1)

In [1]:

```
import pyspark
```

In [2]:

```
import pandas as pd
```

In [24]:

```
from pyspark.sql import SparkSession
spark = SparkSession.newSession
from pyspark.sql.types import StructType,StructField, StringType, IntegerType
from pyspark.sql import Row
import pyspark.sql.functions
```

In [25]:

```
import pandas as pd
d=[['Children', 'First', 6, 0],
    ['Children', 'Second', 24, 0],
    ['Children', 'Third', 27, 52],
    ['Men', 'First', 57, 118],
    ['Men', 'Second', 14, 154],
    ['Men', 'Third', 75, 387],
    ['Men', 'Crew', 192, 693],
    ['Women', 'First', 140, 4],
    ['Women', 'Second', 80, 13],
    ['Women', 'Third', 76, 89],
    ['Women', 'Crew', 20, 3 ]]
t=pd.DataFrame(d,columns=['Sex', 'Class', 'Survived', 'Died'])
```

In [26]:

```
t
```

Out[26]:

	Sex	Class	Survived	Died
0	Children	First	6	0
1	Children	Second	24	0
2	Children	Third	27	52
3	Men	First	57	118
4	Men	Second	14	154
5	Men	Third	75	387
6	Men	Crew	192	693
7	Women	First	140	4
8	Women	Second	80	13
9	Women	Third	76	89
10	Women	Crew	20	3

2)

```
In [28]: t=t[t.Class !="Crew"]#deleting the crew
```

```
In [29]: t
```

Out[29]:

	Sex	Class	Survived	Died
0	Children	First	6	0
1	Children	Second	24	0
2	Children	Third	27	52
3	Men	First	57	118
4	Men	Second	14	154
5	Men	Third	75	387
7	Women	First	140	4
8	Women	Second	80	13
9	Women	Third	76	89

3)

```
In [31]: t['Total_num_people']=t["Survived"]+t['Died']
         t.head()
```

```
<ipython-input-31-2fb5877f9f97>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  t['Total_num_people']=t["Survived"]+t['Died']
```

Out[31]:

	Sex	Class	Survived	Died	Total_num_people
0	Children	First	6	0	6
1	Children	Second	24	0	24
2	Children	Third	27	52	79
3	Men	First	57	118	175
4	Men	Second	14	154	168

4)

```
In [32]: del t["Total_num_people"]
```

5)

In [33]:

```
rslt = t[t['Survived'] > 80]
rslt.head(5)
```

Out[33]:

	Sex	Class	Survived	Died
7	Women	First	140	4

In [34]:

```
t["Total"]=t["Survived"]+ t["Died"]
t["Percentage"]=(t["Survived"]/t["Total"])*100
t[t.Percentage>=80]
```

```
<ipython-input-34-2fa27404cdfe>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/sta
ble/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pyd
ata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-c
opy)
  t["Total"]=t["Survived"]+ t["Died"]
<ipython-input-34-2fa27404cdfe>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/sta
ble/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pyd
ata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-c
opy)
  t["Percentage"]=(t["Survived"]/t["Total"])*100
```

Out[34]:

	Sex	Class	Survived	Died	Total	Percentage
0	Children	First	6	0	6	100.000000
1	Children	Second	24	0	24	100.000000
7	Women	First	140	4	144	97.222222
8	Women	Second	80	13	93	86.021505

In [ ]: