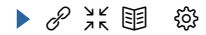




```
%pyspark
# Read in data from S3 Buckets
from pyspark import SparkFiles
url = "https://s3.amazonaws.com/dataviz-curriculum/day_1/demographics.csv"
spark.sparkContext.addFile(url)
df = spark.read.csv(SparkFiles.get("demographics.csv"), sep=",", header=True)

# Show DataFrame
df.show()
```




```
+---+-----+-----+-----+-----+-----+-----+-----+-----+
--+
| id |          name | age | height_meter | weight_kg | children | occupation | academic_degree | salary | location |
--+-----+-----+-----+-----+-----+-----+-----+-----+-----+
--+
| 0 | Darlena Avila | 58 | 1.87 | 53 | 1 | Choreographer | PhD | 68 | South Dakota |
| 1 | Yan Boyd | 65 | 1.8 | 40 | 0 | Cellarman | Bachelor | 73 | Delaware |
| 2 | Joette Lane | 32 | 1.8 | 73 | 1 | Veterinary Surgeon | Master | 69 | South Dakota |
| 3 | Jazmine Hunt | 61 | 1.79 | 89 | 0 | Hawker | PhD | 88 | Louisiana |
| 4 | Remedios Gomez | 23 | 1.64 | 51 | 2 | Choreographer | Bachelor | 83 | West Virginia |
| 5 | Myung Brewer | 20 | 1.68 | 60 | 4 | Window Dresser | Bachelor | 65 | South Dakota |
| 6 | Shaun Lynch | 31 | 1.56 | 62 | 0 | Weaver | Master | 72 | Louisiana |
| 7 | Melodi Mcdowell | 56 | 1.6 | 42 | 0 | Lighthouse Keeper | Master | 65 | Louisiana |
| 8 | Charlesetta Steve... | 30 | 1.62 | 44 | 3 | Millwright | Master | 87 | Louisiana |
| 9 | Merri Charles | 44 | 1.69 | 51 | 5 | Medical Supplier | PhD | 72 | West Virginia |
| 10 | Cassi Meyers | 55 | 1.82 | 72 | 5 | Manicurist | Bachelor | 73 | South Dakota |
```

Run



```
%pyspark
# Print the column names
df.columns
```



```
['id',
 'name',
 'age',
 'height_meter',
 'weight_kg',
 'children',
 'occupation',
 'academic_degree',
 'salary',
 'location']
```

```
%pyspark (/U4G66226D/spaces)
# Print out the first 10 rows
df.show(10)
```



```
+---+-----+---+-----+-----+-----+-----+-----+-----+
--+
| id |          name | age | height_meter | weight_kg | children | occupation | academic_degree | salary | location |
--+-----+---+-----+-----+-----+-----+-----+-----+-----+
| 0 | Darlena Avila | 58 | 1.87 | 53 | 1 | Choreographer | PhD | 68 | South Dakota |
| 1 | Yan Boyd | 65 | 1.8 | 40 | 0 | Cellarman | Bachelor | 73 | Delaware |
| 2 | Joette Lane | 32 | 1.8 | 73 | 1 | Veterinary Surgeon | Master | 69 | South Dakota |
| 3 | Jazmine Hunt | 61 | 1.79 | 89 | 0 | Hawker | PhD | 88 | Louisiana |
| 4 | Remedios Gomez | 23 | 1.64 | 51 | 2 | Choreographer | Bachelor | 83 | West Virginia |
| 5 | Myung Brewer | 20 | 1.68 | 60 | 4 | Window Dresser | Bachelor | 65 | South Dakota |
| 6 | Shaun Lynch | 31 | 1.56 | 62 | 0 | Weaver | Master | 72 | Louisiana |
| 7 | Melodi McDowell | 56 | 1.6 | 42 | 0 | Lighthouse Keeper | Master | 65 | Louisiana |
| 8 | Charlesetta Steve... | 30 | 1.62 | 44 | 3 | Millwright | Master | 87 | Louisiana |
| 9 | Merri Charles | 44 | 1.69 | 51 | 5 | Medical Supplier | PhD | 72 | West Virginia |
+---+-----+---+-----+-----+-----+-----+-----+-----+
--+
only showing top 10 rows
```

Run

Started

Juno ▾



```
%pyspark (/U4G66226D/spaces)
# Select the age, height_meter, and weight_kg columns and use describe to show the summary statistics
df.select(["age", "height_meter", "weight_kg"]).describe().show()
```



```
+-----+-----+-----+-----+
|summary|          age|    height_meter|    weight_kg|
+-----+-----+-----+-----+
|  count|         1000|           1000|           1000|
|   mean|    42.93311.751949999999995|         64.011|
| stddev|14.255445581556843|0.1436897499623555|15.005733939099779|
|   min|          18|           1.5|           38|
|   max|          67|           2|           90|
+-----+-----+-----+-----+
```

```
%pyspark
# Print the schema to see the types
df.printSchema()
```

```
root
 |-- id: string (nullable = true)
 |-- name: string (nullable = true)
 |-- age: string (nullable = true)
 |-- height_meter: string (nullable = true)
 |-- weight_kg: string (nullable = true)
 |-- children: string (nullable = true)
 |-- occupation: string (nullable = true)
 |-- academic_degree: string (nullable = true)
 |-- salary: string (nullable = true)
 |-- location: string (nullable = true)
```

```
%pyspark (/U4G66226D/spaces)
# Rename the Salary column to `Salary (1k)` and show only this new column
df = df.withColumnRenamed('Salary', 'Salary (1k)')
df.select("Salary (1k)").show()
```



```
+-----+
|Salary (1k)|
+-----+
|          68|
|          73|
|          69|
|          88|
|          83|
|          65|
|          72|
|          65|
|          87|
|          72|
|          73|
|          90|
|          78|
|          69|
|          75|
|          77|
|          76|
|          90|
|          79|
|          77|
+-----+
only showing top 20 rows
```

Run

Started

Juno ▾





```
%pyspark (/U4G66226D/spaces)
# Create a new column called `Salary` where the values are the `Salary (1k)` * 1000
# Show the columns `Salary` and `Salary (1k)`
df = df.withColumn("Salary", df["Salary (1k)"] * 1000)
df.select(["Salary", "Salary (1k)"]).show()
```

```
+-----+-----+
| Salary|Salary (1k)|
+-----+-----+
|68000.0|      68|
|73000.0|      73|
|69000.0|      69|
|88000.0|      88|
|83000.0|      83|
|65000.0|      65|
|72000.0|      72|
|65000.0|      65|
|87000.0|      87|
|72000.0|      72|
|73000.0|      73|
|90000.0|      90|
|78000.0|      78|
|69000.0|      69|
|75000.0|      75|
|77000.0|      77|
|76000.0|      76|
|90000.0|      90|
|79000.0|      79|
|77000.0|      77|
+-----+-----+
only showing top 20 rows
```

Interpreter: spark.



Run

Started

Juno ▾

