Mod 2 - Overview - Dataframe- Handling Missing Values

March 4, 2025

0.0.1 Pyspark Handling Missing Values

- Dropping Columns
- Dropping Rows
- Various Parameter In Dropping functionalities
- Handling Missing values by Mean, MEdian And Mode

[3]:
```
from pyspark.sql import SparkSession
spark=SparkSession.builder.appName('Practise').getOrCreate()
```

[4]:
```
df_pyspark=spark.read.csv('test2.csv',header=True,inferSchema=True)
```

[5]:
```
df_pyspark.printSchema()
```

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

[6]:
```
df_pyspark.show()
```

```
+---------+----+----------+------+
|     Name| age|Experience|Salary|
+---------+----+----------+------+
|    Krish|  31|        10| 30000|
|Sudhanshu|  30|         8| 25000|
|    Sunny|  29|         4| 20000|
|     Paul|  24|         3| 20000|
|   Harsha|  21|         1| 15000|
|  Shubham|  23|         2| 18000|
|   Mahesh|null|      null| 40000|
|     null|  34|        10| 38000|
|     null|  36|      null|  null|
+---------+----+----------+------+
```

[7]:

```
##drop the columns
df_pyspark.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|
+----+----------+------+
```

[8]:

```
df_pyspark.show()
```

```
+---------+----+----------+------+
|     Name| age|Experience|Salary|
+---------+----+----------+------+
|    Krish|  31|        10| 30000|
|Sudhanshu|  30|         8| 25000|
|    Sunny|  29|         4| 20000|
|     Paul|  24|         3| 20000|
|   Harsha|  21|         1| 15000|
|  Shubham|  23|         2| 18000|
|   Mahesh|null|      null| 40000|
|     null|  34|        10| 38000|
|     null|  36|      null|  null|
+---------+----+----------+------+
```

[9]:

```
df_pyspark.na.drop().show()
```

```
+---------+---+----------+------+
|     Name|age|Experience|Salary|
+---------+---+----------+------+
|    Krish| 31|        10| 30000|
|Sudhanshu| 30|         8| 25000|
|    Sunny| 29|         4| 20000|
|     Paul| 24|         3| 20000|
|   Harsha| 21|         1| 15000|
|  Shubham| 23|         2| 18000|
+---------+---+----------+------+
```

[14]:

```
### any==how
df_pyspark.na.drop(how="any").show()
```

```
+---------+---+----------+------+
|     Name|age|Experience|Salary|
+---------+---+----------+------+
|    Krish| 31|        10| 30000|
|Sudhanshu| 30|         8| 25000|
|    Sunny| 29|         4| 20000|
|     Paul| 24|         3| 20000|
|   Harsha| 21|         1| 15000|
|  Shubham| 23|         2| 18000|
|         |   |          |      |
+---------+---+----------+------+
```

[17]:

```
##threshold
df_pyspark.na.drop(how="any",thresh=3).show()
```

```
+---------+---+----------+------+
|     Name|age|Experience|Salary|
+---------+---+----------+------+
|    Krish| 31|        10| 30000|
|Sudhanshu| 30|         8| 25000|
|    Sunny| 29|         4| 20000|
|     Paul| 24|         3| 20000|
|   Harsha| 21|         1| 15000|
|  Shubham| 23|         2| 18000|
|     null| 34|        10| 38000|
|         |   |          |      |
+---------+---+----------+------+
```

[19]:

```
##Subset
df_pyspark.na.drop(how="any",subset=['Age']).show()
```

```
+---------+---+----------+------+
|     Name|age|Experience|Salary|
+---------+---+----------+------+
|    Krish| 31|        10| 30000|
|Sudhanshu| 30|         8| 25000|
|    Sunny| 29|         4| 20000|
|     Paul| 24|         3| 20000|
|   Harsha| 21|         1| 15000|
|  Shubham| 23|         2| 18000|
|     null| 34|        10| 38000|
|     null| 36|      null|  null|
|         |   |          |      |
+---------+---+----------+------+
```

[22]:

```
### Filling the Missing Value
df_pyspark.na.fill('Missing Values',['Experience','age']).show()
```

```
+---------+--------------+--------------+------+
|     Name|           age|    Experience|Salary|
+---------+--------------+--------------+------+
|    Krish|            31|            10| 30000|
|Sudhanshu|            30|             8| 25000|
|    Sunny|            29|             4| 20000|
|     Paul|            24|             3| 20000|
|   Harsha|            21|             1| 15000|
|  Shubham|            23|             2| 18000|
|   Mahesh|Missing Values|Missing Values| 40000|
|     null|            34|            10| 38000|
|     null|            36|Missing Values|  null|
|         |              |              |      |
+---------+--------------+--------------+------+
```

[38]:

```
df_pyspark.show()
```

```
+---------+----+----------+------+
|     Name| age|Experience|Salary|
+---------+----+----------+------+
|    Krish|  31|        10| 30000|
|Sudhanshu|  30|         8| 25000|
|    Sunny|  29|         4| 20000|
|     Paul|  24|         3| 20000|
|   Harsha|  21|         1| 15000|
|  Shubham|  23|         2| 18000|
|   Mahesh|null|      null| 40000|
|     null|  34|        10| 38000|
|     null|  36|      null|  null|
|         |    |          |      |
+---------+----+----------+------+
```

[44]:

```
df_pyspark.printSchema()
```

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

[13]:

```python
from pyspark.ml.feature import Imputer

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

[14]:

```python
# Add imputation cols to df
imputer.fit(df_pyspark).transform(df_pyspark).show()
```

```
+---------+----+----------+------+-----------+------------------+--------------+
|     Name| age|Experience|Salary|age_imputed|Experience_imputed|Salary_imputed|
+---------+----+----------+------+-----------+------------------+--------------+
|    Krish|  31|        10| 30000|         31|                10|         30000|
|Sudhanshu|  30|         8| 25000|         30|                 8|         25000|
|    Sunny|  29|         4| 20000|         29|                 4|         20000|
|     Paul|  24|         3| 20000|         24|                 3|         20000|
|   Harsha|  21|         1| 15000|         21|                 1|         15000|
|  Shubham|  23|         2| 18000|         23|                 2|         18000|
|   Mahesh|null|      null| 40000|         29|                 4|         40000|
|     null|  34|        10| 38000|         34|                10|         38000|
|     null|  36|      null|  null|         36|                 4|         20000|
+---------+----+----------+------+-----------+------------------+--------------+
```