

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

```
spark = SparkSession.builder.appName('Big Data Analysis').getOrCreate()
```

```
spark
```

SparkSession - in-memory

SparkContext

[Spark UI](#)

Version

v3.5.1

Master

local[*]

AppName

Big Data Analysis

```
df = spark.read.csv('/content/Car_Purchasing_Data.csv',header=True,inferSchema=True)
```

Display Top 3 Rows in from the dataset

```
df.show(3)
```

```
+-----+-----+-----+-----+-----+-----+-----+-----+
| Customer Name| Customer e-mail|Country|Gender|Age|Annual Salary|Credit Card Debt| Net Worth|Car Purchase Amount|
+-----+-----+-----+-----+-----+-----+-----+-----+
| Martina Avila|cubilia.Curae.Pha...| USA| 0| 42| 62812.09301| 11609.38091|238961.2505| 35321.45877|
| Harlan Barnes| eu.dolor@diam.co.uk| USA| 0| 41| 66646.89292| 9572.957136|530973.9078| 45115.52566|
| Naomi Rodriguez|vulputate.mauris....| USA| 1| 43| 53798.55112| 11160.35506|638467.1773| 42925.70921|
+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 3 rows
```

Display Datatype of each column

```
df.printSchema()
```

```
root
|-- Customer Name: string (nullable = true)
|-- Customer e-mail: string (nullable = true)
|-- Country: string (nullable = true)
|-- Gender: integer (nullable = true)
|-- Age: integer (nullable = true)
|-- Annual Salary: double (nullable = true)
|-- Credit Card Debt: double (nullable = true)
|-- Net Worth: double (nullable = true)
|-- Car Purchase Amount: double (nullable = true)
```

Display Column Names

```
df.columns
```

```
['Customer Name',
 'Customer e-mail',
 'Country',
 'Gender',
 'Age',
 'Annual Salary',
 'Credit Card Debt',
 'Net Worth',
 'Car Purchase Amount']
```

count No Of Rows & Columns Of The Dataset

```
df.count()
```

```
500
```

```
len(df.columns)
```

```
9
```

Get Overall Statistics About The Dataset

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

	summary	Customer Name	Customer e-mail	Country	Gender	Age	Annual Salary	Credit Card Debt
count	500	500	500	500	500	500	500	500
mean	NULL	NULL	NULL	0.506	46.224	62127.239607559975	9607.645048629205	4314
stddev	NULL	NULL	NULL	0.5004647139007097	7.990338855772024	11703.37822774127	3489.1879728381996	1735
min	Abel Stanton	Aenean.gravida@mi...	USA	0	20	20000.0	100.0	
max	Zephania	vulputate.mauris...	USA	1	70	100000.0	20000.0	

Find Unique Values Available In the Gender Column

```
df.toPandas()['Gender'].unique()
```

```
array([0, 1], dtype=int32)
```

Find Total No Of Unique Values Available In the Gender Column

```
len(df.toPandas()['Gender'].unique())
```

```
2
```

How Select Single Column ?

```
df.show()
```

	Customer Name	Customer e-mail	Country	Gender	Age	Annual Salary	Credit Card Debt	Net Worth	Car Purchase Amount
	Martina Avila	cubilia.Currae.Pha...	USA	0	42	62812.09301	11609.38091	238961.2505	35321.45877
	Harlan Barnes	eu.dolor@diam.co.uk	USA	0	41	66646.89292	9572.957136	530973.9078	45115.52566
	Naomi Rodriquez	vulputate.mauris....	USA	1	43	53798.55112	11160.35506	638467.1773	42925.70921
	Jade Cunningham	malesuada@digniss...	USA	1	58	79370.03798	14426.16485	548599.0524	67422.36313
	Cedric Leach	felis.ullamcorper...	USA	1	57	59729.1513	5358.712177	560304.0671	55915.46248
	Carla Hester	mi@Aliquamerat.edu	USA	1	57	68499.85162	14179.47244	428485.3604	56611.99784
	Griffin Rivera	vehicula@at.co.uk	USA	1	47	39814.522	5958.460188	326373.1812	28925.70549
	Orli Casey	nunc.est.mollis@S...	USA	1	50	51752.23445	10985.69656	629312.4041	47434.98265
	Marny Obrien	Phasellus@sedseme...	USA	0	47	58139.2591	3440.823799	630059.0274	48013.6141
	Rhonda Chavez	nec@nuncest.com	USA	1	43	53457.10132	12884.07868	476643.3544	38189.50601
	Jerome	Rowe.ipsum.cursus@dui.org	USA	1	50	73348.70745	8270.707359	612738.6171	59045.51309
	Akeem Gibson	turpis.egestas.Fu...	USA	1	53	55421.65733	10014.96929	293862.5123	42288.81046
	Quin Smith	nulla@ipsum.edu	USA	0	44	37336.3383	10218.32092	430907.1673	28700.0334
	Tatum Moon	Cras.sed.leo@Sedd...	USA	0	48	68304.47298	9466.995128	420322.0702	49258.87571
	Sharon Sharpe	eget.metus@aaliq...	USA	0	55	72776.00382	10597.63814	146344.8965	49510.03356
	Thomas Williams	aliquet.molestie@...	USA	1	53	64662.30061	11326.03434	481433.4324	53017.26723
	Blaine Bender	ultrices.posuere....	USA	0	45	63259.87837	11495.54999	370356.2223	41814.72067
	Stephen Lindsey	erat.eget.ipsu@t...	USA	1	48	52682.06401	12514.52029	549443.5886	43901.71244
	Sloane Mann	at.augue@augue.net	USA	1	52	54503.14423	7377.820914	431098.9998	44633.99241
	Athena Wolf	volutpat.Nulla.fa...	USA	0	59	55368.23716	13272.94647	566022.1306	54827.52403

only showing top 20 rows

```
df.select('Customer Name').show()
```

```
+-----+
| Customer Name|
+-----+
| Martina Avila|
| Harlan Barnes|
| Naomi Rodriquez|
| Jade Cunningham|
| Cedric Leach|
| Carla Hester|
| Griffin Rivera|
```

```

| Orli Casey|
| Marny Obrien|
| Rhonda Chavez|
| Jerome Rowe|
| Akeem Gibson|
| Quin Smith|
| Tatum Moon|
| Sharon Sharpe|
| Thomas Williams|
| Blaine Bender|
| Stephen Lindsey|
| Sloane Mann|
| Athena Wolf|
+-----+
only showing top 20 rows

```

How To Select Multiple Column ?

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

```

+-----+-----+
| Customer Name|Gender|
+-----+-----+
| Martina Avila| 0|
| Harlan Barnes| 0|
| Naomi Rodriquez| 1|
| Jade Cunningham| 1|
| Cedric Leach| 1|
| Carla Hester| 1|
| Griffin Rivera| 1|
| Orli Casey| 1|
| Marny Obrien| 0|
| Rhonda Chavez| 1|
| Jerome Rowe| 1|
| Akeem Gibson| 1|
| Quin Smith| 0|
| Tatum Moon| 0|
| Sharon Sharpe| 0|
| Thomas Williams| 1|
| Blaine Bender| 0|
| Stephen Lindsey| 1|
| Sloane Mann| 1|
| Athena Wolf| 0|
+-----+-----+
only showing top 20 rows

```

Creating A New Column With Marks + 1 And Also Updating existing DataFrame

```
df = df.withColumn('New_Annual Salary',df['Annual Salary']+1000)
```

```
df.show()
```

```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| Customer Name|Customer e-mail|Country|Gender|Age|Annual Salary|Credit Card Debt|Net Worth|Car Purchase Amount|New_Annual
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| Martina Avila|cubilia.Curae.Pha...|USA|0|42|62812.09301|11609.38091|238961.2505|35321.45877|63812
| Harlan Barnes|eu.dolor@diam.co.uk|USA|0|41|66646.89292|9572.957136|530973.9078|45115.52566|67646
| Naomi Rodriquez|vulputate.mauris....|USA|1|43|53798.55112|11160.35506|638467.1773|42925.70921|54798
| Jade Cunningham|malesuada@digniss...|USA|1|58|79370.03798|14426.16485|548599.0524|67422.36313|80370
| Cedric Leach|felis.ullamcorper...|USA|1|57|59729.1513|5358.712177|560304.0671|55915.46248|6072
| Carla Hester|mi@Aliquamerat.edu|USA|1|57|68499.85162|14179.47244|428485.3604|56611.99784|69495
| Griffin Rivera|vehicula@at.co.uk|USA|1|47|39814.522|5958.460188|326373.1812|28925.70549|408
| Orli Casey|nunc.est.mollis@S...|USA|1|50|51752.23445|10985.69656|629312.4041|47434.98265|52752
| Marny Obrien|Phasellus@sedseme...|USA|0|47|58139.2591|3440.823799|630059.0274|48013.6141|5913
| Rhonda Chavez|nec@nuncest.com|USA|1|43|53457.10132|12884.07868|476643.3544|38189.50601|54457
| Jerome|Rowe@ipsum.cursus@dui.org|USA|1|50|73348.70745|8270.707359|612738.6171|59045.51309|74348
| Akeem Gibson|turpis.egestas.Fu...|USA|1|53|55421.65733|10014.96929|293862.5123|42288.81046|56421
| Quin Smith|nulla@ipsum.edu|USA|0|44|37336.3383|10218.32092|430907.1673|28700.0334|3833
| Tatum Moon|Cras.sed.leo@Sedd...|USA|0|48|68304.47298|9466.995128|420322.0702|49258.87571|69304
| Sharon Sharpe| eget.metus@aaliq...|USA|0|55|72776.00382|10597.63814|146344.8965|49510.03356|73776
| Thomas Williams|aliquet.molestie@...|USA|1|53|64662.30061|11326.03434|481433.4324|53017.26723|65662
| Blaine Bender|ultrices.posuere....|USA|0|45|63259.87837|11495.54999|370356.2223|41814.72067|64259
| Stephen Lindsey|erat. eget. ipsum@t...|USA|1|48|52682.06401|12514.52029|549443.5886|43901.71244|53682
| Sloane Mann|at.augue@augue.net|USA|1|52|54503.14423|7377.820914|431098.9998|44633.99241|55503
| Athena Wolf|volutpat.Nulla.fa...|USA|0|59|55368.23716|13272.94647|566022.1306|54827.52403|56368
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 20 rows

```

Rename Customer_Name Column And Give New Name "Name"

```
df.columns
```

```
['Customer Name',
 'Customer e-mail',
 'Country',
 'Gender',
 'Age',
 'Annual Salary',
 'Credit Card Debt',
 'Net Worth',
 'Car Purchase Amount',
 'New_Annual Salary']
```

```
df = df.withColumnRenamed('Customer Name', 'Name')
```

```
df.show()
```

Name	Customer e-mail	Country	Gender	Age	Annual Salary	Credit Card Debt	Net Worth	Car Purchase Amount	New_Annual
Martina Avila	cubilia.Currae.Pha...	USA	0	42	62812.09301	11609.38091	238961.2505	35321.45877	63812
Harlan Barnes	eu.dolor@diam.co.uk	USA	0	41	66646.89292	9572.957136	530973.9078	45115.52566	67646
Naomi Rodriquez	vulputate.mauris....	USA	1	43	53798.55112	11160.35506	638467.1773	42925.70921	54798
Jade Cunningham	malesuada@digniss...	USA	1	58	79370.03798	14426.16485	548599.0524	67422.36313	80370
Cedric Leach	felis.ullamcorper...	USA	1	57	59729.1513	5358.712177	560304.0671	55915.46248	6072
Carla Hester	mi@Aliquam erat...	USA	1	57	68499.85162	14179.47244	428485.3604	56611.99784	69495
Griffin Rivera	vehicula@at.co.uk	USA	1	47	39814.522	5958.460188	326373.1812	28925.70549	408
Orli Casey	nunc.est.mollis@S...	USA	1	50	51752.23445	10985.69656	629312.4041	47434.98265	52752
Marny Obrien	Phasellus@sedseme...	USA	0	47	58139.2591	3440.823799	630059.0274	48013.6141	5913
Rhonda Chavez	nec@nunc est.com	USA	1	43	53457.10132	12884.07868	476643.3544	38189.50601	54457
Jerome Rowe	ipsum.cursus@dui.org	USA	1	50	73348.70745	8270.707359	612738.6171	59045.51309	74348
Akeem Gibson	turpis.egestas.Fu...	USA	1	53	55421.65733	10014.96929	293862.5123	42288.81046	56421
Quin Smith	nulla@ipsum.edu	USA	0	44	37336.3383	10218.32092	430907.1673	28700.0334	3833
Tatum Moon	Cras.sed.leo@Sedd...	USA	0	48	68304.47298	9466.995128	420322.0702	49258.87571	69304
Sharon Sharpe	eget.metus@aaliq...	USA	0	55	72776.00382	10597.63814	146344.8965	49510.03356	73776
Thomas Williams	aliquet.molestie@...	USA	1	53	64662.30061	11326.03434	481433.4324	53017.26723	65662
Blaine Bender	ultrices.posuere....	USA	0	45	63259.87837	11495.54999	370356.2223	41814.72067	64259
Stephen Lindsey	erat.eget.ipsum@t...	USA	1	48	52682.06401	12514.52029	549443.5886	43901.71244	53682
Sloane Mann	at.augue@augue.net	USA	1	52	54503.14423	7377.820914	431098.9998	44633.99241	55503
Athena Wolf	volutpat.Nulla.fa...	USA	0	59	55368.23716	13272.94647	566022.1306	54827.52403	56368

only showing top 20 rows

Display Name Of The Customers Having Net Worth Greater Than 5 Lakhs

```
df.filter(df['Net Worth']>500000).select('Name').show()
```

Name
Harlan Barnes
Naomi Rodriquez
Jade Cunningham
Cedric Leach
Orli Casey
Marny Obrien
Jerome Rowe
Stephen Lindsey
Athena Wolf
Micah Wheeler
Castor Wood
Dahlia Cleveland
Coby Charles
Rachel Ashley
Quincy Bell
Quon Carroll
Leilani Gross
Francesca Cervantes
Charlotte Waller
Rowan Kidd

only showing top 20 rows

Display Name & Gender of the Customer Having Net Worth Greater Than 5 Lakhs

```
df.filter(df['Net Worth']>500000).select(['Name','Gender']).show()
```

```
+-----+
|          Name|Gender|
+-----+
|   Harlan Barnes|    0|
|  Naomi Rodriquez|    1|
|  Jade Cunningham|    1|
|   Cedric Leach|    1|
|   Orli Casey|    1|
|  Marny O'Brien|    0|
|  Jerome Rowe|    1|
| Stephen Lindsey|    1|
|  Athena Wolf|    0|
|  Micah Wheeler|    1|
|  Castor Wood|    1|
|  Dahlia Cleveland|    1|
|  Coby Charles|    0|
| Rachel Ashley|    0|
|  Quincy Bell|    0|
|  Quon Carroll|    0|
|  Leilani Gross|    0|
| Francesca Cervantes|    1|
| Charlotte Waller|    1|
|   Rowan Kidd|    1|
+-----+
only showing top 20 rows
```

Display Name Of The Customers Of The Female Having Net Worth Of Greater Than 5 Lakhs

```
df.filter((df['Net Worth']>500000) & (df['Gender']=='0')).select('Name').show()
```

```
+-----+
|          Name|
+-----+
|   Harlan Barnes|
|  Marny O'Brien|
|  Athena Wolf|
|  Coby Charles|
| Rachel Ashley|
|  Quincy Bell|
|  Quon Carroll|
|  Leilani Gross|
|  Olga Serrano|
| Hedley Greene|
|  Cleo Moore|
|  Yen Santana|
| Emerald Hamilton|
|  Gage Marquez|
|  Ralph Rich|
|  Yasir Tyler|
|  Kadeem Larsen|
| Todd Maldonado|
| Solomon Atkinson|
|   Dean Snider|
+-----+
only showing top 20 rows
```

Display Name Of The Customers Of The Male Having Net Worth Of Greater Than 5 Lakhs

```
df.filter((df['Net Worth']>500000) & (df['Gender']=='1')).select('Name').show()
```

```
+-----+
|          Name|
+-----+
|  Naomi Rodriquez|
|  Jade Cunningham|
|   Cedric Leach|
|   Orli Casey|
|  Jerome Rowe|
| Stephen Lindsey|
|  Micah Wheeler|
+-----+
```

```
|      Castor Wood|
|      Dahlia Cleveland|
|Francesca Cervantes|
|      Charlotte Waller|
|      Rowan Kidd|
|      Lev Kramer|
|      Nissim Acosta|
|      Ila Farrell|
|Ferdinand Weber|
|      Desirae Stone|
|      Travis Burks|
|      Wing Parks|
|      Quamar Cummings|
+-----+
only showing top 20 rows
```

Display Average Annual Salary Of Male & Female Customers

```
df.groupby('Gender').mean().select(['Gender', 'avg(Annual Salary)']).show()
```

```
+-----+-----+
|Gender|avg(Annual Salary)|
+-----+-----+
|      1| 61705.5933223203|
|      0| 62559.12831267206|
+-----+-----+
```

Sort every row of the dataset into Descending Order

```
df.orderBy(df['Annual Salary'].desc()).show(5)
```

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|      Name|      Customer e-mail|Country|Gender|Age|Annual Salary|Credit Card Debt|  Net Worth|Car Purchase Amount|New_Ann
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|      Gemma Hendrix|  lobortis@non.co.uk|  USA|      1| 46|      100000.0|      17452.92179|188032.0778|      58350.31809|
|Flores, Caldwell U.|ut@Etiamvestibulu...|  USA|      0| 41|      92471.17612|      5404.397644|515717.7476|      59096.26978| 9
|      Vaughan|Aliquam@aaliquet.com|  USA|      0| 51|      92455.72807|      9877.169366|285326.3544|      61404.22578| 9
|Daugherty, Veda M.|venenatis.vel.fau...|  USA|      1| 48|      91083.73918|      13148.85597|387538.2487|      60960.83428| 9
|Glenna Graham|sodales@maurisSus...|  USA|      0| 46|      90556.62686|      13872.5667|479586.9387|      61593.52058| 9
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 5 rows
```

Handling Missing Values

```
df.show(5)
```

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|      Name|      Customer e-mail|Country|Gender|Age|Annual Salary|Credit Card Debt|  Net Worth|Car Purchase Amount|New_Annual
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|Martina Avila|cubilia.Curae.Pha...|  USA|      0| 42|      62812.09301|      11609.38091|238961.2505|      35321.45877| 63812
|Harlan Barnes|  eu.dolor@diam.co.uk|  USA|      0| 41|      66646.89292|      9572.957136|530973.9078|      45115.52566| 67646
|Naomi Rodriguez|vulputate.mauris...|  USA|      1| 43|      53798.55112|      11160.35506|638467.1773|      42925.70921| 54798
|Jade Cunningham|malesuada@digniss...|  USA|      1| 58|      79370.03798|      14426.16485|548599.0524|      67422.36313| 80370
|Cedric Leach|felis.ullamcorper...|  USA|      1| 57|      59729.1513|      5358.712177|560304.0671|      55915.46248| 6072
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 5 rows
```

```
from pyspark.sql.functions import mean
```

```
mean1=df.select(mean(df['Net Worth'])).collect()
```

```
mean1[0][0]
```

```
431475.71362506005
```

```
df.fillna(mean1[0][0]).show()
```

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|      Name|      Customer e-mail|Country|Gender|Age|Annual Salary|Credit Card Debt|  Net Worth|Car Purchase Amount|New_Annual
```

Martina Avila	cubilia.Curae.Pha...	USA	0	42	62812.09301	11609.38091	238961.2505	35321.45877	63812
Harlan Barnes	eu.dolor@diam.co.uk	USA	0	41	66646.89292	9572.957136	530973.9078	45115.52566	67646
Naomi Rodriquez	vulputate.mauris...	USA	1	43	53798.55112	11160.35506	638467.1773	42925.70921	54798
Jade Cunningham	malesuada@digniss...	USA	1	58	79370.03798	14426.16485	548599.0524	67422.36313	80370
Cedric Leach	felis.ullamcorper...	USA	1	57	59729.1513	5358.712177	560304.0671	55915.46248	6072
Carla Hester	mi@Aliquamerat.edu	USA	1	57	68499.85162	14179.47244	428485.3604	56611.99784	69495
Griffin Rivera	vehicula@at.co.uk	USA	1	47	39814.522	5958.460188	326373.1812	28925.70549	408
Orli Casey	nunc.est.mollis@S...	USA	1	50	51752.23445	10985.69656	629312.4041	47434.98265	52752
Marny Obrien	Phasellus@sedseme...	USA	0	47	58139.2591	3440.823799	630059.0274	48013.6141	5913
Rhonda Chavez	nec@nuncest.com	USA	1	43	53457.10132	12884.07868	476643.3544	38189.50601	54457
Jerome Rowe	ipsum.cursus@dui.org	USA	1	50	73348.70745	8270.707359	612738.6171	59045.51309	74348
Akeem Gibson	turpis.egestas.Fu...	USA	1	53	55421.65733	10014.96929	293862.5123	42288.81046	56421
Quin Smith	nulla@ipsum.edu	USA	0	44	37336.3383	10218.32092	430907.1673	28700.0334	3833
Tatum Moon	Cras.sed.leo@Sedd...	USA	0	48	68304.47298	9466.995128	420322.0702	49258.87571	69304
Sharon Sharpe	eget.metus@aaliqu...	USA	0	55	72776.00382	10597.63814	146344.8965	49510.03356	73776
Thomas Williams	aliquet.molestie@...	USA	1	53	64662.30061	11326.03434	481433.4324	53017.26723	65662
Blaine Bender	ultrices.posuere....	USA	0	45	63259.87837	11495.54999	370356.2223	41814.72067	64259
Stephen Lindsey	erat.eget.ipsum@t...	USA	1	48	52682.06401	12514.52029	549443.5886	43901.71244	53682
Sloane Mann	at.augue@augue.net	USA	1	52	54503.14423	7377.820914	431098.9998	44633.99241	55503
Athena Wolf	volutpat.Nulla.fa...	USA	0	59	55368.23716	13272.94647	566022.1306	54827.52403	56368

only showing top 20 rows