**Azure Databricks Coding Assessment:**

**Name**: Abhiram Basa

**Date**: 06/01/2024

**Questions:**

1. Create a cluster &Attach the notebook to the cluster and run all commands in the notebook&creates a Data Frame from a Databricks data set&Create a Visualizations in Data bricks notebooks  
   &Rename, duplicate, or remove a visualization or data profile.

2.Explain the copy activity in Azure data factory.

**Answers and hands on output :**

**Answer 1 :**

**Cluster:** Cluster provides a unified platform for various tasks like running notebooks,running pipelines, etc.

* Clusters are usually a virtual machines which helps us to complete the tasks.

**Cluster details:**

* **Cluster size:** 15gb memory
* **Active cores:** 2 cores
* **Cluster mode:** single node

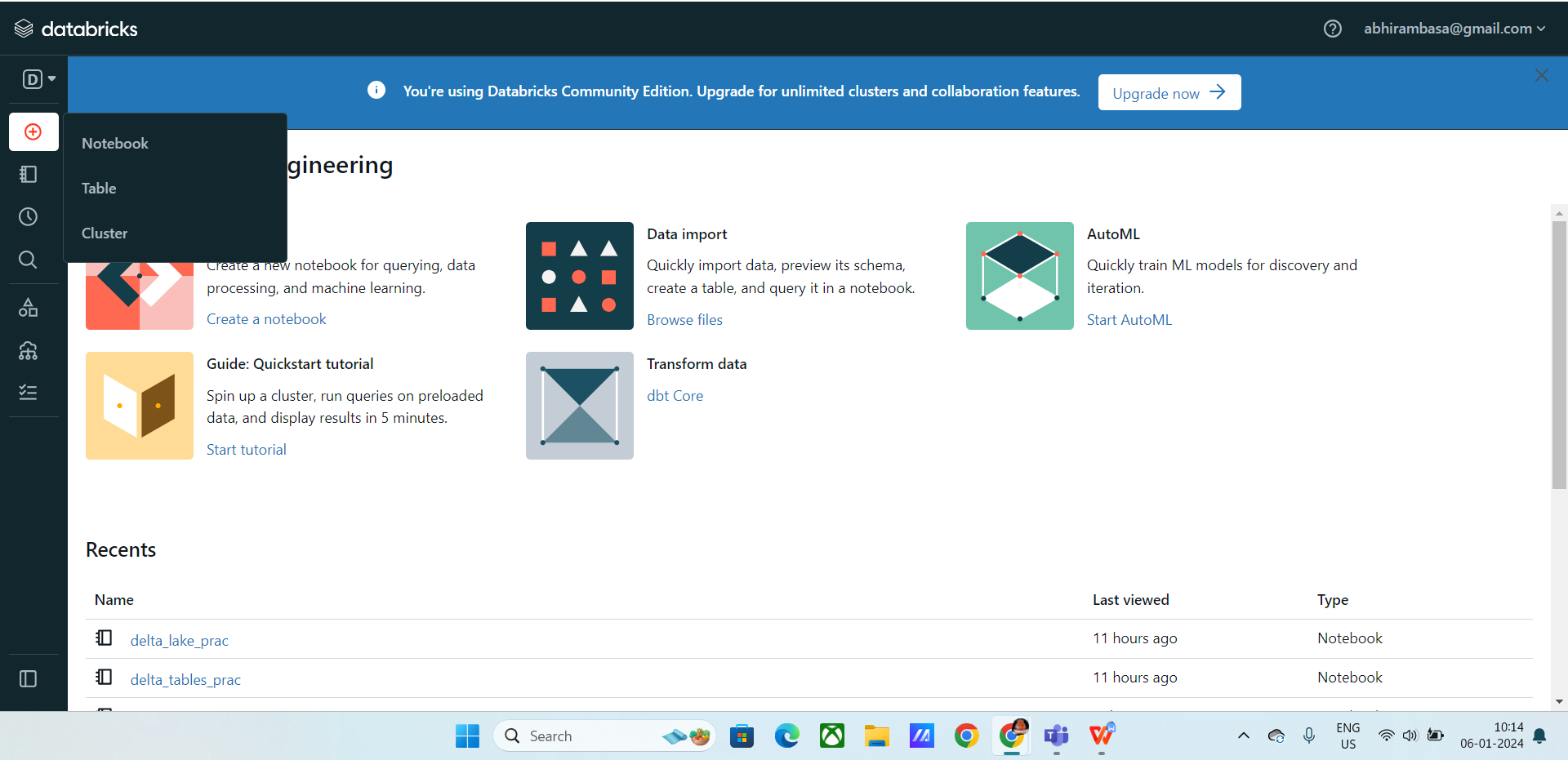
Single mode contains one driver and one worker

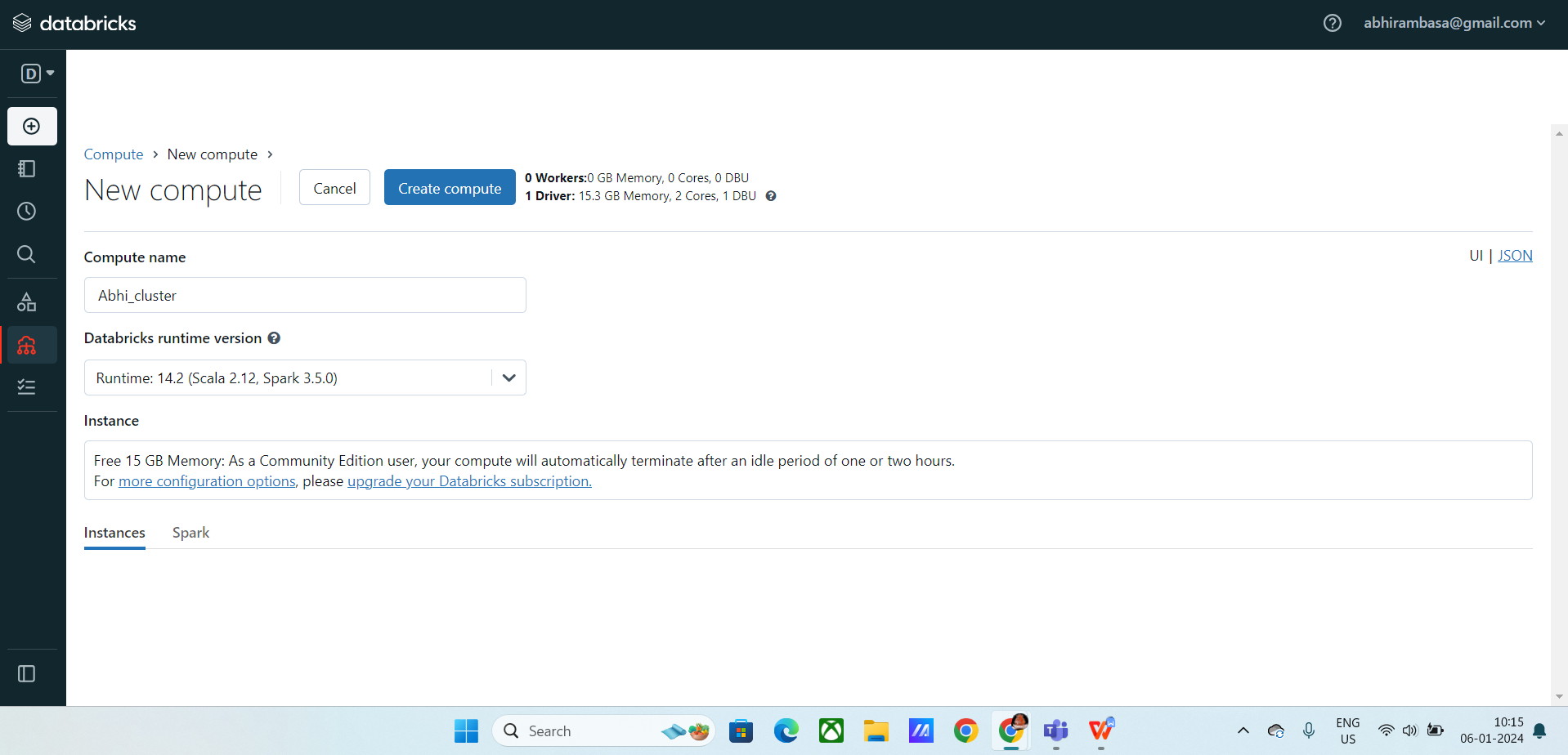
**Driver**: assigns the tasks to the worker

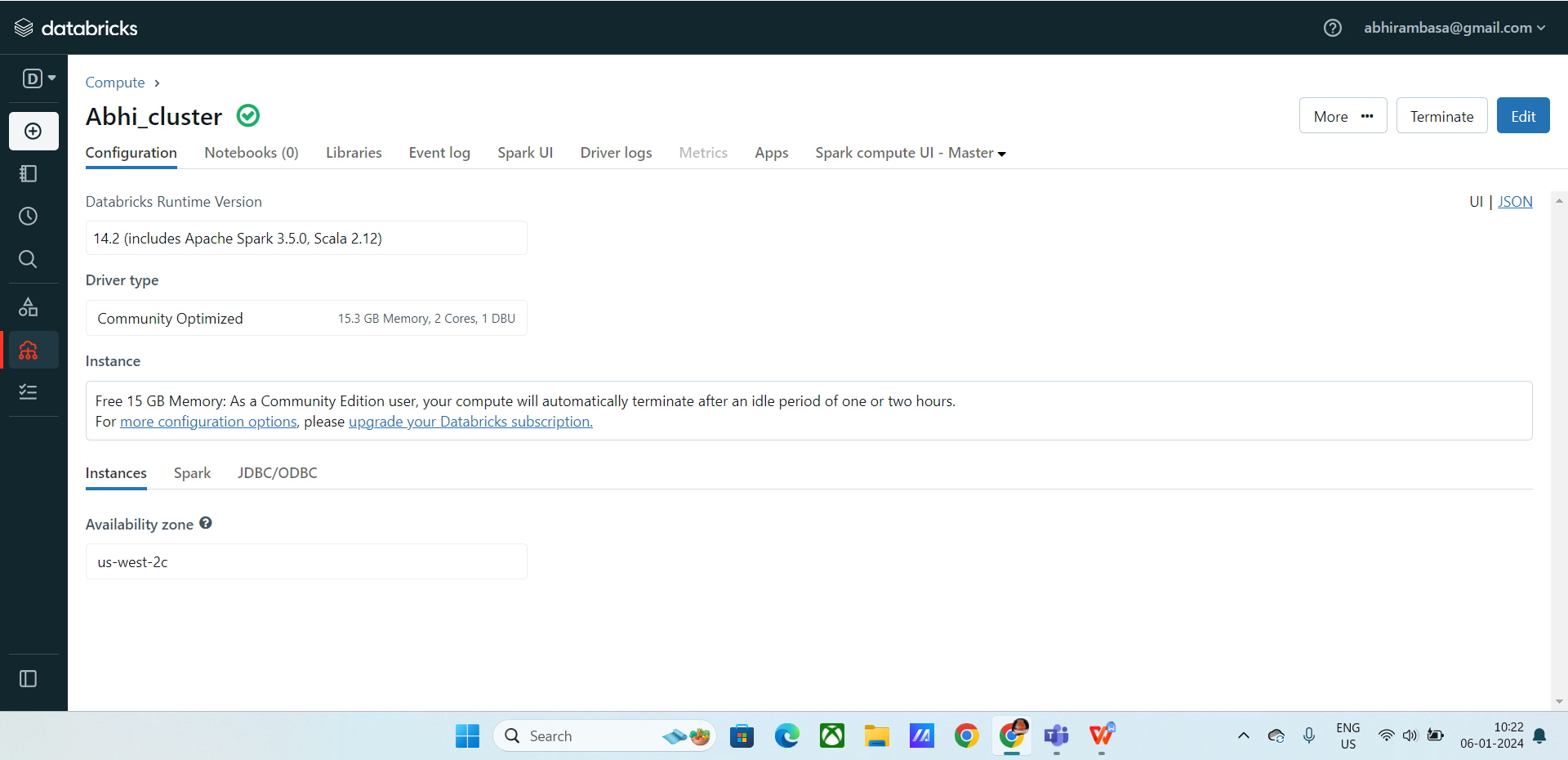
**Worker**: who does the tasks

**How to create a cluster**

* Click on New.
* Click on Cluster.
* Give a name,select the the mode and select the scala version.
* Click on create.

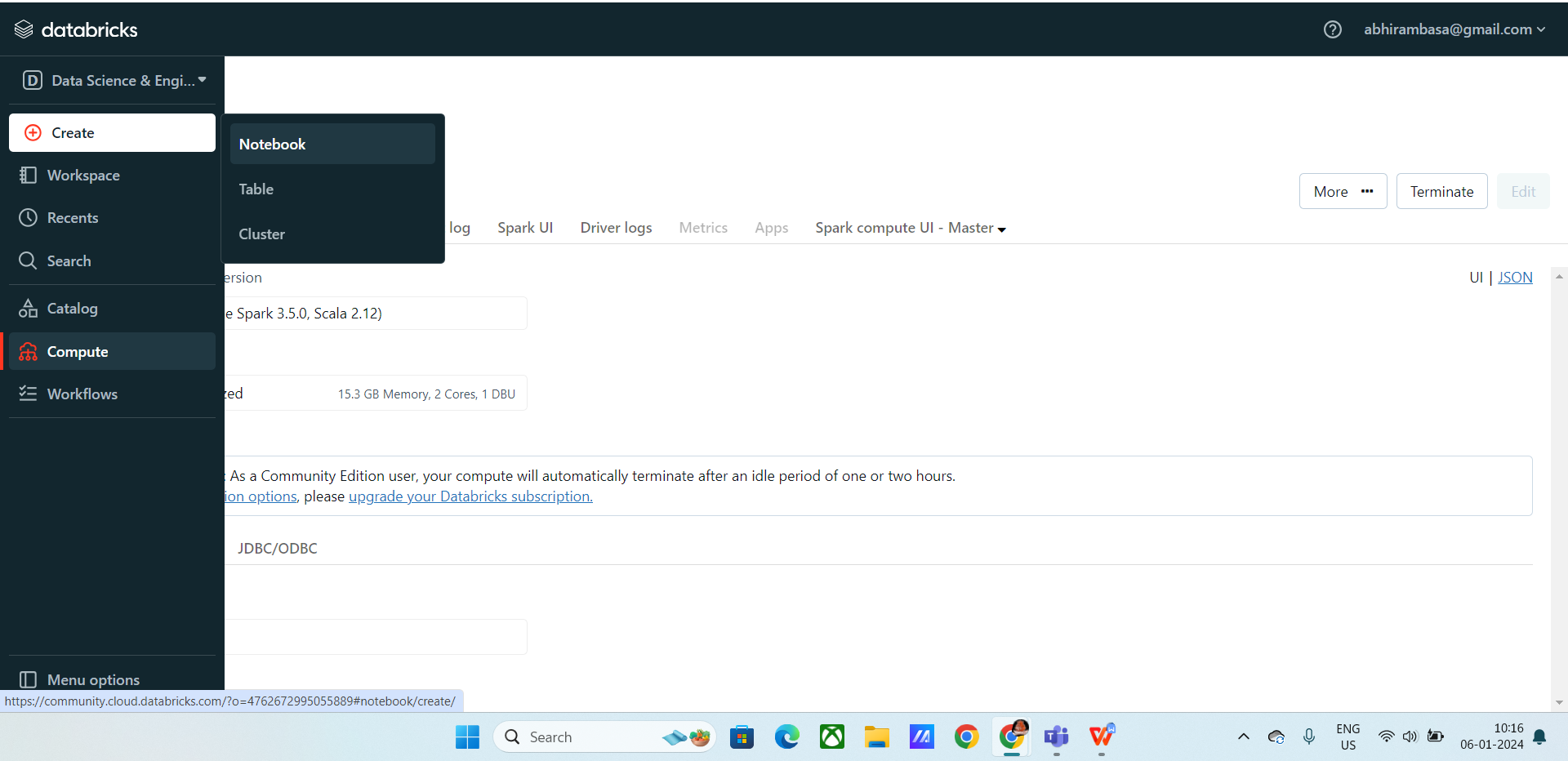


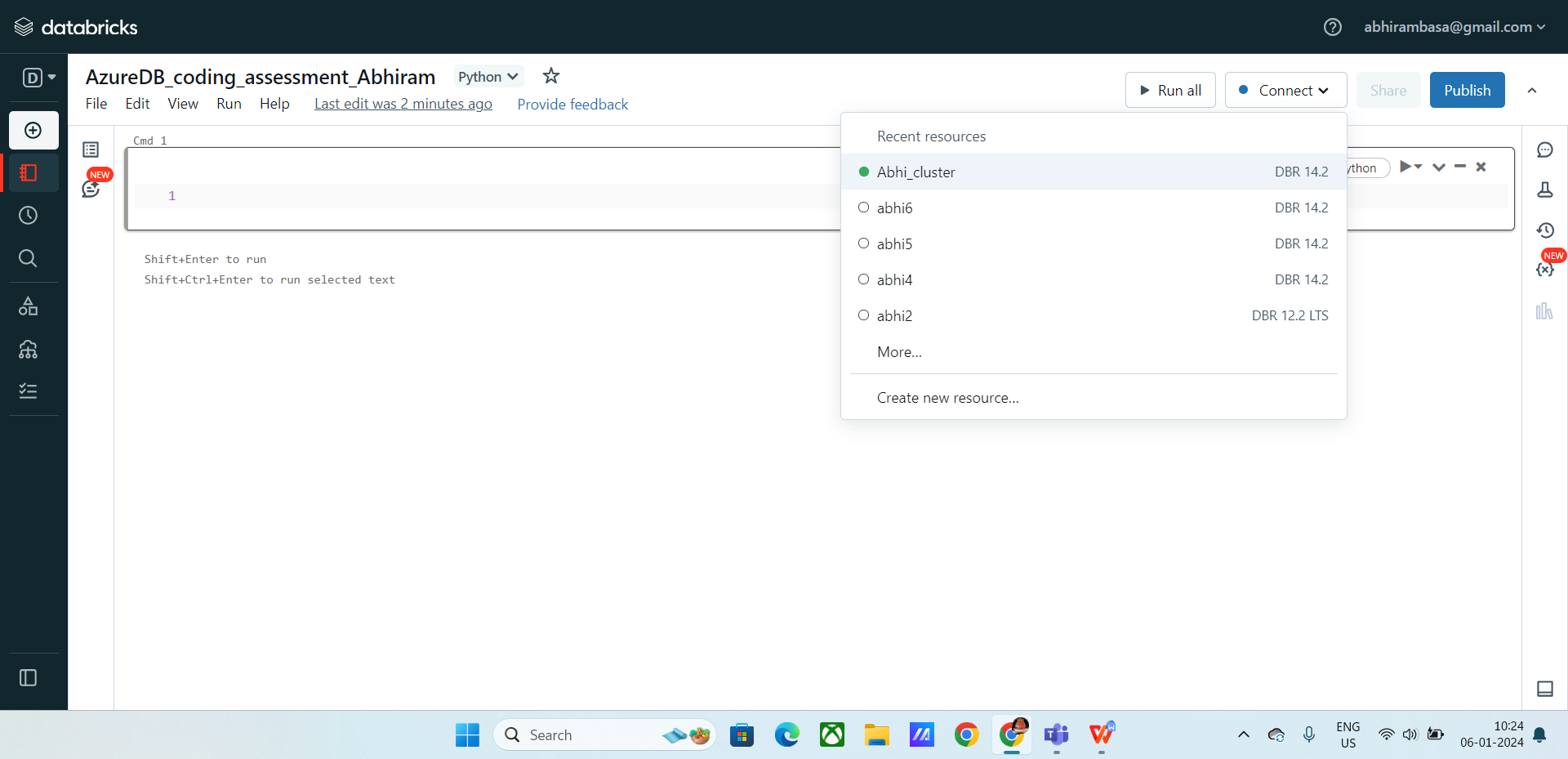




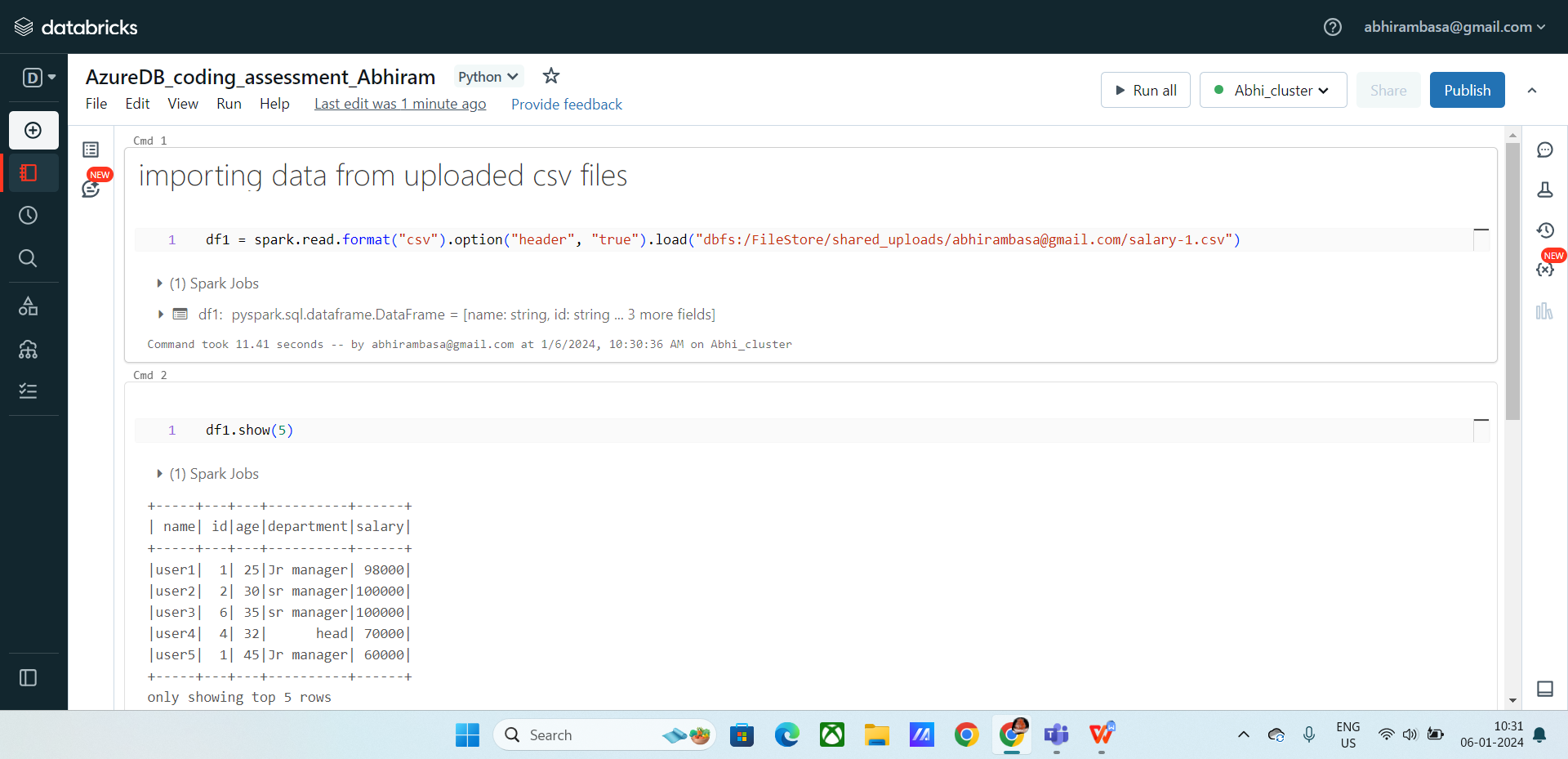
**Attaching a notebook to cluster:**

* Attach the created notebook to the cluster
* Click on top right “connect”
* Select the cluster to connect





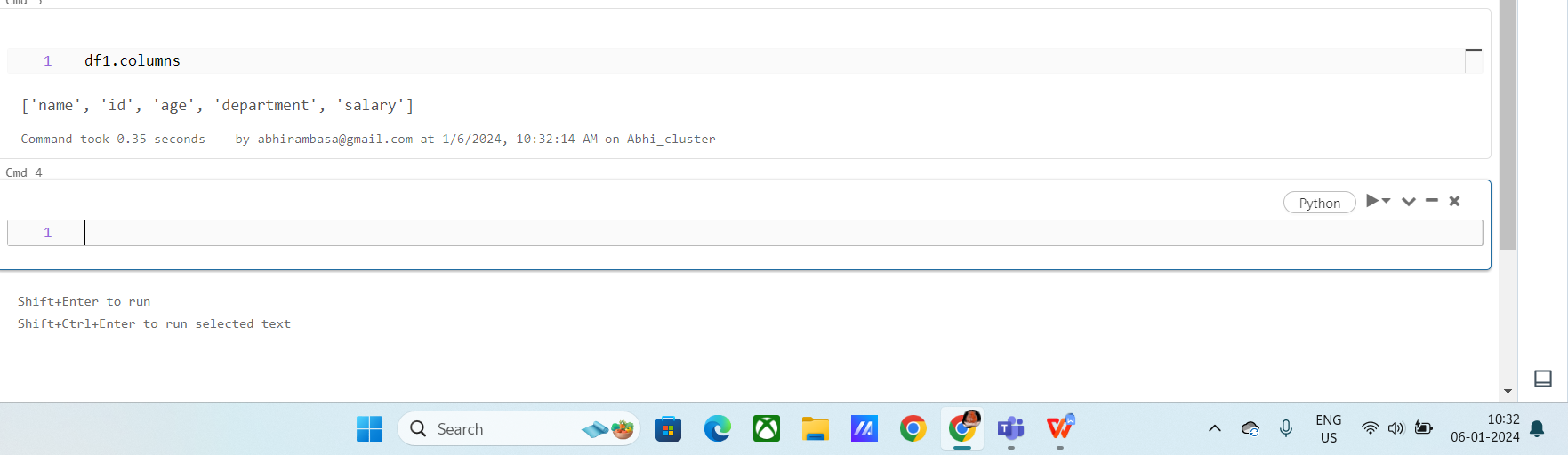
**Running commands**



**Commands:**

Df1.show() : shows the csv file

Df1.columns () : shows all the columns of the table.



**creating a data frame from data brick data sets**

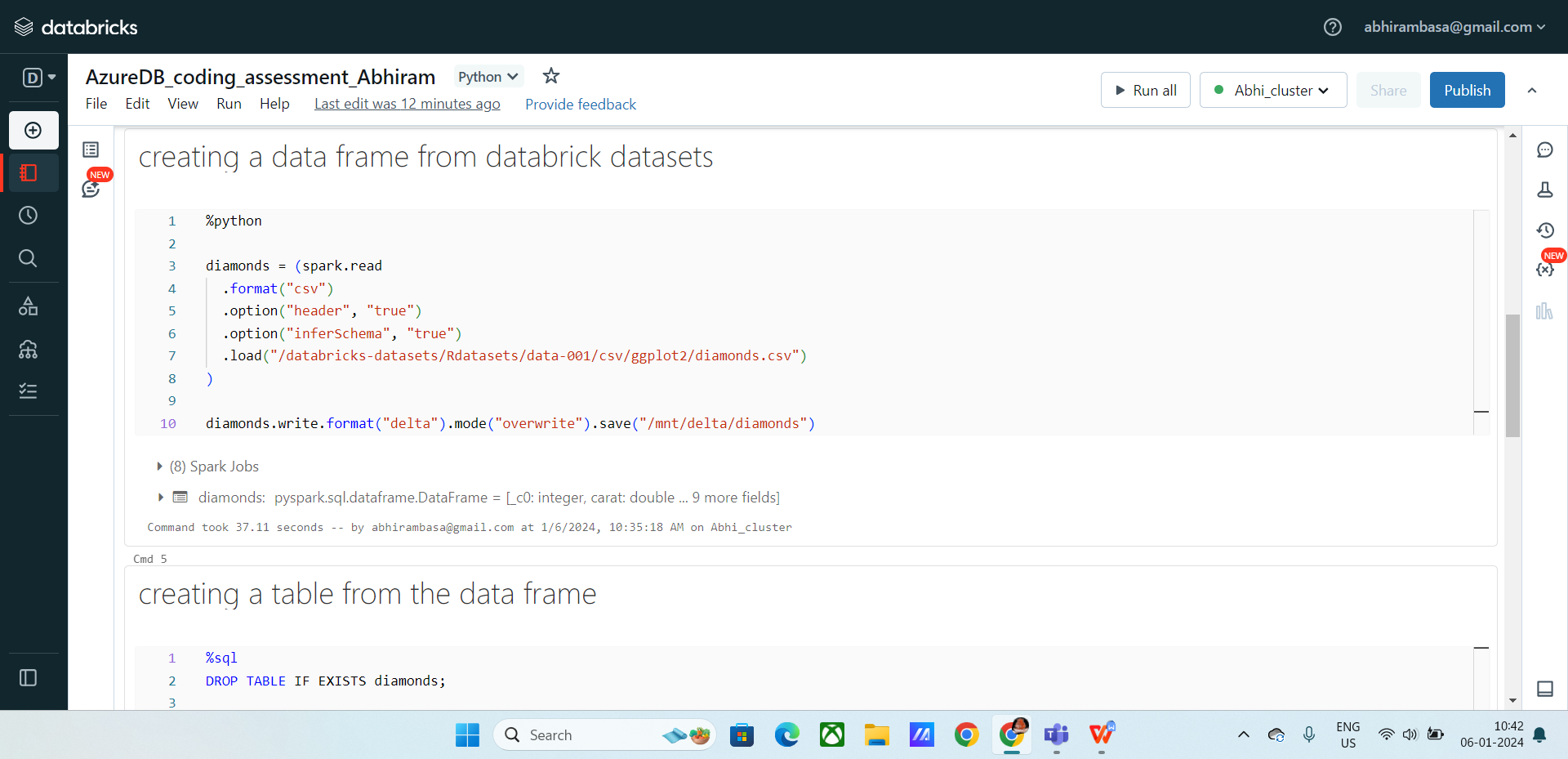
Now import the data bricks data sets from the url

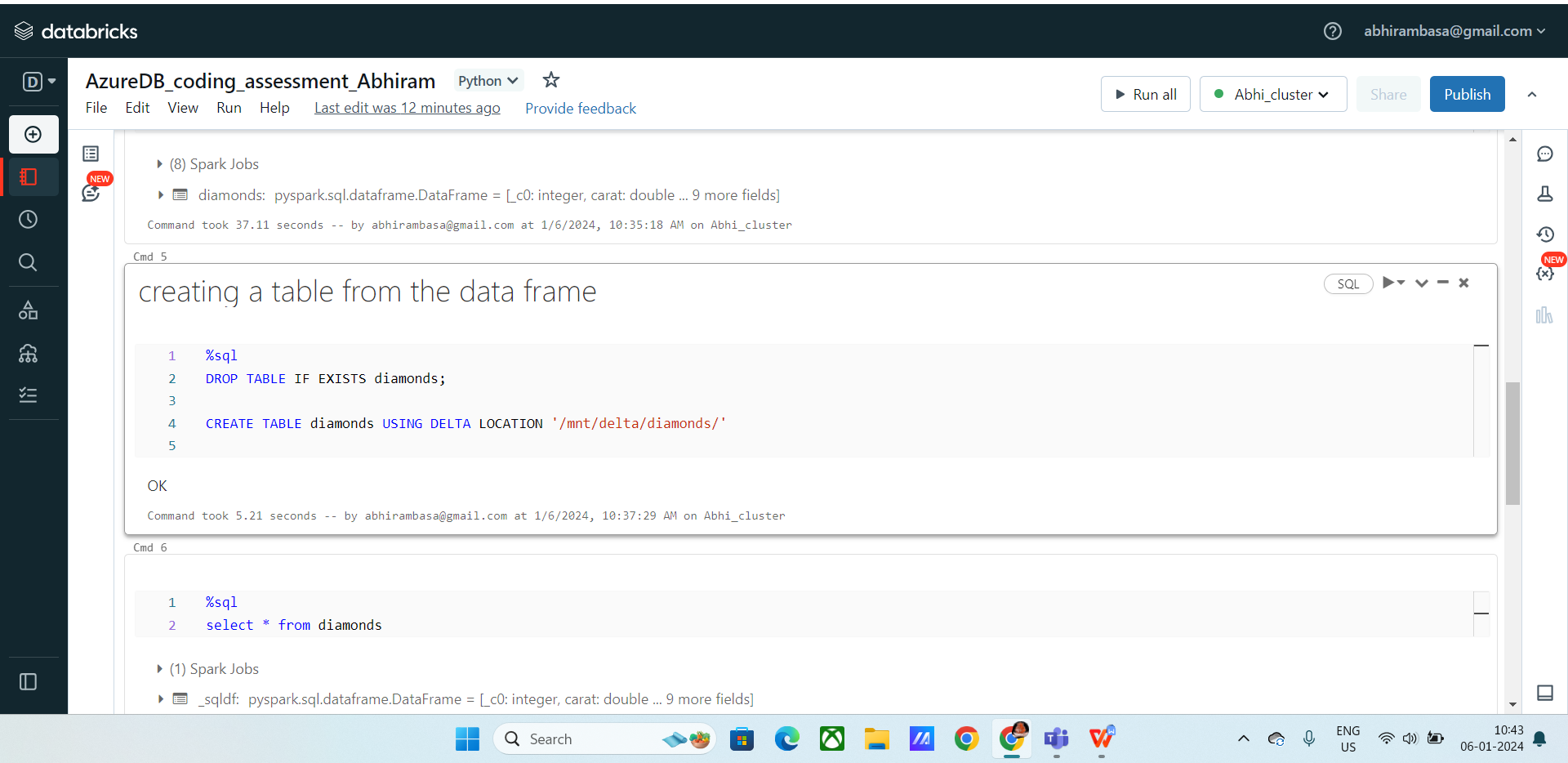
You can create dataframe using the following

**Df.createDataFrame(data, schema)**

**Data Frame :** It contains data in the form of rows and columns, using data frame we can easily manage huge amount of data.

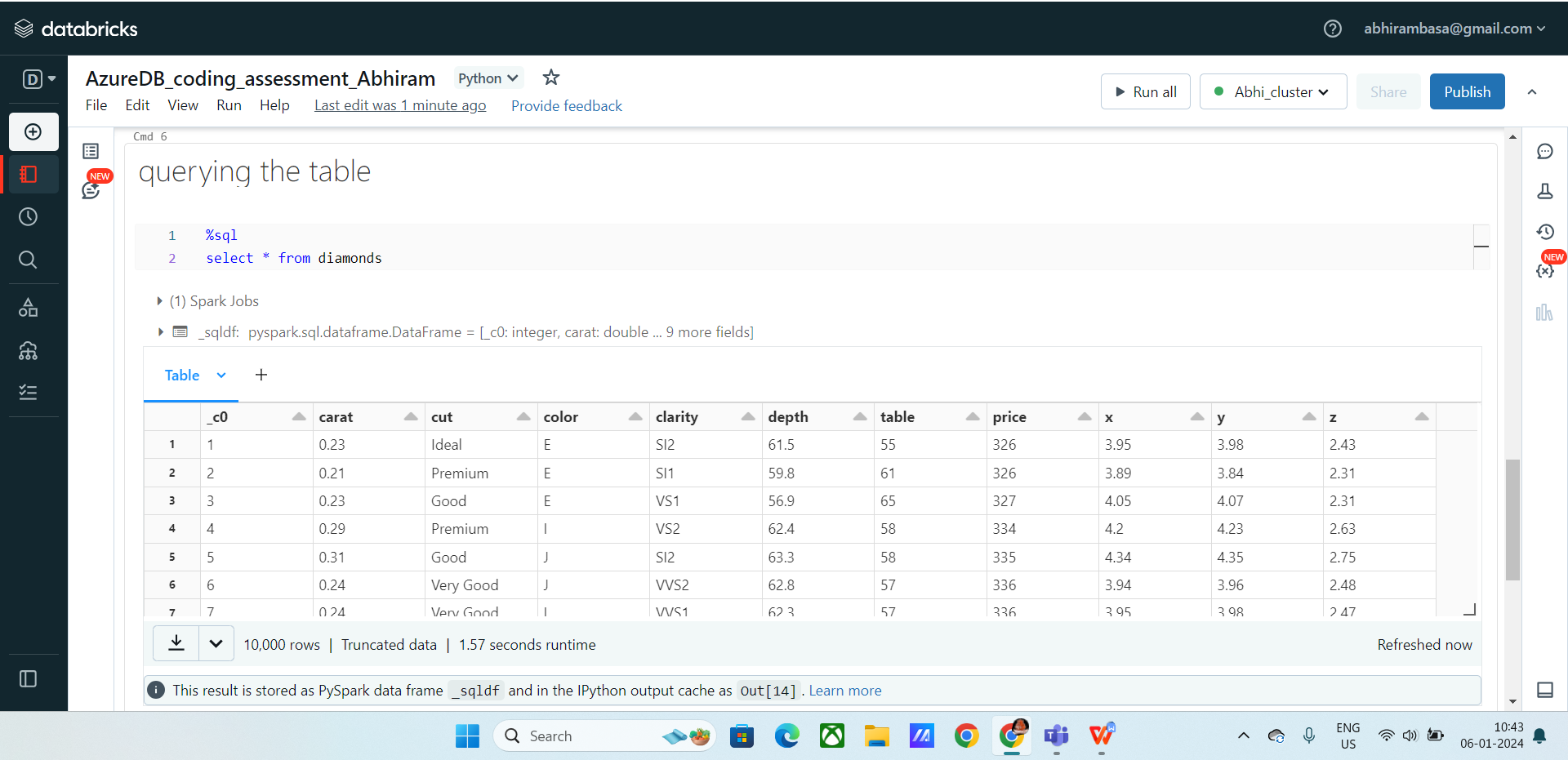
**Schema:** Schema is the structure of data frame columns.



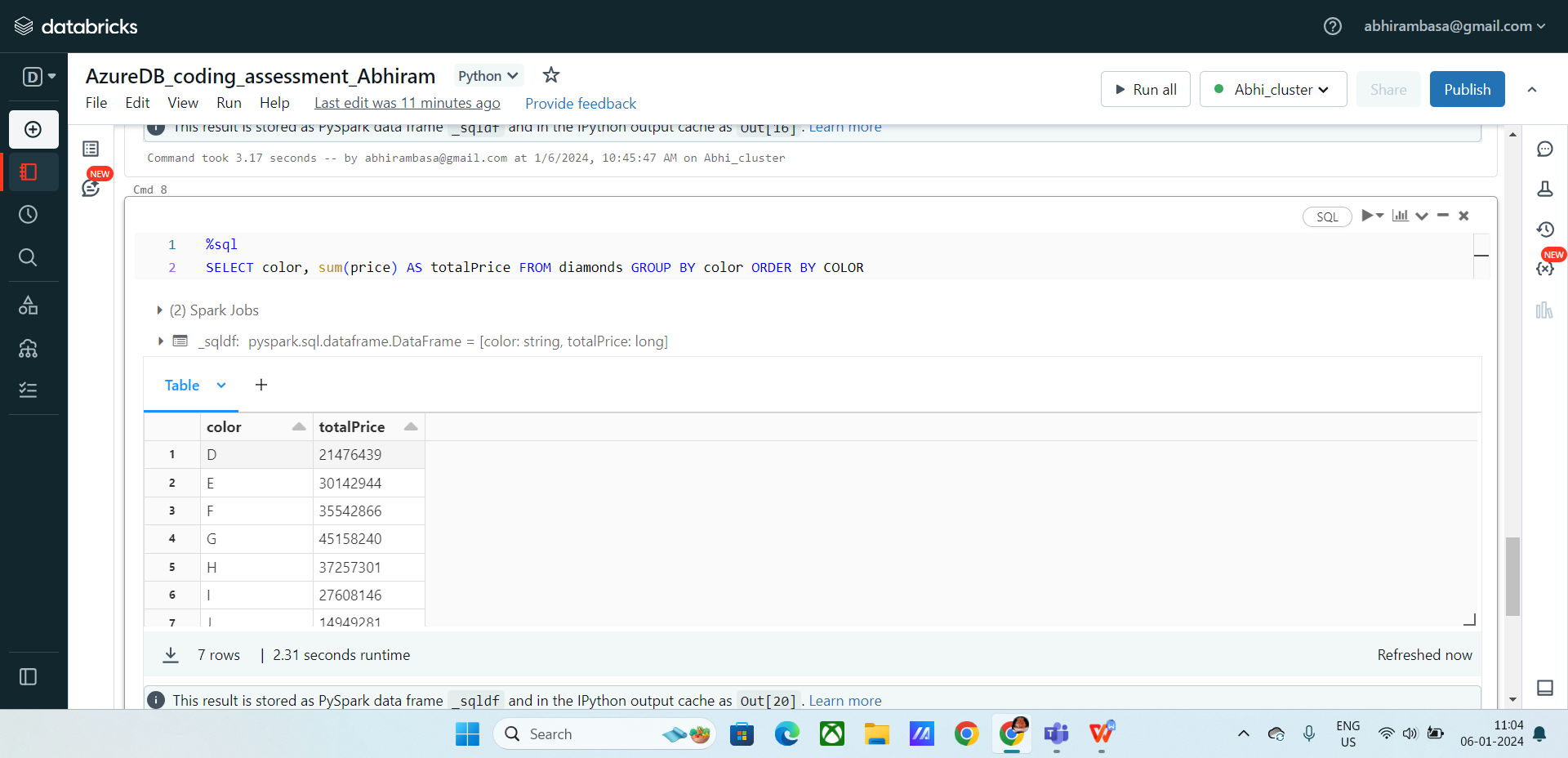


**Querying the table:**

Using %SQL, to execute the sql queries in the notebook.

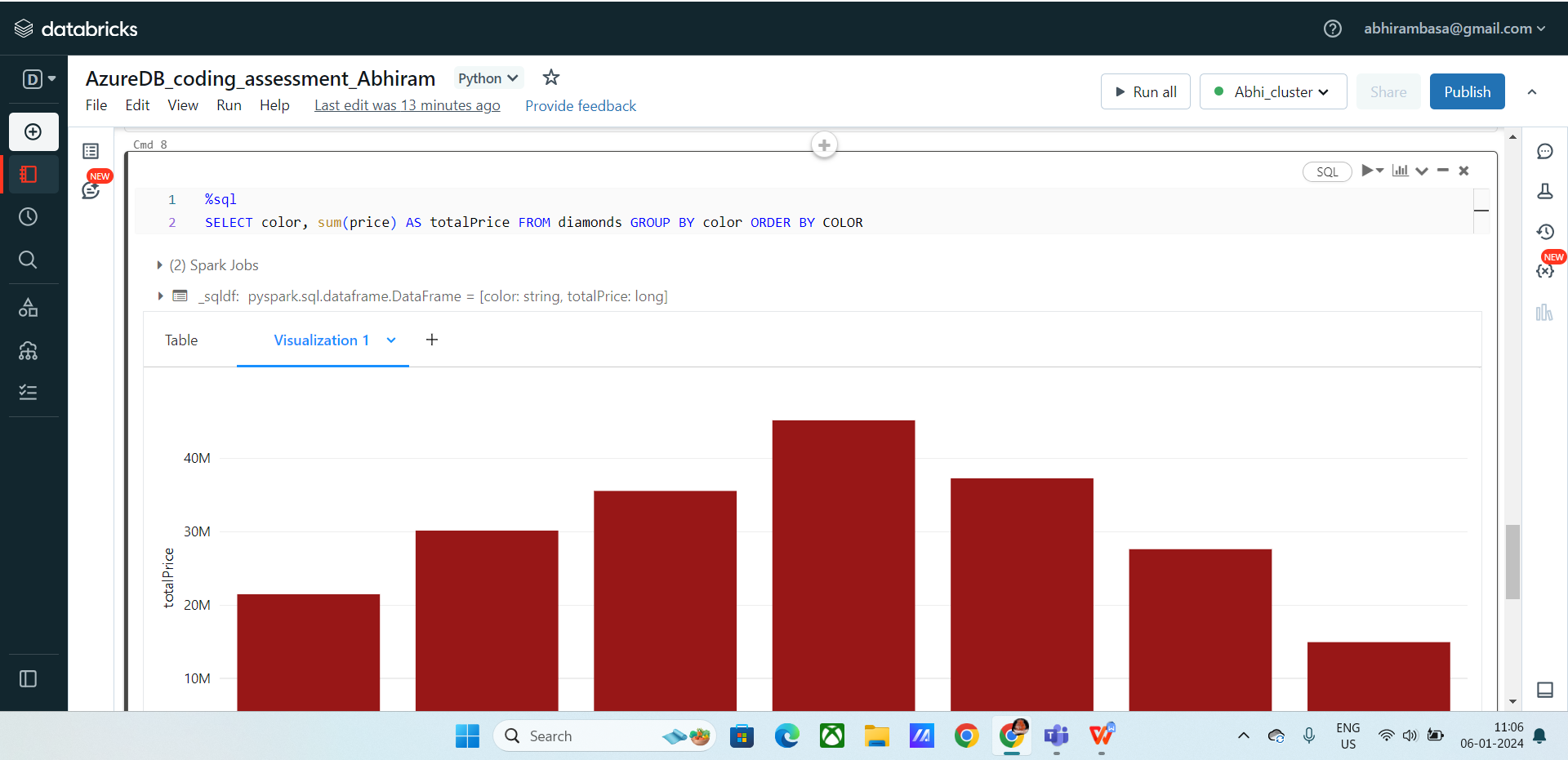


Below I retrieved the data of total prices of diamonds which are grouped by the color, it returns the total price of each color diamonds.

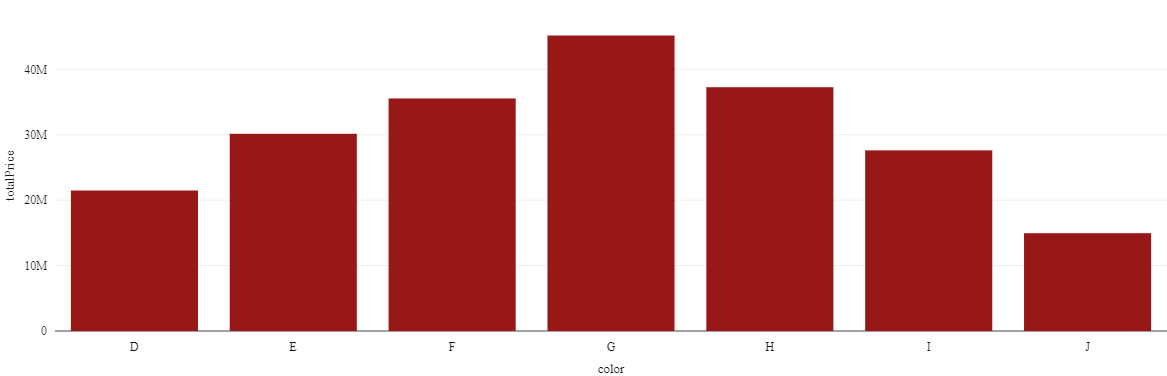


**Creating a visualization for a query:**

* Visualization helps in Analyzing the data or a query in a advanced way.
* In order to Visualize you can use bar graphs, pie charts,etc.
* Click on + icon
* Click on visualization
* Select the columns for X-axis and Y-axis
* Now click on create.

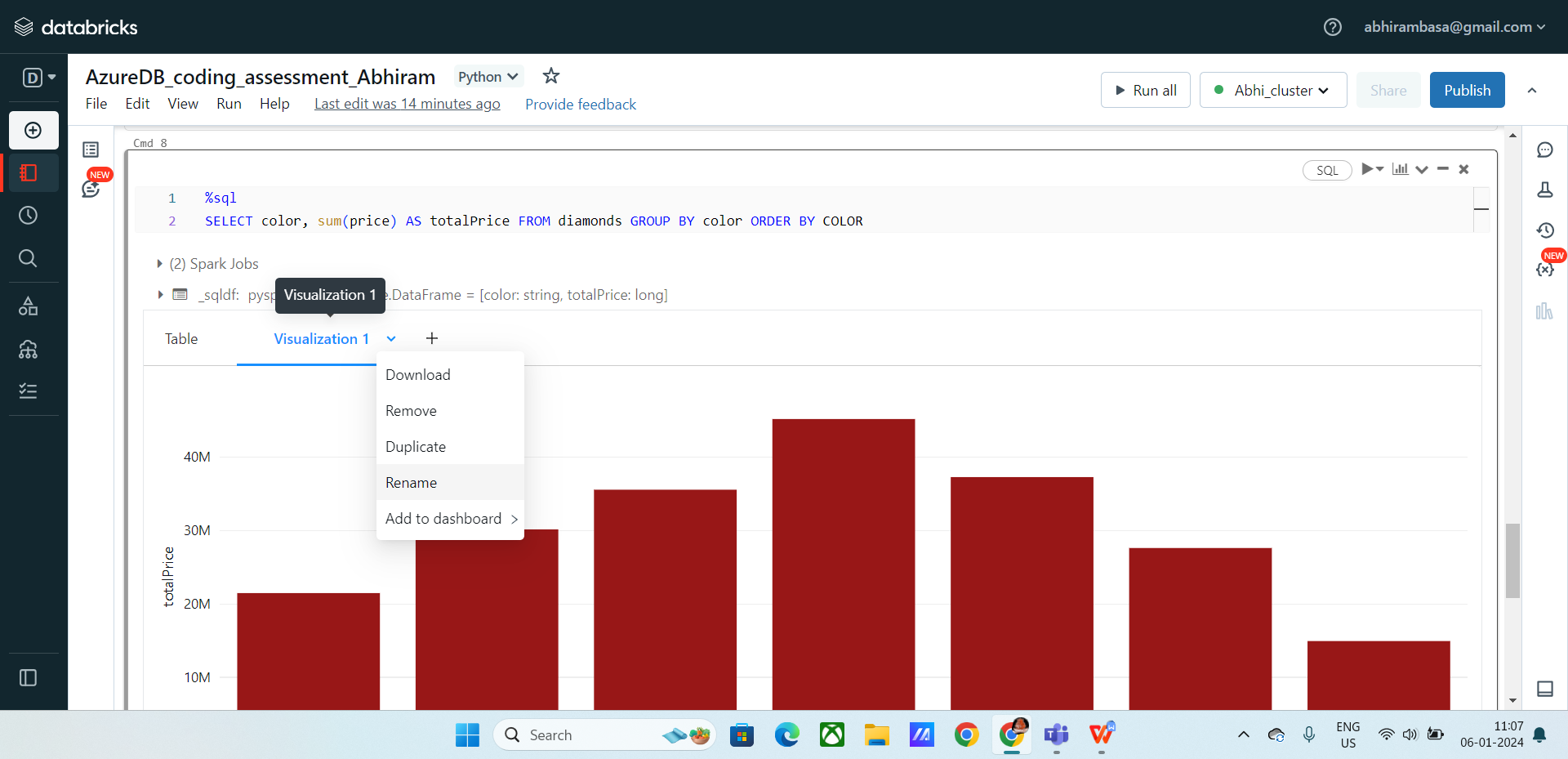


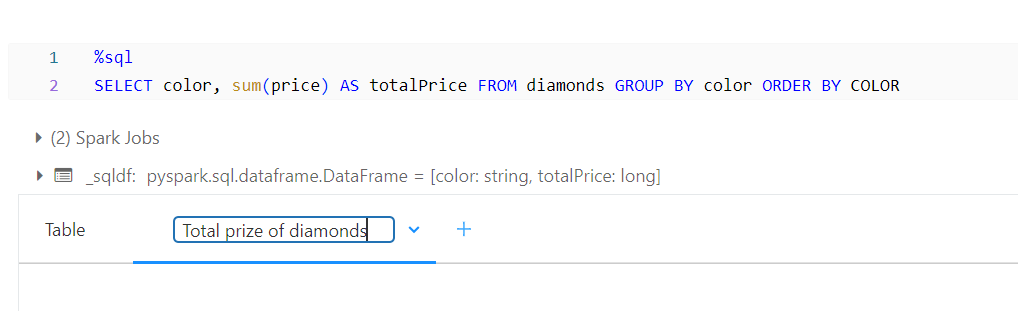
**Visualize output:**



**Rename a visualization:**

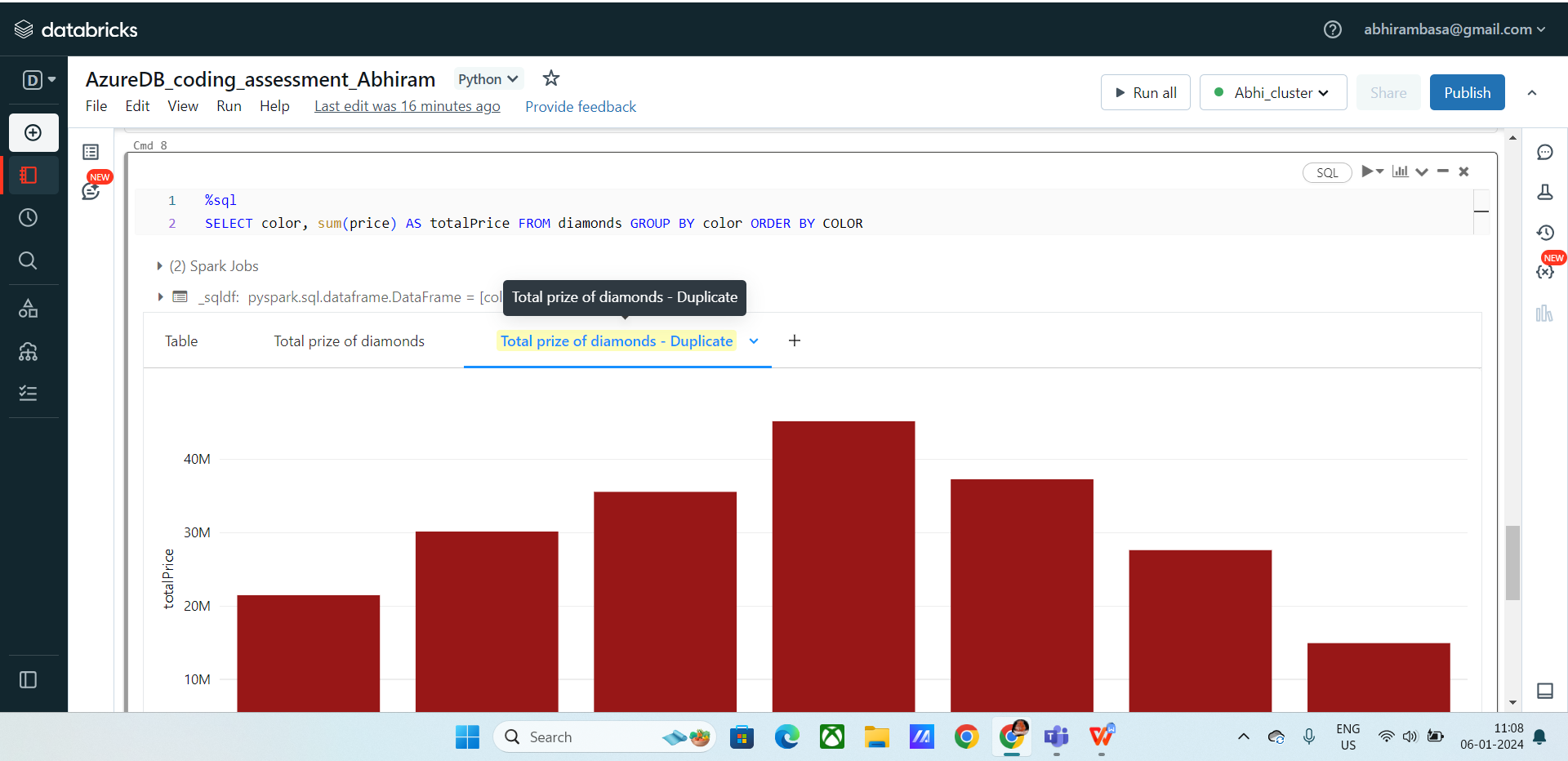
* Click on drop down and click on rename
* Give a name and save





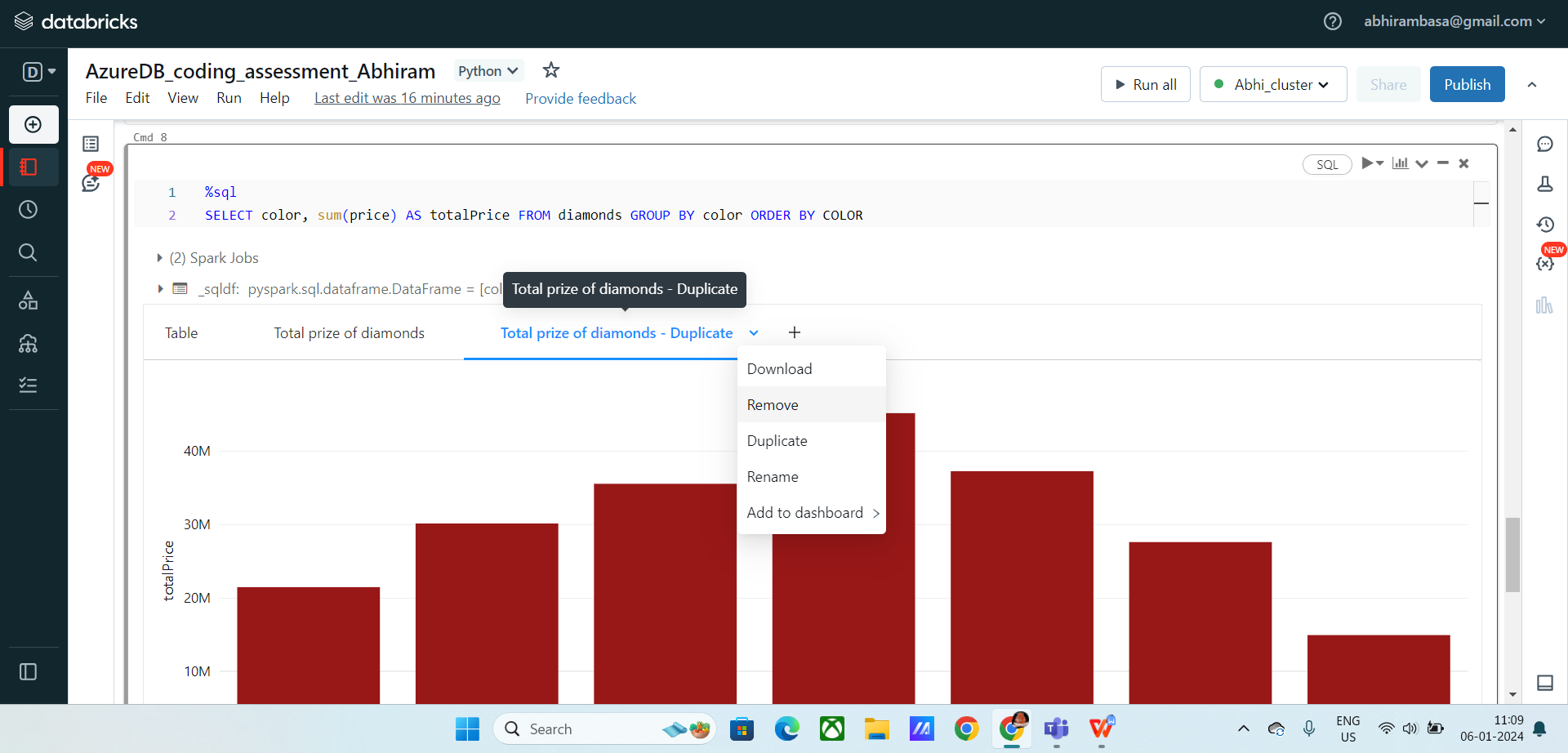
**Duplicate a visualization**

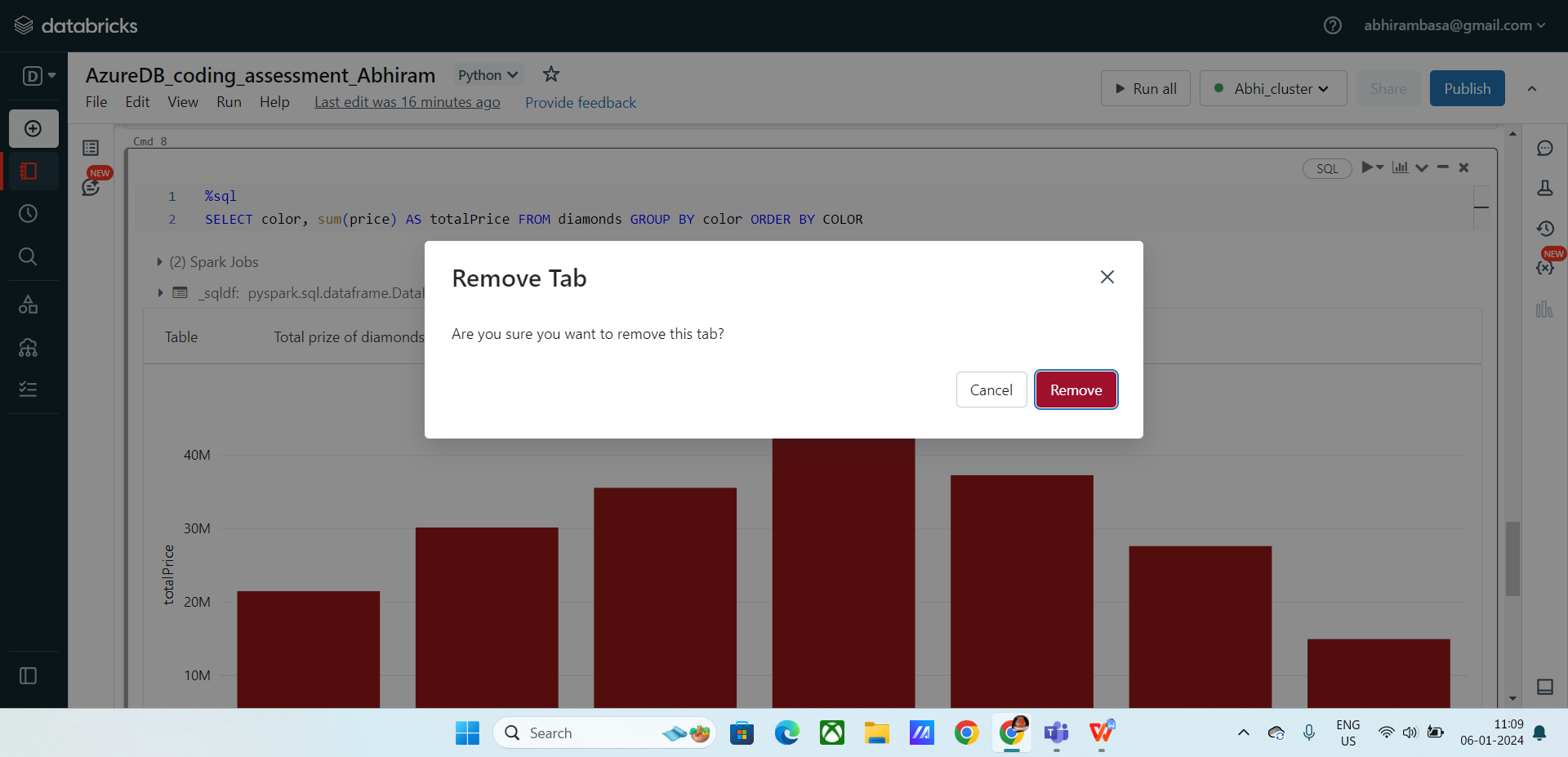
* Click on duplicate from the drop down.
* It will create a copy of the selected visualization.



**Remove a visualization:**

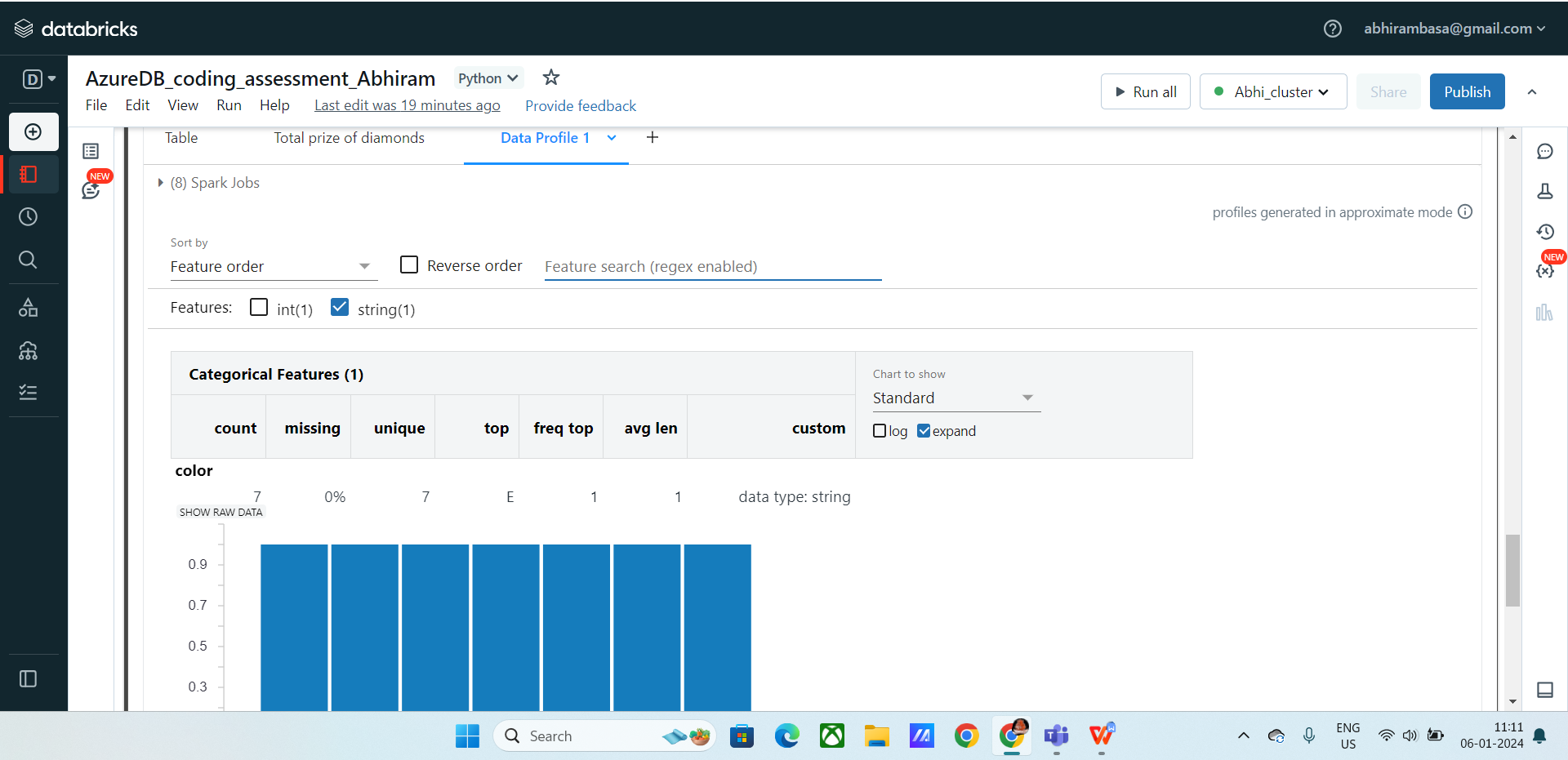
* Click on remove
* It will delete the selected visualization

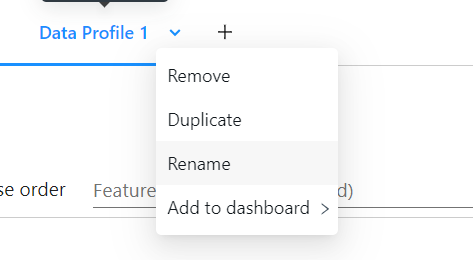




**Creating a data profile**

**Data profile :** Used to analyse the data based on some trends, you can also order and filter the data.





**Answer 2 :** Explain the copy activity in Azure data factory.

**Azure data factory :** It is the cloud-based ETL and data integration service that allows you to create data-driven work flows.

It allows ingestion of data, control the pipelines, control the data flow, schedule and maintain the pipelines.

**Ingestion**: Importing the data from resources to selected destinations, like data warehouse, aure blob storages, Sql databases, etc.

**Pipelines**: These are logical groupings of activities that perform a unit of work.

* A pipeline allows you to manage activities as a set and chain them together to achieve specific data processing goals.
* A pipeline work can be ingestion of data from data resources and exporting them to data warehouses.
* Pipelines are very helpful for automation tasks.

**Steps to copy activity in Azure data factory**

* First create data factory with your subscription.
* Select your resource group.
* After creating, you will be redirected to home page of Azure Data Factory.
* Now the destination is created

**Creating source path:**

* Go to storage account, create an account.
* Now create a storage container, click on create
* Create Azure blob storage, with the default fields
* Now test the connection, after the connection is successful
* Click on create
* Now create a folder in azure blob storage.
* Now click on the azure blob new connection, we need to specify the source path and destination path.
* Choose azure blob storage in the connection section
* Now browse the source path and select the file you want to copy.
* Click on binary copy and click on next.

**Creating destination path :**

* Click on destination path, give the path of the folder that we created in storage account.
* Give a new file for the output to store.
* Now create a task to copy.
* After executing the above task.

**Checking the output:**

* Click on the pipelines runs, you can find the job after successful completion.
* Now go to the destination folder and check the copied file.

Hence , the process of copying data using Azure data factory is complete.