GLRM

Generalized Low Rank Models Of Air Quality Dataset

7107018017 林祐陞

Schema

- 資料狀況確認
- 合併結構相同的年份資料
- na.drop
- 生成missing data for testing
- GLRM
- 應用於原始資料, 補回missing

1

資料狀況確認

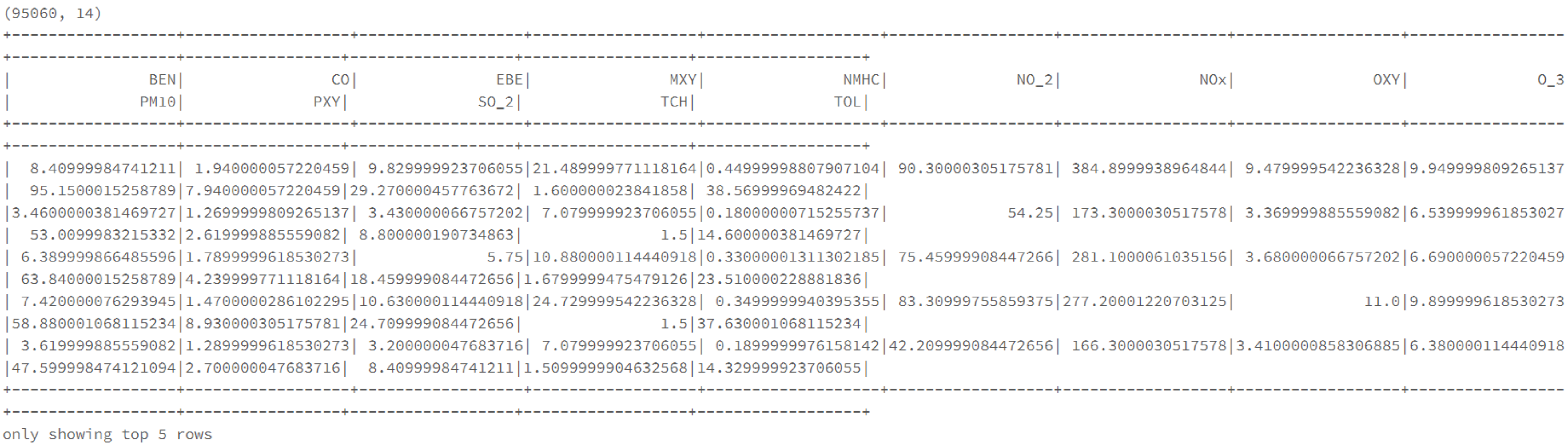
➢ 發現不同年間, 紀錄空汙的資料欄位並不完全相同

```
Years =  2001 Columns =  ['date', 'BEN', 'CO', 'EBE', 'MXY', 'NMHC', 'NO_2', 'NOx', 'OXY', 'O_3', 'PM10', 'PXY', 'SO_2', 'TCH', 'TOL', 'station']
Years =  2002 Columns =  ['date', 'BEN', 'CO', 'EBE', 'MXY', 'NMHC', 'NO_2', 'NOx', 'OXY', 'O_3', 'PM10', 'PXY', 'SO_2', 'TCH', 'TOL', 'station']
Years =  2003 Columns =  ['date', 'BEN', 'CO', 'EBE', 'MXY', 'NMHC', 'NO_2', 'NOx', 'OXY', 'O_3', 'PM10', 'PXY', 'SO_2', 'TCH', 'TOL', 'station']
Years =  2004 Columns =  ['date', 'BEN', 'CO', 'EBE', 'MXY', 'NMHC', 'NO_2', 'NOx', 'OXY', 'O_3', 'PM10', 'PM25', 'PXY', 'SO_2', 'TCH', 'TOL', 'station']
Years =  2005 Columns =  ['date', 'BEN', 'CO', 'EBE', 'MXY', 'NMHC', 'NO_2', 'NOx', 'OXY', 'O_3', 'PM10', 'PM25', 'PXY', 'SO_2', 'TCH', 'TOL', 'station']
Years =  2006 Columns =  ['date', 'BEN', 'CO', 'EBE', 'MXY', 'NMHC', 'NO_2', 'NOx', 'OXY', 'O_3', 'PM10', 'PM25', 'PXY', 'SO_2', 'TCH', 'TOL', 'station']
Years =  2007 Columns =  ['date', 'BEN', 'CO', 'EBE', 'MXY', 'NMHC', 'NO_2', 'NOx', 'OXY', 'O_3', 'PM10', 'PM25', 'PXY', 'SO_2', 'TCH', 'TOL', 'station']
Years =  2008 Columns =  ['date', 'BEN', 'CO', 'EBE', 'MXY', 'NMHC', 'NO_2', 'NOx', 'OXY', 'O_3', 'PM10', 'PM25', 'PXY', 'SO_2', 'TCH', 'TOL', 'station']
Years =  2009 Columns =  ['date', 'BEN', 'CO', 'EBE', 'MXY', 'NMHC', 'NO_2', 'NOx', 'OXY', 'O_3', 'PM10', 'PM25', 'PXY', 'SO_2', 'TCH', 'TOL', 'station']
Years =  2010 Columns =  ['date', 'BEN', 'CO', 'EBE', 'MXY', 'NMHC', 'NO_2', 'NOx', 'OXY', 'O_3', 'PM10', 'PM25', 'PXY', 'SO_2', 'TCH', 'TOL', 'station']
Years =  2011 Columns =  ['date', 'BEN', 'CO', 'EBE', 'NMHC', 'NO', 'NO_2', 'O_3', 'PM10', 'PM25', 'SO_2', 'TCH', 'TOL', 'station']
Years =  2012 Columns =  ['date', 'BEN', 'CO', 'EBE', 'NMHC', 'NO', 'NO_2', 'O_3', 'PM10', 'PM25', 'SO_2', 'TCH', 'TOL', 'station']
Years =  2013 Columns =  ['date', 'BEN', 'CO', 'EBE', 'NMHC', 'NO', 'NO_2', 'O_3', 'PM10', 'PM25', 'SO_2', 'TCH', 'TOL', 'station']
Years =  2014 Columns =  ['date', 'BEN', 'CO', 'EBE', 'NMHC', 'NO', 'NO_2', 'O_3', 'PM10', 'PM25', 'SO_2', 'TCH', 'TOL', 'station']
Years =  2015 Columns =  ['date', 'BEN', 'CO', 'EBE', 'NMHC', 'NO', 'NO_2', 'O_3', 'PM10', 'PM25', 'SO_2', 'TCH', 'TOL', 'station']
Years =  2016 Columns =  ['date', 'BEN', 'CO', 'EBE', 'NMHC', 'NO', 'NO_2', 'O_3', 'PM10', 'PM25', 'SO_2', 'TCH', 'TOL', 'station']
Years =  2017 Columns =  ['date', 'BEN', 'CH4', 'CO', 'EBE', 'NMHC', 'NO', 'NO_2', 'NOx', 'O_3', 'PM10', 'PM25', 'SO_2', 'TCH', 'TOL', 'station']
Years =  2018 Columns =  ['date', 'BEN', 'CH4', 'CO', 'EBE', 'NMHC', 'NO', 'NO_2', 'NOx', 'O_3', 'PM10', 'PM25', 'SO_2', 'TCH', 'TOL', 'station']
```

2

合併相同結構的資料

- 以2001~2003年為例, 其他年份作法相同
- 移除日期和觀測站編號
- 移除該列中若有遺失值的狀況

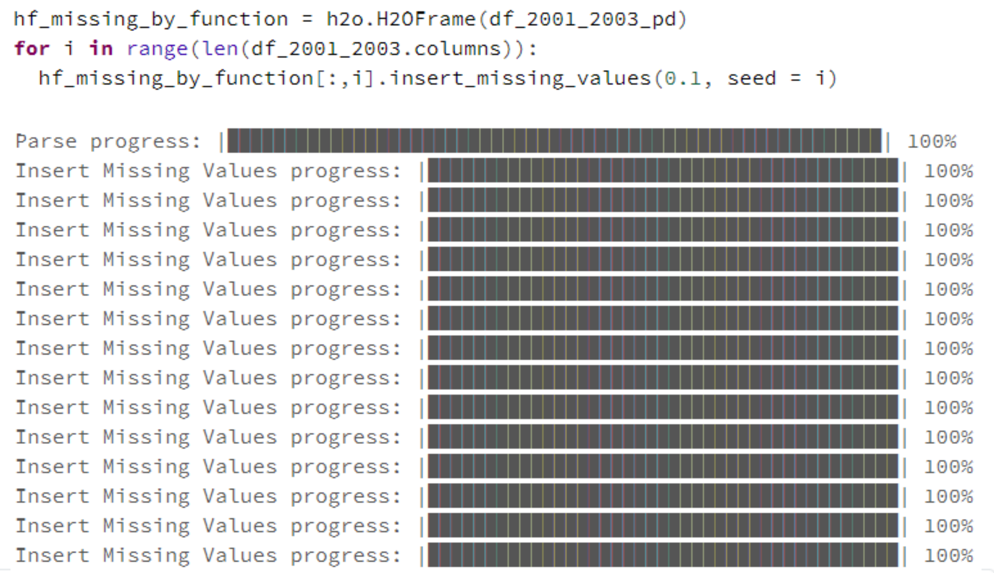


```
(95060, 14)
+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------
+------------------+------------------+------------------+------------------+------------------+
|               BEN|                CO|               EBE|               MXY|              NMHC|              NO_2|               NOx|               OXY|               O_3
|              PM10|               PXY|              SO_2|               TCH|               TOL|
+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------
+------------------+------------------+------------------+------------------+------------------+
|  8.40999984741211| 1.940000057220459| 9.829999923706055|21.489999771118164|0.44999998807907104| 90.30000305175781| 384.8999938964844| 9.479999542236328|9.949999809265137
|  95.1500015258789|7.940000057220459|29.270000457763672| 1.600000023841858| 38.56999969482422|
|3.4600000381469727|1.2699999809265137| 3.430000066757202| 7.079999923706055|0.18000000715255737|             54.25| 173.3000030517578| 3.369999885559082|6.539999961853027
|  53.0099983215332|2.619999885559082| 8.800000190734863|               1.5|14.600000381469727|
| 6.389999866485596|1.7899999618530273|              5.75|10.880000114440918|0.33000001311302185| 75.45999908447266| 281.1000061035156| 3.680000066757202|6.690000057220459
| 63.84000015258789|4.239999771118164|18.459999084472656|1.6799999475479126|23.510000228881836|
| 7.420000076293945|1.4700000286102295|10.630000114440918|24.729999542236328| 0.3499999940395355| 83.30999755859375|277.20001220703125|              11.0|9.899999618530273
|58.880001068115234|8.930000305175781|24.709999084472656|               1.5|37.630001068115234|
| 3.619999885559082|1.2899999618530273| 3.200000047683716| 7.079999923706055| 0.1899999976158142|42.209999084472656| 166.3000030517578|3.4100000858306885|6.380000114440918
|47.599998474121094|2.700000047683716|  8.40999984741211|1.5099999904632568|14.329999923706055|
+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------
+------------------+------------------+------------------+------------------+------------------+
only showing top 5 rows
```

2

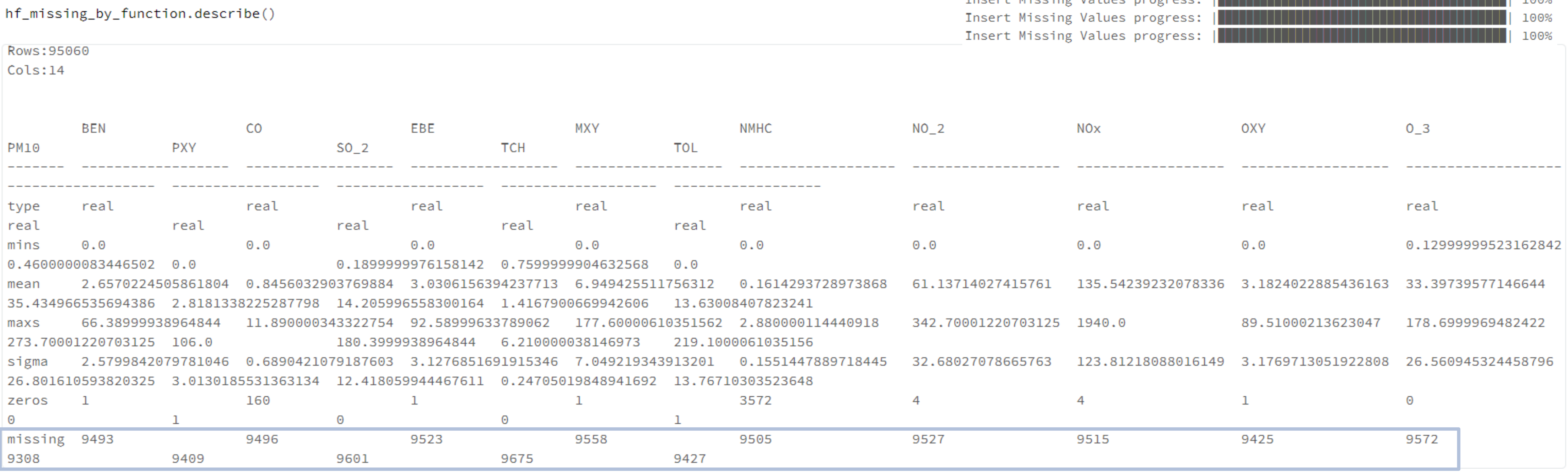
生成missing data

- 發現H2O function是設定機率加入missing
- 導致每個欄位missing數量並非真正10%



```
hf_missing_by_function = h2o.H2OFrame(df_2001_2003_pd)
for i in range(len(df_2001_2003.columns)):
  hf_missing_by_function[:,i].insert_missing_values(0.1, seed = i)
```

```
Parse progress: |████████████████████████████████████████████████| 100%
Insert Missing Values progress: |██████████████████████████████████████| 100%
Insert Missing Values progress: |██████████████████████████████████████| 100%
Insert Missing Values progress: |██████████████████████████████████████| 100%
Insert Missing Values progress: |██████████████████████████████████████| 100%
Insert Missing Values progress: |██████████████████████████████████████| 100%
Insert Missing Values progress: |██████████████████████████████████████| 100%
Insert Missing Values progress: |██████████████████████████████████████| 100%
Insert Missing Values progress: |██████████████████████████████████████| 100%
Insert Missing Values progress: |██████████████████████████████████████| 100%
Insert Missing Values progress: |██████████████████████████████████████| 100%
Insert Missing Values progress: |██████████████████████████████████████| 100%
Insert Missing Values progress: |██████████████████████████████████████| 100%
Insert Missing Values progress: |██████████████████████████████████████| 100%
Insert Missing Values progress: |██████████████████████████████████████| 100%
```

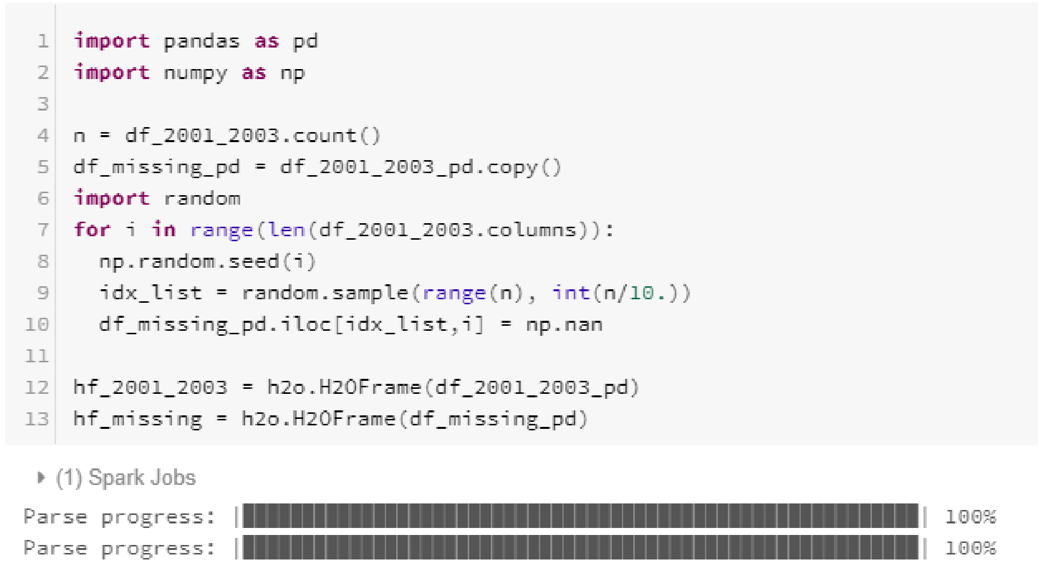


```
hf_missing_by_function.describe()
```

```
Rows:95060
Cols:14


         BEN                 CO                  EBE                 MXY                 NMHC                 NO_2                NOx                 OXY                 O_3
PM10                PXY                 SO_2                TCH                 TOL
-------  ------------------  ------------------  ------------------  ------------------  -------------------  ------------------  ------------------  ------------------  -------------------
------------------  ------------------  ------------------  -------------------  ------------------
type     real                real                real                real                real                 real                real                real                real
real                real                real                real                real
mins     0.0                 0.0                 0.0                 0.0                 0.0                  0.0                 0.0                 0.0                 0.12999999523162842
0.4600000083446502  0.0                 0.1899999976158142  0.7599999904632568  0.0
mean     2.6570224505861804  0.8456032903769884  3.0306156394237713  6.949425511756312   0.1614293728973868   61.13714027415761   135.54239232078336  3.1824022885436163  33.39739577146644
35.434966535694386  2.8181338225287798  14.205996558300164  1.4167900669942606  13.63008407823241
maxs     66.38999938964844   11.890000343322754  92.58999633789062   177.60000610351562  2.880000114440918    342.70001220703125  1940.0              89.51000213623047   178.6999969482422
273.70001220703125  106.0               180.3999938964844   6.210000038146973   219.1000061035156
sigma    2.5799842079781046  0.6890421079187603  3.1276851691915346  7.049219343913201   0.1551447889718445   32.68027078665763   123.81218088016149  3.1769713051922808  26.560945324458796
26.801610593820325  3.0130185531363134  12.418059944467611  0.24705019848941692  13.76710303523648
zeros    1                   160                 1                   1                   3572                 4                   4                   1                   0
0                   1                   0                   0                   1
missing  9493                9496                9523                9558                9505                 9527                9515                9425                9572
9308                9409                9601                9675                9427
```

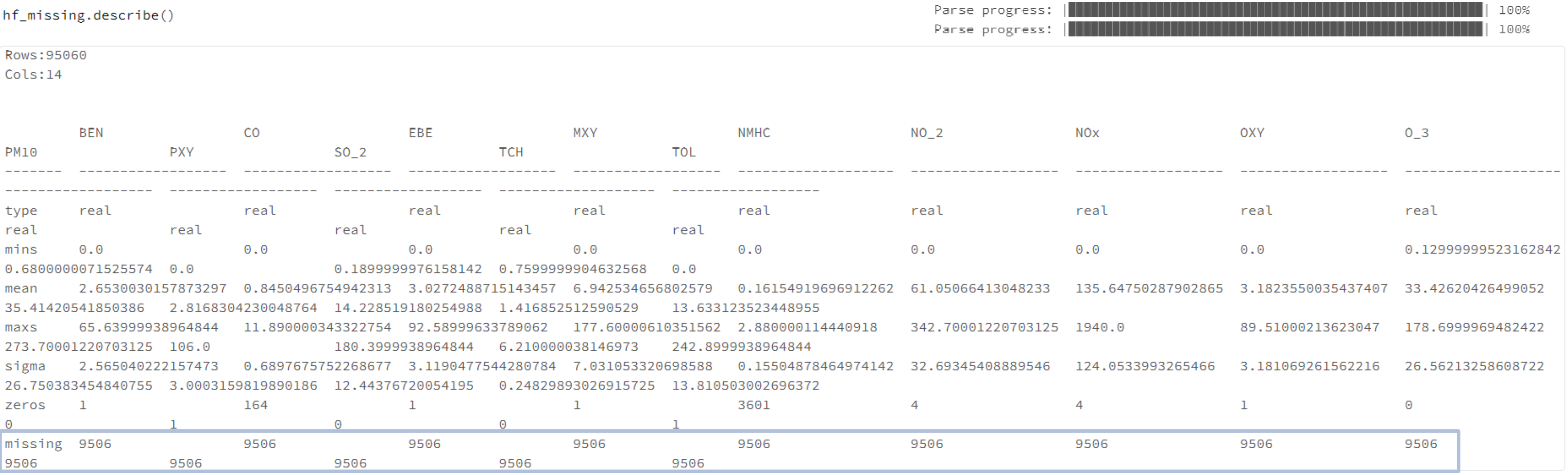
2 生成missing data

➢ 生成隨機index的方式, 移除10%資料量



```python
import pandas as pd
import numpy as np

n = df_2001_2003.count()
df_missing_pd = df_2001_2003_pd.copy()
import random
for i in range(len(df_2001_2003.columns)):
  np.random.seed(i)
  idx_list = random.sample(range(n), int(n/10.))
  df_missing_pd.iloc[idx_list,i] = np.nan

hf_2001_2003 = h2o.H2OFrame(df_2001_2003_pd)
hf_missing = h2o.H2OFrame(df_missing_pd)
```

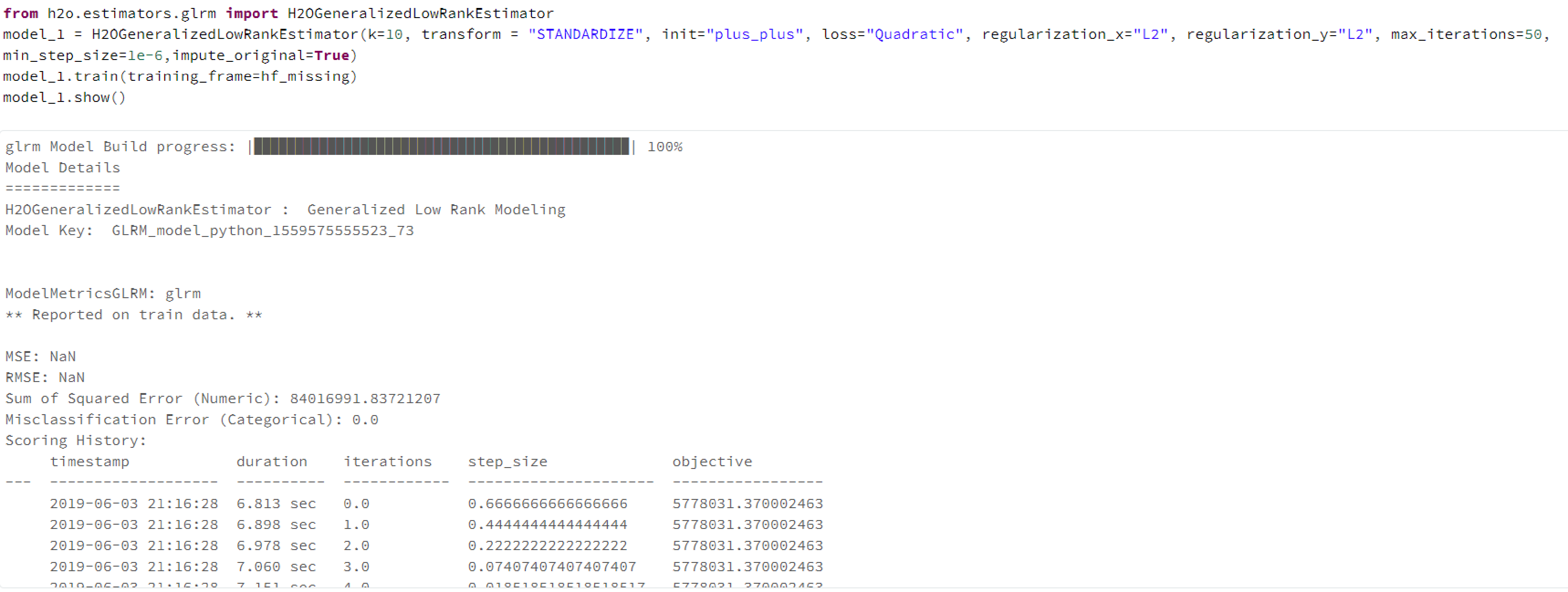
```
▸ (1) Spark Jobs
Parse progress: |████████████████████████████████████████████████████████████| 100%
Parse progress: |████████████████████████████████████████████████████████████| 100%
```



```
hf_missing.describe()
```

```
Rows:95060
Cols:14


        BEN                 CO                  EBE                 MXY                 NMHC                 NO_2                NOx                 OXY                 O_3
PM10              PXY                 SO_2                TCH                  TOL
------- ------------------  ------------------  ------------------  ------------------  -------------------  ------------------  ------------------  ------------------  -------------------
------------------  ------------------  ------------------  -------------------  ------------------
type    real                real                real                real                real                 real                real                real                real
real                real                real                real                 real
mins    0.0                 0.0                 0.0                 0.0                 0.0                  0.0                 0.0                 0.0                 0.1299999523162842
0.6800000071525574  0.0                 0.1899999976158142  0.7599999904632568   0.0
mean    2.6530030157873297  0.8450496754942313  3.0272488715143457  6.942534656802579   0.16154919696912262  61.05066413048233   135.64750287902865  3.1823550035437407  33.42620426499052
35.41420541850386   2.8168304230048764  14.228519180254988  1.416852512590529    13.633123523448955
maxs    65.63999938964844   11.890000343322754  92.58999633789062   177.60000610351562  2.880000114440918    342.70001220703125  1940.0              89.51000213623047   178.6999969482422
273.70001220703125  106.0               180.3999938964844   6.210000038146973    242.8999938964844
sigma   2.565040222157473   0.6897675752268677  3.1190477544280784  7.031053320698588   0.15504878464974142  32.69345408889546   124.0533993265466   3.181069261562216   26.56213258608722
26.750383454840755  3.0003159819890186  12.44376720054195   0.24829893026915725  13.810503002696372
zeros   1                   164                 1                   1                   3601                 4                   4                   1                   0
0                   1                   0                   0                    1
missing 9506                9506                9506                9506                9506                 9506                9506                9506                9506
9506                9506                9506                9506                 9506
```

3 GLRM



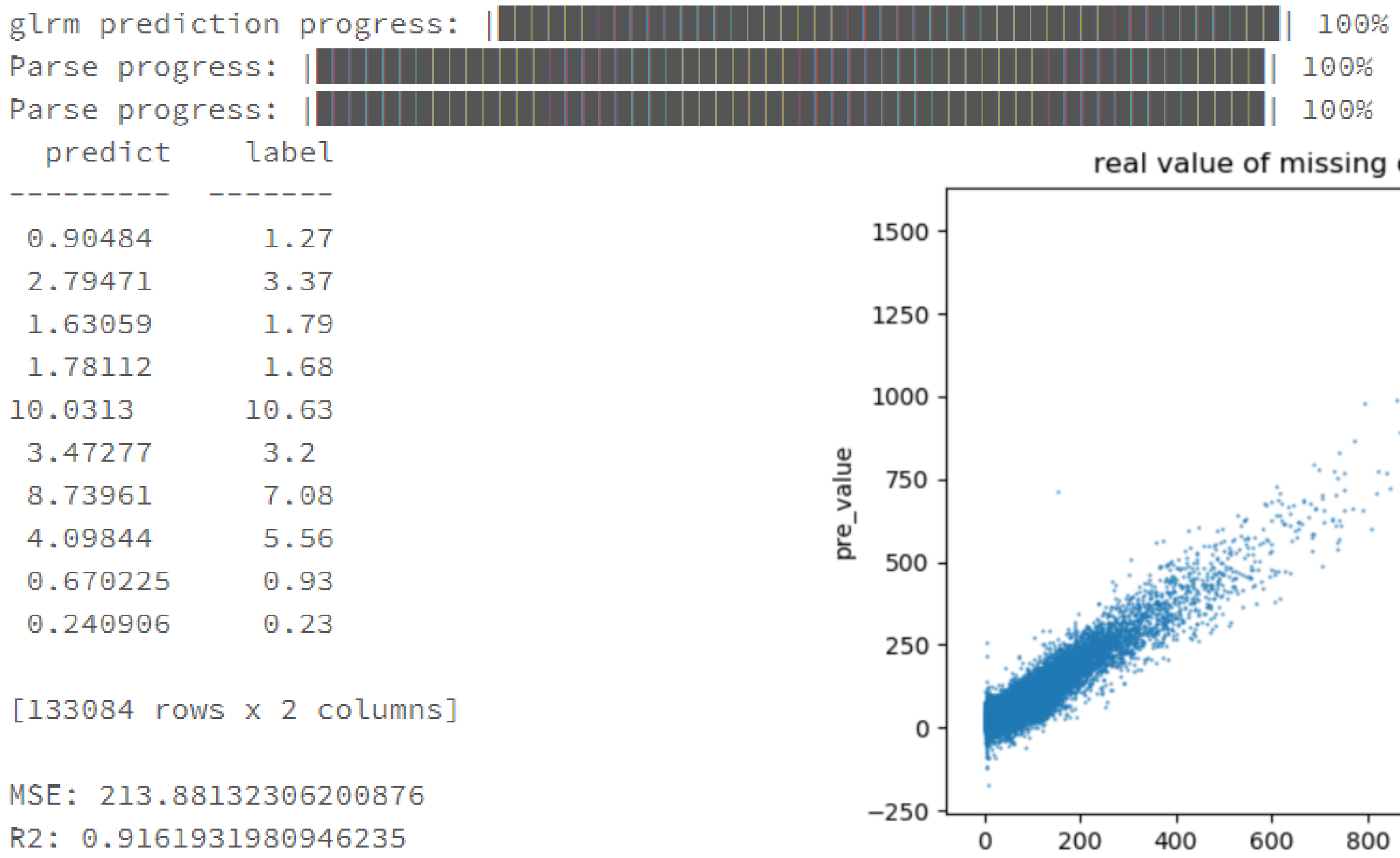
```python
from h2o.estimators.glrm import H2OGeneralizedLowRankEstimator
model_1 = H2OGeneralizedLowRankEstimator(k=10, transform = "STANDARDIZE", init="plus_plus", loss="Quadratic", regularization_x="L2", regularization_y="L2", max_iterations=50,
min_step_size=1e-6,impute_original=True)
model_1.train(training_frame=hf_missing)
model_1.show()
```

```
glrm Model Build progress: |██████████████████████████████████████████████████████| 100%
Model Details
=============
H2OGeneralizedLowRankEstimator :  Generalized Low Rank Modeling
Model Key:  GLRM_model_python_1559575555523_73


ModelMetricsGLRM: glrm
** Reported on train data. **

MSE: NaN
RMSE: NaN
Sum of Squared Error (Numeric): 84016991.83721207
Misclassification Error (Categorical): 0.0
Scoring History:
     timestamp            duration    iterations    step_size              objective
---  -------------------  ----------  ------------  ---------------------  -----------------
     2019-06-03 21:16:28  6.813 sec   0.0           0.6666666666666666     5778031.370002463
     2019-06-03 21:16:28  6.898 sec   1.0           0.4444444444444444     5778031.370002463
     2019-06-03 21:16:28  6.978 sec   2.0           0.2222222222222222     5778031.370002463
     2019-06-03 21:16:28  7.060 sec   3.0           0.07407407407407407    5778031.370002463
     2019-06-03 21:16:28  7.151 sec   4.0           0.018518518518518517   5778031.370002463
```

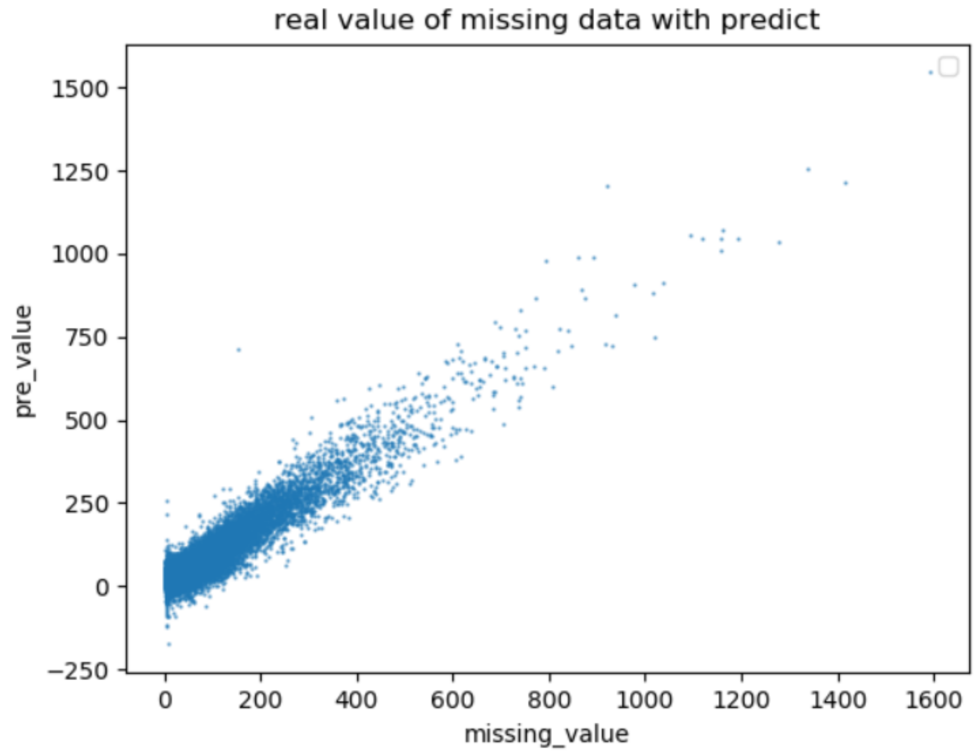
3

GLRM

➢ 觀察預測值跟實際值狀況, 以及MSE、R2

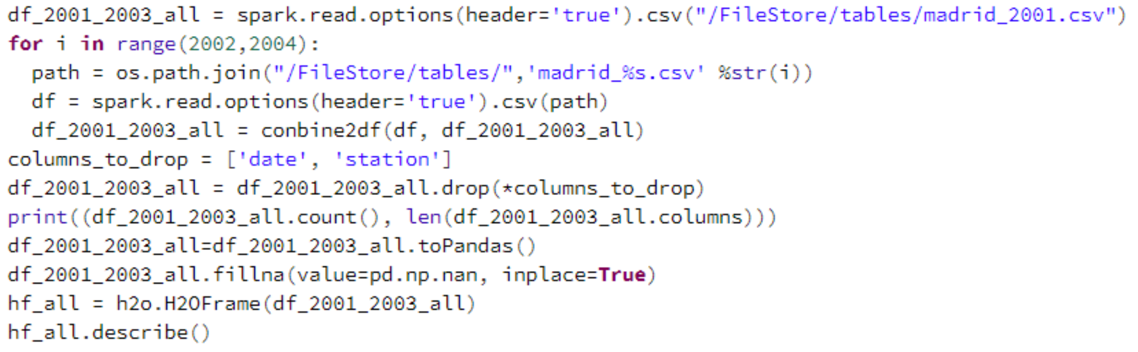


```
glrm prediction progress: |████████████████████████████████████████| 100%
Parse progress: |████████████████████████████████████████| 100%
Parse progress: |████████████████████████████████████████| 100%
  predict     label
---------  -------
 0.90484      1.27
 2.79471      3.37
 1.63059      1.79
 1.78112      1.68
10.0313      10.63
 3.47277      3.2
 8.73961      7.08
 4.09844      5.56
 0.670225     0.93
 0.240906     0.23

[133084 rows x 2 columns]

MSE: 213.88132306200876
R2: 0.9161931980946235
```

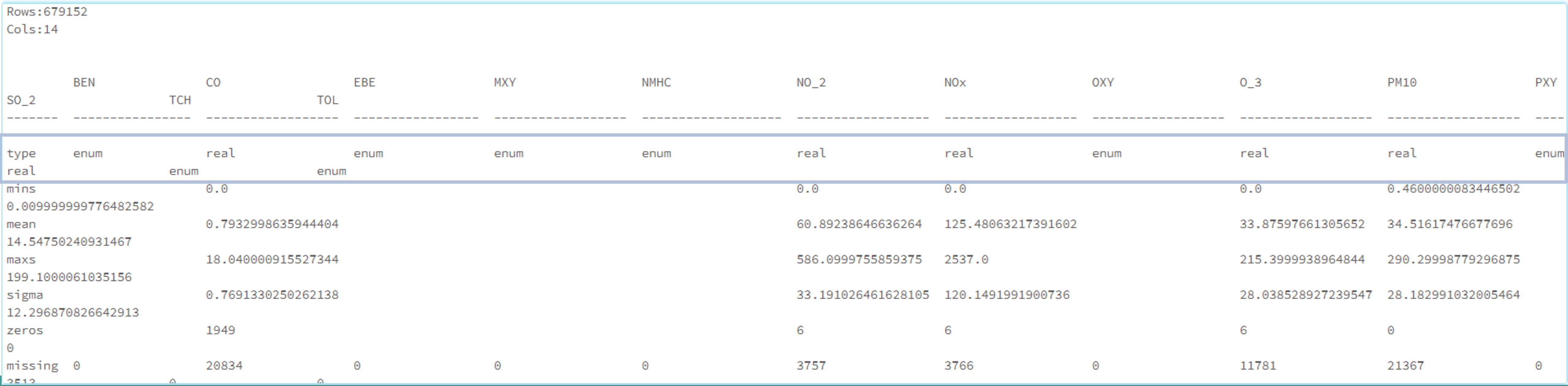


4 原始資料

➢ 應用回去含有missing的原始資料



```python
df_2001_2003_all = spark.read.options(header='true').csv("/FileStore/tables/madrid_2001.csv")
for i in range(2002,2004):
  path = os.path.join("/FileStore/tables/",'madrid_%s.csv' %str(i))
  df = spark.read.options(header='true').csv(path)
  df_2001_2003_all = conbine2df(df, df_2001_2003_all)
columns_to_drop = ['date', 'station']
df_2001_2003_all = df_2001_2003_all.drop(*columns_to_drop)
print((df_2001_2003_all.count(), len(df_2001_2003_all.columns)))
df_2001_2003_all=df_2001_2003_all.toPandas()
df_2001_2003_all.fillna(value=pd.np.nan, inplace=True)
hf_all = h2o.H2OFrame(df_2001_2003_all)
hf_all.describe()
```



```
Rows:679152
Cols:14


        BEN                CO                  EBE                MXY                 NMHC                 NO_2                NOx                 OXY                 O_3                 PM10                PXY
SO_2                TCH                TOL
-------  ----------------  ------------------  -----------------  ------------------  -------------------  ------------------  ------------------  ------------------  ------------------  ------------------  ----
type     enum              real                enum               enum                enum                 real                real                enum                real                real                enum
real                enum               enum
mins                       0.0                                                                             0.0                 0.0                                     0.0                 0.4600000083446502
0.009999999776482582
mean                       0.7932998635944404                                                              60.89238646636264   125.48063217391602                      33.87597661305652   34.51617476677696
14.54750240931467
maxs                       18.040000915527344                                                              586.0999755859375   2537.0                                  215.3999938964844   290.29998779296875
199.1000061035156
sigma                      0.7691330250262138                                                              33.191026461628105  120.1491991900736                       28.038528927239547  28.182991032005464
12.296870826642913
zeros                      1949                                                                            6                   6                                       6                   0
0
missing  0                 20834               0                  0                   0                    3757                3766                0                   11781               21367               0
3513                0                  0
```

4 原始資料

➤ 資料型態設定

```
given_types ={'BEN':'real','CO':'real','EBE':'real','MXY':'real','NMHC':'real','NO_2':'real','NOx':'real','OXY':'real','O_3':'real','PM10':'real','PXY':'real','SO_2':'real','TCH':'real','TOL':'real'}
hf_all = h2o.H2OFrame(df_2001_2003_all, column_types=given_types)
hf_all.describe()
```

```
Parse progress: |████████████████████████████████████████████████████| 100%
Rows:679152
Cols:14


        BEN                CO                 EBE                MXY                NMHC                NO_2               NOx                OXY                O_3                PM10               PX
Y                SO_2               TCH                 TOL
------- ------------------ ------------------ ------------------ ------------------ ------------------- ------------------ ------------------ ------------------ ------------------ ------------------ --
---------------- -------------------- ------------------- ------------------
type    real               real               real               real               real                real               real               real               real               real               re
al               real                 real                real
mins    0.0                0.0                0.0                0.0                0.0                 0.0                0.0                0.0                0.0                0.4600000083446502 0.
0                0.009999999776482582 0.1599999964237213  0.0
mean    2.581386207824226  0.7932998635944404 2.7578531393276013 6.242141480968833  0.1732883057995222  60.89238646636264  125.48063217391602 2.807127976049601  33.87597661305652  34.5161747667769б  2.
571085016477991 14.54750240931467    1.444724870550577   11.907861810575247
maxs    66.38999938964844  18.040000915527344 162.1999969482422  177.60000610351562 4.980000019073486   586.0999755859375  2537.0             103.0              215.3999938964844  290.2999877929687 5 10
6.0              199.1000061035156    6.320000171661377   242.8999938964844
sigma   3.012069349819225  0.7691330250262138 3.1917387012305207 6.878255889944335  0.17126944996481594 33.19102646162810 5 120.1491991900736  3.2683043190333425 28.03852892723954 7 28.182991032005464 2.
9138164039668943 12.296870826642913  0.26881285140334976 13.026110597938137
zeros   1                  1949               1                  1                  8625                6                  6                  1                  6                  0                  1
0                   0                    1
```

5 套用模型

➢ 預測缺失值

```
hf_all_predict = model_1.predict(hf_all)

glrm prediction progress: |████████████████████████████████████████| 100%
```

➢ 轉成pandas dataframe

```
hf_all_predict = hf_all_predict.as_data_frame(use_pandas=True)
hf_all = hf_all.as_data_frame(use_pandas=True)
```

```
hf_all.head()

Out[18]:
   BEN    CO  EBE  MXY  NMHC ...       PM10  PXY       SO_2   TCH  TOL
0  NaN  1.72  NaN  NaN   NaN ...  55.209999  NaN  24.299999   NaN  NaN
1  NaN  1.45  NaN  NaN  0.26 ...  52.389999  NaN  14.230000  1.55  NaN
2  NaN  1.57  NaN  NaN   NaN ...  63.240002  NaN  17.879999   NaN  NaN
3  NaN  2.45  NaN  NaN   NaN ...  67.839996  NaN  24.900000   NaN  NaN
4  NaN  3.26  NaN  NaN   NaN ...  95.779999  NaN  18.750000   NaN  NaN

[5 rows x 14 columns]
```

```
hf_all_predict.head()

Out[17]:
   reconstr_BEN  reconstr_CO      ...       reconstr_TCH  reconstr_TOL
0     14.871348     1.638931      ...           1.984641    148.620780
1     13.153176     1.303795      ...           1.292092    112.591452
2    -24.271660     1.651820      ...           2.499939   -152.156499
3      7.448131     2.722767      ...           1.633151     -9.353156
4     18.068386     3.005149      ...           2.060223    153.798010

[5 rows x 14 columns]
```

6 補回遺漏值

➢ 將原先為missing的部分, 補上predict的數值

```
hf_all[hf_all.isnull()] = hf_all_predict.values
print(hf_all)
```

```
hf_all.head()

Out[18]:
   BEN    CO  EBE  MXY  NMHC ...       PM10  PXY       SO_2   TCH  TOL
0  NaN  1.72  NaN  NaN   NaN ...  55.209999  NaN  24.299999   NaN  NaN
1  NaN  1.45  NaN  NaN  0.26 ...  52.389999  NaN  14.230000  1.55  NaN
2  NaN  1.57  NaN  NaN   NaN ...  63.240002  NaN  17.879999   NaN  NaN
3  NaN  2.45  NaN  NaN   NaN ...  67.839996  NaN  24.900000   NaN  NaN
4  NaN  3.26  NaN  NaN   NaN ...  95.779999  NaN  18.750000   NaN  NaN

[5 rows x 14 columns]
```

```
hf_all_predict.head()

Out[17]:
   reconstr_BEN  reconstr_CO      ...       reconstr_TCH  reconstr_TOL
0     14.871348     1.638931      ...           1.984641    148.620780
1     13.153176     1.303795      ...           1.292092    112.591452
2    -24.271660     1.651820      ...           2.499939   -152.156499
3      7.448131     2.722767      ...           1.633151     -9.353156
4     18.068386     3.005149      ...           2.060223    153.798010

[5 rows x 14 columns]
```

```
              BEN        CO      ...          TCH         TOL
0       14.871348  1.720000      ...     1.984641  148.620780
1       13.153176  1.450000      ...     1.550000  112.591452
2      -24.271660  1.570000      ...     2.499939 -152.156499
3        7.448131  2.450000      ...     1.633151   -9.353156
4       18.068386  3.260000      ...     2.060223  153.798010
5        8.410000  1.940000      ...     1.600000   38.570000
6       14.294153  1.380000      ...     1.490000  120.305082
7        0.425022  1.580000      ...     1.580000   16.097257
8        5.928371 -6.482529      ...    -2.079138  108.989225
9       28.389403  1.920000      ...     0.763445  215.113999
10      15.198524  1.330000      ...     1.800000  119.662176
11      -8.381973  2.180000      ...     2.659169  -76.410493
12      -4.443877  1.140000      ...     1.620000  -25.829356
13      -5.453043  4.680000      ...     3.581642 -115.429274
14       6.970000  1.440000      ...     1.860000   32.299999
15      -7.350760  1.250000      ...     2.334651  -13.051732
16     -11.563627  1.640000      ...     3.536316 -119.051314
17      -8.851091  1.850000      ...     1.850000  -47.399343
18      -9.429507  1.740000      ...     2.581548  -77.686821
19       0.811453  1.540000      ...     1.266664  -28.297347
```