ST117 Individual DRAFT Written Report - Task C

My WARWICK ID 5645242 (Report Pod 041)

Today's date in the format 2025-04-10

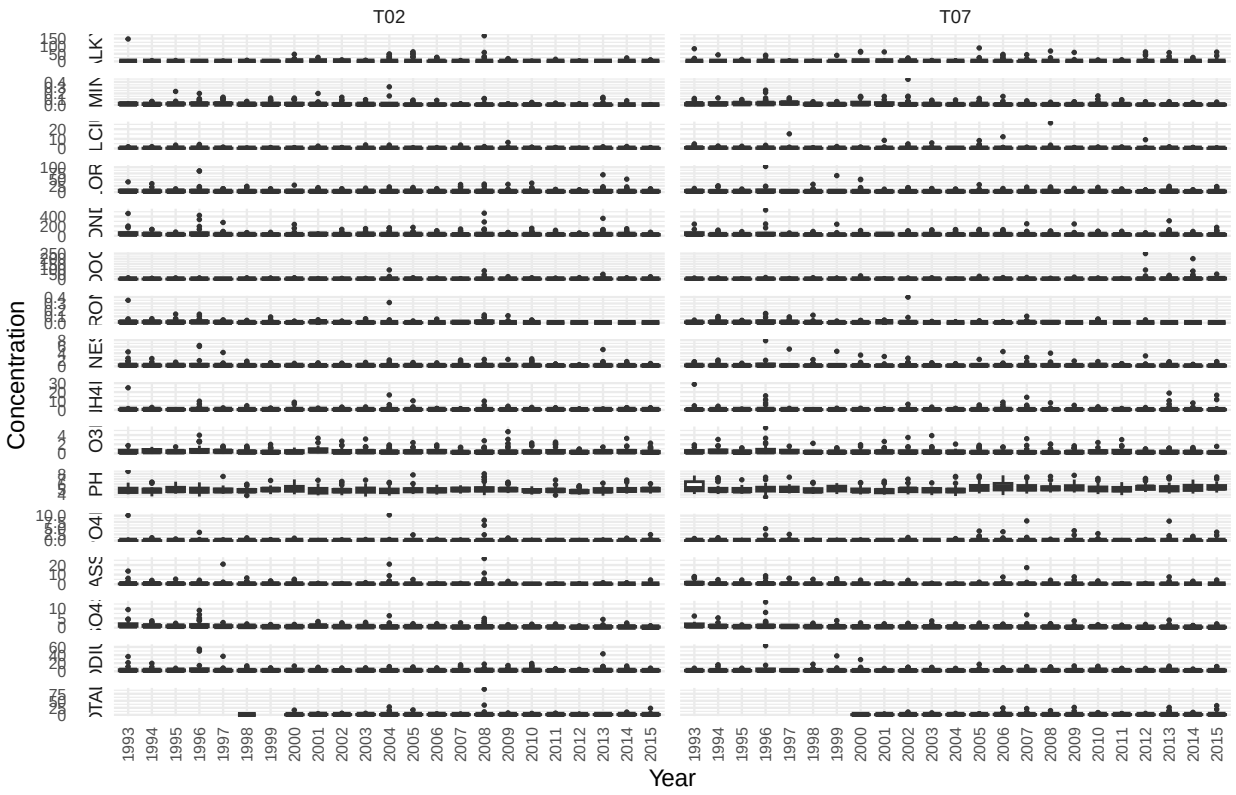
Question 1

Setting up data frames before for phase 2 part C

```
#Import the dataframes from phase 1, precipitation for our pod.
df_precipitation_wide <- readRDS("/Users/danielguo/Desktop/University/Year 1/ST117/5645242_041_WR/df_precipitati

#There is no LCODE or RID for precipitation data set, no averaging is needed

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

#filter out the assigned data for our pod: T02, T07
df_precipitation_Afiltered <- df_precipitation_wide %>%
  filter(SITECODE %in% c("T02", "T07")) %>%
  mutate(
    SDATE = as.Date(SDATE),
    YEAR = as.numeric(format(SDATE, "%Y"))
  )
```

Question 2

```
# load and reshape data
df_precip_tall <- df_precipitation_Afiltered %>%
  pivot_longer(-c(SDATE, SITECODE, YEAR), names_to = "Variables", values_to = "VALUE") %>%
  drop_na() %>%
  mutate(DAYS = as.numeric(SDATE - min(SDATE)))

#plot year boxplots
ggplot(df_precip_tall, aes(x = factor(YEAR), y = VALUE)) +
  geom_boxplot(outlier.size = 0.3) +
  facet_grid(Variables ~ SITECODE, scales = "free_y", switch = "y", labeller = label_wrap_gen()) +
  labs(title = "Yearly Boxplots of Chemical Concentrations by Site",
       x = "Year", y = "Concentration") +
  theme_minimal(base_size = 9) +
  theme(
    axis.text.x = element_text(angle = 90, vjust = 0.5, size = 6),
    axis.text.y = element_text(size = 6),
    strip.text = element_text(size = 7),
    plot.title = element_text(hjust = 0.5, size = 14),
    panel.spacing = unit(0.4, "lines")
  )
```



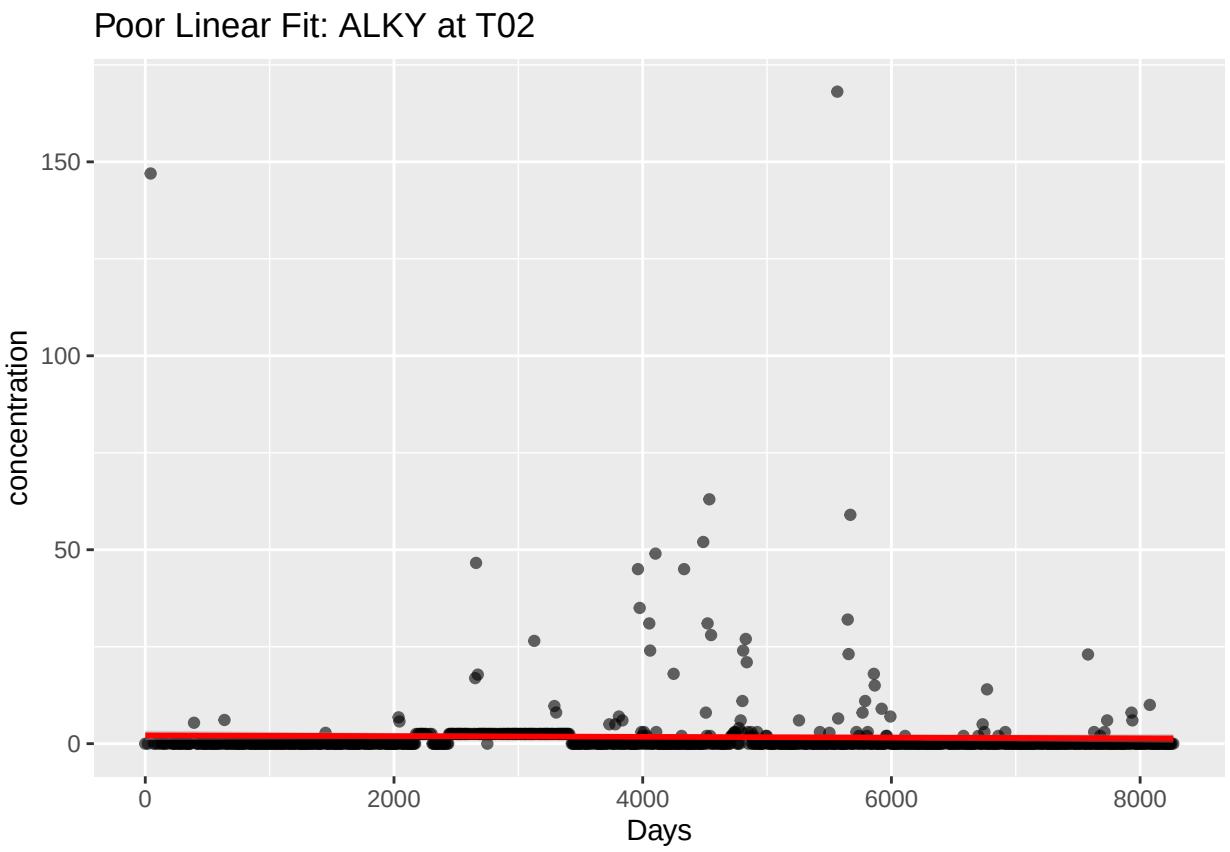

```r
# Compute R2 for each sites and chemicals
mdl_summary <- df_precip_tall %>%
  group_by(SITECODE, Variables) %>%
  group_map(~{
    mod <- lm(VALUE ~ DAYS, data = .x)
    s <- summary(mod)
    data.frame(
      SITECODE = .y$SITECODE,
      Variables = .y$Variables,
      r.squared = s$r.squared,
      p.value = pf(s$fstatistic[1], s$fstatistic[2], s$fstatistic[3], lower.tail = FALSE)
    )
  }) %>%
  bind_rows()

#pick examples for good poor fit
good_fit <- mdl_summary %>% filter(r.squared > 0.7) %>% slice(1)
poor_fit <- mdl_summary %>% filter(r.squared < 0.1) %>% slice(1)

#function to plot fit
plot_fit <- function(data, fit_info, color, title_prefix) {
  ggplot(data, aes(x = DAYS, y = VALUE)) +
    geom_point(alpha = 0.6) +
    geom_smooth(method = "lm", color = color) +
    labs(
      title = paste(title_prefix, "Linear Fit:", fit_info$Variables, "at", fit_info$SITECODE),
      x = "Days",
      y = "concentration"
    )
}

#plot good fit if available, if not say no good fit
if (nrow(good_fit) > 0) {
  good_data <- df_precip_tall %>%
```

```
    filter(SITECODE == good_fit$SITECODE[1], Variables == good_fit$Variables[1])
  print(plot_fit(good_data, good_fit, "blue", "Good"))
} else {
  message("No good fit found R2 > 0.7")
}
```

```
## No good fit found R2 > 0.7
```

```
# Plot poor fit, if not say no poor fit
if (nrow(poor_fit) > 0) {
  poor_data <- df_precip_tall %>%
    filter(SITECODE == poor_fit$SITECODE[1], Variables == poor_fit$Variables[1])
  print(plot_fit(poor_data, poor_fit, "red", "Poor"))
} else {
  message("No poor fit found R2 < 0.1.")
}
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



g # Question 3