Medidas de Similitud o Similaridad

¿Similitud?

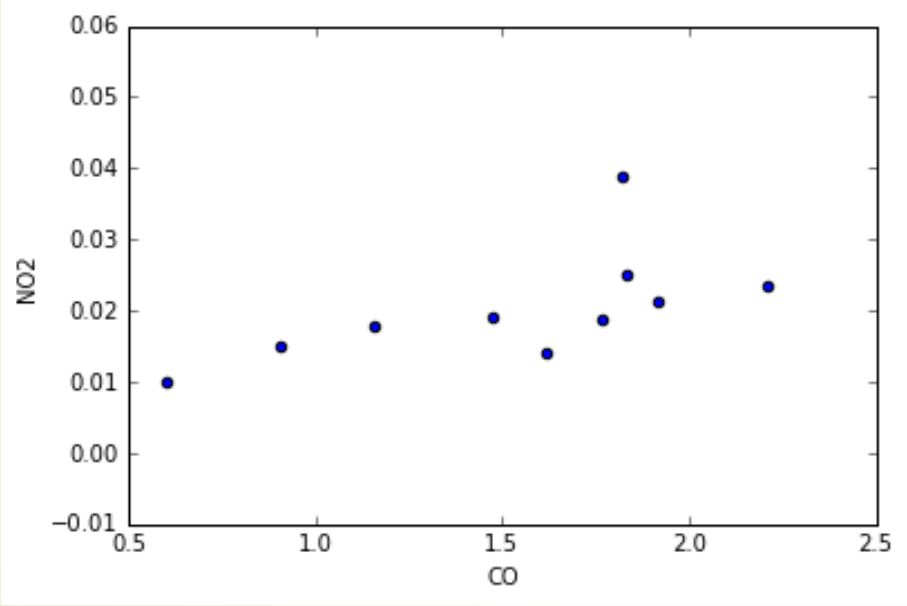
- Similitud se refiere a una forma de medir que tan parecido es un conjunto de datos con respecto a otro.
- El cálculo de la similitud depende del tipo de variable, y del significado o contexto de la misma.
- Las medidas de similitud con más frecuencia son utilizadas para variables con datos cuantitativos y datos binarios o de dos estados.
- Las variables categóricas pueden ser analizadas como variables binarias si se les aplica una transformación.

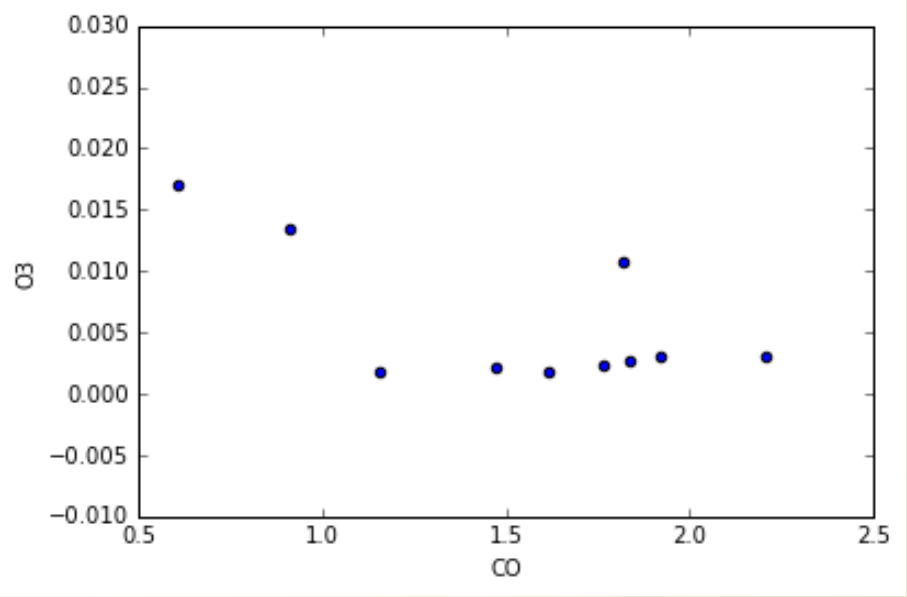
Similitud para datos cuantitativos

CO	NO2	O3	PM10	SO2
0.603	0.01005	0.01695	43.95	0.00165
0.909	0.0151	0.01347	50.04	0.0019
1.835	0.02492	0.00273	46.23	0.00252
2.208	0.02338	0.00307	57.83	0.00298
1.473	0.01925	0.00223	56.66	0.00245
1.156	0.01782	0.00185	46.52	0.00182
1.766	0.01887	0.0024	40.54	0.00205
1.617	0.01397	0.00188	60.1	0.00208
1.917	0.02127	0.00308	57.22	0.00248
1.819	0.03882	0.01082	82.99	0.00228

- CO (**Monóxido de carbono**)
- NO2 (**Dióxido de Nitrógeno**)
- O3 (**Ozono a nivel del suelo**)
- PM10 (**Material Particulado 10 micrómetros**)
- SO2 (**Dióxido de azufre**)

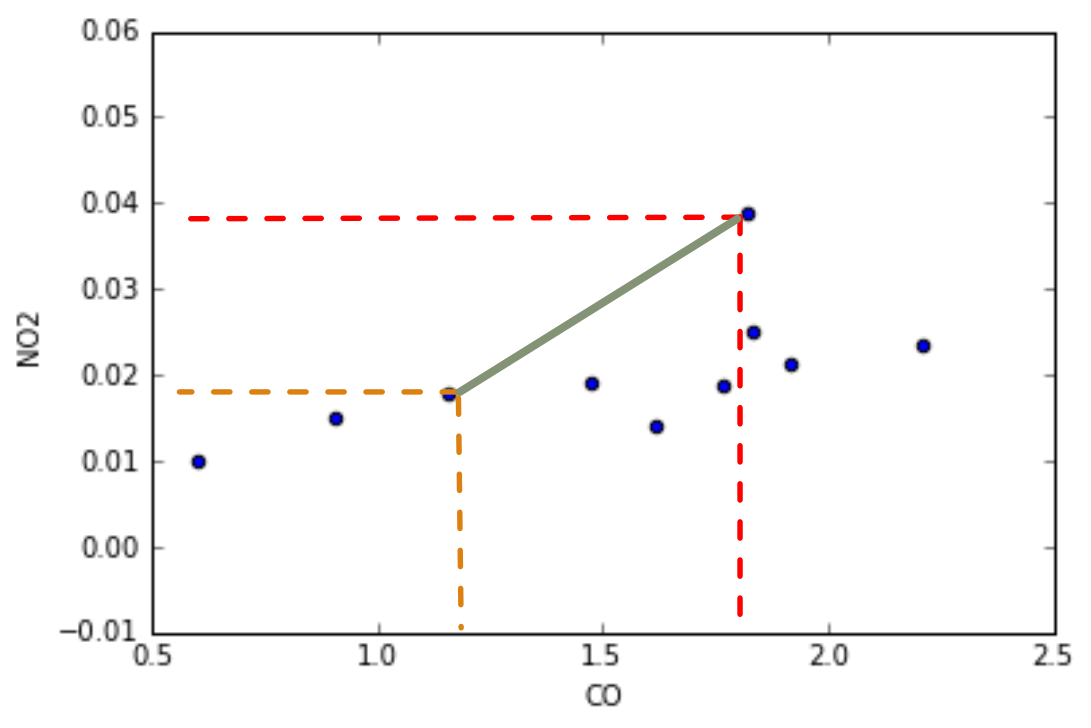
Comparación mediante gráficos





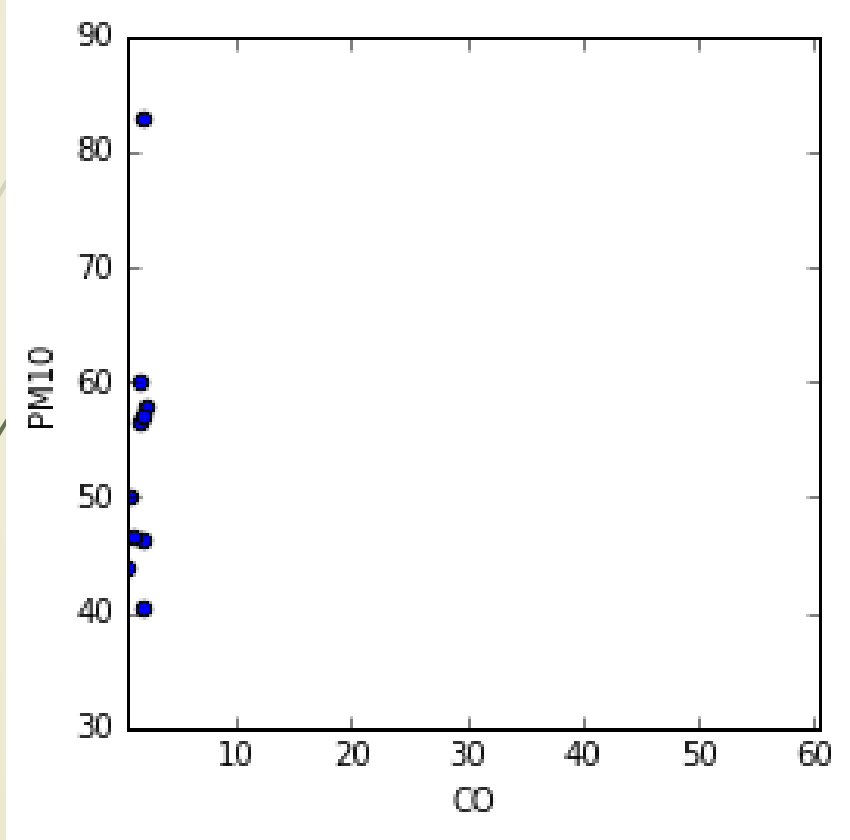
Distancia Euclideana

Distancia Euclideana



- La distancia Euclideana es la generalización del teorema de Pitágoras para más de 2 variables.
- $d = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}$
- Esta medida de similitud puede ser usada para determinar cercanía entre dos muestras o dos variables.
- El problema principal se presenta cuando se tienen variables donde las escalas no son comparables o son muy diferentes.

Distancia Euclideana



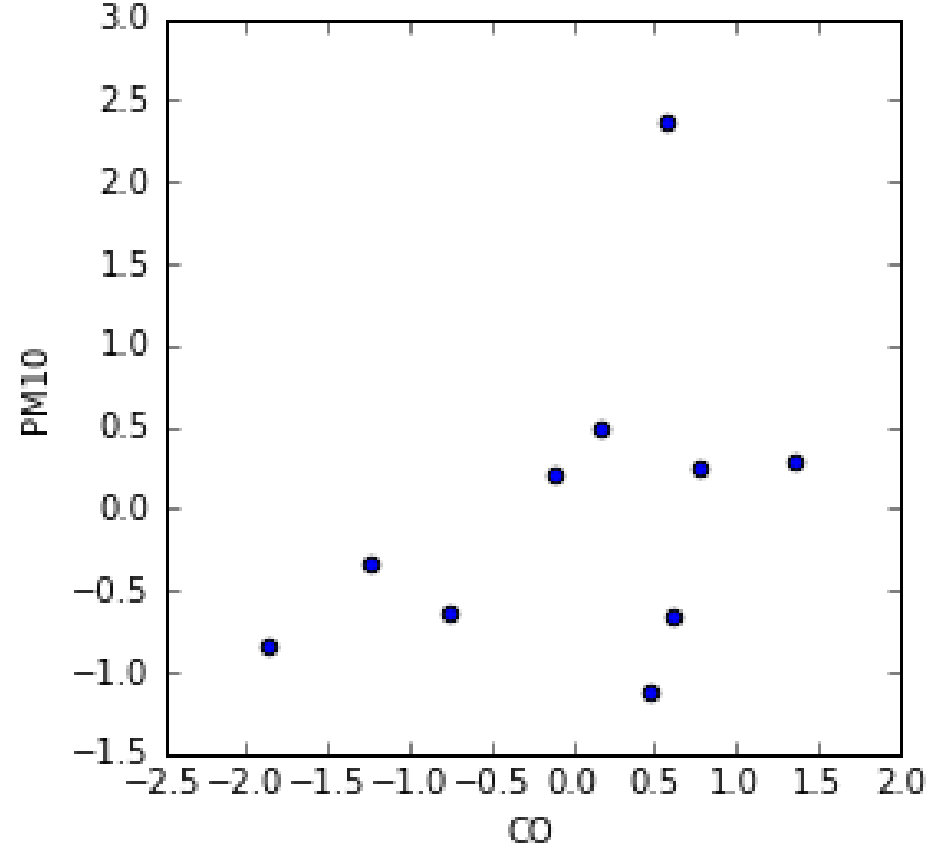
- Una solución es estandarizar los datos de la forma siguiente:
- $x^* = \frac{x-\mu_x}{\sigma_x}$

$$\mu = \begin{bmatrix} \mu_{CO} \\ \mu_{PM10} \end{bmatrix} = \begin{bmatrix} 1.5303 \\ 54.208 \end{bmatrix}$$

CO	PM10
0.603	43.95
0.909	50.04
1.835	46.23
2.208	57.83
1.473	56.66
1.156	46.52
1.766	40.54
1.617	60.1
1.917	57.22
1.819	82.99

CO	PM10
-1.85983221	-0.84590915
-1.24610563	-0.34370729
0.61111924	-0.65789269
1.35922386	0.29868229
-0.11492331	0.20220016
-0.75071195	-0.63397831
0.47272992	-1.12710921
0.1738892	0.48587412
0.77558192	0.24837964
0.57902896	2.37346043

Distancia Euclideana



- Calculando la distancia Euclideana a los nuevos datos nos puede dar información que no tiene mucha dependencia de las escalas de los datos.

- $d^* = \sqrt{(x_0^* - x_1^*)^2 + (y_0^* - y_1^*)^2}$

- Ha esta distancia se le llama distancia Euclideana Estandarizada.

Distancia Euclideana

- Una medida equivalente a la distancia Euclideana Estandarizada es la distancia Euclideana ponderada.

- $d = \sqrt{\frac{1}{\sigma_x^2}(x_0 - x_1)^2 + \frac{1}{\sigma_y^2}(y_0 - y_1)^2}$

Matriz de Similitud

Matriz de similaridad de variables

	V0	V1	V2	V3	V4
V0	**0**	**2.59738013**	**5.47928451**	**3.29339382**	**1.5902661**
V1	**2.59738013**	**0**	**4.46595328**	**2.15890939**	**2.9584916**
V2	**5.47928451**	**4.46595328**	**0**	**4.06882391**	**5.21241236**
V3	**3.29339382**	**2.15890939**	**4.06882391**	**0**	**3.42203797**
V4	**1.5902661**	**2.9584916**	**5.21241236**	**3.42203797**	**0**

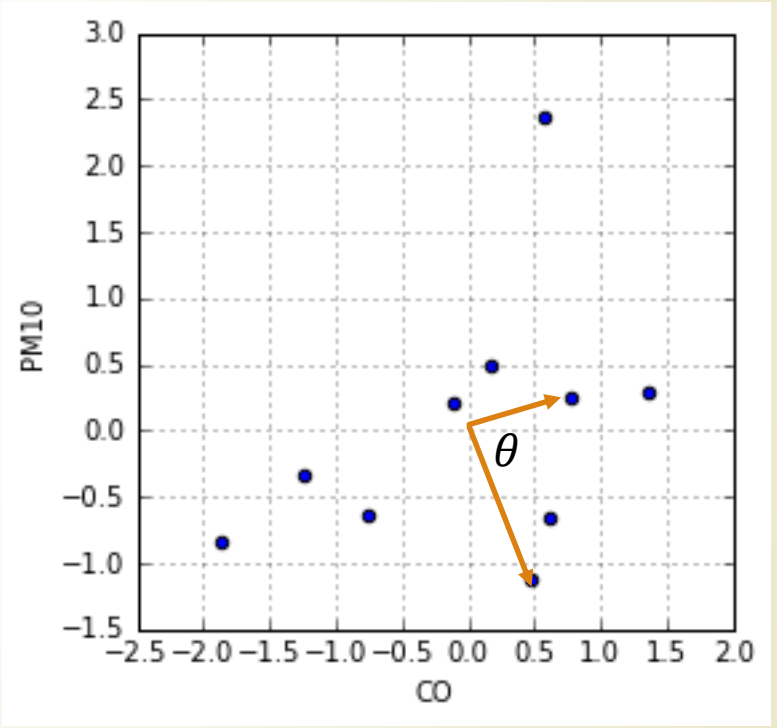
Matriz de similaridad de muestras

	S0	S1	S2	S3	S4	S5	S6	S7	S8	S9
S0	0	1.34691437	4.56208064	5.63232432	4.04003694	3.0923247	3.79551369	3.79790485	4.53183541	5.77777361
S1	1.34691437	0	3.33737108	4.3640639	2.77725954	2.1704896	2.78766751	2.67002765	3.24801882	4.57690636
S2	4.56208064	3.33737108	0	1.68842376	1.35122355	2.40507474	1.49255027	2.1639537	1.03820279	3.83740983
S3	5.63232432	4.3640639	1.68842376	0	2.06191289	3.78842043	2.93511604	2.83266401	1.411171	3.70646943
S4	4.04003694	2.77725954	1.35122355	2.06191289	0	1.90876627	1.76759819	1.21734734	0.94308993	3.7258612
S5	3.0923247	2.1704896	2.40507474	3.78842043	1.90876627	0	1.44941613	1.66584184	2.46867957	4.6688576
S6	3.79551369	2.78766751	1.49255027	2.93511604	1.76759819	1.44941613	0	1.75887818	1.80474361	4.61034677
S7	3.79790485	2.67002765	2.1639537	2.83266401	1.21734734	1.66584184	1.75887818	0	1.52790738	4.05833101
S8	4.53183541	3.24801882	1.03820279	1.411171	0.94308993	2.46867957	1.80474361	1.52790738	0	3.41517986
S9	5.77777361	4.57690636	3.83740983	3.70646943	3.7258612	4.6688576	4.61034677	4.05833101	3.41517986	0

Similaridad basada en el coseno

Similaridad basado en coseno

- La similitud calculada en base al coseno tiene como fundamento el algebra vectorial.
- El producto punto esta definido como: $a \cdot b = \|A\|\|B\|\cos\theta$, donde A y B son vectores.
- La similaridad es calculada como: $\text{similaridad} = cos\theta = \frac{a \cdot b}{\|A\|\|B\|}$
- La distancia es calculada como: $d = 1 - similaridad = 1 - \frac{a \cdot b}{\|A\|\|B\|}$



Similaridad basado en coseno

Matriz de similaridad de variables

	V0	V1	V2	V3	V4
V0	**0**	**0.37479909**	**1.66791993**	**0.60258016**	**0.14049701**
V1	**0.37479909**	**0**	**1.10804104**	**0.25893832**	**0.48625959**
V2	**1.66791993**	**1.10804104**	**0**	**0.91974044**	**1.50940237**
V3	**0.60258016**	**0.25893832**	**0.91974044**	**0**	**0.65057466**
V4	**0.14049701**	**0.48625959**	**1.50940237**	**0.65057466**	**0**

Matriz de similaridad de muestras

	S0	S1	S2	S3	S4	S5	S6	S7	S8	S9
S0	0	0.01241374	1.31617283	1.97791461	0.91015244	0.03742016	0.97027353	1.69652992	1.9931709	1.61796351
S1	0.01241374	0	1.46126773	1.99860501	0.75482546	0.09193713	1.12765075	1.57517583	1.9991796	1.48679678
S2	1.31617283	1.46126773	0	0.49252674	1.97327197	1.04724513	0.06111635	1.46049589	0.57530269	1.55049237
S3	1.97791461	1.99860501	0.49252674	0	1.29602277	1.88468094	0.82015747	0.46882	0.00437937	0.56000548
S4	0.91015244	0.75482546	1.97327197	1.29602277	0	1.18341674	1.99284454	0.34795579	1.20543069	0.27249552
S5	0.03742016	0.09193713	1.04724513	1.88468094	1.18341674	0	0.70050742	1.86491442	1.92438932	1.80789987
S6	0.97027353	1.12765075	0.06111635	0.82015747	1.99284454	0.70050742	0	1.73791603	0.91290646	1.80422921
S7	1.69652992	1.57517583	1.46049589	0.46882	0.34795579	1.86491442	1.73791603	0	0.39193968	0.00544474
S8	1.9931709	1.9991796	0.57530269	0.00437937	1.20543069	1.92438932	0.91290646	0.39193968	0	0.47798224
S9	1.61796351	1.48679678	1.55049237	0.56000548	0.27249552	1.80789987	1.80422921	0.00544474	0.47798224	0

Similaridad basado en coseno (dependencia de escala)

	V0	V1	V2	V3	V4
V0	**0**	**0.37479909**	**1.66791993**	**0.60258016**	**0.14049701**
V1	**0.37479909**	**0**	**1.10804104**	**0.25893832**	**0.48625959**
V2	**1.66791993**	**1.10804104**	**0**	**0.91974044**	**1.50940237**
V3	**0.60258016**	**0.25893832**	**0.91974044**	**0**	**0.65057466**
V4	**0.14049701**	**0.48625959**	**1.50940237**	**0.65057466**	**0**

	V0	V1	V2	V3	V4
V0	**0**	**0.03942439**	**0.4293539**	**0.0410487**	**0.01556278**
V1	**0.03942439**	**0**	**0.33330297**	**0.02860085**	**0.04473988**
V2	**0.4293539**	**0.33330297**	**0**	**0.26793425**	**0.3313351**
V3	**0.0410487**	**0.02860085**	**0.26793425**	**0**	**0.02347511**
V4	**0.01556278**	**0.04473988**	**0.3313351**	**0.02347511**	**0**

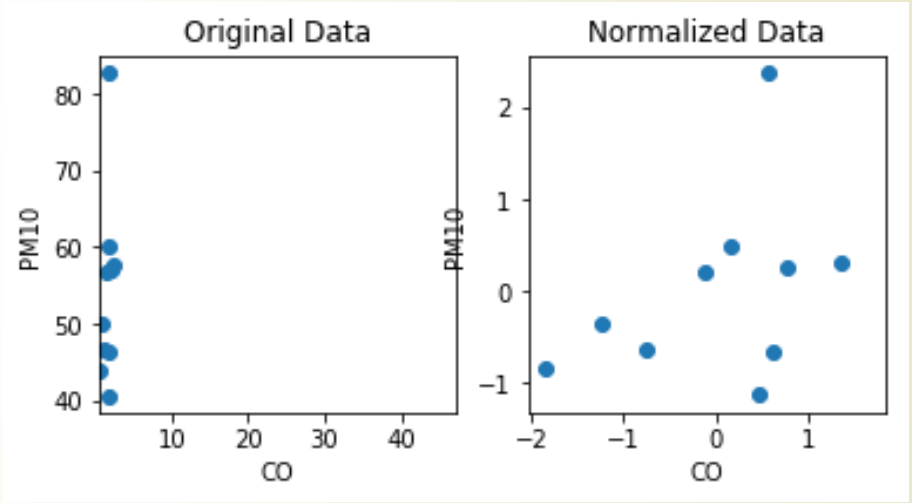
Similaridad basado en coseno (dependencia de escala)

	S0	S1	S2	S3	S4	S5	S6	S7	S8	S9
S0	0	0.01241374	1.31617283	1.97791461	0.91015244	0.03742016	0.97027353	1.69652992	1.9931709	1.61796351
S1	0.01241374	0	1.46126773	1.99860501	0.75482546	0.09193713	1.12765075	1.57517583	1.99916796	1.48679678
S2	1.31617283	1.46126773	0	0.49252674	1.97327197	1.04724513	0.06111635	1.46049589	0.57530269	1.55049237
S3	1.97791461	1.99860501	0.49252674	0	1.29602277	1.88468094	0.82015747	0.46882	0.00437937	0.56000548
S4	0.91015244	0.75482546	1.97327197	1.29602277	0	1.18341674	1.99284454	0.34795579	1.20543069	0.27249552
S5	0.03742016	0.09193713	1.04724513	1.88468094	1.18341674	0	0.70050742	1.86491442	1.92438932	1.80789987
S6	0.97027353	1.12765075	0.06111635	0.82015747	1.99284454	0.70050742	0	1.73791603	0.91290646	1.80422921
S7	1.69652992	1.57517583	1.46049589	0.46882	0.34795579	1.86491442	1.73791603	0	0.39193968	0.00544474
S8	1.9931709	1.99916796	0.57530269	0.00437937	1.20543069	1.92438932	0.91290646	0.39193968	0	0.47798224
S9	1.61796351	1.48679678	1.55049237	0.56000548	0.27249552	1.80789987	1.80422921	0.00544474	0.47798224	0

	S0	S1	S2	S3	S4	S5	S6	S7	S8	S9
S0	0	9.8754E-06	0.00033675	0.00029872	7.5301E-05	6.1884E-05	0.00044444	8.6847E-05	0.00019543	3.3583E-05
S1	9.8754E-06	0	0.0002313	0.00019997	3.0637E-05	2.2317E-05	0.00032182	3.8152E-05	0.00011745	7.0362E-06
S2	0.00033675	0.0002313	0	1.1396E-06	9.3579E-05	0.00010993	7.459E-06	8.1578E-05	1.911E-05	0.00015766
S3	0.00029872	0.00019997	1.1396E-06	0	7.4066E-05	8.8682E-05	1.4429E-05	6.3434E-05	1.0917E-05	0.00013199
S4	7.5301E-05	3.0637E-05	9.3579E-05	7.4066E-05	0	6.577E-07	0.00015388	4.1164E-07	2.8113E-05	8.309E-06
S5	6.1884E-05	2.2317E-05	0.00010993	8.8682E-05	6.577E-07	0	0.00017465	2.11E-06	3.7371E-05	4.2913E-06
S6	0.00044444	0.00032182	7.459E-06	1.4429E-05	0.00015388	0.00017465	0	0.00013837	5.0447E-05	0.00023369
S7	8.6847E-05	3.8152E-05	8.1578E-05	6.3434E-05	4.1164E-07	2.11E-06	0.00013837	0	2.1721E-05	1.2419E-05
S8	0.00019543	0.00011745	1.911E-05	1.0917E-05	2.8113E-05	3.7371E-05	5.0447E-05	2.1721E-05	0	6.6989E-05
S9	3.3583E-05	7.0362E-06	0.00015766	0.00013199	8.309E-06	4.2913E-06	0.00023369	1.2419E-05	6.6989E-05	0

Similaridad basado en coseno

- La similaridad tiene una dependencia en la escala de los datos.
- Pero esa dependencia solo afecta en la magnitud de los índices de similaridad, ya que el coseno depende del ángulo entre los puntos en el espacio.
- La proporción de los índices debe de guardar la misma proporción.



Similaridad basada en correlación

Correlación

- Existen varios tipos de coeficientes de correlación:
  - Coeficiente de correlación de Pearson
  - Coeficiente de correlación de Spearman
  - Correlación canónica
  - Coeficiente de Correlación Intraclase
  - Correlación de Kendall
  - Correlación de Jaspen
- El más conocido es el de Pearson

- $\rho_{x,y} = \frac{E[(x-\mu_x)(y-\mu_y)]}{\sigma_x \sigma_y}$
- El coeficiente de correlación de Pearson no tiene dependencia de la escala de los datos, ya que implícitamente realiza la normalización de los datos.

Correlación

Matriz de similaridad de variables

	V0	V1	V2	V3	V4
V0	**0**	**0.37479909**	**1.66791993**	**0.60258016**	**0.14049701**
V1	**0.37479909**	**0**	**1.10804104**	**0.25893832**	**0.48625959**
V2	**1.66791993**	**1.10804104**	**0**	**0.91974044**	**1.50940237**
V3	**0.60258016**	**0.25893832**	**0.91974044**	**0**	**0.65057466**
V4	**0.14049701**	**0.48625959**	**1.50940237**	**0.65057466**	**0**

Matriz de similaridad de muestras

	S0	S1	S2	S3	S4	S5	S6	S7	S8	S9
S0	0	0.00478325	1.70752283	1.79085145	1.70269887	0.98354635	1.43780599	1.40425373	1.91199485	1.03725271
S1	0.00478325	0	1.73234964	1.81225543	1.66638952	0.94585996	1.51051933	1.41031466	1.93428097	0.96812732
S2	1.70752283	1.73234964	0	0.21260221	0.58375771	1.1356771	0.25212329	1.25079536	0.31479928	1.40082025
S3	1.79085145	1.81225543	0.21260221	0	0.21122927	1.579606	0.54172813	0.72862535	0.07208825	1.5442529
S4	1.70269887	1.66638952	0.58375771	0.21122927	0	1.45874345	1.1484746	0.59443504	0.26465274	1.13078433
S5	0.98354635	0.94585996	1.1356771	1.579606	1.45874345	0	1.00563954	1.27383457	1.39622728	0.1523876
S6	1.43780599	1.51051933	0.25212329	0.54172813	1.1484746	1.00563954	0	1.14290993	0.51116231	1.43717457
S7	1.40425373	1.41031466	1.25079536	0.72862535	0.59443504	1.27383457	1.14290993	0	0.47449784	0.91058237
S8	1.91199485	1.93428097	0.31479928	0.07208825	0.26465274	1.39622728	0.51116231	0.47449784	0	1.32300768
S9	1.03725271	0.96812732	1.40082025	1.5442529	1.13078433	0.1523876	1.43717457	0.91058237	1.32300768	0

Bases de Datos Heterogeneas

Bases de datos heterogéneas.

- Una base de datos heterogénea es una combinación de varios tipos de datos en una sola base de datos.
- Para la obtención de medidas de similitud, se puede hacer uso de variables auxiliares (dummy) para las variables categóricas.
- Accident_Severity y Day_of_Week

Longitude	Latitude	Number_of_Vehicles	Number_of_Casualties	Accident_Severity	Day_of_Week
-0.198465	51.505538	1	1	3	2
-0.178838	51.491836	1	1	3	2
-0.20559	51.51491	1	1	3	2
-0.208327	51.514952	1	1	3	3
-0.206022	51.496572	2	1	2	6
-0.19361	51.500788	2	1	3	5
-0.173519	51.495171	2	1	3	5
-0.163542	51.492497	2	1	3	1
-0.21198	51.513659	2	1	3	6
-0.199786	51.5159	2	2	3	3

Variables Dummy

Accident_Severity
3
3
3
3
2
3
3
3
3
3

Severity_Low	Severity_Medium	Severity_High
0	0	1
0	0	1
0	0	1
0	0	1
0	1	0
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1

Variables Dummy

Longitude	Latitude	Number_of_Vehicles	Number_of_Casualties	Severity 2	Severity 3	Sunday	Monday	Tuesday	Thursday	Friday
-0.198465	51.505538	1	1	0	1	0	1	0	0	0
-0.178838	51.491836	1	1	0	1	0	1	0	0	0
-0.20559	51.51491	1	1	0	1	0	1	0	0	0
-0.208327	51.514952	1	1	0	1	0	0	1	0	0
-0.206022	51.496572	2	1	1	0	0	0	0	0	1
-0.19361	51.500788	2	1	0	1	0	0	0	1	0
-0.173519	51.495171	2	1	0	1	0	0	0	1	0
-0.163542	51.492497	2	1	0	1	1	0	0	0	0
-0.21198	51.513659	2	1	0	1	0	0	0	0	1
-0.199786	51.5159	2	2	0	1	0	0	1	0	0

Variables Dummy

Longitude	Latitude	Number_of_ Vehicles	Number_of_ Casualties	Severity 2	Severity 3	Sunday	Monday	Tuesday	Thursday	Friday
-0.27323522	0.13559719	-1.161895	-0.31622777	-0.31622777	0.31622777	-0.31622777	1.44913767	-0.47434165	-0.47434165	-0.47434165
0.91926386	-1.23487764	-1.161895	-0.31622777	-0.31622777	0.31622777	-0.31622777	1.44913767	-0.47434165	-0.47434165	-0.47434165
-0.70613662	1.07298517	-1.161895	-0.31622777	-0.31622777	0.31622777	-0.31622777	1.44913767	-0.47434165	-0.47434165	-0.47434165
-0.87243152	1.07718602	-1.161895	-0.31622777	-0.31622777	0.31622777	-0.31622777	-0.621059	1.8973666	-0.47434165	-0.47434165
-0.73238412	-0.76118265	0.77459667	-0.31622777	2.84604989	-2.84604989	-0.31622777	-0.621059	-0.47434165	-0.47434165	1.8973666
0.02174532	-0.33949808	0.77459667	-0.31622777	-0.31622777	0.31622777	-0.31622777	-0.621059	-0.47434165	1.8973666	-0.47434165
1.24243615	-0.90131075	0.77459667	-0.31622777	-0.31622777	0.31622777	-0.31622777	-0.621059	-0.47434165	1.8973666	-0.47434165
1.84861964	-1.16876438	0.77459667	-0.31622777	-0.31622777	0.31622777	2.84604989	-0.621059	-0.47434165	-0.47434165	-0.47434165
-1.09438083	0.94786008	0.77459667	-0.31622777	-0.31622777	0.31622777	-0.31622777	-0.621059	-0.47434165	-0.47434165	1.8973666
-0.35349666	1.17200503	0.77459667	2.84604989	-0.31622777	0.31622777	-0.31622777	-0.621059	1.8973666	-0.47434165	-0.47434165

Variables Dummy

- *Coeficiente de similaridad(disimilaridad) general de Gower.*
- *Estandarizar cada variable y multiplicar todas las columnas correspondientes a variables dummy por el factor 1/sqrt(2)=0.7071*
- *Este factor compensa la codificación 0/1.*