



# SALES PREDICTION

**CHILE**  
SEP 2022

# CONTEXTO

El primer proyecto será una predicción de ventas para productos alimenticios que se venden en distintas tiendas.

## OBJETIVO

El objetivo es ayudar a la tienda a comprender las propiedades de los productos y establecimientos que tienen un rol crucial en aumentar las ventas.



# VARIABLES DEL DATAFRAME

Nombre de la variable	Descripción
Item_Identifier	Número de identificación único del producto
Item_Weight	Peso del producto
Item_Fat_Content	Si el producto es bajo en grasas o regular
Item_Visibility	El porcentaje de área total de visualización de todos los productos en la tienda asignados a este producto particular
Item_Type	La categoría a la que pertenece el producto
Item_MRP	Precio Máximo Minorista (precio de lista) del producto
Outlet_Identifier	Número de identificación único de la tienda
Outlet_Establishment_Year	El año en el que se estableció la tienda
Outlet_Size	El tamaño de la tienda en cuanto a la superficie total que cubre
Outlet_Location_Type	El tipo de área donde se ubica la tienda
Outlet_Type	Si la tienda es un almacén o algún tipo de supermercado
Item_Outlet_Sales	Las ventas del producto en la tienda particular. Esta es la variable objetivo que se debe predecir.

# ANÁLISIS EXPLORATORIO

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8523 entries, 0 to 8522
Data columns (total 12 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Item_Identifier                       8523 non-null   object
1   Item_Weight                           7060 non-null   float64
2   Item_Fat_Content                      8523 non-null   object
3   Item_Visibility                      8523 non-null   float64
4   Item_Type                            8523 non-null   object
5   Item_MRP                             8523 non-null   float64
6   Outlet_Identifier                    8523 non-null   object
7   Outlet_Establishment_Year            8523 non-null   int64
8   Outlet_Size                          6113 non-null   object
9   Outlet_Location_Type                 8523 non-null   object
10  Outlet_Type                          8523 non-null   object
11  Item_Outlet_Sales                    8523 non-null   float64
dtypes: float64(4), int64(1), object(7)
memory usage: 799.2+ KB
```

Como se puede observar en la tabla, contamos con datos vacíos en las siguientes variables:

- **Item\_Weight:** 1463 datos vacíos
- **Outlet\_Size:** 2410 datos vacíos

**No se encontró ningún inconveniente con el Dtype de las variables a analizar**

**Tampoco se encontraron valores duplicados dentro de la muestra**

# ANÁLISIS EXPLORATORIO

## CORRECCIÓN BASE

1. **Item\_Weight:** Como no hay ninguna forma de obtener los datos vacíos se procedió a completar los valores faltantes con un 0, no se eliminó la columna de primeras para poder ver primero el análisis de correlaciones

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8523 entries, 0 to 8522
Data columns (total 12 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Item_Identifier                       8523 non-null   object
1   Item_Weight                           8523 non-null   float64
2   Item_Fat_Content                      8523 non-null   object
3   Item_Visibility                       8523 non-null   float64
4   Item_Type                             8523 non-null   object
5   Item_MRP                             8523 non-null   float64
6   Outlet_Identifier                     8523 non-null   object
7   Outlet_Establishment_Year            8523 non-null   int64
8   Outlet_Size                           6113 non-null   object
9   Outlet_Location_Type                 8523 non-null   object
10  Outlet_Type                           8523 non-null   object
11  Item_Outlet_Sales                     8523 non-null   float64
dtypes: float64(4), int64(1), object(7)
memory usage: 799.2+ KB
```



# ANÁLISIS EXPLORATORIO

## CORRECCIÓN BASE

**2. Outlet\_Size:** Se analizó si existía alguna correlación entre la variable Outlet\_Size y otras variables de clasificación de tienda, por ejemplo, Outlet\_Type y Outlet\_Location\_Type

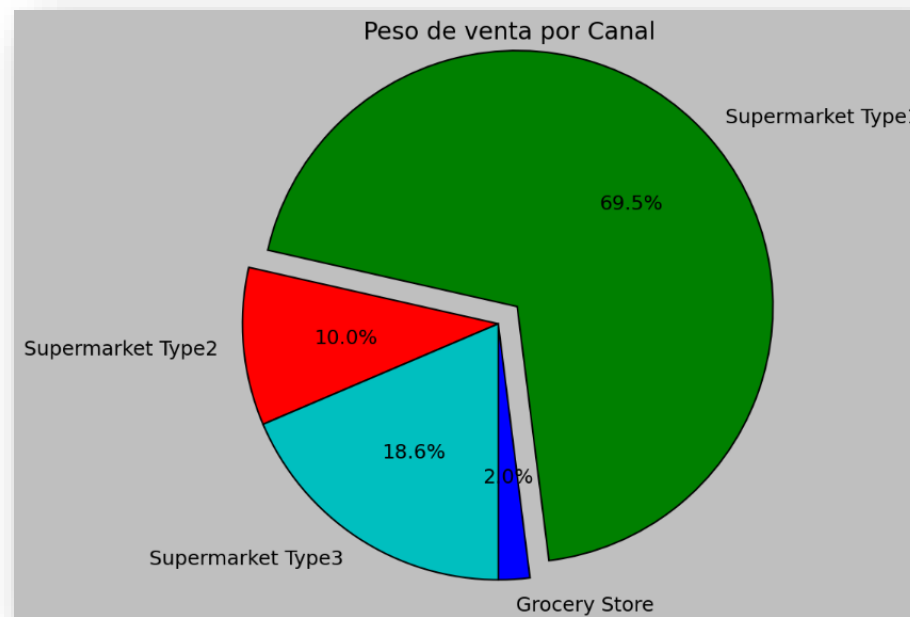
			Item_Outlet_Sales
Outlet_Type	Outlet_Size	Outlet_Location_Type	
Grocery Store	Small	Tier 1	528
Supermarket Type1	High	Tier 3	932
	Medium	Tier 1	930
	Small	Tier 1	930
		Tier 2	930
Supermarket Type2	Medium	Tier 3	928
Supermarket Type3	Medium	Tier 3	935

Dentro de los valores vacíos que se encontraron solo habían Outlet\_Type Supermarket Type 1 (Tier 2) y Grocery Store, por lo que **se procedió a completar los vacíos con Small**

# ANÁLISIS DE GRÁFICOS

## RELEVANCIA CANALES DE VENTA

1. **Supermarket Type 1 es el canal más relevante** : Como se puede observar en el gráfico, tenemos un canal que representa cerca del 70% de las ventas, luego viene Supermarket Type 3, Supermarket Type 2 y, finalmente, Grocery Store.

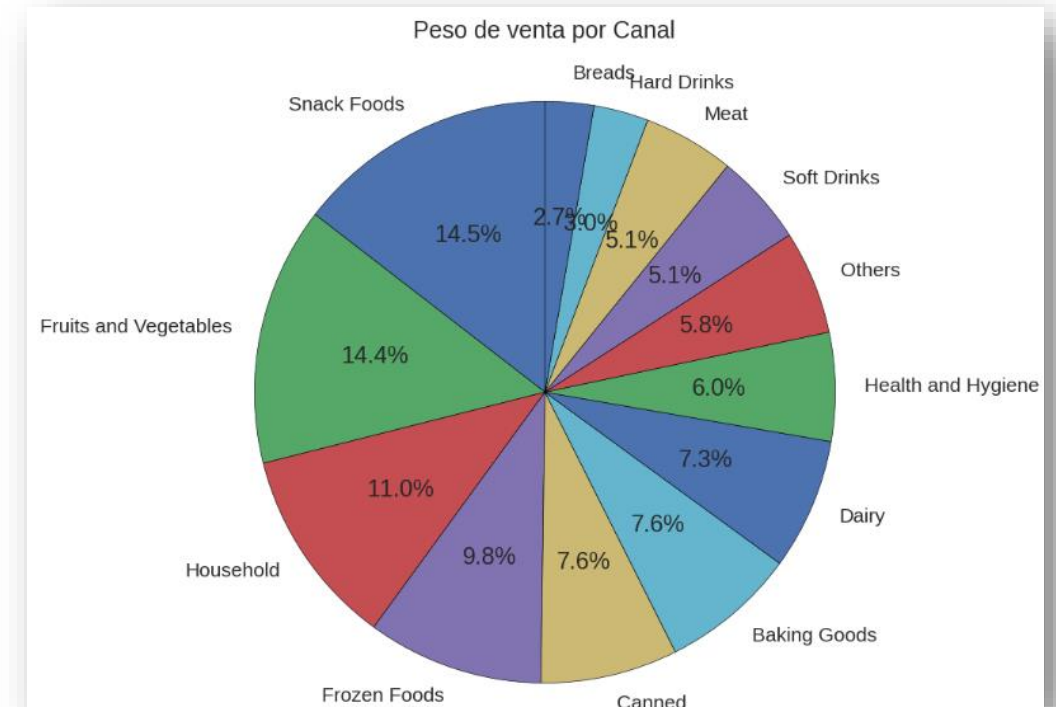
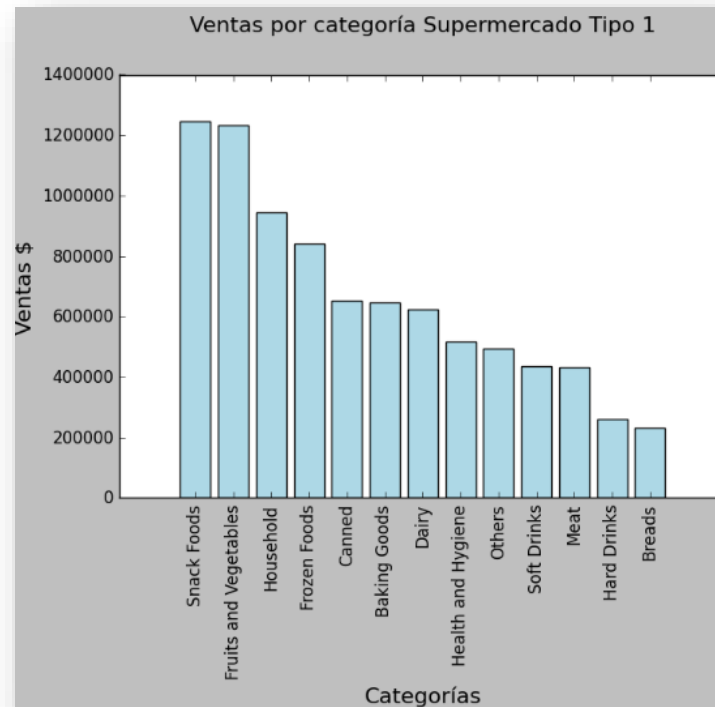


**Como es el canal más relevantes, se procederá solo a analizar este canal dentro de las demás variables**

# ANÁLISIS DE GRÁFICOS

## CATEGORÍAS MÁS RELEVANTES

- Snacks Foods y Fruit & Vegetables se posicionan como las categorías más relevantes del canal:** Solo 4 categorías representan más del 50% de la venta, incluyendo Household y Frozen Foods



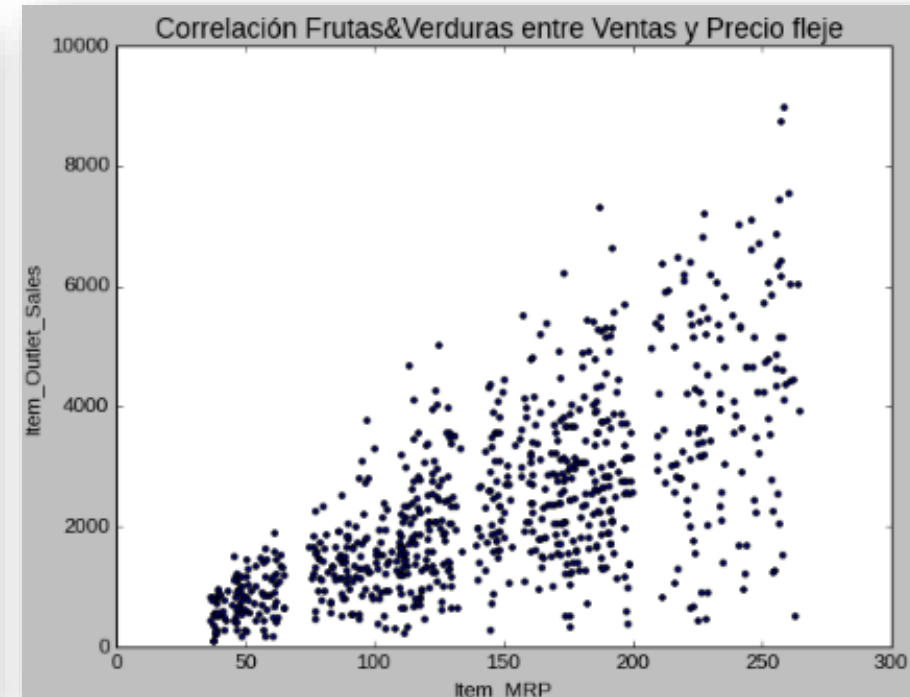
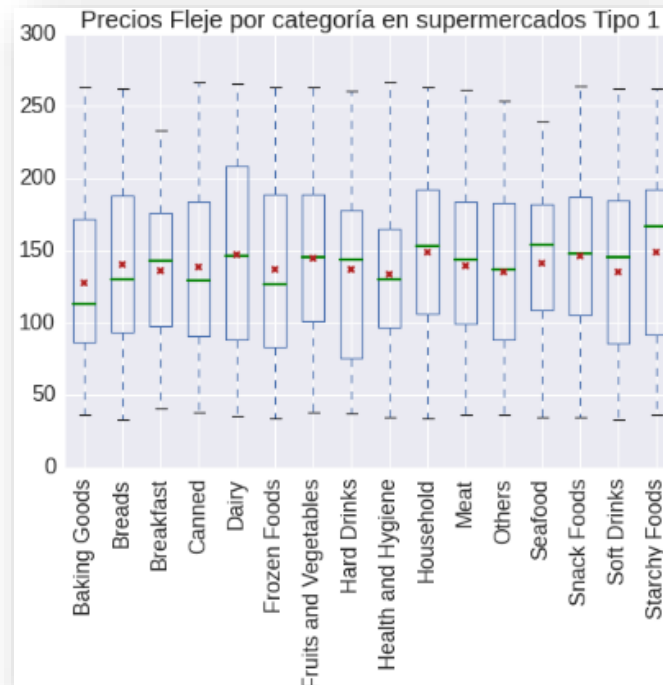
**Como es el canal más relevantes, se procederá solo a analizar este canal dentro de las demás variables**



# ANÁLISIS DE GRÁFICOS

## PRECIO MRP COMO VARIABLE CLAVE DE VENTA

**2. No se observa mucha diferencia entre los precios promedio de venta:** Se puede interpretar que la mayor venta no va necesariamente por la naturaleza del precio, sino que más bien por la naturaleza de cada variables



# MODELO SUPERMERCADO TYPE 1

## CORRELACIONES ENTRE VARIABLES

	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Item_Outlet_Sales	Rank_Item
Item_Weight	1.000000	-0.007025	0.036706	-0.015736	-0.013788	-0.013122	0.019131	-0.022889
Item_Visibility	-0.007025	1.000000	0.001851	0.009889	0.002952	0.004669	-0.013367	0.012632
Item_MRP	0.036706	0.001851	1.000000	-0.004084	0.005673	-0.000200	0.682655	-0.705509
Outlet_Establishment_Year	-0.015736	0.009889	-0.004084	1.000000	0.443472	0.512428	0.012395	-0.012465
Outlet_Size	-0.013788	0.002952	0.005673	0.443472	1.000000	0.607630	-0.006787	0.006293
Outlet_Location_Type	-0.013122	0.004669	-0.000200	0.512428	0.607630	1.000000	0.003622	-0.003263
Item_Outlet_Sales	0.019131	-0.013367	0.682655	0.012395	-0.006787	0.003622	1.000000	-0.944950
Rank_Item	-0.022889	0.012632	-0.705509	-0.012465	0.006293	-0.003263	-0.944950	1.000000

## REGRESIÓN LINEAL

**X = "Item\_MRP", "Item\_Weight"**

**Y= "Item\_Outlet\_Sales"**

**Score= 0,466053**