

# PKA-WEEK1-PROGRAM

January 4, 2024

```
[1]: import pyspark
import os
import sys
from pyspark import SparkContext
from pyspark import SparkConf
os.environ['PYSPARK_PYTHON'] = sys.executable
os.environ['PYSPARK_DRIVER_PYTHON'] = sys.executable

from pyspark.sql import SparkSession
```

```
[2]: spark=SparkSession.builder.appName('DATA ANALYSIS').getOrCreate()
```

```
[3]: df=spark.read.csv('data.csv',header="True", inferSchema="True")
```

```
[13]: df.printSchema()
df.show()
df.head()
df.summary().show()
df.select('age').summary().show()
```

```
root
 |-- name: string (nullable = true)
 |-- age: double (nullable = true)
```

```
+----+----+
|name| age|
+----+----+
| PKA|45.0|
|  GR|43.0|
| DRS|43.0|
|  SR|43.0|
|  SM|17.0|
+----+----+
```

```
+-----+-----+-----+
|summary|name|          age|
+-----+-----+-----+
|  count|  5|          5|
```

	mean	null	38.2
	stddev	null	11.882760622010357
	min	DRS	17.0
	25%	null	43.0
	50%	null	43.0
	75%	null	43.0
	max	SR	45.0
+-----+-----+-----+			

+-----+-----+	
summary	age
+-----+-----+	
count	5
mean	38.2
stddev	11.882760622010357
min	17.0
25%	43.0
50%	43.0
75%	43.0
max	45.0
+-----+-----+	

[ ]: