Sales Data Analysis and Prediction

All column

1. Item\_Identifier: Unique identifier for each item.
2. Item\_Weight: Weight of the item.
3. Item\_Fat\_Content: Indicates the fat content of the item (e.g., low fat, regular).
4. Item\_Visibility: The percentage of total display area of all products in a store allocated to the particular product.
5. Item\_Type: The category of the item.
6. Item\_MRP: Maximum retail price of the item.
7. Outlet\_Identifier: Unique identifier for each outlet/store.
8. Outlet\_Establishment\_Year: The year in which the outlet was established.
9. Outlet\_Size: The size of the outlet (e.g., small, medium, large).
10. Outlet\_Location\_Type: The type of location where the outlet is situated (e.g., urban, suburban, rural).
11. Outlet\_Type: The type of outlet (e.g., grocery store, supermarket).
12. Item\_Outlet\_Sales: The sales of the item in the particular outlet.

dataset contains 8523 rows (data points) and 12 columns (features). This means you have information on 8523 items/outlets, with 12 different attributes recorded for each item/outlet.

Categorical features

1. Item\_Identifier
2. Item\_Fat\_Content
3. Item\_Type
4. Outlet\_Identifier
5. Outlet\_Size
6. Outlet\_Location\_Type
7. Outlet\_Type

Numerical features

1. Item\_Weight
2. Item\_Visibility
3. Item\_MRP
4. Outlet\_Establishment\_Year
5. Item\_Outlet\_Sales
6. **Item\_Weight**: There are 1463 missing values in this column.
7. **Outlet\_Size**: There are 2410 missing values in this column.

Item\_Identifier 0

Item\_Weight 0

Item\_Fat\_Content 0

Item\_Visibility 0

Item\_Type 0

Item\_MRP 0

Outlet\_Identifier 0

Outlet\_Establishment\_Year 0

Outlet\_Size 0

Outlet\_Location\_Type 0

Outlet\_Type 0

Item\_Outlet\_Sales 0

1. **Item\_Weight**:
   * Count: 8523
   * Mean: 12.857645
   * Standard Deviation: 4.226124
   * Minimum: 4.555
   * 25th Percentile (Q1): 9.310
   * Median (50th Percentile or Q2): 12.857645
   * 75th Percentile (Q3): 16.000
   * Maximum: 21.350
2. **Item\_Visibility**:
   * Count: 8523
   * Mean: 0.066132
   * Standard Deviation: 0.051598
   * Minimum: 0.000000
   * 25th Percentile (Q1): 0.026989
   * Median (50th Percentile or Q2): 0.053931
   * 75th Percentile (Q3): 0.094585
   * Maximum: 0.328391
3. **Item\_MRP**:
   * Count: 8523
   * Mean: 140.992782
   * Standard Deviation: 62.275067
   * Minimum: 31.290000
   * 25th Percentile (Q1): 93.826500
   * Median (50th Percentile or Q2): 143.012800
   * 75th Percentile (Q3): 185.643700
   * Maximum: 266.888400
4. **Outlet\_Establishment\_Year**:
   * Count: 8523
   * Mean: 1997.831867
   * Standard Deviation: 8.371760
   * Minimum: 1985
   * 25th Percentile (Q1): 1987
   * Median (50th Percentile or Q2): 1999
   * 75th Percentile (Q3): 2004
   * Maximum: 2009
5. **Item\_Outlet\_Sales**:
   * Count: 8523
   * Mean: 2181.288914
   * Standard Deviation: 1706.499616
   * Minimum: 33.290000
   * 25th Percentile (Q1): 834.247400
   * Median (50th Percentile or Q2): 1794.331000
   * 75th Percentile (Q3): 3101.296400
   * Maximum: 13086.964800

The output **(8523, 11)** represents the shape of your original feature matrix **X**, which contains 8523 rows (samples) and 11 columns (features).

The output **(6818, 11)** represents the shape of your training feature matrix **X\_train**, which contains 6818 rows (samples) and 11 columns (features). This indicates that 80% of your data has been allocated for training.

The output **(1705, 11)** represents the shape of your testing feature matrix **X\_test**, which contains 1705 rows (samples) and 11 columns (features). This indicates that 20% of your data has been allocated for testing