Universidad Nacional Autónoma de México

Facultad de Estudios Superiores de Acatlán

Análisis de Ventas Globales en 2016

Diplomado en Ciencia de Datos

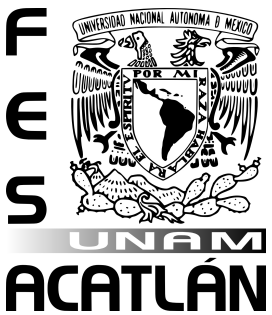
presenta:

Francisco Roman Peña de la Rosa

Módulo 1: Introducción, Manipulación, Exploración y Visualización de Datos

Profesora:

Act. Carla Paola Malerva Reséndiz



Ciudad de México 2021

Introducción

Este reporte corresponde al trabajo realizado con el datset obtenido de **Data World**, también se encuentra disponible en **Tableau Community** y **Kaggle**. Este dataset ha sido utilizado para algunos demos en Tableau, Power BI y QlikView. Sin embargo, en esta ocasión fue seleccionado para aplicar la mayor parte de los métodos y técnicas que comprende la manipulación, exploración y visualización de datos.

Además de aplicar el conocimiento adquirido durante este módulo y los siguientes, otro punto por el cual se consideró trabajar con este dataset fue analizar datos no relacionados con el área de **Relaciones Públicas** y **Análisis de Redes Sociales** que corresponde a las actividades que actualmente desempeño en mí actual trabajo. Otra motivación que lleva a seleccionar este dataset, es que actualmente busco mejorar mí entendimiento de información en **Retail** y **Finanzas**.

El dataset está identificado como **Global Superstore 2016** y contiene un total de 51,357 registros y 23 variables, se encuentra disponible en formato **.xlsx**.

Capítulo 1

Calidad de los Datos

1.1. Conjunto de datos: Global Superstore 2016

Este dataset se encuentra disponible en diferentes fuentes online, algunas de éstas son:

- https://data.world/tableauhelp/superstore-data-sets
- https://public.tableau.com/app/profile/chinmayi8680/viz/Kaggle-SuperStore/Dashboard1

Dentro de estos sitios web podemos ver que la finalidad de la tabla de datos es meramente de aprendizaje y esto permite trabajar con la información de diferentes maneras. Estos datos describen el comportamiento de las ventas de diferentes artículos durante el periodo 2014 a 2016, los mercados de mayor prominencia y el volumen tanto en ventas como en artículos por cada región.

Capítulo 2

Análisis Exploratorio de los Datos EDA

2.1. Diccionario de Datos

A continuación se presenta el arreglo tabular de las variables que integran el dataset trabajado:

Variable	Tipo	Dato	Descripción
v-order-id	Categórica	String	Representa el código de orden de compra
d-order-id	Fecha	Date	Es la fecha en la que se realiza la orden de compra.
d-ship-date	Fecha	Date	Corresponde a la fecha de envío del producto.
v-ship-mode	Categórica	String	Corresponde al modo de envío del producto.
customer-id	Categórica	String	Es el ID del cliente.
t-customer-name	Texto	String	Es el nombre del cliente.
v-segment	Categórica	String	Categoría a la que corresponde el producto.
v-city	Categórica	String	Ciudad.
v-state	Categórica	String	Estado
v-country	Categórica	String	País.
v-postal-code	Categórica	Integer	Código postal.
v-market	Categórica	String	Mercado de venta.
v-product-id	Categórica	String	ID del producto.
v-category	Categórica	String	Categoría principal del producto.
v-sub-category	Categórica	String	Es la subcategoría del producto.
t-product-name	Texto	String	Nombre del producto adquirido.
c-sales	Numérica	Float	Ventas generadas.
c-quantity	Numérica	Integer	Cantidad de productos ordenados-vendidos.
c-discount	Numérica	Float	Descuento aplicado a la venta.
c-profit	Numérica	Float	Utilidad y/o ganancia de la vental artículo(s).
c-shipping-cost	Numérica	Float	Costo del envío.
v-order-priority	Categórica	String	Prioridad de entrega del producto.

2.2. Tipo de Variables

En la siguiente tabla se describe de forma general el tipo de variable que integra el dataset que fue analizado:

Variable	**Tipo**
v-order-id	Categórica
d-order-id	Fecha
d-ship-date	Fecha
v-ship-mode	Categórica
customer-id	Categórica
t-customer-name	Texto
v-segment	Categórica
v-city	Categórica
v-state	Categórica
v-country	Categórica
v-postal-code	Categórica
v-market	Categórica
v-product-id	Categórica
v-category	Categórica
v-sub-category	Categórica
t-product-name	Texto
c-sales	Numérica
c-quantity	Numérica
c-discount	Numérica
c-profit	Numérica
c-shipping-cost	Numérica
v-order-priority	Categórica

2.3. Completitud

El dataset en cuestión tenía con la siguiente dimensión: 51,357 registros y 23 variables en donde nuestra variable objetivo es predecir las ventas por producto. La siguiente tabla muestra la completitud que tienen las variables de forma inicial, es decir, el ratio entre los valores disponibles totales por cada variables y los ausentes:

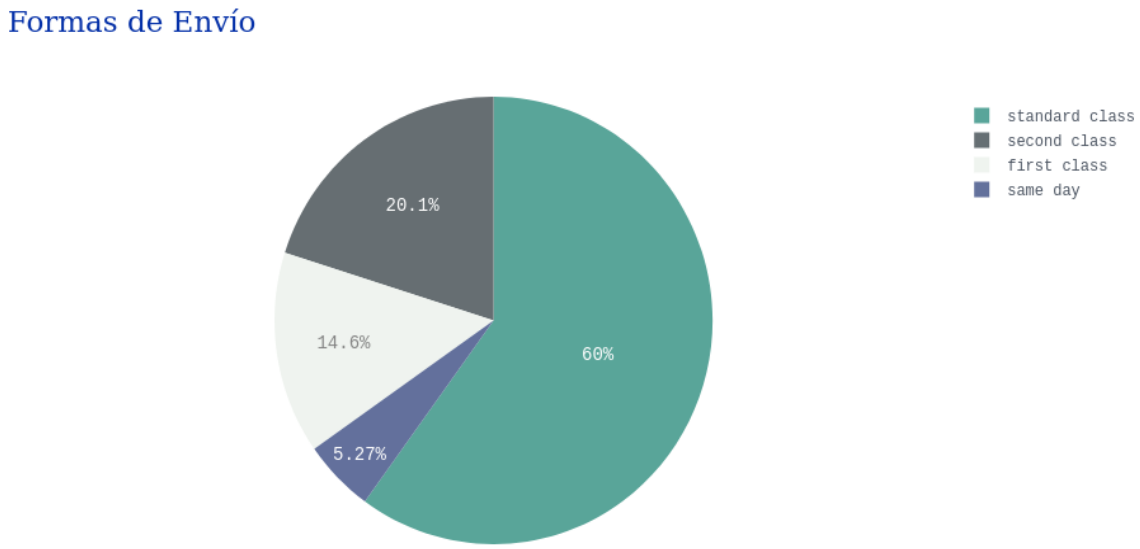
Columna	Total	Completitud
v-postal-code	41303	19.4841
v-market	2	99.9961
t-customer-name	2	99.9961
v-category	1	99.9980
v-segment	1	99.9980
v-order-priority	1	99.9980
v-ship-mode	0	100.0000
v-profit	0	100.0000
c-profit	0	100.0000
c-discount	0	100.0000
c-quantity	0	100.0000
c-sales	0	100.0000
t-product-name	0	100.0000
v-sub-category	0	100.0000
customer-id	0	100.0000
d-order-date	0	100.0000
v-region	0	100.0000
c-shipping-cost	0	100.0000
d-ship-date	0	100.0000
v-country	0	100.0000
v-state	0	100.0000
v-city	0	100.0000
v-product-id	0	100.0000
v-order-id	0	100.0000

De lo anterior podemos ver que la única variable que tiene menos del 80 por ciento de completitud es v-postal-code, y el resto se encuentra en un 100 por ciento, lo cual nos indica que la única variable a eliminar será la antes mencionada.

2.4. Visualización de los Datos

En esta sección vamos a presentar algunas gráficas preliminares que nos ayuden a entender el comportamiento de nuestras variables. La utilidad de la visualización es muy importante, ya que es una manera de interpretar la distribución de información de interés.

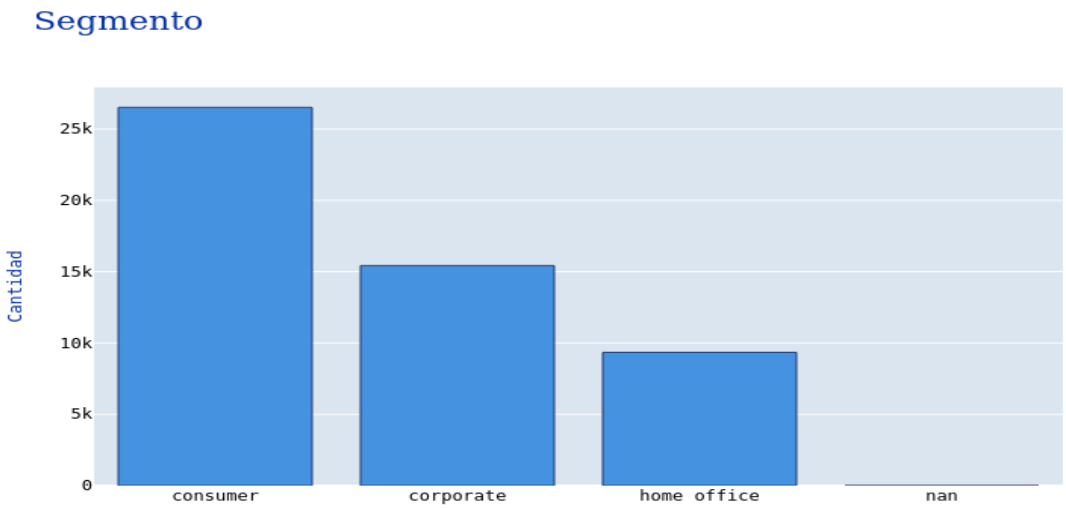
2.4.1. Variables Categóricas



En el gráfico anterior se muestran las formas de envío y se puede observar que la **clase estandard** es la de mayor proporción con el 60 por ciento, sin embargo al mismo tiempo es de nuestro interés conocer como impactan los envíos de esta clase en las las ventas.

	c_sales	c_quantity	c_discount	c_profit	c_shipping_cost
v_ship_mode					
first class	1842003.707800	26070	1117.678000	209871.357200	310013.741540
same day	667201.983900	9230	387.662000	76173.067800	115973.716000
second class	2575785.521080	35743	1450.356000	290880.565080	315532.400360
standard class	7587052.082700	107323	4374.932000	890632.021200	615099.575500

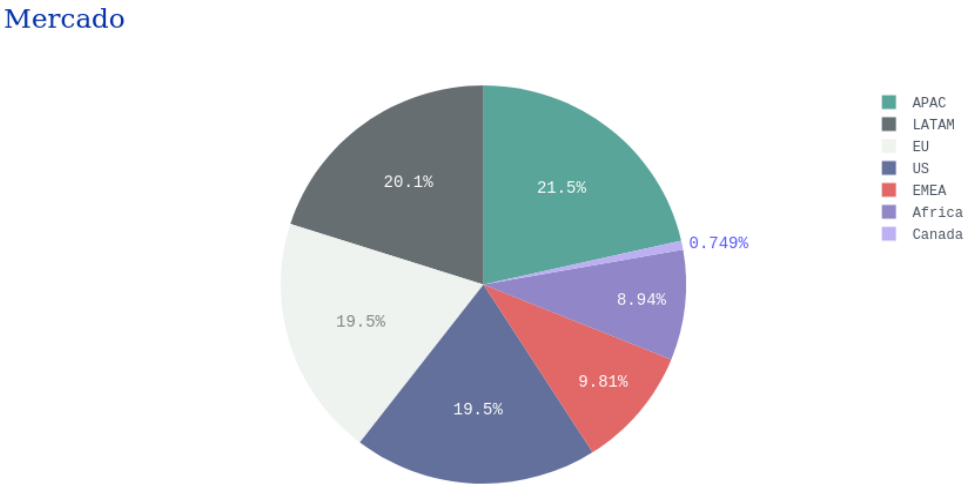
En la tabla anterior podemos ver de forma preliminar (aún no es a versión e interpretación final) que la el envío en clase estandard tiene una mayor cantidad en ventas



La distribución para la variable de segmento muestra que **Consumer** ocupa la primera posición con un 51.7 por ciento seguido de Corporate y Homme Office. De la misma manera que la forma de envío, nos interesa identificar si existe una relación preliminar con las ventas que es nuestra variable objetivo.

	c_sales	c_quantity	c_discount	c_profit	c_shipping_cost
v_segment					
consumer	6528906.713860	92193	3808.842000	747894.660060	699667.799900
corporate	3831422.931060	53579	2205.284000	441905.558660	411427.022000
home office	2308513.610560	32585	1316.502000	276572.842560	244947.901500
nan	3200.040000	9	0.000000	1183.950000	576.710000

A partir de la tabla anterior vemos que **Consumer** en efecto otorga un mayor volumen en ventas, seguido de Corporate y Home Office, respectivamente.

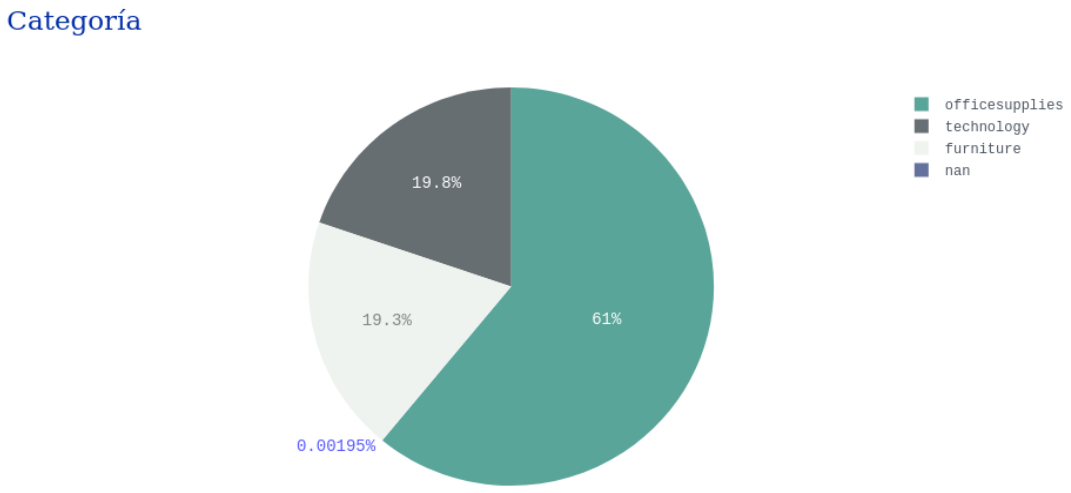


En la gráfica anterior correspondiente a la distribución del **Mercado**, podemos notar que Asia-Pacífico ocupa el primer lugar dentro de esa categoría, seguido de LATAM y US. Sin embargo, para objeto de nuestro análisis se busca conocer si el volumen en ventas mantiene un mismo comportamiento a la propoción de cada mercado.

Esto se muestra en la siguiente tabla donde se comprueba que evidentemente Asia-Pacífico genera un mayor volumen en ventas, así como de utilidad. Y por el contrario LATAM y US se ubican en la penúltima y última posición respectivamente.

	c_sales	c_quantity	c_discount	c_profit	c_shipping_cost
v_market					
APAC	3595332.888600	41245	1637.630000	436993.937000	388591.224000
Africa	787582.011000	10572	718.900000	88656.906000	88614.810000
Canada	66928.170000	833	0.000000	17817.390000	7405.630000
EMEA	808019.991000	11521	986.200000	44645.583000	88855.670000
EU	2941530.751500	37779	1031.550000	371044.846500	309895.624000
LATAM	2167006.247080	38534	1395.258000	221962.007080	234610.703000
US	2305600.836300	37877	1561.090000	286433.021700	238645.710400

Para la variable **Categoría** se realizó un tratamiento semilar a las anteriores, en el siguiente gráfico se muestra su distribución:

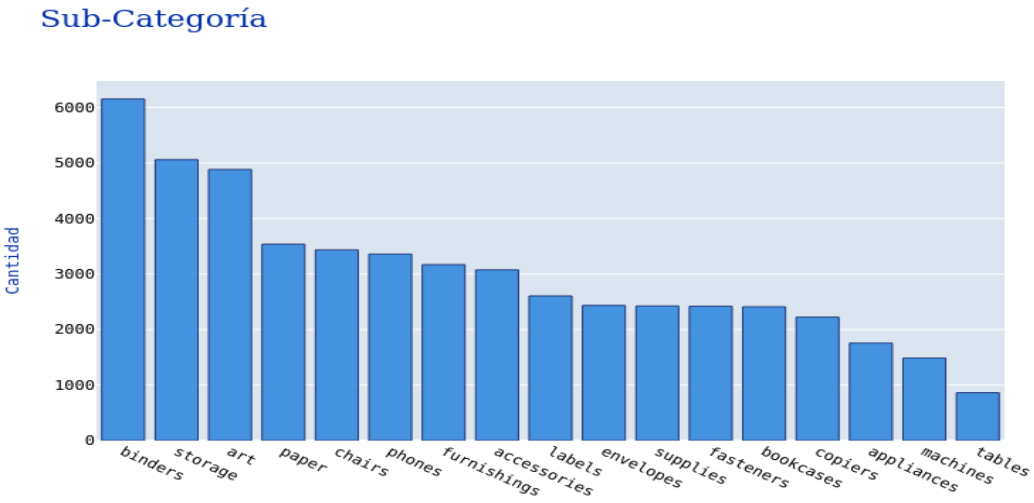


En la tabla que se presenta a continuación se presenta con el objetivo de verificar si **Offfice Supplies** que ocupa el primer lugar con un 61 por ciento, es la categoría de productos que mayor volumen en ventas genera.

	c_sales	c_quantity	c_discount	c_profit	c_shipping_cost
v_category					
furniture	4121230.215900	34973	1660.230000	286068.206000	441704.109000
nan	2036.860000	7	0.000000	366.634800	524.760000
officesupplies	3789513.705500	108195	4297.290000	518795.674300	405925.495000
technology	4759262.514080	35191	1373.108000	662326.496180	508465.069400

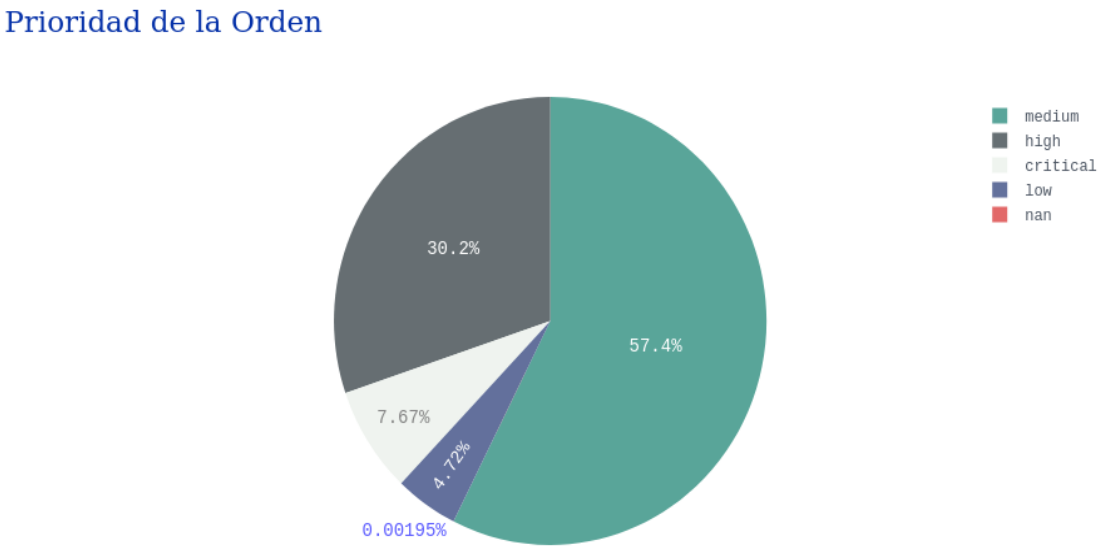
Como podemos ver en la tabla anterior, aún cuando **Office Supplies** es la categoría con mayor proporción dentro de nuestro dataset, se encuentra en segundo lugar en volumen de ventas generado seguida de **Technology**.

En esta sección se muestra el desgloce de las sub-categorías que se derivan de cada una de las anteriores. En la siguiente gráfica tenemos la distribución por los diferentes artículos vendidos:



Como podemos ver en el gráfico de **Sub-Ctegoría**, tenemos que con el 12 % **Binders** ocupa el primer lugar, seguido de **Storage** con un 9.86 %, **Art** con el 9.52 % y **Paper** con el 6.9 %. De esto podemos comprobrar que efectivamente el Top 5 de estos productos pertenecen a Office Supplies, que de acuerdo con el gráfico mostrado en la sección anterior, es la categoría en primer lugar con mayor cantidad de artículos ventidos, pero en segundo en volumen de ventas.

En el siguiente gráfico se muestra la distribución de la prioridad de la orden, en donde podemos ver que **Medium** es la categoría con un mayor porcentaje de ventas con el 57.4 por ciento.



Al mismo tiempo realizamos un contraste entre el gráfico de pastel anterior, con la siguiente tabla con la finalidad de indentificar el volumen en ventas generado por los diferentes tipos en **Prioridad de la Orden**, como podemos visualizar **Medium** es la categoría que mayor cantidad en ventas genera, seguido de **High** y **Critical**.

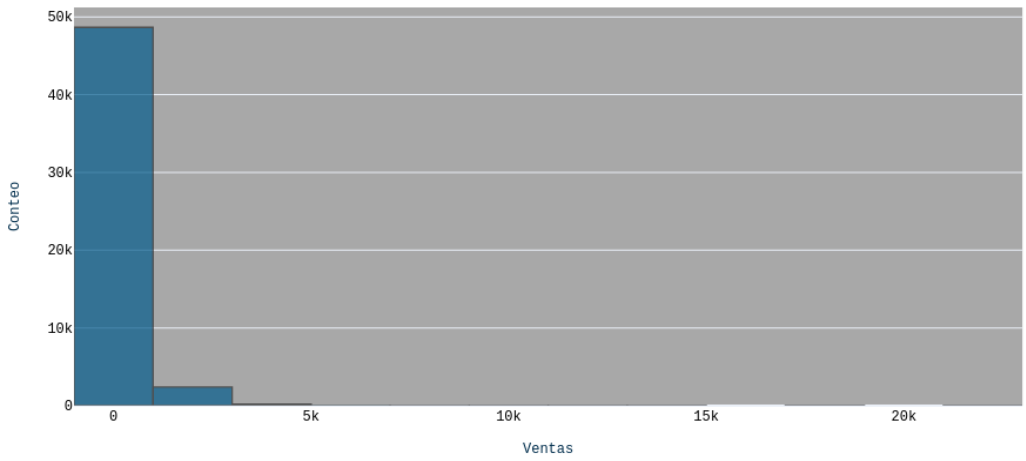
*** Es importante mencionar que las comparaciones que se están llevando a cabo son a nivel del dataset en general, para este caso no se decidió abordar el análisis a nivel producto o segmento de mercado, como lo hemos visto anteriormente. Sin embargo, cabe resaltar que es de gran importancia realizar el análisis a nivel producto para poder realizar la predicción de ventas que es nuestra variable objetivo.

	c_sales	c_quantity	c_discount	c_profit	c_shipping_cost
v_order_priority					
critical	1003240.529780	13468	536.980000	125052.812280	236724.675360
high	3820084.463400	54046	2222.624000	419644.585400	511448.857000
low	565789.068180	8290	344.692000	58289.216180	65108.325000
medium	7280892.374120	102555	4226.332000	864203.762620	542812.816040
nan	2036.860000	7	0.000000	366.634800	524.760000

2.4.2. Variables Continuas

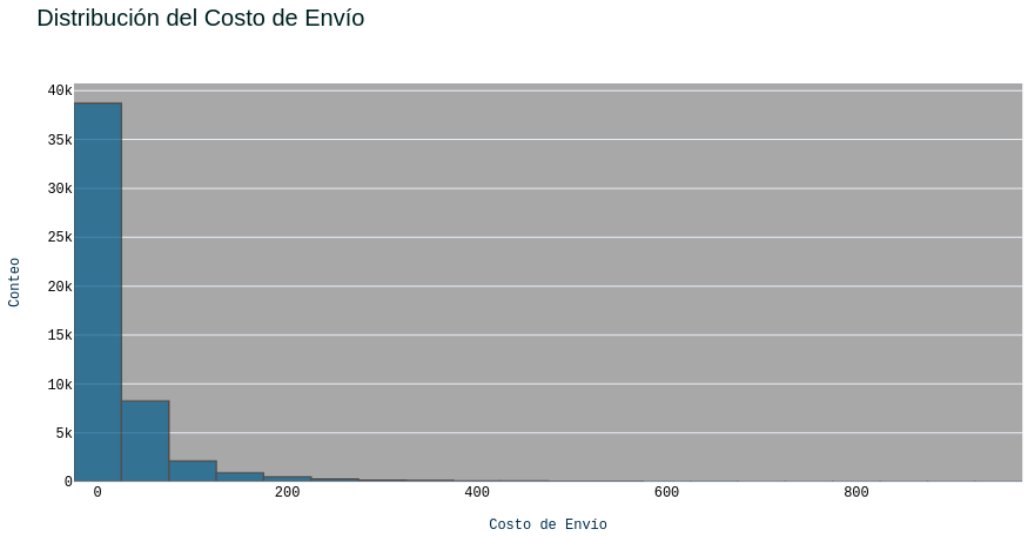
La variable target de nuestro estudio es predecir el volumen en ventas por producto. Para este caso a continuación se muestra la distribución de las ventas:

Distribución del Volumen de Ventas



El histograma de ventas muestra una **distribución simétrica positiva**, y esto se deriva como lo vimos anteriormente debido a que la mayor parte de las ventas se concentra en la categoría **Office Supplies** cuyos rangos de venta se encuentran entre 0.44 USD a los 1000 USD. Por otra parte, es importante resaltar que parte de nuestro análisis previo permitió identificar que **Technology** fue la categoría que mayor volumen de ventas produce.

Por otro lado, en la siguiente figura se muestra la distrubución del **Costo de Envío**. Al igual que la distribución de ventas, el comportamiento de ésta es **simétrica positiva** y el mayor volumen de los envíos no exceden los 25 dólares, dado que como lo vimos anteriormente el mayor mayor porcentaje de artículos vendidos se concentra en **Office Supplies** cuyos costos de envío se mueven en un rango desde los 0.44 a 25 dólares. Los bins siguientes corresponden a los costos de envío de **Tecnology** y **Furniture**.



2.5. Ingeniería de Variables

Para esta sección se utilizó **One-Hot Encodding** para generar las **Dummies** en las **Variables Categóricas**, listadas a continuación:

- v-order-id
- v-ship-mode
- v-segment
- v-city
- v-state
- v-country
- v-market
- v-region
- v-product-id
- v-category
- v-sub-category
- v-order-priority

Sin embargo de éstas, **v-order-id** y **v-product-id** no fueron utilizadas en el proceso de ingeniería de variables, por lo tanto una vez aclarado este punto. Las únicas variables empleadas durante el proceso son:

- v-ship-mode
- v-segment
- v-city ***
- v-state ***
- v-country ***
- v-market
- v-region
- v-category
- v-sub-category
- v-order-priority

En la primera prueba se utilizaron las 10 variables anteriores, obteniendo las siguientes dimensiones para los dataframes **X-train** y **X-test**: (41,038, 4,810) y (10,260, 4,810). En este primer test podemos ver que que la cantidad de columnas en las tablas resultantes no excede el número de registros, y cumplimos con la regla de que el número de características no sea mayor al número de registros.

Sin embargo, es importante mencionar que en una segunda prueba podemos reducir aún más el número de variables dummy si no consideramos: ciudad, estado y país, limitando el análisis únicamente al mercado y a la región de venta de los productos.

Para las variables **Tipo Texto** se utilizó **Count Vectorizer**:

- t-product-name

*** La otra variable de tipo texto presente en el dataset es t-customer-name, sin embargo, el único tratamiento que se le aplicó fue convertirla a minúsculas manteniendo los espacios entre nombre y apellido.

El tratamiento aplicado al nombre del producto fueron:

- Integrar los nombres de los productos vendidos de acuerdo a cada registros en un corpus.

- Identificar número de **Hapaxes** presentes en el **Corpus**.
- Calculamos el ratio entre los hapaxes y el corpus.
- Eliminamos hapaxes tanto en **X-train** y **X-test**.
- Obtención de **Tokens**.
- Lematización.

*** Se eliminaron 105 hapaxes, que como parte del análisis se idenficó que un 68.6 % correspondían a módelos, claves y/o acrónimos en particular que acompañan a los productos, propiamente en productos de la categoría de Tecnología.

*** Otro punto importante de mencionar, es que el modelos de vectorización visto en clase no fue utilizado para el estudio de esta variable, pues se observó que más del 80 % de los productos vendidos se mueven en un rango de 1-2 en frecuencia de compra por parte de los clientes. Esto ocasiona que no se cumpla con la regla de que las palabras aparezcan con al menos el 15 % de frecuencia dentro de cada registro y corpus en general. Por lo anterior no se empleó la vectorización de palabras.

Capítulo 3

Tratamiento y Limpieza de los Datos

3.1. Duplicidad

Para identificar y eliminar duplicados se utilizó el método **df.drop-duplicates** directamente, también se considero el escenario de quitar duplicados por el método de **id** pero no fue aplicado dado que un producto con el mismo id puede ser vendido a diferentes clientes, en distintos mercados y países por mencionar algunos ejemplos. El resultado fue de 59 registros duplicados identificados y eliminados.

3.2. Normalización de Variables Categóricas

Dentro de esta sección se consideró normalizar la variable *v-sub-category* en primera instancia, sin embargo durante el análisis exploratorio se identificaron 17 sub-categorias dentro de la misma que desde la perspectiva del análisis son necesarias para colocar los diferentes productos vendidos con el grado de detalle del datset.

Sin embargo, algunos productos como mesas y sillas pueden ser considerados dentro de muebles y de esta forma estariamos pasamos de 17 a 15 sub-categorías.

3.3. Valores Extremos

Para identificar los **outliers** se utilizaron los 3 métodos vistos durante el módulo: IQR, Percentiles y Z-Score, al final se eligieron: IQR y Percentiles dado que el comportamiento de la distribución de las variables en cuestión no tenía similitud a una distribución nor-

mal. Sin embargo, es importante mencionar que los valores identificados como outliers no necesariamente lo sean, pues como se vió en el EDA el rango de precios entre categorías y sub-categorías tiene variaciones significativas que dependen del tipo y cantidad de productos vendidos. De acuerdo al análisis se tienen registros de ventas de artículos de oficina con precios entre 0.44 y 25 dólares, comparado con el precio de artículos electrónicos de van desde los 100 a 225 dólares.

3.4. Valores Ausentes

Después de visualizar el comportamiento de algunas variables de interés en el dataset y eliminar outliers, se obtiene la siguiente tabla:

	columna	total	completitud
0	v_market	2	99.996101
1	c_profit	0	100.000000
2	c_discount	0	100.000000
3	c_quantity	0	100.000000
4	c_sales	0	100.000000
5	t_product_name	0	100.000000
6	v_sub-category	0	100.000000
7	v_category	0	100.000000
8	v_product_id	0	100.000000
9	v_region	0	100.000000
10	v_order_id	0	100.000000
11	v_country	0	100.000000
12	v_state	0	100.000000
13	v_city	0	100.000000
14	v_segment	0	100.000000
15	t_customer_name	0	100.000000
16	customer_id	0	100.000000
17	v_ship_mode	0	100.000000
18	d_ship_date	0	100.000000
19	d_order_date	0	100.000000
20	c_shipping_cost	0	100.000000
21	v_order_priority	0	100.000000

Como se puede ver, para tener una completitud del 100 % en todas las avraibles, se requiere imputar 2 valores faltantes en **v-market**. Para esto, imputamos utilizando la **Moda**: Aceptamos HO dado que la proporción de categorías es la misma que la general y a través de esto reemplazamos los valores ausentes.

Verificamos el conteo y completitud de **v-market** que se obtuvo después de imputar los valores ausentes:

```
APAC      8888
LATAM     8172
US        8028
EU        7972
EMEA      4058
Africa    3608
Canada     312
Name: v_market, dtype: int64
```

Completitud Después de Imputar Missings

	columna	total	completitud
0	v_order_id	0	100.000000
1	c_profit	0	100.000000
2	c_discount	0	100.000000
3	c_quantity	0	100.000000
4	c_sales	0	100.000000
5	t_product_name	0	100.000000
6	v_sub-category	0	100.000000
7	v_category	0	100.000000
8	v_product_id	0	100.000000
9	v_region	0	100.000000
10	v_market	0	100.000000
11	v_country	0	100.000000
12	v_state	0	100.000000
13	v_city	0	100.000000
14	v_segment	0	100.000000
15	t_customer_name	0	100.000000
16	customer_id	0	100.000000
17	v_ship_mode	0	100.000000
18	d_ship_date	0	100.000000
19	d_order_date	0	100.000000
20	c_shipping_cost	0	100.000000
21	v_order_priority	0	100.000000

Capítulo 4

Reducción de Variables

4.1. Filtro de Alta Correlación

Para la reducción de variables se utilizaron los siguientes enfoques:

- Filtro de Alta Correlación
- Correlación con el Objetivo

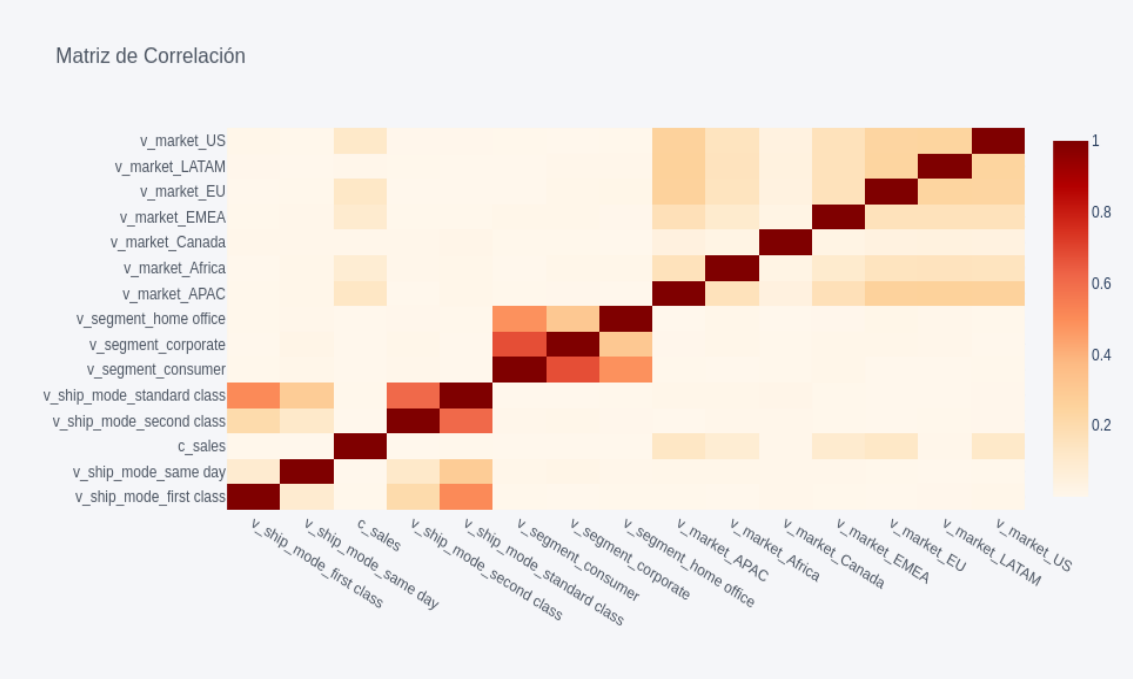
En esta sección se explica las variables utilizadas, el método y el resultado obtenido a traves de un mapa de calor en donde se indica el comportamiento de la correlación entre las variables:

Variables Uilizadas

- v-ship-mode-first class
- v-ship-mode-same day
- v-ship-mode-second class
- v-ship-mode-standard class
- v-segment-consumer
- v-segment-corporate
- v-segment-home office
- v-market-APAC

- v-market-Africa
- v-market-Canada
- v-market-EMEA
- v-market-EU
- v-market-LATAM
- v-market-US
- **c-sales** *target*

En la siguiente figura se muestran los resultados obtenidos (método empleado:**Spearman**):



Al momento de identificar aquellas variables que tienen un valor de correlación superior al 0.7, no tenemos alguna mayor a este valor. Las más cercanas, pero no superiores a dicha cota fueron:

- v-ship-mode-standard class y v-ship-mode-second class con 0.613
- v-segment-corporate y v-segment-consumer con 0.678

Finalmente tenemos como resultados que a través de este método, ninguna variable tiene alta correlación entre si, pues no exceden el valor de 0.7 especificado.

*** Para las tablas finales se seleccionó el Filtro de Alta Correlación, porque en el momento de comparar con el método de Correlación con el Objetivo identificamos que perdemos 10 variables de las 14 utilizadas en el proceso lo cual considero es un alto porcentaje de características que se descartan en el análisis, pero al mismo el segundo método permite identificar que esas 10 variables no estarían contribuyendo objetivamente para efectos de predicción y análisis posteriores. ****consultaría con la profesora si mi argumento es válido y si el método seleccionado fue la mejor opción****.

4.2. Correlación con el Objetivo

A través de este otro método los resultados obtenidos fueron los siguientes:

	c_sales
v_segment_consumer	0.000795
v_ship_mode_second class	0.000913
v_segment_home office	0.001208
v_ship_mode_same day	0.001490
v_segment_corporate	0.001791
v_ship_mode_first class	0.002727
v_ship_mode_standard class	0.003396
v_market_LATAM	0.005513
v_market_Canada	0.008309
v_market_Africa	0.085755

Las condiciones establecidas en el valor de correlación: aquellas variables con un valor menor al 10 % deben ser eliminadas del dataframe de prueba.