PROYECTO INTEGRADOR M0

Carrera: Data Analytics

Cohorte: Data Analytics Part-Time 09

Alumno: Hillebrand, Francisco Javier

28 de Mayo de 2025

Índice

Introducción

Este documento pretende ser la documentación que acompañe a las actividades realizadas durante el proyecto. En él se detallarán todas las consideraciones necesarias que justifiquen cada una de las decisiones tomadas en las distintas actividades del proyecto integrador.

El Documento se dividirá en 4 secciones:

1. Una dedicada a aclarar cuáles son los pasos que se realizaron para limpiar los datasets con los cuales se trabajó.
2. Otra dedicada a comentar las decisiones tomadas a la hora de realizar el DER y el Modelo Relacional del escenario.
3. Una tercera sección en la cual se comentarán todos los insights encontrados y que sean relevantes para el análisis de los datos.

Limpieza de Datos

Antes de adentrarnos en la transformación de cada una de las columnas de cada hoja, es necesario hacer dos aclaraciones generales:

- La configuración regional del archivo se cambió a la opción Argentina, y además se colocó la zona horario de Buenos Aires (GMT - 03:00).
- Para los números se utilizó la coma como separador decimal.
- Para importar los datos utilicé la opción “Importar” en la pestaña “Archivo”, y ahí importé los tres CSV desde mi computadora. Cada CSV los importe en tres hojas distintas, cada una tiene el nombre del CSV que importé.

Hoja “STORE”

Columna	Formato	Aclaración
Store	Texto sin Formato	Si bien se identifican con números, estos no se van a utilizar para realizar cálculos, sino que solo para identificar a las tiendas.
Type	Texto sin Formato	Debido a que son letras, decidí colocarle este formato.
Size	Número	Considere que el tamaño de las tiendas se mide en metros cuadrados, luego de investigar, pueden haber tiendas con 300m^2, y también tiendas con 2.000m^2 o más.

Hoja “SALES”

Columna	Formato	Aclaración
Store	Texto sin Formato	Si bien se identifican con números, estos no se van a utilizar para realizar cálculos, sino que solo para identificar a las tiendas.
Dept	Texto sin Formato	Al igual que Store, aunque los departamentos se identifiquen con números, estos solo serán para identificarlos y no para realizar algún tipo de cálculo.

Columna	Formato	Aclaración
[1]Date	Fecha	Al tratarse de fechas, coloqué el formato dedicado a estos datos.
Weekly_Sales	Moneda (Dólar Estadounidense)	Al tratarse de ventas consideré que estos registros eran valores monetarios. Además, encontré valores negativos, y considere que no pueden haber valores negativos para las ventas, ya que no son perdidas, por ejemplo. Entonces considere que estos valores fueron ingresados de manera incorrecta, y decidí trabajar solo con los valores absolutos del registro. Para esto último decidí utilizar la función buscar y reemplazar, para eliminar el signo negativo en toda la columna.
IsHoliday	Número	Si estos datos se van a almacenar en una base de datos, considero que es más fácil comparar números que palabras, por ende preferí hacer el siguiente cambio: FALSE = 0 TRUE = 1
Location	Texto sin Formato	Son las abreviaciones de los estados en que están las tiendas, por ello el formato es texto

Hoja “FEATURES”

Columna/s	Formato	Aclaración
Store	Texto sin Formato	Si bien se identifican con números, estos no se van a utilizar para realizar cálculos, sino que solo para identificar a las tiendas.
Date	Fecha	Al tratarse de fechas, coloqué el formato dedicado a estos datos.

[1] Comprobé con la función text, que el día de las fechas de esta columna, son viernes, es decir, es el día en que finaliza la semana a la cual se refiere la columna Weekly_Sales

Columna/s	Formato	Aclaración
Temperature	Número	Coloque el formato número personalizado (°F) para especificar que los grados están en grados fahrenheit.
Fuel_Price	Moneda (Dólar Estadounidense)	Detecté que había valores atípicos en esta columna, ya que el precio del combustible en Estados Unidos entre 2011 y 2012, estuvo entre los $3 y $5 aproximadamente, y no $1000 o $2000, como veíamos en los registros. Para solucionar esto, cree una columna al lado, y en ella [2]dividí a los valores mayores o igual a 1.000 (que no corresponden) por 1.000, para así tener los datos normalizados y en la unidad correcta. La columna con los datos originales la oculté.
MarkDown1-5	Moneda (Dólar Estadounidense)	Interprete que esos registros eran las ganancias en productos con las respectivas promociones, por ende, lo tomé como un valor monetario. Los NA (valores no disponibles) los cambie por el valor 0.
CPI	Número	En la columna había valores atípicos, como 129.089, que asumí (luego de investigar como se mide el CPI en Estados Unidos) que deberían ser 129,089. Hice lo siguiente: cree una nueva columna e ingrese la siguiente formula para cada celda, =IF(J2>=1000;J2/1000;J2) entonces aquellos datos que están en miles de puntos, pasarán a estar en cientos de puntos cómo debe ser.
Unemployment	Porcentaje	Primero pasamos la columna a formato numero. Luego, es importante saber que la columna debe estar en

[2] Fórmula aplciada: =IF(D2>=1000;D2/1000;D2)

Columna/s	Formato	Aclaración
		formato porcentaje, por ende todos los calores mayores a mil se dividen por 100.000, y los valores menores a 100 se dividen por 100, así tendremos el porcentaje correcto. La columna con los datos originales la oculte.
IsHoliday	Número	Si estos datos se van a almacenar en una base de datos, considero que es más fácil comparar números que palabras, por ende preferí hacer el siguiente cambio: FALSE = 0 TRUE = 1

Modelado de Datos y Creación de Base de Datos

Siguiendo este link podrá ver el DER y el Modelo Relacional:
[MODELOS DE DATOS PIM0: Lucidchart]

Consideraciones del DER

- El nombre de las entidades respetan los nombres de las columnas de la hoja de cálculo donde se encuentran los datasets
- Consideré que los datos de la columna Weekly_Sales dependen funcionalmente de la fecha, la tienda y del departamento donde se registraron las ventas.
- Consideré que los datos de la columnas Markdown1, Markdown2, Markdown3, Markdown4, Markdown5, dependen funcionalmente de la fecha, la tienda, y además del tipo de promoción. Al hablar de tipo de promoción me refiero a la numeración de los MarkDown, es decir del 1 al 5. Serán estos los posibles valores del campo idMarkdown, en la entidad MARKDOWN.
- Hay relaciones con atributos, estas serán tablas en el Modelo Relacional.
- Los posibles valores del campo idTipo en la entidad TYPE, son: A, B Y C

Consideraciones del Modelo Relacional

- El Modelo Relacional se construyó a partir del DER teniendo en cuenta las siguientes transformaciones:
    - Las relaciones con cardinalidad N:M (muchos a muchos), se convertirán en tablas, y tendrán como PK y FK los id de las entidades que se relacionan. Si tienen atributos, estos serán columnas en las tablas.
    - En las relaciones con cardinalidad 1:N, la PK de la entidad del lado 1 de la relación, será FK en la tabla que representa la entidad del lado N de la relación.
- Las tablas que se originan a partir de las relaciones N:M tienen como nombre la unión de los nombres de las entidades relacionadas.[3]

[3] Por ejemplo, la tabla originada a partir de la relación N:M entre las entidades STORE y DEPARTMENT, se llama STORE_DEPARTMENT

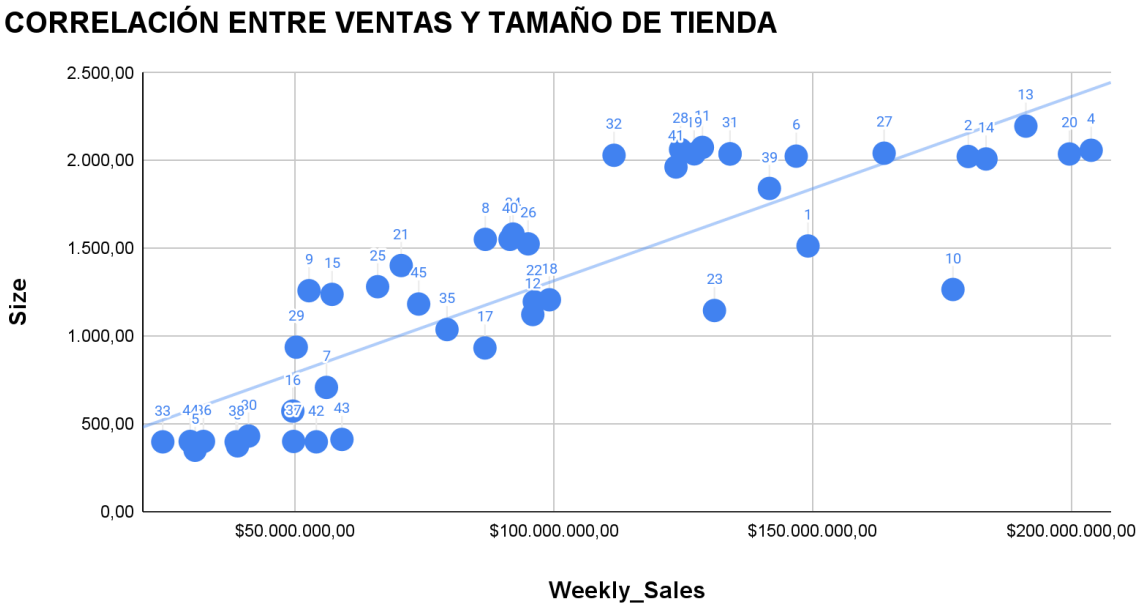
Búsqueda de Insights

En el Warehouse, en la hoja “DATOS INSIGHTS” se encontrarán las tablas dinámicas utilizadas para realizar los gráficos vistos en este documento. En la hoja “INSIGHTS” se pueden ver solo los gráficos.

Aclaración: Siempre que se mencionen las ventas totales, me refiero a la suma de las ventas semanales, más los montos vendidos por cada markdown.

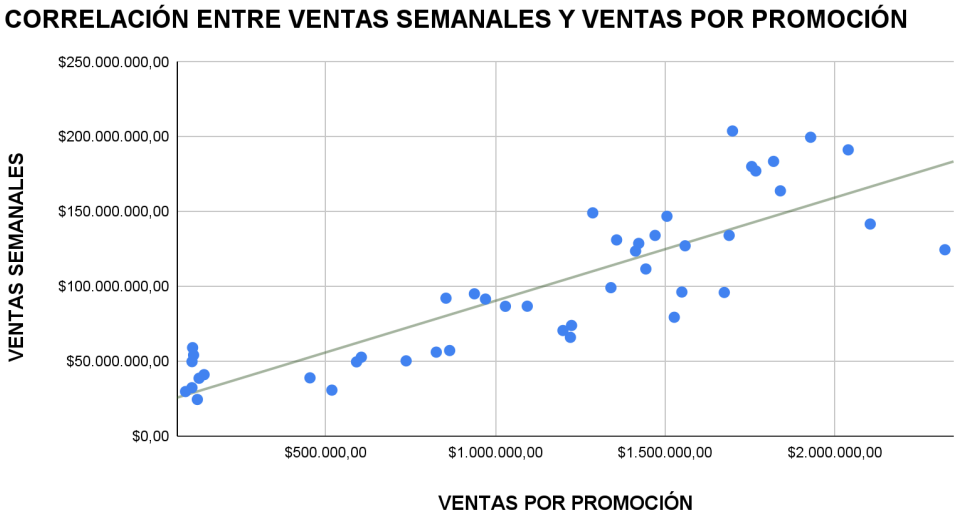
Las tiendas más grandes, ¿Venden más?

Para descubrir esto realicé un gráfico de dispersión donde vemos que puede haber una correlación positiva entre ambas variables, es decir, las tiendas de mayor tamaño suelen tener más ventas. Igualmente faltaría descubrir por qué, ya que esto podría deberse a distintos motivos como la variedad de productos ofrecidos, o bien la cantidad de promociones que ofrecen.



Las tiendas que más venden, son las que más venden a través de promociones (Markdown)

Este gráfico complementa al anterior, las más grandes son las que más venden y al parecer las que más ganan por promoción

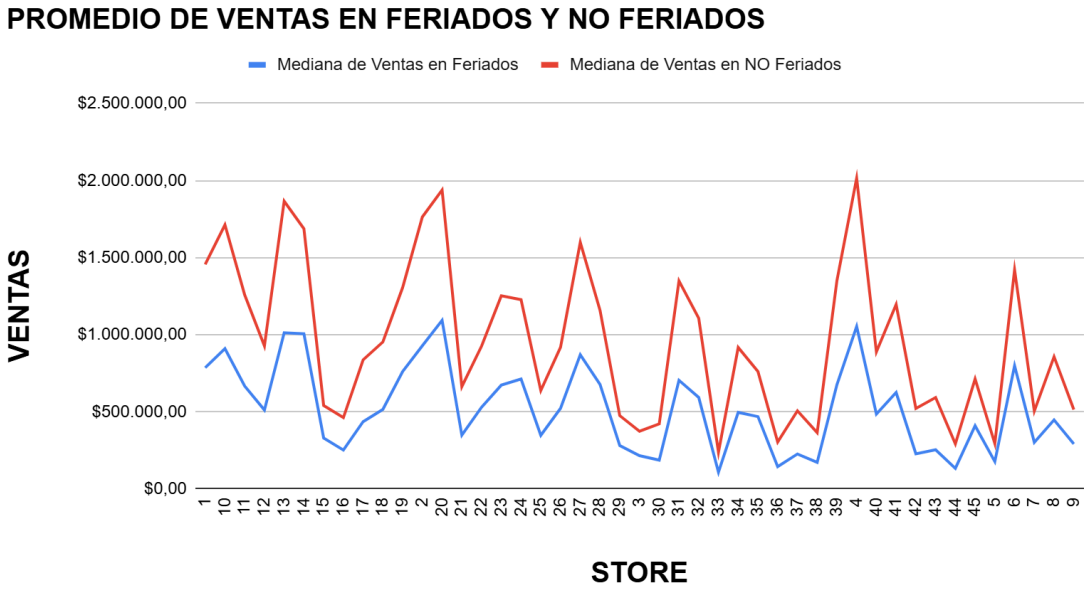


Comparación de Ventas en días Feriados y no Feriados

Para este gráfico utilice 4 tablas dinámicas[4], dos de ellas estaba filtrado por días no feriados y la otra por días no feriados, utilice dos de ellas para calcular la mediana de las ventas totales (suma de las ventas semanales y de las ventas por markdown) en días feriados y las otras para calcular las ventas totales en días no feriados.

Decidí utilizar la mediana, ya que al calcular el promedio y la desviación estándar, noté que esta última era alta en varios casos, por ende no consideré una buena medida central al promedio.

Considero que es normal que se venda menos en feriados ya que hay menos de estos días, igualmente hay que destacar que en los días feriados se suele vender en promedio, aproximadamente la mitad de lo que se suele vender en los días no feriados.

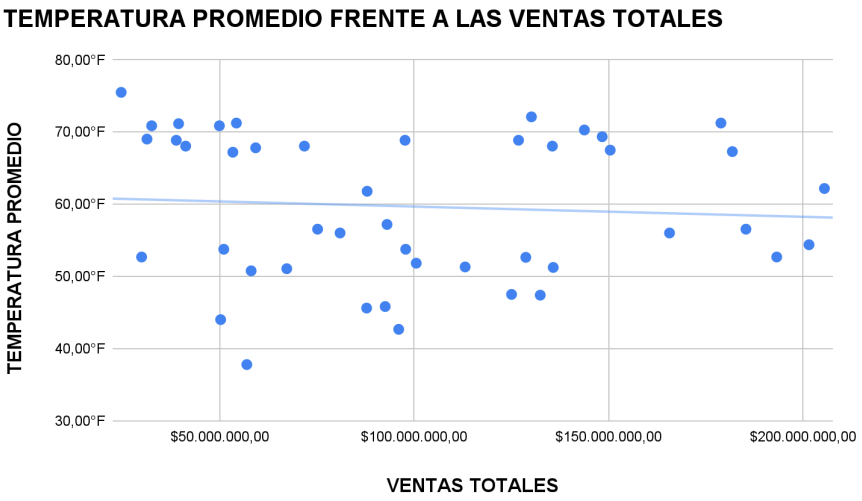


[4] Las tablas dinámicas están en el rango AD1:BW181 de la hoja “DATOS INSIGHTS”

¿La temperatura impacta en las ventas?

En el siguiente gráfico se observa que hay un leve aumento de las ventas a medida que las temperaturas del lugar donde se encuentra la tienda baja.[5]

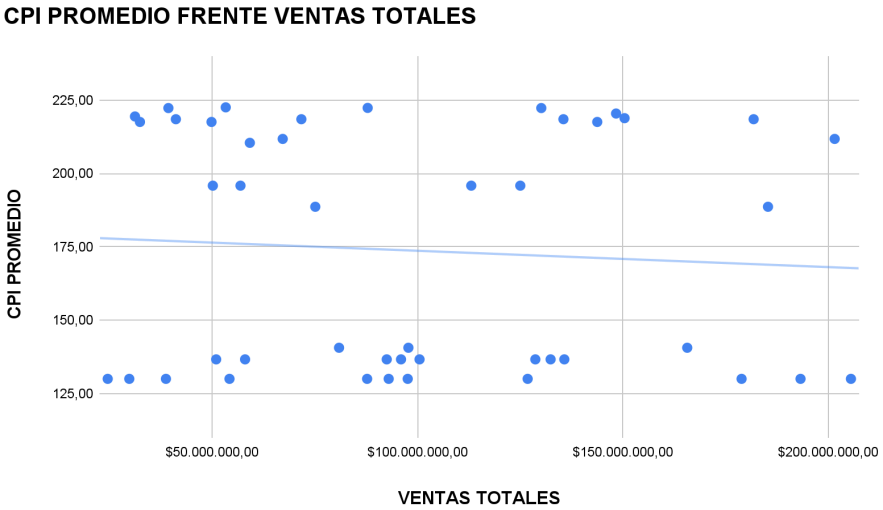
Para este gráfico decidí utilizar el promedio de las temperaturas registradas en el estado donde se encuentra cada tienda, esto debido a que la desviación estándar no era muy alta.[6]



¿El CPI impacta en las ventas?

En el siguiente gráfico se observa que hay una pequeña tendencia a la baja, es decir a medida que cae el CPI aumentan las ventas, pero la correlación es muy baja[7].

Para este gráfico decidí utilizar el promedio del CPI registrados en el estado donde se encuentra cada tienda, esto debido a que la desviación estándar no era muy alta.[8]

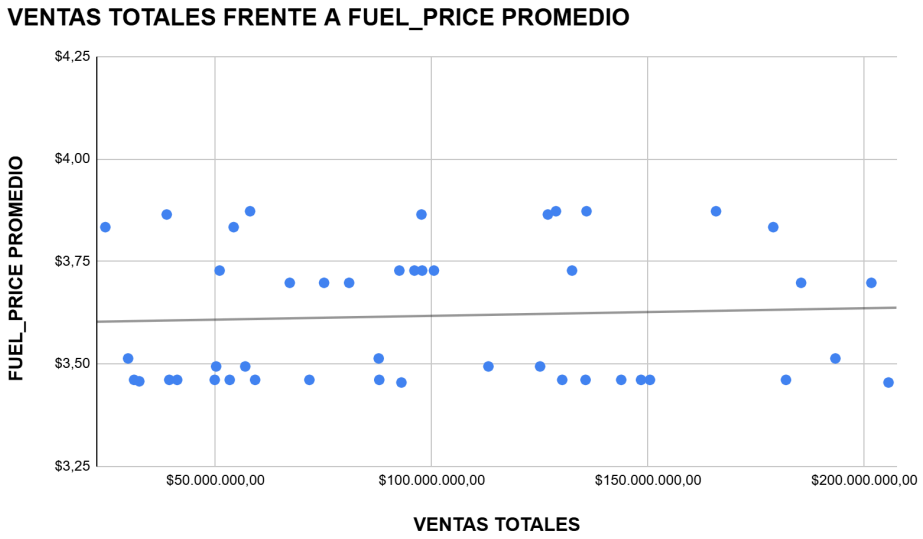


[5] La tabla utilizada para el grafico se encuentra en el rango DU1:DZ46 de la hoja “DATOS INSIGHTS”
[6] La desviación estándar se encuentra en el rango BY108:DR108 de la hoja “DATOS INSIGHTS”
[7] La tabla utilizada para el grafico se encuentra en el rango DU1:DZ46 de la hoja “DATOS INSIGHTS”
[8] La desviación estándar se encuentra en el rango BZ218:DR218 de la hoja “DATOS INSIGHTS”

¿El precio del combustible impacta en las ventas?

Parecido a los insights anteriores, la tendencia es muy baja, al parecer hay un leve aumento de las ventas en las tiendas ubicadas en lugares donde el precio promedio del combustible es más alto.[9]

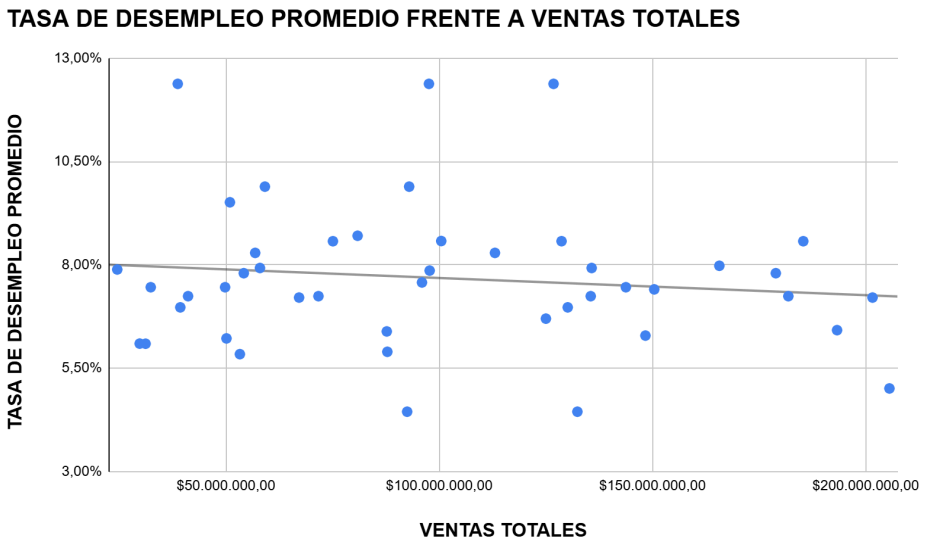
Utilicé el promedio ya que la desviación estándar era baja, por lo cual la consideré una buena medida central.[10]



¿La tasa de desempleo impacta en las ventas?

Hay una leve tendencia a la baja, es decir, en las tiendas ubicadas en estados con una tasa promedio de desempleo, se suele vender un poco más.[11]

De la misma manera, tomé el promedio ya que la desviación estándar no era alta.[12]

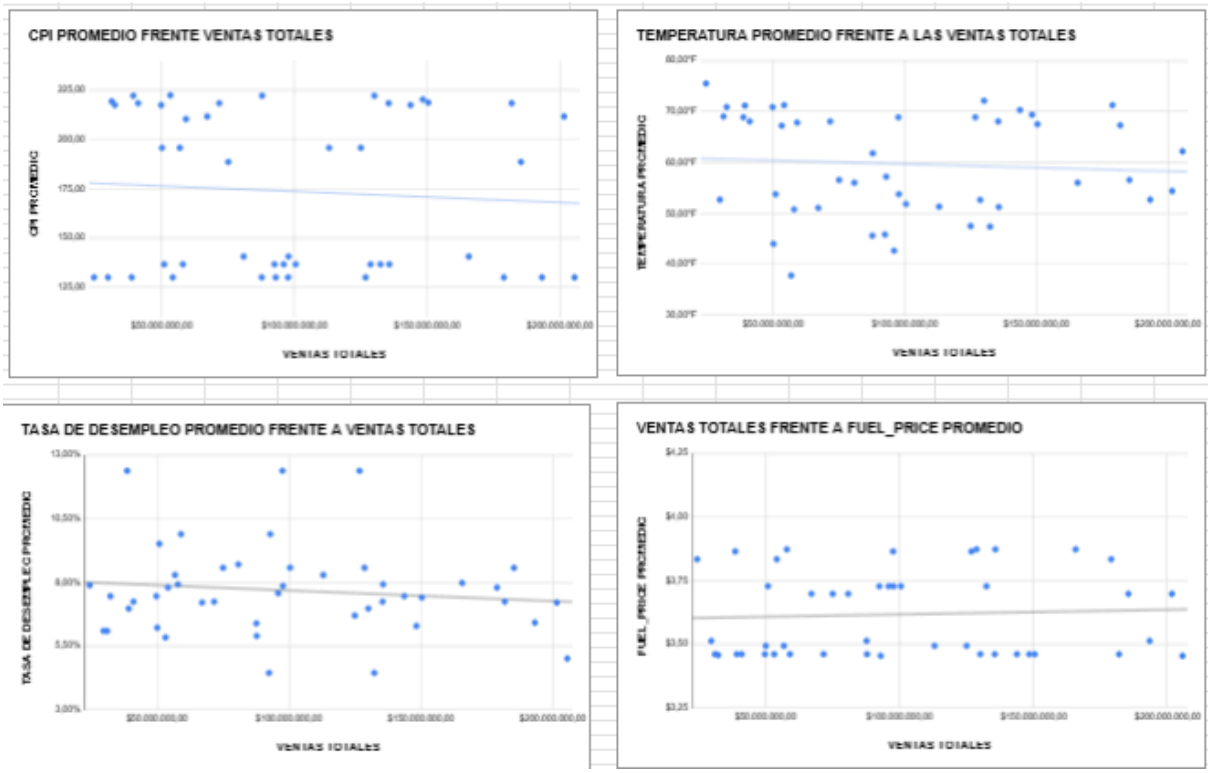


[9] La tabla utilizada para el grafico se encuentra en el rango DU1:EB46 de la hoja “DATOS INSIGHTS”
[10] La desviación estándar se encuentra en el rango BY327:DR327 de la hoja “DATOS INSIGHTS”
[11]La tabla utilizada para el grafico se encuentra en el rango DU1:EB46 de la hoja “DATOS INSIGHTS”
[12] La desviación estándar se encuentra en el rango BY437:DR437 de la hoja “DATOS INSIGHTS”

¿Qué factor externo afecta aún más a las ventas?

Viendo los 4 gráficos anteriores vemos que ninguna de las variables externas (temperatura, precio del combustible, CPI y tasa de desempleo), afecta claramente a las ventas, sino hay una variación pero muy pequeña.

Se puede destacar que solamente el aumento del precio del combustible genera un pequeño aumento en las ventas.

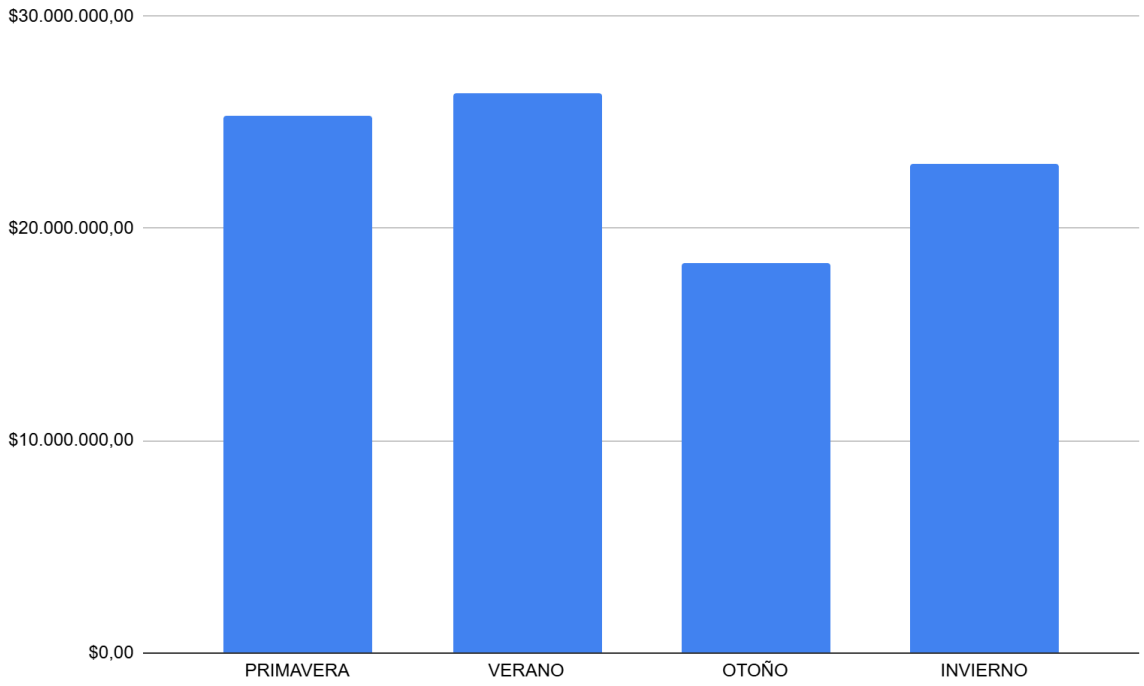


¿En qué estación del año se vende más?

En este gráfico vemos que nuestras tiendas suelen vender un poco más durante la primavera y el verano.[13]

Para este cálculo decidí trabajar con la mediana de las ventas de las 45 tiendas en cada estación del año, trabajé con ella porque al calcular el promedio y la desviación estándar de los datos, noté que esta última era muy alta, y finalmente al calcular la mediana, percibí que el promedio no era representativo como medida de tendencia central, ya que al parecer, el promedio podría estar sesgado por tiendas con ventas atípicamente altas.

[13] En el rango EN55:ER103 de la hoja “DATOS INSIGHT” se puede ver la tabla utilizada para realizar el gráfico.



Total de Ventas por Tienda

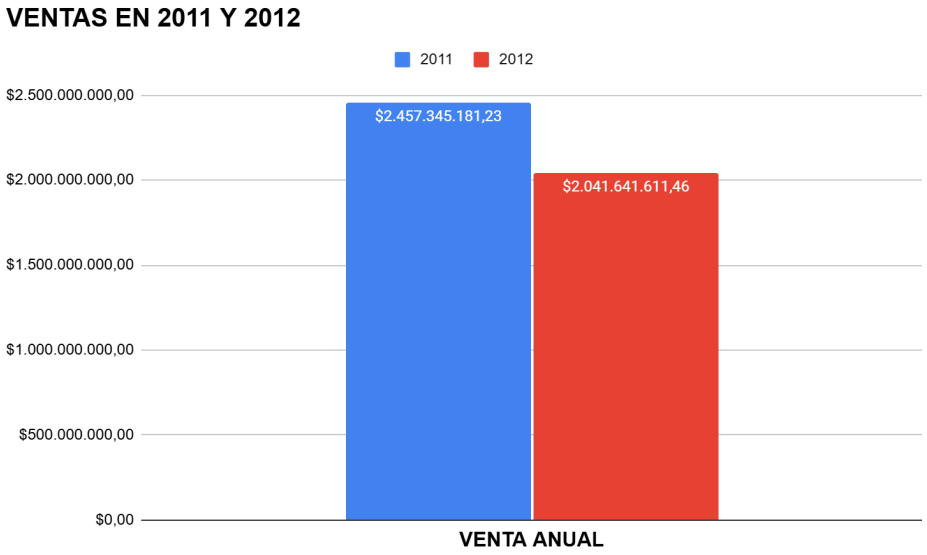
Se muestran las tiendas que más han vendido, si entra en la hoja “INSIGHTS” puede ver el orden de todas las tiendas.

Store	VENTAS TOTALES
4	$205.562.563,47
20	$201.597.626,27
13	$193.284.239,25
14	$185.359.316,57
2	$181.860.343,48
10	$178.920.628,68
27	$165.683.325,30
1	$150.415.232,19
6	$148.352.135,22
39	$143.768.283,93
19	$135.746.669,75
31	$135.524.979,38
23	$132.401.510,74
11	$130.132.333,78
24	$128.684.849,60
28	$126.812.317,99
41	$125.017.981,68
32	$113.072.712,71
18	$100.478.272,20
22	$97.761.076,48

1 2 3

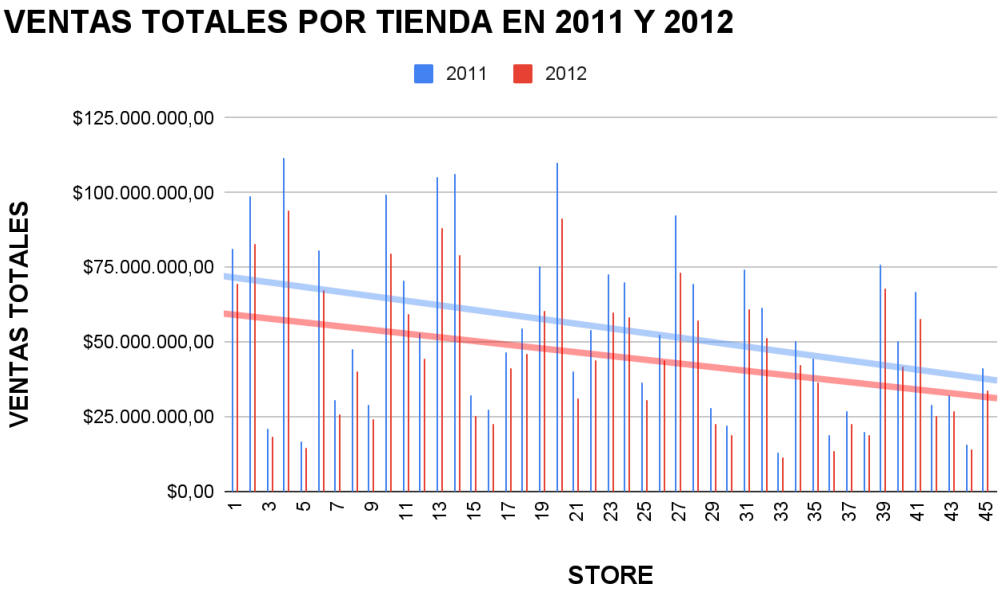
Total de Ventas Anuales

Este gráfico resalta el hecho de que hubo una baja de las ventas en el año 2012 en comparación al año 2011.[14]



Total de Ventas Anuales por Tienda

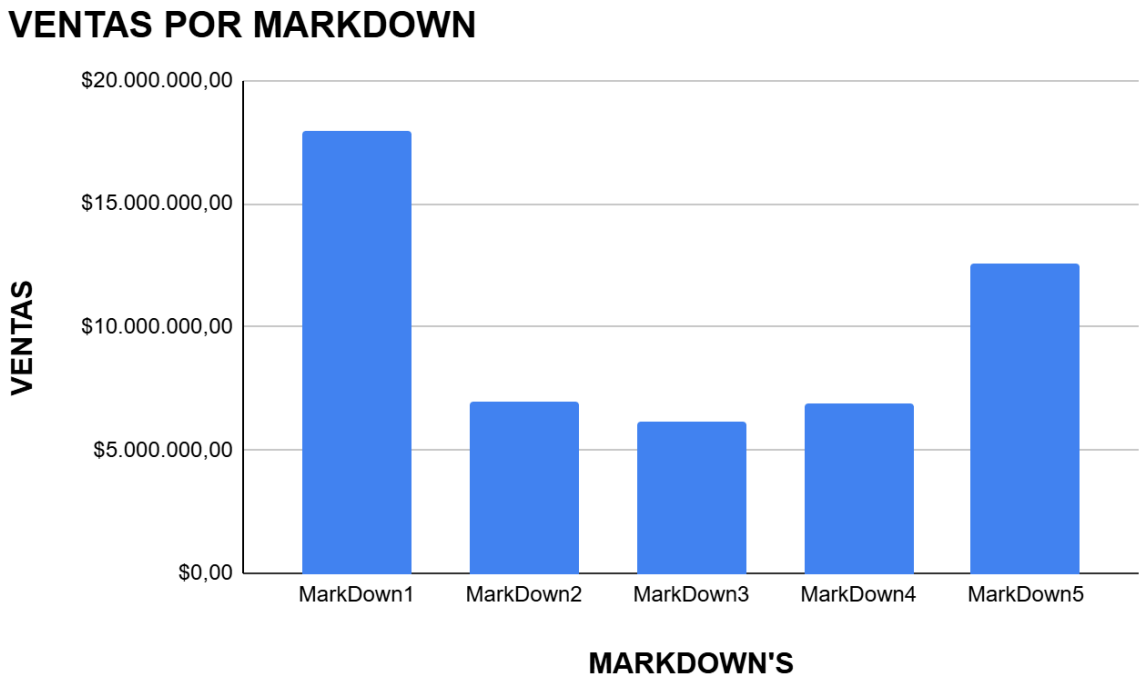
En este gráfico busco que se note que en todas las tiendas vendieron más en 2011 en comparación en 12, es decir, no es un comportamiento solamente general (como se veía en el anterior gráfico), sino que es algo que sucedió en todas las tiendas, las líneas de tendencia así lo demuestran.[15]



[14] En el rango EN1:EO3 de la hoja “DATOS INSIGHT” encontrarán la tabla con la información utilizada para hacer el gráfico.
[15]En el rango EN6:EP51 de la hoja “DATOS INSIGHT” encontrarán la tabla con la información utilizada para hacer el gráfico.

Total de Ventas por MarkDown

Vemos que la promoción más efectiva fue la 1, seguida por la 5. Las demás generaron ingresos parecidos.

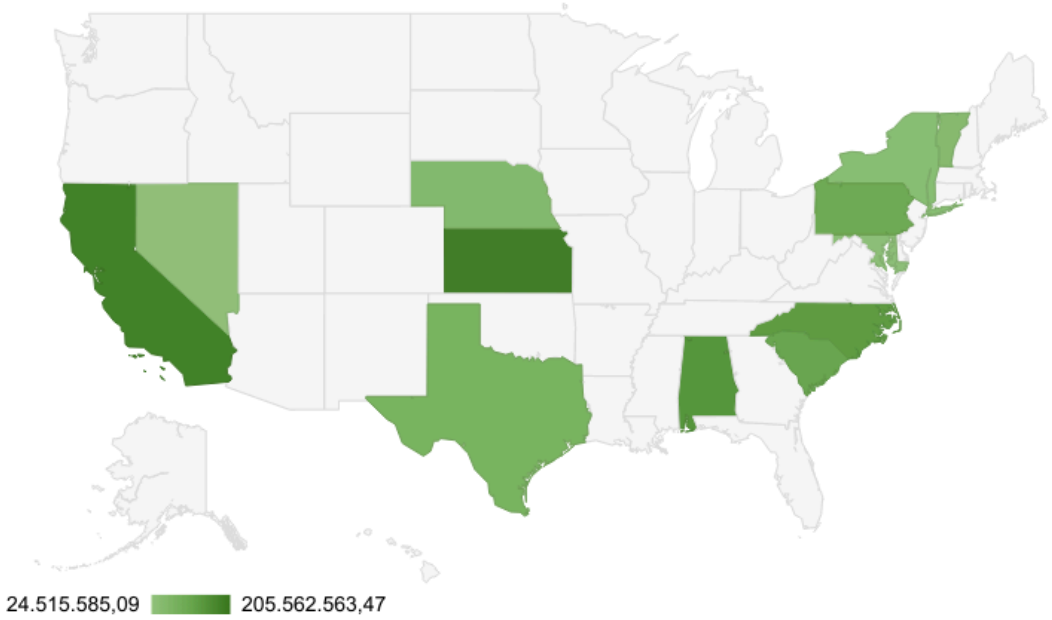


Total de Ventas por Estado

Si se ingresa a la hoja de cálculo se podrá ver el total de ventas por cada estado.

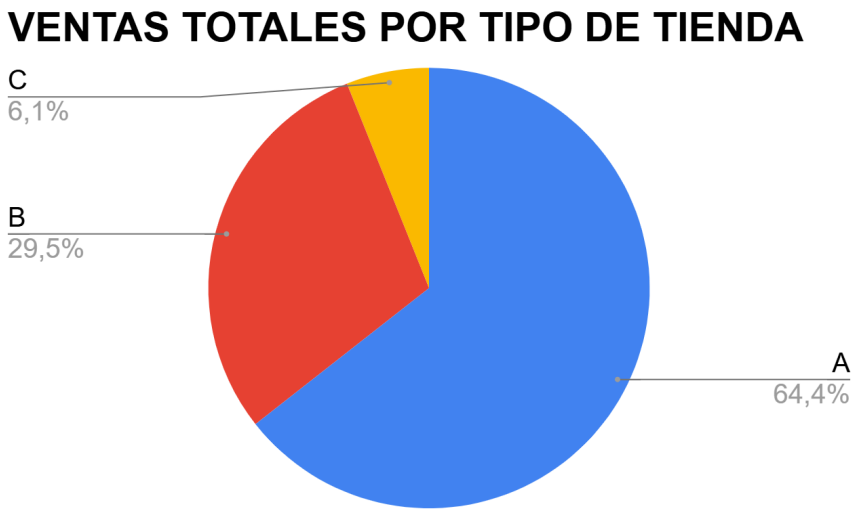
Los estados más claros son aquellos que vendieron menos y los de verde más fuerte son aquellos que más vendieron.[16]

[16] En el rango FL1:FK46 de la hoja “DATOS INSIGHT” encontrarán la tabla con la información utilizada para hacer el gráfico.



Total de Ventas por tipo de Tienda

Como vemos en el gráfico, las tiendas de tipo A son las que más venden, sin embargo hay que tener en cuenta el segundo gráfico, donde vemos que hay más tiendas de ese tipo, esto puede explicar que vendan más.[17]



[17] En el rango FR1:FT4 de la hoja “DATOS INSIGHT” encontrarán la tabla con la información utilizada para hacer los gráficos.


CANTIDAD DE TIPO DE TIENDA
C
13,3%
A
48,9%
B
37,8%

Conclusión

Luego de obtener los datos y realizar una limpieza de los mismos, de manera que estén presentados en el formato correcto (para así poder analizarlos adecuadamente), pude diseñar un modelo de datos de manera que estos mismo puedan ser almacenados en una base de datos relacional, base de datos la cual he intentado que pueda ser escalable por si surgen nuevos tipos de datos que sean necesarios analizar para descubrir insights, KPI's y métricas que nos muestren la actualidad de las ventas de nuestras tiendas de retail.

Gracias a la limpieza de los datos mencionada anteriormente, logré indagar en los datos y encontrar insights que ayuden a entender a la organización que variables impactaban o no en las ventas, variables como la temperatura, el CPI, el tamaño de las tiendas, entre otras.

Finalmente, presenté otros insights que permitan a la organización llevar un control de sus objetivos, de manera que le sean útiles para determinar si se están alcanzando o no estos mismos, por ejemplo, en uno de ellos se pudo validar el hecho de que en el año 2012 se vendió menos que el año anterior, y que esto se dió en todas las tiendas particularmente.

Los insghts fueron presentados en este informe y además en la hoja de cálculo llamada "WAREHOUSE", de modo que la organización tenga los datos y algunas aclaraciones que le sean útiles a la hora de tomar decisiones que lleven a las tiendas de retail al siguiente nivel.