In [38]:
```python
import pandas as pd
from pyspark.sql import SparkSession
import findspark
import numpy as np
```

In [2]:
```python
findspark.init()
findspark.find()
```

Out[2]:
```
'/Users/youngjinseo/anaconda3/lib/python3.10/site-packages/pyspark'
```

In [3]:
```python
spark = SparkSession.builder.appName('practice2').config("spark.driver.bindAddress", "127.0.0.1").getOrCreate()
spark
```

```
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/10/02 15:10:59 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

Out[3]:

SparkSession - in-memory

SparkContext

Spark UI

Version	v3.5.2
Master	local[*]
AppName	practice2

In [4]:
```python
df = spark.read.csv("/Users/youngjinseo/Desktop/파이썬/tested.csv",header = True )
df.show()
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
|        892|       0|     3|    Kelly, Mr. James|  male|34.5|    0|    0|          330911| 7.8292| NULL|       Q|
|        893|       1|     3|Wilkes, Mrs. Jame...|female|  47|    1|    0|          363272|      7| NULL|       S|
|        894|       0|     2|Myles, Mr. Thomas...|  male|  62|    0|    0|          240276| 9.6875| NULL|       Q|
|        895|       0|     3|    Wirz, Mr. Albert|  male|  27|    0|    0|          315154| 8.6625| NULL|       S|
|        896|       1|     3|Hirvonen, Mrs. Al...|female|  22|    1|    1|         3101298|12.2875| NULL|       S|
|        897|       0|     3|Svensson, Mr. Joh...|  male|  14|    0|    0|            7538|  9.225| NULL|       S|
|        898|       1|     3|Connolly, Miss. Kate|female|  30|    0|    0|          330972| 7.6292| NULL|       Q|
|        899|       0|     2|Caldwell, Mr. Alb...|  male|  26|    1|    1|          248738|     29| NULL|       S|
|        900|       1|     3|Abrahim, Mrs. Jos...|female|  18|    0|    0|            2657| 7.2292| NULL|       C|
|        901|       0|     3|Davies, Mr. John ...|  male|  21|    2|    0|       A/4 48871|  24.15| NULL|       S|
|        902|       0|     3|    Ilieff, Mr. Ylio|  male|NULL|    0|    0|          349220| 7.8958| NULL|       S|
|        903|       0|     1|Jones, Mr. Charle...|  male|  46|    0|    0|             694|     26| NULL|       S|
|        904|       1|     1|Snyder, Mrs. John...|female|  23|    1|    0|           21228|82.2667|  B45|       S|
|        905|       0|     2|Howard, Mr. Benjamin|  male|  63|    1|    0|           24065|     26| NULL|       S|
|        906|       1|     1|Chaffee, Mrs. Her...|female|  47|    1|    0|     W.E.P. 5734| 61.175|  E31|       S|
|        907|       1|     2|del Carlo, Mrs. S...|female|  24|    1|    0|   SC/PARIS 2167|27.7208| NULL|       C|
|        908|       0|     2|   Keane, Mr. Daniel|  male|  35|    0|    0|          233734|  12.35| NULL|       Q|
|        909|       0|     3|   Assaf, Mr. Gerios|  male|  21|    0|    0|            2692|  7.225| NULL|       C|
|        910|       1|     3|Ilmakangas, Miss....|female|  27|    1|    0|STON/O2. 3101270|  7.925| NULL|       S|
|        911|       1|     3|"Assaf Khalil, Mr...|female|  45|    0|    0|            2696|  7.225| NULL|       C|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
only showing top 20 rows
```

다양한 조건문

In [8]:
```python
# female 관련 데이터 가져오기
df.filter(df.Sex != "male").show(10)
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
|        893|       1|     3|Wilkes, Mrs. Jame...|female|  47|    1|    0|          363272|      7| NULL|       S|
|        896|       1|     3|Hirvonen, Mrs. Al...|female|  22|    1|    1|         3101298|12.2875| NULL|       S|
|        898|       1|     3|Connolly, Miss. Kate|female|  30|    0|    0|          330972| 7.6292| NULL|       Q|
|        900|       1|     3|Abrahim, Mrs. Jos...|female|  18|    0|    0|            2657| 7.2292| NULL|       C|
|        904|       1|     1|Snyder, Mrs. John...|female|  23|    1|    0|           21228|82.2667|  B45|       S|
|        906|       1|     1|Chaffee, Mrs. Her...|female|  47|    1|    0|     W.E.P. 5734| 61.175|  E31|       S|
|        907|       1|     2|del Carlo, Mrs. S...|female|  24|    1|    0|   SC/PARIS 2167|27.7208| NULL|       C|
|        910|       1|     3|Ilmakangas, Miss....|female|  27|    1|    0|STON/O2. 3101270|  7.925| NULL|       S|
|        911|       1|     3|"Assaf Khalil, Mr...|female|  45|    0|    0|            2696|  7.225| NULL|       C|
|        914|       1|     1|Flegenheim, Mrs. ...|female|NULL|    0|    0|        PC 17598|31.6833| NULL|       S|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
only showing top 10 rows
```

In [11]:
```python
# Age 30보다 같거나 큰 데이터 추출
df.filter(df.Age >= 30).show(10)
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+-----------+------+-----+--------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|     Ticket|  Fare|Cabin|Embarked|
+-----------+--------+------+--------------------+------+----+-----+-----+-----------+------+-----+--------+
|        892|       0|     3|    Kelly, Mr. James|  male|34.5|    0|    0|     330911|7.8292| NULL|       Q|
|        893|       1|     3|Wilkes, Mrs. Jame...|female|  47|    1|    0|     363272|     7| NULL|       S|
|        894|       0|     2|Myles, Mr. Thomas...|  male|  62|    0|    0|     240276|9.6875| NULL|       Q|
|        898|       1|     3|Connolly, Miss. Kate|female|  30|    0|    0|     330972|7.6292| NULL|       Q|
|        903|       0|     1|Jones, Mr. Charle...|  male|  46|    0|    0|        694|    26| NULL|       S|
|        905|       0|     2|Howard, Mr. Benjamin|  male|  63|    1|    0|      24065|    26| NULL|       S|
|        906|       1|     1|Chaffee, Mrs. Her...|female|  47|    1|    0|W.E.P. 5734|61.175|  E31|       S|
|        908|       0|     2|   Keane, Mr. Daniel|  male|  35|    0|    0|     233734| 12.35| NULL|       Q|
|        911|       1|     3|"Assaf Khalil, Mr...|female|  45|    0|    0|       2696| 7.225| NULL|       C|
|        912|       0|     1|Rothschild, Mr. M...|  male|  55|    1|    0|   PC 17603|  59.4| NULL|       C|
+-----------+--------+------+--------------------+------+----+-----+-----+-----------+------+-----+--------+
only showing top 10 rows
```

조건문의 and or 결합

In [18]:
```python
# 성별은 여자 그리고 나이는 30보다 작은 데이터 추출
df.filter((df.Sex == 'female')&(df.Age <= 30)).show(10)
```

```
+-----------+--------+------+--------------------+------+---+-----+-----+----------------+-------+-----+--------+
|PassengerId|Survived|Pclass|                Name|   Sex|Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|
+-----------+--------+------+--------------------+------+---+-----+-----+----------------+-------+-----+--------+
|        896|       1|     3|Hirvonen, Mrs. Al...|female| 22|    1|    1|         3101298|12.2875| NULL|       S|
|        898|       1|     3|Connolly, Miss. Kate|female| 30|    0|    0|          330972| 7.6292| NULL|       Q|
|        900|       1|     3|Abrahim, Mrs. Jos...|female| 18|    0|    0|            2657| 7.2292| NULL|       C|
|        904|       1|     1|Snyder, Mrs. John...|female| 23|    1|    0|           21228|82.2667|  B45|       S|
|        907|       1|     2|del Carlo, Mrs. S...|female| 24|    1|    0|   SC/PARIS 2167|27.7208| NULL|       C|
|        910|       1|     3|Ilmakangas, Miss....|female| 27|    1|    0|STON/O2. 3101270|  7.925| NULL|       S|
|        918|       1|     1|Ostby, Miss. Hele...|female| 22|    0|    1|          113509|61.9792|  B36|       C|
|        929|       1|     3|  Cacic, Miss. Manda|female| 21|    0|    0|          315087| 8.6625| NULL|       S|
|        935|       1|     2|Corbett, Mrs. Wal...|female| 30|    0|    0|          237249|     13| NULL|       S|
|        944|       1|     2|Hocking, Miss. El...|female| 20|    2|    1|           29105|     23| NULL|       S|
+-----------+--------+------+--------------------+------+---+-----+-----+----------------+-------+-----+--------+
only showing top 10 rows
```

In [20]:
```python
#Embarked가 S이고 나이는 20보다 같거나 작은 데이터 추출
df.filter((df.Embarked == 'S') & (df.Age <= 20)).show(10)
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+---------+------+-----+--------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|   Ticket|  Fare|Cabin|Embarked|
+-----------+--------+------+--------------------+------+----+-----+-----+---------+------+-----+--------+
|        897|       0|     3|Svensson, Mr. Joh...|  male|  14|    0|    0|     7538| 9.225| NULL|       S|
|        913|       0|     3|Olsen, Master. Ar...|  male|   9|    0|    1| C 17368|3.1708| NULL|       S|
|        944|       1|     2|Hocking, Miss. El...|female|  20|    2|    1|    29105|    23| NULL|       S|
|        952|       0|     3|     Dika, Mr. Mirko|  male|  17|    0|    0|   349232|7.8958| NULL|       S|
|        954|       0|     3|Bjorklund, Mr. Er...|  male|  18|    0|    0|   347090|  7.75| NULL|       S|
|        979|       1|     3|Badman, Miss. Emi...|female|  18|    0|    0|A/4 31416|  8.05| NULL|       S|
|        981|       0|     2|Wells, Master. Ra...|  male|   2|    1|    1|    29103|    23| NULL|       S|
|        990|       1|     3|Braf, Miss. Elin ...|female|  20|    0|    0|   347471|7.8542| NULL|       S|
|       1001|       0|     2|   Swane, Mr. George|  male|18.5|    0|    0|   248734|    13|    F|       S|
|       1009|       1|     3|Sandstrom, Miss. ...|female|   1|    1|    1|  PP 9549|  16.7|   G6|       S|
+-----------+--------+------+--------------------+------+----+-----+-----+---------+------+-----+--------+
only showing top 10 rows
```

SQL의 like 문장 사용

In [29]:
```python
# Name에서 S로 시작하는 이름
df.filter(df.Name.like("S%")).show(10)
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+--------+--------+-------+--------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|  Ticket|    Fare|  Cabin|Embarked|
+-----------+--------+------+--------------------+------+----+-----+-----+--------+--------+-------+--------+
|        897|       0|     3|Svensson, Mr. Joh...|  male|  14|    0|    0|    7538|   9.225|   NULL|       S|
|        904|       1|     1|Snyder, Mrs. John...|female|  23|    1|    0|   21228| 82.2667|    B45|       S|
|        921|       0|     3|   Samaan, Mr. Elias|  male|NULL|    2|    0|    2662| 21.6792|   NULL|       C|
|        930|       0|     3|     Sap, Mr. Julius|  male|  25|    0|    0|  345768|     9.5|   NULL|       S|
|        939|       0|     3|Shaughnessy, Mr. ...|  male|NULL|    0|    0|  370374|    7.75|   NULL|       Q|
|        942|       0|     1|Smith, Mr. Lucien...|  male|  24|    1|    0|   13695|      60|    C31|       S|
|        973|       0|     1|  Straus, Mr. Isidor|  male|  67|    1|    0|PC 17483|221.7792|C55 C57|       S|
|        992|       1|     1|Stengel, Mrs. Cha...|female|  43|    1|    0|   11778| 55.4417|   C116|       C|
|       1001|       0|     2|   Swane, Mr. George|  male|18.5|    0|    0|  248734|      13|      F|       S|
|       1002|       0|     2|Stanton, Mr. Samu...|  male|  41|    0|    0|  237734| 15.0458|   NULL|       C|
+-----------+--------+------+--------------------+------+----+-----+-----+--------+--------+-------+--------+
only showing top 10 rows
```

In [32]:
```python
#Name에서 중간에 'Mrs' 포함된 데이터 추출
df.filter(df.Name.like("%Mrs%")).show(10)
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+-------------+-------+---------------+--------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|       Ticket|   Fare|          Cabin|Embarked|
+-----------+--------+------+--------------------+------+----+-----+-----+-------------+-------+---------------+--------+
|        893|       1|     3|Wilkes, Mrs. Jame...|female|  47|    1|    0|       363272|      7|           NULL|       S|
|        896|       1|     3|Hirvonen, Mrs. Al...|female|  22|    1|    1|      3101298|12.2875|           NULL|       S|
|        900|       1|     3|Abrahim, Mrs. Jos...|female|  18|    0|    0|         2657| 7.2292|           NULL|       C|
|        904|       1|     1|Snyder, Mrs. John...|female|  23|    1|    0|        21228|82.2667|            B45|       S|
|        906|       1|     1|Chaffee, Mrs. Her...|female|  47|    1|    0|  W.E.P. 5734| 61.175|            E31|       S|
|        907|       1|     2|del Carlo, Mrs. S...|female|  24|    1|    0|SC/PARIS 2167|27.7208|           NULL|       C|
|        911|       1|     3|"Assaf Khalil, Mr...|female|  45|    0|    0|         2696|  7.225|           NULL|       C|
|        914|       1|     1|Flegenheim, Mrs. ...|female|NULL|    0|    0|     PC 17598|31.6833|           NULL|       S|
|        916|       1|     1|Ryerson, Mrs. Art...|female|  48|    1|    3|     PC 17608|262.375|B57 B59 B63 B66|       C|
|        924|       1|     3|Dean, Mrs. Bertra...|female|  33|    1|    2|    C.A. 2315| 20.575|           NULL|       S|
+-----------+--------+------+--------------------+------+----+-----+-----+-------------+-------+---------------+--------+
only showing top 10 rows
```

df의 값이 바깥에 존재하는 list 내에 존재하는지 확인

In [34]:
```python
# 특정 탑승 항구 값을 필터링하여 데이터프레임에서 10개의 레코드를 보여주는 코드
# 'S'와 'C' 값이 포함된 리스트 embark를 생성

embark = ['S','C']
df.filter(df.Embarked.isin(embark)).show(10)
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+---------+-------+-----+--------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|   Ticket|   Fare|Cabin|Embarked|
+-----------+--------+------+--------------------+------+----+-----+-----+---------+-------+-----+--------+
|        893|       1|     3|Wilkes, Mrs. Jame...|female|  47|    1|    0|   363272|      7| NULL|       S|
|        895|       0|     3|    Wirz, Mr. Albert|  male|  27|    0|    0|   315154| 8.6625| NULL|       S|
|        896|       1|     3|Hirvonen, Mrs. Al...|female|  22|    1|    1|  3101298|12.2875| NULL|       S|
|        897|       0|     3|Svensson, Mr. Joh...|  male|  14|    0|    0|     7538|  9.225| NULL|       S|
|        899|       0|     2|Caldwell, Mr. Alb...|  male|  26|    1|    1|   248738|     29| NULL|       S|
|        900|       1|     3|Abrahim, Mrs. Jos...|female|  18|    0|    0|     2657| 7.2292| NULL|       C|
|        901|       0|     3|Davies, Mr. John ...|  male|  21|    2|    0|A/4 48871|  24.15| NULL|       S|
|        902|       0|     3|    Ilieff, Mr. Ylio|  male|NULL|    0|    0|   349220| 7.8958| NULL|       S|
|        903|       0|     1|Jones, Mr. Charle...|  male|  46|    0|    0|      694|     26| NULL|       S|
|        904|       1|     1|Snyder, Mrs. John...|female|  23|    1|    0|    21228|82.2667|  B45|       S|
+-----------+--------+------+--------------------+------+----+-----+-----+---------+-------+-----+--------+
only showing top 10 rows
```

In [39]:
```python
# 20에서 40 사이의 숫자들을 2씩 증가시키는 리스트 생성
# DataFrame에서 Age 열의 값이 age_arrange 리스트에 있는 값들로 필터링

age_arrange = list(np.arange(20, 40, 2))
df.filter(df.Age.isin(age_arrange)).show(10)
```

```
+-----------+--------+------+--------------------+------+---+-----+-----+-------------+-------+-----+--------+
|PassengerId|Survived|Pclass|                Name|   Sex|Age|SibSp|Parch|       Ticket|   Fare|Cabin|Embarked|
+-----------+--------+------+--------------------+------+---+-----+-----+-------------+-------+-----+--------+
|        896|       1|     3|Hirvonen, Mrs. Al...|female| 22|    1|    1|      3101298|12.2875| NULL|       S|
|        898|       1|     3|Connolly, Miss. Kate|female| 30|    0|    0|       330972| 7.6292| NULL|       Q|
|        899|       0|     2|Caldwell, Mr. Alb...|  male| 26|    1|    1|       248738|     29| NULL|       S|
|        907|       1|     2|del Carlo, Mrs. S...|female| 24|    1|    0|SC/PARIS 2167|27.7208| NULL|       C|
|        918|       1|     1|Ostby, Miss. Hele...|female| 22|    0|    1|       113509|61.9792|  B36|       C|
|        923|       0|     2|Jefferys, Mr. Cli...|  male| 24|    2|    0|   C.A. 31029|   31.5| NULL|       S|
|        926|       0|     1|Mock, Mr. Philipp...|  male| 30|    1|    0|        13236|  57.75|  C78|       C|
|        935|       1|     2|Corbett, Mrs. Wal...|female| 30|    0|    0|       237249|     13| NULL|       S|
|        941|       1|     3|"Coutts, Mrs. Wil...|female| 36|    0|    2|   C.A. 37671|   15.9| NULL|       S|
|        942|       0|     1|Smith, Mr. Lucien...|  male| 24|    1|    0|        13695|     60|  C31|       S|
+-----------+--------+------+--------------------+------+---+-----+-----+-------------+-------+-----+--------+
only showing top 10 rows
```

조건문장으로 시작하는지, 끝나는지 아니면 조건문장을 포함하는지 확인

In [40]:
```python
#M 시작한 이름 추출
df.filter(df.Name.startswith("M")).show(10)
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
|        894|       0|     2|Myles, Mr. Thomas...|  male|  62|    0|    0|          240276| 9.6875| NULL|       Q|
|        926|       0|     1|Mock, Mr. Philipp...|  male|  30|    1|    0|           13236|  57.75|  C78|       C|
|        946|       0|     2|Mangiavacchi, Mr....|  male|NULL|    0|    0|     SC/A.3 2861|15.5792| NULL|       C|
|        953|       0|     2|McCrae, Mr. Arthu...|  male|  32|    0|    0|          237216|   13.5| NULL|       S|
|        959|       0|     1|Moore, Mr. Claren...|  male|  47|    0|    0|          113796|   42.4| NULL|       S|
|        962|       1|     3|Mulvihill, Miss. ...|female|  24|    0|    0|          382653|   7.75| NULL|       Q|
|        963|       0|     3|  Minkoff, Mr. Lazar|  male|  21|    0|    0|          349211| 7.8958| NULL|       S|
|        968|       0|     3|    Miles, Mr. Frank|  male|NULL|    0|    0|          359306|   8.05| NULL|       S|
|        989|       0|     3|Makinen, Mr. Kall...|  male|  29|    0|    0|STON/O 2. 3101268|  7.925| NULL|       S|
|       1019|       1|     3| McCoy, Miss. Alicia|female|NULL|    2|    0|          367226|  23.25| NULL|       Q|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
only showing top 10 rows
```

In [42]:
```python
# 끝자리 s로 시작하는 데이터 추출
df.filter(df.Name.endswith("s")).show(10)
```

```
+-----------+--------+------+--------------------+----+----+-----+-----+-----------------+-------+-----+--------+
|PassengerId|Survived|Pclass|                Name| Sex| Age|SibSp|Parch|           Ticket|   Fare|Cabin|Embarked|
+-----------+--------+------+--------------------+----+----+-----+-----+-----------------+-------+-----+--------+
|        892|       0|     3|    Kelly, Mr. James|male|34.5|    0|    0|           330911| 7.8292| NULL|       Q|
|        894|       0|     2|Myles, Mr. Thomas...|male|  62|    0|    0|           240276| 9.6875| NULL|       Q|
|        899|       0|     2|Caldwell, Mr. Alb...|male|  26|    1|    1|           248738|     29| NULL|       S|
|        909|       0|     3|   Assaf, Mr. Gerios|male|  21|    0|    0|             2692|  7.225| NULL|       C|
|        921|       0|     3|   Samaan, Mr. Elias|male|NULL|    2|    0|             2662|21.6792| NULL|       C|
|        923|       0|     2|Jefferys, Mr. Cli...|male|  24|    2|    0|       C.A. 31029|   31.5| NULL|       S|
|        930|       0|     3|     Sap, Mr. Julius|male|  25|    0|    0|           345768|    9.5| NULL|       S|
|        937|       0|     3|Peltomaki, Mr. Ni...|male|  25|    0|    0|STON/O 2. 3101291|  7.925| NULL|       S|
|        970|       0|     2|Aldworth, Mr. Cha...|male|  30|    0|    0|           248744|     13| NULL|       S|
|        977|       0|     3|  Khalil, Mr. Betros|male|NULL|    1|    0|             2660|14.4542| NULL|       C|
+-----------+--------+------+--------------------+----+----+-----+-----+-----------------+-------+-----+--------+
only showing top 10 rows
```

In [43]:
```python
# Mr 포함된 데이터 추출
df.filter(df.Name.contains("Mr")).show(10)
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+---------+-------+-----+--------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|   Ticket|   Fare|Cabin|Embarked|
+-----------+--------+------+--------------------+------+----+-----+-----+---------+-------+-----+--------+
|        892|       0|     3|    Kelly, Mr. James|  male|34.5|    0|    0|   330911| 7.8292| NULL|       Q|
|        893|       1|     3|Wilkes, Mrs. Jame...|female|  47|    1|    0|   363272|      7| NULL|       S|
|        894|       0|     2|Myles, Mr. Thomas...|  male|  62|    0|    0|   240276| 9.6875| NULL|       Q|
|        895|       0|     3|    Wirz, Mr. Albert|  male|  27|    0|    0|   315154| 8.6625| NULL|       S|
|        896|       1|     3|Hirvonen, Mrs. Al...|female|  22|    1|    1|  3101298|12.2875| NULL|       S|
|        897|       0|     3|Svensson, Mr. Joh...|  male|  14|    0|    0|     7538|  9.225| NULL|       S|
|        899|       0|     2|Caldwell, Mr. Alb...|  male|  26|    1|    1|   248738|     29| NULL|       S|
|        900|       1|     3|Abrahim, Mrs. Jos...|female|  18|    0|    0|     2657| 7.2292| NULL|       C|
|        901|       0|     3|Davies, Mr. John ...|  male|  21|    2|    0|A/4 48871|  24.15| NULL|       S|
|        902|       0|     3|    Ilieff, Mr. Ylio|  male|NULL|    0|    0|   349220| 7.8958| NULL|       S|
+-----------+--------+------+--------------------+------+----+-----+-----+---------+-------+-----+--------+
only showing top 10 rows
```

Null 값 & 아닌 값 추출

In [47]:
```python
#결측값 데이터 추출
df.filter(df.Cabin.isNull()).show(10)
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+---------+-------+-----+--------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|   Ticket|   Fare|Cabin|Embarked|
+-----------+--------+------+--------------------+------+----+-----+-----+---------+-------+-----+--------+
|        892|       0|     3|    Kelly, Mr. James|  male|34.5|    0|    0|   330911| 7.8292| NULL|       Q|
|        893|       1|     3|Wilkes, Mrs. Jame...|female|  47|    1|    0|   363272|      7| NULL|       S|
|        894|       0|     2|Myles, Mr. Thomas...|  male|  62|    0|    0|   240276| 9.6875| NULL|       Q|
|        895|       0|     3|    Wirz, Mr. Albert|  male|  27|    0|    0|   315154| 8.6625| NULL|       S|
|        896|       1|     3|Hirvonen, Mrs. Al...|female|  22|    1|    1|  3101298|12.2875| NULL|       S|
|        897|       0|     3|Svensson, Mr. Joh...|  male|  14|    0|    0|     7538|  9.225| NULL|       S|
|        898|       1|     3|Connolly, Miss. Kate|female|  30|    0|    0|   330972| 7.6292| NULL|       Q|
|        899|       0|     2|Caldwell, Mr. Alb...|  male|  26|    1|    1|   248738|     29| NULL|       S|
|        900|       1|     3|Abrahim, Mrs. Jos...|female|  18|    0|    0|     2657| 7.2292| NULL|       C|
|        901|       0|     3|Davies, Mr. John ...|  male|  21|    2|    0|A/4 48871|  24.15| NULL|       S|
+-----------+--------+------+--------------------+------+----+-----+-----+---------+-------+-----+--------+
only showing top 10 rows
```

In [46]:
```python
#결측값 포함되지 않는 데이터 추출
df.filter(df.Cabin.isNotNull()).show(10)
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+-----------+-------+---------------+--------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|     Ticket|   Fare|          Cabin|Embarked|
+-----------+--------+------+--------------------+------+----+-----+-----+-----------+-------+---------------+--------+
|        904|       1|     1|Snyder, Mrs. John...|female|  23|    1|    0|      21228|82.2667|            B45|       S|
|        906|       1|     1|Chaffee, Mrs. Her...|female|  47|    1|    0|W.E.P. 5734| 61.175|            E31|       S|
|        916|       1|     1|Ryerson, Mrs. Art...|female|  48|    1|    3|   PC 17608|262.375|B57 B59 B63 B66|       C|
|        918|       1|     1|Ostby, Miss. Hele...|female|  22|    0|    1|     113509|61.9792|            B36|       C|
|        920|       0|     1|Brady, Mr. John B...|  male|  41|    0|    0|     113054|   30.5|            A21|       S|
|        926|       0|     1|Mock, Mr. Philipp...|  male|  30|    1|    0|      13236|  57.75|            C78|       C|
|        933|       0|     1|Franklin, Mr. Tho...|  male|NULL|    0|    0|     113778|  26.55|            D34|       S|
|        936|       1|     1|Kimball, Mrs. Edw...|female|  45|    1|    0|      11753|52.5542|            D19|       S|
|        938|       0|     1|Chevre, Mr. Paul ...|  male|  45|    0|    0|   PC 17594|   29.7|             A9|       C|
|        940|       1|     1|Bucknell, Mrs. Wi...|female|  60|    0|    0|      11813|76.2917|            D15|       C|
+-----------+--------+------+--------------------+------+----+-----+-----+-----------+-------+---------------+--------+
only showing top 10 rows
```

유일한 값만 필터링

In [53]:
```python
# 'Survived' 열에서 중복된 값을 제거하고, 고유한 값만 출력하는 코드
df.select('Survived').distinct().show()
```

```
+--------+
|Survived|
+--------+
|       0|
|       1|
+--------+
```

기타

In [49]:
```python
## 'Name', 'Sex', 'Age', 'Cabin' 열만 선택하고, 나이가 20 이하인 승객을 필터링한 후, 10개의 레코드를 출력하는 코드
df.select('Name','Sex','Age','Cabin').filter(df.Age <=20).show(10)
```

```
+--------------------+------+----+---------------+
|                Name|   Sex| Age|          Cabin|
+--------------------+------+----+---------------+
|Svensson, Mr. Joh...|  male|  14|           NULL|
|Abrahim, Mrs. Jos...|female|  18|           NULL|
|Olsen, Master. Ar...|  male|   9|           NULL|
|"Katavelas, Mr. V...|  male|18.5|           NULL|
|Hocking, Miss. El...|female|  20|           NULL|
|Rice, Master. Albert|  male|  10|           NULL|
|     Dika, Mr. Mirko|  male|  17|           NULL|
|Bjorklund, Mr. Er...|  male|  18|           NULL|
|Ryerson, Master. ...|  male|  13|B57 B59 B63 B66|
|Burns, Miss. Mary...|female|  18|           NULL|
+--------------------+------+----+---------------+
only showing top 10 rows
```

In [50]:
```python
# 'Name', 'Age', 'Survived' 열만 선택하고, 생존한 승객을 필터링한 후 10개의 레코드를 출력하는 코드
df.select('Name','Age','Survived').filter(df.Survived == '1').show(10)
```

```
+--------------------+----+--------+
|                Name| Age|Survived|
+--------------------+----+--------+
|Wilkes, Mrs. Jame...|  47|       1|
|Hirvonen, Mrs. Al...|  22|       1|
|Connolly, Miss. Kate|  30|       1|
|Abrahim, Mrs. Jos...|  18|       1|
|Snyder, Mrs. John...|  23|       1|
|Chaffee, Mrs. Her...|  47|       1|
|del Carlo, Mrs. S...|  24|       1|
|Ilmakangas, Miss....|  27|       1|
|"Assaf Khalil, Mr...|  45|       1|
|Flegenheim, Mrs. ...|NULL|       1|
+--------------------+----+--------+
only showing top 10 rows
```