

Estudio de INP en México

Sarahi García, Ramón Ruíz*

Centro de Investigación en Matemáticas CIMAT**

15 de Abril de 2024

En este trabajo, realizamos visualización y análisis de datos multidimensionales de los precios de algunos productos que proporciona una las herramientas del INPC del INEGI. Generamos una pequeña base datos con productos específicos, incluyendo limón, leche, camarón, tortillas, electricidad, alquiler, frijol y pollo, en el periodo Agosto-2018 a Febrero-2024. Aplicamos análisis de componentes principales (PCA) para reducir la dimensionalidad de los datos y proyectarlos en un espacio tridimensional. Luego, empleamos el algoritmo K-means para agrupar los datos en un número predeterminado de clusters. Posteriormente, generamos histogramas para cada variable en función de los clusters obtenidos, lo que permite explorar la distribución de los datos dentro de cada cluster. Pudimos observar una cierta estructura en los datos relacionada con el año y el tipo de ciudad al que pertenecen.

I. INTRODUCCIÓN

El Índice Nacional de Precios al Consumidor (INPC) es un indicador de la evaluación y comprensión de la dinámica económica del país. Esta diseñado para estimar la evolución de los precios de los bienes y servicios consumidos por las familias de México.

Existe una complejidad inherente a la medición de las variaciones de precios pues, entre otros factores, existe una amplia gama de artículos y servicios de consumo así como una constante fluctuación en los precios. Además, la asincronía en los cambios de precios y sus variaciones de velocidad agregan un nivel adicional de complejidad al análisis.

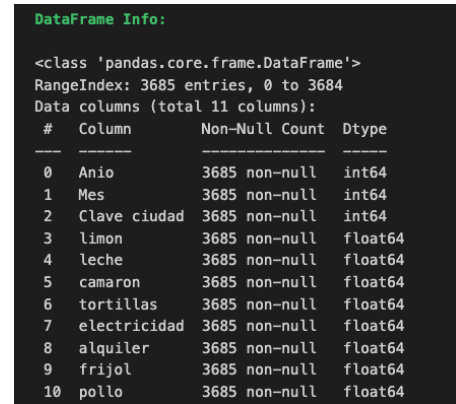
En este contexto, el presente se enfoca en analizar los precios de una amplia gama de productos en diversas ciudades de México. Utilizando datos muestrales, exploramos las fluctuaciones en los precios de nueve productos a lo largo del tiempo y en diferentes ubicaciones geográficas del país. Empleamos técnicas de análisis estadístico y visualización de datos para identificar patrones y relaciones significativas en los datos.

La información recopilada para este análisis proviene de la base de datos proporcionada por una de las herramientas de del INPC del INEGI: **Consulta precios promedio**, que contiene registros de precios de distintas marcas y/o proveedores de una serie de productos y servicios en 55 ciudades que se consumen en todo el país. Las cotizaciones de los productos en esta base de datos son principalmente mensuales y se encuentran desde el mes de Agosto del año 2018.

Hay una extensa cantidad de productos y servicios dis-

ponibles en la página del INEGI, pero nosotros nos centraremos exclusivamente en los siguientes productos mostrados en 1, desde el mes de agosto de 2018 hasta el mes de marzo de 2024 de las 55 ciudades. Para un mejor manejo y representación visual de los datos se modificaron, y/o quitaron algunas de las filas, de modo que para cada producto/servicio tenemos un único proveedor/marca.

El procedimiento de cómo llegamos a esta única tabla de datos se encuentra en el notebook `preprocesamiento.ipynb`, en 1 se encuentra una pequeña tabla con la info de las variables que utilizaremos.



```
DataFrame Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3685 entries, 0 to 3684
Data columns (total 11 columns):
#   Column          Non-Null Count  Dtype  
---  ---
0   Anio             3685 non-null   int64  
1   Mes              3685 non-null   int64  
2   Clave ciudad     3685 non-null   int64  
3   limon            3685 non-null   float64 
4   leche            3685 non-null   float64 
5   camaron          3685 non-null   float64 
6   tortillas        3685 non-null   float64 
7   electricidad     3685 non-null   float64 
8   alquiler         3685 non-null   float64 
9   frijol           3685 non-null   float64 
10  pollo            3685 non-null   float64
```

Figura 1. DataFrame Info

II. ANÁLISIS EXPLORATORIO

Productos

Sobre las variables continuas y cuantitativas (los precios de cada producto), algunas de las características que nos interesan son su promedio, media, máximo, mínimo, etc,

* yesenia.garcia@cimat.mx, ramon.ruiz@cimat.mx

** REP, a cargo de Dr. Johan Van horebeek



que se muestra en la tabla 2.

	count	mean	median	std	min	25%	50%	75%	max
limon	3685	30.468	28.14	13.676	3	21.4	28.14	36.58	94.23
leche	3685	21.462	20.88	6.091	9.6	17.57	20.88	24.5	60
camaron	3685	242.073	228	69.458	100	191.75	228	279	499
tortillas	3685	18.562	18	4.372	9.94	15.5	18	22	30
electricidad	3685	364.326	405.71	113.222	0	243.95	405.71	450.71	540.59
alquiler	3685	106.412	105.2	5.465	99.32	102.18	105.2	109.02	135.78
frijol	3685	34.153	32	11.449	12.75	25.78	32	40	79.89
pollo	3685	79.872	75.75	35.131	15.7	55	75.75	99.9	194

Figura 2. DataFrame Info

Estas estadísticas son sobre el total de datos, es decir, toma en cuenta todas las ciudades, y todos los meses de los 4 años. Más adelante veremos unos gráficos para ciertas ciudades y/o a lo largo del tiempo.

El limón, el frijol, la leche, el alquiler siguen una distribución unimodal y presentan una asimetría hacia la derecha. A continuación se muestran el histograma y boxplot del limon 3 y el alquiler 4.

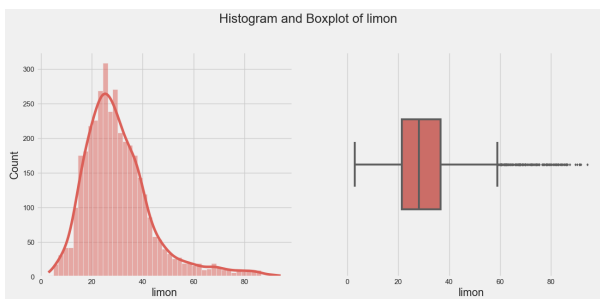


Figura 3. Histograma y Boxplot de precios del Limón

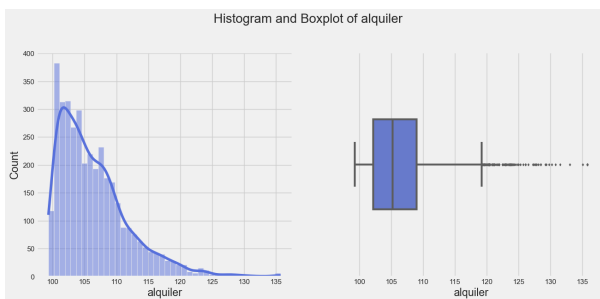


Figura 4. Histograma y Boxplot de precios del alquiler

El camarón presenta una distribución bimodal 5 y los demás productos (electricidad tortillas pollo) presentan distribuciones con una gran varianza (véase tabla 2), por ejemplo, en el caso de las tortillas 6, el precio mínimo es de 9.94, el máximo es 30 y su desviación es 4.37, aproximadamente de un quinto del rango.

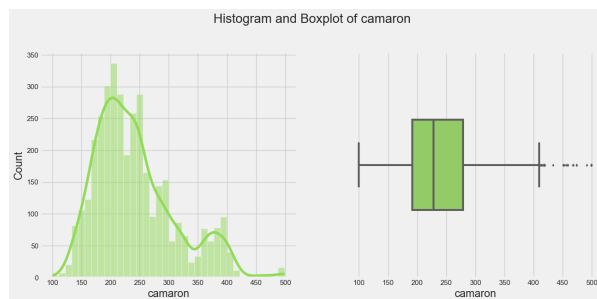


Figura 5. Histograma y Boxplot de precios del camaron

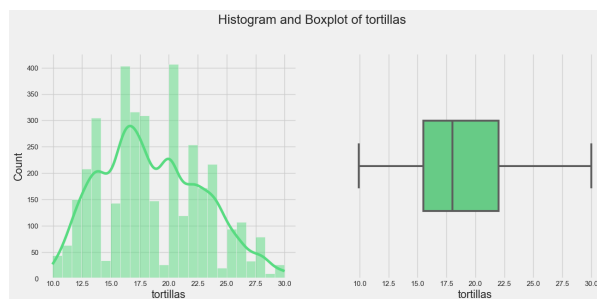


Figura 6. Histograma y Boxplot de precios del tortillas

Los demás histogramas y boxplots pueden consultarse en el notebook `analisis_exploratorio.ipynb`.

Precios en el tiempo

Otra característica interesante sobre los precios de los distintos productos es su evolución en el tiempo. Los siguientes cuatro gráficos muestran esto para cada ciudad. En general, el rango de los precios en función de las ciudades (para un punto fijo en el tiempo) es grande a excepción del limón y la electricidad.

En estas dos imágenes podemos apreciar que todas las gráficas (cada una asociada a una cd. distinta) siguen un patrón bastante definido.

El precio del limón 7 tiene un máximo y un mínimo globales bien definidos por año y sólo unas cuantas ciudades no se apegan a este comportamiento.

Mientras que el precio de la electricidad 8 alcanza un mínimo y luego un máximo de manera periódica, además en estos precios se mantiene por tiempo para volver a subir o bajar después.

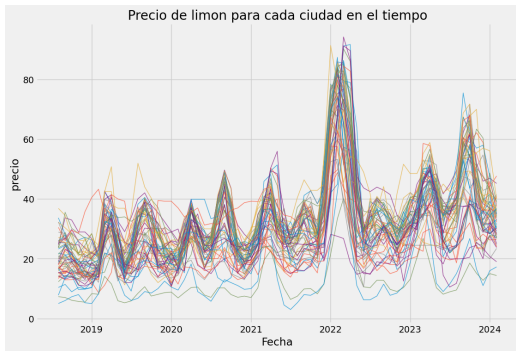
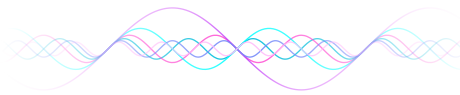


Figura 7. Grafico de la variación de precios del limon para cada ciudad en el periodo agosto de 2018 a febrero de 2024

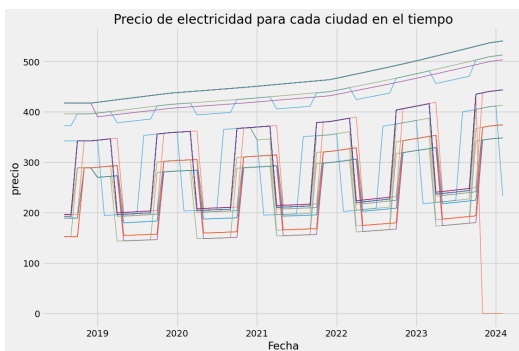


Figura 8. Grafico de la variación de precios del electricidad para cada ciudad en el periodo agosto de 2018 a febrero de 2024

Un caso especial es el del alquiler donde también se muestra una tendencia similar en todas las ciudades, pareciera que los precios suben de manera msomenos lineal y las gráficas parecen no decrecientes. Sin embargo, a diferencia del limón y la electricidad, a medida que avanza el tiempo el rango de precios (en función de las ciudades y a tiempo fijo) es más grande.

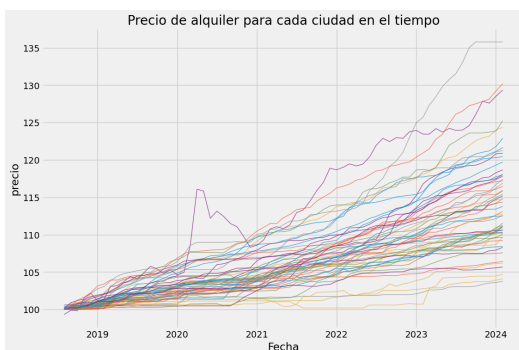


Figura 9. Grafico de la variación de precios del alquiler para cada ciudad en el periodo agosto de 2018 a febrero de 2024

Como mencioamos antes, el rango de precios en ls distintas ciudades de los demás productos es bastante amplio en todo momento y su evolución no parece seguir ningún patrón en conjunto. Como ejemplo de estos casos, se muestra el gráfico de las tortillas 10.

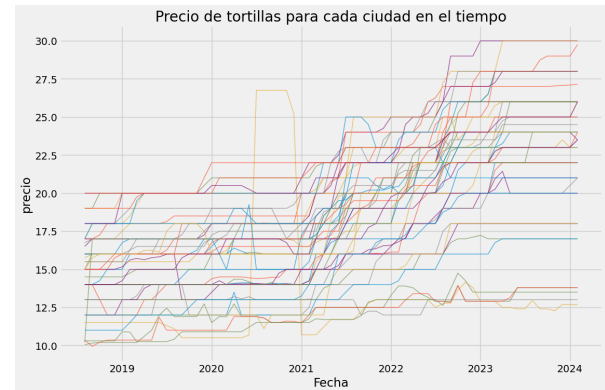


Figura 10. Grafico de la variación de precios del tortillas para cada ciudad en el periodo agosto de 2018 a febrero de 2024

Las demás series (pollo, camarón, leche y frijol) de tiempo tienen un comportamiento anpalogo al de las tortilla y pueden consultarse en el notebook `analisis_exploratorio.ipynb`.

Ciudades

Con los datos disponibles se realizó un resumen obteniendo el promedio de los precios de cada producto agrupando por ciudad (es decir, tomando en cuenta todos los meses de los 4 años para cada ciudad) y sumando estos valores se categorizó a cada ciudad en Caro, Medio o Barato. Se determino si una ciudad pertenecía o no a una categoría usando como referencia los cuantiles de la suma antes mencionada.

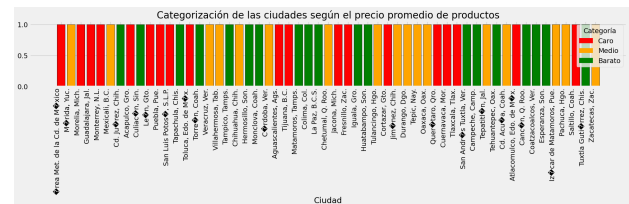
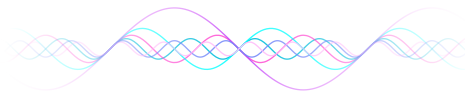


Figura 11. Grafico de la variación de precios del ciudades para cada ciudad en el periodo agosto de 2018 a febrero de 2024

Podemos ver que entre las ciudades más costosas se encuentran algunas como la Ciudad de México, Guadalajara, Monterrey, entre otras ciudades grandes y entre las ciudades más baratas tenemos ciudades más pequeñas



como Huatabampo, La Paz, Campeche, Colima, etc. Las categorías asignadas coinciden con las que se podrían esperar dadas las naturalezas de las respectivas ciudades y sus estilos de vida.

Esta clasificación se agregó como variable categórica a nuestra base de datos para realizar algunas pruebas de clustering más adelante.

III. RESULTADOS

Buscamos alguna clasificación de los datos que pueda ayudarnos a discernir alguna estructura en estos por lo que comenzamos definiendo un vector aleatorio con los precios de los 8 productos y 4 variables aleatorias discretas, una por cada etiqueta disponible en nuestra base de datos: Año, Estación, Mes y Categoría. (Categoría se refiere a la clasificación de la ciudad [11](#))

```
#definimos un vector aleatorio con los precios de los 8 productos
X=df[["limon", "leche", "camaron", "tortillas", "electricidad",
      "alquiler", "frijol", "pollo"]]
#definimos cuatro variables aleatorias discretas
Y1=df['Mes']
Y2=df['Season']
Y3=df['Anio']
Y4=df['categoria']
✓ 0.0s Python
```

Figura 12. Variables aleatorias para realizar PCA y k-means

Aplicamos PCA al vector aleatorio X y graficamos los datos proyectados en las primeras tres componentes principales, coloreando los datos de acuerdo al año [13](#), al mes [14](#), a la categoría [16](#) y a la estación del año [15](#).

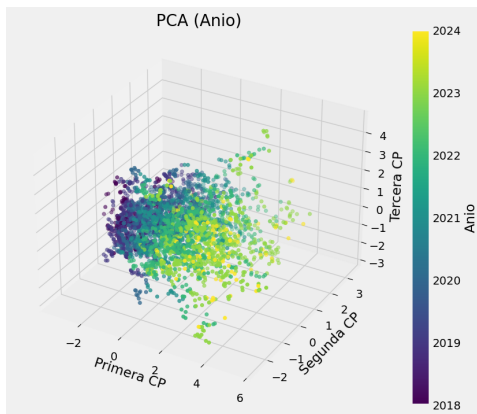


Figura 13. Primeras tres componentes principales coloreadas por año.

En la gráfica de PCA año se aprecia un gradiente de color donde cada color representa un año, es decir, tenemos cierta estructura de acuerdo al año.

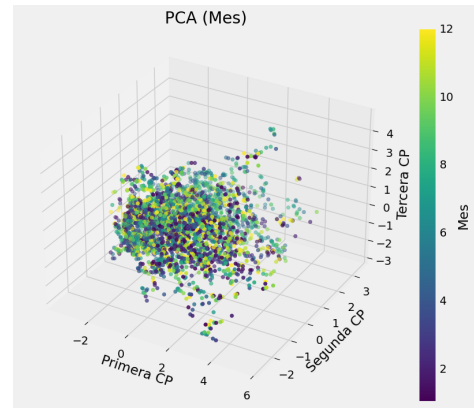


Figura 14. Primeras tres componentes principales coloreadas por mes.

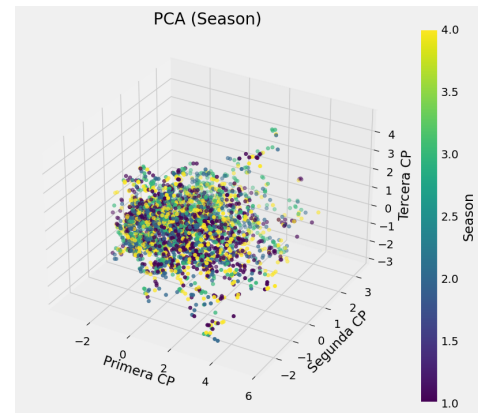


Figura 15. Primeras tres componentes principales coloreadas por estación.

En los dos gráficos anteriores no se observa a simple vista ningún patrón bien definido como en el caso del año.

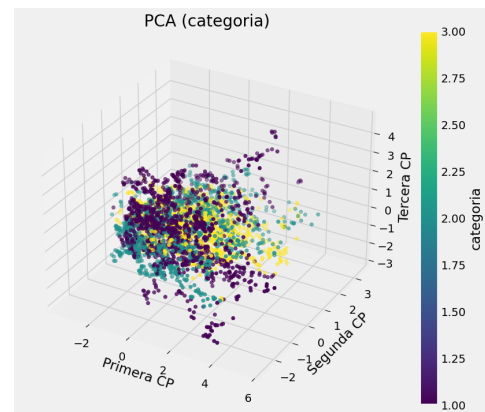
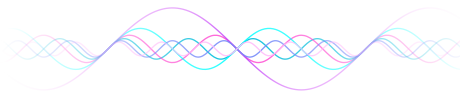


Figura 16. Primeras tres componentes principales coloreadas por categoría (caro, barato, medio).



Finalmente, al colorear con las etiquetas de la variable categoría parece haber algo de agrupamiento, aunque desde el ángulo en que está la imagen no se aprecia adecuadamente.

En el notebook `pca_clustering.ipynb` se encuentran además los gráficos de en 2D de las primeras dos componentes principales y otros gráficos 3D interactivos donde pueden apreciarse más detalladamente los datos y los colores de acuerdo al hue.

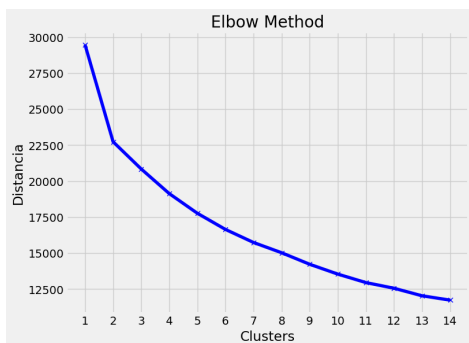


Figura 17. Método del codo para aproximar el número de clusters.

Para ayudarnos a determinar si los datos tienen o no una estructura o se agrupan unos con otros, utilizaremos el algoritmo de k-means sobre los los datos proyectados en las tres primeras componentes principales. Para ello, primero nos apoyamos en el método del codo para tener una idea de donde comenzar el número de clusters.

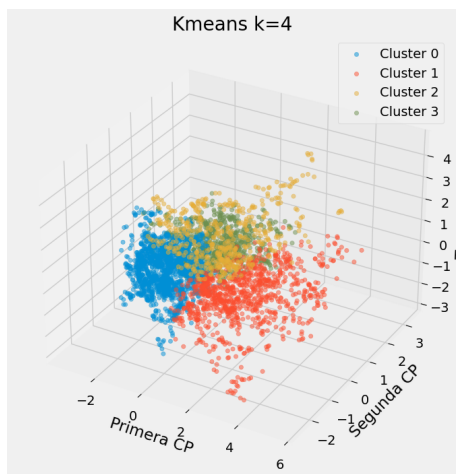


Figura 18. kmeans con 4 clusters sobre los datos reducidos a tres dimensiones

Con base en la gráfica obtenida [17](#) con este método, decidimos comenzar con 4 clusters. En la figura [18](#) se muestran los datos en tres dimensiones coloreados según el cluster determinado con el algoritmo k-means.

Para saber si la clasificación del algoritmo se hizo en función de alguna de las variables categoricas que tenemos disponibles, realizamos gráficos de frecuencia para cada etiqueta del cluster y para cada variable categórica.

En la figura [19](#) se encuentra el histograma de la variable año para cada uno de los cuatro clusters.

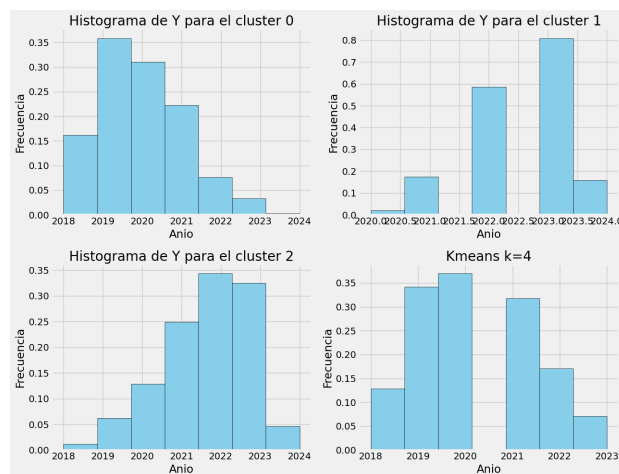


Figura 19. Frecuencias de cada año en por cluster.

En la figura [20](#) se encuentra el histograma de la variable mes para cada uno de los cuatro clusters.

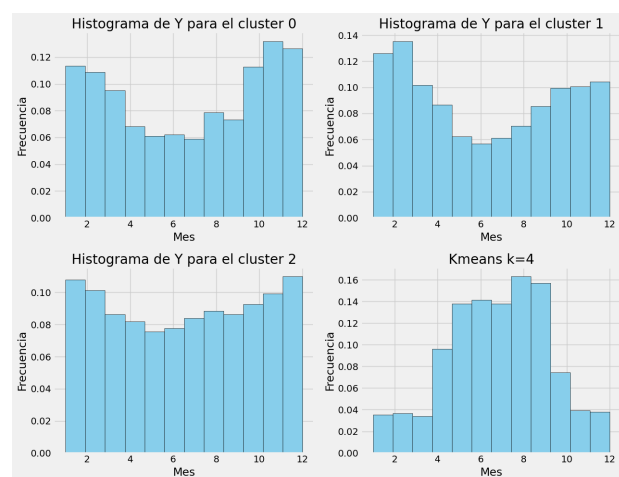


Figura 20. Frecuencias de cada mes en por cluster.

En la figura [21](#) se encuentra el histograma de la variable categoría para cada uno de los cuatro clusters.

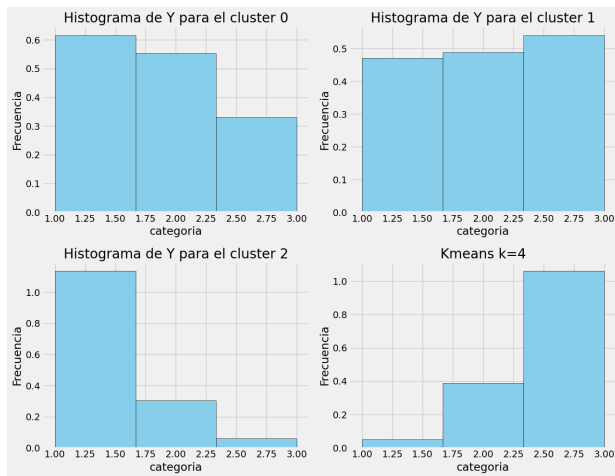
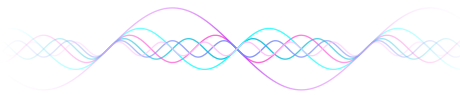


Figura 21. Frecuencias de cada categoría en por cluster.

En la figura 22 se encuentra el histograma de la variable estación para cada uno de los cuatro clusters.

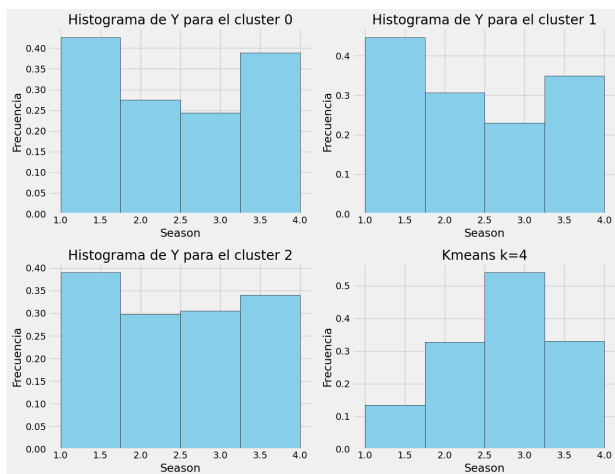


Figura 22. Frecuencias de cada estación en por cluster.

Las dos variables categóricas más discriminantes son año y categoría. Mientras que mes y estación no nos dicen mucho. En los cuatro clusters hay cantidades altas de más de una etiqueta.

Tomando en cuenta el método del codo y los histogramas anteriores se probó el algoritmo k-means para tres clusters sin embargo los histogramas son parecidos a los anteriores, y no se discierne ninguna variable discriminante.

Finalmente, como caso especial debido al aparente agrupamiento que se observa desde el gráfico de pca, se probaron 7 clusters y de igual manera se realizó el gráfico en tres dimensiones.

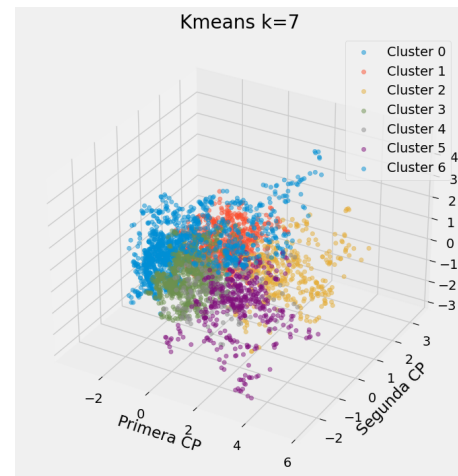


Figura 23. kmeans con 7 clusters sobre los datos reducidos a tres dimensiones

Tal como en 18 también se aprecian grupos definidos, es difícil decir si hay más traslape en uno que otro con este par de gráficos. Sin embargo en los interactivos del notebook `pca_clustering.ipynb` se pueden ver desde distintas perspectivas y parecen clusters más definidos con $k=4$

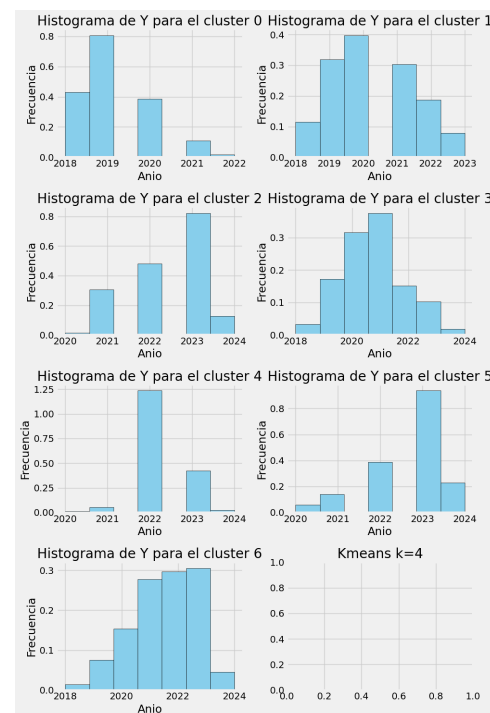
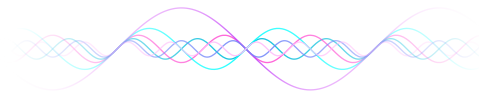


Figura 24. Frecuencias de cada año en por cluster.



IV. CONCLUSIONES

Aplicando una reducción de dimensión a nuestros datos, se puede apreciar una cierta estructura relacionada con el año de los datos. El algoritmo de agrupamiento k-means sugiere además que si una característica importante es *el tipo de ciudad*. Un problema de esta clasificación es que que engloba y resume (de manera muy sencilla) demasiados momentos en el tiempo, sería interesante hacer una clasificación de las ciudades que rescate mejor la información del tiempo.

REFERENCIAS

1. Instituto Nacional de Estadística y Geografía (INEGI). (s/f). Precios promedio. Recuperado de <https://www.inegi.org.mx/app/preciospromedio/?bs=18>
2. Banco de México. (s/f). Recuadros del Informe Trimestral. Recuperado de <https://www.banxico.org.mx/publicaciones-y-prensa/informes-trimestrales/recuadros/%7B1433DE85-D1A1-672C-CAF2-17E95DBA5BC0%7D.pdf>