BS755 / MA575 Lab 3: Data Visualization

Friday, September 22, 2023

Contents

Lab 3: Data Visualization **1**
Setup 1
Data Import and Exploration 2
Data Visualization 4

Lab 3: Data Visualization

In this lab, we will explore data visualization using an air quality data set.

Setup

Setting up global options to supress printing any warnings. Install `rmarkdown` and `knitr` packages.

```
# Global options to suppress warnings in all code chunks
knitr::opts_chunk$set(warning = FALSE)

# Install rmarkdown and knitr packages if they are not already installed.
# install.packages("rmarkdown")
# install.packages("knitr")
```

Set Working Directory

First, set the working directory to the location where your data is stored on your computer. Uncomment and modify the `setwd` line as needed.

```
# setwd("your/directory/path")
```

Install and Load Packages

Install and load the necessary R packages if they are not already installed. Uncomment the `install.packages` lines if needed.

```
# install.packages("car")
# install.packages("ggplot2")
# install.packages("GGally")
library(car)
```

```
## Loading required package: carData
```

```
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

Data Import and Exploration

Read Data

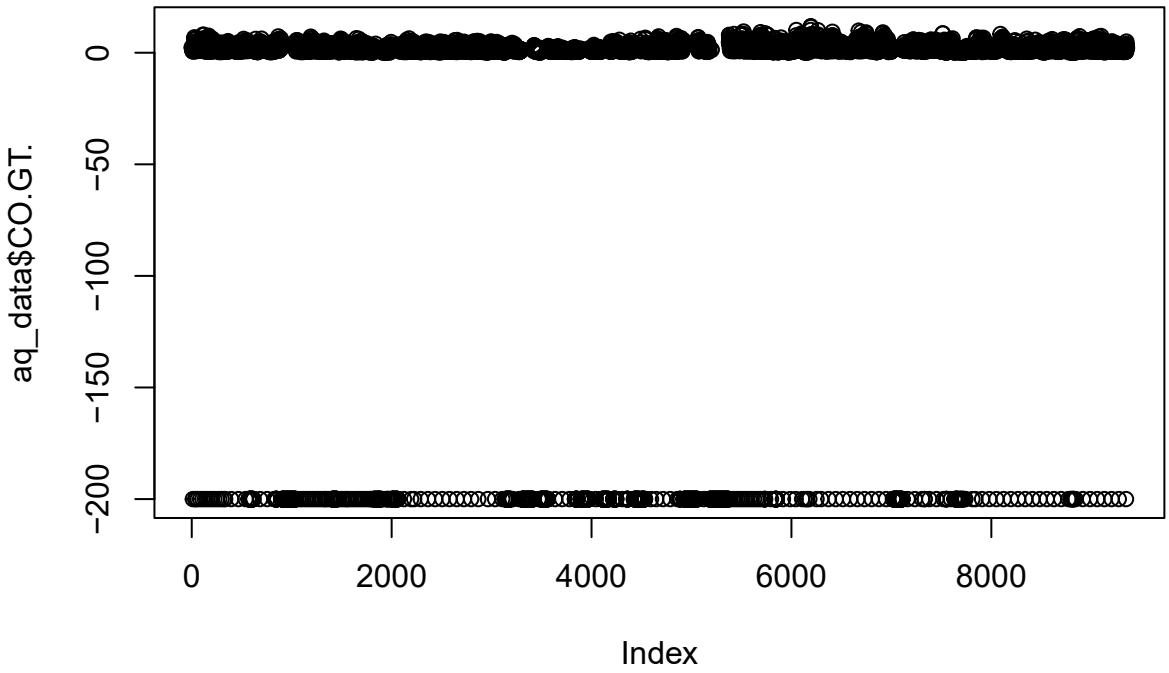
Let's start by reading the data from a CSV file using a semicolon as the delimiter.

```
# Read data from a CSV file using semicolon as delimiter
aq_data <- read.csv("AirQualityData.csv", header=TRUE, as.is=TRUE, sep=';')
```

Check Missing Values

We'll check for missing values in the "CO.GT." column.

```
# Plot Ground CO
plot(aq_data$CO.GT.)
```




```
# Summary of each column
summary(aq_data)
```

```
##      Date               Time               CO.GT.          PT08.S1.CO.
##  Length:9357        Length:9357        Min.   :-200.00   Min.   :-200
##  Class :character   Class :character   1st Qu.:   0.60   1st Qu.: 921
##  Mode  :character   Mode  :character   Median :   1.50   Median :1053
##                                        Mean   : -34.21   Mean   :1049
##                                        3rd Qu.:   2.60   3rd Qu.:1221
##                                        Max.   :  11.90   Max.   :2040
##     NMHC.GT.          C6H6.GT.         PT08.S2.NMHC.       NOx.GT.
##  Min.   :-200.0   Min.   :-200.000   Min.   :-200.0   Min.   :-200.0
##  1st Qu.:-200.0   1st Qu.:   4.000   1st Qu.: 711.0   1st Qu.:  50.0
##  Median :-200.0   Median :   7.900   Median : 895.0   Median : 141.0
##  Mean   :-159.1   Mean   :   1.866   Mean   : 894.6   Mean   : 168.6
##  3rd Qu.:-200.0   3rd Qu.:  13.600   3rd Qu.:1105.0   3rd Qu.: 284.0
##  Max.   :1189.0   Max.   :  63.700   Max.   :2214.0   Max.   :1479.0
##   PT08.S3.NOx.      NO2.GT.          PT08.S4.NO2.   PT08.S5.O3.
##  Min.   :-200   Min.   :-200.00   Min.   :-200   Min.   :-200.0
```

```
##  1st Qu.: 637   1st Qu.:  53.00   1st Qu.:1185   1st Qu.: 700.0
##  Median : 794   Median :  96.00   Median :1446   Median : 942.0
##  Mean   : 795   Mean   :  58.15   Mean   :1391   Mean   : 975.1
##  3rd Qu.: 960   3rd Qu.: 133.00   3rd Qu.:1662   3rd Qu.:1255.0
##  Max.   :2683   Max.   : 340.00   Max.   :2775   Max.   :2523.0
##   Temperature       RelativeHumidity  AbsoluteHumidity
##  Min.   :-200.000   Min.   :-200.00   Min.   :-200.0000
##  1st Qu.:  10.900   1st Qu.:  34.10   1st Qu.:   0.6923
##  Median :  17.200   Median :  48.60   Median :   0.9768
##  Mean   :   9.778   Mean   :  39.49   Mean   :  -6.8376
##  3rd Qu.:  24.100   3rd Qu.:  61.90   3rd Qu.:   1.2962
##  Max.   :  44.600   Max.   :  88.70   Max.   :   2.2310
```

```
# In this data set it looks like missing values are stored as -200
# Check which values in Ground CO are -200
# which(aq_data$CO.GT. == -200)
# The output gives indices/positions in the vector with values -200

# Check how many values in Ground CO are -200
length(which(aq_data$CO.GT. == -200))
```

```
## [1] 1683
```

```
# Exercise: Check other columns in the data set for missing values
```

Data Cleaning

Replace missing data (-200) with NA.

```
# Replace missing data (-200) with NA
aq_data[aq_data == -200] <- NA
# Note that this replaces -200 to NA across all columns in aq_data
```

Data Type Conversion

Check for column types for the columns in the data set.

```
# Check column types for all columns in the data set
str(aq_data)
```

```
## 'data.frame':    9357 obs. of  15 variables:
##  $ Date           : chr  "10/03/2004" "10/03/2004" "10/03/2004" "10/03/2004" ...
##  $ Time           : chr  "18.00.00" "19.00.00" "20.00.00" "21.00.00" ...
##  $ CO.GT.         : num  2.6 2 2.2 2.2 1.6 1.2 1.2 1 0.9 0.6 ...
##  $ PT08.S1.CO.    : int  1360 1292 1402 1376 1272 1197 1185 1136 1094 1010 ...
##  $ NMHC.GT.       : int  150 112 88 80 51 38 31 31 24 19 ...
##  $ C6H6.GT.       : num  11.9 9.4 9 9.2 6.5 4.7 3.6 3.3 2.3 1.7 ...
##  $ PT08.S2.NMHC.  : int  1046 955 939 948 836 750 690 672 609 561 ...
##  $ NOx.GT.        : int  166 103 131 172 131 89 62 62 45 NA ...
##  $ PT08.S3.NOx.   : int  1056 1174 1140 1092 1205 1337 1462 1453 1579 1705 ...
##  $ NO2.GT.        : int  113 92 114 122 116 96 77 76 60 NA ...
##  $ PT08.S4.NO2.   : int  1692 1559 1555 1584 1490 1393 1333 1333 1276 1235 ...
##  $ PT08.S5.O3.    : int  1268 972 1074 1203 1110 949 733 730 620 501 ...
##  $ Temperature    : num  13.6 13.3 11.9 11 11.2 11.2 11.3 10.7 10.7 10.3 ...
##  $ RelativeHumidity: num  48.9 47.7 54 60 59.6 59.2 56.8 60 59.7 60.2 ...
##  $ AbsoluteHumidity: num  0.758 0.726 0.75 0.787 0.789 ...
```

```
# Check column types for specific columsn
# typeof(aq_data$Temperature)
```

Force specific columns to be numeric if they are not already.

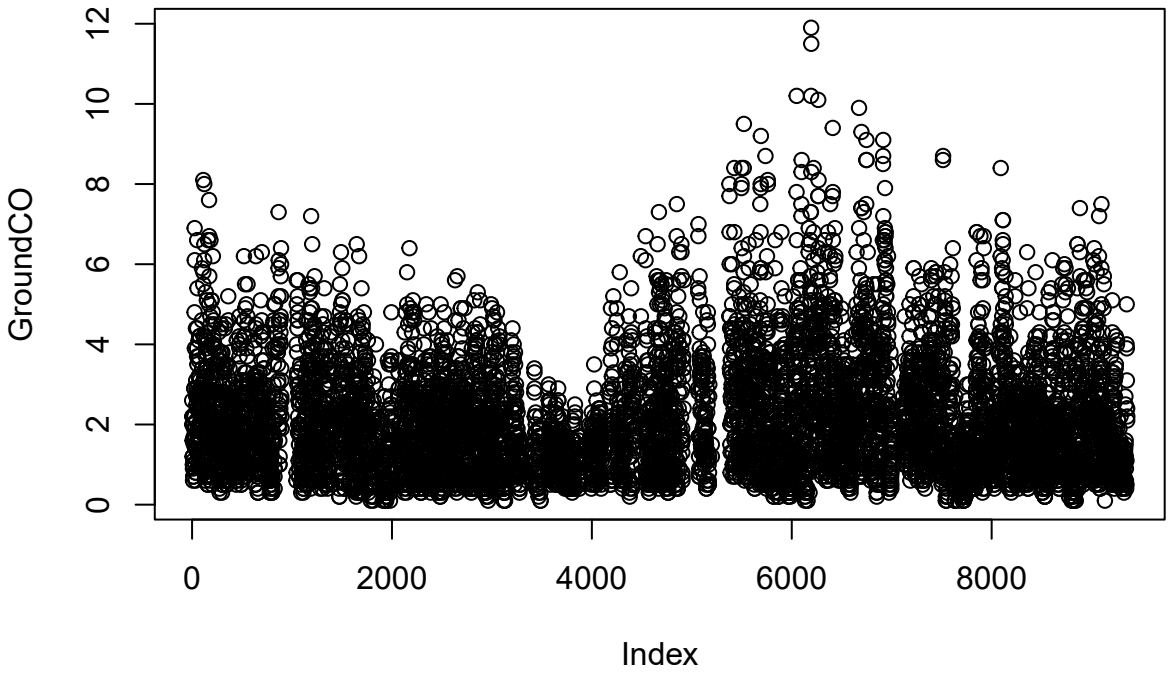
```
# Force specific columns to be numeric if not numeric/integer already.
# Note that change in variable names of PT08.S1.CO. and CO.GT.
Temperature <- as.numeric(as.character(aq_data$Temperature))
RelativeHumidity <- as.numeric(as.character(aq_data$RelativeHumidity))
AbsoluteHumidity <- as.numeric(as.character(aq_data$AbsoluteHumidity))
SensorCO <- as.numeric(as.character(aq_data$PT08.S1.CO.))
GroundCO <- as.numeric(as.character(aq_data$CO.GT.))
```

Data Visualization

Scatterplot of Ground CO

Let's create a scatterplot of Ground CO.

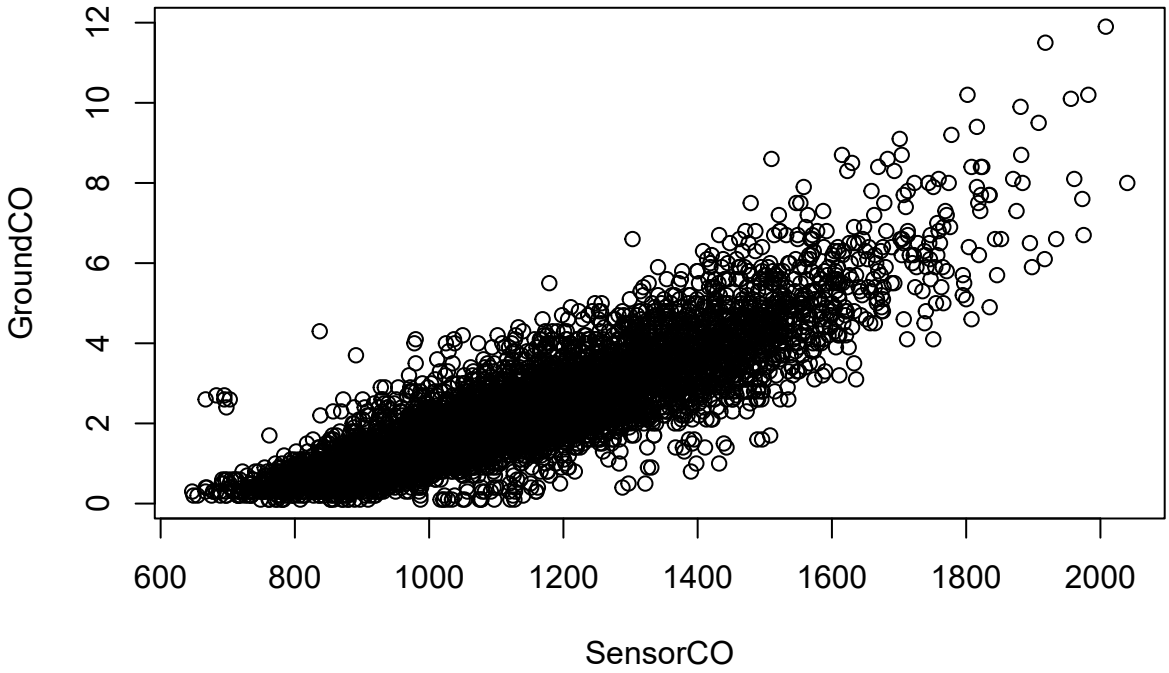
```
# Scatterplot of Ground CO
plot(GroundCO)
```



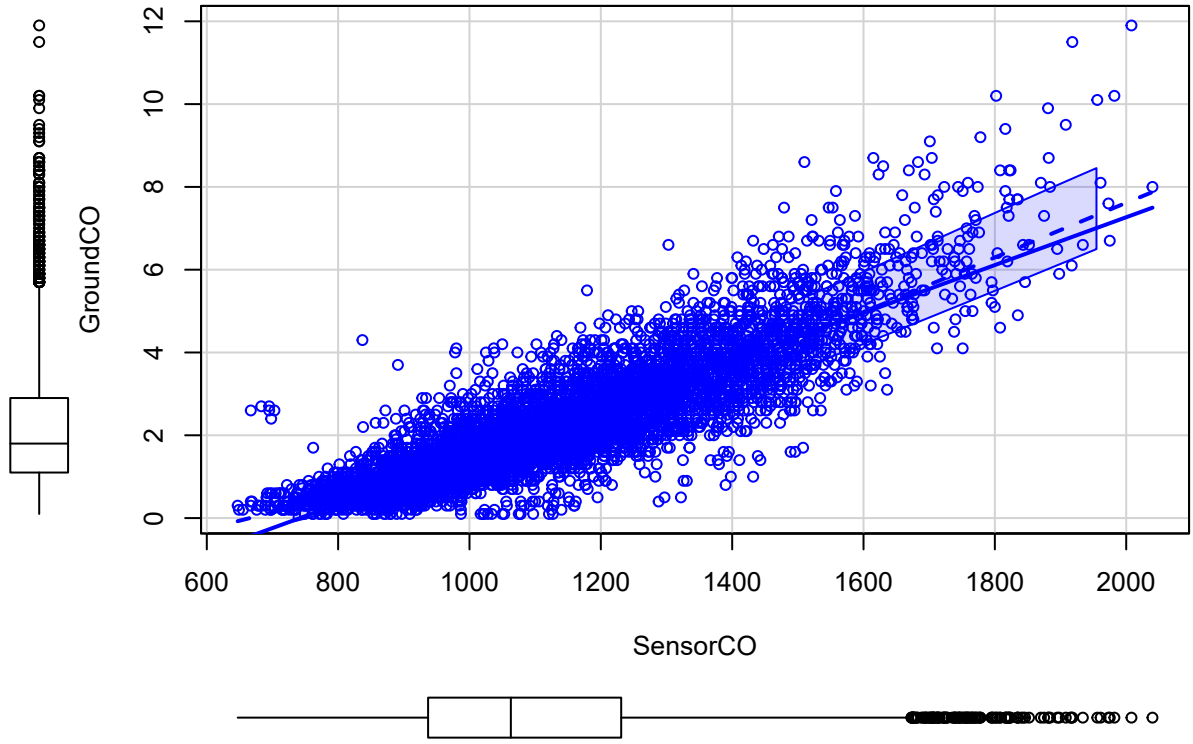
Scatterplot of Ground CO vs. Sensor CO

Now, we'll create a scatterplot of Ground CO vs. Sensor CO.

```
# Plot scatterplot of GroundCO vs SensorCO
plot(SensorCO, GroundCO)
```



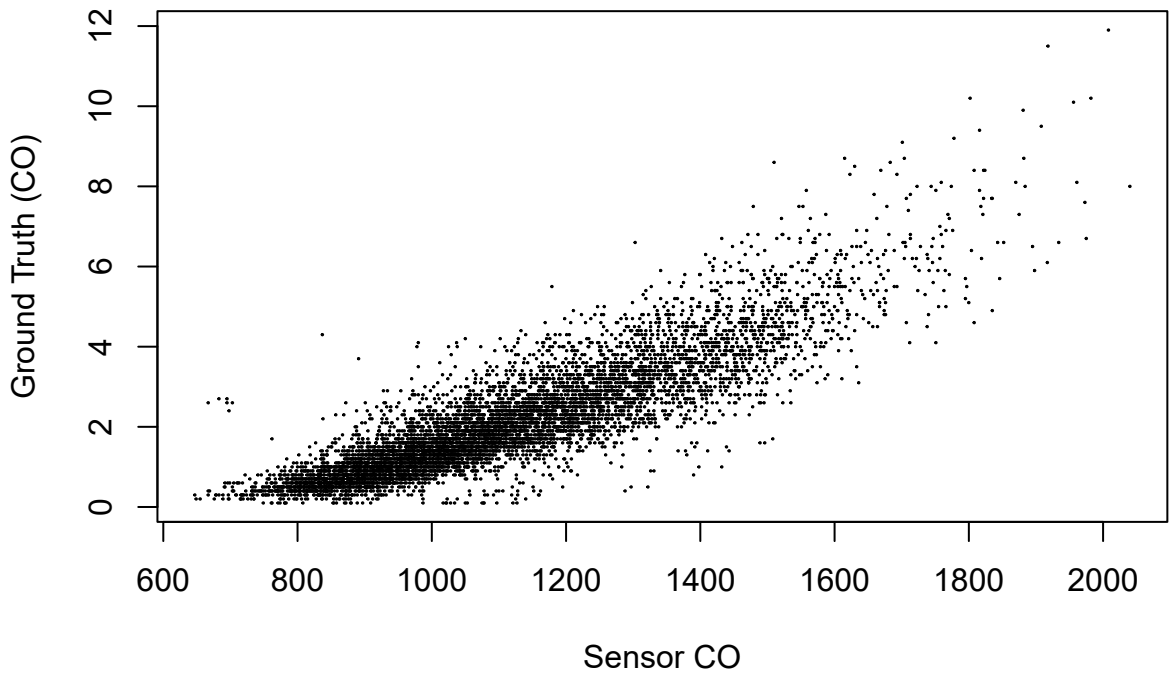

```
# Plot scatterplot of GroundCO vs SensorCO using the scatterplot function ...
# ... from the car package
scatterplot(SensorCO, GroundCO)
```




```
# Documentation for scatterplot function
# ?scatterplot

# Scatterplot of Ground CO vs. Sensor CO
plot(SensorCO, GroundCO,
```

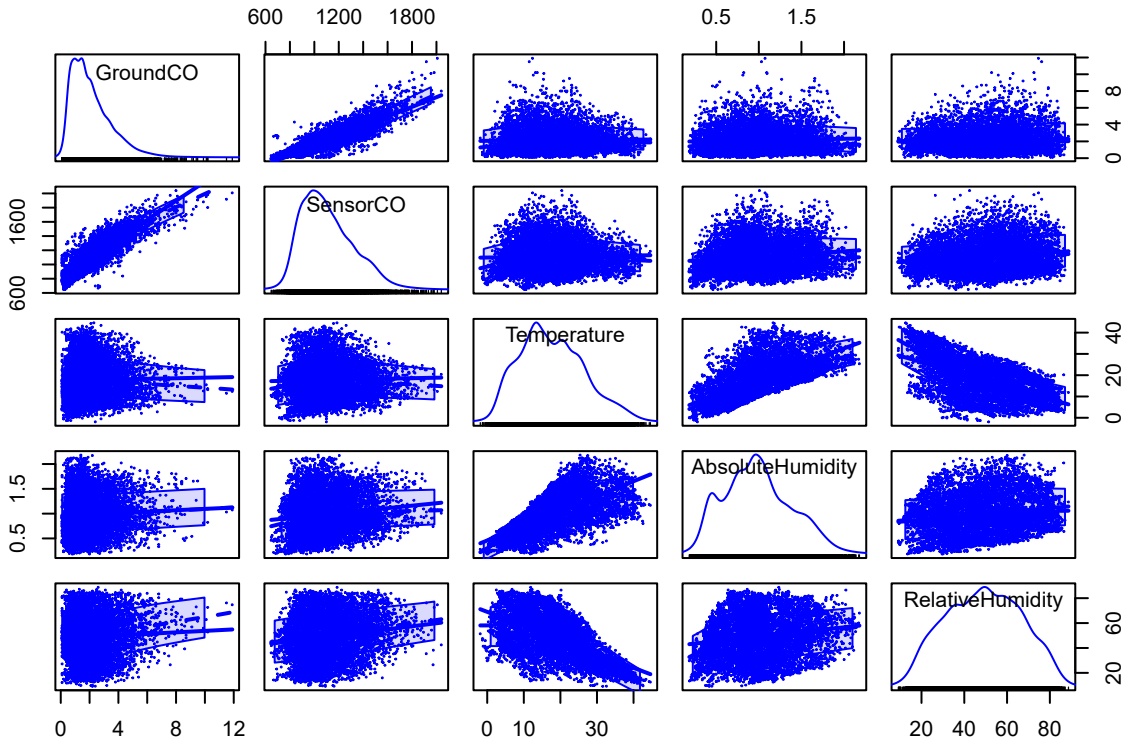
```
     ylab="Ground Truth (CO)", xlab="Sensor CO",
     pch=19, cex=0.1)
```



Scatterplot Matrix

We can also create a scatterplot matrix to visualize relationships among variables.

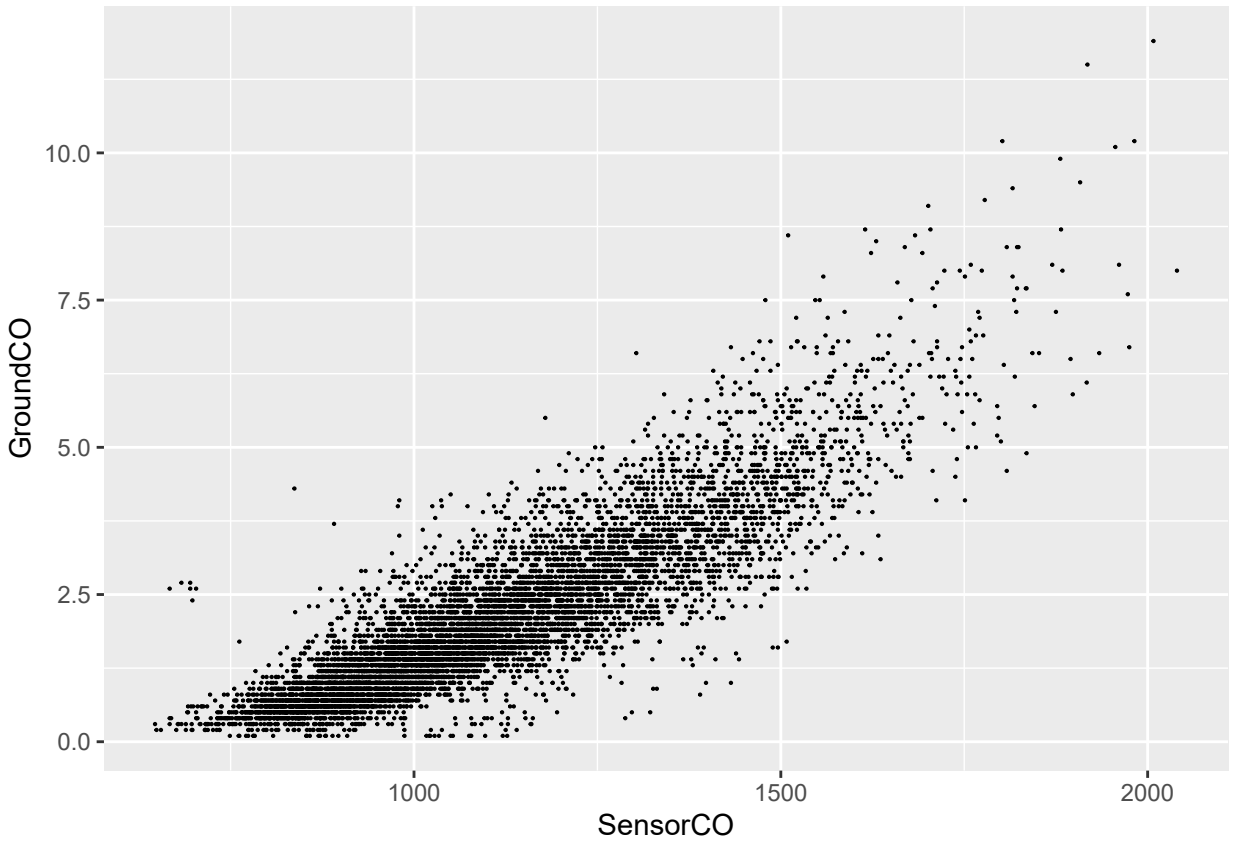
```
# Scatterplot matrix of all the variables using the scatterplotMatrix ...
# ... function from the car package
scatterplotMatrix(~ GroundCO + SensorCO + Temperature + AbsoluteHumidity +
                    RelativeHumidity,
                  pch=19, cex=0.1)
```



Scatterplot using ggplot2

Let's create a scatterplot using the `ggplot2` package.

```
# Scatterplot using ggplot2
# Create a data frame with SensorCO and GroundCO
co_data <- data.frame(SensorCO, GroundCO)
# ggplot function for scatterplot
ggplot(co_data, aes(x=SensorCO, y=GroundCO)) +
  geom_point(size=0.1)
```

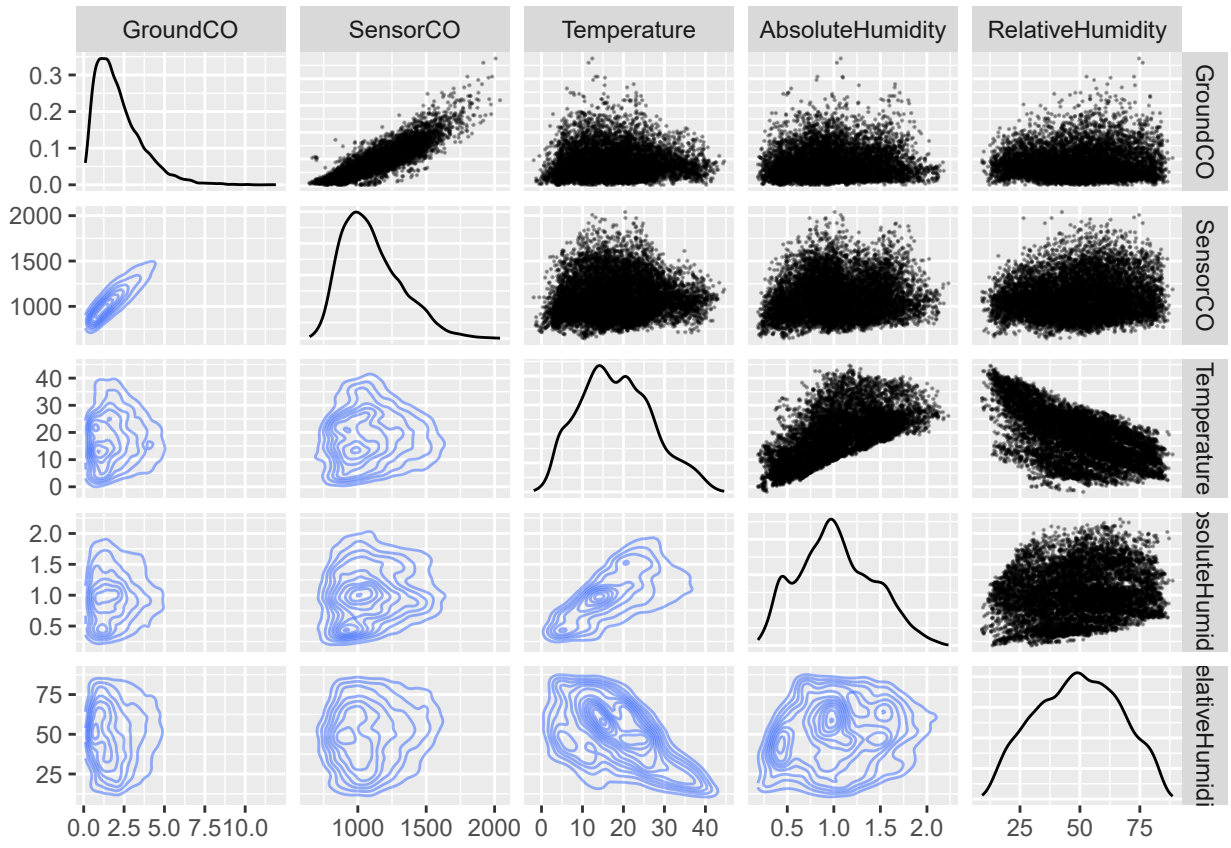


Scatterplot Matrix using ggplot2

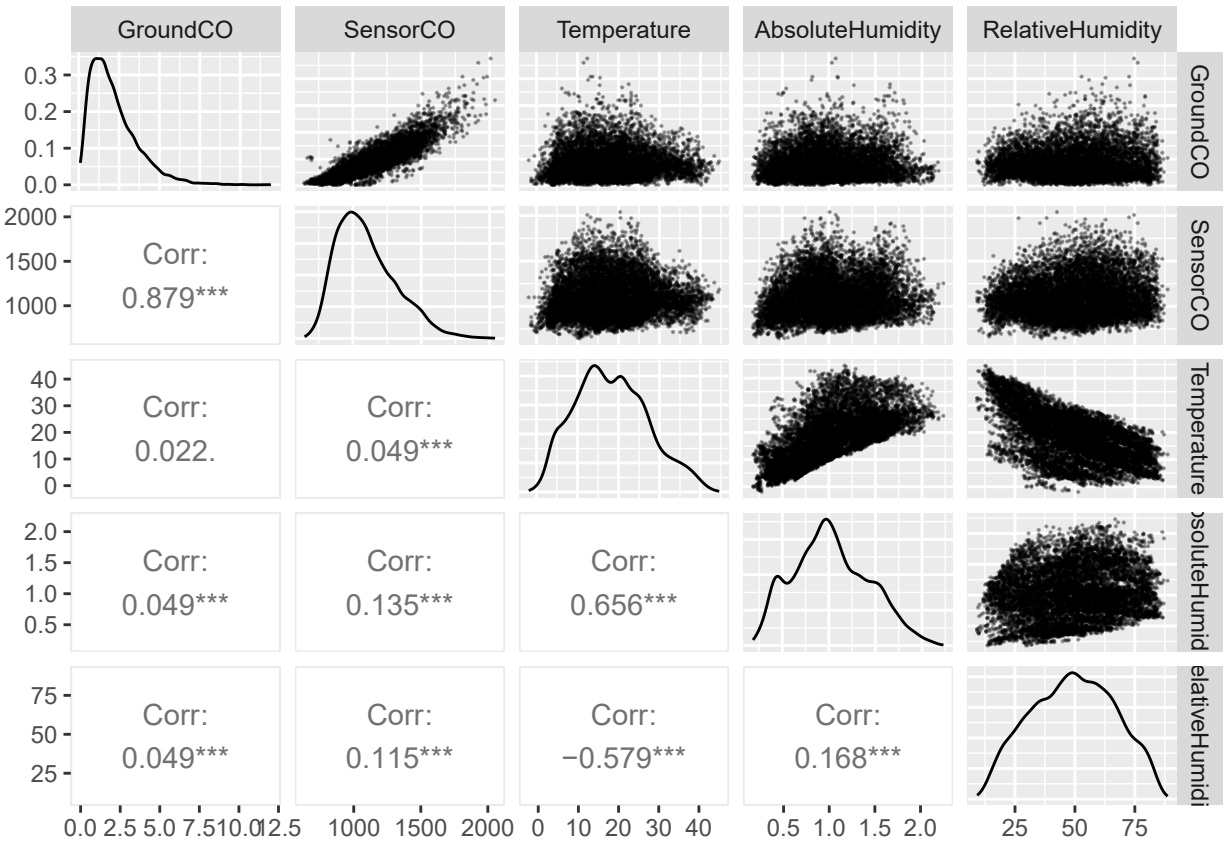
We can create a scatterplot matrix with correlation coefficients using `ggplot2`.

```
# Create a data frame with GroundCO, SensorCO, Temperature, ...
# ... AbsoluteHumidity, RelativeHumidity
co_temp_data <- data.frame(GroundCO, SensorCO, Temperature, AbsoluteHumidity, RelativeHumidity)

# Scatterplot matrix with 2D contours
ggpairs(co_temp_data,
        lower=list(continuous=wrap("density", alpha=0.5), combo="box"),
        upper=list(continuous=wrap("points", alpha=0.3, size=0.1)))
```




```
# Scatterplot matrix with correlation coefficients
ggpairs(co_temp_data,
        upper=list(continuous=wrap("points", alpha=0.3, size=0.1)),
        lower=list(continuous=wrap('cor', size=4)))
```



This concludes our data visualization lab. Remember to save this Rmd script and use an R Markdown editor or RStudio to knit it into a PDF. Ensure you have LaTeX installed to compile this script directly into a PDF instead of HTML; you will need to change the option of `html_document` to `pdf_document` in the header.