

Proyecto: Reckitt

Entregable 3: K-Means

Continuando con el análisis de la marca “Vanish” y la identificación de segmentos clave para identificar áreas de mejora, se procedió a implementar un modelo K-Means para una clasificación de los datos otorgados por Reckitt.

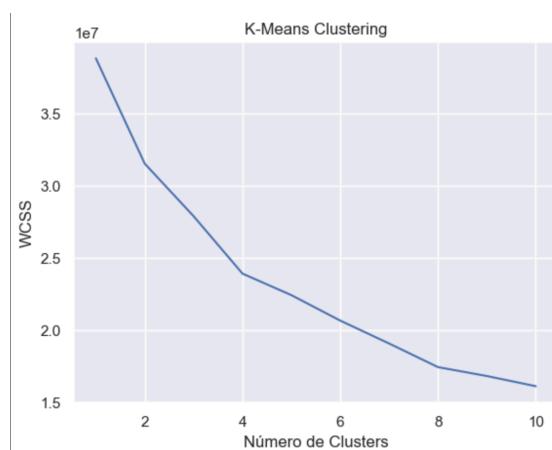
Después de una depuración de variables, datos nulos y selección de datos relevantes, el data frame con el que se procedió a trabajar fue el siguiente:

	WEEK	YEAR	MONTH	ITEM	TOTAL_UNIT_SALES	REGION	CATEGORY	FORMAT	ATTR3	TYPE
0	01-22	2022	1	7501268200001	0.003	TOTAL AUTOS AREA 1		1 LIQUIDO	BAMBINO	FABRIC TREATMENT & SANITIZER
1	01-22	2022	1	7501268200001	0.003	TOTAL AUTOS AREA 1		1 LIQUIDO	GERMICIDA	FABRIC TREATMENT & SANITIZER
2	01-22	2022	1	7501268200001	0.003	TOTAL AUTOS AREA 1		1 LIQUIDO	MASCOTAS	FABRIC TREATMENT & SANITIZER
3	01-22	2022	1	7501268200001	0.003	TOTAL AUTOS AREA 1		1 LIQUIDO	MULTIUSOS	FABRIC TREATMENT & SANITIZER
4	01-22	2022	1	7501268200001	0.003	TOTAL AUTOS AREA 1		1 LIQUIDO	MULTIUSOS	FABRIC TREATMENT & SANITIZER

Las variables categóricas fueron transformadas por medio de “**LabelEncoder**” y debidamente escaladas usando la librería “**StandardScaler**” de **ScikitLearn**.

	YEAR	MONTH	TOTAL_UNIT_SALES	REGION	CATEGORY	FORMAT	ATTR3	TYPE
0	-0.708798	-1.409734	-0.221310	-1.524665	0.0	0.127895	-1.753772	0.0
1	-0.708798	-1.409734	-0.221310	-1.524665	0.0	0.127895	-1.070317	0.0
2	-0.708798	-1.409734	-0.221310	-1.524665	0.0	0.127895	-0.842499	0.0
3	-0.708798	-1.409734	-0.221310	-1.524665	0.0	0.127895	-0.614680	0.0
4	-0.708798	-1.409734	-0.221310	-1.524665	0.0	0.127895	-0.614680	0.0
...
6466101	1.410838	0.456751	-0.102242	1.427222	0.0	1.520062	1.435686	0.0
6466102	1.410838	0.456751	-0.102242	1.427222	0.0	1.520062	1.435686	0.0
6466103	1.410838	0.456751	-0.102242	1.427222	0.0	1.520062	1.435686	0.0
6466104	1.410838	0.456751	-0.102242	1.427222	0.0	1.520062	0.296594	0.0
6466105	1.410838	0.456751	-0.102242	1.427222	0.0	1.520062	1.663505	0.0

6466106 rows x 8 columns

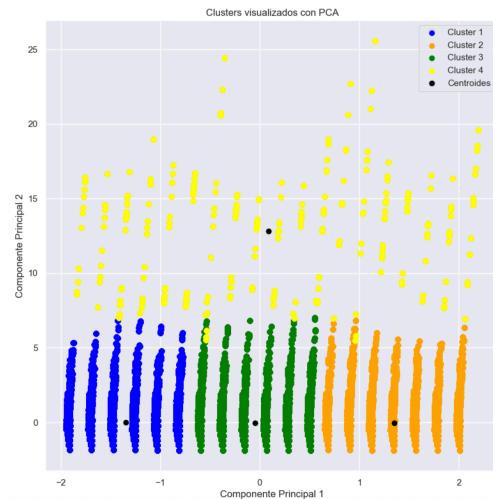


El codo de **Jambú** nos muestra un *codo* óptimo en el valor **4**, dicho valor se utilizó para generar el modelo de K-Means, con un número de **4** clusters.

Después de la generación de clusters y la asignación de grupos, el nuevo data frame generado es el siguiente:

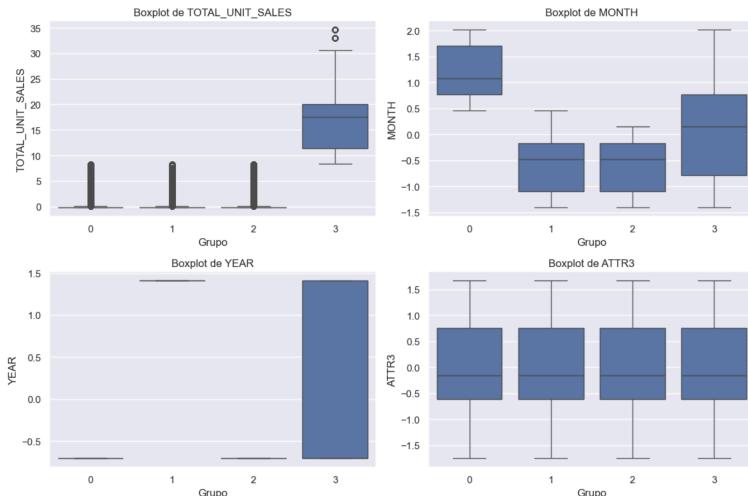
WEEK	YEAR	MONTH	ITEM	TOTAL_UNIT_SALES	REGION	CATEGORY	FORMAT	ATTR3	TYPE	GROUP	ITEM_DESCRIPTION
0	01-22	2022	1 7501268200001	0.003	TOTAL AUTOS AREA 1	1	LIQUIDO	BAMBINO	0	2	LA VALECIANA BOT.PLAST. 1000ML NAL. 750126820...
1	01-22	2022	1 7501268200001	0.003	TOTAL AUTOS AREA 1	1	LIQUIDO	GERMICIDA	0	2	LA VALECIANA BOT.PLAST. 1000ML NAL. 750126820...
2	01-22	2022	1 7501268200001	0.003	TOTAL AUTOS AREA 1	1	LIQUIDO	MASCOTAS	0	2	LA VALECIANA BOT.PLAST. 1000ML NAL. 750126820...
3	01-22	2022	1 7501268200001	0.003	TOTAL AUTOS AREA 1	1	LIQUIDO	MULTIUSOS	0	2	LA VALECIANA BOT.PLAST. 1000ML NAL. 750126820...
4	01-22	2022	1 7501268200001	0.003	TOTAL AUTOS AREA 1	1	LIQUIDO	MULTIUSOS	0	2	LA VALECIANA BOT.PLAST. 1000ML NAL. 750126820...

El diagrama de dispersión de los clusters, con una descomposición previa mediante PCA nos muestra la siguiente separación de grupos:



Si bien no es una visualización común de lo que suelen ser la separación de clusters, la separación de grupos es clara, y nos deja ver como el cuarto cluster parece tener el dominio de los datos.

Con esta premisa, generando diagramas de caja, o, *Boxplots*, observamos precisamente como el **grupo 3**, en este caso, el cluster **número 4**, domina la ventas totales, y por año. Donde únicamente el grupo 1 (**cluster 0**) domino las ventas en mensuales en alguna ocasión.



WEEK	YEAR	MONTH	ITEM	TOTAL_UNIT_SALES	REGION	CATEGORY	FORMAT	ATTR3	TYPE	GROUP	ITEM_DESCRIPTION
32118	01-22	2022	1 0000075000622	371.038	TOTAL AUTOS SCANNING MEXICO	1 LIQUIDO	BAMBINO	0	3		CLORALEX EL RENDIDOR BOT.PLAST. 2000ML NAL 000...
32119	01-22	2022	1 0000075000622	371.038	TOTAL AUTOS SCANNING MEXICO	1 LIQUIDO	GERMICIDA	0	3		CLORALEX EL RENDIDOR BOT.PLAST. 2000ML NAL 000...
32120	01-22	2022	1 0000075000622	371.038	TOTAL AUTOS SCANNING MEXICO	1 LIQUIDO	MASCOTAS	0	3		CLORALEX EL RENDIDOR BOT.PLAST. 2000ML NAL 000...
32121	01-22	2022	1 0000075000622	371.038	TOTAL AUTOS SCANNING MEXICO	1 LIQUIDO	MULTIUSOS	0	3		CLORALEX EL RENDIDOR BOT.PLAST. 2000ML NAL 000...

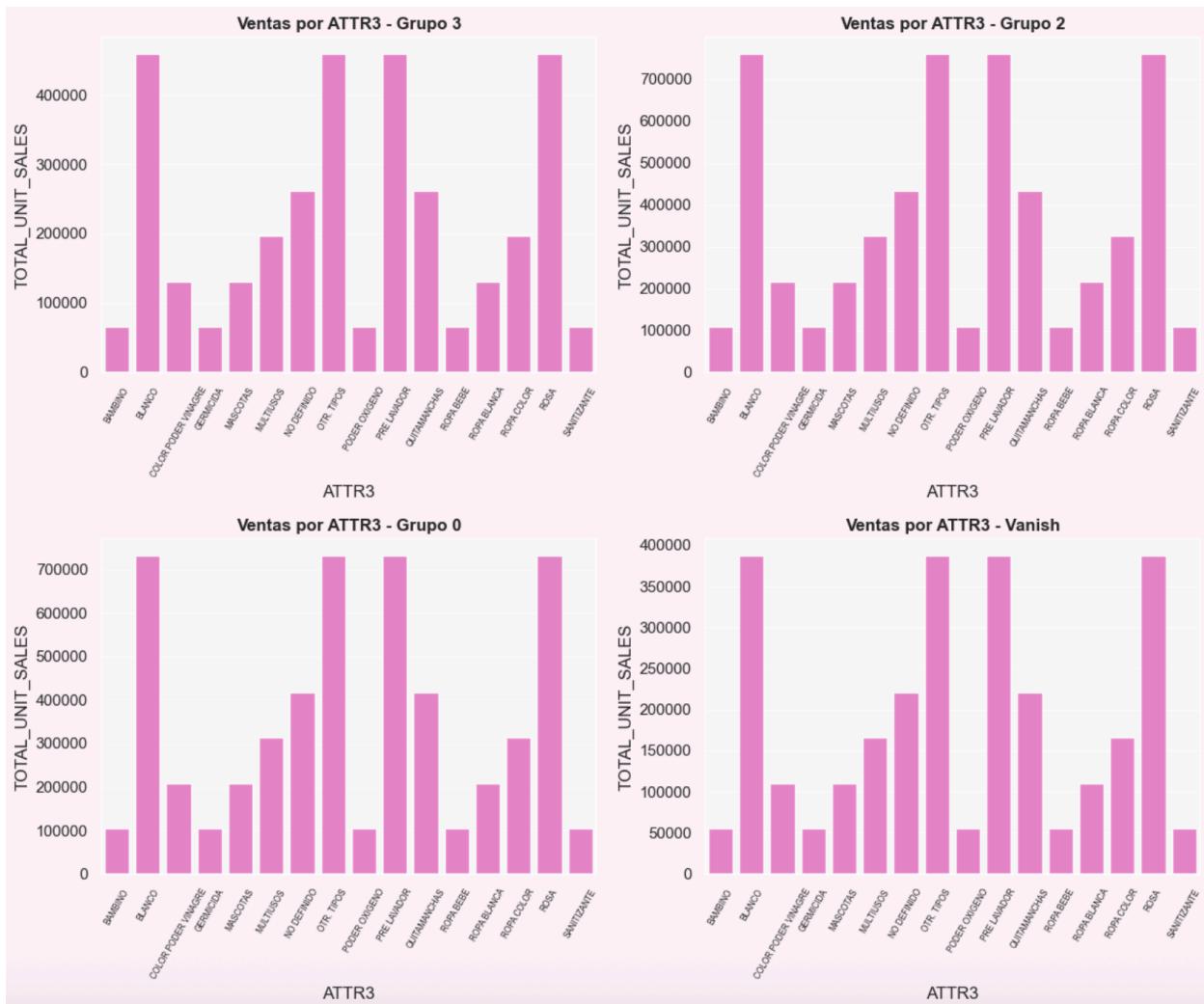
Ahora, un análisis breve del **grupo 3**, nos muestra que la marca líder es *Cloralex*.

Vanish por su parte tiene presencia en los **grupos 0, 1 y 2**.

WEEK	YEAR	MONTH	ITEM	TOTAL_UNIT_SALES	REGION	CATEGORY	FORMAT	ATTR3	TYPE	GROUP	ITEM_DESCRIPTION
901	01-22	2022	1 7501058716422	0.033	TOTAL AUTOS AREA 6	1 POLVO	BAMBINO	0	2		VANISH OXI ACTION GOLD QUITAMANCHA AHORRO DEL ...
902	01-22	2022	1 7501058716422	0.033	TOTAL AUTOS AREA 6	1 POLVO	GERMICIDA	0	2		VANISH OXI ACTION GOLD QUITAMANCHA AHORRO DEL ...
903	01-22	2022	1 7501058716422	0.033	TOTAL AUTOS AREA 6	1 POLVO	MASCOTAS	0	2		VANISH OXI ACTION GOLD QUITAMANCHA AHORRO DEL ...
904	01-22	2022	1 7501058716422	0.033	TOTAL AUTOS AREA 6	1 POLVO	MULTIUSOS	0	2		VANISH OXI ACTION GOLD QUITAMANCHA AHORRO DEL ...
905	01-22	2022	1 7501058716422	0.033	TOTAL AUTOS AREA 6	1 POLVO	MULTIUSOS	0	2		VANISH OXI ACTION GOLD QUITAMANCHA AHORRO DEL ...

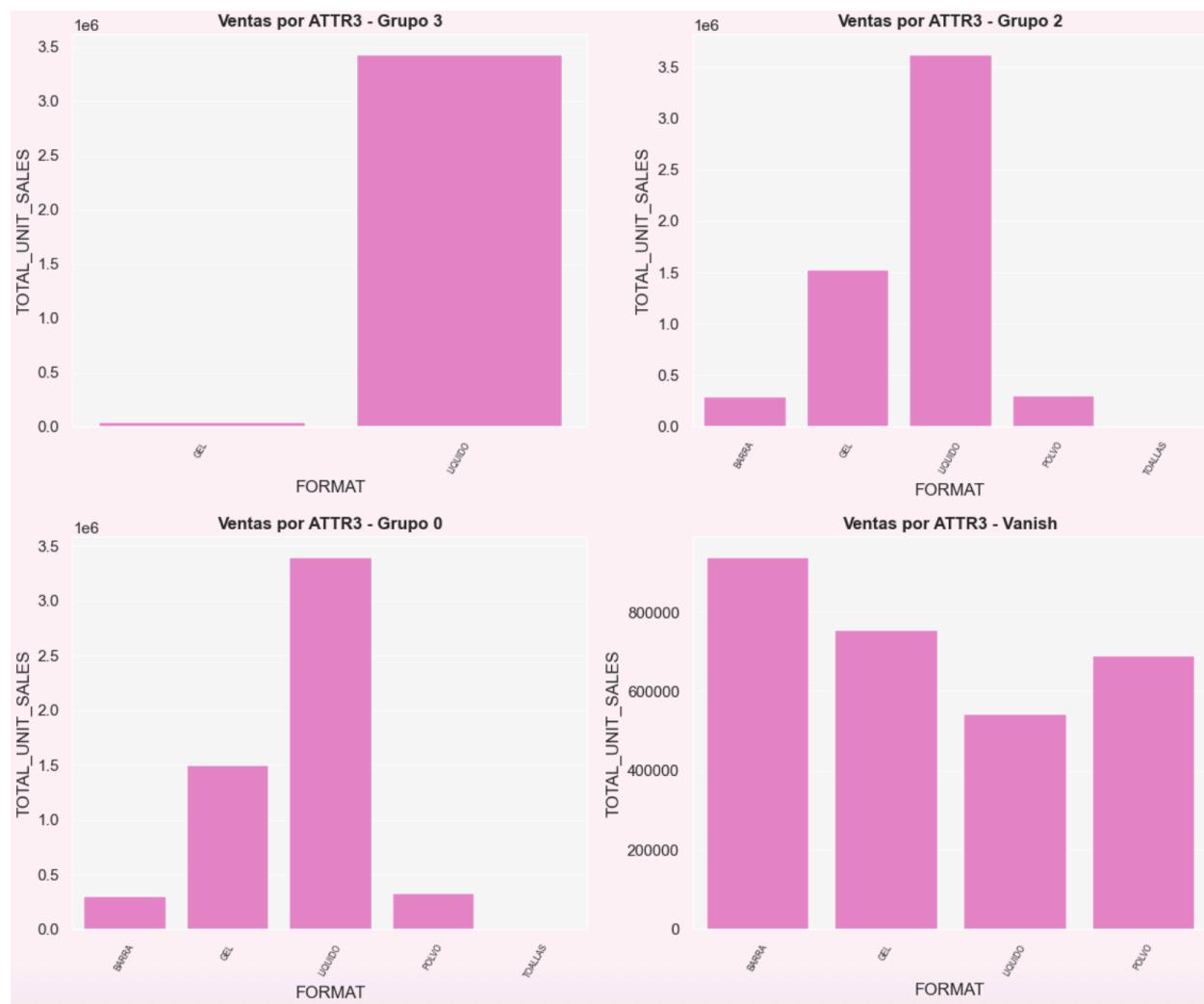
```
vanish_data['GROUP'].unique()
array([2, 0, 1], dtype=int32)
```

Haciendo un chequeo de las **ventas totales**, en base al **atributo** de los productos:

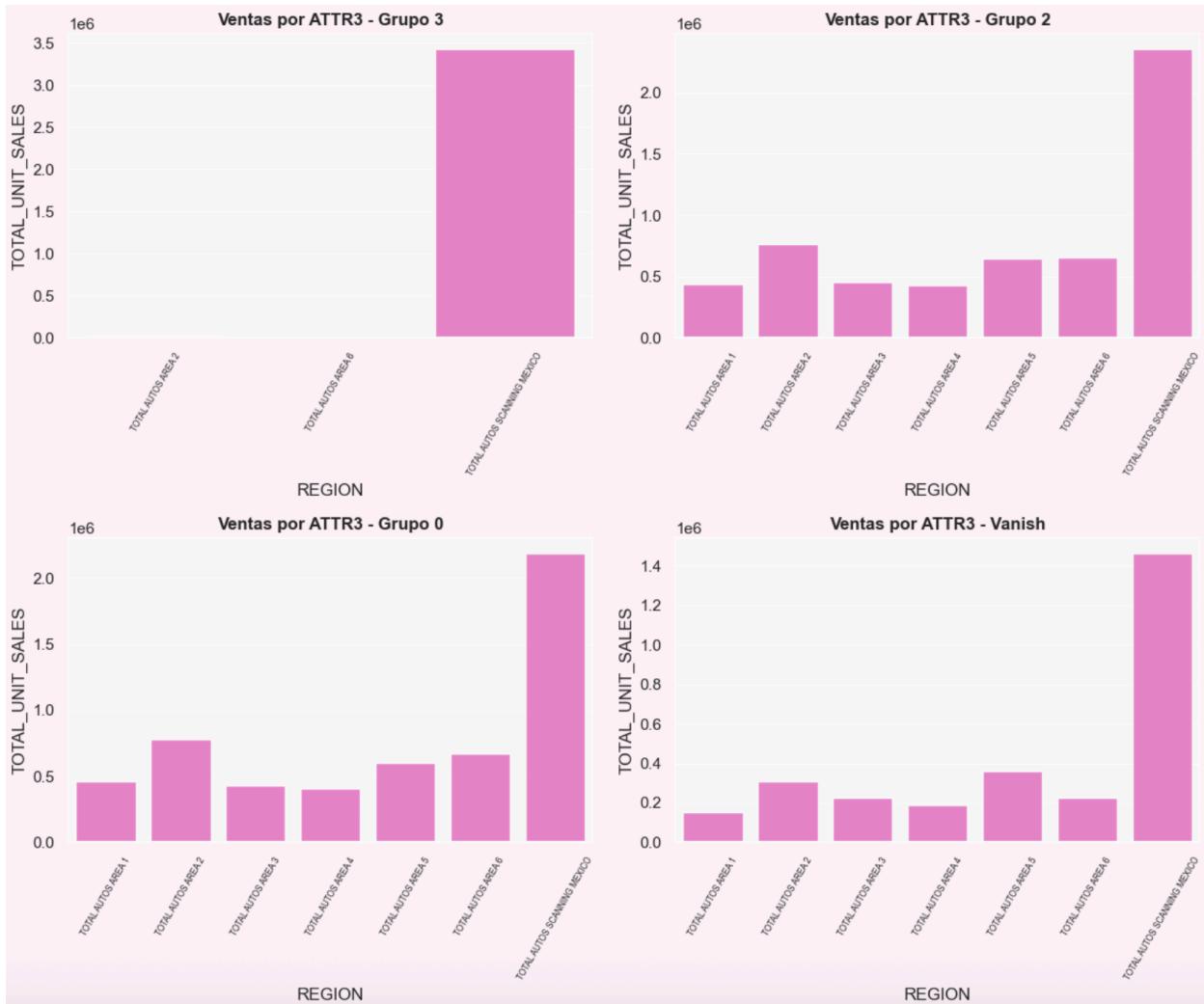


Las ventas en cuanto a los atributos parece ser distribuida de manera casi similar, donde el **Grupo 2 y Grupo 3** parecen tener el dominio por tener mas número de ventas en los formatos ganadores.

Ahora, el análisis del **número total de ventas por formato** nos deja ver que el producto líder *Cloralex* cuenta con únicamente 2 formatos, en los cuales el **Líquido** es su fuerte. Por su parte, *Vanish* cuenta con multiples formatos, pero sus ventas no parecen tener mucha diferencia entre si.



La distribución de **Ventas totales por Región** muestra como la región *Total Autos Scanning México* es la más importante para la distribución de los productos Reckitt, y el área que básicamente hace que *Cloralex* sea líder.



Conclusión

El análisis con el algoritmo K-means demuestra las fortalezas y debilidades de la marca *Vanish*. La marca cuenta con **múltiples formatos**, siendo el formato *Barra* el más destacable, además cuenta con distribución en **múltiples regiones**. Sin embargo, *Cloralex* domina el mercado con **un solo formato** y siendo principalmente distribuida en **una sola región**. Esto nos indica que el éxito de *Cloralex* va más allá que solamente intentar cubrir varios productos y regiones, y se enfoca más en un **estratégico posicionamiento de marca**. Esto se traduce a un reconocimiento de la marca como líder en el mercado.

Vanish podría enfocar sus esfuerzos a **impulsar su formato Barra** y posiblemente, **Gel**, con un fuerte **redireccionamiento de distribución a la region de Total Autos Scanning México** para demostrar que es una **marca líder** en los formatos previamente mencionados, y siendo distribuida en las mismas proporciones que *Cloralex*.