In [1]:
```
from pyspark.sql import SparkSession
import findspark
```

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

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

데이터셋 불러오기

In [3]:
```
spark = SparkSession.builder.appName('practice1').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).
24/09/18 11:46:42 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	`practice1`

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

데이터셋 행 읽기

In [6]:
```
df.show(13)
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+---------+-------+-----+--------+
|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|
+-----------+--------+------+--------------------+------+----+-----+-----+---------+-------+-----+--------+
only showing top 13 rows
```

데이터 선택하기

In [8]:
```
select_df = df.select('Name','Sex')
select_df.show()
```

```
+--------------------+------+
|                Name|   Sex|
+--------------------+------+
|    Kelly, Mr. James|  male|
|Wilkes, Mrs. Jame...|female|
|Myles, Mr. Thomas...|  male|
|    Wirz, Mr. Albert|  male|
|Hirvonen, Mrs. Al...|female|
|Svensson, Mr. Joh...|  male|
|Connolly, Miss. Kate|female|
|Caldwell, Mr. Alb...|  male|
|Abrahim, Mrs. Jos...|female|
|Davies, Mr. John ...|  male|
|    Ilieff, Mr. Ylio|  male|
|Jones, Mr. Charle...|  male|
|Snyder, Mrs. John...|female|
|Howard, Mr. Benjamin|  male|
|Chaffee, Mrs. Her...|female|
|del Carlo, Mrs. S...|female|
|   Keane, Mr. Daniel|  male|
|   Assaf, Mr. Gerios|  male|
|Ilmakangas, Miss....|female|
|"Assaf Khalil, Mr...|female|
+--------------------+------+
only showing top 20 rows
```

데이터 필터

In [9]:
```
filter_df = df.filter(df.Sex == 'male')
filter_df.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|
|        895|       0|     3|    Wirz, Mr. Albert|male|  27|    0|    0|   315154|8.6625| 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|
|        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|
|        905|       0|     2|Howard, Mr. Benjamin|male|  63|    1|    0|    24065|    26| NULL|       S|
|        908|       0|     2|   Keane, Mr. Daniel|male|  35|    0|    0|   233734| 12.35| NULL|       Q|
+-----------+--------+------+--------------------+----+----+-----+-----+---------+------+-----+--------+
only showing top 10 rows
```

데이터 열 삭제

In [10]:
```
drop_df = df.drop('Fare','Cabin')
drop_df.show()
```

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

새로 데이터셋 저장

In [13]:
```
drop_df.write.option("header",True).csv('new_dataset.csv')
```