FinalProject-Kecerdasan Bisnis

Kelompok 1

2024-12-02

0. Import Library

```
if (!requireNamespace("tidyverse")) install.packages("tidyverse")
```

```
## Loading required namespace: tidyverse
```

```
if (!requireNamespace("readr")) install.packages("readr")
```

```
## Loading required namespace: readr
```

```
if (!requireNamespace("writexl")) install.packages("writexl")
```

```
## Loading required namespace: writexl
```

```
if (!requireNamespace("readxl")) install.packages("readxl")
```

```
## Loading required namespace: readxl
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
```

```
## -- 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
```

```
library (readr)
library(writexl)
library(readxl)
```

1. Import Dataset

```
# import data dari website
Enviro <- read.csv("https://data.ca.gov/dataset/0bd5f40b-c59b-4183-be22-d057eae8383c/resource/89b3f4e9-0
```

```
# cek struktur data
str(Enviro)
```

```
## 'data.frame':    8035 obs. of  57 variables:
##  $ Census.Tract                                    : num  6.02e+09 6.07e+09 6.02e+09 6.08e+09 6.02e+0
##  $ Total.Population                                : int  3174 6133 3167 6692 2206 2598 2396 4106 214
##  $ California.County                               : chr  "Fresno " "San Bernardino" "Fresno " "San .
##  $ ZIP                                             : int  93706 91761 93706 95203 93725 90023 95203 9
##  $ Nearby.City...to.help.approximate.location.only.: chr  "Fresno" "Ontario" "Fresno" "Stockton" ...
##  $ Longitude                                       : num  -120 -118 -120 -121 -120 ...
##  $ Latitude                                        : num  36.7 34.1 36.7 37.9 36.7 ...
##  $ CES.3.0.Score                                   : num  94.1 90.7 86 82.5 82 ...
##  $ CES.3.0.Percentile                              : num  100 100 100 100 100 ...
##  $ CES.3.0..Percentile.Range                       : chr  "95-100% (highest scores)" "95-100% (highes
##  $ SB.535.Disadvantaged.Community                  : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ Ozone                                           : num  0.065 0.062 0.062 0.046 0.065 0.046 0.046 0
##  $ Ozone.Pctl                                      : num  98.2 91.1 91.1 53 98.2 ...
##  $ PM2.5                                           : num  15.4 13.3 15.4 12.5 15.4 ...
##  $ PM2.5.Pctl                                      : num  97.2 93.6 97.2 84 97.2 ...
##  $ Diesel.PM                                       : num  48.5 38.6 47.4 24.1 18.8 ...
##  $ Diesel.PM.Pctl                                  : num  95.5 92.1 95.4 73.5 58.2 ...
##  $ Drinking.Water                                  : num  681 905 681 279 1000 ...
##  $ Drinking.Water.Pctl                             : num  80.9 96.1 80.9 29.1 98.6 ...
##  $ Pesticides                                      : num  2.75 1.37 3.03 12.93 3518.41 ...
##  $ Pesticides.Pctl                                 : num  47.8 41.3 48.8 60.6 95.2 ...
##  $ Tox..Release                                    : num  18552 7494 12455 2388 21791 ...
##  $ Tox..Release.Pctl                               : num  97.5 89 95.4 70 98.2 ...
##  $ Traffic                                         : num  909 782 577 1305 435 ...
##  $ Traffic.Pctl                                    : num  63 55.7 39 78.3 24.3 ...
##  $ Cleanup.Sites                                   : num  80.5 66.2 22 50.1 60 ...
##  $ Cleanup.Sites.Pctl                              : num  98.7 97.7 85.1 96.1 97.2 ...
##  $ Groundwater.Threats                             : num  45.8 36 30.2 132.1 54.2 ...
##  $ Groundwater.Threats.Pctl                        : num  89.8 85.6 81.9 98.4 92.1 ...
##  $ Haz..Waste                                      : num  0.795 1.25 0.2 0.795 13.1 ...
##  $ Haz..Waste.Pctl                                 : num  84.3 88.8 60.5 84.3 99.7 ...
##  $ Imp..Water.Bodies                               : int  0 5 0 19 0 7 14 0 7 0 ...
##  $ Imp..Water.Bodies.Pctl                          : num  0 55 0 98.6 0 ...
##  $ Solid.Waste                                     : num  21.8 12 2.5 27 50.8 ...
##  $ Solid.Waste.Pctl                                : num  97.8 92.2 57.2 99.1 99.9 ...
##  $ Pollution.Burden                                : num  80 81.2 71.2 74.5 80.2 ...
##  $ Pollution.Burden.Score                          : num  9.85 10 8.76 9.17 9.88 9.45 8.41 8.24 9.5 9
##  $ Pollution.Burden.Pctl                           : num  100 100 99 99.6 100 ...
##  $ Asthma                                          : num  131.6 60.7 142.1 142.2 90.5 ...
##  $ Asthma.Pctl                                     : num  97.7 69.8 98.3 98.3 89.5 ...
##  $ Low.Birth.Weight                                : num  7.44 7.04 10.16 6.23 4.5 ...
##  $ Low.Birth.Weight.Pctl                           : num  93.8 90.8 99.8 80.7 38.9 ...
##  $ Cardiovascular.Disease                          : num  14.1 12.9 15 14.7 12.8 ...
```

```
##  $ Cardiovascular.Disease.Pctl                    : num  96.3 92.7 97.7 97.2 92.4 ...
##  $ Education                                      : num  53.3 53.3 42.3 40.8 45.1 53.1 46 47.4 50.4
##  $ Education.Pctl                                 : num  95.8 95.8 89.1 87.5 91.1 ...
##  $ Linguistic.Isolation                           : num  16.2 33.4 16.7 15.3 14.7 23.7 27.1 15.8 35
##  $ Linguistic.Isolation.Pctl                      : num  77.5 96.2 78.4 75.1 73.7 ...
##  $ Poverty                                        : num  76.3 72.5 86.8 61.3 66.4 66.4 76.2 74.5 75
##  $ Poverty.Pctl                                   : num  97.1 94.6 99.6 85.6 90.2 ...
##  $ Unemployment                                   : num  17.6 12.3 16.1 19.6 18.6 11.6 14.4 20 28.5
##  $ Unemployment.Pctl                              : num  91.7 71.8 88 95 93.7 ...
##  $ Housing.Burden                                 : num  26 34.1 40.1 21.1 28.1 22 24.3 31.8 31.7 2
##  $ Housing.Burden.Pctl                            : num  79.4 93.8 97.8 63.5 84 ...
##  $ Pop..Char.                                     : num  92.1 87.4 94.6 86.7 80.1 ...
##  $ Pop..Char..Score                               : num  9.55 9.07 9.81 8.99 8.3 8.54 9.53 9.73 8.3
##  $ Pop..Char..Pctl                                : num  99.7 98.1 100 97.7 92.8 ...
```

```
# deskriptif data
summary(Enviro)
```

```
##   Census.Tract       Total.Population California.County       ZIP
##  Min.   :6.001e+09   Min.   :    0    Length:8035        Min.   :   32
##  1st Qu.:6.037e+09   1st Qu.: 3358    Class :character   1st Qu.:91602
##  Median :6.059e+09   Median : 4413    Mode  :character   Median :92691
##  Mean   :6.055e+09   Mean   : 4636                       Mean   :92837
##  3rd Qu.:6.073e+09   3rd Qu.: 5656                       3rd Qu.:94558
##  Max.   :6.115e+09   Max.   :37452                       Max.   :96161
##
##  Nearby.City...to.help.approximate.location.only.   Longitude
##  Length:8035                                      Min.   :-124.3
##  Class :character                                 1st Qu.:-121.5
##  Mode  :character                                 Median :-118.4
##                                                   Mean   :-119.4
##                                                   3rd Qu.:-117.9
##                                                   Max.   :-114.3
##
##     Latitude     CES.3.0.Score   CES.3.0.Percentile CES.3.0..Percentile.Range
##  Min.   :32.55   Min.   : 0.98   Min.   :  0.01     Length:8035
##  1st Qu.:33.92   1st Qu.:14.96   1st Qu.: 25.01     Class :character
##  Median :34.21   Median :25.06   Median : 50.01     Mode  :character
##  Mean   :35.50   Mean   :27.93   Mean   : 50.01
##  3rd Qu.:37.63   3rd Qu.:39.35   3rd Qu.: 75.00
##  Max.   :41.95   Max.   :94.09   Max.   :100.00
##                  NA's   :106     NA's   :106
##  SB.535.Disadvantaged.Community     Ozone           Ozone.Pctl
##  Length:8035                    Min.   :0.02600   Min.   :  0.24
##  Class :character               1st Qu.:0.04000   1st Qu.: 25.87
##  Mode  :character               Median :0.04600   Median : 53.02
##                                 Mean   :0.04743   Mean   : 53.30
##                                 3rd Qu.:0.05500   3rd Qu.: 77.87
##                                 Max.   :0.06800   Max.   :100.00
##
##      PM2.5          PM2.5.Pctl       Diesel.PM       Diesel.PM.Pctl
##  Min.   : 1.651   Min.   :  0.01   Min.   :  0.021   Min.   :  0.01
##  1st Qu.: 8.698   1st Qu.: 30.70   1st Qu.:  8.812   1st Qu.: 25.01
##  Median :10.370   Median : 52.61   Median : 16.448   Median : 50.01
```

```
##  Mean   :10.378   Mean   : 53.59   Mean   : 19.196   Mean   : 50.02
##  3rd Qu.:12.050   3rd Qu.: 81.66   3rd Qu.: 24.646   3rd Qu.: 75.00
##  Max.   :19.600   Max.   :100.00   Max.   :253.731   Max.   :100.00
##  NA's   :19       NA's   :19
##  Drinking.Water    Drinking.Water.Pctl   Pesticides       Pesticides.Pctl
##  Min.   :   6.92   Min.   :  0.01      Min.   :    0.00   Min.   :  0.00
##  1st Qu.: 249.35   1st Qu.: 25.01      1st Qu.:    0.00   1st Qu.:  0.00
##  Median : 479.23   Median : 51.02      Median :    0.00   Median :  0.00
##  Mean   : 472.37   Mean   : 50.34      Mean   :  313.97   Mean   : 17.98
##  3rd Qu.: 664.07   3rd Qu.: 78.57      3rd Qu.:    0.37   3rd Qu.: 30.45
##  Max.   :1245.65   Max.   :100.00      Max.   :91316.19   Max.   :100.00
##  NA's   :18        NA's   :18
##   Tox..Release      Tox..Release.Pctl    Traffic          Traffic.Pctl
##  Min.   :     0.0   Min.   :  0.00    Min.   :   22.41   Min.   :  0.01
##  1st Qu.:    94.8   1st Qu.: 24.85    1st Qu.:  442.08   1st Qu.: 25.01
##  Median :   474.0   Median : 49.90    Median :  699.89   Median : 50.01
##  Mean   :  3182.7   Mean   : 49.90    Mean   :  943.04   Mean   : 50.01
##  3rd Qu.:  3474.2   3rd Qu.: 74.95    3rd Qu.: 1190.08   3rd Qu.: 75.00
##  Max.   :842751.3   Max.   :100.00    Max.   :45687.87   Max.   :100.00
##                                       NA's   :56         NA's   :56
##  Cleanup.Sites    Cleanup.Sites.Pctl Groundwater.Threats
##  Min.   :  0.00   Min.   :  0.00     Min.   :   0.0
##  1st Qu.:  0.00   1st Qu.:  0.00     1st Qu.:   0.2
##  Median :  2.00   Median : 27.29     Median :   5.6
##  Mean   :  8.37   Mean   : 34.45     Mean   :  15.7
##  3rd Qu.: 10.30   3rd Qu.: 63.41     3rd Qu.:  17.8
##  Max.   :323.75   Max.   :100.00     Max.   :1610.2
##
##  Groundwater.Threats.Pctl   Haz..Waste      Haz..Waste.Pctl  Imp..Water.Bodies
##  Min.   :  0.00           Min.   : 0.0000   Min.   :  0.00   Min.   : 0.000
##  1st Qu.:  0.33           1st Qu.: 0.0000   1st Qu.:  0.00   1st Qu.: 0.000
##  Median : 33.60           Median : 0.0500   Median : 25.76   Median : 1.000
##  Mean   : 38.02           Mean   : 0.4534   Mean   : 34.71   Mean   : 3.279
##  3rd Qu.: 66.78           3rd Qu.: 0.2250   3rd Qu.: 63.00   3rd Qu.: 6.000
##  Max.   :100.00           Max.   :28.6950   Max.   :100.00   Max.   :34.000
##
##  Imp..Water.Bodies.Pctl  Solid.Waste     Solid.Waste.Pctl Pollution.Burden
##  Min.   :  0.00         Min.   : 0.000   Min.   :  0.00   Min.   : 8.37
##  1st Qu.:  0.00         1st Qu.: 0.000   1st Qu.:  0.00   1st Qu.:32.24
##  Median : 15.26         Median : 0.200   Median :  9.08   Median :41.80
##  Mean   : 30.68         Mean   : 2.233   Mean   : 27.33   Mean   :41.97
##  3rd Qu.: 63.17         3rd Qu.: 2.250   3rd Qu.: 52.84   3rd Qu.:51.02
##  Max.   :100.00         Max.   :97.800   Max.   :100.00   Max.   :81.19
##
##  Pollution.Burden.Score Pollution.Burden.Pctl     Asthma        Asthma.Pctl
##  Min.   : 1.030         Min.   :  0.01        Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 3.970         1st Qu.: 25.01        1st Qu.: 29.86   1st Qu.: 24.88
##  Median : 5.150         Median : 50.01        Median : 45.27   Median : 49.93
##  Mean   : 5.169         Mean   : 50.01        Mean   : 51.98   Mean   : 49.93
##  3rd Qu.: 6.280         3rd Qu.: 75.00        3rd Qu.: 65.99   3rd Qu.: 74.96
##  Max.   :10.000         Max.   :100.00        Max.   :278.83   Max.   :100.00
##
##  Low.Birth.Weight Low.Birth.Weight.Pctl Cardiovascular.Disease
##  Min.   : 0.000   Min.   :  0.00        Min.   : 0.000
```

```
##  1st Qu.: 3.950   1st Qu.: 24.98        1st Qu.: 6.080
##  Median : 4.920   Median : 50.22        Median : 7.940
##  Mean   : 4.976   Mean   : 50.04        Mean   : 8.266
##  3rd Qu.: 5.930   3rd Qu.: 75.06        3rd Qu.:10.040
##  Max.   :14.890   Max.   :100.00        Max.   :21.260
##  NA's   :222      NA's   :222
##  Cardiovascular.Disease.Pctl   Education     Education.Pctl
##  Min.   :  0.00              Min.   : 0.00   Min.   :  0.00
##  1st Qu.: 25.00              1st Qu.: 6.30   1st Qu.: 25.08
##  Median : 49.96              Median :14.00   Median : 50.00
##  Mean   : 49.98              Mean   :19.12   Mean   : 50.05
##  3rd Qu.: 75.07              3rd Qu.:28.70   3rd Qu.: 74.99
##  Max.   :100.00              Max.   :80.00   Max.   :100.00
##                              NA's   :96      NA's   :96
##  Linguistic.Isolation Linguistic.Isolation.Pctl    Poverty
##  Min.   : 0.00        Min.   :  0.00            Min.   : 0.00
##  1st Qu.: 3.00        1st Qu.: 22.52            1st Qu.:19.20
##  Median : 7.40        Median : 48.34            Median :33.50
##  Mean   :10.42        Mean   : 48.36            Mean   :36.39
##  3rd Qu.:14.90        3rd Qu.: 74.23            3rd Qu.:51.50
##  Max.   :72.30        Max.   :100.00            Max.   :96.20
##  NA's   :242          NA's   :242               NA's   :79
##   Poverty.Pctl     Unemployment    Unemployment.Pctl Housing.Burden
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00    Min.   : 1.00
##  1st Qu.: 25.10   1st Qu.:  6.60   1st Qu.: 25.46    1st Qu.:12.90
##  Median : 50.11   Median :  9.30   Median : 50.27    Median :18.00
##  Mean   : 50.07   Mean   : 10.21   Mean   : 50.32    Mean   :19.33
##  3rd Qu.: 75.02   3rd Qu.: 12.90   3rd Qu.: 75.52    3rd Qu.:24.40
##  Max.   :100.00   Max.   :100.00   Max.   :100.00    Max.   :67.20
##  NA's   :79       NA's   :155      NA's   :155       NA's   :157
##  Housing.Burden.Pctl   Pop..Char.    Pop..Char..Score Pop..Char..Pctl
##  Min.   :  0.03      Min.   : 2.53   Min.   : 0.260   Min.   :  0.01
##  1st Qu.: 25.51      1st Qu.:33.76   1st Qu.: 3.500   1st Qu.: 25.01
##  Median : 50.33      Median :49.96   Median : 5.180   Median : 50.01
##  Mean   : 50.18      Mean   :49.89   Mean   : 5.174   Mean   : 50.01
##  3rd Qu.: 75.01      3rd Qu.:66.45   3rd Qu.: 6.890   3rd Qu.: 75.00
##  Max.   :100.00      Max.   :96.43   Max.   :10.000   Max.   :100.00
##  NA's   :157         NA's   :106     NA's   :106      NA's   :106
```

2. Cleaning

```r
# mengahapus nilai na
Enviro <- na.omit(Enviro)
# menghapus nilai duplikasi
Enviro <- Enviro %>% distinct()
```

3. Pre Processing

Data Kualitas Udara

```
# pilih kolom
air_quality_required_columns <- c("Ozone",
                      "PM2.5",
                      "Diesel.PM",
                      "Ozone.Pctl",
                      "PM2.5.Pctl",
                      "Diesel.PM.Pctl")
# data kulitas udara
air_quality_data <- Enviro %>%
  select(all_of(air_quality_required_columns))
```

Data Pravelansi Penyakit

```
# pilih kolom
disease_prevalence_required_columns <- c("Asthma",
                                         "Cardiovascular.Disease",
                                         "Asthma.Pctl",
                                         "Cardiovascular.Disease.Pctl",
                                         "Pollution.Burden")
# data penyakit
disease_prevalence_data <- Enviro %>%
  select(all_of(disease_prevalence_required_columns))
```

Ekspor Data

```
# #lokasi
# file_output_air <- "dataset/air_quality_data.xlsx"
# file_output_disease <- "dataset/disease_prevalence_data.xlsx"
#
# # Ekspor data kualitas udara ke file XLSX
# write_xlsx(air_quality_data, file_output_air)
#
# # Ekspor data penyakit ke file XLSX
# write_xlsx(disease_prevalence_data, file_output_disease)
```

5. DATA KULITAS UDARA

```
# import data
air_quality <- read_excel("dataset/air_quality_data.xlsx")
```

0. Import Library New

```
if (!require(dplyr)) install.packages("dplyr")
if (!require(ggplot2)) install.packages("ggplot2")
if (!require(caret)) install.packages("caret")
```

```
## Loading required package: caret
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
if (!require(cluster)) install.packages("cluster")
```

```
## Loading required package: cluster
```

```
if (!require(corrplot)) install.packages("corrplot")
```

```
## Loading required package: corrplot
```

```
## corrplot 0.95 loaded
```

```
library(dplyr)
library(ggplot2)
library(caret)
library(cluster)
library(corrplot)
```

1. Observasi Data

```
# tampilkan data
head(air_quality, 5)
```

```
## # A tibble: 5 x 6
##   Ozone PM2.5 Diesel.PM Ozone.Pctl PM2.5.Pctl Diesel.PM.Pctl
##   <dbl> <dbl>     <dbl>      <dbl>      <dbl>          <dbl>
## 1 0.065  15.4      48.5       98.2       97.2           95.5
## 2 0.062  13.3      38.6       91.1       93.6           92.1
## 3 0.062  15.4      47.4       91.1       97.2           95.4
## 4 0.046  12.5      24.1       53.0       84.0           73.5
## 5 0.065  15.4      18.8       98.2       97.2           58.2
```

```r
# struktur data
str(air_quality)
```

```
## tibble [7,557 x 6] (S3: tbl_df/tbl/data.frame)
##  $ Ozone         : num [1:7557] 0.065 0.062 0.062 0.046 0.065 0.046 0.046 0.065 0.046 0.065 ...
##  $ PM2.5         : num [1:7557] 15.4 13.3 15.4 12.5 15.4 ...
##  $ Diesel.PM     : num [1:7557] 48.5 38.6 47.4 24.1 18.8 ...
##  $ Ozone.Pctl    : num [1:7557] 98.2 91.1 91.1 53 98.2 ...
##  $ PM2.5.Pctl    : num [1:7557] 97.2 93.6 97.2 84 97.2 ...
##  $ Diesel.PM.Pctl: num [1:7557] 95.5 92.1 95.4 73.5 58.2 ...
```

```r
# descriptive data
summary(air_quality)
```

```
##      Ozone            PM2.5          Diesel.PM         Ozone.Pctl
##  Min.   :0.0260   Min.   : 1.869   Min.   :  0.021   Min.   :  0.24
##  1st Qu.:0.0400   1st Qu.: 8.698   1st Qu.:  9.267   1st Qu.: 25.87
##  Median :0.0460   Median :10.370   Median : 16.811   Median : 53.02
##  Mean   :0.0474   Mean   :10.436   Mean   : 19.440   Mean   : 53.27
##  3rd Qu.:0.0550   3rd Qu.:12.050   3rd Qu.: 24.760   3rd Qu.: 77.87
##  Max.   :0.0680   Max.   :19.600   Max.   :208.400   Max.   :100.00
##    PM2.5.Pctl     Diesel.PM.Pctl
##  Min.   :  0.02   Min.   : 0.01
##  1st Qu.: 30.70   1st Qu.:26.60
##  Median : 52.61   Median :51.20
##  Mean   : 54.31   Mean   :50.92
##  3rd Qu.: 81.66   3rd Qu.:75.51
##  Max.   :100.00   Max.   :99.99
```

2. Data Cleaning

```r
# Handle missing values
air_quality <- air_quality %>%
  mutate(across(everything(), ~ ifelse(is.na(.), mean(., na.rm = TRUE), .)))

# Remove duplicates
air_quality <- air_quality %>%
  distinct()

# Detect and remove outliers using Z-scores
z_scores <- scale(air_quality %>% select_if(is.numeric))
air_quality <- air_quality[apply(abs(z_scores) < 3, 1, all), ]
```

3. Data Pre Processing

```r
# Mengganti nama kolom
colnames(air_quality) <- c("Ozone_Concentration", "PM2.5_Concentration", "DieselPM_Concentration", "Ozo
```

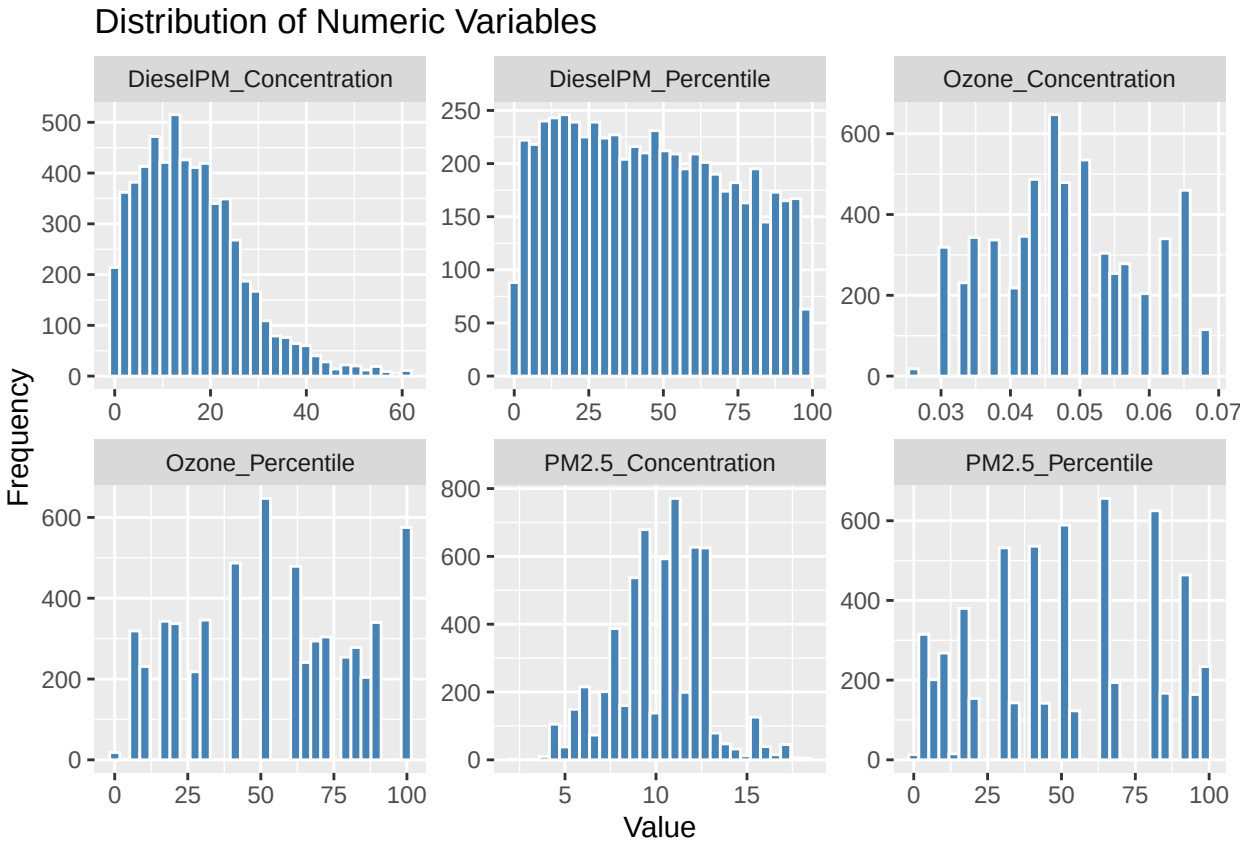
```r
# Ekspor Data Airquality baru
# #lokasi
# file_output_airquality <- "dataset/air_quality.xlsx"
#
# # Ekspor data kualitas udara ke file XLSX baru
# write_xlsx(air_quality, file_output_airquality)
```

Eksplorasi Data Analisis (EDA) Airquality

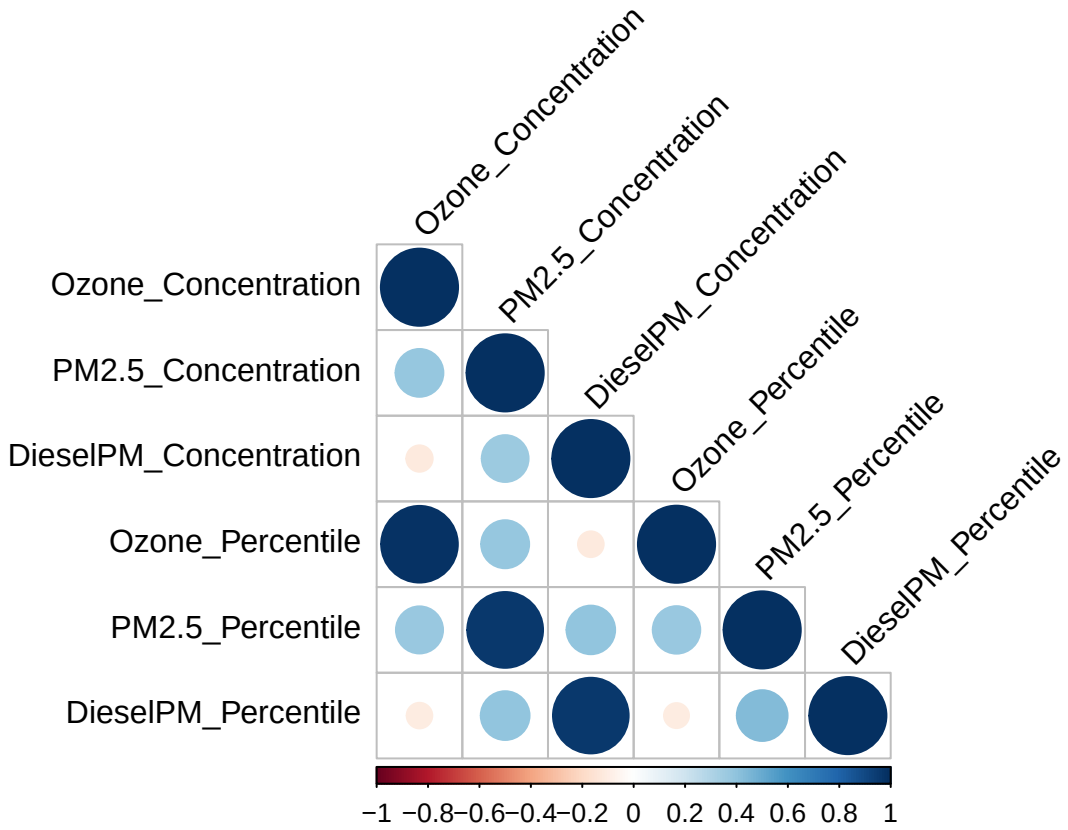
```r
# tampilkan data
head(air_quality, 5)
```

```
## # A tibble: 5 x 6
##   Ozone_Concentration PM2.5_Concentration DieselPM_Concentration
##                 <dbl>               <dbl>                  <dbl>
## 1               0.065                15.4                   48.5
## 2               0.062                13.3                   38.6
## 3               0.062                15.4                   47.4
## 4               0.046                12.5                   24.1
## 5               0.065                15.4                   18.8
## # i 3 more variables: Ozone_Percentile <dbl>, PM2.5_Percentile <dbl>,
## #   DieselPM_Percentile <dbl>
```

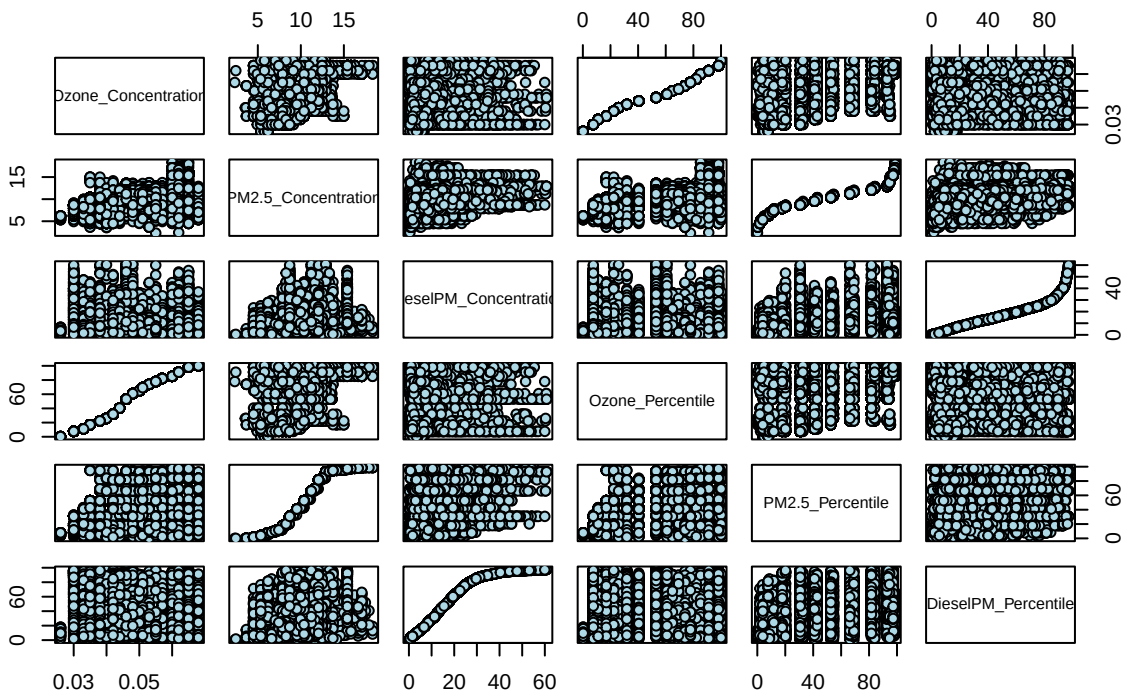
```r
# Visualize distributions of numeric columns
numeric_cols <- air_quality %>% select_if(is.numeric)
numeric_cols_long <- numeric_cols %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Value")

ggplot(numeric_cols_long, aes(x = Value)) +
  geom_histogram(bins = 30, fill = "steelblue", color = "white") +
  facet_wrap(~Variable, scales = "free") +
  labs(title = "Distribution of Numeric Variables", x = "Value", y = "Frequency")
```



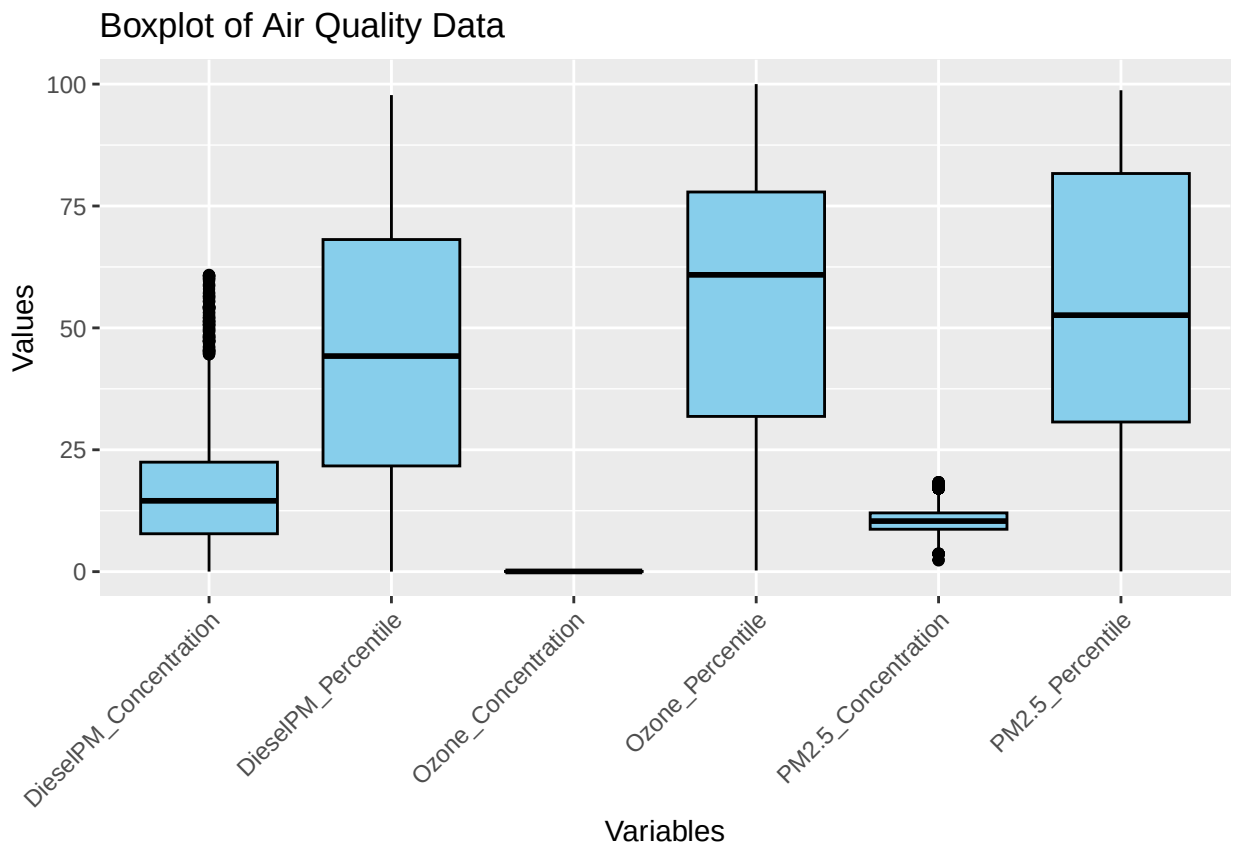

```
# Correlation matrix
cor_matrix <- cor(air_quality %>% select_if(is.numeric))
corrplot(cor_matrix, method = "circle", type = "lower", tl.col = "black", tl.srt = 45)
```




```r
# Pair plot for relationships
pairs(air_quality %>% select_if(is.numeric),
      main = "Pair Plot of Numeric Variables",
      pch = 21, bg = "lightblue")
```




```
# Box plot
ggplot(numeric_cols_long, aes(x = Variable, y = Value)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Boxplot of Air Quality Data", x = "Variables", y = "Values")
```



4. Data Mining

```
# Encoding and scaling
airquality_scaled <- air_quality %>%
  mutate(across(where(is.numeric), scale))
```

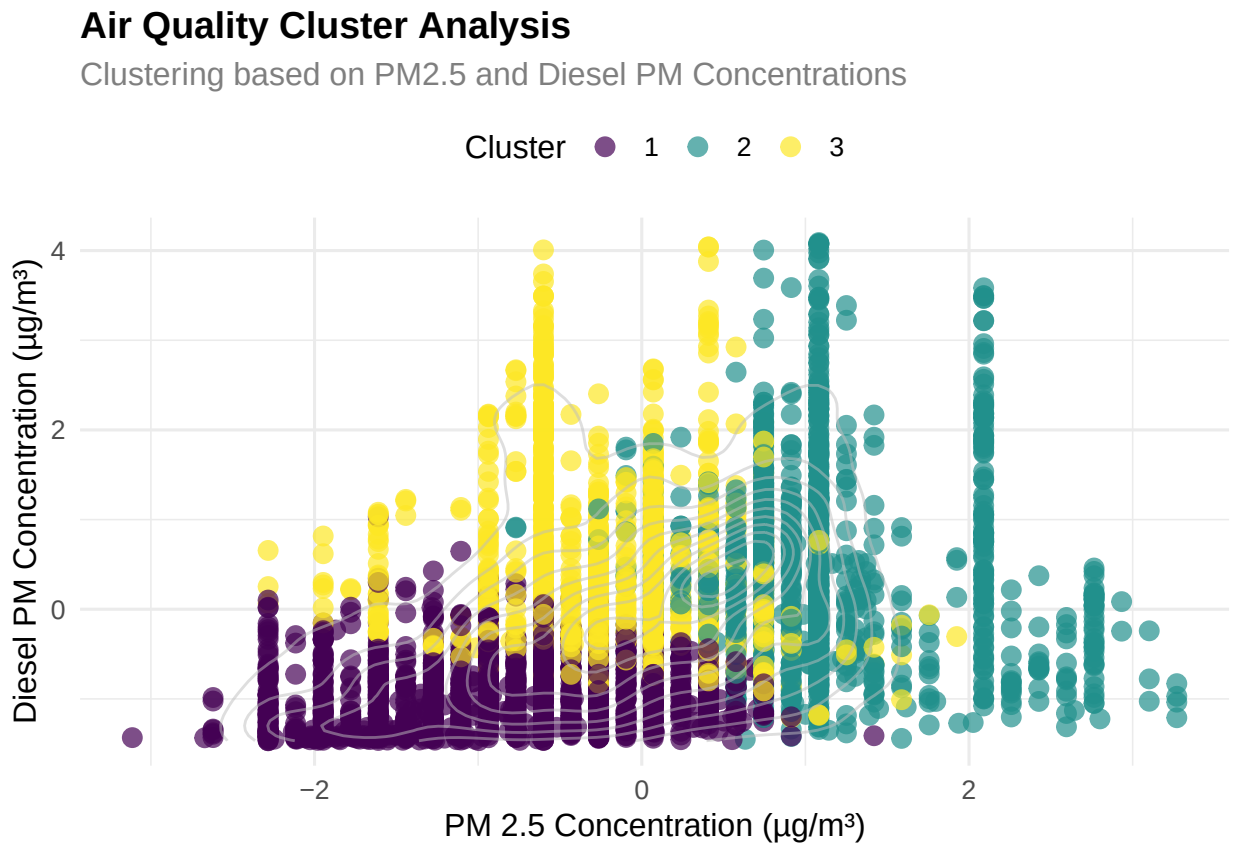
```
set.seed(125)
airquality_kmeans_result <- kmeans(airquality_scaled, centers = 3)
airquality_scaled$Cluster <- as.factor(airquality_kmeans_result$cluster)
```

```
# Train-test split
set.seed(125)
airquality_train_index <- createDataPartition(airquality_scaled$Ozone_Concentration, p = 0.8, list = FA
airquality_train <- airquality_scaled[airquality_train_index, ]
airquality_test <- airquality_scaled[-airquality_train_index, ]
```

Analysis and Visualization

```
# Visualize clusters
airquality_cluster_plot <- ggplot(airquality_train, aes(x = PM2.5_Concentration, y = DieselPM_Concentra
  geom_point(size = 3, alpha = 0.7) +
```

```
  scale_color_viridis_d() +
  geom_density2d(alpha = 0.5, color = "gray") +
  labs(
    title = "Air Quality Cluster Analysis",
    subtitle = "Clustering based on PM2.5 and Diesel PM Concentrations",
    x = "PM 2.5 Concentration (µg/m³)",
    y = "Diesel PM Concentration (µg/m³)"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    legend.position = "top",
    plot.title = element_text(face = "bold"),
    plot.subtitle = element_text(color = "gray50")
  )

print(airquality_cluster_plot)
```



Regression (Linear Regression)

```
airquality_model <- lm(Ozone_Concentration ~ PM2.5_Concentration + DieselPM_Concentration, data = airqua
summary(airquality_model)
```

```
##
```

```
## Call:
## lm(formula = Ozone_Concentration ~ PM2.5_Concentration + DieselPM_Concentration,
##     data = airquality_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3786 -0.6216 -0.1115  0.5088  2.6770
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -0.002459   0.012712  -0.193    0.847
## PM2.5_Concentration     0.501657   0.013602  36.881   <2e-16 ***
## DieselPM_Concentration -0.305520   0.013623 -22.426   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8746 on 4731 degrees of freedom
## Multiple R-squared:  0.2355, Adjusted R-squared:  0.2351
## F-statistic: 728.6 on 2 and 4731 DF,  p-value: < 2.2e-16
```

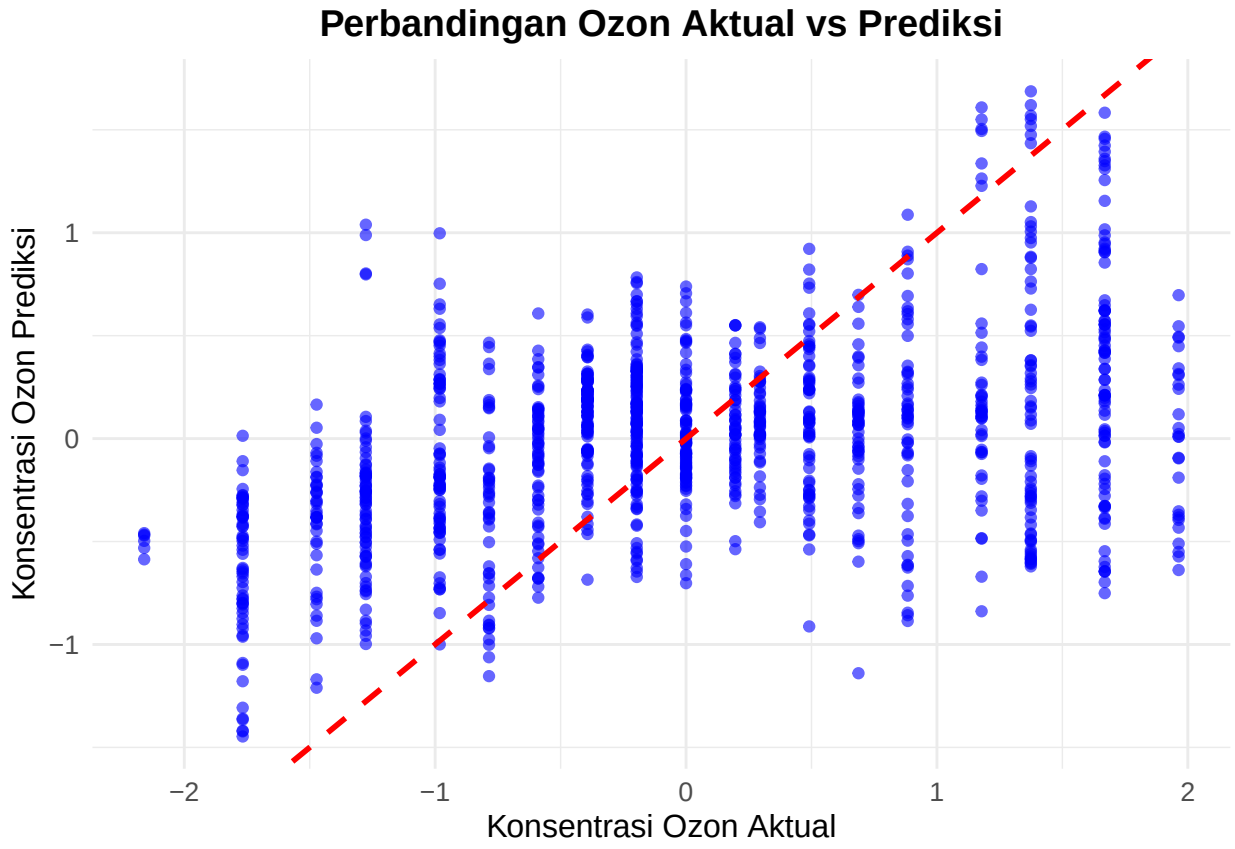
```
# Predict on test data
airquality_predictions <- predict(airquality_model, airquality_test)
airquality_test$Predicted_Ozone <- airquality_predictions

airquality_test %>%
  select(Ozone_Concentration, Predicted_Ozone) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##    Ozone_Concentration[,1] Predicted_Ozone
##                      <dbl>           <dbl>
##  1                  -0.196         -0.588
##  2                   1.67          -0.0186
##  3                   1.37           1.13
##  4                   1.67           0.0143
##  5                   1.67           1.15
##  6                   0.491          0.0968
##  7                  -0.196         -0.672
##  8                   1.67           0.286
##  9                  -0.589          0.0600
## 10                   1.67           0.555
```

```
# Visualize actual vs predicted
ggplot(airquality_test, aes(x = Ozone_Concentration, y = Predicted_Ozone)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_abline(color = "red", linetype = "dashed", size = 1) +
  labs(
    title = "Perbandingan Ozon Aktual vs Prediksi",
    x = "Konsentrasi Ozon Aktual",
    y = "Konsentrasi Ozon Prediksi"
  ) +
  theme_minimal() +
  theme(
```

```
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
    axis.title = element_text(size = 12),
    axis.text = element_text(size = 10)
  )
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```



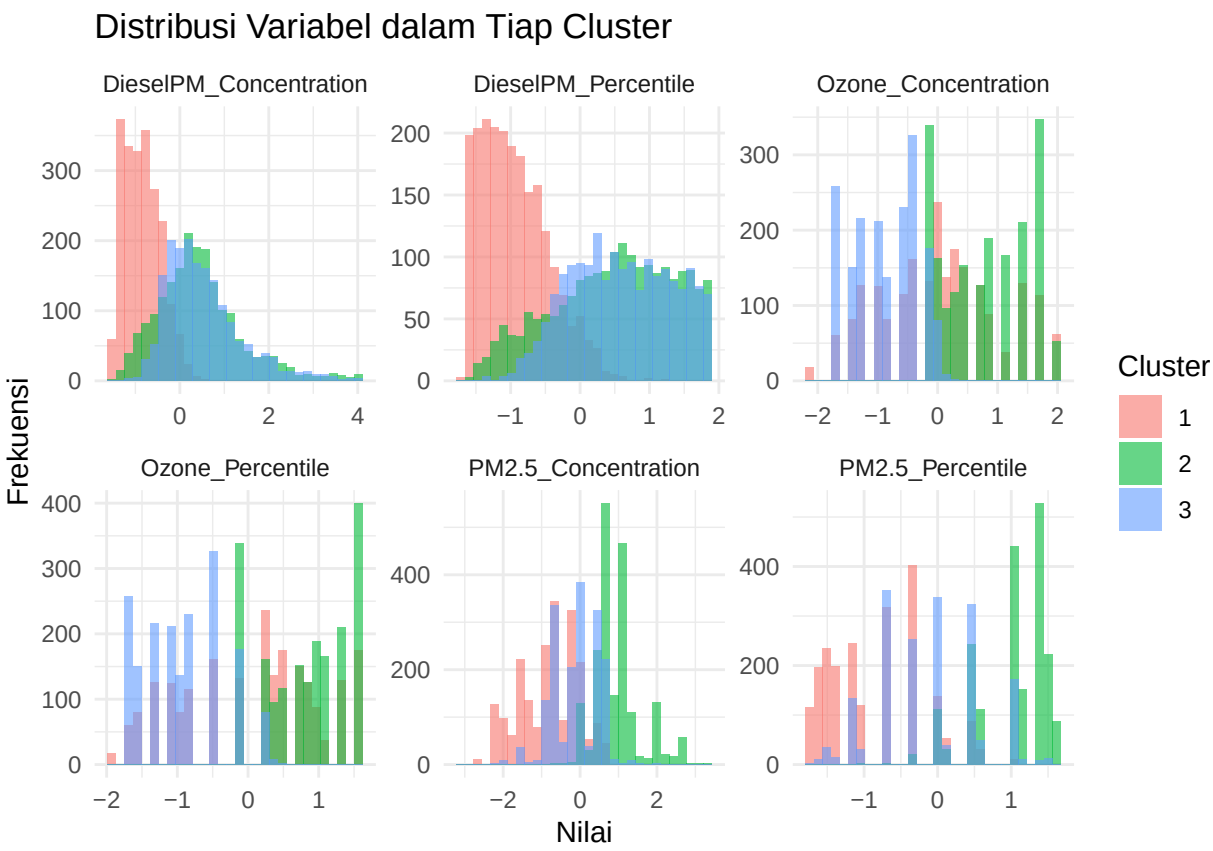
Clustering (K-Means)

```
cluster_summary <- airquality_scaled %>%
  group_by(Cluster) %>%
  arrange(Cluster)

head(cluster_summary, 5)
```

```
## # A tibble: 5 x 7
## # Groups:   Cluster [1]
##   Ozone_Concentration[,1] PM2.5_Concentration[,1] DieselPM_Concentration[,1]
```

```
##                    <dbl>                    <dbl>                    <dbl>
## 1               0.687                   -0.937                   -0.288
## 2               0.196                   -0.264                   -0.184
## 3              -0.196                    0.709                   -1.11
## 4               0.0000332               -0.264                   -0.193
## 5               1.18                    -1.27                     0.0958
## # i 4 more variables: Ozone_Percentile <dbl[,1]>, PM2.5_Percentile <dbl[,1]>,
## #   DieselPM_Percentile <dbl[,1]>, Cluster <fct>
```

```r
# Visualisasi distribusi variabel dalam cluster
numeric_cols_clustered <- airquality_scaled %>%
  pivot_longer(cols = where(is.numeric), names_to = "Variable", values_to = "Value")

ggplot(numeric_cols_clustered, aes(x = Value, fill = Cluster)) +
  geom_histogram(position = "identity", alpha = 0.6, bins = 30) +
  facet_wrap(~Variable, scales = "free") +
  labs(
    title = "Distribusi Variabel dalam Tiap Cluster",
    x = "Nilai",
    y = "Frekuensi"
  ) +
  theme_minimal()
```




```r
# Deskripsi cluster
cluster_descriptions <- cluster_summary %>%
```

```r
  mutate(Description = case_when(
    Ozone_Concentration > 50 & PM2.5_Concentration > 50 ~ "Polusi Tinggi",
    Ozone_Concentration < 30 & PM2.5_Concentration < 30 ~ "Polusi Rendah",
    TRUE ~ "Polusi Sedang"
  ))

cluster_descriptions %>%
  select(Ozone_Concentration, PM2.5_Concentration, Description) %>%
  head(10)
```

```
## Adding missing grouping variables: `Cluster`
```

```
## # A tibble: 10 x 4
## # Groups:   Cluster [1]
##    Cluster Ozone_Concentration[,1] PM2.5_Concentration[,1] Description
##    <fct>                     <dbl>                   <dbl> <chr>
##  1 1                       0.687                   -0.937 Polusi Rendah
##  2 1                       0.196                   -0.264 Polusi Rendah
##  3 1                      -0.196                    0.709 Polusi Rendah
##  4 1                       0.0000332               -0.264 Polusi Rendah
##  5 1                       1.18                    -1.27  Polusi Rendah
##  6 1                       1.67                    -1.61  Polusi Rendah
##  7 1                       0.0000332               -0.264 Polusi Rendah
##  8 1                       0.0000332               -0.264 Polusi Rendah
##  9 1                       0.295                   -0.264 Polusi Rendah
## 10 1                       0.0000332               -0.264 Polusi Rendah
```