Data Quality Report

Index	Nombres	tipo	Valores perdidos	Valores presentes	Valores unicos	Min	Max
Fecha	Fecha	datetime64[n…	0	8760	8045	2015-01-01 00:00:00	2015-12-31 00:00:00
Hora	Hora	object	0	8760	8038	nan	nan
CO	CO	float64	563	8197	2227	0.013	7.467
NO2	NO2	float64	216	8544	2799	0	0.10793
O3	O3	float64	152	8608	3129	0.00115	0.11758
PM10	PM10	float64	244	8516	4624	1.97	263.95
SO2	SO2	float64	138	8622	634	0	0.02968
Unnamed: 7	Unnamed: 7	object	8759	1	1	nan	nan

CO

Índice	Tipo	Tamaño	
0	str	1	Atemajac
1	str	1	Aguilas
2	str	1	Centro
3	str	1	Las Pintas
4	str	1	Loma Dorada
5	str	1	Miravalle
6	str	1	Oblatos
7	str	1	Santa Fe
8	str	1	Tlaquepaque
9	str	1	Vallarta

	0	1	2	3	4	5	6	7	8	9
0	0	41.452	41.2159	45.8992	40.7223	37.4536	44.6021	44.8172	40.9096	51.3707
1	41.452	0	39.4921	37.1437	33.9661	31.1356	37.1977	42.6012	34.8083	31.3483
2	41.2159	39.4921	0	39.1948	37.5856	35.2726	34.3736	44.7654	31.4247	44.414
3	45.8992	37.1437	39.1948	0	33.5505	30.5496	42.298	36.7643	26.8796	37.8597
4	40.7223	33.9661	37.5856	33.5505	0	23.7046	39.5589	35.8517	29.3497	42.1991
5	37.4536	31.1356	35.2726	30.5496	23.7046	0	38.8875	32.8543	25.3541	42.6083
6	44.6021	37.1977	34.3736	42.298	39.5589	38.8875	0	47.0706	35.9512	43.4022
7	44.8172	42.6012	44.7654	36.7643	35.8517	32.8543	47.0706	0	34.9203	49.7131
8	40.9096	34.8083	31.4247	26.8796	29.3497	25.3541	35.9512	34.9203	0	39.4974
9	51.3707	31.3483	44.414	37.8597	42.1991	42.6083	43.4022	49.7131	39.4974	0

Marcados en Negro tenemos los más similares, en este caso son 4-5,5-8,3-8, que corresponden a Loma Dorada con Miravalle, Tlaquepaque con Miravalle y Las pintas con Tlaquepaque. Todavía no se sabe cuál es el parecido entre estas estaciones (pueden ser muy contaminadas, poco contaminadas o medianamente contaminadas, pero resulta ser que todas estas tienen índices muy similares.

PM10

Índice	Tipo	Tamaño	
0	str	1	Atemajac
1	str	1	Aguilas
2	str	1	Centro
3	str	1	Las Pintas
4	str	1	Loma Dorada
5	str	1	Miravalle
6	str	1	Oblatos
7	str	1	Santa Fe
8	str	1	Tlaquepaque
9	str	1	Vallarta

	0	1	2	3	4	5	6	7	8	9
0	0	1910.57	1077.39	3470.6	1061.89	2944.84	824.993	4729.62	2181.34	996.677
1	1910.57	0	2041.52	3803.99	1919.98	3307.64	1861.27	4907.33	2649.52	1895.09
2	1077.39	2041.52	0	3147.93	1319.97	2647.61	987.354	4545.84	1825.5	1305.43
3	3470.6	3803.99	3147.93	0	3586.58	2784.56	3588.99	3453.84	2385.45	3929.64
4	1061.89	1919.98	1319.97	3586.58	0	3106.02	990.551	4846.7	2275.92	1111.87
5	2944.84	3307.64	2647.61	2784.56	3106.02	0	3016.85	3479.17	2182.18	3302.79
6	824.993	1861.27	987.354	3588.99	990.551	3016.85	0	4868.4	2206.77	871.896
7	4729.62	4907.33	4545.84	3453.84	4846.7	3479.17	4868.4	0	3734.13	5172.14
8	2181.34	2649.52	1825.5	2385.45	2275.92	2182.18	2206.77	3734.13	0	2596.23
9	996.677	1895.09	1305.43	3929.64	1111.87	3302.79	871.896	5172.14	2596.23	0

De igual manera se hace la tabla de distancias euclidianas con PM10 y se puede notar que los más similares son Oblatos-Atemajac, Vallarta-Oblatos, Oblatos-Centro. Importante recalcar que Oblatos aparece en los tres menores, parece ser que Oblatos es representativo de algunas regiones cercanas.

Metodología:

Se utilizó como método de similitud la distancia euclidiana. La función de esta es la raíz de la suma de los cuadrados de las diferencias de cada uno de los elementos de los factores.

$$\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}.$$

Este tipo de distancia, también conocida como distancia geométrica es literalmente la distancia entre un punto y otro (Pitágoras si se hablara de un espacio bidimensional)

Código:

```
7
8  import numpy as np
9  import matplotlib.pyplot as plt
10 import scipy.spatial.distance as sc
11 import pandas as pd
12 from cdin import cdinp19 as cd
13
14 #%% Leer datos
15 estaciones = ['Atemajac','Aguilas','Centro','Las Pintas','Loma Dorada','Miravalle','Oblatos','Santa Fe','Tlaquepaque','Vallarta']
16 data = []
17 for estacion in estaciones:
18     tmp = pd.read_excel('../data/contaminacion_2015.xlsx',sheet_name=estacion,index_col=1)
19     tmp = tmp.iloc[:,1:6]
20     data.append(tmp)
21
22 #%% Data Quality Report
23 dqr = cd.dqr(data[0])
24
25 #%% Distancia entre CO's
26 Co = np.zeros((8016,len(data)))
27 for i in range(len(data)):
28     Co[:,i] = data[i].CO[:8016].values
29 Co = pd.DataFrame(Co)
30 Co = Co.dropna()
31
32 #%%
33 Mat_co = sc.squareform(sc.pdist(Co.T,'euclidean'))
34
35 #%%
36 Pm10 = np.zeros((8016,len(data)))
37 for i in range(len(data)):
38     Pm10[:,i] = data[i].PM10[:8016].values
39 Pm10 = pd.DataFrame(Pm10)
40 Pm10 = Pm10.dropna()
41
42 #%%
43 Mat_pm10 = sc.squareform(sc.pdist(Pm10.T,'euclidean'))
44
```