The ANOVA Model

You want to use a model that explains as much of the variability in SalePrice as possible. This mathematical model is a way to represent the relationship between the response and predictor variables in ANOVA.

$$Y_{ik} = \mu + \tau_i + \varepsilon_{ik}$$

Y_{ik} is the response, where k represents the observation number and i indexes the treatments, or in this case, the four heating quality conditions. μ is the overall population mean of the response variable SalePrice, ignoring the heating quality.

τ_i is the effect of each heating quality. Because you have four quality conditions, i has levels 1, 2, 3, or 4. τ_1 is the difference between heating condition one's mean and the overall mean. τ_2 is the difference between heating condition two's mean and the overall mean, and so on.

ε_{ik} is the error term in the model, also known as the unaccounted-for or within group variation. This is a way to represent all variation in SalePrice that hasn't been accounted for due to the different types of heating conditions. Note that this is only one way to parameterize ANOVA model.

The predicted values in ANOVA are the group means, and a residual is the difference between the observed value of the response variable and the predicted value of the response variable. Residuals can be considered observed errors or realizations of the errors. Because residuals are used so frequently when you assess the model assumptions, it's important to understand how they are calculated. Residuals are calculated as the difference between the actual value of the response variable Y and the predicted value of Y from your model.

Because the validity of the p-value depends on the data meeting the assumptions for ANOVA, it's good practice to verify those assumptions in the process of performing the analysis. The first assumption is one of independent observations. Independence implies that the errors, ε_{ik}, in the model are uncorrelated. Good data collection designs can help ensure this assumption.

The second assumption is that the error terms are normally distributed for every group or treatment. The residuals that come from your data are estimates of the error term in the model. Diagnostic plots, including normal quantile-quantile plots and histograms of the residuals, can be used to assess the normality assumption. With a reasonably sized sample and approximately equal groups (or balanced design), only severe departures from normality are considered a problem.

The third assumption is that the error terms have equal variances across treatments. You can use PROC GLM to conduct a formal test for equal variances, and also plot the residuals versus predicted values as a way to graphically verify this assumption. The null hypothesis for this test is that the variances are equal for all populations.