

Trabajo de curso Ciencia de Datos

Consumo de energía en Chicago en 2010

Alberto Ramos Sánchez *

8 de enero de 2021

* alberto.ramos104@alu.ulpgc.es

Índice

1. Introducción	1
2. Descripción de los datos	1
2.1. Descripción de las columnas	1
2.2. Datos perdidos	2
2.3. Presencia de outliers	3
3. Metodología	6
3.1. Métodos aplicados	6
3.1.1. KMeans	6
3.1.2. CMeans	6
3.1.3. Normalización estándar	6
3.1.4. Validación	6
3.1.5. Método Elbow	6
3.1.6. Silhouette	6
3.1.7. TSNE	6
3.2. Implementacion	7
3.2.1. Clase <i>DataLoader</i>	7
3.2.2. Clase <i>KMeansCluster</i>	7
3.2.3. Clase <i>CMeansCluster</i>	7
3.2.4. TSNE	8
3.2.5. Silhouette	8
3.3. Preprocesamiento de los datos	8
4. Resultados	9
4.1. Clustering de energía por tipo de zona agrupando por ciudades	9
4.1.1. Zonas comerciales	9
4.1.2. Zonas residenciales	12
4.1.3. Zonas industriales	13
4.2. Clustering de gas por tipo de zona agrupando por ciudades	16
4.2.1. Zonas comerciales	16
4.2.2. Zonas residenciales	17
4.2.3. Zonas industriales	19
4.3. Clustering de energía observando patrones por bloques	21
4.3.1. Comercial	21
4.3.2. Residencial	22
4.3.3. Industrial	22
4.3.4. Comercial	24
4.3.5. Residencial	24
4.3.6. Industrial	25
4.4. Clustering de energía por media consumida en cada contador por comunidad	26
4.5. Clustering de gas por media consumida en cada contador por comunidad .	27
4.6. Clustering por edad, dimensión del hogar, ocupación total y ocupación de hogares rentados	28
4.7. Predicción de tipo de comunidad a partir del consumo energético utilizando <i>KMeans</i>	32

Índice de figuras

1.	Presencia de zonas con un alto consumo eléctrico.	3
2.	Varianza de consumo eléctrico mensual.	3
3.	Varianza de consumo eléctrico mensual.	4
4.	Presencia de zonas con un alto consumo de gas.	4
5.	Varianza de consumo de gas mensual.	5
6.	Varianza de consumo de gas mensual.	5
7.	Curva de inercia.	7
8.	Agrupamiento de muestras en cada clúster	10
9.	Muestras agrupadas en clusteres	11
10.	Curva <i>silhouette</i>	11
11.	Agrupamiento de muestras en cada clúster	12
12.	Muestras agrupadas en clusteres	13
13.	Curva <i>silhouette</i>	13
14.	Agrupamiento de muestras en cada clúster	14
15.	Muestras agrupadas en clusteres	14
16.	Curva <i>silhouette</i>	15
17.	Agrupamiento de muestras en cada clúster	16
18.	Muestras agrupadas en clusteres	17
19.	Curva <i>silhouette</i>	17
20.	Agrupamiento de muestras en cada clúster	18
21.	Muestras agrupadas en clusteres	18
22.	Curva <i>silhouette</i>	19
23.	Agrupamiento de muestras en cada clúster	19
24.	Muestras agrupadas en clusteres	20
25.	Curva <i>silhouette</i>	20
26.	Tendencia en cada clúster	21
27.	Tendencia en cada clúster	22
28.	Tendencia de consumo en cada bloque	23
29.	Tendencia en cada clúster	24
30.	Tendencia de consumo en cada bloque	25
31.	Tendencia de consumo	25
32.	Tendencia en cada clúster	26
33.	Tendencia en cada clúster	27
34.	Tendencia en cada clúster	29
35.	Tendencia en cada clúster	31

Índice de cuadros

1.	Porcentaje de filas perdidas	2
2.	<i>Accuracy</i> obtenido en la predicción	32

1. Introducción

Este trabajo está dividido en dos partes. En primer lugar, se ha realizado un estudio del consumo energético utilizando el dataset de consumo de energía en *Chicago* en 2010 [1]. En este estudio se han buscado patrones de consumo según distintos criterios. Por otro lado, se ha utilizado el modelo de *k-Means* para realizar la predicción del tipo de zona según el consumo mensual realizado.

Módulos de Python requeridos. En el fichero *requirements.txt* se encuentran todos los módulos instalados en el *virtualenvironment* utilizado para el desarrollo. Sin embargo, los módulos necesarios son:

- *sklearn*
- *skfuzzy*
- *pandas*
- *numpy*
- *matplotlib*
- *seaborn*
- *scipy*
- *plotly*
- *yellowbrick*

2. Descripción de los datos

En este trabajo se ha utilizado el *dataset Chicago Energy Usage 2010* obtenido del repositorio *data.world* [1]. En este dataset se describe el consumo mensual de electricidad y gas de cada bloque de las 77 comunidades de la ciudad de Chicago.

2.1. Descripción de las columnas

Entre las columnas más importantes, se encuentran:

- *COMMUNITY AREA NAME*: nombre de una comunidad de Chicago.
- *CENSUS BLOCK*: número censal. Los registros en blanco son zonas a las que no se agregaron información por privacidad.
- *BUILDING TYPE*: tipo de edificios en ese área (Residential, Commercial, Industrial). Los registros en blanco son zonas con número censal en blanco.
- *BUILDING_SUBTYPE*: subtipo de edificio en ese área (Single Family, Multi 1-7, Multi 7+, Commercial, Industrial, Municipal). Los registros en blanco son zonas con número censal en blanco.
- *KWH JANUARY 2010 ... KWH DECEMBER 2010*: KWH consumidos en cada mes en ese área.

- *TOTAL KWH*: Total de KWH consumidos
- *ELECTRICITY ACCOUNTS*: número de contadores en la zona. Cada contador no corresponde a un edificio.
- *ZERO KWH ACCOUNTS*: contadores que no han consumido energía.
- *THERM JANUARY 2010 ... THERM DECEMBER 2010*: therms consumidos en cada mes en ese área.
- *TOTAL THERMS*: consumo total de gas.
- *GAS ACCOUNTS*: contadores de gas.
- *TOTAL POPULATION*: tamaño de la población
- *TOTAL UNITS*: Número de viviendas/comercios/fábricas según el tipo de zona.
- *AVERAGE BUILDING AGE*: edad media de los edificios
- *AVERAGE HOUSESIZE*: dimensión media del hogar. Se obtiene dividiendo el número de personas en los hogares por el número de hogares.
- *OCCUPIED UNITS*: edificios ocupados
- *OCCUPIED UNITS PERCENTAGE*: porcentaje de edificios ocupados
- *RENTER-OCCUPIED HOUSING UNITS*: unidades rentadas
- *RENTER-OCCUPIED HOUSING PERCENTAGE*: porcentaje de unidades rentadas

2.2. Datos perdidos

Tanto para las columnas de electricidad como gas existen ciertas filas con datos perdidos. En la tabla 1 se muestra el porcentaje de valores perdidos para cada uno. De media un 1.29 % de datos de electricidad y un 2.98 % de datos de gas están vacíos.

	% filas perdidas kWh	% filas perdidas Therms
Enero	1.29	3.33
Febrero	1.29	6.31
Marzo	1.29	2.21
Abril	1.29	2.35
Mayo	1.29	2.77
Junio	1.29	2.64
Julio	1.29	2.71
Agosto	1.29	2.85
Septiembre	1.29	3.40
Octubre	1.29	2.57
Noviembre	1.29	2.33
Diciembre	1.29	2.30
Media	1.29	2.98

Cuadro 1: Porcentaje de filas perdidas

Como existen suficientes datos, se ha optado por eliminarlos. Las filas nulas para la electricidad y el gas no coinciden, por lo que la eliminación se realiza en el momento de la selección —en la clase *DataLoader*— con el fin de mantener la máxima cantidad de datos.

2.3. Presencia de outliers

Consumo de electricidad

Mediante el diagrama de cajas de la figura 1, se puede notar la presencia de ciertas zonas donde el consumo eléctrico es mucho mayor a la media.

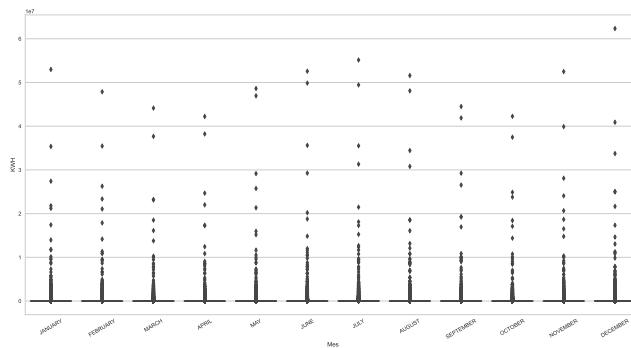


Figura 1: Presencia de zonas con un alto consumo eléctrico.

En la figura 2 se observa un aumento de la varianza en las épocas de verano e invierno, lo que quiere decir que aumenta la diferencia entre el consumo en distintas zonas. Complementando con los valores del diagrama de caja, podemos afirmar que el consumo aumenta en épocas veraniegas e invernales.

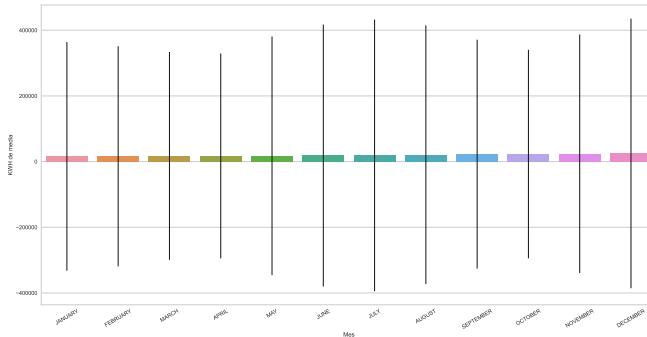


Figura 2: Varianza de consumo eléctrico mensual.

Aún así, según el valor medio se observa un crecimiento del consumo de energía (Figura 3). Sería interesante tener los datos de siguientes años para poder comprobar si esta tendencia continúa en los siguientes años.

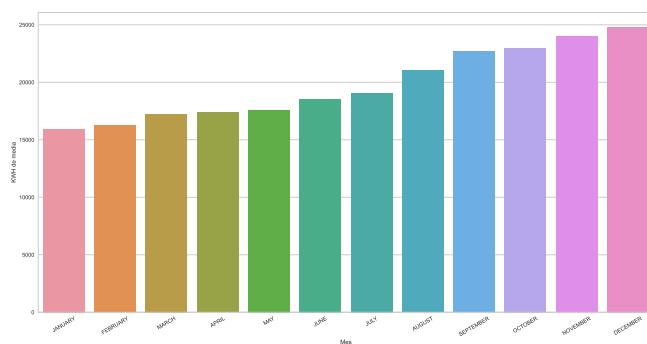


Figura 3: Varianza de consumo eléctrico mensual.

Consumo de gas

Mediante el diagrama de cajas de la figura 4, se puede notar la presencia de ciertas zonas donde el consumo de gas es mucho mayor a la media.

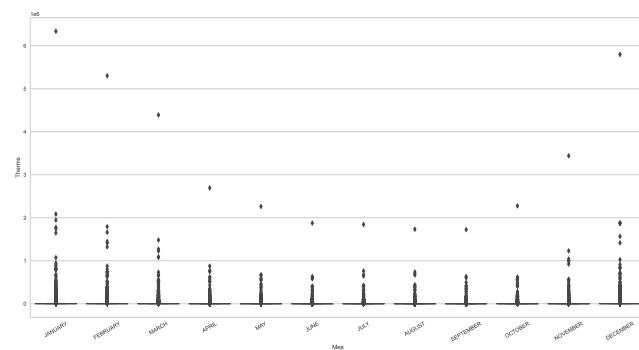


Figura 4: Presencia de zonas con un alto consumo de gas.

En la figura 5 se observa una disminución en la época de verano, lo que quiere decir que en invierno aumenta la diferencia entre el consumo en distintas zonas. Complementando con los valores del diagrama de caja, podemos afirmar que el consumo aumenta en épocas invernales (entre noviembre y marzo).

Aún así, según el valor medio se observa un crecimiento del consumo de gas (Figura 6). Sería interesante tener los datos de siguientes años para poder comprobar si esta tendencia continúa en los siguientes años.

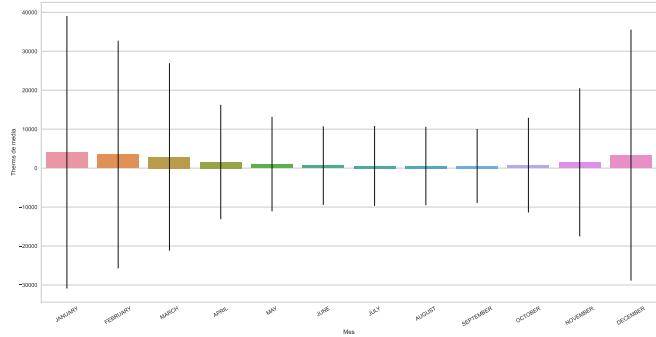


Figura 5: Varianza de consumo de gas mensual.

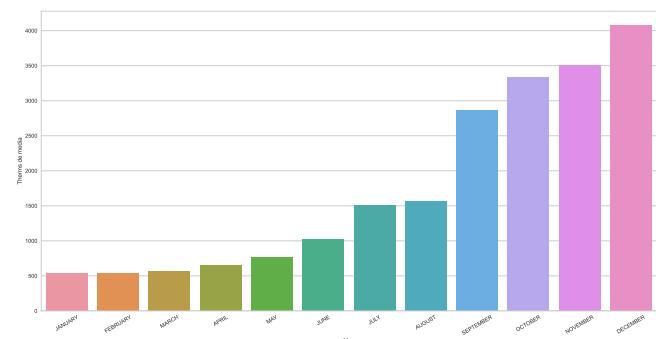


Figura 6: Varianza de consumo de gas mensual.

3. Metodología

3.1. Métodos aplicados

3.1.1. KMeans

Kmeans es un método de clustering cuyo objetivo es realizar la partición de muestras de datos en un número k de grupos —k es un valor arbitrario—. [2]

3.1.2. CMeans

Cmeans es un método de clustering difuso. De forma similar a CMeans, es capaz de dividir en C grupos a un conjunto de muestras, dando un valor de pertenencia a cada grupo para cada una de las muestras. [3]

3.1.3. Normalización estándar

Antes de aplicar *clustering*, realizamos una normalización de los datos utilizando la normalización estándar (1) implementada en la clase *StandardScaler* de *sklearn*. [4]

$$z = \frac{x - \nu}{\sigma} \quad (1)$$

3.1.4. Validación

3.1.5. Método Elbow

El método *elbow* es una técnica que facilita la selección del número óptimo de clústeres. El valor de inercia es la suma de distancias al cuadrado de cada muestra al centroide del clúster al que pertenece.(2)

$$\sum_{j=0}^k \sum_{i=0}^n (x_{ij} - c_j)^2 \quad (2)$$

Representando estos valores en una gráfica podemos estimar el mejor valor de k buscando el punto en el que se produce un cambio brusco en el decrecimiento del valor de inercia. Este punto es el llamado *punto codo*. Intuitivamente, en este punto se considera que aumentar el número de clústeres no disminuirá la distancia de las muestras al centroide. [5]

3.1.6. Silhouette

Silhouette es un método de validación de los agrupamientos realizados por una técnica de clustering. *Silhouette* mide que tan similar es una muestra a las muestras de su propio clúster comparada con otros clústeres.[6]

3.1.7. TSNE

t-Distributed Stochastic Neighbor Embedding (*t-SNE*) es una técnica que facilita la visualización de datos de alta dimensionalidad en dimensionalidades reducidas, preservando la distancia relativa entre las muestras. [7]

3.2. Implementacion

El código y generación de gráficas se encuentra dentro de los *notebooks*: *preprocessing.ipynb*, donde se realiza el preprocessado del *dataset*; *energy_usage_chicago_2010.ipynb*, donde se aplican los distintos métodos a distintos casos; y *prediccion_con_kmeans.ipynb*, donde se aplica *kmeans* con el objetivo de predecir el tipo de zona.

Adicionalmente, se han creado algunos módulos de apoyo, con el que simplificar el código en los *notebooks*.

3.2.1. Clase *DataLoader*

Esta clase se encuentra dentro del módulo *dataloader.py*. Mediante *DataLoader* se simplifica la selección de columnas de consumo mensual mediante las propiedades *energy_cols* y *gas_cols*. Además, esta clase, al solicitar una serie de columnas, devuelve solamente aquellas filas en las que no existe ningún valor nulo. Para ello, se utiliza el método *drop_na* de *pandas*.

3.2.2. Clase *KMeansCluster*

Esta clase se encuentra en el módulo *clustering/kmeans.py*. Utilizando la clase *KElbowVisualizer* del módulo *yellowbrick*, se evalúa el valor de inercia (*distortion score*) en un rango de valores de k pasado por parámetro. Este visualizador devuelve una gráfica (figura 7) donde se muestra el valor de inercia (línea azul) y de tiempo de convergencia (línea verde) para cada k. Además, selecciona el valor de k donde se encuentra el punto codo (línea discontinua negra).

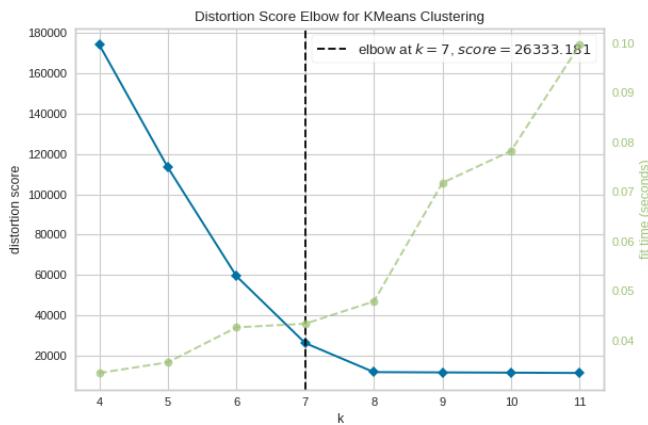


Figura 7: Curva de inercia.

Extraido de: <https://www.scikit-yb.org/en/latest/api/cluster/elbow.html>

En esta clase se utiliza el modelo *KMeans* de la librería *sklearn*. [8] Mediante el método *get_kmeans_of* se puede obtener el modelo *KMeans* del valor k pasado por parámetro.

3.2.3. Clase *CMeansCluster*

Esta clase se encuentra en el módulo *cluster/cmeans.py*. En este caso, se evalúa todas los valores de k para los datos dados. Mediante el método *get_scores*, se obtiene el valor de coeficiente de partición difuso para cada valor de k (*fuzzy partition coefficient* -

fpc). Este valor nos indica como de cohesionado están los clústeres entre si. El valor fpc oscila entre 0 y 1, siendo 1 el mejor resultado.

Una vez encontrada la mejor división, mediante el método `get_kmeans_of` se puede obtener el modelo `CMeans` de un k determinado.

3.2.4. TSNE

La visualización mediante *TSNE* se realiza mediante la función `tsne` del módulo *visualization*. Esta función se apoya en la clase *TSNE* de *sklearn* [9] para reducir a un valor dado de dimensiones una *matriz* recibida. Posteriormente se muestra mediante un gráfico *scatter*.

3.2.5. Silhouette

El análisis *silhouette* se ha aplicado utilizando el método `silhouette_visualizer` de la librería *yellowbrick* [10].

3.3. Preprocesamiento de los datos

En el *notebook preprocessing* se realiza un preprocesamiento de los datos.

Como se destacó en el apartado Descripción de los datos, de media, un 1.29 % de datos de consumo de electricidad y un 2.98 % de datos de gas están vacíos. Como son suficientes datos, se ha optado por eliminar estas filas nulas. Como no todas las filas nulas coinciden en índice, con el fin de mantener el máximo de datos de consumo, se ha optado por eliminar las filas nulas en el momento de la selección con la clase *DataLoader*.

Existe una única columna categórica que deberíamos convertir a numérico. Esta columna es *ELECTRICITY ACCOUNTS*, que contiene el número de contadores de cada zona. Esta columna mezcla valores numéricos y categóricos, por lo que se ha optado por convertir los valores categóricos a numérico. Por ejemplo, un valor catágorico llamado "Less than 4" se ha sustituido por el valor 3.

El resultado se ha volcado en un nuevo fichero csv llamado *energy-usage-2010-clean*.

4. Resultados

4.1. Clustering de energía por tipo de zona agrupando por ciudades

En este caso se ha calculado el valor medio de consumo eléctrico por cada una de las comunidades. En las muestras resultantes se han buscado las posibles agrupaciones de consumo para cada tipo de zona en la ciudad (comercial, industrial y residencial).

4.1.1. Zonas comerciales

En las zonas comerciales se han encontrado 4 grupos de consumo. En el clúster 1, se encuentra la comunidad *Loop*, cuyo consumo es el más alto. En el clúster 2, se encuentran *Near North Side* y *Near South Side*, que quedan en segundo lugar. El resto de comunidades se dividen entre los clústeres 0 y 3. Este resultado era de esperar, pues estas tres comunidades son las zonas centrales de *Chicago*, donde se encuentra la sede del gobierno, y una alta actividad cultural y comercial.

En el gráfico 9 podemos ver como se han agrupado las muestras más cercanas.

La curva *silhouette* obtenida nos indica que las muestras tienen una pertenencia alta a su clúster.

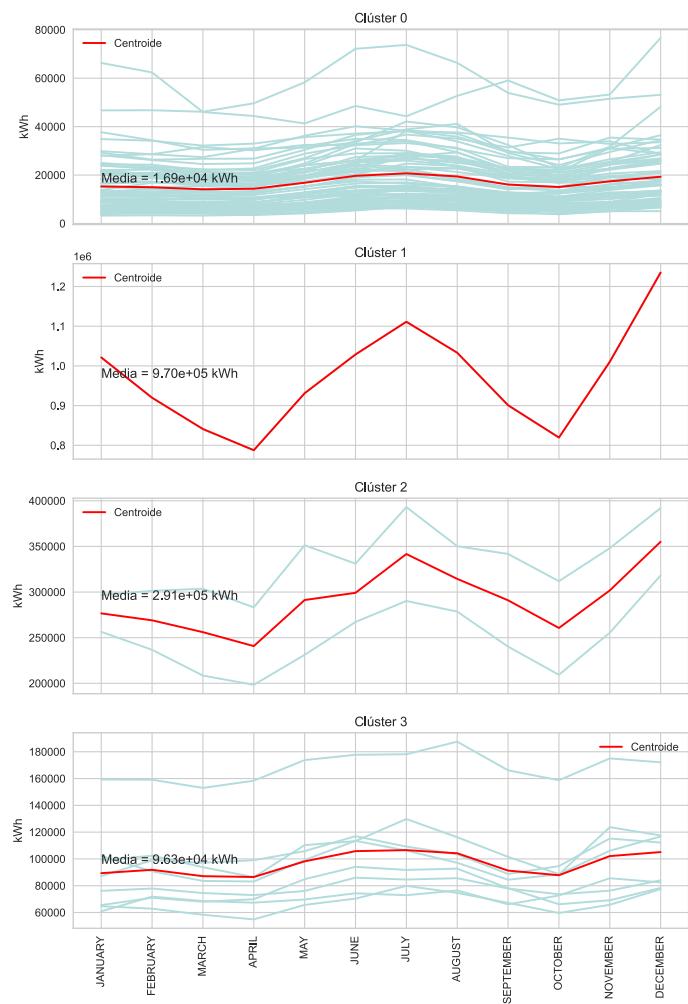


Figura 8: Agrupamiento de muestras en cada clúster

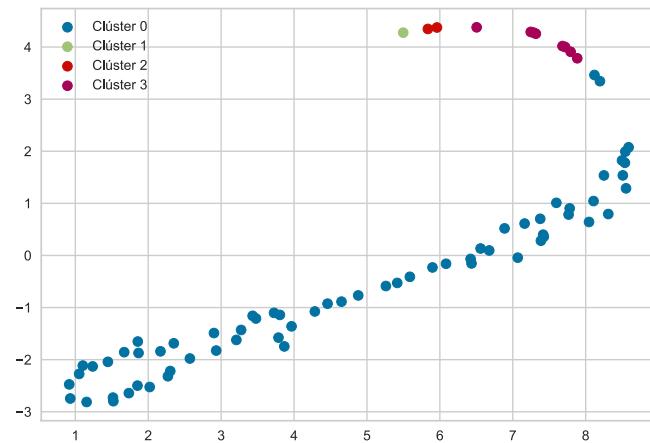


Figura 9: Muestras agrupadas en clusteres

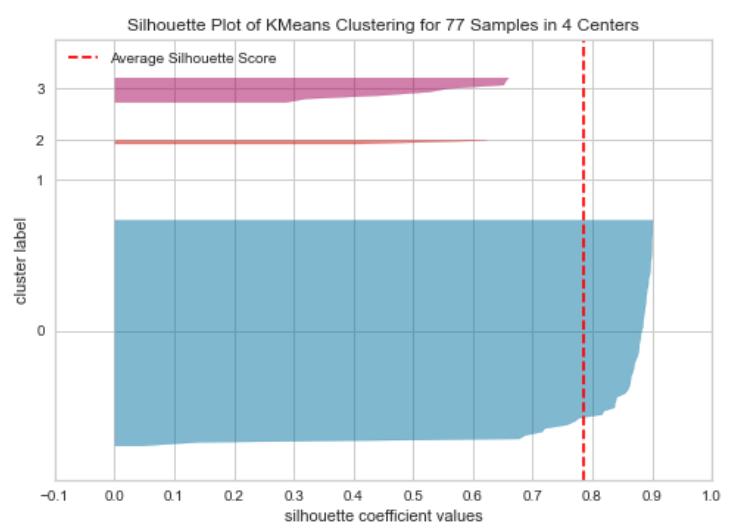


Figura 10: Curva *silhouette*

4.1.2. Zonas residenciales

En las zonas residenciales se han encontrado 4 grupos. Al igual que con las zonas comerciales, en el clúster 1 está *Loop* con un consumo medio más alto. En segundo lugar, en el clúster 2 están *Near North Side* y *Near South Side*.

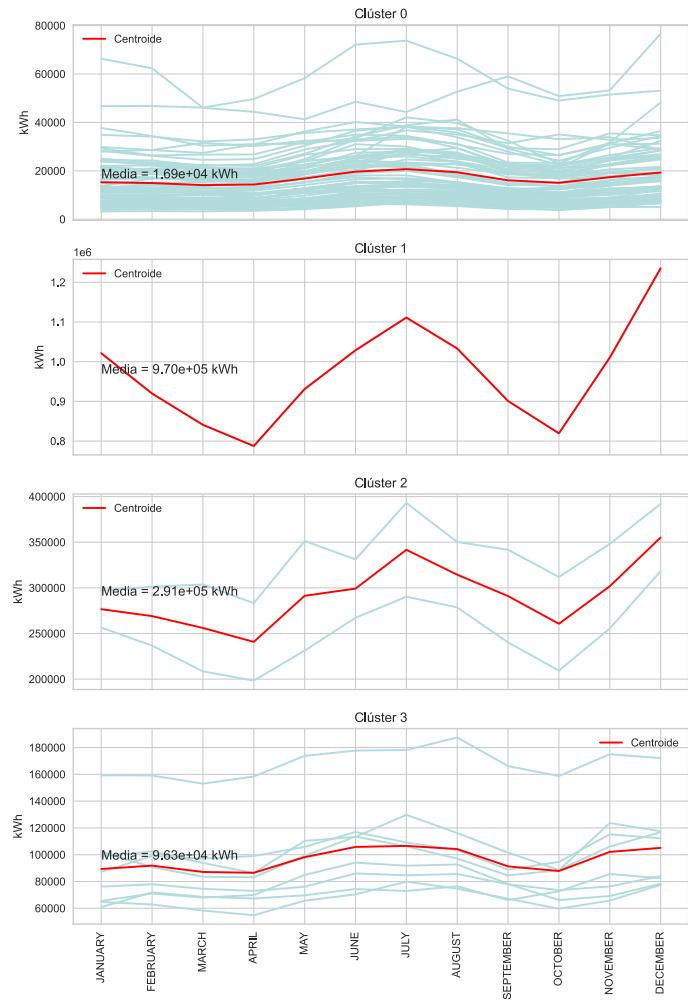


Figura 11: Agrupamiento de muestras en cada clúster

Como se observa en la curva 13, excepto pocas muestras del clúster 0, todas las muestras tienen un alto valor de pertenencia a su clúster.

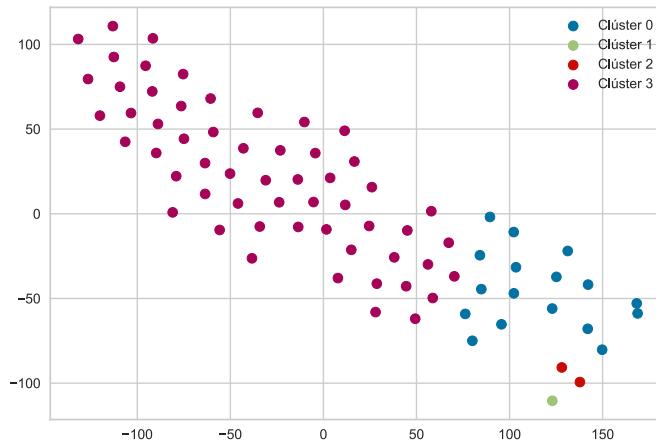


Figura 12: Muestras agrupadas en clusteres

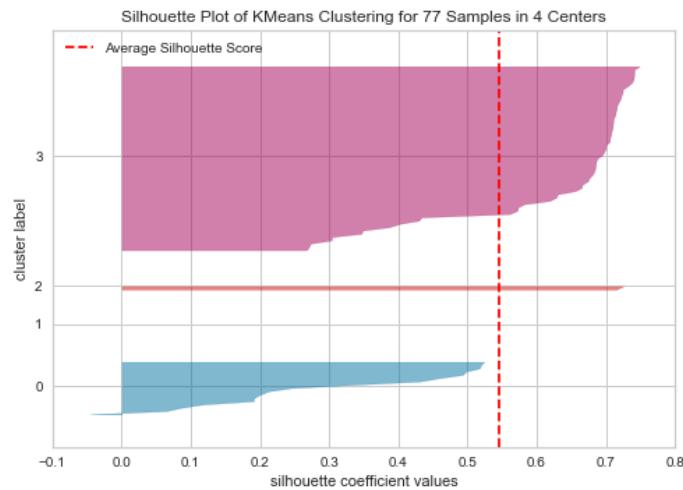


Figura 13: Curva *silhouette*

4.1.3. Zonas industriales

En las zonas industriales se han encontrado 3 grupos. En el clúster 1, con mayor consumo medio, se encontró a la comunidad *Near West Side*. En el clúster 2, en segundo lugar de mayor consumo está la comunidad *South Lawndale*.

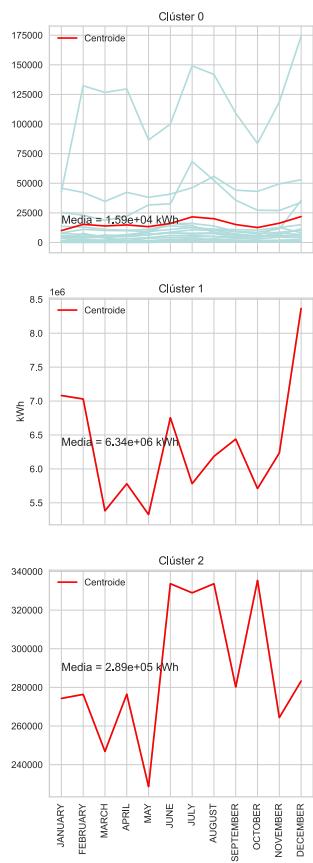


Figura 14: Agrupamiento de muestras en cada clúster

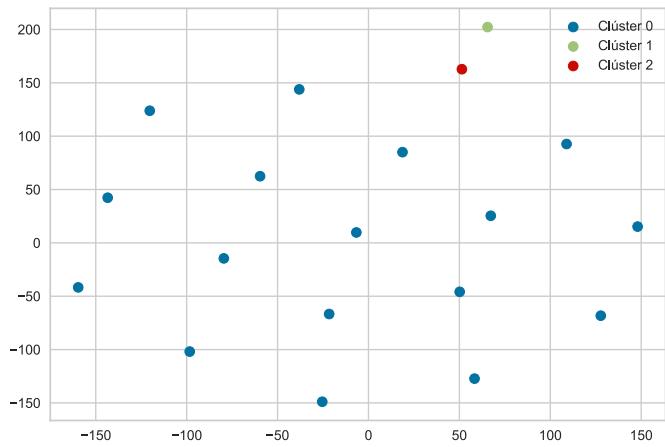


Figura 15: Muestras agrupadas en clusteres

La curva *silhouette* 16 indica una alta pertenencia de las muestras a su clúster.

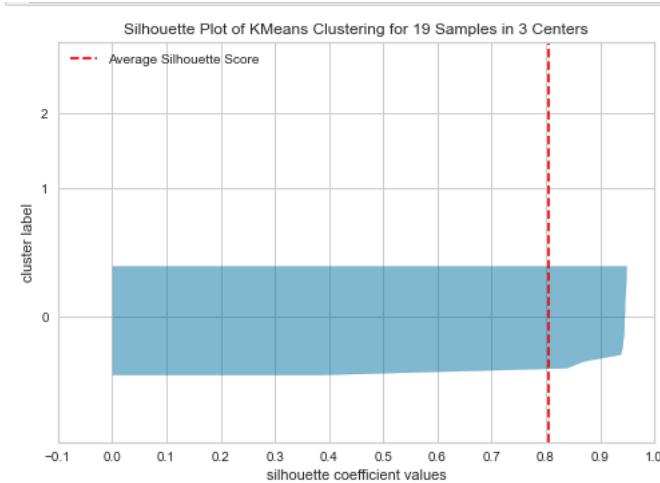


Figura 16: Curva *silhouette*

4.2. Clustering de gas por tipo de zona agrupando por ciudades

4.2.1. Zonas comerciales

Según el consumo de gas, se han encontrado 4 grupos de consumo en zonas comerciales. En el clúster 3 se encuentra *Loop*, con el consumo más alto. En segundo lugar están *Burnside*, *Near North Side*, *Pullman* y *Riverdale*. Se ha encontrado también que la disminución del consumo en épocas de verano es mucho mayor para los clústeres 2 y 3, que para los clústeres 0 y 1.

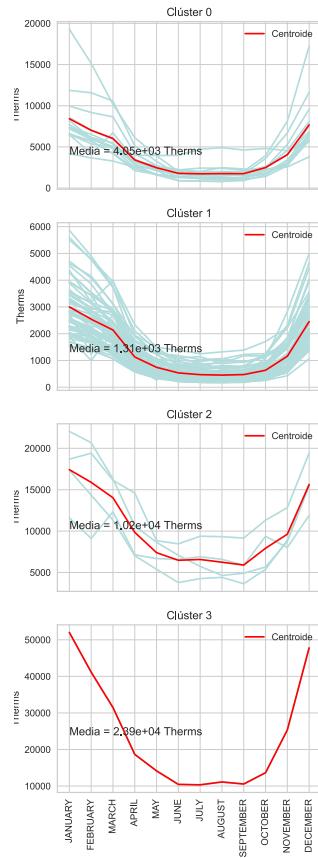


Figura 17: Agrupamiento de muestras en cada clúster

La curva de *silhouette* 19 nos indica una alta pertenencia para la mayoría de muestras, excepto para unas pocas del clúster 0.

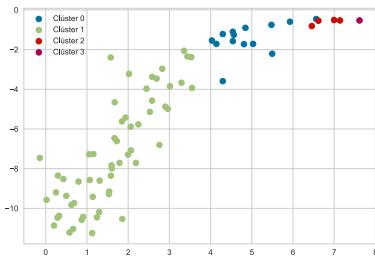


Figura 18: Muestras agrupadas en clusteres

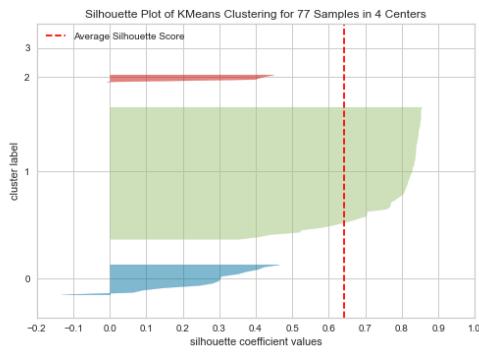


Figura 19: Curva *silhouette*

4.2.2. Zonas residenciales

En las zonas residenciales también se han encontrado 4 grupos. Entre los que más consumen están: *Loop*, en el clúster 2; y *Douglas*, *Hyde Park*, *Kenwood*, *Near South Side* y *O'Hare*, en el clúster 1.

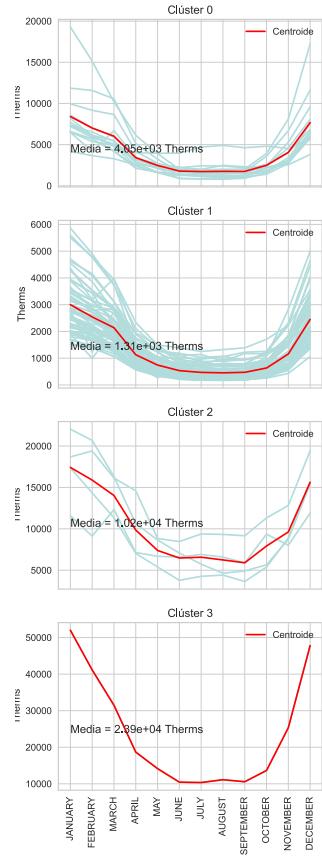


Figura 20: Agrupamiento de muestras en cada clúster

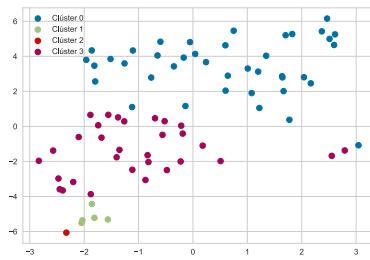


Figura 21: Muestras agrupadas en clusteres

La curva *silhouette* 22 nos indica que ciertas muestras del cluster 3 tienen una baja pertenencia a su clúster y alta a otros.

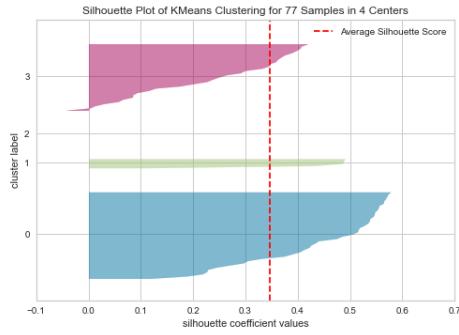


Figura 22: Curva *silhouette*

4.2.3. Zonas industriales

En las zonas industriales se han detectado 5 clústeres. Ordenados de mayor a menor consumo, las comunidades cuyas zonas industriales que más consumen son: *Calumet Heights*, en el clúster 1; *Ashburn*, en el clúster 2; *Chatham*, en el clúster 4; y *Chicago Lawn y Dunning*, en el clúster 3.

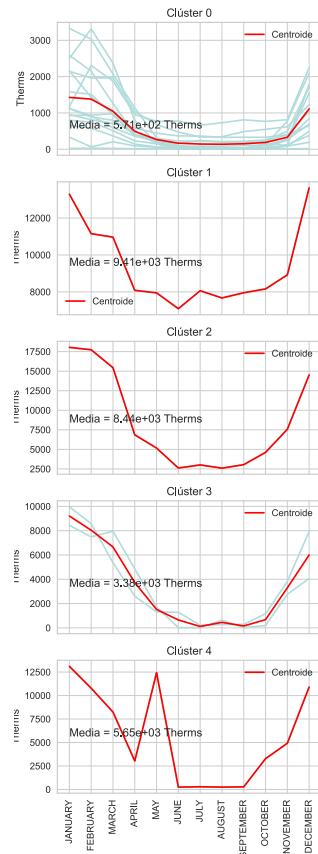


Figura 23: Agrupamiento de muestras en cada clúster

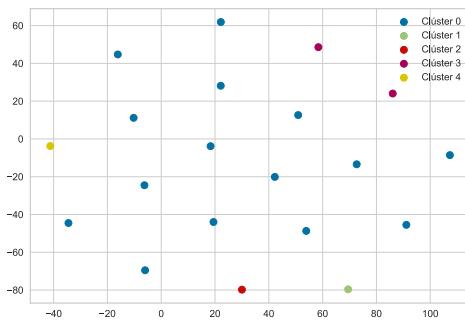


Figura 24: Muestras agrupadas en clusteres

La curva *silhouette* nos muestra un valor de pertenencia más alto para el clúster 0 que el 3. Para los clústeres 1, 2 y 4 no muestra valores porque solamente tienen una muestra.

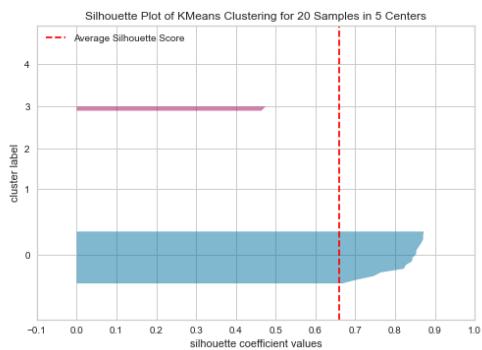


Figura 25: Curva *silhouette*

4.3. Clustering de energía observando patrones por bloques

En este caso seleccionaremos la comunidad con más consumo para cada tipo encontrada anteriormente, y estudiaremos los patrones de consumo entre todos los bloques de dicha comunidad para cada tipo de comunidad (comercial, residencial e industrial).

4.3.1. Comercial

Las zonas comerciales de *Loop* son las que más consumieron electricidad de media. Por lo tanto, seleccionaremos los bloques de dichas zonas para buscar patrones de consumo.

En este caso, se encontraron 3 grupos. El primer grupo (cluster 2) con el consumo más alto tiene un consumo medio de $1e7$ kWh (en el que se encuentran dos bloques censales). El segundo (cluster 1) tiene un consumo promedio de $2.6e6$ kWh. Por último, el clúster 0 tiene el consumo promedio más bajo, con un valor de $4.3e5$ kWh. Además, puede notarse que el consumo en el clúster 1 está más estabilizado que en el clúster 2 (salvo una subida a mitad de año de una muestra.) Se puede observar que, en el clúster 1, el rango entre el menor y mayor valor de su centroide es menor que en el centroide del clúster 2. Lo mismo ocurre en el clúster 0, el consumo es más estable que en el clúster 1 y 2.

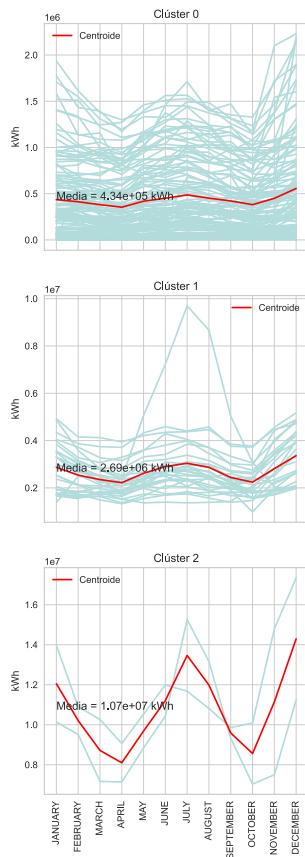


Figura 26: Tendencia en cada clúster

4.3.2. Residencial

Para las zonas residenciales también seleccionamos *Loop* al ser la que más consumió.

En este caso tenemos pocas muestras, por lo que el método *elbow* no es capaz de detectar un valor de k óptimo. Igualmente, seleccionando un valor de k igual a 3, conseguimos agrupar por la cantidad promedio que consume cada bloque. Además, el aumento de consumo en los clústeres 0 y 2 es más notable que en el clúster 1, que se mantiene casi estable durante el año para la muestra con un menor consumo. Esa diferencia en la época de verano crece a medida que crece el consumo promedio.

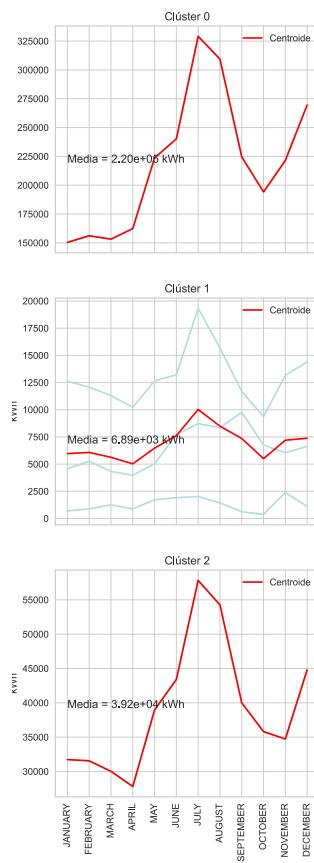


Figura 27: Tendencia en cada clúster

4.3.3. Industrial

En este caso seleccionamos a *Near West Side* por tener las zonas industriales que más electricidad consumen. Como solamente tenemos 3 muestras, visualizamos sus gráficas sin aplicar ningún agrupamiento.

Vemos como en el primer bloque el consumo es más alto a principio y final de año, a diferencia de los otros dos bloques, donde se registran consumos más altos a mitad de año.

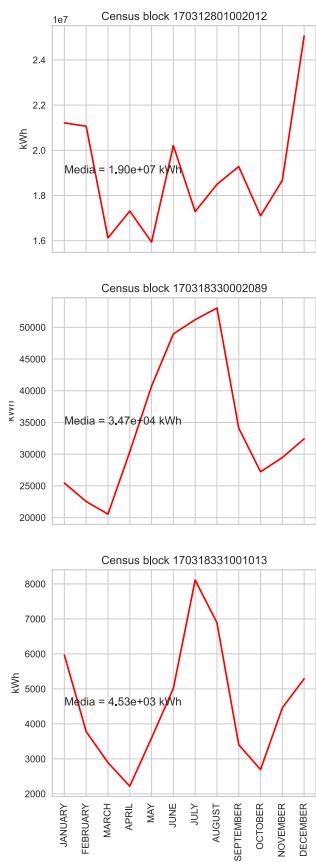


Figura 28: Tendencia de consumo en cada bloque

Igual que antes, seleccionaremos la comunidad con mayor consumo de gas, y buscaremos patrones de consumo entre sus bloques censales (comercial, residencial e industrial).

4.3.4. Comercial

Seleccionamos a *Loop* por ser la comunidad con zonas comerciales que más gas consumieron. En este caso, hemos encontrado 4 grupos. La tendencia en los clústeres 0, 2 y 3, es que a medida que disminuye el consumo medio, es más estable durante el año, siendo la disminución en verano menos notable. Además, vemos que en el clúster 1, la tendencia es justo la contraria que a los demás clústeres, en épocas de verano el consumo aumenta. Sería necesario conocer qué tipo de comercio existe en las muestras del clúster para observar si existe alguna explicación, o en su defecto, ha sido un error de medida.

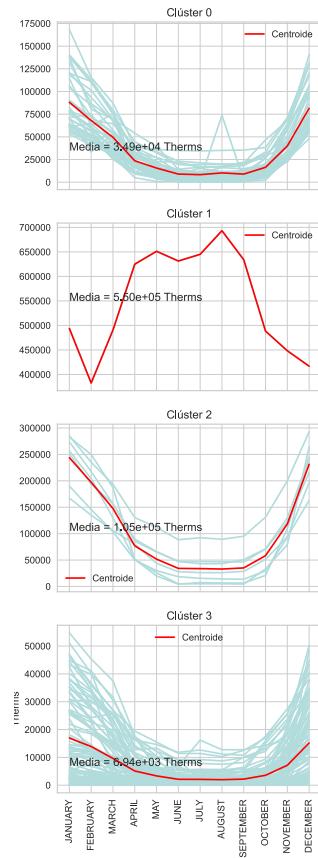


Figura 29: Tendencia en cada clúster

4.3.5. Residencial

Loop también ha sido la comunidad cuyas zonas residenciales han consumido más gas. En la gráfica 30 se muestra la tendencia de consumo de cada bloque.

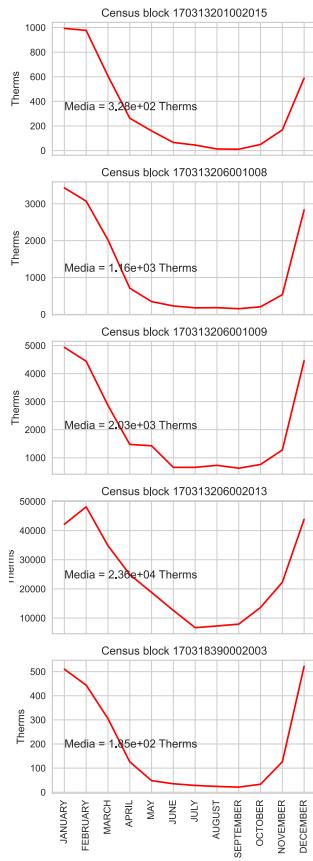


Figura 30: Tendencia de consumo en cada bloque

4.3.6. Industrial

En este caso, *Ashburn* ha sido la comunidad que más gas ha consumido de promedio. En la gráfica 31 se muestra la tendencia de consumo.

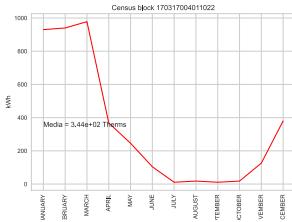


Figura 31: Tendencia de consumo

4.4. Clustering de energía por media consumida en cada contador por comunidad

En este caso, en vez de tomar el consumo total de cada comunidad, calcularemos la media de consumo en cada contador dividiendo el consumo total en cada zona entre el número de contadores que hay en dicha zona.

Como resultado se han encontrado cuatro grupos. En el clúster 1, donde está la comunidad de *Loop*, en los meses de verano e invierno se experimenta una subida del consumo mucho más notable que en el resto de comunidades. En el clúster 2 tenemos a la comunidad *Riverdale*, que está en el segundo puesto como la comunidad que más electricidad consume.

A diferencia, en el clúster 0 y 3, el consumo es más estable durante todo el año. En el clúster 3 están las comunidades *Archer Heights*, *Near North Side*, *Near South Side* y *Woodlawn*. Las 71 comunidades restantes se encuentran en el clúster 0.

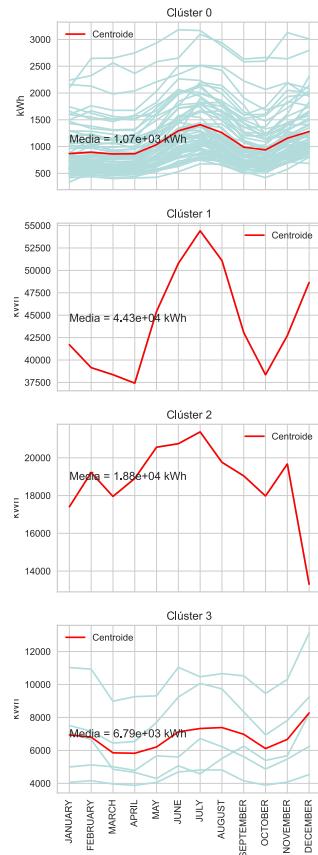


Figura 32: Tendencia en cada clúster

4.5. Clustering de gas por media consumida en cada contador por comunidad

Igual que en el apartado previo, calcularemos la media de consumo en cada contador dividiendo el consumo total en cada zona entre el número de contadores que hay en dicha zona.

Se han encontrado 4 grupos. En el clúster 1 está la comunidad de *Loop*, la cuál es la que más gas consumió. En segundo lugar está el clúster 2 están las comunidades *Near North Side* y *Riverdale*. En siguiente lugar están los clústeres 0 y 3.

A medida que baja el consumo, la diferencia entre el punto máximo y mínimo de consumo es menor. Aproximadamente, observando los valores del gráfico, la diferencia en el clúster 1 es de 8000 therms, en el clúster 2 es de 1500 therms, en el clúster 0 es de 1000 therms, y en el clúster 3 es de 300 therms.

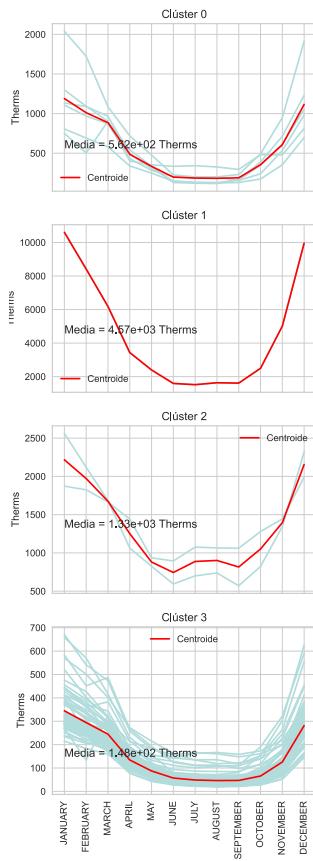


Figura 33: Tendencia en cada clúster

4.6. Clustering por edad, dimensión del hogar, ocupación total y ocupación de hogares rentados

En este caso, se ha aplicado clustering a las muestras de consumo de electricidad y gas, y se han comparado los grupos obtenidos con la edad media de los edificios, con la dimensión del hogar, el porcentaje de ocupación total y de hogares en renta. Se han seleccionado solamente las muestras de tipo residencial.

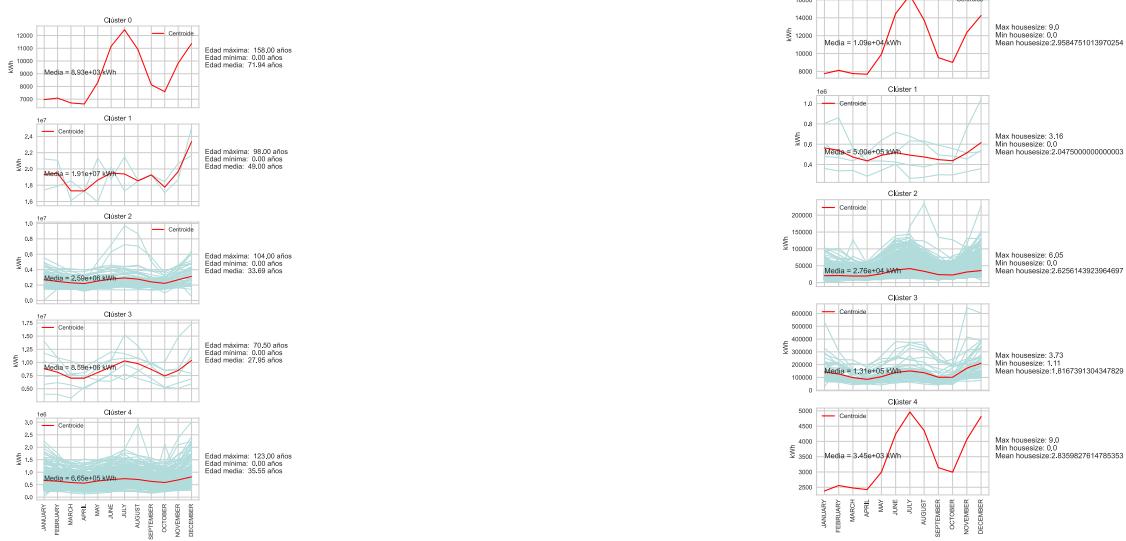
Electricidad

Se han encontrado cinco grupos de consumo de electricidad. En la gráfica 34 se puede observar la tendencia de consumo de electricidad por cada uno de los grupos encontrados.

Por edad de edificio, existen 3 grupos con una edad media similar —y de las más bajas— cuyo consumo es aproximadamente constante durante el año. Estos clústeres son el 2, 3 y 4, con una edad media de 33.69, 27.95 y 35.55 años, respectivamente. En el clúster 1 hay dos zonas que tienen el consumo más alto, y su edad media es de 49 años. Estas zonas son *Near West Side* de tipo industrial, y *Near South Side* de tipo comercial. En el clúster 0 están las zonas que menos consumen y con una edad media mayor que otros clústeres (71.94 años).

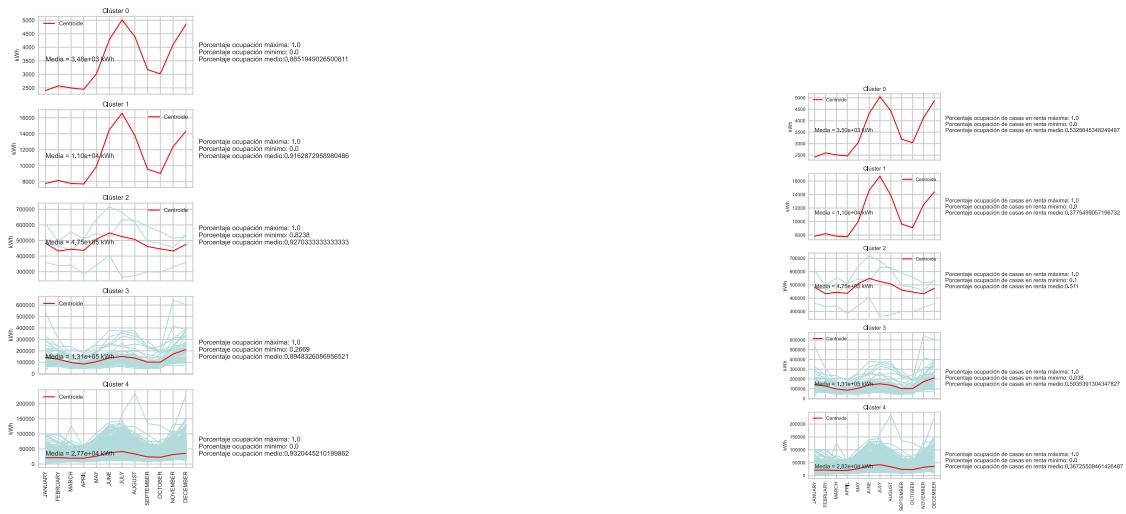
Como se describe en el apartado de Descripción de los datos, la dimensión del hogar es una medida de la relación entre las personas que ocupan hogares y el número de hogares en una zona. Los grupos 1, 2 y 3 tienen un consumo constante durante el año, cuya dimensión de hogar es similar (2.05, 2.63 y 1.82, respectivamente). En los grupos 0 y 1, el aumento de consumo en la época de verano e invierno es mayor, siendo la medida de dimensión de hogar mayor que para los otros grupos. Como conclusión, en las zonas donde existe una mayor dimensión de hogar el aumento es mayor, aunque el consumo medio en el año no es notablemente mayor que en otros grupos.

En cuanto a la ocupación, el clúster 2 es el que más consumo de electricidad promedio realiza. Su ocupación total está en segundo lugar y aproximadamente la mitad de las casas en alquiler. Los clústeres 2, 3 y 4 tienen un consumo promedio similar durante el año, además de ser los más altos. Su ocupación total es también la más alta.



(a) Tendencia por edad de edificio.

(b) Tendencia por dimensión de hogar.



(c) Tendencia por ocupación total.

(d) Tendencia por ocupación en renta.

Figura 34: Tendencia en cada clúster

Gas

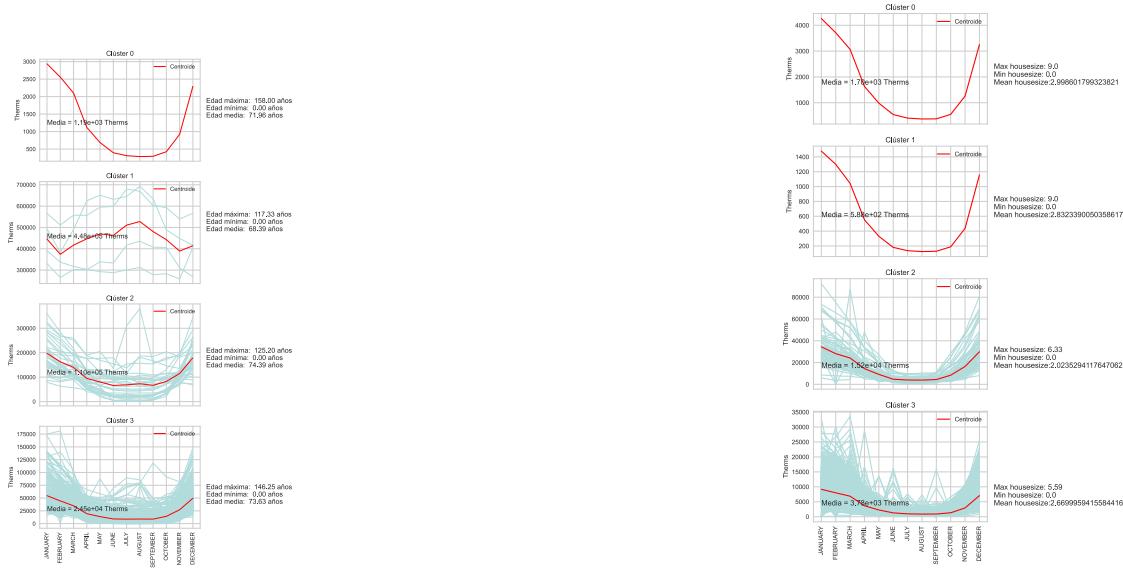
Se han encontrado cuatro grupos de consumo de electricidad, excepto para la ocupación en 35c y 35d, en la que se encontraron cinco grupos. En la gráfica 35 se muestra la tendencia de consumo de gas en cada uno de los grupos.

En el cluster 1 de 35a podemos observar la tendencia que habíamos descubierto anteriormente en apartados anteriores. Existen ciertas zonas cuya tendencia es inversa a la del resto, su consumo aumenta justo en épocas de verano. Estas zonas, de media, tienen una edad menor a la del resto, y la muestra con una edad mayor en este clúster también es menor a las edades máximas en otros clústeres.

En cuanto a la dimensión de hogar, se ha encontrado que el clúster 2, cuyo consumo es el máximo, tienen la menor dimensión de hogar —a diferencia de lo que ocurría con el consumo de electricidad—. Además, este clúster es el que menos muestra tiene, por lo que existen zonas con una población menor que consume mucho más que otras zonas con mayor población —el consumo por habitante es mayor—.

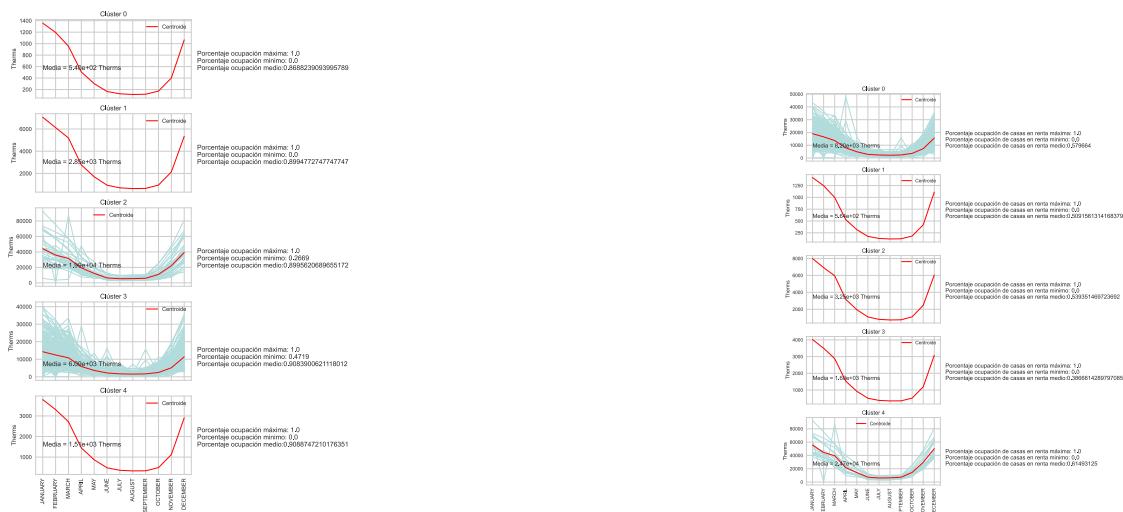
En 35c se han encontrado cinco grupos, aunque observando los datos de consumo medio y de ocupación, podríamos agrupar los clústeres 1 y 4.

Comparando los clústeres 1 y 3 en 35d, se observa que en el grupo 3 el consumo medio es más del doble que para el grupo 2, siendo la ocupación media de casas en alquiler mucho menor. Esto refleja que en el grupo 3 existe un consumo mayor por casa en alquiler que para el grupo 1.



(a) Tendencia por edad de edificio.

(b) Tendencia por dimensión de hogar.



(c) Tendencia por ocupación total.

(d) Tendencia por ocupación en renta.

Figura 35: Tendencia en cada clúster

4.7. Predicción de tipo de comunidad a partir del consumo energético utilizando *KMeans*

Utilizando *KMeans* es posible realizar clasificaciones de datos. Como se puede comprobar en los resultados obtenidos, para poder utilizar *KMeans* como clasificador, es necesario mantener las proporciones de los datos.

En la tabla 2 se muestran los resultados obtenidos.

		Electricidad	Gas
Sin balancear	Train	74.68 %	76.04 %
	Test	75.98 %	76.28 %
Balanceando	Train	37.14 %	36.14 %
	Test	50 %	10 %

Cuadro 2: *Accuracy* obtenido en la predicción

Referencias

- [1] @cityofchicago. Energy usage 2010. <https://data.world/cityofchicago/energy-usage-2010>.
- [2] Mathworks. k-means clustering. <https://es.mathworks.com/help/stats/k-means-clustering.html?lang=en>.
- [3] Mathworks. Fuzzy clustering. <https://es.mathworks.com/help/fuzzy/fuzzy-clustering.html>.
- [4] Scikit learn. sklearn.preprocessing.StandardScaler. <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>.
- [5] Ricardo Moya. Machine learning en python, con ejemplos. selección del número óptimo de clusters. <https://jarroba.com/seleccion-del-numero-optimo-clusters/>.
- [6] @soumya7. Silhouette index – cluster validity index — set 2. <https://www.geeksforgeeks.org/silhouette-index-cluster-validity-index-set-2/>.
- [7] Andre Violante. An introduction to t-sne with python example. <https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1>.
- [8] Scikit learn. sklearn.cluster.KMeans. <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>.
- [9] Scikit learn. sklearn.manifold.TSNE. <https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html>.
- [10] yellowbrick module. <https://www.scikit-yb.org/en/latest/>.