**Introduction:**

For several decades, the number and scale of traffic jams have been increasing globally, particularly in developing countries. The causes are multiple, the main one being the increase in car traffic

This is the problem of determining traffic jams that we are suffering in our time and we are always looking for solutions to avoid congested roads

For this, we are going to create a program using machine learning to determine the amount of traffic from a set of effects that we obtained from a database that relates to a road linking between Minneapolis and Saint Paul in the United States.

**Objective:**

The purpose of this work is to estimate the density of road traffic according to weather conditions, day and time. Using the regression and classification method.

**Dataset presentation:**

We present in this report The Traffic Volume Data Set of the road that connects the city of Mennea police to the city of Saint Paul in the USA.

This database contains 9 attributes and 48 205 values ​​instances; the data has been collected between October 2012 and September 2018.

**Attribue présentation :**

* **Holiday:** National holidays, O there secular holidays, Unofficial holidays, Religious holidays, Substitute holidays.
* **Temp :** temperature in kelvin (float value)
* **rain\_1h :** Numeric Amount in mm of rain that occurred in the hour (float value € [0, 1])
* **snow\_1h :** Numeric Amount in mm of snow that occurred in the hour (float value € [0, 1])
* **clouds\_all :** Numeric Percentage of cloud cover between (integer value € [0, 90]).
* **weather\_main :** textual description of the current weather (Categorical short)
* **weather\_description :** textual description of the current weather (Categorical longer)
* **date\_time :** date and time (float value)
* **traffic\_volume :**  road traffic density (integer value)

**Problems present in the database:**

Poor quality data can have significant impacts. This is why it is important, for all those who create, manipulate or use data, to ensure the quality of the latter.

Here are the errors we found after browsing and analyzing the database.

1. **attribute Holiday**

this column contains the names of national holidays in America and we notice that it has a great impact on the degree of traffic

The problem is the lack of information, as there are 61 national holidays out of here



**None values** represent 99.91 % of all values in attribute holiday

Best solution is deleting this column

* The **weather\_main** attribute and the **weather\_description** attribute carry the same information as the rain\_1h, snow\_1h and clouds\_all attributes.

We can deduce them from the rain\_1h, snow\_1h and clouds\_all attributes.

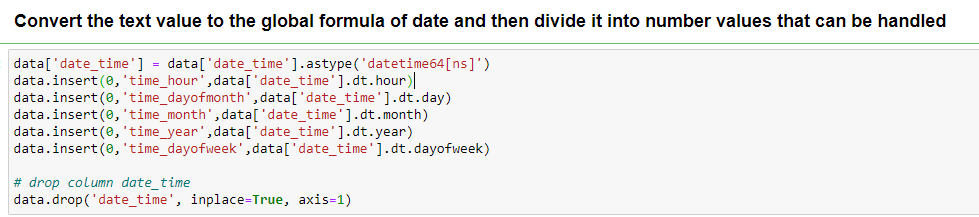
This means that there is redundancy. They are considered **“redundant data”.**

* **date\_time:** Contains multiple values for date and time.

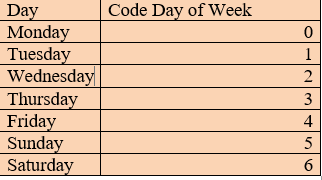
In order to deal with it, we had to divide it into parts; we converted the time-valued text and we extracted: the year; the month; day of the month; day of the week and the hour.

All the minutes values are zeros, there is no need to extract them.

And the solution is to re-divide it and convert it to numeric values using this code



And the result codification day of week is



The **“temp” :** Which represents the temperature measured in the Kelvin unit, for the presence of **outlier** values that cannot be achieved on planet Earth

**Example:**

|  |  |
| --- | --- |
|  | **temp** |
| 11900 | 0.0 |

**0.0 kelvin = -273, 1 °C**

**Fixing these problems:**

Poor quality data can lead to difficulties in extracting information, and ultimately, poor analysis and poor decision making.

To make our database more efficient, we will correct the errors step by step.

For this we will use the **pre-processing** method.

* We convert the temperature **“temp”** to Celsius degree; in order to study it using this code :

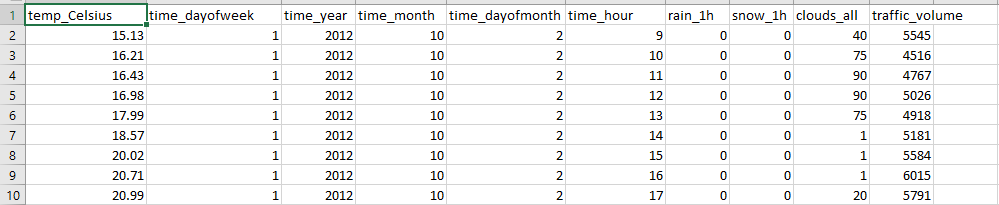


 We noticed the presence of outline values (example), so we delete the line.

* Contains multiple values ​​for date and time.

In order to deal with it, we had to divide it into parts; we converted the time-valued text and we extracted: the year; the month; day of the month; day of the week and the hour.

All the minutes values are zeros, there is no need to extract them.

After executing the code we get:

* To fix the problems present in the first and the third example; we select the relevant attributes and remove the unnecessary attributes using this function :