**How is pollution in the US affected by various specific factors?**

*Team members*

Ana Sofía Galán

Jose Marcos Marmolejo

Juan Eduardo Delgado

Alfredo Velarde

*Hypothesis*

Factors such as population, motor vehicles, industrial activity and wildfires affect pollution levels in the US. We believe that the main air pollutants (NO2, O3, SO2, CO) will increase as the levels of these factors grow in a certain State.

*These were the questions that we brought up:*

* How is pollution measured?
* Which are the main pollutants? How are they measured?
* Is there available information for all the factor that we are considering? Is it reliable?
* How should we compare our information (yearly/monthly, by city/state/county, etc.)?

*Sources of information:*

* Pollution data: Kaggle (extracted from US EPA).
* Population data: US Census.
* Wildfire data: Kaggle (extracted from Forest Service Research Data Archive).
* Vehicle and industrial data: Public databases.

*Data exploration and Cleanup process*

We downloaded data bases from different sources, and all contained a lot of information. Since pollution would be the variable that the rest would be compared to, we decided to take that data base as axis.

The analysis would be made focusing on the 4 main air pollutants (NO2, O3, SO2, CO) and the Air Quality Index (AQI). When we loaded the pollution data base on Jupyter, we learned that there was a lot of missing information for AQI, so we decided to eliminate the variable from our analysis.

With the Pollution data as axis, we then started merging the other databases into the data frame. This included grouping the information under the same criteria (Years and States). Information on motor vehicles and industrial activity didn’t fit the requirements for the analysis: Vehicles was not available divided by state; and Industrial activity was too detailed, and we didn’t know how to filter it properly, considering all the technical details. We decided not to consider those two factors for the analysis.

In the end, we kept only population and wildfires.

*Analysis*

Our analysis will compare how the level of population and occurrences of wildfires relate to the levels of the four main air pollutants, in a given US state, each year from 2010 to 2016.

We gathered all the data in a data frame, indexed by State and Year. With all the data in the single base, we made scatter plots for each of the pollutants against the population and wildfires levels. Graphically, all plots showed almost no linear correlation between our variables.

Since there was no apparent linear correlation, we divided both the population and wildfire variables into bins to normalize the information. We then made scatter plots again (with much less data, after having broken It apart).

*Findings and conclusions*

We concluded that the only pollutant with a significant correlation to our two variables was SO2 (Sulfur dioxide), with R2 values of -0.6876 for population, and -0.8297 with wildfires. And since the correlations are negative, we can conclude that SO2 will be reduced as population and wildfire levels increase.



