In [ ]:

```
#pip install findspark
```

In [1]:

```
import pyspark
import findspark
```

In [2]:

```
from pyspark.context import SparkContext
from pyspark.sql.session import SparkSession
sc = SparkContext('local')
spark = SparkSession(sc)
```

In [3]:

```
import pyspark.sql.functions as f
from pyspark.sql.functions import lit, when, col, regexp_extract,desc
from pyspark.sql import SQLContext
from pyspark.sql import DataFrameStatFunctions as statFunc
from pyspark.sql.functions import explode, col, udf, mean as mean, stddev as stddev
from pyspark.sql.types import IntegerType,StringType
from pyspark.sql.functions import udf
```

In [4]:

```
from pyspark.sql import *
from pyspark.sql.types import *
```

In [5]:

```
df= spark.read.csv("./dataset/Marketing_Analysis.csv",inferSchema=True,header=True)
```

In [6]:

```
df
```

Out[6]:

```
DataFrame[age: int, job: string, marital: string, education: string, default:
string, balance: int, housing: string, loan: string, contact: string, day: in
t, month: string, duration: int, campaign: int, pdays: int, previous: int, po
utcome: string, y: string]
```

In [7]:

```
df.show()
```

```
+---+-----------+--------+---------+-------+-------+-------+----+-------+---
+-----+--------+--------+-----+--------+--------+---+
|age|        job| marital|education|default|balance|housing|loan|contact|day
|month|duration|campaign|pdays|previous|poutcome|  y|
+---+-----------+--------+---------+-------+-------+-------+----+-------+---
+-----+--------+--------+-----+--------+--------+---+
| 58|  management| married| tertiary|     no|   2143|    yes|  no|unknown|  5
|  may|     261|       1|   -1|       0| unknown| no|
| 44|  technician|  single|secondary|     no|     29|    yes|  no|unknown|  5
|  may|     151|       1|   -1|       0| unknown| no|
| 33|entrepreneur| married|secondary|     no|      2|    yes| yes|unknown|  5
|  may|      76|       1|   -1|       0| unknown| no|
| 47| blue-collar| married|  unknown|     no|   1506|    yes|  no|unknown|  5
|  may|      92|       1|   -1|       0| unknown| no|
| 33|     unknown|  single|  unknown|     no|      1|     no|  no|unknown|  5
|  may|     198|       1|   -1|       0| unknown| no|
| 35|  management| married| tertiary|     no|    231|    yes|  no|unknown|  5
|  may|     139|       1|   -1|       0| unknown| no|
| 28|  management|  single| tertiary|     no|    447|    yes| yes|unknown|  5
|  may|     217|       1|   -1|       0| unknown| no|
| 42|entrepreneur|divorced| tertiary|    yes|      2|    yes|  no|unknown|  5
|  may|     380|       1|   -1|       0| unknown| no|
| 58|     retired| married|  primary|     no|    121|    yes|  no|unknown|  5
|  may|      50|       1|   -1|       0| unknown| no|
| 43|  technician|  single|secondary|     no|    593|    yes|  no|unknown|  5
|  may|      55|       1|   -1|       0| unknown| no|
| 41|      admin.|divorced|secondary|     no|    270|    yes|  no|unknown|  5
|  may|     222|       1|   -1|       0| unknown| no|
| 29|      admin.|  single|secondary|     no|    390|    yes|  no|unknown|  5
|  may|     137|       1|   -1|       0| unknown| no|
| 53|  technician| married|secondary|     no|      6|    yes|  no|unknown|  5
|  may|     517|       1|   -1|       0| unknown| no|
| 58|  technician| married|  unknown|     no|     71|    yes|  no|unknown|  5
|  may|      71|       1|   -1|       0| unknown| no|
| 57|    services| married|secondary|     no|    162|    yes|  no|unknown|  5
|  may|     174|       1|   -1|       0| unknown| no|
| 51|     retired| married|  primary|     no|    229|    yes|  no|unknown|  5
|  may|     353|       1|   -1|       0| unknown| no|
| 45|      admin.|  single|  unknown|     no|     13|    yes|  no|unknown|  5
|  may|      98|       1|   -1|       0| unknown| no|
| 57| blue-collar| married|  primary|     no|     52|    yes|  no|unknown|  5
|  may|      38|       1|   -1|       0| unknown| no|
| 60|     retired| married|  primary|     no|     60|    yes|  no|unknown|  5
|  may|     219|       1|   -1|       0| unknown| no|
| 33|    services| married|secondary|     no|      0|    yes|  no|unknown|  5
|  may|      54|       1|   -1|       0| unknown| no|
+---+-----------+--------+---------+-------+-------+-------+----+-------+---
+-----+--------+--------+-----+--------+--------+---+
only showing top 20 rows
```

In [8]:

```
df.printSchema()
```

```
root
 |-- age: integer (nullable = true)
 |-- job: string (nullable = true)
 |-- marital: string (nullable = true)
 |-- education: string (nullable = true)
 |-- default: string (nullable = true)
 |-- balance: integer (nullable = true)
 |-- housing: string (nullable = true)
 |-- loan: string (nullable = true)
 |-- contact: string (nullable = true)
 |-- day: integer (nullable = true)
 |-- month: string (nullable = true)
 |-- duration: integer (nullable = true)
 |-- campaign: integer (nullable = true)
 |-- pdays: integer (nullable = true)
 |-- previous: integer (nullable = true)
 |-- poutcome: string (nullable = true)
 |-- y: string (nullable = true)
```

In [9]:

```
df.columns
```

Out[9]:

```
['age',
 'job',
 'marital',
 'education',
 'default',
 'balance',
 'housing',
 'loan',
 'contact',
 'day',
 'month',
 'duration',
 'campaign',
 'pdays',
 'previous',
 'poutcome',
 'y']
```

In [10]:

```
df.count()
```

Out[10]:

```
45211
```

In [11]:

```
a=df.count()
a
```

Out[11]:

45211

In [12]:

```
yes=df.filter(df.y=='yes').count()
yes
```

Out[12]:

5289

In [13]:

```
no=df.filter(df.y=='no').count()
no
```

Out[13]:

39922

In [14]:

```
#people subscribed
ps=yes/a
ps*100
```

Out[14]:

11.698480458295547

In [15]:

```
#people not subscribed
pns=no/a
pns*100
```

Out[15]:

88.30151954170445

In [16]:

```
#average targeted customer
df.describe("age").show()
```

```
+-------+------------------+
|summary|               age|
+-------+------------------+
|  count|             45211|
|   mean| 40.93621021432837|
| stddev|10.618762040975405|
|    min|                18|
|    max|                95|
+-------+------------------+
```

In [17]:

```
#average balance of customers
bal=df.agg({'balance': 'mean'}).show()
bal
```

```
+------------------+
|      avg(balance)|
+------------------+
|1362.2720576850766|
+------------------+
```

In [18]:

```
# median balance of customers
median=df.approxQuantile('balance',[0.5],0)
print ('The median of Balance is '+str(median))
```

```
The median of Balance is [448.0]
```

In [19]:

```
ba=df.groupby(df['balance']).count().show()
ba
```

```
+-------+-----+
|balance|count|
+-------+-----+
|    148|   39|
|    471|   26|
|   -125|    6|
|   2142|    5|
|    496|   21|
|   1342|    8|
|    463|   20|
|   3749|    1|
|   1088|   15|
|  11317|    1|
|   1238|    9|
|   3175|    3|
|   3997|    1|
|   -362|    2|
|   2366|    8|
|   4519|    4|
|   1959|    4|
|   7982|    2|
|   -565|    3|
|   6397|    1|
+-------+-----+
only showing top 20 rows
```

In [20]:

```
df.select('balance').show()
```

```
+-------+
|balance|
+-------+
|   2143|
|     29|
|      2|
|   1506|
|      1|
|    231|
|    447|
|      2|
|    121|
|    593|
|    270|
|    390|
|      6|
|     71|
|    162|
|    229|
|     13|
|     52|
|     60|
|      0|
+-------+
only showing top 20 rows
```

In [21]:

```
df.groupBy("age").pivot('y').count().show()
```

```
+---+----+----+
|age|  no| yes|
+---+----+----+
| 31|1790| 206|
| 85|   1|   4|
| 65|  38|  21|
| 53| 806|  85|
| 78|  16|  14|
| 34|1732| 198|
| 81|  11|   6|
| 28| 876| 162|
| 76|  16|  16|
| 27| 768| 141|
| 26| 671| 134|
| 44|1043|  93|
| 22|  89|  40|
| 93|null|   2|
| 47| 975| 113|
| 52| 826|  85|
| 86|   5|   4|
| 20|  35|  15|
| 40|1239| 116|
| 94|   1|null|
+---+----+----+
only showing top 20 rows
```

In [22]:

```
#age matters in marketing subscription
df.where(df.y=='yes').groupBy(df.age).count().sort(desc("count")).show()
```

```
+---+-----+
|age|count|
+---+-----+
| 32|  221|
| 30|  217|
| 33|  210|
| 35|  209|
| 31|  206|
| 34|  198|
| 36|  195|
| 29|  171|
| 37|  170|
| 28|  162|
| 38|  144|
| 39|  143|
| 27|  141|
| 26|  134|
| 41|  120|
| 46|  118|
| 40|  116|
| 47|  113|
| 25|  113|
| 42|  111|
+---+-----+
only showing top 20 rows
```

In [23]:

```
#marital status mattered for a subscription
df.groupBy('marital').pivot('y').count().show()
```

```
+--------+-----+----+
| marital|   no| yes|
+--------+-----+----+
|divorced| 4585| 622|
| married|24459|2755|
|  single|10878|1912|
+--------+-----+----+
```

In [24]:

```
df.groupBy("age",'y').pivot("marital").agg(f.count("y")).show()
```

```
+---+---+--------+-------+------+
|age|  y|divorced|married|single|
+---+---+--------+-------+------+
| 78| no|       6|     10|  null|
| 20| no|    null|      2|    33|
| 56|yes|      13|     49|     6|
| 28|yes|       4|     20|   138|
| 29|yes|       5|     33|   133|
| 86|yes|       1|      2|     1|
| 71| no|       3|     25|     1|
| 57| no|     133|    584|    33|
| 79|yes|       2|      8|  null|
| 22|yes|    null|   null|    40|
| 31|yes|      15|     80|   111|
| 42| no|     165|    770|   196|
| 87|yes|       1|      2|  null|
| 59|yes|      16|     66|     6|
| 34|yes|      11|    118|    69|
| 25| no|       6|     84|   324|
| 63| no|       3|     43|     1|
| 23|yes|    null|      2|    42|
| 24| no|       1|     43|   190|
| 64| no|       5|     34|  null|
+---+---+--------+-------+------+
only showing top 20 rows
```

In [25]:

```
#age and marital status together mattered for a subscription
df.where(df.y=='yes').groupBy(df.age).pivot("marital").agg(f.count("y")).show()
```

```
+---+--------+-------+------+
|age|divorced|married|single|
+---+--------+-------+------+
| 31|      15|     80|   111|
| 85|       1|      3|  null|
| 65|       2|     19|  null|
| 53|      18|     60|     7|
| 78|       6|      8|  null|
| 34|      11|    118|    69|
| 81|       2|      4|  null|
| 28|       4|     20|   138|
| 76|       6|     10|  null|
| 27|       2|     29|   110|
| 26|    null|     13|   121|
| 44|      21|     48|    24|
| 22|    null|   null|    40|
| 93|    null|      2|  null|
| 47|      10|     83|    20|
| 52|      10|     67|     8|
| 86|       1|      2|     1|
| 40|      12|     73|    31|
| 20|    null|      1|    14|
| 57|      15|     58|     5|
+---+--------+-------+------+
only showing top 20 rows
```

In [26]:

```
df.where(df.y=='no').groupBy(df.age).pivot("marital").agg(f.count("y")).show()
```

```
+---+--------+-------+------+
|age|divorced|married|single|
+---+--------+-------+------+
| 31|      83|    801|   906|
| 85|    null|      1|  null|
| 65|       7|     31|  null|
| 53|     145|    597|    64|
| 78|       6|     10|  null|
| 34|     138|   1013|   581|
| 81|       6|      5|  null|
| 28|      12|    305|   559|
| 76|       2|     14|  null|
| 26|      20|    157|   494|
| 27|      16|    204|   548|
| 44|     163|    734|   146|
| 22|    null|      9|    80|
| 47|     152|    743|    80|
| 52|     140|    632|    54|
| 86|       1|      4|  null|
| 40|     157|    856|   226|
| 20|    null|      2|    33|
| 94|       1|   null|  null|
| 57|     133|    584|    33|
+---+--------+-------+------+
only showing top 20 rows
```

In [27]:

```
df.groupBy('age',).pivot('y').count().show()
```

```
+---+----+----+
|age|  no| yes|
+---+----+----+
| 31|1790| 206|
| 85|   1|   4|
| 65|  38|  21|
| 53| 806|  85|
| 78|  16|  14|
| 34|1732| 198|
| 81|  11|   6|
| 28| 876| 162|
| 76|  16|  16|
| 27| 768| 141|
| 26| 671| 134|
| 44|1043|  93|
| 22|  89|  40|
| 93|null|   2|
| 47| 975| 113|
| 52| 826|  85|
| 86|   5|   4|
| 20|  35|  15|
| 40|1239| 116|
| 94|   1|null|
+---+----+----+
only showing top 20 rows
```

In [28]:

```
#feature engineering for right age effect on the campaign

fe=df.where(df.y=='yes').groupBy(df.age).count().sort(desc("count")).show()
fe
```

```
+---+-----+
|age|count|
+---+-----+
| 32|  221|
| 30|  217|
| 33|  210|
| 35|  209|
| 31|  206|
| 34|  198|
| 36|  195|
| 29|  171|
| 37|  170|
| 28|  162|
| 38|  144|
| 39|  143|
| 27|  141|
| 26|  134|
| 41|  120|
| 46|  118|
| 40|  116|
| 47|  113|
| 25|  113|
| 42|  111|
+---+-----+
only showing top 20 rows
```

In [29]:

```
df.where(df.y=='yes').groupBy(df.age).count().show()
```

```
+---+-----+
|age|count|
+---+-----+
| 31|  206|
| 85|    4|
| 65|   21|
| 53|   85|
| 78|   14|
| 34|  198|
| 81|    6|
| 28|  162|
| 76|   16|
| 26|  134|
| 27|  141|
| 44|   93|
| 22|   40|
| 93|    2|
| 47|  113|
| 52|   85|
| 86|    4|
| 40|  116|
| 20|   15|
| 57|   78|
+---+-----+
only showing top 20 rows
```

In [30]:

```
ag=df.select('age').show()
ag
```

```
+---+
|age|
+---+
| 58|
| 44|
| 33|
| 47|
| 33|
| 35|
| 28|
| 42|
| 58|
| 43|
| 41|
| 29|
| 53|
| 58|
| 57|
| 51|
| 45|
| 57|
| 60|
| 33|
+---+
only showing top 20 rows
```

In [31]:

```
df.withColumn("ageT", f.when(df.age <= 12, 'Infant').when(( (df.age >= 15) & (df.age <= 30
)), 'YOUNG').when(( (df.age >= 31) & (df.age < 59)), 'MID').when(df.age >= 60, 'OLD').othe
rwise('N/A'))
df.show()
```

```
+---+-----------+--------+---------+-------+-------+-------+----+-------+---
+-----+--------+--------+-----+--------+--------+---+
|age|         job| marital|education|default|balance|housing|loan|contact|day
|month|duration|campaign|pdays|previous|poutcome|  y|
+---+-----------+--------+---------+-------+-------+-------+----+-------+---
+-----+--------+--------+-----+--------+--------+---+
| 58|  management| married| tertiary|     no|   2143|    yes|  no|unknown|  5
|  may|     261|       1|   -1|       0| unknown| no|
| 44|  technician|  single|secondary|     no|     29|    yes|  no|unknown|  5
|  may|     151|       1|   -1|       0| unknown| no|
| 33|entrepreneur| married|secondary|     no|      2|    yes| yes|unknown|  5
|  may|      76|       1|   -1|       0| unknown| no|
| 47| blue-collar| married|  unknown|     no|   1506|    yes|  no|unknown|  5
|  may|      92|       1|   -1|       0| unknown| no|
| 33|     unknown|  single|  unknown|     no|      1|     no|  no|unknown|  5
|  may|     198|       1|   -1|       0| unknown| no|
| 35|  management| married| tertiary|     no|    231|    yes|  no|unknown|  5
|  may|     139|       1|   -1|       0| unknown| no|
| 28|  management|  single| tertiary|     no|    447|    yes| yes|unknown|  5
|  may|     217|       1|   -1|       0| unknown| no|
| 42|entrepreneur|divorced| tertiary|    yes|      2|    yes|  no|unknown|  5
|  may|     380|       1|   -1|       0| unknown| no|
| 58|     retired| married|  primary|     no|    121|    yes|  no|unknown|  5
|  may|      50|       1|   -1|       0| unknown| no|
| 43|  technician|  single|secondary|     no|    593|    yes|  no|unknown|  5
|  may|      55|       1|   -1|       0| unknown| no|
| 41|      admin.|divorced|secondary|     no|    270|    yes|  no|unknown|  5
|  may|     222|       1|   -1|       0| unknown| no|
| 29|      admin.|  single|secondary|     no|    390|    yes|  no|unknown|  5
|  may|     137|       1|   -1|       0| unknown| no|
| 53|  technician| married|secondary|     no|      6|    yes|  no|unknown|  5
|  may|     517|       1|   -1|       0| unknown| no|
| 58|  technician| married|  unknown|     no|     71|    yes|  no|unknown|  5
|  may|      71|       1|   -1|       0| unknown| no|
| 57|    services| married|secondary|     no|    162|    yes|  no|unknown|  5
|  may|     174|       1|   -1|       0| unknown| no|
| 51|     retired| married|  primary|     no|    229|    yes|  no|unknown|  5
|  may|     353|       1|   -1|       0| unknown| no|
| 45|      admin.|  single|  unknown|     no|     13|    yes|  no|unknown|  5
|  may|      98|       1|   -1|       0| unknown| no|
| 57| blue-collar| married|  primary|     no|     52|    yes|  no|unknown|  5
|  may|      38|       1|   -1|       0| unknown| no|
| 60|     retired| married|  primary|     no|     60|    yes|  no|unknown|  5
|  may|     219|       1|   -1|       0| unknown| no|
| 33|    services| married|secondary|     no|      0|    yes|  no|unknown|  5
|  may|      54|       1|   -1|       0| unknown| no|
+---+-----------+--------+---------+-------+-------+-------+----+-------+---
+-----+--------+--------+-----+--------+--------+---+
only showing top 20 rows
```

In [32]:

```
e=df.withColumn("ageT",f.when(( (df.age >= 15) & (df.age <= 30)), 'YOUNG').when(( (df.age
>= 31) & (df.age <= 59)), 'MID').when(df.age >= 60, 'OLD'))
e.show()
```

```
+---+-----------+--------+---------+-------+-------+-------+----+-------+---
+-----+--------+--------+-----+--------+--------+---+-----+
|age|        job| marital|education|default|balance|housing|loan|contact|day
|month|duration|campaign|pdays|previous|poutcome|  y| ageT|
+---+-----------+--------+---------+-------+-------+-------+----+-------+---
+-----+--------+--------+-----+--------+--------+---+-----+
| 58|  management| married| tertiary|     no|   2143|    yes|  no|unknown|  5
|  may|     261|       1|   -1|       0| unknown| no|  MID|
| 44|  technician|  single|secondary|     no|     29|    yes|  no|unknown|  5
|  may|     151|       1|   -1|       0| unknown| no|  MID|
| 33|entrepreneur| married|secondary|     no|      2|    yes| yes|unknown|  5
|  may|      76|       1|   -1|       0| unknown| no|  MID|
| 47| blue-collar| married|  unknown|     no|   1506|    yes|  no|unknown|  5
|  may|      92|       1|   -1|       0| unknown| no|  MID|
| 33|     unknown|  single|  unknown|     no|      1|     no|  no|unknown|  5
|  may|     198|       1|   -1|       0| unknown| no|  MID|
| 35|  management| married| tertiary|     no|    231|    yes|  no|unknown|  5
|  may|     139|       1|   -1|       0| unknown| no|  MID|
| 28|  management|  single| tertiary|     no|    447|    yes| yes|unknown|  5
|  may|     217|       1|   -1|       0| unknown| no|YOUNG|
| 42|entrepreneur|divorced| tertiary|    yes|      2|    yes|  no|unknown|  5
|  may|     380|       1|   -1|       0| unknown| no|  MID|
| 58|     retired| married|  primary|     no|    121|    yes|  no|unknown|  5
|  may|      50|       1|   -1|       0| unknown| no|  MID|
| 43|  technician|  single|secondary|     no|    593|    yes|  no|unknown|  5
|  may|      55|       1|   -1|       0| unknown| no|  MID|
| 41|      admin.|divorced|secondary|     no|    270|    yes|  no|unknown|  5
|  may|     222|       1|   -1|       0| unknown| no|  MID|
| 29|      admin.|  single|secondary|     no|    390|    yes|  no|unknown|  5
|  may|     137|       1|   -1|       0| unknown| no|YOUNG|
| 53|  technician| married|secondary|     no|      6|    yes|  no|unknown|  5
|  may|     517|       1|   -1|       0| unknown| no|  MID|
| 58|  technician| married|  unknown|     no|     71|    yes|  no|unknown|  5
|  may|      71|       1|   -1|       0| unknown| no|  MID|
| 57|    services| married|secondary|     no|    162|    yes|  no|unknown|  5
|  may|     174|       1|   -1|       0| unknown| no|  MID|
| 51|     retired| married|  primary|     no|    229|    yes|  no|unknown|  5
|  may|     353|       1|   -1|       0| unknown| no|  MID|
| 45|      admin.|  single|  unknown|     no|     13|    yes|  no|unknown|  5
|  may|      98|       1|   -1|       0| unknown| no|  MID|
| 57| blue-collar| married|  primary|     no|     52|    yes|  no|unknown|  5
|  may|      38|       1|   -1|       0| unknown| no|  MID|
| 60|     retired| married|  primary|     no|     60|    yes|  no|unknown|  5
|  may|     219|       1|   -1|       0| unknown| no|  OLD|
| 33|    services| married|secondary|     no|      0|    yes|  no|unknown|  5
|  may|      54|       1|   -1|       0| unknown| no|  MID|
+---+-----------+--------+---------+-------+-------+-------+----+-------+---
+-----+--------+--------+-----+--------+--------+---+-----+
only showing top 20 rows
```

In [33]:

```
e.select('ageT','y').show()
```

```
+-----+---+
| ageT|  y|
+-----+---+
|  MID| no|
|  MID| no|
|  MID| no|
|  MID| no|
|  MID| no|
|  MID| no|
|YOUNG| no|
|  MID| no|
|  MID| no|
|  MID| no|
|  MID| no|
|YOUNG| no|
|  MID| no|
|  MID| no|
|  MID| no|
|  MID| no|
|  MID| no|
|  MID| no|
|  OLD| no|
|  MID| no|
+-----+---+
only showing top 20 rows
```

In [34]:

```
e.groupBy('ageT').pivot('y').count().show()
```

```
+-----+-----+----+
| ageT|   no| yes|
+-----+-----+----+
|  MID|32853|3544|
|YOUNG| 5885|1145|
|  OLD| 1184| 600|
+-----+-----+----+
```

In [ ]:

```

```