ST306\_Mini\_Project

S/18/843

2024-01-30

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.3 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.3 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(janitor)

##   
## Attaching package: 'janitor'  
##   
## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

library(skimr)  
library(latex2exp)

local\_data <- read\_csv("../Data/london\_local\_data\_2022.csv")

## Rows: 289069 Columns: 10  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (2): site, code  
## dbl (7): nox, no2, no, pm10, o3, pm2\_5, so2  
## dttm (1): date  
##   
## 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.

view(local\_data)  
local\_sites <- read\_csv("../Data/london\_local\_sites.csv")

## Rows: 67 Columns: 5  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (3): code, site, Parameter\_name  
## dbl (2): latitude, longitude  
##   
## 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.

view(local\_sites)

glimpse(local\_data)

## Rows: 289,069  
## Columns: 10  
## $ site <chr> "Brent - John Keble Primary School", "Brent - John Keble Primary…  
## $ code <chr> "BT6", "BT6", "BT6", "BT6", "BT6", "BT6", "BT6", "BT6", "BT6", "…  
## $ date <dttm> 2022-01-01 00:00:00, 2022-01-01 01:00:00, 2022-01-01 02:00:00, …  
## $ nox <dbl> 13.4, 16.0, 11.1, 7.8, 8.6, 10.1, 16.5, 12.8, 14.6, 16.7, 17.5, …  
## $ no2 <dbl> 10.1, 11.3, 7.0, 5.3, 5.7, 6.9, 12.2, 9.2, 10.2, 11.1, 11.0, 9.1…  
## $ no <dbl> 2.2, 3.0, 2.6, 1.7, 1.9, 2.1, 2.8, 2.4, 2.9, 3.7, 4.2, 4.2, 4.6,…  
## $ pm10 <dbl> 29.9, 17.5, 16.0, 16.5, 14.8, 11.3, 12.9, 11.9, 9.6, 11.6, 14.4,…  
## $ o3 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …  
## $ pm2\_5 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …  
## $ so2 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …

glimpse(local\_sites)

## Rows: 67  
## Columns: 5  
## $ code <chr> "BT8", "BT8", "BT6", "BT6", "CT4", "CT4", "CT3", "CT3",…  
## $ site <chr> "Brent - ARK Franklin Primary Academy", "Brent - ARK Fr…  
## $ latitude <dbl> 51.53240, 51.53240, 51.53780, 51.53780, 51.52023, 51.52…  
## $ longitude <dbl> -0.217719, -0.217719, -0.247793, -0.247793, -0.096106, …  
## $ Parameter\_name <chr> "Nitrogen dioxide", "PM10 particulate matter (Hourly me…

Data Cleaning Part

#remove empty rows and columns  
local\_data <- local\_data %>%  
 remove\_empty(c("cols","rows"))  
view(local\_data)

#replace any missing values in the "local\_data" dataset with the value 0.  
local\_data[is.na(local\_data)] <- 0  
view(local\_data)

skim(local\_data)

Data summary

|  |  |
| --- | --- |
| Name | local\_data |
| Number of rows | 289069 |
| Number of columns | 10 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 2 |
| numeric | 7 |
| POSIXct | 1 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| site | 0 | 1 | 17 | 40 | 0 | 34 | 0 |
| code | 0 | 1 | 3 | 3 | 0 | 34 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| nox | 0 | 1 | 52.63 | 67.70 | -4.5 | 13.3 | 32.8 | 66.2 | 1111.1 | ▇▁▁▁▁ |
| no2 | 0 | 1 | 27.78 | 23.94 | -7.2 | 10.5 | 23.1 | 40.0 | 285.1 | ▇▂▁▁▁ |
| no | 0 | 1 | 16.20 | 31.48 | -3.3 | 0.9 | 5.4 | 17.3 | 631.5 | ▇▁▁▁▁ |
| pm10 | 0 | 1 | 11.28 | 14.01 | -6.5 | 0.0 | 8.4 | 18.0 | 300.8 | ▇▁▁▁▁ |
| o3 | 0 | 1 | 3.32 | 14.14 | -2.0 | 0.0 | 0.0 | 0.0 | 189.6 | ▇▁▁▁▁ |
| pm2\_5 | 0 | 1 | 0.61 | 3.46 | -3.0 | 0.0 | 0.0 | 0.0 | 685.0 | ▇▁▁▁▁ |
| so2 | 0 | 1 | 0.09 | 0.60 | -4.9 | 0.0 | 0.0 | 0.0 | 11.1 | ▁▇▁▁▁ |

**Variable type: POSIXct**

| skim\_variable | n\_missing | complete\_rate | min | max | median | n\_unique |
| --- | --- | --- | --- | --- | --- | --- |
| date | 0 | 1 | 2022-01-01 | 2022-12-31 23:00:00 | 2022-07-02 11:00:00 | 8760 |

glimpse(local\_data)

## Rows: 289,069  
## Columns: 10  
## $ site <chr> "Brent - John Keble Primary School", "Brent - John Keble Primary…  
## $ code <chr> "BT6", "BT6", "BT6", "BT6", "BT6", "BT6", "BT6", "BT6", "BT6", "…  
## $ date <dttm> 2022-01-01 00:00:00, 2022-01-01 01:00:00, 2022-01-01 02:00:00, …  
## $ nox <dbl> 13.4, 16.0, 11.1, 7.8, 8.6, 10.1, 16.5, 12.8, 14.6, 16.7, 17.5, …  
## $ no2 <dbl> 10.1, 11.3, 7.0, 5.3, 5.7, 6.9, 12.2, 9.2, 10.2, 11.1, 11.0, 9.1…  
## $ no <dbl> 2.2, 3.0, 2.6, 1.7, 1.9, 2.1, 2.8, 2.4, 2.9, 3.7, 4.2, 4.2, 4.6,…  
## $ pm10 <dbl> 29.9, 17.5, 16.0, 16.5, 14.8, 11.3, 12.9, 11.9, 9.6, 11.6, 14.4,…  
## $ o3 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…  
## $ pm2\_5 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…  
## $ so2 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…

local\_data <- local\_data %>%  
 separate(date, into = c("date","time"), sep = 10)  
view(local\_data)

local\_data <- local\_data %>%  
 mutate(month = factor(case\_when(  
 month(date)==1 ~ "january",  
 month(date)==2 ~ "february",  
 month(date)==3 ~ "march",  
 month(date)==4 ~ "april",  
 month(date)==5 ~ "may",  
 month(date)==6 ~ "june",  
 month(date)==7 ~ "july",  
 month(date)==8 ~ "august",  
 month(date)==9 ~ "september",  
 month(date)==10 ~ "october",  
 month(date)==11 ~ "november",  
 month(date)==12 ~ "december",  
 )), year = year(date))

local\_data$month <- factor(local\_data$month,   
 levels = c("january","february","march","april","may",  
 "june","july","august","september","october","november","december"))

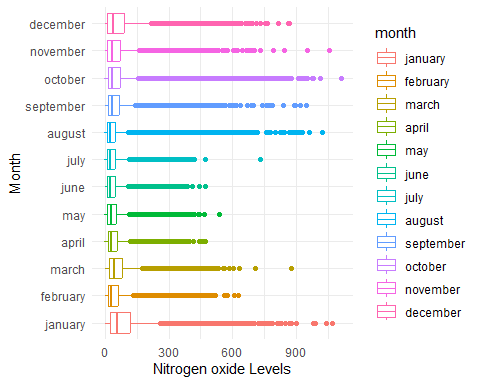
head(local\_data)

## # A tibble: 6 × 13  
## site code date time nox no2 no pm10 o3 pm2\_5 so2 month year  
## <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <dbl>  
## 1 Brent… BT6 2022… "" 13.4 10.1 2.2 29.9 0 0 0 janu… 2022  
## 2 Brent… BT6 2022… " 01… 16 11.3 3 17.5 0 0 0 janu… 2022  
## 3 Brent… BT6 2022… " 02… 11.1 7 2.6 16 0 0 0 janu… 2022  
## 4 Brent… BT6 2022… " 03… 7.8 5.3 1.7 16.5 0 0 0 janu… 2022  
## 5 Brent… BT6 2022… " 04… 8.6 5.7 1.9 14.8 0 0 0 janu… 2022  
## 6 Brent… BT6 2022… " 05… 10.1 6.9 2.1 11.3 0 0 0 janu… 2022

view(local\_data)

Analyzing Part

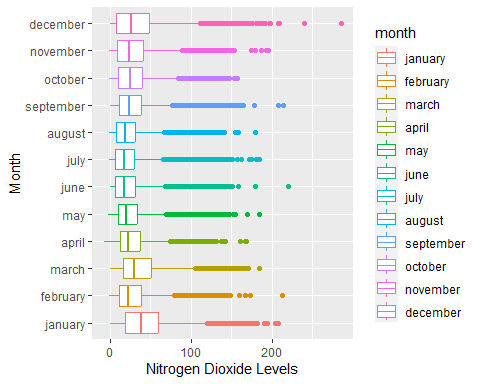
nox\_graph <- local\_data %>%  
 ggplot(aes(x=month, y=nox))+  
 geom\_boxplot(aes(col=month))+  
 coord\_flip()+  
 theme\_minimal()+  
 labs(x = "Month",  
 y = "Nitrogen oxide Levels")  
nox\_graph



postscript(file = "nox\_graph.eps", width = 10, height = 8, horizontal = F)  
nox\_graph  
dev.off()

## png   
## 2

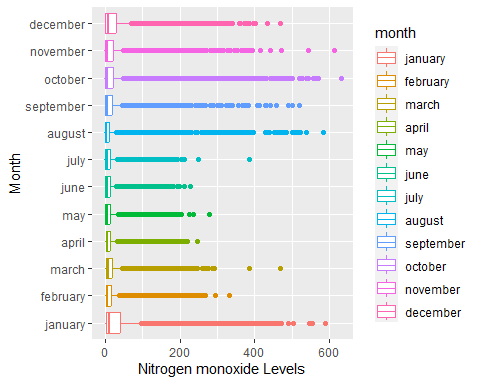
no2\_graph <- local\_data %>%  
 ggplot(aes(x=month, y=no2))+  
 geom\_boxplot(aes(col=month))+  
 coord\_flip()+  
 labs(x = "Month",  
 y = "Nitrogen Dioxide Levels")  
no2\_graph



postscript(file = "no2\_graph.eps", width = 10, height = 8, horizontal = F)  
no2\_graph  
dev.off()

## png   
## 2

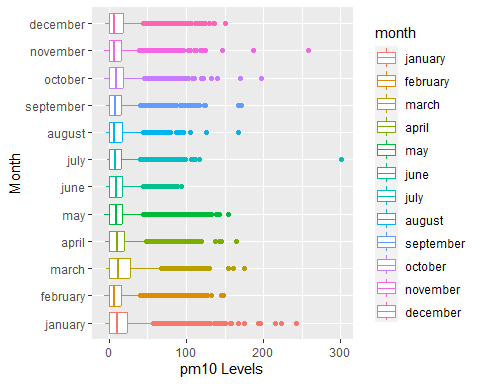
no\_graph <- local\_data %>%  
 ggplot(aes(x=month, y=no))+  
 geom\_boxplot(aes(col=month))+  
 coord\_flip()+  
 labs(x = "Month",  
 y = "Nitrogen monoxide Levels")  
no\_graph



postscript(file = "no\_graph.eps", width = 10, height = 8, horizontal = F)  
no\_graph  
dev.off()

## png   
## 2

pm10\_graph <- local\_data %>%  
 ggplot(aes(x=month, y=pm10))+  
 geom\_boxplot(aes(col=month))+  
 coord\_flip()+  
 labs(x = "Month",  
 y = "pm10 Levels")  
pm10\_graph

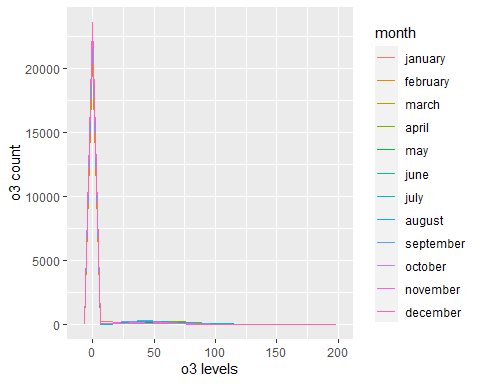


postscript(file = "pm10\_graph.eps", width = 10, height = 8, horizontal = F)  
pm10\_graph  
dev.off()

## png   
## 2

o3\_count\_graph <- local\_data %>%  
 ggplot(aes(x=o3))+  
 geom\_freqpoly(aes(col = month))+  
 labs(x = "o3 levels",  
 y = "o3 count")  
o3\_count\_graph

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



postscript(file = "o3\_count\_graph.eps", width = 10, height = 8, horizontal = F)  
o3\_count\_graph

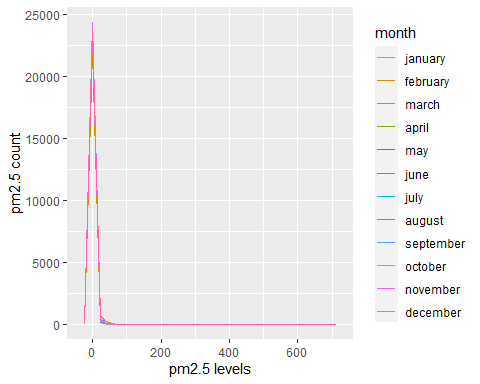
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

dev.off()

## png   
## 2

pm2\_5\_count\_graph <- local\_data %>%  
 ggplot(aes(x=pm2\_5))+  
 geom\_freqpoly(aes(col = month))+  
 labs(x = "pm2.5 levels",  
 y = "pm2.5 count")  
pm2\_5\_count\_graph

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



postscript(file = "pm2\_5\_count\_graph.eps", width = 10, height = 8, horizontal = F)  
pm2\_5\_count\_graph

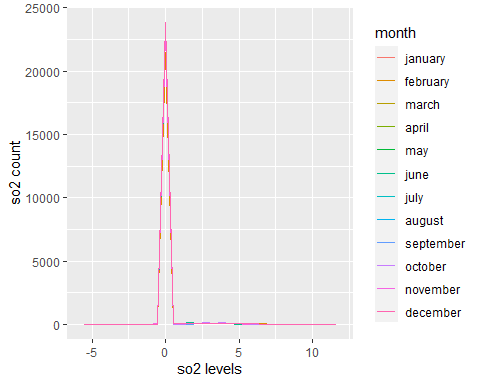
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

dev.off()

## png   
## 2

so2\_count\_graph <- local\_data %>%  
 ggplot(aes(x=so2))+  
 geom\_freqpoly(aes(col = month))+  
 labs(x = "so2 levels",  
 y = "so2 count")  
so2\_count\_graph

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



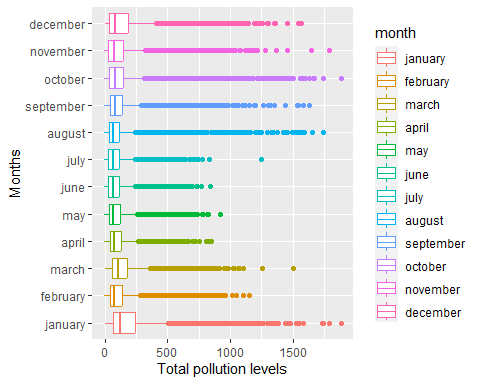
postscript(file = "so2\_count\_graph.eps", width = 10, height = 8, horizontal = F)  
so2\_count\_graph

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

dev.off()

## png   
## 2

Total\_pollution\_graph <- local\_data %>%  
 ggplot(aes(x=month, y = no+no2+nox+pm10+o3+pm2\_5+so2))+  
 geom\_boxplot(aes(col=month))+  
 coord\_flip()+  
 labs(x = "Months",  
 y = "Total pollution levels")  
Total\_pollution\_graph



postscript(file = "Total\_pollution\_graph.eps", width = 10, height = 8, horizontal = F)  
Total\_pollution\_graph  
dev.off()

## png   
## 2

# pollution for each month  
set1 <- local\_data %>%  
 group\_by(month) %>%  
 summarise(NO=sum(no),  
 NO2=sum(no2),  
 NOx=sum(nox),  
 pm10=sum(pm10),  
 O3=sum(o3),  
 pm2\_5=sum(pm2\_5),  
 SO2=sum(so2))  
set1

## # A tibble: 12 × 8  
## month NO NO2 NOx pm10 O3 pm2\_5 SO2  
## <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 january 833924. 1014988. 2293633. 359324. 67611. 25071. 2431.  
## 2 february 311137. 612346. 1089415. 228271. 84328. 10755 2342.  
## 3 march 394539. 869004. 1472031. 413958. 84647. 24516 2412.  
## 4 april 256342. 626266. 1020618. 273625 116097. 15558 2061.  
## 5 may 279166. 581898. 1011325. 260697. 78288. 13570 1968.  
## 6 june 245197. 525196. 902578. 247301. 82322. 7603 1727   
## 7 july 252477. 537435. 924698. 231689. 99910. 14055 1732.  
## 8 august 273364. 547810. 966785. 244677. 131431. 10208 2305.  
## 9 september 377110. 646310. 1224609. 234723. 93936. 13055 1925.  
## 10 october 461804. 664810. 1373086. 273280 46951. 8922 2464.  
## 11 november 401433. 645585. 1261436. 223239. 32976. 12900 1921.  
## 12 december 596718. 757815. 1672765. 268571. 40538. 21490 2790.

# pollution for each site  
set2 <- local\_data %>%  
 group\_by(site) %>%  
 summarise(NO=sum(no),  
 NO2=sum(no2),  
 NOx=sum(nox),  
 pm10=sum(pm10),  
 O3=sum(o3),  
 pm2\_5=sum(pm2\_5),  
 SO2=sum(so2))  
set2

## # A tibble: 34 × 8  
## site NO NO2 NOx pm10 O3 pm2\_5 SO2  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Brent - ARK Franklin Primary A… 1.24e5 2.49e5 4.39e5 1.34e5 0 0 0  
## 2 Brent - John Keble Primary Sch… 1.01e5 2.40e5 3.94e5 1.42e5 0 0 0  
## 3 City of London - Beech Street 1.42e5 3.49e5 5.67e5 1.39e5 0 0 0  
## 4 City of London - Sir John Cass… 4.37e4 1.98e5 2.64e5 1.19e5 0 0 0  
## 5 City of London - Upper Thames … 0 0 0 2.51e2 0 0 0  
## 6 City of London - Walbrook Wharf 3.44e5 4.42e5 9.69e5 0 0 0 0  
## 7 Ealing - Acton Vale 2.82e4 8.58e4 1.29e5 4.73e4 0 0 0  
## 8 Ealing - Hanger Lane Gyratory 5.01e5 4.49e5 1.22e6 1.47e5 0 0 0  
## 9 Ealing - Western Avenue 1.96e5 3.06e5 6.07e5 2.14e5 0 0 0  
## 10 Greenwich - Blackheath 1.06e5 2.25e5 3.87e5 1.42e5 0 0 0  
## # ℹ 24 more rows

set3 <- local\_data%>%  
 group\_by(month) %>%   
 summarise(Total\_pollutant=sum(no)+sum(no2)+sum(nox)+  
 sum(pm10)+sum(o3)+sum(pm2\_5)+sum(so2))  
  
set3 %>%  
 arrange(desc(Total\_pollutant))

## # A tibble: 12 × 2  
## month Total\_pollutant  
## <fct> <dbl>  
## 1 january 4596983.  
## 2 december 3360687.  
## 3 march 3261107.  
## 4 october 2831317.  
## 5 september 2591667.  
## 6 november 2579491.  
## 7 february 2338595.  
## 8 april 2310568.  
## 9 may 2226913.  
## 10 august 2176580.  
## 11 july 2061997.  
## 12 june 2011924.

top\_five\_pollution\_months <- set3 %>%  
 head(5)  
top\_five\_pollution\_months

## # A tibble: 5 × 2  
## month Total\_pollutant  
## <fct> <dbl>  
## 1 january 4596983.  
## 2 february 2338595.  
## 3 march 3261107.  
## 4 april 2310568.  
## 5 may 2226913.

set4 <- local\_data%>%  
 group\_by(site) %>%   
 summarise(Total\_pollutant=sum(no)+sum(no2)+sum(nox)+  
 sum(pm10)+sum(o3)+sum(pm2\_5)+sum(so2))  
  
set4 <- set4 %>%  
 arrange(desc(Total\_pollutant))  
set4

## # A tibble: 34 × 2  
## site Total\_pollutant  
## <chr> <dbl>  
## 1 Lambeth - Brixton Road 2355163.  
## 2 Ealing - Hanger Lane Gyratory 2313841.  
## 3 City of London - Walbrook Wharf 1755354.  
## 4 Lewisham - Loampit Vale 1711119.  
## 5 Ealing - Western Avenue 1322689.  
## 6 Westminster - Oxford Street East 1296355.  
## 7 Hackney - Old Street 1274901.  
## 8 City of London - Beech Street 1196247.  
## 9 Lambeth - Bondway Interchange 1163040.  
## 10 Westminster - Oxford Street 1154628.  
## # ℹ 24 more rows

top\_five\_pollution\_sites <- set4 %>%  
 head(5)  
top\_five\_pollution\_sites

## # A tibble: 5 × 2  
## site Total\_pollutant  
## <chr> <dbl>  
## 1 Lambeth - Brixton Road 2355163.  
## 2 Ealing - Hanger Lane Gyratory 2313841.  
## 3 City of London - Walbrook Wharf 1755354.  
## 4 Lewisham - Loampit Vale 1711119.  
## 5 Ealing - Western Avenue 1322689.