ADVANCED DATA PROCESSING TECHNIQUES

PROJECT REPORT

AIRBNB DATA ANALYSIS USING DATABRICKS

KAVIN M

RA2412033010050

PROJECT OVERVIEW:

This project primarily focuses on analysing Airbnb listings in **London** using **Databricks**. The goal is to explore trends in room types, pricing, availability, and neighbourhood distributions. Leveraged Apache **Spark** in **Databricks Community Edition** to process large-scale data efficiently and extract valuable insights for better decision-making.

AIRBNB:

Airbnb is an online marketplace that connects travellers with hosts offering accommodations. Founded in 2008, it has transformed the hospitality industry by providing **flexible, cost-effective, and unique lodging options**. The platform includes listings ranging from shared rooms to luxury villas, with millions of hosts and guests worldwide.

DATABRICKS:

Databricks is a unified **cloud-based data analytics platform** built on **Apache Spark**. It enables **big data processing, machine learning, and real-time analytics** with an optimized runtime for faster computation.

In this project, we use **Databricks Community Edition**, which offers a free environment for Spark-based analytics.

DATASET OVERVIEW:

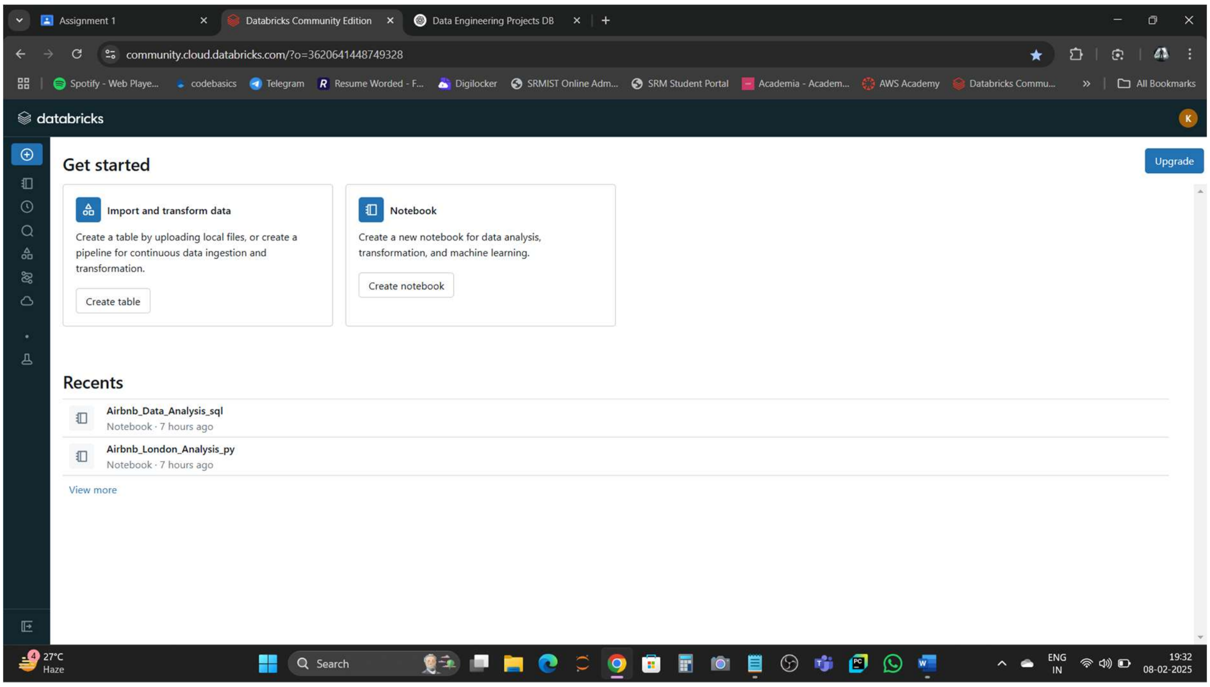
The name of the Dataset is **listings.csv,** which is a csv file contains the information about the accommodating properties and availabilities for the city called **"London, UK".**

This dataset contains several columns:

- **Id**: Unique listing id.
- **Name**: Name of the listing.
- **Neighbourhood**: Location of the listing.
- **Room_type**: Type of room (e.g., Entire home, Private room, Shared room).
- **Price**: Price per night.
- **minimum_nights**: Minimum stay required
- **number_of_reviews**: Total number of reviews.

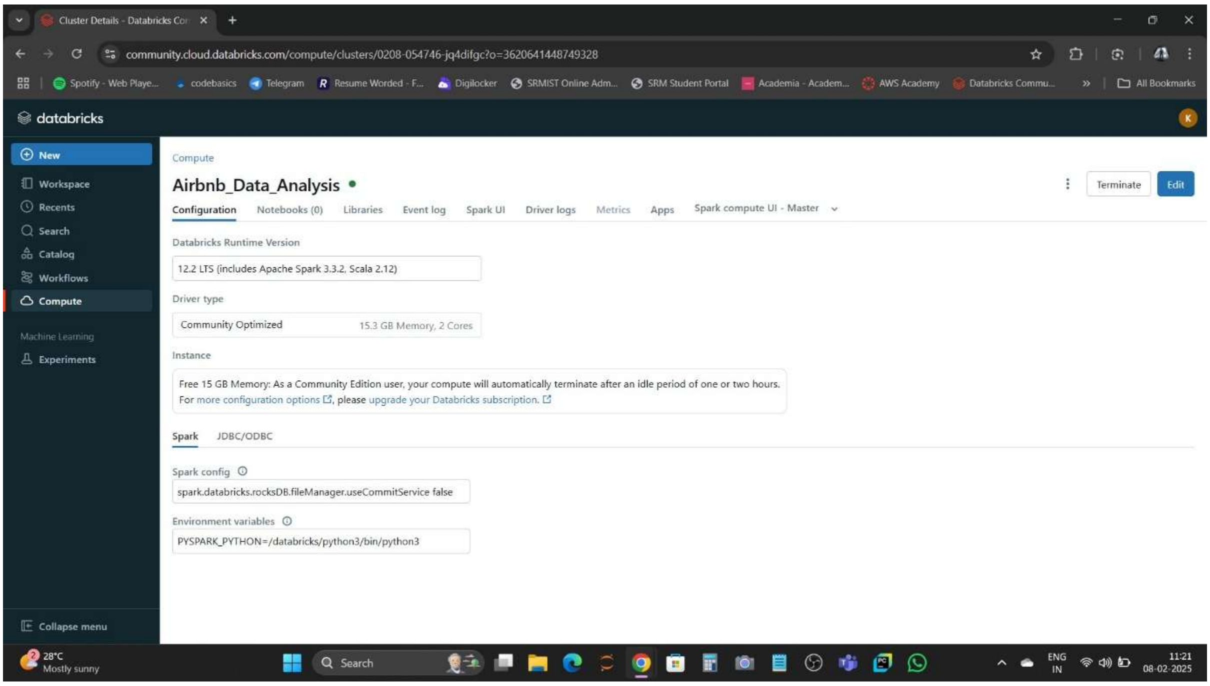
PROJECT SETUP:

1. **Setting Up Databricks Environment:**
    - Go to Databricks Community Edition
    - Sign up and log in.
    - Create a new workspace.



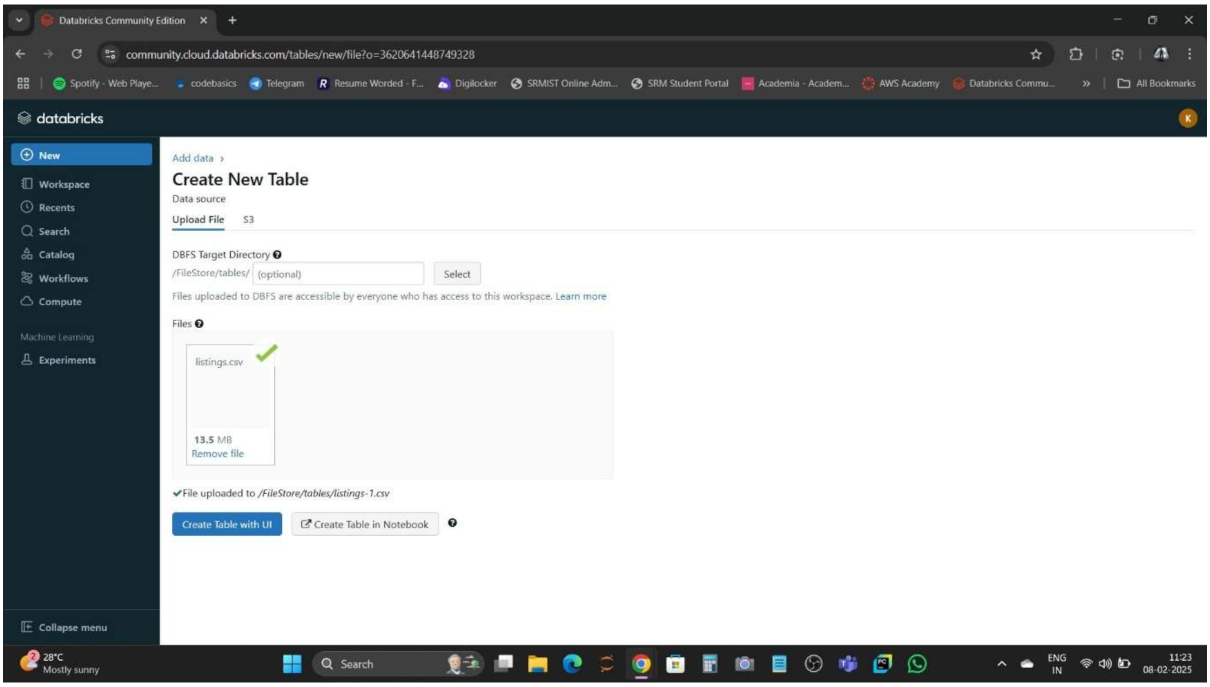
2. Create a New Cluster:

- Click on Compute → Create Cluster.
- Name the cluster (e.g., Airbnb-Analysis).
- Select Single Node Cluster (default).
- Set Runtime Version: Databricks Runtime 12.2 LTS (Scala 2.12, Spark 3.3.2).
- Click Create Cluster and wait for it to start.



3. Download the dataset from the Airbnb site and Upload on Databricks:

- Go to Data → Create Table → Upload listings.csv.
- Select DBFS (Databricks File System) for storage.
- Note the path where the file is stored (e.g., /FileStore/tables/listings.csv).

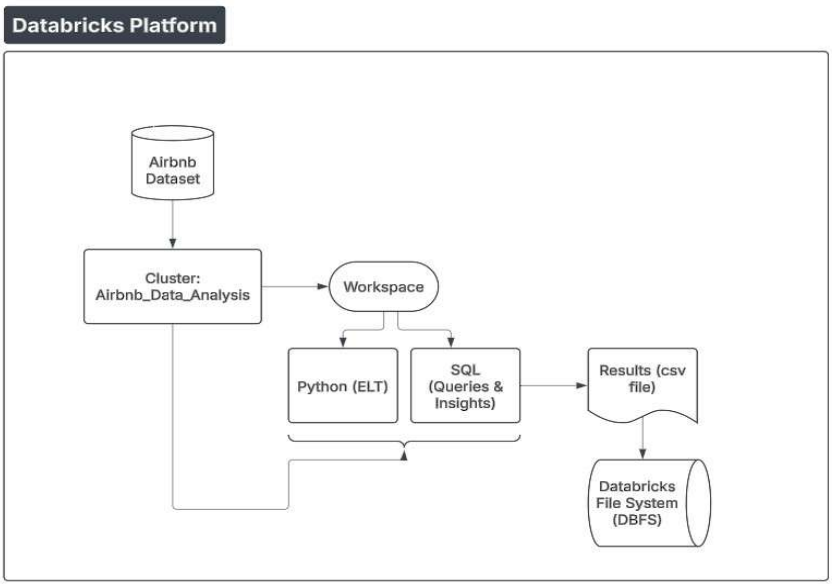


PROJECT DIMENSIONS & WORKFLOW:

Python and SQL are the primary languages used for this project for various purposes.

- Python – For loading, preprocessing and visualizing the data.
- SQL – For generating insights from the data using queries.

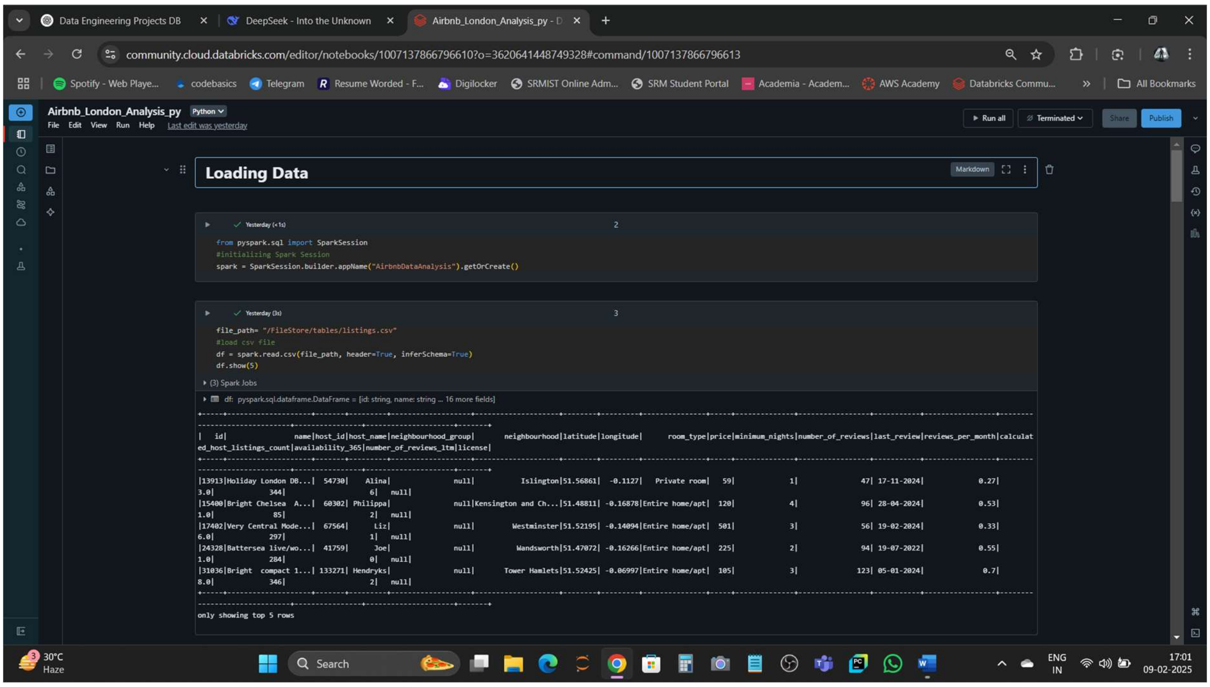
Below flowchart describes the entire project workflow:



IMPLEMENTATION:

Python Notebook:

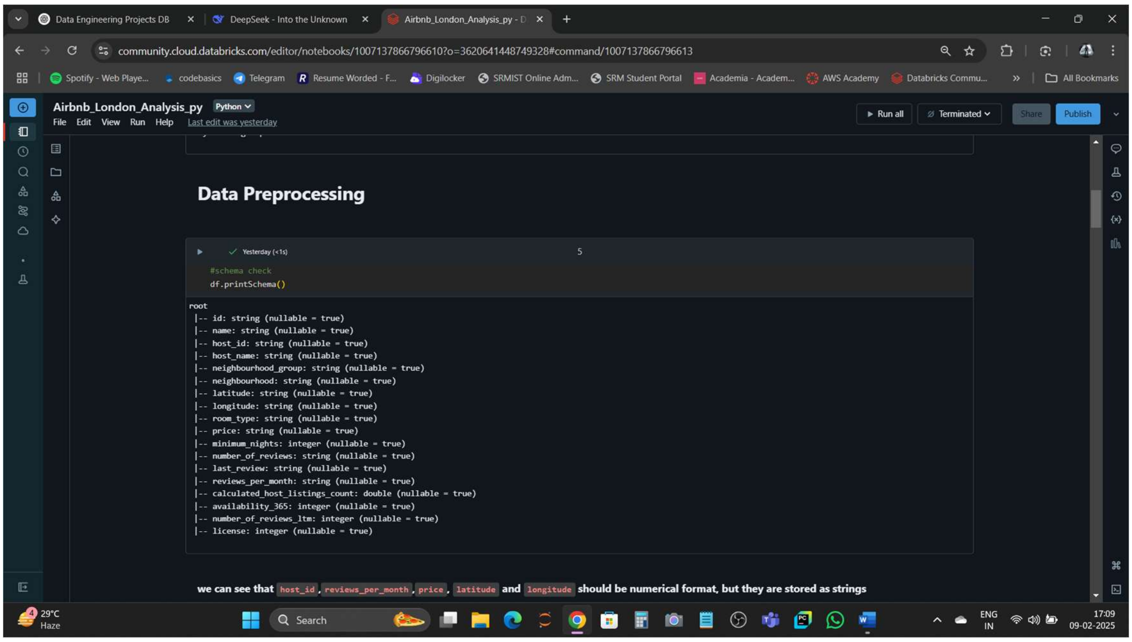
1. Loading data



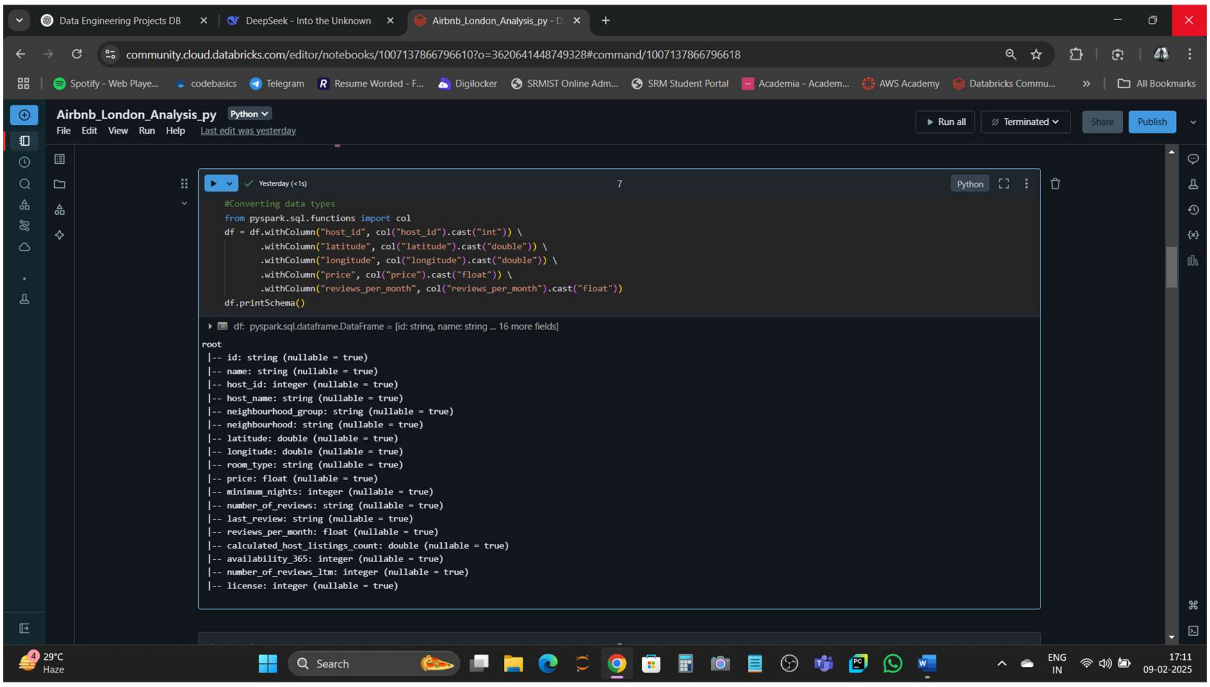
The data has been loaded from the DBFS and the data will initialize the **SparkSession** which allows users to interact with spark functionality, create Dataframes and perform operations.

2. Data Preprocessing

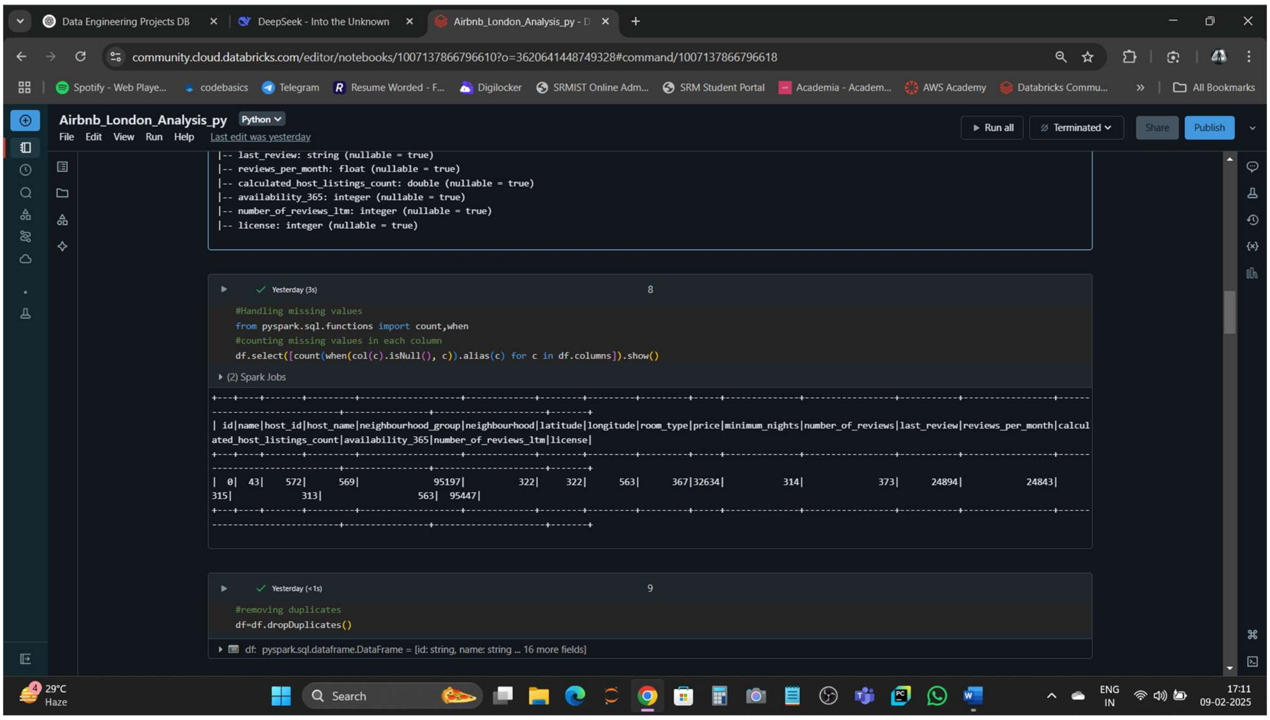
This process involves refining the clumsy data by converting the attributes into their appropriate data types, handling missing values, removing duplicates, evaluating outliers, etc.



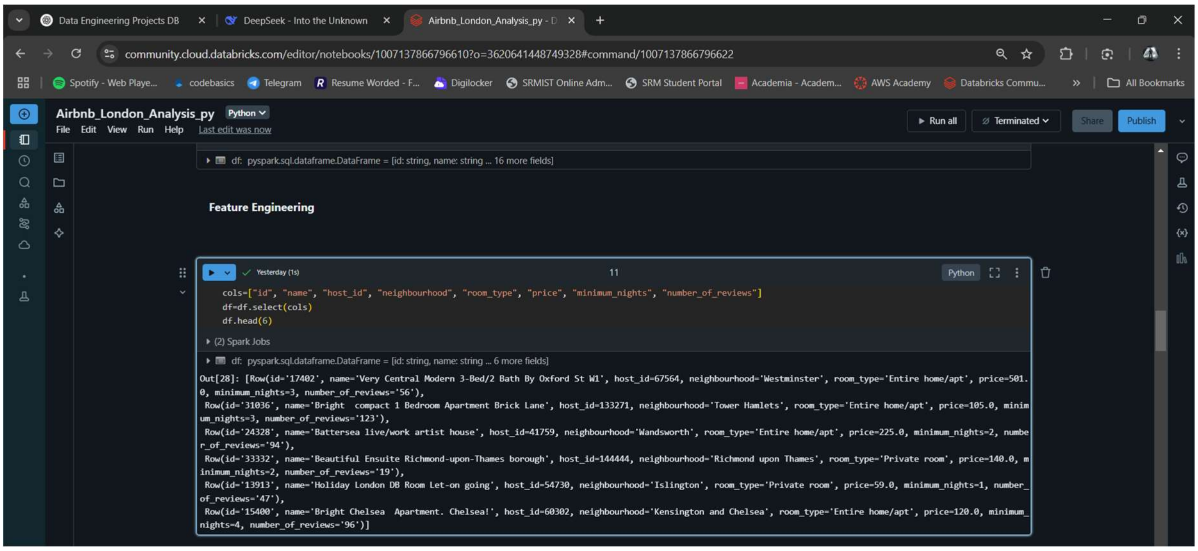
we can see that **host_id,reviews_per_month,price, latitude** and **longitude** should be numerical format, but they are stored as strings.



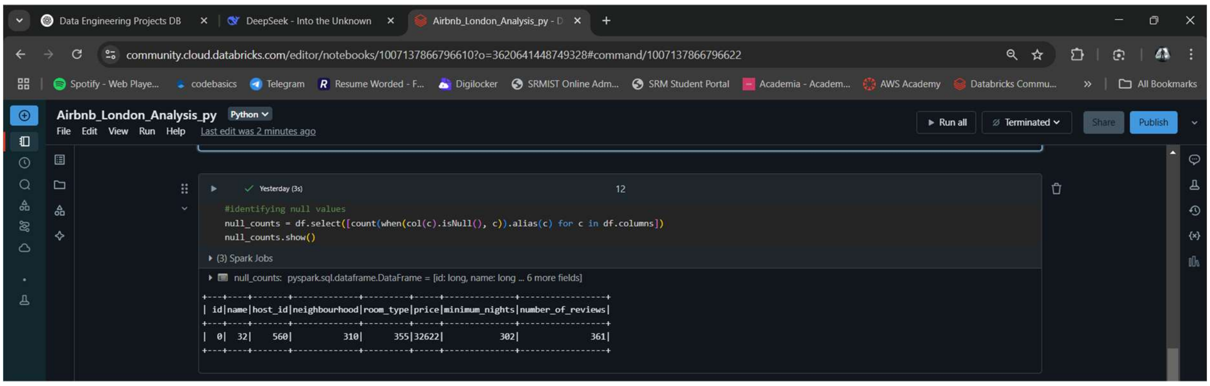
Handling missing values and removing duplicates,



3. **Feature Engineering,**



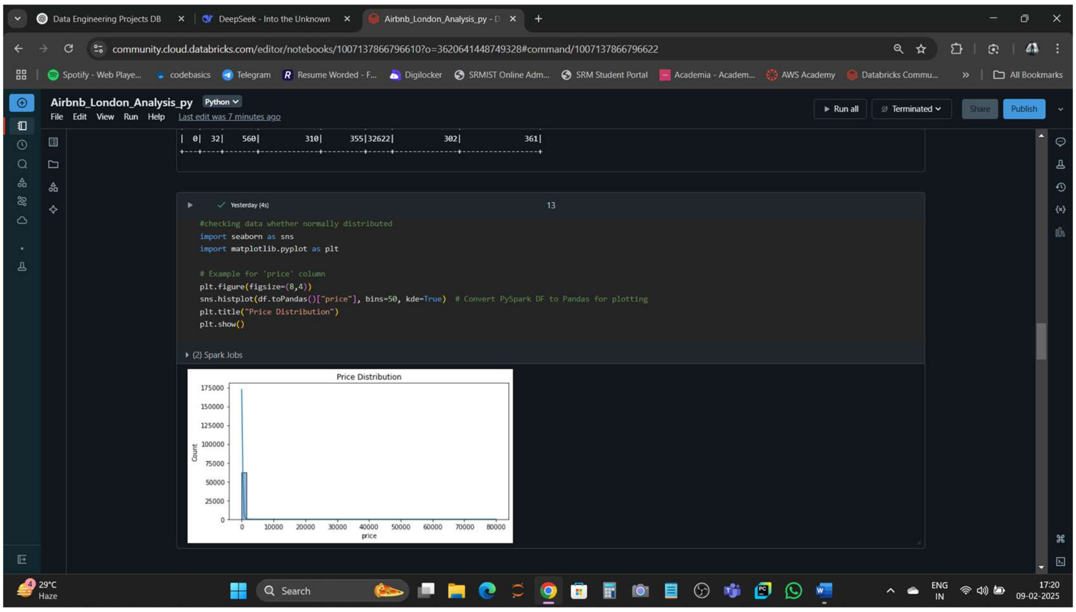
Identifying null values,



It shows that there are several null values in each and every column. There are three ways to ignore or change the null values. They are:

- Mean – If there normal distribution.
- Median – If there any outliers.
- Fill with 0 – If data has any contextual meaning.

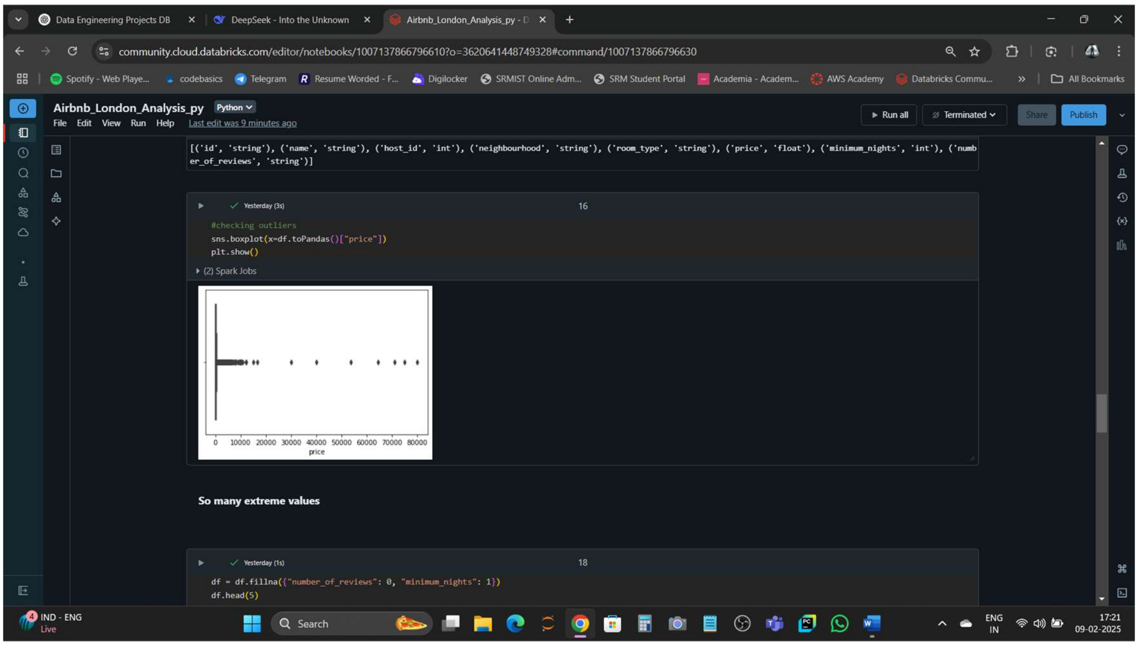
Checking whether the data is normally distributed,



1. If the histogram is **bell-shaped**, the data is likely normally distributed.
2. If it's skewed **left** or **right**, the data is not normally distributed.

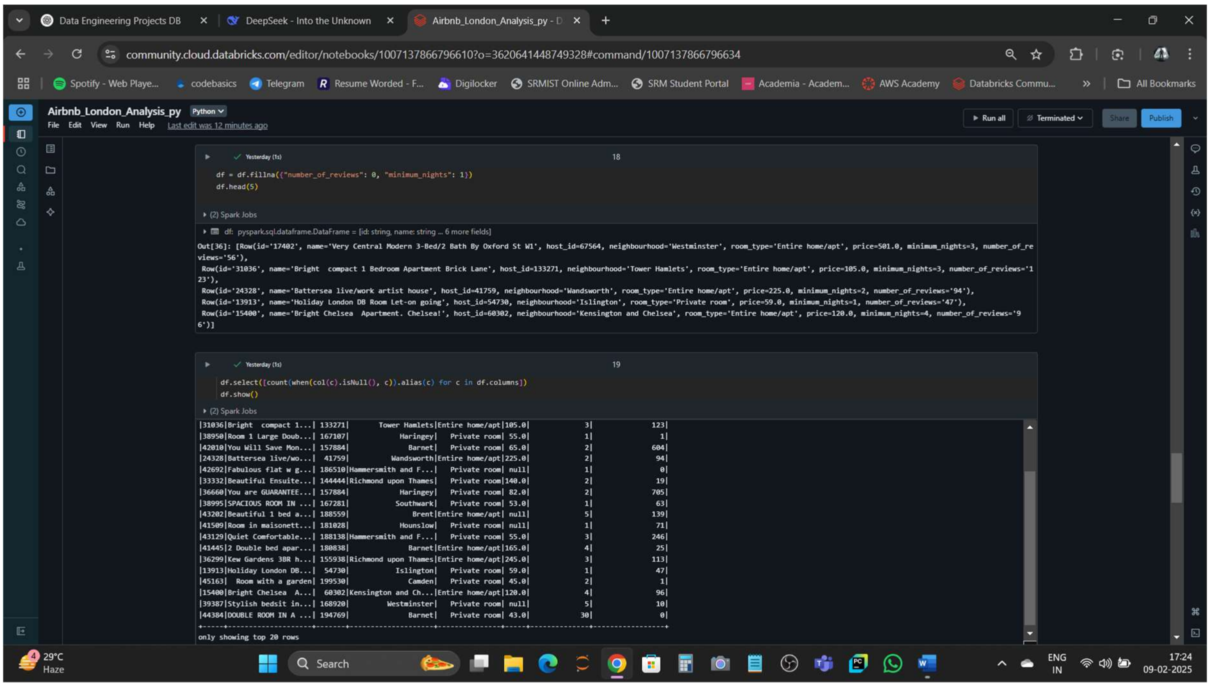
The data has been skewed left in our case. Hence it is not normally distributed.

Checking outliers,

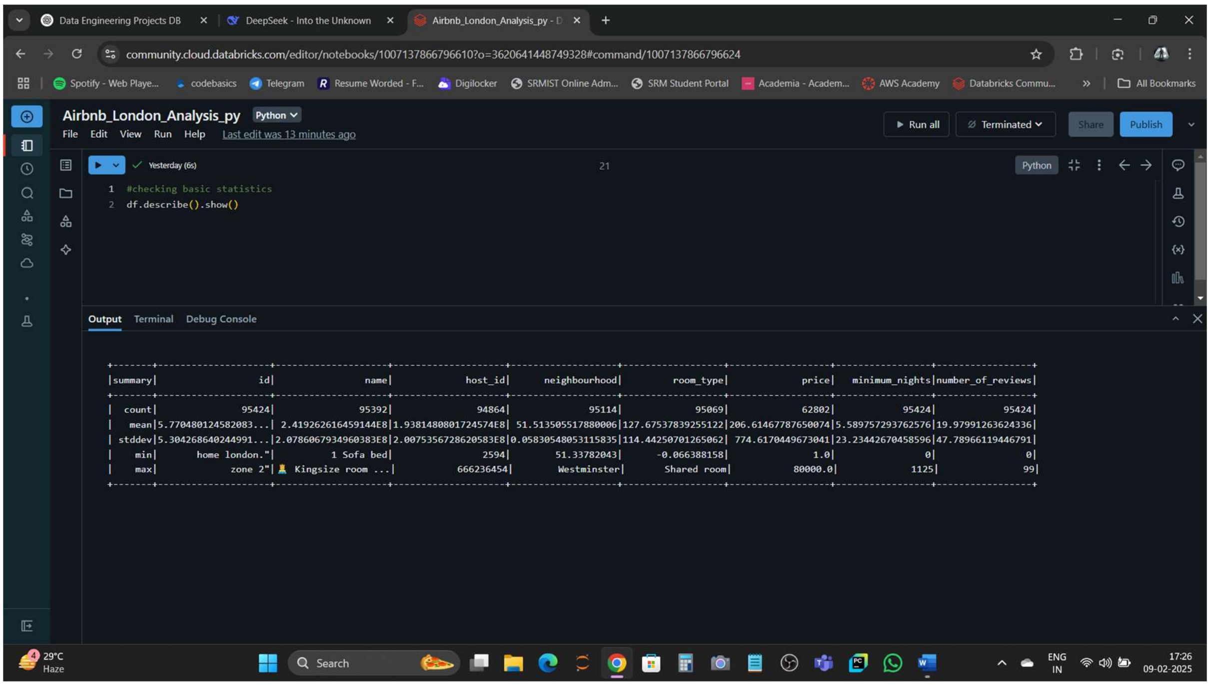


So many extreme values.

By Scaling,

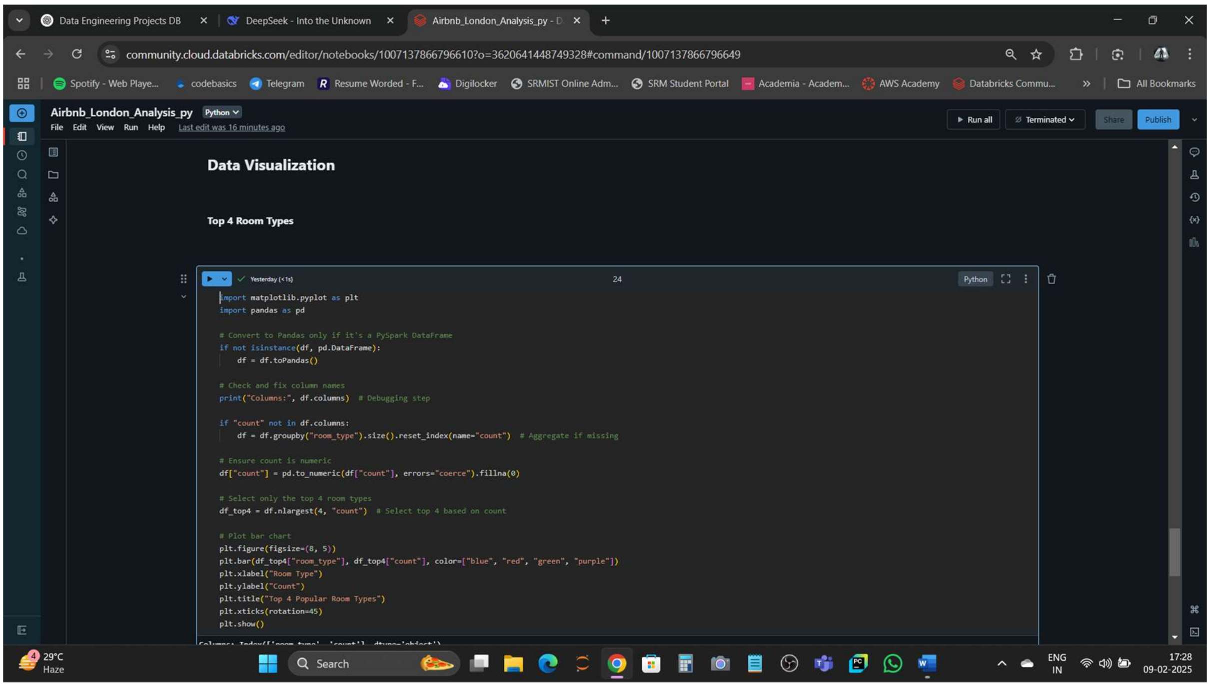


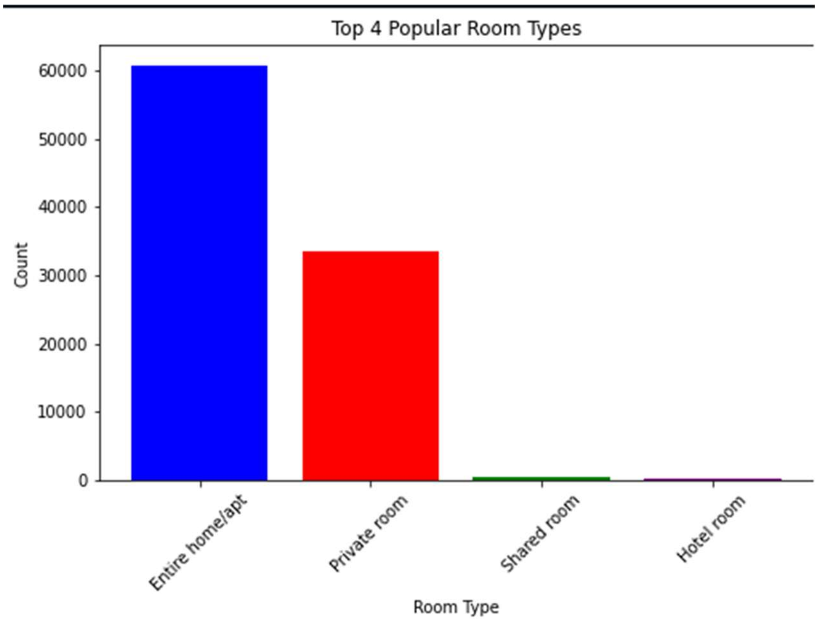
Checking Description of the data,



4. Data Visualization

Visualizing the top 4 room types in the London.

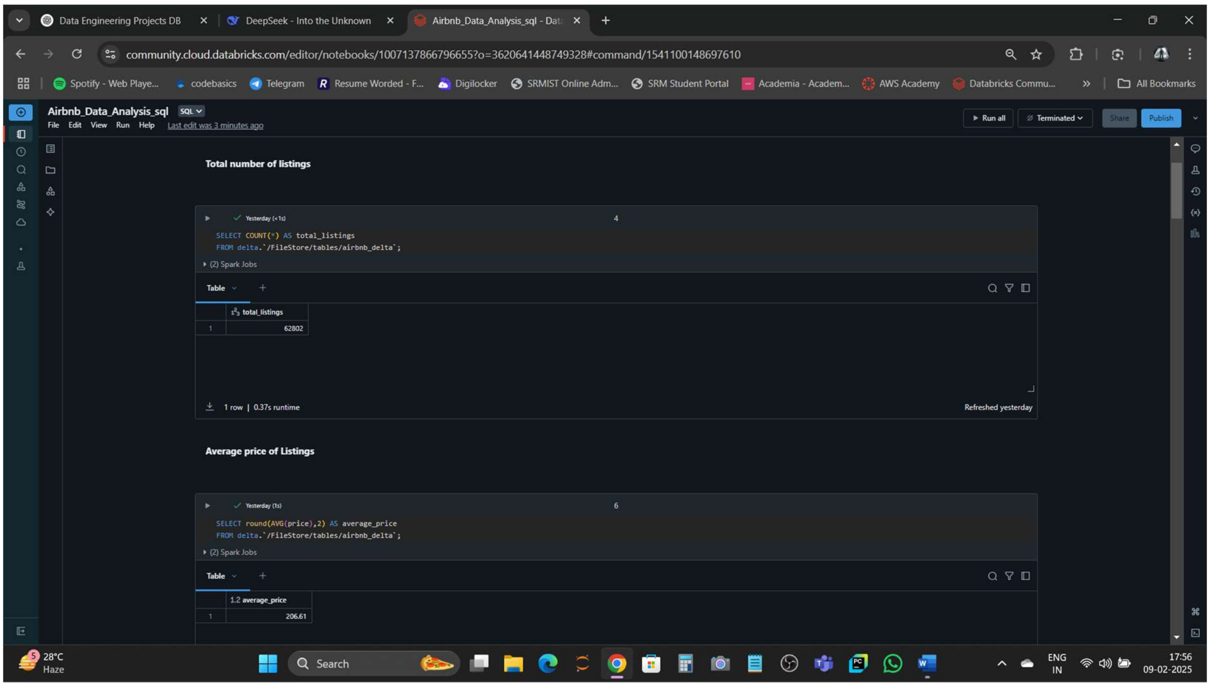


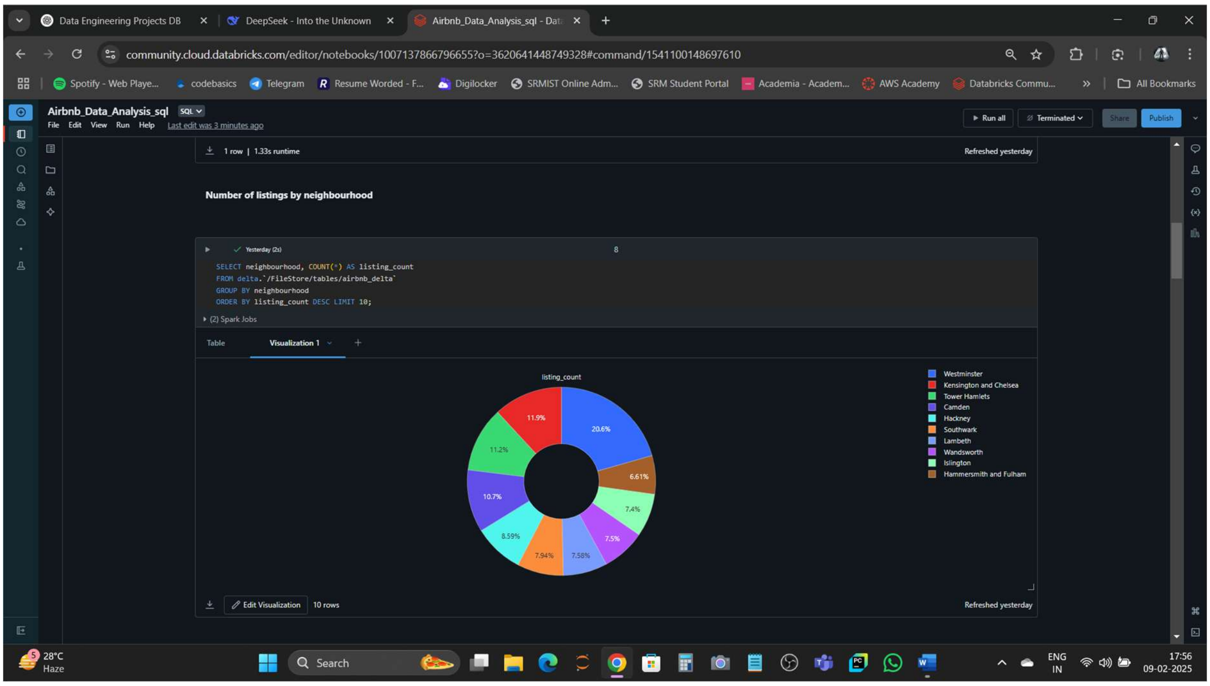


SQL Notebook:

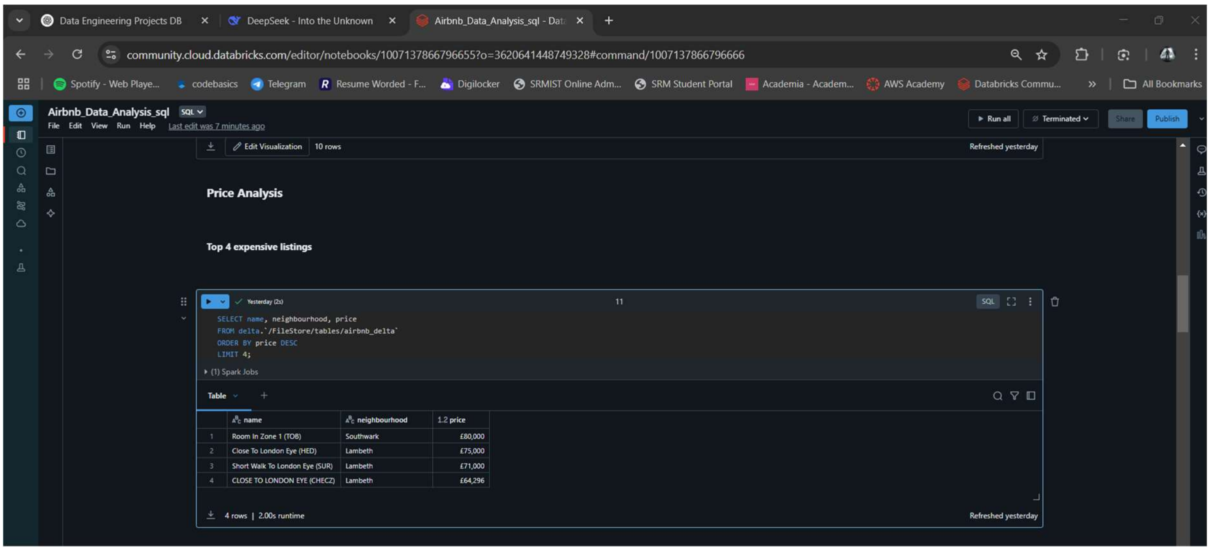
This notebook mainly deals about the further insights such as **Listing Analysis**, **Price analysis**, **Host Analysis**, **Occupancy and Availability analysis** and eventually **Correlation analysis** generated from the data.

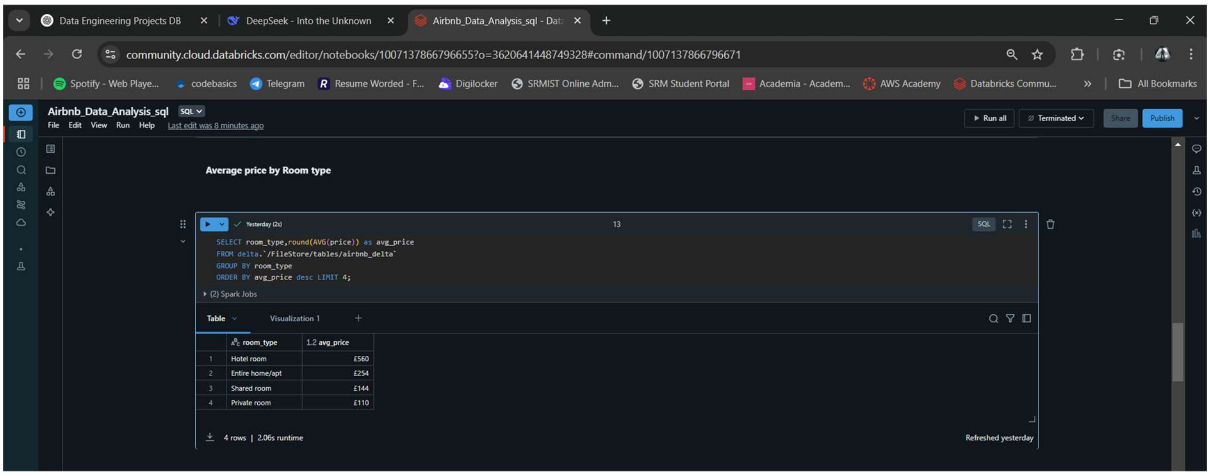
- **Listing Analysis:** Total number of listings, Average price of listings, Number of listings by neighbourhood.



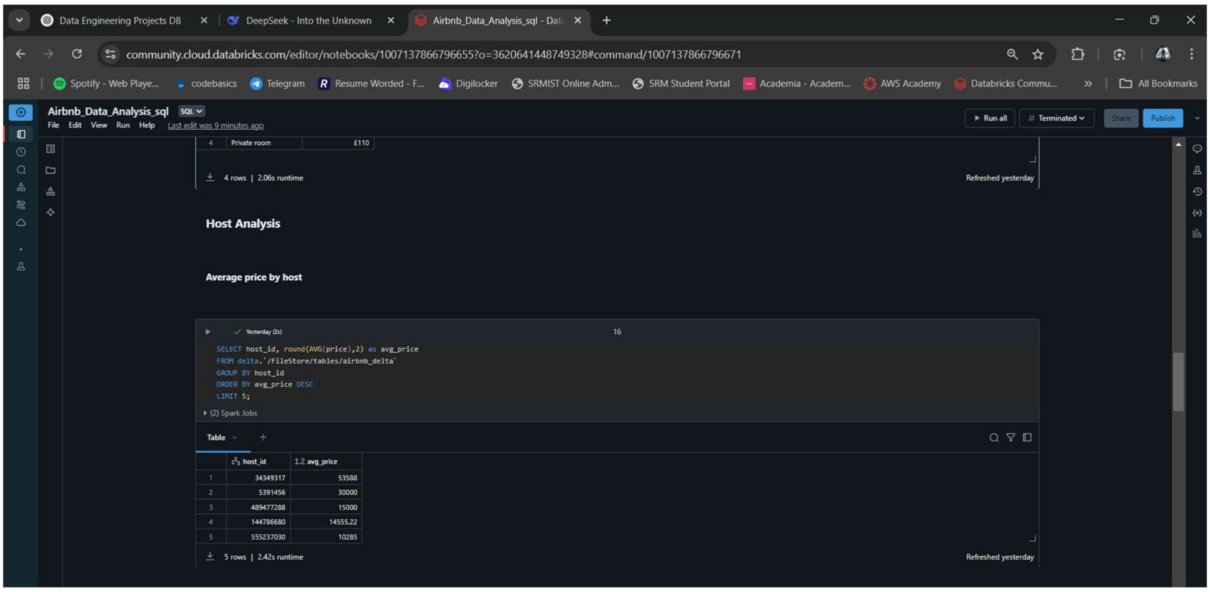


- **Price Analysis:** Top expensive listings, Average price by room type.

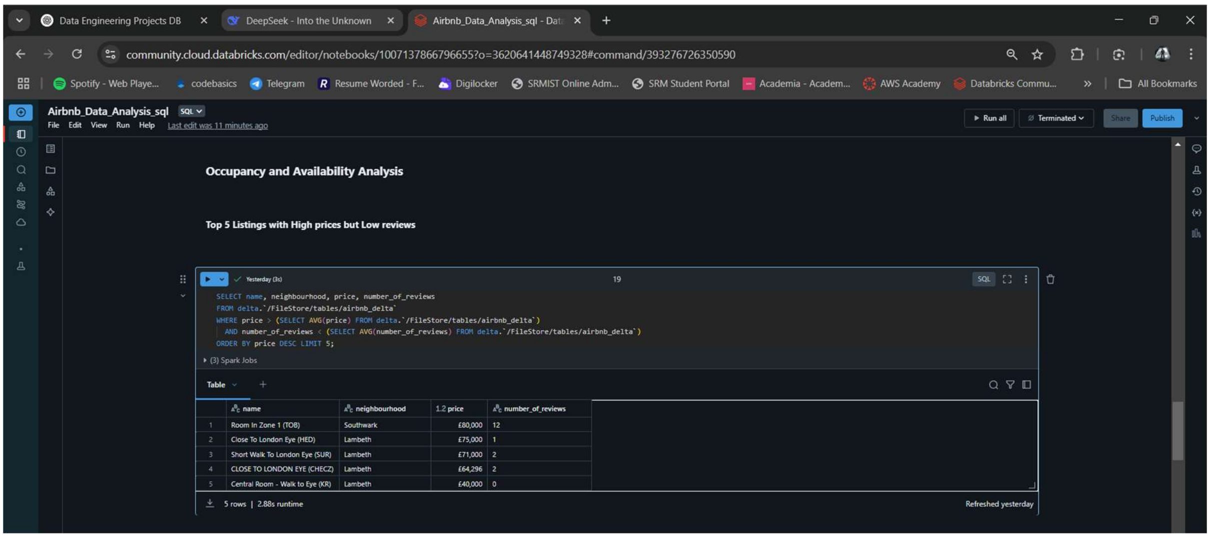


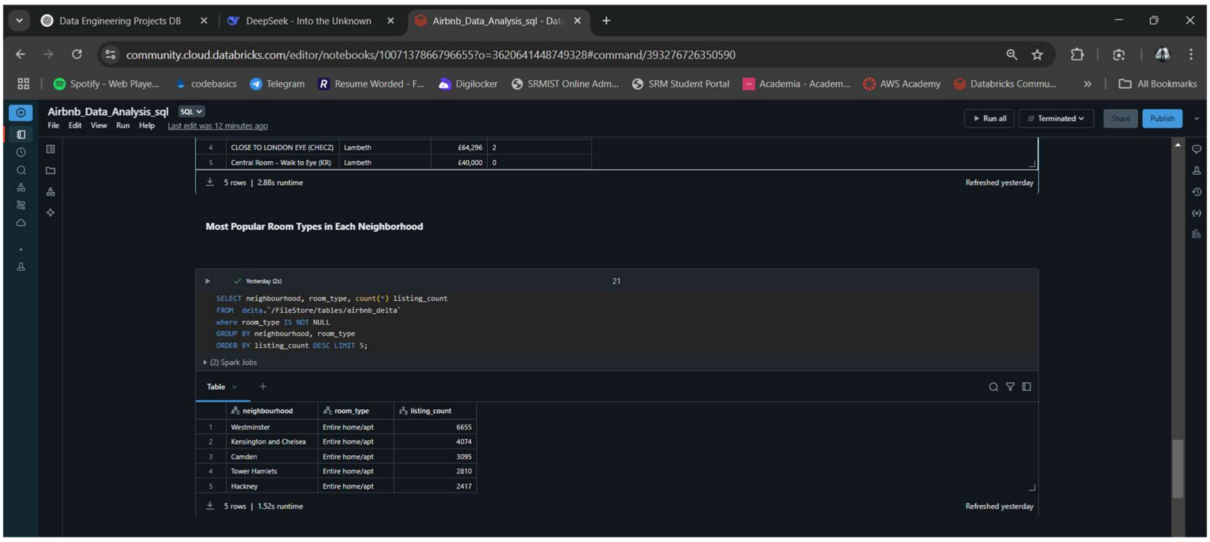


- **Host Analysis:** Average price by host.

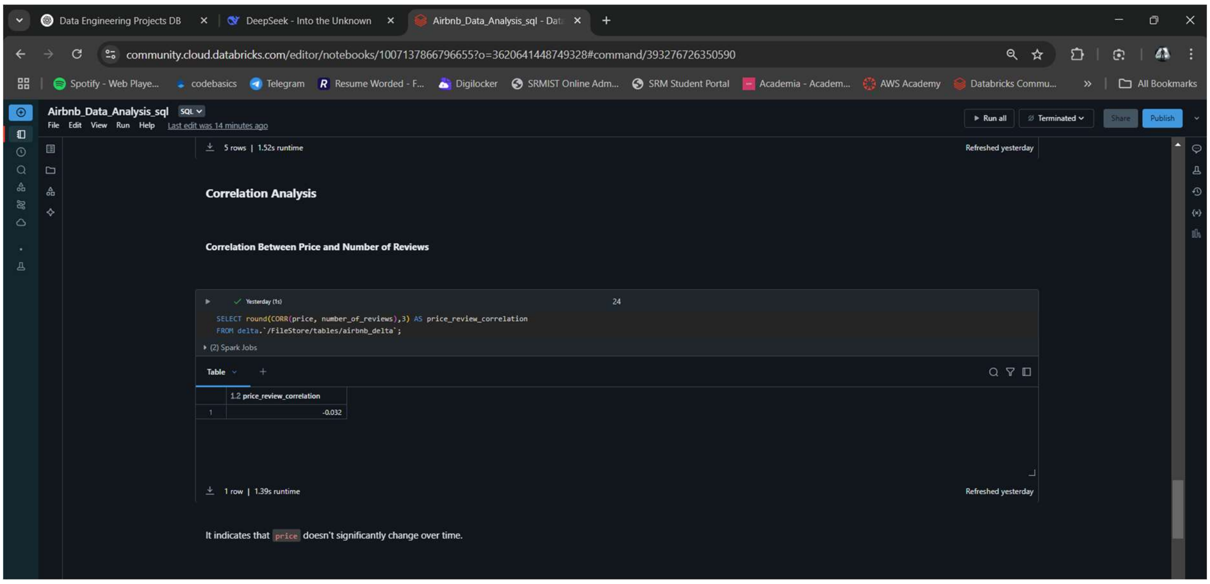


- **Occupancy and Availability Analysis:** Top listings with high prices but low reviews, Most popular room types in each neighbourhood.

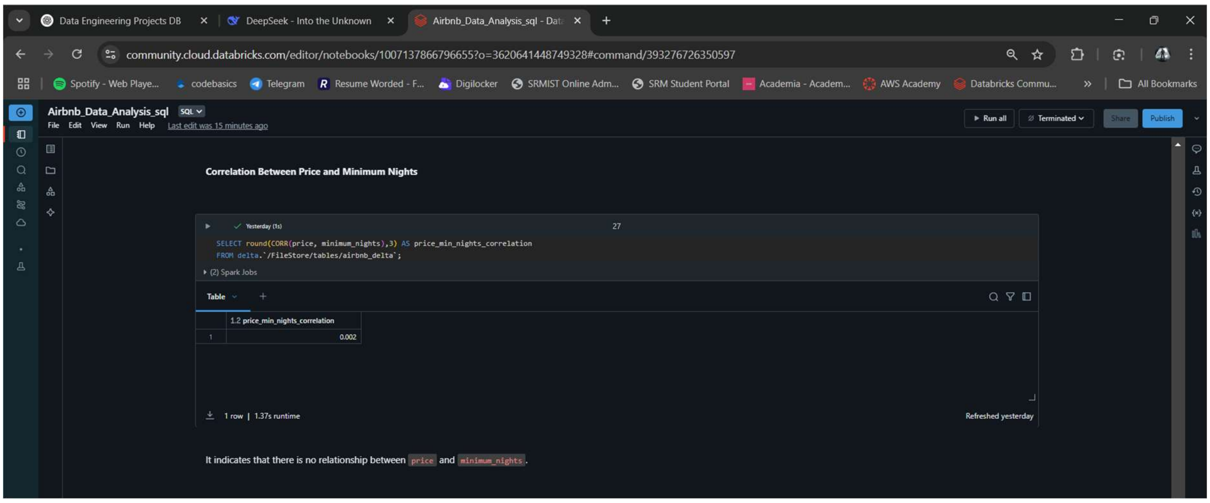




Correlation Analysis: Correlation between Price and Number of reviews, Correlation between Price and Minimum nights.

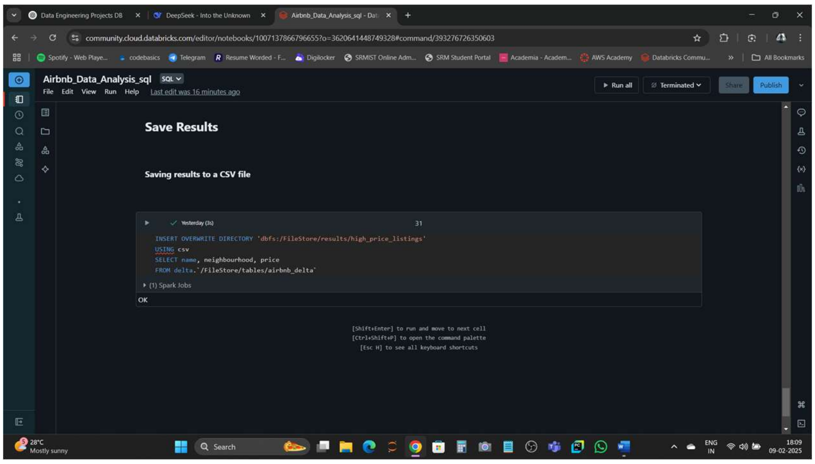


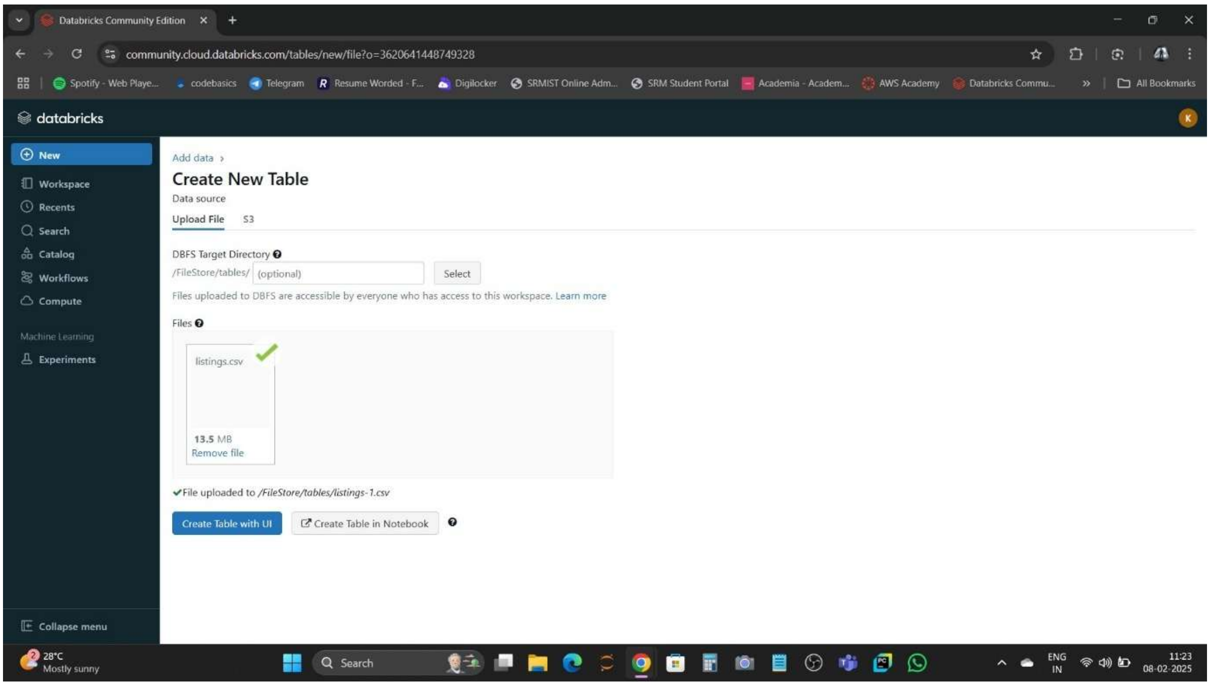
It indicates that **price** doesn't significantly change over time.



It indicates that there is no relationship between **price** and **minimum_nights**.

<u>Saving Results:</u>

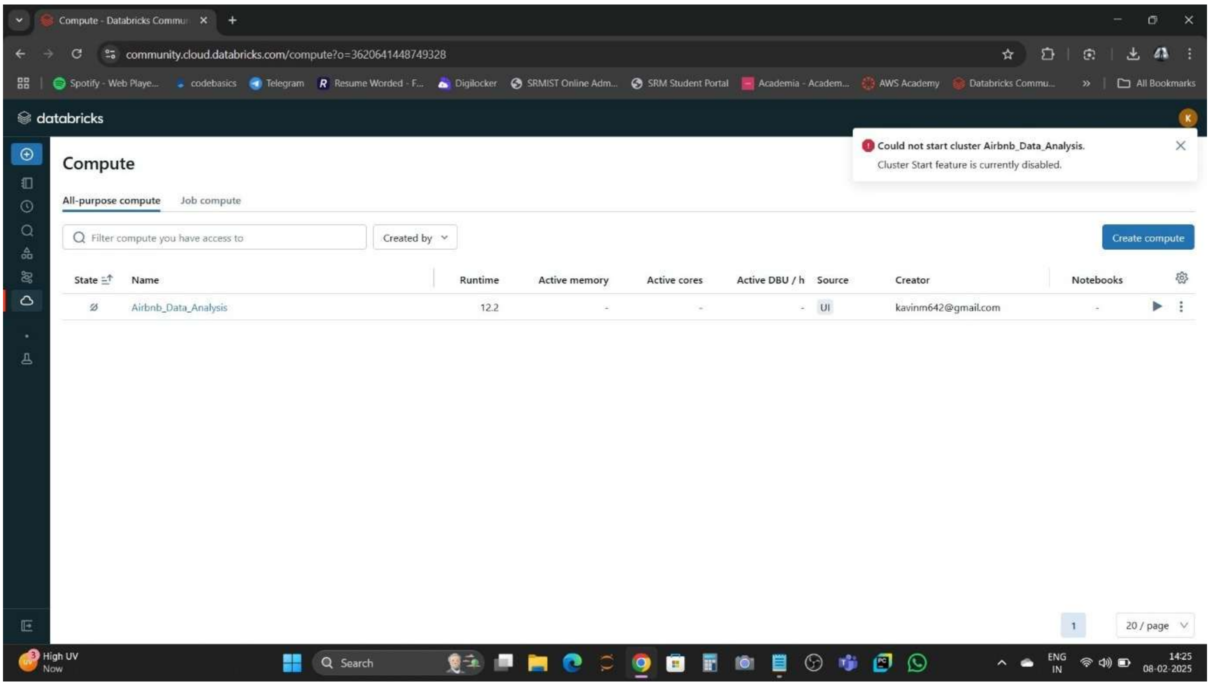




The query extracted Airbnb listings (name, neighbourhood and price). It overwritten any existing files in **high_price_listings.** And finally, the output saved as a csv file which can be viewed and accessed on DBFS (i.e.,) Data Bricks File Store.

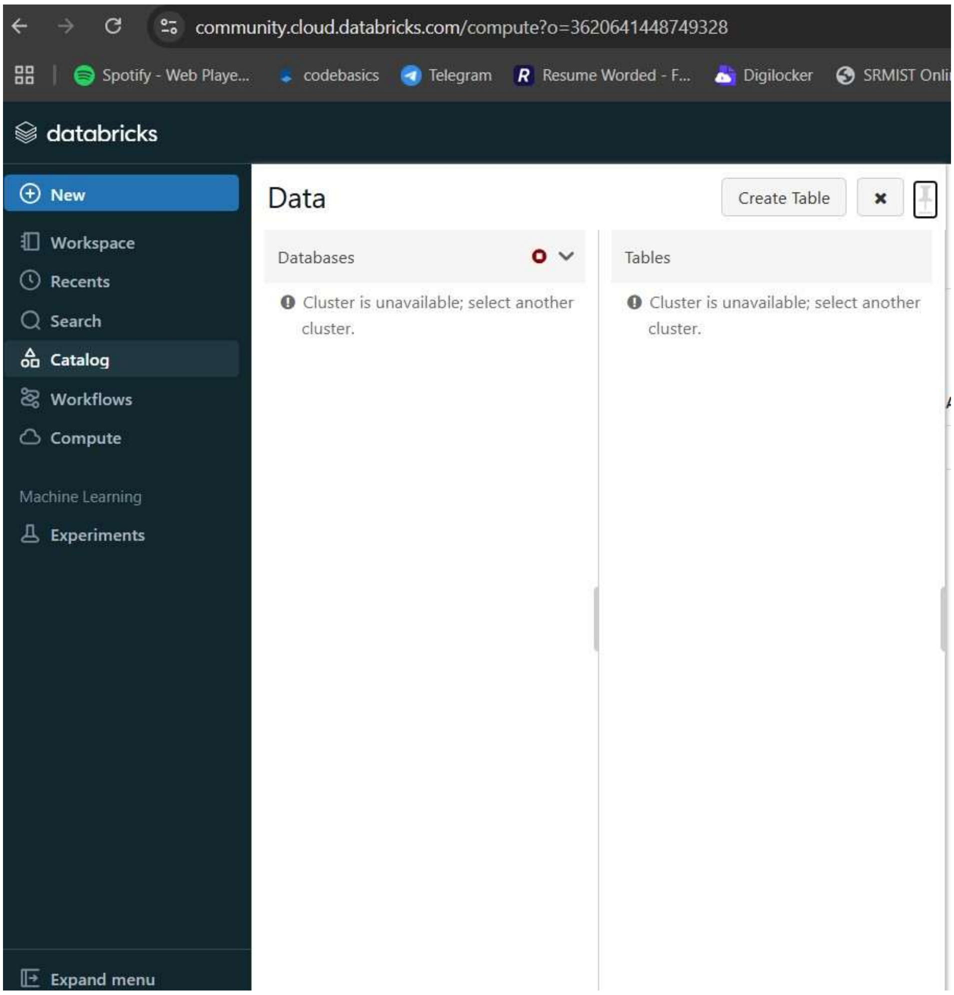
CLUSTER STATUS:

Once the work has been completed, the cluster should be terminated as it consumes more storage and cost if it is a community edition. One major constraint in this community edition is we have certain restrictions such as denial of workflow access, can't restart the cluster once it was terminated and our notebooks will be disabled.



It disabled the cluster as it was terminated once.

Also, we can't able to view the saved report as it will be deleted once the cluster becomes disabled.



PROJECT OUTCOME:

Key findings:

- **Most common room type:** (e.g., Private Room dominates the market).
- **Average price trends:** How prices vary across different neighborhoods.
- **Top neighborhoods:** Areas with the highest number of listings.
- **Availability insights:** Understanding booking patterns.

CONCLUSION:

This analysis provides valuable insights into the Airbnb market in London. Hosts can optimize pricing strategies, while travelers can find budget-friendly locations. Additionally, investors and policymakers can leverage this data to understand market dynamics and improve regulations.

FUTURE WORKS:

- Leverage Machine Learning models to predict price fluctuations.
- Exploring Sentiment Analysis on Airbnb reviews.
- Extend the analysis to include seasonality trends.

LINKS:

- Dataset Link: https://data.insideairbnb.com/united-kingdom/england/london/2024-12-11/data/listings.csv.gz
- Python Notebook: https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/3620641448749328/1007137866796610/7261851981964810/latest.html
- SQL Notebook: https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/3620641448749328/1007137866796655/7261851981964810/latest.html