script

April 6, 2024

1 polars

rust polars URL https://www.statlab.co.jp/seminar/winequality-red2.csv

1.1

[ ]:
```
import polars as pl
```

1.2

csv pandas ,read_csv write_csv

[ ]:
```
df = pl.read_csv(
    "./winequality-red2.csv",
    dtypes={"free_sulfur_dioxide": pl.Float64, "total_sulfur_dioxide": pl.
 ↪Float64},
)
```

int64 float64

1.3

pandas describe()

[ ]:
```
df
```

[ ]:
```
shape: (1_599, 12)
```

fixed_acidi	volatile_ac	citric_acid	residual_su	…	ph	sulphates	alcohol	quality
ty	idity	---	gar		---	---	---	---
---	---	f64	---		f64	f64	f64	i64
f64	f64		f64					

```
 7.4          0.7          0.0          1.9          …  3.51   0.56
9.4      5
 7.8          0.88         0.0          2.6          …  3.2    0.68
9.8      5
 7.8          0.76         0.04         2.3          …  3.26   0.65
9.8      5
 11.2         0.28         0.56         1.9          …  3.16   0.58
9.8      6
 7.4          0.7          0.0          1.9          …  3.51   0.56
9.4      5
 …            …            …            …            …  …      …
…        …
 6.2          0.6          0.08         2.0          …  3.45   0.58
10.5     5
 5.9          0.55         0.1          2.2          …  3.52   0.76
11.2     6
 6.3          0.51         0.13         2.3          …  3.42   0.75
11.0     6
 5.9          0.645        0.12         2.0          …  3.57   0.71
10.2     5
 6.0          0.31         0.47         3.6          …  3.39   0.66
11.0     6
```

[ ]:
```
df.describe()
```

[ ]:
```
shape: (9, 13)
```

```
 statistic   fixed_aci   volatile_   citric_ac   …  ph          sulphates
alcohol     quality
 ---           dity        acidity     id             ---         ---
---         ---
 str           ---         ---         ---            f64         f64
f64         f64
               f64         f64         f64
```

```
 count         1599.0      1599.0      1599.0      …  1599.0      1599.0
1599.0      1599.0
 null_count    0.0         0.0         0.0         …  0.0         0.0
0.0         0.0
 mean          8.319637    0.527821    0.270976    …  3.311113    0.658149
10.422983   5.636023
```

```
  std          1.741096    0.17906     0.194801    …   0.154386   0.169507
1.065668    0.807569
  min          4.6         0.12        0.0         …   2.74       0.33
8.4         3.0
  25%          7.1         0.39        0.09        …   3.21       0.55
9.5         5.0
  50%          7.9         0.52        0.26        …   3.31       0.62
10.2        6.0
  75%          9.2         0.64        0.42        …   3.4        0.73
11.1        6.0
  max          15.9        1.58        1.0         …   4.01       2.0
14.9        8.0
```

1.4

polars drop ph

[ ]:
```
df.drop(["ph"])
```

[ ]:
```
shape: (1_599, 11)
```

```
  fixed_acidi   volatile_a   citric_aci   residual_s   …   density   sulphates
alcohol   quality
  ty            cidity       d            ugar             ---       ---
---       ---
  ---           ---          ---          ---              f64       f64
f64       i64
  f64           f64          f64          f64



  7.4           0.7          0.0          1.9          …   0.9978    0.56
9.4       5
  7.8           0.88         0.0          2.6          …   0.9968    0.68
9.8       5
  7.8           0.76         0.04         2.3          …   0.997     0.65
9.8       5
  11.2          0.28         0.56         1.9          …   0.998     0.58
9.8       6
  7.4           0.7          0.0          1.9          …   0.9978    0.56
9.4       5
  …             …            …            …            …   …         …
…         …
  6.2           0.6          0.08         2.0          …   0.9949    0.58
```

```
10.5      5
 5.9           0.55         0.1          2.2          …   0.99512   0.76
11.2      6
 6.3           0.51         0.13         2.3          …   0.99574   0.75
11.0      6
 5.9           0.645        0.12         2.0          …   0.99547   0.71
10.2      5
 6.0           0.31         0.47         3.6          …   0.99549   0.66
11.0      6
```

1.5

[ ]:
```
df.null_count()
```

[ ]:
```
shape: (1, 12)

 fixed_acidit   volatile_ac   citric_acid   residual_su   …   ph    sulphates
alcohol   quality
 y              idity         ---           gar               ---   ---
---       ---
 ---            ---           u32           ---               u32   u32
u32       u32
 u32            u32                         u32

 0              0             0             0             …   0     0
0         0
```

null 0
citric_acid 0

1.6

[ ]:
```
#
colum_name = "citric_acid"
# 0.0
df_tmp = df.select(colum_name).filter(pl.col(colum_name) != 0.0)
#
mean = df_tmp[colum_name].mean()
```

```
print(mean)
```

```
0.29535787321063395
```

1.7 0

null (fill_null) 0 when...then...otherwise... fill_null

[ ]:
```
#

# citric_acid 0.0
df = df.with_columns(
    pl.when(pl.col("citric_acid") == 0.0)  # citric_acid 0.0
    .then(mean)  #
    .otherwise(pl.col("citric_acid"))  #
    .alias("citric_acid")  #
)
```

null when pl.col("citric_acid").is_null()

[ ]:
```
#
#      null  (otherwise         null   )

# citric_acid 0.0
df = df.with_columns(
    pl.when(pl.col("citric_acid") != mean)  # citric_acid null
    .then(pl.col("citric_acid"))  #
    .alias("citric_acid")  #
)
df
```

[ ]:
```
shape: (1_599, 12)

 fixed_acidi  volatile_ac  citric_acid  residual_su  …  ph    sulphates
alcohol  quality
 ty           idity        ---          gar          ---   ---
---      ---
 ---          ---          f64          ---          f64   f64
f64      i64
 f64          f64                       f64


 7.4          0.7          null         1.9          …  3.51  0.56
9.4      5
 7.8          0.88         null         2.6          …  3.2   0.68
```

```
9.8       5
 7.8            0.76          0.04          2.3           …   3.26   0.65
9.8       5
 11.2           0.28          0.56          1.9           …   3.16   0.58
9.8       6
 7.4            0.7           null          1.9           …   3.51   0.56
9.4       5
 …              …             …             …             …   …      …
…         …
 6.2            0.6           0.08          2.0           …   3.45   0.58
10.5      5
 5.9            0.55          0.1           2.2           …   3.52   0.76
11.2      6
 6.3            0.51          0.13          2.3           …   3.42   0.75
11.0      6
 5.9            0.645         0.12          2.0           …   3.57   0.71
10.2      5
 6.0            0.31          0.47          3.6           …   3.39   0.66
11.0      6
```

[ ]:
```
#

# citric_acid 0.0
df = df.with_columns(
    pl.when(pl.col("citric_acid").is_null())  # citric_acid null
    .then(mean)  #
    .otherwise(pl.col("citric_acid"))  #
    .alias("citric_acid")  #
)
```

[ ]:
```
df
```

[ ]:
```
shape: (1_599, 12)
```

```
 fixed_acidi   volatile_ac   citric_acid   residual_su   …   ph     sulphates
alcohol   quality
 ty            idity         ---           gar               ---    ---
---       ---
 ---           ---           f64           ---               f64    f64
f64       i64
 f64           f64                         f64
```

```
 7.4          0.7          0.295358     1.9          …  3.51   0.56
9.4      5
 7.8          0.88         0.295358     2.6          …  3.2    0.68
9.8      5
 7.8          0.76         0.04         2.3          …  3.26   0.65
9.8      5
 11.2         0.28         0.56         1.9          …  3.16   0.58
9.8      6
 7.4          0.7          0.295358     1.9          …  3.51   0.56
9.4      5
 …            …            …            …            …  …      …
…        …
 6.2          0.6          0.08         2.0          …  3.45   0.58
10.5     5
 5.9          0.55         0.1          2.2          …  3.52   0.76
11.2     6
 6.3          0.51         0.13         2.3          …  3.42   0.75
11.0     6
 5.9          0.645        0.12         2.0          …  3.57   0.71
10.2     5
 6.0          0.31         0.47         3.6          …  3.39   0.66
11.0     6
```

1.8 0

when...then...otherwise...

[ ]:
```
#      null   (otherwise         null    )

# citric_acid 0.0
df = df.with_columns(
    pl.when(pl.col("citric_acid") != mean)  # citric_acid null
    .then(pl.col("citric_acid"))  #
    .alias("citric_acid")  #
)
df
```

[ ]:
```
shape: (1_599, 12)
```

```
 fixed_acidi  volatile_ac  citric_acid  residual_su  …  ph     sulphates
alcohol  quality
 ty           idity        ---          gar             ---    ---
---      ---
 ---          ---          f64          ---             f64    f64
```

```
f64       i64
 f64           f64                       f64


 7.4           0.7           null        1.9           …  3.51   0.56
9.4       5
 7.8           0.88          null        2.6           …  3.2    0.68
9.8       5
 7.8           0.76          0.04        2.3           …  3.26   0.65
9.8       5
 11.2          0.28          0.56        1.9           …  3.16   0.58
9.8       6
 7.4           0.7           null        1.9           …  3.51   0.56
9.4       5
 …             …             …           …             …  …      …
…         …
 6.2           0.6           0.08        2.0           …  3.45   0.58
10.5      5
 5.9           0.55          0.1         2.2           …  3.52   0.76
11.2      6
 6.3           0.51          0.13        2.3           …  3.42   0.75
11.0      6
 5.9           0.645         0.12        2.0           …  3.57   0.71
10.2      5
 6.0           0.31          0.47        3.6           …  3.39   0.66
11.0      6
```

[ ]:
```
df.fill_null(0)
```

[ ]:
```
shape: (1_599, 12)


 fixed_acidi   volatile_ac   citric_acid   residual_su   …  ph     sulphates
alcohol   quality
 ty            idity         ---           gar              ---    ---
---       ---
 ---           ---           f64           ---              f64    f64
f64       i64
 f64           f64                         f64


 7.4           0.7           0.0           1.9           …  3.51   0.56
9.4       5
```

```
  7.8           0.88          0.0           2.6           …  3.2    0.68
9.8       5
  7.8           0.76          0.04          2.3           …  3.26   0.65
9.8       5
  11.2          0.28          0.56          1.9           …  3.16   0.58
9.8       6
  7.4           0.7           0.0           1.9           …  3.51   0.56
9.4       5
  …             …             …             …             …  …      …
…         …
  6.2           0.6           0.08          2.0           …  3.45   0.58
10.5      5
  5.9           0.55          0.1           2.2           …  3.52   0.76
11.2      6
  6.3           0.51          0.13          2.3           …  3.42   0.75
11.0      6
  5.9           0.645         0.12          2.0           …  3.57   0.71
10.2      5
  6.0           0.31          0.47          3.6           …  3.39   0.66
11.0      6
```

1.9

[ ]:
```
df
```

[ ]:
```
shape: (1_599, 12)
```

```
  fixed_acidi   volatile_ac   citric_acid   residual_su   …  ph     sulphates
alcohol   quality
  ty            idity         ---           gar              ---    ---
---       ---
  ---           ---           f64           ---              f64    f64
f64       i64
  f64           f64                         f64



  7.4           0.7           null          1.9           …  3.51   0.56
9.4       5
  7.8           0.88          null          2.6           …  3.2    0.68
9.8       5
  7.8           0.76          0.04          2.3           …  3.26   0.65
9.8       5
  11.2          0.28          0.56          1.9           …  3.16   0.58
```

```
9.8       6
  7.4             0.7           null          1.9           …   3.51   0.56
9.4       5
  …               …             …             …             …   …      …
…         …
  6.2             0.6           0.08          2.0           …   3.45   0.58
10.5      5
  5.9             0.55          0.1           2.2           …   3.52   0.76
11.2      6
  6.3             0.51          0.13          2.3           …   3.42   0.75
11.0      6
  5.9             0.645         0.12          2.0           …   3.57   0.71
10.2      5
  6.0             0.31          0.47          3.6           …   3.39   0.66
11.0      6
```

null

[ ]:
```
#      null   (otherwise          null    )

# citric_acid 0.0
df = df.with_columns(
    pl.when(pl.col("citric_acid") != mean)  # citric_acid null
    .then(pl.col("citric_acid"))  #
    .alias("citric_acid")  #
)
df
```

[ ]:
```
shape: (1_599, 12)
```

```
  fixed_acidi   volatile_ac   citric_acid   residual_su   …   ph     sulphates
alcohol   quality
  ty            idity         ---           gar               ---    ---
---       ---
  ---           ---           f64           ---               f64    f64
f64       i64
  f64           f64                         f64


  7.4             0.7           null          1.9           …   3.51   0.56
9.4       5
  7.8             0.88          null          2.6           …   3.2    0.68
9.8       5
  7.8             0.76          0.04          2.3           …   3.26   0.65
```

```
9.8       5
  11.2          0.28          0.56          1.9           …   3.16   0.58
9.8       6
  7.4           0.7           null          1.9           …   3.51   0.56
9.4       5
  …             …             …             …             …   …      …
…         …
  6.2           0.6           0.08          2.0           …   3.45   0.58
10.5      5
  5.9           0.55          0.1           2.2           …   3.52   0.76
11.2      6
  6.3           0.51          0.13          2.3           …   3.42   0.75
11.0      6
  5.9           0.645         0.12          2.0           …   3.57   0.71
10.2      5
  6.0           0.31          0.47          3.6           …   3.39   0.66
11.0      6
```

citric_acid null

[ ]:
```
#
colum_name = "citric_acid"  #
max_value = df[colum_name].max()
print(max_value)


#


# apply
df = df.with_columns(
    pl.when(pl.col(colum_name).is_null())
    .then(max_value)
    .otherwise(pl.col(colum_name))
    .alias(colum_name)
)
df
```

```
1.0
```

[ ]:
```
shape: (1_599, 12)

 fixed_acidi   volatile_ac   citric_acid   residual_su   …   ph     sulphates
alcohol   quality
 ty            idity         ---           gar               ---    ---
```

```
---       ---
  ---           ---           f64           ---               f64    f64
f64       i64
  f64           f64                         f64


  7.4           0.7           1.0           1.9           …   3.51   0.56
9.4       5
  7.8           0.88          1.0           2.6           …   3.2    0.68
9.8       5
  7.8           0.76          0.04          2.3           …   3.26   0.65
9.8       5
  11.2          0.28          0.56          1.9           …   3.16   0.58
9.8       6
  7.4           0.7           1.0           1.9           …   3.51   0.56
9.4       5
  …             …             …             …             …   …      …
…         …
  6.2           0.6           0.08          2.0           …   3.45   0.58
10.5      5
  5.9           0.55          0.1           2.2           …   3.52   0.76
11.2      6
  6.3           0.51          0.13          2.3           …   3.42   0.75
11.0      6
  5.9           0.645         0.12          2.0           …   3.57   0.71
10.2      5
  6.0           0.31          0.47          3.6           …   3.39   0.66
11.0      6
```

2

2.1 pandas

pyarrow

[ ]:
```
df_pd = df.to_pandas()
df_pd
```

[ ]:
```
      fixed_acidity  volatile_acidity  citric_acid  residual_sugar  chlorides  \
0               7.4             0.700         1.00             1.9      0.076
1               7.8             0.880         1.00             2.6      0.098
2               7.8             0.760         0.04             2.3      0.092
3              11.2             0.280         0.56             1.9      0.075
```

```
4               7.4             0.700         1.00              1.9      0.076
…               …               …             …                 …        …
1594            6.2             0.600         0.08              2.0      0.090
1595            5.9             0.550         0.10              2.2      0.062
1596            6.3             0.510         0.13              2.3      0.076
1597            5.9             0.645         0.12              2.0      0.075
1598            6.0             0.310         0.47              3.6      0.067

      free_sulfur_dioxide  total_sulfur_dioxide  density    ph  sulphates  \
0                    11.0                  34.0  0.99780  3.51       0.56
1                    25.0                  67.0  0.99680  3.20       0.68
2                    15.0                  54.0  0.99700  3.26       0.65
3                    17.0                  60.0  0.99800  3.16       0.58
4                    11.0                  34.0  0.99780  3.51       0.56
…                    …                     …     …        …          …
1594                 32.0                  44.0  0.99490  3.45       0.58
1595                 39.0                  51.0  0.99512  3.52       0.76
1596                 29.0                  40.0  0.99574  3.42       0.75
1597                 32.0                  44.0  0.99547  3.57       0.71
1598                 18.0                  42.0  0.99549  3.39       0.66

      alcohol  quality
0         9.4        5
1         9.8        5
2         9.8        5
3         9.8        6
4         9.4        5
…         …          …
1594     10.5        5
1595     11.2        6
1596     11.0        6
1597     10.2        5
1598     11.0        6

[1599 rows x 12 columns]
```

[ ]:
```
print(type(df_pd))
print(type(df))
```

```
<class 'pandas.core.frame.DataFrame'>
<class 'polars.dataframe.frame.DataFrame'>
```

2.2

pandas

[ ]:
```
import numpy as np
```

```python
def reduce(df):
    start_mem = df.memory_usage().sum() / 1024**2
    print("befor:{:.2f}MB".format(start_mem))

    for col in df.columns:
        col_type = df[col].dtype

        if col_type != object:
            c_min = df[col].min()
            c_max = df[col].max()
            if str(col_type)[:3] == "int":
                if c_min > np.iinfo(np.int8).min and c_max < np.iinfo(np.int8).
 ↪max:
                    df[col] = df[col].astype(np.int8)
                elif c_min > np.iinfo(np.int8).min and c_max < np.iinfo(np.
 ↪int16).max:
                    df[col] = df[col].astype(np.int16)
                elif c_min > np.iinfo(np.int8).min and c_max < np.iinfo(np.
 ↪int32).max:
                    df[col] = df[col].astype(np.int32)
                elif c_min > np.iinfo(np.int8).min and c_max < np.iinfo(np.
 ↪int64).max:
                    df[col] = df[col].astype(np.int64)
            else:
                if (
                    c_min > np.finfo(np.float16).min
                    and c_max < np.finfo(np.float16).max
                ):
                    df[col] = df[col].astype(np.float16)
                elif (
                    c_min > np.finfo(np.float32).min
                    and c_max < np.finfo(np.float32).max
                ):
                    df[col] = df[col].astype(np.float32)
                else:
                    df[col] = df[col].astype(np.float64)

        else:
            pass

    end_mem = df.memory_usage().sum() / 1024**2
    print("after:{:.2f}MB".format(end_mem))
    print("  {:.1f}%".format(100 * (start_mem - end_mem) / start_mem))

    return df
```

[ ]:
```
df_pd = reduce(df_pd)
```

```
befor:0.15MB
after:0.04MB
   76.0%
```