Exploration of Differences in Results between Field and Lab Chlorine Analyses

I performed this exploratory data analysis as part of a larger project focusing on finding trends in results of Total Chlorine Analysis. My specific task was to search for possible correlations between the differences in results between lab and field chlorine analyses vs. differences in time between collection and analysis.

R scripting used for data transformation:

```
#Necessary Package Download

library(magrittr)
library(wqr)
library(dplyr)
library(tidyverse)
library(ggplot2)
library(fuzzyjoin)
library(sqldf)

#Extract db from lims

start <-'2022-10-19'

end <- '2023-10-19'

param<- c("Field-Chlorine Residual Total", "Chlorine Residual Total")

sites2 <- c(1101, 1102, 1103, 1107, 1300, 1301, 1302, 1303, 1401,
1601, 1602, 1603,
            1608, 1701, 1702, 1708, 1710, 1713, 1716, 1718, 1719,
1721, 2401, 2403,
            2501, 2502, 2600, 2601, 2602, 2603, 2706, 2712, 2713,
3800, 3801, 3803,
            3901, 3904, 3906, 3907, 3913, 4001, 4005, 7101, 7204,
7207, 7301, 7302,
            7401, 7502, 7601)

db <- read_LIMS(site = NULL,
```

```
                  parameter = param,

                  start_date = start, end_date = end,

                  sample_class = "Routine Daily") %>%

  filter(!is.na(result)) %>%

  group_by(lims_number, site) %>%

  arrange(lims_number)


#Data Cleaning:
#deleting unnecessary/ redundant fields / variables
#Pivoting parameter field
#Move lims_number to the front to act as primary key
#Cleaning NA values

drop_cols <- c("sample_class", "project_no", "result_as_entered",
"sample_type", "lab_method")
db <- db |>
  select(-one_of(drop_cols)) |>
  pivot_wider(
    names_from = parameter,
    values_from = result
  ) |>
  relocate(lims_number) |>
  na.omit(db)


colnames(db)[4] <- "collection_time"
colnames(db)[7] <- "lcrt"
colnames(db)[8] <- "fcrt"


#Calculating difference between field and lab results

db1 <- db |>
  mutate(
    result_difference = fcrt-lcrt
  )

#Query and create a new database which includes analysis time to find
the difference
```

```
dbtr <- read_LIMS(site = NULL,

                  parameter = param,

                  start_date = start, end_date = end,

                  sample_class = "Routine Daily",

                  select_additional = c("date_time_analyzed" =
"ANALYZED_ON", "ANALYZED_BY")) %>%

  filter(!is.na(result)) %>%

  arrange(lims_number)

#data clean-up and pivot

dbtr <- dbtr %>%
  distinct(lims_number, lab_method, date_time, date_time_analyzed)
%>%
  relocate(lims_number) %>%
  pivot_wider(names_from = lab_method,
              values_from = c(date_time, date_time_analyzed)) %>%
  na.omit(dbtr)

#changing column names for easier referencing

colnames(dbtr)[2] <- "ct_lab"
colnames(dbtr)[3] <- "ct_field"
colnames(dbtr)[4] <- "at_lab"
colnames(dbtr)[5] <- "at_field"

#calculation

dbtr <- dbtr %>%
  mutate(at_difference = difftime(at_lab, at_field, units = "mins"))
%>%
  mutate(at_ct_lab_dif = difftime(at_lab, ct_lab, units = "mins"))
%>%
  mutate(at_ct_field_dif = difftime(at_field, ct_field, units =
"mins"))

#view(dbtr)
```

```
#Inner-join dbtr and db1 tables by lims_number

time_analysis <-
  merge(db1, dbtr, by = "lims_number", all = FALSE)

#filtering result_difference to show only negative difference
accounting for natural chlorine breakdown over time

time_analysis <- time_analysis %>%
  filter(result_difference < 0)

view(time_analysis)

#Export

write.csv(time_analysis,
"C:/Users/Nathan.Ziemecki/Desktop/field_vs_lab/time/time_analysis.csv
", row.names=FALSE)
```

Python scripting used for calculating Pearson's Correlation Coefficients:

```
import pandas as pd
from pathlib import Path

df = pd.read_csv('C:/Users/Nathan.Ziemecki/Desktop/Learning
R/time_analysis.csv')

df1 = df[['result_difference', 'at_difference', 'at_ct_lab_dif',
'at_ct_field_dif']]

correlationsdf1 = df1.corr(method = 'pearson')

filepath =
Path('C:/Users/Nathan.Ziemecki/Desktop/field_vs_lab/time/correlation_
time_dif.csv')

correlationsdf1.to_csv(filepath)
```

Results:

	result_difference	at_difference	at_ct_lab_dif	at_ct_field_dif
result_difference	1	0.015624489	0.015598232	-0.001790684
at_difference	0.015624489	1	0.999931164	0.022715604
at_ct_lab_dif	0.015598232	0.999931164	1	0.03444416
at_ct_field_dif	-0.001790684	0.022715604	0.03444416	1

at_difference: difference between field and lab analysis times
at_ct_lab_dif: difference between sample collection time and lab analysis time
at_ct_field_dif: difference between sample collection time and field analysis time