

AID - Trabajo Práctico Nro. I

6 de febrero de 2022

Ejercicio 1 (Transformaciones de datos).

Seis candidatas son evaluadas para el puesto de recepcionista en una empresa, para ello pasan por dos entrevistas. En la primera las evalúa el responsable de recursos humanos de la empresa (juez 1) y en la segunda el responsable del área de la cual van a depender (juez 2). La asignación de puntajes es en cordialidad, presencia y manejo de idiomas. Los puntajes asignados independientemente por estos jueces se encuentran en el archivo recepcionistas.xls.

- (a) Calcule en promedio por juez de cada una de las aspirantes. ¿Cuál le parece que seleccionaría cada uno de ellos? ¿Existe coincidencia?
- (b) Calcule el promedio de cada una de las aspirantes tomando en cuenta todos los rubros y ambos jueces.
- (c) Transformar las puntuaciones observadas de modo tal que cada una de las seis variables tenga media 0 y dispersión 1. ¿Cuál sería el objetivo de esta transformación?
- (d) Transformar las puntuaciones de modo tal que cada candidata tenga para cada juez media 0 y dispersión 1. ¿Cuál sería el objetivo de esta transformación?

- (e) Grafique los perfiles multivariados de cada una de las candidatas para ambas transformaciones. ¿Qué observa?

Ejercicio 2 (Tipos de variables resúmenes: (Datos: Internet. 2013)).

Se han registrado sobre 1500 individuos las variables siguientes:

- ID: número de identificación del registro.
 - Nacionalidad
 - Edad: cumplida en años
 - Sexo: Masculino (1) Femenino (2)
 - Estatura: en m
 - Sitio: sitio preferido al que se conecta; 1- chat 2 - correo electrónico 3- buscadores 4 – software 5 – música 6 – deportes 7 - otros
 - Uso: Tiempo de uso promedio por día en minutos
 - Temperatura: media anual de la zona de residencia
 - Autos: cantidad de autos en la manzana donde reside
 - Cigarrillos: Cantidad de cigarrillos consumidos mientras utiliza Internet
- (a) Clasificar las variables de la base. Para las variables numéricas construir un gráfico de coordenadas paralelas.
- (b) Construir la tabla de frecuencias de la variable sexo. ¿Hay algún valor que llame la atención? ¿Qué tipo de error considera que es?
- (c) Ordenar los datos por la variable Edad. ¿Encontró algún valor extraño? ¿Qué tipo de error puede ser? item Construir la tabla de frecuencias de la variable Sitio. ¿Encuentra algún valor que le llame la atención? ¿Qué tipo de error puede ser?

- (d) Proceda de forma similar para las variables Temperatura, Autos y Cigarrillos.
 - (e) Elimine de la base los valores que no son posibles y que seguramente corresponde a un error de tipo. Detalle valores/registros que le hayan llamado la atención pero no deban ser eliminados necesariamente.
 - (f) ¿Para cuáles de las variables tiene sentido calcular la media? ¿Y la mediana? ¿Para cuáles de las variables tiene sentido calcular la media? ¿Y la mediana?
 - (g) ¿Cuáles de las variables le parecen simétricas a partir de estos resúmenes? Confirme estas observaciones mediante un boxplot.
 - (h) Calcular la desviación intercuartil y detectar presencia de valores salvajes moderados y severos.
- .

Ejercicio 3. *Gráficos univariados y multivariados (Datos: Gorrones.xls)*

Base de datos: Se han registrado para 49 gorrones las siguientes variables zoo métricas:

- Largo total
- Extensión alar
- Largo del pico y cabeza
- Largo del húmero
- Largo de la quilla del esternón
- Sobrevida (1) Si, (-1) No

- (a) Indicar en cada caso de que tipo de variable se trata.
- (b) Confeccionar un informe para cada variable(univariado).
- (c) Realizar en el caso que corresponda un histograma. Ensayar el número de intervalos que conviene en cada variable, indicar si utiliza algún criterio.
- (d) Realizar un boxplot comparativo para cada una de estas variables particionando por el grupo definido por la supervivencia. ¿Le parece que alguna de estas variables está relacionada con la supervivencia, es decir que toma valores muy distintos en ambos grupos? Analizar en todos los casos la presencia de outliers.
- (e) Construir gráficos bivariados para las todas las variables, particionando por el grupo de supervivencia (un color para cada grupo). ¿Observa alguna regularidad que pueda explicar la supervivencia?
- (f) Construir la matriz de diagramas de dispersión. ¿Considera que algún par de estas medidas están relacionadas? Estudiar si la asociación de algunas de estas medidas es diferente en alguno de los grupos.

Ejercicio 4. *Base de datos razaperros.xls*

Se han registrado respecto de 27 razas de perros las siguientes características:

- Nombre de la raza
- Tamaño: con niveles 1- pequeño, 2- mediano y 3- grande
- Peso: con niveles 1- liviano, 2- medio y 3- pesado
- Velocidad: con niveles 1- lento, 2- mediano y 3- rápido

- Inteligencia: con niveles 1 a 3
 - Afectividad: con niveles 1 a 3
 - Agresividad: con niveles 1 a 3
 - Función: con tres categorías caza, utilitario y compañía.
- (a) Realizar un gráfico de estrellas por raza (utilizando las variables tamaño, peso, velocidad, inteligencia y afectividad).
- (b) Idem por función.
- (c) Idem por agresividad.
- (d) En el primer gráfico se observan estrellas similares. ¿Le parece que las razas son parecidas?

Ejercicio 5. *Matriz de covarianzas: (Datos Gorrones.xls)*

Para esta base de datos, interesa:

- (a) Dimensión de la base de datos (n = número de observaciones, p = cantidad de variables observadas sobre cada individuo).
- (b) Hallar el vector de medias, la matriz de varianzas y covarianzas y la matriz de correlaciones. ¿Qué características tienen estas matrices?
- (c) Explicar que representa el elemento m_{11} de la matriz de varianzas y covarianzas, ídem para el elemento m_{31} .
- (d) Explicar que representa el elemento m_{22} de la matriz de correlaciones, ídem para el elemento m_{13} .
- (e) Relacionar los elementos m_{21} , m_{11} y m_{22} de la matriz de varianzas y covarianzas con el elemento m_{12} de la matriz de correlaciones.

- (f) Hallar una nueva variable e incorporarla en la base de Gorriónes: Diferencia entre el largo total y el largo del húmero. Llamémosla: Diferencia de Largos.
- (g) Calcular nuevamente el vector de medias y las matrices de varianzas y covarianzas y la matriz de correlaciones de la nueva base de datos. Relacionar el nuevo vector de medias con el anterior.
- (h) Hallar la traza de las cuatro matrices. Explicar el significado de cada uno de los resultados. ¿Qué traza/s no aumentan al aumentar una variable? Explique.

Ejercicio 6 (Propiedades de la matriz de Covarianzas (Datos: recepcionistas.xls)).

Para el archivo de se pide:

- (a) Calcular el vector de medias e interpretar los valores.
- (b) Hallar las matrices de varianzas y covarianzas y de correlaciones para la submatriz de puntuaciones del primer juez, ídem para el segundo juez. Ídem para el conjunto total.
- (c) ¿Se puede decir que la suma de las dos primeras submatrices darán como resultado la matriz del grupo total? Si no es así por favor explique por qué no.
- (d) ¿Se cumple esta relación para las trazas? y para el vector de medias? y para los vectores de medianas?

Ejercicio 7 (Medidas de Posición y Escala robustas). (Datos: Internet.2013)

1. Seleccione las variables numéricas del archivo y agregue 5 observaciones que no sean atípicas en forma univariada pero sí lo sean en forma multivariada. Utilice las medidas robustas para detectar estos valores.

2. Ahora agregue cuatro observaciones que sean outliers pero aparezcan enmascaradas. Utilice estrategias robustas para detectar su presencia.