Importing Modules

In [2]:

```
import findspark
findspark.init()
```

In [3]:

```
from pyspark.sql.functions import split
```

In [4]:

```
from pyspark.ml.clustering import KMeans
from pyspark import SparkContext, since
```

In [5]:

```
from pyspark.sql import SQLContext as sc
```

In [6]:

```
from pyspark.context import SparkContext
from pyspark.sql.session import SparkSession
from pyspark.ml.feature import VectorAssembler
```

Creating spark context and starting a session

In [7]:

```
sc = SparkContext.getOrCreate()
spark = SparkSession(sc)
```

Reading the data

In [8]:

```
lines = sc.textFile("F:\Docs\Big data\Assignment\Assignmnet 4\Dataset\pumsb.dat")
```

creating a 2d list from the data read. We are skipping the first attribute.

In [48]:

```
data = []
for line in lines.collect():
    l = str(line)
    feature_list = line.split()
    feature_list = feature_list[1 : :]
    data.append(list(map(lambda x : int(x), feature_list)))
print("The total number of data points are : ",len(data))
```

```
The total number of data points are :  49046
```

Creating a columnname for all the attributes

In [11]:

```
colname = []
for i in range(len(data[0])):
    colname.append("A_" + str(i))
```

In [61]:

```
print(colname)
```

```
['A_0', 'A_1', 'A_2', 'A_3', 'A_4', 'A_5', 'A_6', 'A_7', 'A_8', 'A_9', 'A_10', 'A_11', 'A_12', 'A_
13', 'A_14', 'A_15', 'A_16', 'A_17', 'A_18', 'A_19', 'A_20', 'A_21', 'A_22', 'A_23', 'A_24',
'A_25', 'A_26', 'A_27', 'A_28', 'A_29', 'A_30', 'A_31', 'A_32', 'A_33', 'A_34', 'A_35', 'A_36',
'A_37', 'A_38', 'A_39', 'A_40', 'A_41', 'A_42', 'A_43', 'A_44', 'A_45', 'A_46', 'A_47', 'A_48',
'A_49', 'A_50', 'A_51', 'A_52', 'A_53', 'A_54', 'A_55', 'A_56', 'A_57', 'A_58', 'A_59', 'A_60',
'A_61', 'A_62', 'A_63', 'A_64', 'A_65', 'A_66', 'A_67', 'A_68', 'A_69', 'A_70', 'A_71', 'A_72']
```

Creating a dataframe out of the 2d list we created

In [49]:

```
df = spark.createDataFrame(data, colname)
```

In [50]:

```
print(df.show())
```

```
+---+---+---+---+---+---+---+---+---+---+----+----+----+----+----+----+----+----+----+----+---
+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+---
----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----
---+----+----+----+----+----+----+----+----+----+----+
|A_0|A_1|A_2|A_3|A_4|A_5|A_6|A_7|A_8|A_9|A_10|A_11|A_12|A_13|A_14|A_15|A_16|A_17|A_18|A_19|A_20|A_2
22|A_23|A_24|A_25|A_26|A_27|A_28|A_29|A_30|A_31|A_32|A_33|A_34|A_35|A_36|A_37|A_38|A_39|A_40|A_41|A
A_43|A_44|A_45|A_46|A_47|A_48|A_49|A_50|A_51|A_52|A_53|A_54|A_55|A_56|A_57|A_58|A_59|A_60|A_61|A_62
3|A_64|A_65|A_66|A_67|A_68|A_69|A_70|A_71|A_72|
+---+---+---+---+---+---+---+---+---+---+----+----+----+----+----+----+----+----+----+----+---
+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+---
----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----
---+----+----+----+----+----+----+----+----+----+----+
| 14| 17| 60| 66| 75| 84|125|155|161|163| 168| 170| 180| 184| 188| 198| 205| 260| 265| 277| 290| 3
24|2297|2300|2354|2402|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4502|4503|4518|4525|4567|4680|4780|4786|4799|4815|4853|4933|4937|4940|4946|5113|6465|6857|6867|6921
2|7026|7036|7046|7057|7062|7075|7082|7092|7102|7112|
| 15| 54| 60| 66| 74| 84|111|155|161|167| 168| 170| 180| 184| 188| 198| 252| 260| 265| 277| 291|12
18|2297|2300|2354|2402|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4501|4507|4518|4525|4548|4680|4780|4786|4799|4816|4848|4933|4937|4940|4946|5833|6259|6858|6867|6917
3|7026|7036|7046|7057|7062|7073|7082|7092|7102|7112|
| 14| 17| 60| 66| 73| 84|124|155|161|163| 168| 170| 180| 184| 188| 197| 210| 260| 265| 277| 291| 3
49|1380|2299|2301|2401|3403|3404|4404|4412|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4502|4503|4518|4525|4567|4680|4780|4786|4800|4815|4858|4933|4937|4940|4946|5403|6523|6857|6867|6921
2|7026|7036|7046|7052|7062|7072|7082|7092|7102|7112|
| 15| 17| 60| 66| 73| 84|111|155|161|165| 168| 170| 180| 184| 188| 197| 252| 260| 265| 277| 291|12
93|2297|2299|2301|2401|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4492|4496|4
4502|4507|4518|4525|4567|4680|4780|4786|4800|4815|4853|4933|4937|4940|4946|5514|6205|6858|6867|6921
4|7026|7036|7046|7057|7062|7072|7082|7092|7102|7112|
| 15| 17| 57| 70| 74| 84|111|160|161|167| 168| 170| 180| 184| 188| 197| 252| 260| 265| 278| 291| 3
49|2297|2299|2301|2401|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4502|4504|4518|4525|4567|4680|4780|4786|4799|4820|4863|4933|4937|4940|4946|5794|6325|6857|6867|6899
2|7025|7035|7045|7057|7062|7072|7082|7092|7102|7112|
| 15| 17| 57| 70| 74| 85|111|160|162|167| 169| 170| 180| 184| 188| 197| 252| 260| 265| 278| 287| 3
49|2297|2299|2301|2401|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4502|4504|4523|4526|4527|4627|4727|4785|4798|4807|4833|4936|4939|4940|4952|4953|5946|6856|6868|6869
2|7022|7032|7042|7052|7062|7072|7082|7092|7102|7112|
| 15| 17| 57| 70| 73| 85|111|154|162|167| 169| 170| 180| 184| 188| 197| 252| 260| 265| 278| 285| 3
49|2297|2299|2301|2401|3403|3404|4404|4409|4414|4426|4428|4430|4432|4434|4436|4438|4440|4491|4494|4
4500|4503|4517|4524|4527|4627|4727|4785|4798|4807|4833|4933|4937|4940|4945|4953|5946|6856|6866|6869
2|7022|7032|7042|7052|7062|7072|7082|7092|7102|7112|
| 14| 17| 62| 66| 77| 84|123|155|161|163| 168| 170| 180| 184| 188| 197| 245| 260| 265| 277| 289| 3
20|1348|2300|2303|2403|3403|3404|4404|4411|4424|4426|4428|4430|4432|4434|4437|4438|4443|4493|4496|4
4502|4503|4518|4525|4539|4680|4780|4786|4800|4815|4858|4933|4937|4940|4946|5644|6220|6857|6867|6909
2|7025|7036|7045|7057|7064|7072|7085|7092|7102|7112|
| 15| 17| 62| 66| 76| 84|111|155|161|167| 168| 170| 180| 184| 188| 197| 252| 260| 265| 277| 289| 3
87|1320|2300|2303|2403|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4502|4506|4523|4526|4527|4627|4727|4785|4798|4807|4833|4936|4939|4940|4947|5574|6831|6857|6867|6877
2|7025|7035|7045|7057|7062|7072|7085|7092|7102|7112|
| 14| 17| 59| 66| 73| 84|125|155|161|163| 168| 170| 180| 184| 188| 197| 205| 260| 265| 277| 290| 3
48|1320|2300|2331|2402|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4502|4503|4518|4525|4572|4680|4780|4786|4799|4814|4848|4933|4937|4940|4946|5013|6504|6857|6867|6921
7|7026|7036|7046|7057|7062|7072|7082|7092|7102|7112|
```

```
| 15| 17| 59| 66| 73| 84|111|155|161|166| 168| 170| 180| 184| 188| 197| 252| 260| 265| 277| 291| 3
87|1446|2300|2331|2402|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4502|4507|4518|4525|4543|4680|4780|4786|4799|4816|4853|4933|4937|4940|4946|5815|6101|6857|6867|6905
8|7025|7035|7045|7057|7062|7072|7082|7092|7102|7112|
| 15| 17| 57| 70| 73| 86|125|154|162|167| 169| 170| 180| 184| 188| 197| 229| 260| 265| 278| 284| 4
93|2297|2300|2331|2402|3403|3404|4404|4409|4414|4426|4428|4430|4432|4434|4436|4438|4440|4491|4494|4
4500|4503|4517|4524|4527|4627|4727|4785|4798|4807|4833|4933|4937|4940|4945|4953|5946|6856|6866|6869
2|7022|7032|7042|7052|7062|7072|7082|7092|7102|7112|
| 15| 17| 56| 70| 73| 86|125|154|162|167| 169| 170| 180| 184| 188| 197| 229| 260| 265| 278| 283| 4
93|2297|2300|2331|2402|3403|3404|4404|4409|4414|4426|4428|4430|4432|4434|4436|4438|4440|4491|4494|4
4500|4503|4517|4524|4527|4627|4727|4785|4798|4807|4833|4933|4937|4940|4945|4953|5946|6856|6866|6869
2|7022|7032|7042|7052|7062|7072|7082|7092|7102|7112|
| 15| 17| 56| 70| 73| 86|125|154|162|167| 169| 170| 180| 184| 188| 197| 229| 260| 265| 279| 281| 4
93|2297|2298|2301|2401|3401|3404|4404|4409|4414|4426|4428|4430|4432|4434|4436|4438|4440|4491|4494|4
4500|4503|4517|4524|4527|4627|4727|4785|4798|4807|4833|4933|4937|4940|4945|4953|5946|6856|6866|6869
2|7022|7032|7042|7052|7062|7072|7082|7092|7102|7112|
| 14| 17| 60| 66| 74| 84|111|155|161|163| 168| 170| 180| 184| 188| 198| 252| 260| 265| 278| 292| 3
30|2297|2300|2354|2402|3403|3404|4404|4411|4421|4426|4428|4430|4433|4434|4436|4438|4444|4492|4496|4
4502|4503|4518|4525|4562|4680|4780|4786|4799|4816|4848|4933|4937|4940|4946|5154|6310|6857|6867|6921
7|7026|7036|7046|7057|7062|7072|7082|7092|7106|7112|
| 15| 17| 60| 66| 74| 84|111|155|161|167| 168| 170| 180| 184| 188| 198| 252| 260| 265| 277| 288| 3
20|2297|2300|2354|2402|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4502|4506|4518|4525|4567|4680|4780|4786|4799|4814|4848|4933|4937|4940|4946|5625|6189|6857|6867|6921
2|7026|7036|7046|7057|7062|7072|7082|7092|7102|7112|
| 14| 17| 58| 70| 75| 84|111|160|161|163| 168| 170| 180| 184| 188| 198| 252| 260| 265| 277| 290| 3
30|2297|2300|2354|2402|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4502|4503|4518|4525|4575|4680|4780|4786|4800|4814|4848|4933|4937|4940|4946|5265|6729|6857|6867|6921
0|7026|7036|7046|7057|7062|7072|7082|7092|7102|7112|
| 15| 17| 60| 66| 74| 84|125|155|161|167| 168| 170| 180| 184| 188| 198| 205| 260| 265| 277| 290| 3
48|2297|2299|2301|2401|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4502|4507|4518|4525|4567|4680|4780|4786|4799|4815|4848|4933|4937|4940|4946|5874|6282|6859|6867|6921
2|7026|7036|7046|7057|7062|7073|7082|7092|7102|7112|
| 14| 17| 62| 66| 75| 84|111|155|161|163| 168| 170| 180| 184| 188| 198| 252| 260| 265| 278| 294| 3
20|1387|2299|2301|2401|3403|3404|4404|4411|4422|4426|4428|4430|4432|4435|4436|4438|4443|4493|4496|4
4502|4503|4518|4525|4567|4680|4780|4795|4798|4814|4838|4933|4937|4940|4946|5884|5982|6860|6867|6921
2|7026|7036|7046|7057|7062|7074|7082|7092|7102|7112|
| 15| 17| 59| 68| 73| 84|111|158|161|167| 168| 170| 180| 184| 188| 197| 252| 260| 265| 277| 290| 4
40|1347|2299|2301|2401|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4502|4504|4518|4525|4567|4680|4780|4786|4799|4814|4848|4933|4937|4940|4946|5554|6189|6857|6867|6921
2|7026|7036|7046|7057|7062|7072|7082|7092|7102|7112|
+---+---+---+---+---+---+---+---+---+---+----+----+----+----+----+----+----+----+----+----+----+---
+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+---
----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----
---+----+----+----+----+----+----+----+----+----+----+
only showing top 20 rows

None
```

Aliasing the inbuild KMeans function as kmeans with number of clusters 5 and seed as 1 (random initial points)

In [51]:

```
kmeans = KMeans(k = 5, seed = 1)
```

To make the data readable by the module, we need to transform them. So we are transforming them using the below code. Transformation is nothing but making afeature vector of all the input attribute and storing it as an attribute

In [52]:

```
vecAssembler = VectorAssembler(inputCols=colname, outputCol="features")
```

In [53]:

```
new_df = vecAssembler.transform(df)
new_df.show()
```

```
+---+---+---+---+---+---+---+---+---+---+----+----+----+----+----+----+----+----+----+----+----+---
+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+---
----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----
---+----+----+----+----+----+----+----+----+----+----+--------------------+
|A_0|A_1|A_2|A_3|A_4|A_5|A_6|A_7|A_8|A_9|A_10|A_11|A_12|A_13|A_14|A_15|A_16|A_17|A_18|A_19|A_20|A_2
22|A_23|A_24|A_25|A_26|A_27|A_28|A_29|A_30|A_31|A_32|A_33|A_34|A_35|A_36|A_37|A_38|A_39|A_40|A_41|A
```

```
22|A_23|A_24|A_25|A_26|A_27|A_28|A_29|A_30|A_31|A_32|A_33|A_34|A_35|A_36|A_37|A_38|A_39|A_40|A_41|A
A_43|A_44|A_45|A_46|A_47|A_48|A_49|A_50|A_51|A_52|A_53|A_54|A_55|A_56|A_57|A_58|A_59|A_60|A_61|A_62
3|A_64|A_65|A_66|A_67|A_68|A_69|A_70|A_71|A_72|            features|
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| 14| 17| 60| 66| 75| 84|125|155|161|163| 168| 170| 180| 184| 188| 198| 205| 260| 265| 277| 290| 3
24|2297|2300|2354|2402|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4502|4503|4518|4525|4567|4680|4780|4786|4799|4815|4853|4933|4937|4940|4946|5113|6465|6857|6867|6921
2|7026|7036|7046|7057|7062|7075|7082|7092|7102|7112|[14.0,17.0,60.0,6...|
| 15| 54| 60| 66| 74| 84|111|155|161|167| 168| 170| 180| 184| 188| 198| 252| 260| 265| 277| 291|12
18|2297|2300|2354|2402|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4501|4507|4518|4525|4548|4680|4780|4786|4799|4816|4848|4933|4937|4940|4946|5833|6259|6858|6867|6917
3|7026|7036|7046|7057|7062|7073|7082|7092|7102|7112|[15.0,54.0,60.0,6...|
| 14| 17| 60| 66| 73| 84|124|155|161|163| 168| 170| 180| 184| 188| 197| 210| 260| 265| 277| 291| 3
49|1380|2299|2301|2401|3403|3404|4404|4412|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4502|4503|4518|4525|4567|4680|4780|4786|4800|4815|4858|4933|4937|4940|4946|5403|6523|6857|6867|6921
2|7026|7036|7046|7052|7062|7072|7082|7092|7102|7112|[14.0,17.0,60.0,6...|
| 15| 17| 60| 66| 73| 84|111|155|161|165| 168| 170| 180| 184| 188| 197| 252| 260| 265| 277| 291|12
93|2297|2299|2301|2401|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4492|4496|4
4502|4507|4518|4525|4567|4680|4780|4786|4800|4815|4853|4933|4937|4940|4946|5514|6205|6858|6867|6921
4|7026|7036|7046|7057|7062|7072|7082|7092|7102|7112|[15.0,17.0,60.0,6...|
| 15| 17| 57| 70| 74| 84|111|160|161|167| 168| 170| 180| 184| 188| 197| 252| 260| 265| 278| 291| 3
49|2297|2299|2301|2401|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4502|4504|4518|4525|4567|4680|4780|4786|4799|4820|4863|4933|4937|4940|4946|5794|6325|6857|6867|6899
2|7025|7035|7045|7057|7062|7072|7082|7092|7102|7112|[15.0,17.0,57.0,7...|
| 15| 17| 57| 70| 74| 85|111|160|162|167| 169| 170| 180| 184| 188| 197| 252| 260| 265| 278| 287| 3
49|2297|2299|2301|2401|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4502|4504|4523|4526|4527|4627|4727|4785|4798|4807|4833|4936|4939|4940|4952|4953|5946|6856|6868|6869
2|7022|7032|7042|7052|7062|7072|7082|7092|7102|7112|[15.0,17.0,57.0,7...|
| 15| 17| 57| 70| 73| 85|111|154|162|167| 169| 170| 180| 184| 188| 197| 252| 260| 265| 278| 285| 3
49|2297|2299|2301|2401|3403|3404|4404|4409|4414|4426|4428|4430|4432|4434|4436|4438|4440|4491|4494|4
4500|4503|4517|4524|4527|4627|4727|4785|4798|4807|4833|4933|4937|4940|4945|4953|5946|6856|6866|6869
2|7022|7032|7042|7052|7062|7072|7082|7092|7102|7112|[15.0,17.0,57.0,7...|
| 14| 17| 62| 66| 77| 84|123|155|161|163| 168| 170| 180| 184| 188| 197| 245| 260| 265| 277| 289| 3
20|1348|2300|2303|2403|3403|3404|4404|4411|4424|4426|4428|4430|4432|4434|4437|4438|4443|4493|4496|4
4502|4503|4518|4525|4539|4680|4780|4786|4800|4815|4858|4933|4937|4940|4946|5644|6220|6857|6867|6909
2|7025|7036|7045|7057|7064|7072|7085|7092|7102|7112|[14.0,17.0,62.0,6...|
| 15| 17| 62| 66| 76| 84|111|155|161|167| 168| 170| 180| 184| 188| 197| 252| 260| 265| 277| 289| 3
87|1320|2300|2303|2403|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4502|4506|4523|4526|4527|4627|4727|4785|4798|4807|4833|4936|4939|4940|4947|5574|6831|6857|6867|6877
2|7025|7035|7045|7057|7062|7072|7085|7092|7102|7112|[15.0,17.0,62.0,6...|
| 14| 17| 59| 66| 73| 84|125|155|161|163| 168| 170| 180| 184| 188| 197| 205| 260| 265| 277| 290| 3
48|1320|2300|2331|2402|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4502|4503|4518|4525|4572|4680|4780|4786|4799|4814|4848|4933|4937|4940|4946|5013|6504|6857|6867|6921
7|7026|7036|7046|7057|7062|7072|7082|7092|7102|7112|[14.0,17.0,59.0,6...|
| 15| 17| 59| 66| 73| 84|111|155|161|166| 168| 170| 180| 184| 188| 197| 252| 260| 265| 277| 291| 3
87|1446|2300|2331|2402|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4502|4507|4518|4525|4543|4680|4780|4786|4799|4816|4853|4933|4937|4940|4946|5815|6101|6857|6867|6905
8|7025|7035|7045|7057|7062|7072|7082|7092|7102|7112|[15.0,17.0,59.0,6...|
| 15| 17| 57| 70| 73| 86|125|154|162|167| 169| 170| 180| 184| 188| 197| 229| 260| 265| 278| 284| 4
93|2297|2300|2331|2402|3403|3404|4404|4409|4414|4426|4428|4430|4432|4434|4436|4438|4440|4491|4494|4
4500|4503|4517|4524|4527|4627|4727|4785|4798|4807|4833|4933|4937|4940|4945|4953|5946|6856|6866|6869
2|7022|7032|7042|7052|7062|7072|7082|7092|7102|7112|[15.0,17.0,57.0,7...|
| 15| 17| 56| 70| 73| 86|125|154|162|167| 169| 170| 180| 184| 188| 197| 229| 260| 265| 278| 283| 4
93|2297|2300|2331|2402|3403|3404|4404|4409|4414|4426|4428|4430|4432|4434|4436|4438|4440|4491|4494|4
4500|4503|4517|4524|4527|4627|4727|4785|4798|4807|4833|4933|4937|4940|4945|4953|5946|6856|6866|6869
2|7022|7032|7042|7052|7062|7072|7082|7092|7102|7112|[15.0,17.0,56.0,7...|
| 15| 17| 56| 70| 73| 86|125|154|162|167| 169| 170| 180| 184| 188| 197| 229| 260| 265| 279| 281| 4
93|2297|2298|2301|2401|3401|3404|4404|4409|4414|4426|4428|4430|4432|4434|4436|4438|4440|4491|4494|4
4500|4503|4517|4524|4527|4627|4727|4785|4798|4807|4833|4933|4937|4940|4945|4953|5946|6856|6866|6869
2|7022|7032|7042|7052|7062|7072|7082|7092|7102|7112|[15.0,17.0,56.0,7...|
| 14| 17| 60| 66| 74| 84|111|155|161|163| 168| 170| 180| 184| 188| 198| 252| 260| 265| 278| 292| 3
30|2297|2300|2354|2402|3403|3404|4404|4411|4421|4426|4428|4430|4433|4434|4436|4438|4444|4492|4496|4
4502|4503|4518|4525|4562|4680|4780|4786|4799|4816|4848|4933|4937|4940|4946|5154|6310|6857|6867|6921
7|7026|7036|7046|7057|7062|7072|7082|7092|7106|7112|[14.0,17.0,60.0,6...|
| 15| 17| 60| 66| 74| 84|111|155|161|167| 168| 170| 180| 184| 188| 198| 252| 260| 265| 277| 288| 3
20|2297|2300|2354|2402|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4502|4506|4518|4525|4567|4680|4780|4786|4799|4814|4848|4933|4937|4940|4946|5625|6189|6857|6867|6921
2|7026|7036|7046|7057|7062|7072|7082|7092|7102|7112|[15.0,17.0,60.0,6...|
| 14| 17| 58| 70| 75| 84|111|160|161|163| 168| 170| 180| 184| 188| 198| 252| 260| 265| 277| 290| 3
30|2297|2300|2354|2402|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4502|4503|4518|4525|4575|4680|4780|4786|4800|4814|4848|4933|4937|4940|4946|5265|6729|6857|6867|6921
0|7026|7036|7046|7057|7062|7072|7082|7092|7102|7112|[14.0,17.0,58.0,7...|
| 15| 17| 60| 66| 74| 84|125|155|161|167| 168| 170| 180| 184| 188| 198| 205| 260| 265| 277| 290| 3
48|2297|2299|2301|2401|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4502|4507|4518|4525|4567|4680|4780|4786|4799|4815|4848|4933|4937|4940|4946|5874|6282|6859|6867|6921
```

```
4502|4507|4518|4525|4567|4680|4780|4786|4799|4815|4848|4933|4937|4940|4946|5874|6282|6859|6867|6921
2|7026|7036|7046|7057|7062|7073|7082|7092|7102|7112|[15.0,17.0,60.0,6...|
| 14| 17| 62| 66| 75| 84|111|155|161|163| 168| 170| 180| 184| 188| 198| 252| 260| 265| 278| 294| 3
20|1387|2299|2301|2401|3403|3404|4404|4411|4422|4426|4428|4430|4432|4435|4436|4438|4443|4493|4496|4
4502|4503|4518|4525|4567|4680|4780|4795|4798|4814|4838|4933|4937|4940|4946|5884|5982|6860|6867|6921
2|7026|7036|7046|7057|7062|7074|7082|7092|7102|7112|[14.0,17.0,62.0,6...|
| 15| 17| 59| 68| 73| 84|111|158|161|167| 168| 170| 180| 184| 188| 197| 252| 260| 265| 277| 290| 4
40|1347|2299|2301|2401|3403|3404|4404|4413|4414|4426|4428|4430|4432|4434|4436|4438|4440|4493|4496|4
4502|4504|4518|4525|4567|4680|4780|4786|4799|4814|4848|4933|4937|4940|4946|5554|6189|6857|6867|6921
2|7026|7036|7046|7057|7062|7072|7082|7092|7102|7112|[15.0,17.0,59.0,6...|
+---+---+---+---+---+---+---+---+---+---+----+----+----+----+----+----+----+----+----+----+----+---
+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+---
----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----
---+----+----+----+----+----+----+----+----+----+----+--------------------+
only showing top 20 rows
```

Now we are fitting the model.

In [54]:

```
model = kmeans.fit(new_df.select('features'))
```

We are dropping all the initial columns and showing the only the feature vector of each row with its cluster

In [55]:

```
transformed = model.transform(new_df)
# for name in colname:
#     transformed.drop(name).collect()

transformed = transformed.drop(*colname)
transformed.show()
```

```
+--------------------+----------+
|            features|prediction|
+--------------------+----------+
|[14.0,17.0,60.0,6...|         2|
|[15.0,54.0,60.0,6...|         1|
|[14.0,17.0,60.0,6...|         0|
|[15.0,17.0,60.0,6...|         1|
|[15.0,17.0,57.0,7...|         3|
|[15.0,17.0,57.0,7...|         2|
|[15.0,17.0,57.0,7...|         2|
|[14.0,17.0,62.0,6...|         0|
|[15.0,17.0,62.0,6...|         0|
|[14.0,17.0,59.0,6...|         0|
|[15.0,17.0,59.0,6...|         0|
|[15.0,17.0,57.0,7...|         2|
|[15.0,17.0,56.0,7...|         2|
|[15.0,17.0,56.0,7...|         2|
|[14.0,17.0,60.0,6...|         2|
|[15.0,17.0,60.0,6...|         3|
|[14.0,17.0,58.0,7...|         3|
|[15.0,17.0,60.0,6...|         3|
|[14.0,17.0,62.0,6...|         0|
|[15.0,17.0,59.0,6...|         0|
+--------------------+----------+
only showing top 20 rows
```

Printing the centers of the clusters

In [58]:

```
print("The Centres are : ")
for centre in model.clusterCenters():
    print(centre)
```

```
The Centres are :
[  14.52550679   17.25211628   58.89836266   67.98318111   74.44965471
```

```
   85.43489641  117.41406772  156.0247828   161.28837157  164.8505235
  168.29906438  170.02962798  180.01598352  184.03558699  188.04160169
  197.24353976  240.51681889  260.06042548  265.15187124  277.27216529
  288.77556249  362.94959902 1350.33749165 2299.36372243 2321.71625084
 2415.99548897 3402.80262865 3419.32613054 4404.03246826 4411.61511472
 4414.87157496 4426.01798842 4428.01353308 4430.04834039 4432.0261194
 4434.02511695 4436.04238138 4438.00178213 4440.65170417 4492.36082646
 4495.40198262 4498.46758744 4501.46920249 4504.1192916  4519.03675652
 4525.05012252 4544.40098017 4650.53079751 4750.51253063 4785.88856093
 4798.45316329 4810.96703052 4842.58793718 4933.74966585 4937.50506794
 4940.09751615 4946.58086434 5301.10269548 6156.56309869 6857.30101359
 6866.92169748 6891.24309423 6942.2687681  7023.91089329 7034.41579416
 7043.78697928 7055.1058699  7062.02311205 7072.59890844 7082.39424148
 7092.06582758 7102.26698597 7112.14362887]
[  14.49672392   21.35983708   59.15618913   67.49902603   74.53391181
   84.15937666  123.41809811  156.98034355  161.02904197  164.5723393
  168.03045865  170.07915707  180.01540641  184.02284399  188.06622986
  197.17814769  240.51744289  260.55144324  265.80325837  277.20258544
  290.54258899 1216.25394015 2271.64529839 2299.61342306 2331.12856384
 2457.96847884 3402.83194617 3523.3523995  4404.27749247 4412.57658934
 4414.79829998 4426.06605277 4428.03435452 4430.06764654 4432.02727112
 4434.01965645 4436.02054188 4438.0012396  4440.98229148 4492.91464494
 4495.97060386 4498.98246857 4501.96812467 4504.35452453 4519.12944927
 4525.13423056 4556.15211617 4666.36585798 4766.33681601 4786.50876572
 4798.77403931 4813.91623871 4848.76518505 4933.60031875 4937.39366035
 4940.20842925 4946.29998229 5590.08942801 6380.95714539 6858.09845936
 6867.1087303  6905.72746591 6956.03169825 7025.16858509 7035.37435807
 7045.02815654 7056.22914822 7062.01168762 7072.39826457 7082.14892863
 7092.13015761 7102.13564725 7112.17336639]
[  14.4664973    20.00592781   59.22874987   68.00063512   74.27352599
   86.01227903  120.6263364   155.60294273  161.34592992  164.74076426
  168.36328993  170.06023076  180.0155605   184.03736636  188.18376204
  197.16227374  242.22419816  260.36508945  265.6361808   277.25320207
  287.32941675  373.30951625 2290.37154652 2299.31353869 2320.88673653
 2427.99089658 3402.67301789 3477.97078438 4404.20768498 4411.30782259
 4414.98084048 4426.0084683  4428.00762147 4430.03165026 4432.02222928
 4434.02381708 4436.06361808 4438.00370488 4440.58897004 4492.15179422
 4495.18175082 4498.27469038 4501.29046258 4503.99735366 4519.70308034
 4525.13083519 4536.39832751 4638.44511485 4738.42775484 4785.45390071
 4798.24759183 4808.67661692 4837.99089658 4934.15465227 4937.78564624
 4940.07229808 4947.45633534 4987.63289933 6069.3295226  6856.63501641
 6867.02159416 6880.28262941 6933.32994601 7022.98645073 7033.95215412
 7042.88430189 7054.5321266  7062.05546734 7072.55700222 7082.73917646
 7092.09103419 7102.36805335 7112.11643908]
[  14.5074914    18.31605588   59.48886414   67.10933387   74.49473578
   84.10022272  120.64081798  156.47853817  161.02085442  164.59819802
  168.02176554  170.03644462  180.00850375  184.01255315  188.12026726
  197.27535938  239.39360194  260.26017412  265.52996558  277.15033408
  291.13980563  363.41526625 2291.48916785 2299.56722009 2327.28133225
 2429.67584531 3402.91860701 3455.19143551 4404.1203685  4412.53391375
 4415.08604981 4426.0487953  4428.0308767  4430.08311399 4432.04049403
 4434.03300263 4436.03340757 4438.00111359 4441.05750152 4492.91769589
 4495.97519741 4498.98886414 4501.98258757 4504.40483904 4519.03938044
 4525.20186273 4557.64679085 4667.65296619 4767.60315853 4786.49574813
 4798.788115   4813.83853007 4848.65924276 4933.59232638 4937.39279206
 4940.13514882 4946.29732739 5632.27232233 6316.65296619 6858.26594452
 6867.09880543 6907.66572181 6956.78355942 7025.26958899 7035.5155902
 7045.08068435 7056.32061146 7062.02247419 7072.64922049 7082.23010731
 7092.07299048 7102.21249241 7112.17857866]
[  14.51667212   22.16721151   58.06538084   68.89539065   74.27786858
   87.43935927  119.4187643   155.39293887  161.55148741  164.96126185
  168.58679307  170.0737169   180.03367113  184.08270677  188.06864989
  197.07126512  244.22360248  260.36760379  265.43919582  277.35322001
  285.47777051 1223.39081399 2245.6900948  2299.22850605 2322.07649559
 2440.66884603 3402.50866296 3478.00964367 4404.18649886 4410.54347826
 4414.4610984  4426.00261523 4428.00637463 4430.01405688 4432.00882641
 4434.01095129 4436.030729   4438.00277869 4440.27345538 4491.71542988
 4494.73308271 4497.83000981 4500.84161491 4503.76364825 4518.99231775
 4524.81088591 4530.72392939 4631.86400785 4731.85730631 4785.19303694
 4798.10166721 4807.72409284 4835.23439032 4933.80712651 4937.54968944
 4940.05916966 4946.9228506  4967.57584178 5992.53269042 6856.27786858
 6866.72098725 6873.72507355 6926.68388362 7022.41271657 7033.08303367
 7042.37381497 7053.50196143 7062.01634521 7072.21624714 7082.42301406
 7092.14073227 7102.18617195 7112.09480222]
```

In [ ]: