DATA WRANGLING AIR QUALITY DATASETS

WITH PYTHON
BY DR. ALVIN ANG

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

CONTENTS

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

I. AIR QUALITY DATASETS

https://www.alvinang.sg/s/Air-Quality-No-2.csv

https://www.alvinang.sg/s/Air-Quality-PM-25.csv

https://www.alvinang.sg/s/Data_Wrangling_Air_Quality_Datasets_with_Python_by_Dr_Alvin_Ang.ipynb

Screenshot for Air Quality No. 2 Long.csv:

	A	B	C	D	E	F	G
1	city	country	date.utc	location	parameter	value	unit
2	Paris	FR	2019-06-21 00:00:00+00:00	FR04014	no2	20	µg/m³
3	Paris	FR	2019-06-20 23:00:00+00:00	FR04014	no2	21.8	µg/m³
4	Paris	FR	2019-06-20 22:00:00+00:00	FR04014	no2	26.5	µg/m³
5	Paris	FR	2019-06-20 21:00:00+00:00	FR04014	no2	24.9	µg/m³
6	Paris	FR	2019-06-20 20:00:00+00:00	FR04014	no2	21.4	µg/m³
7	Paris	FR	2019-06-20 19:00:00+00:00	FR04014	no2	25.3	µg/m³
8	Paris	FR	2019-06-20 18:00:00+00:00	FR04014	no2	23.9	µg/m³
9	Paris	FR	2019-06-20 17:00:00+00:00	FR04014	no2	23.2	µg/m³
10	Paris	FR	2019-06-20 16:00:00+00:00	FR04014	no2	19	µg/m³
11	Paris	FR	2019-06-20 15:00:00+00:00	FR04014	no2	19.3	µg/m³

Screenshot for Air Quality pm25 long.csv

	A	B	C	D	E	F	G
1	city	country	date.utc	location	parameter	value	unit
2	Antwerpen	BE	2019-06-18 06:00:00+00:00	BETR801	pm25	18	µg/m³
3	Antwerpen	BE	2019-06-17 08:00:00+00:00	BETR801	pm25	6.5	µg/m³
4	Antwerpen	BE	2019-06-17 07:00:00+00:00	BETR801	pm25	18.5	µg/m³
5	Antwerpen	BE	2019-06-17 06:00:00+00:00	BETR801	pm25	16	µg/m³
6	Antwerpen	BE	2019-06-17 05:00:00+00:00	BETR801	pm25	7.5	µg/m³
7	Antwerpen	BE	2019-06-17 04:00:00+00:00	BETR801	pm25	7.5	µg/m³
8	Antwerpen	BE	2019-06-17 03:00:00+00:00	BETR801	pm25	7	µg/m³

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

II. CONCATENATING THE TWO CSVS

A. IMPORTING AIR QUALITY DATASETS

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

B. CONCATENATING AQ2 AND AQ25 TOGETHER

A2) Concatenating AQ2 and AQ25 Together

```
[4] #Slice out Date / Location / Parameter / Value

    AQ2 = AQ2[["date.utc", "location", "parameter", "value"]]
    AQ2.sample()
```

	date.utc	location	parameter	value
1179	2019-06-13 22:00:00+00:00	London Westminster	no2	15.0

```
#Slice out Date / Location / Parameter / Value

AQ25 = AQ25[["date.utc", "location", "parameter", "value"]]
AQ25.sample()
```

	date.utc	location	parameter	value
865	2019-05-17 15:00:00+00:00	London Westminster	pm25	10.0

```
[6] AQ = pd.concat([AQ25, AQ2], axis = 0)
```

```
AQ
```

	date.utc	location	parameter	value
0	2019-06-18 06:00:00+00:00	BETR801	pm25	18.0
1	2019-06-17 08:00:00+00:00	BETR801	pm25	6.5
2	2019-06-17 07:00:00+00:00	BETR801	pm25	18.5
3	2019-06-17 06:00:00+00:00	BETR801	pm25	16.0
4	2019-06-17 05:00:00+00:00	BETR801	pm25	7.5
...	...	...	...	...
2063	2019-05-07 06:00:00+00:00	London Westminster	no2	26.0
2064	2019-05-07 04:00:00+00:00	London Westminster	no2	16.0
2065	2019-05-07 03:00:00+00:00	London Westminster	no2	19.0
2066	2019-05-07 02:00:00+00:00	London Westminster	no2	19.0
2067	2019-05-07 01:00:00+00:00	London Westminster	no2	23.0

3178 rows × 4 columns

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

C. LOOKING AT THE SHAPE OF THE CONCATTED TABLE

A3) Looking at the Shape of the Concatted Table

```
print("Shape of AQ2: ", AQ2.shape)
print("Shape of AQ25: ", AQ25.shape)
print("Shape of AQ: ", AQ.shape)

#2068 rows + 1110 rows = 3178 rows
#note: 4 columns stick to 4 columns even after concatenating

#in short, this means that Concatenating = Union both tables
#(top down) paste --> AQ25 paste on top of AQ2
```

```
Shape of AQ2:  (2068, 4)
Shape of AQ25:  (1110, 4)
Shape of AQ:  (3178, 4)
```


COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

III. INNER JOIN

A. MERGE BASED ON LOCATION

B. Inner Join

B1) Merge based on Location

```
[9] #AQ2 --> Air Quality No. 2.csv
    #AQ25 --> Air Quality PM 25.csv

    AQ_Merge_1 = pd.merge(AQ25, AQ2, on='location')
    AQ_Merge_1.head()
```

	date.utc_x	location	parameter_x	value_x	date.utc_y	parameter_y	value_y
0	2019-06-18 06:00:00+00:00	BETR801	pm25	18.0	2019-06-17 08:00:00+00:00	no2	41.0
1	2019-06-18 06:00:00+00:00	BETR801	pm25	18.0	2019-06-17 07:00:00+00:00	no2	45.0
2	2019-06-18 06:00:00+00:00	BETR801	pm25	18.0	2019-06-17 06:00:00+00:00	no2	43.5
3	2019-06-18 06:00:00+00:00	BETR801	pm25	18.0	2019-06-17 05:00:00+00:00	no2	42.5
4	2019-06-18 06:00:00+00:00	BETR801	pm25	18.0	2019-06-17 04:00:00+00:00	no2	39.5

B. MERGE BASED ON DATE

B2) Merge based on Date

```
[10] AQ_Merge_2 = pd.merge(AQ25, AQ2, on='date.utc')
     AQ_Merge_2.head()
```

	date.utc	location_x	parameter_x	value_x	location_y	parameter_y	value_y
0	2019-06-18 06:00:00+00:00	BETR801	pm25	18.0	FR04014	no2	51.4
1	2019-06-18 06:00:00+00:00	London Westminster	pm25	7.0	FR04014	no2	51.4
2	2019-06-17 08:00:00+00:00	BETR801	pm25	6.5	FR04014	no2	51.6
3	2019-06-17 08:00:00+00:00	BETR801	pm25	6.5	BETR801	no2	41.0
4	2019-06-17 08:00:00+00:00	BETR801	pm25	6.5	London Westminster	no2	13.0

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

IV. DATE TIME FORMATTING

A. RENAMING THE COLUMN 'DATE.UTC' TO "DATE TIME"

C. Date-Time Formatting

C1) Renaming the Column 'Date.utc' to "DateTime"

```
import pandas as pd

AQ2 = pd.read_csv("https://www.alvinang.sg/s/Air-Quality-No-2.csv")
```

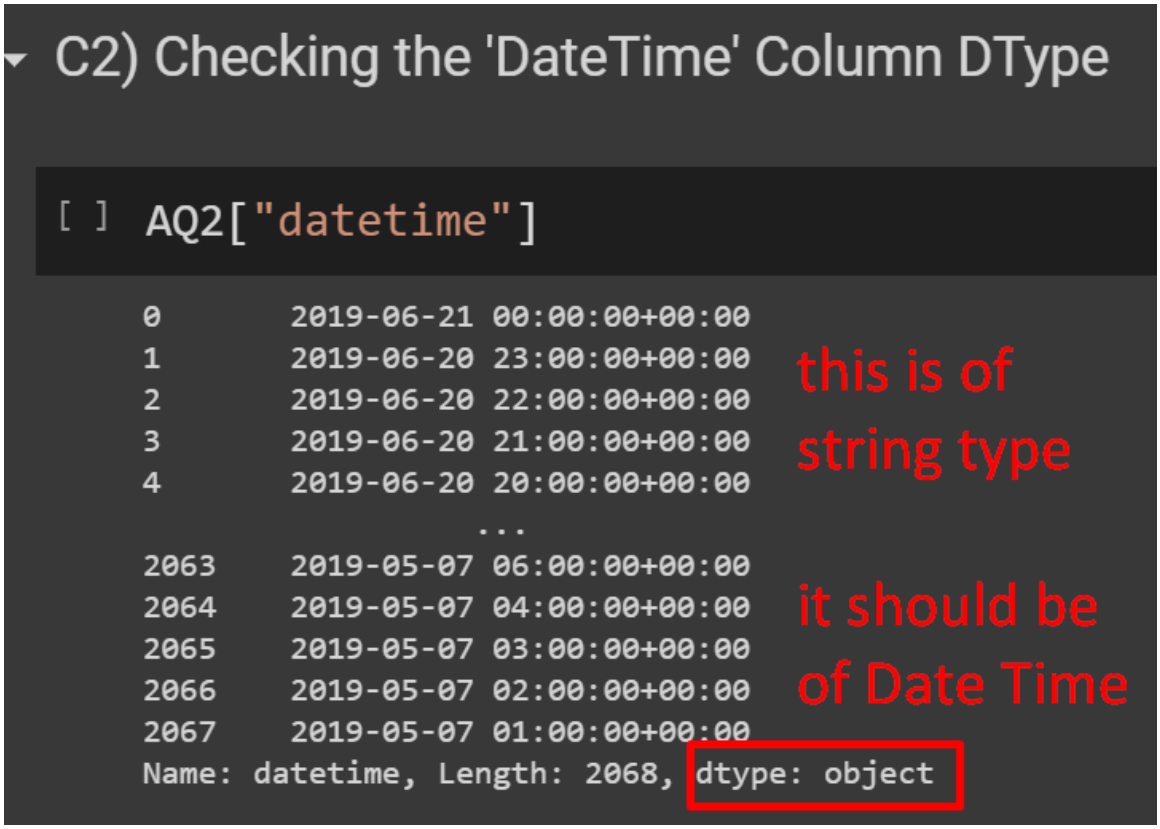
```
[13] AQ2 = AQ2.rename(columns = {"date.utc":"datetime"})
```

```
AQ2.head()
```

	city	country	datetime	location	parameter	value	unit
0	Paris	FR	2019-06-21 00:00:00+00:00	FR04014	no2	20.0	µg/m³
1	Paris	FR	2019-06-20 23:00:00+00:00	FR04014	no2	21.8	µg/m³
2	Paris	FR	2019-06-20 22:00:00+00:00	FR04014	no2	26.5	µg/m³
3	Paris	FR	2019-06-20 21:00:00+00:00	FR04014	no2	24.9	µg/m³
4	Paris	FR	2019-06-20 20:00:00+00:00	FR04014	no2	21.4	µg/m³

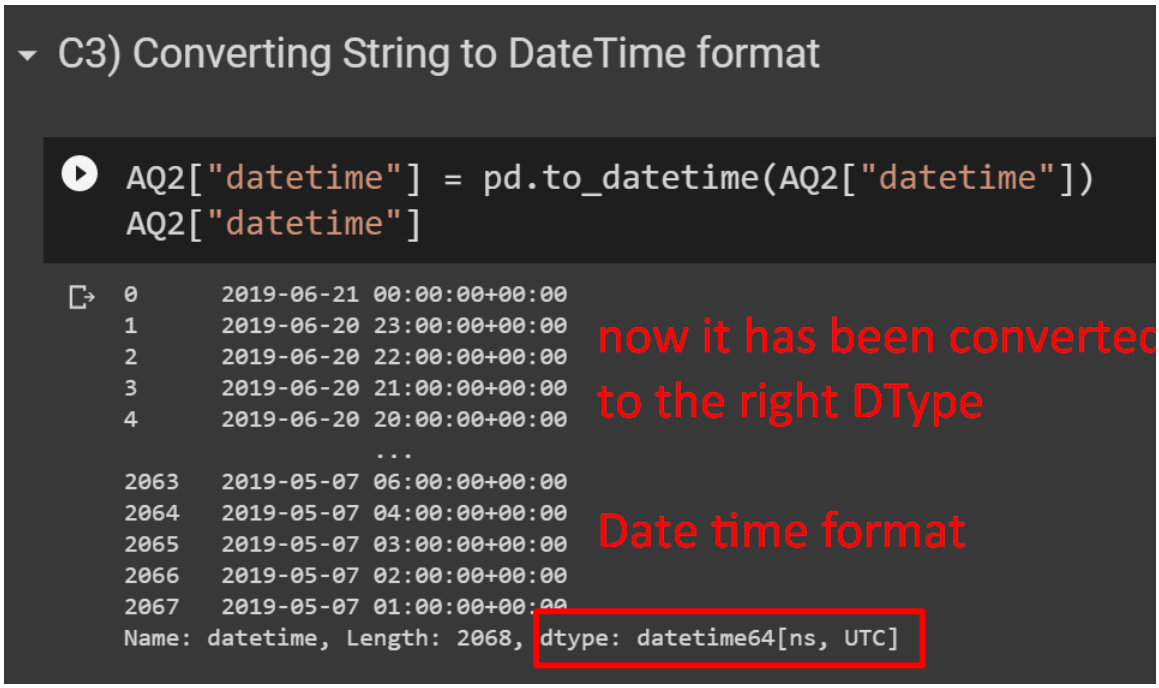
COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

B. CHECKING THE 'DATETIME' COLUMN DTYPE



COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

C. CONVERTING STRING TO DATE TIME FORMAT



COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

D. FINDING THE EARLIEST AND LATEST DATES

C4) Finding the Earliest and Latest Dates

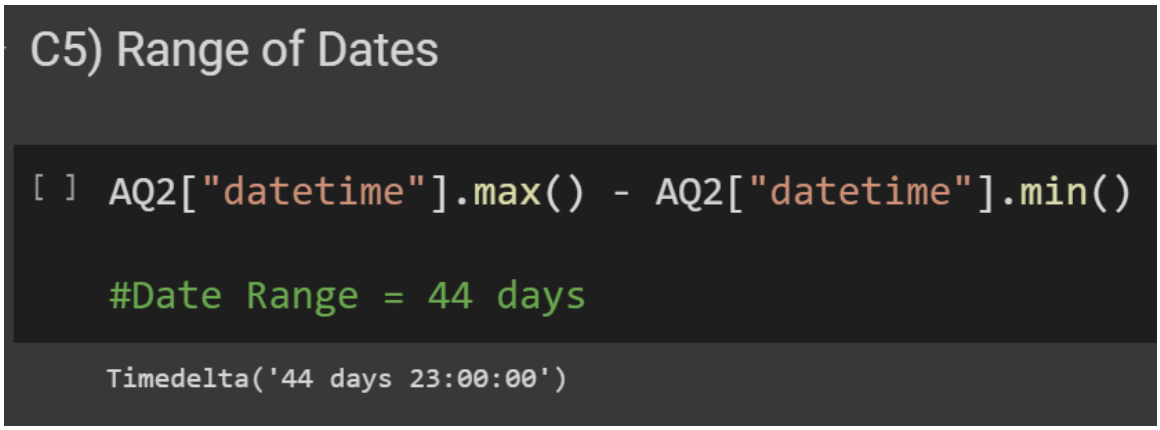
```
AQ2["datetime"].min(),\
AQ2["datetime"].max()

#earliest date = 7/5/2019
#latest date = 21/6/2019
```

```
(Timestamp('2019-05-07 01:00:00+0000', tz='UTC'),
 Timestamp('2019-06-21 00:00:00+0000', tz='UTC'))
```


COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

E. RANGE OF DATES



COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

F. CREATING A NEW COLUMN CALLED MONTH

C6) Creating a New Month Column

```
[ ] AQ2["month"] = AQ2["datetime"].dt.month
```

```
AQ2

#a new month column has been created!
```

new month column

	city	country	datetime	location	parameter	value	unit	month
0	Paris	FR	2019-06-21 00:00:00+00:00	FR04014	no2	20.0	µg/m³	6
1	Paris	FR	2019-06-20 23:00:00+00:00	FR04014	no2	21.8	µg/m³	6
2	Paris	FR	2019-06-20 22:00:00+00:00	FR04014	no2	26.5	µg/m³	6
3	Paris	FR	2019-06-20 21:00:00+00:00	FR04014	no2	24.9	µg/m³	6
4	Paris	FR	2019-06-20 20:00:00+00:00	FR04014	no2	21.4	µg/m³	6
...	...	...	...	...	...	...	...	...
2063	London	GB	2019-05-07 06:00:00+00:00	London Westminster	no2	26.0	µg/m³	5
2064	London	GB	2019-05-07 04:00:00+00:00	London Westminster	no2	16.0	µg/m³	5
2065	London	GB	2019-05-07 03:00:00+00:00	London Westminster	no2	19.0	µg/m³	5
2066	London	GB	2019-05-07 02:00:00+00:00	London Westminster	no2	19.0	µg/m³	5
2067	London	GB	2019-05-07 01:00:00+00:00	London Westminster	no2	23.0	µg/m³	5

2068 rows × 8 columns

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

G. AVERAGE WEEKDAY AIR QUALITY (GROUPBY LOCATION)

C7) Average Weekday Air Quality (Groupby Location)

```
[ ] AQ2.groupby([AQ2["datetime"].dt.weekday, "location"])["value"].mean()
```

```
datetime  location
0         BETR801              27.875000
          FR04014              24.856250
          London Westminster   23.969697
1         BETR801              22.214286
          FR04014              30.999359
          London Westminster   24.885714
2         BETR801              21.125000
          FR04014              29.165753
          London Westminster   23.460432
3         BETR801              27.500000
          FR04014              28.600690
          London Westminster   24.780142
4         BETR801              28.400000
          FR04014              31.617986
          London Westminster   26.446809
5         BETR801              33.500000
          FR04014              25.266154
          London Westminster   24.977612
6         BETR801              21.896552
          FR04014              23.274306
          London Westminster   24.859155
Name: value, dtype: float64
```

THE END

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

ABOUT DR. ALVIN ANG

Dr. Alvin Ang earned his Ph.D., Masters and Bachelor degrees from NTU, Singapore. He is a scientist, entrepreneur, as well as a personal/business advisor. More about him at www.AlvinAng.sg.

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG