

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
- ...

Microsoft Azure

Search resources, services, and docs (G+)

Copilot

Liang_Yutong@student...
CENTRAL EUROPEAN UNIVERSIT...

Home >

LinkedIn_Jobs (job-posting-2024/LinkedIn_Jobs)

SQL database

Search

Overview

Activity log

Tags

Diagnose and solve problems

Query editor (preview)

Mirror database in Fabric (preview)

Settings

Data management

Integrations

Power Platform

Security

Intelligent performance

Monitoring

Automation

Help

Mirror databases in Microsoft Fabric Easily replicate your existing databases in Fabric, and help your team achieve streamlined ETL and operational analytics goals. [Learn more](#)

Essentials

Resource group ([move](#))

[Job Analysis](#)

Status

Paused

Location

France Central

Subscription ([move](#))

[Azure for Students](#)

Subscription ID

I2966144-1c01-4607-93c0-445ca80a100d

Server name

[job-posting-2024.database.windows.net](#)

Connection strings

[Show database connection strings](#)

Pricing tier

Free - General Purpose - Serverless: Gen5, 1 vCore

Overage billing

Disabled

Free monthly vCore amount

88,070 vCore seconds remaining

Earliest restore point

2024-11-25 20:04 UTC

Tags ([edit](#))

[Add tags](#)

Getting started

Monitoring

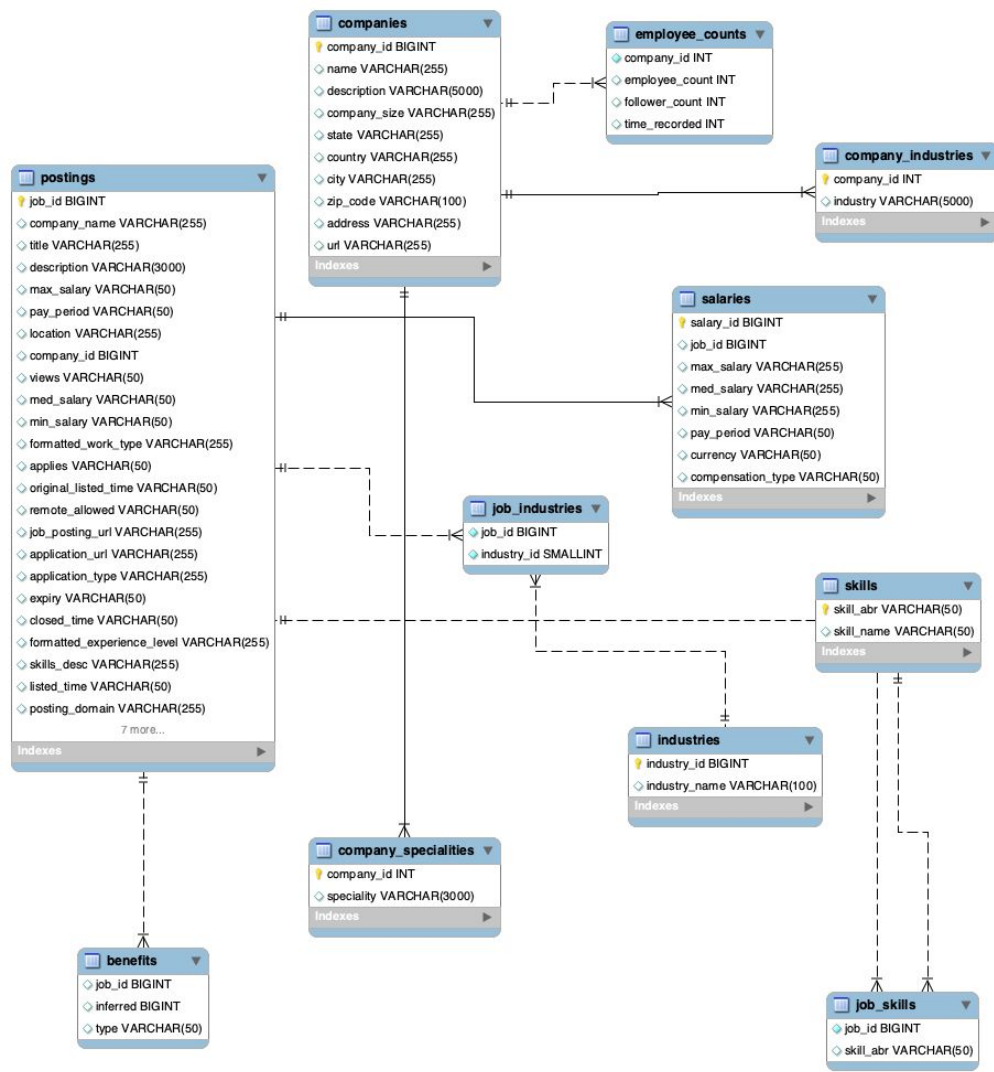
Properties

Features

Notifications (1)

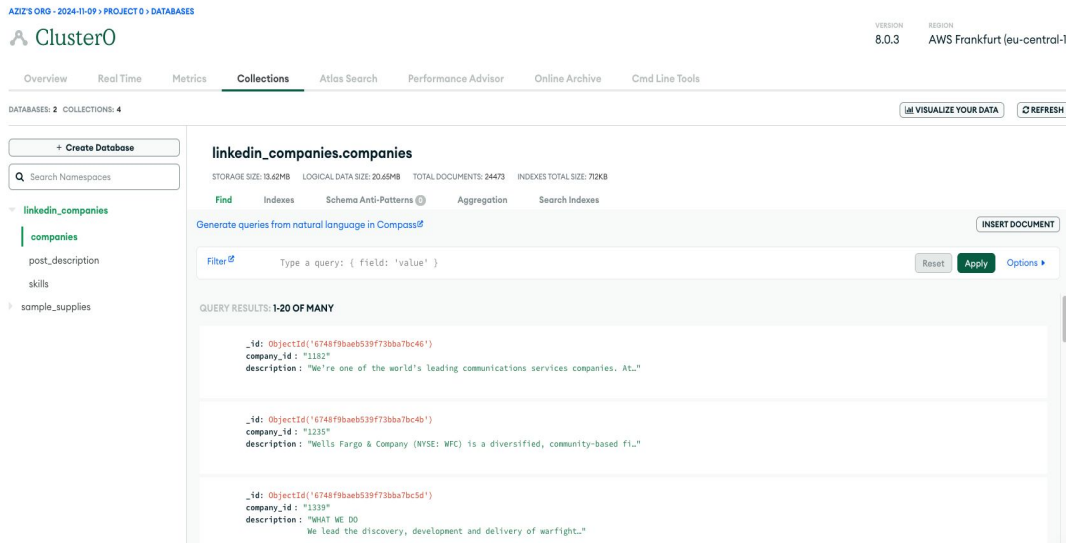
Integrations

Tutorials



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

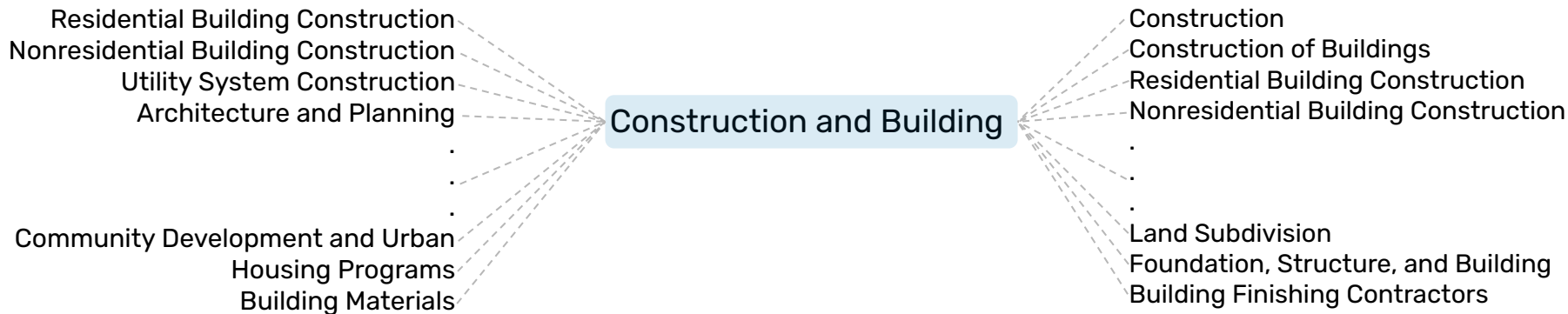
Number of unique industries

LinkedIn data
388 industries

Clusters
25 industries

BLS data (API)
338 industries

Clustering example for one industry



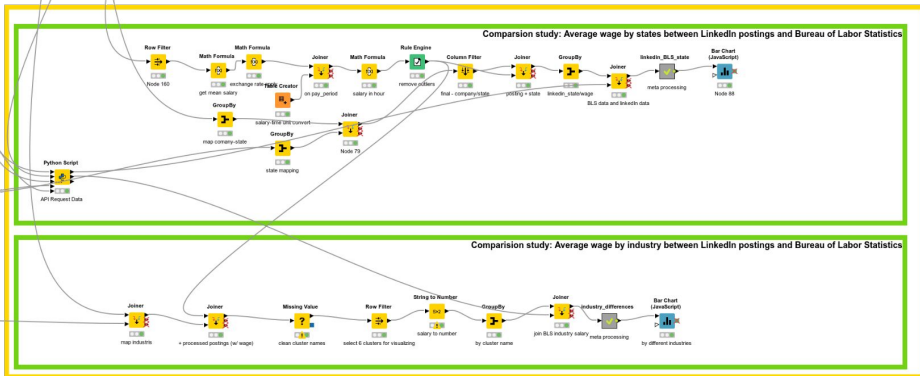
Color coding of boxes:

Azure, MongoDB, joining and cleaning data

Hypothesis 1: Job position level

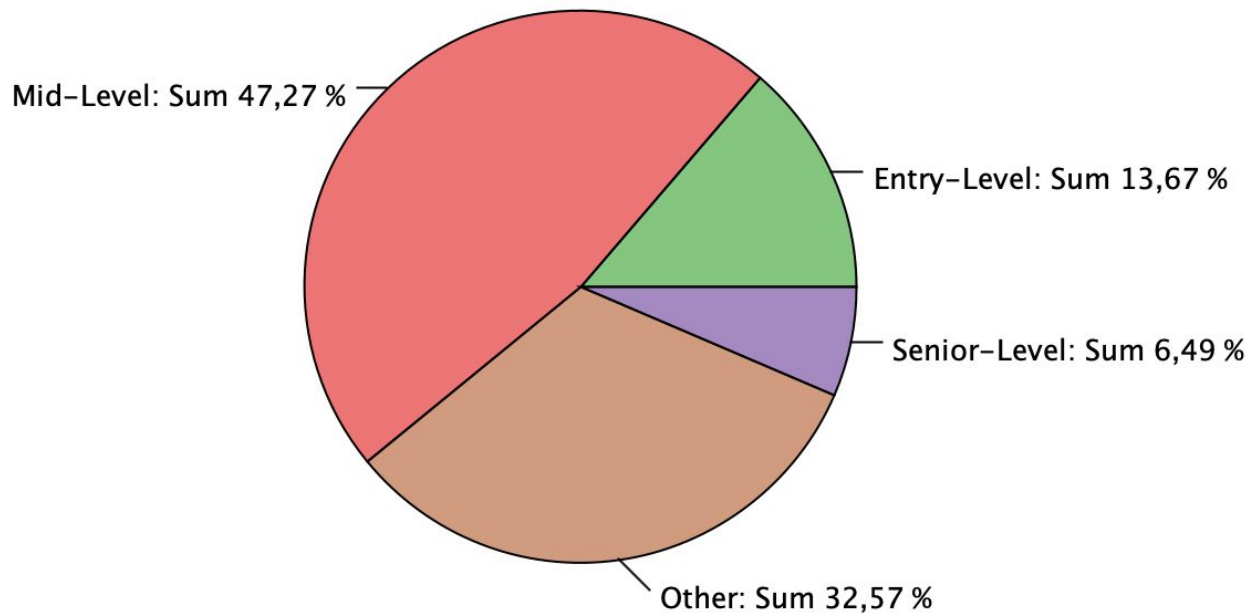
Hypothesis 2: Skill level

Hypothesis 3&4: Salaries by states & industries

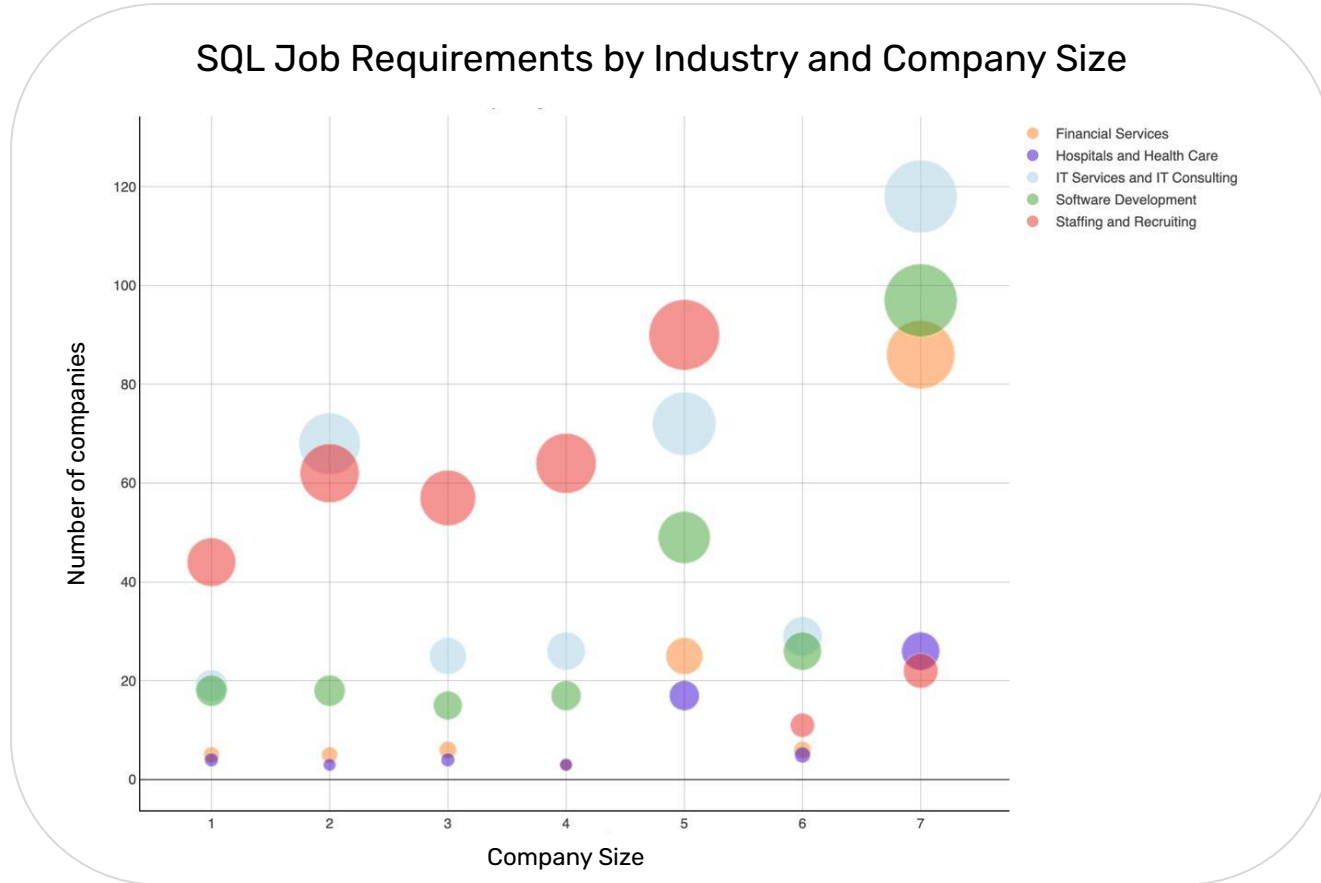


Hypothesis 1: Senior roles are less likely to offer remote work options compared to entry- and mid-level positions. **[True]**

Percentage of Remote Jobs by Experience Level

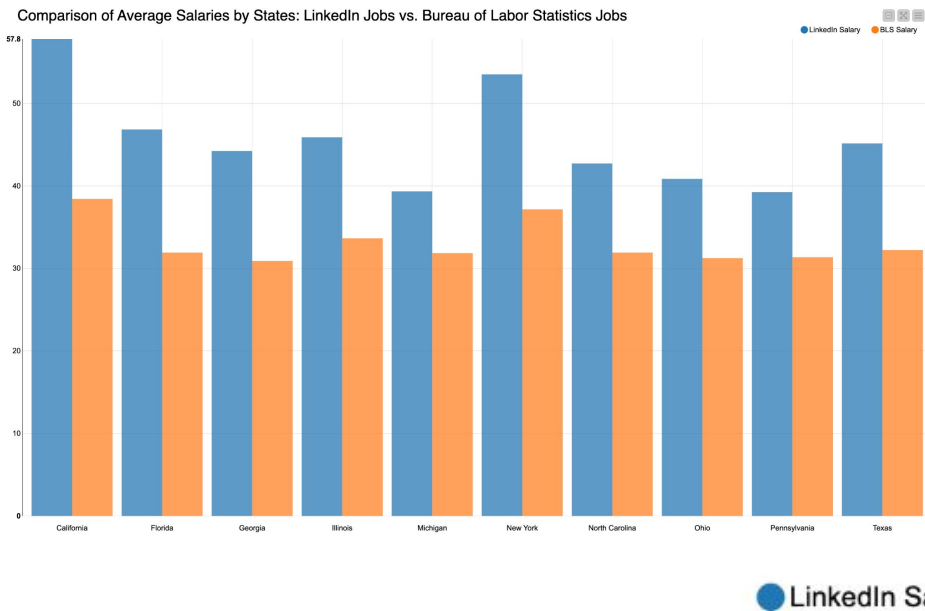


Hypothesis 2: SQL is frequently required in data-centric industries such as finance, healthcare, and e-commerce compared to other sectors. **[True]**

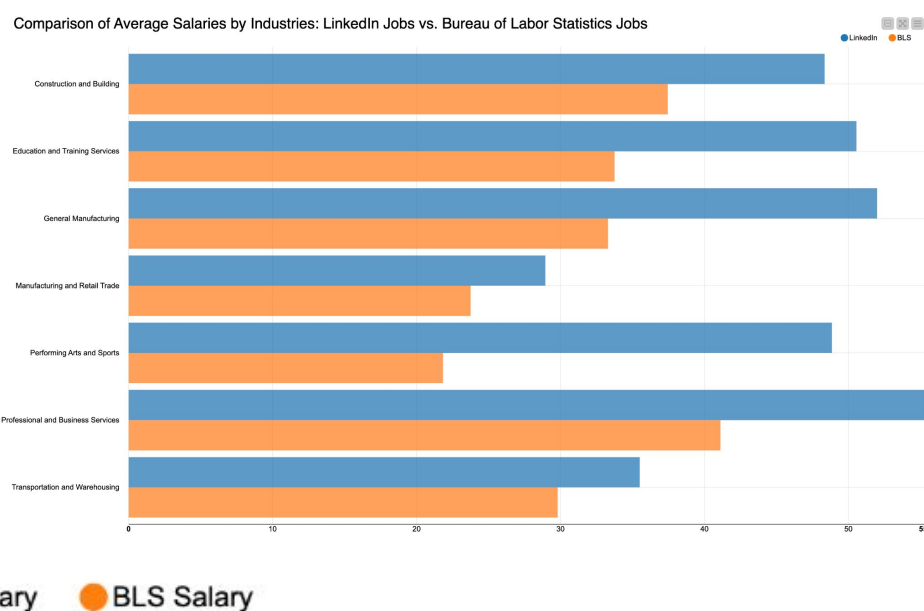


Hypotheses 3 & 4: Average hourly salaries by the U.S. states and industry levels are higher in LinkedIn postings compared to BLS data. (True)

Average hourly salaries by states



Average hourly salaries by industries



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

Term project 2

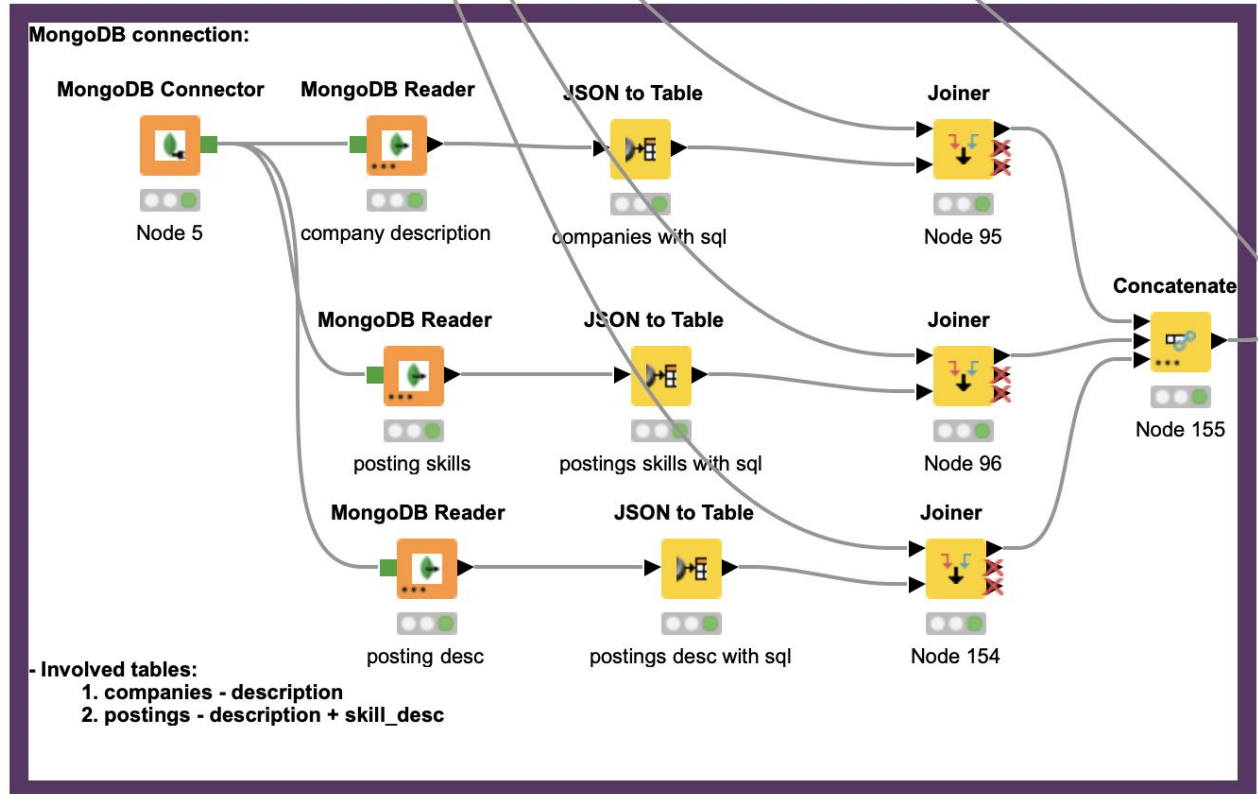
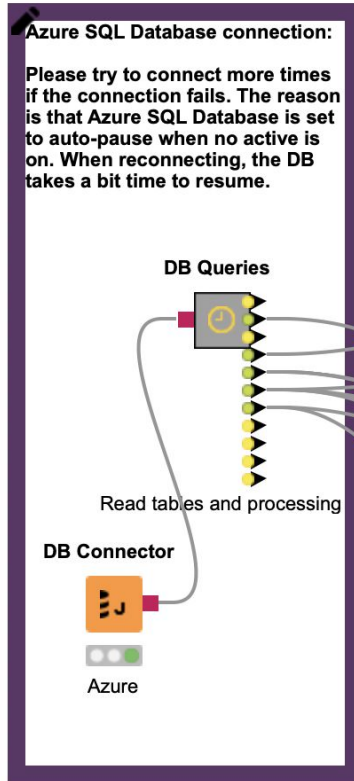
Jobs, Skills, and Trends:
An Analytical Dive into LinkedIn Data

Team:

Azizbek Ussenov
Guillermo Leal
Tatyana Yakushina
Yutong Liang

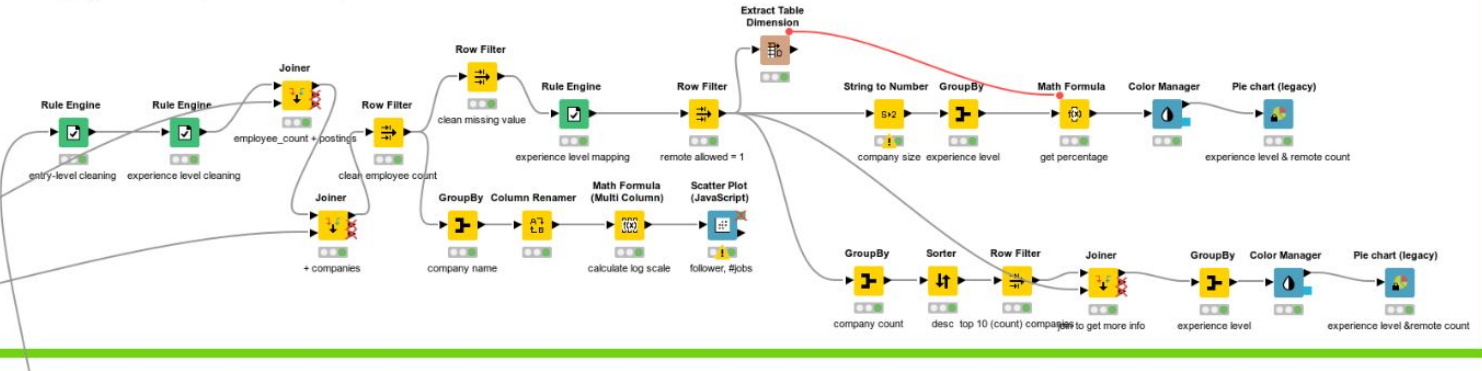
Supervisor: Laszlo Sallo

Appendix. Data preprocessing



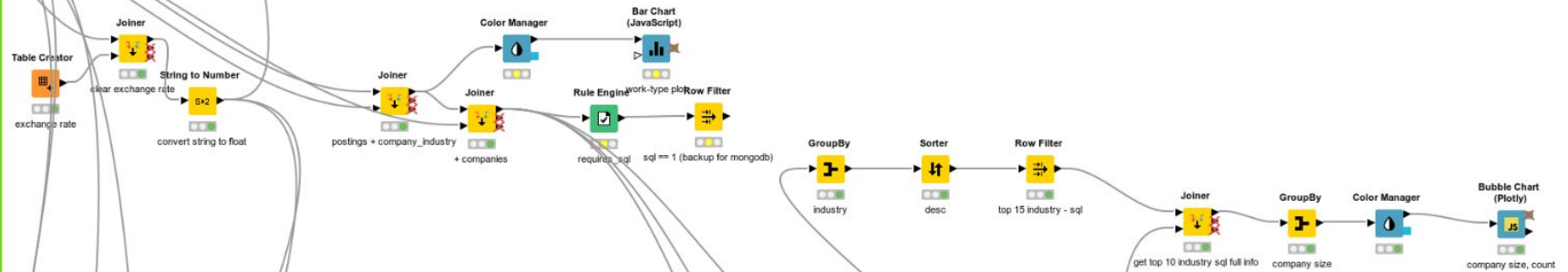
Appendix. Hypothesis 1

Hypothesis 1: Are senior positions employees less likely to work remotely?

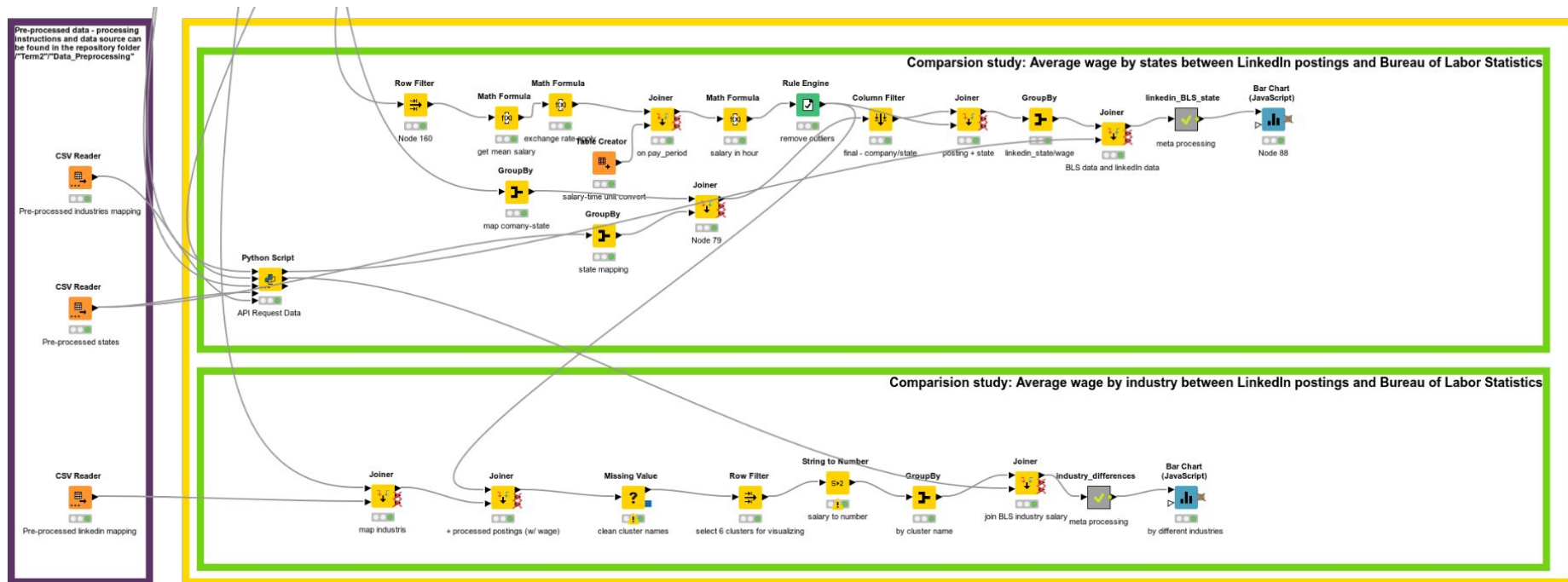


Appendix. Hypothesis 2

Hypothesis 2: Does IT industry demand more SQL skills compare to other industry?



Appendix. Hypotheses 3 & 4



Template slide

- Text for template slide