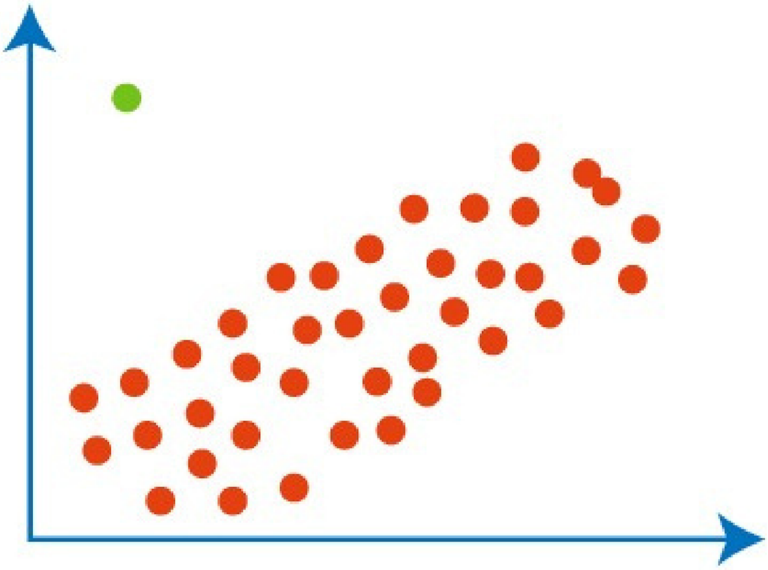
# Day 3: Handling Outliers In Data



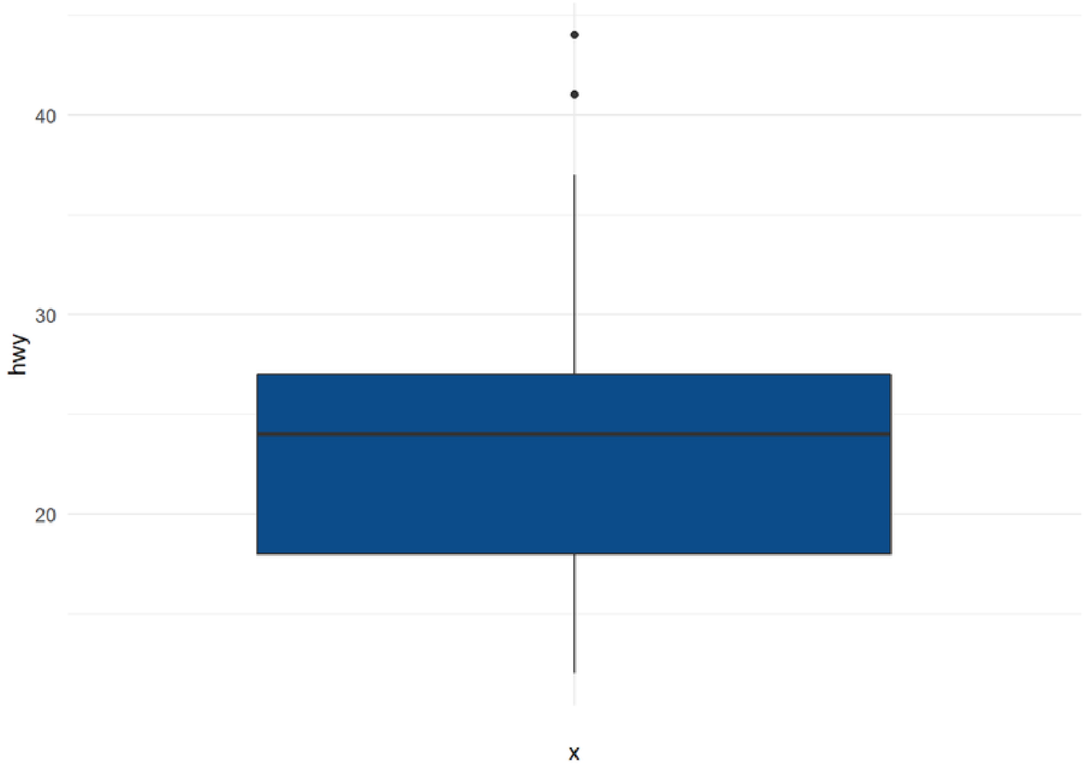
*Author: Alisha Metkari*

# What Is Outlier

- An outlier is a data point that differs from other observations. It is an abnormal observation in the data.
- It can occur due to experimental errors, false data entry, etc.
- Example: Age of human is 130 in data. This is an outlier because the value is unreal for the age of the human.
- The presence of outlier skews the data as their values are different than the normal values. This leads to ungeneralised ML model.
- To avoid this, we need to detect outliers and handle them.
- There are many statistical methods to detect outliers. But before detecting, visualise the outliers.

# Outliers Visualisation Methods

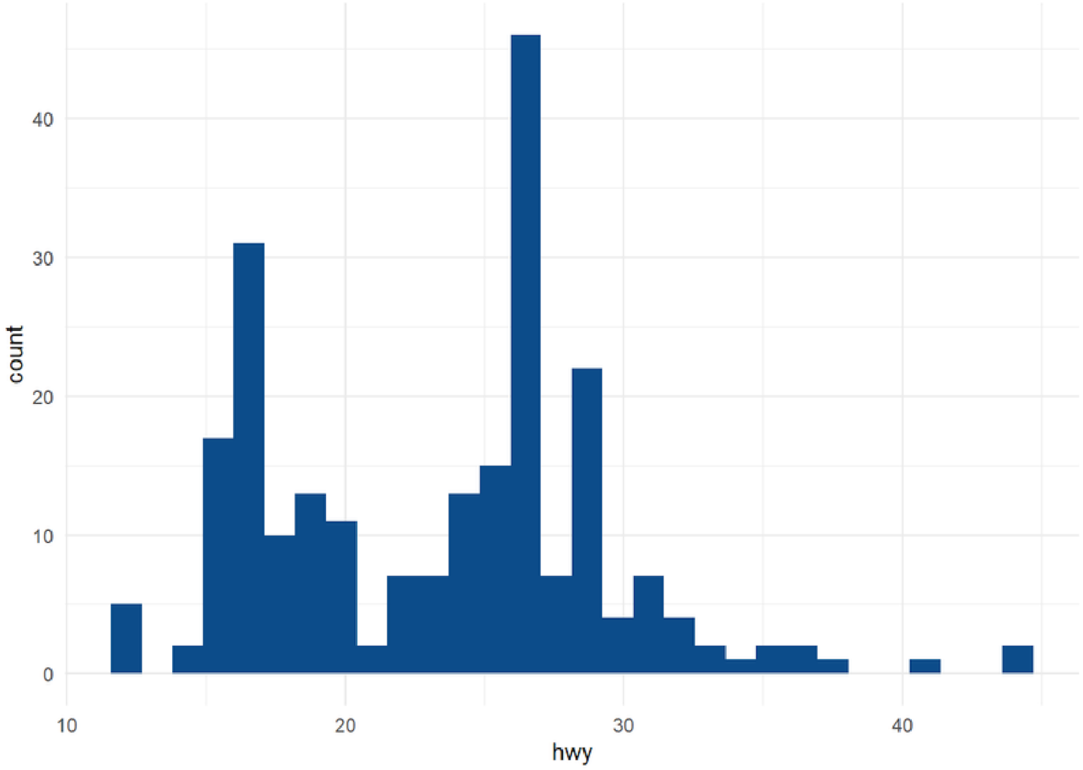
- Box Plot: Any point lying out of the middle line of box plot is an outlier. Box plot follows the IQR rule to show the outliers.



*Author: Alisha Metkari*

# Outliers Visualisation Methods

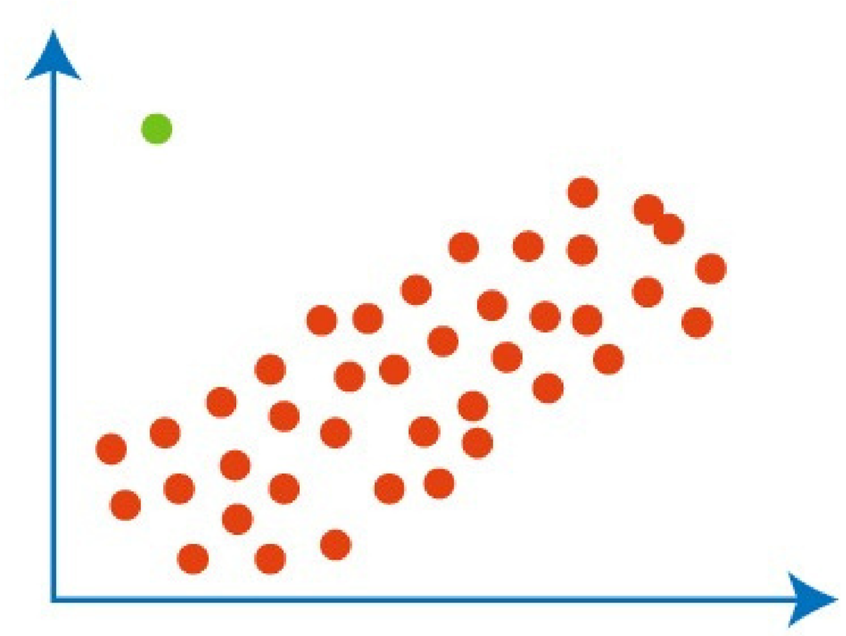
- Histogram: Points that are lying away from the densed part of the histogram are the outliers.



*Author: Alisha Metkari*

# Outliers Visualisation Methods

- Scatter Plot: In a scatter plot, the data points that are lying away from the crowd in isolated place are the outliers.



*Author: Alisha Metkari*

# Outlier Detection Methods

- IQR Method: It uses IQR(Inter Quartile Range) and quartiles. Anything that falls out of the interval [Q1-1.5*IQR, Q3+1.5*IQR] is an outlier. IQR is the difference between third and first quartile.
- Empirical Rule: It states that 68% of data will fall within one standard deviation of the mean, 95% will fall within two standard deviations, and 99.7% will fall within three standard deviations. So, anything falling out of the interval [mean-3*std, mean+3*std] is an outlier.
- Winsorization: It is a process of replacing extreme values in order to limit the effect of outliers. For example, 95% winsorization means replacement of top 5% of data by 95th percentile and bottom 5% of data by 5th percentile.
- Hampler Method: It uses median and MAD(Median Absolute Deviation). Anything that falls out of the interval [median-3*MAD, median+3*MAD] is an outlier.

*Author: Alisha Metkari*

# Ways To Treat Outliers

- Remove outliers if the data is enough to build the model.
- Transform outliers into some other form like log, square root, reciprocal in order to scale all the values in the same range. This will avoid the impact of outliers even if they are present in original data.
- Impute outliers with median, mode, ffill-bfill method. Do not use mean because it is sensitive to outliers.

*Author: Alisha Metkari*

# End Notes

- If outliers are not treated then machine learning model may not generalise well to the data. This is because outlier has unexpected value of the data.
- This leads to the weak performance of the model. Hence treating outliers is the necessary step in data preprocessing.

*Author: Alisha Metkari*