Campus Querétaro

Asignatura

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

Tema

Momento de Retroalimentación: Reto Limpieza del Conjunto de Datos

Integrantes

Carolina Arratia Camacho

a01367552

Frida Lizett Zavala Pérez

a01275226

Fabián González Vera

a01367585

Jazzareth Bernal Martínez

a01367882

Fecha

27 de Agosto del 2023

- Limpien los datos usando herramientas de ETL.
- Expliquen y documenten cada decisión que hayan tomado sobre cómo limpiar los atributos y valores. Sean claros en sus explicaciones y concretos, respuestas ambiguas no serán tomadas en cuenta.
- Suban los scripts utilizados para limpiar los valores y su documentación a su repositorio de equipo del reto.
- Apliquen las transformaciones necesarias a los datos usando herramientas de ETL
- Expliquen y documenten cada decisión que hayan tomado sobre cómo transformar cada variable. Sean claros en sus explicaciones y concretos, respuestas ambiguas no serán tomadas en cuenta.
- Suban los scripts utilizados para transformar los valores y su documentación a su repositorio de equipo del reto.

Nuestro principal objetivo es crear un modelo que logre predecir las ventas unitarias de los productos vendidos en las diferentes tiendas de la marca *Favorita*, con el fin de que los minoristas tengan inventario de los productos al momento adecuado. Por este motivo, debemos examinar la columna de ventas seriamente, para identificar aspectos como la estacionalidad, tendencias, anomalías, similitudes con otras series temporales, etc.

Al graficar las ventas totales diarias de cada una de las tiendas observamos un aspecto importante, que es que en la parte inferior tenemos tiendas cuyas ventas son de 0 hasta cierto periodo de tiempo, esto se debe a que algunas tiendas abrieron mucho después pero dentro del dataset no está indicado simplemente marcan las ventas como 0. En este caso estos registros no aportan ninguna información necesaria para nuestro modelo, al contrario podría afectar la predicción ya que asumirá que durante un periodo de tiempo nuestras ventas fueron menores dado que tenemos valores de 0. Por lo tanto podemos identificar cuantos registros son los que se encuentran en ventas iguales a 0 para cada tienda.
Podemos ver que al filtrar los datos obtenemos que todas las tiendas tiene valores de ventas igual a 0 , esto puede deberse a días donde la tienda cerró o bien vacaciones, sin embargo vemos que los últimos datos se salen del rango de 500 cuando lo normal va de 10 a 100 días

con ventas igual a 0 por lo que es muy probable que estas sean las tiendas que abrieron mucho después y por eso su datos tienen muchas ventas en 0.

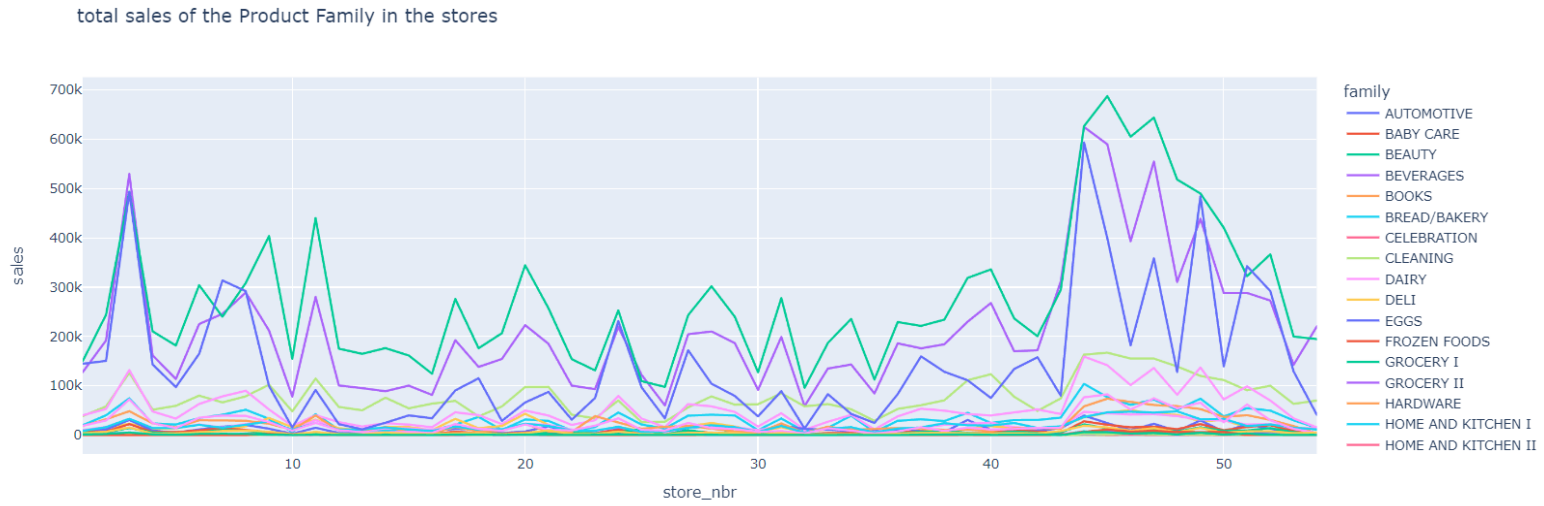
Tienda	Días con ventas de 0
52	1570
22	1015
42	966
21	938
29	812
20	777
53	519

Una vez que observamos las fechas donde dejan de tener días seguidos con ventas de 0 asumimos que es cuando se abren las tiendas y tomamos la misma fecha pero con un año más para utilizarla como referencia de punto de partida de los datos que son significativos para nuestro modelo, ya que si la tienda acaba de abrir el primer año sus ventas no serán muchas y puede afectar el modelo. Con este punto de partida eliminamos las filas de dichas tiendas donde la fecha sea menor a la propuesta, utilizando el siguiente fragmento de código:

```
df_train = df_train [~((df_train .store_nbr == 52) & (df_train .date < "2017-04-20"))]
df_train  = df_train [~((df_train .store_nbr == 22) & (df_train .date < "2015-10-09"))]
df_train  = df_train [~((df_train .store_nbr == 42) & (df_train .date < "2015-08-21"))]
df_train = df_train [~((df_train .store_nbr == 21) & (df_train .date < "2015-07-24"))]
df_train  = df_train [~((df_train .store_nbr == 29) & (df_train .date < "2015-03-20"))]
df_train  = df_train [~((df_train .store_nbr == 20) & (df_train .date < "2015-02-13"))]
df_train  = df_train [~((df_train .store_nbr == 53) & (df_train .date < "2014-05-29"))]
```

dejando nuestro Dataset con 2784540 de 3000000 datos

Por otra parte es importante revisar que no tengamos ninguna anormalidad con respecto a nuestras ventas por familia de productos en las tiendas, para eso graficamos la suma total de las ventas para cada una de las tiendas respecto a la familia de productos existentes, obteniendo la siguiente gráfica:



Como podemos observar en la parte inferior de la gráfica, existen productos cuya sumatoria total de venta de cierta familia de productos es igual a 0, esto es debido a que varias tiendas no ofrecen todos los productos por lo que la predicción de la venta de estos siempre será 0, por lo que no es necesario generar una predicción para estos caso y podemos omitirla de nuestro dataset, sin olvidar que al momento de generar nuestras predicciones tenemos que tomar en cuenta que si las ventas que queremos predecir de una tienda con una familia de productos que se encuentra entre estos nuevos valores su predicción será 0.

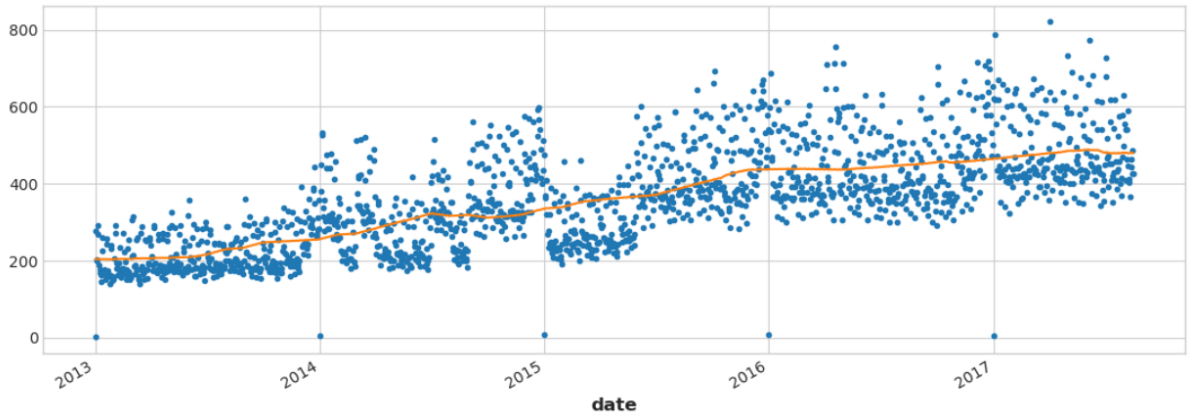
¿El terremoto afectó de manera significativa las ventas de las tiendas?

Dentro de la descripción de datos se nos menciona que: *"Un terremoto de magnitud 7,8 sacudió Ecuador el 16 de abril de 2016. La gente se unió a los esfuerzos de ayuda donando agua y otros productos de primera necesidad, lo que afectó en gran medida las ventas de los supermercados durante varias semanas después del terremoto."*

Por lo que uno de nuestros análisis fue detectar cuánto impacto tienen esta fecha para las nuestro dataset para ver si pueden afectar a nuestro modelo para predecir los datos.Primero vamos a analizar el promedio de ventas que se tuvo en cada año en el mes de Abril a Mayo que es cuando sucedió el terremoto.

Year	Mean Sales
2013	207.83k
2014	241.40k
2015	300.69k
2016	466.80k
2017	485.80k

En este punto podemos ver que si bien hubo una diferencia notable entre el año 2015 al 2016, una vez que graficamos la distribución de las ventas de todos los meses por año (gráfica que se muestra a continuación) podemos observar que el crecimiento en ventas por mes era igual para los demás meses.

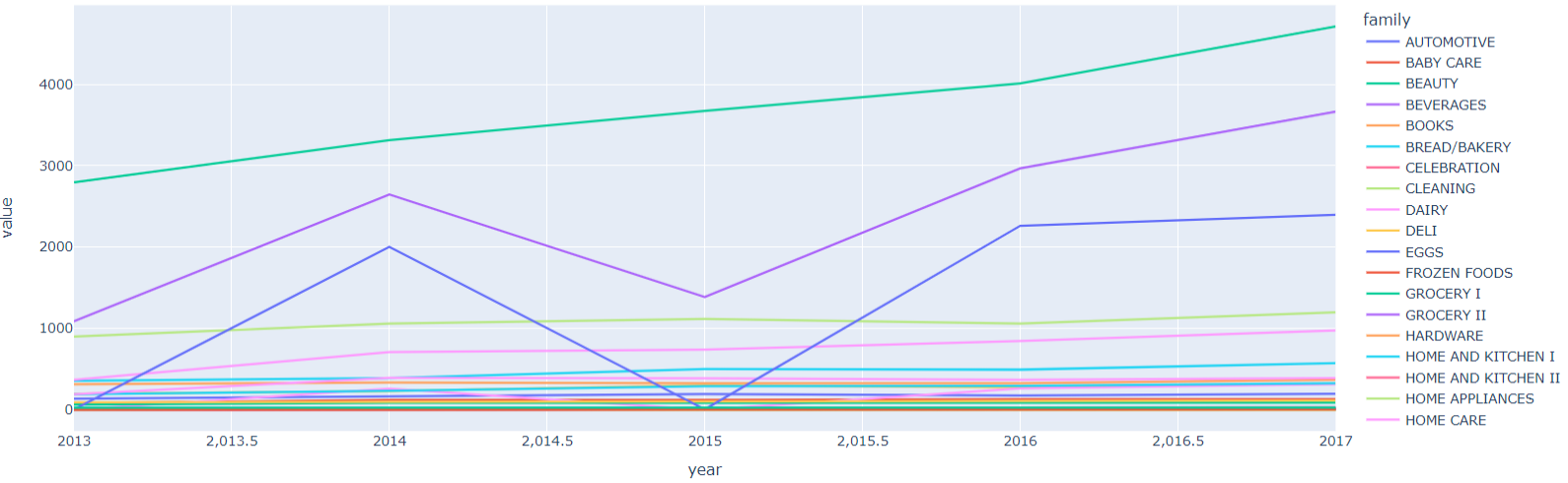


Hasta este punto podemos descartar que se tuvo un impacto relevante para las ventas generales de ese mes comparado a los anteriores años e incluso a los anteriores meses ya que la tendencia de las ventas parece ser lineal positiva.

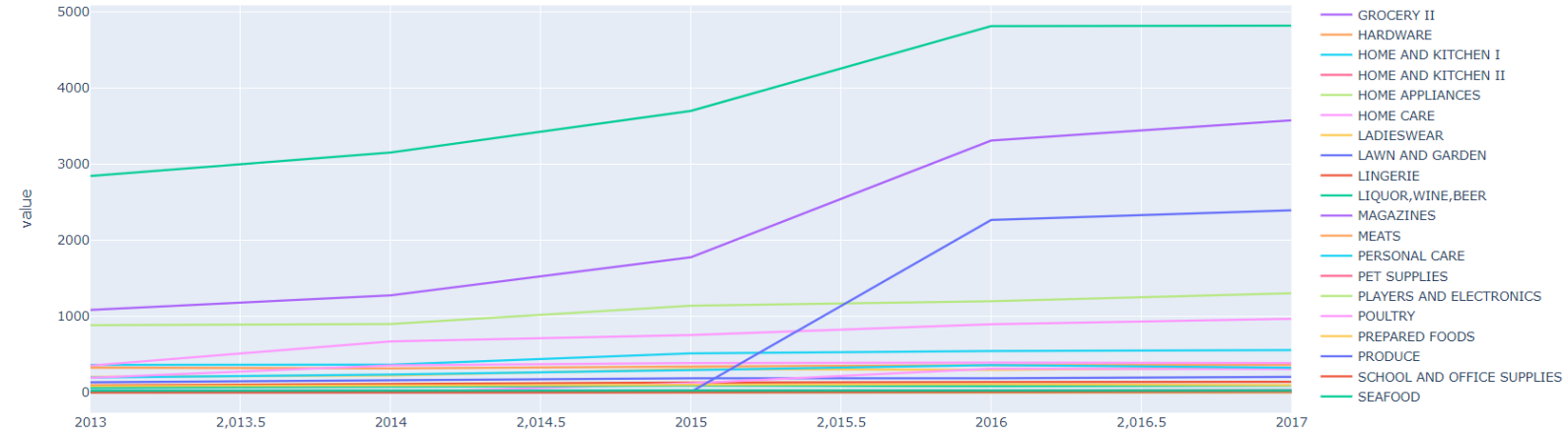
Sin embargo uno de los puntos que tenemos que tomar en cuenta es que no solo se analizan las ventas generales sino las ventas por familia de productos, para identificar si se tuvo o no

una diferencia notable en las ventas de cada familia de productos en ese mes los compararemos con el promedio de venta de el mes anterior y el mes siguiente.

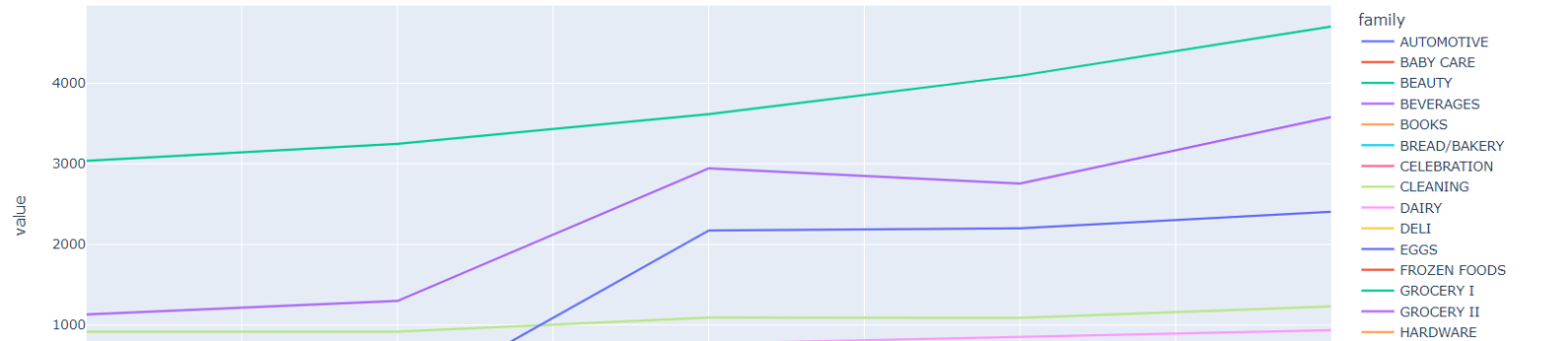
Marzo 2016



Abril-Mayo 2016



Junio 2016

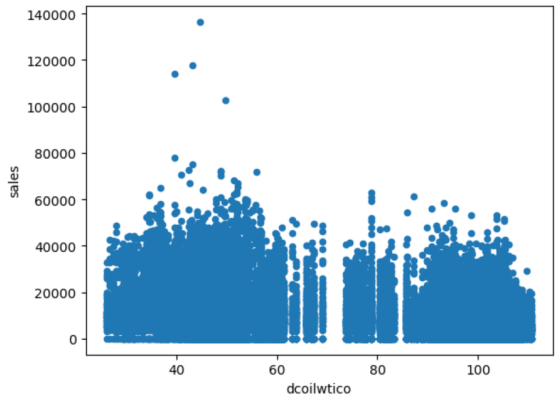


Con las gráficas podemos observar que si bien se tuvo un pequeño aumento en el consumo de cierta familia de productos, la diferencia con los demás meses y años no es lo suficientemente extrema como para afectar el desarrollo de nuestro modelo por lo que no es necesario limpiar estos datos.

¿Es el precio del petróleo un parámetro a tener en cuenta?

Al ver la descripción de este reto se nos dice que Ecuador es un país dependiente del petróleo y puede ser vulnerable a los cambios del precio de este. Es por esto que se quiere averiguar si de verdad el precio del petróleo representa un parámetro considerable que podría afectar las ventas.

Al realizar una gráfica de dispersión comparando las ventas por día con el cambio de precio del petróleo por día, se puede ver que estos no tienen una relación tan estrecha como para considerar al precio del petróleo como algo que repercute de manera importante en las ventas, tampoco existe una aparente correlación, por lo que no será tomado en cuenta.



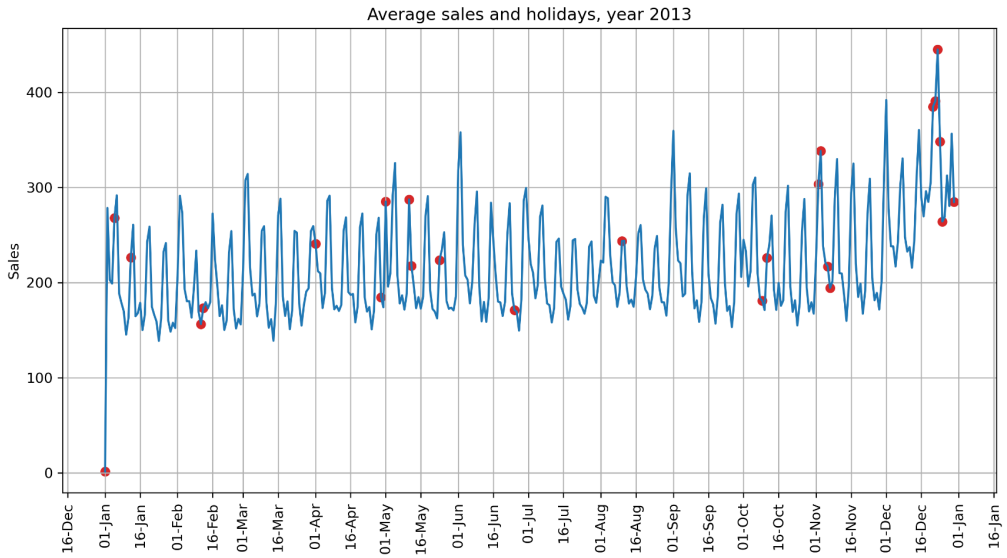
¿Existe alguna relación entre los días festivos y las ventas?

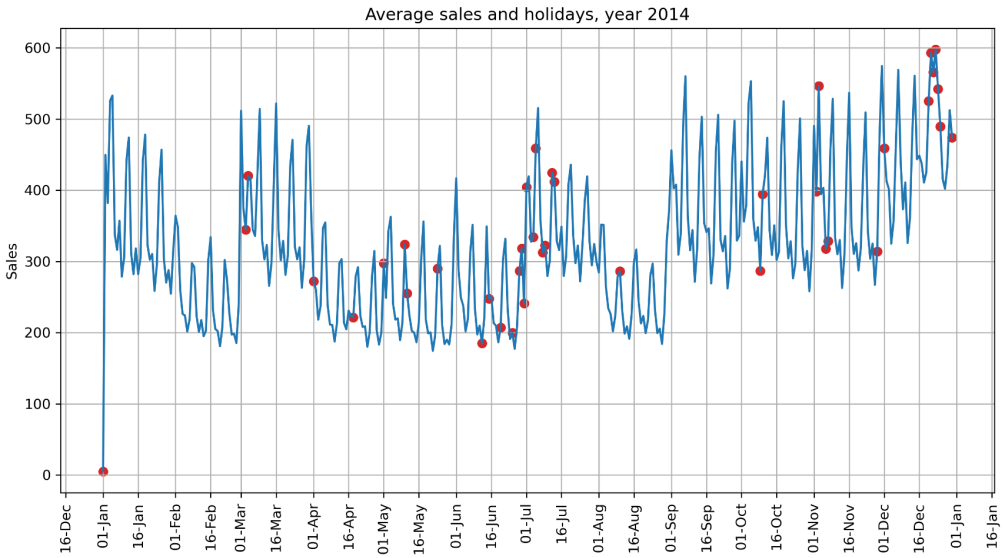
Dentro del Dataset, se incluyen archivos adicionales que pueden contener información suplementaria para el modelo, dentro de estos se encuentra el archivo *holidays_events.csv*, el cual contiene información de eventos festivos y de diversos eventos como lo sería los partidos del mundial de fútbol.

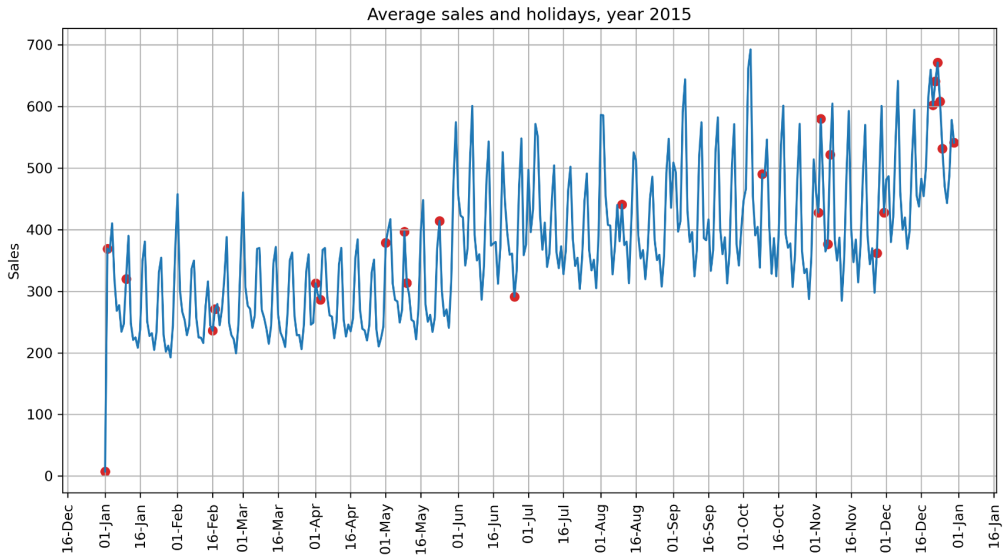
Se buscó analizar si dichos eventos tienen algún efecto en las ventas, para realizar esto primero se condensaron todas las ventas por tienda, familia de producto y fecha, también se removió los datos del terremoto de 2016 de la tabla de holidays, debido a que fue un evento extraordinario.

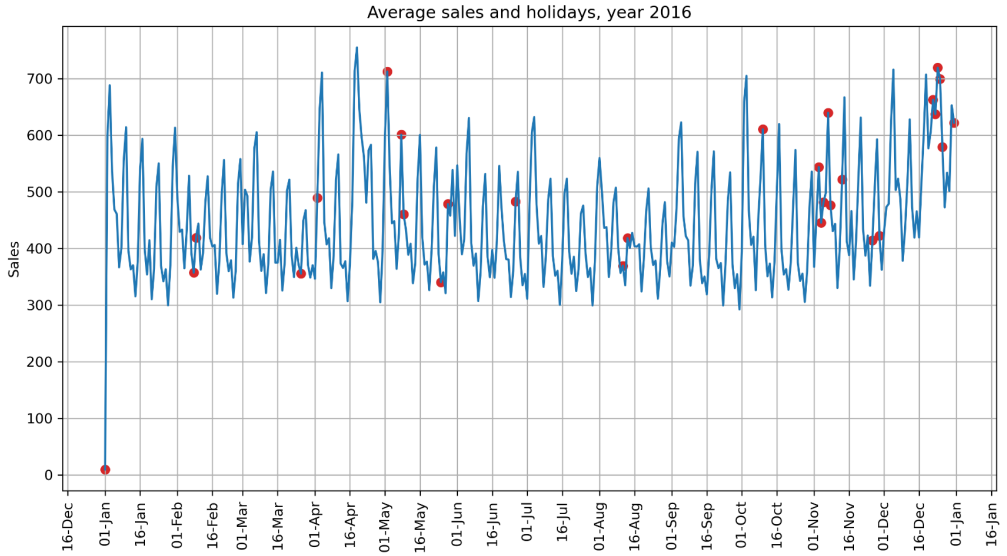
Seguidamente se condensaron las ventas diarias para los años 2013, 2014, 2015 y 2016, debido a que estos son los años de los que tenemos información completa. Para la tabla de holidays nos enfocamos en los eventos que fueron a nivel nacional y regional, esto debido a que consideramos los que pueden tener una mayor efecto en el número total de ventas.

Al graficar los datos, se pueden observar las siguientes situaciones:




Average sales and holidays, year 2014
Sales
0
100
200
300
400
500
600
16-Dec
01-Jan
16-Jan
01-Feb
16-Feb
01-Mar
16-Mar
01-Apr
16-Apr
01-May
16-May
01-Jun
16-Jun
01-Jul
16-Jul
01-Aug
16-Aug
01-Sep
16-Sep
01-Oct
16-Oct
01-Nov
16-Nov
01-Dec
16-Dec
01-Jan
16-Jan





Podemos observar que hay un incremento en el número de ventas diarias alrededor de las fechas en las que sucede un evento, de estos es particularmente interesante que los primeros días del mes de enero se tiene un número de ventas muy bajo, esto se puede asociar al incremento de ventas que tomó lugar en los últimos días del mes de diciembre del año pasado; también está el incremento en ventas alrededor de los días de la madre, el viernes santo, el grito de la independencia, el dia de difuntos y los días de ofertas de Black Friday y Cyber Monday.

Con la ayuda de estas visualizaciones podemos decir que en el periodo de tiempo alrededor de días donde ocurren eventos se pueden observar un aumento en el número de ventas.