

```
In [1]: spark
sc
```

Starting Spark application

ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
0	application_1646933635929_0002	pyspark	idle	<a href="#">Link</a>	<a href="#">Link</a>	✓

SparkSession available as 'spark'.

```
<SparkContext master=yarn appName=livy-session-0>
```

```
In [3]: environment.describe()
```

```
DataFrame[summary: string, Area Code: string, Area: string, Months Code: string, Months: string, Element Code: string, Element: string, Unit: string, Y1961: string, Y1962: string, Y1963: string, Y1964: string, Y1965: string, Y1966: string, Y1967: string, Y1968: string, Y1969: string, Y1970: string, Y1971: string, Y1972: string, Y1973: string, Y1974: string, Y1975: string, Y1976: string, Y1977: string, Y1978: string, Y1979: string, Y1980: string, Y1981: string, Y1982: string, Y1983: string, Y1984: string, Y1985: string, Y1986: string, Y1987: string, Y1988: string, Y1989: string, Y1990: string, Y1991: string, Y1992: string, Y1993: string, Y1994: string, Y1995: string, Y1996: string, Y1997: string, Y1998: string, Y1999: string, Y2000: string, Y2001: string, Y2002: string, Y2003: string, Y2004: string, Y2005: string, Y2006: string, Y2007: string, Y2008: string, Y2009: string, Y2010: string, Y2011: string, Y2012: string, Y2013: string, Y2014: string, Y2015: string, Y2016: string, Y2017: string, Y2018: string, Y2019: string]
```

```
In [4]: environment.printSchema()
```

```
root
|-- Area Code: string (nullable = true)
|-- Area: string (nullable = true)
|-- Months Code: string (nullable = true)
|-- Months: string (nullable = true)
|-- Element Code: string (nullable = true)
|-- Element: string (nullable = true)
|-- Unit: string (nullable = true)
|-- Y1961: string (nullable = true)
|-- Y1962: double (nullable = true)
|-- Y1963: double (nullable = true)
|-- Y1964: double (nullable = true)
|-- Y1965: double (nullable = true)
|-- Y1966: double (nullable = true)
|-- Y1967: double (nullable = true)
|-- Y1968: double (nullable = true)
|-- Y1969: double (nullable = true)
|-- Y1970: double (nullable = true)
|-- Y1971: double (nullable = true)
|-- Y1972: double (nullable = true)
|-- Y1973: double (nullable = true)
|-- Y1974: double (nullable = true)
|-- Y1975: double (nullable = true)
|-- Y1976: double (nullable = true)
|-- Y1977: double (nullable = true)
|-- Y1978: double (nullable = true)
|-- Y1979: double (nullable = true)
|-- Y1980: double (nullable = true)
|-- Y1981: double (nullable = true)
|-- Y1982: double (nullable = true)
|-- Y1983: double (nullable = true)
|-- Y1984: double (nullable = true)
|-- Y1985: double (nullable = true)
|-- Y1986: double (nullable = true)
|-- Y1987: double (nullable = true)
|-- Y1988: double (nullable = true)
|-- Y1989: double (nullable = true)
|-- Y1990: double (nullable = true)
|-- Y1991: double (nullable = true)
|-- Y1992: double (nullable = true)
|-- Y1993: double (nullable = true)
|-- Y1994: double (nullable = true)
|-- Y1995: double (nullable = true)
|-- Y1996: double (nullable = true)
|-- Y1997: double (nullable = true)
|-- Y1998: double (nullable = true)
|-- Y1999: double (nullable = true)
|-- Y2000: double (nullable = true)
|-- Y2001: double (nullable = true)
|-- Y2002: double (nullable = true)
|-- Y2003: double (nullable = true)
|-- Y2004: double (nullable = true)
|-- Y2005: double (nullable = true)
|-- Y2006: double (nullable = true)
|-- Y2007: double (nullable = true)
|-- Y2008: double (nullable = true)
|-- Y2009: double (nullable = true)
|-- Y2010: double (nullable = true)
|-- Y2011: double (nullable = true)
|-- Y2012: double (nullable = true)
|-- Y2013: double (nullable = true)
|-- Y2014: double (nullable = true)
|-- Y2015: double (nullable = true)
|-- Y2016: double (nullable = true)
|-- Y2017: double (nullable = true)
|-- Y2018: double (nullable = true)
|-- Y2019: double (nullable = true)
```

```
In [4]: filter_element=environment.filter(environment['element code']=='7271')
```

```
In [44]: (filter_element).describe().show(5)
```

```
+-----+-----+-----+-----+-----+-----+-----+-----+
| summary | area | area | months | months | element | element | unit | y1961 |
| y1962 | y1963 | y1964 | y1965 | y1966 | y1967 | y1968 | y1969 |
| y1970 | y1971 | y1972 | y1973 | y1974 | y1975 | y1976 | y1977 |
| y1978 | y1979 | y1980 | y1981 | y1982 | y1983 | y1984 | y1985 |
| y1986 | y1987 | y1988 | y1989 | y1990 | y1991 | y1992 | y1993 |
| y1994 | y1995 | y1996 | y1997 | y1998 | y1999 | y2000 | y2001 |
| y2002 | y2003 | y2004 | y2005 | y2006 | y2007 | y2008 | y2009 |
| y2010 | y2011 | y2012 | y2013 | y2014 | y2015 | y2016 | y2017 |
+-----+-----+-----+-----+-----+-----+-----+-----+
| count | 4720 | 4720 | 4720 | 4720 | 4720 | 4720 | 4720 | | | |
| 4057 | 4043 | 4068 | 4063 | 4068 | 4045 | 4094 | 4078 |
| 4069 | 4033 | 4066 | 4063 | 4081 | 4079 | 4124 | 4115 |
| 4048 | 4030 | 4063 | 4038 | 4070 | 4078 | 4078 | 4069 |
| 4064 | 4055 | 4037 | 4325 | 4328 | 4352 | 4370 | 4370 |
| 4391 | 4319 | 4359 | 4333 | 4358 | 4302 | 4337 | 4337 |
| 4383 | 4394 | 4402 | 4446 | 4462 | 4428 | 4406 | 4406 |
| 4410 | 4430 | 4388 | 4428 | 4406 | 4398 | 4389 | 4389 |
| 4402 | 4391 | 4401 | 4401 | 7271 | 0 | null | null |
| mean | 837.5822033898305 | null | 7009.901906779661 | null | 7271.0 | null | null |
| 0.02680823268424937 | -0.03014840465001... | -0.12494904300273428 | -0.22968257107540185 | 0.089441377625794 | -0.12983055419323222 | -0.166969377560821 | 0.11335018404907983 |
| 0.07958172074729604 | -0.17543682399213317 | -0.05171309465424... | 0.19729534432589768 | -0.12973025516403394 | -0.03152985991644... | -0.21419414827671743 | 0.18753221839645806 |
| 0.05503340648622281 | 0.22958837958478064 | 0.2171964198136341 | 0.2164322929466699 | 0.15454224308300343 | 0.3547285359801482 | 0.08378218065468872 | 0.07549702823179809 |
| 0.13852167487684725 | 0.4099071253871257 | 0.43981273840805905 | 0.28240944881889823 | 0.5825003699136863 | 0.3316319048798619 | 0.25795606936416177 | 0.23994500924214396 |
| 0.5646989889705883 | 0.6207112128146471 | 0.3171398314734688 | 0.5750784903912951 | 0.95197247075017 | 0.7373464112624044 | 0.68902592325785 | 0.8097391910739187 |
| 0.923895780493429 | 0.8631204654346326 | 0.7850025034137454 | 0.8817462517037687 | 0.9043398560503844 | 1.063718511876532 | 0.820176829682026 | 0.9446713572401281 |
| 1.0887517006802718 | 0.8701907449200926 | 0.904098961075675 | 0.9785228093947611 | 1.1365946436677246 | 1.3250166866757614 | 1.449424698108909 | 1.298684688778311 |
| 3.131286722842173 | 1.4093372216330877 | 1.298684688778311 | 3.131286722842173 | 1.4093372216330877 | 1.298684688778311 | 3.131286722842173 | 1.4093372216330877 |
| stddev | 1798.3097465403528 | null | 1.031449379256494 | null | 0.0 | null | null |
| 0.7575705698774442 | 1.0073570159504834 | 0.7754948643113622 | 0.6971704228677359 | 0.8428614299155258 | 0.7209283542135018 | 0.7490126794095348 | 0.8241612715325795 |
| 0.6936041133744678 | 0.7048479409128894 | 0.8414191683725599 | 0.7664526896738458 | 0.8126023463580265 | 0.7665975324129207 | 0.7297166393576565 | 0.7561653447527314 |
| 0.693848800641774 | 0.7618242528457145 | 0.7052319271340701 | 0.8145275211695147 | 0.751826435799314 | 0.92444529704423 | 0.7653688568323329 | 0.8840518143577574 |
| 0.8124662807023426 | 0.9480040893771935 | 0.7949241077577988 | 0.9335547621272766 | 0.9130773991778328 | 0.7676871163922145 | 1.0397161259168217 | 0.9975419847324475 |
| 0.9368662016857472 | 0.9432761409736475 | 0.8632656654496923 | 0.9222080733334962 | 0.9699153157652295 | 0.8890735125767881 | 0.986740478964051 | 0.9802468603966342 |
| 1.054938693862054 | 0.8124740613514768 | 1.0903042512059813 | 0.9165798383637689 | 1.0851459942465984 | 0.8481946403315174 | 0.9801440237973198 | 0.9652357654464073 |
| 0.9866350172833419 | 0.9241995811465269 | 1.0232299476699707 | 0.835519882246237 | 0.835519882246237 | 1.0232299476699707 | 0.835519882246237 | 1.0232299476699707 |
| min | 1 | Afghanistan | 7001 | April | 7271 | Temperature change | C | -0.001000 |
| -5.391 | -8.483 | -7.309 | -4.728 | -8.147 | -6.531 | -8.4 | 07 |
| -6.169 | -4.263 | -6.495 | -8.228 | -6.319 | -5.784 | -6.591 | -4.55 |
| -6.101 | -5.437 | -8.411 | -6.345 | -8.75 | -8.963 | -4.51 | -5.311 |
| -4.03 | -4.598 | -5.414 | -7.389 | -5.099 | -4.862 | -6.17 | -4.027 |
| -4.059 | -6.031 | -3.035 | -3.596 | -5.493 | -3.543 | -3.584 | -6.118 |
| -5.025 | -5.171 | -4.981 | -3.19 | -9.334 | -3.543 | -6.072 | -4.854 |
| -5.785 | -3.642 | -5.367 | -4.068 | 0.003000 | 99 | Zimbabwe | 7020 | Sep | Oct | Nov |
| 4.366 | 4.666 | 5.233 | 5.144 | 5.771 | 4.768 | 3.181 | 4.411 |
| 2.999 | 3.665 | 9.475 | 6.304 | 6.91 | 6.15 | 7.689 | 4.875 |
| 5.956 | 5.483 | 4.519 | 6.144 | 5.411 | 6.513 | 9.303 | 5.948 |
| 6.845 | 4.982 | 6.841 | 9.73 | 6.11 | 6.017 | 5.989 | 7.221 |
| 7.27 | 5.637 | 6.816 | 7.017 | 5.836 | 6.092 | 6.719 | 6.171 |
| 7.461 | 7.651 | 11.331 | 7.655 | 8.298 | 6.415 | 7.19 | 8.086 |
| 7.086 | 7.251000 | 10.478 | 7.389 | 9.228 | 7.215000 |
+-----+-----+-----+-----+-----+-----+-----+-----+
| area | y2017 | y2018 | y2019 |
+-----+-----+-----+-----+
| Afghanistan | 1.201 | 1.996 | 2.051000 |
| 2.0493333333333332 |
| Afghanistan | 0.323 | 2.705 | 0.086000 |
| 0.834 | 4.418 | 0.234000 |
| Afghanistan | 1.252 | 1.442 | 0.899000 |
| Afghanistan | 3.28 | 0.855 | 0.647000 |
+-----+-----+-----+-----+
only showing top 5 rows
```

```
In [5]: New_environment=filter_element.select('area','y2017','y2018','y2019')
```

```
In [54]: New_environment.printSchema()
```

```
root
-- area: string (nullable = true)
-- y2017: double (nullable = true)
-- y2018: double (nullable = true)
-- y2019: string (nullable = true)
```

```
In [49]: New_environment.show(5)
```

```
+-----+-----+-----+
| area | y2017 | y2018 | y2019 |
+-----+-----+-----+
| Afghanistan | 1.201 | 1.996 | 2.051000 |
| Afghanistan | 0.323 | 2.705 | 0.086000 |
| Afghanistan | 0.834 | 4.418 | 0.234000 |
| Afghanistan | 1.252 | 1.442 | 0.899000 |
| Afghanistan | 3.28 | 0.855 | 0.647000 |
+-----+-----+-----+
only showing top 5 rows
```

```
In [6]: New_environment.select('area').distinct().show()
```

```
+-----+
| area |
+-----+
| Guyana |
| Turkey |
| Argentina |
| Angola |
| Albania |
| Nicaragua |
| Peru |
| China |
| Somalia |
| Martinique |
| South-Eastern Asia |
| Serbia and Montenegro |
| Croatia |
| Andorra |
| Suriname |
| Norway |
| Guadeloupe |
| Niger |
| Ireland |
| Congo |
+-----+
only showing top 20 rows
```

```
In [56]: print("The Variable", 'area', "has: ", environment.filter(environment['area'].isNull()).count(), "Null Values")
print("The Variable", 'y2017', "has: ", environment.filter(environment['y2017'].isNull()).count(), "Null Values")
print("The Variable", 'y2018', "has: ", environment.filter(environment['y2018'].isNull()).count(), "Null Values")
print("The Variable", 'y2019', "has: ", environment.filter(environment['y2019'].isNull()).count(), "Null Values")
```

```
The Variable area has: 630 Null Values
The Variable y2017 has: 1299 Null Values
The Variable y2018 has: 1317 Null Values
The Variable y2019 has: 1291 Null Values
```

## Shape

```
In [50]: print((New_environment.count(),len(New_environment.columns)))
```

```
(4720, 4)
```

```
In [44]: (environment.select('y2017','y2018','y2019')).describe().show()
```

```
+-----+-----+-----+-----+
| summary | y2017 | y2018 | y2019 |
+-----+-----+-----+-----+
| count | 8357 | 8339 | 8365 |
| mean | 1.0017628335527087 | 1.0108722868449487 | 1.0322196225439515 |
| stddev | 0.8094999814416658 | 0.8729501003990419 | 0.7699041633003846 |
| min | -3.584 | -2.216 | -0.003000 |
| max | 7.389 | 9.228 | 7.215000 |
+-----+-----+-----+-----+
only showing top 5 rows
```

```
In [7]: environment_avg=New_environment.withColumn('Avg3LastYears', (New_environment.y2017+New_environment.y2018+New_environment.y2019)/3).select('area','y2017','y2018','y2019','Avg3LastYears')
```

```
In [56]: environment_avg.show(5)
```

```
+-----+-----+-----+-----+
| area | y2017 | y2018 | y2019 | Avg3LastYears |
+-----+-----+-----+-----+
| Afghanistan | 1.201 | 1.996 | 2.051000 | 2.0493333333333332 |
| Afghanistan | 0.323 | 2.705 | 0.086000 | 0.8226666666666667 |
| Afghanistan | 0.834 | 4.418 | 0.234000 | 1.8286666666666667 |
| Afghanistan | 1.252 | 1.442 | 0.899000 | 1.1976666666666667 |
| Afghanistan | 3.28 | 0.855 | 0.647000 | 1.594 |
+-----+-----+-----+-----+
only showing top 5 rows
```

```
In [8]: environment_avg.sort(environment_avg.Avg3LastYears.desc()).show(5)
```

```
+-----+-----+-----+-----+
| area | y2017 | y2018 | y2019 | Avg3LastYears |
+-----+-----+-----+-----+
| Svalbard and Jan ... | 7.267000 | 9.223000 | 8.841000 | 8.443333333333333 |
| Svalbard and Jan ... | 5.654000 | 8.385000 | 4.831000 | 6.29 |
| Svalbard and Jan ... | 4.062000 | 9.228000 | 4.827000 | 6.038999999999999 |
| Mongolia | 2.405000 | 6.984000 | 4.527000 | 4.638666666666665 |
| Svalbard and Jan ... | 6.705000 | 5.826000 | 1.375000 | 4.635333333333333 |
+-----+-----+-----+-----+
only showing top 5 rows
```

```
In [9]: finalDf=environment_avg
```

```
In [10]: finalDf.coalesce(1).write.format("csv").option("header","true").save("s3://trabajodatalake/refined/Change_climatic/Change_clima_tic_PVSPARK")
```