Coding Challenge-4.1
Azure Databricks

Name-Ajay Chaudhary
Batch-Data Engineering(Batch-1)

Exploratory data analysis (EDA) in Databricks &Visualizing data in Databricks

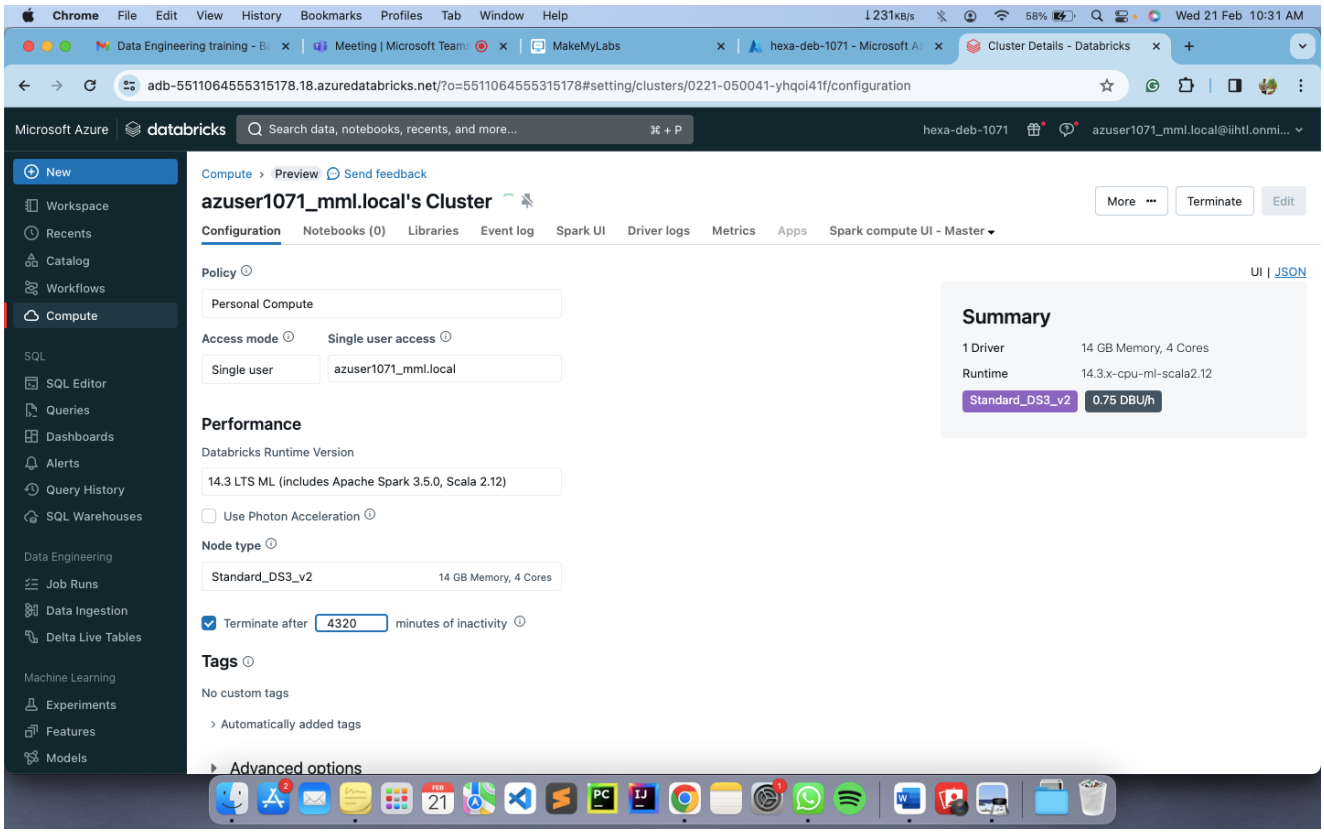
Exploratory data analysis(EDA)-EDA is the process of analyzing datasets to summarize their main characteristics, often employing statistical graphics and other data visualization methods. In Databricks, you can perform EDA using Spark SQL, DataFrame API, and various libraries like Pandas, Matplotlib, Seaborn, etc.

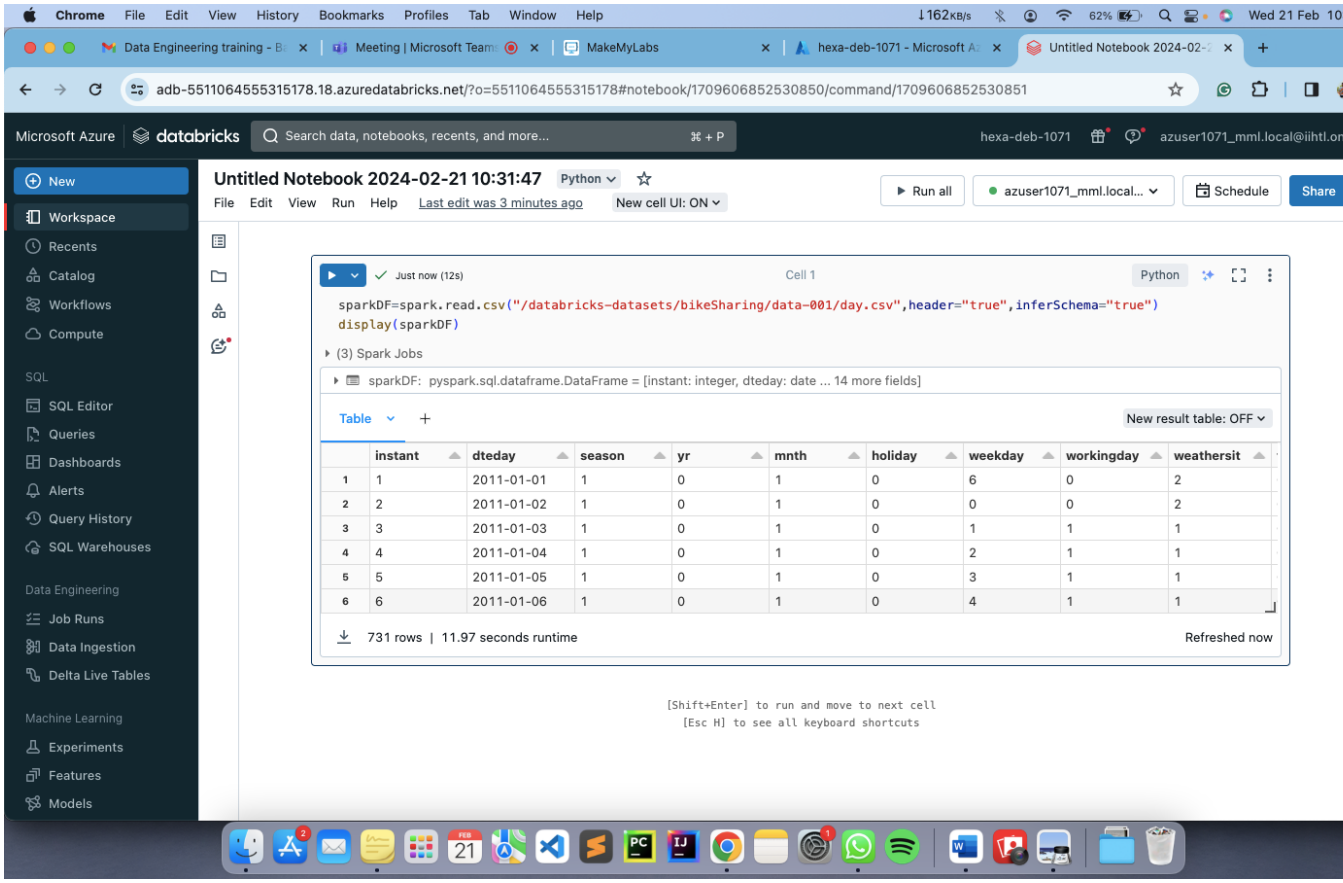
Visualizing data-Visualizing Data: Databricks provides built-in support for visualizing data using various libraries such as Matplotlib, Seaborn, Plotly, and the built-in visualizations of Databricks notebooks. You can create visualizations directly within your Databricks notebook using these libraries or by utilizing the built-in display capabilities of Databricks.

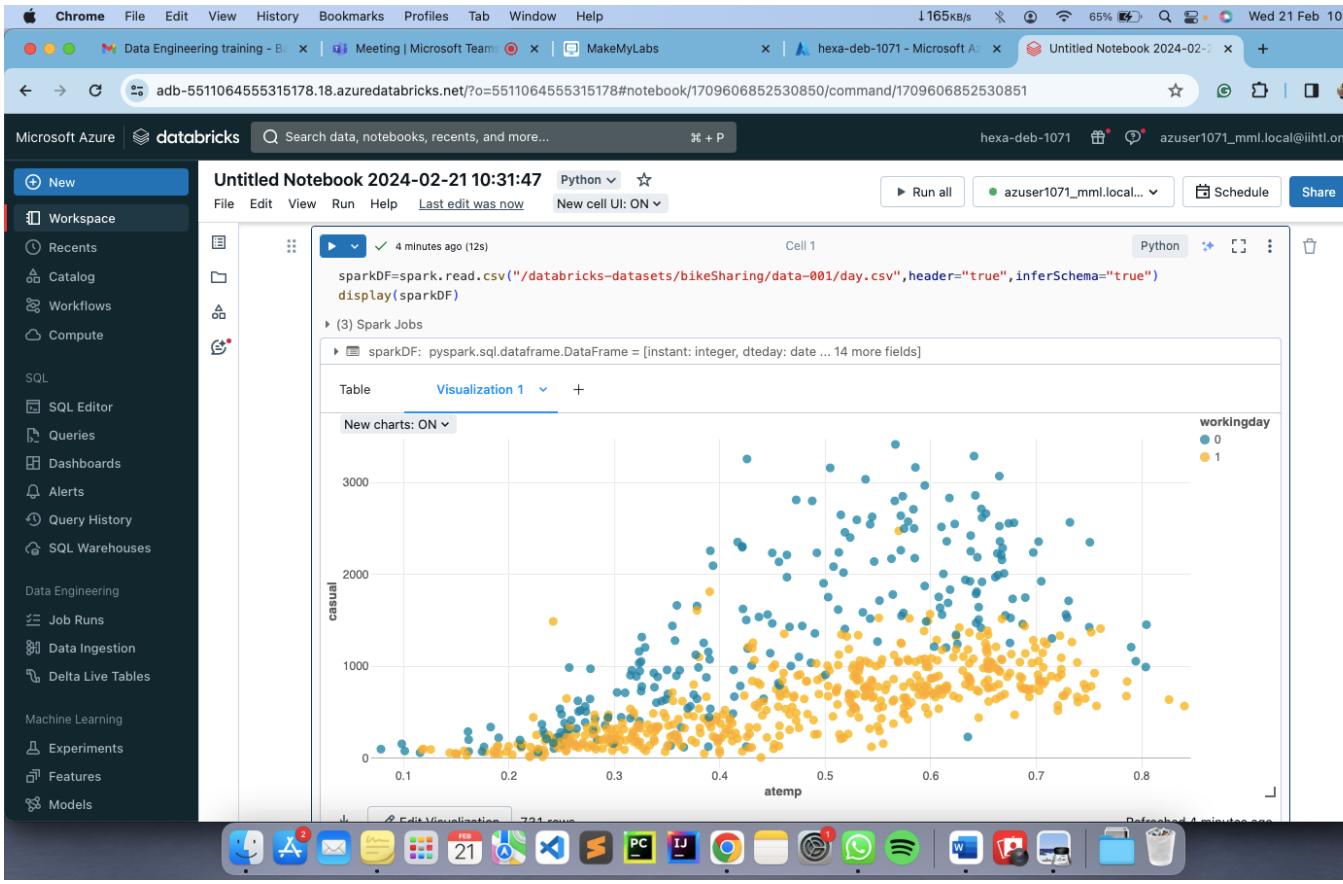
Steps to create visualization-
In order to create visualizations, we need to have data.

- After creating a table
- Click on + symbol
- Click on visualization.
- Select the type of visualization, then select Scatter.

Created a cluster to run the visualization




Chrome File Edit View History Bookmarks Profiles Tab Window Help
Wed 21 Feb
Data Engineering training - B
Meeting | Microsoft Team
MakeMyLabs
hexa-deb-1071 - Microsoft A
Untitled Notebook 2024-02-
adb-5511064555315178.18.azuredatabricks.net/?o=5511064555315178#notebook/1709606852530850/command/1709606852530851
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
⌘ + P
hexa-deb-1071
azuser1071_mml.local@iihtl.or
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Experiments
Features
Models
Untitled Notebook 2024-02-21 10:31:47
Python
File Edit View Run Help
Last edit was 3 minutes ago
New cell UI: ON
Run all
azuser1071_mml.local...
Schedule
Share
Just now (12s)
Cell 1
Python
sparkDF=spark.read.csv("/databricks-datasets/bikeSharing/data-001/day.csv",header="true",inferSchema="true")
display(sparkDF)
(3) Spark Jobs
sparkDF: pyspark.sql.dataframe.DataFrame = [instant: integer, dteday: date ... 14 more fields]
Table
New result table: OFF
instant dteday season yr mnth holiday weekday workingday weathersit
1 1 2011-01-01 1 0 1 0 6 0 2
2 2 2011-01-02 1 0 1 0 0 0 2
3 3 2011-01-03 1 0 1 0 1 1 1
4 4 2011-01-04 1 0 1 0 2 1 1
5 5 2011-01-05 1 0 1 0 3 1 1
6 6 2011-01-06 1 0 1 0 4 1 1
731 rows | 11.97 seconds runtime
Refreshed now
[Shift+Enter] to run and move to next cell
[Esc H] to see all keyboard shortcuts


Chrome File Edit View History Bookmarks Profiles Tab Window Help
Wed 21 Feb
Data Engineering training - B
Meeting | Microsoft Team
MakeMyLabs
hexa-deb-1071 - Microsoft A
Untitled Notebook 2024-02-
adb-5511064555315178.18.azuredatabricks.net/?o=5511064555315178#notebook/1709606852530850/command/1709606852530851
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
⌘ + P
hexa-deb-1071
azuser1071_mml.local@iihtl.or
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Experiments
Features
Models
Untitled Notebook 2024-02-21 10:31:47
Python
File Edit View Run Help
Last edit was now
New cell UI: ON
Run all
azuser1071_mml.local...
Schedule
Share
4 minutes ago (12s)
Cell 1
Python
sparkDF=spark.read.csv("/databricks-datasets/bikeSharing/data-001/day.csv",header="true",inferSchema="true")
display(sparkDF)
(3) Spark Jobs
sparkDF: pyspark.sql.dataframe.DataFrame = [instant: integer, dteday: date ... 14 more fields]
Table
Visualization 1
New charts: ON
workingday
0
1
casual
3000
2000
1000
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
atemp