In [1]:
```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName('CovidAnalysis').getOrCreate()
from numpy import array
from pyspark.sql.types import IntegerType

from pyspark.ml.regression import LinearRegression
```

In [2]:
```
dataset = spark.read.csv("COVID/covid_19_india.csv", inferSchema = True, header = Tr
```

In [3]:
```
dataset
```

Out[3]:
```
DataFrame[Sno: int, Date: string, Time: string, State/UnionTerritory: string, Confir
medIndianNational: string, ConfirmedForeignNational: string, Cured: int, Deaths: in
t, Confirmed: int]
```

In [4]:
```
dataset.show()
```

```
+---+--------+-------+--------------------+-----------------------+-----------------
-------+-----+------+---------+
|Sno|    Date|   Time|State/UnionTerritory|ConfirmedIndianNational|ConfirmedForeignN
ational|Cured|Deaths|Confirmed|
+---+--------+-------+--------------------+-----------------------+-----------------
-------+-----+------+---------+
|  1|30/01/20|6:00 PM|              Kerala|                      1|
0|    0|     0|        1|
|  2|31/01/20|6:00 PM|              Kerala|                      1|
0|    0|     0|        1|
|  3|01/02/20|6:00 PM|              Kerala|                      2|
0|    0|     0|        2|
|  4|02/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
|  5|03/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
|  6|04/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
|  7|05/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
|  8|06/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
|  9|07/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
| 10|08/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
| 11|09/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
| 12|10/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
| 13|11/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
| 14|12/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
| 15|13/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
| 16|14/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
| 17|15/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
| 18|16/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
| 19|17/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
| 20|18/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
+---+--------+-------+--------------------+-----------------------+-----------------
-------+-----+------+---------+
only showing top 20 rows
```

In [5]:
```
dataset.printSchema()
```

```
root
 |-- Sno: integer (nullable = true)
 |-- Date: string (nullable = true)
 |-- Time: string (nullable = true)
 |-- State/UnionTerritory: string (nullable = true)
 |-- ConfirmedIndianNational: string (nullable = true)
 |-- ConfirmedForeignNational: string (nullable = true)
 |-- Cured: integer (nullable = true)
 |-- Deaths: integer (nullable = true)
 |-- Confirmed: integer (nullable = true)
```

In [6]:
```
dataset = dataset.withColumn("ConfirmedIndianNational", dataset["ConfirmedIndianNati

dataset = dataset.withColumn("ConfirmedForeignNational", dataset["ConfirmedForeignNa

dataset = dataset.dropna(subset = ("ConfirmedIndianNational", "ConfirmedForeignNatio
```

In [7]:
```
dataset.show()
```

```
+---+--------+-------+--------------------+-----------------------+-----------------
-------+-----+------+---------+
|Sno|    Date|   Time|State/UnionTerritory|ConfirmedIndianNational|ConfirmedForeignN
ational|Cured|Deaths|Confirmed|
+---+--------+-------+--------------------+-----------------------+-----------------
-------+-----+------+---------+
|  1|30/01/20|6:00 PM|              Kerala|                      1|
0|    0|     0|        1|
|  2|31/01/20|6:00 PM|              Kerala|                      1|
0|    0|     0|        1|
|  3|01/02/20|6:00 PM|              Kerala|                      2|
0|    0|     0|        2|
|  4|02/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
|  5|03/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
|  6|04/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
|  7|05/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
|  8|06/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
|  9|07/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
| 10|08/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
| 11|09/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
| 12|10/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
| 13|11/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
| 14|12/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
| 15|13/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
| 16|14/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
| 17|15/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
| 18|16/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
| 19|17/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
| 20|18/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|
+---+--------+-------+--------------------+-----------------------+-----------------
-------+-----+------+---------+
only showing top 20 rows
```

In [8]:

```
dataset.printSchema()
```

```
root
 |-- Sno: integer (nullable = true)
 |-- Date: string (nullable = true)
 |-- Time: string (nullable = true)
 |-- State/UnionTerritory: string (nullable = true)
 |-- ConfirmedIndianNational: integer (nullable = true)
 |-- ConfirmedForeignNational: integer (nullable = true)
 |-- Cured: integer (nullable = true)
 |-- Deaths: integer (nullable = true)
 |-- Confirmed: integer (nullable = true)
```

In [9]:

```
from pyspark.ml.linalg import Vectors
from pyspark.ml.feature import VectorAssembler
```

In [10]:

```
vector = VectorAssembler(inputCols = ["ConfirmedIndianNational", "ConfirmedForeignNa
```

In [11]:

```
output = vector.transform(dataset)
```

In [12]:

```
output.show()
```

```
+---+--------+-------+--------------------+-----------------------+-----------------
-------+-----+------+---------+------------------+
|Sno|    Date|   Time|State/UnionTerritory|ConfirmedIndianNational|ConfirmedForeignN
ational|Cured|Deaths|Confirmed|    Output Features|
+---+--------+-------+--------------------+-----------------------+-----------------
-------+-----+------+---------+------------------+
|  1|30/01/20|6:00 PM|              Kerala|                      1|
0|    0|     0|        1|(5,[0,4],[1.0,1.0])|
|  2|31/01/20|6:00 PM|              Kerala|                      1|
0|    0|     0|        1|(5,[0,4],[1.0,1.0])|
|  3|01/02/20|6:00 PM|              Kerala|                      2|
0|    0|     0|        2|(5,[0,4],[2.0,2.0])|
|  4|02/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|(5,[0,4],[3.0,3.0])|
|  5|03/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|(5,[0,4],[3.0,3.0])|
|  6|04/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|(5,[0,4],[3.0,3.0])|
|  7|05/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|(5,[0,4],[3.0,3.0])|
|  8|06/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|(5,[0,4],[3.0,3.0])|
|  9|07/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|(5,[0,4],[3.0,3.0])|
| 10|08/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|(5,[0,4],[3.0,3.0])|
| 11|09/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|(5,[0,4],[3.0,3.0])|
| 12|10/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|(5,[0,4],[3.0,3.0])|
| 13|11/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|(5,[0,4],[3.0,3.0])|
| 14|12/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|(5,[0,4],[3.0,3.0])|
| 15|13/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|(5,[0,4],[3.0,3.0])|
| 16|14/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|(5,[0,4],[3.0,3.0])|
| 17|15/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|(5,[0,4],[3.0,3.0])|
| 18|16/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|(5,[0,4],[3.0,3.0])|
| 19|17/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|(5,[0,4],[3.0,3.0])|
| 20|18/02/20|6:00 PM|              Kerala|                      3|
0|    0|     0|        3|(5,[0,4],[3.0,3.0])|
+---+--------+-------+--------------------+------------------------+-----------------
-------+-----+------+---------+-------------------+
only showing top 20 rows
```

In [13]:

```
output.select("Output Features").show()
```

```
+-------------------+
|    Output Features|
+-------------------+
|(5,[0,4],[1.0,1.0])|
|(5,[0,4],[1.0,1.0])|
|(5,[0,4],[2.0,2.0])|
|(5,[0,4],[3.0,3.0])|
|(5,[0,4],[3.0,3.0])|
|(5,[0,4],[3.0,3.0])|
|(5,[0,4],[3.0,3.0])|
|(5,[0,4],[3.0,3.0])|
|(5,[0,4],[3.0,3.0])|
|(5,[0,4],[3.0,3.0])|
|(5,[0,4],[3.0,3.0])|
|(5,[0,4],[3.0,3.0])|
|(5,[0,4],[3.0,3.0])|
|(5,[0,4],[3.0,3.0])|
|(5,[0,4],[3.0,3.0])|
|(5,[0,4],[3.0,3.0])|
|(5,[0,4],[3.0,3.0])|
|(5,[0,4],[3.0,3.0])|
|(5,[0,4],[3.0,3.0])|
|(5,[0,4],[3.0,3.0])|
+-------------------+
only showing top 20 rows
```

In [14]:

```
output.columns
```

Out[14]:

```
['Sno',
 'Date',
 'Time',
 'State/UnionTerritory',
 'ConfirmedIndianNational',
 'ConfirmedForeignNational',
 'Cured',
 'Deaths',
 'Confirmed',
 'Output Features']
```

In [15]:

```
finalized_vector_data = output.select("Date", "Time", "State/UnionTerritory", "Outpu
```

In [16]:

```
finalized_vector_data.show()
```

```
+--------+-------+--------------------+-------------------+---------+
|    Date|   Time|State/UnionTerritory|    Output Features|Confirmed|
+--------+-------+--------------------+-------------------+---------+
|30/01/20|6:00 PM|              Kerala|(5,[0,4],[1.0,1.0])|        1|
|31/01/20|6:00 PM|              Kerala|(5,[0,4],[1.0,1.0])|        1|
|01/02/20|6:00 PM|              Kerala|(5,[0,4],[2.0,2.0])|        2|
|02/02/20|6:00 PM|              Kerala|(5,[0,4],[3.0,3.0])|        3|
|03/02/20|6:00 PM|              Kerala|(5,[0,4],[3.0,3.0])|        3|
|04/02/20|6:00 PM|              Kerala|(5,[0,4],[3.0,3.0])|        3|
|05/02/20|6:00 PM|              Kerala|(5,[0,4],[3.0,3.0])|        3|
|06/02/20|6:00 PM|              Kerala|(5,[0,4],[3.0,3.0])|        3|
|07/02/20|6:00 PM|              Kerala|(5,[0,4],[3.0,3.0])|        3|
|08/02/20|6:00 PM|              Kerala|(5,[0,4],[3.0,3.0])|        3|
|09/02/20|6:00 PM|              Kerala|(5,[0,4],[3.0,3.0])|        3|
|10/02/20|6:00 PM|              Kerala|(5,[0,4],[3.0,3.0])|        3|
|11/02/20|6:00 PM|              Kerala|(5,[0,4],[3.0,3.0])|        3|
|12/02/20|6:00 PM|              Kerala|(5,[0,4],[3.0,3.0])|        3|
|13/02/20|6:00 PM|              Kerala|(5,[0,4],[3.0,3.0])|        3|
|14/02/20|6:00 PM|              Kerala|(5,[0,4],[3.0,3.0])|        3|
|15/02/20|6:00 PM|              Kerala|(5,[0,4],[3.0,3.0])|        3|
|16/02/20|6:00 PM|              Kerala|(5,[0,4],[3.0,3.0])|        3|
|17/02/20|6:00 PM|              Kerala|(5,[0,4],[3.0,3.0])|        3|
|18/02/20|6:00 PM|              Kerala|(5,[0,4],[3.0,3.0])|        3|
+--------+-------+--------------------+-------------------+---------+
only showing top 20 rows
```

In [17]:

```
train_data, test_data = finalized_vector_data.randomSplit([0.75, 0.25])
```

In [18]:

```
regressor = LinearRegression(featuresCol="Output Features", labelCol= "Confirmed")
regressor = regressor.fit(train_data)
```

In [19]:

```
regressor.coefficients
```

Out[19]:

```
DenseVector([-0.0885, -0.0885, -0.0, -0.0, 1.0885])
```

In [20]:

```
regressor.intercept
```

Out[20]:

```
-4.561013232155169e-17
```

In [21]:

```
pred_result = regressor.evaluate(test_data)
```

In [23]:

```
pred_result.predictions.show(40)
```

```
+--------+-------+--------------------+--------------------+---------+--------------
----+
|    Date|   Time|State/UnionTerritory|     Output Features|Confirmed|        predic
tion|
+--------+-------+--------------------+--------------------+---------+--------------
----+
|01/02/20|6:00 PM|              Kerala| (5,[0,4],[2.0,2.0])|        2|
2.0|
|03/03/20|6:00 PM|           Rajasthan| (5,[1,4],[1.0,1.0])|        1|0.999999999999
9999|
|04/02/20|6:00 PM|              Kerala| (5,[0,4],[3.0,3.0])|        3|
3.0|
|04/03/20|6:00 PM|              Kerala|[3.0,0.0,3.0,0.0,...|        3| 2.99999999999
9999|
|05/03/20|6:00 PM|               Delhi| (5,[0,4],[2.0,2.0])|        2|
2.0|
|05/03/20|6:00 PM|              Kerala|[3.0,0.0,3.0,0.0,...|        3| 2.99999999999
9999|
|06/03/20|6:00 PM|           Rajasthan|[1.0,14.0,0.0,0.0...|       15|14.99999999999
9996|
|06/03/20|6:00 PM|           Telengana| (5,[0,4],[1.0,1.0])|        1|
1.0|
|06/03/20|6:00 PM|       Uttar Pradesh| (5,[0,4],[7.0,7.0])|        7| 7.00000000000
0001|
|07/03/20|6:00 PM|               Delhi| (5,[0,4],[3.0,3.0])|        3|
3.0|
|07/03/20|6:00 PM|          Tamil Nadu| (5,[0,4],[1.0,1.0])|        1|
1.0|
|08/02/20|6:00 PM|              Kerala| (5,[0,4],[3.0,3.0])|        3|
3.0|
|08/03/20|6:00 PM|       Uttar Pradesh| (5,[0,4],[7.0,7.0])|        7| 7.00000000000
0001|
|09/03/20|6:00 PM|       Uttar Pradesh| (5,[0,4],[7.0,7.0])|       7| 7.00000000000
0001|
|10/03/20|6:00 PM|           Karnataka| (5,[0,4],[4.0,4.0])|       4|
4.0|
|10/03/20|6:00 PM|       Uttar Pradesh| (5,[0,4],[7.0,7.0])|       7| 7.00000000000
0001|
|12/03/20|6:00 PM|             Haryana|(5,[1,4],[14.0,14...|      14|
14.0|
|12/03/20|6:00 PM|           Karnataka| (5,[0,4],[4.0,4.0])|       4|
4.0|
|12/03/20|6:00 PM|              Kerala|[17.0,0.0,3.0,0.0...|      17|
17.0|
|12/03/20|6:00 PM|              Ladakh| (5,[0,4],[3.0,3.0])|       3|
3.0|
|13/03/20|6:00 PM|   Jammu and Kashmir| (5,[0,4],[1.0,1.0])|       1|
1.0|
|13/03/20|6:00 PM|         Maharashtra|(5,[0,4],[14.0,14...|      14|14.00000000000
0002|
|14/03/20|6:00 PM|      Andhra Pradesh| (5,[0,4],[1.0,1.0])|       1|
1.0|
|14/03/20|6:00 PM|               Delhi|[7.0,0.0,1.0,1.0,...|       7| 6.99999999999
9998|
|14/03/20|6:00 PM|             Haryana|(5,[1,4],[14.0,14...|      14|
14.0|
|14/03/20|6:00 PM|         Maharashtra|(5,[0,4],[14.0,14...|      14|14.00000000000
0002|
|14/03/20|6:00 PM|          Tamil Nadu| (5,[0,4],[1.0,1.0])|       1|
1.0|
|15/02/20|6:00 PM|              Kerala| (5,[0,4],[3.0,3.0])|       3|
3.0|
|15/03/20|6:00 PM|           Karnataka|[6.0,0.0,0.0,1.0,...|       6| 5.99999999999
9998|
|15/03/20|6:00 PM|         Maharashtra|(5,[0,4],[32.0,32...|      32|
32.0|
|15/03/20|6:00 PM|              Punjab| (5,[0,4],[1.0,1.0])|       1|
1.0|
|15/03/20|6:00 PM|          Tamil Nadu| (5,[0,4],[1.0,1.0])|       1|
1.0|
|16/03/20|6:00 PM|              Ladakh| (5,[0,4],[4.0,4.0])|       4|
4.0|
|16/03/20|6:00 PM|              Odisha| (5,[0,4],[1.0,1.0])|       1|
1.0|
|16/03/20|6:00 PM|           Telengana|[3.0,0.0,1.0,0.0,...|       3|2.999999999999
9996|
|17/03/20|6:00 PM|      Andhra Pradesh| (5,[0,4],[1.0,1.0])|       1|
1.0|
|17/03/20|6:00 PM|             Haryana|[1.0,14.0,0.0,0.0...|      15|14.99999999999
9996|
|17/03/20|6:00 PM|           Rajasthan|[2.0,2.0,3.0,0.0,...|       4|3.999999999999
9987|
|17/03/20|6:00 PM|           Telengana|[3.0,2.0,1.0,0.0,...|       5| 4.99999999999
9999|
|17/03/20|6:00 PM|         Uttarakhand| (5,[0,4],[1.0,1.0])|       1|
1.0|
+--------+-------+--------------------+--------------------+--------+--------------
----+
only showing top 40 rows
```

```
In [ ]:
```