In [1]:
```
import findspark
findspark.init()
```

In [2]:
```
import pyspark
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()
```

In [4]:
```
print(spark)
rdd = spark.sparkContext.parallelize([1,2,3,4,5,6])
print("RDD count: " + str(rdd.count()))

rdd = spark.sparkContext.emptyRDD
print(rdd)
rdd2 = spark.sparkContext.parallelize([])
print(rdd2)
```

```
<pyspark.sql.session.SparkSession object at 0x0000014E9CAD6D40>
RDD count: 6
<bound method SparkContext.emptyRDD of <SparkContext master=local[*] appName=pyspark-shell>>
ParallelCollectionRDD[4] at readRDDFromFile at PythonRDD.scala:274
```

In [5]:
```
spark = SparkSession.builder.appName('SparkByExamples.com').getOrCreate()
rdd = spark.sparkContext.parallelize([1,2,3,4,5,6])
rddcollect = rdd.collect()
print("Number of Partitions: " + str(rdd.getNumPartitions()))
print("Action: First element: " + str(rdd.first()))
print(rddcollect)
```

```
Number of Partitions: 8
Action: First element: 1
[1, 2, 3, 4, 5, 6]
```

In [7]:
```
emptyRDD = spark.sparkContext.emptyRDD()
emptyRDD2 = rdd = spark.sparkContext.parallelize([])
print("" + str(emptyRDD2.isEmpty()))
```

```
True
```

In [8]:
```
from pyspark.sql.types import StructType, StructField, StringType, IntegerType
from pyspark.sql.types import ArrayType, DoubleType, BooleanType
from pyspark.sql.functions import col, array_contains
```

In [9]:
```
spark = SparkSession.builder.appName('SparkExamples').getOrCreate()
df = spark.read.csv("D:\github repo\Spark\zipcodes.csv")
df.printSchema()
```

```
root
 |-- _c0: string (nullable = true)
 |-- _c1: string (nullable = true)
 |-- _c2: string (nullable = true)
 |-- _c3: string (nullable = true)
 |-- _c4: string (nullable = true)
 |-- _c5: string (nullable = true)
 |-- _c6: string (nullable = true)
 |-- _c7: string (nullable = true)
 |-- _c8: string (nullable = true)
 |-- _c9: string (nullable = true)
 |-- _c10: string (nullable = true)
 |-- _c11: string (nullable = true)
 |-- _c12: string (nullable = true)
 |-- _c13: string (nullable = true)
 |-- _c14: string (nullable = true)
 |-- _c15: string (nullable = true)
 |-- _c16: string (nullable = true)
 |-- _c17: string (nullable = true)
 |-- _c18: string (nullable = true)
 |-- _c19: string (nullable = true)
```

In [10]:
```
df2 = spark.read.option("header", "True").csv("D:\github repo\Spark\zipcodes.csv")
df2.printSchema()
```

```
root
 |-- RecordNumber: string (nullable = true)
 |-- Zipcode: string (nullable = true)
 |-- ZipCodeType: string (nullable = true)
 |-- City: string (nullable = true)
 |-- State: string (nullable = true)
 |-- LocationType: string (nullable = true)
 |-- Lat: string (nullable = true)
 |-- Long: string (nullable = true)
 |-- Xaxis: string (nullable = true)
 |-- Yaxis: string (nullable = true)
 |-- Zaxis: string (nullable = true)
 |-- WorldRegion: string (nullable = true)
 |-- Country: string (nullable = true)
 |-- LocationText: string (nullable = true)
 |-- Location: string (nullable = true)
 |-- Decommisioned: string (nullable = true)
 |-- TaxReturnsFiled: string (nullable = true)
 |-- EstimatedPopulation: string (nullable = true)
 |-- TotalWages: string (nullable = true)
 |-- Notes: string (nullable = true)
```

In [11]:
```
df3 = spark.read.options(header='True', delimiter=',').csv("D:\github repo\Spark\zipcodes.csv")
df3.printSchema()
```

```
root
 |-- RecordNumber: string (nullable = true)
 |-- Zipcode: string (nullable = true)
 |-- ZipCodeType: string (nullable = true)
 |-- City: string (nullable = true)
 |-- State: string (nullable = true)
 |-- LocationType: string (nullable = true)
 |-- Lat: string (nullable = true)
 |-- Long: string (nullable = true)
 |-- Xaxis: string (nullable = true)
 |-- Yaxis: string (nullable = true)
 |-- Zaxis: string (nullable = true)
 |-- WorldRegion: string (nullable = true)
 |-- Country: string (nullable = true)
 |-- LocationText: string (nullable = true)
 |-- Location: string (nullable = true)
 |-- Decommisioned: string (nullable = true)
 |-- TaxReturnsFiled: string (nullable = true)
 |-- EstimatedPopulation: string (nullable = true)
 |-- TotalWages: string (nullable = true)
 |-- Notes: string (nullable = true)
```

In [12]:
```
dfSchema = spark.read.format("csv").option("header", "True").load("D:\github repo\Spark\zipcodes.csv")
dfSchema.printSchema()
```

```
root
 |-- RecordNumber: string (nullable = true)
 |-- Zipcode: string (nullable = true)
 |-- ZipCodeType: string (nullable = true)
 |-- City: string (nullable = true)
 |-- State: string (nullable = true)
 |-- LocationType: string (nullable = true)
 |-- Lat: string (nullable = true)
 |-- Long: string (nullable = true)
 |-- Xaxis: string (nullable = true)
 |-- Yaxis: string (nullable = true)
 |-- Zaxis: string (nullable = true)
 |-- WorldRegion: string (nullable = true)
 |-- Country: string (nullable = true)
 |-- LocationText: string (nullable = true)
 |-- Location: string (nullable = true)
 |-- Decommisioned: string (nullable = true)
 |-- TaxReturnsFiled: string (nullable = true)
 |-- EstimatedPopulation: string (nullable = true)
 |-- TotalWages: string (nullable = true)
 |-- Notes: string (nullable = true)
```

In [15]:
```
df2.write.option("header", "True").csv("D:\github repo\Spark\zipcode")
```

```
---------------------------------------------------------------------------
Py4JJavaError                             Traceback (most recent call last)
C:\Users\PRATIK~1\AppData\Local\Temp/ipykernel_28220/430991765.py in <module>
----> 1 df2.write.option("header", "True").csv("D:\github repo\Spark\zipcode")

~\AppData\Local\Programs\Python\Python310\lib\site-packages\pyspark\sql\readwriter.py in csv(self, path, mode, compression, s
ep, quote, escape, header, nullValue, escapeQuotes, quoteAll, dateFormat, timestampFormat, ignoreLeadingWhiteSpace, ignoreTra
ilingWhiteSpace, charToEscapeQuoteEscaping, encoding, emptyValue, lineSep)
    953                       charToEscapeQuoteEscaping=charToEscapeQuoteEscaping,
    954                       encoding=encoding, emptyValue=emptyValue, lineSep=lineSep)
--> 955         self._jwrite.csv(path)
    956
    957     def orc(self, path, mode=None, partitionBy=None, compression=None):

~\AppData\Local\Programs\Python\Python310\lib\site-packages\py4j\java_gateway.py in __call__(self, *args)
   1319
   1320         answer = self.gateway_client.send_command(command)
-> 1321         return_value = get_return_value(
   1322             answer, self.gateway_client, self.target_id, self.name)
```

In [16]:
```
spark = SparkSession.builder.appName('SparkEx').getOrCreate()
data = [("James","","Smith","36636","M",60000),
        ("Michael","Rose","","40288","M",70000),
        ("Robert","","Williams","42114","",400000),
        ("Maria","Anne","Jones","39192","F",500000),
        ("Jen","Mary","Brown","","F",0)]
cols = ["firstName", "middleName", "lastName", "dob", "gender", "salary"]
df = spark.createDataFrame(data = data, schema=cols)
df.printSchema()
```

```
root
 |-- firstName: string (nullable = true)
 |-- middleName: string (nullable = true)
 |-- lastName: string (nullable = true)
 |-- dob: string (nullable = true)
 |-- gender: string (nullable = true)
 |-- salary: long (nullable = true)
```

In [17]:
```
df.show(truncate=False)
```

```
+---------+----------+--------+-----+------+------+
|firstName|middleName|lastName|dob  |gender|salary|
+---------+----------+--------+-----+------+------+
|James    |          |Smith   |36636|M     |60000 |
|Michael  |Rose      |        |40288|M     |70000 |
|Robert   |          |Williams|42114|      |400000|
|Maria    |Anne      |Jones   |39192|F     |500000|
|Jen      |Mary      |Brown   |     |F     |0     |
+---------+----------+--------+-----+------+------+
```

In [19]:
```
pandasDF = df.toPandas()
print(pandasDF)
```

```
  firstName middleName  lastName    dob gender  salary
0     James               Smith  36636      M   60000
1   Michael       Rose            40288      M   70000
```

```
2    Robert             Williams  42114         400000
3     Maria       Anne     Jones  39192      F  500000
4       Jen       Mary     Brown             F       0
```

In [20]:
```
data = ["Project Gutenberg's",
        "Alice's Adventures in Wonderland",
        "Project Gutenberg's",
        "Adventures in Wonderland",
        "Project Gutenberg's"]
rdd = spark.sparkContext.parallelize(data)
for i in rdd.collect():
    print(i)
```

```
Project Gutenberg's
Alice's Adventures in Wonderland
Project Gutenberg's
Adventures in Wonderland
Project Gutenberg's
```

In [21]:
```
#FlatMap
rdd2 = rdd.flatMap(lambda x:x.split(" "))
for i in rdd2.collect():
    print(i)
```

```
Project
Gutenberg's
Alice's
Adventures
in
Wonderland
Project
Gutenberg's
Adventures
in
Wonderland
Project
Gutenberg's
```

In [22]:
```
dept = [("Finance",10),
        ("Marketing",20),
        ("Sales",30),
        ("IT",40)
      ]
rdd = spark.sparkContext.parallelize(dept)
df = rdd.toDF()
df.printSchema()
```

```
root
 |-- _1: string (nullable = true)
 |-- _2: long (nullable = true)
```

In [23]:
```
df.show()
```

```
+---------+---+
|       _1| _2|
+---------+---+
|  Finance| 10|
|Marketing| 20|
|    Sales| 30|
|       IT| 40|
+---------+---+
```

In [24]:
```
deptcols = ["deptName", "deptid"]
df2 = rdd.toDF(deptcols)
df2.printSchema()
```

```
root
 |-- deptName: string (nullable = true)
 |-- deptid: long (nullable = true)
```

In [25]:
```
df2.show()
```

```
+---------+------+
| deptName|deptid|
+---------+------+
|  Finance|    10|
|Marketing|    20|
|    Sales|    30|
|       IT|    40|
+---------+------+
```

In [26]:
```
deptDF = spark.createDataFrame(data=dept, schema=deptcols)
deptDF.printSchema()
```

```
root
 |-- deptName: string (nullable = true)
 |-- deptid: long (nullable = true)
```

In [27]:
```
deptDF.show()
```

```
+---------+------+
| deptName|deptid|
+---------+------+
|  Finance|    10|
|Marketing|    20|
|    Sales|    30|
|       IT|    40|
+---------+------+
```

In [28]:
```
from pyspark.sql.types import StructType, StructField, StringType
deptSchema = StructType([StructField('deptName', StringType(), True),
```

```
                     StructField('deptid', StringType(), True)])
deptDF1 = spark.createDataFrame(data=dept, schema=deptSchema)
deptDF1.printSchema()
```

```
root
 |-- deptName: string (nullable = true)
 |-- deptid: string (nullable = true)
```

In [29]:
```
deptDF1.show()
```

```
+---------+------+
| deptName|deptid|
+---------+------+
|  Finance|    10|
|Marketing|    20|
|    Sales|    30|
|       IT|    40|
+---------+------+
```

In [31]:
```
spark: SparkSession = SparkSession.builder.master("local[1]").appName("SparkEx").getOrCreate()
filepath = "D:\github repo\Spark\zipcodes.csv"
df = spark.read.options(header='True', inferSchema='True').csv(filepath)
df.printSchema()
```

```
root
 |-- RecordNumber: integer (nullable = true)
 |-- Zipcode: integer (nullable = true)
 |-- ZipCodeType: string (nullable = true)
 |-- City: string (nullable = true)
 |-- State: string (nullable = true)
 |-- LocationType: string (nullable = true)
 |-- Lat: double (nullable = true)
 |-- Long: double (nullable = true)
 |-- Xaxis: double (nullable = true)
 |-- Yaxis: double (nullable = true)
 |-- Zaxis: double (nullable = true)
 |-- WorldRegion: string (nullable = true)
 |-- Country: string (nullable = true)
 |-- LocationText: string (nullable = true)
 |-- Location: string (nullable = true)
 |-- Decommisioned: boolean (nullable = true)
 |-- TaxReturnsFiled: integer (nullable = true)
 |-- EstimatedPopulation: integer (nullable = true)
 |-- TotalWages: integer (nullable = true)
 |-- Notes: string (nullable = true)
```

In [32]:
```
df.show()
```

```
+------------+-------+-----------+-------------------+-----+--------------+-----+-------+-----+-----+-----+-----------+-------+
--------------------+--------------------+-------------+--------------+-------------------+----------+-------------+
|RecordNumber|Zipcode|ZipCodeType|               City|State|  LocationType|  Lat|   Long|Xaxis|Yaxis|Zaxis|WorldRegion|Country|
LocationText|            Location|Decommisioned|TaxReturnsFiled|EstimatedPopulation|TotalWages|        Notes|
+------------+-------+-----------+-------------------+-----+--------------+-----+-------+-----+-----+-----+-----------+-------+
--------------------+--------------------+-------------+--------------+-------------------+----------+-------------+
|           1|    704|   STANDARD|        PARC PARQUE|   PR|NOT ACCEPTABLE|17.96| -66.22| 0.38|-0.87|  0.3|         NA|     US|
Parc Parque, PR|NA-US-PR-PARC PARQUE|        false|           null|               null|      null|         null|
|           2|    704|   STANDARD|PASEO COSTA DEL SUR|   PR|NOT ACCEPTABLE|17.96| -66.22| 0.38|-0.87|  0.3|         NA|     US|
Paseo Costa Del S...|NA-US-PR-PASEO CO...|        false|           null|               null|      null|         null|
|          10|    709|   STANDARD|       BDA SAN LUIS|   PR|NOT ACCEPTABLE|18.14| -66.26| 0.38|-0.86| 0.31|         NA|     US|
Bda San Luis, PR|NA-US-PR-BDA SAN ...|        false|           null|               null|      null|         null|
|       61391|  76166|     UNIQUE|  CINGULAR WIRELESS|   TX|NOT ACCEPTABLE|32.72| -97.31| -0.1|-0.83| 0.54|         NA|     US|
Cingular Wireless...|NA-US-TX-CINGULAR...|        false|           null|               null|      null|         null|
|       61392|  76177|   STANDARD|         FORT WORTH|   TX|       PRIMARY|32.75| -97.33| -0.1|-0.83| 0.54|         NA|     US|
Fort Worth, TX| NA-US-TX-FORT WORTH|        false|           2126|               4053| 122396986|         null|
|       61393|  76177|   STANDARD|          FT WORTH|   TX|    ACCEPTABLE|32.75| -97.33| -0.1|-0.83| 0.54|         NA|     US|
Ft Worth, TX|   NA-US-TX-FT WORTH|        false|           2126|               4053| 122396986|         null|
|           4|    704|   STANDARD|    URB EUGENE RICE|   PR|NOT ACCEPTABLE|17.96| -66.22| 0.38|-0.87|  0.3|         NA|     US|
Urb Eugene Rice, PR|NA-US-PR-URB EUGE...|        false|           null|               null|      null|         null|
|       39827|  85209|   STANDARD|               MESA|   AZ|       PRIMARY|33.37|-111.64| -0.3|-0.77| 0.55|         NA|     US|
Mesa, AZ|       NA-US-AZ-MESA|        false|          14962|              26883| 563792730|no NWS data, |
|       39828|  85210|   STANDARD|               MESA|   AZ|       PRIMARY|33.38|-111.84|-0.31|-0.77| 0.55|         NA|     US|
Mesa, AZ|       NA-US-AZ-MESA|        false|          14374|              25446| 471000465|         null|
|       49345|  32046|   STANDARD|          HILLIARD|   FL|       PRIMARY|30.69| -81.92| 0.12|-0.85| 0.51|         NA|     US|
Hilliard, FL|   NA-US-FL-HILLIARD|        false|           3922|               7443| 133112149|         null|
|       49346|  34445|     PO BOX|             HOLDER|   FL|       PRIMARY|28.96| -82.41| 0.11|-0.86| 0.48|         NA|     US|
Holder, FL|     NA-US-FL-HOLDER|        false|           null|               null|      null|         null|
|       49347|  32564|   STANDARD|               HOLT|   FL|       PRIMARY|30.72| -86.67| 0.04|-0.85| 0.51|         NA|     US|
Holt, FL|       NA-US-FL-HOLT|        false|          1207|               2190|  36395913|         null|
|       49348|  34487|     PO BOX|          HOMOSASSA|   FL|       PRIMARY|28.78| -82.61| 0.11|-0.86| 0.48|         NA|     US|
Homosassa, FL|  NA-US-FL-HOMOSASSA|        false|           null|               null|      null|         null|
|          10|    708|   STANDARD|       BDA SAN LUIS|   PR|NOT ACCEPTABLE|18.14| -66.26| 0.38|-0.86| 0.31|         NA|     US|
Bda San Luis, PR|NA-US-PR-BDA SAN ...|        false|           null|               null|      null|         null|
|           3|    704|   STANDARD|      SECT LANAUSSE|   PR|NOT ACCEPTABLE|17.96| -66.22| 0.38|-0.87|  0.3|         NA|     US|
Sect Lanausse, PR|NA-US-PR-SECT LAN...|        false|           null|               null|      null|         null|
|       54354|  36275|     PO BOX|      SPRING GARDEN|   AL|       PRIMARY|33.97| -85.55| 0.06|-0.82| 0.55|         NA|     US|
Spring Garden, AL|NA-US-AL-SPRING G...|        false|           null|               null|      null|         null|
|       54355|  35146|   STANDARD|        SPRINGVILLE|   AL|       PRIMARY|33.77| -86.47| 0.05|-0.82| 0.55|         NA|     US|
Springville, AL|NA-US-AL-SPRINGVILLE|        false|           4046|               7845| 172127599|         null|
|       54356|  35585|   STANDARD|        SPRUCE PINE|   AL|       PRIMARY|34.37| -87.69| 0.03|-0.82| 0.56|         NA|     US|
Spruce Pine, AL|NA-US-AL-SPRUCE PINE|        false|            610|               1209|  18525517|         null|
|       76511|  27007|   STANDARD|           ASH HILL|   NC|NOT ACCEPTABLE| 36.4| -80.56| 0.13|-0.79| 0.59|         NA|     US|
Ash Hill, NC|   NA-US-NC-ASH HILL|        false|            842|               1666|  28876493|         null|
|       76512|  27203|   STANDARD|           ASHEBORO|   NC|       PRIMARY|35.71| -79.81| 0.14|-0.79| 0.58|         NA|     US|
Asheboro, NC|   NA-US-NC-ASHEBORO|        false|           8355|              15228| 215474318|         null|
+------------+-------+-----------+-------------------+-----+--------------+-----+-------+-----+-----+-----+-----------+-------+
--------------------+--------------------+-------------+--------------+-------------------+----------+-------------+
only showing top 20 rows
```

In [33]:
```
df.show(truncate=False)
```

```
+------------+-------+-----------+--------------------+-----+--------------+-----+-------+-----+-----+-----+-----------+-------+
---------------------+---------------------------+-------------+--------------+-------------------+----------+-------------
+
|RecordNumber|Zipcode|ZipCodeType|City                |State|LocationType  |Lat  |Long   |Xaxis|Yaxis|Zaxis|WorldRegion|Country|
LocationText          |Location                   |Decommisioned|TaxReturnsFiled|EstimatedPopulation|TotalWages|Notes
+
```

```
|
+------------+-------+-----------+-------------------+-----+--------------+-----+-------+-----+-----+-----+-----------+-------+
---------------------+-----------------------------+------------+--------------+------------------+----------+-------------
+
|1           |704    |STANDARD   |PARC PARQUE        |PR   |NOT ACCEPTABLE|17.96|-66.22 |0.38 |-0.87|0.3  |NA         |US     |
Parc Parque, PR        |NA-US-PR-PARC PARQUE        |false       |null          |null              |null      |null
|
|2           |704    |STANDARD   |PASEO COSTA DEL SUR|PR   |NOT ACCEPTABLE|17.96|-66.22 |0.38 |-0.87|0.3  |NA         |US     |
Paseo Costa Del Sur, PR|NA-US-PR-PASEO COSTA DEL SUR|false       |null          |null              |null      |null
|
|10          |709    |STANDARD   |BDA SAN LUIS       |PR   |NOT ACCEPTABLE|18.14|-66.26 |0.38 |-0.86|0.31 |NA         |US     |
Bda San Luis, PR       |NA-US-PR-BDA SAN LUIS       |false       |null          |null              |null      |null
|
|61391       |76166  |UNIQUE     |CINGULAR WIRELESS  |TX   |NOT ACCEPTABLE|32.72|-97.31 |-0.1 |-0.83|0.54 |NA         |US     |
Cingular Wireless, TX  |NA-US-TX-CINGULAR WIRELESS  |false       |null          |null              |null      |null
|
|61392       |76177  |STANDARD   |FORT WORTH         |TX   |PRIMARY       |32.75|-97.33 |-0.1 |-0.83|0.54 |NA         |US     |
Fort Worth, TX         |NA-US-TX-FORT WORTH         |false       |2126          |4053              |122396986 |null
|
|61393       |76177  |STANDARD   |FT WORTH           |TX   |ACCEPTABLE    |32.75|-97.33 |-0.1 |-0.83|0.54 |NA         |US     |
Ft Worth, TX           |NA-US-TX-FT WORTH           |false       |2126          |4053              |122396986 |null
|
|4           |704    |STANDARD   |URB EUGENE RICE    |PR   |NOT ACCEPTABLE|17.96|-66.22 |0.38 |-0.87|0.3  |NA         |US     |
Urb Eugene Rice, PR    |NA-US-PR-URB EUGENE RICE    |false       |null          |null              |null      |null
|
|39827       |85209  |STANDARD   |MESA               |AZ   |PRIMARY       |33.37|-111.64|-0.3 |-0.77|0.55 |NA         |US     |
Mesa, AZ               |NA-US-AZ-MESA               |false       |14962         |26883             |563792730 |no NWS data,
|
|39828       |85210  |STANDARD   |MESA               |AZ   |PRIMARY       |33.38|-111.84|-0.31|-0.77|0.55 |NA         |US     |
Mesa, AZ               |NA-US-AZ-MESA               |false       |14374         |25446             |471000465 |null
|
|49345       |32046  |STANDARD   |HILLIARD           |FL   |PRIMARY       |30.69|-81.92 |0.12 |-0.85|0.51 |NA         |US     |
Hilliard, FL           |NA-US-FL-HILLIARD           |false       |3922          |7443              |133112149 |null
|
|49346       |34445  |PO BOX     |HOLDER             |FL   |PRIMARY       |28.96|-82.41 |0.11 |-0.86|0.48 |NA         |US     |
Holder, FL             |NA-US-FL-HOLDER             |false       |null          |null              |null      |null
|
|49347       |32564  |STANDARD   |HOLT               |FL   |PRIMARY       |30.72|-86.67 |0.04 |-0.85|0.51 |NA         |US     |
Holt, FL               |NA-US-FL-HOLT               |false       |1207          |2190              |36395913  |null
|
|49348       |34487  |PO BOX     |HOMOSASSA          |FL   |PRIMARY       |28.78|-82.61 |0.11 |-0.86|0.48 |NA         |US     |
Homosassa, FL          |NA-US-FL-HOMOSASSA          |false       |null          |null              |null      |null
|
|10          |708    |STANDARD   |BDA SAN LUIS       |PR   |NOT ACCEPTABLE|18.14|-66.26 |0.38 |-0.86|0.31 |NA         |US     |
Bda San Luis, PR       |NA-US-PR-BDA SAN LUIS       |false       |null          |null              |null      |null
|
|3           |704    |STANDARD   |SECT LANAUSSE      |PR   |NOT ACCEPTABLE|17.96|-66.22 |0.38 |-0.87|0.3  |NA         |US     |
Sect Lanausse, PR      |NA-US-PR-SECT LANAUSSE      |false       |null          |null              |null      |null
|
|54354       |36275  |PO BOX     |SPRING GARDEN      |AL   |PRIMARY       |33.97|-85.55 |0.06 |-0.82|0.55 |NA         |US     |
Spring Garden, AL      |NA-US-AL-SPRING GARDEN      |false       |null          |null              |null      |null
|
|54355       |35146  |STANDARD   |SPRINGVILLE        |AL   |PRIMARY       |33.77|-86.47 |0.05 |-0.82|0.55 |NA         |US     |
Springville, AL        |NA-US-AL-SPRINGVILLE        |false       |4046          |7845              |172127599 |null
|
|54356       |35585  |STANDARD   |SPRUCE PINE        |AL   |PRIMARY       |34.37|-87.69 |0.03 |-0.82|0.56 |NA         |US     |
Spruce Pine, AL        |NA-US-AL-SPRUCE PINE        |false       |610           |1209              |18525517  |null
|
|76511       |27007  |STANDARD   |ASH HILL           |NC   |NOT ACCEPTABLE|36.4 |-80.56 |0.13 |-0.79|0.59 |NA         |US     |
Ash Hill, NC           |NA-US-NC-ASH HILL           |false       |842           |1666              |28876493  |null
|
|76512       |27203  |STANDARD   |ASHEBORO           |NC   |PRIMARY       |35.71|-79.81 |0.14 |-0.79|0.58 |NA         |US     |
Asheboro, NC           |NA-US-NC-ASHEBORO           |false       |8355          |15228             |215474318 |null
|
+------------+-------+-----------+-------------------+-----+--------------+-----+-------+-----+-----+-----+-----------+-------+
---------------------+-----------------------------+------------+--------------+------------------+----------+-------------
+
only showing top 20 rows
```

In [34]:
```
df.na.drop().show()
```

```
+------------+-------+-----------+----+-----+------------+-----+-------+-----+-----+-----+-----------+-------+------------+----
---------+-------------+---------------+------------------+----------+-------------+
|RecordNumber|Zipcode|ZipCodeType|City|State|LocationType|  Lat|   Long|Xaxis|Yaxis|Zaxis|WorldRegion|Country|LocationText|
Location|Decommisioned|TaxReturnsFiled|EstimatedPopulation|TotalWages|        Notes|
+------------+-------+-----------+----+-----+------------+-----+-------+-----+-----+-----+-----------+-------+------------+----
---------+-------------+---------------+------------------+----------+-------------+
|       39827|  85209|   STANDARD|MESA|   AZ|     PRIMARY|33.37|-111.64| -0.3|-0.77| 0.55|         NA|     US|    Mesa, AZ|NA-U
S-AZ-MESA|        false|          14962|             26883| 563792730|no NWS data, |
+------------+-------+-----------+----+-----+------------+-----+-------+-----+-----+-----+-----------+-------+------------+----
---------+-------------+---------------+------------------+----------+-------------+
```

In [35]:
```
df.na.drop(how="any").show(truncate=False)
```

```
+------------+-------+-----------+----+-----+------------+-----+-------+-----+-----+-----+-----------+-------+------------+----
---------+-------------+---------------+------------------+----------+-------------+
|RecordNumber|Zipcode|ZipCodeType|City|State|LocationType|Lat  |Long   |Xaxis|Yaxis|Zaxis|WorldRegion|Country|LocationText|Loca
tion     |Decommisioned|TaxReturnsFiled|EstimatedPopulation|TotalWages|Notes        |
+------------+-------+-----------+----+-----+------------+-----+-------+-----+-----+-----+-----------+-------+------------+----
---------+-------------+---------------+------------------+----------+-------------+
|39827       |85209  |STANDARD   |MESA|AZ   |PRIMARY     |33.37|-111.64|-0.3 |-0.77|0.55 |NA         |US     |Mesa, AZ    |NA-U
S-AZ-MESA|false        |14962          |26883             |563792730 |no NWS data, |
+------------+-------+-----------+----+-----+------------+-----+-------+-----+-----+-----+-----------+-------+------------+----
---------+-------------+---------------+------------------+----------+-------------+
```

In [37]:
```
data = [("James","Smith","USA","CA"),
    ("Michael","Rose","USA","NY"),
    ("Robert","Williams","USA","CA"),
    ("Maria","Jones","USA","FL")
  ]
cols = ["firstName", "lastName", "country", "state"]
df = spark.createDataFrame(data=data, schema=cols)
df.show()
```

```
+---------+--------+-------+-----+
|firstName|lastName|country|state|
```

```
+---------+--------+-------+-----+
|    James|   Smith|    USA|   CA|
|  Michael|    Rose|    USA|   NY|
|   Robert|Williams|    USA|   CA|
|    Maria|   Jones|    USA|   FL|
+---------+--------+-------+-----+
```

In [38]:
```
df.select("firstName", "lastName").show()
```

```
+---------+--------+
|firstName|lastName|
+---------+--------+
|    James|   Smith|
|  Michael|    Rose|
|   Robert|Williams|
|    Maria|   Jones|
+---------+--------+
```

In [39]:
```
#using DataFrame object name
df.select(df.firstName, df.lastName).show()
```

```
+---------+--------+
|firstName|lastName|
+---------+--------+
|    James|   Smith|
|  Michael|    Rose|
|   Robert|Williams|
|    Maria|   Jones|
+---------+--------+
```

In [40]:
```
df.select(df["firstName"], df["lastName"]).show()
```

```
+---------+--------+
|firstName|lastName|
+---------+--------+
|    James|   Smith|
|  Michael|    Rose|
|   Robert|Williams|
|    Maria|   Jones|
+---------+--------+
```

In [41]:
```
#using col function
from pyspark.sql.functions import col
df.select(col("firstName").alias("fName"), col("lastName")).show()
```

```
+-------+--------+
|  fName|lastName|
+-------+--------+
|  James|   Smith|
|Michael|    Rose|
| Robert|Williams|
|  Maria|   Jones|
+-------+--------+
```

In [42]:
```
#show all columns
df.select("*").show()
```

```
+---------+--------+-------+-----+
|firstName|lastName|country|state|
+---------+--------+-------+-----+
|    James|   Smith|    USA|   CA|
|  Michael|    Rose|    USA|   NY|
|   Robert|Williams|    USA|   CA|
|    Maria|   Jones|    USA|   FL|
+---------+--------+-------+-----+
```

In [44]:
```
df.select([col for col in df.columns]).show()
```

```
+---------+--------+-------+-----+
|firstName|lastName|country|state|
+---------+--------+-------+-----+
|    James|   Smith|    USA|   CA|
|  Michael|    Rose|    USA|   NY|
|   Robert|Williams|    USA|   CA|
|    Maria|   Jones|    USA|   FL|
+---------+--------+-------+-----+
```

In [46]:
```
df.select(*cols).show()
```

```
+---------+--------+-------+-----+
|firstName|lastName|country|state|
+---------+--------+-------+-----+
|    James|   Smith|    USA|   CA|
|  Michael|    Rose|    USA|   NY|
|   Robert|Williams|    USA|   CA|
|    Maria|   Jones|    USA|   FL|
+---------+--------+-------+-----+
```

In [49]:
```
df.select(df.columns[:3]).show(3)
```

```
+---------+--------+-------+
|firstName|lastName|country|
+---------+--------+-------+
|    James|   Smith|    USA|
|  Michael|    Rose|    USA|
|   Robert|Williams|    USA|
+---------+--------+-------+
only showing top 3 rows
```

In [50]:
```
df.select(df.columns[2:4]).show(3)
```

```
+-------+-----+
|country|state|
+-------+-----+
|    USA|   CA|
|    USA|   NY|
|    USA|   CA|
|    USA|   FL|
+-------+-----+
```

In [51]:
```
df.select(df.columns[2:4]).show(3)
```

```
+-------+-----+
|country|state|
+-------+-----+
|    USA|   CA|
|    USA|   NY|
|    USA|   CA|
+-------+-----+
only showing top 3 rows
```

In [52]:
```
df.select(df.colRegex("`^.*Name*`")).show()
```

```
+---------+--------+
|firstName|lastName|
+---------+--------+
|    James|   Smith|
|  Michael|    Rose|
|   Robert|Williams|
|    Maria|   Jones|
+---------+--------+
```

In [53]:
```
data = [
        (("James",None,"Smith"),"OH","M"),
        (("Anna","Rose",""),"NY","F"),
        (("Julia","","Williams"),"OH","F"),
        (("Maria","Anne","Jones"),"NY","M"),
        (("Jen","Mary","Brown"),"NY","M"),
        (("Mike","Mary","Williams"),"OH","M")
        ]
```

In [55]:
```
df2 = spark.createDataFrame(data=data)
df2.printSchema()
```

```
root
 |-- _1: struct (nullable = true)
 |    |-- _1: string (nullable = true)
 |    |-- _2: string (nullable = true)
 |    |-- _3: string (nullable = true)
 |-- _2: string (nullable = true)
 |-- _3: string (nullable = true)
```

In [56]:
```
df2.show()
```

```
+--------------------+---+---+
|                  _1| _2| _3|
+--------------------+---+---+
|{James, null, Smith}| OH|  M|
|      {Anna, Rose, }| NY|  F|
| {Julia, , Williams}| OH|  F|
|{Maria, Anne, Jones}| NY|  M|
|  {Jen, Mary, Brown}| NY|  M|
|{Mike, Mary, Will...| OH|  M|
+--------------------+---+---+
```

In [59]:
```
df2.select("name").show(truncate=False)
```

```
---------------------------------------------------------------------------
AnalysisException                         Traceback (most recent call last)
C:\Users\PRATIK~1\AppData\Local\Temp/ipykernel_28220/2068437059.py in <module>
----> 1 df2.select("name").show(truncate=False)

~\AppData\Local\Programs\Python\Python310\lib\site-packages\pyspark\sql\dataframe.py in select(self, *cols)
   1683         [Row(name='Alice', age=12), Row(name='Bob', age=15)]
   1684         """
-> 1685         jdf = self._jdf.select(self._jcols(*cols))
   1686         return DataFrame(jdf, self.sql_ctx)
   1687

~\AppData\Local\Programs\Python\Python310\lib\site-packages\py4j\java_gateway.py in __call__(self, *args)
   1319
   1320         answer = self.gateway_client.send_command(command)
-> 1321         return_value = get_return_value(
   1322             answer, self.gateway_client, self.target_id, self.name)
   1323

~\AppData\Local\Programs\Python\Python310\lib\site-packages\pyspark\sql\utils.py in deco(*a, **kw)
    115                 # Hide where the exception came from that shows a non-Pythonic
    116                 # JVM exception message.
--> 117                 raise converted from None
    118             else:
    119                 raise

AnalysisException: cannot resolve 'name' given input columns: [_1, _2, _3];
'Project ['name]
+- LogicalRDD [_1#1095, _2#1096, _3#1097], false
```

In [60]:
```
df2.select("name.firstName").show()
```

```
---------------------------------------------------------------------------
```

```
AnalysisException                         Traceback (most recent call last)
C:\Users\PRATIK~1\AppData\Local\Temp/ipykernel_28220/2281465888.py in <module>
----> 1 df2.select("name.firstName").show()

~\AppData\Local\Programs\Python\Python310\lib\site-packages\pyspark\sql\dataframe.py in select(self, *cols)
   1683         [Row(name='Alice', age=12), Row(name='Bob', age=15)]
   1684         """
-> 1685         jdf = self._jdf.select(self._jcols(*cols))
   1686         return DataFrame(jdf, self.sql_ctx)
   1687

~\AppData\Local\Programs\Python\Python310\lib\site-packages\py4j\java_gateway.py in __call__(self, *args)
   1319
   1320         answer = self.gateway_client.send_command(command)
-> 1321         return_value = get_return_value(
   1322             answer, self.gateway_client, self.target_id, self.name)
   1323

~\AppData\Local\Programs\Python\Python310\lib\site-packages\pyspark\sql\utils.py in deco(*a, **kw)
    115                 # Hide where the exception came from that shows a non-Pythonic
    116                 # JVM exception message.
--> 117                 raise converted from None
    118             else:
    119                 raise

AnalysisException: cannot resolve 'name.firstName' given input columns: [_1, _2, _3];
'Project ['name.firstName]
+- LogicalRDD [_1#1095, _2#1096, _3#1097], false
```

In [ ]: