

Análisis de la calidad del aire en Valencia

Ángela Alarcón Ballester, Lucía Ponce Salmerón y Carlos Sánchez Polo

2023-11-15

Índice

1	Introducción	1
1.1	Importación y acondicionamiento de los datos	1
1.2	Valores faltantes	3
2	Análisis de las variables	5
2.1	Análisis univariante	5
2.2	Análisis bivariante	6
3	Impacto de las Horas de Desplazamiento Laboral en la Calidad del Aire	6
4	Análisis de outliers	11
5	Conclusiones	13

1 Introducción

En este proyecto analizaremos y trataremos los datos por horas de calidad del aire de las estaciones de la red de vigilancia de la ciudad de Valencia entre los años 2016 y 2020. El conjunto de datos que se utiliza a lo largo de este proyecto se ha obtenido a través del portal de datos abiertos del Ayuntamiento de Valencia (https://valencia.opendatasoft.com/explore/embed/dataset/rvvcca_d_horarios_2016-2020/table/ link).

Comenzaremos realizando una exploración inicial de los datos y responderemos las preguntas que se deriven de ellos.

1.1 Importación y acondicionamiento de los datos

Cargamos la librerías que necesitaremos a lo largo del proyecto:

```
library(pacman)
pacman::p_load(readr,dplyr,tidyr,ggplot2,gridExtra,GGally,pheatmap,purrr,knitr)
```

Cargamos y visualizamos los datos:

```
rvvcca_d_horarios_2016_2020 <-
  read_delim("data/rvvcca_d_horarios_2016-2020.csv",
    delim = ";", escape_double = FALSE,
    col_types = cols(Fecha = col_date(format = "%Y-%m-%d"),
      `Fecha creacion` = col_date(format = "%Y-%m-%d"), ), trim_ws = TRUE)
```

Como se observa, las columnas Fecha y Fecha creación las definimos como fechas en formato “YYYY-MM-DD”.

Convertimos la columna Estacion y Dia de la semana en factor:

```
rvvcca_d_horarios_2016_2020$Estacion <- as.factor(rvvcca_d_horarios_2016_2020$Estacion)
rvvcca_d_horarios_2016_2020$`Dia de la semana` <- factor(rvvcca_d_horarios_2016_2020$
  `Dia de la semana`, levels = c("Lunes", "Martes", "Miercoles", "Jueves", "Viernes",
    "Sabado", "Domingo"))
```

Nos damos cuenta de que la columna Hora está en un formato complejo de manejar, por lo que cambiamos el formato de la columna Hora a uno más manejable:

```
rvvcca_d_horarios_2016_2020$Hora <- format(rvvcca_d_horarios_2016_2020$Hora,
  format="%H:%M:%S")
```

Ordenamos por fecha:

```
indices_orden <- order(rvvcca_d_horarios_2016_2020$Fecha)
rvvcca_d_horarios_2016_2020 <- rvvcca_d_horarios_2016_2020[indices_orden, ]
```

Eliminamos todas las columnas vacías ya que no nos aportan ninguna información. En este caso, solo se eliminará la columna Fecha baja.

```
rvvcca_d_horarios_2016_2020 <- rvvcca_d_horarios_2016_2020[,
  colSums(is.na(rvvcca_d_horarios_2016_2020)) < nrow(rvvcca_d_horarios_2016_2020)]
```

Utilizamos la función glimpse para obtener una visión general de los datos:

```
glimpse(rvvcca_d_horarios_2016_2020)
```

Observamos que contamos con las siguientes variables:

- Información general (Id, Fecha, Día de la semana, Día del mes, Hora, Fecha de creación). Son de tipo double, factor, fecha y caracter.
- Contaminantes atmosféricos (PM1, PM2.5, PM10, NO, NO2, NOx, O3, SO2, CO, NH3, C7H8, C6H6, C8H10) y Ruido. Estas variables miden la concentración de partículas en suspensión de diferentes tamaños y la concentración de óxidos de nitrógeno, ozono y azufre, entre otros. Todos son de tipo double.
- Información meteorológica (Velocidad del viento, Dirección del viento, Temperatura, Humedad relativa, Presión, Radiación, Precipitación, Velocidad máxima del viento). Todos son de tipo double.
- Estacion (Avda. Francia, Boulevard Sur, Molino del Sol, Pista Silla, Politécnico, Viveros, Centro, Consellería Meteo, Nazaret Meteo, Puerto València), la cual indica la ubicación de la estación de monitoreo. Es de tipo factor.

1.2 Valores faltantes

El estudio de los valores faltantes es esencial para garantizar que los análisis de datos sean sólidos y confiables. Al visualizar los datos inicialmente se han identificado valores faltantes, por lo que en este apartado procederemos a analizarlos. Primero, realizamos un resumen de la cantidad de valores faltantes en cada columna:

Id	0	Velocidad del viento	7874
Fecha	0	Direccion del viento	6020
Día de la semana	0	NH3	313088
Día del mes	0	C7H8	313156
Hora	0	C6H6	313635
Estacion	0	Ruido	309508
PM1	286441	C8H10	313053
PM2.5	183359	Temperatura	9033
PM10	183354	Humedad relativa	16357
NO	88740	Presion	6276
NO2	71383	Radiacion	6814
NOx	88738	Precipitacion	7414
O3	87741	Velocidad maxima del viento	8045
SO2	91245	Fecha creacion	0
CO	236277		

Cuadro 1: Número de valores faltantes de cada variable.

Podemos ver en el Cuadro 1 como hay variables con un gran número de valores faltantes, en especial, en los contaminantes atmosféricos y la información meteorológica. Esto es debido a que cada estación dispone de unos sensores distintos de medición de estos fenómenos. Además, el ruido también dispone de muchos valores faltantes, ya que depende de lo tranquila que sea cada estación. Por ello, consideramos oportuno dividir nuestros datos en función de la estación meteorológica asociada, para poder hacer una mejor comparativa, ya que en cada una de ellas se han medido unos valores concretos.

Por tanto, para poder realizar un mejor análisis vamos a agrupar nuestro dataframe por **Estacion**:

```
grupos <- rrvcca_d_horarios_2016_2020 %>% group_by(Estacion)

lista_dataframes <- split(grupos, f = grupos$Estacion)

nombres_dataframes <- unique(rrvcca_d_horarios_2016_2020$Estacion)
```

Limpiamos las columnas vacías de cada dataframe:

```
for (estacion in nombres_dataframes) {
  lista_dataframes[[estacion]] <- lista_dataframes[[estacion]] %>%
    select_if(~!all(is.na(.)))
}
```

Convertimos la lista de dataframes en dataframes independientes:

```
list2env(lista_dataframes, envir = .GlobalEnv)
```

Veamos con más detalle los parámetros que se han medido en cada estación:

```

parametros <- names(rvvcca_d_horarios_2016_2020)

tabla_zonas <- data.frame(Zona = character(0), stringsAsFactors = FALSE)

for (zona in 1:length(lista_dataframes)) {
  fila_zona <- data.frame(Zona = names(lista_dataframes)[zona], stringsAsFactors = FALSE)

  for (parametro in parametros) {
    fila_zona[[parametro]] <- parametro %in% colnames(lista_dataframes[[zona]])
  }

  tabla_zonas <- rbind(tabla_zonas, fila_zona)
}

rownames(tabla_zonas) <- tabla_zonas$Zona
colnames(tabla_zonas)[17] <- "V viento"
colnames(tabla_zonas)[29] <- "V_máx viento"
tabla_zonas <- subset(tabla_zonas, select = -c(Zona,Id,Fecha,`Dia de la semana`,
`Dia del mes`,Hora,Estacion,
`Fecha creacion`))
tabla_numeric <- as.data.frame(sapply(tabla_zonas, as.numeric))

```

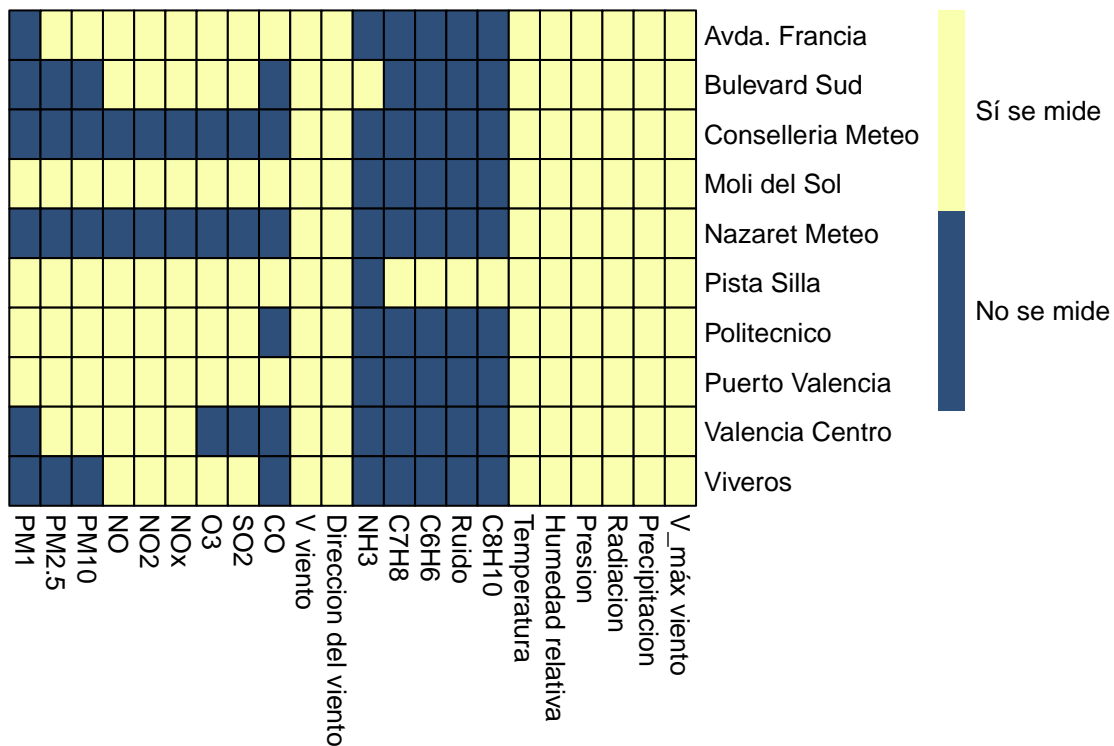


Figura 1: Variables medidas en cada estación.

Podemos ver en la Figura 1 que la información meteorológica se ha medido en todas las estaciones a diferencia de los contaminantes atmosféricos. Como la estación Pista Silla es la que menos columnas vacías tiene, es la que vamos a considerar para nuestro estudio. Veamos cuál es la proporción de valores faltantes en cada una de sus columnas, para ver que variables pueden sernos de utilidad:

PM1	0.636	C8H10	0.097
PM2.5	0.026	Temperatura	0.023
PM10	0.026	Humedad relativa	0.045
NO	0.026	Presion	0.016
NO2	0.026	Radiacion	0.015
NOx	0.026	Precipitacion	0.019
O3	0.065	Velocidad maxima del viento	0.021
SO2	0.070	CO	0.092

Cuadro 2: Proporción de valores faltantes de cada variable para la estación Pista Silla.

En el Cuadro 2 observamos que el número de datos faltantes no es significativo en ninguna de las variables, a excepción de PM1. Por ello, podemos considerar que esta variable no nos aportará información suficiente y podemos eliminarla de nuestro análisis.

Una vez que hemos visualizado y analizado los datos, podemos plantearnos ciertas preguntas. En nuestro caso, hemos considerado que, teniendo en cuenta las variables de las que disponemos, sería interesante realizar un estudio que analice la relación entre la calidad del aire y las horas de mayor cantidad de desplazamientos en automóvil, es decir, las horas de más tráfico en carreteras. ¿Se observará un aumento en la concentración de contaminantes atmosféricos en las horas puntas? ¿Dependerá esto de la estación en la que midamos los datos? ¿Y los fines de semana nos encontraremos ante un aumento o una disminución?

2 Análisis de las variables

2.1 Análisis univariante

Como hemos mencionado, vamos a hacer un estudio de los contaminantes atmosféricos, por lo que únicamente consideraremos estos valores de ahora en adelante. Hemos visto anteriormente que todos los contaminantes atmosféricos sí se miden en la estación Pista Silla. Por tanto, calculamos los estadísticos básicos de estos:

Cuadro 3: Estadísticos básicos de la estación Pista Silla.

Contaminante	mean	sd	median	IQR
NO	19.7068219	28.4383802	10.0	20.0
NO2	33.0132081	23.6815240	28.0	33.0
NOx	63.0690382	62.9793882	44.0	63.0
O3	45.0142906	26.7762457	46.0	42.0
SO2	4.3111563	2.2229034	4.0	1.0
CO	0.1735941	0.1138526	0.1	0.1
C6H6	1.5749097	1.6196948	1.1	1.6
C7H8	5.1178385	6.1959639	3.4	4.4
C8H10	2.3844622	3.3573891	1.3	2.5

Observamos en el Cuadro 3 que el dióxido de nitrógeno (NO2) tiene la media más alta con 33.01, seguido por el dióxido de azufre (SO2) con una media de 4.31. El monóxido de carbono (CO) tiene la concentración más baja con una media de 0.17. Además, se destaca que el rango intercuartílico (IQR) de los óxidos de nitrógeno (NOx) es amplio, indicando una variabilidad significativa en sus niveles.

La desviación estándar (sd) indica la dispersión de los datos con respecto a la media. Por ejemplo, el dióxido de nitrógeno (NO2) tiene una sd de 23.68, lo que sugiere una variabilidad considerable en las concentraciones de NO2. En contraste, el monóxido de carbono (CO) tiene una sd mucho más baja de 0.11, indicando una menor variabilidad en sus niveles.

2.2 Análisis bivalente

A continuación, vamos a realizar un análisis bivalente sobre las variables de interés de la estación meteorológica seleccionada anteriormente con el propósito de identificar patrones y posibles correlaciones entre los contaminantes atmosféricos. Para ello, filtraremos nuestro conjunto de datos para incluir solo las variables de interés y eliminaremos las filas que contienen valores faltantes.

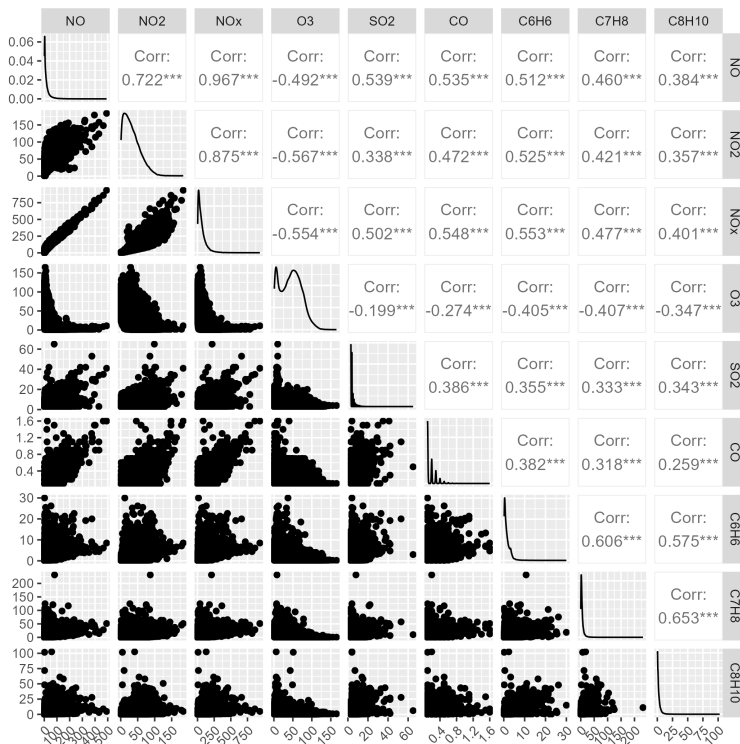


Figura 2: Correlación

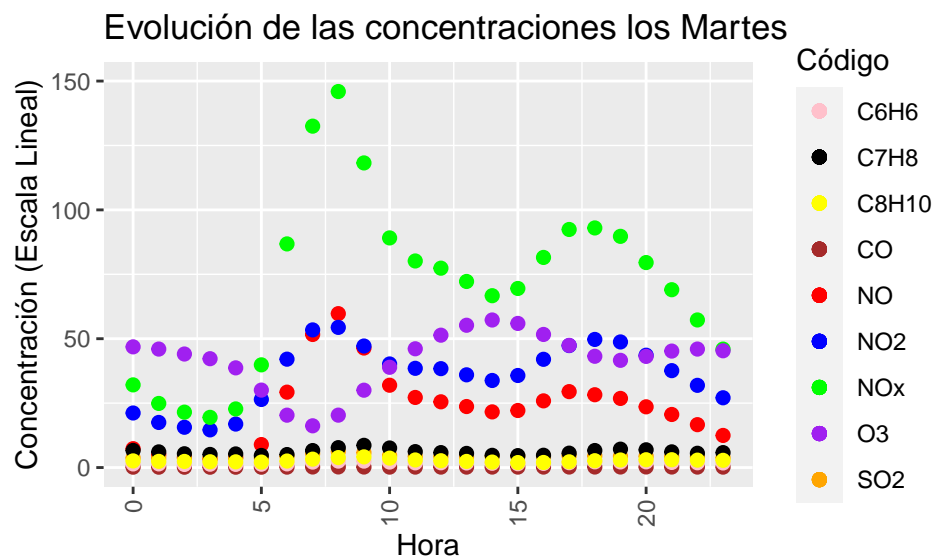
Se evidencia en la Figura 2 una relación lineal clara entre las variables NO2, NO y NOx, con una correlación superior al 0.7 entre ellas. Esta asociación tiene fundamentos sólidos, dado que los óxidos de nitrógeno (NOx) son compuestos químicos formados por átomos de nitrógeno y oxígeno. Entre los óxidos de nitrógeno más prevalentes se encuentran el óxido nítrico (NO) y el dióxido de nitrógeno (NO2), por lo que el término NOx engloba la mezcla de estos óxidos, indicando que un incremento en NO y NO2 conlleva un aumento en NOx.

La justificación de esta relación se encuentra en el proceso de formación de estos compuestos, principalmente durante la combustión de combustibles a elevadas temperaturas, como la que ocurre en motores de vehículos, centrales eléctricas y otras instalaciones industriales. Es relevante destacar que el NO2 es un contaminante atmosférico de considerable importancia.

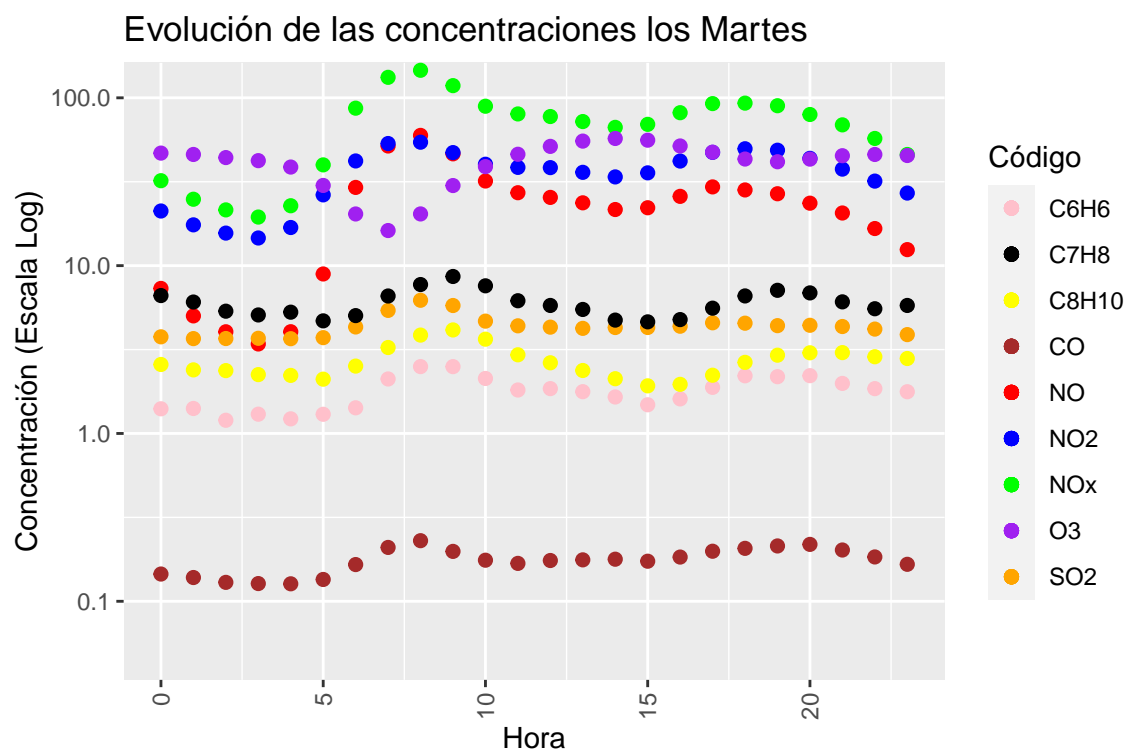
Por consiguiente, a partir de este punto, se podría considerar enfocar el análisis exclusivamente en el NOx para simplificar la investigación, dada su representatividad y la relación intrínseca con los óxidos de nitrógeno.

3 Impacto de las Horas de Desplazamiento Laboral en la Calidad del Aire

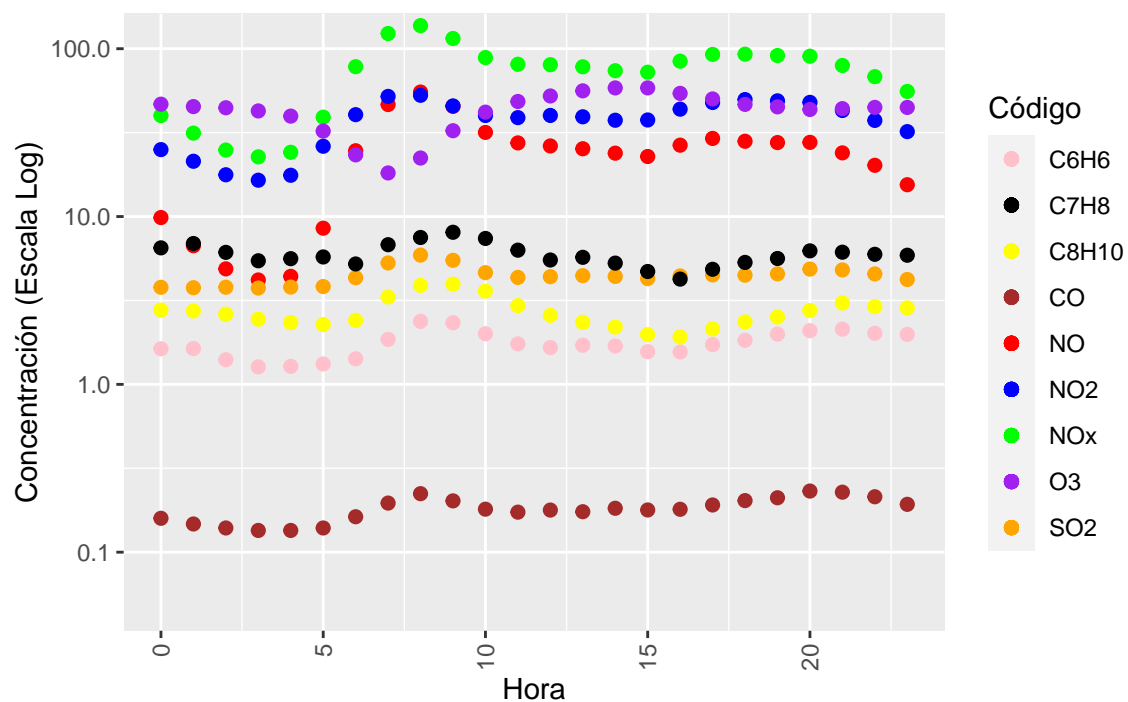
Una vez hecho el análisis podemos comenzar a dar respuesta a nuestras preguntas iniciales. Para ello, vamos a realizar la visualización de la concentración de contaminantes por horas del día, realizando a su vez, la comparación entre días de la semana.



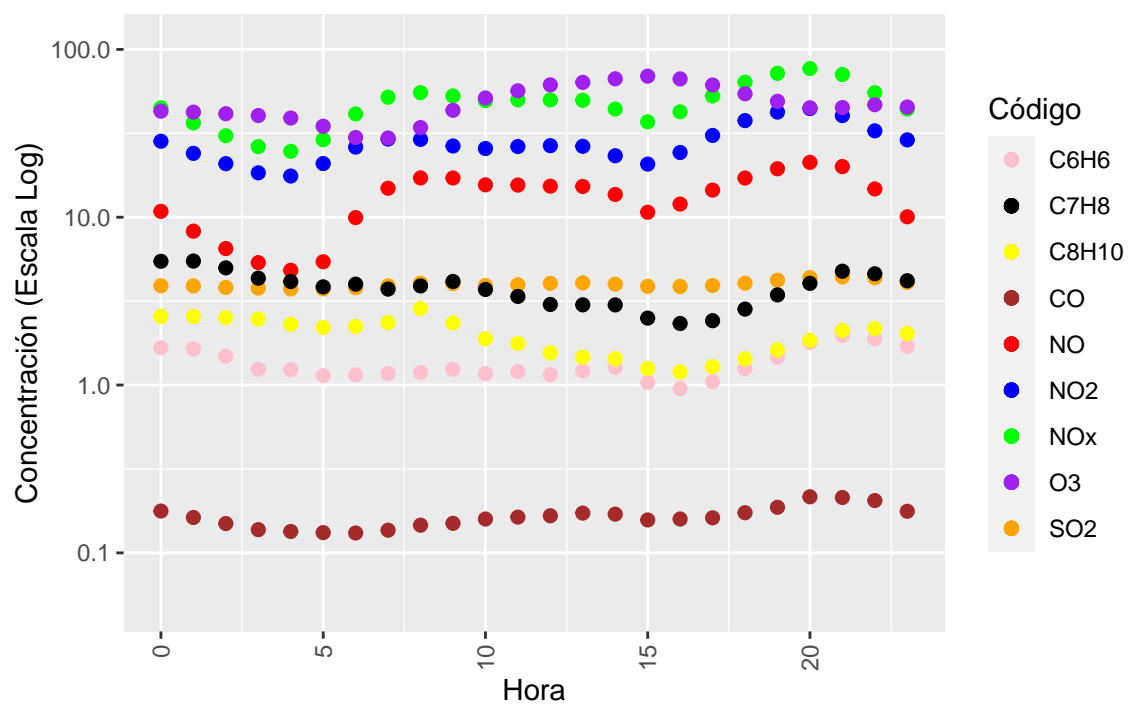
Somos capaces de apreciar los picos de contaminación en horas punta, sin embargo, los valores de las concentraciones de hidrocarburos son bajos, situándose en el extremo inferior de la imagen y solapándose entre ellos. Es por ello que hemos tomado la decisión de representar estos datos en escala logarítmica para varios días. Porque la forma de las gráficas es muy similar en los días laborables, solo mostraremos los martes, pero el resto de gráficas pueden consultarse en la documentación del trabajo.

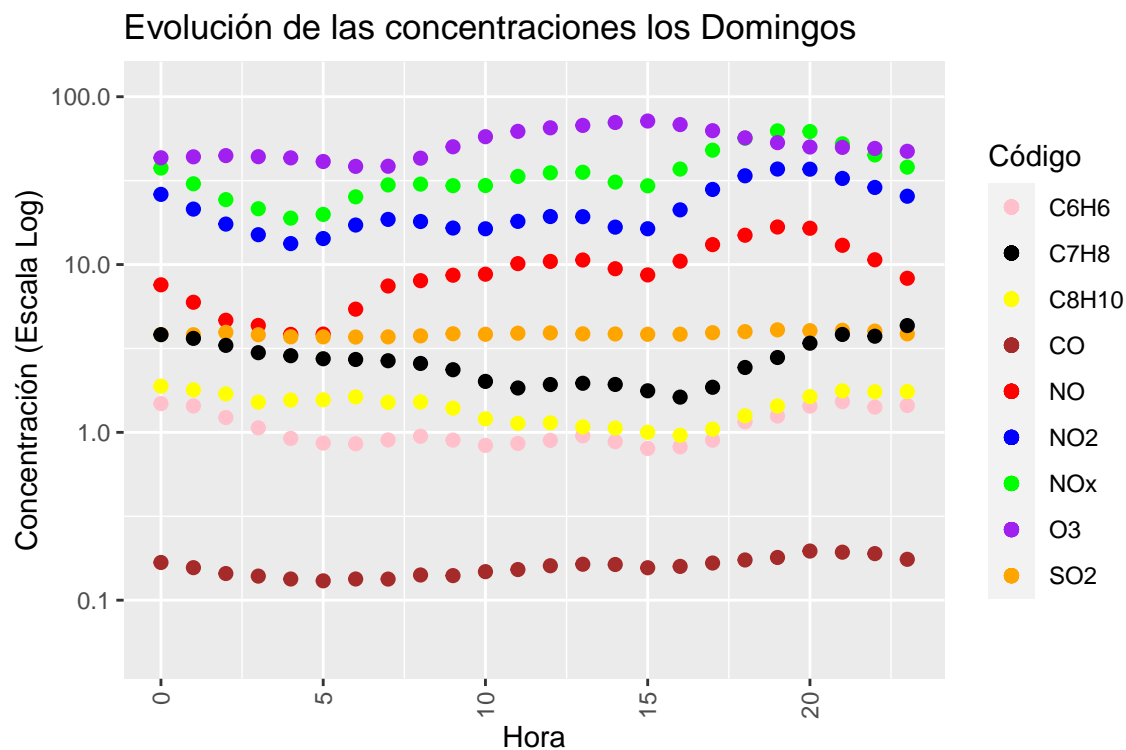


Evolución de las concentraciones los Viernes

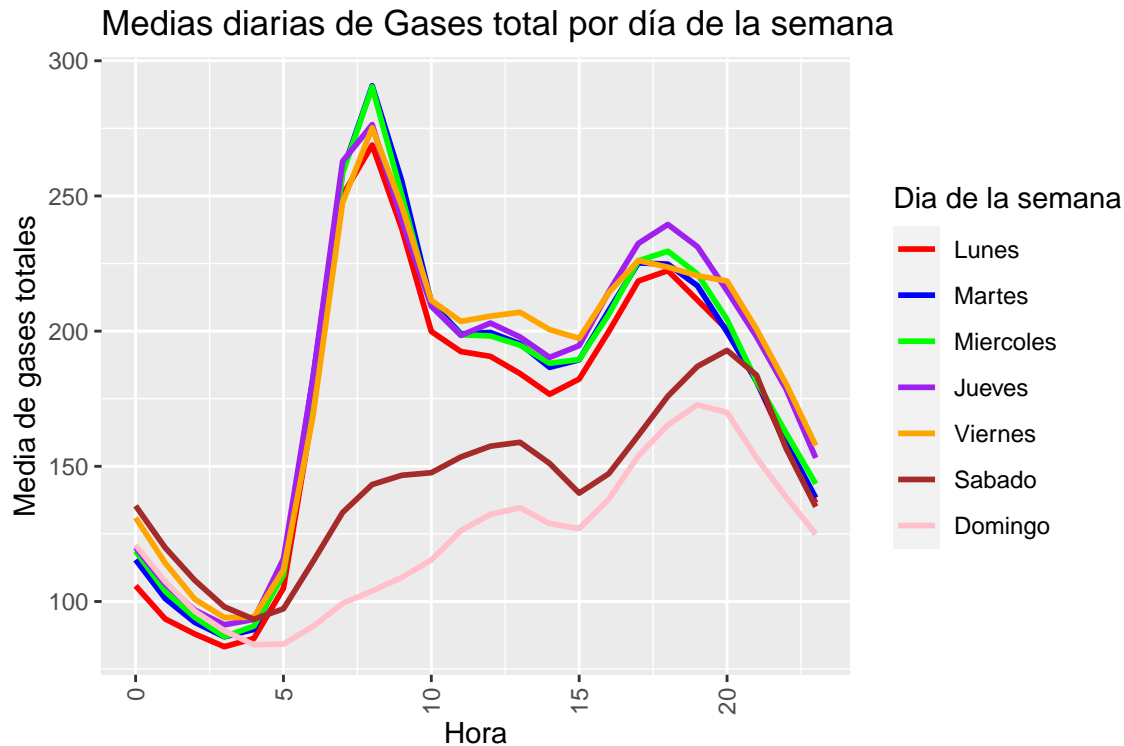


Evolución de las concentraciones los Sábados





Ahora que hemos representado las concentraciones en escala logartítmica, somos capaces percibir los picos de concentración incluso en aquellas variables menos presentes. Es de interés estudiar la variabilidad de las concentraciones en función del día de la semana. Como disponemos de 9 gases para cada día de la semana para cada una de las 24h, hemos decidido representar las medias de concentración por hora de forma general. Véase la siguiente gráfica:



Como comentábamos con anterioridad, los días laborables y los que no son claramente distinguibles. Las concentraciones mantienen levemente la forma, donde se observan picos a las horas puntas en los días laborales. Esto es debido principalmente a la ausencia de necesidad de desplazarse al trabajo, al colegio o a otras zonas de interés.

Podemos comprobar como la conducción de vehículos emite una variedad de gases que tienen un impacto significativo en la calidad del aire y en la salud pública. Entre los principales contaminantes vehiculares nos encontramos los que se registran en nuestro **dataset**. Todos ellos son residuos típicos de los combustibles fósiles. Hemos recopilado información sobre ellos:

- Óxidos de Nitrógeno (NO , NO_2 y NO_x): Producidos durante la combustión a alta temperatura en motores de vehículos. Contribuyen a la formación de ozono troposférico y puede irritar las vías respiratorias.
- Monóxido de Carbono (CO): Resulta de la combustión incompleta de carbono en el motor. Es peligroso porque interfiere con la capacidad de la sangre para transportar oxígeno. Puede tener efectos tóxicos.
- Óxidos de Nitrógeno (NO_x): Contribuye a la formación de smog y afectan la calidad del aire
- Hidrocarburos (C_6H_6 , C_7H_8 y C_8H_{10}): Proviene de la evaporación de combustibles y emisiones de escape. Algunos hidrocarburos son cancerígenos, y otros pueden contribuir a la formación de ozono y smog.

El incremento sustancial de estos contaminantes durante las horas de máximo tráfico se debe a la mayor actividad vehicular y la consiguiente combustión de combustibles fósiles. Está muy claro que los picos de concentración de estos gases no son casuales, ocurren en aquellos momentos de tráfico elevado, las gráficas respaldan esta afirmación. Los picos ocurren:

- En días laborables
- A las horas de entrar a trabajar/estudiar.

Este fenómeno destaca la necesidad de estrategias para mitigar las emisiones vehiculares y mejorar la calidad del aire en entornos urbanos. Un estudio más exhaustivo podría encargarse a expertos químicos.

4 Análisis de outliers

Finalmente, exploraremos la calidad del aire mediante boxplots que representan las concentraciones de dióxido de nitrógeno (NO_2) y ozono (O_3).

Comenzaremos con la variable NO_2 , ya que se trata de una medida representativa de los picos de contaminación del aire. Cada gráfico proporciona una visión detallada de la distribución de esta sustancia en las distintas estaciones de la red de vigilancia atmosférica de la ciudad de València. Comentaremos los resultados al final.

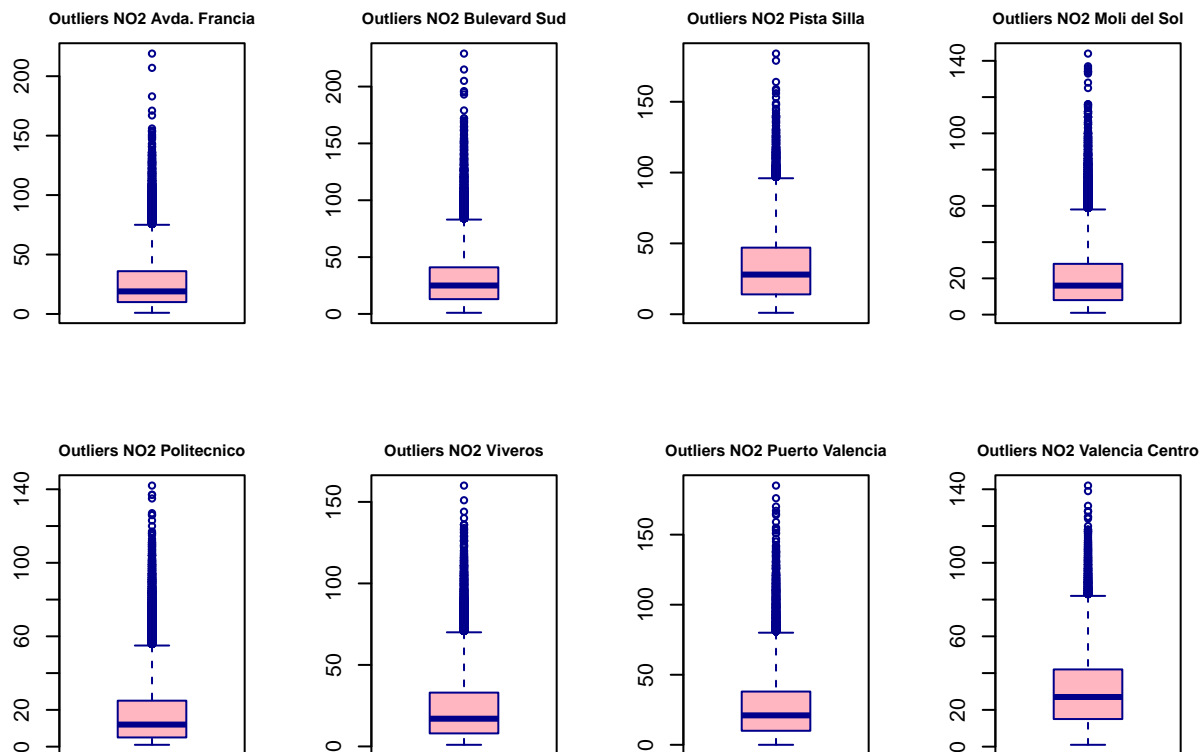


Figura 3: Diagramas de caja de NO_2 en cada estación.

Observamos una notable presencia de outliers en la variable NO_2 en todas las estaciones, destacando en la parte superior de los boxplots. Estos boxplots son estrechos, lo que indica una consistencia en las concentraciones normales de NO_2 , lo cual sugiere que la calidad del aire se mantiene relativamente estable en condiciones regulares. Sin embargo, la presencia de numerosos outliers en la parte superior indica que a menudo se generan concentraciones inusualmente altas de esta sustancia en el aire. Esto coincide con lo obtenido en las gráficas de contaminación por horas, ya que esta variable presentaba picos muy pronunciados a las horas puntas de tráfico de forma regular. Por tanto, podemos deducir que no se trata de valores atípicos, sino valores que reflejan situaciones específicas, como son los picos de tráfico en este caso. Afirmamos por tanto que, aunque estadísticamente considerados como outliers, estos valores son consistentes con los patrones observados en las gráficas de contaminación por horas.

En contraste, la presencia de outliers en la variable de ozono es menos pronunciada. Aunque en las estaciones de Viveros y Consellería Meteo se observa una mayor proporción de outliers, en general, los valores presentan una distribución más uniforme. Los boxplots son más anchos, indicando una dispersión más amplia de las concentraciones de ozono.

También hemos considerado oportuno implementar funciones específicas para la identificación de outliers

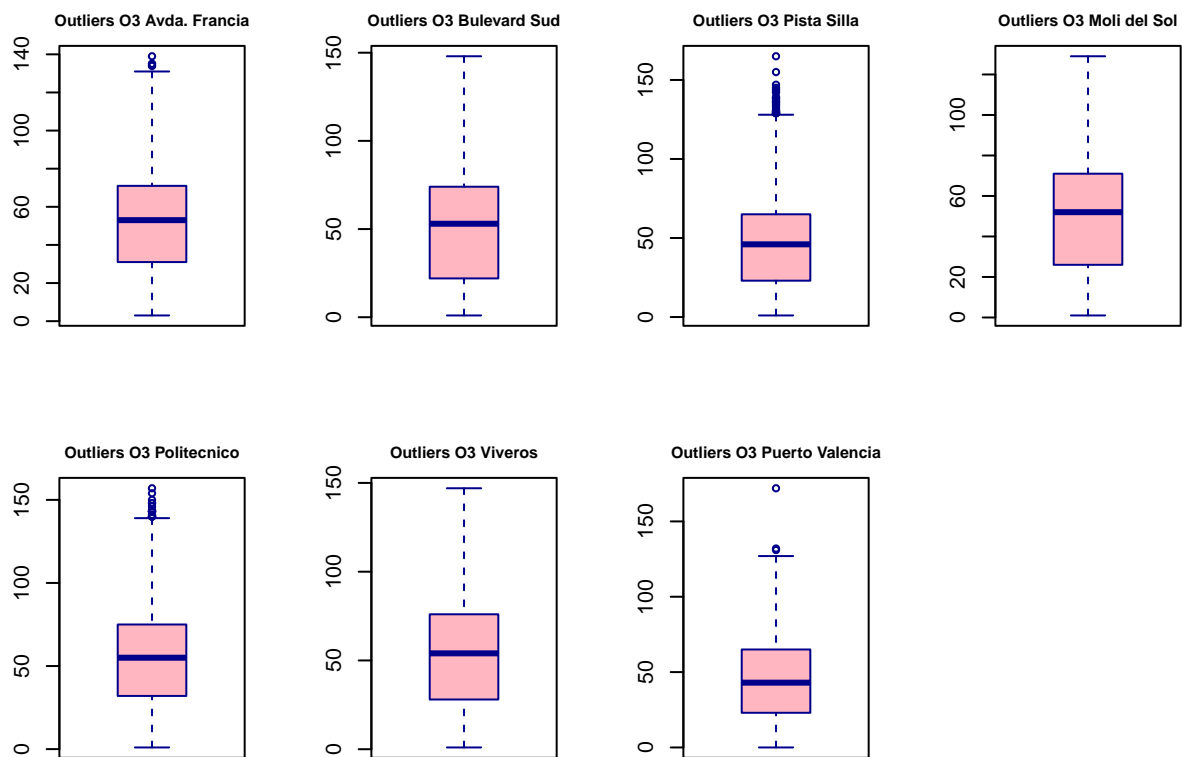


Figura 4: Diagramas de caja de O3 en cada estación.

utilizando distintos métodos estadísticos, como la regla de 3 sigma, el identificador Hampel y la regla de percentiles. Este es otro modo de resaltar valores que podrían considerarse atípicos y podemos comparar con los resultados obtenidos mediante el método boxplot.

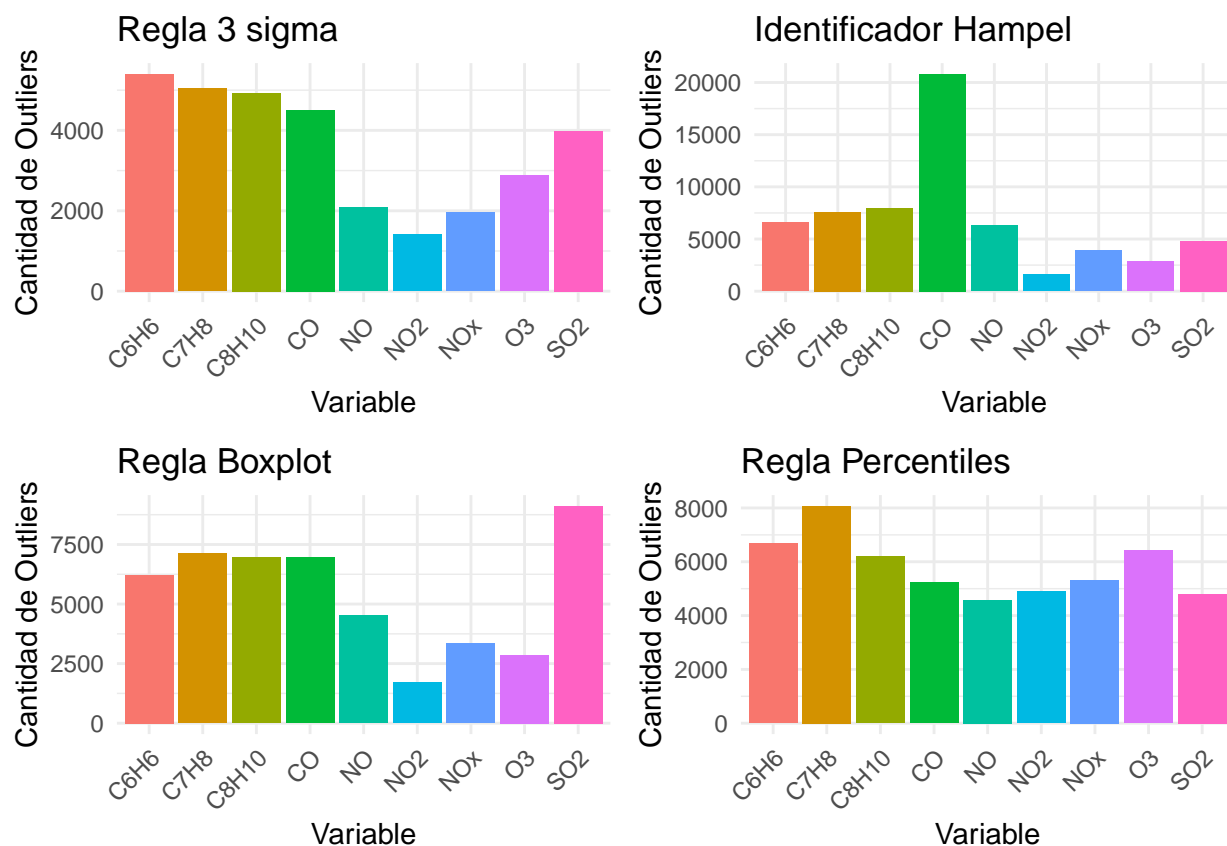


Figura 5: Gráficos de barras para detectar outliers.

La información resultante se presenta de manera clara y comparativa a través de gráficos de barras. Cada barra en estos gráficos representa la cantidad de outliers detectados para una variable específica, empleando diferentes métodos de detección.

En promedio, aplicando la regla de los percentiles obtenemos el mayor número de **outliers**. Esto se debe meramente a la cantidad de datos, pues independientemente de cuales sean sus valores, dejamos fuera a aquellos que no se encuentran entre los percentiles 5 y 95. Sin embargo, el identificador Hampel detecta una cantidad considerablemente mayor de valores anómalos de CO_2 que del resto de variables. Esto puede deberse a que muestra una tendencia muy uniforme salvo en las horas pico, marcándose estos datos como anómalos cuando en realidad son lecturas correctas. Con la regla 3 sigma obtenemos un número menor de outliers. Habría que estudiar si es procedente usar esta regla pues se presuponen distribuciones de datos de tipo gaussiana. Para acabar, la regla Boxplot arroja una cantidad de **outliers** intermedia, respecto al resto.

5 Conclusiones

Nuestra meta principal consistió en determinar si se observaba un incremento en la concentración de contaminantes atmosféricos durante las horas pico. Para lograrlo, examinamos detalladamente la concentración de gases por horas y días en una zona de alto tráfico, empleando una extensa variedad de sensores. Basándonos en los datos analizados en este estudio, afirmamos con confianza que efectivamente existe dicha relación.

Durante los días laborables, la concentración de gases alcanza picos notables en las horas pico, mientras que los sábados y domingos también registran picos de contaminación en las horas convencionales, aunque en menor medida. En estos mismos días, se observan nuevos picos durante la tarde, posiblemente vinculados a desplazamientos de corta duración relacionados con el retorno de segundas residencias de fin de semana, actividades grupales alejadas del lugar de residencia u otras razones lúdico-sociales.

Los compuestos que experimentan el mayor aumento en su concentración son los óxidos de nitrógeno, estrechamente ligados a los combustibles fósiles. A diferencia de los óxidos de nitrógeno, la concentración de O_3 no exhibe la misma dependencia del tráfico. Es probable que su comportamiento a lo largo del día no esté estrechamente vinculado a factores antropológicos.

EXTRA

```
# Vector con los nombres de los compuestos
compuestos <- c("NO", "NO2", "NOx", "O3", "SO2", "CO", "C6H6", "C7H8", "C8H10")

# Bucle for para generar histogramas y guardarlos
for (compuesto in compuestos) {
  # Generar el histograma
  hist_data <- `Pista Silla`[[compuesto]]
  hist_file <- paste("grafs/Histograma_", compuesto, ".png", sep = "")

  # Guardar el histograma como un archivo PNG
  png(hist_file)

  hist(hist_data, main = paste("Histograma de Concentración de", compuesto),
        xlab = "Concentración", ylab = "Frecuencia")

  dev.off()
}

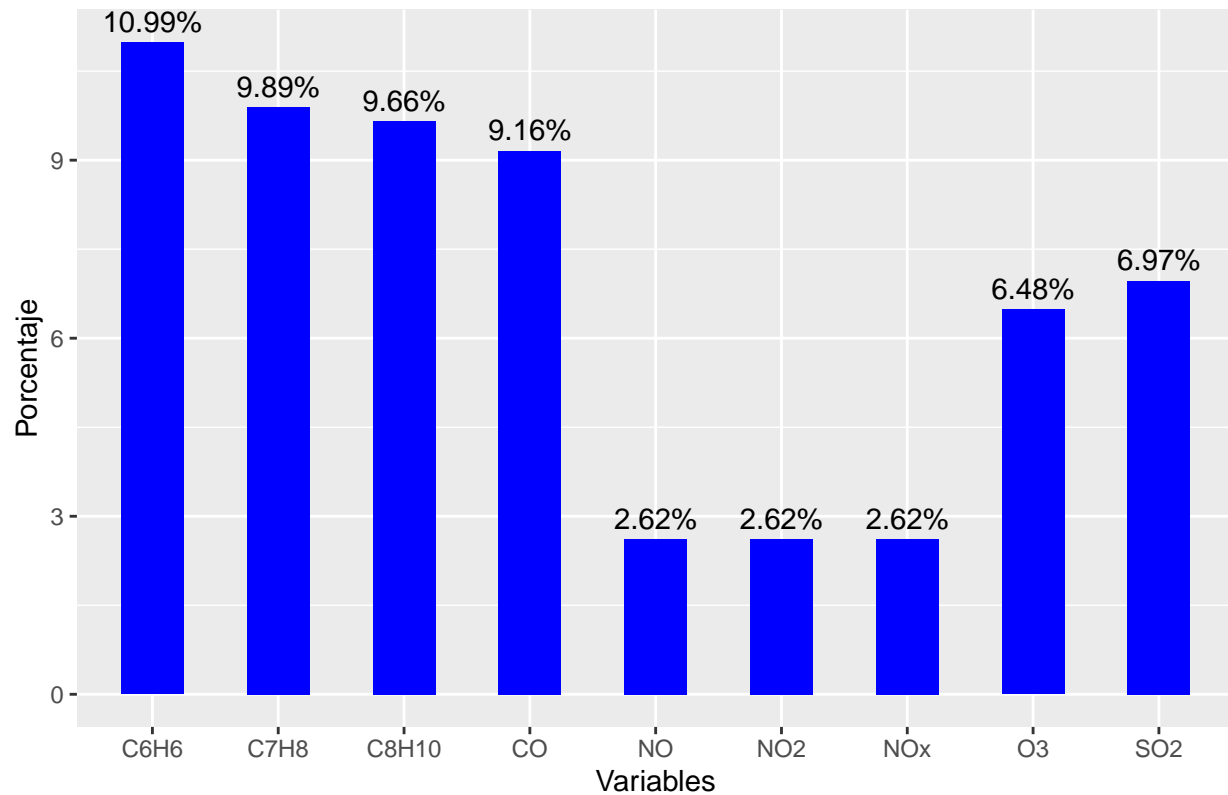
# Calcular el porcentaje de NA para las variables de interés
porcentajes_na <- colMeans(is.na(`Pista Silla`[, compuestos])) * 100

# Crear un gráfico de barras con ggplot2
library(ggplot2)

# Crear un objeto de datos para ggplot2
data_plot <- data.frame(variable = names(porcentajes_na), porcentaje = porcentajes_na)

# Crear el gráfico y ajustar el ancho de las barras
pplot<- ggplot(data_plot, aes(x = variable, y = porcentaje, fill = factor(porcentaje == 100))) +
  geom_bar(stat = "identity", width = 0.5) + # Ajustar el ancho de las barras según sea necesario
  scale_fill_manual(values = c("blue"), guide = 'none') +
  labs(title = "Porcentaje de NA por Variable", y = "Porcentaje", x = "Variables") +
  geom_text(aes(label = paste0(round(porcentaje, 2), "%")), vjust = -0.5)
print(pplot)
```

Porcentaje de NA por Variable



```
ggsave("grafs/Valores_NA.png", pplot, width = 8, height = 6, units = "in")
```

```
library(VIM)
```

```
## Warning: package 'VIM' was built under R version 4.3.2
```

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

```
## The legacy packages mapproj, rgdal, and rgeos, underpinning the sp package,  
## which was just loaded, were retired in October 2023.
```

```
## Please refer to R-spatial evolution reports for details, especially
```

```
## https://r-spatial.org/r/2023/05/15/evolution4.html.
```

```
## It may be desirable to make the sf package available;
```

```
## package maintainers should consider adding sf to Suggests:.
```

```
## VIM is ready to use.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
```

```
##
```

```
## Attaching package: 'VIM'
```

```
## The following object is masked from 'package:datasets':
##
##   sleep

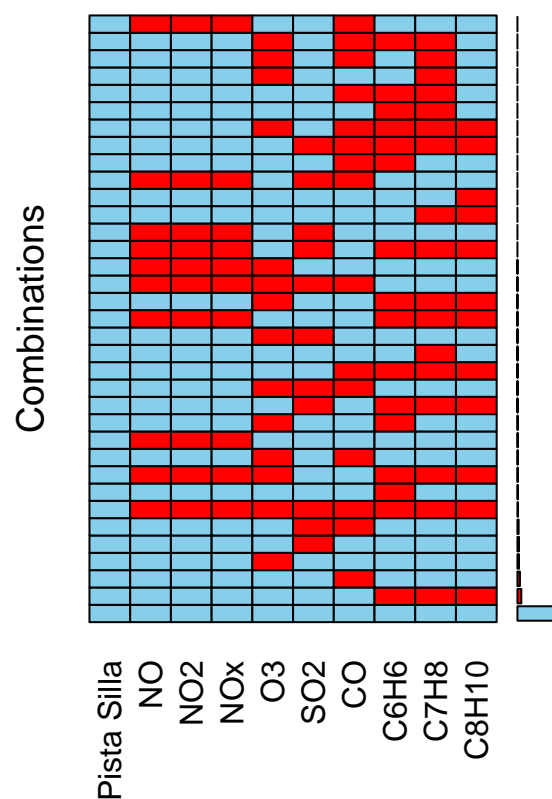
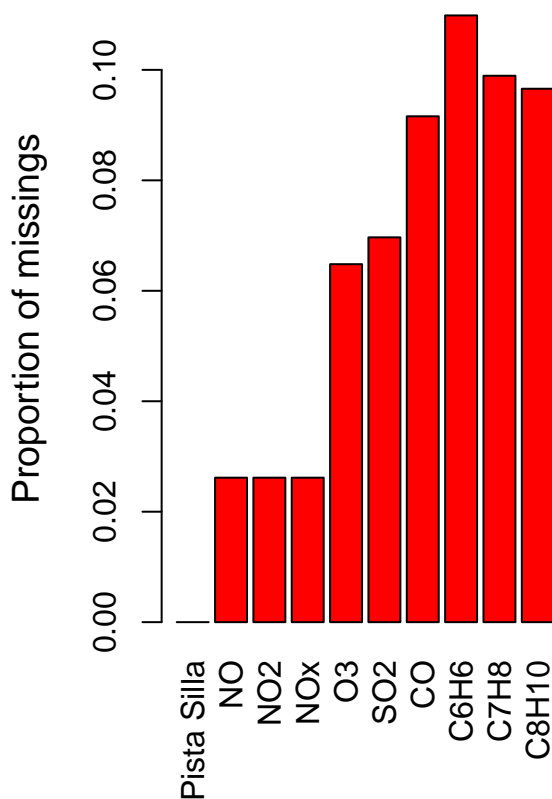
compuestos <- c("NO","NO2", "NOx", "O3", "SO2", "CO", "C6H6", "C7H8", "C8H10")

colnames(`Pista Silla`)[6] <- "Pista Silla"

agregado <- `Pista Silla`%>%
  select(compuestos) %>%
  aggr()

## Warning: Using an external vector in selections was deprecated in tidysselect 1.1.0.
## i Please use 'all_of()' or 'any_of()' instead.
##   # Was:
##   data %>% select(compuestos)
##
##   # Now:
##   data %>% select(all_of(compuestos))
##
## See <https://tidysselect.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.

## Adding missing grouping variables: 'Pista Silla'
```




```
ggsave("grafs/aggr.png", width = 8, height = 6, units = "in")  
colnames(`Pista Silla`)[6] <- "Estacion"
```