

Coursework 2 – Port Talbot Hourly Air Quality Analysis

Deepthi

2025-11-11

```
# Description:
#   This script performs a full statistical analysis on the Port Talbot
#   Hourly Air Quality dataset. It includes data preparation, visualisation,
#   statistical comparison and inference based on hourly pollutant values.
#
# Major Functional Components (Tasks):
#   1. Load dataset, rename variables, create datetime fields.
#   2. Calculate percentage of missing values (invalid data table).
#   3. Visualise a selected pollutant across specified time ranges (function).
#   4. Compare pollutant behaviour over time (function).
#   5. Analyse correlation between O3 and NO2.
#   6. Perform a two-sample t-test (O3 by wind speed group).
#   7. Run ANOVA for NO2 values across different years.
#
# Output:
#   - CSV table showing missing data percentages.
#   - Professional annotated plots saved as PNG files.
#   - Console output for statistical test results.
#
# Usage:
#   Run the script in RStudio. Ensure the dataset file path is set correctly.
#
# Last Updated:   <DD MMM YYYY>
#####

# Load libraries (install if not found)
if (!requireNamespace("pacman", quietly = TRUE)) install.packages("pacman")
pacman::p_load(readr, dplyr, lubridate, ggplot2, scales, ggpubr, tidyr, rstatix, car)

# Load dataset
file_name <- "/Users/mac/Desktop/ASP/Port_Talbot_Hourly_AQ.csv"
PortTalbotAQ <- read_csv(file_name, na = c("", "No data", "NA"))

## Rows: 87672 Columns: 24
## -- Column specification -----
## Delimiter: ","
## chr  (12): Date, Status, Status.1, Status.2, Status.3, Status.4, Status.5, S...
## dbl  (11): Ozone, Nitric oxide, Nitrogen dioxide, Nitrogen oxides as nitroge...
## time (1): Time
##
## 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.
```

```

cat("Dataset Loaded\n\n")

## Dataset Loaded
print(dim(PortTalbotAQ))

## [1] 87672    24

print(names(PortTalbotAQ))

## [1] "Date"
## [2] "Time"
## [3] "Ozone"
## [4] "Status"
## [5] "Nitric oxide"
## [6] "Status.1"
## [7] "Nitrogen dioxide"
## [8] "Status.2"
## [9] "Nitrogen oxides as nitrogen dioxide"
## [10] "Status.3"
## [11] "Sulphur dioxide"
## [12] "Status.4"
## [13] "Carbon monoxide"
## [14] "Status.5"
## [15] "PM10 particulate matter (Hourly measured)"
## [16] "Status.6"
## [17] "PM2.5 particulate matter (Hourly measured)"
## [18] "Status.7"
## [19] "Modelled Wind Direction"
## [20] "Status.8"
## [21] "Modelled Wind Speed"
## [22] "Status.9"
## [23] "Modelled Temperature"
## [24] "Status.10"

#####
# TASK 1: Rename columns and create datetime
#####

PortTalbotAQ <- PortTalbotAQ %>%
  rename(
    date_str = `Date`,
    time     = `Time`,
    O3       = `Ozone`,
    NO       = `Nitric oxide`,
    NO2      = `Nitrogen dioxide`,
    NOx      = `Nitrogen oxides as nitrogen dioxide`,
    SO2      = `Sulphur dioxide`,
    CO       = `Carbon monoxide`,
    PM10     = `PM10 particulate matter (Hourly measured)`,
    PM25     = `PM2.5 particulate matter (Hourly measured)`,
    wind_dir = `Modelled Wind Direction`,
    wind_speed = `Modelled Wind Speed`,
    temp     = `Modelled Temperature`
  ) %>%
  mutate(

```

```

date = dmy(date_str),
datetime = as.POSIXct(paste(date, time), format = "%Y-%m-%d %H:%M"),
year = year(datetime),
month = month(datetime),
day = day(datetime),
hour = hour(datetime)
)

cat("TASK 1 Completed: Dataframe after renaming and datetime conversion\n")

## TASK 1 Completed: Dataframe after renaming and datetime conversion
print(head(PortTalbotAQ, 10))

## # A tibble: 10 x 30
##   date_str   time      O3 Status      NO Status.1    NO2 Status.2    NOx Status.3
##   <chr>     <time> <dbl> <chr>     <dbl> <chr>     <dbl> <chr>     <dbl> <chr>
## 1 01-01-2020 01:00  32.4 V ugm-3  0.652 V ugm-3  12.3 V ugm-3  13.3 V ugm-3
## 2 01-01-2020 02:00  27.1 V ugm-3 NA      V ugm-3  NA      V ugm-3  NA      V ugm-3
## 3 01-01-2020 03:00  34.7 V ugm-3  0.461 V ugm-3  12.3 V ugm-3  13.0 V ugm-3
## 4 01-01-2020 04:00  39.2 V ugm-3  0.403 V ugm-3  7.24 V ugm-3  7.86 V ugm-3
## 5 01-01-2020 05:00  36.1 V ugm-3  0.202 V ugm-3  7.33 V ugm-3  7.64 V ugm-3
## 6 01-01-2020 06:00  33.4 V ugm-3  0.403 V ugm-3  8.89 V ugm-3  9.51 V ugm-3
## 7 01-01-2020 07:00  33.1 V ugm-3  0.374 V ugm-3  10.1 V ugm-3  10.7 V ugm-3
## 8 01-01-2020 08:00  34.1 V ugm-3  0.403 V ugm-3  9.15 V ugm-3  9.77 V ugm-3
## 9 01-01-2020 09:00  28.2 V ugm-3  1.70 V ugm-3  12.7 V ugm-3  15.3 V ugm-3
## 10 01-01-2020 10:00  28.7 V ugm-3  2.59 V ugm-3  13.0 V ugm-3  16.9 V ugm-3
## # i 20 more variables: SO2 <dbl>, Status.4 <chr>, CO <dbl>, Status.5 <chr>,
## #   PM10 <dbl>, Status.6 <chr>, PM25 <dbl>, Status.7 <chr>, wind_dir <dbl>,
## #   Status.8 <chr>, wind_speed <dbl>, Status.9 <chr>, temp <dbl>,
## #   Status.10 <chr>, date <date>, datetime <dtm>, year <dbl>, month <dbl>,
## #   day <int>, hour <int>
print(summary(PortTalbotAQ))

##   date_str      time      O3      Status
## Length:87672 Length:87672 Min.   : -0.1497 Length:87672
## Class :character Class1:hms  1st Qu.: 40.1136 Class :character
## Mode  :character Class2:difftime Median : 54.4826 Mode  :character
##                Mode  :numeric Mean   : 53.0298
##                3rd Qu.: 66.6564
##                Max.   :164.8448
##                NA's   :1740
##      NO      Status.1      NO2      Status.2
## Min.   : -0.059 Length:87672 Min.   : 0.065 Length:87672
## 1st Qu.: 0.864 Class :character 1st Qu.: 6.680 Class :character
## Median : 2.526 Mode  :character Median : 11.058 Mode  :character
## Mean   : 4.371 Mean   : 13.945
## 3rd Qu.: 5.474 3rd Qu.: 17.797
## Max.   :169.557 Max.   :150.520
## NA's   :3214 NA's   :3489
##      NOx      Status.3      SO2      Status.4
## Min.   : 0.261 Length:87672 Min.   : -0.1329 Length:87672
## 1st Qu.: 9.092 Class :character 1st Qu.: 0.4612 Class :character
## Median : 15.585 Mode  :character Median : 0.6745 Mode  :character

```

```
## Mean      : 20.604          Mean      : 2.9400
## 3rd Qu.: 25.553          3rd Qu.: 1.4561
## Max.      :317.217        Max.      :253.8499
## NA's      :3489          NA's      :2100
##          CO              Status.5          PM10              Status.6
## Min.      : -0.012      Length:87672      Min.      : -3.865      Length:87672
## 1st Qu.: 0.051      Class :character      1st Qu.: 10.400      Class :character
## Median : 0.082      Mode  :character      Median : 16.600      Mode  :character
## Mean      : 0.177          Mean      : 23.479
## 3rd Qu.: 0.147          3rd Qu.: 28.020
## Max.      : 9.459          Max.      :574.889
## NA's      :4154          NA's      :5732
##          PM25              Status.7          wind_dir              Status.8
## Min.      : -5.000      Length:87672      Min.      : 0.0      Length:87672
## 1st Qu.: 4.900      Class :character      1st Qu.: 99.5      Class :character
## Median : 7.300      Mode  :character      Median :218.1      Mode  :character
## Mean      : 9.328          Mean      :192.6
## 3rd Qu.: 12.000          3rd Qu.:275.8
## Max.      :208.000          Max.      :360.0
## NA's      :6016          NA's      :1800
##          wind_speed          Status.9          temp              Status.10
## Min.      : 0.000      Length:87672      Min.      : -6.200      Length:87672
## 1st Qu.: 3.300      Class :character      1st Qu.: 6.300      Class :character
## Median : 5.000      Mode  :character      Median : 9.500      Mode  :character
## Mean      : 5.458          Mean      : 9.583
## 3rd Qu.: 7.100          3rd Qu.:13.100
## Max.      :20.800          Max.      :28.500
## NA's      :1800          NA's      :1800
##          date              datetime              year
## Min.      :2015-01-01      Min.      :2015-01-01 01:00:00      Min.      :2015
## 1st Qu.:2017-07-02      1st Qu.:2017-07-02 06:45:00      1st Qu.:2017
## Median :2020-01-01      Median :2020-01-01 12:30:00      Median :2020
## Mean      :2020-01-01      Mean      :2020-01-01 12:30:00      Mean      :2020
## 3rd Qu.:2022-07-02      3rd Qu.:2022-07-02 18:15:00      3rd Qu.:2022
## Max.      :2024-12-31      Max.      :2025-01-01 00:00:00      Max.      :2025
##
##          month              day              hour
## Min.      : 1.000      Min.      : 1.00      Min.      : 0.00
## 1st Qu.: 4.000      1st Qu.: 8.00      1st Qu.: 5.75
## Median : 7.000      Median :16.00      Median :11.50
## Mean      : 6.522      Mean      :15.73      Mean      :11.50
## 3rd Qu.:10.000      3rd Qu.:23.00      3rd Qu.:17.25
## Max.      :12.000      Max.      :31.00      Max.      :23.00
##
```

```
str(PortTalbotAQ)
```

```
## tibble [87,672 x 30] (S3: tbl_df/tbl/data.frame)
## $ date_str   : chr [1:87672] "01-01-2020" "01-01-2020" "01-01-2020" "01-01-2020" ...
## $ time       : 'hms' num [1:87672] 01:00:00 02:00:00 03:00:00 04:00:00 ...
##   .. attr(*, "units")= chr "secs"
## $ O3         : num [1:87672] 32.4 27.1 34.7 39.2 36.1 ...
## $ Status     : chr [1:87672] "V ugm-3" "V ugm-3" "V ugm-3" "V ugm-3" ...
## $ NO         : num [1:87672] 0.652 NA 0.461 0.403 0.202 ...
## $ Status.1   : chr [1:87672] "V ugm-3" "V ugm-3" "V ugm-3" "V ugm-3" ...
```

```
## $ NO2      : num [1:87672] 12.27 NA 12.31 7.24 7.33 ...
## $ Status.2 : chr [1:87672] "V ugm-3" "V ugm-3" "V ugm-3" "V ugm-3" ...
## $ NOx      : num [1:87672] 13.27 NA 13.01 7.86 7.64 ...
## $ Status.3 : chr [1:87672] "V ugm-3" "V ugm-3" "V ugm-3" "V ugm-3" ...
## $ SO2      : num [1:87672] 0.442 0.707 0.928 0.663 0.398 ...
## $ Status.4 : chr [1:87672] "V ugm-3" "V ugm-3" "V ugm-3" "V ugm-3" ...
## $ CO       : num [1:87672] 0.071 0.0867 0.1065 0.0739 0.0651 ...
## $ Status.5 : chr [1:87672] "V mgm-3" "V mgm-3" "V mgm-3" "V mgm-3" ...
## $ PM10     : num [1:87672] 25.1 30.6 27.6 23.6 22.9 20.9 21.2 20.7 23.9 23 ...
## $ Status.6 : chr [1:87672] "V ugm-3 (TEOM FDMS)" "V ugm-3 (TEOM FDMS)" "V ugm-3 (TEOM FDMS)" "V ugm-3 (TEOM FDMS)" ...
## $ PM25     : num [1:87672] 21.1 25 25.9 19.8 20.2 15.9 17.5 18.7 20.8 22.2 ...
## $ Status.7 : chr [1:87672] "V ugm-3 (TEOM FDMS)" "V ugm-3 (TEOM FDMS)" "V ugm-3 (TEOM FDMS)" "V ugm-3 (TEOM FDMS)" ...
## $ wind_dir : num [1:87672] 96.7 101.2 98.9 103.5 97.4 ...
## $ Status.8 : chr [1:87672] "N deg" "N deg" "N deg" "N deg" ...
## $ wind_speed : num [1:87672] 4.6 5.3 5.1 4.8 4.4 4.3 4.8 4.6 4.7 4.6 ...
## $ Status.9 : chr [1:87672] "N ms-1" "N ms-1" "N ms-1" "N ms-1" ...
## $ temp      : num [1:87672] 3 3.1 3.1 3.2 3.4 3.6 3.8 3.9 4 4.1 ...
## $ Status.10 : chr [1:87672] "N degC" "N degC" "N degC" "N degC" ...
## $ date      : Date[1:87672], format: "2020-01-01" "2020-01-01" ...
## $ datetime  : POSIXct[1:87672], format: "2020-01-01 01:00:00" "2020-01-01 02:00:00" ...
## $ year      : num [1:87672] 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...
## $ month     : num [1:87672] 1 1 1 1 1 1 1 1 1 1 ...
## $ day       : int [1:87672] 1 1 1 1 1 1 1 1 1 1 ...
## $ hour      : int [1:87672] 1 2 3 4 5 6 7 8 9 10 ...
```

```
#####
# TASK 2: Invalid data percentage table
#####
```

```
pollutants <- c("O3", "NO", "NO2", "NOx", "SO2", "CO", "PM10", "PM25", "wind_speed", "wind_dir", "temp")
```

```
invalidData <- PortTalbotAQ %>%
  summarise(across(all_of(pollutants), ~ round(mean(is.na(.)) * 100, 2))) %>%
  pivot_longer(everything(), names_to = "pollutant", values_to = "percent_missing")

cat("\nTASK 2: Missing Values Percentage Table\n")
```

```
##
## TASK 2: Missing Values Percentage Table
```

```
print(invalidData)
```

```
## # A tibble: 11 x 2
##   pollutant percent_missing
##   <chr>          <dbl>
## 1 O3            1.98
## 2 NO            3.67
## 3 NO2           3.98
## 4 NOx           3.98
## 5 SO2           2.4
## 6 CO            4.74
## 7 PM10          6.54
## 8 PM25          6.86
## 9 wind_speed    2.05
## 10 wind_dir      2.05
```

```

## 11 temp                2.05
write_csv(invalidData, "invalidData.csv")

#####
# TASK 3: displayPollutantGraph() function
#####

displayPollutantGraph <- function(pollutant, start_month, start_year, end_month, end_year) {

  date_range <- seq.Date(
    from = as.Date(sprintf("%04d-%02d-01", start_year, start_month)),
    to = as.Date(sprintf("%04d-%02d-01", end_year, end_month)),
    by = "month"
  )

  df <- PortTalbotAQ %>%
    filter(date %in% date_range & hour == 12) %>%
    select(date, hour, all_of(pollutant))

  cat("\nTASK 3: Pollutant:", pollutant, "Rows:", nrow(df), "\n")
  print(head(df))

  p <- ggplot(df, aes(x = date, y = .data[[pollutant]])) +
    geom_line(linewidth = 1.3, color = "steelblue") +
    geom_point(size = 2.3, color = "darkred") +
    theme_minimal(base_size = 13) +
    labs(title = paste("Trend of", pollutant, "at noon on the 1st of each month"),
         x = "Date", y = pollutant)

  ggsave(paste0("plot_", pollutant, ".png"), p, width = 11, height = 6, dpi = 300)
  print(p)
}

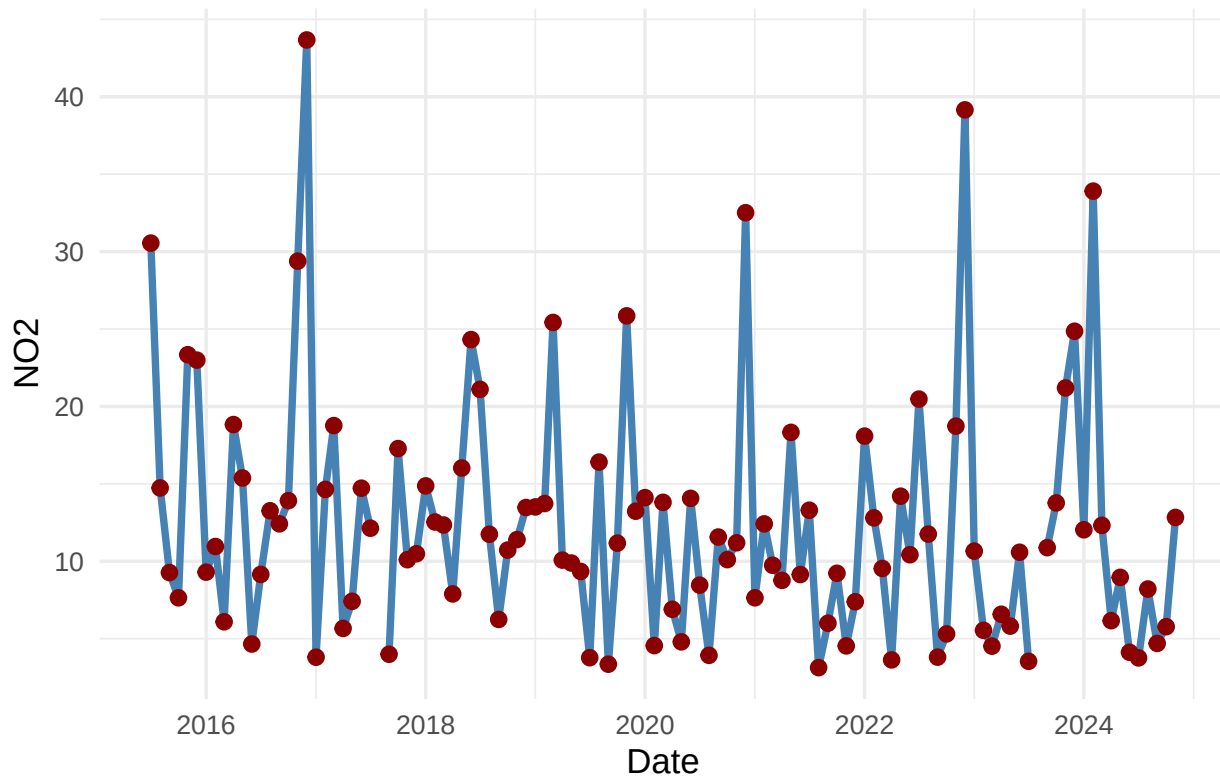
# Required demonstration
displayPollutantGraph("NO2", 7, 2015, 11, 2024)

##
## TASK 3: Pollutant: NO2 Rows: 113
## # A tibble: 6 x 3
##   date      hour  NO2
##   <date>    <int> <dbl>
## 1 2020-01-01    12 14.1
## 2 2020-02-01    12  4.55
## 3 2020-03-01    12 13.8
## 4 2020-04-01    12  6.90
## 5 2020-05-01    12  4.79
## 6 2020-06-01    12 14.1

## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
## Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).

```

Trend of NO2 at noon on the 1st of each month



```
#####
# TASK 4: comparePollutants() function
#####

comparePollutants <- function(pollutant_vector) {

  df <- PortTalbotAQ %>%
    filter(hour == 12 & day == 1) %>%
    select(date, all_of(pollutant_vector)) %>%
    pivot_longer(cols = pollutant_vector, names_to = "pollutant", values_to = "value")

  cat("\nTASK 4: Comparing pollutants\n")
  print(head(df))

  p <- ggplot(df, aes(x = date, y = value, color = pollutant)) +
    geom_line(linewidth = 1.2) +
    theme_minimal(base_size = 13) +
    labs(title = "Pollutant Comparison (Noon on 1st Day of Month)",
         x = "Date", y = "Pollutant Concentration")

  ggsave("compare_pollutants.png", p, width = 11, height = 6, dpi = 300)
  print(p)
}

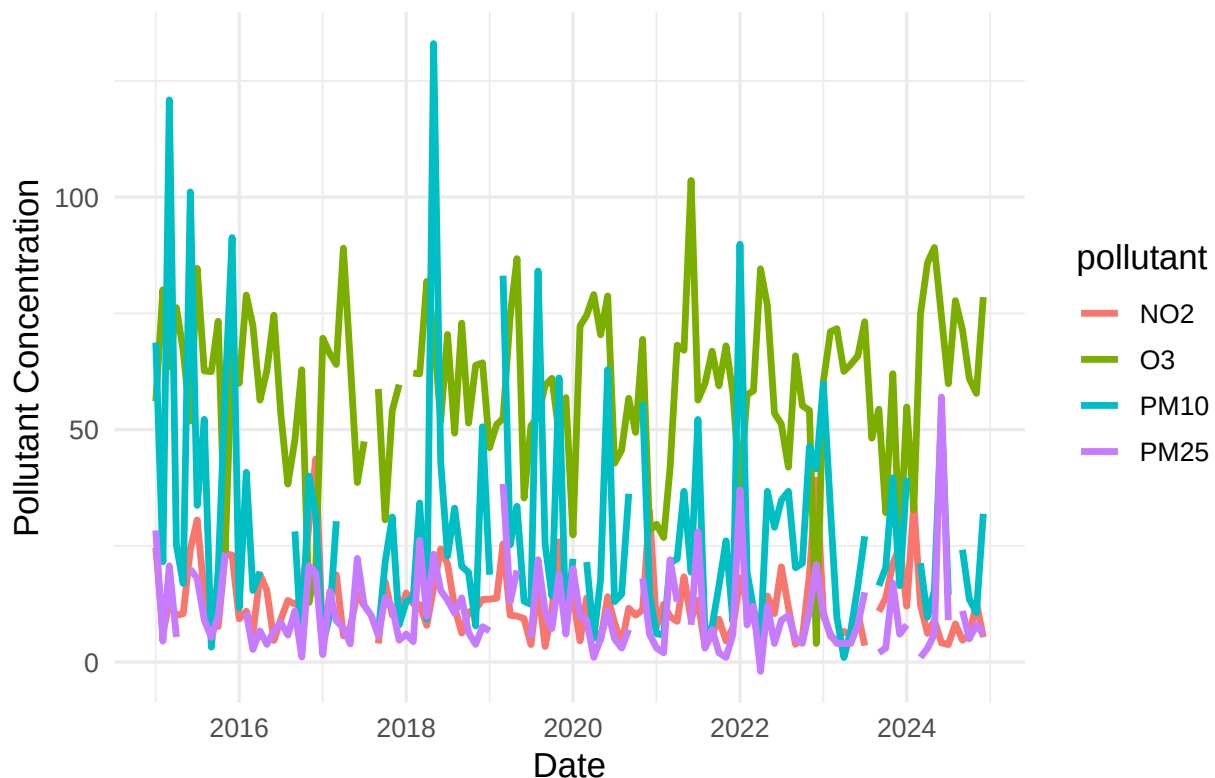
comparePollutants(c("O3", "NO2", "PM10", "PM25"))
```

```
## Warning: Using an external vector in selections was deprecated in tidyslect 1.1.0.
## i Please use `all_of()` or `any_of()` instead.
```

```
## # Was:
## data %>% select(pollutant_vector)
##
## # Now:
## data %>% select(all_of(pollutant_vector))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

##
## TASK 4: Comparing pollutants
## # A tibble: 6 x 3
##   date      pollutant value
##   <date>    <chr>    <dbl>
## 1 2020-01-01 O3      27.3
## 2 2020-01-01 NO2     14.1
## 3 2020-01-01 PM10    22.2
## 4 2020-01-01 PM25    20.1
## 5 2020-02-01 O3      72.3
## 6 2020-02-01 NO2      4.55
```

Pollutant Comparison (Noon on 1st Day of Month)



```
comparePollutants <- function(pollutant_vector) {
  df <- PortTalbotAQ %>%
    filter(hour == 12 & day == 1) %>%
    select(date, all_of(pollutant_vector)) %>%
```

```

    pivot_longer(cols = all_of(pollutant_vector), names_to = "pollutant", values_to = "value")

cat("\nTASK 4: Comparing pollutants\n")
print(head(df))

p <- ggplot(data = df, aes(x = date, y = value, color = pollutant)) +
  geom_line(linewidth = 1.4, alpha = 0.85) +
  theme_classic(base_size = 14) +
  scale_color_manual(values = c(
    "NO2" = "#d62728", # darker red
    "O3" = "#1b9e77", # green
    "PM10" = "#7570b3", # purple-blue
    "PM25" = "#e7298a" # pink
  )) +
  labs(
    title = "Pollutant Trend Comparison (1st Day Each Month at Noon)",
    x = "Date",
    y = "Pollutant Concentration (µg/m³)",
    color = "Pollutant"
  ) +
  theme(
    plot.title = element_text(face = "bold", size = 16),
    legend.position = "bottom",
    panel.grid.major = element_line(color = "grey90")
  )

ggsave("compare_pollutants.png", p, width = 11, height = 6, dpi = 300)
print(p)
}

comparePollutants(c("O3", "NO2", "PM10", "PM25"))

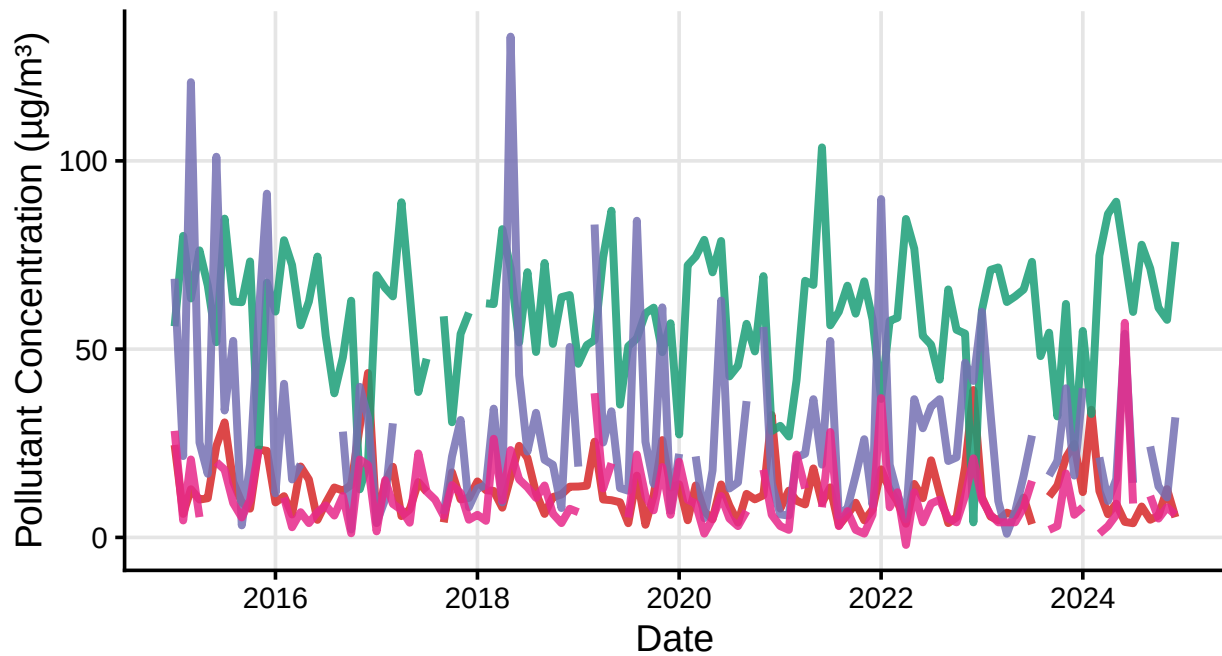
```

```

##
## TASK 4: Comparing pollutants
## # A tibble: 6 x 3
##   date      pollutant value
##   <date>    <chr>    <dbl>
## 1 2020-01-01 O3        27.3
## 2 2020-01-01 NO2        14.1
## 3 2020-01-01 PM10       22.2
## 4 2020-01-01 PM25       20.1
## 5 2020-02-01 O3        72.3
## 6 2020-02-01 NO2         4.55

```

Pollutant Trend Comparison (1st Day Each Month at No



Pollutant — NO2 — O3 — PM10 — PM25

```
#####
# TASK 5: Correlation between O3 and NO2
#####

O3NOCorrelation <- function() {

  df <- PortTalbotAQ %>%
    filter(!is.na(O3) & !is.na(NO2))

  correlation <- cor(df$O3, df$NO2, method = "pearson")

  cat("\nTASK 5: Pearson correlation between O3 and NO2 =", correlation, "\n")

  p <- ggplot(df, aes(x = NO2, y = O3)) +
    geom_point(alpha = 0.35) +
    geom_smooth(method = "lm", se = TRUE, color = "black") +
    theme_minimal(base_size = 13) +
    labs(title = "Correlation between NO2",
         x = "NO2 concentration", y = "O3 concentration")

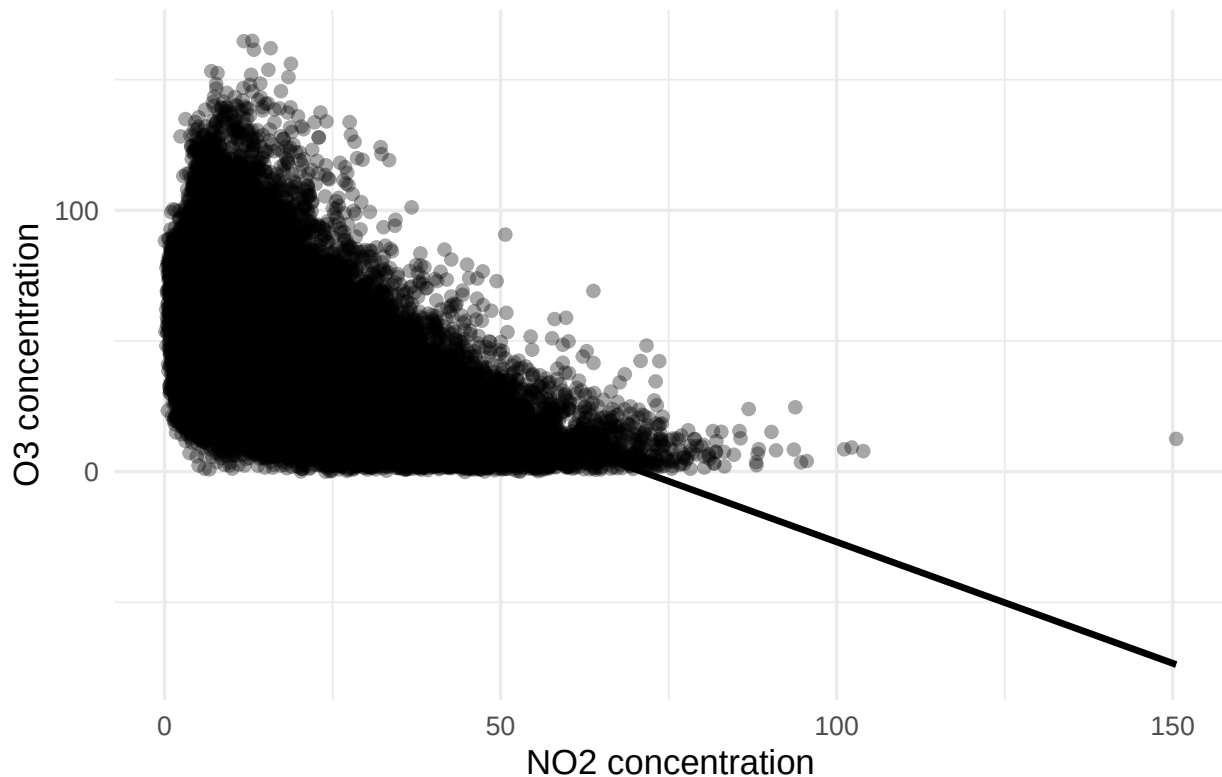
  ggsave("O3_NO2_correlation.png", p, width = 9, height = 6, dpi = 300)
  print(p)
}

O3NOCorrelation()
```

```
##
## TASK 5: Pearson correlation between O3 and NO2 = -0.4858485
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```

Correlation between O3 and NO2



```
#####
# TASK 6: Two-sample T-test
#####
```

```
df_ttest <- PortTalbotAQ %>%
  filter(!is.na(O3) & !is.na(wind_speed)) %>%
  mutate(group = ifelse(wind_speed < 5, "Low wind", "High wind"))

cat("\nTASK 6: Two-sample t-test (O3 by wind speed group)\n")
```

```
##
## TASK 6: Two-sample t-test (O3 by wind speed group)
print(summary(df_ttest))
```

```
##   date_str      time      O3      Status
## Length:84152 Length:84152 Min.   : -0.1497 Length:84152
## Class :character Class1:hms 1st Qu.: 40.0138 Class :character
## Mode  :character Class2:difftime Median : 54.3329 Mode  :character
##                Mode  :numeric Mean   : 52.9290
##                3rd Qu.: 66.5566
##                Max.   :164.8448
##
##      NO      Status.1      NO2      Status.2
## Min.   : -0.0593 Length:84152 Min.   :  0.0648 Length:84152
## 1st Qu.:  0.8636 Class :character 1st Qu.:  6.6630 Class :character
```

```

## Median : 2.5229 Mode :character Median : 11.0406 Mode :character
## Mean : 4.3700 Mean : 13.9189
## 3rd Qu.: 5.4747 3rd Qu.: 17.7589
## Max. :169.5575 Max. :150.5204
## NA's :1872 NA's :2142
## NOx Status.3 SO2 Status.4
## Min. : 0.261 Length:84152 Min. : -0.1329 Length:84152
## 1st Qu.: 9.081 Class :character 1st Qu.: 0.4610 Class :character
## Median : 15.563 Mode :character Median : 0.6713 Mode :character
## Mean : 20.577 Mean : 2.9040
## 3rd Qu.: 25.510 3rd Qu.: 1.4405
## Max. :317.217 Max. :253.8499
## NA's :2142 NA's :785
## CO Status.5 PM10 Status.6
## Min. : -0.0116 Length:84152 Min. : -3.865 Length:84152
## 1st Qu.: 0.0508 Class :character 1st Qu.: 10.300 Class :character
## Median : 0.0816 Mode :character Median : 16.425 Mode :character
## Mean : 0.1745 Mean : 23.256
## 3rd Qu.: 0.1458 3rd Qu.: 27.700
## Max. : 6.6115 Max. :574.889
## NA's :2969 NA's :4901
## PM25 Status.7 wind_dir Status.8
## Min. : -5.000 Length:84152 Min. : 0.0 Length:84152
## 1st Qu.: 4.800 Class :character 1st Qu.: 99.2 Class :character
## Median : 7.200 Mode :character Median :217.7 Mode :character
## Mean : 9.272 Mean :192.5
## 3rd Qu.: 11.900 3rd Qu.:275.9
## Max. :208.000 Max. :360.0
## NA's :5175
## wind_speed Status.9 temp Status.10
## Min. : 0.000 Length:84152 Min. : -6.200 Length:84152
## 1st Qu.: 3.300 Class :character 1st Qu.: 6.300 Class :character
## Median : 4.900 Mode :character Median : 9.500 Mode :character
## Mean : 5.437 Mean : 9.576
## 3rd Qu.: 7.100 3rd Qu.:13.100
## Max. :20.800 Max. :28.500
##
## date datetime year
## Min. :2015-01-01 Min. :2015-01-01 01:00:00.00 Min. :2015
## 1st Qu.:2017-07-02 1st Qu.:2017-07-02 07:45:00.00 1st Qu.:2017
## Median :2020-01-02 Median :2020-01-02 07:30:00.00 Median :2020
## Mean :2020-01-03 Mean :2020-01-03 15:32:28.49 Mean :2020
## 3rd Qu.:2022-07-03 3rd Qu.:2022-07-03 05:15:00.00 3rd Qu.:2022
## Max. :2024-12-31 Max. :2025-01-01 00:00:00.00 Max. :2025
##
## month day hour group
## Min. : 1.000 Min. : 1.0 Min. : 0.0 Length:84152
## 1st Qu.: 4.000 1st Qu.: 8.0 1st Qu.: 5.0 Class :character
## Median : 7.000 Median :16.0 Median :11.0 Mode :character
## Mean : 6.537 Mean :15.7 Mean :11.5
## 3rd Qu.:10.000 3rd Qu.:23.0 3rd Qu.:18.0
## Max. :12.000 Max. :31.0 Max. :23.0
##

```

```

test_result <- t.test(O3 ~ group, data = df_ttest)
print(test_result)

##
## Welch Two Sample t-test
##
## data: O3 by group
## t = 86.531, df = 80001, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group High wind and group Low wind is not e
## 95 percent confidence interval:
## 11.31642 11.84095
## sample estimates:
## mean in group High wind mean in group Low wind
## 58.73648 47.15779

if (test_result$p.value < 0.05) {
  cat("\nInterpretation: Reject H0. Wind speed affects O3 levels.\n")
} else {
  cat("\nInterpretation: Fail to reject H0. No evidence wind speed affects O3 levels.\n")
}

##
## Interpretation: Reject H0. Wind speed affects O3 levels.
#####
# TASK 7: ANOVA (NO2 for Nov-Dec @ 17:00 for 2016 / 2018 / 2020)
#####

anova_df <- PortTalbotAQ %>%
  filter(year %in% c(2016, 2018, 2020) & month %in% c(11, 12) & hour == 17) %>%
  select(NO2, year) %>%
  mutate(year = as.factor(year))

cat("\nTASK 7: ANOVA on NO2 across years\n")

##
## TASK 7: ANOVA on NO2 across years

anova_model <- aov(NO2 ~ year, data = anova_df)
print(summary(anova_model))

##
## Df Sum Sq Mean Sq F value Pr(>F)
## year 2 6866 3433 15.35 7.2e-07 ***
## Residuals 175 39133 224
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 5 observations deleted due to missingness

if (summary(anova_model)[[1]]$`Pr(>F)`[1] < 0.05) {
  tukey <- TukeyHSD(anova_model)
  print(tukey)
  write_csv(as.data.frame(tukey$year), "anova_posthoc_tukey.csv")
}

##
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##

```

```
## Fit: aov(formula = NO2 ~ year, data = anova_df)
##
## $year
##           diff          lwr          upr      p adj
## 2018-2016 -12.7239363 -19.204890 -6.242983 0.0000202
## 2020-2016 -13.5170197 -19.997973 -7.036066 0.0000057
## 2020-2018  -0.7930834  -7.301211  5.715044 0.9553023

p <- ggplot(anova_df, aes(x = year, y = NO2, fill = year)) +
  geom_boxplot() +
  theme_minimal(base_size = 13) +
  labs(title = "NO2 Levels (Nov-Dec @ 17:00)",
       x = "Year", y = "NO2 concentration")

ggsave("NO2_ANOVA.png", p, width = 9, height = 6, dpi = 300)

## Warning: Removed 5 rows containing non-finite outside the scale range
## (`stat_boxplot()`).

print(p)

## Warning: Removed 5 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

