Reactivo 1 – Data Science Project

Camilo Alejandro Hernández Salazar – Jr. Data Scientist

Celular: (+52) 3316039940
Correo: camiloalejandro.ca@gmail.com

I. DATA CLEANING

Al observar la tabla de los datos, podemos ver que posee 20 columnas, de las cuáles solo 5 son numéricas, 2 son del tipo fecha y el resto, contienen datos categóricos. También se puede mencionar que esta tabla puede pertenecer a una empresa donde se comercializan muebles de oficina, así como los aparatos electrónicos y el resto de objeto que se pueden encontrar fácilmente en una oficina.

Al estar la tabla de datos muy bien estructurada y sin datos vacíos, solo queda eliminar las columnas que no nos ayuden a realizar ningún análisis, estas son: *'Row ID','Order ID', 'Customer ID', 'Country', 'Postal Code', 'Product ID', 'Product Name'*.

II. EDA

Al iniciar la exploración de datos, se encontró lo siguiente sobre sus columnas numéricas:

	Sales	Quantity	Discount	Profit
count	9994.000000	9994.000000	9994.000000	9994.000000
mean	229.858001	3.789574	0.156203	28.656896
std	623.245101	2.225110	0.206452	234.260108
min	0.444000	1.000000	0.000000	-6599.978000
25%	17.280000	2.000000	0.000000	1.728750
50%	54.490000	3.000000	0.200000	8.666500
75%	209.940000	5.000000	0.200000	29.364000
max	22638.480000	14.000000	0.800000	8399.976000

gracias a esto, podemos notar que los datos numéricos son muy diferentes entre sí, hablando sobre sus cantidades, por lo que fue necesario hacer una normalización a la hora de entrenar el modelo de *clustering* más tarde. También podemos decir que la distribución es muy grande en cuanto el apartado de ventas (Sales) y el apartado de ganancia (Profit).

Lo primero que se realizó fue un análisis acerca de sus formas de envío, haciendo histogramas para analizar que tipo de envío es el más utilizado, cual es la cantidad de días que tarda cada envío la mayoría de las veces, que segmento de clientes usan cada tipo de envío, cuantos días tardan en llegar los pedidos de acuerdo a cada segmento, que categoría de artículos se envían de que modo y cuantos días tardan en llegar. Esto lo podemos ver en la fig.1.

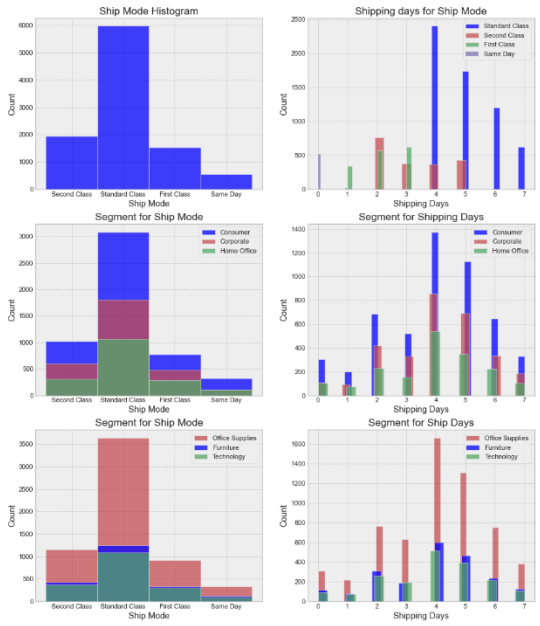

Fig. 1: Hitograma sobre envíos

Dándonos como conclusión que la opción de envío estándar en la más común, aunque sea la que más tarda en que los artículos lleguen a su destino y que los objetos de oficina son los más vendidos y varía bastante los días de envío.

Posteriormente, se hizo una gráfica de barras para mostrar el crecimiento total de año con año acerca de: las ventas, cantidad de artículos vendidos, descuentos otorgados, ganancias y el total de órdenes, tal como se muestra en la fig.2. Con esto solo podemos aseverar que el crecimiento de la empresa en cada uno de estos ámbitos ha ido creciendo conforme pasan los años.

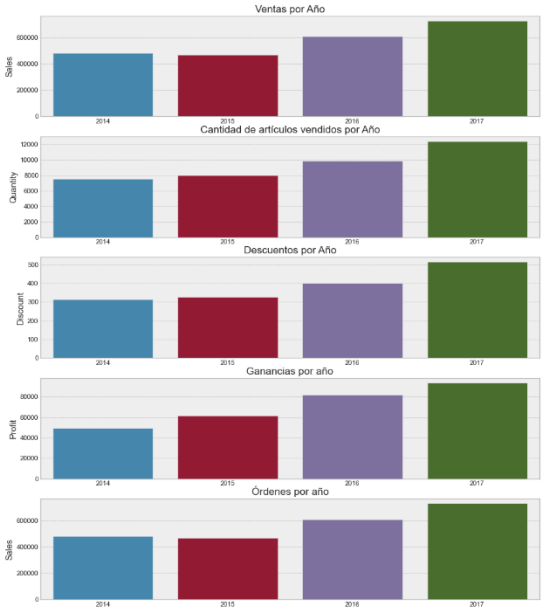

Fig. 2: Gráfica de barras de diversos ámbitos económicos de la empresa, mostrados año a año.

También se realizó una serie de tiempo, de los mismos ámbitos (a excepción de los descuentos por año), tal como expone la fig.3. Aquí podemos observar que parece haber un gran crecimiento económico en la segunda parte del año, ya que también incrementa la demanda.

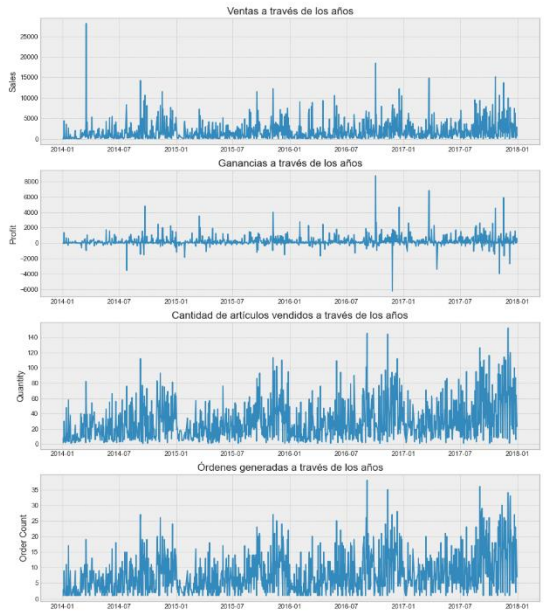

Fig. 3: Series de tiempo de diversos ámbitos económicos y de demanda de la empresa

También se incluyó una creación de tablas donde se analiza, las ventas, cantidad de artículos vendidos, descuentos, ganancia, y número de órdenes por estado, siendo así que los 3 estados donde se encuentran más pérdidas son: Texas, Ohio y Pennsylvania. Mientras que encontramos qué: California, New York y Washington siendo de los estados en los que obtenemos mayor ganancia.

Según las tablas dinámicas creadas con los peores estados en cuestión de ganancias, podemos ver que año a año realizan muchas compras y órdenes de compra, sin embargo, también son los que más aprovechan los descuentos otorgados por la empresa. Esto puede significar que la empresa está realizando una estrategia de venta muy agresiva al aumentar la cantidad de descuentos en sus precios con tal de ganar compradores en dichos estados y que si se quiere reducir el margen de pérdidas económicas, deberían reducir la cantidad de descuentos al año también.

III. MODELLING

Finalmente, en el apartado de modelado, lo primero que se llevó a cabo, fue la conversión de los datos categóricos a numéricos con la función *LebelEncoder* de la librería *sklearn*. Con esta función traducimos cada categoría a un número simple para referenciar cada valor de cada columna.

Posteriormente, se normalizaron los datos con la siguiente ecuación:

$$(X-X.min())/(X.max()-X.min())$$

donde X es nuestro DataFrame con las columnas numéricas y crea una razón de la diferencia del valor actual menos el valor mínimo del DataFrame sobre la diferencia del valor máximo con el valor mínimo.

A continuación, se hizo una especie de *Benchmark*, con una técnica llamada *"Codo de Jambú"*. Esto con el propósito de saber cuál sería el número de grupos óptimo para realizar el entrenamiento y agrupamiento de nuestros datos de clientes. En la siguiente fig. 4, se puede ver como a partir del grupo 4, la caída del comparador WCSS se vuelve mucho más suave, sin embargo, algo similar se ve en el nodo 6, así que podemos usar cualquiera de estas dos cantidades para conseguir un buen *clustering*.

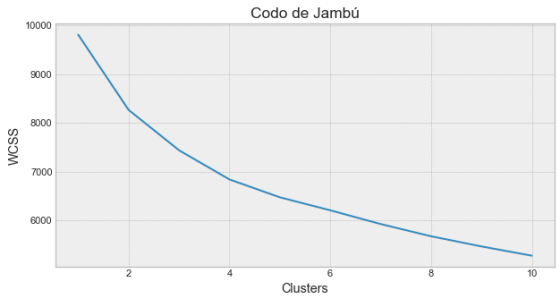

Fig. 4: Muestra gráfica del codo de Jambú.

Por último, realizamos nuestro entrenamiento con el modelo de *KMeans*, al cual introducimos nuestro DataFrame de datos normalizados. Los resultados del entrenamiento los descomponemos en 2 componentes con la función *PCA*, la

cual se encarga de hacer un análisis de componentes principales. De esta manera obtenemos una tabla de los dos componentes elegidos y la categorización del *clustering* realizado previamente. Al graficarlo obtenemos el resultado mostrado en la gráfica de la fig.5, donde se observan los diferentes grupos divididos con sus respectivos colores, creando así un buen agrupamiento de los clientes de acuerdo a sus características de localización y compra.

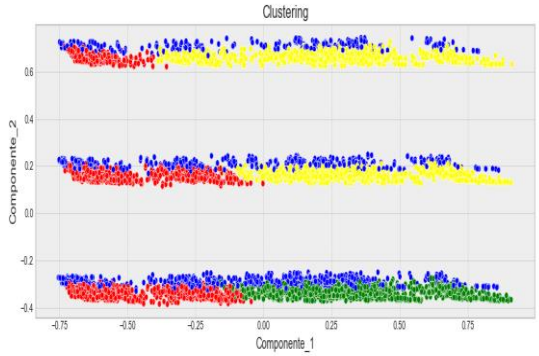

Fig. 5: Resultados del Clustering