# Introduction

`weather_features.csv` contains hourly weather data from 2015 to 2018 for 5 major cities in Spain.

```r
weather_features <- read.csv("weather_features.csv")
df <- data.frame(
  Column = colnames(weather_features),
  Description = c("datetime index localized to CET",
                  "name of city (Barcelona, Bilbao, Madrid, Seville, Valencia)",
                  "temperature (K)",
                  "minimum temperature (K)",
                  "maximum temperature (K)",
                  "pressure (hPa)",
                  "humidity (%)",
                  "wind speed (m/s)",
                  "wind direction (degrees)",
                  "rain in last hour (mm)",
                  "rain in last 3 hours (mm)",
                  "snow in last 3 hours (mm)",
                  "cloud cover (%)",
                  "weather description - code",
                  "weather description - short",
                  "weather description - long",
                  "weather icon")
)
knitr::kable(df, format = "markdown")
```

| Column | Description |
|---|---|
| dt_iso | datetime index localized to CET |
| city_name | name of city (Barcelona, Bilbao, Madrid, Seville, Valencia) |
| temp | temperature (K) |
| temp_min | minimum temperature (K) |
| temp_max | maximum temperature (K) |
| pressure | pressure (hPa) |
| humidity | humidity (%) |
| wind_speed | wind speed (m/s) |
| wind_deg | wind direction (degrees) |
| rain_1h | rain in last hour (mm) |
| rain_3h | rain in last 3 hours (mm) |
| snow_3h | snow in last 3 hours (mm) |
| clouds_all | cloud cover (%) |
| weather_id | weather description - code |
| weather_main | weather description - short |
| weather_description | weather description - long |
| weather_icon | weather icon |

The last 4 columns are non-numerical, and will be dropped. The data for the city of Barcelona has city_name ” Barcelona”, the extra space will be removed.

```r
weather_features <- weather_features[1:13]
weather_features[weather_features$city_name == " Barcelona",]$city_name = "Barcelona"
```

# Missing Values

```r
sapply(weather_features, function(col)
  { sum(sapply(col, function(x) {
    (is.na(x)) + (x == "")
  })) })
```

```
##     dt_iso  city_name       temp   temp_min   temp_max   pressure   humidity
##          0          0          0          0          0          0          0
## wind_speed   wind_deg    rain_1h    rain_3h    snow_3h clouds_all
##          0          0          0          0          0          0
```

Data has no blank or NA cells in any of the columns. Next, we check that there are no missing rows. Since we have hourly data for 3 regular and 1 leap year, we expect there to be $24 \times 365 \times 4 + 24 = 35064$ unique datetime indices.

```r
dates_occurences <- table(weather_features$dt_iso)
length(dates_occurences)
```

```
## [1] 35064
```

This matches our expectations.

Since we have 5 cities, we expect each datetime index to appear 5 times.

```r
sum(sapply(dates_occurences, function(date) { date < 5 }))
```

```
## [1] 0
```

This indicates that we have no missing rows.

# Partition Based on City

This means a total of $35064 \times 5 = 175320$ rows. However, our data has 178396 rows. This indicates the presence of duplicated rows.

To fix this, we first create separate data frames for each city, then remove duplicates.

```r
library(dplyr)
```

```
## Warning: package ’dplyr’ was built under R version 4.4.2
```

```
##
## Attaching package: ’dplyr’
```

```
## The following objects are masked from ’package:stats’:
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
Barcelona <- weather_features[weather_features$city_name == "Barcelona", ] %>% distinct()
Bilbao <- weather_features[weather_features$city_name == "Bilbao", ] %>% distinct()
Madrid <- weather_features[weather_features$city_name == "Madrid", ] %>% distinct()
Seville <- weather_features[weather_features$city_name == "Seville", ] %>% distinct()
Valencia <- weather_features[weather_features$city_name == "Valencia", ] %>% distinct()

dim(Barcelona)
```

```
## [1] 35064    13
```

```
dim(Bilbao)
```

```
## [1] 35064    13
```

```
dim(Madrid)
```

```
## [1] 35064    13
```

```
dim(Seville)
```

```
## [1] 35064    13
```

```
dim(Valencia)
```

```
## [1] 35064    13
```