Weather in Saudi Arabia

(Hokage Heroes)

CONTENTS

01 Introduction

02 Pre-processing

03 Mapreducer

04 Conclusion

Introduction

Vision,Problem Statement, Data description

Vision 2030

Vision 2030 seeks to build a green future.

Where the **Saudi Green Initiative** aims to protect the environment and deal with climate action and energy transition to achieve a green future.
It seeks to increase vegetation cover and help combat desertification through carefully selected afforestation initiatives throughout the Kingdom, assisted by increasing precipitation levels through Cloud seeding technology.

Problem Statement

The plant needs certain conditions for growth, including air temperature. And the growth of plants varies from one season to another due to the change in weather. In addition to that, the climate in Saudi Arabia is different as a result of its wide area. So what would be the appropriate plant to grow in a specific area and a specific season?

This is what we aim to find a solution for,
what is the appropriate temperature depending on the region and plants ?

Our Data

Data description

01

Saudi Arabia climate integrated surface data with the below data observations :

1. Wind
2. Sky condition
3. Visibility
4. Air temperature
5. Sea level pressure

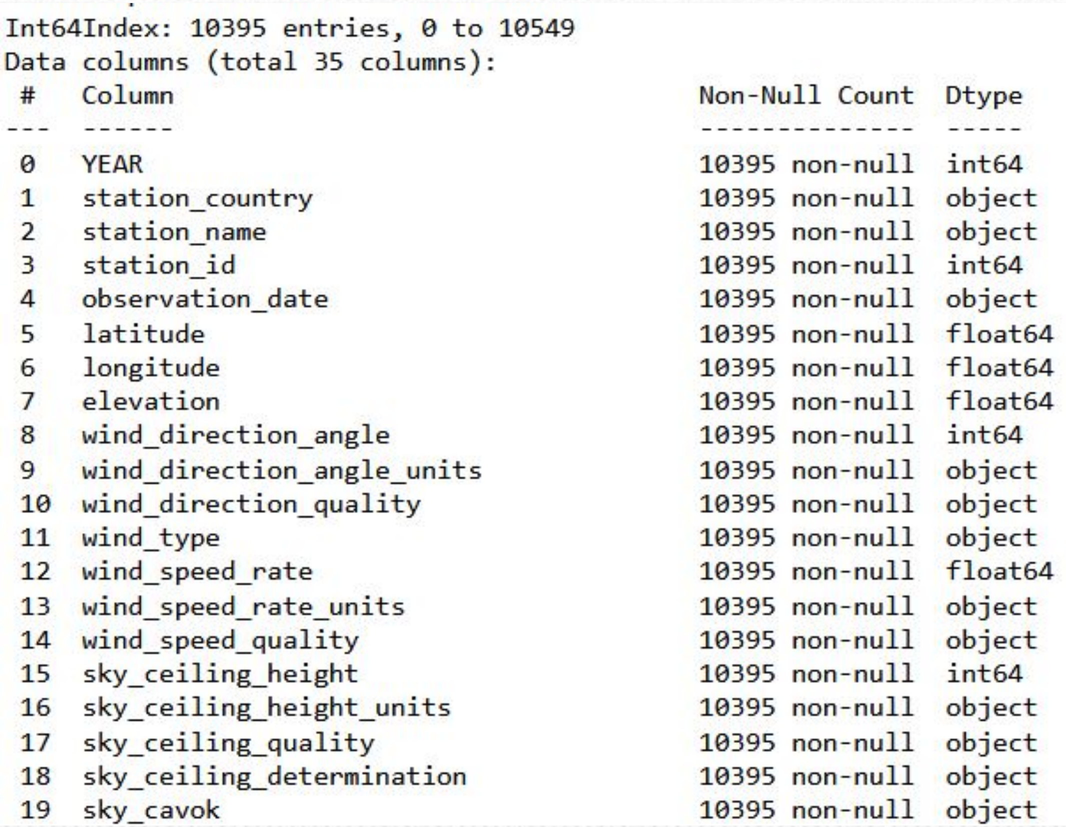
Data description

02

A set of data that contains information about plants and the climate suitable for them

1. Crop name
2. Min/Max/Avg temperature

Dataset Info

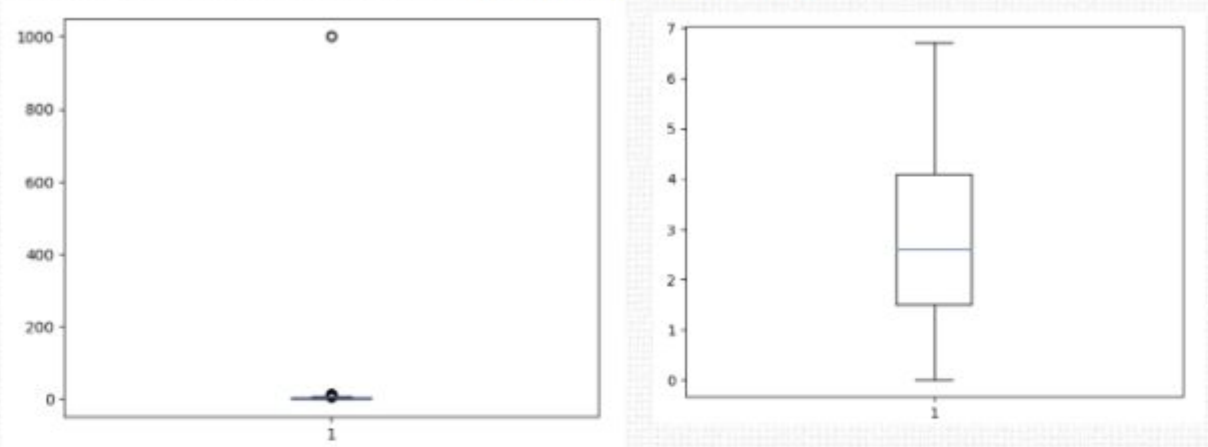


```
Int64Index: 10395 entries, 0 to 10549
Data columns (total 35 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   YEAR                            10395 non-null  int64
 1   station_country                 10395 non-null  object
 2   station_name                    10395 non-null  object
 3   station_id                      10395 non-null  int64
 4   observation_date                10395 non-null  object
 5   latitude                        10395 non-null  float64
 6   longitude                       10395 non-null  float64
 7   elevation                       10395 non-null  float64
 8   wind_direction_angle            10395 non-null  int64
 9   wind_direction_angle_units      10395 non-null  object
 10  wind_direction_quality          10395 non-null  object
 11  wind_type                       10395 non-null  object
 12  wind_speed_rate                 10395 non-null  float64
 13  wind_speed_rate_units           10395 non-null  object
 14  wind_speed_quality              10395 non-null  object
 15  sky_ceiling_height              10395 non-null  int64
 16  sky_ceiling_height_units        10395 non-null  object
 17  sky_ceiling_quality             10395 non-null  object
 18  sky_ceiling_determination       10395 non-null  object
 19  sky_cavok                       10395 non-null  object
```

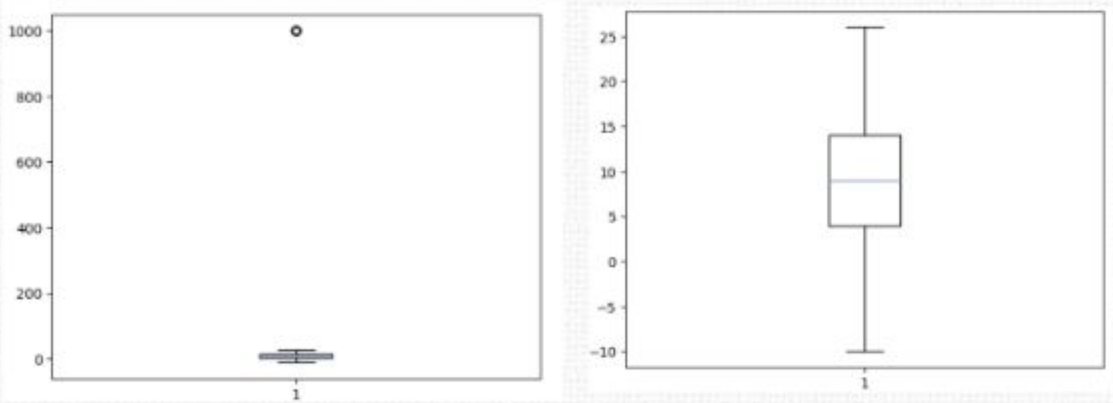
Pre-processing

Outlier

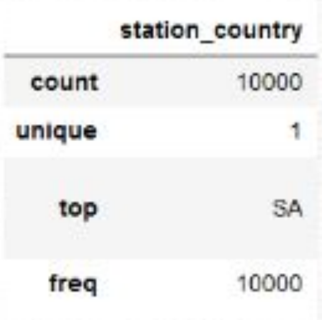
Wind_speed_rate > 100



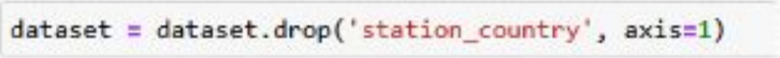
Air_temperature_dew_point > 50



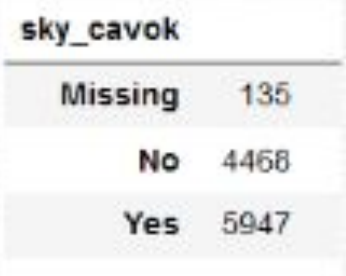
Missing/duplicate Value



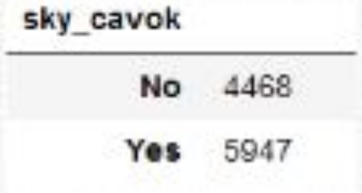
	station_country
count	10000
unique	1
top	SA
freq	10000



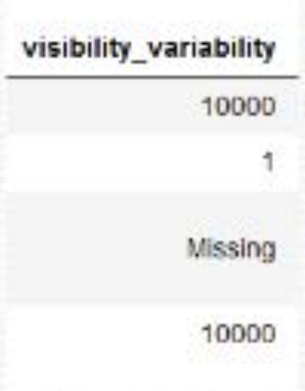
```
dataset = dataset.drop('station_country', axis=1)
```



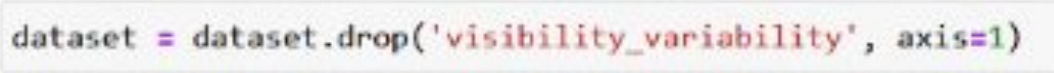
sky_cavok	
Missing	135
No	4468
Yes	5947



sky_cavok	
No	4468
Yes	5947

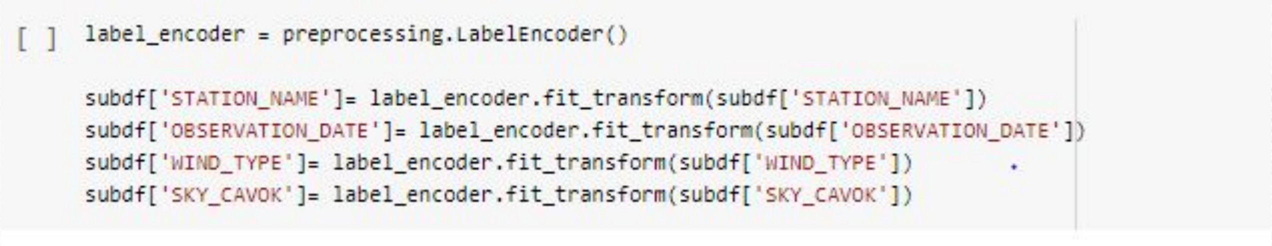


visibility_variability
10000
1
Missing
10000

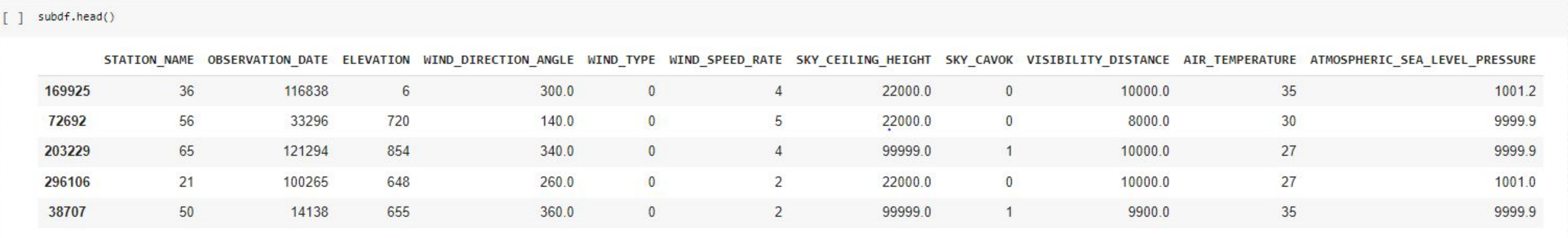


```
dataset = dataset.drop('visibility_variability', axis=1)
```

Feature Engineering-Encoding



```
[ ]  label_encoder = preprocessing.LabelEncoder()

     subdf['STATION_NAME']= label_encoder.fit_transform(subdf['STATION_NAME'])
     subdf['OBSERVATION_DATE']= label_encoder.fit_transform(subdf['OBSERVATION_DATE'])
     subdf['WIND_TYPE']= label_encoder.fit_transform(subdf['WIND_TYPE'])
     subdf['SKY_CAVOK']= label_encoder.fit_transform(subdf['SKY_CAVOK'])
```

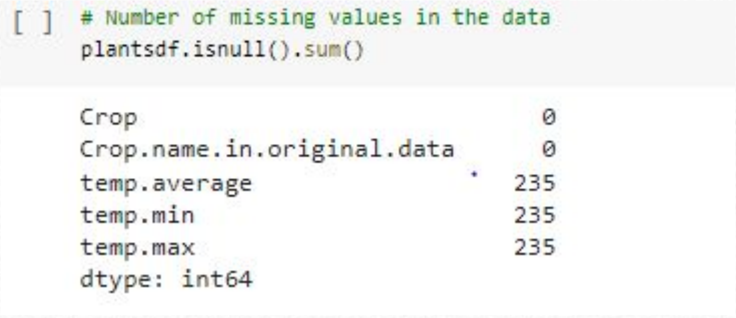


```
[ ]  subdf.head()
```

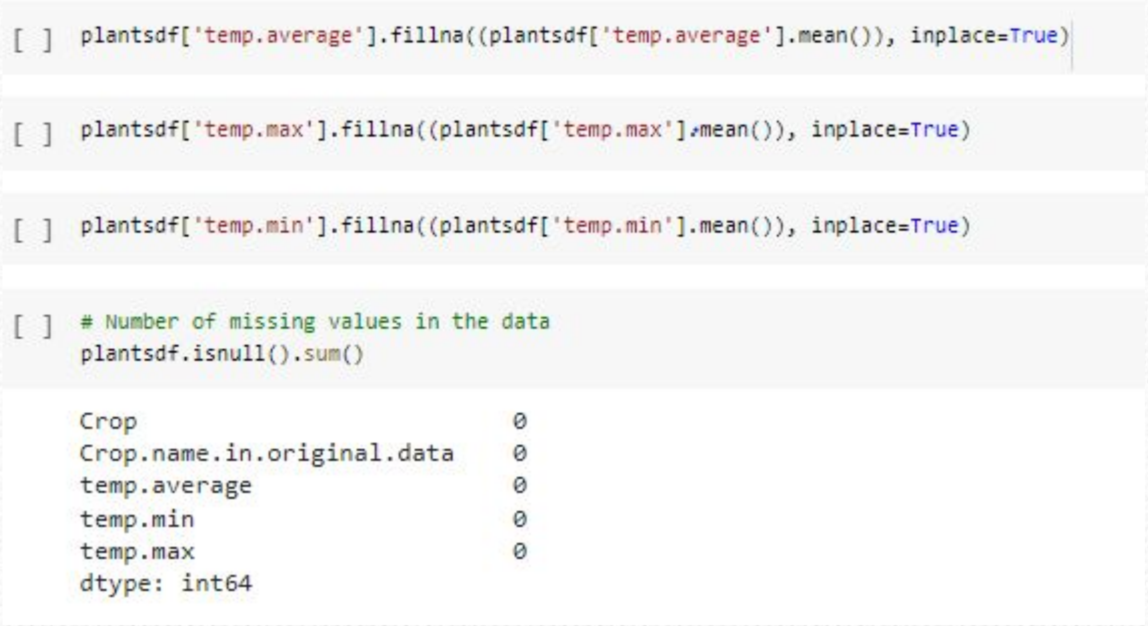
	STATION_NAME	OBSERVATION_DATE	ELEVATION	WIND_DIRECTION_ANGLE	WIND_TYPE	WIND_SPEED_RATE	SKY_CEILING_HEIGHT	SKY_CAVOK	VISIBILITY_DISTANCE	AIR_TEMPERATURE	ATMOSPHERIC_SEA_LEVEL_PRESSURE
169925	36	116838	6	300.0	0	4	22000.0	0	10000.0	35	1001.2
72692	56	33296	720	140.0	0	5	22000.0	0	8000.0	30	9999.9
203229	65	121294	854	340.0	0	4	99999.0	1	10000.0	27	9999.9
296106	21	100265	648	260.0	0	2	22000.0	0	10000.0	27	1001.0
38707	50	14138	655	360.0	0	2	99999.0	1	9900.0	35	9999.9

Missing Value

- Filling the Null value with the mean

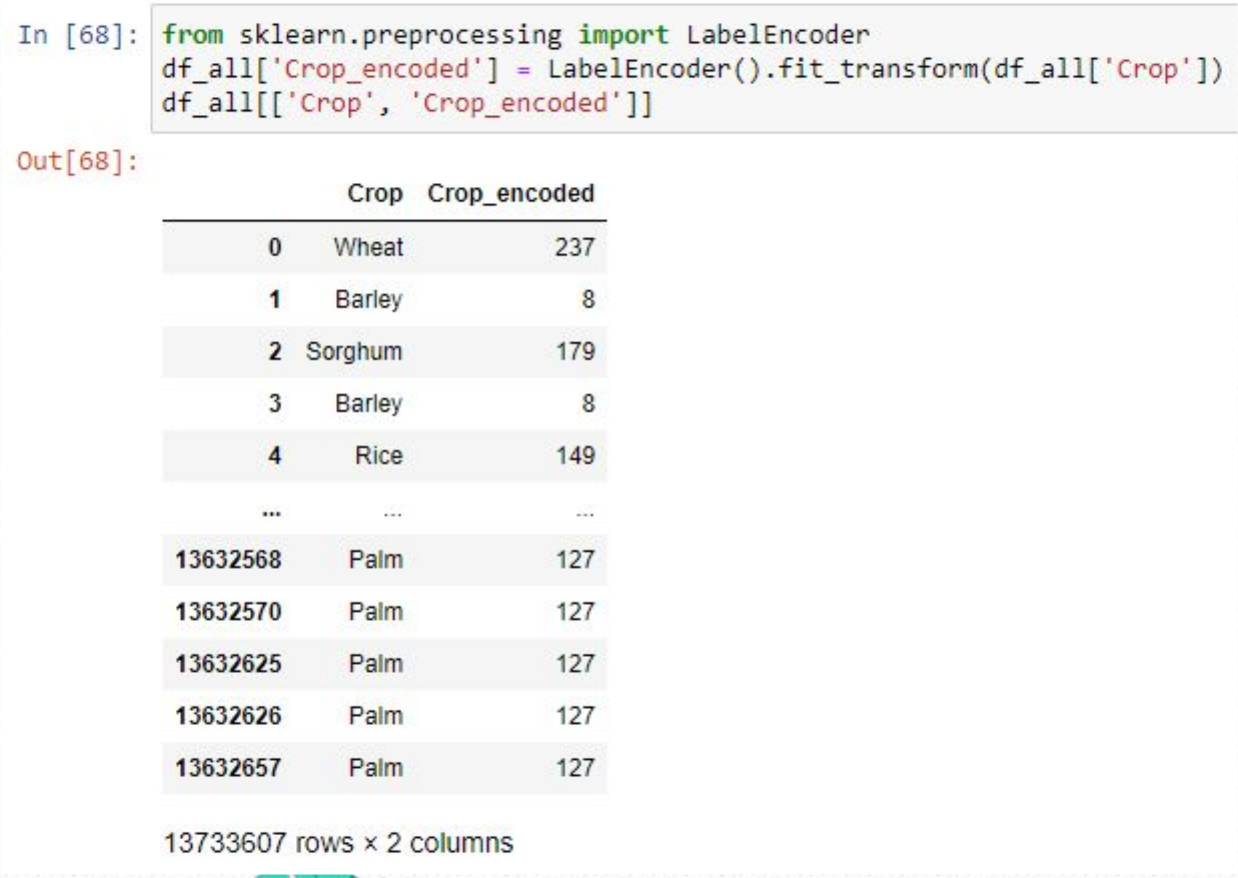


```
[ ]  # Number of missing values in the data
     plantsdf.isnull().sum()

     Crop                            0
     Crop.name.in.original.data      0
     temp.average                  235
     temp.min                      235
     temp.max                      235
     dtype: int64
```



```
[ ]  plantsdf['temp.average'].fillna((plantsdf['temp.average'].mean()), inplace=True)

[ ]  plantsdf['temp.max'].fillna((plantsdf['temp.max'].mean()), inplace=True)

[ ]  plantsdf['temp.min'].fillna((plantsdf['temp.min'].mean()), inplace=True)

[ ]  # Number of missing values in the data
     plantsdf.isnull().sum()

     Crop                          0
     Crop.name.in.original.data    0
     temp.average                  0
     temp.min                      0
     temp.max                      0
     dtype: int64
```

Feature Engineering-Encoding



In [68]:

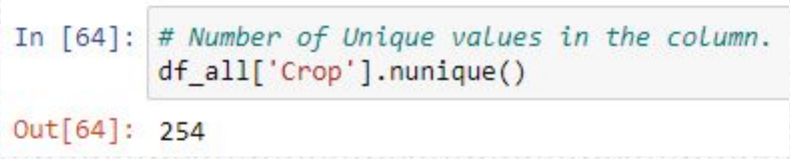
```
from sklearn.preprocessing import LabelEncoder
df_all['Crop_encoded'] = LabelEncoder().fit_transform(df_all['Crop'])
df_all[['Crop', 'Crop_encoded']]
```

Out[68]:

	Crop	Crop_encoded
0	Wheat	237
1	Barley	8
2	Sorghum	179
3	Barley	8
4	Rice	149
...	...	...
13632568	Palm	127
13632570	Palm	127
13632625	Palm	127
13632626	Palm	127
13632657	Palm	127

13733607 rows × 2 columns

- Perform Label Encoder for [Crop] column there is 250 unique value.



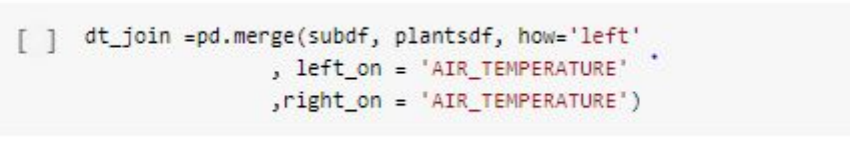
In [64]:

```
# Number of Unique values in the column.
df_all['Crop'].nunique()
```

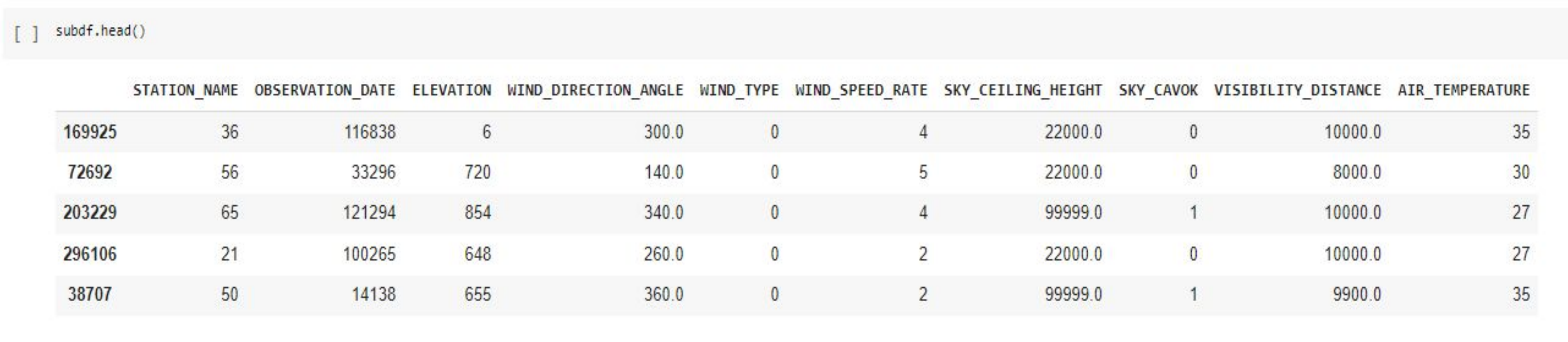
Out[64]:

```
254
```

Join two dataset

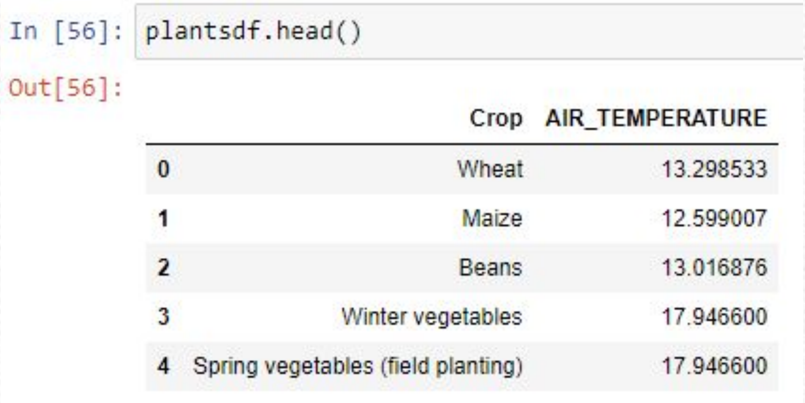


```
[ ]  dt_join =pd.merge(subdf, plantsdf, how='left'
                       , left_on = 'AIR_TEMPERATURE'
                       ,right_on = 'AIR_TEMPERATURE')
```



```
[ ]  subdf.head()
```

	STATION_NAME	OBSERVATION_DATE	ELEVATION	WIND_DIRECTION_ANGLE	WIND_TYPE	WIND_SPEED_RATE	SKY_CEILING_HEIGHT	SKY_CAVOK	VISIBILITY_DISTANCE	AIR_TEMPERATURE
169925	36	116838	6	300.0	0	4	22000.0	0	10000.0	35
72692	56	33296	720	140.0	0	5	22000.0	0	8000.0	30
203229	65	121294	854	340.0	0	4	99999.0	1	10000.0	27
296106	21	100265	648	260.0	0	2	22000.0	0	10000.0	27
38707	50	14138	655	360.0	0	2	99999.0	1	9900.0	35

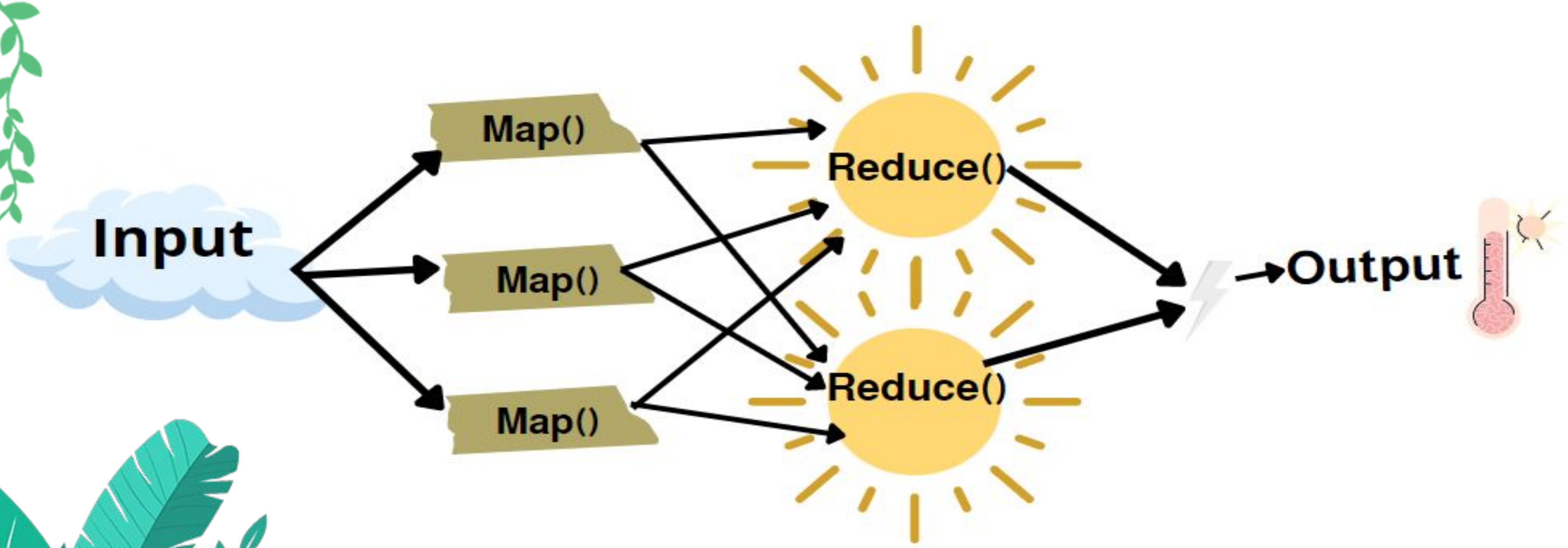


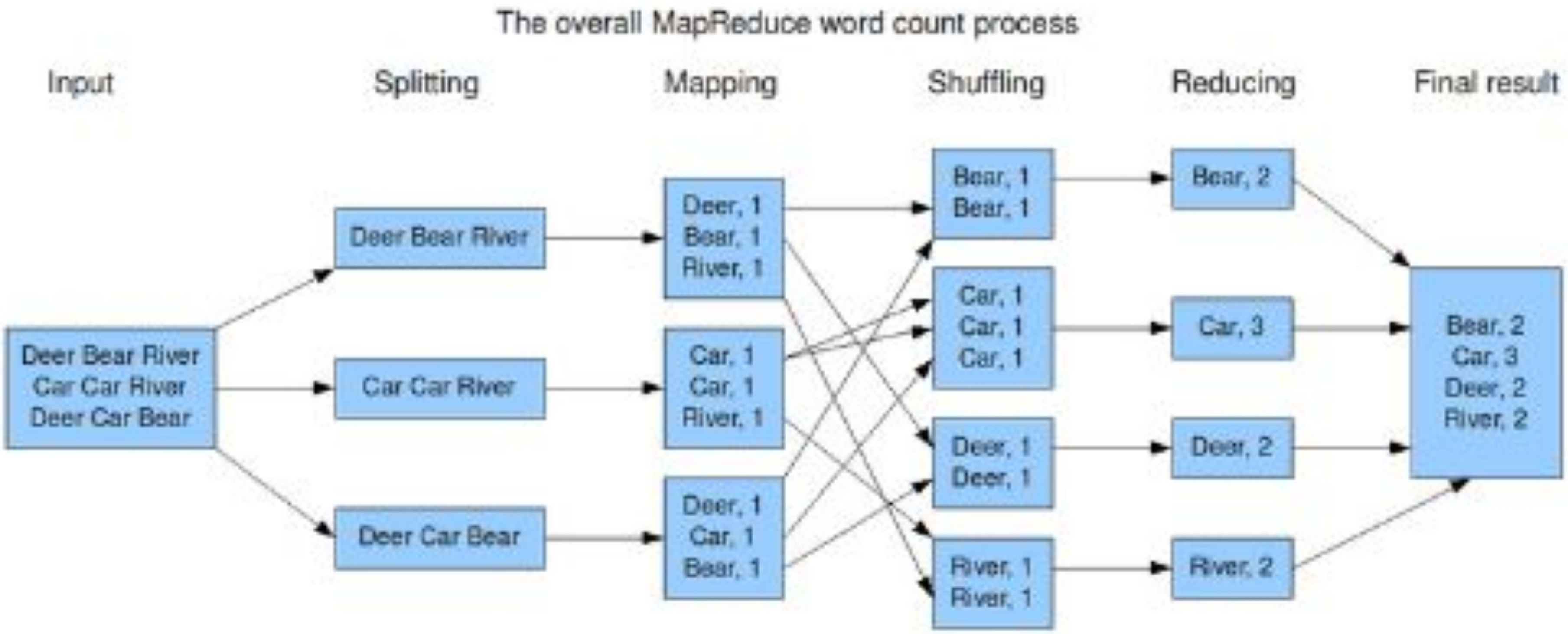
```
In [56]: plantsdf.head()
Out[56]:
```

	Crop	AIR_TEMPERATURE
0	Wheat	13.298533
1	Maize	12.599007
2	Beans	13.016876
3	Winter vegetables	17.946600
4	Spring vegetables (field planting)	17.946600

Mapreducer

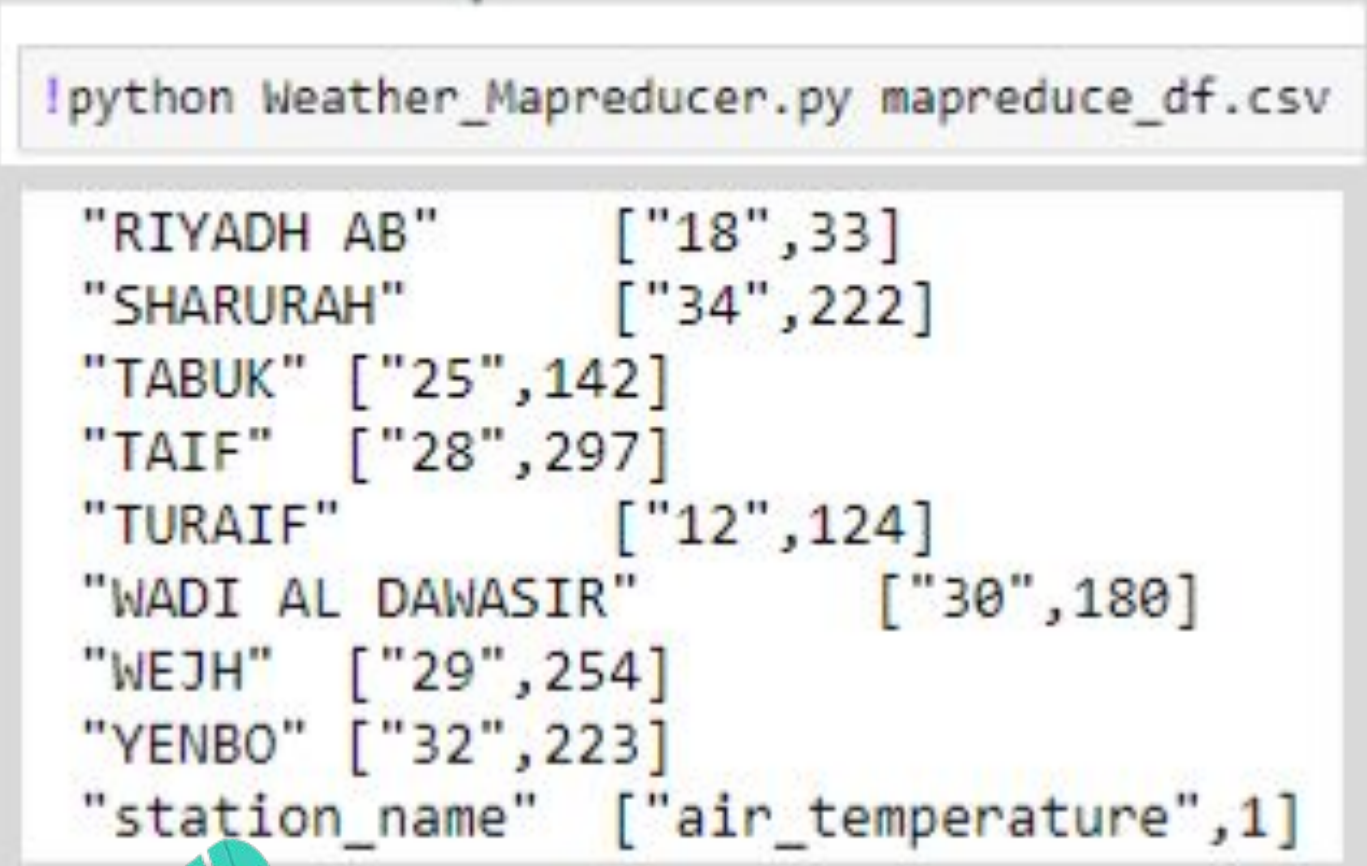
Mapreducer




The overall MapReduce word count process
Input
Splitting
Mapping
Shuffling
Reducing
Final result
Deer Bear River
Car Car River
Deer Car Bear
Deer Bear River
Car Car River
Deer Car Bear
Deer, 1
Bear, 1
River, 1
Car, 1
Car, 1
River, 1
Deer, 1
Car, 1
Bear, 1
Bear, 1
Bear, 1
Car, 1
Car, 1
Car, 1
Deer, 1
Deer, 1
River, 1
River, 1
Bear, 2
Car, 3
Deer, 2
River, 2
Bear, 2
Car, 3
Deer, 2
River, 2

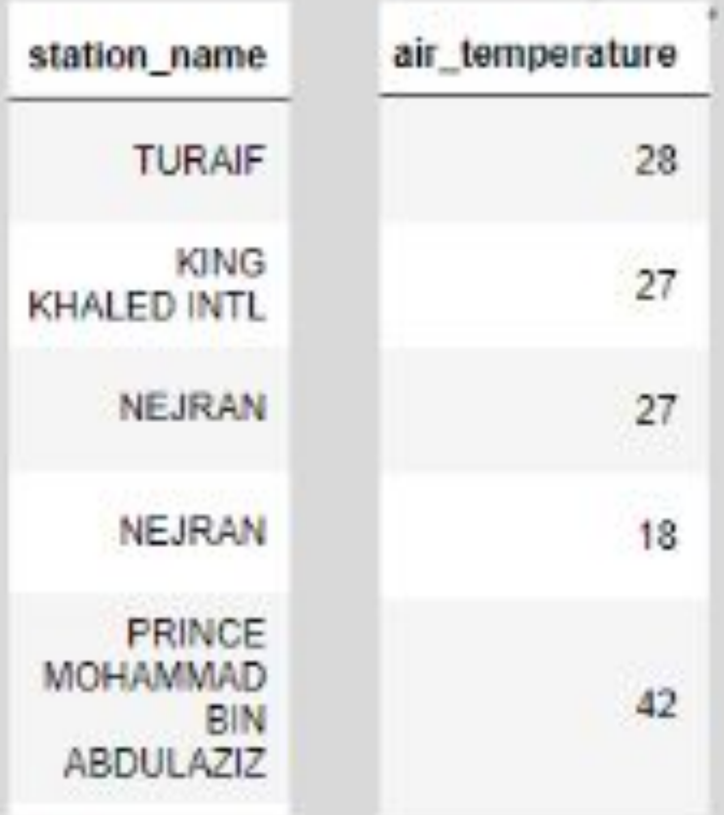
MAP

(station_name,air_temperature)

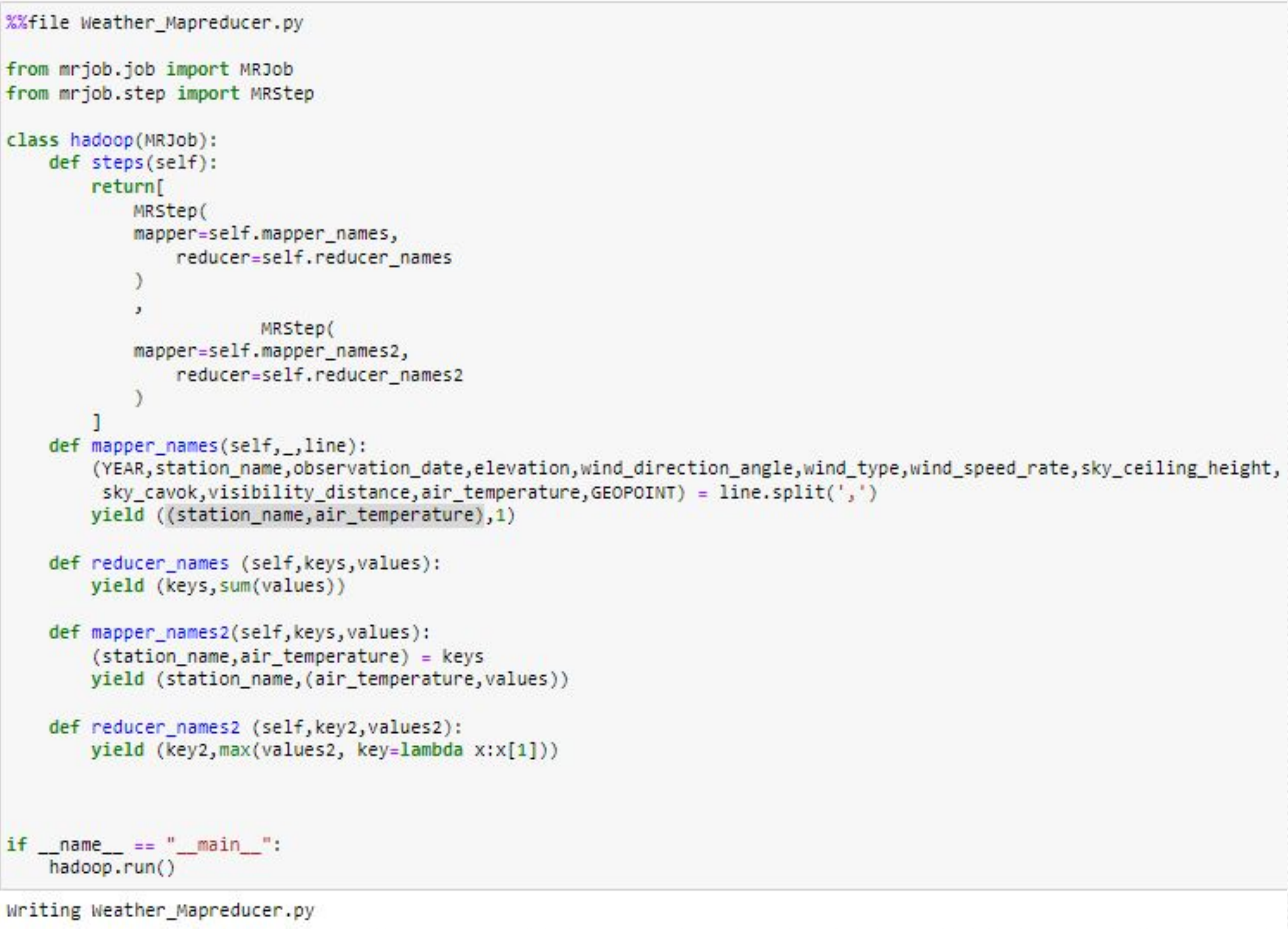


```
!python Weather_Mapreducer.py mapreduce_df.csv
```

```
"RIYADH AB"	["18",33]
"SHARURAH"	["34",222]
"TABUK"	["25",142]
"TAIF"	["28",297]
"TURAIF"	["12",124]
"WADI AL DAWASIR"	["30",180]
"WEJH"	["29",254]
"YENBO"	["32",223]
"station_name"	["air_temperature",1]
```

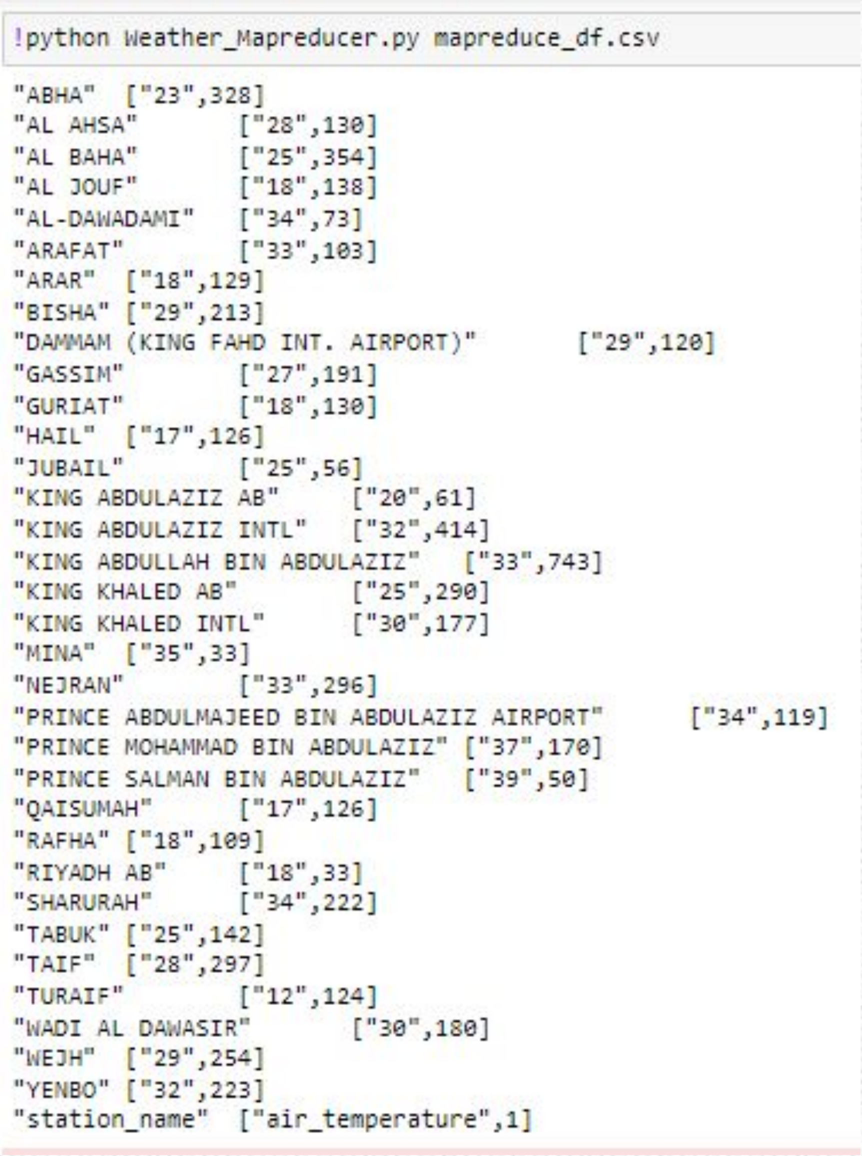


station_name	air_temperature
TURAIF	28
KING KHALED INTL	27
NEJRAN	27
NEJRAN	18
PRINCE MOHAMMAD BIN ABDULAZIZ	42



```
%%file Weather_Mapreducer.py

from mrjob.job import MRJob
from mrjob.step import MRStep

class hadoop(MRJob):
    def steps(self):
        return[
            MRStep(
            mapper=self.mapper_names,
                reducer=self.reducer_names
            )
            ,
                        MRStep(
            mapper=self.mapper_names2,
                reducer=self.reducer_names2
            )
        ]
    def mapper_names(self,_,line):
        (YEAR,station_name,observation_date,elevation,wind_direction_angle,wind_type,wind_speed_rate,sky_ceiling_height,
         sky_cavok,visibility_distance,air_temperature,GEOPOINT) = line.split(',')
        yield ((station_name,air_temperature),1)

    def reducer_names (self,keys,values):
        yield (keys,sum(values))

    def mapper_names2(self,keys,values):
        (station_name,air_temperature) = keys
        yield (station_name,(air_temperature,values))

    def reducer_names2 (self,key2,values2):
        yield (key2,max(values2, key=lambda x:x[1]))



if __name__ == "__main__":
    hadoop.run()
```

```
Writing Weather_Mapreducer.py
```

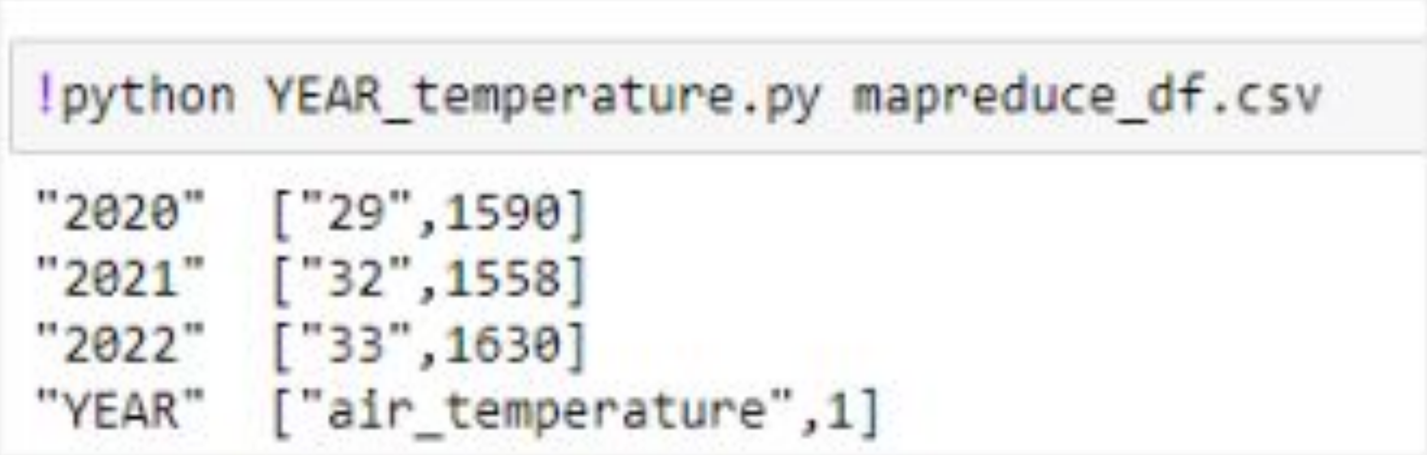


```
!python Weather_Mapreducer.py mapreduce_df.csv
```

```
"ABHA"  ["23",328]
"AL AHSA"       ["28",130]
"AL BAHA"       ["25",354]
"AL JOUF"       ["18",138]
"AL-DAWADAMI"   ["34",73]
"ARAFAT"        ["33",103]
"ARAR"  ["18",129]
"BISHA" ["29",213]
"DAMMAM (KING FAHD INT. AIRPORT)"       ["29",120]
"GASSIM"        ["27",191]
"GURIAT"        ["18",130]
"HAIL"  ["17",126]
"JUBAIL"        ["25",56]
"KING ABDULAZIZ AB"     ["20",61]
"KING ABDULAZIZ INTL"   ["32",414]
"KING ABDULLAH BIN ABDULAZIZ"   ["33",743]
"KING KHALED AB"        ["25",290]
"KING KHALED INTL"      ["30",177]
"MINA"  ["35",33]
"NEJRAN"        ["33",296]
"PRINCE ABDULMAJEED BIN ABDULAZIZ AIRPORT"      ["34",119]
"PRINCE MOHAMMAD BIN ABDULAZIZ" ["37",170]
"PRINCE SALMAN BIN ABDULAZIZ"   ["39",50]
"QAISUMAH"      ["17",126]
"RAFHA" ["18",109]
"RIYADH AB"     ["18",33]
"SHARURAH"      ["34",222]
"TABUK" ["25",142]
"TAIF"  ["28",297]
"TURAIF"        ["12",124]
"WADI AL DAWASIR"       ["30",180]
"WEJH"  ["29",254]
"YENBO" ["32",223]
"station_name"  ["air_temperature",1]
```

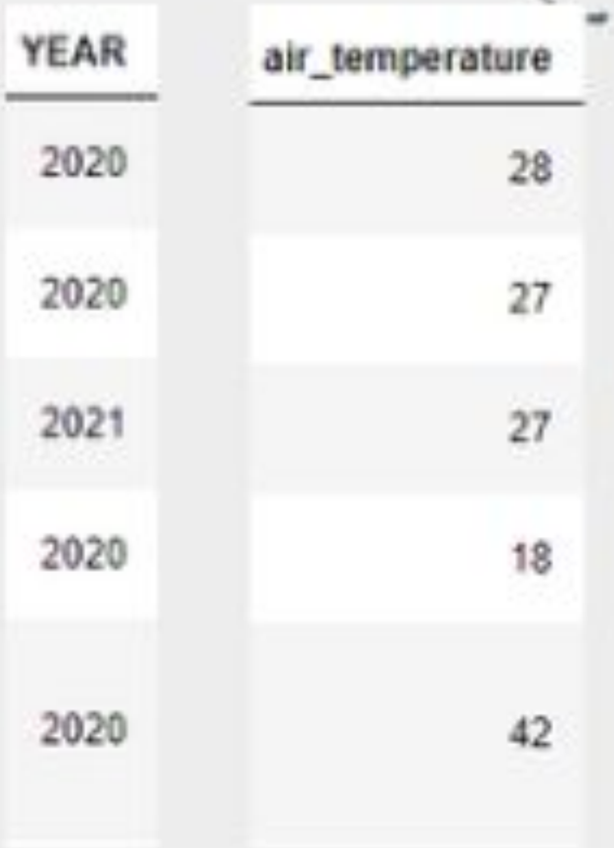
MAP

(YEAR,air_temperature)

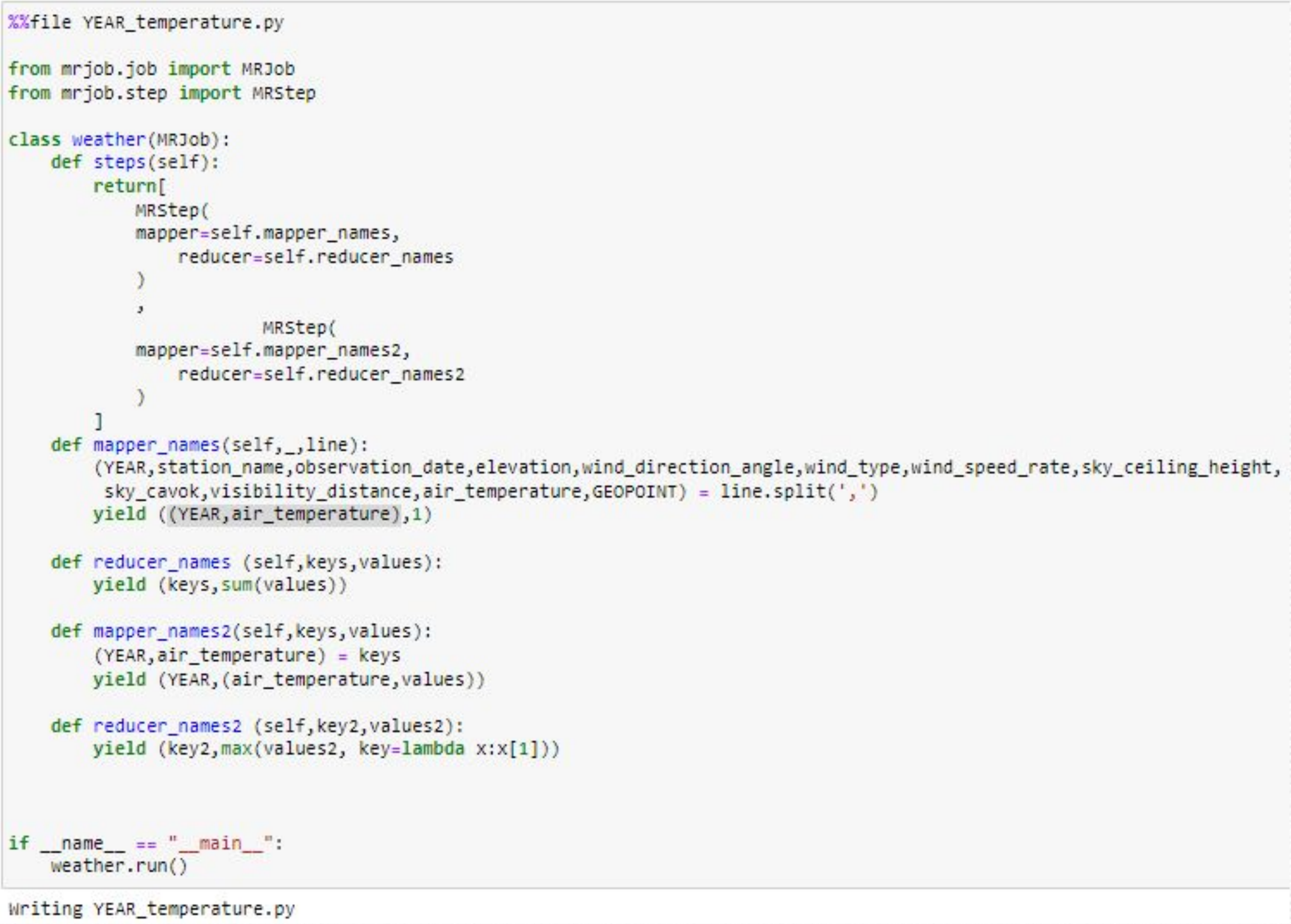


```
!python YEAR_temperature.py mapreduce_df.csv
```

```
"2020"   ["29",1590]
"2021"   ["32",1558]
"2022"   ["33",1630]
"YEAR"   ["air_temperature",1]
```

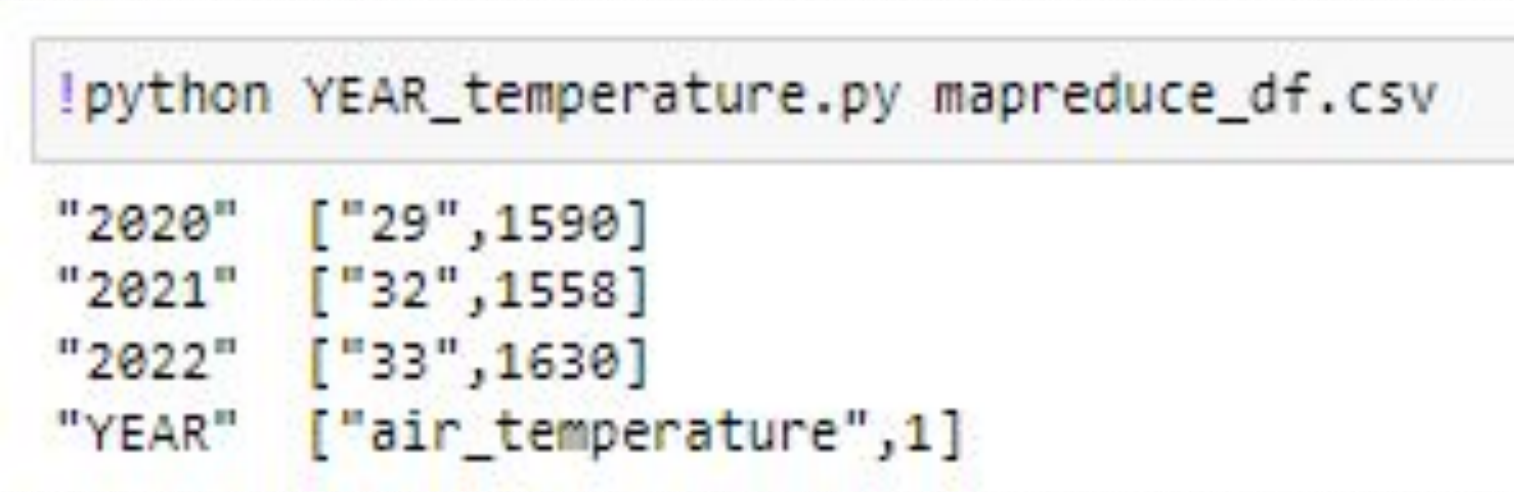


YEAR	air_temperature
2020	28
2020	27
2021	27
2020	18
2020	42



```
%%file YEAR_temperature.py

from mrjob.job import MRJob
from mrjob.step import MRStep

class weather(MRJob):
    def steps(self):
        return[
            MRStep(
            mapper=self.mapper_names,
                reducer=self.reducer_names
            )
            ,
                        MRStep(
            mapper=self.mapper_names2,
                reducer=self.reducer_names2
            )
        ]
    def mapper_names(self,_,line):
        (YEAR,station_name,observation_date,elevation,wind_direction_angle,wind_type,wind_speed_rate,sky_ceiling_height,
         sky_cavok,visibility_distance,air_temperature,GEOPOINT) = line.split(',')
        yield ((YEAR,air_temperature),1)

    def reducer_names (self,keys,values):
        yield (keys,sum(values))

    def mapper_names2(self,keys,values):
        (YEAR,air_temperature) = keys
        yield (YEAR,(air_temperature,values))

    def reducer_names2 (self,key2,values2):
        yield (key2,max(values2, key=lambda x:x[1]))



if __name__ == "__main__":
    weather.run()
```

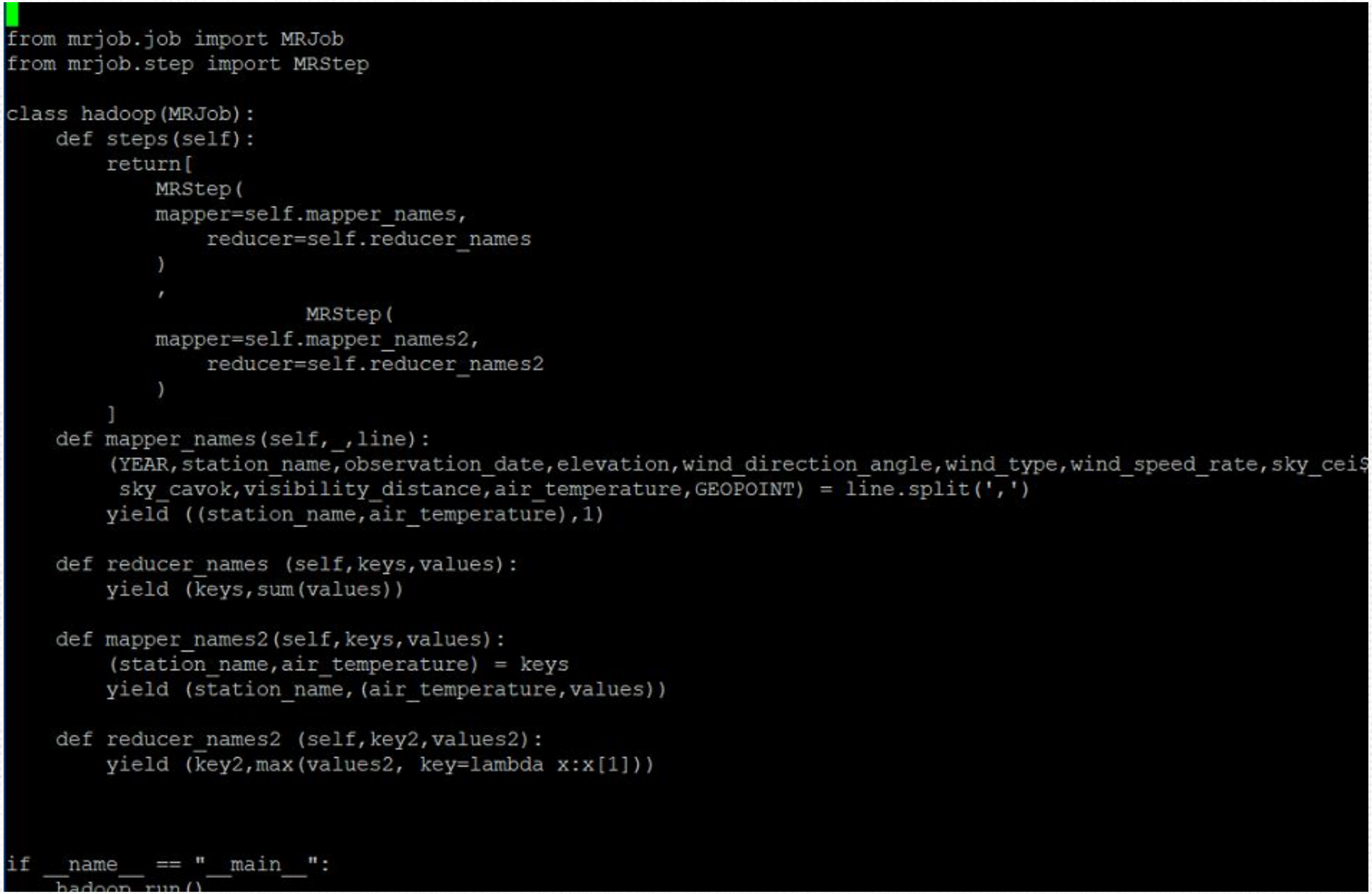
```
Writing YEAR_temperature.py
```



```
!python YEAR_temperature.py mapreduce_df.csv
```

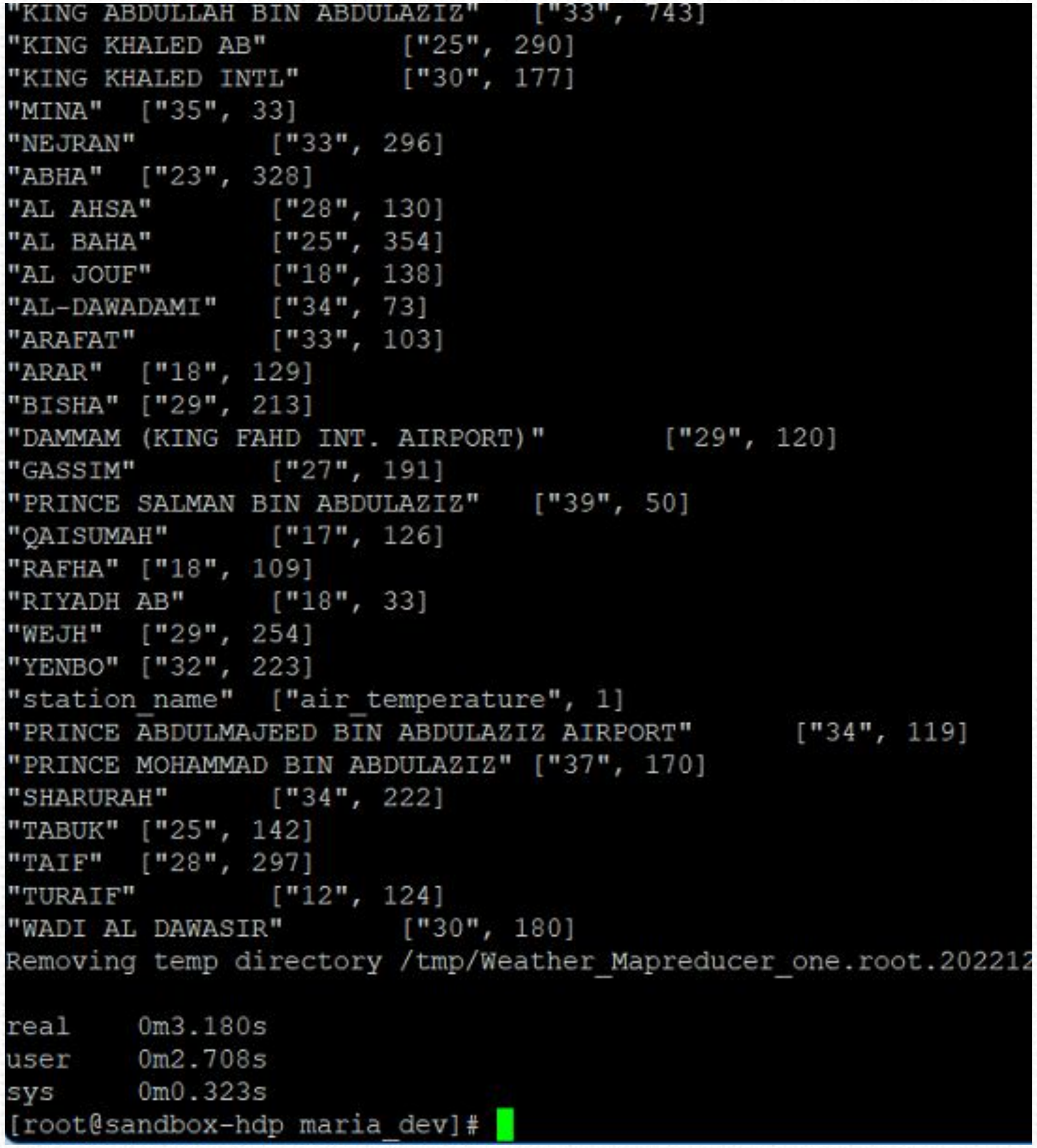
```
"2020"  ["29",1590]
"2021"  ["32",1558]
"2022"  ["33",1630]
"YEAR"  ["air_temperature",1]
```

PuTTY sandbox locally & Hadoop



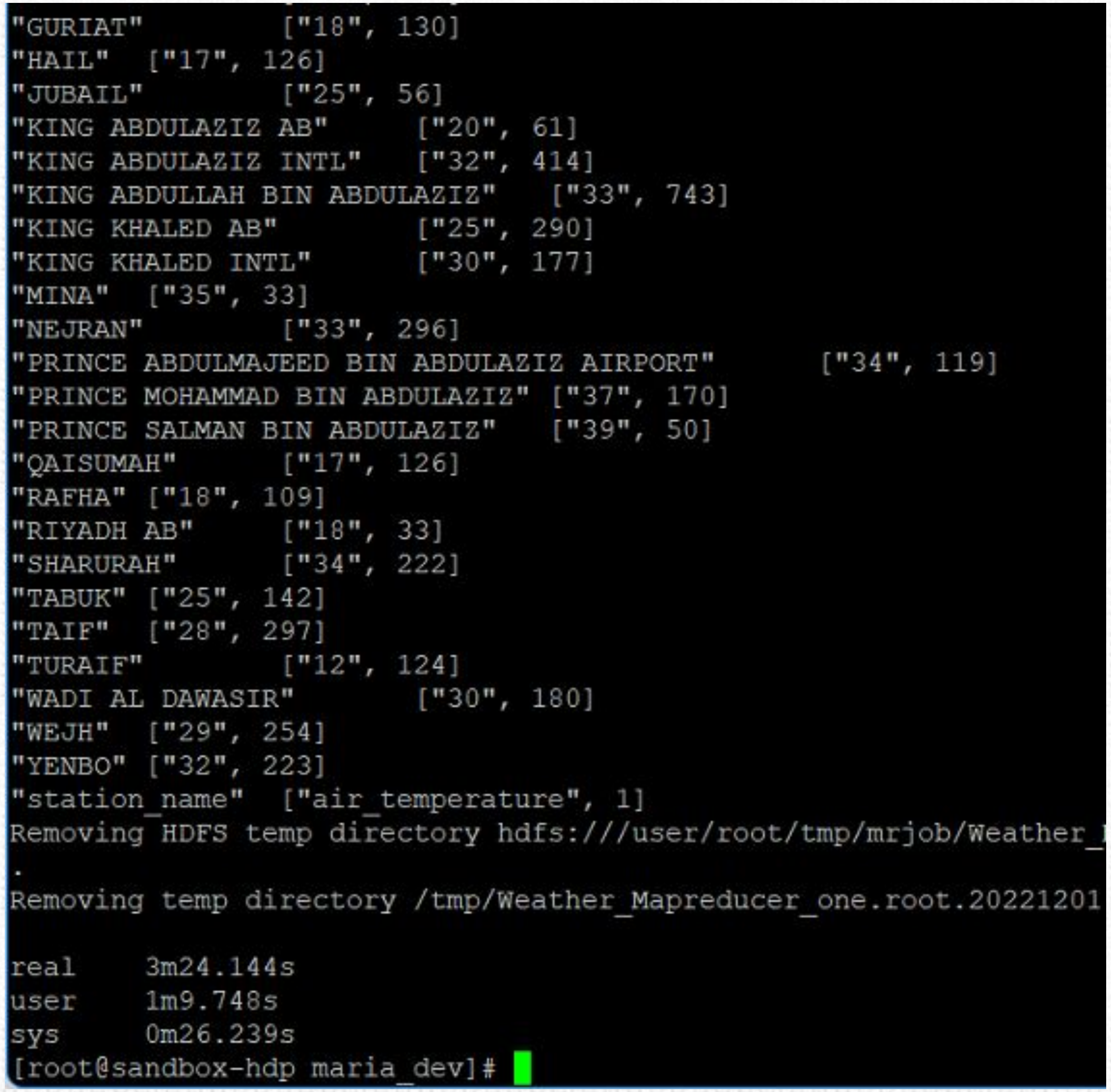
```
from mrjob.job import MRJob
from mrjob.step import MRStep

class hadoop(MRJob):
    def steps(self):
        return[
            MRStep(
            mapper=self.mapper_names,
                reducer=self.reducer_names
            )
            ,
                        MRStep(
            mapper=self.mapper_names2,
                reducer=self.reducer_names2
            )
        ]
    def mapper_names(self,_,line):
        (YEAR,station_name,observation_date,elevation,wind_direction_angle,wind_type,wind_speed_rate,sky_cei$
         sky_cavok,visibility_distance,air_temperature,GEOPOINT) = line.split(',')
        yield ((station_name,air_temperature),1)

    def reducer_names (self,keys,values):
        yield (keys,sum(values))

    def mapper_names2(self,keys,values):
        (station_name,air_temperature) = keys
        yield (station_name,(air_temperature,values))

    def reducer_names2 (self,key2,values2):
        yield (key2,max(values2, key=lambda x:x[1]))



if __name__ == "__main__":
    hadoop.run()
```

Locally

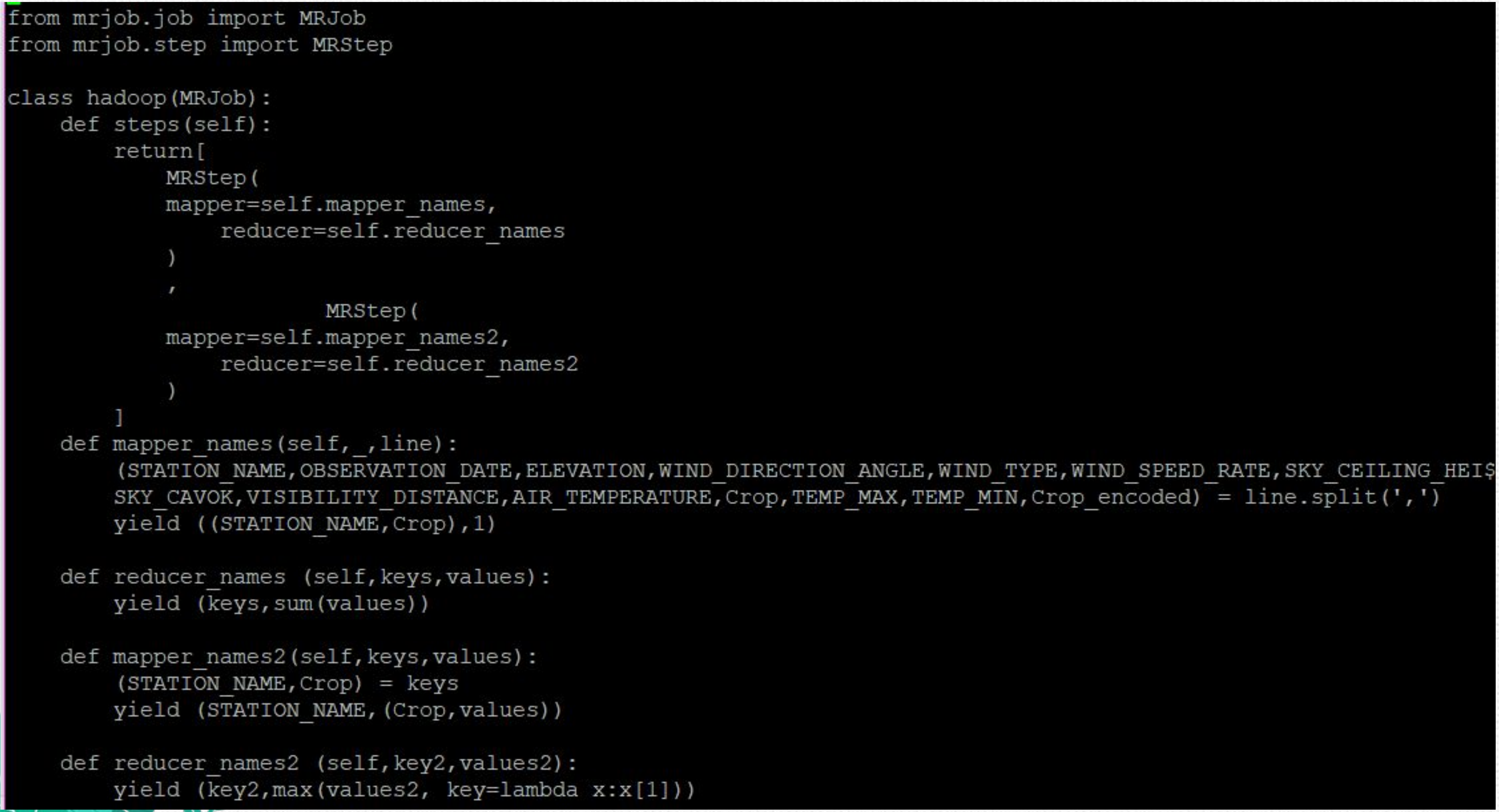


```
"KING ABDULLAH BIN ABDULAZIZ"   ["33", 743]
"KING KHALED AB"        ["25", 290]
"KING KHALED INTL"      ["30", 177]
"MINA"  ["35", 33]
"NEJRAN"        ["33", 296]
"ABHA"  ["23", 328]
"AL AHSA"       ["28", 130]
"AL BAHA"       ["25", 354]
"AL JOUF"       ["18", 138]
"AL-DAWADAMI"   ["34", 73]
"ARAFAT"        ["33", 103]
"ARAR"  ["18", 129]
"BISHA" ["29", 213]
"DAMMAM (KING FAHD INT. AIRPORT)"       ["29", 120]
"GASSIM"        ["27", 191]
"PRINCE SALMAN BIN ABDULAZIZ"   ["39", 50]
"QAISUMAH"      ["17", 126]
"RAFHA" ["18", 109]
"RIYADH AB"     ["18", 33]
"WEJH"  ["29", 254]
"YENBO" ["32", 223]
"station_name"  ["air_temperature", 1]
"PRINCE ABDULMAJEED BIN ABDULAZIZ AIRPORT"      ["34", 119]
"PRINCE MOHAMMAD BIN ABDULAZIZ" ["37", 170]
"SHARURAH"      ["34", 222]
"TABUK" ["25", 142]
"TAIF"  ["28", 297]
"TURAIF"        ["12", 124]
"WADI AL DAWASIR"       ["30", 180]
Removing temp directory /tmp/Weather_Mapreducer_one.root.202212

real    0m3.180s
user    0m2.708s
sys     0m0.323s
[root@sandbox-hdp maria_dev]#
```

Hadoop

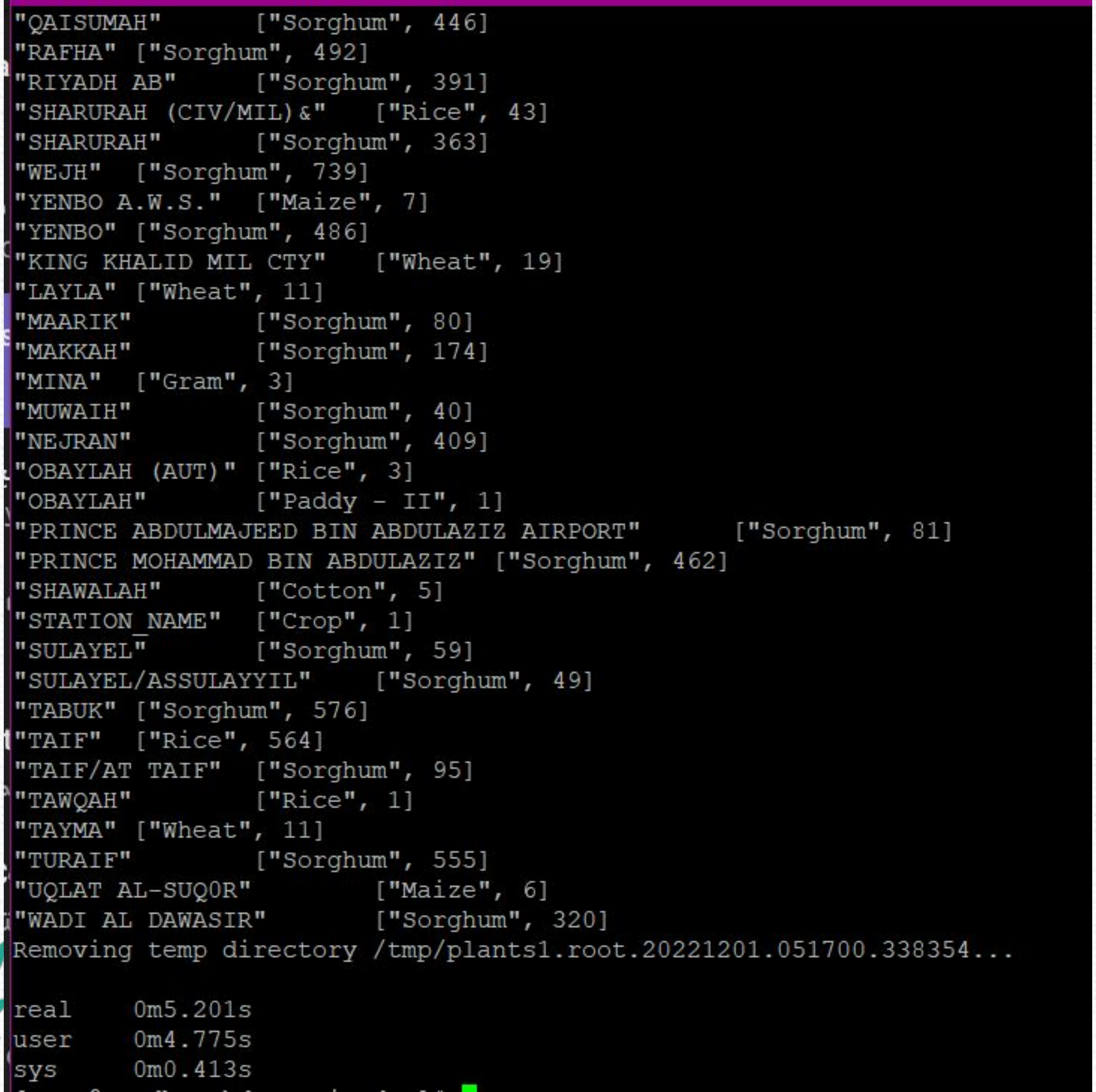


```
"GURIAT"        ["18", 130]
"HAIL"  ["17", 126]
"JUBAIL"        ["25", 56]
"KING ABDULAZIZ AB"     ["20", 61]
"KING ABDULAZIZ INTL"   ["32", 414]
"KING ABDULLAH BIN ABDULAZIZ"   ["33", 743]
"KING KHALED AB"        ["25", 290]
"KING KHALED INTL"      ["30", 177]
"MINA"  ["35", 33]
"NEJRAN"        ["33", 296]
"PRINCE ABDULMAJEED BIN ABDULAZIZ AIRPORT"      ["34", 119]
"PRINCE MOHAMMAD BIN ABDULAZIZ" ["37", 170]
"PRINCE SALMAN BIN ABDULAZIZ"   ["39", 50]
"QAISUMAH"      ["17", 126]
"RAFHA" ["18", 109]
"RIYADH AB"     ["18", 33]
"SHARURAH"      ["34", 222]
"TABUK" ["25", 142]
"TAIF"  ["28", 297]
"TURAIF"        ["12", 124]
"WADI AL DAWASIR"       ["30", 180]
"WEJH"  ["29", 254]
"YENBO" ["32", 223]
"station_name"  ["air_temperature", 1]
Removing HDFS temp directory hdfs:///user/root/tmp/mrjob/Weather_
.
Removing temp directory /tmp/Weather_Mapreducer_one.root.20221201

real    3m24.144s
user    1m9.748s
sys     0m26.239s
[root@sandbox-hdp maria_dev]#
```



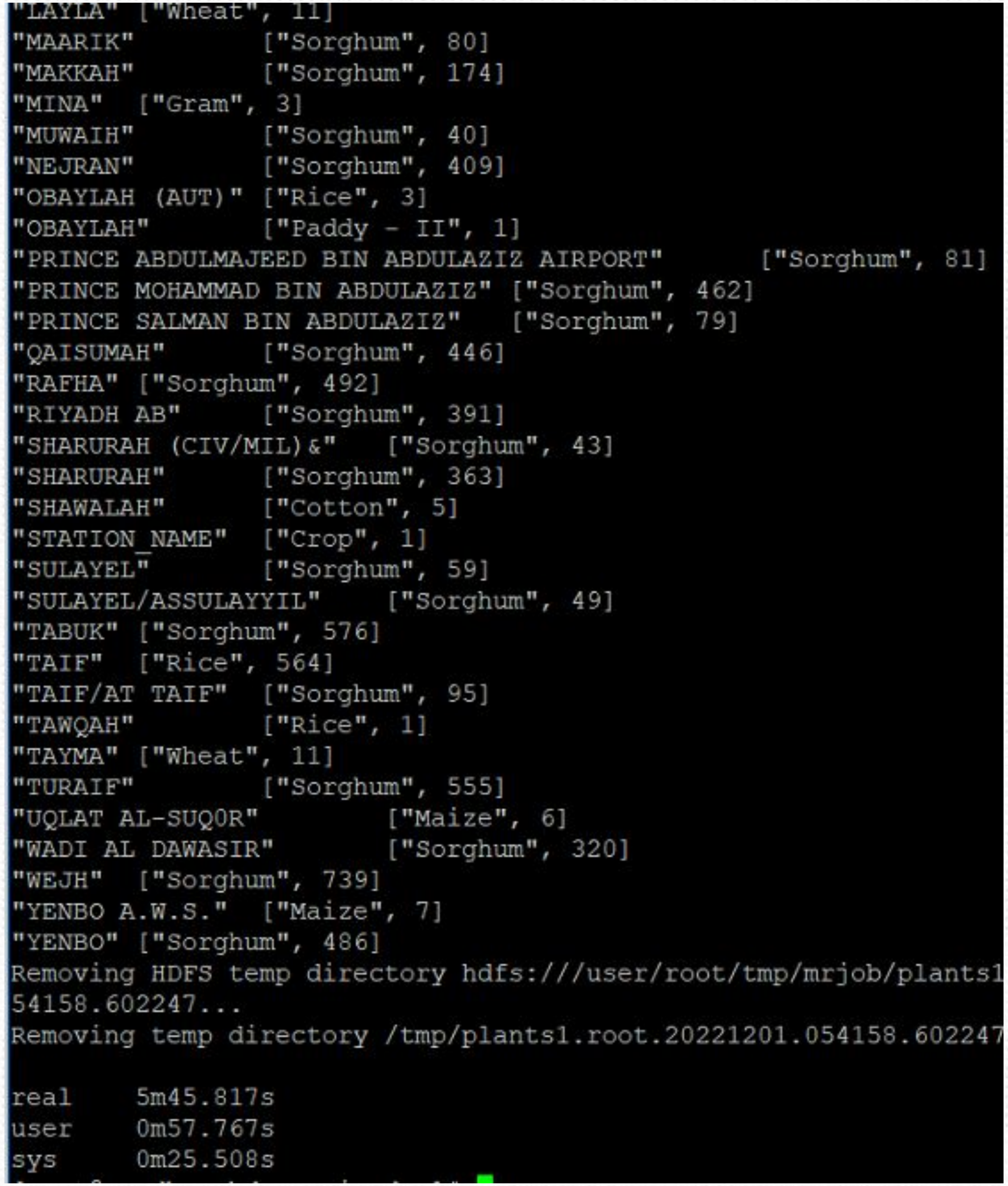
```
from mrjob.job import MRJob
from mrjob.step import MRStep

class hadoop(MRJob):
    def steps(self):
        return[
            MRStep(
            mapper=self.mapper_names,
                reducer=self.reducer_names
            )
            ,
                        MRStep(
            mapper=self.mapper_names2,
                reducer=self.reducer_names2
            )
        ]
    def mapper_names(self,_,line):
        (STATION_NAME,OBSERVATION_DATE,ELEVATION,WIND_DIRECTION_ANGLE,WIND_TYPE,WIND_SPEED_RATE,SKY_CEILING_HEI$
        SKY_CAVOK,VISIBILITY_DISTANCE,AIR_TEMPERATURE,Crop,TEMP_MAX,TEMP_MIN,Crop_encoded) = line.split(',')
        yield ((STATION_NAME,Crop),1)

    def reducer_names (self,keys,values):
        yield (keys,sum(values))

    def mapper_names2(self,keys,values):
        (STATION_NAME,Crop) = keys
        yield (STATION_NAME,(Crop,values))

    def reducer_names2 (self,key2,values2):
        yield (key2,max(values2, key=lambda x:x[1]))
```

Locally



```
"QAISUMAH"      ["Sorghum", 446]
"RAFHA" ["Sorghum", 492]
"RIYADH AB"     ["Sorghum", 391]
"SHARURAH (CIV/MIL)&"   ["Rice", 43]
"SHARURAH"      ["Sorghum", 363]
"WEJH"  ["Sorghum", 739]
"YENBO A.W.S."  ["Maize", 7]
"YENBO" ["Sorghum", 486]
"KING KHALID MIL CTY"   ["Wheat", 19]
"LAYLA" ["Wheat", 11]
"MAARIK"        ["Sorghum", 80]
"MAKKAH"        ["Sorghum", 174]
"MINA"  ["Gram", 3]
"MUWAIH"        ["Sorghum", 40]
"NEJRAN"        ["Sorghum", 409]
"OBAYLAH (AUT)" ["Rice", 3]
"OBAYLAH"       ["Paddy - II", 1]
"PRINCE ABDULMAJEED BIN ABDULAZIZ AIRPORT"      ["Sorghum", 81]
"PRINCE MOHAMMAD BIN ABDULAZIZ" ["Sorghum", 462]
"SHAWALAH"      ["Cotton", 5]
"STATION_NAME"  ["Crop", 1]
"SULAYEL"       ["Sorghum", 59]
"SULAYEL/ASSULAYYIL"    ["Sorghum", 49]
"TABUK" ["Sorghum", 576]
"TAIF"  ["Rice", 564]
"TAIF/AT TAIF"  ["Sorghum", 95]
"TAWQAH"        ["Rice", 1]
"TAYMA" ["Wheat", 11]
"TURAIF"        ["Sorghum", 555]
"UQLAT AL-SUQOR"        ["Maize", 6]
"WADI AL DAWASIR"       ["Sorghum", 320]
Removing temp directory /tmp/plants1.root.20221201.051700.338354...

real    0m5.201s
user    0m4.775s
sys     0m0.413s
```

Hadoop



```
"LAYLA" ["Wheat", 11]
"MAARIK"        ["Sorghum", 80]
"MAKKAH"        ["Sorghum", 174]
"MINA"  ["Gram", 3]
"MUWAIH"        ["Sorghum", 40]
"NEJRAN"        ["Sorghum", 409]
"OBAYLAH (AUT)" ["Rice", 3]
"OBAYLAH"       ["Paddy - II", 1]
"PRINCE ABDULMAJEED BIN ABDULAZIZ AIRPORT"      ["Sorghum", 81]
"PRINCE MOHAMMAD BIN ABDULAZIZ" ["Sorghum", 462]
"PRINCE SALMAN BIN ABDULAZIZ"   ["Sorghum", 79]
"QAISUMAH"      ["Sorghum", 446]
"RAFHA" ["Sorghum", 492]
"RIYADH AB"     ["Sorghum", 391]
"SHARURAH (CIV/MIL)&"   ["Sorghum", 43]
"SHARURAH"      ["Sorghum", 363]
"SHAWALAH"      ["Cotton", 5]
"STATION_NAME"  ["Crop", 1]
"SULAYEL"       ["Sorghum", 59]
"SULAYEL/ASSULAYYIL"    ["Sorghum", 49]
"TABUK" ["Sorghum", 576]
"TAIF"  ["Rice", 564]
"TAIF/AT TAIF"  ["Sorghum", 95]
"TAWQAH"        ["Rice", 1]
"TAYMA" ["Wheat", 11]
"TURAIF"        ["Sorghum", 555]
"UQLAT AL-SUQOR"        ["Maize", 6]
"WADI AL DAWASIR"       ["Sorghum", 320]
"WEJH"  ["Sorghum", 739]
"YENBO A.W.S."  ["Maize", 7]
"YENBO" ["Sorghum", 486]
Removing HDFS temp directory hdfs:///user/root/tmp/mrjob/plants1
54158.602247...
Removing temp directory /tmp/plants1.root.20221201.054158.602247

real    5m45.817s
user    0m57.767s
sys     0m25.508s
```

Thanks!

Any Questions?

- Shouq Alharbi
- Razan Alajlan
- Nada Oteif
- Hayam Alrashed
- Sarah Alrashidi