University of Toronto

Department of Statistical Sciences

Analysis of Factors Associated with Superconductor Critical Temperatures

STA302 — Methods of Data Analysis
Final Research Project

Group Members

Amir Koutahi, Canberk Soytekin, Andrew Hongyi Wu, Bruce Kaiyuan Chen

Contents

1 Introduction (305 words)

Electromagnetism is one of the four fundamental forces governing the universe and manifests itself in many aspects of daily life, including modern electronic devices. A major source of power consumption in electronics is electrical resistance in components, making the reduction of resistance critical for the development of efficient devices.

Superconducting materials exhibit zero electrical resistance. In principle, their use in electronics would allow devices to operate without power loss. In practice, however, superconductors lose this property above a critical temperature. As of February 2026, the highest known critical temperature is 134 K ($-139°$C), observed in the cuprate compound Hg1223 [1]. Consequently, the discovery of room-temperature superconductors remains a central goal of condensed matter physics and electronics manufacturing.

In this project, we address the following research question:

> *Which physical properties of cuprate-based superconductors are correlated with high critical temperatures?*

Recent advances in condensed matter physics have improved understanding of superconducting materials. Plakida [4] argues that cuprate compounds are currently the only materials that qualify as true high-temperature superconductors, as they are the only ones with critical temperatures above the boiling point of liquid nitrogen. Crépel *et al.* [2] suggest that attractive interactions between valence and conduction electrons can contribute to superconductivity, while Slebarski *et al.* [3] report that increased atomic disorder can enhance the critical temperature.

We propose to investigate the relationship between the critical temperatures of cuprate superconductors and their electronic and entropic properties. As of February 2026, no statistical analysis specifically addressing this relationship appears in the literature, leaving the research question open.

A secondary objective of this project is to assess whether multiple linear regression is an appropriate framework for this analysis. The response variable is continuous, and the predictors include both numerical and categorical variables. Since the primary goal is interpretability rather than prediction, linear regression is a suitable methodological choice.

2 Data Description (292 words)

The dataset was obtained from the **UC Irvine Machine Learning Repository** and was originally compiled from laboratory experiments in which superconducting materials were synthesized and their critical temperatures measured [5]. While the original study examines a broad range of superconductors, our analysis focuses specifically on cuprate-based compounds with high critical temperatures [5].

The response variable is the critical temperature (T_c), measured in Kelvin, which represents the threshold below which a material exhibits superconductivity. T_c is continuous, independent across observations, and suitable for linear regression. The empirical distribution of T_c is mildly right-skewed; therefore, a logarithmic transformation is applied to better satisfy normality assumptions.

To align the analysis with the research objective, the dataset was restricted to the top 1000 observations by critical temperature. The full dataset is dominated by low-T_c compounds, which can obscure relationships relevant to high-performance superconductors. This restriction improves interpretability while preserving meaningful variability among high-T_c materials.

Table 1 and Fig. 1 summarize the distributional properties of the predictors. Weighted valence entropy exhibits low dispersion (SD = 0.16) and minimal skewness, indicating similar valence structures across compounds. Atomic radius shows substantial spread (SD = 30.03) with little skewness, reflecting diversity in lattice structure. Electron affinity displays notable variability (SD = 14.23) and right skewness, indicating that strong electron-attraction properties are uncommon. Thermal conductivity is also right-skewed, with a small subset of materials exhibiting high heat transport efficiency. Valence electron range varies across discrete categories, capturing structural differences in valence shell composition relevant to superconducting behavior.

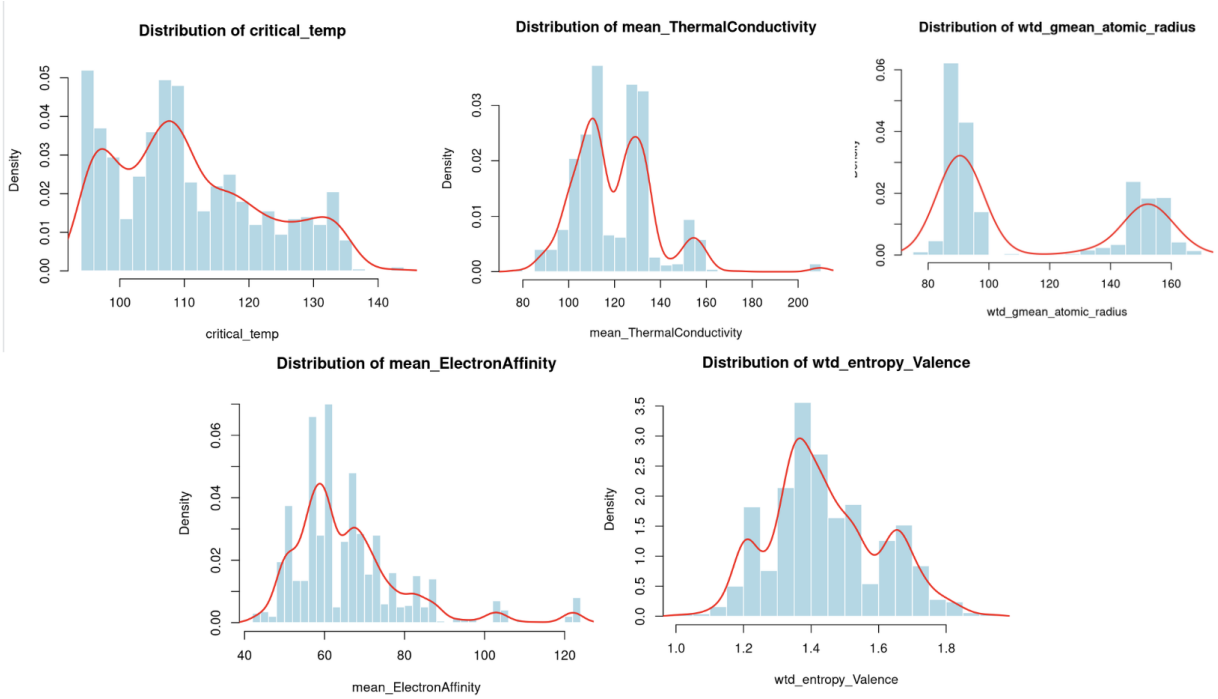


Figure 1: Empirical distributions of the response variable (critical temperature) and selected predictor variables used in the regression analysis.

Table 1: Predictor variables used in the preliminary regression model

Variable	**Description**
Weighted Valence Entropy	Valence electron entropy of constituent elements, weighted by relative proportions, reflecting valence electronic complexity.
Mean Electron Affinity	Average electron affinity of constituent elements, characterizing electronic structure relevant to superconductivity.
Weighted Geometric Mean Atomic Radius	Weighted geometric mean of atomic radii, capturing lattice-scale effects that scale multiplicatively.
Mean Thermal Conductivity	Average thermal conductivity of constituent elements, reflecting heat transport properties.
Valence Electron Range (categorical)	Range of valence electron counts among constituent elements, treated as categorical due to discreteness and nonlinearity.

Table 2: Summary statistics for response and predictor variables used in the preliminary model.

Variable	Mean	SD	Min	Max
Entropy Valence	1.44	0.16	1.01	1.94
Electron Affinity	65.86	14.23	43.63	122.24
Atomic Radius	113.37	30.03	75.56	168.35
Thermal Conductivity	120.18	17.88	77.00	209.61
Valence Electron Range	NA	NA	1	5

3 Preliminary Model Discussion (312 words)

A preliminary multiple linear regression model was fitted using the log-transformed critical temperature, $\log(T_c + 1)$, as the response variable. The model includes four standardized numerical predictors and the categorical predictor `range_Valence`. The fitted model is

$$\begin{aligned}\log(T_c + 1) = \beta_0 &+ \beta_1(\texttt{wtd\_entropy\_Valence}) + \beta_2(\texttt{mean\_ElectronAffinity}) \\ &+ \beta_3(\texttt{wtd\_gmean\_atomic\_radius}) + \beta_4(\texttt{mean\_ThermalConductivity}) \\ &+ \sum_{k=1}^{5} \gamma_k \mathbb{1}\{\texttt{range\_Valence} = k\} + \varepsilon.\end{aligned}$$

The estimated coefficients suggest that higher weighted valence entropy ($\hat{\beta}_1 = 0.063$) and mean thermal conductivity ($\hat{\beta}_4 = 0.022$) are associated with higher values of $\log(T_c + 1)$, while mean electron affinity ($\hat{\beta}_2 = -0.011$) and weighted geometric mean atomic radius ($\hat{\beta}_3 = -0.025$) are associated with lower values of the response. Relative to the reference category of `range_Valence`, most valence-range levels exhibit negative estimated shifts, indicating systematically lower critical temperatures across several electronic regimes, while one level shows a small positive shift. These trends are consistent with the data-driven findings reported in [5].

Table 3: Estimated coefficients for the model.

Term	Estimate
`wtd_entropy_Valence`	0.063
`mean_ElectronAffinity`	−0.011
`wtd_gmean_atomic_radius`	−0.025
`mean_ThermalConductivity`	0.022
`range_Valence1`	−0.052
`range_Valence2`	−0.058
`range_Valence3`	−0.112
`range_Valence4`	−0.113
`range_Valence5`	0.027

Based on Figure 2, the normal Q–Q plot shows residuals that closely follow the reference line, with only minor deviations in the upper tail, indicating that the normality assumption is reasonably satisfied. The residuals versus fitted values and standardized residuals versus fitted values plots do not display a funnel pattern, suggesting that the constant variance assumption is met. However, the smooth trend in the residuals versus fitted values plot exhibits a slight tilt away from zero in parts of the fitted range. In addition, the

presence of a symmetric linear pattern in the lower residuals suggests a mild departure from the linearity assumption.

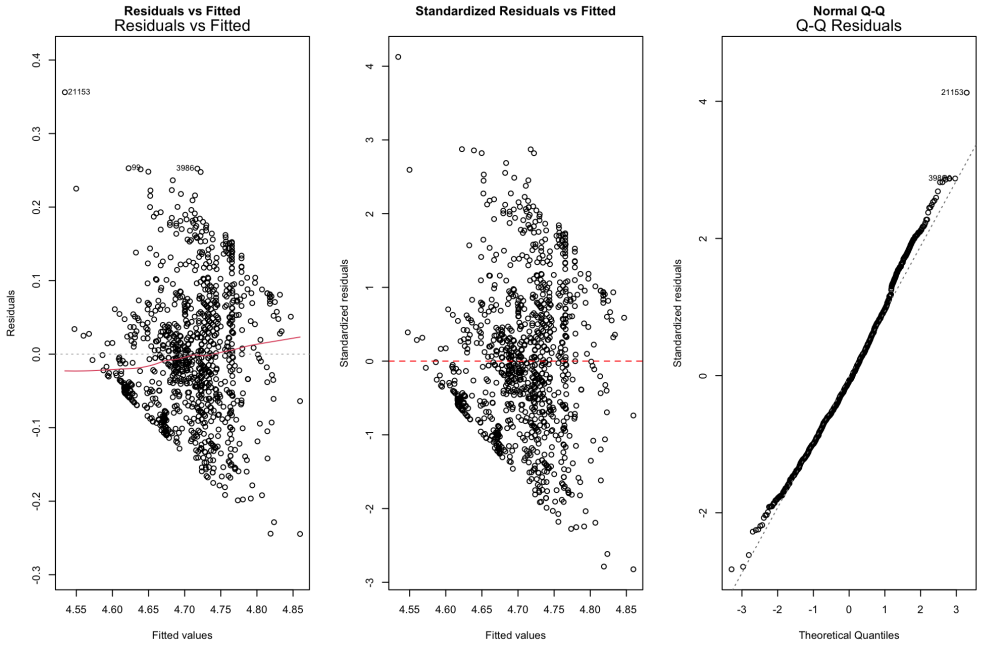


Figure 2: Residual diagnostic plots for the regression model.

4 Model (1011 words)

To identify an appropriate regression model for the critical temperatures of high-T_c cuprate superconductors, we considered three main issues: whether the response should be transformed, whether any observations exerted undue influence on the fitted model, and whether any predictors should be excluded. The final selected model uses the transformed response

$$\log(T_c + 1)$$

and includes all candidate predictors: `wtd_entropy_Valence`, `mean_ElectronAffinity`, `wtd_gmean_atomic_radius`, `mean_ThermalConductivity`, and `range_Valence`.

Choice of Transformation

We first compared a model with the raw response `critical_temp` to a model with the transformed response $\log(T_c + 1)$. The motivation for this transformation was that critical temperature is strictly positive and exhibited right-skewness in the exploratory analysis. A logarithmic transformation is commonly used in this setting because it can reduce skewness, stabilize variance, and improve the approximation to normality required by linear regression.

The numerical comparison between the two models supported the transformation. The raw-response model had adjusted $R^2 = 0.1829$ and AIC $= 6036.811$, whereas the log-response model had adjusted $R^2 = 0.1885$ and AIC $= -1661.363$. Since a smaller AIC indicates a better tradeoff between goodness of fit and

model complexity, the transformed model is strongly preferred. The slight increase in adjusted R^2 also suggests that the predictors explain somewhat more of the variability on the log scale than on the original scale.

Diagnostic plots likewise favored the transformed model. Relative to the raw-response fit, the log-response model showed a more stable spread of residuals and better alignment with the normal Q–Q line. For these reasons, we retained $\log(T_c + 1)$ as the response throughout the remainder of the analysis.

Full Candidate Model

The full candidate model was

$$\begin{aligned}\log(T_c + 1) = \beta_0 &+ \beta_1(\texttt{wtd\_entropy\_Valence}) + \beta_2(\texttt{mean\_ElectronAffinity}) \\ &+ \beta_3(\texttt{wtd\_gmean\_atomic\_radius}) + \beta_4(\texttt{mean\_ThermalConductivity}) \\ &+ \gamma_2 I(\texttt{range\_Valence} = 2) + \gamma_3 I(\texttt{range\_Valence} = 3) \\ &+ \gamma_4 I(\texttt{range\_Valence} = 4) + \gamma_5 I(\texttt{range\_Valence} = 5) + \varepsilon,\end{aligned}$$

where $\texttt{range\_Valence} = 1$ is the reference category.

The fitted model was highly significant overall:

$$F(8, 804) = 24.58, \qquad p < 2.2 \times 10^{-16},$$

with

$$R^2 = 0.1965, \qquad R^2_{\text{adj}} = 0.1885.$$

Thus, the model explains a modest but nontrivial portion of the variation in transformed critical temperature.

Testing Whether Predictors Should Be Excluded

To decide whether any predictors should be removed, we used both partial F-tests and AIC-based model selection. The partial F-tests compare the full model to reduced models obtained by removing one term at a time. The results are shown in Table 4.

Table 4: Partial F-tests comparing the full model to reduced models with one term removed.

Removed term	df	Sum of squares	F-value	p-value
`wtd_entropy_Valence`	1	0.676671	90.4026	< 0.000001
`mean_ElectronAffinity`	1	0.046668	6.2347	0.012726
`wtd_gmean_atomic_radius`	1	0.131309	17.5428	0.000031
`mean_ThermalConductivity`	1	0.208963	27.9172	< 0.000001
`range_Valence`	4	0.630236	21.0497	< 0.000001

All five terms were significant at the 5% level when tested against the full model. In particular, `wtd_entropy_Valence` was the most important numerical predictor, while the categorical factor `range_Valence` was also strongly significant as a block. Although one individual indicator level, `range_Valence2`, was not significant in the coefficient table, the factor as a whole was highly significant. Therefore, it would not be appropriate to remove the entire categorical predictor on the basis of a single nonsignificant level.

We also used backward AIC model selection beginning from the full model. The stepwise procedure retained the full model and removed no predictors. Table 5 compares the full model to several reduced alternatives.

Table 5: Comparison of candidate models. Lower AIC and BIC indicate better fit after accounting for model complexity.

Model	AIC	BIC	R^2	Adjusted R^2
Full model	-1661.363	-1614.356	0.1965	0.1885
No entropy	-1576.733	-1534.426	0.1062	0.0984
No affinity	-1657.083	-1614.776	0.1903	0.1832
No radius	-1645.815	-1603.508	0.1790	0.1718
No thermal	-1635.613	-1593.306	0.1686	0.1614
No range	-1588.391	-1560.187	0.1124	0.1080
AIC selected	-1661.363	-1614.356	0.1965	0.1885

The full model had the smallest AIC and BIC among all candidates, and it also had the largest R^2 and adjusted R^2. This provides further support for retaining all predictors.

Multicollinearity

Before finalizing the model, we checked for collinearity among predictors. The largest pairwise correlation among the numerical predictors was between `wtd_entropy_Valence` and `wtd_gmean_atomic_radius`, with correlation 0.707. This is moderately large, but the VIF diagnostics did not suggest serious multicollinearity. The adjusted GVIF values were

$$1.7146,\ 1.0785,\ 1.4878,\ 1.0708,\ 1.0654$$

for `wtd_entropy_Valence`, `mean_ElectronAffinity`, `wtd_gmean_atomic_radius`, `mean_ThermalConductivity`, and `range_Valence`, respectively. These are all small and well below common concern thresholds, so multicollinearity does not justify excluding any predictors.

Outliers, Leverage, and Influential Observations

We next investigated whether any observations exerted undue influence on the fitted model. We used four standard diagnostics: studentized residuals, leverage, Cook's distance, and DFFITS. The cutoffs used were

$$h_i > \frac{2p}{n} = 0.02214, \qquad D_i > \frac{4}{n} = 0.00492, \qquad |\text{DFFITS}| > 2\sqrt{\frac{p}{n}} = 0.21043,$$

with $n = 813$ observations and p equal to the number of estimated coefficients.

A total of 116 observations were flagged by at least one of these diagnostics. Table 6 summarizes the counts by diagnostic type.

Table 6: Counts of flagged observations by influence diagnostic.

Diagnostic	Count
Outlier flag ($\lvert r_i^* \rvert > 3$)	1
High leverage	85
Cook's distance flag	46
DFFITS flag	46

The most influential observations included indices 42, 637, 198, 805, 643, 724, 725, 740, 717, 709, 752, 674, 713, 774, and 32. Observation 42 was the most extreme, with studentized residual 3.912, leverage 0.0483, Cook's distance 0.0849, and DFFITS 0.8818. This point is a clear outlier candidate and also has high leverage.

To assess the effect of these influential points, we re-fit the model after removing the 15 observations that satisfied the strongest combination of influence criteria. The results are summarized in Table 7.

Table 7: Sensitivity analysis comparing the original full model with the model refit after removing the strongest flagged observations.

Model	n	AIC	BIC	R^2	Adjusted R^2	Residual SE
Original full model	813	-1661.363	-1614.356	0.1965	0.1885	0.0865
Refit after removal	798	-1695.070	-1648.249	0.2472	0.2396	0.0831

The refitted model showed somewhat improved fit, but the signs and significance of the main coefficients remained essentially unchanged. In particular, `wtd_entropy_Valence` and `mean_ThermalConductivity` remained positive and significant, while `mean_ElectronAffinity` and `wtd_gmean_atomic_radius` remained negative and significant. The factor `range_Valence` also remained important. Since the substantive conclusions were stable, these points do not appear to be creating the main scientific conclusions artificially. Moreover, because the analysis specifically targets high-T_c cuprates, extreme observations may be scientifically meaningful rather than erroneous. For that reason, we retained the original full dataset for the primary analysis and treated the refit only as a sensitivity check.

Final Model Choice

The final selected model is the full regression on the transformed response $\log(T_c+1)$. The log transformation was chosen because it improved residual behavior and produced a much smaller AIC than the raw-response model. No predictors were excluded, since all were supported by partial F-tests and the full model was also preferred by AIC and BIC. Influential observations were identified and examined, but because the main results were robust to their removal, they were not omitted from the primary analysis.

5 Final Model Inference and Results (1152 words)

The final model selected in Section 4 was

$$\begin{aligned}\log(T_c+1) = \beta_0 &+ \beta_1(\texttt{wtd\_entropy\_Valence}) + \beta_2(\texttt{mean\_ElectronAffinity}) \\ &+ \beta_3(\texttt{wtd\_gmean\_atomic\_radius}) + \beta_4(\texttt{mean\_ThermalConductivity}) \\ &+ \gamma_2 I(\texttt{range\_Valence} = 2) + \gamma_3 I(\texttt{range\_Valence} = 3) \\ &+ \gamma_4 I(\texttt{range\_Valence} = 4) + \gamma_5 I(\texttt{range\_Valence} = 5) + \varepsilon,\end{aligned}$$

where $\texttt{range\_Valence} = 1$ serves as the reference category.

Coefficient Table

Table 8 summarizes the estimated coefficients, standard errors, 95% confidence intervals, and p-values for the final model.

Table 8: Final model coefficient estimates.

Term	Estimate	Std. Error	95% CI	p-value
Intercept	4.254707	0.051399	[4.153815, 4.355598]	< 0.000001
wtd_entropy_Valence	0.314695	0.033098	[0.249726, 0.379663]	< 0.000001
mean_ElectronAffinity	-0.000538	0.000215	[-0.000961, -0.000115]	0.012726
wtd_gmean_atomic_radius	-0.000620	0.000148	[-0.000910, -0.000329]	0.000031
mean_ThermalConductivity	0.000948	0.000179	[0.000596, 0.001300]	< 0.000001
range_Valence2	-0.002105	0.009083	[-0.019933, 0.015724]	0.816827
range_Valence3	-0.053161	0.008563	[-0.069969, -0.036353]	< 0.000001
range_Valence4	-0.053742	0.012809	[-0.078885, -0.028599]	0.000030
range_Valence5	0.082411	0.017009	[0.049023, 0.115799]	0.000002

Interpretation of Coefficients

Because the response is $\log(T_c + 1)$, each coefficient describes the change in the log of $(T_c + 1)$ associated with a one-unit increase in the corresponding predictor, holding the other predictors fixed. On the original scale, exponentiating a coefficient gives the multiplicative change in $(T_c + 1)$.

The intercept estimate is 4.2547. This corresponds to the expected value of $\log(T_c+1)$ when all numerical predictors are zero and $\texttt{range\_Valence} = 1$. Since zero is not necessarily a meaningful value for all predictors, the intercept is mainly a baseline needed for the linear specification rather than a quantity of direct scientific interest.

The coefficient for wtd_entropy_Valence is 0.3147 and is highly significant ($p < 0.000001$). This is the largest positive numerical coefficient in the model. Exponentiating the estimate gives

$$e^{0.3147} \approx 1.3698,$$

suggesting that a one-unit increase in weighted valence entropy is associated with an approximate 36.98% increase in $(T_c + 1)$, holding all else fixed. This positive relationship indicates that greater valence-electron complexity is associated with higher critical temperatures among the cuprate compounds in this sample.

The coefficient for `mean_ElectronAffinity` is -0.000538 ($p = 0.0127$). Although the magnitude is small, the effect is statistically significant. A one-unit increase in mean electron affinity is associated with a decrease in $\log(T_c + 1)$, corresponding to a multiplicative factor of

$$e^{-0.000538} \approx 0.999462.$$

Thus, higher electron affinity is associated with slightly lower critical temperatures after controlling for the other variables.

The coefficient for `wtd_gmean_atomic_radius` is -0.000620 ($p = 0.000031$). This indicates that larger weighted geometric mean atomic radius is associated with lower $\log(T_c + 1)$, holding other predictors fixed. The corresponding multiplicative factor is

$$e^{-0.000620} \approx 0.999380,$$

so the effect per one-unit increase is small in size but statistically reliable.

The coefficient for `mean_ThermalConductivity` is 0.000948 and is highly significant ($p < 0.000001$). This positive coefficient indicates that materials with larger mean thermal conductivity tend to have higher critical temperatures, all else equal. On the original scale,

$$e^{0.000948} \approx 1.000949,$$

which corresponds to an increase of about 0.095% in $(T_c + 1)$ per one-unit increase in thermal conductivity.

The categorical predictor `range_Valence` is interpreted relative to the reference category `range_Valence` = 1. The estimated coefficient for `range_Valence2` is -0.0021 and is not significant ($p = 0.8168$), so there is no evidence that this level differs from the reference category. In contrast, `range_Valence3` and `range_Valence4` have significantly negative coefficients, -0.0532 and -0.0537, respectively. Their multiplicative effects are

$$e^{-0.0532} \approx 0.9482, \qquad e^{-0.0537} \approx 0.9477,$$

indicating that, relative to the reference category, these levels are associated with approximately 5.18% and 5.23% lower values of (T_c+1), holding the other predictors fixed. By contrast, `range_Valence5` has a positive coefficient 0.0824 ($p = 0.000002$), corresponding to

$$e^{0.0824} \approx 1.0859,$$

or about an 8.59% increase in $(T_c + 1)$ relative to the reference category.

Overall, the coefficient pattern suggests that higher critical temperature in this subset of cuprate superconductors is associated with larger weighted valence entropy and greater mean thermal conductivity, while larger electron affinity and atomic radius are associated with lower critical temperature. The categorical valence-range structure also appears important: some valence ranges are associated with systematically lower critical temperatures, whereas the highest range category considered here is associated with higher values.

What the Results Suggest About the Research Question

The research question asked which physical properties of cuprate-based superconductors are associated with high critical temperatures. The final model suggests that several properties are statistically associated with $\log(T_c + 1)$, and hence with critical temperature itself. Among the numerical predictors, weighted valence entropy emerges as the strongest positive associate, while mean thermal conductivity also shows a positive contribution. In contrast, mean electron affinity and weighted geometric mean atomic radius are negatively associated with critical temperature.

These results suggest that both electronic complexity and structural features matter. Since the model is observational, these findings should be interpreted as conditional associations rather than causal effects. Nevertheless, the regression does provide evidence that the selected electronic and lattice-related variables are informative for distinguishing higher-T_c cuprate compounds within this dataset.

Comparison of Results with Findings from the Literature

The results obtained from our statistical analysis are statistically significant and in line with the findings from the literature. As discussed, Crépel *et al.* [2] suggest that attractive interactions between valence and conduction electrons can contribute to superconductivity, while Slebarski *et al.* [3] report that increased atomic disorder can enhance the critical temperature.

Critical temperatures being 8.59% higher on average (relative to the reference category) in compounds with a valence range of 5 suggests that highly localized $\ell = 1$ electrons in s orbitals may be hybridizing with the easily ionizable (and potentially conducting) $\ell = 5$ valence electrons in the compounds. This would contribute to increased critical temperatures via exotic quantum electronic interactions, in line with the findings of Crépel *et al.* [2].

Furthermore, an average 36.98% increase in (T_c+1) per one-unit increase in the weighted valence entropy implies that increased atomic disorder can significantly enhance the critical temperatures observed in cuprate superconductors, confirming the findings reported in Slebarski *et al.* [3].

Model Performance

The overall performance of the final model is summarized in Table 9.

Table 9: Final model performance metrics.

Metric	Value
R^2	0.196501
Adjusted R^2	0.188506
Residual standard error	0.086516
AIC	-1661.363335
BIC	-1614.356024
F-statistic	24.577900
Model p-value	< 0.000001
Sample size	813

The R^2 value indicates that the model explains about 19.65% of the variation in $\log(T_c + 1)$, and the adjusted R^2 is similar at 18.85%. This means that the included predictors capture a meaningful but limited

portion of the variability in critical temperature. This is not unexpected, since superconductivity is governed by complex chemical, crystallographic, and quantum-mechanical mechanisms, many of which are not represented in the current model.

The residual standard error of 0.0865 indicates that predictions on the log scale are typically within that distance of the observed values. The very small model p-value confirms that the predictors are jointly associated with the response. Finally, the strongly negative AIC and BIC values are primarily useful for comparing candidate models; within the candidate set considered here, the final model had the smallest AIC and BIC and was therefore preferred.

In summary, the final model is statistically significant, reasonably interpretable, and clearly preferable to the reduced alternatives considered during model selection. Its main limitation is not lack of significance, but rather that a substantial proportion of variability remains unexplained, indicating that additional physical descriptors would likely be needed for a more complete account of superconducting critical temperature.

6 Discussion and Conclusion (1003 words)

Conclusion

We investigated the electronic and entropic physical properties associated with high critical temperatures in cuprate-based conductors. To that end, we conducted the analyses required for statistical inference on the *Superconductivity Data* dataset obtained from the **UC Irvine Machine Learning Repository**. To align the analysis with the research objective, observations corresponding to 1000 cuprate compounds with the highest critical temperatures were used in the project.

The physical properties whose relationship with the critical temperature was investigated were mean thermal conductivity, weighted geometric mean of the atomic radii, mean electron affinity, weighted valence entropy and the valence range, with the valence range taking integer values between 1 through 5 and being treated as a categorical variable.

The appropriate linear regression model was identified to be the multiple regression model consisting of the regression terms for each predictor variable, where the response variable (critical temperature) was log-transformed to reduce skewness, stabilize variance, and improve the approximation to normality required by linear regression. The final candidate model was observed to be highly statistically significant, with a p-value smaller than 2.2×10^{-16}. Moreover, the model was determined to explain a modest but non-trivial portion of the variation in the log-transformed critical temperatures, with its R^2 and R^2_{adj} scores being 0.1965 and 0.1885, respectively.

The predictor variables were also determined to be highly statistically significant by means of partial F-tests and AIC-based model selection methods. All predictors were observed to be significant at the $\alpha = 0.05$ level when tested against the complete model using partial F-tests. Furthermore, the AIC and BIC scores for the complete model (-1661.363 and -1614.356, respectively) were determined to be the smallest among all candidates. In addition to the variables being statistically significant, no serious multicollinearity was observed between the predictor variables as a result of the VIF diagnostics.

The model was observed to be robust to influential points. Two trials were carried out to assess the effect of influential points on the model. In one model, 15 observations that satisfied the strongest combination of influence criteria (leverage, Cook's distance, and DFFITS metrics) were removed from the dataset, while the other model was fit to all observations to serve as a control. The model that had the 15 influential points

removed from its dataset showed a marginally improved fit based on its AIC, BIC, R^2, and R^2_{adj} metrics. However, the signs and the significance of the main coefficients were observed to be practically unchanged. Therefore, the influential points were kept in the dataset, owing to the fact that extreme observations can be meaningful in a scientific context and that these points did not appear to have a large influence on the main conclusions drawn from the model. In conclusion, a statistically significant, scientifically-grounded, and interpretable model was successfully fit to the superconductivity data obtained from the UC Irvine Machine Learning Repository.

Discussion of Key Findings

As discussed, all predictor variables were determined to have a statistically significant relationship with the critical temperature at the $\alpha = 0.05$ significance level. In terms of magnitude, the weighted valence entropy (`wtd_entropy_Valence`) and the valence range (`valence_Range`) were determined to have the largest effect on the critical temperature. A one-unit increase in the weighted valence entropy was observed to correspond to a 36.98% increase in the critical temperature on average. Likewise, critical temperatures were 8.59% higher on average (relative to the reference category) in compounds with a valence range of 5. Findings from the literature report that increased atomic disorder and attractive interactions between valence and conduction electrons can enhance the critical temperature, agreeing with the key findings of our project.

The analysis of the other predictor variables yielded statistically significant, yet physically trivial results. One-unit increases in the mean electron affinity, weighted geometric mean of the atomic radii, and mean thermal conductivity were observed to correspond to -0.0538%, -0.062%, and 0.095% changes in the critical temperature, respectively. Despite these changes being statistically significant, they are unlikely to be substantial in the context of physical research.

Suggestions and Recommendations

The R^2 and R^2_{adj} metrics for the final model (0.1965 and 0.1885, respectively) indicate that the model explains only a limited proportion of the total variability in the critical temperature. While this may initially appear to be a weakness, it is consistent with the underlying physics of superconductivity, which is governed by complex, highly non-linear quantum mechanical interactions between electrons, lattice vibrations, and material-specific structural properties. As such, a purely linear modeling framework is unlikely to fully capture the richness of these relationships.

To improve the explanatory and predictive capability of the model, several extensions can be considered. First, the inclusion of non-linear transformations of the predictor variables (such as polynomial terms or logarithmic transformations beyond the response variable) may better capture curvature and threshold effects present in the data. Second, the incorporation of interaction terms between predictors could allow the model to reflect coupled physical mechanisms, such as the interplay between atomic structure and electronic properties.

In addition, expanding the set of predictor variables may provide further insight. Variables that more directly characterize quantum or crystallographic properties—such as lattice structure parameters, electron density measures, or phonon-related features—could be particularly valuable. The current dataset, while informative, may not fully represent all relevant physical drivers of superconductivity.

Alternative modeling approaches should also be considered. Methods such as generalized additive models (GAMs), tree-based models, or other machine learning techniques may be more suitable for capturing com-

plex, non-linear relationships without requiring explicit specification of functional forms. These approaches could complement the current linear regression analysis by improving predictive performance while still allowing for partial interpretability.

Finally, increasing the size and diversity of the dataset would likely enhance model performance. A larger dataset with broader coverage of material types and physical properties would reduce sampling limitations and allow more robust estimation of relationships. Overall, while the current model provides statistically significant and interpretable results, there remains substantial opportunity for refinement through more advanced modeling techniques and richer data.

Potential Improvements in Analyses

Several improvements could strengthen the analysis in future work.

First, although restricting the data to the top 1000 compounds by critical temperature made the study more focused on high-T_c cuprates, this choice may affect generalizability. It would be useful to compare the current results with models fit to alternative subsets, or to the full cuprate dataset, to check whether the conclusions remain stable.

Second, the current model assumes linear relationships between the predictors and $\log(T_c + 1)$. Since the diagnostic plots suggest mild departures from linearity, future analyses could include polynomial terms, interaction terms, or spline-based methods to better capture more complex structure in the data.

Third, the model includes only a limited number of predictors, which likely contributes to the modest R^2. Adding variables such as crystallographic, compositional, or doping-related descriptors could improve the explanatory power of the model and provide a more complete picture of the factors associated with superconducting critical temperature.

Finally, model assessment could be strengthened by adding cross-validation or robust regression methods. These would help evaluate how well the model generalizes and whether the results are sensitive to unusual observations. Since the analysis is observational, future work should also be careful to distinguish statistical association from physical causation.

7 Author Contributions

Part I

- **Amir Koutahi** — Data preprocessing and model construction.
- **Canberk Soytekin** — Introduction, literature review, bibliography, and formulation of the research question.
- **Andrew Wu** — Data description.
- **Bruce Chen** — Data description.

Part II

- **Amir Koutahi** — Model development Section 4 and R Markdown implementation.
- **Canberk Soytekin** — Discussion and conclusion Section 6 and results part of Section 5.
- **Andrew Wu** — Final model inference, results interpretation Section 5, and revisions.
- **Bruce Chen** — Final model inference, results interpretation Section 5, and revisions.

Disclaimer: Word counts for each section were computed using the online tool wordcounter.net. Numerical values appearing in tables were excluded from the reported word counts.

References

[1] M. Stephens, "Characterizing a top superconductor," *Physics*, vol. 18, p. 90, 2025.

[2] V. Crépel, T. Cea, L. Fu, and F. Guinea, "Unconventional superconductivity due to interband polarization," *Phys. Rev. B*, vol. 105, p. 094506, 2022.

[3] A. Slebarski *et al.*, "Enhancing superconductivity of $Y_5Rh_6Sn_{18}$ by atomic disorder," *Phys. Rev. B*, vol. 102, p. 054514, 2020.

[4] N. Plakida, *High-Temperature Cuprate Superconductors*. Berlin, Germany: Springer, 2010, p. 6.

[5] K. Hamidieh, "A data-driven statistical model for predicting the critical temperature of a superconductor," *Computational Materials Science*, vol. 154, pp. 346–354, 2018.

Analysis of Factors Associated with Superconductor Critical Temperatures

Amir Koutahi, Canberk Soytekin, Andrew Hongyi Wu, Bruce Kaiyuan Chen

Description

This R Markdown document presents a complete regression-based statistical analysis of superconducting materials using the dataset `Cleaned_Data.csv`. The purpose of the analysis is to investigate how selected material characteristics are associated with superconducting critical temperature, with particular attention to compounds containing both copper (Cu) and oxygen (O), since these are especially relevant in the study of high-temperature superconductors. The document is written to provide a full workflow, beginning with data preparation and ending with model interpretation and diagnostic assessment.

The analysis begins by importing the dataset and restricting attention to observations whose chemical formula information includes both Cu and O. To focus the study on the most relevant high-temperature materials, the data are then ordered by decreasing critical temperature and only the top 1000 observations are retained. After this filtering step, the variable `range_Valence` is converted to a categorical factor with levels 1 through 5, and a transformed response variable, `log_Tc = log(critical_temp + 1)`, is created. This log transformation is introduced to reduce skewness, stabilize variance, and improve the suitability of the data for linear modeling.

Once the data are prepared, the document constructs a reduced working dataset containing the response and predictor variables of interest: `critical_temp`, `log_Tc`, `wtd_entropy_Valence`, `mean_ElectronAffinity`, `wtd_gmean_atomic_radius`, `mean_ThermalConductivity`, and `range_Valence`. Missing observations are removed so that all subsequent analyses are performed on a complete-case dataset. The script also reports the final number of observations used, ensuring transparency about the effective sample size.

The next stage of the document provides descriptive statistics for the main quantitative variables, including means, standard deviations, minima, and maxima. These summaries give an overall numerical picture of the filtered sample before modeling begins. In addition, the script visually compares the distribution of the original response variable `critical_temp` with the transformed response `log_Tc` using histograms and normal Q-Q plots. This step is important because it motivates whether a transformation of the response leads to a more appropriate modeling framework.

The core modeling section compares two multiple linear regression models: one using the raw response `critical_temp`, and another using the transformed response `log_Tc`. Both models include the same explanatory variables, namely weighted entropy of valence, mean electron affinity, weighted geometric mean atomic radius, mean thermal conductivity, and the categorical variable `range_Valence`. For each model, the output includes the full regression summary as well as several model quality measures such as AIC, BIC, R^2, adjusted R^2, and residual standard error. Diagnostic plots are also generated for both models. This comparison is used to determine whether the transformed-response model provides a better statistical fit and better satisfies regression assumptions.

After selecting the log-response model as the main modeling framework, the document treats this as the full model and studies it in more detail. The full model summary is reported, along with both the standard ANOVA table and a Type II ANOVA table. These outputs allow assessment of the contribution of each predictor while accounting for the others. The analysis is designed not only to identify statistically significant variables but also to compare their relative importance within the fitted model.

To further evaluate the necessity of each predictor, the document carries out a series of partial F-tests. In each case, one predictor is removed from the full model and the reduced model is compared to the full one. Separate reduced models are created for removing `wtd_entropy_Valence`, `mean_ElectronAffinity`, `wtd_gmean_atomic_radius`, `mean_ThermalConductivity`, and `range_Valence`. The results are summarized in a table containing degrees of freedom, sums of squares, F-statistics, and p-values. This part of the analysis provides a formal test of whether each variable contributes meaningfully to explaining variation in the response.

The script then applies AIC-based model selection using backward elimination through `stepAIC`. This procedure starts from the full log-response model and removes predictors when doing so improves the Akaike Information Criterion. The selected model is then summarized and compared with the full model and several reduced candidate models using AIC, BIC, R^2, and adjusted R^2. This step offers a balance between explanatory power and model simplicity, allowing the final model choice to be justified from a model-selection perspective rather than only from individual p-values.

An additional part of the analysis examines multicollinearity among the quantitative predictors. A correlation matrix is computed for the numerical explanatory variables, and variance inflation factors (VIFs) are calculated for the full model. These checks are included to determine whether strong linear relationships among predictors may be inflating uncertainty in coefficient estimates or complicating interpretation.

The document also includes a detailed influence and outlier analysis. Several standard regression diagnostics are computed, including standardized residuals, studentized residuals, leverage values, Cook's distance, and DFFITS. Practical cutoffs are defined for each of these influence measures, and observations exceeding these thresholds are flagged. A table of flagged observations is produced, and counts are reported for each type of diagnostic flag. In addition, several influence-related plots are generated, including residual plots, leverage plots, Cook's distance plots, and an influence plot where point size reflects Cook's distance. This section is important because unusual or highly influential observations can strongly affect regression results and may distort conclusions if left unchecked.

To assess sensitivity of the analysis to influential observations, the document identifies especially problematic points and refits the full model after removing them. The refitted model is summarized and compared against the original full model using sample size, AIC, BIC, R^2, adjusted R^2, and residual standard error. This sensitivity analysis helps determine whether the main conclusions are robust or overly dependent on a small subset of influential cases.

The final chosen model is then taken to be the AIC-selected model. For this final model, the document reports the regression summary, a Type II ANOVA table, standard regression diagnostic plots, a coefficient table with 95% confidence intervals, and a table of overall model performance measures. These include R^2, adjusted R^2, residual standard error, AIC, BIC, the overall F-statistic, the overall model p-value, and the sample size. Together, these outputs provide a concise but complete statistical summary of the selected model.

Because the final response variable is on the log scale, the document also includes a back-transformed effect table. This table converts coefficient estimates into multiplicative effects on `Tc + 1` and corresponding percentage changes, which makes the results easier to interpret in practical terms. A separate interpretation helper table is also produced to summarize the direction of each effect and whether it is statistically significant at the 5% level. This is useful for translating technical regression output into conclusions that are easier to discuss in a written report.

Finally, the script reports a residual summary for the final model, including the minimum, first quartile, median, mean, third quartile, and maximum residual. This provides one more compact check of model behavior and helps support the diagnostic assessment given earlier.

Overall, this R Markdown file is designed to function as a full statistical analysis pipeline. It includes data filtering, exploratory summaries, transformation assessment, model fitting, hypothesis testing, model comparison, multicollinearity checks, influence diagnostics, sensitivity analysis, final model selection, and practical interpretation of results. Its goal is not only to fit a model, but also to justify that model carefully

and evaluate whether its conclusions are statistically reliable and substantively meaningful in the context of superconducting materials.

```
library(dplyr)
library(ggplot2)
library(car)
library(MASS)
```

```
options(scipen = 999)

df <- read.csv("Cleaned_Data.csv", stringsAsFactors = FALSE)

df <- df[grepl("Cu", df$CJ) & grepl("O", df$CJ), ]
df <- df[order(df$critical_temp, decreasing = TRUE), ][1:1000, ]

df$range_Valence <- factor(df$range_Valence, levels = 1:5)
df$log_Tc <- log(df$critical_temp + 1)

vars_needed <- c(
  "critical_temp",
  "log_Tc",
  "wtd_entropy_Valence",
  "mean_ElectronAffinity",
  "wtd_gmean_atomic_radius",
  "mean_ThermalConductivity",
  "range_Valence"
)

df <- df[, vars_needed]
df <- na.omit(df)

cat("Number of observations used:", nrow(df), "\n\n")
```

```
## Number of observations used: 813
```

```
# 2. Summary statistics

summary_stats <- data.frame(
  Variable = c(
    "critical_temp",
    "log_Tc",
    "wtd_entropy_Valence",
    "mean_ElectronAffinity",
    "wtd_gmean_atomic_radius",
    "mean_ThermalConductivity"
  ),
  Mean = sapply(df[, c(
    "critical_temp",
    "log_Tc",
    "wtd_entropy_Valence",
    "mean_ElectronAffinity",
    "wtd_gmean_atomic_radius",
    "mean_ThermalConductivity"
  )], mean),
  SD = sapply(df[, c(
```

```
    "critical_temp",
    "log_Tc",
    "wtd_entropy_Valence",
    "mean_ElectronAffinity",
    "wtd_gmean_atomic_radius",
    "mean_ThermalConductivity"
  )], sd),
  Min = sapply(df[, c(
    "critical_temp",
    "log_Tc",
    "wtd_entropy_Valence",
    "mean_ElectronAffinity",
    "wtd_gmean_atomic_radius",
    "mean_ThermalConductivity"
  )], min),
  Max = sapply(df[, c(
    "critical_temp",
    "log_Tc",
    "wtd_entropy_Valence",
    "mean_ElectronAffinity",
    "wtd_gmean_atomic_radius",
    "mean_ThermalConductivity"
  )], max)
)

summary_stats[, -1] <- round(summary_stats[, -1], 4)

cat("SUMMARY STATISTICS\n")
```

```
## SUMMARY STATISTICS
```

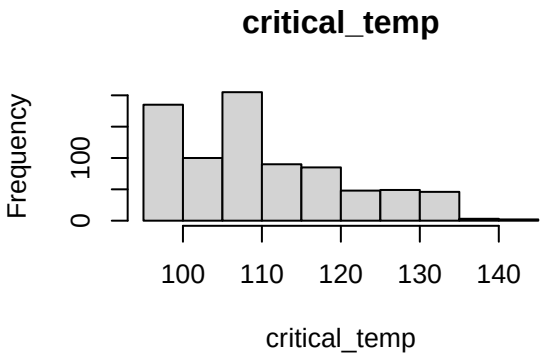
```
print(summary_stats)
```

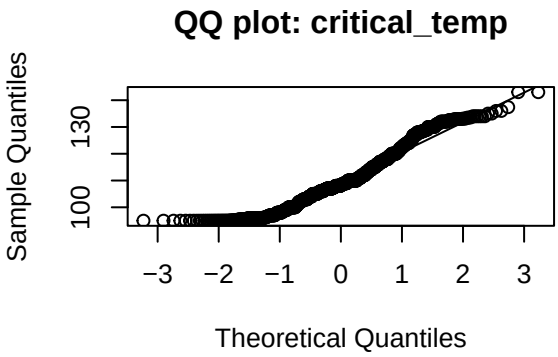
```
##                                       Variable     Mean      SD     Min
## critical_temp                     critical_temp 110.1667 10.8915 95.0000
## log_Tc                                   log_Tc   4.7064  0.0960  4.5643
## wtd_entropy_Valence         wtd_entropy_Valence   1.4749  0.1573  1.0114
## mean_ElectronAffinity     mean_ElectronAffinity  66.9585 15.1994 43.6314
## wtd_gmean_atomic_radius wtd_gmean_atomic_radius 114.5595 30.5234 75.5574
## mean_ThermalConductivity mean_ThermalConductivity 118.7720 18.1175 77.0044
##                                Max
## critical_temp            143.0000
## log_Tc                     4.9698
## wtd_entropy_Valence        1.9350
## mean_ElectronAffinity    122.2440
## wtd_gmean_atomic_radius  168.3544
## mean_ThermalConductivity 209.6053
```

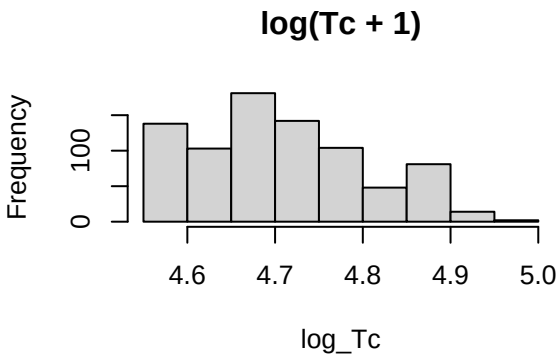
```
cat("\n")
```

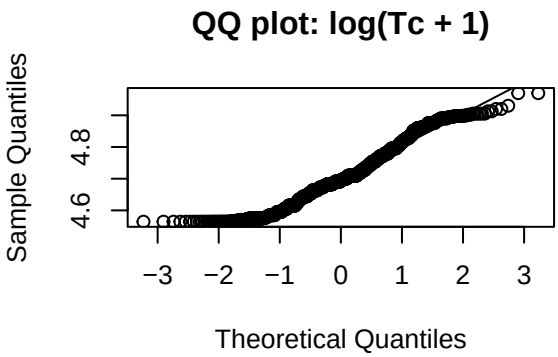
```
# Transformation check

par(mfrow = c(2, 2))
hist(df$critical_temp, main = "critical_temp", xlab = "critical_temp")
qqnorm(df$critical_temp, main = "QQ plot: critical_temp")
```

```r
qqline(df$critical_temp)

hist(df$log_Tc, main = "log(Tc + 1)", xlab = "log_Tc")
qqnorm(df$log_Tc, main = "QQ plot: log(Tc + 1)")
qqline(df$log_Tc)
```










```r
# Raw vs log-response models

fit_raw <- lm(
  critical_temp ~
    wtd_entropy_Valence +
    mean_ElectronAffinity +
    wtd_gmean_atomic_radius +
    mean_ThermalConductivity +
    range_Valence,
  data = df
)

fit_log <- lm(
  log_Tc ~
    wtd_entropy_Valence +
    mean_ElectronAffinity +
    wtd_gmean_atomic_radius +
    mean_ThermalConductivity +
    range_Valence,
  data = df
)

cat("RAW RESPONSE MODEL\n")
```

```
## RAW RESPONSE MODEL
```

```
print(summary(fit_raw))
```

```
##
## Call:
## lm(formula = critical_temp ~ wtd_entropy_Valence + mean_ElectronAffinity +
##     wtd_gmean_atomic_radius + mean_ThermalConductivity + range_Valence,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.032  -6.828  -1.219   5.685  37.901
##
## Coefficients:
##                           Estimate Std. Error t value             Pr(>|t|)
## (Intercept)              62.05094    5.84908  10.609 < 0.0000000000000002 ***
## wtd_entropy_Valence      34.08935    3.76646   9.051 < 0.0000000000000002 ***
## mean_ElectronAffinity    -0.06063    0.02451  -2.473               0.0136 *
## wtd_gmean_atomic_radius  -0.06680    0.01684  -3.966       0.000079511596 ***
## mean_ThermalConductivity  0.09762    0.02042   4.781       0.000002077534 ***
## range_Valence2           -0.21906    1.03359  -0.212               0.8322
## range_Valence3           -6.10483    0.97443  -6.265       0.000000000607 ***
## range_Valence4           -6.10841    1.45764  -4.191       0.000030904843 ***
## range_Valence5           10.01753    1.93563   5.175       0.000000287603 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.845 on 804 degrees of freedom
## Multiple R-squared:  0.1909, Adjusted R-squared:  0.1829
## F-statistic: 23.72 on 8 and 804 DF,  p-value: < 0.00000000000000022
```

```
cat("\nAIC raw model:", AIC(fit_raw), "\n")
```

```
##
## AIC raw model: 6036.811
```

```
cat("BIC raw model:", BIC(fit_raw), "\n")
```

```
## BIC raw model: 6083.819
```

```
cat("Adjusted R-squared raw model:", summary(fit_raw)$adj.r.squared, "\n\n")
```

```
## Adjusted R-squared raw model: 0.1828685
```

```
cat("LOG RESPONSE MODEL\n")
```

```
## LOG RESPONSE MODEL
```

```
print(summary(fit_log))
```

```
##
## Call:
## lm(formula = log_Tc ~ wtd_entropy_Valence + mean_ElectronAffinity +
##     wtd_gmean_atomic_radius + mean_ThermalConductivity + range_Valence,
##     data = df)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.23536 -0.06136 -0.00827  0.05294  0.32731
##
## Coefficients:
##                            Estimate Std. Error t value             Pr(>|t|)
## (Intercept)               4.2547066  0.0513988  82.778 < 0.0000000000000002 ***
## wtd_entropy_Valence       0.3146948  0.0330978   9.508 < 0.0000000000000002 ***
## mean_ElectronAffinity    -0.0005379  0.0002154  -2.497               0.0127 *
## wtd_gmean_atomic_radius  -0.0006198  0.0001480  -4.188       0.000031198767 ***
## mean_ThermalConductivity  0.0009481  0.0001794   5.284       0.000000163182 ***
## range_Valence2           -0.0021045  0.0090827  -0.232               0.8168
## range_Valence3           -0.0531612  0.0085628  -6.208       0.000000000857 ***
## range_Valence4           -0.0537417  0.0128090  -4.196       0.000030244718 ***
## range_Valence5            0.0824111  0.0170094   4.845       0.000001518918 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08652 on 804 degrees of freedom
## Multiple R-squared:  0.1965, Adjusted R-squared:  0.1885
## F-statistic: 24.58 on 8 and 804 DF,  p-value: < 0.00000000000000022
```

```
cat("\nAIC log model:", AIC(fit_log), "\n")
```

```
##
## AIC log model: -1661.363
```

```
cat("BIC log model:", BIC(fit_log), "\n")
```

```
## BIC log model: -1614.356
```

```
cat("Adjusted R-squared log model:", summary(fit_log)$adj.r.squared, "\n\n")
```

```
## Adjusted R-squared log model: 0.1885057
```

```
transformation_comparison <- data.frame(
  Model = c("Raw response", "Log response"),
  AIC = c(AIC(fit_raw), AIC(fit_log)),
  BIC = c(BIC(fit_raw), BIC(fit_log)),
  R_squared = c(summary(fit_raw)$r.squared, summary(fit_log)$r.squared),
  Adj_R_squared = c(summary(fit_raw)$adj.r.squared, summary(fit_log)$adj.r.squared),
  Residual_SE = c(summary(fit_raw)$sigma, summary(fit_log)$sigma)
)
transformation_comparison[, -1] <- round(transformation_comparison[, -1], 4)

cat("TRANSFORMATION COMPARISON\n")
```

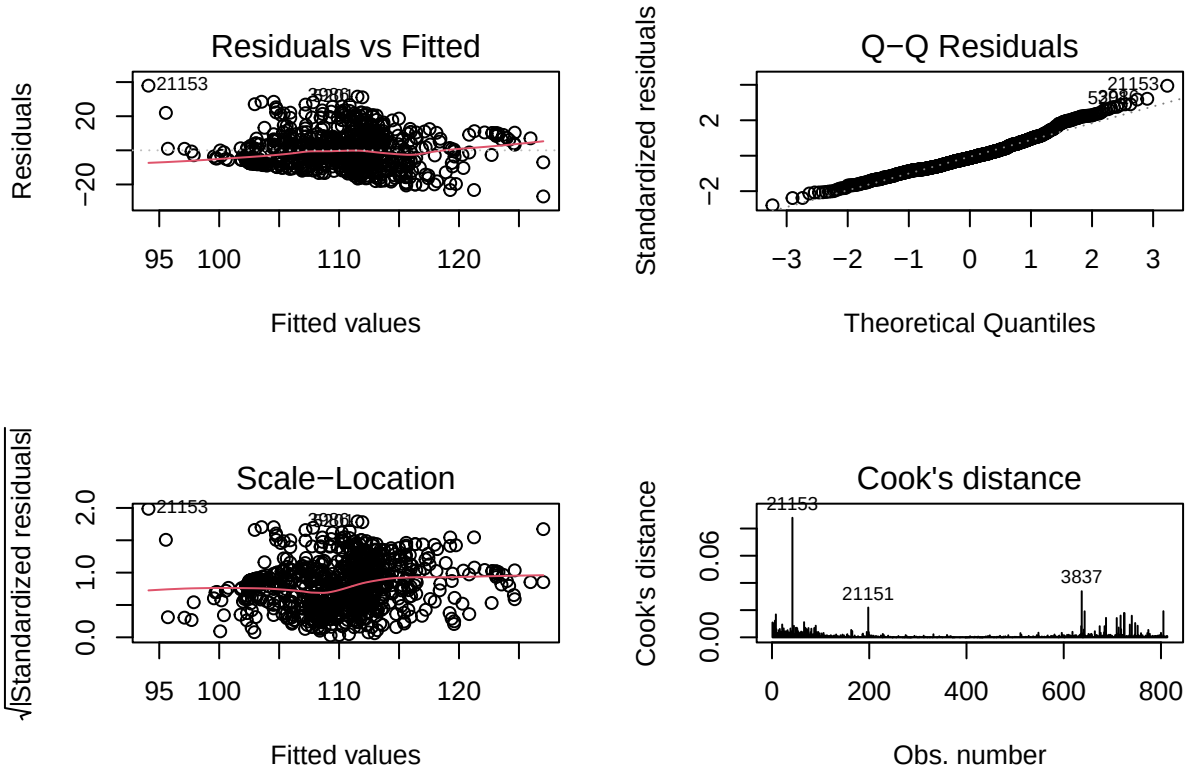
```
## TRANSFORMATION COMPARISON
```

```
print(transformation_comparison)
```

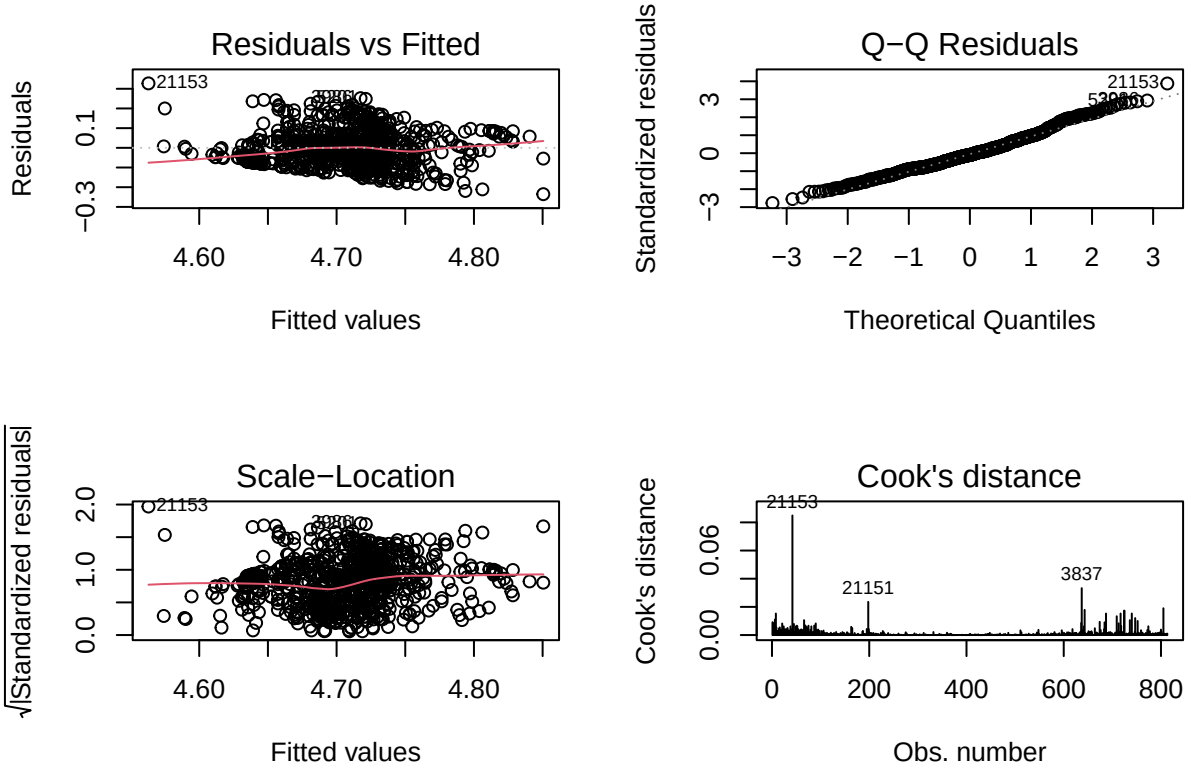
```
##          Model       AIC       BIC R_squared Adj_R_squared Residual_SE
## 1 Raw response  6036.811  6083.819    0.1909        0.1829      9.8454
## 2 Log response -1661.363 -1614.356    0.1965        0.1885      0.0865
```

```
cat("\n")
```

```
par(mfrow = c(2, 2))
plot(fit_raw, which = 1:4)
```




```
par(mfrow = c(2, 2))
plot(fit_log, which = 1:4)
```




```
# Full model
```

```r
fit_full <- fit_log

cat("FULL MODEL SUMMARY\n")
```

```
## FULL MODEL SUMMARY
```

```r
print(summary(fit_full))
```

```
##
## Call:
## lm(formula = log_Tc ~ wtd_entropy_Valence + mean_ElectronAffinity +
##     wtd_gmean_atomic_radius + mean_ThermalConductivity + range_Valence,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23536 -0.06136 -0.00827  0.05294  0.32731
##
## Coefficients:
##                            Estimate Std. Error t value             Pr(>|t|)
## (Intercept)               4.2547066  0.0513988  82.778 < 0.0000000000000002 ***
## wtd_entropy_Valence       0.3146948  0.0330978   9.508 < 0.0000000000000002 ***
## mean_ElectronAffinity    -0.0005379  0.0002154  -2.497               0.0127 *
## wtd_gmean_atomic_radius  -0.0006198  0.0001480  -4.188       0.000031198767 ***
## mean_ThermalConductivity  0.0009481  0.0001794   5.284       0.000000163182 ***
## range_Valence2           -0.0021045  0.0090827  -0.232               0.8168
## range_Valence3           -0.0531612  0.0085628  -6.208       0.000000000857 ***
## range_Valence4           -0.0537417  0.0128090  -4.196       0.000030244718 ***
## range_Valence5            0.0824111  0.0170094   4.845       0.000001518918 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08652 on 804 degrees of freedom
## Multiple R-squared:  0.1965, Adjusted R-squared:  0.1885
## F-statistic: 24.58 on 8 and 804 DF,  p-value: < 0.00000000000000022
```

```r
cat("\nFULL MODEL ANOVA\n")
```

```
##
## FULL MODEL ANOVA
```

```r
print(anova(fit_full))
```

```
## Analysis of Variance Table
##
## Response: log_Tc
##                           Df Sum Sq  Mean Sq F value                Pr(>F)
## wtd_entropy_Valence        1 0.3074 0.307440 41.0737       0.0000000002496 ***
## mean_ElectronAffinity      1 0.2550 0.254972 34.0640       0.0000000077317 ***
## wtd_gmean_atomic_radius    1 0.0396 0.039606  5.2913               0.02169 *
## mean_ThermalConductivity   1 0.2395 0.239488 31.9953       0.0000000214830 ***
## range_Valence              4 0.6302 0.157559 21.0497 < 0.00000000000000022 ***
## Residuals                804 6.0180 0.007485
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
cat("\nType II ANOVA\n")
```

```
##
## Type II ANOVA
```

```r
print(Anova(fit_full, type = 2))
```

```
## Anova Table (Type II tests)
##
## Response: log_Tc
##                            Sum Sq  Df F value                 Pr(>F)
## wtd_entropy_Valence        0.6767   1 90.4026 < 0.00000000000000022 ***
## mean_ElectronAffinity      0.0467   1  6.2347               0.01273 *
## wtd_gmean_atomic_radius    0.1313   1 17.5428          0.0000311988 ***
## mean_ThermalConductivity   0.2090   1 27.9172          0.0000001632 ***
## range_Valence              0.6302   4 21.0497 < 0.00000000000000022 ***
## Residuals                  6.0180 804
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
cat("\nR-squared:", summary(fit_full)$r.squared, "\n")
```

```
##
## R-squared: 0.1965007
```

```r
cat("Adjusted R-squared:", summary(fit_full)$adj.r.squared, "\n")
```

```
## Adjusted R-squared: 0.1885057
```

```r
cat("AIC:", AIC(fit_full), "\n")
```

```
## AIC: -1661.363
```

```r
cat("BIC:", BIC(fit_full), "\n\n")
```

```
## BIC: -1614.356
```

```r
# Partial F tests

fit_no_entropy <- lm(
  log_Tc ~ mean_ElectronAffinity + wtd_gmean_atomic_radius +
    mean_ThermalConductivity + range_Valence,
  data = df
)

fit_no_affinity <- lm(
  log_Tc ~ wtd_entropy_Valence + wtd_gmean_atomic_radius +
    mean_ThermalConductivity + range_Valence,
  data = df
)

fit_no_radius <- lm(
  log_Tc ~ wtd_entropy_Valence + mean_ElectronAffinity +
    mean_ThermalConductivity + range_Valence,
  data = df
)

fit_no_thermal <- lm(
```

```
  log_Tc ~ wtd_entropy_Valence + mean_ElectronAffinity +
    wtd_gmean_atomic_radius + range_Valence,
  data = df
)

fit_no_range <- lm(
  log_Tc ~ wtd_entropy_Valence + mean_ElectronAffinity +
    wtd_gmean_atomic_radius + mean_ThermalConductivity,
  data = df
)

cat("PARTIAL F TEST: remove wtd_entropy_Valence\n")
```

```
## PARTIAL F TEST: remove wtd_entropy_Valence
```

```
pf_entropy <- anova(fit_no_entropy, fit_full)
print(pf_entropy)
```

```
## Analysis of Variance Table
##
## Model 1: log_Tc ~ mean_ElectronAffinity + wtd_gmean_atomic_radius + mean_ThermalConductivity +
##     range_Valence
## Model 2: log_Tc ~ wtd_entropy_Valence + mean_ElectronAffinity + wtd_gmean_atomic_radius +
##     mean_ThermalConductivity + range_Valence
##   Res.Df    RSS Df Sum of Sq      F                Pr(>F)
## 1    805 6.6947
## 2    804 6.0180  1   0.67667 90.403 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("\n")
```

```
cat("PARTIAL F TEST: remove mean_ElectronAffinity\n")
```

```
## PARTIAL F TEST: remove mean_ElectronAffinity
```

```
pf_affinity <- anova(fit_no_affinity, fit_full)
print(pf_affinity)
```

```
## Analysis of Variance Table
##
## Model 1: log_Tc ~ wtd_entropy_Valence + wtd_gmean_atomic_radius + mean_ThermalConductivity +
##     range_Valence
## Model 2: log_Tc ~ wtd_entropy_Valence + mean_ElectronAffinity + wtd_gmean_atomic_radius +
##     mean_ThermalConductivity + range_Valence
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    805 6.0647
## 2    804 6.0180  1  0.046668 6.2347 0.01273 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("\n")
```

```
cat("PARTIAL F TEST: remove wtd_gmean_atomic_radius\n")
```

```
## PARTIAL F TEST: remove wtd_gmean_atomic_radius
```

```r
pf_radius <- anova(fit_no_radius, fit_full)
print(pf_radius)
```

```
## Analysis of Variance Table
##
## Model 1: log_Tc ~ wtd_entropy_Valence + mean_ElectronAffinity + mean_ThermalConductivity +
##     range_Valence
## Model 2: log_Tc ~ wtd_entropy_Valence + mean_ElectronAffinity + wtd_gmean_atomic_radius +
##     mean_ThermalConductivity + range_Valence
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    805 6.1493
## 2    804 6.0180  1   0.13131 17.543 0.0000312 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
cat("\n")
```

```r
cat("PARTIAL F TEST: remove mean_ThermalConductivity\n")
```

```
## PARTIAL F TEST: remove mean_ThermalConductivity
```

```r
pf_thermal <- anova(fit_no_thermal, fit_full)
print(pf_thermal)
```

```
## Analysis of Variance Table
##
## Model 1: log_Tc ~ wtd_entropy_Valence + mean_ElectronAffinity + wtd_gmean_atomic_radius +
##     range_Valence
## Model 2: log_Tc ~ wtd_entropy_Valence + mean_ElectronAffinity + wtd_gmean_atomic_radius +
##     mean_ThermalConductivity + range_Valence
##   Res.Df   RSS Df Sum of Sq      F       Pr(>F)
## 1    805 6.227
## 2    804 6.018  1   0.20896 27.917 0.0000001632 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
cat("\n")
```

```r
cat("PARTIAL F TEST: remove range_Valence\n")
```

```
## PARTIAL F TEST: remove range_Valence
```

```r
pf_range <- anova(fit_no_range, fit_full)
print(pf_range)
```

```
## Analysis of Variance Table
##
## Model 1: log_Tc ~ wtd_entropy_Valence + mean_ElectronAffinity + wtd_gmean_atomic_radius +
##     mean_ThermalConductivity
## Model 2: log_Tc ~ wtd_entropy_Valence + mean_ElectronAffinity + wtd_gmean_atomic_radius +
##     mean_ThermalConductivity + range_Valence
##   Res.Df    RSS Df Sum of Sq     F                Pr(>F)
## 1    808 6.6482
## 2    804 6.0180  4   0.63024 21.05 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
cat("\n")
```

```r
partial_f_summary <- data.frame(
  Removed_Term = c(
    "wtd_entropy_Valence",
    "mean_ElectronAffinity",
    "wtd_gmean_atomic_radius",
    "mean_ThermalConductivity",
    "range_Valence"
  ),
  Df = c(
    pf_entropy$Df[2],
    pf_affinity$Df[2],
    pf_radius$Df[2],
    pf_thermal$Df[2],
    pf_range$Df[2]
  ),
  Sum_of_Sq = c(
    pf_entropy$`Sum of Sq`[2],
    pf_affinity$`Sum of Sq`[2],
    pf_radius$`Sum of Sq`[2],
    pf_thermal$`Sum of Sq`[2],
    pf_range$`Sum of Sq`[2]
  ),
  F_value = c(
    pf_entropy$F[2],
    pf_affinity$F[2],
    pf_radius$F[2],
    pf_thermal$F[2],
    pf_range$F[2]
  ),
  P_value = c(
    pf_entropy$`Pr(>F)`[2],
    pf_affinity$`Pr(>F)`[2],
    pf_radius$`Pr(>F)`[2],
    pf_thermal$`Pr(>F)`[2],
    pf_range$`Pr(>F)`[2]
  )
)

partial_f_summary[, -1] <- lapply(partial_f_summary[, -1], function(x) round(x, 6))
cat("PARTIAL F TEST SUMMARY TABLE\n")
```

```
## PARTIAL F TEST SUMMARY TABLE
```

```r
print(partial_f_summary)
```

```
##               Removed_Term Df Sum_of_Sq   F_value  P_value
## 1      wtd_entropy_Valence  1  0.676671 90.402601 0.000000
## 2    mean_ElectronAffinity  1  0.046668  6.234738 0.012726
## 3  wtd_gmean_atomic_radius  1  0.131309 17.542769 0.000031
## 4 mean_ThermalConductivity  1  0.208963 27.917241 0.000000
## 5            range_Valence  4  0.630236 21.049715 0.000000
```

```r
cat("\n")
```

```r
# 7. AIC model selection

fit_scope <- lm(
  log_Tc ~
    wtd_entropy_Valence +
    mean_ElectronAffinity +
    wtd_gmean_atomic_radius +
    mean_ThermalConductivity +
    range_Valence,
  data = df
)

fit_aic_backward <- stepAIC(fit_scope, direction = "backward", trace = TRUE)
```

```
## Start:  AIC=-3970.56
## log_Tc ~ wtd_entropy_Valence + mean_ElectronAffinity + wtd_gmean_atomic_radius +
##     mean_ThermalConductivity + range_Valence
##
##                            Df Sum of Sq    RSS     AIC
## <none>                                  6.0180 -3970.6
## - mean_ElectronAffinity     1   0.04667 6.0647 -3966.3
## - wtd_gmean_atomic_radius   1   0.13131 6.1493 -3955.0
## - mean_ThermalConductivity  1   0.20896 6.2270 -3944.8
## - range_Valence             4   0.63024 6.6482 -3897.6
## - wtd_entropy_Valence       1   0.67667 6.6947 -3885.9
```

```r
cat("\nAIC-SELECTED MODEL SUMMARY\n")
```

```
##
## AIC-SELECTED MODEL SUMMARY
```

```r
print(summary(fit_aic_backward))
```

```
##
## Call:
## lm(formula = log_Tc ~ wtd_entropy_Valence + mean_ElectronAffinity +
##     wtd_gmean_atomic_radius + mean_ThermalConductivity + range_Valence,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23536 -0.06136 -0.00827  0.05294  0.32731
##
## Coefficients:
##                             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)                4.2547066  0.0513988  82.778 < 0.0000000000000002 ***
## wtd_entropy_Valence        0.3146948  0.0330978   9.508 < 0.0000000000000002 ***
## mean_ElectronAffinity     -0.0005379  0.0002154  -2.497               0.0127 *
## wtd_gmean_atomic_radius   -0.0006198  0.0001480  -4.188       0.000031198767 ***
## mean_ThermalConductivity   0.0009481  0.0001794   5.284       0.000000163182 ***
## range_Valence2            -0.0021045  0.0090827  -0.232               0.8168
## range_Valence3            -0.0531612  0.0085628  -6.208       0.000000000857 ***
## range_Valence4            -0.0537417  0.0128090  -4.196       0.000030244718 ***
## range_Valence5             0.0824111  0.0170094   4.845       0.000001518918 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08652 on 804 degrees of freedom
## Multiple R-squared:  0.1965, Adjusted R-squared:  0.1885
## F-statistic: 24.58 on 8 and 804 DF,  p-value: < 0.00000000000000022
```

```
cat("\nAIC of selected model:", AIC(fit_aic_backward), "\n")
```

```
##
## AIC of selected model: -1661.363
```

```
cat("BIC of selected model:", BIC(fit_aic_backward), "\n")
```

```
## BIC of selected model: -1614.356
```

```
cat("Adjusted R-squared of selected model:", summary(fit_aic_backward)$adj.r.squared, "\n\n")
```

```
## Adjusted R-squared of selected model: 0.1885057
```

```
# Compare candidate models

model_comparison <- data.frame(
  Model = c(
    "Full model",
    "No entropy",
    "No affinity",
    "No radius",
    "No thermal",
    "No range",
    "AIC selected"
  ),
  AIC = c(
    AIC(fit_full),
    AIC(fit_no_entropy),
    AIC(fit_no_affinity),
    AIC(fit_no_radius),
    AIC(fit_no_thermal),
    AIC(fit_no_range),
    AIC(fit_aic_backward)
  ),
  BIC = c(
    BIC(fit_full),
    BIC(fit_no_entropy),
    BIC(fit_no_affinity),
    BIC(fit_no_radius),
    BIC(fit_no_thermal),
    BIC(fit_no_range),
    BIC(fit_aic_backward)
  ),
  R_squared = c(
    summary(fit_full)$r.squared,
    summary(fit_no_entropy)$r.squared,
    summary(fit_no_affinity)$r.squared,
    summary(fit_no_radius)$r.squared,
    summary(fit_no_thermal)$r.squared,
    summary(fit_no_range)$r.squared,
```

```
    summary(fit_aic_backward)$r.squared
  ),
  Adj_R_squared = c(
    summary(fit_full)$adj.r.squared,
    summary(fit_no_entropy)$adj.r.squared,
    summary(fit_no_affinity)$adj.r.squared,
    summary(fit_no_radius)$adj.r.squared,
    summary(fit_no_thermal)$adj.r.squared,
    summary(fit_no_range)$adj.r.squared,
    summary(fit_aic_backward)$adj.r.squared
  )
)

model_comparison[, -1] <- round(model_comparison[, -1], 4)

cat("MODEL COMPARISON TABLE\n")
```

```
## MODEL COMPARISON TABLE
```

```
print(model_comparison)
```

```
##          Model       AIC       BIC R_squared Adj_R_squared
## 1   Full model -1661.363 -1614.356    0.1965        0.1885
## 2   No entropy -1576.733 -1534.426    0.1062        0.0984
## 3  No affinity -1657.083 -1614.776    0.1903        0.1832
## 4    No radius -1645.815 -1603.508    0.1790        0.1718
## 5   No thermal -1635.613 -1593.306    0.1686        0.1614
## 6     No range -1588.391 -1560.187    0.1124        0.1080
## 7 AIC selected -1661.363 -1614.356    0.1965        0.1885
```

```
cat("\n")
```

```
# 9. Multicollinearity

num_vars <- df[, c(
  "wtd_entropy_Valence",
  "mean_ElectronAffinity",
  "wtd_gmean_atomic_radius",
  "mean_ThermalConductivity"
)]

cat("CORRELATION MATRIX\n")
```

```
## CORRELATION MATRIX
```

```
print(round(cor(num_vars), 3))
```

```
##                          wtd_entropy_Valence mean_ElectronAffinity
## wtd_entropy_Valence                    1.000                -0.121
## mean_ElectronAffinity                 -0.121                 1.000
## wtd_gmean_atomic_radius                0.707                 0.046
## mean_ThermalConductivity              -0.231                 0.061
##                          wtd_gmean_atomic_radius mean_ThermalConductivity
## wtd_entropy_Valence                        0.707                  -0.231
## mean_ElectronAffinity                      0.046                   0.061
## wtd_gmean_atomic_radius                    1.000                  -0.038
```

```
## mean_ThermalConductivity                  -0.038                    1.000
```

```
cat("\n")
```

```
cat("VIF VALUES\n")
```

```
## VIF VALUES
```

```
print(vif(fit_full))
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## wtd_entropy_Valence      2.939799  1        1.714584
## mean_ElectronAffinity    1.163076  1        1.078460
## wtd_gmean_atomic_radius  2.213537  1        1.487796
## mean_ThermalConductivity 1.146563  1        1.070777
## range_Valence            1.660283  4        1.065425
```

```
cat("\n")
```

```
# Outliers / leverage / influence

n <- nrow(df)
p <- length(coef(fit_full))

std_resid <- rstandard(fit_full)
stud_resid <- rstudent(fit_full)
lev <- hatvalues(fit_full)
cooks <- cooks.distance(fit_full)
dffits_vals <- dffits(fit_full)

lev_cutoff <- 2 * p / n
cook_cutoff <- 4 / n
dffits_cutoff <- 2 * sqrt(p / n)

influence_table <- data.frame(
  Index = 1:n,
  Fitted = fitted(fit_full),
  Residual = resid(fit_full),
  Std_Resid = std_resid,
  Stud_Resid = stud_resid,
  Leverage = lev,
  CooksD = cooks,
  DFFITS = dffits_vals
)

influence_table$Outlier_Flag <- abs(influence_table$Stud_Resid) > 3
influence_table$Leverage_Flag <- influence_table$Leverage > lev_cutoff
influence_table$Cook_Flag <- influence_table$CooksD > cook_cutoff
influence_table$DFFITS_Flag <- abs(influence_table$DFFITS) > dffits_cutoff

flagged <- influence_table %>%
  filter(Outlier_Flag | Leverage_Flag | Cook_Flag | DFFITS_Flag)

cat("INFLUENCE CUT-OFFS\n")
```

```
## INFLUENCE CUT-OFFS
```

```r
cat("Leverage cutoff:", lev_cutoff, "\n")
```

```
## Leverage cutoff: 0.02214022
```

```r
cat("Cook's distance cutoff:", cook_cutoff, "\n")
```

```
## Cook's distance cutoff: 0.004920049
```

```r
cat("DFFITS cutoff:", dffits_cutoff, "\n\n")
```

```
## DFFITS cutoff: 0.2104292
```

```r
cat("NUMBER OF FLAGGED OBSERVATIONS:", nrow(flagged), "\n\n")
```

```
## NUMBER OF FLAGGED OBSERVATIONS: 116
```

```r
if (nrow(flagged) > 0) {
  flagged <- flagged[order(-flagged$CooksD), ]
  cat("TOP FLAGGED OBSERVATIONS\n")
  print(head(flagged, 20))
  cat("\n")
}
```

```
## TOP FLAGGED OBSERVATIONS
##       Index   Fitted   Residual Std_Resid Stud_Resid    Leverage      CooksD
## 21153    42 4.563042  0.3273069  3.878072   3.912425 0.048340289 0.084882231
## 3837    637 4.850476 -0.2353552 -2.772842  -2.784463 0.037499876 0.033284089
## 21151   198 4.574943  0.1999696  2.355084   2.361780 0.036796077 0.023542561
## 6481    805 4.744403 -0.1800549 -2.120380  -2.125011 0.036646606 0.019003482
## 6083    643 4.778660 -0.1635399 -1.930604  -1.933890 0.041340077 0.017858724
## 6486    724 4.751353 -0.1735540 -2.043101  -2.047151 0.035964843 0.017303046
## 8648    725 4.751263 -0.1734638 -2.042047  -2.046090 0.035972819 0.017289173
## 6480    740 4.732753 -0.1580425 -1.862579  -1.865449 0.038118347 0.015275637
## 6485    717 4.744314 -0.1613893 -1.900577  -1.903676 0.036656018 0.015271880
## 10591   687 4.806151 -0.2110315 -2.466577  -2.474423 0.022068378 0.015254881
## 3684      8 4.681122  0.2241529  2.616707   2.626286 0.019648869 0.015248376
## 6484    709 4.732719 -0.1477515 -1.741301  -1.743509 0.038122955 0.013352795
## 8384    747 4.793750 -0.2190393 -2.552736  -2.561550 0.016361885 0.012043907
## 5319      6 4.699136  0.2135193  2.487692   2.495768 0.015797301 0.011036945
## 3489    736 4.757729 -0.1830178 -2.137322  -2.142087 0.020396838 0.010568420
## 3683     66 4.681122  0.1864126  2.176136   2.181215 0.019648869 0.010545939
## 9352    752 4.743254 -0.1685430 -1.970450  -1.973996 0.022551582 0.009953396
## 4923    674 4.720665 -0.1205071 -1.421660  -1.422565 0.040075656 0.009375465
## 3986      1 4.717439  0.2523739  2.930928   2.944879 0.009438195 0.009094424
## 8379    684 4.779134 -0.1840138 -2.145562  -2.150392 0.017297740 0.009003407
##           DFFITS Outlier_Flag Leverage_Flag Cook_Flag DFFITS_Flag
## 21153  0.8817790         TRUE          TRUE      TRUE        TRUE
## 3837  -0.5496116        FALSE          TRUE      TRUE        TRUE
## 21151  0.4616162        FALSE          TRUE      TRUE        TRUE
## 6481  -0.4144626        FALSE          TRUE      TRUE        TRUE
## 6083  -0.4015922        FALSE          TRUE      TRUE        TRUE
## 6486  -0.3954054        FALSE          TRUE      TRUE        TRUE
## 8648  -0.3952458        FALSE          TRUE      TRUE        TRUE
## 6480  -0.3713553        FALSE          TRUE      TRUE        TRUE
## 6485  -0.3713428        FALSE          TRUE      TRUE        TRUE
## 10591 -0.3717105        FALSE         FALSE      TRUE        TRUE
## 3684   0.3718091        FALSE         FALSE      TRUE        TRUE
```

```
## 6484  -0.3471024        FALSE          TRUE      TRUE        TRUE
## 8384  -0.3303710        FALSE         FALSE      TRUE        TRUE
## 5319   0.3161938        FALSE         FALSE      TRUE        TRUE
## 3489  -0.3090960        FALSE         FALSE      TRUE        TRUE
## 3683   0.3087994        FALSE         FALSE      TRUE        TRUE
## 9352  -0.2998388        FALSE          TRUE      TRUE        TRUE
## 4923  -0.2906658        FALSE          TRUE      TRUE        TRUE
## 3986   0.2874559        FALSE         FALSE      TRUE        TRUE
## 8379  -0.2852997        FALSE         FALSE      TRUE        TRUE
```

```r
# Additional summary for writeup
influence_counts <- data.frame(
  Metric = c("Outlier_Flag", "Leverage_Flag", "Cook_Flag", "DFFITS_Flag"),
  Count = c(
    sum(influence_table$Outlier_Flag),
    sum(influence_table$Leverage_Flag),
    sum(influence_table$Cook_Flag),
    sum(influence_table$DFFITS_Flag)
  )
)

cat("INFLUENCE FLAG COUNTS\n")
```

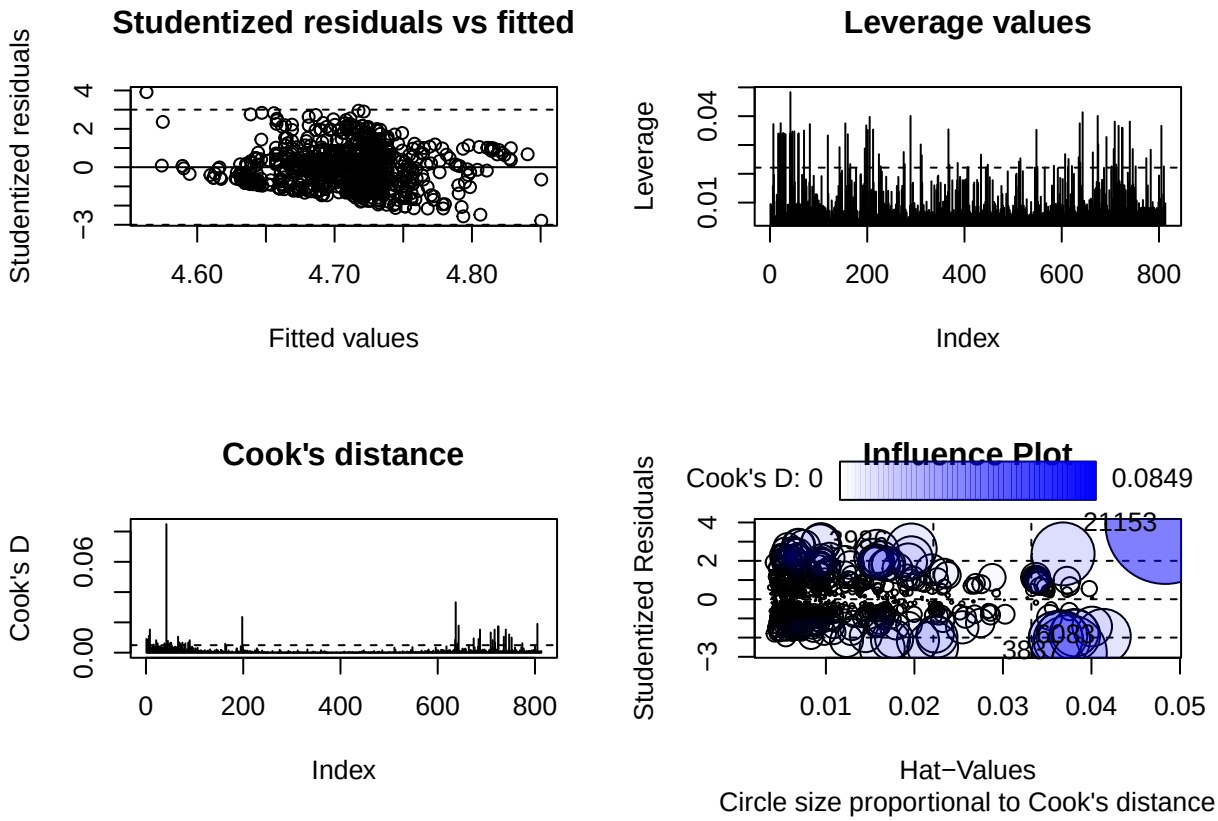
```
## INFLUENCE FLAG COUNTS
```

```r
print(influence_counts)
```

```
##          Metric Count
## 1  Outlier_Flag     1
## 2 Leverage_Flag    85
## 3     Cook_Flag    46
## 4   DFFITS_Flag    46
```

```r
cat("\n")
```

```r
# Influence plots

par(mfrow = c(2, 2))

plot(fitted(fit_full), rstudent(fit_full),
     xlab = "Fitted values", ylab = "Studentized residuals",
     main = "Studentized residuals vs fitted")
abline(h = c(-3, 0, 3), lty = c(2, 1, 2))

plot(lev, type = "h", main = "Leverage values", ylab = "Leverage")
abline(h = lev_cutoff, lty = 2)

plot(cooks, type = "h", main = "Cook's distance", ylab = "Cook's D")
abline(h = cook_cutoff, lty = 2)

influencePlot(fit_full, main = "Influence Plot", sub = "Circle size proportional to Cook's distance")
```




```
##          StudRes         Hat       CookD
## 3986   2.944879 0.009438195 0.009094424
## 21153  3.912425 0.048340289 0.084882231
## 3837  -2.784463 0.037499876 0.033284089
## 6083  -1.933890 0.041340077 0.017858724
```

```
# refit after removing strongest problematic points

remove_idx <- flagged %>%
  filter(Cook_Flag & (Outlier_Flag | Leverage_Flag)) %>%
  pull(Index)

cat("INDICES TO CONSIDER REMOVING\n")
```

```
## INDICES TO CONSIDER REMOVING
```

```
print(remove_idx)
```

```
##  [1]  42 637 198 805 643 724 725 740 717 709 752 674 713 774  32
```

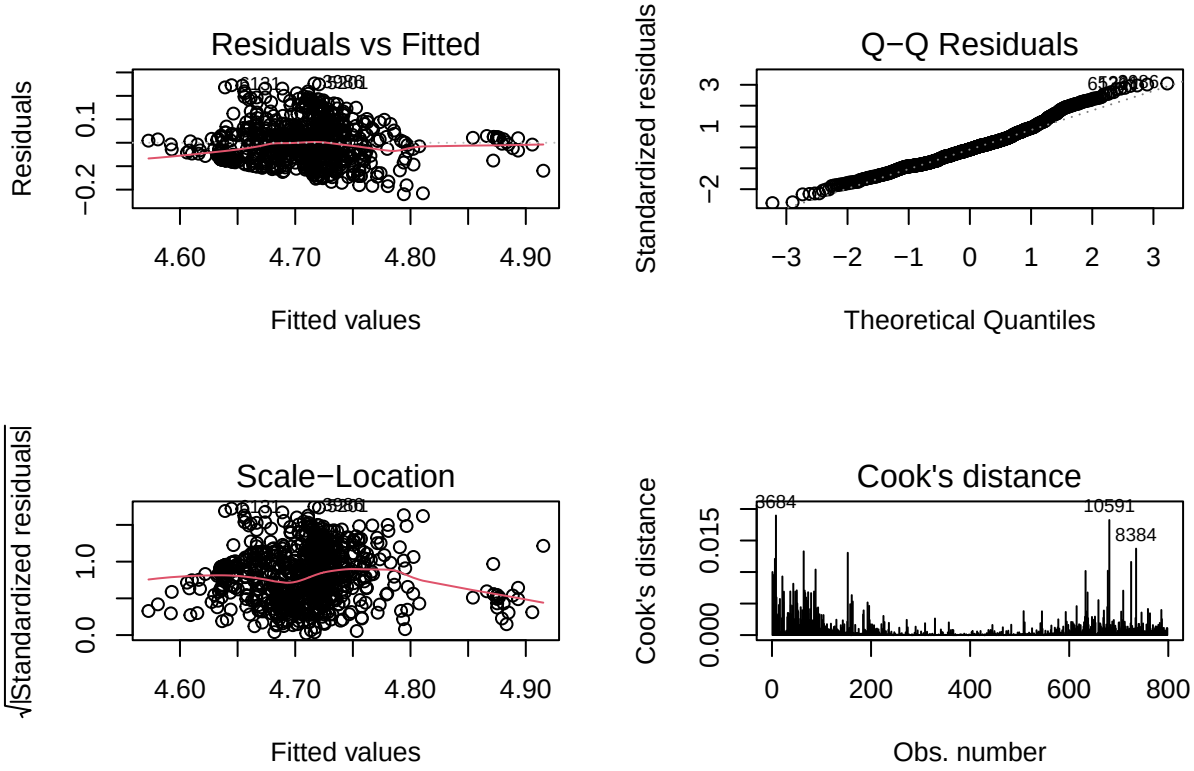
```
cat("\n")
```

```
fit_full_clean <- NULL

if (length(remove_idx) > 0) {
  df_clean <- df[-remove_idx, ]

  fit_full_clean <- lm(
    log_Tc ~
      wtd_entropy_Valence +
```

```
      mean_ElectronAffinity +
      wtd_gmean_atomic_radius +
      mean_ThermalConductivity +
      range_Valence,
    data = df_clean
  )

  cat("REFIT MODEL AFTER REMOVAL\n")
  print(summary(fit_full_clean))
  cat("\nAIC cleaned model:", AIC(fit_full_clean), "\n")
  cat("BIC cleaned model:", BIC(fit_full_clean), "\n")
  cat("Adjusted R-squared cleaned model:", summary(fit_full_clean)$adj.r.squared, "\n\n")

  par(mfrow = c(2, 2))
  plot(fit_full_clean, which = 1:4)

  sensitivity_comparison <- data.frame(
    Model = c("Original full model", "Refit after removing flagged points"),
    N = c(nrow(df), nrow(df_clean)),
    AIC = c(AIC(fit_full), AIC(fit_full_clean)),
    BIC = c(BIC(fit_full), BIC(fit_full_clean)),
    R_squared = c(summary(fit_full)$r.squared, summary(fit_full_clean)$r.squared),
    Adj_R_squared = c(summary(fit_full)$adj.r.squared, summary(fit_full_clean)$adj.r.squared),
    Residual_SE = c(summary(fit_full)$sigma, summary(fit_full_clean)$sigma)
  )

  sensitivity_comparison[, -1] <- round(sensitivity_comparison[, -1], 4)

  cat("SENSITIVITY COMPARISON TABLE\n")
  print(sensitivity_comparison)
  cat("\n")
}
```

```
## REFIT MODEL AFTER REMOVAL
##
## Call:
## lm(formula = log_Tc ~ wtd_entropy_Valence + mean_ElectronAffinity +
##     wtd_gmean_atomic_radius + mean_ThermalConductivity + range_Valence,
##     data = df_clean)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.219506 -0.060391 -0.007492  0.044004  0.253111
##
## Coefficients:
##                             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)                4.2483836  0.0505759  84.000 < 0.0000000000000002 ***
## wtd_entropy_Valence        0.3183607  0.0328322   9.697 < 0.0000000000000002 ***
## mean_ElectronAffinity     -0.0005002  0.0002094  -2.388               0.0172 *
## wtd_gmean_atomic_radius   -0.0005659  0.0001443  -3.922    0.000095271682454 ***
## mean_ThermalConductivity   0.0008964  0.0001773   5.057    0.000000529911919 ***
## range_Valence2            -0.0029602  0.0087402  -0.339               0.7349
## range_Valence3            -0.0560636  0.0083126  -6.744    0.000000000029712 ***
## range_Valence4            -0.0604636  0.0127703  -4.735    0.000002600401972 ***
```

```
## range_Valence5            0.1432058  0.0192261   7.449    0.000000000000248 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08309 on 789 degrees of freedom
## Multiple R-squared:  0.2472, Adjusted R-squared:  0.2396
## F-statistic: 32.39 on 8 and 789 DF,  p-value: < 0.00000000000000022
##
##
## AIC cleaned model: -1695.07
## BIC cleaned model: -1648.249
## Adjusted R-squared cleaned model: 0.2395686
```




```
## SENSITIVITY COMPARISON TABLE
##                                 Model   N       AIC       BIC R_squared
## 1                 Original full model 813 -1661.363 -1614.356    0.1965
## 2 Refit after removing flagged points 798 -1695.070 -1648.249    0.2472
##   Adj_R_squared Residual_SE
## 1        0.1885      0.0865
## 2        0.2396      0.0831
```

```r
# Final model diagnostics

fit_final <- fit_aic_backward

cat("FINAL CHOSEN MODEL\n")
```

```
## FINAL CHOSEN MODEL
```

```r
print(summary(fit_final))
```

```
##
## Call:
## lm(formula = log_Tc ~ wtd_entropy_Valence + mean_ElectronAffinity +
##     wtd_gmean_atomic_radius + mean_ThermalConductivity + range_Valence,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23536 -0.06136 -0.00827  0.05294  0.32731
##
## Coefficients:
##                            Estimate Std. Error t value             Pr(>|t|)
## (Intercept)               4.2547066  0.0513988  82.778 < 0.0000000000000002 ***
## wtd_entropy_Valence       0.3146948  0.0330978   9.508 < 0.0000000000000002 ***
## mean_ElectronAffinity    -0.0005379  0.0002154  -2.497               0.0127 *
## wtd_gmean_atomic_radius  -0.0006198  0.0001480  -4.188       0.000031198767 ***
## mean_ThermalConductivity  0.0009481  0.0001794   5.284       0.000000163182 ***
## range_Valence2           -0.0021045  0.0090827  -0.232               0.8168
## range_Valence3           -0.0531612  0.0085628  -6.208       0.000000000857 ***
## range_Valence4           -0.0537417  0.0128090  -4.196       0.000030244718 ***
## range_Valence5            0.0824111  0.0170094   4.845       0.000001518918 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08652 on 804 degrees of freedom
## Multiple R-squared:  0.1965, Adjusted R-squared:  0.1885
## F-statistic: 24.58 on 8 and 804 DF,  p-value: < 0.00000000000000022
```

```
cat("\nFINAL MODEL TYPE II ANOVA\n")
```

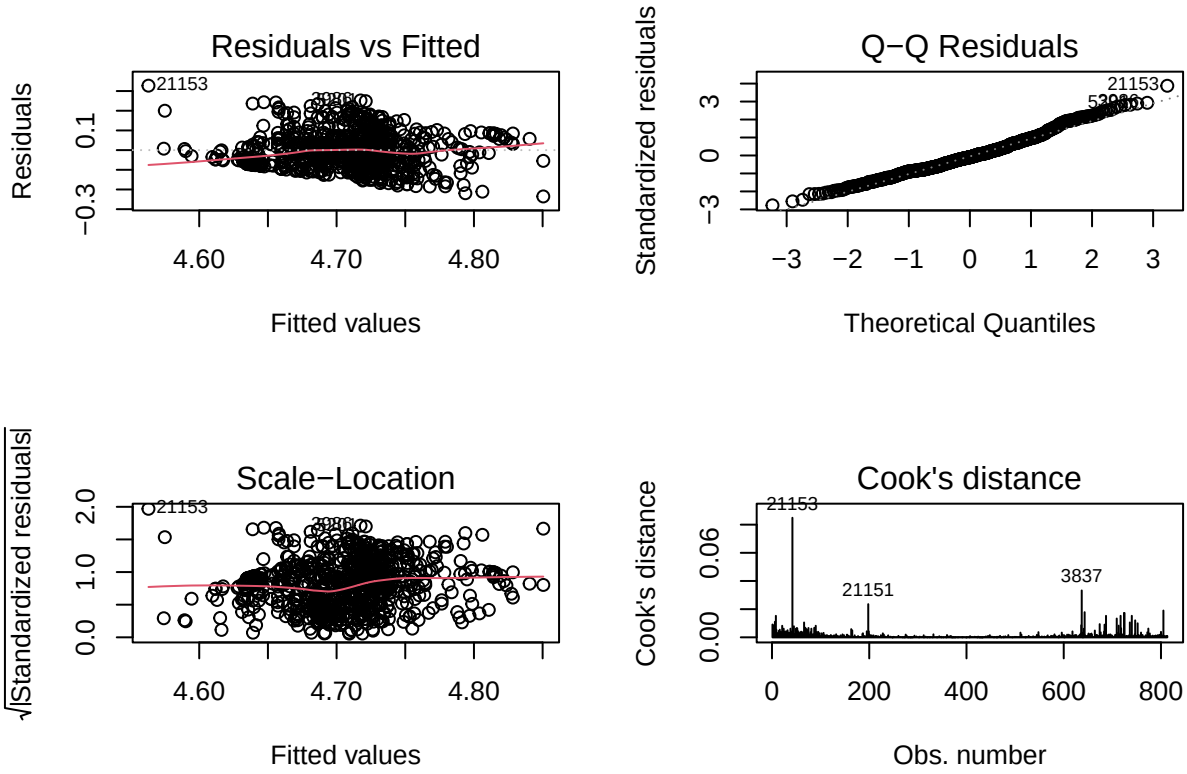
```
##
## FINAL MODEL TYPE II ANOVA
```

```
print(Anova(fit_final, type = 2))
```

```
## Anova Table (Type II tests)
##
## Response: log_Tc
##                           Sum Sq  Df F value                Pr(>F)
## wtd_entropy_Valence       0.6767   1 90.4026 < 0.00000000000000022 ***
## mean_ElectronAffinity     0.0467   1  6.2347               0.01273 *
## wtd_gmean_atomic_radius   0.1313   1 17.5428          0.0000311988 ***
## mean_ThermalConductivity  0.2090   1 27.9172          0.0000001632 ***
## range_Valence             0.6302   4 21.0497 < 0.00000000000000022 ***
## Residuals                 6.0180 804
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("\n")
```

```
par(mfrow = c(2, 2))
plot(fit_final, which = 1:4)
```




```
# Final model coefficient table

coef_mat <- summary(fit_final)$coefficients
ci_mat <- confint(fit_final)

final_model_table <- data.frame(
  Term = rownames(coef_mat),
  Estimate = coef_mat[, "Estimate"],
  Std_Error = coef_mat[, "Std. Error"],
  t_value = coef_mat[, "t value"],
  p_value = coef_mat[, "Pr(>|t|)"],
  CI_Lower_95 = ci_mat[, 1],
  CI_Upper_95 = ci_mat[, 2],
  row.names = NULL
)

final_model_table[, -1] <- lapply(final_model_table[, -1], function(x) round(x, 6))

cat("FINAL MODEL COEFFICIENT TABLE\n")
```

```
## FINAL MODEL COEFFICIENT TABLE
```

```
print(final_model_table)
```

```
##                       Term  Estimate Std_Error   t_value  p_value CI_Lower_95
## 1              (Intercept)  4.254707  0.051399 82.778374 0.000000    4.153815
## 2      wtd_entropy_Valence  0.314695  0.033098  9.508028 0.000000    0.249726
## 3    mean_ElectronAffinity -0.000538  0.000215 -2.496946 0.012726   -0.000961
## 4  wtd_gmean_atomic_radius -0.000620  0.000148 -4.188409 0.000031   -0.000910
## 5 mean_ThermalConductivity  0.000948  0.000179  5.283677 0.000000    0.000596
## 6           range_Valence2 -0.002105  0.009083 -0.231704 0.816827   -0.019933
```

```
## 7           range_Valence3 -0.053161  0.008563 -6.208364 0.000000   -0.069969
## 8           range_Valence4 -0.053742  0.012809 -4.195608 0.000030   -0.078885
## 9           range_Valence5  0.082411  0.017009  4.845039 0.000002    0.049023
##   CI_Upper_95
## 1    4.355598
## 2    0.379663
## 3   -0.000115
## 4   -0.000329
## 5    0.001300
## 6    0.015724
## 7   -0.036353
## 8   -0.028599
## 9    0.115799
```

```
cat("\n")
```

```
# Final model performance metrics

final_metrics <- data.frame(
  Metric = c(
    "R-squared",
    "Adjusted R-squared",
    "Residual standard error",
    "AIC",
    "BIC",
    "F-statistic",
    "Model p-value",
    "Sample size"
  ),
  Value = c(
    summary(fit_final)$r.squared,
    summary(fit_final)$adj.r.squared,
    summary(fit_final)$sigma,
    AIC(fit_final),
    BIC(fit_final),
    unname(summary(fit_final)$fstatistic[1]),
    pf(
      summary(fit_final)$fstatistic[1],
      summary(fit_final)$fstatistic[2],
      summary(fit_final)$fstatistic[3],
      lower.tail = FALSE
    ),
    nobs(fit_final)
  )
)

final_metrics$Value <- round(final_metrics$Value, 6)

cat("FINAL MODEL PERFORMANCE METRICS\n")
```

```
## FINAL MODEL PERFORMANCE METRICS
```

```
print(final_metrics)
```

```
##                    Metric        Value
## 1               R-squared     0.196501
```

```
## 2      Adjusted R-squared     0.188506
## 3 Residual standard error     0.086516
## 4                     AIC -1661.363335
## 5                     BIC -1614.356024
## 6             F-statistic    24.577900
## 7           Model p-value     0.000000
## 8             Sample size   813.000000
```

```
cat("\n")
```

```
# Back-transformed effect multipliers
#     For log(Tc + 1), exp(beta) is the
#     multiplicative change in (Tc + 1)
#     for a 1-unit increase in predictor

effect_table <- data.frame(
  Term = rownames(coef_mat),
  Estimate = coef_mat[, "Estimate"],
  Multiplier_on_Tc_plus_1 = exp(coef_mat[, "Estimate"]),
  Percent_Change_in_Tc_plus_1 = 100 * (exp(coef_mat[, "Estimate"]) - 1),
  row.names = NULL
)

effect_table[, -1] <- lapply(effect_table[, -1], function(x) round(x, 6))

cat("BACK-TRANSFORMED EFFECT TABLE\n")
```

```
## BACK-TRANSFORMED EFFECT TABLE
```

```
print(effect_table)
```

```
##                      Term  Estimate Multiplier_on_Tc_plus_1
## 1             (Intercept)  4.254707               70.436150
## 2      wtd_entropy_Valence  0.314695                1.369841
## 3    mean_ElectronAffinity -0.000538                0.999462
## 4  wtd_gmean_atomic_radius -0.000620                0.999380
## 5 mean_ThermalConductivity  0.000948                1.000949
## 6           range_Valence2 -0.002105                0.997898
## 7           range_Valence3 -0.053161                0.948227
## 8           range_Valence4 -0.053742                0.947677
## 9           range_Valence5  0.082411                1.085902
##   Percent_Change_in_Tc_plus_1
## 1                6943.615047
## 2                  36.984116
## 3                  -0.053776
## 4                  -0.061965
## 5                   0.094856
## 6                  -0.210229
## 7                  -5.177281
## 8                  -5.232318
## 9                   8.590213
```

```
cat("\n")
```

```
# 17. Human-readable interpretation helper

interpretation_table <- data.frame(
```

```
  Term = c(
    "wtd_entropy_Valence",
    "mean_ElectronAffinity",
    "wtd_gmean_atomic_radius",
    "mean_ThermalConductivity",
    "range_Valence2",
    "range_Valence3",
    "range_Valence4",
    "range_Valence5"
  ),
  Direction = c(
    ifelse(coef(fit_final)["wtd_entropy_Valence"] > 0, "Positive", "Negative"),
    ifelse(coef(fit_final)["mean_ElectronAffinity"] > 0, "Positive", "Negative"),
    ifelse(coef(fit_final)["wtd_gmean_atomic_radius"] > 0, "Positive", "Negative"),
    ifelse(coef(fit_final)["mean_ThermalConductivity"] > 0, "Positive", "Negative"),
    ifelse(coef(fit_final)["range_Valence2"] > 0, "Positive", "Negative"),
    ifelse(coef(fit_final)["range_Valence3"] > 0, "Positive", "Negative"),
    ifelse(coef(fit_final)["range_Valence4"] > 0, "Positive", "Negative"),
    ifelse(coef(fit_final)["range_Valence5"] > 0, "Positive", "Negative")
  ),
  Significant_5pct = c(
    coef_mat["wtd_entropy_Valence", "Pr(>|t|)"] < 0.05,
    coef_mat["mean_ElectronAffinity", "Pr(>|t|)"] < 0.05,
    coef_mat["wtd_gmean_atomic_radius", "Pr(>|t|)"] < 0.05,
    coef_mat["mean_ThermalConductivity", "Pr(>|t|)"] < 0.05,
    coef_mat["range_Valence2", "Pr(>|t|)"] < 0.05,
    coef_mat["range_Valence3", "Pr(>|t|)"] < 0.05,
    coef_mat["range_Valence4", "Pr(>|t|)"] < 0.05,
    coef_mat["range_Valence5", "Pr(>|t|)"] < 0.05
  )
)

cat("INTERPRETATION HELPER TABLE\n")
```

```
## INTERPRETATION HELPER TABLE
```

```
print(interpretation_table)
```

```
##                                           Term Direction Significant_5pct
## wtd_entropy_Valence           wtd_entropy_Valence  Positive             TRUE
## mean_ElectronAffinity       mean_ElectronAffinity  Negative             TRUE
## wtd_gmean_atomic_radius   wtd_gmean_atomic_radius  Negative             TRUE
## mean_ThermalConductivity mean_ThermalConductivity  Positive             TRUE
## range_Valence2                     range_Valence2  Negative            FALSE
## range_Valence3                     range_Valence3  Negative             TRUE
## range_Valence4                     range_Valence4  Negative             TRUE
## range_Valence5                     range_Valence5  Positive             TRUE
```

```
cat("\n")
```

```
# Residual summaries for final writeup

resid_summary <- data.frame(
  Statistic = c("Min", "Q1", "Median", "Mean", "Q3", "Max"),
  Residual = round(c(
```

```
    min(resid(fit_final)),
    quantile(resid(fit_final), 0.25),
    median(resid(fit_final)),
    mean(resid(fit_final)),
    quantile(resid(fit_final), 0.75),
    max(resid(fit_final))
  ), 6),
  row.names = NULL
)

cat("FINAL MODEL RESIDUAL SUMMARY\n")
```

```
## FINAL MODEL RESIDUAL SUMMARY
```

```
print(resid_summary)
```

```
##   Statistic  Residual
## 1       Min -0.235355
## 2        Q1 -0.061359
## 3    Median -0.008269
## 4      Mean  0.000000
## 5        Q3  0.052936
## 6       Max  0.327307
```

```
cat("\n")
```