lab1_compile

October 15, 2020

1 Laboratory 1

Authors:
- Roberto Pérez
- Arturo Bringas
- Edgar Bazo
- Mariana Lugo

2 Imports

Python libraries

[1]:
```
import sys
import pandas as pd
import numpy as np
import re
import unicodedata
from pandas_profiling import ProfileReport
import plotly.graph_objects as go
import plotly.express as px
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.figure_factory as ff
import probscale
from scipy import stats
```

Acillary modules

[2]:
```
%reload_ext autoreload
%autoreload 2

from utils.lab1_funcs import *
```

3 Loading data

Data downloaded from -> https://datos.cdmx.gob.mx/explore/dataset/consumo-agua/export/

[3]:
```
df = pd.read_csv("data/consumo-agua.csv")
```

[4]:
```
df
```

[4]:

```
                          Geo Point  \
0      19.4552601937,-99.1126617526
1      19.4552601937,-99.1126617526
2      19.4557195871,-99.1135822797
3      19.4596467168,-99.1044693641
4      19.4741606185,-99.1467497317
…                               …
71097  19.4485642979,-99.1399395353
71098  19.4493393649,-99.1457191092
71099  19.4483923147,-99.1459300721
71100  19.4475868325,-99.1425094385
71101  19.4474017534,-99.1397251034

                                              Geo Shape  consumo_total_mixto  \
0      {"type": "MultiPolygon", "coordinates": [[[[-9…               159.72
1      {"type": "MultiPolygon", "coordinates": [[[[-9…                 0.00
2      {"type": "MultiPolygon", "coordinates": [[[[-9…                 0.00
3      {"type": "MultiPolygon", "coordinates": [[[[-9…                 0.00
4      {"type": "MultiPolygon", "coordinates": [[[[-9…                56.72
…                                                    …                    …
71097  {"type": "MultiPolygon", "coordinates": [[[[-9…                  NaN
71098  {"type": "MultiPolygon", "coordinates": [[[[-9…                71.30
71099  {"type": "MultiPolygon", "coordinates": [[[[-9…               759.16
71100  {"type": "MultiPolygon", "coordinates": [[[[-9…               402.65
71101  {"type": "MultiPolygon", "coordinates": [[[[-9…                41.20

       anio             nomgeo  consumo_prom_dom  consumo_total_dom  \
0      2019  Gustavo A. Madero         42.566364             468.23
1      2019  Gustavo A. Madero         35.936667             107.81
2      2019  Gustavo A. Madero         24.586000             122.93
3      2019  Gustavo A. Madero          0.000000               0.00
4      2019       Azcapotzalco         67.436250             539.49
…       …                  …                 …                  …
71097  2019        Cuauhtémoc         20.053112            3930.41
71098  2019        Cuauhtémoc         21.126615            9549.24
71099  2019        Cuauhtémoc         27.527778            4707.25
71100  2019        Cuauhtémoc         30.605000             550.89
71101  2019        Cuauhtémoc         22.507710            8552.94

                alcaldia             colonia  consumo_prom_mixto  \
0      GUSTAVO A. MADERO      7 DE NOVIEMBRE           53.240000
1      GUSTAVO A. MADERO      7 DE NOVIEMBRE            0.000000
2      GUSTAVO A. MADERO      7 DE NOVIEMBRE            0.000000
3      GUSTAVO A. MADERO  NUEVA TENOCHTITLAN            0.000000
4           AZCAPOTZALCO            PROHOGAR           56.720000
```

```
…                 …                 …                 …
71097         CUAUHTEMOC            GUERRERO               NaN
71098         CUAUHTEMOC            GUERRERO         35.650001
71099         CUAUHTEMOC            GUERRERO         94.894999
71100         CUAUHTEMOC            GUERRERO        100.662498
71101         CUAUHTEMOC            GUERRERO         13.733333

       consumo_total  consumo_prom  consumo_prom_no_dom  bimestre  \
0             631.00     42.066667             3.050000         3
1             115.13     28.782500             7.320000         3
2             197.96     32.993333            75.030000         3
3             253.53     84.510000            84.510000         3
4             839.35     76.304545           121.570000         3
…               …           …                   …       …
71097        4286.28     19.307568            13.687308         1
71098        9796.12     20.976702            13.506923         1
71099        5692.81     29.344381            15.093334         1
71100         963.15     41.876087             9.610000         1
71101        9000.07     21.951366            15.034444         1

       consumo_total_no_dom    gid indice_des
0                      3.05  57250       ALTO
1                      7.32  57253      MEDIO
2                     75.03  57255    POPULAR
3                    253.53  57267       BAJO
4                    243.14  57330       BAJO
…                     …    …        …
71097                355.87    233       BAJO
71098                175.59    238    POPULAR
71099                226.40    239       BAJO
71100                  9.61    244       BAJO
71101                405.93    247       BAJO

[71102 rows x 17 columns]
```

4 Exploratory Data Analysis (EDA)

4.1 Data profiling

4.2 ¿Cuántas variables tenemos?

[5]:
```
count_vars(df)
```

```
Número de variables en los datos --> 17
```

4.3 ¿Cuántas observaciones tenemos?

[6]:
```
count_obs(df)
```

```
Número de observaciones en los datos --> 71102
```

4.4 ¿Cuántas observaciones únicas tenemos por variable?

[7]:
```
count_unique_obs(df)
```

[7]:
```
Geo Point               22930
Geo Shape               22922
consumo_total_mixto     24339
anio                        1
nomgeo                     17
consumo_prom_dom        52060
consumo_total_dom       47051
alcaldia                   16
colonia                  1340
consumo_prom_mixto      31911
consumo_total           56015
consumo_prom            62214
consumo_prom_no_dom     37440
bimestre                    3
consumo_total_no_dom    27336
gid                     71102
indice_des                  4
dtype: int64
```

4.5 ¿Cuántas variables numéricas tenemos?

Tenemos 8 variables numéricas

[8]:
```
vars_num = ["consumo_total",
    "consumo_total_dom",
    "consumo_total_no_dom",
    "consumo_total_mixto",
    "consumo_prom",
    "consumo_prom_dom",
    "consumo_prom_no_dom",
    "consumo_prom_mixto"]
```

[9]:
```
count_type_vars(vars_num, "numerica")
```

```
Número de variables de tipo numerica --> 8
```

```
            Variable(s)
1         consumo_total
2     consumo_total_dom
```

```
3  consumo_total_no_dom
4   consumo_total_mixto
5           consumo_prom
6      consumo_prom_dom
7   consumo_prom_no_dom
8    consumo_prom_mixto
```

```
None
```

4.6 ¿Cuántas variables de fecha tenemos?

- Para efectos de este ejercicio, no hay ninguna variable de tipo fecha, o que consideremos de fecha.

4.7 ¿Cuántas variables categóricas tenemos?

[10]:
```
cat_vars = [
    "anio",
    "nomgeo",
    "alcaldia",
    "colonia",
    "bimestre",
    "indice_des",
]
```

[11]:
```
count_type_vars(cat_vars, "categórica")
```

```
Número de variables de tipo categórica --> 6
```

```
  Variable(s)
1        anio
2      nomgeo
3    alcaldia
4     colonia
5    bimestre
6  indice_des
```

```
None
```

4.8 ¿Cuántas variables de texto tenemos?

- Para efectos de este ejercicio, no hay ninguna variable de tipo texto, o que consideremos de texto. Se podría considerar a la variable `gid` como un identificador de texto.

[12]:
```
vars_text = ["gid"]
```

[13]:
```
count_type_vars(vars_text, "texto")
```

```
Número de variables de tipo texto --> 1
```

```
  Variable(s)
1         gid

None
```

4.9 Generea el profiling de cada variable

4.9.1 Numeric data profiling

- ☒ Tipo de dato: float, integer
- ☒ Número de observaciones
- ☒ Mean
- ☒ Desviación estándar
- ☒ Cuartiles: 25%, 50%, 75%
- ☒ Valor máximo
- ☒ Valor mínimo
- ☒ Número de observaciones únicos
- ☒ Top 5 observaciones repetidas
- ☒ Número de observaciones con valores faltantes
- ☒ ¿Hay redondeos? -> Se observa que los datos de los totales en el consumo (dom, mixto, no_dom) están a dos decimales desde la fuente. Los datos de los promedios del consumo (dom, mixto, no_dom) tienen seis decimales desde la fuente. No podemos asegurar que los datos, tanto de los totales, como de los promedios, están redondeados.

Función para perfil de datos numérico

[14]:
```
## Data profiling compacted in function
data_profiling_numeric(df, vars_num)
```

```
**********
** General description of data **
**********

             consumo_total consumo_total_dom consumo_total_no_dom  \
dtype              float64           float64               float64
count_unique         56015             47051                 27336
missing_v                0              4820                     0
count                71102             66282                 71102
mean               1695.85           1186.26                436.06
std                 3555.7           2771.04               2126.15
min                      0                 0                     0
25%                340.952           161.635                 10.98
50%                896.175           604.185                54.055
75%                 1808.9           1261.45                230.43
max                 119727           95060.7                119727

             consumo_total_mixto consumo_prom consumo_prom_dom  \
dtype                    float64      float64          float64
count_unique               24339        62214            52060
missing_v                   8327            0             4820
```

count	62775	71102	66282
mean	174.36	111.217	29.1324
std	312.664	1069.95	64.5659
min	0	0	0
25%	0	23.0101	18.6905
50%	79.94	31.6938	26.4142
75%	233.32	45.4849	36.2466
max	23404.4	89691.8	7796.41

	consumo_prom_no_dom	consumo_prom_mixto
dtype	float64	float64
count_unique	37440	31911
missing_v	0	8327
count	71102	62775
mean	126.76	50.6362
std	1095.82	130.409
min	0	0
25%	6.27542	0
50%	19.28	33.4517
75%	54.1869	61.2165
max	89691.8	11702.2

None

****
** Top repeated variables **
****

	consumo_total			consumo_total_dom			\
	value	count	part_notnull	value	count	part_notnull	
top_1	0.00	2451	3.45	0.00	9861	14.88	
top_2	3.05	70	0.10	1.22	37	0.06	
top_3	1.22	68	0.10	10.98	21	0.03	
top_4	3.66	42	0.06	25.62	20	0.03	
top_5	6.71	41	0.06	3.66	20	0.03	

	consumo_total_no_dom			consumo_total_mixto		\
	value	count	part_notnull	value	count	
top_1	0.00	8109	11.40	0.0	17715	
top_2	1.22	402	0.57	36.0	74	
top_3	1.83	316	0.44	17.7	61	
top_4	3.05	302	0.42	36.6	59	
top_5	7.93	219	0.31	18.3	54	

	consumo_prom	consumo_prom_dom	\

```
      part_notnull        value count part_notnull            value count
top_1        28.22         0.00  2451         3.45             0.00  9861
top_2         0.12         1.22    62         0.09             1.22    33
top_3         0.10         3.05    55         0.08            14.64    23
top_4         0.09         4.27    43         0.06            10.98    22
top_5         0.09         6.71    39         0.05            15.25    22

                  consumo_prom_no_dom                   consumo_prom_mixto  \
      part_notnull               value count part_notnull              value
top_1        14.88                0.00  8109        11.40               0.00
top_2         0.05                1.22   330         0.46              36.00
top_3         0.03                1.83   290         0.41              29.28
top_4         0.03                3.05   260         0.37              36.60
top_5         0.03                4.27   216         0.30              23.80


       count part_notnull
top_1  17715        28.22
top_2     58         0.09
top_3     57         0.09
top_4     53         0.08
top_5     49         0.08

None
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
```

4.9.2 Categoric data profiling

Profiling: Variables categóricas

- ☒ Número de categorías
- ☒ Valor de las categorías
- ☒ Moda
- ☒ Valores faltantes
- ☒ Número de observaciones con valores faltantes
- ☒ Proporción de observaciones por categoría
- ☒ Top 1, top 2, top 3 (moda 1, moda 2, moda 3)
- ☒ Faltas de ortografía ?

[15]:
```
#data profiling function
data_profiling_categ(df,cat_vars)
```

```
**********
Variable Categorica anio
**********

Info                      anio
Num_Registros            71102
```

```
Num_de_categorias              1
Moda                        2019
Valores_faltantes              0
Top1                [2019, 71102]
Top2                           0
Top3                           0
```

None
Valores de las categorias y sus proporciones

```
           Observaciones proporción
Categoría
2019               71102     100.0%
```

None

****
Variable Categorica nomgeo
****

```
Info                                    nomgeo
Num_Registros                            71102
Num_de_categorias                           17
Moda                                Iztapalapa
Valores_faltantes                            0
Top1                       [Iztapalapa, 10515]
Top2                [Gustavo A. Madero, 10058]
Top3                        [Cuauhtémoc, 7313]
```

None
Valores de las categorias y sus proporciones

Categoría	Observaciones	proporción
Iztapalapa	10515	14.8%
Gustavo A. Madero	10058	14.1%
Cuauhtémoc	7313	10.3%
Benito Juárez	6049	8.5%
Venustiano Carranza	5179	7.3%
Miguel Hidalgo	5110	7.2%
Coyoacán	4947	7.0%
Azcapotzalco	4216	5.9%
Álvaro Obregón	4140	5.8%
Iztacalco	3469	4.9%
Xochimilco	2450	3.4%
Talpan	2140	3.0%
Tláhuac	1955	2.7%
Tlalpan	1064	1.5%
La Magdalena Contreras	955	1.3%

```
Cuajimalpa de Morelos             892       1.3%
Milpa Alta                        650       0.9%
```

None

```
**********
Variable Categorica alcaldia
**********
```

Info	alcaldia
Num_Registros	71102
Num_de_categorias	16
Moda	IZTAPALAPA
Valores_faltantes	0
Top1	[IZTAPALAPA, 10515]
Top2	[GUSTAVO A. MADERO, 10058]
Top3	[CUAUHTEMOC, 7313]

None

Valores de las categorias y sus proporciones

Categoría	Observaciones	proporción
IZTAPALAPA	10515	14.8%
GUSTAVO A. MADERO	10058	14.1%
CUAUHTEMOC	7313	10.3%
BENITO JUAREZ	6049	8.5%
VENUSTIANO CARRANZA	5179	7.3%
MIGUEL HIDALGO	5110	7.2%
COYOACAN	4947	7.0%
AZCAPOTZALCO	4216	5.9%
ALVARO OBREGON	4140	5.8%
IZTACALCO	3469	4.9%
TLALPAN	3204	4.5%
XOCHIMILCO	2450	3.4%
TLAHUAC	1955	2.7%
MAGDALENA CONTRERAS	955	1.3%
CUAJIMALPA	892	1.3%
MILPA ALTA	650	0.9%

None

```
**********
Variable Categorica colonia
**********
```

Info	colonia

```
Num_Registros                          71102
Num_de_categorias                       1340
Moda                                  CENTRO
Valores_faltantes                          0
Top1                          [CENTRO, 1139]
Top2                [AGRICOLA ORIENTAL, 837]
Top3                        [ROMA NORTE, 602]
```

None
Valores de las categorias y sus proporciones

Categoría	Observaciones	proporción
CENTRO	1139	1.6%
AGRICOLA ORIENTAL	837	1.2%
ROMA NORTE	602	0.8%
MOCTEZUMA 2A SECCION	558	0.8%
JARDIN BALBUENA	498	0.7%
…	…	…
HUIZACHITO	2	0.0%
CALZADA JALALPA	2	0.0%
PIRU SECC. I	2	0.0%
U. HAB. NUEVA ROSITA	1	0.0%
SANTISIMA TRINIDAD	1	0.0%

[1340 rows x 2 columns]

None

****
Variable Categorica bimestre
****

```
Info                 bimestre
Num_Registros           71102
Num_de_categorias           3
Moda                        2
Valores_faltantes           0
Top1                [2, 23942]
Top2                [3, 23822]
Top3                [1, 23338]
```

None
Valores de las categorias y sus proporciones

Categoría	Observaciones	proporción
2	23942	33.7%
3	23822	33.5%

```
1                   23338      32.8%

None



*******************************
Variable Categorica indice_des
*******************************

Info                      indice_des
Num_Registros                  71102
Num_de_categorias                  4
Moda                            BAJO
Valores_faltantes                  0
Top1                   [BAJO, 29248]
Top2                [POPULAR, 16539]
Top3                   [ALTO, 15516]

None
Valores de las categorias y sus proporciones

           Observaciones proporción
Categoría
BAJO                29248      41.1%
POPULAR             16539      23.3%
ALTO                15516      21.8%
MEDIO                9799      13.8%

None
```

4.9.3 Additional data profiling

- En este ejercicio particular no se contó con variables del siguiente tipo para hacer el profiling:
  - Imágen
  - Audio

4.10 ¿Qué conocemos ahora de este set de datos por variable?

1. ¿Cuántas alcadías tienes?
   - Hay un total de 16 alcaldías
2. ¿Cuántos `nomgeo` tienes?
   - Hay un total de 17 nomgeo
3. ¿Identificas algún error?
   - Hay una categoría con un error ortográfico: Existen 2140 observaciones en la columna **nomgeo** que dice: Talpan y 1064 observaciones que dicen: Tlalpan

4.11 Transformar el nombre de las columnas a formato estándar: minúsculas, sin espacios en blanco -cambiar por guiónes bajos-, sin signos de puntuación

[16]:
```
df_cleancols = clean_col_names(df)
```

[17]:
```
df_cleancols
```

[17]:

```
                          geo_point  \
0      19.4552601937,-99.1126617526
1      19.4552601937,-99.1126617526
2      19.4557195871,-99.1135822797
3      19.4596467168,-99.1044693641
4      19.4741606185,-99.1467497317
…                               …
71097  19.4485642979,-99.1399395353
71098  19.4493393649,-99.1457191092
71099  19.4483923147,-99.1459300721
71100  19.4475868325,-99.1425094385
71101  19.4474017534,-99.1397251034

                                              geo_shape  consumo_total_mixto  \
0      {"type": "MultiPolygon", "coordinates": [[[[-9…               159.72
1      {"type": "MultiPolygon", "coordinates": [[[[-9…                 0.00
2      {"type": "MultiPolygon", "coordinates": [[[[-9…                 0.00
3      {"type": "MultiPolygon", "coordinates": [[[[-9…                 0.00
4      {"type": "MultiPolygon", "coordinates": [[[[-9…                56.72
…                                                    …                    …
71097  {"type": "MultiPolygon", "coordinates": [[[[-9…                  NaN
71098  {"type": "MultiPolygon", "coordinates": [[[[-9…                71.30
71099  {"type": "MultiPolygon", "coordinates": [[[[-9…               759.16
71100  {"type": "MultiPolygon", "coordinates": [[[[-9…               402.65
71101  {"type": "MultiPolygon", "coordinates": [[[[-9…                41.20

       anio             nomgeo  consumo_prom_dom  consumo_total_dom  \
0      2019  Gustavo A. Madero         42.566364             468.23
1      2019  Gustavo A. Madero         35.936667             107.81
2      2019  Gustavo A. Madero         24.586000             122.93
3      2019  Gustavo A. Madero          0.000000               0.00
4      2019       Azcapotzalco         67.436250             539.49
…     …                    …                 …                  …
71097  2019        Cuauhtémoc         20.053112            3930.41
71098  2019        Cuauhtémoc         21.126615            9549.24
71099  2019        Cuauhtémoc         27.527778            4707.25
71100  2019        Cuauhtémoc         30.605000             550.89
71101  2019        Cuauhtémoc         22.507710            8552.94
```

```
                alcaldia             colonia  consumo_prom_mixto  \
0      GUSTAVO A. MADERO      7 DE NOVIEMBRE           53.240000
1      GUSTAVO A. MADERO      7 DE NOVIEMBRE            0.000000
2      GUSTAVO A. MADERO      7 DE NOVIEMBRE            0.000000
3      GUSTAVO A. MADERO  NUEVA TENOCHTITLAN            0.000000
4           AZCAPOTZALCO            PROHOGAR           56.720000
…                    …                   …                   …
71097         CUAUHTEMOC            GUERRERO                 NaN
71098         CUAUHTEMOC            GUERRERO           35.650001
71099         CUAUHTEMOC            GUERRERO           94.894999
71100         CUAUHTEMOC            GUERRERO          100.662498
71101         CUAUHTEMOC            GUERRERO           13.733333

       consumo_total  consumo_prom  consumo_prom_no_dom  bimestre  \
0             631.00     42.066667             3.050000         3
1             115.13     28.782500             7.320000         3
2             197.96     32.993333            75.030000         3
3             253.53     84.510000            84.510000         3
4             839.35     76.304545           121.570000         3
…               …           …                   …       …
71097        4286.28     19.307568            13.687308         1
71098        9796.12     20.976702            13.506923         1
71099        5692.81     29.344381            15.093334         1
71100         963.15     41.876087             9.610000         1
71101        9000.07     21.951366            15.034444         1

       consumo_total_no_dom    gid indice_des
0                      3.05  57250       ALTO
1                      7.32  57253      MEDIO
2                     75.03  57255    POPULAR
3                    253.53  57267       BAJO
4                    243.14  57330       BAJO
…                     …    …         …
71097                355.87    233       BAJO
71098                175.59    238    POPULAR
71099                226.40    239       BAJO
71100                  9.61    244       BAJO
71101                405.93    247       BAJO

[71102 rows x 17 columns]
```

4.12 Transformación de variables geoespaciales

- Agregar la variable `latitud` y `longitud` generadas a partir de la column `geo_point`.
- Pasar la variable `latitud` y `longitud` a numérica -si no la tomó como numérica-.
- Eliminar la columna `geo_point` -una vez que creaste la variable `latitud` y `longitud`.
- Eliminar la columna `geo_shape`.

- Cambiar a minúsculas las columnas `alcaldía`, `colonia` e `indice_des`.

[18]:
```
df_geotransform = geo_transformation(df_cleancols, "geo_point", "geo_shape")
df_geotransform
```

[18]:

	consumo_total_mixto	anio	nomgeo	consumo_prom_dom	\
0	159.72	2019	Gustavo A. Madero	42.566364	
1	0.00	2019	Gustavo A. Madero	35.936667	
2	0.00	2019	Gustavo A. Madero	24.586000	
3	0.00	2019	Gustavo A. Madero	0.000000	
4	56.72	2019	Azcapotzalco	67.436250	
…	…	…	…	…	
71097	NaN	2019	Cuauhtémoc	20.053112	
71098	71.30	2019	Cuauhtémoc	21.126615	
71099	759.16	2019	Cuauhtémoc	27.527778	
71100	402.65	2019	Cuauhtémoc	30.605000	
71101	41.20	2019	Cuauhtémoc	22.507710	

	consumo_total_dom	alcaldia	colonia	\
0	468.23	GUSTAVO A. MADERO	7 DE NOVIEMBRE	
1	107.81	GUSTAVO A. MADERO	7 DE NOVIEMBRE	
2	122.93	GUSTAVO A. MADERO	7 DE NOVIEMBRE	
3	0.00	GUSTAVO A. MADERO	NUEVA TENOCHTITLAN	
4	539.49	AZCAPOTZALCO	PROHOGAR	
…	…	…	…	
71097	3930.41	CUAUHTEMOC	GUERRERO	
71098	9549.24	CUAUHTEMOC	GUERRERO	
71099	4707.25	CUAUHTEMOC	GUERRERO	
71100	550.89	CUAUHTEMOC	GUERRERO	
71101	8552.94	CUAUHTEMOC	GUERRERO	

	consumo_prom_mixto	consumo_total	consumo_prom	consumo_prom_no_dom	\
0	53.240000	631.00	42.066667	3.050000	
1	0.000000	115.13	28.782500	7.320000	
2	0.000000	197.96	32.993333	75.030000	
3	0.000000	253.53	84.510000	84.510000	
4	56.720000	839.35	76.304545	121.570000	
…	…	…	…	…	
71097	NaN	4286.28	19.307568	13.687308	
71098	35.650001	9796.12	20.976702	13.506923	
71099	94.894999	5692.81	29.344381	15.093334	
71100	100.662498	963.15	41.876087	9.610000	
71101	13.733333	9000.07	21.951366	15.034444	

	bimestre	consumo_total_no_dom	gid	indice_des	latitud	longitud
0	3	3.05	57250	ALTO	19.455260	-99.112662
1	3	7.32	57253	MEDIO	19.455260	-99.112662

```
2             3                 75.03  57255    POPULAR  19.455720 -99.113582
3             3                253.53  57267       BAJO  19.459647 -99.104469
4             3                243.14  57330       BAJO  19.474161 -99.146750
…           …                     …      …          …         …         …
71097         1                355.87    233       BAJO  19.448564 -99.139940
71098         1                175.59    238    POPULAR  19.449339 -99.145719
71099         1                226.40    239       BAJO  19.448392 -99.145930
71100         1                  9.61    244       BAJO  19.447587 -99.142509
71101         1                405.93    247       BAJO  19.447402 -99.139725

[71102 rows x 17 columns]
```

4.13 Geospatial data profiling

[19]:
```
geo_vars = ["latitud", "longitud"]

geo_vars_precision(df_geotransform, geo_vars)
```

```
                 No. of entries - latitud  No. of entries - longitud
No. of decimals
10                                  64044                      64152
9                                    6274                       6316
8                                     712                        580
7                                      69                         39
6                                       3                         12

None
```

- Cambiar a minúsculas las columnas `alcaldía`, `colonia` e `indice_des`.

[20]:
```
vars_lower=["indice_des", "alcaldia", "colonia","nomgeo"]

df_lower_values = convert_lower(df_geotransform, vars_lower)
```

4.14 Corrección de observaciones seleccionadas

- Nótese que hay una entrada llamada “talpan” que debería ser “tlalpan”

[21]:
```
df_lower_values["nomgeo"].value_counts()
```

[21]:
```
iztapalapa               10515
gustavo a. madero        10058
cuauhtémoc                7313
benito juárez             6049
venustiano carranza       5179
miguel hidalgo            5110
coyoacán                  4947
azcapotzalco              4216
```

```
álvaro obregón             4140
iztacalco                  3469
xochimilco                 2450
talpan                     2140
tláhuac                    1955
tlalpan                    1064
la magdalena contreras      955
cuajimalpa de morelos       892
milpa alta                  650
Name: nomgeo, dtype: int64
```

[22]:
```
dicc_cor = {
    "nomgeo": {
        "talpan": "tlalpan"
    }
}
```

[23]:
```
df_correct = correct_selected_entries(df_lower_values, dicc_cor)
```

[24]:
```
df_correct["nomgeo"].value_counts()
```

[24]:
```
iztapalapa               10515
gustavo a. madero        10058
cuauhtémoc                7313
benito juárez             6049
venustiano carranza       5179
miguel hidalgo            5110
coyoacán                  4947
azcapotzalco              4216
álvaro obregón            4140
iztacalco                 3469
tlalpan                   3204
xochimilco                2450
tláhuac                   1955
la magdalena contreras     955
cuajimalpa de morelos      892
milpa alta                 650
Name: nomgeo, dtype: int64
```

5 Review Changes

5.1 ¿Cuántas variables tenemos?

[25]:
```
count_vars(df_correct)
```

```
Número de variables en los datos --> 17
```

5.2 ¿Cuántas observaciones tenemos?

[26]:
```
count_obs(df_correct)
```

```
Número de observaciones en los datos --> 71102
```

5.3 ¿Cuántas observaciones únicas tenemos por variable?

[27]:
```
count_unique_obs(df_correct)
```

[27]:
```
consumo_total_mixto     24339
anio                        1
nomgeo                     16
consumo_prom_dom        52060
consumo_total_dom       47051
alcaldia                   16
colonia                  1340
consumo_prom_mixto      31911
consumo_total           56015
consumo_prom            62214
consumo_prom_no_dom     37440
bimestre                    3
consumo_total_no_dom    27336
gid                     71102
indice_des                  4
latitud                 22930
longitud                22930
dtype: int64
```

5.4 ¿Cuántas variables numéricas tenemos?

[28]:
```
count_type_vars(vars_num, "numerica")
```

```
Número de variables de tipo numerica --> 8

            Variable(s)
1         consumo_total
2     consumo_total_dom
3  consumo_total_no_dom
4   consumo_total_mixto
5          consumo_prom
6      consumo_prom_dom
7   consumo_prom_no_dom
8    consumo_prom_mixto

None
```

5.5 ¿Cuántas variables de fecha tenemos?

- Para efectos de este ejercicio, no hay ninguna variable de tipo fecha, o que consideremos de fecha.

5.6 ¿Cuántas variables categóricas tenemos?

[29]:
```
count_type_vars(cat_vars, "categórica")
```

```
Número de variables de tipo categórica --> 6

  Variable(s)
1        anio
2      nomgeo
3    alcaldia
4     colonia
5    bimestre
6  indice_des

None
```

5.7 ¿Cuántas variables de texto tenemos?

- Para efectos de este ejercicio, no hay ninguna variable de tipo texto, o que consideremos de texto. Se podría considerar a la variable `gid` como un identificador de texto.

[30]:
```
count_type_vars(vars_text, "texto")
```

```
Número de variables de tipo texto --> 1

  Variable(s)
1         gid

None
```

5.8 Genera el profiling de cada variable

[31]:
```
## Data profiling compacted in function
data_profiling_numeric(df_correct, vars_num)
```

```
**********************************
** General description of data **
**********************************

             consumo_total consumo_total_dom consumo_total_no_dom  \
dtype              float64           float64              float64
count_unique         56015             47051                27336
missing_v                0              4820                    0
count                71102             66282                71102
mean               1695.85           1186.26               436.06
std                 3555.7           2771.04              2126.15
```

```
min                     0                0                    0
25%               340.952          161.635                10.98
50%               896.175          604.185                54.055
75%                1808.9          1261.45                230.43
max                119727          95060.7                119727

            consumo_total_mixto consumo_prom consumo_prom_dom  \
dtype                   float64      float64          float64
count_unique              24339        62214            52060
missing_v                  8327            0             4820
count                     62775        71102            66282
mean                     174.36      111.217          29.1324
std                      312.664      1069.95          64.5659
min                            0            0                0
25%                            0      23.0101          18.6905
50%                        79.94      31.6938          26.4142
75%                       233.32      45.4849          36.2466
max                      23404.4      89691.8          7796.41

            consumo_prom_no_dom consumo_prom_mixto
dtype                   float64            float64
count_unique              37440              31911
missing_v                     0               8327
count                     71102              62775
mean                     126.76            50.6362
std                      1095.82            130.409
min                            0                  0
25%                      6.27542                  0
50%                        19.28            33.4517
75%                       54.1869            61.2165
max                       89691.8            11702.2
```

None

```
****************************
** Top repeated variables **
****************************

      consumo_total                     consumo_total_dom                     \
              value count part_notnull             value count part_notnull
top_1          0.00  2451         3.45              0.00  9861        14.88
top_2          3.05    70         0.10              1.22    37         0.06
top_3          1.22    68         0.10             10.98    21         0.03
top_4          3.66    42         0.06             25.62    20         0.03
top_5          6.71    41         0.06              3.66    20         0.03
```

```
      consumo_total_no_dom                    consumo_total_mixto         \
                     value count part_notnull               value  count
top_1                 0.00  8109        11.40                 0.0  17715
top_2                 1.22   402         0.57                36.0     74
top_3                 1.83   316         0.44                17.7     61
top_4                 3.05   302         0.42                36.6     59
top_5                 7.93   219         0.31                18.3     54

                   consumo_prom                    consumo_prom_dom        \
      part_notnull        value count part_notnull            value count
top_1        28.22         0.00  2451         3.45             0.00  9861
top_2         0.12         1.22    62         0.09             1.22    33
top_3         0.10         3.05    55         0.08            14.64    23
top_4         0.09         4.27    43         0.06            10.98    22
top_5         0.09         6.71    39         0.05            15.25    22

                   consumo_prom_no_dom                    consumo_prom_mixto  \
      part_notnull               value count part_notnull               value
top_1        14.88                0.00  8109        11.40                0.00
top_2         0.05                1.22   330         0.46               36.00
top_3         0.03                1.83   290         0.41               29.28
top_4         0.03                3.05   260         0.37               36.60
top_5         0.03                4.27   216         0.30               23.80


       count part_notnull
top_1  17715        28.22
top_2     58         0.09
top_3     57         0.09
top_4     53         0.08
top_5     49         0.08
None
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
```

[32]:
```
df_correct["nomgeo"].value_counts()
```

[32]:
```
iztapalapa               10515
gustavo a. madero        10058
cuauhtémoc                7313
benito juárez             6049
venustiano carranza       5179
miguel hidalgo            5110
coyoacán                  4947
azcapotzalco              4216
```

```
álvaro obregón             4140
iztacalco                  3469
tlalpan                    3204
xochimilco                 2450
tláhuac                    1955
la magdalena contreras      955
cuajimalpa de morelos       892
milpa alta                  650
Name: nomgeo, dtype: int64
```

[33]:
```
#data profiling function
data_profiling_categ(df_correct,cat_vars)
```

```
**********
Variable Categorica anio
**********

Info                         anio
Num_Registros               71102
Num_de_categorias               1
Moda                         2019
Valores_faltantes               0
Top1                [2019, 71102]
Top2                            0
Top3                            0

None
Valores de las categorias y sus proporciones

           Observaciones proporción
Categoría
2019               71102     100.0%

None



**********
Variable Categorica nomgeo
**********

Info                                    nomgeo
Num_Registros                            71102
Num_de_categorias                           16
Moda                                iztapalapa
Valores_faltantes                            0
Top1                       [iztapalapa, 10515]
Top2                [gustavo a. madero, 10058]
Top3                        [cuauhtémoc, 7313]
```

None
Valores de las categorias y sus proporciones

Categoría	Observaciones	proporción
iztapalapa	10515	14.8%
gustavo a. madero	10058	14.1%
cuauhtémoc	7313	10.3%
benito juárez	6049	8.5%
venustiano carranza	5179	7.3%
miguel hidalgo	5110	7.2%
coyoacán	4947	7.0%
azcapotzalco	4216	5.9%
álvaro obregón	4140	5.8%
iztacalco	3469	4.9%
tlalpan	3204	4.5%
xochimilco	2450	3.4%
tláhuac	1955	2.7%
la magdalena contreras	955	1.3%
cuajimalpa de morelos	892	1.3%
milpa alta	650	0.9%

None

****
Variable Categorica alcaldia
****

Info	alcaldia
Num_Registros	71102
Num_de_categorias	16
Moda	iztapalapa
Valores_faltantes	0
Top1	[iztapalapa, 10515]
Top2	[gustavo a. madero, 10058]
Top3	[cuauhtemoc, 7313]

None
Valores de las categorias y sus proporciones

Categoría	Observaciones	proporción
iztapalapa	10515	14.8%
gustavo a. madero	10058	14.1%
cuauhtemoc	7313	10.3%
benito juarez	6049	8.5%
venustiano carranza	5179	7.3%
miguel hidalgo	5110	7.2%

coyoacan	4947	7.0%
azcapotzalco	4216	5.9%
alvaro obregon	4140	5.8%
iztacalco	3469	4.9%
tlalpan	3204	4.5%
xochimilco	2450	3.4%
tlahuac	1955	2.7%
magdalena contreras	955	1.3%
cuajimalpa	892	1.3%
milpa alta	650	0.9%

None

****
Variable Categorica colonia
****

Info	colonia
Num_Registros	71102
Num_de_categorias	1340
Moda	centro
Valores_faltantes	0
Top1	[centro, 1139]
Top2	[agricola oriental, 837]
Top3	[roma norte, 602]

None
Valores de las categorias y sus proporciones

Categoría	Observaciones	proporción
centro	1139	1.6%
agricola oriental	837	1.2%
roma norte	602	0.8%
moctezuma 2a seccion	558	0.8%
jardin balbuena	498	0.7%
…	…	…
calzada jalalpa	2	0.0%
huizachito	2	0.0%
piru secc. i	2	0.0%
u. hab. nueva rosita	1	0.0%
santisima trinidad	1	0.0%

[1340 rows x 2 columns]

None

```
*********************************
Variable Categorica bimestre
*********************************

Info                 bimestre
Num_Registros           71102
Num_de_categorias           3
Moda                        2
Valores_faltantes           0
Top1                [2, 23942]
Top2                [3, 23822]
Top3                [1, 23338]

None
Valores de las categorias y sus proporciones

           Observaciones proporción
Categoría
2                  23942      33.7%
3                  23822      33.5%
1                  23338      32.8%

None



*********************************
Variable Categorica indice_des
*********************************

Info                     indice_des
Num_Registros                 71102
Num_de_categorias                 4
Moda                           bajo
Valores_faltantes                 0
Top1                   [bajo, 29248]
Top2                [popular, 16539]
Top3                   [alto, 15516]

None
Valores de las categorias y sus proporciones

           Observaciones proporción
Categoría
bajo               29248      41.1%
popular            16539      23.3%
alto               15516      21.8%
medio               9799      13.8%

None
```

5.9 Data Profiling con Pandas-Profiling

[34]:
```
profile = ProfileReport(df_correct, title="Pandas Profiling Report",␣
↪explorative = True)
```

[35]:
```
profile
```

```
HBox(children=(HTML(value='Summarize dataset'), FloatProgress(value=0.0, max=31.
↪0), HTML(value='')))


HBox(children=(HTML(value='Generate report structure'), FloatProgress(value=0.0,␣
↪max=1.0), HTML(value='')))


HBox(children=(HTML(value='Render HTML'), FloatProgress(value=0.0, max=1.0),␣
↪HTML(value='')))


<IPython.core.display.HTML object>
```

[35]:

[36]:
```
profile.to_file("Profile_variables.html")
```

```
HBox(children=(HTML(value='Export report to file'), FloatProgress(value=0.0,␣
↪max=1.0), HTML(value='')))
```

6 Graphic Exploratory Data Analysis (GEDA)

6.1 Análisis Univariado

6.1.1 *Variables Categóricas*

Barplots La función que diseñamos para crear gráficas de barras está pensada para que el usuario especifique la variable que desea visualizar. - La lista de variables que se pueden graficar con esta función es: - “anio” - “nomgeo”, - “alcaldia”, - “colonia”, - “bimestre”, - “indice_des”,

[37]:
```
df_plot = df_correct.copy()
```

[38]:
```
barplot_cat(df_plot, "alcaldia", tops=10)
```

```
Otras_categs contiene la siguiente información:
    -> 6 categorías (37.50%)
    -> su conteo de valores representa el (14.21%) del conteo total
```

6.1.2 *Varibles Númericas*

Se definen las listas de las variables a explorar:

[39]:
```
vars_num_tot= [
    "consumo_total",
    "consumo_total_dom",
    "consumo_total_no_dom",
    "consumo_total_mixto",
    "indice_des"]

vars_num_prom= [
    "consumo_prom",
    "consumo_prom_dom",
    "consumo_prom_no_dom",
    "consumo_prom_mixto",
    "indice_des"]

vars_num_i= [
    "consumo_total",
    "consumo_total_dom",
    "consumo_total_no_dom",
    "consumo_total_mixto",
    "consumo_prom",
    "consumo_prom_dom",
    "consumo_prom_no_dom",
    "consumo_prom_mixto",
    "indice_des",
    "alcaldia",
    "bimestre"]
```

Histogramas La función que diseñamos para crear histogramas está pensada para que el usuario especifique la variable que desea visualizar. - La lista de variables que se pueden graficar con esta función es: - 'consumo_total', - 'consumo_total_dom', - 'consumo_total_no_dom', - 'consumo_total_mixto', - 'consumo_prom', - 'consumo_prom_dom', - 'consumo_prom_no_dom', - 'consumo_prom_mixto'

[40]:
```
histograms_numeric_total(df_plot, "consumo_total")
```

A partir de la exploración de los datos numéricos por medio de histogramas, se notó la presencia de muchos datos atípicos.

6.1.3 Distribución del consumo de agua por índice de desarrollo.

Se muestran los histogramas del consumo (variables numéricas) por cada categoría del índice de desarrollo(`indice_des`).

Con el objetivo de tener una mejor observación del comportamiento de la distribución del consumo (totales y promedios), se transformaron las variables en escala logarítmica.

Sin embargo, de acuerdo con el data profiling en el EDA, se observó que para todos los consumos (totales y promedio), el valor top 1 es *cero*. Por lo tanto, la distribuciones logarítmicas no mostrarán estos valores.

Con base en lo anterior, y dado que la granularidad de los datos es a nivel manzana, ¿Existen muchas manzanas sin consumo de agua? ¿Es correcto lo anterior o es un error? Contestaremos la pregunta más adelante.

La función que diseñamos para crear histogramas está pensada para que el usuario especifique la variable que desea visualizar. - La lista de variables que se pueden graficar con esta función es: - 'consumo_total', - 'consumo_total_dom', - 'consumo_total_no_dom', - 'consumo_total_mixto', - 'consumo_prom', - 'consumo_prom_dom', - 'consumo_prom_no_dom', - 'consumo_prom_mixto'

[41]:
```
histograms_numeric(df_plot, "consumo_prom_mixto", "indice_des")
```

```
/Users/rp_mbp/.pyenv/versions/itam_intro_to_ds/lib/python3.7/site-
packages/pandas/core/series.py:726: RuntimeWarning:

divide by zero encountered in log
```

Boxplots Los boxplots reafirman que para todas las variables numéricas existen outliers, incluso por cada categoría del índice desarrollo.

La función que diseñamos para crear gráficas de baja y brazos está pensada para que el usuario especifique la variable que desea visualizar. - La lista de variables que se pueden graficar con esta función es: - 'consumo_total', - 'consumo_total_dom', - 'consumo_total_no_dom', - 'consumo_total_mixto', - 'consumo_prom', - 'consumo_prom_dom', - 'consumo_prom_no_dom', - 'consumo_prom_mixto'

[42]:
```
box_plot_num(df_plot, "indice_des", "consumo_total_mixto")
```

Scatterplots La función que diseñamos para crear scatterplots está pensada para que el usuario especifique las 2 variables que desea visualizar. - La lista de variables que se pueden graficar con esta función es: - "consumo_total_mixto", - "consumo_total_dom", - "consumo_total_no_dom", - "consumo_total"

- Variables totales del consumo:

[43]:
```
scatterPlotFacet(df_plot, "consumo_total_dom", "consumo_total_no_dom",␣
↪"indice_des", "bimestre")
```

- Variables promedio del consumo:
- La lista de variables que se pueden graficar con esta función es:
  - "consumo_prom_dom",
  - "consumo_prom_mixto",
  - "consumo_prom_no_dom",
  - "consumo_prom"

[44]:
```
scatterPlotFacet(df_plot, "consumo_prom_dom", "consumo_prom_no_dom",␣
↪"indice_des", "bimestre")
```

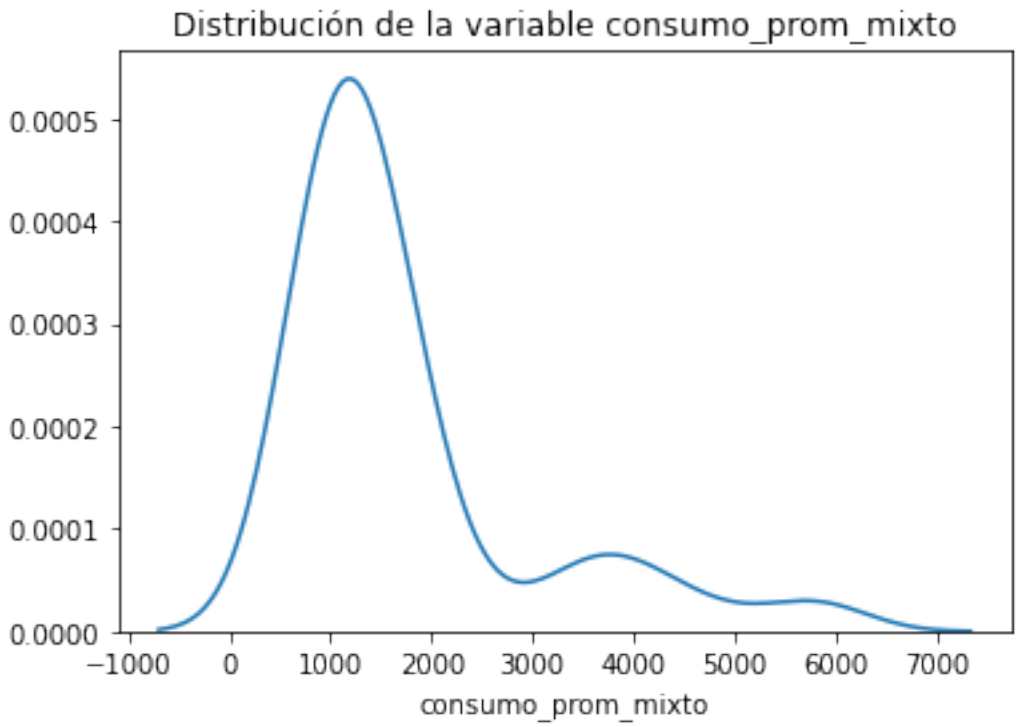
Rugplot La función que diseñamos para crear gráficas de tapete está pensada para que el usuario especifique la variable que desea visualizar. - La lista de variables que se pueden graficar con esta función es: - 'consumo_total', - 'consumo_total_dom', - 'consumo_total_no_dom', - 'consumo_total_mixto', - 'consumo_prom', - 'consumo_prom_dom', - 'consumo_prom_no_dom', - 'consumo_prom_mixto'

[45]:
```
rugplot_num(df_plot, "consumo_prom_no_dom")
```

Density Estimate La función que diseñamos para crear gráficas de densidad está pensada para que el usuario especifique la variable que desea visualizar. - La lista de variables que se pueden graficar con esta función es: - 'consumo_total', - 'consumo_total_dom', - 'consumo_total_no_dom', - 'consumo_total_mixto', - 'consumo_prom', - 'consumo_prom_dom', - 'consumo_prom_no_dom', - 'consumo_prom_mixto'

[46]:
```
distplot_num(df_plot, "consumo_prom_mixto", 85)
```

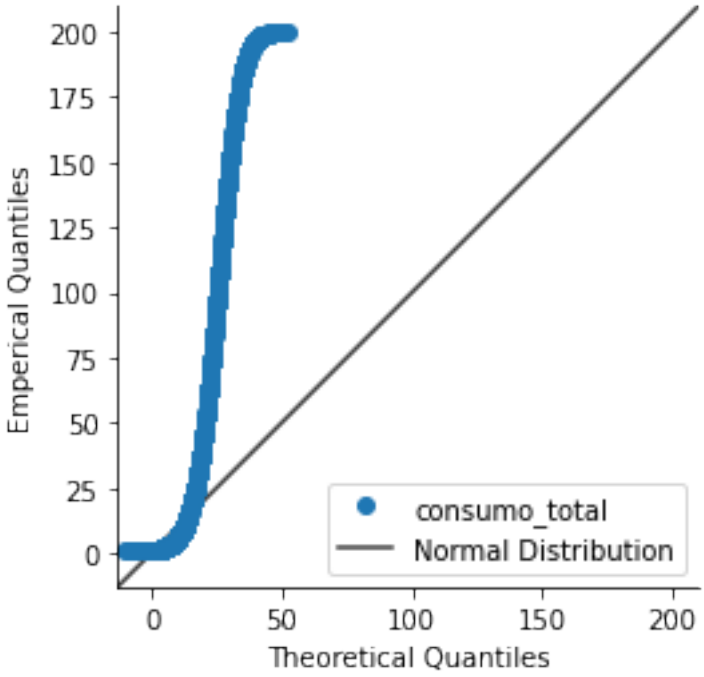
```
/Users/rp_mbp/.pyenv/versions/itam_intro_to_ds/lib/python3.7/site-
packages/seaborn/distributions.py:2551: FutureWarning:

`distplot` is a deprecated function and will be removed in a future version.
Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `kdeplot` (an axes-level function for kernel density
plots).
```



QQ-Plot La función que diseñamos para crear gráficas qq está pensada para que el usuario especifique la variable que desea visualizar. - La lista de variables que se pueden graficar con esta función es: - 'consumo_total', - 'consumo_total_dom', - 'consumo_total_no_dom', - 'consumo_total_mixto', - 'consumo_prom', - 'consumo_prom_dom', - 'consumo_prom_no_dom', - 'consumo_prom_mixto'

[47]:
```
qq_plot(data = df_plot, variable = "consumo_total", ymin = 1, ymax = 200)
```

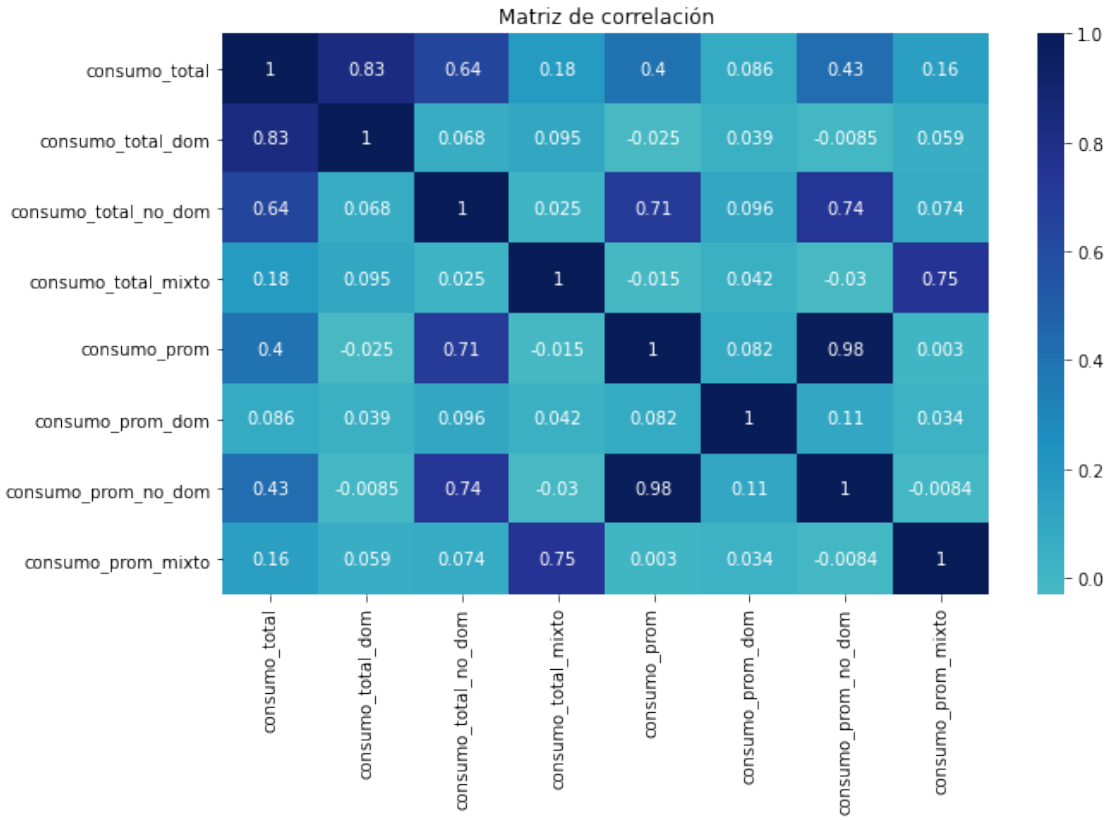
```
Porcentaje de datos conservado 0.15581840173272202
```



Matriz de Correlación

[48]:
```
corr_plot(data = df, variables = vars_num, title = "Matriz de correlación")
```

[48]:
```
Text(0.5, 1.0, 'Matriz de correlación')
```



6.2 Análisis Multivariado

Distribución del consumo de agua por categoría del ínidice de desarrollo en cada alcaldía.

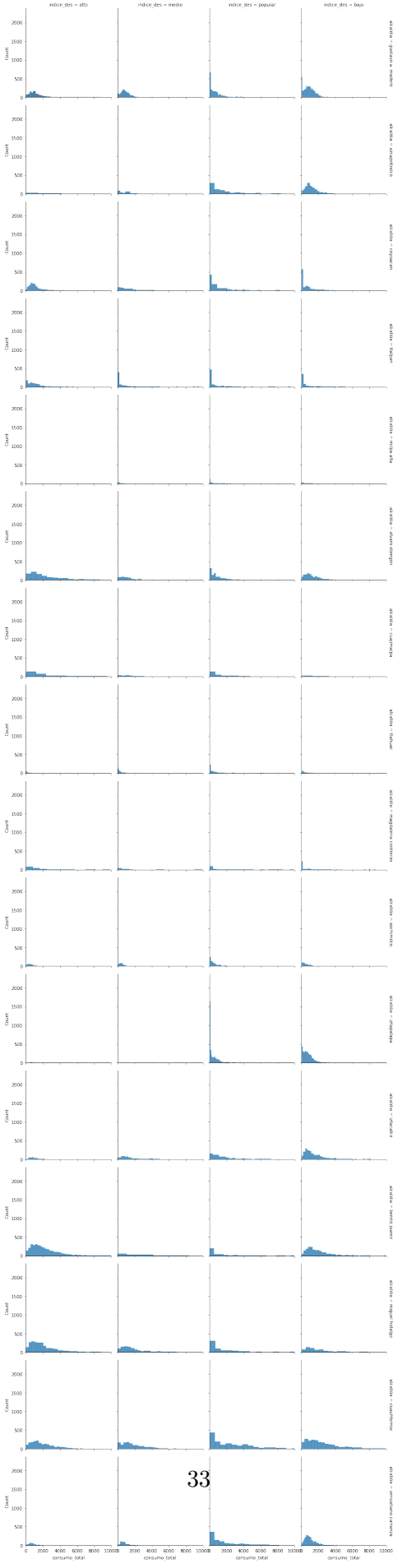
- Conteo de registros por alcaldía por `indice_des`

[49]:
```
create_heatmap(df_plot, "alcaldia", "indice_des", "gid")
```

Con el mapa de calor por alcaldía podemos ver que sí hay ubicaciones específicas con una tendencia clara a ser clasificadas con un indice_des particular (e.g. si el registro es de Azcapotzalco, tenderá a ser clasificado como "bajo"). (el mapa de calor muestra la proporción de conteos de cada clasificación por alcaldía)

Histogramas de distribución del consumo por índice de desarrollo humano y variables categóricas. La función que diseñamos para crear este grupo de histogramas está pensada para que el usuario especifique la variable de consumo que desea visualizar. - La lista de variables que se pueden graficar con esta función es: - 'consumo_total', - 'consumo_total_dom', - 'consumo_total_no_dom', - 'consumo_total_mixto', - 'consumo_prom', - 'consumo_prom_dom', - 'consumo_prom_no_dom', - 'consumo_prom_mixto'

[50]:
```
histograms_numeric_rv_cat(df_plot, "consumo_total", "indice_des", "alcaldia")
```

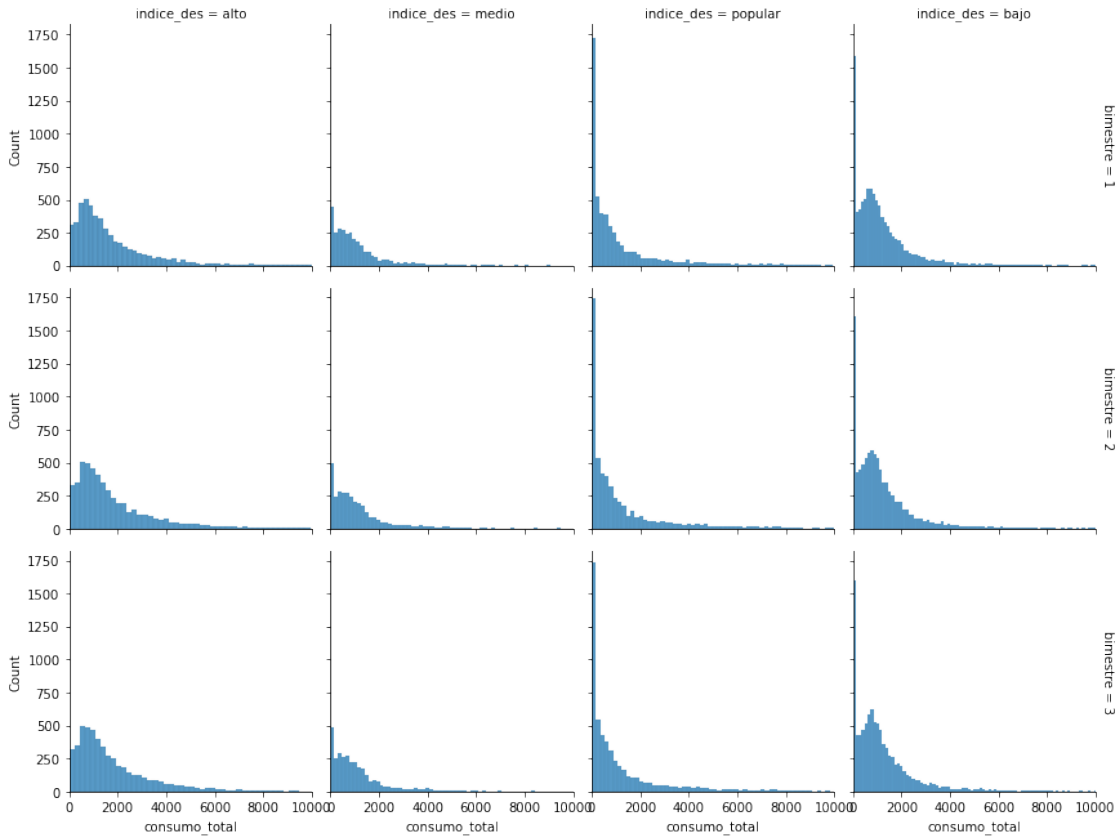

indice_des = alto
indice_des = medio
indice_des = popular
indice_des = bajo
Count
consumo_total

Boxplot del consumo por alcaldía: La función que diseñamos para crear esta visualización de histogramas está pensada para que el usuario especifique la variable de consumo que desea visualizar. - La lista de variables que se pueden graficar con esta función es: - 'consumo_total', - 'consumo_total_dom', - 'consumo_total_no_dom', - 'consumo_total_mixto', - 'consumo_prom', - 'consumo_prom_dom', - 'consumo_prom_no_dom', - 'consumo_prom_mixto'

[51]:
```
box_plot_num_location(df_plot, "consumo_total_no_dom", "iztapalapa")
```

Distribución del consumo de agua por categoría del ínidice de desarrollo por bimestre. La función que diseñamos para crear esta visualización de histogramas está pensada para que el usuario especifique la variable de consumo que desea visualizar. - La lista de variables que se pueden graficar con esta función es: - 'consumo_total', - 'consumo_total_dom', - 'consumo_total_no_dom', - 'consumo_total_mixto', - 'consumo_prom', - 'consumo_prom_dom', - 'consumo_prom_no_dom', - 'consumo_prom_mixto'

[52]:
```
histograms_numeric_rv_cat(df_plot, "consumo_total", "indice_des", "bimestre")
```

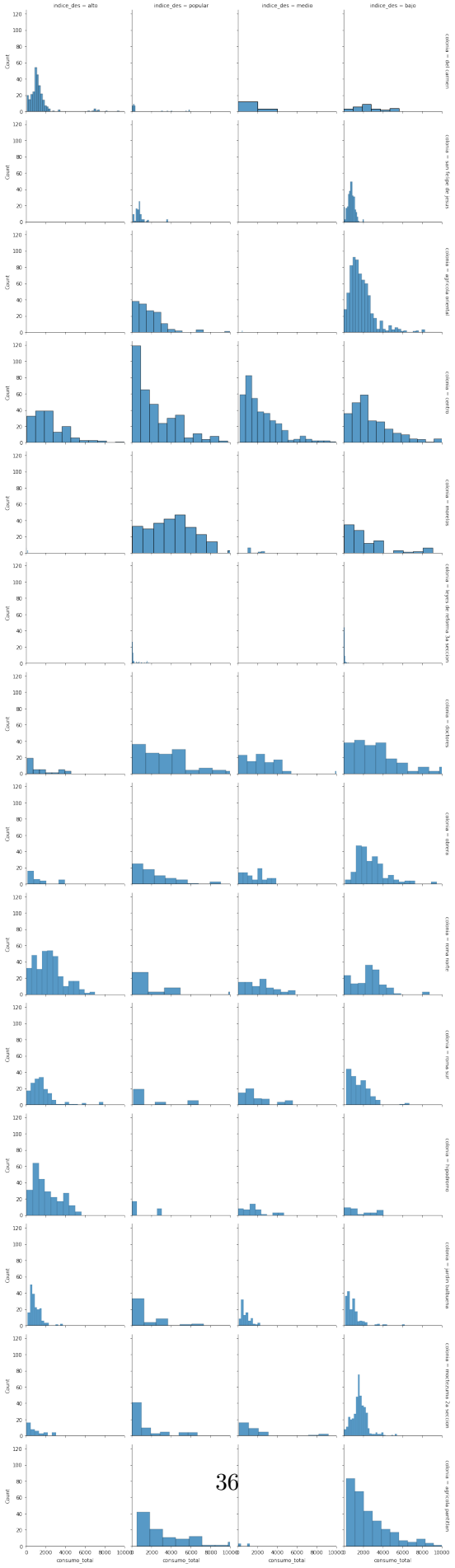


Distribución del consumo de agua por categoría del ínidice de desarrollo por colonia. Se enlistan las colonias top 15 con mayor número de observaciones:

[53]:
```
df_colonia = df_plot[df_plot["colonia"].isin(colonia_top_15)]
```

La función que diseñamos para crear esta visualización de histogramas está pensada para que el usuario especifique la variable de consumo que desea visualizar. - La lista de variables que se pueden graficar con esta función es: - 'consumo_total', - 'consumo_total_dom', - 'consumo_total_no_dom', - 'consumo_total_mixto', - 'consumo_prom', - 'consumo_prom_dom', - 'consumo_prom_no_dom', - 'consumo_prom_mixto'

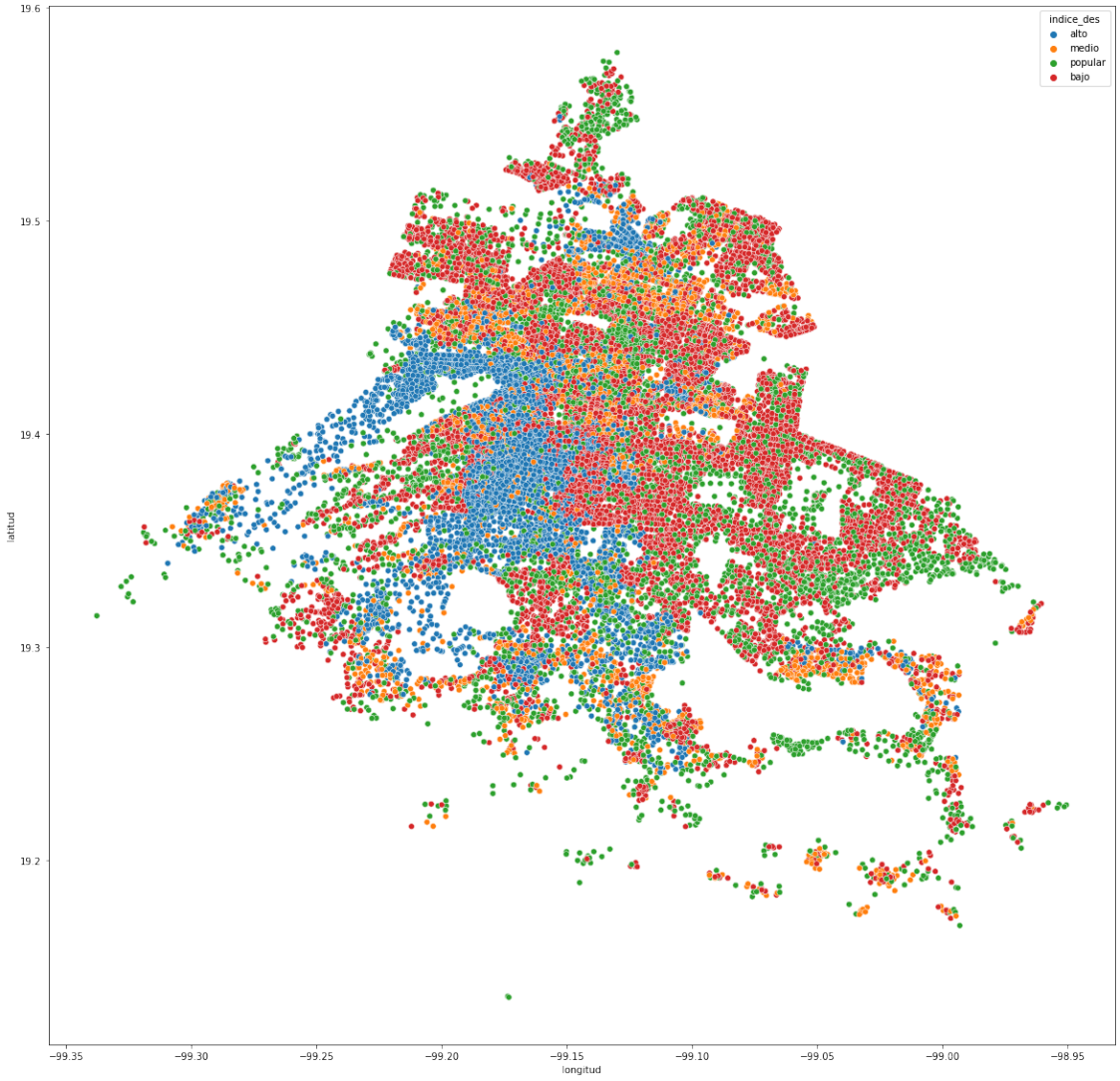
[54]:
```
histograms_numeric_rv_cat(df_colonia, "consumo_total", "indice_des", "colonia")
```


indice_des = alto
indice_des = popular
indice_des = medio
indice_des = bajo
colonia = del carmen
colonia = los reyes de juárez
colonia = agrícola oriental
colonia = centro
colonia = moderna
colonia = tepeyac de zaragoza la ampliación
colonia = aquiahuac
colonia = obrera
colonia = nueva santa maría
colonia = nueva vallejo
colonia = ampliación san pablo xalpa
colonia = moctezuma 2a sección
colonia = agrícola pantitlán
Count
consumo_total

Distribución espacial de la variable de respuesta `indice_des`

[55]:
```
scatter_map(df_plot)
```

```
/Users/rp_mbp/.pyenv/versions/itam_intro_to_ds/lib/python3.7/site-
packages/seaborn/_decorators.py:43: FutureWarning:

Pass the following variables as keyword args: x, y. From version 0.12, the only
valid positional argument will be `data`, and passing other arguments without an
explicit keyword will result in an error or misinterpretation.
```



Consistencia de la clasificación de la variable `colonia` con el `indice_des`. Lo que nos interesa averiguar con este analisis es si las distintas colonias están clasificadas con una sola etiqueta de índice de desarrollo (e.g. la colonia "Navidad" siempre es clasificada como "popular")

[56]:
```
colonia_devidx_consistency(df_plot)
```

Con esta gráfica nos damos cuenta de que no hay mucha consistencia en la clasificación. Solo el 21% de las las colonias fueron consistentemente clasificadas con una sola etiqueta de la variable `indice_des`.