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

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

In [3]:
```
df=spark.read.option("header","true").csv('D:\\tips.csv',inferSchema=True)
df.show()
```

```
+----------+----+------+------+---+------+----+
|total_bill| tip|   sex|smoker|day|  time|size|
+----------+----+------+------+---+------+----+
|     16.99|1.01|Female|    No|Sun|Dinner|   2|
|     10.34|1.66|  Male|    No|Sun|Dinner|   3|
|     21.01|3.5 |  Male|    No|Sun|Dinner|   3|
|     23.68|3.31|  Male|    No|Sun|Dinner|   2|
|     24.59|3.61|Female|    No|Sun|Dinner|   4|
|     25.29|4.71|  Male|    No|Sun|Dinner|   4|
|      8.77|2.01|  Male|    No|Sun|Dinner|   2|
|     26.88|3.12|  Male|    No|Sun|Dinner|   4|
|     15.04|1.96|  Male|    No|Sun|Dinner|   2|
|     14.78|3.23|  Male|    No|Sun|Dinner|   2|
|     10.27|1.71|  Male|    No|Sun|Dinner|   2|
|     35.26| 5.0|Female|    No|Sun|Dinner|   4|
|     15.42|1.57|  Male|    No|Sun|Dinner|   2|
|     18.43| 3.0|  Male|    No|Sun|Dinner|   4|
|     14.83|3.02|Female|    No|Sun|Dinner|   2|
|     21.58|3.92|  Male|    No|Sun|Dinner|   2|
|     10.33|1.67|Female|    No|Sun|Dinner|   3|
|     16.29|3.71|  Male|    No|Sun|Dinner|   3|
|     16.97| 3.5|Female|    No|Sun|Dinner|   3|
|     20.65|3.35|  Male|    No|Sat|Dinner|   3|
+----------+----+------+------+---+------+----+
only showing top 20 rows
```

In [4]:
```
df.printSchema()
```

```
root
 |-- total_bill: double (nullable = true)
 |-- tip: double (nullable = true)
 |-- sex: string (nullable = true)
 |-- smoker: string (nullable = true)
 |-- day: string (nullable = true)
 |-- time: string (nullable = true)
 |-- size: integer (nullable = true)
```

In [5]:
```
from pyspark.ml.feature import StringIndexer
from pyspark.ml.feature import VectorAssembler
```

In [15]:
```
indexer=StringIndexer(inputCol='sex',outputCol='Gender_Col')
df1=indexer.fit(df).transform(df)
df1.show()
```

```
+----------+----+------+------+---+------+----+----------+
|total_bill| tip|   sex|smoker|day|  time|size|Gender_Col|
+----------+----+------+------+---+------+----+----------+
|     16.99|1.01|Female|    No|Sun|Dinner|   2|       1.0|
|     10.34|1.66|  Male|    No|Sun|Dinner|   3|       0.0|
|     21.01|3.5 |  Male|    No|Sun|Dinner|   3|       0.0|
|     23.68|3.31|  Male|    No|Sun|Dinner|   2|       0.0|
|     24.59|3.61|Female|    No|Sun|Dinner|   4|       1.0|
|     25.29|4.71|  Male|    No|Sun|Dinner|   4|       0.0|
|      8.77|2.01|  Male|    No|Sun|Dinner|   2|       0.0|
|     26.88|3.12|  Male|    No|Sun|Dinner|   4|       0.0|
|     15.04|1.96|  Male|    No|Sun|Dinner|   2|       0.0|
|     14.78|3.23|  Male|    No|Sun|Dinner|   2|       0.0|
|     10.27|1.71|  Male|    No|Sun|Dinner|   2|       0.0|
|     35.26| 5.0|Female|    No|Sun|Dinner|   4|       1.0|
|     15.42|1.57|  Male|    No|Sun|Dinner|   2|       0.0|
|     18.43| 3.0|  Male|    No|Sun|Dinner|   4|       0.0|
|     14.83|3.02|Female|    No|Sun|Dinner|   2|       1.0|
|     21.58|3.92|  Male|    No|Sun|Dinner|   2|       0.0|
|     10.33|1.67|Female|    No|Sun|Dinner|   3|       1.0|
|     16.29|3.71|  Male|    No|Sun|Dinner|   3|       0.0|
|     16.97| 3.5|Female|    No|Sun|Dinner|   3|       1.0|
|     20.65|3.35|  Male|    No|Sat|Dinner|   3|       0.0|
+----------+----+------+------+---+------+----+----------+
only showing top 20 rows
```

In [16]:
```
indexer=StringIndexer(inputCol='smoker',outputCol='Smoker_Col')
df2=indexer.fit(df1).transform(df1)
df2.show()
```

```
+----------+----+------+------+---+------+----+----------+----------+
|total_bill| tip|   sex|smoker|day|  time|size|Gender_Col|Smoker_Col|
+----------+----+------+------+---+------+----+----------+----------+
|     16.99|1.01|Female|    No|Sun|Dinner|   2|       1.0|       0.0|
|     10.34|1.66|  Male|    No|Sun|Dinner|   3|       0.0|       0.0|
|     21.01|3.5 |  Male|    No|Sun|Dinner|   3|       0.0|       0.0|
|     23.68|3.31|  Male|    No|Sun|Dinner|   2|       0.0|       0.0|
|     24.59|3.61|Female|    No|Sun|Dinner|   4|       1.0|       0.0|
|     25.29|4.71|  Male|    No|Sun|Dinner|   4|       0.0|       0.0|
|      8.77|2.01|  Male|    No|Sun|Dinner|   2|       0.0|       0.0|
|     26.88|3.12|  Male|    No|Sun|Dinner|   4|       0.0|       0.0|
|     15.04|1.96|  Male|    No|Sun|Dinner|   2|       0.0|       0.0|
|     14.78|3.23|  Male|    No|Sun|Dinner|   2|       0.0|       0.0|
|     10.27|1.71|  Male|    No|Sun|Dinner|   2|       0.0|       0.0|
|     35.26| 5.0|Female|    No|Sun|Dinner|   4|       1.0|       0.0|
|     15.42|1.57|  Male|    No|Sun|Dinner|   2|       0.0|       0.0|
|     18.43| 3.0|  Male|    No|Sun|Dinner|   4|       0.0|       0.0|
|     14.83|3.02|Female|    No|Sun|Dinner|   2|       1.0|       0.0|
|     21.58|3.92|  Male|    No|Sun|Dinner|   2|       0.0|       0.0|
|     10.33|1.67|Female|    No|Sun|Dinner|   3|       1.0|       0.0|
|     16.29|3.71|  Male|    No|Sun|Dinner|   3|       0.0|       0.0|
|     16.97| 3.5|Female|    No|Sun|Dinner|   3|       1.0|       0.0|
|     20.65|3.35|  Male|    No|Sat|Dinner|   3|       0.0|       0.0|
+----------+----+------+------+---+------+----+----------+----------+
only showing top 20 rows
```

In [18]:
```
indexer=StringIndexer(inputCol='day',outputCol='Day_Col')
df3=indexer.fit(df2).transform(df2)
df3.show()
```

```
+----------+----+------+------+---+------+----+----------+----------+-------+
|total_bill| tip|   sex|smoker|day|  time|size|Gender_Col|Smoker_Col|Day_Col|
+----------+----+------+------+---+------+----+----------+----------+-------+
|     16.99|1.01|Female|    No|Sun|Dinner|   2|       1.0|       0.0|    1.0|
|     10.34|1.66|  Male|    No|Sun|Dinner|   3|       0.0|       0.0|    1.0|
|     21.01|3.5 |  Male|    No|Sun|Dinner|   3|       0.0|       0.0|    1.0|
|     23.68|3.31|  Male|    No|Sun|Dinner|   2|       0.0|       0.0|    1.0|
|     24.59|3.61|Female|    No|Sun|Dinner|   4|       1.0|       0.0|    1.0|
|     25.29|4.71|  Male|    No|Sun|Dinner|   4|       0.0|       0.0|    1.0|
|      8.77|2.01|  Male|    No|Sun|Dinner|   2|       0.0|       0.0|    1.0|
|     26.88|3.12|  Male|    No|Sun|Dinner|   4|       0.0|       0.0|    1.0|
|     15.04|1.96|  Male|    No|Sun|Dinner|   2|       0.0|       0.0|    1.0|
|     14.78|3.23|  Male|    No|Sun|Dinner|   2|       0.0|       0.0|    1.0|
|     10.27|1.71|  Male|    No|Sun|Dinner|   2|       0.0|       0.0|    1.0|
|     35.26| 5.0|Female|    No|Sun|Dinner|   4|       1.0|       0.0|    1.0|
|     15.42|1.57|  Male|    No|Sun|Dinner|   2|       0.0|       0.0|    1.0|
|     18.43| 3.0|  Male|    No|Sun|Dinner|   4|       0.0|       0.0|    1.0|
|     14.83|3.02|Female|    No|Sun|Dinner|   2|       1.0|       0.0|    1.0|
|     21.58|3.92|  Male|    No|Sun|Dinner|   2|       0.0|       0.0|    1.0|
|     10.33|1.67|Female|    No|Sun|Dinner|   3|       1.0|       0.0|    1.0|
|     16.29|3.71|  Male|    No|Sun|Dinner|   3|       0.0|       0.0|    1.0|
|     16.97| 3.5|Female|    No|Sun|Dinner|   3|       1.0|       0.0|    1.0|
|     20.65|3.35|  Male|    No|Sat|Dinner|   3|       0.0|       0.0|    0.0|
+----------+----+------+------+---+------+----+----------+----------+-------+
only showing top 20 rows
```

In [19]:
```
indexer=StringIndexer(inputCol='time',outputCol='Time_Col')
df3=indexer.fit(df3).transform(df3)
df3.show()
```

```
+----------+----+------+------+---+------+----+----------+----------+-------+--------+
|total_bill| tip|   sex|smoker|day|  time|size|Gender_Col|Smoker_Col|Day_Col|Time_Col|
+----------+----+------+------+---+------+----+----------+----------+-------+--------+
|     16.99|1.01|Female|    No|Sun|Dinner|   2|       1.0|       0.0|    1.0|     0.0|
|     10.34|1.66|  Male|    No|Sun|Dinner|   3|       0.0|       0.0|    1.0|     0.0|
|     21.01|3.5 |  Male|    No|Sun|Dinner|   3|       0.0|       0.0|    1.0|     0.0|
|     23.68|3.31|  Male|    No|Sun|Dinner|   2|       0.0|       0.0|    1.0|     0.0|
|     24.59|3.61|Female|    No|Sun|Dinner|   4|       1.0|       0.0|    1.0|     0.0|
|     25.29|4.71|  Male|    No|Sun|Dinner|   4|       0.0|       0.0|    1.0|     0.0|
|      8.77|2.01|  Male|    No|Sun|Dinner|   2|       0.0|       0.0|    1.0|     0.0|
|     26.88|3.12|  Male|    No|Sun|Dinner|   4|       0.0|       0.0|    1.0|     0.0|
|     15.04|1.96|  Male|    No|Sun|Dinner|   2|       0.0|       0.0|    1.0|     0.0|
|     14.78|3.23|  Male|    No|Sun|Dinner|   2|       0.0|       0.0|    1.0|     0.0|
|     10.27|1.71|  Male|    No|Sun|Dinner|   2|       0.0|       0.0|    1.0|     0.0|
|     35.26| 5.0|Female|    No|Sun|Dinner|   4|       1.0|       0.0|    1.0|     0.0|
|     15.42|1.57|  Male|    No|Sun|Dinner|   2|       0.0|       0.0|    1.0|     0.0|
|     18.43| 3.0|  Male|    No|Sun|Dinner|   4|       0.0|       0.0|    1.0|     0.0|
|     14.83|3.02|Female|    No|Sun|Dinner|   2|       1.0|       0.0|    1.0|     0.0|
|     21.58|3.92|  Male|    No|Sun|Dinner|   2|       0.0|       0.0|    1.0|     0.0|
|     10.33|1.67|Female|    No|Sun|Dinner|   3|       1.0|       0.0|    1.0|     0.0|
|     16.29|3.71|  Male|    No|Sun|Dinner|   3|       0.0|       0.0|    1.0|     0.0|
|     16.97| 3.5|Female|    No|Sun|Dinner|   3|       1.0|       0.0|    1.0|     0.0|
|     20.65|3.35|  Male|    No|Sat|Dinner|   3|       0.0|       0.0|    0.0|     0.0|
+----------+----+------+------+---+------+----+----------+----------+-------+--------+
only showing top 20 rows
```

In [21]:
```
featureAss=VectorAssembler(inputCols=['tip','Gender_Col','Smoker_Col','Day_Col','Time_Col'],outputCol='inputfeatures')
outputframe=featureAss.transform(df3)
outputframe.show(5,False)
```

```
+----------+----+------+------+---+------+----+----------+----------+-------+--------+-----------------------+
|total_bill|tip |sex   |smoker|day|time  |size|Gender_Col|Smoker_Col|Day_Col|Time_Col|inputfeatures          |
+----------+----+------+------+---+------+----+----------+----------+-------+--------+-----------------------+
|16.99     |1.01|Female|No    |Sun|Dinner|2   |1.0       |0.0       |1.0    |0.0     |[1.01,1.0,0.0,1.0,0.0] |
|10.34     |1.66|Male  |No    |Sun|Dinner|3   |0.0       |0.0       |1.0    |0.0     |[5,[0,3],[1.66,1.0]]   |
|21.01     |3.5 |Male  |No    |Sun|Dinner|3   |0.0       |0.0       |1.0    |0.0     |[5,[0,3],[3.5,1.0]]    |
|23.68     |3.31|Male  |No    |Sun|Dinner|2   |0.0       |0.0       |1.0    |0.0     |[5,[0,3],[3.31,1.0]]   |
|24.59     |3.61|Female|No    |Sun|Dinner|4   |1.0       |0.0       |1.0    |0.0     |[3.61,1.0,0.0,1.0,0.0] |
+----------+----+------+------+---+------+----+----------+----------+-------+--------+-----------------------+
only showing top 5 rows
```

In [22]:
```
finaldata=outputframe.select('inputfeatures','total_bill')
finaldata.show(5,False)
```

```
+-----------------------+----------+
|inputfeatures          |total_bill|
+-----------------------+----------+
|[1.01,1.0,0.0,1.0,0.0] |16.99     |
|(5,[0,3],[1.66,1.0])   |10.34     |
|(5,[0,3],[3.5,1.0])    |21.01     |
|(5,[0,3],[3.31,1.0])   |23.68     |
|[3.61,1.0,0.0,1.0,0.0] |24.59     |
+-----------------------+----------+
only showing top 5 rows
```

In [23]:
```
finaldata.printSchema()
```

```
root
 |-- inputfeatures: vector (nullable = true)
 |-- total_bill: double (nullable = true)
```

In [24]:
```
train,test=finaldata.randomSplit([.70,.30])
```

In [25]:
```
from pyspark.ml.regression import LinearRegression
lr=LinearRegression(featuresCol='inputfeatures',labelCol='total_bill')
lrt=lr.fit(train)
```

In [26]:
```
print(lrt.coefficients)
print(lrt.intercept)
```

```
[4.209948833744898,-0.9914012916996 74,0.12930420148679753,-0.11589705579207694,-1.8480744357668 57]
0.08244403593093 2
```

In [27]:
```
res=lrt.evaluate(test)
res.predictions.show(5,False)
```

```
C:\Users\user\anaconda3\lib\site-packages\pyspark\sql\context.py:125: FutureWarning: Deprecated in 3.0.0. Use SparkSession.builder.getOrCreate() instead.
  warnings.warn(
+---------------+----------+------------------+
|inputfeatures  |total_bill|prediction        |
+---------------+----------+------------------+
|(5,[0],[1.25])|10.07     |13.343755078112054|
|(5,[0],[1.97])|12.02     |16.37427023840838 |
|(5,[0],[2.0]) |12.69     |16.500541703420273|
|(5,[0],[2.01])|20.23     |16.542632191758177|
|(5,[0],[2.24])|16.04     |17.510713423519505|
+---------------+----------+------------------+
only showing top 5 rows
```

In [28]:
```
print("R2",res.r2)
print("Mean Absolute Error is",res.meanAbsoluteError)
print("Root Mean Square Error(RMSE) is",res.rootMeanSquaredError)
```

```
R2 0.426611616588921
Mean Absolute Error is 5.269890426575595
Root Mean Square Error(RMSE) is 7.048244434152945
```

In [ ]:
```
```