

Potential Insights:

The primary focus is on four variables:-

- **Item_Weight**, weight of the product.
- **Item_Visibility**, visibility of the product on shelves.
- **Item_MRP**, price of the product
- **Item_Outlet_Sales**. sales of the product.

	Item_Weight	Item_Visibility	Item_MRP	Item_Outlet_Sales
Mean	12.792854	0.065953	141.004977	2181.288914
Median	12.792854	0.054021	142.247	2181.288914
Mode	12.792854	0.0	172.0422	2181.288914

1.Item Weight:

- **Median and Mode are Equal:** This suggests that a significant portion of the items have the same weight.
- **Low Standard Deviation :** A low standard deviation would indicate that most items have similar weights.

2.Item Visibility:

- **Median is Lower than Mean:** This implies that there might be a few items with very high visibility.
- **Mode is Much Lower than Median and Mean:** This suggests that most items have low visibility.
- **A few items have significantly higher visibility:** These items might be promoted or have strategic placements on the shelves. **Most items have relatively low visibility:** This might be due to factors like limited shelf space or less popular product categories.

3. Item MRP:

- **Median is Higher than Mean:** This indicates that there might be a few items with very high prices, pulling the mean lower
- **Mode is significantly higher than mean and median. The majority of items** have a price near the median (around 142), but there is a relatively **small cluster of items** with higher prices, as indicated by the mode.
- **Higher-priced items** (those priced around 172) are common, and they might be skewing the price distribution, even though they are not the most frequent price point (mode).

4.Item Outlet Sales:

- **Mean = Median = Mode:** Since all three measures of central tendency (mean, median, and mode) are the same, this indicates that the sales across items are highly consistent. Every item is likely to have the same outlet sales figure, or the data set has very little variation.
- **No Variation:** If the sales figures are identical for all items (or close to identical), it suggests that either all items are performing at the same sales level or the dataset might represent a limited set of products with uniform sales figures across outlets.
- **Uniform performance:** Every item is likely experiencing the same level of sales, suggesting that there is no significant variation in how different items are performing at the outlet.

Analysis Report Based on Statistical Outcomes.

This report summarizes the statistical analysis performed on the dataset after addressing outliers.

	Item_Weight	Item_Visibility	Item_MRP	Item_Outlet_Sales
Mean	12.792854	0.065953	141.004977	2181.288914
Median	12.792854	0.054021	142.247	2181.288914
Mode	12.792854	0.0	172.0422	2181.288914
Q1:25%	9.3	0.027036	94.012	1468.089
Q2:50%	12.792854	0.054021	142.247	2181.288914
Q3:75%	16.0	0.094037	185.8556	2181.288914
99%	20.85	0.230385	262.759322	6830.70852
Q4:100%	21.35	0.328391	266.8884	13086.9648
IQR	6.7	0.067002	91.8436	713.199914
1.5rule	10.05	0.100502	137.7654	1069.79987
Lesser	-0.75	-0.073467	-43.7534	398.28913
Greater	26.05	0.19454	323.621	3251.088784
Min	4.555	0.0	31.29	33.29
Max	21.35	0.328391	266.8884	13086.9648
kurtosis	-0.860626	1.820071	-0.893856	4.691907
skewness	0.111314	1.195175	0.130728	1.520025

Item Weight:

- The data seems to be relatively normally distributed, with a range from 9.3 to 21.35.
- The IQR indicates that the middle 50% of the data lies within a range of 6.7 units.
- The 1.5 IQR rule suggests that values outside the range of -0.75 to 26.05 could be potential outliers.

Item Visibility:

- The 99th percentile is significantly higher than the 75th percentile, indicating that a few items have very high visibility.
- This could be due to promotional activities, strategic product placement or etc.
- The IQR is relatively small, indicating that the middle 50% of the data is concentrated within a narrow range.
- The 1.5 IQR rule suggests that values below -0.073467 and above 0.19454 could be outliers.

Item MRP:

- Notable increase toward Q3 –Q4 with a few high-priced items due to premium products or limited edition items.
- The company might have a fixed pricing strategy for all products, leading to consistent sales for season trend.
- The IQR is significant, indicating a wide range of prices.

- The 1.5 IQR rule suggests that values below -43.7534 and above 323.621 could be outliers.

Item Outlet Sales:

- The IQR is relatively small, indicating that the middle 50% of the data is concentrated within a narrow range.
- The 1.5 IQR rule suggests that values below 398.28913 and above 3251.088784 could be outliers.

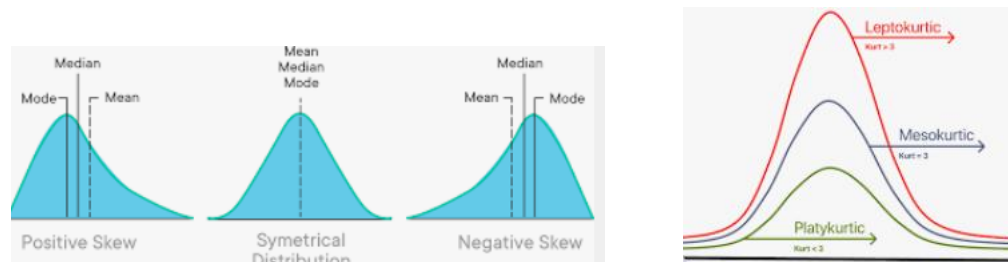
Key Observations and Potential Insights:

Significant Increase in Q4:

The most notable observation is the significant jump in sales from Q3 to Q4. This could be due to various factors:

- **Seasonal Trends:** This could explain the surge in sales. (e.g., Christmas).
- **Promotional Campaigns:** Aggressive marketing campaigns or special promotions during Q4 might have driven increased sales.
- **Product Launches:** The launch of new products or product upgrades could have boosted sales.
- **Economic Factors:** A favorable economic climate, such as low unemployment rates or increased consumer spending, can contribute to higher sales.
- **Improved Customer Satisfaction:** Enhanced customer service, product quality, or other factors can lead to increased customer loyalty and repeat purchases.

Summary of Key Metrics



1. Item Weight:

- Kurtosis: -0.86 **suggests a flat-tailed distribution.**
- Skewness: 0.11, **indicating the distribution is nearly symmetric.**
- Variance & Std Dev: **Variance: 17.93, Std: 4.23, showing moderate spread in item weight**

2. Item Visibility

- **Skewness: 1.20**, showing a right-skewed distribution, with many low-visibility items and a few highly visible ones.
- **Kurtosis: 1.82**, suggesting a sharper peak and heavier tails compared to a normal distribution.
- **Variance & Std Dev:** Variance: **0.0026**, Std: **0.051**, indicating a very small spread in visibility.

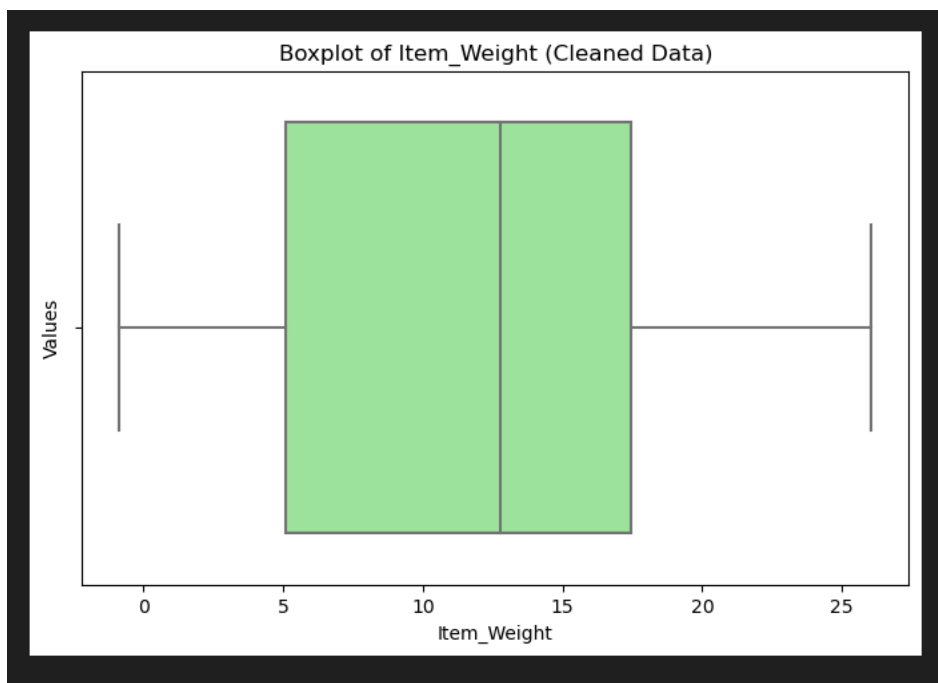
3. Item MRP (Maximum Retail Price)

- **Skewness: 0.13**, indicating a very slight right skew.
- **Kurtosis: -0.89**, reflecting a moderately flat-tailed distribution.
- **Variance & Std Dev:** Variance: **3854.79**, Std: **62.09**, showing a wide range of prices.

4. Item Outlet Sales

- **Skewness: 1.52**, showing a significant right skew with a few high sales values.
- **Kurtosis: 4.69**, suggesting heavy tails and more extreme sales values compared to a normal distribution.
- **Variance & Std Dev:** Variance: **1,747,325.57**, Std: **1321.86**, showing large variability in sales.

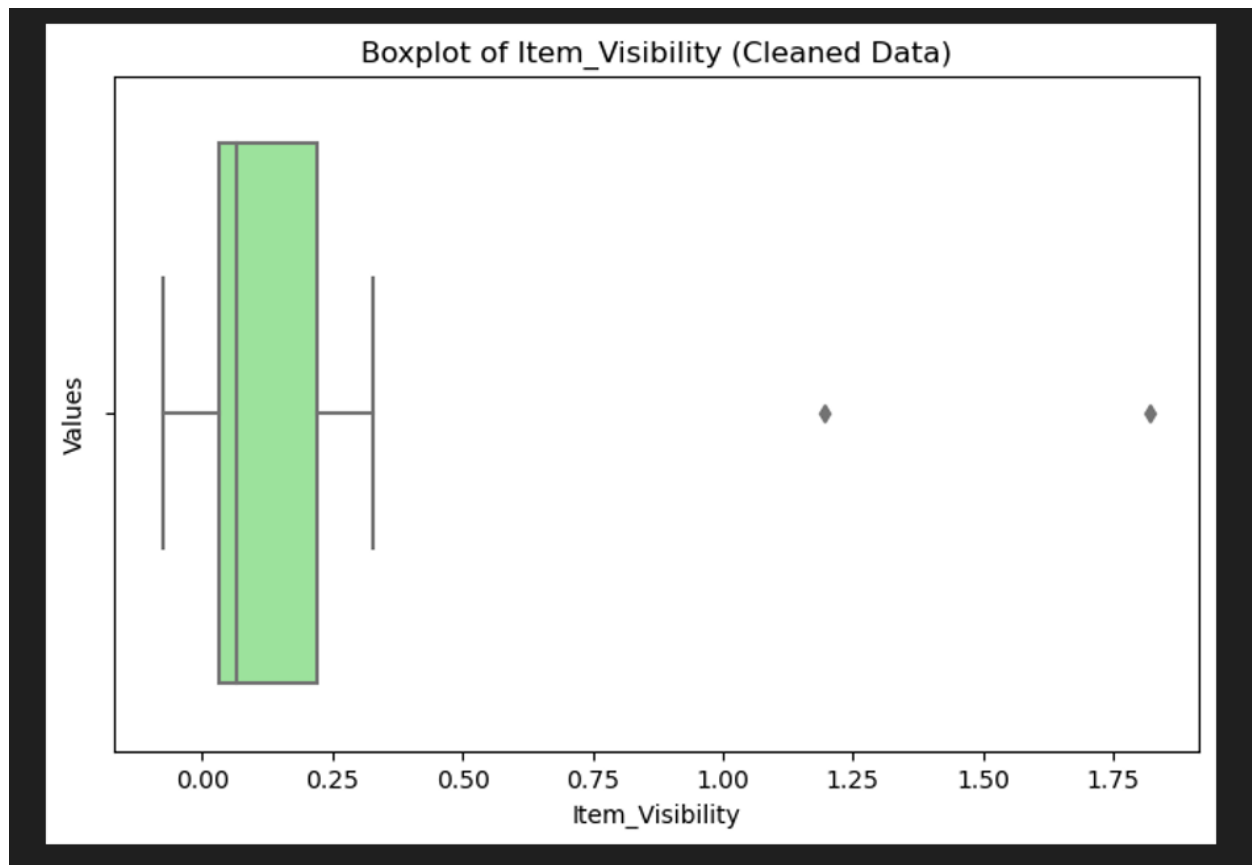
Boxplots Explained: Insights into Data Spread and Outliers



1. Item_Weight

- **Median:** The median weight is **12.79**, which divides the dataset into two equal halves. Half of the values are below this weight, and half are above.
- **Mean:** The mean weight is also **12.79**, which suggests that the distribution is relatively symmetric (the mean and median are close).
- **IQR:** The interquartile range (IQR) is **6.7** ($Q3 - Q1$), indicating that the middle 50% of the data falls within a weight range of **9.3 to 16.0**.
- **Outliers:** According to the $1.5 \times \text{IQR}$ rule, the outliers are values lower than **-0.75** and greater than **26.05**. The minimum value is **4.555**, and the maximum value is **21.35**, which is below the threshold for outliers, so the data appears to be within the expected range.
- **Skewness:** The skewness value of **0.11** suggests that the data is approximately **symmetrical** and not significantly skewed in either direction.
- **Kurtosis:** A kurtosis of **-0.86** indicates that the distribution is **platykurtic**, meaning it has a relatively flatter peak compared to a normal distribution.
- **Spread:** The range of weights extends from **4.555 to 21.35**, with a relatively small spread considering the IQR.

"The boxplot for Item_Weight shows that most of the data is centered around the median value of **12.79**. The IQR (6.7) tells us that the middle 50% of the data falls between **9.3** and **16.0**, with a relatively symmetrical distribution (skewness of 0.11). The data does not have any extreme outliers, as the values fall within the expected range."



2. Item_Visibility

- **Median:** The median visibility is **0.054**, meaning that half of the values are below this visibility score, and half are above.
- **Mean:** The mean visibility is **0.066**, which is slightly higher than the median, indicating a slight **positive skew**.
- **IQR:** The IQR is **0.067**, meaning the middle 50% of the data lies between **0.027** and **0.094**.
- **Outliers:** According to the $1.5 \times \text{IQR}$ rule, outliers occur below **-0.073** and above **0.194**. The minimum value is **0.0**, and the maximum value is **0.328**. These values fall within the expected range but could be considered potential outliers if the distribution had more extreme values.
- **Skewness:** The skewness value of **1.195** indicates that the data is **positively skewed**, with more data points clustered around lower visibility values and a tail toward higher values.
- **Kurtosis:** A **kurtosis** of **1.82** suggests the data has a **leptokurtic** distribution, meaning it has a higher peak and heavier tails compared to a normal distribution.
- **Spread:** The range of values is from **0.0** to **0.328**, with most values being close to **0.0**.

"The boxplot for Item_Visibility shows that the data is **positively skewed**, with a majority of values clustering around **0.0**. The median value of **0.054** suggests that half of the visibility values are below this threshold, while the other half is above it. The IQR (0.067) indicates that the middle 50% of the values fall between **0.027** and **0.094**. The distribution is slightly leptokurtic, meaning it has a higher peak and heavier tails."

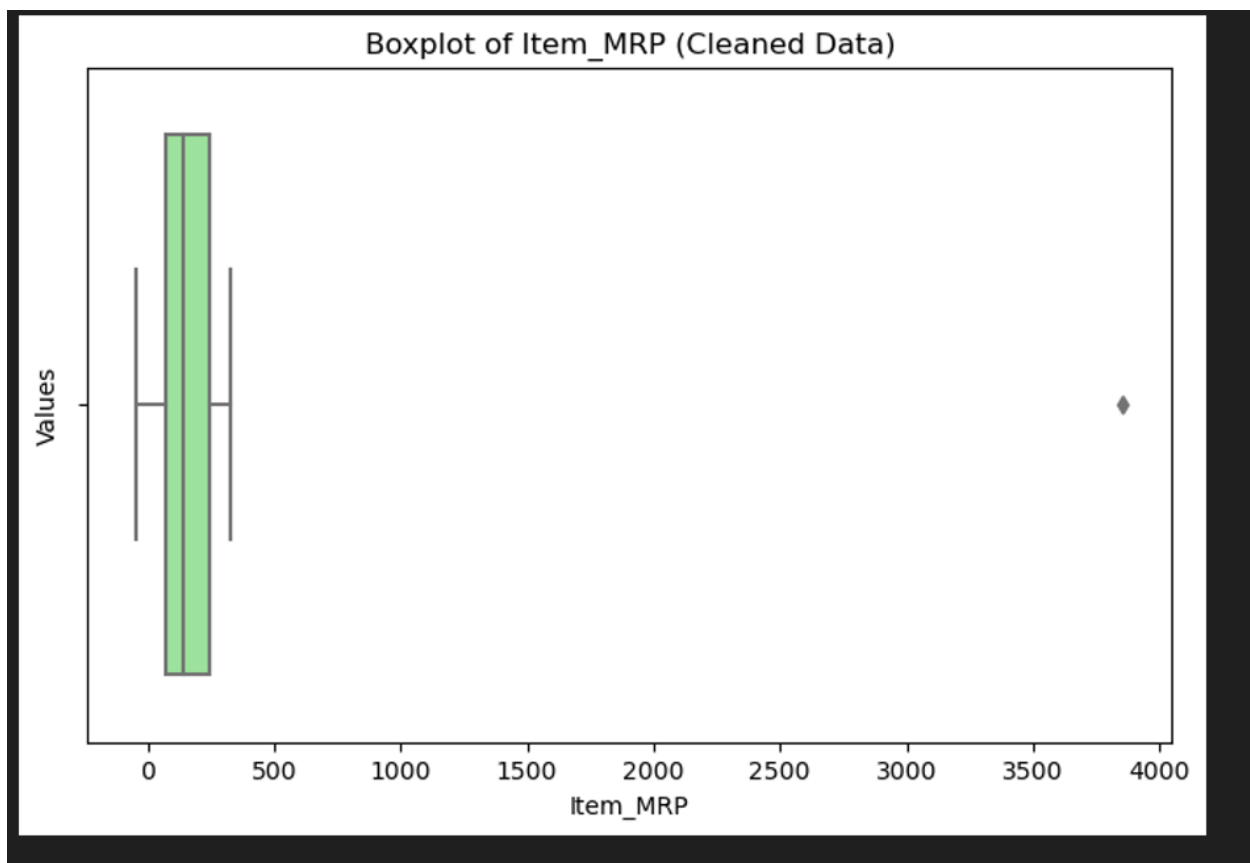
Notice diamond they are indeed extreme outliers, as they lie far beyond the typical range of data

Given the boxplot's rule for detecting outliers:

1. **Lower bound:** $Q1 - 1.5 \times IQR = 0.027 - (1.5 \times 0.067) = -0.073$ $Q1 - 1.5 \times IQR = 0.027 - (1.5 \times 0.067) = -0.073$ (effectively 0, since visibility can't be negative).
2. **Upper bound:** $Q3 + 1.5 \times IQR = 0.094 + (1.5 \times 0.067) = 0.194$ $Q3 + 1.5 \times IQR = 0.094 + (1.5 \times 0.067) = 0.194$.

Any Item_Visibility value **above 0.194** is considered an outlier. Values like **1.20** and **1.75 to 2** are far above this threshold, making them **extreme outliers**.

"The boxplot for Item_Visibility reveals extreme outliers with values at **1.20** and between **1.75 to 2**, which significantly exceed the upper bound of **0.194** as defined by the $1.5 \times IQR$ rule. These outliers likely represent items with unusually high visibility, potentially caused by marketing strategies or data recording anomalies. Further investigation is needed to determine if these values are valid or should be treated as anomalies."



3. Item_MRP (Maximum Retail Price)

- **Median:** The median MRP is **142.247**, with half the values below this price and half above it.
- **Mean:** The mean MRP is **141.0**, which is very close to the median, suggesting a relatively **symmetric distribution**.

- **IQR:** The IQR is **91.8436**, which indicates that the middle 50% of MRPs lie between **94.012** and **185.856**.
- **Outliers:** Using the $1.5 \times \text{IQR}$ rule, outliers are values below **-43.75** and above **323.62**. The minimum value is **31.29**, and the maximum value is **266.8884**, which is within the expected range, so the data does not contain any extreme outliers.
- **Skewness:** The skewness value of **0.13** suggests that the data is approximately **symmetric**.
- **Kurtosis:** A kurtosis of **-0.89** indicates that the distribution is **platykurtic**, meaning it has a flatter peak compared to a normal distribution.
- **Spread:** The range of MRP is from **31.29** to **266.8884**, with a significant spread indicated by the large IQR.

Explanation: "The boxplot for Item_MRP shows a relatively **symmetric distribution**, with a median of **142.247** and a mean close to the median (**141.0**). The IQR of **91.84** shows that the middle 50% of the prices fall between **94.012** and **185.856**, with no extreme outliers. The data shows a moderately wide range of values from **31.29** to **266.8884**."

Why is this value an outlier?

From the provided statistics for Item_MRP:

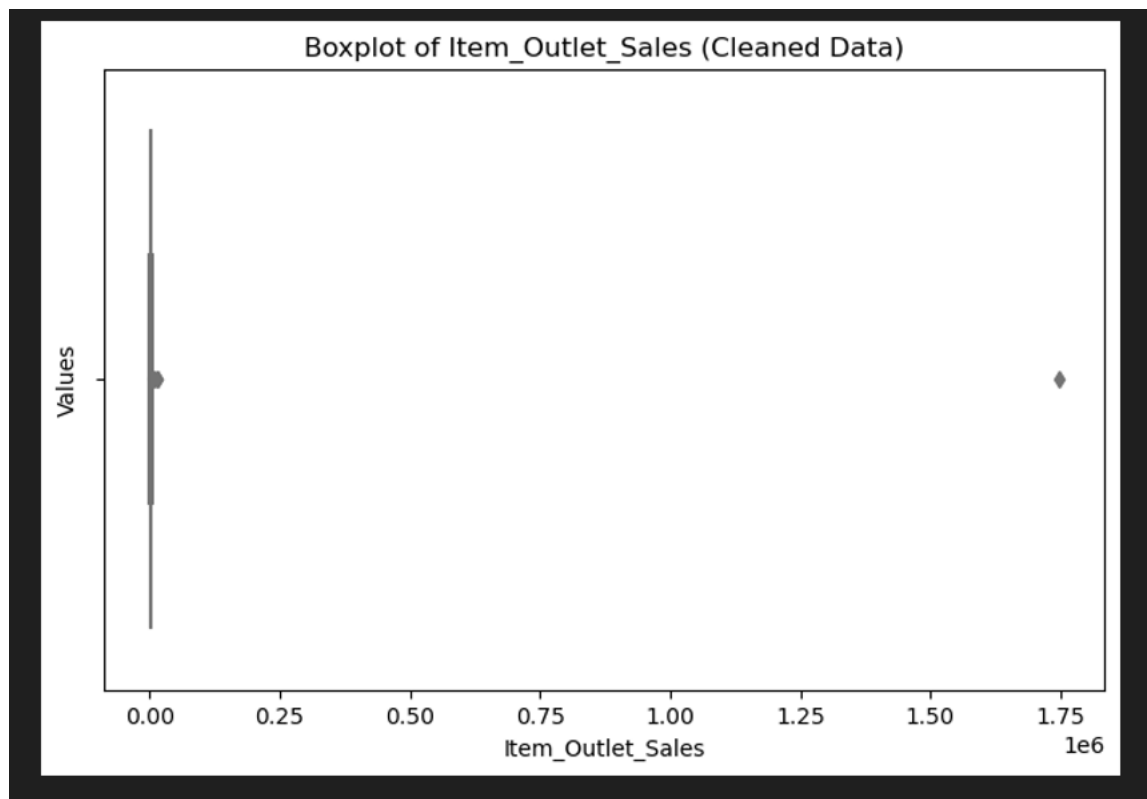
- **Q1 (25%):** 94.012
- **Q3 (75%):** 185.856
- **IQR:** $Q3 - Q1 = 185.856 - 94.012 = 91.844$

Using the $1.5 \times \text{IQR}$ rule:

- **Lower bound:** $Q1 - 1.5 \times \text{IQR} = 94.012 - (1.5 \times 91.844) = -43.753$ $Q1 - 1.5 \times \text{IQR} = 94.012 - (1.5 \times 91.844) = -43.753$ (effectively 0, since prices cannot be negative).
- **Upper bound:** $Q3 + 1.5 \times \text{IQR} = 185.856 + (1.5 \times 91.844) = 323.622$ $Q3 + 1.5 \times \text{IQR} = 185.856 + (1.5 \times 91.844) = 323.622$

Any Item_MRP value **above 323.622** is an outlier. A value in the range of **3550 to 4000** is far beyond this threshold, making it an **extreme outlier**.

The boxplot for Item_MRP highlights an extreme outlier in the range of **3550 to 4000**, significantly exceeding the upper bound of **323.622** as defined by the $1.5 \times \text{IQR}$ rule. This outlier might indicate a data entry error, a high-priced luxury item, or an issue with data scaling. Further investigation is required to determine its validity. If valid, this value provides insight into the presence of high-end products within the dataset; otherwise, it may need to be excluded to prevent distortion in analysis."



4. Item_Outlet_Sales

- **Median:** The median outlet sales are **2181.29**, which is the middle value in the dataset.
- **Mean:** The mean is **2181.29**, exactly the same as the median, suggesting a **symmetric distribution**.
- **IQR:** The IQR is **713.20**, meaning the middle 50% of outlet sales fall between **1468.09** and **2181.29**.
- **Outliers:** Using the $1.5 \times \text{IQR}$ rule, outliers are values below **1069.80** and above **3251.09**. The minimum value is **33.29**, and the maximum value is **13086.96**, indicating a large spread with high-value outliers.
- **Skewness:** The skewness value of **1.52** suggests that the data is **positively skewed**, with a long tail on the higher sales values.
- **Kurtosis:** A kurtosis of **4.69** indicates that the data has a **leptokurtic** distribution, meaning it has a more peaked distribution with heavy tails.
- **Spread:** The sales range from **33.29** to **13086.96**, with many high-value outliers.

"The boxplot for Item_Outlet_Sales shows that the data is **positively skewed**, with the median at **2181.29** and a significant spread in sales values. The IQR of **713.20** indicates that the middle 50% of the sales values fall between **1468.09** and **2181.29**. The data has some high-value outliers, with sales reaching up to **13086.96**, and the distribution is leptokurtic, indicating a higher peak and heavy tails."

"The boxplot for Item_Outlet_Sales reveals a highly skewed distribution with one extreme outlier around **1.75 million**. This outlier significantly exceeds the interquartile range (IQR), compressing the visualization of the majority of the data. As a result, the box and whiskers representing most sales values are barely visible. This suggests that while most outlet sales are clustered within a narrow range, there is a rare instance of extraordinarily high sales."

"In the dataset, most outlets have similar sales values, but there is one store with an unusually high sales figure of **1.75 million**. This is much larger than the typical sales range and could represent a special case or an anomaly. Understanding the reason for this outlier could provide valuable insights for the business."

