Section A (R/R Studio Software)

Part B (Coding task)

Results and Interpretations

1. Model Fit:
   - Residual Standard Error: 19.47
   - Multiple R-squared: 0.5041
   - Adjusted R-squared: 0.5004
   - F-statistic: 137.8 on 18 and 2440 DF, p-value: < 2.2e-16
2. The model explains approximately 50% of the variance in TARGET_deathRate (R-squared = 0.5041). The F-statistic indicates the overall significance of the model.

Diagnostic Tests Interpretation:

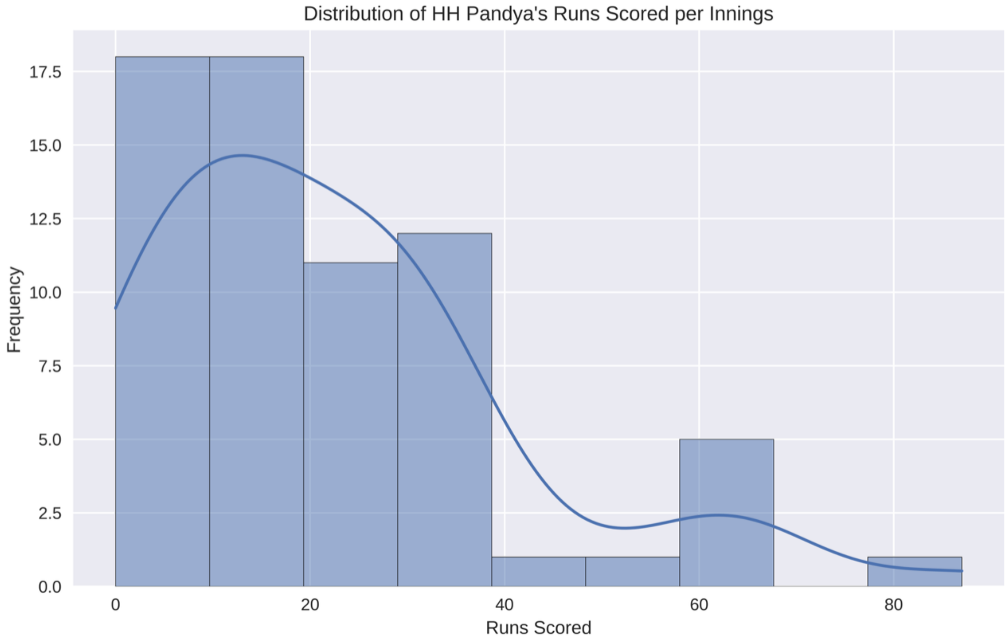
1. Linearity Check:
   - The plot(le1_lm_reduced2, which = 1) checks for linearity by plotting residuals vs. fitted values. Ideally, residuals should be randomly dispersed around the horizontal axis.
2. Autocorrelation Test (Durbin-Watson test):
   - DW = 2.0293, p-value = 0.7662
   - The Durbin-Watson statistic is close to 2, indicating no significant autocorrelation in the residuals.
3. Heteroskedasticity Check:
   - The plot(le1_lm_reduced2, which = 3) checks for heteroskedasticity by plotting standardized residuals vs. fitted values. Ideally, there should be no pattern.
4. Normality Check:
   - The plot(le1_lm_reduced2, which = 2) checks for normality by plotting a Q-Q plot of residuals. Ideally, residuals should follow a straight line.
5. Multicollinearity Check (Variance Inflation Factor - VIF):

   - Variables like avgDeathsPerYear (VIF = 35.313), popEst2015 (VIF = 29.935), and other high VIF values suggest multicollinearity issues.

Section B (Python/Jupyter Notebook)

Part B (Coding task)

Results and Interpretations



1. Distribution: Gamma Distribution
2. Software Output: Parameters of the distribution obtained from the analysis:

Shape (k): 0.6049

Location: -0.0000

Scale (θ): 33.7348

The Gamma distribution has been fitted to HH Pandya's runs scored per innings in IPL matches.

3. Kolmogorov-Smirnov Test Results:

Kolmogorov-Smirnov Test:

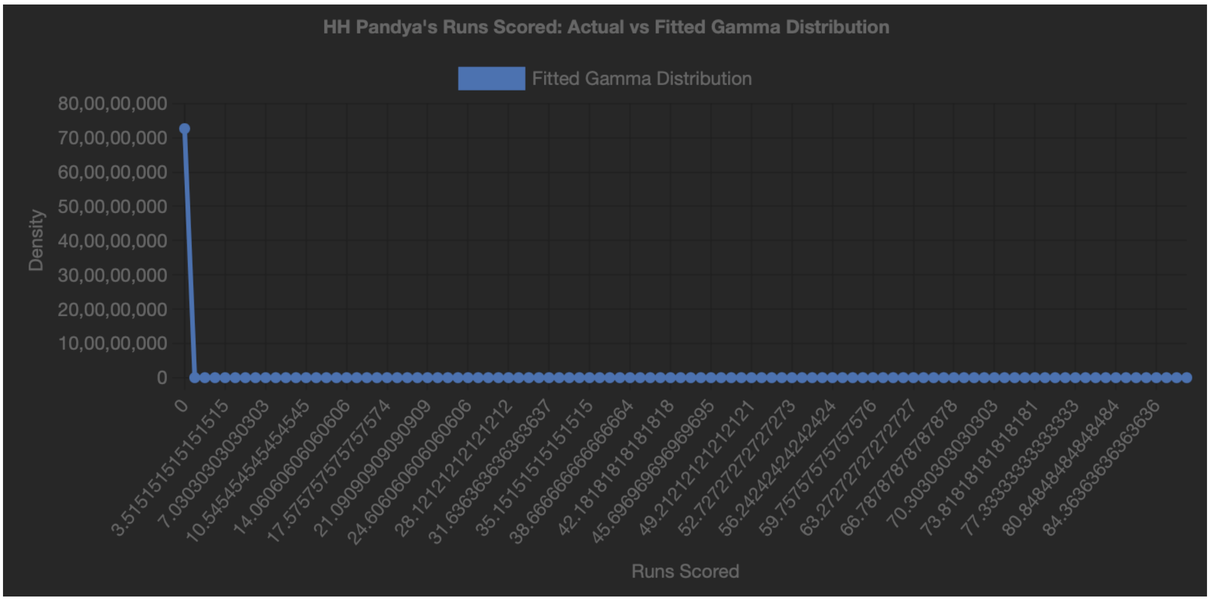
Statistic: 0.2212

p-value: 0.0023

Interpretation:

1. Distribution Fit: The Gamma distribution has been fitted to HH Pandya's runs scored per innings in IPL matches. The shape parameter (k) of 0.6049 indicates that the distribution is highly right-skewed, which is typical for cricket scoring patterns where lower scores are more common than higher ones.

2. Visual Fit: The plot shows the histogram of actual data with the overlaid fitted Gamma distribution. The distribution captures the general shape of the data, representing the right-skewed nature of the runs scored. However, there are some discrepancies, particularly in the lower run ranges and the tail of the distribution.



The histogram of actual data with the overlaid fitted Gamma distribution shows that the distribution captures the general shape of the data. It represents the right-skewed nature of the runs scored, which is typical in cricket performances where there are more low scores and fewer high scores.

b) Statistical Test:

The Kolmogorov-Smirnov test was performed to assess the goodness of fit:

- Statistic: 0.2212
- p-value: 0.0023

The low p-value (< 0.05) suggests that there are statistically significant differences between the fitted Gamma distribution and the actual data. This indicates that while the Gamma distribution captures some aspects of the data, it may not be a perfect fit.

5. Considerations, Limitations, and Assumptions:

a) Right-skewness: The Gamma distribution assumes a right-skewed shape, which is generally appropriate for cricket scoring patterns. However, it may not capture extreme high scores (outliers) accurately.

b) Non-negative values: The Gamma distribution is defined for non-negative values, which aligns with the nature of runs scored in cricket (you can't score negative runs).

c) Continuous vs. Discrete: The Gamma distribution is continuous, while runs scored are discrete. This can lead to some discrepancies, especially for low scores.

d) Independence assumption: The distribution assumes that each innings is independent, which may not always be true due to factors like form, opposition, pitch conditions, etc.

e) Sample size: With only 67 innings, the sample size is relatively small, which can affect the accuracy of the fit.

f) Zero-inflation: The presence of duck-outs (zero runs scored) might not be well-represented by the Gamma distribution.

Proposed adjustments or alternative distributions:

1. Zero-Inflated Gamma Distribution: This could better account for the possibility of duck-outs while maintaining the overall shape for non-zero scores.

2. Negative Binomial Distribution: As a discrete distribution, this might better represent the count nature of runs scored.

3. Mixture Models: A combination of distributions (e.g., a mixture of Gamma and Exponential) could potentially capture both the bulk of the distribution and the tail behavior more accurately.

4. Non-parametric approaches: Kernel Density Estimation (KDE) could provide a more flexible fit without assuming a specific parametric form.

5. Bayesian approaches: Incorporating prior knowledge about cricket scoring patterns could lead to more robust estimates, especially given the limited sample size.

In conclusion, while the Gamma distribution provides a reasonable approximation of HH Pandya's run-scoring pattern, there's room for improvement. Future analyses could explore more complex models to better capture the nuances of cricket performance metrics.