PROYECTO MÉTODOS NUMÉRICOS.

Base de datos encuesta.

Materia: Métodos numéricos.

Grupo "4Y"

Integrantes:

Emiliano Espinoza perales.

Haro Reyes Rogelio.

1.- Descripción del problema

Se realizó una encuesta de manera independiente con el objetivo de recolectar información actual referente al estado/características físicas de un total de 75 personas seleccionadas de manera aleatoria. Información en forma de variables que tiene como objetivo obtener la correlación entre las mismas y más específicamente de la mayoría de estas variables con una variable específica como lo podría ser el peso.

Cada registro representa una persona encuestada. A continuación, se describen las variables de estudio...

2.- Descripción de las variables de estudio.

- PESO: Variable numérica de los kg que pesa.

#VARIABLE DEPENDIENTE

- ALTURA: Variable numérica de los cm que mide.

#VARIABLE INDEPENDIENTE

- EDAD: Variable numérica de los años de edad.

#VARIABLE INDEPENDIENTE

- COMPLEXIÓN: Variable numérica (Rango 1:3) que representa el tipo de cuerpo.

#VARIABLE INDEPENDIENTE

- NACIONALIDAD: Variable numérica (Rango 1:3) que representa la nacionalidad.

#VARIABLE INDEPENDIENTE

- SEXO: Variable Cualitativa (1 = si la persona es hombre, 2 = si la persona es mujer).

#VARIABLE INDEPENDIENTE

3.- Análisis exploratorio de los datos.

Leer/cargar archivo .csv (valores separados por comas)

archivoleido <- read.csv("BDD-METODOSNUMERICOS-OFF.csv")

Imprimir una vista de los primeros 6 registros

head(archivoleido)

```
  PESO..Kg. ALTURA..Cm. EDAD COMPLEXION.FISICA NACIONALIDAD SEXO
1        62         175   20                 1            1    1
2        79         172   19                 2            1    1
3       102         179   23                 3            2    2
4        67         165   42                 1            3    2
5        80         185   27                 2            2    1
6       110         165   50                 3            1    2
>
```

Se hace posible la vinculación con las variables/columnas de nuestra base de datos...

Simplificando a la expresión "archivoleido$nombre_variable"

attach(archivoleido)

Se desvinculan las variables/columnas de la base de datos que 'attach()' hizo... posible referenciar de manera individual

detach(archivoleido)

Gráficas de Dispersión.-

Diagrama de dispersión de ALTURA en relación al número de registro.

plot(ALTURA..Cm., xlab="No. de Registro", main = "Diagrama de Dispersión (Altura)")

EXPLICACIÓN. - En el diagrama de dispersión de los valores de la Altura en función al número de registro se puede visualizar una gran dispersión entre la mayoría de los valores en la tabla. Podemos observar que del registro 1 al 25 hay valores de estatura variados, la mayoría por encima de los de menor valor, al igual que existe el valor mayor (190). Los registros de entre 25 y 40 muestran valores de estatura medios. Finalmente, los valores de estatura de los registros de 40 a 75 presentan el comportamiento más variado, conteniendo el menor valor (155), valores medios y algunos de los mayores.

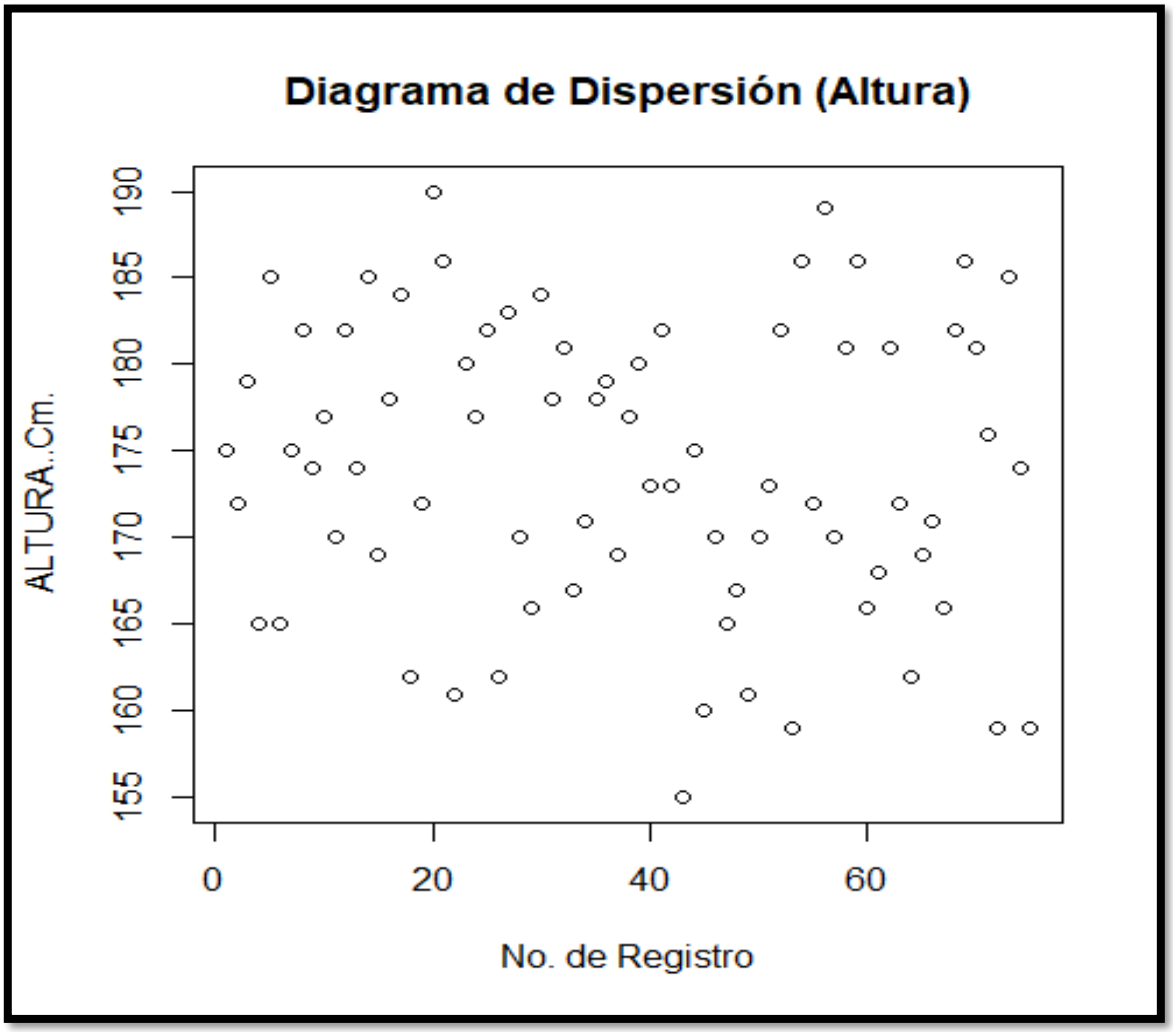


Diagrama de dispersión de EDAD en relación al número de registro.

plot(EDAD, xlab="No. de Registro", main = "Diagrama de Dispersión (Edad)")

EXPLICACIÓN. - Podemos visualizar que del registro 1 al 35 existen los valores de edad mayores, al igual que valores de edad menores y medios que van creciendo. Se observa la mayor dispersión en este rango. Del registro 35 al 75 se siguen mostrando los valores de edad menores y medios que continúan incrementando.

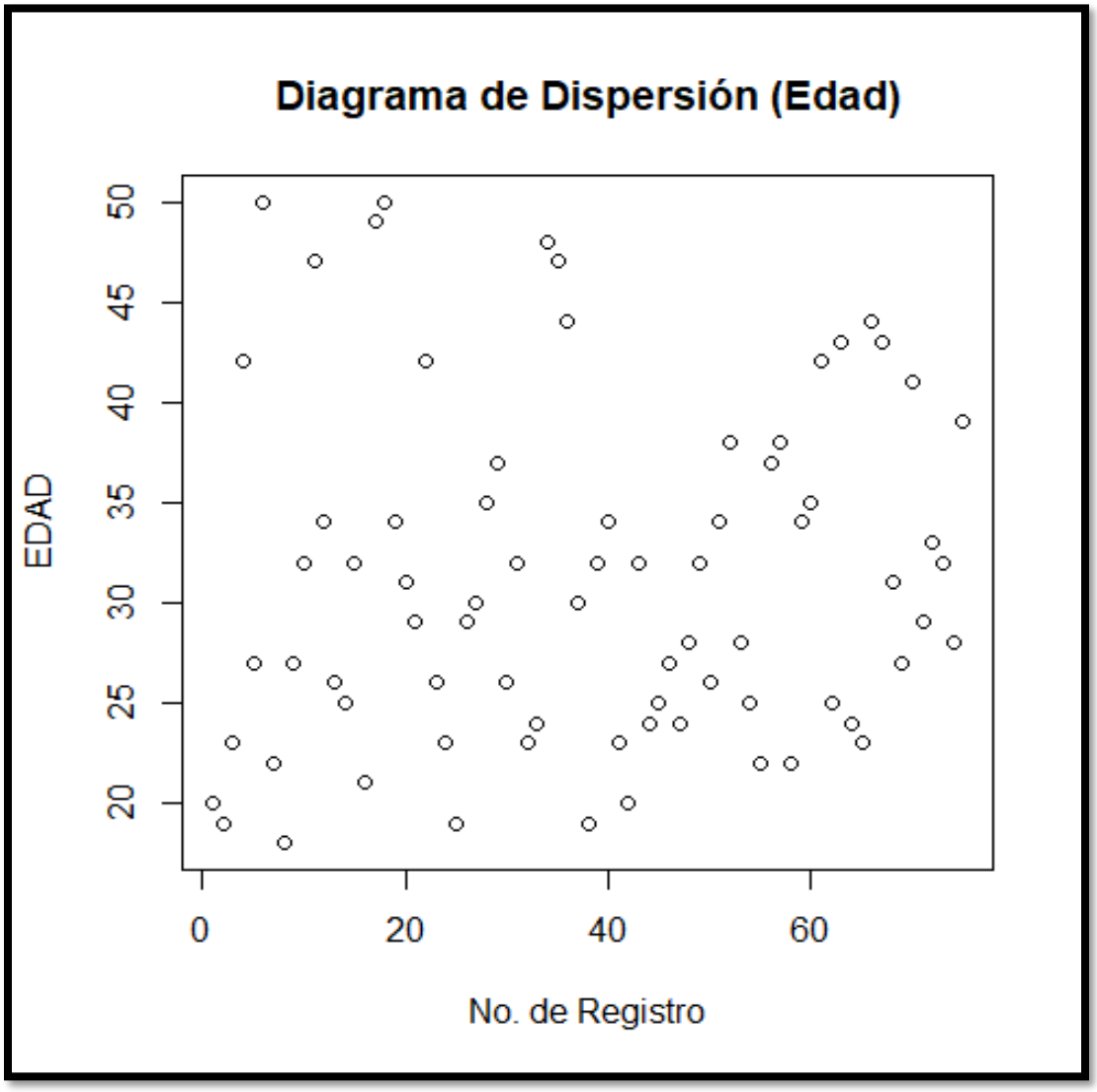


Diagrama de dispersión de COMPLEXIÓN FÍSICA en relación al número de registro.

plot(COMPLEXION.FISICA, xlab="No. de Registro", main = "Diagrama de Dispersión (Complexión Física)")

EXPLICACIÓN. - Al ser una variable de tres valores posibles, el análisis se hace de manera que visualicemos la cantidad de cada tipo de complexión en cierto rango del número de registros. Con esto, vemos que:

- En los registros del 0 al 20: Predominan ectomorfos y hay menos mesomorfos.
- En los registros del 20 al 40: Predominan mesomorfos sobre ectomorfos y endomorfos.
- En los registros del 40 al 60: Predominan ectomorfos y mesomorfos sobre endomorfos.
- En los registros del 60 al 75: Predominan mesomorfos y endomorfos sobre ectomorfos.

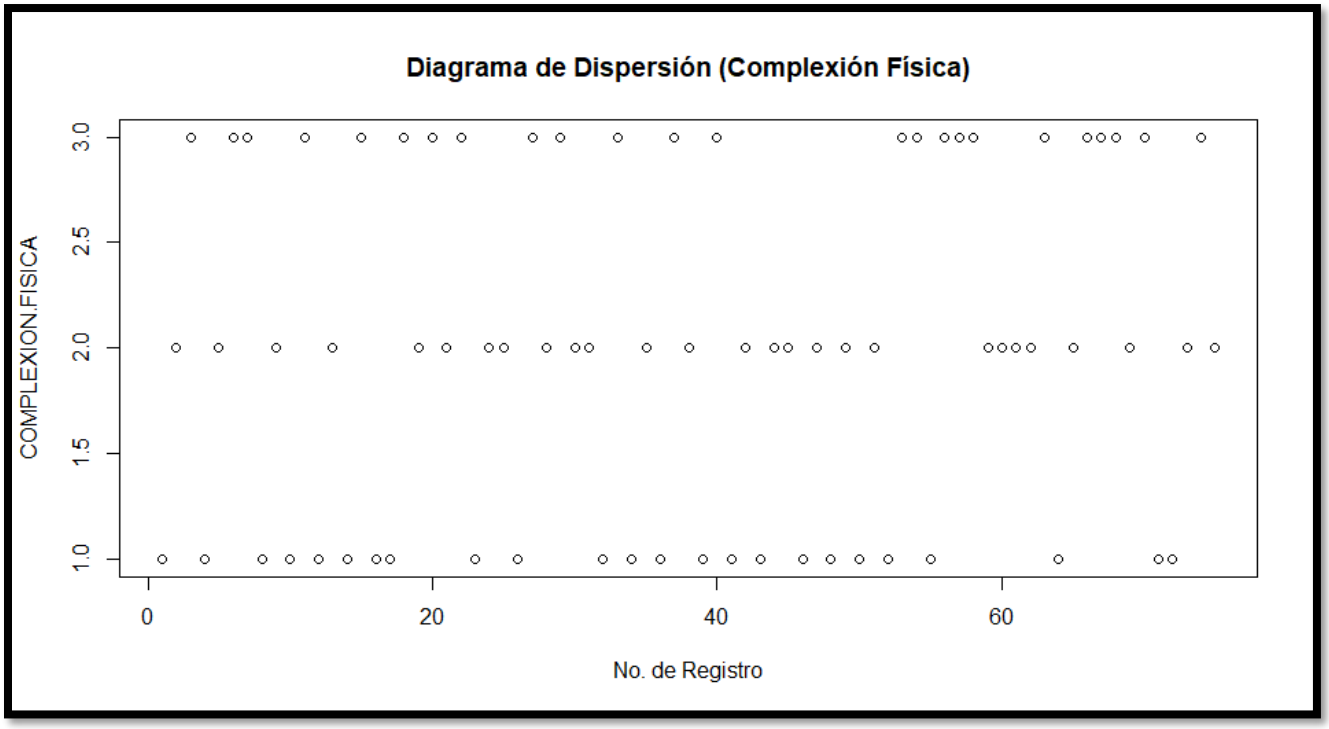


Diagrama de dispersión de NACIONALIDAD en relación al número de registro.

plot(NACIONALIDAD, xlab="No. de Registro", main="Diagrama de Dispersión (Nacionalidad)")

EXPLICACIÓN. –

- En los registros del 0 al 20: No existe diferencia significativa, pero hay más europeos.
- En los registros del 20 al 40: Sin diferencia significativa, hay más americanos y menos europeos.
- En los registros del 40 al 60: Hay más europeos y menos personas americanas.
- En los registros del 60 al 75: Hay más personas asiáticas y europeas que americanas.

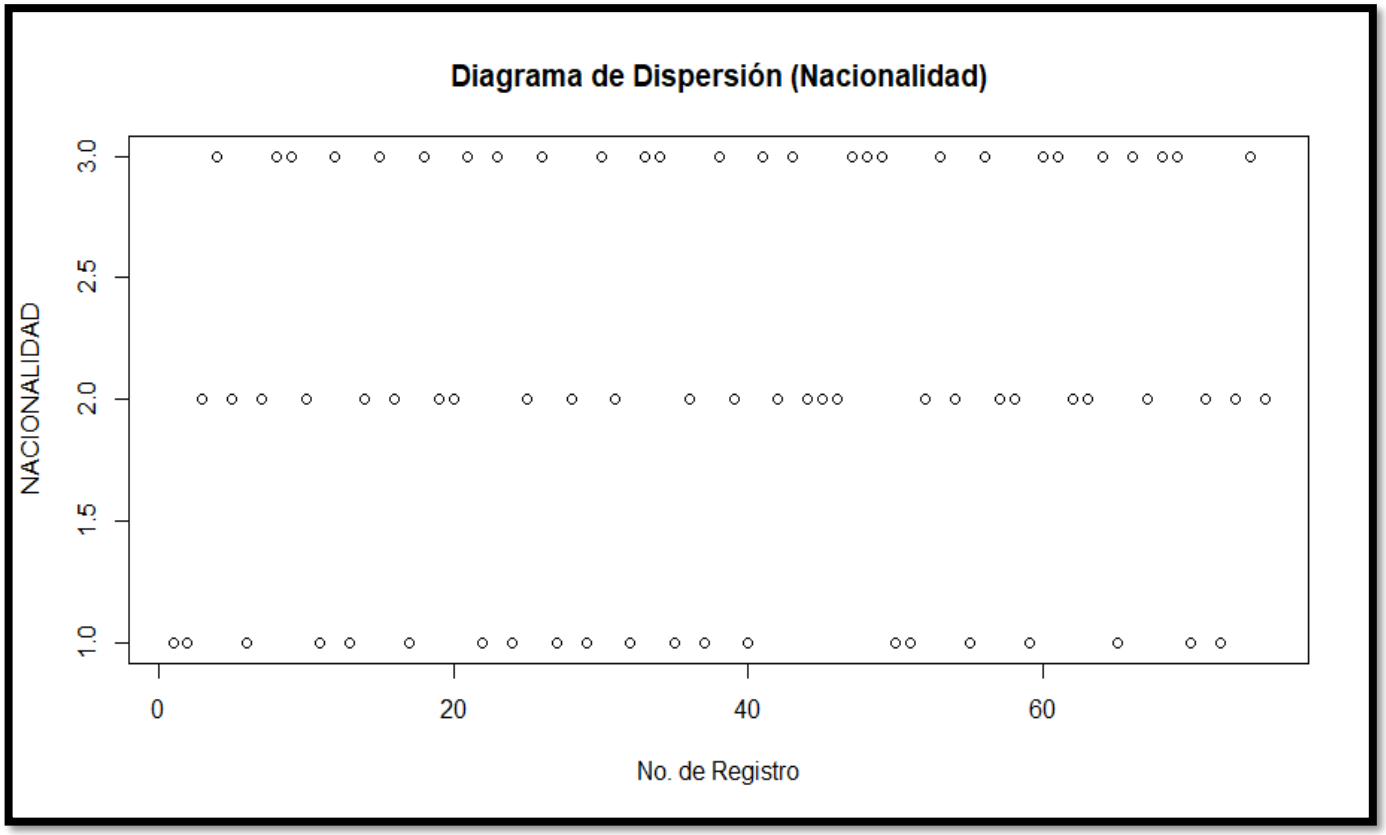
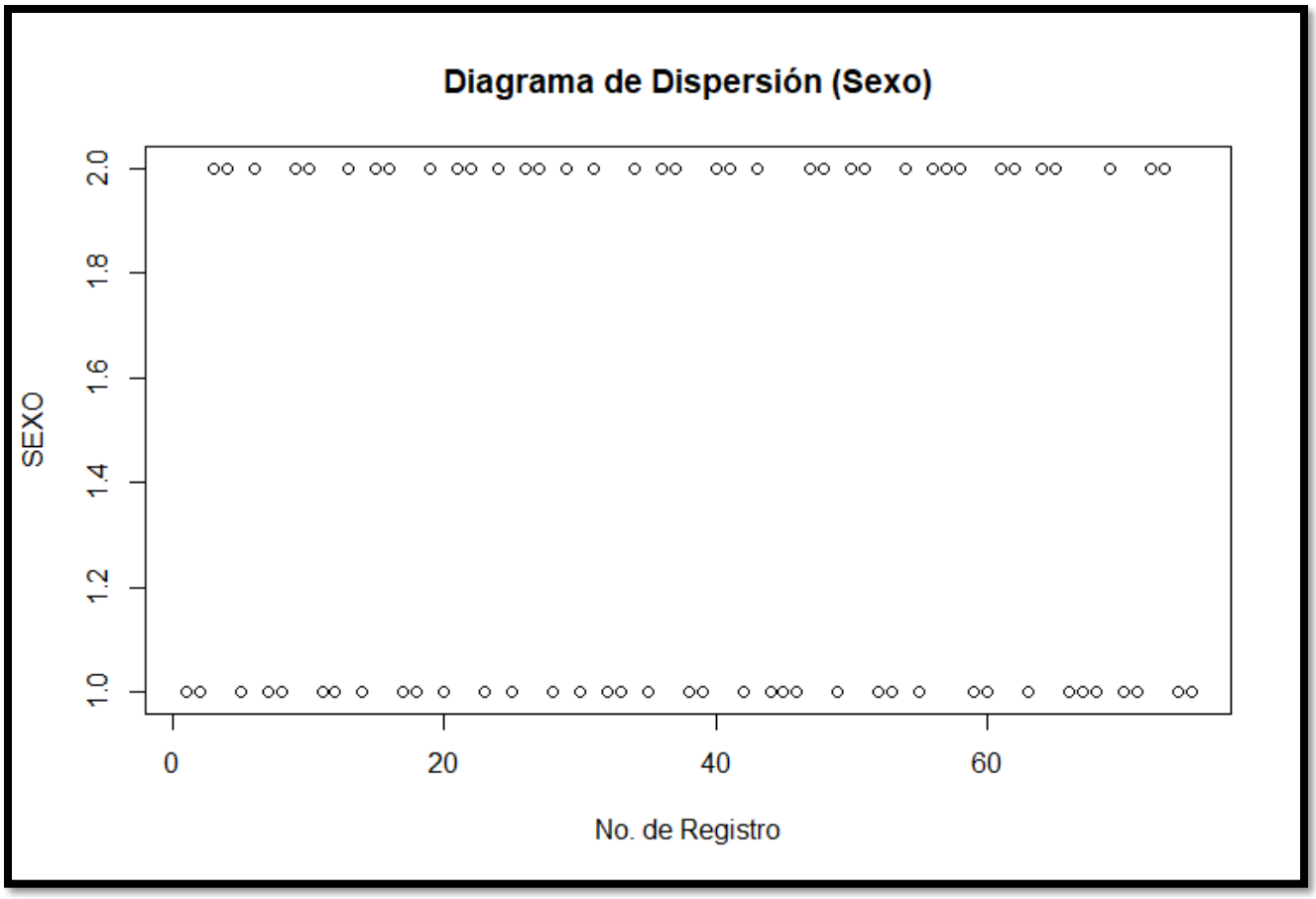


Diagrama de dispersión de SEXO en relación al número de registro.

plot(SEXO, xlab="No. de Registro", main = "Diagrama de Dispersión (Sexo)")

EXPLICACIÓN. -

- En los registros del 0 al 20: Sin diferencia significativa, hay más hombres que mujeres.
- En los registros del 20 al 40: Sin diferencia significativa, hay más mujeres que hombres.
- En los registros del 40 al 60: Sin diferencia significativa, hay más mujeres que hombres.
- En los registros del 60 al 75: Sin diferencia significativa, hay más hombres que mujeres.



4.- Análisis de correlaciones.

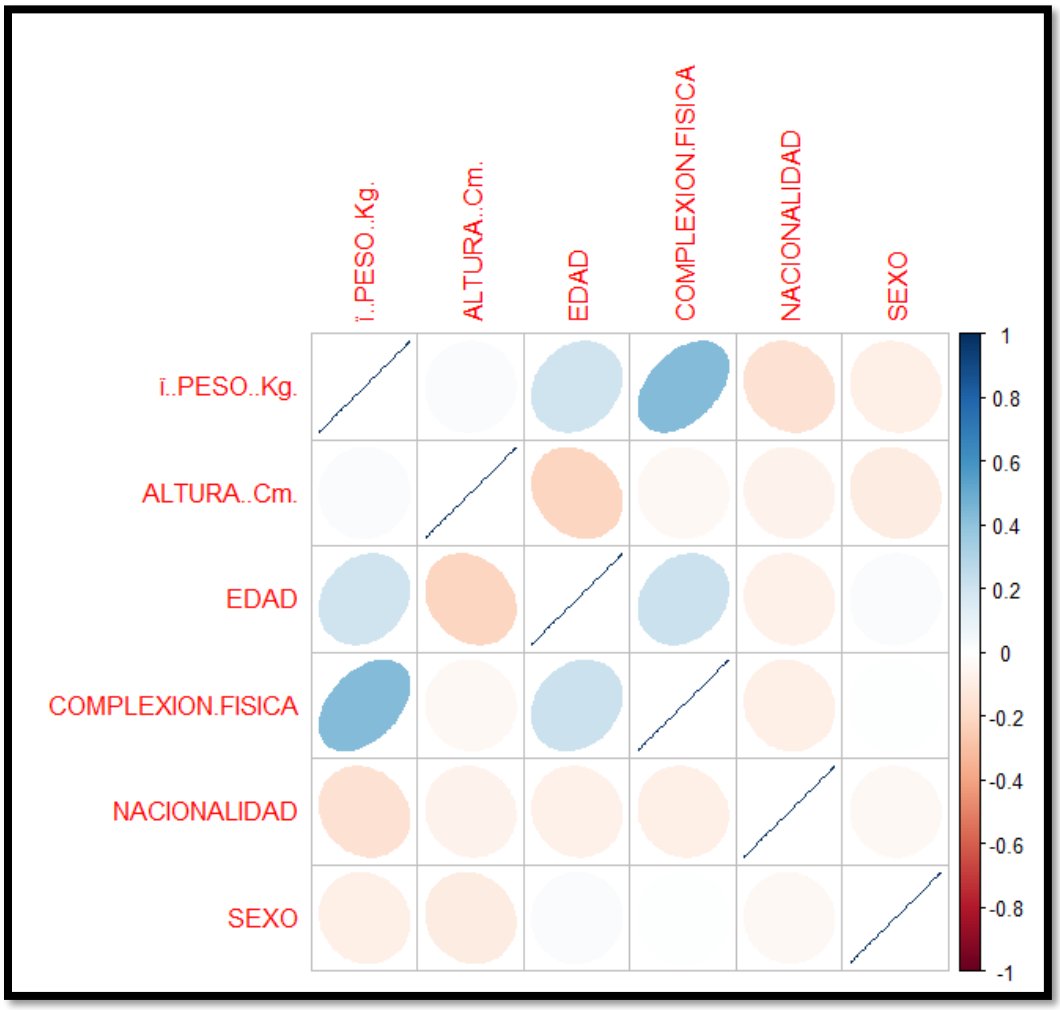
Llamar a la librería 'corrplot'

library(corrplot)

Usar comando 'corrplot' para generar gráfica visual de la correlación entre nuestras variables independientes, considerando un rango de valores de -1 a 1 (valor de correlación) que se representa con un gradiente de colores: azul, blanco y rojo. La figura escogida fue la elipse.

corrplot(cor(archivoleido), method = "ellipse")

EXPLICACIÓN.- Podemos observar que, la mayoría de las variables no tienen un nivel significativo de correlación entre sí. Al igual que existen valores positivos y negativos de correlación. La mayor correlación positiva es entre la COMPLEXIÓN FÍSICA y el PESO, lo que indica que ambas variables (sus valores) incrementan una con la otra. La mayor correlación negativa es entre la EDAD y la ALTURA (QUE ES MÍNIMA), lo que indica que, si la variable de EDAD incrementa, la variable ALTURA disminuye.



5.- Ajuste del modelo y cálculo de los coeficientes o parámetros

Ajuste de modelo.-

Variable Dependiente ~ Variable(s) Independiente(s)

lm(formula = ï..PESO..Kg. ~ ALTURA..Cm. + EDAD + COMPLEXION.FISICA + NACIONALIDAD + SEXO)

modeloarchivoleido <-
lm(ï..PESO..Kg.~ALTURA..Cm.+EDAD+COMPLEXION.FISICA+NACIONALIDAD+SEXO)

Determinar coeficientes.-

Con 'summary()' calculamos los coeficientes y determinamos cuales de nuestras variables independientes tienen mayor influencia en la variable dependiente de acuerdo a el número de asteriscos presentes en cada una. Tenemos también valores como el estimado (Estimate) y el error(Std. Error).

summary(modeloarchivoleido)

ANÁLISIS.-

```
#Coefficients:
#                 Estimate Std. Error t value Pr(>|t|)
#(Intercept)       50.25600   42.67590   1.178 0.242994
#ALTURA..Cm.        0.09059    0.22294   0.406 0.685731 (NO influye en 'PESO')
#EDAD               0.23950    0.22528   1.063 0.291432 (NO influye en 'PESO')
#COMPLEXION.FISICA  8.64831    2.31385   3.738 0.000379 *** (SÍ influye en 'PESO')
#NACIONALIDAD      -2.42596    2.28934  -1.060 0.292986 (NO influye en 'PESO')
#SEXO              -3.04380    3.63245  -0.838 0.404954 (NO influye en 'PESO')
#Multiple R-squared:  0.2281, Adjusted R-squared:  0.1722
```

```
Call:
lm(formula = ï..PESO..Kg. ~ ALTURA..Cm. + EDAD + COMPLEXION.FISICA +
    NACIONALIDAD + SEXO)

Residuals:
    Min      1Q  Median      3Q     Max
-48.583 -10.129  -2.389   9.066  34.714

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      50.25600   42.67590   1.178 0.242994
ALTURA..Cm.       0.09059    0.22294   0.406 0.685731
EDAD              0.23950    0.22528   1.063 0.291432
COMPLEXION.FISICA 8.64831    2.31385   3.738 0.000379 ***
NACIONALIDAD     -2.42596    2.28934  -1.060 0.292986
SEXO             -3.04380    3.63245  -0.838 0.404954
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.63 on 69 degrees of freedom
Multiple R-squared:  0.2281,    Adjusted R-squared:  0.1722
F-statistic: 4.078 on 5 and 69 DF,  p-value: 0.002658
```

Coeficientes del modelo.-

- 50.25600

- 0.09059
- 0.23950
- 8.64831 (Estadísticamente significativo)
- -2.42596
- -3.04380 (No es estadísticamente significativo)

- 0.2281, coeficiente de determinación múltiple, determina que tan cerca están los datos de la línea de regresión ajustada.

6.- Predicciones

Después de recabar 75 registros de manera aleatoria de las características físicas de personas de 3 nacionalidades, el encuestador desea predecir el nivel de peso en kg para la próxima persona de cierta nacionalidad a encuestar, y así poder saber si existe un patrón en el rango de peso de acuerdo a las demás características físicas.

DATOS. -

-Altura:180

-Edad: 29

-Complexión: 1

-Nacionalidad: 3

-Sexo: 1

predict(object = modeloarchivoleido, newdata = data.frame(ALTURA..Cm.=180,EDAD=29,

COMPLEXION.FISICA=1, NACIONALIDAD=3, SEXO=1), interval="predict", level=0.95)

```
> predict(object = modeloarchivoleido, newdata = data.frame(ALTURA..Cm.=180,EDAD=29,
+         COMPLEXION.FISICA=1, NACIONALIDAD=3, SEXO=1), interval="predict", level=0.95)
       fit      lwr      upr
1 71.83509 39.62261 104.0476
```

RESULTADO. -

El resultado de la predicción del PESO de una persona de altura 180, de 29 años, de complexión física delgada (ectomorfo), de nacionalidad asiática y de sexo 'hombre'... ha sido de 71 kg.

Conclusión. -

Este valor de peso ha sido mayormente determinado por la COMPLEXIÓN FÍSICA.Y se da a entender que, de acuerdo a los datos que tenemos, las personas asiáticas tienden a tener valores de peso más regulados/normales.

7.- Conclusiones del modelo de regresión lineal

Después de ajustar el modelo de regresión lineal, sabemos que nuestras variables (independientes) que tuvieron una repercusión sobre la variable dependiente o de respuesta (PESO) fueron: COMPLEXIÓN FÍSICA solamente. En cuanto a las variables NACIONALIDAD y SEXO podemos decir que fue el caso contrario, ya que, al existir una correlación negativa, el efecto de relación fue inverso, (aunque mínimo).

8.- Ajuste del modelo de regresión logística

ASIGNAR VARIABLE DEPENDIENTE E INDEPENDIENTE.-

Variable dependiente: SEXO

Variable independiente: COMPLEXIÓN FÍSICA

Para continuar con la investigación acerca de los patrones de características físicas específicas de acuerdo a alguna otra característica en especial, el encuestador está interesado en conocer la probabilidad de que una persona sea hombre o mujer dependiendo de su complexión física.

Convertir nuestra variable categórica en factor con 'as.factor()'.-

SEXO <- as.factor(SEXO)

modelo_logistico <- glm(formula = SEXO~COMPLEXION.FISICA,

family = binomial(link = "logit"))

Con 'summary.gml()' se muestra qué variables influyen

en la variable dependiente, dependiendo del número de asteriscos.

summary.glm(modelo_logistico)

```
#Call:
#glm(formula = SEXO ~ COMPLEXION.FISICA, family = binomial(link = "logit"))

#Coefficients:
#                   Estimate Std. Error z value Pr(>|z|)
#(Intercept)       -2.667e-02  6.219e-01  -0.043    0.966
#COMPLEXION.FISICA  2.564e-16  2.887e-01   0.000    1.000
```

CONCLUSIÓN.- La variable COMPLEXION.FISICA no influye en el correspondiente SEXO de los encuestados, lo que quiere decir que hay variedad/no se sigue un patrón de que cierto sexo tenga mayormente una cierta complexión. Dado esto, la probabilidad es casi nula.

9.- Predicciones del modelo logístico y conclusiones

Predecir dado un valor de complexión física de 3 (Endomorfo)

predict(modelo_logistico, newdata = data.frame(COMPLEXION.FISICA = 3))

RESULTADO= -0.02666825

Conclusiones del modelo de regresión logística.- La variable independiente COMPLEXIÓN FÍSICA no influye en la variable dependiente SEXO por tanto no es posible calcular una probabilidad coherente.

10. Coherencia entre análisis

A partir de todas las operaciones previamente realizadas y del análisis de la información recabada en la encuesta/base de datos, podemos decir que el hecho de que no exista la suficiente correlación entre todas y cada una de las variables es el hecho de que la información presenta datos muy variados entre los 75 registros, en otras palabras, no existe un 'sesgo' entre la información que se presenta en los registros y cómo esta se va relacionando a lo largo de la tabla, por lo que se puede decir que hay gran variedad de casos a pesar de ser solamente 75 filas.

Los modelos de regresión nos ayudaron a enfocar el análisis de esta correlación entre una variable dependiente o de respuesta y varias variables independientes o 'de regresión', haciéndonos concluir que hubo un mínimo valor de correlaciones, solo una variable tuvo relevancia. La otra utilidad de contar con estos modelos fue que pudimos aplicar predicciones a partir de los datos con los que ya contábamos.