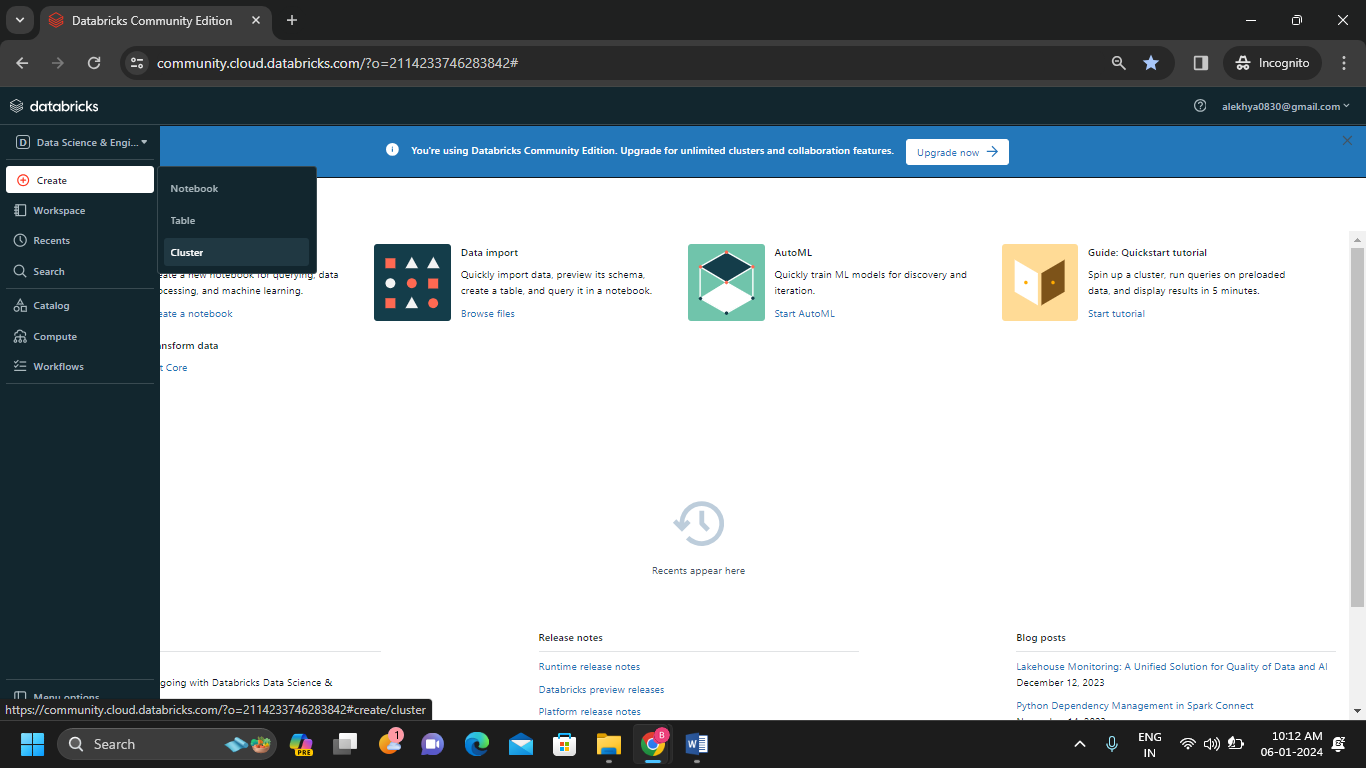
**Alekhya Krishna Balivada Azure Databricks Coding Assessment - (06-01-2024)**

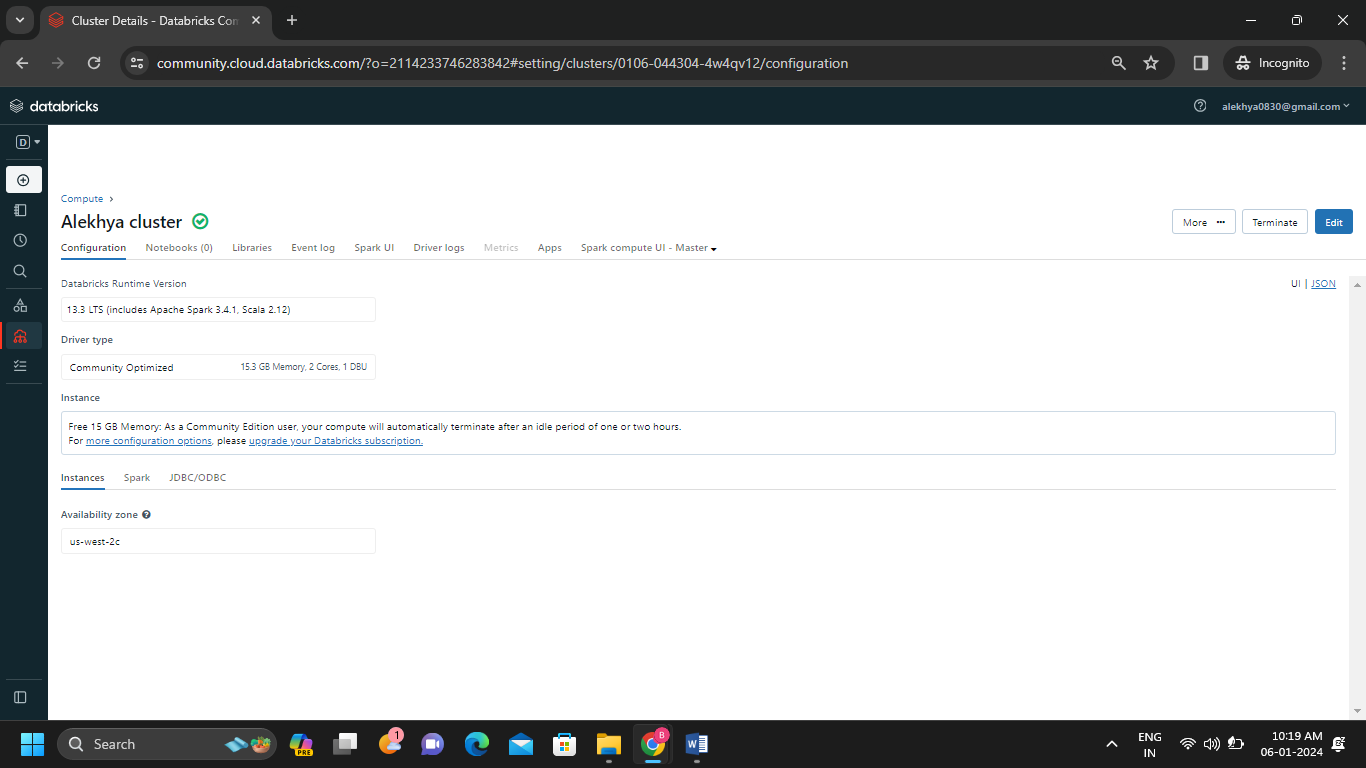
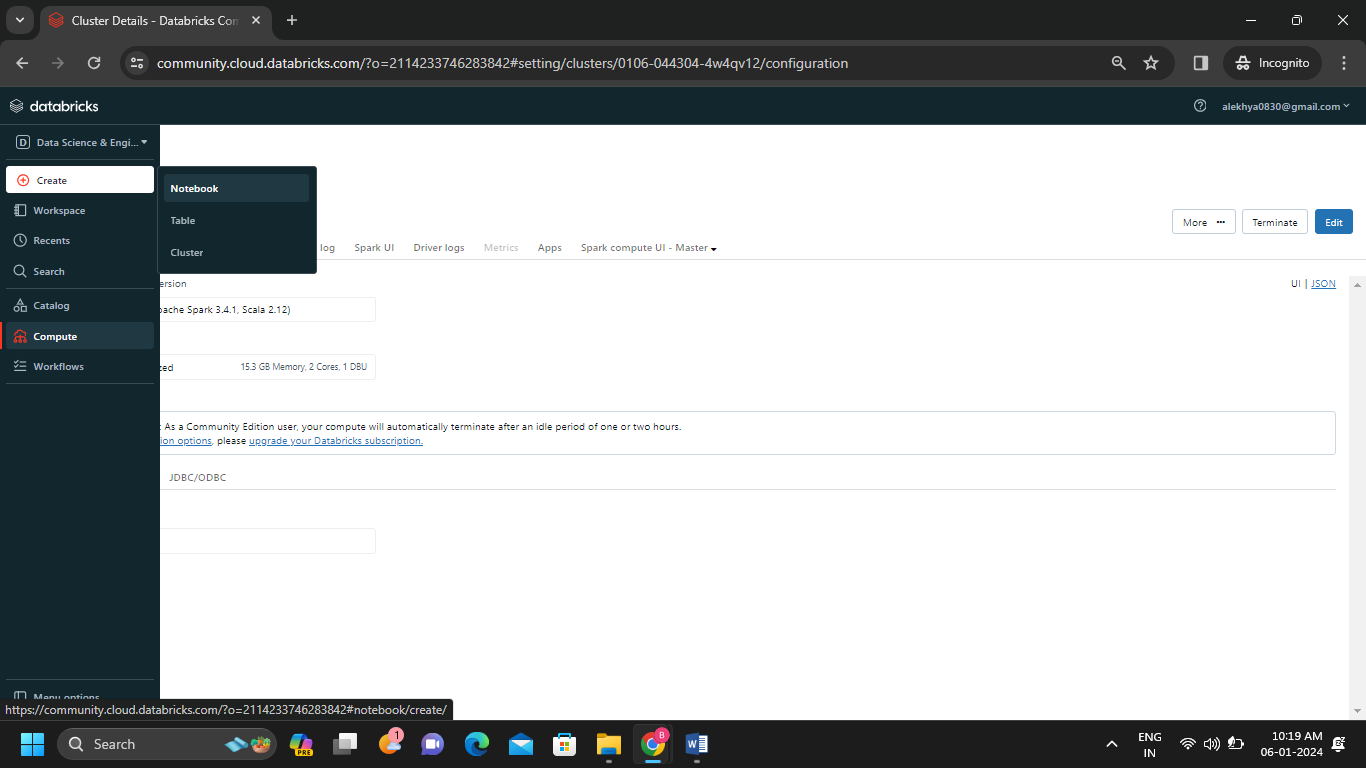
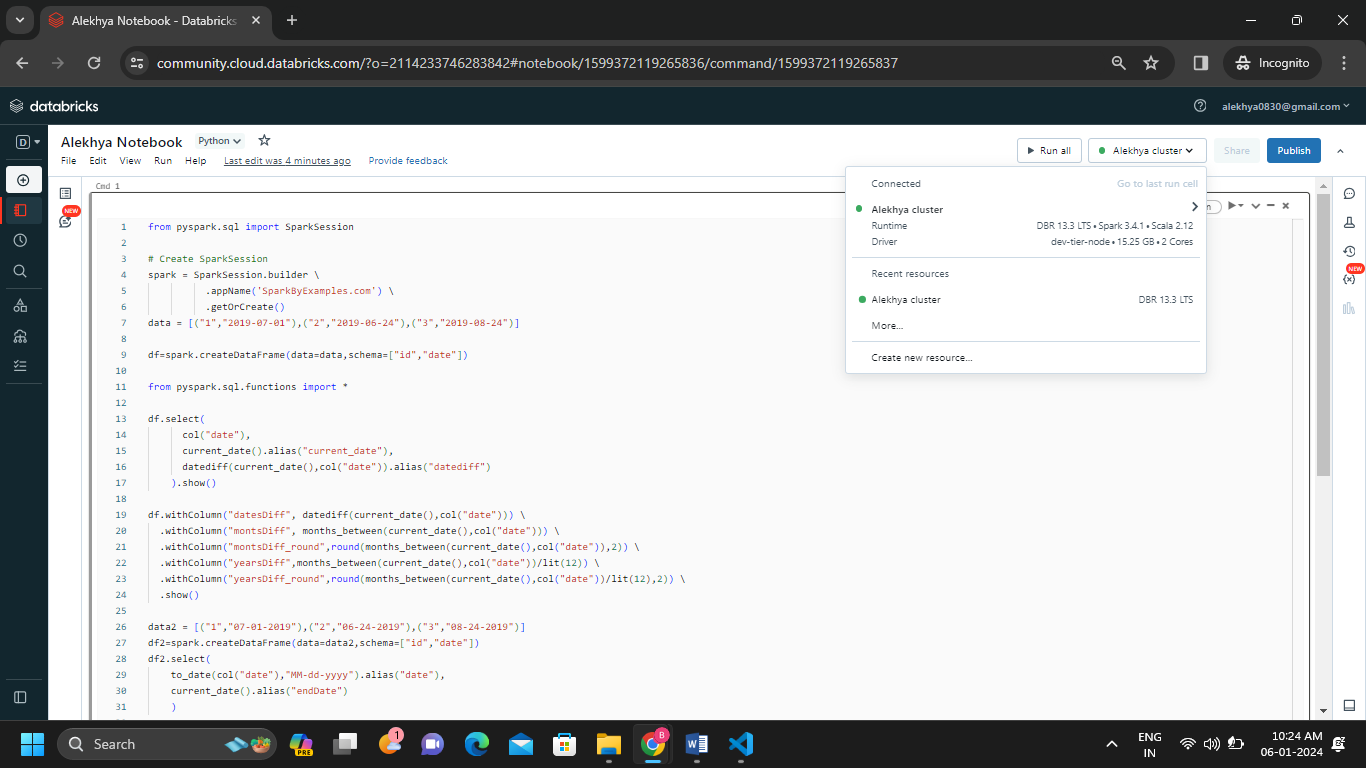
|  |  |
| --- | --- |
| Name | Alekhya Krishna Balivada |
| Date | 06-01-2024 |
| Type | Coding Assessment |
| Topic | 1Q. Cluster creation, running commands on notebook, Dataframe creation, Creating visualization.  2Q. Copying data on Azure Data Factory. |

**1Q. Create a cluster &Attach the notebook to the cluster and run all commands in the notebook&creates a DataFrame from a Databricks dataset&Create a Visualizations in Databricks notebooks**

**Ans:**

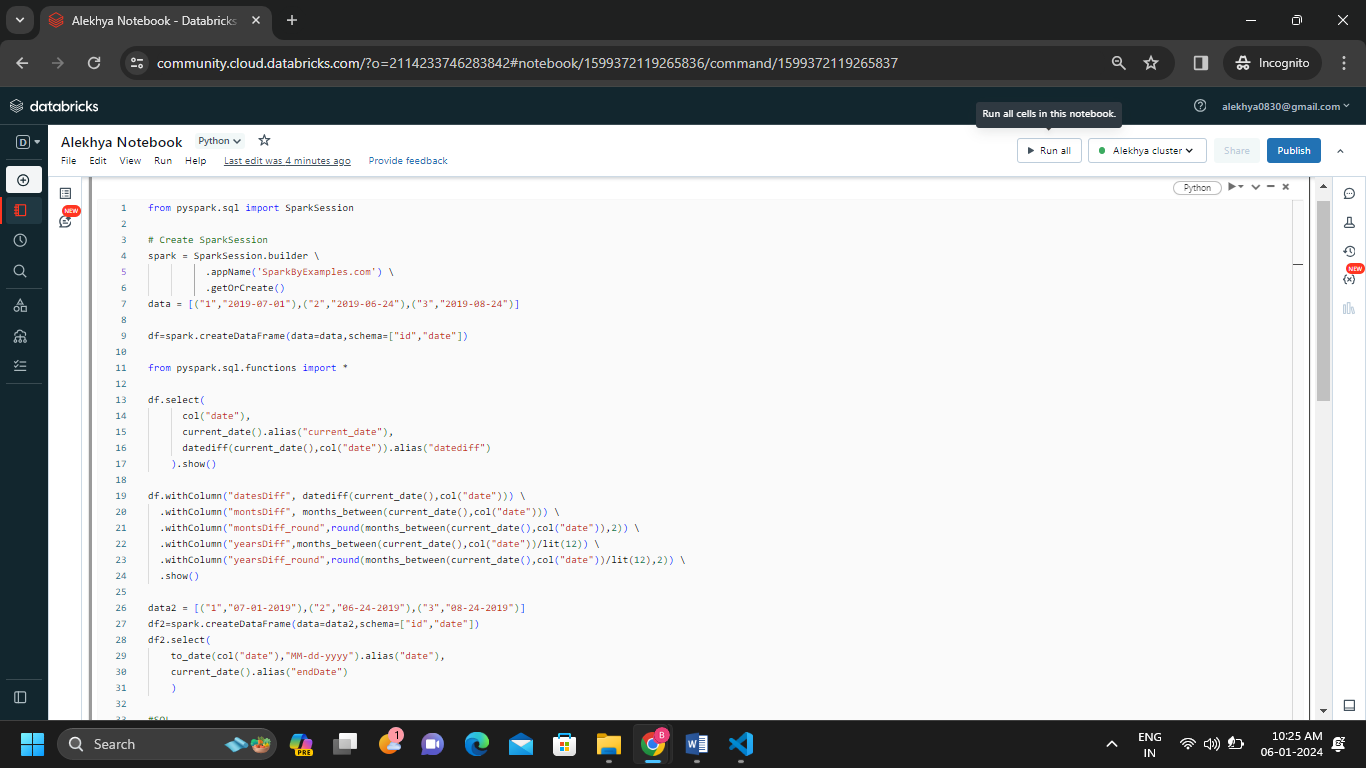
1. **Create a cluster & Attach the notebook to the cluster and run all commands in the notebook**

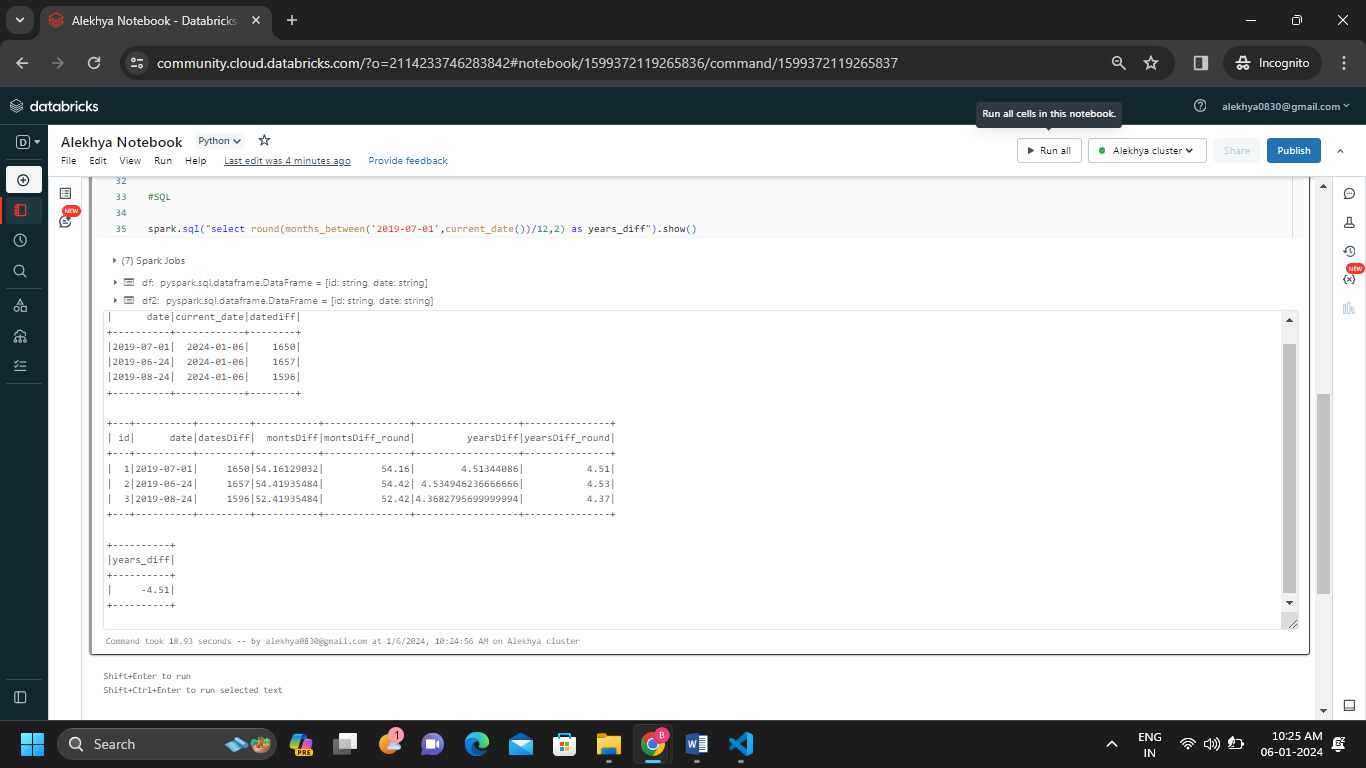
**Step-1:** Create the cluster with name “Alekhya cluster” by clicking on create cluster.

* We can see that the cluster is created.
* Now create a notebook with name “Alekhya notebook” to run the commands.
* Notebook is created and it is connected to the cluster which we created in step-1.

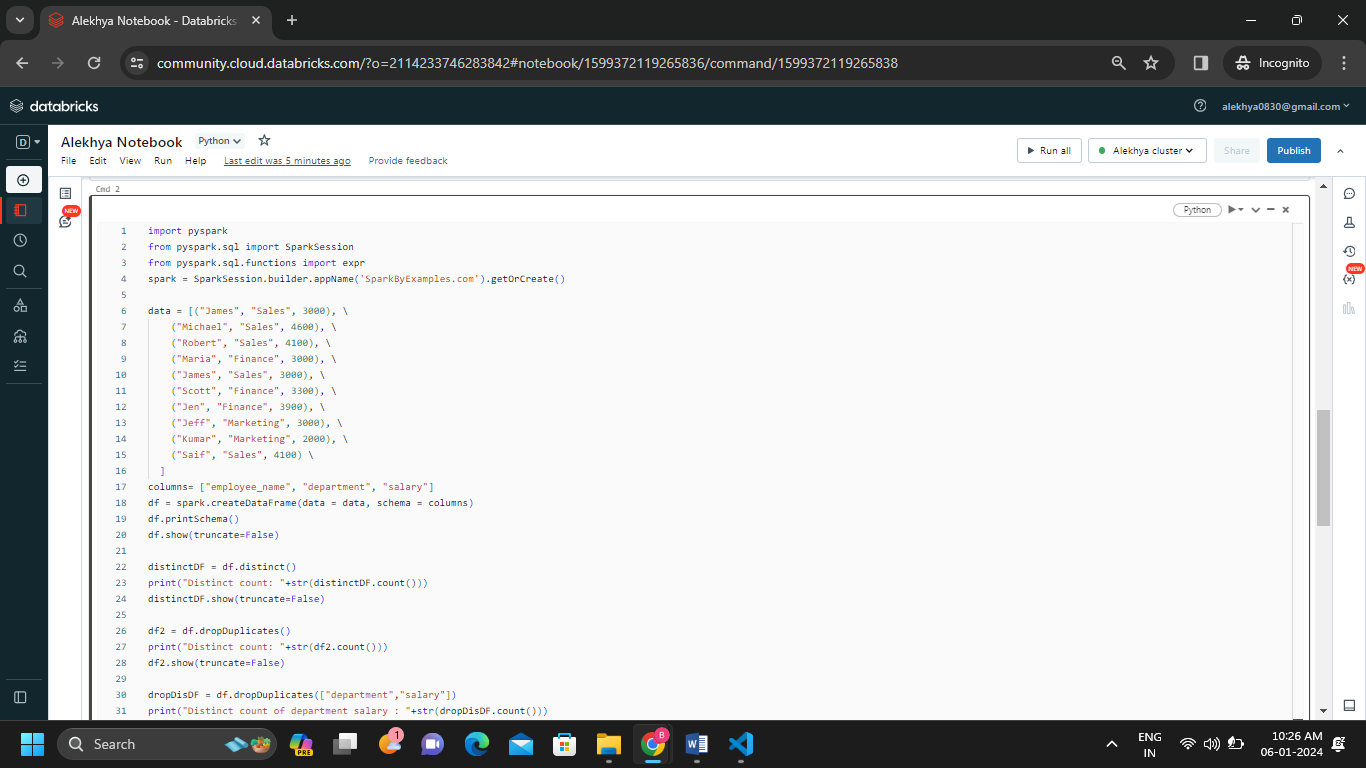
**Step-2:** Now run some commands in the created notebook which is connected to the cluster we created.

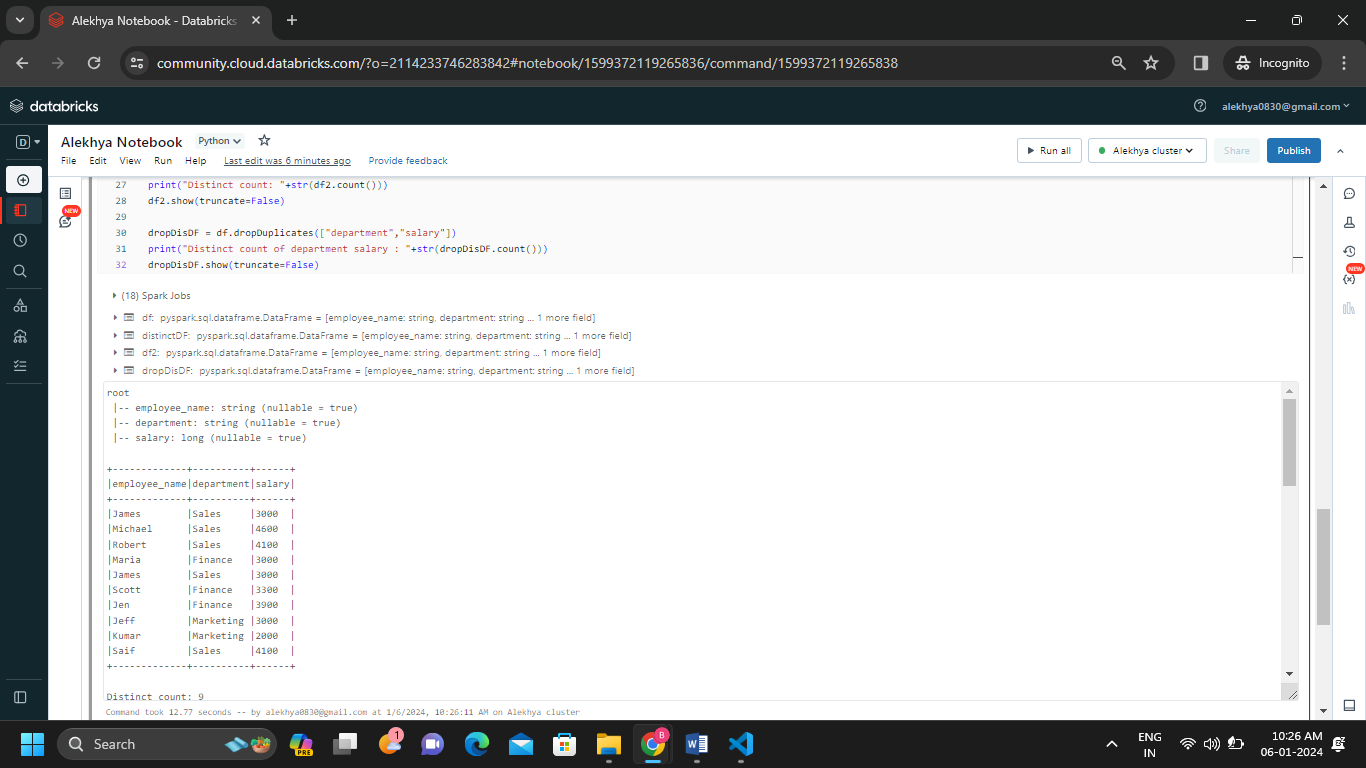
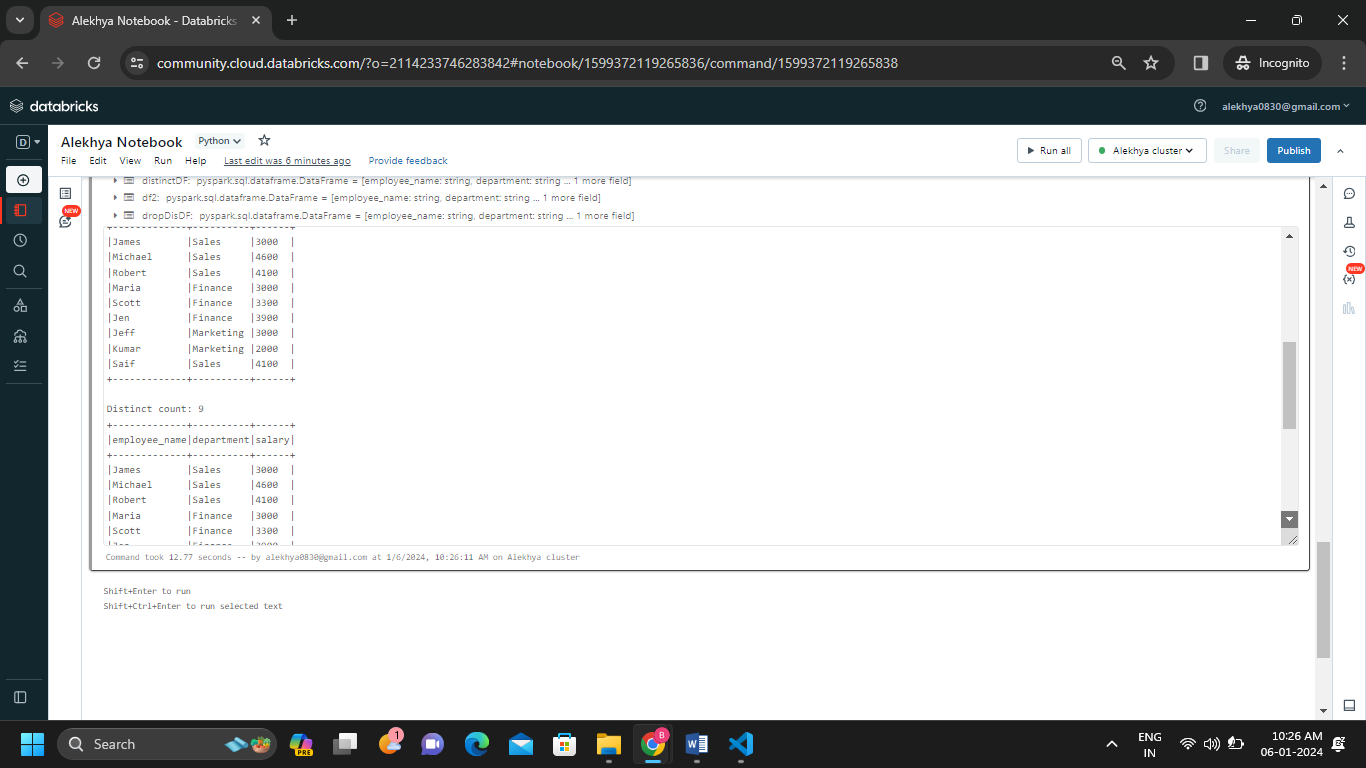
* I have run a code on the notebook which I created and got the output.

**Code-1**

****

**Code-2:**





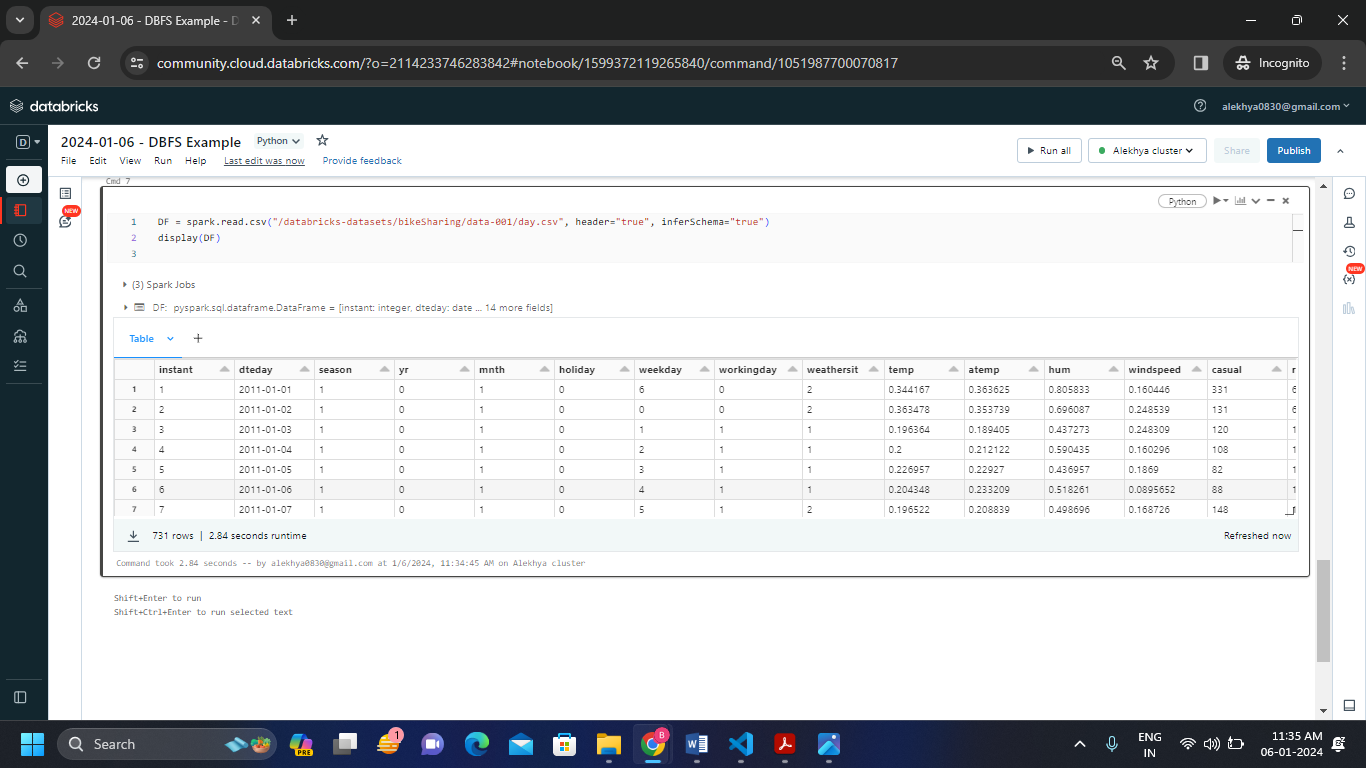
**B) creates a DataFrame from a Databricks dataset**

In Data bricks, we can create a DataFrame from various data sources like CSV files, Parquet files, JSON files, databases and so on.

Run the below commands the we get the output

DF = spark.read.csv("/databricks-datasets/bikeSharing/data-001/day.csv", header="true", inferSchema="true")

display(DF)

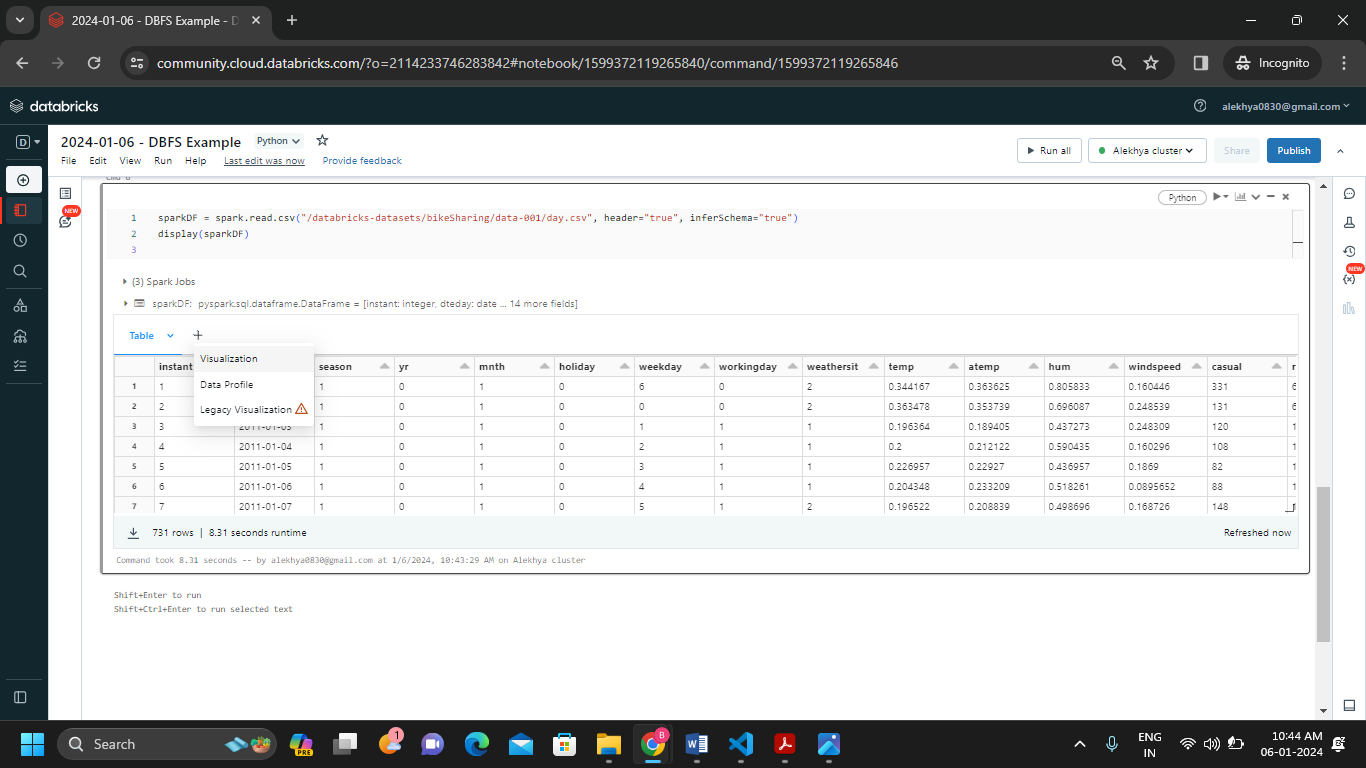
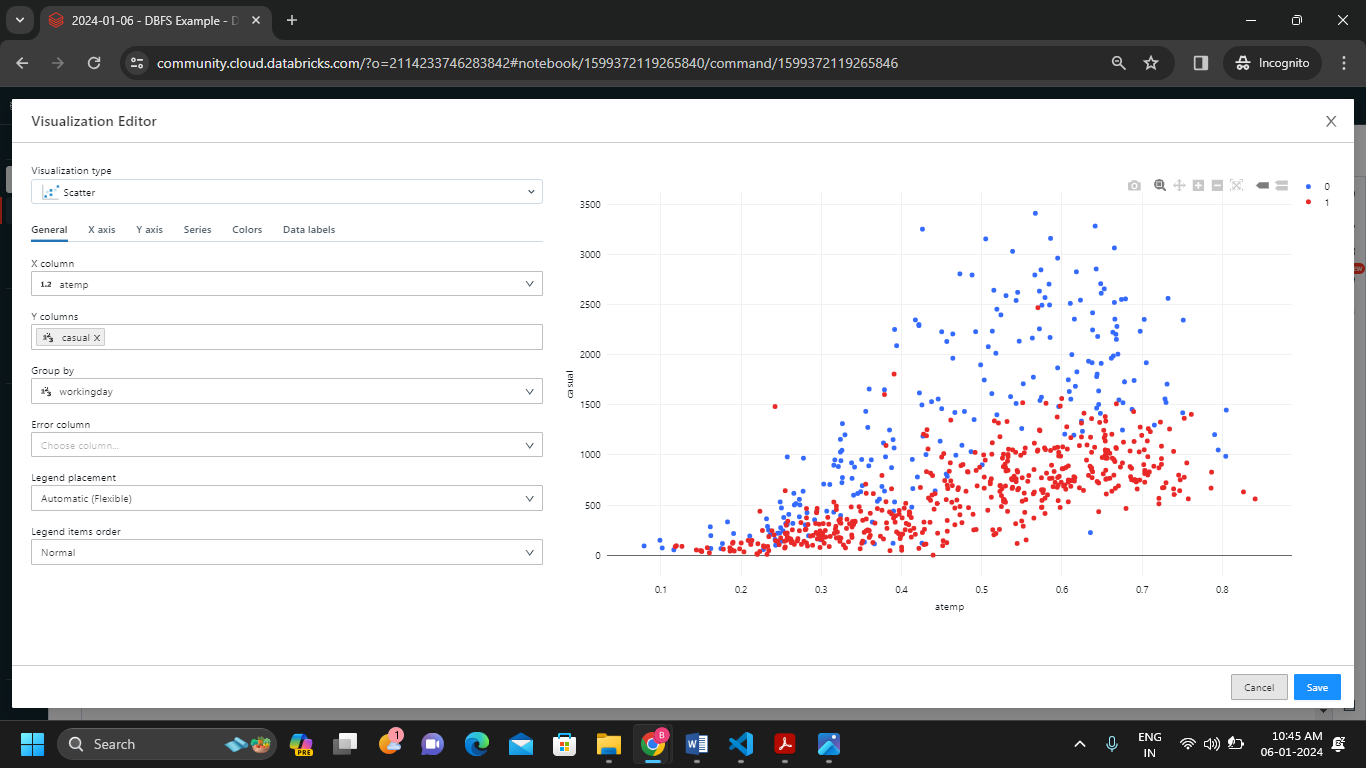
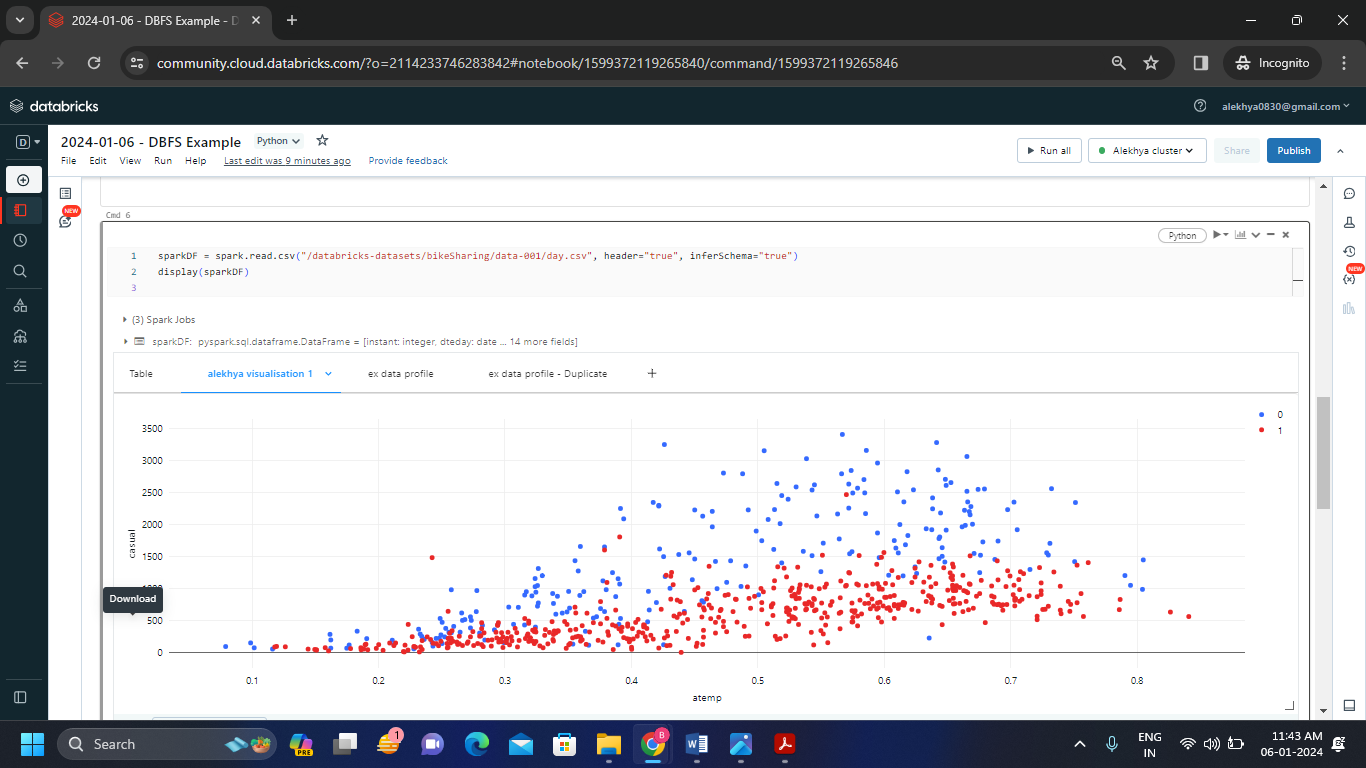


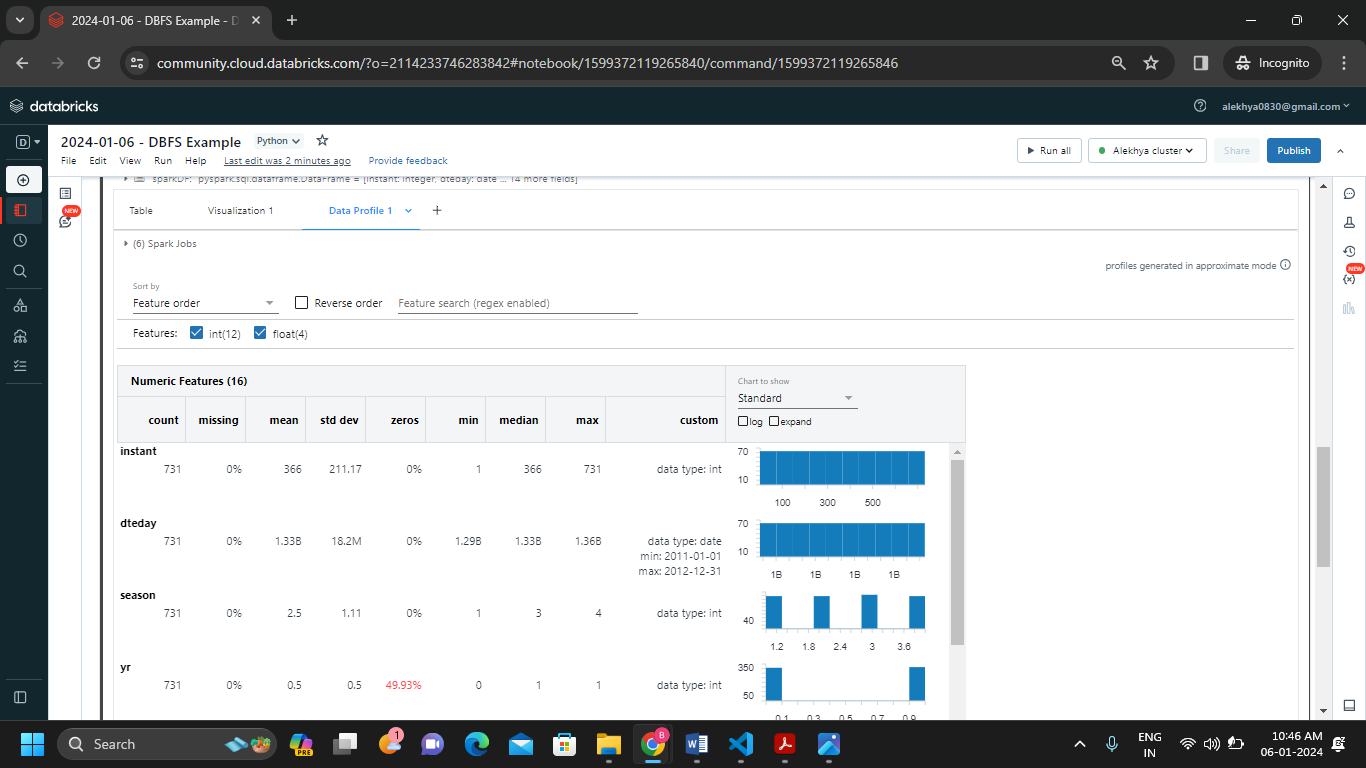
**C) Create a Visualizations in Databricks notebooks &Rename, duplicate, or remove a visualization or data profile.**

**Step-1:** Run the below commands the we get the output

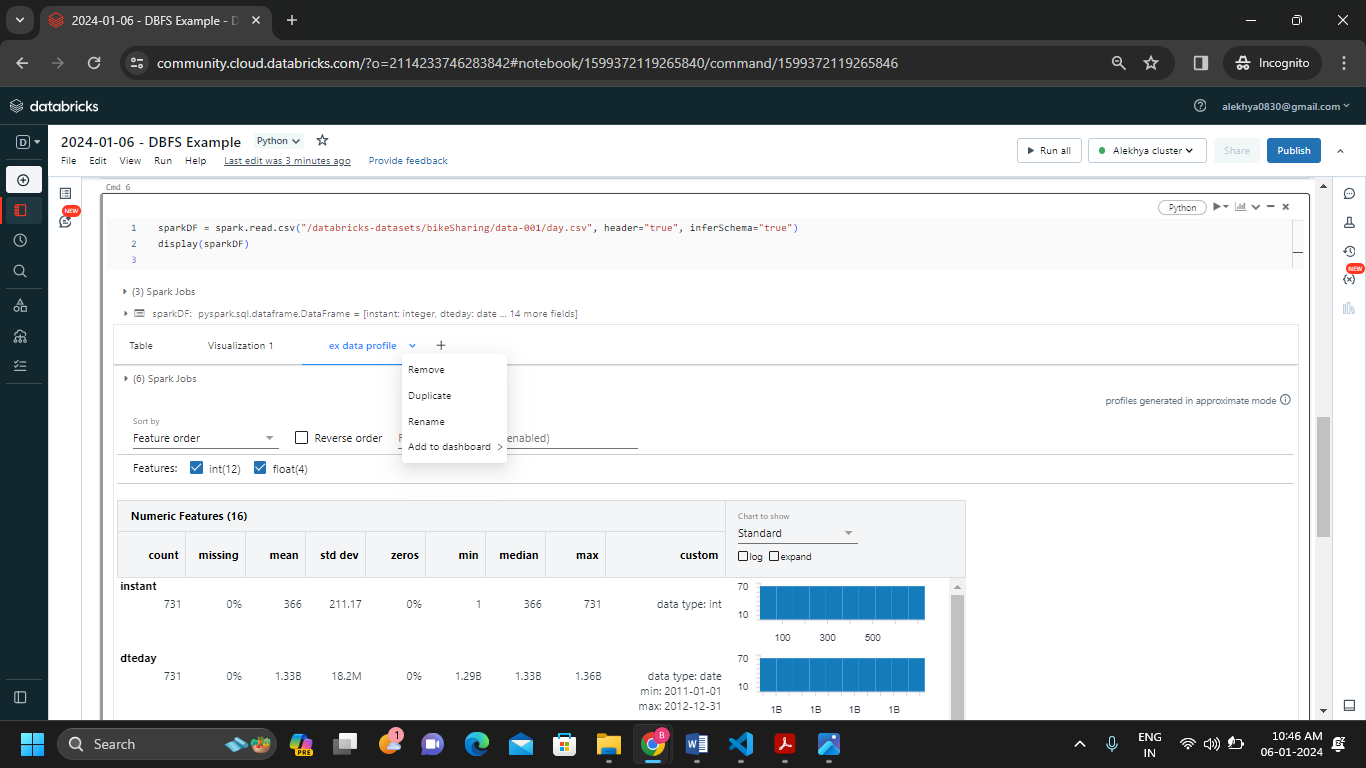
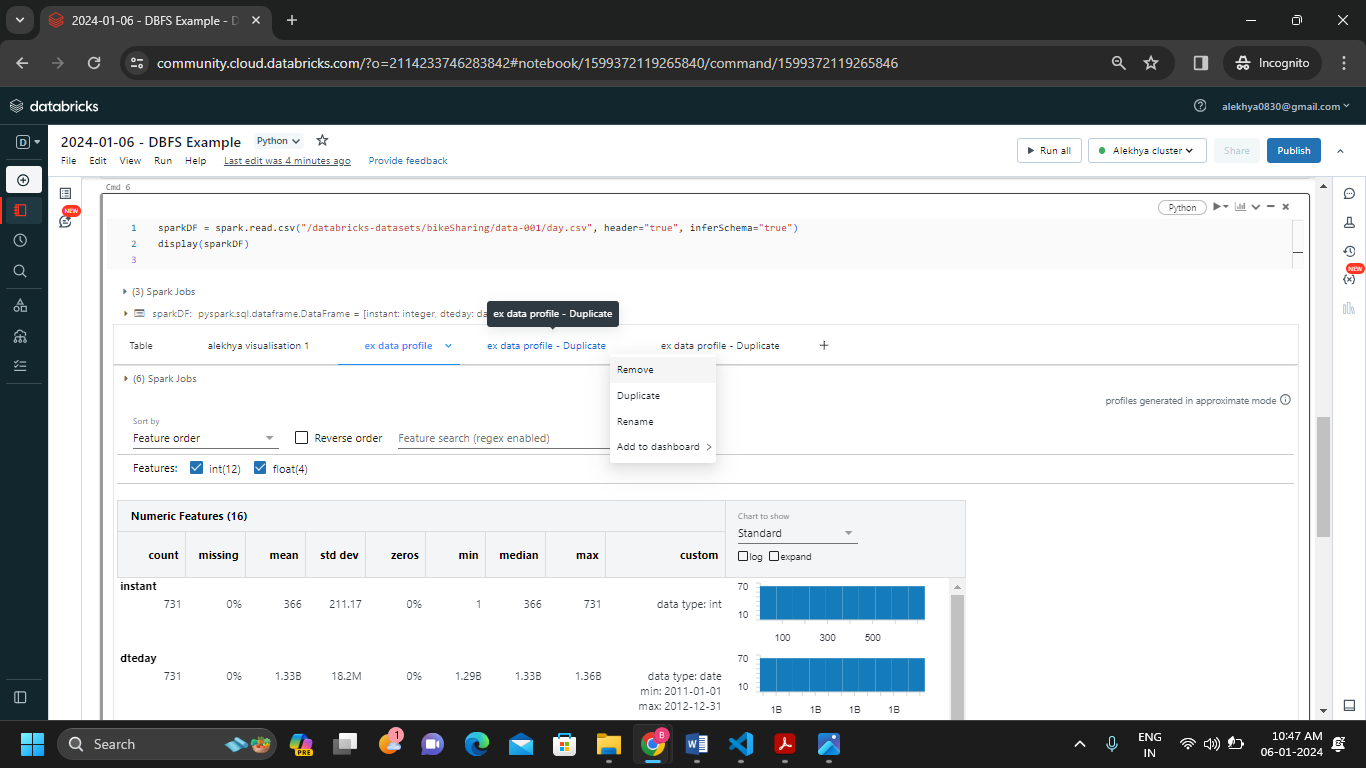
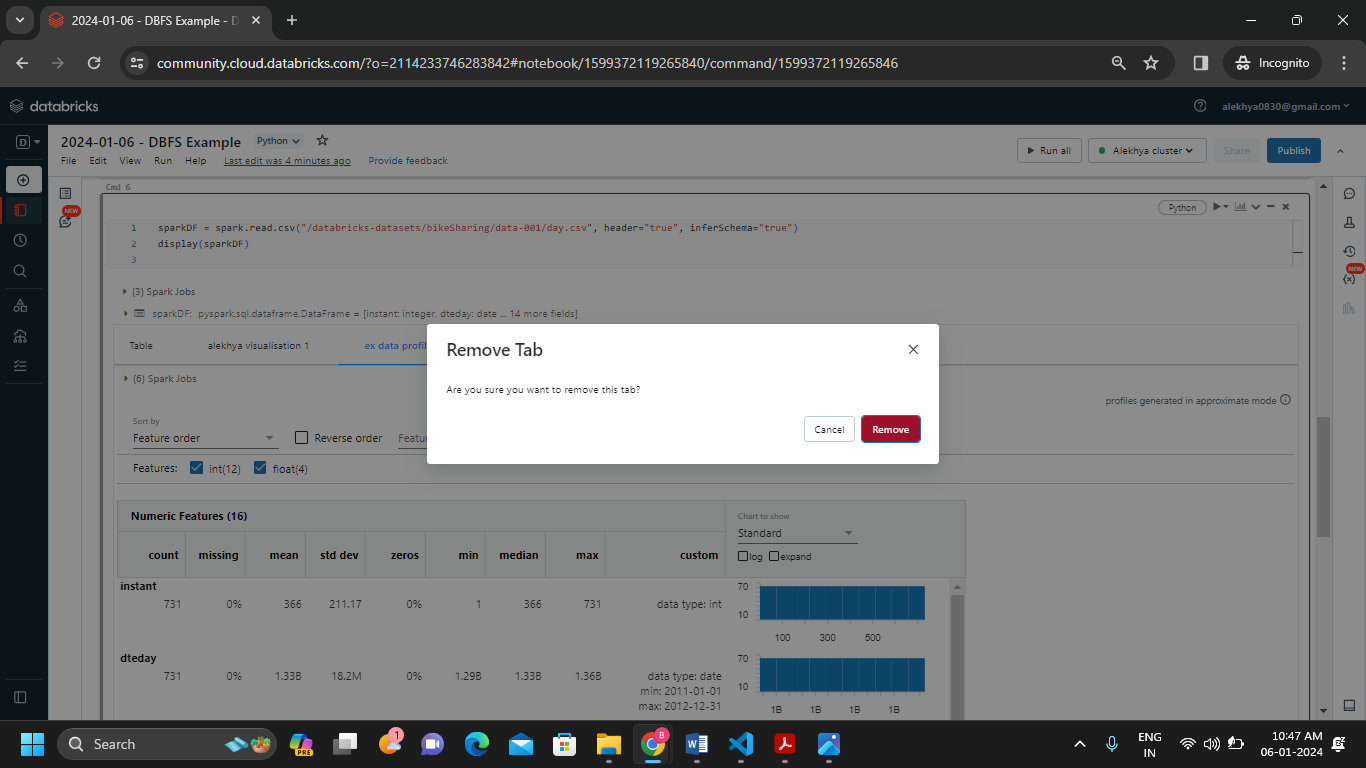
sparkDF = spark.read.csv("/databricks-datasets/bikeSharing/data-001/day.csv", header="true", inferSchema="true")

display(sparkDF)

* Run the below commands the we get the output
* Now click on the + just beside the table to get the visualization of the output in a barchar or scatter or anything else by choosing the option.
* ****We get the visualization graph by giving the respected columns and the graph type in the visualization editor and now save it.
* We can see the visualization graph just beside the table output.
* Now create the data profile by clicking on + just like visualization which is shown as below.



**Rename, duplicate, or remove a visualization or data profile.**

* Now click on + at the data profile or visualization we can see the remove, duplicate and rename options .
* We renamed the visualization and data profile by clicking on rename option
* We add the duplicate data profile also by clicking on duplicate option
* We can remove the data profile by clicking on remove option.

**2Q. Explain the copy activity in Azure data factory.**

**Ans:** The Copy Activity is a key element of Azure Data Factory (ADF) that facilitates data movement from a source to a destination. It is an essential component of building ADF data pipelines that facilitate the extraction, transformation, and loading (ETL) of many data sources.

* To perform the copy data activity we need to have source and destination storage accounts.
* Source account contains the container with a file where we uploaded.
* Destination account also contains a container which is empty at the starting then after creating the pipeline and dataset on Azure data factory we can get the copied file from source container.

To perform the copy activity we need to follow the certain steps.

**Step-1:** Create an Azure Data Factory by clicking on Data Factory.

**Step-2:** Launch the Azure Data Factory Studio**.**

**Step-3:** Create 2 storage accounts and then create 2 new containers in each storage account. Now upload a csv file in one container and keep another container empty.

**Step-4:** Now in the data factory create a new pipeline and click on copy data.

**Step-5:** Fill the necessary data in general block

**Step-6:** In the source block create new dataset (dataset-1).

**Step-7:** As we uploaded the csv file in container click on Azure Blob Storage and select csv format.

**Step-8:** Now click on new link service then link the first storage account where we uploaded the csv file and also give the path of the container 1 then click on test connection and create the dataset.

**Step-9:** In the sink block create a new dataset(dataset-2) to get the copied file which we uploaded in container1.

**Step-10:** Now click on new link service then link the storage account 2 where the copy of csv file will be pasted and also give the path of the container 2 which consists of no files then click on test connection and create the dataset.

**Step-11:** Now click on validate and we can see the pipeline validation output.

**Step-12:** Now debug the copied file and we can see the activity status of the copied data

**Step-13:** We can see the input code and output code of the copied data

**Step-14:** We also can see the details the operations we have done above

**Step-15:** Hence by following the above steps we successfully can copy the csv file from one storage account to another storage account using Azure Data Factory Pipelining.