Actividad 1. Ingreso a Universidad

Edgar Anuar Sánchez Hernández

2023-01-16

Actividad

Como se menciona en el material de la unidad 3, es necesario realizar un análisis exploratorio de los datos con el fin de tener información relevante en cuanto el nivel que tienen los aspirantes a ingresar a la maestría. Por lo tanto, basados en los datos proporcionados y aplicando las herramientas analizadas en esta unidad, responda las siguientes preguntas:

a) identifique y defina con claridad el objeto o unidad de interés del presente estudio.

Para esto primero cargamos la base de datos

```
library(readxl)
datosexcel <- read_excel("Calif_ingreso.xlsx")
dfdatos <- data.frame(datosexcel)
head(dfdatos)
```

```
##   Nombre   sexo geodif ancompl alg anreal estad
## 1   Jose Hombre     36      58  43     36    37
## 2  Maria  Mujer     31      42  41     40    29
## 3   Luis Hombre     76      78  69     66    81
## 4  Elena  Mujer     46      56  52     56    40
## 5 Franco Hombre     12      42  38     38    28
## 6  Julio Hombre     39      46  51     54    41
```

Respuesta a)

Se observa que las unidades de estudio son estudiantes que desean ingresar a la maestría en Matemáticas y que realizaron su examen de admisión.

b)

Para cada una de las variables consideradas en este estudio mencione su tipo (categórica/numérica) y su escala de medición (nominal/ordinal/intervalo/razón).

Respuesta b):

Las variables de estudio, su tipo y escala de medición son las siguientes:

1. Sexo. Tipo: categórica. Escala: nominal.

2. geodif: puntuación obtenida por el candidato en el examen de Gometría Diferencial. Tipo: numérica. Escala: intervalo.

3. ancompl: puntuación obtenida por el candidato en el examen de Análisis Complejo. Tipo: numérica. Escala: intervalo.

4. alg: puntuación obtenida por el candidato en el examen de Álgebra. Tipo: numérica. Escala: intervalo.

5. anreal: puntuación obtenida por el candidato en el examen de Análisis Real. Tipo: numérica. Escala: intervalo.

6. estad: puntuación obtenida por el candidato en el examen de Estadística. Tipo: numérica. Escala: intervalo.

c)

Al realizar un análisis descriptivo de la base de datos

i. ¿En alguna de las variables bajo estudio es posible identificar casos que podrían ser considerados como atípicos (outliers)? Justifique su respuesta.

Respuesta i)

En este caso las variables numéricas, que son las puntuaciones en cada materia del examen de admisión, podrían presentar casos atípicos. Pero es muy poco probable, ya que estos outliers serían puntuaciones muy altas o muy bajas, comparadas con la media. Estos outliers representarían, ya sea la presencia de un alumno superdotado o con capacidades superiores a las del resto de alumnos, o bien por el contrario, podría representar a un alumno con rendimiendo demasiado bajo, quizás un alumno que no sea de la Lic. en Matemáticas o incluso que nisiquiera sea egresado de alguna carrera relacionada con la ciencia. Otra posible explicación de un outlier sería un error en la captura de su calificación.

ii. ¿En cuáles variables se presentan y cuántos casos tendría?. Es necesario presentar evidencia que permita comprobar la veracidad de su respuesta.

Respuesta ii)

Para identificar outliers realizaremos un diagrama de cajas y bigotes con R tanto en general como para sólo hombres y sólo mujeres. Entendiendo que los casos atípicos se mostrarían como puntos fuera de los "bigotes" de las cajas. Ya que la longitud de los "bigotes" de la gráfica se encuentran a una distancia de $1.5RI$ desde Q_1 para el mínimo y desde Q_3 para el máximo.

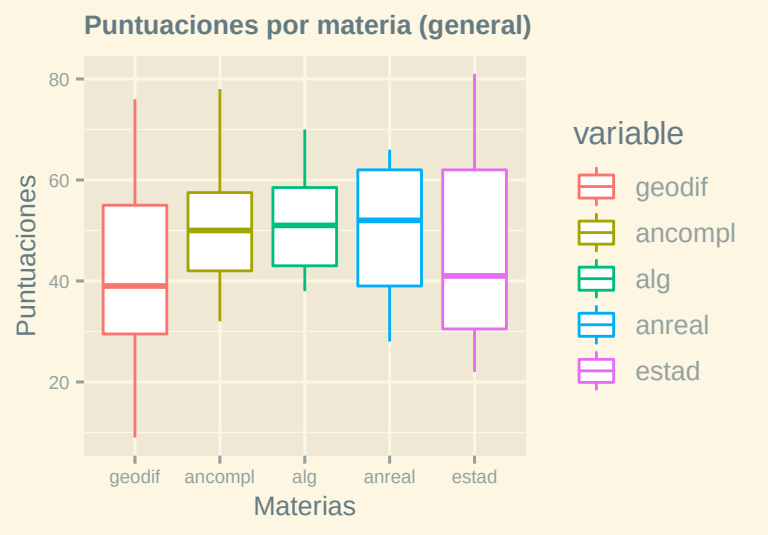
```
library(ggplot2)
library("ggthemes")
library(reshape2)
library(dplyr)

datos_mod <- melt(dfdatos, id = c("Nombre", "sexo"))
head(datos_mod)
```

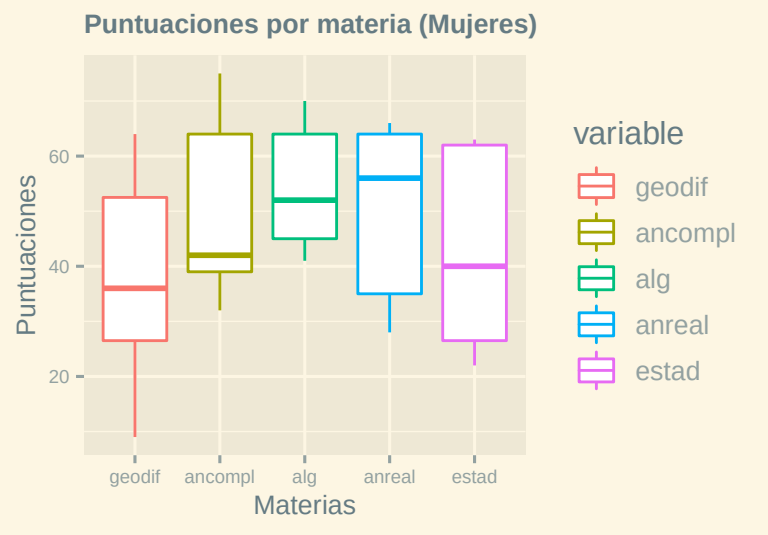
```
##   Nombre   sexo variable value
## 1   Jose Hombre   geodif    36
## 2  Maria  Mujer   geodif    31
```

```
## 3   Luis Hombre   geodif    76
## 4  Elena  Mujer   geodif    46
## 5 Franco Hombre   geodif    12
## 6  Julio Hombre   geodif    39
```

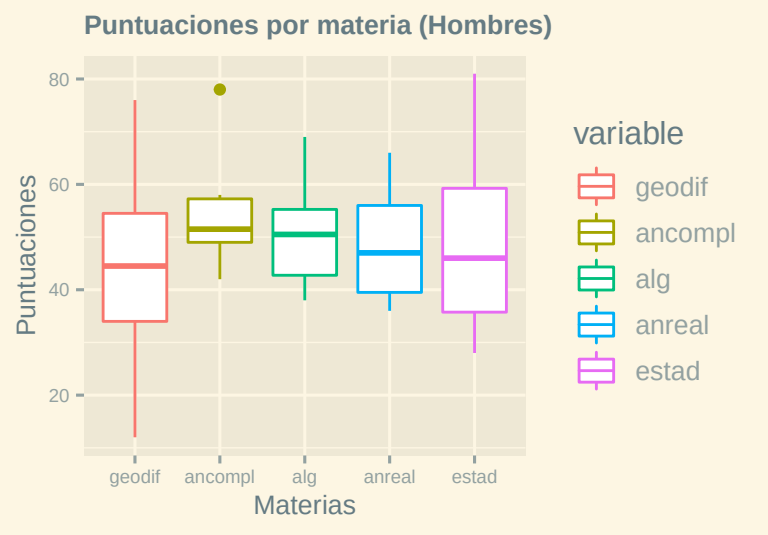
```
ggplot(datos_mod) +
  geom_boxplot(aes(x=variable, y=value, color=variable)) +
  theme_solarized_2() +
  theme(axis.text=element_text(size=7), axis.title=element_text(size=10),
        plot.title=element_text(size=10, face="bold")) +
  ggtitle("Puntuaciones por materia (general)") +
  labs(y= "Puntuaciones", x = "Materias")
```



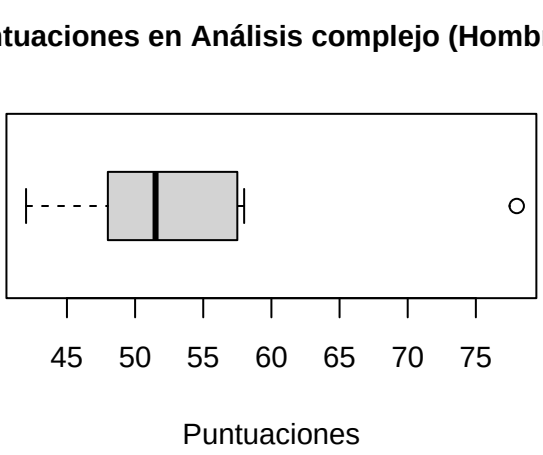

```
ggplot(filter(datos_mod, sexo == "Mujer")) +
  geom_boxplot(aes(x=variable, y=value, color=variable)) +
  theme_solarized_2() +
  theme(axis.text=element_text(size=7), axis.title=element_text(size=10),
        plot.title=element_text(size=10, face="bold")) +
  ggtitle("Puntuaciones por materia (Mujeres)") +
  labs(y= "Puntuaciones", x = "Materias")
```




```
ggplot(filter(datos_mod, sexo == "Hombre")) +
  geom_boxplot(aes(x=variable, y=value, color=variable)) +
  theme_solarized_2() +
  theme(axis.text=element_text(size=7), axis.title=element_text(size=10),
        plot.title=element_text(size=10, face="bold")) +
  ggtitle("Puntuaciones por materia (Hombres)") +
  labs(y= "Puntuaciones", x = "Materias")
```




```
ancompl_h <- boxplot(filter(datos_mod, sexo == "Hombre", variable == "ancompl")["value"],
                     xlab = "Puntuaciones", main = "Puntuaciones en Análisis complejo (Hombres)",
                     horizontal = T, cex.main = 0.9, cex.axis = 0.9, cex.lab = 0.9)
```



En las gráficas se observa que para el caso de sólo hombres, aparece un dato atípico (outliers). Una puntuación demasiada alta comparada con el resto. El valor de dicha puntuación es 78. Se aconseja revisar esa puntuación para verificar que no se trate de un error en la captura. En caso de no tratarse de un error, se trataría de un estudiante hombre que es sobresaliente en el área de análisis complejo.

d)

Explorando más profundamente los datos, ¿hay evidencia que existen variables con alta correlación entre ellas? Justifique la respuesta con evidencia estadística y en caso de ser afirmativa mencione los dos pares de variables más altamente correlacionadas (positiva o negativamente).

Respuesta

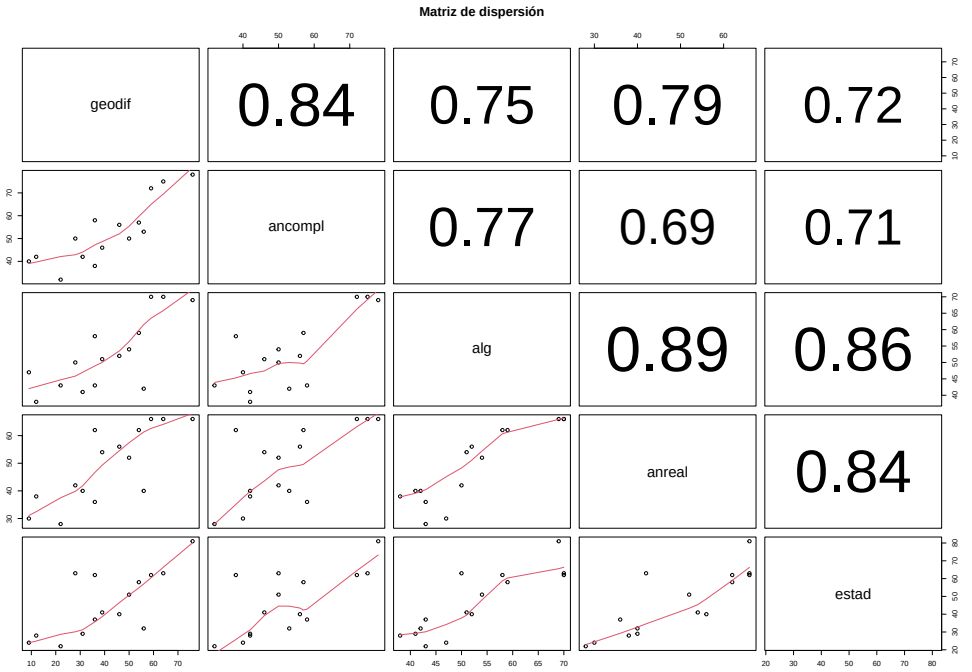
Graficamos la matriz de dispersión junto con los coeficientes de correlación por pares y obtenemos lo siguiente:

```
panel.cor <- function(x, y, digits=2, prefix="", cex.cor, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr=c(0,1,0,1))
  r <- abs(cor(x, y))
  txt <- format(c(r, 0.123456789), digits=digits)[1]
  txt <- paste(prefix, txt, sep="")
  if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex=cex.cor * r)
}

# panel.cor <- function(x1, y, digits=2, prefix="", cex.cor){
#   par(usr = c(0, 1, 0, 1))
#   r <- abs(cor(x1, y))
#   txt <- format(c(r, 0.123456789), digits=digits)[1]
#   txt <- paste(prefix, txt, sep="")
#   if(missing(cex.cor))
#   cex <- 0.5/strwidth(txt)
#   text(0.5, 0.5, txt, cex = cex)
```

```
# }


pairs(~ geodif + ancompl + alg + anreal + estad, data = dfdatos,
      lower.panel=panel.smooth, upper.panel=panel.cor, main='Matriz de dispersión')
```




```
#par(fig=c(0,1,0,1))
```

Observamos que todos los pares de variables numéricas tienen cierto grado de correlación positiva. Las variables con mayor nivel de correlación son: álgebra-análisis (cor=0.89) real y estadística-álgebra (cor=0.86).

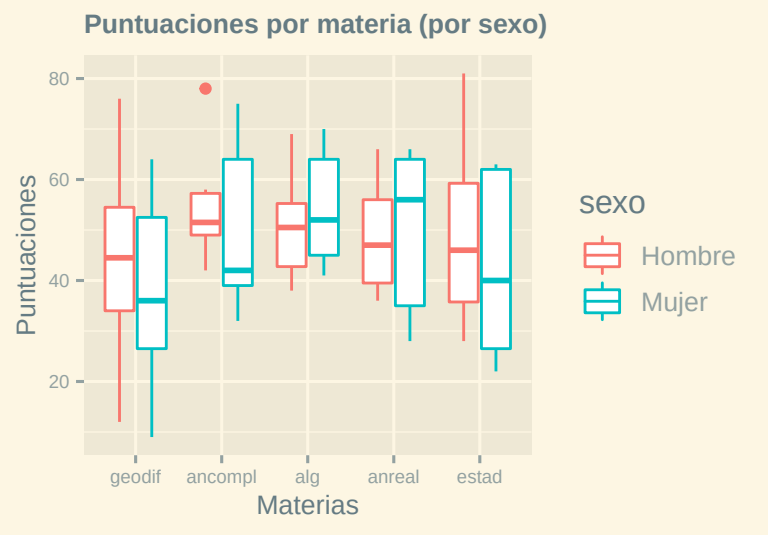
e)

A los integrantes del comité de ingreso les gustaría conocer si resulta verdadera la afirmación de que, a partir de los datos recolectados, los hombres alcanzan mejores puntuaciones en las cinco áreas que las mujeres. ¿Hay evidencia que apoyen esta afirmación? Justifique su respuesta con evidencia verificable.

Respuesta

Para responder esta pregunta, generamos una gráfica con boxplots lado a lado de hombres y de mujeres:

```
ggplot(datos_mod) +
  geom_boxplot(aes(x=variable, y=value, color=sexo)) +
  theme_solarized_2() +
  theme(axis.text=element_text(size=7), axis.title=element_text(size=10),
        plot.title=element_text(size=10, face="bold")) +
  ggtitle("Puntuaciones por materia (por sexo)") +
  labs(y= "Puntuaciones", x = "Materias")
```



Como se puede observar, en promedio, los hombres obtuvieron mejores puntuaciones que las mujeres en geometría diferencial, análisis complejo y estadística. Sin embargo, las mujeres obtuvieron, en promedio, mejores puntuaciones que los hombres tanto en álgebra como en análisis real. Por lo tanto no hay evidencia de que los hombres alcanzaron mejores calificaciones en las 5 áreas que las mujeres.

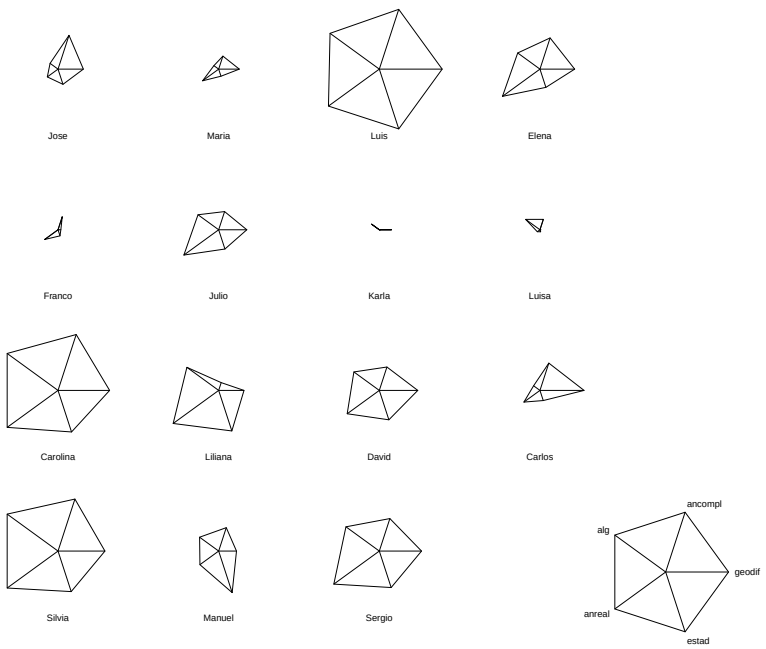
f)

En el caso de que se dictamine el ingreso de los 15 aspirantes, una de las acciones que implementará la coordinación del programa para nivelar las habilidades y conocimiento de los nuevos estudiantes y combatir la deserción escolar es ofrecer cursos propedéuticos "personalizado" a los aspirantes dependiendo de los resultados obtenidos. Debido al limitado número de profesores del departamento es prácticamente imposible asignar a cada aspirante un tutor aunado a que los cursos se estarían impartiendo durante el verano. La estrategia que se pretende seguir se basa en agrupar aquellos aspirantes que tuvieron un rendimiento similar en las cinco áreas evaluadas y asignarles un tutor que los ayude a revisar y alcanzar el nivel esperado en todas las materias. A partir de los datos proporcionados, ¿existirán candidatos que presentan comportamiento similar (conocimiento) en las cinco áreas que permitan formar los grupos de preparación? Justifique su respuesta con evidencia estadística y en caso de existir similitud entre candidatos, mencione tres ejemplos de agrupaciones.

Respuesta

Para responder esta pregunta, elaboramos una gráfica de polígonos para cada alumno, donde cada vértice represente su puntuación en cada área del examen. De esta manera podremos identificar alumnos con rendimientos similares, cuando sus polígonos sean similares.

```
#library(symbols)
stars(scale(dfdatos[3:7]), key.loc = c(11,2), radius = T,
      labels = dfdatos$Nombre, flip.labels = F, len = 0.9)
```



Como se observa en el gráfico, sí existen alumnos con conocimiento similar en las 5 áreas. Por ejemplo, Luis, Carolina, Silvia y Sergio tienen aproximadamente un buen dominio en las 5 materias, mientras que David, Liliana, Elena, Julio y Manuel, tienen un conocimiento ligeramente inferior al del primer grupo en las 5 áreas, incluso con una tendencia predominante al análisis real. Finalmente, el resto de aspirantes: José, María, Franco, Karla, Luisa y Carlos; tienen el menor nivel de comocimiento, incluso la mayoría presenta lagunas, o áreas de conocimiento donde obtuvieron un rendimiento bastante bajo, comparado con el del resto de alumnos. Este es el ejemplo de 3 posibles agrupaciones con base en su rendimiento en las 5 áreas de conocimiento.