In [1]:
```
from pyspark.sql import SparkSession
spark=SparkSession.builder.appName('Dataframe').getOrCreate()
```

In [2]:
```
spark
```

Out[2]:

SparkSession - in-memory

SparkContext

Spark UI (http://Arnak:4040)

Version

v3.5.1

Master

local[*]

AppName

Dataframe

In [4]:
```
master2=spark.read.csv(r"D:\NIT\FEBRUARY\26 feb (time series and spark)\sample
```

In [5]:
```
master2
```

Out[5]:
```
DataFrame[Name: string, age: int, Experience: int, Salary: int]
```

In [6]:
```
master2.show()
```

```
+--------+----+----------+------+
|    Name| age|Experience|Salary|
+--------+----+----------+------+
|    jack|  31|        10| 30000|
|    alex|  30|         8| 25000|
|caroline|  29|         4| 20000|
|    paul|  24|         3| 20000|
|  sandra|  21|         1| 15000|
|casandra|  23|         2| 18000|
|     dan|NULL|      NULL| 40000|
|    NULL|  34|        10| 38000|
|    NULL|  36|      NULL|  NULL|
+--------+----+----------+------+
```

In [7]:
```
master2.drop('Name').show()
```

```
+----+----------+------+
| age|Experience|Salary|
+----+----------+------+
|  31|        10| 30000|
|  30|         8| 25000|
|  29|         4| 20000|
|  24|         3| 20000|
|  21|         1| 15000|
|  23|         2| 18000|
|NULL|      NULL| 40000|
|  34|        10| 38000|
|  36|      NULL|  NULL|
+----+----------+------+
```

In [9]:
```
master2.show()
```

```
+--------+----+----------+------+
|    Name| age|Experience|Salary|
+--------+----+----------+------+
|    jack|  31|        10| 30000|
|    alex|  30|         8| 25000|
|caroline|  29|         4| 20000|
|    paul|  24|         3| 20000|
|  sandra|  21|         1| 15000|
|casandra|  23|         2| 18000|
|     dan|NULL|      NULL| 40000|
|    NULL|  34|        10| 38000|
|    NULL|  36|      NULL|  NULL|
+--------+----+----------+------+
```

In [11]:
```
master2.na.drop().show()
```

```
+--------+---+----------+------+
|    Name|age|Experience|Salary|
+--------+---+----------+------+
|    jack| 31|        10| 30000|
|    alex| 30|         8| 25000|
|caroline| 29|         4| 20000|
|    paul| 24|         3| 20000|
|  sandra| 21|         1| 15000|
|casandra| 23|         2| 18000|
+--------+---+----------+------+
```

In [12]:
```
master2.na.drop(how="any").show()
```

```
+--------+---+----------+------+
|    Name|age|Experience|Salary|
+--------+---+----------+------+
|    jack| 31|        10| 30000|
|    alex| 30|         8| 25000|
|caroline| 29|         4| 20000|
|    paul| 24|         3| 20000|
|  sandra| 21|         1| 15000|
|casandra| 23|         2| 18000|
+--------+---+----------+------+
```

In [16]:
```
master2.na.drop(how="any",thresh=3).show()
```

```
+--------+---+----------+------+
|    Name|age|Experience|Salary|
+--------+---+----------+------+
|    jack| 31|        10| 30000|
|    alex| 30|         8| 25000|
|caroline| 29|         4| 20000|
|    paul| 24|         3| 20000|
|  sandra| 21|         1| 15000|
|casandra| 23|         2| 18000|
|    NULL| 34|        10| 38000|
+--------+---+----------+------+
```

In [17]:
```
master2.na.drop(how="any",subset=['Age']).show()
```

```
+--------+---+----------+------+
|    Name|age|Experience|Salary|
+--------+---+----------+------+
|    jack| 31|        10| 30000|
|    alex| 30|         8| 25000|
|caroline| 29|         4| 20000|
|    paul| 24|         3| 20000|
|  sandra| 21|         1| 15000|
|casandra| 23|         2| 18000|
|    NULL| 34|        10| 38000|
|    NULL| 36|      NULL|  NULL|
+--------+---+----------+------+
```

In [19]:
```
master2.na.fill('Missing Values',['Experience','age']).show()
```

```
+--------+----+----------+------+
|    Name| age|Experience|Salary|
+--------+----+----------+------+
|    jack|  31|        10| 30000|
|    alex|  30|         8| 25000|
|caroline|  29|         4| 20000|
|    paul|  24|         3| 20000|
|  sandra|  21|         1| 15000|
|casandra|  23|         2| 18000|
|     dan|NULL|      NULL| 40000|
|    NULL|  34|        10| 38000|
|    NULL|  36|      NULL|  NULL|
+--------+----+----------+------+
```

In [21]:
```
master2.show()
```

```
+--------+----+----------+------+
|    Name| age|Experience|Salary|
+--------+----+----------+------+
|    jack|  31|        10| 30000|
|    alex|  30|         8| 25000|
|caroline|  29|         4| 20000|
|    paul|  24|         3| 20000|
|  sandra|  21|         1| 15000|
|casandra|  23|         2| 18000|
|     dan|NULL|      NULL| 40000|
|    NULL|  34|        10| 38000|
|    NULL|  36|      NULL|  NULL|
+--------+----+----------+------+
```

In [23]:
```
master2.printSchema()
```

```
root
 |-- Name: string (nullable = true)
 |-- age: integer (nullable = true)
 |-- Experience: integer (nullable = true)
 |-- Salary: integer (nullable = true)
```

In [24]:
```
from pyspark.ml.feature import Imputer

imputer = Imputer(
    inputCols=['age', 'Experience', 'Salary'],
    outputCols=["{}_imputed".format(c) for c in ['age', 'Experience', 'Salary'
    ).setStrategy("median")
```

In [26]:
```
imputer.fit(master2).transform(master2).show()
```

```
+--------+----+----------+------+-----------+-----------------+-------------
-+
|    Name| age|Experience|Salary|age_imputed|Experience_imputed|Salary_impute
d|
+--------+----+----------+------+-----------+-----------------+-------------
-+
|    jack|  31|        10| 30000|         31|                10|          3000
0|
|    alex|  30|         8| 25000|         30|                 8|          2500
0|
|caroline|  29|         4| 20000|         29|                 4|          2000
0|
|    paul|  24|         3| 20000|         24|                 3|          2000
0|
|  sandra|  21|         1| 15000|         21|                 1|          1500
0|
|casandra|  23|         2| 18000|         23|                 2|          1800
0|
|     dan|NULL|      NULL| 40000|         29|                 4|          4000
0|
|    NULL|  34|        10| 38000|         34|                10|          3800
0|
|    NULL|  36|      NULL|  NULL|         36|                 4|          2000
0|
+--------+----+----------+------+-----------+-----------------+-------------
-+
```

In [28]:
```
from pyspark.ml.feature import Imputer

imputer_mode = Imputer(
    inputCols=['age', 'Experience', 'Salary'],
    outputCols=["{}_imputed".format(c) for c in ['age', 'Experience', 'Salary'
    ).setStrategy("mode")
```

In [29]:
```
imputer_mode.fit(master2).transform(master2).show()
```

```
+--------+----+----------+------+----------+-----------------+-------------
-+
|    Name| age|Experience|Salary|age_imputed|Experience_imputed|Salary_impute
d|
+--------+----+----------+------+----------+-----------------+-------------
-+
|    jack|  31|        10| 30000|         31|                10|         3000
0|
|    alex|  30|         8| 25000|         30|                 8|         2500
0|
|caroline|  29|         4| 20000|         29|                 4|         2000
0|
|    paul|  24|         3| 20000|         24|                 3|         2000
0|
|  sandra|  21|         1| 15000|         21|                 1|         1500
0|
|casandra|  23|         2| 18000|         23|                 2|         1800
0|
|     dan|NULL|      NULL| 40000|         21|                10|         4000
0|
|    NULL|  34|        10| 38000|         34|                10|         3800
0|
|    NULL|  36|      NULL|  NULL|         36|                10|         2000
0|
+--------+----+----------+------+----------+-----------------+-------------
-+
```

In [30]:
```
imputer_mean = Imputer(
    inputCols=['age', 'Experience', 'Salary'],
    outputCols=["{}_imputed".format(c) for c in ['age', 'Experience', 'Salary'
    ).setStrategy("mean")
```

In [31]:
```
imputer_mean.fit(master2).transform(master2).show()
```

```
+--------+----+----------+------+-----------+-----------------+-------------
-+
|    Name| age|Experience|Salary|age_imputed|Experience_imputed|Salary_impute
d|
+--------+----+----------+------+-----------+-----------------+-------------
-+
|    jack|  31|        10| 30000|         31|                10|         3000
0|
|    alex|  30|         8| 25000|         30|                 8|         2500
0|
|caroline|  29|         4| 20000|         29|                 4|         2000
0|
|    paul|  24|         3| 20000|         24|                 3|         2000
0|
|  sandra|  21|         1| 15000|         21|                 1|         1500
0|
|casandra|  23|         2| 18000|         23|                 2|         1800
0|
|     dan|NULL|      NULL| 40000|         28|                 5|         4000
0|
|    NULL|  34|        10| 38000|         34|                10|         3800
0|
|    NULL|  36|      NULL|  NULL|         36|                 5|         2575
0|
+--------+----+----------+------+-----------+-----------------+-------------
-+
```

In [ ]: