```
!pip install pyspark
```

```
Collecting pyspark
  Downloading pyspark-3.5.1.tar.gz (317.0 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 317.0/317.0 MB 3.0 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-packages (from pyspark) (0.10.9.7)
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.5.1-py2.py3-none-any.whl size=317488491 sha256=0ecd3a7f411dd552d024c6e4b27dcd2797486659d
  Stored in directory: /root/.cache/pip/wheels/80/1d/60/2c256ed38dddce2fdd93be545214a63e02fbd8d74fb0b7f3a6
Successfully built pyspark
Installing collected packages: pyspark
Successfully installed pyspark-3.5.1
```

```
import pyspark
```

```
import pandas as pd
pd.read_csv('/content/train.csv')
```

	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	Postal Code	Region	Prod
0	1	CA-2017-152156	08/11/2017	11/11/2017	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420.0	South	FUR-B 10001
1	2	CA-2017-152156	08/11/2017	11/11/2017	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420.0	South	FUR-C 10000
2	3	CA-2017-138688	12/06/2017	16/06/2017	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	California	90036.0	West	OFF-10000
3	4	US-2016-108966	11/10/2016	18/10/2016	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311.0	South	FUR-10000
4	5	US-2016-108966	11/10/2016	18/10/2016	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311.0	South	OFF-10000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
9795	9796	CA-2017-125920	21/05/2017	28/05/2017	Standard Class	SH-19975	Sally Hughsby	Corporate	United States	Chicago	Illinois	60610.0	Central	OFF-10003
9796	9797	CA-2016-128608	12/01/2016	17/01/2016	Standard Class	CS-12490	Cindy Schnelling	Corporate	United States	Toledo	Ohio	43615.0	East	OFF-A 10001
9797	9798	CA-2016-128608	12/01/2016	17/01/2016	Standard Class	CS-12490	Cindy Schnelling	Corporate	United States	Toledo	Ohio	43615.0	East	TEC-P 10004
9798	9799	CA-2016-128608	12/01/2016	17/01/2016	Standard Class	CS-12490	Cindy Schnelling	Corporate	United States	Toledo	Ohio	43615.0	East	TEC-P 10000
9799	9800	CA-2016-128608	12/01/2016	17/01/2016	Standard Class	CS-12490	Cindy Schnelling	Corporate	United States	Toledo	Ohio	43615.0	East	TEC-A 10000

9800 rows × 18 columns

```
from pyspark.sql import SparkSession
```

```
spark=SparkSession.builder.appName('Practice').getOrCreate()
```

```
spark
```

SparkSession - in-memory

SparkContext

Spark UI

Version
v3.5.1
Master
local[*]
AppName
Practice

```
df_pyspark=spark.read.csv('/content/train.csv')
```

```
df_pyspark=spark.read.option('header','true').csv('/content/train.csv',inferSchema=True)
```

```
df_pyspark.printSchema()
```

```
root
 |-- Row ID: integer (nullable = true)
 |-- Order ID: string (nullable = true)
 |-- Order Date: string (nullable = true)
 |-- Ship Date: string (nullable = true)
 |-- Ship Mode: string (nullable = true)
 |-- Customer ID: string (nullable = true)
 |-- Customer Name: string (nullable = true)
 |-- Segment: string (nullable = true)
 |-- Country: string (nullable = true)
 |-- City: string (nullable = true)
 |-- State: string (nullable = true)
 |-- Postal Code: integer (nullable = true)
 |-- Region: string (nullable = true)
 |-- Product ID: string (nullable = true)
 |-- Category: string (nullable = true)
 |-- Sub-Category: string (nullable = true)
 |-- Product Name: string (nullable = true)
 |-- Sales: string (nullable = true)
```

```
df_pyspark=spark.read.csv('/content/train.csv' , header=True,inferSchema=True)
df_pyspark.show()
```

```
+------+--------------+----------+----------+--------------+-----------+------------------+-----------+-------------+---------------+---
|Row ID|      Order ID|Order Date| Ship Date|     Ship Mode|Customer ID|     Customer Name|    Segment|      Country|           City|
+------+--------------+----------+----------+--------------+-----------+------------------+-----------+-------------+---------------+---
|     1|CA-2017-152156|08/11/2017|11/11/2017|  Second Class|   CG-12520|       Claire Gute|   Consumer|United States|      Henderson|
|     2|CA-2017-152156|08/11/2017|11/11/2017|  Second Class|   CG-12520|       Claire Gute|   Consumer|United States|      Henderson|
|     3|CA-2017-138688|12/06/2017|16/06/2017|  Second Class|   DV-13045|   Darrin Van Huff|  Corporate|United States|    Los Angeles|
|     4|US-2016-108966|11/10/2016|18/10/2016|Standard Class|   SO-20335|    Sean O'Donnell|   Consumer|United States|Fort Lauderdale|
|     5|US-2016-108966|11/10/2016|18/10/2016|Standard Class|   SO-20335|    Sean O'Donnell|   Consumer|United States|Fort Lauderdale|
|     6|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|     7|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|     8|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|     9|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|    10|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|    11|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|    12|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|    13|CA-2018-114412|15/04/2018|20/04/2018|Standard Class|   AA-10480|      Andrew Allen|   Consumer|United States|        Concord|Nor
|    14|CA-2017-161389|05/12/2017|10/12/2017|Standard Class|   IM-15070|      Irene Maddox|   Consumer|United States|        Seattle|
|    15|US-2016-118983|22/11/2016|26/11/2016|Standard Class|   HP-14815|     Harold Pawlan|Home Office|United States|     Fort Worth|
|    16|US-2016-118983|22/11/2016|26/11/2016|Standard Class|   HP-14815|     Harold Pawlan|Home Office|United States|     Fort Worth|
|    17|CA-2015-105893|11/11/2015|18/11/2015|Standard Class|   PK-19075|         Pete Kriz|   Consumer|United States|        Madison|
|    18|CA-2015-167164|13/05/2015|15/05/2015|  Second Class|   AG-10270|   Alejandro Grove|   Consumer|United States|    West Jordan|
|    19|CA-2015-143336|27/08/2015|01/09/2015|  Second Class|   ZD-21925|Zuschuss Donatelli|   Consumer|United States|  San Francisco|
|    20|CA-2015-143336|27/08/2015|01/09/2015|  Second Class|   ZD-21925|Zuschuss Donatelli|   Consumer|United States|  San Francisco|
+------+--------------+----------+----------+--------------+-----------+------------------+-----------+-------------+---------------+---
only showing top 20 rows
```

```
type(df_pyspark)
```

```
pyspark.sql.dataframe.DataFrame
def __init__(jdf: JavaObject, sql_ctx: Union['SQLContext', 'SparkSession'])

A distributed collection of data grouped into named columns.

.. versionadded:: 1.3.0

.. versionchanged:: 3.4.0
    Supports Spark Connect.
```

```
df_pyspark.columns
```

```
['Row ID',
 'Order ID',
 'Order Date',
 'Ship Date',
 'Ship Mode',
 'Customer ID',
 'Customer Name',
 'Segment',
```

```
 'Country',
 'City',
 'State',
 'Postal Code',
 'Region',
 'Product ID',
 'Category',
 'Sub-Category',
 'Product Name',
 'Sales']
```

```
df_pyspark.show()
```

```
+------+--------------+----------+----------+--------------+-----------+------------------+-----------+-------------+---------------+---
|Row ID|      Order ID|Order Date| Ship Date|     Ship Mode|Customer ID|     Customer Name|    Segment|      Country|           City|
+------+--------------+----------+----------+--------------+-----------+------------------+-----------+-------------+---------------+---
|     1|CA-2017-152156|08/11/2017|11/11/2017|  Second Class|   CG-12520|       Claire Gute|   Consumer|United States|      Henderson|
|     2|CA-2017-152156|08/11/2017|11/11/2017|  Second Class|   CG-12520|       Claire Gute|   Consumer|United States|      Henderson|
|     3|CA-2017-138688|12/06/2017|16/06/2017|  Second Class|   DV-13045|   Darrin Van Huff|  Corporate|United States|    Los Angeles|
|     4|US-2016-108966|11/10/2016|18/10/2016|Standard Class|   SO-20335|    Sean O'Donnell|   Consumer|United States|Fort Lauderdale|
|     5|US-2016-108966|11/10/2016|18/10/2016|Standard Class|   SO-20335|    Sean O'Donnell|   Consumer|United States|Fort Lauderdale|
|     6|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|     7|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|     8|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|     9|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|    10|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|    11|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|    12|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|    13|CA-2018-114412|15/04/2018|20/04/2018|Standard Class|   AA-10480|      Andrew Allen|   Consumer|United States|        Concord|Nor
|    14|CA-2017-161389|05/12/2017|10/12/2017|Standard Class|   IM-15070|      Irene Maddox|   Consumer|United States|        Seattle|
|    15|US-2016-118983|22/11/2016|26/11/2016|Standard Class|   HP-14815|     Harold Pawlan|Home Office|United States|     Fort Worth|
|    16|US-2016-118983|22/11/2016|26/11/2016|Standard Class|   HP-14815|     Harold Pawlan|Home Office|United States|     Fort Worth|
|    17|CA-2015-105893|11/11/2015|18/11/2015|Standard Class|   PK-19075|         Pete Kriz|   Consumer|United States|        Madison|
|    18|CA-2015-167164|13/05/2015|15/05/2015|  Second Class|   AG-10270|   Alejandro Grove|   Consumer|United States|    West Jordan|
|    19|CA-2015-143336|27/08/2015|01/09/2015|  Second Class|   ZD-21925|Zuschuss Donatelli|   Consumer|United States|  San Francisco|
|    20|CA-2015-143336|27/08/2015|01/09/2015|  Second Class|   ZD-21925|Zuschuss Donatelli|   Consumer|United States|  San Francisco|
+------+--------------+----------+----------+--------------+-----------+------------------+-----------+-------------+---------------+---
only showing top 20 rows
```

```
df_pyspark.select(['Sales', 'Customer ID']).show()
```

```
+--------+-----------+
|   Sales|Customer ID|
+--------+-----------+
|  261.96|   CG-12520|
|  731.94|   CG-12520|
|   14.62|   DV-13045|
|957.5775|   SO-20335|
|  22.368|   SO-20335|
|   48.86|   BH-11710|
|    7.28|   BH-11710|
| 907.152|   BH-11710|
|  18.504|   BH-11710|
|   114.9|   BH-11710|
|1706.184|   BH-11710|
| 911.424|   BH-11710|
|  15.552|   AA-10480|
| 407.976|   IM-15070|
|   68.81|   HP-14815|
|   2.544|   HP-14815|
|  665.88|   PK-19075|
|    55.5|   AG-10270|
|    8.56|   ZD-21925|
|  213.48|   ZD-21925|
+--------+-----------+
only showing top 20 rows
```

```
df_pyspark.dtypes
```

```
[('Row ID', 'int'),
 ('Order ID', 'string'),
 ('Order Date', 'string'),
 ('Ship Date', 'string'),
 ('Ship Mode', 'string'),
 ('Customer ID', 'string'),
 ('Customer Name', 'string'),
 ('Segment', 'string'),
 ('Country', 'string'),
```

```
 ('City', 'string'),
 ('State', 'string'),
 ('Postal Code', 'int'),
 ('Region', 'string'),
 ('Product ID', 'string'),
 ('Category', 'string'),
 ('Sub-Category', 'string'),
 ('Product Name', 'string'),
 ('Sales', 'string')]
```

```
df_pyspark.describe().show()
```

```
+-------+------------------+--------------+----------+----------+--------------+-----------+------------------+-----------+-------------
|summary|            Row ID|      Order ID|Order Date| Ship Date|     Ship Mode|Customer ID|     Customer Name|    Segment|      Country
+-------+------------------+--------------+----------+----------+--------------+-----------+------------------+-----------+-------------
|  count|              9800|          9800|      9800|      9800|          9800|       9800|              9800|       9800|         9800
|   mean|            4900.5|          NULL|      NULL|      NULL|          NULL|       NULL|              NULL|       NULL|         NULL
| stddev|2829.1606529145706|          NULL|      NULL|      NULL|          NULL|       NULL|              NULL|       NULL|         NULL
|    min|                 1|CA-2015-100006|01/01/2018|01/01/2016|   First Class|   AA-10315|     Aaron Bergman|   Consumer|United States
|    max|              9800|US-2018-169551|31/12/2017|31/12/2018|Standard Class|   ZD-21925|Zuschuss Donatelli|Home Office|United States
+-------+------------------+--------------+----------+----------+--------------+-----------+------------------+-----------+-------------
```

```
df_pyspark.drop('Sales').show()
```

```
+------+--------------+----------+----------+--------------+-----------+------------------+-----------+-------------+---------------+---
|Row ID|      Order ID|Order Date| Ship Date|     Ship Mode|Customer ID|     Customer Name|    Segment|      Country|           City|
+------+--------------+----------+----------+--------------+-----------+------------------+-----------+-------------+---------------+---
|     1|CA-2017-152156|08/11/2017|11/11/2017|  Second Class|   CG-12520|       Claire Gute|   Consumer|United States|      Henderson|
|     2|CA-2017-152156|08/11/2017|11/11/2017|  Second Class|   CG-12520|       Claire Gute|   Consumer|United States|      Henderson|
|     3|CA-2017-138688|12/06/2017|16/06/2017|  Second Class|   DV-13045|   Darrin Van Huff|  Corporate|United States|    Los Angeles|
|     4|US-2016-108966|11/10/2016|18/10/2016|Standard Class|   SO-20335|    Sean O'Donnell|   Consumer|United States|Fort Lauderdale|
|     5|US-2016-108966|11/10/2016|18/10/2016|Standard Class|   SO-20335|    Sean O'Donnell|   Consumer|United States|Fort Lauderdale|
|     6|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|     7|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|     8|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|     9|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|    10|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|    11|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|    12|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|    13|CA-2018-114412|15/04/2018|20/04/2018|Standard Class|   AA-10480|      Andrew Allen|   Consumer|United States|        Concord|Nor
|    14|CA-2017-161389|05/12/2017|10/12/2017|Standard Class|   IM-15070|      Irene Maddox|   Consumer|United States|        Seattle|
|    15|US-2016-118983|22/11/2016|26/11/2016|Standard Class|   HP-14815|     Harold Pawlan|Home Office|United States|     Fort Worth|
|    16|US-2016-118983|22/11/2016|26/11/2016|Standard Class|   HP-14815|     Harold Pawlan|Home Office|United States|     Fort Worth|
|    17|CA-2015-105893|11/11/2015|18/11/2015|Standard Class|   PK-19075|         Pete Kriz|   Consumer|United States|        Madison|
|    18|CA-2015-167164|13/05/2015|15/05/2015|  Second Class|   AG-10270|   Alejandro Grove|   Consumer|United States|    West Jordan|
|    19|CA-2015-143336|27/08/2015|01/09/2015|  Second Class|   ZD-21925|Zuschuss Donatelli|   Consumer|United States|  San Francisco|
|    20|CA-2015-143336|27/08/2015|01/09/2015|  Second Class|   ZD-21925|Zuschuss Donatelli|   Consumer|United States|  San Francisco|
+------+--------------+----------+----------+--------------+-----------+------------------+-----------+-------------+---------------+---
only showing top 20 rows
```

```
df_pyspark.na.drop().show()
```

```
+------+--------------+----------+----------+--------------+-----------+------------------+-----------+-------------+---------------+---
|Row ID|      Order ID|Order Date| Ship Date|     Ship Mode|Customer ID|     Customer Name|    Segment|      Country|           City|
+------+--------------+----------+----------+--------------+-----------+------------------+-----------+-------------+---------------+---
|     1|CA-2017-152156|08/11/2017|11/11/2017|  Second Class|   CG-12520|       Claire Gute|   Consumer|United States|      Henderson|
|     2|CA-2017-152156|08/11/2017|11/11/2017|  Second Class|   CG-12520|       Claire Gute|   Consumer|United States|      Henderson|
|     3|CA-2017-138688|12/06/2017|16/06/2017|  Second Class|   DV-13045|   Darrin Van Huff|  Corporate|United States|    Los Angeles|
|     4|US-2016-108966|11/10/2016|18/10/2016|Standard Class|   SO-20335|    Sean O'Donnell|   Consumer|United States|Fort Lauderdale|
|     5|US-2016-108966|11/10/2016|18/10/2016|Standard Class|   SO-20335|    Sean O'Donnell|   Consumer|United States|Fort Lauderdale|
|     6|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|     7|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|     8|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|     9|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|    10|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|    11|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|    12|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|    13|CA-2018-114412|15/04/2018|20/04/2018|Standard Class|   AA-10480|      Andrew Allen|   Consumer|United States|        Concord|Nor
|    14|CA-2017-161389|05/12/2017|10/12/2017|Standard Class|   IM-15070|      Irene Maddox|   Consumer|United States|        Seattle|
|    15|US-2016-118983|22/11/2016|26/11/2016|Standard Class|   HP-14815|     Harold Pawlan|Home Office|United States|     Fort Worth|
|    16|US-2016-118983|22/11/2016|26/11/2016|Standard Class|   HP-14815|     Harold Pawlan|Home Office|United States|     Fort Worth|
|    17|CA-2015-105893|11/11/2015|18/11/2015|Standard Class|   PK-19075|         Pete Kriz|   Consumer|United States|        Madison|
|    18|CA-2015-167164|13/05/2015|15/05/2015|  Second Class|   AG-10270|   Alejandro Grove|   Consumer|United States|    West Jordan|
|    19|CA-2015-143336|27/08/2015|01/09/2015|  Second Class|   ZD-21925|Zuschuss Donatelli|   Consumer|United States|  San Francisco|
|    20|CA-2015-143336|27/08/2015|01/09/2015|  Second Class|   ZD-21925|Zuschuss Donatelli|   Consumer|United States|  San Francisco|
+------+--------------+----------+----------+--------------+-----------+------------------+-----------+-------------+---------------+---
only showing top 20 rows
```

```
df_pyspark.filter("Sales=0").show()
```

```
+------+--------------+----------+----------+--------------+-----------+----------------+-----------+-------------+-------------+-------
|Row ID|      Order ID|Order Date| Ship Date|     Ship Mode|Customer ID|   Customer Name|    Segment|      Country|         City|
+------+--------------+----------+----------+--------------+-----------+----------------+-----------+-------------+-------------+-------
|  2107|US-2015-152723|26/09/2015|26/09/2015|      Same Day|   HG-14965|   Henry Goldwyn|  Corporate|United States|     Mesquite|
|  2762|CA-2018-126536|12/10/2018|14/10/2018|   First Class|   NK-18490|    Neil Knudson|Home Office|United States|San Francisco|  Calif
|  4102|US-2018-102288|19/06/2018|23/06/2018|Standard Class|   ZC-21910|Zuschuss Carroll|   Consumer|United States|      Houston|
|  4712|CA-2015-112403|31/03/2015|31/03/2015|      Same Day|   JO-15280|   Jas O'Carroll|   Consumer|United States| Philadelphia|Pennsyl
|  7549|CA-2015-103492|10/10/2015|15/10/2015|Standard Class|   CM-12715|  Craig Molinari|  Corporate|United States|   Huntsville|
|  8034|CA-2016-119690|25/06/2016|28/06/2016|   First Class|   MV-17485|   Mark Van Huff|   Consumer|United States|      Houston|
|  8659|CA-2017-168361|21/06/2017|25/06/2017|Standard Class|   KB-16600|     Ken Brennan|  Corporate|United States|      Chicago|    Ill
|  9293|CA-2018-124114|02/03/2018|02/03/2018|      Same Day|   RS-19765|  Roland Schwarz|  Corporate|United States|         Waco|
+------+--------------+----------+----------+--------------+-----------+----------------+-----------+-------------+-------------+-------
```

```
df_pyspark.filter("Sales>=200").select(['Customer ID', 'Customer Name','Sales']).show()
```

```
+-----------+------------------+--------+
|Customer ID|     Customer Name|   Sales|
+-----------+------------------+--------+
|   CG-12520|       Claire Gute|  261.96|
|   CG-12520|       Claire Gute|  731.94|
|   SO-20335|    Sean O'Donnell|957.5775|
|   BH-11710|   Brosina Hoffman| 907.152|
|   BH-11710|   Brosina Hoffman|1706.184|
|   BH-11710|   Brosina Hoffman| 911.424|
|   IM-15070|      Irene Maddox| 407.976|
|   PK-19075|         Pete Kriz|  665.88|
|   ZD-21925|Zuschuss Donatelli|  213.48|
|   EB-13870|       Emily Burns| 1044.63|
|   TB-21520|   Tracy Blumstein| 3083.43|
|   GH-14485|         Gene Hale|1097.544|
|   SN-20710|      Steve Nguyen|532.3992|
|   SN-20710|      Steve Nguyen| 212.058|
|   SN-20710|      Steve Nguyen| 371.168|
|   PO-18865| Patrick O'Donnell|  211.96|
|   JM-15265|    Janet Molinari| 1029.95|
|   TB-21055|   Ted Butterfield|  208.56|
|   TB-21055|   Ted Butterfield|  319.41|
|   PS-18970|    Paul Stevenson| 213.115|
+-----------+------------------+--------+
only showing top 20 rows
```

##Groupby

Counting the number of the customers from the different states

```
df_pyspark.groupby('State').count().show(99)
```

```
+--------------------+-----+
|               State|count|
+--------------------+-----+
|                Utah|   53|
|           Minnesota|   89|
|                Ohio|  454|
|              Oregon|  122|
|            Arkansas|   60|
|               Texas|  973|
|        North Dakota|    7|
|        Pennsylvania|  582|
|         Connecticut|   82|
|            Nebraska|   38|
|             Vermont|   11|
|              Nevada|   39|
|          Washington|  504|
|            Illinois|  483|
|            Oklahoma|   66|
|District of Columbia|   10|
|            Delaware|   93|
|          New Mexico|   37|
|       West Virginia|    4|
|            Missouri|   66|
|        Rhode Island|   55|
```

```
|             Georgia|  177|
|             Montana|   15|
|            Michigan|  253|
|            Virginia|  224|
|      North Carolina|  247|
|             Wyoming|    1|
|              Kansas|   24|
|          New Jersey|  122|
|            Maryland|  105|
|             Alabama|   61|
|             Arizona|  223|
|                Iowa|   26|
|       Massachusetts|  135|
|             Kentucky|  137|
|           Louisiana|   41|
|         Mississippi|   53|
|           Tennessee|  183|
|       New Hampshire|   27|
|             Florida|  373|
|             Indiana|  135|
|               Idaho|   21|
|      South Carolina|   42|
|        South Dakota|   12|
|          California| 1946|
|            New York| 1097|
|           Wisconsin|  105|
|            Colorado|  179|
|               Maine|    8|
+--------------------+-----+
```

Calulationg the total number of the Sales amount

```
df_pyspark.agg({'Sales':'sum'}).show()
```

```
+------------------+
|        sum(Sales)|
+------------------+
|2237133.1626999523|
+------------------+
```

Maxium amount of the sales

```
df_pyspark.agg({'Sales':'max'}).show()
```

```
+----------+
|max(Sales)|
+----------+
|    999.98|
+----------+
```

```
training = spark.read.csv('/content/train.csv', header=True,inferSchema=True)
```

```
training.show()
```

```
+------+--------------+----------+----------+--------------+-----------+------------------+-----------+-------------+---------------+---
|Row ID|      Order ID|Order Date| Ship Date|     Ship Mode|Customer ID|     Customer Name|    Segment|      Country|           City|
+------+--------------+----------+----------+--------------+-----------+------------------+-----------+-------------+---------------+---
|     1|CA-2017-152156|08/11/2017|11/11/2017|  Second Class|   CG-12520|       Claire Gute|   Consumer|United States|      Henderson|
|     2|CA-2017-152156|08/11/2017|11/11/2017|  Second Class|   CG-12520|       Claire Gute|   Consumer|United States|      Henderson|
|     3|CA-2017-138688|12/06/2017|16/06/2017|  Second Class|   DV-13045|   Darrin Van Huff|  Corporate|United States|    Los Angeles|
|     4|US-2016-108966|11/10/2016|18/10/2016|Standard Class|   SO-20335|    Sean O'Donnell|   Consumer|United States|Fort Lauderdale|
|     5|US-2016-108966|11/10/2016|18/10/2016|Standard Class|   SO-20335|    Sean O'Donnell|   Consumer|United States|Fort Lauderdale|
|     6|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|     7|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|     8|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|     9|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|    10|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|    11|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|    12|CA-2015-115812|09/06/2015|14/06/2015|Standard Class|   BH-11710|   Brosina Hoffman|   Consumer|United States|    Los Angeles|
|    13|CA-2018-114412|15/04/2018|20/04/2018|Standard Class|   AA-10480|      Andrew Allen|   Consumer|United States|        Concord|Nor
|    14|CA-2017-161389|05/12/2017|10/12/2017|Standard Class|   IM-15070|      Irene Maddox|   Consumer|United States|        Seattle|
|    15|US-2016-118983|22/11/2016|26/11/2016|Standard Class|   HP-14815|     Harold Pawlan|Home Office|United States|     Fort Worth|
|    16|US-2016-118983|22/11/2016|26/11/2016|Standard Class|   HP-14815|     Harold Pawlan|Home Office|United States|     Fort Worth|
|    17|CA-2015-105893|11/11/2015|18/11/2015|Standard Class|   PK-19075|         Pete Kriz|   Consumer|United States|        Madison|
|    18|CA-2015-167164|13/05/2015|15/05/2015|  Second Class|   AG-10270|   Alejandro Grove|   Consumer|United States|    West Jordan|
```

```
|    19|CA-2015-143336|27/08/2015|01/09/2015|  Second Class|   ZD-21925|Zuschuss Donatelli|   Consumer|United States|  San Francisco|
|    20|CA-2015-143336|27/08/2015|01/09/2015|  Second Class|   ZD-21925|Zuschuss Donatelli|   Consumer|United States|  San Francisco|
+------+--------------+----------+----------+--------------+-----------+------------------+-----------+-------------+---------------+---
only showing top 20 rows
```

```
from pyspark.ml.feature import VectorAssembler
featureassembler=VectorAssembler(inputCols=["Product ID","Customer ID"],outputCol="Independent")
```

```
output=featureassembler.transform(training)
```

```
---------------------------------------------------------------------------
IllegalArgumentException                  Traceback (most recent call last)
<ipython-input-28-97ed59724aa3> in <cell line: 1>()
----> 1 output=featureassembler.transform(training)

 3 frames 

/usr/local/lib/python3.10/dist-packages/pyspark/errors/exceptions/captured.py in deco(*a, **kw)
```