Chicago 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)
chi_com_illinois_air_quality <- read_csv("Data/chi_com,-illinois-air-quality.csv")
```

```
## Rows: 2468 Columns: 3
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): date
## dbl (2): pm25, o3
##
## 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.
```

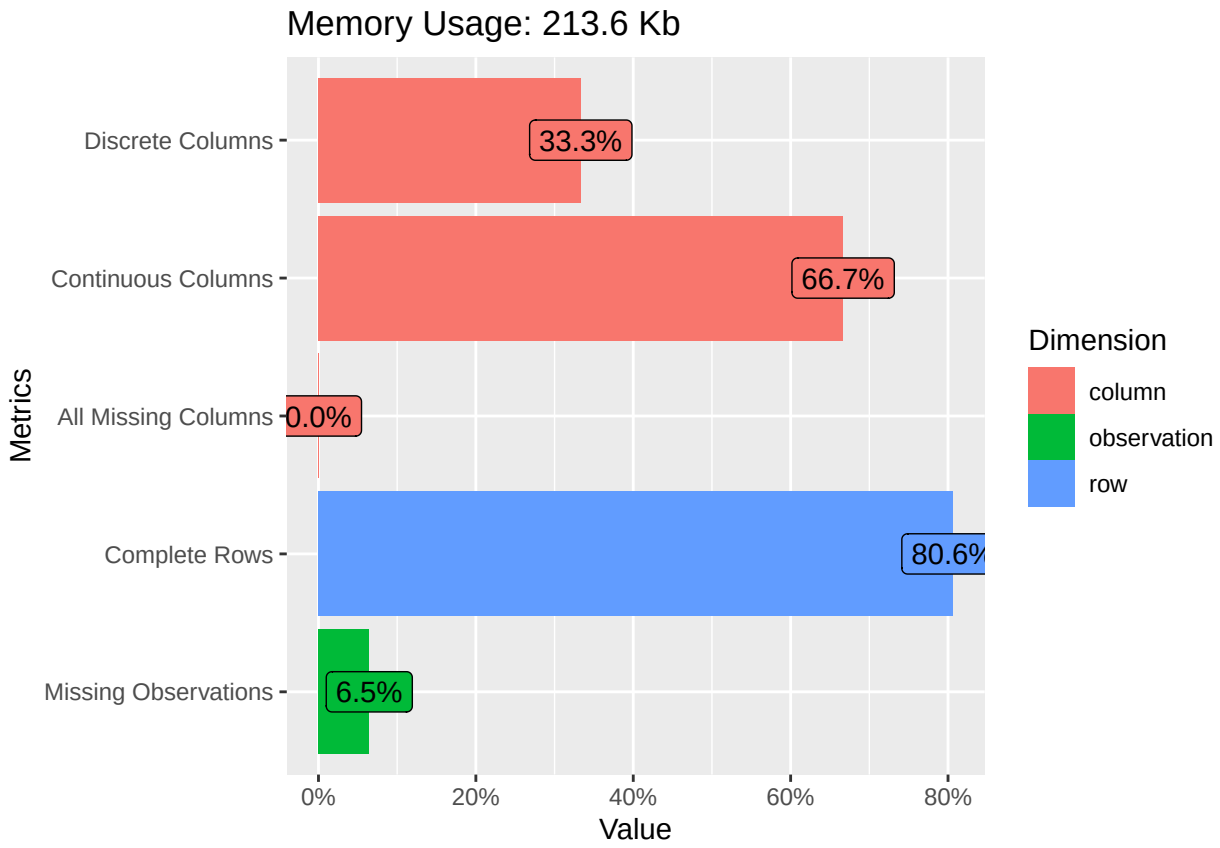
#summarize data statistically

```
summary(chi_com_illinois_air_quality)
```

```
##      date                pm25             o3
##  Length:2468        Min.   : 4.00   Min.   :  2.00
##  Class :character   1st Qu.:26.00   1st Qu.: 21.00
##  Mode  :character   Median :36.00   Median : 28.00
##                     Mean   :36.91   Mean   : 29.63
##                     3rd Qu.:47.00   3rd Qu.: 36.00
##                     Max.   :97.00   Max.   :104.00
##                     NA's   :55      NA's   :423
```

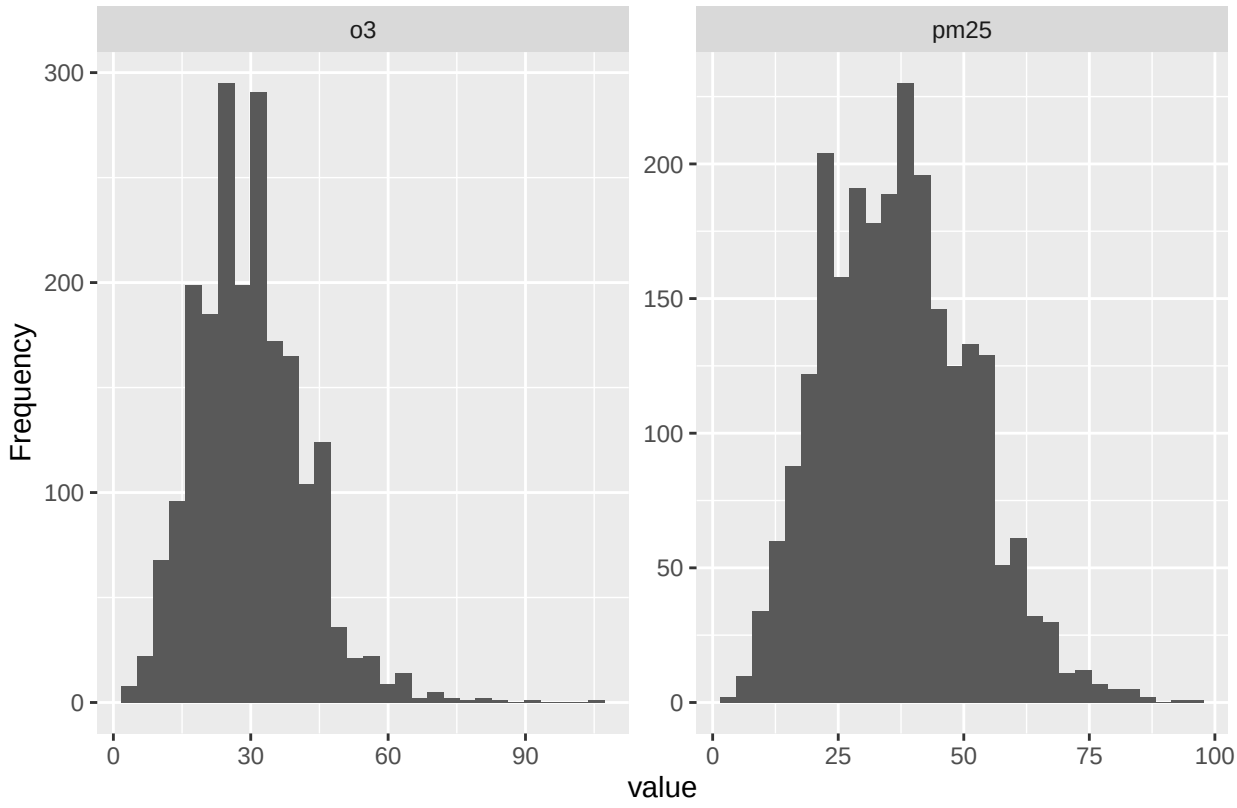
```
plot_str(chi_com_illinois_air_quality)
```

```
plot_intro(chi_com_illinois_air_quality)
```



#Histogram of pollutants

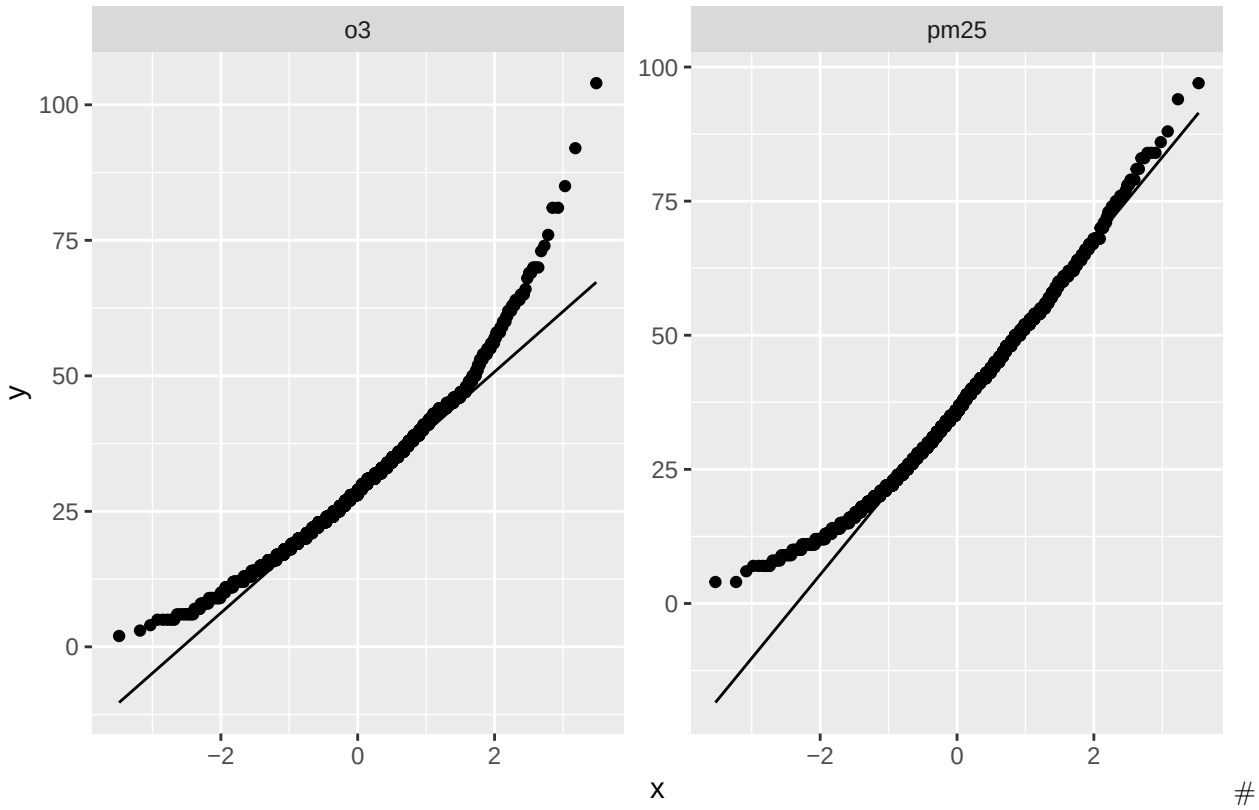
```
plot_histogram(chi_com_illinois_air_quality)
```




```
qq_plot <- plot_qq(chi_com_illinois_air_quality)
```

```
## Warning: Removed 478 rows containing non-finite values (stat_qq).
```

```
## Warning: Removed 478 rows containing non-finite values (stat_qq_line).
```



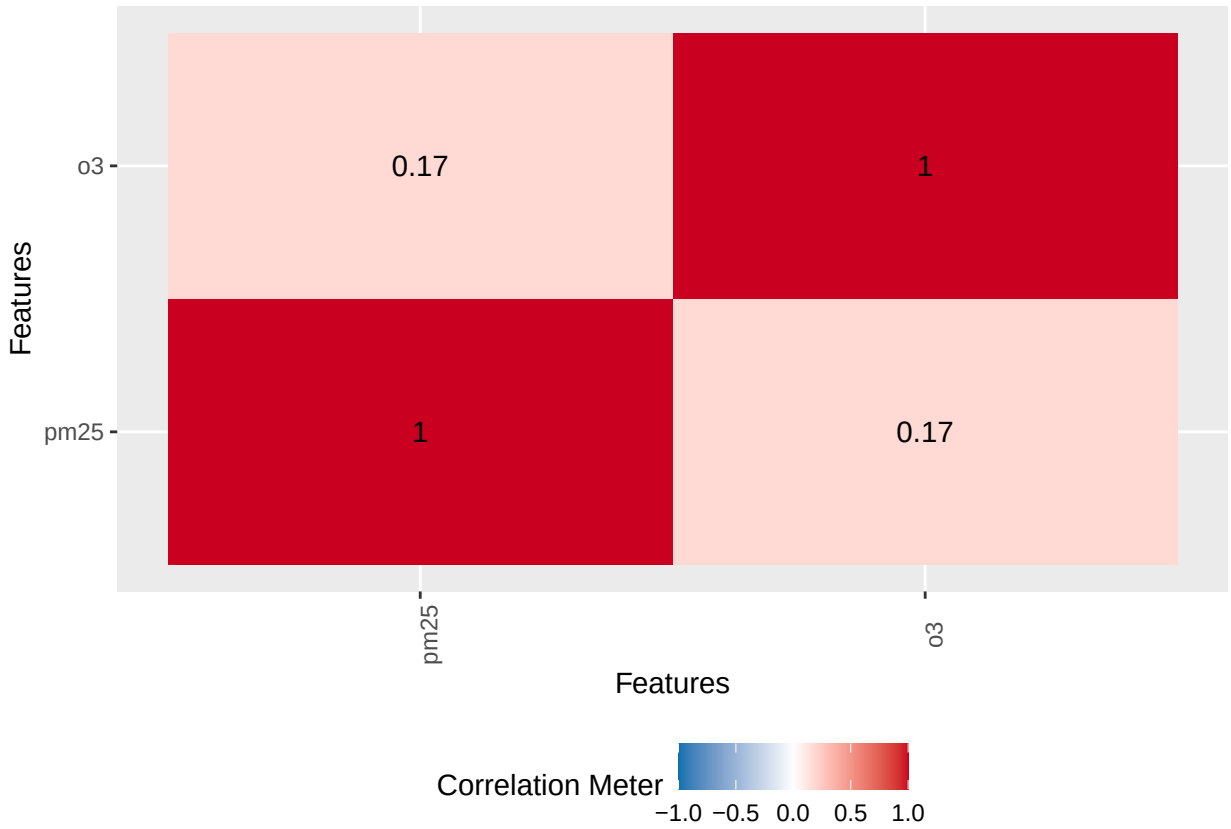
o3 appears skewed on both ends

```
#log_qq_data <- update_columns(qq_plot, 'o3', function(x) log(x + 1))
#plot_qq(log_qq_data[3], sampled_rows = 1000L)
```

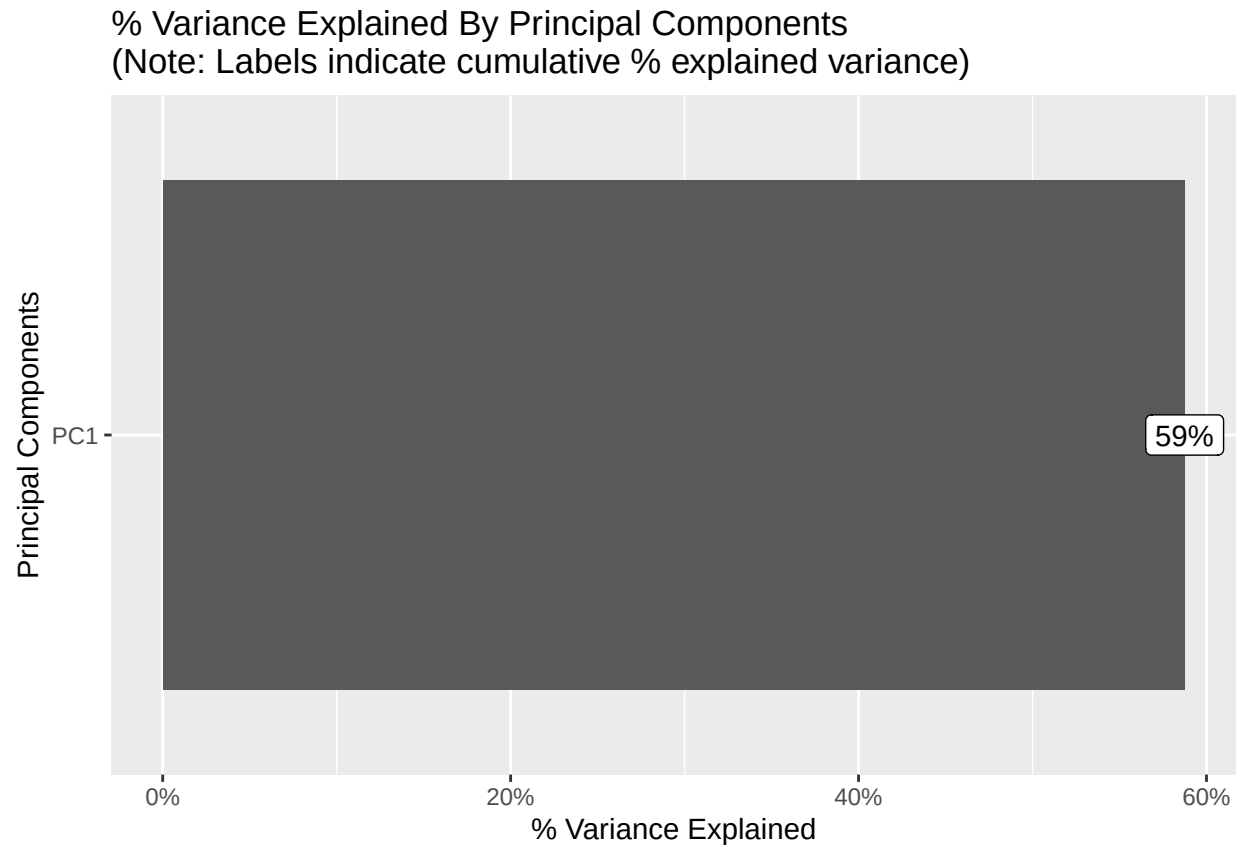
#correlation matrix excluding NAs

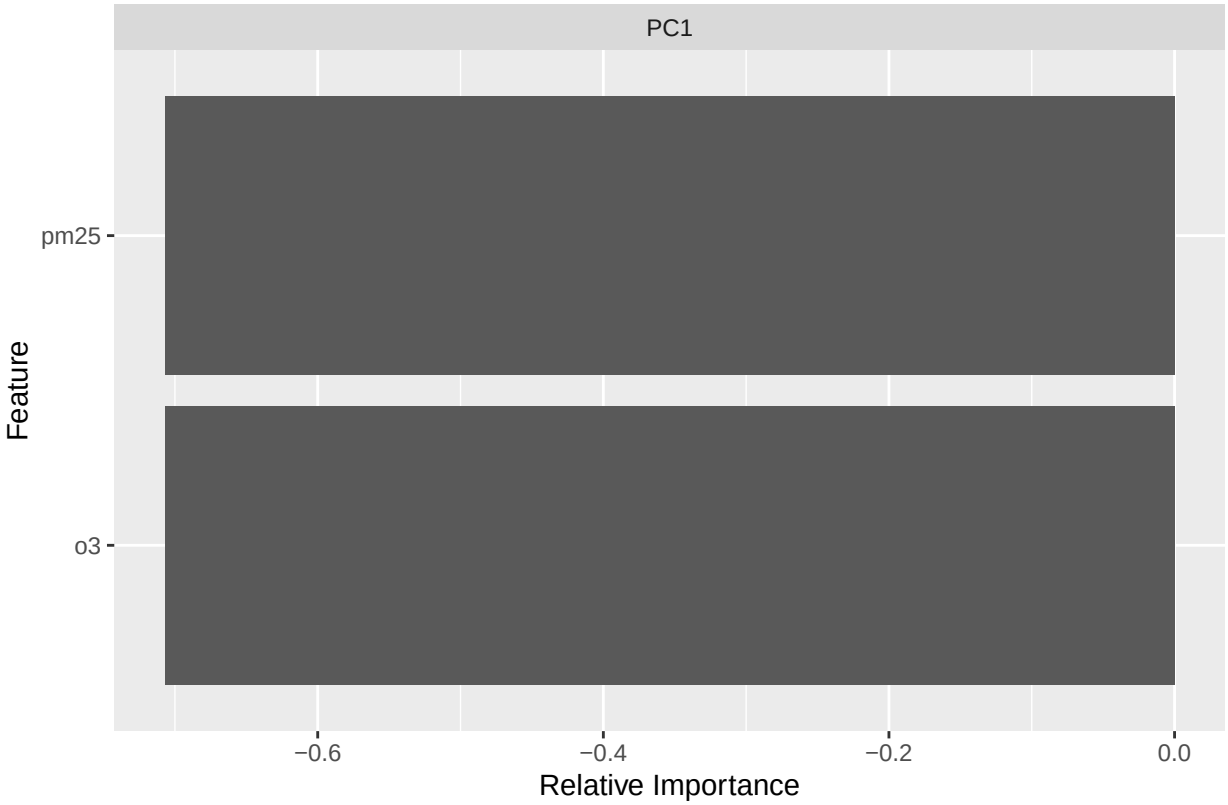
```
plot_correlation(na.omit(chi_com_illinois_air_quality), maxcat = 5L)
```

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



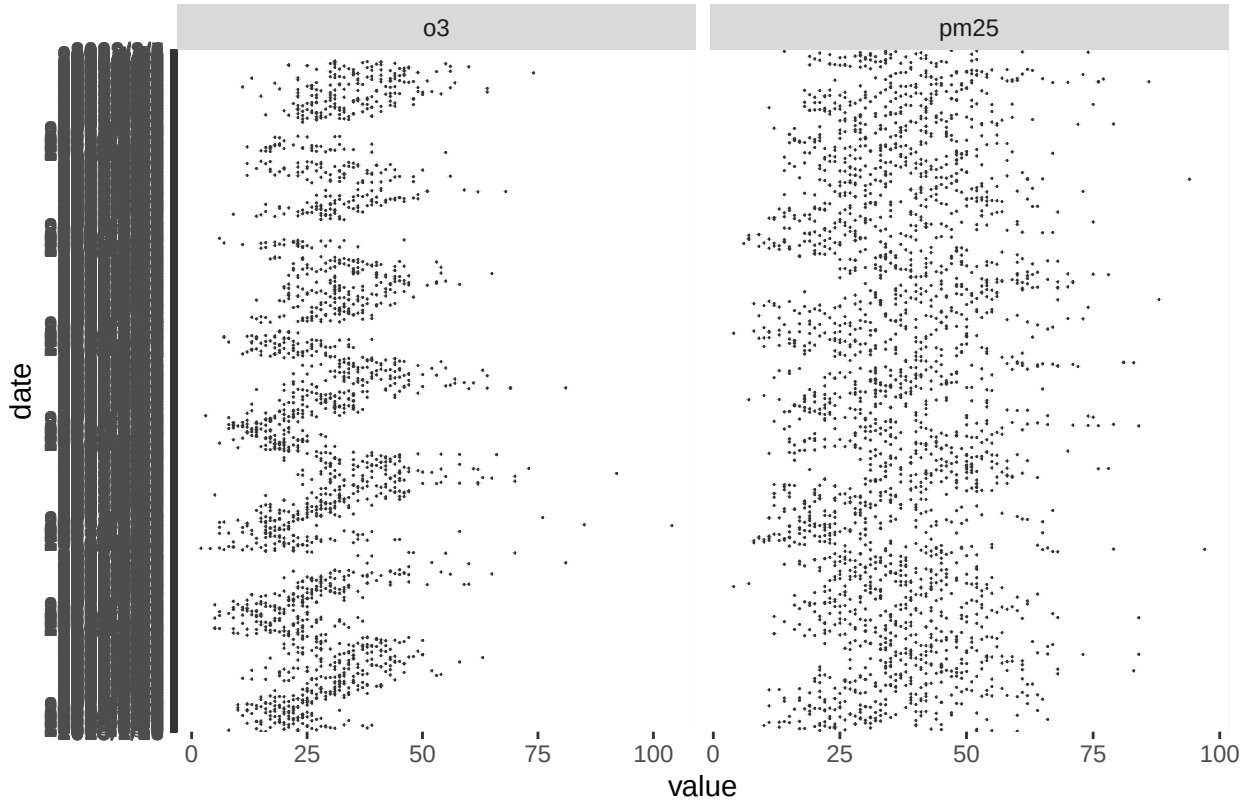

```
pca_df <- na.omit(chi_com_illinois_air_quality[, c("pm25", "o3")])
plot_prcomp(pca_df, variance_cap = 0.9, nrow = 2L, ncol = 2L)
```


% Variance Explained By Principal Components
(Note: Labels indicate cumulative % explained variance)
Principal Components
PC1
59%
0%
20%
40%
60%
% Variance Explained


PC1
Feature
pm25
o3
−0.6
−0.4
−0.2
0.0
Relative Importance

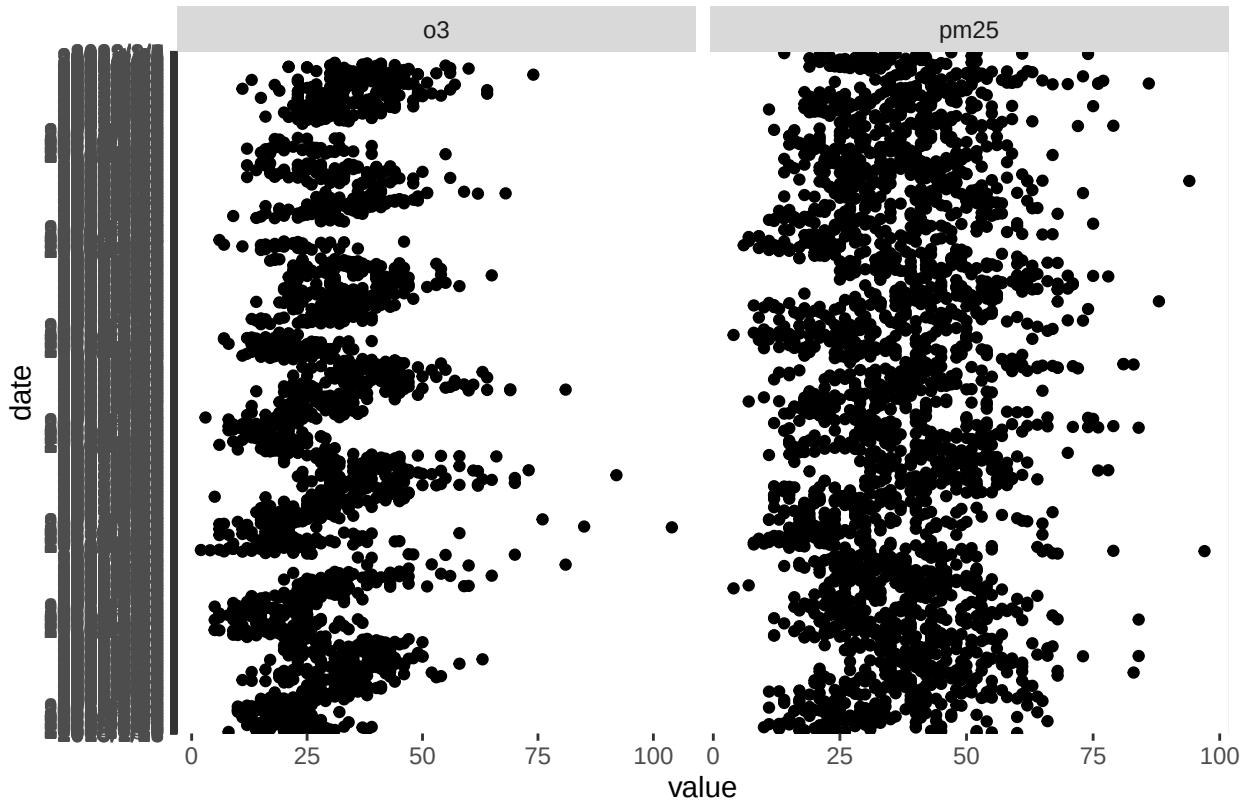
```
plot_boxplot(chi_com_illinois_air_quality, by = 'date')
```

```
## Warning: Removed 478 rows containing non-finite values (stat_boxplot).
```




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

```
## Warning: Removed 478 rows containing missing values (geom_point).
```


o3
pm25
date
0
25
50
75
100
value