Topic:
Exploring the Factors Affecting the Higher Heating Value of Low Rank Coals Based on Multiple Linear Regression.

Group Members:
Tingjun Kang(tk3041), Linke Wu(lw3106), Yuhao Wang(yw3924)

Introduction:
In recent years, global warming has emerged as an issue that cannot be overlooked. The primary factor contributing to global warming is the emission of greenhouse gases (CO_2, $N_2O\ldots$). The world now faces a paradigm shift towards decarbonization, a transition intimately linked to our economy, environment, and collective fate. Reducing carbon emissions has become an urgent priority for numerous countries. This concern is highly relevant to the efficiency of coal's thermal output. Enhanced efficiency in coal's thermal output means less coal is wasted. Consequently, this reduction in coal usage will decrease overall global greenhouse gas emissions. Therefore, we believe it is imperative to investigate the factors influencing the calorific value of coal. Understanding these factors is vital for our ability to control environmental change and enhance productivity.

Research Question:
While reviewing journal articles in the field of fuel processing technology, we encountered several papers that closely align with our core concept. These studies focused on the higher heating values (HHV) of low-rank coals, which are those that have undergone minimal metamorphism during formation. The HHV is a crucial aspect of calorific value, which is a physical measure of the heat released by a unit mass or volume of fuel when it burns completely. The findings of these papers are pertinent to our inquiry as they identify key factors influencing the HHV of coal and provide valuable data. Consequently, we intend to analyze whether these factors significantly impact the calorific value by constructing multiple linear models.

Data Collection:
We sourced a relevant open-source dataset from the website below. Data describes the relationship between higher heating value **HHV** (Y, MJ/kg) and 4 predictors: **Moisture, Ash, Volatile Matter,** and **Fixed Carbon Rate** for 50 samples of coal. This dataset will be instrumental in our analysis, allowing us to apply the principles of multivariate linear regression to the factors influencing the calorific value (HHV) of coal.

Website:
http://users.stat.ufl.edu/~winner/datasets.html
https://www.sciencedirect.com/science/article/pii/S0378382008002269?via%3Dihub

Data Description:
1. General description:

The dataset comprises 51 rows and 6 columns. Excluding the first row, which likely serves as a header, the remaining 50 rows represent 50 distinct types of coal. Each column corresponds to different properties of the coal, providing a comprehensive overview of various characteristics that could influence its calorific value.

2. Columns description:
   - **No.coal:** This is a unique identifier for each coal sample. It is a numerical column likely ranging from 1 to 50, given there are 50 samples.

   - **Moisture_wt:** Represents the moisture content in the coal samples, expressed as a percentage by weight. The values vary, for example, from 47.0 to 43.6 in the first few entries, indicating the diversity in moisture content across different coal samples.

   - **Ash_wt:** This column indicates the ash content in the coal, also expressed as a percentage by weight. Similar to moisture, this value varies among samples, as seen in the range from 7.3 to 13.9 in the first few samples.

   - **Volatile_Matter_wt:** Reflects the amount of volatile matter in the coal samples, again as a percentage by weight. This is an essential factor in coal quality, with values like 25.8 and 23.1 in the initial entries.

   - **Fixed_Carbon_wt:** This measures the fixed carbon rate in the coal samples, a critical component for combustion processes. It is also expressed as a percentage by weight.

   - **HHV (Higher Heating Value):** The primary variable of interest, HHV, is measured in Mega Joules per kilogram (MJ/kg). It represents the amount of energy released during combustion. The values in this column will be the focus of analysis to determine how they correlate with moisture, ash, volatile matter, and fixed carbon content.

Exploratory data analysis:

1. Numerical Summaries:

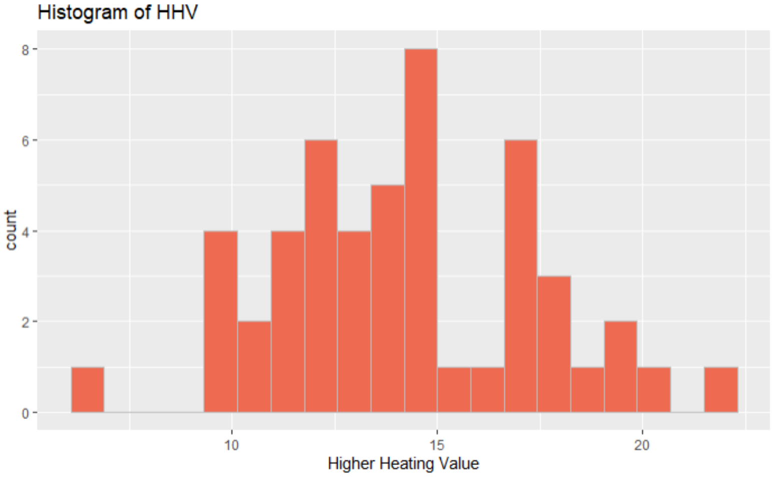
Table0: Basic Numerical Analysis of Statistical Data

Table	Variables	Min value	Max value	Median	Mean	IQR	Standard deviation
Explanatory variable	Moisture (wt.%)	6.1	47	25.7	26.61	15.3	10.42
	Ash(wt.%)	5.8	48	17.7	20.35	11.95	10.44
	Volatile Matter (wt.%)	8.9	35	26.25	26.15	6.625	5.39
	Fixed Carbon (wt.%)	8.9	42.7	26.3	26.93	10.225	7.02
Response variable	HHV (Mj/Kg)	6.4	21.81	14.14	14.21	4.6525	3.16

This table was generated from calculations performed in RStudio. It presents basic data information, highlighting that the moisture content ranges from 6.1 to 47. The table also clearly delineates the ranges for other variables. The significant differences between the minimum and maximum values indicate substantial attribute variations among different coals. Furthermore, the median, mean, and interquartile range (IQR) for various characteristics of different coals have been calculated. For variables such as 'Volatile Matter' and 'High Heating Value (HHV),' the median and mean values are notably close, with a difference of less than or equal to 0.1. The standard deviation for these data points has also been computed. The smallest standard deviation, approximately 3.16, is observed in HHV, suggesting minor variations in HHV among different coals. In contrast, the standard deviation for Ash is the highest, around 10.44, indicating significant variability in the ash content produced after the combustion of different coals.

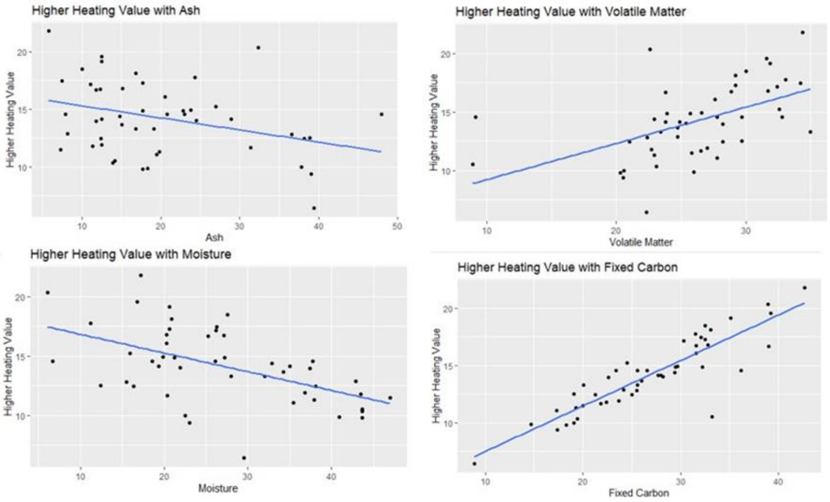
2. Graphical Summaries:

Graph: Histogram of Higher Heating Value (HHV)



The figure above displays a histogram of the Higher Heating Value (HHV), which also serves as the response variable in the analysis. It is crucial to draw this graph to ensure that the response variable approximates a normal distribution, a necessary condition for the validity of subsequent linear regression analysis. This histogram exhibits a symmetric skew and is unimodal, indicating a single peak in its distribution. The majority of HHV values are concentrated around 13.5 MJ/Kg, marking the central value of the histogram. This central value is represented by approximately 8 coal samples. Overall, the HHV of all the analyzed coal samples ranges between about 6.5 and 21.5 MJ/Kg.

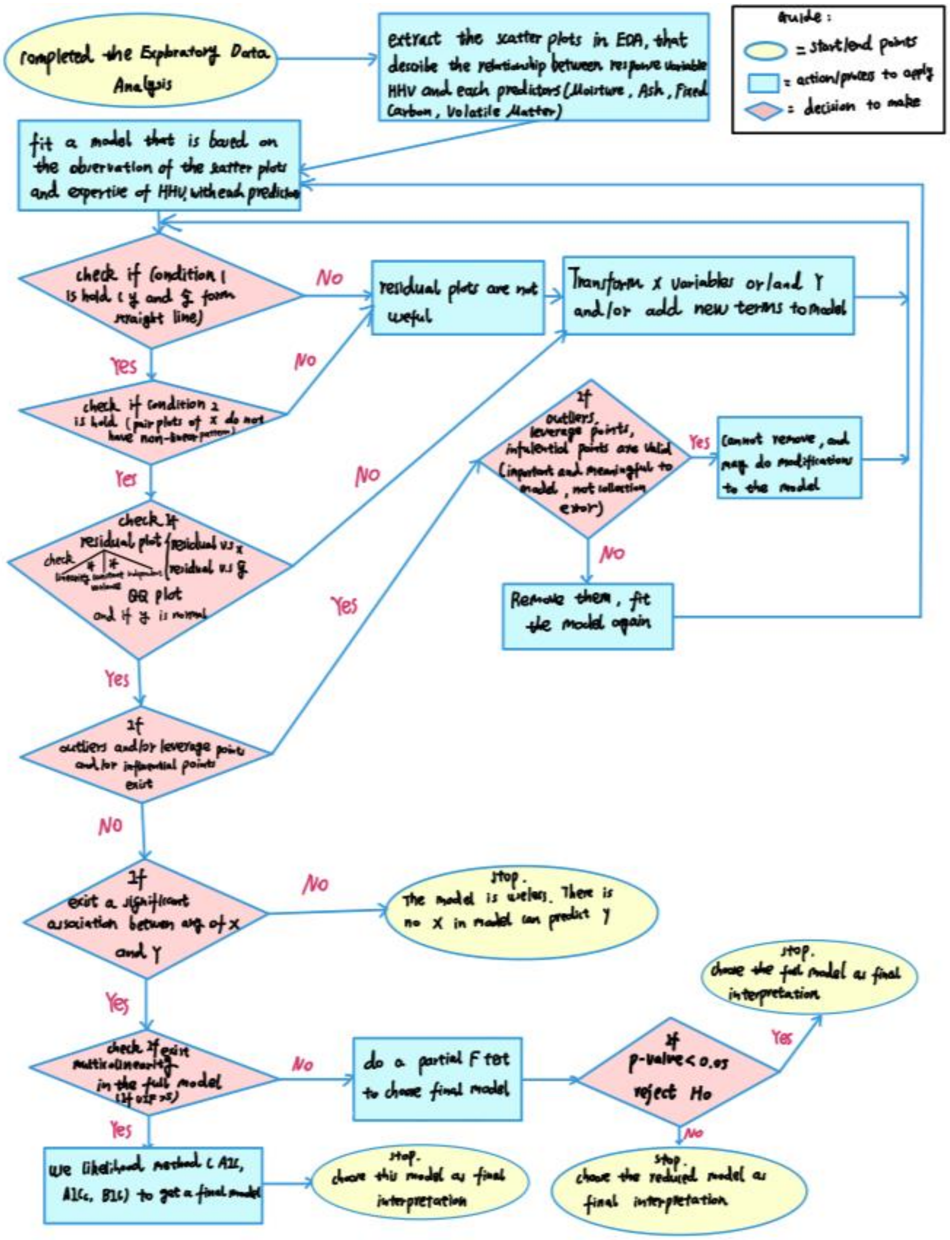
Graph: Scatterplots of HHV and Four Predictors



The scatterplots illustrate the relationship between the response variable (HHV) and other explanatory variables. It is evident that these four variables exhibit a linear relationship with HHV. Fixed Carbon and Volatile Matter demonstrate a positive correlation with HHV, while Moisture and Ash exhibit a negative correlation. Additionally, all these figures display constant variance with respect to HHV.

Method: Multiple Linear Regression and Model Selection (AIC, BIC…etc.)

After exploratory data analysis(EDA), We found that the Higher Heating Value (HHV) is approximately normally distributed, satisfying the normality assumption of multiple linear regression. In our project, we will apply multiple linear regression models to determine statistically significant factors affecting the HHV. Utilizing model selection methods, we aim to identify the most effective model, thereby gaining insights into optimizing coal's calorific value. We also built a flow chart to illustrate how to fit the model and to conduct model selection.


Guide :
= start/end points
= action/process to apply
= decision to make
completed the Exploratory Data Analysis
extract the scatter plots in EDA, that descibe the relationship between response variable HHV and each predictors (Moisture, Ash, Fixed Carbon, Volatile Matter)
fit a model that is based on the observation of the scatter plots and expertise of HHV, with each predictor
check if condition 1 is hold (y and ŷ form straight line)
No
residual plots are not useful
Transform X variables or/and Y and/or add new terms to model
Yes
No
check if condition 2 is hold (pair plots of x do not have non-linear pattern)
Yes
check if residual plot {residual v.s x, residual v.s ŷ; check linearity, constant variance, independent; QQ plot and if ŷ is normal
No
Yes
If outliers, leverage points, influential points are valid (important and meaningful to model, not collection error)
Yes
cannot remove, and may do modifications to the model
No
Remove them, fit the model again
If outliers and/or leverage points and/or influential points exist
No
If exist a significant association between any of X and Y
No
stop. The model is useless. There is no X in model can predict Y
Yes
check if exist multicollinearity in the full model (if VIF > 5)
No
do a partial F test to chose final model
If p-value < 0.05 reject Ho
Yes
stop. chose the full model as final interpretation
No
stop. chose the reduced model as final interpretation
Yes
use likelihood method (AIC, AICc, BIC) to get a final model
stop. chose this model as final interpretation

Frist, we randomly allocate 80% of the dataset to a training set, used for model fitting (training model). The remaining 20% forms the testing set, employed for model evaluation. After refitting the model with the testing set, we obtain a testing model. For model validation, we first ensure minimal difference in the estimated regression coefficients between the training and testing models. Additionally, the same predictors should be significant in both models. We also aim for a similar adjusted R-squared between the two models. Finally, it's important that no new violations appear in the testing model.

We then establishment of models through both manual and automatic selection, including the application of stepwise selection in R studio.

- For manual selection model (Model_1):

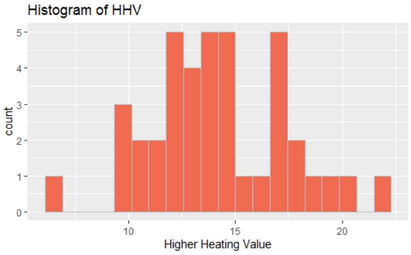
We manually removed some insignificant predictors with P-values larger than 0.05(i.e. T-Test of $H_0: \beta_i = 0$), as indicated in the summary of our initial model. Furthermore, we checked for multicollinearity among the predictors by calculating the Variance Inflation Factor (VIF). To assess multicollinearity, we calculated the VIF for each predictor in R Studio. If a predictor's VIF was greater than 5, it was considered for removal. It is important to remove only one predictor at a time, and then recalculate the VIF for the new model. Based on the updated VIF, we then decided whether to remove another predictor. This process helped us in building a model through the manual selection of variables.

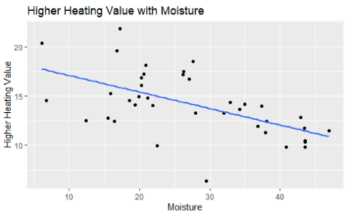
- For automated selection model (Model_2):

In addition, we also have a method for building an automated selection model. We perform stepwise selection on our initial model in R Studio. This stepwise selection process provides a systematic way to select a model from a large set of predictors. It is designed to choose the model with the lowest AIC/AICc or BIC, which are likelihood methods that estimate the quality of each model relative to the others. Models with smaller AIC/AICc or BIC values are preferred. Therefore, by applying the methods mentioned above, we can develop two types of models: one manually selected and the other automatically selected.

After building the manually selected model and the automated selected model for the training set, it's essential to check for model violations and diagnostics. Firstly, we examine the histogram of the response variable to verify its normality. For checking multicollinearity, we draw pair plots of predictors for both models to check if either predictors have high linear correlation. Next, we plot Residual vs. Predictor and Residuals vs. Fitted Values to confirm the absence of discernible patterns, thereby validating linearity, uncorrelated errors, and constant variance. Furthermore, we use a Q-Q plot to recheck normality, looking for a straight diagonal line of points with small deviations at the ends. Last, we need to decide whether to remove leverage points, outliers, and influential points in both models.

Results section:





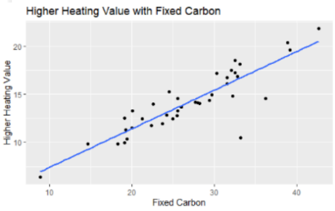


Figure: EDA for Training set

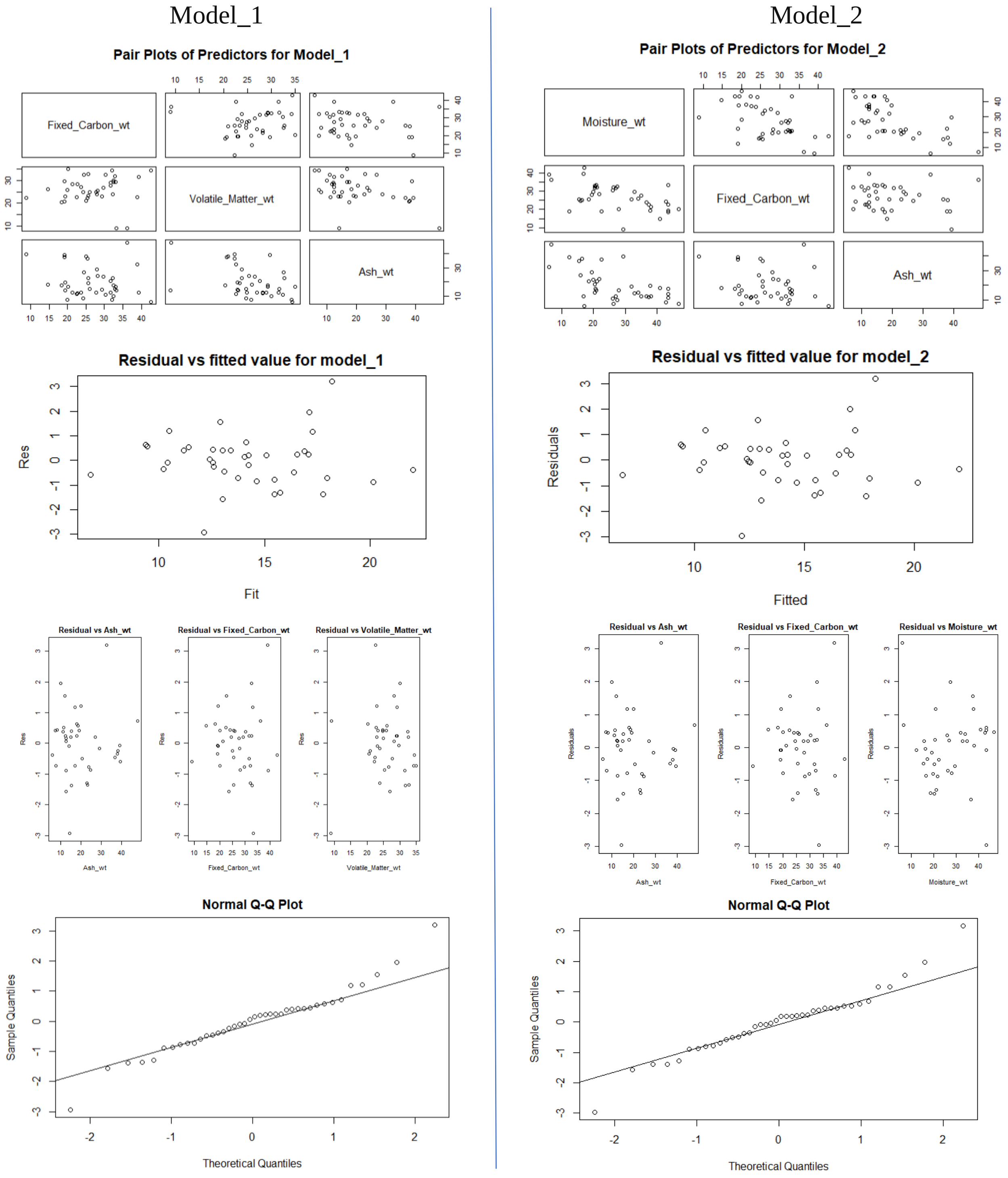
We selected some figures from the Exploratory Data Analysis (EDA) of the training set, which demonstrated that our response variable HHV has good normality. The scatterplots revealed that four predictors - Moisture, Ash, Volatile Matter, and Fixed Carbon - have a linear relationship with HHV. In a later stage, we also conducted EDA for the testing set and, fortunately, obtained similar images and conclusions (refer to Appendix 1). This consistency indicates that the dataset is highly suitable.

Table1: Comparing Two Models

Manual selected model (Model_1)		Automated selected model (Model_2)	
Variable:	VIF	variable	VIF
Volatile Matter	1.24	Moisture	4.65
Fixed Carbon	1.04	Fixed Carbon	3.06
Ash	1.27	Ash	3.45
Likelihood methods:	A:Training B:Testing	Likelihood methods	A:Training B:Testing
Adj.R^2	A: 0.9496 B: 0.9619	Adj.R^2	A: 0.9497 B: 0.9577
AIC	A: -24.41 B: -2.09	AIC	A: -24.50 B: -1.37
AICc	A: -22.74 B: 17.91	AICc	A: -22.84 B: 18.63
BIC	A: -11.97 B: 1.63	BIC	A: -12.06 B: 2.36

After we addressed the multicollinearity in the initial model, we obtained the manually selected model. This model demonstrates that all predictors' Variance Inflation Factors (VIFs) are lower than 5 and relatively small. Subsequently, applying stepwise selection, we obtained an automated selected model. Although all predictors' VIFs in this model are also smaller than 5, they are larger than those in the manually selected model. We also noted that the automated model's adjusted R-squared is higher than that of the manual model in the training set, and its AIC, AICc, and BIC are lower in the training set. However, after refitting with the testing set, these values showed the opposite results. Despite the automated model's smaller AIC, BIC, and AICc, the higher VIFs raise concerns. Therefore, through this comparison, we cannot definitively determine which model is superior. Further testing is required to discern this.

Figure: Model Violations and Diagnostics for Training Set



Additionally, we processed Model Diagnostics by creating relevant graphs for both Model_1 and Model_2. All images from these tests passed the criteria for model violations. This indicates that both models satisfy normality, linearity, uncorrelated errors, and constant variance. Therefore, no transformation is necessary. We also performed similar checks for the testing set (refer to

Appendix 2), which upheld all assumptions. Although the Q-Q plot of Model_1 suggests that some points may not slightly better satisfy normality, these observations are still insufficient to conclusively determine our final model.

Table2: Leverage points, outliers, influential points of two models

Item	Leverage points (No.coal)	outliers (No.coal)	Influential points (No.coal)
Model_1(trainng)	5,36,41	5,28	5
Model_2(trainng)	5,36,41	5,28	5
Model_1(testing)	N/A	N/A	3
Model_2(testing)	N/A	N/A	3

After identifying the leverage points, outliers, and influential points of both models in the training and testing sets, we discovered that these points are the same in both sets. Furthermore, upon examining these points in the dataset, we decided not to remove them. We realized that they represent different types of coal, each with distinct properties. Their presence is not a result of inherent data issues or errors in data collection.

Table3: Estimated regression coefficients and significant

Estimated Regression Coefficients	Intercept	Ash	Fixed Carbon	Moisture	Volatile Matter
Model_1(training)	-3.19 ***	0.02 .	0.38 ***		0.25 ***
Model_1(testing)	-3.98	0.01	0.34 ***		0.33 **
Model_2(training)	22.11 ***	-0.23 ***	0.13 ***	-0.25 ***	
Model_2(testing)	29.12 ***	-0.32 ***	0.007	-0.33 **	

Table 3 and the adjusted R-squared in Table 1 collectively validate our model. We observed that the estimated regression coefficients are similar between the corresponding training and testing models for each approach. Additionally, the adjusted R-squared values for each pair of models are nearly identical. Moreover, the various residual plots previously created for the testing model show no new model violations, leading us to believe that our model validation is largely successful. However, some predictors in the testing model are no longer significant, which we attribute to limitations in our model. This will be further discussed in the final part of the article.

Thus, our earlier comparisons confirm that there isn't a significant difference between the two models based on the detection results. The only notable observation is that some scatter plots of

Model_1 align more closely with our hypothesis. Therefore, we next used these two models to predict the data in the testing set and calculated the Sum of squared errors (SSE) for each result. The SSE for Model_1 with the prediction data is 3.97, which is smaller than that of Model_2 (4.02). Combining all the aforementioned conclusions and the SSE calculation results, we chose Model_1 as our final model, which is the manually selected model.

Discussion section: Limitations and Improvements

The final model represent in formula is:

$$\hat{y} = -3.197 + 0.023\text{Ash\_wt} + 0.385\text{Fixed\_Carbon\_wt} + 0.252\text{Volatile\_Matter\_wt}$$

Ash_wt represents the weight of ash in the coal, Fixed_Carbon_wt indicates the weight of fixed carbon, and Volatile_Matter_wt is the weight of volatile matter. This suggests that with a one percent increase in the weight of ash, the average of HHV (Higher Heating Value) would increase by 0.023 units. Other variables have similar interpretations. We recognize certain limitations in our model. One significant factor is the size of our dataset, which comprises only 50 data points. Consequently, our testing set contains just 10 data points, potentially introducing considerable variation. Additionally, the distribution of the testing data somewhat differs from that of the training data, which may impact the final model's accuracy. Despite these limitations, we have identified the main factors influencing the HHV of coal. By increasing the proportion of these three variables, we can enhance the HHV of coal, thereby improving the efficiency of coal usage and contributing to environmental betterment.

References:

Dashti, A., Noushabadi, A. S., Raji, M., Razmi, A., Ceylan, S., & Mohammadi, A. H. (2019). Estimation of biomass higher heating value (HHV) based on the proximate analysis: Smart Modeling and Correlation. Fuel, 257, 115931. https://doi.org/10.1016/j.fuel.2019.115931

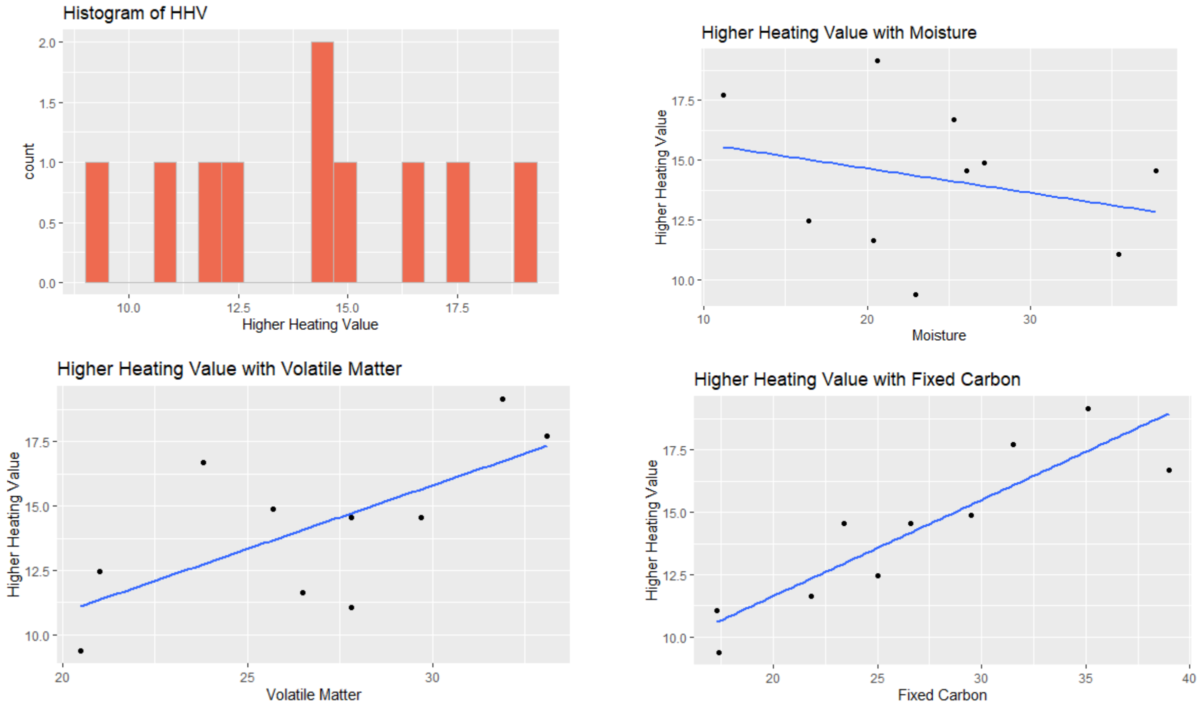
Elliott, D. C. (1980). Decarboxylation as a means of upgrading the heating value of low-rank coals. Fuel, 59(11), 805–806. https://doi.org/10.1016/0016-2361(80)90261-6

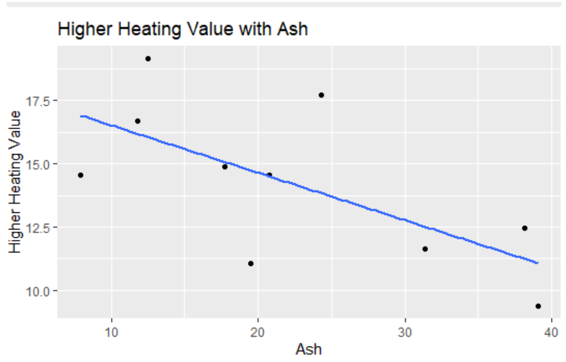
Go, A. W., & Conag, A. T. (2018). A unified semi-empirical model for estimating the higher heating value of coals based on proximate analysis. Combustion Science and Technology, 190(12), 2203–2223. https://doi.org/10.1080/00102202.2018.1497612

Akkaya, A. V. (2009). Proximate analysis based multiple regression models for higher heating value estimation of low rank coals. Fuel Processing Technology, 90(2), 165–170. https://doi.org/10.1016/j.fuproc.2008.08.016

Appendix1:

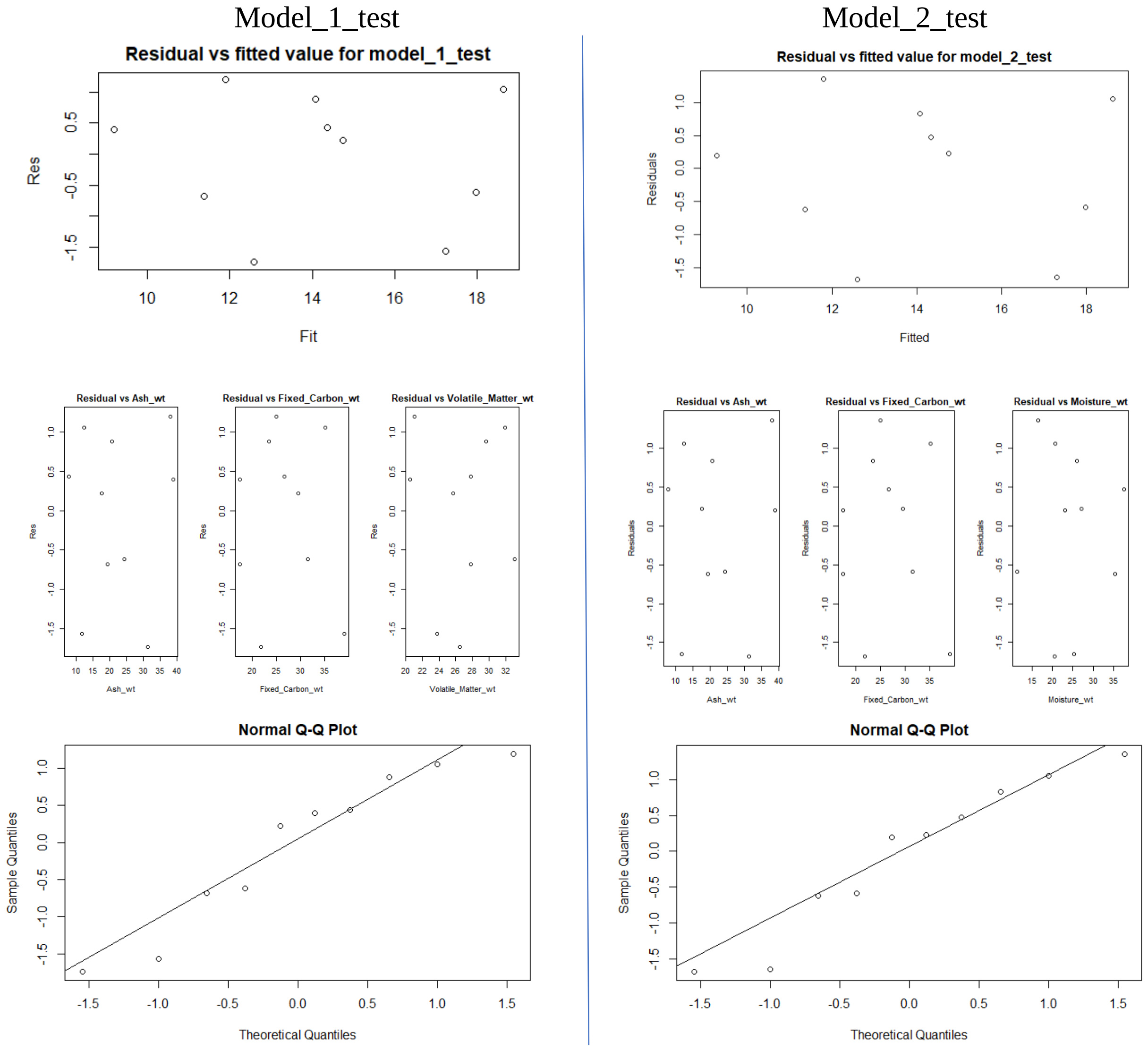
Figure : EDA of Testing Data Set





Apendix2:

Figure: Residual plots of Two testing models.



Model_1_test

Model_2_test