

# Azure Databricks Coding Challenge

Pratik Wani

## Question 1

### ❖ Exploratory data analysis (EDA) in Databricks and Visualizing data in Databricks

- Exploratory Data Analysis (EDA) in Databricks refers to the process of exploring datasets to understand their structure, patterns, and gain insights into the data.
- Different Steps in EDA
  - Data Upload
  - Data Reading
  - Data Statistics
  - Data Visualization
- I created the workspace and cluster in the azure databricks and uploaded a csv file for EDA and Data Visualization

## ○ EDA tasks:

CodingChallenge-EDA and DV 2024-02-21 10:28:42 Python ☆

File Edit View Run Help Last edit was 2 minutes ago New cell UI: OFF

Run all Pratik Wani's Cluster Schedule Share

Cmd 1

1 from pyspark.sql import SparkSession

Command took 0.54 seconds -- by pratikwani116@outlook.com at 2/21/2024, 10:30:55 AM on Pratik Wani's Cluster

Cmd 2

1 spark=SparkSession.Builder().appName('CodingChallenge').getOrCreate()

Command took 0.25 seconds -- by pratikwani116@outlook.com at 2/21/2024, 10:31:29 AM on Pratik Wani's Cluster

Cmd 3

1 # EDA - CSV file Upload  
2  
3 dataframe=spark.read.csv("/FileStore/tables/CCsv.csv")

▶ (1) Spark Jobs  
▶ dataframe: pyspark.sql.dataframe.DataFrame = [\_c0: string, \_c1: string ... 2 more fields]

Command took 13.95 seconds -- by pratikwani116@outlook.com at 2/21/2024, 10:33:47 AM on Pratik Wani's Cluster

Cmd 4

CodingChallenge-EDA and DV 2024-02-21 10:28:42 Python ☆

File Edit View Run Help Last edit was 3 minutes ago New cell UI: OFF

Run all Pratik Wani's Cluster Schedule Share

Command took 13.95 seconds -- by pratikwani116@outlook.com at 2/21/2024, 10:33:47 AM on Pratik Wani's Cluster

Cmd 4

1 # EDA - Data Analysis  
2  
3 dataframe.show()

▶ (1) Spark Jobs

_c0	_c1	_c2	_c3
Country	Population (milli...	GDP per capita (USD)	Life expectancy (...)
USA	328.2	65246	78.9
China	1439.3	10261	76.9
Japan	126.5	40849	84.6
Germany	83.2	48039	81.2
India	1380.0	2155	69.7
UK	68.2	42943	81.3
France	65.3	41463	82.3
Italy	60.4	34483	83.5
Brazil	211.0	8824	75.9
Canada	38.0	46462	82.3
Australia	25.4	55871	83.9
South Korea	51.8	32775	83.5

Command took 0.73 seconds -- by pratikwani116@outlook.com at 2/21/2024, 10:34:11 AM on Pratik Wani's Cluster

Cmd 5



Cmd 5

```
1 # EDA - Statistical analysis
2
3 dataframe.describe().show()
```

(2) Spark Jobs

summary	_c0	_c1	_c2	_c3
count	13	13	13	13
mean	NULL	323.10833333333335	35780.916666666664	80.33333333333333
stddev	NULL	514.8427093577104	19458.065882854728	4.337643544039956
min	Australia	126.5	10261	69.7
max	USA	Population (milli...	GDP per capita (USD)	Life expectancy (...)

Command took 3.25 seconds -- by pratikwani116@outlook.com at 2/21/2024, 10:37:04 AM on Pratik Wani's Cluster

Cmd 6

## ○ Data Visualization:

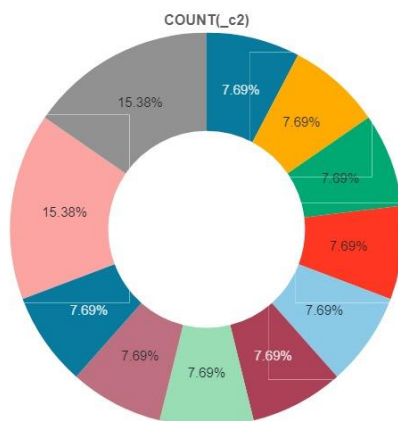


```
1 # EDA - Data Visualization
2
3 display(dataframe)
4
```

(1) Spark Jobs

Table Visualization 1 +

New charts: ON



\_c3

- 83.5
- 82.3
- 83.9
- 75.9
- 81.3
- 69.7
- 81.2
- 84.6
- 76.9
- 78.9
- Life expectancy (years)



Edit Visualization

13 rows

Refreshed 3 minutes ago