Segmentación y Predicción de los Ingresos de los organismos Operadores de los servicios de Agua Potable, Alcantarillado y Saneamiento en México

Proyecto final

Jessica Briseño

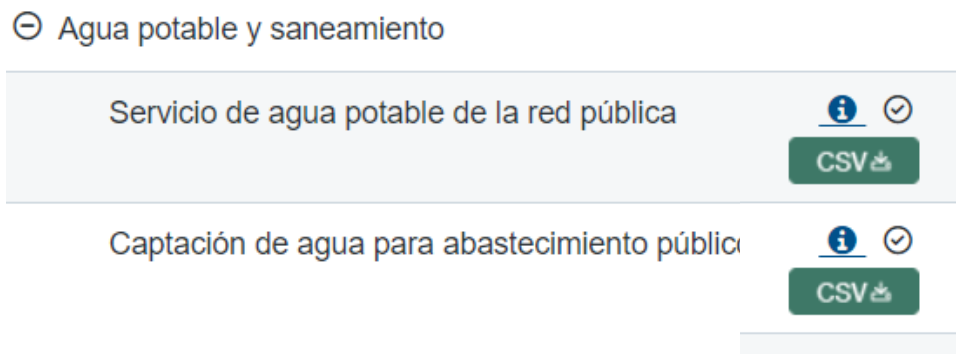
jevabrir@gmail.com

Introducción

En México quienes se encargan de la prestación de los servicios de agua potable, alcantarillado y saneamiento se les conoce como OAPAS (Operador de Agua Potable, Alcantarillado y Saneamiento).

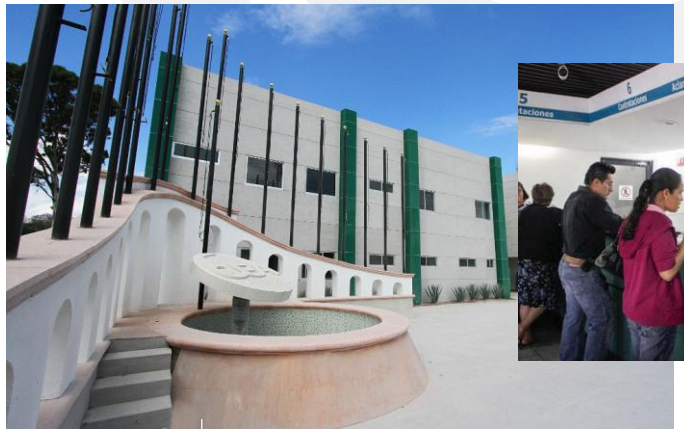
El INEGI realizó durante el año 2023 el Censo Nacional de Gobiernos Municipales y Demarcaciones Territoriales de la Ciudad de México con el que se puede analizar la gestión y desempeño de la administración pública municipal y de las demarcaciones de la Ciudad de México.

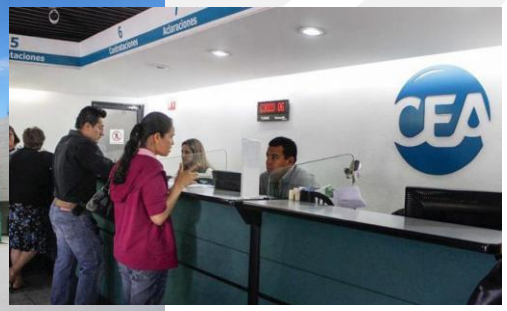
Módulo 5 contiene información sobre la gestión, administración, características técnicas y ambientales de la prestación de los servicios municipales de agua potable y saneamiento (INEGI, 2021).



Planteamiento del problema

En el Censo Gubernamental Municipal se reporta un total de **2469** folios que corresponden con los OAPAS en los diferentes municipios de México.





Este tipo de organismos presentan diferencias tanto en el tamaño (volumen servido o el número de clientes) como en su estructura, lo que ha dificultado el diseño de un método para clasificarlos

Es por lo que resulta esencial analizar y seleccionar algunos elementos del Censo Gubernamental Municipal, que tienen importancia para agrupar y caracterizar a los OAPAS en México, con el propósito lograr una comprensión de la relación entre los OAPAS clasificados y diseñar políticas públicas acordes a cada grupo.

Objetivo

El principal objetivo de este proyecto es implementar el algoritmo de aprendizaje no supervisado K-Means para agrupar y caracterizar a los cerca de 2400 prestadores de los servicios de agua potable, alcantarillado y saneamiento que se encuentran en México, a partir de los datos abiertos del Censo Nacional de Gobiernos Municipales y Demarcaciones Territoriales de la Ciudad de México 2022 (INEGI,2022), para lograr una comprensión de la relación de las variables que estos prestadores de servicios tienen en común.

Para lograr este objetivo se desarrollan los siguientes objetivos específicos:

1. Descargar los diversos archivos con formato .csv correspondientes al Censo Nacional de Gobiernos Municipales y Demarcaciones Territoriales de la Ciudad de México 2022 de la página de INEGI, para diseñar una dataset con los datos requeridos para el análisis.
2. Realizar un análisis exploratorio de los datos para revisar los valores faltantes, identificar datos atípicos, imputar valores y seleccionar las variables para llevar el análisis.
3. Realizar el preprocesamiento necesario, así como implementar los algoritmos K-Means, Hierarchical y DBScan para segmentar a los prestadores de los servicios de agua potable y alcantarillado.
4. Utilizando la métrica de rendimiento “Silhouette score”, identificar el mejor algoritmo de clasificación e implementarlo para clasificar a los OAPAS.
5. Realizar la caracterización y descripción de la segmentación de los OAPAS.
6. Implementar el método de regresión lineal múltiple para generar un modelo matemático para estimar el ingreso de los OAPAS en función de cuatro variables de estudio.

Metodología

1 Creación de dataset

- Importar datos Censo INEGI
- Creación columnas de Estado y Municipio (diccionarios)

2 Análisis Exploratorio

- Revisión de valores y valores nulos
- Descripción estadística
- Visualización de datos
- Detección de datos atípicos y eliminación de valores
- Imputación de valores
- Visualización de datos

3 Análisis de correlaciones

- Correlación

4 Segmentación OAPAS

- Estandarización de los datos
- Segmentación con Hierarchical
- Segmentación con DBScan
- Segmentación con K-Means
- Selección del modelo
- Asignación del clústeres
- Evaluación (silhouette_score)
- Visualización
- Caracterización de las Clases

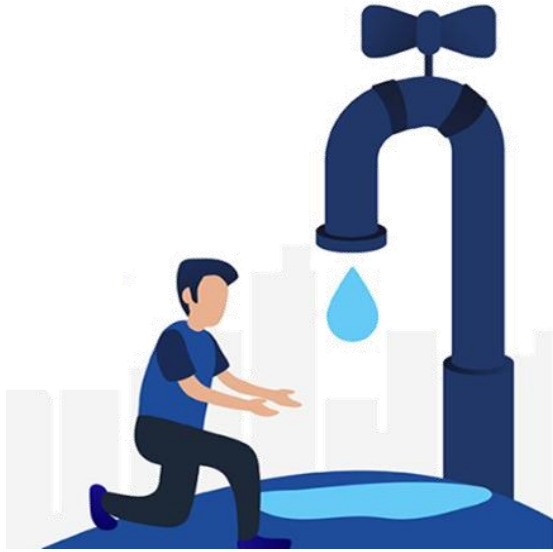
5 Modelo de regresión

- Datos de entrenamiento y prueba
- Estandarización
- Entrenamiento del Modelo
- Evaluación del Modelo
- Visualización

6 Conclusiones

JupyterLab

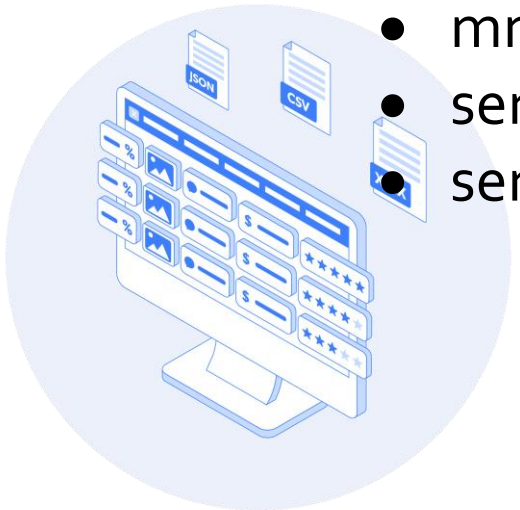
- **Pandas**
- **Numpy**
- **Matplotlib**
- **Seaborn**
- **Sklearn**
  - PCA
  - StandardScaler
  - Kmeans
  - silhouette_score



Resultados

1. Creación dataset

Para la creación de dataset se descargaron de la página de datos abiertos de INEGI diversos archivos que fueron analizados utilizando las librerías de Pandas, Numpy y Matplotlib de Python. Los siguientes archivos con formatos de archivo separado por comas (csv) o Excel (xls) contienen la información base para la conformación del dataset:



- admncion_cngmd2021.xls
- entidad_cngmd2021.csv
- mnpio_cngmd2021.csv
- servagua_cngmd2021.xls
- servdren_cngmd2021.xls

2469 filas
73 columnas

Las variables de este archivo que conforman el dataset a analizar son:

- **folio**: Indica el identificador de cada cuestionario compuesto por la clave de entidad y por la clave de municipio o delegación
- **totl_tom:** Número total de tomas que cubre el servicio de agua entubada de la red pública
- **conx_tot**: Número total de conexiones a la red de drenaje y alcantarillado por tipo de usuario.
- **Pobl_aPot**: Porcentaje de población que contaba con acceso al servicio de agua de la red pública.
- **Pob_dren**: Porcentaje de la población que tenía acceso al servicio de drenaje y alcantarillado de la red pública.
- **total_in**: Ingreso por el suministro de agua potable y saneamiento durante el año 2020

- **CVE_Ent**
- **Name_Ent**
  Nombre de la entidad federativa
- **Name_Mun**
  Nombre del municipio.

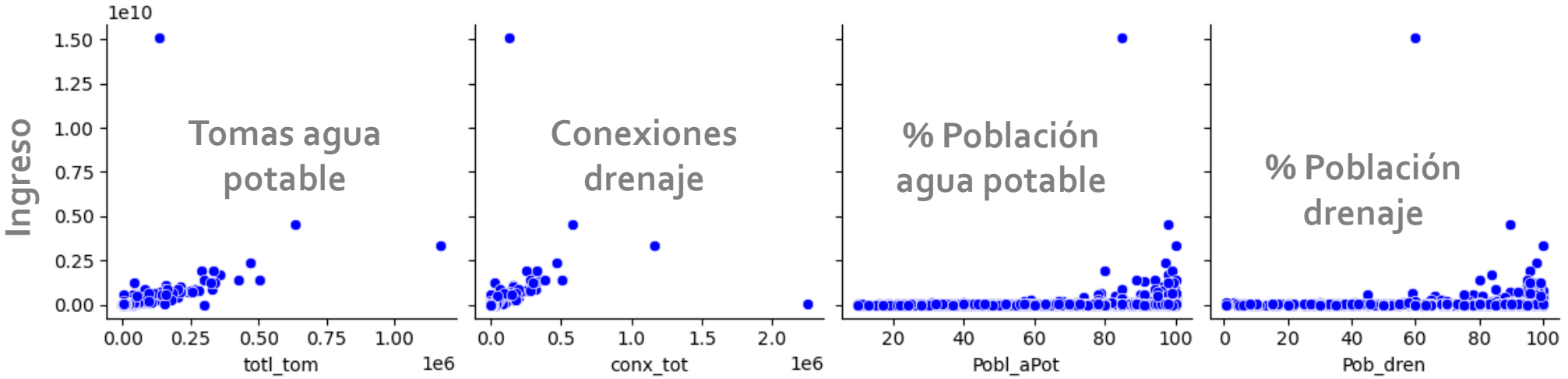
2469 filas, 9 columnas

2. Análisis Exploratorio

Utilizando el dataset se realizó un análisis exploratorio de los datos para revisar los valores faltantes, identificar datos atípicos e los imputar valores.

	Valores faltantes
totl_tom	131
conx_tot	708
Pob_dren	563
Pobl_aPot	57
total_in	692
folio	0
cve_ent	0
Name_Ent	0
Name_Mun	0

	count	mean	std	min	25%	50%	75%	max
totl_tom	2,338.000	11,617.353	46,772.576	4.000	711.000	2,051.500	6,000.000	1,170,135.000
conx_tot	1,761.000	13,956.785	72,134.443	3.000	680.000	2,051.000	6,726.000	2,250,000.000
Pob_dren	1,906.000	73.816	24.843	0.500	60.000	80.000	94.000	100.000
Pobl_aPot	2,412.000	77.460	24.791	10.000	70.000	89.000	96.000	100.000
total_in	1,777.000	50,885,566.946	414,168,338.008	321.000	160,000.000	1,280,000.000	8,320,415.370	15,094,274,894.000



2.1 Valores atípicos

La detección de los valores atípicos se llevó a cabo la realización de tablas de información en orden ascendente y descendente para cada una de las variables de estudio, así como diversas gráficas.

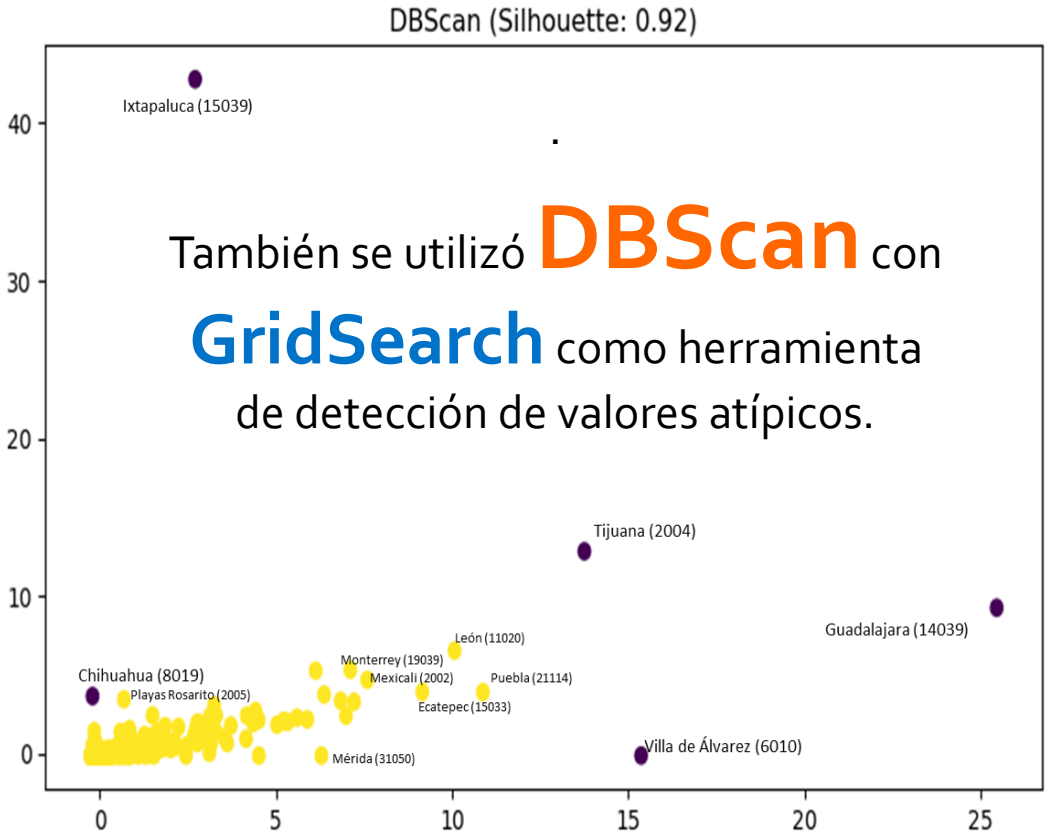
Los datos abiertos de INEGI se detectaron un total de:

75
valores atípicos

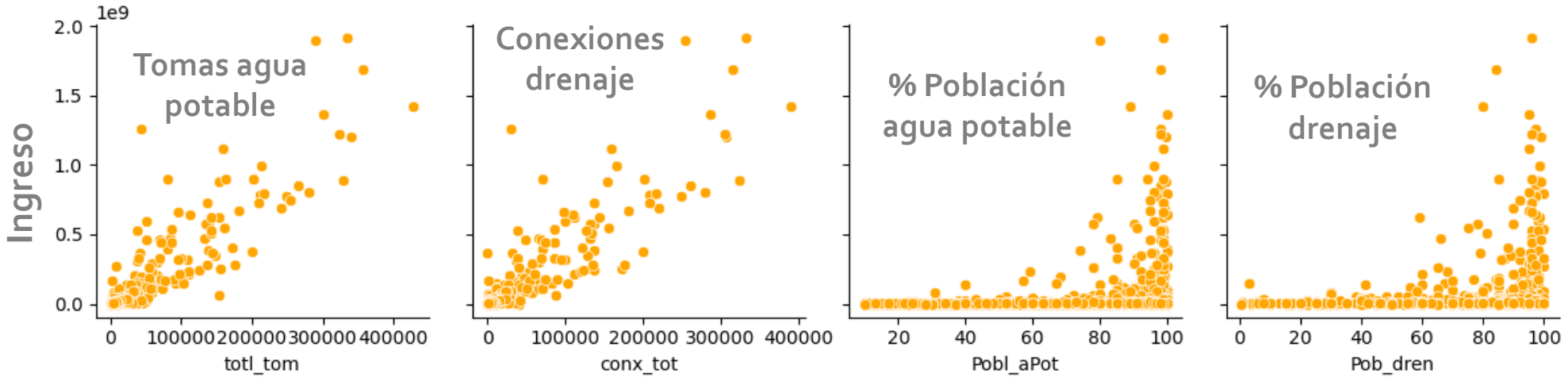
5 de los cuales corresponden a OAPAS en donde se encuentran algunos de los núcleos de población más importantes de México y que corresponden a los municipios de

- **Tijuana**
- **Guadalajara**
- **Chihuahua**
- **Puebla**
- **León.**

Dataset
2394 filas, 9 columnas

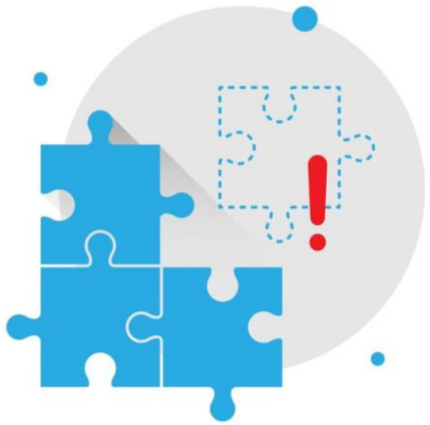


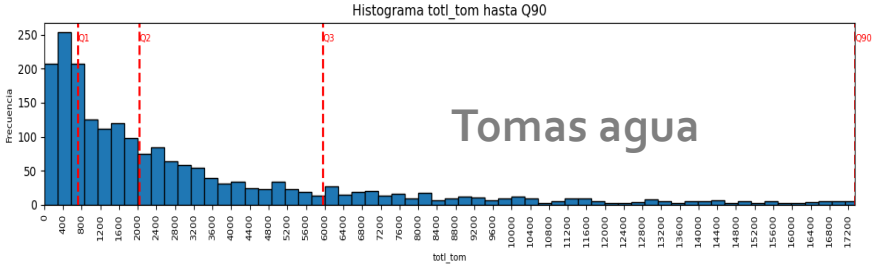
También se utilizó **DBScan** con **GridSearch** como herramienta de detección de valores atípicos.

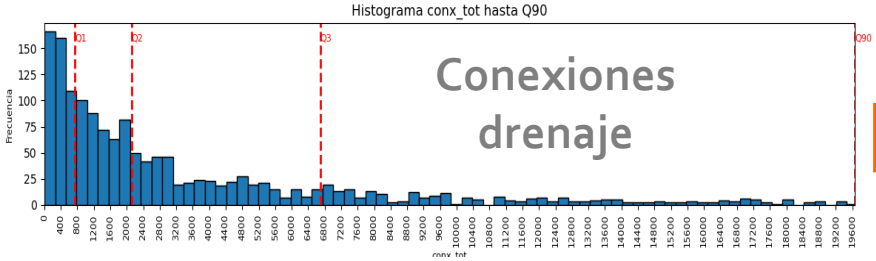


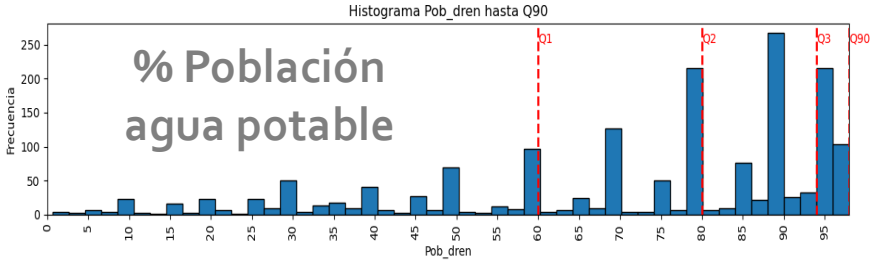
2.2 Imputación de valores

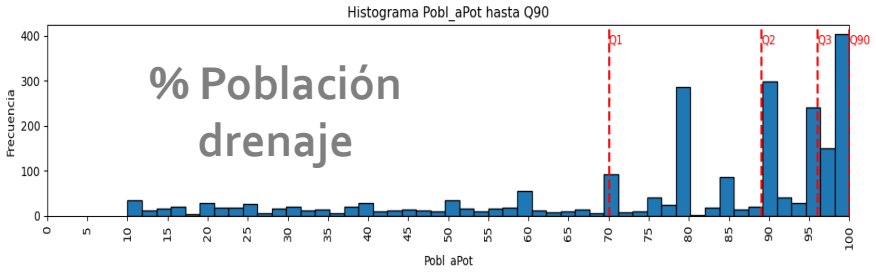
En este trabajo se lleva a cabo la imputación de valores faltantes con el objetivo de no perder valores importantes para el análisis.

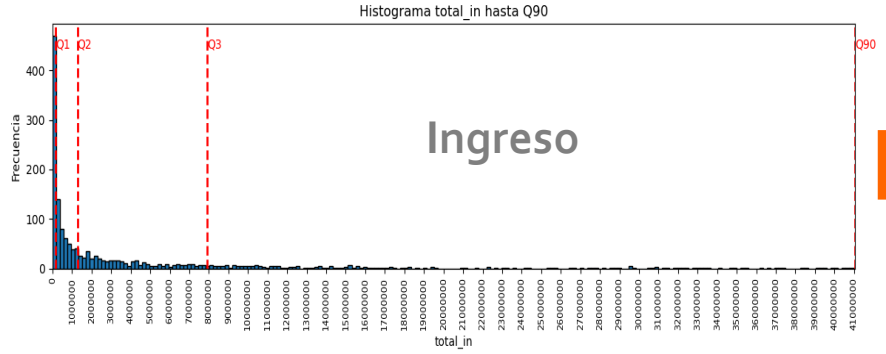


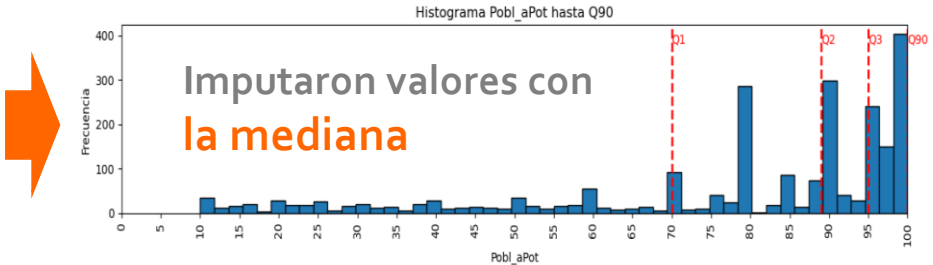


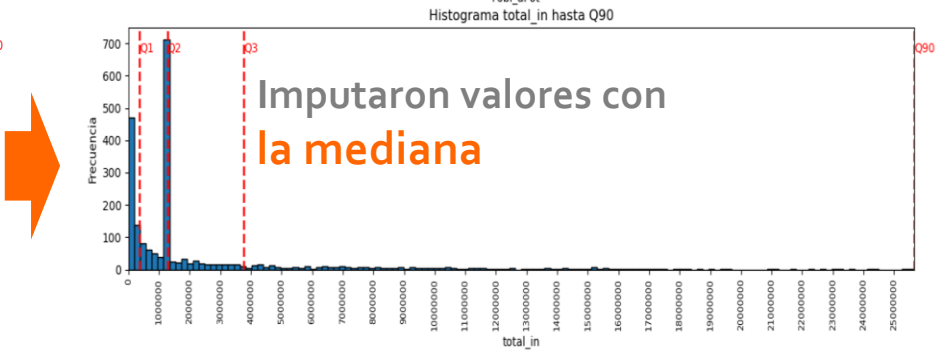








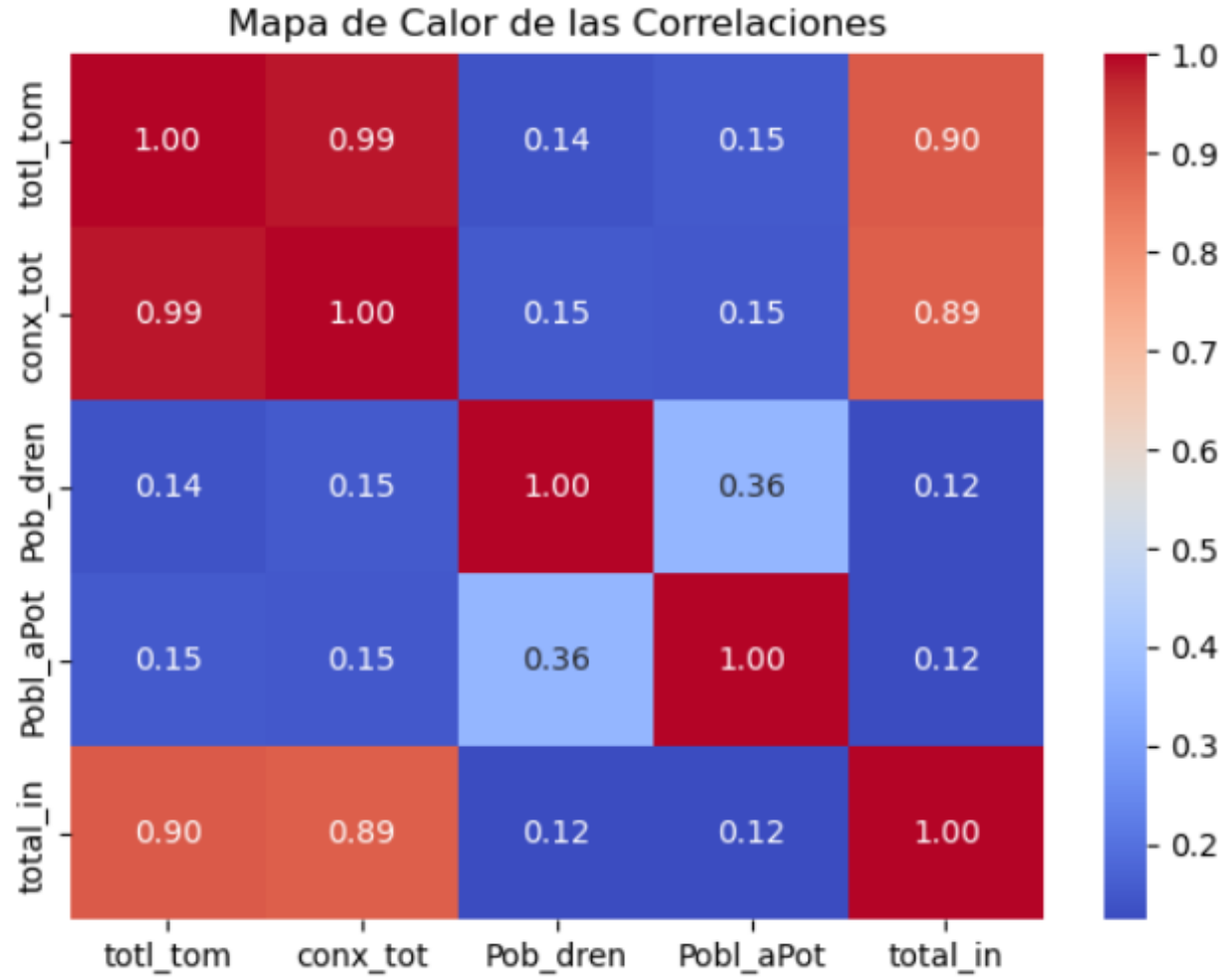


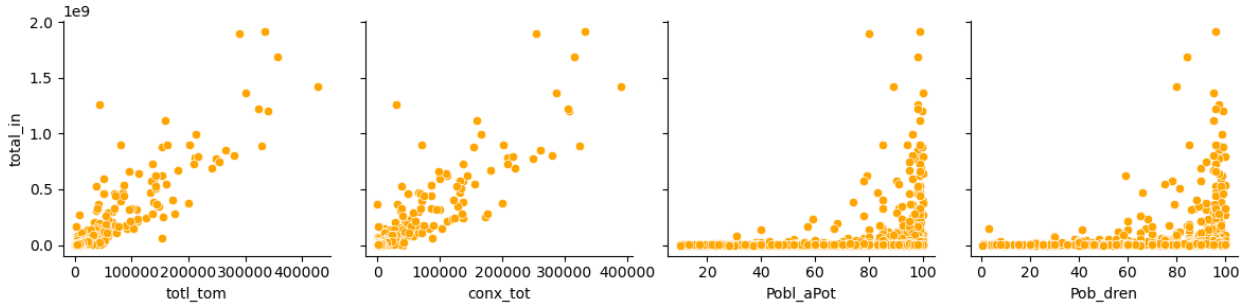


3. Análisis de correlaciones

Con el objetivo de cuantificar el grado de correlación existente entre las cinco variables de estudio se llevó a cabo el cálculo de la matriz de correlación

	totl_tom	conx_tot	Pob_dren	Pobl_aPot	total_in
totl_tom	1.000	0.986	0.139	0.154	0.899
conx_tot	0.986	1.000	0.152	0.149	0.892
Pob_dren	0.139	0.152	1.000	0.361	0.124
Pobl_aPot	0.154	0.149	0.361	1.000	0.124
total_in	0.899	0.892	0.124	0.124	1.000





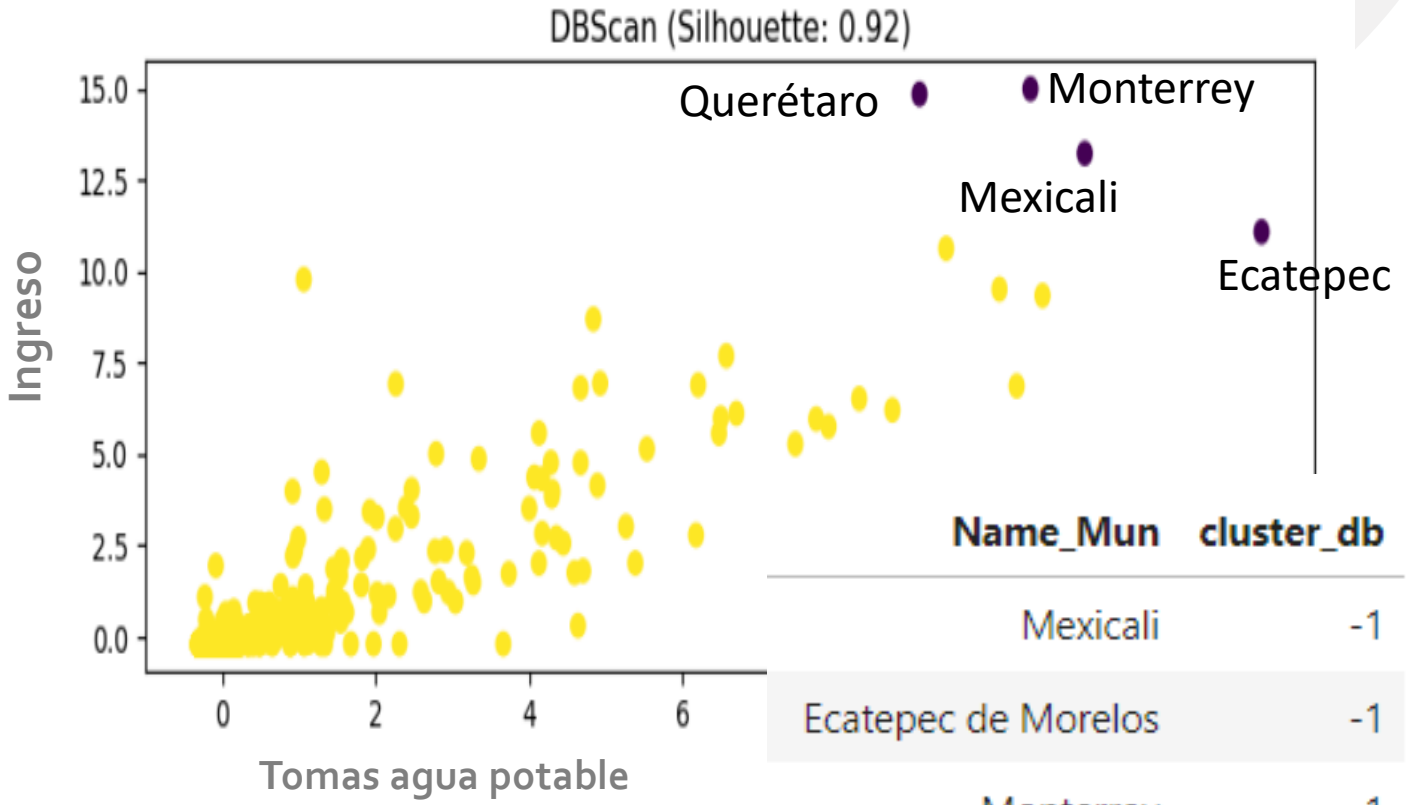
4. Modelos de segmentación

Con el objetivo de disminuir la diferencia en los valores de los variables, se realizó **la estandarización de estas utilizando el StandarScaler de la librería de sklearn**, posteriormente se implementaron los algoritmos de clasificación **K-Means, Hierarchical y DBScan** para segmentar a los prestadores de los servicios de agua potable y alcantarillado.

4.1 Segmentación con DBScan

La implementación del algoritmo DBScan se llevó a cabo utilizando la herramienta GridSearch para determinar los mejores parámetros del modelo. En este sentido se determinó que los mejores parámetros son:

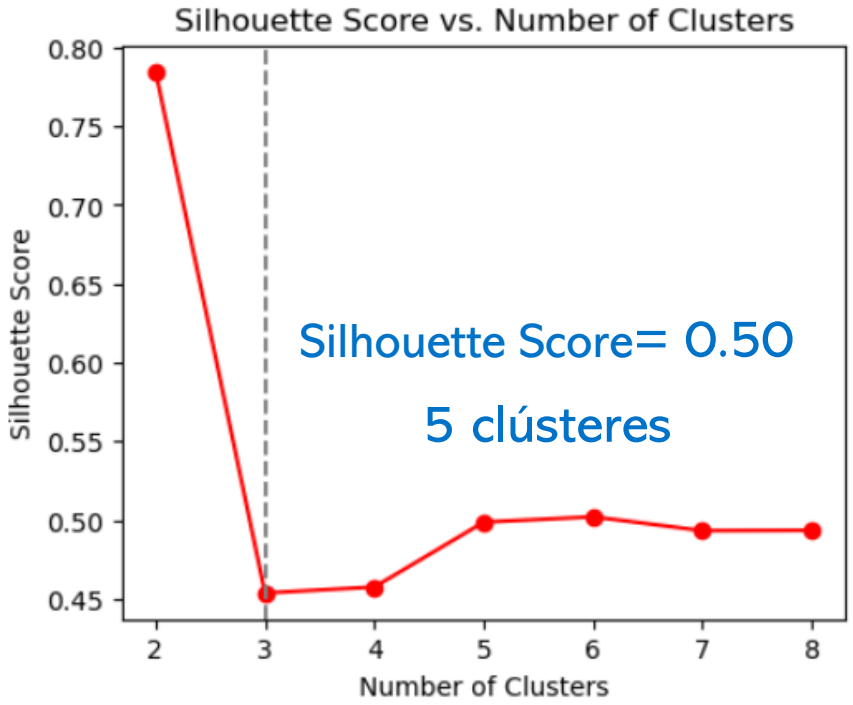
- **eps= 7**
- **min_samples=40**
- **Clúster= 1**



Name_Mun	cluster_db
Mexicali	-1
Ecatepec de Morelos	-1
Monterrey	-1
Querétaro	-1

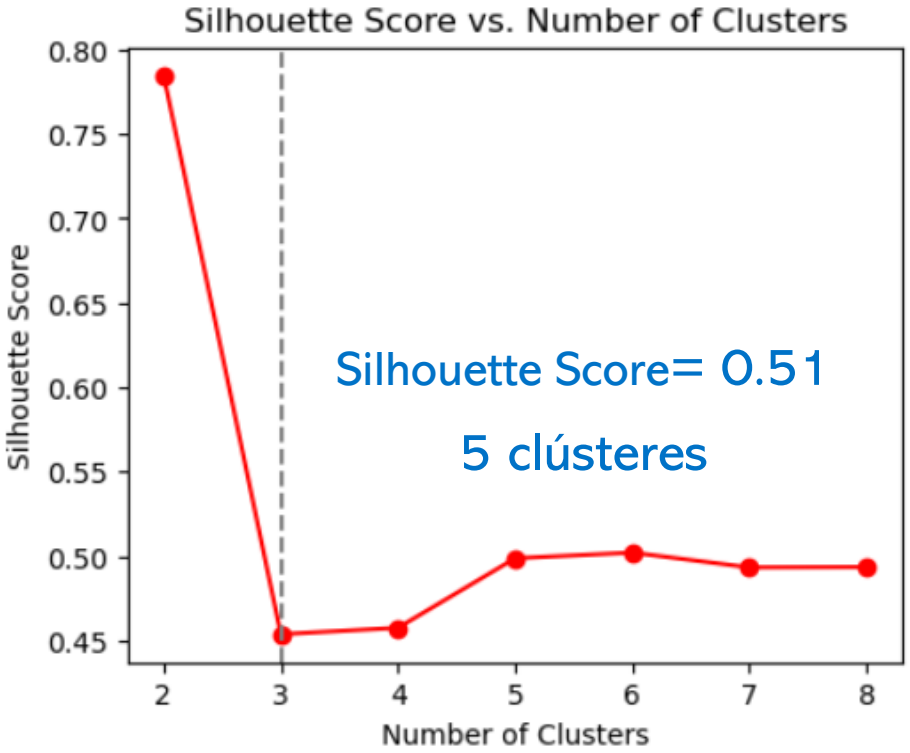
4.2 Segmentación con DBScan

La implementación del algoritmo se llevó a cabo con la librería **hierarchy de scipy** y se evaluaron utilizando la métrica **Silhouette Score** en función del número de clústeres.



4.3 Segmentación con KMeans

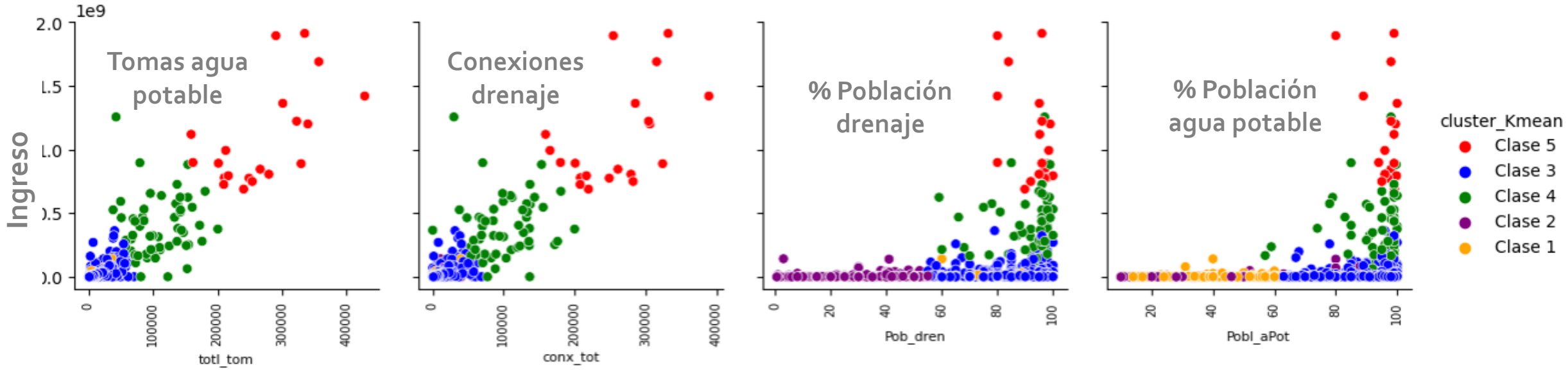
La implementación del algoritmo se llevó a cabo con la librería **KMeans de sklearn** y se evaluaron utilizando la métrica **Silhouette Score** en función del número de clústeres.



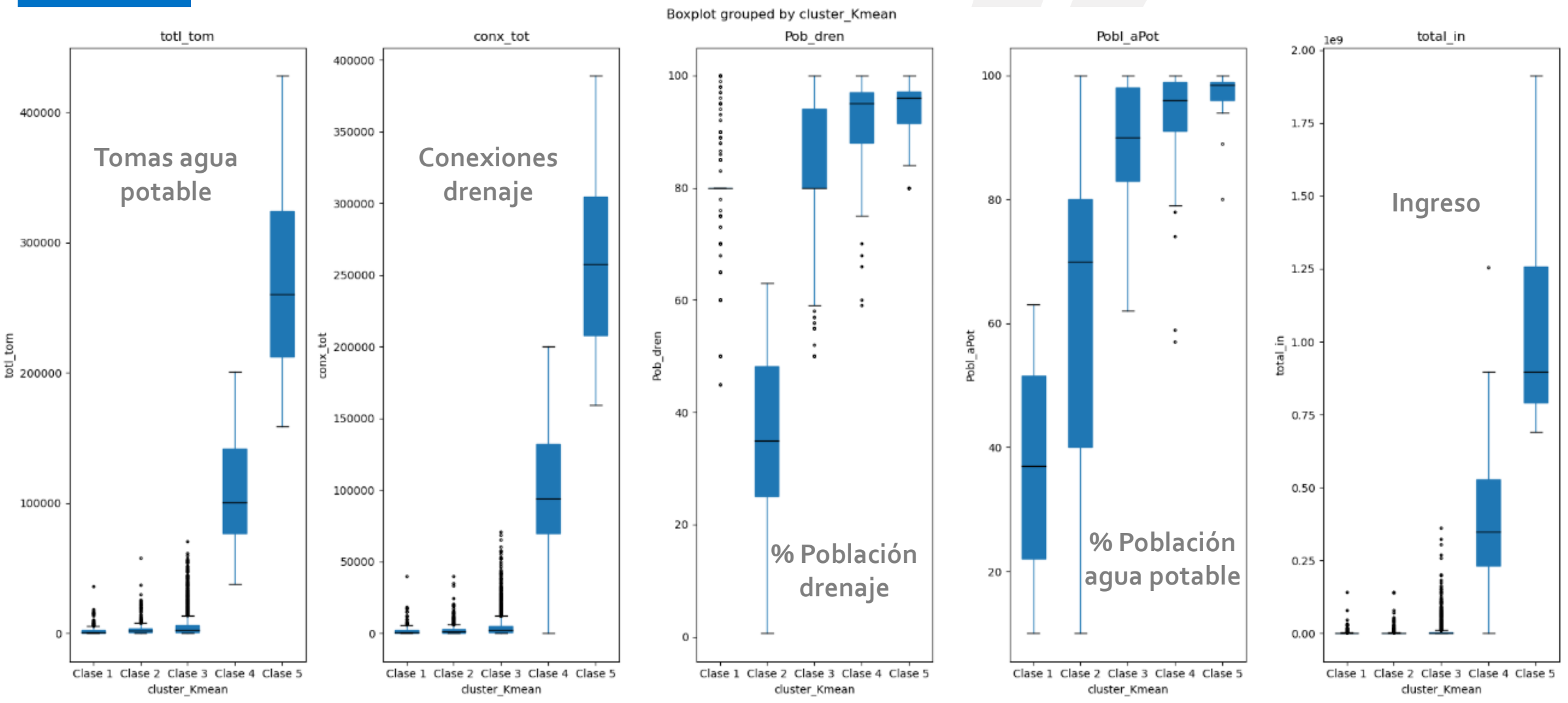
En la figura se observa que un número adecuado de clúster es 3. No obstante lo anterior, en este trabajo se decidió realizar la segmentación de los OAPAS en cinco grupos para una mejor caracterización

Se decidió realizar la segmentación de los OAPAS empleando KMeans

4.4 Segmentación OAPAS con KMeans



Clúster	Descripción	Núm. de OAPAS	Tomas de agua potable	Conexiones drenaje	% Pob. servida agua potable	% Pob. servida drenaje	Ingresos totales pesos
Clase 1	Muy baja cobertura de agua potable y alcantarillado	347	30 a 36,269 Prom. 2,125	38 a 39,983 Prom. 2,177	10% a 63% Prom. 36%	45% a 100% Prom. 80%	$321 a $141,716,159 Prom. $2,654,816 pesos
Clase 2	Baja cobertura de agua potable y alcantarillado	399	79 a 58,183 Prom. 3,653	11a 40,000 Prom. 2,656	10% a 80% Prom. 62%	0.6% a 63% Prom. 34%	3,390 a $142,660,111 pesos Prom= $4,052,319 pesos
Clase 3	Media cobertura de agua potable y alcantarillado	1,571	4 a 70,645 Prom. 5,910	4 a 70,645 Prom. 5,419	62% a 100% Prom. 90%	50% a 100% Prom. 84%	$1,000 a $363,176,392 Prom. $8,955,772
Clase 4	Alta cobertura de agua potable y alcantarillado	57	38,063 a 200,470 Prom.108,525	74 a 200,040 Prom. 99,790	57% a 100% Prom. 93%	59% a 100% Prom. 91%	$1,271,509 a $1,256,686,180 Prom. $391,604,8256
Clase 5	Muy alta cobertura de agua potable y alcantarillado	20	159,115 a 428,144 Prom. 268,229	159,115 a 389,221 Prom. 257,066	80% a 100% Prom. 97%	80% a 100% Prom. 93%	$689,929,731 a $1,913,074,129 Prom. $1,083,648,155
OAPAS	Mega-ciudades Tijuana,Chihuahua,León Guadalajara	5	de 469,797 a 1,170,135 Prom. 695,939	de 326,729 a 1,164,305 Prom. 609,597	89.6% a 100% Prom. 96%	91% a 100% Prom. 96%	$1,326,602,408 a $4,559,694,557 Prom. $2,588,810,855

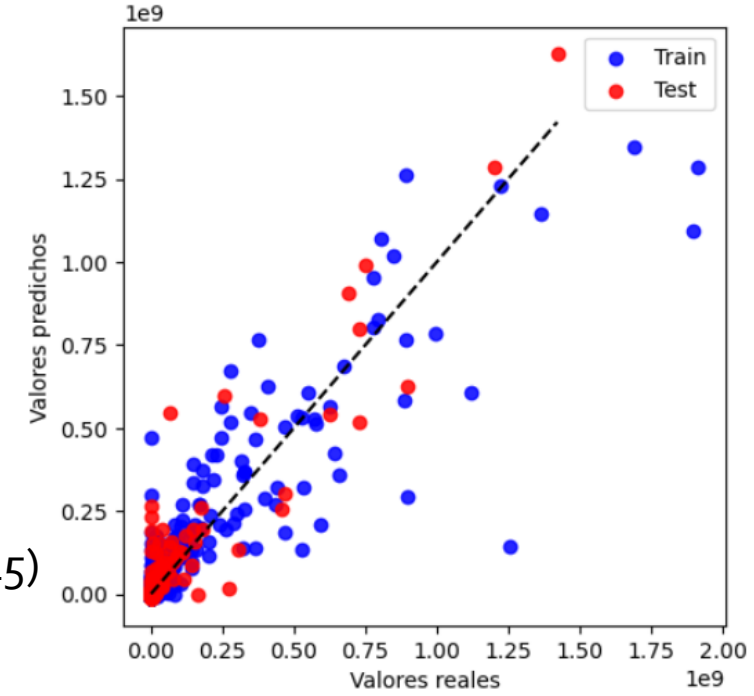


Clúster	Descripción	Núm. de OAPAS
Clase 1	Muy baja cobertura de agua potable y alcantarillado	347
Clase 2	Baja cobertura de agua potable y alcantarillado	399
Clase 3	Media cobertura de agua potable y alcantarillado	1,571

Clúster	Descripción	Núm. de OAPAS
Clase 4	Alta cobertura de agua potable y alcantarillado	57
Clase 5	Muy alta cobertura de agua potable y alcantarillado	20

5. Modelo de regresión estimar Ingreso

En este apartado se muestran los resultados del modelo matemático calculado utilizando la biblioteca de sklearn para estimar el ingreso de los OAPAS en función de la cantidad de tomas y conexiones de agua potable, así como el porcentaje de población servida de agua potable y drenaje.



- **Estandarización** de las variables con StandarScaler de la librería de sklearn.
- Entrenamiento Modelo con **datos entrenamiento** 80% de las observaciones (1915)
- Evaluación Modelo con con **datos prueba** el 20% como datos de prueba (479)

$$\boldsymbol{I} = 3{,}303.4\,\boldsymbol{totl_{tom}} + 573.7\,\boldsymbol{conx_{tot}} - 91{,}795.6\,\boldsymbol{Pbl_{aPot}} + 859.7\,\boldsymbol{Pob_{dren}} - 3{,}670{,}323$$

Donde I es el ingreso total municipal de un Organismo operadore de agua en pesos, $totl_{tom}$ es el número total de tomas de agua potable, $conx_{tot}$ es el número de total de conexiones de drenaje, Pbl_{aPot} es el porcentaje de la población con servicio de agua potable y Pob_{dren} es el porcentaje de la población con servicio de drenaje.

R^2 del modelo ajustado en los datos de prueba: 0.84 lo que indica un buen ajuste.

Error absoluto medio (MAE) = 19,742,077 pesos

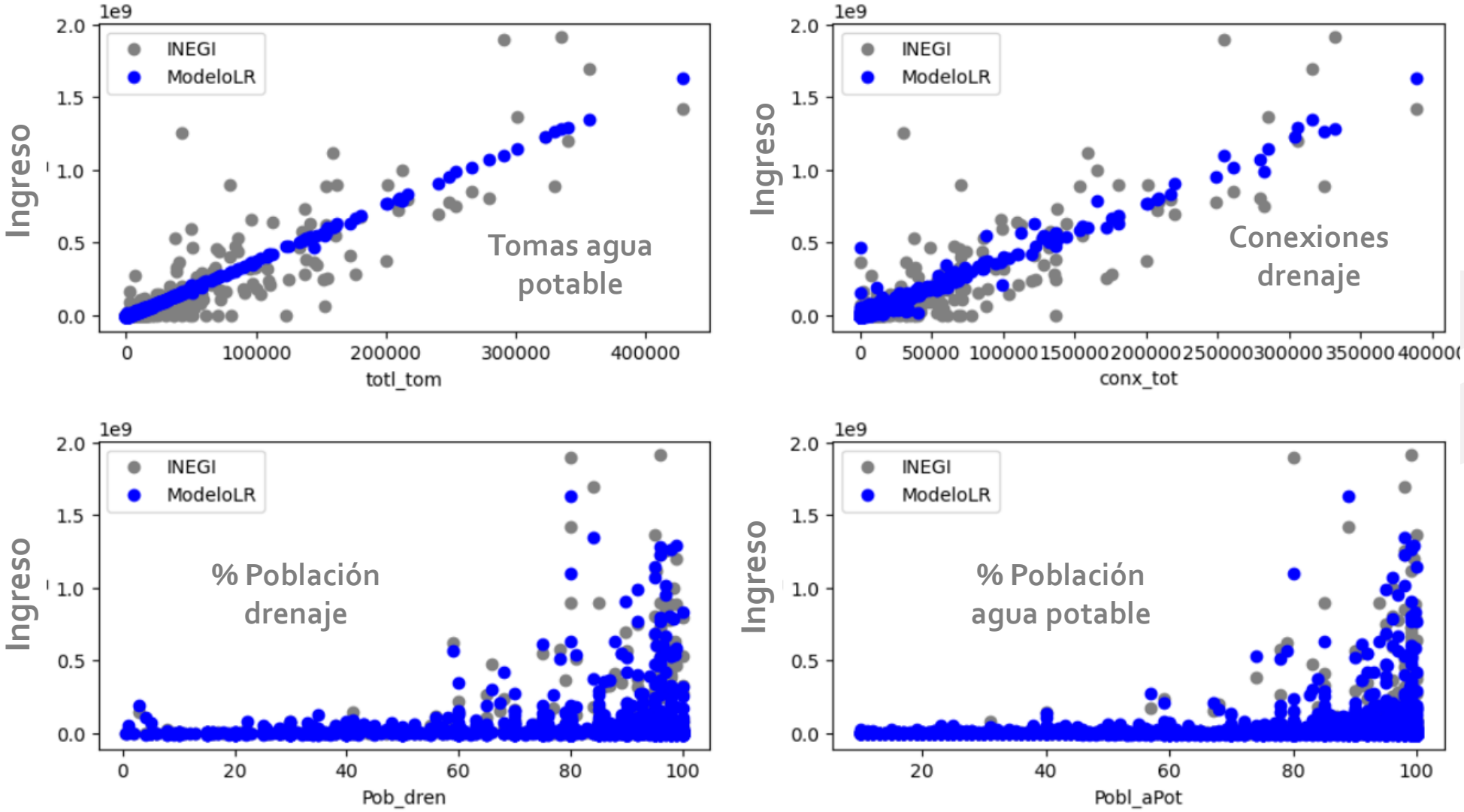
Raíz del error cuadrático medio (RMSE) = 50,203,927 pesos

Contoso

$$I = 3{,}303.4\ totl_{tom} + 573.7\ conx_{tot} - 91{,}795.6\ Pbl_{aPot} + 859.7\ Pob_{dren} - 3{,}670{,}323$$

Donde I es el ingreso total municipal de un Organismo operadore de agua en pesos, $totl_{tom}$ es el número total de tomas de agua potable, $conx_{tot}$ es el número de total de conexiones de drenaje, Pbl_{aPot} es el porcentaje de la población con servicio de agua potable y Pob_{dren} es el porcentaje de la población con servicio de drenaje.

Name_Mun_x	folio	Estimación Ingreso		total_in	totl_tom	conx_tot	Pobl_aPot	Pob_dren
Aguascalientes	1001	$ 1,261,913,007	42%	$ 889,622,823	329,552	324,115	99	98
Mexicali	2002	$ 1,347,678,870	-20%	$ 1,690,761,382	356,964	315,632	98	84
Saltillo	5030	$ 1,016,420,711	20%	$ 846,102,145	266,156	261,084	98	97
Torreón	5035	$ 801,942,547	3%	$ 779,119,591	210,439	208,214	99	97.8
Ecatepec de Morelos	15033	$ 1,625,855,613	14%	$ 1,420,984,395	428,144	389,221	89	80
Naucalpan de Juárez	15057	$ 604,228,522	-46%	$ 1,120,061,885	159,115	159,115	99	95.21
Nezahualcóyotl	15058	$ 827,429,761	4%	$ 795,183,597	216,702	216,702	99.8	100
Toluca	15106	$ 784,932,090	-21%	$ 993,679,907	212,640	165,409	96	98.5
Pachuca de Soto	13048	$ 907,542,835	32%	$ 689,929,732	240,447	219,526	99.12	89.87
Morelia	16053	$ 952,395,009	23%	$ 776,019,641	248,868	248,868	97	97
Apodaca	19006	$ 799,295,568	10%	$ 726,300,455	209,704	207,835	99	96
Guadalupe	19026	$ 767,601,155	-14%	$ 893,093,811	201,316	200,888	99	96
Monterrey	19039	$ 1,284,674,815	-33%	$ 1,913,074,129	335,097	331,865	99	96
Querétaro	22014	$ 1,094,567,389	-42%	$ 1,895,036,863	290,449	254,564	80	80
Benito Juárez	23005	$ 1,146,025,441	-16%	$ 1,363,658,394	301,185	285,618	100	95.1
San Luis Potosí	24028	$ 1,071,356,971	33%	$ 806,832,385	279,528	279,528	96	95
Culiacán	25006	$ 1,285,962,167	7%	$ 1,201,402,177	339,960	306,183	99.5	99
Hermosillo	26030	$ 1,227,809,852	0%	$ 1,223,248,221	322,638	304,325	98	96
Matamoros	28022	$ 625,799,895	-30%	$ 898,915,604	161,859	180,138	94	80
Reynosa	28032	$ 988,212,848	32%	$ 749,935,955	253,819	282,480	95	92



Los resultados del MAE y el RMSE indican que el modelo tiene un error mínimo para estimar el ingreso en los Clase 5, en donde los ingresos tienen un rango de $689,929,731 a $1,913,074,129 pesos. Sin embargo, es recomendable revisar sus estimaciones con mayor detalle para verificar su correcta implementación.

6. Conclusiones

Los resultados de la segmentación de los OAPAS realizada con el método de aprendizaje no supervisado KMeans sugieren que la segmentación con cinco clases permite determinar con buenos resultados algunas de las principales características de estos prestadores de servicios en México. Sin embargo, existen factores importantes como **datos con errores presentes en el Censo de INEGI que dificultaron el obtener una segmentación con mayor precisión**. Por otro lado, en este trabajo sólo se consideraron para la segmentación el número de tomas de agua potable, el número de conexiones de drenaje, el porcentaje de la población servida con agua potable y alcantarillado, así como el ingreso de los OAPAS, factores que aunque son importantes, se recomienda realizar otros análisis considerando factores como la cantidad de empleados del organismo operador, el número de habitantes, los ingresos de los diferentes rubros, entre otras variables que están presentes en el censo de INEGI. **Reto la imputación de valores.**

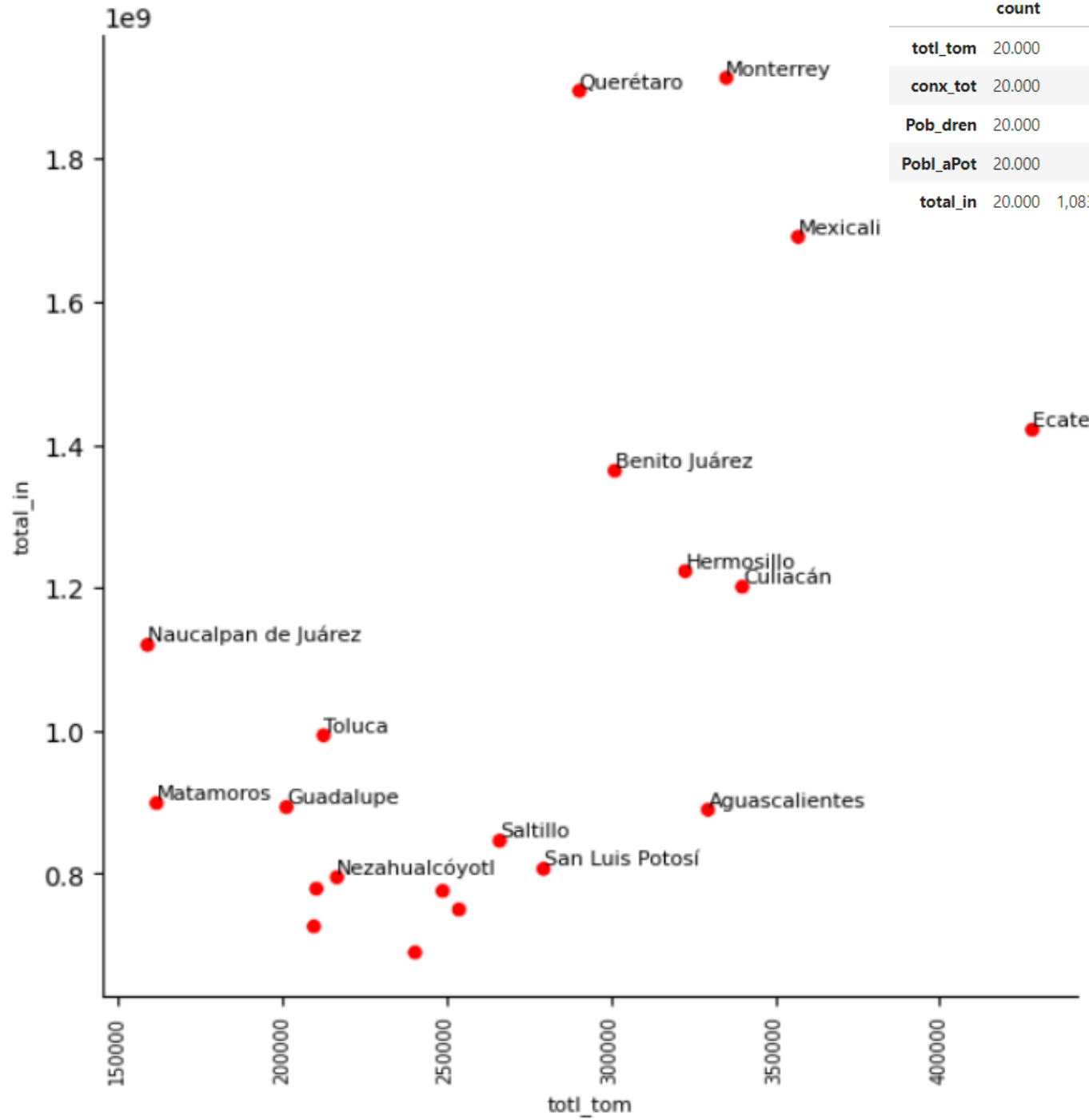
Otra conclusión importante en este trabajo es que se logró determinar un modelo matemático para estimar el ingreso de los OAPAS en México en función del número de tomas de agua potable, el número de conexiones de drenaje, así como del porcentaje de la población servida con agua potable y alcantarillado con un coeficiente de determinación R2 de 0.84 calculada en los datos de prueba, lo que indica que el modelo **presenta un buen ajuste**. Sin embargo, **debido a la gran dispersión presente en las observaciones, el modelo podría se útil para estimar el ingreso sólo en los OAPAS de los Clase 5,** en los que los ingresos tienen un rango de $689,929,731 a $1,913,074,129 pesos.

Les agradeceré sus comentarios

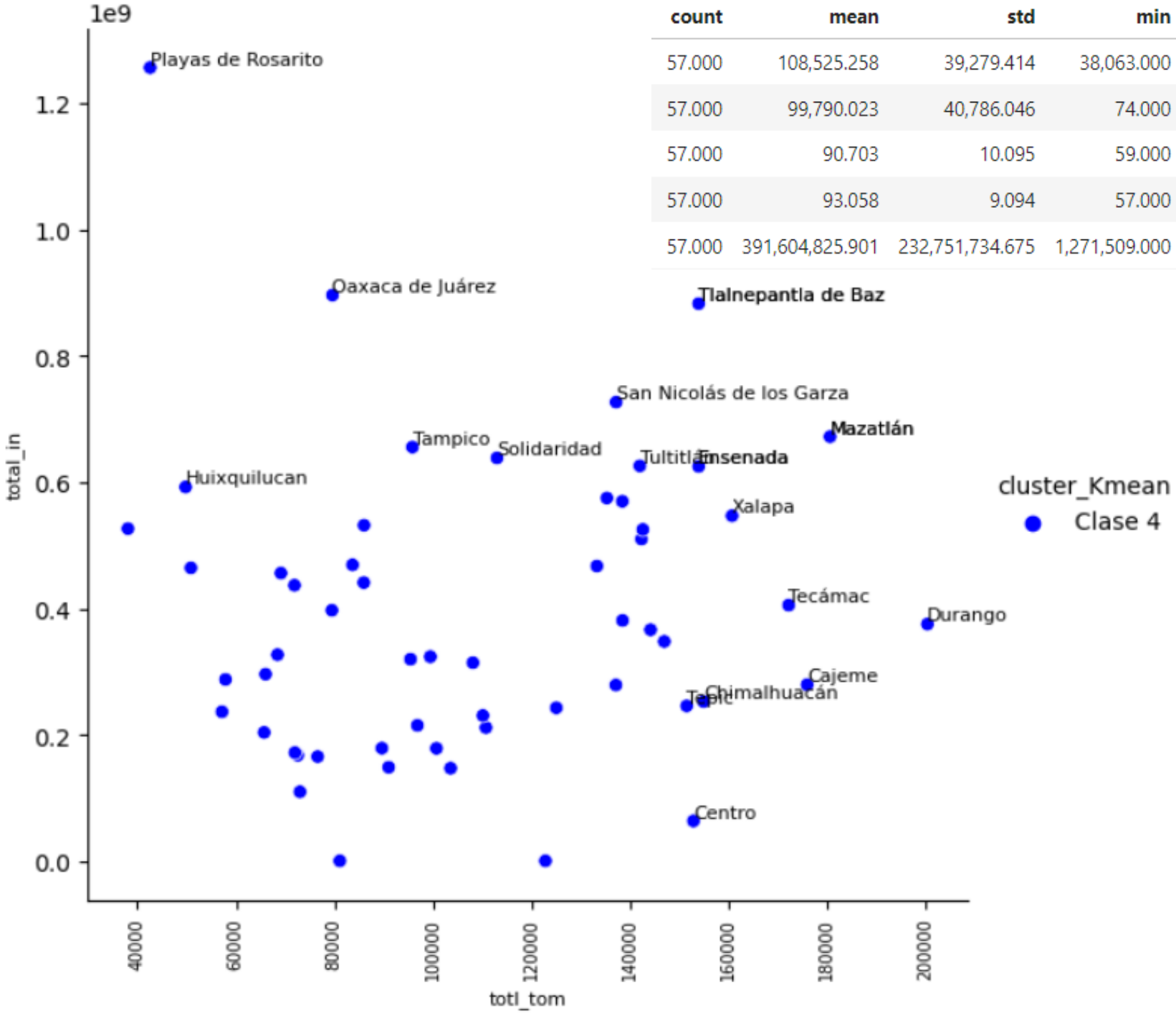
Gracias

jevabrir@gmail.com

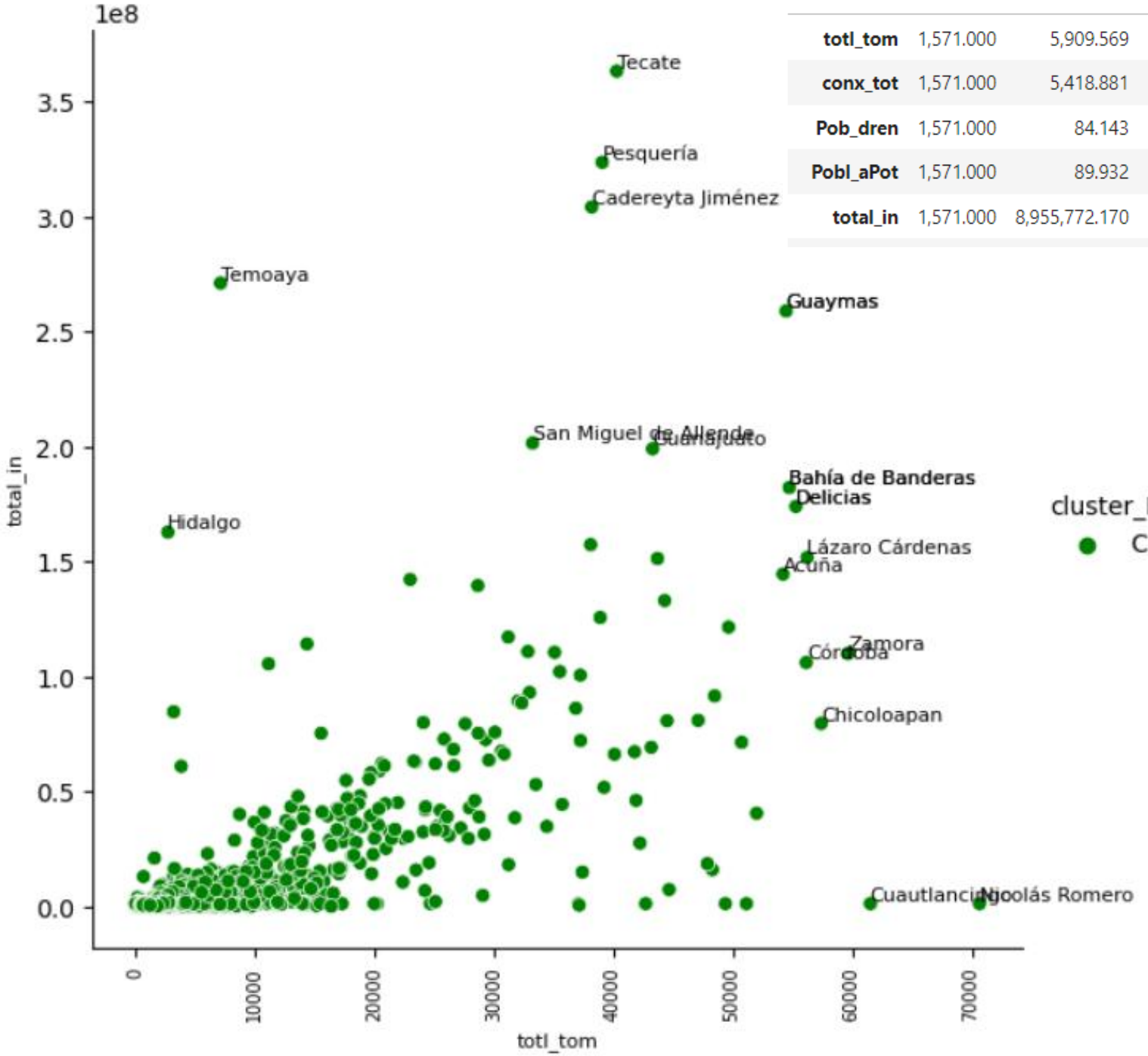
	count	mean	std	min	25%	50%	75%	max
totl_tom	20.000	268,229.100	69,984.273	159,115.000	212,089.750	259,987.500	324,366.500	428,144.000
conx_tot	20.000	257,065.522	61,816.588	159,115.000	208,119.250	257,824.000	304,789.500	389,221.000
Pob_dren	20.000	93.124	6.661	80.000	91.468	96.000	97.200	100.000
Pobl_aPot	20.000	96.671	4.704	80.000	96.000	98.500	99.000	100.000
total_in	20.000	1,083,648,154.624	387,682,696.940	689,929,731.810	791,167,595.438	896,004,707.675	1,258,350,764.348	1,913,074,128.970

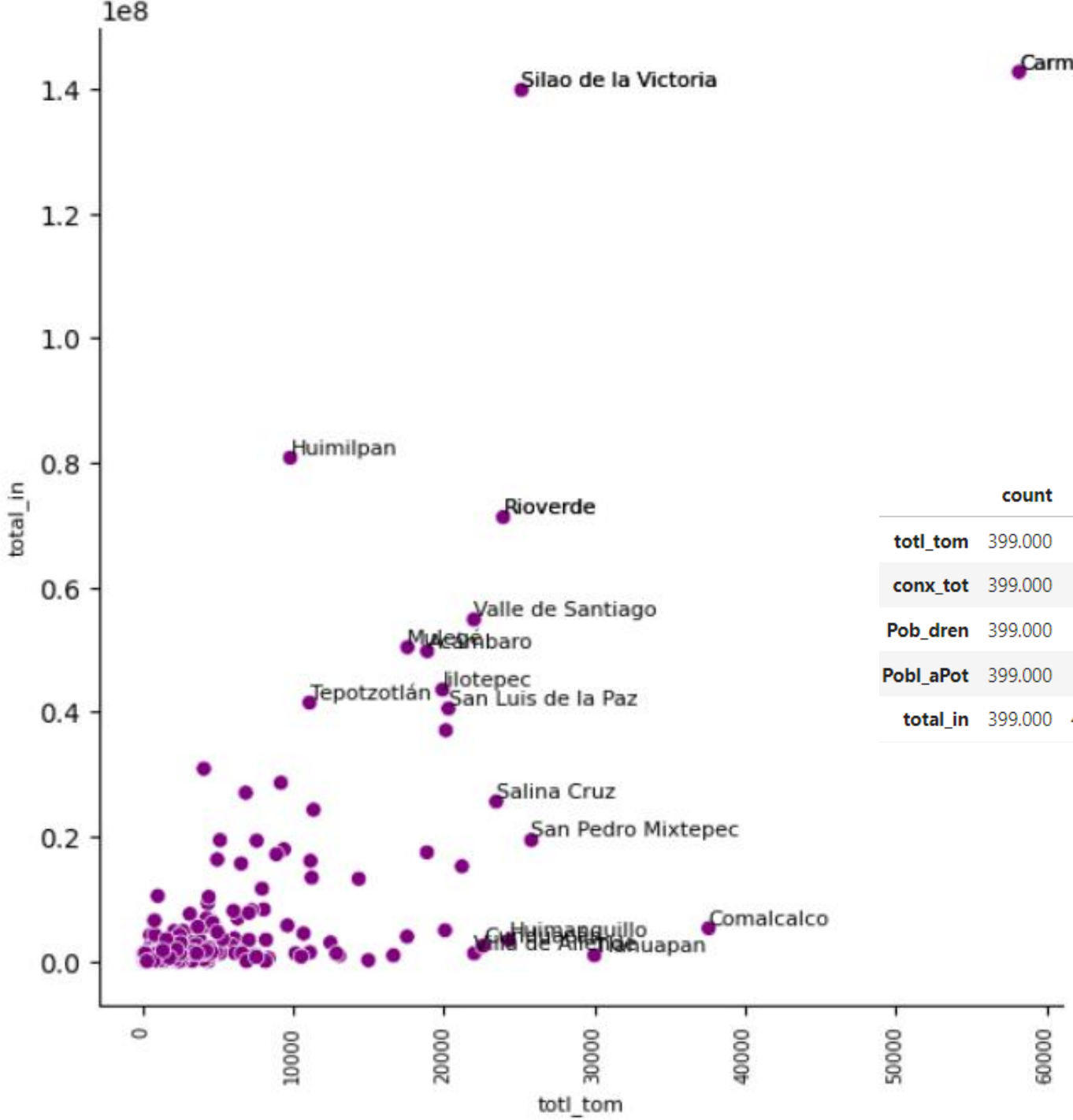


	count	mean	std	min	25%	50%	75%	max
	57.000	108,525.258	39,279.414	38,063.000	76,602.000	100,736.000	142,048.000	200,470.000
	57.000	99,790.023	40,786.046	74.000	69,862.000	94,363.000	132,397.000	200,040.000
	57.000	90.703	10.095	59.000	88.000	95.000	97.000	100.000
	57.000	93.058	9.094	57.000	91.000	96.000	99.000	100.000
	57.000	391,604,825.901	232,751,734.675	1,271,509.000	231,225,911.000	348,488,763.900	527,094,536.390	1,256,686,180.000

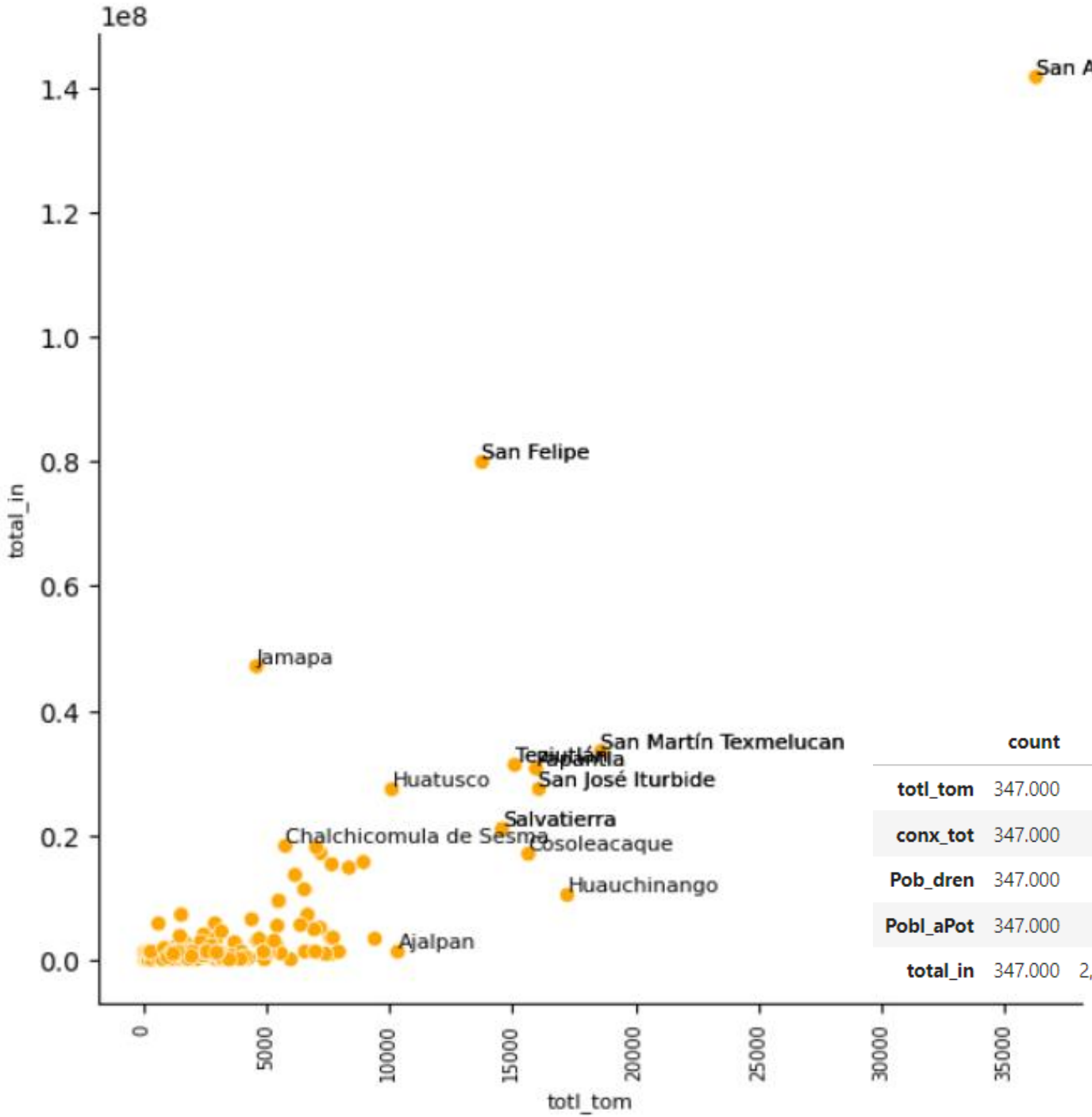


	count	mean	std	min	25%	50%	75%	max
totl_tom	1,571.000	5,909.569	9,224.634	4.000	910.500	2,516.000	6,099.000	70,645.000
conx_tot	1,571.000	5,418.881	8,835.853	4.000	797.680	2,213.000	5,362.500	70,645.000
Pob_dren	1,571.000	84.143	10.482	50.000	80.000	80.000	94.000	100.000
Pobl_aPot	1,571.000	89.932	9.072	62.000	83.000	90.000	98.000	100.000
total_in	1,571.000	8,955,772.170	26,852,362.221	1,000.000	481,000.000	1,271,509.000	4,587,818.585	363,176,392.800





	count	mean	std	min	25%	50%	75%	max
totl_tom	399.000	3,653.638	5,906.286	79.000	690.000	1,762.000	3,679.000	58,183.000
conx_tot	399.000	2,655.520	4,617.663	11.000	400.000	1,177.000	2,686.500	40,000.000
Pob_dren	399.000	34.827	15.286	0.600	25.000	35.000	48.215	63.000
Pobl_aPot	399.000	61.986	26.547	10.000	40.000	70.000	80.000	100.000
total_in	399.000	4,052,318.967	13,136,540.625	3,390.000	315,455.035	1,271,509.000	1,584,605.560	142,660,111.250



	count	mean	std	min	25%	50%	75%	max
totl_tom	347.000	2,125.500	3,492.313	30.000	343.500	850.000	2,475.000	36,269.000
conx_tot	347.000	2,177.743	3,683.194	38.587	321.820	800.000	2,476.927	39,983.000
Pob_dren	347.000	79.791	9.679	45.000	80.000	80.000	80.000	100.000
Pobl_aPot	347.000	36.376	16.304	10.000	22.000	37.000	51.500	63.000
total_in	347.000	2,654,815.858	9,987,260.579	321.000	93,400.000	1,271,509.000	1,271,509.000	141,761,159.290