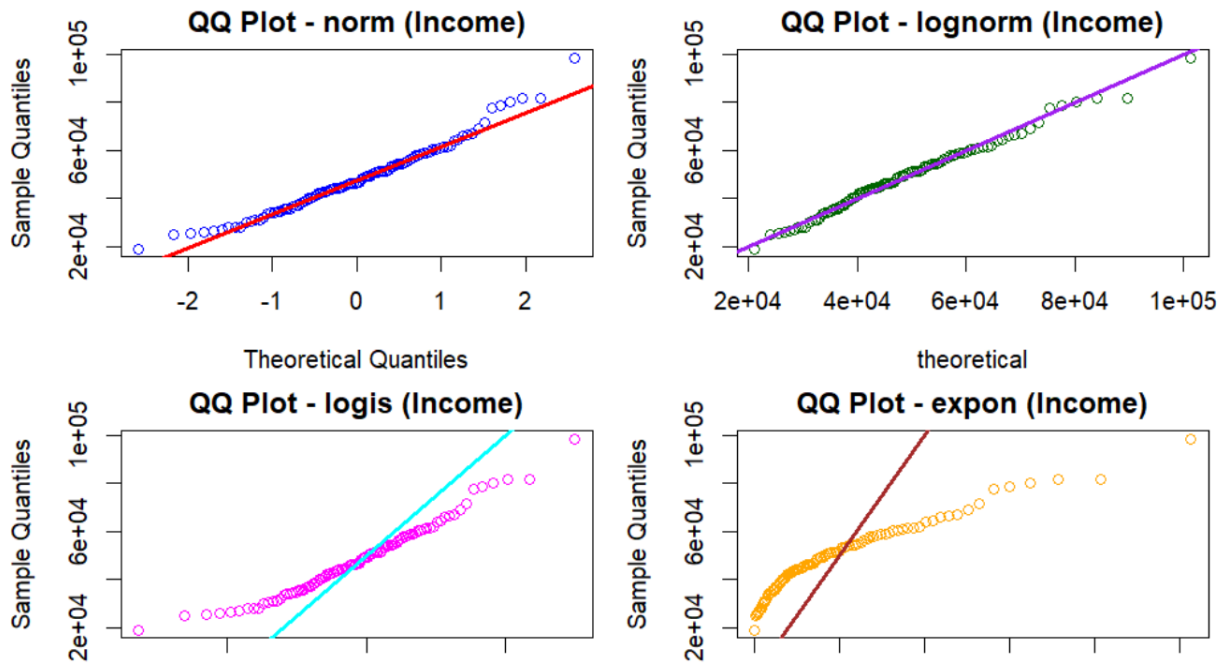
Analysis Report On The Best Fitting Distributions For Customer Data Variables

Given the visual data from the QQ plots below for the variables (**Income, Height, and CreditScore**) and considering the four different distributions (**normal, log-normal, logistic, and exponential**), I will analyze and provide the reasoning for selecting the best-fitting distribution for each variable. The analysis involves examining the alignment of the data points with the 45-degree line that represents the theoretical distribution.

1. Income Analysis



The QQ plot for the normal distribution concerning Income shows the data points closely aligned with the reference line, especially in the central quantiles. This indicates that the Income data is symmetrically distributed and adheres closely to a bell curve, which is characteristic of normal distributions.

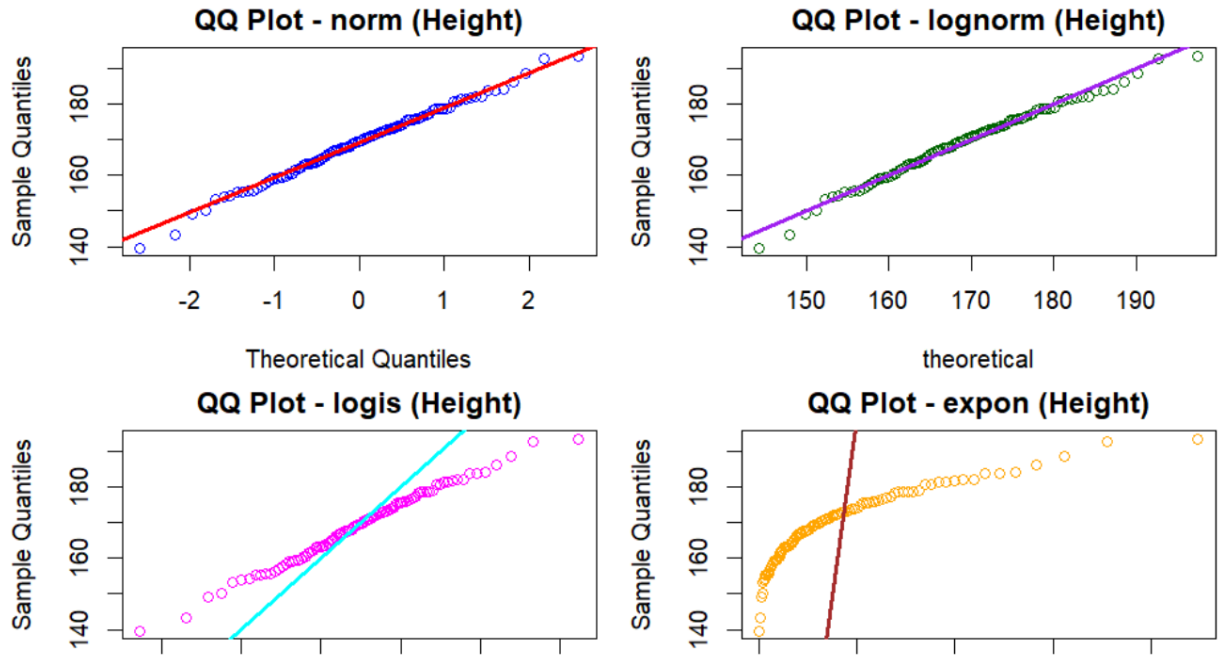
The log-normal distribution appears less fitting, as the QQ plot shows a significant divergence of the data points from the reference line in the higher quantiles, indicating a potential right skew in the data that the log-normal distribution is not capturing as accurately.

In the logistic distribution QQ plot, while the central quantiles follow the reference line, the tails, especially the upper tail, show some deviation. This suggests that while the logistic distribution captures the central tendency well, it may not fully account for the tail behaviour.

The exponential distribution is a poor fit for the Income data, as evidenced by the pronounced curvature of the data points in its QQ plot, indicating a significant right skew that is not characteristic of Income data which is usually not zero-bound.

Best Fit for Income: Normal Distribution

2. Height Analysis



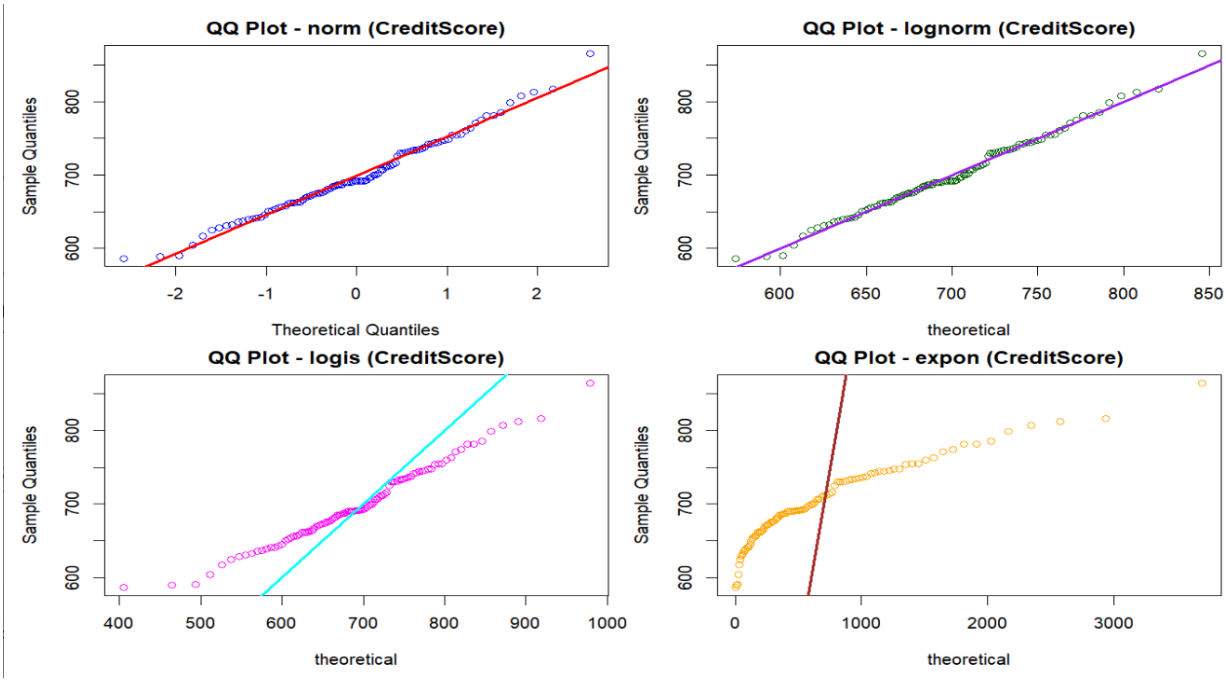
The QQ plot for the normal distribution regarding Height shows a close alignment of data points with the reference line throughout most of the distribution, suggesting that the Height data is normally distributed.

For the log-normal distribution, the QQ plot indicates a slight deviation in the upper quantiles, suggesting the presence of outliers or a slight positive skew in the Height data.

The logistic distribution QQ plot for Height also displays a good alignment with the reference line in the central quantiles but shows slight deviations at the tails.
The exponential distribution does not appear to be a good fit for the Height data, as the QQ plot shows a clear divergence from the reference line, which indicates that the exponential model does not adequately represent the distribution of Height.

Best Fit for Height: Normal Distribution

3. CreditScore Analysis



The normal distribution QQ plot for CreditScore reveals some deviations from the reference line, particularly in the lower and upper quantiles, suggesting the presence of outliers or heavy tails.

The log-normal distribution QQ plot shows a better alignment in the central quantiles but displays a significant deviation in the upper tail, indicating a skew that is not fully captured by this distribution.

The logistic distribution QQ plot demonstrates a better fit in the central region of the distribution compared to the normal distribution but, like the log-normal, shows a deviation in the tails.

The exponential distribution QQ plot shows a severe misalignment with the reference line, indicating that it is not a suitable model for CreditScore, which is not expected to be a zero-bound or heavily right-skewed metric.

Best Fit for CreditScore: Logistic Distribution

Conclusion

Based on the analysis of the QQ plots, the normal distribution provides the best fit for both Income and Height, as the data points align closely with the theoretical quantiles, especially within the interquartile range. For CreditScore, the logistic distribution appears to provide a slightly better fit than the normal distribution in the central quantiles while also accounting for the heavy tails. This analysis assumes that the central tendency and dispersion are the primary attributes of interest for determining the best fit.

Recommendations

While the QQ plots are a valuable tool for visual distribution assessment, it is recommended that additional statistical tests, such as the Kolmogorov-Smirnov test or the Shapiro-Wilk test for normality, be conducted to further validate these findings. Moreover, considering the real-world implications and the nature of the data is crucial, as the best statistical fit may not always align with the practical expectations or theoretical understanding of the variable in question.

Citation:

1. Poulson, B. (2023). R for Data Science: Analysis and Visualization [Video]. LinkedIn Learning. https://www.linkedin.com/learning/r-for-data-science-analysis-and-visualization.