

```
!pip install pyspark
```

```
Requirement already satisfied: pyspark in /usr/local/lib/python3.7/dist-packages (3.2.0)
Requirement already satisfied: py4j==0.10.9.2 in /usr/local/lib/python3.7/dist-packages (0.10.9.2)
```

```
# Import SparkSession
from pyspark.sql import SparkSession
```

```
# create a object SparkSession
spark = SparkSession.builder.appName('Kmeans_App').getOrCreate()
```

```
#Load Data
data1 = spark.read.option('header','True').csv('/content/pubg.csv')

data2 = spark.read.option('header','True').csv('/content/Responses.csv')
```

```
a = SparkSession.builder
```

```
type(a)
```

```
pyspark.sql.session.SparkSession.Builder
```

```
data1.show()
```

Id	groupId	matchId	assists	boosts	damageDealt	DBNOs	headshots
2f262dd9795e60	78437bcd91d40e	d5db3a49eb2955	0	0	0	0	
a32847cf5bf34b	85b7ce5a12e10b	65223f05c7fdb4	0	0	163.2	1	
1b1900a9990396	edf80d6523380a	1cadec4534f30a	0	3	278.7	2	
f589dd03b60bf2	804ab5e5585558	c4a5676dc91604	0	0	191.9	1	
c23c4cc5b78b35	b3e2cd169ed920	cd595700a01bfa	0	0	100	1	
fd034582dd4d2e	9b8930ae086a	6f6e52b15ddf21	0	1	200	2	
c60b5633f4dcc8	7c0f817f6627c7	3232c1e0fec04b	0	3	638.2	4	
f0ba8246b6980f	7318b5204462cb	112e9711f86001	0	0	27.94	0	
79c5d5eda1c72e	a85b81198dfc06	ef5fc25e28ffb1	1	4	275.8	3	
94834a28e52abd	bc513cde35fa54	f36a754a9b88f7	1	1	530.4	4	
f051dcc9b0b3ce	d203c0e3d8c321	89a6a8738190b4	0	0	20.59	0	
f02c2f34accf08	22ed911205c815	559dac9580b92a	1	0	62.72	0	
6701c06774d409	cdb79f944d585b	9f3dec5ffba4e	0	0	0	0	
4e4aef4ae5f5	a9dfa1c736c889	ac92da38bb19ad	0	2	13.83	0	
d26b4b75c5229d	130ea20c924e8c	14fb1c1b26e9a4	0	0	25.8	0	
c5473a410326a8	8a25860cd71a23	88cffe1ae97aff	1	1	594	2	
321fe9f3c71131	cb3471586d99b4	1cf664f7c75122	0	0	50.31	0	
f8933f3ee2e431	114b20e9d7504b	d5fcb7a3981d33	0	0	0	0	
c70c7337cd46b4	73870d831717aa	e6602141e44281	0	0	25.8	0	
d6c231133b5d57	928733f3037f92	b4baee11351ae6	0	0	30.96	0	

```
only showing top 20 rows
```



```
data2.tail(5)
```

```
[Row(Age='25', Gender='Male', Do you play PUBG game='No', How long have you been play
Row(Age='24', Gender='Female', Do you play PUBG game='No', How long have you been p]
Row(Age='24', Gender='Female', Do you play PUBG game='No', How long have you been p]
Row(Age='22', Gender='Male', Do you play PUBG game='No', How long have you been play
Row(Age='23', Gender='Female', Do you play PUBG game='No', How long have you been p]
```

```
data1.printSchema()
```

```
root
|-- Id: string (nullable = true)
|-- groupId: string (nullable = true)
|-- matchId: string (nullable = true)
|-- assists: string (nullable = true)
|-- boosts: string (nullable = true)
|-- damageDealt: string (nullable = true)
|-- DBNOs: string (nullable = true)
|-- headshotKills: string (nullable = true)
|-- heals: string (nullable = true)
|-- killPlace: string (nullable = true)
|-- killPoints: string (nullable = true)
|-- kills: string (nullable = true)
|-- killStreaks: string (nullable = true)
|-- longestKill: string (nullable = true)
|-- matchDuration: string (nullable = true)
|-- matchType: string (nullable = true)
|-- maxPlace: string (nullable = true)
|-- numGroups: string (nullable = true)
|-- rankPoints: string (nullable = true)
|-- revives: string (nullable = true)
|-- rideDistance: string (nullable = true)
|-- roadKills: string (nullable = true)
|-- swimDistance: string (nullable = true)
|-- teamKills: string (nullable = true)
|-- vehicleDestroys: string (nullable = true)
|-- walkDistance: string (nullable = true)
|-- weaponsAcquired: string (nullable = true)
|-- winPoints: string (nullable = true)
|-- winPlacePerc: string (nullable = true)
```

```
data2.printSchema()
```

```
root
|-- Age: string (nullable = true)
|-- Gender: string (nullable = true)
|-- Do you play PUBG game: string (nullable = true)
|-- How long have you been playing this game: string (nullable = true)
|-- How often do you play this Game: string (nullable = true)
|-- How much time you spent daily: string (nullable = true)
|-- How affect this game on students6: string (nullable = true)
|-- Positive effects of playing PUBG: string (nullable = true)
|-- Negative effects of playing PUBG: string (nullable = true)
|-- What are reasons that you dont play PUBG: string (nullable = true)
```

```
|-- How affect this game on students10: string (nullable = true)
|-- According to you are there positive effects of playing PUBG: string (nullable =
```

```
data1 = data1.withColumn('Id',data1.Id.astype('string'))\
.withColumn('groupId',data1.groupId.astype('string'))\
.withColumn('matchId',data1.matchId.astype('string'))\
.withColumn('assists',data1.assists.astype('int'))\
.withColumn('boosts',data1.boosts.astype('int'))\
.withColumn('damageDealt',data1.damageDealt.astype('float'))\
.withColumn('DBNOs',data1.DBNOs.astype('int'))\
.withColumn('headshotKills',data1.headshotKills.astype('int'))\
.withColumn('heals',data1.heals.astype('int'))\
.withColumn('killPlace',data1.killPlace.astype('int'))\
.withColumn('killPoints',data1.killPoints.astype('int'))\
.withColumn('kills',data1.kills.astype('int'))\
.withColumn('killStreaks',data1.killStreaks.astype('int'))\
.withColumn('longestKill',data1.longestKill.astype('float'))\
.withColumn('matchDuration',data1.matchDuration.astype('int'))\
.withColumn('matchType',data1.matchType.astype('string'))\
.withColumn('maxPlace',data1.maxPlace.astype('int'))\
.withColumn('numGroups',data1.numGroups.astype('int'))\
.withColumn('rankPoints',data1.rankPoints.astype('int'))\
.withColumn('revives',data1.revives.astype('int'))\
.withColumn('rideDistance',data1.rideDistance.astype('float'))\
.withColumn('roadKills',data1.roadKills.astype('int'))\
.withColumn('swimDistance',data1.swimDistance.astype('int'))\
.withColumn('teamKills',data1.teamKills.astype('int'))\
.withColumn('vehicleDestroys',data1.vehicleDestroys.astype('int'))\
.withColumn('walkDistance',data1.walkDistance.astype('float'))\
.withColumn('weaponsAcquired',data1.weaponsAcquired.astype('int'))\
.withColumn('winPoints',data1.winPoints.astype('int'))\
.withColumn('winPlacePerc',data1.winPlacePerc.astype('float'))
```

```
data1.printSchema()
```

```
root
|-- Id: string (nullable = true)
|-- groupId: string (nullable = true)
|-- matchId: string (nullable = true)
|-- assists: integer (nullable = true)
|-- boosts: integer (nullable = true)
|-- damageDealt: float (nullable = true)
|-- DBNOs: integer (nullable = true)
|-- headshotKills: integer (nullable = true)
|-- heals: integer (nullable = true)
|-- killPlace: integer (nullable = true)
|-- killPoints: integer (nullable = true)
|-- kills: integer (nullable = true)
|-- killStreaks: integer (nullable = true)
```

```
|-- longestKill: float (nullable = true)
|-- matchDuration: integer (nullable = true)
|-- matchType: string (nullable = true)
|-- maxPlace: integer (nullable = true)
|-- numGroups: integer (nullable = true)
|-- rankPoints: integer (nullable = true)
|-- revives: integer (nullable = true)
|-- rideDistance: float (nullable = true)
|-- roadKills: integer (nullable = true)
|-- swimDistance: integer (nullable = true)
|-- teamKills: integer (nullable = true)
|-- vehicleDestroys: integer (nullable = true)
|-- walkDistance: float (nullable = true)
|-- weaponsAcquired: integer (nullable = true)
|-- winPoints: integer (nullable = true)
|-- winPlacePerc: float (nullable = true)
```

```
data2 = data2.withColumn('Age',data2.Age.astype('int'))
```

```
data2.printSchema()
```

```
root
|-- Age: integer (nullable = true)
|-- Gender: string (nullable = true)
|-- Do you play PUBG game: string (nullable = true)
|-- How long have you been playing this game: string (nullable = true)
|-- How often do you play this Game: string (nullable = true)
|-- How much time you spent daily: string (nullable = true)
|-- How affect this game on students6: string (nullable = true)
|-- Positive effects of playing PUBG: string (nullable = true)
|-- Negative effects of playing PUBG: string (nullable = true)
|-- What are reasons that you dont play PUBG: string (nullable = true)
|-- How affect this game on students10: string (nullable = true)
|-- According to you are there positive effects of playing PUBG: string (nullable =
```

```
data1.select('Id','kills','killPoints').show()
```

```
+-----+-----+-----+
|          Id|kills|killPoints|
+-----+-----+-----+
|2f262dd9795e60|    0|    1126|
|a32847cf5bf34b|    1|    1309|
|1b1900a9990396|    2|         0|
|f589dd03b60bf2|    1|         0|
|c23c4cc5b78b35|    0|    1332|
|fd034582dd4d2e|    0|         0|
|c60b5633f4dcc8|    8|         0|
|f0ba8246b6980f|    0|         0|
|79c5d5eda1c72e|    4|         0|
|94834a28e52abd|    5|    1502|
|f051dcc9b0b3ce|    0|         0|
|f02c2f34accf08|    0|         0|
|6701c06774d409|    0|    1299|
```

```
|4e4aef4aeee5f5| 1| 0|
|d26b4b75c5229d| 0| 1267|
|c5473a410326a8| 2| 0|
|321fe9f3c71131| 0| 0|
|f8933f3ee2e431| 0| 0|
|c70c7337cd46b4| 0| 1183|
|d6c231133b5d57| 0| 1130|
+-----+-----+
only showing top 20 rows
```

```
data2.select('Age','Gender','How affect this game on students6').show()
```

```
+---+-----+-----+
|Age|Gender|How affect this game on students6|
+---+-----+-----+
| 21|Female|Positive|
| 24| Male|Negative|
| 22| Male|Negative|
| 23| Male|Positive|
| 21|Female|Negative|
| 24| Male|Positive|
| 22| Male|Negative|
| 22|Female|Negative|
| 24| Male|Negative|
| 26| Male|Negative|
| 18|Female|Negative|
| 18| Male|Positive|
| 20|Female|Positive|
| 19| Male|Positive|
| 21| Male|Positive|
| 22| Male|Positive|
| 23| Male|Positive|
| 25| Male|Negative|
| 23|Female|Positive|
| 25| Male|Positive|
+---+-----+-----+
only showing top 20 rows
```

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

```
assembeler = VectorAssembler(
    inputCols = ['kills','headshotKills'],
    outputCol = 'features'
)
```

```
data1.printSchema()
```

```
root
|-- Id: string (nullable = true)
|-- groupId: string (nullable = true)
|-- matchId: string (nullable = true)
|-- assists: integer (nullable = true)
|-- boosts: integer (nullable = true)
|-- damageDealt: float (nullable = true)
```

```

|-- DBNOs: integer (nullable = true)
|-- headshotKills: integer (nullable = true)
|-- heals: integer (nullable = true)
|-- killPlace: integer (nullable = true)
|-- killPoints: integer (nullable = true)
|-- kills: integer (nullable = true)
|-- killStreaks: integer (nullable = true)
|-- longestKill: float (nullable = true)
|-- matchDuration: integer (nullable = true)
|-- matchType: string (nullable = true)
|-- maxPlace: integer (nullable = true)
|-- numGroups: integer (nullable = true)
|-- rankPoints: integer (nullable = true)
|-- revives: integer (nullable = true)
|-- rideDistance: float (nullable = true)
|-- roadKills: integer (nullable = true)
|-- swimDistance: integer (nullable = true)
|-- teamKills: integer (nullable = true)
|-- vehicleDestroys: integer (nullable = true)
|-- walkDistance: float (nullable = true)
|-- weaponsAcquired: integer (nullable = true)
|-- winPoints: integer (nullable = true)
|-- winPlacePerc: float (nullable = true)

```

```
data1 = assembler.transform(data1)
```

```
data1.show()
```

Id	groupId	matchId	assists	boosts	damageDealt	DBNOs	headshots
2f262dd9795e60	78437bcd91d40e	d5db3a49eb2955	0	0	0.0	0	
a32847cf5bf34b	85b7ce5a12e10b	65223f05c7fdb4	0	0	163.2	1	
1b1900a9990396	edf80d6523380a	1cadec4534f30a	0	3	278.7	2	
f589dd03b60bf2	804ab5e5585558	c4a5676dc91604	0	0	191.9	1	
c23c4cc5b78b35	b3e2cd169ed920	cd595700a01bfa	0	0	100.0	1	
fd034582dd4d2e	9b8930ae0086a	6f6e52b15ddf21	0	1	200.0	2	
c60b5633f4dcc8	7c0f817f6627c7	3232c1e0fec04b	0	3	638.2	4	
f0ba8246b6980f	7318b5204462cb	112e9711f86001	0	0	27.94	0	
79c5d5eda1c72e	a85b81198dfc06	ef5fc25e28ffb1	1	4	275.8	3	
94834a28e52abd	bc513cde35fa54	f36a754a9b88f7	1	1	530.4	4	
f051dcc9b0b3ce	d203c0e3d8c321	89a6a8738190b4	0	0	20.59	0	
f02c2f34accf08	22ed911205c815	559dac9580b92a	1	0	62.72	0	
6701c06774d409	cdb79f944d585b	9f3dec5ffba4e	0	0	0.0	0	
4e4aef4ae05f5	a9dfa1c736c889	ac92da38bb19ad	0	2	13.83	0	
d26b4b75c5229d	130ea20c924e8c	14fb1c1b26e9a4	0	0	25.8	0	
c5473a410326a8	8a25860cd71a23	88cffe1ae97aff	1	1	594.0	2	
321fe9f3c71131	cb3471586d99b4	1cf664f7c75122	0	0	50.31	0	
f8933f3ee2e431	114b20e9d7504b	d5fcb7a3981d33	0	0	0.0	0	
c70c7337cd46b4	73870d831717aa	e6602141e44281	0	0	25.8	0	
d6c231133b5d57	928733f3037f92	b4baee11351ae6	0	0	30.96	0	

only showing top 20 rows

```
data1.select('features').show()
```

```
+-----+
| features|
+-----+
|(2,[],[])|
|[1.0,1.0]|
|[2.0,1.0]|
|[1.0,0.0]|
|(2,[],[])|
|(2,[],[])|
|[8.0,1.0]|
|(2,[],[])|
|[4.0,0.0]|
|[5.0,0.0]|
|(2,[],[])|
|(2,[],[])|
|(2,[],[])|
|[1.0,0.0]|
|(2,[],[])|
|[2.0,1.0]|
|(2,[],[])|
|(2,[],[])|
|(2,[],[])|
|(2,[],[])|
+-----+
```

only showing top 20 rows

```
from pyspark.ml.clustering import KMeans
from pyspark.ml.evaluation import ClusteringEvaluator
```

```
kmeans = KMeans().setK(3).setSeed(1)
```

```
model = kmeans.fit(data1)
```

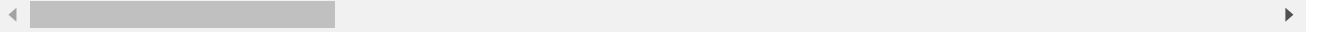
```
pred = model.transform(data1)
```

```
pred.show()
```

```
+-----+-----+-----+-----+-----+-----+-----+-----+
|          Id|      groupId|      matchId|assists|boosts|damageDealt|DBNOs|heads|
+-----+-----+-----+-----+-----+-----+-----+-----+
|2f262dd9795e60|78437bcd91d40e|d5db3a49eb2955|0|0|0.0|0|
|a32847cf5bf34b|85b7ce5a12e10b|65223f05c7fdb4|0|0|163.2|1|
|1b1900a9990396|edf80d6523380a|1cadec4534f30a|0|3|278.7|2|
|f589dd03b60bf2|804ab5e5585558|c4a5676dc91604|0|0|191.9|1|
|c23c4cc5b78b35|b3e2cd169ed920|cd595700a01bfa|0|0|100.0|1|
|fd034582dd4d2e|9b8930ae086a|6f6e52b15ddf21|0|1|200.0|2|
|c60b5633f4dcc8|7c0f817f6627c7|3232c1e0fec04b|0|3|638.2|4|
|f0ba8246b6980f|7318b5204462cb|112e9711f86001|0|0|27.94|0|
|79c5d5eda1c72e|a85b81198dfc06|ef5fc25e28ffb1|1|4|275.8|3|
|94834a28e52abd|bc513cde35fa54|f36a754a9b88f7|1|1|530.4|4|
|f051dcc9b0b3ce|d203c0e3d8c321|89a6a8738190b4|0|0|20.59|0|
|f02c2f34accf08|22ed911205c815|559dac9580b92a|1|0|62.72|0|
```


6701c06774d409	cdb79f944d585b	9f3decb5ffba4e	0	0	0.0	0
4e4aef4aeee5f5	a9dfa1c736c889	ac92da38bb19ad	0	2	13.83	0
d26b4b75c5229d	130ea20c924e8c	14fb1c1b26e9a4	0	0	25.8	0
c5473a410326a8	8a25860cd71a23	88cffe1ae97aff	1	1	594.0	2
321fe9f3c71131	cb3471586d99b4	1cf664f7c75122	0	0	50.31	0
f8933f3ee2e431	114b20e9d7504b	d5fcb7a3981d33	0	0	0.0	0
c70c7337cd46b4	73870d831717aa	e6602141e44281	0	0	25.8	0
d6c231133b5d57	928733f3037f92	b4baee11351ae6	0	0	30.96	0

+-----+-----+-----+-----+-----+-----+-----+
only showing top 20 rows



```
evaluator = ClusteringEvaluator()
```

```
res = evaluator.evaluate(pred)
```

```
res
```

```
0.7617710366356096
```

✓ 0s completed at 18:09

