```
!pip install pyspark
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-w
Collecting pyspark
  Downloading pyspark-3.2.1.tar.gz (281.4 MB)
     |████████████████████████████████| 281.4 MB 34 kB/s
Collecting py4j==0.10.9.3
  Downloading py4j-0.10.9.3-py2.py3-none-any.whl (198 kB)
     |████████████████████████████████| 198 kB 64.0 MB/s
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.2.1-py2.py3-none-any.whl size=
  Stored in directory: /root/.cache/pip/wheels/9f/f5/07/7cd8017084dce4e93e84e9
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9.3 pyspark-3.2.1
```

```
from pyspark.sql import SparkSession
spark=SparkSession.builder.appName('DataFrame').getOrCreate()
```

```
spark
```

SparkSession - in-memory

SparkContext

Spark UI

Version
    `v3.2.1`
Master
    `local[*]`
AppName
    `DataFrame`

```
df=spark.read.csv('student.csv')
df.show()
```

```
+------+----+---+------+
|   _c0| _c1|_c2|   _c3|
+------+----+---+------+
|stu_id|name|age|course|
|   100| aaa| 10|Devops|
|   101| bbb| 20|    DE|
|   102| aaa| 30|Devops|
|   103| bbb| 40|    DE|
|   104| aaa| 50|Devops|
|   105| bbb| 60|    DE|
|   106| aaa| 70|Devops|
|   107| bbb| 80|    DE|
|   108| aaa| 90|Devops|
|   109| bbb|100|    DE|
+------+----+---+------+
df=spark.read.csv('student.csv',header=True,inferSchema=True)
df.show()
```

```
+------+----+---+------+
|stu_id|name|age|course|
+------+----+---+------+
|   100| aaa| 10|Devops|
|   101| bbb| 20|    DE|
|   102| aaa| 30|Devops|
|   103| bbb| 40|    DE|
|   104| aaa| 50|Devops|
|   105| bbb| 60|    DE|
|   106| aaa| 70|Devops|
|   107| bbb| 80|    DE|
|   108| aaa| 90|Devops|
|   109| bbb|100|    DE|
+------+----+---+------+
```

```
df.printSchema()
```

```
root
 |-- stu_id: integer (nullable = true)
 |-- name: string (nullable = true)
 |-- age: integer (nullable = true)
 |-- course: string (nullable = true)
```

```
df.head(3)
```

```
[Row(stu_id=100, name='aaa', age=10, course='Devops'),
 Row(stu_id=101, name='bbb', age=20, course='DE'),
 Row(stu_id=102, name='aaa', age=30, course='Devops')]
```

```
df.show()
```

```
+------+----+---+------+
|stu_id|name|age|course|
+------+----+---+------+
|   100| aaa| 10|Devops|
|   101| bbb| 20|    DE|
|   102| aaa| 30|Devops|
|   103| bbb| 40|    DE|
|   104| aaa| 50|Devops|
|   105| bbb| 60|    DE|
|   106| aaa| 70|Devops|
|   107| bbb| 80|    DE|
|   108| aaa| 90|Devops|
|   109| bbb|100|    DE|
+------+----+---+------+
```

```
df.select(['stu_id','age']).show()
```

```
+------+---+
|stu_id|age|
+------+---+
|   100| 10|
|   101| 20|
|   102| 30|
|   103| 40|
|   104| 50|
|   105| 60|
|   106| 70|
|   107| 80|
|   108| 90|
|   109|100|
+------+---+
```

```
df.dtypes
```

```
[('stu_id', 'int'), ('name', 'string'), ('age', 'int'), ('course', 'string')]
```

```
df.schema
```

```
StructType(List(StructField(stu_id,IntegerType,true),StructField(name,StringTy
```

```
df.count()
```

```
10
```

```
df.describe().show()
```

```
+-------+------------------+----+------------------+------+
|summary|            stu_id|name|               age|course|
+-------+------------------+----+------------------+------+
|  count|                10|  10|                10|    10|
|   mean|             104.5|null|              55.0|  null|
| stddev|3.0276503540974917|null|30.276503540974915|  null|
|    min|               100| aaa|                10|    DE|
|    max|               109| bbb|               100|Devops|
+-------+------------------+----+------------------+------+
```

```
df=df.withColumn('Age after 10 years',df['age']+10)
df.show()
```

```
+------+----+---+------+------------------+
|stu_id|name|age|course|Age after 10 years|
+------+----+---+------+------------------+
|   100| aaa| 10|Devops|                20|
|   101| bbb| 20|    DE|                30|
|   102| aaa| 30|Devops|                40|
|   103| bbb| 40|    DE|                50|
|   104| aaa| 50|Devops|                60|
|   105| bbb| 60|    DE|                70|
|   106| aaa| 70|Devops|                80|
|   107| bbb| 80|    DE|                90|
|   108| aaa| 90|Devops|               100|
|   109| bbb|100|    DE|                110|
+------+----+---+------+------------------+
```

```
df=df.drop('Age after 10 years')
df.show()
```

```
+------+----+------+
|stu_id|name|course|
+------+----+------+
|   100| aaa|Devops|
|   101| bbb|    DE|
|   102| aaa|Devops|
|   103| bbb|    DE|
|   104| aaa|Devops|
|   105| bbb|    DE|
|   106| aaa|Devops|
|   107| bbb|    DE|
|   108| aaa|Devops|
|   109| bbb|    DE|
+------+----+------+
```

```
df.show()
```

```
+------+----+---+------+
|stu_id|name|age|course|
+------+----+---+------+
|   100| aaa| 10|Devops|
|   101| bbb| 20|    DE|
|   102| aaa| 30|Devops|
|   103| bbb| 40|    DE|
|   104| aaa| 50|Devops|
|   105| bbb| 60|    DE|
|   106| aaa| 70|Devops|
|   107| bbb| 80|    DE|
|   108| aaa| 90|Devops|
|   109| bbb|100|    DE|
+------+----+---+------+
```

```
df=spark.read.csv('student1.csv',header=True,inferSchema=True)
df.show()
```

```
+------+----+----+---------+
|stu_id|name| age|coursefee|
+------+----+----+---------+
|   110| aaa|  10|     1000|
|   111| bbb|  20|     2000|
|   112| aaa|  30|     1000|
|   113| bbb|null|     2000|
|   104| aaa|  50|     1000|
|   105| bbb|  60|     2000|
|   106| aaa|  70|     1000|
|   107| bbb|  80|     2000|
|  null| aaa|  90|     1000|
|  null| bbb| 100|     2000|
+------+----+----+---------+
```

```
df.na.drop().show()
```

```
+------+----+---+---------+
|stu_id|name|age|coursefee|
+------+----+---+---------+
|   110| aaa| 10|     1000|
|   111| bbb| 20|     2000|
|   112| aaa| 30|     1000|
|   104| aaa| 50|     1000|
|   105| bbb| 60|     2000|
|   106| aaa| 70|     1000|
|   107| bbb| 80|     2000|
+------+----+---+---------+
```

```
df.na.drop(how="any").show()
```

```
+------+----+---+---------+
|stu_id|name|age|coursefee|
+------+----+---+---------+
|   110| aaa| 10|     1000|
|   111| bbb| 20|     2000|
|   112| aaa| 30|     1000|
|   104| aaa| 50|     1000|
|   105| bbb| 60|     2000|
|   106| aaa| 70|     1000|
|   107| bbb| 80|     2000|
+------+----+---+---------+
```

```
df.na.drop(how="any",thresh=1).show()
```

```
+------+----+----+---------+
|stu_id|name| age|coursefee|
+------+----+----+---------+
|   110| aaa|  10|     1000|
|   111| bbb|  20|     2000|
|   112| aaa|  30|     1000|
|   113| bbb|null|     2000|
|   104| aaa|  50|     1000|
|   105| bbb|  60|     2000|
|   106| aaa|  70|     1000|
|   107| bbb|  80|     2000|
|  null| aaa|  90|     1000|
|  null| bbb| 100|     2000|
+------+----+----+---------+
```

```
df.na.drop(how="any",subset=['stu_id']).show()
```

```
+------+----+----+---------+
|stu_id|name| age|coursefee|
+------+----+----+---------+
|   110| aaa|  10|     1000|
|   111| bbb|  20|     2000|
|   112| aaa|  30|     1000|
|   113| bbb|null|     2000|
|   104| aaa|  50|     1000|
|   105| bbb|  60|     2000|
|   106| aaa|  70|     1000|
|   107| bbb|  80|     2000|
+------+----+----+---------+
```

```
df.show()
```

```
+------+----+----+---------+
|stu_id|name| age|coursefee|
+------+----+----+---------+
|   110| aaa|  10|     1000|
|   111| bbb|  20|     2000|
|   112| aaa|  30|     1000|
|   113| bbb|null|     2000|
|   104| aaa|  50|     1000|
|   105| bbb|  60|     2000|
|   106| aaa|  70|     1000|
|   107| bbb|  80|     2000|
|  null| aaa|  90|     1000|
|  null| bbb| 100|     2000|
+------+----+----+---------+
```

```
from pyspark.ml.feature import Imputer
imputer = Imputer(
inputCols=['stu_id','age'] ,
outputCols= ["{}_imputed".format(c) for c in ['stu_id','age']]).setStrategy("mean")
```

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

```
+------+----+----+---------+--------------+-----------+
|stu_id|name| age|coursefee|stu_id_imputed|age_imputed|
+------+----+----+---------+--------------+-----------+
|   110| aaa|  10|     1000|           110|         10|
|   111| bbb|  20|     2000|           111|         20|
|   112| aaa|  30|     1000|           112|         30|
|   113| bbb|null|     2000|           113|         56|
|   104| aaa|  50|     1000|           104|         50|
|   105| bbb|  60|     2000|           105|         60|
|   106| aaa|  70|     1000|           106|         70|
|   107| bbb|  80|     2000|           107|         80|
|  null| aaa|  90|     1000|           108|         90|
|  null| bbb| 100|     2000|           108|        100|
+------+----+----+---------+--------------+-----------+
```

```
df.write.option('header','true').saveAsTable("stu")
```

```
spark.sql("select * from stu")
```

```
DataFrame[stu_id: int, name: string, age: int, coursefee: int]
```

```
spark.sql("select * from stu").show()
```

```
+------+----+----+---------+
|stu_id|name| age|coursefee|
+------+----+----+---------+
|   110| aaa|  10|     1000|
|   111| bbb|  20|     2000|
|   112| aaa|  30|     1000|
|   113| bbb|null|     2000|
|   104| aaa|  50|     1000|
|   105| bbb|  60|     2000|
|   106| aaa|  70|     1000|
|   107| bbb|  80|     2000|
|  null| aaa|  90|     1000|
|  null| bbb| 100|     2000|
+------+----+----+---------+
```

```
df.select(["stu_id","name"]).show()
```

```
+------+----+
|stu_id|name|
+------+----+
|   110| aaa|
|   111| bbb|
|   112| aaa|
|   113| bbb|
|   104| aaa|
|   105| bbb|
|   106| aaa|
|   107| bbb|
|  null| aaa|
|  null| bbb|
+------+----+
```

```
spark.sql("select count(*) from stu").show()
```

```
+--------+
|count(1)|
+--------+
|      10|
+--------+
```

```
spark.sql("select max(age) from stu").show()
```

```
+--------+
|max(age)|
+---------+
|      100|
+---------+
```

```
df.groupBy("name").count().show()
```

```
+----+-----+
|name|count|
+----+-----+
| aaa|    5|
| bbb|    5|
+----+-----+
```

✓ 0s completed at 8:04 PM ● ✕