STA 141A Final Project

Donggyun Ha, 914319683
Chan Woong Joo, 913053322
Sungwon Lee, 912978026
Jong Chan Park, 913274897

Contents

We splitted the amount of work by analyzing each question in Numeric Data Analysis.
For every other part of the project, All of us worked together without excluding anyone.

Data Description

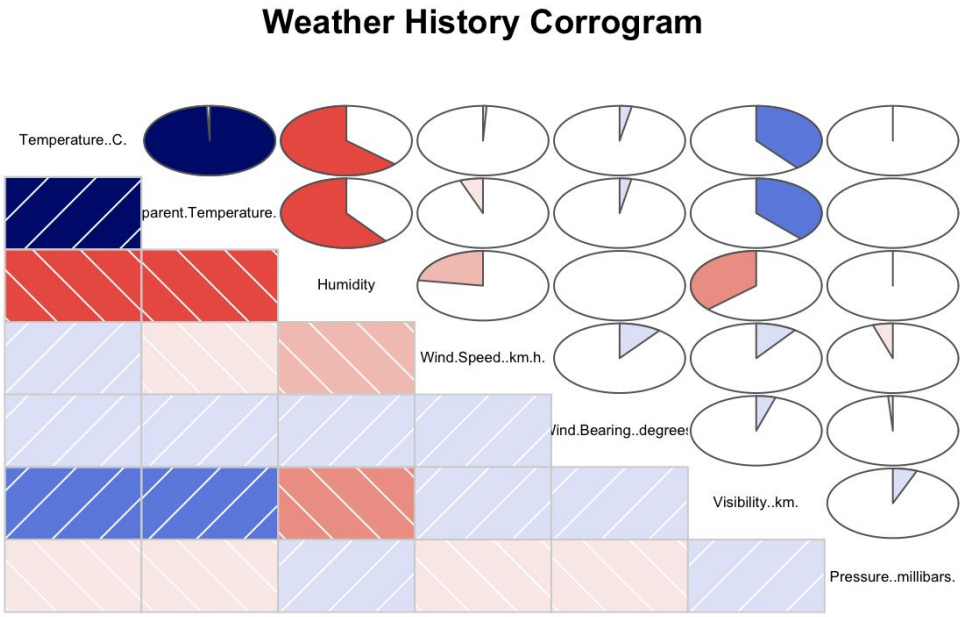
The data we retrieved was taken from the website: https://www.kaggle.com/budincsevity/szeged-weather. It is a weather record of Szeged, Hungary from the year 2006 to 2016. It contains both numerical and categorical data that are going to be analyzed throughout the project. The data are collected as a dataframe weather.History.csv, which contains the following columns:

1.**Formatted.date** : Time of data in this row
2.**Summary** : Data
3.**Precip.type :** 3 types of weather (Sunny=null, Rainy, Snowy days)
4.**Temperature** : Temperature (°C)
5.**Apparant Temperature** : The perceived temperature (°C)
6.**Humidity** : Humidity (g/m3)
7.**Wind.Speed**: Wind speed (km/hr)
8.**Wind.Bearing.degrees** : Wind direction
9. **Visibility** : Visibility (km)

10.**Pressure.millibars** : Pressure (Hpa)
11.**Daily.summary** : Weather condition in the morning, afternoon and evening.

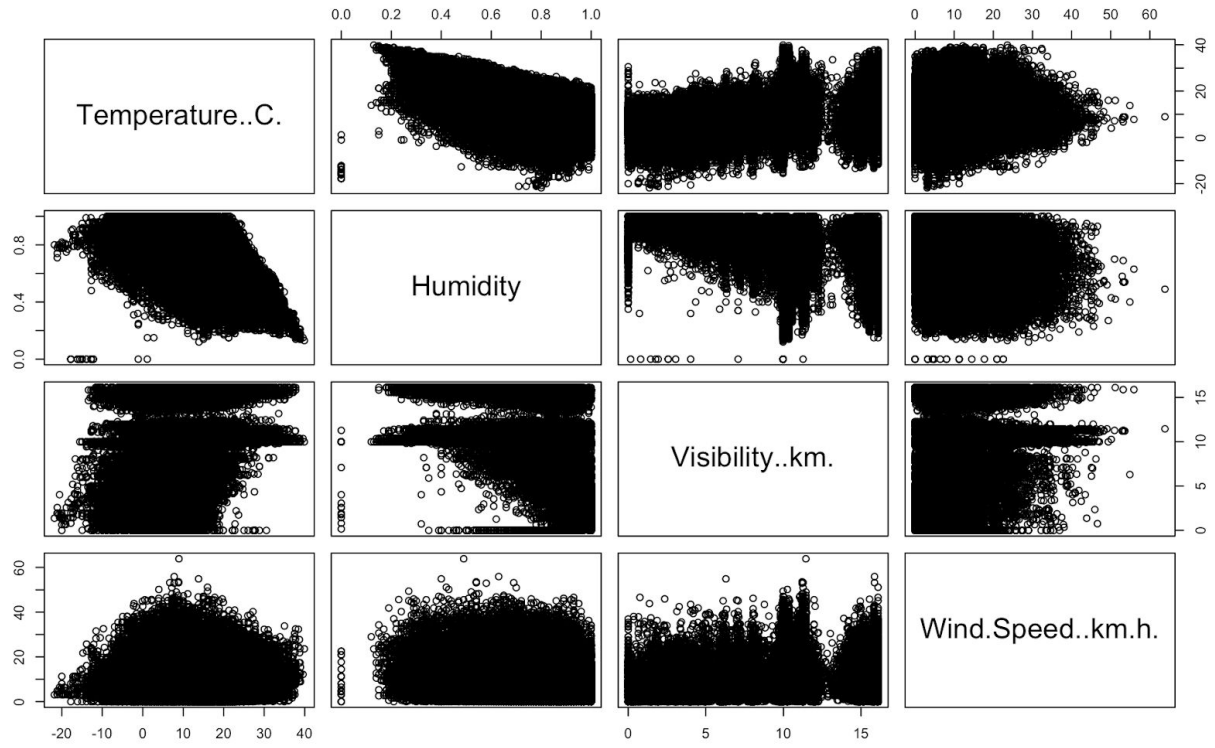
Analysis

Numerical variable Analysis



Figure[1]: Corrogram of Weather History

When analyzing numerical data, correlogram provides a quick visual representation to begin with. The plot gives an overview of the relationship between each variable and its relationship. Temperature and apparent temperature have the strongest positive correlation compared to other variables, and there is also a positive correlation between temperature and visibility. Furthermore, there was a negative correlation between temperature and humidity, and humidity and visibility. We are able to get an idea that some variables are more interesting than others. For example, we were interested in the relationship between **temperature and humidity**, **humidity and visibility**, **temperature and wind speed**, and **temperature and visibility**.

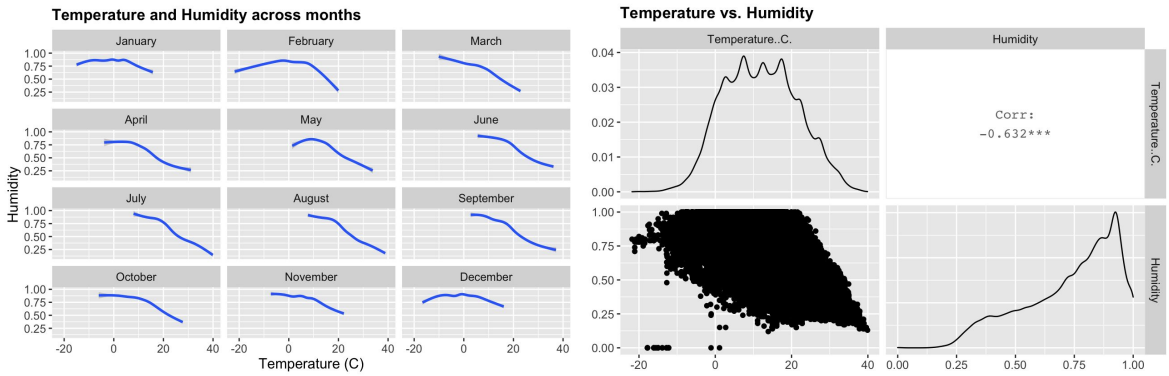


Figure[2]: Temperature vs. Humidity vs. Visibility vs. Wind Speed

The pair graphs provide a visual representation of how data points are distributed. As the correlogram explained, there is a correlation between each variable, and that we may be able to build a predictive model.

Numeric Data Analysis

Question 1. What is the relationship between **Temperature** and **Humidity** in Szeged? (Monthly changes too)



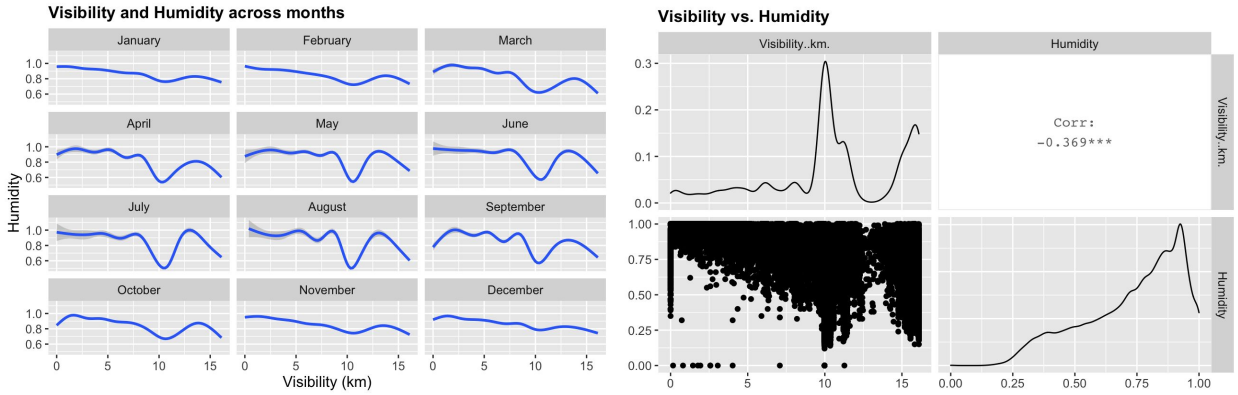
Figure[3]: ggplot and ggpairs of Humidity vs Temperature(in month for ggplot)

First, we want to observe the relationship between Temperature and Humidity across the month factor. From the histogram above, we can see that the humidity is skewed to the left. Also, we can see that the correlation between the Temperature and Humidity is -0.632 which is fairly high. Also, since the correlation ratio is negative, we can see that the temperature and humidity are in a negative relationship. From the graph on the bottom left, When the temperature goes up, the humidity goes down. It shows the negative relationship throughout the month.

However, at certain points from -20 to 0 degrees celsius, the humidity goes up when the temperature goes up. Also, from the graph on the right, we can observe that the humidity increases up to a certain point, and then drops as the temperature increases. In the graph, the humidity curve starts to descend dramatically at approximately 10°C. This is the reason why the histogram of the humidity is skewed to the left. Furthermore, we can clearly see that the curve descends gradually in the winter season compared to April to October.

Therefore, We can say that there is a negative linear relationship between the temperature and the humidity.

Question 2. What is the relationship between **Visibility** and **Humidity** in Szeged? (Monthly changes too)

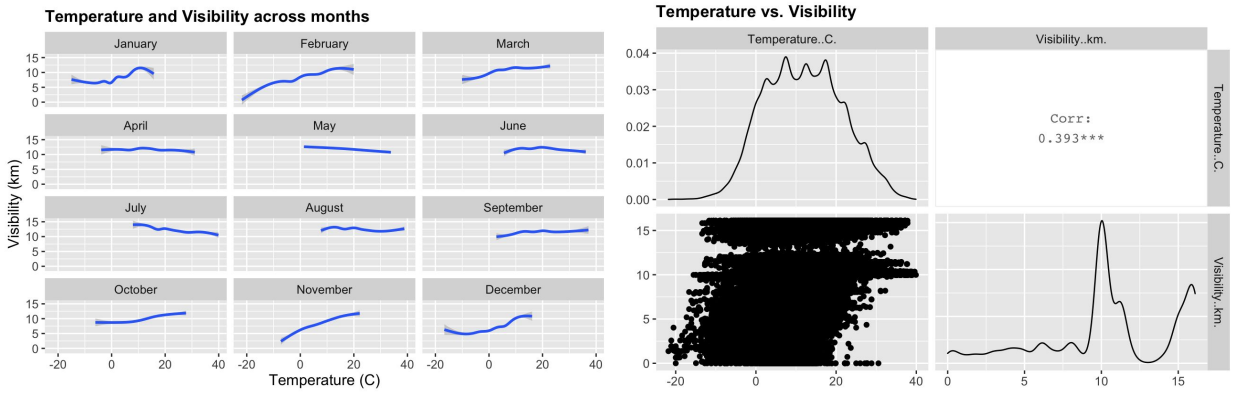


Figure[4]: ggplot and ggpairs of Visibility vs Humidity(in month for ggplot)

We can see from the plots above that there is not much difference between the plot of humidity and visibility of each month. Also, the correlation ratio is -0.369 which is fairly low and it has a negative relationship. What's noticeable is that the humidity significantly drops from May to September around 10km of visibility. This tells us that during the summer time in szeged, the visibility drops significantly due to other factors such as (comparatively) high temperature or change of the weather conditions that happens during the hot summer months.

Also, We can notice that winter months (November, December, January, February) have lower values of visibility compared to other months. However it is a very small difference that we might not notice. Although there is not much change when humidity is low, as humidity increases, visibility decreases exponentially. In conclusion, we can assume that there is a weak correlation between humidity and visibility.

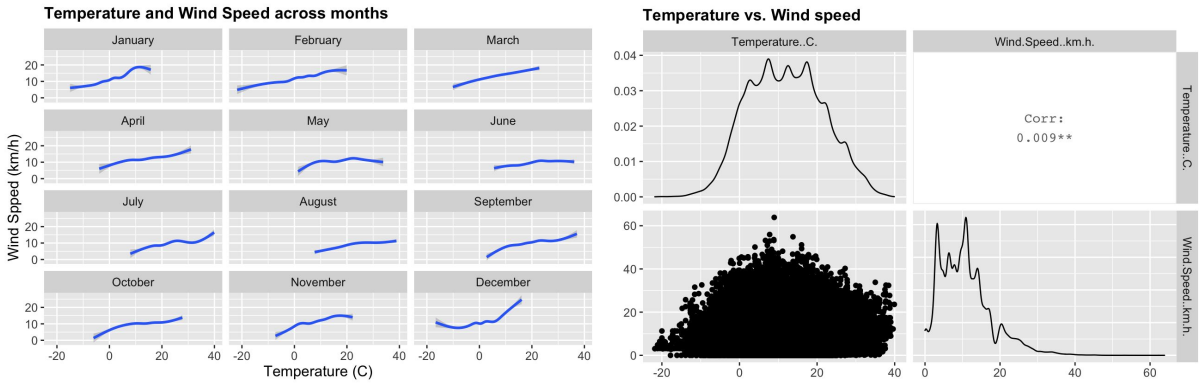
Question 3. What is the relationship between **Temperature** and **Visibility** in Szeged?(Monthly changes too)



Figure[5]: ggplot and ggparis of Visibility vs Temperature(in month for ggplot)

If we take a look at the graph above, the graph provides interesting information about the temperature and visibilities. If we take a look at the bottom left graph of temperature vs visibilities, we can see that the spreadness of the observations shows weak linear regression shape heading upper right side. Although the visibility around 13km is slightly cut off due to the lack of observations, the graph is somewhat clear that the two variables are correlated each other. The correlation ratio is 0.393 which is slightly low. It has a positive correlation. Like we mentioned in question 2, we can observe that the visibility goes up around 10 km in summer. As we predicted in question 2, it has a relationship with the temperature above 15 celsius degrees. All in all, we could find out weak positive correlation from the above graph.

Question 4. What is the relationship between **Temperature** and **Wind speed** in Szeged? (Monthly changes too)



Figure[6]: ggplot and ggpairs of Temperature vs Wind Speed(in month for ggplot)

If we take a look at the graph above, we can see that the relationship between wind speed and Temperature is not that strong. The correlation between the wind speed and visibility is 0.009 which is 0.9% of correlation. What we can see is that most observations are evenly spreaded through the temperature -20 ~20 degrees (in Celsius). If we take a look at the monthly graph, we can observe that the wind speed drops when the temperature drops also, the wind speed goes up significantly when the temperature goes high. Especially, we could observe the inflection point at 20 degrees Celsius. In conclusion, we can conclude that the relationship between temperature and wind speed is not significant and it has a positive relationship.

Predictive Model Using Multiple Regression

Through analyzing the relationship between certain numerical variables, we were interested in building a regression model for temperature. In order to better construct a predictive model for temperature, we transformed the temperature using log transformation.

Full Model

We first derived a predictive model for temperature with month as a factor. The full model contains all the predictor variables with month as a factor.

```
temp.lm1 = lm(log.temp ~  Apparent.Temperature..C. + Humidity + Wind.Speed..km.h. +
          Wind.Bearing..degrees. + Visibility..km. + Loud.Cover + Pressure..millibars.+ month +
          Apparent.Temperature..C.*month + Humidity * month + Wind.Speed..km.h. * month+
          Wind.Bearing..degrees. * month +
          Visibility..km. * month + Pressure..millibars. * month)
```

```
Residual standard error: 0.03529 on 96369 degrees of freedom
Multiple R-squared:  0.9837,    Adjusted R-squared:  0.9837
F-statistic: 7.013e+04 on 83 and 96369 DF,  p-value: < 2.2e-16
```

The R-squared value, coefficient of determination, indicates that the model with full predictor explains the temperature by 98%. However, although we obtained high R-squared value, the predictive model is too complicated, with too many predictor variables to be taken into account. Therefore, it is necessary to eliminate some of the predictor variables from the full model.

The full model contains predictor variables that are not necessary in explaining the dependent variable: temperatures. From the summary of the models, we decided to drop the variables that are not significant. For example, Loud Cover, Wind Bearing and month, and Pressure and month

presented high p-values that are unnecessary. After deleting unnecessary variables, we derived a refined model to predict temperature.

Reduced Final Model

```
temp.lm5 = lm(log.temp ~ Humidity +  Visibility..km. + month + Humidity * month +
           Visibility..km. * month)
```

```
Residual standard error: 0.1351 on 96417 degrees of freedom
Multiple R-squared:  0.7611,    Adjusted R-squared:  0.761
F-statistic:  8776 on 35 and 96417 DF,  p-value: < 2.2e-16
```

After eliminating unnecessary predictor variables, we reduced the full model into a simple and reliable model for predicting temperature. The model contains predictor variables of humidity, visibility, month, humidity with month as a factor, and visibility with month as a factor. With an R-squared value of 0.7611, the reduced final model may be used for predicting temperature in Szeged.

Summary

In conclusion, using the observations of 10 years of weather observations in Szeged, we conducted various statistical methods to figure out if there are some relationships between the temperature and other variables, to find the most accurate predictive model. For the predictive model, we used the month as a factor. Among 12 variables, 2 variables were two most significant variables correlated with the temperature which are **Humidity** and **Visibility**. We dropped all the other insignificant or less significant variables according to p-value. Therefore, We could predict that if the temperature in Szeged goes up, it is likely that the humidity goes down and visibility goes up, which means temperature has negative linear relationship with humidity and positive linear relationship with visibility

Appendix

- R Code

```
# Packility, Wind Speed of each pair
pairs( ~ Temperature..C.+Humidity+Wind.Speed..km.h.+Visibility..km.+ Wind.Speed..km.h.,
      data = wh, main = "Weather History Scatter Plot")



#Transform date into months
wh$Formatted.Date = as.Date(wh$Formatted.Date)
wh$Formatted.Date = months(wh$Formatted.Date)
wh$Formatted.Date

# We have received information about faceting grid according to monthly order from the
following website:
# https://www.neonscience.org/dc-time-series-plot-facets-ndvi-r

#Make them into a factor and add it
wh$month_name = factor(wh$Formatted.Date, levels = c('January','February','March',
                                        'April','May','June','July',
                                        'August','September','October',
                                        'November','December'))

####### ggplots

#ggplots of Temperature and Humidity for each month
plot1 = ggplot(wh, aes(Temperature..C. , Humidity))  + facet_wrap(~month_name, nc = 3) +
  geom_smooth() + ggtitle("Temperature and Humidity across months") + xlab("Temperature
(C)") +
  ylab("Humidity") +theme(plot.title = element_text(lineheight=.8, face="bold",size = 13)) +
  theme(text = element_text(size=12))

#ggplots of Visibility and Humidity for each month
plot2 = ggplot(wh, aes(Visibility..km., Humidity))  + facet_wrap(~month_name, nc = 3) +
  geom_smooth() +ggtitle("Visibility and Humidity across months") + xlab("Visibility (km)") +
  ylab("Humidity") + theme(plot.title = element_text(lineheight=.8, face="bold",size = 13)) +
  theme(text = element_text(size=12))

#ggplots of Temperature and Visibility for each month
plot3 = ggplot(wh, aes(Temperature..C., Visibility..km.))  + facet_wrap(~month_name, nc = 3) +
```

```
  geom_smooth() + ggtitle("Temperature and Visibility across months") + xlab("Temperature
(C)") +
  ylab("Visibility (km)") + theme(plot.title = element_text(lineheight=.8, face="bold",size = 13)) +
  theme(text = element_text(size=12))

#ggplots of Temperature and Wind Speed for each month
plot4 = ggplot(wh, aes(Temperature..C., Wind.Speed..km.h.))  + facet_wrap(~month_name, nc
= 3) +
  geom_smooth() + ggtitle("Temperature and Wind Speed across months") + xlab("Temperature
(C)") +
  ylab("Wind Spped (km/h)") + theme(plot.title = element_text(lineheight=.8, face="bold",size =
13)) +
  theme(text = element_text(size=12))

#Show ggplots
plot1
plot2
plot3
plot4


####### ggpairs

#ggpairs of Temperature vs. Humidity
pair1 = ggpairs(data= wh, columns = c(4,6), title = "Temperature vs. Humidity") +
  theme(plot.title = element_text(lineheight=.8, face="bold",size = 13)) +
  theme(text = element_text(size=12))

#ggpairs of Visibility vs. Humidity
pair2 = ggpairs(data= wh, columns = c(9,6), title = "Visibility vs. Humidity") +
  theme(plot.title = element_text(lineheight=.8, face="bold",size = 13)) +
  theme(text = element_text(size=12))

#ggpairs of Temperature and Visibility
pair3 = ggpairs(data= wh, columns = c(4,9), title = "Temperature vs. Visibility") +
  theme(plot.title = element_text(lineheight=.8, face="bold",size = 13)) +
  theme(text = element_text(size=12))

#ggpairs of Temperature vs. Wind speed
pair4 = ggpairs(data= wh, columns = c(4,7), title = "Temperature vs. Wind speed") +
  theme(plot.title = element_text(lineheight=.8, face="bold",size = 13)) +
  theme(text = element_text(size=12))
```

```
#Show ggpairs
pair1
pair2
pair3
pair4

#####Predictive model

qplot(Temperature..C., data  = na.omit(wh), binwidth = 0.3)
skewness(wh$Temperature..C.)
qplot(log(Temperature..C.+ 25), data = na.omit(wh), binwidth = 1/8)
log.temp = log(Temperature..C. + 25)
month = as.factor(wh$Formatted.Date)



# Temperature predictive model with month factor
# First.
temp.lm1 = lm(log.temp ~  Apparent.Temperature..C. + Humidity + Wind.Speed..km.h. +
            Wind.Bearing..degrees. + Visibility..km. + Loud.Cover + Pressure..millibars.+ month +
            Apparent.Temperature..C.*month + Humidity * month + Wind.Speed..km.h. * month+
            Wind.Bearing..degrees. * month +
            Visibility..km. * month + Pressure..millibars. * month)




summary(temp.lm1)

# Second.
# drop Loud cover, wind.Bearing..degrees * month, pressure * month
temp.lm2 = lm(log.temp ~  Apparent.Temperature..C. + Humidity + Wind.Speed..km.h. +
            Wind.Bearing..degrees. + Visibility..km. + Pressure..millibars.+ month +
            Apparent.Temperature..C.*month + Humidity * month + Wind.Speed..km.h. * month+
            Visibility..km. * month)

summary(temp.lm2)



# Third
# drop Apparent.Temperature..C., Apparent.Temperature..C. * month
temp.lm3 = lm(log.temp ~ Humidity + Wind.Speed..km.h. +
            Wind.Bearing..degrees. + Visibility..km. + Pressure..millibars.+ month +
            Humidity * month + Wind.Speed..km.h. * month+
```

```
            Visibility..km. * month)

summary(temp.lm3)

# but it is still too complicated
# so we drop Wind.Bearing..degrees

temp.lm4 = lm(log.temp ~ Humidity + Wind.Speed..km.h. +
            Visibility..km. + month +
            Humidity * month + Wind.Speed..km.h. * month+
            Visibility..km. * month)

summary(temp.lm4)

# Final
# drop wind Wind.Speed..km.h.
temp.lm5 = lm(log.temp ~ Humidity +
            Visibility..km. + month +
            Humidity * month +
            Visibility..km. * month)

summary(temp.lm5)
```