PROJECT-2 REPORT

BigData Analysis Project – Weather

SUBMITTED BY

SUSHMITA JADHAV

M12951946

TABLE OF CONTENTS

1. INTRODUCTION
2. ANALYSIS
   2.1. FINDING AVERAGE TMIN, TMAX FOR EACH YEAR
   2.2. FINDING MAXIMUM TMAX AND MINIMUM TMIN FOR EACH YEAR
   2.3. FIVE HOTTEST WEATHER STATIONS FOR EACH YEAR
   2.4. FIVE COLDEST WEATHER STATIONS FOR EACH YEAR
   2.5. HOTTEST DAY IN THE ENTIRE DATASET
   2.6. COLDEST DAY IN THE ENTIRE DATASET
   2.7. BONUS TASK 1 – MEDIAN TMIN, TMAX FOR EACH YEAR
   2.8. BONUS TASK 2 – MEDIAN TMIN, TMAX FOR ENTIRE DATASET
3. Why Spark, not other tools?
4. APPENDIX
   4.1. SPARK CODE
   4.2. OUTPUT OF THE PROGRAM

1. INTRODUCTION

This project performs Data Analysis and processing using Apache Spark. The Project will use the weather dataset from https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/by_year/.

This project will use only 19 years of data (2000 - 2019) for all the stations starting with US and elements TMAX, TMIN. The dataset is available on the CEAS Hadoop directory /user/tatavag/weather

The following information serves as a definition of each field in one line of data covering one station-day. Each field described below is separated by a comma (,) and follows the order presented in this document.

ID = 11-character station identification code

YEAR/MONTH/DAY = 8-character date in YYYYMMDD format (e.g. 19860529 = May 29, 1986)

ELEMENT = 4-character indicator of element type

DATA VALUE = 5-character data value for ELEMENT

M-FLAG = 1-character Measurement Flag

Q-FLAG = 1-character Quality Flag

S-FLAG = 1-character Source Flag

OBS-TIME = 4-character time of observation in hour-minute format (i.e. 0700 =7:00 am)

The assigned tasks to be performed are:

1. Average TMIN, TMAX for each year excluding abnormalities or missing data.
2. Maximum TMAX, Minimum TMIN for each year excluding abnormalities or missing data.
3. 5 hottest, 5 coldest weather stations for each year excluding abnormalities or missing data.
4. Hottest and coldest day and corresponding weather stations in the entire dataset.

Bonus tasks:
1. Median TMIN, TMAX for each year and corresponding weather stations.
2. Median TMIN, TMAX for the entire dataset

The analysis is mainly done using Apache Spark(Spark Data frames) and Apache Spark(SparkSQL). The PySpark program outputs the analysis results for each year starting from 2000 to 2019 with run() and for entire dataset with runFullDataset() commands on PySpark Prompt.

I have also tried hands on Impala cluster on AWS to get plot of the analysis

NOTE: ALL THE TEMPERATURES ARE REPRESENTED IN TENTHS OF DEGREE CELSIUS

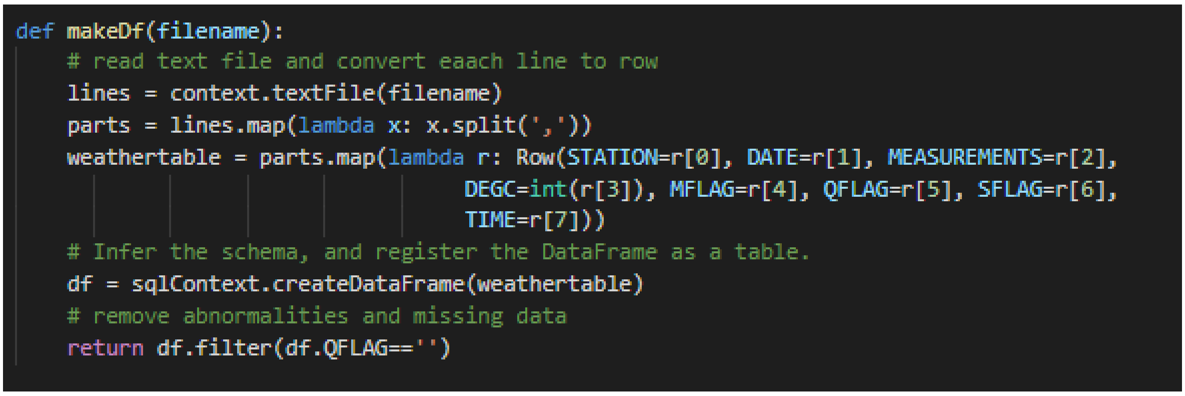
Tenths of degree Celsius value is equal to degree Celsius value multiplied by 10. It should not be misunderstood with Celsius.

225.0 tenths of degree Celsius = 22.5 degree Celsius

2. ANALYSIS

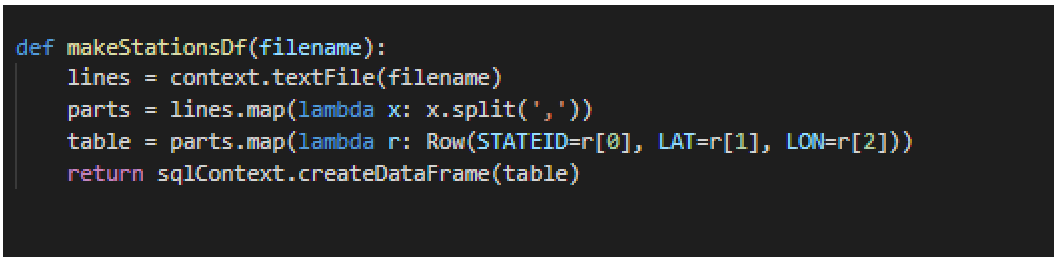
For the analysis the input for each year is a csv file located at “hdfs:/user/tatavag/weather/20**.csv”

The following snippet is used to create a dataframe to carry out the analysis



```
def makeDf(filename):
    # read text file and convert eaach line to row
    lines = context.textFile(filename)
    parts = lines.map(lambda x: x.split(','))
    weathertable = parts.map(lambda r: Row(STATION=r[0], DATE=r[1], MEASUREMENTS=r[2],
                                            DEGC=int(r[3]), MFLAG=r[4], QFLAG=r[5], SFLAG=r[6],
                                            TIME=r[7]))
    # Infer the schema, and register the DataFrame as a table.
    df = sqlContext.createDataFrame(weathertable)
    # remove abnormalities and missing data
    return df.filter(df.QFLAG=='')
```

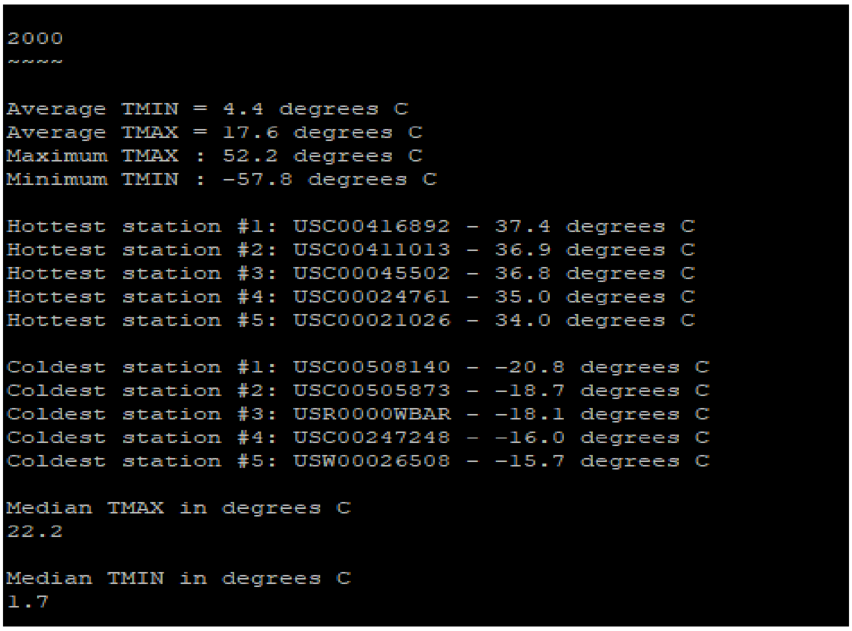
Snippet used to create a dataframe with information about stations is given below:



```
def makeStationsDf(filename):
    lines = context.textFile(filename)
    parts = lines.map(lambda x: x.split(','))
    table = parts.map(lambda r: Row(STATEID=r[0], LAT=r[1], LON=r[2]))
    return sqlContext.createDataFrame(table)
```

Output is displayed for each year on the console which answers the analysis queries for the corresponding year when run() is called and for the entire dataset when runFullDataset() is called.

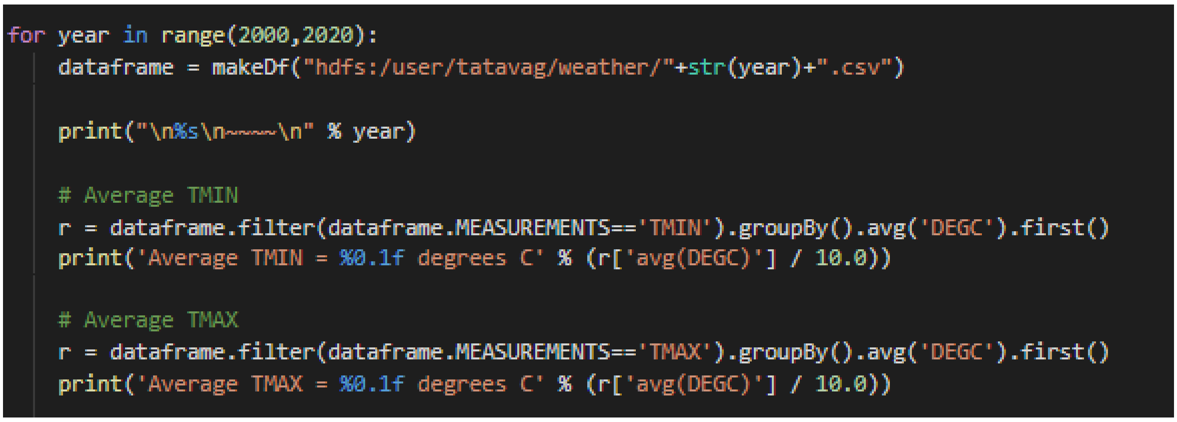
Output for each year looks as shown below:



```
2000
~~~~

Average TMIN = 4.4 degrees C
Average TMAX = 17.6 degrees C
Maximum TMAX : 52.2 degrees C
Minimum TMIN : -57.8 degrees C

Hottest station #1: USC00416892 - 37.4 degrees C
Hottest station #2: USC00411013 - 36.9 degrees C
Hottest station #3: USC00045502 - 36.8 degrees C
Hottest station #4: USC00024761 - 35.0 degrees C
Hottest station #5: USC00021026 - 34.0 degrees C

Coldest station #1: USC00508140 - -20.8 degrees C
Coldest station #2: USC00505873 - -18.7 degrees C
Coldest station #3: USR0000WBAR - -18.1 degrees C
Coldest station #4: USC00247248 - -16.0 degrees C
Coldest station #5: USW00026508 - -15.7 degrees C

Median TMAX in degrees C
22.2

Median TMIN in degrees C
1.7
```

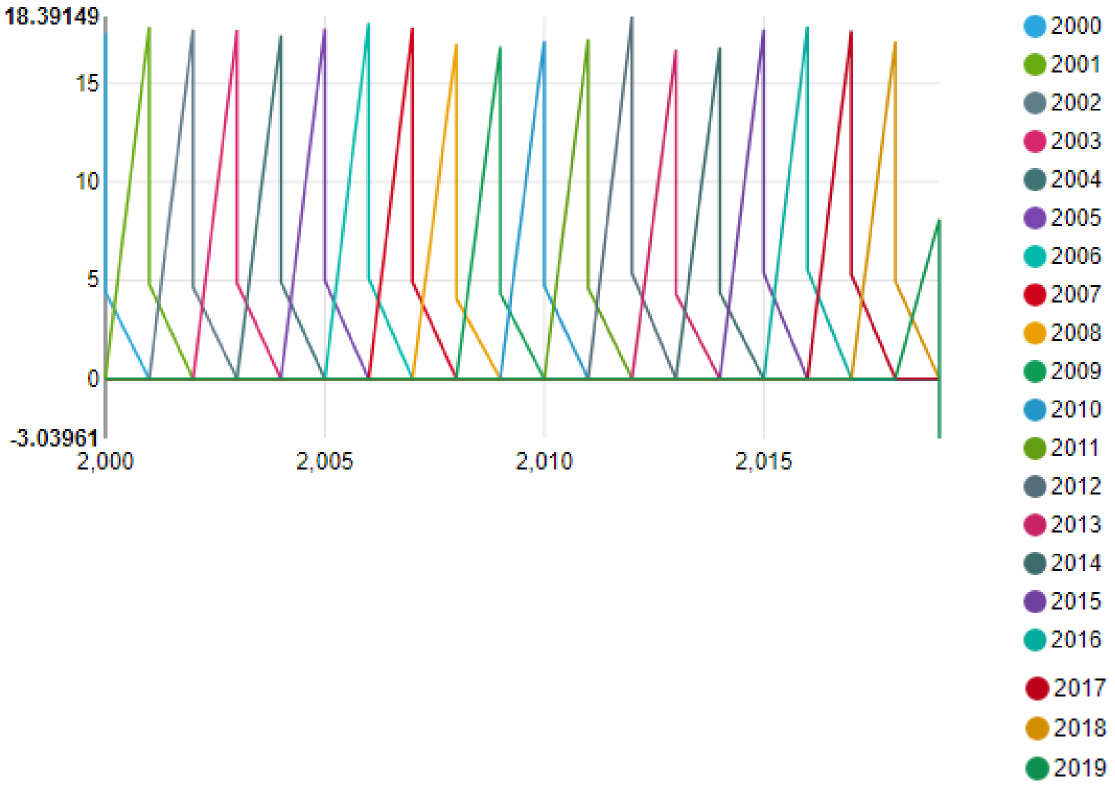
2.1 FINDING AVERAGE TMIN, TMAX FOR EACH YEAR

Average TMIN and Average TMAX for each year are obtained as follows:



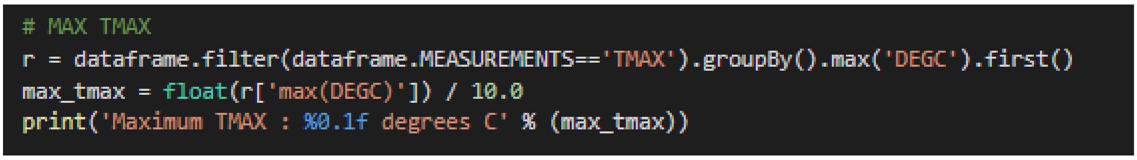
```
for year in range(2000,2020):
    dataframe = makeDf("hdfs:/user/tatavag/weather/"+str(year)+".csv")

    print("\n%s\n~~~~~\n" % year)

    # Average TMIN
    r = dataframe.filter(dataframe.MEASUREMENTS=='TMIN').groupBy().avg('DEGC').first()
    print('Average TMIN = %0.1f degrees C' % (r['avg(DEGC)'] / 10.0))

    # Average TMAX
    r = dataframe.filter(dataframe.MEASUREMENTS=='TMAX').groupBy().avg('DEGC').first()
    print('Average TMAX = %0.1f degrees C' % (r['avg(DEGC)'] / 10.0))
```

Output Plot:



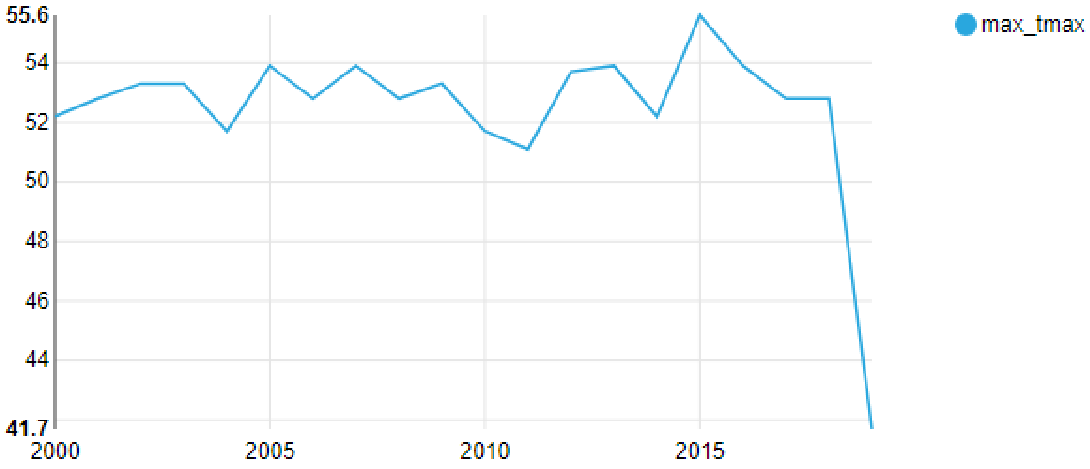
2.2 FINDING MAXIMUM TMAX AND MINIMUM TMIN FOR EACH YEAR

Max-TMAX

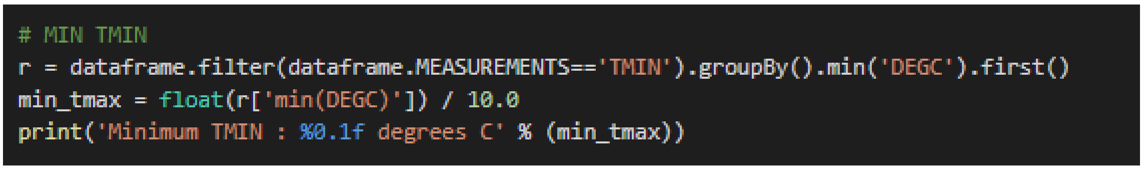


```
# MAX TMAX
r = dataframe.filter(dataframe.MEASUREMENTS=='TMAX').groupBy().max('DEGC').first()
max_tmax = float(r['max(DEGC)']) / 10.0
print('Maximum TMAX : %0.1f degrees C' % (max_tmax))
```

Output Plot:

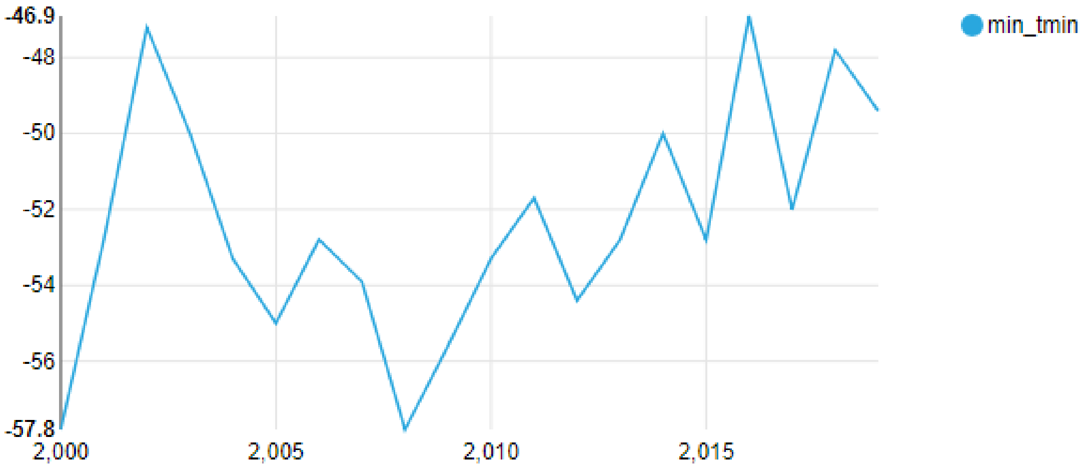


MIN-TMIN:



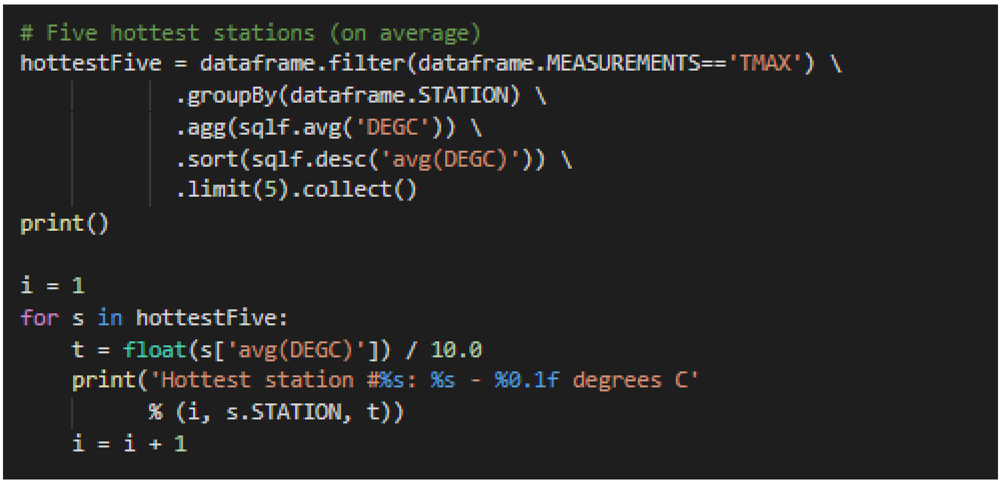
```
# MIN TMIN
r = dataframe.filter(dataframe.MEASUREMENTS=='TMIN').groupBy().min('DEGC').first()
min_tmax = float(r['min(DEGC)']) / 10.0
print('Minimum TMIN : %0.1f degrees C' % (min_tmax))
```

Output Plot:



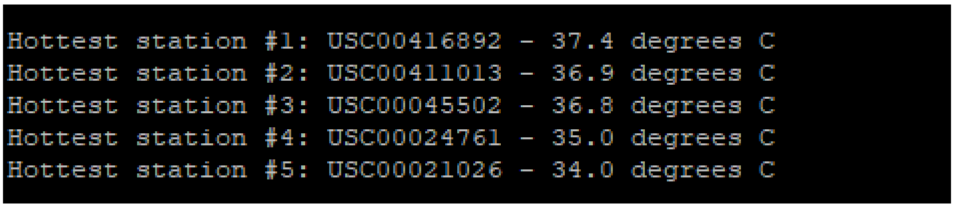
2.3 FIVE HOTTEST WEATHER STATIONS FOR EACH YEAR

Five hottest stations are obtained as shown below: Records are filtered where measurements column data is ‘TMAX’. The records are grouped on station name and average of temperature for each station is found. Five stations with highest average temperatures is displayed.



```python
# Five hottest stations (on average)
hottestFive = dataframe.filter(dataframe.MEASUREMENTS=='TMAX') \
            .groupBy(dataframe.STATION) \
            .agg(sqlf.avg('DEGC')) \
            .sort(sqlf.desc('avg(DEGC)')) \
            .limit(5).collect()
print()

i = 1
for s in hottestFive:
    t = float(s['avg(DEGC)']) / 10.0
    print('Hottest station #%s: %s - %0.1f degrees C'
          % (i, s.STATION, t))
    i = i + 1
```

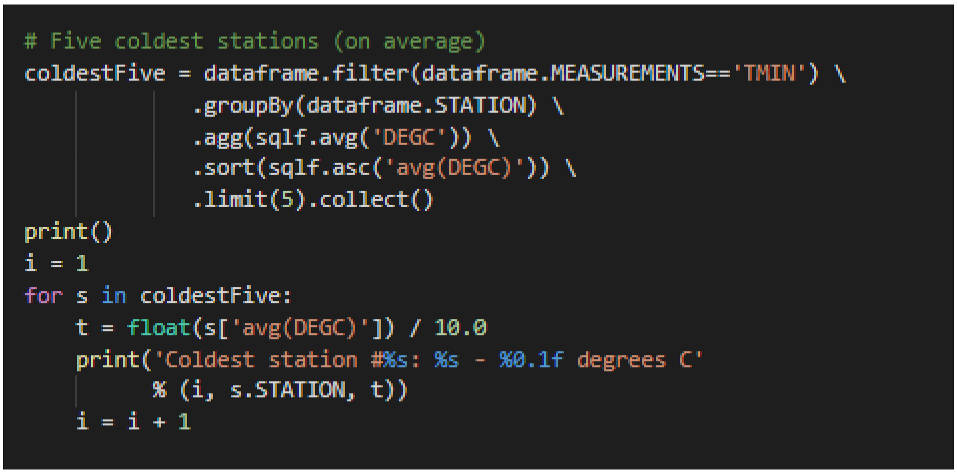
Output: Output for each year is shown on the console as below:



```
Hottest station #1: USC00416892 - 37.4 degrees C
Hottest station #2: USC00411013 - 36.9 degrees C
Hottest station #3: USC00045502 - 36.8 degrees C
Hottest station #4: USC00024761 - 35.0 degrees C
Hottest station #5: USC00021026 - 34.0 degrees C
```

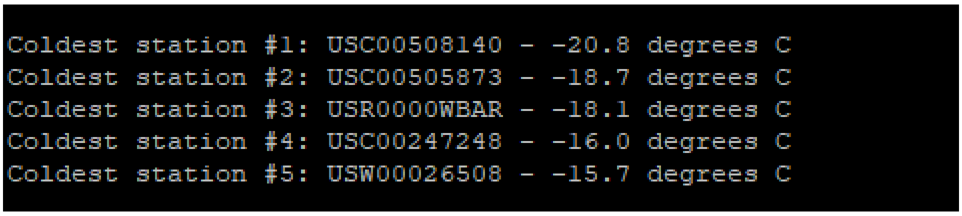
2.4 FIVE COLDEST WEATHER STATIONS FOR EACH YEAR

Same as above, five records with lowest average temperature is diplayed.



```python
# Five coldest stations (on average)
coldestFive = dataframe.filter(dataframe.MEASUREMENTS=='TMIN') \
            .groupBy(dataframe.STATION) \
            .agg(sqlf.avg('DEGC')) \
            .sort(sqlf.asc('avg(DEGC)')) \
            .limit(5).collect()
print()
i = 1
for s in coldestFive:
    t = float(s['avg(DEGC)']) / 10.0
    print('Coldest station #%s: %s - %0.1f degrees C'
          % (i, s.STATION, t))
    i = i + 1
```

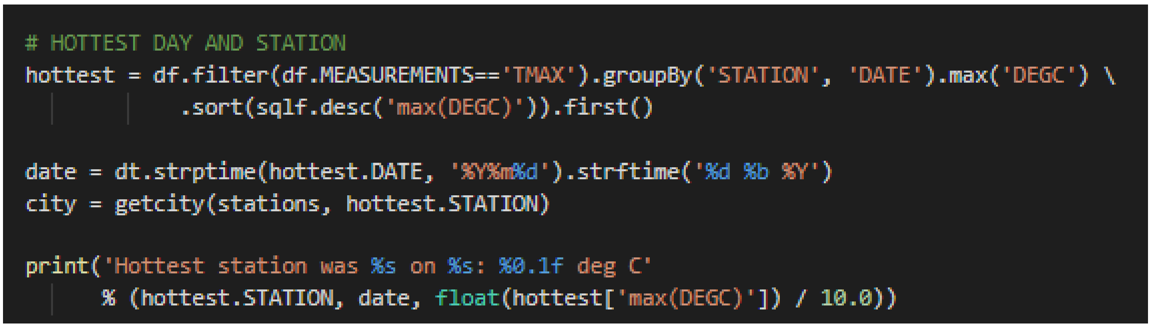
Output: Output for each year is shown on the console as below:



```
Coldest station #1: USC00508140 - -20.8 degrees C
Coldest station #2: USC00505873 - -18.7 degrees C
Coldest station #3: USR0000WBAR - -18.1 degrees C
Coldest station #4: USC00247248 - -16.0 degrees C
Coldest station #5: USW00026508 - -15.7 degrees C
```

2.5 HOTTEST DAY IN THE ENTIRE DATASET

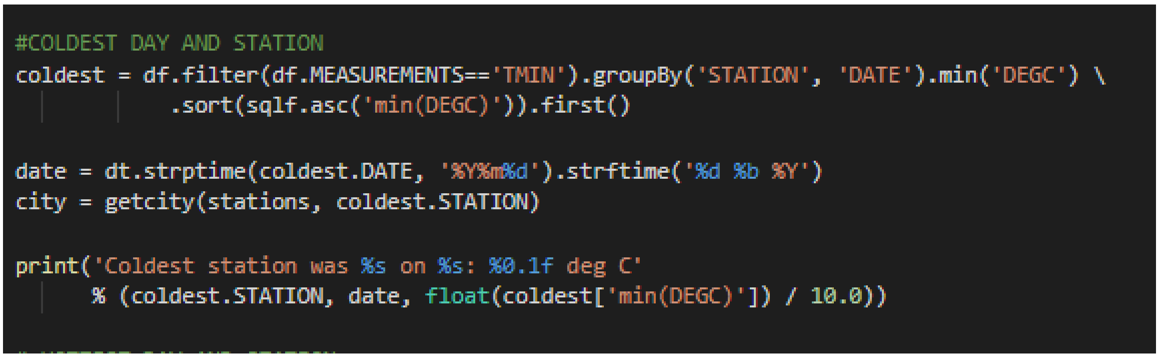
For the entire dataset, a dataframe is created with all the csv files of years in range 2000-2019 and hottest day is obtained as shown:



```
# HOTTEST DAY AND STATION
hottest = df.filter(df.MEASUREMENTS=='TMAX').groupBy('STATION', 'DATE').max('DEGC') \
            .sort(sqlf.desc('max(DEGC)')).first()

date = dt.strptime(hottest.DATE, '%Y%m%d').strftime('%d %b %Y')
city = getcity(stations, hottest.STATION)

print('Hottest station was %s on %s: %0.1f deg C'
      % (hottest.STATION, date, float(hottest['max(DEGC)']) / 10.0))
```

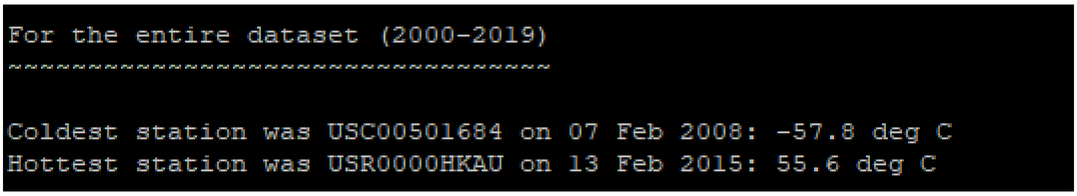
2.6 COLDEST DAY IN THE ENTIRE DATASET

For the entire dataset, a dataframe is created with all the csv files of years in range 2000-2019 and coldest day is obtained as shown:



```
#COLDEST DAY AND STATION
coldest = df.filter(df.MEASUREMENTS=='TMIN').groupBy('STATION', 'DATE').min('DEGC') \
            .sort(sqlf.asc('min(DEGC)')).first()

date = dt.strptime(coldest.DATE, '%Y%m%d').strftime('%d %b %Y')
city = getcity(stations, coldest.STATION)

print('Coldest station was %s on %s: %0.1f deg C'
      % (coldest.STATION, date, float(coldest['min(DEGC)']) / 10.0))
```

Output of 2.5 & 2.6:

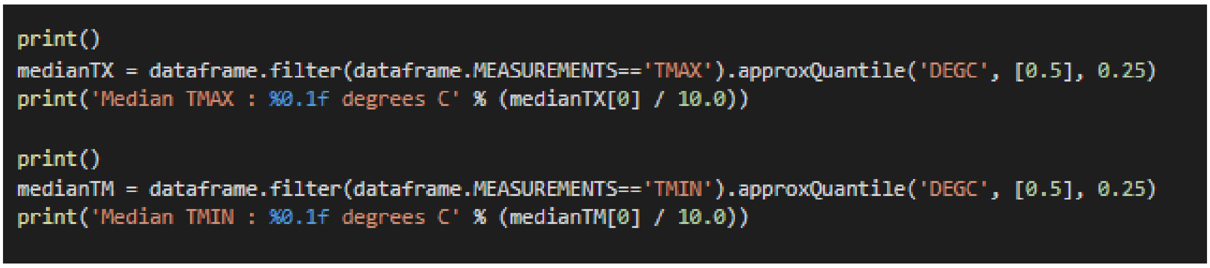


```
For the entire dataset (2000-2019)
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Coldest station was USC00501684 on 07 Feb 2008: -57.8 deg C
Hottest station was USR0000HKAU on 13 Feb 2015: 55.6 deg C
```

2.7 BONUS TASK 1 – MEDIAN TMIN, TMAX FOR EACH YEAR

To obtain median Tmin and Tmax for each year, approxQuantile method from PySpark SQL functions has been used. This method takes three arguments, first one gives the column name of the dataframe, second one is a list of quantile probabilities (0.5 is for median), and third one is for relative error

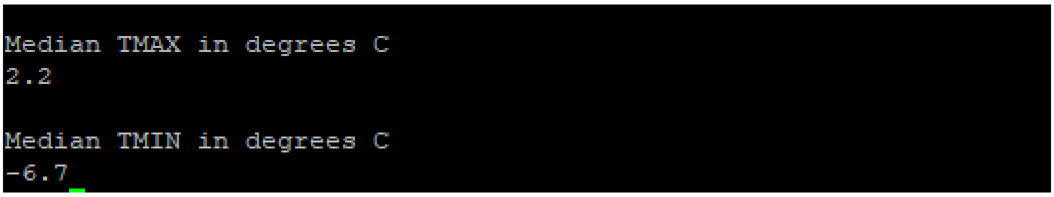
Snippet used to obtain the median Tmin and Tmax for each year is:



```
print()
medianTX = dataframe.filter(dataframe.MEASUREMENTS=='TMAX').approxQuantile('DEGC', [0.5], 0.25)
print('Median TMAX : %0.1f degrees C' % (medianTX[0] / 10.0))

print()
medianTM = dataframe.filter(dataframe.MEASUREMENTS=='TMIN').approxQuantile('DEGC', [0.5], 0.25)
print('Median TMIN : %0.1f degrees C' % (medianTM[0] / 10.0))
```

Output:

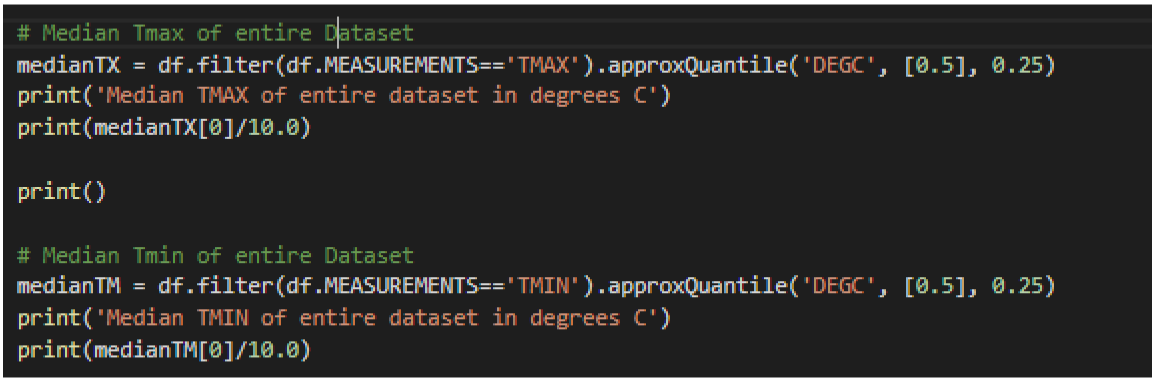
Output is displayed for each year is in the format shown below:



```
Median TMAX in degrees C
2.2

Median TMIN in degrees C
-6.7
```

2.8 BONUS TASK 2 – MEDIAN TMIN, TMAX FOR ENTIRE DATASET

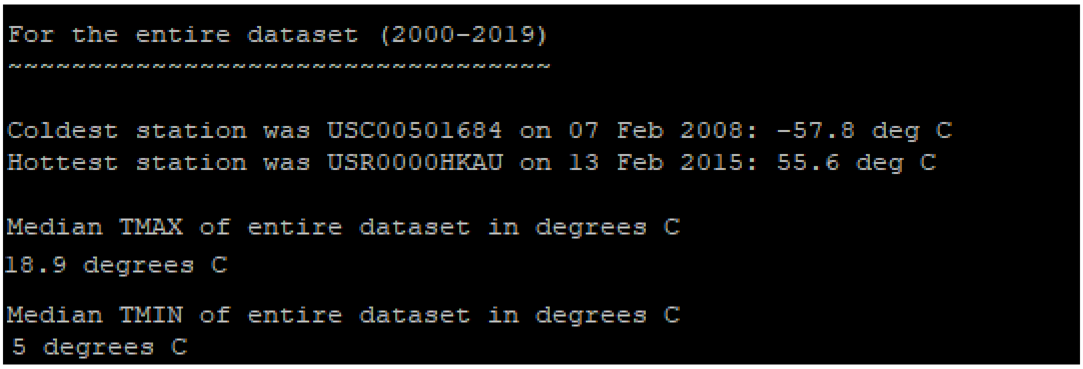
To obtain median Tmin and Tmax, the same approxQuantile method is used but on the dataframe that is created by merging the csv files of all years in range 2000-2019



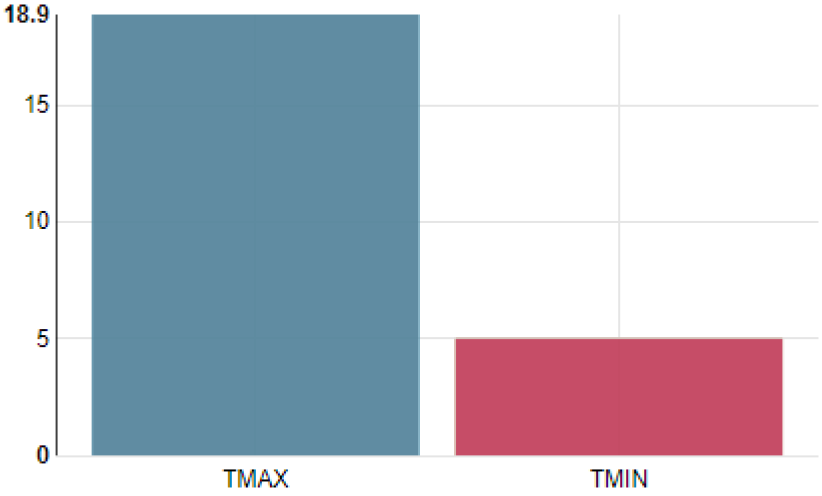
```
# Median Tmax of entire Dataset
medianTX = df.filter(df.MEASUREMENTS=='TMAX').approxQuantile('DEGC', [0.5], 0.25)
print('Median TMAX of entire dataset in degrees C')
print(medianTX[0]/10.0)

print()

# Median Tmin of entire Dataset
medianTM = df.filter(df.MEASUREMENTS=='TMIN').approxQuantile('DEGC', [0.5], 0.25)
print('Median TMIN of entire dataset in degrees C')
print(medianTM[0]/10.0)
```

Output:

Output is displayed for entire dataset is in the format shown below:



```
For the entire dataset (2000-2019)
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Coldest station was USC00501684 on 07 Feb 2008: -57.8 deg C
Hottest station was USR0000HKAU on 13 Feb 2015: 55.6 deg C

Median TMAX of entire dataset in degrees C
18.9 degrees C
Median TMIN of entire dataset in degrees C
 5 degrees C
```



3. Why Spark, not other tools?

What is Spark?

Spark is defined as a distributed data processing engine for general-purpose that is useful in wide range of circumstances. On top of the Spark core data processing engine, there are libraries for SQL, machine learning, graph computation, and stream processing, which can be used together in an application. Spark supports programming languages which include: Java, Python, R and Scala. Spark is mainly used by application developers and data scientists to rapidly query, perform analysis, and transform data as required at scale.

To get the results of the analysis I have created Spark Dataframes with required columns and performed queries on the dataframe using Spark SQL functions. I have chosen Spark because of the following key reasons:

It is Simple: There is a set of APIs that are made available by Spark which can be used for interacting with data quickly and easily at scale. Proper documentation of these APIs is provided which helps application developers and data scientists to quickly put Spark to work.

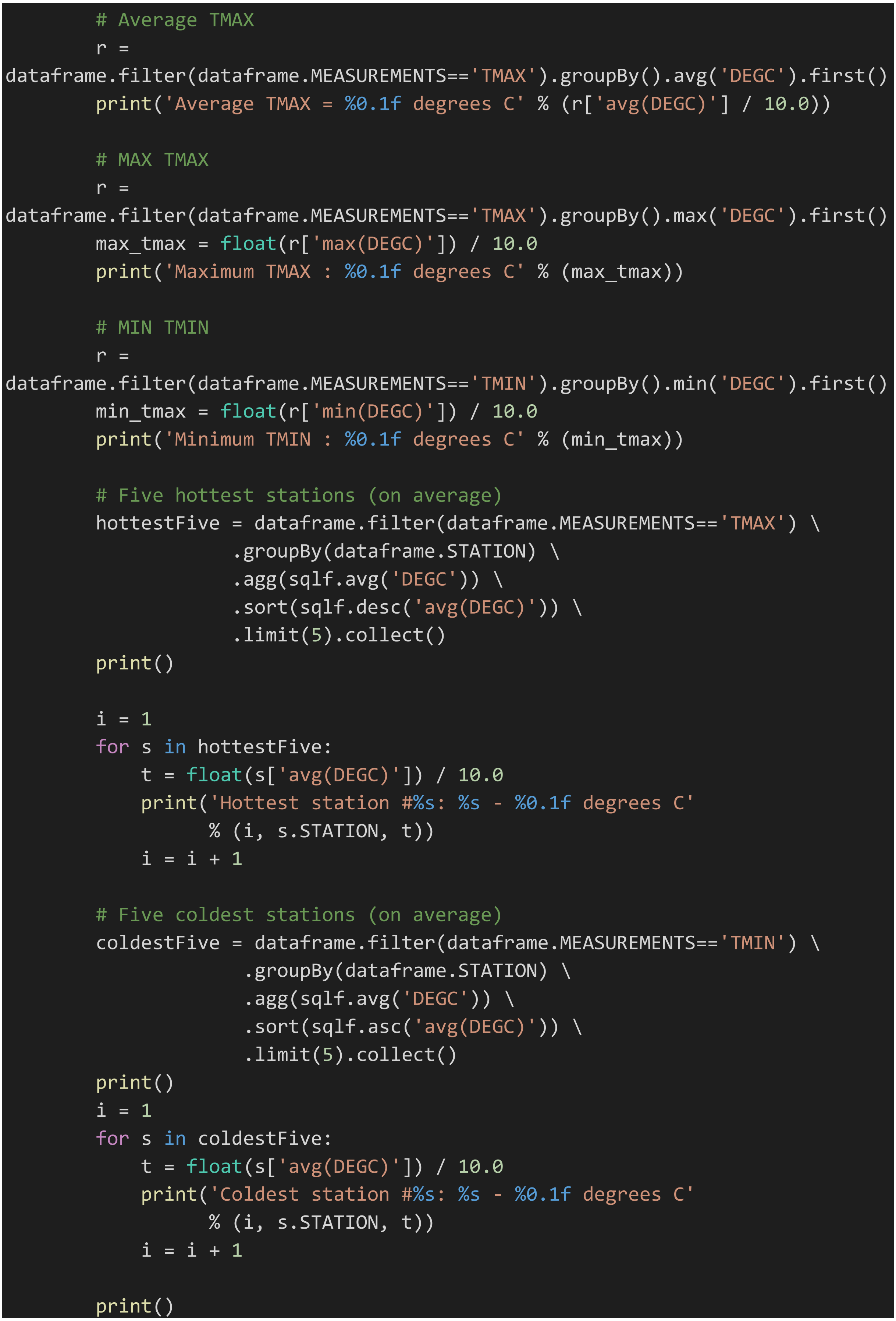
It is Fast: Spark is fast in operating in both memory and on disk. There are claims that Spark can be much faster, almost 100 times faster than Hadoop's MapReduce. Spark SQL contains a library of functions that lets users use any constructs while writing Spark pipelines.

Its Support: Spark supports a range of programming languages, including Java, Python, R, and Scala. It provides support for tight integration with many leading storage solutions of Hadoop ecosystem. It is also growing providers in commercial market like Databricks, IBM, and delivers very comprehensive support for many Spark-based solutions.

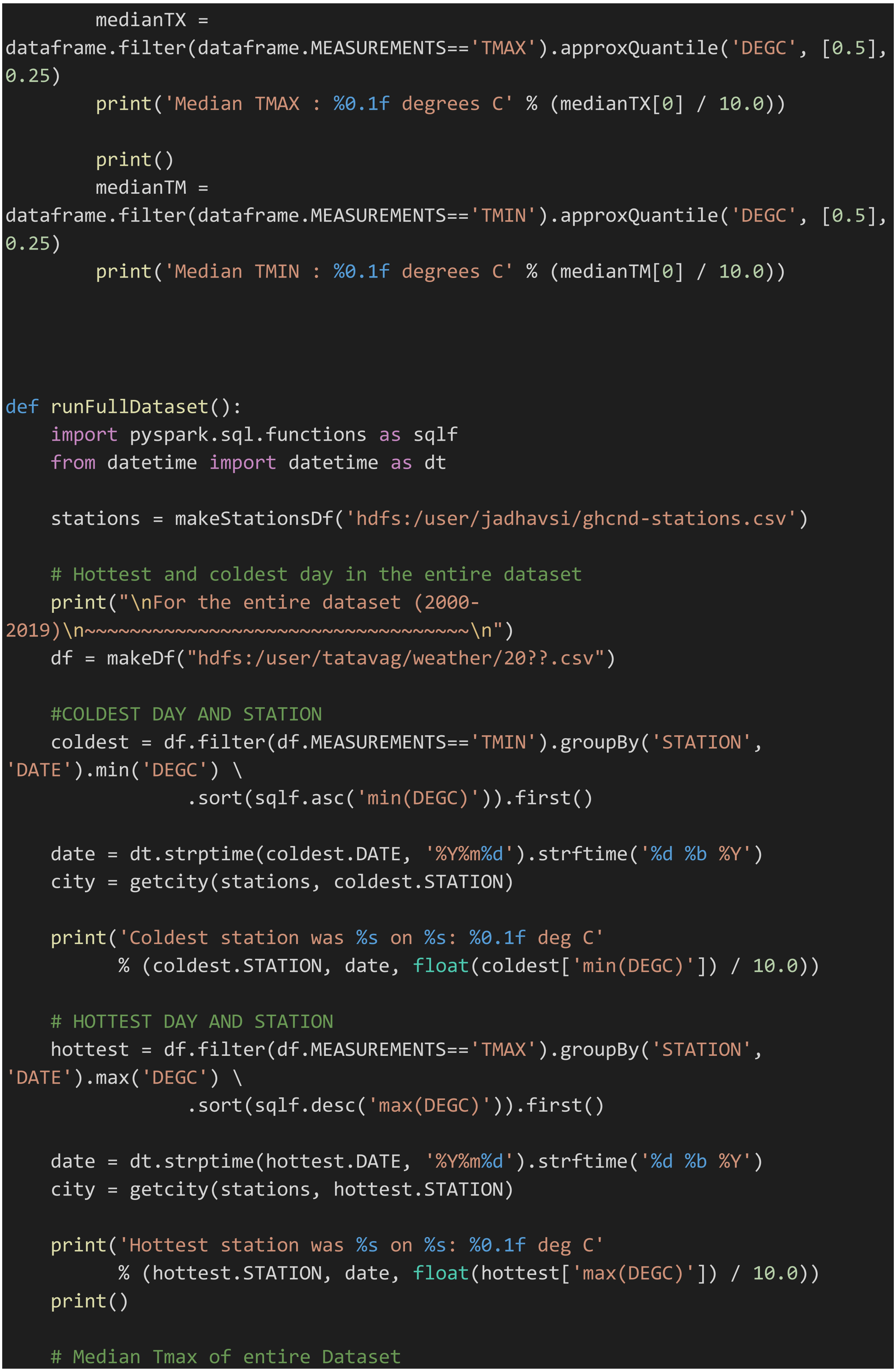
It is developer friendly: Spark SQL can be seen to be a developer-friendly Spark based API which is aimed to make the programming easier. Developers enjoy the concise and fluid way it can be programmed.

SPARK CODE:

```
#!/usr/bin/env python
from __future__ import print_function
from pyspark import SparkContext
from pyspark.sql import SQLContext, Row

context = SparkContext.getOrCreate()
sqlContext = SQLContext(context)

def makeDf(filename):
    # read text file and convert eaach line to row
    lines = context.textFile(filename)
    parts = lines.map(lambda x: x.split(','))
    weathertable = parts.map(lambda r: Row(STATION=r[0], DATE=r[1],
MEASUREMENTS=r[2],
                                   DEGC=int(r[3]), MFLAG=r[4], QFLAG=r[5],
SFLAG=r[6],
                                   TIME=r[7]))
    # Infer the schema, and register the DataFrame as a table.
    df = sqlContext.createDataFrame(weathertable)
    # remove abnormalities and missing data
    return df.filter(df.QFLAG=='')


def makeStationsDf(filename):
    lines = context.textFile(filename)
    parts = lines.map(lambda x: x.split(','))
    table = parts.map(lambda r: Row(STATEID=r[0], LAT=r[1], LON=r[2]))
    return sqlContext.createDataFrame(table)


def run():
    import pyspark.sql.functions as sqlf

    stations = makeStationsDf('hdfs:/user/jadhavsi/ghcnd-stations.csv')

    for year in range(2000,2020):
        dataframe = makeDf("hdfs:/user/tatavag/weather/"+str(year)+".csv")

        print("\n%s\n~~~~\n" % year)

        # Average TMIN
        r =
dataframe.filter(dataframe.MEASUREMENTS=='TMIN').groupBy().avg('DEGC').first()
        print('Average TMIN = %0.1f degrees C' % (r['avg(DEGC)'] / 10.0))
```



```
        # Average TMAX
        r = 
dataframe.filter(dataframe.MEASUREMENTS=='TMAX').groupBy().avg('DEGC').first()
        print('Average TMAX = %0.1f degrees C' % (r['avg(DEGC)'] / 10.0))

        # MAX TMAX
        r = 
dataframe.filter(dataframe.MEASUREMENTS=='TMAX').groupBy().max('DEGC').first()
        max_tmax = float(r['max(DEGC)']) / 10.0
        print('Maximum TMAX : %0.1f degrees C' % (max_tmax))

        # MIN TMIN
        r = 
dataframe.filter(dataframe.MEASUREMENTS=='TMIN').groupBy().min('DEGC').first()
        min_tmax = float(r['min(DEGC)']) / 10.0
        print('Minimum TMIN : %0.1f degrees C' % (min_tmax))

        # Five hottest stations (on average)
        hottestFive = dataframe.filter(dataframe.MEASUREMENTS=='TMAX') \
                    .groupBy(dataframe.STATION) \
                    .agg(sqlf.avg('DEGC')) \
                    .sort(sqlf.desc('avg(DEGC)')) \
                    .limit(5).collect()
        print()

        i = 1
        for s in hottestFive:
            t = float(s['avg(DEGC)']) / 10.0
            print('Hottest station #%s: %s - %0.1f degrees C'
                  % (i, s.STATION, t))
            i = i + 1

        # Five coldest stations (on average)
        coldestFive = dataframe.filter(dataframe.MEASUREMENTS=='TMIN') \
                     .groupBy(dataframe.STATION) \
                     .agg(sqlf.avg('DEGC')) \
                     .sort(sqlf.asc('avg(DEGC)')) \
                     .limit(5).collect()
        print()
        i = 1
        for s in coldestFive:
            t = float(s['avg(DEGC)']) / 10.0
            print('Coldest station #%s: %s - %0.1f degrees C'
                  % (i, s.STATION, t))
            i = i + 1

        print()
```



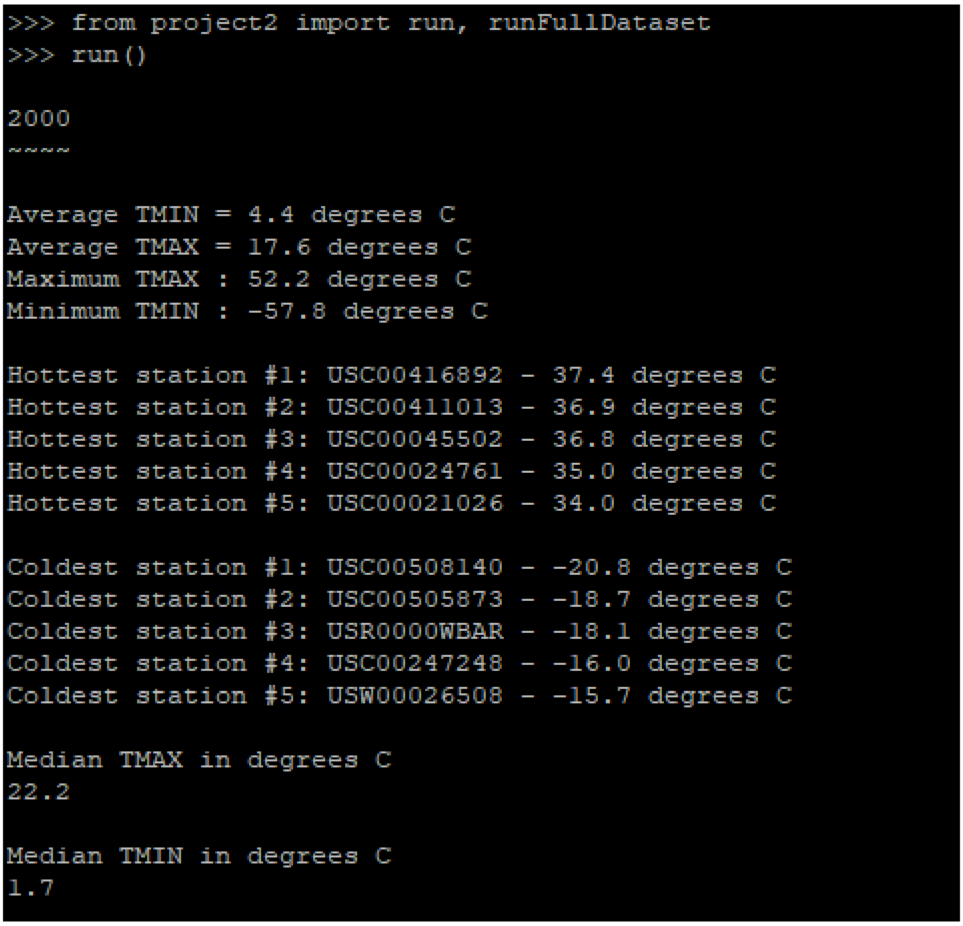
```
        medianTX =
dataframe.filter(dataframe.MEASUREMENTS=='TMAX').approxQuantile('DEGC', [0.5],
0.25)
        print('Median TMAX : %0.1f degrees C' % (medianTX[0] / 10.0))

        print()
        medianTM =
dataframe.filter(dataframe.MEASUREMENTS=='TMIN').approxQuantile('DEGC', [0.5],
0.25)
        print('Median TMIN : %0.1f degrees C' % (medianTM[0] / 10.0))




def runFullDataset():
    import pyspark.sql.functions as sqlf
    from datetime import datetime as dt

    stations = makeStationsDf('hdfs:/user/jadhavsi/ghcnd-stations.csv')

    # Hottest and coldest day in the entire dataset
    print("\nFor the entire dataset (2000-
2019)\n~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~\n")
    df = makeDf("hdfs:/user/tatavag/weather/20??.csv")

    #COLDEST DAY AND STATION
    coldest = df.filter(df.MEASUREMENTS=='TMIN').groupBy('STATION',
'DATE').min('DEGC') \
                .sort(sqlf.asc('min(DEGC)')).first()

    date = dt.strptime(coldest.DATE, '%Y%m%d').strftime('%d %b %Y')
    city = getcity(stations, coldest.STATION)

    print('Coldest station was %s on %s: %0.1f deg C'
          % (coldest.STATION, date, float(coldest['min(DEGC)']) / 10.0))

    # HOTTEST DAY AND STATION
    hottest = df.filter(df.MEASUREMENTS=='TMAX').groupBy('STATION',
'DATE').max('DEGC') \
                .sort(sqlf.desc('max(DEGC)')).first()

    date = dt.strptime(hottest.DATE, '%Y%m%d').strftime('%d %b %Y')
    city = getcity(stations, hottest.STATION)

    print('Hottest station was %s on %s: %0.1f deg C'
          % (hottest.STATION, date, float(hottest['max(DEGC)']) / 10.0))
    print()

    # Median Tmax of entire Dataset
```

```
    medianTX = df.filter(df.MEASUREMENTS=='TMAX').approxQuantile('DEGC',
[0.5], 0.25)
    print('Median TMAX of entire dataset in degrees C')
    print(medianTX[0]/10.0)

    print()

    # Median Tmin of entire Dataset
    medianTM = df.filter(df.MEASUREMENTS=='TMIN').approxQuantile('DEGC',
[0.5], 0.25)
    print('Median TMIN of entire dataset in degrees C')
    print(medianTM[0]/10.0)

if __name__ == '__main__':
    run()
```

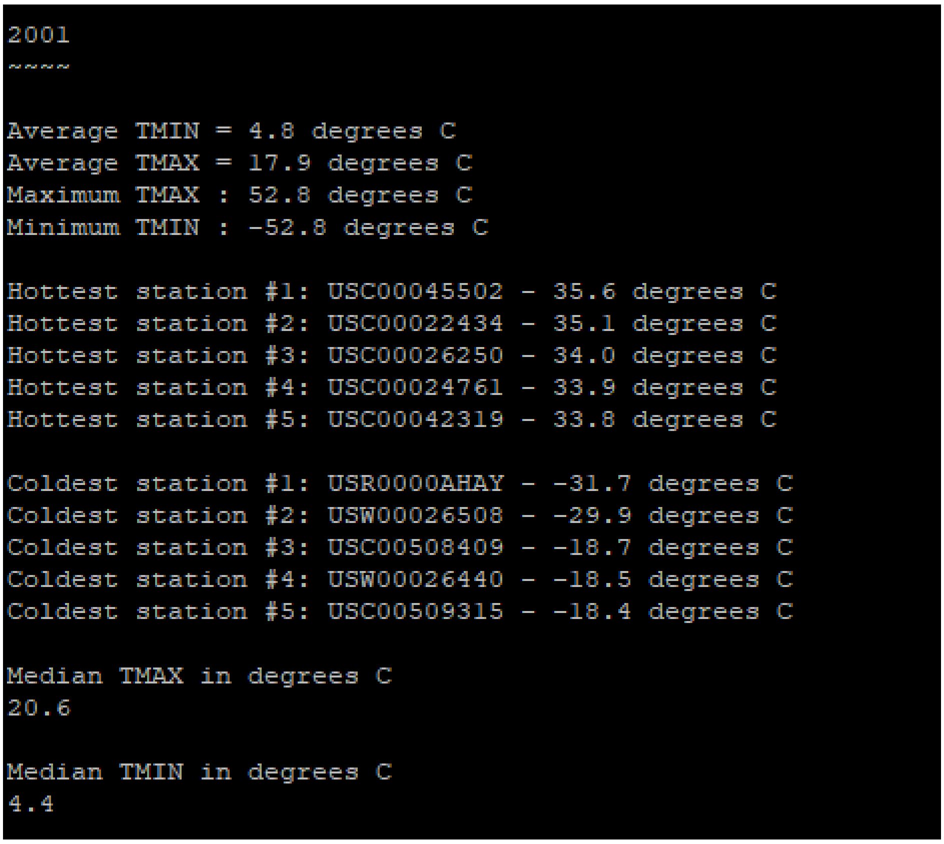
OUTPUT:

To run the above program give the following commands on PySpark prompt

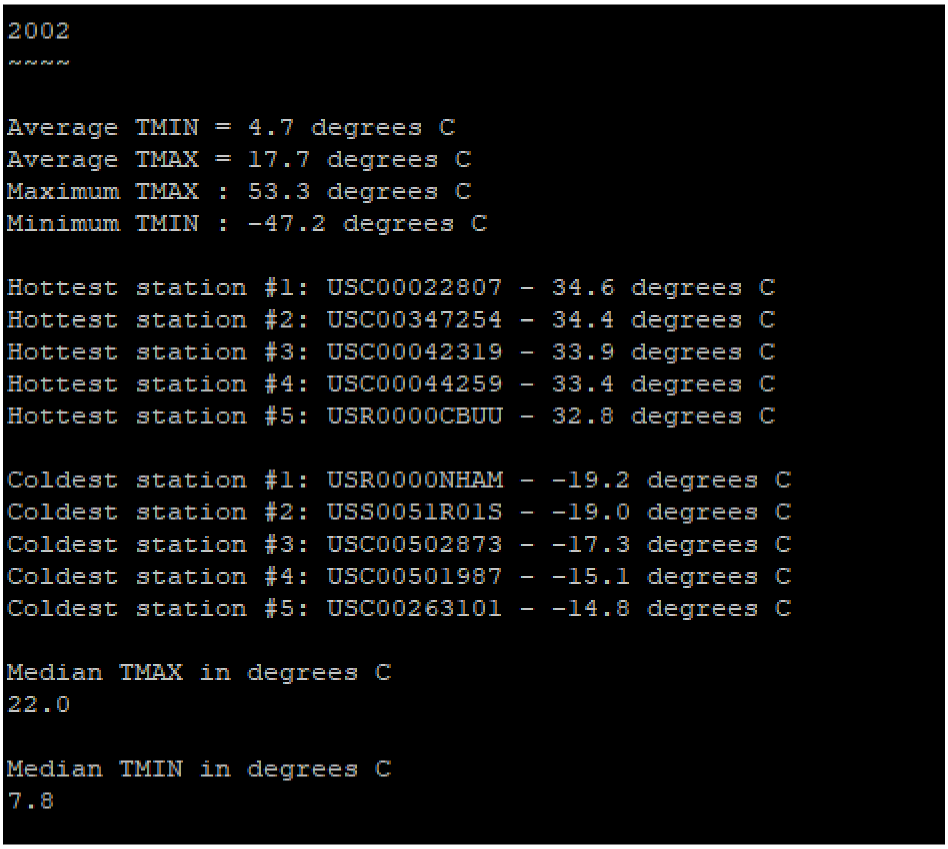
>>> from project2 import run, runFullDataset
>>> run ()



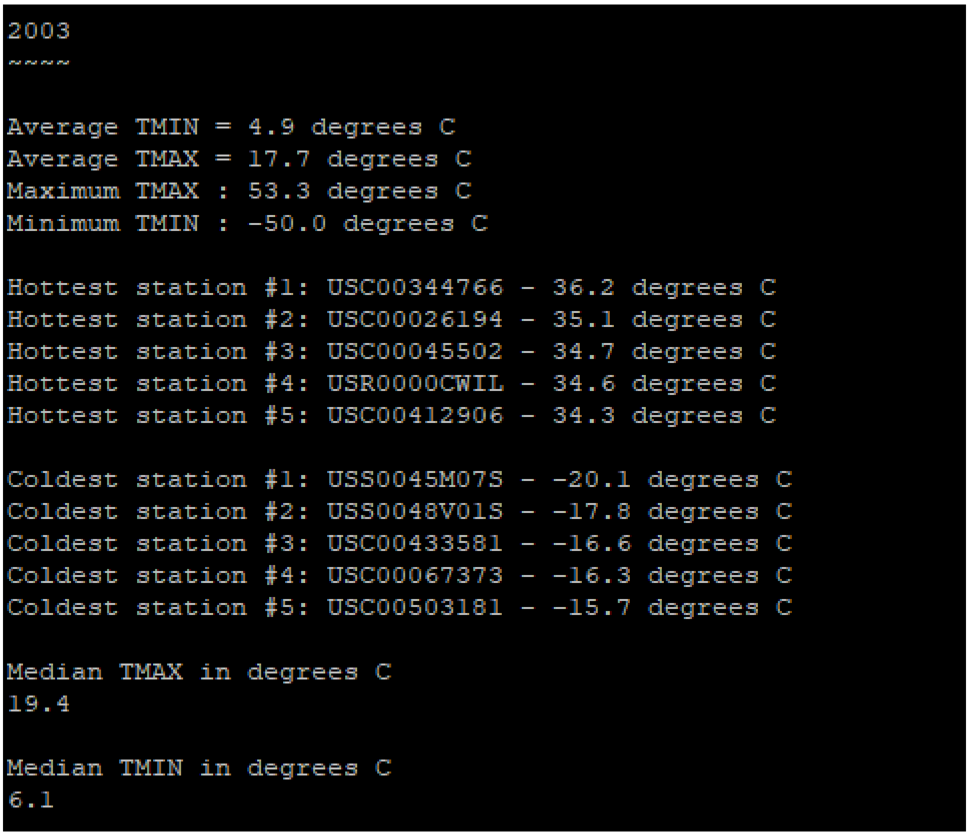
```
>>> from project2 import run, runFullDataset
>>> run()

2000
~~~~

Average TMIN = 4.4 degrees C
Average TMAX = 17.6 degrees C
Maximum TMAX : 52.2 degrees C
Minimum TMIN : -57.8 degrees C

Hottest station #1: USC00416892 - 37.4 degrees C
Hottest station #2: USC00411013 - 36.9 degrees C
Hottest station #3: USC00045502 - 36.8 degrees C
Hottest station #4: USC00024761 - 35.0 degrees C
Hottest station #5: USC00021026 - 34.0 degrees C

Coldest station #1: USC00508140 - -20.8 degrees C
Coldest station #2: USC00505873 - -18.7 degrees C
Coldest station #3: USR0000WBAR - -18.1 degrees C
Coldest station #4: USC00247248 - -16.0 degrees C
Coldest station #5: USW00026508 - -15.7 degrees C

Median TMAX in degrees C
22.2

Median TMIN in degrees C
1.7
```



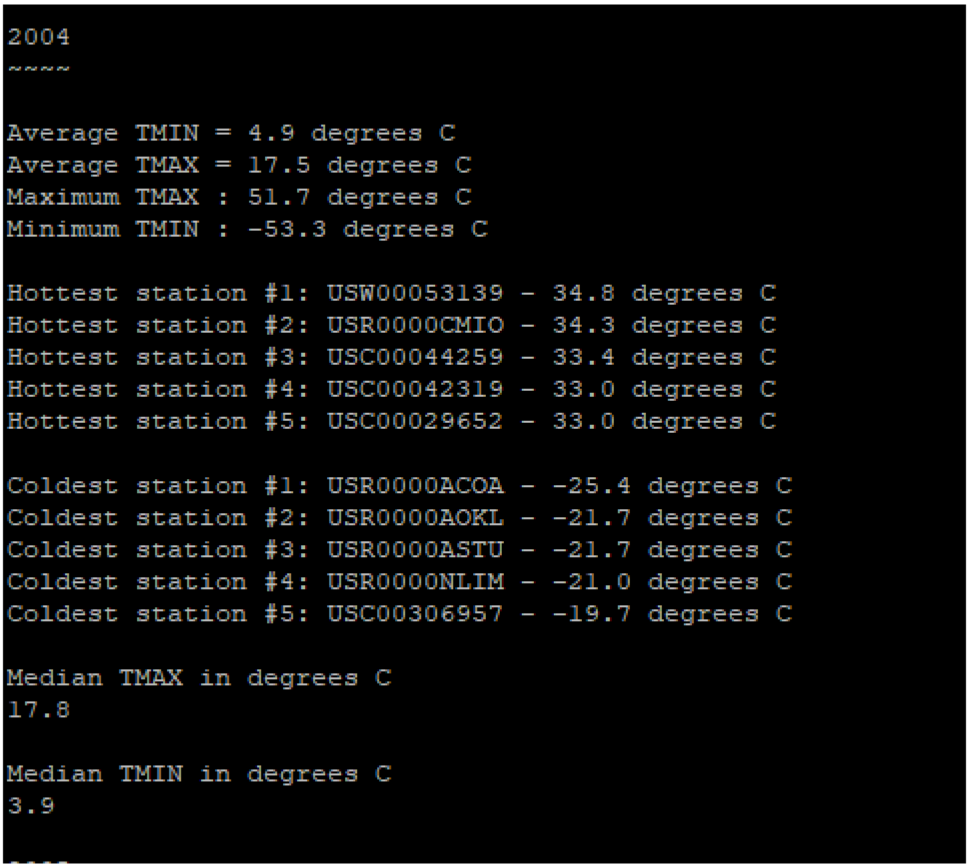
```
2001
~~~~

Average TMIN = 4.8 degrees C
Average TMAX = 17.9 degrees C
Maximum TMAX : 52.8 degrees C
Minimum TMIN : -52.8 degrees C

Hottest station #1: USC00045502 - 35.6 degrees C
Hottest station #2: USC00022434 - 35.1 degrees C
Hottest station #3: USC00026250 - 34.0 degrees C
Hottest station #4: USC00024761 - 33.9 degrees C
Hottest station #5: USC00042319 - 33.8 degrees C

Coldest station #1: USR0000AHAY - -31.7 degrees C
Coldest station #2: USW00026508 - -29.9 degrees C
Coldest station #3: USC00508409 - -18.7 degrees C
Coldest station #4: USW00026440 - -18.5 degrees C
Coldest station #5: USC00509315 - -18.4 degrees C

Median TMAX in degrees C
20.6

Median TMIN in degrees C
4.4
```



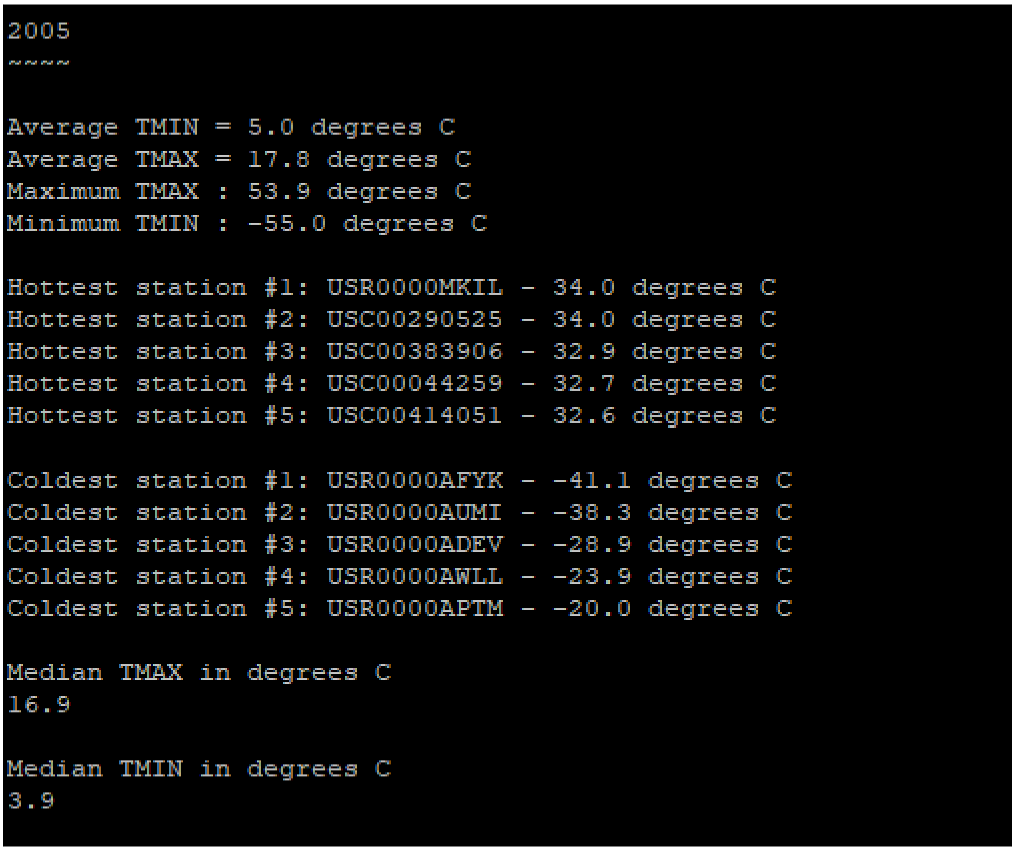
```
2002
~~~~

Average TMIN = 4.7 degrees C
Average TMAX = 17.7 degrees C
Maximum TMAX : 53.3 degrees C
Minimum TMIN : -47.2 degrees C

Hottest station #1: USC00022807 - 34.6 degrees C
Hottest station #2: USC00347254 - 34.4 degrees C
Hottest station #3: USC00042319 - 33.9 degrees C
Hottest station #4: USC00044259 - 33.4 degrees C
Hottest station #5: USR0000CBUU - 32.8 degrees C

Coldest station #1: USR0000NHAM - -19.2 degrees C
Coldest station #2: USS0051R01S - -19.0 degrees C
Coldest station #3: USC00502873 - -17.3 degrees C
Coldest station #4: USC00501987 - -15.1 degrees C
Coldest station #5: USC00263101 - -14.8 degrees C

Median TMAX in degrees C
22.0

Median TMIN in degrees C
7.8
```



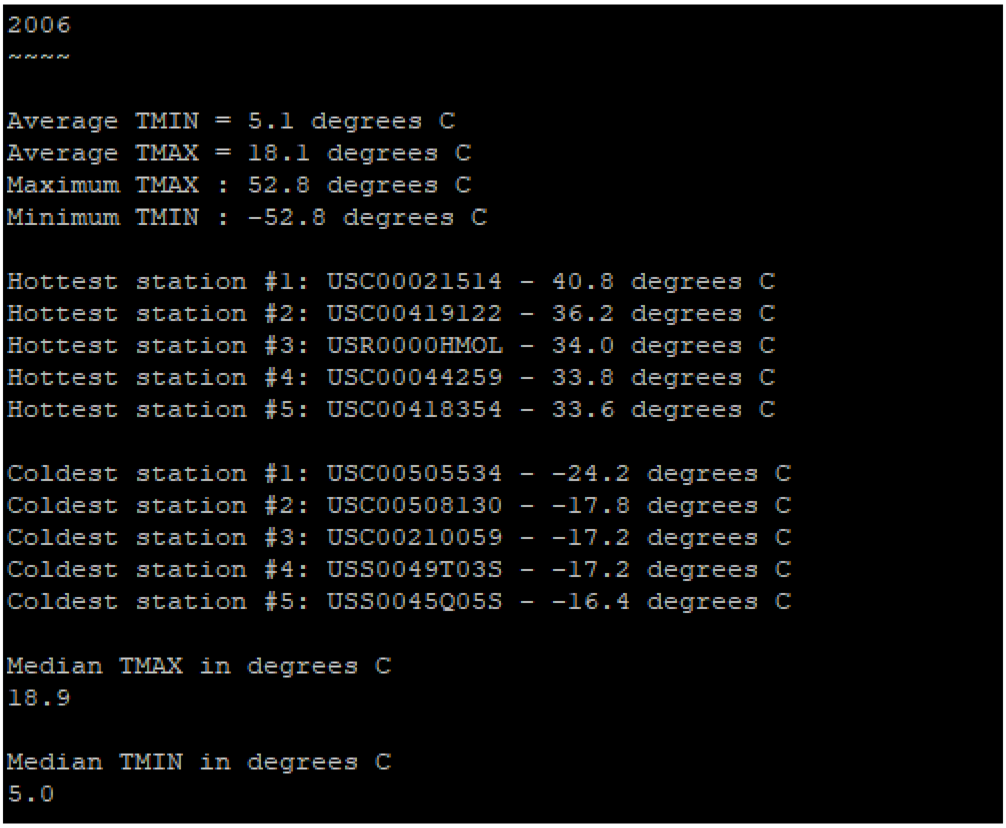
```
2003
~~~~

Average TMIN = 4.9 degrees C
Average TMAX = 17.7 degrees C
Maximum TMAX : 53.3 degrees C
Minimum TMIN : -50.0 degrees C

Hottest station #1: USC00344766 - 36.2 degrees C
Hottest station #2: USC00026194 - 35.1 degrees C
Hottest station #3: USC00045502 - 34.7 degrees C
Hottest station #4: USR0000CWIL - 34.6 degrees C
Hottest station #5: USC00412906 - 34.3 degrees C

Coldest station #1: USS0045M07S - -20.1 degrees C
Coldest station #2: USS0048V01S - -17.8 degrees C
Coldest station #3: USC00433581 - -16.6 degrees C
Coldest station #4: USC00067373 - -16.3 degrees C
Coldest station #5: USC00503181 - -15.7 degrees C

Median TMAX in degrees C
19.4

Median TMIN in degrees C
6.1
```



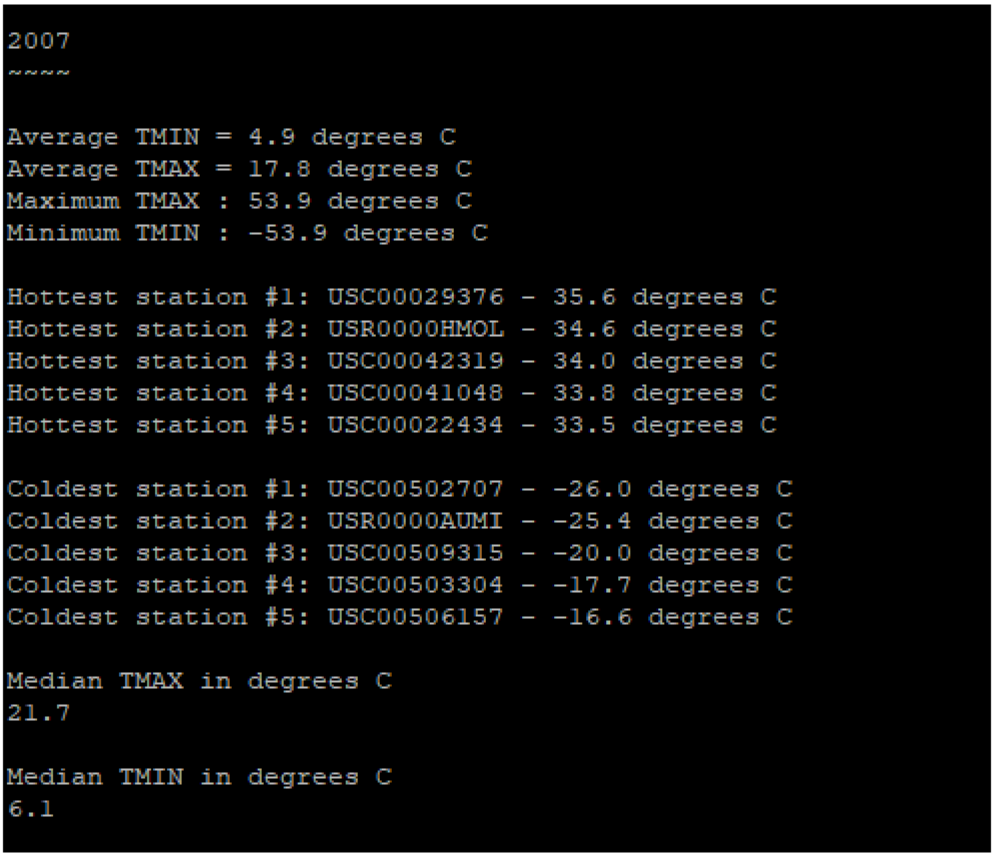
```
2004
~~~~

Average TMIN = 4.9 degrees C
Average TMAX = 17.5 degrees C
Maximum TMAX : 51.7 degrees C
Minimum TMIN : -53.3 degrees C

Hottest station #1: USW00053139 - 34.8 degrees C
Hottest station #2: USR0000CMIO - 34.3 degrees C
Hottest station #3: USC00044259 - 33.4 degrees C
Hottest station #4: USC00042319 - 33.0 degrees C
Hottest station #5: USC00029652 - 33.0 degrees C

Coldest station #1: USR0000ACOA - -25.4 degrees C
Coldest station #2: USR0000AOKL - -21.7 degrees C
Coldest station #3: USR0000ASTU - -21.7 degrees C
Coldest station #4: USR0000NLIM - -21.0 degrees C
Coldest station #5: USC00306957 - -19.7 degrees C

Median TMAX in degrees C
17.8

Median TMIN in degrees C
3.9
```



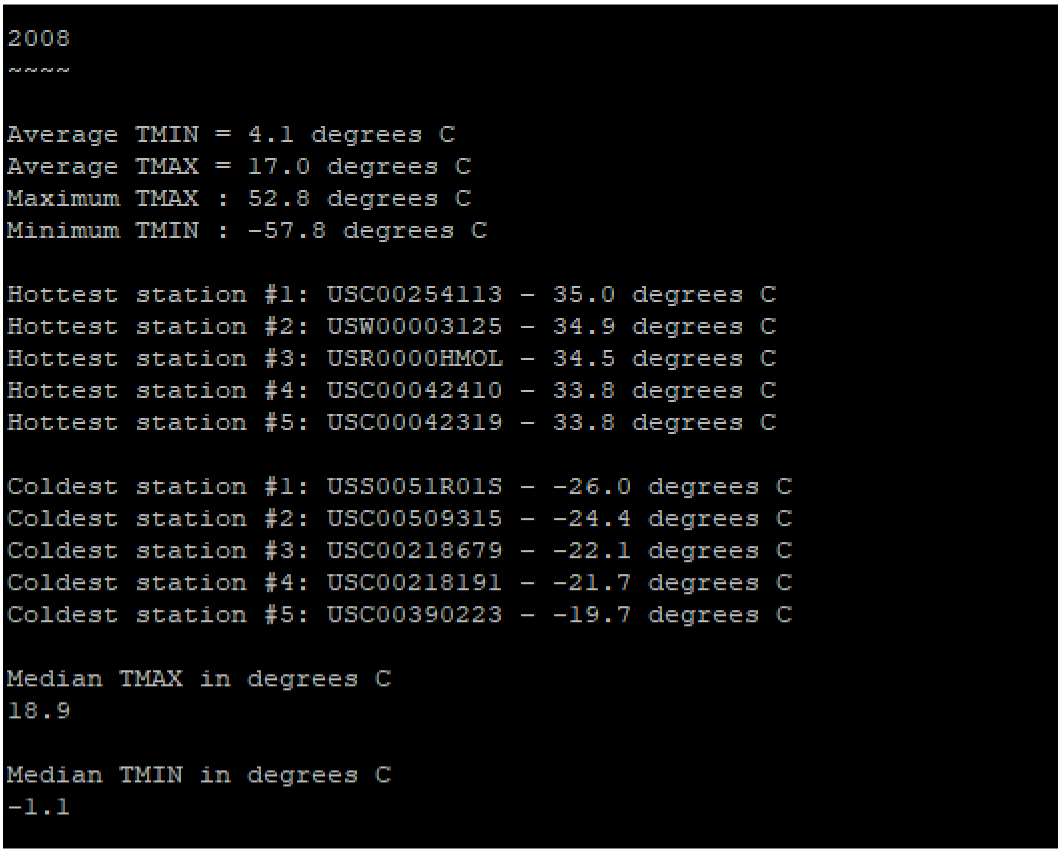
```
2005
~~~~

Average TMIN = 5.0 degrees C
Average TMAX = 17.8 degrees C
Maximum TMAX : 53.9 degrees C
Minimum TMIN : -55.0 degrees C

Hottest station #1: USR0000MKIL - 34.0 degrees C
Hottest station #2: USC00290525 - 34.0 degrees C
Hottest station #3: USC00383906 - 32.9 degrees C
Hottest station #4: USC00044259 - 32.7 degrees C
Hottest station #5: USC00414051 - 32.6 degrees C

Coldest station #1: USR0000AFYK - -41.1 degrees C
Coldest station #2: USR0000AUMI - -38.3 degrees C
Coldest station #3: USR0000ADEV - -28.9 degrees C
Coldest station #4: USR0000AWLL - -23.9 degrees C
Coldest station #5: USR0000APTM - -20.0 degrees C

Median TMAX in degrees C
16.9

Median TMIN in degrees C
3.9
```



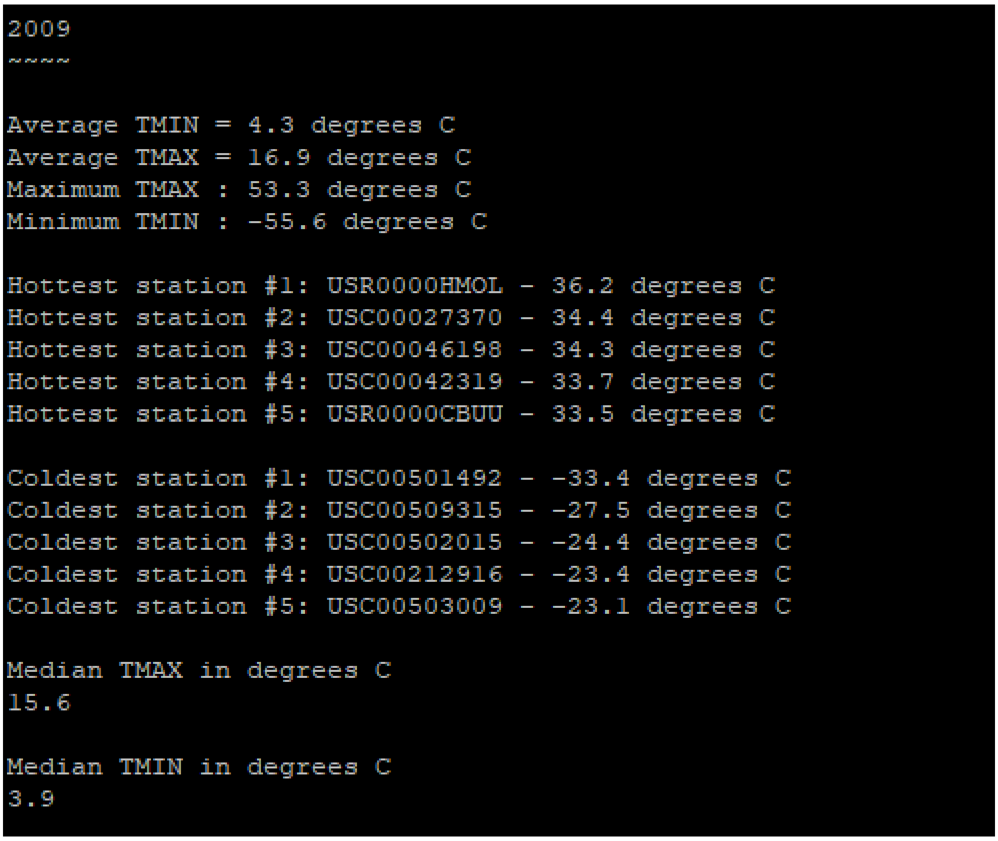
```
2006
~~~~

Average TMIN = 5.1 degrees C
Average TMAX = 18.1 degrees C
Maximum TMAX : 52.8 degrees C
Minimum TMIN : -52.8 degrees C

Hottest station #1: USC00021514 - 40.8 degrees C
Hottest station #2: USC00419122 - 36.2 degrees C
Hottest station #3: USR0000HMOL - 34.0 degrees C
Hottest station #4: USC00044259 - 33.8 degrees C
Hottest station #5: USC00418354 - 33.6 degrees C

Coldest station #1: USC00505534 - -24.2 degrees C
Coldest station #2: USC00508130 - -17.8 degrees C
Coldest station #3: USC00210059 - -17.2 degrees C
Coldest station #4: USS0049T03S - -17.2 degrees C
Coldest station #5: USS0045Q05S - -16.4 degrees C

Median TMAX in degrees C
18.9

Median TMIN in degrees C
5.0
```



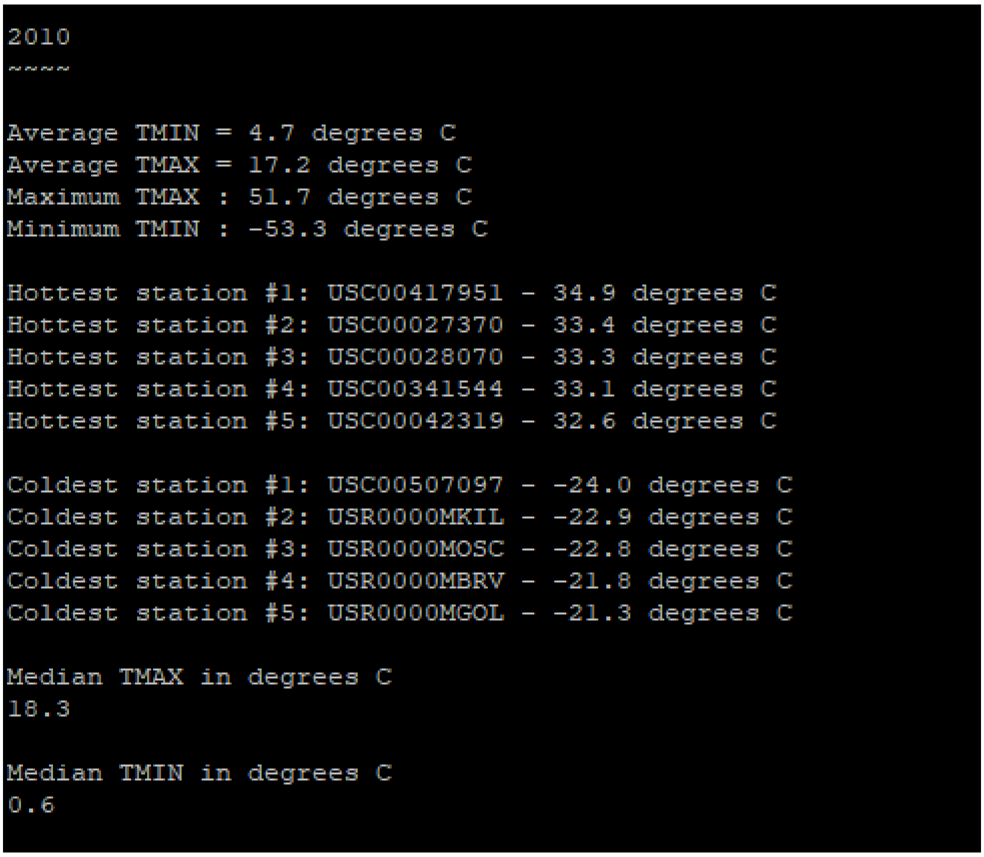
```
2007
~~~~

Average TMIN = 4.9 degrees C
Average TMAX = 17.8 degrees C
Maximum TMAX : 53.9 degrees C
Minimum TMIN : -53.9 degrees C

Hottest station #1: USC00029376 - 35.6 degrees C
Hottest station #2: USR0000HMOL - 34.6 degrees C
Hottest station #3: USC00042319 - 34.0 degrees C
Hottest station #4: USC00041048 - 33.8 degrees C
Hottest station #5: USC00022434 - 33.5 degrees C

Coldest station #1: USC00502707 - -26.0 degrees C
Coldest station #2: USR0000AUMI - -25.4 degrees C
Coldest station #3: USC00509315 - -20.0 degrees C
Coldest station #4: USC00503304 - -17.7 degrees C
Coldest station #5: USC00506157 - -16.6 degrees C

Median TMAX in degrees C
21.7

Median TMIN in degrees C
6.1
```



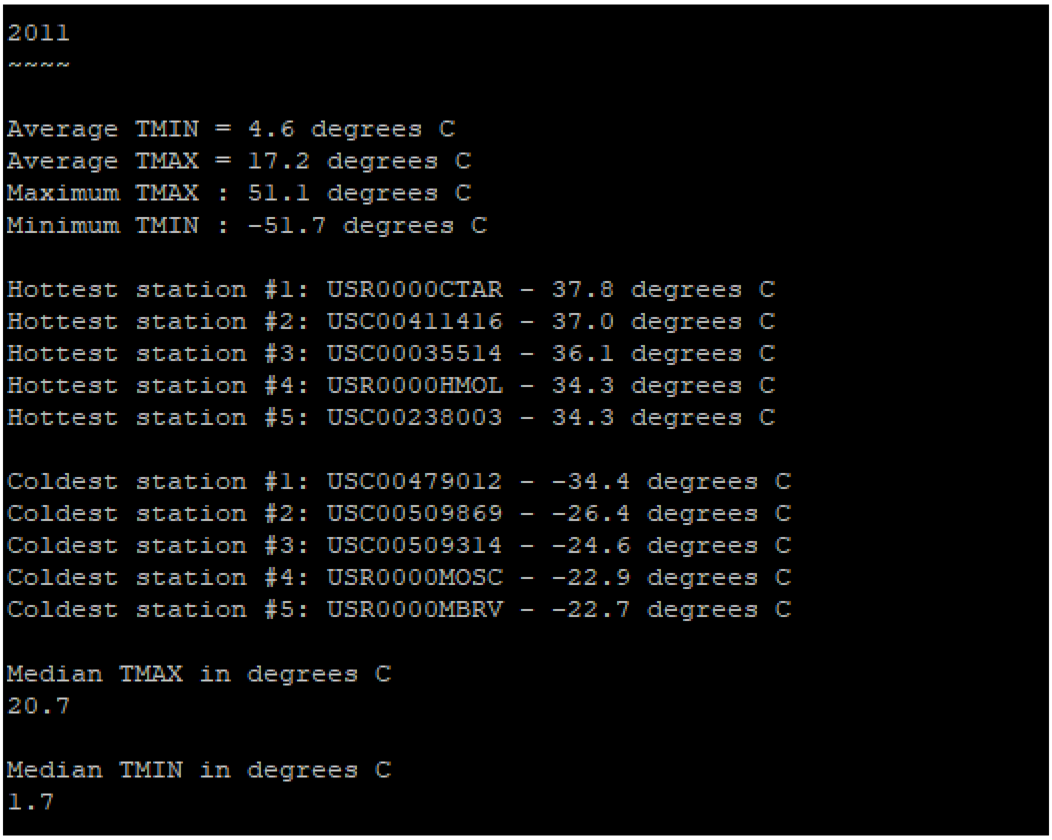
```
2008
~~~~

Average TMIN = 4.1 degrees C
Average TMAX = 17.0 degrees C
Maximum TMAX : 52.8 degrees C
Minimum TMIN : -57.8 degrees C

Hottest station #1: USC00254113 - 35.0 degrees C
Hottest station #2: USW00003125 - 34.9 degrees C
Hottest station #3: USR0000HMOL - 34.5 degrees C
Hottest station #4: USC00042410 - 33.8 degrees C
Hottest station #5: USC00042319 - 33.8 degrees C

Coldest station #1: USS0051R01S - -26.0 degrees C
Coldest station #2: USC00509315 - -24.4 degrees C
Coldest station #3: USC00218679 - -22.1 degrees C
Coldest station #4: USC00218191 - -21.7 degrees C
Coldest station #5: USC00390223 - -19.7 degrees C

Median TMAX in degrees C
18.9

Median TMIN in degrees C
-1.1
```



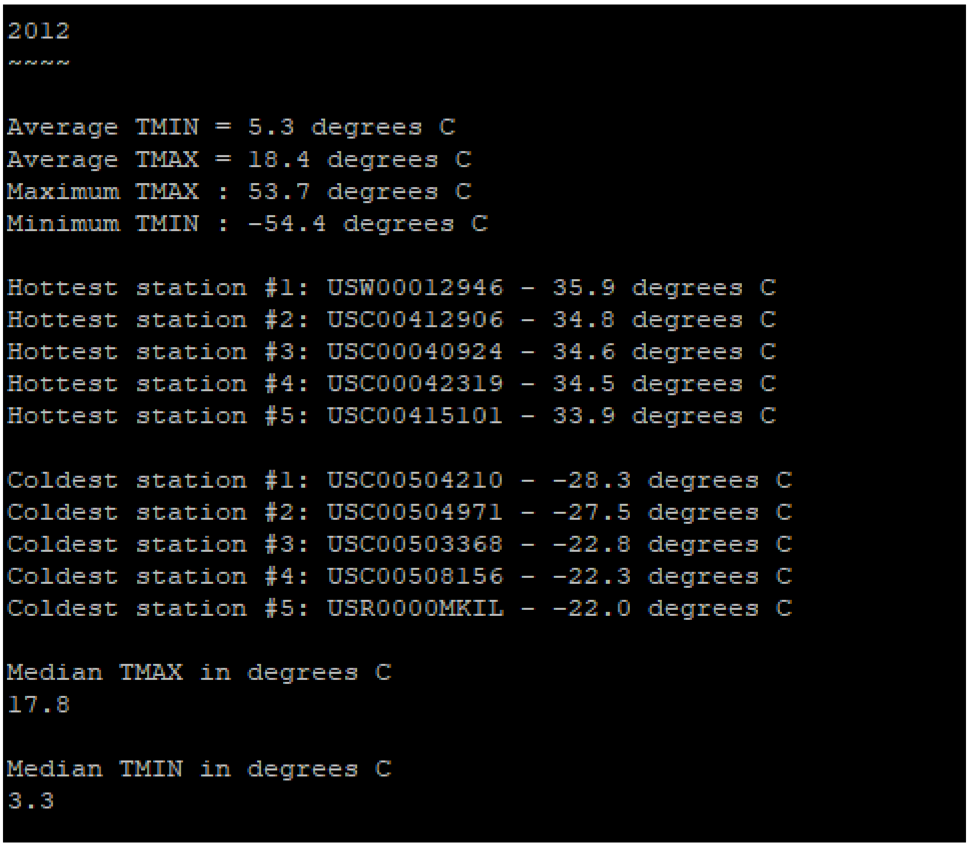
```
2009
~~~~

Average TMIN = 4.3 degrees C
Average TMAX = 16.9 degrees C
Maximum TMAX : 53.3 degrees C
Minimum TMIN : -55.6 degrees C

Hottest station #1: USR0000HMOL - 36.2 degrees C
Hottest station #2: USC00027370 - 34.4 degrees C
Hottest station #3: USC00046198 - 34.3 degrees C
Hottest station #4: USC00042319 - 33.7 degrees C
Hottest station #5: USR0000CBUU - 33.5 degrees C

Coldest station #1: USC00501492 - -33.4 degrees C
Coldest station #2: USC00509315 - -27.5 degrees C
Coldest station #3: USC00502015 - -24.4 degrees C
Coldest station #4: USC00212916 - -23.4 degrees C
Coldest station #5: USC00503009 - -23.1 degrees C

Median TMAX in degrees C
15.6

Median TMIN in degrees C
3.9
```



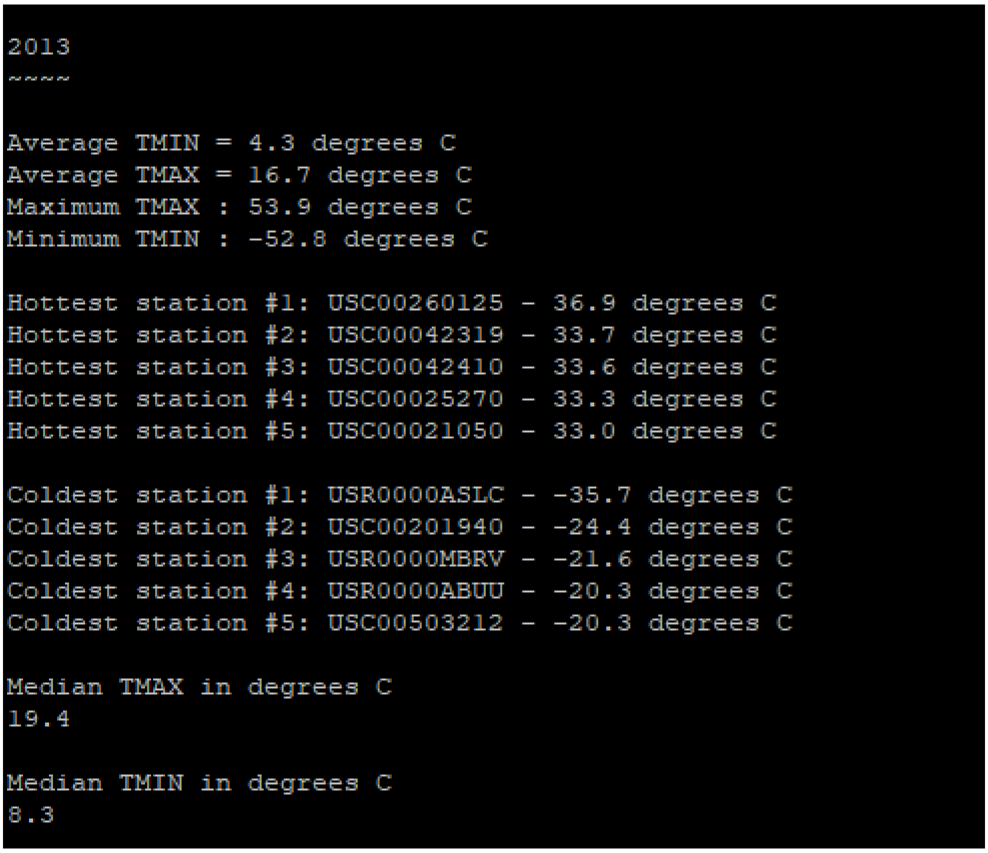
```
2010
~~~~

Average TMIN = 4.7 degrees C
Average TMAX = 17.2 degrees C
Maximum TMAX : 51.7 degrees C
Minimum TMIN : -53.3 degrees C

Hottest station #1: USC00417951 - 34.9 degrees C
Hottest station #2: USC00027370 - 33.4 degrees C
Hottest station #3: USC00028070 - 33.3 degrees C
Hottest station #4: USC00341544 - 33.1 degrees C
Hottest station #5: USC00042319 - 32.6 degrees C

Coldest station #1: USC00507097 - -24.0 degrees C
Coldest station #2: USR0000MKIL - -22.9 degrees C
Coldest station #3: USR0000MOSC - -22.8 degrees C
Coldest station #4: USR0000MBRV - -21.8 degrees C
Coldest station #5: USR0000MGOL - -21.3 degrees C

Median TMAX in degrees C
18.3

Median TMIN in degrees C
0.6
```



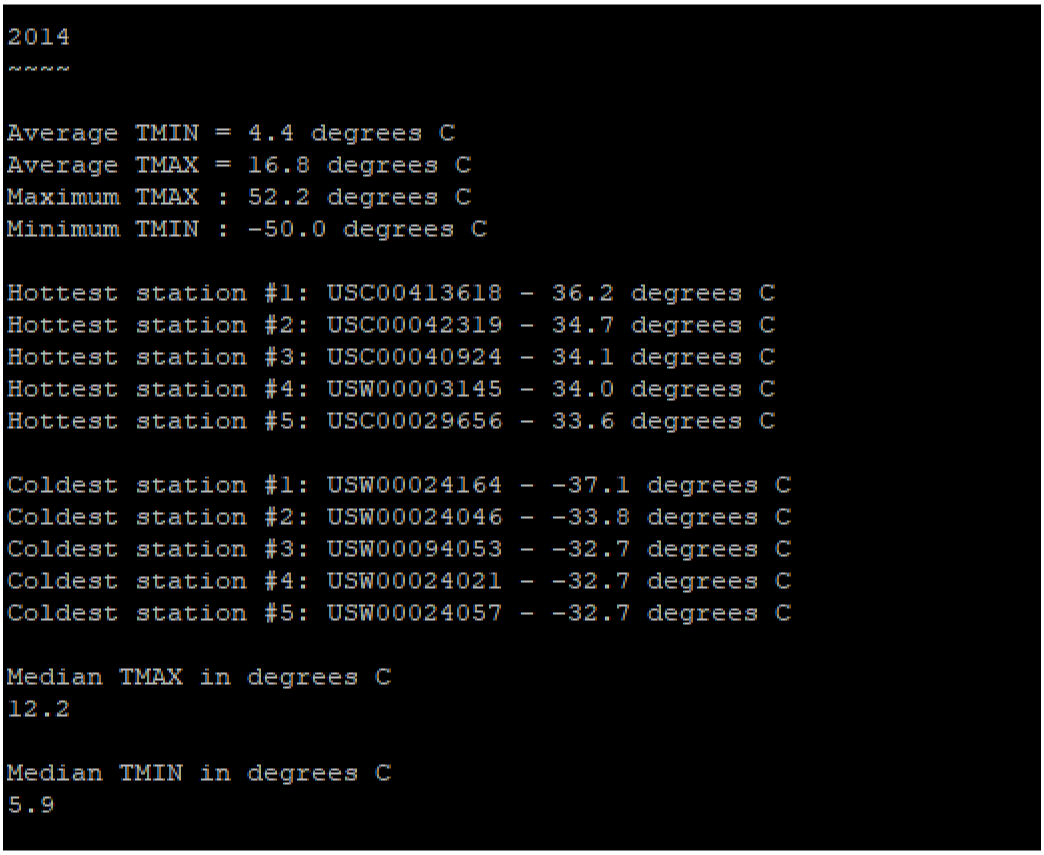
```
2011
~~~~

Average TMIN = 4.6 degrees C
Average TMAX = 17.2 degrees C
Maximum TMAX : 51.1 degrees C
Minimum TMIN : -51.7 degrees C

Hottest station #1: USR0000CTAR - 37.8 degrees C
Hottest station #2: USC00411416 - 37.0 degrees C
Hottest station #3: USC00035514 - 36.1 degrees C
Hottest station #4: USR0000HMOL - 34.3 degrees C
Hottest station #5: USC00238003 - 34.3 degrees C

Coldest station #1: USC00479012 - -34.4 degrees C
Coldest station #2: USC00509869 - -26.4 degrees C
Coldest station #3: USC00509314 - -24.6 degrees C
Coldest station #4: USR0000MOSC - -22.9 degrees C
Coldest station #5: USR0000MBRV - -22.7 degrees C

Median TMAX in degrees C
20.7

Median TMIN in degrees C
1.7
```



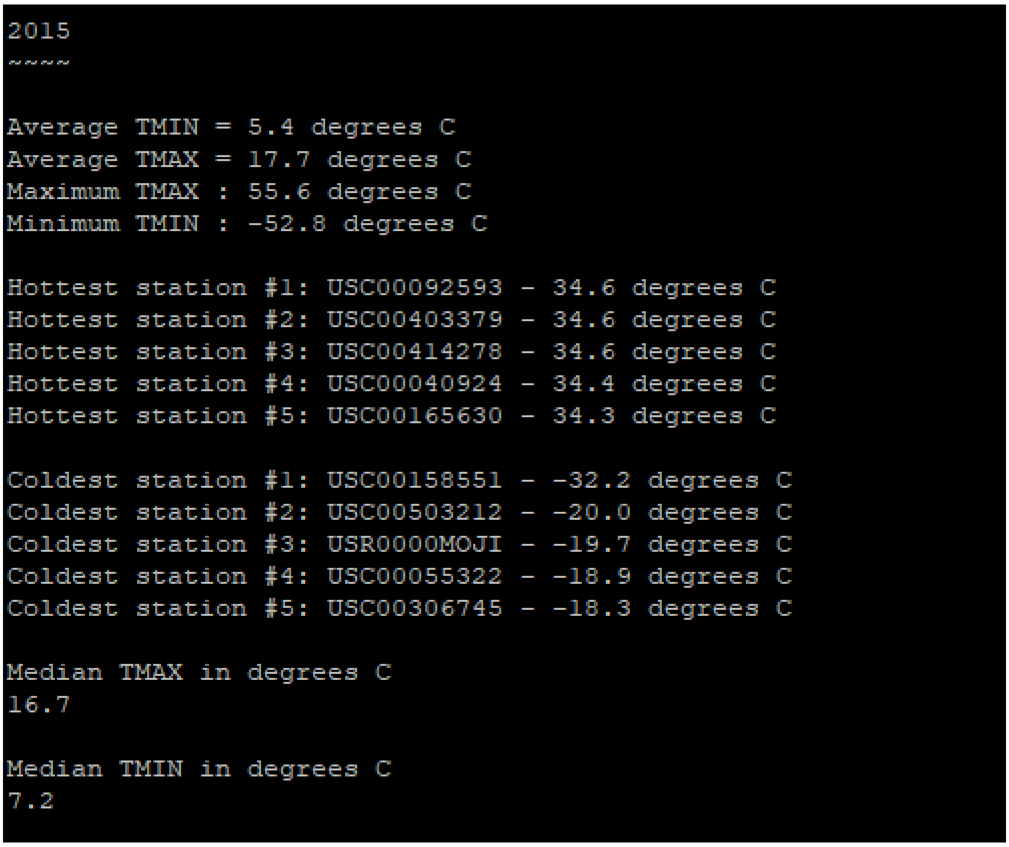
```
2012
~~~~

Average TMIN = 5.3 degrees C
Average TMAX = 18.4 degrees C
Maximum TMAX : 53.7 degrees C
Minimum TMIN : -54.4 degrees C

Hottest station #1: USW00012946 - 35.9 degrees C
Hottest station #2: USC00412906 - 34.8 degrees C
Hottest station #3: USC00040924 - 34.6 degrees C
Hottest station #4: USC00042319 - 34.5 degrees C
Hottest station #5: USC00415101 - 33.9 degrees C

Coldest station #1: USC00504210 - -28.3 degrees C
Coldest station #2: USC00504971 - -27.5 degrees C
Coldest station #3: USC00503368 - -22.8 degrees C
Coldest station #4: USC00508156 - -22.3 degrees C
Coldest station #5: USR0000MKIL - -22.0 degrees C

Median TMAX in degrees C
17.8

Median TMIN in degrees C
3.3
```



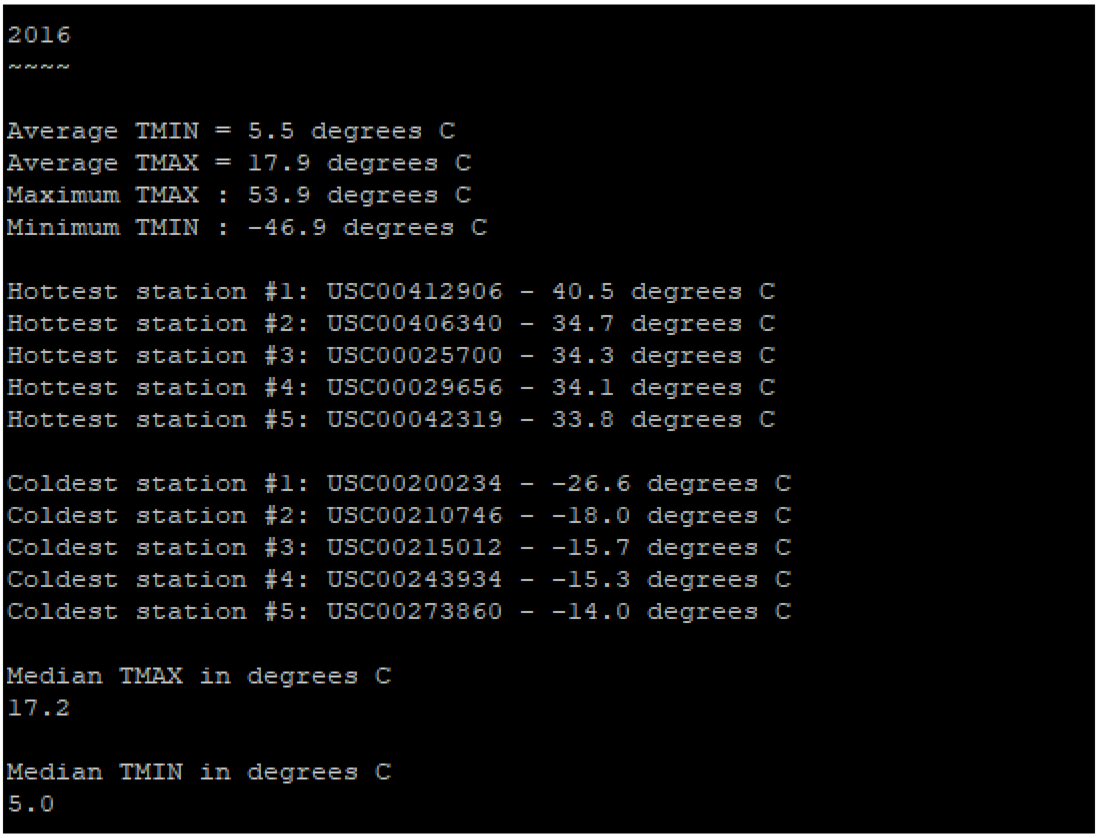
```
2013
~~~~

Average TMIN = 4.3 degrees C
Average TMAX = 16.7 degrees C
Maximum TMAX : 53.9 degrees C
Minimum TMIN : -52.8 degrees C

Hottest station #1: USC00260125 - 36.9 degrees C
Hottest station #2: USC00042319 - 33.7 degrees C
Hottest station #3: USC00042410 - 33.6 degrees C
Hottest station #4: USC00025270 - 33.3 degrees C
Hottest station #5: USC00021050 - 33.0 degrees C

Coldest station #1: USR0000ASLC - -35.7 degrees C
Coldest station #2: USC00201940 - -24.4 degrees C
Coldest station #3: USR0000MBRV - -21.6 degrees C
Coldest station #4: USR0000ABUU - -20.3 degrees C
Coldest station #5: USC00503212 - -20.3 degrees C

Median TMAX in degrees C
19.4

Median TMIN in degrees C
8.3
```



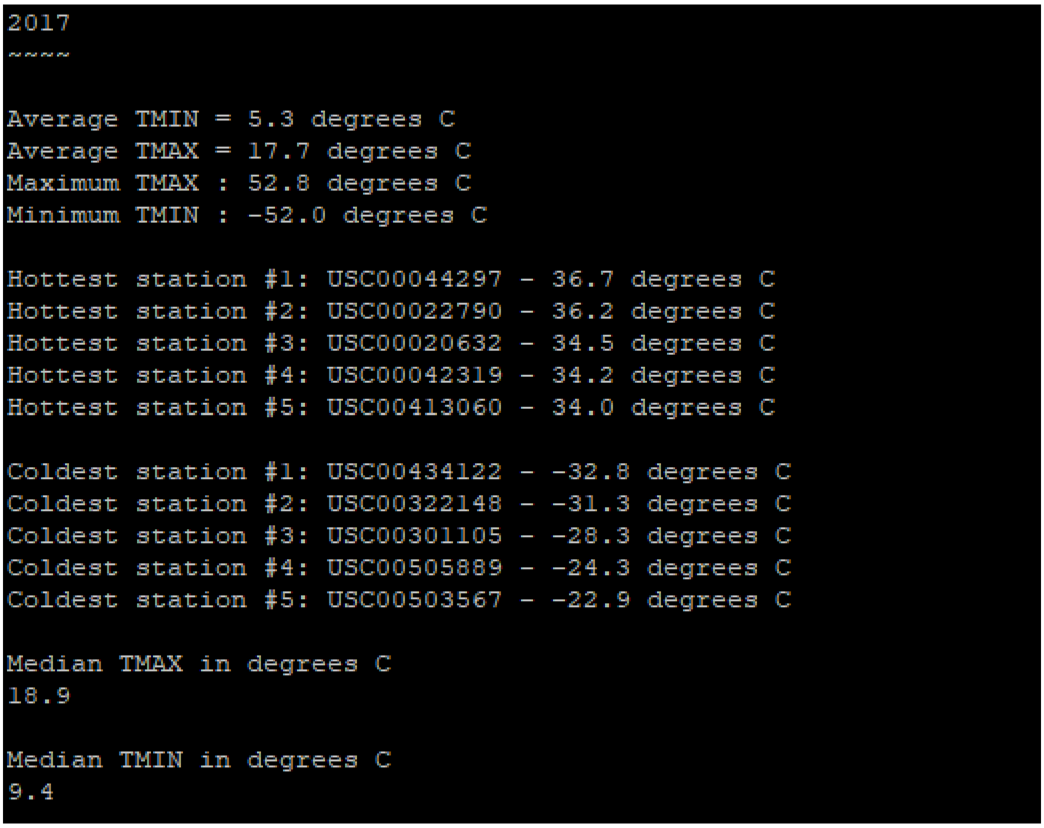
```
2014
~~~~

Average TMIN = 4.4 degrees C
Average TMAX = 16.8 degrees C
Maximum TMAX : 52.2 degrees C
Minimum TMIN : -50.0 degrees C

Hottest station #1: USC00413618 - 36.2 degrees C
Hottest station #2: USC00042319 - 34.7 degrees C
Hottest station #3: USC00040924 - 34.1 degrees C
Hottest station #4: USW00003145 - 34.0 degrees C
Hottest station #5: USC00029656 - 33.6 degrees C

Coldest station #1: USW00024164 - -37.1 degrees C
Coldest station #2: USW00024046 - -33.8 degrees C
Coldest station #3: USW00094053 - -32.7 degrees C
Coldest station #4: USW00024021 - -32.7 degrees C
Coldest station #5: USW00024057 - -32.7 degrees C

Median TMAX in degrees C
12.2

Median TMIN in degrees C
5.9
```



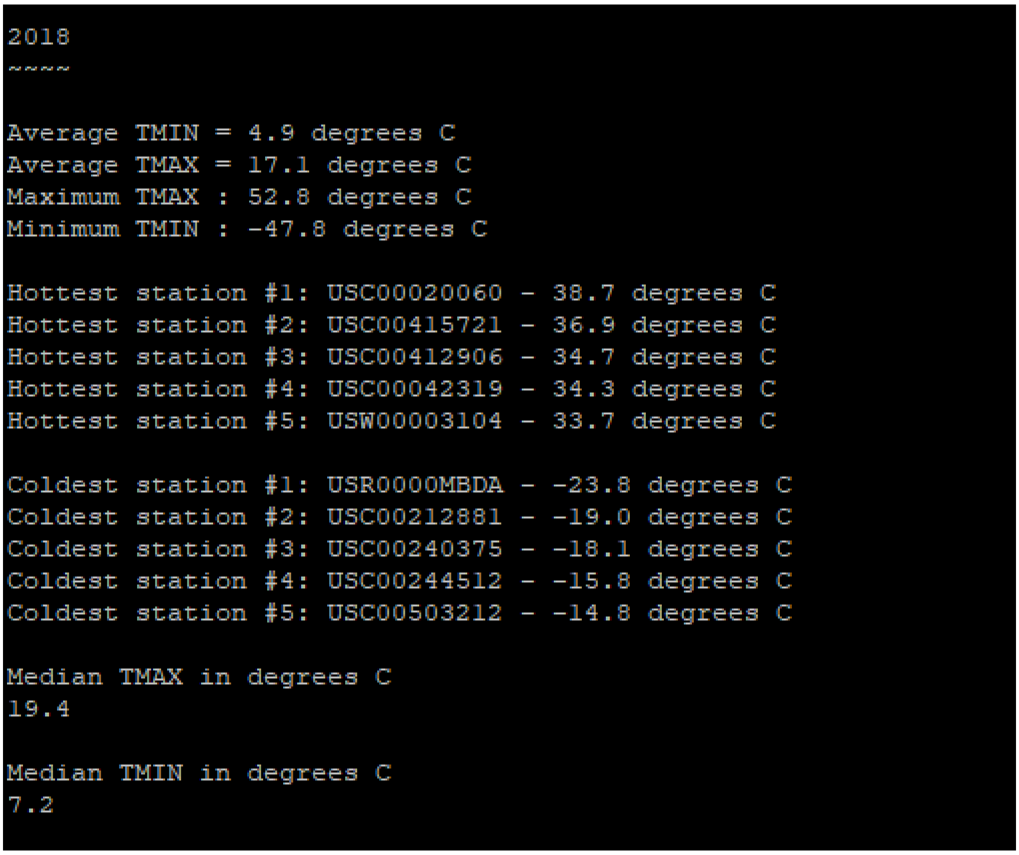
```
2015
~~~~

Average TMIN = 5.4 degrees C
Average TMAX = 17.7 degrees C
Maximum TMAX : 55.6 degrees C
Minimum TMIN : -52.8 degrees C

Hottest station #1: USC00092593 - 34.6 degrees C
Hottest station #2: USC00403379 - 34.6 degrees C
Hottest station #3: USC00414278 - 34.6 degrees C
Hottest station #4: USC00040924 - 34.4 degrees C
Hottest station #5: USC00165630 - 34.3 degrees C

Coldest station #1: USC00158551 - -32.2 degrees C
Coldest station #2: USC00503212 - -20.0 degrees C
Coldest station #3: USR0000MOJI - -19.7 degrees C
Coldest station #4: USC00055322 - -18.9 degrees C
Coldest station #5: USC00306745 - -18.3 degrees C

Median TMAX in degrees C
16.7

Median TMIN in degrees C
7.2
```



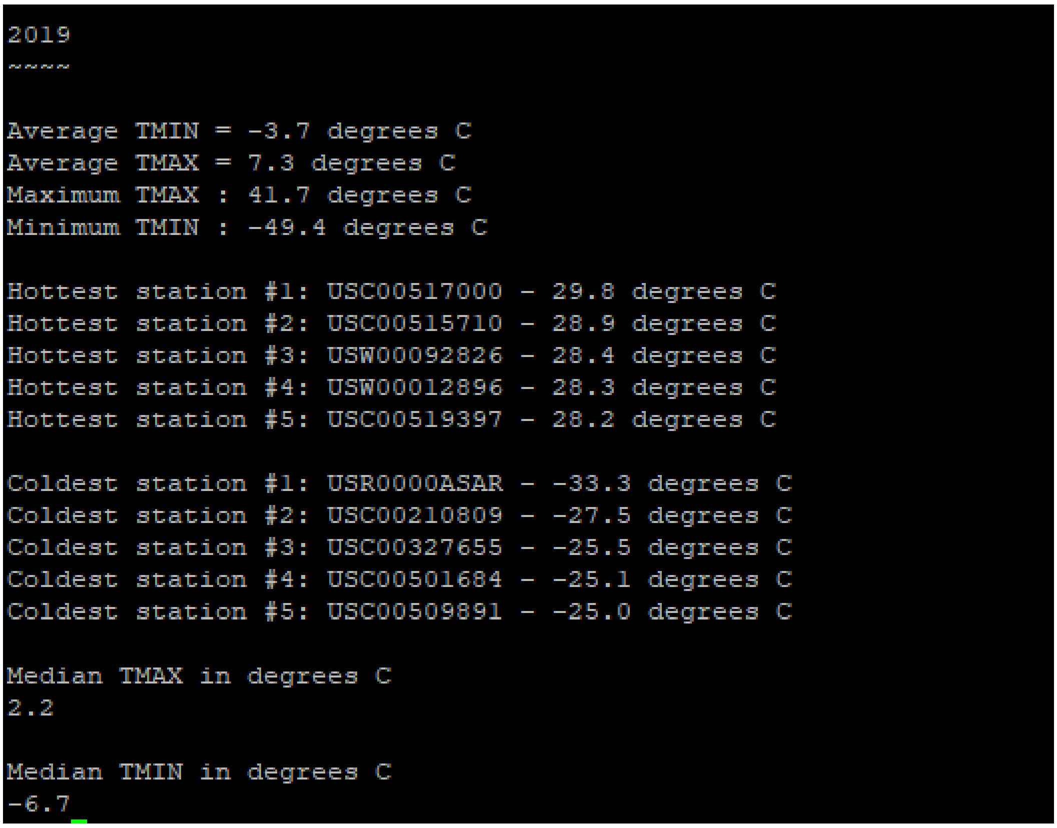
```
2016
~~~~

Average TMIN = 5.5 degrees C
Average TMAX = 17.9 degrees C
Maximum TMAX : 53.9 degrees C
Minimum TMIN : -46.9 degrees C

Hottest station #1: USC00412906 - 40.5 degrees C
Hottest station #2: USC00406340 - 34.7 degrees C
Hottest station #3: USC00025700 - 34.3 degrees C
Hottest station #4: USC00029656 - 34.1 degrees C
Hottest station #5: USC00042319 - 33.8 degrees C

Coldest station #1: USC00200234 - -26.6 degrees C
Coldest station #2: USC00210746 - -18.0 degrees C
Coldest station #3: USC00215012 - -15.7 degrees C
Coldest station #4: USC00243934 - -15.3 degrees C
Coldest station #5: USC00273860 - -14.0 degrees C

Median TMAX in degrees C
17.2

Median TMIN in degrees C
5.0
```



```
2017
~~~~

Average TMIN = 5.3 degrees C
Average TMAX = 17.7 degrees C
Maximum TMAX : 52.8 degrees C
Minimum TMIN : -52.0 degrees C

Hottest station #1: USC00044297 - 36.7 degrees C
Hottest station #2: USC00022790 - 36.2 degrees C
Hottest station #3: USC00020632 - 34.5 degrees C
Hottest station #4: USC00042319 - 34.2 degrees C
Hottest station #5: USC00413060 - 34.0 degrees C

Coldest station #1: USC00434122 - -32.8 degrees C
Coldest station #2: USC00322148 - -31.3 degrees C
Coldest station #3: USC00301105 - -28.3 degrees C
Coldest station #4: USC00505889 - -24.3 degrees C
Coldest station #5: USC00503567 - -22.9 degrees C

Median TMAX in degrees C
18.9

Median TMIN in degrees C
9.4
```



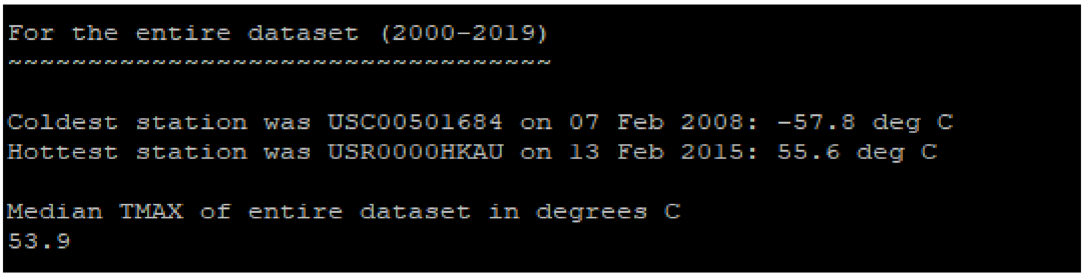
```
2018
~~~~

Average TMIN = 4.9 degrees C
Average TMAX = 17.1 degrees C
Maximum TMAX : 52.8 degrees C
Minimum TMIN : -47.8 degrees C

Hottest station #1: USC00020060 - 38.7 degrees C
Hottest station #2: USC00415721 - 36.9 degrees C
Hottest station #3: USC00412906 - 34.7 degrees C
Hottest station #4: USC00042319 - 34.3 degrees C
Hottest station #5: USW00003104 - 33.7 degrees C

Coldest station #1: USR0000MBDA - -23.8 degrees C
Coldest station #2: USC00212881 - -19.0 degrees C
Coldest station #3: USC00240375 - -18.1 degrees C
Coldest station #4: USC00244512 - -15.8 degrees C
Coldest station #5: USC00503212 - -14.8 degrees C

Median TMAX in degrees C
19.4

Median TMIN in degrees C
7.2
```



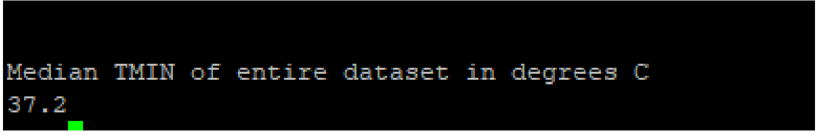
```
2019
~~~~

Average TMIN = -3.7 degrees C
Average TMAX = 7.3 degrees C
Maximum TMAX : 41.7 degrees C
Minimum TMIN : -49.4 degrees C

Hottest station #1: USC00517000 - 29.8 degrees C
Hottest station #2: USC00515710 - 28.9 degrees C
Hottest station #3: USW00092826 - 28.4 degrees C
Hottest station #4: USW00012896 - 28.3 degrees C
Hottest station #5: USC00519397 - 28.2 degrees C

Coldest station #1: USR0000ASAR - -33.3 degrees C
Coldest station #2: USC00210809 - -27.5 degrees C
Coldest station #3: USC00327655 - -25.5 degrees C
Coldest station #4: USC00501684 - -25.1 degrees C
Coldest station #5: USC00509891 - -25.0 degrees C

Median TMAX in degrees C
2.2

Median TMIN in degrees C
-6.7
```

To obtain results for 2.5, 2.6 and 2.8, use the following command on PySpark Prompt:

>>> runFullDataset()



```
For the entire dataset (2000-2019)
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Coldest station was USC00501684 on 07 Feb 2008: -57.8 deg C
Hottest station was USR0000HKAU on 13 Feb 2015: 55.6 deg C

Median TMAX of entire dataset in degrees C
53.9
```



```
Median TMIN of entire dataset in degrees C
37.2
```

References:

https://spark.apache.org/docs/2.3.0/api/python/pyspark.sql.html#
https://www.edureka.co/blog/pyspark-dataframe-tutorial/
https://www.edureka.co/blog/spark-sql-tutorial/