**DATA CLEANING REPORT [NIMET DATASET]**

Handling Missing Values & Outliers

**Temperature & Relative Humidity NIMET Dataset**

The missing values in the temperature column was imputed with the mean temperature of that month. The temperature column contains continuous values. The missing values in the humidity column was imputed with the mode temperature of that month. The humidity column contains discrete values.

In January, 2010, there are two missing values in the RH% 0600Z column [Day 12 & 20]. The mode is 96. The missing values are replaced with the mode 96.

In November 2011, there is one missing value in the RH%1200Z column [Day 07]. It is imputed with the mode is 74 for that month.

In December, 2011 there is one missing value in the RH%1500Z column [Day 12]. It is imputed with the mode of 81.

In June, 2012, there is one missing value in the RH%1200Z column [Day 12]. It is imputed with the mode is 80.

In October, 2012, there is one missing value in the RH%0600Z column [Day 19]. It is imputed with the mode is 96. There is one missing value in the RH%0900Z column [Day 19]. The mode is 85.

In April, 2013, there is one missing value in the RH%1500Z column [Day 18]. The mode is 72.

In December, 2013, there is one missing value in the RH%0600Z column [Day 08]. The mode is 97.

In May, 2014, there are three missing values in the RH%0600Z column [Day 13, 14 & 25]. They are imputed with the mode of 98.

In September, 2014, there is one missing value in the RH%0600Z column [Day 28]. It is imputed with the mode of 98.

In November, 2014, there is one missing value in the RH%0600Z column [Day 29]. It is imputed with the mode of 97.

In June 2015 [Day 29], the values of min temp, rain, RH%0600Z and RH%0900Z are missing. The min temp is imputed with a mean of 23.3. The rainfall is imputed with a mean of 19.6. The RH%0600Z is imputed with a mode of 97. The RH%0900Z is imputed with a mode of 85.

In July, 2015, there is a missing value in the RH%01200Z [Day 01]. It is imputed with a mode of 96. There is one missing value in the max temp column [Day 10]. It is imputed with a mean of 28.0. There are two missing values in the min temp column [Day 10 & 11]. They are imputed with a mean of 23.2.

In September 2015, there are two missing values in the Max Temp column [Day 3 & 30]. They are imputed with the mean of 28.7. There is one missing value in the Min Temp column [Day 20]. It is imputed with a mean of 23.3. There are two missing values in the RH%0600Z column [Day 20 & 26]. They are imputed with a mode of 98.

In October 2015, there is a missing value in the Max Temp column [Day 2]. It is imputed with the mean of 29.7. There are three missing values in the RH%0600Z column [Day 29, 30 & 31]. They are imputed with a mode of 96.

In November 2015, there is a missing value in the Max Temp column [Day 1]. It is imputed with the mean of 32.2. There are four missing values in the RH%0600Z column [Day 01, 02, 28 & 30]. They are imputed with a mode of 96.

In December 2015, there are eight missing values in the RH%0600Z column [Day 01, 02, 03, 04, 05, 09, 10, & 13]. They are imputed with a mode of 90.

In January, 2016, there is a missing value in the RH%0600Z column [Day 24]. It is imputed with a mode of 96.

In February, 2016, there are four missing values in the RH%0600Z column [Day 05, 11, 13, 14]. They are imputed with a mode of 94. There is a missing value in the Min Temp column [Day 24]. It is imputed with the mean of 24.4.

In April, 2016, there is a missing value in the RH%1200Z column [Day 26]. It is imputed with a mode of 70.

In May, 2016, there is a missing value in the Max Temp column [Day 31]. It is imputed with the mean of 31.3.

In June, 2016, there is a missing value in the Min Temp column [Day 21]. It is imputed with the mean of 22.9. There is a missing value in the RH%0600Z column [Day 21]. It is imputed with a mode of 96.

In July, 2016, there are four missing values in the RH%0600Z column [Day 06, 07, 09, 15]. It is imputed with a mode of 96. There is a missing value in the RH%1200Z column [Day 30]. It is imputed with a mode of 88. There are two missing values in the Max Temp column [Day 26 & 27]. It is imputed with the mean of 28.5.

In September, 2016, there are two missing values in the Max Temp column [Day 04 & 28]. It is imputed with the mean of 28.8. There is a missing value in the Min Temp column [Day 04]. It is imputed with the mean of 23.1. There are two missing values in the RH%0600Z column [Day 06, 07, 09, 15]. They are imputed with a mode of 96.

In December, 2016, there are four missing values in the Max Temp column [Day 02, 10, 11 & 13]. It is imputed with the mean of 33.0.

In January, 2017, there are three missing values in the Max Temp column [Day 17, 19, & 20]. It is imputed with the mean of 32.8.

In February, 2017, there is a missing value in the Min Temp column [Day 18]. It is imputed with the mean of 23.4. There is a missing value in the RH%0600Z column [Day 22]. It is imputed with a mode of 96.

In March, 2017, there are two missing values in the Max Temp column [Day 18 & 31]. It is imputed with the mean of 33.9. There are two missing values in the Min Temp column [Day 21 & 23]. It is imputed with a mean of 23.8.

In April, 2017, there are two missing values in the RH%0600Z column [Day 10 & 11]. It is imputed with a mode of 96.

In May, 2017, there is a missing value in the Max Temp column [Day 14]. It is imputed with the mean of 30.6. There is a missing value in the Min Temp column [Day 14]. It is imputed with a mean of 23.5. There are six missing values in the RH%0600Z column [Day 13, 14, 17, 21,23 & 27]. It is imputed with a mode of 96.

In June, 2017, there are four missing values in the Max Temp column [Day 04, 08, 20 & 23]. They are imputed with the mean of 29.9. There is a missing value in the Min Temp column [Day 04]. It is imputed with a mean of 23.5.

In July, 2017, there are six missing values in the Max Temp column [Day 01, 02, 10, 26, 30 & 31]. They are imputed with the mean of 28.3.

In August, 2017, there are 18 missing values in the Min Temp column [Day 13, 15-31]. They are imputed with the mean of 22.9.

In September, 2017, all the values in the Min Temp column are missing. They are imputed with the value of 23.3. 23.3 is the average of the min temperature from Jan 2010 to August 2017.

In October, 2017, there is a missing value in the Max Temp column [Day 02]. It is imputed with the mean of 30.6. All the values in the Min Temp column are missing. They are imputed with the value of 23.3. There is a missing value in the Rain column [Day 01]. It Is imputed with a mean of 8.6. There is a missing value in the RH%0600Z column [Day 12]. It is imputed with a mode of 96.

In November, 2017, all the values in the Min Temp column are missing. They are imputed with the value of 23.3. There is a missing value in the RH%0600Z column [Day 23.]. It is imputed with a mode of 96.

In December, 2017, all the values in the Min Temp column are missing. They are imputed with the value of 23.3. There are two missing values in the RH%0600Z column [Day 18 & 30]. It is imputed with a mode of 96. There is a missing value in the Max Temp column [Day 27]. It is imputed with the mean of 32.5.

**VAPOUR PRESSURE NIMET DATASET**

In January 2010 [Day 12 & 20], there are two missing values in the Vapour Pressure 0600Z column. They are imputed with the mean of the column for that month. The mean is 29.8.

In December 2011 [Day 16], there is a missing value in the Vapour Pressure 0600Z column. It is imputed with the mean of the column for that month. The mean is 27.5.

On 28th January 2012, the value of the Max Temp is 3.5. It is corrected to 35.

In June 2012 [Day 12], there is a missing value in the Vapour Pressure 1200Z column. It is imputed with the mean of the column for that month. The mean is 30.4.

In April 2013 [Day 18], there is a missing value in the Vapour Pressure 1500Z column. It is imputed with the mean of the column for that month. The mean is 31.7.

In May 2013 [Day 18], there is a missing value in the Vapour Pressure 1500Z column. It is imputed with the mean of the column for that month. The mean is 31.7.

In May 2014 [Day 13, 14, & 25], there are three missing values in the Vapour Pressure 0600Z column. It is imputed with the mean of the column for that month. The mean is 30.1.

In November 2014 [Day 29], there is a missing value in the Vapour Pressure 0600Z column. It is imputed with the mean of the column for that month. The mean is 29.1.

In June 2015 [Day 29], there is a missing value in the Vapour Pressure 0600Z column. It is imputed with the mean of the column for that month. The mean is 29.1. There is a missing value in the Vapour Pressure 0900Z column. It is imputed with the mean of the column for that month. The mean is 30.4. There is a missing value in the Vapour Pressure 1200Z column. It is imputed with the mean of the column for that month. The mean is 30.7.

In July 2015 [Day 29], there are 14 missing values in the Vapour Pressure 0600Z column. It is imputed with the mean of the column for that month. The mean is 28.9. There is a missing value in the Vapour Pressure 1200Z column. It is imputed with the mean of the column for that month. The mean is 30.1.

In September 2015 [Day 20 & 26], there are 2 missing values in the Vapour Pressure 0600Z column. It is imputed with the mean of the column for that month. The mean is 29.3.

In October 2015 [Day 29, 30 & 31], there are 3 missing values in the Vapour Pressure 0600Z column. It is imputed with the mean of the column for that month. The mean is 29.4.

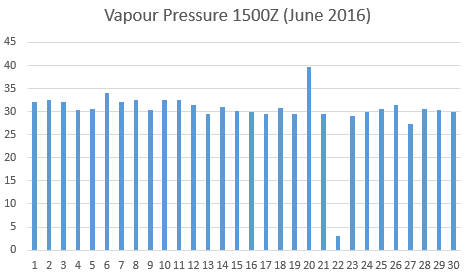
In November 2015 [Day 01, 02, 28, & 30], there are 4 missing values in the Vapour Pressure 0600Z column. They are imputed with the mean of the column for that month. The mean is 30.0.

In December 2015 [Day 01, 02, 03. 04. 05, 09, 10, & 13], there are 8 missing values in the Vapour Pressure 0600Z column. They are imputed with the mean of the column for that month. The mean is 20.0.

In January 2016 [Day 24], there is a missing value in the Vapour Pressure 0600Z column. It is imputed with the mean of the column for that month. The mean is 23.3.

In February 2016 [Day 05, 11, 13, & 14], there are 4 missing values in the Vapour Pressure 0600Z column. They are imputed with the mean of the column for that month. The mean is 30.1.

In June 2016 [Day 21], there is a missing value in the Vapour Pressure 0600Z column. It is imputed with the mean of the column for that month. The mean is 28.7.



In Day 22, the value of Vapour Pressure 1500Z is 3. This is an outlier. This value is replaced with the mean of the column, 31.1.

In July 2016 [Day 06, 07, 09, & 15], there are 4 missing values in the Vapour Pressure 0600Z column. They are imputed with the mean of the column for that month. The mean is 29.1.

In September 2016 [Day 08] there is a missing value in the Vapour Pressure 0600Z column. It is imputed with the mean of the column for that month. The mean is 29.1.

In February 2017 [Day 22] there is a missing value in the Vapour Pressure 0600Z column. It is imputed with the mean of the column for that month. The mean is 29.4.

In April 2017 [Day 10 & 11] there are missing values in the Vapour Pressure 0600Z column. They are imputed with the mean of the column for that month. The mean is 30.

In May 2017 [Day 13, 14, 17, 21, 23 & 27] there are missing values in the Vapour Pressure 0600Z column. They are imputed with the mean of the column for that month. The mean is 30.

In October 2017 [Day 12] there is a missing value in the Vapour Pressure 0600Z column. It is imputed with the mean of the column for that month. The mean is 30.

In November 2017 [Day 23] there is a missing value in the Vapour Pressure 0600Z column. It is imputed with the mean of the column for that month. The mean is 29.

In December 2017 [Day 18 & 30] there are missing values in the Vapour Pressure 0600Z column. They are imputed with the mean of the column for that month. The mean is 28.5.

On 8th July 2010, there is an outlier of 312 in the Vapour Pressure 1200Z. It was corrected to 31.2.

On 19th February, 2014, the Vapour Pressure 0900Z value is 321.2. It was corrected to 32.1

On 12th May 2014, there is an outlier or 89 in the Vapour Pressure 1500Z column. It was replaced with the value of the next column 31.2.

On the 17th August 2014, there is an outlier of 96 in the Vapour Pressure 0600Z column. It was replaced with the value of the next cell 28.0.

On the 18th Feb 2015, there is an outlier of 285 in the Vapour Pressure 1500Z column. It was corrected to 28.5.

On 6th March 2015, there is an outlier of 267 in the Vapour Pressure 1500Z column. It was corrected to 26.7.

On 22nd November 2015, the value of Vapour Pressure 1500Z is 2.0. This is an outlier. It is replaced with the next value in that column 30.

On 3rd August 2016, there is an outlier of 2706 in the Vapour Pressure 0600Z column. It was corrected to 27.06. There is also an outlier of 2207 in the Vapour Pressure 0900Z column. It was corrected to 22.07.

On 2nd August 2017, there is an outlier of 286.0 in the Vapour Pressure 1500Z column. It was corrected to 28.6.

**NOTE:** The data for September 2014 is missing. To resolve this, the data for August was used for September, except for 31/08/14. It is because September has 30 days.

Trailing Spaces are removed from the dataset.

**EXPLORATORY DATA ANALYSIS**



Figure 1: Statistical Description of the dataset

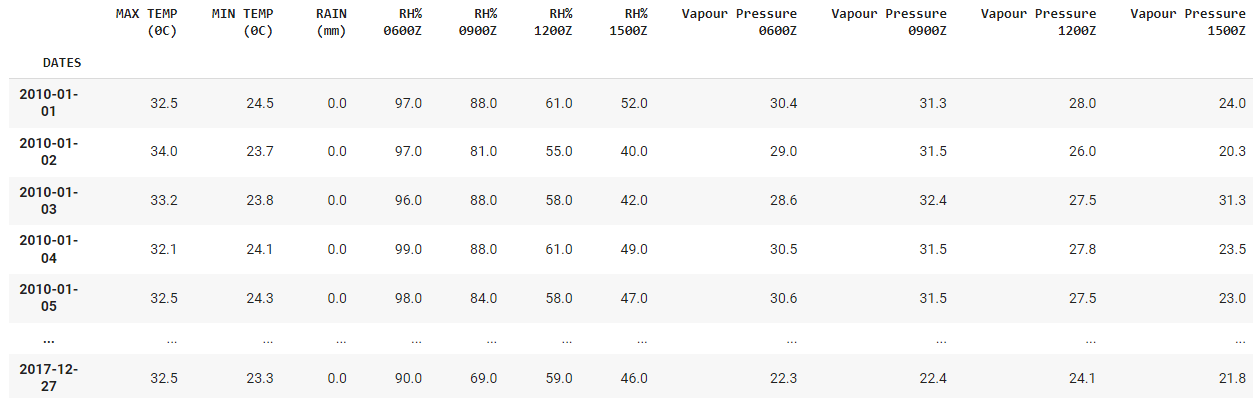


Figure 2: Clean dataset Preview

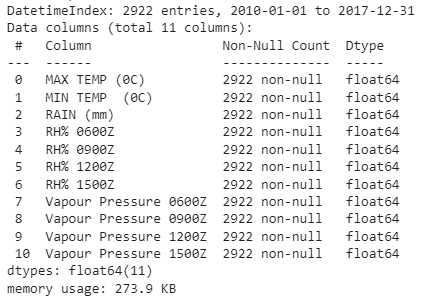


Figure 3: Data types of variables

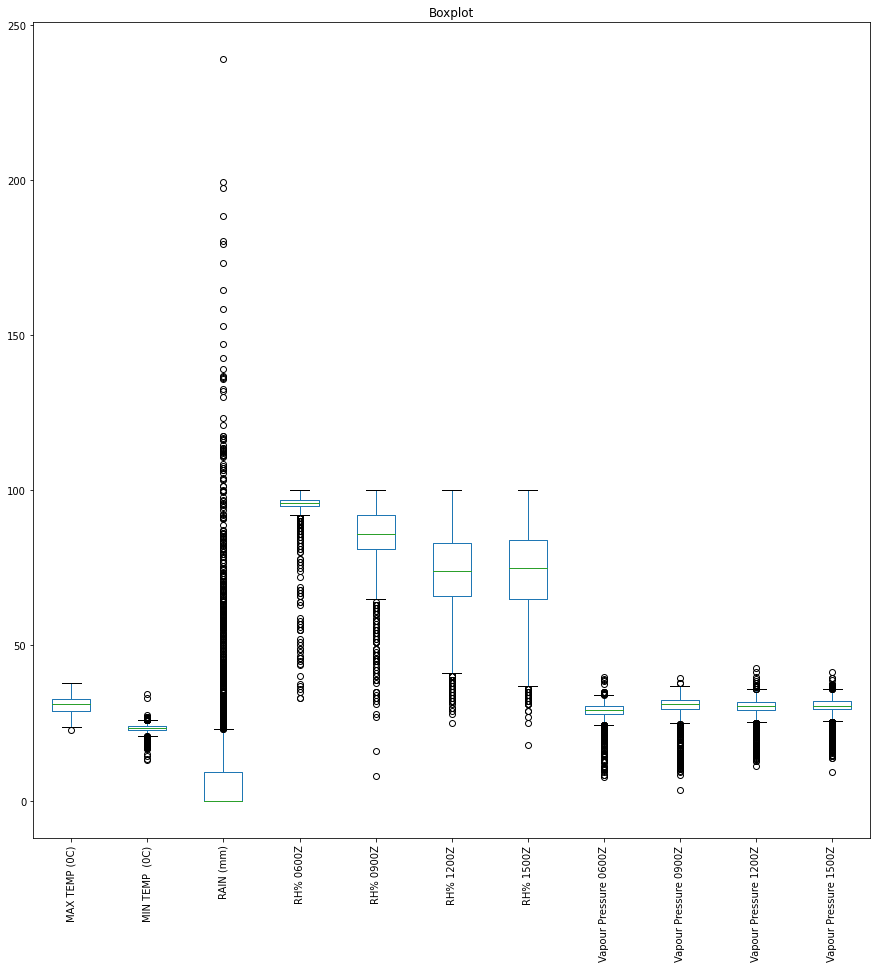


Figure 2: Boxplot of variables in the dataset

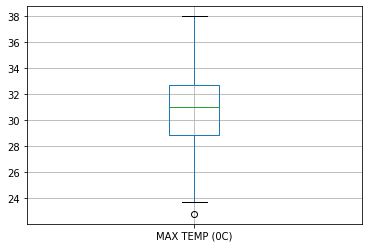


Figure 3: Boxplot of Max Temp variable

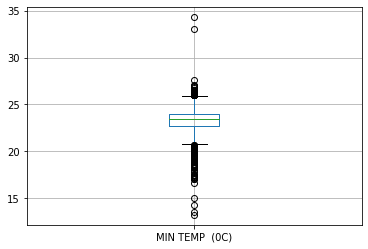


Figure 3: Boxplot of Min Temp variable

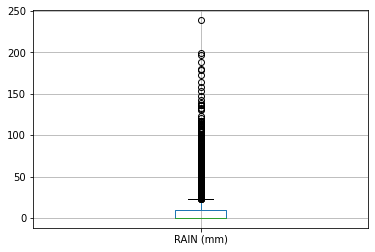


Figure 4: Boxplot of Rain variable

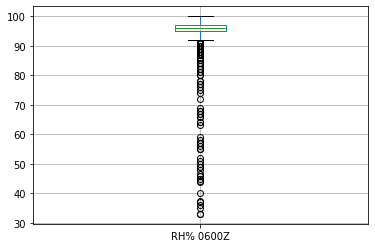


Figure 5: Boxplot of Relative Humidity 0600Z variable

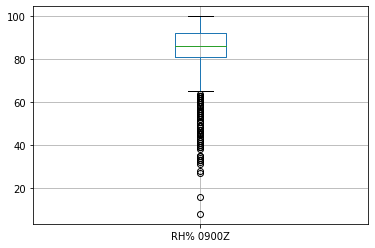


Figure 6: Boxplot of Relative Humidity 0900Z variable

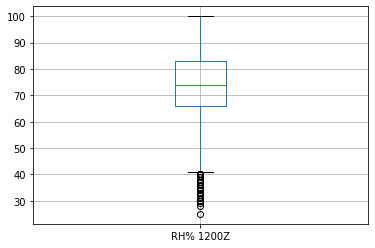


Figure 7: Boxplot of Relative Humidity 1200Z variable

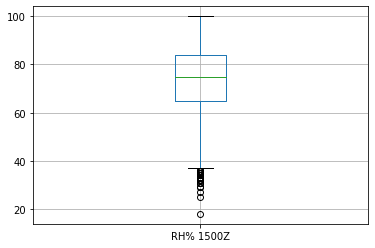


Figure 8: Boxplot of Relative Humidity 1500Z variable

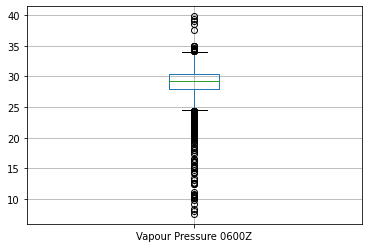


Figure 9: Boxplot of Vapour Pressure 0600Z variable

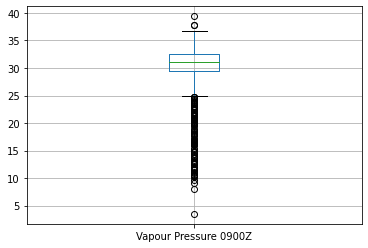


Figure 10: Boxplot of Vapour Pressure 0900Z variable

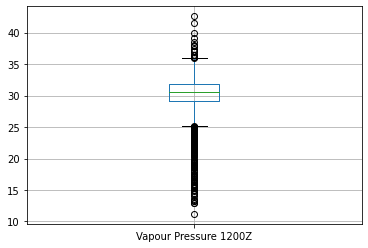


Figure 11: Boxplot of Vapour Pressure 1200Z variable

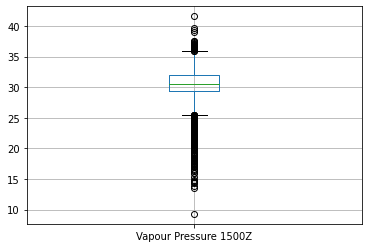


Figure 12: Boxplot of Vapour Pressure 1500Z variable

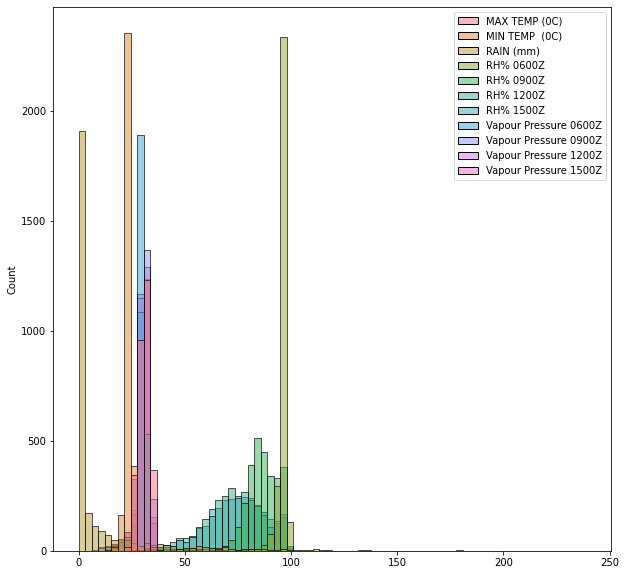


Figure 13: Histogram of all the variables in dataset

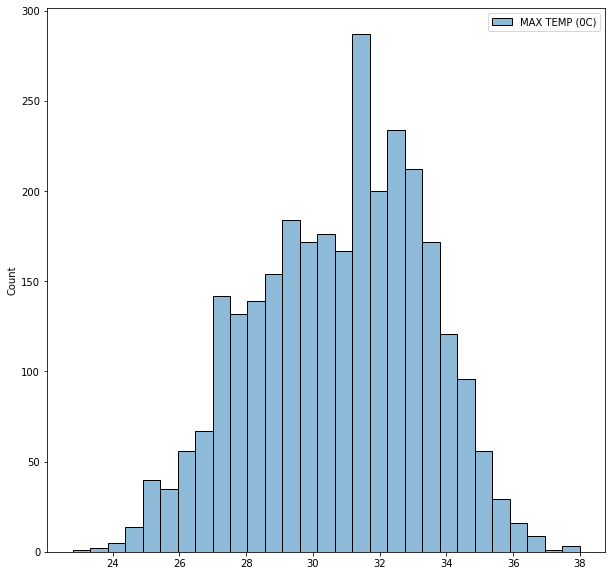


Figure 14: Histogram of all the Max Temp variable

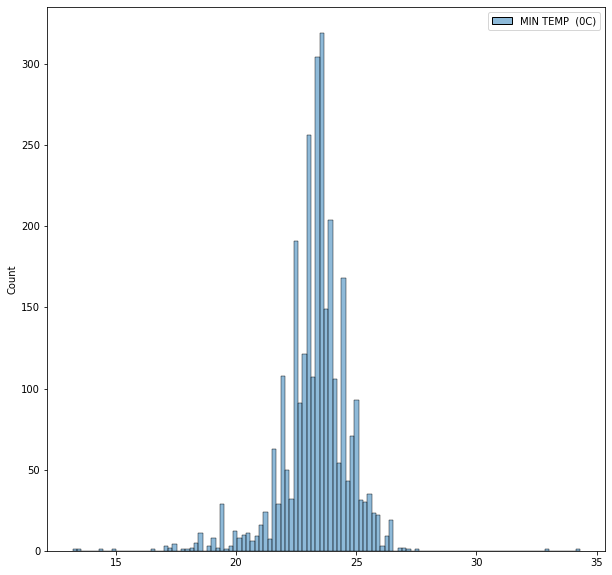


Figure 15: Histogram of all the Min Temp variable

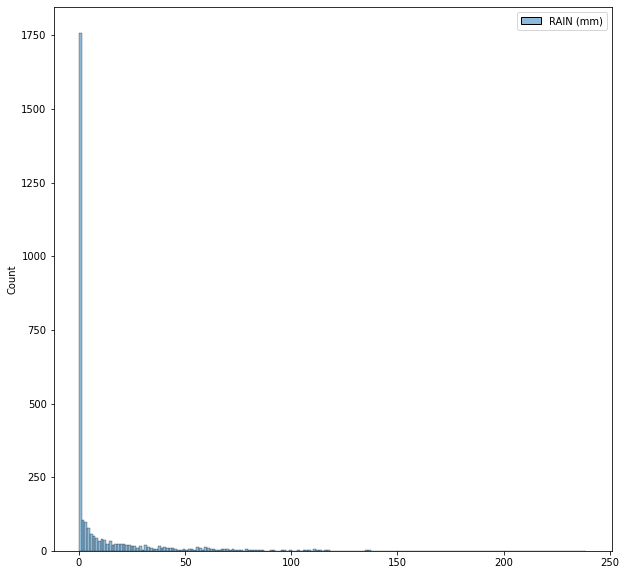


Figure 16: Histogram of all the Rain variable

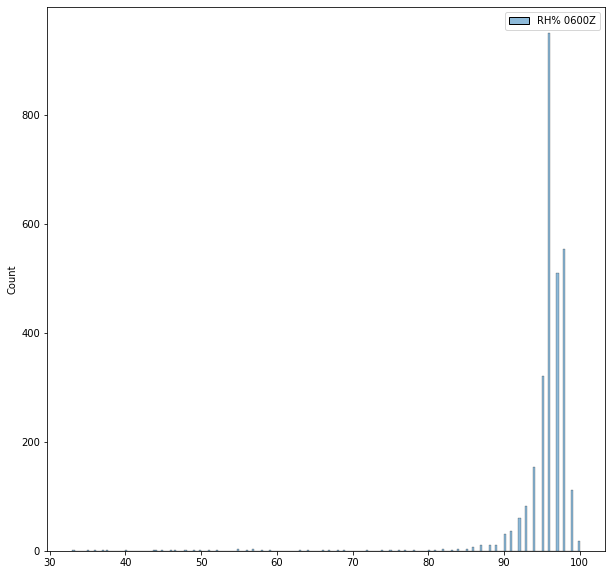


Figure 17: Histogram of all the Relative Humidity 0600Z variable

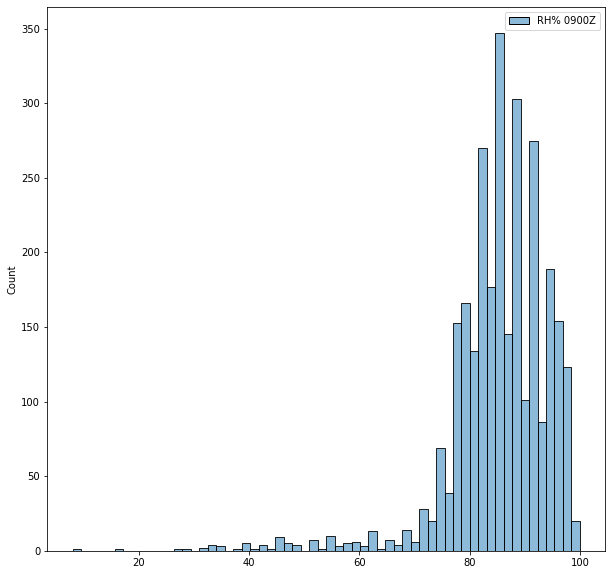


Figure 18: Histogram of all the Relative Humidity 0900Z variable

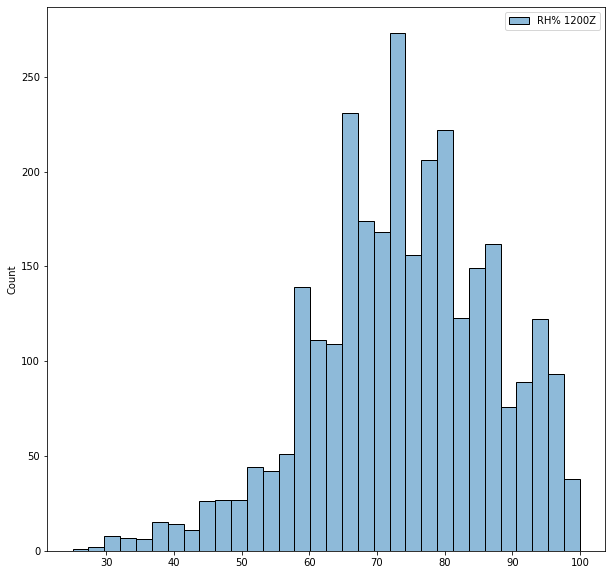


Figure 19: Histogram of all the Relative Humidity 1200Z variable

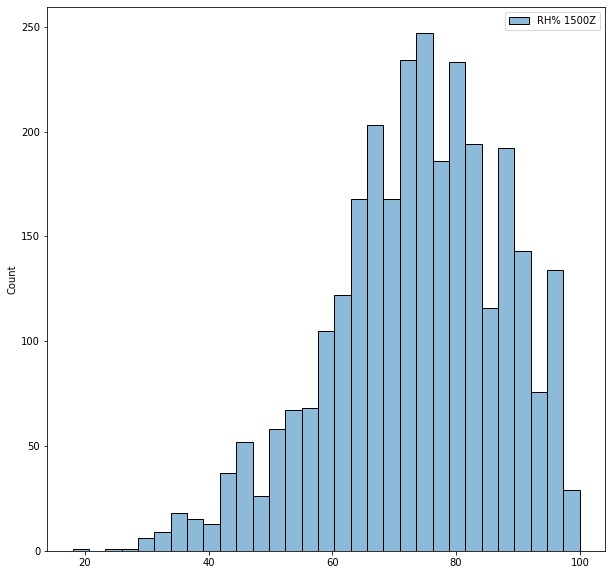


Figure 20: Histogram of all the Relative Humidity 1500Z variable

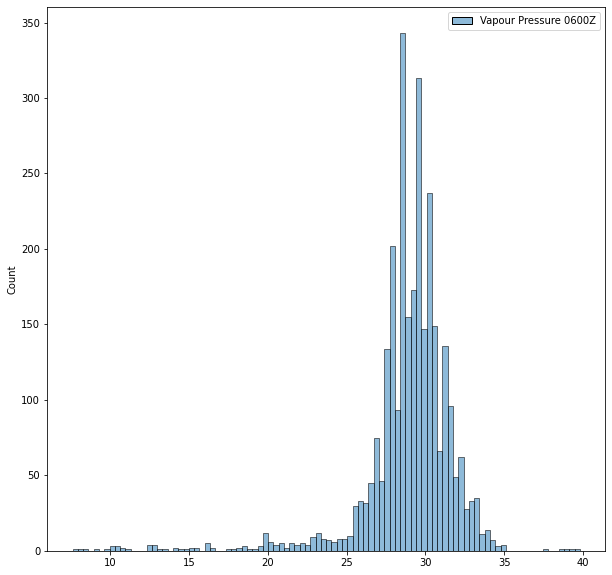


Figure 21: Histogram of all the Vapour Pressure 0600Z variable

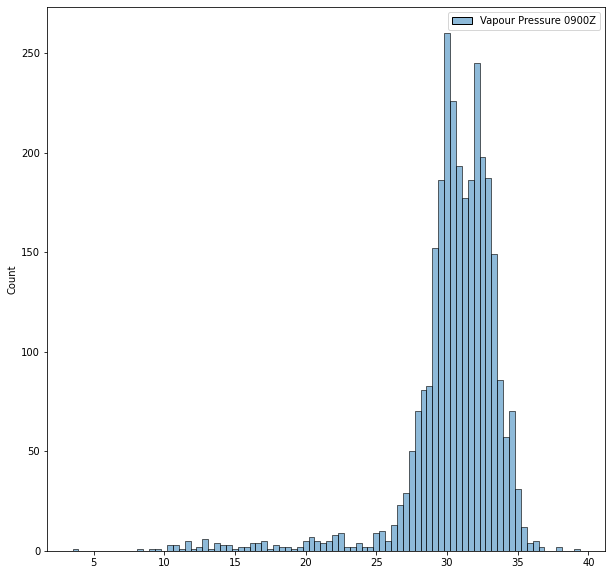


Figure 22: Histogram of all the Vapour Pressure 0900Z variable

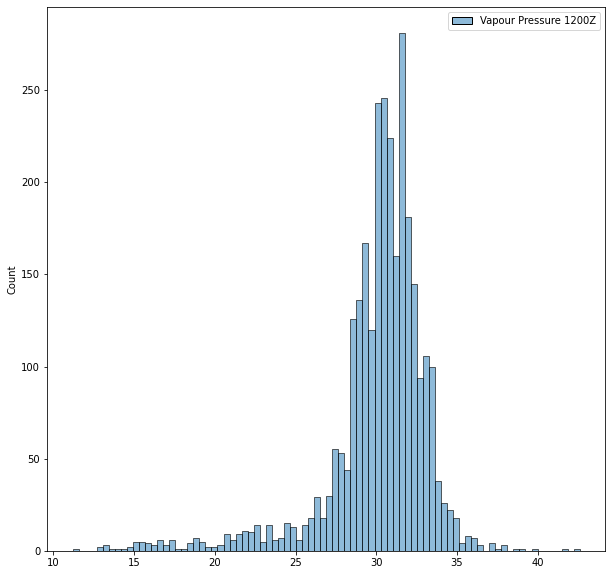


Figure 23: Histogram of all the Vapour Pressure 1200Z variable

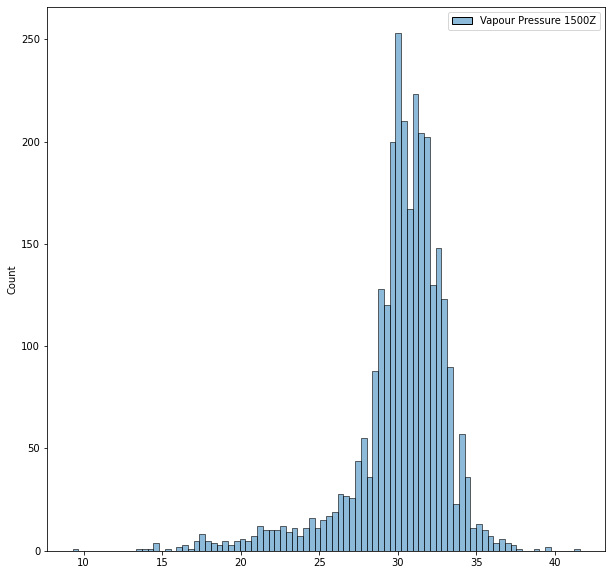


Figure 24: Histogram of all the Vapour Pressure 1500Z variable

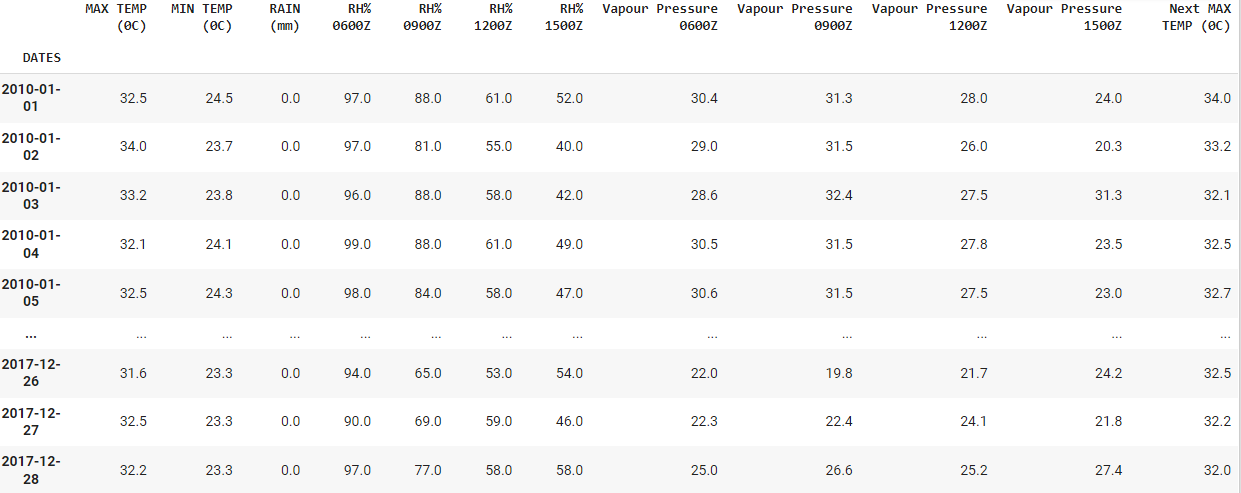


Figure 27: Dataset with the Next Max Temp as Target variable

A target variable column is created containing the value of the next day’s maximum temperature. The goal is to train a model to predict what the maximum temperature will be the next day.

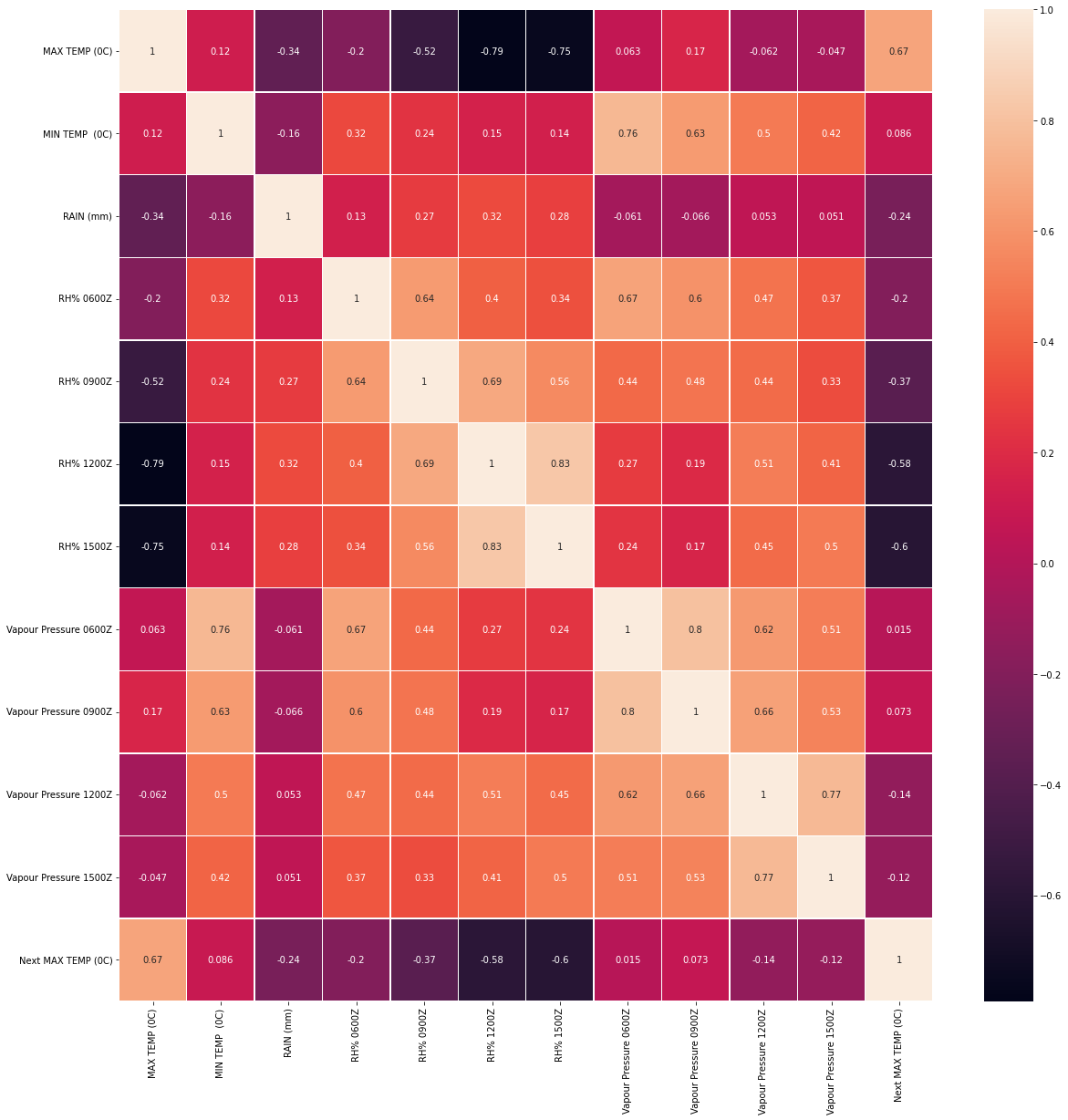


Figure 28: Pearson correlation of variables

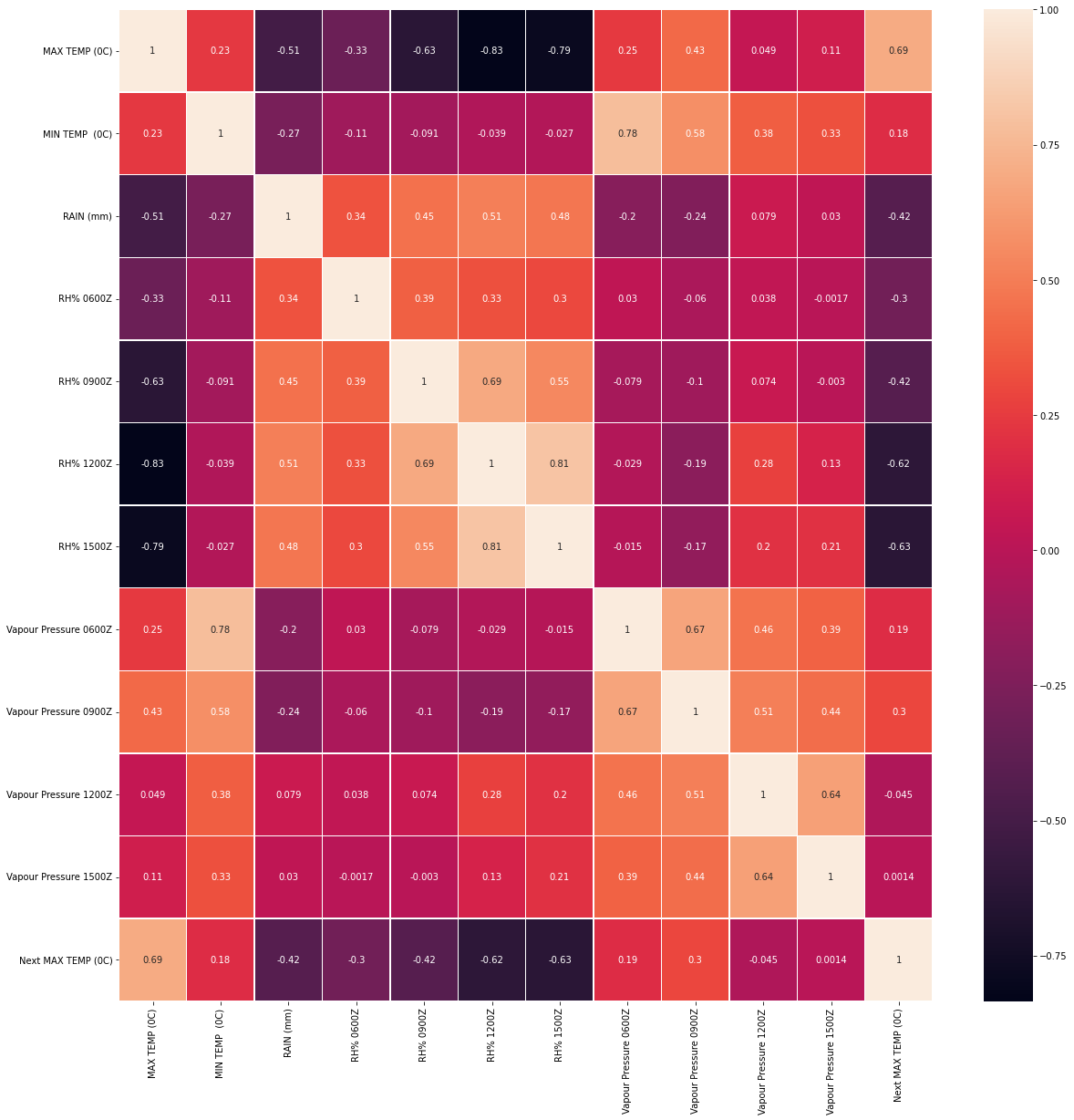


Figure 29: Spearman correlation of variables

[Input spearman correlation formula]

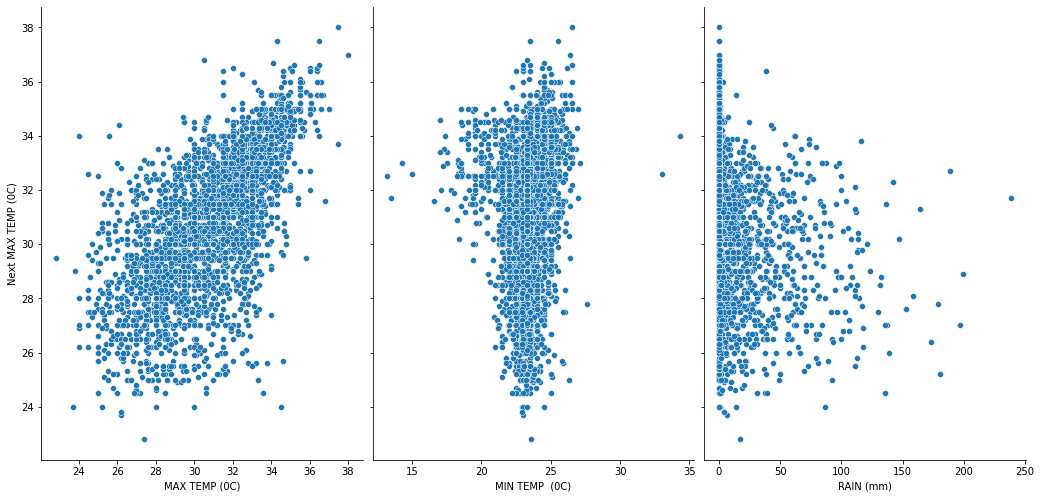


Figure 30: Scatter plot of Next Max Temp against Max Temp, Min Temp and Rain variables

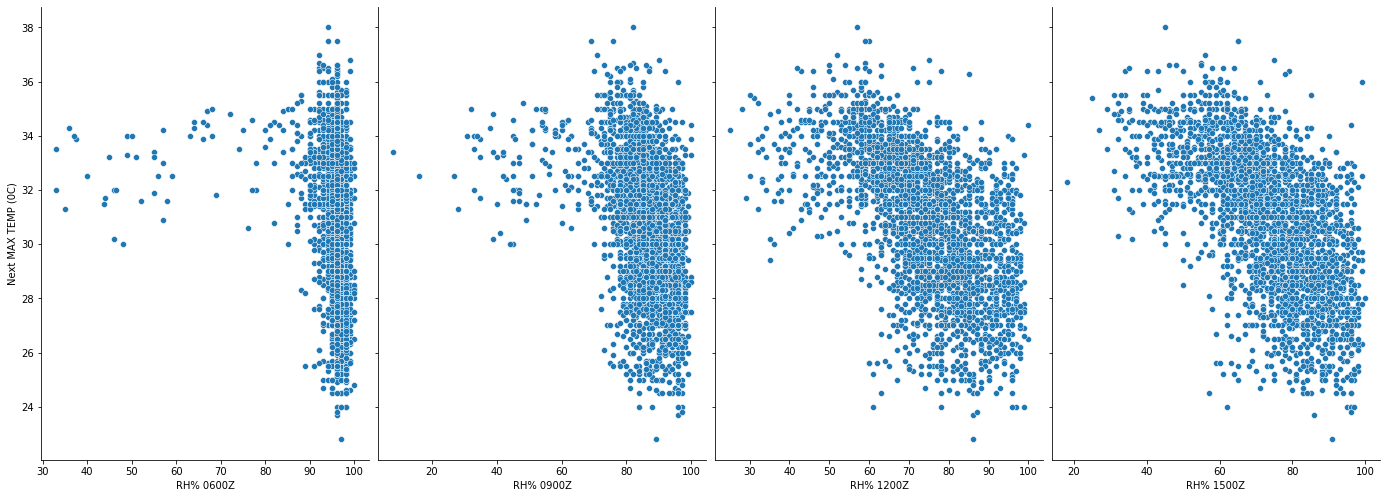


Figure 31: Scatter plot of Next Max Temp against RH0600z, RH0900z, RH1200z and RH1500z variables

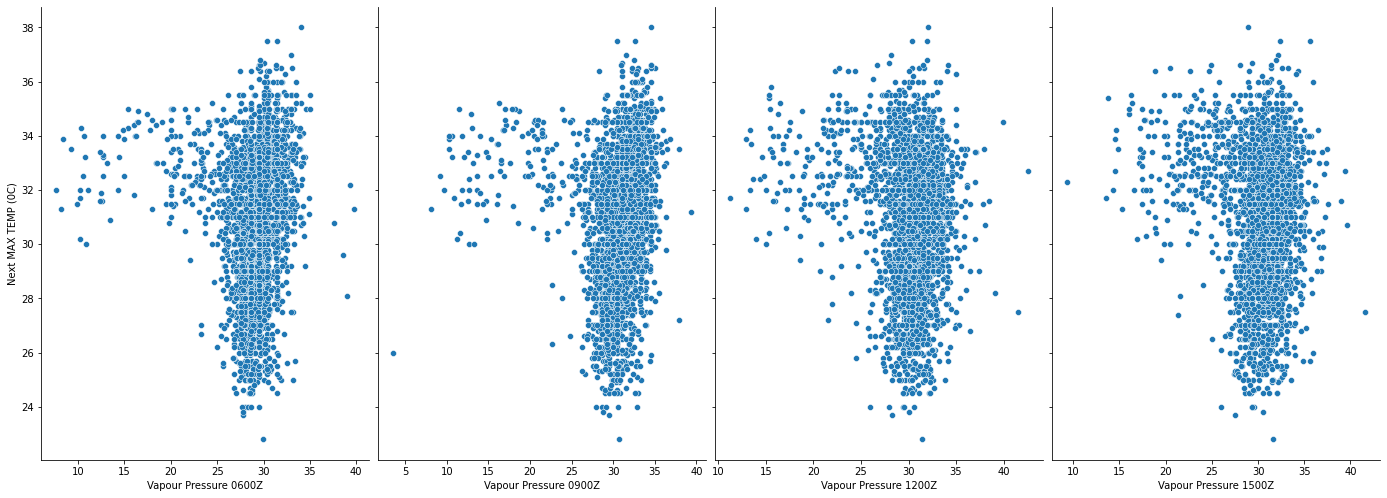
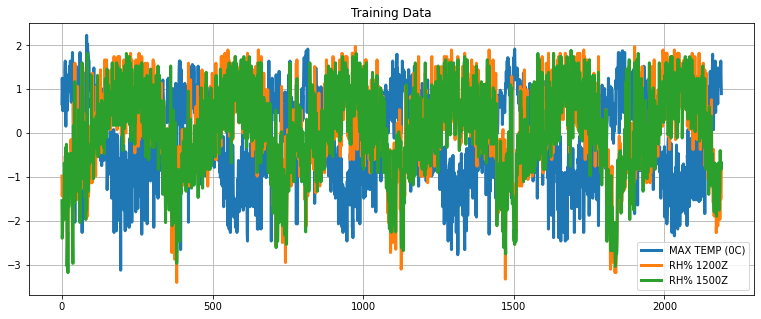


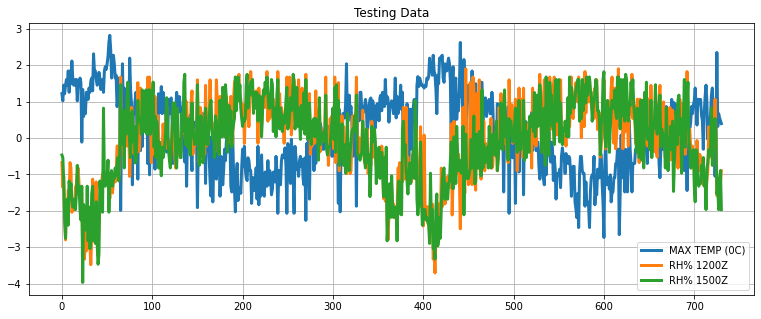
Figure 31: Scatter plot of Next Max Temp against VP0600z, VP0900z, VP1200z and VP1500z variables

The three features that will be used for predicting the Next Max Temp are ‘Max Temp’, ‘RH1200Z’, ‘RH1500Z’. Max Temp has a positive correlation of 0.67 with Next Max Temp. RH1200Z has a negative correlation of -0.58 with Next Max Temp. RH1500Z has a negative correlation of -0.60 with Next Max Temp.

Splitting Data into Training Set and Test Set

Before splitting the dataset, standardization is applied to it. This is to ensure all the features are of the same variance. The dataset is split in the ration of 80:20.



Figure 32: Line Plot of the Max Temp, RH% 1200Z & RH% 1500Z features after applying standardization in training set and test set.

Model Training  
The algorithms used for training are Ridge Regression, Lasso Regression, Decision Tree Regression, XGBoost Regressor.

**Ridge Regression**

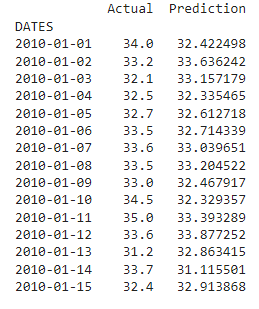


Figure 33: Actual result vs predicted result of the ridge regression model.

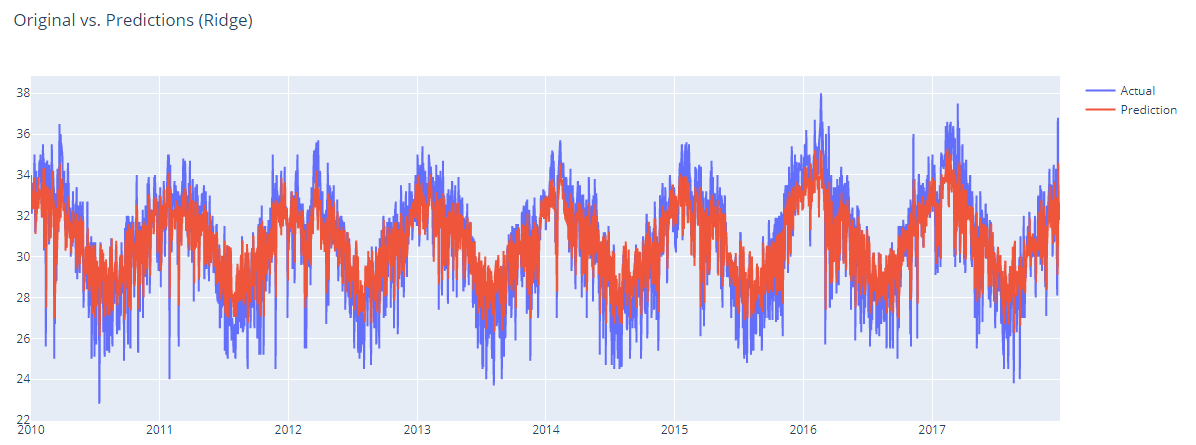


Figure 34: Line Plots of actual result vs predicted result of the ridge regression model.

**Lasso Regression**

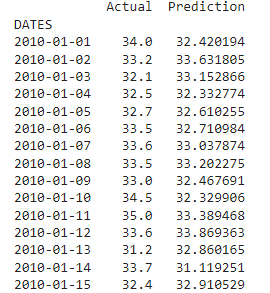


Figure 35: Actual result vs predicted result of the lasso regression model.

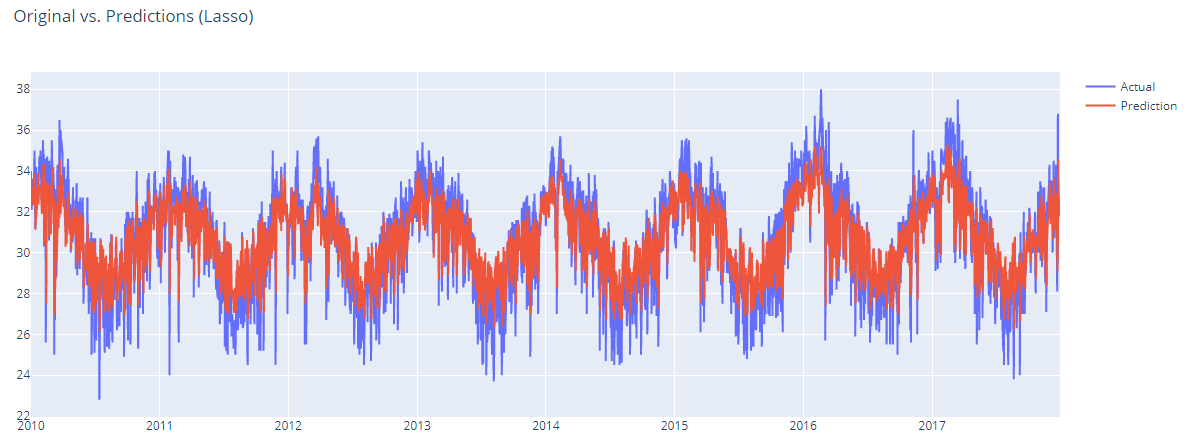


Figure 36: Line Plots of actual result vs predicted result of the lasso regression model.

**Decision Tree Regression**

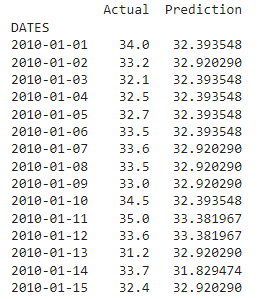


Figure 37: Actual result vs predicted result of the decision tree regression model.

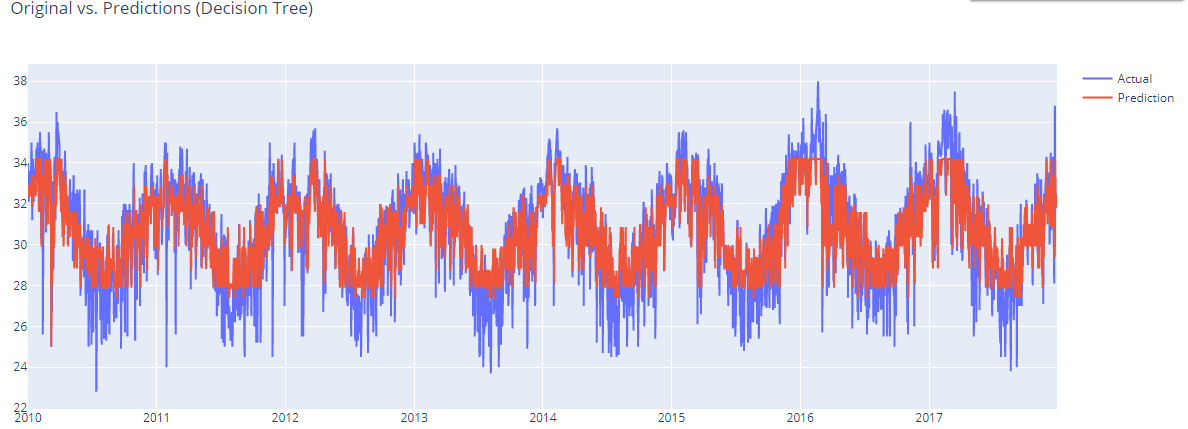


Figure 38: Line Plots of actual result vs predicted result of the decision tree regression model.

**XGBoost Model**

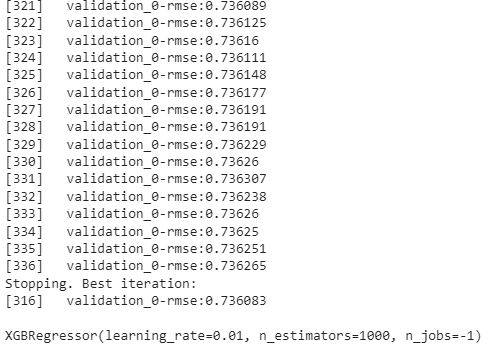
****

Figure 39: Training iteration of the XGBoost regression model.

During the training process, we have the lowest root mean squared error (RMSE) of the validation set in the 316th iteration.

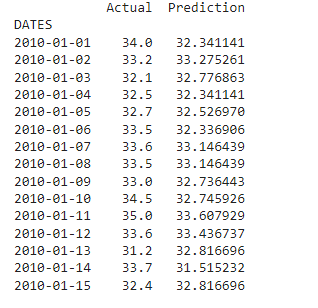


Figure 40: Actual result vs predicted result of the XGBoost regression model.

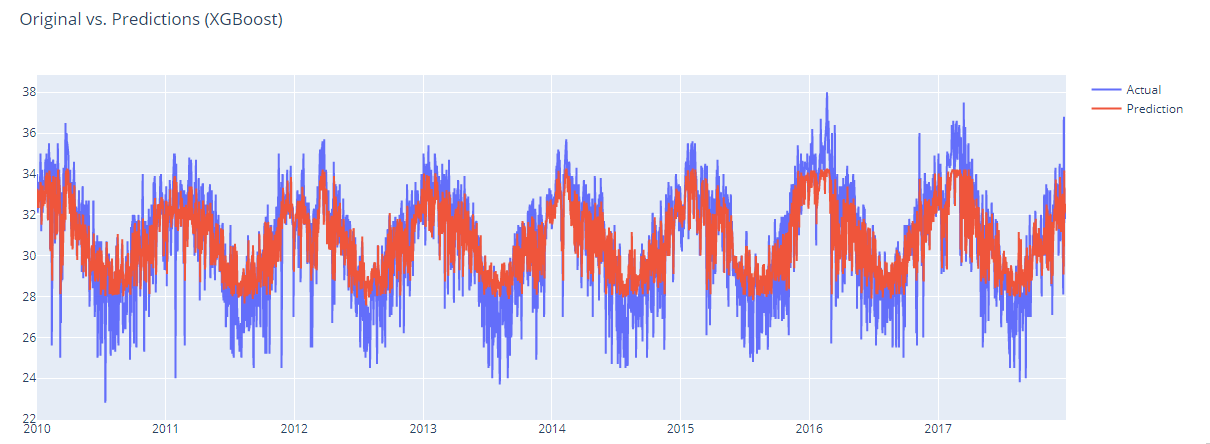


Figure 41: Line plot of actual result vs predicted result of the XGBoost regression model.