Wind Statistics

Introduction:

The data have been modified to contain some missing values, identified by NaN.
Using pandas should make this exercise easier, in particular for the bonus question.

You should be able to perform all of these operations without using a for loop or other looping construct.

1. The data in 'wind.data' has the following format:

```
"""
Yr Mo Dy   RPT   VAL   ROS   KIL   SHA   BIR   DUB   CLA   MUL   CLO   BEL   MAL
61  1  1 15.04 14.96 13.17  9.29   NaN  9.87 13.67 10.25 10.83 12.58 18.50 15.04
61  1  2 14.71   NaN 10.83  6.50 12.62  7.67 11.50 10.04  9.79  9.67 17.54 13.83
61  1  3 18.50 16.88 12.33 10.13 11.17  6.17 11.25   NaN  8.50  7.67 12.75 12.71
"""
```

```
'\nYr Mo Dy   RPT   VAL   ROS   KIL   SHA   BIR   DUB   CLA   MUL   CLO   BEL   MAL\n61  1  1 15.04 14.96 13.17  9.29   NaN  9.87 13.67
10.25 10.83 12.58 18.50 15.04\n61  1  2 14.71   NaN 10.83  6.50 12.62  7.67 11.50 10.04  9.79  9.67 17.54 13.83\n61  1  3 18.50 16.88
12.33 10.13 11.17  6.17 11.25   NaN  8.50  7.67 12.75 12.71\n'
```

The first three columns are year, month and day. The remaining 12 columns are average windspeeds in knots at 12 locations in Ireland on that day.

More information about the dataset go here.

Step 1. Import the necessary libraries

```
import pandas as pd
```

Step 2. Import the dataset from this address

```
data = pd.read_csv('wind_stats.tsv', sep='\s+')
```

Step 3. Assign it to a variable called data and replace the first 3 columns by a proper datetime index.

```
data['date'] = pd.to_datetime(data[['Yr', 'Mo', 'Dy']].astype(str).agg('-'.join, axis=1), format='%y-%m-%d')
data = data.set_index('date').drop(['Yr', 'Mo', 'Dy'], axis=1)
step3 = data.head()
print(step3)
```

```
               RPT    VAL    ROS    KIL    SHA   BIR    DUB    CLA    MUL  \
date
2061-01-01  15.04  14.96  13.17   9.29    NaN  9.87  13.67  10.25  10.83
2061-01-02  14.71    NaN  10.83   6.50  12.62  7.67  11.50  10.04   9.79
2061-01-03  18.50  16.88  12.33  10.13  11.17  6.17  11.25    NaN   8.50
2061-01-04  10.58   6.63  11.75   4.58   4.54  2.88   8.63   1.79   5.83
2061-01-05  13.33  13.25  11.42   6.17  10.71  8.21  11.92   6.54  10.92

               CLO    BEL    MAL
date
2061-01-01  12.58  18.50  15.04
2061-01-02   9.67  17.54  13.83
2061-01-03   7.67  12.75  12.71
2061-01-04   5.88   5.46  10.88
2061-01-05  10.34  12.92  11.83
```

Step 4. Year 2061? Do we really have data from this year? Create a function to fix it and apply it.

```
def fix_year(year):
    if year > 2000:
        return year - 100
    return year
```

```
data.index = data.index.map(lambda x: x.replace(year=fix_year(x.year)))
step4 = data.index[:5]
print(step4)
```

```
DatetimeIndex(['1961-01-01', '1961-01-02', '1961-01-03', '1961-01-04',
               '1961-01-05'],
              dtype='datetime64[ns]', name='date', freq=None)
```

Step 5. Set the right dates as the index. Pay attention at the data type, it should be datetime64[ns].

```
step5 = data.index.dtype
print(step5)
```

```
datetime64[ns]
```

Step 6. Compute how many values are missing for each location over the entire record.

They should be ignored in all calculations below.

```
step6 = data.isna().sum()
print(step6)
```

```
RPT    6
VAL    3
ROS    2
KIL    5
SHA    2
BIR    0
DUB    3
CLA    2
MUL    3
CLO    1
BEL    0
MAL    4
dtype: int64
```

Step 7. Compute how many non-missing values there are in total.

```
step7 = data.notna().sum().sum()
print(step7)
```

```
78857
```

Step 8. Calculate the mean windspeeds of the windspeeds over all the locations and all the times.

A single number for the entire dataset.

```
step8 = data.mean().mean()
print(step8)
```

```
10.227982360836938
```

Step 9. Create a DataFrame called loc_stats and calculate the min, max and mean windspeeds and standard deviations of the windspeeds at each location over all the days

A different set of numbers for each location.

```
loc_stats = pd.DataFrame({
    'min': data.min(),
    'max': data.max(),
    'mean': data.mean(),
    'std': data.std()
})
step9 = loc_stats
print(step9)
```

```
       min    max       mean       std
RPT  0.67  35.80  12.362987  5.618413
VAL  0.21  33.37  10.644314  5.267356
ROS  1.50  33.84  11.660526  5.008450
KIL  0.00  28.46   6.306468  3.605811
SHA  0.13  37.54  10.455834  4.936125
BIR  0.00  26.16   7.092254  3.968683
DUB  0.00  30.37   9.797343  4.977555
CLA  0.00  31.08   8.495053  4.499449
MUL  0.00  25.88   8.493590  4.166872
CLO  0.04  28.21   8.707332  4.503954
BEL  0.13  42.38  13.121007  5.835037
MAL  0.67  42.54  15.599079  6.699794
```

Step 10. Create a DataFrame called day_stats and calculate the min, max and mean windspeed and standard deviations of the windspeeds across all the locations at each day.

A different set of numbers for each day.

```
day_stats = pd.DataFrame({
    'min': data.min(axis=1),
    'max': data.max(axis=1),
    'mean': data.mean(axis=1),
    'std': data.std(axis=1)
})
step10 = day_stats.head()
print(step10)
```

```
              min    max       mean       std
date
1961-01-01  9.29  18.50  13.018182  2.808875
1961-01-02  6.50  17.54  11.336364  3.188994
1961-01-03  6.17  18.50  11.641818  3.681912
1961-01-04  1.79  11.75   6.619167  3.198126
1961-01-05  6.17  13.33  10.630000  2.445356
```

Step 11. Find the average windspeed in January for each location.

Treat January 1961 and January 1962 both as January.

```
step11 = data[data.index.month == 1].mean()
print(step11)
```

```
RPT    14.847325
VAL    12.914560
ROS    13.299624
KIL     7.199498
SHA    11.667734
BIR     8.054839
DUB    11.819355
CLA     9.512047
MUL     9.543208
CLO    10.053566
BEL    14.550520
MAL    18.028763
dtype: float64
```

Step 12. Downsample the record to a yearly frequency for each location.

```
step12 = data.resample('YE').mean()
print(step12)
```

```
                 RPT        VAL        ROS       KIL        SHA       BIR  \
date
1961-12-31  12.299583  10.351796  11.362369  6.958227  10.881763  7.729726
1962-12-31  12.246923  10.110438  11.732712  6.960440  10.657918  7.393068
1963-12-31  12.813452  10.836986  12.541151  7.330055  11.724110  8.434712
1964-12-31  12.363661  10.920164  12.104372  6.787787  11.454481  7.570874
1965-12-31  12.451370  11.075534  11.848767  6.858466  11.024795  7.478110
1966-12-31  13.461973  11.557205  12.020630  7.345726  11.805041  7.793671
1967-12-31  12.737151  10.990986  11.739397  7.143425  11.630740  7.368164
1968-12-31  11.835628  10.468197  11.409754  6.477678  10.760765  6.067322
1969-12-31  11.166356   9.723699  10.902000  5.767973   9.873918  6.189973
```

```
1970-12-31  12.600329  10.726932  11.730247  6.217178  10.567370  7.609452
1971-12-31  11.273123   9.095178  11.088329  5.241507   9.440329  6.097151
1972-12-31  12.463962  10.561311  12.058333  5.929699   9.430410  6.358825
1973-12-31  11.828466  10.680493  10.680493  5.547863   9.640877  6.548740
1974-12-31  13.643096  11.811781  12.336356  6.427041  11.110986  6.809781
1975-12-31  12.008575  10.293836  11.564712  5.269096   9.190082  5.668521
1976-12-31  11.737842  10.203115  10.761230  5.109426   8.846339  6.311038
1977-12-31  13.099616  11.144493  12.627836  6.073945  10.003836  8.586438
1978-12-31  12.504356  11.044274  11.380000  6.082356  10.167233  7.650658

                 DUB        CLA       MUL        CLO        BEL        MAL
date
1961-12-31   9.733923   8.858788  8.647652   9.835577  13.502795  13.680773
1962-12-31  11.020712   8.793753  8.316822   9.676247  12.930685  14.323956
1963-12-31  11.075699  10.336548  8.903589  10.224438  13.638877  14.999014
1964-12-31  10.259153   9.467350  7.789016  10.207951  13.740546  14.910301
1965-12-31  10.618712   8.879918  7.907425   9.918082  12.964247  15.591644
1966-12-31  10.579808   8.835096  8.514438   9.768959  14.265836  16.307260
1967-12-31  10.652027   9.325616  8.645014   9.547425  14.774548  17.135945
1968-12-31   8.859180   8.255519  7.224945   7.832978  12.808634  15.017486
1969-12-31   8.564493   7.711397  7.924521   7.754384  12.621233  15.762904
1970-12-31   9.609890   8.334630  9.297616   8.289808  13.183644  16.456027
1971-12-31   8.385890   6.757315  7.915370   7.229753  12.208932  15.025233
1972-12-31   9.704508   7.680792  8.357295   7.515273  12.727377  15.028716
1973-12-31   8.482110   7.614274  8.245534   7.812411  12.169699  15.441096
1974-12-31  10.084603   9.896986  9.331753   8.736356  13.252959  16.947671
1975-12-31   8.562603   7.843836  8.797945   7.382822  12.631671  15.307863
1976-12-31   9.149126   7.146202  8.883716   7.883087  12.332377  15.471448
1977-12-31  11.523205   8.378384  9.098192   8.821616  13.459068  16.590849
1978-12-31   9.489342   8.800466  9.089753   8.301699  12.967397  16.771370
```

Step 13. Downsample the record to a monthly frequency for each location.

```
step13 = data.resample('ME').mean()
print(step13)
```

```
                 RPT        VAL        ROS       KIL        SHA        BIR  \
date
1961-01-31  14.841333  11.988333  13.431613  7.736774  11.072759   8.588065
1961-02-28  16.269286  14.975357  14.441481  9.230741  13.852143  10.937500
1961-03-31  10.890000  11.296452  10.752903  7.284000  10.509355   8.866774
1961-04-30  10.722667   9.427667   9.998000  5.830667   8.435000   6.495000
1961-05-31   9.860968   8.850000  10.818065  5.905333   9.490323   6.574839
...               ...        ...        ...       ...        ...        ...
1978-08-31   9.645161   8.259355   9.032258  4.502903   7.368065   5.935161
1978-09-30  10.913667  10.895000  10.635000  5.725000  10.372000   9.278333
1978-10-31   9.897742   8.670968   9.295806  4.721290   8.525161   6.774194
1978-11-30  16.151667  14.802667  13.508000  7.317333  11.475000   8.743000
1978-12-31  16.175484  13.748065  15.635161  7.094839  11.398710   9.241613

                 DUB        CLA        MUL        CLO        BEL        MAL
date
1961-01-31  11.184839   9.245333   9.085806  10.107419  13.880968  14.703226
1961-02-28  11.890714  11.846071  11.821429  12.714286  18.583214  15.411786
1961-03-31   9.644194   9.829677  10.294138  11.251935  16.410968  15.720000
1961-04-30   6.925333   7.094667   7.342333   7.237000  11.147333  10.278333
1961-05-31   7.604000   8.177097   8.039355   8.499355  11.900323  12.011613
...               ...        ...        ...        ...        ...        ...
1978-08-31   5.650323   5.417742   7.241290   5.536774  10.466774  12.054194
1978-09-30  10.790333   9.583000  10.069333   8.939000  15.680333  19.391333
1978-10-31   8.115484   7.337742   8.297742   8.243871  13.776774  17.150000
1978-11-30  11.492333   9.657333  10.701333  10.676000  17.404667  20.723000
1978-12-31  12.077419  10.194839  10.616774  11.028710  13.859677  21.371613

[216 rows x 12 columns]
```

Step 14. Downsample the record to a weekly frequency for each location.

```
step14 = data.resample('W').mean()
print(step14)
```

```
                 RPT        VAL        ROS        KIL        SHA        BIR  \
date
1961-01-01  15.040000  14.960000  13.170000   9.290000        NaN   9.870000
1961-01-08  13.541429  11.486667  10.487143   6.417143   9.474286   6.435714
1961-01-15  12.468571   8.967143  11.958571   4.630000   7.351429   5.072857
1961-01-22  13.204286   9.862857  12.982857   6.328571   8.966667   7.417143
1961-01-29  19.880000  16.141429  18.225714  12.720000  17.432857  14.828571
...               ...        ...        ...        ...        ...        ...
```

```
1978-12-03  14.934286  11.232857  13.941429   5.565714  10.215714   8.618571
1978-12-10  20.740000  19.190000  17.034286   9.777143  15.287143  12.774286
1978-12-17  16.758571  14.692857  14.987143   6.917143  11.397143   7.272857
1978-12-24  11.155714   8.008571  13.172857   4.004286   7.825714   6.290000
1978-12-31  14.951429  11.801429  16.035714   6.507143   9.660000   8.620000

                 DUB        CLA        MUL        CLO        BEL        MAL
date
1961-01-01  13.670000  10.250000  10.830000  12.580000  18.500000  15.040000
1961-01-08  11.061429   6.616667   8.434286   8.497143  12.481429  13.238571
1961-01-15   7.535714   6.820000   5.712857   7.571429  11.125714  11.024286
1961-01-22   9.257143   7.875714   7.145714   8.124286   9.821429  11.434286
1961-01-29  15.528571  15.160000  14.480000  15.640000  20.930000  22.530000
...               ...        ...        ...        ...        ...        ...
1978-12-03   9.642857   7.685714   9.011429   9.547143  11.835714  18.728571
1978-12-10  14.437143  12.488571  13.870000  14.082857  18.517143  23.061429
1978-12-17  10.208571   7.967143   9.168571   8.565714  11.102857  15.562857
1978-12-24   7.798571   8.667143   7.151429   8.072857  11.845714  18.977143
1978-12-31  13.708571  10.477143  10.868571  11.471429  12.947143  26.844286

[940 rows x 12 columns]
```

Step 15. Calculate the min, max and mean windspeeds and standard deviations of the windspeeds across all locations for each week (assume that the first week starts on January 2 1961) for the first 52 weeks.

```
weekly = data.resample('W', origin=pd.Timestamp('1961-01-02')).mean()
weekly_stats = pd.DataFrame({
    'min': weekly.min(axis=1),
    'max': weekly.max(axis=1),
    'mean': weekly.mean(axis=1),
    'std': weekly.std(axis=1)
})
step15 = weekly_stats.iloc[:52]
print(step15)
```

```
                 min        max       mean       std
date
1961-01-01   9.290000  18.500000  13.018182  2.808875
1961-01-08   6.417143  13.541429   9.847659  2.601705
1961-01-15   4.630000  12.468571   8.353214  2.719649
1961-01-22   6.328571  13.204286   9.368413  2.224531
1961-01-29  12.720000  22.530000  16.958095  2.915635
1961-02-05   8.247143  16.827143  11.800357  2.807310
1961-02-12  10.774286  21.832857  15.891548  3.147412
1961-02-19   9.542857  21.167143  13.726825  3.105819
1961-02-26   8.524286  16.304286  12.604286  2.364323
1961-03-05   7.834286  17.842857  11.766766  2.535336
1961-03-12   6.881429  16.701429  10.612579  2.746233
1961-03-19   7.084286  19.350000  11.756310  3.320318
1961-03-26   6.648571  18.134286  10.462857  3.071975
1961-04-02   7.300000  13.900000  10.268433  1.883742
1961-04-09   5.958571  13.607143   9.412381  2.399840
1961-04-16   4.947143   9.482857   6.845595  1.803831
1961-04-23   7.768571  13.620000  10.146667  1.895943
1961-04-30   4.801429  10.117143   7.445000  1.864130
1961-05-07   9.952857  17.548571  13.164048  2.223149
1961-05-14   5.295714  10.421429   8.059802  1.491250
1961-05-21   4.258571  12.042857   7.470258  2.327906
1961-05-28   3.748333  11.697143   7.177956  2.370473
1961-06-04   6.310000  13.597143   9.244643  2.167131
1961-06-11   5.214286  12.250000   8.459048  2.136900
1961-06-18   6.520000  15.351429  10.173810  2.676791
1961-06-25   5.478571  17.410000  10.066548  3.494271
1961-07-02   6.507143  14.535714   9.528810  2.311594
1961-07-09   7.220000  15.987143  10.580099  2.505843
1961-07-16   8.412857  16.680000  11.666190  2.483875
1961-07-23   2.715714   8.415714   5.350952  1.726225
1961-07-30   5.727143  13.761429   9.431071  2.279097
1961-08-06   6.238571  13.760000   9.446786  2.099734
1961-08-13   5.078571  10.934286   8.199206  1.855421
1961-08-20   8.600000  16.626667  12.668413  2.466378
1961-08-27   7.108571  16.485714  12.061786  2.649263
1961-09-03   5.642857  13.664286   8.437381  2.300787
1961-09-10   5.685714  11.034286   8.186786  1.866173
1961-09-17  10.442857  19.878571  14.510833  3.058931
1961-09-24   4.851667  11.018571   7.247937  1.958569
1961-10-01   8.370000  16.208571  11.771091  2.750518
1961-10-08   5.262857  11.410000   8.034167  2.148981
1961-10-15   6.578571  15.260000   9.828115  2.680505
1961-10-22  10.721429  23.641429  15.479643  3.509702
```

```
1961-10-29   8.408571  18.404286  12.688631  3.139748
1961-11-05   7.541429  19.195714  11.612857  3.253959
1961-11-12   4.220000  10.858571   7.487262  2.516385
1961-11-19   4.501429  16.988571   7.647024  3.534858
1961-11-26   4.970000  12.732857   8.755516  2.472631
1961-12-03   6.245714  14.725714   9.754762  2.724343
1961-12-10   8.624286  17.362857  13.145357  2.904971
1961-12-17   8.697143  15.112857  12.703095  2.110713
1961-12-24   5.517143  15.757143   9.315714  3.174787
```