In the data curation process for the heart disease study, I made a strategic decision to exclude certain features due to their excessive missing values, which could compromise the integrity of the analysis. These features included variables such as restckm, pncaden, and earlobe, among others. Their high volume of nulls presents significant challenges for any robust statistical analysis, leading to their removal to maintain data quality.

In dealing with missing values within our dataset, I adopted a median imputation strategy. This approach was chosen due to its robustness against outliers, ensuring that our imputed values are not skewed by extreme, atypical data points. By replacing missing values with the median, we maintain the central tendency of the distribution for each variable, thus preserving the dataset's integrity for accurate analysis. This method was applied across all numerical features with missing data, thus enabling a more reliable and comprehensive evaluation of potential risk factors for heart disease, such as blood pressure and cholesterol levels.

The statistical correlation analysis yielded interesting findings. For instance, trestbps and trestbpd, which pertain to resting systolic and diastolic blood pressures, respectively, demonstrated a strong correlation coefficient of 0.915. This is a predictable outcome, given that systolic and diastolic readings are both components of overall blood pressure measurements. Furthermore, tpeakbps and tpeakbpd, related to peak blood pressures achieved, exhibited a similarly high correlation of 0.896. This insight aligns with medical expectations, as both readings are taken during the peak exercise phase of a stress test, therefore are expected to move in tandem.

Moreover, the correlation between thalach (maximum heart rate achieved) and thalrest (resting heart rate) was found to be 0.881, signifying a strong relationship between heart rate measurements taken under resting conditions versus those achieved during maximum exertion.

The Q-Q plot analysis for the variable trestbpd revealed a non-normal distribution with pronounced deviations at the tails, particularly on the lower end. This suggests the presence of outliers, or extreme values, which are more abundant than what would be anticipated in a normal distribution. The central tendency of the data, however, did align closely with the normal distribution's expected values, indicating a degree of normality amidst the outliers.

For graphical insights, the scatter plot analysis of trestbps against chol hinted at a potential positive relationship, albeit slightly dispersed. Both variables are acknowledged as key risk factors for heart disease, and the visualization supports this association.

Lastly, the scatter plot examining the relationship between age and cholesterol prominently showed a cluster of cases surfacing post-40 years of age, which correlates with established medical understanding that cholesterol concerns often intensify with advancing age.