Facultad de Ingeniería

9558 - CIENCIA DE DATOS

Primer Cuatrimestre de 2024

MACHINE LEARNING

Marcela Jazmin Cruz
110066

Gualdi Martina
110513

Introducción

En el presente informe se llevará a cabo un análisis exploratorio de los dataset cuya información incluye:

- ID de shops
- ID de ítems
- Categorías
- Precio de ítems
- Fecha
- Cantidad de ítems vendidos en un dia

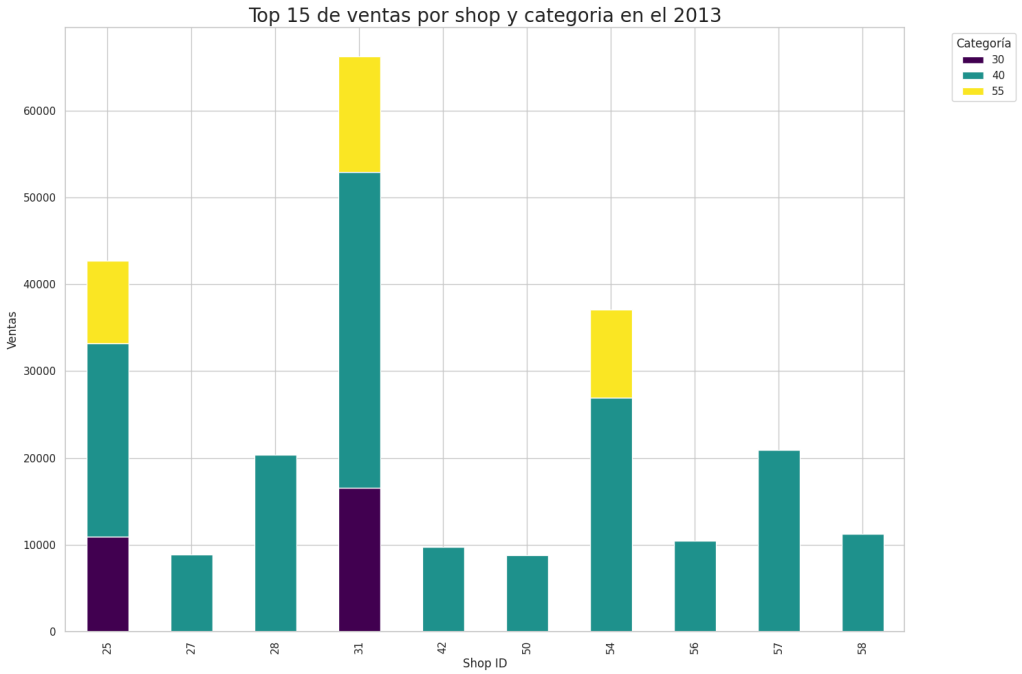
A lo largo de los años

Como primer punto decidimos ver cómo fueron las ventas a lo largo de los 3 años que brinda el dataset: 2013, 2014 y 2015. En el siguiente gráfico podemos observar la tendencia negativa que tiene la cantidad de ventas al año. Mientras que de 2013 a 2014 las ventas totales cayeron un 15%, entre 2014 y 2015 las ventas cayeron en un 42%.

Para interiorizarnos en cada uno de los años decidimos analizar el top 15 de ventas por shop y categoría.

Año 2013

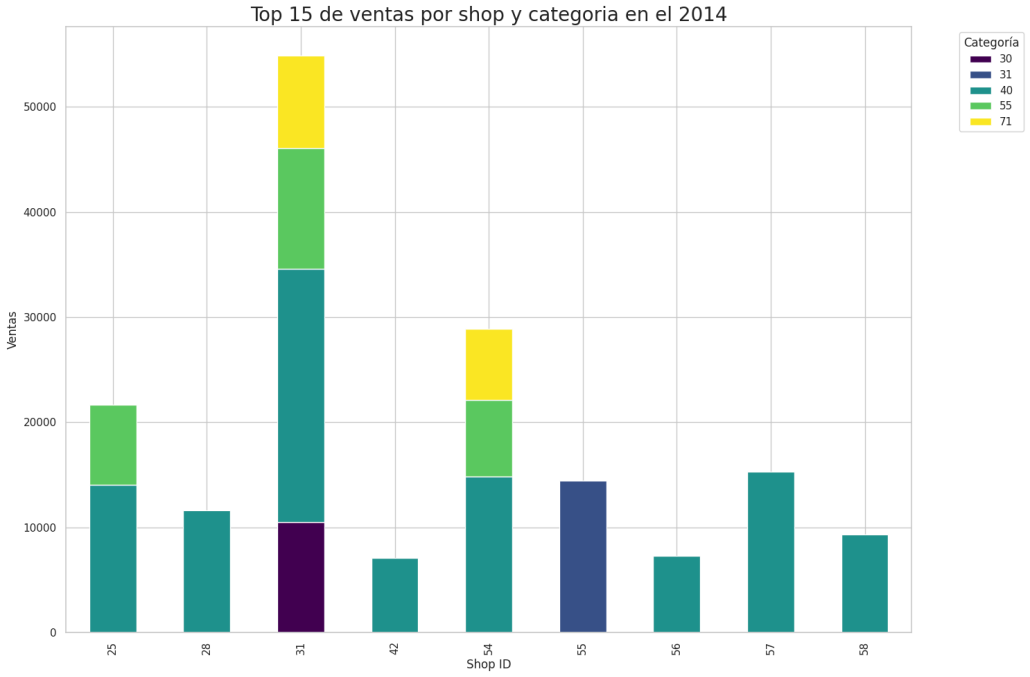
En el año 2013, el año con mayor cantidad de ventas, obtenemos el siguiente resultado.



Podemos apreciar que los shops con mayor cantidad de ventas en ese año fueron: **25, 27, 28, 31, 42, 50, 54, 56, 57, 58.** Además se observa en el gráfico podemos decir que la categoría más vendida fue la **40**, la misma corresponde a **Кино - DVD**.

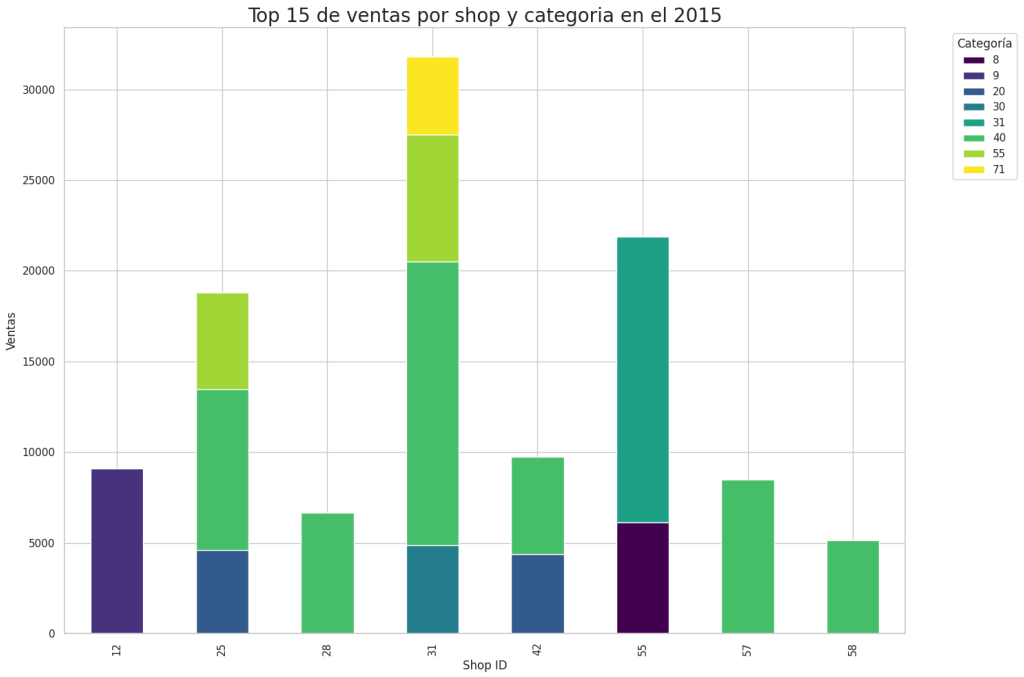
Año 2014

Por otro lado, en el año 2014 se observa el siguiente gráfico. En este caso los shops con mayores ventas fueron: **25, 28, 31, 42, 54, 55, 56, 57, 58**. Y la categoría más vendida, al igual que el 2015 fue la **40**, correspondiente a **Кино - DVD**.



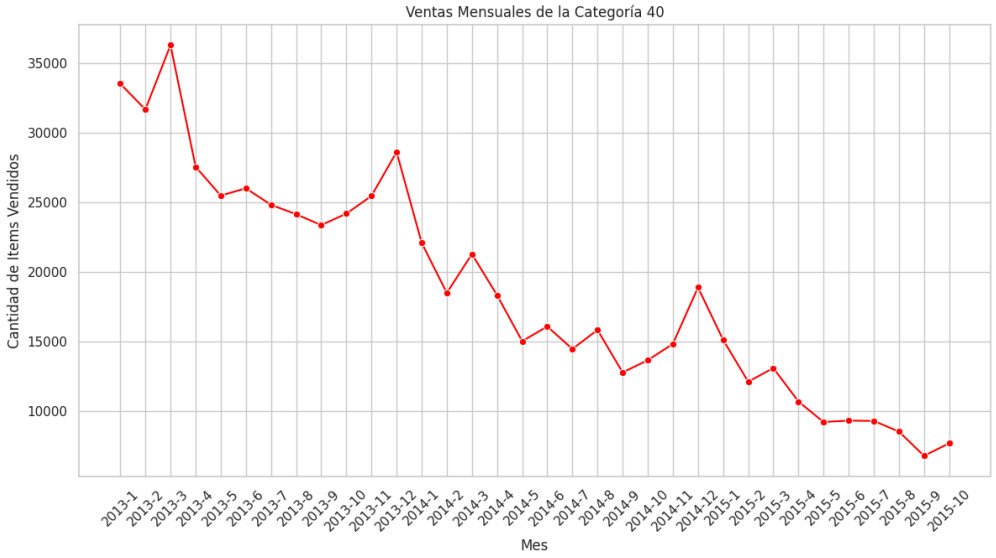
Año 2015

Como se puede apreciar, en este año al igual que los otros, la categoría más vendida fue la **40**, correspondiente a **Кино - DVD**. Y los shops con mayores ventas fueron: **12, 25, 28, 31, 42, 55, 57, 58**. De esta manera, ahora nos interesa saber cómo se fue comportando la categoría más vendida frente a diferentes parámetros.



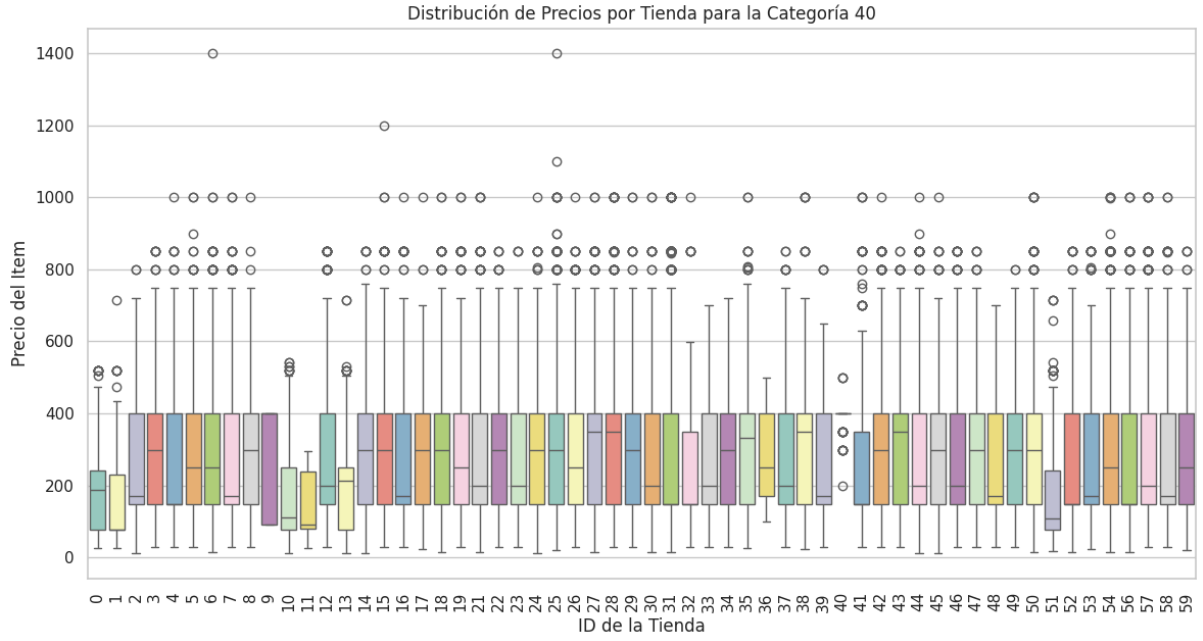
La categoría más vendida

Como se mencionó anteriormente, la categoría que más ventas obtuvo fue en todos los casos la **40**, correspondiente a **Кино - DVD**. Por este motivo consideramos interesante inspeccionar la misma un poco más allá. Una de las preguntas que nos surgen es si las ventas de esta categoría disminuyeron al igual que las ventas totales o si se mantuvieron un poco más estables.



En el gráfico que se muestra podemos confirmar que a medida que fue pasando el tiempo las ventas decayeron más allá de ser la categoría más vendida en esos 3 años.

Otro punto de cuestionamiento que encontramos fue si esta categoría se vendía similarmente en todas las tiendas o había algunas específicas en las que se vendía más o menos, estos fueron los resultados obtenidos de dicho análisis:

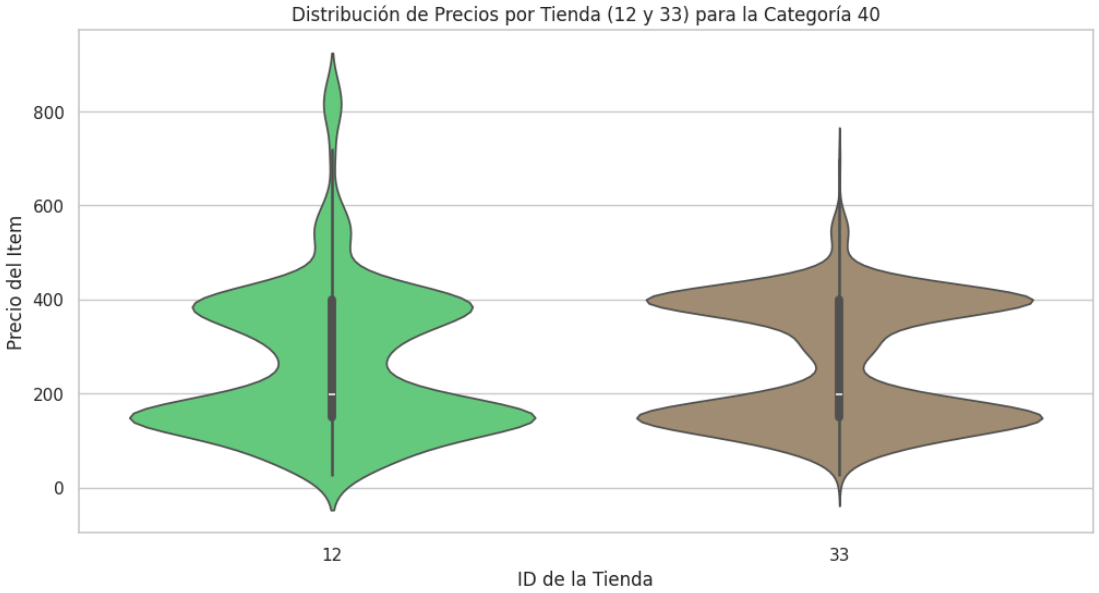


El boxplot de precios por tienda para la categoría 40 revela variaciones significativas en los precios de los ítems vendidos entre diferentes tiendas. Algunas tiendas muestran una mayor dispersión de precios (como por ejemplo la 12 o la 33), lo que indica que venden ítems con una amplia gama de los mismos, mientras que otras tiendas tienen precios más uniformes y concentrados (como por ejemplo la 0 y la 10).

Tiendas con outliers extremos (como por ejemplo la 25 y la 15) sugieren la presencia de productos significativamente más caros o más baratos en comparación con el resto de su inventario. Además, la mediana de los precios varía entre tiendas, lo que puede reflejar diferencias en estrategias de precios, tipos de productos vendidos, o la demanda del mercado local.

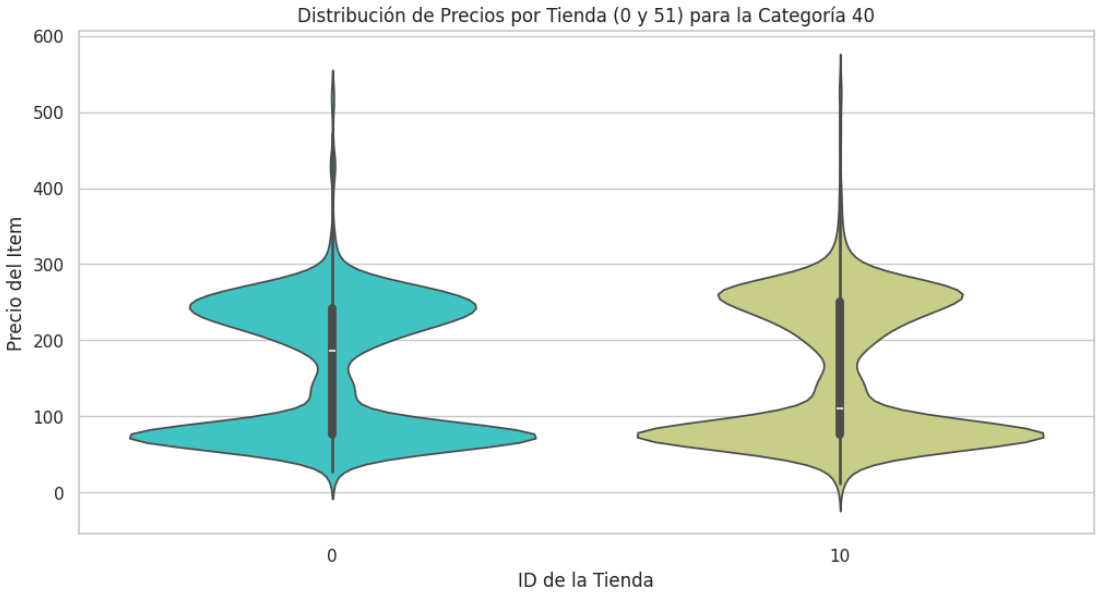
Por esta razón decidimos ahora llevar a cabo un análisis un poco más puntual sobre estas cuatro tiendas mencionadas.

Tiendas con gran dispersión de precios



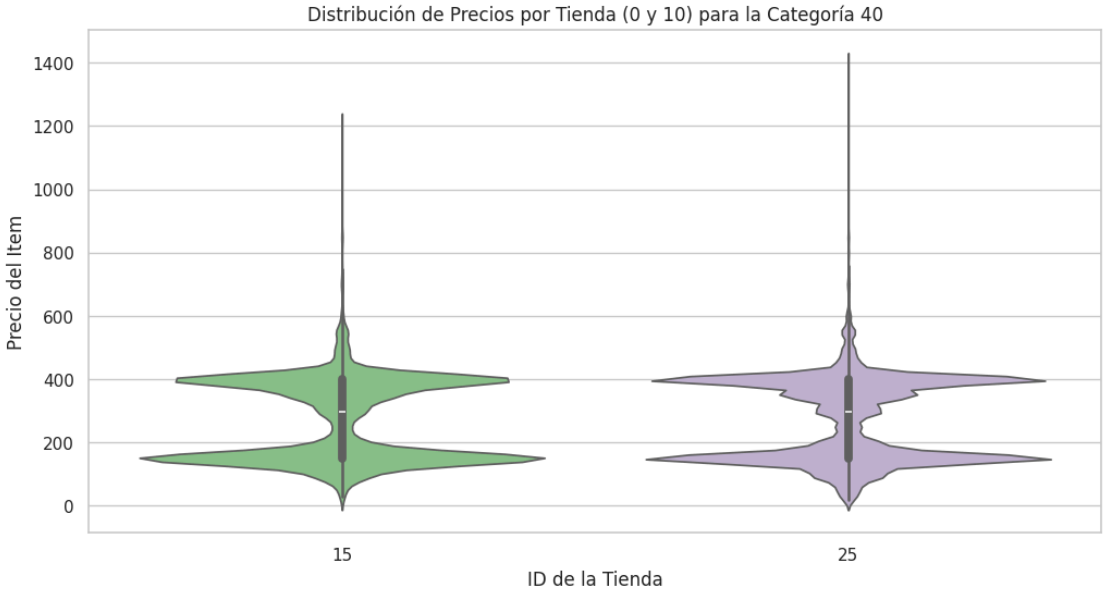
Las tiendas que muestran una gran dispersión de precios indican que tienen una amplia variedad de productos con diferentes rangos de precios. La diversidad en los precios puede ser una estrategia y también puede reflejar una falta de enfoque en un segmento específico del mercado.

Tiendas con precios más concentrados



Las tiendas con precios más concentrados tienen una menor variabilidad en sus precios, lo que sugiere que venden productos que son más uniformes en cuanto a costo. Estas tiendas podrían estar centradas en un tipo particular de producto o en una estrategia de precios

Tiendas con outliers extremos

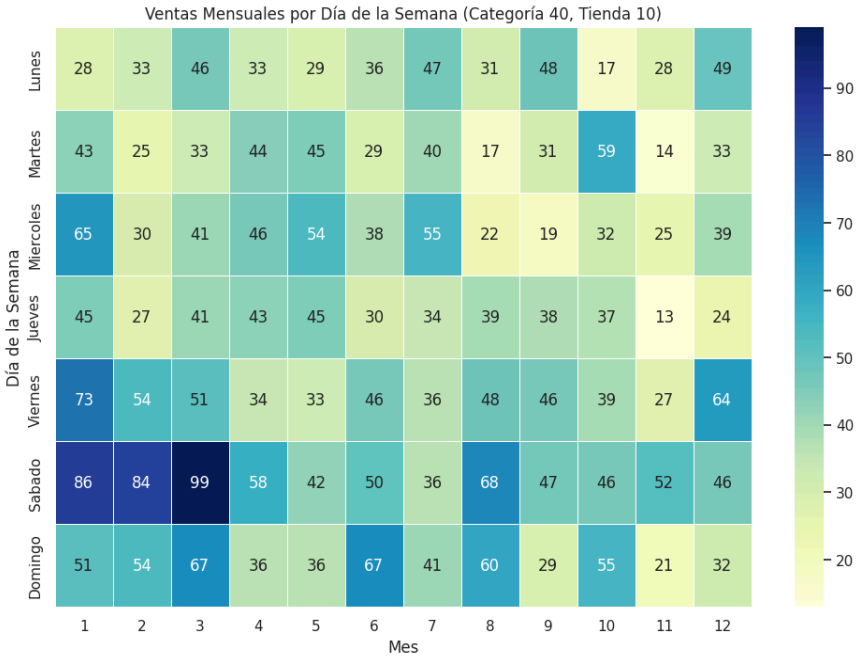


Las tiendas con outliers extremos tienen precios que se desvían significativamente del rango típico, tanto hacia precios muy altos como muy bajos. Esto podría reflejar casos como por ejemplo: promociones especiales, productos de lujo, o liquidaciones de stock.

Influencia de los días de la semana

Continuando con el análisis de la categoría más vendida, tomando la misma como ejemplo, nuestro nuevo objetivo es analizar cómo se comportan las ventas dependiendo el día de la semana. Esto lo llevaremos a cabo analizando una de las tiendas de cada una de las tres secciones anteriormente desarrolladas.

Tienda 10



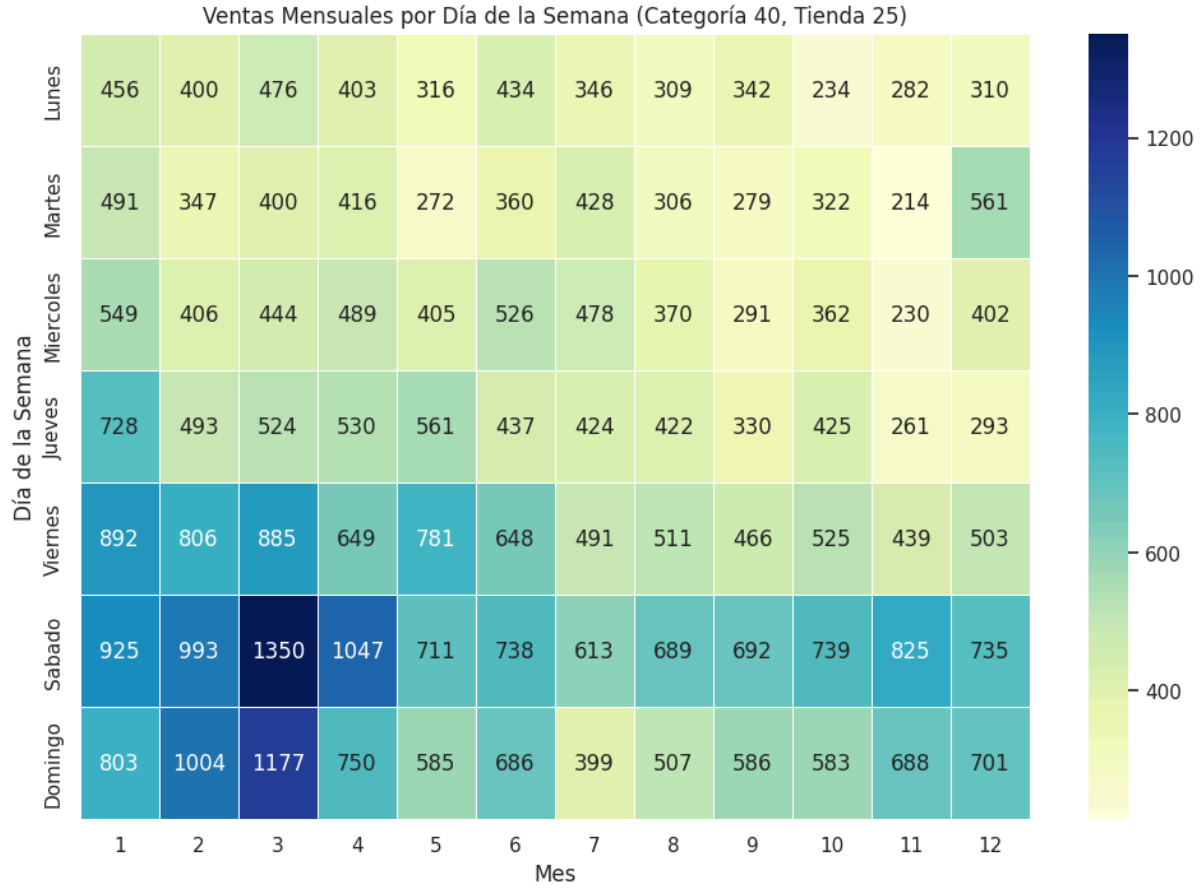
Como se puede apreciar en la imagen el gráfico representa para cada mes cuántas ventas hubieron cada día de la semana, en dicha tienda y categoría. Con este grafico podemos hacer algunas observaciones:

- Notablemente los días en los que más compras se realizan son los sábados y domingos y en algunos casos los viernes. En la mayoría de los casos no se encuentran grandes valores de compras entre semana. Un motivo por el cual puede suceder esto es que los fines de semana las personas pueden encontrar más momentos libres para realizar ciertas actividades como hacer compras.

- En el mes de diciembre los valores son bastante parejos todos los días. Esto puede deberse a distintos motivos, uno de ellos que al ser época festiva las personas hacen compras sin importar el día en el que lo hace.

- En el caso de la tienda 10, en el violín plot pudimos notar que es una de las tiendas con precios uniformes, por esta razón puede ser que las ventas a lo largo de los meses no sean tan variadas, es decir, que no hay valores tan extremos en ningún caso. De esta manera podemos decir que los precios podrían llegar a influir en las ventas de la tienda.

Tienda 25

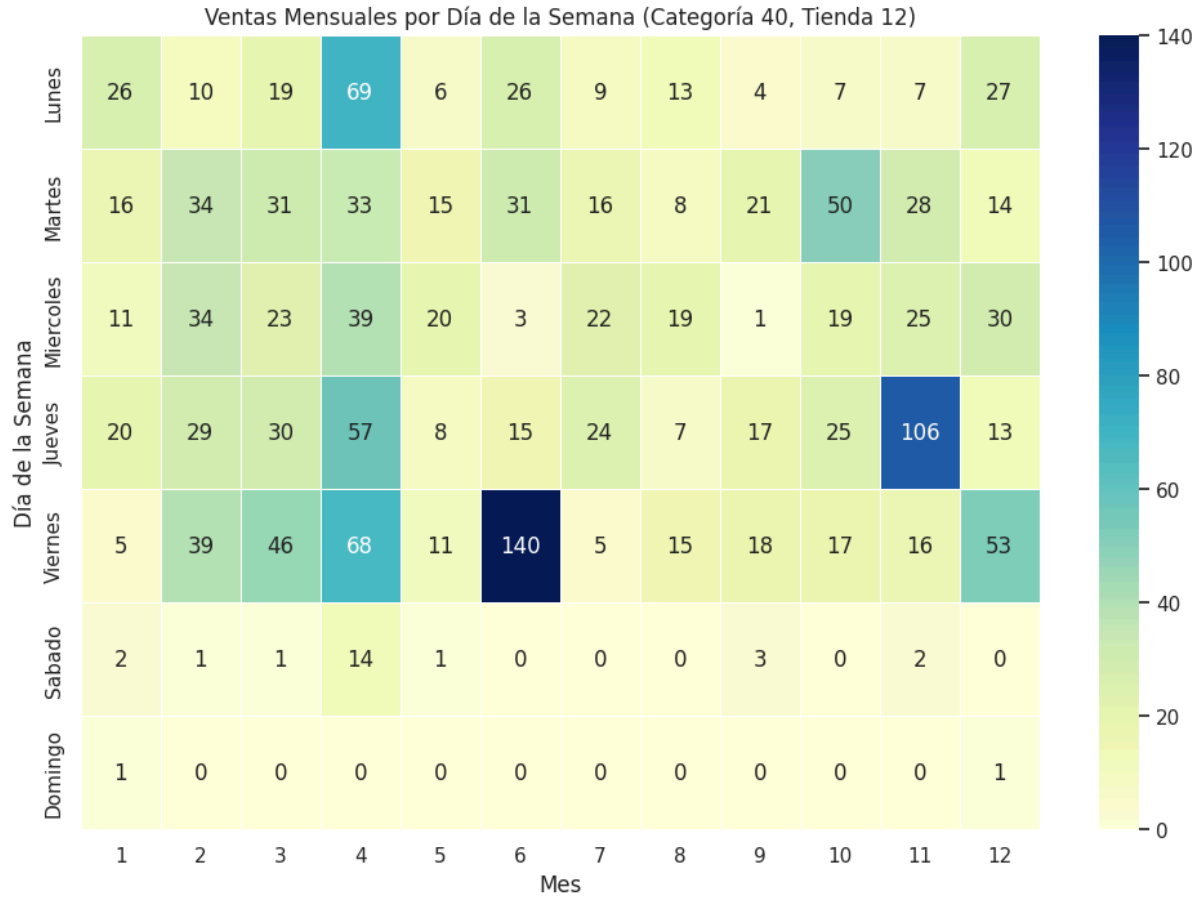


En este caso podemos notar una amplia diferencia con el gráfico de la tienda 10, los números de ventas. Las ventas totales por mes son notablemente mayores en comparación a la tienda anterior. Por ejemplo, en el mes de Diciembre podemos notar que en este caso hay un total de 3505 ventas mientras que en la tienda 10 hay un total de 287 ventas. Esto puede deberse, por ejemplo, al tipo de producto que se venda en cada tienda.

Más allá de esta diferencia que mencionamos anteriormente los puntos principales mencionados en el apartado anterior se respetan de igual manera solo a pesar de que los valores son mayores en este caso.

La tienda 25 pertenece a las tiendas que tienen un outlier muy grande, de esta manera podemos explicar el porque hay meses en los que las ventas totales son 3631 y en otros fueron 5256. La gran variedad de precios podría derivar en una amplia diferencia de ventas totales en los diferentes meses.

Tienda 12



En este caso podemos notar algo completamente diferente a los casos anteriores. Los valores son muy bajos y en muchos de los casos tienen 0 ventas. Los valores, nuevamente, pueden deberse al tipo de producto, al igual que las ventas en 0. Otra de los motivos que podemos tomar en cuenta es la ubicación de la tienda, si se encuentra en una zona menos poblada/transitada puede que las ventas sean menores a las de una tienda ubicada en una zona céntrica.

En el caso de la tienda 12 los valores más altos se alcanzan en el día viernes del mes 6 y el día jueves del mes 11, esto quiere decir que en esta tienda no se respetan los parámetros principales utilizados en las dos tiendas anteriormente detalladas.

La tienda 12 pertenece al grupo de las tiendas que tienen precios más dispersos, sin embargo en este caso no logramos obtener información muy precisa ya que los valores de las ventas son muy bajos e incluso 0.

Conclusiones del análisis

A modo de conclusión nos gustaría remarcar los aspectos que pudimos distinguir para poder predecir las futuras ventas, o mejor dicho, qué factores pueden hacer que las ventas próximas sean mayores o peores.

1. **Categoría**

Como bien se mencionó anteriormente, en los tres años que involucra el dataset encontramos como categoría más vendida **40**, correspondiente a **Кино - DVD**. Guiándonos por el nombre de la categoría podemos deducir que los productos dentro del área de informática, electrónica o similares tienen más demanda que otros tipos de productos.

2. **El paso del tiempo**

En el primer apartado se mostró como el paso del tiempo tendía a disminuir las ventas totales incluso en la categoría más vendida. De esta manera lo que podemos intuir o predecir para futuras ventas es que cuanto más pase el tiempo (a largo plazo) las ventas totales van a bajar, puede ocurrir que ocasionalmente aumenten de un mes a otro como se mostró en el gráfico “ventas mensuales de la categoría 40” sin embargo la tendencia es igualmente negativa.

3. **Tienda**

Al momento de inspeccionar las tiendas y sus ventas pudimos notar que sus valores son muy variantes. Por esta razón antes de concluir vamos a mencionar a qué ciudad corresponde cada una de las tiendas de las cuales se realizaron los tres heatmap anteriores.

La tienda 10 **(Жуковский ул. Чкалова 39м?)** está ubicada en Zhukovskiy, Moskovskaya oblast', Rusia.

La tienda 25 **(Москва ТРК "Атриум")** está ubicada en Zemlyanoy Val St, 33, Moscow.

En el caso de la tienda 12 **(Интернет-магазин ЧС)** no se ha logrado encontrar información sobre la ubicación de dicha tienda.

Esto nos logra confirmar una de nuestras hipótesis, la ubicación importa. La tienda con menos ventas fue la 12 la cual se desconoce la ubicación, mientras que la 25, la de mayor ventas, está ubicada en la capital del país por lo cual es un lugar muy transitado tanto por ciudadanos como por turistas. Por otro lado la tienda 10 que se encuentra en un valor intermedio

entre las otras dos en cuanto a cantidad de ventas, está ubicada en una ciudad pero probablemente menos transitada que Moscow. Entonces, a modo de conclusión, podemos predecir que a mejor ubicación de tienda mayores serán las ventas de la misma.

Modelos

Baseline

En primer lugar, consideramos primordial hacer un ordenamiento de datos. Luego, para definir las variables del modelo, se seleccionan las características (features) y el objetivo (target). Las características incluyen identificadores de artículos, categorías, tiendas, y variables temporales como el año, mes, día y día de la semana, además de estadísticas agregadas como la media de ventas por categoría y tienda, y la media de ventas por tienda. Estas características se almacenan en el df `X`. El objetivo, que es el número de ventas diarias del artículo, se almacena en la Serie `y`. El siguiente paso es dividir los datos en conjuntos de entrenamiento y validación. Se asigna el 80% de los datos al conjunto de entrenamiento y el 20% restante al conjunto de validación.

Después, se crea una instancia del modelo de regresión lineal y se ajusta el modelo a los datos de entrenamiento. Una vez entrenado, se pueden obtener los coeficientes y el intercepto del modelo.

Para evaluar el modelo, se realizan predicciones sobre el conjunto de validación, almacenando los resultados en `y_pred`. El rendimiento del modelo se mide mediante el coeficiente de determinación, el error cuadrático medio y el error absoluto medio.

Random forest

Se procede a la limpieza de datos eliminando aquellas filas que contienen cantidades de ventas diarias negativas y precios negativos, así como registros con valores nulos en la columna date_block_num.

Para mejorar la capacidad predictiva del modelo, se crean dos nuevas características: mean_category_shop, que representa la media de ventas diarias de una categoría de productos en una tienda específica, y mean_items_shop, que refleja la media de ventas diarias de un producto específico en una tienda. Estas agregaciones pueden capturar tendencias y comportamientos de venta a nivel de tienda y categoría.

Posteriormente, se definen las variables independientes (X3) y la variable dependiente (y3) para el modelo. Las variables independientes incluyen el identificador del producto, el identificador de la tienda, el mes y el día de la semana, mientras que la variable dependiente es la cantidad de ventas diarias. Los datos se dividen en conjuntos de entrenamiento y validación usando la función train_test_split, con un 20% de los datos reservados para la validación.

Antes de entrenar el modelo, se realiza una estandarización de las características utilizando Standard Scaler de Scikit-learn.

Se entrena un modelo de RandomForestRegressor con hiperparámetros específicos que controlan la profundidad máxima de los árboles, el número mínimo de muestras

requeridas para dividir un nodo y el número mínimo de muestras en una hoja. Estos parámetros se seleccionan para equilibrar la complejidad del modelo y su capacidad de generalización.

Finalmente, se realizan predicciones sobre el conjunto de validación y se calcula la raíz del error cuadrático medio (RMSE) para evaluar el rendimiento del modelo.

XG Boost

Se implementa un modelo de machine learning utilizando XG Boost para predecir ventas. Inicialmente se realiza una exploración y limpieza de datos, la cual incluye la verificación de valores únicos en diferentes columnas, el manejo de valores faltantes y la combinación de diferentes dfs para crear un conjunto de datos completo y coherente. A continuación se crean nuevas características a partir de los datos existentes, como agregaciones mensuales de ventas y la transformación de variables categóricas en variables numéricas.

El modelo XG Boost se utiliza para la predicción debido a su eficiencia y rendimiento en tareas de predicción con grandes conjuntos de datos. Se dividen los datos en conjuntos de entrenamiento y prueba, y se configura el modelo con parámetros específicos para optimizar su rendimiento. Luego, el modelo se entrena utilizando los datos de entrenamiento.

Finalmente, se evalúa el rendimiento del modelo utilizando métricas como el RMSE (Root Mean Squared Error) en los datos de prueba para determinar su precisión.