Cherry Blossom Prediction

Exploratory Data Analysis (EDA)

Alex Yinan Guo

Contents

Setup

Load packages and read in data sets.

```
library(GGally)
library(rjson)
library(tidyverse)
library(lubridate)
library(cowplot)
library(ggpubr)
library(knitr)
library(mice)
library(mctest)
```

```
library(lme4)
library(quantreg)
library(AER)
library(MASS)
library(dplyr)
library(reshape2)

df <- read.csv(file = '../data/processed/clean_data.csv')
df_non_impute <- read.csv(file = '../data/processed/processed_data.csv')
```

Data Summaries

Before Imputation

```
summary(df_non_impute)
```

```
##       id              location           country             city
##  Length:7028        Length:7028        Length:7028        Length:7028
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##       lat              long              alt               year
##  Min.   :24.34   Min.   :-77.04   Min.   :   0.00   Min.   : 812
##  1st Qu.:34.69   1st Qu.:131.62   1st Qu.:   6.30   1st Qu.:1964
##  Median :35.44   Median :135.68   Median :  30.00   Median :1983
##  Mean   :36.73   Mean   :125.80   Mean   :  75.46   Mean   :1935
##  3rd Qu.:38.89   3rd Qu.:139.75   3rd Qu.:  44.00   3rd Qu.:1999
##  Max.   :47.57   Max.   :145.59   Max.   :1390.00   Max.   :2021
##
##   bloom_date          bloom_doy          tmax             tmin
##  Length:7028        Min.   :  8.0   Min.   :-1.231   Min.   :-6.892
##  Class :character   1st Qu.: 94.0   1st Qu.:15.181   1st Qu.: 6.601
##  Mode  :character   Median :100.0   Median :19.071   Median :10.235
##                     Mean   :101.3   Mean   :17.957   Mean   : 9.665
##                     3rd Qu.:110.0   3rd Qu.:20.593   3rd Qu.:12.037
##                     Max.   :160.0   Max.   :32.300   Max.   :26.500
##                                     NA's   :1390     NA's   :1375
##       prcp          agdd_winter      tmax_winter      tmin_winter
##  Min.   :  0.300   Min.   :   0.0   Min.   :-5.361   Min.   :-19.2316
##  1st Qu.:  3.403   1st Qu.: 110.7   1st Qu.: 4.226   1st Qu.: -3.1965
##  Median :  4.649   Median : 280.3   Median : 8.774   Median :  0.3411
##  Mean   :  6.210   Mean   : 320.8   Mean   : 7.800   Mean   : -0.2942
##  3rd Qu.:  6.927   3rd Qu.: 462.9   3rd Qu.:10.871   3rd Qu.:  2.2145
##  Max.   :122.900   Max.   :1713.3   Max.   :24.032   Max.   : 18.4540
##  NA's   :1376      NA's   :1486     NA's   :1494     NA's   :1494
##   prcp_winter     co2_percapita      co2_emission
##  Min.   :  0.00   Min.   : 0.0003   Min.   :1.099e+04
##  1st Qu.: 16.30   1st Qu.: 5.5023   1st Qu.:3.589e+08
##  Median : 26.37   Median : 7.9740   Median :9.143e+08
##  Mean   : 36.59   Mean   : 7.3841   Mean   :8.547e+08
```

```
##  3rd Qu.: 47.28   3rd Qu.: 9.5871   3rd Qu.:1.223e+09
##  Max.   :525.25   Max.   :22.2365   Max.   :6.135e+09
##  NA's   :1494     NA's   :749       NA's   :749
```

```
apply(df_non_impute, 2, function(x) length(unique(x)))
```

```
##            id      location       country          city           lat
##           109           105             4           105           112
##          long           alt          year    bloom_date     bloom_doy
##           110           106           837          3359           127
##          tmax          tmin          prcp   agdd_winter   tmax_winter
##          5053          5080          4825          4129          5028
##   tmin_winter   prcp_winter co2_percapita  co2_emission
##          5034          4965           403           402
```

Data After Imputation

```
head(df)
```

```
##   country         city year bloom_doy      lat     long alt     tmax     tmin
## 1   Japan Ishigakijima 1962        51 24.33667 124.1644 5.7 23.80000 21.05040
## 2   Japan Ishigakijima 1963        46 24.33667 124.1644 5.7 23.80000 21.05040
## 3   Japan Ishigakijima 1967        44 24.33667 124.1644 5.7 23.80000 21.05040
## 4   Japan Ishigakijima 1971        41 24.33667 124.1644 5.7 23.80000 21.05040
## 5   Japan Ishigakijima 1972        51 24.33667 124.1644 5.7 23.80000 21.05040
## 6   Japan Ishigakijima 1973        37 24.33667 124.1644 5.7 26.78116 21.81633
##       prcp agdd_winter tmax_winter tmin_winter prcp_winter co2_percapita
## 1 21.97045       22.35    17.80000    15.01875    98.40000         3.0680
## 2 21.97045       22.35    17.80000    15.01875    98.40000         3.3700
## 3 21.97045       22.35    17.80000    15.01875    98.40000         4.8532
## 4 21.97045       22.35    17.80000    15.01875    98.40000         7.4844
## 5 21.97045       22.35    17.80000    15.01875    98.40000         7.8922
## 6  6.74517      649.50    22.93487    17.36111    48.49867         8.3473
```

```
tail(df)
```

```
##          country    city year bloom_doy     lat     long alt     tmax     tmin
## 6177 Switzerland Liestal 2016        97 47.4814 7.730519 350 17.77065 7.578804
## 6178 Switzerland Liestal 2017        85 47.4814 7.730519 350 16.61808 6.177260
## 6179 Switzerland Liestal 2018        98 47.4814 7.730519 350 17.63562 7.103836
## 6180 Switzerland Liestal 2019        86 47.4814 7.730519 350 17.05753 6.315890
## 6181 Switzerland Liestal 2020        77 47.4814 7.730519 350 17.61284 6.630055
## 6182 Switzerland Liestal 2021        87 47.4814 7.730519 350 15.70822 5.805753
##          prcp agdd_winter tmax_winter tmin_winter prcp_winter co2_percapita
## 6177 1.685870      141.35    6.819355   1.5451613    23.83871      4.677000
## 6178 2.740822      235.55    5.344240  -1.6786098    15.42166      4.515500
## 6179 2.196438      337.85    6.214555   0.5637097    27.30415      4.324300
## 6180 2.327945      374.60    7.906413   0.2789939    20.28763      4.276600
## 6181 2.223224      475.65    9.360289   1.1466073    21.00222      3.731900
## 6182 3.298630      372.05    7.178418   0.4391705    33.85637      2.381433
```

```
summary(df)
```

```
##    country              city                year        bloom_doy
##  Length:6182        Length:6182        Min.   :1950   Min.   :  8.0
##  Class :character   Class :character   1st Qu.:1972   1st Qu.: 93.0
```

```
##  Mode  :character   Mode  :character   Median :1987   Median : 99.0
##                                        Mean   :1987   Mean   :100.9
##                                        3rd Qu.:2002   3rd Qu.:110.0
##                                        Max.   :2021   Max.   :160.0
##       lat             long             alt              tmax
##  Min.   :24.34   Min.   :-77.04   Min.   :   0.0   Min.   :-1.231
##  1st Qu.:34.34   1st Qu.:130.98   1st Qu.:   5.7   1st Qu.:15.445
##  Median :35.54   Median :136.19   Median :  26.5   Median :19.180
##  Mean   :36.84   Mean   :126.61   Mean   :  77.2   Mean   :18.178
##  3rd Qu.:39.70   3rd Qu.:140.10   3rd Qu.:  51.1   3rd Qu.:20.695
##  Max.   :47.57   Max.   :145.59   Max.   :1390.0   Max.   :32.300
##       tmin             prcp          agdd_winter      tmax_winter
##  Min.   :-6.892   Min.   :  0.300   Min.   :   0.0   Min.   :-5.361
##  1st Qu.: 6.732   1st Qu.:  3.472   1st Qu.: 107.2   1st Qu.: 4.469
##  Median :10.368   Median :  4.799   Median : 277.6   Median : 8.940
##  Mean   : 9.929   Mean   :  6.389   Mean   : 320.1   Mean   : 8.089
##  3rd Qu.:12.197   3rd Qu.:  7.179   3rd Qu.: 459.1   3rd Qu.:11.041
##  Max.   :26.500   Max.   :122.900   Max.   :1713.3   Max.   :24.032
##   tmin_winter        prcp_winter     co2_percapita
##  Min.   :-19.2316   Min.   :  0.00   Min.   : 1.238
##  1st Qu.: -3.0480   1st Qu.: 16.64   1st Qu.: 5.963
##  Median :  0.4671   Median : 27.20   Median : 7.980
##  Mean   :  0.0480   Mean   : 37.12   Mean   : 7.504
##  3rd Qu.:  2.3828   3rd Qu.: 47.83   3rd Qu.: 9.736
##  Max.   : 18.4540   Max.   :525.25   Max.   :22.236
```

```
apply(df, 2, function(x) length(unique(x)))
```

```
##       country          city          year     bloom_doy           lat
##             4           105            72           127           112
##          long           alt          tmax          tmin          prcp
##           110           106          5417          5429          5177
##   agdd_winter   tmax_winter   tmin_winter   prcp_winter co2_percapita
##          4592          5516          5522          5453           272
```

```
unique(df$country)
```

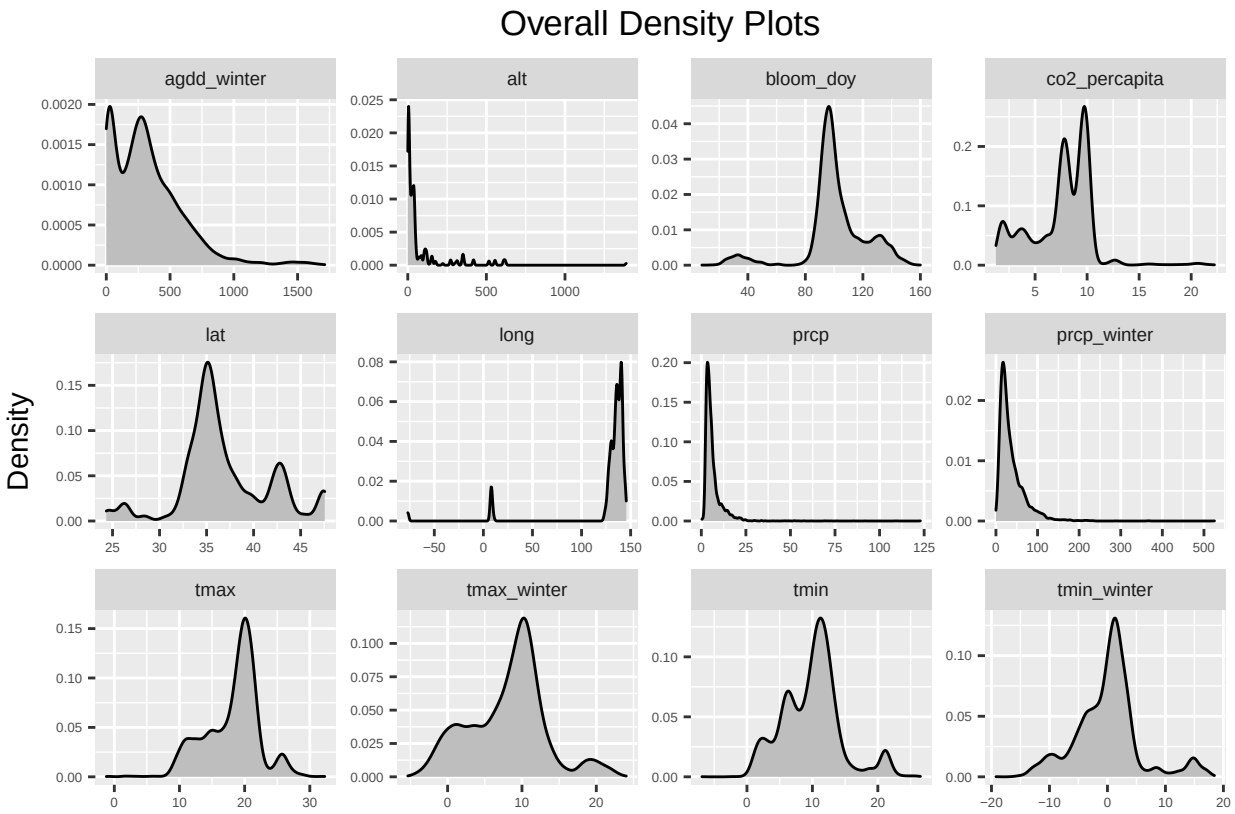
```
## [1] "Japan"         "Switzerland"   "South Korea"   "United States"
```

Density Plots

Overall Density Plots

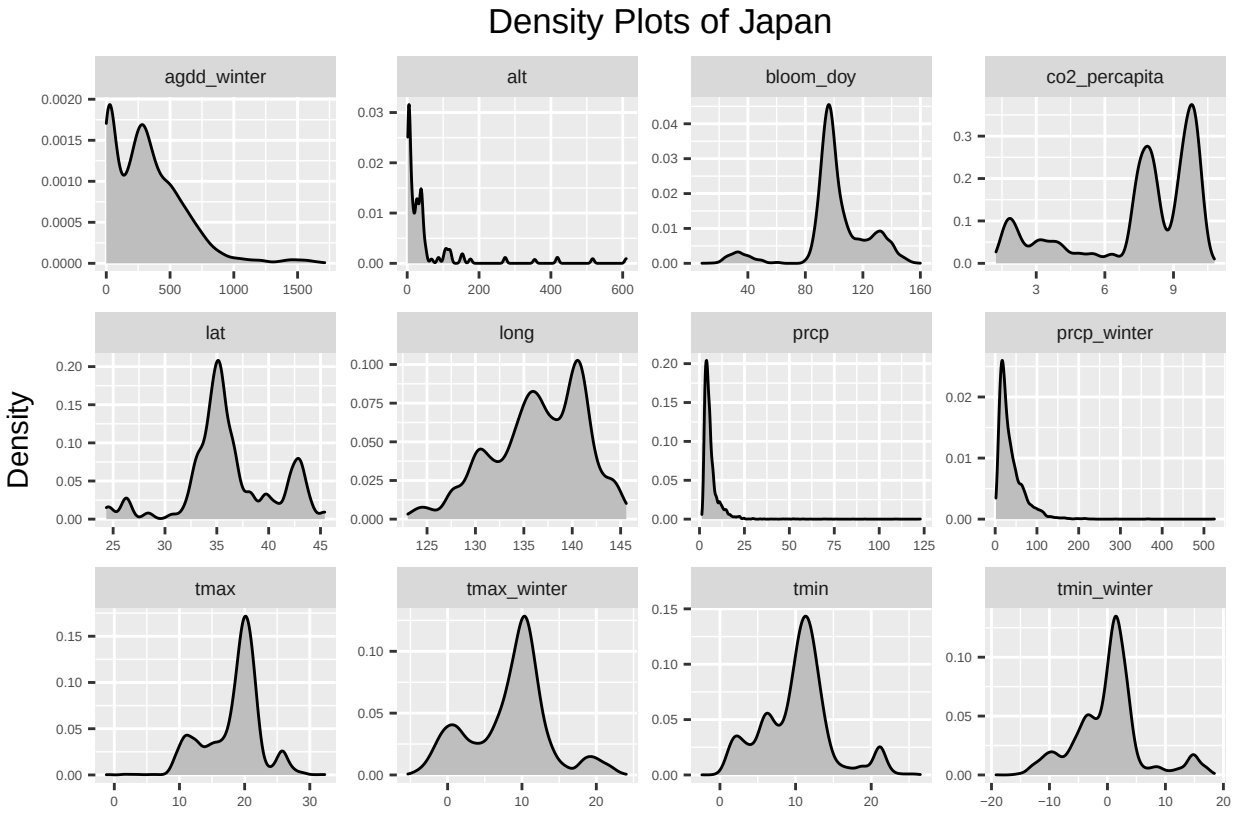
```
df %>%
  select_if(is.numeric) %>%
  dplyr::select(-year) %>%
  pivot_longer(everything()) %>%
  ggplot(aes(x=value)) +
  geom_density(fill='grey') +
  facet_wrap(~name, scales='free') +
  theme(strip.text = element_text(size=7),
        axis.text.x = element_text(size=5),
        axis.text.y = element_text(size=5),
        plot.title = element_text(hjust = 0.5)) +
  labs(x ="",
```

```
       y = "Density",
       title = 'Overall Density Plots')
```



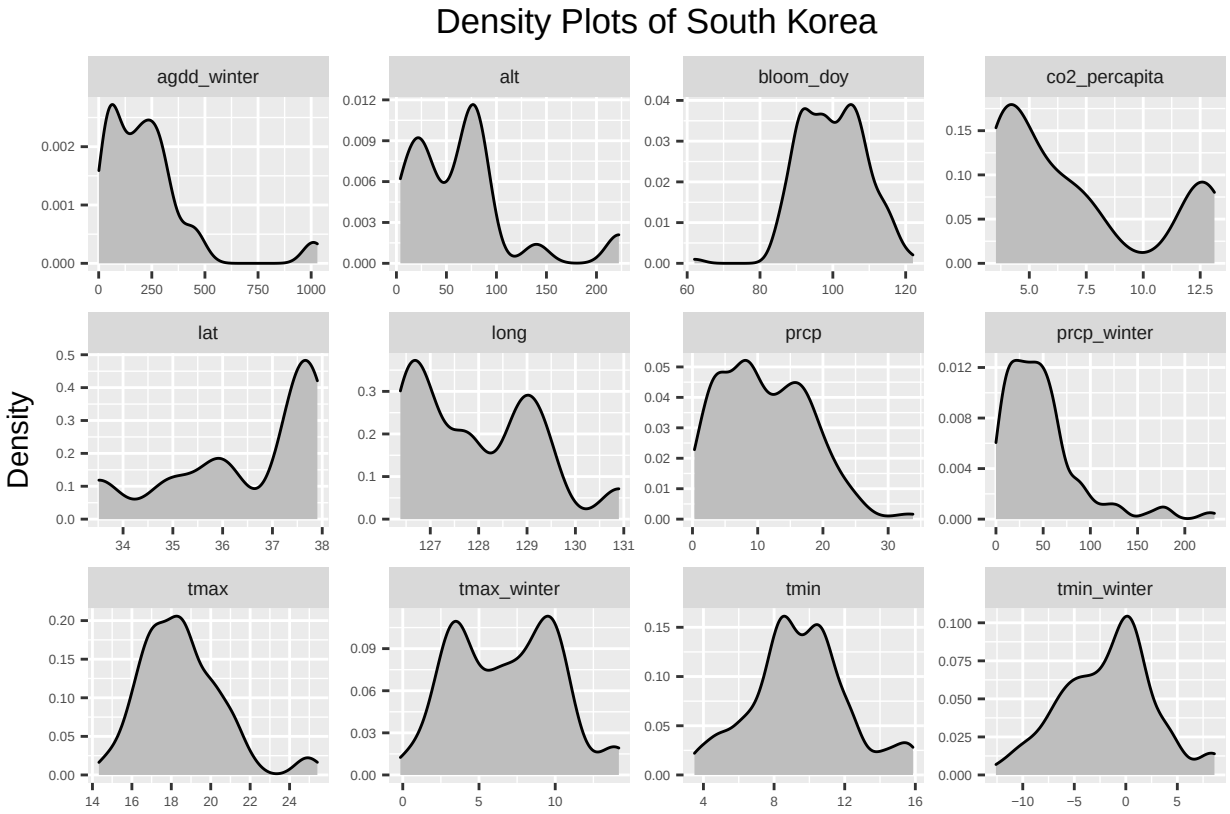
Density Plots of Japan

```
df %>%
  filter(country=='Japan') %>%
  select_if(is.numeric) %>%
  dplyr::select(-year) %>%
  pivot_longer(everything()) %>%
  ggplot(aes(x=value)) +
  geom_density(fill='grey') +
  facet_wrap(~name, scales='free') +
  theme(strip.text = element_text(size=7),
        axis.text.x = element_text(size=5),
        axis.text.y = element_text(size=5),
        plot.title = element_text(hjust = 0.5)) +
  labs(x ="",
       y = "Density",
       title = 'Density Plots of Japan')
```



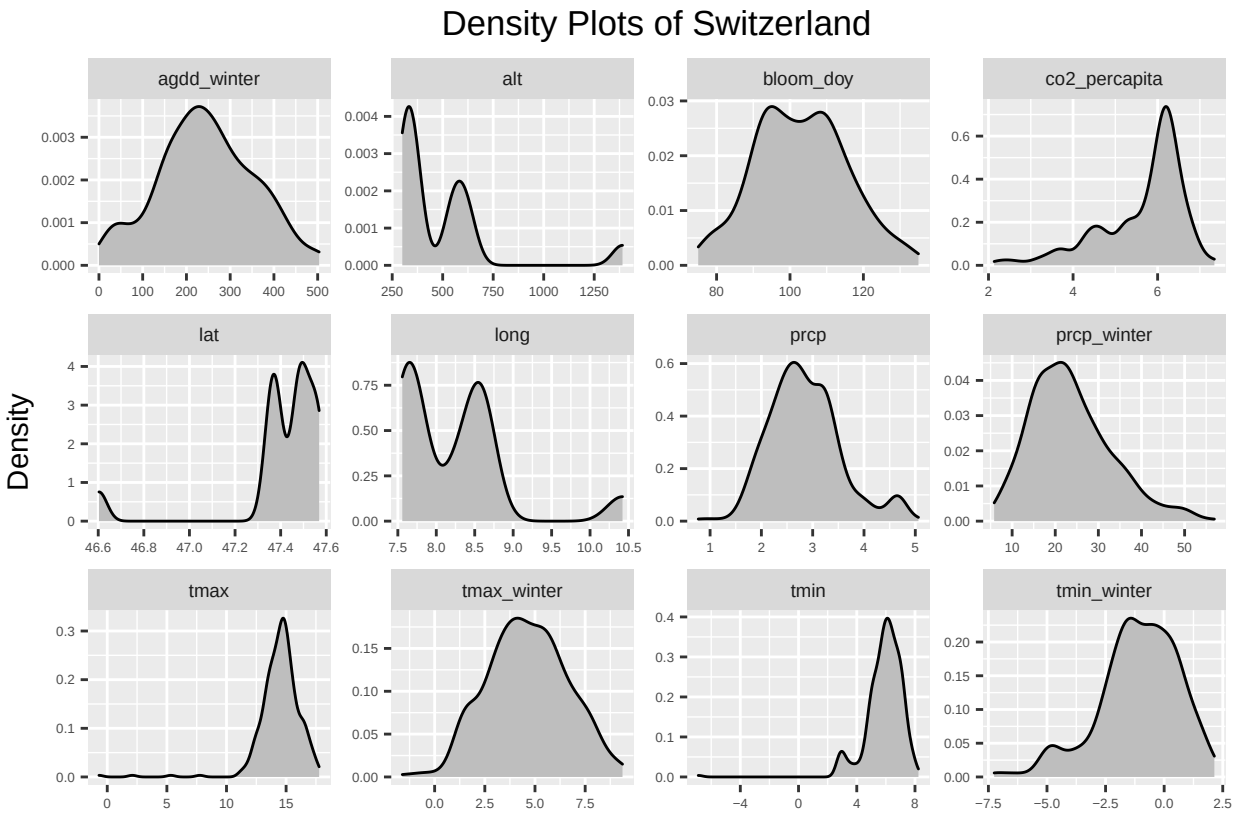
Density Plots of South Korea

```
df %>%
  filter(country=='South Korea') %>%
  select_if(is.numeric) %>%
  dplyr::select(-year) %>%
  pivot_longer(everything()) %>%
  ggplot(aes(x=value)) +
  geom_density(fill='grey') +
  facet_wrap(~name, scales='free') +
  theme(strip.text = element_text(size=7),
        axis.text.x = element_text(size=5),
        axis.text.y = element_text(size=5),
        plot.title = element_text(hjust = 0.5)) +
  labs(x ="",
       y = "Density",
       title = 'Density Plots of South Korea')
```



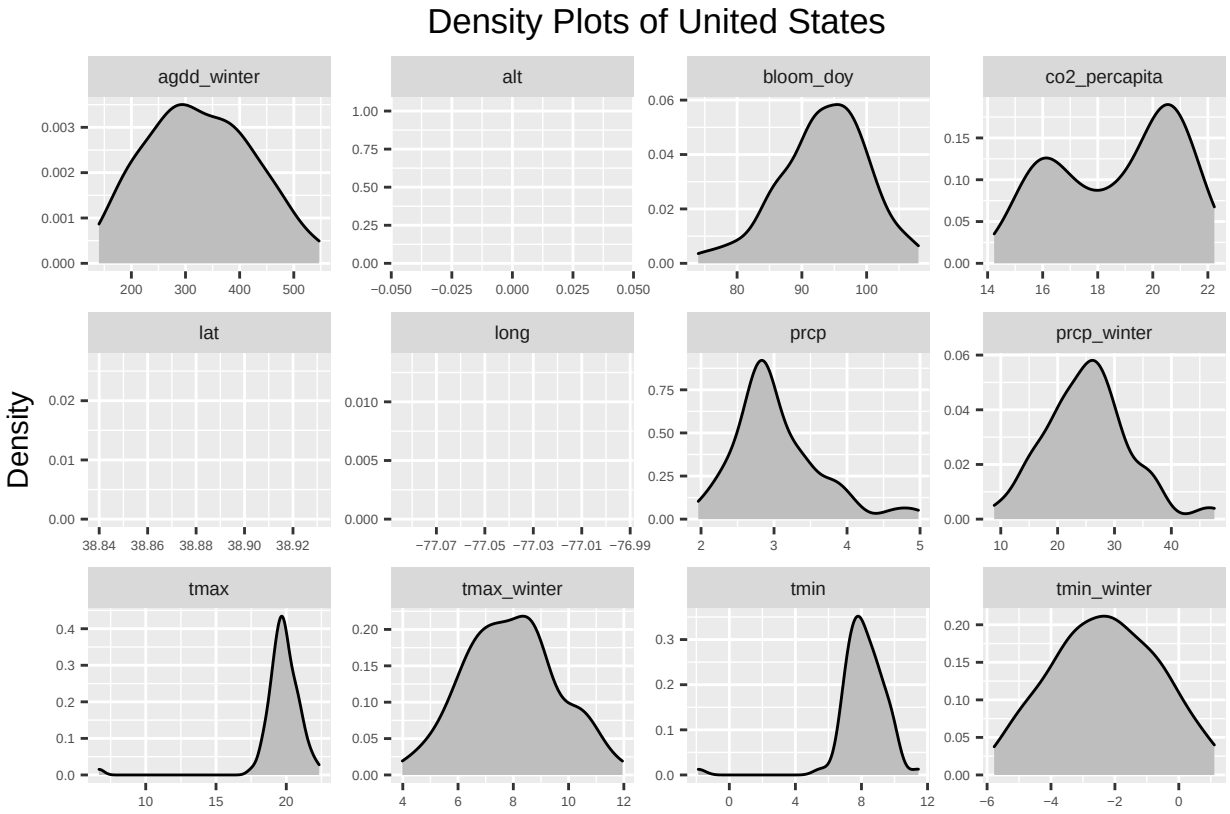
Density Plots of Switzerland

```
df %>%
  filter(country=='Switzerland') %>%
  select_if(is.numeric) %>%
  dplyr::select(-year) %>%
  pivot_longer(everything()) %>%
  ggplot(aes(x=value)) +
  geom_density(fill='grey') +
  facet_wrap(~name, scales='free') +
  theme(strip.text = element_text(size=7),
        axis.text.x = element_text(size=5),
        axis.text.y = element_text(size=5),
        plot.title = element_text(hjust = 0.5)) +
  labs(x ="",
       y = "Density",
       title = 'Density Plots of Switzerland')
```



Density Plots of United States

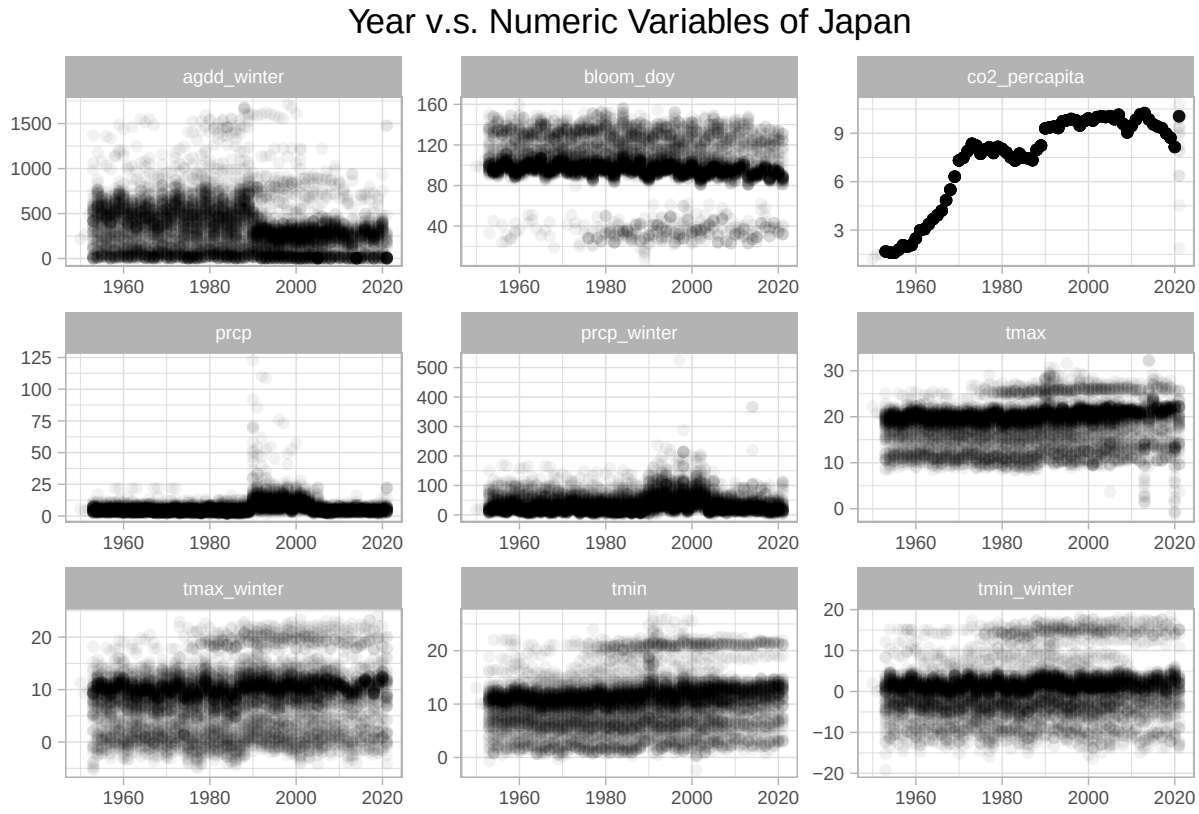
```
df %>%
  filter(country=='United States') %>%
  select_if(is.numeric) %>%
  dplyr::select(-year) %>%
  pivot_longer(everything()) %>%
  ggplot(aes(x=value)) +
  geom_density(fill='grey') +
  facet_wrap(~name, scales='free') +
  theme(strip.text = element_text(size=7),
        axis.text.x = element_text(size=5),
        axis.text.y = element_text(size=5),
        plot.title = element_text(hjust = 0.5)) +
  labs(x ="",
       y = "Density",
       title = 'Density Plots of United States')
```



Year v.s. Numeric Variables

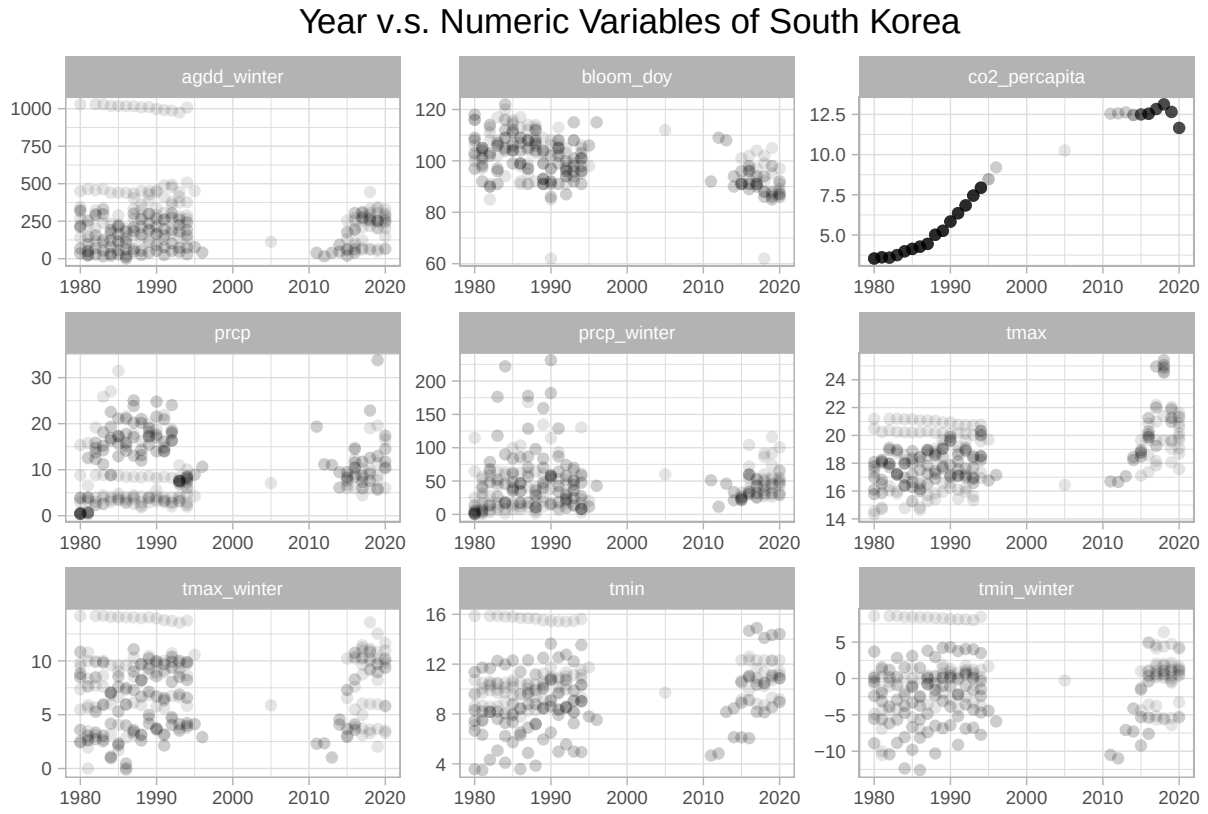
Year v.s. Numeric Variables of Japan

```
par(mfrow=c(3, 4))

df %>%
  filter(country=='Japan') %>%
  select_if(is.numeric) %>%
  pivot_longer(-c(year, lat, long, alt)) %>%
  ggplot(aes(x=year, y=value)) +
  geom_point(alpha = 0.05) +
  facet_wrap(~name, scales='free') +
  labs(x ="",
       y = "",
       title="Year v.s. Numeric Variables of Japan") +
  theme_light() +
  theme(strip.text = element_text(size=7),
        axis.text.x = element_text(size=7),
        axis.text.y = element_text(size=7),
        plot.title = element_text(hjust = 0.5))
```



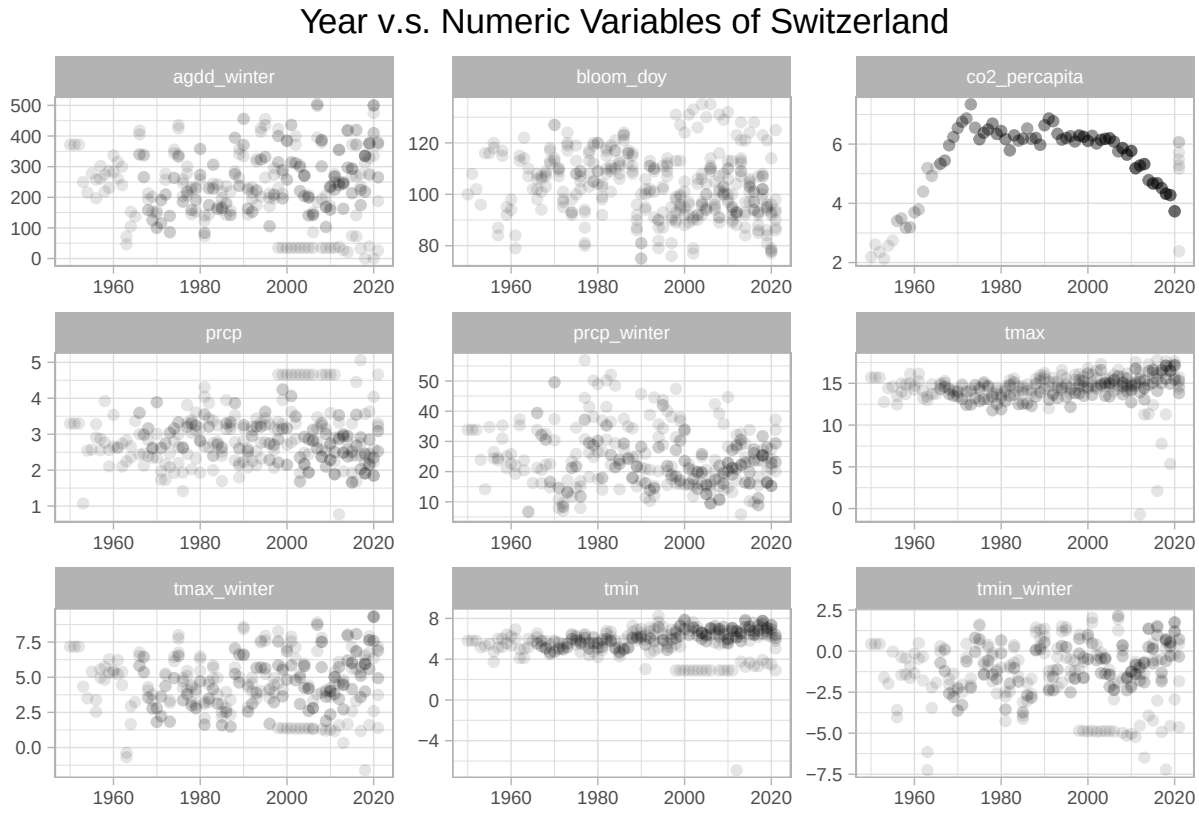
Year v.s. Numeric Variables of South Korea

```
par(mfrow=c(3, 4))

df %>%
  filter(country=='South Korea') %>%
  select_if(is.numeric) %>%
  pivot_longer(-c(year, lat, long, alt)) %>%
  ggplot(aes(x=year, y=value)) +
  geom_point(alpha = 0.1) +
  facet_wrap(~name, scales='free') +
  labs(x ="",
       y = "",
       title="Year v.s. Numeric Variables of South Korea") +
  theme_light() +
  theme(strip.text = element_text(size=7),
        axis.text.x = element_text(size=7),
        axis.text.y = element_text(size=7),
        plot.title = element_text(hjust = 0.5))
```



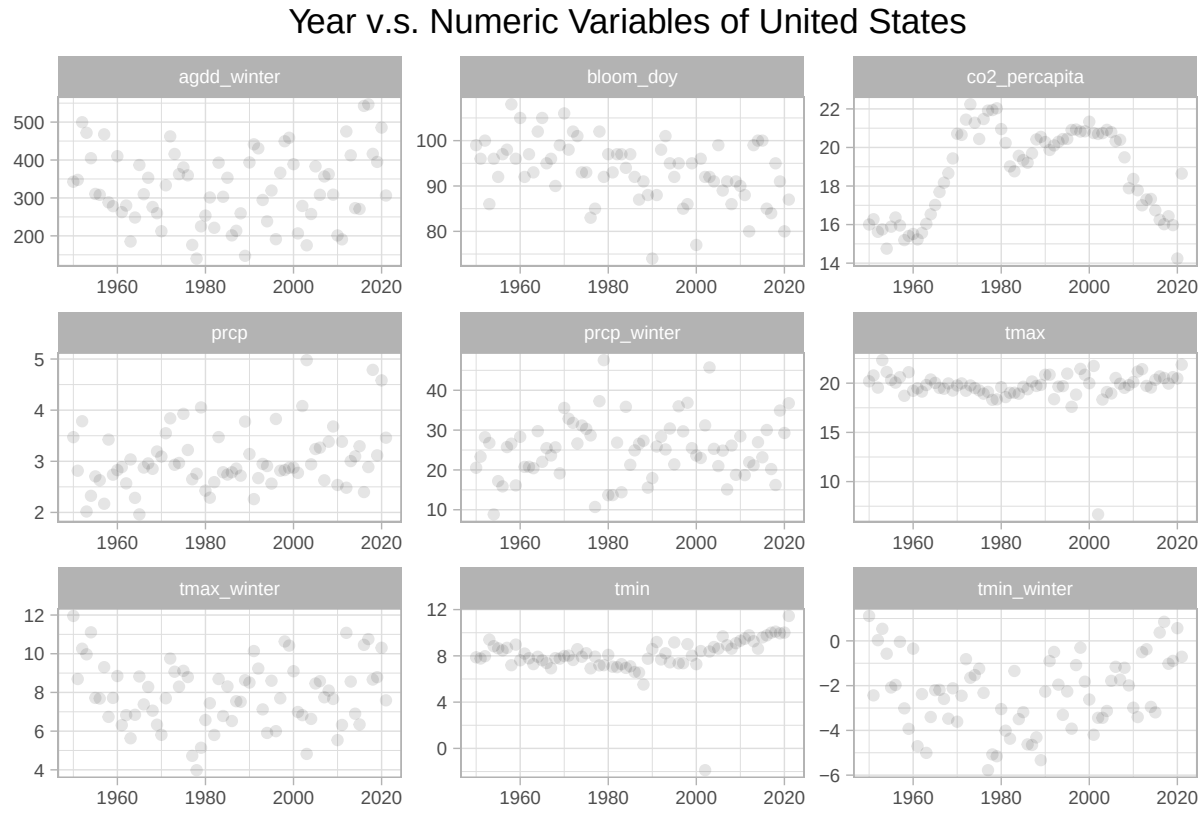
Year v.s. Numeric Variables of Switzerland

```
par(mfrow=c(3, 4))

df %>%
  filter(country=='Switzerland') %>%
  select_if(is.numeric) %>%
  pivot_longer(-c(year, lat, long, alt)) %>%
  ggplot(aes(x=year, y=value)) +
  geom_point(alpha = 0.1) +
  facet_wrap(~name, scales='free') +
  labs(x ="",
       y = "",
       title="Year v.s. Numeric Variables of Switzerland") +
  theme_light() +
  theme(strip.text = element_text(size=7),
        axis.text.x = element_text(size=7),
        axis.text.y = element_text(size=7),
        plot.title = element_text(hjust = 0.5))
```



Year v.s. Numeric Variables of United States

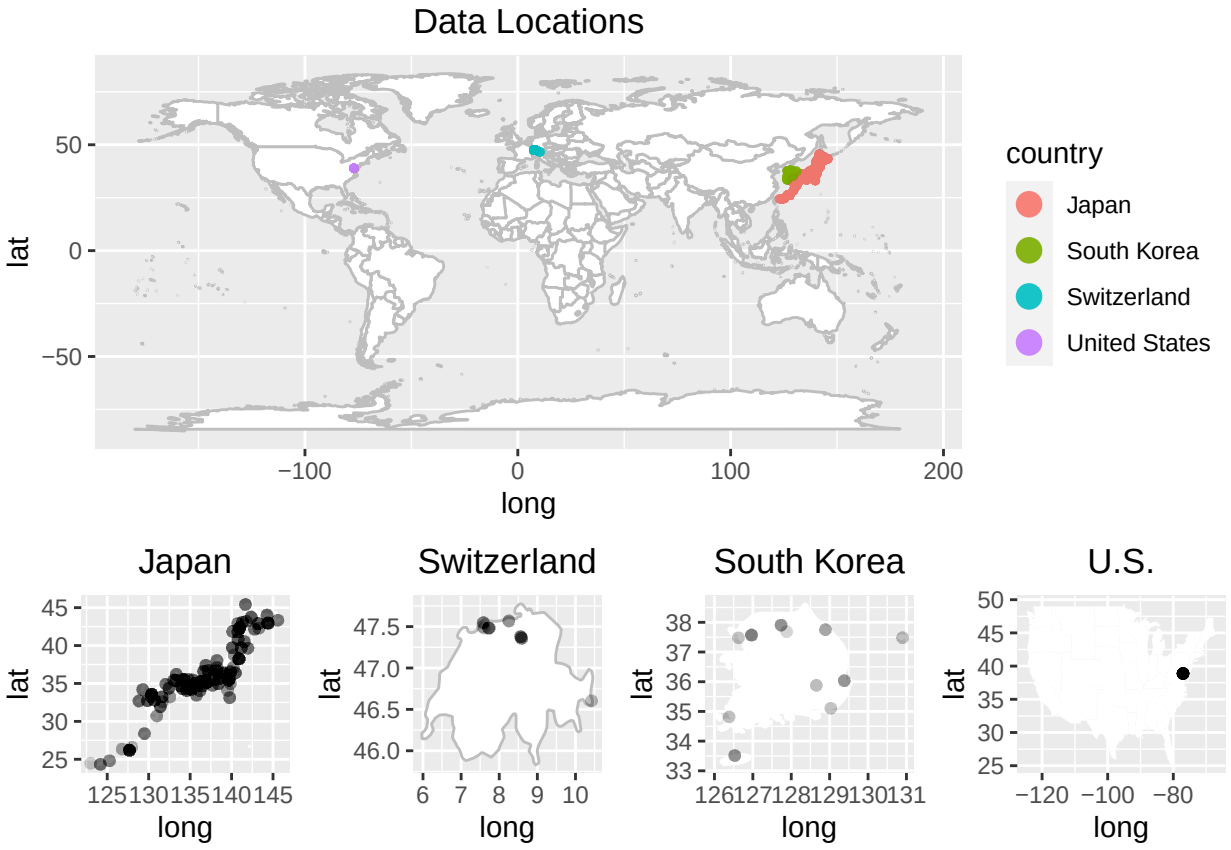
```
par(mfrow=c(3, 4))

df %>%
  filter(country=='United States') %>%
  select_if(is.numeric) %>%
  pivot_longer(-c(year, lat, long, alt)) %>%
  ggplot(aes(x=year, y=value)) +
  geom_point(alpha = 0.1) +
  facet_wrap(~name, scales='free') +
  labs(x ="",
       y = "",
       title="Year v.s. Numeric Variables of United States") +
  theme_light() +
  theme(strip.text = element_text(size=7),
        axis.text.x = element_text(size=7),
        axis.text.y = element_text(size=7),
        plot.title = element_text(hjust = 0.5))
```



Visualization of Data Locations

```
world <- map_data("world")
all_plot <- ggplot() +
  geom_map(
    data = world, map = world,
    aes(long, lat, map_id = region, fill = region),
    color = "gray", fill = "white") +
  geom_point(
    data = df,
    aes(long, lat, color=country),
    alpha=0.1, size=1) +
  guides(colour = guide_legend(override.aes = list(alpha=0.9,size=4))) +
  labs(title='Data Locations') +
  theme(plot.title = element_text(hjust = 0.5))

jp <- map_data('world', 'Japan')
jp_plot <- ggplot() +
  geom_map(
    data = jp, map = jp,
    aes(long, lat, map_id = region),
    color = "white", fill = "white"
  ) +
  geom_point(
```

```
    data = df %>% filter(country=='Japan'),
    aes(long, lat), alpha=0.01) +
  ggtitle("Japan") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))

swiss <- map_data('world', 'Switzerland')
swiss_plot <- ggplot() +
  geom_map(
    data = swiss, map = swiss,
    aes(long, lat, map_id = region),
    color = "gray", fill = "white"
  ) +
  geom_point(
    data = df %>% filter(country=='Switzerland'),
    aes(long, lat), alpha=0.01) +
  ggtitle("Switzerland") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))

sk <- map_data('world', 'South Korea')
sk_plot <- ggplot() +
  geom_map(
    data = sk, map = sk,
    aes(long, lat, map_id = region),
    color = "white", fill = "white"
  ) +
  geom_point(
    data = df %>% filter(country=='South Korea'),
    aes(long, lat), alpha=0.01) +
  ggtitle("South Korea") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))

usa <- map_data("state")
usa_plot <- ggplot() +
  geom_map(
    data = usa, map = usa,
    aes(long, lat, map_id = region),
    fill = "white", size = 0.1
  ) +
  geom_point(
    data = df %>% filter(country=='United States'),
    aes(long, lat), alpha=0.1) +
  ggtitle("U.S.") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))

individual_plots <- ggarrange(jp_plot, swiss_plot, sk_plot, usa_plot,
          ncol = 4, nrow = 1)

ggarrange(all_plot, individual_plots,
          ncol = 1, nrow = 2, heights = c(2.5,1.5))
```



Basic Linear Models

```
df <- read.csv("../data/processed/clean_data.csv")
summary(df)
```

```
##    country              city                year        bloom_doy
##  Length:6182        Length:6182        Min.   :1950   Min.   :  8.0
##  Class :character   Class :character   1st Qu.:1972   1st Qu.: 93.0
##  Mode  :character   Mode  :character   Median :1987   Median : 99.0
##                                        Mean   :1987   Mean   :100.9
##                                        3rd Qu.:2002   3rd Qu.:110.0
##                                        Max.   :2021   Max.   :160.0
##       lat             long              alt              tmax
##  Min.   :24.34   Min.   :-77.04   Min.   :   0.0   Min.   :-1.231
##  1st Qu.:34.34   1st Qu.:130.98   1st Qu.:   5.7   1st Qu.:15.445
##  Median :35.54   Median :136.19   Median :  26.5   Median :19.180
##  Mean   :36.84   Mean   :126.61   Mean   :  77.2   Mean   :18.178
##  3rd Qu.:39.70   3rd Qu.:140.10   3rd Qu.:  51.1   3rd Qu.:20.695
##  Max.   :47.57   Max.   :145.59   Max.   :1390.0   Max.   :32.300
##       tmin             prcp          agdd_winter      tmax_winter
##  Min.   :-6.892   Min.   :  0.300   Min.   :   0.0   Min.   :-5.361
##  1st Qu.: 6.732   1st Qu.:  3.472   1st Qu.: 107.2   1st Qu.: 4.469
##  Median :10.368   Median :  4.799   Median : 277.6   Median : 8.940
##  Mean   : 9.929   Mean   :  6.389   Mean   : 320.1   Mean   : 8.089
##  3rd Qu.:12.197   3rd Qu.:  7.179   3rd Qu.: 459.1   3rd Qu.:11.041
```

```
##  Max.   :26.500   Max.   :122.900   Max.   :1713.3   Max.   :24.032
##   tmin_winter        prcp_winter     co2_percapita
##  Min.   :-19.2316   Min.   :  0.00   Min.   : 1.238
##  1st Qu.: -3.0480   1st Qu.: 16.64   1st Qu.: 5.963
##  Median :  0.4671   Median : 27.20   Median : 7.980
##  Mean   :  0.0480   Mean   : 37.12   Mean   : 7.504
##  3rd Qu.:  2.3828   3rd Qu.: 47.83   3rd Qu.: 9.736
##  Max.   : 18.4540   Max.   :525.25   Max.   :22.236
```

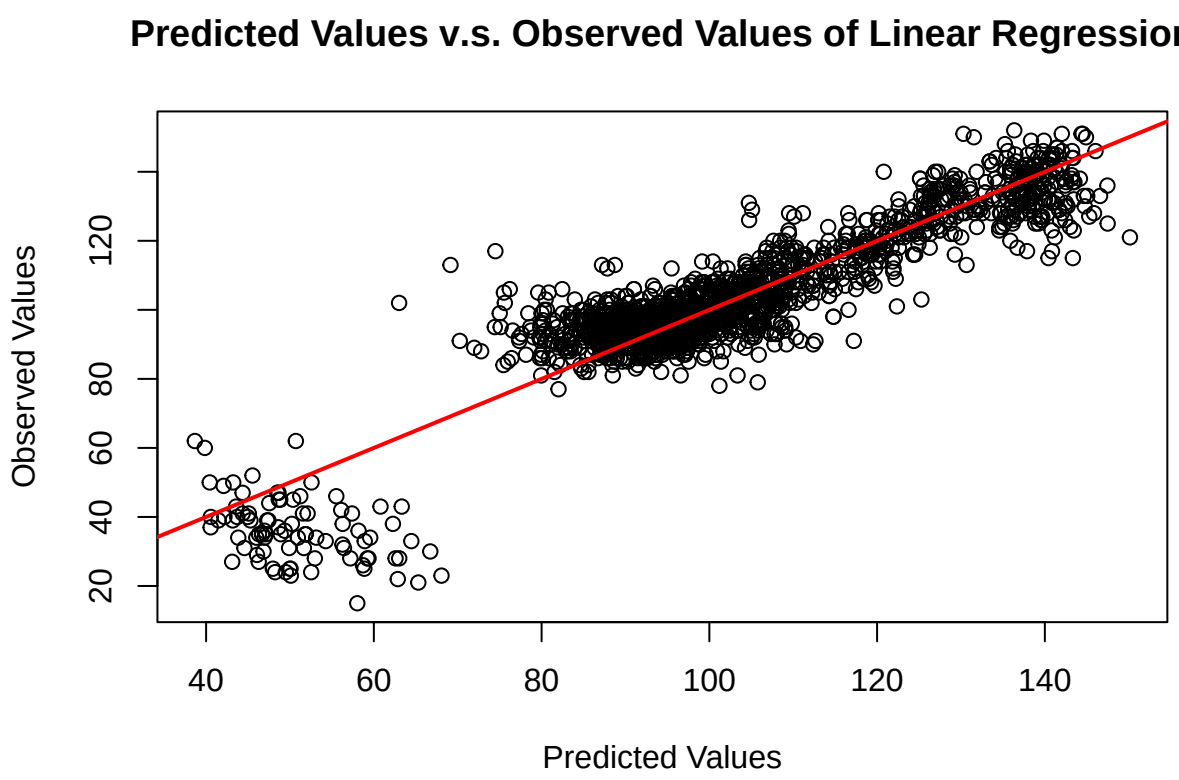
```
df <- df %>% na.omit
train <- df[sort(sample(nrow(df), nrow(df)*.7)),]
test <- df[-sort(sample(nrow(df), nrow(df)*.7)),]
```

Fit Linear Regression

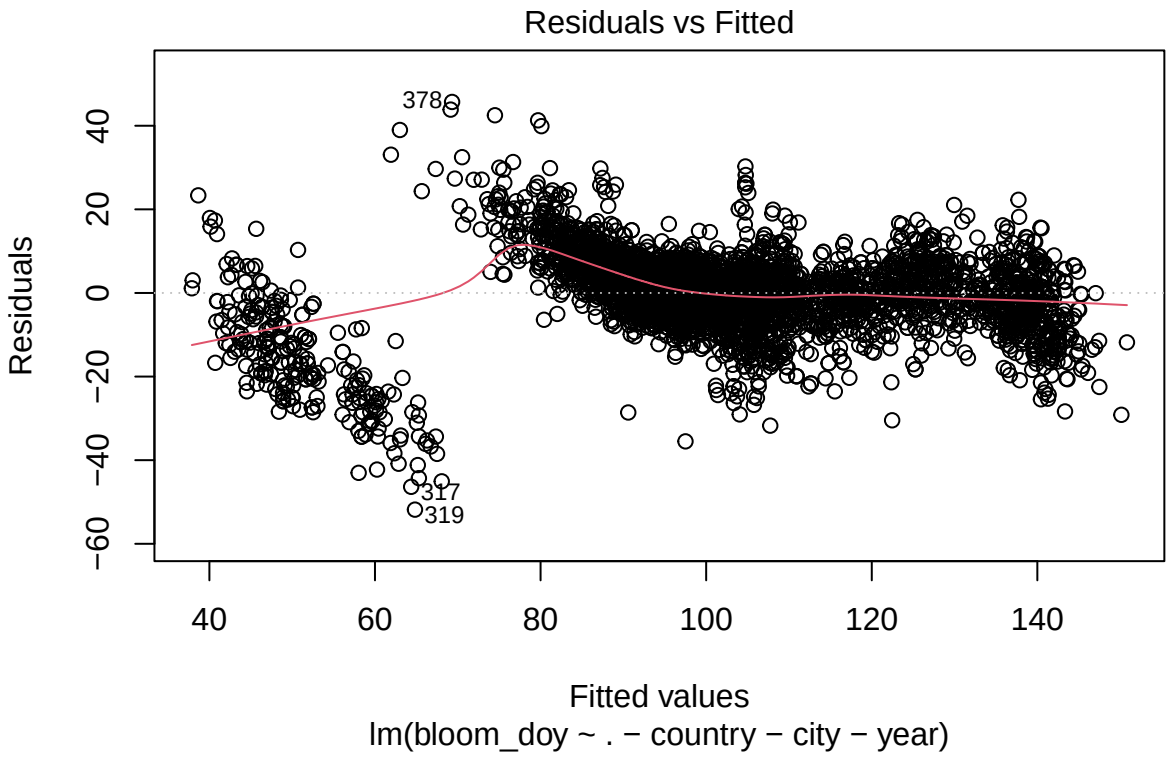
```
lm <- lm(bloom_doy~.-country-city-year, data = train)
summary(lm)
```

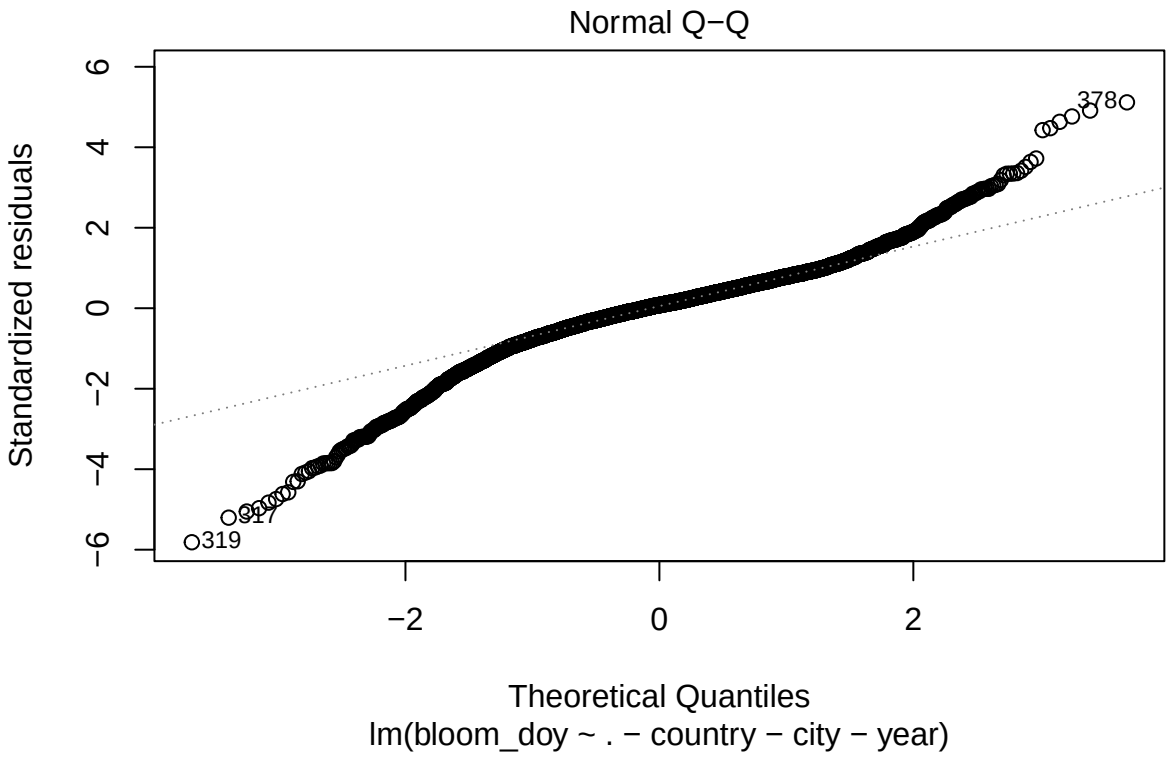
```
##
## Call:
## lm(formula = bloom_doy ~ . - country - city - year, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.829  -3.993   0.704   4.930  45.688
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   46.8151363  4.8828104   9.588  < 2e-16 ***
## lat            1.6304238  0.0869469  18.752  < 2e-16 ***
## long           0.1372419  0.0055343  24.798  < 2e-16 ***
## alt           -0.0089975  0.0010863  -8.283  < 2e-16 ***
## tmax          -0.6404403  0.1041418  -6.150 8.47e-10 ***
## tmin          -0.8974683  0.1619025  -5.543 3.15e-08 ***
## prcp           0.1393194  0.0317438   4.389 1.17e-05 ***
## agdd_winter    0.0077328  0.0008757   8.830  < 2e-16 ***
## tmax_winter   -0.6859738  0.1133954  -6.049 1.58e-09 ***
## tmin_winter   -0.9495250  0.1176218  -8.073 8.84e-16 ***
## prcp_winter   -0.0193137  0.0053406  -3.616 0.000302 ***
## co2_percapita  0.1177370  0.0529925   2.222 0.026351 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.938 on 4315 degrees of freedom
## Multiple R-squared:  0.8373, Adjusted R-squared:  0.8368
## F-statistic:  2018 on 11 and 4315 DF,  p-value: < 2.2e-16
```

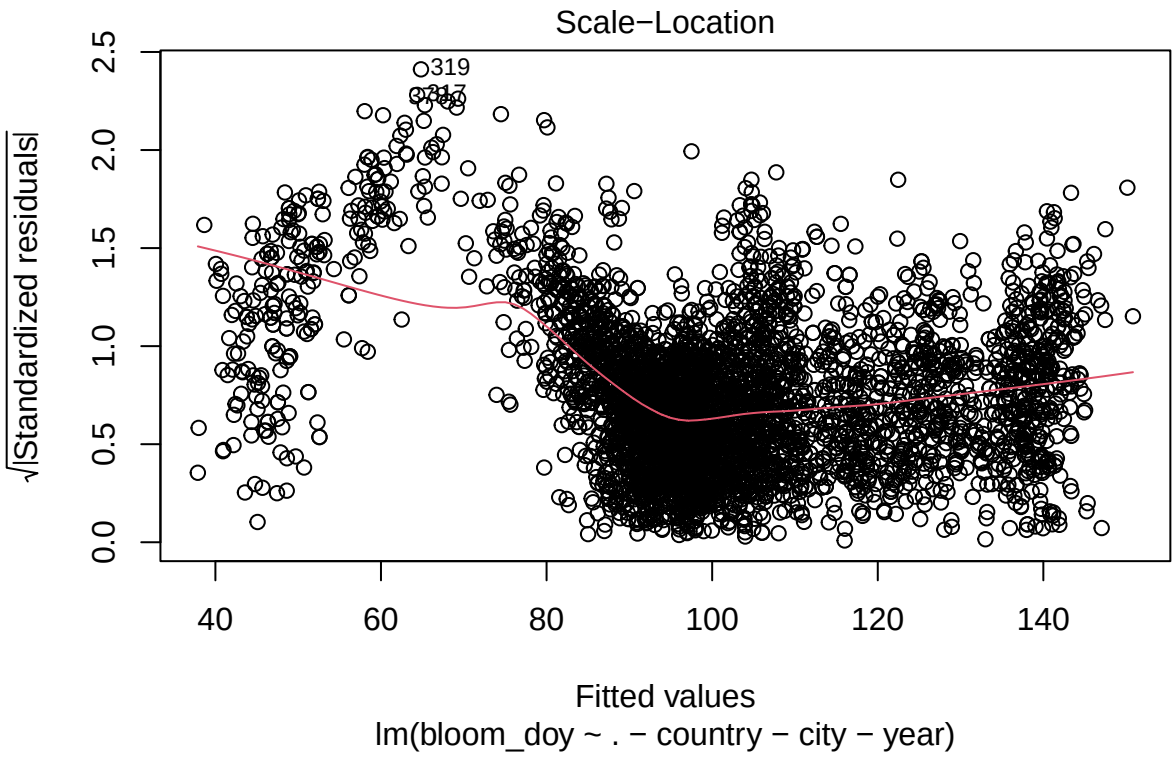
```
plot(predict(lm, test),
     test$bloom_doy,
     xlab = "Predicted Values",
     ylab = "Observed Values",
     main = 'Predicted Values v.s. Observed Values of Linear Regression')
abline(a=0, b=1, lib = 1, col = "red", lwd = 2)
```

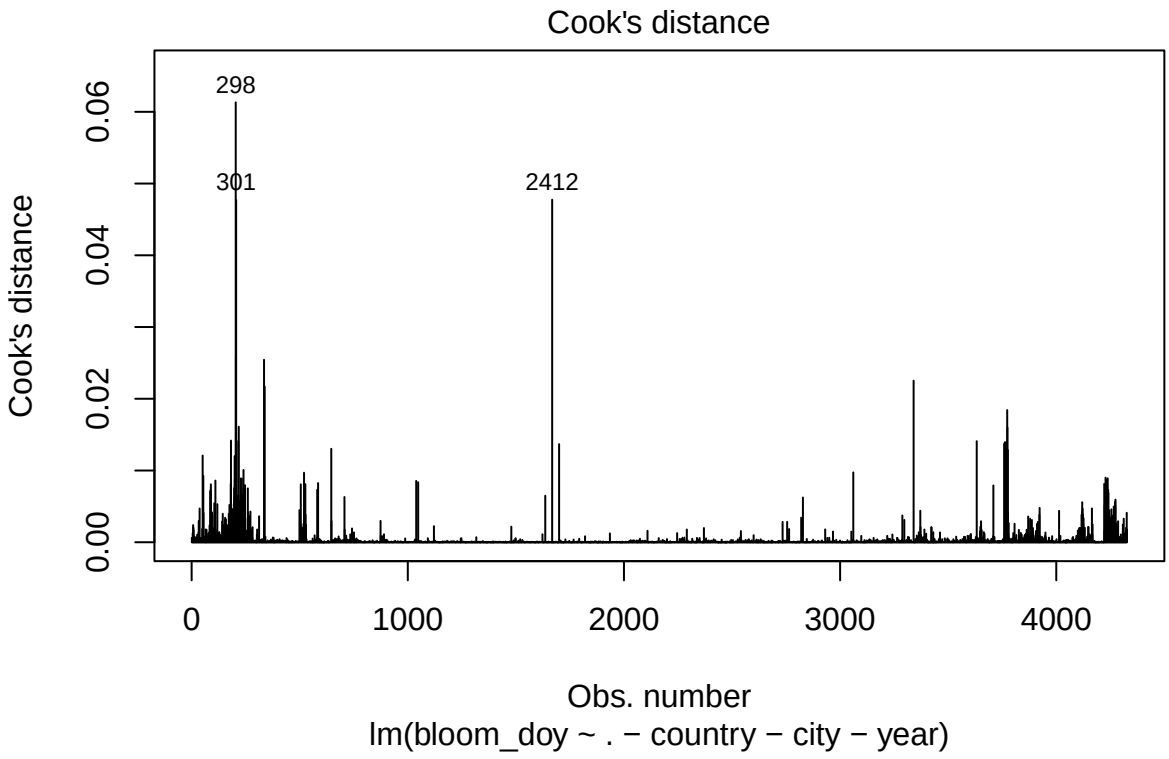


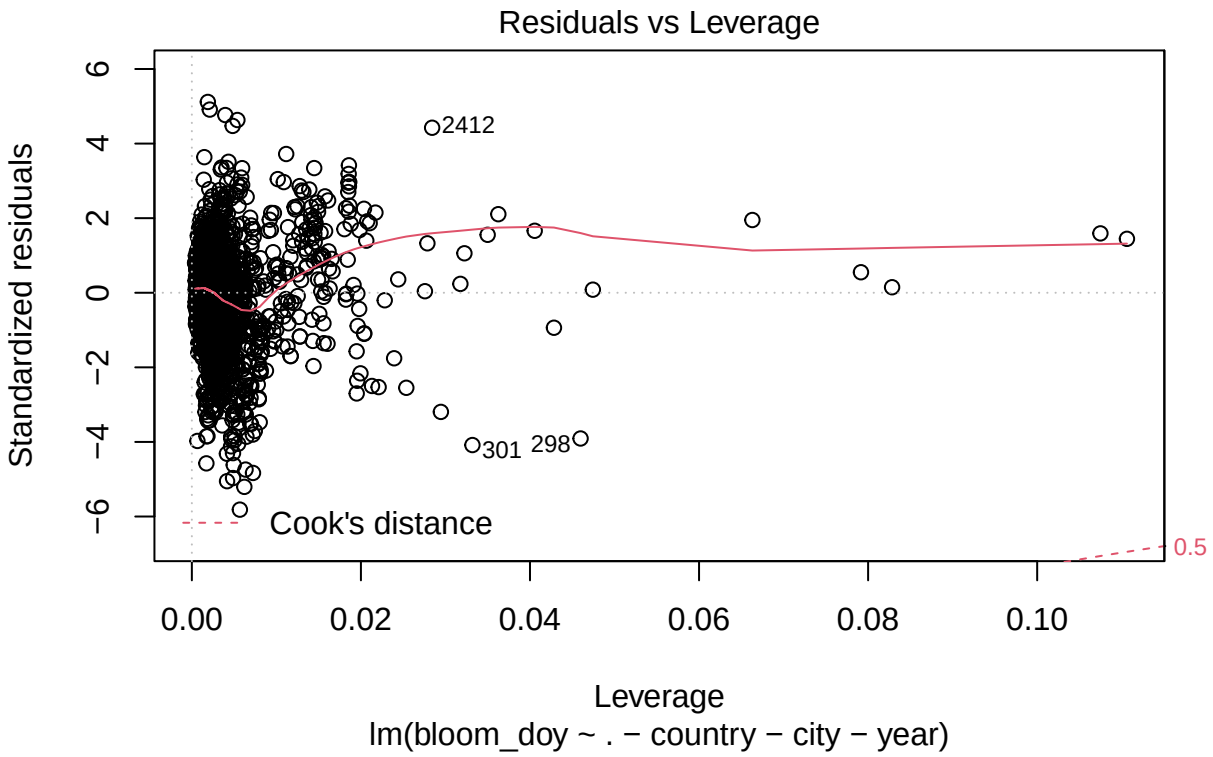

```
plot(lm, which = c(1,2,3,4,5))
```


Residuals vs Fitted
Residuals
Fitted values
lm(bloom_doy ~ . − country − city − year)
378
317
319
40
20
0
−20
−40
−60
40
60
80
100
120
140


Normal Q−Q
Standardized residuals
6
4
2
0
−2
−4
−6
378
317
319
−2
0
2
Theoretical Quantiles
lm(bloom_doy ~ . − country − city − year)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
319
40
60
80
100
120
140
Fitted values
lm(bloom_doy ~ . − country − city − year)


Cook's distance
298
301
2412
Cook's distance
0.00
0.02
0.04
0.06
0
1000
2000
3000
4000
Obs. number
lm(bloom_doy ~ . − country − city − year)



Confidence Intervals

The confidence interval reflects the uncertainty around the mean predictions. To display the 95% confidence intervals around the mean the predictions, specify the option interval = "confidence".

For example, the 95% confidence interval associated with the first observation in the test data set is (43.02558, 47.14276). This means that, according to our model, a tree with same parameters with the first observation has, on average, a bloom_doy between 43 and 47.

```
lm_confidence_intervals <- predict(lm, newdata = test, interval = "confidence")
head(lm_confidence_intervals)
```

```
##         fit      lwr      upr
## 4  45.09240 43.22755 46.95725
## 6  40.54306 39.41630 41.66982
## 15 42.18923 41.13035 43.24812
## 23 42.08405 41.14056 43.02753
## 30 39.84616 38.86796 40.82436
## 37 43.11125 42.10332 44.11918
```

Prediction Intervals

The prediction interval gives uncertainty around a single value. In the same way, as the confidence intervals, the prediction intervals can be computed as follow.

The 95% prediction intervals associated with the first observation in the test dataset is (27.33479, 62.83354). This means that, according to our model, 95% of trees with same parameters have bloom_doy between 27 and 63.

```r
lm_prediction_intervals <- predict(lm, newdata = test, interval = "prediction")
head(lm_prediction_intervals)
```

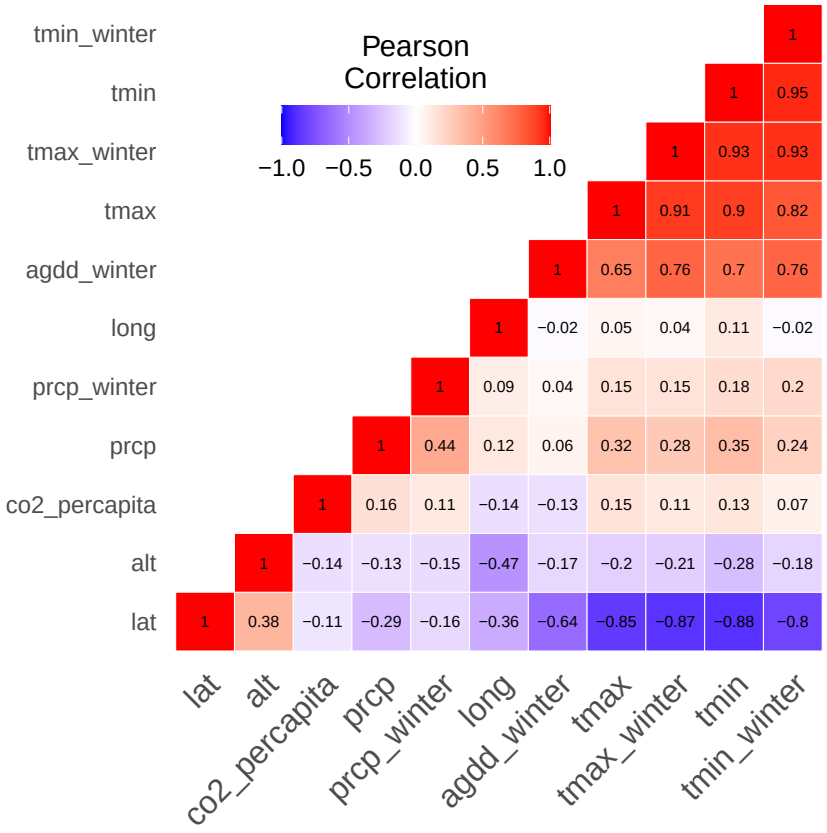
```
##         fit      lwr      upr
## 4  45.09240 27.46952 62.71528
## 6  40.54306 22.98294 58.10317
## 15 42.18923 24.63334 59.74513
## 23 42.08405 24.53474 59.63336
## 30 39.84616 22.29495 57.39737
## 37 43.11125 25.55836 60.66414
```

Check multicollinearity

First we check the correlation matrix of independent variables. We found that there are serveral variables that have high correlation with each other.

```r
X <- df %>% dplyr::select(-c(country, city, year, bloom_doy))
cormat <- round(cor(X),2)

reorder_cormat <- function(cormat){
  # Use correlation between variables as distance
  dd <- as.dist((1-cormat)/2)
  hc <- hclust(dd)
  cormat <-cormat[hc$order, hc$order]
}

# Get lower triangle of the correlation matrix
get_lower_tri<-function(cormat){
    cormat[upper.tri(cormat)] <- NA
    return(cormat)
}

# Get upper triangle of the correlation matrix
get_upper_tri <- function(cormat){
    cormat[lower.tri(cormat)]<- NA
    return(cormat)
}

# Reorder the correlation matrix
cormat <- reorder_cormat(cormat)
upper_tri <- get_upper_tri(cormat)
# Melt the correlation matrix
melted_cormat <- melt(upper_tri, na.rm = TRUE)
# Create a ggheatmap
ggheatmap <- ggplot(melted_cormat, aes(Var2, Var1, fill = value))+
 geom_tile(color = "white")+
 scale_fill_gradient2(low = "blue", high = "red", mid = "white",
   midpoint = 0, limit = c(-1,1), space = "Lab",
    name="Pearson\nCorrelation") +
  theme_minimal()+ # minimal theme
 theme(axis.text.x = element_text(angle = 45, vjust = 1,
    size = 12, hjust = 1))+
 coord_fixed() +
geom_text(aes(Var2, Var1, label = value), color = "black", size = 2) +
```

```
theme(
  axis.title.x = element_blank(),
  axis.title.y = element_blank(),
  panel.grid.major = element_blank(),
  panel.border = element_blank(),
  panel.background = element_blank(),
  axis.ticks = element_blank(),
  legend.justification = c(1, 0),
  legend.position = c(0.6, 0.7),
  legend.direction = "horizontal")+
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                title.position = "top", title.hjust = 0.5))
# Print the heatmap
print(ggheatmap)
```



The 'mctest' package in R provides the Farrar-Glauber test and other relevant tests for multicollinearity. Here we check the overall and individual diagnostic checking for multicollinearity respectively. The calculated value of the Chi-square test statistic is found to be 49102.7, implying the presence of multicollinearity in the model.

Next, we check Farrar–Glauber test (F-test) for the location of the multicollinearity. The VIF, TOL and Wi columns provide the diagnostic output for variance inflation factor, tolerance and Farrar-Glauber F-test respectively. Variables such as `tmin`, `tmin_winter`, and `tmax_winter` show high correlation with other features.

```
options(scipen=999)
omcdiag(lm)
```

```
##
## Call:
## omcdiag(mod = lm)
##
##
## Overall Multicollinearity Diagnostics
##
##                        MC Results detection
## Determinant |X'X|:         0.0000         1
## Farrar Chi-Square:     48935.8277         1
## Red Indicator:             0.4692         0
## Sum of Lambda Inverse:   102.2116         1
## Theil's Method:           -1.1328         0
## Condition Number:        126.9978         1
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
```

```
imcdiag(lm)
```

```
##
## Call:
## imcdiag(mod = lm)
##
##
## All Individual Multicollinearity Diagnostics Result
##
##                    VIF    TOL         Wi         Fi Leamer    CVIF Klein   IND1
## lat             9.5653 0.1045  3696.8002  4108.5074 0.3233 -0.5028     1 0.0002
## long            2.3903 0.4184   600.0517   666.8786 0.6468 -0.1256     0 0.0010
## alt             1.5151 0.6600   222.3185   247.0777 0.8124 -0.0796     0 0.0015
## tmax           10.0890 0.0991  3922.8109  4359.6887 0.3148 -0.5303     1 0.0002
## tmin           28.7112 0.0348 11960.1496 13292.1341 0.1866 -1.5092     1 0.0001
## prcp            1.7086 0.5853   305.8530   339.9154 0.7650 -0.0898     0 0.0014
## agdd_winter     2.9890 0.3346   858.4583   954.0636 0.5784 -0.1571     0 0.0008
## tmax_winter    18.6529 0.0536  7618.9712  8467.4850 0.2315 -0.9805     1 0.0001
## tmin_winter    23.7788 0.0421  9831.3116 10926.2106 0.2051 -1.2499     1 0.0001
## prcp_winter     1.4906 0.6709   211.7407   235.3219 0.8191 -0.0784     0 0.0016
## co2_percapita   1.3208 0.7571   138.4681   153.8891 0.8701 -0.0694     0 0.0018
##                  IND2
## lat            1.3606
## long           0.8837
## alt            0.5166
## tmax           1.3688
## tmin           1.4665
## prcp           0.6302
## agdd_winter    1.0111
## tmax_winter    1.4379
## tmin_winter    1.4555
## prcp_winter    0.5001
## co2_percapita 0.3691
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
```

```
## * all coefficients have significant t-ratios
##
## R-square of y on all x: 0.8373
##
## * use method argument to check which regressors may be the reason of collinearity
## ===================================
```

Mix Effects Model

```
suppressWarnings(suppressMessages(
  print(full_mixed_model <- lmer(bloom_doy ~ tmin_winter+
                                   prcp +
                                   agdd_winter +
                                   prcp_winter +
                                   co2_percapita +
                                   (tmin_winter +
                                      prcp +
                                      agdd_winter +
                                      prcp_winter +
                                      co2_percapita | city),
                                 data = train
                                 ))))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: bloom_doy ~ tmin_winter + prcp + agdd_winter + prcp_winter +
##     co2_percapita + (tmin_winter + prcp + agdd_winter + prcp_winter +
##     co2_percapita | city)
##    Data: train
## REML criterion at convergence: 27572.42
## Random effects:
##  Groups   Name          Std.Dev.  Corr
##  city     (Intercept)   17.113429
##           tmin_winter    1.247535  0.00
##           prcp           0.050872  0.02  1.00
##           agdd_winter    0.004213  0.43 -0.20 -0.20
##           prcp_winter    0.008996  0.10 -0.46 -0.46 -0.64
##           co2_percapita  0.142850 -0.23 -0.04 -0.05  0.76 -0.84
##  Residual                5.327073
## Number of obs: 4327, groups:  city, 105
## Fixed Effects:
##   (Intercept)    tmin_winter           prcp    agdd_winter    prcp_winter
##    100.151188      -1.528839      -0.022415       0.005131      -0.003842
## co2_percapita
##     -0.279891
## optimizer (nloptwrap) convergence code: 0 (OK) ; 0 optimizer warnings; 1 lme4 warnings
```

```
summary(full_mixed_model)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: bloom_doy ~ tmin_winter + prcp + agdd_winter + prcp_winter +
##     co2_percapita + (tmin_winter + prcp + agdd_winter + prcp_winter +
##     co2_percapita | city)
##    Data: train
##
## REML criterion at convergence: 27572.4
```

```
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -6.1780 -0.6006  0.0241  0.5917  4.3199
##
## Random effects:
##  Groups   Name          Variance     Std.Dev.  Corr
##  city     (Intercept)   292.86945655 17.113429
##           tmin_winter     1.55634381  1.247535  0.00
##           prcp            0.00258799  0.050872  0.02  1.00
##           agdd_winter     0.00001775  0.004213  0.43 -0.20 -0.20
##           prcp_winter     0.00008092  0.008996  0.10 -0.46 -0.46 -0.64
##           co2_percapita   0.02040614  0.142850 -0.23 -0.04 -0.05  0.76 -0.84
##  Residual                 28.37770471  5.327073
## Number of obs: 4327, groups:  city, 105
##
## Fixed effects:
##                   Estimate  Std. Error t value
## (Intercept)   100.1511877   1.7860289  56.075
## tmin_winter    -1.5288387   0.1529149  -9.998
## prcp           -0.0224151   0.0250522  -0.895
## agdd_winter     0.0051310   0.0009446   5.432
## prcp_winter    -0.0038424   0.0044622  -0.861
## co2_percapita  -0.2798909   0.0402592  -6.952
##
## Correlation of Fixed Effects:
##             (Intr) tmn_wn prcp   agdd_w prcp_w
## tmin_winter  0.044
## prcp        -0.076  0.091
## agdd_winter -0.039 -0.283  0.185
## prcp_winter -0.030 -0.149 -0.426  0.024
## co2_percapt -0.271 -0.146 -0.050  0.454 -0.080
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

```
head(coef(full_mixed_model)$city)
```

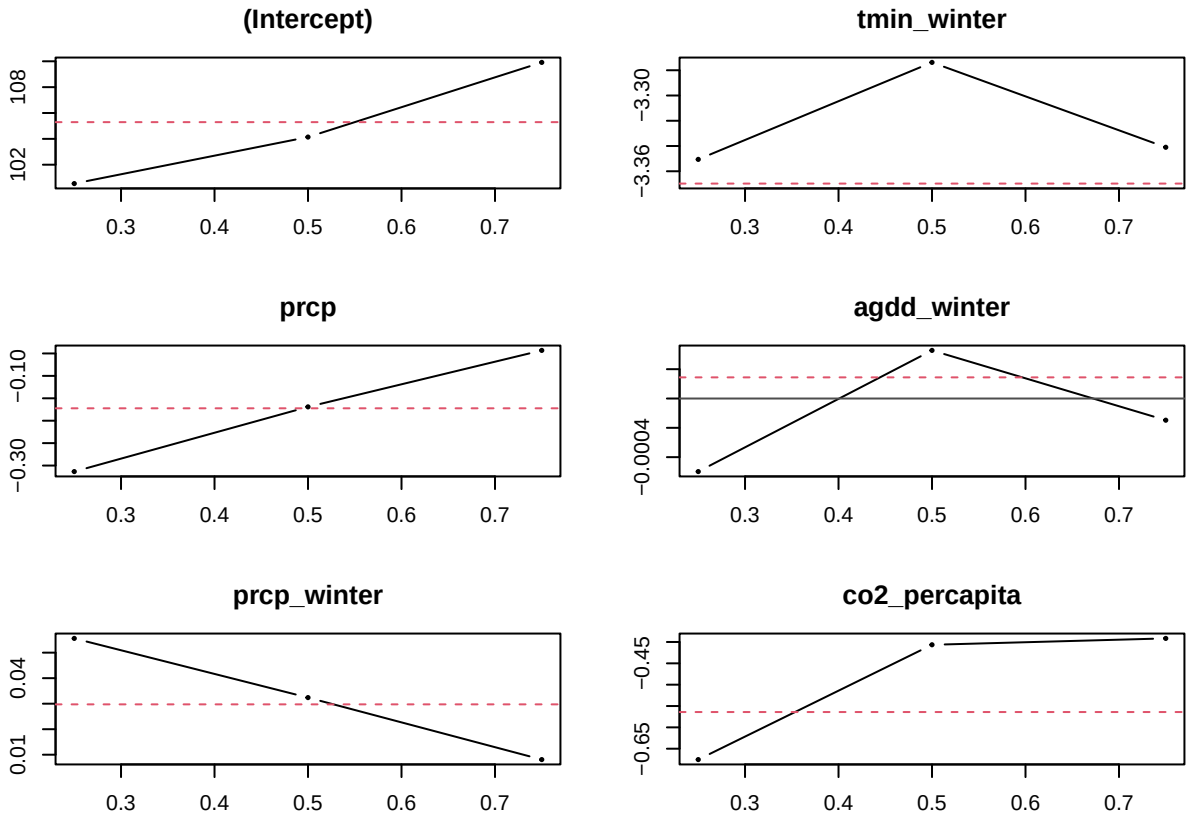
```
##                 (Intercept) tmin_winter        prcp agdd_winter  prcp_winter
## Abashiri          138.34745   0.1154472  0.04682133 0.007509464 -0.006193204
## Aikawa            111.11285  -2.4613460 -0.05997413 0.008648131 -0.003762338
## Akita             111.42374  -1.7789375 -0.03204712 0.007322146 -0.004190768
## Aomori            113.28283  -2.4340776 -0.05860437 0.007421905 -0.000738001
## Asahikawa         128.82712  -0.3438183  0.02785285 0.003621707  0.001582381
## Basel-Binningen    97.69248  -4.0845166 -0.12637100 0.002352252  0.013668125
##                 co2_percapita
## Abashiri           -0.3776682
## Aikawa             -0.2287468
## Akita              -0.2699398
## Aomori             -0.2904325
## Asahikawa          -0.4802303
## Basel-Binningen    -0.4198080
```

```
rq <- rq(bloom_doy~tmin_winter+prcp+agdd_winter+prcp_winter+co2_percapita,
         data = train,
         tau=c(0.25, 0.5, 0.75))
```

```
summary(rq)
```

```
##
## Call: rq(formula = bloom_doy ~ tmin_winter + prcp + agdd_winter + prcp_winter +
##     co2_percapita, tau = c(0.25, 0.5, 0.75), data = train)
##
## tau: [1] 0.25
##
## Coefficients:
##               Value     Std. Error t value   Pr(>|t|)
## (Intercept)   100.54006   0.83194  120.84997   0.00000
## tmin_winter    -3.35059   0.06159  -54.40091   0.00000
## prcp           -0.31346   0.06775   -4.62679   0.00000
## agdd_winter    -0.00050   0.00151   -0.32903   0.74215
## prcp_winter     0.05558   0.00744    7.46816   0.00000
## co2_percapita  -0.67546   0.07100   -9.51410   0.00000
##
## Call: rq(formula = bloom_doy ~ tmin_winter + prcp + agdd_winter + prcp_winter +
##     co2_percapita, tau = c(0.25, 0.5, 0.75), data = train)
##
## tau: [1] 0.5
##
## Coefficients:
##               Value     Std. Error t value   Pr(>|t|)
## (Intercept)   104.13908   0.64275  162.02101   0.00000
## tmin_winter    -3.27378   0.04988  -65.62700   0.00000
## prcp           -0.16914   0.05457   -3.09938   0.00195
## agdd_winter     0.00033   0.00102    0.32213   0.74737
## prcp_winter     0.03238   0.00582    5.56060   0.00000
## co2_percapita  -0.40652   0.05480   -7.41809   0.00000
##
## Call: rq(formula = bloom_doy ~ tmin_winter + prcp + agdd_winter + prcp_winter +
##     co2_percapita, tau = c(0.25, 0.5, 0.75), data = train)
##
## tau: [1] 0.75
##
## Coefficients:
##               Value      Std. Error t value    Pr(>|t|)
## (Intercept)    109.91384    0.58766  187.03520    0.00000
## tmin_winter     -3.34100    0.02897 -115.33964    0.00000
## prcp            -0.04320    0.05685   -0.75993    0.44734
## agdd_winter     -0.00015    0.00077   -0.19206    0.84771
## prcp_winter      0.00809    0.00543    1.49060    0.13614
## co2_percapita   -0.39162    0.05736   -6.82750    0.00000
```

```
plot(rq)
```




```
head(round(predict(rq, newdata = test), 0))
```

```
##    tau= 0.25 tau= 0.50 tau= 0.75
## 4         44        51        57
## 6         37        45        49
## 15        42        49        53
## 23        38        46        50
## 30        32        41        45
## 37        39        47        52
```