Proyecto1

July 18, 2021

0.0.1 Proyecto 1 - Ciencia de datos

Integrentes:

- Karina Valladares, 18005
- Alexa Bravo, 18831
- José Eduardo López, 181045

[1]:
```
#importamos las librerías que vamos a utilizar

import pandas as pd #para trabajar con la base de datos y el data frame
from pandas_profiling import * #para generar code book
```

[2]:
```
df = pd.read_csv("data_mineduc.csv", sep = ";") #leeemos el .csv (NOTA: hay que␣
↪separar por punto y coma porque así lo genera Tableau)
```

[3]:
```
df
```

[3]:
```
               CODIGO DISTRITO   DEPARTAMENTO                    MUNICIPIO  \
0      10-01-0001-42   10-001  SUCHITEPEQUEZ                  MAZATENANGO
1      10-01-0002-42      NaN  SUCHITEPEQUEZ                  MAZATENANGO
2      10-01-0003-42   10-018  SUCHITEPEQUEZ                  MAZATENANGO
3      10-01-0004-42   10-019  SUCHITEPEQUEZ                  MAZATENANGO
4      10-01-0005-42   10-019  SUCHITEPEQUEZ                  MAZATENANGO
…                ...      …            …                             …
33361  08-08-2420-41   08-022    TOTONICAPAN  SAN BARTOLO AGUAS CALIENTES
33362  08-08-2421-43   08-022    TOTONICAPAN  SAN BARTOLO AGUAS CALIENTES
33363  08-08-2446-45   08-022    TOTONICAPAN  SAN BARTOLO AGUAS CALIENTES
33364  08-08-2447-45   08-022    TOTONICAPAN  SAN BARTOLO AGUAS CALIENTES
33365  08-08-2506-41   08-022    TOTONICAPAN  SAN BARTOLO AGUAS CALIENTES

                                      ESTABLECIMIENTO  \
0                                            EODP NO.1
1          EPDP ANEXA A COLEGIO 'REPUBLICA DE GUATEMALA'
2                                    COLEGIO "FROEBEL"
3      INSTITUTO PRIVADO MIXTO TECNOLOGICO DEL SUROCC…
4                  COLEGIO PRIVADO MIXTO LICEO MAZATECO
…                                                   …
```

```
33361                                COPB ANEXO A EORM
33362                                             EORM
33363  NUCLEO FAMILIAR EDUCATIVO PARA EL DESARROLLO N…
33364  NUCLEO FAMILIAR EDUCATIVO PARA EL DESARROLLO N…
33365                                COPB ANEXO A EORM

                                            DIRECCION           TELEFONO  \
0      11 CALLE FINAL ZONA 1  FRENTE INST. TECNICO IN…           58359230
1                               3A AVE. 7A CALLE ZONA 1                NaN
2              11. CALLE FINAL ZONA 3 COLONIA SAN ANDRES           41341472
3                                        FINCA CHOJOJA  78722582-52932245
4                                    6A AVE. 2-21 ZONA 1           40281538
…                                                    …                …
33361                       PARAJE CHUIAJ ALDEA TZANJON           40574929
33362                       PARAJE CHUIAJ ALDEA TZANJON           40574929
33363               CASERIO CHOCACULEU ALDEA PARRAXCHAJ           50495174
33364                                ALDEA CHOCANULEU             47702945
33365                 PARAJE PACHOC ALDEA TIERRA BLANCA           46293218

                                      SUPERVISOR  \
0                      RITA ELENA GONZALEZ GARCIA
1                                             NaN
2                     JUANA MARIA CIFUENTES PEREZ
3      REYNA DEL ROSARIO CHAVEZ BARROW DE PACHECO
4      REYNA DEL ROSARIO CHAVEZ BARROW DE PACHECO
…                                             …
33361               JOSÉ FELICIANO PÉREZ CHAMPET
33362               JOSÉ FELICIANO PÉREZ CHAMPET
33363               JOSÉ FELICIANO PÉREZ CHAMPET
33364               JOSÉ FELICIANO PÉREZ CHAMPET
33365               JOSÉ FELICIANO PÉREZ CHAMPET

                                       DIRECTOR                NIVEL  \
0                                           NaN             PARVULOS
1                                           NaN             PARVULOS
2      LIGIA ALEJANDRA CIFUENTES GODOY DE LOPEZ             PARVULOS
3               JOSE FELICIANO ESCOBAR MARTINEZ             PARVULOS
4                  SANDRY ZIOMARA MANCIO MENDEZ             PARVULOS
…                                          …                    …
33361               JULIO CRISTOBAL PUAC TUMAX  PREPRIMARIA BILINGUE
33362               JULIO CRISTOBAL PUAC TUMAX              PRIMARIA
33363        ZOILA RADARANY CHAMORRO VELÁSQUEZ                BASICO
33364               JAMES DANIEL SONTAY TORRES                BASICO
33365    ERICKA ANABELLA XILOJ PEREZ DE CAPRIEL  PREPRIMARIA BILINGUE

        SECTOR    AREA                STATUS   MODALIDAD     JORNADA  \
0      OFICIAL  URBANA               ABIERTA  MONOLINGUE    MATUTINA
```

```
1       PRIVADO  URBANA  CERRADA TEMPORALMENTE  MONOLINGUE    MATUTINA
2       PRIVADO  URBANA                 ABIERTA  MONOLINGUE    MATUTINA
3       PRIVADO  URBANA                 ABIERTA  MONOLINGUE    MATUTINA
4       PRIVADO  URBANA                 ABIERTA  MONOLINGUE    MATUTINA
…         …      …                      …          …          …
33361   OFICIAL   RURAL                 ABIERTA    BILINGUE    MATUTINA
33362   OFICIAL   RURAL                 ABIERTA    BILINGUE    MATUTINA
33363   OFICIAL   RURAL                 ABIERTA  MONOLINGUE  VESPERTINA
33364   OFICIAL   RURAL                 ABIERTA  MONOLINGUE  VESPERTINA
33365   OFICIAL   RURAL                 ABIERTA    BILINGUE    MATUTINA

                   PLAN  DEPARTAMENTAL
0       DIARIO(REGULAR)  SUCHITEPÉQUEZ
1       DIARIO(REGULAR)  SUCHITEPÉQUEZ
2       DIARIO(REGULAR)  SUCHITEPÉQUEZ
3       DIARIO(REGULAR)  SUCHITEPÉQUEZ
4       DIARIO(REGULAR)  SUCHITEPÉQUEZ
…                  …           …
33361   DIARIO(REGULAR)    TOTONICAPÁN
33362   DIARIO(REGULAR)    TOTONICAPÁN
33363   DIARIO(REGULAR)    TOTONICAPÁN
33364   DIARIO(REGULAR)    TOTONICAPÁN
33365   DIARIO(REGULAR)    TOTONICAPÁN

[33366 rows x 17 columns]
```

[7]:
```
df.isna().sum() #revisamos si hay datos faltantes en nuestras variables
```

[7]:
```
CODIGO                0
DISTRITO           1676
DEPARTAMENTO         22
MUNICIPIO            22
ESTABLECIMIENTO      28
DIRECCION           303
TELEFONO           8374
SUPERVISOR         1678
DIRECTOR          12810
NIVEL                22
SECTOR               22
AREA                 22
STATUS               22
MODALIDAD            22
JORNADA              22
PLAN                 22
DEPARTAMENTAL        22
dtype: int64
```

[8]:
```
df.isnull().sum() #revisamos si hay datos faltantes en nuestras variables
```

[8]:
```
CODIGO                0
DISTRITO           1676
DEPARTAMENTO         22
MUNICIPIO            22
ESTABLECIMIENTO      28
DIRECCION           303
TELEFONO           8374
SUPERVISOR         1678
DIRECTOR          12810
NIVEL                22
SECTOR               22
AREA                 22
STATUS               22
MODALIDAD            22
JORNADA              22
PLAN                 22
DEPARTAMENTAL        22
dtype: int64
```

0.0.2 Descripción de los datos

- Cantidad de columnas: 17 (variables)
- Cantidad de filas: 33366 (registros)

0.0.3 Lista de variables

- CODIGO: 0 faltantes (no se requiere limpieza). Es un número asignado por el Ministerio de Educación al establecimiento.
- DISTRITO: 1676 faltantes (complicado de limpiar). Sección asignada, por medio de un código particular.
- DEPARTAMENTO: 22 faltantes (fácil de limpiar). El nombre del departamento al que pertenece la institución.
- MUNICIPIO: 22 faltantes (fácil de limpiar). Nivel municipal.
- ESTABLECIMIENTO: 28 faltantes (fácil de limpiar). Nombre de la institución educativa.
- DIRECCION: 303 faltantes (dificultad media de limpieza). Dirección del establecimiento.
- TELEFONO: 8,374 faltantes (complicado de limpiar). Número telefónico del establecimiento.
- SUPERVISOR: 1,678 faltantes (complicado de limpiar). Trabajador asignado por el Ministerio de Educación según el distrito al que pertenece.
- DIRECTOR: 12,810 faltantes (complicado de limpair). Alto mando en el establecimiento.
- NIVEL: 22 faltantes (fácil de limpiar). Alto nivel de educación impartida.
- SECTOR: 22 faltantes (fácil de limpiar). Si pertenece al sector privado o es público.

- AREA: 22 faltantes (fácil de limpiar). Si está ubicado en un área rural o urbana.
- STATUS: 22 faltantes (fácil de limpiar). Si el establecimiento está abierto, o cerrado.
- MODALIDAD: 22 faltantes (fácil de limpiar). Monolingue, bilingue o poliglota.
- JORNADA: 22 faltantes (fácil de limpiar). Si es matutina, vespertina, nocturna o cualquier combinación de estas.
- PLAN: 22 faltantes (fácil de limpiar). Si está en un plan diario, fines de semana, sábado o a distancia.
- DEPARTAMENTAL: 22 faltantes (fácil de limpiar). La jurisdicción a la que pertenece el establecimiento.

0.0.4 Estrategia de limpieza

- Para las variables DEPARTAMENTO, MUNICIPIO, ESTABLECIMIENTO, NIVEL, SECTOR, AREA, STATUS, MODALIDAD, JORNADA, PLAN, DEPARTAMENTAL tienen pocos datos faltantes por lo que podemos simplemente eliminar los registros dado que son relativamente pocos (menos de 300) comparado con la cantidad total de registros (33, 366).
- DIRECCION: revisar errores ortográficos y duplicados.
- SUPERVISOR: intentar limpiar las columnas para solo dejar palabras clave. Revisar errores ortográficas y duplicados.
- DIRECTOR: eliminar la variable dado que hay demasiados datos faltantes y no se considera relevante para un análisis posterior.
- TELEFONO: asegurarse de que quede únicamente un número de teléfono por columna, eliminar los guiones o espacios en blanco. Separar los números de teléfono de forma adecuada.
- AREA: pasar a binaria (1 rural, 0 para urbana).
- STATUS: pasar a binaria (1 abierto, 0 para cerrado).
- MODALIDAD: pasar a una variable numérica (2 políglota, 1 bilingüe, 0 monolingüe).

[14]:
```
profile = ProfileReport(df, title = "Reporte Proyecto", explorative = True)␣
↪#generamos el Codebook, le ponemos un titulo y que sea interactivo
```

[15]:
```
profile.to_file("Codebook.html") #imprimimos el HTML
```

```
Summarize dataset:   0%|          | 0/30 [00:00<?, ?it/s]

Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]

Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]

Export report to file:   0%|          | 0/1 [00:00<?, ?it/s]
```

0.0.5 Link a Github

https://github.com/karinaValladares/CC3066-Proyecto-1