Data Analytics

Forest Fire Detection in the French mediterranean area

Manèle El Zoghlami

September, 2022

Table of content

Introduction

As the climate crisis takes its toll on the world, some regions have been suffering the consequences more than others. Floods, storms, extreme droughts, the impact of climate change is now undeniable.
In France, in the span of a few years, one of the most spectacular ways those changes have been reflected in nature is through fire forests. If the phenomenon is far from being a new one, in those last few years hectares of forest have been destroyed by the flames at an alarming rate.
If climate change cannot be single-handedly held as responsible for those natural hazards , it has been setting a propitious environment for those fires to start and spread at a concerning rate, the drought acting as a fuel.

For those reasons, it is now imperative for authorities throughout the world to implement the necessary means to prevent as much as possible fires from starting and spreading.

With this goal in mind, the Prométhée project was created in France in 1973. The very-prone-to-forest-fires region that is the Mediterranean area of France has been equipped with this tool to make an inventory of the region's fires, causes, spatial information, etc.

In that mindset, my business case will consist of analyzing the main reasons for forest fires in the Mediterranean area of France between March 2020 and September 2022, the most devastated departments of the region and analyzing the potential increase of those forest fires in the last years. I will provide some insights as to what could be done to slow down the spread of forest fires. Finally, I will do some forecasting to try and predict what the situation will be next year.
I will also use another dataset from INSEE to provide some insights in SQL about the towns and departments that have been the most hit by forest fires.

Data and data sources

When it comes to the data collection I wanted to find a reliable data source that would be updated regularly when it comes to the forest fires in France. I also needed a data source that would provide some information regarding the causes of the fire and the extent of the damages.
The data source I chose and that fitted those criterias was the one created by the french government in 1973 called Prométhée (Prometheus in English). This database was created to make the inventory of forest fires in the mediterranean region because of its sensitivity. This database is a collaborative one in which the services that participate in the prevention of fires are also the ones feeding the database. The database is also continuously updated and the final track record is stopped on the 31st of January every year. All those conditions made this database a very reliable one for me to collect my data from.
This dataset is composed of the ID of the fire forest, the type of fires - in this case the only type is forest-fire, the town in which it occurred, the size of the area that was burnt and what triggered it.

To enhance my analysis, I chose a second dataset to do some queries on in SQL. This second dataset came from the INSEE (Institut national de la statistique et des études économiques). This database provides information about each town in France, their population, their surface area, department, altitude, geolocation, etc.

Data collection

I wanted first to collect the data through the pandas read csv function. But after inspecting the data provided by the csv file on the Prométhée website I've realized that it was drastically different from what was shown on the website. For example, the column of causes was absent from the file. I therefore decided to collect the data through web scraping.

Data cleaning and Exploratory data analysis

To clean and explore the dataset about forest fires I started by importing the libraries needed to clean, explore and visualize : Pandas, Numpy, matplotlib and seaborn. I then inspected the data through pandas different features like ".info()", ".describe()", ".unique()", etc. and repeated this process throughout the data cleaning to make sure the data was ready for analysis.

I started inspecting my dataset by looking for missing values with ".isnull().sum()". At first glance the data doesn't have any missing values. But from what I saw on the Prométhée website a lot of the causes seemed to be listed under "-" when the cause was unknown. I decided to inspect this value .

```
forest.loc[forest['Cause'] == '-']
```

	ID	Type	Date	Department	Town	Area (ha)	Cause
0	983	Forêt	2020-03-31	2B	Vallecalle	0.3000	-
3	1036	Forêt	2020-04-02	7	Malarce-sur-la-Thines	0.0002	-
5	1397	Forêt	2020-04-02	11	Armissan	0.0300	-
6	1052	Forêt	2020-04-03	48	Saint-Pierre-des-Tripiers	2.1700	-
7	1045	Forêt	2020-04-03	7	Meyras	2.0000	-
...	...	...	...	...	...	...	...
4274	6218	Forêt	2022-08-30	7	Saint-Julien-Labrousse	0.05	-
4275	6113	Forêt	2022-08-31	13	Martigues	0.002	-
4276	5827	Forêt	2022-08-31	34	Courniou	0.2118	-
4277	5858	Forêt	2022-08-31	13	Marseille	0.015	-
4282	6114	Forêt	2022-09-01	13	Trets	0.03	-

1857 rows × 7 columns

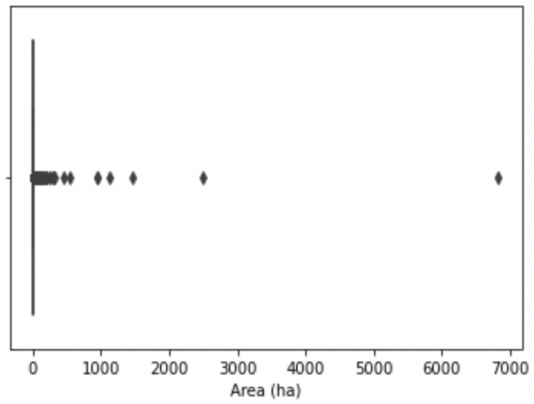
It turns out that 1857 out of 4288, or 43.3% of my data in the Cause column is missing.

I first try to inspect the missing values according to departments and towns. My first idea was to replace the missing values with the most frequent cause of forest fire (mode) of every department. After further inspection I realize that the process will not be worth it and will mess with the overall accuracy of the dataset. I decided to get rid of all the rows with missing values and create a new dataset from it.

Then, I inspect the burnt area column (Area (ha)). I realize that a lot of values are written in a strange orthograph, what was supposed to be "2 500 000" for example is "2\xa0500,0000". I realize that it is because those numbers are string objects. I start by normalizing them with a lambda function :

```
forest_cause['Area (ha)'] = forest_cause['Area (ha)'].apply(lambda x: str(x).replac'e(u'\xa0', u''))
```

I then convert the column to numeric values. Now that it is clean and numeric, I inspect it for outliers since it is the column most likely to have them.



I find outliers, but decide to keep them after some reflection and inspection of the data. Indeed, from a purely statistical perspective those values are outliers but from my analysis perspective they are useful as they indicate the extent of a burnt area in a specific department and commune.

Now that my data is cleaned I prepare it for its exportation to SQL. Since the database I'm going to join it with in SQL has the name of the communes in upper case I decide to make that transformation and export it.
Once the data is exported to SQL I encounter an unexpected issue ; SQL is not able to read my csv file because of the presence of french accents. I return to my python file and perform an extensive erasure of all accents :

```
forest_cause['Commune'] = forest_cause['Commune'].str.replace(u"Ê", "E")
forest_cause['Commune'] = forest_cause['Commune'].str.replace(u"É", "E")
forest_cause['Commune'] = forest_cause['Commune'].str.replace(u"È", "E")
forest_cause['Commune'] = forest_cause['Commune'].str.replace(u"Â", "A")
forest_cause['Commune'] = forest_cause['Commune'].str.replace(u"À", "A")
forest_cause['Commune'] = forest_cause['Commune'].str.replace(u"Ç", "C")
forest_cause['Commune'] = forest_cause['Commune'].str.replace(u"Î", "I")
forest_cause['Commune'] = forest_cause['Commune'].str.replace(u"Ï", "I")
forest_cause['Commune'] = forest_cause['Commune'].str.replace(u"Û", "U")
forest_cause['Commune'] = forest_cause['Commune'].str.replace(u"Ü", "U")
forest_cause['Commune'] = forest_cause['Commune'].str.replace(u"Ô", "O")

forest_cause['Cause'] = forest_cause['Cause'].str.replace(u"é", "e")
forest_cause['Cause'] = forest_cause['Cause'].str.replace(u"è", "e")
forest_cause['Cause'] = forest_cause['Cause'].str.replace(u"ê", "e")
```

The data is now ready for SQL and for the exploratory data analysis through visualization.

Hectares lost in forest fires between 2020 and 2022 in the mediterranean region

This past year has been a particularly destructive one when it comes to forest fires, I've noticed how the topic was very recurrent in the news during the summer. I decided to investigate if this year was significantly worse than the last one.
I first investigate the year between September 2021 and September 2022 :

```
#Total hectares lost in forest fires between Sept 2021 and Sept 2022 in the area
```

```
forest_2022 = forest_cause.loc[(forest_cause['Date'] >= '2021-09-01') & (forest_cause['Date'] < '2022-09-01')]
forest_2022['Area (ha)'].sum()
```

```
12208.140800000001
```

```
forest_2022['Area (ha)'].mean()
```

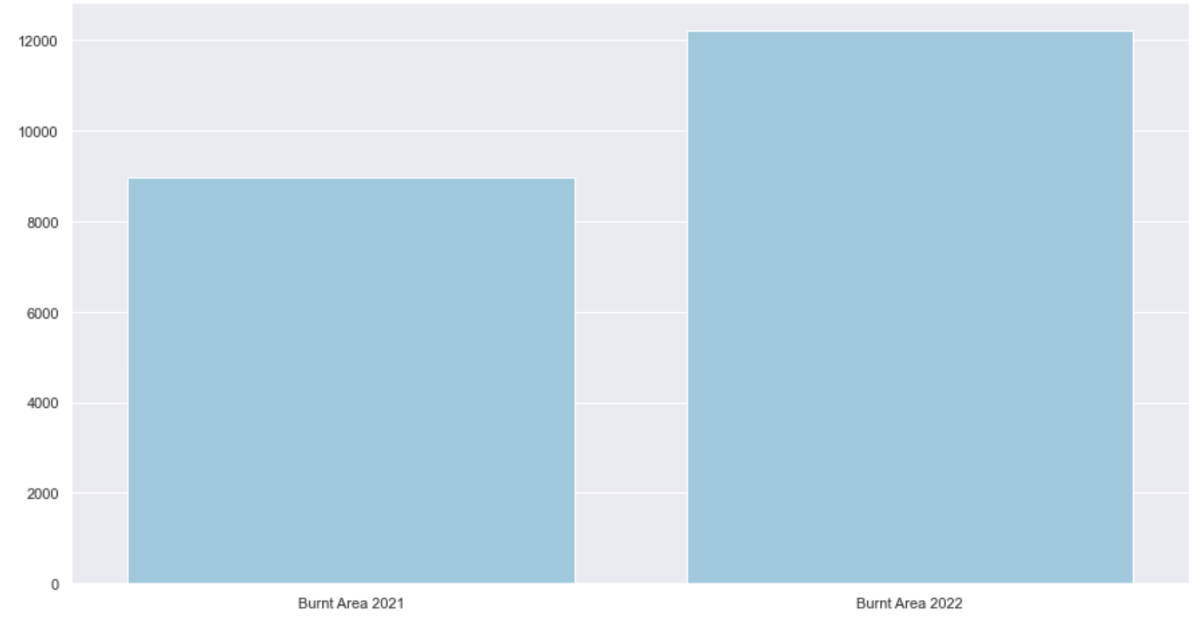
```
11.910381268292687
```

I then plot the results :

```
sum_burnt_2022 = forest_2022['Area (ha)'].sum()
sum_burnt_2021 = forest_2021['Area (ha)'].sum()
sum_burnt = pd.DataFrame({'Burnt Area 2021': sum_burnt_2021, 'Burnt Area 2022': sum_burnt_2022}
```

```
sns.set(style="darkgrid")
sns.set(rc = {'figure.figsize':(15,8)})
sns.barplot(data=sum_burnt, color="skyblue")
```

```
<AxesSubplot:>
```



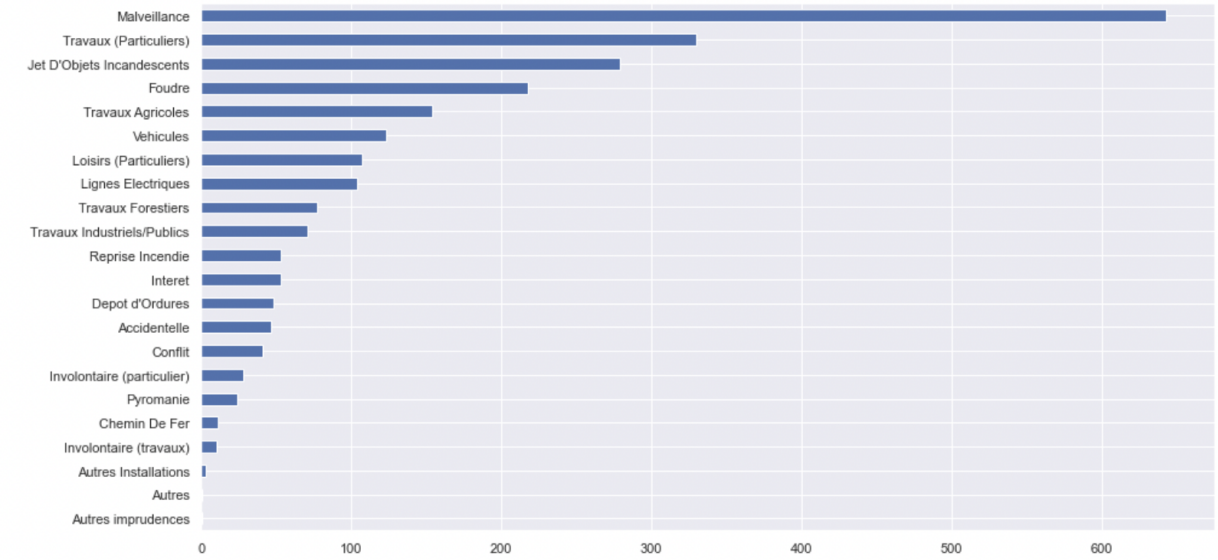
I can confidently conclude from this plot that the forest losses increase between 2021 and 2022.

Most frequent causes of fire

I then plot the most frequent causes of forest fire in the area between 2020 and 2022 :

```
forest_cause.Cause.value_counts().sort_values().plot(kind = 'barh')
```

```
<AxesSubplot:>
```

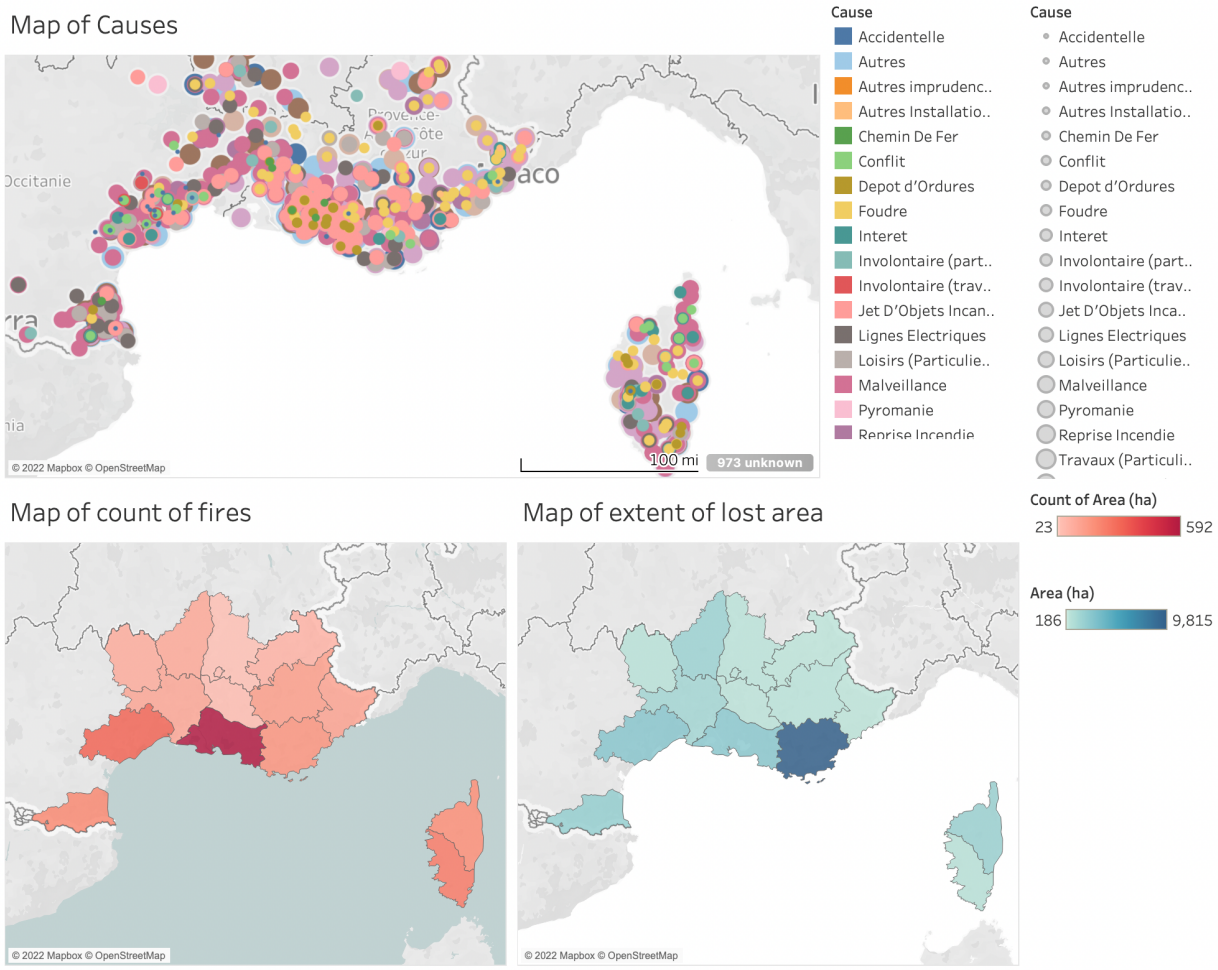


We can see from this graph that the most frequent cause is malicious intent, which means that the vast majority of forest fires are triggered by people who do it on purpose. At first I was confused as to what was the difference between malintent and pyromania was but I found an explanation in the french journal Marianne that explains in their article about this summer's forest fires that pyromania is a pathology that makes the pyromaniac seek the fires' spectacle at all cost, pushing them to set fire to forests. On the other hand malicious intent fires are set by people who were ill-intentioned in the first and decided for example to commit an act of revenge by setting fire to someone's house, bin, or enterprise, starting a fire that would spread to the forest.

Visualization in Tableau

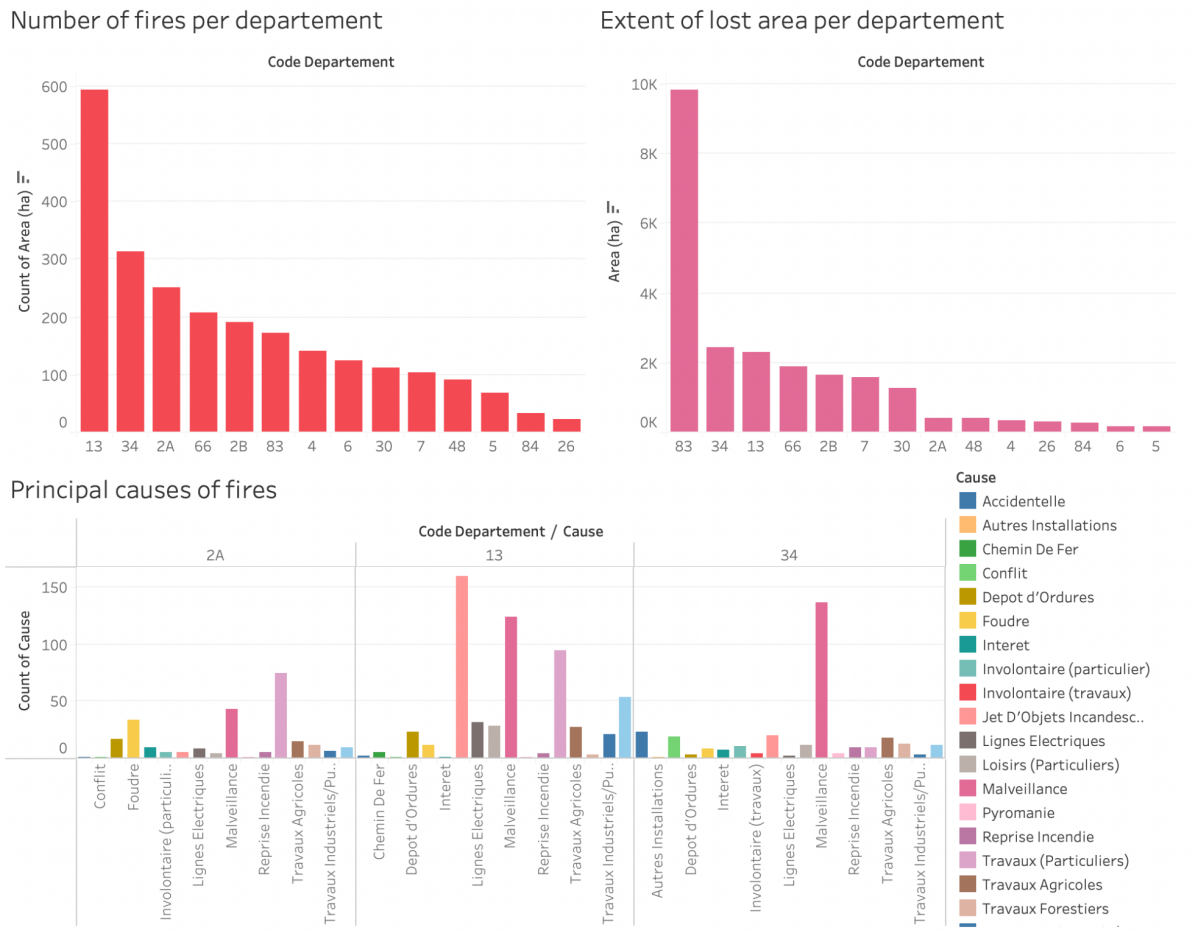
I then plotted more data in Tableau to get more insights from my data.

Map of the forest fires



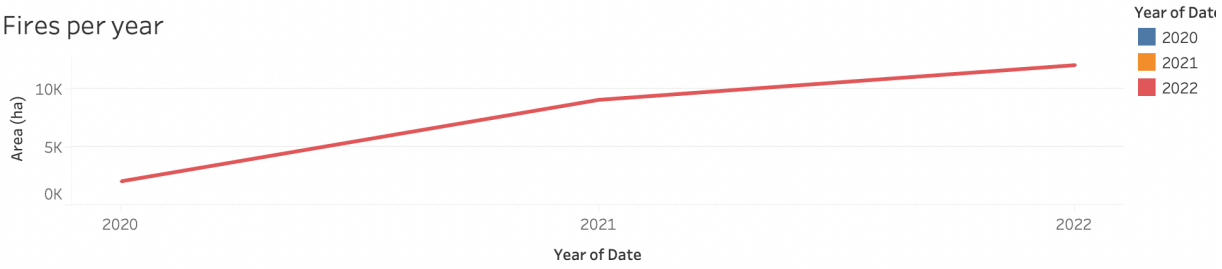
This dashboard gives some precious insights about the mediterranean area. The first map describes the principal causes of forest fires. The bigger the dot the more recurrent the cause, the causes also have their own color code. We can see how malicious-intent is a very widespread cause of fire, but also throwing of incandescent objects.
The map of count of fires shows the number of forest fires per department. We can see how the Bouche-Du-Rhône department was the most impacted in the last three years. The third map shows the extent of the burnt area per region. Here we can see how the Var department has lost the most hectares.

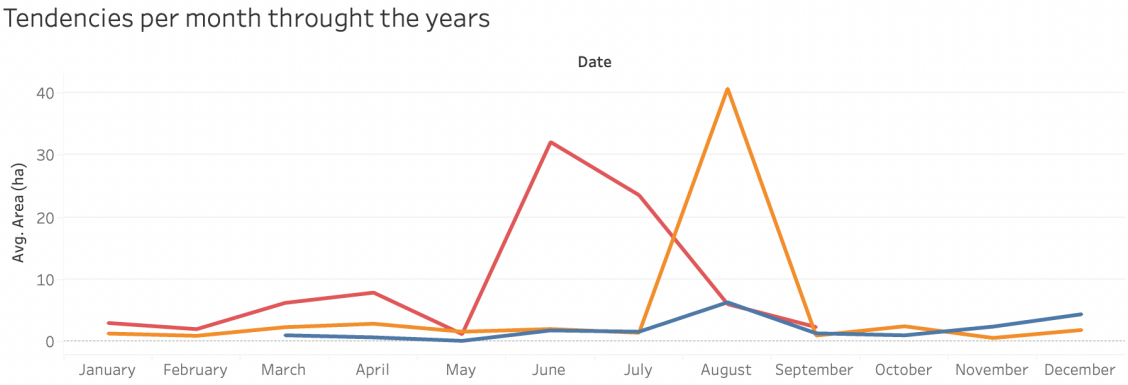
Bar plots of the forest fires



Then I plotted some bar plots to compare my data. In the first bar plot I displayed the number of fires per department to have a clear comparison of the most impacted departments. We can see in this graph that Bouche-du-Rhone, Hérault and Corse-du-Sud are the three most impacted departments. In the second plot I display the departments that lost the biggest portions of lands. In this graph the Var department gets the first place. This can be explained by two ravaging forest fires that destroyed 30 and 240 hectares. Lastly, in the third plot I decided to plot the principal reasons for the fires in the three most impacted departments. We can see that for the 13 (Bouche du Rhône) it is the throwing of incandescent objects that are the main causes, for the 2A (Corse-du-Sud) it is personal construction projects and in the 34 (Hérault) it is malicious intentions**.** Those plots offer precious information for authorities since they could use them to target prevention around the particular cause that impacts their department, for example legal consequences could be reinforced regarding the throwing of incandescent objects.

Graphs for time relations





Finally I decided to create some graphs to represent the time relations for the data. In the first graph we can see that throughout the years the number of forest fires has been drastically increasing. We can also see on the second graph that the worst month out of all the years was the month of August 2021, followed by the month of June 2022. We can also see on this graph that the forest fires started sooner and stronger this year, the consequences of climate change are more obvious this year. Finally, the last graph shows how the months of June and August have been the worst ones in the last three years. This information could be used by authorities to increase the alertness for the most risky departments during those periods.

Database selection

SQL :

- Small amount of data
- Fast updating
- Slower querying than NoSQL
- Object Oriented Programming unfriendly

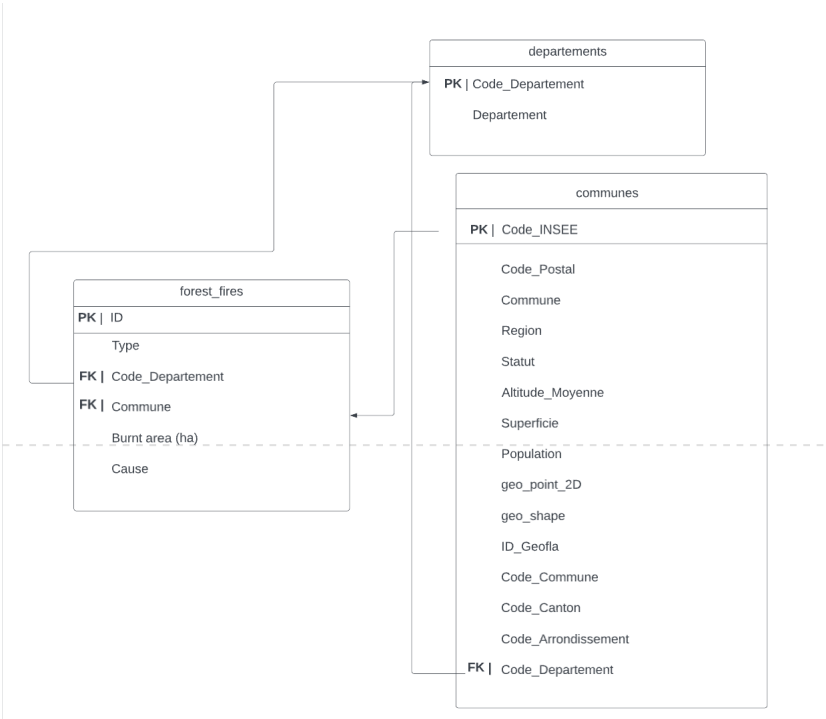
NoSQL :

- Big amount of data
- Slow updating
- Faster querying than NoSQL
- Object Oriented Programming friendly

After doing my data cleaning and visualization I had to decide on which database to choose to further my work. I decided to use SQL because it allowed me to work directly on my cleaned data, plus I only needed to perform simple queries. Even though NoSQL is more Object Oriented Programming friendly I knew that because I would only do simple queries I wouldn't need an OOP friendly approach.

Entities. ERD

This entity relationship diagram encompasses my three tables. The first table stems from the dataframe I've cleaned in Python, the second table was provided by the INSEE and shows various information about all the communes(towns) in France. It will help me to get more information about the communes present in my forest_fires table. From the table about communes I created a third table about departments only that will enable me to treat communes and departments separately in my queries.



Each table has its own primary key, which is unique. Each table also has foreign keys that will enable me to join them together. Code_Departement and Commune are the elements that are present in every table.

Database

```
9  ALTER TABLE df_forest_fire3
10 RENAME TO forest_fires;
11
12 ALTER TABLE forest_fires
13 RENAME COLUMN `Area (ha)` TO Area_ha;
```

I start by altering the tables to give them more insightful names, I also rename the columns that might be impractical in future queries.

I then create my departments table. For that I use columns that already existed in my communes table.

```
12  CREATE TABLE dep
13  AS (SELECT Code_Departement, Departement
14      FROM communes);
```

```
15  select dep.Code_Departement,dep.Departement, forest_fires.Commune,
16  forest_fires.Date, forest_fires.Area_ha, forest_fires.Cause
17  from forest_fires
18  left join dep
19  on forest_fires.Code_Departement=dep.Code_Departement;
20
```

100% 1:16

Result Grid Filter Rows: Search Export:

Code_Departeme...	Departement	Commune	Date	Area_ha	Cause
13	BOUCHES-DU-RHONE	FARE-LES-OLIVIERS (LA)	2020-04-02	0.1	Vehicules
13	BOUCHES-DU-RHONE	SAINT-CANNAT	2020-04-02	0.12	Travaux (Particuliers)
26	DROME	GRIGNAN	2020-04-03	0.3	Malveillance

I can now make my first queries. I start by joining the tables departments and forest_fires

Now I want to see the communes with the most population, so I join the table forest_fires and communes and order by population.

```
21  select forest_fires.Date, forest_fires.Commune, forest_fires.Area_ha,
22  communes.Population, communes.Altitude_Moyenne, communes.Code_Departement
23  from forest_fires
24  inner join communes
25  on forest_fires.Commune=communes.Commune
26  order by communes.Population DESC;
27  #Communes with the most population
```

100% 1:28

Result Grid Filter Rows: Search Export:

Date	Commune	Area_ha	Population	Altitude_Moyen...	Code_Departeme...
2021-06-12	BEZIERS	4.1598	71	45	34
2020-07-08	VALENCE	0.02	64.4	147	26
2020-07-28	HYERES	1.52	54.7	60	83
2020-07-03	ARLES	0.008	53	4	13

I notice that the four most populated towns are part of the most impacted departments by the forest fire. I can deduce that there is some correlation between a very populated commune and forest fires.

I now make a query to display the departments with the highest number of forest fires :

```
38  select communes.Code_Departement, communes.Departement, COUNT(DISTINCT(forest_fires.Area_ha)),
39  communes.Population, communes.Altitude_Moyenne
40  from forest_fires
41  inner join communes
42  on forest_fires.Commune=communes.Commune
43  group by  communes.Code_Departement,  communes.Departement, communes.Population, communes.Altitude_Moyenne
44  order by COUNT(DISTINCT(forest_fires.Area_ha)) DESC;
45  #Departement with the most fires
```

100% 1:46

Result Grid Filter Rows: Search Export:

Code_Departeme...	Departement	COUNT(DISTINCT(forest_fires.Area_...	Population	Altitude_Moyen...
13	BOUCHES-DU-RHONE	14	46.6	48
4	ALPES-DE-HAUTE-PROVENCE	8	17.2	868
30	GARD	7	0.6000000000000001	444
34	HERAULT	7	0.5	194

I now make a query to display the towns with the highest count of forest fires. I discover that Martigues was the most devastated town, followed by Digne-les-bains and Saint-Jean-de-la-Blaquiere. There is no correlation with the altitude.

```
29  select forest_fires.Commune, COUNT(DISTINCT(forest_fires.Area_ha)), communes.Population,
30  communes.Altitude_Moyenne, communes.Code_Departement
31  from forest_fires
32  inner join communes
33  on forest_fires.Commune=communes.Commune
34  group by forest_fires.Commune, communes.Population, communes.Altitude_Moyenne, communes.Code_Departement
35  order by COUNT(DISTINCT(forest_fires.Area_ha)) DESC;
36  #Towns with the most fires
```

100% 1:28

Result Grid Filter Rows: Search Export:

Commune	COUNT(DISTINCT(forest_fires.Area_...	Population	Altitude_Moyen...	Code_Departeme...
MARTIGUES	14	46.6	48	13
DIGNE-LES-BAINS	8	17.2	868	4
SAINT-JEAN-DE-LA-BLAQUIERE	7	0.5	194	34
SAINTE-CECILE-D'ANDORGE	7	0.6000000000000001	444	30
ALLAUCH	6	18.6	341	13
MIRAMAS	6	25.4	56	13

Finally I make a query to display the three tables joined together :

```
49  SELECT
50    dep.Code_Departement,
51    dep.Departement,
52    forest_fires.Commune,
53    forest_fires.Date,
54    forest_fires.Area_ha,
55    forest_fires.Cause,
56    communes.Altitude_Moyenne,
57    communes.Superficie,
58    communes.Population,
59    communes.geo_point_2d
60  FROM dep
61  JOIN forest_fires
62    ON dep.Code_Departement = forest_fires.Code_Departement
63  JOIN communes
64    ON forest_fires.Code_Departement = communes.Code_Departement;
65    #Joining all the tables together
```

Code_Departeme...	Departement	Commune	Date	Area_ha	Cause	Altitude_Moyen...	Superficie	Population	geo_point_2d
26	DROME	MOTTE-CHALANCON (LA)	2021-04-07	0.52	Travaux Forestiers	606	1582	0.2	44.72641161157575,5.304076177947098
26	DROME	PUY-SAINT-MARTIN	2021-04-04	2	Malveillance	606	1582	0.2	44.72641161157575,5.304076177947098
26	DROME	MONTGUERS	2020-08-26	289.39	Accidentelle	606	1582	0.2	44.72641161157575,5.304076177947098
26	DROME	MONTELIMAR	2020-08-25	3.96	Travaux Agricoles	606	1582	0.2	44.72641161157575,5.304076177947098

Conclusion

If the past few years have been essential regarding climate change awareness, this year in particular seems to have been deemed more alarming than the previous ones. This summer, all over French news were distressing footages of vast portions of forests going up in smoke. Witnessing the destruction of my country's environment made me want to investigate the extent of the damages through my newly acquired skills in data analytics.

This project was a very insightful one. I've discovered through my analysis that the first causes of forest fires were man-made, with ill-intentions being the most likely cause for a forest fire followed closely by the throwing of incandescent objects. It confirmed what I've learned in this article that forest fires had human origins in 90% of cases. My analysis also enabled me to spot which departments needed the most vigilance. The departments of the Corse-du-Sud, Bouche-du-Rhône and Herault are the most hit by forest fires. But in terms of lost area, the Var is the department that registered the biggest losses. The queries I've done in SQL also made me realize that it was, unfortunately, the most populated towns that were impacted by the forest fires. Through bar plots I was also able to notice what was the main cause of forest fires in the three departments that registered the most forest fires in those last three years. This information could be very valuable in implementing fire fighting strategies in those departments.

Resources

Prométhée forest fires database :
https://www.promethee.com/

INSEE's communes database :
https://public.opendatasoft.com/explore/dataset/correspondance-code-insee-code-postal/table/

"Incendies de l'été 90% des feux sont d'origine humaine", Marianne, le 11/08/2022 :
https://www.marianne.net/societe/ecologie/incendies-de-lete-90-des-feux-sont-dorigine-humaine