# STAT 480 Ka Ki Lai (kakilai2)
# Homework 6 Report

## Code for data preparation:

```
#pig
#records = LOAD 'input/ncdc/19011910.txt' AS (usaf:chararray, wban:int, year: int, temp: int);
#station = LOAD 'input/ncdc/stationlistshort.txt' AS (usaf:chararray, wban:int, name:chararray);
```

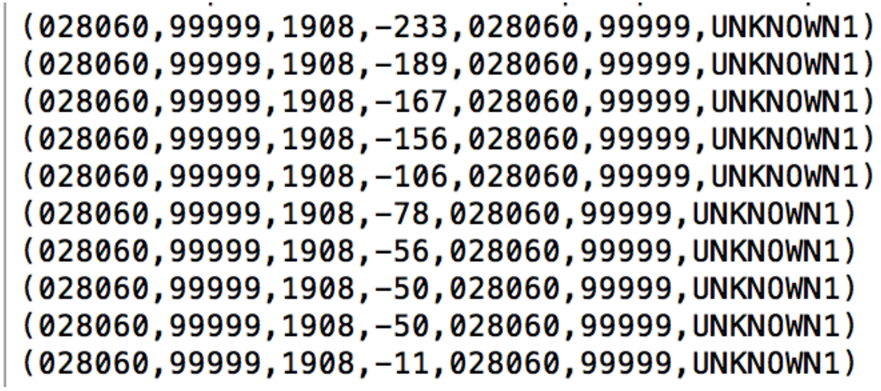
## Exercise 1:

The following code join the observed temperature data with the station name data so that the location name will be included within each observation in the relation:

```
# recordnstation = JOIN records BY $0, station BY $0;
#lim_result = LIMIT recordnstation 10;
#DESCRIBE lim_result;
#DUMP lim_result;
```

The result of first 10 entries is shown below:

Column Names: 1. usaf, 2. wban, 3. year, 4. temperature, 5. usaf, 6. wban 7. location name



```
(028060,99999,1908,-233,028060,99999,UNKNOWN1)
(028060,99999,1908,-189,028060,99999,UNKNOWN1)
(028060,99999,1908,-167,028060,99999,UNKNOWN1)
(028060,99999,1908,-156,028060,99999,UNKNOWN1)
(028060,99999,1908,-106,028060,99999,UNKNOWN1)
(028060,99999,1908,-78,028060,99999,UNKNOWN1)
(028060,99999,1908,-56,028060,99999,UNKNOWN1)
(028060,99999,1908,-50,028060,99999,UNKNOWN1)
(028060,99999,1908,-50,028060,99999,UNKNOWN1)
(028060,99999,1908,-11,028060,99999,UNKNOWN1)
```

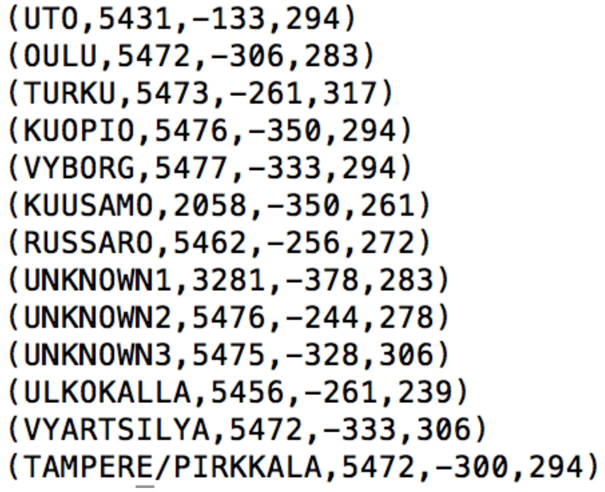
From the above result above, we can see that the location name has been added to each observation accordingly.

## Exercise 2:

*Code:*

```
#C = GROUP recordnstation BY name;
#MaxMin_TEMP = FOREACH C GENERATE group,COUNT($1),MIN(recordnstation.temp),
#MAX(recordnstation.temp) ;
#DESCRIBE MaxMin_TEMP;
#DUMP MaxMin_TEMP;
```

The number of trusted temperature observations, the minimum and maximum temperatures by station are shown as below:
Column Names : 1. Station, 2. Number of temperature observations, 3. Min Temp, 4. Max Temp

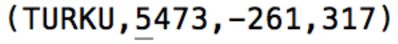


```
(UTO,5431,-133,294)
(OULU,5472,-306,283)
(TURKU,5473,-261,317)
(KUOPIO,5476,-350,294)
(VYBORG,5477,-333,294)
(KUUSAMO,2058,-350,261)
(RUSSARO,5462,-256,272)
(UNKNOWN1,3281,-378,283)
(UNKNOWN2,5476,-244,278)
(UNKNOWN3,5475,-328,306)
(ULKOKALLA,5456,-261,239)
(VYARTSILYA,5472,-333,306)
(TAMPERE/PIRKKALA,5472,-300,294)
```

**Exercise 3:**

*Code*
*#ord  = ORDER MaxMin_TEMP by $3 DESC;*
*#max_record = LIMIT ord 1;*
*#DUMP max_record;*

*Output:*
*Column names: 1. Location name, 2. Count of observations, 3. Min Temp, 4. Max Temp*

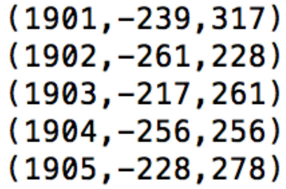


```
(TURKU,5473,-261,317)
```

*Hence, the location with highest max temp is TURKU.*

*#filtered = FILTER recordnstation BY name == max_record.$0;*
*#grp_records = GROUP filtered BY year;*
*#maxmintemp = FOREACH grp_records GENERATE group,MIN(filtered.temp),*
*#MAX(filtered.temp) ;*
*#DESCRIBE maxmintemp;*
*#DUMP maxmintemp;*

Column Names: 1. Year, 2. Min Temp, 3. Max Temp (for the station with highest maximum temperature)



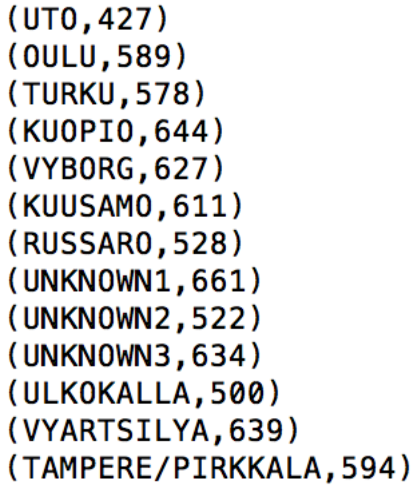
```
(1901,-239,317)
(1902,-261,228)
(1903,-217,261)
(1904,-256,256)
(1905,-228,278)
```

**Exercise 4:**

*code*

*#range = FOREACH MaxMin_TEMP GENERATE group, $3-$2;*
*#DESCRIBE range;*
*#DUMP range;*

The temperature range for each location is shown below:
Col names: 1. Station, 2. temperature range



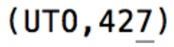
```
(UTO,427)
(OULU,589)
(TURKU,578)
(KUOPIO,644)
(VYBORG,627)
(KUUSAMO,611)
(RUSSARO,528)
(UNKNOWN1,661)
(UNKNOWN2,522)
(UNKNOWN3,634)
(ULKOKALLA,500)
(VYARTSILYA,639)
(TAMPERE/PIRKKALA,594)
```

The following code find the station name and temperature range for the station with smallest temperature range for the time period:

*#ord_range = ORDER range by $1;*
*#min_range = LIMIT ord_range 1;*
*#DESCRIBE min_range;*
*#DUMP min_range;*

Column names:

1. Station with minimum temperature range, 2. range of temperature



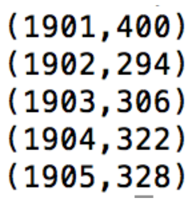
```
(UTO,427)
```

Hence, the station with minimum temperature range is UTO with a range of 427.

To obtain that station's temperature ranges by year:
Code:

*#filtered_mr = FILTER recordnstation BY name == min_range.$0;*
*#grp_mr = GROUP filtered_mr BY year;*
*#range_mr = FOREACH grp_mr GENERATE group, MAX(filtered_mr.temp) -*
*#MIN(filtered_mr.temp);*
*#DESCRIBE range_mr;*
*#DUMP range_mr;*

Hence, the station's temperature ranges by year is as below:
Column Names: 1. year, 2. Range of Temperature



```
(1901,400)
(1902,294)
(1903,306)
(1904,322)
(1905,328)
```