Introduction

Dataset Background

The chosen dataset is a comma-separated values (csv) file that was downloaded from Kaggle.com. It is titled “Weather in Szeged 2006-2016.” It contains an hourly summary of weather in Szeged, Hungary from the years 2006-2016. With almost 100000 rows and 12 columns, this dataset is quite large, representing variables that contribute to weather such as Precipitation Type, Apparent Temperature, Humidity, Wind Speed, Visibility, and more. This dataset was chosen due to its readability and interpretability.

Motivation

The goal of this project was to build a model such that prediction error measured by mean-square error is minimized. Cross-validation was a requirement to assess the performance of the model. Methodology for solving this problem was decided by taking topics covered in the course as well as knowledge of R into consideration. Linear regression proved to be the best way to create a model for the dataset that was chosen. After some research, it was determined that multiple linear regression would be necessary due to the nature of the dataset. Combining the requirements for the project and knowledge of linear regression in R, response variables and a prediction variable were determined. The goal became build a multiple linear regression model that predicts apparent temperature based on temperature, wind speed, humidity, and pressure. Repeated K-fold Cross Validation was chosen to assess the model and determine Mean Squared Error.

Notations

Mean Squared Error will be denoted by the acronym MSE, with Root Mean Squared Error being RMSE.

Mean Absolute Error will be denoted by MAE.

R-squared is denoted R^2.

Let x_1, x_2 , x_3, and x_4 represent the response variables: Temperature, Wind Speed, Humidity, and Pressure, respectively.

Apparent Temperature is the predictor variable, represented by $\hat{Y}$.

Algorithm Implementation

Choosing the Multiple Linear Regression Model

The linear regression model was chosen because it seemed to fit the dataset well. This decision was based on three things: checking the correlation between variables, realizing that the predictor variable was normally distributed, and noticing that some response variables showed linearity. The correlation between variables was checked using the corr() function. A histogram visualization, shown in Figure 1, represented that the predictor variable was normally distributed. Lastly, one of the plots shown in Figure 2 represented linearity. The predictor variable is apparent temperature, which is the temperature equivalent perceived by humans. After some general research, it was realized that multiple linear regression would have to be used because there are multiple factors, or response variables, that affect apparent temperature.

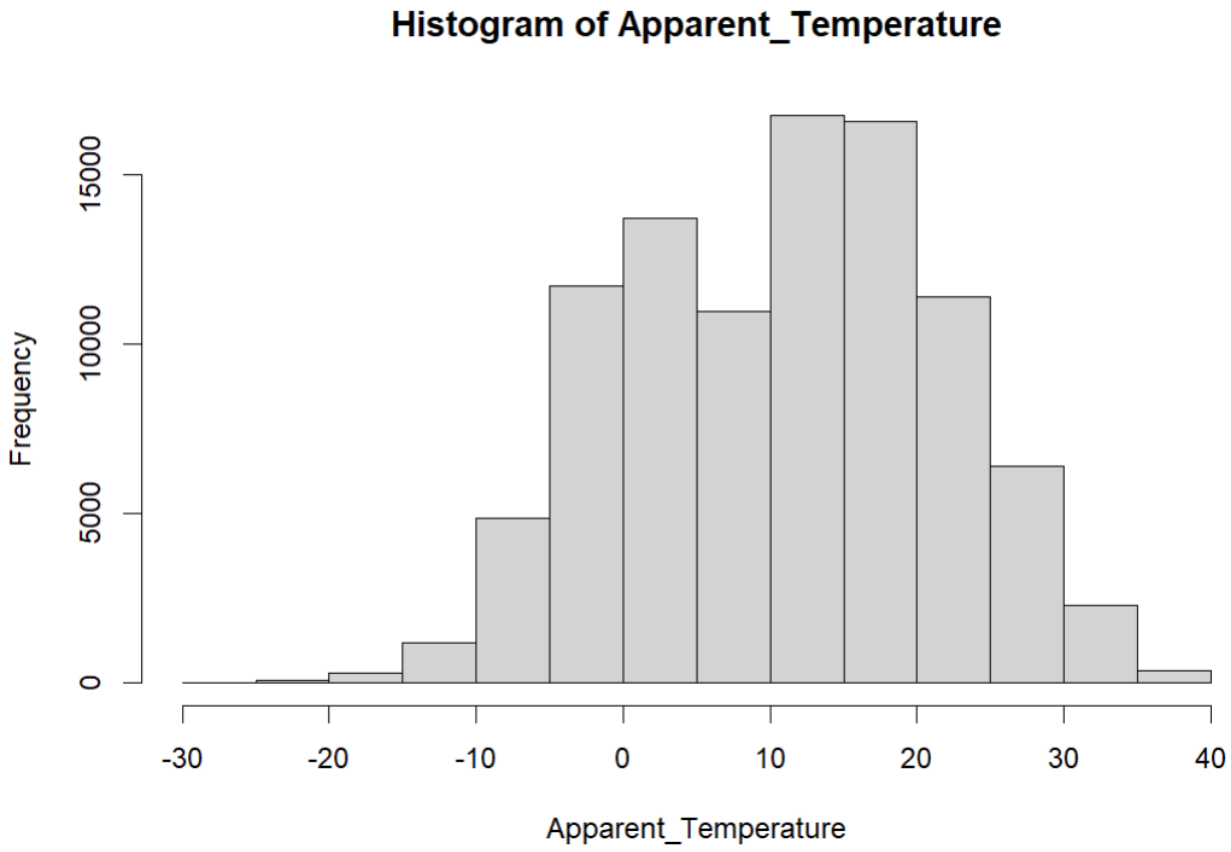

Figure 2

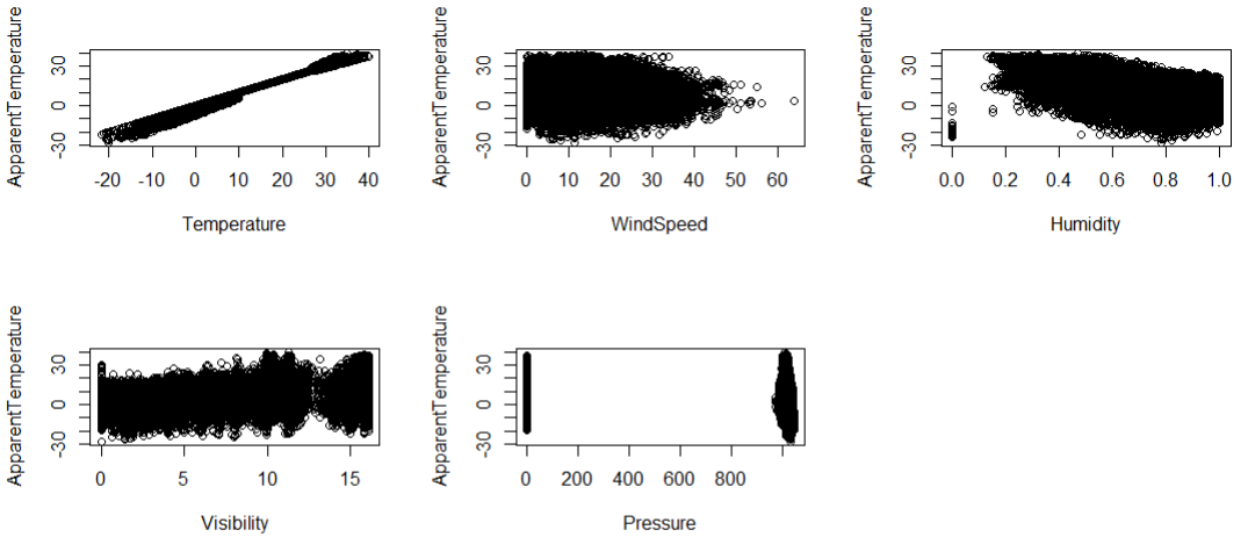

Figure 2

Managing Large Dataset in R

Because the dataset chosen was quite large, importing it into R became a slightly time-consuming process. Once the data was imported, it was then shortened to a dataframe with only the predictor variable and potential response variables. This was done in order to make the process of coding the linear regression more interpretable. These potential response variables were chosen after some general weather research was performed. It was found that typically, apparent temperature is dependent on the initial five response variables chosen. The shortened dataframe had hourly data on apparent temperature, temperature, humidity, wind speed, visibility, and pressure. A preview of this dataframe is shown below in Figure 3.

	ApparentTemperature	Temperature	Humidity	WindSpeed	Visibility	Pressure
1	7.388889	9.472222	0.89	14.1197	15.8263	1015.13
2	7.227778	9.355556	0.86	14.2646	15.8263	1015.63
3	9.377778	9.377778	0.89	3.9284	14.9569	1015.94
4	5.944444	8.288889	0.83	14.1036	15.8263	1016.41
5	6.977778	8.755556	0.83	11.0446	15.8263	1016.51
6	7.111111	9.222222	0.85	13.9587	14.9569	1016.66

Figure 3

Multiple Linear Regression Algorithm

Because the goal of the model was to predict apparent temperature, the shortened dataframe was used to create the first regression iteration (shown below in Figure 4). This model proved to be extremely useful because it showed which potential response variables will be significant for the final model. When a variable's p-value is less than the significance level, the variable is not significant for the model. As represented in Fig. 4, the Visibility response variable proves to be

insignificant for predicting apparent temperature, so it was not included in the final iteration of the model.

The final iteration of the multiple linear regression algorithm successfully ran and revealed the methodology for predicting apparent temperature (shown below in Figure 5). A mathematical equation for apparent temperature is created based on this model.

```
Call:
lm(formula = ApparentTemperature ~ Temperature + WindSpeed +
    Humidity + Visibility + Pressure, data = dataframe)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3242 -0.7158 -0.1071  0.6840  5.3697

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept) -2.528e+00  3.969e-02  -63.681  < 2e-16 ***
Temperature  1.126e+00  4.903e-04 2296.393  < 2e-16 ***
WindSpeed   -9.469e-02  5.260e-04 -180.022  < 2e-16 ***
Humidity     1.056e+00  2.419e-02   43.670  < 2e-16 ***
Visibility  -2.506e-04  9.189e-04   -0.273    0.785
Pressure     1.960e-04  2.983e-05    6.571 5.02e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.079 on 96447 degrees of freedom
Multiple R-squared:  0.9898,    Adjusted R-squared:  0.9898
F-statistic: 1.875e+06 on 5 and 96447 DF,  p-value: < 2.2e-16
```

Figure 4

```
Call:
lm(formula = ApparentTemperature ~ Temperature + WindSpeed +
    Humidity + Pressure, data = dataframe)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3229 -0.7156 -0.1073  0.6837  5.3715

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept) -2.530e+00  3.883e-02  -65.154  < 2e-16 ***
Temperature  1.126e+00  4.772e-04 2359.502  < 2e-16 ***
WindSpeed   -9.470e-02  5.249e-04 -180.420  < 2e-16 ***
Humidity     1.057e+00  2.393e-02   44.182  < 2e-16 ***
Pressure     1.954e-04  2.975e-05    6.569  5.1e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.079 on 96448 degrees of freedom
Multiple R-squared:  0.9898,    Adjusted R-squared:  0.9898
F-statistic: 2.344e+06 on 4 and 96448 DF,  p-value: < 2.2e-16
```

Figure 5

Results

Multiple Linear Regression Results

From the final iteration of the model, the mathematical equation for predicting apparent temperature is as follows:

$$\hat{Y} = -2.5300486304 + 1.1259596402x_1 - 0.0946955500x_2 + 1.0573056860x_3 + 0.0001954374x_4$$

This equation was retrieved from the intercept and coefficients column of the final model.

Assessing the Performance of the Model

Using R-squared from Linear Regression Results

As represented in Figures 4 and 5, the R-squared value for the final model was relatively high at 98.98% which represents that the model explains variation in the response variable around its mean very well.

Repeated K-fold Cross Validation

In order to assess the performance of the model, Repeated K-Fold Cross Validation was used. The command used randomly split the data into 10 subsets, or 10 folds. It then reserved one subset and trained the model on the other 9. For each of these tests, prediction error was recorded, and the average of these 10 errors was calculated in order to calculate the cross-validation error. The cross-validation error, given by Mean Squared Error was around 1.17.

Conclusion

Through the final Multiple Linear Regression model, apparent temperature for Szeged, Hungary can be predicted using the mathematical formula. Using the R-squared representation of the model's performance, it can be said that the final model explains variation very well. The Mean Squared Error representation of the model's performance represents a prediction error of around 1.17 degrees Celsius. Predicting apparent temperature has proven to be beneficial when predicting the human body's comfort level when outside. In the future, it would be advantageous to explore other factors that may contribute to apparent temperature, as this may help reduce the prediction error, thus creating a better prediction. Improving on this model in the future would be done by using subset selection rather than researching which variables may contribute to

apparent temperature because it may return a more accurate model. In addition, subset selection would have eliminated running the first model to see which variables are significant and which were not.