Predicting a Building's Energy Efficiency

Contents

[1. Introduction 1](#_Toc173798512)

[1.1 overview of the problem 1](#_Toc173798513)

[1.2 description of the data set 2](#_Toc173798514)

[1.3 goal of the project 2](#_Toc173798515)

[1.4 plan of action 2](#_Toc173798516)

[2. Azure Databricks 3](#_Toc173798517)

[3. Data Visualization 3](#_Toc173798518)

[Analysis of Categorical and Numerical Variables 5](#_Toc173798519)

[Examination of the dependent variable 5](#_Toc173798520)

[4. Model (Neural Network) 7](#_Toc173798521)

[5. Results 8](#_Toc173798522)

# 1. Introduction

## 1.1 overview of the problem

With record high and low temperatures across the globe, it is becoming increasingly important to be efficient when it comes to heating and cooling our buildings. Whether you are trying to reduce the cost of your energy bill or you're trying to reduce your carbon footprint, improving the energy efficacy of your building can save you some money and even help the environment. We will look at a data set that can help us with both!

When it comes to efficient building design, the computation of the heating load (HL) and the cooling load (CL) is required to determine the specifications of the heating and cooling equipment needed to maintain comfortable indoor air conditions. To estimate the necessary heating and cooling capacities, architects and building designers need information about the characteristics of the building and the conditioned space (for example, occupancy and activity level). For this reason, we will investigate the effect of eight input variables (RC), surface area, wall area, roof area, overall height, orientation, glazing area, and glazing area distribution, to determine the output variables HL and CL of residential buildings.

We will use R squared (R2 score) to evaluate our model performance. R-squared is a statistical measure of how close the data are to the fitted regression line. It is essential to create predictions that are close to the actual values. In this case, we want to achieve a high R squared; the higher the R squared, the better the model fits the data.

## 1.2 description of the data set

We will use this data set from the University of California, Irvine Machine Learning Repository.

The following is UCI's information on the data set:

We perform energy analysis using 12 different building shapes simulated in Ecotect. The buildings differ concerning the glazing area, the glazing area distribution, and the orientation, among other parameters. We simulate various settings as functions of the aforementioned characteristics to obtain 768 building shapes. The dataset comprises 768 samples and 8 features, aiming to predict two real-valued responses. If the response is rounded to the nearest integer, it can also be used as a multi-class classification problem.

**Dataset Information:**

We perform energy analysis using 12 different building shapes simulated in Ecotect. The buildings differ in terms of the glazing area, the glazing area distribution, and the orientation, amongst other parameters. We simulate various settings as functions of the characteristics above to obtain 768 building shapes. The dataset comprises 768 samples and 8 features, aiming to predict two real-valued responses. If the response is rounded to the nearest integer, it can also be used as a multi-class classification problem.

**Attribute Information:**

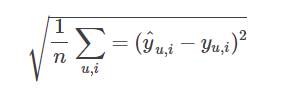
The dataset contains eight attributes (or features, denoted by X1...X8) and two responses (or outcomes, denoted by y1 and y2). The aim is to use the eight features to predict each of the two responses.

* X1 Relative Compactness
* X2 Surface Area
* X3 Wall Area
* X4 Roof Area
* X5 Overall Height
* X6 Orientation
* X7 Glazing Area
* X8 Glazing Area Distribution
* y1 Heating Load
* y2 Cooling Load

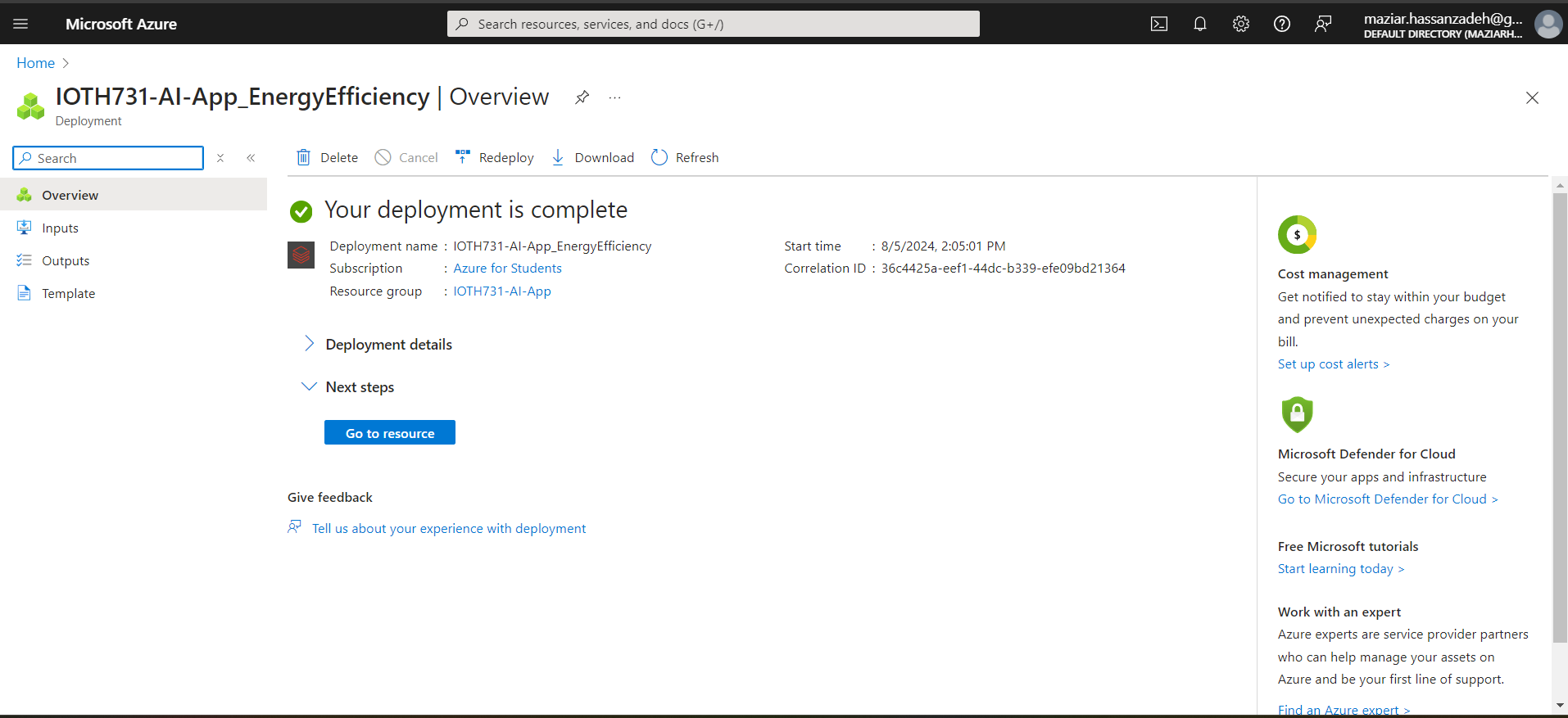
## 1.3 goal of the project

This project aims to create multiple regression models that predict both the heating and cooling loads. We will use the best-performing model.

## 1.4 plan of action

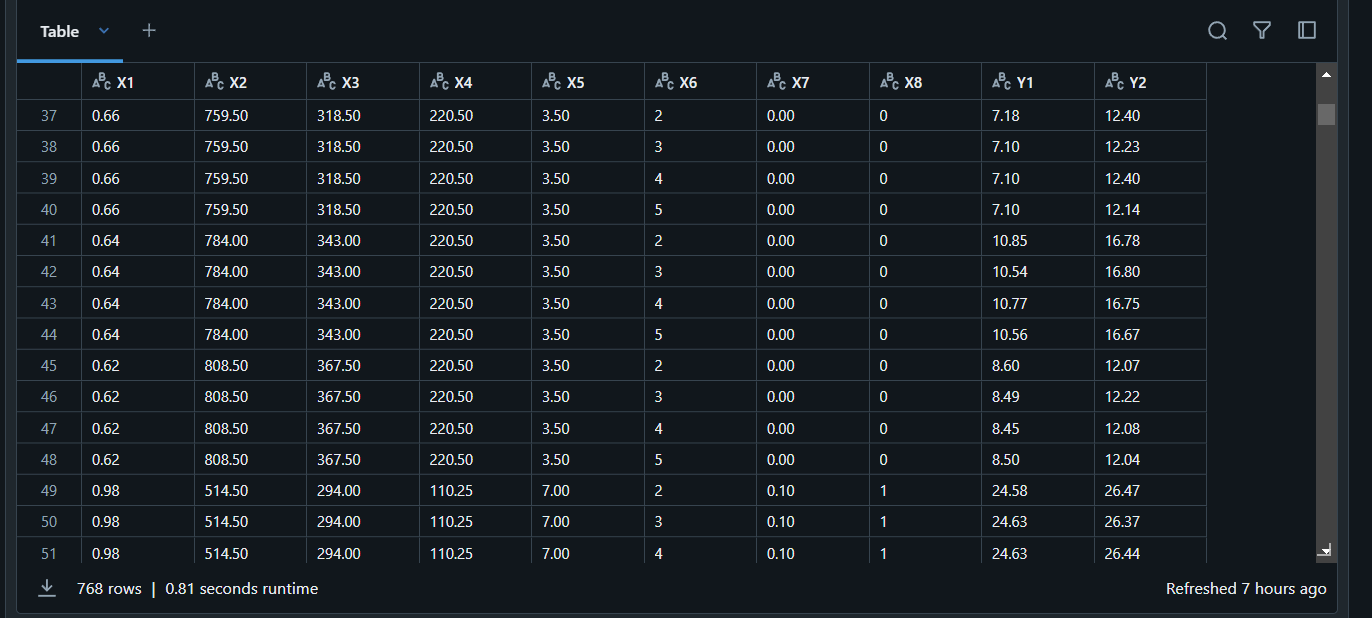
We will download and clean the data. We will use visualization tools to get a better understanding of the data that we are working with. Then, we will create the following regression models: linear regression, ridge regression, random forest, and *ensemble*. The models will be evaluated by using the Root Mean Squared Error (RMSE)

# 2. Azure Databricks

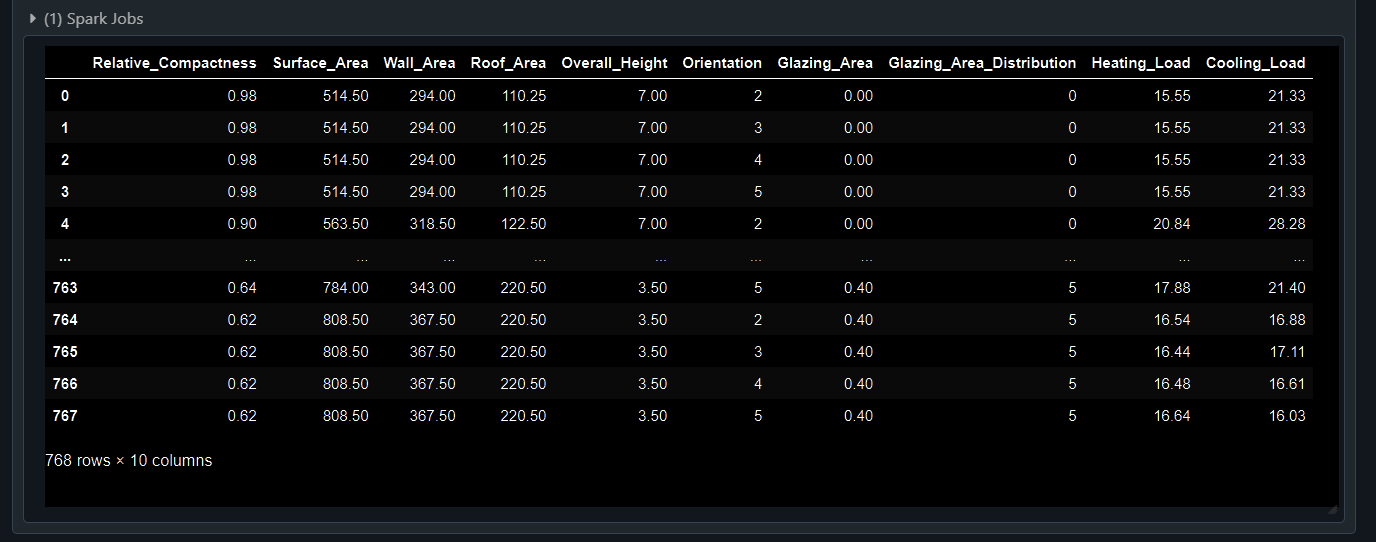


# 3. Data Visualization

Data is in Spark format in the Databricks:

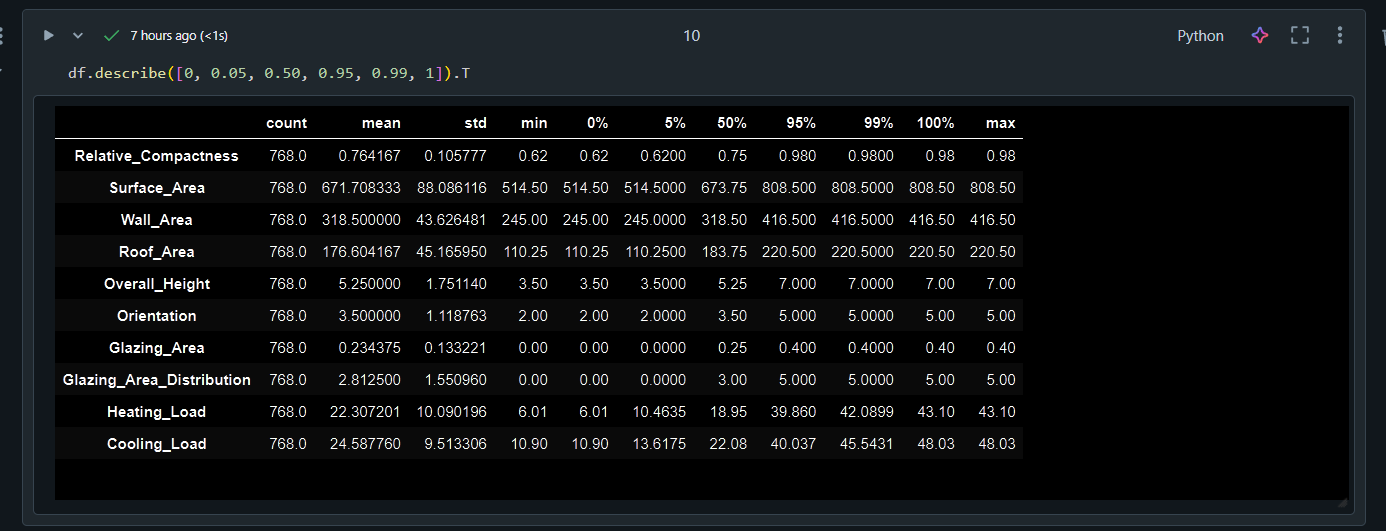


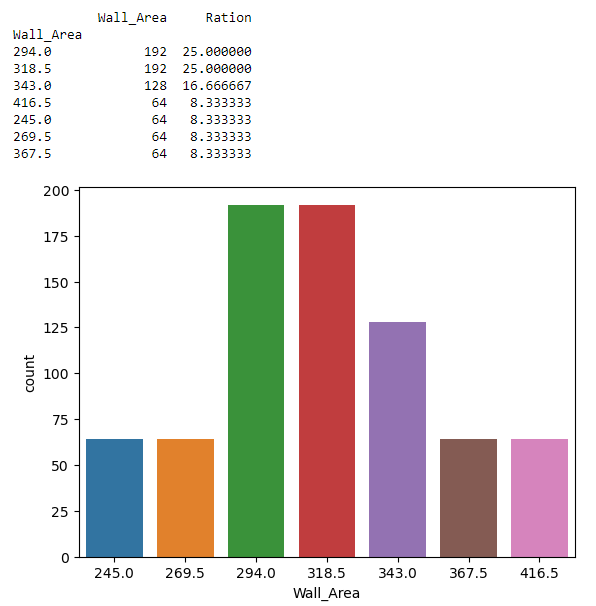
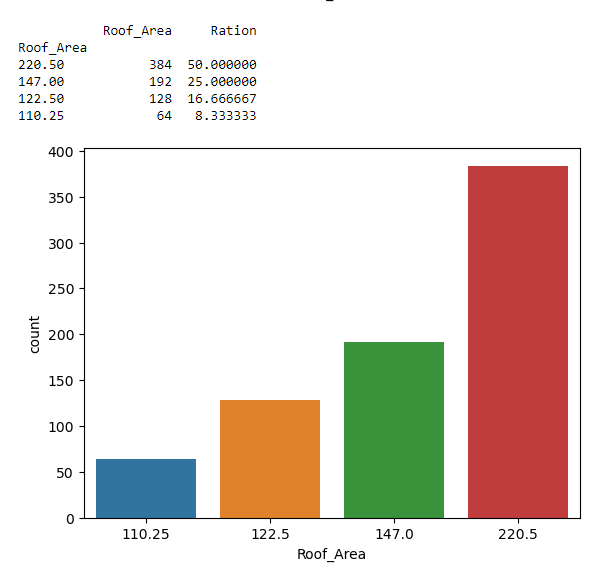
We need to convert data to Pandas DataFrame,   
features’ data types to the original Data Types,  
And, rename the columns to meaningful names



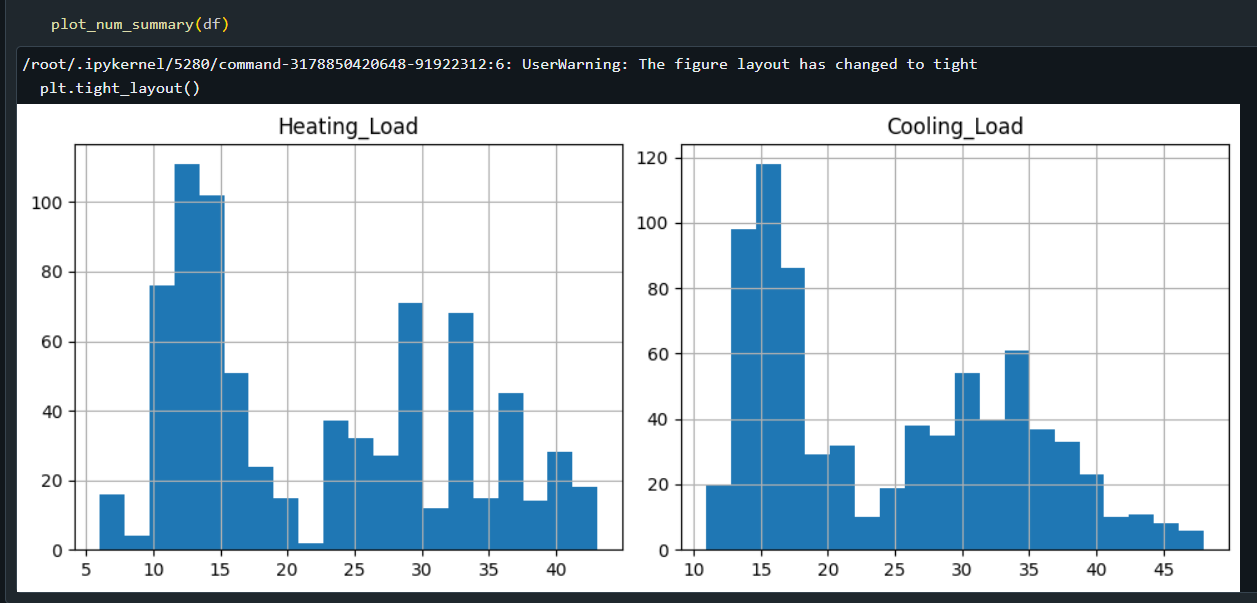


Data describe:

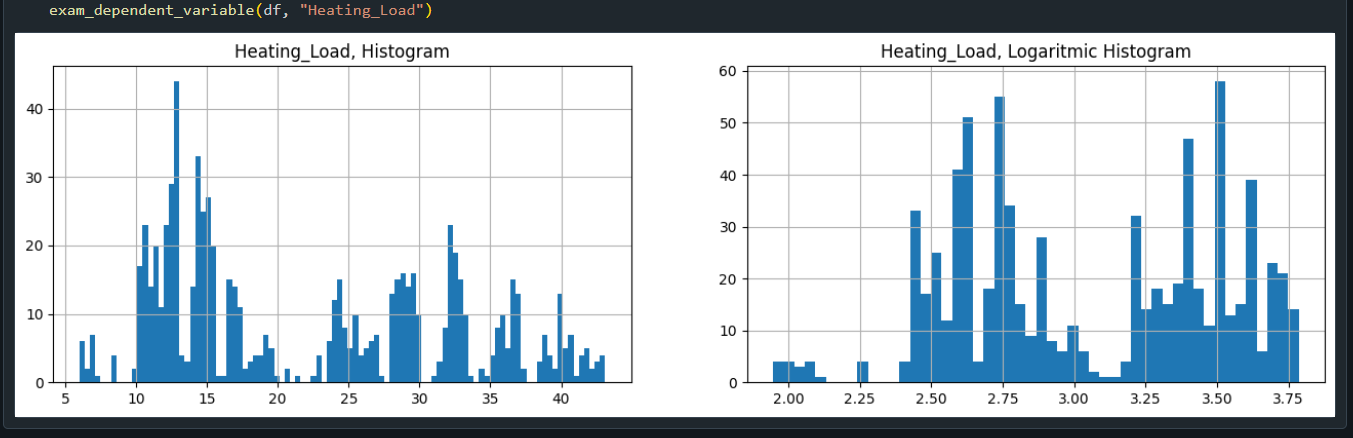


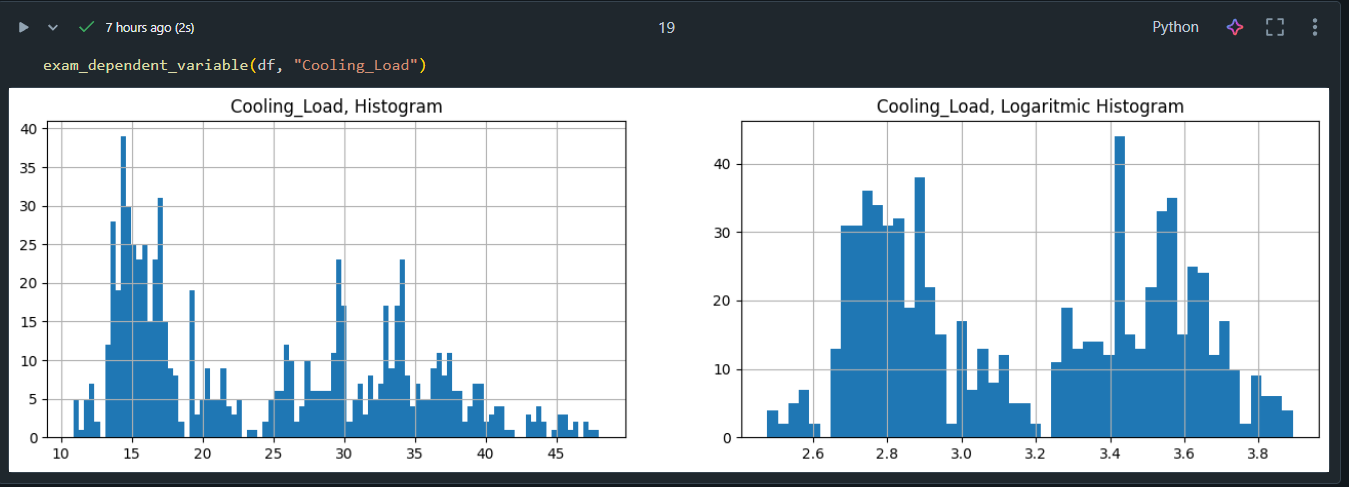
 

### Analysis of Categorical and Numerical Variables

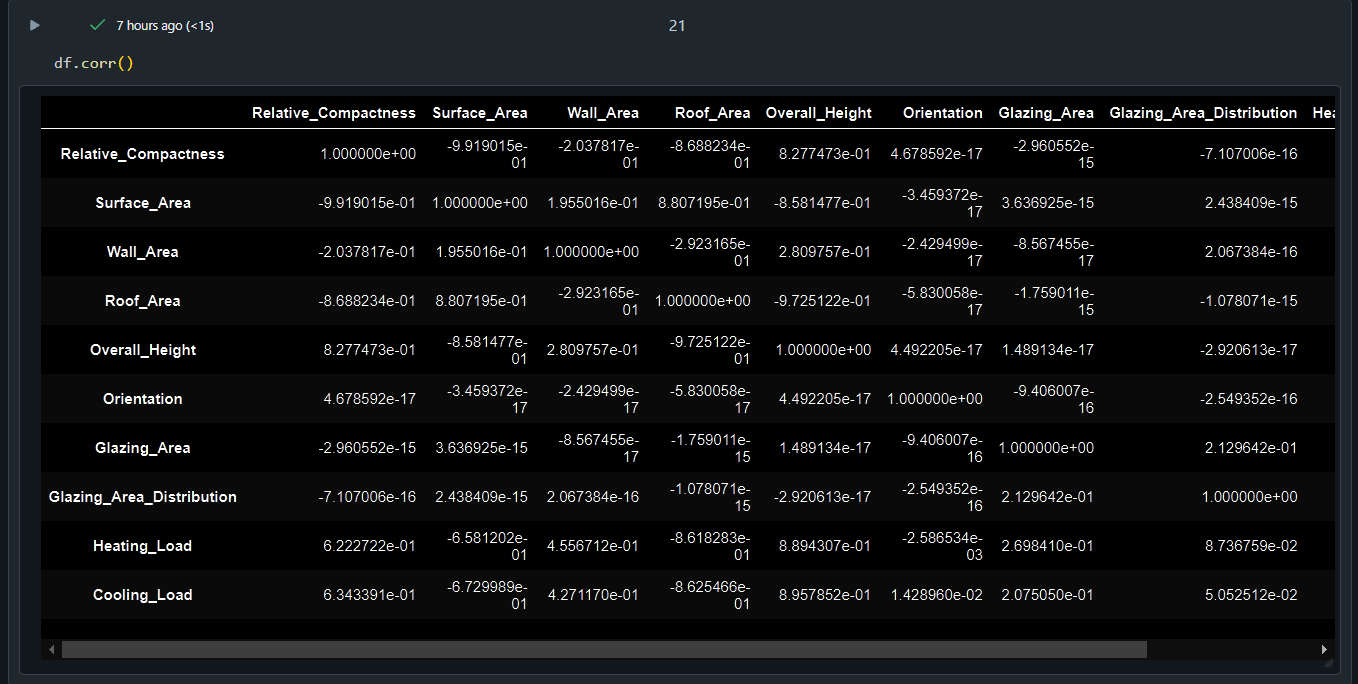


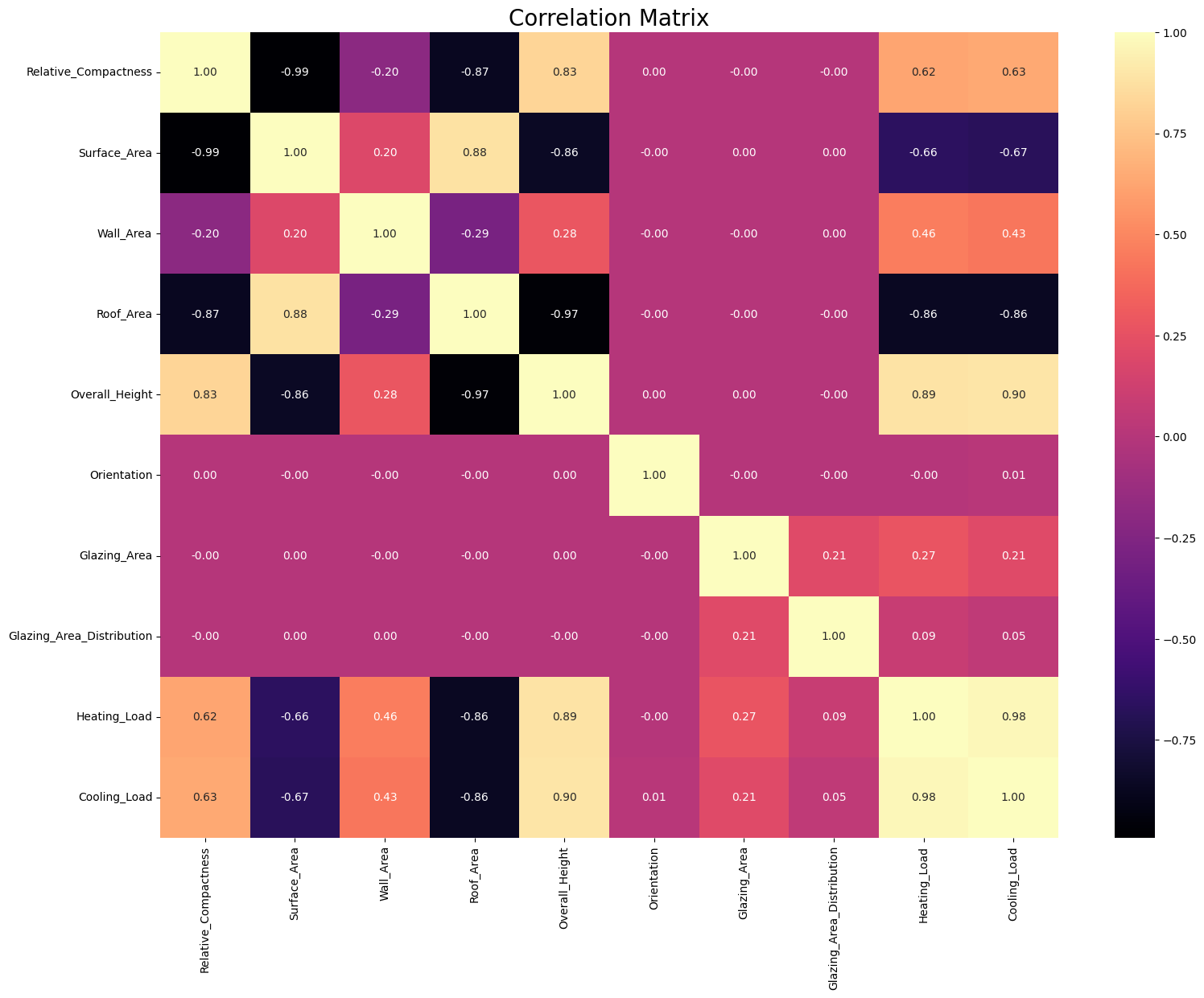
### Examination of the dependent variable





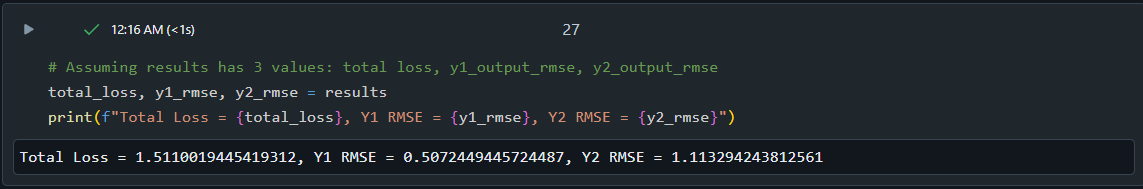
Correlation





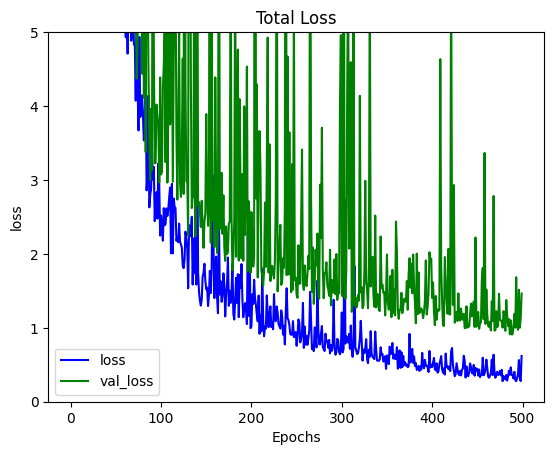
# 4. Model (Neural Network)

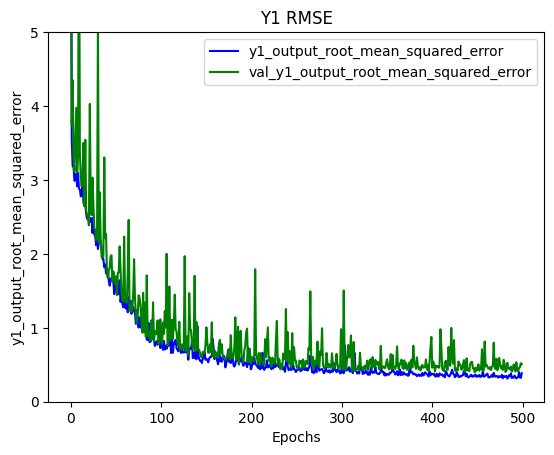




# 5. Results

Loss and Metric Over Epochs:





True vs. Predicted Scatter Plot:

