Removing the Singularities

a) Detecting the singularities:

R-Code:

```
data<-read.csv("1.csv",TRUE)
fit<-
lm(data$Air.Pollution~data$Temperature+data$Relative.Humidity+data$Heat.I
ndex+data$Carbon.Monoxide+data$Noise.Pollution)
summary(fit)
```

Output:

```
Call:
lm(formula = data$Air.Pollution ~ data$Temperature + data$Relative.Humidit
y + 
    data$Heat.Index + data$Carbon.Monoxide + data$Noise.Pollution)

Coefficients:
           (Intercept)        data$Temperature  data$Relative.Humidity  
             13.202107               -0.003891                0.003028  
       data$Heat.Index    data$Carbon.Monoxide    data$Noise.Pollution  
                    NA                0.374373                0.262325  

> summary(fit)
Call:
lm(formula = data$Air.Pollution ~ data$Temperature + data$Relative.Humidit
y + 
    data$Heat.Index + data$Carbon.Monoxide + data$Noise.Pollution)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.8556  -3.4118  -0.1456   3.2127  26.5532 

Coefficients: (1 not defined because of singularities)
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            13.202107   0.213371  61.874   <2e-16 ***
data$Temperature       -0.003891   0.004463  -0.872   0.3834    
data$Relative.Humidity  0.003028   0.001794   1.688   0.0915 .  
data$Heat.Index               NA         NA      NA       NA    
data$Carbon.Monoxide    0.374373   0.005751  65.098   <2e-16 ***
data$Noise.Pollution    0.262325   0.004015  65.334   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.464 on 30879 degrees of freedom
Multiple R-squared:  0.5489,	Adjusted R-squared:  0.5489 
F-statistic:  9394 on 4 and 30879 DF,  p-value: < 2.2e-16
```

Inference:

By finding the present of singularities:
We need to find the variables among which there is more correlation.

b) Finding the most correlated variable:

R - Code:

```
d<-
cbind(data$Temperature,data$Relative.Humidity,data$Heat.Index,data$Carbon.Monoxide,
data$Noise.Pollution)

cor(d)
```

Output:

```
              [,1]         [,2]         [,3]         [,4]          [,5]
[1,]  1.0000000000  0.537889412  0.996426740 -0.002864487 -0.000697518
[2,]  0.5378894123  1.000000000  0.607169774 -0.002928457 -0.003740822
[3,]  0.9964267399  0.607169774  1.000000000 -0.002993283 -0.001032225
[4,] -0.0028644871 -0.002928457 -0.002993283  1.000000000  0.773613603
[5,] -0.0006975168 -0.003740820 -0.001032227  0.773613605  1.000000000
```

Inference:

From the above output it is inferred that, there is stronger relation between the third variable (Heat Index) and the first variable (Temperature)

While determining which factor to be eliminated among the two variables i.e Heat Index and Temperature. Linear Regression model is used to determine it,

c) Removing the singularities:

R - Code:

```
setwd("G:/no food waste intern")
data<-read.csv("1.csv",TRUE)
#data
fit1<-lm(data$Air.Pollution~data$Heat.Index)
summary(fit1)
fit2<-lm(data$Air.Pollution~data$Temperature)
summary(fit2)
```

Output:

```
> fit1<-lm(data$Air.Pollution~data$Heat.Index)> > summary(fit1)
Call:
lm(formula = data$Air.Pollution ~ data$Heat.Index)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.043  -4.382   0.368   5.199  36.319 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)     45.0617662  0.1328498 339.193   <2e-16 ***
data$Heat.Index -0.0007376  0.0048001  -0.154    0.878    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.646 on 30882 degrees of freedom
Multiple R-squared:  7.647e-07,	Adjusted R-squared:  -3.162e-05 
F-statistic: 0.02361 on 1 and 30882 DF,  p-value: 0.8779
G-
> 
> fit2<-lm(data$Air.Pollution~data$Temperature)
> 
 > summary(fit2)
Call:
lm(formula = data$Air.Pollution ~ data$Temperature)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.043  -4.382   0.367   5.201  36.319 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      45.079626   0.176404 255.547   <2e-16 ***
data$Temperature -0.001217   0.005602  -0.217    0.828    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.646 on 30882 degrees of freedom
Multiple R-squared:  1.528e-06,	Adjusted R-squared:  -3.085e-05 
F-statistic: 0.04719 on 1 and 30882 DF,  p-value: 0.828
```

Inference:

While considering the above two Linear models. We must look at the Multiple R squared value which shows the factor that determines the dependent variable the most.
From the above output, it is evident that the Temperature explains the Air pollution more than the Heat index, So that Heat index can be eliminated.

Correlation

R-Code:

```
setwd("F:/no food waste intern/day-3")
data<-read.csv("ocd.csv",TRUE)
data$Year=NULL
data$Month=NULL
data$Date=NULL
data
new=cbind(data)
new
c<-cor(new)
c
```

Output:

```
                 Temperature   Humidity Carbon.Monoxide Air.Pollution No
ise.Pollution


Temperature     1.000000000 0.48341323     0.001605298    0.01346443      0.01547675
Humidity        0.483413227 1.00000000     0.036872460    0.04523131      0.01605351
Carbon.Monoxide 0.001605298 0.03687246     1.000000000    0.90707296      0.54755120
Air.Pollution   0.013464427 0.04523131     0.907072958    1.00000000      0.54623214
Noise.Pollution 0.015476748 0.01605351     0.547551200    0.54623214      1.00000000
```

Inference:

The above one is a correlation matrix after removing the singularities.

Regression:

R – Code:

```
setwd("F:/no food waste intern/day-4")
data<-read.csv("lm.csv",TRUE)
data
lm1<-
lm(data$Average.of.Temperature~data$Average.of.Relative.Humidity)
lm1
summary(lm1)
lm2<-
lm(data$Average.of.Air.Pollution~data$Average.of.Noise.Pollution)
lm2
summary(lm2)
```

Output:

```
> lm1<-lm(data$Average.of.Temperature~data$Average.of.Relative.Humidity)
> lm1

Call:
lm(formula = data$Average.of.Temperature ~ data$Average.of.Relative.Humidi
ty)

Coefficients:
                      (Intercept)  data$Average.of.Relative.Humidity
                        30.219866                           0.007826

> summary(lm1)

Call:
lm(formula = data$Average.of.Temperature ~ data$Average.of.Relative.Humidi
ty)

Residuals:
     Min       1Q   Median       3Q      Max
-0.17971 -0.09250 -0.01504  0.07180  0.31421

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      30.219866   3.513224   8.602 4.42e-09 **
*
data$Average.of.Relative.Humidity  0.007826   0.055110   0.142    0.888
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1312 on 26 degrees of freedom
Multiple R-squared:  0.0007751,       Adjusted R-squared:  -0.03766
F-statistic: 0.02017 on 1 and 26 DF,  p-value: 0.8882

> lm2<-lm(data$Average.of.Air.Pollution~data$Average.of.Noise.Pollution)
> lm2

Call:
lm(formula = data$Average.of.Air.Pollution ~ data$Average.of.Noise.Polluti
on)

Coefficients:
                    (Intercept)  data$Average.of.Noise.Pollution
                         14.303                            0.573

> summary(lm2)

Call:
lm(formula = data$Average.of.Air.Pollution ~ data$Average.of.Noise.Polluti
on)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9687 -0.4511  0.1738  0.3308  0.5849
```

```
Coefficients:
                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    14.30337    0.55221   25.90   <2e-16 ***
data$Average.of.Noise.Pollution  0.57302    0.01015   56.47   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4693 on 26 degrees of freedom
Multiple R-squared:  0.9919,	Adjusted R-squared:  0.9916 
F-statistic:  3189 on 1 and 26 DF,  p-value: < 2.2e-16
```

Inference:

From the summary of the two models it is inferred that, temperature is less dependent on humidity. Whereas Noise pollution is more dependent on Air pollution.

Multiple Linear Regression

R-Code:

a<-read.csv("ocd.csv",TRUE)

fit<-lm(a$Air.Pollution~a$Temperature+a$Humidity+a$Carbon.Monoxide+a$Air.Pollution+a$Noise.Pollution)

fit

summary(fit)

Output:

```
 > summary(fit)
Call:
lm(formula = a$Air.Pollution ~ a$Temperature + a$Humidity + a$Carbon.Monox
ide + 
    a$Air.Pollution + a$Noise.Pollution)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.80453 -0.50740 -0.00439  0.55019  2.72718 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -18.738937   1.155342 -16.219  < 2e-16 ***
a$Temperature     0.002069   0.005769   0.359 0.719961    
a$Humidity        0.001042   0.002199   0.474 0.635782    
a$Carbon.Monoxide  1.281859   0.028863  44.411  < 2e-16 ***
a$Noise.Pollution  0.057168   0.015794   3.620 0.000318 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.7946 on 650 degrees of freedom
Multiple R-squared:  0.8265,   Adjusted R-squared:  0.8254 
F-statistic: 773.9 on 4 and 650 DF,  p-value: < 2.2e-16
```

Inference:

The above summary shows that with the attributes such as temperature, humidity, temperature, carbon mono oxide and noise pollution we could predict air pollution. From the multiple R squared value it is evident that we could determine the air pollution.