Philly Pollution

Ally Racho

2/8/2022

#load EDA library

```
library(DataExplorer)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
camden_spruce_st_newjersey_usa_air_quality <- read_csv("Data/camden-spruce st, newjersey, usa-air-quali
```

```
## Rows: 2859 Columns: 6
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): date
## dbl (5): pm25, o3, no2, so2, co
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

#summarize data statistically

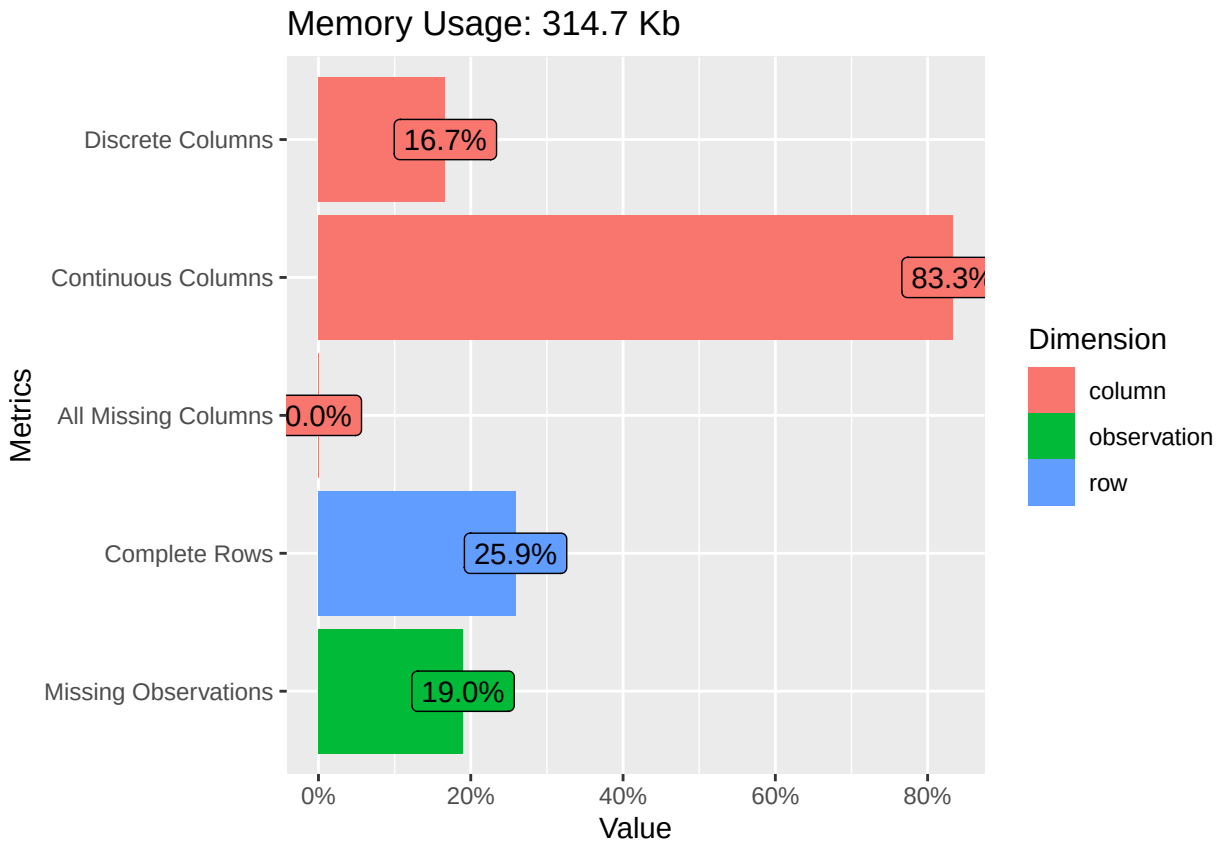
```
summary(camden_spruce_st_newjersey_usa_air_quality)
```

```
##      date                pm25             o3             no2
##  Length:2859        Min.   :  4.00   Min.   : 1.00   Min.   : 1.000
##  Class :character   1st Qu.: 28.00   1st Qu.:22.00   1st Qu.: 5.000
##  Mode  :character   Median : 38.00   Median :28.00   Median : 8.000
##                     Mean   : 39.68   Mean   :29.09   Mean   : 9.684
##                     3rd Qu.: 49.00   3rd Qu.:36.00   3rd Qu.:13.000
##                     Max.   :123.00   Max.   :89.00   Max.   :45.000
##                     NA's   :578      NA's   :20      NA's   :413
##       so2               co
##  Min.   : 1.000   Min.   : 1.000
##  1st Qu.: 1.000   1st Qu.: 1.000
##  Median : 1.000   Median : 2.000
##  Mean   : 1.349   Mean   : 2.276
```

```
##  3rd Qu.: 1.000   3rd Qu.: 3.000
##  Max.   :56.000   Max.   :11.000
##  NA's   :2044     NA's   :202
```

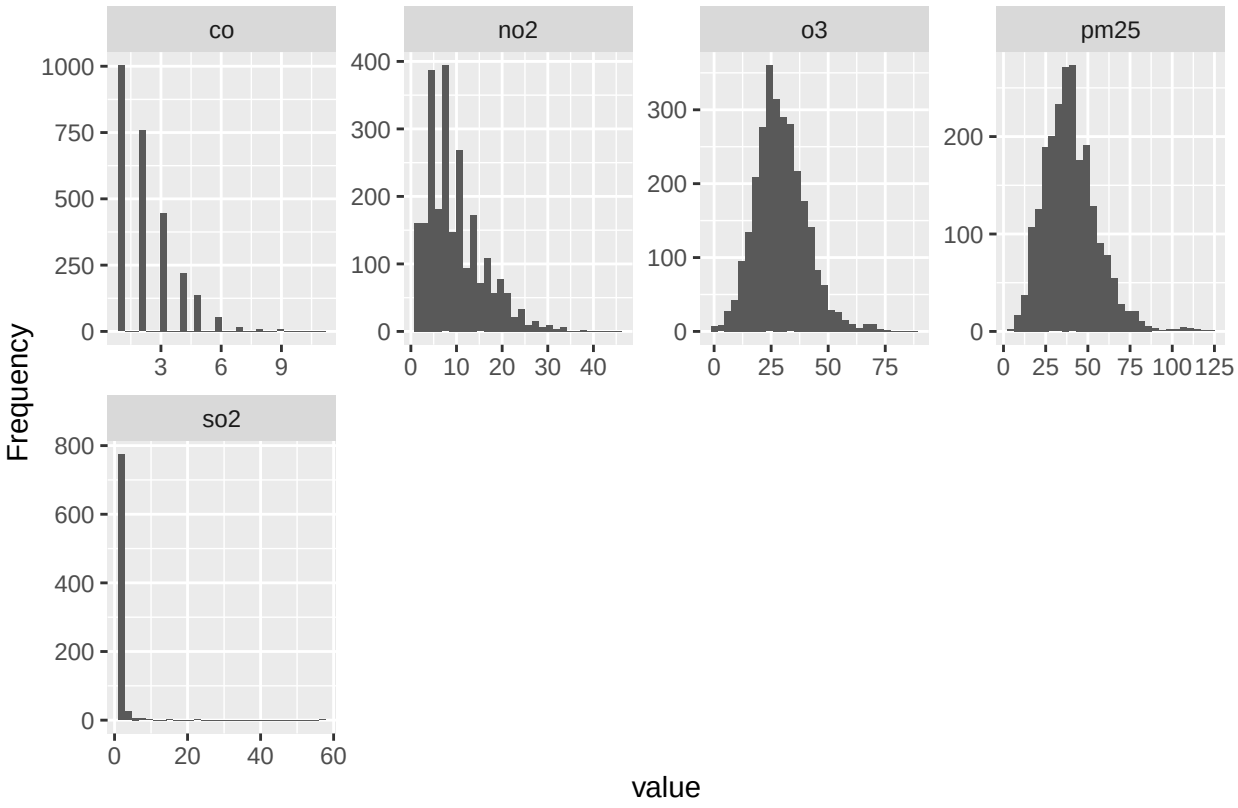
```
plot_str(camden_spruce_st_newjersey_usa_air_quality)
```

```
plot_intro(camden_spruce_st_newjersey_usa_air_quality)
```



#Histogram of pollutants

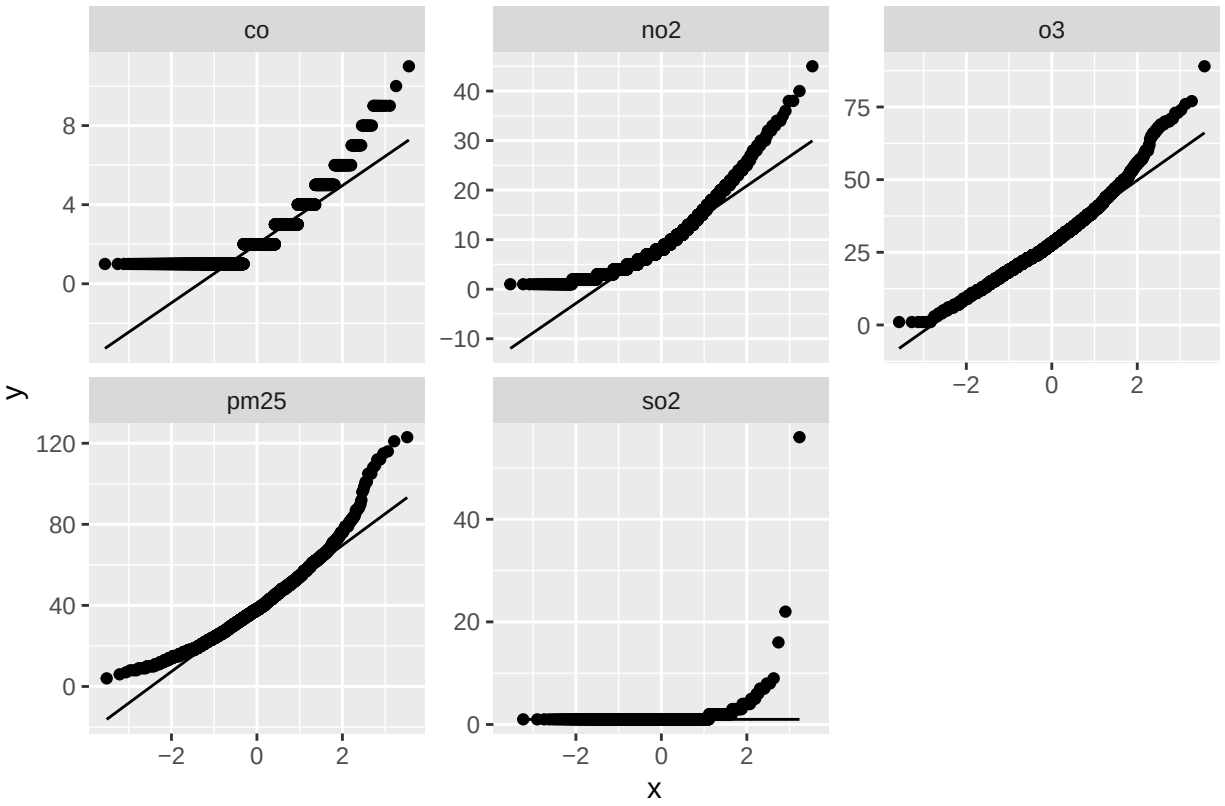
```
plot_histogram(camden_spruce_st_newjersey_usa_air_quality)
```




```
qq_plot <- plot_qq(camden_spruce_st_newjersey_usa_air_quality)
```

```
## Warning: Removed 3257 rows containing non-finite values (stat_qq).
```

```
## Warning: Removed 3257 rows containing non-finite values (stat_qq_line).
```



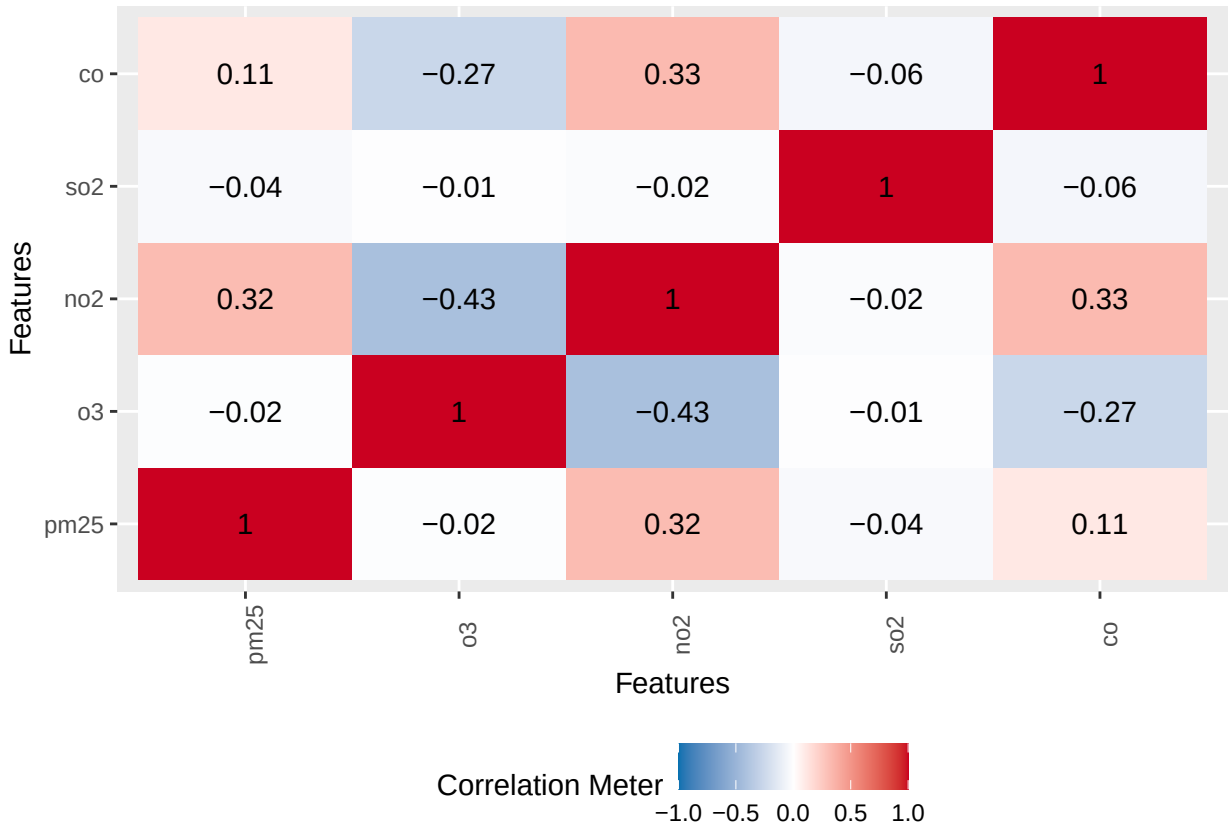
#co, no2, pm25 appear skewed on both tails – log transform and replot

```
#log_qq_data <- update_columns(qq_plot, c(1, 2, 4), function(x) log(x + 1))
#plot_qq(log_qq_data[, 1:2], sampled_rows = 1000L)
#plot_qq(log_qq_data[, 4], sampled_rows = 1000L)
```

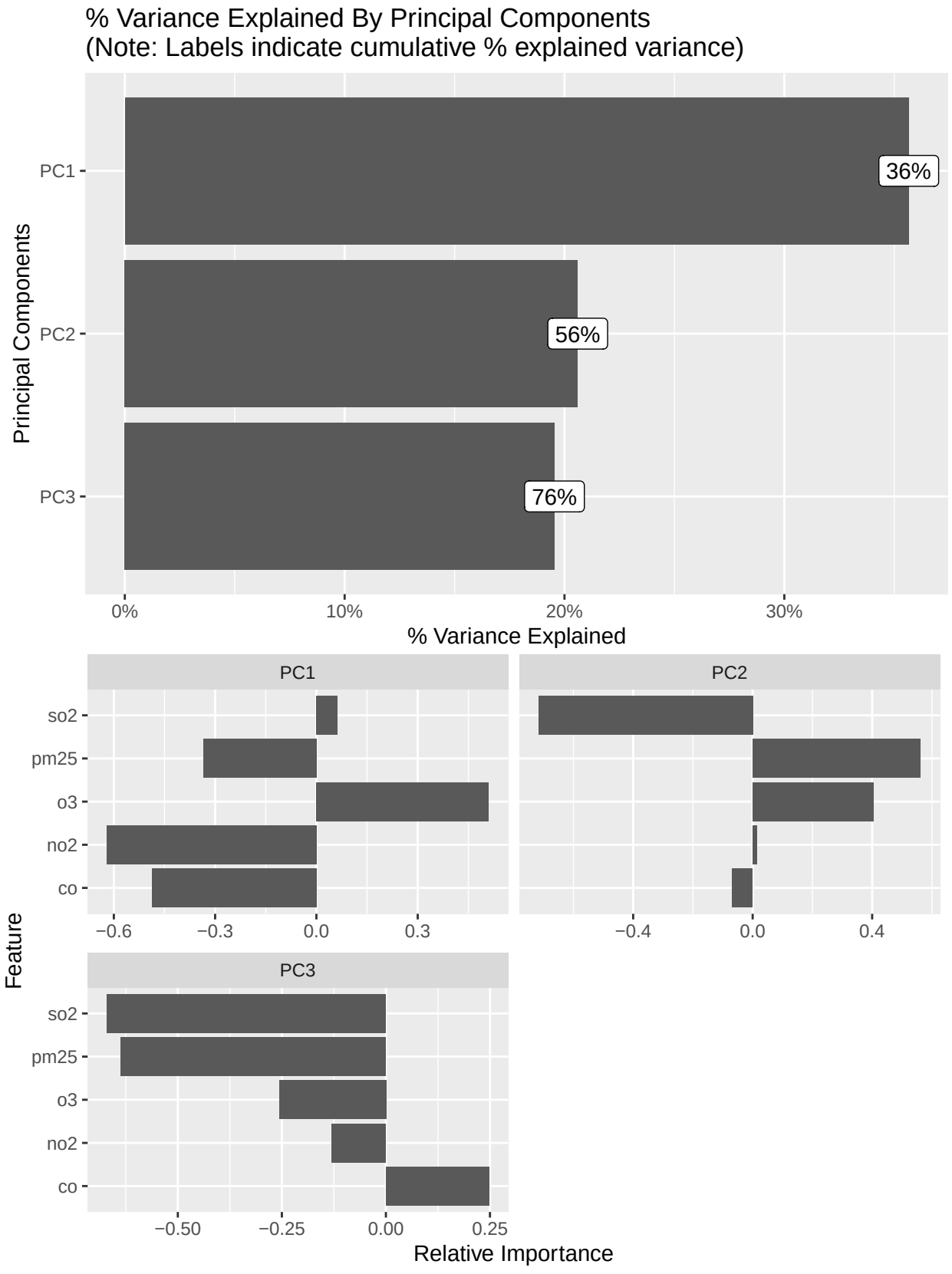
#correlation matrix excluding NAs

```
plot_correlation(na.omit(camden_spruce_st_newjersey_usa_air_quality), maxcat = 5L)
```

```
## Warning in dummify(data, maxcat = maxcat): Ignored all discrete features since
## `maxcat` set to 5 categories!
```

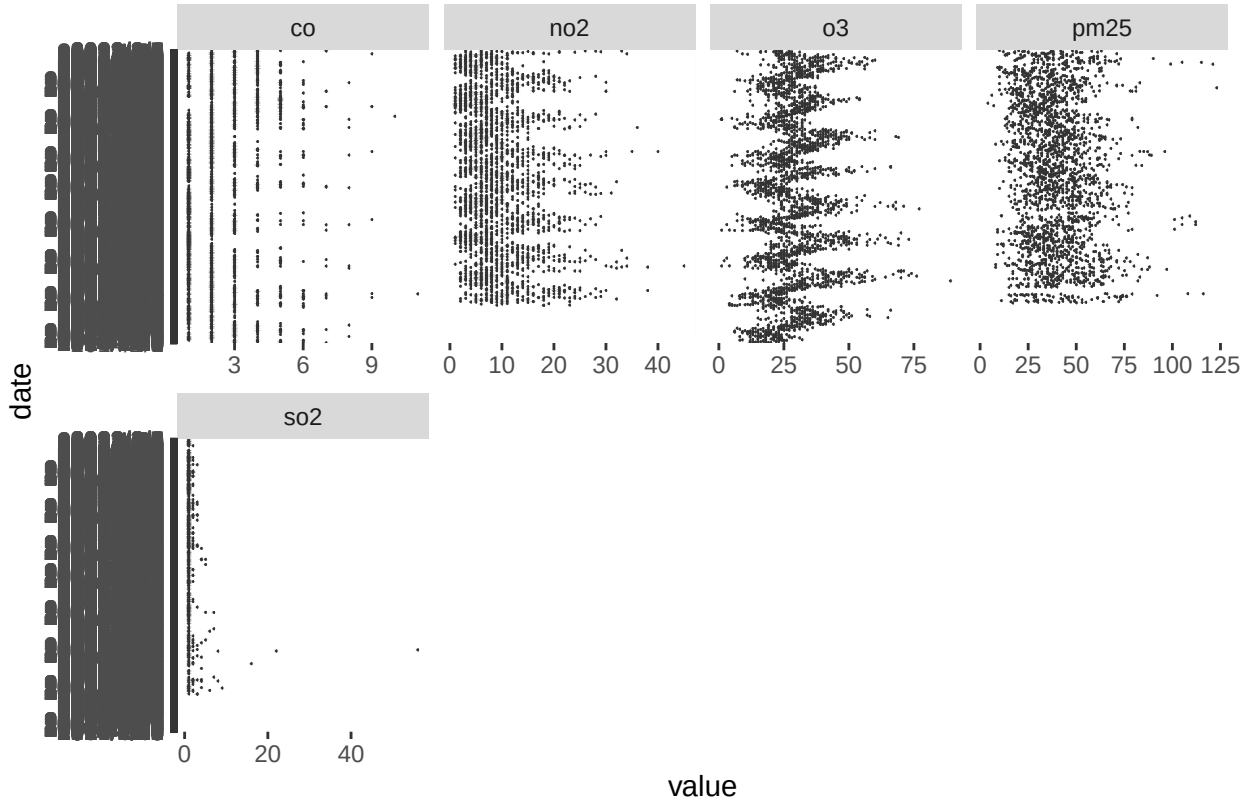



```
pca_df <- na.omit(camden_spruce_st_newjersey_usa_air_quality[, c("pm25", "o3", "no2", "so2", "co")])
plot_prcomp(pca_df, variance_cap = 0.9, nrow = 2L, ncol = 2L)
```


% Variance Explained By Principal Components
(Note: Labels indicate cumulative % explained variance)
Principal Components
PC1
PC2
PC3
36%
56%
76%
0%
10%
20%
30%
% Variance Explained
PC1
PC2
PC3
Feature
so2
pm25
o3
no2
co
−0.6
−0.3
0.0
0.3
−0.4
0.0
0.4
−0.50
−0.25
0.00
0.25
Relative Importance

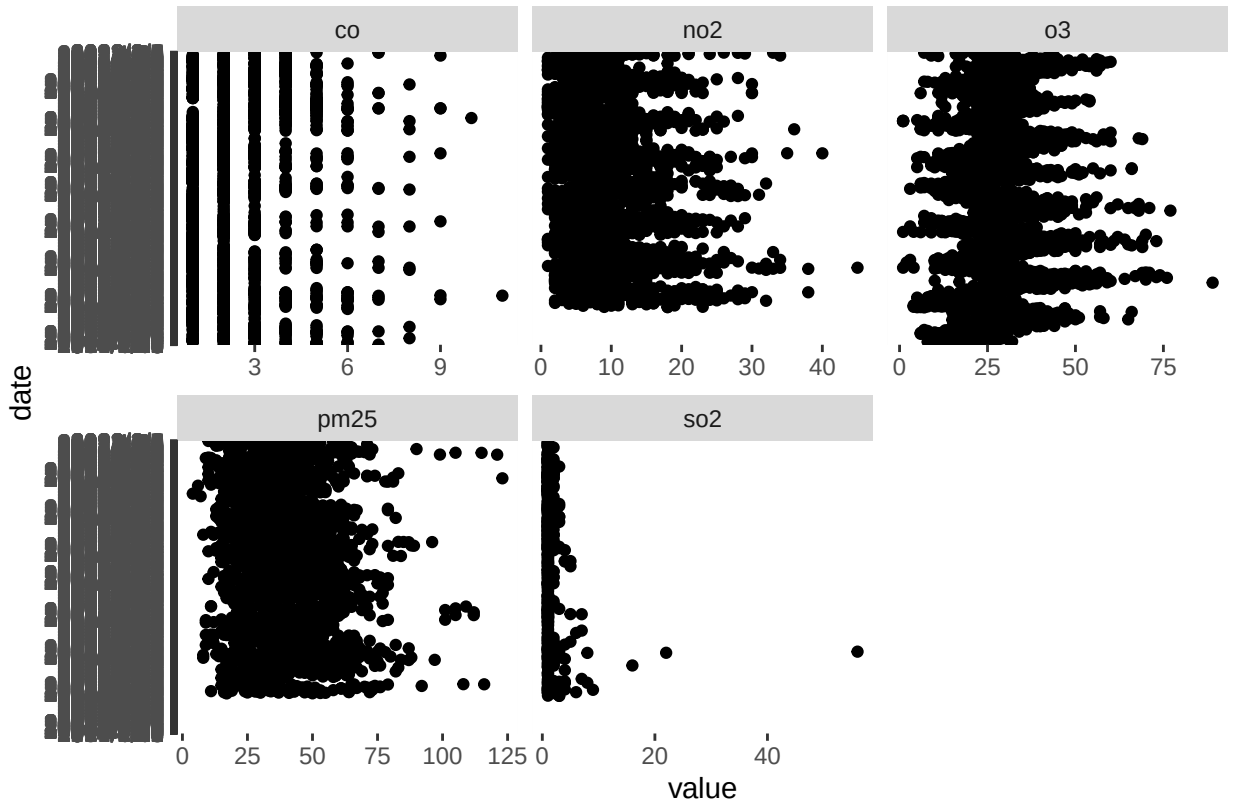
```
plot_boxplot(camden_spruce_st_newjersey_usa_air_quality, by = 'date')
```

```
## Warning: Removed 3257 rows containing non-finite values (stat_boxplot).
```




```
plot_scatterplot(camden_spruce_st_newjersey_usa_air_quality, by = 'date')
```

```
## Warning: Removed 3257 rows containing missing values (geom_point).
```


co
no2
o3
pm25
so2
date
value
3
6
9
0
10
20
30
40
0
25
50
75
0
25
50
75
100
125
0
20
40