Coding challenge

Name : Chakradhar Bhogapurapu

Date : 06/01/2024

Mail : [chakribhogapurapu@gmail.com](mailto:chakribhogapurapu@gmail.com)

**Question 1**

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

**Answer**

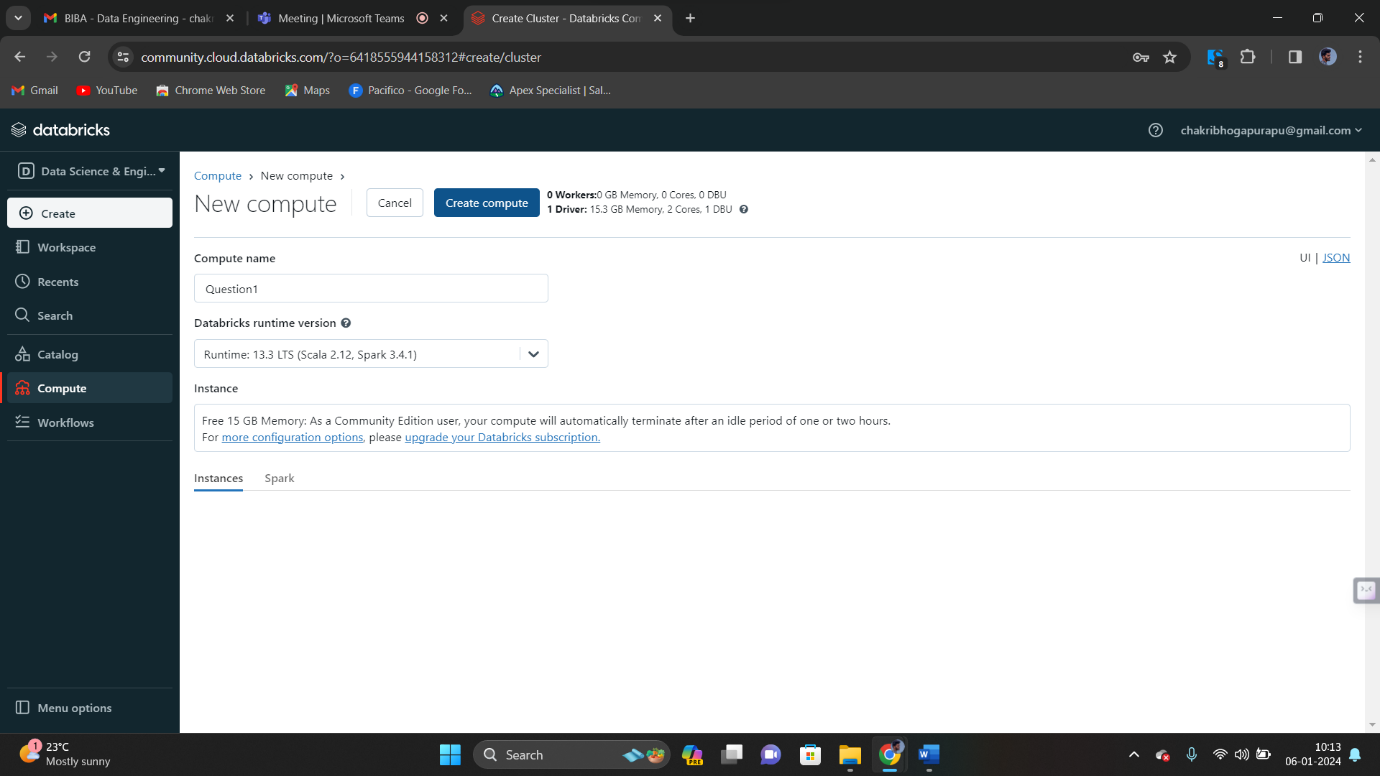
🡪Creating a cluster and attach the notebook to the cluster

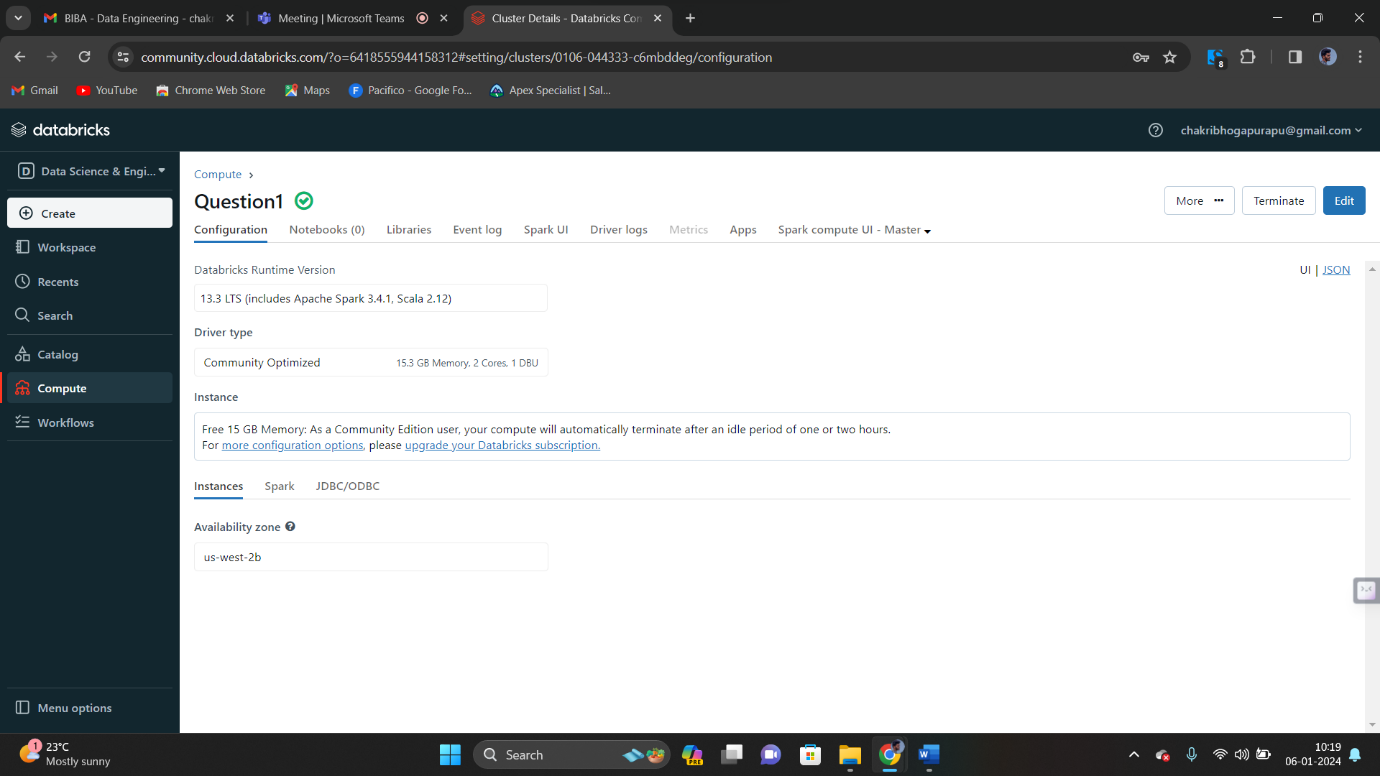
In azure databricks community addition, create a cluster

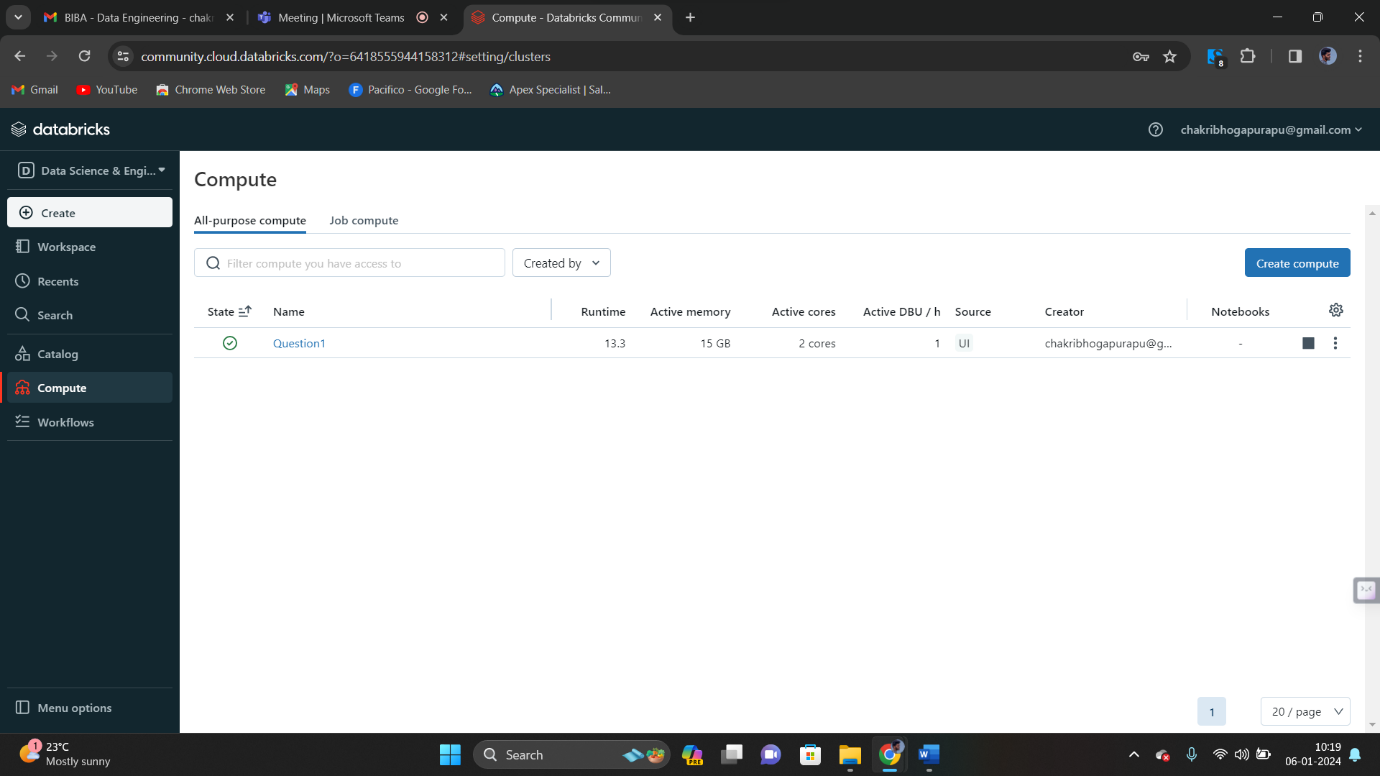
Give the cluster name as “Question1”

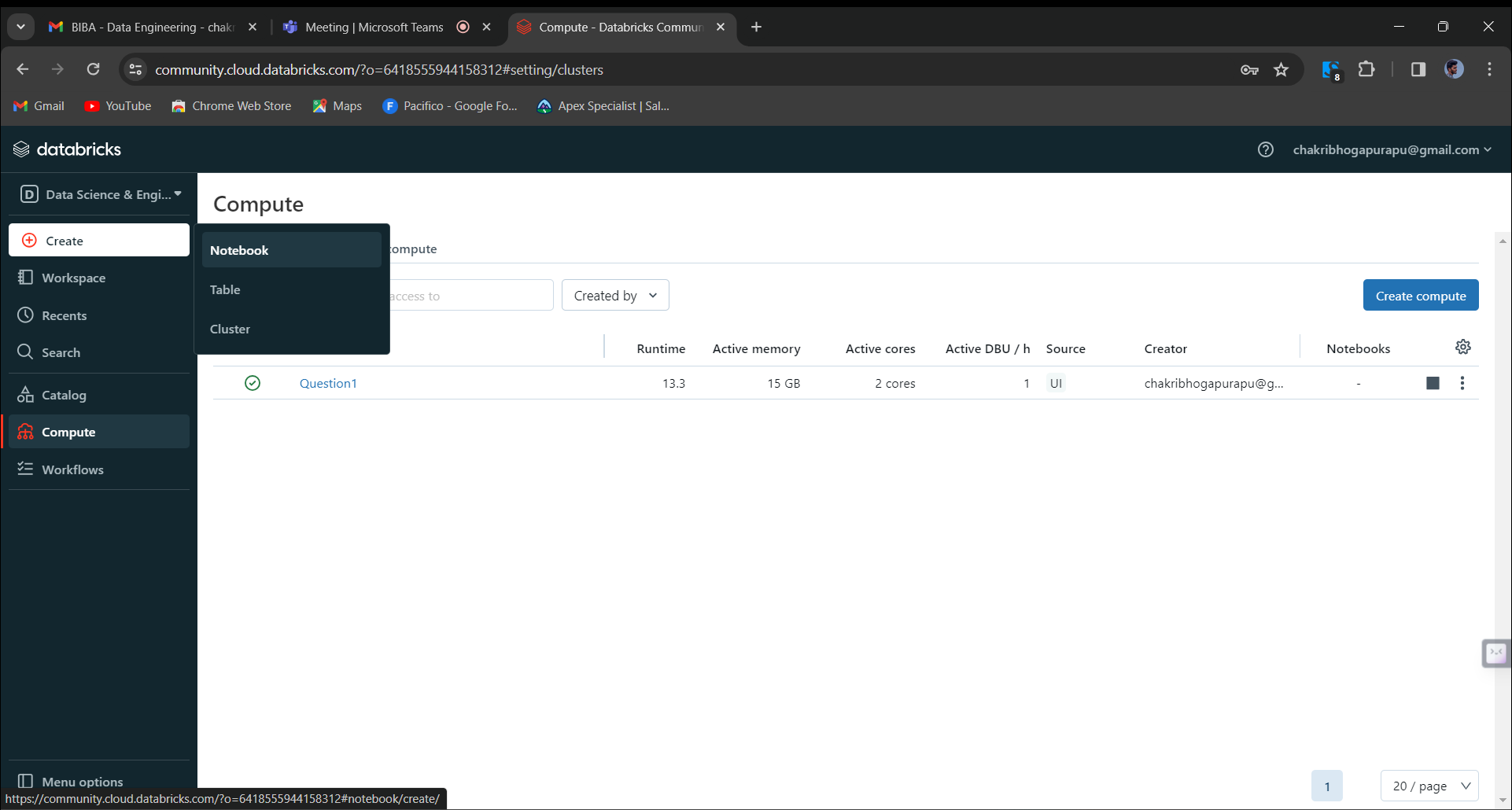
Give databricks runtime version as 13.3 LTS.

Click on create compute



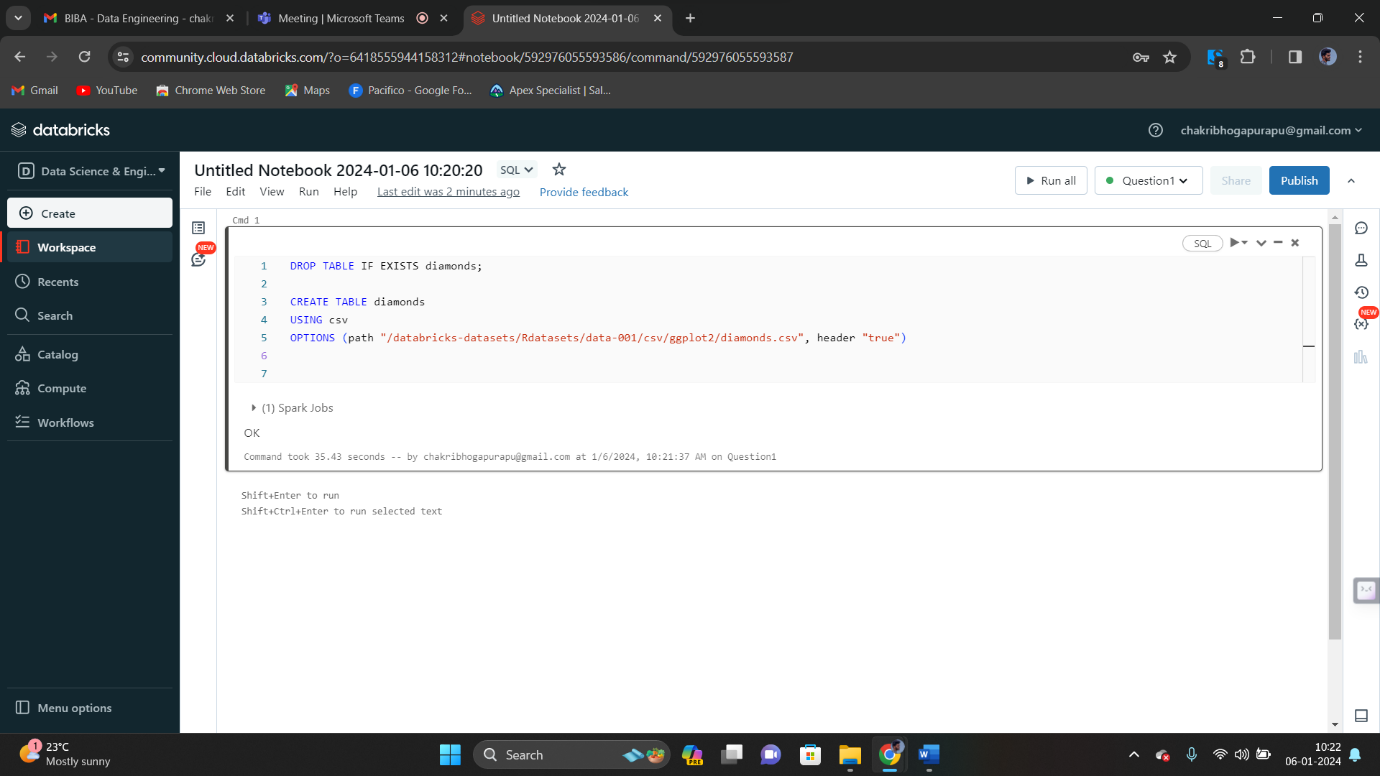
Our cluster is created as shown below. 

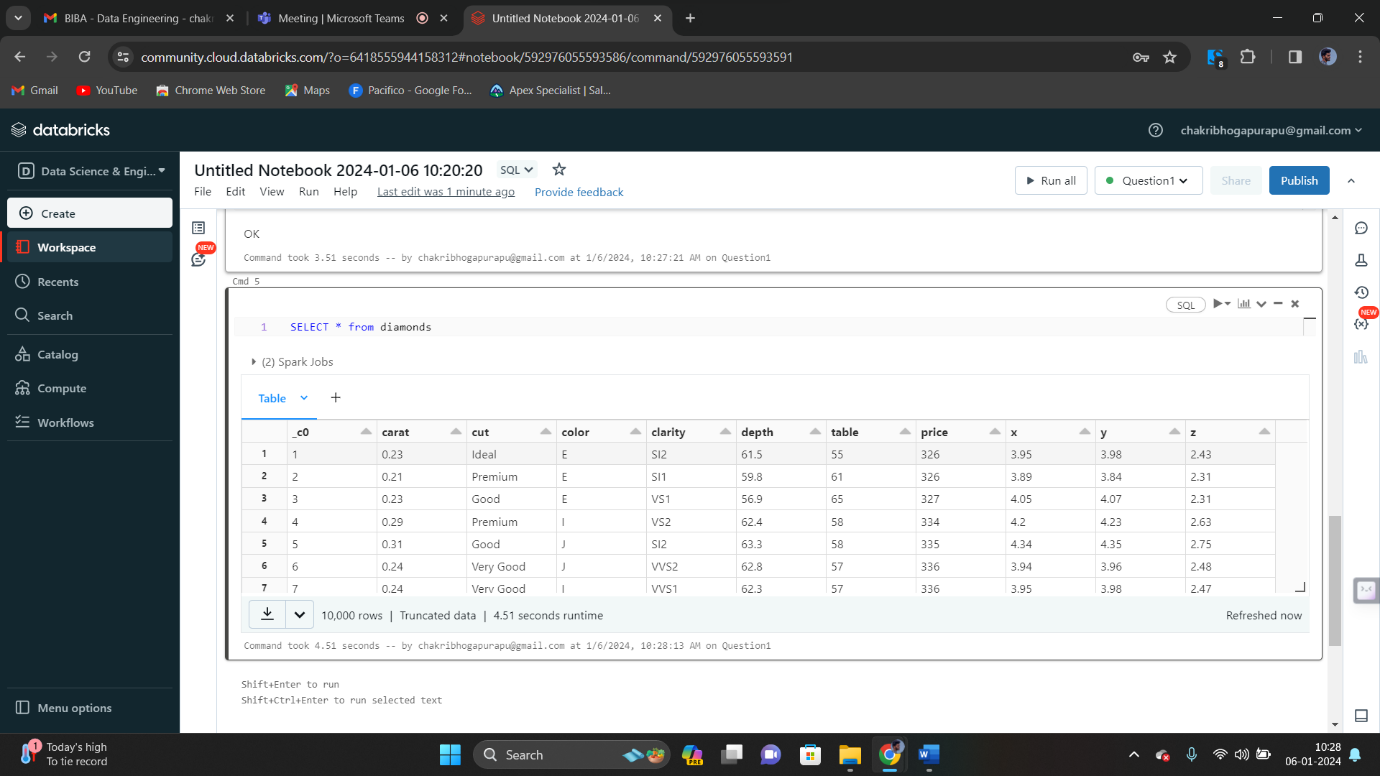
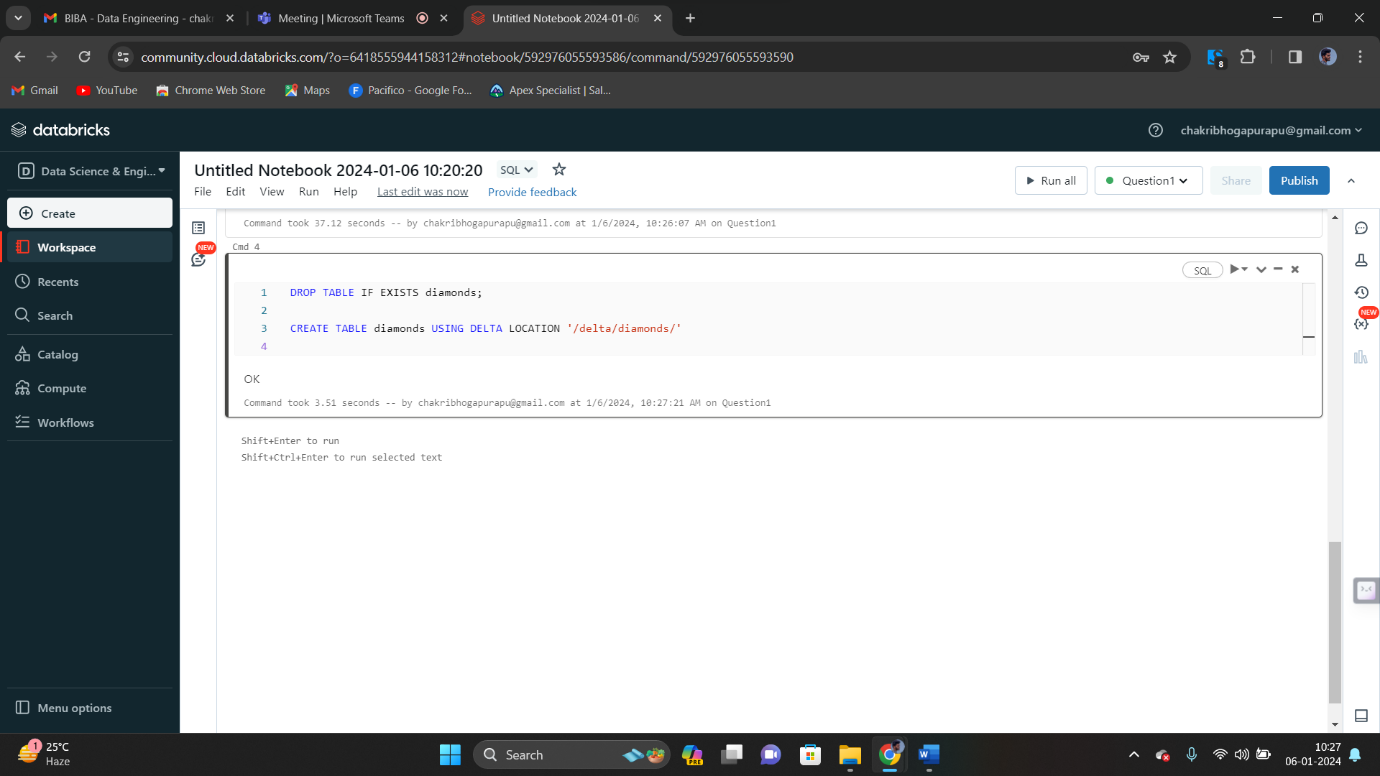
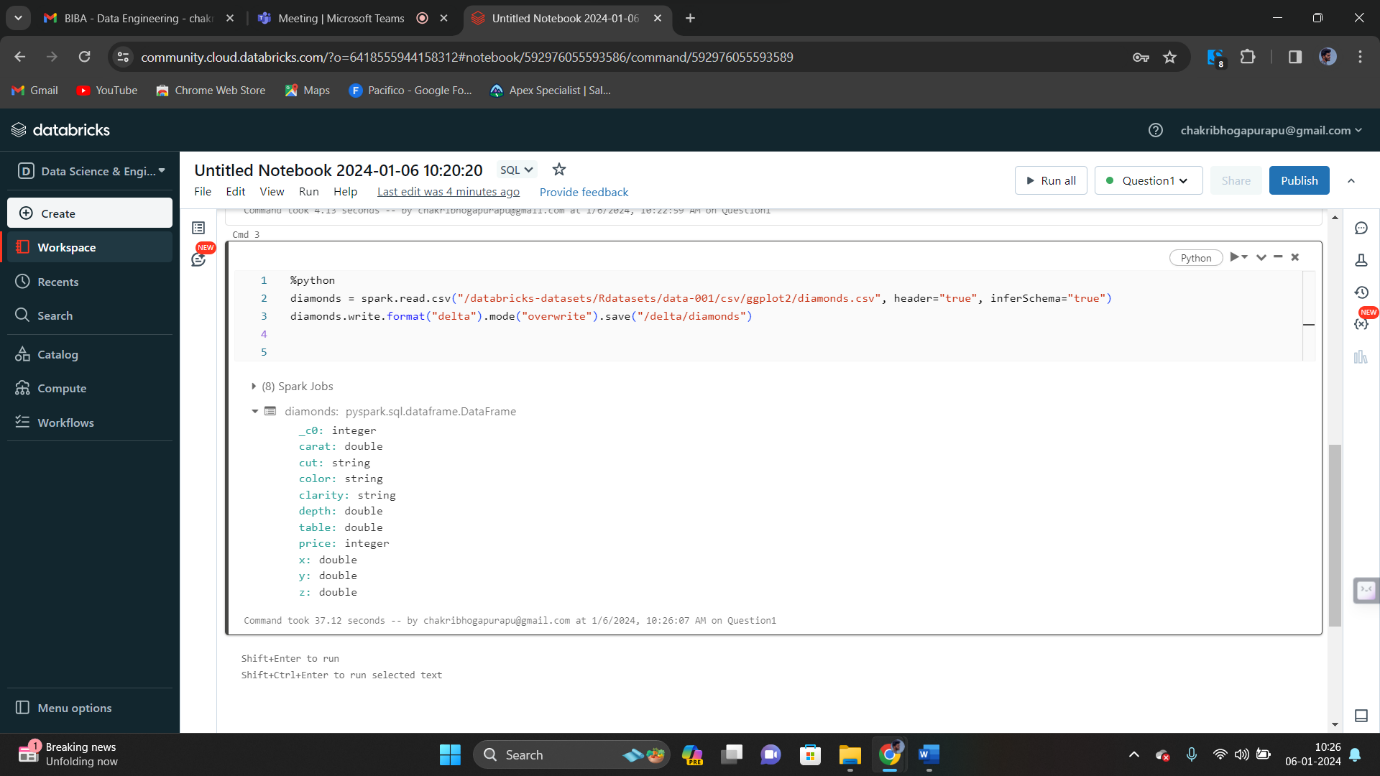
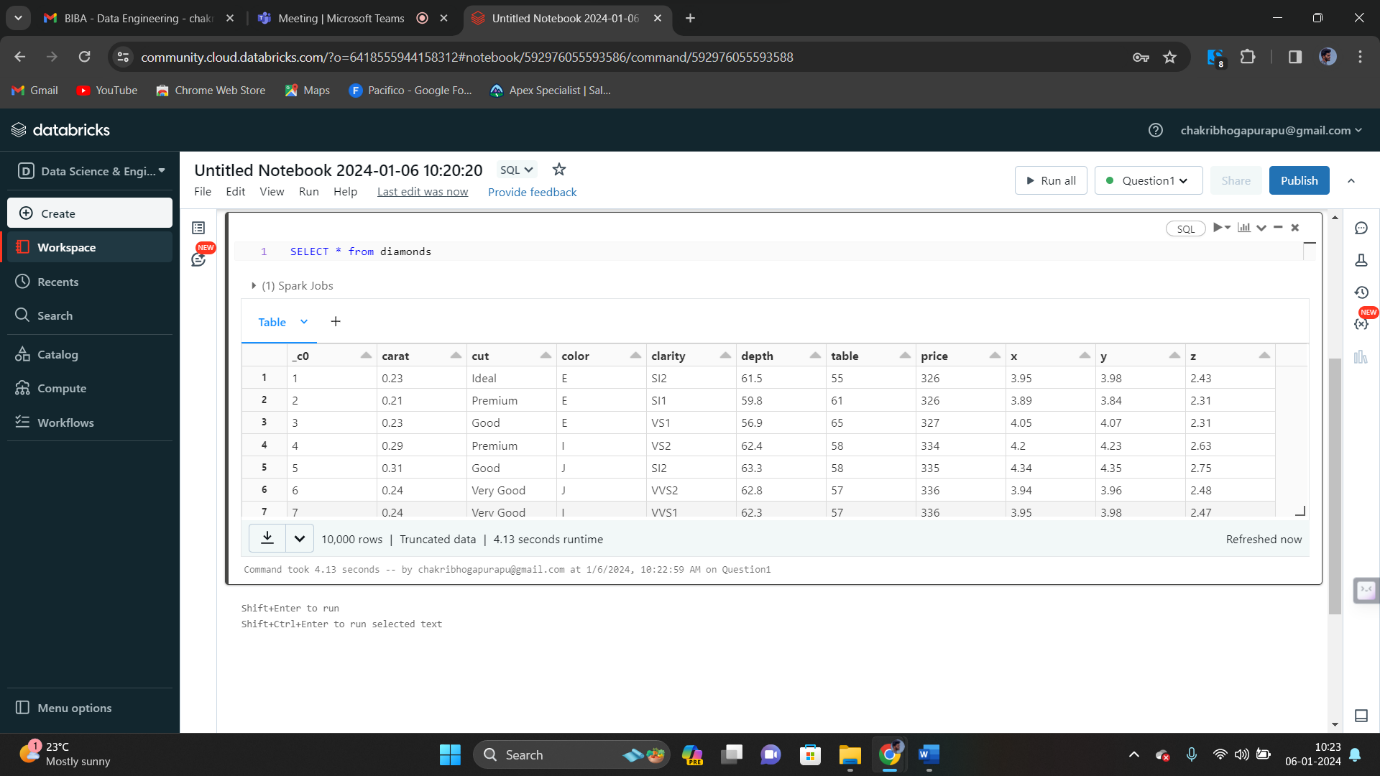
Click on our cluster

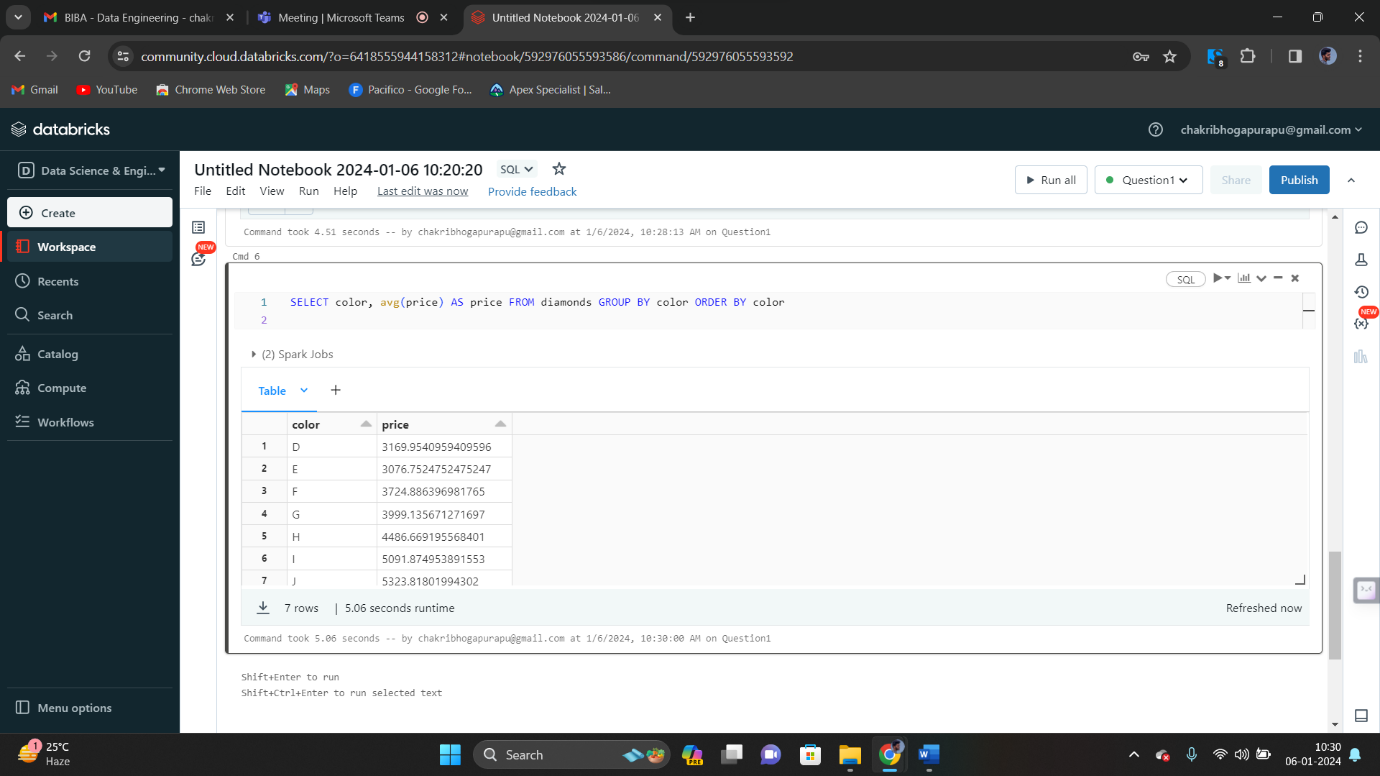
Click on crate button. Now create a notebook. We can give name for our notebook. If not we can leave like that. 

Run the command to fetch the data from the csv file named as “diamonds.csv”

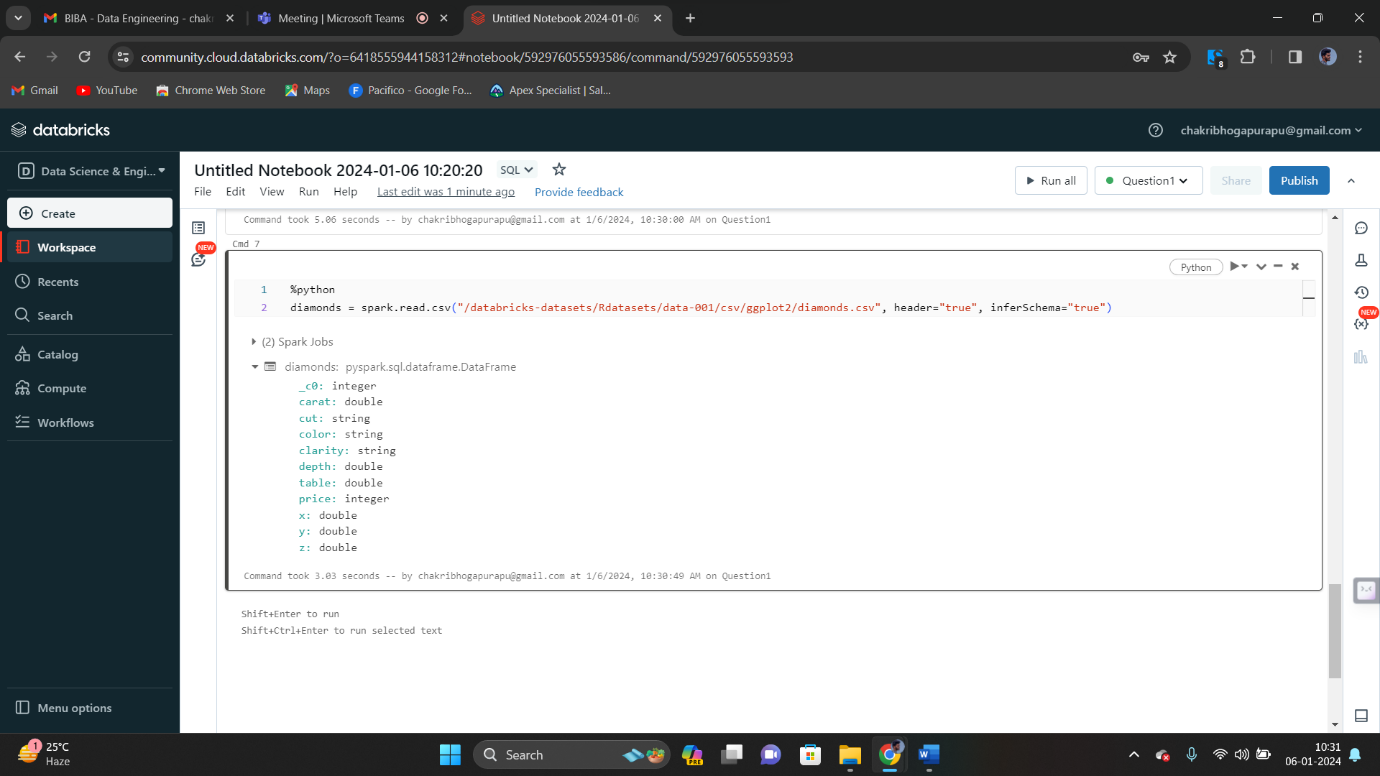
Provide the path in the command to fetch the data.

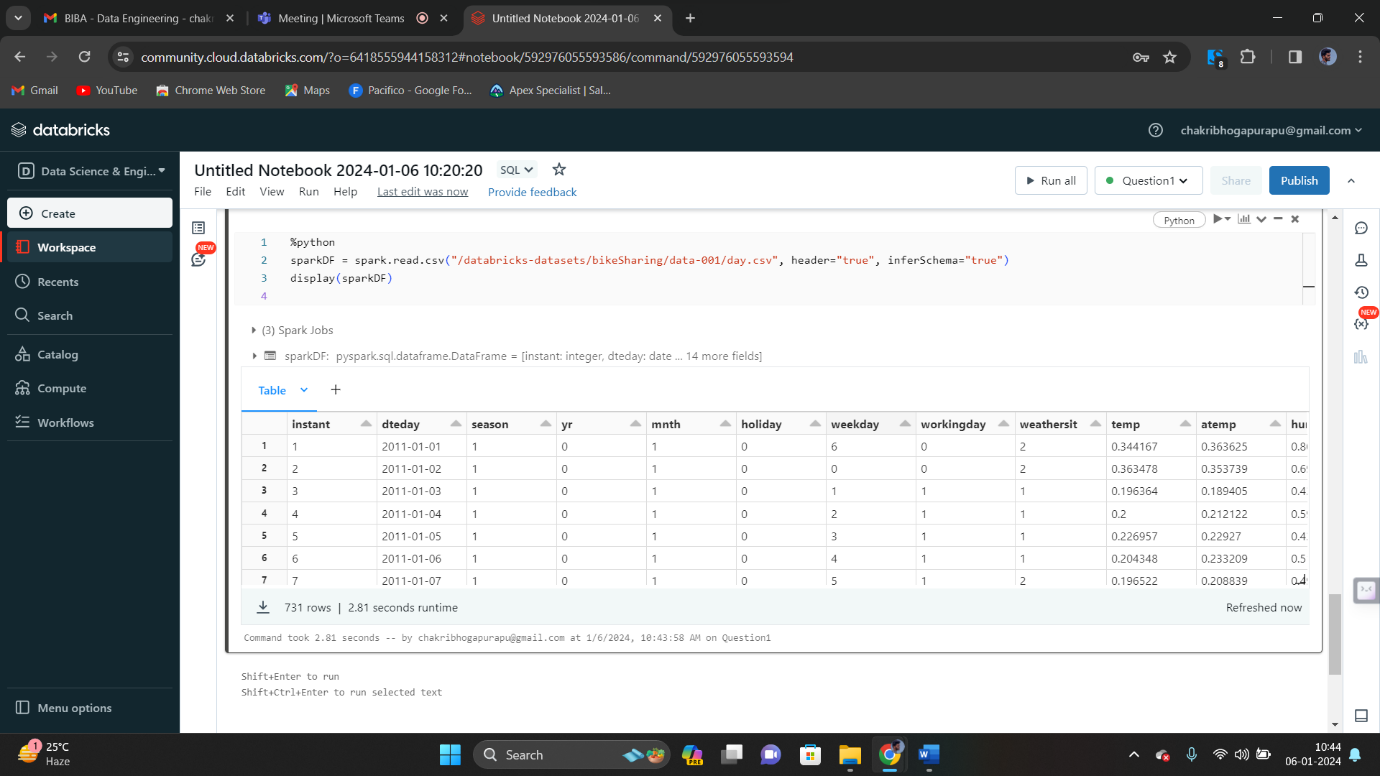




Manipulate data and display the result.

Creating a data frame from a data bricks dataset.

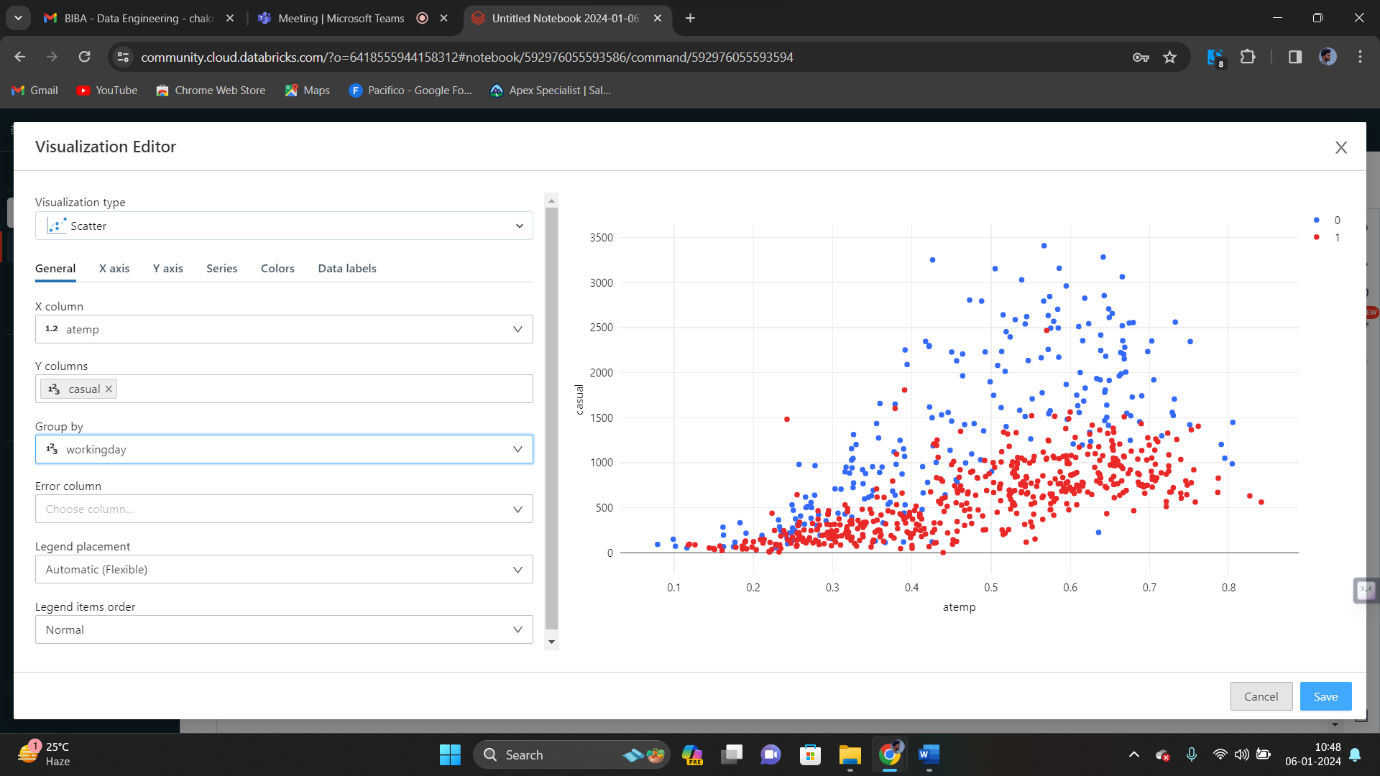


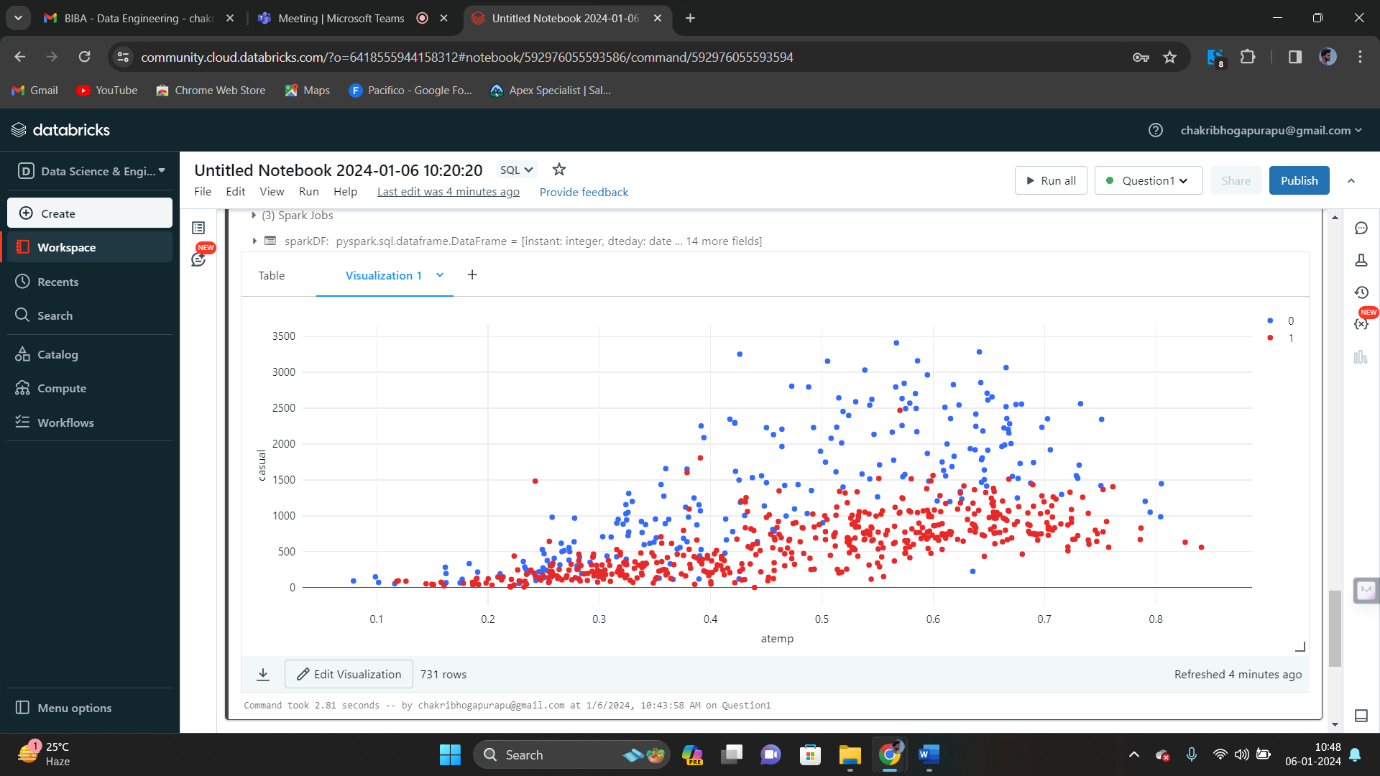
🡪Create a visualization

Now click on “+” and click on visualization

In visualization, select visualization type as “Scatter”.

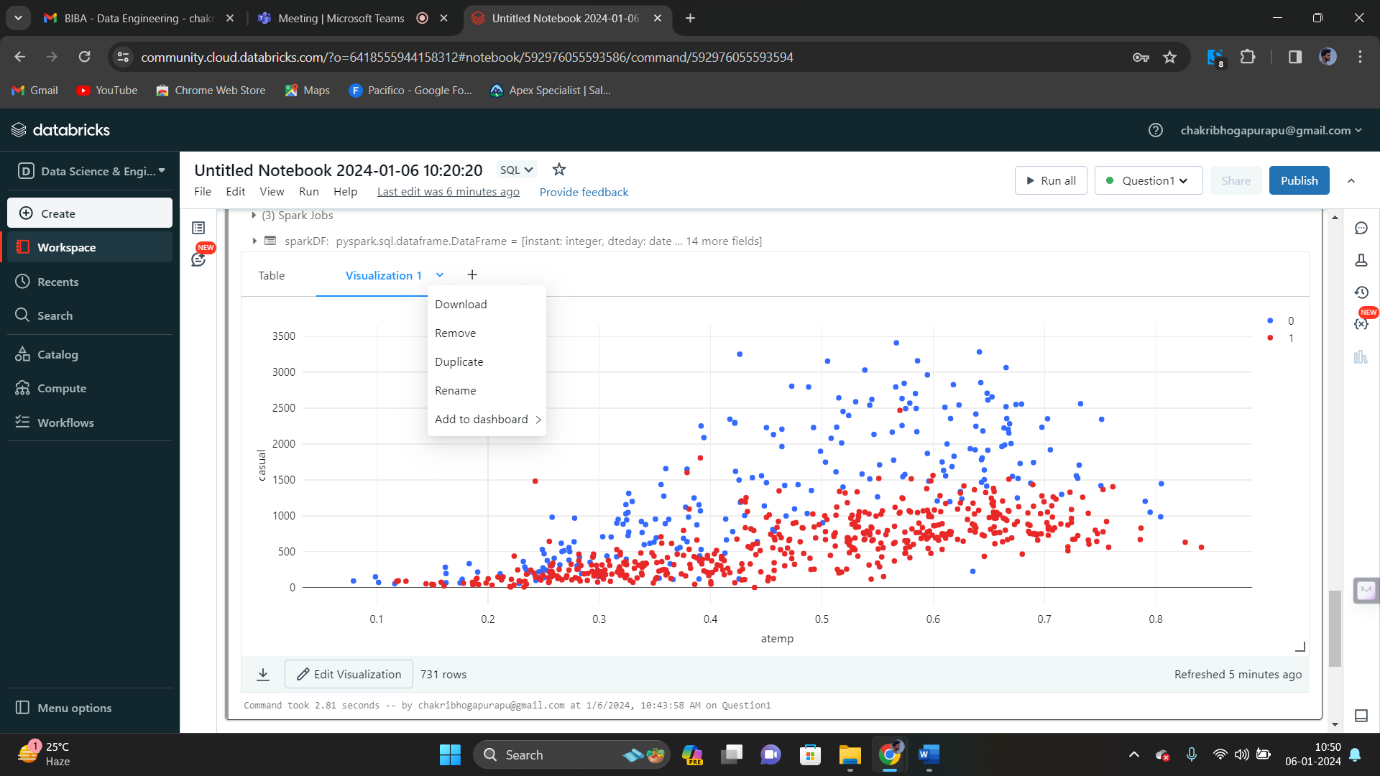
Select x column as “atemp” and y column as “casual” and groupby as “workingday”

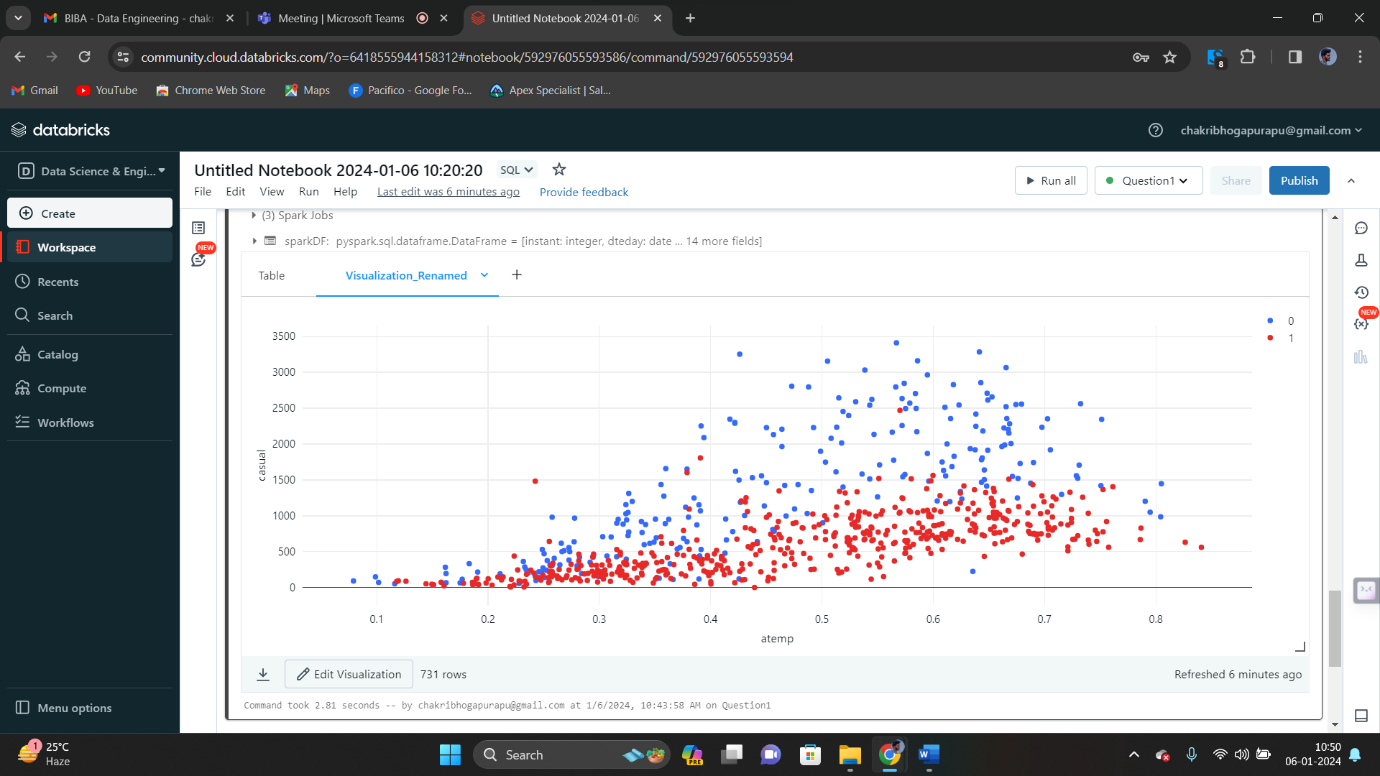
Now click on save. 

We get output as shown below

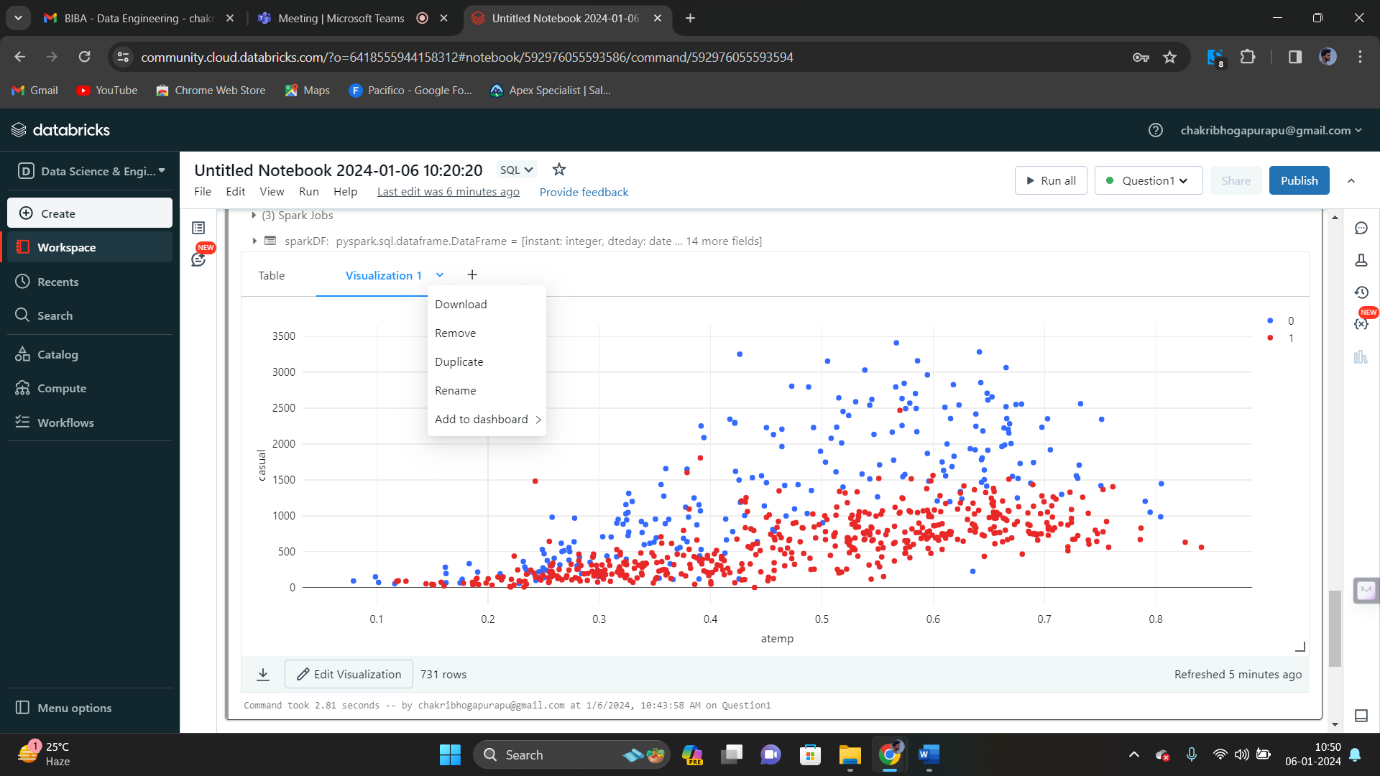
🡪Rename the visualization

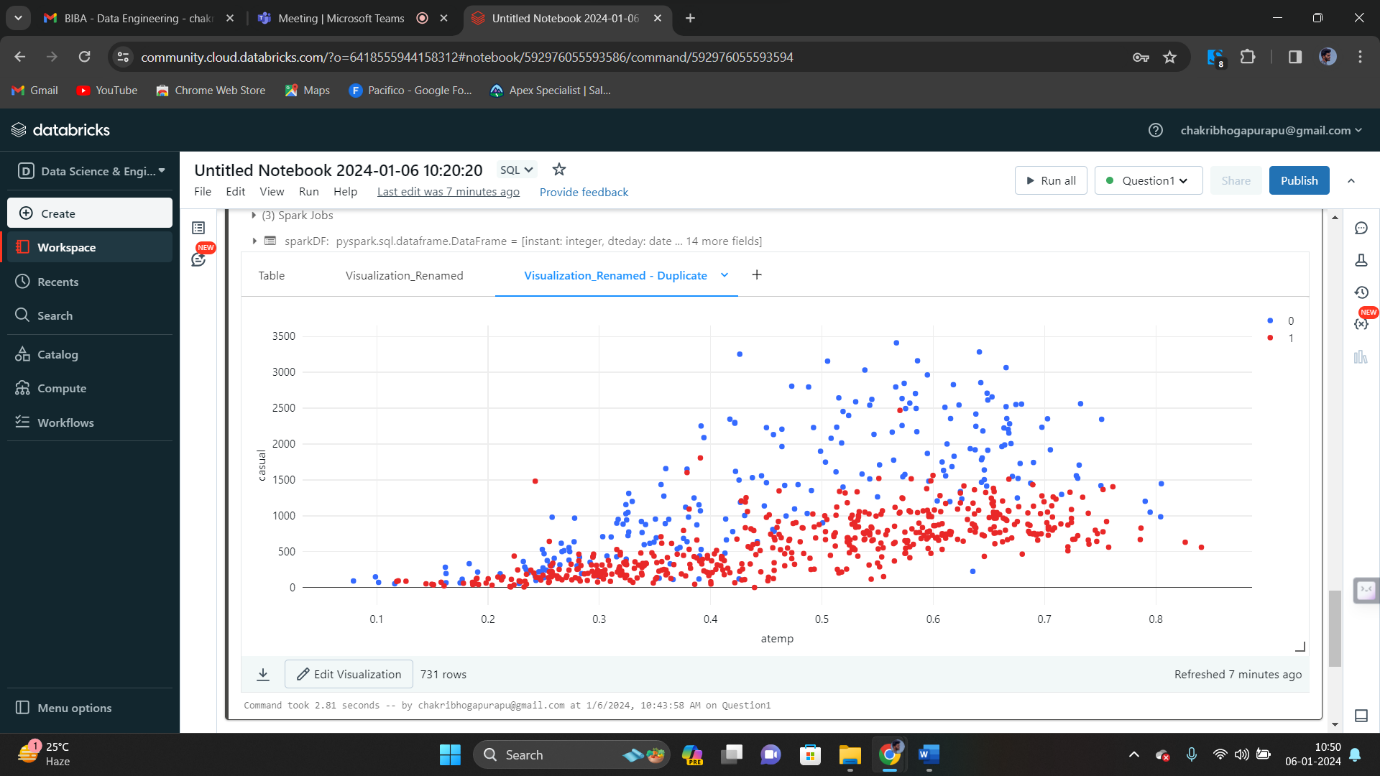
To rename the visualization, click on down arrow. Now there is an option called as “Rename”. Now click on it



Rename “Visualization 1” as “Visualization\_Renamed”

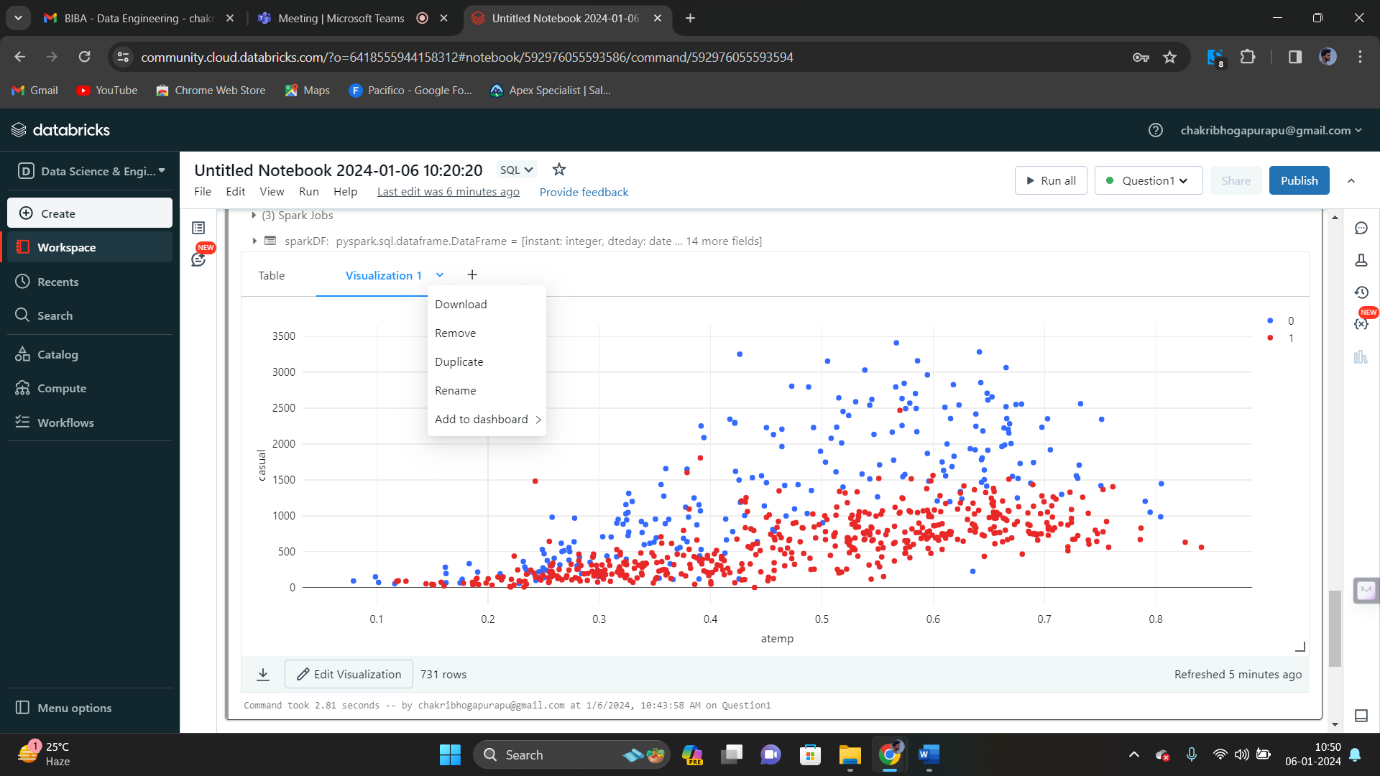
🡪Duplicate

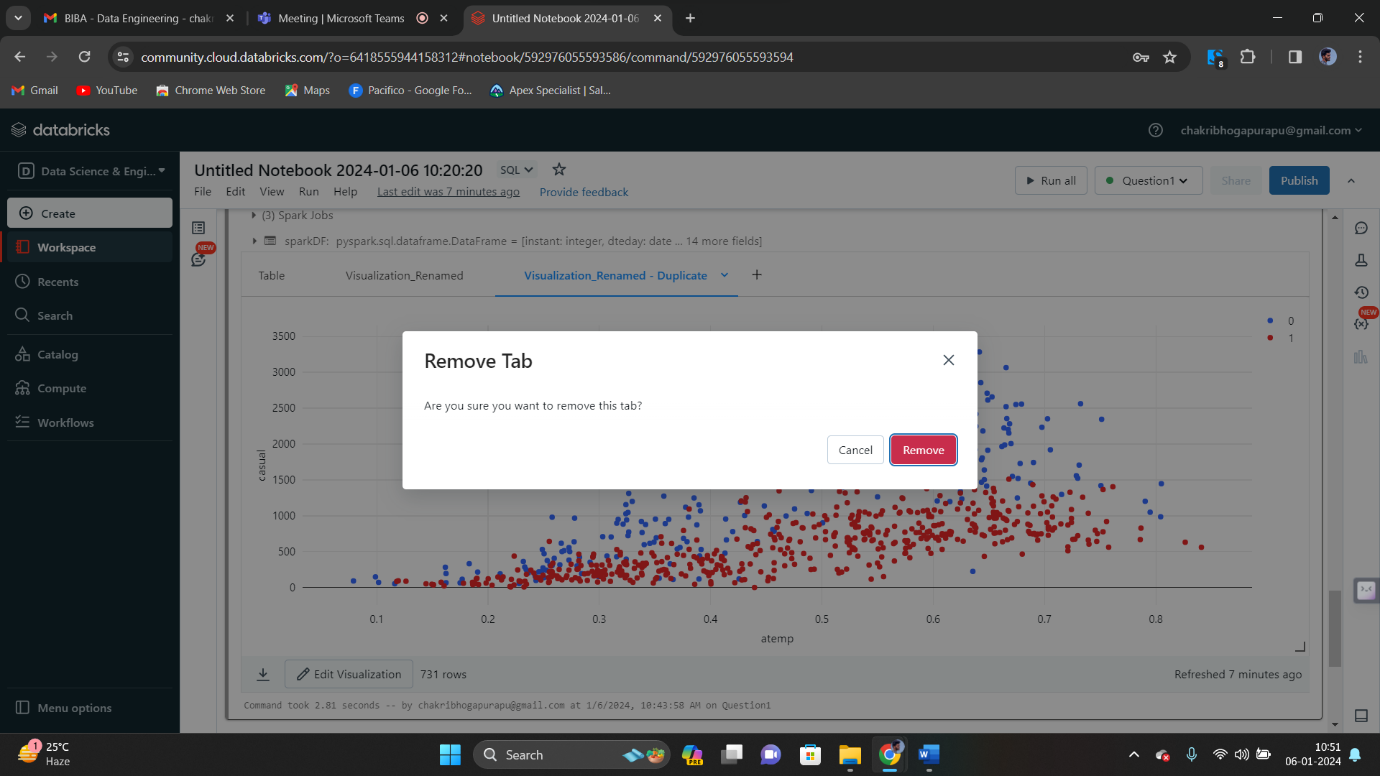
Now click on dropdown arrow. We can see the option called “Duplicate”. Click on it.

Now we get the duplicated visualization as shown below. 

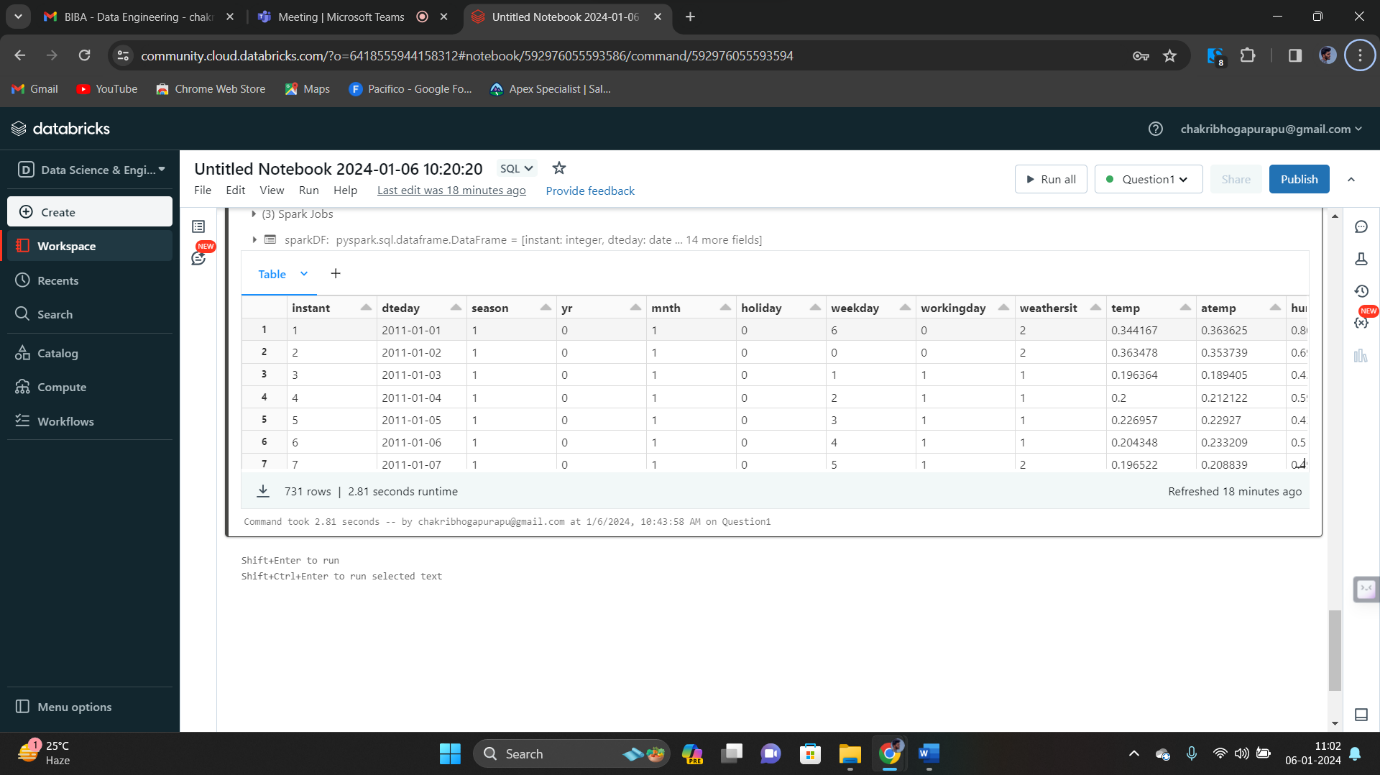
🡪Remove

Again click on the dropdown arrow. Now click on “Remove”.



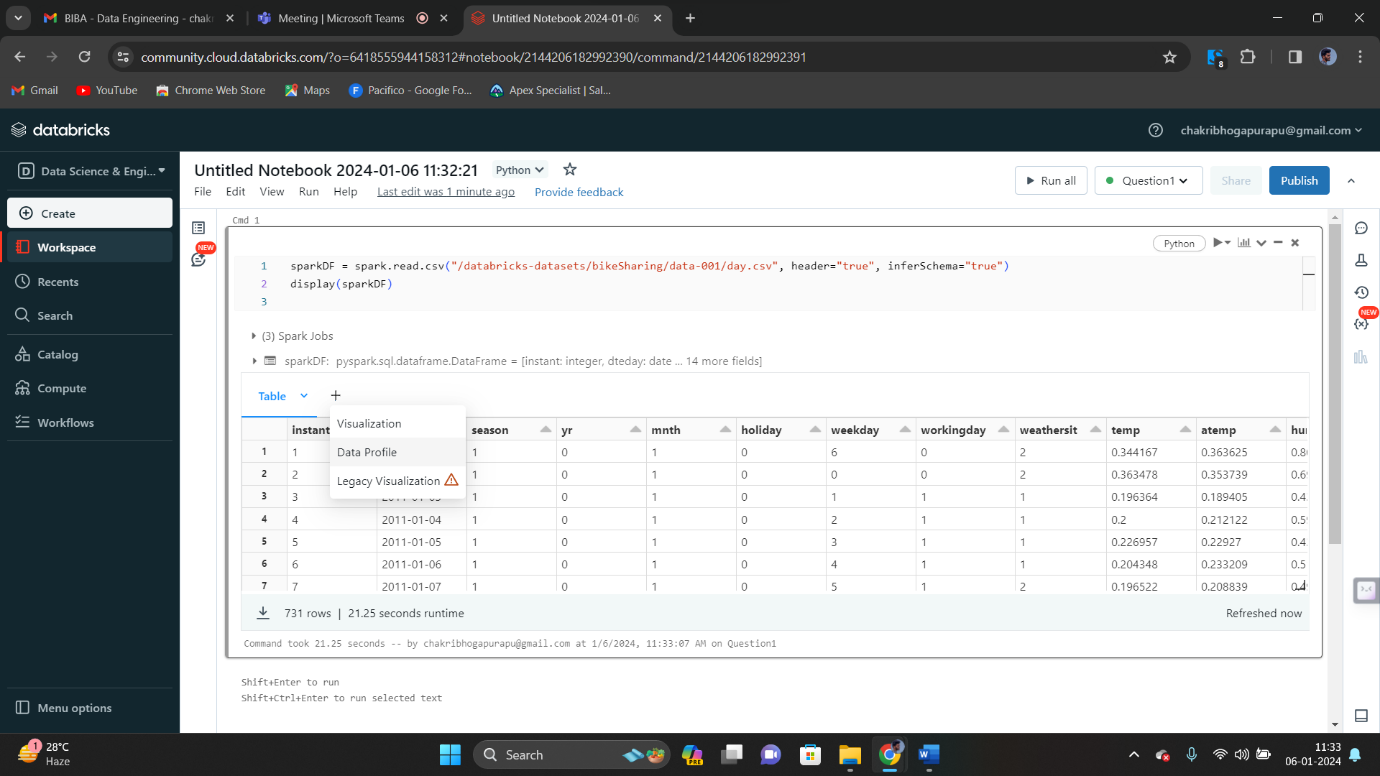
Now we get a pop up to remove the “visualization”. 

Now the visualization is removed.

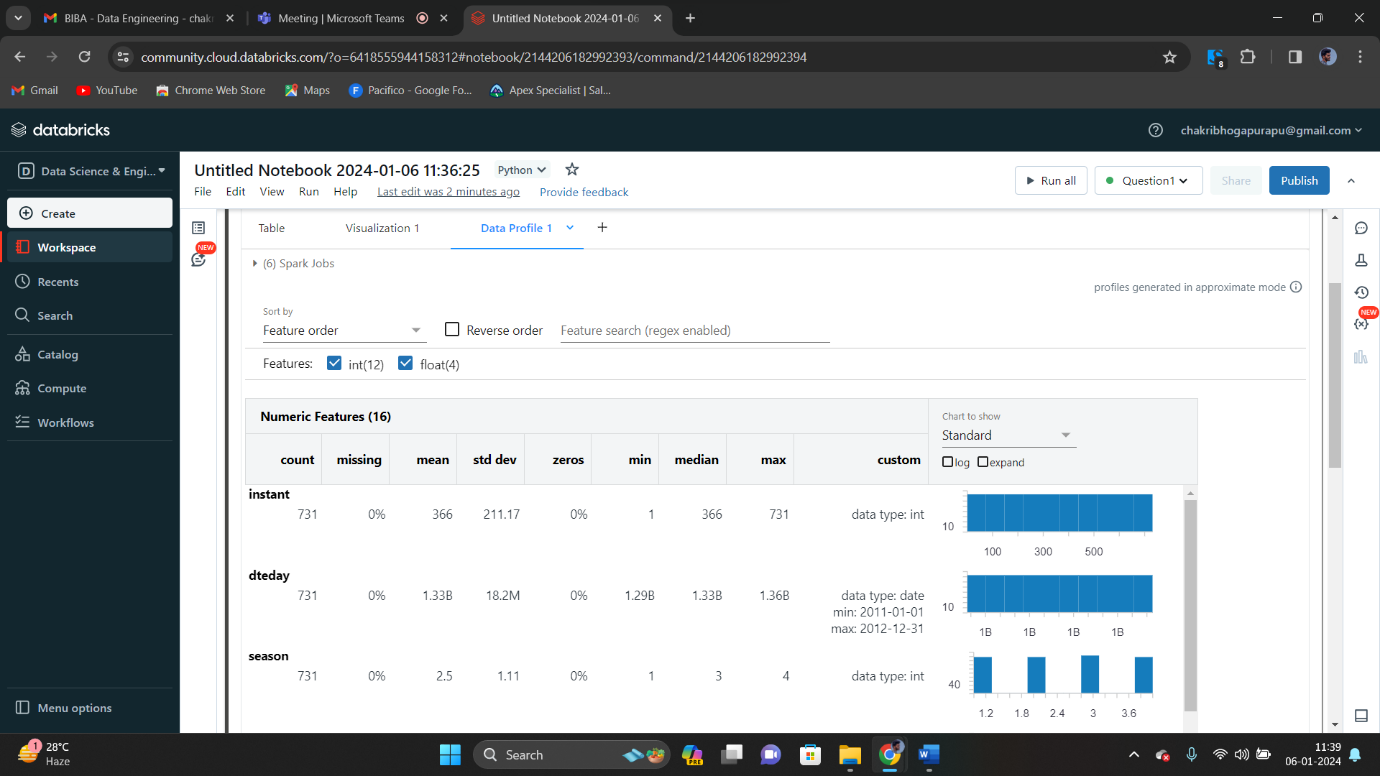


From the above image, we can see that there is no visualization after removing it.

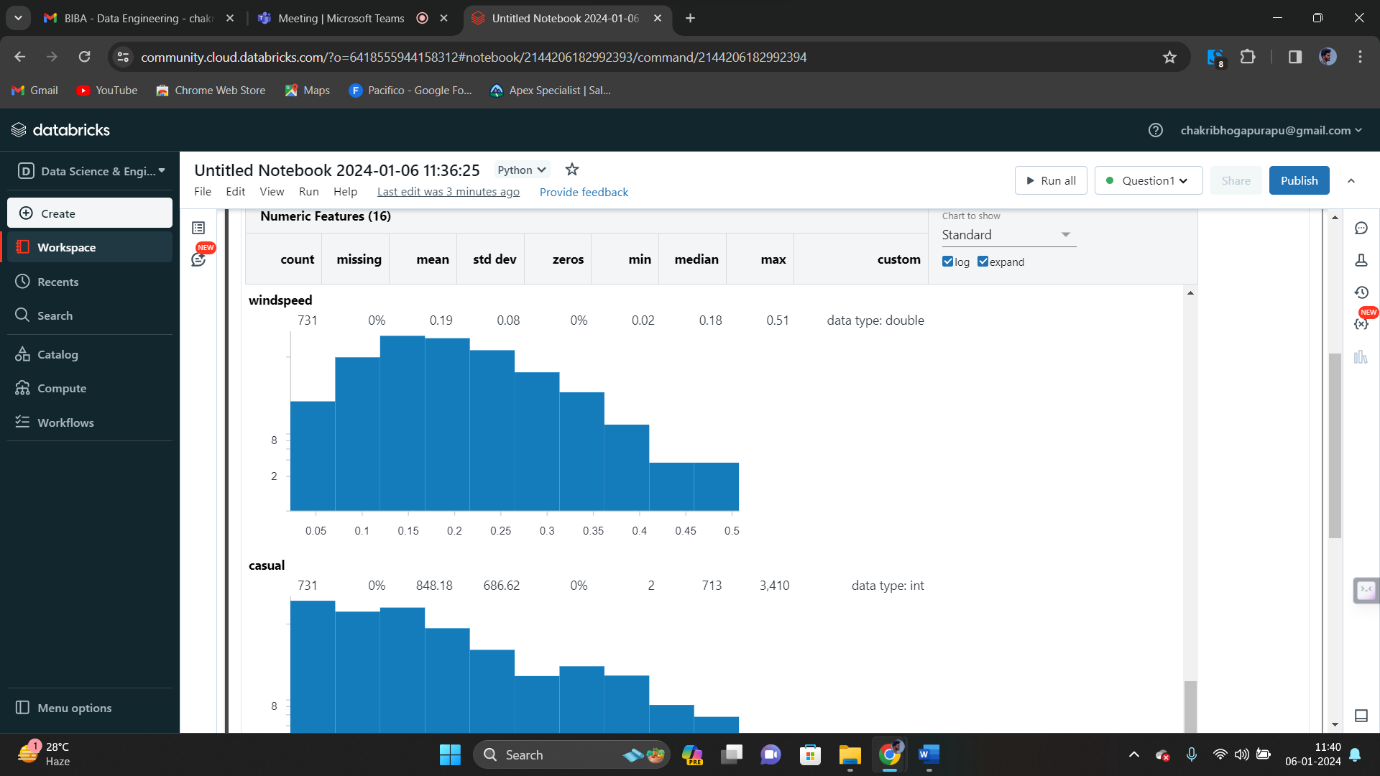
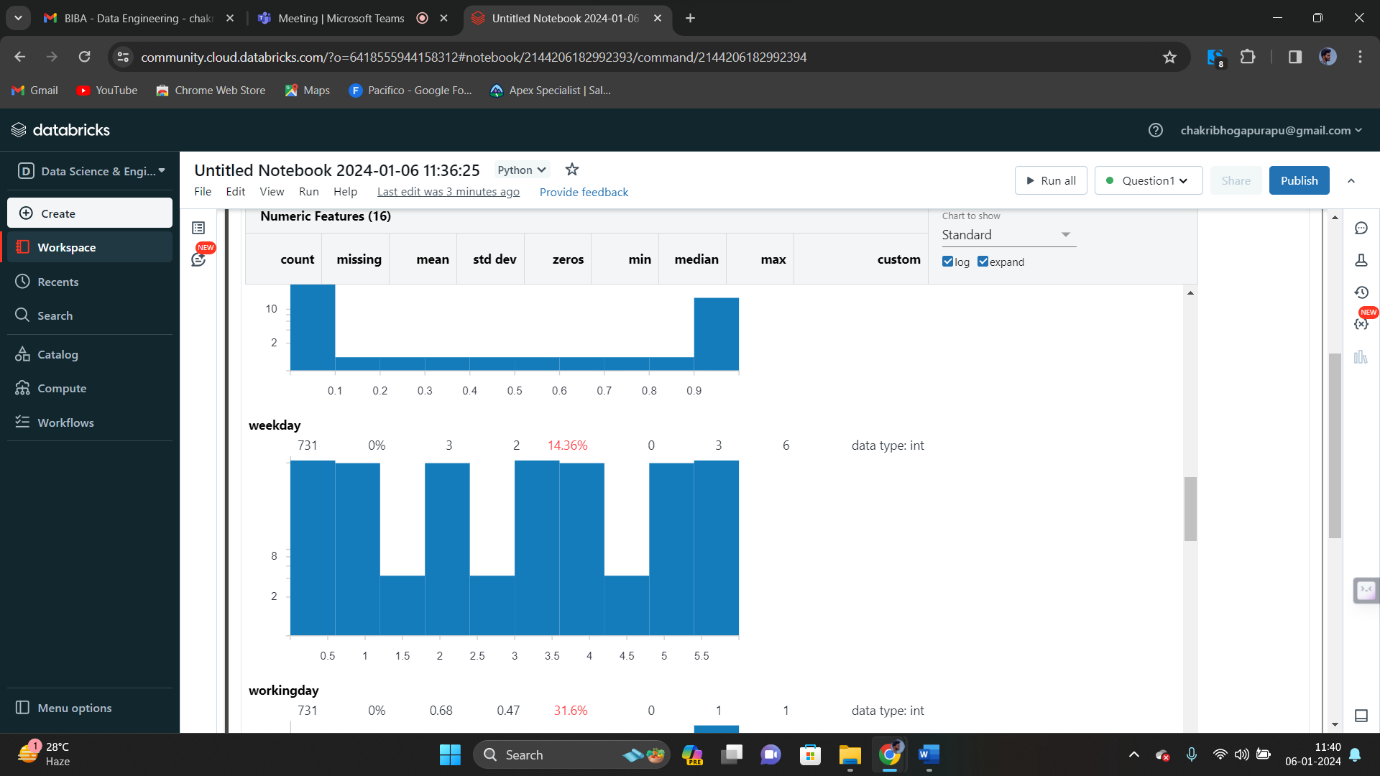
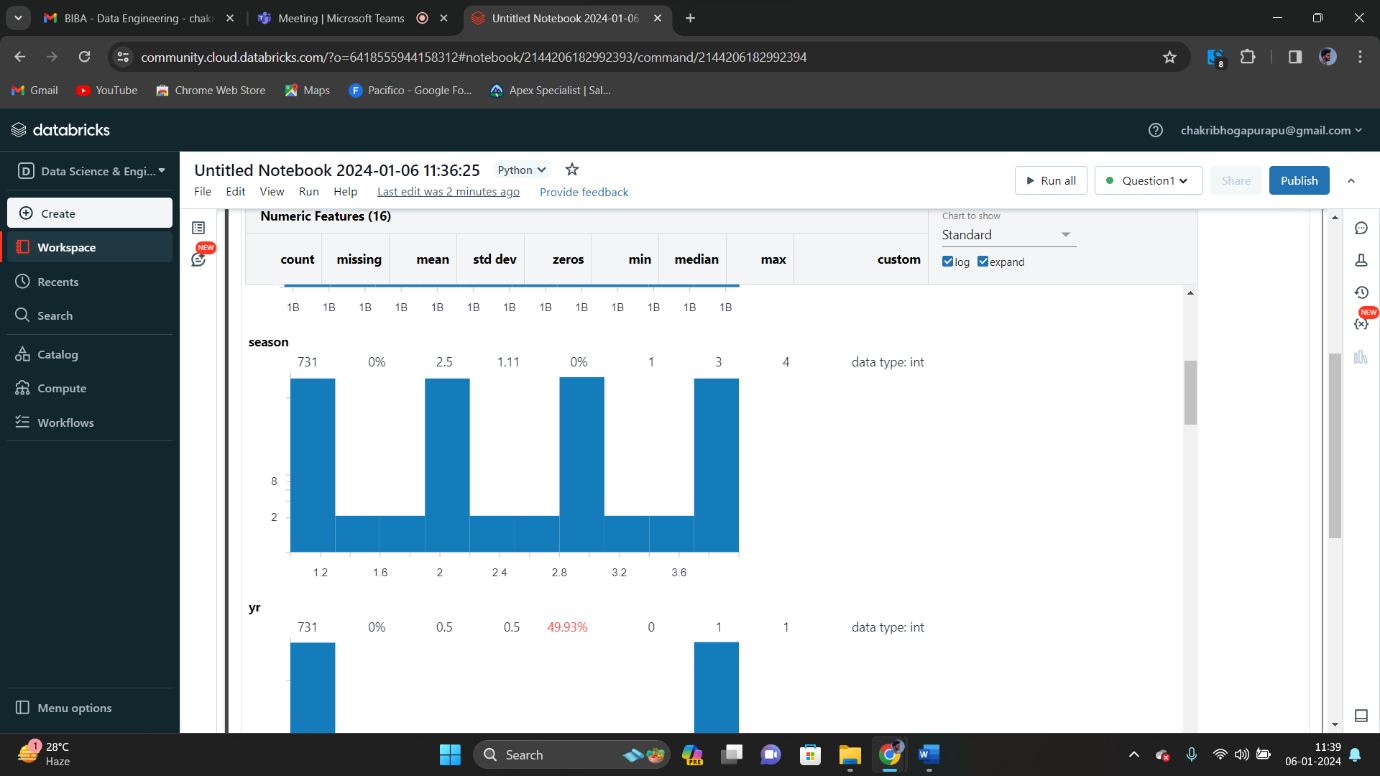
🡪Data profile

Click on data profile

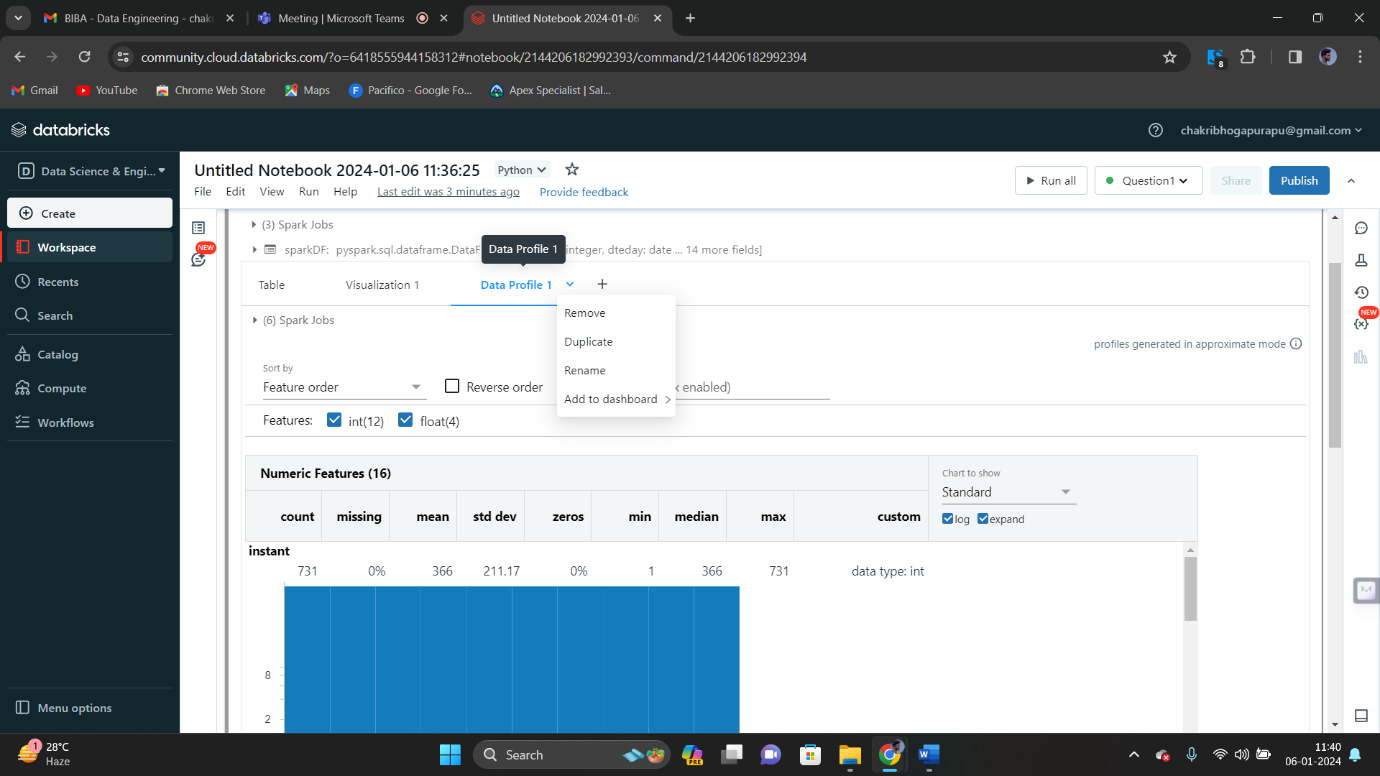
We get



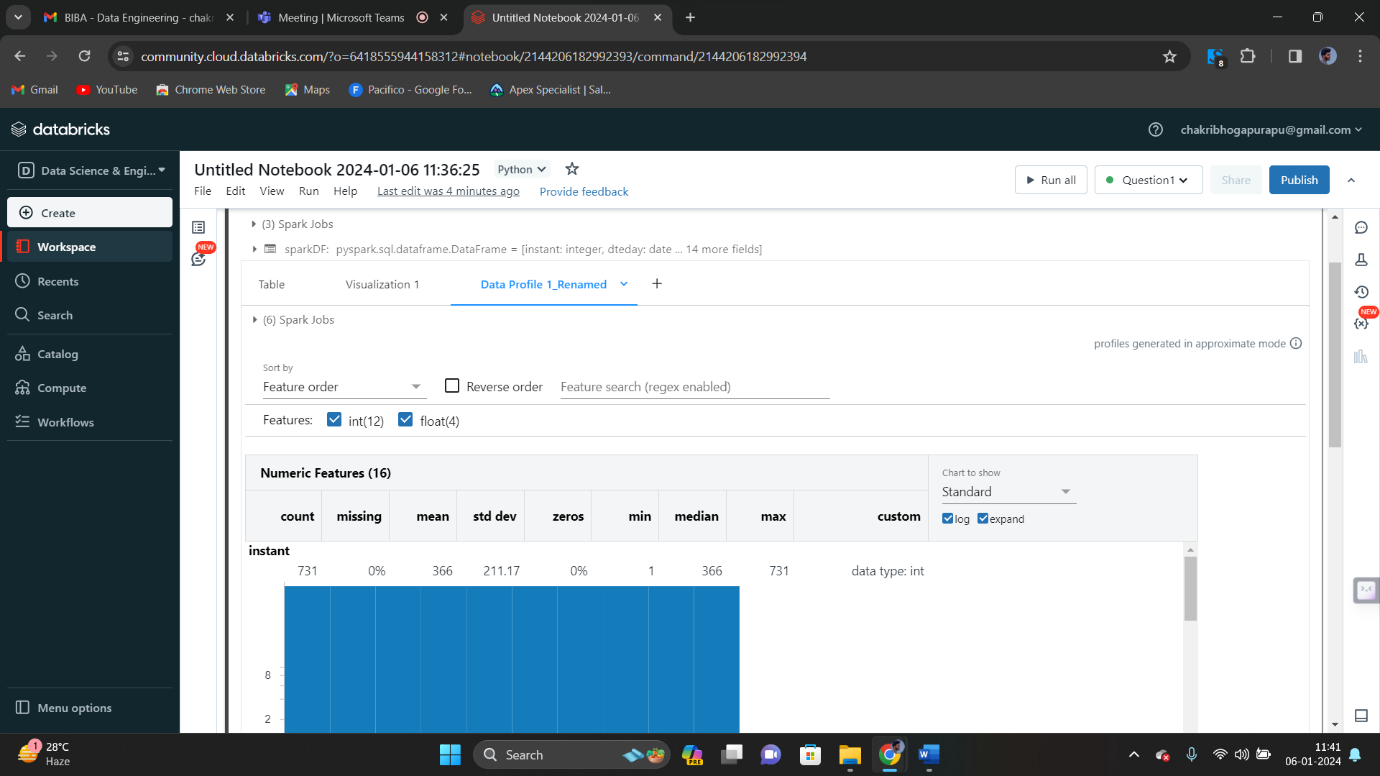
Click on log and expand. We get output as below.



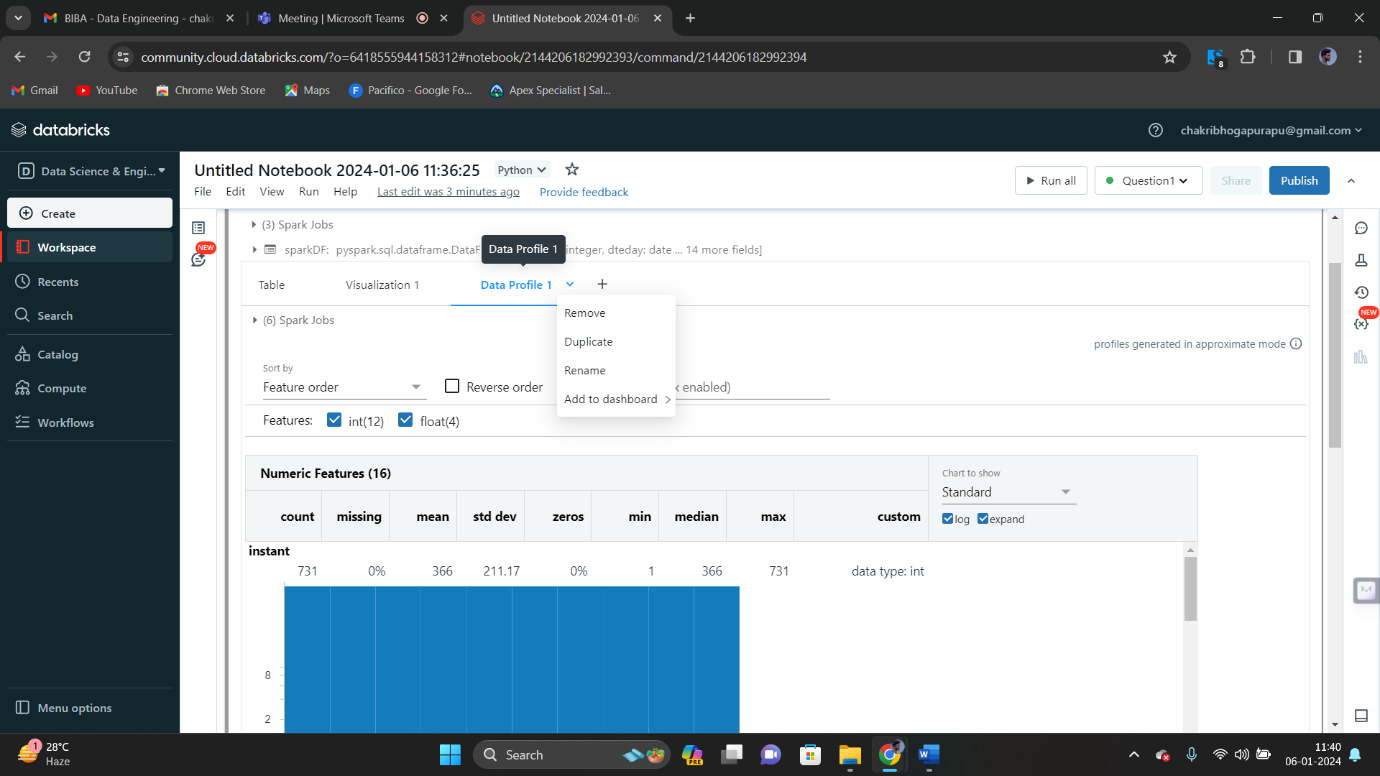
🡪Rename

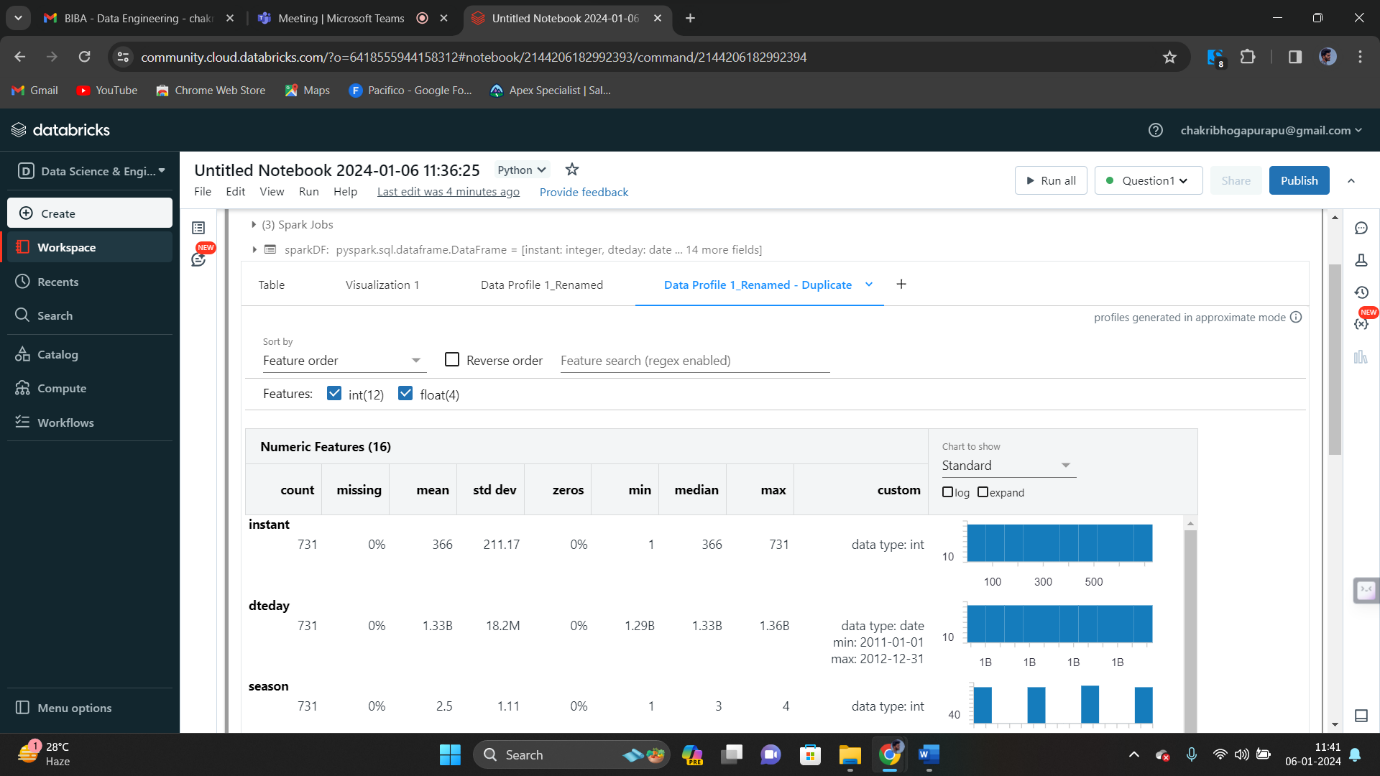
Click on rename option. Now rename “Data Profile 1” as “Data Profile 1\_Renamed” 

Now we get the the name of dayta profile renamed.

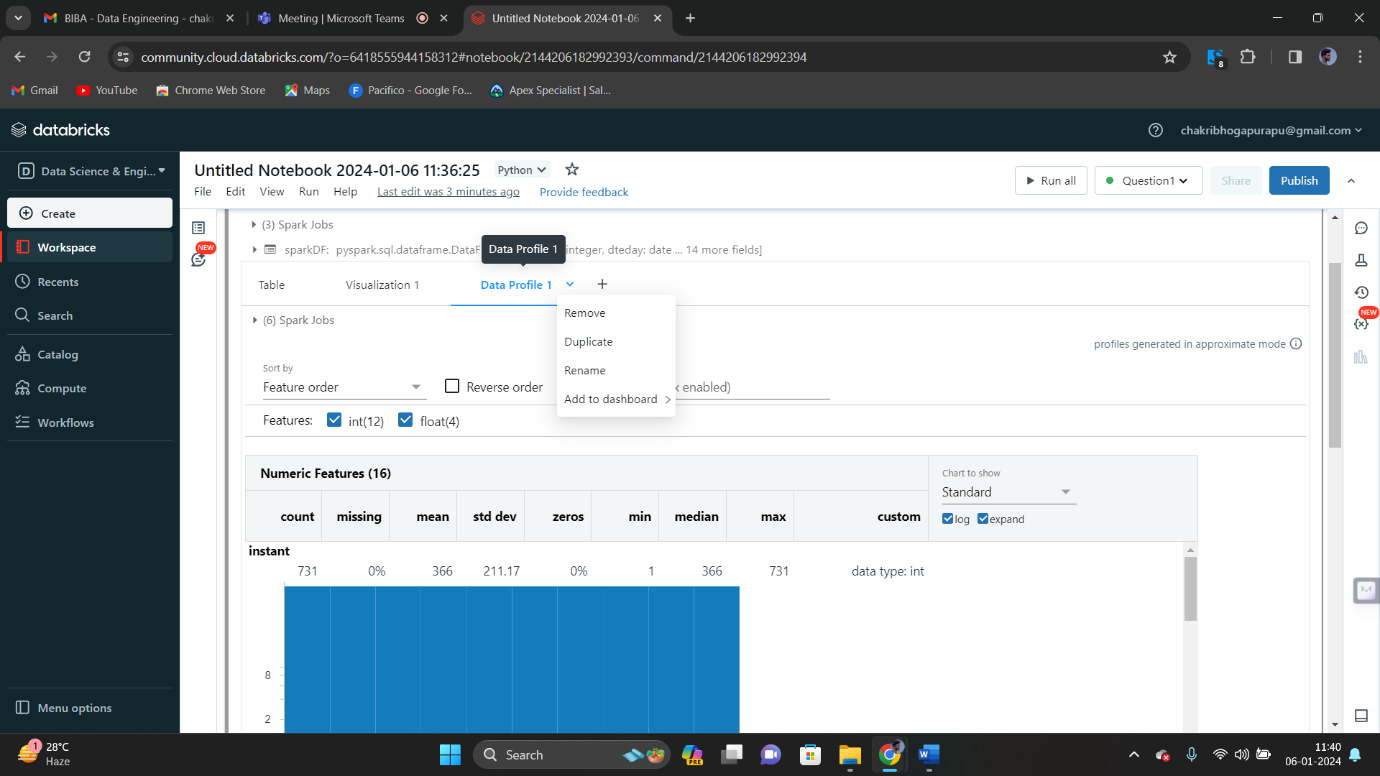


🡪Duplicate

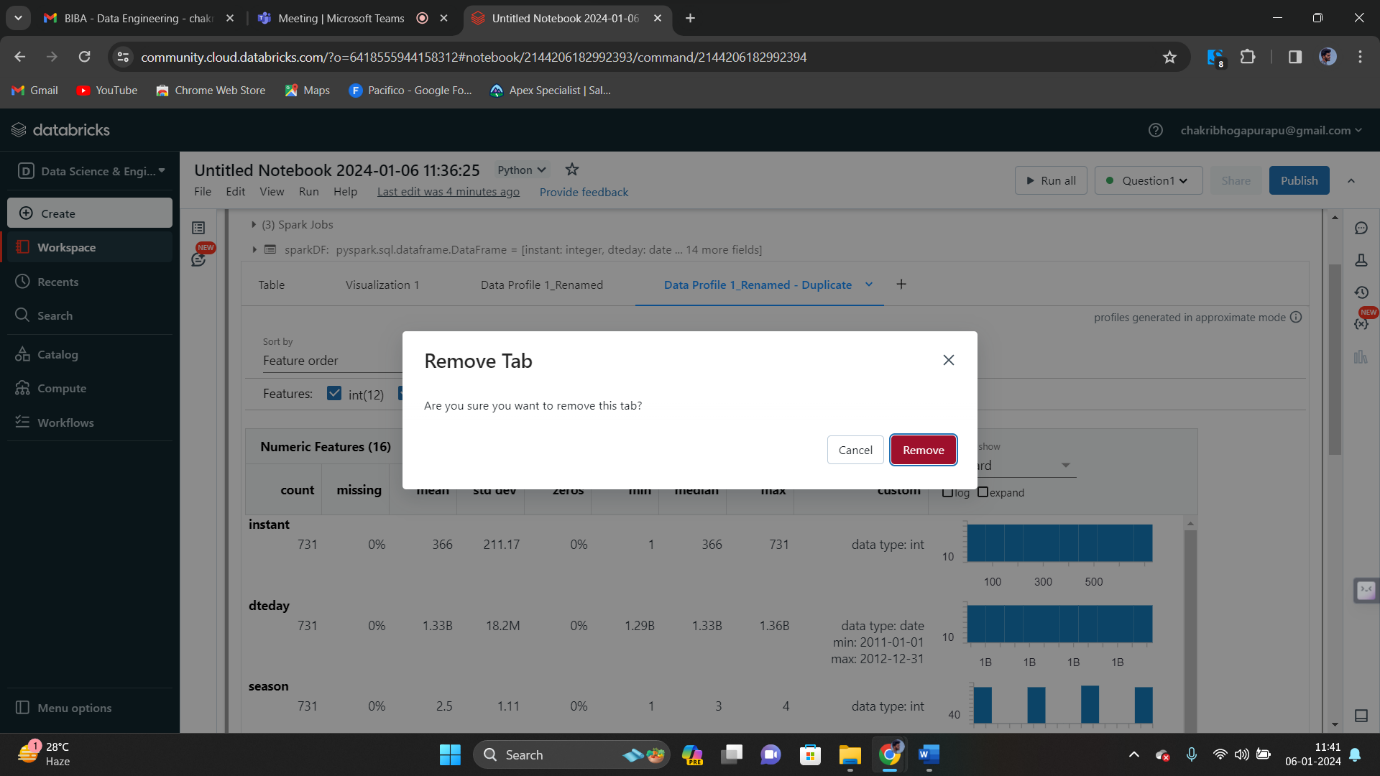
Click on duplicate option as shown below

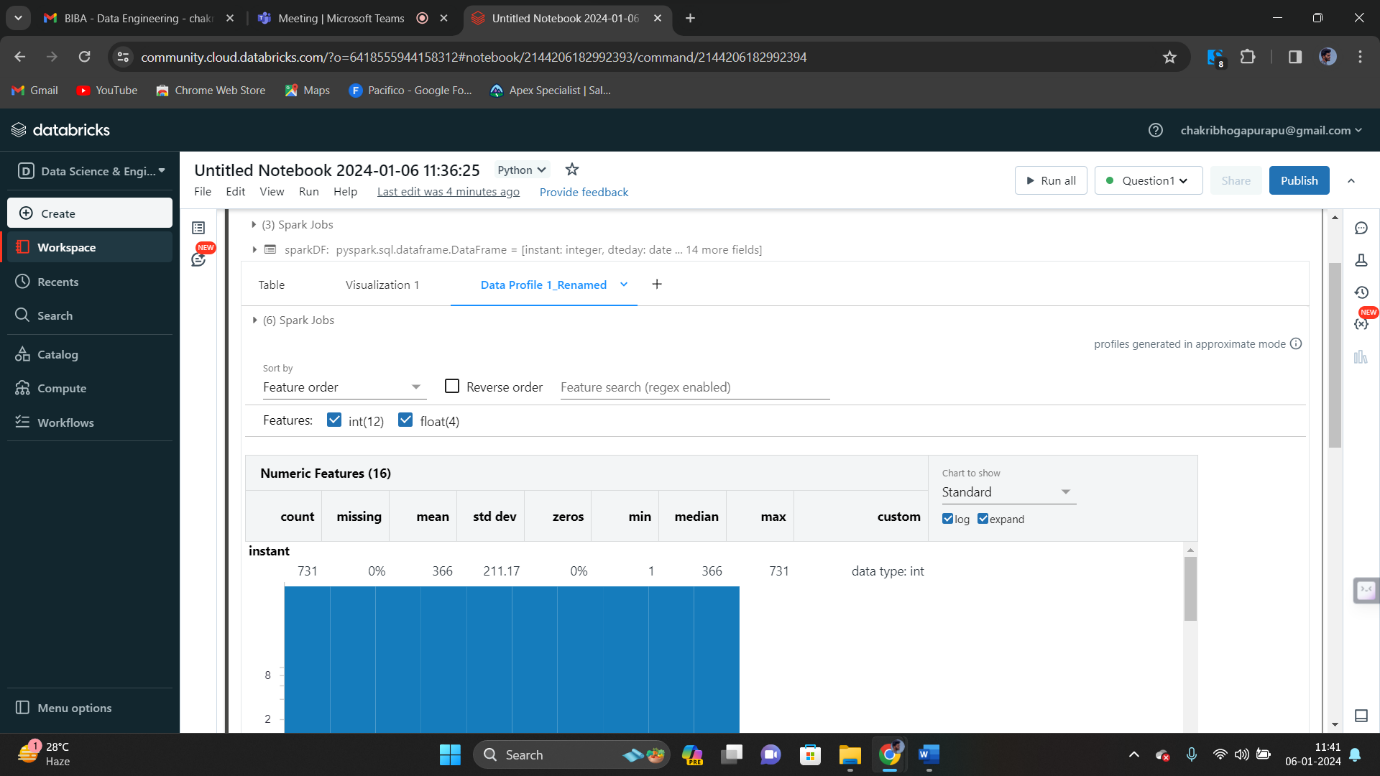
Now we get the duplicate date profile as shown below. 

🡪Remove

Click on remove option from the drop down.

Now we get the pop up to remove the data profile.



Click on “Remove”, our data profile will be removed permanently. Now there will be no data frame as shown in below image.

**Question 2**

Explain the copy activity in Azure data factory.

**Answer**

Copy activity in Azure data factory.

Copy activity is a type of data movement operation that can define within a data pipeline that was created. Data factory is a cloud based data integration service that allows to crate, schedule and manage pipelines.

Features of Copy Activity in Azure Data Factory

* Flexible Data Movement : Copy activity supports various data formats, column mapping, etc. to ensure flexibility in moving data in between source and destination.
* Data Transformation : It allows to transform data during the copy process by using mapping, filtering, etc.
* Monitoring and Logging : Data factory provides monitoring and logging capabilities for copy activity allowing to track progress of data movement operations.
* Integration with Data Management Gateway : If we are uploading data from the local storage, Integration with data management gateway facilitates for secure and efficient data movement.

To use copy activity, we create a pipeline in azure data factory and add a copy activity within that pipeline. We then configure the source and destination data sets, specify the data movement settings and define any required data transformations.

The copy activity in azure data factory is used to copy data between the different data stores that are located in local and in the cloud. After copying the data, we can use other activities to transform and analyse it.

🡪We need to create a blob storage in our resource group. We can store our csv files in folders in blob storage.

We need to create a data factory from the azure portal.

After creating it, we need to use the copy data tool to copy data.

In order to start the copy data tool, we need to create our own data factory.

🡪In azure data factory home page, click on Ingest. And after entering into it, click on “Build-in copy task”. Now click on next. On it, specify a name for the connection. Select yout Azure subscriotion from the azure subscription list and select the storage account of yours that was created by you. Now click on test connection and select create option.

Now choose the file that was uploaded earlier by you. Click on browse to navigate to that csv file. Now click on ok.

🡪Now complete the destination configuration. To do it, click on AzureBlobStorage connection that was created in connection block.

Now give the output path.

🡪Now review all settings and deployment.

Click on settings, and give the name for your pipeline. Click on next and on summary page, and on deployment complete page, select “Monitor” to monitor the pipeline.

🡪Now monitor the running results.