jupyter Dataease_Assign Last Checkpoint: an hour ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run Code

In [2]:
```
import findspark

findspark.init()
```

In [3]:
```
import pyspark

from pyspark.sql import SparkSession

spark = SparkSession.builder.getOrCreate()

df = spark.sql("select 'spark' as hello ")

df.show()
```

```
+-----+
|hello|
+-----+
|spark|
+-----+
```

jupyter Dataease_Assign Last Checkpoint: an hour ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run Code

```
+-----+
```

In [4]:
```
#Reading Data FIles :

df=spark.read.csv("Downloads\startup.csv",inferSchema=True,header=True)
df1=spark.read.parquet("Downloads\consumerInternet.parquet")


df3=df.unionAll(df1)

df3.printSchema()
```

```
root
 |-- Sr_No: string (nullable = true)
 |-- Date: string (nullable = true)
 |-- Startup_Name: string (nullable = true)
 |-- Industry_Vertical: string (nullable = true)
 |-- SubVertical: string (nullable = true)
 |-- City: string (nullable = true)
 |-- Investors_Name: string (nullable = true)
 |-- InvestmentnType: string (nullable = true)
 |-- Amount_in_USD: string (nullable = true)
 |-- Remarks: string (nullable = true)
```

jupyter Dataease_Assign Last Checkpoint: an hour ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run Code

In [6]:
```
# How many startups are there in Pune City?
df3.createOrReplaceTempView("df3")

PuneStartups = spark.sql("SELECT COUNT(Startup_Name) FROM df3 WHERE City = 'Pune'")
PuneStartups.show()
```

```
+------------------+
|count(Startup_Name)|
+------------------+
|               105|
+------------------+
```

In [7]:
```
#How many startups in Pune got their Seed/ Angel Funding?

Startups = spark.sql("SELECT COUNT(Startup_Name) FROM df3 WHERE InvestmentnType like 'Seed%/%' AND City = 'Pune'")
Startups.show()
```

```
+------------------+
|count(Startup_Name)|
+------------------+
|                 6|
+------------------+
```

jupyter Dataease_Assign Last Checkpoint: an hour ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run Code

In [32]:
```
#What is the total amount raised by startups in PuneCity? Hint - use regex_replace to get rid of null

PuneCityAmount = spark.sql("SELECT SUM(regexp_replace(Amount_in_USD, 'N/A', '00')) AS Amount FROM df3 WHERE City = 'Pune'")
PuneCityAmount.show()
```

```
+------+
|Amount|
+------+
|   0.0|
+------+
```

In [8]:
```
#What are the top 5 Industry_Vertical which has the highest number of startups in India?

TopFive = spark.sql("SELECT Industry_Vertical, COUNT(Startup_Name) FROM df3 GROUP BY Industry_Vertical ORDER BY COUNT(Startup
TopFive.show()
```

```
+-----------------+------------------+
|Industry_Vertical|count(Startup_Name)|
+-----------------+------------------+
|Consumer Internet|               941|
|       Technology|               478|
|        eCommerce|               186|
|              nan|               171|
|       Healthcare|                70|
+-----------------+------------------+
```

jupyter Dataease_Assign Last Checkpoint: an hour ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run Code

In [8]:
```
#What are the top 5 Industry_Vertical which has the highest number of startups in India?

TopFive = spark.sql("SELECT Industry_Vertical, COUNT(Startup_Name) FROM df3 GROUP BY Industry_Vertical ORDER BY COUNT(Startup
TopFive.show()
```

```
+-----------------+--------------------+
|Industry_Vertical|count(Startup_Name)|
+-----------------+--------------------+
|Consumer Internet|                 941|
|       Technology|                 478|
|        eCommerce|                 186|
|              nan|                 171|
|       Healthcare|                  70|
+-----------------+--------------------+
```

In [88]:
```
TopInvestor=spark.sql("SELECT Investors_Name  , COUNT(Amount_in_USD) FROM df3 GROUP BY Investors_Name ORDER BY COUNT(Amount_i
TopInvestor.show()
```

```
+--------------------+--------------------+
|      Investors_Name|count(Amount_in_USD)|
+--------------------+--------------------+
|Undisclosed Inves...|                  39|
|Undisclosed inves...|                  30|
|          Ratan Tata|                  25|
+--------------------+--------------------+
```