```
import pyspark
```

```
from pyspark import SparkContext
from pyspark.sql import SQLContext
```

```
s = SparkContext.getOrCreate()
sc = SQLContext(s)
```

```
/usr/local/lib/python3.12/dist-packages/pyspark/sql/context.py:113: FutureWarning: Deprecated in 3.0.0. Use SparkSession.builder.ge
  warnings.warn(
```

```
df = sc.read.csv("MentalHealthSurvey.csv", header=True, inferSchema=True)
df.show()
```

```
+------+---+----------+-------------+----------------+-------------+-------+------------------+---------------------+--------------
|gender|age|university| degree_level|    degree_major|academic_year|   cgpa|residential_status|campus_discrimination|sports_engagem
+------+---+----------+-------------+----------------+-------------+-------+------------------+---------------------+--------------
|  Male| 20|        PU|Undergraduate|    Data Science|     2nd year|3.0-3.5|        Off-Campus|                   No|        No Spo
|  Male| 20|       UET| Postgraduate|Computer Science|     3rd year|3.0-3.5|        Off-Campus|                   No|        1-3 ti
|  Male| 20|      FAST|Undergraduate|Computer Science|     3rd year|2.5-3.0|        Off-Campus|                   No|        1-3 ti
|  Male| 20|       UET|Undergraduate|Computer Science|     3rd year|2.5-3.0|         On-Campus|                   No|        No Spo
|Female| 20|       UET|Undergraduate|Computer Science|     3rd year|3.0-3.5|        Off-Campus|                  Yes|        No Spo
|Female| 20|       UET|Undergraduate|Computer Science|     3rd year|3.0-3.5|        Off-Campus|                   No|        No Spo
|  Male| 26|        PU| Postgraduate|    Data Science|     1st year|2.5-3.0|         On-Campus|                  Yes|        1-3 ti
|  Male| 22|        PU|Undergraduate|    Data Science|     2nd year|3.0-3.5|        Off-Campus|                  Yes|        No Spo
|  Male| 20|   COMSATS|Undergraduate|Computer Science|     3rd year|2.5-3.0|        Off-Campus|                  Yes|        1-3 ti
|  Male| 23|   COMSATS|Undergraduate|Computer Science|     3rd year|2.5-3.0|        Off-Campus|                   No|        No Spo
|  Male| 20|   COMSATS|Undergraduate|Computer Science|     2nd year|3.0-3.5|         On-Campus|                   No|        No Spo
|  Male| 20|   COMSATS|Undergraduate|Computer Science|     3rd year|3.0-3.5|        Off-Campus|                   No|        1-3 ti
|  Male| 21|   COMSATS|Undergraduate|Computer Science|     3rd year|3.5-4.0|         On-Campus|                   No|        No Spo
|  Male| 19|        PU|Undergraduate|    Data Science|     1st year|3.0-3.5|        Off-Campus|                   No|        No Spo
|Female| 20|        PU|Undergraduate|    Data Science|     1st year|3.0-3.5|        Off-Campus|                   No|        No Spo
|Female| 19|        PU|Undergraduate|    Data Science|     1st year|3.0-3.5|        Off-Campus|                  Yes|        No Spo
|Female| 20|        PU|Undergraduate|    Data Science|     1st year|3.0-3.5|        Off-Campus|                   No|        No Spo
|  Male| 19|        PU|Undergraduate|    Data Science|     1st year|2.5-3.0|        Off-Campus|                   No|        No Spo
|Female| 19|        PU|Undergraduate|    Data Science|     1st year|3.5-4.0|        Off-Campus|                   No|        1-3 ti
|  Male| 19|        PU|Undergraduate|    Data Science|     1st year|3.5-4.0|        Off-Campus|                  Yes|        1-3 ti
+------+---+----------+-------------+----------------+-------------+-------+------------------+---------------------+--------------
only showing top 20 rows
```

```
df.filter(df["gender"] == "Male").count()
```

```
63
```

```
df.filter((df["gender"] == "Male") & (df.university == "COMSATS")).count()
```

```
8
```

```
df.filter((df["gender"] == "Male") & ((df["age"] >= 20) & (df["age"] <= 22)) & (df["university"] == "PU")).count()
```

```
19
```

```
from pyspark.sql.functions import avg

df.select(avg("age")).show()
```

```
+------------------+
|          avg(age)|
+------------------+
|19.942528735632184|
+------------------+
```

```
from pyspark.sql.functions import *

df.select(stddev("age")).show()
```

```
+------------------+
|       stddev(age)|
+------------------+
|1.6236358985749892|
+------------------+
from pyspark.sql.functions import *

df.select(min("age")).show()
```

```
+--------+
|min(age)|
+--------+
|      17|
+--------+
```

```
from pyspark.sql.functions import *

df.select(max("age")).show()
```

```
+--------+
|max(age)|
+--------+
|      26|
+--------+
```

```
data = sc.createDataFrame(df.tail(25))
data.show(truncate=False)
```

```
+------+---+----------+-------------+----------------------+-------------+-------+------------------+---------------------+--------
|gender|age|university|degree_level |degree_major          |academic_year|cgpa   |residential_status|campus_discrimination|sports_e
+------+---+----------+-------------+----------------------+-------------+-------+------------------+---------------------+--------
|Male  |20 |UMT       |Undergraduate|Computer Science      |3rd year     |3.5-4.0|Off-Campus        |No                   |7+ times
|Female|19 |PU        |Undergraduate|Computer Science      |2nd year     |3.0-3.5|Off-Campus        |No                   |No Sport
|Male  |20 |COMSATS   |Undergraduate|Computer Science      |3rd year     |2.5-3.0|Off-Campus        |Yes                  |1-3 time
|Male  |20 |UMT       |Undergraduate|Computer Science      |3rd year     |3.5-4.0|On-Campus         |No                   |No Sport
|Male  |20 |FAST      |Undergraduate|Computer Science      |3rd year     |2.5-3.0|Off-Campus        |No                   |1-3 time
|Male  |20 |FAST      |Undergraduate|Computer Science      |3rd year     |3.5-4.0|On-Campus         |No                   |No Sport
|Male  |20 |FAST      |Undergraduate|Computer Science      |3rd year     |2.5-3.0|On-Campus         |No                   |7+ times
|Male  |19 |UOL       |Undergraduate|Computer Science      |3rd year     |2.5-3.0|Off-Campus        |Yes                  |4-6 time
|Male  |22 |UET       |Undergraduate|Computer Science      |3rd year     |2.5-3.0|Off-Campus        |No                   |No Sport
|Male  |20 |UET       |Undergraduate|Computer Science      |3rd year     |3.0-3.5|Off-Campus        |No                   |7+ times
|Female|20 |UET       |Undergraduate|Computer Science      |3rd year     |3.5-4.0|Off-Campus        |No                   |No Sport
|Female|19 |UET       |Undergraduate|Computer Science      |3rd year     |3.0-3.5|On-Campus         |No                   |No Sport
|Male  |21 |PU        |Undergraduate|Data Science          |1st year     |3.5-4.0|Off-Campus        |No                   |7+ times
|Male  |26 |KUST      |Undergraduate|Data Science          |4th year     |3.5-4.0|Off-Campus        |Yes                  |1-3 time
|Female|19 |UET       |Undergraduate|Computer Science      |3rd year     |2.5-3.0|Off-Campus        |Yes                  |No Sport
|Male  |22 |PU        |Undergraduate|Information Technology|4th year     |3.0-3.5|Off-Campus        |No                   |4-6 time
|Male  |22 |PU        |Undergraduate|Information Technology|4th year     |3.0-3.5|Off-Campus        |Yes                  |No Sport
|Female|21 |PU        |Undergraduate|Information Technology|4th year     |3.5-4.0|Off-Campus        |Yes                  |No Sport
|Female|21 |PU        |Undergraduate|Information Technology|4th year     |3.0-3.5|Off-Campus        |No                   |No Sport
|Male  |22 |PU        |Undergraduate|Information Technology|4th year     |2.5-3.0|Off-Campus        |No                   |No Sport
+------+---+----------+-------------+----------------------+-------------+-------+------------------+---------------------+--------
only showing top 20 rows
```

```
df.orderBy(desc("study_satisfaction")).show()
```

```
+------+---+----------+-------------+----------------+-------------+-------+------------------+---------------------+--------------
|gender|age|university| degree_level|    degree_major|academic_year|   cgpa|residential_status|campus_discrimination|sports_engagem
+------+---+----------+-------------+----------------+-------------+-------+------------------+---------------------+--------------
|  Male| 20|        PU|Undergraduate|    Data Science|     2nd year|3.0-3.5|        Off-Campus|                   No|        No Spo
|  Male| 20|       UET| Postgraduate|Computer Science|     3rd year|3.0-3.5|        Off-Campus|                   No|        1-3 ti
|  Male| 20|      FAST|Undergraduate|Computer Science|     3rd year|2.5-3.0|        Off-Campus|                   No|        1-3 ti
|  Male| 20|   COMSATS|Undergraduate|Computer Science|     2nd year|3.0-3.5|         On-Campus|                   No|        No Spo
|  Male| 21|   COMSATS|Undergraduate|Computer Science|     3rd year|3.5-4.0|         On-Campus|                   No|        No Spo
|Female| 19|        PU|Undergraduate|    Data Science|     1st year|3.0-3.5|        Off-Campus|                  Yes|        No Spo
|Female| 20|        PU|Undergraduate|    Data Science|     1st year|3.0-3.5|        Off-Campus|                   No|        No Spo
|  Male| 19|        PU|Undergraduate|    Data Science|     1st year|2.5-3.0|        Off-Campus|                   No|        No Spo
|  Male| 19|        PU|Undergraduate|    Data Science|     1st year|3.5-4.0|        Off-Campus|                  Yes|        1-3 ti
|Female| 20|        PU|Undergraduate|    Data Science|     1st year|3.0-3.5|        Off-Campus|                   No|        1-3 ti
|  Male| 21|        PU|Undergraduate|    Data Science|     2nd year|3.5-4.0|         On-Campus|                   No|        1-3 ti
|  Male| 20|       UET|Undergraduate|Computer Science|     3rd year|3.5-4.0|        Off-Campus|                   No|        1-3 ti
|Female| 19|        PU|Undergraduate|    Data Science|     1st year|3.5-4.0|        Off-Campus|                   No|        1-3 ti
|  Male| 19|        PU|Undergraduate|    Data Science|     1st year|3.0-3.5|        Off-Campus|                   No|         7+ ti
|  Male| 20|        PU|Undergraduate|    Data Science|     1st year|3.5-4.0|        Off-Campus|                   No|        4-6 ti
|  Male| 18|        PU|Undergraduate|    Data Science|     1st year|3.5-4.0|        Off-Campus|                   No|        No Spo
|  Male| 18|        PU|Undergraduate|    Data Science|     1st year|1.5-2.0|         On-Campus|                   No|        No Spo
|Female| 19|        PU|Undergraduate|    Data Science|     1st year|3.0-3.5|         On-Campus|                   No|        No Spo
|Female| 19|        PU|Undergraduate|    Data Science|     1st year|3.0-3.5|        Off-Campus|                   No|        4-6 ti
|  Male| 19|        PU|Undergraduate|Computer Science|     1st year|2.0-2.5|         On-Campus|                   No|        1-3 ti
```

```
+------+---+----------+-------------+----------------+-------------+-------+------------------+---------------------+--------------
only showing top 20 rows
```

```
df.orderBy(asc("study_satisfaction")).show()
```

```
+------+---+----------+-------------+--------------------+-------------+-------+------------------+---------------------+----------
|gender|age|university| degree_level|        degree_major|academic_year|   cgpa|residential_status|campus_discrimination|sports_eng
+------+---+----------+-------------+--------------------+-------------+-------+------------------+---------------------+----------
|  Male| 20|        PU|Undergraduate|    Computer Science|     1st year|1.5-2.0|        Off-Campus|                  Yes|         7
|Female| 18|        PU|Undergraduate|        Data Science|     1st year|2.5-3.0|        Off-Campus|                   No|        No
|  Male| 22|        PU|Undergraduate|Information Techn...|     4th year|2.5-3.0|        Off-Campus|                   No|        No
|  Male| 19|        PU|Undergraduate|        Data Science|     1st year|2.0-2.5|         On-Campus|                   No|        No
|Female| 19|        PU|Undergraduate|        Data Science|     2nd year|2.5-3.0|         On-Campus|                   No|        No
|  Male| 18|        PU|Undergraduate|        Data Science|     1st year|3.0-3.5|        Off-Campus|                   No|        No
|  Male| 20|       UET|Undergraduate|    Computer Science|     3rd year|2.5-3.0|         On-Campus|                   No|        No
|Female| 20|       UET|Undergraduate|    Computer Science|     3rd year|3.0-3.5|        Off-Campus|                  Yes|        No
|  Male| 22|        PU|Undergraduate|        Data Science|     2nd year|3.0-3.5|        Off-Campus|                  Yes|        No
|  Male| 20|   COMSATS|Undergraduate|    Computer Science|     3rd year|2.5-3.0|        Off-Campus|                  Yes|        1-
|  Male| 23|   COMSATS|Undergraduate|    Computer Science|     3rd year|2.5-3.0|        Off-Campus|                   No|        No
|  Male| 19|        PU|Undergraduate|        Data Science|     1st year|3.0-3.5|        Off-Campus|                   No|        No
|Female| 20|        PU|Undergraduate|        Data Science|     1st year|3.0-3.5|        Off-Campus|                   No|        No
|  Male| 20|        PU|Undergraduate|        Data Science|     2nd year|2.5-3.0|        Off-Campus|                   No|        No
|  Male| 23|        PU|Undergraduate|        Data Science|     2nd year|2.5-3.0|        Off-Campus|                  Yes|        1-
|  Male| 18|        PU|Undergraduate|    Computer Science|     1st year|3.5-4.0|        Off-Campus|                   No|         7
|  Male| 21|        PU|Undergraduate|Software Engineering|     4th year|3.5-4.0|        Off-Campus|                  Yes|        1-
|  Male| 17|        PU|Undergraduate|Information Techn...|     1st year|0.0-0.0|        Off-Campus|                   No|         7
|  Male| 21|        PU|Undergraduate|        Data Science|     2nd year|2.5-3.0|         On-Campus|                  Yes|        4-
|  Male| 21|      FAST|Undergraduate|        Data Science|     3rd year|2.0-2.5|        Off-Campus|                   No|        1-
+------+---+----------+-------------+--------------------+-------------+-------+------------------+---------------------+----------
only showing top 20 rows
```

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

```
+------+-----+
|gender|count|
+------+-----+
|Female|   24|
|  Male|   63|
+------+-----+
```

```
df.groupBy("gender", "university").count().show()
```

```
+------+----------+-----+
|gender|university|count|
+------+----------+-----+
|  Male|        VU|    1|
|  Male|      KUST|    1|
|Female|   COMSATS|    1|
|Female|       UET|    5|
|  Male|       UET|    5|
|  Male|      FAST|    6|
|  Male|       UMT|    2|
|  Male|   COMSATS|    8|
|Female|        PU|   18|
|  Male|        PU|   38|
|  Male|      NUST|    1|
|  Male|       UOL|    1|
+------+----------+-----+
```

```
df.filter(df["gender"] == "Male").select(max("age")).show()
```

```
+--------+
|max(age)|
+--------+
|      26|
+--------+
```

```
df.filter(df["gender"] == "Male").select(sum("age")).show()
```

```
+--------+
|sum(age)|
+--------+
|    1271|
+--------+
df.filter(df["gender"] == "Female").select(sum("age")).show()
```

```
+--------+
|sum(age)|
+--------+
|     464|
+--------+
```

```
df.select(sum("age")).show()
```

```
+--------+
|sum(age)|
+--------+
|    1735|
+--------+
```

```
mean_v = df.select(mean("age")).collect()[0][0]
print(mean_v)
df1 = df.replace(26, mean_v, subset=["age"])
df1.show()
```

```
19.942528735632184
+------+---+----------+-------------+----------------+-------------+-------+------------------+---------------------+--------------
|gender|age|university| degree_level|    degree_major|academic_year|   cgpa|residential_status|campus_discrimination|sports_engagem
+------+---+----------+-------------+----------------+-------------+-------+------------------+---------------------+--------------
|  Male| 20|        PU|Undergraduate|    Data Science|     2nd year|3.0-3.5|        Off-Campus|                   No|        No Spo
|  Male| 20|       UET| Postgraduate|Computer Science|     3rd year|3.0-3.5|        Off-Campus|                   No|        1-3 ti
|  Male| 20|      FAST|Undergraduate|Computer Science|     3rd year|2.5-3.0|        Off-Campus|                   No|        1-3 ti
|  Male| 20|       UET|Undergraduate|Computer Science|     3rd year|2.5-3.0|         On-Campus|                   No|        No Spo
|Female| 20|       UET|Undergraduate|Computer Science|     3rd year|3.0-3.5|        Off-Campus|                  Yes|        No Spo
|Female| 20|       UET|Undergraduate|Computer Science|     3rd year|3.0-3.5|        Off-Campus|                   No|        No Spo
|  Male| 19|        PU| Postgraduate|    Data Science|     1st year|2.5-3.0|         On-Campus|                  Yes|        1-3 ti
|  Male| 22|        PU|Undergraduate|    Data Science|     2nd year|3.0-3.5|        Off-Campus|                  Yes|        No Spo
|  Male| 20|   COMSATS|Undergraduate|Computer Science|     3rd year|2.5-3.0|        Off-Campus|                  Yes|        1-3 ti
|  Male| 23|   COMSATS|Undergraduate|Computer Science|     3rd year|2.5-3.0|        Off-Campus|                   No|        No Spo
|  Male| 20|   COMSATS|Undergraduate|Computer Science|     2nd year|3.0-3.5|         On-Campus|                   No|        No Spo
|  Male| 20|   COMSATS|Undergraduate|Computer Science|     3rd year|3.0-3.5|        Off-Campus|                   No|        1-3 ti
|  Male| 21|   COMSATS|Undergraduate|Computer Science|     3rd year|3.5-4.0|         On-Campus|                   No|        No Spo
|  Male| 19|        PU|Undergraduate|    Data Science|     1st year|3.0-3.5|        Off-Campus|                   No|        No Spo
|Female| 20|        PU|Undergraduate|    Data Science|     1st year|3.0-3.5|        Off-Campus|                   No|        No Spo
|Female| 19|        PU|Undergraduate|    Data Science|     1st year|3.0-3.5|        Off-Campus|                  Yes|        No Spo
|Female| 20|        PU|Undergraduate|    Data Science|     1st year|3.0-3.5|        Off-Campus|                   No|        No Spo
|  Male| 19|        PU|Undergraduate|    Data Science|     1st year|2.5-3.0|        Off-Campus|                   No|        No Spo
|Female| 19|        PU|Undergraduate|    Data Science|     1st year|3.5-4.0|        Off-Campus|                   No|        1-3 ti
|  Male| 19|        PU|Undergraduate|    Data Science|     1st year|3.5-4.0|        Off-Campus|                  Yes|        1-3 ti
+------+---+----------+-------------+----------------+-------------+-------+------------------+---------------------+--------------
only showing top 20 rows
```

```
mean_value = df.agg(mean("age").alias("Mean_Age")).collect()[0]["Mean_Age"]

print("Mean age:", mean_value)
```

```
Mean age: 19.942528735632184
```