

# Coding challenge

## Azure Databricks

Name: Rohan Vinayak Chaudhari

Email: [chaudharirohan24@gmail.com](mailto:chaudharirohan24@gmail.com)

Batch: Data Engineering 1

**Question1:** Exploratory data analysis (EDA) in Databricks & Visualizing data in Databricks.

**Spark Session:**

**Uploaded a Titanic dataset for EDA:**

The screenshot displays the Microsoft Azure Databricks workspace. The left sidebar contains navigation options: 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, and Machine Learning. The main area shows a notebook titled 'EDA 2024-02-21 10:13:03' with a Python language selector. The notebook contains two code cells. Cell 1, executed at 10:13 AM, shows the Spark session details: SparkSession - hive, SparkContext, Spark UI, Version v3.5.0, Master local[\*], 4], and AppName Databricks Shell. Cell 2, executed at 10:24 AM, contains the code `df=spark.read.csv('/FileStore/tables/output_file.csv', header=True, inferSchema=True)` and its output: (2) Spark Jobs, df: pyspark.sql.dataframe.DataFrame = [PassengerId: integer, Survived: integer ... 10 more fields].

## Viewing the data:

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P hexa-deb-1076 azuser1076\_mml.local@ihhtl.onmicr...

EDA 2024-02-21 10:13:03 Python ☆

File Edit View Run Help Last edit was 8 minutes ago New cell UI: ON

Run all azuser1076\_mml.local's... Schedule Share

df.show()

(1) Spark Jobs

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen ...	male	22.0	1	0	A/5 21171	7.25	NULL	S
2	1	1	Cumings, Mrs. Joh...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. ...	female	26.0	0	0	STON/O2. 3101282	7.925	NULL	S
4	1	1	Futrelle, Mrs. Ja...	female	35.0	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. Willia...	male	35.0	0	0	373450	8.05	NULL	S
6	0	3	Horan, Mr. James	male	NULL	0	0	330877	8.4583	NULL	Q
7	0	1	McCarthy, Mr. Tim...	male	54.0	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. ...	male	2.0	3	1	349909	21.075	NULL	S
9	1	3	Johnson, Mrs. Osc...	female	27.0	0	2	347742	11.1333	NULL	S
10	1	2	Nasser, Mrs. Nich...	female	14.0	1	0	237736	30.0708	NULL	C
11	1	3	Sandstrom, Miss. ...	female	4.0	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. El...	female	58.0	0	0	113783	26.55	C103	S
13	0	3	Saunderscock, Mr. ...	male	20.0	0	0	A/5. 2151	8.05	NULL	S
14	0	3	Andersson, Mr. An...	male	39.0	1	5	347082	31.275	NULL	S
15	0	3	Vestrom, Miss. Hu...	female	14.0	0	0	350406	7.8542	NULL	S
16	1	2	Hewlett, Mrs. (Ma...	female	55.0	0	0	248706	16.0	NULL	S
17	0	3	Rice, Master. Eugene	male	2.0	4	1	382652	29.125	NULL	Q
18	1	2	Williams, Mr. Cha...	male	NULL	0	0	244373	13.0	NULL	S

## Describing the whole dataset:

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P hexa-deb-1076 azuser1076\_mml.local@ihhtl.onmicr...

EDA 2024-02-21 10:13:03 Python ☆

File Edit View Run Help Last edit was 9 minutes ago New cell UI: ON

Run all azuser1076\_mml.local's... Schedule Share

df.describe().show()

(2) Spark Jobs

summary	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	P
arch	Ticket	Fare	Cabin	Embarked				
count	891	891	891	891	891	714	891	
mean	446.0	0.3838383838383838	2.308641975308642	NULL	NULL	29.69911764705882	0.5238078563411896	0.3815937149270
stddev	257.3538420152301	0.48659245426485753	0.8360712409770491	NULL	NULL	14.526497332334035	1.1027434322934315	0.806057221129
min	110152	1	0	1	Andersson, Mr. A...	female	0.42	0
max	891	0.0	A10	1	3	van Melkebeke, Mr...	male	80.0

## Filtering Cabin column which contain null values:

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P hexa-deb-1076 azuser1076\_mml.local@iitl.onmicr...

EDA 2024-02-21 10:13:03 Python ☆

File Edit View Run Help Last edit was 9 minutes ago New cell UI: ON ▾ ▶ Run all azuser1076\_mml.local's... Schedule Share

1025 AM (<1s) Cell 5

```
df = df.filter(df.Cabin.isNotNull())
```

df: pyspark.sql.dataframe.DataFrame = [PassengerId: integer, Survived: integer ... 10 more fields]

1025 AM (<1s) Cell 6 Python

```
df.show()
```

(1) Spark Jobs

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	1	Cumings, Mrs. Joh...	female	38.0	1	0	PC 17599	71.2833	C85	C
4	1	1	Futrelle, Mrs. Ja...	female	35.0	1	0	113803	53.1	C123	S
7	0	1	McCarthy, Mr. Tim...	male	54.0	0	0	17463	51.8625	E46	S
11	1	3	Sandstrom, Miss. ...	female	4.0	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. El...	female	58.0	0	0	113783	26.55	C103	S
22	1	2	Beesley, Mr. Lawr...	male	34.0	0	0	248698	13.0	D56	S
24	1	1	Sloper, Mr. Willi...	male	28.0	0	0	113788	35.5	A6	S
28	0	1	Fortune, Mr. Char...	male	19.0	3	2	19950	263.0	C23 C25 C27	S
32	1	1	Spencer, Mrs. Mil...	female	NULL	1	0	PC 17569	146.5208	B78	C
53	1	1	Harper, Mrs. Henr...	female	49.0	1	0	PC 17572	76.7292	D33	C
55	0	1	Ostby, Mr. Engelh...	male	65.0	0	1	113509	61.9792	B30	C
56	1	1	Woolner, Mr. Hugh	male	NULL	0	0	19947	35.5	C52	S
62	1	1	Icard, Miss. Amelie	female	38.0	0	0	113572	80.0	B28	NULL
63	0	1	Harris, Mr. Henry...	male	45.0	1	0	36973	83.475	C83	S
67	1	2	Nye, Mrs. (Elizab...	female	29.0	0	0	C.A. 29395	10.5	F33	S

## Filtering Age column which contain null values:

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P hexa-deb-1076 azuser1076\_mml.local@iitl.onmicr...

EDA 2024-02-21 10:13:03 Python ☆

File Edit View Run Help Last edit was 10 minutes ago New cell UI: ON ▾ ▶ Run all azuser1076\_mml.local's... Schedule Share

1028 AM (<1s) Cell 7

```
df = df.filter(df.Age.isNotNull())
```

df: pyspark.sql.dataframe.DataFrame = [PassengerId: integer, Survived: integer ... 10 more fields]

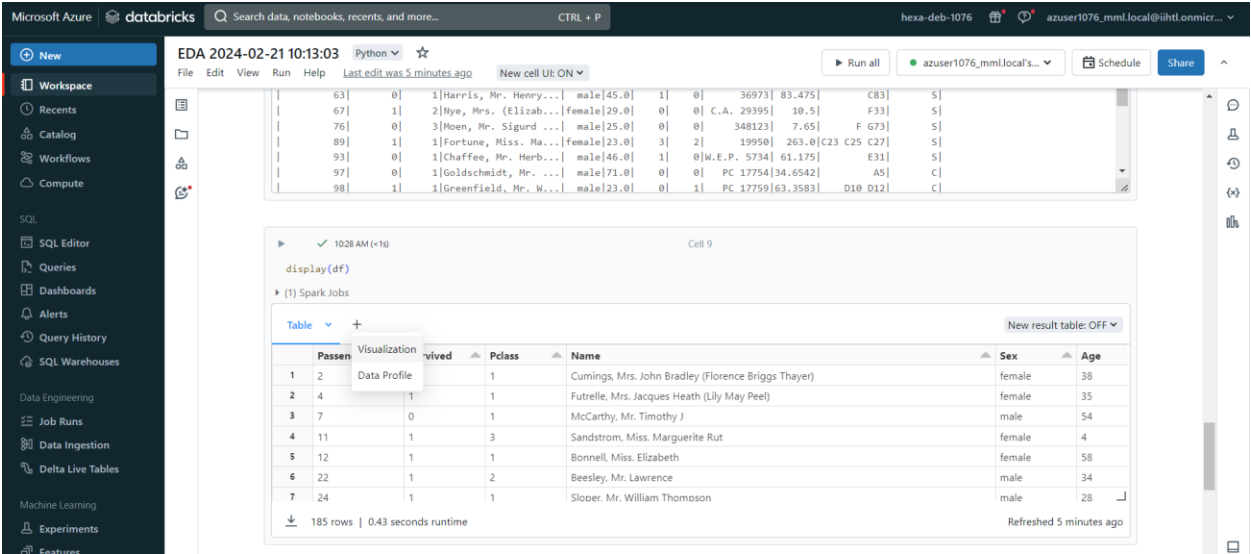
1027 AM (<1s) Cell 8

```
df.show()
```

(1) Spark Jobs

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	1	Cumings, Mrs. Joh...	female	38.0	1	0	PC 17599	71.2833	C85	C
4	1	1	Futrelle, Mrs. Ja...	female	35.0	1	0	113803	53.1	C123	S
7	0	1	McCarthy, Mr. Tim...	male	54.0	0	0	17463	51.8625	E46	S
11	1	3	Sandstrom, Miss. ...	female	4.0	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. El...	female	58.0	0	0	113783	26.55	C103	S
22	1	2	Beesley, Mr. Lawr...	male	34.0	0	0	248698	13.0	D56	S
24	1	1	Sloper, Mr. Willi...	male	28.0	0	0	113788	35.5	A6	S
28	0	1	Fortune, Mr. Char...	male	19.0	3	2	19950	263.0	C23 C25 C27	S
53	1	1	Harper, Mrs. Henr...	female	49.0	1	0	PC 17572	76.7292	D33	C
55	0	1	Ostby, Mr. Engelh...	male	65.0	0	1	113509	61.9792	B30	C
62	1	1	Icard, Miss. Amelie	female	38.0	0	0	113572	80.0	B28	NULL
63	0	1	Harris, Mr. Henry...	male	45.0	1	0	36973	83.475	C83	S
67	1	2	Nye, Mrs. (Elizab...	female	29.0	0	0	C.A. 29395	10.5	F33	S

# Using Display method for Analyzing the data:



## Analysis

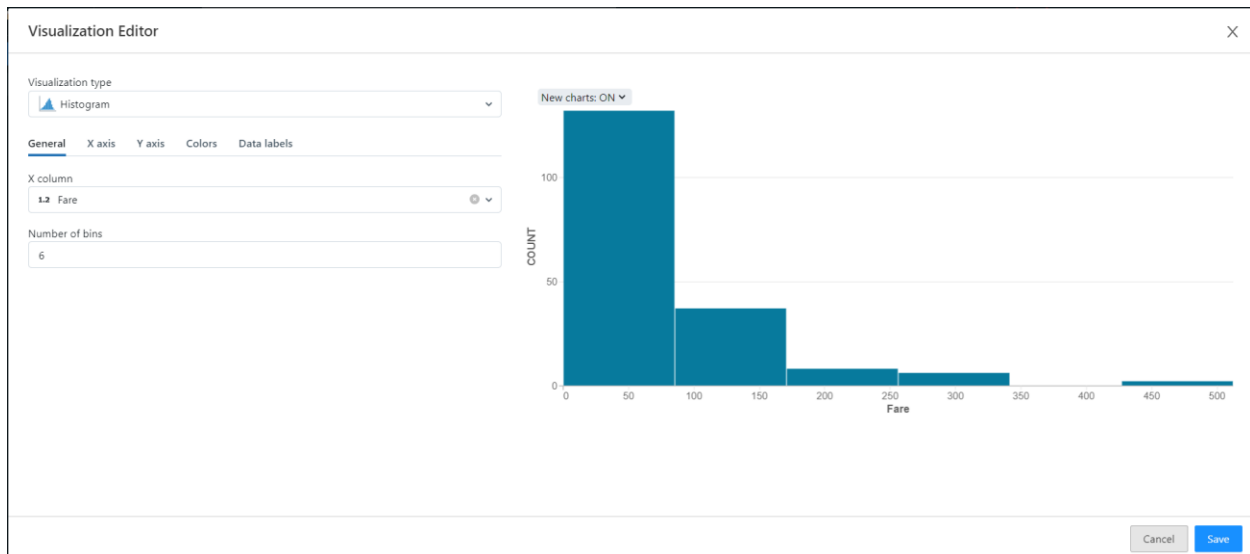
Analyzing How many people survived of particular age group and grouped by Gender:

### Bar Graph



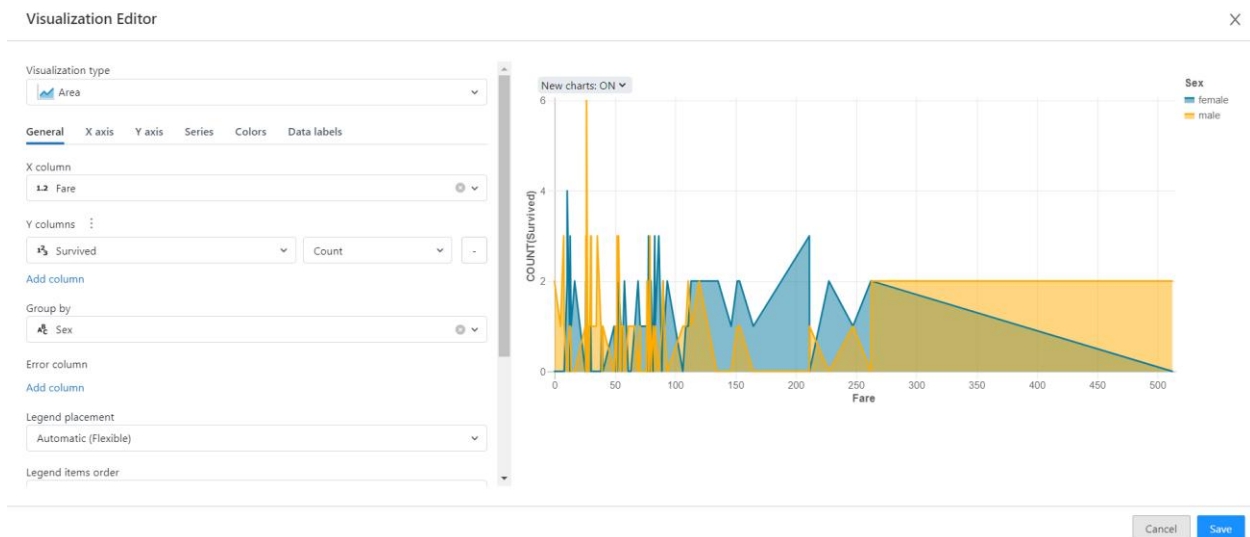
Count of people lying in particular Ticket price:

## Histogram:



Analysis for fare vs survived where we will come to know that higher ticket price passenger survived more or not:

## Area chart:



Here we will come to know about the proportion of the gender:

### Pie chart:

