Data Engineering Batch 1 **Date: 21/02/2024**

Name: Pradip Bochare

Azure Databricks Coding Assessment Question 1

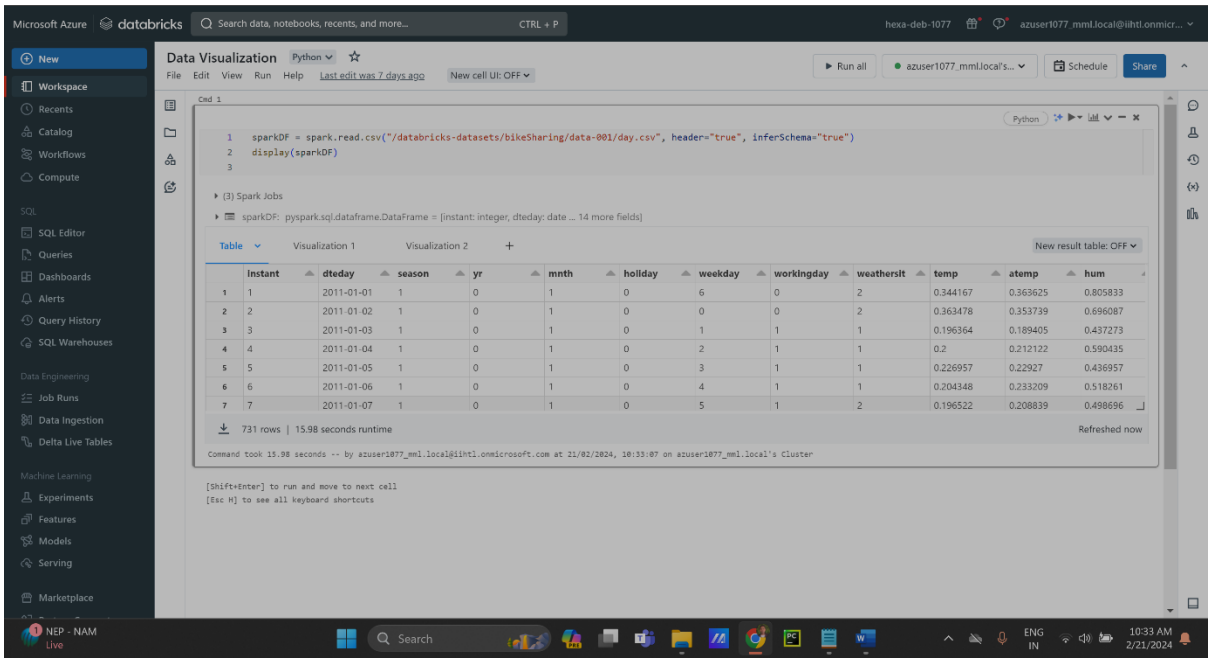
Exploratory data analysis (EDA) in Databricks &Visualizing data in Databricks

❖ Exploratory data analysis (EDA) 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.
- Utilize PySpark for data loading, manipulation, and analysis within Databricks notebooks.
- Leverage PySpark functions such as show(), describe(), and printSchema() to explore the dataset.
- Perform data cleaning and preprocessing using PySpark DataFrame operations like dropna() and withColumn().
- Use Databricks built-in visualizations or external libraries like Matplotlib and Seaborn for data visualization.
- Visualize data distributions, trends, and relationships through histograms, scatter plots, etc. Utilize Seaborn for more advanced visualizations and styling options.
- Databricks provides seamless integration with PySpark, enabling scalable and efficient data analysis workflows.

❖ Visualizing data in Databricks

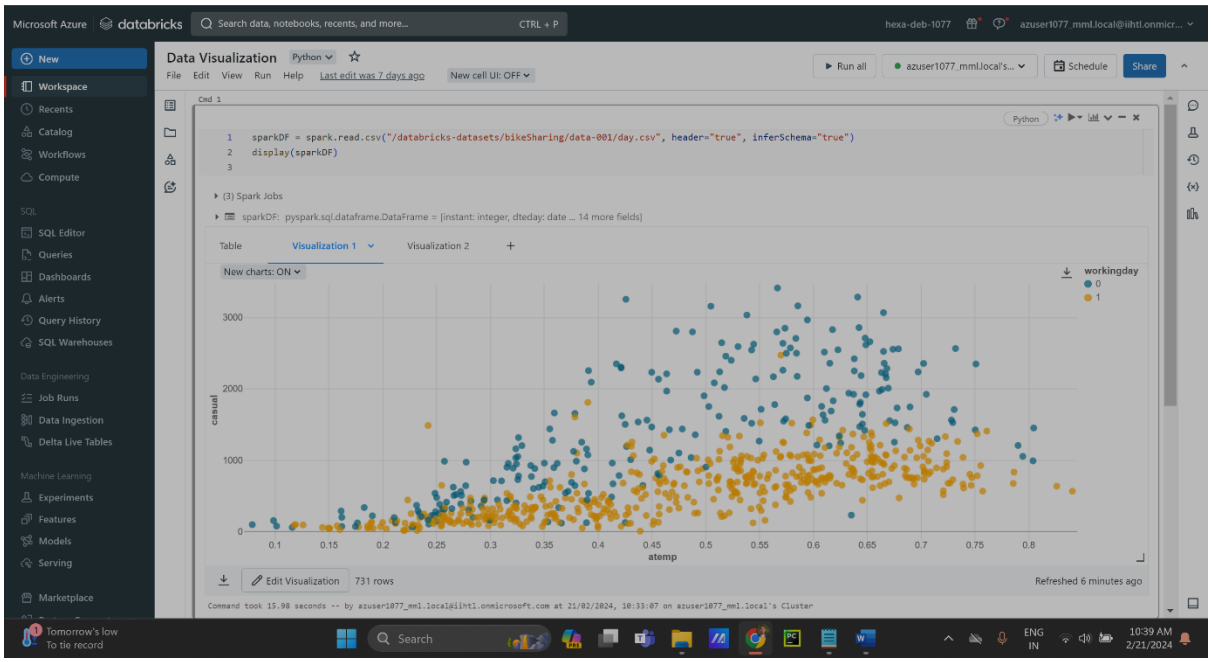
- Here we are using data frame to perform visualization on it

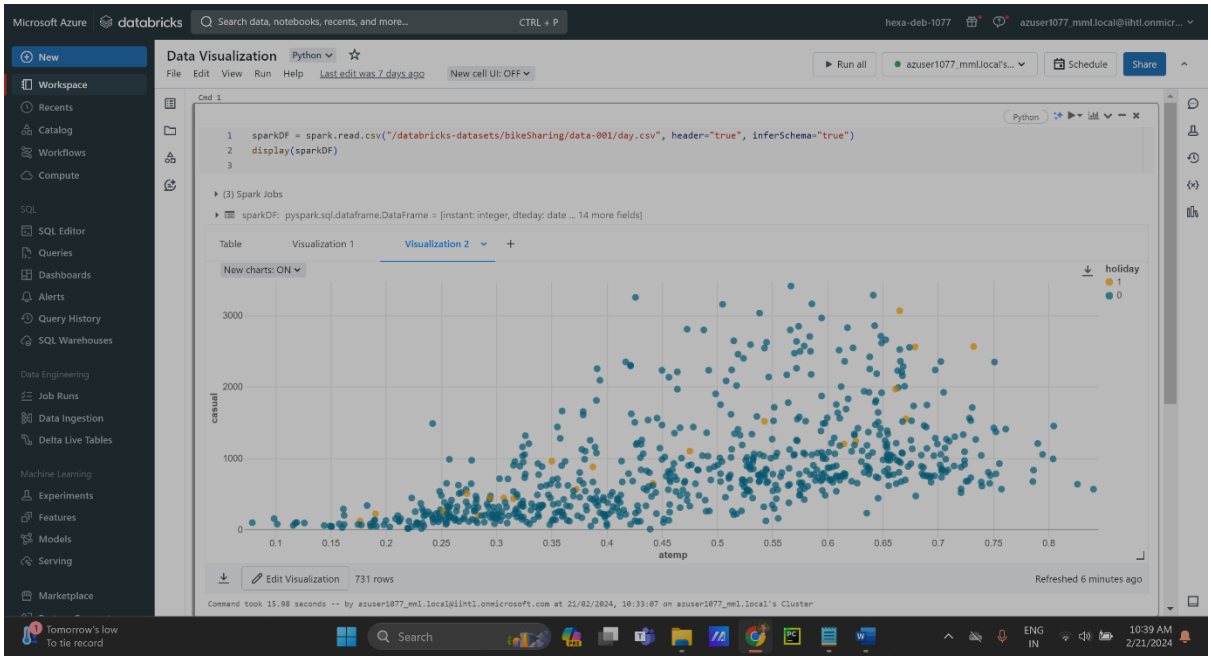


- Giving Parameters for Scatter plot according to columns in table

- Giving Parameters for Bubble plot according to columns in table

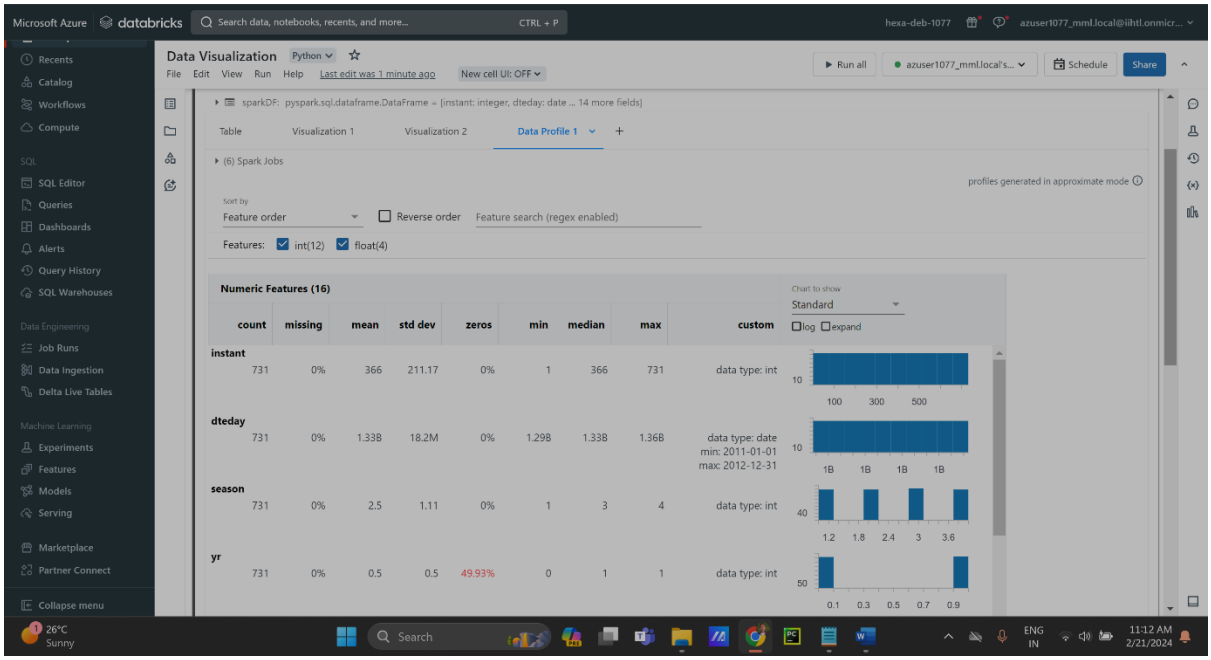
- Visualization In Azure Databricks

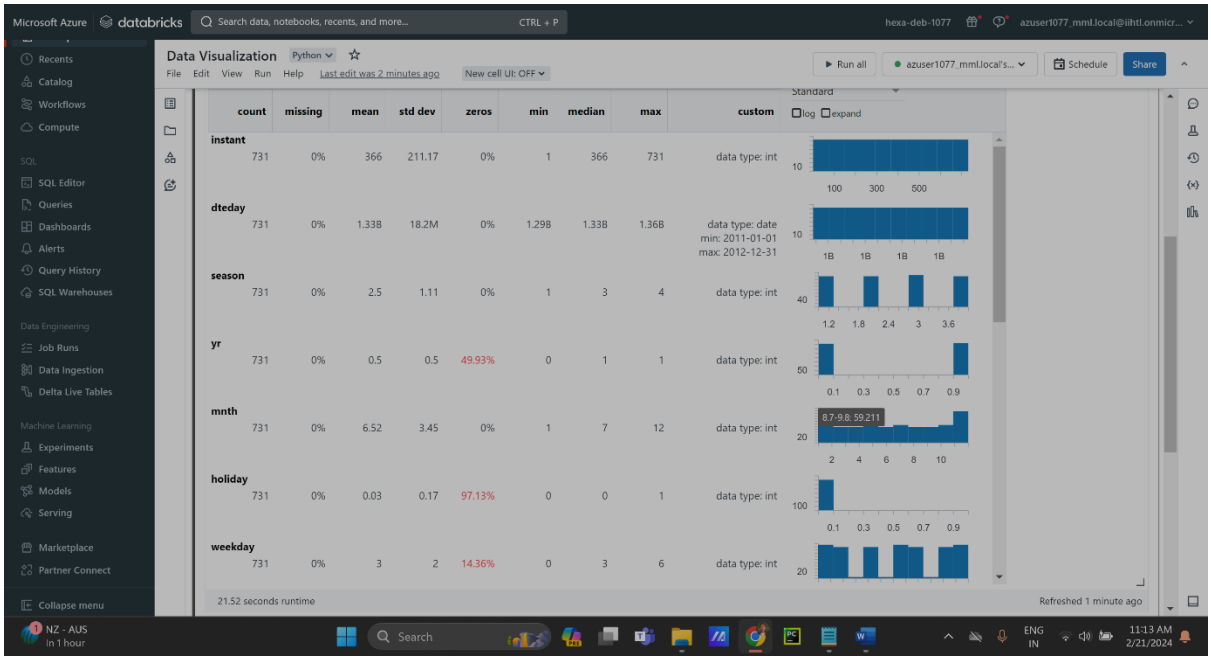


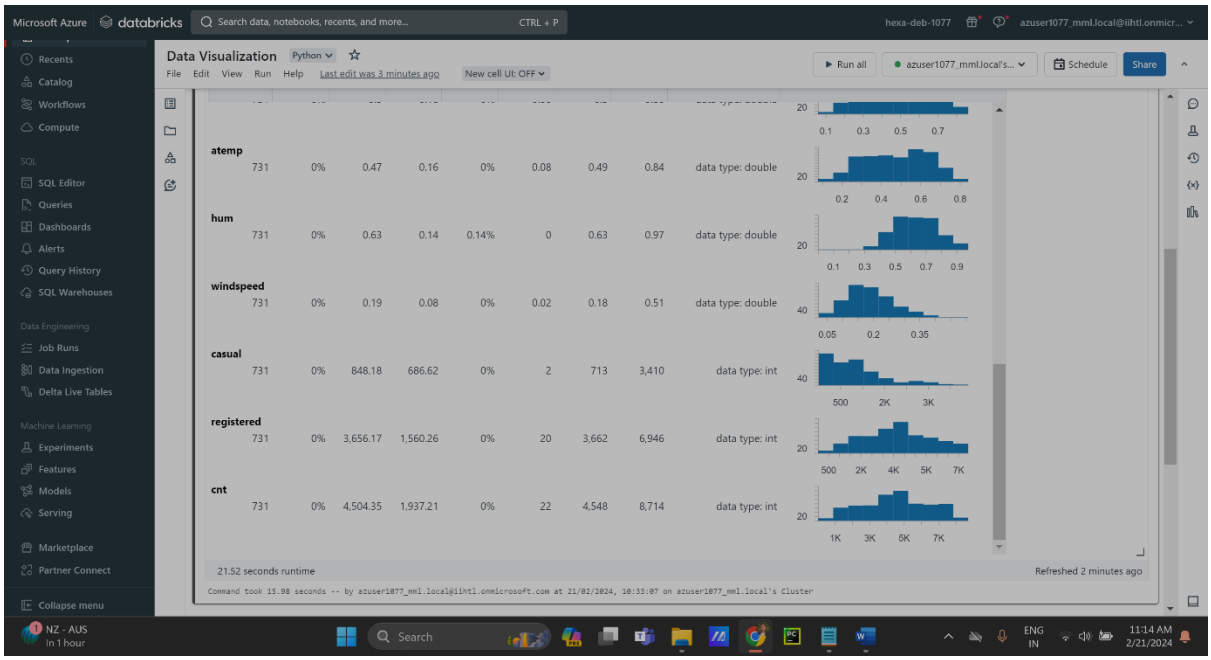


- After creating the visualization, we can also create data profile.

Data profile : Used to analyse the data based on some trends, you can also order and filter the data.




Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
Data Visualization
Python
File Edit View Run Help Last edit was 2 minutes ago
New cell UI: OFF
Run all
Schedule
Share
Standard
log expand
count missing mean std dev zeros min median max custom
instant 731 0% 366 211.17 0% 1 366 731 data type: int
dteday 731 0% 1.33B 18.2M 0% 1.29B 1.33B 1.36B data type: date min: 2011-01-01 max: 2012-12-31
season 731 0% 2.5 1.11 0% 1 3 4 data type: int
yr 731 0% 0.5 0.5 49.93% 0 1 1 data type: int
mnth 731 0% 6.52 3.45 0% 1 7 12 data type: int
holiday 731 0% 0.03 0.17 97.13% 0 0 1 data type: int
weekday 731 0% 3 2 14.36% 0 3 6 data type: int
8.7-9.8: 59.211
21.52 seconds runtime
Refreshed 1 minute ago
NZ - AUS
In 1 hour
Search
ENG
IN
11:13 AM
2/21/2024


Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
Data Visualization
Python
File Edit View Run Help Last edit was 3 minutes ago
New cell UI: OFF
Run all
Schedule
Share
atemp 731 0% 0.47 0.16 0% 0.08 0.49 0.84 data type: double
hum 731 0% 0.63 0.14 0.14% 0 0.63 0.97 data type: double
windspeed 731 0% 0.19 0.08 0% 0.02 0.18 0.51 data type: double
casual 731 0% 848.18 686.62 0% 2 713 3,410 data type: int
registered 731 0% 3,656.17 1,560.26 0% 20 3,662 6,946 data type: int
cnt 731 0% 4,504.35 1,937.21 0% 22 4,548 8,714 data type: int
21.52 seconds runtime
Refreshed 2 minutes ago
Command took 15.98 seconds -- by azuser1077_mml.local@iihtl.onmicrosoft.com at 21/02/2024, 10:33:07 on azuser1077_mml.local's Cluster
NZ - AUS
In 1 hour
Search
ENG
IN
11:14 AM
2/21/2024