

PRA 2 - Tipología y ciclo de vida de los datos aula 1

Alejandro Zarza Roa

2022-05-18

```
library(car)

## Loading required package: carData

data_cities <- read.csv("data/datos_agregados_ciudades.csv")
summary(data_cities)

##      Municipio      Provincia      Comunidad.Autónoma Habitantes..2022.
## Length:101      Length:101      Length:101      Min.   : 75877
## Class :character Class :character Class :character 1st Qu.: 89039
## Mode  :character Mode  :character Mode  :character Median : 130577
##                                     Mean  : 219018
##                                     3rd Qu.: 212730
##                                     Max.   :3280782
## Carril.Bici      Precios.alquiler  Distancia.a.Madrid..km.
## Length:101      Length:101      Min.   : 0.0
## Class :character Class :character 1st Qu.: 319.0
## Mode  :character Mode  :character Median : 466.0
##                                     Mean  : 485.9
##                                     3rd Qu.: 614.0
##                                     Max.   :2029.0
## Tiempo.medio..de.viaje.a.Madrid..en.coche
## Length:101
## Class :character
## Mode  :character
##
##
## Tiempo.medio.de.viaje.a.Madrid.en.tren.o.autobús Distancia.a.Barcelona..km.
## Length:101      Min.   : 0.0
## Class :character 1st Qu.: 527.0
## Mode  :character Median : 638.0
##                                     Mean  : 732.9
##                                     3rd Qu.: 995.0
##                                     Max.   :2563.0
## Tiempo.medio..de.viaje.a.Barcelona..en.coche
## Length:101
## Class :character
## Mode  :character
##
##
## Tiempo.medio.de.viaje.a.Barcelona.en.tren.o.autobús
## Length:101
## Class :character
## Mode  :character
##
```

```

##
##
## Aeropuerto.Internacional.más.cercano
## Length:101
## Class :character
## Mode :character
##
##
##
## Número.de.conexiones.del.aeropuerto.internacional.más.cercano
## Min. : 3.00
## 1st Qu.: 16.00
## Median : 70.00
## Mean : 86.68
## 3rd Qu.:170.00
## Max. :174.00
## Distancia.al.aeropuerto.internacional.más.cercano..km.
## Length:101
## Class :character
## Mode :character
##
##
##
## Temperatura.máxima.media..°C. Temperatura.mínima.media..°C.
## Min. :22.00 Min. :-1.000
## 1st Qu.:28.00 1st Qu.: 1.000
## Median :30.00 Median : 5.000
## Mean :29.89 Mean : 4.465
## 3rd Qu.:33.00 3rd Qu.: 7.000
## Max. :36.00 Max. :16.000
## Número.de.visitantes.internacionales.al.año Precio.medio.vivienda..EUR.m2.
## Mode:logical Length:101
## NA's:101 Class :character
## Mode :character
##
##
##
## Patrimonio.de.la.Humanidad X..días.calidad.del.aire.Buena
## Length:101 Length:101
## Class :character Class :character
## Mode :character Mode :character
##
##
##
## X..días.calidad.del.aire.Razonablemente.Buena X..días.calidad.del.aire.Regular
## Length:101 Length:101
## Class :character Class :character
## Mode :character Mode :character
##
##
##
## X..días.calidad.del.aire.Desfavorable
## Length:101
## Class :character
## Mode :character

```

```
##
##
##
## X..días.calidad.del.aire.Muy.Desfavorable
## Length:101
## Class :character
## Mode :character
##
##
##
## X..días.calidad.del.aire.Extremadamente.Desfavorable
## Length:101
## Class :character
## Mode :character
##
##
##
## Ponderación.calidad.del.aire Horas.de.Sol.medias.anuales
## Length:101 Length:101
## Class :character Class :character
## Mode :character Mode :character
##
##
##
```

Descripción del dataset

¿Por qué es importante y qué pregunta/problema pretende responder?

El dataset elegido para el desarrollo de esta práctica es “Ciudades para teletrabajar en España”, se puede consultar en

<https://www.kaggle.com/datasets/amerono/ciudades-para-teletrabajar-en-espaa>. Este Dataset contiene información sobre diferentes ciudades de España y diversas variables relacionadas con su idoneidad para el teletrabajo. Estas variables resultan útiles para aquellos que se plantean trabajar de forma remota en España.

Los campos que contiene este dataset son:

- Nombre del municipio (municipios españoles de más de 75.000 habitantes)
- Provincia
- Comunidad Autónoma
- Población (fuente de datos: INE, 2022)
- Kilómetros de carril bici en el municipio por cada 100.000 habitantes (fuentes de datos: El País, La Sexta, ayuntamientos)
- Precios de viviendas en alquiler por metro cuadrado (fuente: Idealista, marzo 2023)
- Distancia a Madrid en km (fuente: Google Maps)

- Tiempo medio de viaje en coche a Madrid (fuente: Google Maps)
- Tiempo medio de viaje en tren o autobús a Madrid (fuente: Google Maps)
- Distancia a Barcelona en km (fuente: Google Maps)
- Tiempo medio de viaje en coche a Barcelona (fuente: Google Maps)
- Tiempo medio de viaje en tren o autobús a Barcelona (fuente: Google Maps)
- Aeropuerto internacional más cercano (fuentes: AENA, Google Maps)
- Número de conexiones del aeropuerto internacional más cercano (fuente: AENA)
- Distancia en km al aeropuerto internacional más cercano (fuente: Google Maps)
- Temperatura máxima media en °C (fuente: Weatherspark)
- Temperatura mínima media en °C (fuente: Weatherspark)
- Número de visitantes internacionales (turistas) al año (pendiente de completar)
- Precio medio de venta de viviendas por metro cuadrado (fuente: Idealista, marzo 2023)
- Patrimonio de la Humanidad: Si dispone o no de uno o varios Patrimonios (fuente: Wikipedia)
- Porcentaje de Días con calidad de aire “Buena” (fuente: Ministerio para la Transición Ecológica y Reto Demográfico)
- Porcentaje de Días con calidad de aire “Razonablemente Buena” (fuente: Ministerio para la Transición Ecológica y Reto Demográfico)
- Porcentaje de Días con calidad de aire “Regular” (fuente: Ministerio para la Transición Ecológica y Reto Demográfico)
- Porcentaje de Días con calidad de aire “Desfavorable” (fuente: Ministerio para la Transición Ecológica y Reto Demográfico)
- Porcentaje de Días con calidad de aire “Muy Desfavorable” (fuente: Ministerio para la Transición Ecológica y Reto Demográfico)
- Porcentaje de Días con calidad de aire “Extremadamente Desfavorable” (fuente: Ministerio para la Transición Ecológica y Reto Demográfico)
- Ponderación calidad del aire: Cálculo propio a partir de los 6 valores anteriores Horas de sol medias anuales (fuente: AEMET)

Recopilando datos sobre aspectos relevantes como la calidad de vida, la conectividad a internet, el costo de vida, la seguridad, el acceso a servicios y comodidades, entre otros, el objetivo principal de este conjunto de datos es identificar las ciudades en España más propicias para el teletrabajo. Mediante el análisis de esta información, los teletrabajadores potenciales pueden tomar decisiones informadas sobre dónde establecerse o realizar viajes de trabajo, teniendo en cuenta factores clave

que impactan en su experiencia de trabajo remoto. En resumen, este dataset busca proporcionar una guía sólida para aquellos que desean encontrar entornos ideales para el teletrabajo en España.

El dataset “Ciudades para teletrabajar en España” no solo ofrece información relevante para la elección de ciudades adecuadas para el teletrabajo, sino que también desempeña un papel fundamental en la comprensión de las necesidades y demandas de los teletrabajadores. Esto permite a las autoridades locales, empresas y profesionales adaptar sus políticas y servicios para satisfacer estas necesidades de manera efectiva. Además, el dataset proporciona una base sólida para realizar análisis exhaustivos del impacto económico y social del teletrabajo en diferentes regiones de España, lo que contribuye a una toma de decisiones más informada en torno al fomento y desarrollo del teletrabajo en el país.

Integración y selección de los datos de interés a analizar

En este punto, se realiza un descarte de las variables que no sean de interés

- Tiempo medio de viaje en coche a Madrid (fuente: Google Maps)
- Tiempo medio de viaje en tren o autobús a Madrid (fuente: Google Maps)
- Tiempo medio de viaje en coche a Barcelona (fuente: Google Maps)
- Tiempo medio de viaje en tren o autobús a Barcelona (fuente: Google Maps)

Se descartan estas cuatro ya que son variables que se pueden deducir aproximadamente en función de la distancia en kilómetros, ya que por lo general las comunicaciones son similares.

Debido a la ausencia de valores se descarta también el Número de visitantes internacionales (turistas) al año. Por último, también eliminaremos Patrimonio de la Humanidad, debido a que no resulta de interés.

```
columnas_a_borrar <- c("Tiempo.medio.de.viaje.a.Madrid.en.coche",  
"Tiempo.medio.de.viaje.a.Madrid.en.tren.o.autobús",  
"Tiempo.medio.de.viaje.a.Barcelona.en.coche",  
"Tiempo.medio.de.viaje.a.Barcelona.en.tren.o.autobús",  
"Número.de.visitantes.internacionales.al.año", "Patrimonio.de.la.Humanidad")  
  
data_cities <- data_cities[, -which(names(data_cities) %in% columnas_a_borrar)]
```

Limpieza de los datos

Gestión de ceros o elementos vacíos

```
# Verificar si hay valores nulos por columna en el dataset  
nulos <- colSums(data_cities == "N/D", na.rm = TRUE)  
  
# Imprimir los resultados  
print(nulos)  
  
## Municipio  
## 0  
## Provincia
```

##	0
##	Comunidad.Autónoma
##	0
##	Habitantes..2022.
##	0
##	Carril.Bici
##	24
##	Precios.alquiler
##	1
##	Distancia.a.Madrid..km.
##	0
##	Distancia.a.Barcelona..km.
##	0
##	Aeropuerto.Internacional.más.cercano
##	0
##	Número.de.conexiones.del.aeropuerto.internacional.más.cercano
##	0
##	Distancia.al.aeropuerto.internacional.más.cercano..km.
##	0
##	Temperatura.máxima.media..°C.
##	0
##	Temperatura.mínima.media..°C.
##	0
##	Precio.medio.vivienda..EUR.m2.
##	56
##	X..días.calidad.del.aire.Buena
##	14
##	X..días.calidad.del.aire.Razonablemente.Buena
##	14
##	X..días.calidad.del.aire.Regular
##	14
##	X..días.calidad.del.aire.Desfavorable
##	14
##	X..días.calidad.del.aire.Muy.Desfavorable
##	14
##	X..días.calidad.del.aire.Extremadamente.Desfavorable
##	14
##	Ponderación.calidad.del.aire
##	14
##	Horas.de.Sol.medias.anuales
##	3

Las columnas que tienen valores N/D son Carril.Bici, Precios.alquiler, Precio.medio.vivienda.EUR.m2., X.días.calidad.del.aire.Buena, X.días.calidad.del.aire.Razonablemente.Buena, X.días.calidad.del.aire.Regular,X.días.calidad.del.aire.Desfavorable,X.días.calidad.del.aire.Muy.Desfavorable,X.días.calidad.del.aire.Extremadamente.Desfavorable,Ponderación.calidad.del.aire, Horas.de.Sol.medias.anuales.

Como hemos visto al inicio, muchos de los valores que deberían ser numéricos están puestos de tipo character. Por lo tanto, antes de gestionar los valores extremos vamos a transformar dichas variables.

```
data$cities$Carril.Bici <- as.numeric(data$cities$Carril.Bici)
```

```

## Warning: NAs introducidos por coerción

data_cities$Precios.alquiler <- as.numeric(data_cities$Precios.alquiler)

## Warning: NAs introducidos por coerción

data_cities$Distancia.al.aeropuerto.internacional.más.cercano..km <-
as.numeric(data_cities$Distancia.al.aeropuerto.internacional.más.cercano..km)

## Warning: NAs introducidos por coerción

data_cities$Temperatura.mínima.media..°C. <-
as.numeric(data_cities$Temperatura.mínima.media..°C.)
data_cities$Precio.medio.vivienda..EUR.m2. <-
as.numeric(data_cities$Precio.medio.vivienda..EUR.m2.)

## Warning: NAs introducidos por coerción

data_cities$X..días.calidad.del.aire.Buena <-
as.numeric(data_cities$X..días.calidad.del.aire.Buena)

## Warning: NAs introducidos por coerción

data_cities$Distancia.al.aeropuerto.internacional.más.cercano..km. <-
as.numeric(data_cities$Distancia.al.aeropuerto.internacional.más.cercano..km.)

## Warning: NAs introducidos por coerción

data_cities$X..días.calidad.del.aire.Razonablemente.Buena <-
as.numeric(data_cities$X..días.calidad.del.aire.Razonablemente.Buena)

## Warning: NAs introducidos por coerción

data_cities$X..días.calidad.del.aire.Regular <-
as.numeric(data_cities$X..días.calidad.del.aire.Regular)

## Warning: NAs introducidos por coerción

data_cities$X..días.calidad.del.aire.Desfavorable <-
as.numeric(data_cities$X..días.calidad.del.aire.Desfavorable)

## Warning: NAs introducidos por coerción

data_cities$X..días.calidad.del.aire.Muy.Desfavorable <-
as.numeric(data_cities$X..días.calidad.del.aire.Muy.Desfavorable)

## Warning: NAs introducidos por coerción

data_cities$X..días.calidad.del.aire.Extremadamente.Desfavorable <-
as.numeric(data_cities$X..días.calidad.del.aire.Extremadamente.Desfavorable)

## Warning: NAs introducidos por coerción

data_cities$Ponderación.calidad.del.aire <-
as.numeric(data_cities$Ponderación.calidad.del.aire)

## Warning: NAs introducidos por coerción

```

```
data_cities$Horas.de.Sol.medias.anuales <- as.numeric(gsub("\\.", "",
data_cities$Horas.de.Sol.medias.anuales))
```

Warning: NAs introducidos por coerción

Debido a la importancia que le voy a dar en este estudio, voy a descartar las ciudades de las que no tenemos métricas acerca de la calidad del aire.

```
data_cities <- subset(data_cities, !Ponderación.calidad.del.aire %in% "N/D")
```

Cambio las ciudades que no tienen datos de los kilómetros de carril bici por cero.

```
data_cities$Carril.Bici <- ifelse(data_cities$Carril.Bici == "N/D", 0,
data_cities$Carril.Bici)
```

Ahora únicamente quedaria gestionar los valores de los precios de la vivienda, del alquiler y las horas de sol medias anuales. Para ello, voy a hacer una media entre los valores de la misma Provincia.

```
data_cities$Precios.alquiler <- with(data_cities, ifelse(Precios.alquiler == "N/D",
ave(as.numeric(Precios.alquiler),
Provincia,
FUN = function(x) mean(x, na.rm =
TRUE)),
Precios.alquiler))
```

```
data_cities$Precio.medio.vivienda..EUR.m2. <- as.numeric(gsub("\\.", "",
data_cities$Precio.medio.vivienda..EUR.m2.))
```

```
data_cities$Precio.medio.vivienda..EUR.m2. <- with(data_cities,
ifelse(Precio.medio.vivienda..EUR.m2. == "N/D",
ave(as.numeric(Precio.medio.vivienda..EUR.m2.),
Provincia,
FUN = function(x) mean(x, na.rm =
TRUE)),
ifelse(as.numeric(Precio.medio.vivienda..EUR.m2.) < 100,
ave(as.numeric(Precio.medio.vivienda..EUR.m2.),
Provincia,
FUN = function(x) mean(x[x > 100],
na.rm = TRUE)),
Precio.medio.vivienda..EUR.m2.)))
```

```
data_cities$Horas.de.Sol.medias.anuales <- with(data_cities,
ifelse(Horas.de.Sol.medias.anuales == "N/D",
ave(as.numeric(Horas.de.Sol.medias.anuales),
Provincia,
FUN = function(x) mean(x[x > 100],
na.rm=TRUE)),
```



```

ifelse(as.numeric(Horas.de.Sol.medias.anuales) < 100,
ave(as.numeric(Horas.de.Sol.medias.anuales),
Provincia,
FUN = function(x) mean(x[x > 100],
na.rm = TRUE)),
Horas.de.Sol.medias.anuales)))

```

```

nulos <- colSums(data_cities == "N/D", na.rm = TRUE)
print(nulos)

```

```

## Municipio
## 0
## Provincia
## 0
## Comunidad.Autónoma
## 0
## Habitantes..2022.
## 0
## Carril.Bici
## 0
## Precios.alquiler
## 0
## Distancia.a.Madrid..km.
## 0
## Distancia.a.Barcelona..km.
## 0
## Aeropuerto.Internacional.más.cercano
## 0
## Número.de.conexiones.del.aeropuerto.internacional.más.cercano
## 0
## Distancia.al.aeropuerto.internacional.más.cercano..km.
## 0
## Temperatura.máxima.media..°C.
## 0
## Temperatura.mínima.media..°C.
## 0
## Precio.medio.vivienda..EUR.m2.
## 0
## X..días.calidad.del.aire.Buena
## 0
## X..días.calidad.del.aire.Razonablemente.Buena
## 0
## X..días.calidad.del.aire.Regular
## 0
## X..días.calidad.del.aire.Desfavorable
## 0
## X..días.calidad.del.aire.Muy.Desfavorable
## 0
## X..días.calidad.del.aire.Extremadamente.Desfavorable
## 0
## Ponderación.calidad.del.aire
## 0

```

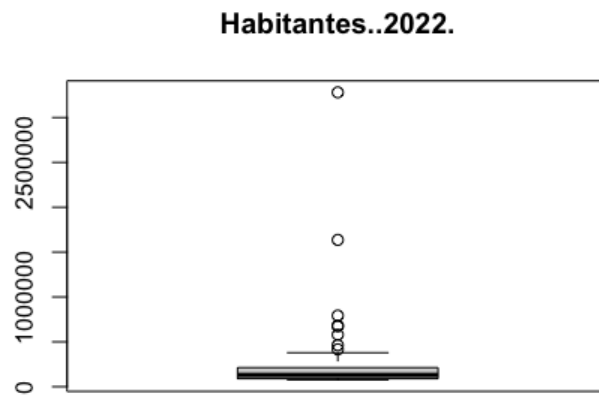
```
##                                     Horas.de.Sol.medias.anuales
##                                     0
##      Distancia.al.aeropuerto.internacional.más.cercano..km
##                                     0
```

Gestión de valores extremos

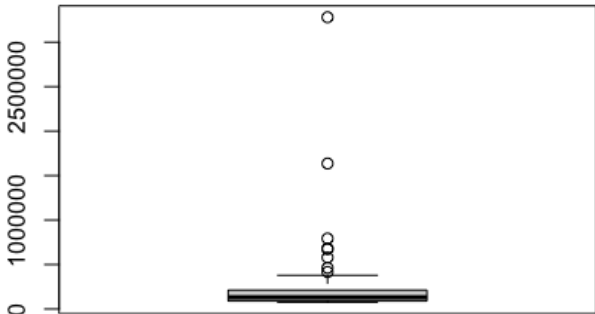
```
variables_numericas <- sapply(data_cities, is.numeric)

nombres_variables_numericas <- names(variables_numericas[variables_numericas])

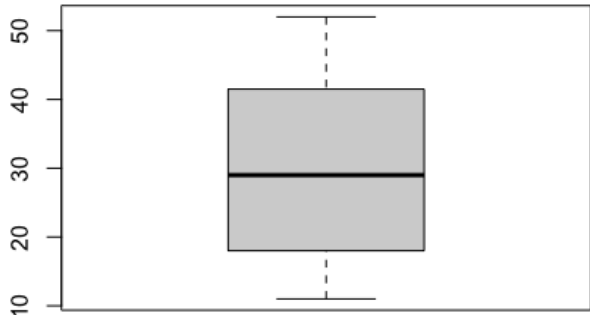
for (i in 1:length(nombres_variables_numericas)) {
  variable <- nombres_variables_numericas[i]
  boxplot(data_cities[,variable], main = variable, ylab = "")
}
```



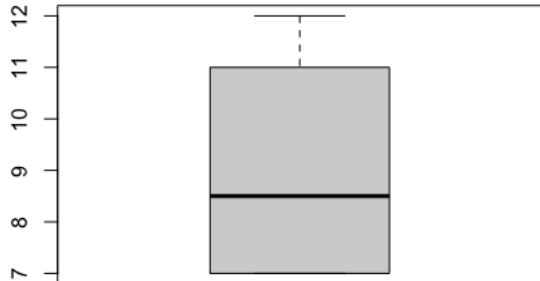
Habitantes..2022.



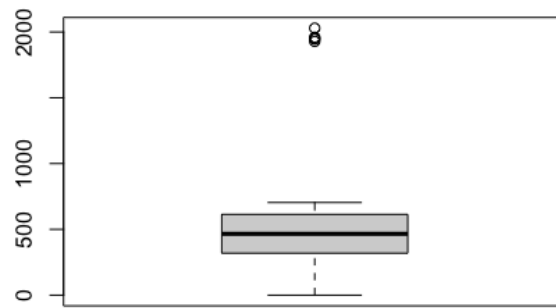
Carril.Bici



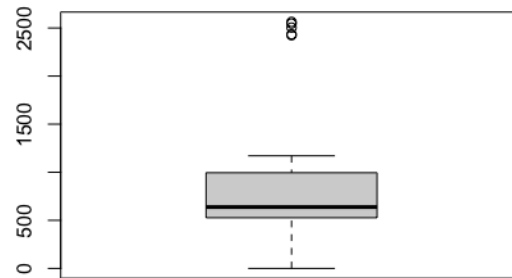
Precios.alquiler



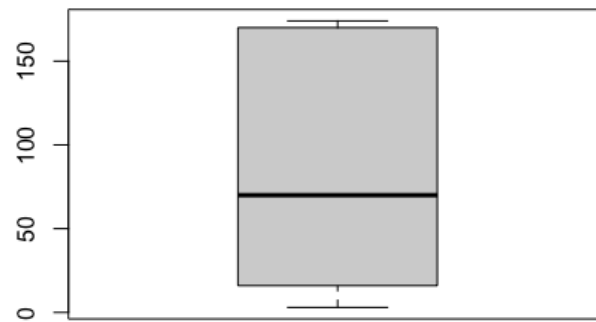
Distancia.a.Madrid..km.



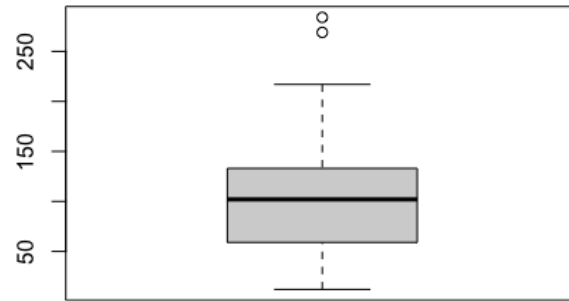
Distancia.a.Barcelona..km.



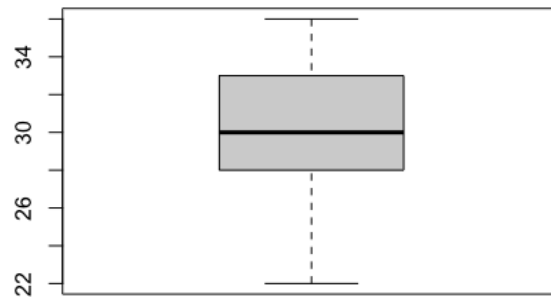
numero.de.conexiones.del.aeropuerto.internacional.más



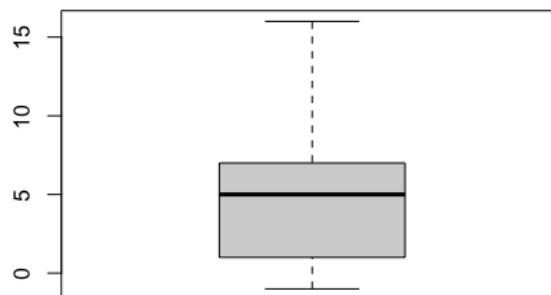
Distancia.al.aeropuerto.internacional.más.cercano..l



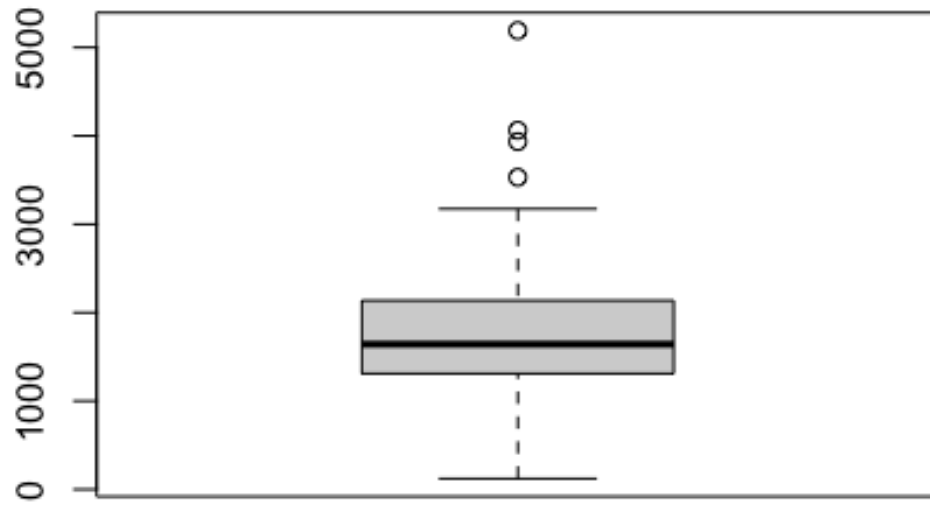
Temperatura.máxima.media..°C.



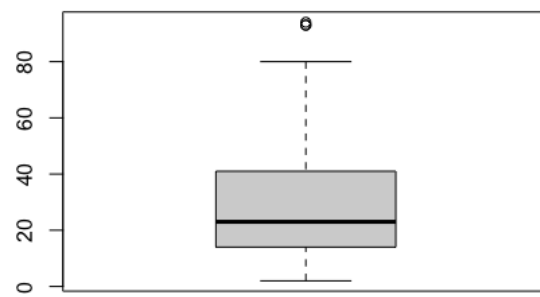
Temperatura.mínima.media..°C.



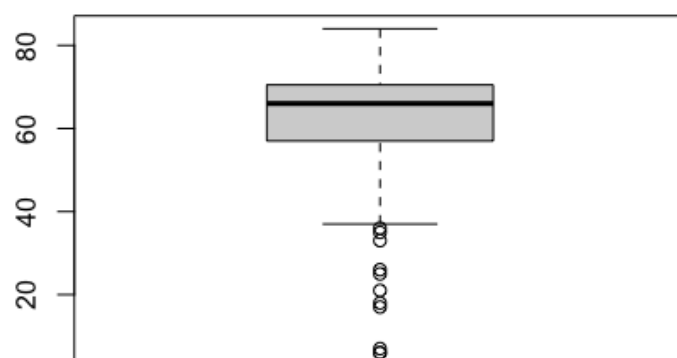
Precio.medio.vivienda..EUR.m2.



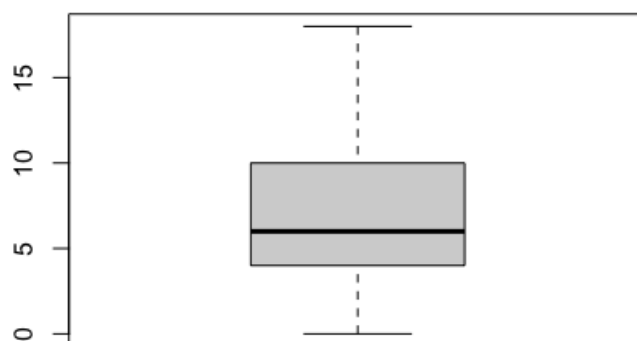
X..días.calidad.del.aire.Buena



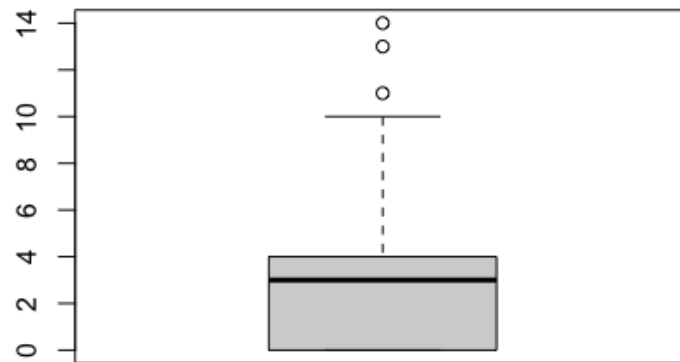
X..días.calidad.del.aire.Razonablemente.Buena



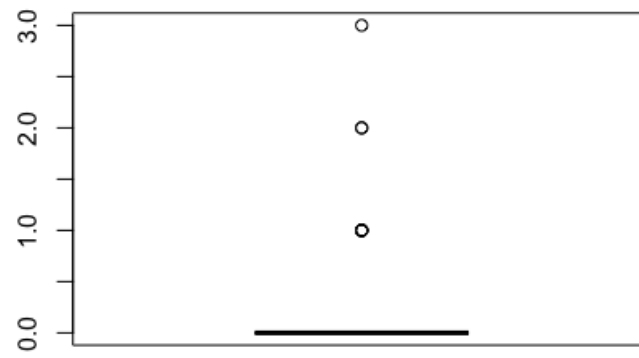
X..días.calidad.del.aire.Regular



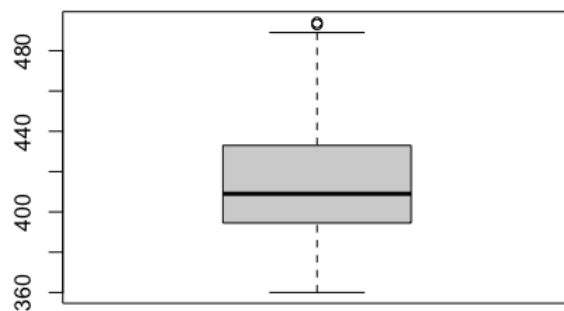
X..días.calidad.del.aire.Desfavorable



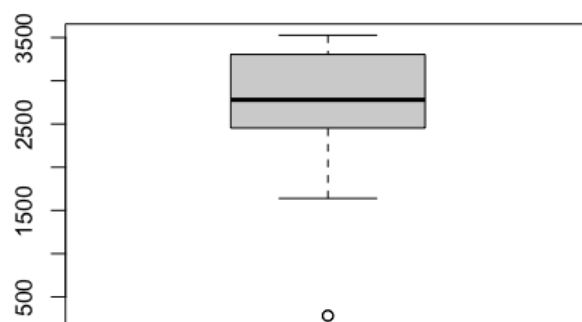
X..días.calidad.del.aire.Muy.Desfavorable



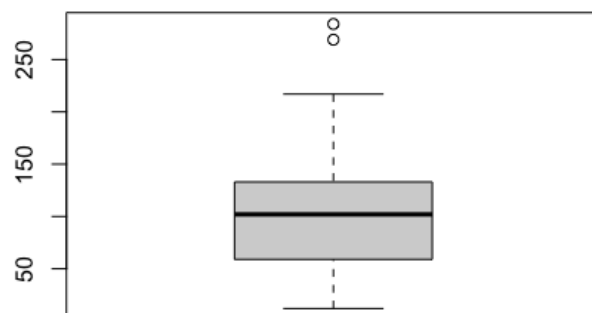
Ponderación.calidad.del.aire



Horas.de.Sol.medias.anuales



Distancia.al.aeropuerto.internacional.más.cercano..



Se observan algunos valores que se encuentran bastante fuera de los diagramas. Sin embargo esto se debe a diversos factores como la diferencia de habitantes en una población como Barcelona o Madrid con los municipios más pequeños, ya que se han considerado ciudades con un valor mínimo de 75000 habitantes.

Por otro lado también tenemos diferencias muy significativas en las distancias a Barcelona y a Madrid ya que también se incluyen poblaciones en Islas Canarias o Ceuta.

Análisis de los datos.

Las variables que se van a utilizar para este apartado de análisis son:

- Ponderación calidad del aire: Es importante considerarla, ya que la calidad del aire tiene un gran impacto en la salud y el bienestar de los trabajadores remotos.
- Precios alquiler: Esta variable es importante para analizar el costo de vida en cada ciudad.
- Precio medio vivienda (EUR/m2): Al igual que los precios de alquiler, el precio medio de la vivienda puede proporcionar información valiosa sobre el costo de vida en cada ciudad.
- Distancia al aeropuerto internacional más cercano (km): Un alto porcentaje de trabajadores remotos o bien son extranjeros, o precisan viajar de vez en cuando a la oficina.
- Carril Bici: Esta variable puede indicar la disponibilidad de infraestructura para el ciclismo y promover un medio de transporte saludable y respetuoso con el medio ambiente.
- Horas de Sol medias: Estas horas de sol pueden tener un impacto en el estado de ánimo y la productividad de los trabajadores remotos.

```
variables <- c("Ponderación.calidad.del.aire", "Precios.alquiler",
"Precio.medio.vivienda..EUR.m2.",
"Distancia.al.aeropuerto.internacional.más.cercano..km.",
"Carril.Bici", "Horas.de.Sol.medias.anuales")
data_analisis <- data_cities[, variables]

summary(data_analisis)

## Ponderación.calidad.del.aire Precios.alquiler Precio.medio.vivienda..EUR.m2.
## Min. :360.0 Min. : 7.00 Min. : 122
## 1st Qu.:394.5 1st Qu.: 7.25 1st Qu.:1310
## Median :409.0 Median : 8.50 Median :1642
## Mean :415.1 Mean : 9.00 Mean :1821
## 3rd Qu.:433.0 3rd Qu.:10.50 3rd Qu.:2136
## Max. :494.0 Max. :12.00 Max. :5191
## NA's :14 NA's :95 NA's :58
## Distancia.al.aeropuerto.internacional.más.cercano..km. Carril.Bici
## Min. : 12.0 Min. :11.00
## 1st Qu.: 59.0 1st Qu.:19.00
## Median :102.0 Median :29.00
## Mean :105.9 Mean :30.00
## 3rd Qu.:133.0 3rd Qu.:37.25
## Max. :284.0 Max. :52.00
## NA's :74 NA's :93
```

```
## Horas.de.Sol.medias.anuales
## Min.      : 282
## 1st Qu.:2453
## Median :2780
## Mean      :2784
## 3rd Qu.:3301
## Max.      :3527
## NA's      :11
```

Se hace la comprobación de la normalidad y homogeneidad de la varianza.

```
for (variable in data_analisis) {

  shapiro.test(variable)

}

# Convertir variables a numérico para tratar de corregir el error
data_analisis$Ponderación.calidad.del.aire <-
as.numeric(data_analisis$Ponderación.calidad.del.aire)
data_analisis$Precios.alquiler <- as.numeric(data_analisis$Precios.alquiler)
data_analisis$Precio.medio.vivienda..EUR.m2. <-
as.numeric(data_analisis$Precio.medio.vivienda..EUR.m2.)
data_analisis$Distancia.al.aeropuerto.internacional.más.cercano..km. <-
as.numeric(data_analisis$Distancia.al.aeropuerto.internacional.más.cercano..km.)
data_analisis$Carril.Bici <- as.numeric(data_analisis$Carril.Bici)
data_analisis$Horas.de.Sol.medias <- as.numeric(data_analisis$Horas.de.Sol.medias)

# Prueba de homogeneidad de varianza

# LeveneTest(data_analisis)

data_analisis$Carril.Bici <- as.numeric(data_analisis$Carril.Bici)
data_analisis$Precios.alquiler <- as.numeric(data_analisis$Precios.alquiler)

# Prueba de contraste de hipótesis (t-test)
# t.test(data_analisis$Precios.alquiler ~ data_analisis$Carril.Bici)

# Correlación de variables (coeficiente de correlación de Pearson)
cor(data_analisis$Ponderación.calidad.del.aire,
data_analisis$Precio.medio.vivienda..EUR.m2., method = "pearson")

## [1] NA

summary(data_analisis)

## Ponderación.calidad.del.aire Precios.alquiler Precio.medio.vivienda..EUR.m2.
## Min.      :360.0           Min.      : 7.00      Min.      : 122
## 1st Qu.:394.5           1st Qu.: 7.25      1st Qu.:1310
## Median :409.0           Median : 8.50      Median :1642
## Mean      :415.1          Mean      : 9.00      Mean      :1821
## 3rd Qu.:433.0           3rd Qu.:10.50     3rd Qu.:2136
## Max.      :494.0          Max.      :12.00     Max.      :5191
## NA's      :14            NA's      :95       NA's      :58
```

```
## Distancia.al.aeropuerto.internacional.más.cercano..km. Carril.Bici
## Min. : 12.0 Min. :11.00
## 1st Qu.: 59.0 1st Qu.:19.00
## Median :102.0 Median :29.00
## Mean :105.9 Mean :30.00
## 3rd Qu.:133.0 3rd Qu.:37.25
## Max. :284.0 Max. :52.00
## NA's :74 NA's :93
## Horas.de.Sol.medias.anuales Horas.de.Sol.medias
## Min. : 282 Min. : 282
## 1st Qu.:2453 1st Qu.:2453
## Median :2780 Median :2780
## Mean :2784 Mean :2784
## 3rd Qu.:3301 3rd Qu.:3301
## Max. :3527 Max. :3527
## NA's :11 NA's :11

# Regresión Lineal
# modelo_regresion <- lm(data_analisis$Precios.alquiler ~
data_analisis$Distancia.al.aeropuerto.internacional.más.cercano..km. +
data_analisis$Carril.Bici + data_analisis$Horas.de.Sol.medias, data = data_analisis)
# summary(modelo_regresion)

# exportamos el dataset resultante

write.csv(data_analisis, file = "data/dataset_limpio.csv", row.names = FALSE)
```

Resolución del problema

Debido a problemas con la presencia de valores NA en un dataset no puedo llegar a una conclusión final con estos datos. Por lo tanto voy a plantear diferentes puntos de vista y análisis que tenía en mente.

Lo he querido apoyar en tres factores fundamentales a la hora de elegir vivir en una ciudad determinada:

- Evaluación de la calidad del aire: Si la variable “Ponderación calidad del aire” muestra diferencias significativas entre las ciudades, se pueden identificar las ciudades con la mejor y peor calidad del aire para tomar decisiones relacionadas con la ubicación del teletrabajo.
- Análisis de precios: Comparando las variables “Precios alquiler” y “Precio medio vivienda (EUR/m²)”, se puede determinar si hay una correlación entre los precios de alquiler y los precios de venta de viviendas en las ciudades. Esto puede ayudar a tomar decisiones sobre el tipo de vivienda a considerar en cada ciudad.
- Accesibilidad y transporte: Evaluando la variable “Distancia al aeropuerto internacional más cercano (km)” y el dato de “Carril Bici”, se puede analizar la disponibilidad y facilidad de transporte en las ciudades. Esto puede ser relevante para aquellos que requieren viajar con frecuencia o desean utilizar la bicicleta como medio de transporte.