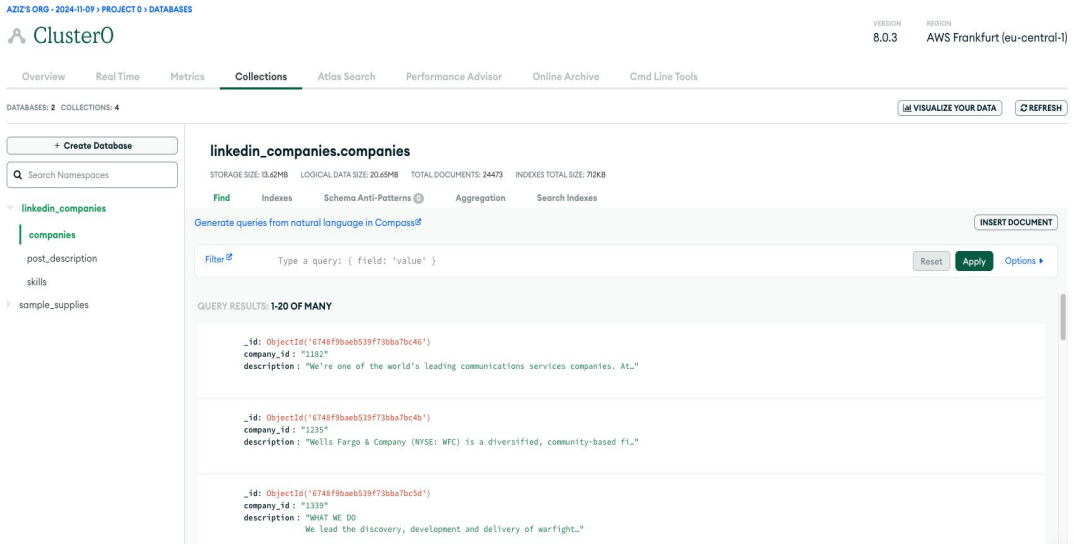
- Azure free SQL Database subscription
- Data imported via Azure Data Studio
- …





MongoDB

- Extracted relevant columns from *companies.csv* and *posting.csv* files
- Converted CSV files into JSON files by chunking into ***N*** chunks to adhere to MongoDB's 16MB document size limit
- Uploaded to MongoDB database by creating collections



Data Sources

- LinkedIn Job Postings, Kaggle
  - Job postings and company data

- API data: U.S. Bureau of Labor Statistics (BLS)
  - BLS data segmented by occupation and industry hourly wages
  - Updates are provided monthly and annually, enabling up-to-date analysis
  - combined BLS wage data with other analyses for clearer economic insights.

For further analysis, both datasets were joined in 2 ways:

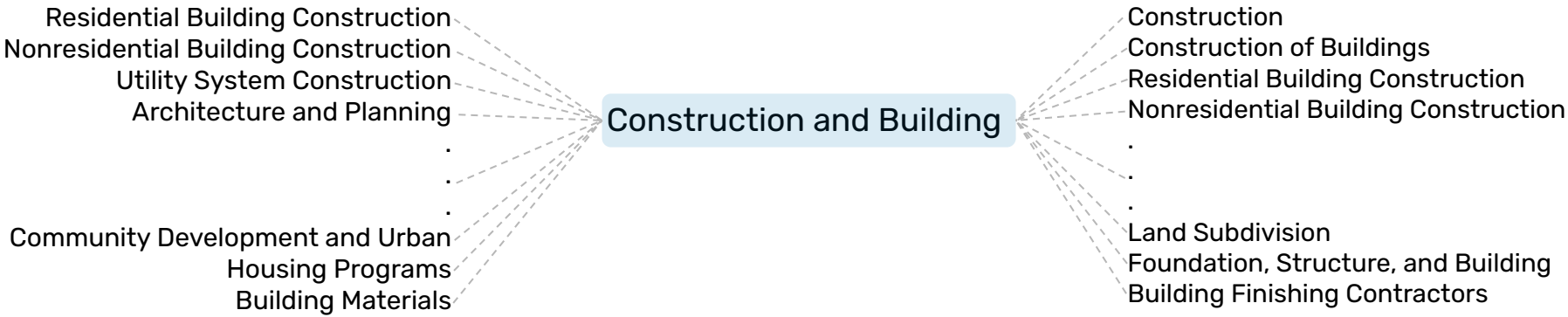
- by U.S. state level: 51 states;
- by industry level: 25 unique industries;

Clustering Linkedin and BLS data by industries

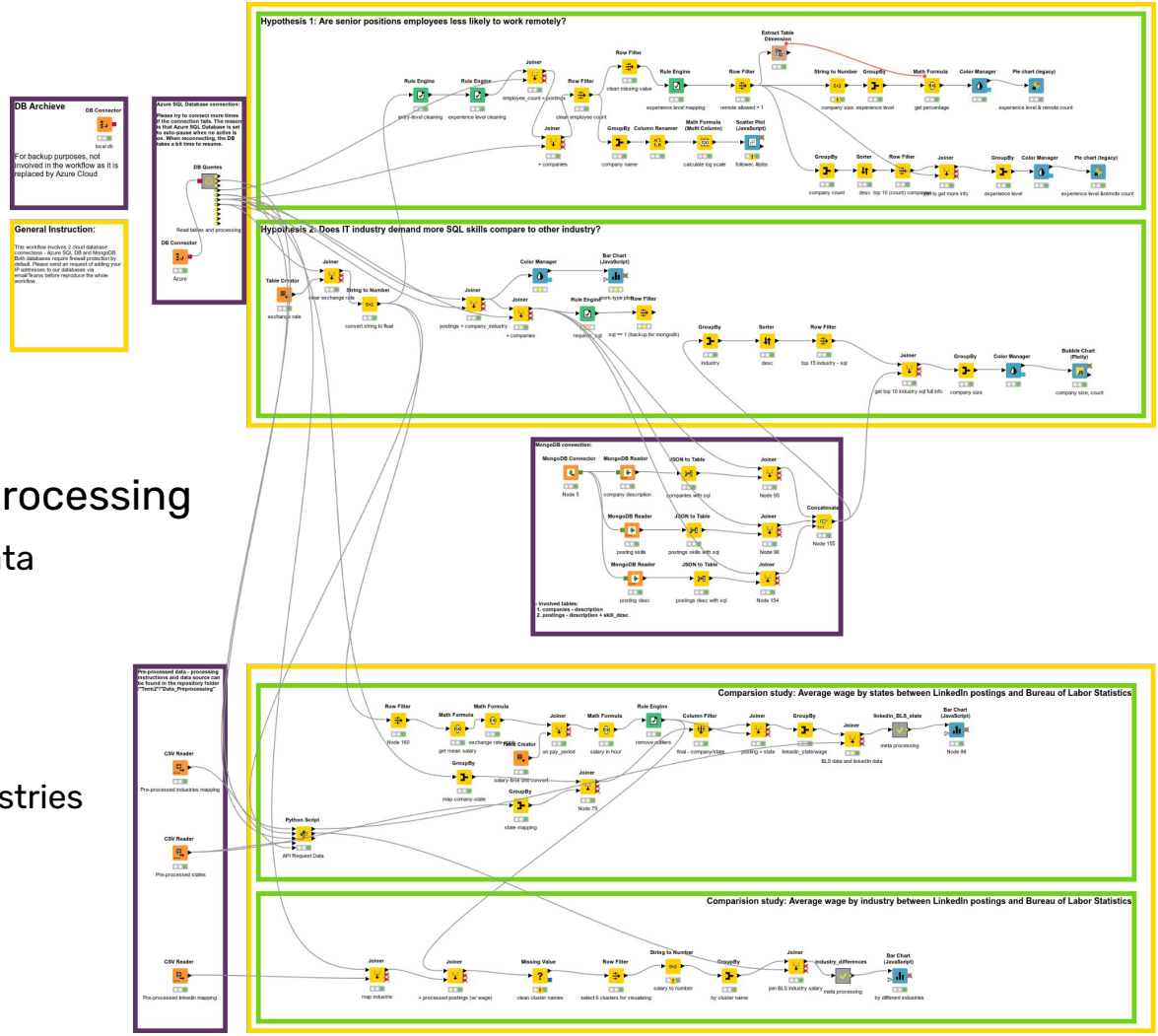
Number of unique industries

LinkedIn data	Clusters	BLS data (API)
388 industries	25 industries	338 industries

Clustering example for one industry



Knime workflow



Color coding of boxes:

purple - data connection & preprocessing

Azure, MongoDB, joining and cleaning data

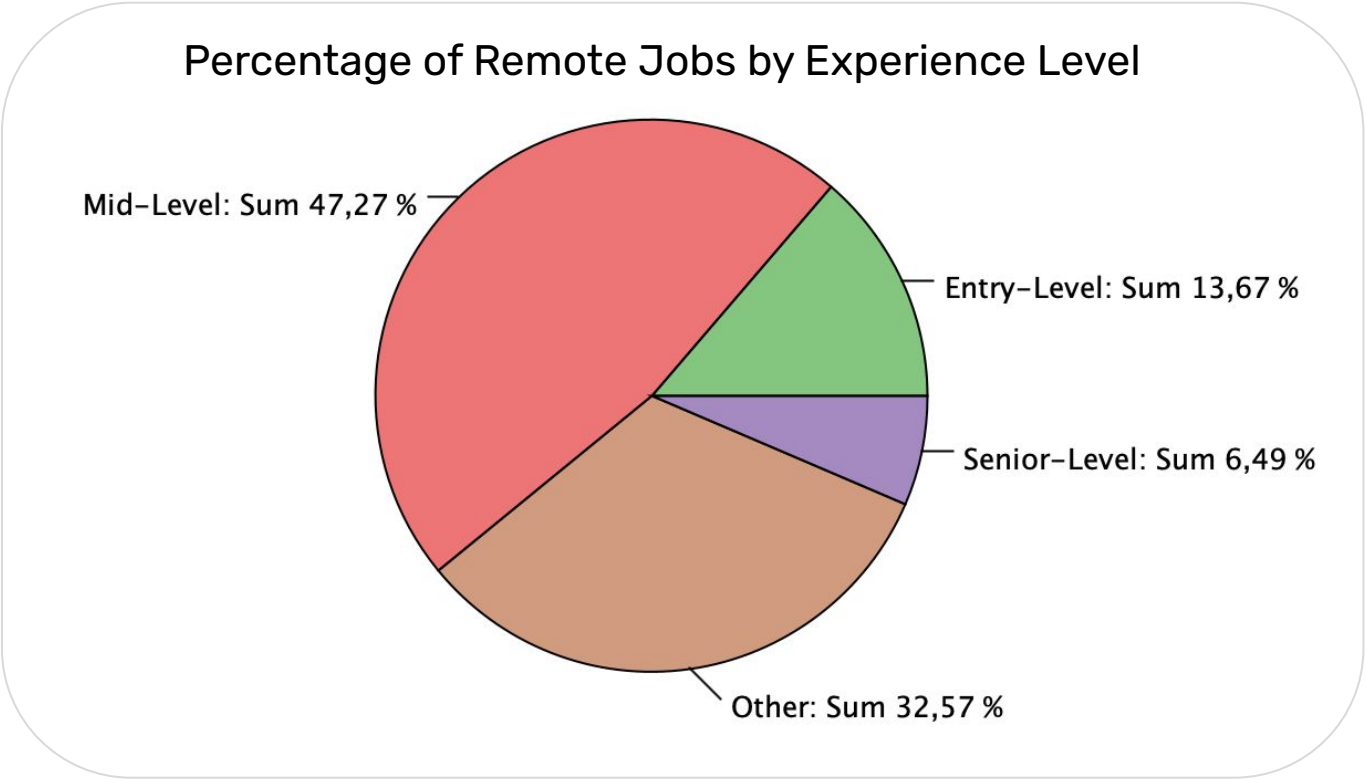
green - data analysis

Hypothesis 1: Job position level

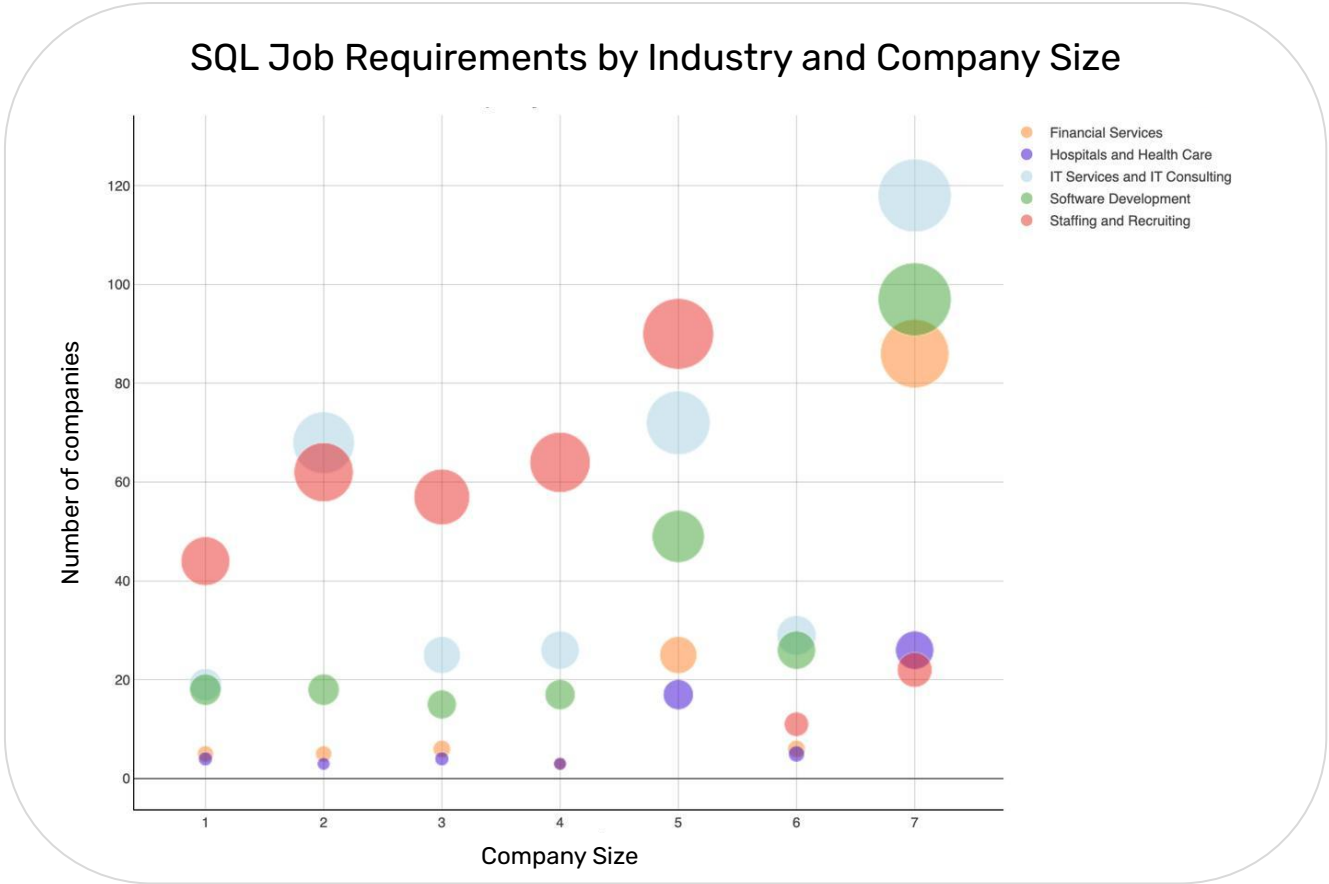
Hypothesis 2: Skill level

Hypothesis 3&4: Salaries by states & industries

Hypothesis 1: Senior roles are less likely to offer remote work options compared to entry- and mid-level positions. (True)

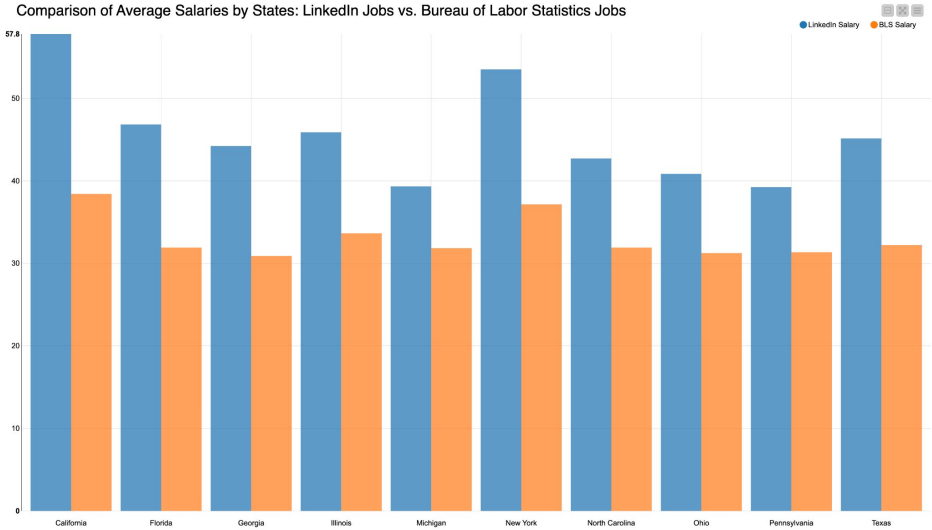


Hypothesis 2: SQL is frequently required in data-centric industries such as finance, healthcare, and e-commerce compared to other sectors. (True)

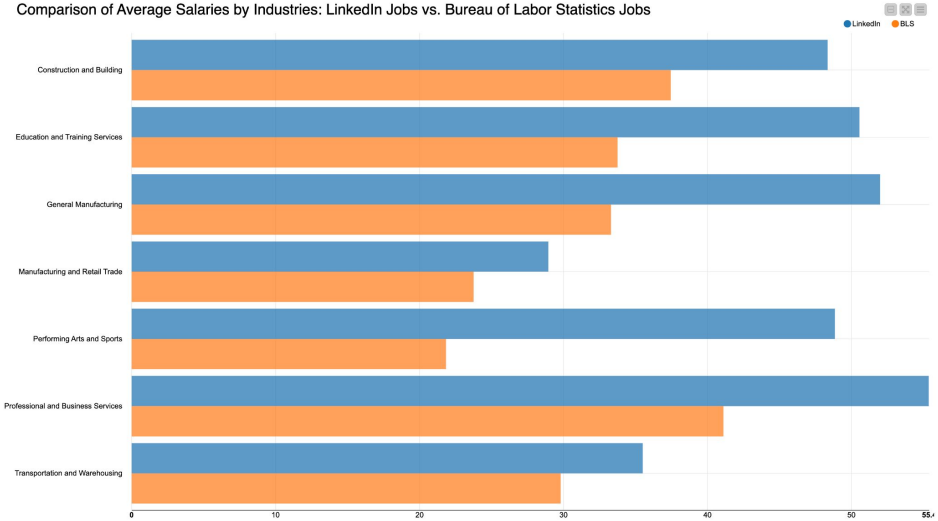


Hypotheses 3 & 4: Average hourly salaries by the U.S. states and industry levels are higher in LinkedIn postings compared to BLS data. (True)

Average hourly salaries by states



Average hourly salaries by industries



LinkedIn Salary BLS Salary

Intuition: Discrepancy suggest that linkedin may reflect real-time market trends more closely or exhibit bias toward higher-paying positions.

Central European University
Data Engineering 1: SQL and Different Shapes of Data
Fall 2024/25

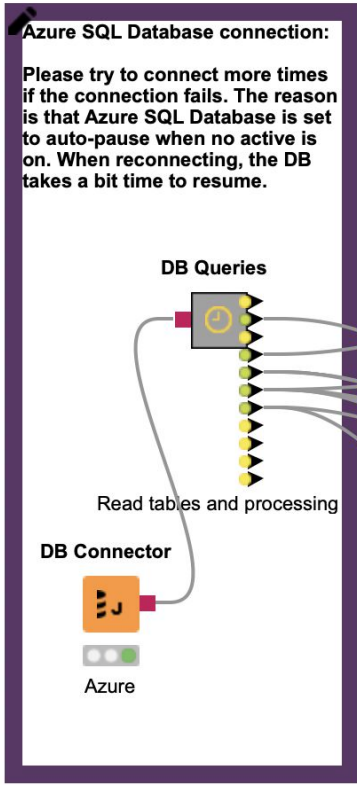
Term project 2

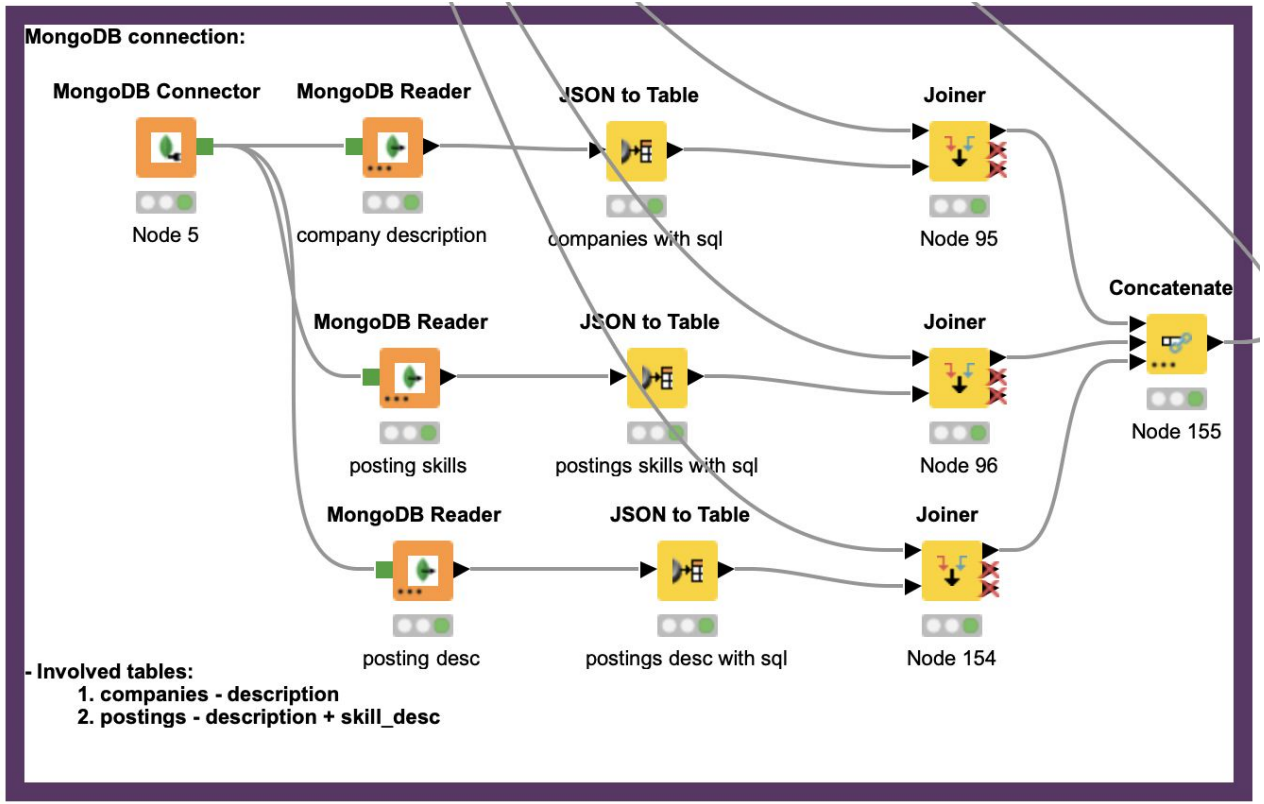
Jobs, Skills, and Trends:
An Analytical Dive into LinkedIn Data

Team:
Azizbek Ussenov
Guillermo Leal
Tatyana Yakushina
Yutong Liang

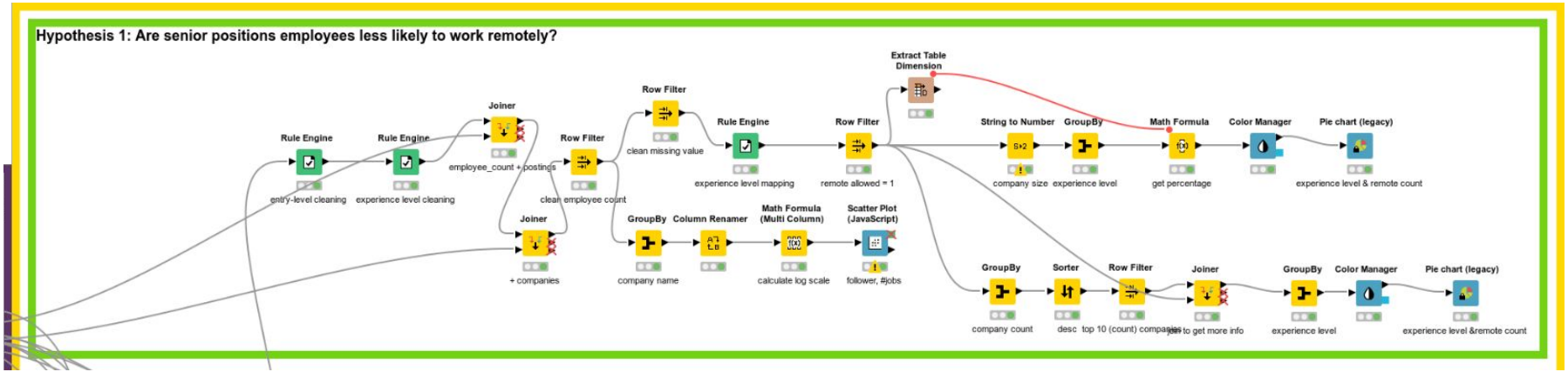
Supervisor: Laszlo Sallo

Appendix. Data preprocessing

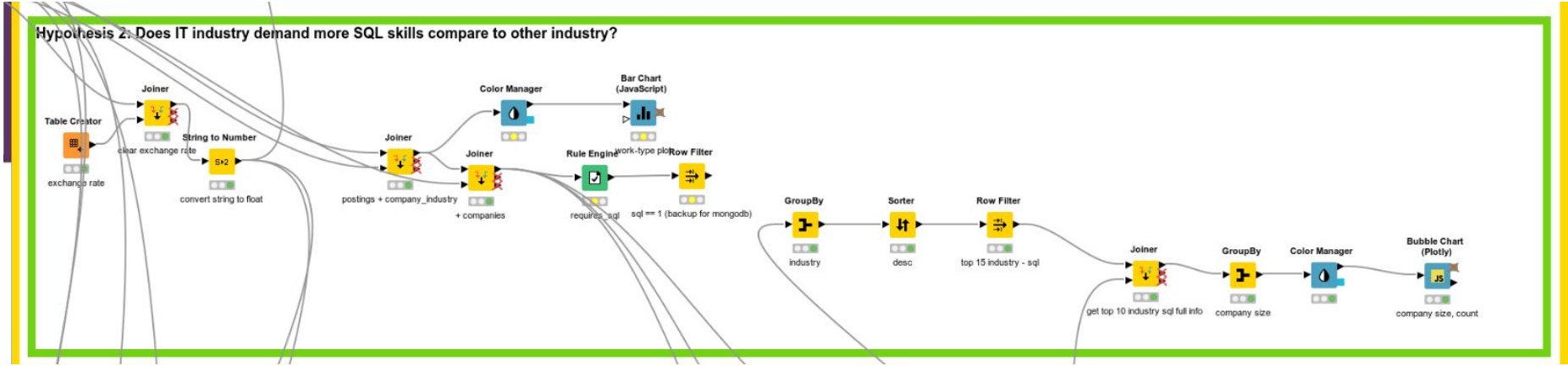




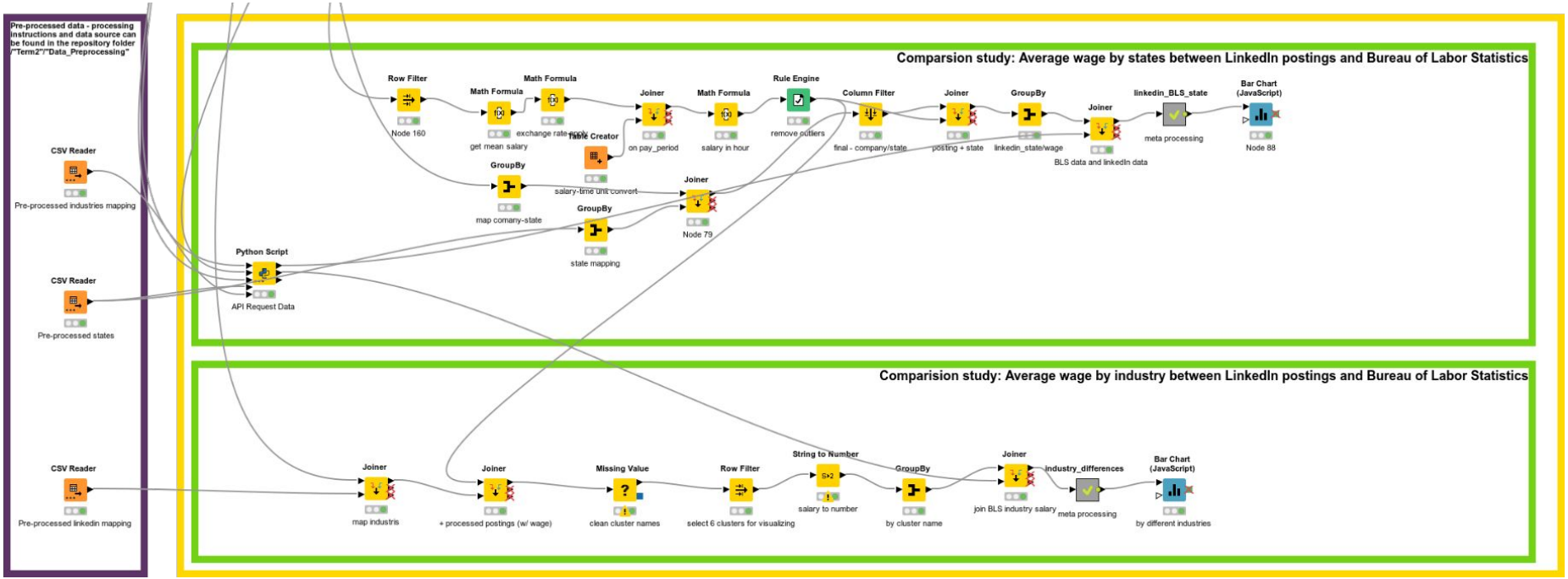
Appendix. Hypothesis 1



Appendix. Hypothesis 2



Appendix. Hypotheses 3 & 4



Template slide

- Text for template slide