Analysis of California Residential Water Demand and Supply

Steve

2025-04-18

```
# Install required packages (uncomment if needed)
# install.packages("readr")
# install.packages("dplyr")
# install.packages("tidyr")
# install.packages("lubridate")
# install.packages("ggplot2")
# install.packages("corrplot")
# install.packages("ggcorrplot")

# Load required libraries
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.4.3
```

```
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
```

```
library(tidyr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

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

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.4.3
```

```
## corrplot 0.95 loaded
```

```r
# Load required packages
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats 1.0.0     v stringr 1.5.1
## v purrr   1.0.2     v tibble  3.2.1
## -- 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
```

```r
library(lubridate)

# 1. Read and Clean Data
df <- read.csv("C:/Users/HomePC/OneDrive/Desktop/CaRDS.csv", check.names = FALSE) %>%
  pivot_longer(
    cols = -c(PWSID, Variable),
    names_to = "Date",
    values_to = "Value"
  ) %>%
  mutate(Date = as.Date(Date)) %>%
  filter(!is.na(Value))  # Remove missing values
```

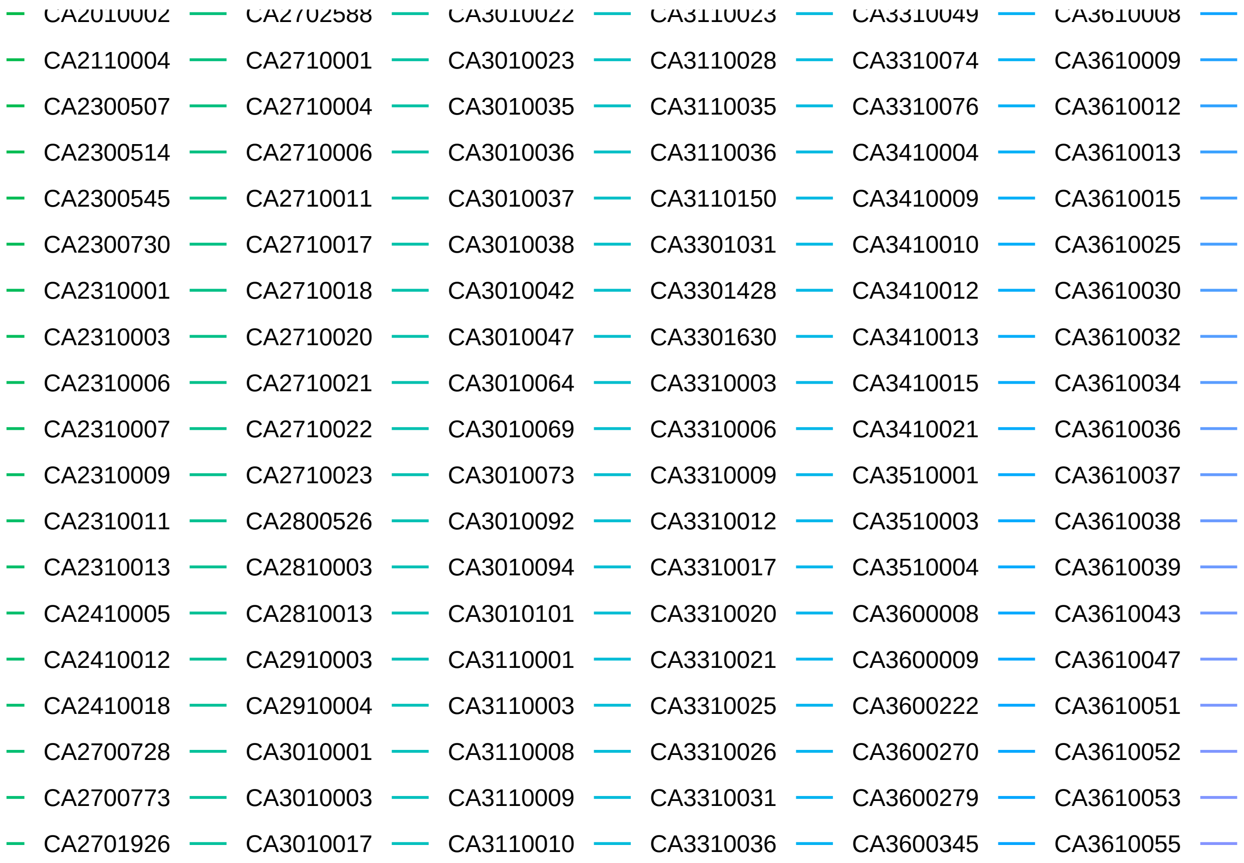
```r
# 2. Basic Exploration
# View structure
glimpse(df)
```

```
## Rows: 218,160
## Columns: 4
## $ PWSID    <chr> "CA0110005", "CA0110005", "CA0110005", "CA0110005", "CA011000~
## $ Variable <chr> "PDSI", "PDSI", "PDSI", "PDSI", "PDSI", "PDSI", "PDSI", "PDSI~
## $ Date     <date> 2013-01-01, 2013-02-01, 2013-03-01, 2013-04-01, 2013-05-01, ~
## $ Value    <dbl> -1.33, -2.02, -2.76, -3.23, -3.72, -4.14, -4.39, -4.31, -4.02~
```

```r
# Summary statistics
df %>%
  group_by(PWSID, Variable) %>%
  summarise(
    Mean = mean(Value, na.rm = TRUE),
    Median = median(Value, na.rm = TRUE),
    SD = sd(Value, na.rm = TRUE),
    .groups = "drop"
  )
```

```
## # A tibble: 2,020 x 5
##    PWSID     Variable          Mean        Median      SD
##    <chr>     <chr>            <dbl>         <dbl>   <dbl>
##  1 CA0110005 PDSI          -2.44e 0         -2.46 2.49e 0
##  2 CA0110005 demand         2.92e 9 2854036302    5.98e 8
##  3 CA0110005 precipitation  4.00e 1          6.24 6.38e 1
##  4 CA0110005 supply         5.00e 9 4891900000    1.03e 9
##  5 CA0110005 temperature    1.53e 1         15.2  2.64e 0
##  6 CA0110006 PDSI          -2.44e 0         -2.46 2.49e 0
##  7 CA0110006 demand         2.13e 8  227687086    8.86e 7
##  8 CA0110006 precipitation  3.48e 1          9.18 5.34e 1
##  9 CA0110006 supply         1.35e11  510466242    2.06e11
## 10 CA0110006 temperature    1.53e 1         15.4  3.56e 0
## # i 2,010 more rows
```

```
# 3. Time Series Visualization
# Plot PDSI over time
df %>%
  filter(Variable == "PDSI") %>%
  ggplot(aes(x = Date, y = Value, color = PWSID)) +
  geom_line() +
  labs(title = "PDSI Index Over Time", y = "PDSI") +
  theme_minimal()
```

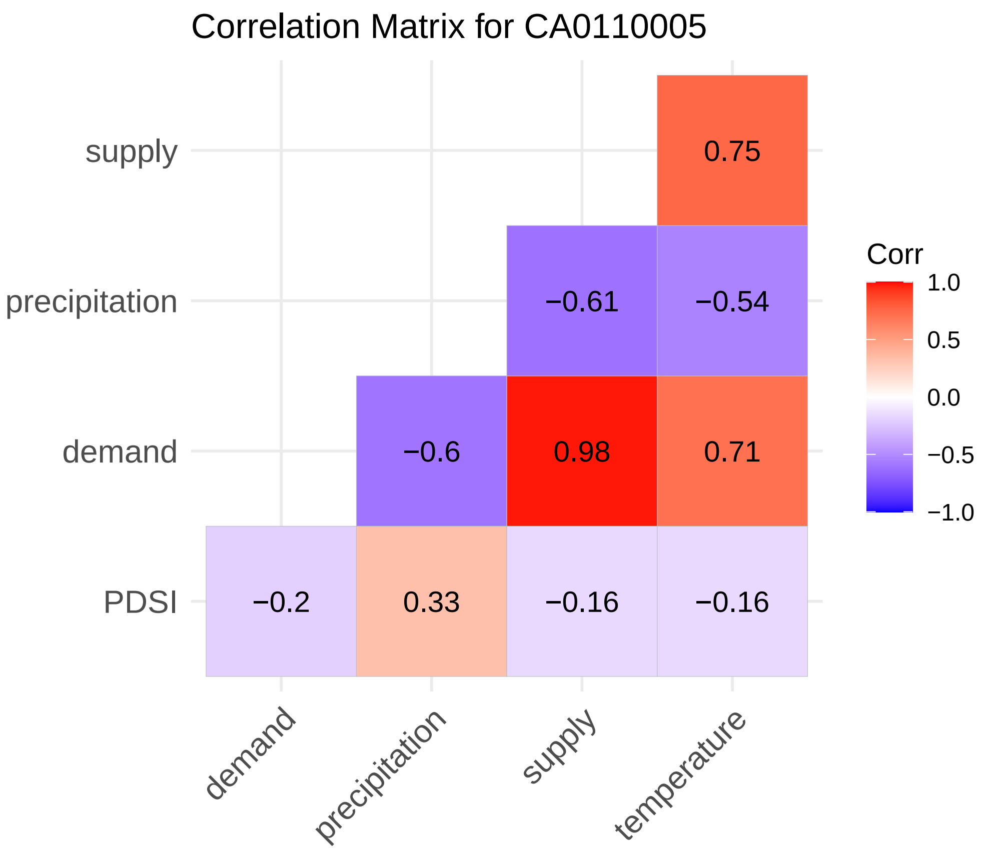


```r
# 4. Correlation Analysis for a single PWSID

library(ggcorrplot)
```

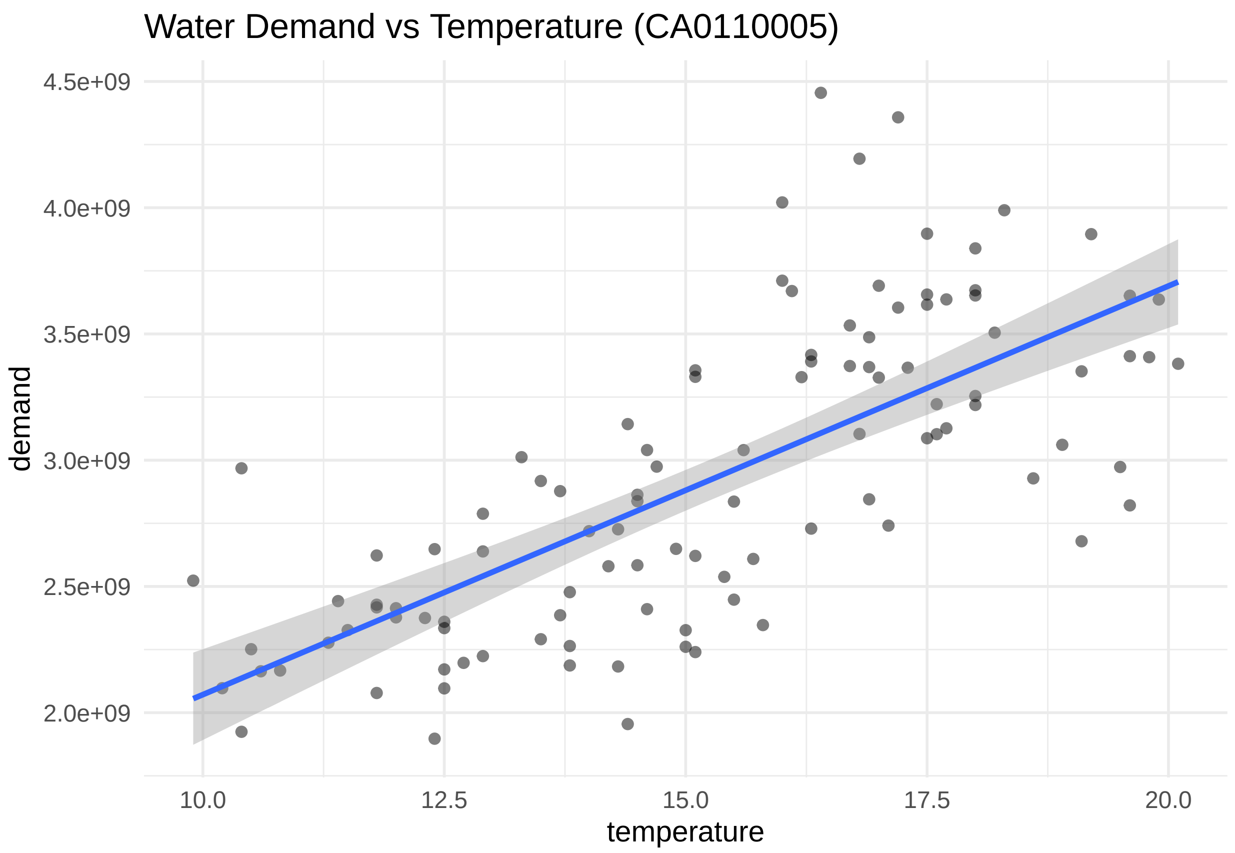
```
## Warning: package 'ggcorrplot' was built under R version 4.4.3
```

```r
df %>%
  filter(PWSID == "CA0110005") %>%
  select(-PWSID) %>%
  pivot_wider(names_from = Variable, values_from = Value) %>%
  select(-Date) %>%
  cor(use = "complete.obs") %>%
  ggcorrplot::ggcorrplot(
    type = "lower",
    lab = TRUE,
    title = "Correlation Matrix for CA0110005"
  )
```

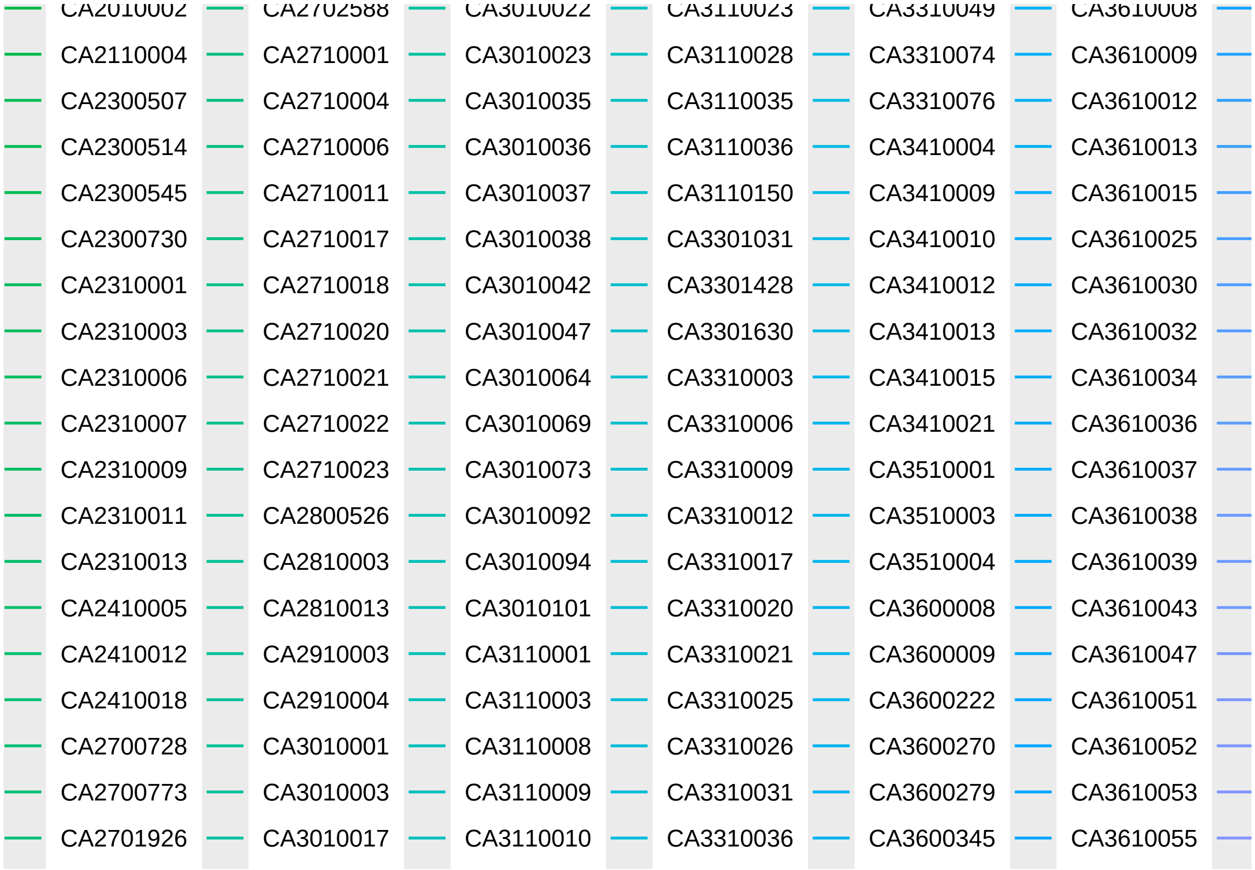


```
# 5. Demand vs Temperature Analysis
df %>%
  filter(PWSID == "CA0110005",
         Variable %in% c("demand", "temperature")) %>%
  pivot_wider(names_from = Variable, values_from = Value) %>%
  ggplot(aes(x = temperature, y = demand)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm") +
  labs(title = "Water Demand vs Temperature (CA0110005)") +
  theme_minimal()
```

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



```r
# 6. Monthly Aggregations
# Average monthly demand across years
df %>%
  filter(Variable == "demand") %>%
  mutate(Month = month(Date, label = TRUE)) %>%
  group_by(PWSID, Month) %>%
  summarise(Avg_Demand = mean(Value, na.rm = TRUE), .groups = "drop") %>%
  ggplot(aes(x = Month, y = Avg_Demand, color = PWSID, group = PWSID)) +
  geom_line() +
  labs(title = "Seasonal Demand Patterns", y = "Average Demand")
```



```r
# 7. Annual Trends
# Annual precipitation trends
df %>%
  filter(Variable == "precipitation") %>%
  mutate(Year = year(Date)) %>%
  group_by(PWSID, Year) %>%
  summarise(Total_Precipitation = sum(Value, na.rm = TRUE), .groups = "drop") %>%
  ggplot(aes(x = Year, y = Total_Precipitation, color = PWSID)) +
  geom_line() +
  geom_point() +
  labs(title = "Annual Precipitation Trends", y = "Total Precipitation")
```

CA2010002
CA2110004
CA2300507
CA2300514
CA2300545
CA2300730
CA2310001
CA2310003
CA2310006
CA2310007
CA2310009
CA2310011
CA2310013
CA2410005
CA2410012
CA2410018
CA2700728
CA2700773
CA2701926
CA2702588
CA2710001
CA2710004
CA2710006
CA2710011
CA2710017
CA2710018
CA2710020
CA2710021
CA2710022
CA2710023
CA2800526
CA2810003
CA2810013
CA2910003
CA2910004
CA3010001
CA3010003
CA3010017
CA3010022
CA3010023
CA3010035
CA3010036
CA3010037
CA3010038
CA3010042
CA3010047
CA3010064
CA3010069
CA3010073
CA3010092
CA3010094
CA3010101
CA3110001
CA3110003
CA3110008
CA3110009
CA3110010
CA3110023
CA3110028
CA3110035
CA3110036
CA3110150
CA3301031
CA3301428
CA3301630
CA3310003
CA3310006
CA3310009
CA3310012
CA3310017
CA3310020
CA3310021
CA3310025
CA3310026
CA3310031
CA3310036
CA3310049
CA3310074
CA3310076
CA3410004
CA3410009
CA3410010
CA3410012
CA3410013
CA3410015
CA3410021
CA3510001
CA3510003
CA3510004
CA3600008
CA3600009
CA3600222
CA3600270
CA3600279
CA3600345
CA3610008
CA3610009
CA3610012
CA3610013
CA3610015
CA3610025
CA3610030
CA3610032
CA3610034
CA3610036
CA3610037
CA3610038
CA3610039
CA3610043
CA3610047
CA3610051
CA3610052
CA3610053
CA3610055