Name : Parth Nandedkar
Date : 21 Feb 2024
Topics : Azure Databricks
Batch : Data Engineering Batch-1

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

Exploratory Data Analysis (EDA) is a crucial step in the data analysis process, and Databricks provides a powerful platform for performing EDA and visualising data. Here's a general guide on how you can perform EDA and visualise data using Databricks:

Loading Data:

Begin by loading your dataset into Databricks. You can do this from various sources such as Azure Data Lake Storage, Azure Blob Storage, Azure SQL Database, etc.
Databricks supports multiple file formats like CSV, Parquet, JSON, etc. You can use the appropriate reader to load your data into a DataFrame.

Understanding Data:

Once the data is loaded, you can use DataFrame operations to explore its structure and contents. Methods like display, show, describe, schema, etc., can be helpful.
Check for missing values, data types, summary statistics, unique values, etc., to get a better understanding of your data.

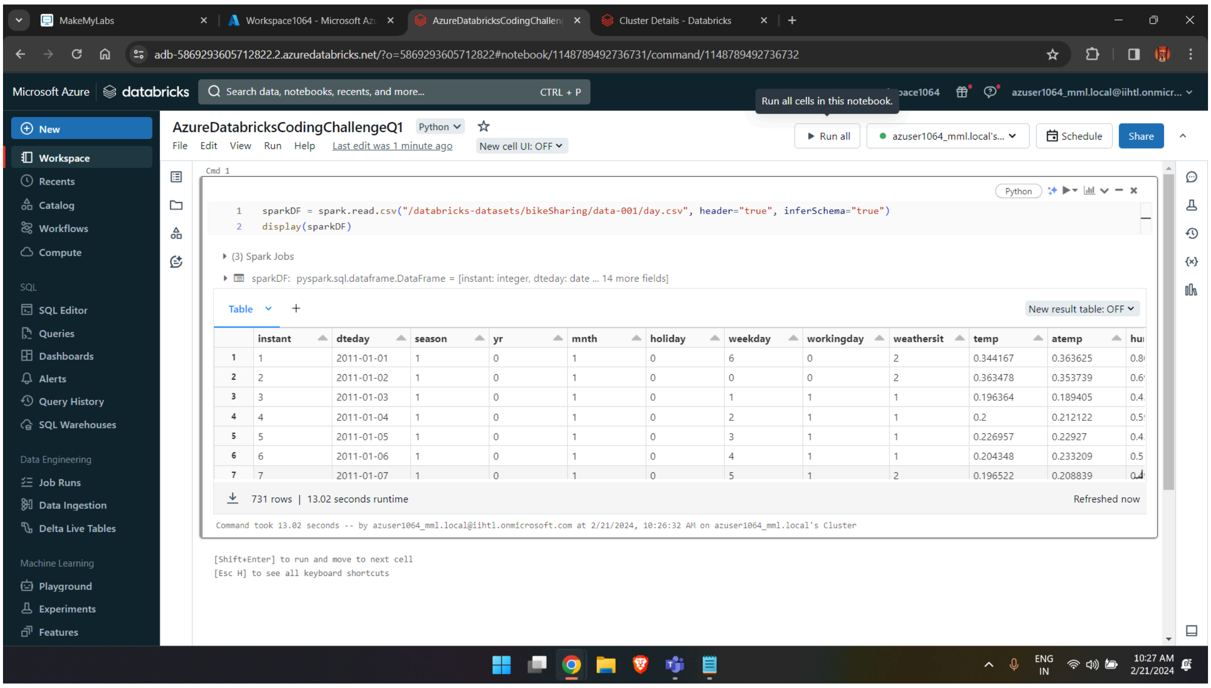
Data Visualization:

Databricks supports various visualisation libraries such as Matplotlib, Seaborn, Plotly, etc., which you can use directly in your notebooks.
You can create different types of plots like histograms, scatter plots, bar plots, line plots, etc., to visualise the distribution, relationships, and patterns in your data.

After creating a cluster and creating a new notebook.
By using PySpark using **spark.read.csv()** we have stored/converted the csv data which was already present in the databricks workspace into dataframe.

sparkDF is a data frame which stores all the values.

display() is the method to see whether the data is stored or not.



MakeMyLabs
Workspace1064 - Microsoft Az
AzureDatabricksCodingChalle
Cluster Details - Databricks
adb-5869293605712822.2.azuredatabricks.net/?o=5869293605712822#notebook/1148789492736731/command/1148789492736732
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
Run all cells in this notebook.
pace1064
azuser1064_mml.local@iihtl.onmicr...
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
Playground
Experiments
Features
AzureDatabricksCodingChallengeQ1
Python
File
Edit
View
Run
Help
Last edit was 1 minute ago
New cell UI: OFF
Run all
azuser1064_mml.local's...
Schedule
Share
Cmd 1
Python
Table
New result table: OFF
ENG
IN
10:27 AM
2/21/2024

```
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]

	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hu
1	1	2011-01-01	1	0	1	0	6	0	2	0.344167	0.363625	0.8
2	2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739	0.6
3	3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405	0.4
4	4	2011-01-04	1	0	1	0	2	1	1	0.2	0.212122	0.5
5	5	2011-01-05	1	0	1	0	3	1	1	0.226957	0.22927	0.4
6	6	2011-01-06	1	0	1	0	4	1	1	0.204348	0.233209	0.5
7	7	2011-01-07	1	0	1	0	5	1	2	0.196522	0.208839	0.4

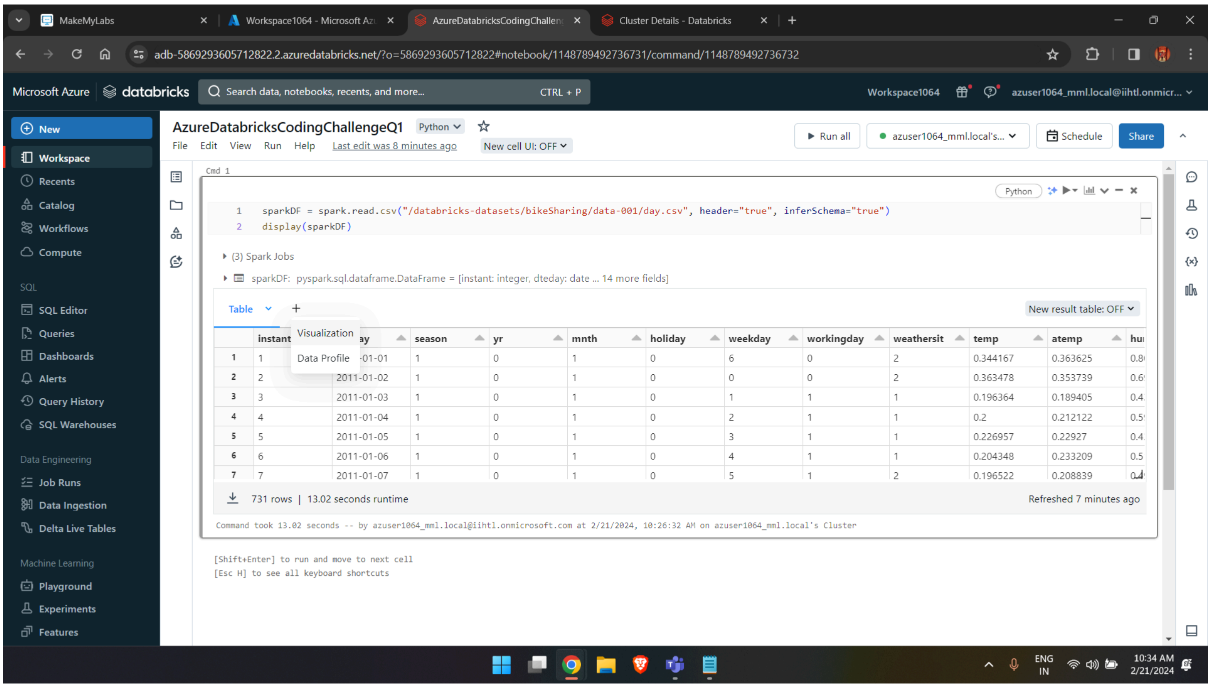
731 rows | 13.02 seconds runtime

Refreshed now

```
Command took 13.02 seconds -- by azuser1064_mml.local@iihtl.onmicrosoft.com at 2/21/2024, 10:26:32 AM on azuser1064_mml.local's Cluster
```

```
[Shift+Enter] to run and move to next cell
[Esc H] to see all keyboard shortcuts
```

As per the above screenshot we can say that the data is stored in the data frames and we can use this data for visualisation purposes.



AzureDatabricksCodingChallengeQ1

Cmd 1

```
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 | Data Profile | New result table: OFF

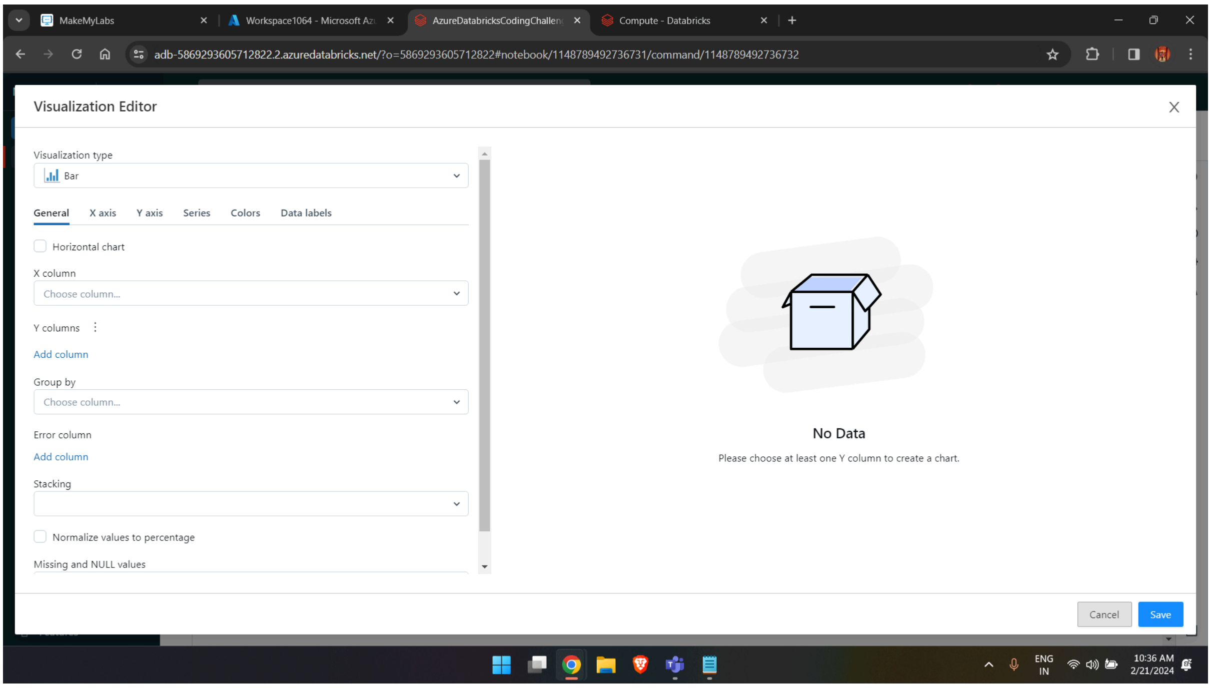
	instant	ay	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hu
1	1	-01-01	1	0	1	0	6	0	2	0.344167	0.363625	0.8
2	2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739	0.6
3	3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405	0.4
4	4	2011-01-04	1	0	1	0	2	1	1	0.2	0.212122	0.5
5	5	2011-01-05	1	0	1	0	3	1	1	0.226957	0.22927	0.4
6	6	2011-01-06	1	0	1	0	4	1	1	0.204348	0.233209	0.5
7	7	2011-01-07	1	0	1	0	5	1	2	0.196522	0.208839	0.4

731 rows | 13.02 seconds runtime | Refreshed 7 minutes ago

```
Command took 13.02 seconds -- by azuser1064_mml.local@iihtl.onmicrosoft.com at 2/21/2024, 10:26:32 AM on azuser1064_mml.local's Cluster
```

```
[Shift+Enter] to run and move to next cell
[Esc H] to see all keyboard shortcuts
```

On table creation there is an option named as Visualization where there are options like below



Visualization Editor

Visualization type: Bar

General | X axis | Y axis | Series | Colors | Data labels

□ Horizontal chart

X column: Choose column...

Y columns: Add column

Group by: Choose column...

Error column: Add column

Stacking

□ Normalize values to percentage

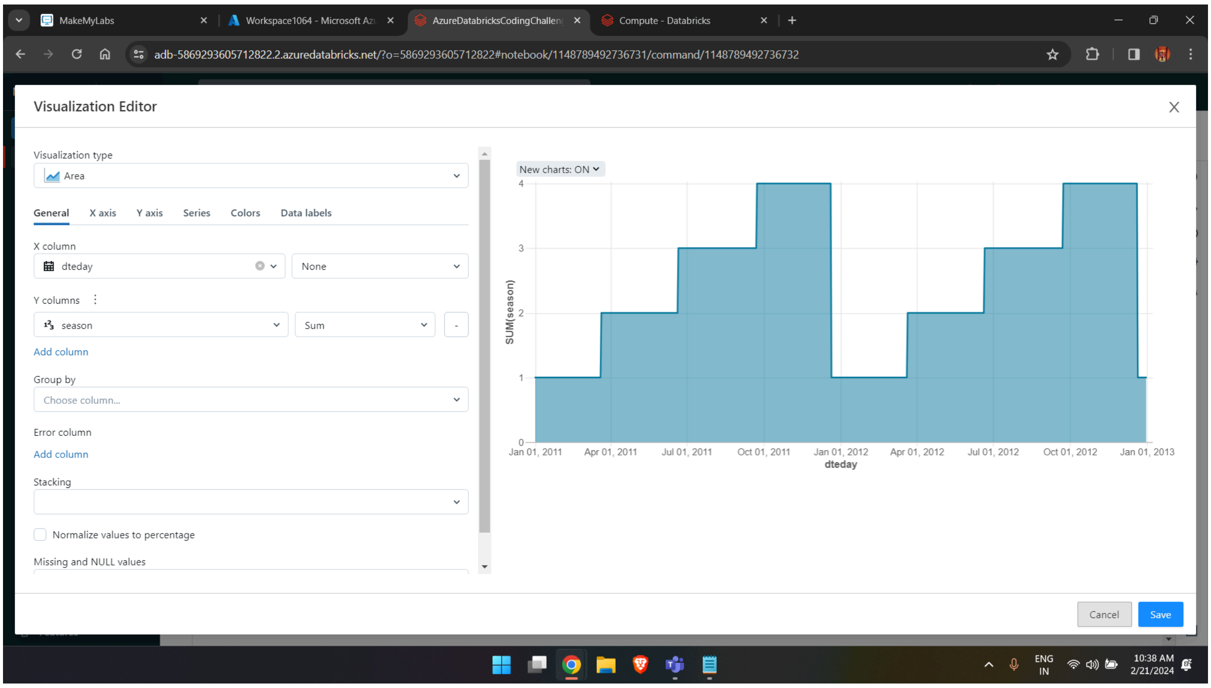
Missing and NULL values

No Data

Please choose at least one Y column to create a chart.

Cancel | Save

Now according to data we can use the different visualisations so as to analyse the data.



Visualization Editor

Visualization type: Area

General | X axis | Y axis | Series | Colors | Data labels

X column: dteday | None

Y columns: season | Sum

Add column

Group by: Choose column...

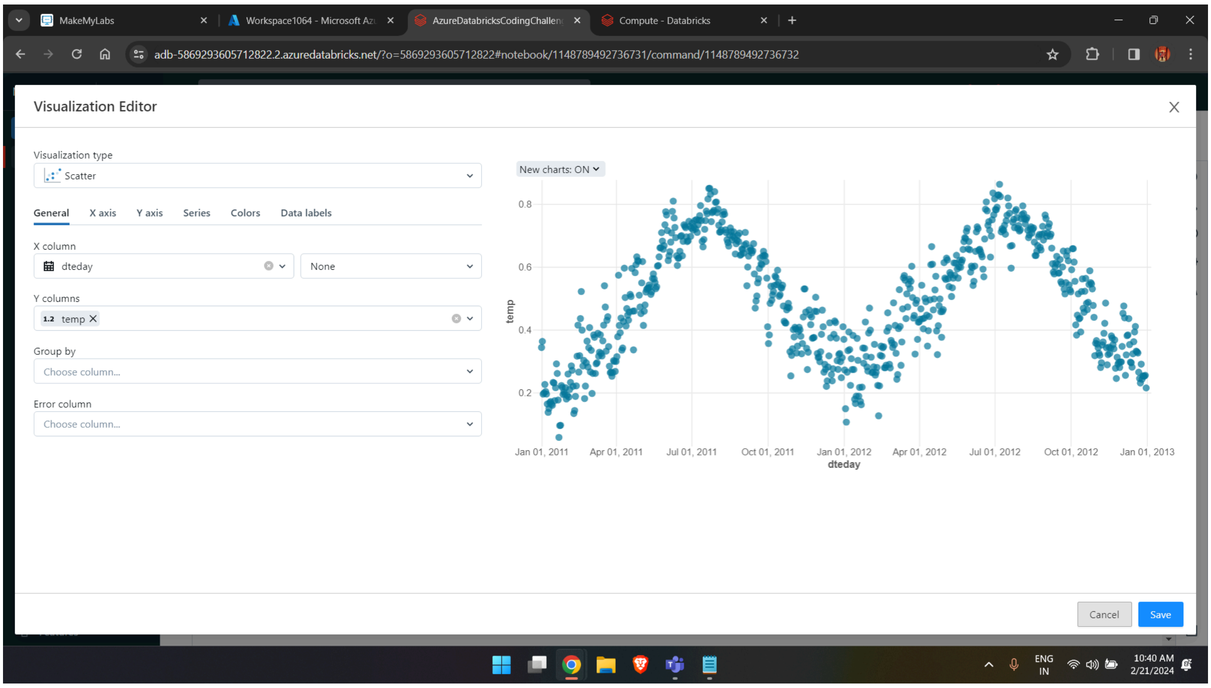
Error column: Add column

Stacking:

□ Normalize values to percentage

Missing and NULL values

This is the area wise graph of the every day present in the data and based on season.



Visualization Editor

Visualization type: Scatter

General | X axis | Y axis | Series | Colors | Data labels

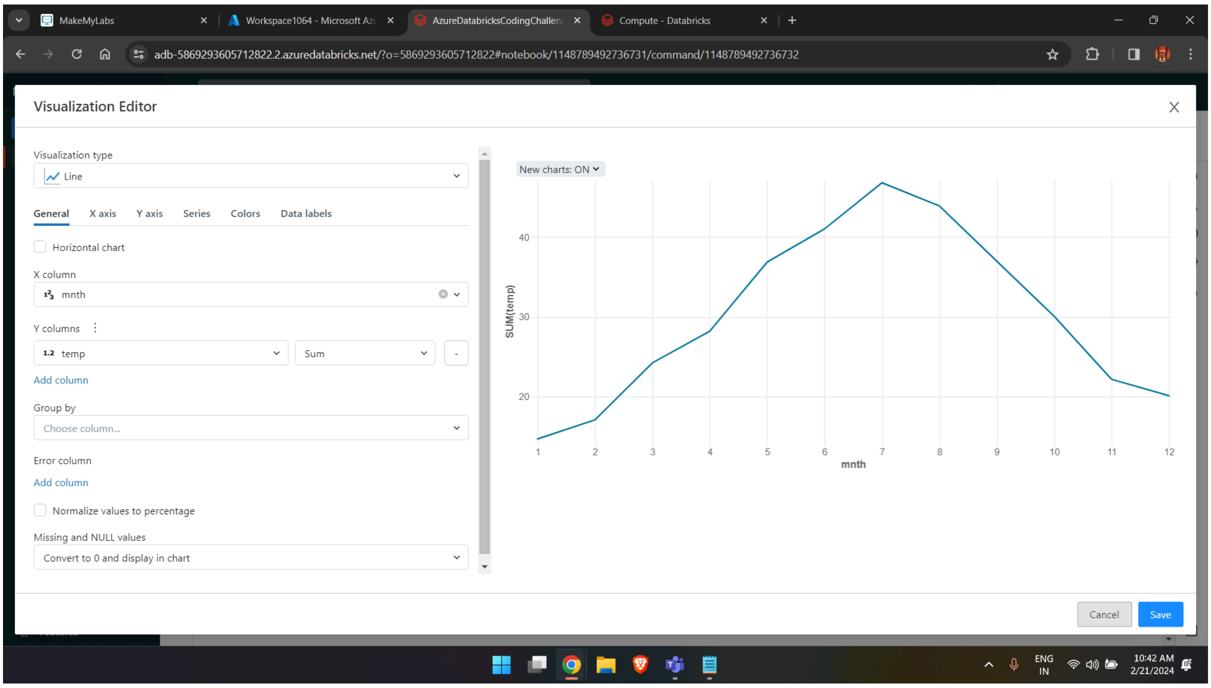
X column: dteday | None

Y columns: 1.2 temp ✕

Group by: Choose column...

Error column: Choose column...

Above is the scattered graph of per day according to temperature.


MakeMyLabs
Workspace1064 - Microsoft Az
AzureDatabricksCodingChallen
Compute - Databricks
adb-5869293605712822.2.azuredatabricks.net/?o=5869293605712822#notebook/1148789492736731/command/1148789492736732
New charts: ON
SUM(temp)
40
30
20
1
2
3
4
5
6
7
8
9
10
11
12
mnth
10:42 AM
2/21/2024
ENG
IN

Visualization Editor

Visualization type: Line

General | X axis | Y axis | Series | Colors | Data labels

☐ Horizontal chart

X column: mnth

Y columns: temp | Sum

Add column

Group by: Choose column...

Error column

Add column

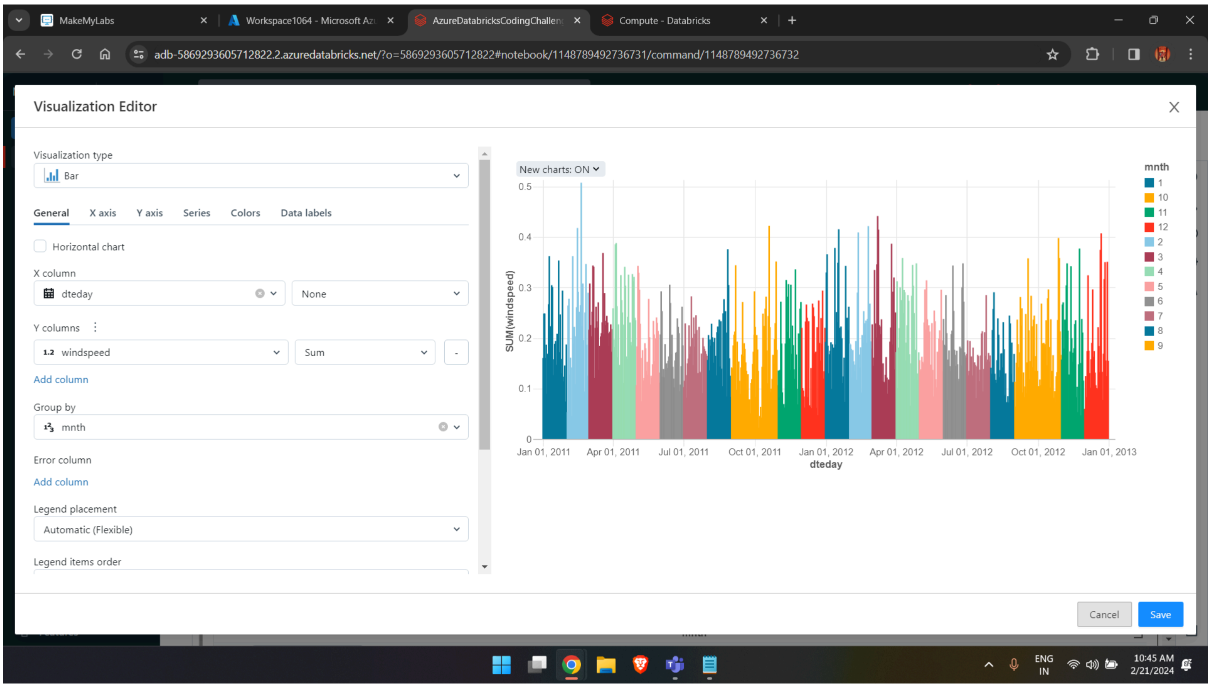
☐ Normalize values to percentage

Missing and NULL values: Convert to 0 and display in chart

Cancel | Save

Above is the line graph for the same.

We can analyse that during month 6-9 the temperature is at its peak and before that temp is rising and after that temperature is falling.



Visualization Editor

Visualization type: Bar

General | X axis | Y axis | Series | Colors | Data labels

☐ Horizontal chart

X column: dteday | None

Y columns: windspeed | Sum

Add column

Group by: mnth

Error column

Add column

Legend placement: Automatic (Flexible)

Legend items order

Cancel | Save

In above visualisation the wind speed can be analysed for every day according to month the colours are different as we have used group by months.