ST117 Individual DRAFT Written Report - Part EDA

My WARWICK ID 5645242 (Report Pod 041)

2025-04-08

Step 5 Summary of aspects of the datasets that are most relevant for statistical analyses

For all three datasets: As a quality check, a standard quality control solution is analysed alongside the collected samples. Multiple levels of quality control were applied: Quality Codes

Stream water chemistry data (1992-2015)

When, Where and How was the data collected

Dip samples are collected weekly at designated ECN sites across the UK, representing different landscapes and ecosystems. (Including T02, T04, T05, T06, T07, T08, T11, T12). A clean 250 ml bottle is used, rinsed beforehand to avoid contamination. Conductivity and pH are measured on unfiltered water. Filtered samples are analysed for key ions, alkalinity, and dissolved organic carbon. Continuous monitoring of temperature, pH, conductivity, and turbidity is planned.

Why was the data collected

Aims to detect changes in water quality caused by factors such as climate change, pollution, and land use. We can assess ecosystem health, acidification, and nutrient budgets.

Precipitation chemistry data (1992-2015)

When, Where and How was the data collected

The data was collected weekly from T01 to T12 terrestrial ECN sites across the UK, including upland and lowland regions. Precipitation samples are collected using a continuously open bulk collector. The collector gathers both wet and dry deposition. At each collection, the bottle is replaced, and the funnel is cleaned/swapped with a clean one. Samples are measured and analysed for volume, pH, conductivity, and conc. of dissolved ions and potential contamination factors are recorded.

Why was the data collected

The aim is to assess the impact of atmospheric pollutant on ecosystems. It support the detection of enviromental changes linked to pollution and contributes to researches on acid rain and atmospheric deposition which would help policy makinf.

Soil solution chemistry data (1992-2015)

When, Where and How was the data collected

The samples are collected fortnightly from T01 to T12 terrestrial ECN sites across the UK, with different regions. The data is collected using suction lysimeters installed at two depths in the soil, shallow and deep.

These samplers mimic plant roots and draw in soil water, which is later analysed for pH, conductivity, nutrients, dissolved ions etc.

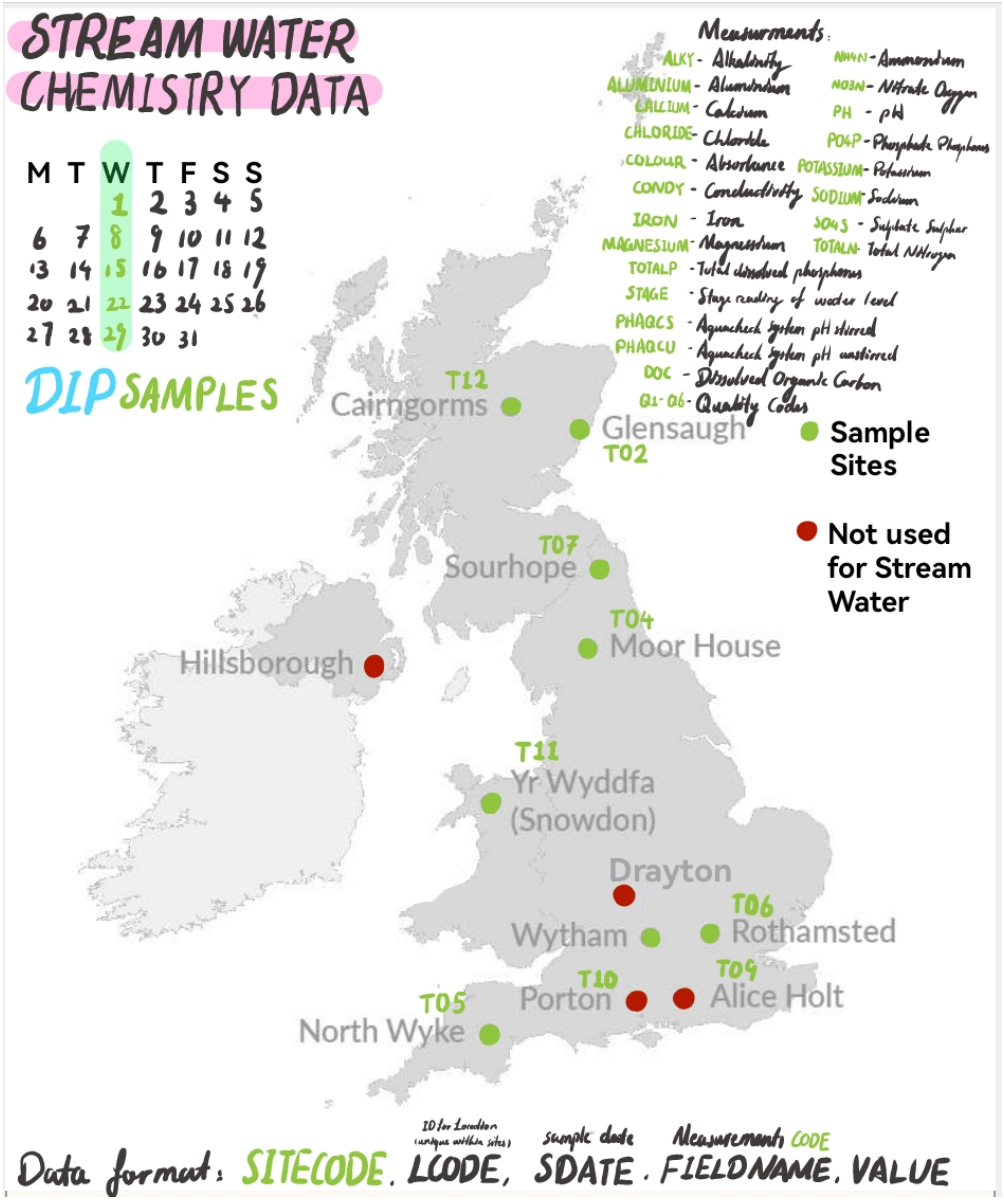
Why was the data collected

Aims to assess temporal changes in nutrient availability and soil chemistry. This helps in understanding its responsiveness to climate, pollution, or land use changes.

Step 6: Diagrams about data structure

Stream water

```
knitr::include_graphics("/Users/danielguo/Desktop/University/Year 1/ST117/5645242_041_WR/Phase 1 Diagra
```



Precipitation

```
knitr::include_graphics("/Users/danielguo/Desktop/University/Year 1/ST117/5645242_041_WR/Phase 1 Diagra
```

Soil solution

```
knitr::include_graphics("/Users/danielguo/Desktop/University/Year 1/ST117/5645242_041_WR/Phase 1 Diagra
```

Step 7: Data loading and preprocessing

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats 1.0.0     v stringr 1.5.1
## v purrr   1.0.4     v tibble  3.2.1
## v readr   2.1.5
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
#read and create individual data frames for each chemistry data:
df_stream <- read.csv("/Users/danielguo/Desktop/University/Year 1/ST117/5645242_041_WR/data/Stream wate
```

```r
df_precipitation <- read.csv("/Users/danielguo/Desktop/University/Year 1/ST117/5645242_041_WR/data/Preci
df_soil <- read.csv("/Users/danielguo/Desktop/University/Year 1/ST117/5645242_041_WR/data/Soil solution/

#remove the irrelevant variable for this assignment for stream water data
columns_to_remove_wc <- c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "COLOUR")
df_stream_wide <- df_stream %>%
  pivot_wider(
    names_from = FIELDNAME,
    values_from = VALUE
  ) %>%
  select(-all_of(columns_to_remove_wc)) #converting the dataframe to a format showing them as separate
df_stream_wide <- df_stream_wide %>%
  mutate(SDATE = dmy(SDATE)) #converting the date format for easier later use
head(df_stream_wide) #print the first few rows
```

```
## # A tibble: 6 x 23
##   SITECODE LCODE SDATE      ALUMINIUM TOTALN CHLORIDE   DOC  IRON MAGNESIUM
##   <chr>    <int> <date>         <dbl>  <dbl>    <dbl> <dbl> <dbl>     <dbl>
## 1 T04          3 1992-10-06      0.04   0.28      3.9  10.3  0.53      0.9
## 2 T04          1 1992-10-06      0.16   0.55      3.4  16    0.93      0.59
## 3 T04          2 1992-10-06      0.1    0.5       3.6  19.9  0.82      0.31
## 4 T04          3 1992-10-15      0.06   0.31      4.4  11.7  0.45      0.68
## 5 T04          1 1992-10-15      0.12   0.41      4.2  15    0.71      0.44
## 6 T04          2 1992-10-15      0.1    0.4       4.2  16.6  0.52      0.33
## # i 14 more variables: NH4N <dbl>, NO3N <dbl>, PH <dbl>, PO4P <dbl>,
## #   POTASSIUM <dbl>, SO4S <dbl>, SODIUM <dbl>, STAGE <dbl>, CALCIUM <dbl>,
## #   PHAQCS <dbl>, PHAQCU <dbl>, ALKY <dbl>, CONDY <dbl>, TOTALP <dbl>
```

```r
#same process will be applied for the other two dataframes:
columns_to_remove_pc <- c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "COLOUR")
df_precipitation_wide <- df_precipitation %>%
  pivot_wider(
    names_from = FIELDNAME,
    values_from = VALUE
  ) %>%
  select(-all_of(columns_to_remove_pc))
df_precipitation_wide <- df_precipitation_wide %>%
  mutate(SDATE = dmy(SDATE))
head(df_precipitation_wide)
```

```
## # A tibble: 6 x 22
##   SITECODE SDATE      VOLUME ALUMINIUM SODIUM   SO4S POTASSIUM  PO4P    PH  NO3N
##   <chr>    <date>      <dbl>     <dbl>  <dbl>  <dbl>     <dbl> <dbl> <dbl> <dbl>
## 1 T04      1992-10-06    611         0   0.34  0.5        0.05     0  4.61  0.3
## 2 T04      1992-10-15    467         0   2.1   0.59       0.1      0  4.97  0.11
## 3 T04      1992-10-21     63         0   3.68  0.75       0.16     0  4.74  0.26
## 4 T04      1992-10-28    842         0   0.7   0.34       0.05     0  4.9   0.11
## 5 T06      1992-10-28    358        NA   1.08  0.613      0       NA  4.56  0.12
## 6 T08      1992-10-28    332        NA  NA    NA         NA       NA NA    NA
## # i 12 more variables: NH4N <dbl>, MAGNESIUM <dbl>, IRON <dbl>, DOC <dbl>,
## #   CHLORIDE <dbl>, CALCIUM <dbl>, TOTALN <dbl>, PHAQCS <dbl>, PHAQCU <dbl>,
## #   CONDY <dbl>, ALKY <dbl>, TOTALP <dbl>
```

```r
columns_to_remove_ss <- c("Q1", "Q2", "Q3", "Q4", "Q5", "COLOUR")
df_soil_wide <- df_soil %>%
```

```
  pivot_wider(
    names_from = FIELDNAME,
    values_from = VALUE
  ) %>%
  select(-all_of(columns_to_remove_ss))
df_soil_wide <- df_soil_wide %>%
  mutate(SDATE = dmy(SDATE))
head(df_soil_wide)
```

```
## # A tibble: 6 x 24
##   SITECODE SDATE      RID   VACUUM VOLUME  SO4S SODIUM TOTALN ALUMINIUM CALCIUM
##   <chr>    <date>     <chr>  <dbl>  <dbl> <dbl>  <dbl>  <dbl>     <dbl>   <dbl>
## 1 T04      1992-10-06 A3D     0.27    360    NA     NA     NA        NA      NA
## 2 T04      1992-10-06 F4S     0.16    500    NA     NA     NA        NA      NA
## 3 T04      1992-10-06 A3S     0.24    390    NA     NA     NA        NA      NA
## 4 T04      1992-10-06 B1D     0.23    410    NA     NA     NA        NA      NA
## 5 T04      1992-10-06 B1S     0.12    540    NA     NA     NA        NA      NA
## 6 T04      1992-10-06 C5D     0.21    440    NA     NA     NA        NA      NA
## # i 14 more variables: CHLORIDE <dbl>, DOC <dbl>, IRON <dbl>, MAGNESIUM <dbl>,
## #   NH4N <dbl>, NO3N <dbl>, PH <dbl>, PO4P <dbl>, POTASSIUM <dbl>,
## #   PHAQCS <dbl>, PHAQCU <dbl>, ALKY <dbl>, CONDY <dbl>, TOTALP <dbl>
```

```
#save the dataframes to import in phase 2
saveRDS(df_stream_wide, "df_stream_wide.rds")
saveRDS(df_precipitation_wide, "df_precipitation_wide.rds")
saveRDS(df_soil_wide, "df_soil_wide.rds")
```

Step 8: DQA (Data Quality Analysis)

Check for Uniqueness

```
# Check for uniqueness: find the number of duplicated rows
sum(duplicated(df_stream_wide))
```

```
## [1] 0
```

```
sum(duplicated(df_precipitation_wide))
```

```
## [1] 0
```

```
sum(duplicated(df_soil_wide))
```

```
## [1] 0
```

Check for Timeliness

```
# Check for timeliness: sample date range and frequency

range(df_stream_wide$SDATE, na.rm = TRUE) #Finding the range for stream water sample dates
```

```
## [1] "1992-10-06" "2015-12-31"
```

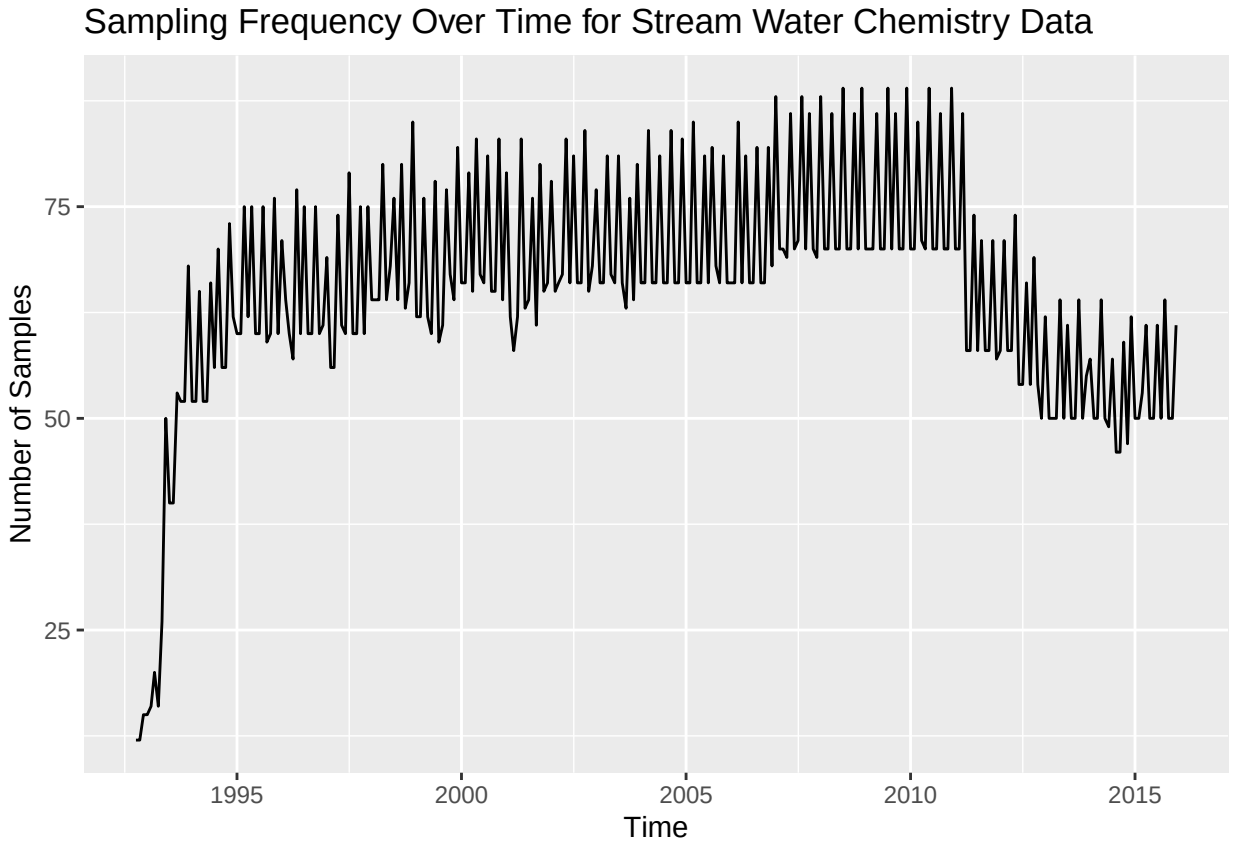
```
range(df_precipitation_wide$SDATE, na.rm = TRUE) #Finding the range for precipitation sample dates
```

```
## [1] "1992-10-06" "2015-12-31"
```

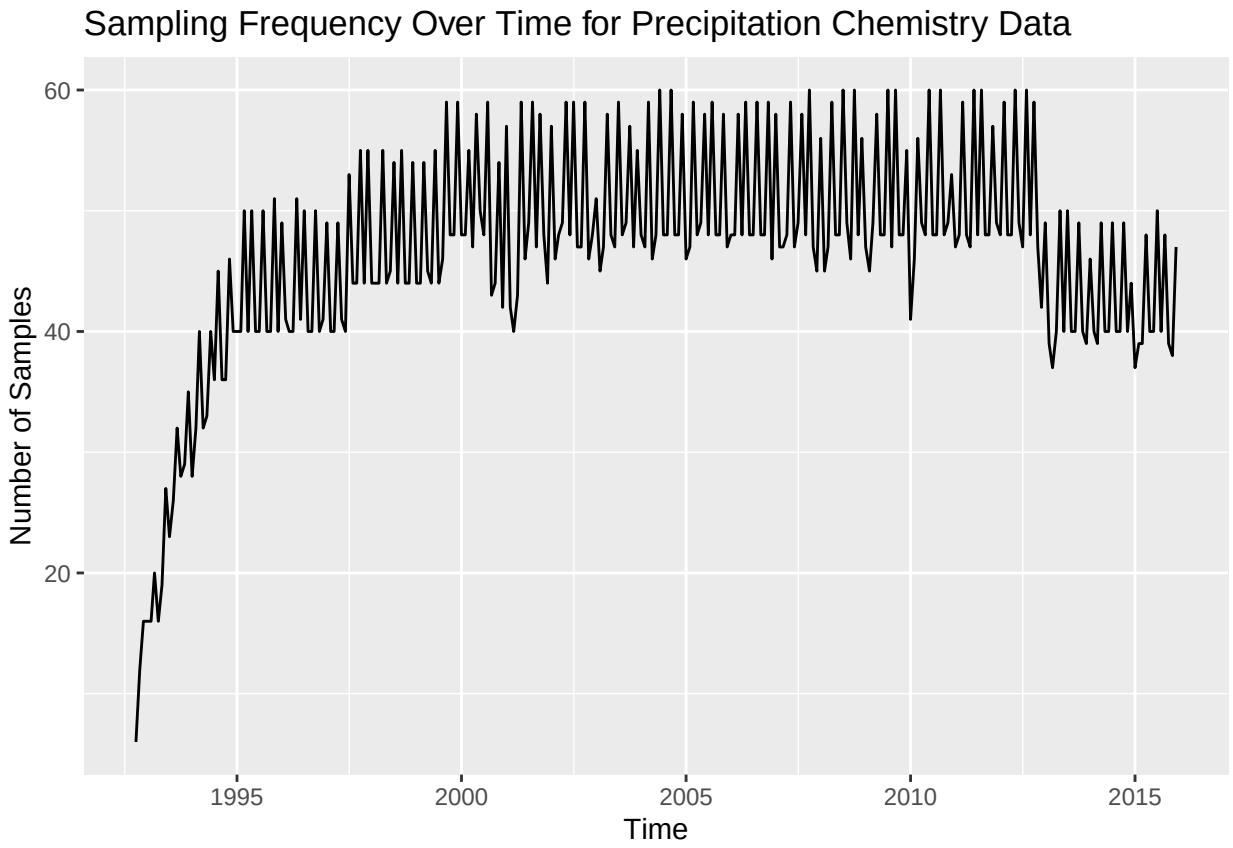
```r
range(df_soil_wide$SDATE, na.rm = TRUE) #Finding the range for soil solution sample dates
```

```
## [1] "1992-10-06" "2015-12-23"
```

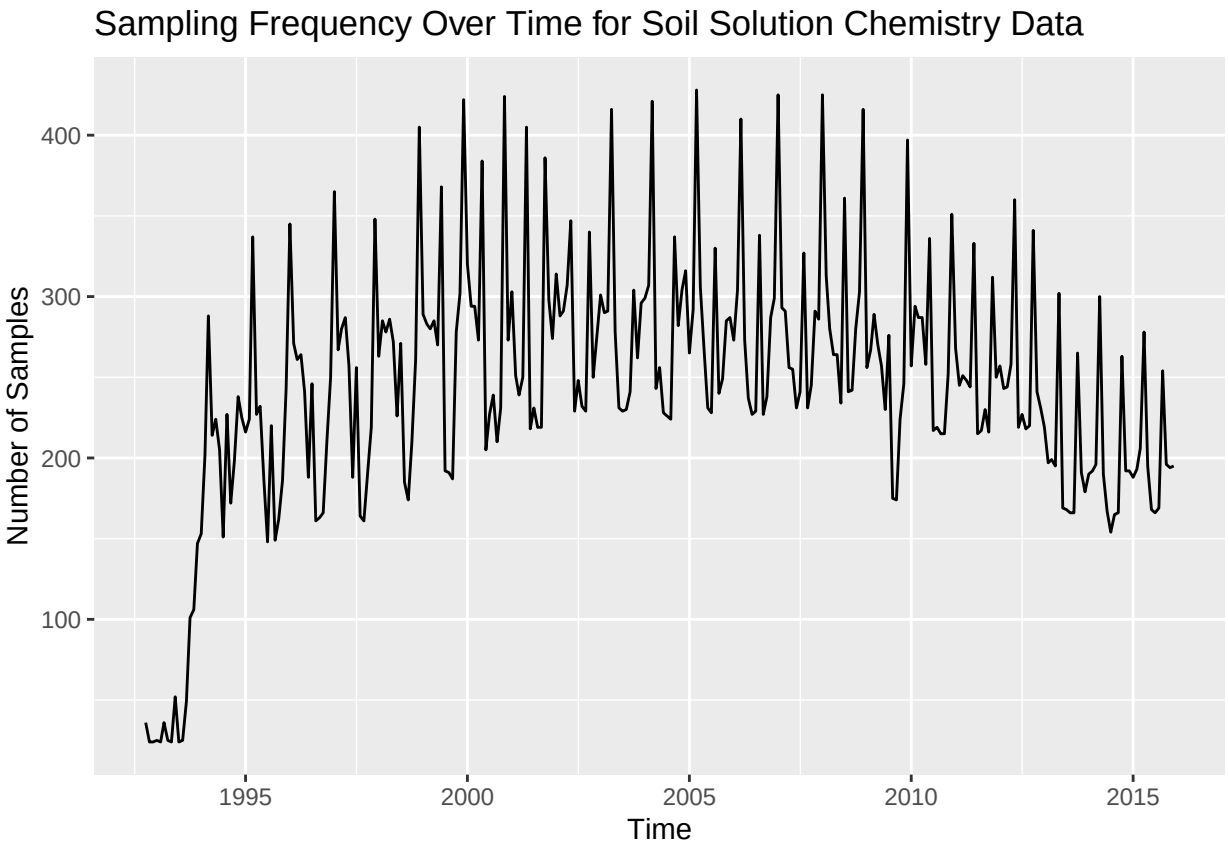
```r
df_stream_wide %>%
  # Create a new column "Month" by extracting the year and month from the SDATE column
  mutate(Month = format(SDATE, "%Y-%m")) %>%
  # Count the number of records for each Month value
  count(Month) %>%
  # Create a line plot using ggplot2
  ggplot(aes(x = as.Date(paste0(Month, "-01")), y = n)) +
  # Plot the number of samples as a line over time
  geom_line() +
  #labels
  labs(title = "Sampling Frequency Over Time for Stream Water Chemistry Data", x = "Time", y = "Number
```




```r
#Same process is used for the other two data frames
df_precipitation_wide %>%
  mutate(Month = format(SDATE, "%Y-%m")) %>%
  count(Month) %>%
  ggplot(aes(x = as.Date(paste0(Month, "-01")), y = n)) +
  geom_line() +
  labs(title = "Sampling Frequency Over Time for Precipitation Chemistry Data", x = "Time", y = "Number
```




```
df_soil_wide %>%
  mutate(Month = format(SDATE, "%Y-%m")) %>%
  count(Month) %>%
  ggplot(aes(x = as.Date(paste0(Month, "-01")), y = n)) +
  geom_line() +
  labs(title = "Sampling Frequency Over Time for Soil Solution Chemistry Data", x = "Time", y = "Number
```



Check for Validity

```
# Check for Validity:
#Check for dates before 1992 or after 2015 for all data frames
df_stream_wide %>%
  filter(SDATE < as.Date("1992-01-01") | SDATE > as.Date("2015-12-31"))
```

```
## # A tibble: 0 x 23
## # i 23 variables: SITECODE <chr>, LCODE <int>, SDATE <date>, ALUMINIUM <dbl>,
## #   TOTALN <dbl>, CHLORIDE <dbl>, DOC <dbl>, IRON <dbl>, MAGNESIUM <dbl>,
## #   NH4N <dbl>, NO3N <dbl>, PH <dbl>, PO4P <dbl>, POTASSIUM <dbl>, SO4S <dbl>,
## #   SODIUM <dbl>, STAGE <dbl>, CALCIUM <dbl>, PHAQCS <dbl>, PHAQCU <dbl>,
## #   ALKY <dbl>, CONDY <dbl>, TOTALP <dbl>
```

```
df_precipitation_wide %>%
  filter(SDATE < as.Date("1992-01-01") | SDATE > as.Date("2015-12-31"))
```

```
## # A tibble: 0 x 22
## # i 22 variables: SITECODE <chr>, SDATE <date>, VOLUME <dbl>, ALUMINIUM <dbl>,
## #   SODIUM <dbl>, SO4S <dbl>, POTASSIUM <dbl>, PO4P <dbl>, PH <dbl>,
## #   NO3N <dbl>, NH4N <dbl>, MAGNESIUM <dbl>, IRON <dbl>, DOC <dbl>,
## #   CHLORIDE <dbl>, CALCIUM <dbl>, TOTALN <dbl>, PHAQCS <dbl>, PHAQCU <dbl>,
## #   CONDY <dbl>, ALKY <dbl>, TOTALP <dbl>
```

```
df_soil_wide %>%
  filter(SDATE < as.Date("1992-01-01") | SDATE > as.Date("2015-12-31"))
```

```
## # A tibble: 0 x 24
```

```
## # i 24 variables: SITECODE <chr>, SDATE <date>, RID <chr>, VACUUM <dbl>,
## #   VOLUME <dbl>, SO4S <dbl>, SODIUM <dbl>, TOTALN <dbl>, ALUMINIUM <dbl>,
## #   CALCIUM <dbl>, CHLORIDE <dbl>, DOC <dbl>, IRON <dbl>, MAGNESIUM <dbl>,
## #   NH4N <dbl>, NO3N <dbl>, PH <dbl>, PO4P <dbl>, POTASSIUM <dbl>,
## #   PHAQCS <dbl>, PHAQCU <dbl>, ALKY <dbl>, CONDY <dbl>, TOTALP <dbl>
```

```
# Check for any pH values below 0 or above 14
df_stream_wide %>% filter(PH < 0 | PH > 14)
```

```
## # A tibble: 0 x 23
## # i 23 variables: SITECODE <chr>, LCODE <int>, SDATE <date>, ALUMINIUM <dbl>,
## #   TOTALN <dbl>, CHLORIDE <dbl>, DOC <dbl>, IRON <dbl>, MAGNESIUM <dbl>,
## #   NH4N <dbl>, NO3N <dbl>, PH <dbl>, PO4P <dbl>, POTASSIUM <dbl>, SO4S <dbl>,
## #   SODIUM <dbl>, STAGE <dbl>, CALCIUM <dbl>, PHAQCS <dbl>, PHAQCU <dbl>,
## #   ALKY <dbl>, CONDY <dbl>, TOTALP <dbl>
```

```
df_precipitation_wide %>% filter(PH < 0 | PH > 14)
```

```
## # A tibble: 0 x 22
## # i 22 variables: SITECODE <chr>, SDATE <date>, VOLUME <dbl>, ALUMINIUM <dbl>,
## #   SODIUM <dbl>, SO4S <dbl>, POTASSIUM <dbl>, PO4P <dbl>, PH <dbl>,
## #   NO3N <dbl>, NH4N <dbl>, MAGNESIUM <dbl>, IRON <dbl>, DOC <dbl>,
## #   CHLORIDE <dbl>, CALCIUM <dbl>, TOTALN <dbl>, PHAQCS <dbl>, PHAQCU <dbl>,
## #   CONDY <dbl>, ALKY <dbl>, TOTALP <dbl>
```

```
df_soil_wide %>% filter(PH < 0 | PH > 14)
```

```
## # A tibble: 0 x 24
## # i 24 variables: SITECODE <chr>, SDATE <date>, RID <chr>, VACUUM <dbl>,
## #   VOLUME <dbl>, SO4S <dbl>, SODIUM <dbl>, TOTALN <dbl>, ALUMINIUM <dbl>,
## #   CALCIUM <dbl>, CHLORIDE <dbl>, DOC <dbl>, IRON <dbl>, MAGNESIUM <dbl>,
## #   NH4N <dbl>, NO3N <dbl>, PH <dbl>, PO4P <dbl>, POTASSIUM <dbl>,
## #   PHAQCS <dbl>, PHAQCU <dbl>, ALKY <dbl>, CONDY <dbl>, TOTALP <dbl>
```

```
#Check for any negative units except Alkalinity
df_stream_wide %>%
  select(-ALKY) %>% # remove ALKY temporarily
  filter(if_any(where(is.numeric), ~ . < 0))
```

```
## # A tibble: 0 x 22
## # i 22 variables: SITECODE <chr>, LCODE <int>, SDATE <date>, ALUMINIUM <dbl>,
## #   TOTALN <dbl>, CHLORIDE <dbl>, DOC <dbl>, IRON <dbl>, MAGNESIUM <dbl>,
## #   NH4N <dbl>, NO3N <dbl>, PH <dbl>, PO4P <dbl>, POTASSIUM <dbl>, SO4S <dbl>,
## #   SODIUM <dbl>, STAGE <dbl>, CALCIUM <dbl>, PHAQCS <dbl>, PHAQCU <dbl>,
## #   CONDY <dbl>, TOTALP <dbl>
```

```
df_precipitation_wide %>%
  select(-ALKY) %>% # remove ALKY temporarily
  filter(if_any(where(is.numeric), ~ . < 0))
```

```
## # A tibble: 1 x 21
##   SITECODE SDATE      VOLUME ALUMINIUM SODIUM  SO4S POTASSIUM  PO4P    PH  NO3N
##   <chr>    <date>      <dbl>     <dbl>  <dbl> <dbl>     <dbl> <dbl> <dbl> <dbl>
## 1 T03      2013-07-03    534        NA   0.15    NA      0.07 0.004  6.71  0.06
## # i 11 more variables: NH4N <dbl>, MAGNESIUM <dbl>, IRON <dbl>, DOC <dbl>,
## #   CHLORIDE <dbl>, CALCIUM <dbl>, TOTALN <dbl>, PHAQCS <dbl>, PHAQCU <dbl>,
## #   CONDY <dbl>, TOTALP <dbl>
```

```
df_soil_wide %>%
  select(-ALKY) %>%    # remove ALKY temporarily
  filter(if_any(where(is.numeric), ~ . < 0))
```

```
## # A tibble: 0 x 23
## # i 23 variables: SITECODE <chr>, SDATE <date>, RID <chr>, VACUUM <dbl>,
## #   VOLUME <dbl>, SO4S <dbl>, SODIUM <dbl>, TOTALN <dbl>, ALUMINIUM <dbl>,
## #   CALCIUM <dbl>, CHLORIDE <dbl>, DOC <dbl>, IRON <dbl>, MAGNESIUM <dbl>,
## #   NH4N <dbl>, NO3N <dbl>, PH <dbl>, PO4P <dbl>, POTASSIUM <dbl>,
## #   PHAQCS <dbl>, PHAQCU <dbl>, CONDY <dbl>, TOTALP <dbl>
```

Note There exist a negative for the variable of magnesium in the precipitation data frame.

Check for Accuracy

```
#Check for accuracy, use summary and compare to real life
summary(df_stream_wide)
```

```
##    SITECODE             LCODE           SDATE               ALUMINIUM
##  Length:18257       Min.   :1.000   Min.   :1992-10-06   Min.   :0.000
##  Class :character   1st Qu.:1.000   1st Qu.:1999-03-24   1st Qu.:0.000
##  Mode  :character   Median :2.000   Median :2004-08-25   Median :0.032
##                     Mean   :2.165   Mean   :2004-07-22   Mean   :0.048
##                     3rd Qu.:3.000   3rd Qu.:2009-10-14   3rd Qu.:0.070
##                     Max.   :7.000   Max.   :2015-12-31   Max.   :1.280
##                                                          NA's   :3858
##      TOTALN          CHLORIDE            DOC               IRON
##  Min.   : 0.000   Min.   :  0.000   Min.   :  0.000   Min.   : 0.000
##  1st Qu.: 0.300   1st Qu.:  4.199   1st Qu.:  2.500   1st Qu.: 0.004
##  Median : 0.453   Median :  7.140   Median :  5.700   Median : 0.040
##  Mean   : 1.685   Mean   : 13.421   Mean   :  8.122   Mean   : 0.208
##  3rd Qu.: 0.890   3rd Qu.: 19.400   3rd Qu.: 10.800   3rd Qu.: 0.290
##  Max.   :47.600   Max.   :526.000   Max.   :355.000   Max.   :22.000
##  NA's   :5885     NA's   :3003      NA's   :4650      NA's   :3857
##    MAGNESIUM           NH4N             NO3N              PH
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   :3.720
##  1st Qu.: 0.446   1st Qu.: 0.000   1st Qu.: 0.048   1st Qu.:5.940
##  Median : 2.135   Median : 0.020   Median : 0.130   Median :7.070
##  Mean   : 2.433   Mean   : 0.046   Mean   : 1.440   Mean   :6.685
##  3rd Qu.: 4.308   3rd Qu.: 0.040   3rd Qu.: 0.542   3rd Qu.:7.700
##  Max.   :13.200   Max.   :12.000   Max.   :45.600   Max.   :8.930
##  NA's   :3002     NA's   :3403     NA's   :3037     NA's   :2663
##       PO4P          POTASSIUM           SO4S             SODIUM
##  Min.   :0.0000   Min.   :  0.000   Min.   :  0.000   Min.   :  0.01
##  1st Qu.:0.0000   1st Qu.:  0.260   1st Qu.:  0.760   1st Qu.:  2.82
##  Median :0.0000   Median :  0.600   Median :  1.930   Median :  6.75
##  Mean   :0.0192   Mean   :  1.014   Mean   :  7.993   Mean   :  8.18
##  3rd Qu.:0.0120   3rd Qu.:  1.070   3rd Qu.: 11.855   3rd Qu.: 13.00
##  Max.   :1.5930   Max.   :707.110   Max.   :313.000   Max.   :102.32
##  NA's   :3150     NA's   :2999      NA's   :3149      NA's   :3019
##      STAGE          CALCIUM           PHAQCS          PHAQCU
##  Min.   :  0.0   Min.   :  0.00   Min.   : 0.00   Min.   : 0.00
##  1st Qu.: 27.0   1st Qu.:  2.45   1st Qu.: 4.00   1st Qu.: 4.04
##  Median : 80.0   Median :  9.60   Median : 4.05   Median : 4.08
```

```
##  Mean   :113.8   Mean   : 47.05   Mean   : 4.45   Mean   : 4.52
##  3rd Qu.:150.0   3rd Qu.:123.00   3rd Qu.: 4.10   3rd Qu.: 4.17
##  Max.   :810.0   Max.   :506.00   Max.   :78.80   Max.   :40.10
##  NA's   :8627    NA's   :3018     NA's   :7733    NA's   :7735
##       ALKY             CONDY            TOTALP
##  Min.   :  -5.00   Min.   :   0.2   Min.   :0.000
##  1st Qu.:   6.60   1st Qu.:  42.6   1st Qu.:0.000
##  Median :  37.75   Median :  99.0   Median :0.007
##  Mean   :  96.53   Mean   : 279.2   Mean   :0.015
##  3rd Qu.: 214.00   3rd Qu.: 668.0   3rd Qu.:0.014
##  Max.   :1000.00   Max.   :7668.0   Max.   :1.390
##  NA's   :6979      NA's   :3249     NA's   :10700
```

```
summary(df_precipitation_wide)
```

```
##    SITECODE             SDATE                VOLUME         ALUMINIUM
##  Length:13008       Min.   :1992-10-06   Min.   :   0.0   Min.   : 0.000
##  Class :character   1st Qu.:1999-07-21   1st Qu.:  65.0   1st Qu.: 0.000
##  Mode  :character   Median :2004-12-01   Median : 216.0   Median : 0.004
##                     Mean   :2004-11-01   Mean   : 353.7   Mean   : 0.082
##                     3rd Qu.:2010-03-18   3rd Qu.: 477.0   3rd Qu.: 0.015
##                     Max.   :2015-12-31   Max.   :4356.0   Max.   :40.900
##                                          NA's   :232      NA's   :5391
##      SODIUM            SO4S          POTASSIUM           PO4P
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.709   1st Qu.: 0.300   1st Qu.: 0.040   1st Qu.: 0.000
##  Median : 1.358   Median : 0.490   Median : 0.110   Median : 0.001
##  Mean   : 2.093   Mean   : 0.732   Mean   : 0.352   Mean   : 0.059
##  3rd Qu.: 2.480   3rd Qu.: 0.801   3rd Qu.: 0.242   3rd Qu.: 0.007
##  Max.   :63.230   Max.   :52.489   Max.   :89.100   Max.   :48.960
##  NA's   :3939     NA's   :3580     NA's   :3961     NA's   :3491
##       PH              NO3N             NH4N            MAGNESIUM
##  Min.   :2.130   Min.   :  0.000   Min.   :   0.000   Min.   :-0.320
##  1st Qu.:4.740   1st Qu.:  0.150   1st Qu.:   0.138   1st Qu.: 0.084
##  Median :5.170   Median :  0.279   Median :   0.297   Median : 0.160
##  Mean   :5.343   Mean   :  0.504   Mean   :   1.522   Mean   : 0.322
##  3rd Qu.:5.800   3rd Qu.:  0.533   3rd Qu.:   0.610   3rd Qu.: 0.300
##  Max.   :9.120   Max.   :122.338   Max.   :1264.000   Max.   :99.401
##  NA's   :2345    NA's   :3358      NA's   :3422       NA's   :3957
##      IRON             DOC            CHLORIDE         CALCIUM
##  Min.   : 0.000   Min.   :  0.000   Min.   :  0.000   Min.   :  0.000
##  1st Qu.: 0.000   1st Qu.:  0.940   1st Qu.:  1.120   1st Qu.:  0.170
##  Median : 0.003   Median :  1.600   Median :  2.288   Median :  0.320
##  Mean   : 0.030   Mean   :  2.935   Mean   :  3.850   Mean   :  1.097
##  3rd Qu.: 0.014   3rd Qu.:  3.000   3rd Qu.:  4.420   3rd Qu.:  0.640
##  Max.   :29.000   Max.   :250.000   Max.   :336.140   Max.   :412.000
##  NA's   :5351     NA's   :6277      NA's   :3362      NA's   :3946
##      TOTALN            PHAQCS           PHAQCU          CONDY
##  Min.   :  0.000   Min.   : 0.000   Min.   : 0.000   Min.   :   0.00
##  1st Qu.:  0.350   1st Qu.: 4.020   1st Qu.: 4.040   1st Qu.:  14.20
##  Median :  0.620   Median : 4.070   Median : 4.120   Median :  22.10
##  Mean   :  1.186   Mean   : 4.642   Mean   : 4.782   Mean   :  32.64
##  3rd Qu.:  1.145   3rd Qu.: 5.260   3rd Qu.: 5.520   3rd Qu.:  35.90
##  Max.   :189.630   Max.   :80.100   Max.   :40.100   Max.   :6360.00
##  NA's   :7270      NA's   :7303     NA's   :7092     NA's   :2796
```

```
##       ALKY             TOTALP
##  Min.   : -9.400   Min.   :0.000
##  1st Qu.:  0.000   1st Qu.:0.000
##  Median :  0.000   Median :0.000
##  Mean   :  3.277   Mean   :0.024
##  3rd Qu.:  2.180   3rd Qu.:0.010
##  Max.   :380.000   Max.   :4.280
##  NA's   :6960      NA's   :11611
```

```
summary(df_soil_wide)
```

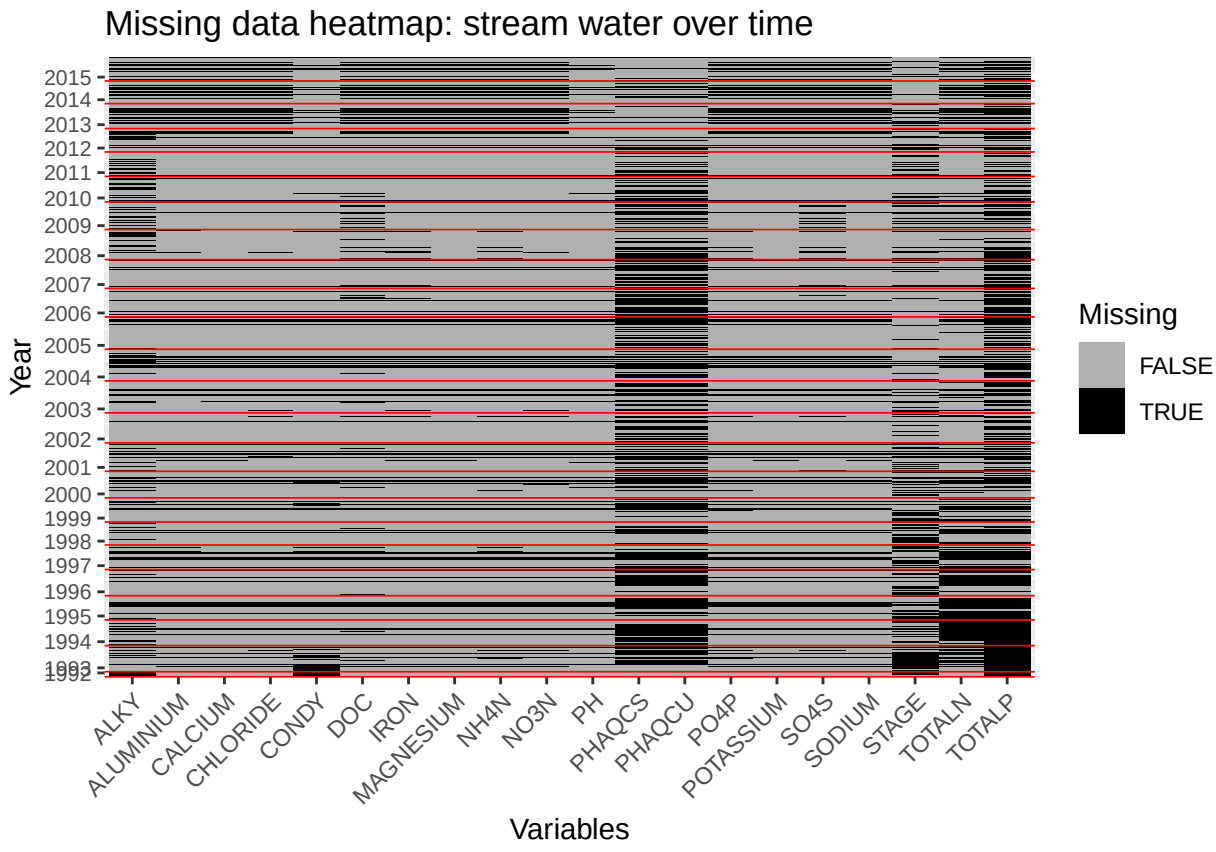
```
##    SITECODE             SDATE                RID                VACUUM
##  Length:68058       Min.   :1992-10-06   Length:68058       Min.   :0.000
##  Class :character   1st Qu.:1999-10-06   Class :character   1st Qu.:0.000
##  Mode  :character   Median :2004-10-20   Mode  :character   Median :0.200
##                     Mean   :2004-10-08                      Mean   :0.213
##                     3rd Qu.:2009-11-04                      3rd Qu.:0.380
##                     Max.   :2015-12-23                      Max.   :0.900
##                                                             NA's   :7225
##      VOLUME            SO4S            SODIUM            TOTALN
##  Min.   :   0.0   Min.   :  0.00   Min.   :   0.00   Min.   :  0.00
##  1st Qu.:   0.0   1st Qu.:  0.32   1st Qu.:   2.64   1st Qu.:  0.20
##  Median :  50.0   Median :  0.94   Median :   3.85   Median :  0.45
##  Mean   : 114.3   Mean   :  1.52   Mean   :   4.51   Mean   :  0.80
##  3rd Qu.: 165.0   3rd Qu.:  2.00   3rd Qu.:   5.61   3rd Qu.:  0.75
##  Max.   :1189.0   Max.   :101.57   Max.   :1161.00   Max.   :505.00
##  NA's   :6757     NA's   :40153    NA's   :41829     NA's   :49576
##    ALUMINIUM        CALCIUM          CHLORIDE           DOC
##  Min.   : 0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 0.04   1st Qu.:  0.41   1st Qu.:  3.22   1st Qu.:  3.23
##  Median : 0.19   Median :  0.78   Median :  5.10   Median : 12.01
##  Mean   : 0.39   Mean   :  5.05   Mean   :  7.29   Mean   : 12.69
##  3rd Qu.: 0.64   3rd Qu.:  3.72   3rd Qu.:  8.80   3rd Qu.: 18.60
##  Max.   :15.13   Max.   :567.40   Max.   :166.79   Max.   :333.00
##  NA's   :45948   NA's   :41830    NA's   :39475    NA's   :46305
##       IRON         MAGNESIUM          NH4N              NO3N
##  Min.   : 0.00   Min.   : 0.00   Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 0.01   1st Qu.: 0.34   1st Qu.:  0.00   1st Qu.:  0.01
##  Median : 0.07   Median : 0.51   Median :  0.02   Median :  0.03
##  Mean   : 0.16   Mean   : 1.05   Mean   :  0.23   Mean   :  1.09
##  3rd Qu.: 0.19   3rd Qu.: 1.09   3rd Qu.:  0.07   3rd Qu.:  0.08
##  Max.   :23.73   Max.   :55.00   Max.   :916.00   Max.   :190.25
##  NA's   :45322   NA's   :41813   NA's   :40692    NA's   :39972
##        PH             PO4P          POTASSIUM          PHAQCS
##  Min.   :3.20    Min.   :  0.00   Min.   :   0.00   Min.   :3.02
##  1st Qu.:4.47    1st Qu.:  0.00   1st Qu.:   0.04   1st Qu.:4.03
##  Median :5.00    Median :  0.00   Median :   0.21   Median :4.09
##  Mean   :5.40    Mean   :  0.04   Mean   :   0.57   Mean   :4.83
##  3rd Qu.:6.27    3rd Qu.:  0.01   3rd Qu.:   0.46   3rd Qu.:5.83
##  Max.   :9.00    Max.   :626.00   Max.   :1334.00   Max.   :8.18
##  NA's   :36423   NA's   :39446    NA's   :41910     NA's   :51943
##      PHAQCU            ALKY             CONDY             TOTALP
##  Min.   :0.68    Min.   :-11.20   Min.   :   0.00   Min.   :0.00
##  1st Qu.:4.05    1st Qu.:  0.00   1st Qu.:  32.30   1st Qu.:0.00
##  Median :4.16    Median :  0.00   Median :  43.60   Median :0.01
```

```
##  Mean   :5.01    Mean   : 12.23   Mean   :  70.96   Mean   :0.02
##  3rd Qu.:6.00    3rd Qu.:  3.00   3rd Qu.:  70.60   3rd Qu.:0.01
##  Max.   :8.45    Max.   :477.00   Max.   :2412.00   Max.   :1.14
##  NA's   :51474   NA's   :52418    NA's   :37974     NA's   :63277
```

Check for Completeness

```
#Completness
# Stream Water
# Prepare the data for plotting
df_stream_missing <- df_stream_wide %>%
  select(SDATE, where(is.numeric), -LCODE) %>%   # Keep SDATE + numeric columns, remove LCODE
  pivot_longer(-SDATE, names_to = "Variables", values_to = "Value") %>%
  mutate(Missing = is.na(Value)) %>%
  arrange(SDATE) %>%
  mutate(Row = as.numeric(factor(SDATE, levels = unique(SDATE))))

# Year break lines for visual clarity
year_breaks_stream <- df_stream_missing %>%
  distinct(SDATE, Row) %>%
  mutate(Year = as.numeric(format(SDATE, "%Y"))) %>%
  group_by(Year) %>%
  summarise(Row = min(Row), .groups = "drop")

# Plot
ggplot(df_stream_missing, aes(x = Variables, y = Row, fill = Missing)) +
  geom_tile(color = NA) +
  scale_fill_manual(values = c("FALSE" = "grey69", "TRUE" = "black")) +
  scale_y_reverse() +
  geom_hline(data = year_breaks_stream, aes(yintercept = Row), color = "red", linewidth = 0.3) +
  scale_y_continuous(
    breaks = year_breaks_stream$Row + 10,  # adjust this based on your density
    labels = year_breaks_stream$Year,
    expand = c(0, 0)
  ) +
  labs(title = "Missing data heatmap: stream water over time",
       x = "Variables", y = "Year", fill = "Missing")+
  theme(axis.text.y = element_text(size = 8),
        axis.text.x = element_text(angle = 45, hjust = 1))
```

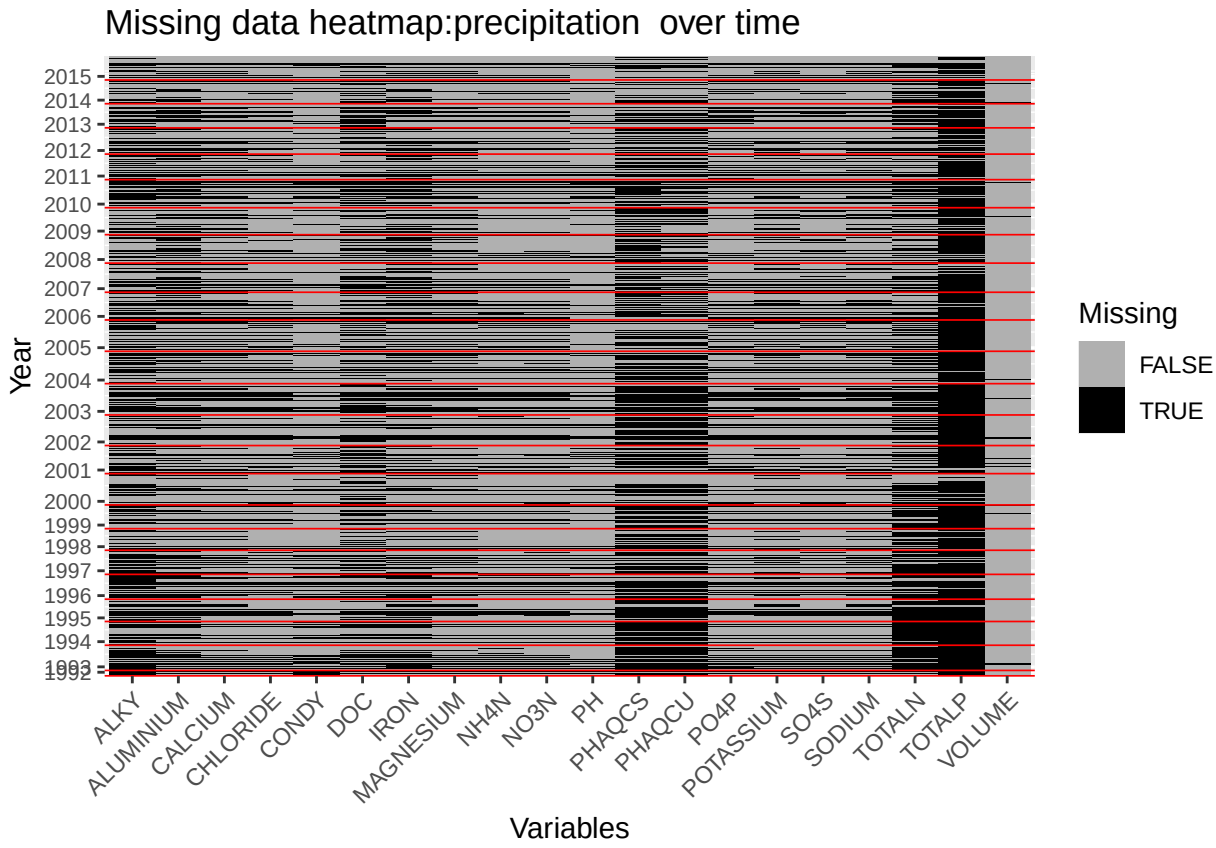
```
## Scale for y is already present.
## Adding another scale for y, which will replace the existing scale.
```




```r
#Same process is used for the other two data frames, copy pasted and changed the df labels
# Precipitation
df_precipitation_missing <- df_precipitation_wide %>%
  select(SDATE, where(is.numeric)) %>%
  pivot_longer(-SDATE, names_to = "Variables", values_to = "Value") %>%
  mutate(Missing = is.na(Value)) %>%
  arrange(SDATE) %>%
  mutate(Row = as.numeric(factor(SDATE, levels = unique(SDATE))))

year_breaks_precipitation <- df_precipitation_missing %>%
  distinct(SDATE, Row) %>%
  mutate(Year = as.numeric(format(SDATE, "%Y"))) %>%
  group_by(Year) %>%
  summarise(Row = min(Row), .groups = "drop")

ggplot(df_precipitation_missing, aes(x = Variables, y = Row, fill = Missing)) +
  geom_tile(color = NA) +
  scale_fill_manual(values = c("FALSE" = "grey69", "TRUE" = "black")) +
  scale_y_reverse() +
  geom_hline(data = year_breaks_precipitation, aes(yintercept = Row), color = "red", linewidth = 0.3) +
  scale_y_continuous(
    breaks = year_breaks_precipitation$Row + 10,
    labels = year_breaks_precipitation$Year,
    expand = c(0, 0)
  ) +
  labs(title = "Missing data heatmap:precipitation  over time",
       x = "Variables", y = "Year", fill = "Missing")+
```

```
  theme(axis.text.y = element_text(size = 8),
        axis.text.x = element_text(angle = 45, hjust = 1))
```

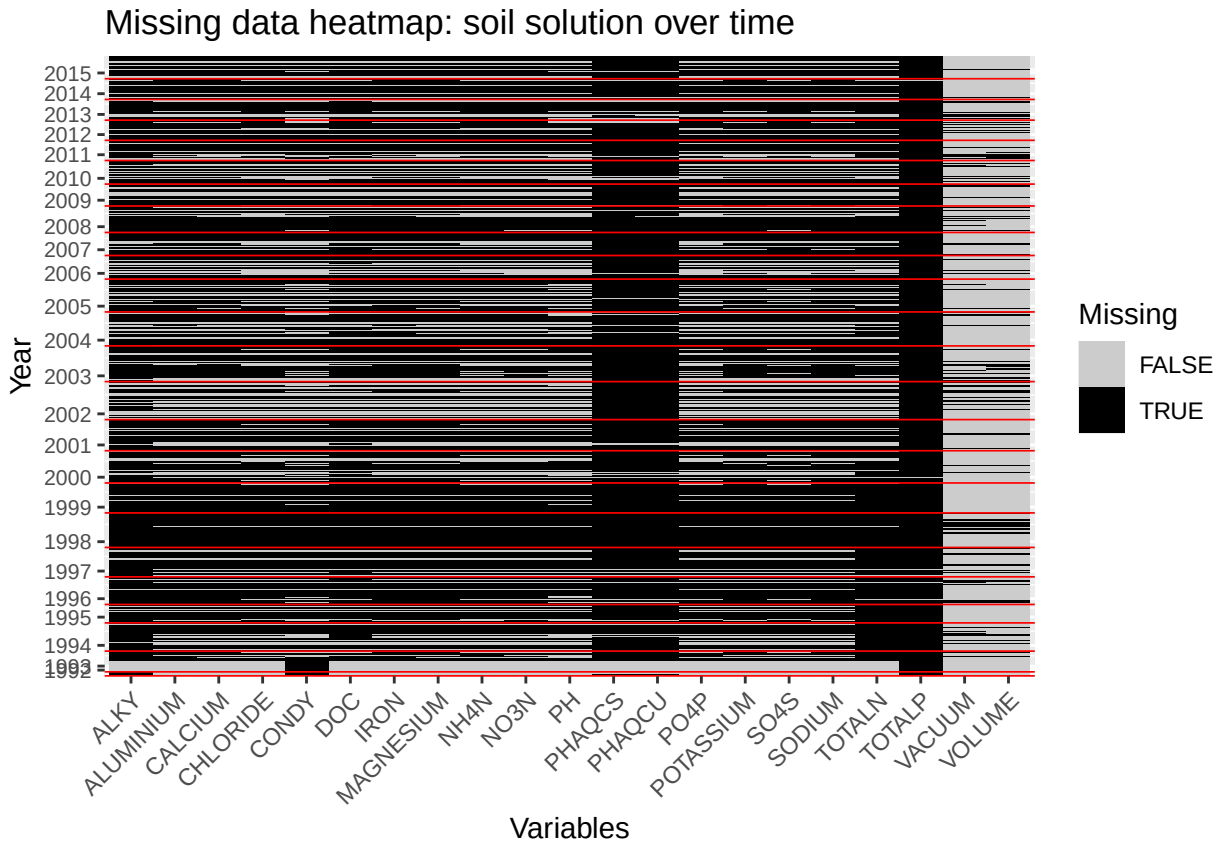
```
## Scale for y is already present.
## Adding another scale for y, which will replace the existing scale.
```




```
#Soil Solution
df_ss_missing <- df_soil_wide %>%
  select(SDATE, where(is.numeric)) %>%
  pivot_longer(-SDATE, names_to = "Variables", values_to = "Value") %>%
  mutate(Missing = is.na(Value)) %>%
  arrange(SDATE) %>%
  mutate(Row = as.numeric(factor(SDATE, levels = unique(SDATE))))

year_breaks_ss <- df_ss_missing %>%
  distinct(SDATE, Row) %>%
  mutate(Year = as.numeric(format(SDATE, "%Y"))) %>%
  group_by(Year) %>%
  summarise(Row = min(Row), .groups = "drop")

ggplot(df_ss_missing, aes(x = Variables, y = Row, fill = Missing)) +
  geom_tile(color = NA) +
  scale_fill_manual(values = c("FALSE" = "grey80", "TRUE" = "black")) +
  scale_y_reverse() +
  geom_hline(data = year_breaks_ss, aes(yintercept = Row), color = "red", linewidth = 0.3) +
  scale_y_continuous(
    breaks = year_breaks_ss$Row + 10,
```

```
    labels = year_breaks_ss$Year,
    expand = c(0, 0)
  ) +
  labs(title = "Missing data heatmap: soil solution over time",
       x = "Variables", y = "Year", fill = "Missing")+
  theme(axis.text.y = element_text(size = 8),
        axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Scale for y is already present.
## Adding another scale for y, which will replace the existing scale.
```



Stream water

Example code chunk showing some initial steps. This is generic example code chunks using unrelated data to illustrate the richness of plotting functions in R and show how you can arrange figures on a page using par().

Uniqueness: From the r code to check for the number of duplicated row, we can see that the dataframe has a uniqueness of 100%.

Consistency: consistent

Timeliness: The range is from the 6th of October 1992 to the 31st of December 2015.

From the Sample Frequency plot, we note that we may have different no. weeks in a month, thus resulting the fluctuation. The frequency started low, though slowly increasing after 1993. However, the frequency started to decrease in late 2000s.

Validity: Valid since no out of range value and false formatting was spotted.

Accuracy: Accurate as it matches reality

User needs and trade-offs: Policymakers may want long-term trends to help them make policies. Scientists may need long-term trends to study environmental change, which is critical with regards to climate change.

Completeness: Overall, the data is not that complete, though there is more grey than black. Overall, the completeness varies across the the years and variables, with PHAQCS (Aquacheck system pH stirred), PHAQCU (Aquacheck system pH unstirred), and TOTALP (Total dissolved phosphorus) barely having any completion, while it became less complete since the middle of 2012. However, one variable that is almost complete is PH (pH)

Precipitation

Uniqueness: From the r code to check for the number of duplicated row, we can see that the dataframe has a uniqueness of 100%.

Consistency: consistent

Timeliness: The range is from the 6th of October 1992 to the 31st of December 2015.

From the Sample Frequency plot, we note that we may have different no. weeks in a month, thus resulting the fluctuation. The frequency started low, though slowly increasing after 1993. However, the frequency started to decrease after 2010.

Validity: Less valid as there exist a negative (-0.32) for the variable of magnesium.

Accuracy: Accurate as it matches reality

User needs and trade-offs: Policymakers may want long-term trends to help them make policies. Scientists may need long-term trends to study environmental change, which is critical with regards to climate change.

Completeness: The data is relatively less complete compared to the stream water data, though there is more grey than black. Overall, the completeness varies across the the years and variables: ALKY (Alkalinity), DOC (Dissolved organic carbon), PHAQCS (Aquacheck system pH stirred), PHAQCU (Aquacheck system pH unstirred), TOTALN (Total nitrogen) and TOTALP (Total dissolved phosphorus) are the variables with the completion, while it is less complete during the first few years of the data, 2003 and, 2013. However, two sets of variables are almost complete, which are: PH (pH) and VOLUME (Volume of sample collected)

Soil solution

Uniqueness: From the r code to check for the number of duplicated row, we can see that the dataframe has a uniqueness of 100%.

Consistency: consistent

Timeliness: The range is from the 6th of October 1992 to the 23rd of December 2015, which is valid even though it ends earlier compared to the previous dataframes, since this is only sampled fortnightly.

From the Sample Frequency plot, we note that we may have different no. weeks in a month, thus resulting the fluctuation. The frequency started low, though slowly increasing after 1995. However, the frequency started to decrease after 2010.

Validity: Valid since no out of range value and false formatting was spotted.

Accuracy: Accurate as it matches reality

User needs and trade-offs: Policymakers may want long-term trends to help them make policies. Scientists may need long-term trends to study environmental change, which is critical with regards to climate change.

Completeness: The data is way less complete in comparison to the previous two comparison, with more black (missing data) than grey (completed data). Overall, the completeness varies across the the years and variables: with PHAQCS (Aquacheck system pH stirred), PHAQCU (Aquacheck system pH unstirred), and TOTALP (Total dissolved phosphorus) barely having any completion, while it has little completion in the

second half of 1990s, 2001, 2003, 2008 and 2014-2015. However, two sets of variables are almost complete, which are: VACUUM (Residual vacuum at time of sampling) and VOLUME (Volume of sample collected)

Step 9: Exploratory Data Analysis (EDA)

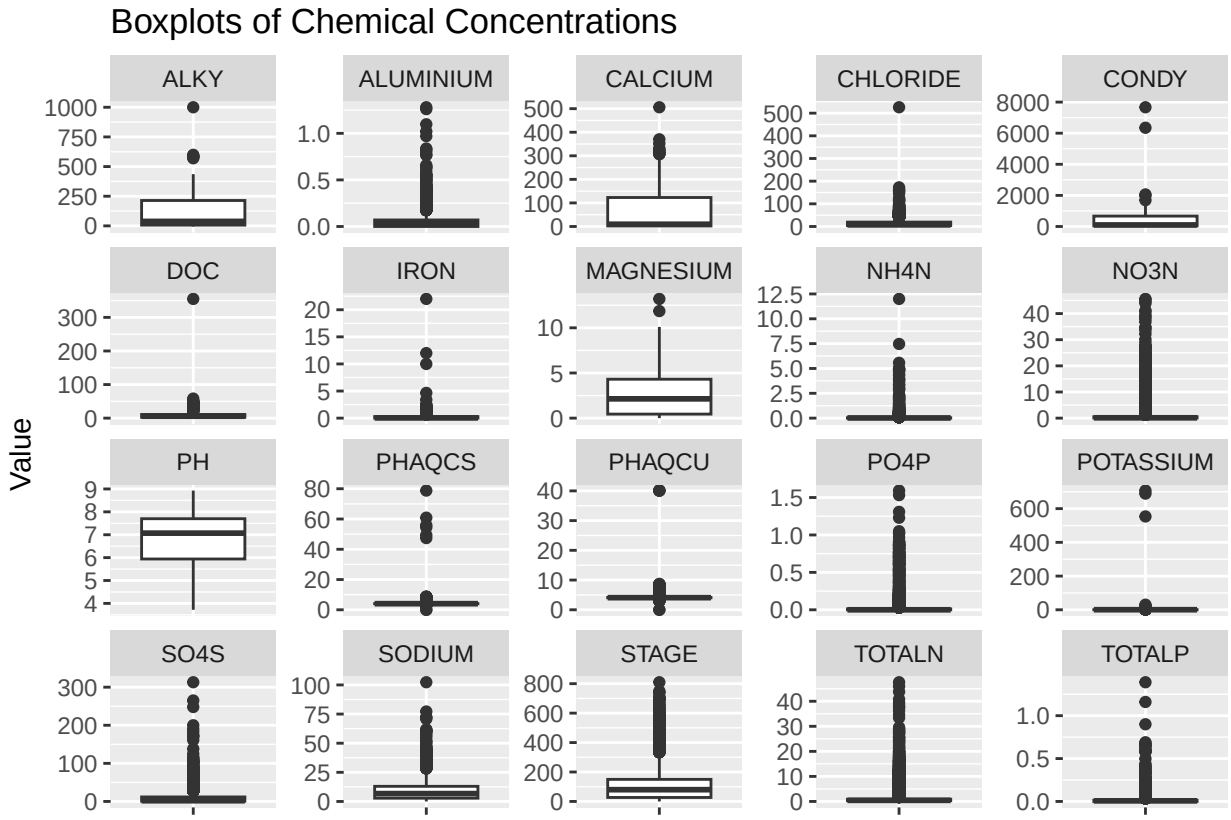
Stream water

```
df_stream_wide %>%
  select(where(is.numeric), -LCODE) %>%
  summarise(across(everything(), list(
    mean = ~mean(.x, na.rm = TRUE), #calculate mean excluding
    sd = ~sd(.x, na.rm = TRUE),
    var = ~var(.x, na.rm = TRUE)
), .names = "{.col}_{.fn}"))
```

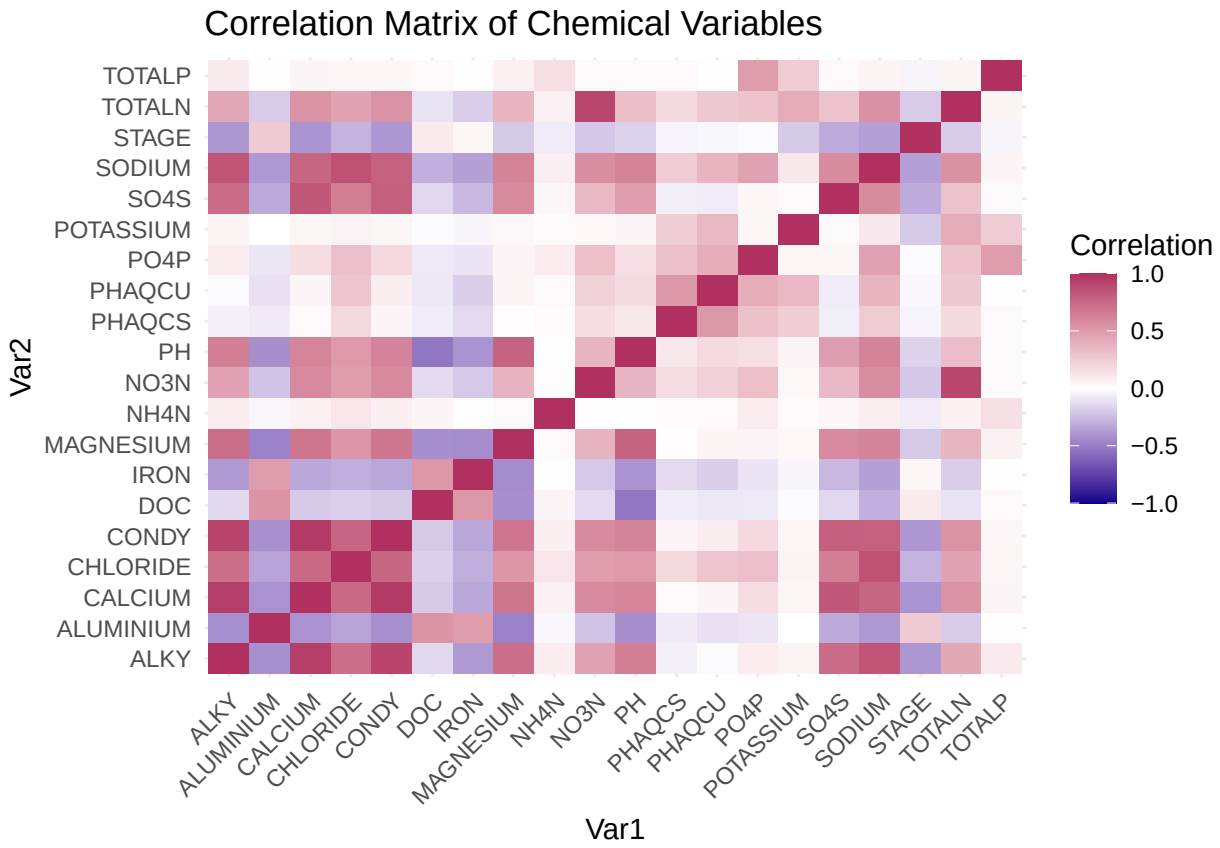
```
## # A tibble: 1 x 60
##   ALUMINIUM_mean ALUMINIUM_sd ALUMINIUM_var TOTALN_mean TOTALN_sd TOTALN_var
##            <dbl>        <dbl>         <dbl>       <dbl>     <dbl>      <dbl>
## 1         0.0484       0.0630       0.00397        1.69      3.36       11.3
## # i 54 more variables: CHLORIDE_mean <dbl>, CHLORIDE_sd <dbl>,
## #   CHLORIDE_var <dbl>, DOC_mean <dbl>, DOC_sd <dbl>, DOC_var <dbl>,
## #   IRON_mean <dbl>, IRON_sd <dbl>, IRON_var <dbl>, MAGNESIUM_mean <dbl>,
## #   MAGNESIUM_sd <dbl>, MAGNESIUM_var <dbl>, NH4N_mean <dbl>, NH4N_sd <dbl>,
## #   NH4N_var <dbl>, NO3N_mean <dbl>, NO3N_sd <dbl>, NO3N_var <dbl>,
## #   PH_mean <dbl>, PH_sd <dbl>, PH_var <dbl>, PO4P_mean <dbl>, PO4P_sd <dbl>,
## #   PO4P_var <dbl>, POTASSIUM_mean <dbl>, POTASSIUM_sd <dbl>, ...
```

```
df_stream_tall <- df_stream_wide %>%
  select(SDATE, where(is.numeric), -LCODE) %>%
  pivot_longer(-SDATE, names_to = "Chemical_Variables", values_to = "Value")

ggplot(df_stream_tall, aes(x = " ", y = Value)) +
  geom_boxplot() +
  facet_wrap(~ Chemical_Variables, scales = "free_y") +
  labs(title = "Boxplots of Chemical Concentrations",
       x = NULL, y = "Value")
```

```
## Warning: Removed 93716 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

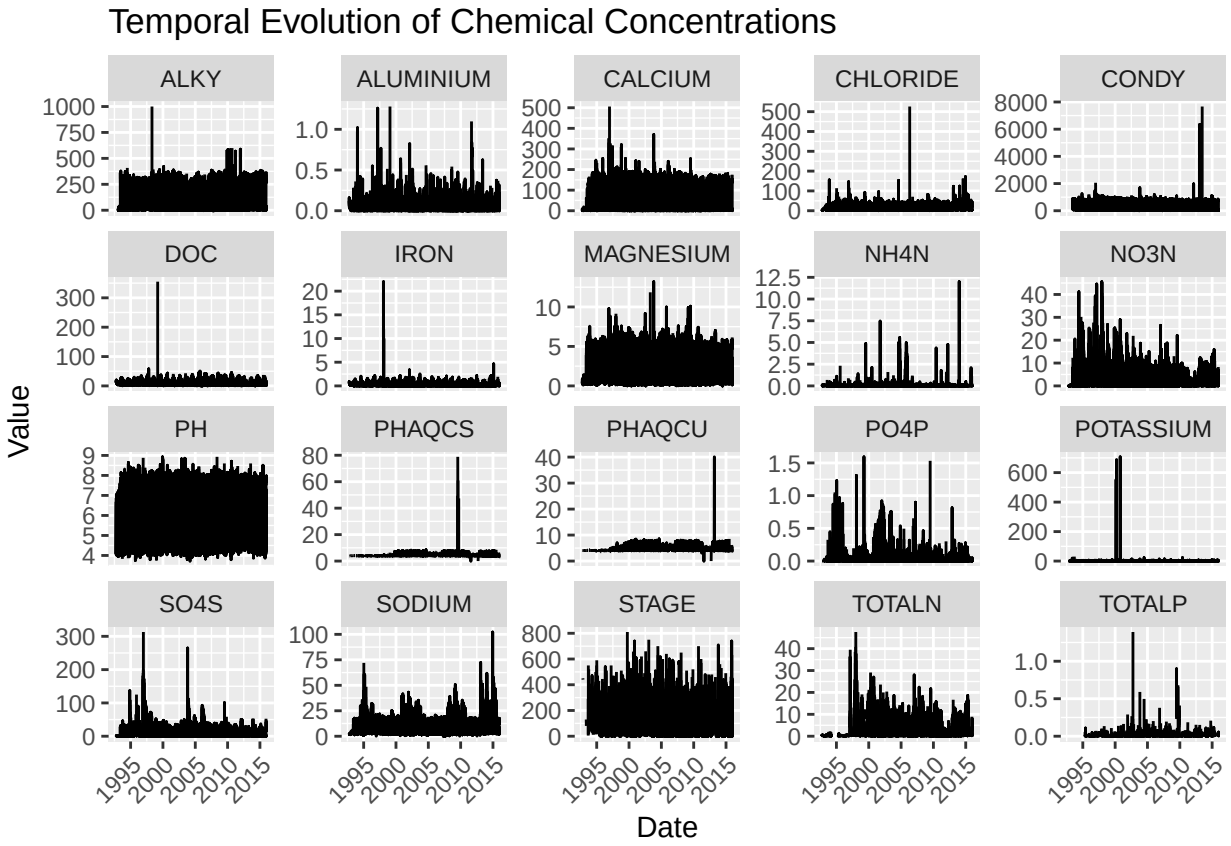



```r
# Calculate the correlation matrix
cormat_stream <- df_stream_wide %>%
  select(where(is.numeric), -LCODE) %>%
  cor(use = "pairwise.complete.obs")

# Reshape the correlation matrix into a long format using pivot_longer
cor_long_stream <- as.data.frame(cormat_stream) %>%
  rownames_to_column(var = "Var1") %>%
  pivot_longer(cols = -Var1, names_to = "Var2", values_to = "value")

# Plot the heatmap
ggplot(cor_long_stream, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "darkblue", high = "maroon", midpoint = 0, limit = c(-1, 1)) +
  labs(title = "Correlation Matrix of Chemical Variables",
       fill = "Correlation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



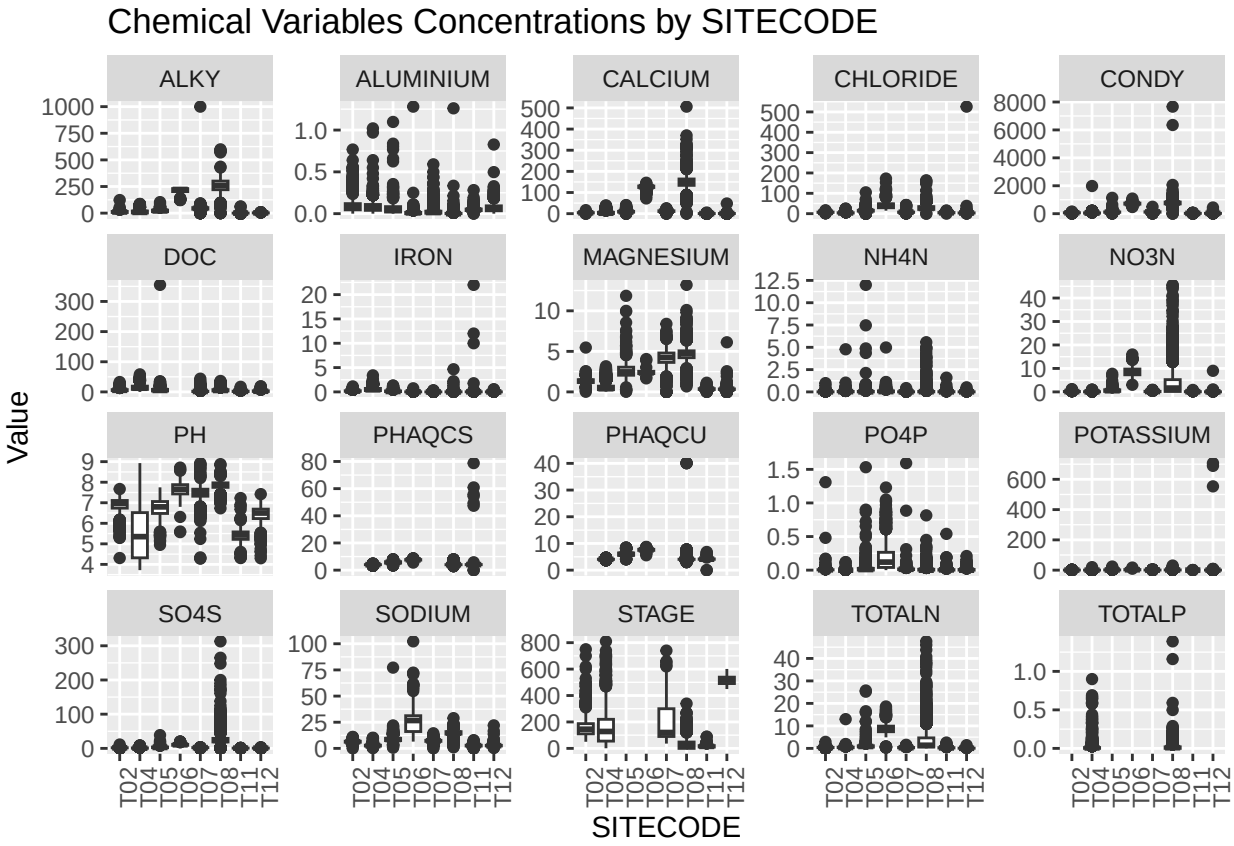

```
ggplot(df_stream_tall, aes(x = as.Date(SDATE, format = "%d-%b-%y"), y = Value)) +
  geom_line(aes(group = 1)) +
  facet_wrap(~ Chemical_Variables, scales = "free_y") +
  labs(title = "Temporal Evolution of Chemical Concentrations", x = "Date") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 51 rows containing missing values or values outside the scale range
## (`geom_line()`).
```




```
df_stream_tall_spatial <- df_stream_wide %>%
  select(SDATE, SITECODE, where(is.numeric), -LCODE) %>%
  pivot_longer(cols = -c(SDATE, SITECODE), names_to = "Chemical", values_to = "Value")

ggplot(df_stream_tall_spatial, aes(x = SITECODE, y = Value)) +
  geom_boxplot() +
  facet_wrap(~ Chemical, scales = "free_y") +
  labs(title = "Chemical Variables Concentrations by SITECODE") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

```
## Warning: Removed 93716 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



From the numerical summaries and visualisations, most chemical variables have a median close to 0 (except pH), and have many outliers.

Precipitation

Similar as stream water

Soil solution

Similar as stream water