

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

```
spark = SparkSession \
    .builder \
    .appName("Python Spark SQL basic example") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()
```

```
spark
```



SparkSession - in-memory

SparkContext

[Spark UI](#)

Version

v3.5.5

Master

local[*]

AppName

Python Spark SQL basic example

```
import os
os.getcwd()
```



'/content'

```
df = spark.read.csv("diabetes.csv", header=True,inferSchema=True)
```

```
df.show(20)
```



Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50.0	1
1	85	66	29	0	26.6	0.351	31.0	0
8	183	64	0	0	23.3	0.672	32.0	1
1	89	NULL	23	94	NULL	NULL	NaN	0
0	137	40	35	168	43.1	2.288	33.0	1
5	NULL	74	0	0	25.6	0.201	30.0	0
3	78	50	32	88	31	0.248	26.0	1
10	115	0	0	0	35.3	0.134	29.0	0
2	197	70	45	543	30.5	0.158	53.0	1
8	125	96	0	0	0	0.232	54.0	1
4	110	?	0	0	37.6	0.191	30.0	0
10	168	74	0	0	38	0.537	34.0	1
10	139	80	0	0	27.1	1.441	57.0	0
1	189	60	23	846	30.1	0.398	59.0	1
5	166	72	19	175	25.8	0.587	51.0	1
7	100	0	0	0	30	0.484	32.0	1
0	118	84	47	230	45.8	0.551	31.0	1
7	##	74	0	0	29.6	0.254	31.0	1
1	103	30	38	83	43.3	0.183	33.0	0
1	115	70	30	96	34.6	0.529	32.0	1

only showing top 20 rows

```
df.head(5)
```



```
[Row(Pregnancies=6, Glucose='148', BloodPressure='72', SkinThickness=35, Insulin=0, BMI='33.6', DiabetesPedigreeFunction=0.627,
Age=50.0, Outcome=1),
 Row(Pregnancies=1, Glucose='85', BloodPressure='66', SkinThickness=29, Insulin=0, BMI='26.6', DiabetesPedigreeFunction=0.351,
Age=31.0, Outcome=0),
 Row(Pregnancies=8, Glucose='183', BloodPressure='64', SkinThickness=0, Insulin=0, BMI='23.3', DiabetesPedigreeFunction=0.672,
Age=32.0, Outcome=1),
 Row(Pregnancies=1, Glucose='89', BloodPressure=None, SkinThickness=23, Insulin=94, BMI=None, DiabetesPedigreeFunction=None,
Age=nan, Outcome=0),
 Row(Pregnancies=0, Glucose='137', BloodPressure='40', SkinThickness=35, Insulin=168, BMI='43.1', DiabetesPedigreeFunction=2.288,
Age=33.0, Outcome=1)]
```

```
type(df)
```



```
pyspark.sql.dataframe.DataFrame
def __init__(jdf: JavaObject, sql_ctx: Union['SQLContext', 'SparkSession'])
```

```
+-----+-----+-----+-----+
|      ML|      F|    150.0|      60|
|PySpark|      M|    75.0|      50|
+-----+-----+-----+-----+
```

Notes

```
df.printSchema()
```



```
root
|-- Pregnancies: integer (nullable = true)
|-- Glucose: string (nullable = true)
|-- BloodPressure: string (nullable = true)
|-- SkinThickness: integer (nullable = true)
|-- Insulin: integer (nullable = true)
|-- BMI: string (nullable = true)
|-- DiabetesPedigreeFunction: double (nullable = true)
|-- Age: double (nullable = true)
|-- Outcome: integer (nullable = true)
```

```
df.select("Glucose").show(5)
```



```
+-----+
|Glucose|
+-----+
|   148|
|    85|
|   183|
|    89|
|   137|
+-----+
only showing top 5 rows
```

```
df1 = df.select(df['Glucose'],df['Age']+1)
```

```
df1.show()
```



```
+-----+-----+
|Glucose|(Age + 1)|
+-----+-----+
|   148|    51.0|
|    85|    32.0|
|   183|    33.0|
|    89|     NaN|
|   137|    34.0|
|   NULL|    31.0|
|    78|    27.0|
|   115|    30.0|
|   197|    54.0|
|   125|    55.0|
|   110|    31.0|
|   168|    35.0|
|   139|    58.0|
|   189|    60.0|
|   166|    52.0|
|   100|    33.0|
|   118|    32.0|
|    ##|    32.0|
|   103|    34.0|
|   115|    33.0|
+-----+-----+
only showing top 20 rows
```

```
df['Glucose']
```



```
Column<'Glucose'>
```

```
df[['Glucose','Age']]
```



```
DataFrame[Glucose: string, Age: double]
```

```
df.dtypes
```



```
[('Pregnancies', 'int'),
 ('Glucose', 'string'),
 ('BloodPressure', 'string'),
```

```
( 'SkinThickness', 'int'),
( 'Insulin', 'int'),
( 'BMI', 'string'),
( 'DiabetesPedigreeFunction', 'double'),
( 'Age', 'double'),
( 'Outcome', 'int')]
```

```
dist_bp = df.select('BloodPressure').distinct()
```

```
dist_bp.show()
```

```
↗ +-----+
|BloodPressure|
+-----+
|          54|
|          64|
|          30|
|          85|
|          52|
|           0|
|          98|
|         110|
|          96|
|         100|
|          70|
|          61|
|          75|
|          46|
|          78|
|          60|
|          90|
|          68|
|         104|
|         102|
+-----+
only showing top 20 rows
```

```
df.select('Pregnancies').distinct().count()
```

```
↗ 17
```

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

```
↗ +-----+-----+-----+-----+-----+-----+
|summary|Pregnancies|Glucose|BloodPressure|SkinThickness|Insulin|BMI|DiabetesF|
+-----+-----+-----+-----+-----+-----+
|count|768|767|767|768|768|767| |
|mean|3.8450520833333335|120.91906005221932|69.07963446475196|20.536458333333332|79.79947916666667|31.993211488250633|0.47|
|stddev|3.36957806269887|32.009944976772395|19.363065113384383|15.952217567727642|115.24400235133803|7.892243623420615|0.33|
|min|0|##|0|0|0|0|
|max|17|99|?|99|846|?|
+-----+-----+-----+-----+-----+-----+
◀────────────────────────────────────────────────────────────────────────────────▶
```

```
df.describe(['BMI', 'Glucose']).show()
```

```
↗ +-----+-----+
|summary|BMI|Glucose|
+-----+-----+
|count|767|767|
|mean|31.993211488250633|120.91906005221932|
|stddev|7.892243623420615|32.009944976772395|
|min|0|##|
|max|?|99|
+-----+-----+
```

```
df.describe(['Age']).show()
```

```
↗ +-----+
|summary|Age|
+-----+
|count|768|
|mean|NaN|
|stddev|NaN|
|min|21.0|
|max|NaN|
+-----+
```

```
df['Age']
```

Column<'Age'>

```
#adding column
df_new = df.withColumn('New Age',df.Age+2)
```

```
df_new.show()
```

```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|Pregnancies|Glucose|BloodPressure|SkinThickness|Insulin| BMI|DiabetesPedigreeFunction| Age|Outcome|New Age|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|      6|    148|          72|         35|      0|33.6|          0.627|50.0|      1|    52.0|
|      1|     85|          66|         29|      0|26.6|          0.351|31.0|      0|    33.0|
|      8|    183|          64|          0|      0|23.3|          0.672|32.0|      1|    34.0|
|      1|     89|        NULL|         23|     94|NULL|          NULL|NaN|      0|     NaN|
|      0|    137|          40|         35|    168|43.1|          2.288|33.0|      1|    35.0|
|      5|    NULL|          74|          0|      0|25.6|          0.201|30.0|      0|    32.0|
|      3|     78|          50|         32|     88| 31|          0.248|26.0|      1|    28.0|
|     10|    115|          0|          0|      0|35.3|          0.134|29.0|      0|    31.0|
|      2|    197|          70|         45|    543|30.5|          0.158|53.0|      1|    55.0|
|      8|    125|          96|          0|      0| 0|          0.232|54.0|      1|    56.0|
|      4|    110|          ?|          0|      0|37.6|          0.191|30.0|      0|    32.0|
|     10|    168|          74|          0|      0| 38|          0.537|34.0|      1|    36.0|
|     10|    139|          80|          0|      0|27.1|          1.441|57.0|      0|    59.0|
|      1|    189|          60|         23|    846|30.1|          0.398|59.0|      1|    61.0|
|      5|    166|          72|         19|    175|25.8|          0.587|51.0|      1|    53.0|
|      7|    100|          0|          0|      0| 30|          0.484|32.0|      1|    34.0|
|      0|    118|          84|         47|    230|45.8|          0.551|31.0|      1|    33.0|
|      7|     ##|          74|          0|      0|29.6|          0.254|31.0|      1|    33.0|
|      1|    103|          30|         38|     83|43.3|          0.183|33.0|      0|    35.0|
|      1|    115|          70|         30|     96|34.6|          0.529|32.0|      1|    34.0|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

```

only showing top 20 rows

```
#df.select(isnull())
```

```
df_new= df_new.drop('New Age')
```

```
df_new.show()
```

```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|Pregnancies|Glucose|BloodPressure|SkinThickness|Insulin| BMI|DiabetesPedigreeFunction| Age|Outcome|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|      6|    148|          72|         35|      0|33.6|          0.627|50.0|      1|
|      1|     85|          66|         29|      0|26.6|          0.351|31.0|      0|
|      8|    183|          64|          0|      0|23.3|          0.672|32.0|      1|
|      1|     89|        NULL|         23|     94|NULL|          NULL|NaN|      0|
|      0|    137|          40|         35|    168|43.1|          2.288|33.0|      1|
|      5|    NULL|          74|          0|      0|25.6|          0.201|30.0|      0|
|      3|     78|          50|         32|     88| 31|          0.248|26.0|      1|
|     10|    115|          0|          0|      0|35.3|          0.134|29.0|      0|
|      2|    197|          70|         45|    543|30.5|          0.158|53.0|      1|
|      8|    125|          96|          0|      0| 0|          0.232|54.0|      1|
|      4|    110|          ?|          0|      0|37.6|          0.191|30.0|      0|
|     10|    168|          74|          0|      0| 38|          0.537|34.0|      1|
|     10|    139|          80|          0|      0|27.1|          1.441|57.0|      0|
|      1|    189|          60|         23|    846|30.1|          0.398|59.0|      1|
|      5|    166|          72|         19|    175|25.8|          0.587|51.0|      1|
|      7|    100|          0|          0|      0| 30|          0.484|32.0|      1|
|      0|    118|          84|         47|    230|45.8|          0.551|31.0|      1|
|      7|     ##|          74|          0|      0|29.6|          0.254|31.0|      1|
|      1|    103|          30|         38|     83|43.3|          0.183|33.0|      0|
|      1|    115|          70|         30|     96|34.6|          0.529|32.0|      1|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

```

only showing top 20 rows

```
df_null = df.filter(df['BloodPressure'].isNull())
```

```
df_null.show()
```

```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|Pregnancies|Glucose|BloodPressure|SkinThickness|Insulin| BMI|DiabetesPedigreeFunction|Age|Outcome|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|      1|     89|        NULL|         23|     94|NULL|          NULL|NaN|      0|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

```

```
df.select('Pregnancies').distinct().show()
```

```

+-----+
|Pregnancies|

```

```
+-----+
|      12|
|       1|
|      13|
|       6|
|       3|
|       5|
|      15|
|       9|
|      17|
|       4|
|       8|
|       7|
|      10|
|      11|
|      14|
|       2|
|       0|
+-----+
```

```
df.show(20)
```

```
→ +-----+-----+-----+-----+-----+-----+-----+-----+-----+
|Pregnancies|Glucose|BloodPressure|SkinThickness|Insulin| BMI|DiabetesPedigreeFunction| Age|Outcome|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|      6|    148|      72|      35|    0|33.6|      0.627|50.0|    1|
|     1|     85|     66|     29|    0|26.6|      0.351|31.0|    0|
|     8|    183|     64|      0|    0|23.3|      0.672|32.0|    1|
|     1|     89|    NULL|     23|    94|NULL|      NULL|NaN|    0|
|     0|    137|     40|     35|   168|43.1|      2.288|33.0|    1|
|     5|    NULL|     74|      0|    0|25.6|      0.201|30.0|    0|
|     3|     78|     50|     32|    88| 31|      0.248|26.0|    1|
|    10|    115|      0|      0|    0|35.3|      0.134|29.0|    0|
|     2|    197|     70|     45|   543|30.5|      0.158|53.0|    1|
|     8|    125|     96|      0|    0| 0|      0.232|54.0|    1|
|     4|    110|      ?|      0|    0|37.6|      0.191|30.0|    0|
|    10|    168|     74|      0|    0| 38|      0.537|34.0|    1|
|    10|    139|     80|      0|    0|27.1|      1.441|57.0|    0|
|     1|    189|     60|     23|   846|30.1|      0.398|59.0|    1|
|     5|    166|     72|     19|   175|25.8|      0.587|51.0|    1|
|     7|    100|      0|      0|    0| 30|      0.484|32.0|    1|
|     0|    118|     84|     47|   230|45.8|      0.551|31.0|    1|
|     7|     ##|     74|      0|    0|29.6|      0.254|31.0|    1|
|     1|    103|     30|     38|    83|43.3|      0.183|33.0|    0|
|     1|    115|     70|     30|    96|34.6|      0.529|32.0|    1|
+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 20 rows
```

```
df1 = df.na.drop(how='any')
```

```
df1.show(20)
```

```
→ +-----+-----+-----+-----+-----+-----+-----+-----+-----+
|Pregnancies|Glucose|BloodPressure|SkinThickness|Insulin| BMI|DiabetesPedigreeFunction| Age|Outcome|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|      6|    148|      72|      35|    0|33.6|      0.627|50.0|    1|
|     1|     85|     66|     29|    0|26.6|      0.351|31.0|    0|
|     8|    183|     64|      0|    0|23.3|      0.672|32.0|    1|
|     0|    137|     40|     35|   168|43.1|      2.288|33.0|    1|
|     3|     78|     50|     32|    88| 31|      0.248|26.0|    1|
|    10|    115|      0|      0|    0|35.3|      0.134|29.0|    0|
|     2|    197|     70|     45|   543|30.5|      0.158|53.0|    1|
|     8|    125|     96|      0|    0| 0|      0.232|54.0|    1|
|     4|    110|      ?|      0|    0|37.6|      0.191|30.0|    0|
|    10|    168|     74|      0|    0| 38|      0.537|34.0|    1|
|    10|    139|     80|      0|    0|27.1|      1.441|57.0|    0|
|     1|    189|     60|     23|   846|30.1|      0.398|59.0|    1|
|     5|    166|     72|     19|   175|25.8|      0.587|51.0|    1|
|     7|    100|      0|      0|    0| 30|      0.484|32.0|    1|
|     0|    118|     84|     47|   230|45.8|      0.551|31.0|    1|
|     7|     ##|     74|      0|    0|29.6|      0.254|31.0|    1|
|     1|    103|     30|     38|    83|43.3|      0.183|33.0|    0|
|     1|    115|     70|     30|    96|34.6|      0.529|32.0|    1|
|     3|    126|     88|     41|   235|39.3|      0.704|27.0|    0|
|     8|     99|     84|      0|    0| ?|      0.388|50.0|    0|
+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 20 rows
```

```
df1 = df.na.drop(how='any', thresh=7)
```

```
df1.show(20)
```

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50.0	1
1	85	66	29	0	26.6	0.351	31.0	0
8	183	64	0	0	23.3	0.672	32.0	1
0	137	40	35	168	43.1	2.288	33.0	1
5	NULL	74	0	0	25.6	0.201	30.0	0
3	78	50	32	88	31	0.248	26.0	1
10	115	0	0	0	35.3	0.134	29.0	0
2	197	70	45	543	30.5	0.158	53.0	1
8	125	96	0	0	0	0.232	54.0	1
4	110	?	0	0	37.6	0.191	30.0	0
10	168	74	0	0	38	0.537	34.0	1
10	139	80	0	0	27.1	1.441	57.0	0
1	189	60	23	846	30.1	0.398	59.0	1
5	166	72	19	175	25.8	0.587	51.0	1
7	100	0	0	0	30	0.484	32.0	1
0	118	84	47	230	45.8	0.551	31.0	1
7	##	74	0	0	29.6	0.254	31.0	1
1	103	30	38	83	43.3	0.183	33.0	0
1	115	70	30	96	34.6	0.529	32.0	1
3	126	88	41	235	39.3	0.704	27.0	0

only showing top 20 rows

```
df1 = df.na.drop(how='any',subset=['Glucose'])
```

```
df1.show(20)
```

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50.0	1
1	85	66	29	0	26.6	0.351	31.0	0
8	183	64	0	0	23.3	0.672	32.0	1
1	89	NULL	23	94	NULL	NULL	NaN	0
0	137	40	35	168	43.1	2.288	33.0	1
3	78	50	32	88	31	0.248	26.0	1
10	115	0	0	0	35.3	0.134	29.0	0
2	197	70	45	543	30.5	0.158	53.0	1
8	125	96	0	0	0	0.232	54.0	1
4	110	?	0	0	37.6	0.191	30.0	0
10	168	74	0	0	38	0.537	34.0	1
10	139	80	0	0	27.1	1.441	57.0	0
1	189	60	23	846	30.1	0.398	59.0	1
5	166	72	19	175	25.8	0.587	51.0	1
7	100	0	0	0	30	0.484	32.0	1
0	118	84	47	230	45.8	0.551	31.0	1
7	##	74	0	0	29.6	0.254	31.0	1
1	103	30	38	83	43.3	0.183	33.0	0
1	115	70	30	96	34.6	0.529	32.0	1
3	126	88	41	235	39.3	0.704	27.0	0

only showing top 20 rows

```
df1 = df.na.fill("Missing",subset=["BloodPressure","Glucose"])
```

```
df1.show(20)
```

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50.0	1
1	85	66	29	0	26.6	0.351	31.0	0
8	183	64	0	0	23.3	0.672	32.0	1
1	89	Missing	23	94	NULL	NULL	NaN	0
0	137	40	35	168	43.1	2.288	33.0	1
5	Missing	74	0	0	25.6	0.201	30.0	0
3	78	50	32	88	31	0.248	26.0	1
10	115	0	0	0	35.3	0.134	29.0	0
2	197	70	45	543	30.5	0.158	53.0	1
8	125	96	0	0	0	0.232	54.0	1
4	110	?	0	0	37.6	0.191	30.0	0
10	168	74	0	0	38	0.537	34.0	1
10	139	80	0	0	27.1	1.441	57.0	0
1	189	60	23	846	30.1	0.398	59.0	1
5	166	72	19	175	25.8	0.587	51.0	1
7	100	0	0	0	30	0.484	32.0	1
0	118	84	47	230	45.8	0.551	31.0	1
7	##	74	0	0	29.6	0.254	31.0	1
1	103	30	38	83	43.3	0.183	33.0	0
1	115	70	30	96	34.6	0.529	32.0	1

only showing top 20 rows

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

```
imputer = Imputer(inputCols=['Age'], outputCols=['Imputed Age'],strategy="mean")
```

```
imputer.fit(df).transform(df).show(20)
```

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	Imputed Age
6	148	72	35	0	33.6	0.627	50.0	1	50.0
1	85	66	29	0	26.6	0.351	31.0	0	31.0
8	183	64	0	0	23.3	0.672	32.0	1	32.0
1	89	NULL	23	94	NULL	NULL	NaN	0	33.25684485006519
0	137	40	35	168	43.1	2.288	33.0	1	33.0
5	NULL	74	0	0	25.6	0.201	30.0	0	30.0
3	78	50	32	88	31	0.248	26.0	1	26.0
10	115	0	0	0	35.3	0.134	29.0	0	29.0
2	197	70	45	543	30.5	0.158	53.0	1	53.0
8	125	96	0	0	0	0.232	54.0	1	54.0
4	110	?	0	0	37.6	0.191	30.0	0	30.0
10	168	74	0	0	38	0.537	34.0	1	34.0
10	139	80	0	0	27.1	1.441	57.0	0	57.0
1	189	60	23	846	30.1	0.398	59.0	1	59.0
5	166	72	19	175	25.8	0.587	51.0	1	51.0
7	100	0	0	0	30	0.484	32.0	1	32.0
0	118	84	47	230	45.8	0.551	31.0	1	31.0
7	##	74	0	0	29.6	0.254	31.0	1	31.0
1	103	30	38	83	43.3	0.183	33.0	0	33.0
1	115	70	30	96	34.6	0.529	32.0	1	32.0

only showing top 20 rows

```
df.filter("Age>50").show()
```

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1	89	NULL	23	94	NULL	NULL	NaN	0
2	197	70	45	543	30.5	0.158	53.0	1
8	125	96	0	0	0	0.232	54.0	1
10	139	80	0	0	27.1	1.441	57.0	0
1	189	60	23	846	30.1	0.398	59.0	1
5	166	72	19	175	25.8	0.587	51.0	1
11	143	94	33	146	36.6	0.254	51.0	1
13	145	82	19	110	22.2	0.245	57.0	0
5	109	75	26	0	36	0.546	60.0	0
4	111	72	47	207	37.1	1.39	56.0	1
9	171	110	24	240	45.4	0.721	54.0	1
8	176	90	34	300	33.7	0.467	58.0	1
2	109	92	0	0	42.7	0.845	54.0	0
4	134	72	0	0	23.8	0.277	60.0	1
4	146	92	0	0	31.2	0.539	61.0	1
5	132	80	0	0	26.8	0.186	69.0	0
0	105	84	0	0	27.9	0.741	62.0	1
3	128	78	0	0	21.1	0.268	55.0	0
5	147	78	0	0	33.7	0.218	65.0	0
8	181	68	36	495	30.1	0.615	60.0	1

only showing top 20 rows

```
df.filter("Age>50").select(['Age','Insulin','Glucose','Outcome']).show()
```

Age	Insulin	Glucose	Outcome
NaN	94	89	0
53.0	543	197	1
54.0	0	125	1
57.0	0	139	0
59.0	846	189	1
51.0	175	166	1
51.0	146	143	1
57.0	110	145	0
60.0	0	109	0
56.0	207	111	1
54.0	240	171	1
58.0	300	176	1
54.0	0	109	0
60.0	0	134	1
61.0	0	146	1

```

|69.0|    0|    132|    0|
|62.0|    0|    105|    1|
|55.0|    0|    128|    0|
|65.0|    0|    147|    0|
|60.0|  495|    181|    1|
+---+-----+-----+-----+
only showing top 20 rows

```

```
df.filter(~(df['Age']>50) & (df['Glucose']>175)).select(['Age','Insulin','Glucose','Outcome']).show()
```

```

↵ +---+-----+-----+-----+
  | Age|Insulin|Glucose|Outcome|
  +---+-----+-----+-----+
  |32.0|    0|    183|    1|
  |41.0|    0|    196|    1|
  |26.0|   70|    180|    0|
  |25.0|    0|    180|    1|
  |41.0|  304|    187|    1|
  |43.0|    0|    188|    1|
  |36.0|  130|    179|    1|
  |41.0|    0|    194|    1|
  |41.0|    0|    184|    1|
  |21.0|  478|    177|    1|
  |31.0|  744|    197|    0|
  |23.0|    0|    179|    1|
  |49.0|    0|    184|    1|
  |24.0|  375|    193|    0|
  |34.0|  130|    191|    0|
  |29.0|    0|    182|    1|
  |37.0|    0|    179|    0|
  |41.0|    0|    180|    1|
  |41.0|    0|    178|    1|
  |29.0|  249|    196|    1|
  +---+-----+-----+-----+
only showing top 20 rows

```

Start coding or [generate](#) with AI.

Start coding or [generate](#) with AI.

Start coding or [generate](#) with AI.

```
df.filter((df['Age']>50) & (df['Glucose']>175)).select(['Age','Insulin','Glucose','Outcome']).show()
```

```

↵ +---+-----+-----+-----+
  | Age|Insulin|Glucose|Outcome|
  +---+-----+-----+-----+
  |53.0|   543|    197|    1|
  |59.0|   846|    189|    1|
  |58.0|   300|    176|    1|
  |60.0|   495|    181|    1|
  |57.0|   280|    196|    1|
  |60.0|    0|    179|    0|
  |51.0|   192|    181|    1|
  |59.0|    0|    194|    1|
  |67.0|    0|    194|    0|
  |55.0|   145|    195|    1|
  |53.0|   207|    187|    1|
  |62.0|    0|    197|    1|
  |52.0|   156|    176|    1|
  |66.0|    0|    190|    1|
  +---+-----+-----+-----+

```

```
df.groupBy(['Age']).count().show()
```

```

↵ +---+-----+
  | Age|count|
  +---+-----+
  |67.0|    3|
  |70.0|    1|
  |69.0|    2|
  | NaN|    1|
  |49.0|    5|
  |29.0|   29|
  |64.0|    1|
  |47.0|    6|
  |42.0|   18|
  |44.0|    8|
  |35.0|   10|
  |62.0|    4|
  |39.0|   12|

```



```
|34.0| 14|
|37.0| 19|
|25.0| 48|
|36.0| 16|
|41.0| 22|
|23.0| 38|
|50.0| 8|
+-----+
only showing top 20 rows
```

```
df.groupBy('BloodPressure').max().show()
```

```
↗ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+
|BloodPressure|max(Pregnancies)|max(SkinThickness)|max(Insulin)|max(DiabetesPedigreeFunction)|max(Age)|max(Outcome)|
+-----+ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+
|          54|          7|          32|          175|          0.748|          62.0|          1|
|          64|          8|          46|          415|          1.731|          41.0|          1|
|          30|          1|          42|          99|          0.496|          33.0|          1|
|          85|         12|          54|         100|          1.213|          56.0|          1|
|          52|          9|          43|          540|          1.699|          42.0|          1|
|           0|         13|          30|           0|          0.933|          72.0|          1|
|          98|         10|          41|          58|          1.321|          34.0|          1|
|         110|          9|          46|         240|          0.721|          54.0|          1|
|         NULL|          1|          23|          94|          NULL|          NaN|           0|
|          96|          8|           0|           0|          0.268|          54.0|          1|
|         100|          8|          39|         240|          1.021|          45.0|          1|
|          70|         15|          99|         744|          2.329|          63.0|          1|
|          61|          6|           0|           0|          0.151|          55.0|           0|
|          75|          5|          32|           0|          0.572|          60.0|          1|
|          46|          2|          21|          335|          0.654|          22.0|           0|
|          78|         14|          63|          293|          2.42|          67.0|          1|
|          60|         13|          46|         846|          1.072|          67.0|          1|
|          90|         13|          51|         680|          0.805|          66.0|          1|
|          68|         11|          49|         579|          1.391|          60.0|          1|
|         104|          5|          25|           0|          0.435|          52.0|          1|
+-----+ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+
only showing top 20 rows
```

```
df.groupBy('BloodPressure').mean().show()
```

```
↗ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+
|BloodPressure| avg(Pregnancies)|avg(SkinThickness)| avg(Insulin)|avg(DiabetesPedigreeFunction)| avg(Age)| avg
+-----+ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+
|          54| 2.727272727272727|16.363636363636363| 76.81818181818181| 0.4329090909090909|29.181818181818183|0.181818181
|          64|2.4186046511627906|23.348837209302324| 90.13953488372093| 0.48055813953488374| 26.13953488372093| 0.30232558
|          30|          1.0|          40.0|          91.0|          0.3395|          29.5|
|          85| 5.833333333333333|31.0|22.666666666666668| 0.5155|36.833333333333336|
|          52|2.4545454545454546|16.454545454545453| 91.45454545454545| 0.48700000000000004|25.181818181818183| 0.27272727
|           0|3.6285714285714286|1.5142857142857142|           0.0| 0.38842857142857146|30.714285714285715|0.457142857
|          98| 5.333333333333333|13.666666666666666|19.333333333333332| 0.8506666666666667|31.666666666666668| 0.66666666
|         110| 4.333333333333333|33.666666666666664|123.33333333333333| 0.5733333333333334|          39.0| 0.66666666
|         NULL|          1.0|          23.0|          94.0|          NULL|          NaN|
|          96|          4.0|           0.0|           0.0|          0.22975|          31.5|
|         100|3.6666666666666665|36.666666666666664|          104.0| 0.7133333333333333|          34.0| 0.33333333
|          70| 4.087719298245614| 24.05263157894737| 94.96491228070175| 0.438578947368421|31.982456140350877|0.403508771
|          61|          6.0|           0.0|           0.0|          0.151|          55.0|
|          75|          2.625|          13.875|           0.0|          0.382625|          33.25|
|          46|          1.5|          20.0|          209.0| 0.41500000000000004|          22.0|
|          78| 4.888888888888889|23.733333333333334| 66.75555555555556| 0.4885111111111111| 39.08888888888889|0.377777777
|          60|2.6486486486486487| 20.18918918918919|105.70270270270271| 0.4364594594594594|29.135135135135137| 0.18918918
|          90| 4.545454545454546|23.545454545454547|116.13636363636364| 0.4235454545454545| 38.13636363636363|
|          68| 3.488888888888889|23.133333333333333| 89.08888888888889| 0.47006666666666674|31.555555555555557|0.26666666
|         104|          2.5|          12.5|           0.0|          0.293|          46.5|
+-----+ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+
only showing top 20 rows
```

```
df.agg({"Insulin":'max'}).show()
```

```
↗ +-----+
|max(Insulin)|
+-----+
|          846|
+-----+
```

Start coding or [generate](#) with AI.

