```
%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.option('header', 'true').csv(SparkFiles.get("demographics.csv"), inferSchema=True, sep=',')

# Show DataFrame
df.show()
```

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

```
| 11|     Shawnee Harmon| 66|       1.63|       78|       5| Medical Physicist|          PhD|    90|     Delawar
```

```
%pyspark
# What occupation had the highest salary?
df.orderBy(df["Salary"].desc()).select("occupation", "Salary").limit(1).show()
```

```
+-----------------+------+
|       occupation|Salary|
+-----------------+------+
|Medical Physicist|    90|
+-----------------+------+
```

```
%pyspark
# What occupation had the lowest salary?
df.orderBy(df["Salary"]).select("occupation", "Salary").limit(1).show()
```

```
+--------------+------+
|    occupation|Salary|
+--------------+------+
|Window Dresser|    65|
+--------------+------+
```

```
%pyspark
# What is the mean salary of this dataset?
from pyspark.sql.functions import mean
df.select(mean("Salary")).show()
```

```
+-----------+
|avg(Salary)|
+-----------+
|     77.738|
+-----------+
```

```
%pyspark (/U4G66226D/spaces)
# What is the max and min of the Salary column?
from pyspark.sql.functions import max, min
df.select(max("Salary"), min("Salary")).show()
```

```
+-----------+-----------+
|max(Salary)|min(Salary)|
+-----------+-----------+
|         90|         65|
+-----------+-----------+
```

```
%pyspark (/U4G66226D/spaces)
# Show all of the occupations where salaries were above 80k
from pyspark.sql.functions import count
df.filter("Salary > 80").select("occupation").show()
```

```
+--------------------+
|          occupation|
+--------------------+
|              Hawker|
|       Choreographer|
|          Millwright|
|   Medical Physicist|
|           Scientist|
|     Claims Adjustor|
| Planning Technician|
|       Booking Clerk|
|      Sub-Postmaster|
|        Shelf Filler|
|             Chemist|
|        Betting Shop|
|     Hire Car Driver|
|    Heating Engineer|
|    Vehicle Assessor|
|   Building Surveyor|
|Advertising Contr...|
|   Medical Physicist|
|            Labourer|
|   Technical Analyst|
+--------------------+
only showing top 20 rows
```

(/U4G66226D/spaces)

```
%pyspark
# BONUS
# What is the average age and height for each academic degree type?
# HINT: You will need to use `groupby` to solve this
avg_df = df.groupBy("academic_degree").avg()
avg_df.select("academic_degree", "avg(age)", "avg(height_meter)").show()
```

```
+---------------+------------------+------------------+
|academic_degree|          avg(age)| avg(height_meter)|
+---------------+------------------+------------------+
|            PhD| 43.159763313609 47|1.7438165680473379|
|         Master|43.139318885448915|1.7549226006191951|
|       Bachelor| 42.51032448377581| 1.757227138643069|
+---------------+------------------+------------------+
```

Interpreter: spark.

Run

Started

Juno