# 4DVST (Data Visualization Tools)

Jean-Baptiste, Miguel, Olivier

Table des matières

[Contexte du projet 1](#_Toc162517824)

[Description du cahier des charges 1](#_Toc162517825)

[Description du dataset 1](#_Toc162517826)

[Champs du dataset 2](#_Toc162517827)

[Composition de l’équipe 3](#_Toc162517828)

[Choix des technologies 3](#_Toc162517829)

## Contexte du projet

Dans la continuation de notre formation SUPINFO en Master 1 en spécialisation Data Engineer, il nous a été demandé d’accomplir un travail de visualisation de données, à la suite du cours qui nous a été dispensé à ce sujet.

## 

## Description du cahier des charges

Nous avons reçu un cahier des charges assez précis, mais qui laisse tout de même place à de la créativité et à de l’interprétation.

Nous avions quatre visualisations à faire, qu’il faut regrouper dans un dashboard, puis il a fallu faire un data storytelling qui interprète les visualisations de données.

## 

## Description du dataset

Un dataset nous à été fourni dans le contexte de ce projet. Il s’agit de données synthétiques de la production scientifique mondiale de 1996 à 2014.

### Champs du dataset

Le dataset comporte neuf champs qui vont être détaillés ci-dessous.

* Documents

Le nombre de documents produit durant l’année.

* Citable.documents

Le nombre de documents citables produit durant l’année.

* Citations

Le nombre de citations de documents durant l’année.

* Self.Citations

Le nombre de citations interne au pays (citation au sein d’un pays pour un document produit dans le même pays).

* Citations.per.Document

Le nombre moyen de citations d’un document durant l’année.

* H.index

Indice qui permet de quantifier la production scientifique d’un scientifique. Il est calculé comme ceci : un chercheur avec un indice de *h* a publié *h* articles qui ont été cités au moins *h* fois.

Il s’agit ici d’une moyenne du hindex de chaque scientifique d’un pays pour une année.

* Year

L’année.

* Rank

Rang du pays en termes de production de document par rapport aux autres pays du dataset pour une année.

* Country

Le nom du pays.

## Composition de l’équipe

L’équipe initiale était constituée de Jean-Baptiste et de Miguel, elle a été rejointe par Olivier en cours de projet.

En termes de distribution des rôles, Jean-Baptiste et Miguel ont collaborés sur les quatre visualisations ainsi que le dashboard. Olivier, quant à lui, avant de rejoindre l’équipe, à fait les quatre visualisations de son coté, pour ensuite rejoindre l’équipe et partager son travail avec les autres afin de s’en inspirer (ce travail n’a pas pu être intégré au projet car il est basé sur une technologie différente).

Le data storytelling ainsi que la documentation est le fruit de la collaboration des trois membres de l’équipe.

## Choix des technologies

Comme mentionné précédemment, il y a deux versions du projet, en ce qui s’agit des visualisations.

Les visualisations faites par Olivier, qui ne constituent pas ceux du projet final, ont été fait en Python avec les librairies numpy, pandas, geopandas et matplotlib. Ces technologies sont bien implantées dans le monde de la visualisation de données, ce qui fait qu’elles sont très bien documentées. En revanche, leur bémol est leur interface peu esthétique et la difficulté de la mise en place de dashboards.

Les visualisations gardées pour la version finale du projet ont été réalisés avec Python et les librairies streamlit, pandas et geopandas. En comparaison avec matplotlib, strealit propose une interface web épuré et donne la posibilité de regrouper les visualisations en dashboards facilement. Il nous a donc parut pertinent de garder cette version du projet plutôt que celle utilisant matplotlib.

Nous utilisons le gestionnaire de dépendances Poetry, qui nous permet de facilement déployer un projet identique sur différentes machines et de créer un manifest du projet qui liste les dépendances.