```
pip install pyspark
```

```
Requirement already satisfied: pyspark in /usr/local/lib/python3.7/dist-packages (3.1
Requirement already satisfied: py4j==0.10.9 in /usr/local/lib/python3.7/dist-packages
```

```
import pyspark
```

```
import pandas as pd
type(pd.read_csv('/content/text1.csv'))
```

```
pandas.core.frame.DataFrame
```

```
from pyspark.sql import SparkSession
```

```
spark = SparkSession.builder.appName('Practise').getOrCreate()
```

```
spark
```

SparkSession - in-memory

SparkContext

Spark UI

Version
    v3.1.2
Master
    local[*]
AppName
    Practise

```
df_pyspark=spark.read.csv('/content/text1.csv')
```

```
df_pyspark.show()
```

```
+------+---+
|   _c0|_c1|
+------+---+
|  Name|Age|
|  Yash| 23|
|Tamizh| 23|
|Madhan| 23|
+------+---+
```

```
df_pyspark= spark.read.option('header','true').csv('/content/text1.csv')
```

```
type(df_pyspark)
```

```
pyspark.sql.dataframe.DataFrame
```

```
df_pyspark.head(3)
```

```
[Row(Name='Yash', Age='23'),
 Row(Name='Tamizh', Age='23'),
 Row(Name='Madhan', Age='23')]
```

```
pd.read_csv('/content/text1.csv')
```

	Name	Age	Experience
0	Yash	23	2
1	Tamizh	23	3
2	Madhan	23	2

```
spark
```

SparkSession - in-memory

SparkContext

Spark UI

Version
    `v3.1.2`
Master
    `local[*]`
AppName
    `Practise`

```
## read the dataset
```

```
df_pyspark=spark.read.option('header','true').csv('/content/text1.csv',inferSchema=True)
```

```
df_pyspark.printSchema()
```

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

```
df_pyspark=spark.read.csv('/content/text1.csv',header=True,inferSchema=True)
```

```
df_pyspark.show()
```

```
+------+---+----------+
|  Name|Age|Experience|
+------+---+----------+
|  Yash| 23|         2|
|Tamizh| 23|         3|
|Madhan| 23|         2|
+------+---+----------+
```

```
type(df_pyspark)
```

```
pyspark.sql.dataframe.DataFrame
```

```
df_pyspark.columns
```

```
['Name', 'Age', 'Experience']
```

```
df_pyspark.select('Name').show()
```

```
+------+
|  Name|
+------+
|  Yash|
|Tamizh|
|Madhan|
+------+
```

```
df_pyspark.select(['Name','Experience']).show()
```

```
+------+----------+
|  Name|Experience|
+------+----------+
|  Yash|         2|
|Tamizh|         3|
|Madhan|         2|
+------+----------+
```

```
df_pyspark.dtypes
```

```
[('Name', 'string'), ('Age', 'int'), ('Experience', 'int')]
```

```
df_pyspark.describe().show()
```

```
+-------+------+----+------------------+
|summary|  Name| Age|        Experience|
+-------+------+----+------------------+
|  count|     3|   3|                 3|
|   mean|  null|23.0|2.3333333333333335|
| stddev|  null| 0.0|0.5773502691896258|
|    min|Madhan|  23|                 2|
|    max|  Yash|  23|                 3|
+-------+------+----+------------------+
```

```
##adding columns
```

```
df_pyspark=df_pyspark.withColumn('Experience after two years',df_pyspark['Experience']+2)
```

```
df_pyspark.show()
```

```
+------+---+----------+--------------------------+
|  Name|Age|Experience|Experience after two years|
+------+---+----------+--------------------------+
|  Yash| 23|         2|                         4|
|Tamizh| 23|         3|                         5|
|Madhan| 23|         2|                         4|
+------+---+----------+--------------------------+
```

```
## Drop the column
```

```
df_pyspark = df_pyspark.drop('Experience after two years')
```

```
df_pyspark.show()
```

```
+------+---+----------+
|  Name|Age|Experience|
+------+---+----------+
|  Yash| 23|         2|
|Tamizh| 23|         3|
|Madhan| 23|         2|
+------+---+----------+
```

```
df_pyspark.withColumnRenamed('Name','Newname').show()
```

```
+-------+---+----------+
|Newname|Age|Experience|
+-------+---+----------+
|   Yash| 23|         2|
| Tamizh| 23|         3|
| Madhan| 23|         2|
+-------+---+----------+
```

```
pd.read_csv('/content/text1.csv')
```

	Name	Age	Experience	Salary
0	Yash	23.0	2.0	25000.0
1	Tamizh	23.0	3.0	28000.0

```
df_pyspark =spark.read.csv('/content/text1.csv',header=True)
```

3	Deepak	25.0	5.0	21000.0

```
df_pyspark.show()
```

```
+--------+----+----------+------+
|    Name| Age|Experience|Salary|
+--------+----+----------+------+
|    Yash|  23|         2| 25000|
|  Tamizh|  23|         3| 28000|
|  Madhan|  23|         2| 20000|
|  Deepak|  25|         5| 21000|
|Nishanth|  24|         2| 26000|
|     Anu|  23|         1| 10000|
|   Durga|null|      null| 20000|
|    null|  25|         3| 30000|
|    null|  30|      null|  null|
+--------+----+----------+------+
```

```
##drop the columns
```

```
df_pyspark.drop('Name').show()
```

```
+----+----------+------+
| Age|Experience|Salary|
+----+----------+------+
|  23|         2| 25000|
|  23|         3| 28000|
|  23|         2| 20000|
|  25|         5| 21000|
|  24|         2| 26000|
|  23|         1| 10000|
|null|      null| 20000|
|  25|         3| 30000|
|  30|      null|  null|
+----+----------+------+
```

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

```
+--------+---+----------+------+
|    Name|Age|Experience|Salary|
+--------+---+----------+------+
|    Yash| 23|         2| 25000|
|  Tamizh| 23|         3| 28000|
|  Madhan| 23|         2| 20000|
|  Deepak| 25|         5| 21000|
|Nishanth| 24|         2| 26000|
|     Anu| 23|         1| 10000|
+--------+---+----------+------+
```

```
df_pyspark.na.drop(how='all').show()
```

```
+--------+----+----------+------+
|    Name| Age|Experience|Salary|
+--------+----+----------+------+
|    Yash|  23|         2| 25000|
|  Tamizh|  23|         3| 28000|
|  Madhan|  23|         2| 20000|
|  Deepak|  25|         5| 21000|
|Nishanth|  24|         2| 26000|
|     Anu|  23|         1| 10000|
|   Durga|null|      null| 20000|
|    null|  25|         3| 30000|
|    null|  30|      null|  null|
+--------+----+----------+------+
```

```
df_pyspark.na.drop(how='any').show()
```

```
+--------+---+----------+------+
|    Name|Age|Experience|Salary|
+--------+---+----------+------+
|    Yash| 23|         2| 25000|
|  Tamizh| 23|         3| 28000|
|  Madhan| 23|         2| 20000|
|  Deepak| 25|         5| 21000|
|Nishanth| 24|         2| 26000|
|     Anu| 23|         1| 10000|
+--------+---+----------+------+
```

```
#threshold
```

```
df_pyspark.na.drop(how='all',thresh=3).show()
```

```
+--------+---+----------+------+
|    Name|Age|Experience|Salary|
+--------+---+----------+------+
|    Yash| 23|         2| 25000|
|  Tamizh| 23|         3| 28000|
|  Madhan| 23|         2| 20000|
|  Deepak| 25|         5| 21000|
|Nishanth| 24|         2| 26000|
|     Anu| 23|         1| 10000|
|    null| 25|         3| 30000|
+--------+---+----------+------+
```

```
#subset
```

```
df_pyspark.na.drop(how='all',subset=['Name']).show()
```

```
+--------+----+----------+------+
|    Name| Age|Experience|Salary|
+--------+----+----------+------+
|    Yash|  23|         2| 25000|
|  Tamizh|  23|         3| 28000|
|  Madhan|  23|         2| 20000|
|  Deepak|  25|         5| 21000|
|Nishanth|  24|         2| 26000|
|     Anu|  23|         1| 10000|
|   Durga|null|      null| 20000|
+--------+----+----------+------+
```

```
##filling the missing value
```

```
df_pyspark.show()
```

```
+--------+----+----------+------+
|    Name| Age|Experience|Salary|
+--------+----+----------+------+
|    Yash|  23|         2| 25000|
|  Tamizh|  23|         3| 28000|
|  Madhan|  23|         2| 20000|
|  Deepak|  25|         5| 21000|
|Nishanth|  24|         2| 26000|
|     Anu|  23|         1| 10000|
|   Durga|null|      null| 20000|
|    null|  25|         3| 30000|
|    null|  30|      null|  null|
+--------+----+----------+------+
```

```
df_pyspark.na.fill('Missing Values').show()
```

```
+--------------+--------------+--------------+--------------+
|          Name|           Age|    Experience|        Salary|
+--------------+--------------+--------------+--------------+
|          Yash|            23|             2|         25000|
|        Tamizh|            23|             3|         28000|
|        Madhan|            23|             2|         20000|
|        Deepak|            25|             5|         21000|
|      Nishanth|            24|             2|         26000|
|           Anu|            23|             1|         10000|
|         Durga|Missing Values|Missing Values|         20000|
|Missing Values|            25|             3|         30000|
|Missing Values|            30|Missing Values|Missing Values|
+--------------+--------------+--------------+--------------+
```

```
#filling the null values with mean, mode, median
```

```
df_pyspark.dtypes
```

```
[('Name', 'string'),
 ('Age', 'string'),
 ('Experience', 'string'),
 ('Salary', 'string')]
```

```
df_pyspark = spark.read.csv('/content/text1.csv',header=True,inferSchema=True)
```

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

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

```
imputer.fit(df_pyspark).transform(df_pyspark).show()
```

```
+--------+----+----------+------+-----------+------------------+--------------+
|    Name| Age|Experience|Salary|Age_imputed|Experience_imputed|Salary_imputed|
+--------+----+----------+------+-----------+------------------+--------------+
|    Yash|  23|         2| 25000|         23|                 2|         25000|
|  Tamizh|  23|         3| 28000|         23|                 3|         28000|
|  Madhan|  23|         2| 20000|         23|                 2|         20000|
|  Deepak|  25|         5| 21000|         25|                 5|         21000|
|Nishanth|  24|         2| 26000|         24|                 2|         26000|
|     Anu|  23|         1| 10000|         23|                 1|         10000|
|   Durga|null|      null| 20000|         23|                 2|         20000|
|    null|  25|         3| 30000|         25|                 3|         30000|
|    null|  30|      null|  null|         30|                 2|         21000|
+--------+----+----------+------+-----------+------------------+--------------+
```

```
#filter
```

```
df_pyspark=spark.read.csv('/content/text1.csv',header=True,inferSchema=True)
```

```
df_pyspark.show()
```

```
+--------+---+----------+------+
|    Name|Age|Experience|Salary|
+--------+---+----------+------+
|    Yash| 23|         2| 25000|
|  Tamizh| 23|         3| 28000|
|  Madhan| 23|         2| 20000|
|  Deepak| 25|         5| 21000|
|Nishanth| 24|         2| 26000|
|     Anu| 23|         1| 10000|
|   Durga| 25|         2| 20000|
+--------+---+----------+------+
```

```
df_pyspark.filter('Salary<=20000').show()
```

```
+------+---+----------+------+
|  Name|Age|Experience|Salary|
+------+---+----------+------+
|Madhan| 23|         2| 20000|
|   Anu| 23|         1| 10000|
| Durga| 25|         2| 20000|
+------+---+----------+------+
```

```
df_pyspark.filter(df_pyspark['Salary']<=20000).show()
```

```
+------+---+----------+------+
|  Name|Age|Experience|Salary|
+------+---+----------+------+
|Madhan| 23|         2| 20000|
|   Anu| 23|         1| 10000|
| Durga| 25|         2| 20000|
+------+---+----------+------+
```

```
df_pyspark.filter((df_pyspark['Salary']<=20000) |
                  (df_pyspark['Salary']>=10000)).show()
```

```
+--------+---+----------+------+
|    Name|Age|Experience|Salary|
+--------+---+----------+------+
|    Yash| 23|         2| 25000|
|  Tamizh| 23|         3| 28000|
|  Madhan| 23|         2| 20000|
|  Deepak| 25|         5| 21000|
|Nishanth| 24|         2| 26000|
|     Anu| 23|         1| 10000|
|   Durga| 25|         2| 20000|
+--------+---+----------+------+
```

```
df_pyspark.filter(~(df_pyspark['Salary']<=20000)).show()
```

```
+--------+---+----------+------+
|    Name|Age|Experience|Salary|
+--------+---+----------+------+
|    Yash| 23|         2| 25000|
|  Tamizh| 23|         3| 28000|
|  Deepak| 25|         5| 21000|
|Nishanth| 24|         2| 26000|
+--------+---+----------+------+
```

```
df_pyspark=spark.read.csv('/content/text2.csv',header=True,inferSchema=True)
```

```
df_pyspark.show()
```

```
+------+----------+------+
|  Name|Department|Salary|
+------+----------+------+
|  Yash|        DS| 25000|
|  Yash|      Mech| 28000|
|Madhan|    Design| 20000|
|Madhan|    Design| 21000|
|Madhan|    Design| 26000|
|Rasega|        IT| 10000|
|  Pavi|  Digital| 20000|
|  Pavi|       IT| 25000|
|Rasega|     Data| 40000|
|Rasega|    Cloud| 30000|
+------+---------+------+
```

```
#GroupBy
```

```
df_pyspark.groupBy('Name').sum().show()
```

```
+------+-----------+
|  Name|sum(Salary)|
+------+-----------+
|Rasega|      80000|
|  Pavi|      45000|
|Madhan|      67000|
|  Yash|      53000|
+------+-----------+
```

```
df_pyspark.groupBy('Department').max().show()
```

```
+----------+-----------+
|Department|max(Salary)|
+----------+-----------+
|      Data|      40000|
|   Digital|      20000|
|    Design|      26000|
|        IT|      25000|
|      Mech|      28000|
|        DS|      25000|
|     Cloud|      30000|
+----------+-----------+
```

```
df_pyspark.groupBy('Department').mean().show()
```

```
+----------+------------------+
|Department|       avg(Salary)|
+----------+------------------+
|      Data|           40000.0|
|   Digital|           20000.0|
|    Design|22333.333333333332|
|        IT|           17500.0|
|      Mech|           28000.0|
|        DS|           25000.0|
|     Cloud|           30000.0|
+----------+------------------+
```

```
df_pyspark.groupBy('Department').count().show()
```

```
+----------+-----+
|Department|count|
+----------+-----+
|      Data|    1|
|   Digital|    1|
|    Design|    3|
|        IT|    2|
|      Mech|    1|
|        DS|    1|
|     Cloud|    1|
+----------+-----+
```

```
df_pyspark.agg({'Salary':'sum'}).show()
```

```
+-----------+
|sum(Salary)|
+-----------+
|     245000|
+-----------+
```