Informe limpieza

Lilia Rojas_2020

2/11/2020

#Informe de Análisis

Este documento pretende aclarar los pasos que se dieron en la limpieza de la data y explicar el proceso.

##Conversión de factores

Fue necesario adaptar modificar algunos datos en las bases de datos (BD), con la finalidad de hacerlo más fáciles de procesar para nuestros propósitos. En la BD "bd_col" fue necesario cambiar las definiciones de la modalidad Sena, para obtener las respuestas en sólo dos categorías: "Sí" para las modalidades que incluyen programación y "No para las que no la incluyen.

```
db_col$upz_fct <- factor(fct_collapse(db_col$upz,
                                      No = c("ARTES GRAFICAS Y DISEÑO","CIENCIAS DEL DEPORTE","CISCO",
                                             "N.A", "PENSAMIENTO AMBIENTAL", "PREINGENIERIAS", "TÉCNICO
                                             "TÉCNICO EN DESARROLLO DE OPERACIONES LOGÍSTICAS EN LA CAD
                                             "TECNICO EN PREPRENSA DIGITAL PARA MEDIOS IMPRESOS", "TÉCN
                                             "TÉCNICO EN VENTA DE PRODUCTOS Y SERVICIOS", "N.A"),
                                      Si = c("TÉCNICO EN DISEÑO E INTEGRACIÓN DE MULTIMEDIA", "TÉCNICO
))

save(db_col, file="db_col.RData")

db_test$upz_fct <- factor(fct_collapse(db_test$upz,
                                      No = c("Ciencias del deporte","CISCO","Humanidades y medios de co
                                      Si = c("Técnico en diseño e integración multimedia", "Técnico en
))
```

En el caso de la BD "db_test" también se hizo una conversión similar para convertir las respuestas en: "1" para las correctas y "0" para las incorrectas.

```
#--Respuestas A

db_test$PPregunta.7 <- factor(fct_collapse(db_test$Pregunta.7,
                                           "1" = "A",
                                           "0" = c("B", "C", "D", " ")))

db_test$PPregunta.12 <- factor(fct_collapse(db_test$Pregunta.12,
                                           "1" = "A",
                                           "0" = c("B", "C", "D", " ")))

db_test$PPregunta.14 <- factor(fct_collapse(db_test$Pregunta.14,
                                           "1" = "A",
                                           "0" = c("B", "C", "D", " ")))
```

```
db_test$PPregunta.18 <- factor(fct_collapse(db_test$Pregunta.18,
                                            "1" = "A",
                                            "0" = c("B", "C", "D", " ")))
db_test$PPregunta.21 <- factor(fct_collapse(db_test$Pregunta.21,
                                            "1" = "A",
                                            "0" = c("B", "C", "D", " ")))
db_test$PPregunta.23 <- factor(fct_collapse(db_test$Pregunta.23,
                                            "1" = "A",
                                            "0" = c("B", "C", "D", " ")))
db_test$PPregunta.27 <- factor(fct_collapse(db_test$Pregunta.27,
                                            "1" = "A",
                                            "0" = c("B", "C", "D", " ")))



#--REspuestas B
db_test$PPregunta.1 <- factor(fct_collapse(db_test$Pregunta.1,
                                            "1" = "B",
                                            "0" = c("A", "C", "D", " ")))

db_test$PPregunta.8 <- factor(fct_collapse(db_test$Pregunta.8,
                                            "1" = "B",
                                            "0" = c("A", "C", "D", " ")))
db_test$PPregunta.13 <- factor(fct_collapse(db_test$Pregunta.13,
                                            "1" = "B",
                                            "0" = c("A", "C", "D", " ")))
db_test$PPregunta.17 <- factor(fct_collapse(db_test$Pregunta.17,
                                            "1" = "B",
                                            "0" = c("A", "C", "D", " ")))
db_test$PPregunta.19 <- factor(fct_collapse(db_test$Pregunta.19,
                                            "1" = "B",
                                            "0" = c("A", "C", "D", " ")))
db_test$PPregunta.22 <- factor(fct_collapse(db_test$Pregunta.22,
                                            "1" = "B",
                                            "0" = c("A", "C", "D", " ")))
db_test$PPregunta.25 <- factor(fct_collapse(db_test$Pregunta.25,
                                            "1" = "B",
                                            "0" = c("A", "C", "D", " ")))
db_test$PPregunta.26 <- factor(fct_collapse(db_test$Pregunta.26,
                                            "1" = "B",
                                            "0" = c("A", "C", "D", " ")))



#--Respuestas C
db_test$PPregunta.2 <- factor(fct_collapse(db_test$Pregunta.2,
                                            "1" = "C",
                                            "0" = c("A", "B", "D", " ")))

db_test$PPregunta.5 <- factor(fct_collapse(db_test$Pregunta.5,
                                            "1" = "C",
                                            "0" = c("A", "B", "D", " ")))

db_test$PPregunta.10 <- factor(fct_collapse(db_test$Pregunta.10,
```

```
                                          "1" = "C",
                                          "0" = c("A", "B", "D", " ")))
db_test$PPregunta.11 <- factor(fct_collapse(db_test$Pregunta.11,
                                          "1" = "C",
                                          "0" = c("A", "B", "D", " ")))
db_test$PPregunta.20 <- factor(fct_collapse(db_test$Pregunta.20,
                                          "1" = "C",
                                          "0" = c("A", "B", "D", " ")))
db_test$PPregunta.24 <- factor(fct_collapse(db_test$Pregunta.24,
                                          "1" = "C",
                                          "0" = c("A", "B", "D", " ")))
db_test$PPregunta.28 <- factor(fct_collapse(db_test$Pregunta.28,
                                          "1" = "C",
                                          "0" = c("A", "B", "D", " ")))


#--REspuestas D

db_test$PPregunta.3 <- factor(fct_collapse(db_test$Pregunta.3,
                                          "1" = "D",
                                          "0" = c("A", "B", "C", " ")))
db_test$PPregunta.4 <- factor(fct_collapse(db_test$Pregunta.4,
                                          "1" = "D",
                                          "0" = c("A", "B", "C", " ")))
db_test$PPregunta.6 <- factor(fct_collapse(db_test$Pregunta.6,
                                          "1" = "D",
                                          "0" = c("A", "B", "C", " ")))
db_test$PPregunta.9 <- factor(fct_collapse(db_test$Pregunta.9,
                                          "1" = "D",
                                          "0" = c("A", "B", "C", " ")))
db_test$PPregunta.15 <- factor(fct_collapse(db_test$Pregunta.15,
                                          "1" = "D",
                                          "0" = c("A", "B", "C", " ")))
db_test$PPregunta.16 <- factor(fct_collapse(db_test$Pregunta.16,
                                           "1" = "D",
                                           "0" = c("A", "B", "C", " ")))
```

A renglón seguido se incorpora la sumatoria de los datos y se guarda como “db_test1”

```
db_test <- data.frame(db_test)
write.xlsx(db_test, file="db_test.xlsx")

db_test1 <- "db_test1.xlsx"
db_test1 <- data.frame(read_excel(db_test1))
save(db_test1, file="db_tes1.RData")

db_col1 <- data.frame(db_col)
db_test1 <- data.frame(db_test1)

#Copias por seguridad
write.xlsx(db_col1, file="db_col1.xlsx")
write.xlsx(db_test1, file="db_test1.xlsx")
save(db_col1, file="db_col1.RData")
```

```
save(db_test1, file="db_test1.RData")
```

El siguiente paso es mezclar las tablas, (en este caso las copias, para evitar perder los datos en caso de un mal proceso), exminar la tabla y eliminar las columnas que tienen datos incompletos o que no aportan a el análisis. Una vez obtenida la DB "db_basica" se guarda como archivo y como RData, por si es necesario recuperar algún dato guardado en la base consolidada.Las dimensiones de la BD son 75 registros por 75 variables.

```
db_basica <- merge (db_col1, db_test1, by="apellidos", all.y = TRUE)
db_basica <- select(db_basica, -Ejemplo.I, -Ejemplo.II, -Ejemplo.III, -No, -sexo.x, -curso.x, -cfk.x, -
write.xlsx(db_basica, file="db_basica.xlsx")

save(db_basica, file="db_basica.RData")
dim(db_basica) #75 registros 75 variables
```

```
[1] 75 75
```

Finalmente, se crea una base de trabajo seleccionando las variables de interés. La llamamos "db_limpia", pues no tiene variables ajenas al análisis que se va a ahacer. Sinembargo, siempre podremos incluir otras variables disponibles en la "db_basica", según las necesidades del análisis. Por eso, para salvaguardar su integridad, se guarda en formato .xlsx y .RData.

```
db_limpia <-  select(db_basica, result=puntaje_total,sexo= sexo.y, cfk =cfk.y, upz=upz_fct.y, edad = eda
write.xlsx(db_limpia, file="db_limpia.xlsx")
save(db_limpia, file="db_limpia.RData")
```

La BD "db_limpia" será la BD usada para el proceso de análisis.