|  |
| --- |
| **Data Set: Regression Datasheet (Calories)** |
| **Question / Problem:**  • Perform a Multiple Linear Regression Analysis to determine a Regression Equation for Calories burned.  • Comment on the process and decisions made through the process |
| **Variables:**  Duration: Independent Variable, Numerical  Calories: Dependent Variable, Numerical  Pulse: Independent Variable, Numerical  Maxpulse: Independent Variable, Numerical |
| **Techniques used:**  Correlation Analysis:   * Checked the correlation between each independent variable and Calories Burned to determine relationships. * Used a correlation matrix to identify any strong multicollinearity among independent variables. * **Findings:** * Strong positive correlation between Duration and Calories burned (r=0.868r = 0.868). * Weak correlations between Pulse (r=−0.0034r = -0.0034) and Max Pulse (r=0.0797r = 0.0797) with Calories burned. * Moderate correlation between Pulse and Max Pulse, (r=0.7888r = 0.7888), indicating possible multicollinearity.   Multiple Linear Regression:   * Used Data Analysis -> Regression in Excel. * Input Calories Burned (Y) as the dependent variable. * Selected Duration, Pulse, and Max Pulse as independent variables. * Chose Residual Plots and Normal Probability Plots for diagnostic checking.   Diagnostics   * Checked R2R^2: Indicates how well the model explains the variability in Calories Burned. * Evaluated P-values: Retained variables with P<0.05P < 0.05. * Investigated Multicollinearity: Removed variables if collinearity caused instability. |
| **Analysis and Visualisation:**   * Generated a correlation matrix to observe the relationship between variables. * Expected to find a positive correlation between exercise duration and calories burned. * Checked for multicollinearity (if Pulse and Max Pulse were highly correlated**A table with numbers and a black line    AI-generated content may be incorrect.** * Strong positive correlation (0.868) between Duration and Calories Burned, suggesting that exercise duration is a strong predictor. * Weak correlation between Pulse (-0.0034) and Calories Burned, indicating Pulse may not significantly impact Calories. * Low correlation (0.0797) between Max Pulse and Calories, suggesting it may not be a strong predictor. * Moderate correlation (0.7888) between Pulse and Max Pulse, which could indicate multicollinearity (further checks required).   Used Excel’s Regression Tool with:   * Y Range (Dependent Variable): Calories Burned * X Range (Independent Variables): Duration, Pulse, Max Pulse * Selected Labels for clarity * Checked Residual Plots and Normal Probability Plot   A screenshot of a spreadsheet  AI-generated content may be incorrect.  A table of numbers and a number of calories  AI-generated content may be incorrect.  A graph with blue dots  AI-generated content may be incorrect.  A graph with blue dots  AI-generated content may be incorrect.  A graph with blue dots and numbers  AI-generated content may be incorrect.  A graph showing a normal probability plot  AI-generated content may be incorrect.  R² and Adjusted R²   * R² value indicates how well the independent variables explain Calories Burned. * A high R² (close to 1) means a strong predictive model.   Coefficients & P-values   * Significant Variables: Variables with P-values < 0.05 are retained. * Insignificant Variables: If P-value > 0.05, the variable may be removed, and regression is re-run.   Multicollinearity Check   * Since Pulse and Max Pulse have moderate correlation (0.7888), one may be removed if it causes instability in the model.   Skewness Check: SKEW = 0.549, which is slightly skewed but not significantly.  **A close up of numbers  AI-generated content may be incorrect.**    **Answer**  Duration has a strong positive correlation (r=0.868) and a low P-value with calories burned indicating it's a critical predictor.  Pulse and Max Pulse have weak or insignificant correlations, suggesting they may not impact the outcome significantly. |
| **Considerations:**   * Duration is a strong predictor (correlation = 0.868), so it should be included in the model. * Pulse and Max Pulse have weak correlations with Calories Burned and may be removed if their P-values are high. * Skewness (0.549) is acceptable, indicating a roughly normal distribution of residuals. * If Pulse and Max Pulse show multicollinearity, one may be removed to improve the model. * The Pulse Residual Plot is very clumped together so that diagram isn’t really a good reparation of the data. |