ETL Workshop I

Presented by:

Juan José Rendón Jaramillo

Teacher:

Javier Alejandro Vergara Zorrila

Universidad Autónoma De Occidente

ELT

08/03/2024

Introduction

In today's technology-driven world, managing and analyzing data efficiently is crucial for businesses to derive meaningful insights. This workshop leverages SQL Alchemy as an Object-Relational Mapping (ORM) tool to interact with a database seamlessly, allowing for efficient data handling, transformation, and analysis. This document outlines the entire process from data loading to exploratory analysis, highlighting key findings and the rationale behind categorizing data for better visualization and understanding.

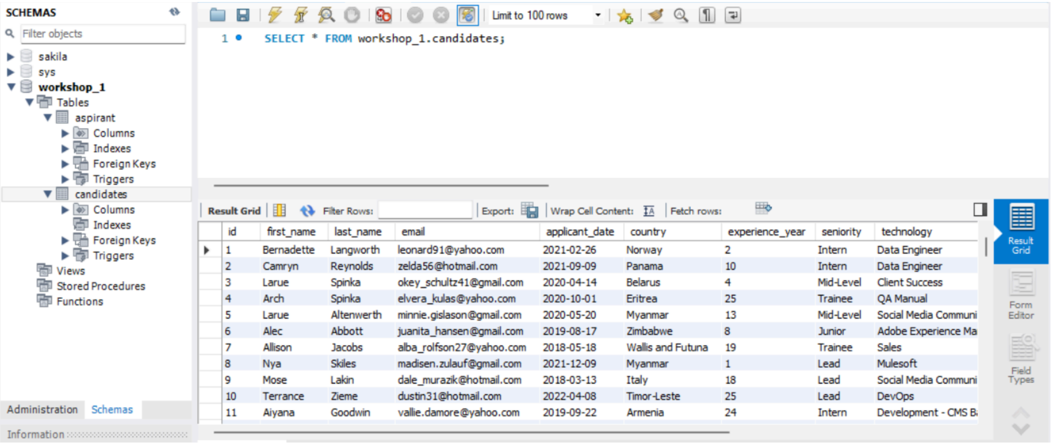
SQLAlchemy: Bridging the Gap with ORM

SQLAlchemy serves as a powerful tool for database operations, providing a high-level ORM to communicate with the database in a more pythonic manner. The ORM layer allows developers to represent database tables as classes and rows as objects, abstracting away the complexities of SQL queries.

Database Creation and Data Loading

The initial step involves defining a database schema that represents the data model using SQL Alchemy's declarative base. Each class in the ORM correlates to a table in the database, with attributes representing table columns. This structure simplifies the process of creating a database and ensures consistency in data handling.

Upon defining the schema, data loading is performed by instantiating objects of the defined classes and committing them to the database. This process translates complex data structures into database-compatible formats, effectively storing the data for subsequent retrieval and analysis.

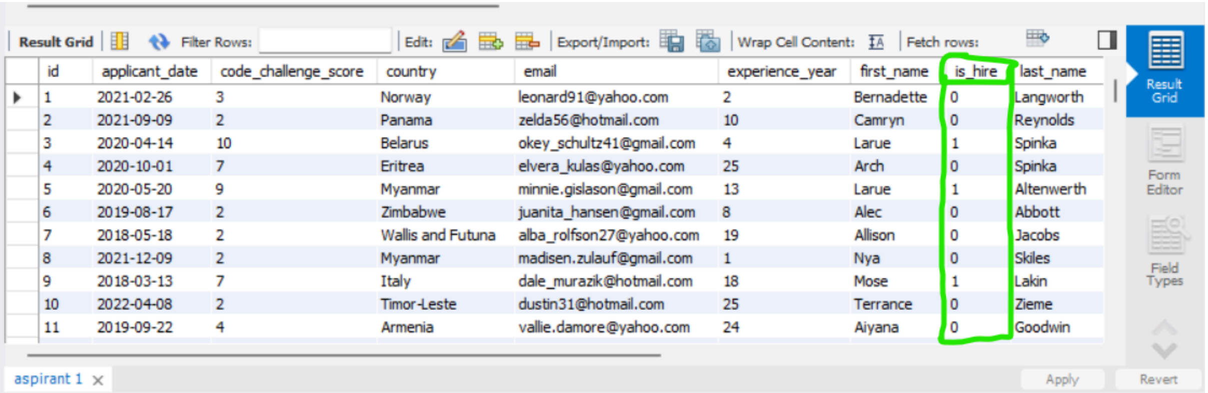


```
SELECT * FROM workshop_1.candidates;
```

id	first_name	last_name	email	applicant_date	country	experience_year	seniority	technology
1	Bernadette	Langworth	leonard91@yahoo.com	2021-02-26	Norway	2	Intern	Data Engineer
2	Camryn	Reynolds	zelda56@hotmail.com	2021-09-09	Panama	10	Intern	Data Engineer
3	Larue	Spinka	okey_schultz41@gmail.com	2020-04-14	Belarus	4	Mid-Level	Client Success
4	Arch	Spinka	elvera_kulas@yahoo.com	2020-10-01	Eritrea	25	Trainee	QA Manual
5	Larue	Altenwerth	minnie.gislason@gmail.com	2020-05-20	Myanmar	13	Mid-Level	Social Media Communi
6	Alec	Abbott	juanita_hansen@gmail.com	2019-08-17	Zimbabwe	8	Junior	Adobe Experience Ma
7	Allison	Jacobs	alba_rolfson27@yahoo.com	2018-05-18	Wallis and Futuna	19	Trainee	Sales
8	Nya	Skiles	madisen.zulauf@gmail.com	2021-12-09	Myanmar	1	Lead	Mulesoft
9	Mose	Lakin	dale_murazik@hotmail.com	2018-03-13	Italy	18	Lead	Social Media Communi
10	Terrance	Zieme	dustin31@hotmail.com	2022-04-08	Timor-Leste	25	Lead	DevOps
11	Aiyana	Goodwin	vallie.damore@yahoo.com	2019-09-22	Armenia	24	Intern	Development - CMS Ba

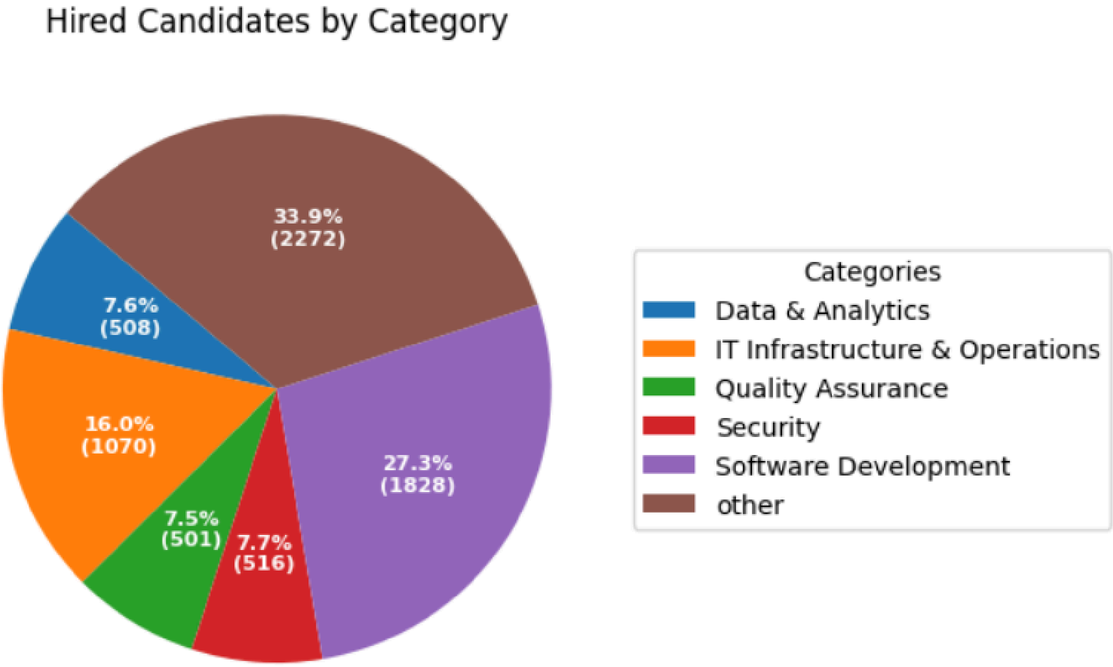
Data Extraction and Exploratory Data Analysis (EDA)

With the data successfully loaded into the database, SQLAlchemy facilitates its extraction through session queries. This step involves retrieving unprocessed data from the database, serving as a foundation for comprehensive exploratory data analysis (EDA).



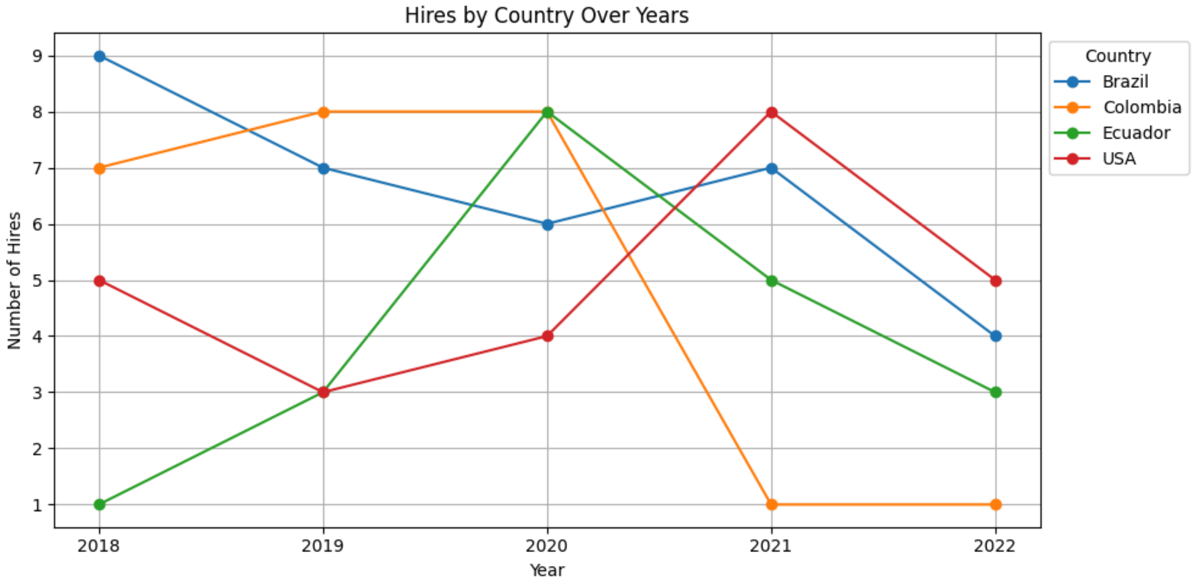
id	applicant_date	code_challenge_score	country	email	experience_year	first_name	is_hire	last_name
1	2021-02-26	3	Norway	leonard91@yahoo.com	2	Bernadette	0	Langworth
2	2021-09-09	2	Panama	zelda56@hotmail.com	10	Camryn	0	Reynolds
3	2020-04-14	10	Belarus	okey_schultz41@gmail.com	4	Larue	1	Spinka
4	2020-10-01	7	Eritrea	elvera_kulas@yahoo.com	25	Arch	0	Spinka
5	2020-05-20	9	Myanmar	minnie.gislason@gmail.com	13	Larue	1	Altenwerth
6	2019-08-17	2	Zimbabwe	juanita_hansen@gmail.com	8	Alec	0	Abbott
7	2018-05-18	2	Wallis and Futuna	alba_rolfson27@yahoo.com	19	Allison	0	Jacobs
8	2021-12-09	2	Myanmar	madisen.zulauf@gmail.com	1	Nya	0	Skiles
9	2018-03-13	7	Italy	dale_murazik@hotmail.com	18	Mose	1	Lakin
10	2022-04-08	2	Timor-Leste	dustin31@hotmail.com	25	Terrance	0	Zieme
11	2019-09-22	4	Armenia	vallie.damore@yahoo.com	24	Aiyana	0	Goodwin

The EDA process is pivotal in understanding the underlying patterns and trends within the data. By employing various statistical and visualization techniques, we can uncover insights that inform decision-making. This workshop focused on categorization for pie chart visualization to enhance the clarity and interpretability of the data, especially given the wide range of roles encompassed in the dataset.



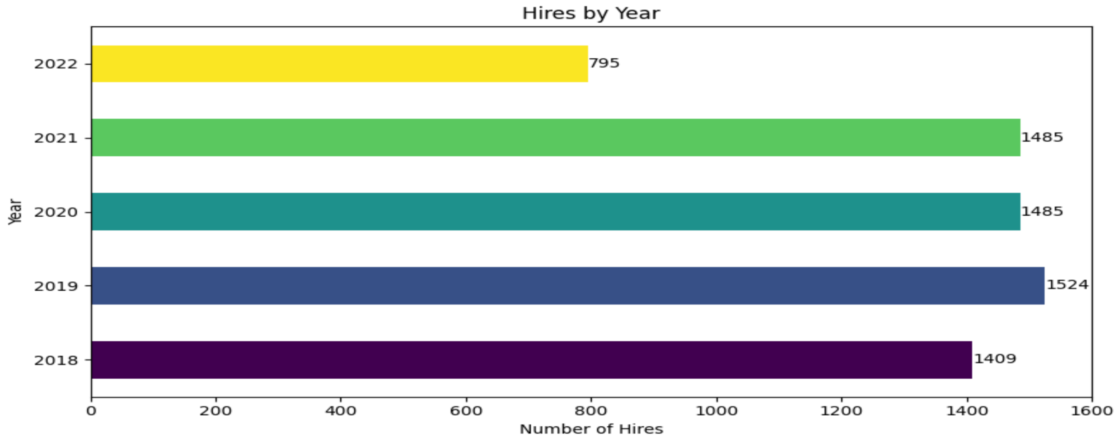
The hiring data reveals a distinct pattern, with the "Other" category leading in terms of the number of hires, followed closely by "Software Development." This observation suggests a diversified demand across various niche roles and specialized positions that don't fall into the more traditionally defined categories. The prominence of the "Other" category could be indicative of evolving industry needs, spotlighting emerging technologies, interdisciplinary roles, or specialized skills that are gaining traction but have not yet been distinctly categorized within the broader tech ecosystem.

Software Development maintaining a strong second position underscores the continuous, foundational demand for development skills in the tech industry. This aligns with the ongoing need for new software, applications, and systems, reflecting the sector's relentless innovation and growth. Software development's critical role in driving technological advancement ensures its position as a cornerstone of tech hiring.

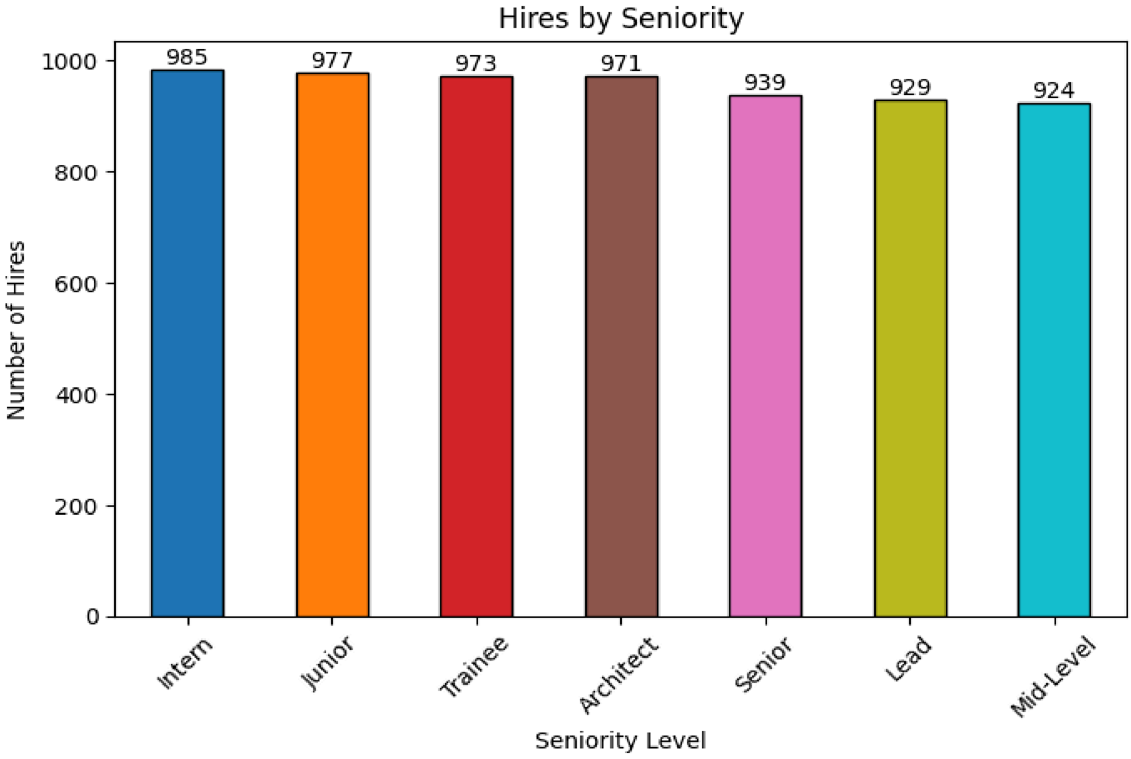


Categorization and analysis by countries have highlighted specific hiring patterns in key markets such as Brazil, Colombia, Ecuador, and the United States, suggesting that regional and economic factors play a significant role in hiring dynamics. This geographic variability underscores the need for tailored and context-aware hiring strategies.

Given the incomplete data for 2022, it's prudent to focus our conclusions on observed trends up to 2021. These trends underline the varied responses of different countries to global technological and economic changes. While some countries demonstrated steady growth or faced downturns after a peak, others showed more fluctuating patterns. These differences highlight the importance of considering local factors, such as economic conditions, policy environments, and industry needs, in understanding and forecasting tech hiring trends. The absence of complete data for 2022 necessitates a cautious approach to drawing conclusions for that year, emphasizing the need for ongoing analysis as more data becomes available.



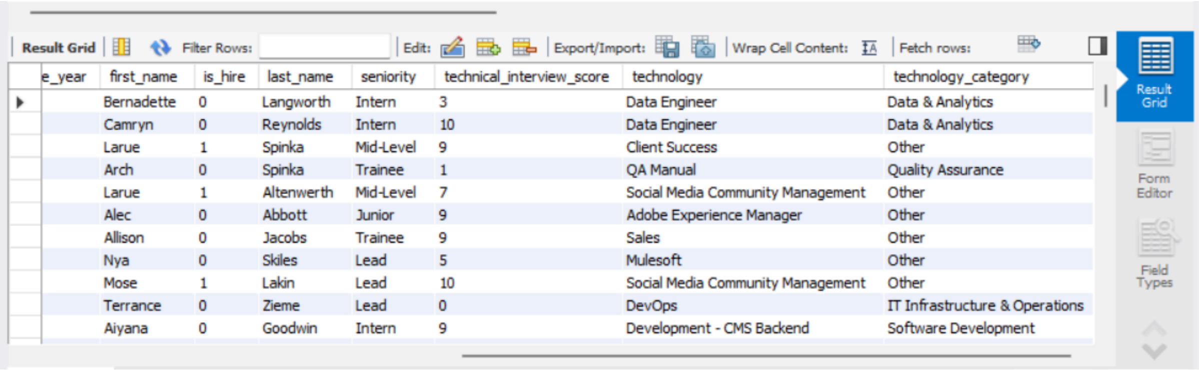
The analysis of hiring trends over the years indicates a dynamic and evolving job market within the technology sector. Starting in 2018, there was a noticeable increase in hires, rising from approximately 1,409 to about 1,524 in 2019, indicating a robust growth phase. This trend, however, stabilizes in 2020 and 2021, with both years observing an equal number of hires at 1,485, suggesting a period of market equilibrium or external factors influencing hiring patterns. The data for 2022, while incomplete, signals a pause in the previous growth trend, potentially due to emerging challenges or changes within the industry.



In terms of seniority levels, the predominance of hires at the intern, junior, trainee, and architect levels indicates a strong inclination towards investing in emerging and developing talent, while senior, lead, and mid-level positions show a lesser proportion of hires, perhaps reflecting a more competitive market or the pursuit of internal talent for promotions also assuming that the greater seniority, the lower the demand will be for it.

Post-EDA Data Handling

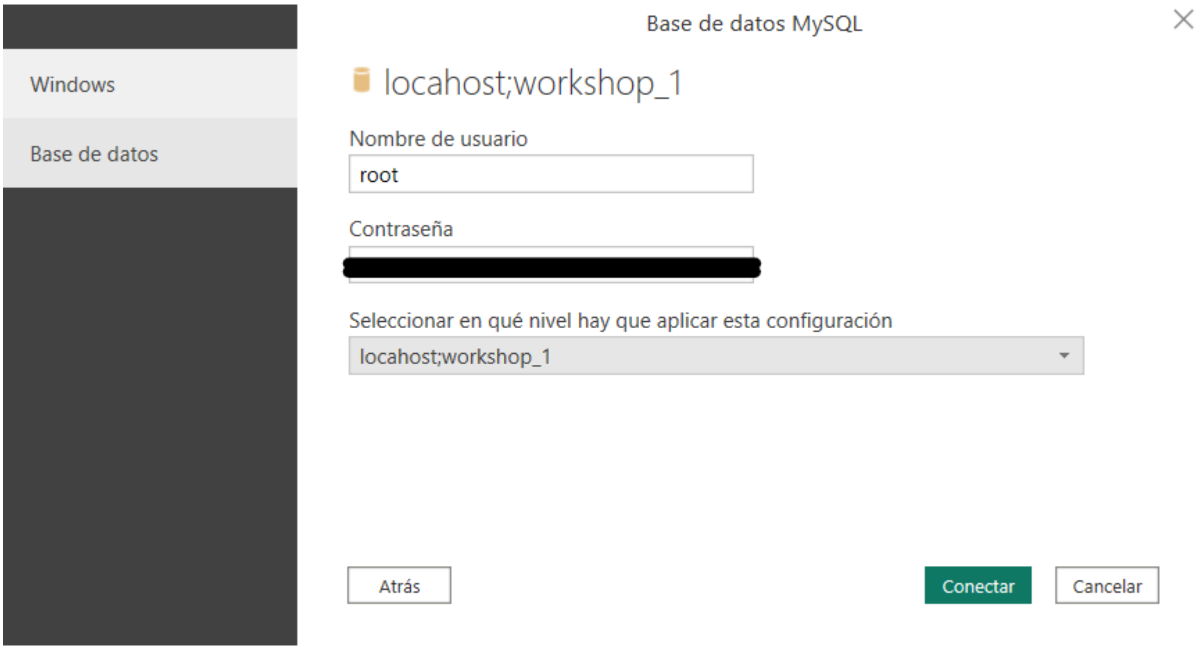
Following the comprehensive EDA, the processed data was stored in a newly created table within the database. This step was essential for organizing the analyzed data in a structured manner, facilitating easier access and visualization.



e_year	first_name	is_hire	last_name	seniority	technical_interview_score	technology	technology_category
	Bernadette	0	Langworth	Intern	3	Data Engineer	Data & Analytics
	Camryn	0	Reynolds	Intern	10	Data Engineer	Data & Analytics
	Larue	1	Spinka	Mid-Level	9	Client Success	Other
	Arch	0	Spinka	Trainee	1	QA Manual	Quality Assurance
	Larue	1	Altenwerth	Mid-Level	7	Social Media Community Management	Other
	Alec	0	Abbott	Junior	9	Adobe Experience Manager	Other
	Allison	0	Jacobs	Trainee	9	Sales	Other
	Nya	0	Skiles	Lead	5	Mulesoft	Other
	Mose	1	Lakin	Lead	10	Social Media Community Management	Other
	Terrance	0	Zieme	Lead	0	DevOps	IT Infrastructure & Operations
	Aiyana	0	Goodwin	Intern	9	Development - CMS Backend	Software Development

Visualization

First, the connection is made from Power BI to the table which stores all the processed information.



Base de datos MySQL

locahost;workshop_1

Nombre de usuario

root

Contraseña

[redacted]

Seleccionar en qué nivel hay que aplicar esta configuración

locahost;workshop_1

The final phase involved creating visualizations to represent the EDA findings effectively. Various charts, including pie charts, bar charts, and multi line charts, were generated to illustrate hires by technology, year, seniority, and country over the years.

These visualizations provide a clear and engaging way to communicate the analysis outcomes and it can be viewed on the dashboard within the repository where the entire project is located.

Key Findings

The exploratory data analysis revealed several key insights:

- The categorization of job roles into broader groups was essential for a coherent visualization in pie charts, preventing clutter and enabling a clear distinction between sectors.

- Hiring trends varied significantly by year and geography, with notable fluctuations in different regions.

- The "Other" category, comprising a mix of roles, indicated a demand for versatile skill sets bridging technical and business domains.

- Software Development remained a prominent category, underscoring the ongoing need for technical expertise in creating and optimizing software solutions.

Conclusion

The workshop demonstrated the effective use of SQL Alchemy for ORM-based data management and the importance of exploratory data analysis (EDA) in uncovering key industry trends. Through the process of categorization and visualization, complex data was rendered accessible, revealing valuable insights into hiring practices and the demand for various skill sets within the technology sector.