BIO3SA3_Project_Group6

JY

12/11/2020

Group 6, Lu Qiao, Jiayi Mo

```
df = read.csv(file.choose(), stringsAsFactors = F)#import data frame
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

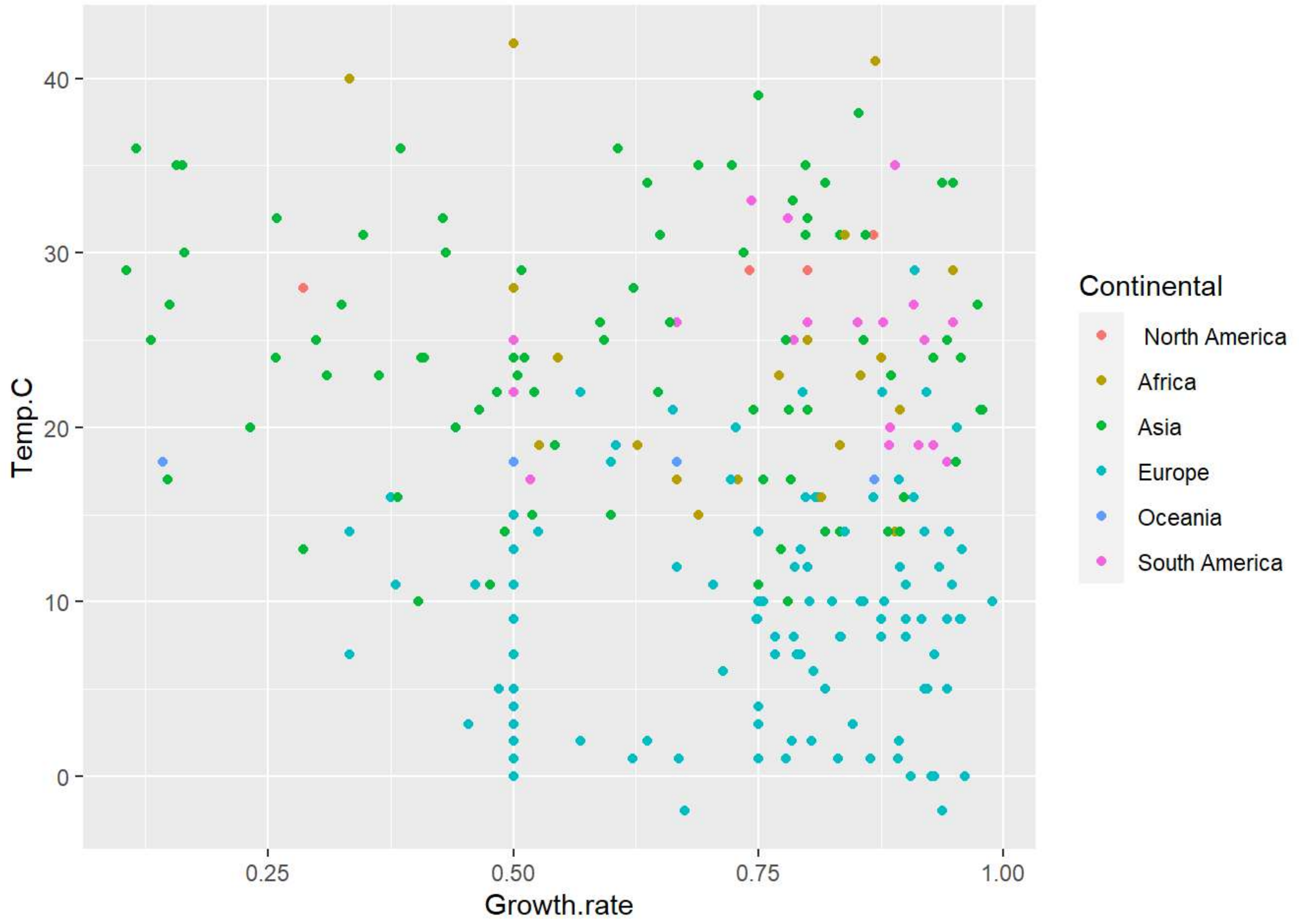
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

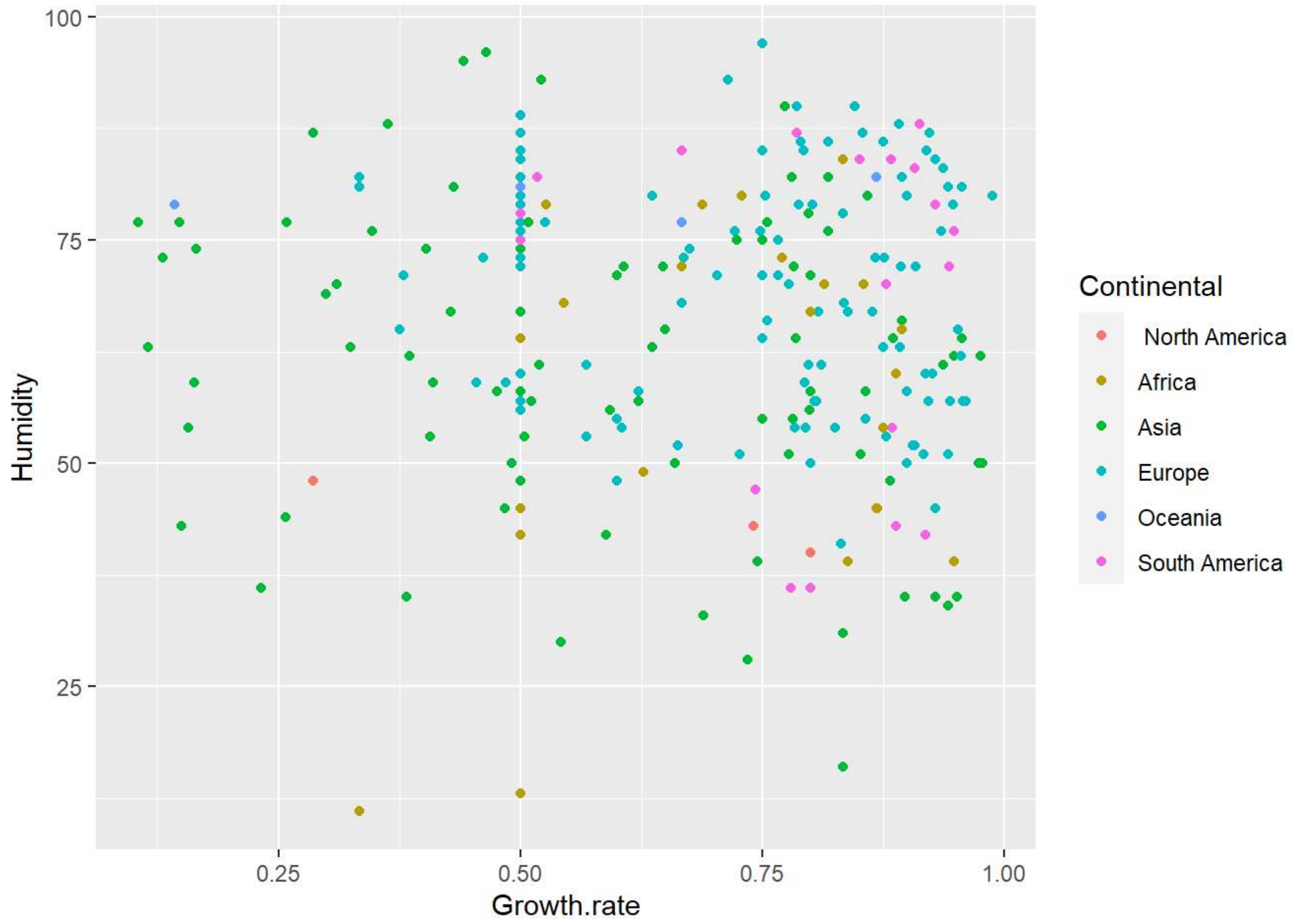
```
head(df)
```

```
##       Country Continental Day confirmed Deaths   City Humidity SunHour Temp.C
## 1 Afghanistan        Asia   1         0      0  Kabul       51     8.7      9
## 2 Afghanistan        Asia   7         1      0  Kabul       74     7.2      9
## 3 Afghanistan        Asia  14         1      0  Kabul       39     8.7     10
## 4 Afghanistan        Asia  21        11      0  Kabul       31    11.6     14
## 5 Afghanistan        Asia  30        24      0  Kabul       61     9.2     15
## 6     Albania      Europe   1         0      0 Tirana       54     8.7     12
##   Windspeed Population
## 1         8   17.47931
## 2         5   17.47931
## 3         7   17.47931
## 4         5   17.47931
## 5         5   17.47931
## 6         9   14.87242
```

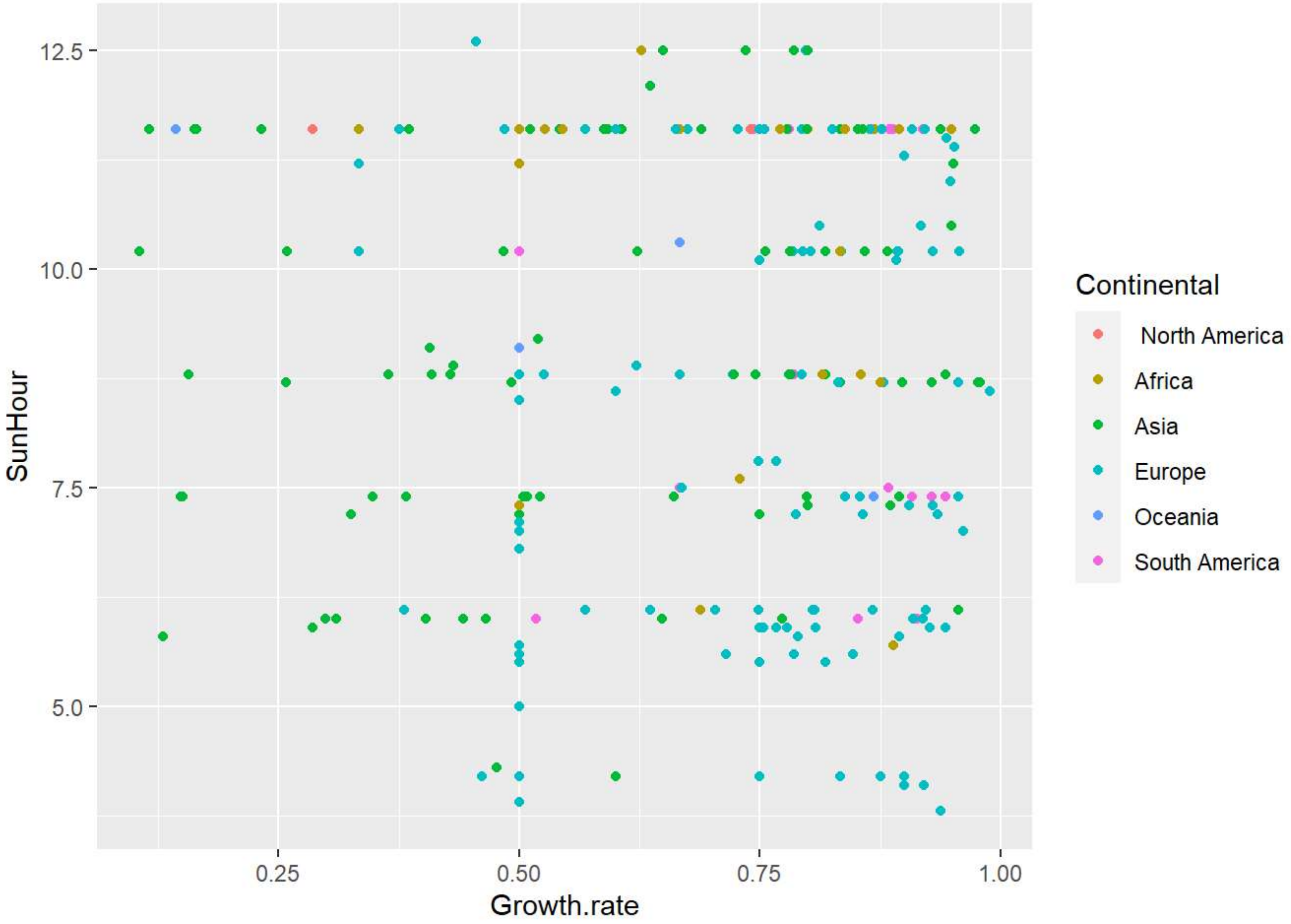
```
#we found a mistake of day's name, so we replace it here
df$Day[df$Day == 30] <- 28

#delete countries most confirmed cases are 0 by looking at df data table
test<-df[-c(6:10,16:20,31:35,51:60,71:85,96:105,111:115,131:145,160:164,170:179,
           185:204,215:219,225:259,300:304,310:324,330:334,345:354,379:374,380:399,
           405:409,415:424,430:439,465:474,500:504,515:524,540:544,560:564,580:589,
           605:609,615:624,634:638,649:663,669:673),]
df1<-na.omit(test)#delete empty/NA values

cnt<- data.frame(unique(df$Country))#original number of countries
cnt<- data.frame(unique(df1$Country))#numbers of country after deletion

#calculate growth rate
#credit from: https://community.rstudio.com/t/growth-rate-calculation-in-r/38675/2
dg<-subset(df1, select= c(confirmed))#create a new data frame named dg that contain confirmed
cases only
#build a function and apply to dg
Y <- function(x)x+1#confirmed +1 because log 0 is error
dg1<-data.frame(lapply(dg,Y))
# we did not subset original day from df1 directly, because it is not convenient for calculat
ion
dg1$day <- c(1:349)#add a column of day,

growth_rate = dg1 %>%
    arrange(day) %>%
    mutate(Diff_day = day - lag(day),
           Diff_growth = confirmed - lag(confirmed),
           rate_percent = (Diff_growth / Diff_day)/confirmed)
options(scipen=999)#disable scientific rotation

gr<-subset(growth_rate, select= c(rate_percent))#subset rate
df2<-cbind(df1,gr)
df3 <- df2[df2$rate_percent>0,]#exclude value <=0
df4<-na.omit(df3)#delete empty/NA values
colnames(df4)[12] <- "Growth.rate"#rename growth_percent to Growth.rate


library(ggplot2)

#Scatter plot of growth rate vs. variables, labeled by Continental
ggplot(df4, aes(x=Growth.rate, y=Temp.C, color=Continental)) + geom_point()
```



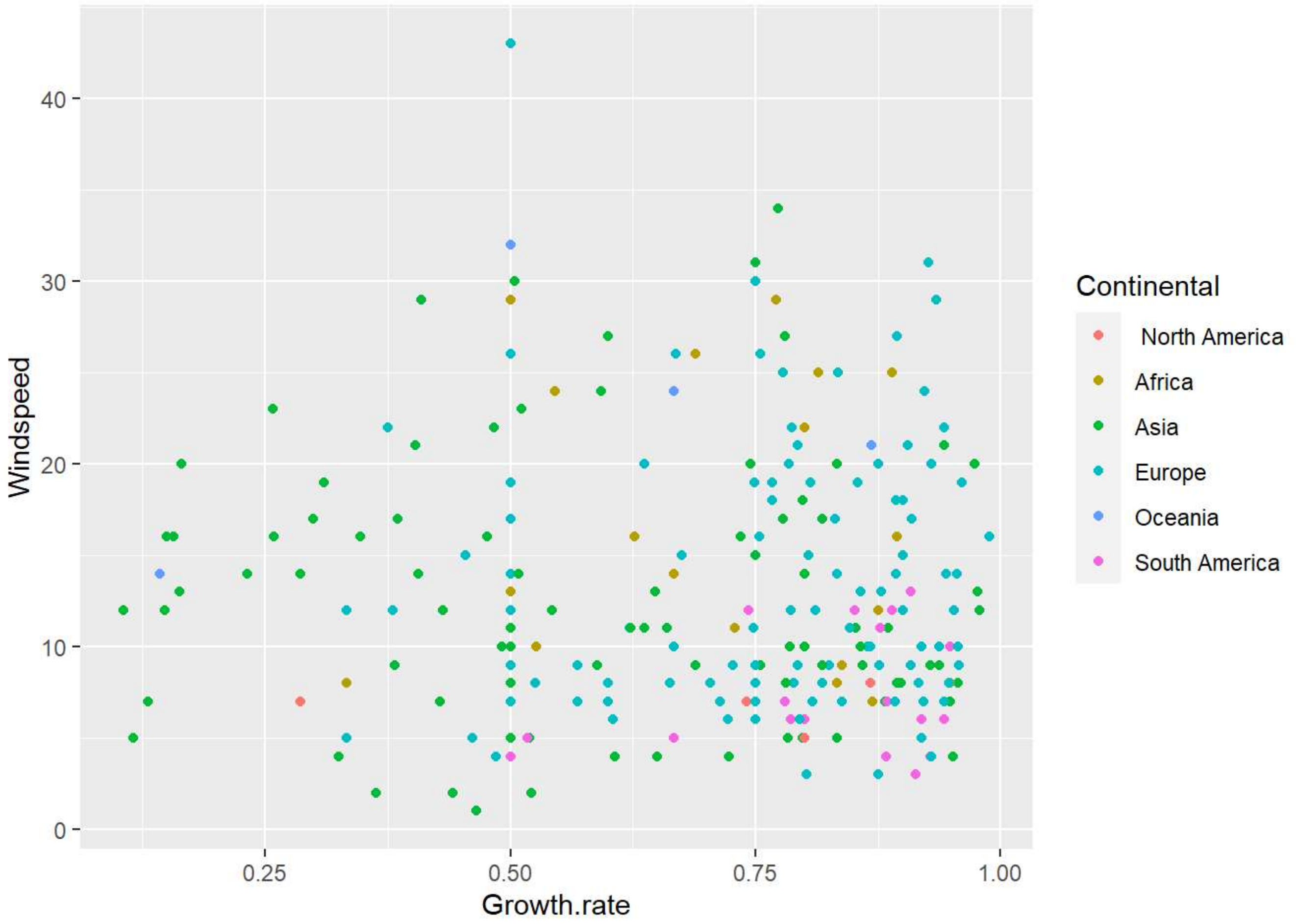
```
ggplot(df4, aes(x=Growth.rate, y=Humidity, color=Continental)) + geom_point()
```



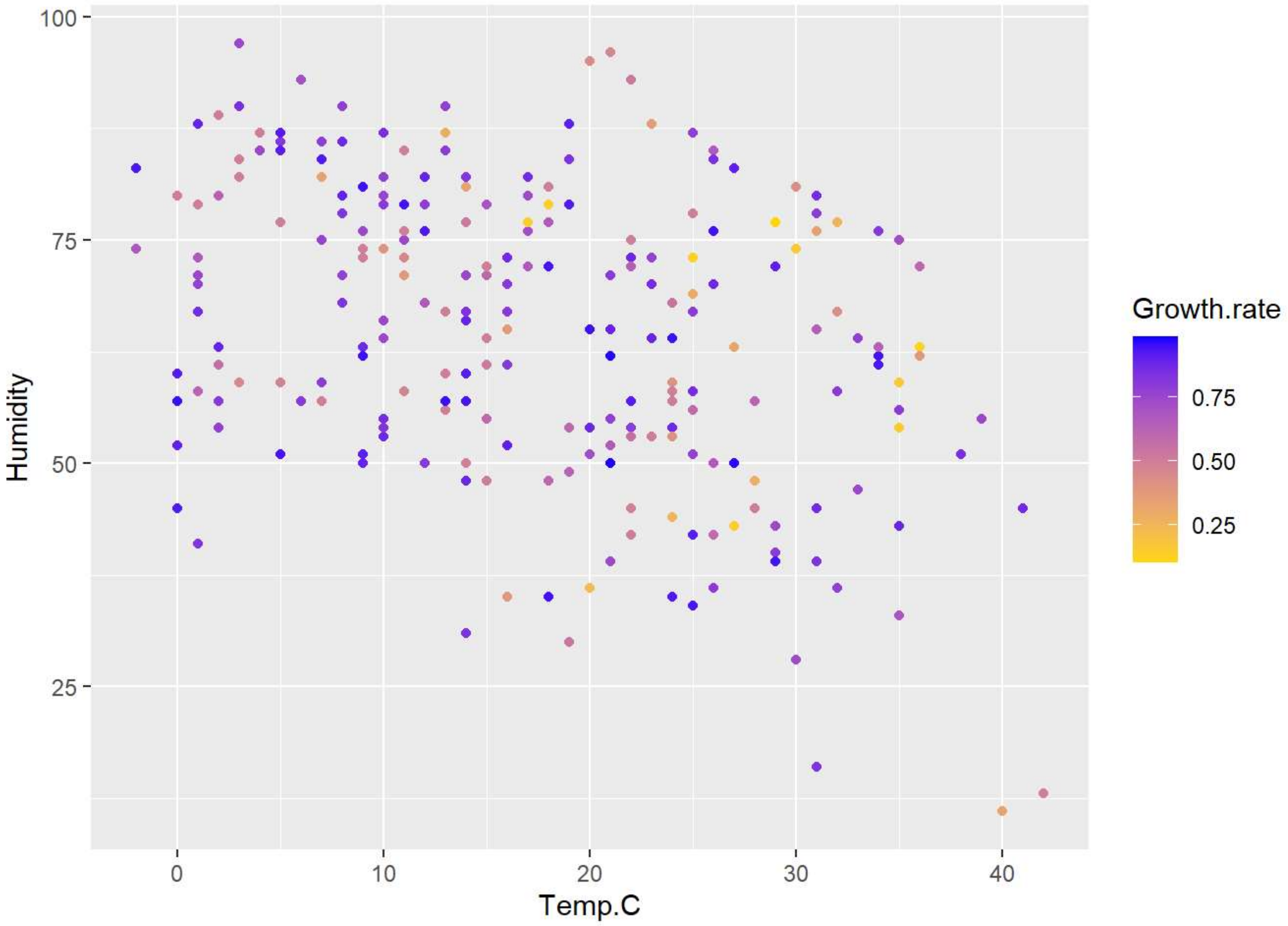
```
ggplot(df4, aes(x=Growth.rate, y=SunHour, color=Continental)) + geom_point()
```



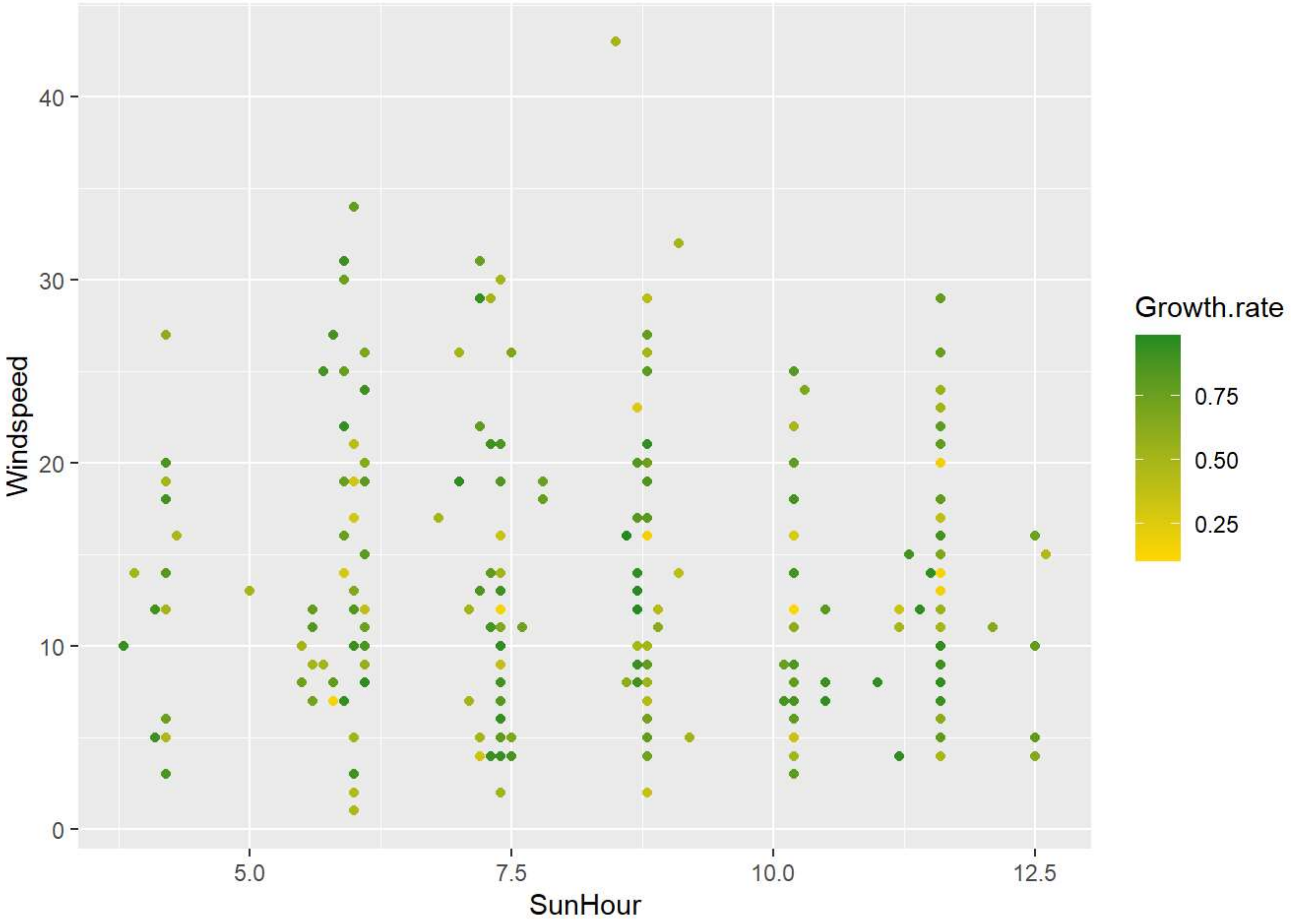
```
ggplot(df4, aes(x=Growth.rate, y=Windspeed, color=Continental)) + geom_point()
```



```
#scatter plot of Temp and humidity, labeled by growth rate
ggplot(df4, aes(x=Temp.C, y=Humidity, color=Growth.rate)) +
    geom_point()+scale_color_gradient(low="gold", high="blue")
```



```
#scatter plot of sunlight and wind speed
ggplot(df4, aes(x=SunHour, y=Windspeed, color=Growth.rate)) +
    geom_point()+scale_color_gradient(low="gold", high="forestgreen")
```



```
##PCA
library(vegan)  #bstick and screeplot come from vega
```

```
## Loading required package: permute
```

```
## Loading required package: lattice
```

```
## This is vegan 2.5-7
```

```
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
df.pca<-prcomp(df4[,c(7,8,9,10)],scale=TRUE)
#PCA on correlation matrix,
summary(df.pca) #eigenvalues
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4
## Standard deviation     1.3983 1.0051 0.7631 0.6726
## Proportion of Variance 0.4888 0.2525 0.1456 0.1131
## Cumulative Proportion  0.4888 0.7413 0.8869 1.0000
```

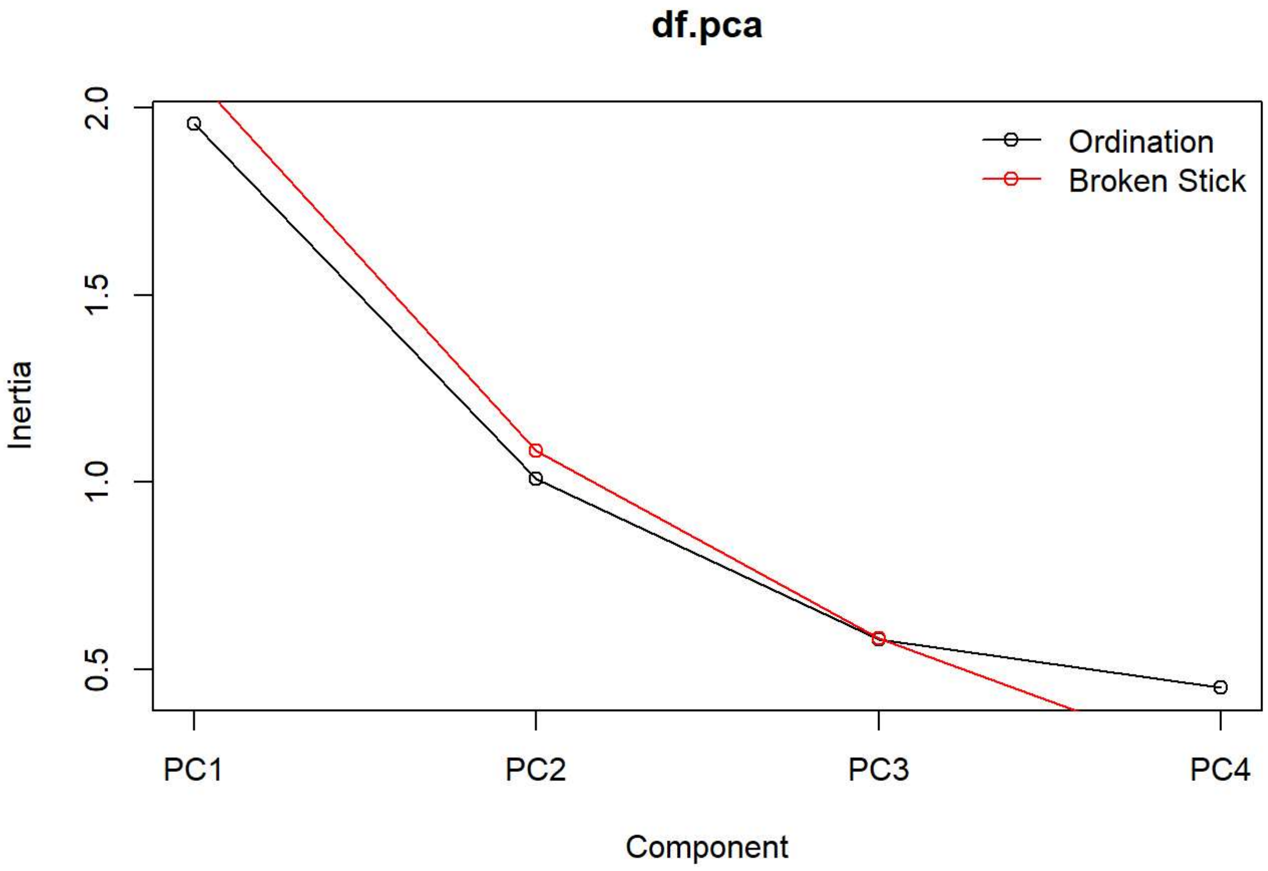
```
print(df.pca)#eigenvectors eigenvalues
```

```
## Standard deviations (1, .., p=4):
## [1] 1.3982521 1.0050795 0.7631294 0.6725621
##
## Rotation (n x k) = (4 x 4):
##                  PC1        PC2         PC3           PC4
## Humidity  -0.5276576  0.3753306 -0.62425109  0.4370525827
## SunHour    0.5982021 -0.1323053 -0.03230965  0.7896870744
## Temp.C     0.5618803  0.1397048 -0.69238295 -0.4305565457
## Windspeed -0.2191397 -0.9067000 -0.36037830 -0.0006522439
```

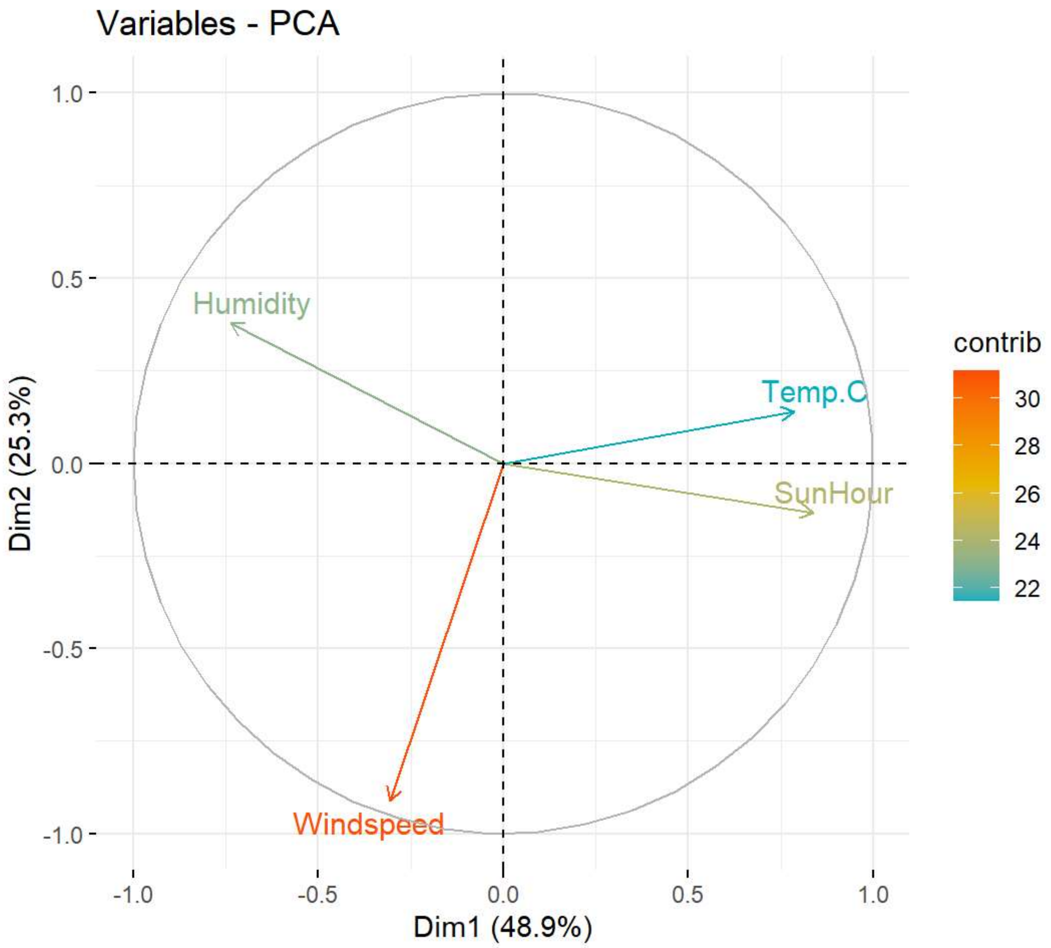
```
bstick(df.pca)
```

```
##       PC1       PC2       PC3       PC4
## 2.0833333 1.0833333 0.5833333 0.2500000
```

```
screeplot(df.pca, bstick = TRUE, type = "lines")
```



```
#PCA variation plot
fviz_pca_var(df.pca,
             axes = c(1,2),
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE,     # Avoid text overlapping
)
```



```
##We found that PC1 and PC2 is useful in our data set
##So we choose all four variables

##linear regression
##Using linear regression to test the relationship between covid19 growth rate and our choose
n variables
lm1 = lm(Growth.rate ~ Humidity + SunHour + Windspeed + Temp.C, data=df4)
summary(lm1)#only temperature is significant
```

```
## 
## Call:
## lm(formula = Growth.rate ~ Humidity + SunHour + Windspeed + Temp.C, 
##     data = df4)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.54228 -0.17686  0.05207  0.17319  0.32553 
## 
## Coefficients:
##               Estimate Std. Error t value         Pr(>|t|)    
## (Intercept)  0.8544344  0.1122967   7.609 0.00000000000065 ***
## Humidity    -0.0013041  0.0009856  -1.323         0.187074    
## SunHour      0.0062039  0.0069169   0.897         0.370671    
## Windspeed   -0.0023305  0.0019304  -1.207         0.228541    
## Temp.C      -0.0058671  0.0016048  -3.656         0.000315 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2133 on 237 degrees of freedom
## Multiple R-squared:  0.05627,    Adjusted R-squared:  0.04034 
## F-statistic: 3.533 on 4 and 237 DF,  p-value: 0.008022
```

```
#we want to delete the most insignificant variables and to see if there are any change to the result
lm2 = lm(Growth.rate ~ Humidity + Windspeed + Temp.C, data=df4)#delete sunhour
summary(lm2)#Wind speed is the most insignificant, temp still significant
```

```
## 
## Call:
## lm(formula = Growth.rate ~ Humidity + Windspeed + Temp.C, data = df4)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.52853 -0.17186  0.05466  0.17358  0.32539 
## 
## Coefficients:
##               Estimate Std. Error t value             Pr(>|t|)    
## (Intercept)  0.9238954  0.0812870  11.366 < 0.0000000000000002 ***
## Humidity    -0.0016563  0.0009037  -1.833             0.068064 .  
## Windspeed   -0.0024248  0.0019267  -1.258             0.209449    
## Temp.C      -0.0053283  0.0014875  -3.582             0.000413 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2132 on 238 degrees of freedom
## Multiple R-squared:  0.05306,    Adjusted R-squared:  0.04113 
## F-statistic: 4.446 on 3 and 238 DF,  p-value: 0.004628
```

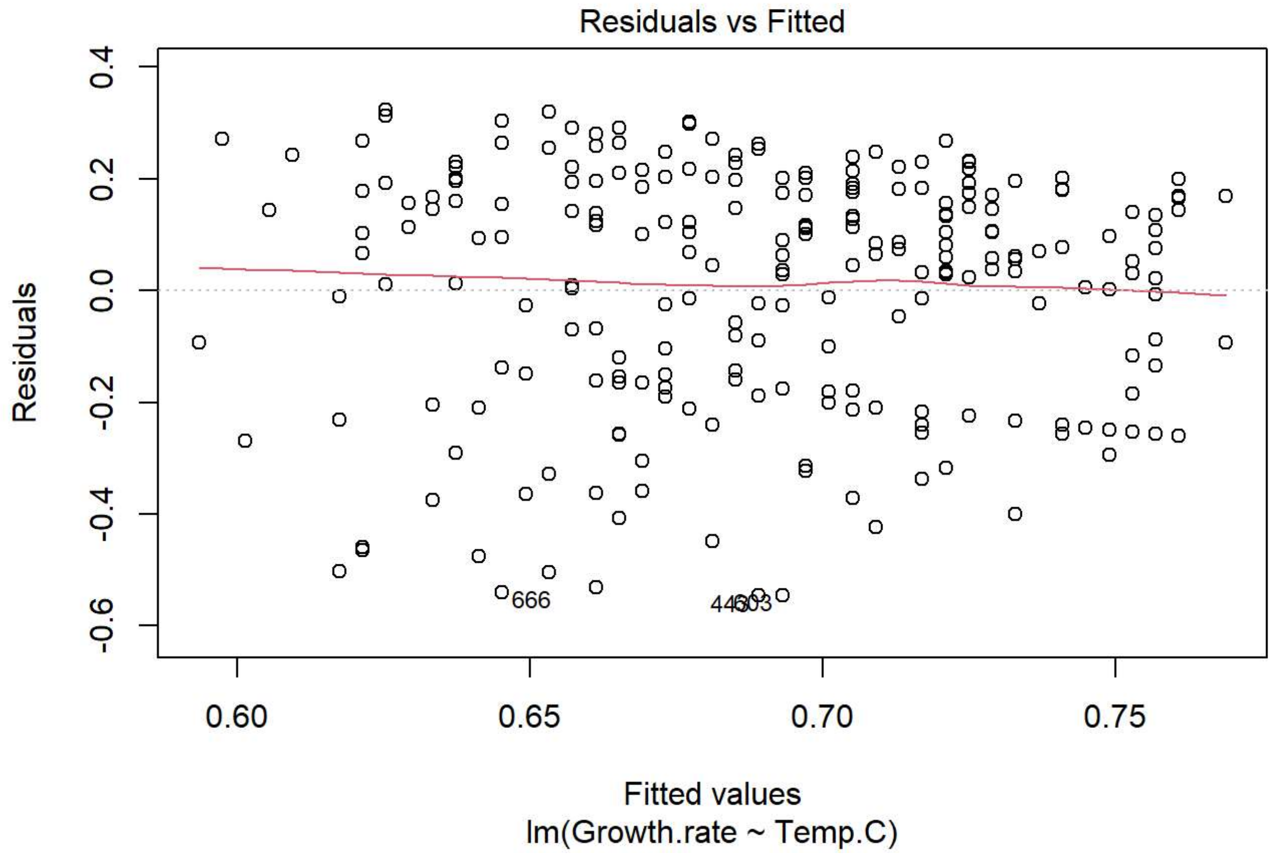
```
lm3 = lm(Growth.rate ~ Humidity + Temp.C, data=df4)#delete wind speed
summary(lm3)#Humidity is the most insignificant, temp still significant
```

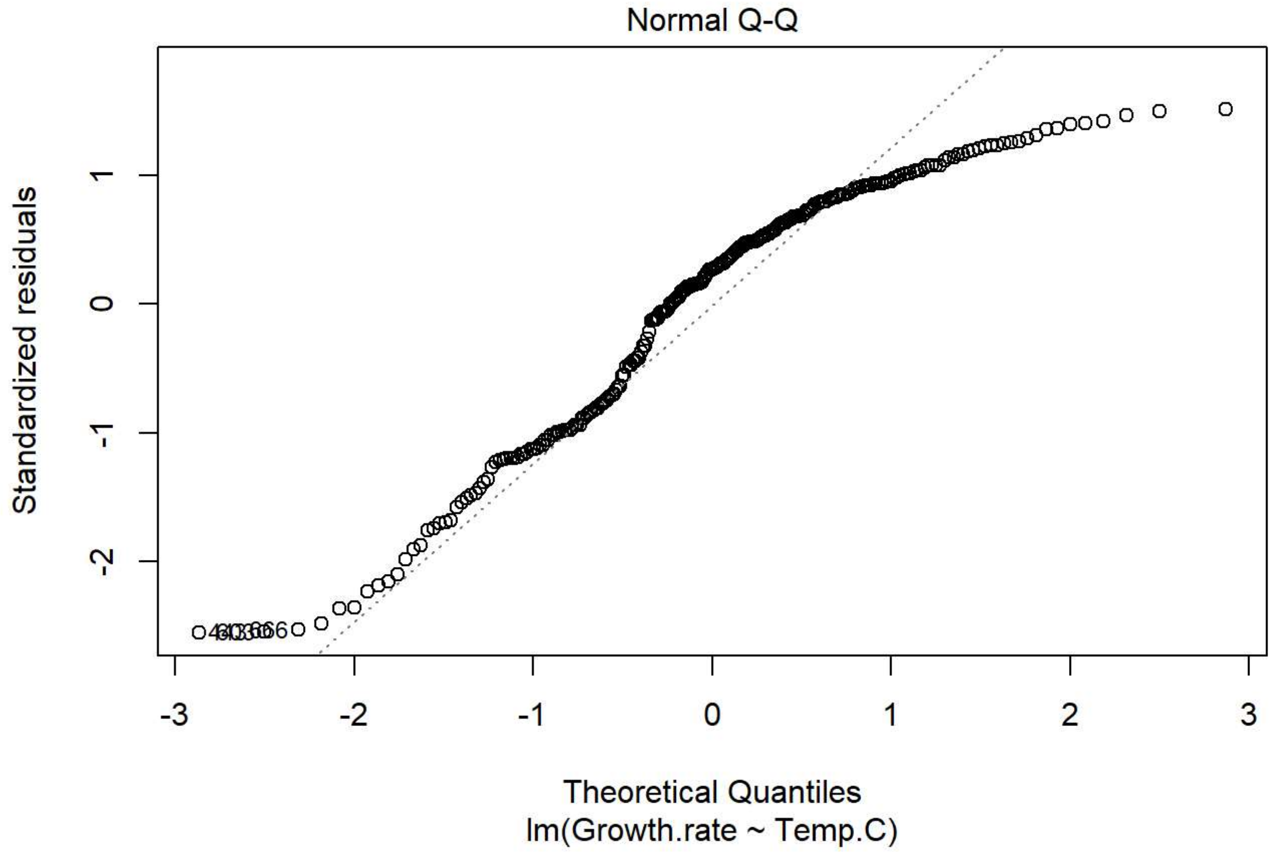
```
##
## Call:
## lm(formula = Growth.rate ~ Humidity + Temp.C, data = df4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5293 -0.1728  0.0478  0.1712  0.3332
##
## Coefficients:
##                Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  0.8789668  0.0731168  12.021 < 0.0000000000000002 ***
## Humidity    -0.0015723  0.0009023  -1.743             0.082700 .
## Temp.C      -0.0048898  0.0014479  -3.377             0.000854 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2134 on 239 degrees of freedom
## Multiple R-squared:  0.04676,    Adjusted R-squared:  0.03879
## F-statistic: 5.862 on 2 and 239 DF,  p-value: 0.00327
```

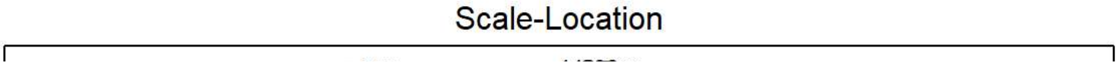
```
lm4 = lm(Growth.rate ~ Temp.C, data=df4)#delete humidity
#temperature is the only significant variable, so we test how it fits to a linear regression
 when dependent is growth rate
summary(lm4)
```

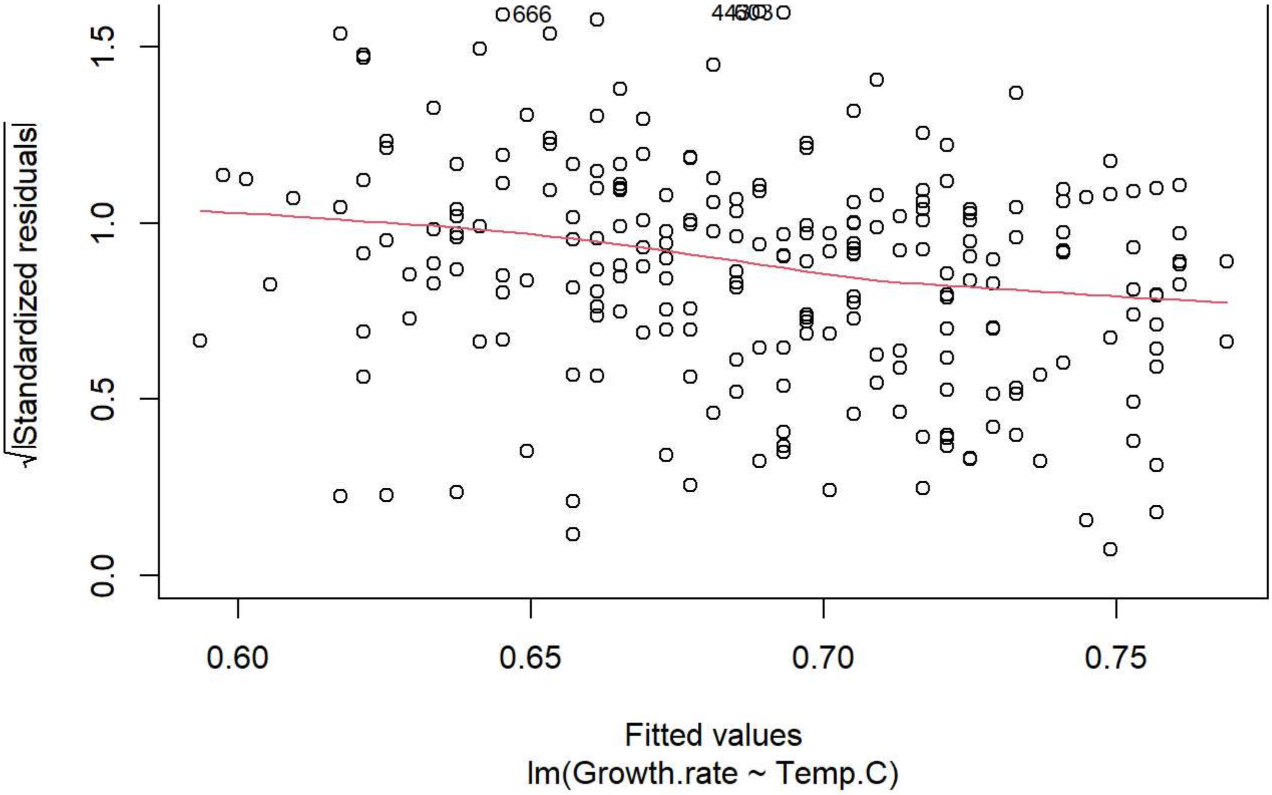
```
##
## Call:
## lm(formula = Growth.rate ~ Temp.C, data = df4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5462 -0.1782  0.0582  0.1765  0.3231
##
## Coefficients:
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  0.760756   0.027393  27.772 < 0.0000000000000002 ***
## Temp.C      -0.003981   0.001356  -2.935              0.00366 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2143 on 240 degrees of freedom
## Multiple R-squared:  0.03465,    Adjusted R-squared:  0.03063
## F-statistic: 8.615 on 1 and 240 DF,  p-value: 0.003658
```

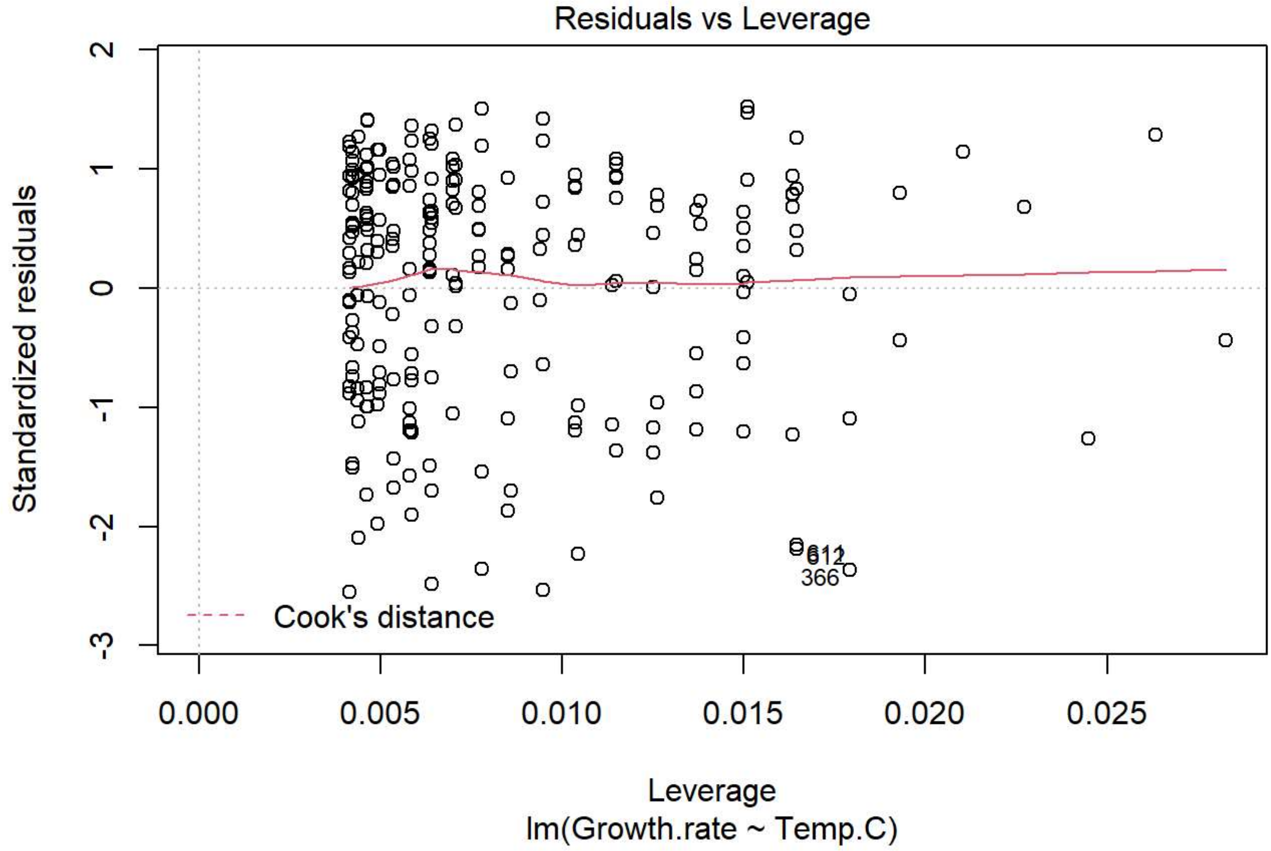
```
plot(lm4)#plot growth rate vs. temperature
```


Residuals vs Fitted
Residuals
0.4
0.2
0.0
-0.2
-0.4
-0.6
666
0.60
0.65
0.70
0.75
Fitted values
lm(Growth.rate ~ Temp.C)


Normal Q-Q
Standardized residuals
1
0
-1
-2
666
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Growth.rate ~ Temp.C)


Scale-Location


√|Standardized residuals|
0.0
0.5
1.0
1.5
0.60
0.65
0.70
0.75
Fitted values
lm(Growth.rate ~ Temp.C)
666
4403


Residuals vs Leverage
Standardized residuals
-3
-2
-1
0
1
2
0.000
0.005
0.010
0.015
0.020
0.025
Leverage
lm(Growth.rate ~ Temp.C)
Cook's distance
612
366

```
##AIC
step(lm1)
```

```
## Start:  AIC=-742.97
## Growth.rate ~ Humidity + SunHour + Windspeed + Temp.C
##
##             Df Sum of Sq    RSS     AIC
## - SunHour    1   0.03658 10.814 -744.15
## - Windspeed  1   0.06628 10.844 -743.49
## - Humidity   1   0.07961 10.857 -743.19
## <none>                   10.778 -742.97
## - Temp.C     1   0.60787 11.386 -731.69
##
## Step:  AIC=-744.15
## Growth.rate ~ Humidity + Windspeed + Temp.C
##
##             Df Sum of Sq    RSS     AIC
## - Windspeed  1   0.07197 10.886 -744.54
## <none>                   10.814 -744.15
## - Humidity   1   0.15266 10.967 -742.76
## - Temp.C     1   0.58306 11.398 -733.44
##
## Step:  AIC=-744.54
## Growth.rate ~ Humidity + Temp.C
##
##            Df Sum of Sq    RSS     AIC
## <none>                  10.886 -744.54
## - Humidity  1   0.13831 11.025 -743.49
## - Temp.C    1   0.51955 11.406 -735.26
```

```
##
## Call:
## lm(formula = Growth.rate ~ Humidity + Temp.C, data = df4)
##
## Coefficients:
## (Intercept)     Humidity       Temp.C
##    0.878967    -0.001572    -0.004890
```

```
##AIC shows that growth rate vs temperature + humidity could be a good lm model
```