Práctica 2: Limpieza y análisis de datos

Daniel Cabello Vázquez

Enero 2020

Table of Contents

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder? 1
1.1 Objetivo 1
1.2 Descripción del Dataset 2
2. Integración y selección de los datos de interés a analizar. 3
3. Limpieza de los datos. 5
3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos? 5
3.2 Identificación y tratamiento de valores extremos. 7
4. Análisis de los datos y presentación de los resultados 13
4.1 Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar) 13
4.2 Comprobación de la normalidad y homogeneidad de la varianza 15
4.2.1 Test de normalidad de Kolmogorov-Smirnov 15
4.2.2 Homogeneidad de la varianza 16
4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos 17
4.3.1 Análisis de correlación 17
4.3.2 Contraste de hipótesis 19
4.3.3 Análisis de tendencia de la temperatura media mensual 1997-2015 24
6. Conclusiones 27

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

1.1 Objetivo

A partir de información de la estación meteorológica del aeropuerto de Barajas se nos plantea las siguientes cuestiones:

- ¿Qué factores climatológicos influyen más sobre la visibilidad y en qué sentido lo hace? Esta variable tiene gran importancia a la hora de regular el tráfico aéreo del aropuerto de Barajas.
- ¿Se ha producido un cambio estadísticamente significativo en la temperatura media de los últimos 5 años repecto al del periodo 1997-2001?
- ¿Cuál ha sido la tendencia en la temperatura media, máxima y mínima anual y estacional para el periodo comprendido entre 1997 y 2015?

1.2 Descripción del Dataset

Datos climatológicos diarios de la estación meteorológica del aeropuerto de Barajas recogidos entre 1997 y 2015. Comprende 23 variables y 6812 observaciones.

El conjunto de datos fue descargado de la web de Kaggle:
https://www.kaggle.com/juliansimon/weather_madrid_lemd_1997_2015.csv
Gathered web https://www.wunderground.com/ The Weather Company, LLC

Las variables son las siguientes:

- CET: Fecha de la observación meteorológica en formato de año-mes-día.
- Max TemperatureC: Temperatura máxima del aire registrada durante el día indicado, en grados Celsius.
- Mean TemperatureC: Temperatura media del aire registrada durante el día indicado, en grados Celsius.
- Min TemperatureC: Temperatura mínima del aire registrada durante el día indicado, en grados Celsius.
- Dew PointC: Punto de rocío, temperatura a la que empieza a condensarse el vapor de agua contenido en el aire. en grados Celsius.
- MeanDew PointC: Promedio del Punto de rocio, en grados celsius.
- Min DewpointC: Punto de rocío mínimo, en grados celsius.
- Max Humidity: Humedad relativa máxima durante el día indicado, en porcentaje.
- Mean Humidity: Humedad relativa media durante el día indicado, en porcentaje.
- Min Humidity: Humedad relativa mínima durante el día indicado, en porcentaje.
- Max Sea Level PressurehPa: Presión atmosférica máxima durante el día indicado a nivel de mar, en hPa.
- Mean Sea Level PressurehPa: Presión atmosférica media durante el día indicado al nivel de mar, en hPa.
- Min Sea Level PressurehPa: Presión atmosférica mínima durante el día indicado al nivel de mar, en hPa.
- Max VisibilityKm: Visibilidad máxima durante el día indicado, en km.
- Mean VisibilityKm: visibilidad media durante el día indicado, en Km.
- Min VisibilitykM: visibilidad mínimz durante el día indicado, en Km.
- Max Wind SpeedKm/h: velocidad máxima del viento, en Km/hora.

- Mean Wind SpeedKm/h: velocidad media del viento, en Km/hora.
- Max Gust SpeedKm/h: velocidad máxima de razhas, en Km/hora.
- Precipitationmm: Precipitación acumulada durante el día indicado, en mm.
- CloudCover: Nubosidad.
- Events: enetos meteorológico durante el día indicado.
- WindDirDegrees: Dirección del viento, en grado.

2. Integración y selección de los datos de interés a analizar.

Se cargan de datos del archivo csv en un dataframe (data_weather)

```
# Carga de datos
data_weather <- read.csv("DATA/weather_madrid_1997_2015.csv")

# Se visualizan los primeros registros
head(data_weather)
```

```
##        CET Max.TemperatureC Mean.TemperatureC Min.TemperatureC
Dew.PointC
## 1 1997-1-1               7                 4               2
5
## 2 1997-1-2               7                 3               0
6
## 3 1997-1-3               5                 3               2
5
## 4 1997-1-4               7                 3              -1
-2
## 5 1997-1-5               2                 0              -1
2
## 6 1997-1-6               7                 3               1
2
##   MeanDew.PointC Min.DewpointC Max.Humidity Mean.Humidity Min.Humidity
## 1              3             2          100            95           76
## 2              3             0          100            92           71
## 3              1            -1          100            85           70
## 4             -3            -4           86            63           49
## 5              0            -3          100            95           86
## 6             -1            -3          100            82           57
##   Max.Sea.Level.PressurehPa Mean.Sea.Level.PressurehPa
## 1                      1010                       1008
## 2                      1007                       1003
## 3                      1005                        999
## 4                      1012                       1010
## 5                      1012                       1008
## 6                      1014                       1010
##   Min.Sea.Level.PressurehPa Max.VisibilityKm Mean.VisibilityKm
Min.VisibilitykM
## 1                      1004               10                 9
```

```
4
## 2                      997               10                  9
4
## 3                      996               10                 10
7
## 4                     1005               10                 10
10
## 5                     1005               10                  5
1
## 6                     1008               10                 10
10
##   Max.Wind.SpeedKm.h Mean.Wind.SpeedKm.h Max.Gust.SpeedKm.h
Precipitationmm
## 1                 13                    6                  NA
0
## 2                 26                    8                  47
0
## 3                 27                   19                  NA
0
## 4                 27                   19                  40
0
## 5                 14                    6                  NA
0
## 6                 11                    5                  NA
0
##   CloudCover    Events WindDirDegrees
## 1          6                    229
## 2          5      Rain          143
## 3          6 Rain-Snow          256
## 4          2                    284
## 5          7      Snow            2
## 6          4                     64
```

Se revisa si los Tipos de datos asignados a cada variable es la adecuada:

```
sapply(data_weather,function(x) class(x))
```

```
##                       CET           Max.TemperatureC
##                  "factor"                  "integer"
##          Mean.TemperatureC           Min.TemperatureC
##                  "integer"                  "integer"
##                 Dew.PointC             MeanDew.PointC
##                  "integer"                  "integer"
##              Min.DewpointC               Max.Humidity
##                  "integer"                  "integer"
##              Mean.Humidity               Min.Humidity
##                  "integer"                  "integer"
##  Max.Sea.Level.PressurehPa Mean.Sea.Level.PressurehPa
##                  "integer"                  "integer"
##  Min.Sea.Level.PressurehPa           Max.VisibilityKm
##                  "integer"                  "integer"
```

```
##          Mean.VisibilityKm           Min.VisibilitykM
##                  "integer"                  "integer"
##         Max.Wind.SpeedKm.h        Mean.Wind.SpeedKm.h
##                  "integer"                  "integer"
##         Max.Gust.SpeedKm.h            Precipitationmm
##                  "integer"                  "numeric"
##                 CloudCover                     Events
##                  "integer"                   "factor"
##             WindDirDegrees
##                  "integer"
```

Los tipos de datos de algunas variables no son las adecuados, por lo que se cambiarán para facilitar los análisis posteriores.

- El tipo de dato del campo 'CET' es 'factor', pero interesa que sea tratado como fecha con el formato "año/mes/día".
- Las variables relacionadas con temperatura se los trata como integer pero le asignaremos el tipo numérico para poder trabajar con decimales.

```
data_weather$CET <- as.Date.factor(data_weather$CET, "%Y-%m-%d")

data_weather[,2:7] <- apply(data_weather[,2:7],2, function(x)
as.numeric(x))


class(data_weather$CET)
```

```
## [1] "Date"
```

3. Limpieza de los datos.

3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

En el dataset, en ningún caso, los ceros indican ausencia de valores sino que es en sí mismo un valor cuantitativo de la variable(temperatura, percipitación, punto de rocío…). Los valores nulos aparecen en blanco, sin dato alguno, identificándose y asignándose como 'NA' durante la carga del dataset.

Por otra parte, en el campo 'Precipitationmm' se han detectado anomalías en los valores de precipitación, haciéndolos poco fiables e invalidando su uso para los análisis. Normalmente el cero indicaría ausencia de precipitación, pero se ha evidenciado que existen largos periodos sin precipitación a pesar de que se producen eventos de lluvía (según el campo 'Events'). Por tanto, este campo será eliminada del dataset.

A continuación, se contabilizan el nº de nulos existentes en cada campos:

```
sapply(data_weather, function(x) sum(is.na(x)))
```

```
##                          CET           Max.TemperatureC
##                            0                          2
##          Mean.TemperatureC           Min.TemperatureC
##                            3                          2
##                  Dew.PointC             MeanDew.PointC
##                            2                          2
##              Min.DewpointC               Max.Humidity
##                            2                          2
##              Mean.Humidity               Min.Humidity
##                            2                          2
##  Max.Sea.Level.PressurehPa Mean.Sea.Level.PressurehPa
##                            0                          0
##  Min.Sea.Level.PressurehPa           Max.VisibilityKm
##                            0                        940
##          Mean.VisibilityKm           Min.VisibilitykM
##                          940                        940
##         Max.Wind.SpeedKm.h        Mean.Wind.SpeedKm.h
##                            0                          0
##         Max.Gust.SpeedKm.h             Precipitationmm
##                         3306                          0
##                 CloudCover                     Events
##                         1372                          0
##             WindDirDegrees
##                            0
```

Los campos con más valores nulos se eliminarán del dataset: los campos de visibilidad máxima y mínima con 940 registros vacíos, la nubosidad (cloudcover) con 1.372 registros vacíos y la velocidad máxima de las rachas (Max.Gust.SpeedKm.h) con 3.306 registros vacíos. La opción de eliminar los registros con valores vacíos no es una opción porque invalidaría los análisis estadísticos de la serie histórica y estos campos no son esenciales para los objetivos planteados. No se elimina el campo de visibiliad media (Mean.VisibilityKm) ya que se utilizará para hacer el análisis de correlación para ver que variable climática tiene más influencia en la visibilidad.

Se eliminan los campos:

```
library(dplyr)

# Eliminación de campos del dataset
data_weather <- select(data_weather, -Max.VisibilityKm, -
Min.VisibilitykM, -Max.Gust.SpeedKm.h, -CloudCover, -Precipitationmm)
```

Las variables relativas a la temperatura, punto de rocío y humedad presentan 2 ó 3 valores nulos cada uno. En estos casos se realiza una imputación de valores basada en los k vecinos más próximos (Knn - Imputation). Para esta imputación se utiliza la función KNN() del paquete VIM.

```
library(VIM)

# Imputación de valores
```

```
data_weather$Max.TemperatureC <- kNN(data_weather)$Max.TemperatureC
data_weather$Mean.TemperatureC <- kNN(data_weather)$Mean.TemperatureC
data_weather$Min.TemperatureC <- kNN(data_weather)$Min.TemperatureC
data_weather$Dew.PointC <- kNN(data_weather)$Dew.PointC
data_weather$MeanDew.PointC <- kNN(data_weather)$MeanDew.PointC
data_weather$Min.DewpointC <- kNN(data_weather)$Min.DewpointC
data_weather$Max.Humidity <- kNN(data_weather)$Max.Humidity
data_weather$Mean.Humidity <- kNN(data_weather)$Mean.Humidity
data_weather$Min.Humidity <- kNN(data_weather)$Min.Humidity

# Contabilización de valores nulo
sapply(data_weather, function(x) sum(is.na(x)))
```

```
##                          CET           Max.TemperatureC
##                            0                          0
##          Mean.TemperatureC           Min.TemperatureC
##                            0                          0
##                 Dew.PointC             MeanDew.PointC
##                            0                          0
##              Min.DewpointC               Max.Humidity
##                            0                          0
##              Mean.Humidity               Min.Humidity
##                            0                          0
##  Max.Sea.Level.PressurehPa Mean.Sea.Level.PressurehPa
##                            0                          0
##  Min.Sea.Level.PressurehPa          Mean.VisibilityKm
##                            0                        940
##         Max.Wind.SpeedKm.h        Mean.Wind.SpeedKm.h
##                            0                          0
##                     Events             WindDirDegrees
##                            0                          0
```

Ahora ya no hay valores nulos en ningún campo del dataset a excepción de 'Mean.VisibilityKm' por la razón comentada anteriormente.

3.2 Identificación y tratamiento de valores extremos.

Los valores mínimo y máximo y los cuartiles permiten ver en qué rango de valores se mueve cada variable, y, si estos, se encuentran dentro de lo aceptable para cada una de las variables. En los datos que se muestran abajo no se aprecia valores anómalos que se salga de lo posible. Las temperaturas, porcentajes de humedad, la presión del aire, la precipitación y la velocidad y dirección del viento se sitúan dentro del rango de lo aceptable.

```
# Obtención de los cuartiles y el valor máximo y mínimo de las variables cuantitativas
summary(data_weather[,c(-1,-16)])
```

```
##  Max.TemperatureC Mean.TemperatureC Min.TemperatureC    Dew.PointC
##  Min.   : 0.00    Min.   :-3.00     Min.   :-10.000   Min.   :-12.000
```

```
##  1st Qu.:13.00    1st Qu.: 8.00     1st Qu.:  3.000   1st Qu.:  5.000
##  Median :20.00    Median :14.00     Median :  9.000   Median :  8.000
##  Mean   :21.04    Mean   :14.66     Mean   :  8.641   Mean   :  8.121
##  3rd Qu.:29.00    3rd Qu.:21.00     3rd Qu.: 14.000   3rd Qu.: 12.000
##  Max.   :41.00    Max.   :32.00     Max.   : 28.000   Max.   : 20.000
##
##  MeanDew.PointC    Min.DewpointC      Max.Humidity    Mean.Humidity
##  Min.   :-15.000   Min.   :-22.000   Min.   : 16.00   Min.   : 15.00
##  1st Qu.:  2.000   1st Qu.: -2.000   1st Qu.: 68.00   1st Qu.: 41.00
##  Median :  6.000   Median :  2.000   Median : 87.00   Median : 59.00
##  Mean   :  4.977   Mean   :  1.451   Mean   : 81.14   Mean   : 57.97
##  3rd Qu.:  8.000   3rd Qu.:  5.000   3rd Qu.: 94.00   3rd Qu.: 74.00
##  Max.   : 16.000   Max.   : 14.000   Max.   :100.00   Max.   :100.00
##
##   Min.Humidity    Max.Sea.Level.PressurehPa Mean.Sea.Level.PressurehPa
##  Min.   :  4.00   Min.   : 994              Min.   : 986
##  1st Qu.: 19.00   1st Qu.:1017              1st Qu.:1014
##  Median : 32.00   Median :1020              Median :1018
##  Mean   : 34.73   Mean   :1021              Mean   :1018
##  3rd Qu.: 47.25   3rd Qu.:1024              3rd Qu.:1022
##  Max.   :100.00   Max.   :1047              Max.   :1043
##
##  Min.Sea.Level.PressurehPa Mean.VisibilityKm Max.Wind.SpeedKm.h
##  Min.   : 965              Min.   : 0.00     Min.   :  0.00
##  1st Qu.:1011              1st Qu.:10.00     1st Qu.: 14.00
##  Median :1015              Median :10.00     Median : 21.00
##  Mean   :1015              Mean   :11.72     Mean   : 21.95
##  3rd Qu.:1019              3rd Qu.:10.00     3rd Qu.: 27.00
##  Max.   :1041              Max.   :31.00     Max.   :182.00
##                            NA's   :940
##                Events     WindDirDegrees
##                   :5014   Min.   : -1.0
##  Rain             :1140   1st Qu.: 66.0
##  Rain-Thunderstorm: 247   Median :223.0
##  Fog              : 233   Mean   :197.2
##  Fog-Rain         :  69   3rd Qu.:299.0
##  Thunderstorm     :  45   Max.   :360.0
##  (Other)          :  64
```

Si se definen los valores extremos o atípicos (outliers) como aquellos valores que exceden 3 veces el rango intercuartílico (RIC) a partir del cuartil 1 (Q1) o del cuartil 3 (q3), es decir, caen fuera del intervalo Q1-3RIC - Q3+3RIC, entonces se tienen los siguientes valores extremos para cada campo:

```
# Valores extremos de cada campo
sapply(data_weather[,c(-1,-17)], function(x) boxplot.stats(x)$out[1:10])
```

```
##       Max.TemperatureC Mean.TemperatureC Min.TemperatureC Dew.PointC
##  [1,]               NA                NA               NA        -10
##  [2,]               NA                NA               NA         -8
```

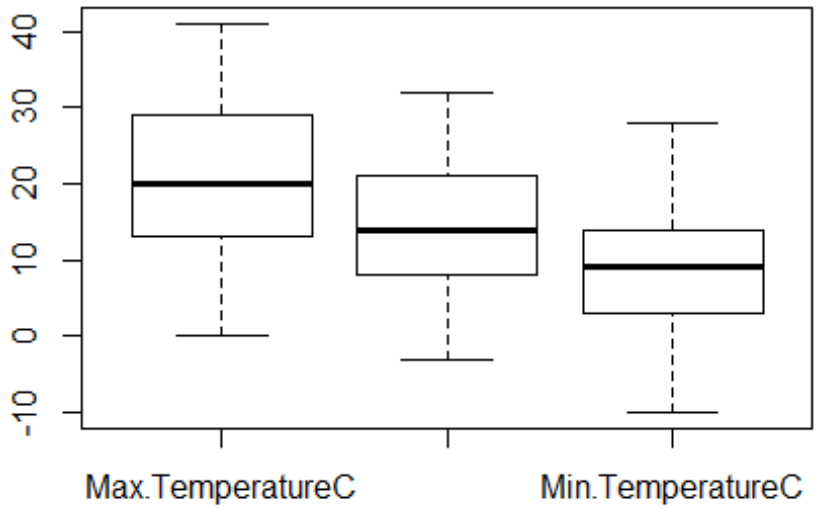
```
##  [3,]               NA                NA               NA         -6
##  [4,]               NA                NA               NA         -9
##  [5,]               NA                NA               NA         -8
##  [6,]               NA                NA               NA         -7
##  [7,]               NA                NA               NA         -8
##  [8,]               NA                NA               NA         -7
##  [9,]               NA                NA               NA         -6
## [10,]               NA                NA               NA         -6
##       MeanDew.PointC Min.DewpointC Max.Humidity Mean.Humidity
Min.Humidity
##  [1,]            -15           -22           16            NA
93
##  [2,]            -12           -18           26            NA
100
##  [3,]             -9           -13           28            NA
93
##  [4,]            -12           -18           26            NA
100
##  [5,]             -9           -13           28            NA
93
##  [6,]            -12           -13           NA            NA
93
##  [7,]            -11           -16           NA            NA
93
##  [8,]             -9           -13           NA            NA
93
##  [9,]            -11           -15           NA            NA
93
## [10,]            -11           -14           NA            NA
93
##       Max.Sea.Level.PressurehPa Mean.Sea.Level.PressurehPa
##  [1,]                      1005                       999
##  [2,]                      1035                       999
##  [3,]                      1035                      1001
##  [4,]                      1036                       992
##  [5,]                      1035                      1000
##  [6,]                      1036                       998
##  [7,]                      1003                      1000
##  [8,]                      1005                      1001
##  [9,]                      1005                      1036
## [10,]                      1003                      1036
##       Min.Sea.Level.PressurehPa Mean.VisibilityKm Max.Wind.SpeedKm.h
##  [1,]                       997                 9                48
##  [2,]                       996                 9                48
##  [3,]                      1032                 5                58
##  [4,]                      1032                 7                47
##  [5,]                       997                 8                48
##  [6,]                       993                 6                48
##  [7,]                       981                 9                47
##  [8,]                       998                 9                47
```

```
##  [9,]                      998                 7                  47
## [10,]                      997                 6                  55
##       Mean.Wind.SpeedKm.h WindDirDegrees
##  [1,]                  19             NA
##  [2,]                  19             NA
##  [3,]                  23             NA
##  [4,]                  19             NA
##  [5,]                  23             NA
##  [6,]                  29             NA
##  [7,]                  24             NA
##  [8,]                  24             NA
##  [9,]                  24             NA
## [10,]                  21             NA
```

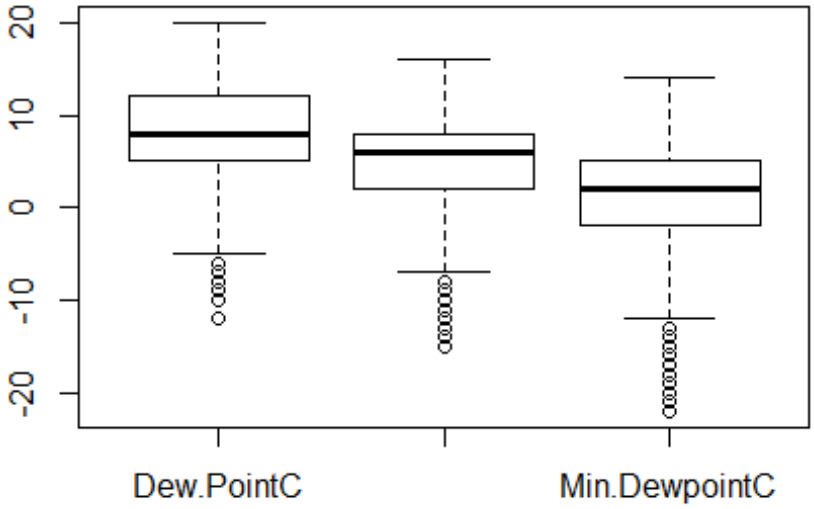
Puesto que estos valores extremos se encuadran dentro del rango posible de valores de cada atributo, no es necesario hacer un tratamiento previo de los mismos. En todo caso, se pueden aplicar análisis estadísticos robustos para minimizar o evitar efectos indeseados de estos valores extremos sobre los estadísticos (de tendencia central, dispersión…) y de su propagación en pruebas estadísticas.

El rango de valores y sus valores extremos se pueden representar de una forma más visual mediante diagramas de caja y bigotes (boxplot):

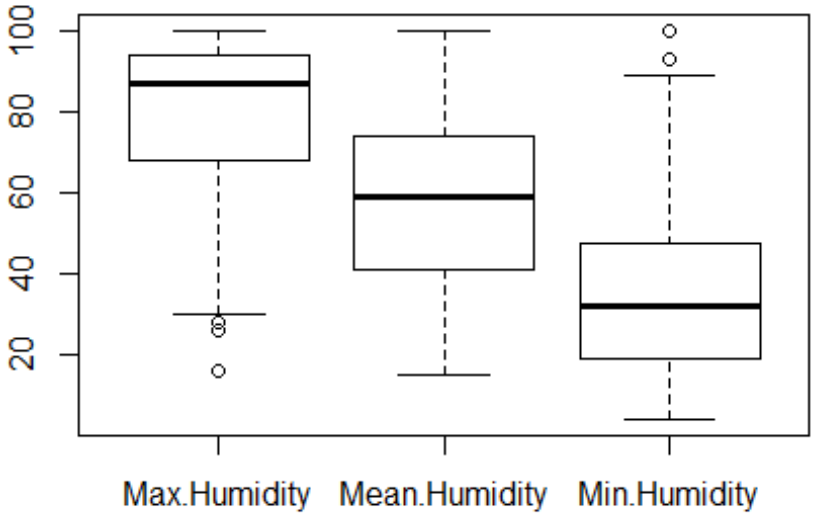
```
boxplot(data_weather[,2:4])
```



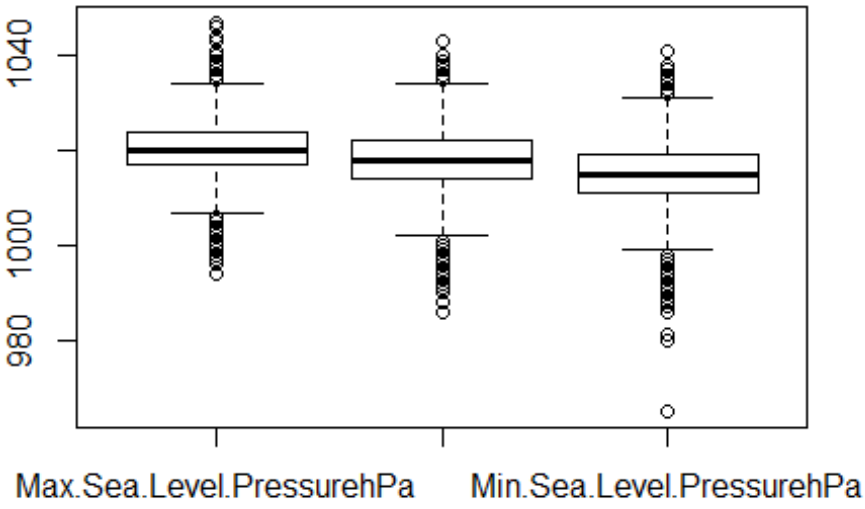

```
boxplot(data_weather[,5:7])
```



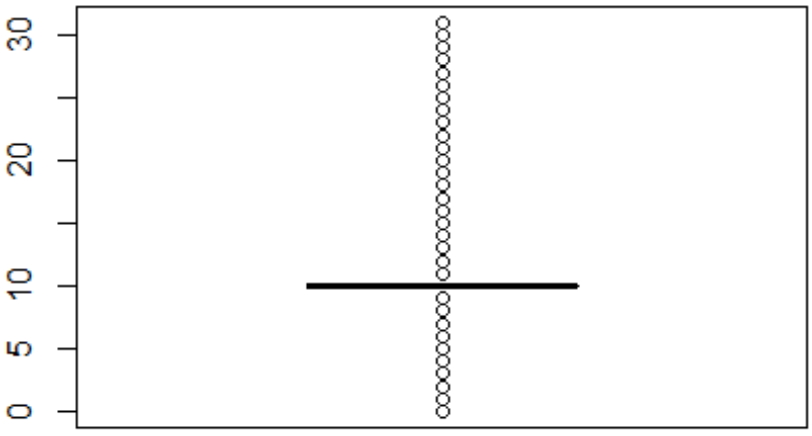

```
boxplot(data_weather[,8:10])
```



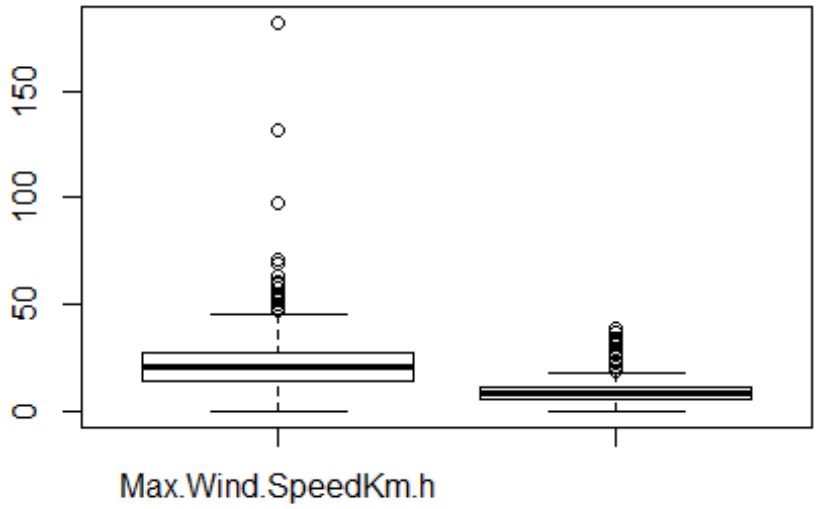

```
boxplot(data_weather[,11:13])
```




```
boxplot(data_weather[,14])
```




```
boxplot(data_weather[,15:16])
```



Una vez finalizada la carga, integración y limpieza de los datos se guarda el nuevo conjunto de datos en un nuevo archivo csv:

```
write.csv(data_weather, "DATA/weather_madrid_1997_2015_clean.csv")
```

4. Análisis de los datos y presentación de los resultados

4.1 Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Carga de los datos limpios

```
data_weather_clean <- data_weather
```

Análisis 1. Se compara si la temperatura media en el mes de enero en los primeros 5 años (1997-2001) del registro es significativamente mayor que la de los 5 últimos años (2011-2015). Se utilizará un test paramétrico de constraste de hipótesis para verificarlo. En la hipótesis nula no existe diferencia significativa en la temperatura media del mes de enero mientras que en la hipótesis alternativa si hay una diferencia significativa. Los datos a a seleccionar son los registros de temperatura media (Mean.TemperatureC) del mes de enero en los años 1999-2003 y 2011-2015.

```
data_ch <-data_weather_clean %>%
  select(CET, Mean.TemperatureC, Min.TemperatureC, Max.TemperatureC) %>%
  filter((CET >= "1997-01-01" & CET <= "2001-12-31") | CET >= "2011-01-
```

```
01")

data_ch <- mutate(data_ch, grupo = if_else(CET >= "1997-01-01" & CET <=
"2001-12-31",1, 2),
                  month = as.integer(format(CET, "%m")),
                  year = as.integer(format(CET, "%Y")))
```

Análisis 2. Análisis de correlación entre las variables de visibilidad media y las demás variables climáticas cuantitativas para ver cual es la que más influye. Solo se seleccionarán aquellos registros o filas en los que no existan nulos en el campo 'Mean.Visibilitykm'.

```
# Se elimina filas con nulos en el campo 'Mean.VisibilityKm'
data_cor <-
data_weather_clean[!is.na(data_weather_clean$Mean.VisibilityKm),]
```

Análisis 3. Análisis de tendencia de la temperatura media, mínima y máxima mensual

Se extraen el año y el mes del campo fecha (CET) y se pone en dos nuevos campos (year, month), luego se agrupan los datos de temperatura por año y mes obteniendo la media mensual de cada año. En el dataset original no existen registros en los meses de marzo y abril del 2000 por lo que no se han podido detectar antes mediante la búsqueda de valores vacíos. Para solucionarlo, se ha optado por añadir los registros correspondientes a esas fechas imputando el valor promedio calculado a partir los valores de los dos anteriores y los dos posteriores de cada mes (marzo y abril).

```
# Se agrupan los datos por año y mes calculando la media mensual
data_month <- data_weather_clean %>%
  mutate(month = as.integer(format(CET, "%m")), year =
as.integer(format(CET, "%Y"))) %>%
  group_by(year,month) %>%
  summarise(Mean.Temp_C = round(mean(Mean.TemperatureC),2),
            Min.Temp_C = round(mean(Min.TemperatureC),2),
            Max.Temp_C = round(mean(Max.TemperatureC),2))

# Se añaden los registros para febrero y marzo de 2000 no incluidos en el
dataset original
mar_2000 <- list(2000,3,10.71,5.10,16.83)
apr_2000 <- list(2000,4,11.85,5.96,18.24)
data_month <- rbind.data.frame(data_month, mar_2000, apr_2000)

# Se reordenan las filas por año y mes
data_month <- arrange(data_month, year, month)


data_year <- data_month %>%
  group_by(year) %>%
  summarise(Mean.Temp_C = round(mean(Mean.Temp_C),2),
            Min.Temp_C = round(mean(Min.Temp_C),2),
            Max.Temp_C = round(mean(Max.Temp_C),2))
```

```
data_year
```

```
## # A tibble: 19 x 4
##     year Mean.Temp_C Min.Temp_C Max.Temp_C
##    <dbl>       <dbl>      <dbl>      <dbl>
##  1  1997        15.6       9.82       21.8
##  2  1998        14.5       8.45       21.1
##  3  1999        14.2       8.37       20.5
##  4  2000        14.2       8.13       20.7
##  5  2001        13.6       7.85       19.8
##  6  2002        14.1       8.51       20.2
##  7  2003        14.6       8.99       20.6
##  8  2004        13.8       8.27       19.7
##  9  2005        14.4       8.1        21.1
## 10  2006        15.0       9.12       21.4
## 11  2007        13.5       7.44       19.9
## 12  2008        14.0       8.1        20.3
## 13  2009        15.1       8.93       21.6
## 14  2010        14.3       8.97       20.2
## 15  2011        15.4       9.18       22.0
## 16  2012        14.5       8.02       21.2
## 17  2013        14.5       8.15       20.9
## 18  2014        15.8       9.52       22.0
## 19  2015        15.7       8.96       22.6
```

4.2 Comprobación de la normalidad y homogeneidad de la varianza.

4.2.1 Test de normalidad de Kolmogorov-Smirnov

```
library(dplyr)
library(nortest)
data_n <- select(data_weather_clean, -CET, -Events)

a <- sapply(data_n, function(x) if (lillie.test(x)$p.value < 0.01)
{c(lillie.test(x)$p.value, "Normal")} else {c(lillie.test(x)$p.value, "No
Normal")})
a
```

```
##      Max.TemperatureC        Mean.TemperatureC       Min.TemperatureC
## [1,] "1.61524536849871e-155" "1.80997197860294e-138"
"5.42182454064599e-91"
## [2,] "Normal"                "Normal"                "Normal"
##      Dew.PointC              MeanDew.PointC          Min.DewpointC
## [1,] "4.12667486006265e-95" "6.68773067604575e-146"
"7.29027951070114e-66"
## [2,] "Normal"                "Normal"                "Normal"
##      Max.Humidity Mean.Humidity          Min.Humidity
## [1,] "0"          "1.05985926151013e-69" "2.22792112390749e-103"
```

```
## [2,] "Normal"     "Normal"               "Normal"
##      Max.Sea.Level.PressurehPa Mean.Sea.Level.PressurehPa
## [1,] "8.602329103562e-64"      "5.01147671979378e-71"
## [2,] "Normal"                  "Normal"
##      Min.Sea.Level.PressurehPa Mean.VisibilityKm Max.Wind.SpeedKm.h
## [1,] "3.39955365680883e-55"    "0"               "2.11222051854002e-
199"
## [2,] "Normal"                  "Normal"          "Normal"
##      Mean.Wind.SpeedKm.h WindDirDegrees
## [1,] "0"                 "2.36798372894893e-269"
## [2,] "Normal"            "Normal"
```

Según la prueba de normalidad de Lilliefors todas las variables cuantitativas siguen una distribución normal, el pvalor de todas las variables es bastante inferior al nivel de significación fijado de 0,05.

4.2.2 Homogeneidad de la varianza

Se comprueba la homogeneidad de la varianza de la temperatura media, mínima y máxima (Mean.TemperatureC) para dos intervalos de tiempo de 5 años (1997-2001 y 2011-2015) sobre los que se va aplicar un constraste de hipótesis

```
# Test de barlett
bartlett.test(Mean.TemperatureC ~ grupo, data = data_ch)

##
##  Bartlett test of homogeneity of variances
##
## data:  Mean.TemperatureC by grupo
## Bartlett's K-squared = 3.9536, df = 1, p-value = 0.04677

bartlett.test(Max.TemperatureC ~ grupo, data = data_ch)

##
##  Bartlett test of homogeneity of variances
##
## data:  Max.TemperatureC by grupo
## Bartlett's K-squared = 4.9997, df = 1, p-value = 0.02535

bartlett.test(Min.TemperatureC ~ grupo, data = data_ch)

##
##  Bartlett test of homogeneity of variances
##
## data:  Min.TemperatureC by grupo
## Bartlett's K-squared = 2.4811, df = 1, p-value = 0.1152
```

La homocedasticidad o igualdad de varianzas se cumplen para la temperatura media y la máxima considerando un nivel de significación de 0.05, pero no para las temperaturas mínimas aunque tampoco se aleja mucho

4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos.

4.3.1 Análisis de correlación

Nos interesa saber qué variables cuantitativas influyen más en la visibilidad media y si estas son significativas. Para cuantificarlo se hará un análisis de correlación utilizando el coeficiente de correlación de Pearson. Este coeficiciente se aplica a variables aleatorias cuantitativas continuas y es independiente de la escala de medida de las variables por lo que no hace falta normalizarlas. Hay que tener en cuenta que el coeficiente mide la correlación lineal entre dos variables y el hecho de que el coeficiente sea bajo solo refleja no linealidad entre las variables.

```
library(dplyr)

# Se seleccionan las variables cuantitativas
data_cor <- select(data_cor, -CET, -Events)

# Se crea una matriz vacía donde se guardaran los resultados
corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")

# Se realiza el análisis de correlación entre las humedad relativa media
y el resto de variables
for (i in 1:(ncol(data_cor)-1)) {
  t_test <- cor.test(data_cor[,i],data_cor[,"Mean.VisibilityKm"], method
= "pearson")
  corr_coef = t_test$estimate
  p_val = t_test$p.value

  pair = matrix(ncol = 2, nrow = 1)
  pair[1][1] = corr_coef
  pair[2][1] = p_val

  corr_matrix <- rbind(corr_matrix,pair)
  rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(data_cor)[i]
}

# Se muestran los resultados del análisis de correlación
corr_matrix
```

```
##                              estimate       p-value
## Max.TemperatureC           0.299083023 1.249667e-121
## Mean.TemperatureC          0.251678141  1.607670e-85
## Min.TemperatureC           0.152723910  5.592970e-32
## Dew.PointC                -0.007271701  5.774511e-01
## MeanDew.PointC            -0.048571115  1.965889e-04
## Min.DewpointC             -0.102657570  3.127514e-15
```

```
## Max.Humidity                -0.300060599 1.884518e-122
## Mean.Humidity               -0.394684320 3.755946e-218
## Min.Humidity                -0.476237185  0.000000e+00
## Max.Sea.Level.PressurehPa    0.067798640  1.990546e-07
## Mean.Sea.Level.PressurehPa   0.039258589  2.622257e-03
## Min.Sea.Level.PressurehPa   -0.081809980  3.425022e-10
## Mean.VisibilityKm            1.000000000  0.000000e+00
## Max.Wind.SpeedKm.h           0.033245197  1.084334e-02
## Mean.Wind.SpeedKm.h          0.048793945  1.836823e-04
```

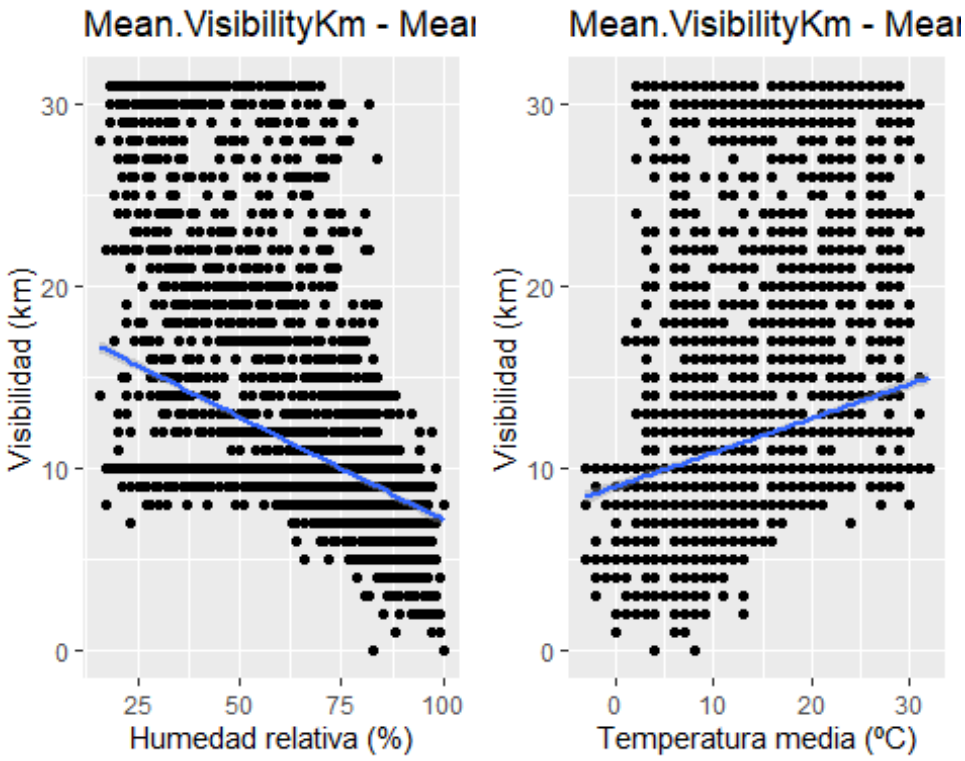
Los resultados muestran que las variables que más influyen en la visibiliad media son:

- las variables que miden la humedad relativa del aire, cuanto mayor es la humedad menor es la visibilidad en km ya que es más probable que se produzca lluvia, llovizna, niebla o neblina,
- le siguen las variables de temperatura que muestran una correlación positiva con la visibilidad (cuanto mayor es la temperatura mayor es la visibilidad pues el aire puede contener una mayor cantidad de agua sin que esta condense),
- en tercer lugar se encuentra la temperatura de rocío mostrando una correlación negativa, pues cuando mayor es la temperatura a la que el vapor de agua empieza a condensar menor es la visibilidad del aire y, cuanto más baja, menor la probabilidad de que el vapor condense favorenciendo una mayor visibilidad.
- La influencia de la velocidad media del viento es pequeña pero aun así significativa (p-value = 1.8368e-04), con el viento se mejora la visibiliad. Lo mismo se puede decir de la presión a nivel del mar, pero su relación es más indirecta influyendo en las condiciones climáticas asociados a anticiclones y borrascas.
- Hay que tener en cuenta que en el análisis no se consideran otros factores importantes que no se recogen en el dataset como las partículas en supensión reduciendo la visibilidad (brua, calima) o la posición del sol.

Esta correlación se puede visualizar mediante un diagrama de dispersión y su recta de regresión lineal, en los siguiente gráficos se representa la visibilidad frente a la humedad relativa y la temperatura media.

```
library(ggplot2)
library(gridExtra)

d1 <- ggplot(data_cor, aes(x=Mean.Humidity, y=Mean.VisibilityKm)) +
geom_point() + ggtitle("Mean.VisibilityKm - Mean.Humidity") +
xlab("Humedad relativa (%)") + ylab("Visibilidad (km)") +
geom_smooth(method=lm)

d2 <- ggplot(data_cor, aes(x=Mean.TemperatureC, y=Mean.VisibilityKm)) +
geom_point() + ggtitle("Mean.VisibilityKm - Mean.TemperatureC") +
xlab("Temperatura media (ºC)") + ylab("Visibilidad (km)") +
```

```
geom_smooth(method=lm)

grid.arrange(d1, d2, nrow=1, ncol= 2)
```



4.3.2 Contraste de hipótesis

Se quiere saber si la temperatura media anual y mensual del periodo 1997-2001 (1) es la misma que para el periodo 2011-2015 (2) o si, por el contrario, se ha producido un cambio significativo en las mismas.

Para hacer la comparación se utiliza la prueba de t-Student, un test paramétrico de constraste de hipótesis que permite comprobar la igualdad de las medias de dos muestras que siguen una distribución normal. El test se aplica comparando tanto los dos periodos en su conjunto como cada mes de los dos periodos.

El test trabaja con las siguientes hipótesis:

- Hipótesis hula (h0): no existe diferencia significativa en la temperatura media
- Hipótesis alternativa (h1): existe una diferencia significativa en la temperatura.

1. Si se compara la temperatura media de los dos periodos tenemos el siguiente resultado:

```
# test t-Student
t.test(Mean.TemperatureC ~ grupo, data = data_ch)

##
##  Welch Two Sample t-test
```

```
##
## data:  Mean.TemperatureC by grupo
## t = -3.0857, df = 3552.8, p-value = 0.002047
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.2713013 -0.2834249
## sample estimates:
## mean in group 1 mean in group 2
##        14.44881        15.22618
```

El test rechaza la hipótesis nula y acepta la hipótesis alternativa de que hay un cambio estadísticamente significativo (p-value = 0.002) en la temperatura media entre el periodo 1997-2001 (14,45 ºC) y el periodo 2011-2015 (15,23 ºC). Se constata un incremento en la temperatura media con un intervalo de confianza entre 1.27 y 0.28 ºC.

2. Si se compara la temperatura media de cada mes que hay en los dos periodos se obtiene los siguientes resultados

```
# Se crea una matriz vacía donde se guardarán los resultados
ch_matrix <- matrix(nc = 5, nr = 0)
colnames(ch_matrix) <- c("month", "men_temp_1997-2001","men_temp_2011-
2015" , "p-value", "h1")

# Contraste de hipótesis para cada mes
for (i in 1:12) {
  ch <- t.test(Mean.TemperatureC ~ grupo, data = data_ch[data_ch$month ==
i,])
  group1_mean = ch$estimate[1]
  group2_mean = ch$estimate[2]
  p_val = ch$p.value
  h1 = if(ch$p.value < 0.05) {"aceptada"} else {"rechazada"}

  pair = matrix(ncol = 5, nrow = 1)
  pair[1][1] = i
  pair[2][1] = group1_mean
  pair[3][1] = group2_mean
  pair[4][1] = p_val
  pair[5][1] = h1

  ch_matrix <- rbind(ch_matrix,pair)

}

ch_matrix
```

```
##       month men_temp_1997-2001 men_temp_2011-2015 p-value
##  [1,] "1"   "5.8"               "6.12258064516129" "0.305071479145916"
##  [2,] "2"   "8.13432835820896" "6.48936170212766" "1.37898095246891e-
06"
```

```
##  [3,] "3"   "11.4596774193548" "10.0129032258065" "5.94192048968991e-
06"
##  [4,] "4"   "12.7333333333333" "13.7533333333333"
"0.00603557488957751"
##  [5,] "5"   "16.0789473684211" "17.7032258064516"
"0.000133135952270861"
##  [6,] "6"   "21.1901408450704" "22.5266666666667"
"0.000371425515054444"
##  [7,] "7"   "24.0322580645161" "25.8645161290323" "6.52163910890356e-
08"
##  [8,] "8"   "24.7518248175182" "25.5290322580645"
"0.00834935072024491"
##  [9,] "9"   "20.3866666666667" "21.16"            "0.0182059119044383"
## [10,] "10"  "15.1032258064516" "16.1741935483871"
"0.00183407075756502"
## [11,] "11"  "8.44666666666667" "10.4"             "1.14270467659307e-
07"
## [12,] "12"  "5.21935483870968" "6.41290322580645"
"0.00021739966432379"
##       h1
##  [1,] "rechazada"
##  [2,] "aceptada"
##  [3,] "aceptada"
##  [4,] "aceptada"
##  [5,] "aceptada"
##  [6,] "aceptada"
##  [7,] "aceptada"
##  [8,] "aceptada"
##  [9,] "aceptada"
## [10,] "aceptada"
## [11,] "aceptada"
## [12,] "aceptada"
```

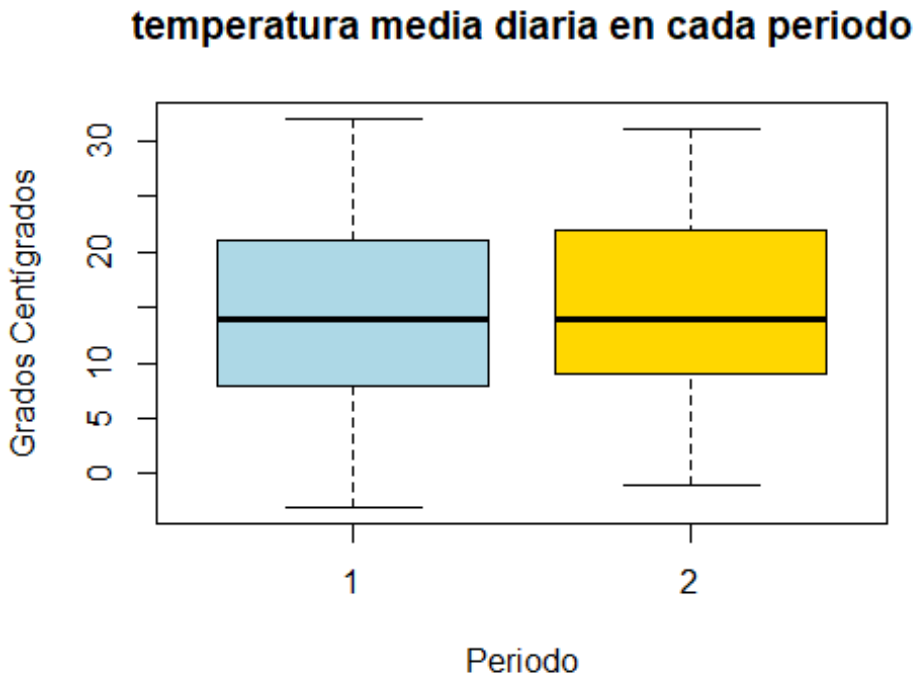
A excepción del mes de enero, en el que la hipótesis nula es aceptada, en los demás meses el test rechaza la hipótesis nula y acepta la hipótesis aternativa de que hay un cambio significativo en la temperatura media del mes. En los meses de febrero y marzo se produce una disminución de la temperatura media mientras que en los meses de abril a diciembre se produce un incremento en las temperaturas medias.

Representación gráfica

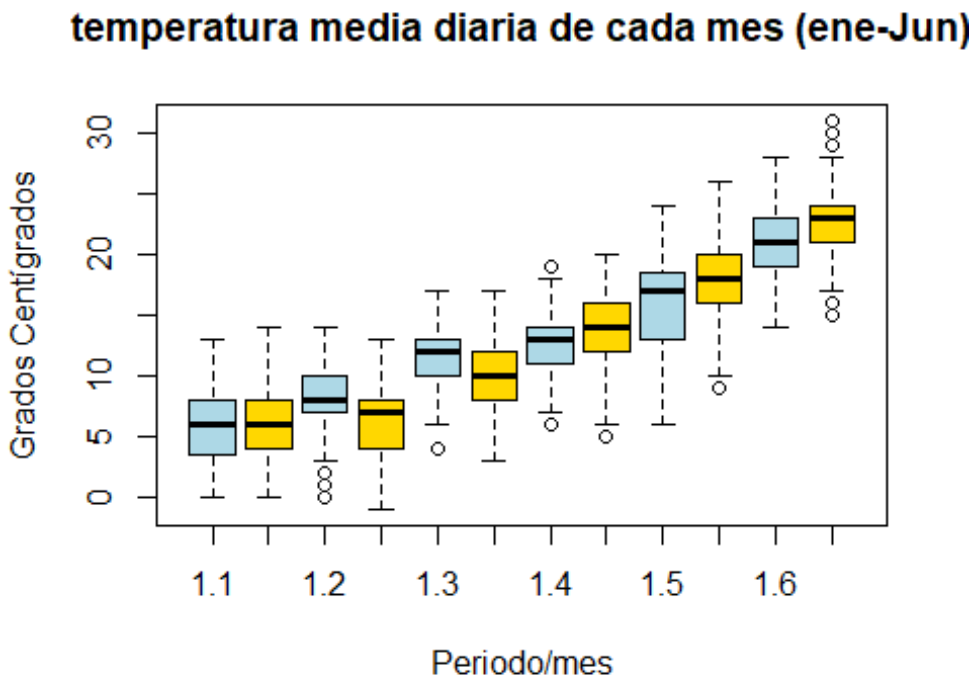
Una forma de visualizarlo es utilizando diagramas de caja y bigotes, en el que se comparan la distribución de la temperatura media diaria del periodo 1997-2001 (1, en azul) y el periodo 2011-2015 (2, en amarillo)

```
boxplot(Mean.TemperatureC ~ grupo,
        data = data_ch,
        col = (c("lightblue", "gold")),
        main = "temperatura media diaria en cada periodo",
```

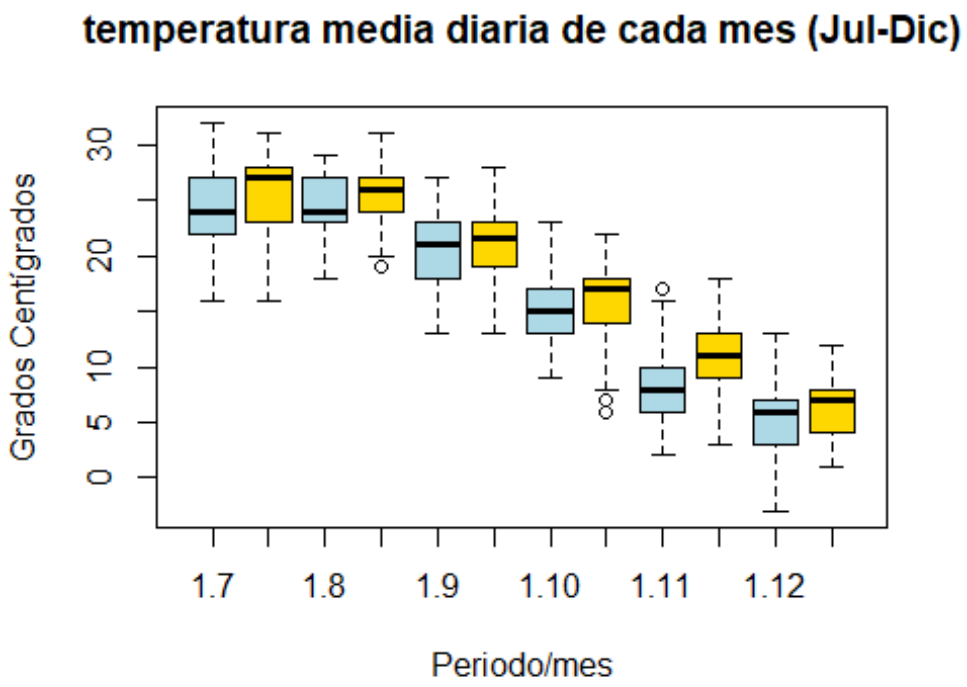
```
        xlab = "Periodo",
        ylab = "Grados Centígrados")
```




```
boxplot(Mean.TemperatureC ~ grupo + month,
        data = data_ch[data_ch$month <= 6,],
        col = (c("lightblue", "gold")),
        main = "temperatura media diaria de cada mes (ene-Jun)",
        xlab = "Periodo/mes",
        ylab = "Grados Centígrados")
```




```r
boxplot(Mean.TemperatureC ~ grupo + month,
        data = data_ch[data_ch$month > 6,],
        col = (c("lightblue", "gold")),
        main = "temperatura media diaria de cada mes (Jul-Dic)",
        xlab = "Periodo/mes",
        ylab = "Grados Centígrados")
```

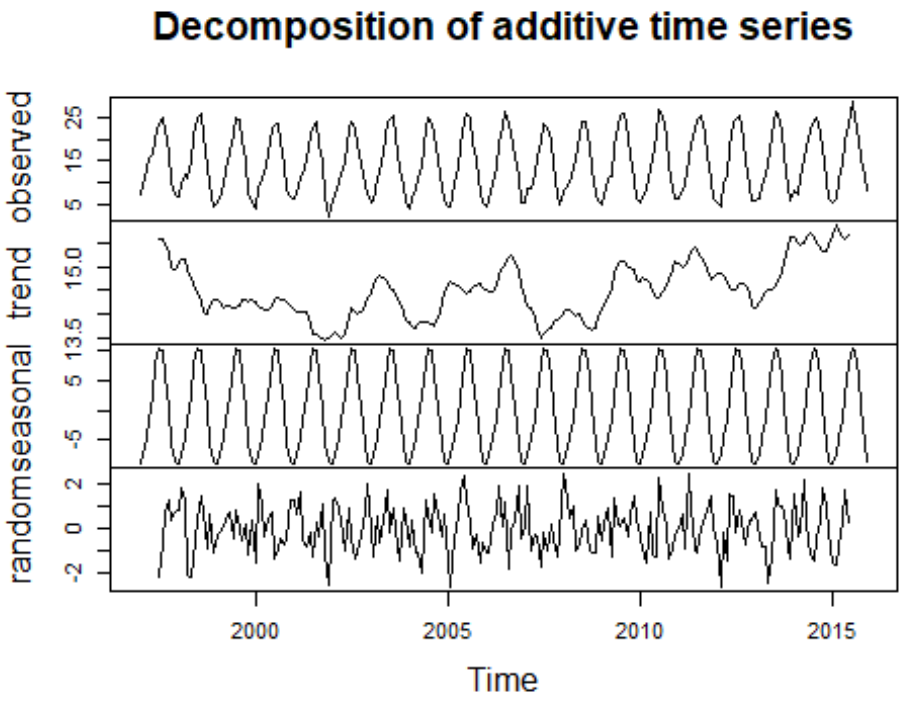


4.3.3 Análisis de tendencia de la temperatura media mensual 1997-2015

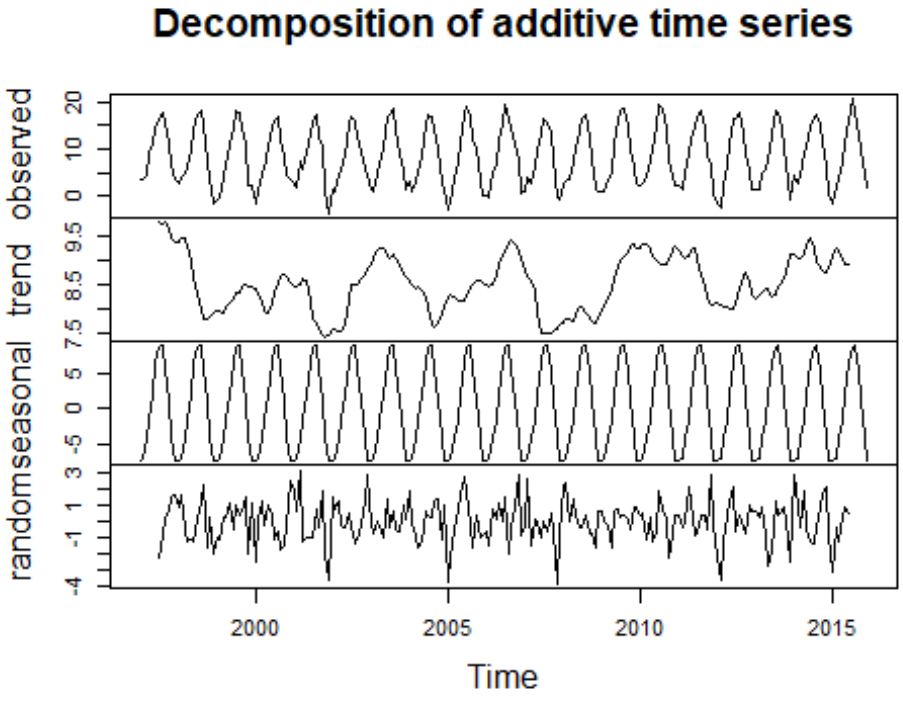
Con los datasets mensuales y anuales generados en el apartado 4.1 se realiza un análisis de tendencias de la temperatura media, mínima y máxima para el periodo 1997-2015. Para determinar si existe una tendencia en las temperaturas de la serie histórica se ha utilizado el test de Mann-Kendall, un test no paramétrico que puede manejar patrones estacionales dentro de los datos.

En primer lugar se va a descomponer la serie temporal en sus componentes constituyentes: El componente de tendencias, el componente aleatorio y el componente estacional.

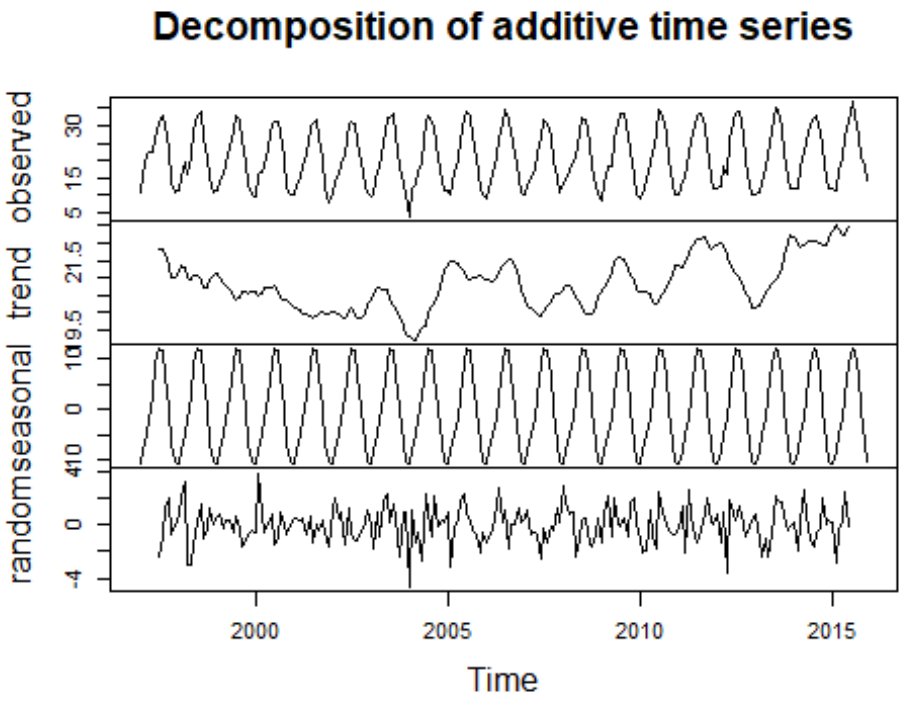
```
if(!require(mice)){install.packages("mice")}
if(!require(Kendall)){install.packages("Kendall")}
if(!require(trend)){install.packages("trend")}
library(mice)
library(Kendall)
library(trend)

# Se convierten los datos mensuales en un objeto de serie temporal ts
TS_mean = ts(data_month$Mean.Temp_C, frequency = 12, start = c(1997,1))
TS_min = ts(data_month$Min.Temp_C, frequency = 12, start = c(1997,1))
TS_max = ts(data_month$Max.Temp_C, frequency = 12, start = c(1997,1))

# Descomposición de los objetos de las series temporales
plot(decompose(TS_mean)) # TEMPERATURA MEDIA MENSUAL
```




```r
plot(decompose(TS_min)) #  MEDIA DE LAS TEMPERATURA MÁXIMAS
```




```r
plot(decompose(TS_max)) #  MEDIA DE LAS TEMPERATURA MÁXIMAS
```



Si se observa el gráfico que muestra la tendencia (trend) se aprecia que en los primeros años de la serie hay un descenso de las temperaturas, para luego estabilizarse y fluctuar hasta que a partir de 2005 sigue una tendencia positiva con fluctuaciones. A la hora de interpretar los resultados del test hay que tener en cuenta que la tendencia es relativa a la ventana temporal que se analiza y no puede extrapolarse de manera alegre, sobre todo si el intervalo temporal es pequeño como el utilizado aquí (desde un punto de vista climático)

el Test de Mann-Kendall da los siguientes resultados para medias anuales y mensuales

```
# Test de Mann-Kendall de serie anual
MK_mean_year = MannKendall(data_year$Mean.Temp_C)
MK_min_year = MannKendall(data_year$Min.Temp_C)
MK_max_year = MannKendall(data_year$Max.Temp_C)

#Test de Mann-Kendall de serie estacional
SMK_mean = SeasonalMannKendall(TS_mean)
SMK_min = SeasonalMannKendall(TS_min)
SMK_max = SeasonalMannKendall(TS_max)
```

Los resultados obtenidos con el test de Mann-Kendall son los siguientes:

Temperatura media: * Anual: tau = 0.246, 2-sided pvalue =0.15121 * Estacional: tau = 0.117, 2-sided pvalue =0.015335

Media de las Temperaturas Máximas: * Anual: tau = 0.287, 2-sided pvalue =0.093092 * Estacional: tau = 0.164, 2-sided pvalue =0.000713

Media de las Temperaturas Mínimas: * Anual: tau = 0.0587, 2-sided pvalue =0.75271 * Estacional: tau = 0.0327, 2-sided pvalue =0.49838

En los datos de la serie de temperatura anual (media, mínimo y máximo) no se aprecia una clara tendencia para el periodo analizado, no es significativa, los p-valores son amyores al nivel de significación de 0.05.

Por otra parte, si se observa una tendencia estacional significativa en la serie de temperaturas medias mensuales y en el de las máximas mensuales, no así en el de las mínimas.

6. Conclusiones

En los objetivos se plantean tres cuestiones a responder a partir de la serie de datos climáticos 1997-2015 de la estación metereológica del aeropuerto de Barajas:

1. ***¿Qué factores climatológicos influyen más sobre la visibilidad y en qué sentido lo hace?*** Para responder a esta cuestión se ha llevado a cabo un análisis de correlación entre la variable visibilidad y las demás variables cuantitativas aplicando el coeficiente de correlación de pearson. Los resultados muestran que la variable más influyente es la humedad relativa, seguida de lejos por la temperatura y por el punto de rocío. Todas son estadísticamente significativas aunque los coeficientes no sean muy altos. La humedad relativa y el punto de rocío se correlaciona de forma negativa con la visibilidad, de tal manera que a mayor humedad relativa y mayor temperatura de rocío menor es la visibilidad. En cambio, la temperatura se correlaciona de manera positiva, cuanto mayor es la temperatura más se favorece la visibilidad.

2. ***¿Se ha producido un cambio estadísticamente significativo en la temperatura media de los últimos 5 años repecto al del periodo 1997-2001?*** Para averiguarlo se ha realizado un test paramétrico de constraste de hipótesis para comprobar la igualdad de las temperaturas medias de ambos periodos. En este caso, se ha aplicado el test t-Student comparando tanto la temperatura media de ambos periodos como el de cada uno de los meses. Los resultados muestran que hay un aumento estadísticamente significativo en la temperatura media entre 1997-2001 y 2011-2015. Cuando se compara la temperatura media de cada mes, los meses de febrero y marzo muestra un descenso en la temperatura media entre ambos periodos mientras que en los meses que van de abril a diciembre se detecta un aumento en la tempertura medio, todos estadísticamente significativos. Solo en el mes de enero el cambio de la temepratura media no es significativo.

3. ***¿Cuál ha sido la tendencia en la temperatura media, máxima y mínima anual y estacional para el periodo comprendido entre 1997 y 2015?*** Para responder a esta pregunta se ha hecho un análisis de tendencias utilizando el test de Mann-Kendall que permite separar la tendencia en las temperaturas de las fluctuaciones estacionales propias. los resultados no muestran una tendencia

clara dentro de la ventana temporal analizada, las temperaturas anuales (media, maxima y mínima) no muestran tendencias significativas. En cambio, la media de las temperaturas máximas y las temperaturas medias estacionales si muestran una tendencia positiva estadisticamente significativa según el test.

Los análisis llevados a cabo han respondido, en parte, a las cuestiones planteadas siendo una primera aproximación. Así, por ejemplo, el análisis de correlación solo tiene en cuenta la influencia de las variables una a una y no en su conjunto. Habría que hacer análisis adicionales para ir más allá como un análisis de regresión múltiple. Los resultados del constraste de hipótesis y el análisis de tendencia para ver si hay cambios significativos en las temperaturas medias o una tendencia son muy dependientes de las ventanas temporales que se utilicen, por lo que sus conclusiones no se pueden generalizar. Se necesitan series temporales más largas para que sean representativas.