In [18]:
```
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 [19]:
```
dataset = spark.read.csv("COVID/StatewiseTestingDetails.csv", inferSchema = True, he
```

In [20]:
```
dataset
```

Out[20]:
```
DataFrame[Date: string, State: string, TotalSamples: double, Negative: string, Posit
ive: double]
```

In [21]:
```
dataset.show()
```

```
+----------+--------------------+------------+--------+--------+
|      Date|               State|TotalSamples|Negative|Positive|
+----------+--------------------+------------+--------+--------+
|2020-04-17|Andaman and Nicob...|      1403.0|    1210|    12.0|
|2020-04-24|Andaman and Nicob...|      2679.0|    null|    27.0|
|2020-04-27|Andaman and Nicob...|      2848.0|    null|    33.0|
|2020-05-01|Andaman and Nicob...|      3754.0|    null|    33.0|
|2020-05-16|Andaman and Nicob...|      6677.0|    null|    33.0|
|2020-05-19|Andaman and Nicob...|      6965.0|    null|    33.0|
|2020-05-20|Andaman and Nicob...|      7082.0|    null|    33.0|
|2020-05-21|Andaman and Nicob...|      7167.0|    null|    33.0|
|2020-05-22|Andaman and Nicob...|      7263.0|    null|    33.0|
|2020-05-23|Andaman and Nicob...|      7327.0|    null|    33.0|
|2020-05-24|Andaman and Nicob...|      7327.0|    null|    33.0|
|2020-05-25|Andaman and Nicob...|      7363.0|    null|    33.0|
|2020-05-26|Andaman and Nicob...|      7448.0|    null|    33.0|
|2020-05-27|Andaman and Nicob...|      7499.0|    null|    33.0|
|2020-05-28|Andaman and Nicob...|      7519.0|    null|    33.0|
|2020-05-29|Andaman and Nicob...|      7567.0|    null|    33.0|
|2020-05-30|Andaman and Nicob...|      7567.0|    null|    33.0|
|2020-05-31|Andaman and Nicob...|      7706.0|    null|    33.0|
|2020-06-01|Andaman and Nicob...|      7805.0|    null|    33.0|
|2020-06-02|Andaman and Nicob...|      8086.0|    null|    33.0|
+----------+--------------------+------------+--------+--------+
only showing top 20 rows
```

In [41]:
```
dataset = dataset.withColumn("Negative", dataset["Negative"].cast(IntegerType()))
dataset = dataset.dropna(subset = ("Negative", "TotalSamples", "Positive"))
```

In [42]:
```
dataset.show()
```

```
+----------+--------------------+------------+--------+--------+
|      Date|               State|TotalSamples|Negative|Positive|
+----------+--------------------+------------+--------+--------+
|2020-04-17|Andaman and Nicob...|      1403.0|    1210|    12.0|
|2020-04-02|      Andhra Pradesh|      1800.0|    1175|   132.0|
|2020-04-10|      Andhra Pradesh|      6374.0|    6009|   365.0|
|2020-04-11|      Andhra Pradesh|      6958.0|    6577|   381.0|
|2020-04-12|      Andhra Pradesh|      6958.0|    6553|   405.0|
|2020-04-13|      Andhra Pradesh|      8755.0|    8323|   432.0|
|2020-04-14|      Andhra Pradesh|     10505.0|   10032|   473.0|
|2020-04-15|      Andhra Pradesh|     11613.0|   11088|   525.0|
|2020-04-16|      Andhra Pradesh|     20235.0|   19701|   534.0|
|2020-04-18|      Andhra Pradesh|     21450.0|   20487|   603.0|
|2020-04-19|      Andhra Pradesh|     26958.0|   26311|   647.0|
|2020-04-20|      Andhra Pradesh|     30733.0|   30011|   722.0|
|2020-04-21|      Andhra Pradesh|     35755.0|   34998|   757.0|
|2020-04-22|      Andhra Pradesh|     41512.0|   40699|   813.0|
|2020-04-23|      Andhra Pradesh|     48032.0|   47139|   893.0|
```

```
|2020-04-24|      Andhra Pradesh|     54338.0|   53383|   955.0|
|2020-04-25|      Andhra Pradesh|     61266.0|   60250|  1016.0|
|2020-04-26|      Andhra Pradesh|     68034.0|   66937|  1097.0|
|2020-04-27|      Andhra Pradesh|     74551.0|   73374|  1177.0|
|2020-04-28|      Andhra Pradesh|     80334.0|   79075|  1259.0|
+----------+--------------------+------------+--------+--------+
only showing top 20 rows
```

In [43]:
```
dataset.printSchema()
```

```
root
 |-- Date: string (nullable = true)
 |-- State: string (nullable = true)
 |-- TotalSamples: double (nullable = true)
 |-- Negative: integer (nullable = true)
 |-- Positive: double (nullable = true)
```

In [44]:
```
dataset.printSchema()
```

```
root
 |-- Date: string (nullable = true)
 |-- State: string (nullable = true)
 |-- TotalSamples: double (nullable = true)
 |-- Negative: integer (nullable = true)
 |-- Positive: double (nullable = true)
```

In [45]:
```
from pyspark.ml.linalg import Vectors
from pyspark.ml.feature import VectorAssembler
```

In [49]:
```
vector = VectorAssembler(inputCols = ["TotalSamples", "Negative", "Positive"], outpu
```

In [50]:
```
output = vector.transform(dataset)
```

In [51]:
```
output.show()
```

```
+----------+--------------------+------------+--------+--------+--------------------
+
|      Date|               State|TotalSamples|Negative|Positive|     Output Features
|
+----------+--------------------+------------+--------+--------+--------------------
+
|2020-04-17|Andaman and Nicob...|      1403.0|    1210|    12.0|[1403.0,1210.0,12.0]
|
|2020-04-02|      Andhra Pradesh|      1800.0|    1175|   132.0|[1800.0,1175.0,13...
|
|2020-04-10|      Andhra Pradesh|      6374.0|    6009|   365.0|[6374.0,6009.0,36...
|
|2020-04-11|      Andhra Pradesh|      6958.0|    6577|   381.0|[6958.0,6577.0,38...
|
|2020-04-12|      Andhra Pradesh|      6958.0|    6553|   405.0|[6958.0,6553.0,40...
|
|2020-04-13|      Andhra Pradesh|      8755.0|    8323|   432.0|[8755.0,8323.0,43...
|
|2020-04-14|      Andhra Pradesh|     10505.0|   10032|   473.0|[10505.0,10032.0,...
|
|2020-04-15|      Andhra Pradesh|     11613.0|   11088|   525.0|[11613.0,11088.0,...
|
|2020-04-16|      Andhra Pradesh|     20235.0|   19701|   534.0|[20235.0,19701.0,...
|
|2020-04-18|      Andhra Pradesh|     21450.0|   20487|   603.0|[21450.0,20487.0,...
|
|2020-04-19|      Andhra Pradesh|     26958.0|   26311|   647.0|[26958.0,26311.0,...
|
|2020-04-20|      Andhra Pradesh|     30733.0|   30011|   722.0|[30733.0,30011.0,...
|
```

```
|2020-04-21|      Andhra Pradesh|     35755.0|   34998|   757.0|[35755.0,34998.0,...
|
|2020-04-22|      Andhra Pradesh|     41512.0|   40699|   813.0|[41512.0,40699.0,...
|
|2020-04-23|      Andhra Pradesh|     48032.0|   47139|   893.0|[48032.0,47139.0,...
|
|2020-04-24|      Andhra Pradesh|     54338.0|   53383|   955.0|[54338.0,53383.0,...
|
|2020-04-25|      Andhra Pradesh|     61266.0|   60250|  1016.0|[61266.0,60250.0,...
|
|2020-04-26|      Andhra Pradesh|     68034.0|   66937|  1097.0|[68034.0,66937.0,...
|
|2020-04-27|      Andhra Pradesh|     74551.0|   73374|  1177.0|[74551.0,73374.0,...
|
|2020-04-28|      Andhra Pradesh|     80334.0|   79075|  1259.0|[80334.0,79075.0,...
|
+----------+--------------------+------------+--------+--------+--------------------
+
only showing top 20 rows
```

In [52]:
```
output.select("Output Features").show()
```

```
+--------------------+
|     Output Features|
+--------------------+
|[1403.0,1210.0,12.0]|
|[1800.0,1175.0,13...|
|[6374.0,6009.0,36...|
|[6958.0,6577.0,38...|
|[6958.0,6553.0,40...|
|[8755.0,8323.0,43...|
|[10505.0,10032.0,...|
|[11613.0,11088.0,...|
|[20235.0,19701.0,...|
|[21450.0,20487.0,...|
|[26958.0,26311.0,...|
|[30733.0,30011.0,...|
|[35755.0,34998.0,...|
|[41512.0,40699.0,...|
|[48032.0,47139.0,...|
|[54338.0,53383.0,...|
|[61266.0,60250.0,...|
|[68034.0,66937.0,...|
|[74551.0,73374.0,...|
|[80334.0,79075.0,...|
+--------------------+
only showing top 20 rows
```

In [53]:
```
output.columns
```

Out[53]:
```
['Date', 'State', 'TotalSamples', 'Negative', 'Positive', 'Output Features']
```

In [81]:
```
finalized_vector_data = output.select("Date", "State", "Output Features", "Positive"
```

In [82]:
```
finalized_vector_data.show()
```

```
+----------+--------------------+--------------------+--------+
|      Date|               State|     Output Features|Positive|
+----------+--------------------+--------------------+--------+
|2020-04-17|Andaman and Nicob...|[1403.0,1210.0,12.0]|    12.0|
|2020-04-02|      Andhra Pradesh|[1800.0,1175.0,13...|   132.0|
|2020-04-10|      Andhra Pradesh|[6374.0,6009.0,36...|   365.0|
|2020-04-11|      Andhra Pradesh|[6958.0,6577.0,38...|   381.0|
|2020-04-12|      Andhra Pradesh|[6958.0,6553.0,40...|   405.0|
|2020-04-13|      Andhra Pradesh|[8755.0,8323.0,43...|   432.0|
|2020-04-14|      Andhra Pradesh|[10505.0,10032.0,...|   473.0|
```

```
|2020-04-15|      Andhra Pradesh|[11613.0,11088.0,...|   525.0|
|2020-04-16|      Andhra Pradesh|[20235.0,19701.0,...|   534.0|
|2020-04-18|      Andhra Pradesh|[21450.0,20487.0,...|   603.0|
|2020-04-19|      Andhra Pradesh|[26958.0,26311.0,...|   647.0|
|2020-04-20|      Andhra Pradesh|[30733.0,30011.0,...|   722.0|
|2020-04-21|      Andhra Pradesh|[35755.0,34998.0,...|   757.0|
|2020-04-22|      Andhra Pradesh|[41512.0,40699.0,...|   813.0|
|2020-04-23|      Andhra Pradesh|[48032.0,47139.0,...|   893.0|
|2020-04-24|      Andhra Pradesh|[54338.0,53383.0,...|   955.0|
|2020-04-25|      Andhra Pradesh|[61266.0,60250.0,...|  1016.0|
|2020-04-26|      Andhra Pradesh|[68034.0,66937.0,...|  1097.0|
|2020-04-27|      Andhra Pradesh|[74551.0,73374.0,...|  1177.0|
|2020-04-28|      Andhra Pradesh|[80334.0,79075.0,...|  1259.0|
+----------+--------------------+--------------------+--------+
only showing top 20 rows
```

In [83]:
```
train_data, test_data = finalized_vector_data.randomSplit([0.75, 0.25])
```

In [84]:
```
regressor = LinearRegression(featuresCol="Output Features", labelCol= "Positive")
regressor = regressor.fit(train_data)
```

In [85]:
```
regressor.coefficients
```

Out[85]:
```
DenseVector([0.0, -0.0, 1.0])
```

In [86]:
```
regressor.intercept
```

Out[86]:
```
-1.267084913451375e-11
```

In [87]:
```
pred_result = regressor.evaluate(test_data)
```

In [91]:
```
pred_result.predictions.show(40)
```

```
+----------+--------------------+--------------------+--------+--------------------+
|      Date|               State|     Output Features|Positive|          prediction|
+----------+--------------------+--------------------+--------+--------------------+
|2020-04-02|        Chhattisgarh|  [1232.0,921.0,9.0]|     9.0|   8.999999999990523|
|2020-04-02|                 Goa|   [220.0,197.0,5.0]|     5.0|   4.999999999987515|
|2020-04-02|              Kerala|[8456.0,7622.0,28...|   286.0|  285.99999999999295|
|2020-04-03|             Haryana| [1325.0,938.0,44.0]|    44.0|   43.99999999999097|
|2020-04-03|           Karnataka|[4587.0,4281.0,12...|   128.0|  127.99999999998909|
|2020-04-03|         Tamil Nadu|[3684.0,2789.0,41...|   411.0|  410.99999999999244|
|2020-04-05|       Uttar Pradesh|[5255.0,4796.0,27...|   278.0|   277.9999999999891|
|2020-04-06|        Maharashtra|[17563.0,15808.0,...|   868.0|   867.9999999999964|
|2020-04-06|            Mizoram|      [58.0,0.0,1.0]|     1.0|  0.9999999999879382|
|2020-04-07|              Delhi|[9041.0,7308.0,57...|   576.0|   575.9999999999995|
|2020-04-07|           Karnataka|[6580.0,5942.0,17...|   175.0|  174.99999999999207|
|2020-04-07|         Uttarakhand|[1289.0,1092.0,32.0]|    32.0|  31.999999999989054|
|2020-04-08|           Karnataka|[6967.0,6473.0,18...|   181.0|  180.99999999999045|
|2020-04-09|Dadra and Nagar H...|     [80.0,80.0,0.0]|     0.0|-1.26738045601980...|
|2020-04-09|              Kerala|[12710.0,11469.0,...|   357.0|   356.9999999999964|
|2020-04-09|              Punjab|[3192.0,2777.0,13...|   130.0|  129.99999999999028|
|2020-04-09|         Tamil Nadu|[7267.0,5824.0,83...|   834.0|   833.9999999999939|
|2020-04-10|      Andhra Pradesh|[6374.0,6009.0,36...|   365.0|  364.99999999998715|
|2020-04-10|          Chandigarh|  [223.0,199.0,19.0]|    19.0|  18.999999999987377|
|2020-04-10|        Chhattisgarh|[3473.0,3322.0,18.0]|    18.0|  17.999999999988628|
|2020-04-10|             Gujarat|[7718.0,7237.0,37...|   378.0|  377.99999999998823|
|2020-04-10|   Jammu and Kashmir|[2961.0,2754.0,20...|   207.0|  206.99999999998724|
|2020-04-10|         Maharashtra|[30000.0,28865.0,...|  1135.0|  1134.9999999999864|
|2020-04-11|Dadra and Nagar H...|   [211.0,211.0,0.0]|     0.0|-1.26786440697560...|
|2020-04-11|               Delhi|[11709.0,10218.0,...|  1069.0|  1068.9999999999914|
|2020-04-11|   Jammu and Kashmir|[3206.0,2982.0,22...|   224.0|  223.99999999998724|
|2020-04-11|           Karnataka|[8560.0,8231.0,21...|   215.0|  214.99999999998826|
|2020-04-11|         Maharashtra|[31841.0,30477.0,...|  1761.0|   1760.999999999982|
|2020-04-11|         Tamil Nadu|[9842.0,7779.0,96...|   969.0|   968.9999999999989|
```

```
|2020-04-12|              Assam|[3138.0,2973.0,29.0]|    29.0|  28.999999999988674|
|2020-04-12|             Punjab|[4281.0,3590.0,17...|   170.0|  169.99999999999278|
|2020-04-13|          Chandigarh|  [296.0,263.0,21.0]|    21.0|   20.99999999998745|
|2020-04-13|            Gujarat|[14251.0,12970.0,...|   572.0|   571.9999999999945|
|2020-04-13|   Jammu and Kashmir|[4065.0,3795.0,27...|   270.0|   269.9999999999872|
|2020-04-13|             Odisha|[4170.0,4116.0,54.0]|    54.0|   53.99999999998719|
|2020-04-13|             Punjab|[4480.0,3858.0,17...|   176.0|  175.99999999999199|
|2020-04-13|           Rajasthan|[31804.0,28657.0,...|   847.0|   847.0000000000109|
|2020-04-13|         Uttarakhand|[1998.0,1665.0,35.0]|    35.0|   34.99999999999046|
|2020-04-14|        Chhattisgarh|[4812.0,4319.0,33.0]|    33.0|   32.99999999999209|
|2020-04-14|              Delhi|[16282.0,13748.0,...|  1561.0|  1560.9999999999975|
+----------+--------------------+--------------------+--------+--------------------+
only showing top 40 rows
```

In [ ]: