Practical 4: Working with PySpark

```
!pip install pyspark
```

```
import pyspark

import pandas as pd

pd.read_csv('data.csv')
```

	Name	Age	Experience	Salary
0	Willetta	27.0	52.0	62860.0
1	Merrie	45.0	34.0	57591.0
2	Joleen	43.0	42.0	61343.0
3	Nananne	25.0	33.0	75478.0
4	Jennica	56.0	10.0	71299.0
5	Lizzie	55.0	27.0	62050.0
6	Kaja	28.0	18.0	52610.0
7	Aubrie	55.0	18.0	77702.0
8	Ginnie	71.0	26.0	88700.0
9	Correy	47.0	25.0	83288.0
10	Rochette	24.0	12.0	56123.0
11	Hayley	55.0	51.0	72758.0
12	Mathilda	42.0	44.0	62825.0
13	Veda	61.0	31.0	68511.0
14	Jessy	59.0	41.0	85482.0
15	John	30.0	5.0	50000.0
16	Mike	25.0	NaN	NaN
17	NaN	40.0	NaN	NaN
18	Sarah	NaN	3.0	40000.0
19	Sarah	NaN	NaN	NaN

```
from pyspark.sql import SparkSession

spark = SparkSession.builder.appName('Practise').getOrCreate()

# inferSchema determines the datatype of each column.

df = spark.read.csv('data.csv', header=True, inferSchema=True)

df.show()
```

```
+--------+----+----------+------+
|    Name| Age|Experience|Salary|
+--------+----+----------+------+
|Willetta|  27|        52| 62860|
|  Merrie|  45|        34| 57591|
|  Joleen|  43|        42| 61343|
| Nananne|  25|        33| 75478|
| Jennica|  56|        10| 71299|
|  Lizzie|  55|        27| 62050|
|    Kaja|  28|        18| 52610|
|  Aubrie|  55|        18| 77702|
|  Ginnie|  71|        26| 88700|
|  Correy|  47|        25| 83288|
|Rochette|  24|        12| 56123|
|  Hayley|  55|        51| 72758|
|Mathilda|  42|        44| 62825|
|    Veda|  61|        31| 68511|
|   Jessy|  59|        41| 85482|
|    John|  30|         5| 50000|
|    Mike|  25|      NULL|  NULL|
|    NULL|  40|      NULL|  NULL|
|   Sarah|NULL|         3| 40000|
|   Sarah|NULL|      NULL|  NULL|
+--------+----+----------+------+
```

```
# Type of the data

print(type(df))
```

```
<class 'pyspark.sql.dataframe.DataFrame'>
```

```
# Check Schema

df.printSchema()
```

```
root
 |-- Name: string (nullable = true)
 |-- Age: integer (nullable = true)
 |-- Experience: integer (nullable = true)
 |-- Salary: integer (nullable = true)
```

```
# Get Column Names

df.columns
```

```
['Name', 'Age', 'Experience', 'Salary']
```

```
# First 3 Data Values

df.head(3)
```

```
[Row(Name='Willetta', Age=27, Experience=52, Salary=62860),
 Row(Name='Merrie', Age=45, Experience=34, Salary=57591),
 Row(Name='Joleen', Age=43, Experience=42, Salary=61343)]
```

```
# Selecting Specific Columns

df.select(['Name', 'Age']).show()
```

```
+--------+----+
|    Name| Age|
+--------+----+
|Willetta|  27|
|  Merrie|  45|
|  Joleen|  43|
| Nananne|  25|
| Jennica|  56|
|  Lizzie|  55|
|    Kaja|  28|
|  Aubrie|  55|
|  Ginnie|  71|
|  Correy|  47|
|Rochette|  24|
|  Hayley|  55|
|Mathilda|  42|
|    Veda|  61|
|   Jessy|  59|
|    John|  30|
|    Mike|  25|
|    NULL|  40|
|   Sarah|NULL|
|   Sarah|NULL|
+--------+----+
```

```
# Check Datatypes

df.dtypes
```

```
[('Name', 'string'), ('Age', 'int'), ('Experience', 'int'), ('Salary', 'int')]
```

```
# Describe dataset

df.describe().show()
```

```
+-------+--------+-----------------+-----------------+-----------------+
|summary|    Name|              Age|       Experience|           Salary|
+-------+--------+-----------------+-----------------+-----------------+
|  count|      19|               18|               17|               17|
|   mean|    NULL|43.77777777777778|27.764705882352942|66389.41176470589|
| stddev|    NULL|14.63901135988258| 15.29513571272214|13310.58934673266|
|    min|  Aubrie|               24|                3|            40000|
|    max|Willetta|               71|               52|            88700|
+-------+--------+-----------------+-----------------+-----------------+
```

```
# Adding columns in data frame

df = df.withColumn('Experience After 2 Years', df['Experience'] + 2)

df.show()
```

```
+--------+----+----------+------+-----------------------+
|    Name| Age|Experience|Salary|Experience After 2 Years|
+--------+----+----------+------+-----------------------+
|Willetta|  27|        52| 62860|                     54|
|  Merrie|  45|        34| 57591|                     36|
|  Joleen|  43|        42| 61343|                     44|
| Nananne|  25|        33| 75478|                     35|
| Jennica|  56|        10| 71299|                     12|
|  Lizzie|  55|        27| 62050|                     29|
|    Kaja|  28|        18| 52610|                     20|
|  Aubrie|  55|        18| 77702|                     20|
|  Ginnie|  71|        26| 88700|                     28|
|  Correy|  47|        25| 83288|                     27|
|Rochette|  24|        12| 56123|                     14|
|  Hayley|  55|        51| 72758|                     53|
|Mathilda|  42|        44| 62825|                     46|
|    Veda|  61|        31| 68511|                     33|
|   Jessy|  59|        41| 85482|                     43|
|    John|  30|         5| 50000|                      7|
|    Mike|  25|      NULL|  NULL|                   NULL|
|    NULL|  40|      NULL|  NULL|                   NULL|
|   Sarah|NULL|         3| 40000|                      5|
|   Sarah|NULL|      NULL|  NULL|                   NULL|
+--------+----+----------+------+-----------------------+
```

```
# Drop the column

df = df.drop('Experience After 2 Years')

df.show()
```

```
+--------+----+----------+------+
|    Name| Age|Experience|Salary|
+--------+----+----------+------+
|Willetta|  27|        52| 62860|
|  Merrie|  45|        34| 57591|
|  Joleen|  43|        42| 61343|
| Nananne|  25|        33| 75478|
| Jennica|  56|        10| 71299|
|  Lizzie|  55|        27| 62050|
|    Kaja|  28|        18| 52610|
|  Aubrie|  55|        18| 77702|
|  Ginnie|  71|        26| 88700|
|  Correy|  47|        25| 83288|
|Rochette|  24|        12| 56123|
|  Hayley|  55|        51| 72758|
|Mathilda|  42|        44| 62825|
|    Veda|  61|        31| 68511|
|   Jessy|  59|        41| 85482|
|    John|  30|         5| 50000|
|    Mike|  25|      NULL|  NULL|
|    NULL|  40|      NULL|  NULL|
|   Sarah|NULL|         3| 40000|
|   Sarah|NULL|      NULL|  NULL|
+--------+----+----------+------+
```

```
# Rename the columns
```

```
df.withColumnRenamed('Name', 'New Name').show()
```

```
+--------+----+----------+------+
|New Name| Age|Experience|Salary|
+--------+----+----------+------+
|Willetta|  27|        52| 62860|
|  Merrie|  45|        34| 57591|
|  Joleen|  43|        42| 61343|
| Nananne|  25|        33| 75478|
| Jennica|  56|        10| 71299|
|  Lizzie|  55|        27| 62050|
|    Kaja|  28|        18| 52610|
|  Aubrie|  55|        18| 77702|
|  Ginnie|  71|        26| 88700|
|  Correy|  47|        25| 83288|
|Rochette|  24|        12| 56123|
|  Hayley|  55|        51| 72758|
|Mathilda|  42|        44| 62825|
|    Veda|  61|        31| 68511|
|   Jessy|  59|        41| 85482|
|    John|  30|         5| 50000|
|    Mike|  25|      NULL|  NULL|
|    NULL|  40|      NULL|  NULL|
|   Sarah|NULL|         3| 40000|
|   Sarah|NULL|      NULL|  NULL|
+--------+----+----------+------+
```

```
df.na.drop().show()
```

```
+--------+---+----------+------+
|    Name|Age|Experience|Salary|
+--------+---+----------+------+
|Willetta| 27|        52| 62860|
|  Merrie| 45|        34| 57591|
|  Joleen| 43|        42| 61343|
| Nananne| 25|        33| 75478|
| Jennica| 56|        10| 71299|
|  Lizzie| 55|        27| 62050|
|    Kaja| 28|        18| 52610|
|  Aubrie| 55|        18| 77702|
|  Ginnie| 71|        26| 88700|
|  Correy| 47|        25| 83288|
|Rochette| 24|        12| 56123|
|  Hayley| 55|        51| 72758|
|Mathilda| 42|        44| 62825|
|    Veda| 61|        31| 68511|
|   Jessy| 59|        41| 85482|
|    John| 30|         5| 50000|
+--------+---+----------+------+
```

```
df.na.drop(how="all").show()
```

```
+--------+----+----------+------+
|    Name| Age|Experience|Salary|
+--------+----+----------+------+
|Willetta|  27|        52| 62860|
|  Merrie|  45|        34| 57591|
|  Joleen|  43|        42| 61343|
| Nananne|  25|        33| 75478|
| Jennica|  56|        10| 71299|
|  Lizzie|  55|        27| 62050|
|    Kaja|  28|        18| 52610|
|  Aubrie|  55|        18| 77702|
|  Ginnie|  71|        26| 88700|
|  Correy|  47|        25| 83288|
|Rochette|  24|        12| 56123|
|  Hayley|  55|        51| 72758|
|Mathilda|  42|        44| 62825|
|    Veda|  61|        31| 68511|
|   Jessy|  59|        41| 85482|
|    John|  30|         5| 50000|
|    Mike|  25|      NULL|  NULL|
|    NULL|  40|      NULL|  NULL|
|   Sarah|NULL|         3| 40000|
|   Sarah|NULL|      NULL|  NULL|
+--------+----+----------+------+
```

```
df.na.drop(how="any").show()
```

```
+--------+---+----------+------+
|    Name|Age|Experience|Salary|
+--------+---+----------+------+
|Willetta| 27|        52| 62860|
|  Merrie| 45|        34| 57591|
|  Joleen| 43|        42| 61343|
| Nananne| 25|        33| 75478|
| Jennica| 56|        10| 71299|
|  Lizzie| 55|        27| 62050|
|    Kaja| 28|        18| 52610|
|  Aubrie| 55|        18| 77702|
|  Ginnie| 71|        26| 88700|
|  Correy| 47|        25| 83288|
|Rochette| 24|        12| 56123|
|  Hayley| 55|        51| 72758|
|Mathilda| 42|        44| 62825|
|    Veda| 61|        31| 68511|
|   Jessy| 59|        41| 85482|
|    John| 30|         5| 50000|
+--------+---+----------+------+
```

```
df.na.drop(how="any", thresh=2).show()
```

```
+--------+----+----------+------+
|    Name| Age|Experience|Salary|
+--------+----+----------+------+
|Willetta|  27|        52| 62860|
|  Merrie|  45|        34| 57591|
|  Joleen|  43|        42| 61343|
| Nananne|  25|        33| 75478|
| Jennica|  56|        10| 71299|
|  Lizzie|  55|        27| 62050|
|    Kaja|  28|        18| 52610|
|  Aubrie|  55|        18| 77702|
|  Ginnie|  71|        26| 88700|
|  Correy|  47|        25| 83288|
|Rochette|  24|        12| 56123|
|  Hayley|  55|        51| 72758|
|Mathilda|  42|        44| 62825|
|    Veda|  61|        31| 68511|
|   Jessy|  59|        41| 85482|
|    John|  30|         5| 50000|
|    Mike|  25|      NULL|  NULL|
|   Sarah|NULL|         3| 40000|
+--------+----+----------+------+
```

```
df.na.drop(how="any", thresh=3).show()
```

```
+--------+----+----------+------+
|    Name| Age|Experience|Salary|
+--------+----+----------+------+
|Willetta|  27|        52| 62860|
|  Merrie|  45|        34| 57591|
|  Joleen|  43|        42| 61343|
| Nananne|  25|        33| 75478|
| Jennica|  56|        10| 71299|
|  Lizzie|  55|        27| 62050|
|    Kaja|  28|        18| 52610|
|  Aubrie|  55|        18| 77702|
|  Ginnie|  71|        26| 88700|
|  Correy|  47|        25| 83288|
|Rochette|  24|        12| 56123|
|  Hayley|  55|        51| 72758|
|Mathilda|  42|        44| 62825|
|    Veda|  61|        31| 68511|
|   Jessy|  59|        41| 85482|
|    John|  30|         5| 50000|
|   Sarah|NULL|         3| 40000|
+--------+----+----------+------+
```

```
df.na.drop(how="any", subset=['Experience']).show()
```

```
+--------+----+----------+------+
|    Name| Age|Experience|Salary|
+--------+----+----------+------+
|Willetta|  27|        52| 62860|
|  Merrie|  45|        34| 57591|
|  Joleen|  43|        42| 61343|
| Nananne|  25|        33| 75478|
| Jennica|  56|        10| 71299|
|  Lizzie|  55|        27| 62050|
|    Kaja|  28|        18| 52610|
|  Aubrie|  55|        18| 77702|
|  Ginnie|  71|        26| 88700|
|  Correy|  47|        25| 83288|
|Rochette|  24|        12| 56123|
|  Hayley|  55|        51| 72758|
|Mathilda|  42|        44| 62825|
|    Veda|  61|        31| 68511|
|   Jessy|  59|        41| 85482|
|    John|  30|         5| 50000|
|   Sarah|NULL|         3| 40000|
+--------+----+----------+------+
```

```
df.na.fill({'Name' : 'Missing', 'Salary': 0, 'Experience' : 0, 'Age' : 0}).show()
```

```
+--------+---+----------+------+
|    Name|Age|Experience|Salary|
+--------+---+----------+------+
|Willetta| 27|        52| 62860|
|  Merrie| 45|        34| 57591|
|  Joleen| 43|        42| 61343|
| Nananne| 25|        33| 75478|
| Jennica| 56|        10| 71299|
|  Lizzie| 55|        27| 62050|
|    Kaja| 28|        18| 52610|
|  Aubrie| 55|        18| 77702|
|  Ginnie| 71|        26| 88700|
|  Correy| 47|        25| 83288|
|Rochette| 24|        12| 56123|
|  Hayley| 55|        51| 72758|
|Mathilda| 42|        44| 62825|
|    Veda| 61|        31| 68511|
|   Jessy| 59|        41| 85482|
|    John| 30|         5| 50000|
|    Mike| 25|         0|     0|
| Missing| 40|         0|     0|
|   Sarah|  0|         3| 40000|
|   Sarah|  0|         0|     0|
+--------+---+----------+------+
```