Barton Springs Dataset Test

Group 7

2023-10-04

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
library(openxlsx)
library(readxl)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data = read.csv("~/Data-Science-G7/CopyOfBarton_Springs_Salamanders_DO_and_Flow.csv")
```

```
#  a subset with only relevant columns for DISSOLVED OXYGEN
dissolved_oxygen_data <- data[data$PARAMETER == "DISSOLVED OXYGEN" & data$UNIT == "MG/L", c("WATERSHED"

#  a subset with only relevant columns for TOTAL SALAMANDER
total_salamander_data <- data[data$PARAM_TYPE == "Salamanders", c("WATERSHED", "SAMPLE_DATE", "SITE_NAM

total_salamander_data$SITE_NAME <- as.factor(total_salamander_data$SITE_NAME)

#  RESULT to numeric
total_salamander_data$RESULT <- as.numeric(total_salamander_data$RESULT)

#  a linear model
model <- lm(RESULT ~ SITE_NAME, data = total_salamander_data)

#  the summary of the regression
summary(model)
```

```
##
## Call:
## lm(formula = RESULT ~ SITE_NAME, data = total_salamander_data)
```

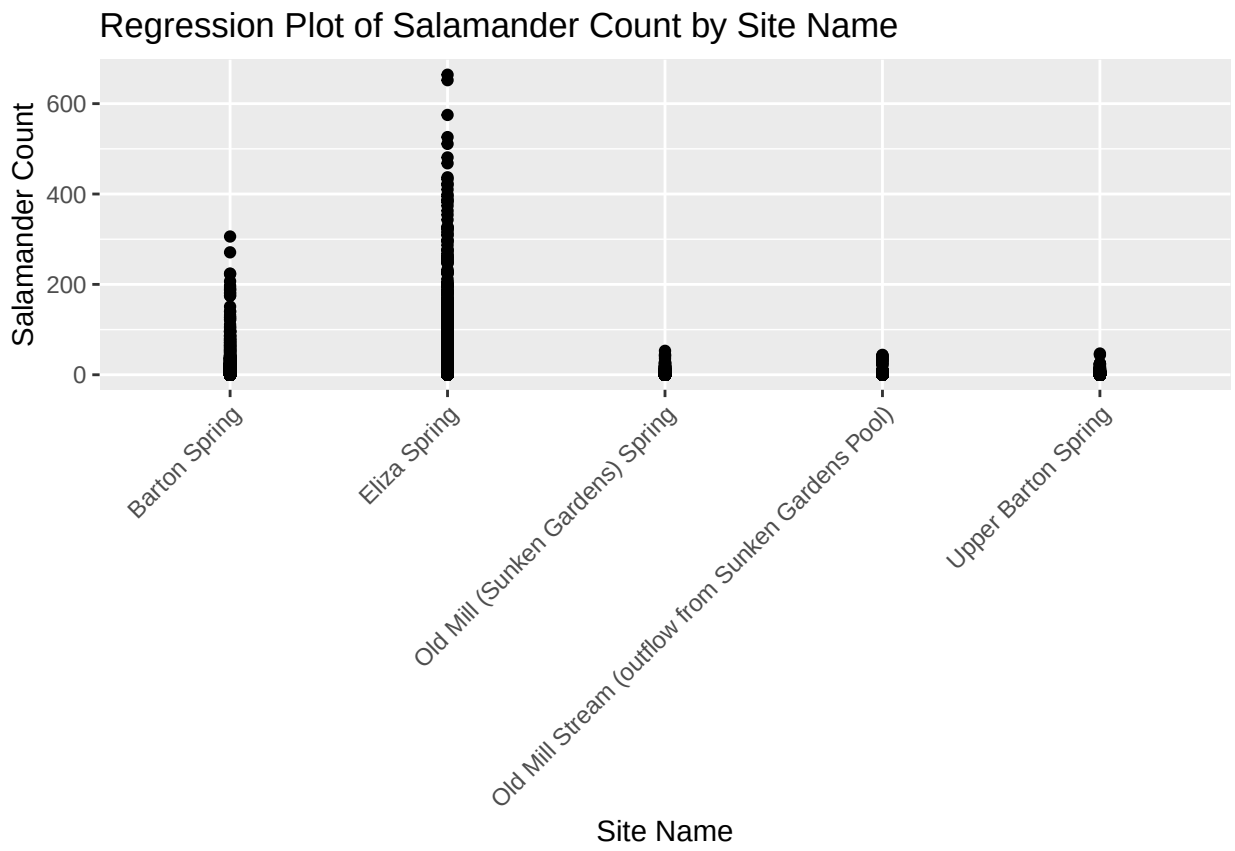
```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.28  -2.67  -0.40  -0.31 649.72
##
## Coefficients:
##                                                                 Estimate Std. Error
## (Intercept)                                                       2.6714     0.2084
## SITE_NAMEEliza Spring                                            11.6044     0.2890
## SITE_NAMEOld Mill (Sunken Gardens) Spring                        -2.3564     0.2597
## SITE_NAMEOld Mill Stream (outflow from Sunken Gardens Pool)      -2.0501     0.5396
## SITE_NAMEUpper Barton Spring                                     -2.2761     0.3106
##                                                                 t value Pr(>|t|)
## (Intercept)                                                      12.817  < 2e-16
## SITE_NAMEEliza Spring                                            40.158  < 2e-16
## SITE_NAMEOld Mill (Sunken Gardens) Spring                        -9.072  < 2e-16
## SITE_NAMEOld Mill Stream (outflow from Sunken Gardens Pool)      -3.799 0.000145
## SITE_NAMEUpper Barton Spring                                     -7.329 2.36e-13
##
## (Intercept)                                                     ***
## SITE_NAMEEliza Spring                                           ***
## SITE_NAMEOld Mill (Sunken Gardens) Spring                       ***
## SITE_NAMEOld Mill Stream (outflow from Sunken Gardens Pool)     ***
## SITE_NAMEUpper Barton Spring                                    ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.69 on 39296 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.08246,    Adjusted R-squared:  0.08237
## F-statistic: 882.9 on 4 and 39296 DF,  p-value: < 2.2e-16
```

```r
# load the ggplot2 library if not already installed
if (!require(ggplot2)) {
  install.packages("ggplot2")
  library(ggplot2)
}


#  ggplot code here
gg <- ggplot(total_salamander_data, aes(x = SITE_NAME, y = RESULT)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Regression Plot of Salamander Count by Site Name",
       x = "Site Name",
       y = "Salamander Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
gg
```

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

```
## Warning: Removed 12 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 12 rows containing missing values (`geom_point()`).
```




```
#  the ggplot as a PNG file
ggsave("salamander_regression_plot.png", gg, width = 10, height = 6, units = "in")
```

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

```
## Warning: Removed 12 rows containing non-finite values (`stat_smooth()`).
## Removed 12 rows containing missing values (`geom_point()`).
```

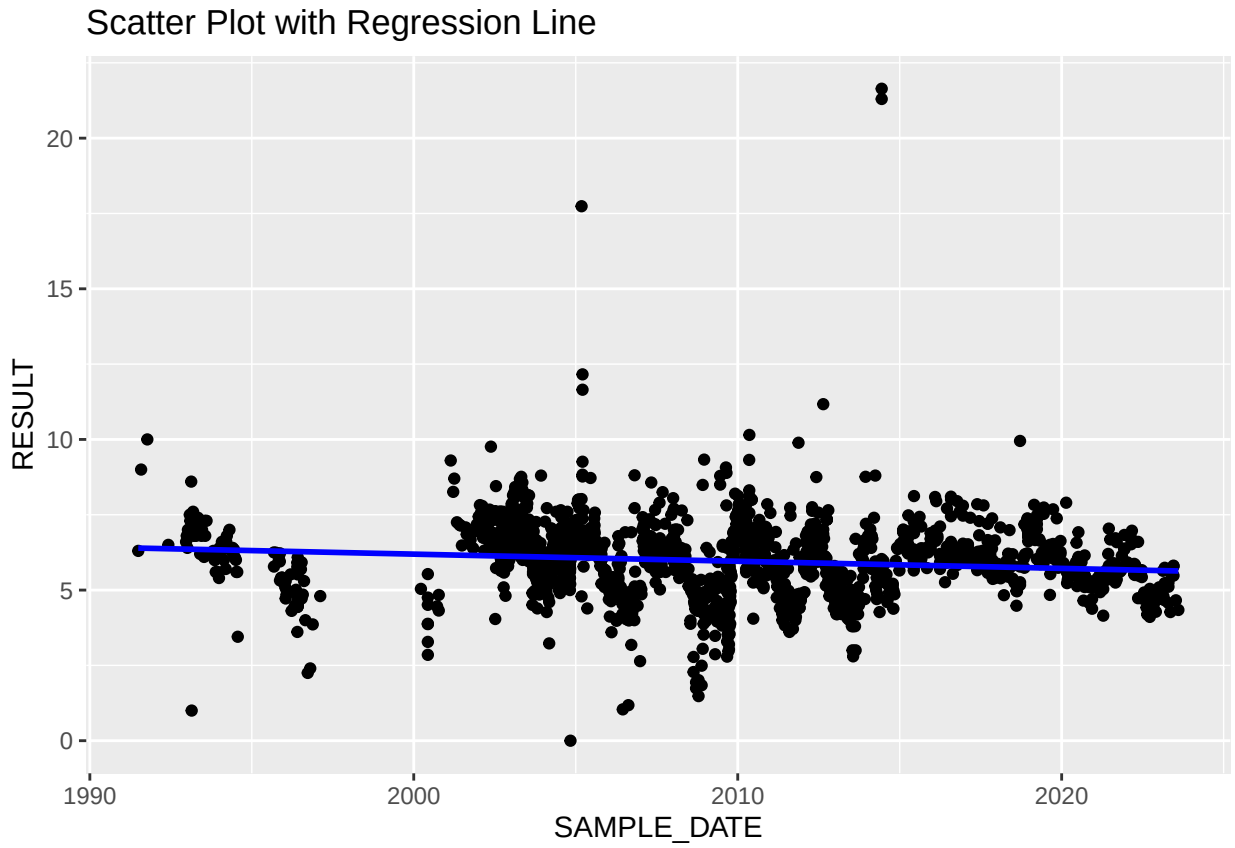
```
#  categorical variables to factors
dissolved_oxygen_data$WATERSHED <- as.factor(dissolved_oxygen_data$WATERSHED)
dissolved_oxygen_data$SITE_NAME <- as.factor(dissolved_oxygen_data$SITE_NAME)
dissolved_oxygen_data$UNIT <- as.factor(dissolved_oxygen_data$UNIT)
dissolved_oxygen_data$RESULT <- as.numeric(dissolved_oxygen_data$RESULT)

# Convert SAMPLE_DATE to Date
dissolved_oxygen_data$SAMPLE_DATE <- as.Date(dissolved_oxygen_data$SAMPLE_DATE, format = "%m/%d/%Y %I:%M


# Fit a linear model
model_date_vs_dissolved_oxygen <- lm(RESULT ~ SAMPLE_DATE, data = dissolved_oxygen_data)

# Load the ggplot2 library
library(ggplot2)

#  a scatter plot with a regression line
plot <- ggplot(dissolved_oxygen_data, aes(x = SAMPLE_DATE, y = RESULT)) +
  geom_point() +
```

```
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Scatter Plot with Regression Line",
       x = "SAMPLE_DATE",
       y = "RESULT")
plot
```

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

```
## Warning: Removed 1 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```




```
# Save the plot as a PNG file
ggsave("scatter_plot_regression_line.png", plot, width = 10, height = 6, units = "in")
```

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

```
## Warning: Removed 1 rows containing non-finite values (`stat_smooth()`).
## Removed 1 rows containing missing values (`geom_point()`).
```

```
#  summary statistics
summary(model_date_vs_dissolved_oxygen)
```

```
##
## Call:
## lm(formula = RESULT ~ SAMPLE_DATE, data = dissolved_oxygen_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -6.0779 -0.7865  0.0298  0.7157 15.7894
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.902e+00  1.848e-01  37.356  < 2e-16 ***
## SAMPLE_DATE -6.480e-05  1.272e-05  -5.095 3.85e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.316 on 1818 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.01408,    Adjusted R-squared:  0.01354
## F-statistic: 25.96 on 1 and 1818 DF,  p-value: 3.85e-07
```