```
>>> df=sc.read.format('csv')\
...     .option('header','True')\
   .opti...     .option('delimiter',';')\
...     .load('/home/uttam/futurense_hadoop-pyspark/labs/dataset/bankmarket/bankmarketdata.csv')
>>> 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
```

```
>>> success = float(df.filter(df.y =='yes').count() / df.count()*100)
>>> success
11.698480458295547
>>> failure = 100 - success
>>> failure
88.30151954170445
```

```
>>> df.select(max("age"), min("age"), round(mean("age"),2)).show()
+--------+--------+------------------+
|max(age)|min(age)|round(avg(age), 2)|
+--------+--------+------------------+
|      95|      18|             40.94|
+--------+--------+------------------+
```

```
>>> df.select(round(mean("balance"),2).alias("avg_balance"), round(percentile_approx("balance", 0.5),2).alias("median_balance")).show()
+-----------+--------------+
|avg_balance|median_balance|
+-----------+--------------+
|    1362.27|         448.0|
+-----------+--------------+
```

```
>>> df.filter(df.y == 'yes').groupBy("age").count().sort('count', ascending=False).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
```

```
>>> age_cat_dict = {20:'Teen',40:'Young',60:'Middle',80:'Seniors',100:'Seniors'}
age_cat_udf = udf(lambd>>> age_cat_udf = udf(lambda age: age_cat_dict[age],StringType())
>>> df_age_cat = df.select('age','y').filter(df.y == 'yes').withColumn('age_cat', age_cat_udf(ceil(df['age']/20) * 20)).groupBy('age_cat').count()
>>> df_age_cat.show()
+-------+-----+
|age_cat|count|
+-------+-----+
|   Teen|   33|
| Middle| 1830|
|Seniors|  502|
|  Young| 2924|
+-------+-----+
```

```
>>> df.filter(df.y == 'yes').groupBy("marital").count().sort('count', ascending=False).show()
+--------+-----+
| marital|count|
+--------+-----+
| married| 2755|
|  single| 1912|
|divorced|  622|
+--------+-----+
```

```
>>> df.filter(df.y == 'yes').groupBy("age","marital").count().sort('count', ascending=False).show()
+---+-------+-----+
|age|marital|count|
+---+-------+-----+
| 30| single|  151|
| 28| single|  138|
| 29| single|  133|
| 32| single|  124|
| 26| single|  121|
| 34|married|  118|
| 31| single|  111|
| 27| single|  110|
| 35|married|  101|
| 36|married|  100|
| 25| single|   99|
| 37|married|   98|
| 33| single|   97|
| 33|married|   97|
| 32|married|   87|
| 39|married|   87|
| 38|married|   86|
| 35| single|   84|
| 47|married|   83|
| 31|married|   80|
+---+-------+-----+
only showing top 20 rows
```