ST117 Individual DRAFT Written Report - Task A

My WARWICK ID 5645242 (Report Pod 041)

2025-04-08

Setting up data frames before for phase 2 part A

```
#Import the dataframes from phase 1
df_stream_wide <- readRDS("/Users/danielguo/Desktop/University/Year 1/ST117/5645242_041_WR/df_stream_wide.rds")
df_precipitation_wide <- readRDS("/Users/danielguo/Desktop/University/Year 1/ST117/5645242_041_WR/df_precipitati
df_soil_wide <- readRDS("/Users/danielguo/Desktop/University/Year 1/ST117/5645242_041_WR/df_soil_wide.rds")

#Averaging the repeated sampling by the variable LCODE
df_stream_avg <- df_stream_wide %>%
  group_by(SDATE, SITECODE, LCODE) %>%
  summarise(across(where(is.numeric), mean, na.rm = TRUE), .groups = "drop")
```

```
## Warning: There was 1 warning in `summarise()`.
## i In argument: `across(where(is.numeric), mean, na.rm = TRUE)`.
## i In group 1: `SDATE = 1992-10-06`, `SITECODE = "T04"`, `LCODE = 1`.
## Caused by warning:
## ! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.
## Supply arguments directly to `.fns` through an anonymous function instead.
##
##   # Previously
##   across(a:b, mean, na.rm = TRUE)
##
##   # Now
##   across(a:b, \(x) mean(x, na.rm = TRUE))
```

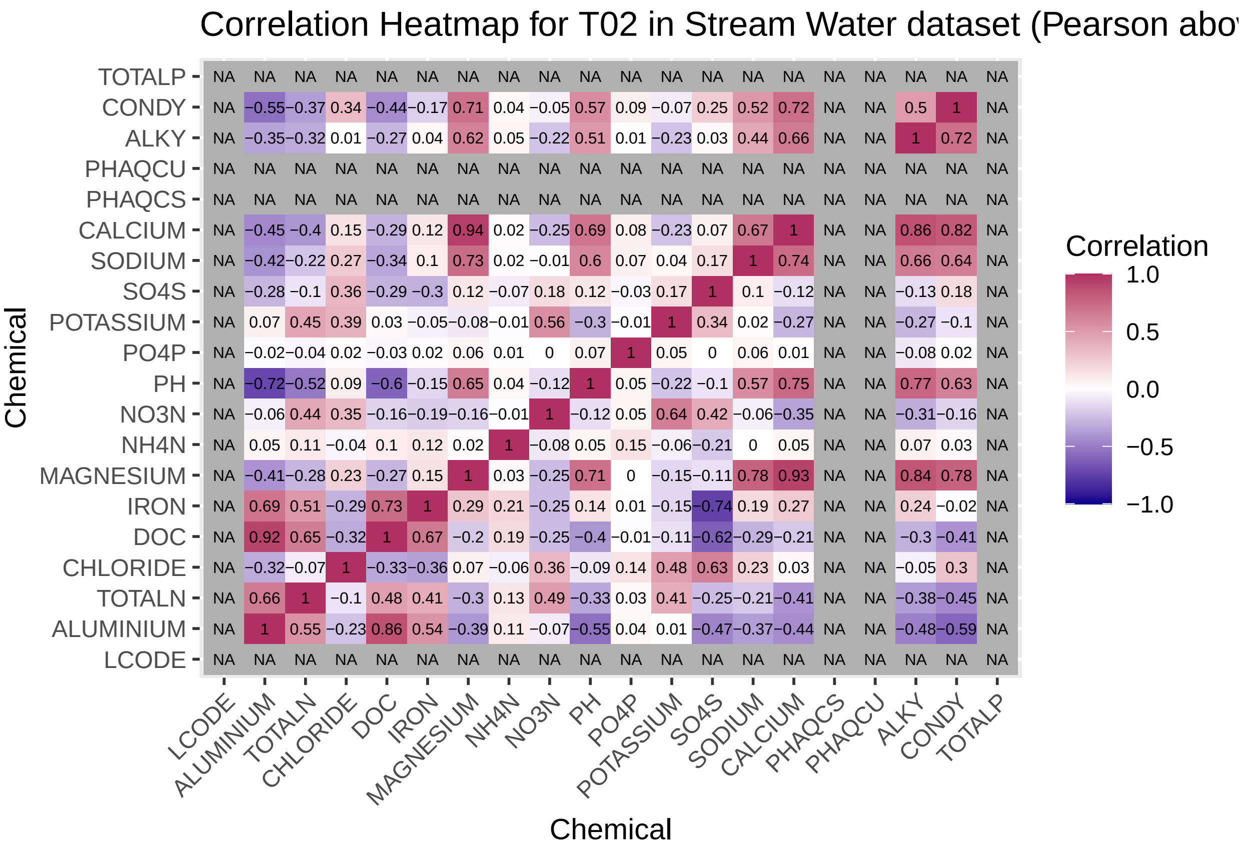
```
#weighted averaging considering the sampling volume (VOLUME) by the variable RID
df_soil_avg <- df_soil_wide %>%
  group_by(SDATE, SITECODE, RID) %>%
  summarise(across(where(is.numeric), ~ weighted.mean(., w = VOLUME, na.rm = TRUE)), .groups = "drop")

#STAGE, VACUUM, and VOLUME are only contained in some of the datasets and may be dropped unless needed.
df_stream_avg <- df_stream_avg %>% select(-STAGE)
df_precipitation_wide <- df_precipitation_wide %>% select(-VOLUME)
df_soil_avg <- df_soil_avg %>% select(-VACUUM, -VOLUME)

#filter out the assigned data for our pod: 2002-2008: T02, T04, T06
df_stream_Afiltered <- df_stream_avg %>%
  filter(format(SDATE, "%Y") %in% 2002:2008,
         SITECODE %in% c("T02", "T04", "T06"))
df_precipitation_Afiltered <- df_precipitation_wide %>%
  filter(format(SDATE, "%Y") %in% 2002:2008,
         SITECODE %in% c("T02", "T04", "T06"))
df_soil_Afiltered <- df_soil_avg %>%
  filter(format(SDATE, "%Y") %in% 2002:2008,
         SITECODE %in% c("T02", "T04", "T06"))
```

Question 1

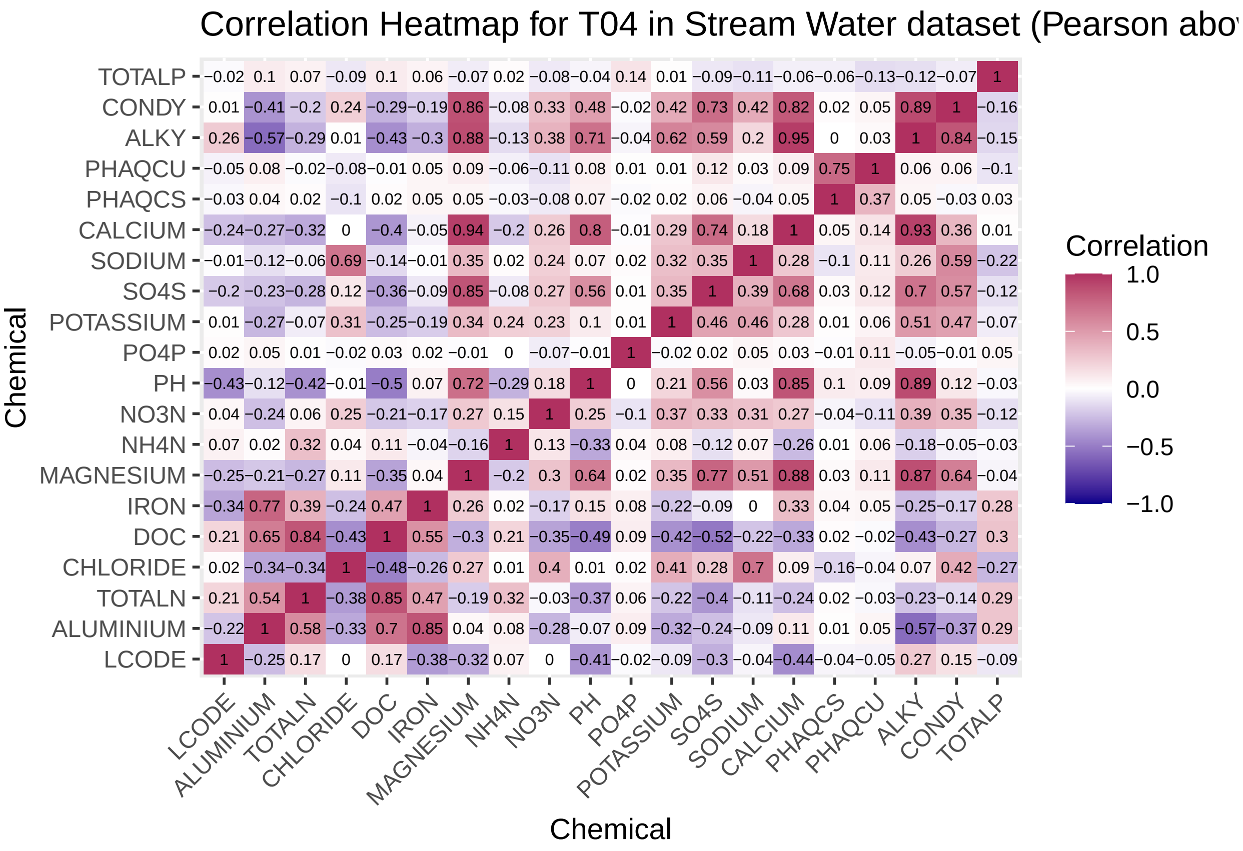
Question 2

```
plot_and_find_differences <- function(sitecode, df, dataname = "data") {
  # Filter for the selected site
  df_site <- df %>%
    filter(SITECODE == sitecode) %>%
    ungroup() %>%
    select(where(is.numeric))

  #Pearson and Spearman correlation matrices
  correlation_pearson <- suppressWarnings(round(cor(df_site, method = "pearson", use = "pairwise.complete.obs"),
  correlation_spearman <- suppressWarnings(round(cor(df_site, method = "spearman", use = "pairwise.complete.obs'

  # Combine: Pearson upper triangle, Spearman lower triangle
  correlation_combined <- correlation_pearson
  correlation_combined[lower.tri(correlation_combined)] <- correlation_spearman[lower.tri(correlation_spearman)]

  # Long format
  correlation_long <- as.data.frame(as.table(correlation_combined)) %>%
    rename(Variables1 = Var1, Variables2 = Var2, Correlation = Freq)

  # plot
  heatmap_plot <- ggplot(correlation_long, aes(Variables1, Variables2, fill = Correlation)) +
    geom_tile() +
    scale_fill_gradient2(low = "darkblue", high = "maroon",
                         midpoint = 0, limit = c(-1, 1), na.value = "grey69") +
    geom_text(aes(label = ifelse(is.na(Correlation), "NA", round(Correlation, 2))), size = 2) + #adjusting to th
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    labs(
      title = paste("Correlation Heatmap for", sitecode, "in", dataname, "dataset (Pearson above, Spearman below
      x = "Chemical", y = "Chemical"
    )
   # identify large absolute differences bigger than 0,5
  abs_diff <- abs(abs(correlation_pearson) - abs(correlation_spearman))
  diff_table <- as.data.frame(as.table(abs_diff)) %>%
    rename(Variables1 = Var1, Variables2 = Var2, AbsDiff = Freq) %>%
    filter(AbsDiff > 0.5) %>%
    mutate(Site = sitecode) %>%
    arrange(desc(AbsDiff))

  return(list(
    plot = heatmap_plot,
    differences_table = diff_table
  ))
}

result_T02_stream <- plot_and_find_differences("T02", df_stream_Afiltered, "Stream Water")
result_T04_stream <- plot_and_find_differences("T04", df_stream_Afiltered, "Stream Water")
result_T06_stream <- plot_and_find_differences("T06", df_stream_Afiltered, "Stream Water")
result_T02_precipitation <- plot_and_find_differences("T02", df_precipitation_Afiltered, "Precipitation")
result_T04_precipitation <- plot_and_find_differences("T04", df_precipitation_Afiltered, "Precipitation")
result_T06_precipitation <- plot_and_find_differences("T06", df_precipitation_Afiltered, "Precipitation")
result_T02_soil <- plot_and_find_differences("T02", df_soil_Afiltered, "Soil Solution")
result_T04_soil <- plot_and_find_differences("T04", df_soil_Afiltered, "Soil Solution")
result_T06_soil <- plot_and_find_differences("T06", df_soil_Afiltered, "Soil Solution")


# View the plot
```

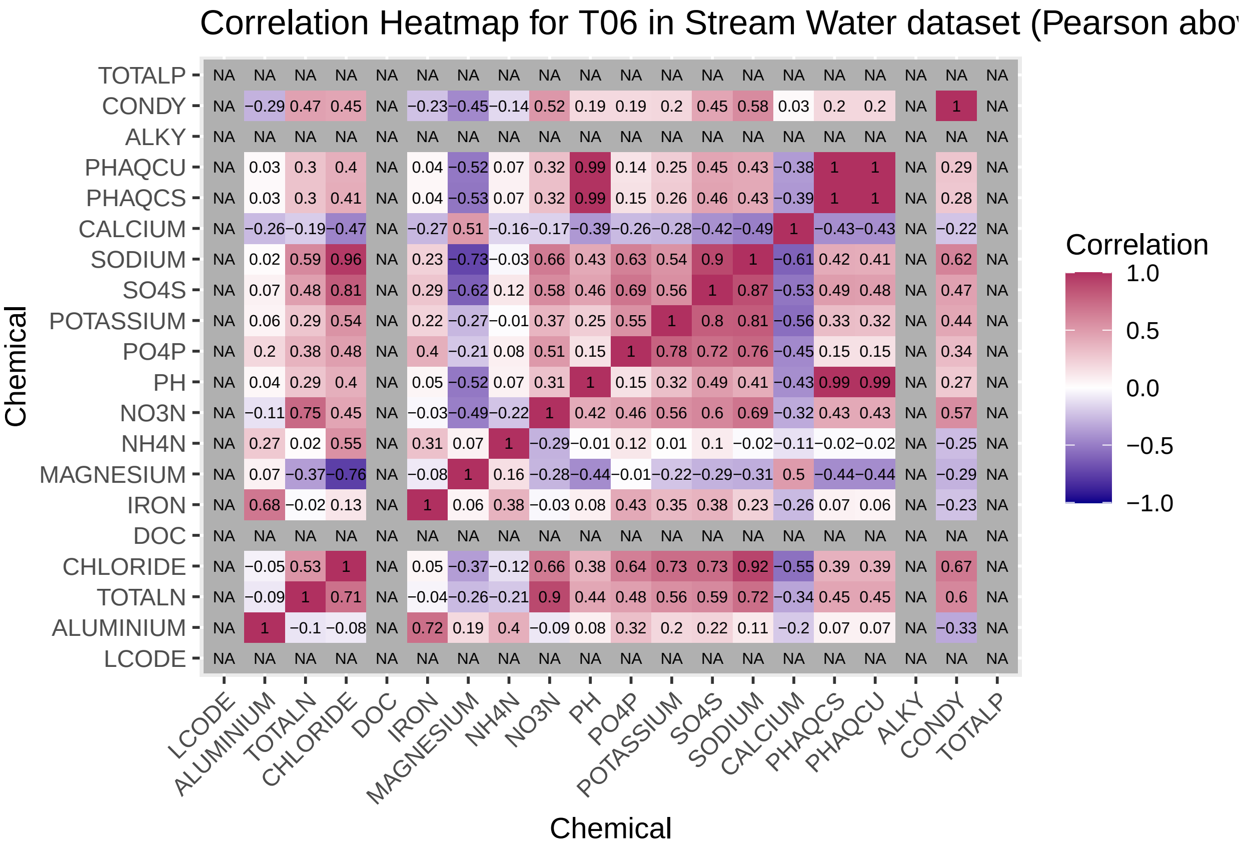
```
print(result_T02_stream$plot)
```



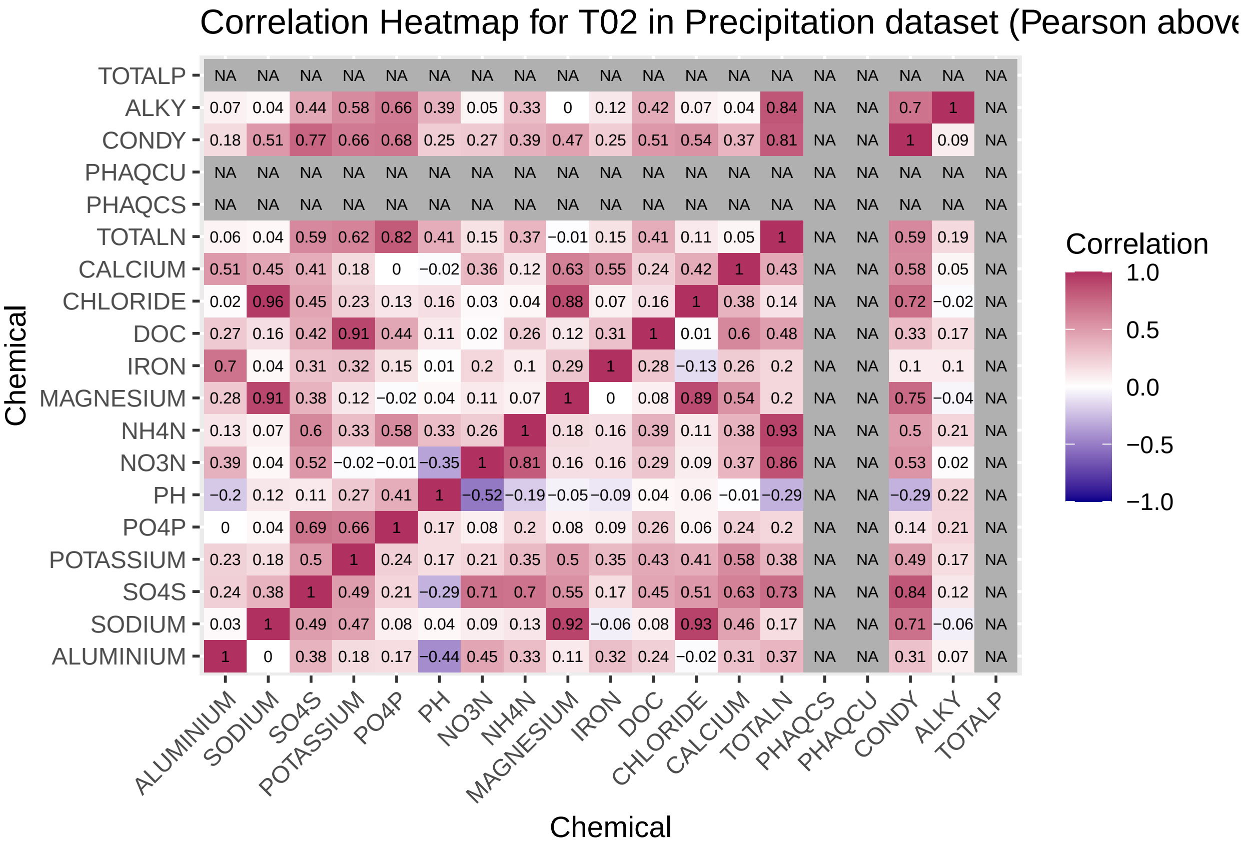
```
print(result_T04_stream$plot)
```



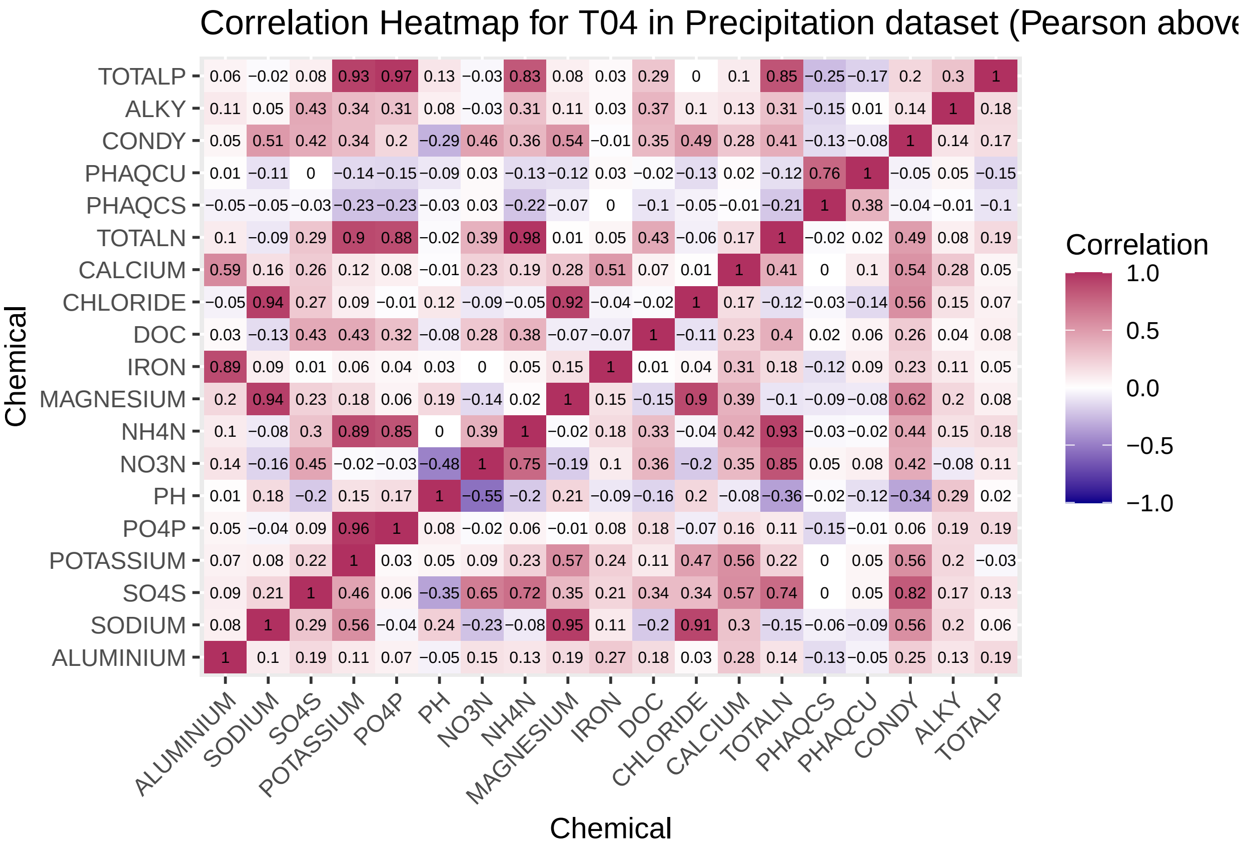
```
print(result_T06_stream$plot)
```



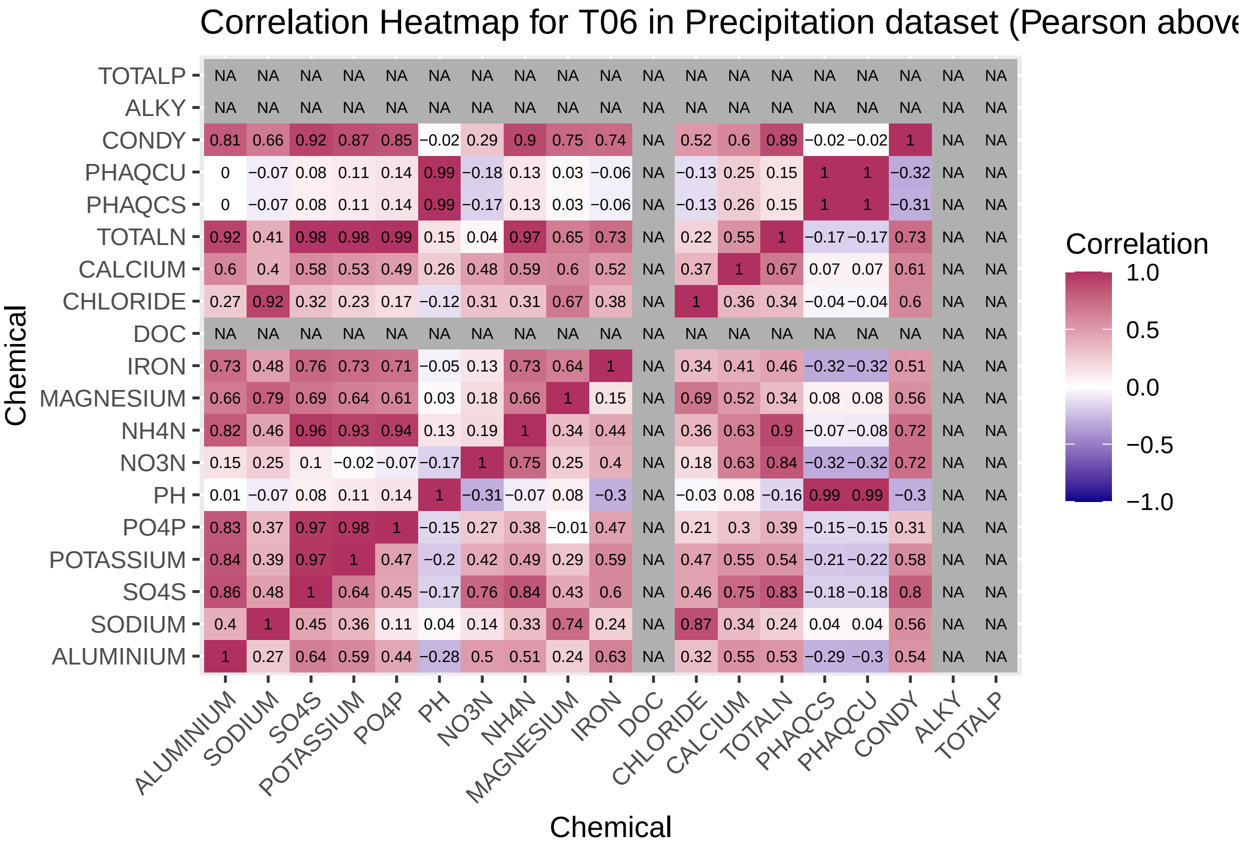
```r
print(result_T02_precipitation$plot)
```



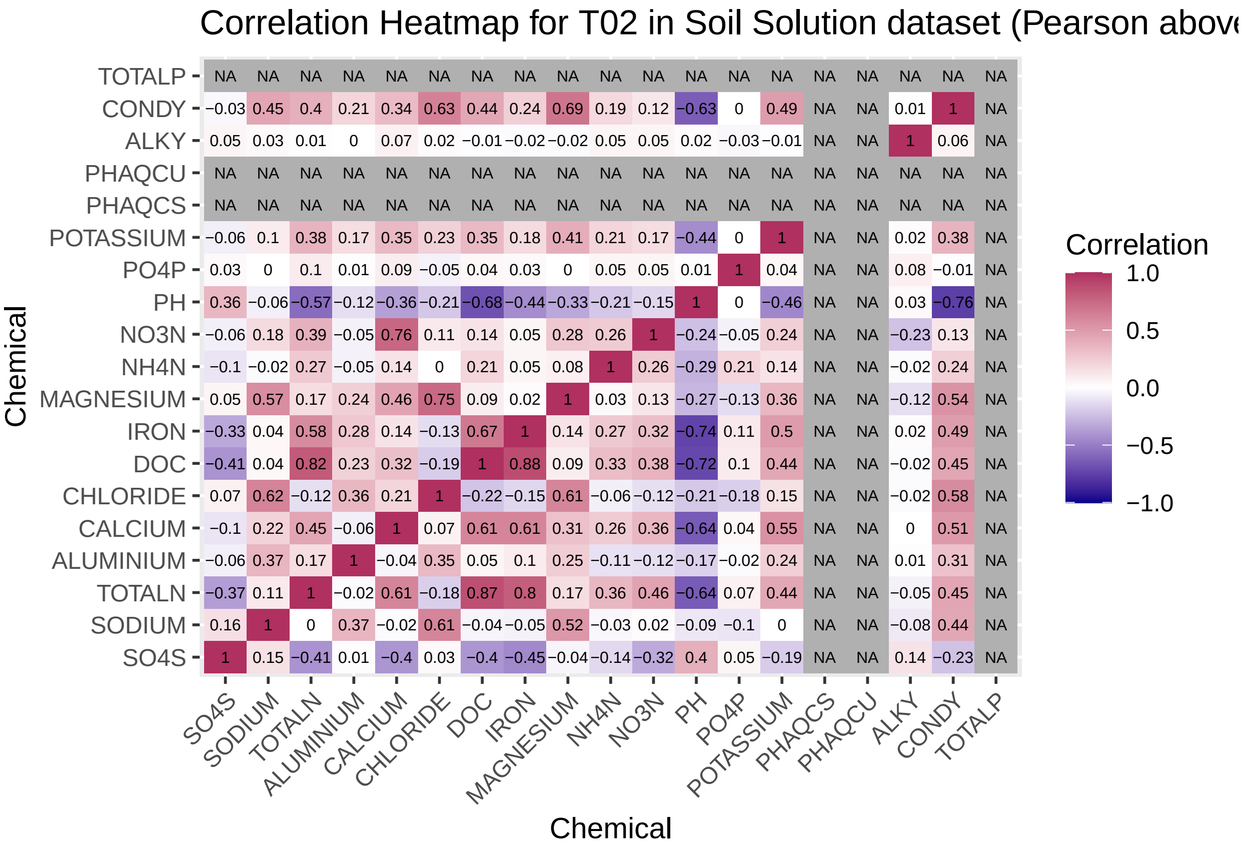
```r
print(result_T04_precipitation$plot)
```



```
print(result_T06_precipitation$plot)
```

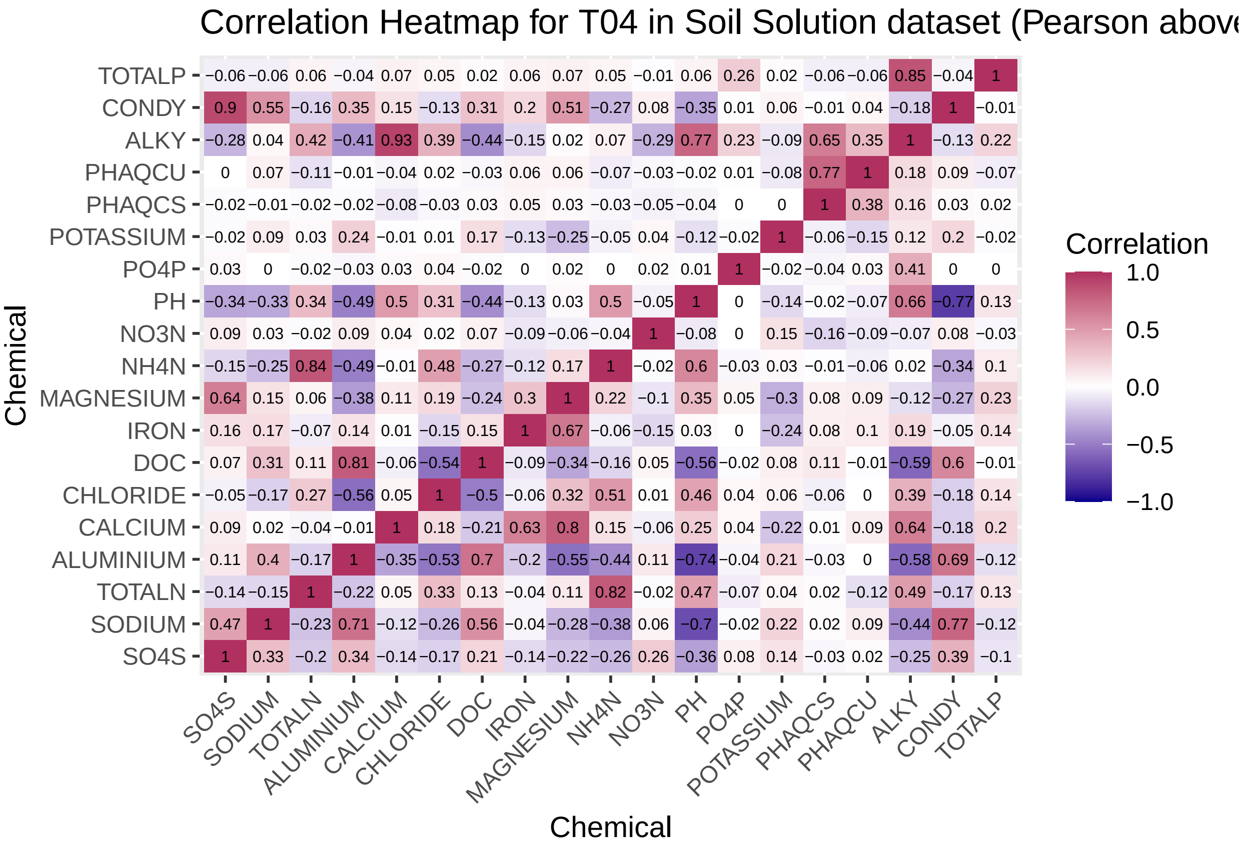


```
print(result_T02_soil$plot)
```



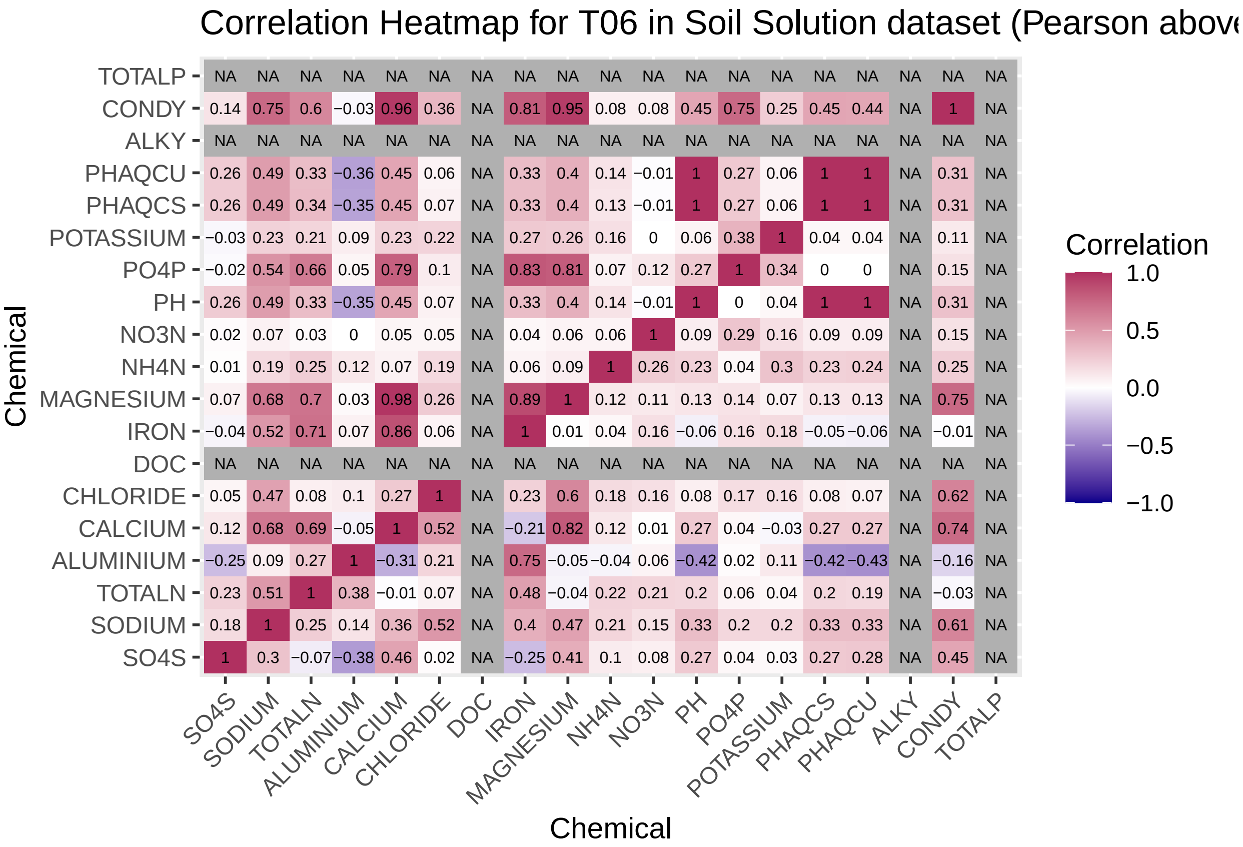
Chemical	SO4S	SODIUM	TOTALN	ALUMINIUM	CALCIUM	CHLORIDE	DOC	IRON	MAGNESIUM	NH4N	NO3N	PH	PO4P	POTASSIUM	PHAQCS	PHAQCU	ALKY	CONDY	TOTALP
TOTALP	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CONDY	−0.03	0.45	0.4	0.21	0.34	0.63	0.44	0.24	0.69	0.19	0.12	−0.63	0	0.49	NA	NA	0.01	1	NA
ALKY	0.05	0.03	0.01	0	0.07	0.02	−0.01	−0.02	−0.02	0.05	0.05	0.02	−0.03	−0.01	NA	NA	1	0.06	NA
PHAQCU	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PHAQCS	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
POTASSIUM	−0.06	0.1	0.38	0.17	0.35	0.23	0.35	0.18	0.41	0.21	0.17	−0.44	0	1	NA	NA	0.02	0.38	NA
PO4P	0.03	0	0.1	0.01	0.09	−0.05	0.04	0.03	0	0.05	0.05	0.01	1	0.04	NA	NA	0.08	−0.01	NA
PH	0.36	−0.06	−0.57	−0.12	−0.36	−0.21	−0.68	−0.44	−0.33	−0.21	−0.15	1	0	−0.46	NA	NA	0.03	−0.76	NA
NO3N	−0.06	0.18	0.39	−0.05	0.76	0.11	0.14	0.05	0.28	0.26	1	−0.24	−0.05	0.24	NA	NA	−0.23	0.13	NA
NH4N	−0.1	−0.02	0.27	−0.05	0.14	0	0.21	0.05	0.08	1	0.26	−0.29	0.21	0.14	NA	NA	−0.02	0.24	NA
MAGNESIUM	0.05	0.57	0.17	0.24	0.46	0.75	0.09	0.02	1	0.03	0.13	−0.27	−0.13	0.36	NA	NA	−0.12	0.54	NA
IRON	−0.33	0.04	0.58	0.28	0.14	−0.13	0.67	1	0.14	0.27	0.32	−0.74	0.11	0.5	NA	NA	0.02	0.49	NA
DOC	−0.41	0.04	0.82	0.23	0.32	−0.19	1	0.88	0.09	0.33	0.38	−0.72	0.1	0.44	NA	NA	−0.02	0.45	NA
CHLORIDE	0.07	0.62	−0.12	0.36	0.21	1	−0.22	−0.15	0.61	−0.06	−0.12	−0.21	−0.18	0.15	NA	NA	−0.02	0.58	NA
CALCIUM	−0.1	0.22	0.45	−0.06	1	0.07	0.61	0.61	0.31	0.26	0.36	−0.64	0.04	0.55	NA	NA	0	0.51	NA
ALUMINIUM	−0.06	0.37	0.17	1	−0.04	0.35	0.05	0.1	0.25	−0.11	−0.12	−0.17	−0.02	0.24	NA	NA	0.01	0.31	NA
TOTALN	−0.37	0.11	1	−0.02	0.61	−0.18	0.87	0.8	0.17	0.36	0.46	−0.64	0.07	0.44	NA	NA	−0.05	0.45	NA
SODIUM	0.16	1	0	0.37	−0.02	0.61	−0.04	−0.05	0.52	−0.03	0.02	−0.09	−0.1	0	NA	NA	−0.08	0.44	NA
SO4S	1	0.15	−0.41	0.01	−0.4	0.03	−0.4	−0.45	−0.04	−0.14	−0.32	0.4	0.05	−0.19	NA	NA	0.14	−0.23	NA

```
print(result_T04_soil$plot)
```



Chemical	SO4S	SODIUM	TOTALN	ALUMINIUM	CALCIUM	CHLORIDE	DOC	IRON	MAGNESIUM	NH4N	NO3N	PH	PO4P	POTASSIUM	PHAQCS	PHAQCU	ALKY	CONDY	TOTALP
TOTALP	−0.06	−0.06	0.06	−0.04	0.07	0.05	0.02	0.06	0.07	0.05	−0.01	0.06	0.26	0.02	−0.06	−0.06	0.85	−0.04	1
CONDY	0.9	0.55	−0.16	0.35	0.15	−0.13	0.31	0.2	0.51	−0.27	0.08	−0.35	0.01	0.06	−0.01	0.04	−0.18	1	−0.01
ALKY	−0.28	0.04	0.42	−0.41	0.93	0.39	−0.44	−0.15	0.02	0.07	−0.29	0.77	0.23	−0.09	0.65	0.35	1	−0.13	0.22
PHAQCU	0	0.07	−0.11	−0.01	−0.04	0.02	−0.03	0.06	0.06	−0.07	−0.03	−0.02	0.01	−0.08	0.77	1	0.18	0.09	−0.07
PHAQCS	−0.02	−0.01	−0.02	−0.02	−0.08	−0.03	0.03	0.05	0.03	−0.03	−0.05	−0.04	0	0	1	0.38	0.16	0.03	0.02
POTASSIUM	−0.02	0.09	0.03	0.24	−0.01	0.01	0.17	−0.13	−0.25	−0.05	0.04	−0.12	−0.02	1	−0.06	−0.15	0.12	0.2	−0.02
PO4P	0.03	0	−0.02	−0.03	0.03	0.04	−0.02	0	0.02	0	0.02	0.01	1	−0.02	−0.04	0.03	0.41	0	0
PH	−0.34	−0.33	0.34	−0.49	0.5	0.31	−0.44	−0.13	0.03	0.5	−0.05	1	0	−0.14	−0.02	−0.07	0.66	−0.77	0.13
NO3N	0.09	0.03	−0.02	0.09	0.04	0.02	0.07	−0.09	−0.06	−0.04	1	−0.08	0	0.15	−0.16	−0.09	−0.07	0.08	−0.03
NH4N	−0.15	−0.25	0.84	−0.49	−0.01	0.48	−0.27	−0.12	0.17	1	−0.02	0.6	−0.03	0.03	−0.01	−0.06	0.02	−0.34	0.1
MAGNESIUM	0.64	0.15	0.06	−0.38	0.11	0.19	−0.24	0.3	1	0.22	−0.1	0.35	0.05	−0.3	0.08	0.09	−0.12	−0.27	0.23
IRON	0.16	0.17	−0.07	0.14	0.01	−0.15	0.15	1	0.67	−0.06	−0.15	0.03	0	−0.24	0.08	0.1	0.19	−0.05	0.14
DOC	0.07	0.31	0.11	0.81	−0.06	−0.54	1	−0.09	−0.34	−0.16	0.05	−0.56	−0.02	0.08	0.11	−0.01	−0.59	0.6	−0.01
CHLORIDE	−0.05	−0.17	0.27	−0.56	0.05	1	−0.5	−0.06	0.32	0.51	0.01	0.46	0.04	0.06	−0.06	0	0.39	−0.18	0.14
CALCIUM	0.09	0.02	−0.04	−0.01	1	0.18	−0.21	0.63	0.8	0.15	−0.06	0.25	0.04	−0.22	0.01	0.09	0.64	−0.18	0.2
ALUMINIUM	0.11	0.4	−0.17	1	−0.35	−0.53	0.7	−0.2	−0.55	−0.44	0.11	−0.74	−0.04	0.21	−0.03	0	−0.58	0.69	−0.12
TOTALN	−0.14	−0.15	1	−0.22	0.05	0.33	0.13	−0.04	0.11	0.82	−0.02	0.47	−0.07	0.04	0.02	−0.12	0.49	−0.17	0.13
SODIUM	0.47	1	−0.23	0.71	−0.12	−0.26	0.56	−0.04	−0.28	−0.38	0.06	−0.7	−0.02	0.22	0.02	0.09	−0.44	0.77	−0.12
SO4S	1	0.33	−0.2	0.34	−0.14	−0.17	0.21	−0.14	−0.22	−0.26	0.26	−0.36	0.08	0.14	−0.03	0.02	−0.25	0.39	−0.1

```
print(result_T06_soil$plot)
```



Question 3

```
# View the table of large differences bigger than 0.5 for all sites for each data:
kable(result_T02_stream$differences_table, caption = "Large Pearson-Spearman Differences > 0.5 for T02 for Strea
```

Table 1: Large Pearson–Spearman Differences > 0.5 for T02 for Stream Water data

Variables1	Variables2	AbsDiff	Site

```
kable(result_T04_stream$differences_table, caption = "Large Pearson-Spearman Differences > 0.5 for T04 for Strea
```

Table 2: Large Pearson–Spearman Differences > 0.5 for T04 for Stream Water data

Variables1	Variables2	AbsDiff	Site

```
kable(result_T06_stream$differences_table, caption = "Large Pearson-Spearman Differences > 0.5 for T06 for Strea
```

Table 3: Large Pearson–Spearman Differences > 0.5 for T06 for Stream Water data

Variables1	Variables2	AbsDiff	Site

```
kable(result_T02_precipitation$differences_table, caption = "Large Pearson-Spearman Differences > 0.5 for T02 fo
```

Table 4: Large Pearson–Spearman Differences > 0.5 for T02 for Precipitation data

Variables1	Variables2	AbsDiff	Site
TOTALN	NO3N	0.71	T02
NO3N	TOTALN	0.71	T02
ALKY	TOTALN	0.65	T02
TOTALN	ALKY	0.65	T02
TOTALN	PO4P	0.62	T02
PO4P	TOTALN	0.62	T02
ALKY	CONDY	0.61	T02
CONDY	ALKY	0.61	T02
TOTALN	NH4N	0.56	T02
NH4N	TOTALN	0.56	T02
NH4N	NO3N	0.55	T02
NO3N	NH4N	0.55	T02
CONDY	PO4P	0.54	T02
PO4P	CONDY	0.54	T02

```r
kable(result_T04_precipitation$differences_table, caption = "Large Pearson-Spearman Differences > 0.5 for T04 fo
```

Table 5: Large Pearson–Spearman Differences > 0.5 for T04 for Precipitation data

Variables1	Variables2	AbsDiff	Site
PO4P	POTASSIUM	0.93	T04
POTASSIUM	PO4P	0.93	T04
TOTALP	POTASSIUM	0.90	T04
POTASSIUM	TOTALP	0.90	T04
NH4N	PO4P	0.79	T04
PO4P	NH4N	0.79	T04
TOTALP	PO4P	0.78	T04
PO4P	TOTALP	0.78	T04
TOTALN	PO4P	0.77	T04
PO4P	TOTALN	0.77	T04
TOTALN	POTASSIUM	0.68	T04
POTASSIUM	TOTALN	0.68	T04
NH4N	POTASSIUM	0.66	T04
POTASSIUM	NH4N	0.66	T04
TOTALP	TOTALN	0.66	T04
TOTALN	TOTALP	0.66	T04
TOTALP	NH4N	0.65	T04
NH4N	TOTALP	0.65	T04
IRON	ALUMINIUM	0.62	T04
ALUMINIUM	IRON	0.62	T04

```r
kable(result_T06_precipitation$differences_table, caption = "Large Pearson-Spearman Differences > 0.5 for T06 fo
```

Table 6: Large Pearson–Spearman Differences > 0.5 for T06 for Precipitation data

Variables1	Variables2	AbsDiff	Site
TOTALN	NO3N	0.80	T06
NO3N	TOTALN	0.80	T06
NO3N	SO4S	0.66	T06
SO4S	NO3N	0.66	T06
MAGNESIUM	PO4P	0.60	T06
TOTALN	PO4P	0.60	T06

Variables1	Variables2	AbsDiff	Site
PO4P	MAGNESIUM	0.60	T06
PO4P	TOTALN	0.60	T06
NH4N	NO3N	0.56	T06
NO3N	NH4N	0.56	T06
NH4N	PO4P	0.56	T06
PO4P	NH4N	0.56	T06
CONDY	PO4P	0.54	T06
PO4P	CONDY	0.54	T06
PO4P	SO4S	0.52	T06
SO4S	PO4P	0.52	T06
PO4P	POTASSIUM	0.51	T06
POTASSIUM	PO4P	0.51	T06

```
kable(result_T02_soil$differences_table, caption = "Large Pearson-Spearman Differences > 0.5 for T02 for Soil So
```

Table 7: Large Pearson–Spearman Differences > 0.5 for T02 for Soil Solution data

Variables1	Variables2	AbsDiff	Site

```
kable(result_T04_soil$differences_table, caption = "Large Pearson-Spearman Differences > 0.5 for T04 for Soil So
```

Table 8: Large Pearson–Spearman Differences > 0.5 for T04 for Soil Solution data

Variables1	Variables2	AbsDiff	Site
MAGNESIUM	CALCIUM	0.69	T04
CALCIUM	MAGNESIUM	0.69	T04
TOTALP	ALKY	0.63	T04
ALKY	TOTALP	0.63	T04
IRON	CALCIUM	0.62	T04
CALCIUM	IRON	0.62	T04
CONDY	SO4S	0.51	T04
SO4S	CONDY	0.51	T04

```
kable(result_T06_soil$differences_table, caption = "Large Pearson-Spearman Differences > 0.5 for T06 for Soil So
```

Table 9: Large Pearson–Spearman Differences > 0.5 for T06 for Soil Solution data

Variables1	Variables2	AbsDiff	Site
MAGNESIUM	IRON	0.88	T06
IRON	MAGNESIUM	0.88	T06
CONDY	IRON	0.80	T06
IRON	CONDY	0.80	T06
PO4P	CALCIUM	0.75	T06
CALCIUM	PO4P	0.75	T06
CALCIUM	TOTALN	0.68	T06
IRON	ALUMINIUM	0.68	T06
TOTALN	CALCIUM	0.68	T06
ALUMINIUM	IRON	0.68	T06
PO4P	MAGNESIUM	0.67	T06
MAGNESIUM	PO4P	0.67	T06
PO4P	IRON	0.67	T06
IRON	PO4P	0.67	T06

Variables1	Variables2	AbsDiff	Site
MAGNESIUM	TOTALN	0.66	T06
TOTALN	MAGNESIUM	0.66	T06
IRON	CALCIUM	0.65	T06
CALCIUM	IRON	0.65	T06
PO4P	TOTALN	0.60	T06
TOTALN	PO4P	0.60	T06
CONDY	PO4P	0.60	T06
PO4P	CONDY	0.60	T06
CONDY	TOTALN	0.57	T06
TOTALN	CONDY	0.57	T06

Question 4