¿Caro o Barato? Análisis de Precios de Almacén en un Contexto Inflacionario

Diplomatura en Ciencia de Datos, Aprendizaje Automático y sus aplicaciones

Práctico 4 - Analisis No Supervisado

Dario Yvanoff

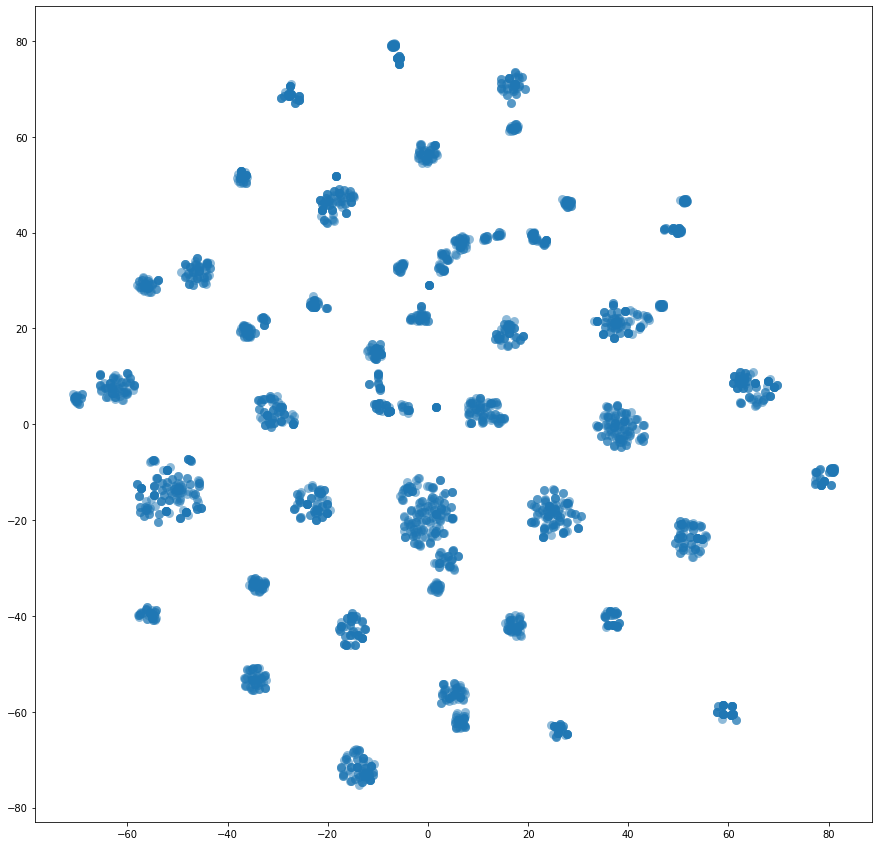
Leandro Sepulveda

Gonzalo Torres

En este practico se aplicaron algoritmos de aprendizaje automático no supervisado (clusterización), a los fines de lograr categorizar los diferentes productos que componían el dataset con que se trabaja, dado que el dataset original no contenía las categorías correspondientes según es informado en la página Precios Claros

Para realizar el curado de los datos, se utilizó el script desarrollado en practicos anteriores, lo que nos otorgó como resultado una “*sparse matriz*” (una matriz compuesta principalmente por ceros), dada la forma elegida de categorizar los productos. Además, como se comentó en informes anteriores, nuestra base presentaba alta dimensionalidad (+240 columnas), lo cual dificultaría el cómputo de diferentes modelos que aplicásemos.

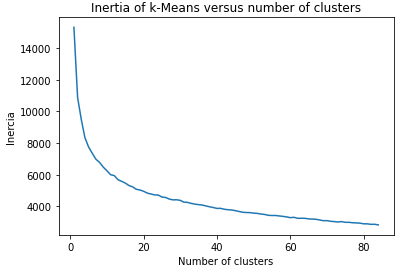
A fin de evaluar si existía algún tipo de patron, se realizó también una proyección t-SNE, la cual se puede visualizar a continuación.



En ella notamos que existirían algunos cluster, y sería el objetivo intentar determinar como se podrían conformar.

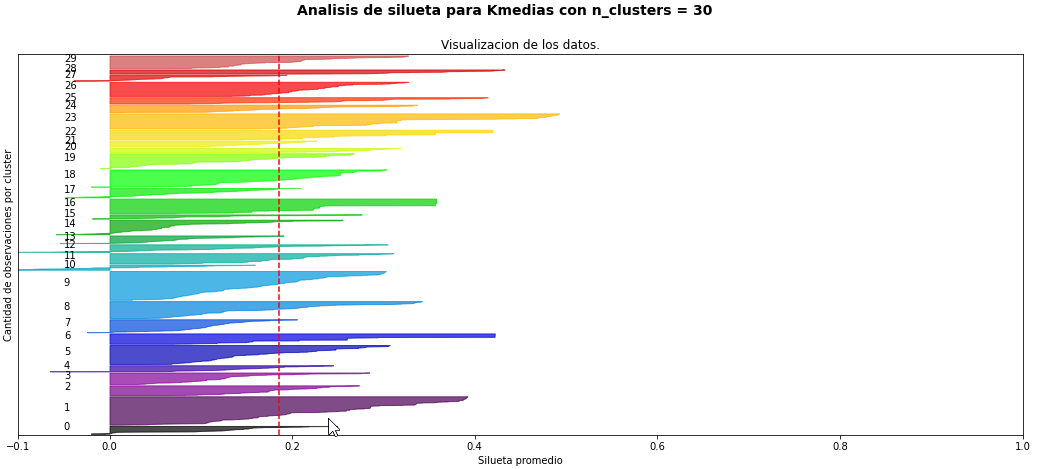
A fin de disminuir el problema de dimensionalidad, se aplicó una reducción utilizando PCA con 100 componentes, lo cual capturó el 95% de la variabilidad de los datos, manteniendose esta como la nueva base de datos para posterior trabajo.

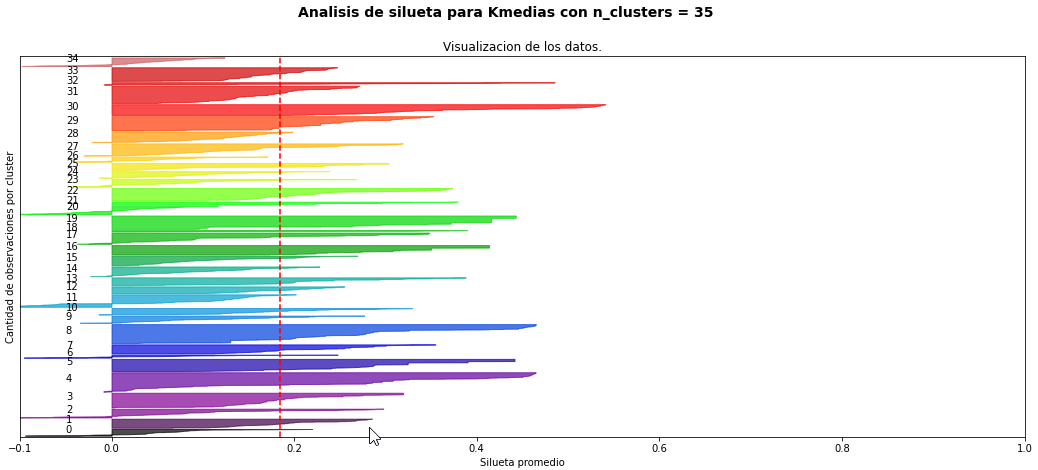
El primer modelo de clusterización utilizado fue KMeans. En el mismo notamos, mediante la utilización del “gráfico del codo” que la cantidad de clúster adecuados se encontraban entre 20 y 40.

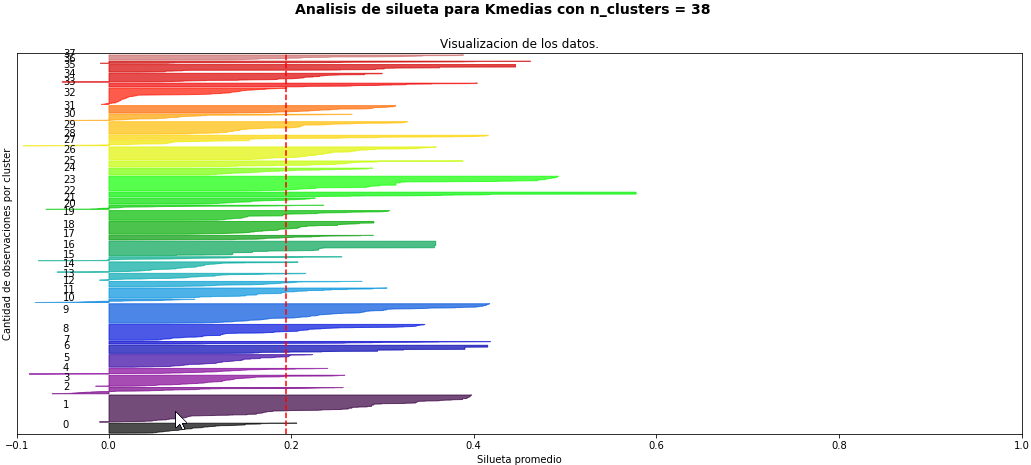


Estas elecciones de cantidad de clúster fueron además evaluadas utilizado la silueta asociada a cada una. Los resultados no fueron los esperados, ya que no cumplían con los estares adecuados para una correcta clusterización (no presentar valores negativos y que todos los valores estuviesen por encima del promedio).

Lo anteriormente descripto puede observarse en las siguientes imágenes.





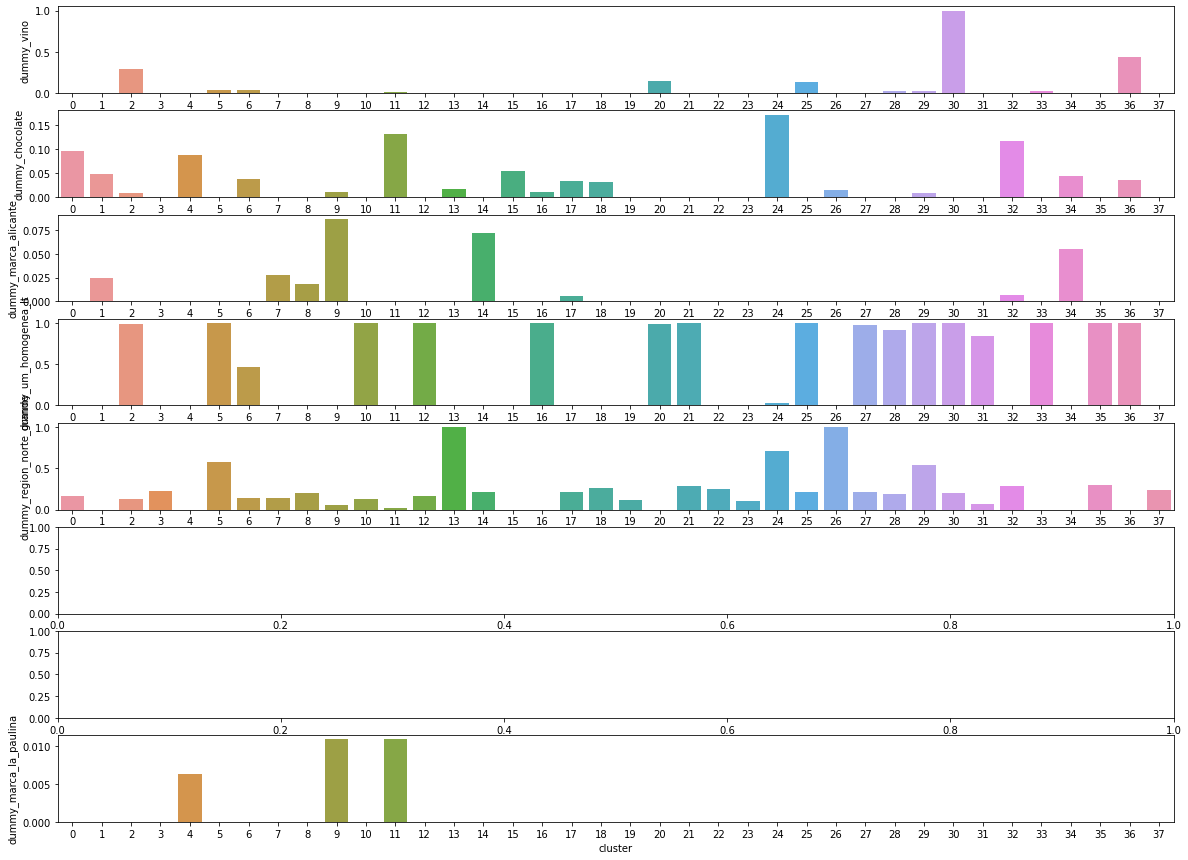


Resultado de la clusterización en 38 conjuntos (ya que se consideró era el valor más adecuado), no se lograron identificar características particulares que permitiesen asociar a los diferentes conjuntos.

A continuación se muestran los resultados pertenecientes al cluster 2 y 9.

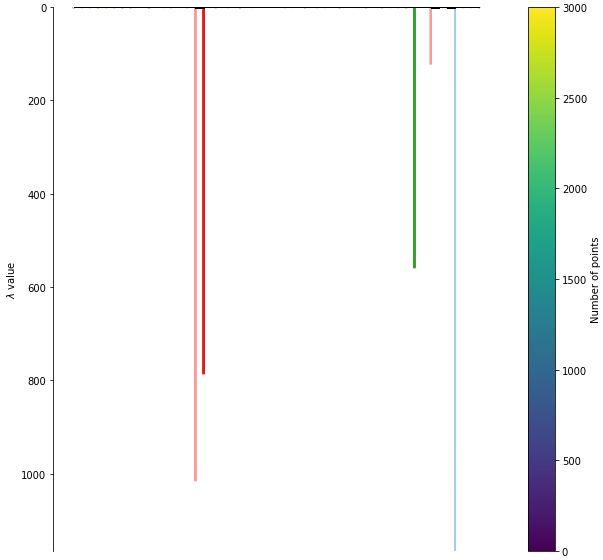


Mostramos a continuación la frecuencia de algunas palabras claves (las más frecuentes) u se observa que en general no hay claridad respecto a la pertenencia de estas a un cluster puntual.



El segundo modelo de clusterización que se aplicó fue uno de tipo jerarquico, HDBSCAN. Al igual que en KMeans, los resultados obtenidos no fueron satisfactorios.

Resumimos sus resultados en el siguiente árbol, uqe no presenta la forma esperada, si no que está muy concentrado en algunos puntos que agrupan productos muy disimiles.



Como conclusión del análisis realizado, notamos que si bien resulta posible realizar una clusterización de los datos, en línea con lo que se observó al principio (grafico t-SNE), estos clusters conformados no cumplen con nuestro objetivo de ser capaces de re-construir la característica Categoría, la cual se perdió en el proceso de extracción de datos.

Entendemos que esta falencia puede surgir del modo en que se clasificaron los productos, mediante la utilización de variables dummies para capturar la información contenida en los nombres, lo cual creó un dataset dominado por valores ceros. Comprendemos que en esta etapa, quizas lo más conveniente hubiese sido aplicar algoritmos de procesamiento de lenguaje natural para lograr mejorar la base para posteriores análisis.