

Name: Krushnakumar Patle

Email: [krishnapatle128@gmail.com](mailto:krishnapatle128@gmail.com)

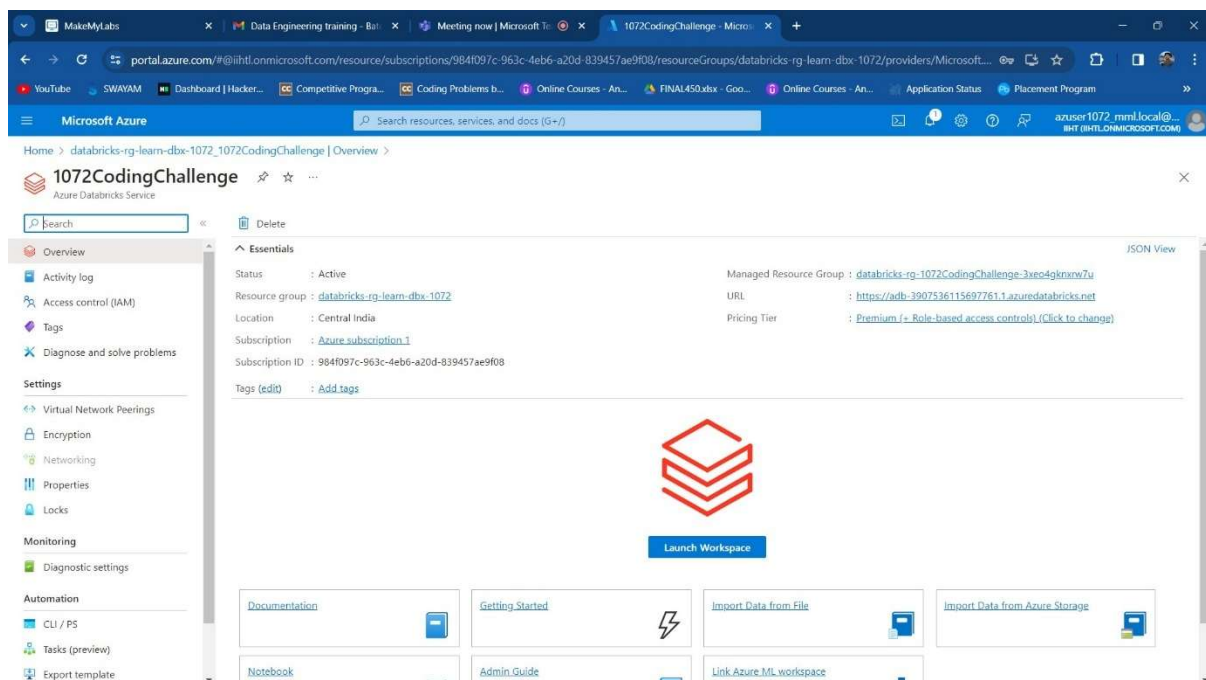
Batch: Data Engineering Batch-1

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

The screenshot shows the Databricks workspace interface. A notebook titled "Untitled Notebook 2024-02-21 10:27:29" is open. The left sidebar contains a navigation pane with options like Workspace, Recents, Catalog, Workflows, Compute, SQL Editor, Queries, Dashboards, Alerts, Query History, SQL Warehouses, Data Engineering, Job Runs, Data Ingestion, Delta Live Tables, Machine Learning, Experiments, Features, and Models. The main area displays a Python cell with the following code:

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

Below the code, a table visualization is shown with the following columns: Instant, dteday, season, yr, mnth, holiday, weekday, workingday, weathersit, and temp. The table contains 7 rows of data.

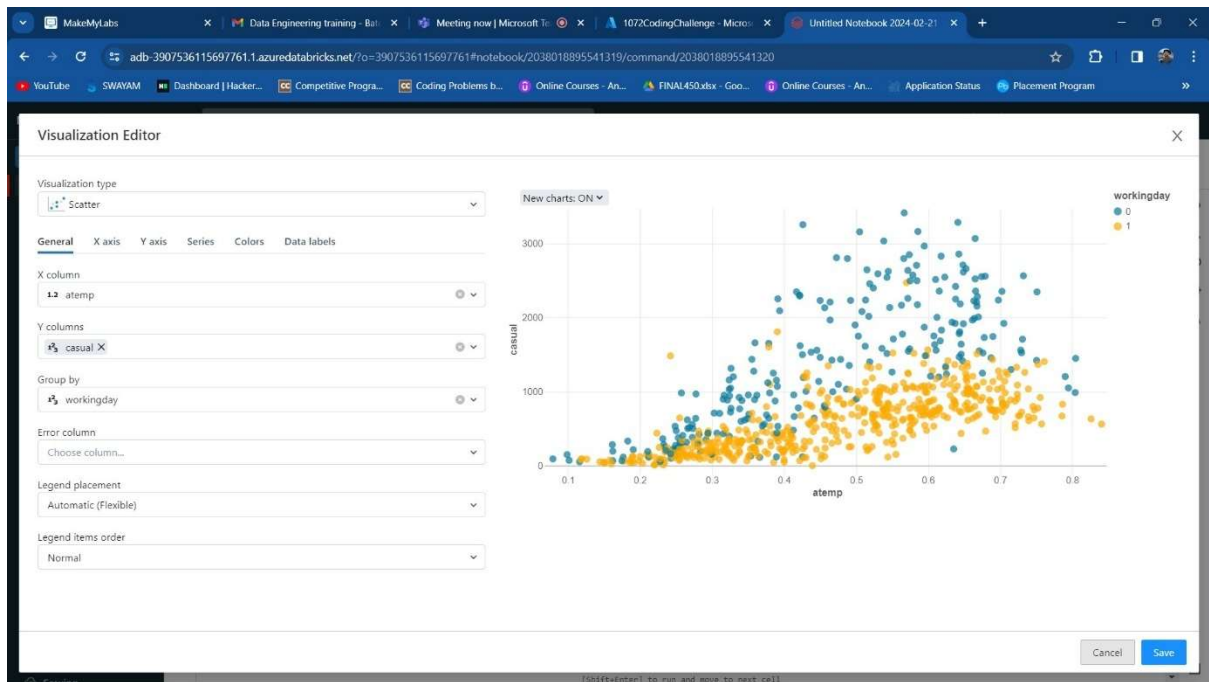
	Instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp
1	1	2011-01-01	1	0	1	0	6	0	2	0.344167
2	2	2011-01-02	1	0	1	0	0	0	2	0.363478
3	3	2011-01-03	1	0	1	0	1	1	1	0.196364
4	4	2011-01-04	1	0	1	0	2	1	1	0.2
5	5	2011-01-05	1	0	1	0	3	1	1	0.226957
6	6	2011-01-06	1	0	1	0	4	1	1	0.204348
7	7	2011-01-07	1	0	1	0	5	1	2	0.196522

The table has 731 rows and 1457 seconds runtime. A "Visualization" menu is open, showing options for "Table" and "Data Profile".

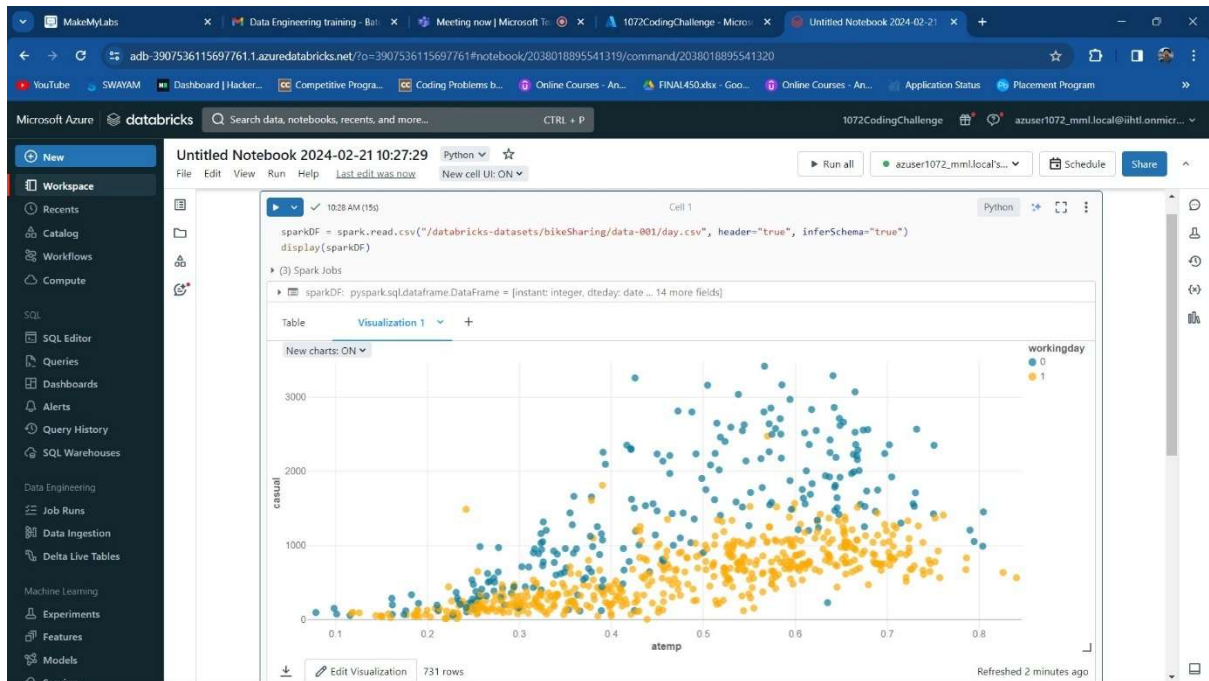
### 3. Select visualize option to visualize data

The screenshot shows the "Visualization Editor" window. The "Visualization type" is set to "Bar". The "General" tab is selected, showing options for "Horizontal chart", "X column", "Y columns", "Group by", "Error column", "Stacking", "Normalize values to percentage", and "Missing and NULL values". The "X column" is set to "Choose column..." and the "Y columns" are set to "Add column". The "Group by" is set to "Choose column...". The "Error column" is set to "Add column". The "Stacking" is set to "Choose column...". The "Normalize values to percentage" checkbox is unchecked. The "Missing and NULL values" section shows "Choose column..." and "Choose column..." options. The main area displays a "No Data" message with a box icon and the text "Please choose at least one Y column to create a chart.".

### 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

