02_Preprocesamiento

September 5, 2024

[ ]:
```
# Importamos librerías
import pandas as pd
import numpy as  np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
import unicodedata
warnings.filterwarnings('ignore')
```

[ ]:
```
df = pd.read_csv('data/data_raw.csv')
df.head()
```

[ ]:
```
                                              IDHASH  \
0  E4287C2FE19F63C5E6641955147E36684A5A2FF8064676…
1  2DC37F0B9727B6591EC72D7A942647797A200F45D47C9E…
2  3B8677B90781D7BB8F2F967C05FA2DBBE153BBB682DF05…
3  FA366704D9E9F6FB5E5F55C1FB0CEEE973C626A5616F55…
4  E31CF8F30F3AE60B3D8A14F6E1020E9AD26EE975F0823B…

                     COLEGIO COLEGIO_DEPA COLEGIO_PROV  \
0       LA DIVINA PROVIDENCIA         LIMA         LIMA
1           86019 LA LIBERTAD       ÁNCASH       HUARAZ
2  0113 DANIEL ALOMIAS ROBLES         LIMA         LIMA
3     SEBASTIAN SALAZAR BONDY         LIMA         LIMA
4           TRILCE LOS OLIVOS         LIMA         LIMA

             COLEGIO_DIST COLEGIO_PAIS  COLEGIO_ANIO_EGRESO  \
0              SURQUILLO         PERÚ                 2020
1                 HUARAZ         PERÚ                 2017
2  SAN JUAN DE LURIGANCHO         PERÚ                 2016
3            SANTA ANITA         PERÚ                 2018
4   SAN MARTÍN DE PORRES         PERÚ                 2020

                       ESPECIALIDAD  ANIO_POSTULA  CICLO_POSTULA  …  \
0           INGENIERÍA DE SISTEMAS          2021              1  …
1  INGENIERÍA DE TELECOMUNICACIONES          2021              1  …
2              INGENIERÍA MECÁNICA          2021              1  …
3           INGENIERÍA ELECTRÓNICA          2021              1  …
```

```
4                      ARQUITECTURA          2021              1  …

           DOMICILIO_DIST ANIO_NACIMIENTO NACIMIENTO_PAIS  NACIMIENTO_DEPA  \
0  SAN JUAN DE MIRAFLORES            2004            PERÚ             LIMA
1                  HUARAZ            2001            PERÚ           ÁNCASH
2  SAN JUAN DE LURIGANCHO            2000            PERÚ             LIMA
3             SANTA ANITA            2002            PERÚ             LIMA
4           INDEPENDENCIA            2004            PERÚ             LIMA

  NACIMIENTO_PROV    NACIMIENTO_DIST       SEXO CALIF_FINAL INGRESO  \
0            LIMA  VILLA EL SALVADOR  MASCULINO        3.11      NO
1          HUARAZ             HUARAZ   FEMENINO        3.58      NO
2            LIMA        JESÚS MARÍA  MASCULINO        8.04      NO
3            LIMA               LIMA  MASCULINO       10.82      NO
4            LIMA              COMAS   FEMENINO        6.21      NO

                                          MODALIDAD
0  EXTRAORDINARIO1 - DEPORTISTAS CALIFICADOS DE A…
1                                        ORDINARIO
2                                        ORDINARIO
3           EXTRAORDINARIO2 - INGRESO DIRECTO CEPRE
4                                        ORDINARIO

[5 rows x 22 columns]
```

[ ]:
```
# 1. Creamos una copia para editar el dataframe
df_1 = df.copy()
```

[ ]:
```
df_1.columns
```

[ ]:
```
Index(['IDHASH', 'COLEGIO', 'COLEGIO_DEPA', 'COLEGIO_PROV', 'COLEGIO_DIST',
       'COLEGIO_PAIS', 'COLEGIO_ANIO_EGRESO', 'ESPECIALIDAD', 'ANIO_POSTULA',
       'CICLO_POSTULA', 'DOMICILIO_DEPA', 'DOMICILIO_PROV', 'DOMICILIO_DIST',
       'ANIO_NACIMIENTO', 'NACIMIENTO_PAIS', 'NACIMIENTO_DEPA',
       'NACIMIENTO_PROV', 'NACIMIENTO_DIST', 'SEXO', 'CALIF_FINAL', 'INGRESO',
       'MODALIDAD'],
      dtype='object')
```

[ ]:
```
# 2.1 Cambiamos el nombre de las columnas para evitar que haya columnas␣
 ↪repetidas
df_1['COLEGIO_DEPA'] = df_1['COLEGIO_DEPA'].replace({
    'ÁNCASH': 'ANCASH',
    'JUNÍN': 'JUNIN',
    'HUÁNUCO': 'HUANUCO',
    'SAN MARTÍN': 'SAN MARTIN',
    'APURÍMAC': 'APURIMAC'
})
```

```
## Verificamos los valores únicos de la columna de Departamentos
df_1['COLEGIO_DEPA'].unique()
```

[ ]:
```
array(['LIMA', 'ANCASH', 'VENEZUELA', 'AYACUCHO', 'ICA', 'CAJAMARCA',
       'PIURA', 'AMAZONAS', 'TACNA', 'HUANUCO', 'JUNIN', 'CALLAO',
       'UCAYALI', 'LA LIBERTAD', 'APURIMAC', 'LAMBAYEQUE', 'SAN MARTIN',
       nan, 'HUANCAVELICA', 'CUSCO', 'AREQUIPA', 'PUNO', 'MOQUEGUA',
       'PASCO', 'MADRE DE DIOS', 'LORETO', 'ARGENTINA', 'TUMBES',
       'ECUADOR', 'CANADA', 'ESPAÑA', 'CHILE', 'ITALIA', 'COLOMBIA'],
      dtype=object)
```

[ ]:
```
# 2.2 Cambiamos el nombre de las columnas para evitar que haya columnas␣
 ↪repetidas
df_1['COLEGIO_PROV'] = df_1['COLEGIO_PROV'].replace({
    'PROV. CONST. DEL CALLAO': 'PROV CONST DEL CALLAO',
    'HUÁNUCO': 'HUANUCO',
    'SÁNCHEZ CARRIÓN': 'SANCHEZ CARRION',
    'CAÑETE': 'CANETE',
    'SAN ROMÁN': 'SAN ROMAN',
    'LA CONVENCIÓN': 'LA CONVENCION',
    'ASUNCIÓN': 'ASUNCION',
    'OYÓN': 'OYON',
    'HUAMALÍES': 'HUAMALIES',
    'AZÁNGARO': 'AZANGARO',
    'JAÉN': 'JAEN',
    'BONGARÁ': 'BONGARA',
    'MORROPÓN': 'MORROPON',
    'CARLOS FERMÍN FITZCARRALD': 'CARLOS FERMIN FITZCARRALD',
    'DANIEL ALCIDES CARRIÓN': 'DANIEL ALCIDES CARRION',
    'VILCAS HUAMÁN': 'VILCAS HUAMAN',
    'MARISCAL CÁCERES': 'MARISCAL CACERES',
    'VIRÚ': 'VIRU',
    'HUANCANÉ': 'HUANCANE',
    'PÁUCAR DEL SARA SARA': 'PAUCAR DEL SARA SARA',
    'LA UNIÓN': 'LA UNION',
    'JULCÁN': 'JULCAN',
    'ANTONIO RAIMONDI': 'ANTONIO RAYMONDI',
    'RODRÍGUEZ DE MENDOZA': 'RODRIGUEZ DE MENDOZA',
    'MARAÑÓN': 'MARANON',
    'CONTUMAZÁ': 'CONTUMAZA',
    'ITALIA.': 'ITALIA',
    'CARABOBO.': 'CARABOBO',
    'MARISCAL RAMÓN CASTILLA': 'MARISCAL RAMON CASTILLA',
    'VÍCTOR FAJARDO': 'VICTOR FAJARDO',
    'SÁNCHEZ CARRIÓN': 'SANCHEZ CARRION',
    'LAURICOCHA.': 'LAURICOCHA',
```

```
    'CALGARY.': 'CALGARY',
    'CONDORCANQUI.': 'CONDORCANQUI',
    'ECUADOR.': 'ECUADOR',
    'SUCRE.': 'SUCRE',
    'ATALAYA.': 'ATALAYA',
    'BOLÍVAR': 'BOLIVAR',
    'CASTROVIRREYNA.': 'CASTROVIRREYNA',
    'URUBAMBA.': 'URUBAMBA',
    'HUAROCHIRI.': 'HUAROCHIRI',
    'CAMANA.': 'CAMANA',
    'LAMBAYEQUE.': 'LAMBAYEQUE',
    'FERREÑAFE.': 'FERREÑAFE',
    'CARAVELÍ.': 'CARAVELI',
    'CHEPÉN': 'CHEPEN',
    'LA CONVENCION.': 'LA CONVENCION',
    'CARAVELI.': 'CARAVELI',
    'MAYNAS.': 'MAYNAS',
    'HUALGAYOC.': 'HUALGAYOC',
    'LORETO.': 'LORETO',
    'CUTERVO.': 'CUTERVO',
    'CAMANÁ.': 'CAMANA',
    'QUISPICANCHI.': 'QUISPICANCHI',
    'YUNGAY.': 'YUNGAY',
    'TRUJILLO.': 'TRUJILLO',
    'LAMPA.': 'LAMPA',
    'ESPINAR.': 'ESPINAR',
    'TAYACAJA.': 'TAYACAJA',
    'TARMA.': 'TARMA',
    'BOLOGNESI.': 'BOLOGNESI',
    'CHANCHAMAYO.': 'CHANCHAMAYO',
    'SULLANA.': 'SULLANA',
    'HUANCAYO.': 'HUANCAYO',
    'HUARI.': 'HUARI',
    'CHICLAYO.': 'CHICLAYO',
    'TOCACHE.': 'TOCACHE',
    'PISCO.': 'PISCO',
    'CHANCHAMAYO.': 'CHANCHAMAYO',
    'PROV CONST DEL CALLAO.': 'PROV CONST DEL CALLAO',
    'HUAROCHIRÍ': 'HUAROCHIRI',
    'VILCAS HUAMAN': 'VILCAS HUAMAN',
    'CAMANÁ': 'CAMANA',
    'LA UNIÒN': 'LA UNION',
    'PÀUCAR DEL SARA SARA': 'PAUCAR DEL SARA SARA',
    'SAN MARTÍN': 'SAN MARTIN',
    'CARAVELÍ': 'CARAVELI',
    'CELENDÍN': 'CELENDIN',
    'HUAYTARÁ': 'HUAYTARA',
```

```
    'CONCEPCIÓN': 'CONCEPCION',
    'VILCAS HUAMAN': 'VILCAS HUAMÁN',
    'LA UNION': 'LA UNIÒN',
    'JUNIN': 'JUNÍN',
    'PAUCAR DEL SARA SARA' : 'PÀUCAR DEL SARA SARA'

})

## Verificamos los valores únicos de la columna de Provincias
df_1['COLEGIO_PROV'].unique()
```

[ ]:
```
array(['LIMA', 'HUARAZ', 'VENEZUELA', 'HUAMANGA', 'CHINCHA', 'CELENDIN',
       'SULLANA', 'CHACHAPOYAS', 'POMABAMBA', 'TACNA', 'HUANUCO', 'TARMA',
       'PROV CONST DEL CALLAO', 'CORONEL PORTILLO', 'HUAURA',
       'SANCHEZ CARRION', 'ANDAHUAYLAS', 'JAUJA', 'CHICLAYO', 'TOCACHE',
       'HUARI', nan, 'HUARAL', 'PISCO', 'CAJAMARCA', 'CUTERVO',
       'HUANCAYO', 'SAN IGNACIO', 'TAYACAJA', 'BARRANCA', 'CANETE',
       'SATIPO', 'NASCA', 'SAN MARTIN', 'LAMBAYEQUE', 'HUAROCHIRI',
       'CUSCO', 'CASTROVIRREYNA', 'AREQUIPA', 'LUYA', 'CHOTA',
       'SAN ROMAN', 'HUANTA', 'ILO', 'JAEN', 'PUNO', 'CANTA', 'CARHUAZ',
       'CHANCHAMAYO', 'TRUJILLO', 'RECUAY', 'PASCO', 'HUANCAVELICA',
       'SANTA', 'HUALGAYOC', 'ICA', 'OXAPAMPA', 'TALARA', 'ABANCAY',
       'TAMBOPATA', 'BONGARA', 'TAHUAMANU', 'AZANGARO', 'MOYOBAMBA',
       'YAUYOS', 'OYON', 'CHUPACA', 'CONCEPCION', 'MAYNAS', 'PADRE ABAD',
       'MANU', 'LEONCIO PRADO', 'JUNÍN', 'MELGAR', 'GRAU', 'DOS DE MAYO',
       'COTABAMBAS', 'HUANCA SANCOS', 'ACOBAMBA', 'BOLOGNESI',
       'MARISCAL NIETO', 'CARABAYA', 'EL COLLAO', 'AYMARAES', 'YAULI',
       'SAN MIGUEL', 'ARGENTINA', 'VIRU', 'LA MAR', 'PIURA', 'ZARUMILLA',
       'CANCHIS', 'PARINACOCHAS', 'HUANCANE', 'LA CONVENCION', 'CARAVELI',
       'RODRIGUEZ DE MENDOZA', 'ESPINAR', 'ASUNCION', 'HUAMALIES',
       'HUAYLAS', 'MORROPON', 'ISLAY', 'SANTA CRUZ', 'LUCANAS',
       'CAJATAMBO', 'FERREÑAFE', 'UTCUBAMBA', 'TUMBES', 'SIHUAS', 'AMBO',
       'CAYLLOMA', 'SANDIA', 'ANTA', 'CARLOS FERMIN FITZCARRALD', 'BAGUA',
       'HUALLAGA', 'ANTONIO RAYMONDI', 'PACHITEA', 'RIOJA', 'ECUADOR',
       'LORETO', 'PÀUCAR DEL SARA SARA', 'MARISCAL LUZURIAGA', 'MARANON',
       'PALLASCA', 'YUNGAY', 'YAROWILCA', 'PACASMAYO',
       'DANIEL ALCIDES CARRION', 'SUCRE', 'ATALAYA', 'BOLIVAR', 'CALGARY',
       'CASMA', 'HUARMEY', 'URUBAMBA', 'UCAYALI', 'YUNGUYO', 'CHINCHEROS',
       'CHEPEN', 'ANGARAES', 'SECHURA', 'CANAS', 'PAITA', 'CHURCAMPA',
       'PALPA', 'FERRENAFE', 'CORONGO', 'HUACAYBAMBA', 'QUISPICANCHI',
       'MARISCAL CACERES', 'CAMANA', 'LAMPA', 'JORGE BASADRE', 'PATAZ',
       'VILCAS HUAMAN', 'SAN ANTONIO DE PUTINA', 'CANGALLO', 'CASTILLA',
       'CAJABAMBA', 'ALTO AMAZONAS', 'VILCAS HUAMÁN', 'CALCA',
       'SANTIAGO DE CHUCO', 'AYACUCHO', 'CARACAS', 'PUERTO INCA',
       'ACOMAYO', 'VICTOR FAJARDO', 'OTUZCO', 'SAN MARCOS', 'HUANCABAMBA',
       'CHUCUITO', 'CHUMBIVILCAS', 'MOHO', 'LA UNIÒN', 'LAURICOCHA',
       'EL DORADO', 'BELLAVISTA', 'JULCAN', 'PICOTA', 'CARABOBO',
```

```
       'LA UNION', 'CHILE', 'OCROS', 'PAUCAR DEL SARA SARA', 'AYABACA',
       'ANTABAMBA', 'HUAYTARA', 'LAMAS', 'CONTUMAZA',
       'MARISCAL RAMON CASTILLA', 'SAN PABLO', 'CONDESUYOS', 'PARURO',
       'PAUCARTAMBO', 'CONDORCANQUI', 'ITALIA', 'ASCOPE', 'COLOMBIA'],
      dtype=object)
```

[ ]:
```
# 2.3 Cambiamos el nombre de las columnas para evitar que haya columnas␣
 ↪repetidas
## Debido a la gran cantidad de distritos, utilizamos una función para␣
 ↪reemplazar tildes y caracteres especiales

def eliminar_tildes(texto):
    if isinstance(texto, str):
        return ''.join(c for c in unicodedata.normalize('NFD', texto) if␣
 ↪unicodedata.category(c) != 'Mn')
    else:
        return texto

## Reemplazar los nombres de distritos utilizando la función
df_1['COLEGIO_DIST'] = df_1['COLEGIO_DIST'].apply(eliminar_tildes)

## Verificamos los valores únicos de la columna de Distritos
df_1['COLEGIO_DIST'].unique()
```

[ ]:
```
array(['SURQUILLO', 'HUARAZ', 'SAN JUAN DE LURIGANCHO', 'SANTA ANITA',
       'SAN MARTIN DE PORRES', 'VENEZUELA', 'VILLA EL SALVADOR',
       'LOS OLIVOS', 'AYACUCHO', 'SAN MIGUEL', 'CHINCHA ALTA',
       'CARABAYLLO', 'CELENDIN', 'SULLANA', 'LINCE', 'CHACHAPOYAS',
       'COMAS', 'SAN JUAN DE MIRAFLORES', 'LIMA', 'POMABAMBA',
       'CORONEL GREGORIO ALBARRACIN LANCHIPA', 'ATE', 'HUANUCO', 'TARMA',
       'CALLAO', 'CALLERIA', 'HUACHO', 'HUAMACHUCO', 'BARRANCO', 'SAYAN',
       'INDEPENDENCIA', 'CHORRILLOS', 'HUANCARAMA', 'JAUJA',
       'LA VICTORIA', 'PUENTE PIEDRA', 'MIRAFLORES', 'ANDAHUAYLAS',
       'RIMAC', 'TOCACHE', 'HUARI', nan, 'CHANCAY', 'TUMAN', 'PISCO',
       'VILLA MARIA DEL TRIUNFO', 'CAJAMARCA', 'CIENEGUILLA', 'ANCON',
       'QUEROCOTILLO', 'HUANCAYO', 'TABACONAS', 'BELLAVISTA', 'CHIRINOS',
       'EL TAMBO', 'QUICHUAS', 'EL AGUSTINO', 'HUARAL', 'LURIGANCHO',
       'LA MOLINA', 'BARRANCA', 'SANTIAGO DE SURCO', 'MAGDALENA DEL MAR',
       'JESUS MARIA', 'CHACLACAYO', 'SAN ISIDRO', 'QUILMANA',
       'VENTANILLA', 'PARAMONGA', 'SATIPO', 'NASCA',
       'LA BANDA DE SHILCAYO', 'BRENA', 'TUCUME', 'CHICLA', 'OLMOS',
       'WANCHAQ', 'AURAHUA', 'AREQUIPA', 'TACNA', 'LAMUD', 'CHOTA',
       'JULIACA', 'HUANTA', 'GROCIO PRADO', 'ILO', 'JAEN', 'PUNO',
       'SANTA ROSA DE QUIVES', 'CARHUAZ', 'CHANCHAMAYO', 'TRUJILLO',
       'PACHACAMAC', 'CATAC', 'TINYAHUARCO', 'CHICLAYO', 'LAMBRAS',
       'LAMBAYEQUE', 'HUANCAVELICA', 'CUTERVO', 'ASCENSION',
       'SAN ANTONIO', 'CHIMBOTE', 'SAN VICENTE DE CANETE', 'BAMBAMARCA',
```

'ICA', 'LURIN', 'MARCONA', 'MORALES', 'POLVORA', 'VILLA RICA',
'PARINAS', 'NUEVO IMPERIAL', 'ABANCAY', 'RAHUAPAMPA', 'LA PERLA',
'TAMBOPATA', 'MORO', 'TINTAY PUNCU', 'JAZAN', 'IBERIA',
'POMAHUACA', 'PUEBLO LIBRE', 'ARAPA', 'ASILLO', 'PUEBLO NUEVO',
'MOYOBAMBA', 'CATAHUASI', 'DANIEL HERNANDEZ', 'OYON', 'CHUPACA',
'NAVAN', 'PICHANAQUI', 'PUNTA HERMOSA', 'PANGOA', 'CONCEPCION',
'IQUITOS', 'SANTA MARIA', 'IRAZOLA', 'HUEPETUHE', 'RUPA-RUPA',
'TAPO', 'CARHUAMAYO', 'AYAVIRI', 'CURASCO', 'QUIVILLA',
'MARISCAL CASTILLA', 'SIMON BOLIVAR', 'MARA',
'SANTIAGO DE LUCANAMARCA', 'HUAYUCACHI', 'TARAPOTO', 'PAUCARA',
'SAN RAMON', 'SAN LUIS', 'VISTA ALEGRE', 'MANGAS', 'MOQUEGUA',
'CUSCO', 'PUNCHANA', 'HUABAL', 'SANCOS', 'MACUSANI', 'PARCONA',
'LA UNION', 'CHONGOS BAJO', 'YANACANCHA', 'CHUCUITO', 'ASUNCION',
'MONSEFU', 'ILAVE', 'SAN PEDRO DE CHUNAN', 'SANTA EULALIA',
'CARAYBAMBA', 'SANTA ROSA DE SACCO', 'YAUYOS', 'LA FLORIDA',
'SAN ANDRES', 'IMPERIAL', 'ARGENTINA', 'VIRU', 'PADRE ABAD',
'TAMBO', 'HUAMBOS', 'COTARUSE', 'PUNTA NEGRA', 'PAUCARTAMBO',
'SAN BORJA', 'ZARUMILLA', 'CARMEN DE LA LEGUA REYNOSO',
'SAN AGUSTIN', 'ACOLLA', 'SICUANI', 'CHAUPIMARCA', 'MARIATANA',
'CORACORA', 'MALA', 'CHALHUANCA', 'TARACO', 'PICHARI', 'JESUS',
'CARAVELI', 'CARAMPOMA', 'MAZAMARI', 'SAN NICOLAS', 'ESPINAR',
'SAN JERONIMO DE TUNAN', 'HUARIBAMBA', 'LUYA', 'CHACAS', 'LLATA',
'ANDAYMARCA', 'CHETO', 'QUERECOTILLO', 'CARAZ', 'PULLO', 'SUCRE',
'SAN BARTOLO', 'PAUCARCOLLA', 'PACHAS', 'ACOBAMBA', 'CHONTABAMBA',
'NUEVO CHIMBOTE', 'MORROPON', 'SAN MIGUEL DE ACO', 'CHILCA',
'ISLAY', 'HUARIACA', 'SANTA CRUZ', 'SAN SEBASTIAN', 'HUANDO',
'PUQUIO', 'CAJATAMBO', 'TICRAPO', 'CIRCA', 'PALCA', 'CALLAYUC',
'CHIQUIAN', 'PAZOS', 'FERRENAFE', 'SAN MARCOS', 'MI PERU', 'CUMBA',
'MOLLENDO', 'SAN JERONIMO', 'TUMBES', 'CHAVIN DE PARIARCA',
'SAN JUAN', 'CHUPA', 'SAN RAFAEL', 'LLAPA', 'HUMAY', 'CURIMANA',
'CANTA', 'REQUE', 'SAN CRISTOBAL', 'MAJES', 'CUYOCUYO',
'HUAROCONDO', 'HERMILIO VALDIZAN', 'NANCHOC', 'PIURA', 'NIEPOS',
'SAN JUAN BAUTISTA', 'COPORAQUE', 'JOSE LEONARDO ORTIZ', 'TIABAYA',
'PAUCARPATA', 'BAGUA', 'MOCHE', 'AZANGARO',
'JOSE LUIS BUSTAMANTE Y RIVERO', 'POZUZO', 'VEINTISEIS DE OCTUBRE',
'OXAPAMPA', 'ALTO SAPOSOA', 'JOSE CRESPO Y CASTILLO', 'CHACCHO',
'LUNAHUANA', 'ACORIA', 'UMARI', 'RIOJA', 'LA TINGUINA', 'ACOCHACA',
'CARMEN ALTO', 'TUMAY HUARACA', 'RICARDO PALMA', 'PIMENTEL',
'MATUCANA', 'SAPALLANGA', 'HUALLANCA', 'ECUADOR', 'HUASAHUASI',
'SAN DAMIAN', 'NAUTA', 'SANTA ROSA', 'PAUSA', 'CHIPAO', 'PAMPAS',
'LLAMA', 'CHOLON', 'SANTIAGO', 'HUANDOVAL', 'AWAJUN', 'YUNGAY',
'YAULI', 'PAMPAMARCA', 'OCROS', 'PISCOBAMBA', 'GUADALUPE',
'YANAHUANCA', 'CASTILLA', 'PUCUSANA', 'HUACANA', 'HUAURA', 'NUNOA',
'LA PECA', 'YANAHUARA', 'UCHIZA', 'CAMPORREDONDO', 'MIRGAS',
'COLCABAMBA', 'CHINCHA BAJA', 'HUANCAN', 'LA OROYA', 'SEPAHUA',
'SAN FRANCISCO DE ASIS DE YARUSYACAN', 'LAGUNAS', 'SAN IGNACIO',
'AMARILIS', 'PERENE', 'BOLIVAR', 'CALGARY', 'CASMA', 'HUARMEY',

'URUBAMBA', 'VARGAS GUERRA', 'QUICHES', 'HUAYLLACAYAN',
'SACSAMARCA', 'MAGDALENA', 'UCUNCHA', 'YUNGUYO', 'PILCOMAYO',
'SAN PEDRO DE LLOC', 'CHAVIN DE HUANTAR', 'CORTEGANA', 'SUNAMPE',
'NUEVO PROGRESO', 'MANTA', 'ANCO_HUALLO', 'BAGUA GRANDE', 'SOLOCO',
'SIVIA', 'ALTO LARAN', 'CHIVAY', 'ALTO SELVA ALEGRE', 'ANTA',
'CHALLHUAHUACHO', 'ATAVILLOS ALTO', 'TONGOD', 'ANCO',
'TRES DE DICIEMBRE', 'CHEPEN', 'MUQUIYAUYO', 'JESUS NAZARENO',
'LIRCAY', 'PACANGA', 'TALAVERA', 'SAN JUAN DEL ORO', 'HUANCAPON',
'CHANGUILLO', 'LARAMATE', 'SECHURA', 'CACRA', 'LLUMPA', 'YANAOCA',
'COCHARCAS', 'COJATA', 'VICHAYAL', 'PACHAMARCA', 'COPA', 'JAYANCA',
'LOS BANOS DEL INCA', 'SIHUAS', 'CHALAMARCA',
'SANTA MARIA DEL VALLE', 'YARINACOCHA', 'CHURUBAMBA',
'ATAVILLOS BAJO', 'OCOBAMBA', 'CHACAPALPA', 'CAYMA', 'SOCABAYA',
'SACHACA', 'INGENIO', 'SANTO TORIBIO', 'LLAMELLIN', 'ACO',
'LA PAMPA', 'EL MANTARO', 'KIMBIRI', 'COCHABAMBA', 'PINRA',
'CERRO AZUL', 'MANCOS', 'PACORA', 'PAROBAMBA', 'SICAYA',
'HUACAYBAMBA', 'SANTA ROSA DE OCOPA', 'LAS LOMAS', 'SUPE PUERTO',
'COSPAN', 'PARACAS', 'OLLANTAYTAMBO', 'SAN PEDRO DE CORIS',
'ALTO DE LA ALIANZA', 'LEIMEBAMBA', 'SUPE', 'URCOS', 'JUNIN',
'JUANJUI', 'CAMANA', 'MOYA', 'SAUSA', 'LAMPA',
'SAN JUAN DE RONTOY', 'LA ARENA', 'ORURILLO', 'SANTA ANA',
'PAUCARBAMBA', 'PACOBAMBA', 'ANDARAPA', 'CHINGAS', 'TINTA',
'AUCARA', 'SAN MARCOS DE ROCCHAC', 'CHINCHEROS',
'SAN SALVADOR DE QUIJE', 'SAN PEDRO DE CHANA', 'SORITOR',
'PACHACUTEC', 'LA MORADA', 'ACRAQUIA', 'SALCAHUASI', 'ILABAYA',
'TUNAN MARCA', 'TAYABAMBA', 'CHAGLLA', 'LAS PIEDRAS', 'PALCAMAYO',
'COTABAMBAS', 'TUPAC AMARU INCA', 'LEONCIO PRADO', 'HUANTAN',
'SAN MATEO', 'ANTONIO RAYMONDI', 'HUAROCHIRI', 'VICE',
'VILCAS HUAMAN', 'CATACAOS', 'LA COIPA', 'SALAS', 'MOCHUMI',
'SHUNTE', 'VITIS', 'MONZON', 'HUALMAY', 'HUARANGO',
'NUEVA CAJAMARCA', 'MARCO', 'CHALA', 'CATACHE',
'SANTA CRUZ DE COCACHACRA', 'POMACANCHA', 'JUSTO APU SAHUARAURA',
'PEDRO VILCA APAZA', 'ORCOTUNA', 'PALPA', 'RIO TAMBO',
'SAN PEDRO DE LARCAY', 'LOS MOROCHUCOS', 'ECHARATE',
'LONYA GRANDE', 'LOS CHANKAS', 'ANDABAMBA', 'PAMPA HERMOSA',
'PUCAYACU', 'PALLASCA', 'VICTOR LARCO HERRERA', 'POMACOCHA',
'PIMPINGOS', 'VEGUETA', 'ORCOPAMPA', 'SOCOS', 'JANJAILLO',
'EL ALTO', 'CAJABAMBA', 'ELIAS SOPLIN VARGAS', 'HUAMBO',
'YURIMAGUAS', 'MARCABAL', 'SAN LUIS DE LUCMA', 'YANAS', 'SANTA',
'INAMBARI', 'QUEROBAMBA', 'CERRO COLORADO', 'ANANEA',
'HUANCABAMBA', 'JANGAS', 'LONGAR', 'LLOCHEGUA', 'SINGA',
'HUANCA-HUANCA', 'GORGOR', 'PALLANCHACRA', 'ANDAHUAYLILLAS',
'PISAC', 'COASA', 'MOLINOS', 'SANTA ANA DE HUAYCAHUACHO',
'CHONGOYAPE', 'CHIGUIRIP', 'CORRALES', 'ZUNIGA', 'EL INGENIO',
'CHUPAMARCA', 'CACHACHI', 'SAN CLEMENTE', 'PAITA', 'INCAHUASI',
'APARICIO POMARES', 'SANTIAGO DE CHUCO', 'TANTARA', 'CARACAS',
'SAN JOSE DEL ALTO', 'ANDRES AVELINO CACERES DORREGARAY',

'PACARAN', 'YURACYACU', 'SANTA ANA DE TUSI', 'CHUQUIBAMBILLA',
'PUERTO INCA', 'CONSTITUCION', 'PACLLON', 'AUCALLAMA', 'PACUCHA',
'COCABAMBA', 'ATICO', 'NAMORA', 'CALCA', 'CHALACO', 'POMACANCHI',
'HUAMANQUIQUIA', 'OTUZCO', 'MASMA', 'LOS AQUIJES', 'ANRA',
'PARCOY', 'CALETA DE CARQUIN', 'ELEAZAR GUZMAN BARRON',
'RANRAHIRCA', 'CHINCHAO', 'SAN SALVADOR', 'ASIA', 'PEDRO GALVEZ',
'ACOMAYO', 'SAN MIGUEL DE EL FAIQUE', 'HUANCARAY', 'PITIPO',
'JOSE MARIA ARGUEDAS', 'COCACHACRA', 'JOSE GALVEZ', 'URANMARCA',
'DESAGUADERO', 'SAN JOSE', 'SURCUBAMBA', 'PUCYURA', 'PATAPO',
'TINICACHI', 'LOS ORGANOS', 'MACATE', 'SAN JUAN DE TANTARANCHE',
'CABANILLAS', 'MALVAS', 'YAUTAN', 'SAMUEL PASTOR', 'SANTO TOMAS',
'EL CARMEN', 'ACZO', 'AHUAYCHA', 'PATIVILCA', 'CUSIPATA', 'YAUCA',
'ZEPITA', 'TURPAY', 'MICAELA BASTIDAS', 'YAMBRASBAMBA',
'SAN PEDRO DE PILLAO', 'PACOCHA', 'PUCARA', 'JULCAN', 'HUALGAYOC',
'TILALI', 'COTAHUASI', 'JIRCAN', 'SANTA MARIA DE CHICMO',
'LUYANDO', 'PAUCAS', 'HUACHAC', 'ETEN', 'CALANGO', 'NEPENA',
'HUASTA', 'CODO DEL POZUZO', 'PALCAZU', 'NINABAMBA', 'CURA MORI',
'PACHANGARA', 'UMACHIRI', 'RAYMONDI', 'SAN MIGUEL DE CAURI',
'PACCHA', 'HUAMANGUILLA', 'MORROPE', 'AQUIA', 'HUANZA',
'CONTAMANA', 'SANTILLANA', 'SAN JUAN DE ISCOS', 'CALAMARCA',
'SAN HILARION', 'NAMBALLE', 'NAGUANAGUA', 'MARIANO DAMASO BERAUN',
'TINTAY', 'HUANUHUANU', 'HUANCANE', 'CHILE', 'LA JOYA', 'HUANCAPI',
'CONGAS', 'CHAPARRA', 'HUACHON', 'CURAHUASI', 'TICAPAMPA',
'SAN LUIS DE SHUARO', 'HUATA', 'ANTIOQUIA', 'MOLINOPAMPA',
'TAMBO GRANDE', 'HUACHIS', 'LUCANAS', 'LIMABAMBA', 'JUMBILLA',
'COISHCO', 'MUSGA', 'TARICA', 'CURGOS', 'CORONGO', 'BELEN',
'CHECRAS', 'CORONEL CASTANEDA', 'PAICO', 'CHURCAMPA', 'CABANA',
'CASHAPAMPA', 'VILCABAMBA', 'MOLINO', 'HUACRACHUCO', 'JULI',
'CANCHABAMBA', 'CANGALLO', 'HUACCANA', 'PITUMARCA', 'PACASMAYO',
'MANUEL ANTONIO MESONES MURO', 'HUALHUAS', 'PAIMAS', 'SALLIQUE',
'LA JALCA', 'LA PUNTA', 'HUAYLLAY', 'HUACACHI', 'LUCRE',
'PARDO MIGUEL', 'SILLAPATA', 'PILLCO MARCA', 'MOTUPE',
'SANTO DOMINGO DE LOS OLLEROS', 'GUADALUPITO', 'QUILLO', 'LAJAS',
'OROPESA', 'RANRACANCHA', 'LA ESPERANZA', 'HUAYTARA', 'AYAPATA',
'TICLACAYAN', 'QUISQUI (KICHKI)', 'CHULUCANAS', 'SURCO',
'UCHUMARCA', 'SAN MIGUEL DE ACOS', 'ALONSO DE ALVARADO', 'AMBO',
'CANCHAQUE', 'CONTUMAZA', 'SANTIAGO DE TUCUMA', 'CULLHUAS',
'ANTABAMBA', 'SAN JUAN DE LA VIRGEN', 'JUAN ESPINOZA MEDRANO',
'COLQUIOC', 'VILAVILA', 'TRITA', 'CHARACATO',
'DANIEL ALOMIA ROBLES', 'UCO', 'PEBAS', 'PULAN', 'SAN PABLO',
'YANAQUIHUA', 'HUAYANA', 'CCAPI', 'OLLACHEA', 'QUIQUIJANA',
'MACARI', 'TABALOSOS', 'HUAMALI', 'MARIANO MELGAR', 'FLORIDA',
'PUCALA', 'COYA', 'LA CRUZ', 'SAUCE', 'PAMPAS DE HOSPITAL',
'YUCAY', 'COATA', 'ACARI', 'LA LIBERTAD DE PALLAN', 'EL ORO',
'CHALLABAMBA', 'NIEVA', 'JOSE DOMINGO CHOQUEHUANCA', 'PANAO',
'POMATA', 'ITALIA', 'CHILLIA', 'CONGALLA', 'RIO NEGRO',
'SAUCEPAMPA', 'PUTINZA', 'CASA GRANDE', 'SALAVERRY', 'MACHUPICCHU',

```
       'COLOMBIA', 'LURICOCHA', 'CHUPURO', 'SANTO DOMINGO DE ACOBAMBA',
       'RIPAN', 'CUPI', 'CONCHUCOS', 'SAN JOSE DE LOS MOLINOS',
       'ARAMANGO', 'CHILETE', 'COYLLURQUI', 'SAN ANTONIO DE CUSICANCHA',
       'CAPILLAS', 'COPANI', 'QUICACHA', 'OMIA', 'CAYNARACHI', 'ALLAUCA'],
      dtype=object)
```

[ ]:
```
# 4. Crear una nueva columna para las modalidades resumidas en solo 13␣
 ↪categorías, para una mejor visualización en el Dashboard
## Pasamos de 45 modalidades a 13, para poder graficarlo mejor en el Dashboard␣
 ↪de Power BI

print(f'Tenemos {len(df_1.MODALIDAD.unique())} modalidades en la columna␣
 ↪original')

modalidades = {
    'ORDINARIO': 'Ingreso Ordinario',

    'EXTRAORDINARIO1 - DEPORTISTAS CALIFICADOS DE ALTO NIVEL( Iniciar␣
 ↪estudios)': 'Ingreso por Deportistas Calificados',
    'EXTRAORDINARIO 1 - DEPORTISTAS CALIFICADOS DE ALTO NIVEL( Iniciar␣
 ↪estudios)': 'Ingreso por Deportistas Calificados',
    'EXTRAORDINARIO - DEPORTISTA CALIFICADO DE ALTO NIVEL( Iniciar estudios)':␣
 ↪'Ingreso por Deportistas Calificados',

    'EXTRAORDINARIO1 - CONVENIO ANDRES BELLO (iniciar estudios)': 'Ingreso por␣
 ↪Convenios',
    'EXTRAORDINARIO 1 - CONVENIO ANDRÉS BELLO (iniciar estudios)': 'Ingreso por␣
 ↪Convenios',
    'EXTRAORDINARIO1 - CONVENIO ANDRES BELLO (continuar estudios)': 'Ingreso␣
 ↪por Convenios',
    'EXTRAORDINARIO 1 - CONVENIO ANDRÉS BELLO (continuar estudios)': 'Ingreso␣
 ↪por Convenios',
    'EXTRAORDINARIO 1 - CONVENIO DIPLOMATICO': 'Ingreso por Convenios',

    'EXTRAORDINARIO - DOS PRIMEROS ALUMNOS': 'Ingreso por Excelencia Académica',
    'EXTRAORDINARIO 1 - DOS PRIMEROS ALUMNOS': 'Ingreso por Excelencia␣
 ↪Académica',
    'EXTRAORDINARIO1 - DOS PRIMEROS ALUMNOS': 'Ingreso por Excelencia␣
 ↪Académica',

    'EXTRAORDINARIO - VÍCTIMA DEL TERRORISMO (iniciar estudios)': 'Ingreso por␣
 ↪Víctimas de Terrorismo',
    'EXTRAORDINARIO 1 - VÍCTIMA DEL TERRORISMO (iniciar estudios)': 'Ingreso␣
 ↪por Víctimas de Terrorismo',
    'EXTRAORDINARIO1 - VICTIMAS DEL TERRORISMO (iniciar estudios)': 'Ingreso␣
 ↪por Víctimas de Terrorismo',
```

```
  'EXTRAORDINARIO 1 - VICTIMAS DEL TERRORISMO (continuar estudios)': 'Ingreso␣
↪por Víctimas de Terrorismo',
  'EXTRAORDINARIO - VICTIMA DEL TERRORISMO (continuar estudios)': 'Ingreso␣
↪por Víctimas de Terrorismo',
  'EXTRAORDINARIO1 - VICTIMAS DEL TERRORISMO (continuar estudios)': 'Ingreso␣
↪por Víctimas de Terrorismo',

  'EXTRAORDINARIO - TRASLADO EXTERNO': 'Ingreso por Traslado Externo',
  'EXTRAORDINARIO1 - TRASLADO EXTERNO': 'Ingreso por Traslado Externo',
  'EXTRAORDINARIO 1 - TRASLADO EXTERNO': 'Ingreso por Traslado Externo',
  'EXTRAORDINARIO 1 - TRASLADO EXTERNO PARA ESTUDIANTES PROVENIENTES DE␣
↪UNIVERSIDADES NO LICENCIADAS': 'Ingreso por Traslado Externo',
  'EXTRAORDINARIO1 - TRASLADO EXTERNO PARA ESTUDIANTES PROVENIENTES DE␣
↪UNIVERSIDADES NO LICENCIADAS': 'Ingreso por Traslado Externo',
  'EXTRAORDINARIO - TRASLADO EXTERNO PARA ESTUDIANTES PROVENIENTES DE␣
↪UNIVERSIDADES NO LICENCIADAS': 'Ingreso por Traslado Externo',

  'EXTRAORDINARIO - TITULADOS O GRADUADOS UNI': 'Ingreso por Titulados o␣
↪Graduados',
  'EXTRAORDINARIO1 - TITULADOS O GRADUADOS': 'Ingreso por Titulados o␣
↪Graduados',
  'EXTRAORDINARIO1 - TITULADOS O GRADUADOS UNI': 'Ingreso por Titulados o␣
↪Graduados',
  'EXTRAORDINARIO 1 - TITULADOS O GRADUADOS EN OTRA UNIVERSIDAD': 'Ingreso␣
↪por Titulados o Graduados',
  'EXTRAORDINARIO - TITULADOS O GRADUADOS EN OTRA UNIVERSIDAD': 'Ingreso por␣
↪Titulados o Graduados',
  'EXTRAORDINARIO 1 - TITULADO O GRADUADO UNI': 'Ingreso por Titulados o␣
↪Graduados',

  'EXTRAORDINARIO - DIPLOMADO CON BACHILLERATO INTERNACIONAL': 'Ingreso por␣
↪Bachillerato Internacional',
  'EXTRAORDINARIO 1 - DIPLOMADOS CON BACHILLERATO': 'Ingreso por Bachillerato␣
↪Internacional',
  'EXTRAORDINARIO1 - DIPLOMADOS CON BACHILLERATO INTERNACIONAL': 'Ingreso por␣
↪Bachillerato Internacional',
  'EXTRAORDINARIO 1 - DIPLOMADOS CON BACHILLERATO': 'Ingreso por Bachillerato␣
↪Internacional',

  'EXTRAORDINARIO - PERSONA CON DISCAPACIDAD (iniciar estudios)': 'Ingreso␣
↪por Discapacidad',
  'EXTRAORDINARIO 1 - PERSONAS CON DISCAPACIDAD (iniciar estudios)': 'Ingreso␣
↪por Discapacidad',
  'EXTRAORDINARIO1 - PERSONAS CON DISCAPACIDAD (iniciar estudios)': 'Ingreso␣
↪por Discapacidad',
```

```
    'EXTRAORDINARIO INGRESO DIRECTO CEPRE-UNI': 'Ingreso Directo CEPRE',
    'EXTRAORDINARIO1 - INGRESO DIRECTO CEPRE': 'Ingreso Directo CEPRE',
    'EXTRAORDINARIO INGRESO DIRECTO CEPRE-UNI O CEPRE-UNI INTENSIVO': 'Ingreso␣
 ↪Directo CEPRE',
    'EXTRAORDINARIO 2 - INGRESO DIRECTO CEPRE': 'Ingreso Directo CEPRE',
    'EXTRAORDINARIO2 - INGRESO DIRECTO CEPRE': 'Ingreso Directo CEPRE',

    'TALENTO BECA 18': 'Ingreso por Beca',
    'EXTRAORDINARIO2 - TALENTO BECA 18': 'Ingreso por Beca',

    'INGRESO ESCOLAR NACIONAL': 'Ingreso Escolar Nacional',

    'INTERESADO': 'Otros'
}

df_1['MODALIDAD_RESUM'] = df_1['MODALIDAD'].map(modalidades)

df_1['MODALIDAD_RESUM'] = df_1['MODALIDAD_RESUM'].fillna('Otros')

## Creamos MODALIDADES_RESUM donde se resumen las modalidades a solo 13
print(f'Tenemos {len(df_1.MODALIDAD_RESUM.unique())} modalidades en la columna␣
 ↪resumida. Estas son: {df_1.MODALIDAD_RESUM.unique()}')
```

```
Tenemos 45 modalidades en la columna original
Tenemos 13 modalidades en la columna resumida. Estas son: ['Ingreso por
Deportistas Calificados' 'Ingreso Ordinario'
 'Ingreso Directo CEPRE' 'Ingreso por Convenios'
 'Ingreso por Excelencia Académica' 'Ingreso por Víctimas de Terrorismo'
 'Ingreso por Traslado Externo' 'Ingreso por Titulados o Graduados'
 'Ingreso Escolar Nacional' 'Ingreso por Beca' 'Otros'
 'Ingreso por Bachillerato Internacional' 'Ingreso por Discapacidad']
```

[ ]:
```
# 5. Creamos una nueva variable categórica según la calificación final de los␣
 ↪estudiantes
## Menos de 11: DESAPROBADO
## Entre 11 y 15: APROBADO
## Más de 15: SOBRESALIENTE
labels = ["DESAPROBADO", "APROBADO", "SOBRESALIENTE"]
bins = [0, 11, 15, df_1["CALIF_FINAL"].max()]

## Creamos CALIF_CATEGORICO
df_1["CALIF_CATEGORICO"] = pd.cut(df_1["CALIF_FINAL"],
                                  bins=bins,
                                  labels=labels)
```

[ ]:
```
# 6. Analizamos el nuevo dataframe y lo exportamos a Excel
df_1.head()
```

[ ]:

```
                                                IDHASH  \
0  E4287C2FE19F63C5E6641955147E36684A5A2FF8064676…
1  2DC37F0B9727B6591EC72D7A942647797A200F45D47C9E…
2  3B8677B90781D7BB8F2F967C05FA2DBBE153BBB682DF05…
3  FA366704D9E9F6FB5E5F55C1FB0CEEE973C626A5616F55…
4  E31CF8F30F3AE60B3D8A14F6E1020E9AD26EE975F0823B…

                     COLEGIO COLEGIO_DEPA COLEGIO_PROV  \
0       LA DIVINA PROVIDENCIA         LIMA         LIMA
1           86019 LA LIBERTAD       ANCASH       HUARAZ
2  0113 DANIEL ALOMIAS ROBLES         LIMA         LIMA
3     SEBASTIAN SALAZAR BONDY         LIMA         LIMA
4           TRILCE LOS OLIVOS         LIMA         LIMA

             COLEGIO_DIST COLEGIO_PAIS  COLEGIO_ANIO_EGRESO  \
0               SURQUILLO         PERÚ                 2020
1                  HUARAZ         PERÚ                 2017
2  SAN JUAN DE LURIGANCHO         PERÚ                 2016
3             SANTA ANITA         PERÚ                 2018
4    SAN MARTIN DE PORRES         PERÚ                 2020

                       ESPECIALIDAD  ANIO_POSTULA  CICLO_POSTULA  …  \
0            INGENIERÍA DE SISTEMAS          2021              1  …
1  INGENIERÍA DE TELECOMUNICACIONES          2021              1  …
2               INGENIERÍA MECÁNICA          2021              1  …
3            INGENIERÍA ELECTRÓNICA          2021              1  …
4                      ARQUITECTURA          2021              1  …

  NACIMIENTO_PAIS NACIMIENTO_DEPA NACIMIENTO_PROV    NACIMIENTO_DIST  \
0            PERÚ            LIMA            LIMA  VILLA EL SALVADOR
1            PERÚ          ÁNCASH          HUARAZ             HUARAZ
2            PERÚ            LIMA            LIMA        JESÚS MARÍA
3            PERÚ            LIMA            LIMA               LIMA
4            PERÚ            LIMA            LIMA              COMAS

        SEXO CALIF_FINAL INGRESO  \
0  MASCULINO        3.11      NO
1   FEMENINO        3.58      NO
2  MASCULINO        8.04      NO
3  MASCULINO       10.82      NO
4   FEMENINO        6.21      NO

                                          MODALIDAD  \
0  EXTRAORDINARIO1 - DEPORTISTAS CALIFICADOS DE A…
1                                          ORDINARIO
2                                          ORDINARIO
3            EXTRAORDINARIO2 - INGRESO DIRECTO CEPRE
```

```
4                                          ORDINARIO

                      MODALIDAD_RESUM  CALIF_CATEGORICO
0  Ingreso por Deportistas Calificados       DESAPROBADO
1                    Ingreso Ordinario       DESAPROBADO
2                    Ingreso Ordinario       DESAPROBADO
3                Ingreso Directo CEPRE       DESAPROBADO
4                    Ingreso Ordinario       DESAPROBADO

[5 rows x 24 columns]
```

[ ]:
```
## Lo exportamos a Excel
df_1.to_excel('data/data_preprocessed.xlsx', index=False)
```