Analysis for Dataset 'Infrared Thermography'

799 DSAN CapStone Project

Zhen Li

2024-06-01

Abstract:

In public gathering areas such as airports, train stations, hospitals, and schools, there is a high likelihood of becoming transmission hubs for infectious diseases. If potential patients can be quickly identified and further screened, it could greatly reduce the risk of spreading infectious diseases. Fever is typically a very common symptom of illness, especially for diseases that can severely threaten public health, such as the recently experienced global outbreak of the COVID19.

When assessing fever symptoms, the oral temperature is the true value we aim to obtain. Measuring oral temperature for large crowds involves hygiene and time efficiency issues. Without direct contact with the human body, infrared thermography can quickly capture facial temperature. However, the measurement values of infrared thermography can be affected by other factors, thereby reducing the reference value of its readings. The goal of this project is to attempt to establish a good model that accurately predicts the oral temperature of individuals using quickly obtained facial temperature from certain areas and other easily measurable parameters.

If the model can provide good predictive performance, it can preliminary and rapidly screen out individuals with fever symptoms for further examination, playing a very positive role in preventing the spread of pandemic diseases.

Data Understanding:

- **The data source**

The original dataset, FLIR_groups1and2.csv, is available for download at https://doi.org/10.13026/3bhc-9065 (Wang et al., 2023). This file contains 1,020 observations and four rounds of measurements across 120 variables. For this project, which aims to develop a model for rapid screening, we have retained 32 variables from the first round of measurements, focusing on infrared thermography and environmental parameters, for our analysis.

- **Variables list**

As listed below, oral temperature is our target value. We aim to predict this target value based on the measurements of the feature (input) variables.

Variable Name	Role	Type	Description
SubjectID	ID	Categorical	Subject ID
T_OR_Max1	Target	Continuous	Oral temperature
Gender	Feature	Categorical	Male or Female
Age	Feature	Categorical	Age ranges in categories
Ethnicity	Feature	Categorical	American Indian or Alaska Native, Asian, Black or African America, Hispanic/Latino, Multiracial, Native Hawaiian or other Pacific Islander, white.
T_atm	Feature	Continuous	Ambiant temperature
Humidity	Feature	Continuous	Relative humidity
Distance	Feature	Continuous	Distance between the subjects and the IRTs
T_offset1	Feature	Continuous	Temperature difference between the set and measured blackbody temperature
Max1R13_1	Feature	Continuous	Max value of a circle with diameter of 13 pixels from the right canthus point to the face centerline
Max1L13_1	Feature	Continuous	Max value of a circle with diameter of 13 pixels from the left canthus point to the face centerline
aveAllR13_1	Feature	Continuous	Average value of a circle with diameter of 13 pixels from the right canthus point to the face centerline

Variable Name	Role	Type	Description
aveAllL13_1	Feature	Continuous	Average value of a circle with diameter of 13 pixels from the left canthus point to the face centerline
T_RC1	Feature	Continuous	Average temperature of the highest four pixels in a square of 24x24 pixels around the right canthus, with 2/3 toward the face center (dry area, 16x24 pixels) and 1/3 away from the face center (wet area, 8x24 pixels).
T_RC_Dry1	Feature	Continuous	Average temperature of the highest four pixels in the right canthus dry area, a rectangle of 16x24 pixels.
T_RC_Wet1	Feature	Continuous	Average temperature of the highest four pixels in the right canthus wet area, a rectangle of 8x24 pixels.
T_RC_Max1	Feature	Continuous	Max value of a square of 24x24 pixels around the right canthus, with 2/3 toward the face center (dry area, 16x24 pixels) and 1/3 away from the face center (wet area, 8x24 pixels).

Variable Name	Role	Type	Description
T_LC1	Feature	Continuous	Average temperature of the highest four pixels in a square of 24x24 pixels around the left canthus, with 2/3 toward the face center (dry area, 16x24 pixels) and 1/3 away from the face center (wet area, 8x24 pixels).
T_LC_Dry1	Feature	Continuous	Average temperature of the highest four pixels in the left canthus dry area, a rectangle of 16x24 pixels.
T_LC_Wet1	Feature	Continuous	Average temperature of the highest four pixels in the left canthus wet area, a rectangle of 16x24 pixels.
T_LC_Max1	Feature	Continuous	Max value of a circle with diameter of 13 pixels from the left canthus point to the face centerline
RCC1	Feature	Continuous	Average value of a square of 3x3 pixels centered at the right canthus point.
LCC1	Feature	Continuous	Average value of a square of 3x3 pixels centered at the left canthus point.
canthiMax1	Feature	Continuous	Max value in the extended canthi area
T_FHCC1	Feature	Continuous	Average temperature within Center point of forehead, a square of 3x3 pixels

Variable Name	Role	Type	Description
T_FHRC1	Feature	Continuous	Average temperature within Right point of the forehead, a square of 3x3 pixels
T_FHLC1	Feature	Continuous	Average temperature within Left point of the forehead, a square of 3x3 pixels
T_FHBC1	Feature	Continuous	Average temperature within Bottom point of the forehead, a square of 3x3 pixels
T_FHTC1	Feature	Continuous	Average temperature within Top point of the forehead, a square of 3x3 pixels
T_FH_Max1	Feature	Continuous	Maximum temperature within the extended forehead area
T_FHC_Max1	Feature	Continuous	Max value in the Center point of forehead, a square of 3x3 pixels
T_Max1	Feature	Continuous	Maximum temperature within the whole face region

- **Measure Guidance**

Figure 1 illustrates the methods used to measure all the temperatures. The image is provided by one of the authors who donated the original dataset.

Figure 1

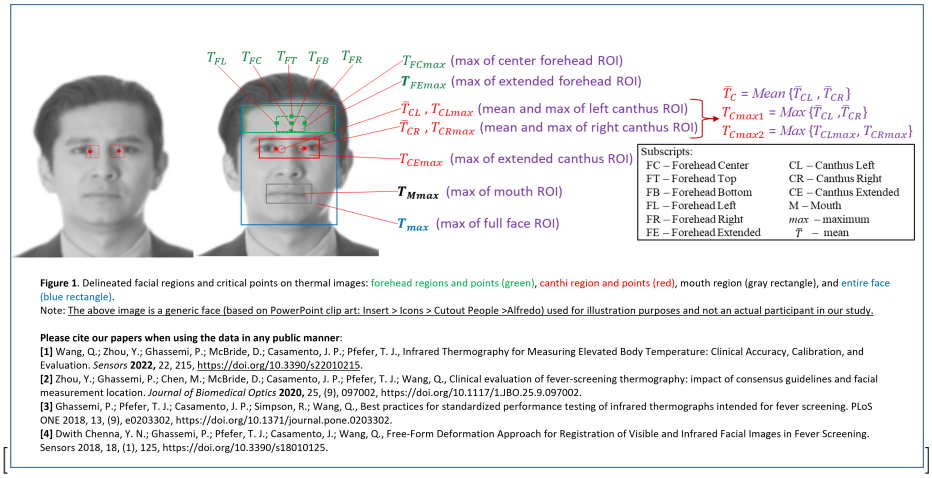
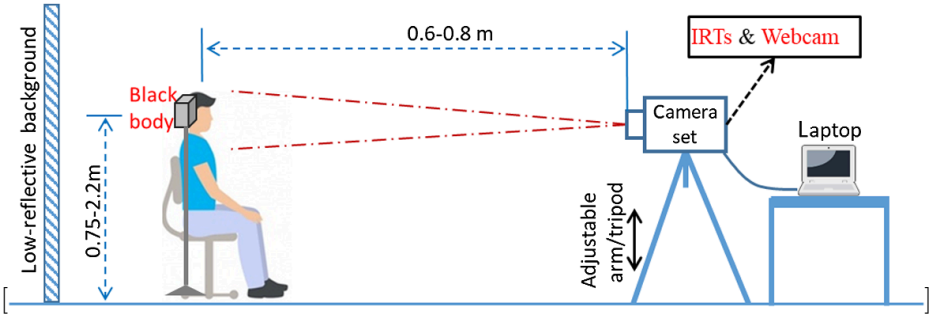


Figure 1. Delineated facial regions and critical points on thermal images: forehead regions and points (green), canthi region and points (red), mouth region (gray rectangle), and entire face (blue rectangle).
Note: The above image is a generic face (based on PowerPoint clip art: Insert > Icons > Cutout People >Alfredo) used for illustration purposes and not an actual participant in our study.

Please cite our papers when using the data in any public manner:
[1] Wang, Q.; Zhou, Y.; Ghassemi, P.; McBride, D.; Casamento, J. P.; Pfefer, T. J., Infrared Thermography for Measuring Elevated Body Temperature: Clinical Accuracy, Calibration, and Evaluation. *Sensors* **2022,** 22, 215, https://doi.org/10.3390/s22010215.
[2] Zhou, Y.; Ghassemi, P.; Chen, M.; McBride, D.; Casamento, J. P.; Pfefer, T. J.; Wang, Q., Clinical evaluation of fever-screening thermography: impact of consensus guidelines and facial measurement location. *Journal of Biomedical Optics* **2020**, 25, (9), 097002, https://doi.org/10.1117/1.JBO.25.9.097002.
[3] Ghassemi, P.; Pfefer, T. J.; Casamento, J. P.; Simpson, R.; Wang, Q., Best practices for standardized performance testing of infrared thermographs intended for fever screening. PLoS ONE 2018, 13, (9), e0203302, https://doi.org/10.1371/journal.pone.0203302.
[4] Dwith Chenna, Y. N.; Ghassemi, P.; Pfefer, T. J.; Casamento, J.; Wang, Q., Free-Form Deformation Approach for Registration of Visible and Infrared Facial Images in Fever Screening. Sensors 2018, 18, (1), 125, https://doi.org/10.3390/s18010125.

Figure 2 serves as a guide for taking measurements. Additionally, environmental parameters will be recorded as input variables.

Figure 2



Data Prepariation:

- **Load Necessary Library**

The code listed below includes all the R packages used in this analysis. Please note that some functions defined in different packages share the same names; the function in the last loaded package will be used by default. To use a specific package's function, use the '::' operator.

```
library(tidyverse) #include "dplyr" "ggplot2"
library(Hmisc)
library(psych)
library(scales)
library(caTools) # seed
library(patchwork) # multiplot
library(car) # vif
```

```
library(reshape2) # melt() to convert wide format to long format
library(caret)
library(glmnet)
library(randomForest)
```

- **Brief Look at the Selected Data**

Firstly, let's examine the selected variables for our analysis. As listed below, there are a total of 32 variables and 1,020 observations. We notice that some variables, such as 'T_offset1', have missing values. Additionally, some categorical data, like 'Age', have overlapping ranges such as '21-25', '26-30', and '21-30', which should be combined into a single group to reduce complexity.

```
data_selected
```

```
 32  Variables      1020  Observations
---------------------------------------------------------------------------
T_offset1
       n  missing distinct
    1003       17      196

lowest : -0.06 -0.07 -0.08 -0.10 -0.11, highest: 2.55  2.62  2.89  2.90  3.58
---------------------------------------------------------------------------
Max1R13_1
       n  missing distinct
    1004       16      257

lowest : 33.35 33.61 33.76 33.82 33.88, highest: 37.92 37.93 38.11 38.28 38.52
---------------------------------------------------------------------------
Max1L13_1
       n  missing distinct
    1004       16      249

lowest : 33.55 33.73 33.77 33.84 33.85, highest: 37.91 37.94 38.03 38.05 38.16
---------------------------------------------------------------------------
aveAllR13_1
       n  missing distinct
    1004       16      315

lowest : 29.95 31.62 31.99 32.02 32.08, highest: 37.18 37.28 37.64 37.69 37.74
---------------------------------------------------------------------------
aveAllL13_1
       n  missing distinct
    1004       16      294

lowest : 31.54 32.21 32.53 32.67 32.69, highest: 37.10 37.32 37.55 37.57 37.78
```

```
---------------------------------------------------------------------------
T_RC1
       n  missing distinct
    1003       17      244

lowest : 33.34 33.69 33.89 34.18 34.19, highest: 37.89 37.91 38.10 38.26 38.51
---------------------------------------------------------------------------
T_RC_Dry1
       n  missing distinct
    1003       17      249

lowest : 33.34 33.53 33.64 33.73 33.85, highest: 37.89 37.91 38.10 38.25 38.51
---------------------------------------------------------------------------
T_RC_Wet1
       n  missing distinct
    1003       17      260

lowest : 33.30 33.46 33.69 33.88 33.89, highest: 37.74 37.83 38.06 38.22 38.41
---------------------------------------------------------------------------
T_RC_Max1
       n  missing distinct
    1003       17      247

lowest : 33.35 33.76 33.90 34.20 34.22, highest: 37.92 37.93 38.13 38.28 38.52
---------------------------------------------------------------------------
T_LC1
       n  missing distinct
    1003       17      247

lowest : 33.78 33.82 33.90 33.92 33.95, highest: 37.88 37.91 38.00 38.03 38.16
---------------------------------------------------------------------------
T_LC_Dry1
       n  missing distinct
    1003       17      250

lowest : 33.67 33.72 33.77 33.82 33.90, highest: 37.86 37.91 38.00 38.03 38.16
---------------------------------------------------------------------------
T_LC_Wet1
       n  missing distinct
    1003       17      272

lowest : 33.12 33.43 33.53 33.58 33.61, highest: 37.73 37.82 37.86 37.90 37.93
---------------------------------------------------------------------------
T_LC_Max1
       n  missing distinct
    1003       17      246
```

```
lowest : 33.79 33.85 33.92 33.95 34.02, highest: 37.92 37.94 38.03 38.05 38.21
---------------------------------------------------------------------------
RCC1
       n  missing distinct
    1003       17      270

lowest : 32.78 33.16 33.30 33.36 33.40, highest: 37.52 37.76 37.77 37.85 38.31
---------------------------------------------------------------------------
LCC1
       n  missing distinct
    1003       17      279

lowest : 32.57 32.82 32.96 33.19 33.29, highest: 37.50 37.59 37.64 37.76 37.91
---------------------------------------------------------------------------
canthiMax1
       n  missing distinct
    1003       17      242

lowest : 33.85 34.30 34.40 34.42 34.44, highest: 37.94 38.03 38.13 38.28 38.52
---------------------------------------------------------------------------
canthi4Max1
       n  missing distinct
    1003       17      247

lowest : 33.82 34.28 34.34 34.39 34.41, highest: 37.92 38.00 38.10 38.26 38.51
---------------------------------------------------------------------------
T_FHCC1
       n  missing distinct
    1003       17      289

lowest : 30.02 30.93 31.21 31.22 31.65, highest: 36.61 36.70 36.77 36.89 37.18
---------------------------------------------------------------------------
T_FHRC1
       n  missing distinct
    1003       17      300

lowest : 29.54 30.48 30.60 30.92 31.05, highest: 36.54 36.60 36.74 36.84 37.24
---------------------------------------------------------------------------
T_FHLC1
       n  missing distinct
    1003       17      308

lowest : 30.21 31.10 31.17 31.47 31.68, highest: 36.54 36.57 36.84 37.06 37.07
---------------------------------------------------------------------------
T_FHBC1
```

```
       n  missing distinct
    1003       17      307

lowest : 30.26 30.74 31.35 31.42 31.50, highest: 36.50 36.52 36.68 37.03 37.27
---------------------------------------------------------------------------
T_FHTC1
       n  missing distinct
    1003       17      320

lowest : 27.70 30.04 30.82 31.02 31.07, highest: 36.57 36.74 36.75 36.76 37.44
---------------------------------------------------------------------------
T_FH_Max1
       n  missing distinct
    1003       17      244

lowest : 33.08 33.10 33.18 33.46 33.53, highest: 37.45 37.50 37.53 37.64 38.02
---------------------------------------------------------------------------
T_FHC_Max1
       n  missing distinct
    1003       17      275

lowest : 31.63 31.70 32.35 32.48 32.75, highest: 37.00 37.06 37.15 37.38 37.58
---------------------------------------------------------------------------
T_Max1
       n  missing distinct
    1003       17      229

lowest : 33.85 34.47 34.52 34.58 34.60, highest: 37.94 38.03 38.13 38.28 38.52
---------------------------------------------------------------------------
T_OR_Max1
       n  missing distinct
    1003       17      283

lowest : 32.18 32.33 32.53 32.62 32.68, highest: 37.31 37.36 37.40 37.50 37.76
---------------------------------------------------------------------------
Gender
       n  missing distinct
    1020        0        2

Value      Female   Male
Frequency     606    414
Proportion  0.594  0.406
---------------------------------------------------------------------------
Age
       n  missing distinct
    1020        0        8
```

```
Value        >60 18-20 21-25 21-30 26-30 31-40 41-50 51-60
Frequency      3   534   355    10    67    31     9    11
Proportion 0.003 0.524 0.348 0.010 0.066 0.030 0.009 0.011
---------------------------------------------------------------------------
Ethnicity
       n  missing distinct
    1020        0        6

American Indian or Alaskan Native (4, 0.004), Asian (260, 0.255), Black or
African-American (143, 0.140), Hispanic/Latino (57, 0.056), Multiracial (50,
0.049), White (506, 0.496)
---------------------------------------------------------------------------
T_atm
       n  missing distinct
    1020        0       78

lowest : 20.2 20.3 20.5 20.6 20.7, highest: 28.2 28.5 28.6 28.7 29.1
---------------------------------------------------------------------------
Humidity
       n  missing distinct
    1020        0      353

lowest : 10.2 10.3 10.6 10.7 10.8, highest: 60.3 60.5 61   61.2 9.9
---------------------------------------------------------------------------
Distance
       n  missing distinct
    1018        2       33

lowest : 0.54 0.55 0.56 0.57 0.58, highest: 0.84 0.85 0.9  0.92 79
---------------------------------------------------------------------------
```

- **Dealing the Missing Data and Combine Age Group**

After removing the missing values and combining age groups, the cleaned data consists of 32 variables and 1,001 observations. While the data quality has improved, some values still appear to be incorrect. For instance, the highest value of 'Distance' is 79, which seems unreasonable.

```
data_cleaned

 32  Variables      1001  Observations
---------------------------------------------------------------------------
T_offset1
       n  missing distinct
    1001        0      196
```

```
lowest : -0.06 -0.07 -0.08 -0.10 -0.11, highest: 2.55  2.62  2.89  2.90  3.58
-------------------------------------------------------------------------------
Max1R13_1
       n  missing distinct
    1001        0      257

lowest : 33.35 33.61 33.76 33.82 33.88, highest: 37.92 37.93 38.11 38.28 38.52
-------------------------------------------------------------------------------
Max1L13_1
       n  missing distinct
    1001        0      249

lowest : 33.55 33.73 33.77 33.84 33.85, highest: 37.91 37.94 38.03 38.05 38.16
-------------------------------------------------------------------------------
aveAllR13_1
       n  missing distinct
    1001        0      315

lowest : 29.95 31.62 31.99 32.02 32.08, highest: 37.18 37.28 37.64 37.69 37.74
-------------------------------------------------------------------------------
aveAllL13_1
       n  missing distinct
    1001        0      294

lowest : 31.54 32.21 32.53 32.67 32.69, highest: 37.10 37.32 37.55 37.57 37.78
-------------------------------------------------------------------------------
T_RC1
       n  missing distinct
    1001        0      244

lowest : 33.34 33.69 33.89 34.18 34.19, highest: 37.89 37.91 38.10 38.26 38.51
-------------------------------------------------------------------------------
T_RC_Dry1
       n  missing distinct
    1001        0      249

lowest : 33.34 33.53 33.64 33.73 33.85, highest: 37.89 37.91 38.10 38.25 38.51
-------------------------------------------------------------------------------
T_RC_Wet1
       n  missing distinct
    1001        0      260

lowest : 33.30 33.46 33.69 33.88 33.89, highest: 37.74 37.83 38.06 38.22 38.41
-------------------------------------------------------------------------------
T_RC_Max1
       n  missing distinct
```

```
    1001        0      247

lowest : 33.35 33.76 33.90 34.20 34.22, highest: 37.92 37.93 38.13 38.28 38.52
---------------------------------------------------------------------------
T_LC1
       n  missing distinct
    1001        0      247

lowest : 33.78 33.82 33.90 33.92 33.95, highest: 37.88 37.91 38.00 38.03 38.16
---------------------------------------------------------------------------
T_LC_Dry1
       n  missing distinct
    1001        0      250

lowest : 33.67 33.72 33.77 33.82 33.90, highest: 37.86 37.91 38.00 38.03 38.16
---------------------------------------------------------------------------
T_LC_Wet1
       n  missing distinct
    1001        0      272

lowest : 33.12 33.43 33.53 33.58 33.61, highest: 37.73 37.82 37.86 37.90 37.93
---------------------------------------------------------------------------
T_LC_Max1
       n  missing distinct
    1001        0      246

lowest : 33.79 33.85 33.92 33.95 34.02, highest: 37.92 37.94 38.03 38.05 38.21
---------------------------------------------------------------------------
RCC1
       n  missing distinct
    1001        0      270

lowest : 32.78 33.16 33.30 33.36 33.40, highest: 37.52 37.76 37.77 37.85 38.31
---------------------------------------------------------------------------
LCC1
       n  missing distinct
    1001        0      279

lowest : 32.57 32.82 32.96 33.19 33.29, highest: 37.50 37.59 37.64 37.76 37.91
---------------------------------------------------------------------------
canthiMax1
       n  missing distinct
    1001        0      242

lowest : 33.85 34.30 34.40 34.42 34.44, highest: 37.94 38.03 38.13 38.28 38.52
---------------------------------------------------------------------------
```

```
canthi4Max1
       n  missing distinct
    1001        0      247

lowest : 33.82 34.28 34.34 34.39 34.41, highest: 37.92 38.00 38.10 38.26 38.51
--------------------------------------------------------------------------------
T_FHCC1
       n  missing distinct
    1001        0      289

lowest : 30.02 30.93 31.21 31.22 31.65, highest: 36.61 36.70 36.77 36.89 37.18
--------------------------------------------------------------------------------
T_FHRC1
       n  missing distinct
    1001        0      300

lowest : 29.54 30.48 30.60 30.92 31.05, highest: 36.54 36.60 36.74 36.84 37.24
--------------------------------------------------------------------------------
T_FHLC1
       n  missing distinct
    1001        0      308

lowest : 30.21 31.10 31.17 31.47 31.68, highest: 36.54 36.57 36.84 37.06 37.07
--------------------------------------------------------------------------------
T_FHBC1
       n  missing distinct
    1001        0      306

lowest : 30.26 30.74 31.35 31.42 31.50, highest: 36.50 36.52 36.68 37.03 37.27
--------------------------------------------------------------------------------
T_FHTC1
       n  missing distinct
    1001        0      319

lowest : 27.70 30.04 30.82 31.02 31.07, highest: 36.57 36.74 36.75 36.76 37.44
--------------------------------------------------------------------------------
T_FH_Max1
       n  missing distinct
    1001        0      244

lowest : 33.08 33.10 33.18 33.46 33.53, highest: 37.45 37.50 37.53 37.64 38.02
--------------------------------------------------------------------------------
T_FHC_Max1
       n  missing distinct
    1001        0      275
```

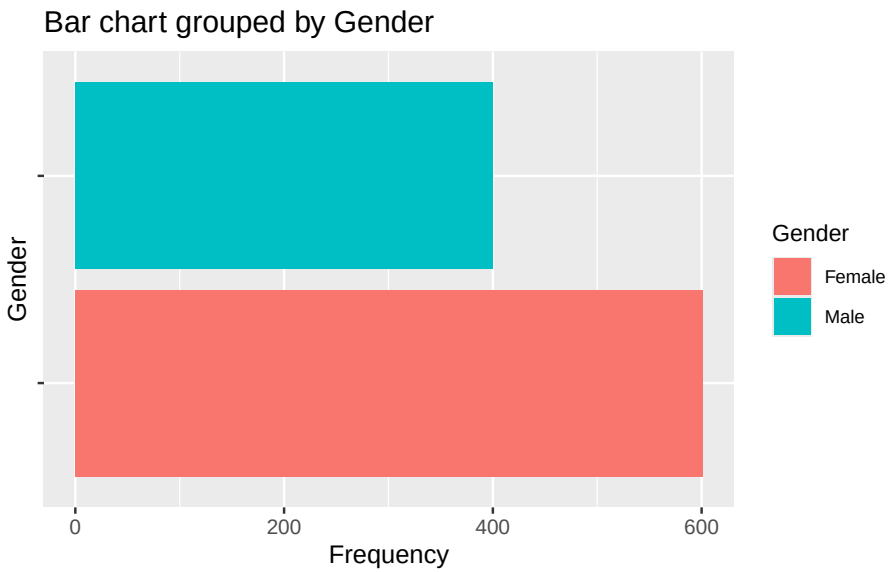
```
lowest : 31.63 31.70 32.35 32.48 32.75, highest: 37.00 37.06 37.15 37.38 37.58
---------------------------------------------------------------------------
T_Max1
       n  missing distinct
    1001        0      229

lowest : 33.85 34.47 34.52 34.58 34.60, highest: 37.94 38.03 38.13 38.28 38.52
---------------------------------------------------------------------------
T_OR_Max1
       n  missing distinct
    1001        0      283

lowest : 32.18 32.33 32.53 32.62 32.68, highest: 37.31 37.36 37.40 37.50 37.76
---------------------------------------------------------------------------
Gender
       n  missing distinct
    1001        0        2

Value      Female   Male
Frequency     601    400
Proportion    0.6    0.4
---------------------------------------------------------------------------
Age
       n  missing distinct
    1001        0        6

Value      18-20 21-30 31-40 41-50 51-60   60+
Frequency    524   423    31     9    11     3
Proportion 0.523 0.423 0.031 0.009 0.011 0.003
---------------------------------------------------------------------------
Ethnicity
       n  missing distinct
    1001        0        6

American Indian or Alaskan Native (4, 0.004), Asian (255, 0.255), Black or
African-American (143, 0.143), Hispanic/Latino (57, 0.057), Multiracial (49,
0.049), White (493, 0.493)
---------------------------------------------------------------------------
T_atm
       n  missing distinct
    1001        0       78

lowest : 20.2 20.3 20.5 20.6 20.7, highest: 28.2 28.5 28.6 28.7 29.1
---------------------------------------------------------------------------
Humidity
       n  missing distinct
```

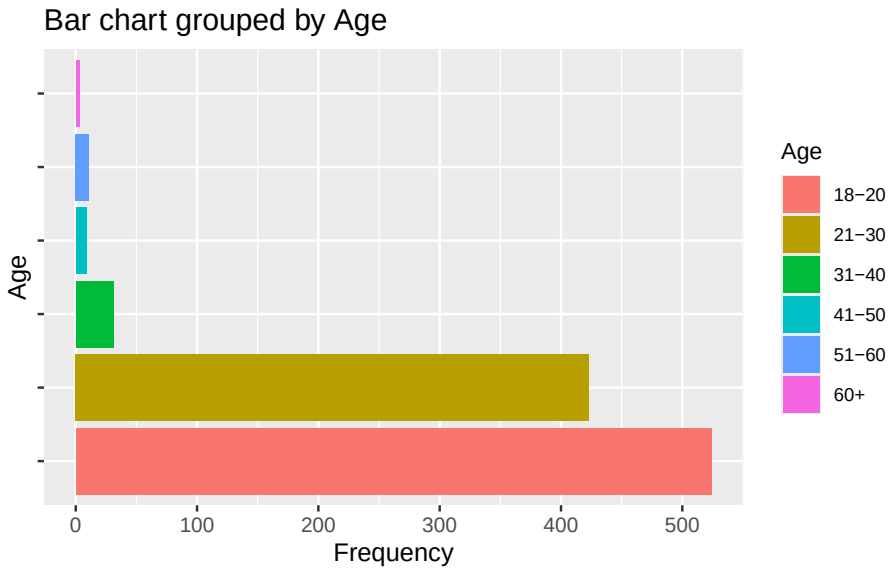
```
    1001        0      350

lowest : 10.2 10.3 10.6 10.7 10.8, highest: 60.3 60.5 61   61.2 9.9
------------------------------------------------------------------------------
Distance
       n  missing distinct
    1001        0       33

lowest : 0.54 0.55 0.56 0.57 0.58, highest: 0.84 0.85 0.9  0.92 79
------------------------------------------------------------------------------
```

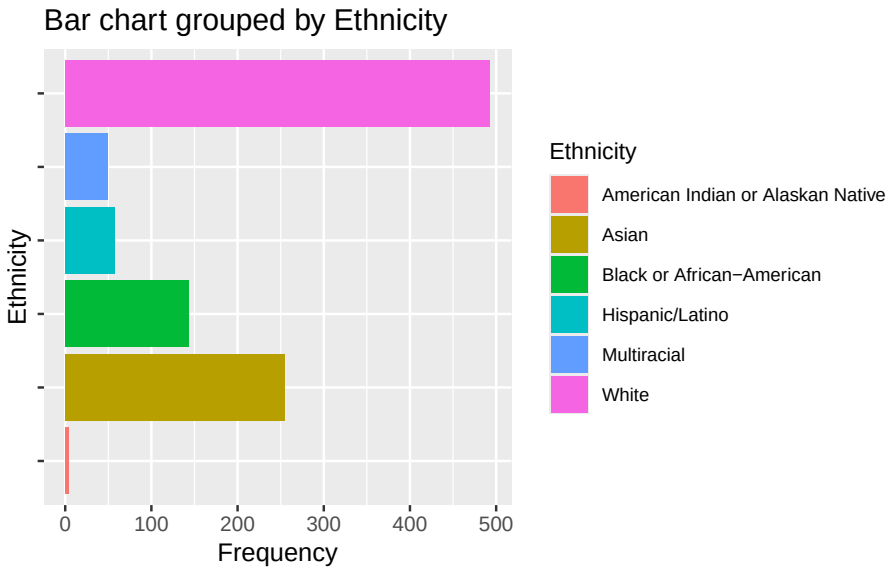
- **Plot Categorical Data**

To better understand the data, we begin by plotting the categorical data.

1. The test subjects are fairly evenly distributed between males and females, with a ratio of 6:4.
2. The dataset is predominantly composed of individuals under 30, suggesting that the model built on this data may not be suitable for all age groups.
3. Similarly, the distribution of ethnic groups is unbalanced, indicating that the model based on this dataset may not be appropriate for all ethnicities.







- **Process with Categorical Data to Numerical**

To build the model, we convert categorical data to numerical values.

1. We set "gender is female" as the baseline, meaning that when "gender is female," the value of Gender_M is zero.
2. Based on the plot presented above, we will not consider the age group.
3. For the same reason, we will not consider the ethnicity group.

- **Self Defined Descriptive statistics**

To better describe the numerical data from the original measurements, we define a function that provides basic information such as sample size, mean, standard deviation, skewness, and kurtosis.

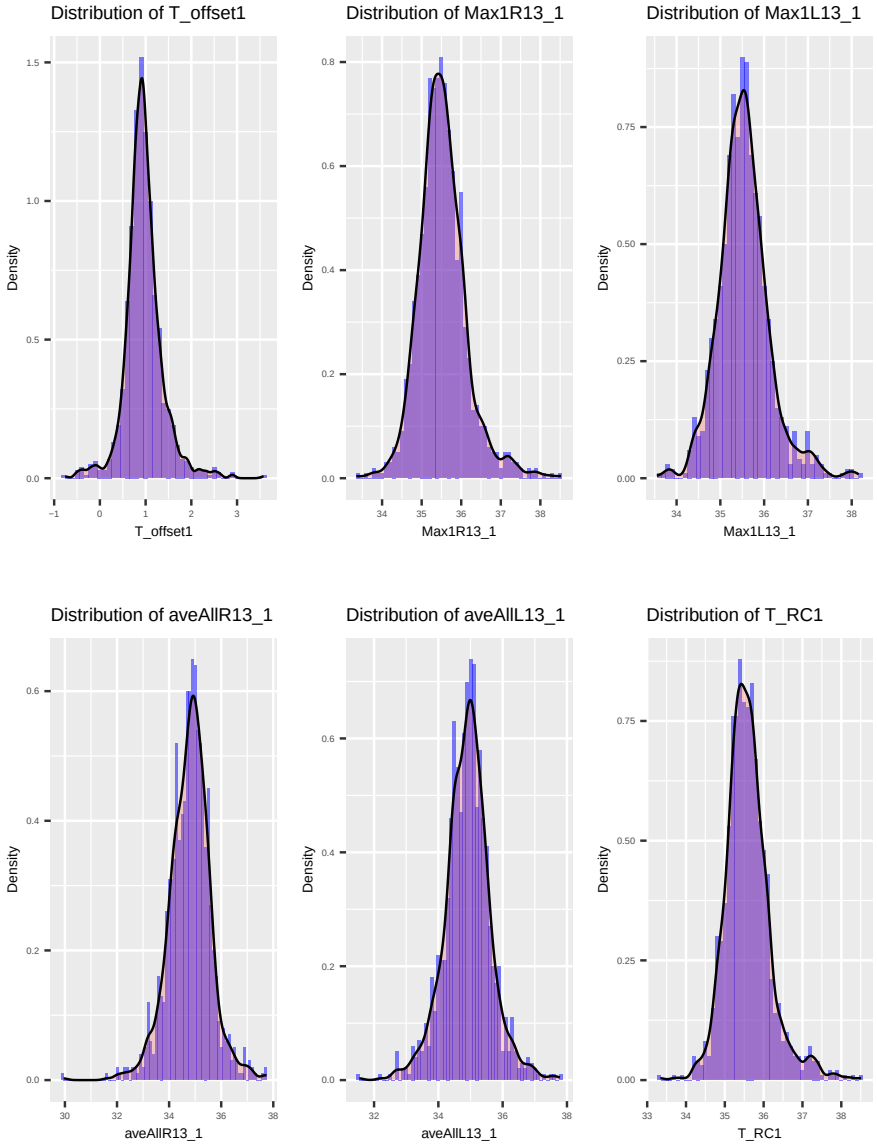
1. Sample Size (n): Each variable has 1001 observations, indicating a substantial amount of data.
2. Mean: The means of the variables range from around 24 to 36, suggesting the data are centered around these values. The temperature-related variables generally have means around 35, while 'Humidity' has a lower mean (28.76), and 'Distance' has a very low mean (0.731).
3. Standard Deviation (stdev): The standard deviation values indicate the spread of the data. For most temperature-related variables, the standard deviations are around 0.5 to 0.9, indicating relatively low variability. 'Humidity' has a higher variability (stdev of 13.07), which is notable. 'Distance' also shows substantial variability (stdev of 2.48).
4. Skewness: Positive skewness values indicate a right-skewed distribution (e.g., T_offset1, Max1R13_1, etc.). Negative skewness values indicate a left-skewed distribution (e.g., aveAllR13_1, T_FHCC1). Most temperature variables show slight to moderate skewness, indicating some asymmetry in the data distribution.
5. Kurtosis: Kurtosis values close to 3 indicate a normal distribution. Values less than 3 indicate a flatter distribution (platykurtic), while values greater than 3 indicate a more peaked distribution (leptokurtic). The kurtosis values in this dataset range from around 1.64 to 4.88, with most being around 2, suggesting generally platykurtic distributions, except T_FHTC1 which is leptokurtic.
6. Outliers: High skewness and kurtosis in some variables (e.g., T_FHTC1 with kurtosis of 4.88 and skewness of -1.33, Distance with kurtosis of 992.51 and skewness of 31.51) suggest the presence of outliers.

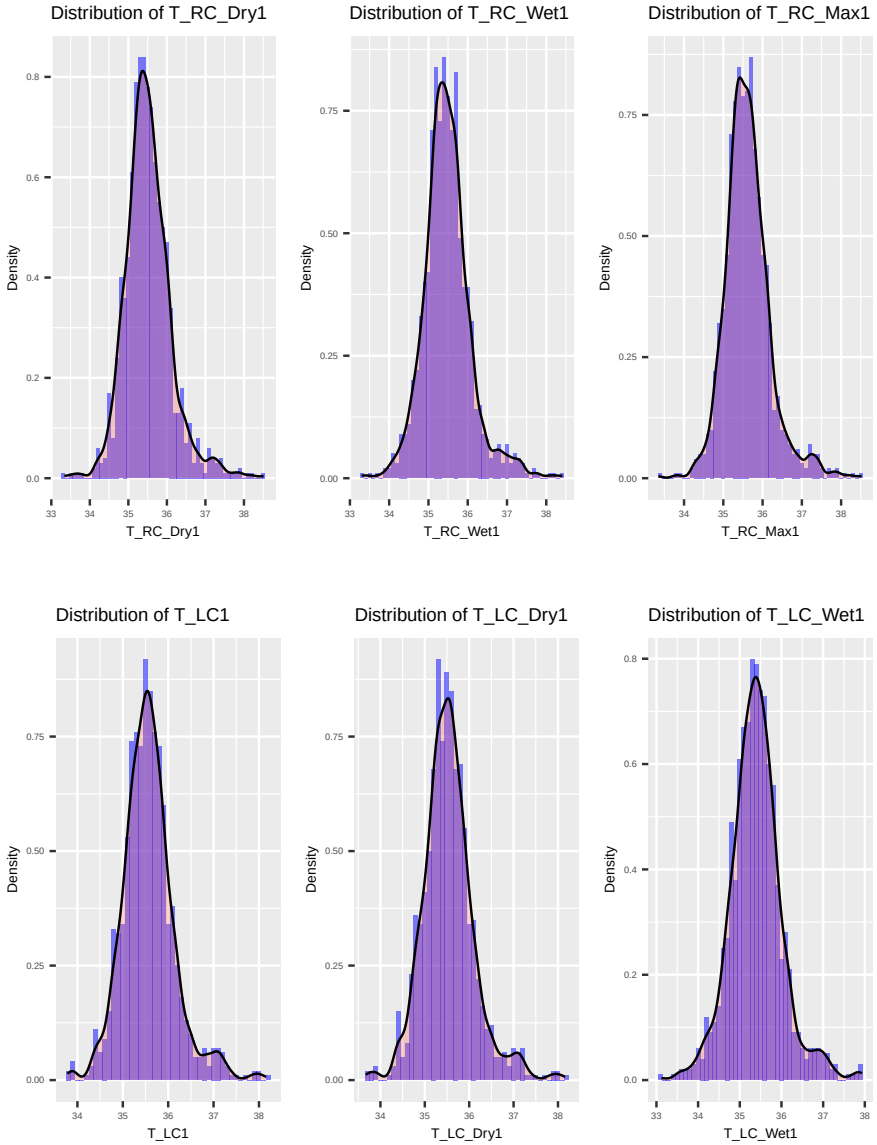
```
             T_offset1    Max1R13_1    Max1L13_1  aveAllR13_1   aveAllL13_1
n        1001.0000000 1001.0000000 1001.0000000 1001.0000000 1001.00000000
mean        0.9603896   35.5275325   35.5399401   34.7882517   34.90594406
stdev       0.4185555    0.6103086    0.5870455    0.8010727    0.71848071
skew        0.6325190    0.7788724    0.6380632   -0.2935453   -0.04830282
kurtosis    4.5039029    2.4011231    2.0659957    2.3833188    1.64233439
                T_RC1    T_RC_Dry1    T_RC_Wet1    T_RC_Max1        T_LC1
n        1001.0000000 1001.0000000 1001.0000000 1001.0000000 1001.0000000
mean       35.5922877   35.5142258   35.4787812   35.6221279   35.5701399
stdev       0.5837500    0.6031786    0.6059537    0.5830048    0.5743379
skew        0.8807364    0.8053378    0.7359167    0.8820616    0.7096845
kurtosis    2.7000235    2.5329753    2.3889692    2.7033613    2.0980680
            T_LC_Dry1    T_LC_Wet1    T_LC_Max1         RCC1         LCC1
```

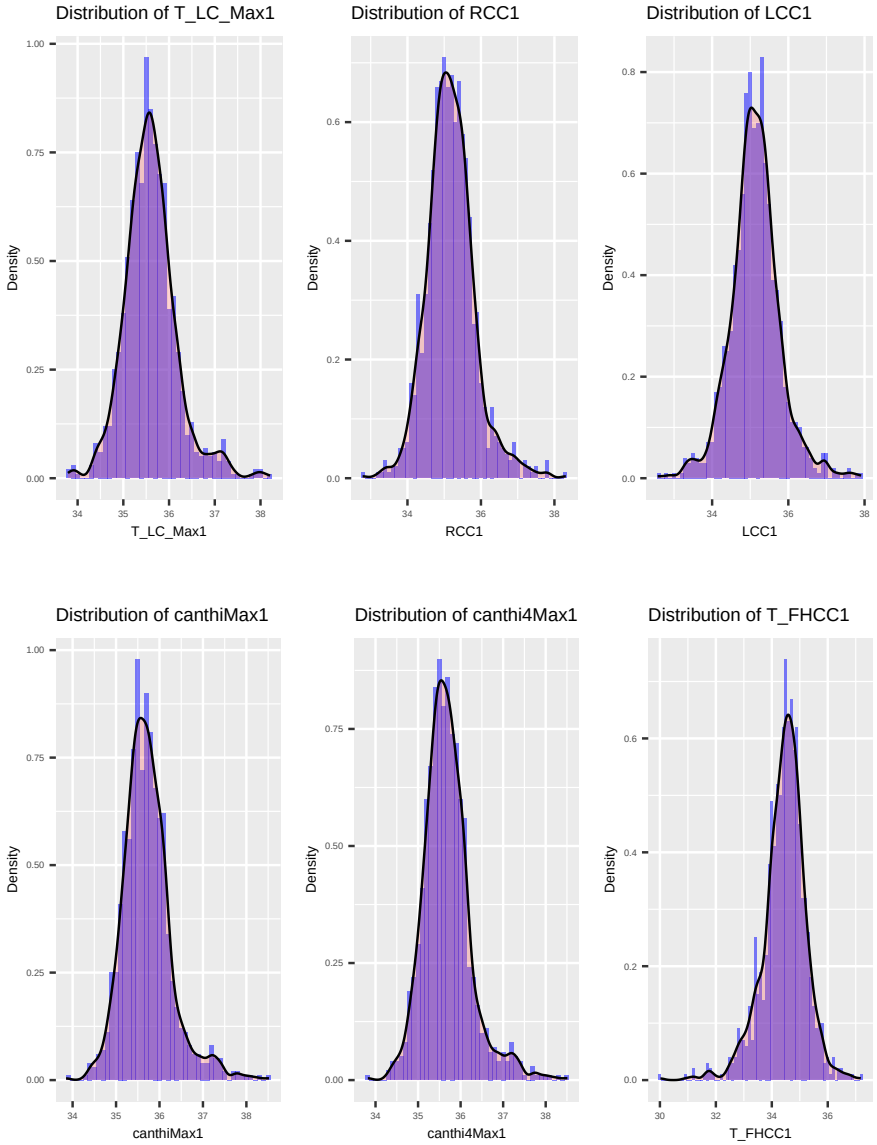
n	1001.0000000	1001.0000000	1001.0000000	1001.0000000	1001.0000000
mean	35.5362438	35.3946953	35.6015684	35.1816783	35.1327273
stdev	0.5804356	0.6229043	0.5741002	0.6536831	0.6546552
skew	0.6848343	0.4178882	0.7082639	0.5745866	0.2512231
kurtosis	2.1166300	1.6974900	2.0939820	1.8732465	1.8038732
	canthiMax1	canthi4Max1	T_FHCC1	T_FHRC1	T_FHLC1
n	1001.0000000	1001.0000000	1001.0000000	1001.0000000	1001.0000000
mean	35.7172128	35.6874426	34.4402498	34.4359740	34.4469930
stdev	0.5610792	0.5610754	0.7711031	0.7874477	0.7736153
skew	0.9910312	0.9952600	-0.7274469	-0.8856128	-0.6746324
kurtosis	2.6084703	2.6364560	2.6940562	3.5849533	2.2154211
	T_FHBC1	T_FHTC1	T_FH_Max1	T_FHC_Max1	T_Max1
n	1001.0000000	1001.0000000	1001.0000000	1001.0000000	1001.0000000
mean	34.3452947	34.4324975	35.3561538	34.9995405	35.8206094
stdev	0.7701425	0.9141997	0.5627050	0.6428375	0.5308622
skew	-0.6207007	-1.3304576	0.2122662	-0.4226499	1.0612231
kurtosis	2.4991055	4.8861606	2.2781230	2.6153255	2.9267537
	T_OR_Max1	T_atm	Humidity	Distance	
n	1001.0000000	1001.0000000	1001.0000000	1001.000000	
mean	35.3290809	24.1204795	28.7625375	0.731009	
stdev	0.6971657	1.3455966	13.0695289	2.477248	
skew	-0.2928126	0.5047649	0.6835440	31.508324	
kurtosis	1.7341084	1.2645024	-0.6031495	992.512306	

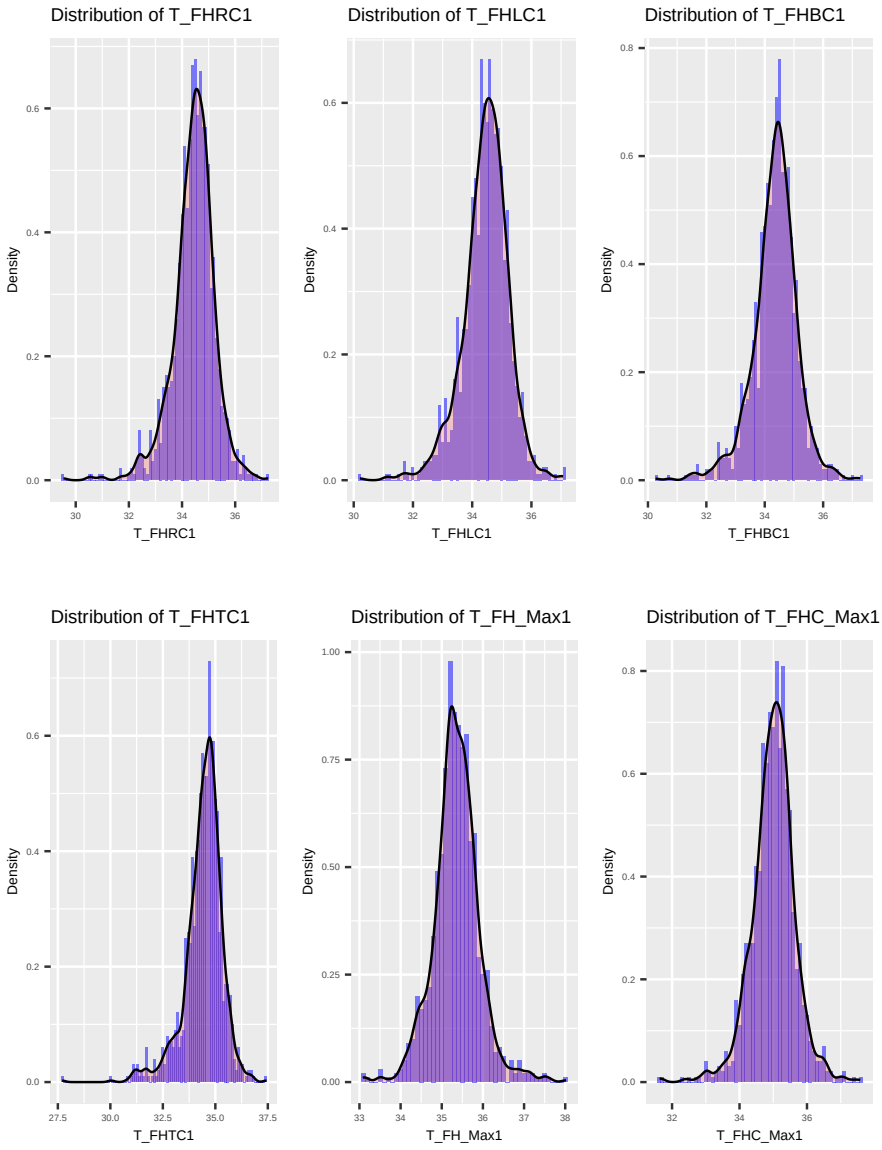
- **Plot Numerical Data**

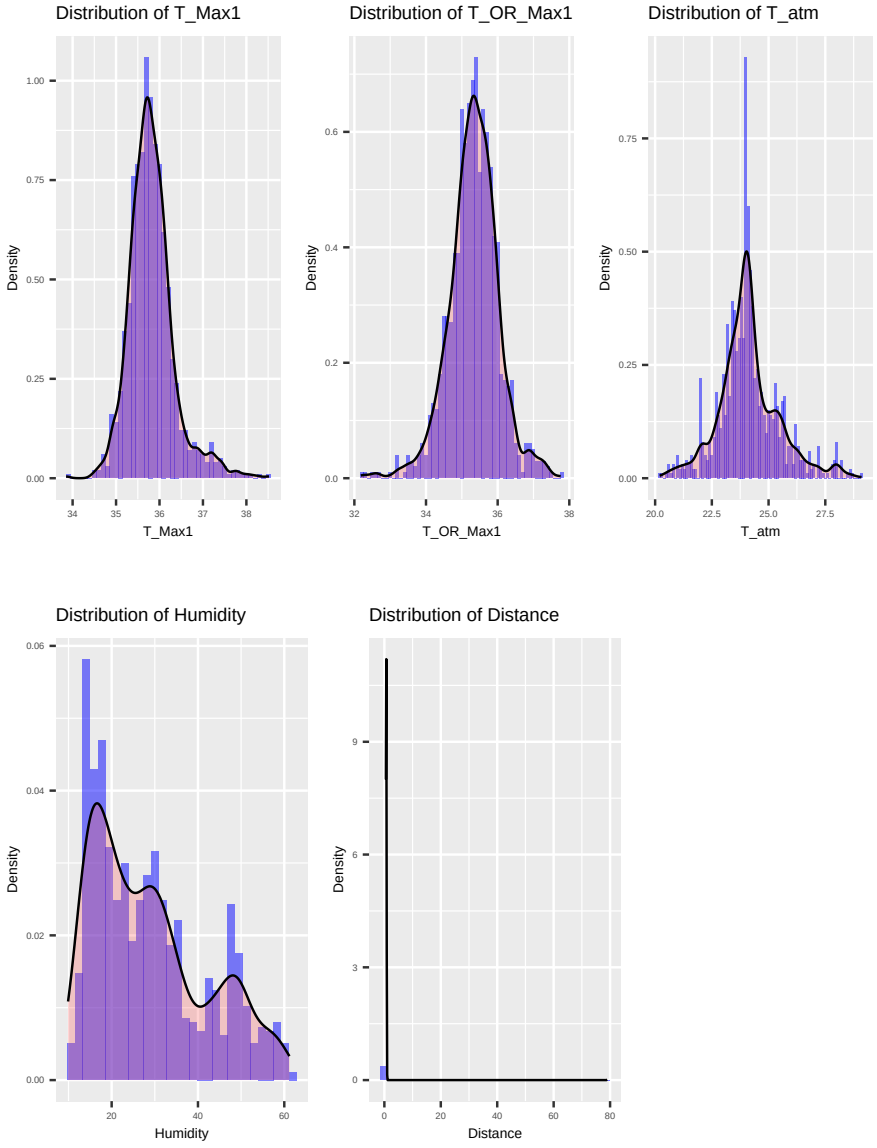
The plot of the numerical data confirms the findings from the descriptive statistics. For example, the distribution of the 'Distance' variable is skewed by outliers.


Distribution of T_offset1
Density
T_offset1
Distribution of Max1R13_1
Density
Max1R13_1
Distribution of Max1L13_1
Density
Max1L13_1
Distribution of aveAllR13_1
Density
aveAllR13_1
Distribution of aveAllL13_1
Density
aveAllL13_1
Distribution of T_RC1
Density
T_RC1


Distribution of T_RC_Dry1
Density
T_RC_Dry1
Distribution of T_RC_Wet1
Density
T_RC_Wet1
Distribution of T_RC_Max1
Density
T_RC_Max1
Distribution of T_LC1
Density
T_LC1
Distribution of T_LC_Dry1
Density
T_LC_Dry1
Distribution of T_LC_Wet1
Density
T_LC_Wet1


Distribution of T_LC_Max1
Density
T_LC_Max1
Distribution of RCC1
Density
RCC1
Distribution of LCC1
Density
LCC1
Distribution of canthiMax1
Density
canthiMax1
Distribution of canthi4Max1
Density
canthi4Max1
Distribution of T_FHCC1
Density
T_FHCC1


Distribution of T_FHRC1
Density
T_FHRC1
Distribution of T_FHLC1
Density
T_FHLC1
Distribution of T_FHBC1
Density
T_FHBC1
Distribution of T_FHTC1
Density
T_FHTC1
Distribution of T_FH_Max1
Density
T_FH_Max1
Distribution of T_FHC_Max1
Density
T_FHC_Max1



- **Dealing the Outlier**

To mitigate the impact of outliers, we consider filtering the data by setting lower and upper bounds based on the distance from the quantiles.

Interquartile Range (IQR)

$$\text{IQR} = Q3 - Q1$$

Lower Bound

$$lower\_bound = Q1 - 1.5 \times IQR$$

Upper Bound

$$upper\_bound = Q3 + 1.5 \times IQR$$

- **Self Defined Descriptive statistics w/o Outliers**

After removing the outliers from the dataset, 675 observations remain.

1. The skewness values are close to zero, indicating that the distribution is symmetrical with no significant skew to either the left or the right.
2. A kurtosis value close to zero suggests that the tails of the distribution are similar to those of a normal distribution, exhibiting neither heavy-tailed (leptokurtic) nor light-tailed (platykurtic) behavior.

	T_offset1	Max1R13_1	Max1L13_1	aveAllR13_1	aveAllL13_1
n	675.0000000	675.00000000	675.00000000	675.00000000	675.00000000
mean	0.9498667	35.51145185	35.52610370	34.81300741	34.92672593
stdev	0.2682311	0.41865791	0.39710012	0.57617561	0.50162978
skew	0.1699537	-0.03307242	-0.02561885	-0.49406712	-0.29366306
kurtosis	0.3269626	-0.19921366	-0.26949398	0.04932116	-0.01710634

	T_RC1	T_RC_Dry1	T_RC_Wet1	T_RC_Max1	T_LC1
n	675.00000000	675.00000000	675.00000000	675.00000000	675.00000000
mean	35.56583704	35.49743704	35.44700741	35.59620741	35.54657778
stdev	0.38532772	0.40890871	0.40533818	0.38524972	0.38751164
skew	0.07701227	0.02018552	-0.01964001	0.07536223	-0.04932664
kurtosis	-0.28719747	-0.23966447	-0.24502362	-0.28975583	-0.28779348

	T_LC_Dry1	T_LC_Wet1	T_LC_Max1	RCC1	LCC1
n	675.00000000	675.0000000	675.00000000	675.000000000	675.00000000
mean	35.51885926	35.3719407	35.57788148	35.163911111	35.12720000
stdev	0.39090328	0.4236659	0.38905781	0.451885392	0.44866007
skew	-0.03237607	-0.1009501	-0.04058744	-0.006771046	-0.06312115
kurtosis	-0.26949748	-0.3098021	-0.26418090	-0.366994525	-0.15111713

	canthiMax1	canthi4Max1	T_FHCC1	T_FHRC1	T_FHLC1
n	675.0000000	675.00000000	675.0000000	675.0000000	675.0000000
mean	35.6853778	35.65548148	34.5260444	34.5206222	34.5346815
stdev	0.3686775	0.36702291	0.4778338	0.4948308	0.5034498
skew	0.0559423	0.05071196	-0.2723006	-0.3076018	-0.2310515
kurtosis	-0.3832508	-0.37831776	-0.1206053	-0.1587055	-0.2403635

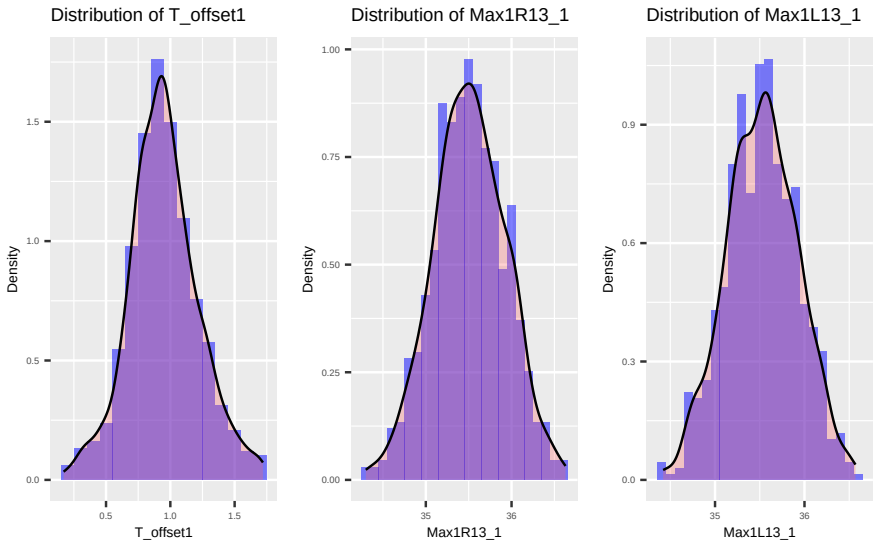
	T_FHBC1	T_FHTC1	T_FH_Max1	T_FHC_Max1	T_Max1
n	675.0000000	675.0000000	675.00000000	675.0000000	675.00000000
mean	34.4235111	34.6015407	35.38080000	35.0522519	35.77832593
stdev	0.4660465	0.5055869	0.35305023	0.4049863	0.33395039
skew	-0.2096625	-0.3028192	-0.08386341	-0.1270698	0.07534275
kurtosis	-0.2399384	-0.2332168	-0.15807008	-0.3130280	-0.26250809

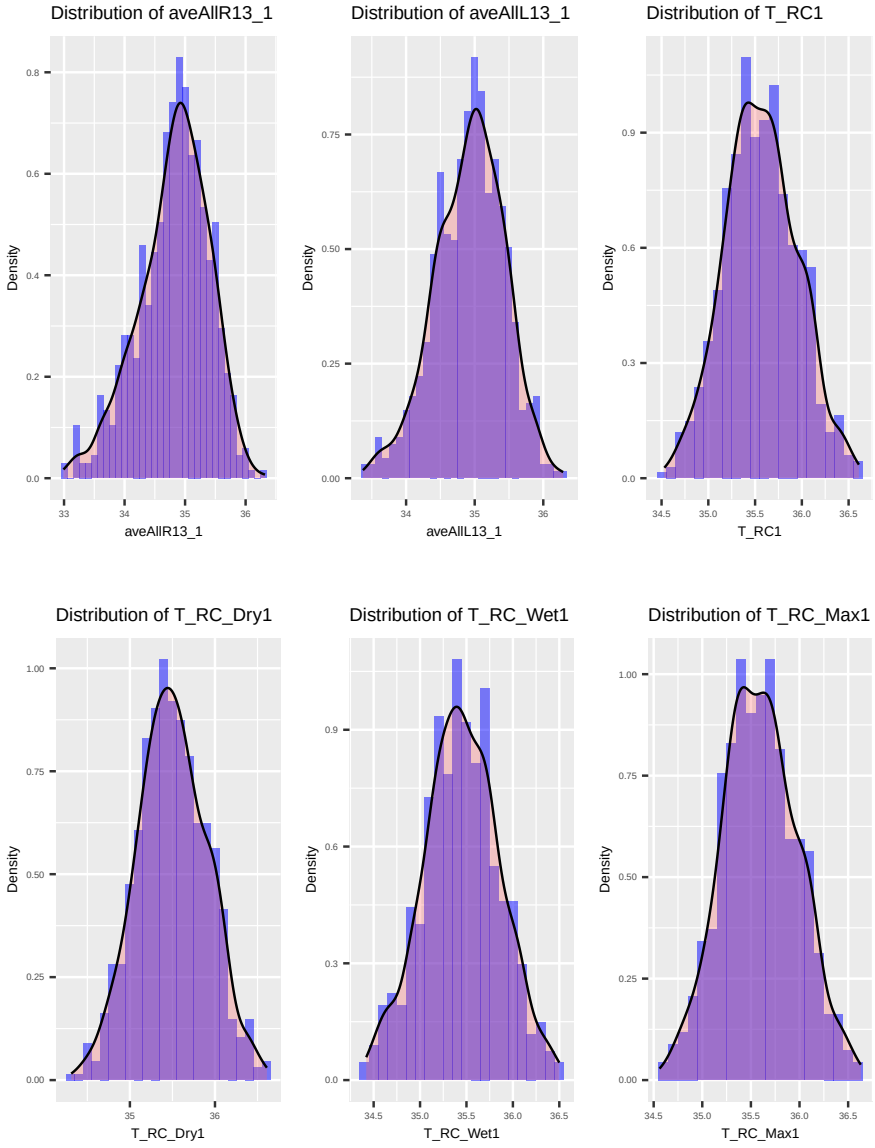
	T_OR_Max1	T_atm	Humidity	Distance

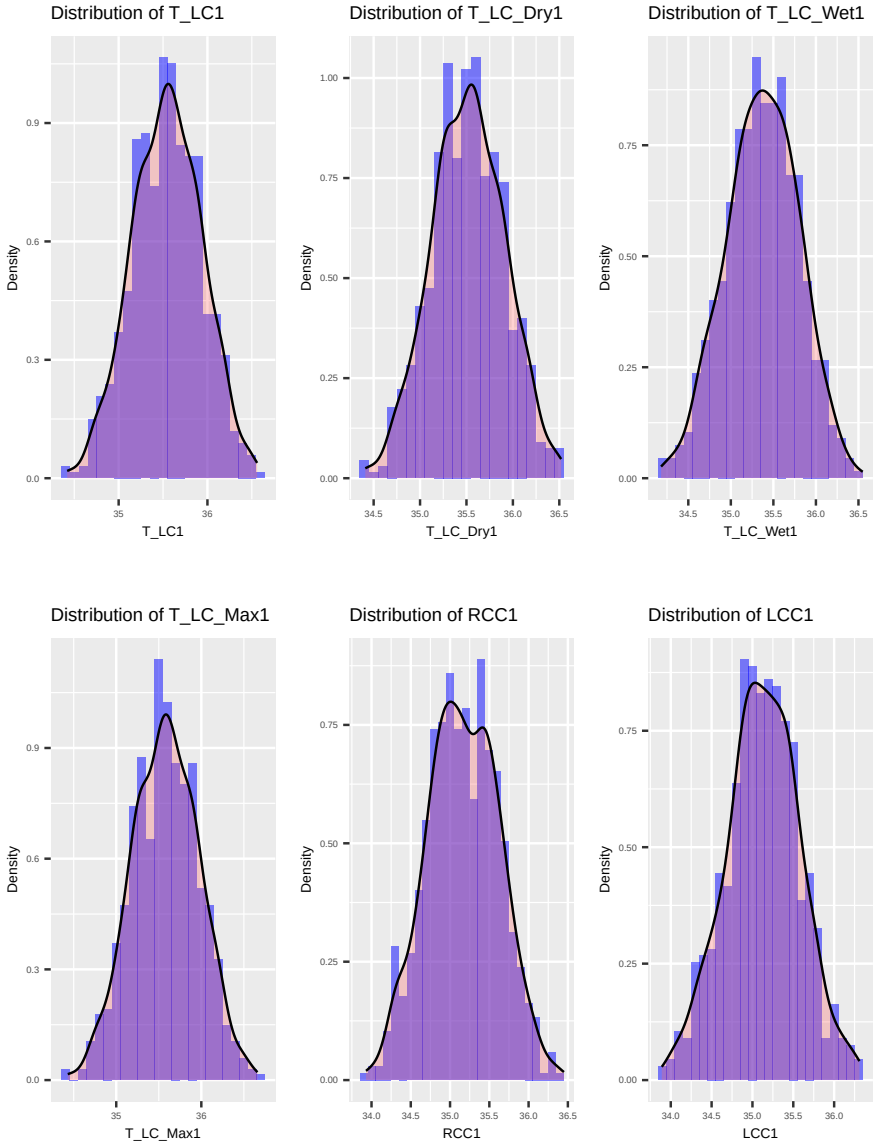
n	675.0000000	675.00000000	675.0000000	675.00000000
mean	35.3155852	24.03822222	27.8800000	0.65139259
stdev	0.5023698	0.96898021	12.6608324	0.06765345
skew	-0.1805682	0.10449759	0.7551421	0.85037793
kurtosis	-0.2809447	0.08506333	-0.4852465	-0.45633757

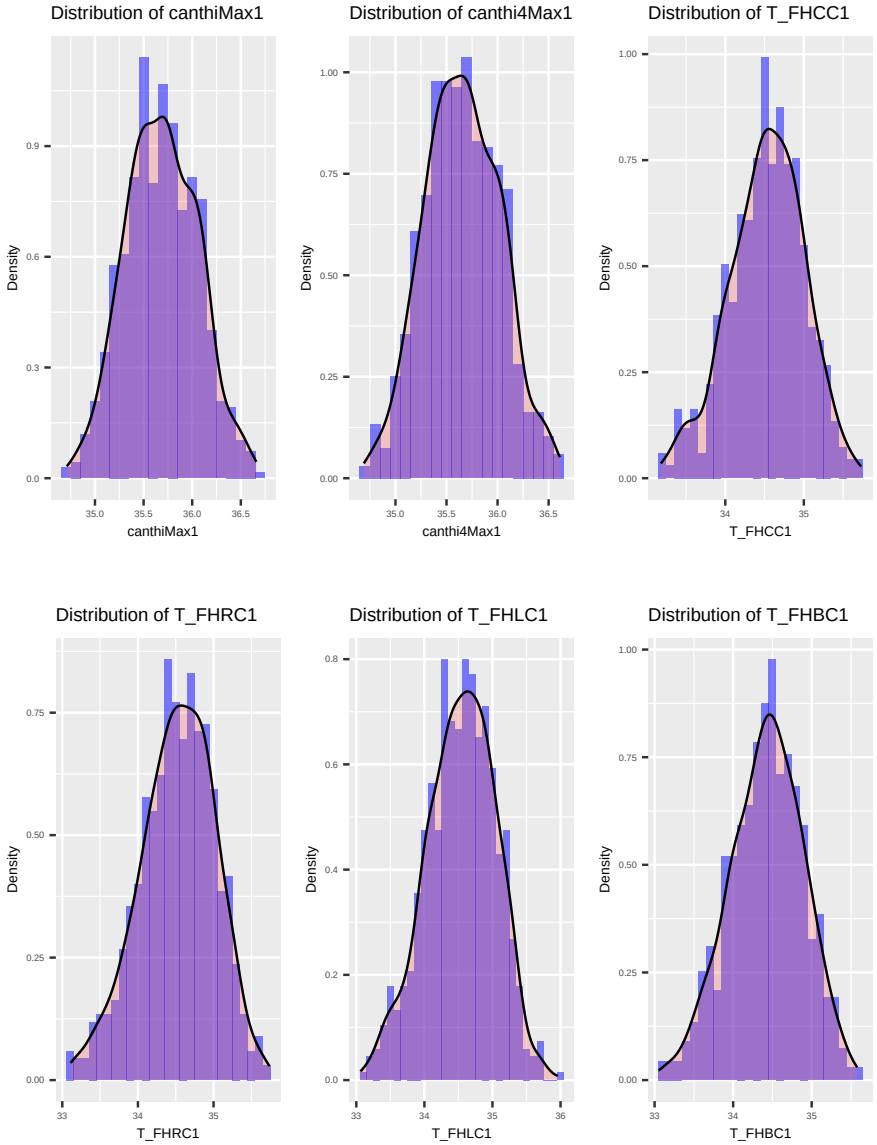
- **Plot Numerical Data w/o Outliers**

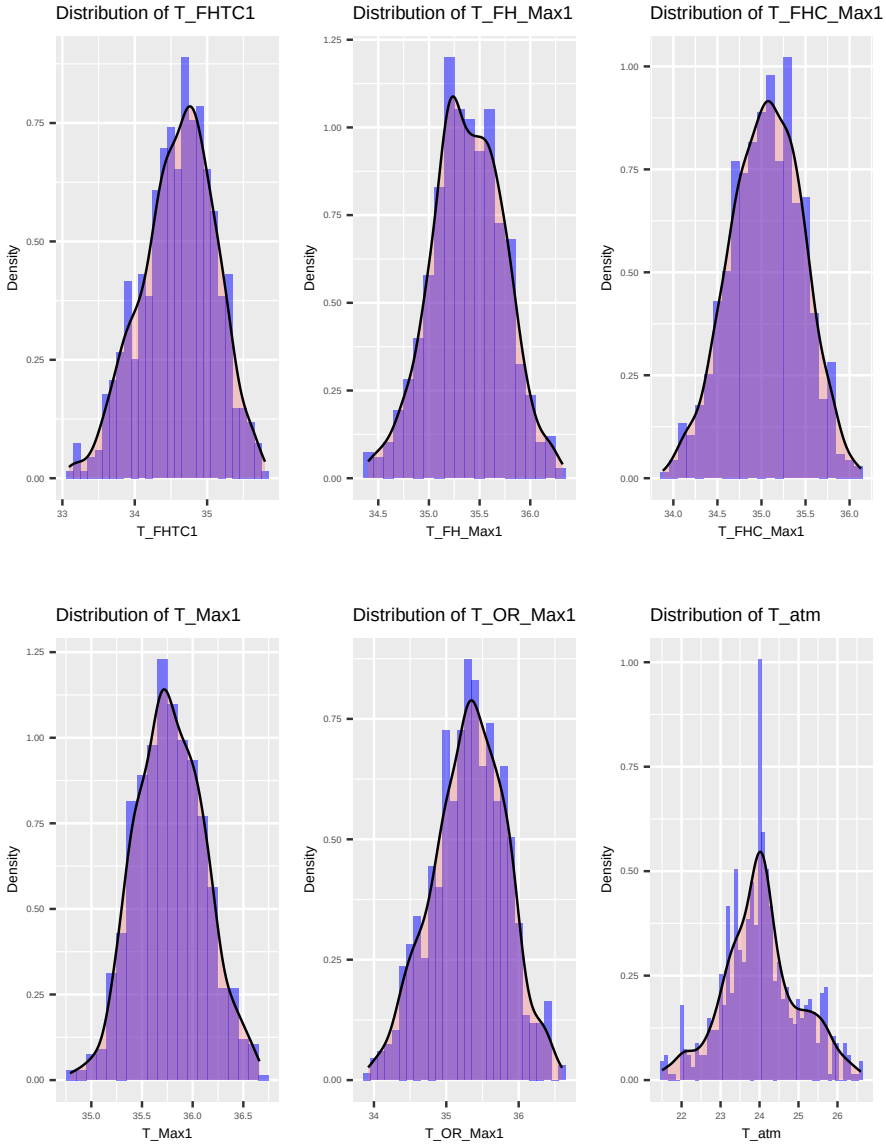
The plot of the numerical data confirms the findings of the descriptive statistics. The plots exhibit concentrated ranges.

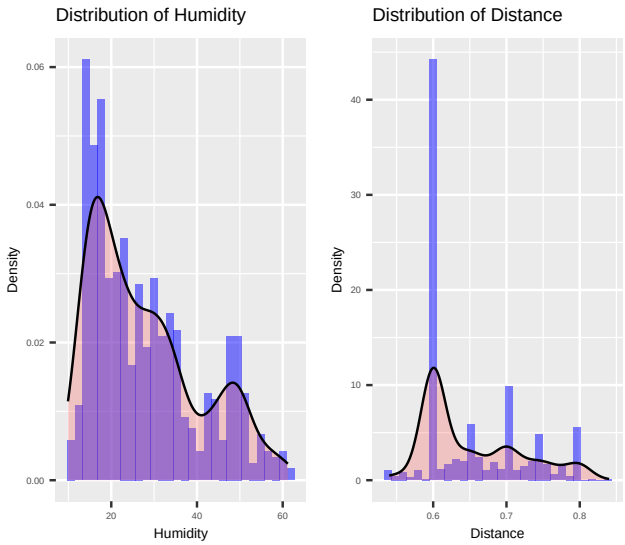



Distribution of aveAllR13_1
Density
aveAllR13_1
Distribution of aveAllL13_1
Density
aveAllL13_1
Distribution of T_RC1
Density
T_RC1
Distribution of T_RC_Dry1
Density
T_RC_Dry1
Distribution of T_RC_Wet1
Density
T_RC_Wet1
Distribution of T_RC_Max1
Density
T_RC_Max1


Distribution of T_LC1
Density
T_LC1
Distribution of T_LC_Dry1
Density
T_LC_Dry1
Distribution of T_LC_Wet1
Density
T_LC_Wet1
Distribution of T_LC_Max1
Density
T_LC_Max1
Distribution of RCC1
Density
RCC1
Distribution of LCC1
Density
LCC1


Distribution of canthiMax1
Density
canthiMax1
Distribution of canthi4Max1
Density
canthi4Max1
Distribution of T_FHCC1
Density
T_FHCC1
Distribution of T_FHRC1
Density
T_FHRC1
Distribution of T_FHLC1
Density
T_FHLC1
Distribution of T_FHBC1
Density
T_FHBC1

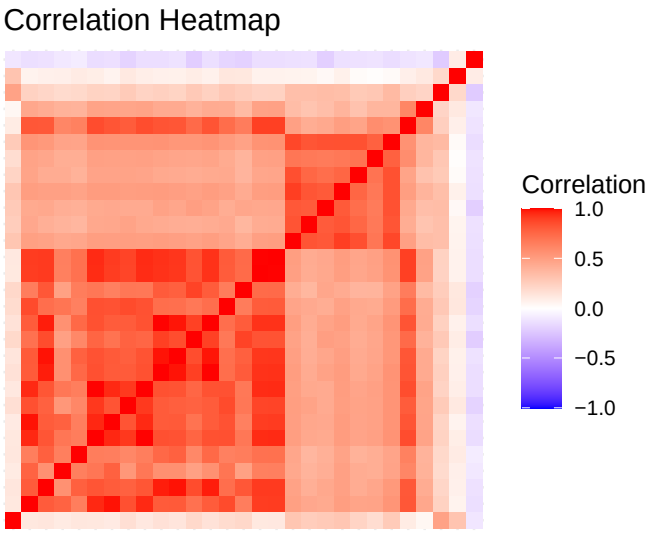

Distribution of T_FHTC1
Density
T_FHTC1
Distribution of T_FH_Max1
Density
T_FH_Max1
Distribution of T_FHC_Max1
Density
T_FHC_Max1
Distribution of T_Max1
Density
T_Max1
Distribution of T_OR_Max1
Density
T_OR_Max1
Distribution of T_atm
Density
T_atm



- **Correlation of Numerical Data**

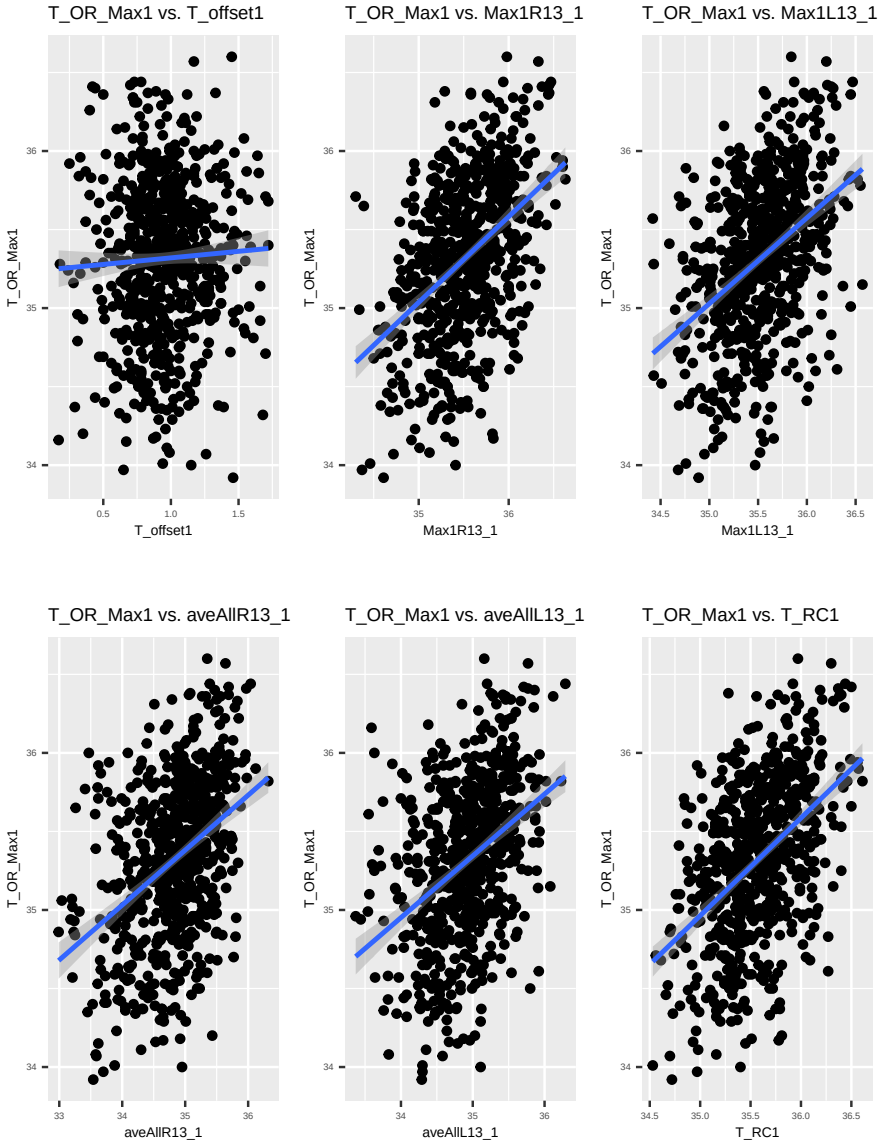
We also want to examine the correlations among the dependent variables. A correlation of 1 indicates the strongest positive relationship, -1 indicates the strongest negative relationship, and 0 indicates no relationship.

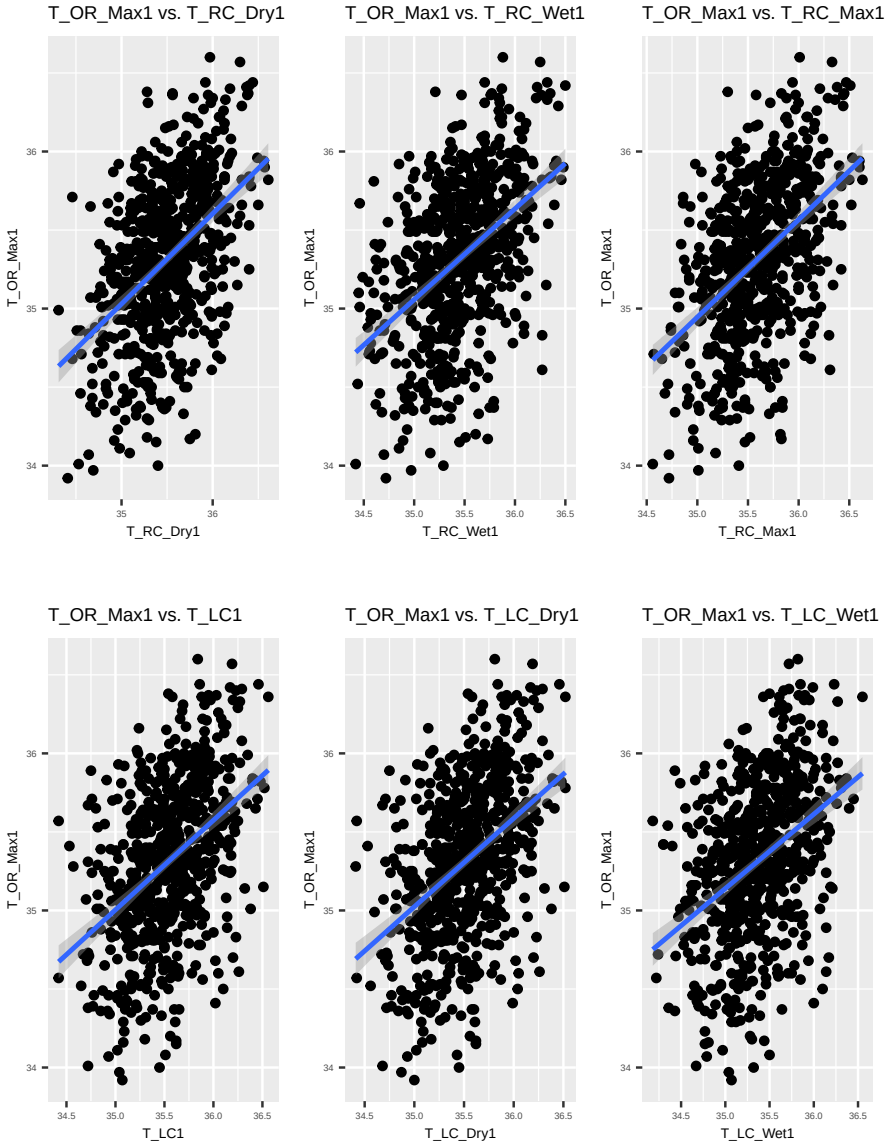
- **Correlation Matrix Heatmap**

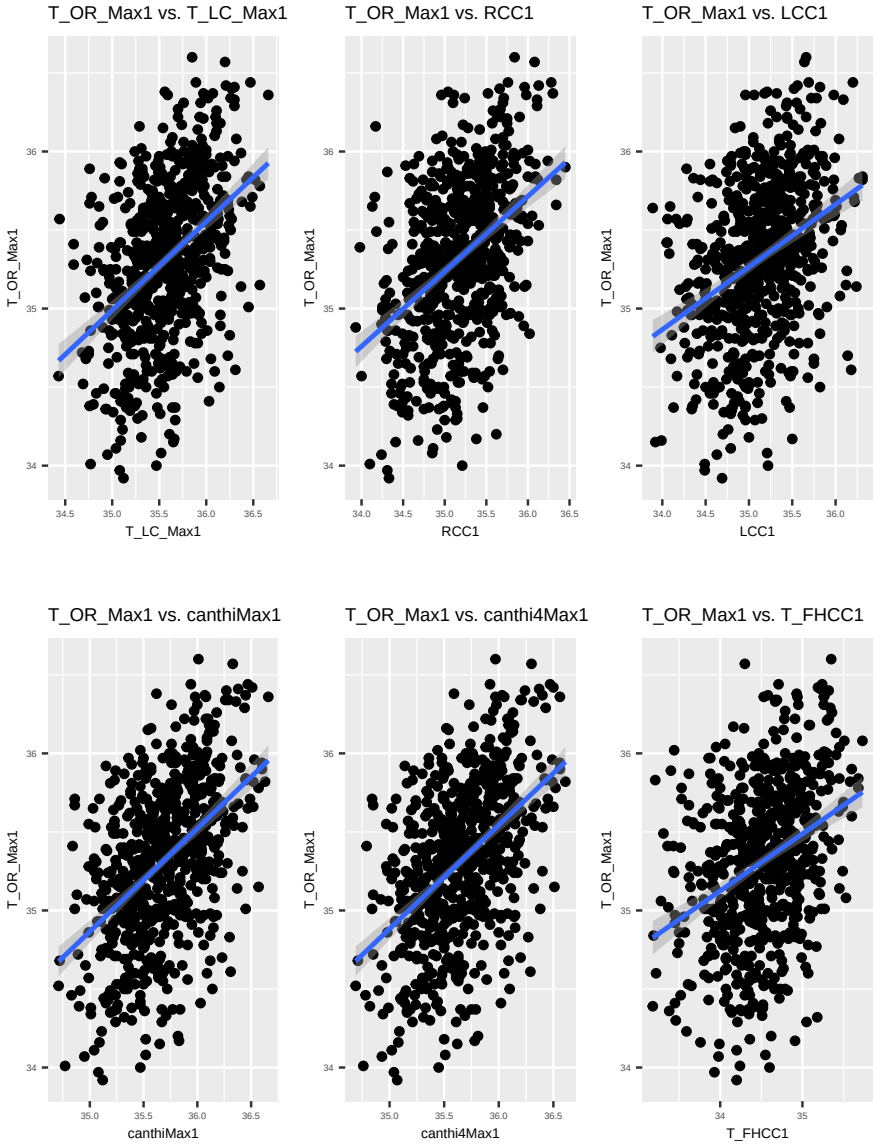


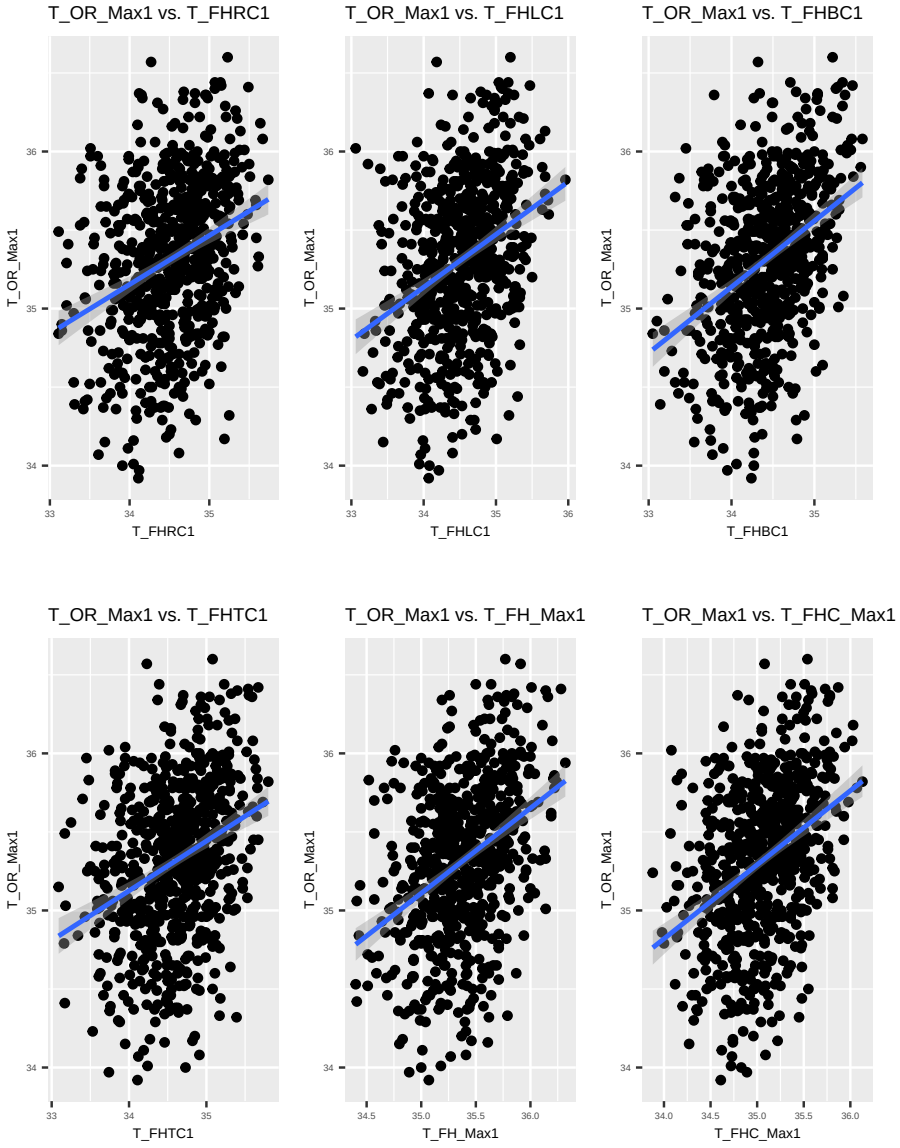
- **Oral Temperature VS. Dependent Variables**

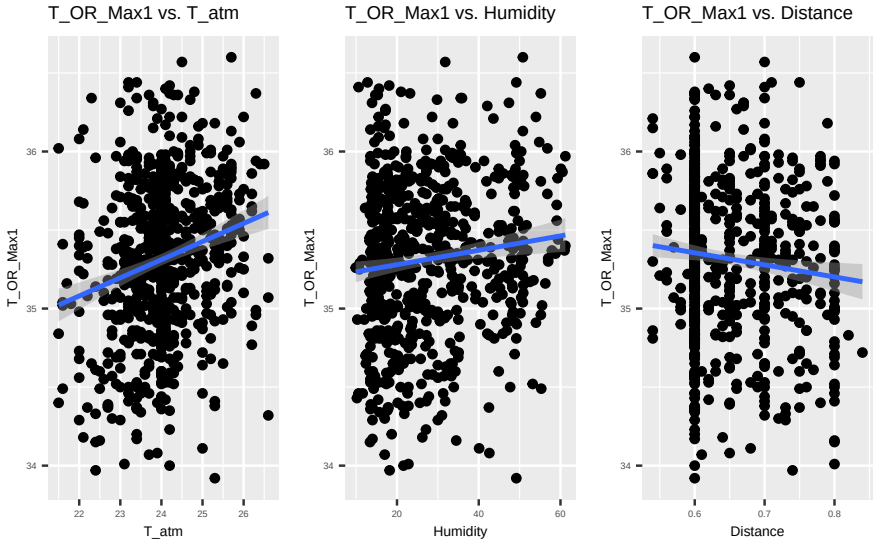
We want to examine the relationship between the target value and each individual dependent variable, without considering the other dependent variables.




T_OR_Max1 vs. T_RC_Dry1
T_OR_Max1 vs. T_RC_Wet1
T_OR_Max1 vs. T_RC_Max1
T_OR_Max1 vs. T_LC1
T_OR_Max1 vs. T_LC_Dry1
T_OR_Max1 vs. T_LC_Wet1
T_OR_Max1
T_RC_Dry1
T_RC_Wet1
T_RC_Max1
T_LC1
T_LC_Dry1
T_LC_Wet1


T_OR_Max1 vs. T_LC_Max1
T_OR_Max1 vs. RCC1
T_OR_Max1 vs. LCC1
T_OR_Max1 vs. canthiMax1
T_OR_Max1 vs. canthi4Max1
T_OR_Max1 vs. T_FHCC1
T_OR_Max1
T_LC_Max1
RCC1
LCC1
canthiMax1
canthi4Max1
T_FHCC1


T_OR_Max1 vs. T_FHRC1
T_OR_Max1 vs. T_FHLC1
T_OR_Max1 vs. T_FHBC1
T_OR_Max1 vs. T_FHTC1
T_OR_Max1 vs. T_FH_Max1
T_OR_Max1 vs. T_FHC_Max1
T_OR_Max1
T_FHRC1
T_FHLC1
T_FHBC1
T_FHTC1
T_FH_Max1
T_FHC_Max1



- **Standardization of Data**

The purpose of standardization is to ensure equal weightage of features. We use the z-score to standardize the data, giving it a mean of 0 and a standard deviation of 1.

$$StandardizedValue = \frac{x - \mu}{\sigma}$$

Modeling:

- **Split Data for Training and Testing by the Ration 80:20**

By setting the training/testing ratio to 80/20, we aim to ensure there is sufficient data to fit the model while retaining enough unseen data to test the created model.

- **Full Liner Regression Model**

We begin with a full linear model using all the available predictors.

1. Significant predictors:

a. T_offset1: Negative impact on T_OR_Max1 (p-value = 0.00572).
b. T_RC_Wet1: Positive impact (p-value = 0.01794).
c. T_FHBC1: Positive impact (p-value = 0.00447).
d. T_Max1: Highly significant positive impact (p-value < 2e-16).
e. T_atm: Positive impact (p-value = 0.04110).
f. Humidity: Positive impact (p-value = 0.04047).

2. Model Performance:

a. Multiple R-squared of 0.4829 indicates that approximately 48.29% of the variability in T_OR_Max1 is explained by the model.
b. Adjusted R-squared of 0.4545 accounts for the number of predictors in the model, providing a more conservative estimate.
c. Since the value of R-squared is less than 0.5, we want to improve our model performance.

```
Call:
lm(formula = T_OR_Max1 ~ ., data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4305 -0.4729  0.1262  0.5641  1.3968

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0403524  0.0426431  -0.946  0.34444
T_offset1   -0.1103248  0.0397636  -2.775  0.00572 **
Max1R13_1   -0.2336011  0.2270822  -1.029  0.30409
Max1L13_1    0.0599152  0.2300158   0.260  0.79459
aveAllR13_1  0.1027192  0.0672097   1.528  0.12703
aveAllL13_1  0.0604321  0.0642277   0.941  0.34719
T_RC1        0.7608109  0.7872684   0.966  0.33429
T_RC_Dry1    0.2819361  0.2354084   1.198  0.23159
T_RC_Wet1    0.2801456  0.1179915   2.374  0.01794 *
T_RC_Max1   -0.8971967  0.7390849  -1.214  0.22532
T_LC1        0.7005650  0.7427485   0.943  0.34601
T_LC_Dry1    0.0003762  0.3089001   0.001  0.99903
T_LC_Wet1   -0.0180984  0.1121093  -0.161  0.87181
T_LC_Max1   -0.8506549  0.6715548  -1.267  0.20582
RCC1        -0.0981621  0.0867169  -1.132  0.25816
LCC1        -0.0438804  0.0826842  -0.531  0.59585
canthiMax1   0.8335121  0.7723943   1.079  0.28102
canthi4Max1 -1.2996385  0.7928805  -1.639  0.10178
T_FHCC1     -0.1186885  0.0989227  -1.200  0.23075
T_FHRC1     -0.0726441  0.0680752  -1.067  0.28641
T_FHLC1      0.0980432  0.0700637   1.399  0.16230
T_FHBC1      0.2483280  0.0869825   2.855  0.00447 **
T_FHTC1      0.0050607  0.0746947   0.068  0.94601
T_FH_Max1   -0.1127790  0.0590556  -1.910  0.05671 .
T_FHC_Max1   0.0251103  0.0920137   0.273  0.78504
T_Max1       0.9546431  0.0806797  11.833  < 2e-16 ***
T_atm        0.0857441  0.0418758   2.048  0.04110 *
Humidity     0.0741832  0.0361173   2.054  0.04047 *
```

```
Distance    -0.0241198  0.0357459  -0.675  0.50013
Gender_M     0.0580021  0.0727974   0.797  0.42595
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7652 on 527 degrees of freedom
Multiple R-squared:  0.4829,    Adjusted R-squared:  0.4545
F-statistic: 16.97 on 29 and 527 DF,  p-value: < 2.2e-16
```

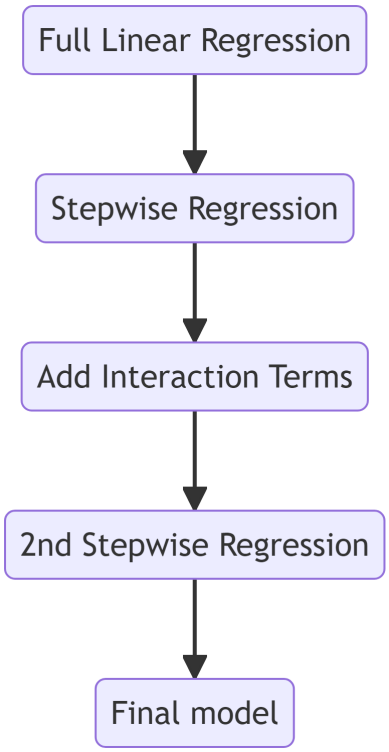
- **Full Regression and Stepwised Regression Model**

1. To build a comprehensive regression model, we add interaction terms (for two predictors only) to the full linear model. This increases the R-squared value, bringing it closer to 1. However, it also significantly increases the model's complexity. In our case, with 435 factors (29 dependent variables and 406 interaction terms) involved, this suggests potential overfitting, where the model fits the training data very well but may not generalize well to new, unseen data.

2.A stepwise method (backward) provides a way to remove predictors from the full regression model, reducing complexity. However, the resulting model formula remains overly complicated.

- **Improved Regression Model**

1. We begin with the full linear regression model.
2. Using the stepwise method, we removed predictors from the full linear regression model. The predictors retained in the model are: T_offset1, aveAllR13_1, aveAllL13_1, T_RC_Wet1, T_LC_Max1, RCC1, canthi4Max1, T_FHCC1, T_FHLC1, T_FHBC1, T_FH_Max1, T_Max1, T_atm, Humidity, and Distance.
3. Add interaction terms for the retained predictors.
4. Run the stepwise method again to determine the final model.



- **Summary of the Improved Regression Model**

1. The **predictors** kept are: T_offset1 ,Max1R13_1, aveAllR13_1, T_RC_Dry1, T_RC_Wet1, canthi4Max1, T_FHBC1, T_FH_Max1, T_Max1, and T_atm.
2. The **interaction terms** kept are: T_offset1:aveAllR13_1, T_offset1:T_FHBC1, T_offset1:T_atm, aveAllR13_1:T_FHBC1, aveAllR13_1:Humidity, T_RC_Dry1:canthi4Max1, T_RC_Dry1:T_FHBC1, T_RC_Dry1:T_Max1, T_RC_Dry1:Humidity, T_RC_Wet1:canthi4Max1, T_RC_Wet1:T_Max1, T_RC_Wet1:T_atm, T_RC_Wet1:Humidity, canthi4Max1:T_FH_Max1, T_FHBC1:Humidity, T_FH_Max1:T_Max1, T_FH_Max1:T_atm, and T_Max1:T_atm.
3. The Multiple R-squared of 0.5667 indicates that approximately 56.67% of the variability in T_OR_Max1 is explained by the model.
4. The F-statistic and associated p-value ($< 2.2e\text{-}16$) indicate that the overall model is statistically significant.
5. Variance Inflation Factor (VIF) values indicate the degree of multicollinearity in the model. High VIF values in this model, such as T_RC_Dry1 of 41.72 and Max1R13_1 of 41.99, suggest that predictors has a high

correlation with other predictors.

6. To address multicollinearity effectively in the regression models., we will try applying PCA, or by using Ridge regression.

```
Call:
lm(formula = T_OR_Max1 ~ T_offset1 + Max1R13_1 + aveAllR13_1 +
    T_RC_Dry1 + T_RC_Wet1 + canthi4Max1 + T_FHBC1 + T_FH_Max1 +
    T_Max1 + T_atm + Humidity + T_offset1:aveAllR13_1 + T_offset1:T_FHBC1 +
    T_offset1:T_atm + aveAllR13_1:T_FHBC1 + aveAllR13_1:Humidity +
    T_RC_Dry1:canthi4Max1 + T_RC_Dry1:T_FHBC1 + T_RC_Dry1:T_Max1 +
    T_RC_Dry1:Humidity + T_RC_Wet1:canthi4Max1 + T_RC_Wet1:T_Max1 +
    T_RC_Wet1:T_atm + T_RC_Wet1:Humidity + canthi4Max1:T_FH_Max1 +
    T_FHBC1:Humidity + T_FH_Max1:T_Max1 + T_FH_Max1:T_atm + T_Max1:T_atm,
    data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1722 -0.4166  0.1103  0.4980  1.4805

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.16842    0.04254   3.959 8.57e-05 ***
T_offset1             -0.09034    0.03696  -2.445  0.01483 *
Max1R13_1             -0.37278    0.19174  -1.944  0.05240 .
aveAllR13_1            0.07544    0.05265   1.433  0.15249
T_RC_Dry1              0.36768    0.19162   1.919  0.05555 .
T_RC_Wet1              0.12638    0.07013   1.802  0.07209 .
canthi4Max1           -0.97059    0.11841  -8.197 1.88e-15 ***
T_FHBC1                0.14525    0.04800   3.026  0.00260 **
T_FH_Max1             -0.08115    0.04659  -1.742  0.08214 .
T_Max1                 1.37546    0.08940  15.385  < 2e-16 ***
T_atm                  0.07611    0.03726   2.043  0.04159 *
Humidity               0.10073    0.03351   3.006  0.00277 **
T_offset1:aveAllR13_1  0.08233    0.04110   2.003  0.04566 *
T_offset1:T_FHBC1     -0.05869    0.03934  -1.492  0.13631
T_offset1:T_atm       -0.07065    0.03049  -2.317  0.02088 *
aveAllR13_1:T_FHBC1   -0.10754    0.05236  -2.054  0.04051 *
aveAllR13_1:Humidity  -0.10652    0.05750  -1.853  0.06451 .
T_RC_Dry1:canthi4Max1 -0.41477    0.14505  -2.859  0.00441 **
T_RC_Dry1:T_FHBC1      0.10885    0.05778   1.884  0.06013 .
T_RC_Dry1:T_Max1       0.29565    0.13922   2.124  0.03417 *
T_RC_Dry1:Humidity     0.18327    0.07477   2.451  0.01456 *
T_RC_Wet1:canthi4Max1 -0.21746    0.13200  -1.647  0.10006
T_RC_Wet1:T_Max1       0.19441    0.13013   1.494  0.13578
T_RC_Wet1:T_atm        0.17655    0.05490   3.216  0.00138 **
```

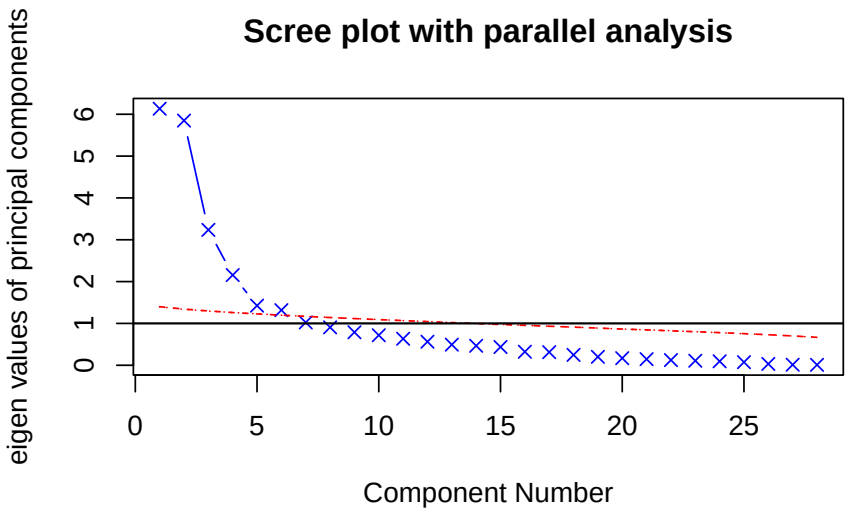
```
T_RC_Wet1:Humidity    -0.17378    0.05891  -2.950  0.00332 **
canthi4Max1:T_FH_Max1  0.42607    0.07210   5.910 6.16e-09 ***
T_FHBC1:Humidity       0.08976    0.03981   2.255  0.02455 *
T_FH_Max1:T_Max1      -0.27661    0.06898  -4.010 6.96e-05 ***
T_FH_Max1:T_atm       -0.06392    0.03826  -1.671  0.09535 .
T_Max1:T_atm          -0.14386    0.05969  -2.410  0.01630 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7005 on 527 degrees of freedom
Multiple R-squared:  0.5667,    Adjusted R-squared:  0.5428
F-statistic: 23.76 on 29 and 527 DF,  p-value: < 2.2e-16
```

```
            T_offset1             Max1R13_1           aveAllR13_1
             1.588778             41.989761              3.036011
            T_RC_Dry1             T_RC_Wet1           canthi4Max1
            41.719901              5.685309             15.777151
              T_FHBC1             T_FH_Max1                T_Max1
             2.727546              2.508126              9.123867
                 T_atm              Humidity T_offset1:aveAllR13_1
             1.542628              1.262523              1.707703
   T_offset1:T_FHBC1       T_offset1:T_atm   aveAllR13_1:T_FHBC1
             2.120347              1.383515              3.426110
aveAllR13_1:Humidity T_RC_Dry1:canthi4Max1     T_RC_Dry1:T_FHBC1
             3.389069             35.459508              4.699374
    T_RC_Dry1:T_Max1    T_RC_Dry1:Humidity T_RC_Wet1:canthi4Max1
            33.710904              6.067767             29.592164
    T_RC_Wet1:T_Max1       T_RC_Wet1:T_atm    T_RC_Wet1:Humidity
            27.806219              3.514764              3.653173
canthi4Max1:T_FH_Max1      T_FHBC1:Humidity      T_FH_Max1:T_Max1
             6.890518              1.840454              6.873623
     T_FH_Max1:T_atm          T_Max1:T_atm
             1.864243              3.790260
```

- **Address high VIF for the Model with PCA**

Principal Component Analysis (PCA) transforms the predictors into a set of uncorrelated components, reducing the number of variables without losing much information. Using too many predictors can lead to overfitting and complicate the interpretation of the analysis.

1. As shown in the scree plot, the first six components have eigenvalues greater than 1. The curve begins to flatten at the 7th component (elbow), indicating that the first six components should be retained.

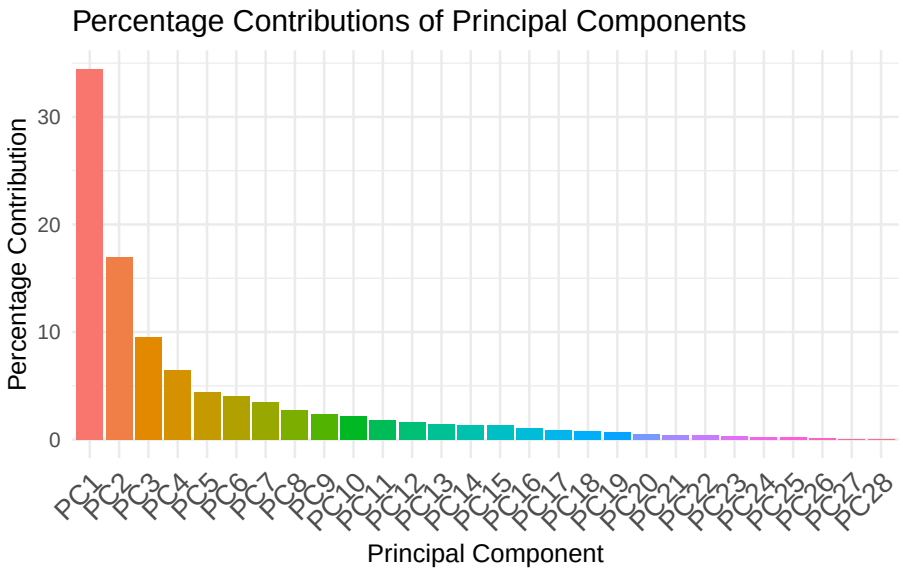


2.The first six components contribute to more than 75% of the total variance.

```
[1] "Accumulated contribution up to component 1 :  34.47 %"
[1] "Accumulated contribution up to component 2 :  51.43 %"
[1] "Accumulated contribution up to component 3 :  60.94 %"
[1] "Accumulated contribution up to component 4 :  67.42 %"
[1] "Accumulated contribution up to component 5 :  71.8 %"
[1] "Accumulated contribution up to component 6 :  75.81 %"
[1] "Accumulated contribution up to component 7 :  79.34 %"
[1] "Accumulated contribution up to component 8 :  82.1 %"
[1] "Accumulated contribution up to component 9 :  84.46 %"
[1] "Accumulated contribution up to component 10 :  86.66 %"
[1] "Accumulated contribution up to component 11 :  88.48 %"
[1] "Accumulated contribution up to component 12 :  90.09 %"
[1] "Accumulated contribution up to component 13 :  91.49 %"
[1] "Accumulated contribution up to component 14 :  92.85 %"
[1] "Accumulated contribution up to component 15 :  94.15 %"
[1] "Accumulated contribution up to component 16 :  95.18 %"
[1] "Accumulated contribution up to component 17 :  96.09 %"
[1] "Accumulated contribution up to component 18 :  96.91 %"
[1] "Accumulated contribution up to component 19 :  97.64 %"
[1] "Accumulated contribution up to component 20 :  98.14 %"
[1] "Accumulated contribution up to component 21 :  98.58 %"
[1] "Accumulated contribution up to component 22 :  98.97 %"
[1] "Accumulated contribution up to component 23 :  99.3 %"
[1] "Accumulated contribution up to component 24 :  99.57 %"
[1] "Accumulated contribution up to component 25 :  99.82 %"
```

```
[1] "Accumulated contribution up to component 26 :  99.92 %"
[1] "Accumulated contribution up to component 27 :  99.97 %"
[1] "Accumulated contribution up to component 28 :  100 %"
```

3. The contributions of the first six components are also shown in the plot.



4. The linear model with six components shows great VIF values, close to 1 (all less than 5). However, the R-squared value drops to 0.321, meaning only 32.1% of the variations are explained. Adding up to 20 components increases the R-squared value to over 0.5, but this makes the model too complex to use.

```
Call:
lm(formula = T_OR_Max1 ~ ., data = train_data_pca)

Residuals:
     Min       1Q   Median       3Q      Max
-2.56805 -0.53953  0.04293  0.62484  1.89648

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.010084   0.046457   0.217   0.8282
PC1         -0.007296   0.013137  -0.555   0.5788
PC2          0.235647   0.014889  15.827   <2e-16 ***
PC3         -0.039088   0.019824  -1.972   0.0491 *
PC4         -0.018475   0.024275  -0.761   0.4470
PC5          0.055235   0.029022   1.903   0.0575 .
```

```
PC6          -0.030685   0.030082  -1.020   0.3082
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8583 on 550 degrees of freedom
Multiple R-squared:  0.321, Adjusted R-squared:  0.3136
F-statistic: 43.34 on 6 and 550 DF,  p-value: < 2.2e-16
```

```
     PC1      PC2      PC3      PC4      PC5      PC6
1.016655 1.002594 1.000763 1.006119 1.011774 1.001972
```

- **Address high VIF with Ridge Regression**

Ridge Regression adds a penalty equal to the square of the magnitude of the coefficients, helping to reduce the impact of multicollinearity.

```
Ridge Regression Equation:

T_OR_Max1 = 0.08592975 +

(-0.0947 * T_offset1) +
(-0.0980 * Max1R13_1) +
(0.0915 * aveAllR13_1) +
(0.0285 * T_RC_Dry1) +
(0.0537 * T_RC_Wet1) +
(-0.2444 * canthi4Max1) +
(0.1146 * T_FHBC1) +
(-0.0107 * T_FH_Max1) +
(0.7387 * T_Max1) +
(0.0816 * T_atm) +
(0.0954 * Humidity) +
(0.0633 * T_offset1:aveAllR13_1) +
(-0.0413 * T_offset1:T_FHBC1) +
(-0.0616 * T_offset1:T_atm) +
(-0.0877 * aveAllR13_1:T_FHBC1) +
(-0.0216 * aveAllR13_1:Humidity) +
(-0.1070 * T_RC_Dry1:canthi4Max1) +
(0.0682 * T_RC_Dry1:T_FHBC1) +
(0.0109 * T_RC_Dry1:T_Max1) +
(0.0773 * T_RC_Dry1:Humidity) +
(-0.0570 * T_RC_Wet1:canthi4Max1) +
(0.0374 * T_RC_Wet1:T_Max1) +
(0.1129 * T_RC_Wet1:T_atm) +
(-0.0950 * T_RC_Wet1:Humidity) +
(0.2323 * canthi4Max1:T_FH_Max1) +
(0.0393 * T_FHBC1:Humidity) +
(-0.0784 * T_FH_Max1:T_Max1) +
(-0.0311 * T_FH_Max1:T_atm) +
(-0.0984 * T_Max1:T_atm)
```

Evaluate the model

1. Root Mean Squared Error (RMSE) is a common metric used to evaluate the performance of regression models. It measures the average magnitude of the errors between the predicted and actual values. The formula for RMSE is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y_i} - y_i)^2}$$

2. How you can calculate RMSE for training and testing data:

a. Train your regression model using the training data.
b. Make predictions on both the training and testing datasets.
c. Calculate RMSE using the actual and predicted values for both datasets.

3. We also calculate R Squared to see how many percentage of the variability in T_OR_Max1 is explained by the model. The formular for R Squared is:

$$R^2 = 1 - \frac{RSS}{TSS}$$

4. Model Accuracy: RMSE values close to zero indicate a better fit. In this case, the RMSE values are relatively low, suggesting that the model's predictions are close to the actual values.

5. Model Performance: The RMSE on the training data (0.7212) is very close to the RMSE on the testing data (0.7401). This indicates consistent performance across both datasets, suggesting that the model is well-generalized and not overfitting. R Squared value shoes approximately 51.4% of the variability in T_OR_Max1 is explained by the model.

```
Caculated value based on Ridge model:

R-squared : 0.5144992

RMSE on Training Data: 0.7211868

RMSE on Testing Data: 0.7401346
```

Review the model

1. Model Structure: The model predicts the variable T_OR_Max1 based on multiple predictors and their interactions. It is a ridge regression model, used to address multicollinearity among the predictors by adding a penalty to the size of the coefficients. This results in a more stable and generalizable model.

2. Coefficient: The coefficients have been shrunk towards zero to prevent overfitting.
3. Interaction Terms: Interaction terms are included to capture the combined effects of predictors. Regularization helps manage the complexity introduced by these terms, ensuring the model remains interpretable and generalizable.
4. Regularization Term: The 'glmnet' package in R is used to generate the ridge regression model. Cross-validation helps to find the optimal value of the regularization parameter 'lambda,' which controls the trade-off between fitting the training data and keeping the coefficients small.

Reference

Wang, Q., Zhou, Y., Ghassemi, P., Chenna, D., Chen, M., Casamento, J., Pfefer, T., & McBride, D. (2023). *Facial and oral temperature data from a large set of human subject volunteers (version 1.0.0)*. PhysioNet. https://doi.org/10.13026/3bhc-9065