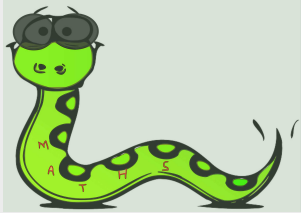

# Statistique en mouvement : leçons du passé et perspectives d'avenir

## Apports de la programmation en Python avec Pandas pour l'enseignement de la statistique et des probabilités dans le secondaire

Bro Frédéric

Lycée Henri-Moissan (Académie Créteil)

14 mars 2018


# Plan de la conférence

# Traiter des statistiques

**Objectifs :** A partir d'un jeu de données statistiques de **n'importe quelle taille**, on souhaite :

- Analyser et résumer
- modéliser
- prendre des décisions en statistique inférentielle

Avec Python, on utilisera le module **pandas**.
Voyons comment utiliser **pandas** au Lycée et dès la seconde.

# Créer des cartes dynamiques

- Les open data fournissent de plus en plus les coordonnées GPS.
- Grâce au module **folium** , on pourra représenter ces données sur des cartes « dynamiques » et donner plus de relief à la compréhension de ces statistiques.

`Petit Guide folium`

# Intérêt ?

## 1er exemple dans l'histoire de croisement de données + cartographie

- Le choléra sévit à Londres et atteint son pic en 1854.
- Le mode de propagation était inconnu.
- Sur même une carte, **John Snow** médecin et épidémiologiste pointa les adresses des morts du choléra et l'emplacement des pompes à eaux. `Résultat obtenu`
- En enlevant le bras de la pompe à **broad Street**, le nombre de morts diminua.
- La bactérie **Vibrio cholerae** ne se propageait pas par inhalation mais par ingestion !
  Ce ne fut que 30 ans plus tard que l'on isola cette bactérie et confirma la thèse de Snow.

# Installation de folium

Pour installer le module **folium**

## Utilisation de conda

1. Ouvrir Anaconda prompt
2. Écrire :

```
conda install -c conda-forge folium
```

Ou

## Utilisation de pip (via pyzo)

Ecrire dans le Shells :

```
pip install folium
```

# Plan de la conférence

1. Pandas & folium

2. **Présentation des activités**
   - Le titanic
   - Le pollen à Paris
   - Challenge Data Scientist
   - Les séismes dans le monde
   - Statistique sportive

# Fichier csv

- On trouve de plus en plus des données statistiques appelées **open data**.
- Elles sont souvent sous forme de tableaux, enregistrées au format **csv**.
- Pour séparer les éléments d'une même ligne du tableau, il est utilisé :
  - soit un point-virgule « ; »
  - soit une virgule « , »
  - soit un espace de tabulation « »

# Exemple

**stat.csv**

```
Valeurs ;Effectifs
1 ;4
2 ;6
3 ;2
4 ;3
```

Il correspond au tableau :

| Valeurs | Effectifs |
|---|---|
| 1 | 4 |
| 2 | 6 |
| 3 | 2 |
| 4 | 3 |

# Activité 1

## Le Tinanic

Le fichier **`titanic.csv`** contient les données relatives aux passagers ayant embarqué sur le Titanic, lors de son voyage inaugural (reliant Southampton à New-York en avril 1912).

**Objectifs :**

- Étudier la proportion de survivants par
  - classe
  - sexe
- La règle « les femmes et les enfants d'abord » fut-elle respectée ?

# Activité 2

## Le pollen à Paris

Tout au long de l'année, la Mairie de Paris assure la surveillance du patrimoine arboré et recense chacun d'eux dans le fichier csv :

**`Arbres.csv`**

Certains arbres provoquent des fortes allergies au pollen. Voici la liste de ces arbres :

- Frêne
- Olivier
- Noisetier de Byzance
- Bouleau
- Charme
- Cyprès
- Mûrier
- Platane
- Aulne

# Activité 2

**Objectifs :**

- Déterminer la proportions des arbres à fort potentiel allergisant.
- Conseiller les personnes ayant des allergies.

`Carte montrant la répartition des arbres allergisants`

# Activité 3

## Challenge Data Scientist

Le terme « data scientist » a été inventé par Dhanurjay Patil (LinkedIn) et Jeff Hammerbacher (Facebook) en cherchant comment caractériser les métiers des données pour afficher des offres d'emploi :

« Analyste, ça fait trop Wall Street ;
statisticien, çà agace les économistes ;
chercheur scientifique, çà fait trop académique.
Pourquoi pas "data scientist" ? »

# Activité 3

## Problématique

D'après l'article de challenge disponible via ce lien URL

https ://www.challenges.fr/economie/la-concurrence-inquietante-d-airbnb-pour-les-hoteliers_416014

on peut lire « La concurrence inquiétante d'AirBnb pour les hôteliers ».

Explorer, seul ou en groupe, le fichier :

**`Airbnb_Paris.csv`**

On pourra répondre aux questions suivantes :

# Activité 3

1. Lister les variables et déterminer leur type.
2. Calculer les indicateurs statistiques de chaque variable quantitative.
   Tous les logements sont ils loués ?
   Que dire des variables disponibilité et avis_par_mois ?
3. Étudier ce que rapporte un logement à l'année.
   *(On s'intéresse aux logements dont la disponibilité est ..... et le nombre d'avis est ..... )*
4. Où les hôtels ont-ils le plus de concurrence ?
   (Utiliser graphique(s), carte(s) pour répondre à cette question).

`Carte montrant la répartition des logements`

# Activité 4

## Les séismes de magnitude supérieure à 5

Le relevé de tous les séismes qui se sont produits dans le monde depuis ces 30 derniers jours, est accessible via l'URL :

http ://earthquake.usgs.gov/earthquakes/feed/v1.0/summary/all_month.csv

Pour chaque séisme, est noté :

- sa latitude et longitude (exprimées en degré)
- le temps (exprimé en UTC : Coordinated Universal Time)
  *Exemple :* `2018-02-21T17:59:34` `.564Z`

  $\Updownarrow$

  21 février 2018 à 17h59m34s
- sa magnitude (exprimée selon l'échelle de Richter)
- sa profondeur (exprimée en km)

# Activité 4

**Objectifs :**

- Analyser ce jeu de données.
- Modéliser le temps d'attente entre deux séismes de magnitude supérieure à 5.
- Calculer la proportion de séismes qui se sont produits ces 30 derniers jours, dans les pays de l'Asie du sud-est *(*Indonésie, Singapore, *etc)*.
  Cette zone appelée aussi **`Insulide`** correspond à :
  - une latitude allant de -13° à 15°
  - une longitude allant de 90° à 170°.

  Ce lieu est le carrefour de plusieurs plaques géologiques et est un lieu réputé sensible.
- Calculer la probabilité d'avoir au moins un séisme sur une période souhaitée.

`Carte montrant ces séismes`

# Activité 5

## Statistique sportive

En 2008, le cycliste américain Christian Vande Velde participa la 10-ième étape du tour de France.
Sur son vélo, une balise GPS a été fixée.
Cette balise enregistre régulièrement, durant la course :

- sa position GPS : latitude et longitude (exprimées en degré)
- son altitude (exprimée en mètre)

Les données sont collectées dans un tableau via le fichier **`velo.csv`**.

**Objectifs :**

Les axes d'études de cette statistique « vivante » seront :

- la distance parcourue par le cycliste
- sa vitesse
- l'évaluation de la pente lors de son déplacement

`Carte montrant le circuit`

# Installation de geopy

Pour installer le module **folium**

## Utilisation de conda

1. Ouvrir Anaconda prompt
2. Écrire :

```
conda install -c conda-forge geopy
```

Ou

## Utilisation de pip (via pyzo)

Ecrire dans le Shells de pyzo :

```
pip install geopy
```

# Merci pour votre attention.

**Sitographie :**


1. kaggle.com
2. https ://opendata.paris.fr/explore/dataset/les-arbres/
3. http ://insideairbnb.com/
4. http ://earth-quake.usgs.gov/earthquakes/feed/v1.0/summary/all_month.csv
5. https ://www.guidevtt.com/Membres/public_details_rando.php ?randoID=1396


**Je fais des stats...**