MDPI

Article

ProyectoAED2023

Carlos, Diego, Miguel

[1] Universidad de Valencia - ETSE, UV Avinguda de l'Universitat, 46100 Burjassot, Valencia; leutnant@fh-muenster.de
[2] ;
* Correspondence:

Simple Summary: A Simple summary goes here.

Abstract: Análisis exploratorio de un conjunto de datos sobre la calidad del aire de la ciudad de Valencia entre 2004 y 2022

Keywords: Contaminación, Valencia, datos, Análisis Exploratorio Datos

1. Introducción

En el presente informe planteamos el análisis exploratorio de un conjunto de datos sobre la calidad del aire de la ciudad de Valencia entre 2004 y 2022. El dataset empleado continene observaciones obtenidas de distintas estaciones de la red de vigilancia de Valencia. Las observaciones están compuestas por variables respectivas a diversas moleculas y elementos presentes en el aire junto a otras de tipo meteorológico como la velocidad del viento, la temperatura, etc.

El procedimiento a seguir comenzará con la correcta importación del dataset, a lo que seguirá un previo estudio de los datos con el objetivo de conocer la estructura del dataset y sus peculiaridades. Continuaremos con la preparación de los datos para resolver las preguntas planteadas y escogerán las variables de interés y los periodos temporales sobre los que se realizará el análisis univariante y bivariante, por otro lado también se gestionarán los outliers y datos faltantes. ç

Una vez preparado el dataset y estudiadas sus variables, procederemos a responder a las preguntas planteadas mediante diversas metodologías. Todo el proceso irá acompañado de las explicaciones pertinentes y finalizaremos con una conclusión del trabajo realizado.

2. Objetivos

El objetivo principal de este trabajo es familizarse con las herramientas y metodologías aprendidas para la carga, manipulación y análisis exploratorio de un conjunto de datos. Como objetivos principales tenemos: analisis univariante y bivariante, detección e imputación de NA's, detección y gestión de outliers. Por otro lado, planteamos otros objetivos en forma de preguntas:

- ¿Existe algún tipo de influencia sobre los niveles de gases contaminantes debido al cambio en el tráfico hacia el centro (creación carriles bici y reducción de carriles) de la ciudad en los últimos años?
- ¿Existe alguna correlación entre la calidad del aire y el día de la semana/año?
- ¿Existe cierta evolución de la contaminación sonora (años/zonas)?
- ¿Existe cierta evolución de la temperatura? (A medias, se puede hacer algo más o incluso relacionarlo con las precipitaciones. No se me ocurre cómo pero se le podría dar una vuelta)
- ¿Existe cierta evolución de la contaminación. Como medir la contaminación?

Citation: ProyectoAED2023. *Journal Not Specified* **2023**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

3. Análisis exploratorio de los datos

3.1. Importación de los datos

Para la importación de nuestro dataset fue necesario establecer ";" como el separador de las variables. Por otro lado, definimos el formato de las variables de forma previa ya que tras una exploración visual percibimos variables de tipo *factor* y *fecha*. Con el objetivo de estudiar las características generales de nuestro dataset realizamos un sumario con el tipo de cada variable *type*, cuantos valores distintos tiene *levels*, su valor más frecuente *topLevel*, cuántas veces aparece *topCount*, y en qué proporción *topFreq* y por último la proporción de datos faltantes de cada variable *missFreq*. En el Anexo 1, en la Tabla 1, encontramos las características generales obtenidas. A contrinuación explicamos cada variable:

- *Id* es un identificador para cada una de las filas y no aporta ninguna información concreta.
- *Fecha* es una variable tipo *Date* que nos proporciona la fecha en la que se tomaron los datos que componen la observación. Tomaremos esta variable para la ordenación ascendente de los datos.
- *Dia_de_la_semana* y *Dia_del_mes* son variables de tipo factor que indican en que día de la semana y en qué día del mes se realiza la medida. Puede ser extraído a partir de *Fecha* usando las funciones *wday()* y *day()*, de la librería *lubridate*.
- *Estacion* es una variable de tipo factor cuyos niveles son las distintas estaciones meteorológicas donde se tomaron las mediciones de las variables numéricas, haciendo un total de 13 estaciones.
- *PM1, PM2.5, PM10* son variables de tipo numérico con datos sobre la concentración de materiales particuldos (PM) de menos de 1, 2.5 y 10 micrómetros de diámetro respectivamente.
- *NO, NO2, NOx, O2, SO2, CO, NH3* son variables con datos sobre la concentración de estas moléculas inorgánicas consideradas contaminantes en el aire. Respecto a los óxidos de nitrógeno (NOx), estos son un grupo de gases compuestos por oxígeno y nitrógeno, es decir, NO y NO2 forman parte de este grupo y por lo tanto los valores de las variables estarán altamente correlacionados.
- *C7H8, C6H6, C8H10* son variables con datos sobre la concentración de estas moléculas orgánicas también consideradas contaminantes en el aire.
- *Vel_viento, Dir_viento, Temperatura, Humidad_rel, Presion, Radiacion_solar, Precipitacion, Max_vel_viento* son variables numéricas con mediciones de estas distintas condiciones ambientales al momento de medir las concentraciones de moléculas contaminantes en el aire.
- *As, Ni, Cd, Pb* son variables numéricas con datos de otras concentraciones de gases y metales contaminantes en el aire.
- *B(a)p* es una variable booleana que sólo presenta una observación. Elresto de datos son faltants y no podemos saber lo que representa esta variable.
- *Fecha_creación* y *Fecha_baja* son variables de tipo *Date* que parecen estar relacionadas con la creacion del dataset y que no parecen aportar información sobre nuestros datos. *Fecha_creación* solo presenta dos entradas distintas y *Fecha_baja* sólo presenta NA's por lo que estas variables parecen prescindibles.

En el sumario mencionado anteriormente se pueden observar grandes porcentajes de NA's en todas las variables. Esto junto al conocimiento de que los datos provienen de distintas estaciones, nos hace pensar que el dataset está compuesto por la unión de diversos datasets de los que provienen las mismas o distintas variables. En la siguiente sección vamos a analizar más profundamente los datos faltantes.

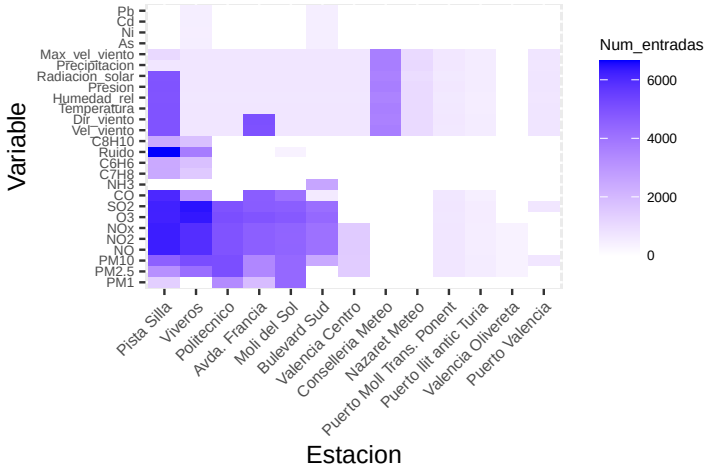

Figure 1

3.2. Análisis de datos faltantes y especiales

En esta sección, previamente al análisis univariante y bivariante de las variables, vamos a realizar un análisis más profundo de la estructura del dataset. El dataset parece estar compuesto por la unión de un conjunto datasets, posiblemente por las distintas estaciones de las que provienen los datos. Con el objetivo de obtener un conjunto de datos consistente, analizaremos el porcentaje de NA's que hay por cada variable, luego veremos que datos proporcionan las estaciones y con que frecuencia, también comprobaremos cuando comenzaron a generar datos las estaciones y durante que periodos de tiempo se han obtenido las variables.

En el Anexo 1, en la Tabla 1 podemos confirmar los elevados porcentajes NA's. Para comprobar el origen de los NA's en el mapa de calor de la Figura 1, podemos ver qué variables mide cada estación meteorológica, además de la cantidad de datos que aporta cada una de ellas sobre cada varaible numérica. Con este gráfico verificamos cada estación obtiene conjuntos distintos de variables y con frecuencias distintas, lo que hace pensar que las variables se toman en distintos periodos de tiempo.

Como podemos ver en la Figura 2, las estaciones se van incorporando al estudio con el paso del tiempo, esto explica la diferencia de observaciones obtenida por cada una de ellas. Por otro lado, nos interesa saber durante que periodos de tiempo se obtienen las variables. A partir del gráfico de la Figura 3 podemos ver el periodo de tiempo en el que las variables comienzan a ser medidas. Tras esta primera visualización de nuestros datos, podemos concluir que la gran cantidad de datos faltantes se debe a una combinación del hecho de que no todas las estaciones miden todas las variables, sumado a que los datos de ciertas variables se empiezan a medir posteriormente que el resto.

En nuestro caso vamos a escoger un período de 10 años, desde 2012 a 2022, ya que en este período de tiempo hay un número significativo de estaciones recogiendo datos y la mayoría de las variables son medidas en este intervalo de tiempo de forma consistente. No obstante, decidimos descartar las variables *Pb*, *Cd*, *Ni*, *As*, *B(a)p* y *NH3* debido a que se comienzan a medir después del 2012 y resultaría poco razonable imputar los datos de estas variables en esos años. De la misma forma, también descartamos las variables *C7H8*, *C6H6*, y *C8H10* debido a que presentan grandes períodos de tiempo con ausencia de datos intercalados entre 2012 y 2022, por lo que también sería poco razonable imputar estos datos.

3.3. Análisis univariante

3.3.1. Características univariantes

En esta sección analizaremos las variables de interés de forma individual con el objetivo de conocer sus magnitudes, es decir, las unidades de medida que representan, y como se distribuyen. Comenzamos con un summary de las variables numéricas para

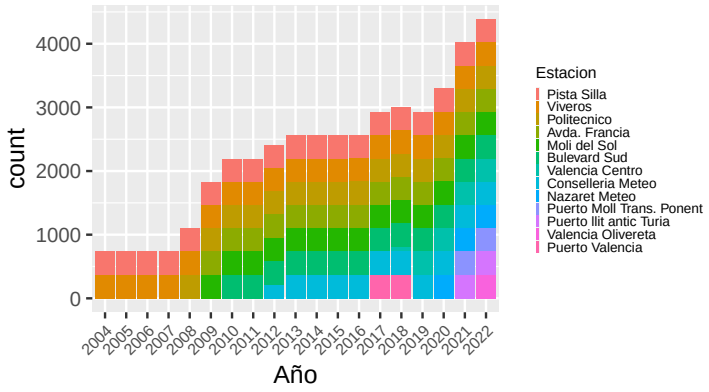

Figure 2

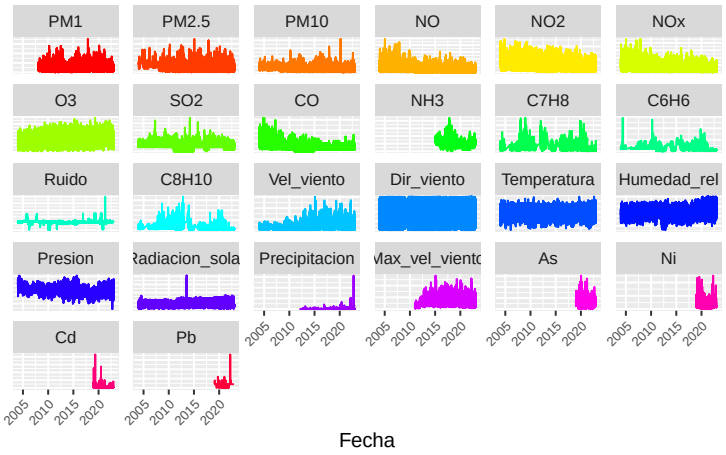

Figure 3

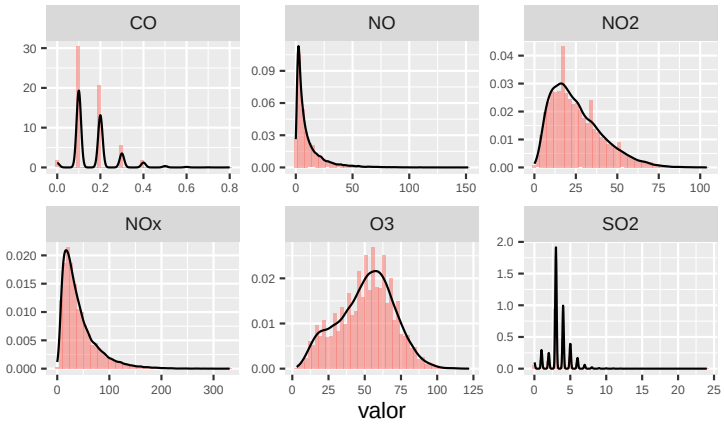

Figure 4

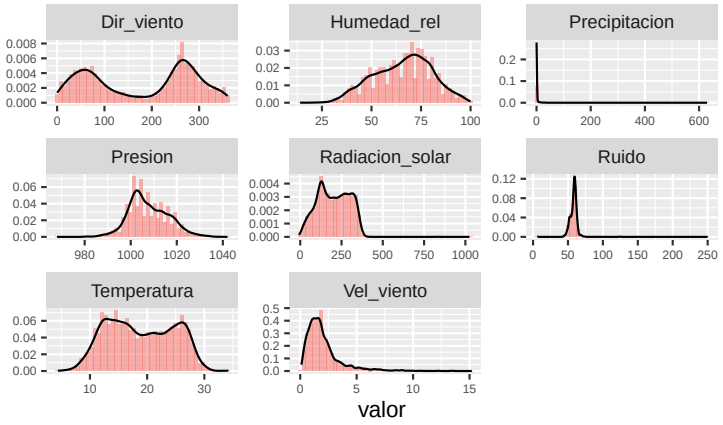

Figure 5

analizar sus magnitudes y como se distribuyen los datos. El resultado de este sumario lo podemos ver en el Anexo 1 Tabla 2. Mediante el sumario anterior podemos averiguar cuales son las unidades de medida empleadas contrastando con información externa, ya que en la fuente de los datos no se proporcionan.

- Gases (NO, NO2, NOx, SO2, CO, O3): microgramos por metro cubico (μg/m3)
- Temperatura: grados centígrados (C^o)
- Humedad: porcentaje (%)
- Presión: hectopascales (hPa)
- Velocidad del viento: kilometros/hora (m/s)
- Dirección del viento: ángulo de 0 a 360 (^{o})
- Precipitación: litros de agua por metro cuadrado (mm)
- Radiación solar: Vatios por metro cuadrado (W/m2)
- Ruido: decibelios (dB)

También es interesante observar las distribuciones de estas variables. Esto podemos observarlo en las figuras 4 y 5.

3.3.2. Detección y eliminación de outliers con métodos univariantes

Compararemos el funcionamiento de lass reglas tresSigma, hampel, boxplot y percentiles. En la Figura 6 podemos observar cómo la regla del percentil siempre elimina datos (el 2.5% superior e inferior). Podemos usar esto como comparador de el resto de métodos. Es visible el agresivo comportamiento de la regla boxplot y el excesivamente poco agresivo método de la regla 3-sigma. Hemos decidido utilizar la regla de Hampel para la detección

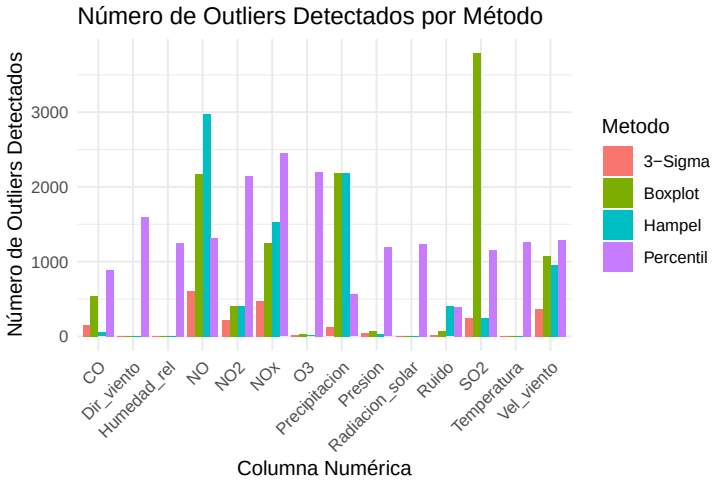

Figure 6

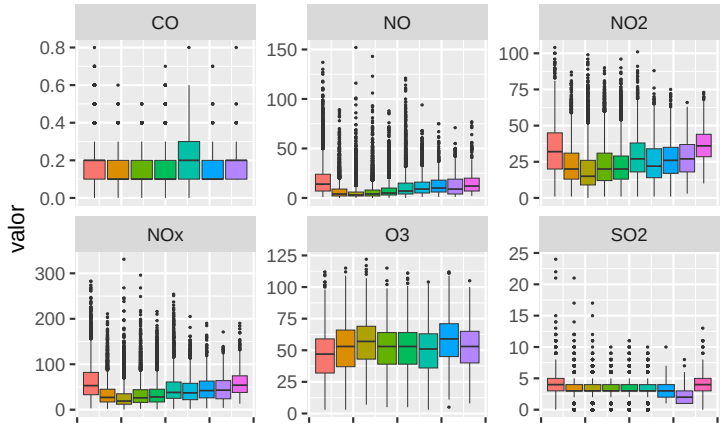

Figure 7

de los outliers, ya que esta es una regla robusta a distribuciones que no sean Gaussianas. En cuanto a qué hacer con estos outliers, hemos decidido sustituirlos por NA para su posterior inputación.

3.4. Analisis bivariante

3.4.1. Características bivariantes

Mediante boxplots vamos a analizar como se distribuyen los datos respecto a las estaciones. En la figura 7 se observan las referentes a gases y en la figura 8 el resto. A partir de ambas figuras podemos concluir con que generalmente la estación no influye en los valores de las variables. Esta información nos es útil para la imputación de NA's, ya que podemos emplear datos de otras estaciones.

También visualizaremos la evolución mediante boxplots de el valor de las variables a lo largo de los años. En la Figura 9 se puede apreciar una leve reducción a lo largo de los años de los gases considerados como contaminantes a excepción del SO2 que parece mantenerse constante. Por otro lado vemos un leve aumento del ozono (O3), por lo que podría existir una relación con la disminución de gases contaminantes. Continuamos con las variables meteorológicas. Como era de esperar, en la Figura 10 vemos como las variables meteorológicas no han sufrido variaciones a lo largo de los años. Por otro lado si que se aprecia un leve aumento en el ruido.

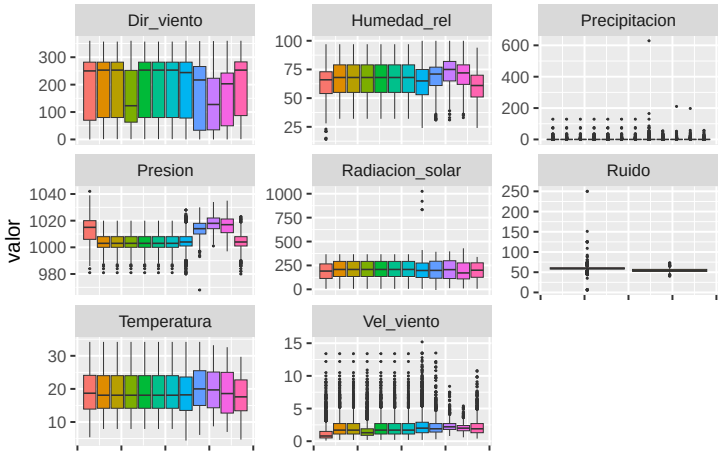

Figure 8

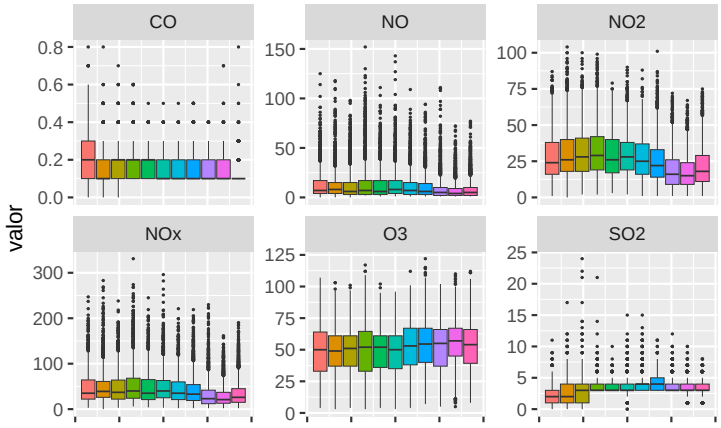

Figure 9

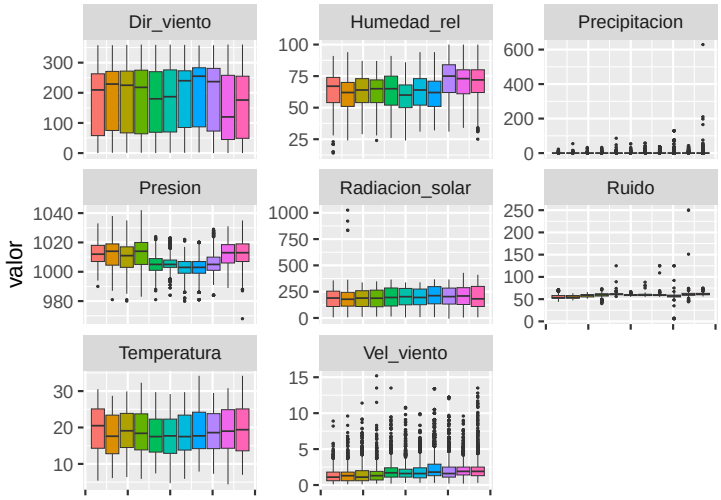

Figure 10

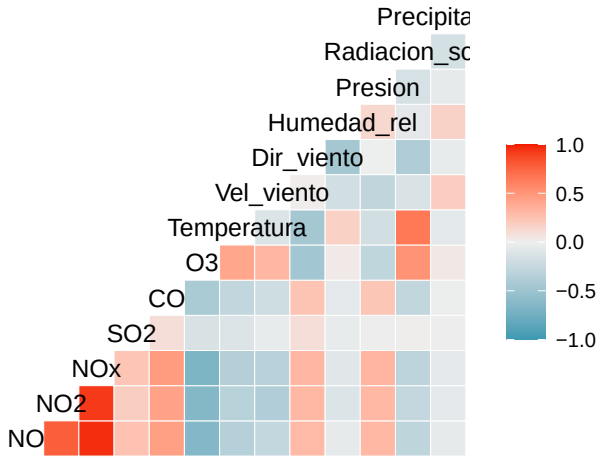

Figure 11

3.4.2. Imputación de datos anómalos

Ahora que conocemos las unidades de nuestras variables tras haberlas analizado, sabemos cuál es el rango de valores puede tomar cada una. Para tratar estos datos anómalos, en primer lugar se convertirán en NA's y a continuación se les imputará un valor igual que el resto de datos faltantes. En el anáslisis univariante vimos que los datos diarios de las variables en todas las estaciones de Valencia suelen ser parecido. Decidimos usar este hecho para imputar los datos faltantes en cada día con la media de esta variable sobre todas las estaciones.

3.4.3. Correlaciones

Basándonos en la Figura 11, podemos extraer las siguientes conclusiones:

- Existe una correlación negativa importante entre NOs y Ozono. Esto podría deberse a que la presencia de óxidos de nitrógeno puede contribuir a la degradación del ozono en la atmósfera, lo que tiene implicaciones para la calidad del aire, la contaminación y el efecto hivernadero.
- Correlación positiva de NOx con el resto de NO (NO, NO2...). Como era de esperar y como se ha mencionado en la definición de las variables, NOx representa el conjunto de oxidos de nitrogeno, entre ellos el NO y el NO2.
- Correlación positiva entre Temperatura y Radiación Solar. Esto es coherente con las estaciones más cálidas que a menudo experimentan más horas de sol y mayor radiación solar.
- Correlación entre Radiación Solar y Ozono. Esto tiene sentido ya que la capa de ozono filtra la mayor parte de la radiación ultravioleta proveniente del sol, por lo tanto están muy relacionados entre ellos.

3.4.4. Detección y eliminación de outliers con métodos multivariable

La distancia de Mahalanobis mide la distancia de un punto de datos a un conjunto de datos multivariado centrado y escalado según la matriz de covarianza. Los puntos que tienen distancias de Mahalanobis más grandes son considerados como outliers, ya que están más lejos de la distribución típica de los datos. En otras palabras: con estos métodos multivariable no nos fijamos en que un valor sea anómalo, sino una combinación de estos, siendo representados en un espacio vectorial y midiendo la distancia de cada punto (muestra) con el resto. Lo que haremos al encontrar un outlier será eliminarlo directamente, pues toda la muestra será anómala.

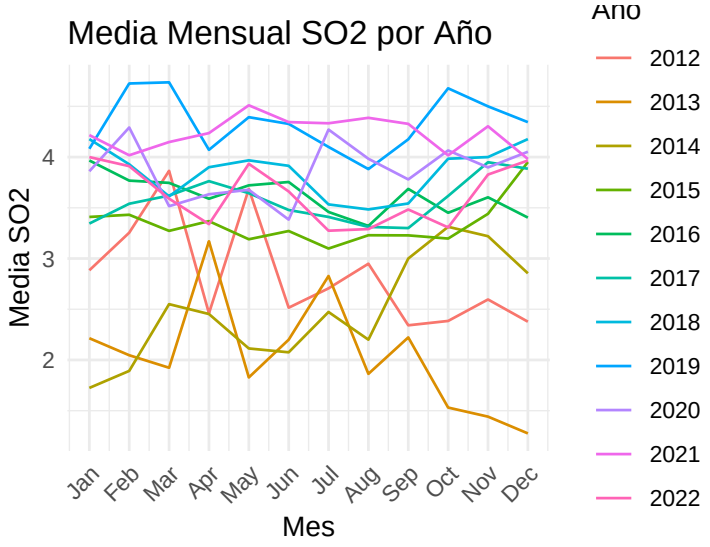

Figure 12

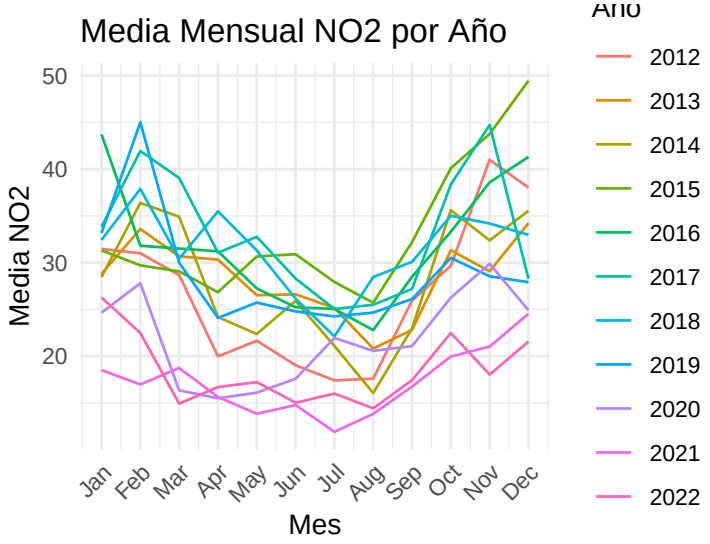

Figure 13

4. Resolución de preguntas planteadas sobre los datos

4.1. Influencia de carril bici

Para estudiar la influencia de un carril bici por el centro de Valencia, estudiaremos la evolución de gases contaminantes en diversas estaciones de la ciudad. Mediremos la evolución de los gases SO2 y NO2, relacionados con la combustión de carburantes fósiles, el tráfico rodado y las emisiones de determinadas industrias y grandes instalaciones de combustión. Además, las estaciones deberían ser las más céntricas, ya que ahí el efecto del carril bici y las restricciones de acceso deberían hacerse más notables. Observaremos las siguientes estaciones: Avda. Francia, Bulevard Sud, Valencia Centro y Olivereta

Si bien, en la Figura 12 no se observan cambios significativos en los niveles de SO2, en la Figura 13 sí que podemos observar una evolución en los niveles de NO2 a lo largo de los años. Los niveles más bajos se dan en los últimos 3 años. Además, se puede apreciar una reducción periódica de los niveles de NO2 en los meses de verano, que coincide con el momento del año donde menos gente vive en la ciudad y menos tráfico hay. En conclusión parece que las medidas de transporte que se han tomado en la ciudad han influenciado en la reducción de gases contaminantes.

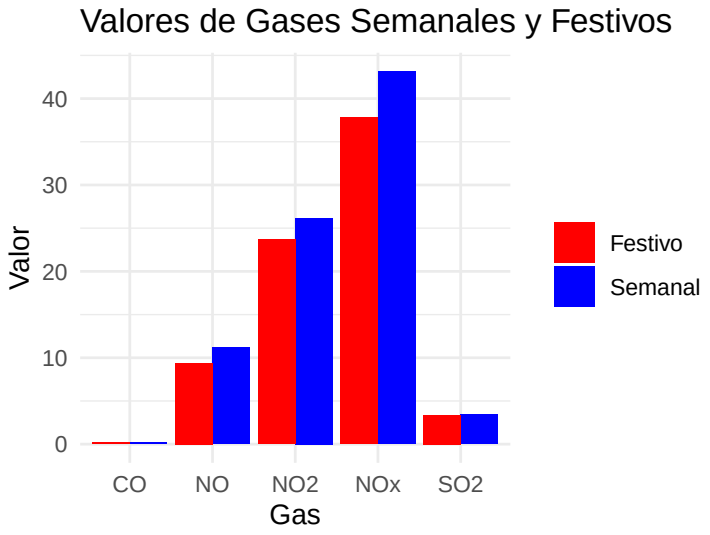

Figure 14

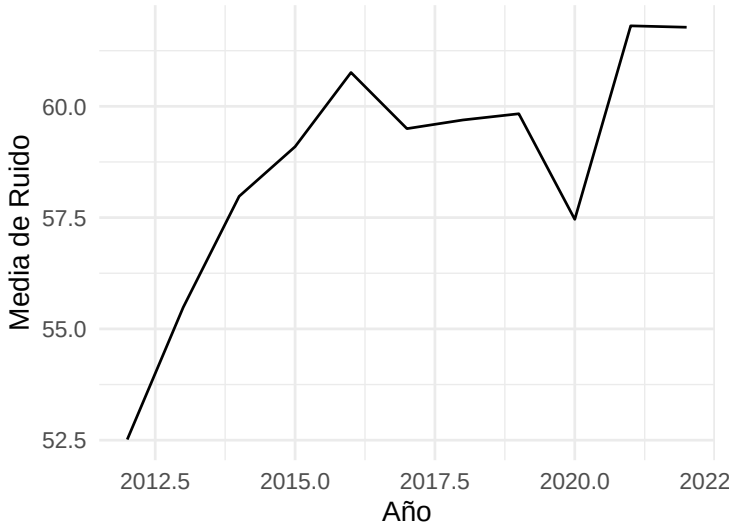

Figure 15

4.2. Relación entre la calidad del aire y el dia de la semana

En esta pregunta observamos si, en general la calidad del aire se ve influenciada por el día de la semana. Queremos sobre todo fijarnos en la distinción entre días laborables y de descanso. Podemos comprobar cómo, efectivamente, existe una disminución en los gases contaminantes los días festivos con respecto a los días laborables.

4.3. Evolución de la contaminación sonora a lo largo de los años

En la Figura 15 podemos observar un incremento en el ruido medio a lo largo de los años. En los datos del 2020 podemos observar un decrecimiento de los mismos. Podría estudiarse si este efecto es debido al Covid-19.

Con la figura 16 comprobaremos si este ruido medido depende de la zona dónde la midamos. Podemos observar que la desviación es despreciaable. Sí que podemos determinar, gracias a contrastar estos datos con los de la gráfica boxplot **??**, es que la evolución en el ruido no viene dada por una tendencia natural. Más bien, el gráfico nos indica que existe un aumento significativo de ruido en días concretos.

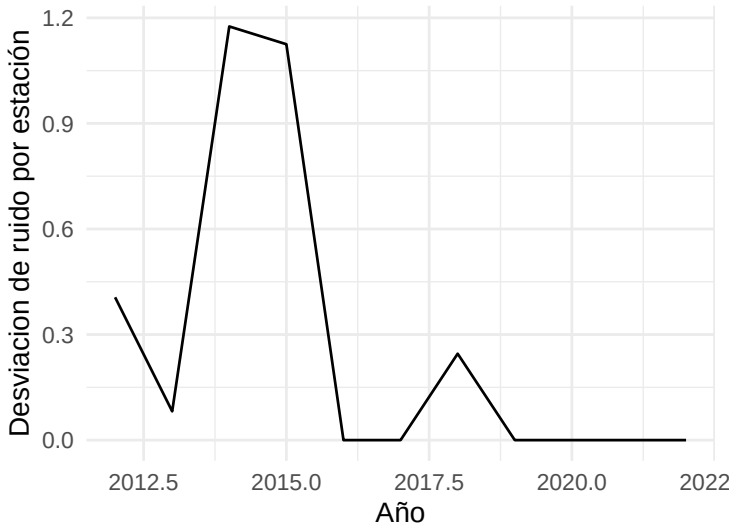

Figure 16

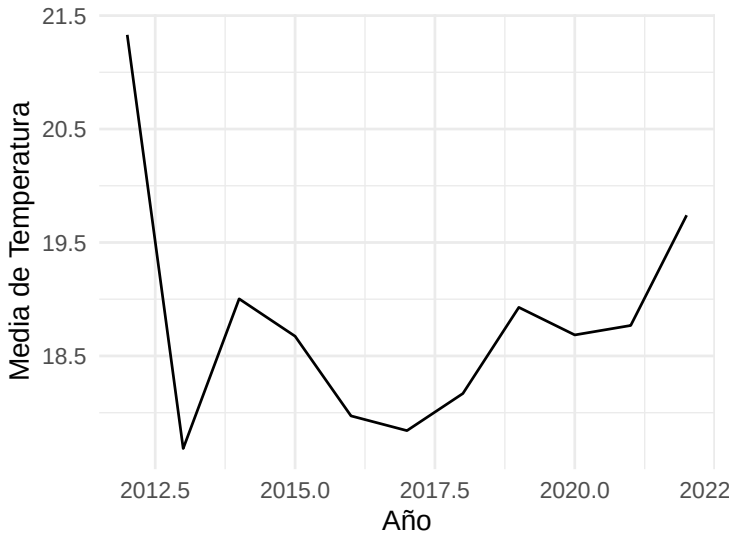

Figure 17

4.4. Evolución de la temperatura a lo largo de los años

En la Figura 17 se puede observar el aumento generalizado de la temperatura a lo largo de los años. Es curioso que, pese a que en el mundo existe un aumento generalizado de la temperatura, en Valencia estuvo bajando esta tendencia y no es hasta el año 2017 que comenzaron a subir de nuevo. Pese a lo mencionado, bastante visible el hecho de que nos encontramos en una tendencia creciente.

4.5. Cómo se ha comportado la contaminación a lo largo de los años

Estudiaremos el comportamiento de todos los gases contaminantes en todas las estaciones. Gracias a superponer todas estas evoluciones en un mismo gráfico, podemos tomar perspectiva y comparar el comportamiento entre los diferentes gases.

Esta vez, en la Figrua 18 podemos que en general las medidas de estos se han mantenido bastante estables o decrecientes en el tiempo. Con las gráficas de la figura **??** podemos contrastar la información que nos otorga esta gráfica. Por ejemplo:

- El CO sí parece decrecer en el último año. Sin embargo, por los outliers y el rango de valores que se maneja, en el gráfico presente puede no apreciarse esta evolución.

- Por ejemplo, en evoluciones como la del NO2 podemos reforzar la fiabilidad observando cómo la tendencia en su representación de boxplot es similar. Incluso parece haber consistencia en los datos que la regla boxplot marca como outliers.
- En el gas SO2, observamos una tendencia estable a lo largo de los años. No obstante, comparando con su contraparte representada en boxplot ganamos la información de que el gas ha ido ganando estabilidad, habiendo cada vez menos outliers con el paso del tiempo.

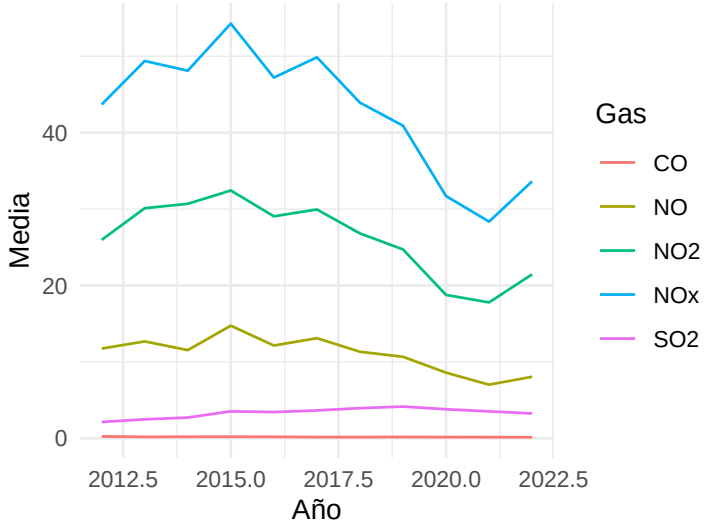

Figure 18

5. Conclusion

En definitiva, a lo largo de este análisis hemos utilizado métodos de exploración visual para ayudarnos a comprender la estructura interna de nuestros datos. También se ha realizado un análisis univariante de las variables obteniendo estadísticos que nos ayudaron averiguar sus unidades de medida, las cuales eran previamente desconocidas. También se han determinado relaciones entre nuestras variables en el análisis bivariante, que pueden servir para plantear estudios futuros. Además, utilizando la información obtenida durante el análisis se han descartado algunas variables y se han imputado los valores anómalos del resto. Finalmente, se han respondido algunas preguntas que surgieron durante el proceso de análisis utilizando un dataset sin valores anómalos.

6. Anexo 1

Table 1. Sumario variables

variable	type	levels	topLevel	topCount	topFrec	missFrec
Id	numeric	43388	1	1	0	0
Fecha	Date	6940	2022-01-01	12	0	0
Dia_de_la_semana	factor	7	Sabado	6205	0.143	0
Dia_del_mes	factor	31	16	1426	0.033	0
Estacion	factor	13	Pista Silla	6940	0.16	0
PM1	numeric	179	4	978	0.023	0.748
PM2.5	numeric	221	9	1567	0.036	0.463
PM10	numeric	479	13	1104	0.025	0.336
NO	numeric	178	2	3890	0.09	0.23
NO2	numeric	120	16	961	0.022	0.23
NOx	numeric	338	21	647	0.015	0.23
O3	numeric	116	56	688	0.016	0.241
SO2	numeric	22	3	15273	0.352	0.237
CO	numeric	18	0.1	7358	0.17	0.548
NH3	numeric	21	5	469	0.011	0.941
C7H8	numeric	235	1	163	0.004	0.908
C6H6	numeric	76	1	370	0.009	0.902
Ruido	numeric	67	62	1135	0.026	0.751
C8H10	numeric	182	1	240	0.006	0.901
Vel_viento	numeric	117	0.9	972	0.022	0.535
Dir_viento	numeric	362	264	177	0.004	0.536
Temperatura	numeric	282	16.5	146	0.003	0.637
Humedad_rel	numeric	82	71	521	0.012	0.644
Presion	numeric	64	1002	850	0.02	0.634
Radiacion_solar	numeric	401	139	104	0.002	0.639
Precipitacion	numeric	209	0	9545	0.22	0.73
Max_vel_viento	numeric	397	3.5	335	0.008	0.724
As	numeric	50	0.28	288	0.007	0.978
Ni	numeric	405	0.75	11	0	0.978
Cd	numeric	64	0.03	124	0.003	0.978
Pb	numeric	6	0.01	929	0.021	0.978
B(a)p	logical	2	FALSE	1	0	1
Fecha_creacion	Date	2	2022-12-04	39008	0.899	0
Fecha_baja	Date	1	NA	0	0	1

Table 2. Sumario de variables numericas

	min	Q1.25%	median	mean	dt	Q3.75%	max	n	IQR
NO	0.0	3.0	6.0	10.83	13.36	13.0	152.0	25107	10.0
NO2	0.0	14.0	22.0	25.58	15.42	35.0	104.0	25105	21.0
NOx	0.0	18.0	32.0	41.93	33.85	55.0	331.0	25107	37.0
SO2	0.0	3.0	3.0	3.40	1.43	4.0	24.0	24064	1.0
CO	0.0	0.1	0.1	0.16	0.09	0.2	0.8	12261	0.1
O3	3.0	38.0	53.0	50.83	19.15	64.0	122.0	23958	26.0
Temperatura	4.4	13.9	18.5	18.86	5.75	24.1	34.2	12950	10.2
Vel_viento	0.1	1.0	1.6	2.01	1.56	2.4	15.2	16261	1.4
Dir_viento	0.0	67.0	220.0	179.31	110.28	274.0	360.0	16226	207.0
Humedad_rel	14.0	55.0	67.0	65.81	14.51	76.0	100.0	12652	21.0
Presion	968.0	1002.0	1007.0	1007.73	8.56	1014.0	1042.0	13037	12.0
Radiacion_solar	-7.0	126.0	200.0	200.75	94.03	281.0	1025.0	12841	155.0
Precipitacion	0.0	0.0	0.0	1.26	8.93	0.0	629.0	11725	0.0
Ruido	6.0	55.0	59.0	58.08	6.77	61.0	250.0	5157	6.0