Arvid Edenheim (arved490)
Jonatan Bjurenfalk (jonbj587)

TDDE31 - Big Data Analytics

BDA1 Spark

Sample results are given along with each answer. See attached files for further results and code.

1a)

Max, Top 20

```
86200, 1975:  36.1
95160, 1975:  35.8
96550, 1975:  35.6
106100, 1975:  35.5
75240, 1975:  35.4
63600, 1992:  35.4
63050, 1992:  35.2
97390, 1975:  35.0
97200, 1975:  35.0
75240, 1992:  35.0
98210, 1975:  35.0
97190, 1975:  35.0
96030, 1975:  35.0
76000, 1992:  35.0
85040, 1992:  35.0
96350, 1975:  35.0
98040, 1975:  35.0
85220, 1975:  35.0
95350, 1975:  35.0
82110, 1975:  34.9
```

Min, Top 20

```
95530, 2010:  15.2
99090, 1979:  13.1
53220, 1984:  12.0
117160, 2001:  8.0
89560, 2010:  7.9
104390, 1998:  7.5
84390, 1986:  5.3
71140, 2009:  4.9
107530, 1970:  4.1
149160, 1951:  3.4
65640, 1955:  3.4
65640, 1957:  3.3
71140, 2007:  1.6
71140, 2004:  1.6
81350, 1966:  1.5
71140, 2008:  0.5
71500, 2008:  0.2
```

53220, 1985: 0.2
163950, 2013: 0.1
83210, 1982: -0.2
The reason for the highest average being 15 degrees in the minimum top 20 is that there are only two data points, both being in august, in the year 2010. We assume that the same goes for the other high averages.

1b)
In 1b) we get the same results as in 1a).

Python Elapsed time: 1872.5362420082092s

Spark Elapsed time: 1367.294s

The time difference is almost 500 seconds. This is because spark distributes the workload across multiple nodes. The nature of parallelization in MapReduce.

2a) Using all readings.
1950-03: 81
1950-04: 352
1950-05: 2802
1950-06: 4886
1950-07: 5811
1950-08: 5954
1950-09: 3612
1950-10: 1248
1950-11: 2
1950-12: 1
1951-02: 1
1951-04: 690
1951-05: 3345
1951-06: 9918
1951-07: 12578
1951-08: 13933
1951-09: 9601
1951-10: 3169
1951-11: 70
1951-12: 6

2b) Using only distinct readings
1950-03: 26
1950-04: 36
1950-05: 46
1950-06: 47
1950-07: 49
1950-08: 49

```
1950-09: 50
1950-10: 46
1950-11: 2
1950-12: 1
1951-02: 1
1951-04: 88
1951-05: 98
1951-06: 110
1951-07: 111
1951-08: 112
1951-09: 112
1951-10: 113
1951-11: 22
1951-12: 5
```

We can see that when we only use distinct readings for each month we get fewer readings for each month and year.

3)

```
1967-06, 102190, 15.066666666666661
1975-08, 102190, 17.29354838709677
2002-08, 102190, 17.300000000000004
2004-09, 102190, 9.330000000000002
1986-06, 102200, 15.218333333333334
1991-05, 102210, 8.16653225806452
2000-01, 102390, -4.7984615384615354
2005-05, 102390, 7.657819225251075
1962-06, 102540, 12.281111111111109
1967-10, 102540, 4.554838709677419
1974-11, 102540, -1.5955555555555552
1974-12, 102540, -4.446236559139784
1978-01, 102540, -7.055913978494625
1983-03, 102540, -1.3462365591397853
1990-02, 102540, 1.5000000000000004
1991-01, 102540, -8.16236559139785
1992-04, 102540, 2.3455555555555554
1995-03, 102540, -1.420430107526882
2008-10, 102540, 4.827956989247313
2010-02, 102540, -11.007142857142856
```

Looking at the averages for each month and year we can see that they seem to be what we expect them to be given a specific month.

4)

The result file in this exercise was empty as expected.

5)

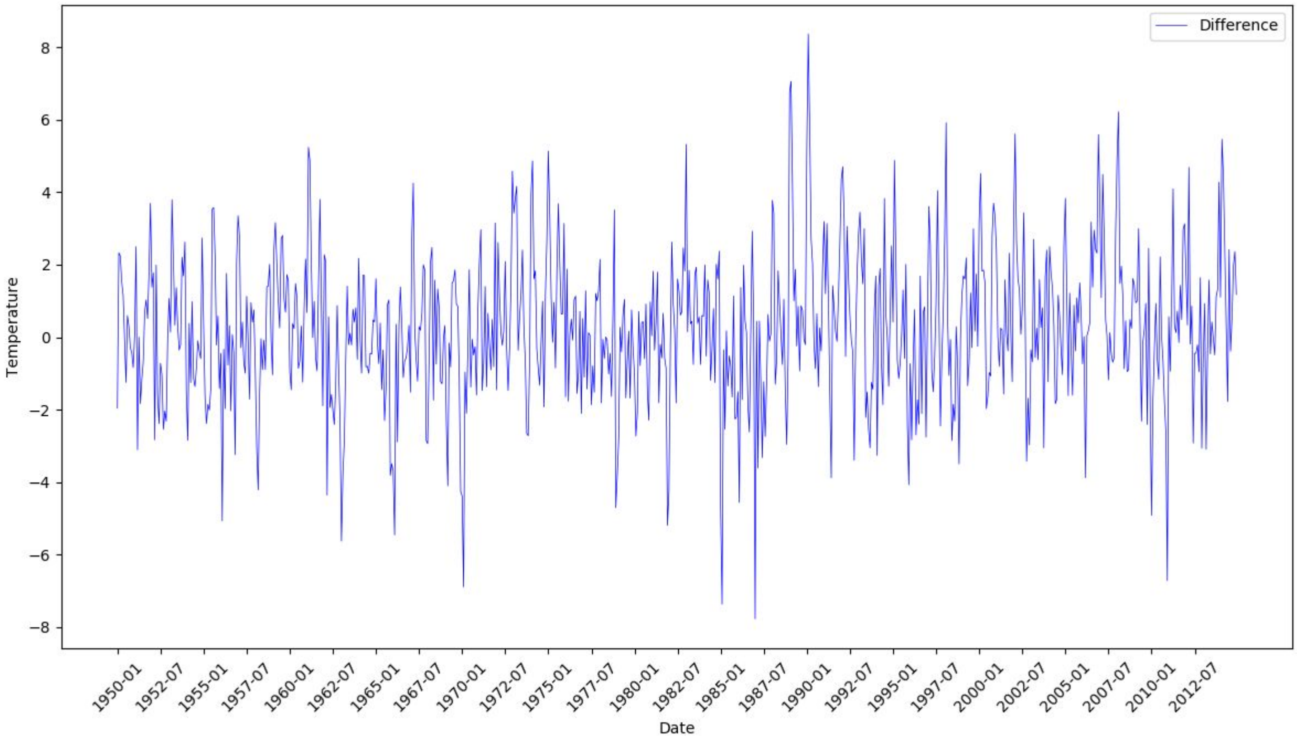
```
2014-09, 48.45000000000001
2009-05, 54.166666666666686
2009-08, 61.566666666666684
2016-04, 26.900000000000006
1998-05, 38.36666666666669
2002-02, 47.583333333333364
2016-02, 21.5625
1997-03, 9.549999999999999
1999-01, 61.93333333333339
2009-03, 34.48333333333334
2011-12, 42.133333333333375
2015-09, 101.29999999999998
2015-10, 2.2625
2006-12, 29.733333333333334
2008-11, 46.750000000000036
1994-06, 45.10000000000002
1999-05, 27.38333333333334
2004-01, 26.400000000000016
2004-09, 37.20000000000001
2006-08, 148.08333333333334
```

6)

Difference:

```
1950-01, -1.9604396182945512
1950-02, 2.3301122486854053
1950-03, 2.238629067922681
1950-04, 1.5325254517862152
1950-05, 1.151662889551627
1950-06, -0.028923429254035682
1950-07, -1.248976765983695
1950-08, 0.5964857468890745
1950-09, 0.2924376138185121
1950-10, -0.28796165209301083
1950-11, -0.4332336009459741
1950-12, -0.8393748128889749
1951-01, 0.048391026866735576
1951-02, 2.499089877320752
1951-03, -3.1077703051999532
1951-04, -0.000966126813521484
1951-05, -1.8327917385350272
1951-06, -1.1717586629139074
1951-07, -0.6652758568146844
1951-08, 0.6926069431800403
```

Long term average for each month:
01: -3.084520059124802
02: -3.7600898773207523
03: -0.5515845335097252
04: 4.408099548213785
05: 10.3026054031313
06: 15.628297124452367
07: 16.813538884109967
08: 15.950072957564368
09: 11.663187386181486
10: 7.114977781125273
11: 2.297608600945975
12: -1.147923574207799



In the graph we can see how the average temperature for each month and year differ from the long term monthly average. There is also a small increase in the difference each year from 1950 to 2012.