SummaryStatistics

November 29, 2017

In [29]:
```
from pyspark.sql.types import StructType, StructField, FloatType, LongType, StringType
from pyspark.shell import spark

feats = []
f = open('features.txt')
for line_num, line in enumerate(f):
    if line_num == 0:
        # Timestamp
        feats.append(StructField(line.strip(), LongType(), True))
    elif line_num == 1:
        # Geohash
        feats.append(StructField(line.strip(), StringType(), True))
    else:
        # Other features
        feats.append(StructField(line.strip(), FloatType(), True))

schema = StructType(feats)
```

In [30]:
```
df = spark.read.format('csv').option('sep', '\t').schema(schema).load('inputs/nam_mini
```

In [31]:
```
len(df.columns)
```

Out[31]:
```
58
```

In [34]:
```
import pyspark.sql.functions as sf
for i in range(2,5):
    summary_values = df.select(sf.max(df.columns[i]).alias("Max"),
              sf.min(df.columns[i]).alias("Min"),
              sf.avg(df.columns[i]).alias("Avg"),
              sf.stddev(df.columns[i]).alias("Std_Dev"))
    print("Feature: " + df.columns[i] + "\n")
    print("Max value: ")
    summary_values.select(summary_values.Max).show()
    print("\n")
    print("Min value: ")
    summary_values.select(summary_values.Min).show()
    print("\n")
    print("Average: ")
```

```
            summary_values.select(summary_values.Avg).show()
            print("\n")
            print("Std. Dev: ")
            summary_values.select(summary_values.Std_Dev).show()
            print("\n")
```

```
Feature: geopotential_height_lltw

Max value:
+--------+
|     Max|
+--------+
|4902.578|
+--------+



Min value:
+---------+
|      Min|
+---------+
|-5817.172|
+---------+



Average:
+------------------+
|               Avg|
+------------------+
|1571.8196034524215|
+------------------+



Std. Dev:
+------------------+
|           Std_Dev|
+------------------+
|1850.1112628212659|
+------------------+



Feature: water_equiv_of_accum_snow_depth_surface

Max value:
+------+
```

```
|   Max|
+------+
|5501.0|
+------+
```

```
Min value:
+---+
|Min|
+---+
|0.0|
+---+
```

```
Average:
+------------------+
|               Avg|
+------------------+
|20.543379364668635|
+------------------+
```

```
Std. Dev:
+-----------------+
|          Std_Dev|
+-----------------+
|68.09250334632227|
+-----------------+
```

```
Feature: drag_coefficient_surface

Max value:
+-------+
|    Max|
+-------+
|11800.0|
+-------+
```

```
Min value:
+---+
|Min|
```

```
+---+
|0.0|
+---+
```

Average:

```
+-------------------+
|                Avg|
+-------------------+
|0.08109455386769765|
+-------------------+
```

Std. Dev:

```
+------------------+
|           Std_Dev|
+------------------+
|18.554648814462873|
+------------------+
```