%pyspark

a١

```
# 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()
   l idl
                     namelagelheight_meterlweight_kglchildrenl
                                                                occupation|academic_degree|salary|
                                                                                                  locatio
    -+
    1 01
              Darlena Avilal 581
                                    1.871
                                               53 l
                                                       11
                                                             Choreographer
                                                                                    PhDI
                                                                                           681 South Dakot
   a١
    1 11
                  Yan Boydl 651
                                     1.81
                                               401
                                                       Ø1
                                                                 Cellarmanl
                                                                               Bachelori
                                                                                           73 I
                                                                                                  Delawar
   еl
   1 21
                Joette Lanel 321
                                     1.81
                                               73 I
                                                       1|Veterinary Surgeon|
                                                                                           691 South Dakot
                                                                                 Masterl
   a١
      31
               Jazmine Huntl 61
                                    1.791
                                               891
                                                       Ø1
                                                                    Hawkerl
                                                                                    PhDI
                                                                                           881
                                                                                                 Louisian
   a١
    1 41
             Remedios Gomez | 23|
                                    1.641
                                               51 l
                                                       21
                                                             Choreographer
                                                                               Bachelori
                                                                                           83|West Virgini
   a١
               Myung Brewerl 201
      51
                                                            Window Dresserl
                                                                                Bachelori
                                                                                           651 South Dakot
                                    1.681
                                               60 l
                                                       41
   a١
   1 61
                Shaun Lynchl 311
                                    1.561
                                               62 I
                                                       Ø1
                                                                   Weaverl
                                                                                           72 I
                                                                                                 Louisian
                                                                                 Masterl
   a١
   1 71
            Melodi Mcdowell| 56|
                                                       0| Lighthouse Keeper|
                                                                                                 Louisian
                                     1.61
                                               421
                                                                                 Masterl
                                                                                           65 I
   al
      8|Charlesetta Steve...| 30|
                                    1.621
                                                       31
                                                                Millwrightl
                                                                                           87 I
                                               441
                                                                                                 Louisian
                                                                                 Masterl
   al
   1 91
              Merri Charles | 44|
                                    1.691
                                               51 l
                                                       51 Medical Supplier
                                                                                    PhDI
                                                                                           72|West Virgini
Run al
                                                                                  Started
                                                                                           Juno ✓ 
   | 10|
               Cassi Meyers | 55|
                                    1.821
                                               721
                                                       51
                                                                Manicuri st L
                                                                                           731 South Dakot
                                                                                Bachelorl
```

▶ ♂ 光 圃 袋

...

# Print the column names

78 I

5| Medical Physicist|

PhDI

901

Delawar

Interpreter: spark.pyspark. FINISHED Took 27 sec 981 millisec. Updated by ars0107 on February 01 2019, 12:26:56 PM (CST)

 $Q_{V}$ 

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

%pyspark

df.columns

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

l idl nl						occupationlaca	J		
++ -+   0	Darlena Avila	·	1.87	531	+ 1l				South Dakot
al   1	Yan Boydl		1.81	401	01	Cellarmanl	Bachelori	731	
el I 21	Joette Lanel	321	1.8	731	1 Ve	eterinary Surgeonl	Masterl	691	South Dakot
al I 31	Jazmine Huntl	611	1.79	891	01	Hawkerl	PhDI	881	Louisian
al I 41 al	Remedios Gomezl	231	1.641	511	21	Choreographerl	Bachelor	8311	Vest Virgini
1 51 1 31	Myung Brewerl	201	1.68	601	41	Window Dresserl	Bachelor	651	South Dakot
1 61 al	Shaun Lynchl	311	1.561	621	01	Weaverl	Masterl	721	Louisian
   7  	Melodi McdowellI	561	1.61	421	0 l l	ighthouse Keeperl	Masterl	651	Louisian
	arlesetta Steve	301	1.621	441	31	Millwright	Masterl	871	Louisian
l 91 al	Merri CharlesI	441	1.691	511	51	Medical Supplier	PhDI	72 I V	Vest Virgini

Started Juno V 🔯 😕 👓

## %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()

```
Isummary
                  agel
                         height_meterl
                                           weight_kal
                 10001
 countl
                                10001
                                               10001
                42.933|1.7519499999999995|
  meanl
                                             64.011
| stddev|14.255445581556843|0.1436897499623555|15.005733939099779|
   minl
                   181
                                 1.51
                   67 I
                                  21
                                                901
   maxl
```

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

## root

- I-- id: string (nullable = true)
- I-- name: string (nullable = true)
- I-- age: string (nullable = true)
- I-- height\_meter: string (nullable = true)
- I-- weight\_kg: string (nullable = true)
- I-- children: string (nullable = true)
- I-- occupation: string (nullable = true)
- I-- academic\_degree: string (nullable = true)
- I-- salary: string (nullable = true)
- I-- 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 I 73 I 69 I 88 I 83 I 65 I 72 I 65 I 87 I 72 I 73 I 901 78 I 69 I 75 I 77 I 76 I 901 79 I 77 I only showing top 20 rows

•••

```
%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)|
+----+
168000.01
                  681
173000.01
                  73 l
169000.01
                  69 I
188000.01
                  881
                  831
183000.01
165000.01
                  65 l
172000.01
                  721
165000.01
                  65 l
                  871
187000.01
172000.01
                  721
173000.01
                  73 l
                  901
190000.01
178000.01
                  78 l
169000.01
                  69 I
175000.01
                  75 l
177000.01
                  771
                  761
176000.01
190000.01
                  901
179000.01
                  79 I
177000.01
                  771
+----+
only showing top 20 rows
                                                                                                              Q_{V} \vee
Interpreter: spark.
```

•••