# Análisis exploratorio de datos del suministro de agua de la CDMX

Análisis exploratorio de datos del suministro de agua de la Ciudad de México a octubre de 2020. Este análisis fue efectuado empleando la información de los primeros 3 bimestres del año 2019 por el concepto de suministro de agua a nivel manzana en metros cúbicos, considerando la facturación por servicio de consumo medido y promedio otorgada por el Gobierno de la Ciudad de México y disponible en el portal de Datos Abiertos Ciudad de México.

Este análisis fue efecutado por el equipo de Ciencia de Datos conformado por:


- Carlos Bautista (125761)
- Enrique Ortiz (150644)
- Mario Heredia (197863)
- José Antonio Lechuga (192610)


## Contenido

En el presente reporte se abordan las siguientes preguntas y actividades:

- Exploración Inicial
- Limpieza y manipulación de datos
- Detección de errores en los datos
- Estudio de datos faltantes
- Comportamiento del consumo bimestral
- Estudio de alcadías y colonias
- Estudio de consumo total
- Estudio de consumo promedio
- Características de los datos para la definición de un modelo

# Exploración inicial

Con el objetivo de relizar una exploración inicial, se contestaron las siguientes preguntas:

- ¿Cuántas variables existen?
- ¿Cuántas observaciones existen?
- ¿Cuántas observaciones únicas se tienen por variables?
- ¿Cuántas variables numéricas existen?
- ¿Cuántas variables de fecha tenemos?
- ¿Cuántas variables categóricas existen?
- ¿Cuántas variables de texto hay?
- ¿Qué se conoce hasta ahora de este set de datos por variable?
- ¿Cuántas alcaldías hay? ¿Cuántos nomgeo existen? ¿Se identifica algún error?

## ¿Cuántas variables existen? ¿Cuántas observaciones existen?

```
Tenemos 71102 observaciones para 17 variables.
```

## ¿Cuántas observaciones únicas se tienen por variables?

| Variable | Observaciones unicas |
|---|---|
| Geo Point | 22,930 |
| Geo Shape | 22,922 |
| consumo_total_mixto | 24,339 |
| anio | 1 |
| nomgeo | 17 |
| consumo_prom_dom | 52,060 |
| consumo_total_dom | 47,051 |
| alcaldia | 16 |
| colonia | 1,340 |
| consumo_prom_mixto | 31,911 |
| consumo_total | 56,015 |
| consumo_prom | 62,214 |
| consumo_prom_no_dom | 37,440 |
| bimestre | 3 |
| consumo_total_no_dom | 27,336 |
| gid | 71,102 |
| indice_des | 4 |

## ¿Cuántas variables numéricas existen?

```
Geo Point                object
Geo Shape                object
consumo_total_mixto     float64
anio                      int64
nomgeo                   object
consumo_prom_dom        float64
consumo_total_dom       float64
alcaldia                 object
colonia                  object
consumo_prom_mixto      float64
consumo_total           float64
consumo_prom            float64
consumo_prom_no_dom     float64
bimestre                  int64
consumo_total_no_dom    float64
gid                       int64
indice_des               object
dtype: object

Tenemos 11 variables numéricas.
```

## ¿Cuántas variables de fecha tenemos?

Ninguna. Se puede ver en la impresión de dtypes. La variable `anio` indica el año, pero eso no es suficiente para ser una fecha (no califica como date, time o datetime).

```
Tenemos 0 variables de fecha.
```

## ¿Cuántas variables categóricas existen?

Las siguientes pueden ser consideradas variables categóricas: `nomgeo`, `alcaldia`, `colonia`, e `indice_des`. `bimestre` fue contada como numérica, pero es cierto que puede ser analizada como categórica, al igual que año (dado que solo hay cifras para 2019). Por lo tanto, tenemos entre 4 y 6 variables categóricas, dependiendo de si tomamos a bimestre y año como categórica o no.

## ¿Cuántas variables de texto hay?

Aunque `nomgeo`, `alcaldia`, `colonia` e `indice_des` son texto, no contienen como tal texto a analizar (como palabras u oraciones). Diríamos que son categóricas.

## ¿Qué se conoce hasta ahora de este set de datos por variable?

- `Geo Point` indica la latitud y longitud de la manzana
- `Geo Shape` es un diccionario con información geográfica
- `nomgeo` y `alcaldia` son equivalentes (e indican el nombre de la alcaldía). Hay 16 distintos
- `colonia` indica el nombre de la colonia
- `anio` indica el año, y solo hay datos para 2019
- `bimestre` indica el bimestre del año, y solo hay datos para el bimestre 1, 2 y 3

## ¿Cuántas alcaldías hay? ¿Cuántos `nomgeo` existen? ¿Se identifica algún error?

```
Se tiene 16 alcaldías únicas y 17 nomgeo únicos.
```

Sí el error está en la variable nomgeo: se registró "Talpan" en vez de "Tlalpan".

```
['Talpan', 'Tlalpan']
```

# Limpieza y manipulación de datos

Para realizar una limpieza de los datos, se realizaron las siguientes actividades:

- Transformación del nombre de las columnas a formato estándar: minúsculas, sin espacios en blanco, sin signos de puntuación ni acentos
- Agregar las variables `latitud` y `longitud` generadas a partir de la columna `geo_point`
- Transformar la variables `latitud` y `longitud` a numéricas
- Eliminar la columna `geo_point`
- Eliminar la columna `geo_shape`
- Cambiar a minúsculas las columnas `alcaldia`, `colonia`, `indice_des` y `nomgeo`
- Sustituir los valores de `nomgeo` donde se escribió "Talpan" en lugar de "Tlalpan"
- Homogenizar los valores de `nomgeo` y `alcaldia` sustituyendo en `nomgeo` las entradas con valor de "La Magdalena Contreras" y "Cuajimalpa de Morelos" por "Magdalena Contreras" y "Cuajimalpa" respectivamente.

## Identificación de variables numéricas, categóricas, texto y fechas

- **Variables numéricas:**
  - `consumo_total_mixto`
  - `anio` (de nuevo, que puede ser contada como categórica)
  - `consumo_promo_dom`
  - `consumo_total_dom`
  - `consumo_prom_mixto`
  - `consumo_total`
  - `consumo_prom`
  - `consumo_prom_no_dom`
  - `consumo_total_no_dom`
  - `gid`
  - `bimestre` (una vez mas) puede ser o no contada como numérica
- **Variables de fecha:** ninguna.
- **Variables de texto:**
  - `latitud` y `longitud` son texto para pandas, aunque en realidad son coordenadas
  - No hay variables que contengan texto a analizar
- **Variables categóricas:**
  - `nomgeo`
  - `alcaldia`
  - `colonia`
  - `indices_des`
  - `bimestre`

```
Tenemos 13 variables numéricas.
Tenemos 0 variables de fecha.
```

Un profiling detallado de todas las variablaes, se incluye en conjunto con este reporte con el nombre **profiling_cliente.html**.

# Detección de errores en los datos

Con el objetivo de detectar posibles en los datos, se contestaron las siguientes preguntas:

- ¿Existen consumos negativos?
- ¿Existen valores igual a cero para el consumo?
- ¿Existen observaciones donde la suma de los consumos no equivalga al consumo total?
- ¿Existen observaciones donde el total de tomas de agua no sea la suma de los 3 tipos de tomas?
- ¿En cuántas observaciones no coinciden los valores de `alcaldia` y `nomgeo` ?

## ¿Existen consumos negativos?

| Tipo de consumo | Porcentaje de consumos negativos |
|---|---|
| Consumo total mixto | 0 |
| Consumo promedio mixto | 0 |
| Consumo total doméstico | 0 |
| Consumo promedio doméstico | 0 |
| Consumo total no doméstico | 0 |
| Consumo promedio no doméstico | 0 |
| Consumo total | 0 |
| Consumo total promedio | 0 |

## ¿Existen valores igual a cero para el consumo?

| Tipo de consumo | Porcentaje de consumos igual a cero |
|---|---|
| Consumo total mixto | 24.9 |
| Consumo promedio mixto | 24.9 |
| Consumo total doméstico | 13.9 |
| Consumo promedio doméstico | 13.9 |
| Consumo total no doméstico | 11.4 |
| Consumo promedio no doméstico | 11.4 |
| Consumo total | 3.4 |
| Consumo total promedio | 3.4 |

Es posible verificar que para cada tipo de consumo, el número de observaciones faltantes es el mismo tanto para el dato de total como para el promedio.

## ¿Existen observaciones donde la suma de los consumos no equivalga al consumo total?

Debido a que el consumo total (i.e. `consumo_total`) se calcula de la siguiente manera:

$$consumo\_total = consumo\_total\_domestico + consumo\_total\_no\_domestico + co$$

se comprobó el número de observaciones que divergían de dicho cálculo obteniendo los siguientes resultados:

### ¿Cúantos valores divergen al nivel de unidades?

```
0
```

### ¿Cúantos valores divergen al nivel de décimas?

```
6
```

## ¿Cúantos valores divergen al nivel de centésimas?

```
149
```

## ¿Existen observaciones donde el total de tomas de agua no sea la suma de los 3 tipos de tomas?

Considerando a los consumos promedios como:

$$consumo\_promedio = \frac{consumo\_total}{no\_tomas}$$

Se comprobó que el número total de tomas equivalga a la suma de las tomas de agua domésticas, no domésticas y mixtas obteniendo los siguientes resultados:

## ¿En cuántas observaciones el número de tomas de agua no cuadra?

```
5650
```

## ¿Qué porcentaje representa?

```
9.1%
```

## ¿Cómo se distribuyen?

| Diferencia en tomas | No observaciones | Fracción |
|---|---|---|
| 0.0 | 56,564 | 90.9 |
| 1.0 | 4,742 | 7.6 |
| 2.0 | 707 | 1.1 |
| 3.0 | 139 | 0.2 |
| 4.0 | 34 | 0.1 |
| 5.0 | 16 | 0.0 |
| 6.0 | 1 | 0.0 |
| 7.0 | 6 | 0.0 |
| 8.0 | 3 | 0.0 |
| 16.0 | 2 | 0.0 |

### ¿En cuántas observaciones no coinciden los valores de `alcaldia` y `nomgeo`?

0

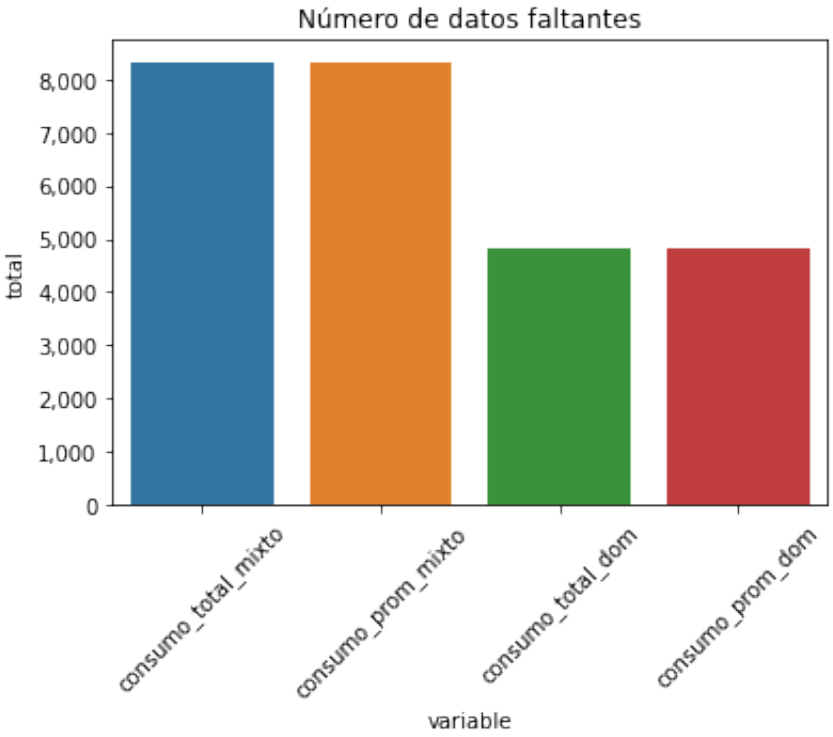
## Datos faltantes

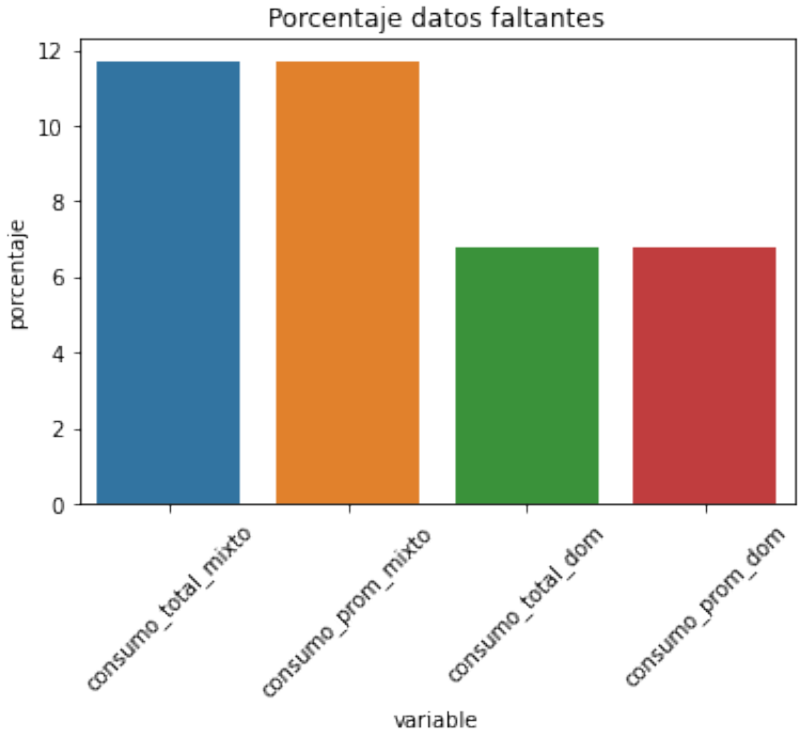
Con base en los resultados obtenidos del Data Profiling, generamos un cuadro que condense la información sobre las variables que contienen *missing values*. Podemos observar que estas variables se refieren a algún tipo de consumo, en específico las de consumo mixto y consumo doméstico:

| Concepto | Número de datos faltantes |
|---|---|
| consumo_total_mixto | 8,327 |
| anio | 0 |
| nomgeo | 0 |
| consumo_prom_dom | 4,820 |
| consumo_total_dom | 4,820 |
| alcaldia | 0 |
| colonia | 0 |
| consumo_prom_mixto | 8,327 |
| consumo_total | 0 |
| consumo_prom | 0 |
| consumo_prom_no_dom | 0 |
| bimestre | 0 |
| consumo_total_no_dom | 0 |
| gid | 0 |
| indice_des | 0 |
| latitud | 0 |
| longitud | 0 |

Dado que el resto de variables tienen información completa para el total de las observaciones, podríamos sospechar que los *missing values* se expliquen por manzanas donde no existe ese tipo de consumo particular.

| Variable | Total | Porcentaje |
|---|---|---|
| consumo_total_mixto | 8,327 | 11.7 |
| consumo_prom_mixto | 8,327 | 11.7 |
| consumo_total_dom | 4,820 | 6.8 |
| consumo_prom_dom | 4,820 | 6.8 |




Recordamos que el consumo total de una manzana se compone de la suma de las tomas de agua de tipo doméstico, no doméstico y mixto. Esto lo podemos comprobar, y si ambas sumas son iguales, podríamos afirmar que los missing values efectivamente corresponden a manzanas donde no existen tomas de dichas características.

```
120578129.0 120578129.0
```

Y efectivamente vemos que las sumas coinciden y que la diferencia es cero:

0.0

# Comportamiento del consumo bimestral

Con el objetivo de detectar posibles en los datos, se contestaron las siguientes preguntas:

- ¿De cuántos bimestres tenemos información?
- ¿Cúantas observaciones se tienen por bimestre?
- ¿Existen diferencias considerables entre el consumo por bimestre?

## ¿De cuántos bimestres tenemos información?

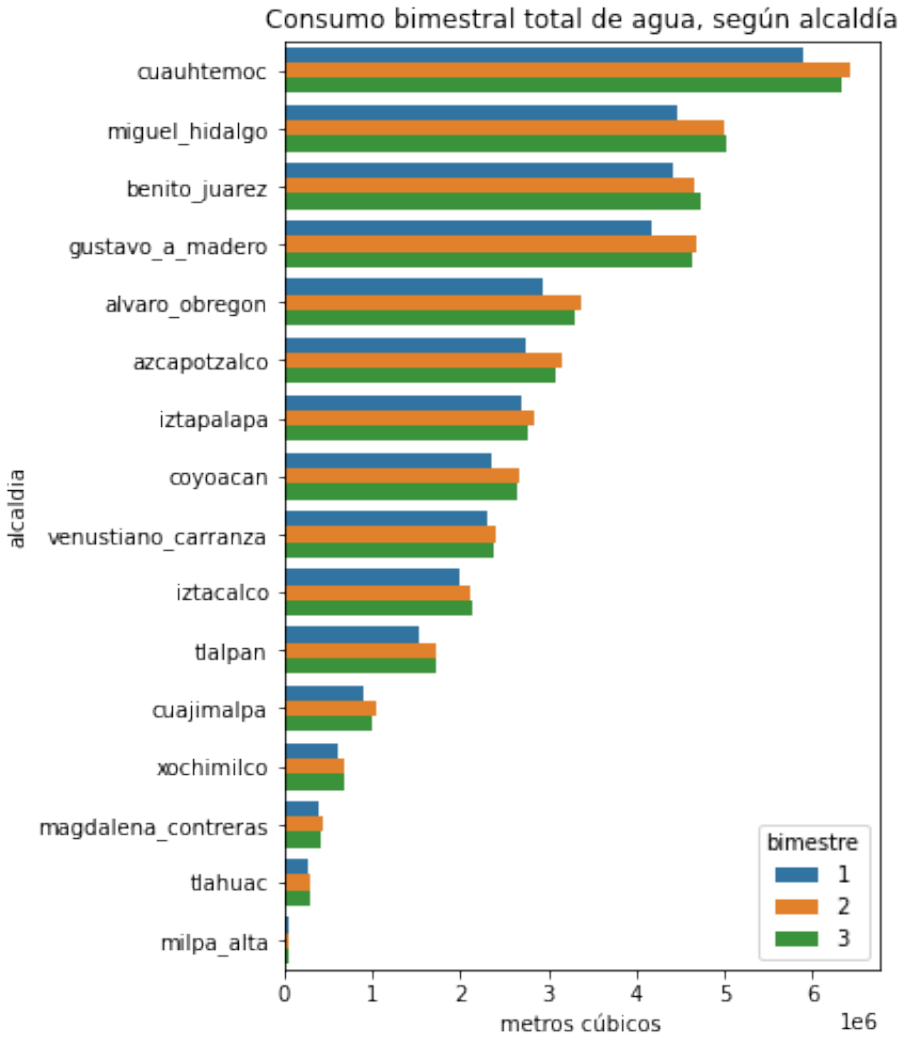
3

## ¿Cúantas observaciones se tienen por bimestre?

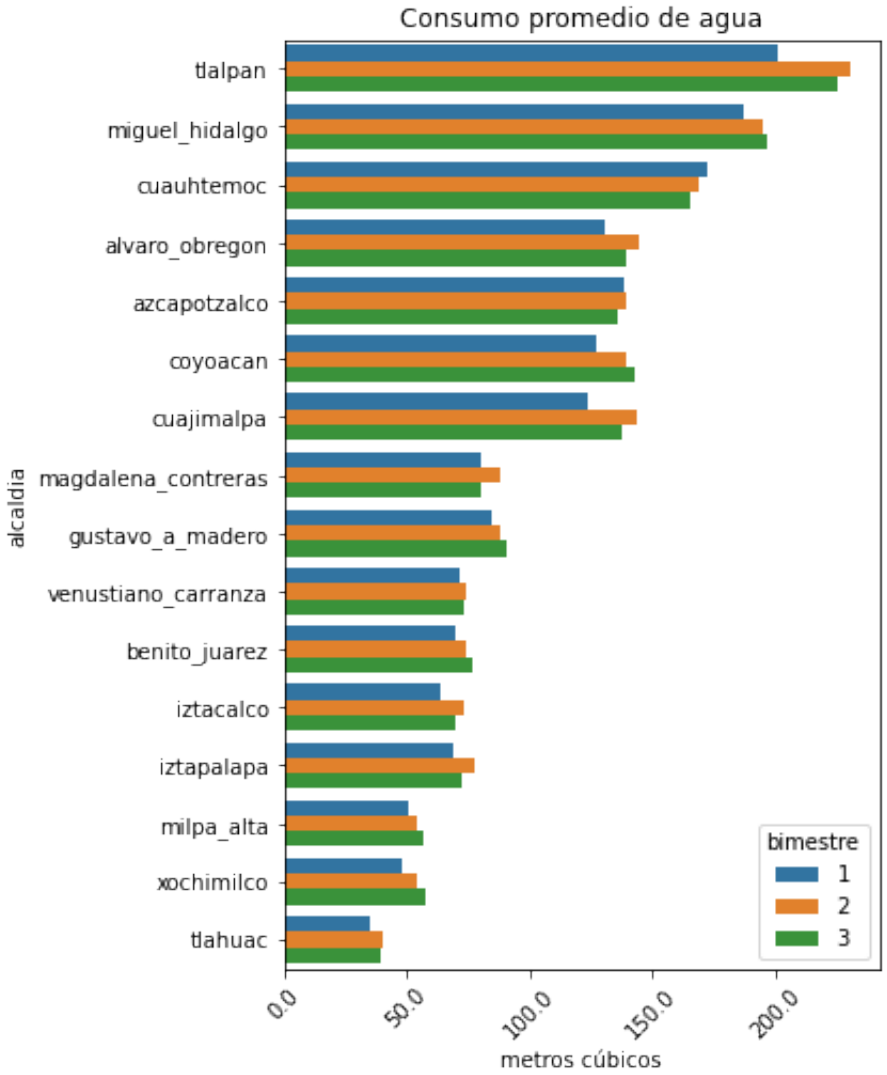
| Bimestre | Observaciones |
|---|---|
| 1 | 23,338 |
| 2 | 23,942 |
| 3 | 23,822 |

## ¿Existen diferencias considerables entre el consumo por bimestre?

Podemos graficar para cada alcaldía el consumo total por bimestre para identificar si el comportamiento es muy distinto entre bimestres:



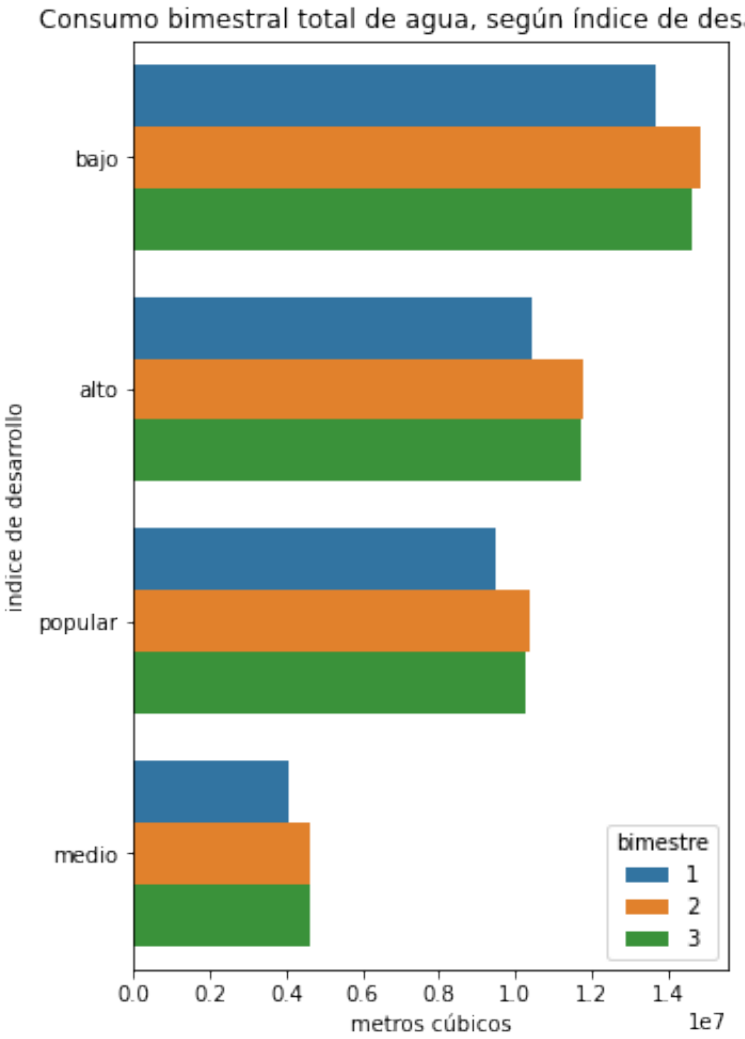

Hacemos lo mismo con el consumo promedio bimestral:



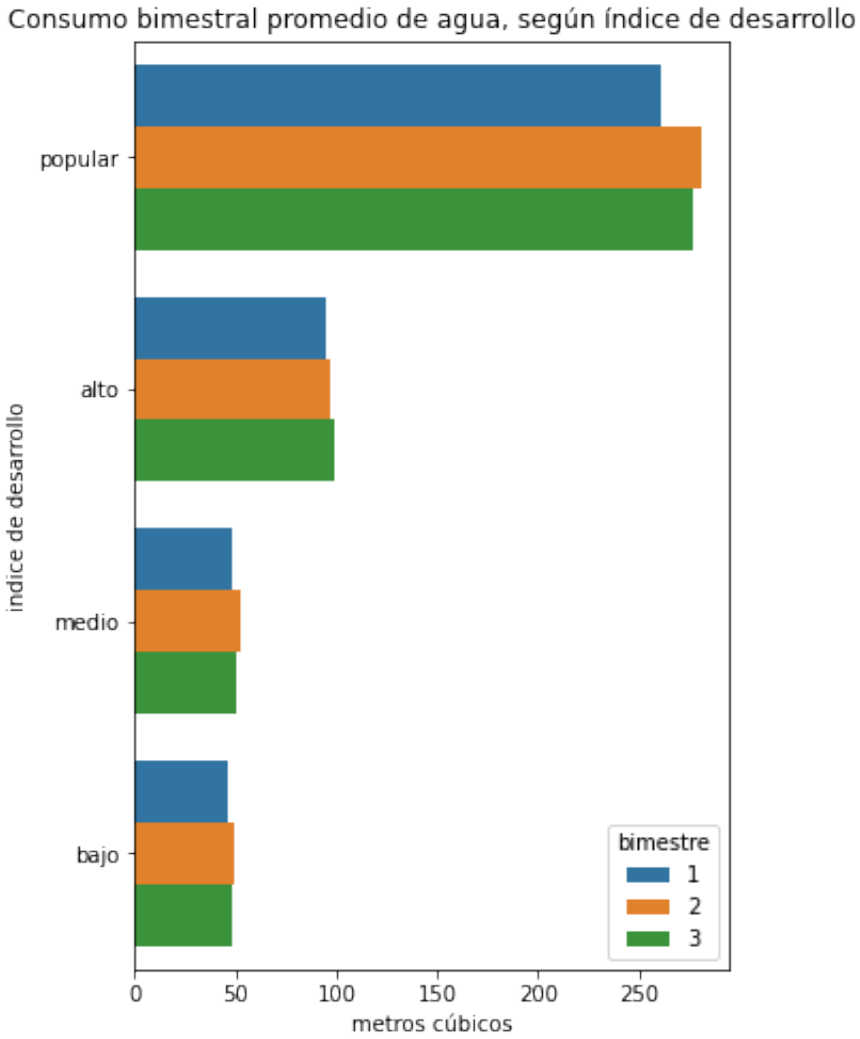

Podemos observar que, en términos generales, el consumo (total y promedio) del primer brimestre es ligeramente más bajo respecto los bimestres dos y tres, sin embargo la diferencia no parece ser significativa.

En ese sentido, por simplicidad y ante la falta de información sobre años previos que permitan verificar estacionalidad en los datos, el análisis y el GEDA de las variables de consumo se realizará para el total del año.

Asimismo, verificamos a través del mismo procedimiento el consumo total por bimestre según el índice de desarrollo y tampoco observamos diferencias considerables entre bimestres:




Lo mismo para el consumo promedio bimestral:




# Estudio de alcaldías y colonias de la CDMX

Para analizar las colonias de la Ciudad de México se contestaron las siguientes preguntas:

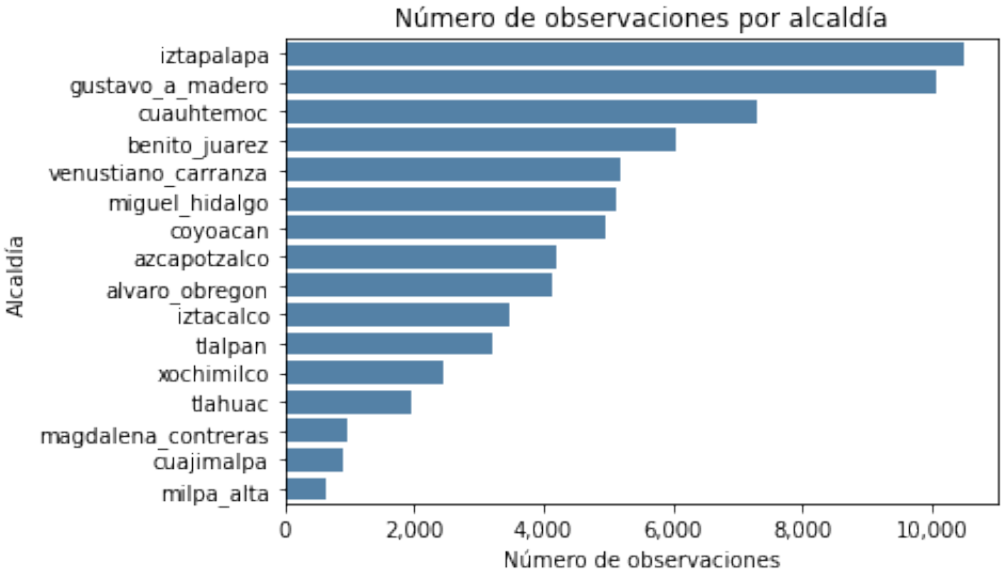
- ¿Cuántas alcaldías existen?
- ¿Se tienen números de observaciones similares para cada alcadía?
- ¿Cuántas colonias diferentes existen?
- ¿Cuántas colonias se tienen por alcaldía?
- ¿Coincide este dato con los datos oficiales?
- ¿Cuantás colonias existen por índice de desarrollo?

## ¿Cuántas alcaldías existen?

```
Existen 16 alcaldías diferentes

Existen 16 valores para 'nomgeo' diferentes
```

## ¿Se tienen números de observaciones similares para cada alcadía?

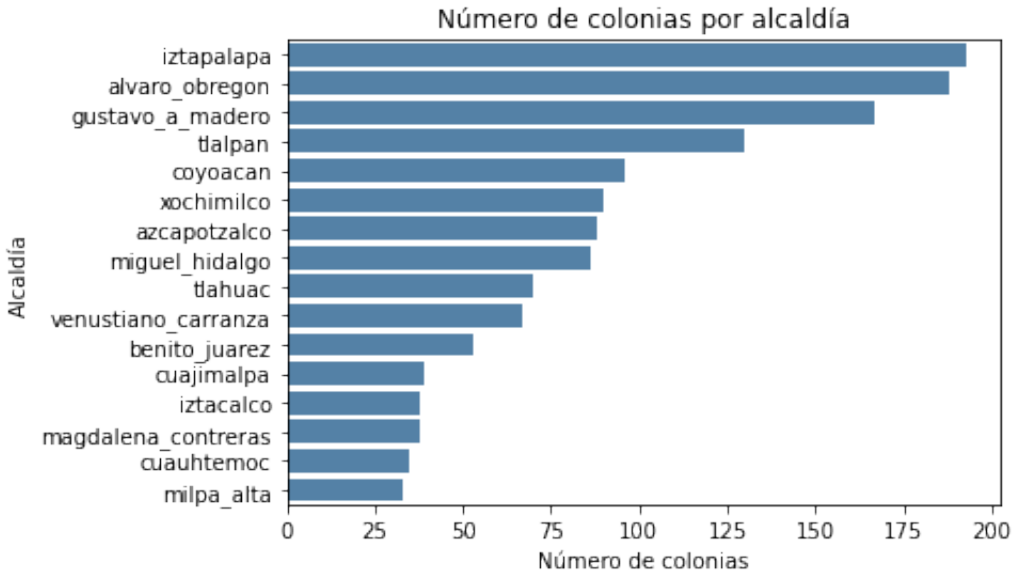



## ¿Cuántas colonias diferentes existen?

```
1340
```

## ¿Cúantas colonias se tienen por alcadía?

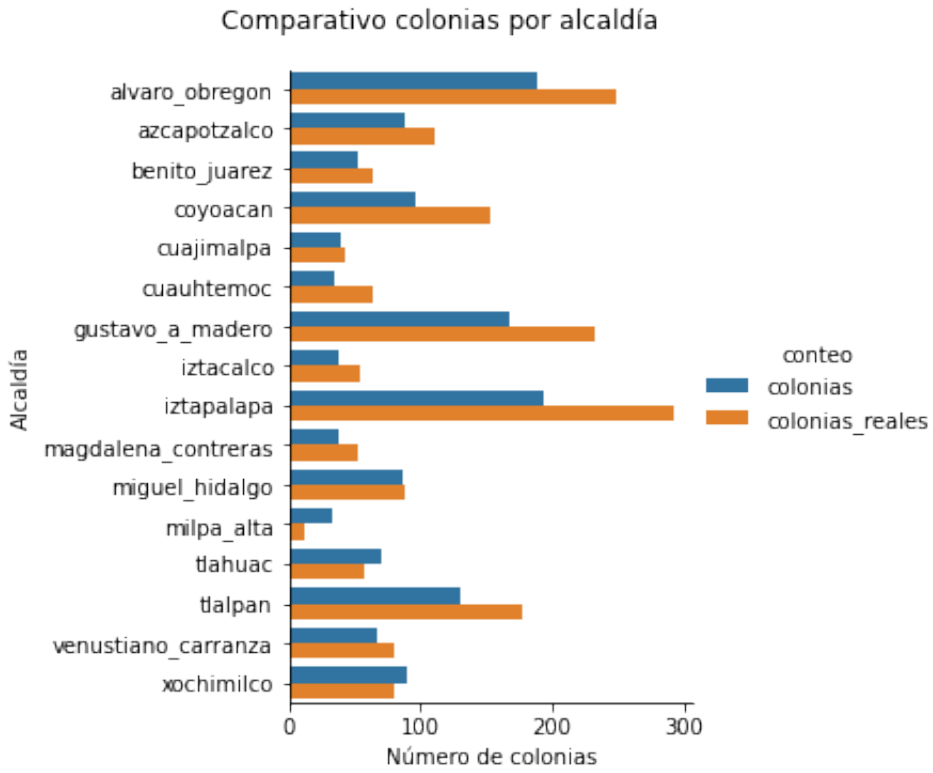
| Alcaldía | Numero de colonias |
|---|---|
| Álvaro Obregón | 188 |
| Azcapotzalco | 88 |
| Benito Juárez | 53 |
| Coyoacán | 96 |
| Cuajimalpa | 39 |
| Cuauhtémoc | 35 |
| Gustavo A. Madero | 167 |
| Iztacalco | 38 |
| Iztapalapa | 193 |
| Magdalena Contreras | 38 |
| Miguel Hidalgo | 86 |
| Milpa Alta | 33 |
| Tláhuac | 70 |
| Tlalpan | 130 |
| Venustiano Carranza | 67 |
| Xochimilco | 90 |




## ¿Coincide este dato con los datos oficiales?

Actualizando con datos reales de la CDMX:

| Alcaldía | Colonias | Colonias Reales | Fracción |
|---|---|---|---|
| Álvaro Obregón | 188 | 249 | 76.0 |
| Azcapotzalco | 88 | 111 | 79.0 |
| Benito Juárez | 53 | 64 | 83.0 |
| Coyoacán | 96 | 153 | 63.0 |
| Cuajimalpa | 39 | 43 | 91.0 |
| Cuauhtémoc | 35 | 64 | 55.0 |
| Gustavo A. Madero | 167 | 232 | 72.0 |
| Iztacalco | 38 | 55 | 69.0 |
| Iztapalapa | 193 | 293 | 66.0 |
| Magdalena Contreras | 38 | 52 | 73.0 |
| Miguel Hidalgo | 86 | 88 | 98.0 |
| Milpa Alta | 33 | 12 | 275.0 |
| Tláhuac | 70 | 58 | 121.0 |
| Tlalpan | 130 | 178 | 73.0 |
| Venustiano Carranza | 67 | 80 | 84.0 |
| Xochimilco | 90 | 80 | 112.0 |

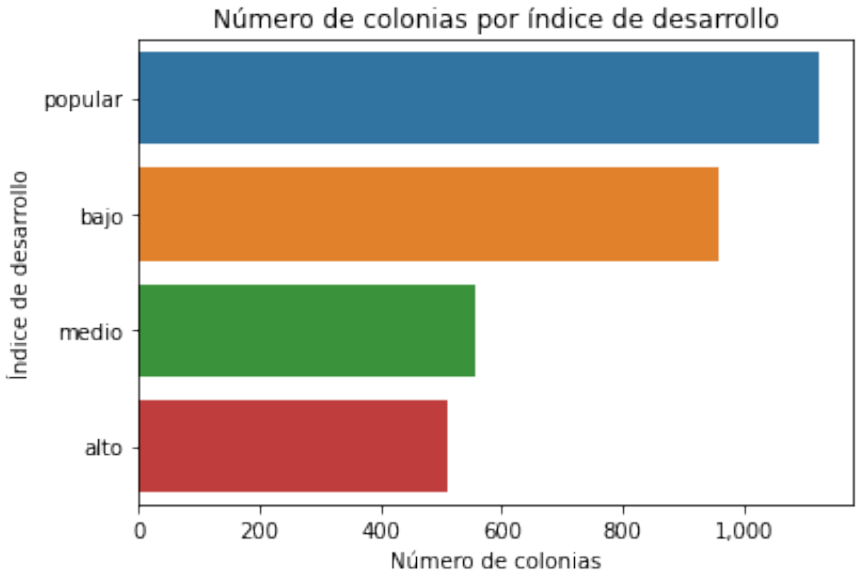



Existen diferencias significativas entre el número de `colonias` detectadas en los datos empleados y las `colonias_reales` reportadas en los datos oficiales de la CDMX. En el caso de tener valores inferiores, puede deberse a la falta de captura de consumos en dichos lugares. Por otro lado, existen valores superiores para algunas alcaldías como es el caso de Milpa Alta. En los datos oficiales, se reportan una sola colonia **Villa Milpa Alta**, mientras que en los datos esta se fragmenta en:

```
- pblo_villa_milpa_alta_bo_la_concepcion
- pblo_villa_milpa_alta_bo_los_angeles
- pblo_villa_milpa_alta_bo_santa_martha
- pblo_villa_milpa_alta_bo_san_mateo
- pblo_villa_milpa_alta_bo_santa_cruz
- villa_milpa_alta_centro
- pblo_villa_milpa_alta_bo_san_agustin
- pblo_villa_milpa_alta_bo_la_luz
```

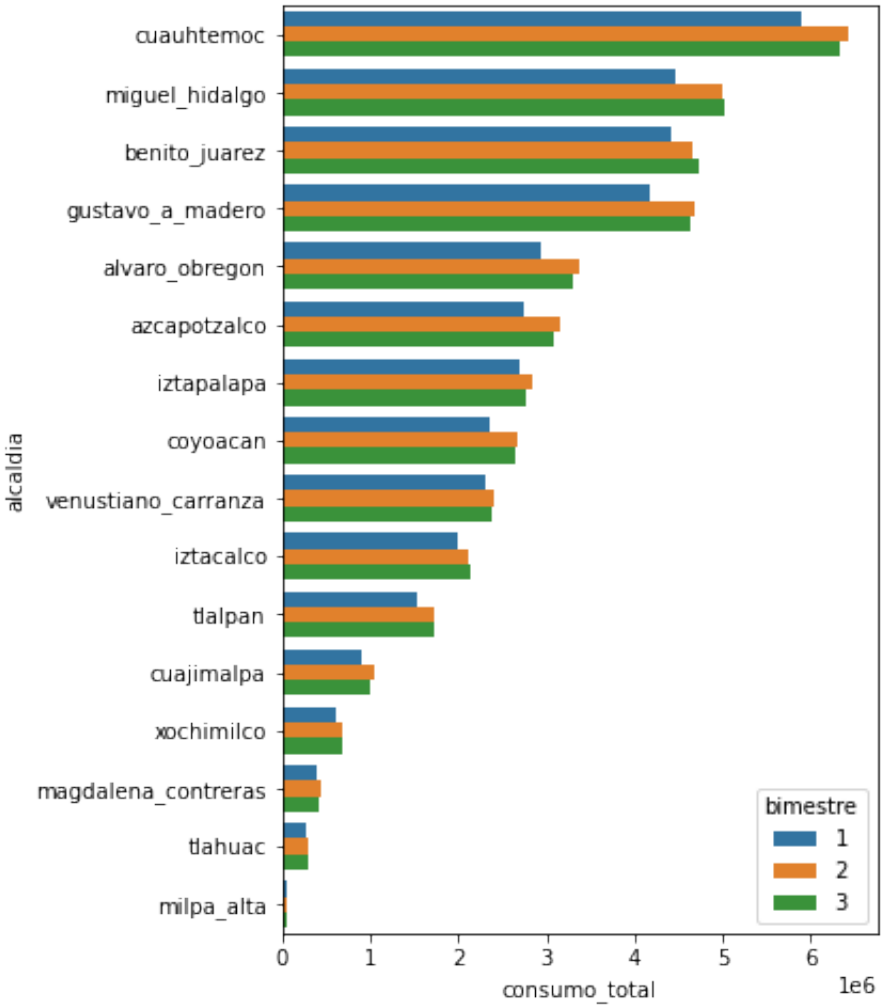
## ¿Cuantás colonias existen por índice de desarrollo?

| Índice de desarrollo | Número de colonias |
|---|---|
| Alto | 512 |
| Medio | 558 |
| Popular | 1,124 |
| Bajo | 958 |




# Consumo total

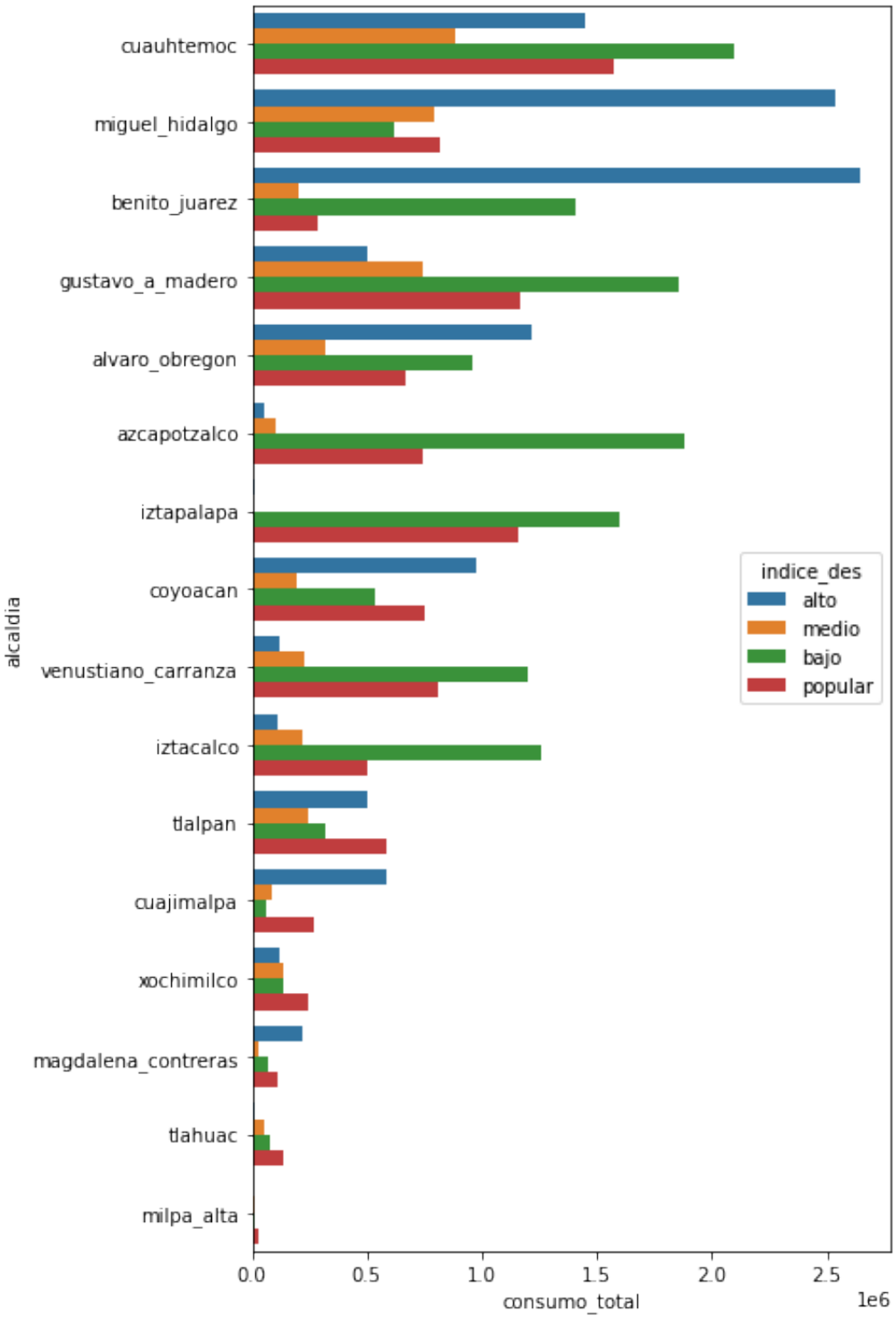
## ¿Existen diferencias importantes en cuanto a consumo total, entre bimestres?



Como se puede apreciar, el consumo total no varía por bimestre. Para simplificar el análisis exploratorio, los 3 bimestres serán resumidos en uno sólo por medio del promedio. De esta manera, tendremos una sóla observación por manzana.

Una vez que se tienen los datos promediados a lo largo del semestre, se prosigue con el análisis del consumo total, por alcaldía y por índice de desarrollo.

## ¿Cómo se distribuye el consumo entre niveles de desarrollo, según la alcaldía?

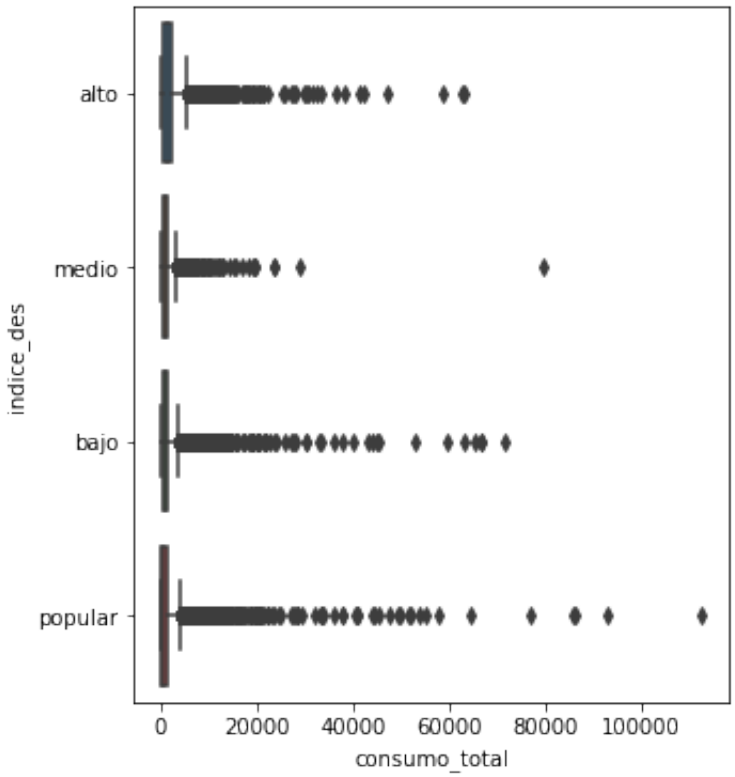


- En las alcaldías Miguel Hidalgo, Benito Juárez y Cuajimalpa la mayor parte del consumo total proviene de manzanas correspondientes a índice alto.
- En las alcaldías Cuahtémoc, Álvaro Obregón, Tlalpan, Coyoacán y Xochimilco el consumo total se divide de forma relativamente equitativa entre 3 o 4 níveles del índice.
- En Azcapotzalco, Iztapalapa, Iztacalco, Venustiano Carranza y Milpa Alta el consumo total proviene predominantemente de manzanas con índice bajo o popular.

## Distribución de consumo total por índice de desarrollo

## ¿Existe un nivel de índice de desarrollo que presente consumos totales más altos que otros niveles?

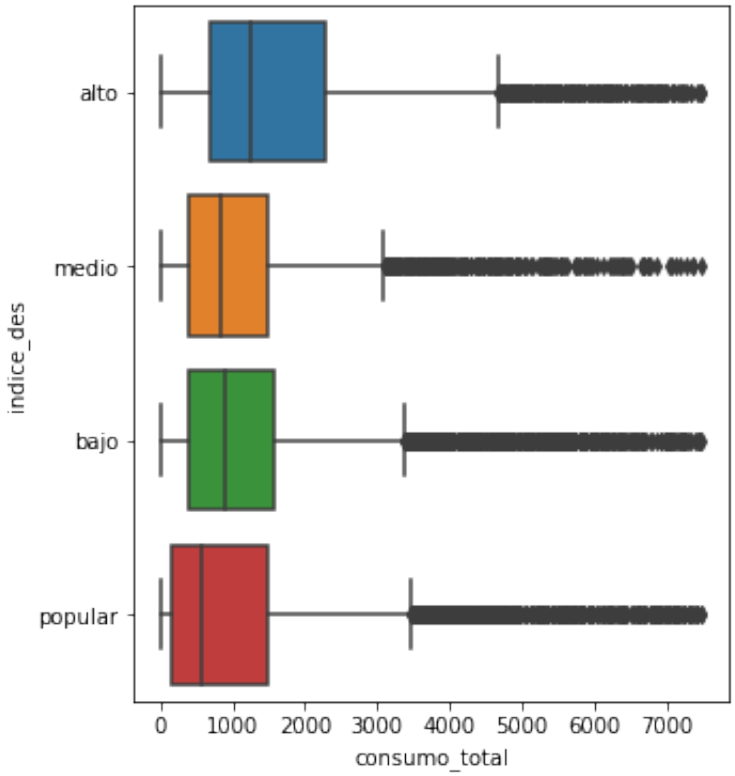
Para contestar esta pregunta, un boxplot resulta útil.




Los boxplots se ven muy "aplastados" debido a la gran dispersión de valores que existe. Sin embargo es útil para ver que en el nivel popular contiene manzanas con algunos de los valores bimestrales más altos rebasando los 80,000 $m^3$ en más de una ocasión.

Para ver la distribución general, sin considerar valores tan extremos, se realiza un filtro en el cual sólo se considerarán las manzanas con un consumo total bimestral menor a 7,500 $m^3$

```
<AxesSubplot:xlabel='consumo_total', ylabel='indice_des'>
```

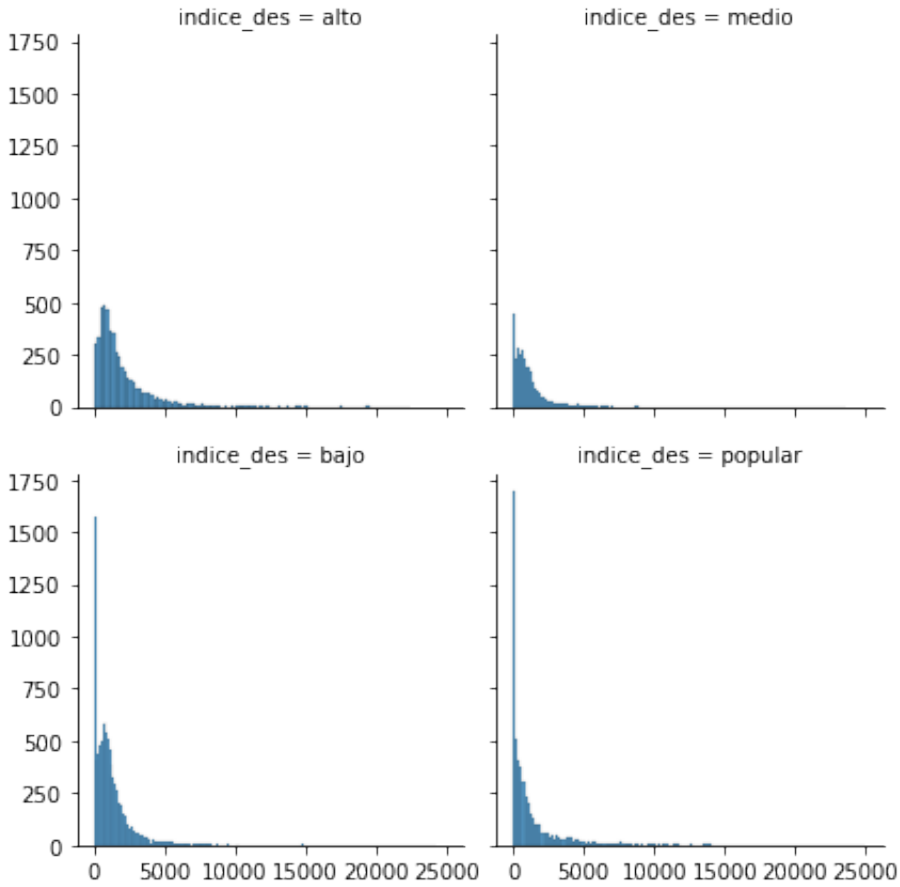



Aunque la mediana y cuartil superior del nivel alto son mayores a aquellas correspondientes a otros niveles, la gran dispersión de los datos no permite observar un resultado importante en este sentido. Por lo que, parece ser, que los niveles de consumo total por colonia se distribuyen de forma similar entre niveles de índice de desarrollo

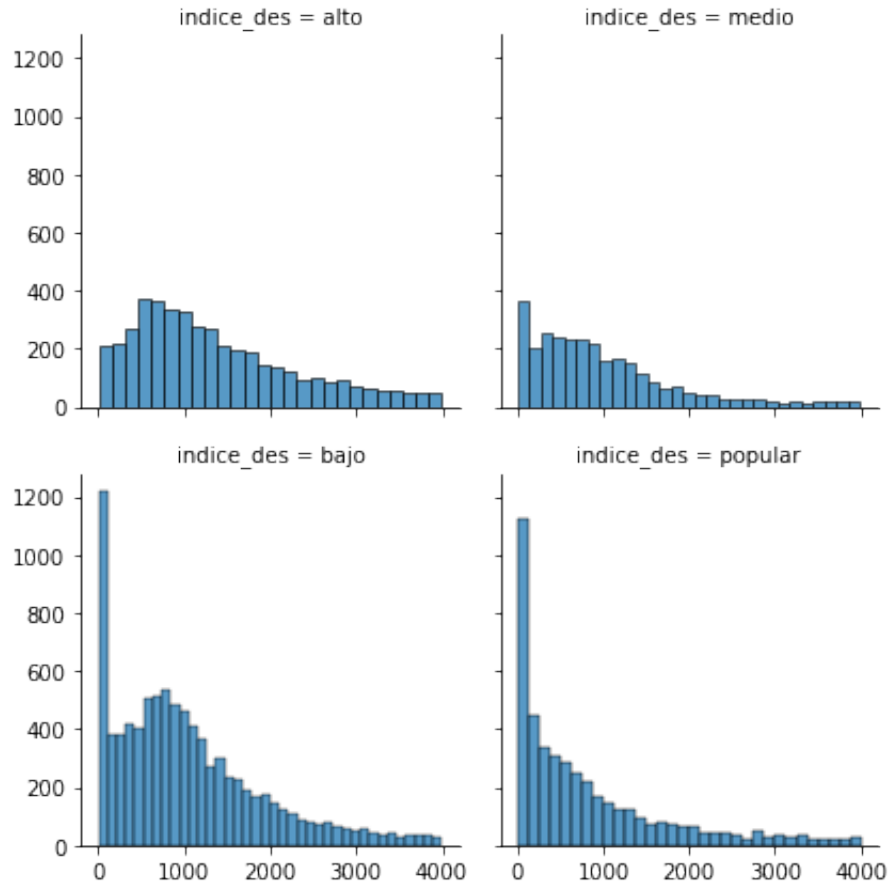
Un boxplot es útil para ver el rango, y los cuartiles de los datos, sin embargo, no permite apreciar la densidad o concentración de niveles de consumo total.

## ¿Qué tan frecuentes son los consumos totales tan altos?

## ¿Qué valores de consumo total son los más comunes, por nivel de desarrollo?




Los valores extremos no son nada comunes. Después de consumos totales de 5000, la cola de la distribución es muy delgada. Para apreciar mejor los valores más bajos, se realiza un filtro de consumos totales menores a 7500. Tambien se quitan consumos iguales a 0.

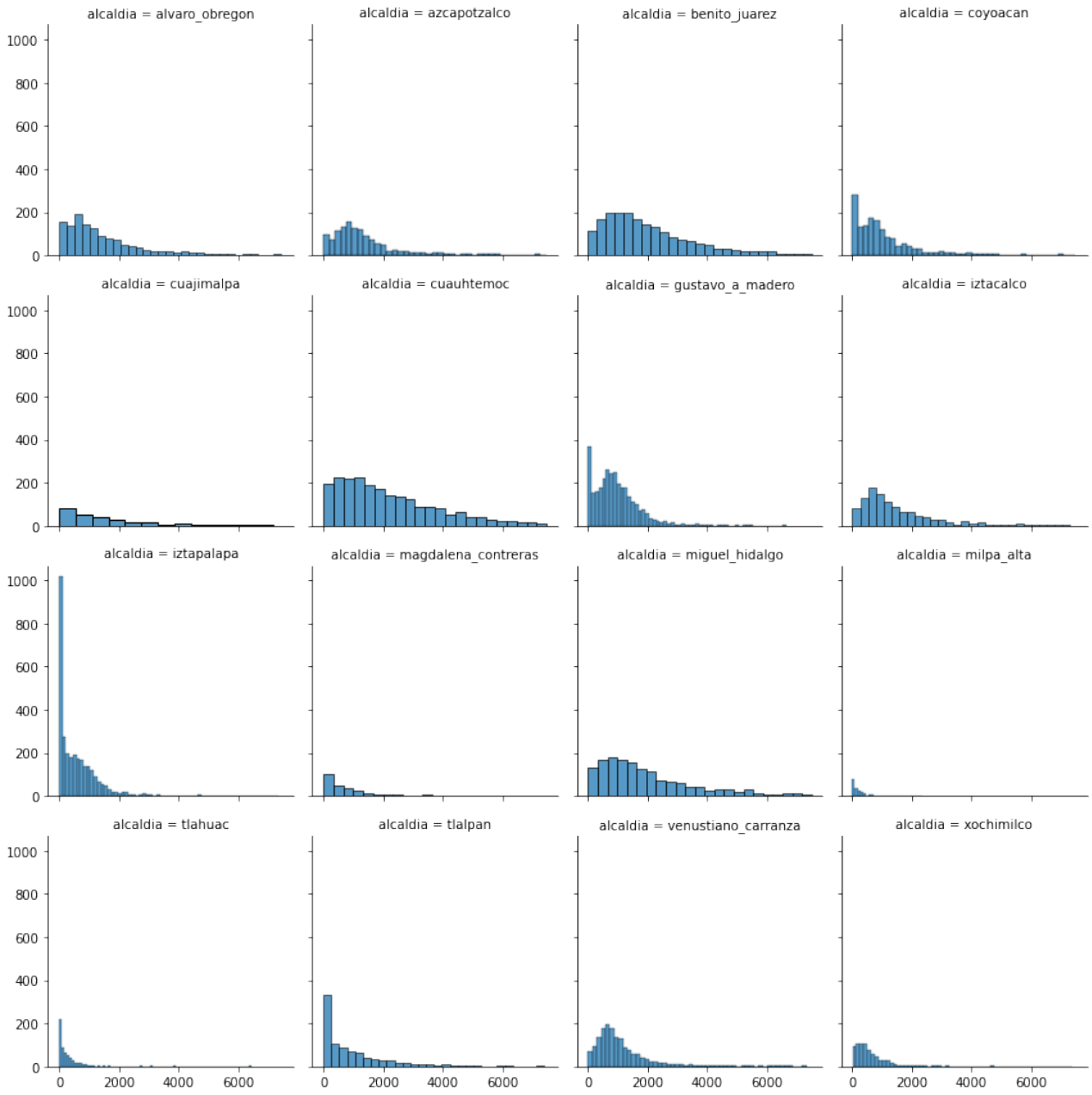



Las distribuciones de consumo total por nivel de desarrollo presentan comportamientos distintos entre sí, aunque con algunas similitudes.

- Para niveles de desarrollo bajo, popular y, en menor magnitud, mediano son muy comunes las manzanas con consumos cercanos a 0. Mientras que en indice alto esto no sucede.
- Quitando el "bin" más bajo de cada distribución, se observa que la moda del nivel alto y bajo son similares y cercanas a los 750 $m^3$.
- Mientrás más alto es el consumo, menor es el valor de la densidad en el caso del nivel popular.

## ¿Qué tan frecuentes son los consumos totales tan altos?

## ¿Qué valores de consumo total son los más comunes, por alcaldía?

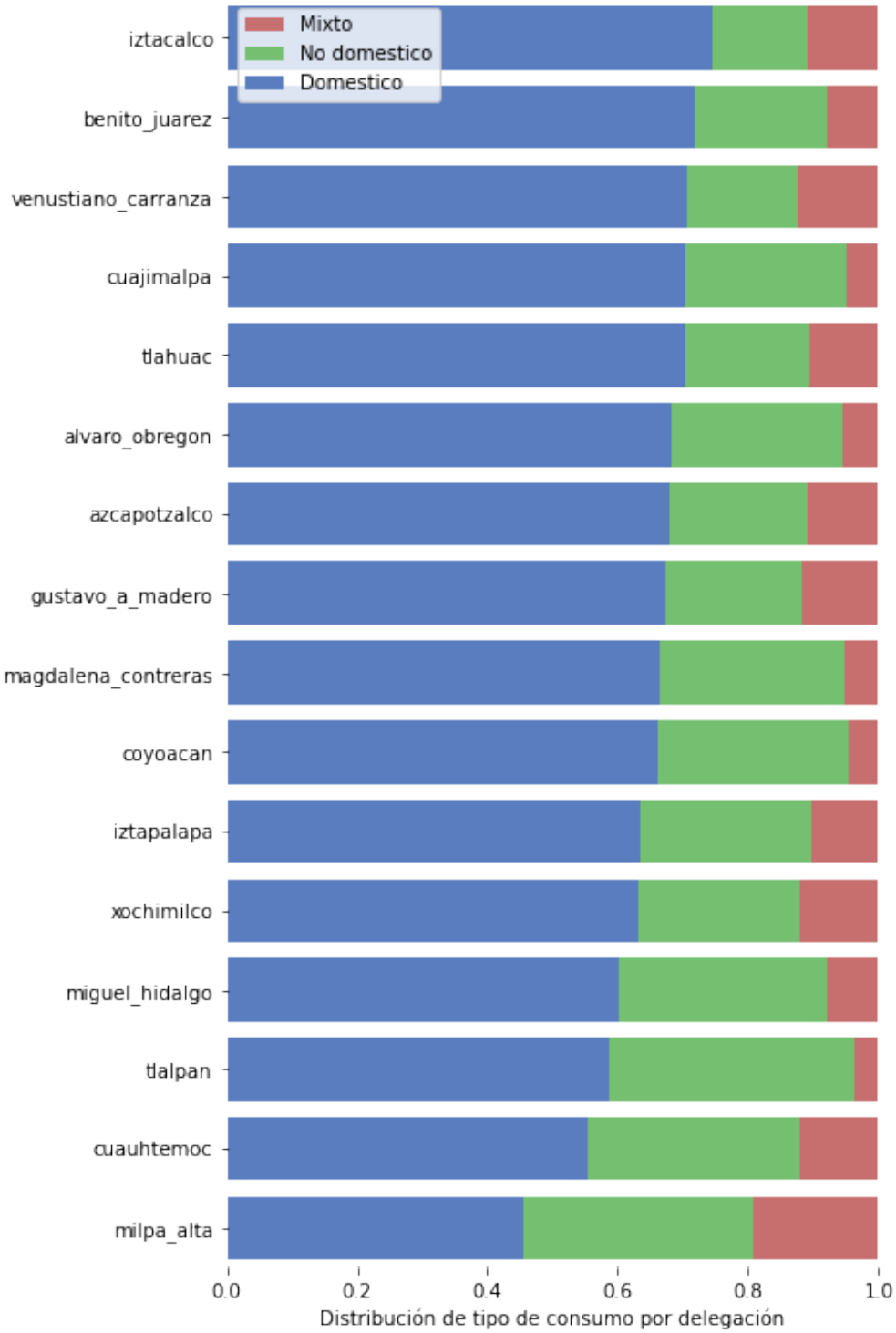



- En Iztapalapa se encuentra la gran mayoría de los registros cercanos a 0 que se observaron anteriormente.
- El consumo general de Milpa Alta es muy pequeño, como lo sugerían varias gráficas anteriormente.

## ¿Cómo se divide el consumo total entre doméstico, no doméstico y mixto por alcaldía?

## ¿Qué alcaldía presenta mayor proporción de consumo doméstico? ¿No doméstico?

Una gráfica de barras del porcentaje de cada tipo de consumo es útil para contestar dicha pregunta.

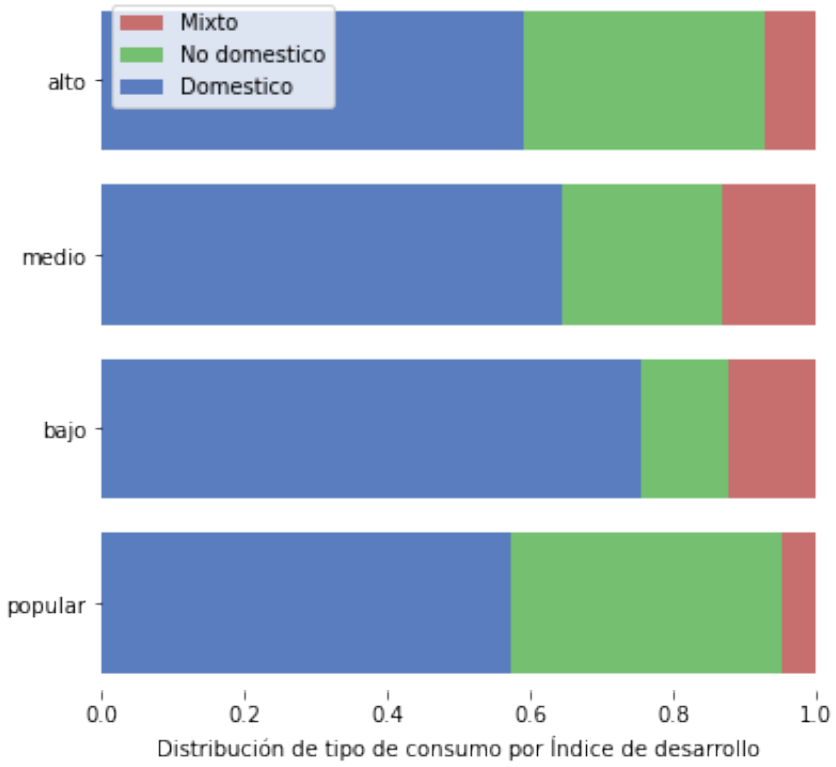



Iztacalco tiene la mayor proporción de consumo doméstico, mientras que Milpa Alta y Cuauhtemoc muestran las menores.

Tlalpan presenta la mayor proporción de consumo no doméstico, mientras que Iztacalco la menor.

### ¿Cómo se divide el consumo total entre doméstico, no doméstico y mixto por alcaldía?

### ¿Qué alcaldía presenta mayor proporción de consumo doméstico? ¿No doméstico?




- El consumo doméstico es más común en manzanas de nivel de desarrollo bajo.
- Los níveles alto y popular presentan proporciones similares en los 3 tipos de consumo. Tienen las proporciones más bajas de consumo doméstico y más altas de consumo no doméstico
- Proporciones relativamente altas de consumo mixto se ven en niveles medio y bajo

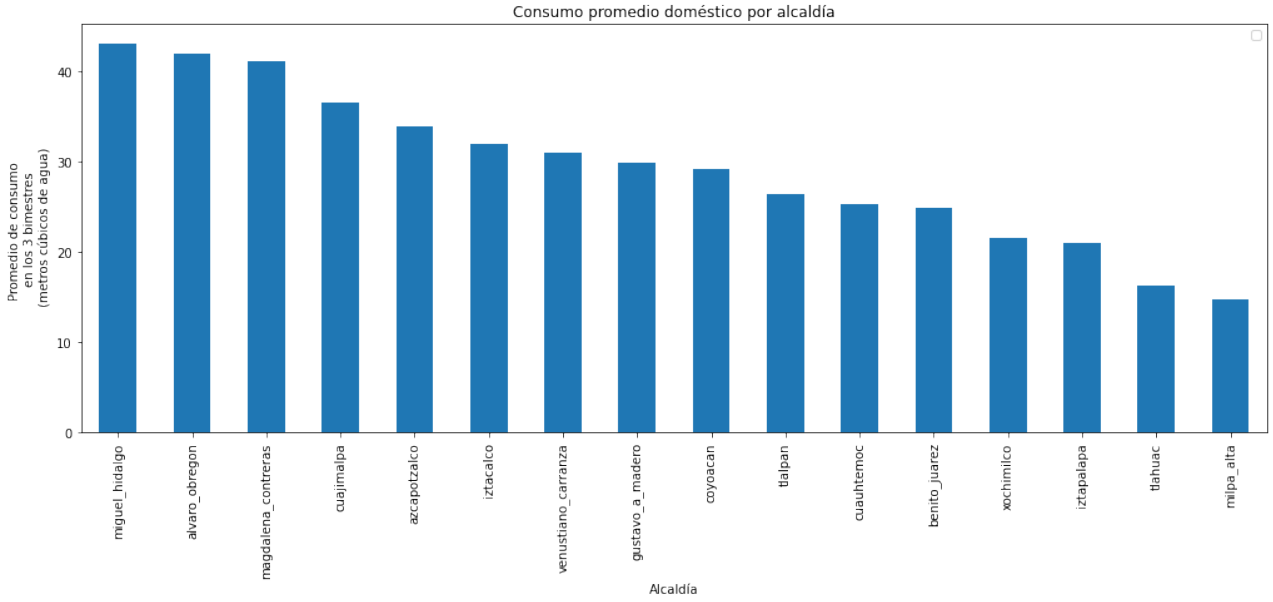
# Consumo promedio

En esta sección analizaremos el consumo promedio de las tres variables observadas para el total: consumo doméstico, no doméstico y mixto.

## Consumo promedio doméstico por colonia.

En la siguiente gráfica observamos que hay una variación considerable en el consumo doméstico por colonia. Colonias como Miguel Hidalgo o Álvaro Obregón tienen un nivel de consumo considerablemente mayor al de colonias como Tláhuac o Milpa Alta.

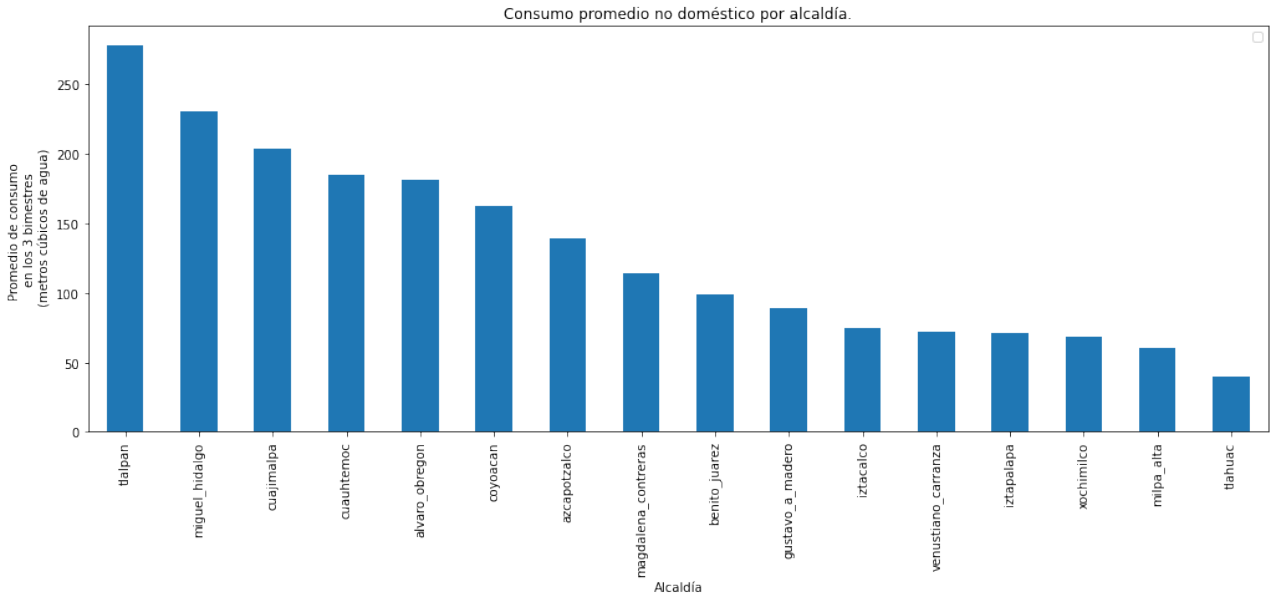
Se decidió utilizar el promedio de los tres bimestres por fines de claridad y porque el consumo tuvo poca variación de bimestre a bimestre en cada colonia (lo cual se mostrará en la siguiente gráfica).




El conjunto de datos no presenta unidades, pero el promedio del consumo promedio doméstico por colonia es de 29.29 unidades, con una desviación estándar de 8.65 unidades, un mínimo de 14.8 (Milpa Alta) y un máximo de 43 (Miguel Hidalgo).

## Consumo promedio no doméstico por colonia

A continuación observamos la misma gráfica para el promedio del consumo no doméstico.

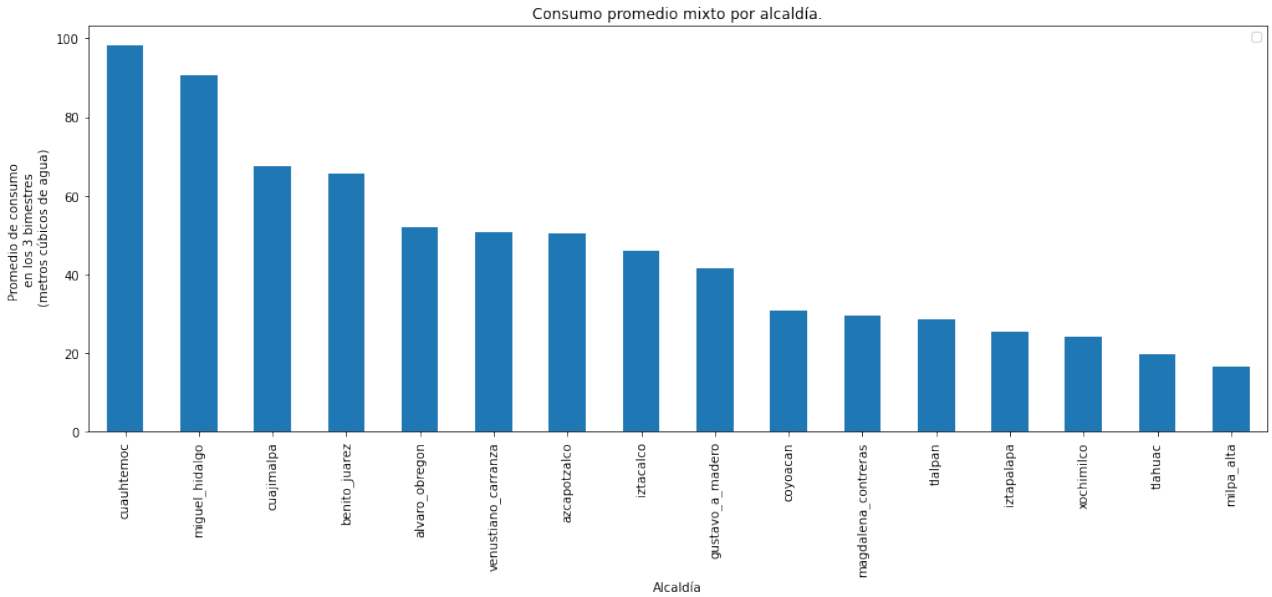



Es interesante notar que algunas colonias, como Tlalpan o Cuauhtémoc, no son preponderantes en el consumo doméstico pero sí en el no doméstico. Esto es lo que habríamos de esperar, pues xisten zonas residenciales y zonas industriales en el país.

El promedio del consumo no doméstico es de 129 unidades, más de 4 veces el promedio del consumo doméstico. La varianza es mayor en este caso: con una desviación estándar de 70 unidades, observamos colonias como Tláhuac, que cuenta con tan solo 40 unidades de consumo y colonias como Tlalpan, con 278 unidades de consumo en promedio para los tres bimestres.

## Consumo promedio mixto por colonia.

El consumo mixto está definido como aquel que ocurre en zonas donde simultáneamente se usa el agua para consumo doméstico y no doméstico. ¿Cómo se ve la distribución por colonia en en este caso?



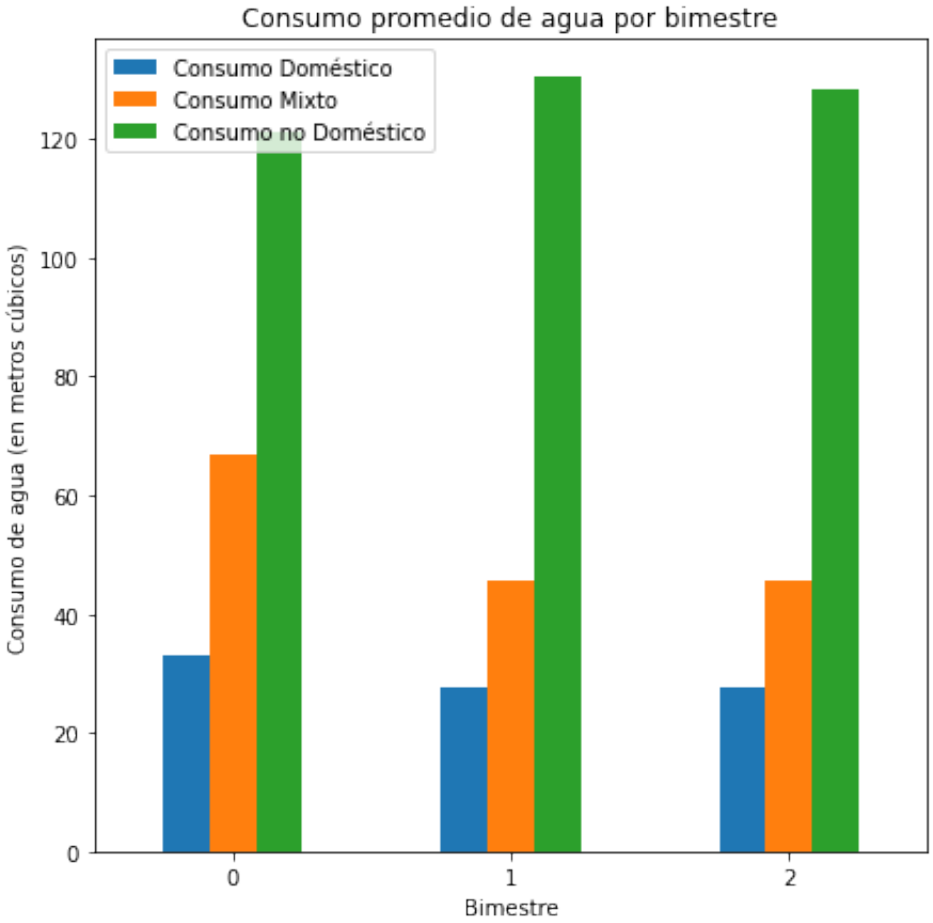

En este caso, el mínimo es de 16.63 unidades y el máximo de 98.25. El promedio del consumo promedio mixto de agua es de 46 unidades.

La siguiente tabla, que muestra el ranking de cada colonia en estos tres rubros, nos permite ver que algunas colonias, como Miguel Hidalgo o Cuajimalpa, tienen una actividad importante en estos tres campos, mientras que otras, como Milpa Alta y Tláhuac, tienen un consumo relativamente bajo. Por supuesto, existen otras, como Cuahutémoc o Tlalpan, que tienen un consumo alto en algunas categorías y bajo en otras.

| Colonia | Ranking Doméstico | Ranking No Doméstico | Ranking Mixto |
|---|---|---|---|
| Miguel Hidalgo | 1 | 2 | 2 |
| Álvaro Obregón | 2 | 5 | 5 |
| La Magdalena Contreras | 3 | 8 | 11 |
| Cuajimalpa de Morelos | 4 | 3 | 3 |
| Azcapotzalco | 5 | 7 | 7 |
| Iztacalco | 6 | 11 | 8 |
| Venustiano Carranza | 7 | 12 | 6 |
| Gustavo A. Madero | 8 | 10 | 9 |
| Coyoacán | 9 | 6 | 10 |
| Tlalpan | 10 | 1 | 12 |
| Cuauhtémoc | 11 | 4 | 1 |
| Benito Juárez | 12 | 9 | 4 |
| Xochimilco | 13 | 14 | 14 |
| Iztapalapa | 14 | 13 | 13 |
| Tláhuac | 15 | 16 | 15 |
| Milpa Alta | 16 | 15 | 16 |

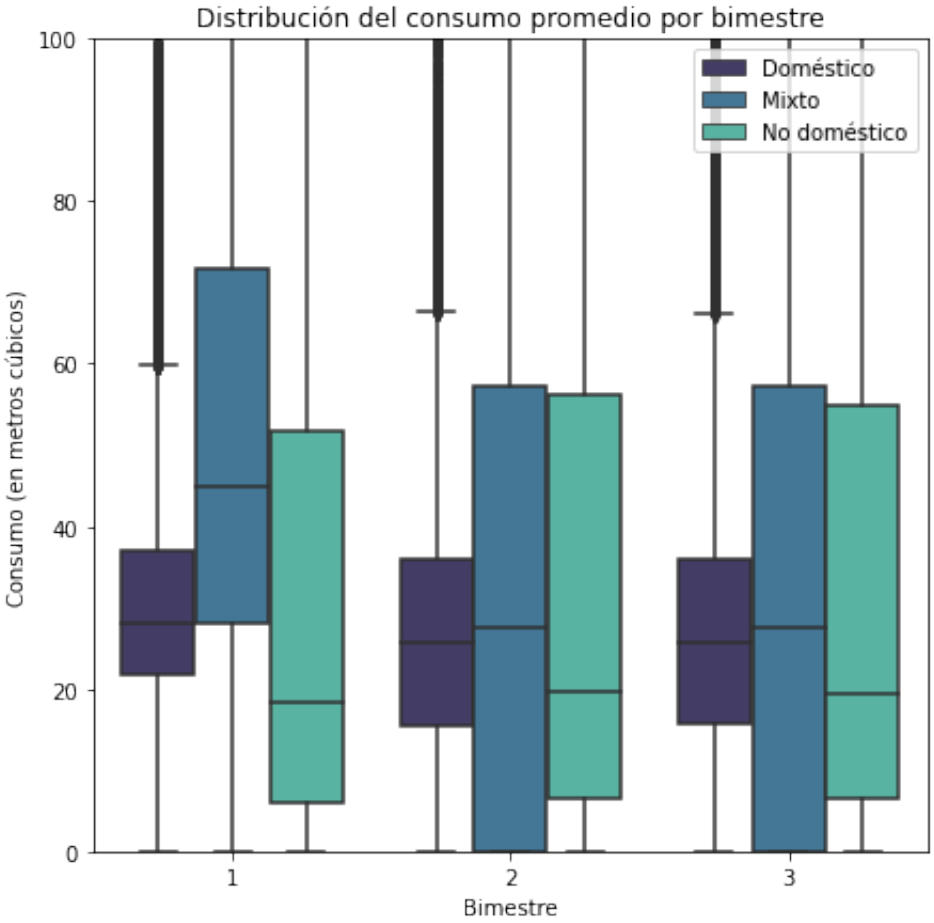
Nota: La tabla fue hecha a mano.

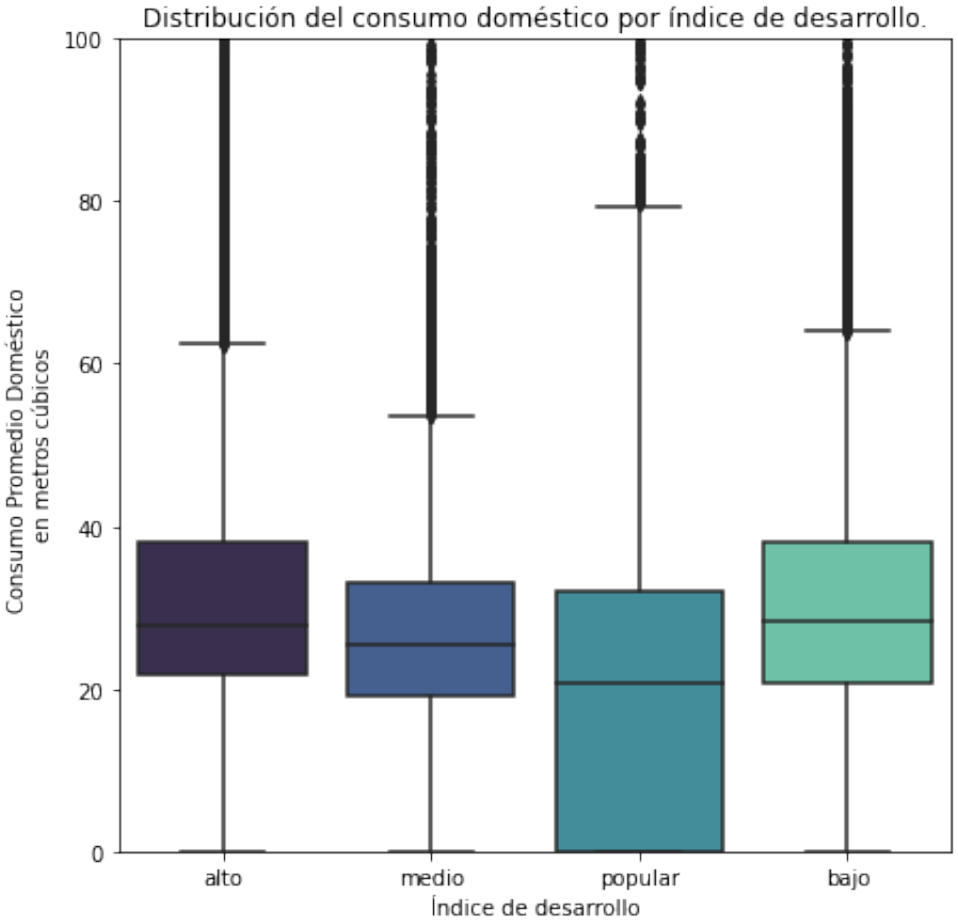
En la siguiente gráfica podemos observar que el consumo promedio, tanto del sector doméstico como del no doméstico, tuvo pocos cambios de bimestre a bimestre. El consumo mixto disminuyó de 67 a 45 unidades del primer al segundo bimestre pero después se mantuvo estable.



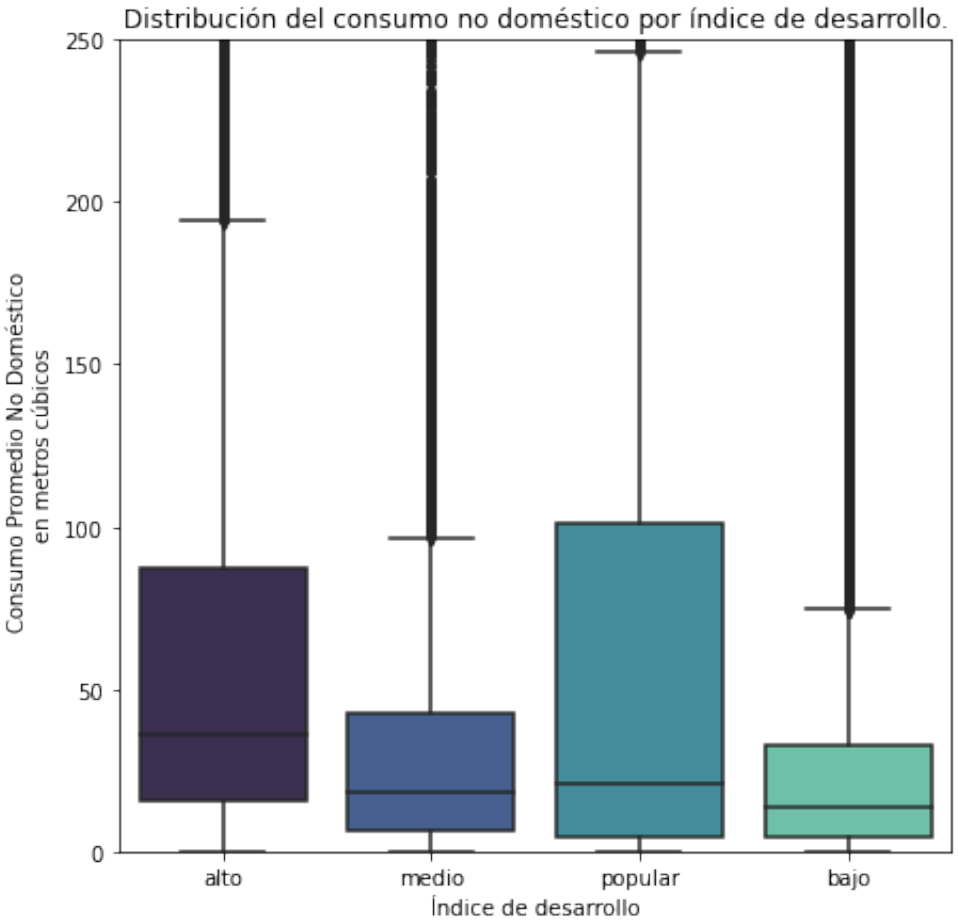

## Distribución del consumo promedio por bimestre y por índice de desarrollo.

Curiosamente, a pesar de que algunas colonias tienen un consumo mucho mayor que otras, podemos notar que sus distribuciones son relativamente similares, con un medianta cercana a las 28 unidades. Es relevante también mencionar que el sector con índice de desarrollo popular tiene una varianza mayor en su consumo promedio de agua.

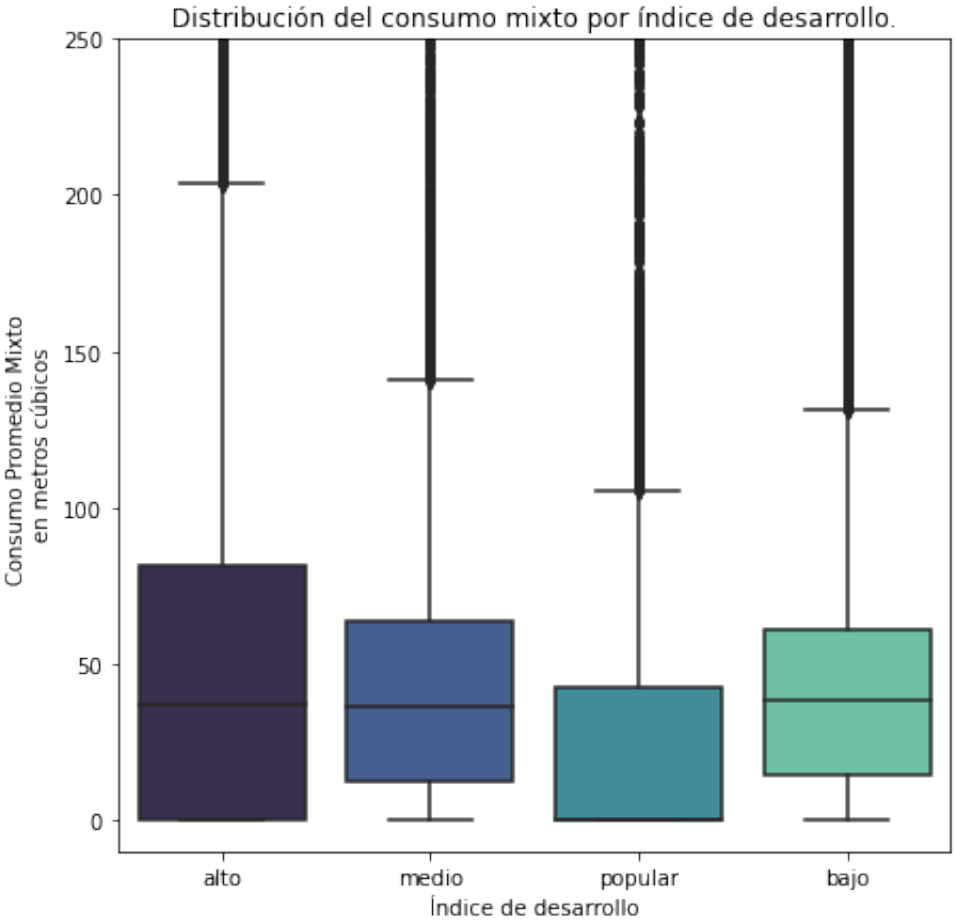

Distribución del consumo promedio por bimestre
Doméstico
Mixto
No doméstico
Consumo (en metros cúbicos)
0
20
40
60
80
100
1
2
3
Bimestre


Distribución del consumo doméstico por índice de desarrollo.
Consumo Promedio Doméstico
en metros cúbicos
0
20
40
60
80
100
alto
medio
popular
bajo
Índice de desarrollo

Para el sector no doméstico, las distribuciones varían un poco más. Notamos que el sector con alto índice de desarrollo tiene una mediana notablemente mayor que la de los demás sectores, aunque sigue siendo cierto que el sector popular tiene mayor varianza en su consumo promedio.




Finalmente, realizamos el mismo análisis para el consumo mixto.

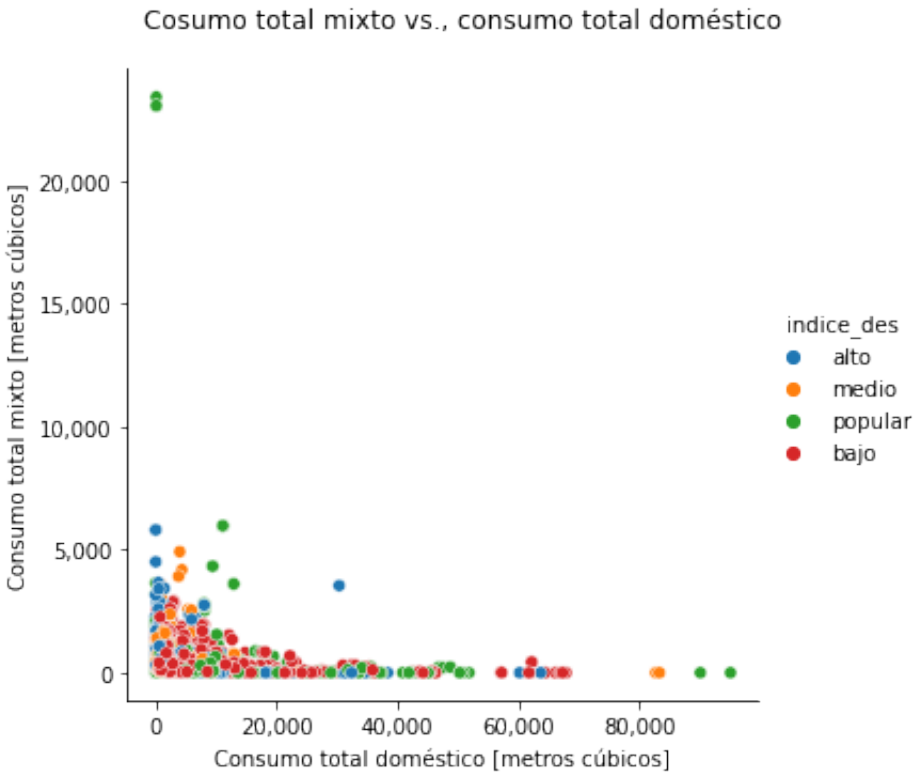

Distribución del consumo mixto por índice de desarrollo.
Consumo Promedio Mixto en metros cúbicos
250
200
150
100
50
0
alto
medio
popular
bajo
Índice de desarrollo

# Características de los datos para la definición de un modelo

Como se solicitó para el presente proyecto, se estudió la factibilidad de emplear las variables del set de datos para predecir el índice de desarrollo de la manzana en cuestión (i.e. `indice_des`). Para ello, de las siguientes variables:

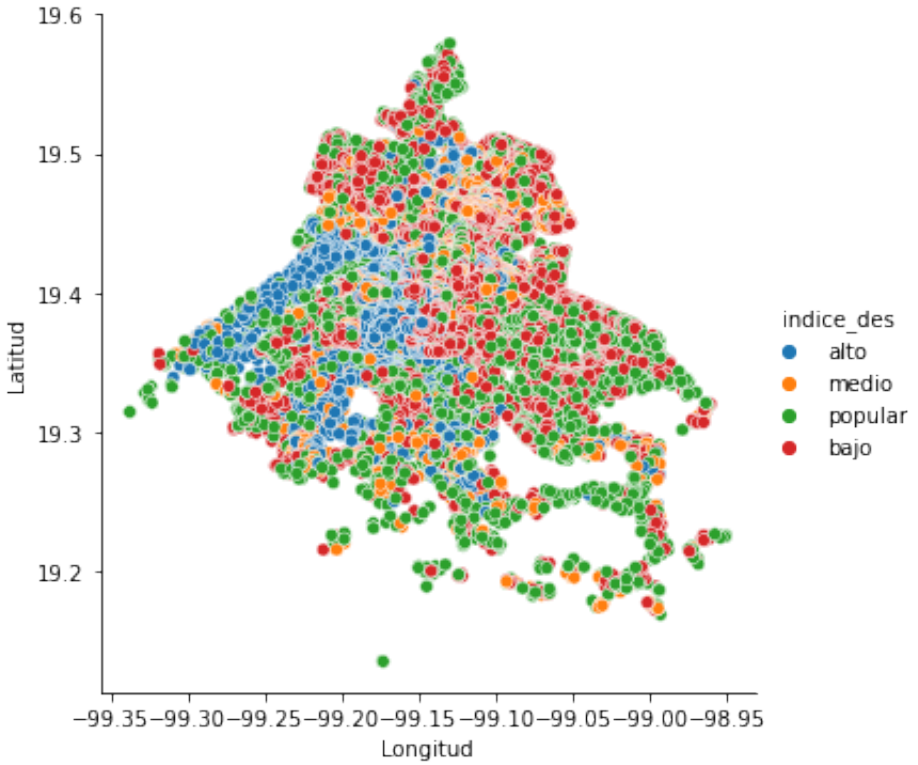
- Año
- Bimestre
- GID
- Latitud
- Longitud
- Consumo total
- Consumo total doméstico
- Consumo total no doméstico
- Consumo total mixto
- Consumo promedio
- Consumo promedio doméstico
- Consumo promedio no doméstico
- Consumo promedio mixto

Se realizaron gráficas de dispersión con coloración por índice de desarrollo para determinar si con las variables originales se podía obtener una segmentación de las categorías. Todas las gráficas realizadas (a excepción de latitud contra longitud que se muestra más adelante) no lograron este propósito. Un ejemplo del resultado típico de muestra a continuación.




Como es posible observar, no existe una segmentación adecuada de las categorías de índice de desarrollo. Las demás gráficas se omiten para evitar aglomerar el reporte.

El único caso de parcial éxito es la gráfica de longitud contra latitud que se muestra a continuación:



Por lo anterior, se considera que la información no es suficiente ni completa para generar un modelo correcto para predecir el índice de desarrollo. Nos gustaría contar con información adicional como número de tomas, tarifa promedio y población por manzana.

En general se concluye que:

- Las con las que contamos no permiten una segmentación de los datos
- Es conveneiente tener datos de población por manzana para perfilar el tipo de consumo per cápita
- Las gráficas de cajas y brazos demuestran que las distribuciones son más o menos similares y no hay patrones evidentes con los datos con los que se cuentan
- Es relevante contar con más contexto