# U03 Ejercicio 01

En este ejercicio vamos a analizar los datos meteorológicos de 6 estaciones situadas en algunas ciudades españolas. Los datos se han obtenido a partir del paquete weatherData. Se trata de datos a nivel diarío y recogen las siguientes variables:

- estacion: localización de la estación
- id_estacion: código abreviado de la estación
- Date: fecha

- Max_TemperatureC: temperatuta máxima diaria
- Mean_TemperatureC: temperatura media diaria
- Min_TemperatureC: temperatura máxima diaria
- Max_Wind_SpeedKm_h: velocidad del viento máxima diaria
- Mean_Wind_SpeedKm_h: velocidad del viento media diaria
- WindDirDegrees: dirección del viento promedio diaria
- Precipitationmm: Precipitación en mm de altura (es equivalente a $l/m^2$ )
- CloudCover: Grado de copertura del cielo (escala 1:8)
- Mean_Humidity: Humedad media (%)
- Events: cadena textual con los eventos meteorológicos ocurridos, separados por guiones
- Fog: niebla, TRUE/FALSE
- Hail: granizo, TRUE/FALSE
- Rain: lluvia, TRUE/FALSE
- Snow: nieve, TRUE/FALSE
- Thunderstorm: tormenta, TRUE/FALSE
- Tornado: tornado, TRUE/FALSE

Para cargar los datos hacemos lo siguiente:

```
library(reshape2)
library(tidyverse)
# Compruba la localización de los ficheros en tu ordenador y adapta la ruta si es necesario
estaciones <- read.delim("../../Datasets/estaciones_meteo.txt",sep="\t")
estaciones <- estaciones %>%  filter(estacion %in% c("MADRID","BARCELONA","SEVILLA","ZARAGOZA","BILBAO"
meteo_data <- read.delim("../../Datasets/meteo_data.csv",sep=";",stringsAsFactors = FALSE)
meteo_data <- meteo_data %>% mutate(Date=as.Date(Date))
meteo_data <- meteo_data %>% filter(estacion  %in% estaciones$estacion)
```

Veamos un resumen de lo que contienen estos datos

```
summary(meteo_data)
```

```
##    estacion         id_estacion            Date
##  Length:2196        Length:2196        Min.   :2016-01-01
##  Class :character   Class :character   1st Qu.:2016-04-01
##  Mode  :character   Mode  :character   Median :2016-07-01
##                                        Mean   :2016-07-01
##                                        3rd Qu.:2016-10-01
##                                        Max.   :2016-12-31
##
##  Max_TemperatureC Mean_TemperatureC Min_TemperatureC Max_Wind_SpeedKm_h
##  Min.   : 1.00    Min.   :-1.00     Min.   :-6.00    Min.   : 6.00
##  1st Qu.:15.00    1st Qu.:11.00     1st Qu.: 7.00    1st Qu.:16.00
##  Median :20.00    Median :16.00     Median :11.00    Median :23.00
##  Mean   :21.43    Mean   :16.25     Mean   :11.11    Mean   :24.53
```

```
##  3rd Qu.:27.00    3rd Qu.:21.00     3rd Qu.:16.00    3rd Qu.:29.00
##  Max.   :44.00    Max.   :33.00     Max.   :26.00    Max.   :72.00
##
##  Mean_Wind_SpeedKm_h WindDirDegrees Precipitationmm    CloudCover
##  Min.   : 2.00       Min.   :  1    Min.   : 0.000   Min.   :1.000
##  1st Qu.: 6.00       1st Qu.:114    1st Qu.: 0.000   1st Qu.:1.000
##  Median :10.00       Median :233    Median : 0.000   Median :3.000
##  Mean   :10.87       Mean   :208    Mean   : 1.483   Mean   :3.236
##  3rd Qu.:14.00       3rd Qu.:292    3rd Qu.: 0.315   3rd Qu.:5.000
##  Max.   :47.00       Max.   :360    Max.   :78.990   Max.   :8.000
##                                                      NA's   :356
##  Mean_Humidity       Events             Fog             Hail
##  Min.   : 19.00   Length:2196        Mode :logical   Mode :logical
##  1st Qu.: 56.00   Class :character   FALSE:1909      FALSE:2189
##  Median : 70.00   Mode  :character   TRUE :287       TRUE :7
##  Mean   : 67.41                      NA's :0         NA's :0
##  3rd Qu.: 80.00
##  Max.   :100.00
##
##     Rain             Snow          Thunderstorm     Tornado
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:1497      FALSE:2189      FALSE:2092      FALSE:2195
##  TRUE :699       TRUE :7         TRUE :104       TRUE :1
##  NA's :0         NA's :0         NA's :0         NA's :0
##
##
##
```

# Distribución de frecuencias

En primer lugar vamos a estudiar las distribuciones de frecuencia

Para la estación de Zaragoza y la variable Mean_TemperatureC, dibuja un histograma. Fija la anchura de las categorías a un grado centígrado.

Ahora dibuja de forma conjunta los histogramas para la variable Mean_TemperatureC (Pista si dibujas el histograma con ggplot usa position="identity": geom_histogram(aes(...),position="identity"))

¿Observas la diferencia entre las distribuciones en las diferentes estaciones? Para poder apreciar mejor las diferencias entre estaciones, representa el mismo gráfico pero:

- usando polígonos de frecuencias (geom_freqpoly())
- usando densidades de frecuencia ajustadas (geom_density())

Ahora, dibuja las densidades para la variable Precipitationmm

## Gráficos con varias variables - facets

¿Puedes dibujar las distribuciones de varias variables en un solo gráfico?

Para que el gráfico sea interpretable usa solo las variables Mean_TemperatureC, Mean_Wind_SpeedKm_h,WindDirDegrees, Precipitationmm ,CloudCover ,Mean_Humidity

Pista: Para ello conviene primero transformar los datos a formato long mediante la función melt del paquete reshape2

### Boxplots

Otra forma de representar las distribuciones de una variable, más simplificada, pero más apropiada para la comparación, es el boxplot.

Dibuja un boxplot que el eje x presente las estaciones y en el eje y las distribuciones de la variable Max_TemperatureC.

Ahora dibuja un gráfico con facets para ver los boxplots de las diferentes variables

### Variables categóricas

En el conjunto de datos meteorológicos, tenemos también una serie de variables categóricas que nos dicen si en un día determinado se ha producido un determinado fenómeno (lluvia, nieve, tormenta, etc)

¿Cual es la estación con más eventos de lluvia?

No te parece que alguna ciudad tiene más días de lluvia que lo que sugiere la variable Precipitationmm Cuenta para cada estación el número de días con precipitación mayor que 0

¿Y de niebla?

¿Eres capaz de mostrar los conteos de todas estas variables en un solo gráfico?

## Medidas de posicion

### Comparación media,mediana, media truncada

Calcula la media, la mediana y la media truncada para cada estación de la variable Mean_Wind_SpeedKm_h

Ahora hazlo en un solo paso para las variables Mean_TemperatureC, Mean_Wind_SpeedKm_h, WindDirDegrees, Precipitationmm ,CloudCover ,Mean_Humidity

(Pista: Puedes usar la transformación a datos long que has hecho en el ejercicio 1 de la unidad)

¿Puedes hacerlo también sin transformar los datos a tipo long?

(Pista: usa la función summarise_each del paquete dplyr)

Selecciona del resultado anterior solo las medias normales de las variables temperatura.

(Pista: Usa la funcion matches dentro de select)

Ahora representa los resultados de forma gráfica.

#### Medias de ángulos (Para alumnos Top)

Los ángulos son tipo de variable especial, ya que son ciclícos, es decir están limitados a 360º y la distancia entre ellos no se calcula de forma convencional. La distancia entre un ángulo de 2º y 358º no es 356º sino 4º. Para calcular su media debe hacerse de forma vectorial. ¿Se te ocurre como hacerlo?

## Medidas de dispersión

Calculemos la dispersión de las diferentes variables, por estación. Para ello calcularemos tres medidas distintas:

- Desviación típica
- Rango intercuártilico (IQR)

Calculalos para cada estación para la variable Precipitacionmm

Calcula para las variables Mean_TemperatureC, Mean_Wind_SpeedKm_h,WindDirDegrees, Precipitationmm ,CloudCover ,Mean_Humidity en un solo paso

Muestra los resultados en una sola gráfica

### Cuantiles y boxplots personalizados

Para las variables de tipo numérico que has utilizado antes, calcula los cuantiles 0.05,0.25,0.5,0.75 y 0.95. Calcula también calculamos la media y la desviación típica

Esta informacíon puede mostrarse en un solo gráfico. ¿Se te ocurre como?

(Pista: puedes usar un boxplot personalizado para mostrar los cuantiles, puntos para la media y una barra de error para la sd)

## Forma de las distribuciones

Calcula ahora diversas medidas de forma en el mismo data frame:

- Coeficiente de variación
- Skewness
- Kurtosis

Veamos el resultado de forma gŕafica para una medida, por ejemplo el coeficiente de variación (cv)

Ahora, esto ya es de alumnos top, muestra todas las medidas para todas las estaciones y variables en una única gráfica

## Correlación

Calcula la matriz de correlación entre las variables numéricas para la estación de Barcelona

Responde a las siguientes preguntas:

¿Como cambia la temperatura cuando la nubosidad aumenta? ¿Con que temperatura correlaciona más con la máxima, mínima o media? ¿Que influye más en la copertura del cielo la humedad o el viento? ¿En el mismo sentido?

Representala la matriz de correlación gráficamente mediante la función ggpairs del paquete GGally