# **Experiment 1**

Aim: Introduction to Data science and Data preparation using Pandas steps.

# Theory:

Data Science is a multidisciplinary field that involves collecting, processing, analyzing, and visualizing data to extract meaningful insights. It uses techniques from statistics, machine learning, and programming to solve real-world problems. One of the essential steps in data science is data preparation, which ensures that raw data is cleaned and structured for analysis.

Key Steps in Data Preparation Using Pandas:

- 1. Loading Data
- 2. Exploring Data
- 3. Handling Missing Values
- 4. Data Transformation
- 5. Feature Engineering

#### **Dataset Overview:**

The dataset consists of 12 columns, each providing key insights into retail products, their pricing, visibility, and sales performance across different outlets. Below is a breakdown of the dataset's columns and their significance:

- **Item\_Identifier:** A unique code assigned to each product, helping distinguish different items in the inventory.
- **Item\_Weight:** The weight of a product, which can influence logistics, storage, and customer purchasing decisions.
- **Item\_Fat\_Content:** Categorizes products as Low Fat or Regular Fat, reflecting their nutritional composition and target consumers.
- **Item\_Visibility**: Measures how prominently an item is displayed on store shelves, impacting its likelihood of being purchased.
- Item\_Type: Classifies products into broad categories like Dairy, Beverages, or Snacks, aiding in trend analysis across different product segments.
- **Item\_MRP:** The maximum retail price, which plays a crucial role in determining affordability and customer demand.
- Outlet\_Identifier: A unique code assigned to each retail outlet, linking products to specific store locations.
- Outlet\_Establishment\_Year: The year an outlet was established, useful for analyzing how store maturity influences sales performance.
- Outlet\_Size: Defines store sizes as Small, Medium, or Large, affecting customer foot traffic and product demand.

- Outlet\_Location\_Type: Indicates whether a store is located in an Urban, Suburban, or Tier 3 area, capturing demographic influences on sales.
- **Outlet\_Type:** Differentiates between store formats, such as Grocery Stores and Supermarkets, highlighting variations in business models.
- **Item\_Outlet\_Sales:** The target variable, representing the total revenue generated by a product at a particular outlet.

### **Problem Statement:**

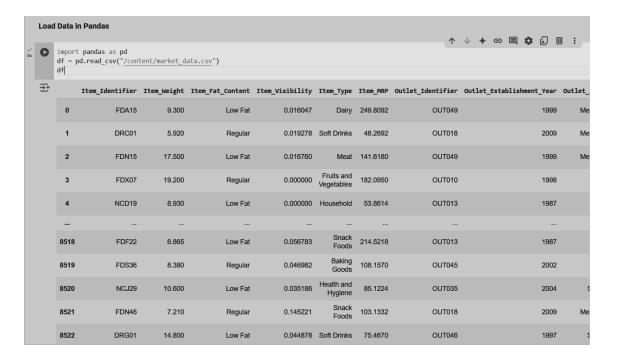
The given dataset provides comprehensive details about retail product sales, focusing on the relationship between product attributes, outlet characteristics, and sales performance. This analysis aims to address the following key objectives:

- **Product Performance**: Identifying products or product types that drive the highest sales and those underperforming.
- Outlet Insights: Understanding the impact of outlet size, location, and type on overall sales performance.
- Pricing Analysis: Investigating how pricing strategies, such as maximum retail price (MRP), influence customer purchasing behavior.
- **Possible sales Prediction**: Developing models to predict item-level sales and provide actionable insights for inventory and pricing strategies.

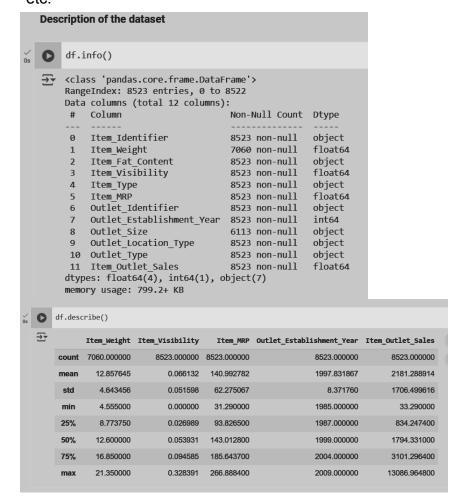
By preprocessing the dataset and applying statistical analysis, the goal is to extract meaningful patterns that can guide data-driven decisions in retail operations.

## Steps:

Load Data in Pandas
 The first step is to load our excel sheet or dataset using read csv method.



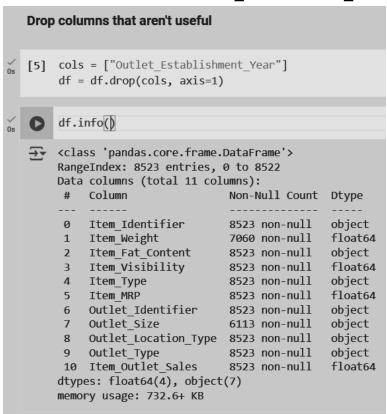
 Description of the dataset df.info() provides information about the dataset like what are the different columns or features present in the dataset, what are their respective data types etc.



The describe method helps us understand the instances or the value in every row. It gives information like number of values, mean, max value for every column.

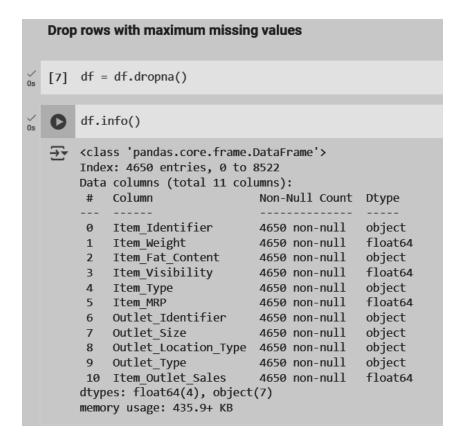
# 3. Drop columns that aren't useful

It may not be necessary that all features present in our dataset will contribute to our analysis. There would be columns which are not required and which increase the size of data unnecessarily. In such cases we drop entire such columns. In our dataset the feature "Outlet Establishment Year" is not useful, so we drop it.

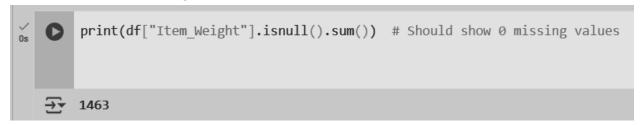


# 4. Drop rows with maximum missing values.

In our dataset there are instances that have some missing values. We want a proper dataset where there are no missing values. In order to do this, we simply remove the rows with null values in them. To do this we use the dropna method. After deleting all null values rows, we observe that the number of non null values for every feature is now the same.



### Take care of missing values



Analysis of our dataset shows that "Item\_Weight" has 1463 missing values. This can cause problems in our analysis. So to address this, we will first categorize items by their types(as similar items may have similar weights), then we find their respective mean and fill null values with the mean.

```
df['Item_Weight'] = df.groupby('Item_Type')['Item_Weight'].transform(lambda x: x.fillna(x.mean()))
print(df["Item_Weight"].isnull().sum()) # Should show 0 missing values
```

### 6. Find out outliers (manually)

We will use the IQR method to calculate outliers for "Item\_Outlet\_Sale": First Quartile Q1 = 834.2474
Second Quartile Q2 = 3101.2964

IQR = Q3-Q1 = 2267.049

Lower Bound = 834.2474 - 1.5x2267.049 = -2566.3261

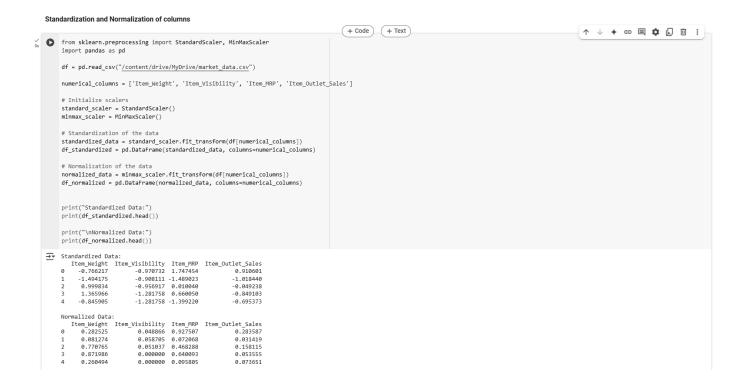
Upper Bound = 3101.2964 - 1.5x2267.049 = 6501.8699

So any value less than -2566.3261 or greater than 6501.8699 can be considered as outliers. (Last column represents Item Outlet Sales)

Item Iden	Item Weight	tem Fat	Item Visib Ite	m Type Iten	n MRP Outl	et Ide Out	et Est Outle	t Siz Outle	t LocOut	let Tyr Item Outlet Sales
ED A 1 E	- 1		0.010047 D		2 0002 011	_	1000 14-4	_	_	272E 420
FDA46	13.0	Low ⊦aτ	0.11/818	Snack Foo	192.9130	JU1049	1999	ıvıeaıum	Her 1	Supermark 2527.377
FDC02	21.35	Low Fat	0.069103	Canned	259.9278	OUT018	2009	Medium	Tier 3	Supermark 6768.523
EDLEA	10.15	nl	0.042270	C	43C FOAC	OLITO13	1007	ridul.	T: 1	C 272 E420
NCP30	20.5	Low Fat	0.032835	Household	40.2822	OUT045	2002	2	Tier 2	Supermark 707.0796
FDY25		Low Fat	0.03381	Canned	180.5976	OUT027	1985	Medium	Tier 3	Supermark 7968.294
NCHSA	12 5	Low Fat	N N72660	Household	160 202	ULITUAE	100	llem2	Tior 1	Sunarmark 1/128 178
NCR53		Low Fat	0.144338	Health and	224.4404	OUT027	1985	Medium	Tier 3	Supermark 6976.252
FDCF3		I F-4			101 7016		1000		T: 2	C 101 201C
110011		LOW TUE		ricultii uiit				Jilluli	TICL Z	Supermun SZ 10, 150
FDY56	16.35	Regular			227.6062	OUT017	2007		Tier 2	Supermark 7222.598
r		-								
FDH19		Low Fat	0.032928	Meat	1/3.1/38	001027	1985	Medium	Her 3	Supermark /298.5
FDY55	16.75	Low Fat	0.081253	Fruits and	256.4988	OUT013	1987	High	Tier 3	Supermark 7452.965
EUAUS	17.6	Dogular	n n76553	Most	110 5202	OUT017	2007	,	Tior 2	Supermark 450 0000
FDO23	17.85	Low Fat	0.14/024	Breads	93.7436	OUTUIN	200	Medium	Her 3	Supermark 1134.523
DRE60	9.395	Low Fat	0.159658	Soft Drinks	224.972	OUT045	200	2	Tier 2	Supermark 7696.648
DRP47	15.75	Low Fat	0.141399	Hard Drink	250.5382	OUT017	200	7	Tier 2	Supermark 2775.72
FUU55	10.2	Low Fat	U.U35984	Fruits and	Z0U.0Z/8	UU1045	2002	!	Her 2	Supermark 4425.5/3
FDN58		Regular	0.056597	Snack Foo	230.9984	OUT027	1985	Medium	Tier 3	Supermark 9267.936
EDC44		1 F-4	0.3047	Г Г	76 067	OLITO40	1000	CII	T: 4	CC+ 220 701
FDI44	10.1	Low Fat	0.100389	Fruits and	/७.ሀ328	UU 1U49	1995	ivieaium	Her 1	Supermark 1853.587
FDW56		Low Fat	0.070557	Fruits and	191.2162	OUT027	1985	Medium	Tier 3	Supermark 7504.232
ED 404	4.5	в I	0.0546	c 1	FO 4004	OLITO40	2000	K # 11	T' 3	
FUE43	12.1	LOW Fat	U.U4U3ZZ	Fruits and	1/8.5002	OTITION	2005	viedium	Her 3	Supermark 3332.100
FDR35		Low Fat	0.020597	Breads	200.0742	OUT027	1985	Medium	Tier 3	Supermark 8958.339
EDTEO		1 F-4	0.000000	Caralli Ear	100 2010	OLITO27	1000	N A = 41	T1 2	C 2522 444

# 7. Standardization and Normalization

Standardization and normalization are crucial in data preprocessing because they bring features to a common scale, improving model performance and convergence. Standardization (Z-score scaling) centers data around zero with unit variance, while Normalization (Min-Max scaling) rescales data to a fixed range (e.g., 0 to 1), ensuring fair weight distribution across features.



#### Conclusion:

In this experiment, we explored the fundamental steps of data preparation using Pandas in Data Science. We loaded and analyzed a retail dataset, identified missing values, and handled them effectively. We also detected and removed outliers using the IQR method to ensure clean and reliable data. Additionally, we discussed feature selection, data transformation, and the importance of standardization and normalization in preparing data for further analysis.