Escuela Nacional de Estadística e Informática

SOFTWARE "R"

Lima – Perú

www.inei.gob.pe

Análisis de Clasificación y Agrupamiento

1. Introducción

Todos los campos científicos tienen la necesidad de conglomerar o agrupar objetos similares. Los botánicos agrupan las plantas, los historiadores agrupan los eventos, y los químicos agrupan los elementos y los fenómenos. No debería ser una sorpresa que cuando los administradores de mercadotecnia, sociólogos, psicólogos, etc. tratan de volverse más científicos deben, encontrar una necesidad de procedimientos que agrupen a los objetos.

Por ejemplo, una meta de los administradores de mercadotecnia consiste en identificar segmentos similares para que se puedan desarrollar programas de mercadotécnica para cada uno de estos segmentos o grupos. Por consiguiente, es útil agrupar a los clientes. Se podrían agrupar con la base en los beneficios de producto que buscan. Por consiguiente, los estudiantes podrían ser agrupados sobre la base de los beneficios que buscan de una universidad: se podrían agrupar a los clientes mediante sus estilos de vida. El resultado podría ser un grupo al cual le gusten las actividades externas, otro que disfrute del entrenamiento, y un tercero que se interese en la cocina y en la jardinería. Cada segmento puede tener necesidades distintas de productos y puede responder en forma diferente a los enfoques de publicidad.

Otro ejemplo consistiría en que un investigador social tenga la necesidad de establecer grupos de poblaciones con niveles de bienestar similares dentro de cada grupo, y distintos entre grupos. Por consiguiente, le permitiría recomendar políticas de ayuda a los más necesitados, toda vez que podría conocer las carencias de estos grupos menos favorecidos.

La agrupación es entonces el proceso de dividir un conjunto de datos en grupos mutuamente excluyentes, de tal forma que cada miembro de un grupo esté lo mas cerca posible de otros y grupos diferentes estén lo más lejos posible de otros, donde la distancia se mide con respecto a las variables especificadas, que se quieren predecir.

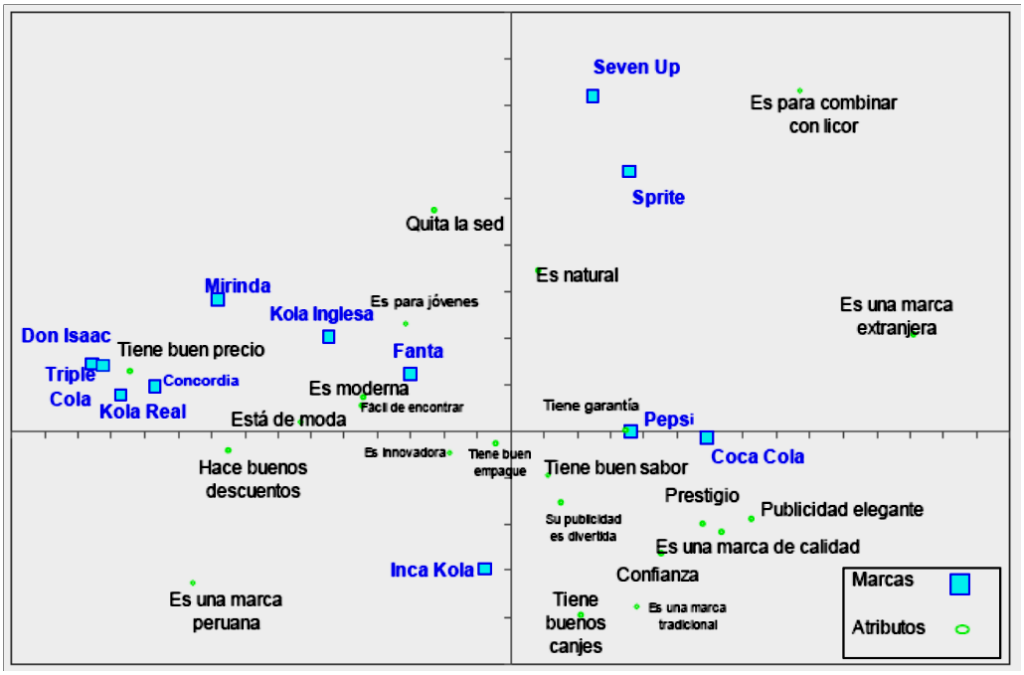
Igualmente, podríamos pensar en clasificar a un objeto o individuo en base a una serie características que permiten decidir a que grupo o clase pertenece, con la finalidad de otorgarle o no un beneficio. Por ejemplo, pensemos en la clasificación de una persona en si es o no susceptible de otorgarle un préstamo de dinero. Por consiguiente, pensaríamos en un modelo probabilístico que nos permita realizar dicha clasificación basados en el conocimiento de, por ejemplo: su nivel de estudios, su nivel de ingreso, si cuenta o no con una tarjeta de crédito, si posee automóvil, etc. Estas variables, que nos permiten tomar una decisión, deben ser previamente definidas haciendo uso de alguna técnica de exploración multivariada.

Como se ha descrito, surge la necesidad de agrupar y/o clasificar objetos o individuos basados en múltiples variables de distinta naturaleza; lo que hace necesario el conocimiento de técnicas estadísticas de agrupación y clasificación y el manejo computacional de algún software estadístico que nos permite construir los grupos más óptimos posibles.

2. Ejemplos de agrupación

- ***Posicionamiento de gaseosas***

Se desea posicionar las marcas de bebidas gaseosas por una serie de atributos que definen lo que podríamos llamar "personalidad de la gaseosa". Así, aplicamos una técnica de agrupación multivariada que nos permite explorar que gaseosa o gaseosas esta relacionada a que atributo.

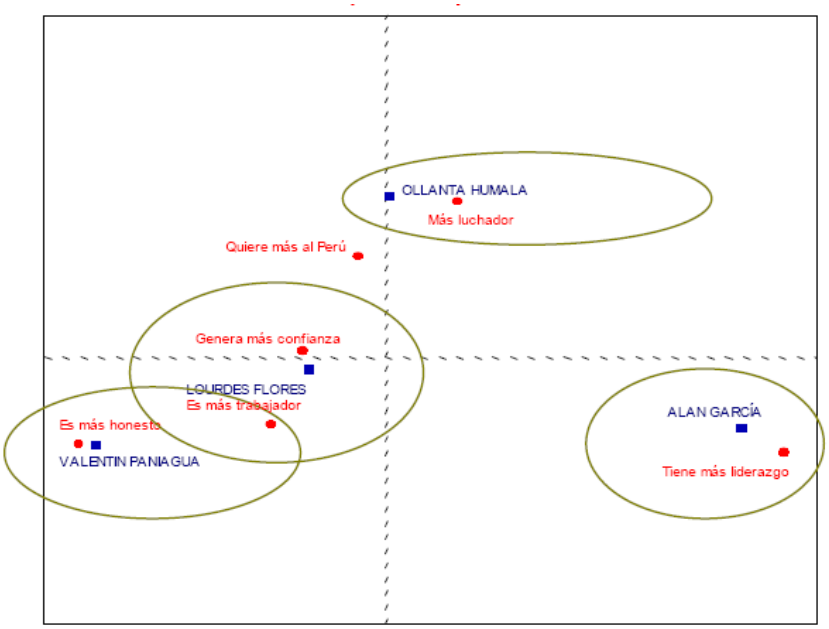


Observemos como las gaseosas Coca Cola y Pepsi son consideradas bebidas de garantía, con buen sabor, de prestigio, tienen publicidad elegante, de confianza, etc. Mientras que las gaseosas Don Issac, Triple Cola, Kola Real Mirinda y otras, son consideradas bebidas que tienen buen precio, hacen buenos descuentos.

Se ha logrado entonces agrupar a las gaseosas por una serie de atributos percibidos por la población consumidora.

- ***Percepción de cómo son los candidatos a la presidencia de la república***

Se desea conocer como percibe la población a los candidatos que optan por ser presidente de la república del Perú. Buscamos entonces construir un mapa perceptual, que nos permita observar, cuales son los atributos asociados a cada candidato.



El gráfico, llamado también "Mapa Perceptual" nos permite explorar cual es la percepción que tienen los votantes con respecto a cuatro candidatos a la presidencia de la República. Vemos que el candidato Ollanta Humala es visto como una persona luchadora, mientras que la candidata Lourdes Flores es vista como una persona trabajadora y que genera confianza. El candidato Valentín Paniagua es conocida como una persona honesta y trabajadora. Y por último el candidato Alan García es considerado una persona con liderazgo.

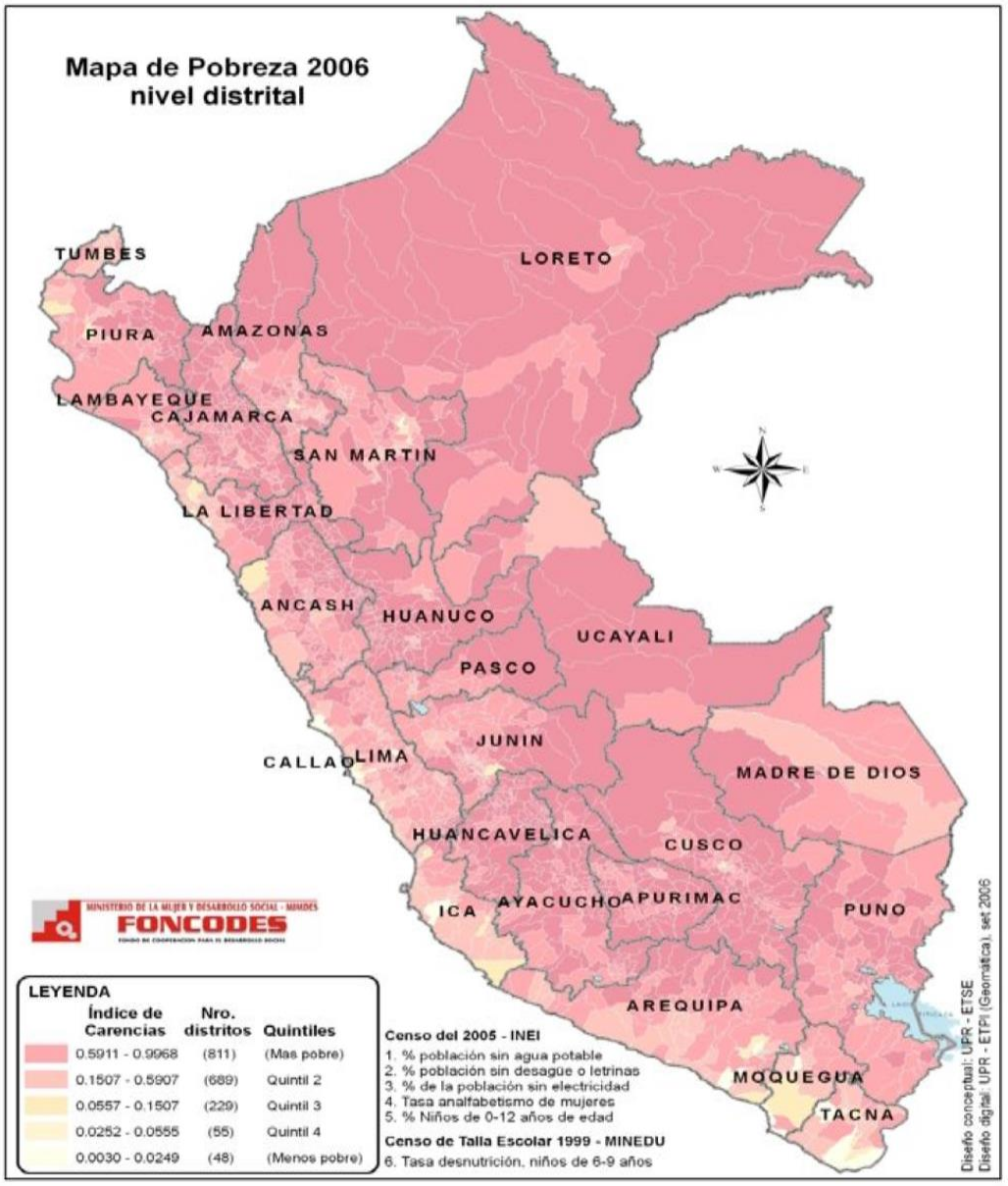
Hemos conseguido entonces agrupar una serie de atributos alrededor de cuatro candidatos a la presidencia dela república.

- ***Construcción de un indicador de carencia que nos permita elaborar un mapa de pobreza de los distritos del país.***

Se desea agrupar a los 1833 distritos del país, en cinco categorías de pobreza, desde muy pobres hasta menos pobres. Lo que nos permitiría elaborar un mapa de pobreza departamental.

Para lograr esta agrupación, el investigador elabora un índice de carencia, para cada distrito, que resume un grupo de variables que definen carencia de algunos servicios básicos, infraestructura o educación (pueden ser mas).

Basados en este índice, que es una variable escalar, se construyen quintiles, y se denomina al de menor quintil, como distritos menos pobres y los de mayor quintil, como distritos más pobres. Así, se genera un mapa de pobreza como el que sigue:



3. Agrupamiento de variables basados en correlaciones

Cuando se miden varias variables en cada una de un gran número de unidades experimentales, a menudo resulta interesante ver si estas variables están interrelacionadas y como lo están. Para examinar estas interrelaciones, cuando varias parejas de variables están intensamente correlacionadas entre sí, se podría intentar la partición de las variables respuesta en grupos, de modo que las variables dentro de un grupo tengan elevadas correlaciones entre sí y las que se encuentran en grupos diferentes tengan bajas correlaciones. Es frecuente que una partición de ese tipo revele aspectos importantes de los datos que resulten útiles de considerar al decidir cómo interpretar estos últimos. Como ilustración, consideremos el ejemplo siguiente:

Cuarenta y ocho individuos que habían presentado solicitud de trabajo a una gran empresa fueron entrevistados y clasificados en relación con 15 criterios. Los aspirantes se clasificaron según la forma de su letra en la solicitud (FL), su aspecto (APP), su capacidad académica (AA), su amabilidad (LA), su autoconfianza (SC), su lucidez (LC), su honestidad (HON), su arte de vender (SMS), su experiencia (EXP), su empuje (DRV), su ambición (AMB), su capacidad para captar conceptos (GSP), su potencial (POT), su entusiasmo para trabajar en grupo (KJ) y su conveniencia (SUIT). Cada criterio se evaluó en una escala del 0 a 10, con 0 como una calificación muy insatisfactoria y con 10 como una calificación muy alta. La evaluación de cada uno de estos individuos es estos 15 criterios se muestra en la tabla siguiente:

ID	FL	APP	AA	LA	SC	LC	HON	SMS	EXP	DRV	AMB	GSP	POT	KJ	SUIT
1	6	7	2	5	8	7	8	8	3	8	9	7	5	7	10
2	9	10	5	8	10	9	9	10	5	9	9	8	8	8	10
3	7	8	3	6	9	8	9	7	4	9	9	8	6	8	10
4	5	6	8	5	6	5	9	2	8	4	5	8	7	6	5
5	6	8	8	8	4	4	9	5	8	5	5	8	8	7	7
6	7	7	7	6	8	7	10	5	9	6	5	8	6	6	6
7	9	9	8	8	8	8	8	8	10	8	10	8	9	8	10
8	9	9	9	8	9	9	8	8	10	9	10	9	9	9	10
9	9	9	7	8	8	8	8	5	9	8	9	8	8	8	10
10	4	7	10	2	10	10	7	10	3	10	10	10	9	3	10
11	4	7	10	0	10	8	3	9	5	9	10	8	10	2	5
12	4	7	10	4	10	10	7	8	2	8	8	10	10	3	7
13	6	9	8	10	5	4	9	4	4	4	5	4	7	6	8
14	8	9	8	9	6	3	8	2	5	2	6	6	7	5	6
15	4	8	8	7	5	4	10	2	7	5	3	6	6	4	6
16	6	9	6	7	8	9	8	9	8	8	7	6	8	6	10
17	8	7	7	7	9	5	8	6	6	7	8	6	6	7	8
18	6	8	8	4	8	8	6	4	3	3	6	7	2	6	4
19	6	7	8	4	7	8	5	4	4	2	6	8	3	5	4
20	4	8	7	8	8	9	10	5	2	6	7	9	8	8	9
21	3	8	6	8	8	8	10	5	3	6	7	8	8	5	8
22	9	8	7	8	9	10	10	10	3	10	8	10	8	10	8
23	7	10	7	9	9	9	10	10	3	9	9	10	9	10	8
24	9	8	7	10	8	10	10	10	2	9	9	9	9	10	8
25	6	9	7	7	4	5	9	3	2	4	4	4	4	5	4
26	7	8	7	8	5	4	8	2	3	4	5	6	5	5	6
27	2	10	7	9	8	9	10	5	3	5	6	7	6	4	5
28	6	3	5	3	5	3	5	0	0	3	3	0	0	5	0
29	4	3	4	3	3	0	0	0	0	4	4	0	0	5	0
30	4	6	5	6	9	4	10	3	1	3	3	2	2	7	3
31	5	5	4	7	8	4	10	3	2	5	5	3	4	8	3
32	3	3	5	7	7	9	10	3	2	5	3	7	5	5	2
33	2	3	5	7	7	9	10	3	2	2	3	6	4	5	2
34	3	4	6	4	3	3	8	1	1	3	3	3	2	5	2
35	6	7	4	3	3	0	9	0	1	0	2	3	1	5	3
36	9	8	5	5	6	6	8	2	2	2	4	5	6	6	3

37	4	9	6	4	10	8	8	9	1	3	9	7	5	3	2
38	4	9	6	6	9	9	7	9	1	2	10	8	5	5	2
39	10	6	9	10	9	10	10	10	10	10	8	10	10	10	10

ID	FL	APP	AA	LA	SC	LC	HON	SMS	EXP	DRV	AMB	GSP	POT	KJ	SUIT
40	10	6	9	10	9	10	10	10	10	10	10	10	10	10	10
41	10	7	8	0	2	1	2	0	10	2	0	3	0	0	10
42	10	3	8	0	1	1	0	0	10	0	0	0	0	0	10
43	3	4	9	8	2	4	5	3	6	2	1	3	3	3	8
44	7	7	7	6	9	8	8	6	8	8	10	8	8	6	5
45	9	6	10	9	7	7	10	2	1	5	5	7	8	4	5
46	9	8	10	10	7	9	10	3	1	5	7	9	9	4	4
47	0	7	10	3	5	0	10	0	0	2	2	0	0	0	0
48	0	6	10	1	5	0	10	0	0	2	2	0	0	0	0

(Esta tabla también se adjunta en formato Excel y SPSS)

Supongamos que la empresa desea ofrecer empleo a los seis mejores solicitantes (candidatos).
¿Cómo los seleccionaría? ¿Qué indicador le facilitaría la decisión?

Una manera simple para hacerlo sería colocar a los 48 nombres en una caja, mezclarlos y, a continuación, extraer al azar seis nombres. Este método daría a todos igual posibilidad de ser seleccionados. La mayoría de los empleadores no aprobarían este método porque no se usa ninguno de los datos reunidos en la entrevista.

Un segundo método sería calcular una calificación promedio (PROM) para cada individuo, mediante el cálculo del promedio de las 15 variables. De donde se podría calcular el indicador:

Para cada individuo y seleccionar a los seis que tuvieran las calificaciones promedio más altas. Es obvio que este método da un peso igual a cada uno de los 15 criterios y, como consecuencia, este método de cálculo de una calificación sencilla para cada individuo no siempre es deseable, pues la existencia de correlaciones fuertes entre variables, dan redundancia de información al entrar al indicador con el mismo peso.

Una tercer posibilidad, probablemente la preferida para muchos empleadores, sería ponderar cada uno de los criterios de acuerdo con la importancia que da ese empleador a cada uno de ellos y, a continuación, calcular un promedio ponderado para cada individuo. De este modo, se podría calcular una calificación ponderada como:

En donde $w_1 + w_2 + \ldots + w_{15} = 1$, y en donde w_i mide la importancia relativa del i-ésimo criterio, i = 1,2,…,15. Note que el requisito de que la suma de los pesos sea 1 en realidad no es necesario; esto sencillamente garantiza que los promedios ponderados resultantes tengan un valor entre 0 y 10 para estos datos.

Para examinar todavía más si los enfoques que se acaban de describir son razonables, podríamos buscar las interrelaciones entre las variables. Considere las correlaciones entre todas las parejas de estas 15 variables, que se obtienen con el procedimiento correlations de SPSS.

```
CORRELATIONS
 /VARIABLES=FL APP AA LA SC LC HON SMS EXP DRV AMB GSP POT KJ SUIT
 /PRINT=TWOTAIL NOSIG
 /MISSING=PAIRWISE.
```

Los resultados del procedimiento se dan en la tabla de correlaciones siguiente, en donde se han subrayado todas las correlaciones con valores de 0.5 o superior.

	FL	APP	AA	LA	SC	LC	HON	SMS	EXP	DRV	AMB	GSP	POT	KJ	SUIT
FL	1.00	0.24	0.04	0.31	0.09	0.23	-0.11	0.27	0.55	0.35	0.30	0.34	0.37	0.47	0.59
APP	0.24	1.00	0.12	0.38	0.43	0.37	0.35	0.49	0.14	0.34	0.55	0.51	0.51	0.28	0.38
AA	0.04	0.12	1.00	0.00	0.00	0.08	-0.03	0.05	0.27	0.09	0.04	0.20	0.29	-0.32	0.14
LA	0.31	0.38	0.00	1.00	0.30	0.48	0.65	0.36	0.14	0.39	0.36	0.50	0.61	0.69	0.33
SC	0.09	0.43	0.00	0.30	1.00	0.81	0.41	0.80	0.02	0.70	0.84	0.72	0.67	0.48	0.25
LC	0.23	0.37	0.08	0.48	0.81	1.00	0.36	0.82	0.15	0.70	0.77	0.88	0.78	0.53	0.42
HON	-0.11	0.35	-0.03	0.65	0.41	0.36	1.00	0.24	-0.16	0.28	0.22	0.39	0.42	0.45	0.00
SMS	0.27	0.49	0.05	0.36	0.80	0.82	0.24	1.00	0.26	0.81	0.87	0.78	0.75	0.56	0.56
EXP	0.55	0.14	0.27	0.14	0.02	0.15	-0.16	0.26	1.00	0.34	0.18	0.30	0.35	0.21	0.69
DRV	0.35	0.34	0.09	0.39	0.70	0.70	0.28	0.81	0.34	1.00	0.79	0.71	0.79	0.61	0.62
AMB	0.30	0.55	0.04	0.36	0.84	0.77	0.22	0.87	0.18	0.79	1.00	0.79	0.78	0.57	0.44
GSP	0.34	0.51	0.20	0.50	0.72	0.88	0.39	0.78	0.30	0.71	0.79	1.00	0.88	0.55	0.53
POT	0.37	0.51	0.29	0.61	0.67	0.78	0.42	0.75	0.35	0.79	0.78	0.88	1.00	0.54	0.57
KJ	0.47	0.28	-0.32	0.69	0.48	0.53	0.45	0.56	0.21	0.61	0.57	0.55	0.54	1.00	0.40
SUIT	0.59	0.38	0.14	0.33	0.25	0.42	0.00	0.56	0.69	0.62	0.44	0.53	0.57	0.40	1.00

A continuación, divida estas variables en subgrupos en los que las variables estén intensamente correlacionadas entre sí, mientras que las que estén en subgrupos diferentes tengan baja correlación. En este proceso, es probable que sea más fácil empezar por formar el primer subgrupo con las dos variables que estén lo más intensamente relacionadas entre sí. La correlación entre LC y GSP es 0.88, como lo es la correlación entre POT y GSP. Asimismo,
la correlación entre LC y POT es 0.78. De este modo, con certeza, estas tres variables deben estar en el mismo subgrupo, porque las tres están muy correlacionadas entre sí. SMS debe incluirse en este subgrupo, puesto que tiene correlaciones de 0.82, 0.78 y 0.75 con LC, GSP y POT, respectivamente. AMB también debe incluirse en este subgrupo, ya que tiene correlaciones de 0.76, 0.86, 0.78 y 0.77 con las otras variables que están en el subgrupo. Un examen adicional de la matriz de correlaciones revelas que las variables DRV y SC también deben incluirse en este subgrupo. Note que todas las variables en el primer grupo tienen

correlaciones entre sí que son por lo menos de 0.67. Ninguna de las otras variables parece que pertenezca a este primer grupo, de modo que ahora empecemos a buscar un segundo grupo de variables.

Se puede formar un segundo grupo de variables con FL, EXP y SUIT. Las correlaciones entre parejas sucesivas de estas variables son 0.55, 0359 y 0.69, respectivamente. Ninguna de las demás variables pertenece a este subgrupo.

Se puede establecer un tercer grupo de variables con KJ y LA, las cuales tienen una correlación de 0.69 entre sí. HON tiene una correlación de 0.65 con LA, pero solo de 0.45 con KJ. Se podría pensar en si HON debe incluirse en un subgrupo con LA y KJ. Debido a que estos datos provienen de mediciones hechas en "personas", yo me inclinaría a incluir HON en este subgrupo. ¿Qué haría usted?.

Para resumir, los grupos finales de variables son:

Grupo 1: SC, LC, SMS, DRV, AMB,
GSP y POT Grupo 2: FL, EXP y SUIT
Grupo 3: LA,
HON y KJ Grupo 4:
AA
Grupo 5: APP

Una pregunta que se podría hacer es: "¿esta formación de grupos de variables tiene alguna influencia sobre cómo podrían seleccionarse a los solicitantes a quienes se les hace una oferta de trabajo?"

Resulta interesante hacer notar que, en principio, la mayoría de los lectores probablemente creyeron que se estaba midiendo 15 características diferentes de cada solicitante mediante estas 15 variables, pero los agrupamientos precedentes harían que la mayoría de los investigadores crean que solo se midieron cinco características diferentes. Como resultado, muchos creerían que cada solicitante debe evaluarse respecto a estas cinco características.
Entonces, con el fin de decidir a quienes hacerles las ofertas de trabajo, podrían calcularse una nueva calificación global para cada individuo tomando un promedio de estas cinco nuevas características o un promedio ponderado de estas.

Para evaluar cada individuo respecto a cada una de estas características subyacentes, se podrían promediar sencillamente las variables que se encuentran en cada grupo. De este modo se tendría:

Entonces se podría calcular una calificación para cada individuo mediante el uso de un promedio ponderado

Otra manera de calcular una calificación para cada una de las nuevas variables, sería elegir una de las variables de cada grupo para que representara la característica subyacente para cada uno de estos grupos.