CLUSTERIZACIÓN DE EMPRESAS DE DISTRIBUCIÓN ELÉCTRICA DE LATAM Y EL CARIBE

CODERHOUSE
CURSO DATA SCIENCE
ESTUDIANTE: Marvin Cruz
PROFESORA: Estefanía Susanj
TUTOR: Giuliano Crena

INTRODUCCIÓN

Objetivo

Generar una segmentación de empresas, que en este caso se consideran clientes potenciales, para orientar adecuadamente las estrategias comerciales de una empresa de soluciones técnicas y consultoría del sector energía que pretende expandir sus operaciones al sector LATAM y El Caribe.

Contexto comercial

La compañía requiere mejorar sus estrategias de ventas, por lo que ha solicitado la realización del estudio con datos estadísticos obtenidos en la base de datos del Banco Mundial. Con el análisis de datos los insights presentados, la gerencia de planificación comercial fortalecerá las políticas comerciales en la región.

Audiencia

La audiencia para este proyecto es personal del sector técnico-comercial de una empresa de servicios de consultoría y provisión de equipos de baja, media y alta tensión para empresas de distribución eléctrica.

INTRODUCCIÓN

Hipótesis:

Los clientes potenciales de la compañía pueden agruparse en función de parámetros como el índice de pérdidas de energía, cantidad de empleados, energía transportada y otros indicadores técnicos.

La orientación de la estrategia comercial para empresas dentro del mismo país puede ser diferente.

Preguntas formuladas:

¿Es posible agrupar las empresas eléctricas en diferentes clusters para orientar las estrategias de ventas?

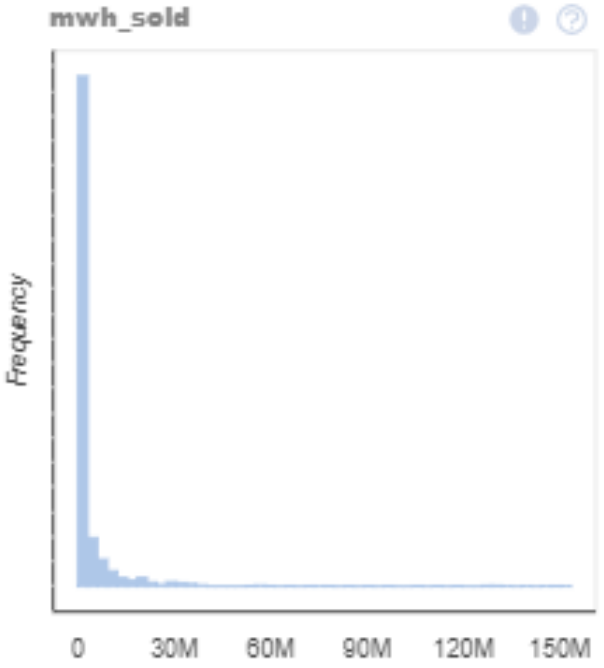
¿Es posible agregar a los grupos definidos un factor de agrupamiento por zona geográfica?

¿Es posible mantener la misma estrategia comercial a nivel país?

EDA

Análisis univariado:

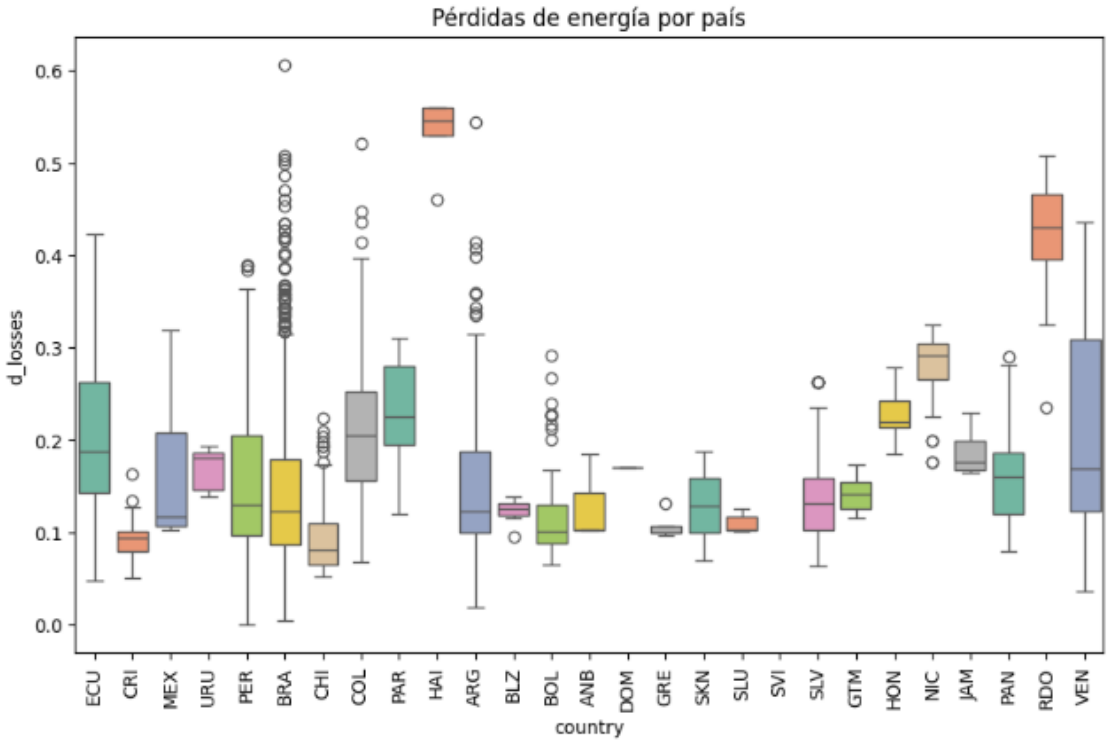
- Se aprecia que en general no se cuenta con una distribución normal.
- Distribuciones con un sesgo pronunciado a la izquierda.
- Se cuenta con 24 variables numéricas, 3 categóricas y 1 georeferenciada (País)
- Se confirma que el Datset contiene una cantidad importante de Missing Values.



Fuente de datos:

Para llevar adelante el proyecto se dispone de un data set con información relevada por el Banco Mundial, descargada del sitio: https://energydata.info/

EDA



Análisis bivariado:

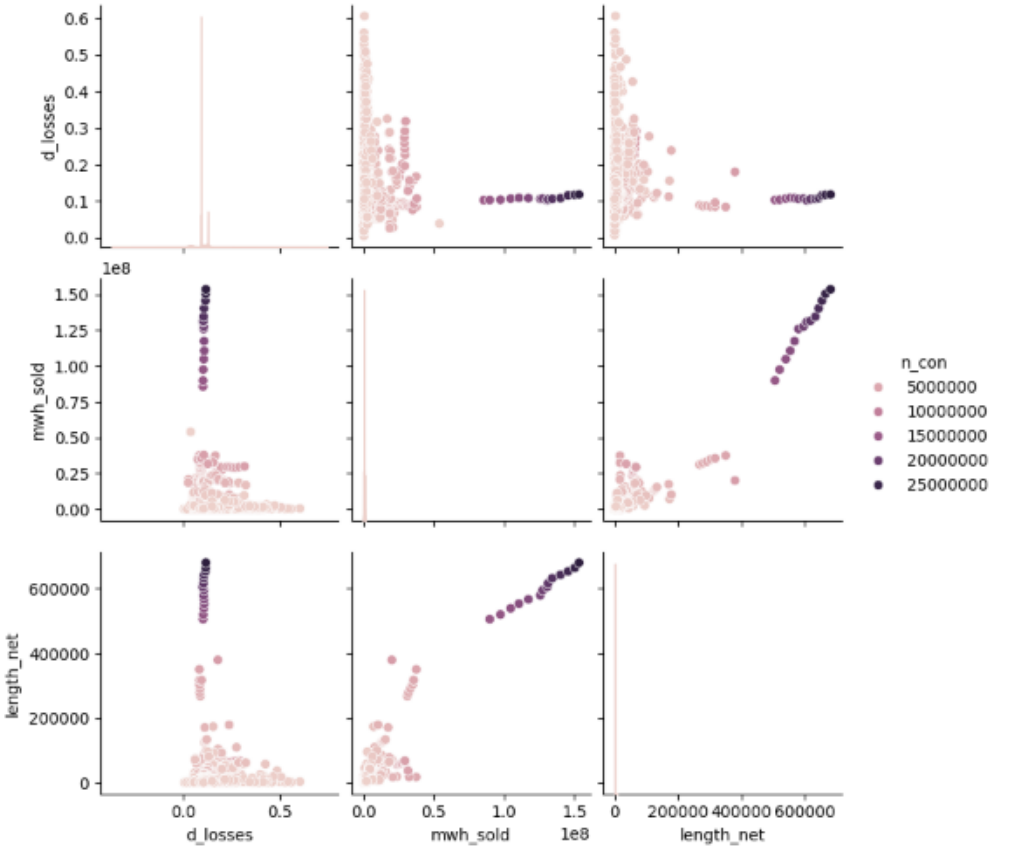
Se observa que Brasil posee la mayor cantidad de outlyers, seguido por Argentina y Bolivia. Para el caso de Haiti se considera que tiene la mayor cantidad de Perdidas de energía de la region, con un promedio, y valores característicos del boxplot excepcionalmente altos comparados con el resto de paises.

Por otra parte, vemos que San Vicente y las Granadinas no tiene información sobre este parámetro.

EDA

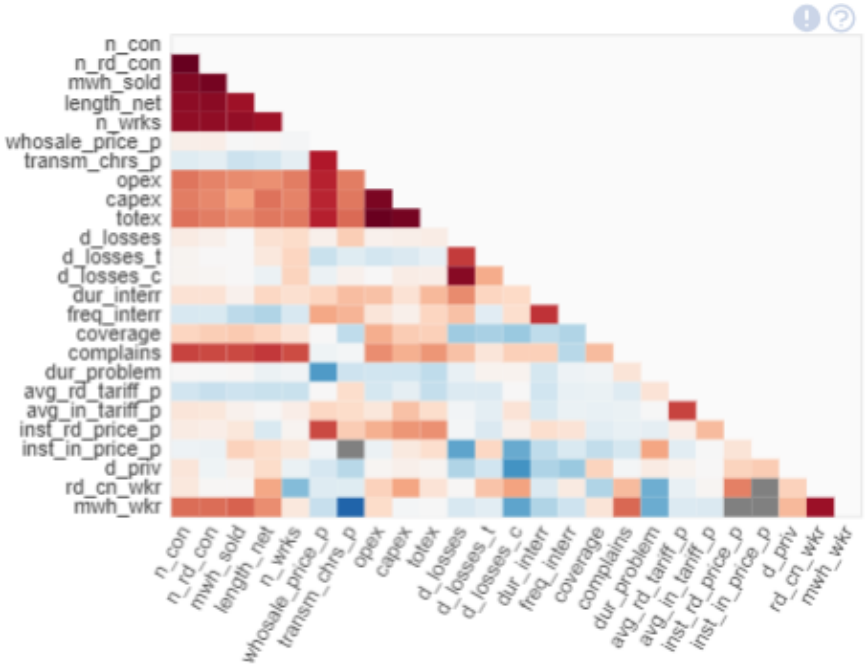
Análisis multivariado:

- El agrupamiento de pérdida de energía es concentrado en valores de 0.10 a 0.20, independiente de las dos variables de los ejes.
- Existen valores de pérdida de energía muy en rango, pero con valores de venta de energía y km de red muy elevados, igual comportamiento se aprecia sobre la cuarta variable (hue) Esto indica que las empresas eléctricas de mayor tamaño, así como los paises con mayor consumo mantienen una operación que se considera dentro del rango de pérdidas aceptable.
- Se tiene una alta correlación entre la energía circulante y los knm totales de red.
- La distribución de pérdidas mantiene valores en general distribuidos alrededor del 10-20 % con una linea de datos atipica siempre alrededor de esos valores para la variable pérdidas

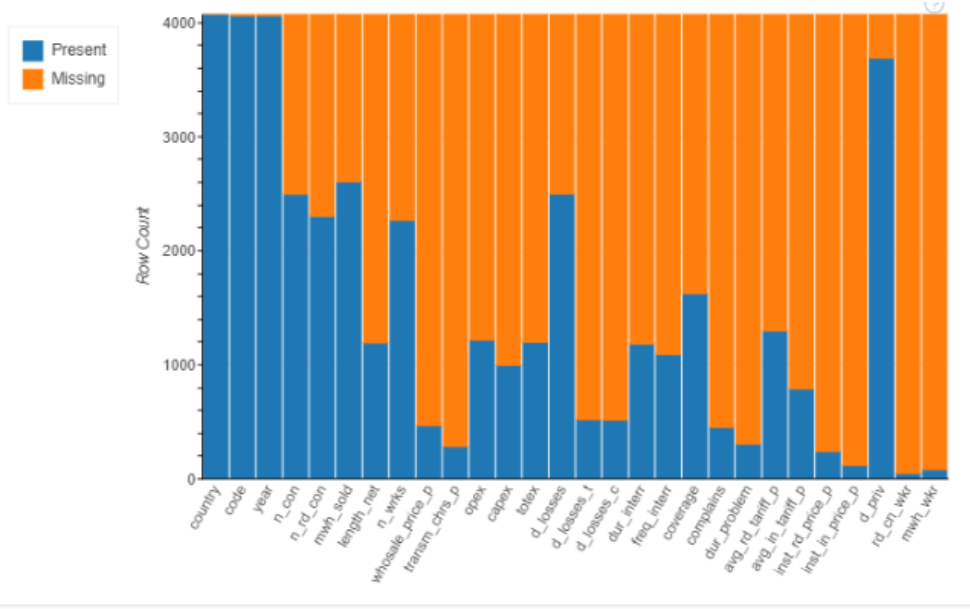


EDA

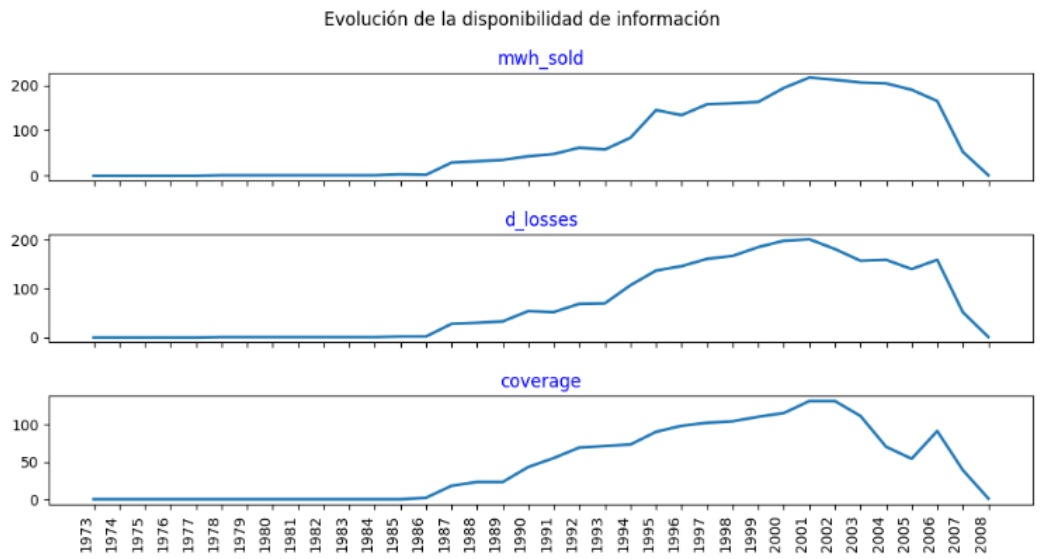
Matriz de correlación:



Missing values:



EDA



Missing values:

En los últimos años la disponibilidad de la información se ha ido incrementando en todos los sectores, influenciado por el monitoreo de indicadores por parte de los organismos de control así como por la incursión de la gestión basada en sistemas de la información.

En conclusión, la presencia de muchos Missing values no es un impedimento para contnuar con el análisis, pero debe ser considerada al momento de definir los criterios para la implementación del modelo de Machine Learning elegid

MODELO NO SUPERVISADO

KMEANS:

```
Número de clusters: 5.00
Indice de Silhouette: 0.44
Índice de Calinski-Harabasz: 141.11
Índice de Davies-Bouldin: 0.99
```

DBSCAN:

```
Número de clusters: 4
Puntos sin clasificar (Noise): 141
Indice de Silhouette: 0.03
Índice de Calinski-Harabasz: 34.37
Índice de Davies-Bouldin: 1.57
```

SPECTRAL CLUSTERING:

```
Número de clusters: 5
Indice de Silhouette: 0.26
Índice de Calinski-Harabasz: 107.82
Índice de Davies-Bouldin: 1.29
```

HDBSCAN:

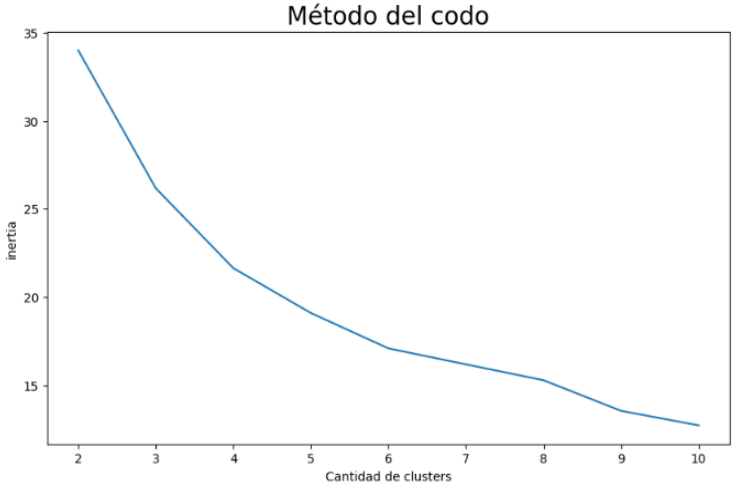
```
Número de clusters: 5
Puntos sin clasificar (Noise): 106.00
Indice de Silhouette: 0.10
Índice de Calinski-Harabasz: 43.36
Índice de Davies-Bouldin: 2.37
```

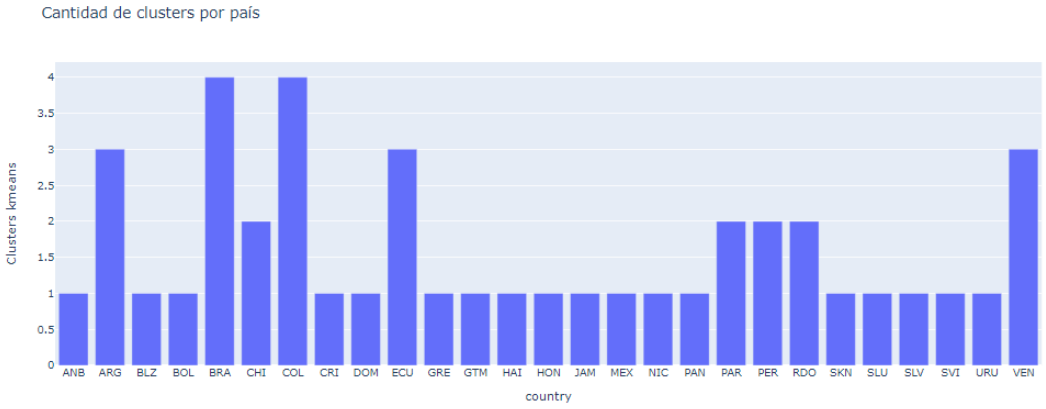
Se observa que el modelo que tiene mejores métricas es Kmeans.

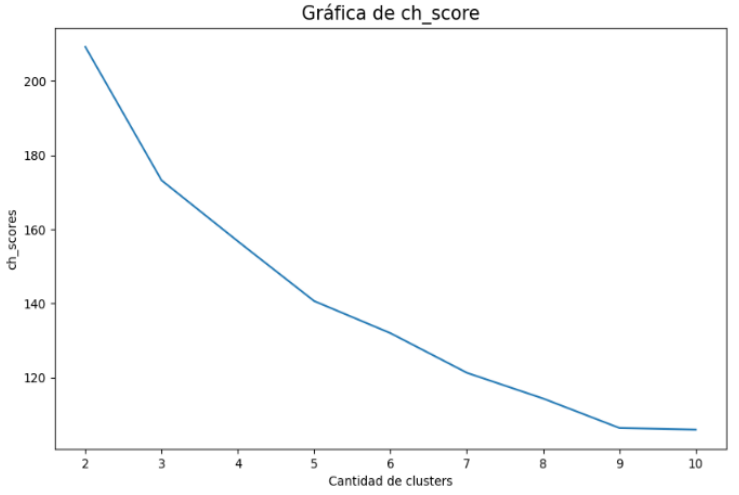
BDSCAN y HBDSCAN muestran una gran cantidad de datos considerados ruido, lo cual es no deseado para el objetivo que se pretende alcanzar.

MODELO FINAL

Cantidad de Clusters: 5.00
Indice de Silhouette: 0.43
Índice de Calinski-Harabasz: 136.85
Índice de Davies-Bouldin: 1.05







CONCLUSIONES

- **Se confirma que es posible agrupar las empresas eléctricas en diferentes clusters para orientar las estrategias de ventas aprovechando el aprendizaje no supervisado.**

- **La orientación de la estrategia comercial para empresas dentro del mismo país puede ser diferente.**

- **No es recomendable manejar la misma estrategia comercial para empresas del mismo pais, se recomienda estudiar la misma estrategia para empresas correspondientes a los clusters generados, esto implica que la misma estrategia puede ser adecuada para clientes en distintos paises.**