Name: Krushnakumar Patle

Email: krishnapatle128@gmail.com

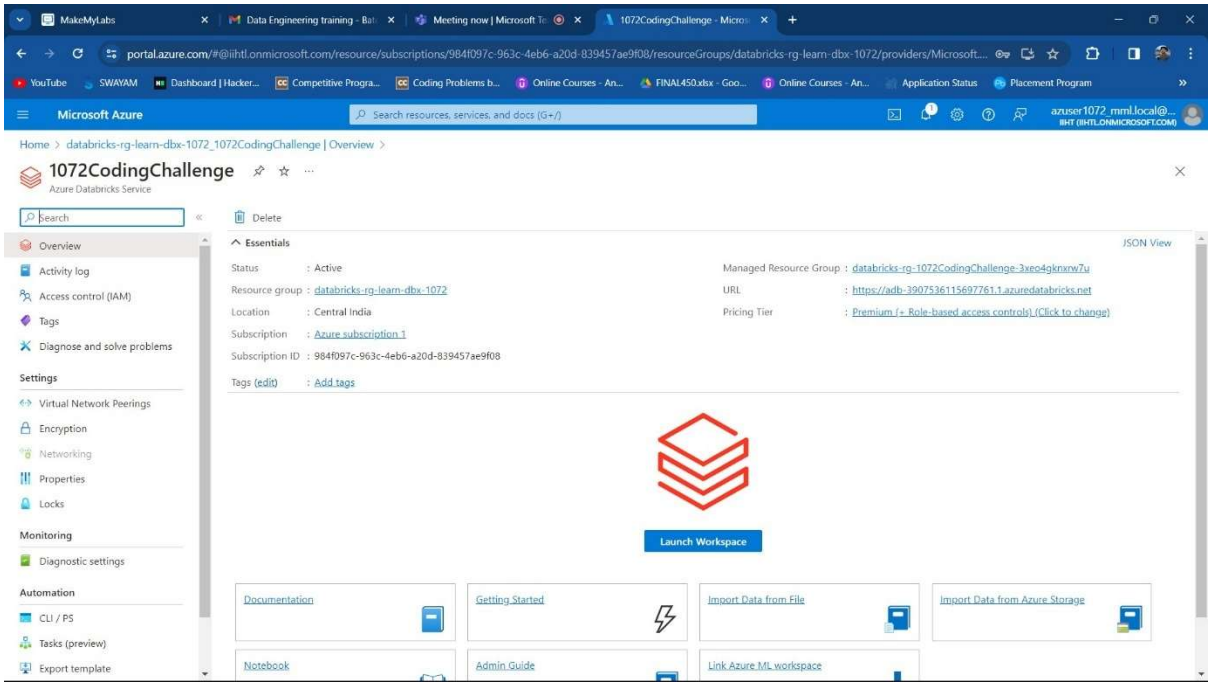
Batch: Data Engineering Batch-1

Azure Databricks Coding Challenge

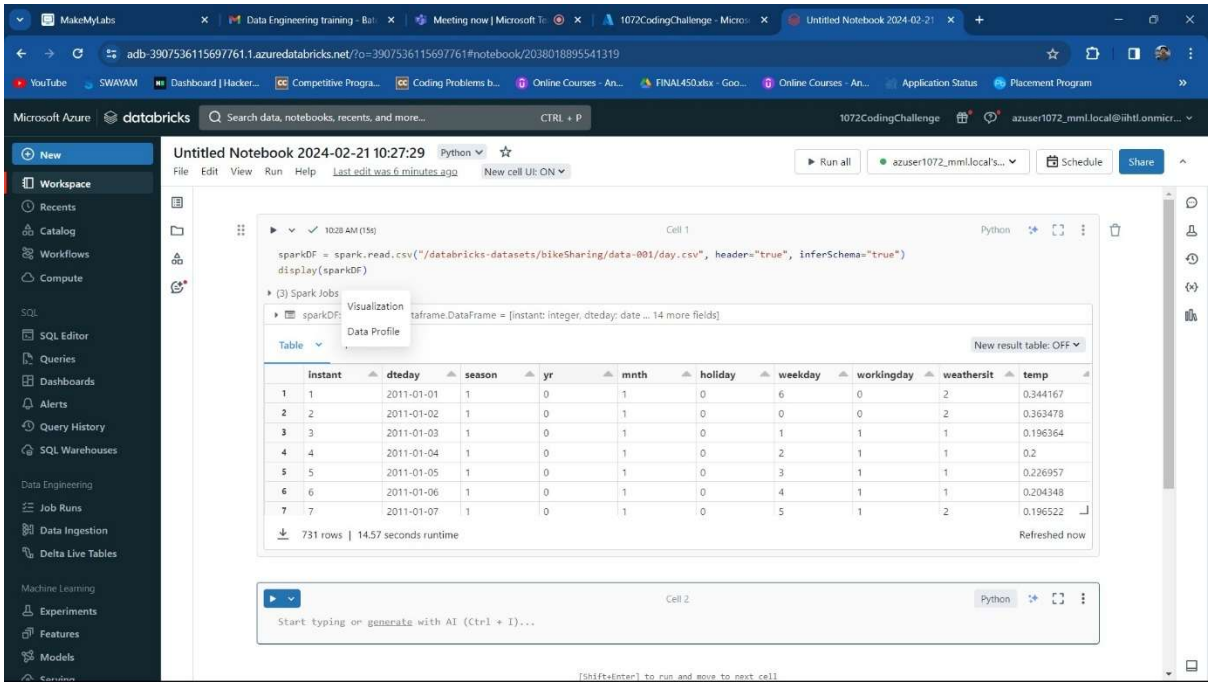
Q1. Exploratory data analysis (EDA) in Databricks &Visualizing data in Databricks

- Exploratory data analysis (EDA) includes methods for exploring data sets to summarize their main characteristics and identify any problems with the data. Using statistical methods and visualizations, you can learn about a data set to determine its readiness for analysis and inform what techniques to apply for data preparation. EDA can also influence which algorithms you choose to apply for training ML models.
- Azure Databricks has built-in analysis and visualization tools in both Databricks SQL and in Databricks Runtime.
- The goal of EDA is to understand the data's characteristics, identify patterns, and outliers, and generate hypotheses for further analysis or modeling. EDA is crucial for making informed decisions in data-driven projects.

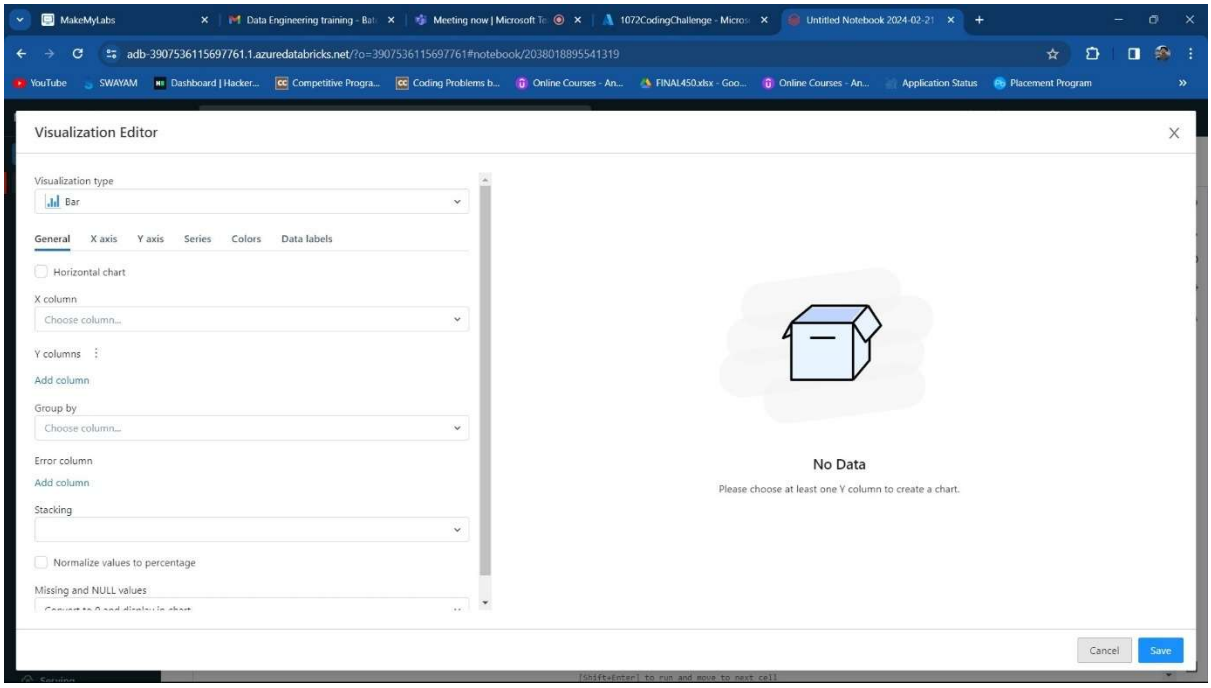
1. Create azure databricks



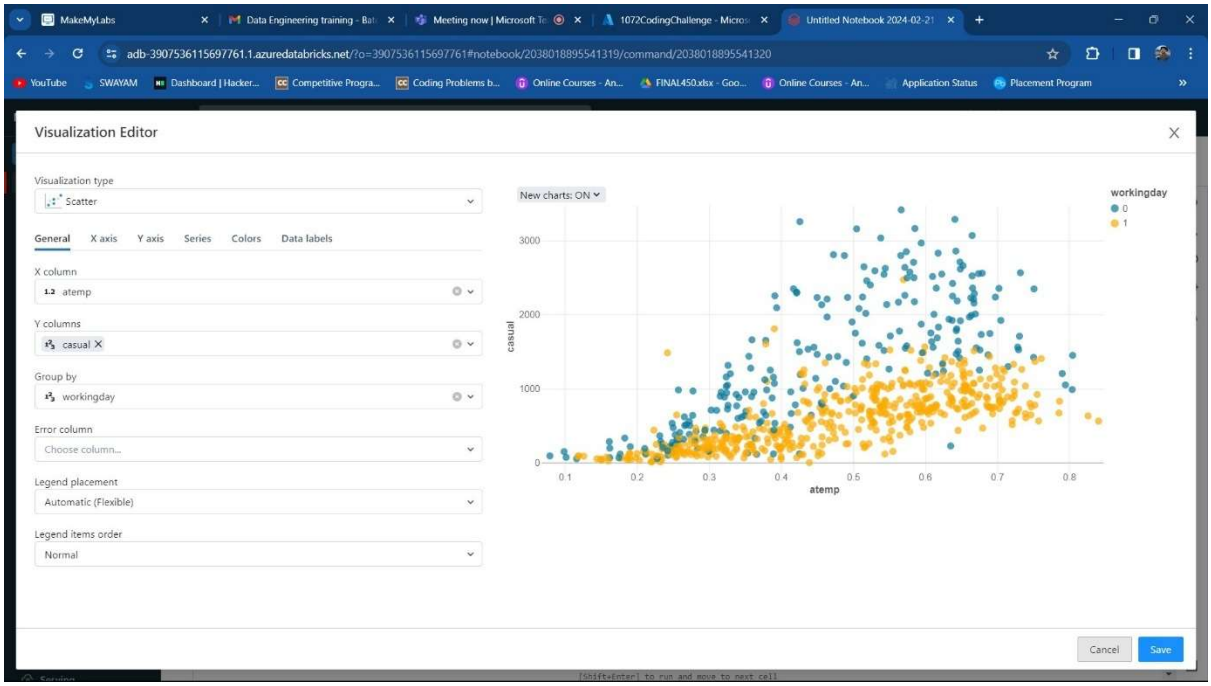
2. Load the data for visualization



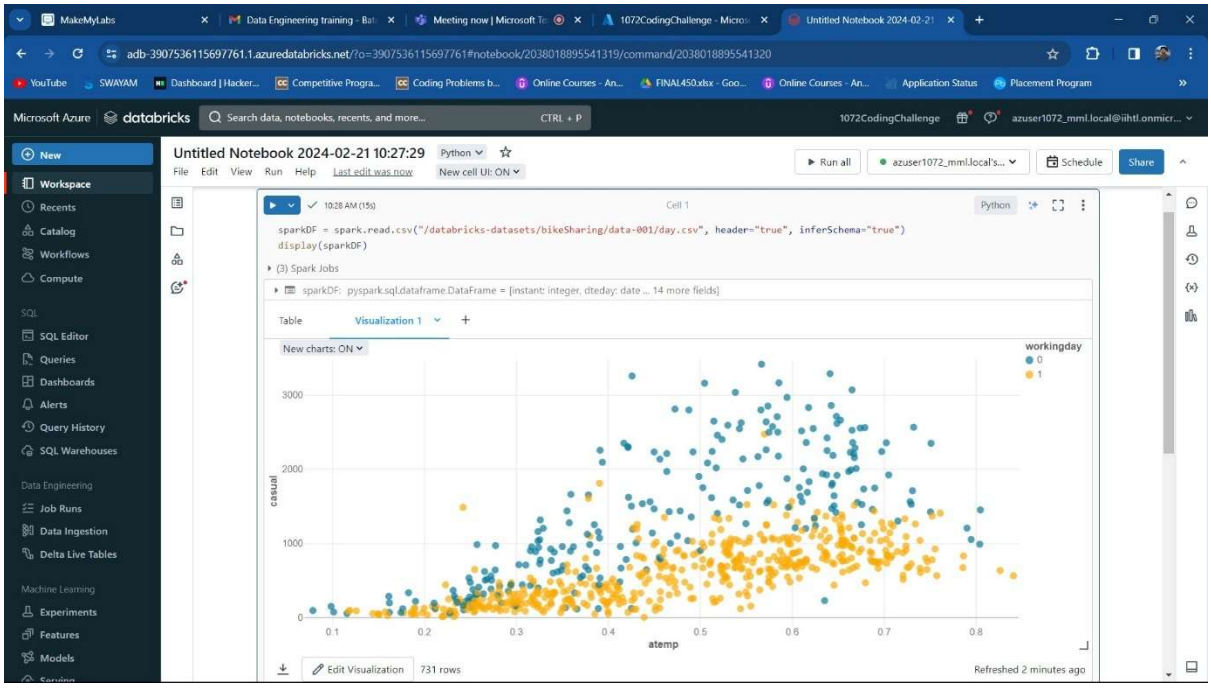
3. Select visualize option to visualize data



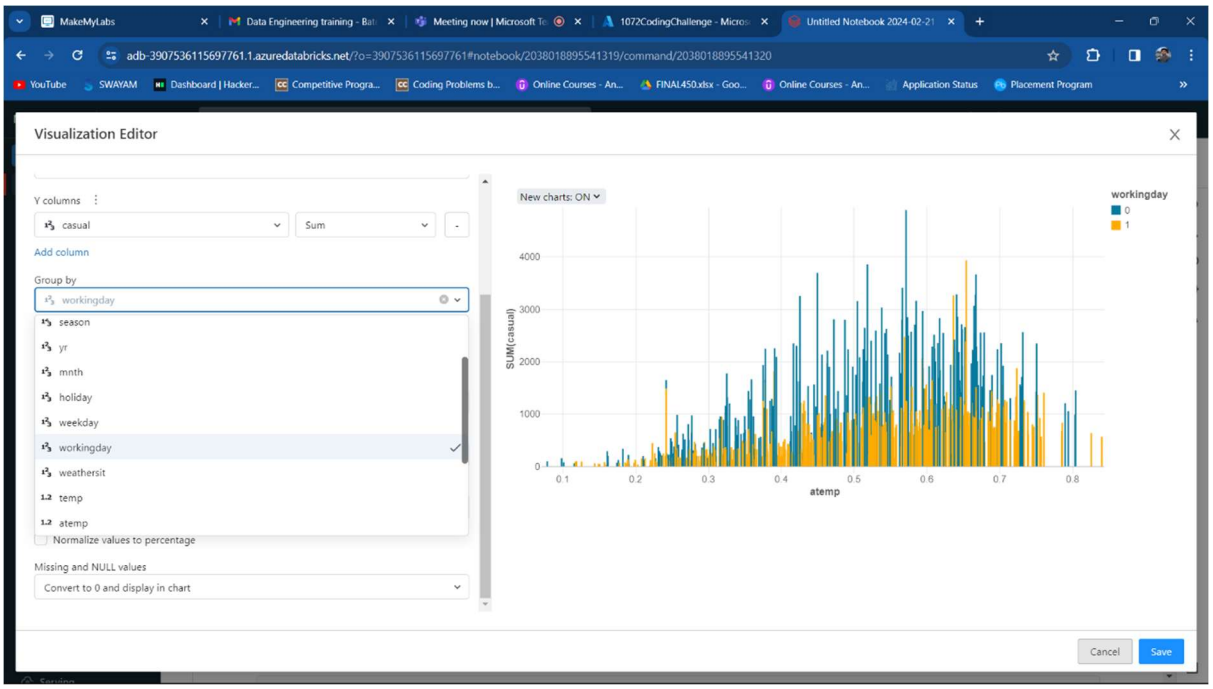
4. Select visulaize type which you want and give x and y values

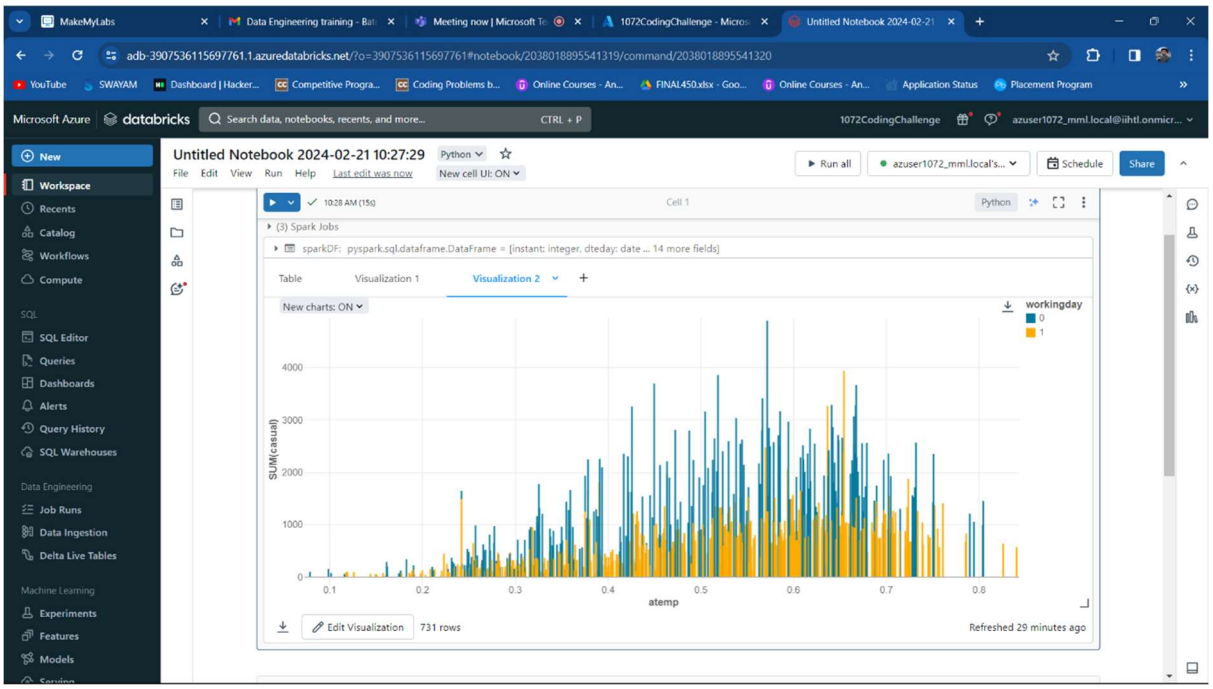


5. Your visualization graph is ready now



6. This is bar visualization of data file


Visualization Editor
Y columns
casual
Sum
Add column
Group by
workingday
season
yr
mnth
holiday
weekday
workingday
weathersit
temp
atemp
Normalize values to percentage
Missing and NULL values
Convert to 0 and display in chart
New charts: ON
workingday
0
1
SUM(casual)
atemp
Cancel
Save


Microsoft Azure
databricks
Search data, notebooks, recents, and more...
Untitled Notebook 2024-02-21 10:27:29
Python
File Edit View Run Help
Last edit was now
New cell UI: ON
Run all
Schedule
Share
Workspace
Recents
Catalog
Workflows
Compute
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Job Runs
Data Ingestion
Delta Live Tables
Experiments
Features
Models
(3) Spark Jobs
sparkDF: pyspark.sql.dataframe.DataFrame = [instant: integer, dteday: date ... 14 more fields]
Table
Visualization 1
Visualization 2
New charts: ON
workingday
SUM(casual)
atemp
Edit Visualization
731 rows
Refreshed 29 minutes ago