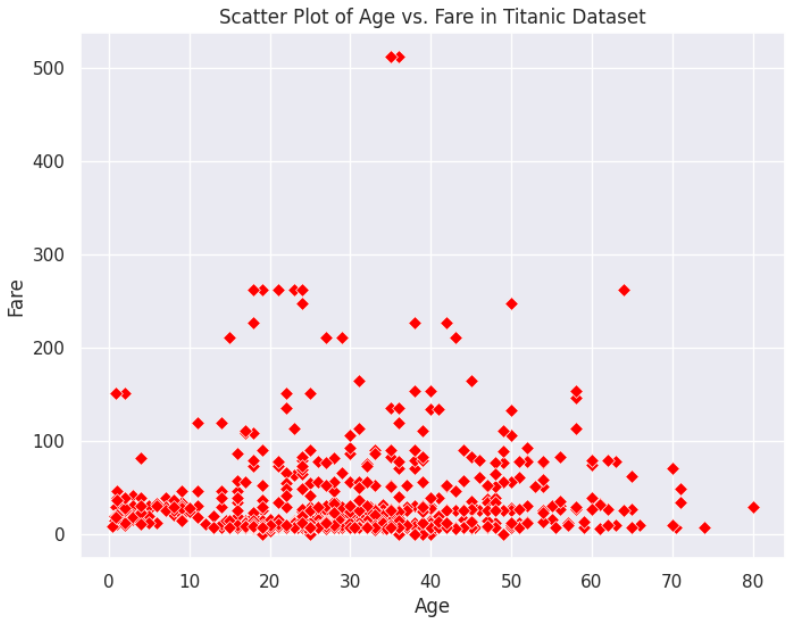
How to interpret scatter plot for Bivariate analysis – numerical vs numerical



A. Interpretation of the Scatter Plot Components:

- **Horizontal Axis (X-axis):** Represents one of the numerical variables, in this case, "Age" of the passengers, ranging from approximately 0 to 80 years.
- **Vertical Axis (Y-axis):** Represents the other numerical variable, "Fare" paid by the passengers, ranging from 0 to over 500.
- **Points:** Each red diamond-shaped point on the plot represents a single passenger in the Titanic dataset. The position of the point is determined by the passenger's age (x-coordinate) and the fare they paid (y-coordinate).

B. Interpreting the Relationship Between Age and Fare:

By examining the pattern of the points, we can try to understand if there's any relationship or correlation between a passenger's age and the fare they paid:

- **Overall Distribution:** The points are scattered across the plot, indicating a wide range of ages and fares among the passengers. The majority of

the points are concentrated in the lower fare range (below 100) and span across all ages.

- **Density:** The density of points can give an idea of where most of the data lies. There's a high density of points in the lower-left quadrant (younger ages and lower fares) and along the bottom (all ages with very low fares).
- **Trend or Correlation:** We look for any discernible pattern or trend in the scattering of the points:
    - **Positive Correlation:** If higher values of one variable tend to occur with higher values of the other (points generally move from the bottom-left to the top-right), it suggests a positive correlation.
    - **Negative Correlation:** If higher values of one variable tend to occur with lower values of the other (points generally move from the top-left to the bottom-right), it suggests a negative correlation.
    - **No Clear Correlation:** If the points are scattered randomly without any apparent pattern, it suggests little or no linear correlation between the variables.

In this particular scatter plot, there doesn't seem to be a strong linear correlation between "Age" and "Fare." Passengers of all ages appear to have paid a wide range of fares, and passengers who paid similar fares also span a wide range of ages.

- **Outliers:** Points that lie far away from the main cluster of points can be considered outliers. In this plot, there are a few points with very high fares (above 200, with one exceeding 500) at various ages. These represent passengers who paid significantly more than the majority. There also seem to be some passengers with age 0 (likely infants) who paid various fares.

Scatter plots are the best choice for visualizing the relationship between two numerical variables when you want to:

- **Explore the type and strength of the correlation:** Determine if there's a positive, negative, or no linear relationship between the variables. You can also get a sense of how strong that relationship

might be (how tightly clustered the points are around a potential trend).

- **Identify patterns and trends:** Look for non-linear relationships, clusters of data points, or other structures in the data that might not be evident from summary statistics alone.
- **Detect outliers:** Easily spot data points that deviate significantly from the general pattern. These outliers can be important for identifying unusual cases or potential errors in the data.
- **Assess the variability of one variable with respect to another:** See how the values of one variable are spread out for different values of the other variable.
- **Provide a visual representation of the raw data points:** Unlike aggregated visualizations, scatter plots show each individual data point, preserving the granularity of the data.

In summary, scatter plots are invaluable for initially exploring the relationship between two numerical variables. They allow you to visually assess correlation, identify patterns, detect outliers, and understand the underlying distribution of the data in a two-dimensional space.