Data Processes Project

Alejandro Basco Plaza, Miriam Zaragoza Pastor, Jonatan Ruedas Mora, Raúl Cruz Benita and Laura Sánchez de Rojas Huerta

20/12/2019

Contents

Abstract

Introduction and related work

Exploratory data analysis

Dataset description

Before exposing the results of the exploratory data analysis carried the data, important information and background about the dataset is introduced below:

The dataset used for this work was obtained from Kaggle's website. In this page, we can find a lot of information and datasets about the air quality in Madrid.

The data in this data set has been collected from the original files provided by Madrid Open Data. Decide soluciones organization processed them and uploaded to the Kaggle's website.

Originally, Madrid Open Data get the data from 24 automatic stations around Madrid. Those stations, which are set around all the districts of the city, record information about pollution in the area. The data was generated from those stations whose pollution sensors measure air quality in Madrid city.

Regarding the observations, all csv files include twelve month data, except the last one that only has data until May. According to this, we have the followings rows:

CSV	Observations
Stations.csv	24 rows
Madrid2001.csv	217872 rows
Madrid2002.csv	217296 rows

CSV	Observations
Madrid2003.csv	243984 rows
Madrid2004.csv	245496 rows
Madrid2005.csv	237000 rows
Madrid2006.csv	230568 rows
Madrid2007.csv	225120 rows
Madrid2008.csv	226392 rows
Madrid2009.csv	215688 rows
Madrid2010.csv	209448 rows
Madrid2011.csv	209928 rows
Madrid2012.csv	210720 rows
Madrid2013.csv	209880 rows
Madrid2014.csv	210024 rows
Madrid2015.csv	210096 rows
Madrid2016.csv	209496 rows
Madrid2017.csv	210120 rows
Madrid2018.csv	69096 rows

We can also differentiate two domains whose columns represent different data:

On the one hand, we have Stations.csv with six columns, which contains information about the stations previously mentioned, in which the data about the pollution levels of the air is collected. On the other hand, we have Madrid20xx.csv, which contain the data about Madrid's air pollution itself. Some of these csv files have 14 columns, others 16 and others 17. We have the following columns in the csv files:

CSV	Features
Stations.csv	6 columns
Madrid2001.csv	16 columns
Madrid2002.csv	16 columns
Madrid2003.csv	16 columns
Madrid2004.csv	17 columns
Madrid2005.csv	17 columns
Madrid2006.csv	17 columns
Madrid2007.csv	17 columns
Madrid2008.csv	17 columns
Madrid2009.csv	17 columns
Madrid2010.csv	17 columns
Madrid2011.csv	14 columns
Madrid2012.csv	14 columns
Madrid2013.csv	14 columns
Madrid2014.csv	14 columns
Madrid2015.csv	14 columns
Madrid2016.csv	14 columns
Madrid2017.csv	16 columns
Madrid2018.csv	16 columns

Data wrangling

Before getting to work with the previously described data, some steps were performed in order to make the data more manageable:

Selected Features

In order to perform both the exploratory data analysis and the prediction methods, we will focus on the five gases that harm humans the most when they are present, and that relate the most to pollution: 1. CO

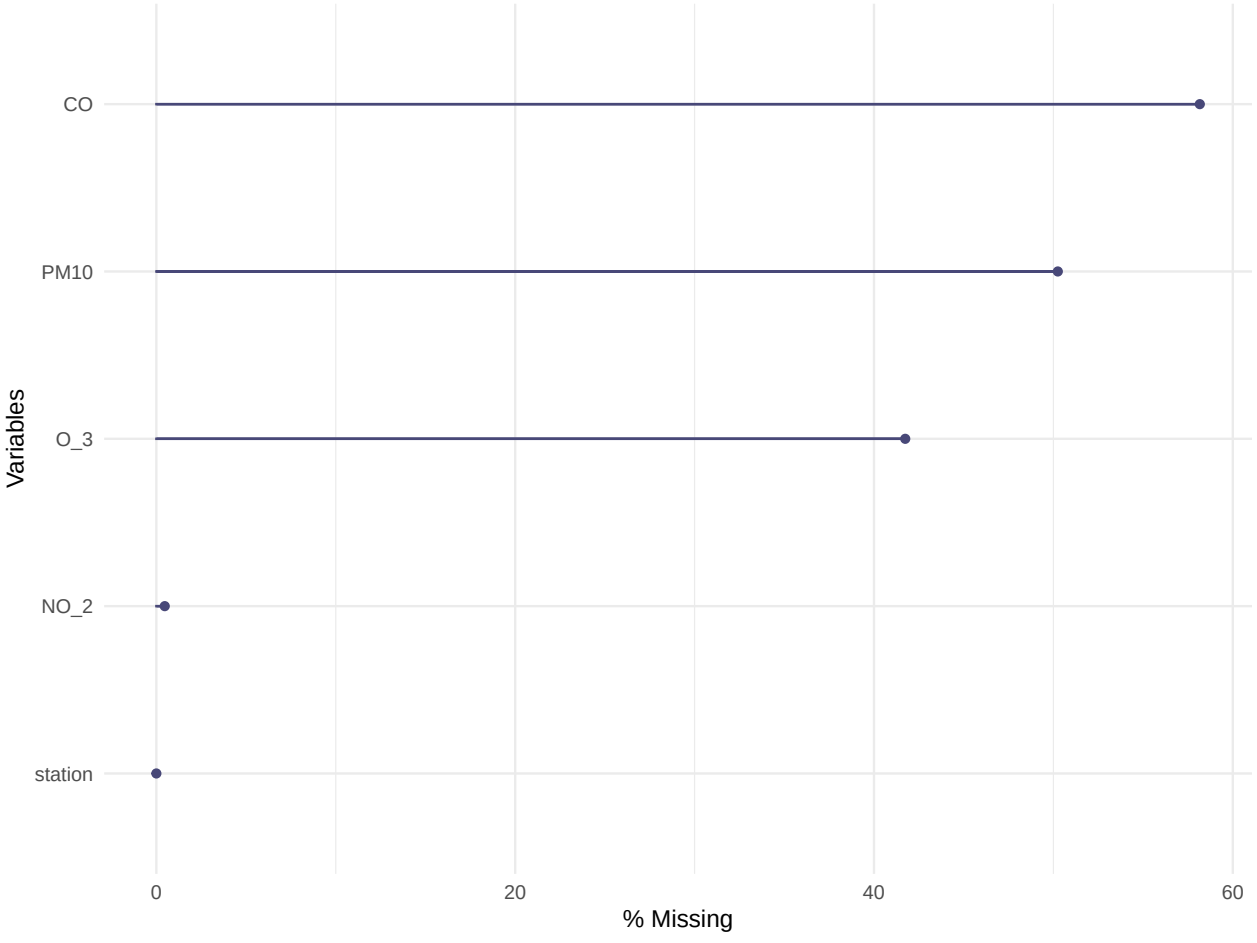
2. O3
3. SO2
4. PMH10
5. NO2

Datasets unification

In order to work with the data, all the yearly datasets were unified in a single one, adding one additional feature to the existing date column that all of the datasets present: the year. Thus, after this transformation, a single dataset contained all the information about Madrid's air pollution in the last 17 years, instead of 17 different datasets (one per year). Two additional columns were added: the *only_month* and *only_year* columns, wihic stand for the month and the year in which the observation was recorded, respectively.

Missing values treatment

As we can observe in the plot below, there is a high amount of missing values in the variables that we concern about:



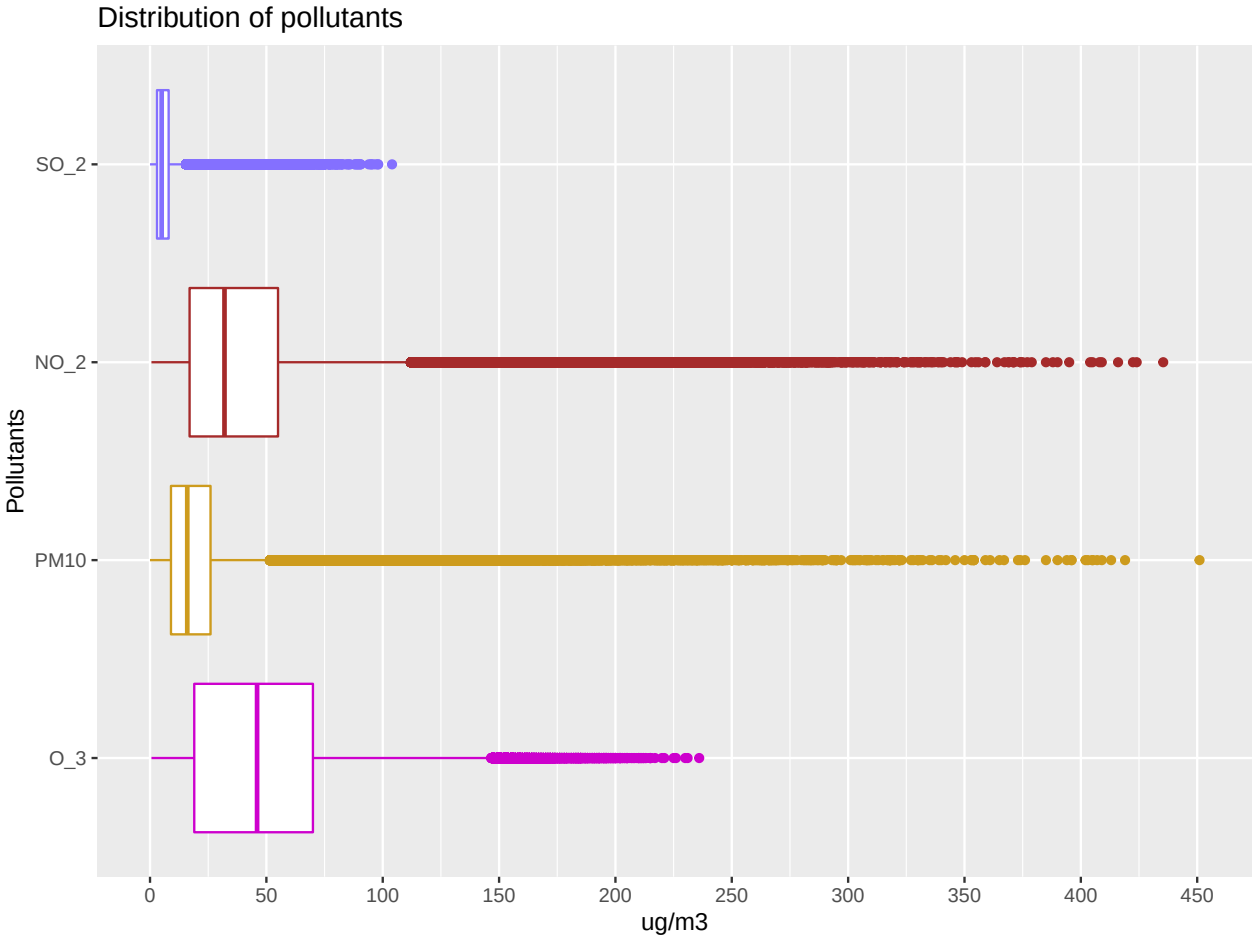
Thus, in order to deal with missing values, the mean of each one of the variables for each year, month and station is computed, so that all the missing values for each one of the columns in that same year, month and station are imputed with the corresponding mean.

Descriptive statistics analysis

After the data wrangling step, we are ready to perform the exploratory data analysis.

To start the analysis, the distribution of the variables in our dataset in terms of center, dispersion, and linear relationships descriptive measures, is analysed below.

The plot below is useful to visualize the center measures of the five selected variables:



As we can see, there are many outliers or unexpected values for all of the five variables.

Strength of relationships

Prediction

Results

Discussion and future work