Linear Regression:

The goal of the model created in this project is to predict the compressive strength of concrete based on the components of the mixture. This can allow a user to know in advance what to expect the strength of the concrete to be in order to ensure that first of all the concrete is sufficiently capable of bearing the load it will be put under but can also prevent the use of excess materials creating concrete that is far stronger than what is required for the task minimising waste which can lower the overall cost of a project.

The data used for the model is the Concrete Compressive Strength Data Set from the UCI Machine Learning Repository. This data provided all values in a numerical format so no alteration was necessary on the data values. The labels were reduced to shorthand versions of themselves in the original data set so that they would be easy to access in R for example “Cement (component 1)(kg in a m^3 mixture)” was reduced to Cement for ease of use.

The data was placed in a linear regression model for this analysis and the summary produced by R Studio on the first model using all input variables showed that each of the input variables were statistically significant. The highest Adjusted R-squared value was achieved using this model compared to a model that removes any of the variables.

From the results of the model while it can predict the compressive strength with some degree of accuracy the adjusted r-squared value of the model at just over 0.59 is still quite low for a model to be used in practice as it would introduce a significant amount of errors compared to the actual values. This may indicate that the data would be better suited for another form of modelling which could increase the adjusted r-squared value.

Polynomial Regression:

The goal of the models created during this analysis is to predict the heating and cooling load of a structure based on some the features of said structure. This information will be useful during the planning phase of a building as it will allow the engineers and architects to understand the insulation of the building so they can plan other aspects such as the heating and air conditioning the will be required for the building or adjust some factors in the building if they have certain standards to meet or they wish to exceed for factors such as how environmentally friendly the building is in terms of insulation.

The data set used for the model is the Energy Efficiency Data Set from the UCI machine learning repository. No additional preparation was needed for the data before use. The labels however were in the format of X1-8 for the input variables and Y1 and Y2 for the output variables.

The input variables for this data set were

1. Relative Compactness
2. Surface Area
3. Wall Area
4. Roof Area
5. Overall Height
6. Orientation
7. Glazing Area
8. Glazing Area Distribution

The output variables for this data set were

1. Heating Load
2. Cooling Load

As there are two different output variables that are both relevant to the data set there were two models created. Roof Area provided the highest r-squared value from the data set when using polynomial regression. The results provided a good r squared value at degree 2 but the value could be improved slightly by increasing the degree to 3. The accuracy could be further improved by adding additional input values with a degree of 3. Overall height was not used in the model as this could cause errors when creating the model. For both output variables orientation and glazing area distribution proved to be statistically insignificant so these were removed from the models.

The first model using heating load as the output variable using relative compactness, surface area, wall area, roof area and glazing area as the input variables resulted in and adjusted r-squared value of 0.9896. The second model using cooling load as the output variable and all the input variables remained the same as the first model. The adjusted r squared value for this model was 0.966. The second model explains slightly less of the variation but with both being above .95 they explain the vast majority of the values. When the same values are placed in a linear regression model, the first model has an adjusted r squared value of .8979 and the second model has an adjusted r squared value of .8669. The values show that while linear regression can produce good models with this data set there is significant improvement to be made in the r squared value by using polynomial regression instead. Overall from the results of the two models they are suitable for use in predicting new values based on the results from them.