# Data Analytics 2022-2023 – Travail 1 – Feux de forêt aux U.S.A.

## Définition de l'objet de recherche

Y a-t-il une relation entre la taille d'un feu et sa cause ?

## Fouille de données

### Description du dataset

Le dataset utilisé provient de Kaggle mis en ligne il y a deux ans et intitulé « *U.S. Wildfire data (plus other attributes)* » et son auteur est l'utilisateur Capcloudcoder. Il s'agit d'un sous-dataset basé sur un dataset qui recense plus de 1.8 million de feux de forêts aux Etats Unis (échantillon de 50 000 feux pris aux hasard), combiné avec des informations météorologiques.

Le lien du dataset se trouve dans les sources.

### Description des données

Le dataset recense énormément d'informations, dont des informations qui ne nous intéressent pas. Par exemple des informations liées à la situation géographique, à la végétation, ou encore aux conditions météorologiques. Parmi les données récoltées, nous avons donc décidé de ne garder que les colonnes suivantes :

- Fire_size
- Stat_cause_descr
- Disc_clean_date

Que nous avons récupéré dans un nouveau dataset sous la forme suivante :

- Fire size : la taille du feu en hectares
- Cause : la cause du feu (parmi 11 causes différentes)
    - Incendie criminel
    - Combustion de débris
    - Feu de camp
    - Feux d'artifice
    - Enfants
    - Foudre
    - Fumeur
    - Utilisation d'équipement
    - Chemin de fer
    - Divers
    - Manquant/non défini
- Date : date de découverte du feu (MM/JJ/YYYY)

## Nettoyage des données

Pour exploiter les données nous avons dû effectuer quelques changements, tels que la conversion de la date par exemple. Nous avons également dû nous nous assurer que le dataset ne comportait pas de données manquantes et gérer les cas exceptionnels.

### *Gestion des données manquantes*

Nous nous sommes tout d'abord assurés que le dataset ne comportait pas de données manquantes qui pourraient altérer les résultats de nos tests. Nous nous sommes vite rendu compte qu'une portion des données était manquante (environ 6%).

```
#Pourcentage de data manquante
missing_values_count = fires.isnull().sum()
total_cells = np.product(fires.shape)
total_missing = missing_values_count.sum()
percent_missing = (total_missing/total_cells) * 100

percent_missing
```

```
5.948804194925951
```

Mais après une rapide inspection, nous nous sommes rendu compte que les colonnes concernées étaient des colonnes facultatives et que nous ne prendrions pas en compte lors de nos tests.

```
missing_values_count = fires.isnull().sum()
missing_values_count
```

```
Unnamed: 0             0
Unnamed: 0.1           0
fire_name          29454
fire_size              0
fire_size_class        0
stat_cause_descr       0
latitude               0
longitude              0
state                  0
disc_clean_date        0
cont_clean_date    27890
discovery_month        0
disc_date_final    26659
cont_date_final    29735
putout_time        27890
disc_date_pre          0
disc_pre_year          0
disc_pre_month         0
wstation_usaf          0
dstation_m             0
```

### *Gestion des données incohérentes*

Nous nous sommes également assurés qu'il n'y avait pas de données incohérentes (telles que des erreurs de typographie dans les causes des feux).

```
causes = fires['stat_cause_descr'].unique()
causes.sort()
causes
```

```
array(['Arson', 'Campfire', 'Children', 'Debris Burning', 'Equipment Use',
       'Fireworks', 'Lightning', 'Miscellaneous', 'Missing/Undefined',
       'Powerline', 'Railroad', 'Smoking', 'Structure'], dtype=object)
```

# Analyse

## Variable quantitative

Nous avons décidé d'analyser la variable fire_size (la taille du feu en hectares), qui semblait la plus pertinente à analyser compte tenu de notre question de recherche.

Pour ce faire, nous avons généré les indicateurs de statistique descriptive concernant cette variable.

| | Fire size |
|---|---|
| count | 55001.00000 |
| mean | 1957.92153 |
| std | 14368.30238 |
| min | 0.51000 |
| 25% | 1.17000 |
| 50% | 4.00000 |
| 75% | 20.00000 |
| max | 606945.00000 |

*Analyse et interprétation :*

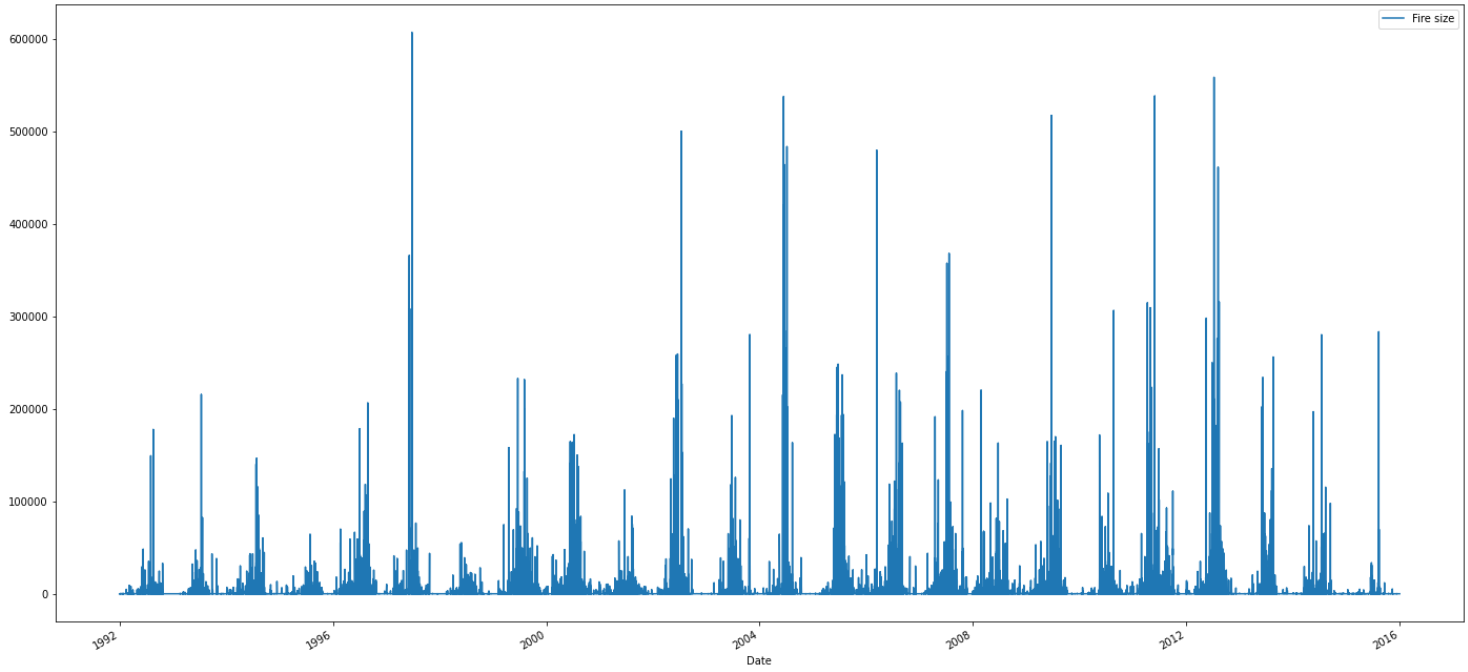
Parmi les 55000 feux compris dans le dataset, la taille moyenne d'un feu est de 19 hectares, pour un écart-type de 14 368. C'est un écart-type de taille importante, indiquant une grande variation entre la taille des différents feux. Cette différence en taille est observable sur le graphique ci-dessous, reprenant la taille des feux par année.

On peut voir que 75% des feux ont une taille inférieure ou égale à 20 hectares, là où la taille maximum d'un feu est de 607 000 hectares. Il y a donc une minorité de feux de taille importante.

On peut également constater que, la moitié des feux a une taille inférieure ou égale à 4, tandis que la moyenne des feux est de 20, ce qui nous indique que les données ne suivent pas une distribution normale.

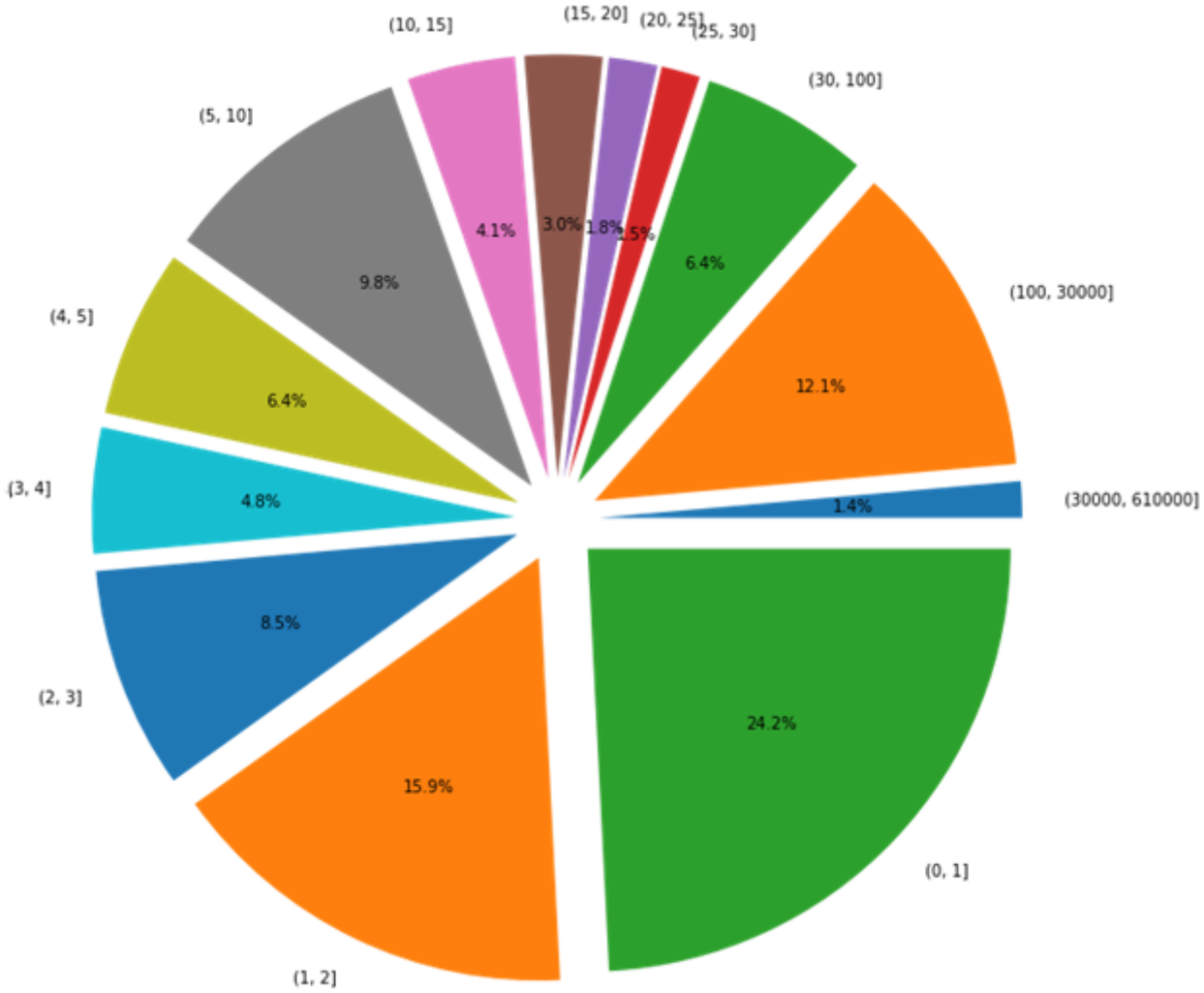
*Graphique 1 (linéaire) :*

Ce graphique linéaire permet de visualiser la différence d'écart-type observer ci-dessus. On peut également observer un schéma récurent : il semble y avoir une période de l'année où les feux sont de taille plus importante.



*Graphique 2 (circulaire) :*

On peut observer ici que les feux dont la taille est comprise entre 1 et 100 représentent 86,5% des données, tandis que les feux dont la taille est comprise entre 100 et 30 000 représentent 12,1% des données et ceux dont la taille est comprise entre 30 000 et 61 000 représentent seulement 1,4% des données. Cela est un indicateur de plus qui démontre que nos données ne suivent pas une distribution normale.



## Variable qualitative

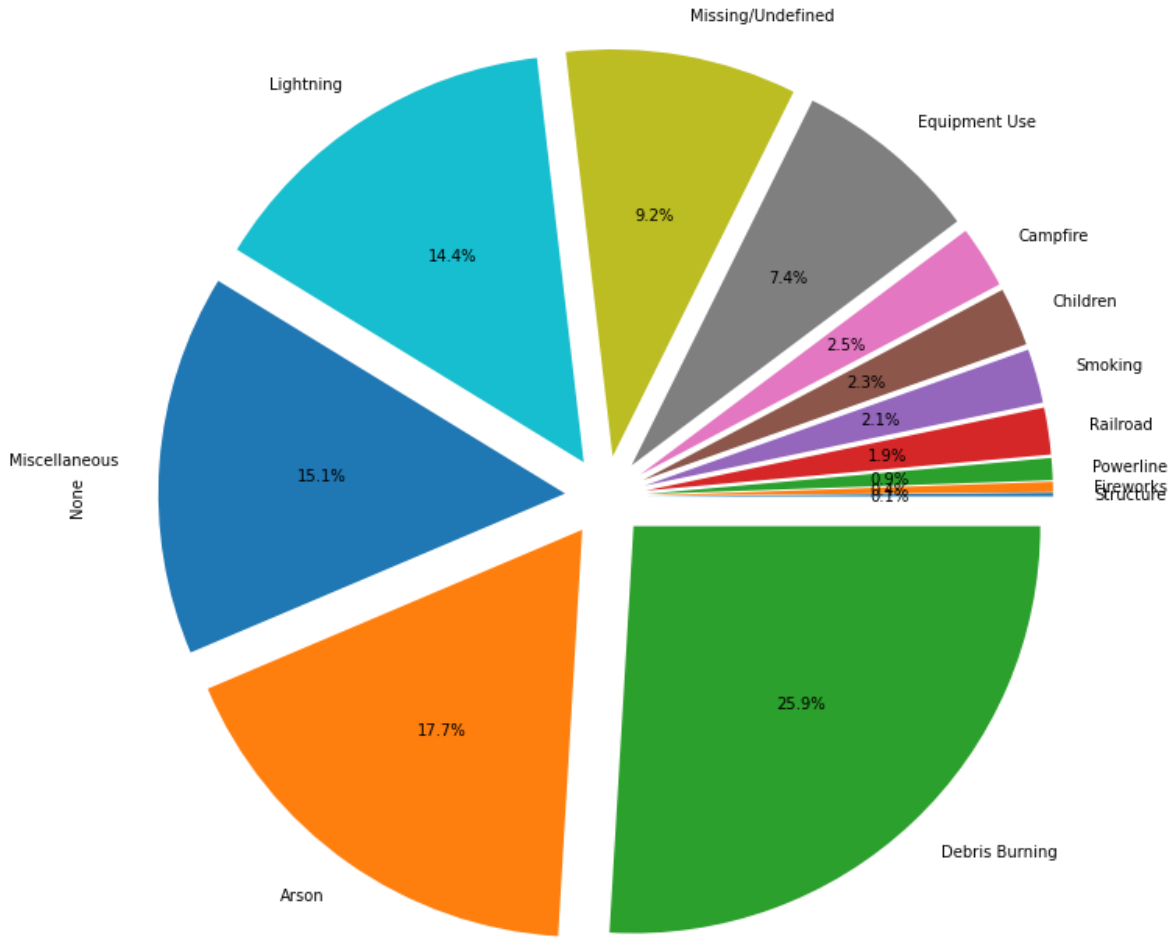
La variable concernée est la cause du feu.

Pour rappel, chaque cause du feu vient d'une liste de 11 cause différentes reprises ci-dessous :

- Incendie criminel
- Combustion de débris
- Feu de camp
- Feux d'artifice
- Enfants
- Foudre
- Fumeur
- Utilisation d'équipement
- Chemin de fer
- Divers
- Manquant/non défini

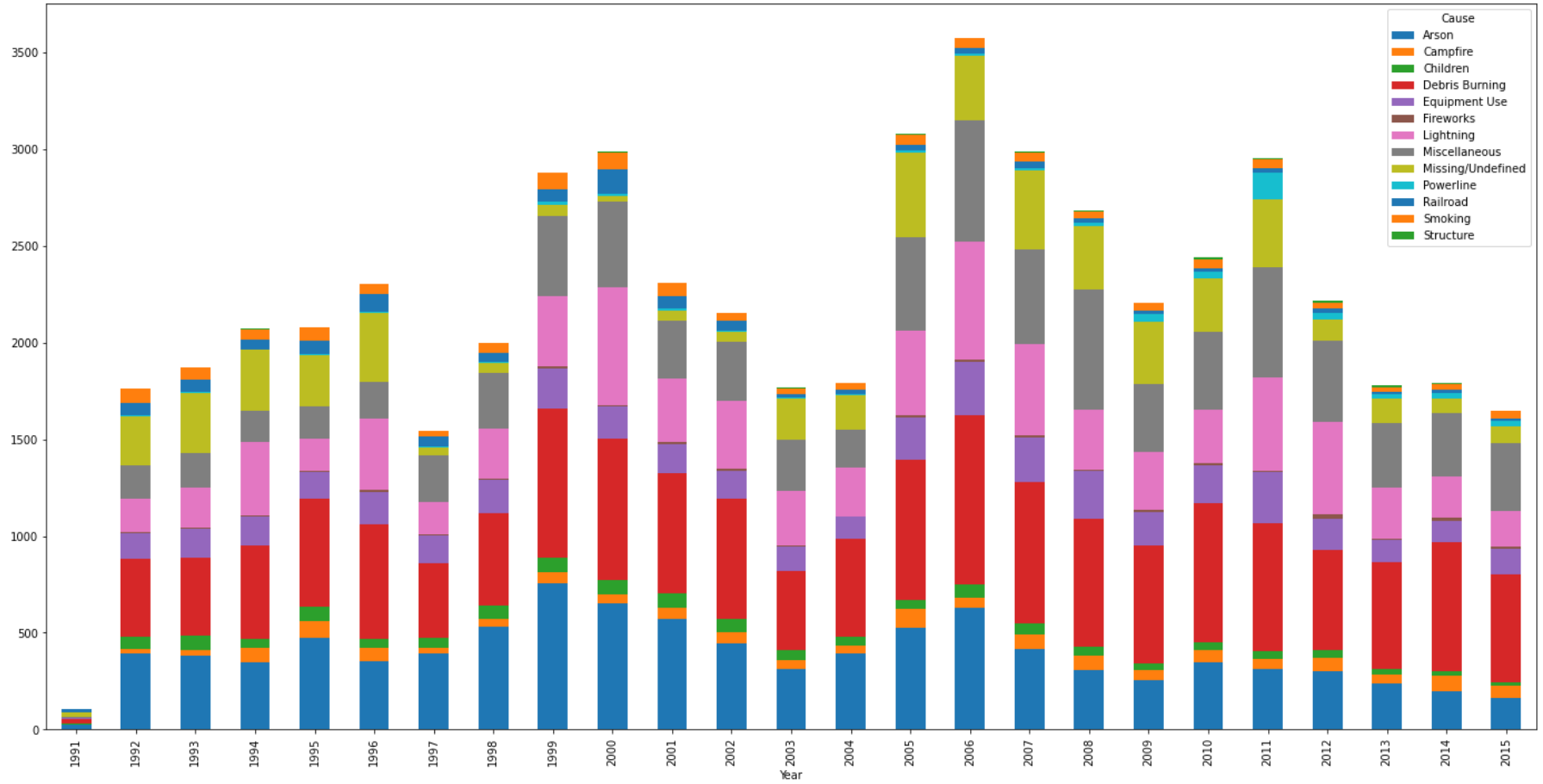
*Graphique 1 (circulaire) :*

Ce graphique nous permet de visualiser la proportion de feux en fonction de chaque cause. Les causes principales sont donc : la combustion de débris, les incendies criminels et les éclaires. Ce qui semble logique étant donné la nature moins contrôlable par l'Homme de ceux-ci en comparaison aux autres, tels que les feux de camp, les fumeurs ou les enfants.



*Graphique 2 (graphique en bâtonnets) :*

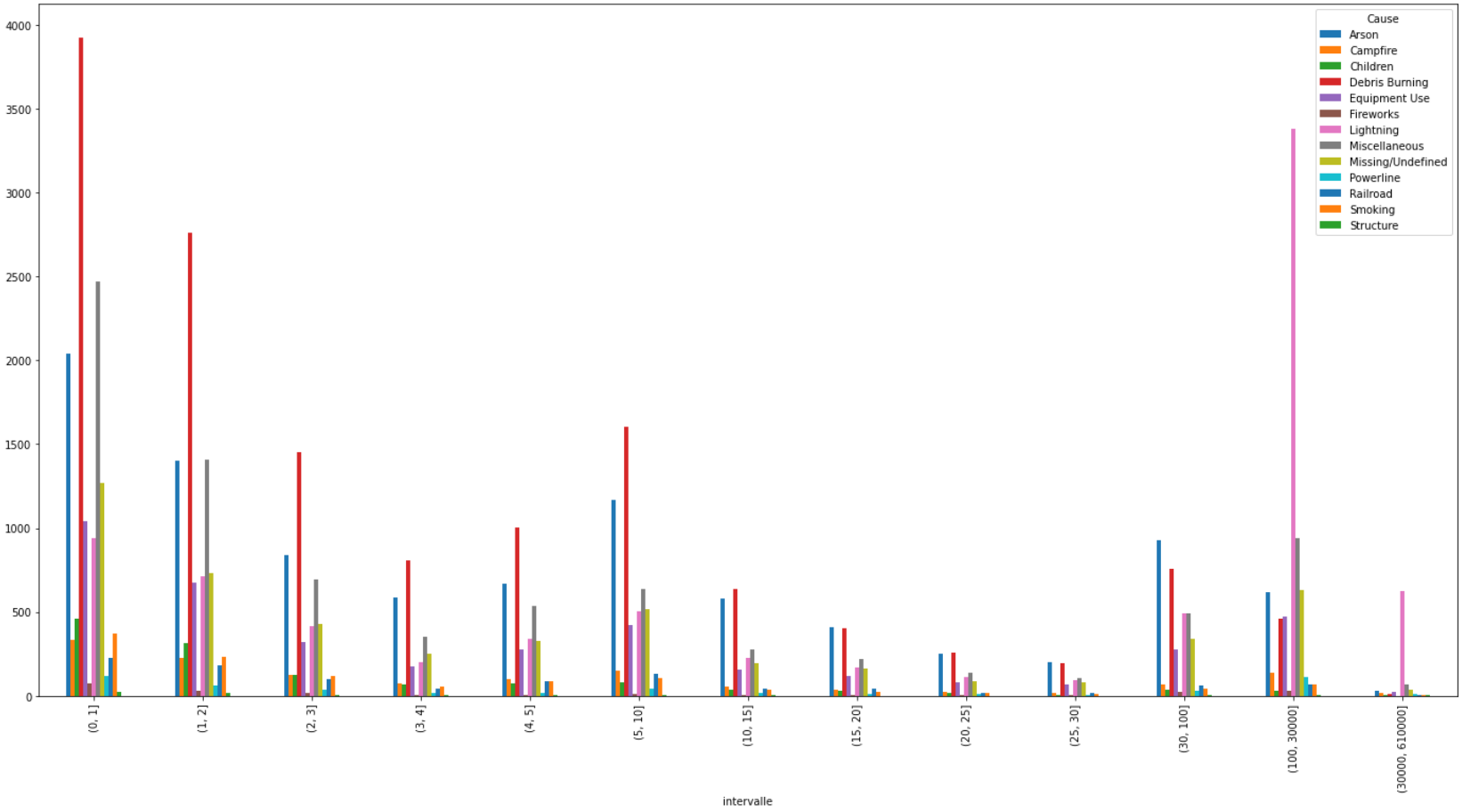
Ce graphique en bâtonnets nous permet de visualiser le nombre de feux par année et leurs causes. Nous pouvons constater que la proportion des causes reste relativement stable d'année en année. La cause qui varie le plus étant « Manquant/non défini ». Les autres causes semblent augmenter proportionnellement avec le nombre de feux. Il semble également être cohérent avec les données présentées dans le graphique circulaire ci-dessus.



## Tableaux croisés

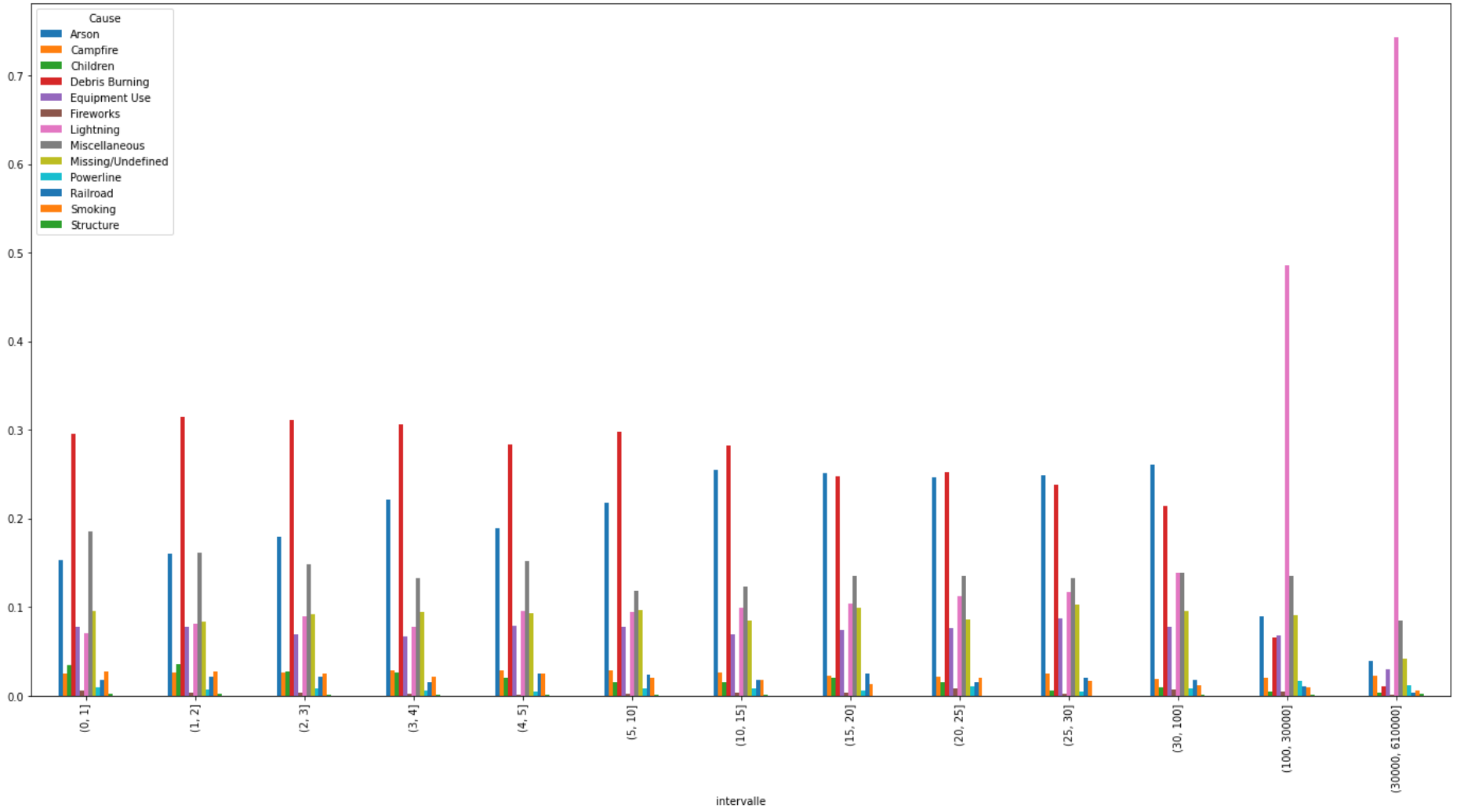
*Absolu*

Ici, nous pouvons voir en abscisse : les feux par intervalle de taille et en ordonnées la moyenne des tailles de ceux-ci. Le tout catégorisé par leur cause. Nous pouvons constater que la cause les plus récurrentes d'un feu change drastiquement en fonction de sa taille. On observe par exemple que la combustion de débris cause de feux qui, en moyenne, sont de tailles relativement petites. Tandis que les feux de tailles importantes sont en général causés par des éclairs. Une piste d'explication serait que les feux causés par la foudre sont moins contrôlables que ceux causés par des débris.



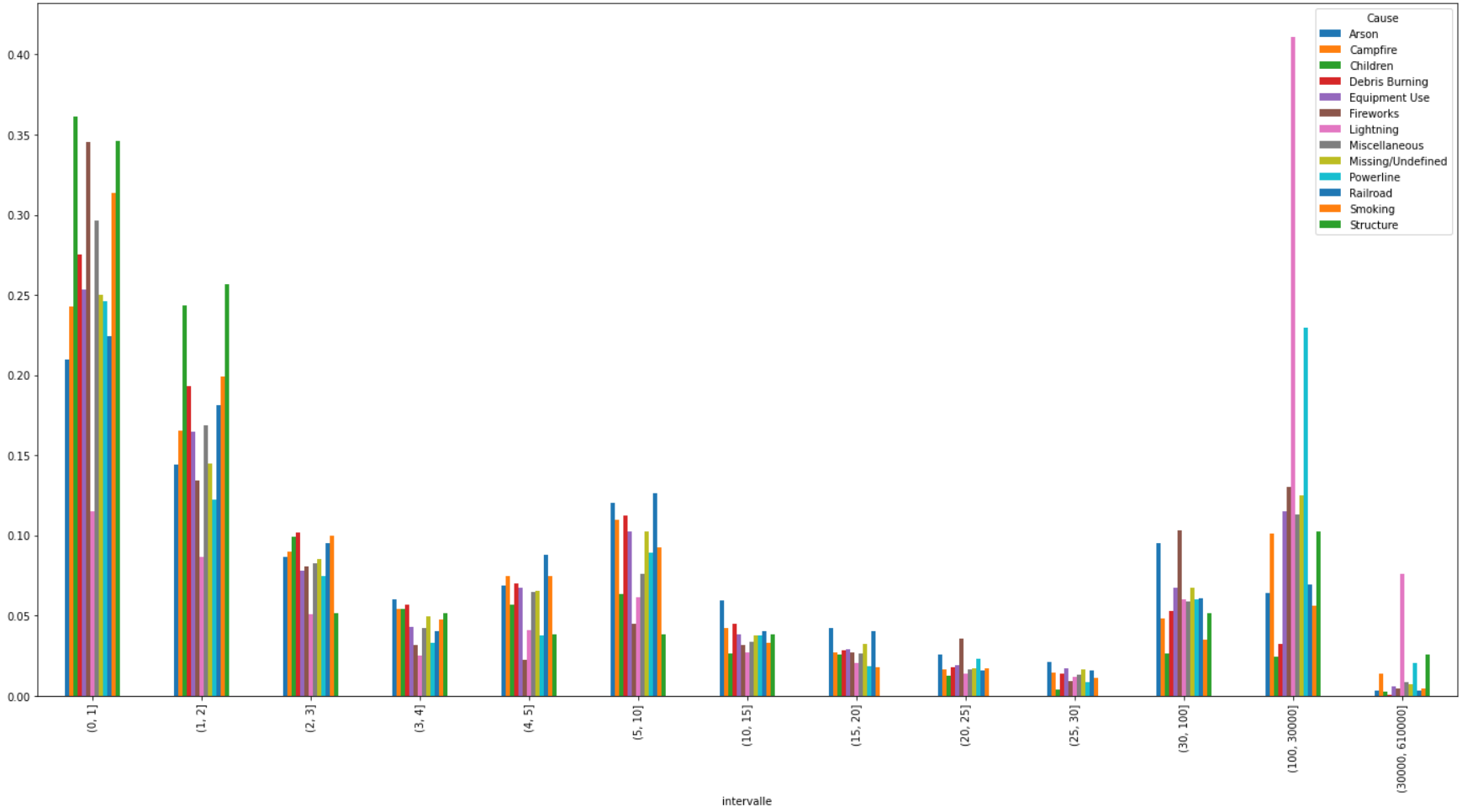
## *Relatif index*

Ce graphique lui, est normalisé sur l'index, il met de nouveau en évidence la relation entre la cause du feu et sa taille. Il représente pour chaque intervalle la proportion donc chaque cause est responsable.



*Index colonne*

Celui-ci est normalisé sur la colonne. Chaque barre indique la proportion d'une cause par intervalle de taille de feu. Il corrobore nos graphiques précédents et nous permet de visualiser le lien entre la taille d'un feu et sa cause, qui, pour certaines causes de feux, est flagrante.



## Test d'hypothèse

*Fixer les variables et les hypothèses*

Pour faire ce test, nous avons prélevé deux échantillons aléatoires dans le dataset, chacun de taille 500, afin d'effectuer un test de moyenne. L'un des dataset comprenant des feux ayant eu lieu en juin, juillet et août, et l'autre comprenant des feux ayant eu lieu le reste de l'année.

Rest of year :

| | fire_size |
|---|---|
| count | 500.000000 |
| mean | 22.740120 |
| std | 103.070703 |
| min | 0.510000 |
| 25% | 1.000000 |
| 50% | 3.000000 |
| 75% | 10.000000 |
| max | 1400.000000 |

June/Jully/August :

| | fire_size |
|---|---|
| count | 500.000000 |
| mean | 90.554320 |
| std | 928.235095 |
| min | 0.520000 |
| 25% | 1.000000 |
| 50% | 3.000000 |
| 75% | 10.000000 |
| max | 19520.000000 |

$H_0 = m_1 - m_2 = 0$ (La taille des feux reste similaire toute l'année).

$H_1 = m_1 - m_2 < 0$ (La taille des feux est plus grande en juin, juillet et aout que le reste de l'année).

Nous effectuerons donc un test unilatéral droit visant à tester l'égalité des moyennes des tailles de feu des deux échantillons.

*Préciser le seuil de signification alpha*

$\alpha = 0.05$

*Préciser la loi de probabilité*

Utilisation de la loi normale standardisée, $n$ étant supérieur à 30.

*Déterminer le seuil de rejet et la zone de non-refus*

Région de refus : ]1.645 ; +∞ [

Région de non-refus :] -∞ ; 1.645]

*Calculer la grandeur expérimentale*

$$\frac{d - E(d)}{\sqrt{var\ d}} = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} =$$

$$\frac{90.55 - 22.74 - 0}{\sqrt{\frac{928.23^2}{500} + \frac{103.07^2}{500}}} = 1{,}6239$$

*Conclusion*

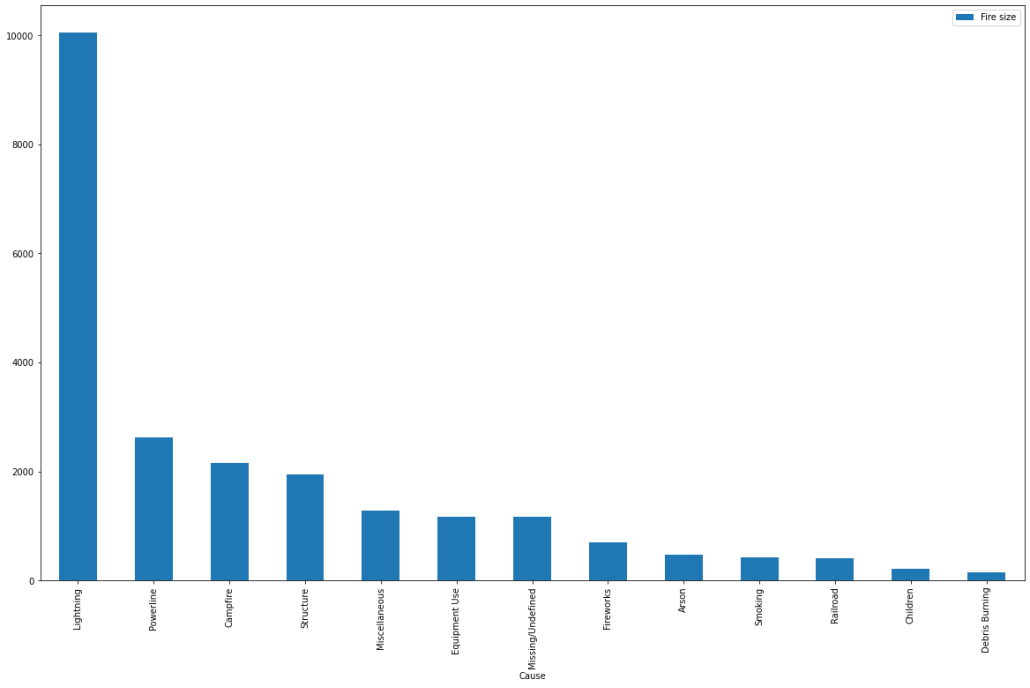
1,6239 appartient à la zone de non-refus, on ne peut donc pas rejeter $H_0$ au profit de $H_1$.

Au seuil de signification $\alpha$ = 0.05, la taille moyenne des feux en juin, juillet et août semble ne pas être plus grande que la taille moyenne des feux le reste de l'année.

## Réponse à la question de recherche

Après analyse de notre dataset, nous pouvons observer une relation entre la cause d'un feu et sa taille. Nous avons produit un dernier graphique, mettant en lien la taille moyenne d'un feu et sa cause. Les résultats sont cohérents avec ce que nous avons observé. Il y a également probablement un lien entre la contrôlabilité de la cause d'un feu et sa taille.

Le test d'hypothèse ne semble pas indiquer que les feux sont en moyenne plus gros en été, mais il semble y avoir quelques feux de grandes tailles en été, ce qui est cohérent avec l'écart type de la taille des feux qui y sont observés. Il pourrait être intéressant d'analyser le lien entre la période de l'année et le type de cause.



## Sources

Dataset (téléchargé le 5/10/2022) : https://www.kaggle.com/datasets/capcloudcoder/us-wildfire-data-plus-other-attributes

Short, Karen C. 2017. Spatial wildfire occurrence data for the United States, 1992-2015 [FPAFOD20170508]. 4th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.4

NOAA National Centers for Environmental Information (2001): Integrated Surface Hourly [1992-2015] - ftp://ftp.ncdc.noaa.gov/pub/data/noaa/

Meiyappan, Prasanth, and Atul K. Jain. "Three distinct global estimates of historical land-cover change and land-use conversions for over 200 years." Frontiers of Earth Science 6.2 (2012): 122-139.

"World Cities Database." Simplemaps, simplemaps.com/data/world-cities.