Initiate a spark Session & Import Data ¶

In [1]:
```
import findspark
findspark.init('/home/ubuntu/spark-2.4.5-bin-hadoop2.7')
import pyspark
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName('lr_example').getOrCreate()
```

In [2]:
```
df = spark.read.csv('cruise_ship_info.csv', header=True, inferSchema=True)
```

In [3]:
```
df.show()
```

```
+----------+----------+---+-----------------+----------+------+------+--
---------------+----+
|  Ship_name|Cruise_line|Age|          Tonnage|passengers|length|cabins|pa
ssenger_density|crew|
+----------+----------+---+-----------------+----------+------+------+--
---------------+----+
|    Journey|    Azamara|  6|30.276999999999997|      6.94|  5.94|  3.55|
42.64|3.55|
|      Quest|    Azamara|  6|30.276999999999997|      6.94|  5.94|  3.55|
42.64|3.55|
|Celebration|   Carnival| 26|           47.262|     14.86|  7.22|  7.43|
31.8| 6.7|
|   Conquest|   Carnival| 11|            110.0|     29.74|  9.53| 14.88|
36.99|19.1|
|    Destiny|   Carnival| 17|          101.353|     26.42|  8.92| 13.21|
38.36|10.0|
|    Ecstasy|   Carnival| 22|           70.367|     20.52|  8.55|  10.2|
34.29| 9.2|
|    Elation|   Carnival| 15|           70.367|     20.52|  8.55|  10.2|
34.29| 9.2|
|    Fantasy|   Carnival| 23|           70.367|     20.56|  8.55| 10.22|
34.23| 9.2|
|Fascination|   Carnival| 19|           70.367|     20.52|  8.55|  10.2|
34.29| 9.2|
|    Freedom|   Carnival|  6|110.23899999999999|      37.0|  9.51| 14.87|
29.79|11.5|
|      Glory|   Carnival| 10|            110.0|     29.74|  9.51| 14.87|
36.99|11.6|
|    Holiday|   Carnival| 28|           46.052|     14.52|  7.27|  7.26|
31.72| 6.6|
|Imagination|   Carnival| 18|           70.367|     20.52|  8.55|  10.2|
34.29| 9.2|
|Inspiration|   Carnival| 17|           70.367|     20.52|  8.55|  10.2|
34.29| 9.2|
|     Legend|   Carnival| 11|             86.0|     21.24|  9.63| 10.62|
40.49| 9.3|
|   Liberty*|   Carnival|  8|            110.0|     29.74|  9.51| 14.87|
36.99|11.6|
|    Miracle|   Carnival|  9|             88.5|     21.24|  9.63| 10.62|
41.67|10.3|
|   Paradise|   Carnival| 15|           70.367|     20.52|  8.55|  10.2|
34.29| 9.2|
|      Pride|   Carnival| 12|             88.5|     21.24|  9.63| 11.62|
41.67| 9.3|
|  Sensation|   Carnival| 20|           70.367|     20.52|  8.55|  10.2|
34.29| 9.2|
+----------+----------+---+-----------------+----------+------+------+--
---------------+----+
only showing top 20 rows
```

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

```
root
 |-- Ship_name: string (nullable = true)
 |-- Cruise_line: string (nullable = true)
 |-- Age: integer (nullable = true)
 |-- Tonnage: double (nullable = true)
 |-- passengers: double (nullable = true)
 |-- length: double (nullable = true)
 |-- cabins: double (nullable = true)
 |-- passenger_density: double (nullable = true)
 |-- crew: double (nullable = true)
```

In [5]:
```
df.describe('Age',
 'Tonnage',
 'passengers',
 'length').show()
```

```
+-------+-----------------+-----------------+-----------------+-----------------+
|summary|              Age|          Tonnage|       passengers|           length|
+-------+-----------------+-----------------+-----------------+-----------------+
|  count|              158|              158|              158|              158|
|   mean|15.689873417721518| 71.28467088607599|18.45740506329114|8.130632911392404|
| stddev| 7.615691058751413|37.229540025907866|9.677094775143416|1.793473548054825|
|    min|                4|            2.329|             0.66|             2.79|
|    max|               48|            220.0|             54.0|            11.82|
+-------+-----------------+-----------------+-----------------+-----------------+
```

In [6]:
```
df.describe('cabins',
 'passenger_density',
 'crew').show()
```

```
+-------+-----------------+-----------------+-----------------+
|summary|           cabins|passenger_density|             crew|
+-------+-----------------+-----------------+-----------------+
|  count|              158|              158|              158|
|   mean| 8.830000000000005|39.90094936708861|7.794177215189873|
| stddev|4.4714172221480615| 8.63921711391542|3.503486564627034|
|    min|             0.33|             17.7|             0.59|
|    max|             27.0|            71.43|             21.0|
+-------+-----------------+-----------------+-----------------+
```

Feature Engineering

In [7]:
```
from pyspark.ml.feature import StringIndexer, VectorAssembler, OneHotEncoderE
```

In [8]:
```
indexer = StringIndexer(inputCol='Cruise_line', outputCol='cruise_line_index'
```

In [9]:
```
df = indexer.fit(df).transform(df)
```

In [10]:
```
df.show()
```

```
+----------+----------+---+------------------+----------+------+------+--
---------------+----+-----------------+
|  Ship_name|Cruise_line|Age|           Tonnage|passengers|length|cabins|pa
ssenger_density|crew|cruise_line_index|
+----------+----------+---+------------------+----------+------+------+--
---------------+----+-----------------+
|    Journey|    Azamara|  6|30.276999999999997|      6.94|  5.94|  3.55|
42.64|3.55|             16.0|
|      Quest|    Azamara|  6|30.276999999999997|      6.94|  5.94|  3.55|
42.64|3.55|             16.0|
|Celebration|   Carnival| 26|            47.262|     14.86|  7.22|  7.43|
31.8| 6.7|              1.0|
|   Conquest|   Carnival| 11|             110.0|     29.74|  9.53| 14.88|
36.99|19.1|              1.0|
|    Destiny|   Carnival| 17|           101.353|     26.42|  8.92| 13.21|
38.36|10.0|              1.0|
|    Ecstasy|   Carnival| 22|            70.367|     20.52|  8.55|  10.2|
34.29| 9.2|              1.0|
|    Elation|   Carnival| 15|            70.367|     20.52|  8.55|  10.2|
34.29| 9.2|              1.0|
|    Fantasy|   Carnival| 23|            70.367|     20.56|  8.55| 10.22|
34.23| 9.2|              1.0|
|Fascination|   Carnival| 19|            70.367|     20.52|  8.55|  10.2|
34.29| 9.2|              1.0|
|    Freedom|   Carnival|  6|110.23899999999999|      37.0|  9.51| 14.87|
29.79|11.5|              1.0|
|      Glory|   Carnival| 10|             110.0|     29.74|  9.51| 14.87|
36.99|11.6|              1.0|
|    Holiday|   Carnival| 28|            46.052|     14.52|  7.27|  7.26|
31.72| 6.6|              1.0|
|Imagination|   Carnival| 18|            70.367|     20.52|  8.55|  10.2|
34.29| 9.2|              1.0|
|Inspiration|   Carnival| 17|            70.367|     20.52|  8.55|  10.2|
34.29| 9.2|              1.0|
|     Legend|   Carnival| 11|              86.0|     21.24|  9.63| 10.62|
40.49| 9.3|              1.0|
|   Liberty*|   Carnival|  8|             110.0|     29.74|  9.51| 14.87|
36.99|11.6|              1.0|
|    Miracle|   Carnival|  9|              88.5|     21.24|  9.63| 10.62|
41.67|10.3|              1.0|
|   Paradise|   Carnival| 15|            70.367|     20.52|  8.55|  10.2|
34.29| 9.2|              1.0|
|      Pride|   Carnival| 12|              88.5|     21.24|  9.63| 11.62|
41.67| 9.3|              1.0|
|  Sensation|   Carnival| 20|            70.367|     20.52|  8.55|  10.2|
34.29| 9.2|              1.0|
+----------+----------+---+------------------+----------+------+------+--
---------------+----+-----------------+
only showing top 20 rows
```

In [11]:
```
encoder = OneHotEncoderEstimator(inputCols=['cruise_line_index'], outputCols=
```

In [12]:
```
df = encoder.fit(df).transform(df)
```

In [13]:
```
df.show()
```

```
+----------+----------+---+-----------------+----------+------+------+--
---------------+----+-----------------+---------------+
|  Ship_name|Cruise_line|Age|          Tonnage|passengers|length|cabins|pa
ssenger_density|crew|cruise_line_index| cruise_encoded|
+----------+----------+---+-----------------+----------+------+------+--
---------------+----+-----------------+---------------+
|    Journey|    Azamara|  6|30.276999999999997|      6.94|  5.94|  3.55|
42.64|3.55|             16.0|(19,[16],[1.0])|
|      Quest|    Azamara|  6|30.276999999999997|      6.94|  5.94|  3.55|
42.64|3.55|             16.0|(19,[16],[1.0])|
|Celebration|   Carnival| 26|           47.262|     14.86|  7.22|  7.43|
31.8| 6.7|              1.0| (19,[1],[1.0])|
|   Conquest|   Carnival| 11|            110.0|     29.74|  9.53| 14.88|
36.99|19.1|              1.0| (19,[1],[1.0])|
|    Destiny|   Carnival| 17|          101.353|     26.42|  8.92| 13.21|
38.36|10.0|              1.0| (19,[1],[1.0])|
|    Ecstasy|   Carnival| 22|           70.367|     20.52|  8.55|  10.2|
34.29| 9.2|              1.0| (19,[1],[1.0])|
|    Elation|   Carnival| 15|           70.367|     20.52|  8.55|  10.2|
34.29| 9.2|              1.0| (19,[1],[1.0])|
|    Fantasy|   Carnival| 23|           70.367|     20.56|  8.55| 10.22|
34.23| 9.2|              1.0| (19,[1],[1.0])|
|Fascination|   Carnival| 19|           70.367|     20.52|  8.55|  10.2|
34.29| 9.2|              1.0| (19,[1],[1.0])|
|    Freedom|   Carnival|  6|110.23899999999999|      37.0|  9.51| 14.87|
29.79|11.5|              1.0| (19,[1],[1.0])|
|      Glory|   Carnival| 10|            110.0|     29.74|  9.51| 14.87|
36.99|11.6|              1.0| (19,[1],[1.0])|
|    Holiday|   Carnival| 28|           46.052|     14.52|  7.27|  7.26|
31.72| 6.6|              1.0| (19,[1],[1.0])|
|Imagination|   Carnival| 18|           70.367|     20.52|  8.55|  10.2|
34.29| 9.2|              1.0| (19,[1],[1.0])|
|Inspiration|   Carnival| 17|           70.367|     20.52|  8.55|  10.2|
34.29| 9.2|              1.0| (19,[1],[1.0])|
|     Legend|   Carnival| 11|             86.0|     21.24|  9.63| 10.62|
40.49| 9.3|              1.0| (19,[1],[1.0])|
|   Liberty*|   Carnival|  8|            110.0|     29.74|  9.51| 14.87|
36.99|11.6|              1.0| (19,[1],[1.0])|
|    Miracle|   Carnival|  9|             88.5|     21.24|  9.63| 10.62|
41.67|10.3|              1.0| (19,[1],[1.0])|
|   Paradise|   Carnival| 15|           70.367|     20.52|  8.55|  10.2|
34.29| 9.2|              1.0| (19,[1],[1.0])|
|      Pride|   Carnival| 12|             88.5|     21.24|  9.63| 11.62|
41.67| 9.3|              1.0| (19,[1],[1.0])|
|  Sensation|   Carnival| 20|           70.367|     20.52|  8.55|  10.2|
34.29| 9.2|              1.0| (19,[1],[1.0])|
+----------+----------+---+-----------------+----------+------+------+--
---------------+----+-----------------+---------------+
only showing top 20 rows
```

In [14]:
```
df.columns
```

Out[14]:
```
['Ship_name',
 'Cruise_line',
 'Age',
 'Tonnage',
 'passengers',
 'length',
 'cabins',
 'passenger_density',
 'crew',
 'cruise_line_index',
 'cruise_encoded']
```

In [15]:
```
assembler = VectorAssembler(inputCols=['Age','Tonnage','passengers','length',
                            outputCol='features')
```

In [16]:
```
df = assembler.transform(df)
```

In [17]:
```
df.show()
```

```
+----------+----------+---+-----------------+---------+------+------+--
--------------+----+----------------+--------------+-------------------
+
|  Ship_name|Cruise_line|Age|          Tonnage|passengers|length|cabins|pa
ssenger_density|crew|cruise_line_index| cruise_encoded|            features
|
+----------+----------+---+-----------------+---------+------+------+--
--------------+----+----------------+--------------+-------------------
+
|    Journey|    Azamara|  6|30.276999999999997|      6.94|  5.94|  3.55|
42.64|3.55|            16.0|(19,[16],[1.0])|(25,[0,1,2,3,4,5,...|
|      Quest|    Azamara|  6|30.276999999999997|      6.94|  5.94|  3.55|
42.64|3.55|            16.0|(19,[16],[1.0])|(25,[0,1,2,3,4,5,...|
|Celebration|   Carnival| 26|           47.262|     14.86|  7.22|  7.43|
31.8| 6.7|             1.0| (19,[1],[1.0])|(25,[0,1,2,3,4,5,...|
|   Conquest|   Carnival| 11|            110.0|     29.74|  9.53| 14.88|
36.99|19.1|             1.0| (19,[1],[1.0])|(25,[0,1,2,3,4,5,...|
|    Destiny|   Carnival| 17|          101.353|     26.42|  8.92| 13.21|
38.36|10.0|             1.0| (19,[1],[1.0])|(25,[0,1,2,3,4,5,...|
|    Ecstasy|   Carnival| 22|           70.367|     20.52|  8.55|  10.2|
34.29| 9.2|             1.0| (19,[1],[1.0])|(25,[0,1,2,3,4,5,...|
|    Elation|   Carnival| 15|           70.367|     20.52|  8.55|  10.2|
34.29| 9.2|             1.0| (19,[1],[1.0])|(25,[0,1,2,3,4,5,...|
|    Fantasy|   Carnival| 23|           70.367|     20.56|  8.55| 10.22|
34.23| 9.2|             1.0| (19,[1],[1.0])|(25,[0,1,2,3,4,5,...|
|Fascination|   Carnival| 19|           70.367|     20.52|  8.55|  10.2|
34.29| 9.2|             1.0| (19,[1],[1.0])|(25,[0,1,2,3,4,5,...|
|    Freedom|   Carnival|  6|110.23899999999999|      37.0|  9.51| 14.87|
29.79|11.5|             1.0| (19,[1],[1.0])|(25,[0,1,2,3,4,5,...|
|      Glory|   Carnival| 10|            110.0|     29.74|  9.51| 14.87|
36.99|11.6|             1.0| (19,[1],[1.0])|(25,[0,1,2,3,4,5,...|
|    Holiday|   Carnival| 28|           46.052|     14.52|  7.27|  7.26|
31.72| 6.6|             1.0| (19,[1],[1.0])|(25,[0,1,2,3,4,5,...|
|Imagination|   Carnival| 18|           70.367|     20.52|  8.55|  10.2|
34.29| 9.2|             1.0| (19,[1],[1.0])|(25,[0,1,2,3,4,5,...|
|Inspiration|   Carnival| 17|           70.367|     20.52|  8.55|  10.2|
34.29| 9.2|             1.0| (19,[1],[1.0])|(25,[0,1,2,3,4,5,...|
|     Legend|   Carnival| 11|             86.0|     21.24|  9.63| 10.62|
40.49| 9.3|             1.0| (19,[1],[1.0])|(25,[0,1,2,3,4,5,...|
|   Liberty*|   Carnival|  8|            110.0|     29.74|  9.51| 14.87|
36.99|11.6|             1.0| (19,[1],[1.0])|(25,[0,1,2,3,4,5,...|
|    Miracle|   Carnival|  9|             88.5|     21.24|  9.63| 10.62|
41.67|10.3|             1.0| (19,[1],[1.0])|(25,[0,1,2,3,4,5,...|
|   Paradise|   Carnival| 15|           70.367|     20.52|  8.55|  10.2|
34.29| 9.2|             1.0| (19,[1],[1.0])|(25,[0,1,2,3,4,5,...|
|      Pride|   Carnival| 12|             88.5|     21.24|  9.63| 11.62|
41.67| 9.3|             1.0| (19,[1],[1.0])|(25,[0,1,2,3,4,5,...|
|  Sensation|   Carnival| 20|           70.367|     20.52|  8.55|  10.2|
34.29| 9.2|             1.0| (19,[1],[1.0])|(25,[0,1,2,3,4,5,...|
+----------+----------+---+-----------------+---------+------+------+--
--------------+----+----------------+--------------+-------------------
+
only showing top 20 rows
```

In [18]:
```
df.select('features').head(1)
```

Out[18]:
```
[Row(features=SparseVector(25, {0: 6.0, 1: 30.277, 2: 6.94, 3: 5.94, 4: 3.55, 5: 42.64, 22: 1.0}))]
```

In [19]:
```
df_final = df.select('features', 'crew')
```

In [20]:
```
train_data, test_data = df_final.randomSplit([0.7,0.3])
```

Model Building & Evaluation

In [21]:
```
from pyspark.ml.regression import LinearRegression
```

In [22]:
```
lr = LinearRegression(featuresCol='features', labelCol='crew')
```

In [23]:
```
lr_model = lr.fit(train_data)
```

In [24]:
```
model_result = lr_model.evaluate(test_data)
```

In [25]:
```
model_result.meanAbsoluteError
```

Out[25]:
```
0.7095596157752411
```

In [26]:
```
model_result.meanSquaredError
```

Out[26]:
```
1.2341143531978442
```

In [27]:
```
model_result.rootMeanSquaredError
```

Out[27]:
```
1.1109069957461986
```

In [28]:
```
model_result.r2
```

Out[28]:
```
0.8787118487653475
```

In [29]:
```
test_data_pred = test_data.select('features')
```

In [30]:
```
test_pred = lr_model.transform(test_data_pred)
```

In [31]:
```
test_pred.show()
```

```
+--------------------+------------------+
|            features|        prediction|
+--------------------+------------------+
|(25,[0,1,2,3,4,5,...|13.388820224936202|
|(25,[0,1,2,3,4,5,...|12.297736354437026|
|(25,[0,1,2,3,4,5,...| 7.625286172766864|
|(25,[0,1,2,3,4,5,...| 7.607497912342315|
|(25,[0,1,2,3,4,5,...| 7.224343652098771|
|(25,[0,1,2,3,4,5,...| 8.486678479003468|
|(25,[0,1,2,3,4,5,...|13.013026295725084|
|(25,[0,1,2,3,4,5,...|12.977449774875984|
|(25,[0,1,2,3,4,5,...|11.011138762687219|
|(25,[0,1,2,3,4,5,...|12.070053525326832|
|(25,[0,1,2,3,4,5,...|11.584630651943742|
|(25,[0,1,2,3,4,5,...| 9.498999372646576|
|(25,[0,1,2,3,4,5,...| 9.445634591372928|
|(25,[0,1,2,3,4,5,...| 9.427846330948379|
|(25,[0,1,2,3,4,5,...| 7.026741480431232|
|(25,[0,1,2,3,4,5,...| 8.718511040769252|
|(25,[0,1,2,3,4,5,...|11.420015657107275|
|(25,[0,1,2,3,4,5,...| 8.540628436523761|
|(25,[0,1,2,3,4,5,...| 7.181143826984111|
|(25,[0,1,2,3,4,5,...| 8.273381885091037|
+--------------------+------------------+
only showing top 20 rows
```

In [32]:
```
test_data.select('crew').show()
```

```
+-----+
| crew|
+-----+
| 13.6|
|11.85|
|  6.6|
| 7.65|
|  7.2|
| 8.22|
| 11.6|
| 11.6|
|  9.3|
| 11.5|
| 10.0|
|  9.2|
|  9.2|
|  9.2|
|  6.6|
|  9.0|
|12.38|
|  9.0|
| 6.96|
| 8.42|
+-----+
only showing top 20 rows
```

Thank You!