Project Management Sales Prediction

En este proyecto analizamos los datos de ventas de una empresa para identificar patrones y hacer recomendaciones para mejorar las ventas.

Objetivo

Ayudar al distribuidor a comprender las propiedades de los productos y los puntos de venta que influyen en el modelo de predicción.

Explorar los datos

Preparar los datos para el modelado, incluyendo la selección de características relevantes y la normalización de los datos..

Seleccionar un modelo

Una vez que se comprenden los objetivos del negocio y se han explorado los datos, se puede seleccionar un modelo apropiado .

Entrenar y ajustar el modelo

Luego se debe entrenar el modelo con los datos y ajustarlo para obtener el mejor rendimiento.

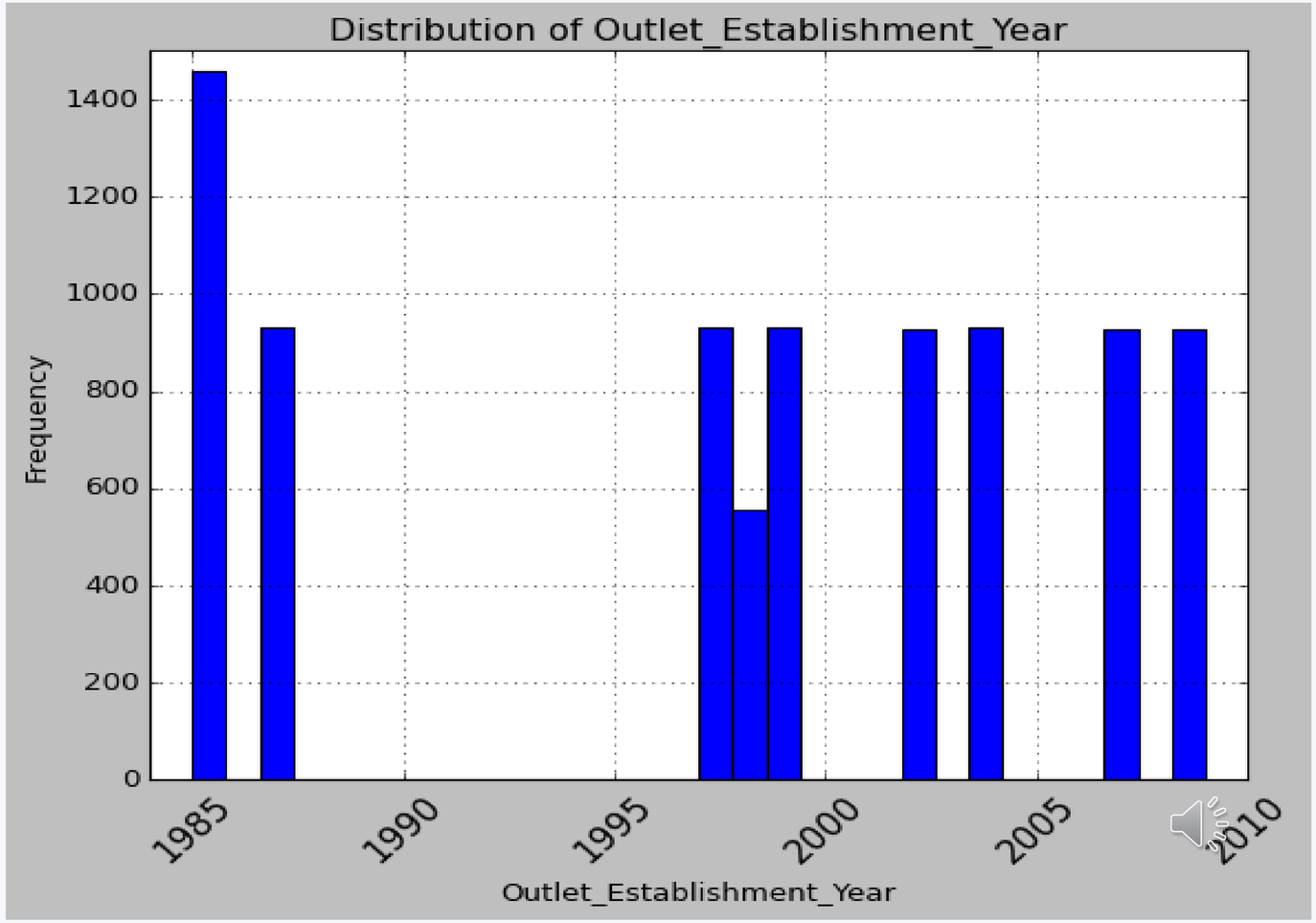
Evaluar el modelo y Comunicar los resultados

Es importante evaluar el modelo y se debe comunicar los resultados al cliente de una manera que sea fácil de entender.

Explorar los datos

Write here your awesome subtitle

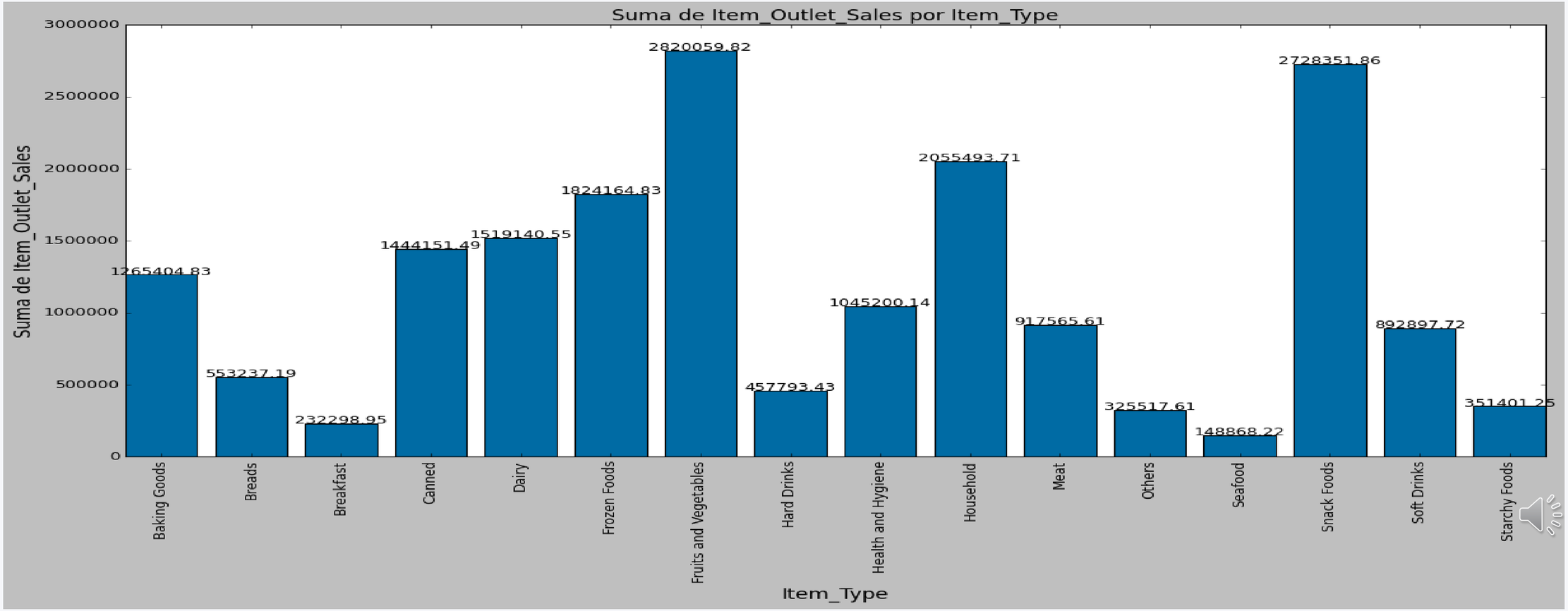
En el análisis del Histograma podemos ver las distribuciones de las ventas por el año de inauguración de las tiendas y el conjunto de datos reveló que la tienda más antigua vende más que todas las otras.



Explorar los datos

Write here your awesome subtitle

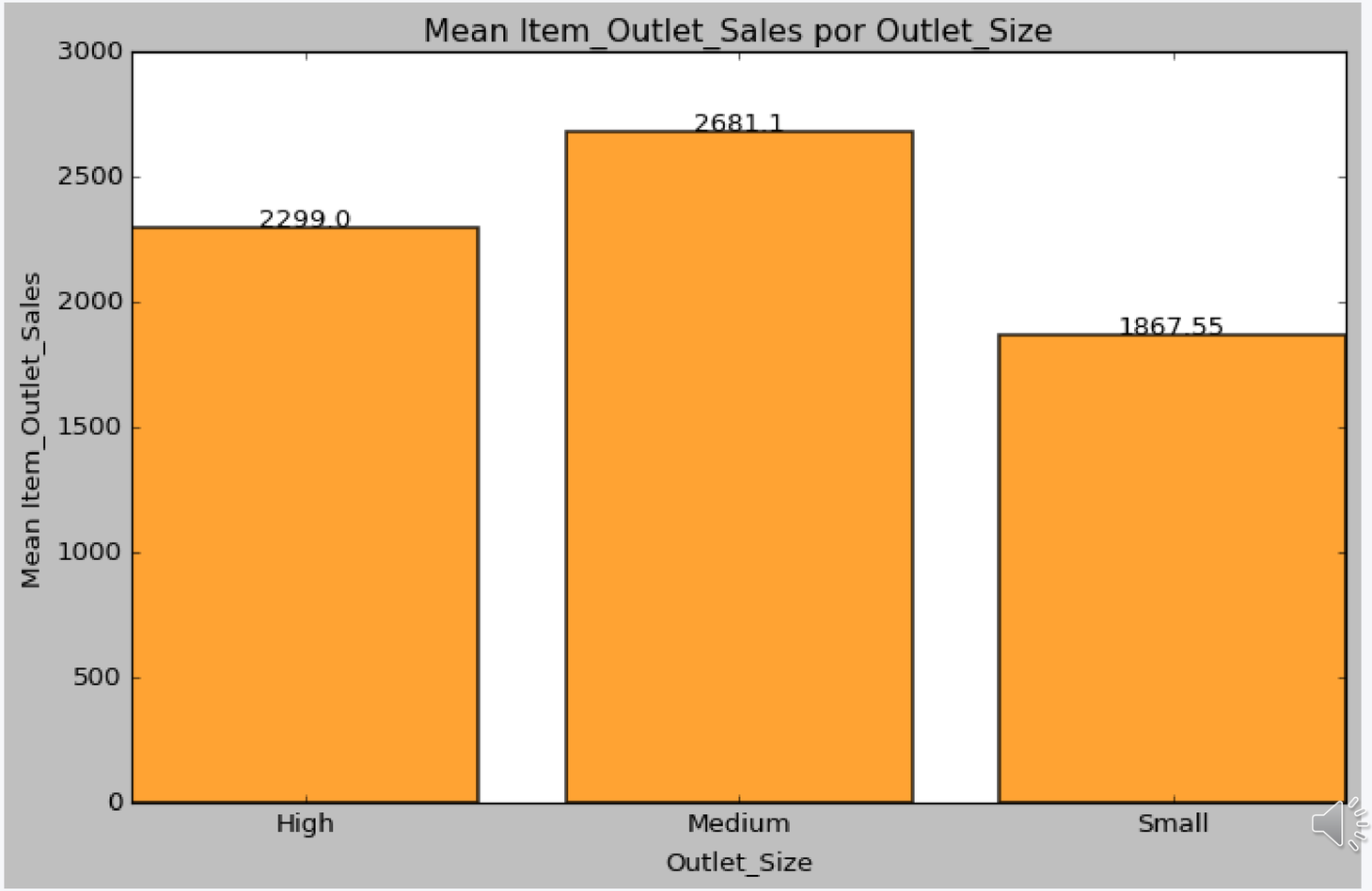
En el análisis del grafico de barras podemos ver las distribuciones de las ventas por el tipo de articulo y el grafico verifica que Frutas/Vegetales y Snack_Food son los tipos que más vende.



Explorar los datos

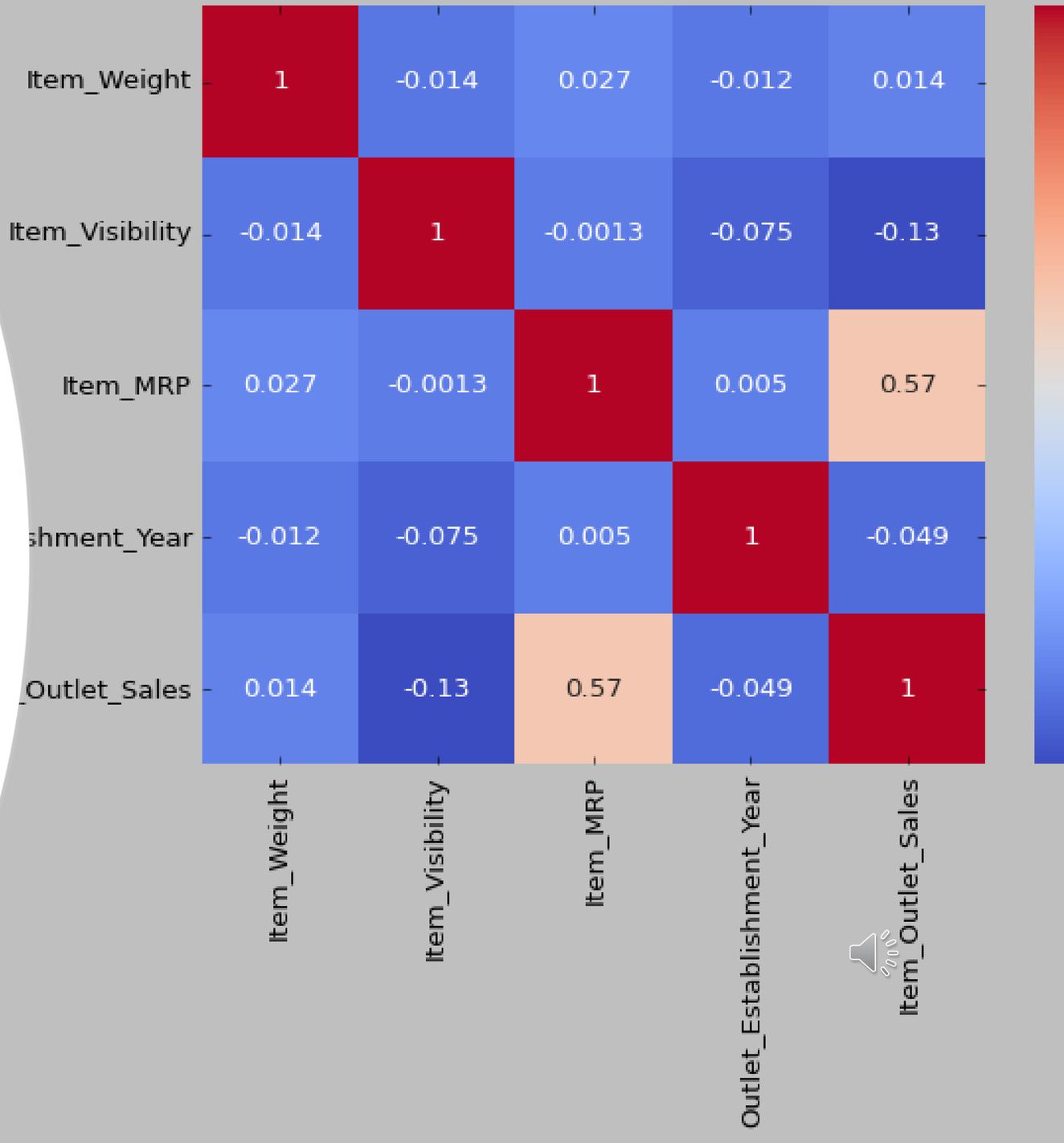
Write here your awesome subtitle

En el análisis del grafico de barras podemos ver el promedio de las ventas por el tamaño de la tienda y el grafico verifica que las tiendas de tamaño mediano tiene mejor promedio de ventas.



Explorar los datos

- El conjunto de datos contiene información sobre el peso de los productos, su contenido de grasa, el tipo de producto y la ubicación de la tienda, entre otros.
- Hay 8523 entradas en el conjunto de datos, con algunas columnas que tienen valores faltantes. En el pré-procesamiento verifiqué la columna "Item_Identifier" no tiene una relación significativa con la variable objetivo, decidí eliminarla por no afectar la capacidad del modelo para predecir la variable objetivo. también convertí la columna 'Outlet_Establishment_Year' en objeto por que consideré los valores como categóricos, pese que son años, en la columna 'Item_Fat_Content' sustituí valores únicos que representan la misma categoría por un solo nombre(Low Fat, estaba escrito de 3 formas distintas) Identifiqué el objetivo (X) y las características (y): Asigné la columna "Item_Outlet_Sales" como el objetivo y el resto de las variables relevantes como el matriz de características, Realice un train test split, Cree un pipeline de preprocesamiento para preparar el conjunto de datos para el aprendizaje automático.
- La matriz de correlación verifica una baja correlación entre las columnas con excepción de la columna Item_MRP.



Seleccionar un modelo

Write here your awesome subtitle

1 - Regresión lineal: Calculamos el R^2 del modelos, que indica cuánto de la variación en las ventas se puede explicar por las variables que usamos en el modelo. También calculamos el Raíz del error cuadrático medio (RECM), que indica cuánto se desvía nuestra predicción de las ventas reales. El modelo de regresión lineal obtuvo un R^2 de 0.56 y el RECM para el modelo de regresión lineal fue de 1139.10

```
[ ] #A continuación, se muestra el código para obtener los datos de R^2 después de ajustar nuestro modelo:
    train_score = reg.score(X_train_df, y_train)
    print(train_score)
    test_score = reg.score(X_test_df, y_test)
    print(test_score)
```

```
    0.5615547614077183
    0.5671059423458024
```

```
[ ] #Raíz del error cuadrático medio (RECM): raíz cuadrada de la media de los errores al cuadrado.
    rmse_train = np.sqrt(mean_squared_error(y_train, train_preds))
    rmse_test = np.sqrt(mean_squared_error(y_test, test_preds))
    print(rmse_train)
    print(rmse_test)
```

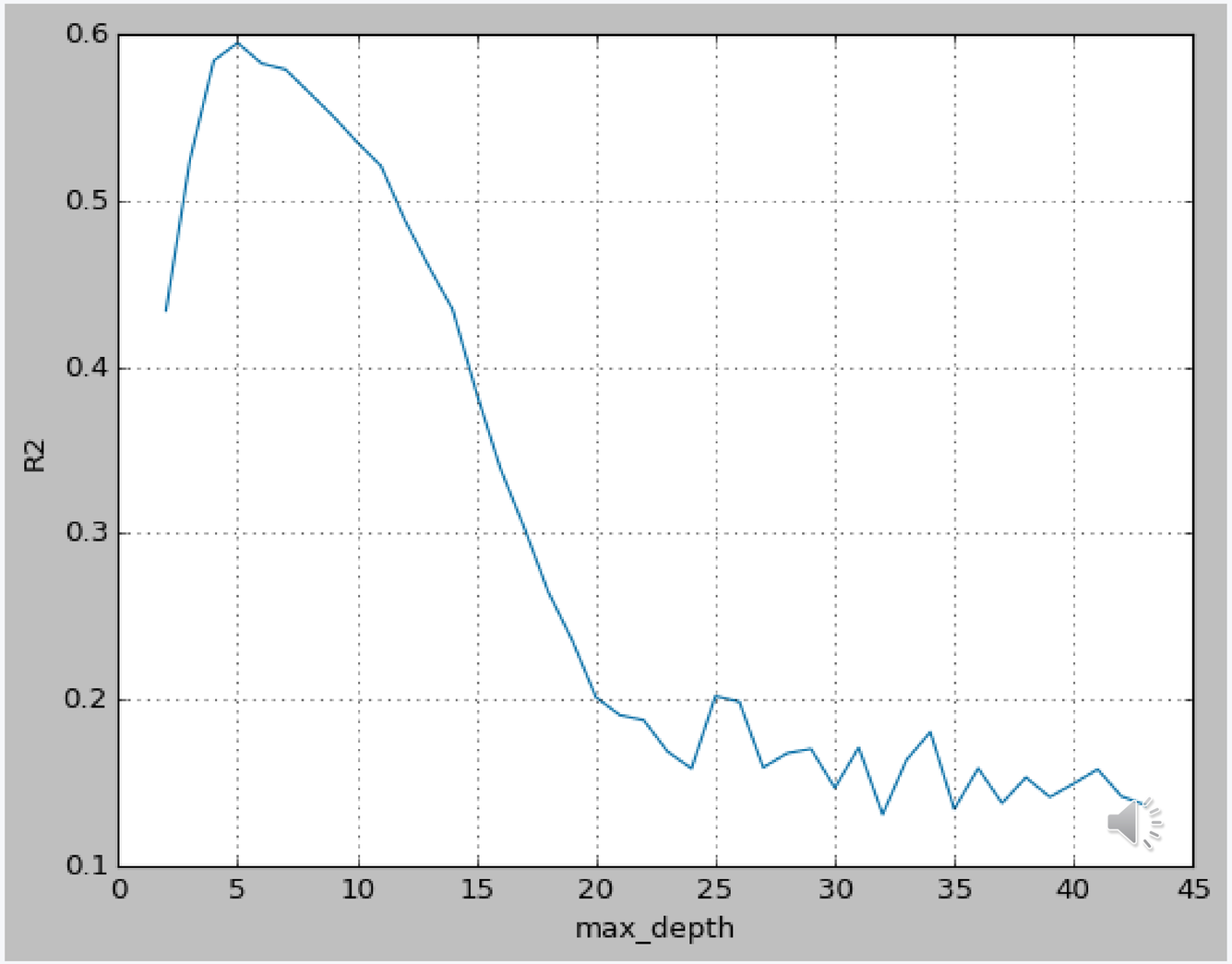
```
    1139.1045880315098
    1092.8608663020173
```

Seleccionar un modelo

Write here your awesome subtitle

2 - Árbol de regresión: Primero hice una lista de valores para probar la profundidad del árbol de regresión según el grafico abajo pudimos definir la profundidad de 5 que es el punto con mayor R^2.

Pero el modelo de árbol de regresión obtuvo el mismo resultado que la regresión lineal para R^2 de 0.56 . el RECM para el modelo de árbol de regresión fue de 0.24 en las datos de entrenamiento y 1543.38 en los datos de test



Evaluar el modelo y Comunicar los resultados

Recomendaciones:

Podemos ver que los productos más comunes son frutas y verduras, seguidos de Snacks.

Esto puede ser útil para la planificación de inventario y marketing.

Podemos ver que hay una relación positiva entre el tamaño y la antigüedad de las tiendas con las ventas.

Esto sugiere que necesita mejorar su estrategia de puntos de ventas.

Resumen:

Analizamos los datos de ventas de un supermercado y usamos dos modelos diferentes para predecir las ventas.

En el caso del modelo de regresión lineal, el RECM en el conjunto de prueba es de 1139.1045880315098, mientras que en el caso del modelo de árbol de regresión el RECM es de 0.24. A partir de estos resultados, podemos concluir que el modelo de árbol de regresión es mejor que el modelo de regresión lineal, ya que tiene un RECM menor en el conjunto de prueba y R^2 fue de 0.56 para los dos modelos . La diferencia entre los resultados de los modelos de entrenamiento y prueba puede ser causada por la complejidad del modelo, la cantidad de datos de entrenamiento, la calidad de los datos, entre otros factores.