Exercise 1

Discuss the implementation of the Air Quality data project.

What went well? What skills do you feel you need to develop further?

In this project, I implemented a complete ETL (Extract, Transform, Load) process to extract data from an external source, transform and normalize it, and finally load it into an SQL database as the final destination.

What went well?

I believe the data transformation and normalization were successfully executed. The final database contains only the necessary columns required for analysis, making it efficient and relevant for the task.

The data is now easy to query, visualize, and modify. The coding is well-structured and optimized throughout the entire pipeline, and all tools used in the process work perfectly.

Overall, I think the exercise was well-executed. I went through 4 iterations of the coding and transformation process before arriving at the final version. Most of the initial issues were related to the API fetching process, which caused several errors early on the project.

Challenges Faced

This particular technique wasn't heavily practiced in class, and it took some time to identify the correct API URL. Although it was listed on the website's homepage, I initially struggled to locate it and ended up researching it through the browser's developer tools. This experience taught me a lot, and I now feel confident in applying API data fetching and transformation techniques in future projects.

Another significant challenge was normalizing the geometry column. Across all attempts at completing the exercise, the most difficult issue to overcome—particularly toward the end—was identifying the correct encoding needed to read the DataFrame properly. Without the correct encoding, the MinMaxScaler could not be applied successfully. Resolving this encoding issue was key to completing the data normalization process effectively.

Exercise 1

Future Development

For future development, the dataset could be expanded by including additional meaningful columns, such as pollution levels. This would enable a more comprehensive analysis of pollutant concentrations across the geolocated points.

Another potential enhancement would be the creation of new features specifically designed for machine learning applications. By engineering relevant features, the dataset could support predictive modeling or classification tasks, opening the door to more advanced data-driven insights and automated decision-making processes.

Exercise 2

AO2.3

Discuss how you ensured your code is well optimised.

What optimisations does PySpark incorporate?

To ensure my code is well optimized, I took several steps throughout the development process:

1. **Data Handling Efficiency**
   I focused on minimizing unnecessary data movement and redundancy by filtering out irrelevant columns, keeping only the essential dataset needed for transformation. This approach reduced the processing load and enhanced the overall efficiency of the ETL pipeline.

2. **Data Normalization**
   I ensured the data was properly normalized and included in the DataFrame, maintaining consistency for the subsequent steps of the exercise. By fitting the normalized data into the DataFrame, I ensured all values were scaled appropriately. This is crucial for ensuring uniformity and accuracy in the analysis, and it helped facilitate smoother transformations and modeling in the later stages, ensuring the data was in a suitable format for further processing.

Exercise 2

AO2.3

3. **Testing**
   I conducted thorough testing at each stage of the process to ensure that the data was fitted correctly and the model was processed and clean in all stages. This testing helped me verify the integrity and performance of the code throughout the pipeline.

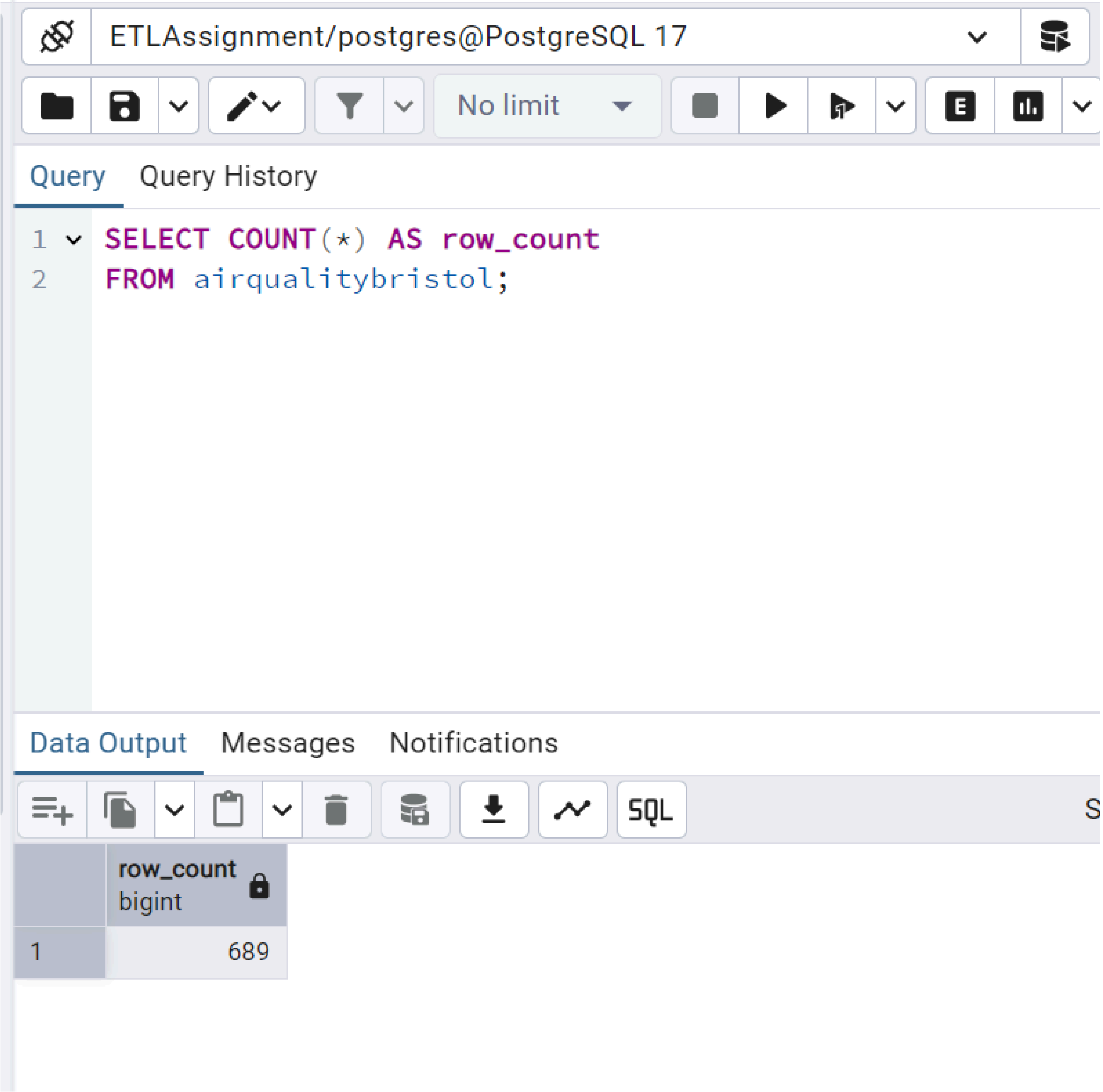
Exercise 3

Create at least two SQL queries for your database to explore the data you transformed to demonstrate the Loading process worked correctly.

Include screenshots of each query and their results

In the first picture I am counting the row number in the dataset.

Exercise 3

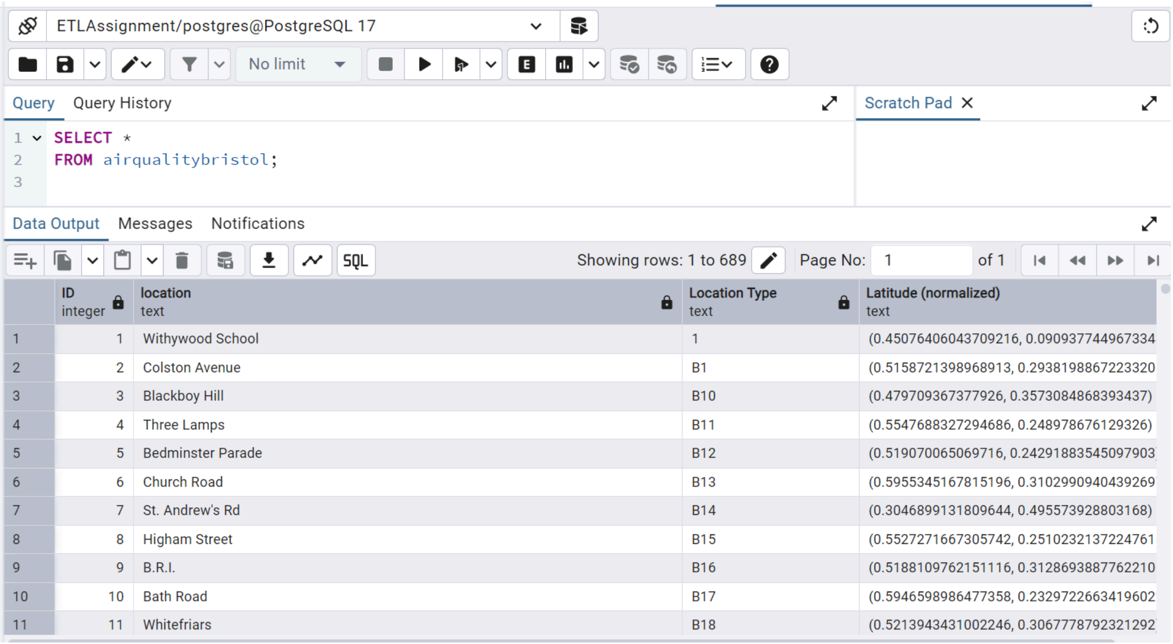


```
SELECT COUNT(*) AS row_count
FROM airqualitybristol;
```

	row_count bigint
1	689

In the second picture you can see the normalized values in the Latitude column on the right.

Exercise 3



```
SELECT *
FROM airqualitybristol;
```

Showing rows: 1 to 689

	ID integer	location text	Location Type text	Latitude (normalized) text
1	1	Withywood School	1	(0.45076406043709216, 0.090937744967334
2	2	Colston Avenue	B1	(0.5158721398968913, 0.2938198867223320
3	3	Blackboy Hill	B10	(0.479709367377926, 0.3573084868393437)
4	4	Three Lamps	B11	(0.5547688327294686, 0.248978676129326)
5	5	Bedminster Parade	B12	(0.519070065069716, 0.24291883545097903
6	6	Church Road	B13	(0.5955345167815196, 0.3102990940439269
7	7	St. Andrew's Rd	B14	(0.3046899131809644, 0.495573928803168)
8	8	Higham Street	B15	(0.5527271667305742, 0.2510232137224761
9	9	B.R.I.	B16	(0.5188109762151116, 0.3128693887762210
10	10	Bath Road	B17	(0.5946598986477358, 0.2329722663419602
11	11	Whitefriars	B18	(0.5213943431002246, 0.3067778792321292

Exercise 4

AO5.1

As a Data Engineer, you will have to keep up to date with new and emerging technologies in the industry. Explain how horizon scanning and the delphi method can be employed to identify and evaluate upcoming technologies.

What technologies and tools do you expect to rise in prominence in the next five years? Why?

An **Horizon Scanning** is a process used to systematically explore and identify emerging trends or opportunities that could affect the future. This involves researching and scanning any environment such as: web, books, articles, to find signals of change. Horizon scan helps organizations to anticipate changes.

The **Delphi Method** is a structured technique used to gather information on a particular topic that involves questioning if opinions are correct to allow participants to provide feedback usually in anonymity to reduce any influence and allow a more honest feedback.

Exercise 4 AO5.1

Below you can find the result of my research, referencing the opinions and articles of the source of the information.

With the breakthroughs of Big Data in the early 2000s (Bornmann and Haunschild, 2020), and the progression into distributed computing with improved hardware, there was a significant shift toward what we call today as a fast and ever-evolving application of data science and machine learning.

Some of the latest applications, techniques, and tools involve the use of predictive analytics, recommendation systems, natural language processing, image and video analysis, personalized marketing, supply chain optimization and energy management.

The incorporation of Generative AI into industry has been furthermore an advancement in every sector, especially in medicine. “Using Generative AI even helps identify chronic conditions” (Analytics Insight, 2024). This is not only revolutionizing the healthcare system but every industry, making it essential for businesses today to align with data scientists for their decision-making (PSMGT, 2025).

Reference list:

Bornmann, L. and Haunschild, R., 2020. The Evolution of Big Data Research: A Bibliometric Analysis Based on Studies Published in Sociology Journals. Scientometrics, [online] 124(2), pp.595–610. Available at: https://link.springer.com/article/10.1007/s11192-020-03371-2 (Accessed 25 March 2025).

Analytics Insight (2024) *Generative AI in healthcare: Personalizing treatment for better outcomes*. Available at: https://www.analyticsinsight.net/generative-ai/generative-ai-in-healthcare-personalizing-treatment-for-better-outcomes (Accessed: 25 March 2025).

Exercise 5

AO3.3

Analyse the differences between data warehouses and data lakes.

Consider scenarios where each would be more appropriate

The Differences Between Data Warehouses and Data Lakes

The differences between **Data Warehouses** and **Data Lakes** are essential for businesses that rely on robust data storage, management, and analysis. Both serve important roles but are designed for different use cases. Below is a breakdown:

A **Data Warehouse** is primarily designed to store **structured data**. This data is typically organized into tables and schemas. Here are some key characteristics of a data warehouse:

- **Structured Data**: Data in a warehouse is well-organized, typically in relational databases, and is ready for analysis and reporting.
- **Historical Data**: Data warehouses are ideal for storing **historical data**, allowing businesses to access data over long periods of time. This enables tracking of trends and analysis over extended periods.
- **Optimized for Queries**: Data warehouses are optimized for fast querying and reporting. This makes them crucial for **Business Intelligence (BI)** environments, where data is queried for insights and decision-making.
- **BI Tool Integration**: A data warehouse integrates seamlessly with BI tools, making it easy to generate reports and dashboards that support business analysis and decision-making.

Data Lake: Scalable Storage for Raw, Unstructured Data

A **Data Lake**, on the other hand, is designed to store vast amounts of **raw data**, including both **unstructured** and **semi-structured** data. Here are the key features of a data lake:

- **Data in Native Format**: Data is stored in its **raw form**, with no preprocessing or structuring required. This includes data in unstructured formats such as **text**, **video**, and **audio**.
- **Real-Time Machine Learning**: Since data is stored in its raw format, it can be processed in real-time, enabling **real-time exploration** and uncovering

new patterns or insights for machine learning applications.

- **Scalable for Large Data Volumes**: Data lakes are designed to handle massive amounts of data—often in the **petabyte** range. This scalability makes them ideal for storing vast datasets from various sources.