Graphical Analysis

Let us perform a graphical analysis of the available data and have a look at its statistcal characteristics.

Load data

We have separated the original excel file into separate .csv files to ease the import. Since temperature and weather data from google trends are monthly values, we decided to set them to the 15th of each month to match with the gas data. We omit here the last 12 lines of the gas.csv which correspond to the values to predict.

```
library(readr)
gas <- na.omit(read_csv("data/gas.csv", col_types = cols(date = col_date(format = "%Y-%m-%d"))))
temp <- read_csv("data/temp.csv", col_types = cols(date = col_date(format = "%Y-%m-%d")))
google <- read_csv("data/google.csv", col_types = cols(date = col_date(format = "%Y-%m-%d")))
```

Let us have a look at the statistical summary of these datasets.

```
summary(gas)
```

```
##       date               gas_cons        log_gas_cons
##  Min.   :1973-01-15   Min.   : 102770   Min.   :11.54
##  1st Qu.:1983-11-22   1st Qu.: 146251   1st Qu.:11.89
##  Median :1994-09-30   Median : 315840   Median :12.66
##  Mean   :1994-09-29   Mean   : 395710   Mean   :12.64
##  3rd Qu.:2005-08-07   3rd Qu.: 636806   3rd Qu.:13.36
##  Max.   :2016-06-15   Max.   :1037197   Max.   :13.85
```

The temperature data is taken from the National Center for Environmental Information.

```
summary(temp)
```

```
##       date             av_temp_val     av_temp_ano      min_temp_val
##  Min.   :1895-01-15   Min.   :21.90   Min.   :-8.590   Min.   :12.52
##  1st Qu.:1925-07-22   1st Qu.:37.96   1st Qu.:-0.990   1st Qu.:26.95
##  Median :1956-01-30   Median :52.73   Median : 0.170   Median :40.11
##  Mean   :1956-01-29   Mean   :52.16   Mean   : 0.171   Mean   :40.19
##  3rd Qu.:1986-08-07   3rd Qu.:66.72   3rd Qu.: 1.450   3rd Qu.:54.01
##  Max.   :2017-02-15   Max.   :76.80   Max.   : 8.910   Max.   :63.55
##   min_temp_ano      max_temp_val    max_temp_ano        prec_val
##  Min.   :-8.6200   Min.   :31.26   Min.   :-9.2300   Min.   :0.540
##  1st Qu.:-0.9400   1st Qu.:48.55   1st Qu.:-1.2400   1st Qu.:2.140
##  Median : 0.1900   Median :65.34   Median : 0.1600   Median :2.510
##  Mean   : 0.1854   Mean   :64.13   Mean   : 0.1542   Mean   :2.501
##  3rd Qu.: 1.3800   3rd Qu.:79.78   3rd Qu.: 1.6800   3rd Qu.:2.880
##  Max.   : 8.4700   Max.   :90.84   Max.   :10.0900   Max.   :4.440
##     prec_ano         cool_days_val   cool_days_ano     heat_days_val
##  Min.   :-1.620000   Min.   :  1.0   Min.   :-72.000   Min.   :   3.0
##  1st Qu.:-0.310000   1st Qu.: 10.0   1st Qu.: -5.000   1st Qu.:  56.0
##  Median :-0.010000   Median : 41.0   Median : -1.000   Median : 312.0
##  Mean   : 0.006603   Mean   :101.0   Mean   :  1.931   Mean   : 385.9
##  3rd Qu.: 0.310000   3rd Qu.:184.8   3rd Qu.:  7.000   3rd Qu.: 692.0
##  Max.   : 2.130000   Max.   :405.0   Max.   : 90.000   Max.   :1184.0
##  heat_days_ano
```

```
##  Min.   :-258.000
##  1st Qu.: -26.000
##  Median :  -2.000
##  Mean   :  -3.763
##  3rd Qu.:  15.000
##  Max.   : 259.000
```

```
summary(google)
```

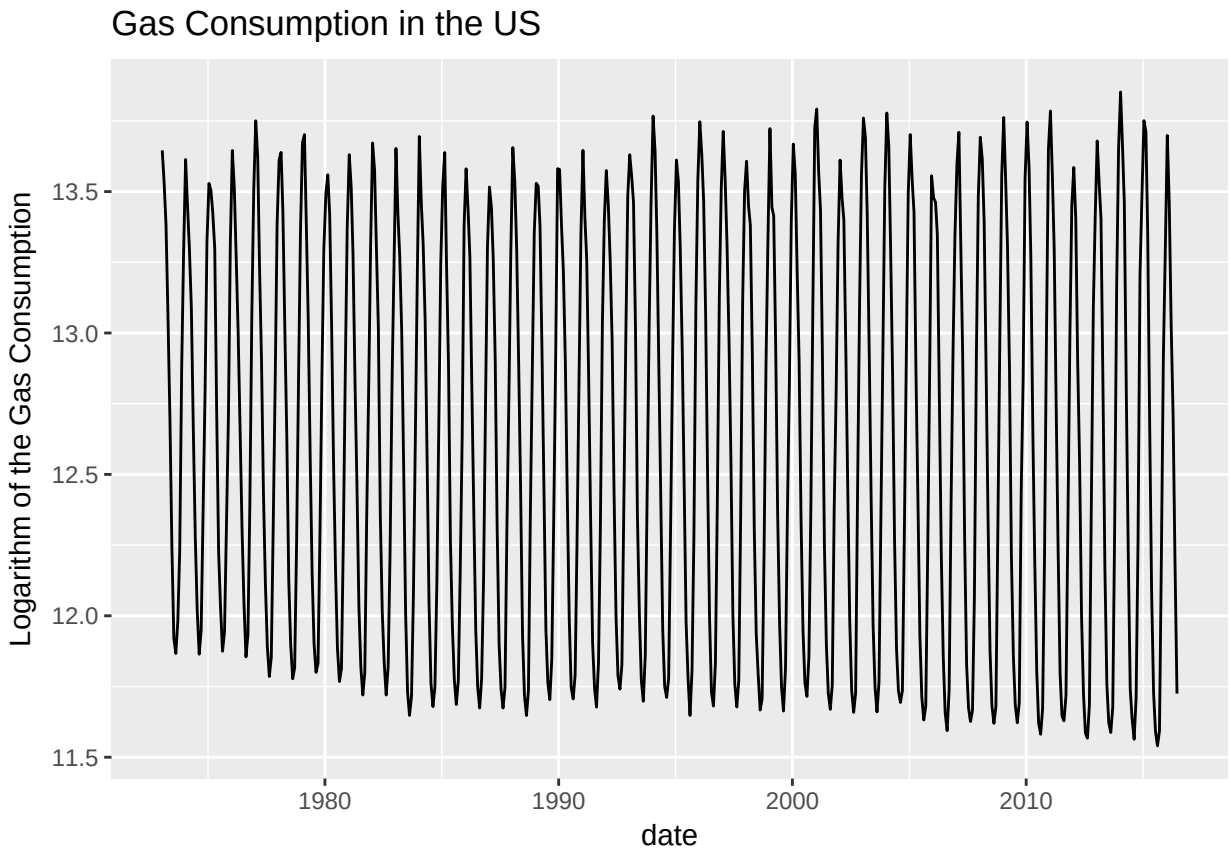
```
##       date               heatwave    extreme_weather   snow_storm
##  Min.   :2004-01-15   Min.   :2.00   Min.   :1.000   Min.   :  1.000
##  1st Qu.:2007-04-30   1st Qu.:3.00   1st Qu.:2.000   1st Qu.:  2.000
##  Median :2010-08-15   Median :3.00   Median :2.000   Median :  3.000
##  Mean   :2010-08-15   Mean   :3.27   Mean   :2.252   Mean   :  8.767
##  3rd Qu.:2013-11-30   3rd Qu.:4.00   3rd Qu.:3.000   3rd Qu.:  9.000
##  Max.   :2017-03-15   Max.   :8.00   Max.   :8.000   Max.   :100.000
```

We first notice that these datasets span over different periods of time: while gas consumption, our target value goes from January 1973 to July 2016, temperature data start in January 1895 and end in February 2017. Finally, data from Google Trends go from january 2004 and up to march 2017.

Plots

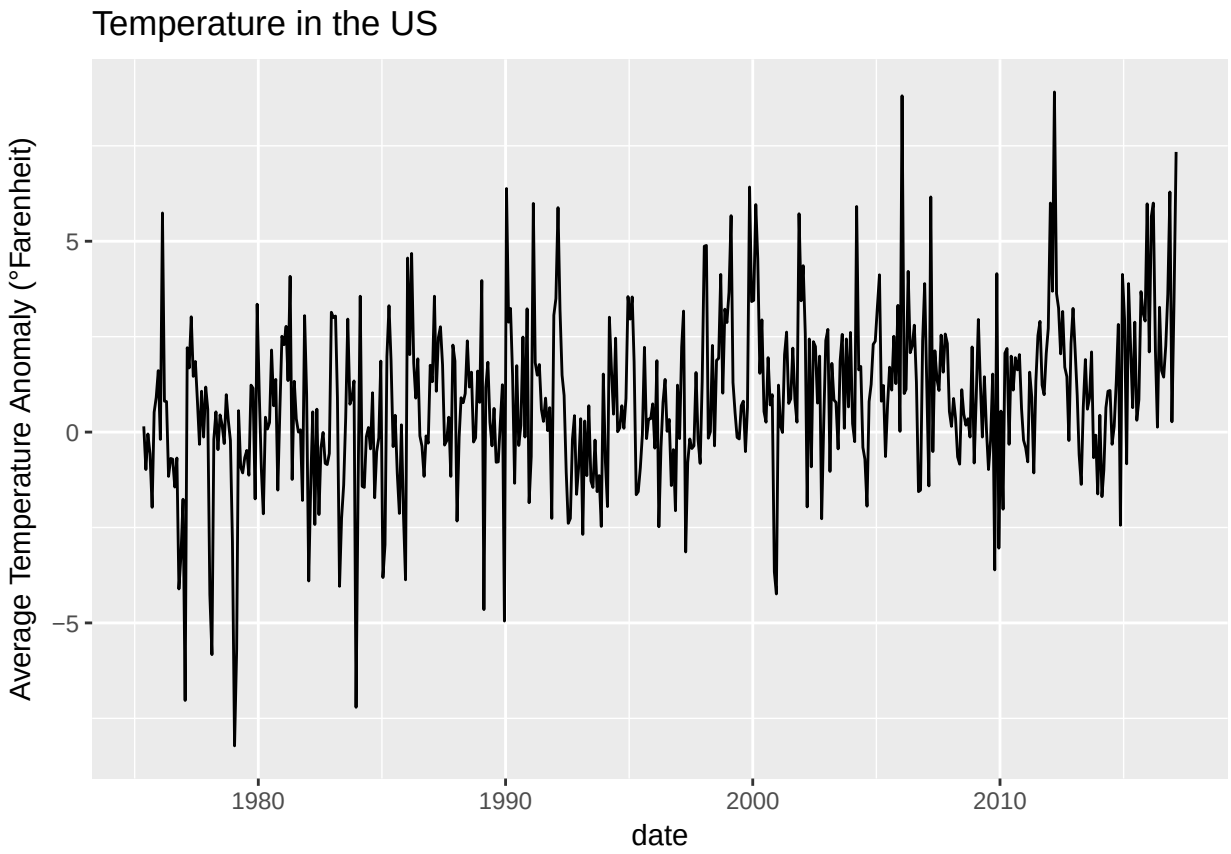
Gas consumption

```
library(ggplot2)
#ggplot(gas, aes(x=date, y=gas_cons)) + geom_line() +
#            labs(title = "Gas Consumption in the US",  y = "Gas Consumption")
ggplot(gas, aes(x=date, y=log_gas_cons)) + geom_line() +
            labs(title = "Gas Consumption in the US",  y = "Logarithm of the Gas Consumption")
```

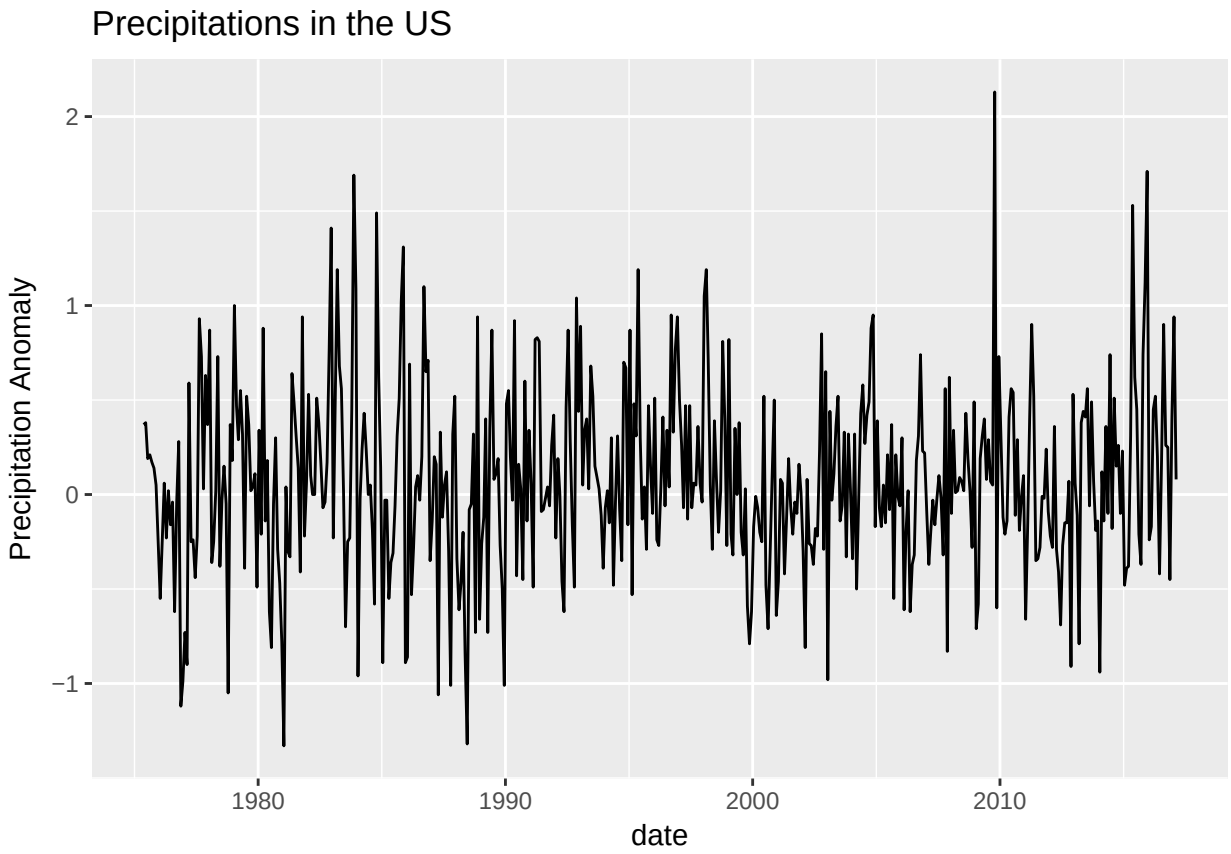


Temperature

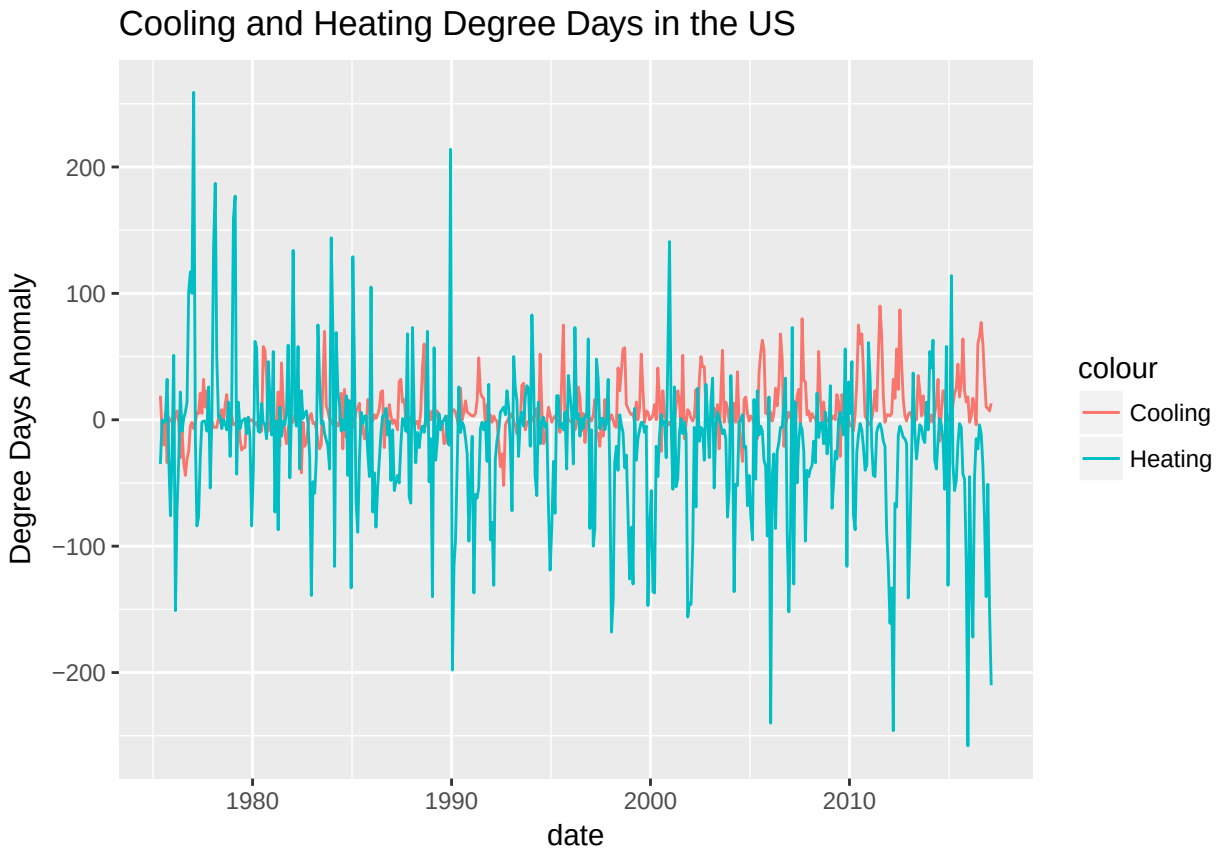
```
temp_trunc = temp[temp$date >= 1973-01-15,]
ggplot(temp_trunc)+geom_line(aes(x=date, y=av_temp_ano))+
    labs(title = "Temperature in the US",  y = "Average Temperature Anomaly (°Farenheit)")
```




```
ggplot(temp_trunc)+geom_line(aes(x=date, y=prec_ano))+
    labs(title = "Precipitations in the US",  y = "Precipitation Anomaly")
```

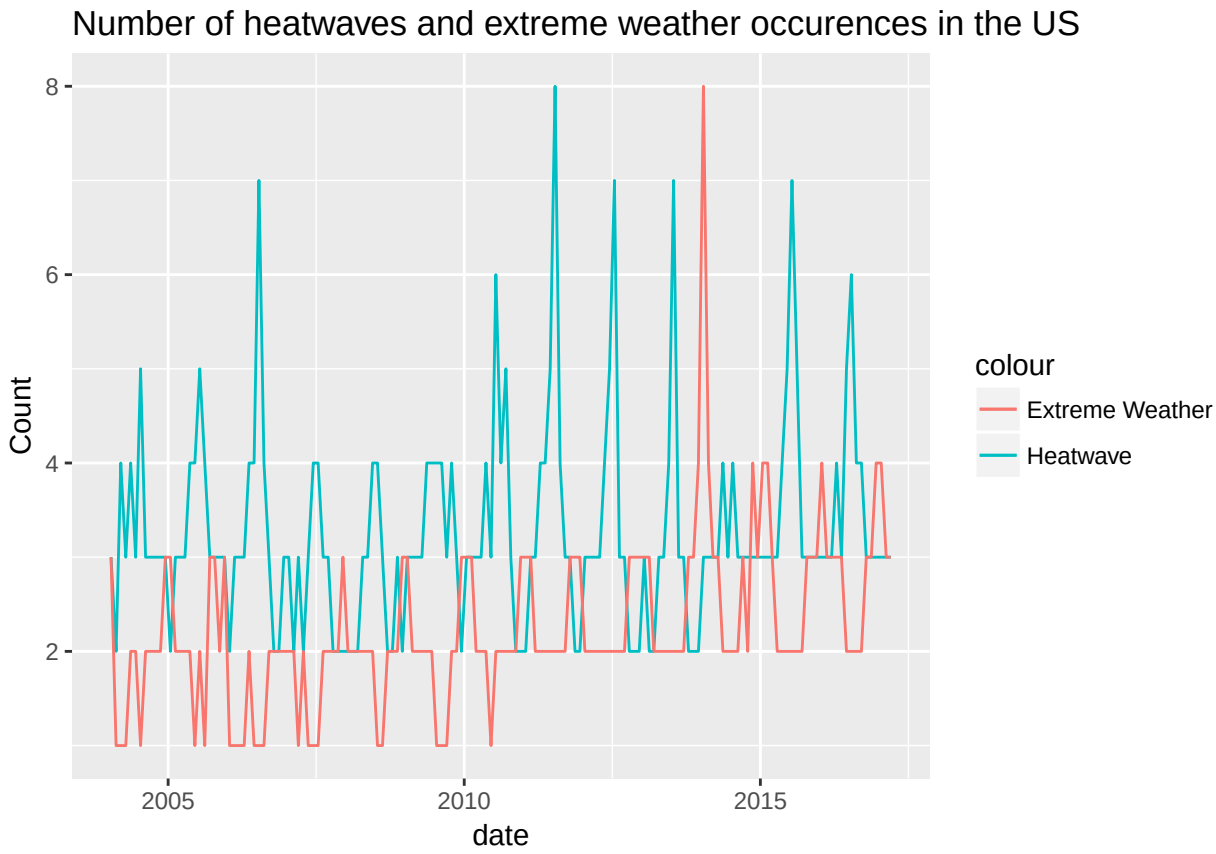



```
ggplot(temp_trunc)+geom_line(aes(x=date, y=cool_days_ano, colour = "Cooling"))+
    geom_line(aes(x=date, y=heat_days_ano, colour ="Heating"))+
    labs(title = "Cooling and Heating Degree Days in the US",  y = "Degree Days Anomaly")
```

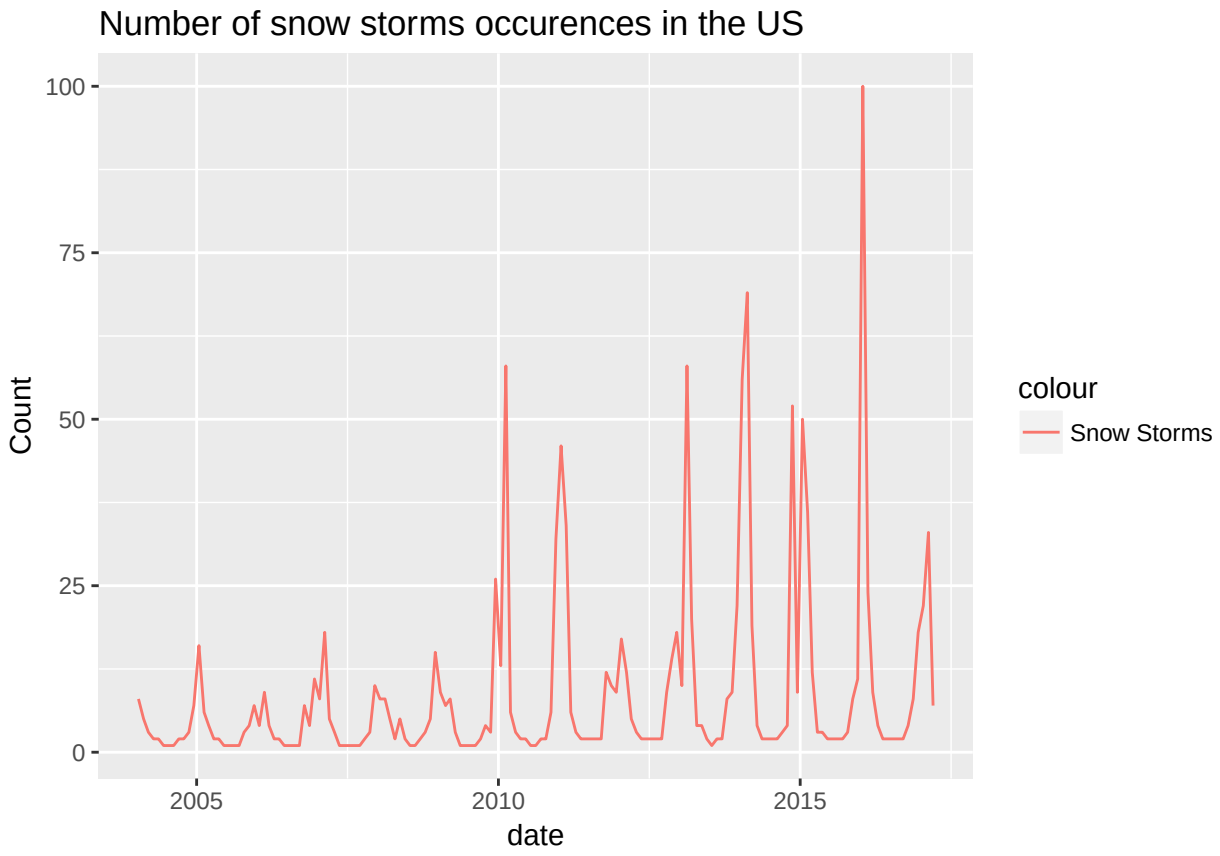


Google trends

```
ggplot(google)+geom_line(aes(x=date, y=heatwave, colour="Heatwave"))+
    geom_line(aes(x=date, y=extreme_weather, colour = "Extreme Weather"))+
    labs(title = "Number of heatwaves and extreme weather occurences in the US",  y = "Count")
```




```
ggplot(google)+geom_line(aes(x=date, y=snow_storm, colour="Snow Storms"))+
    labs(title = "Number of snow storms occurences in the US",  y = "Count")
```

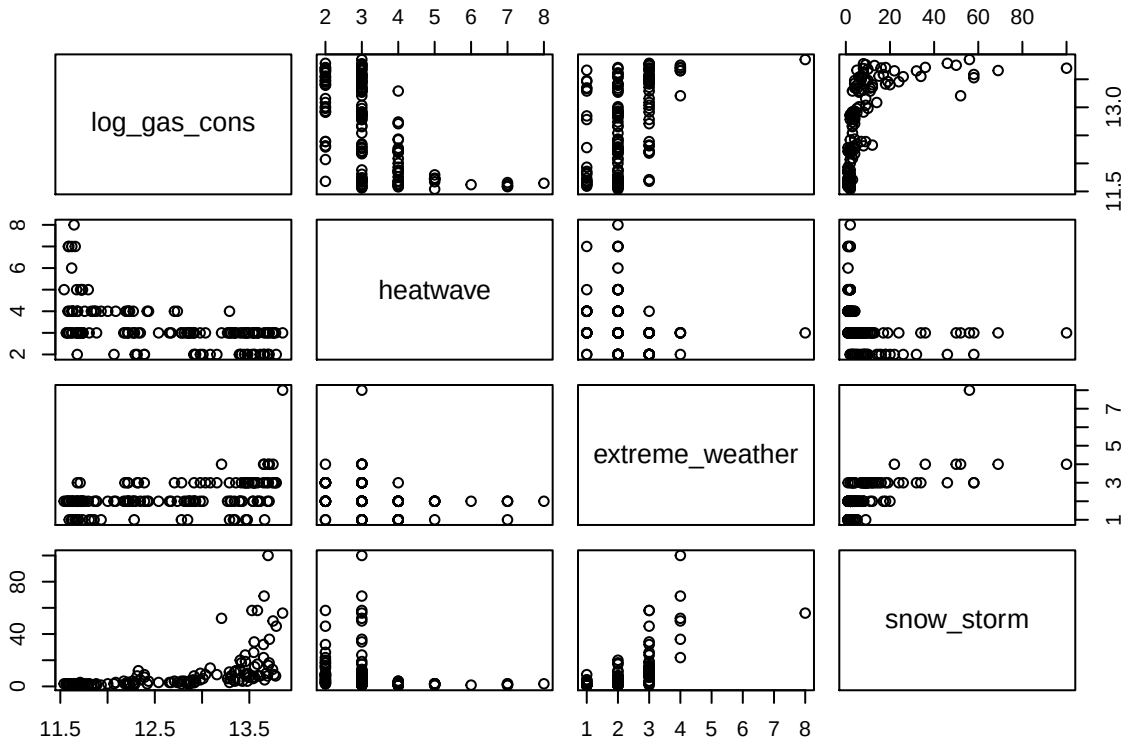


Covariance analysis

Let us merge the dataframes

```
gas_temp = merge(gas, temp, by = "date", all.x = TRUE)
gas_goo = merge(gas, google, by = "date", all = FALSE)
all = merge(gas_temp, google, by= "date",  all = FALSE)
```

```
pairs(gas_goo[,-c(1,2)])
```


log_gas_cons
heatwave
extreme_weather
snow_storm
2 3 4 5 6 7 8
0 20 40 60 80
11.5 13.0
2 4 6 8
1 3 5 7
0 40 80
11.5 12.5 13.5
1 2 3 4 5 6 7 8