Collecte de Données et Data Visualization

Compréhension des objectifs, présentation de la démarche

Plan

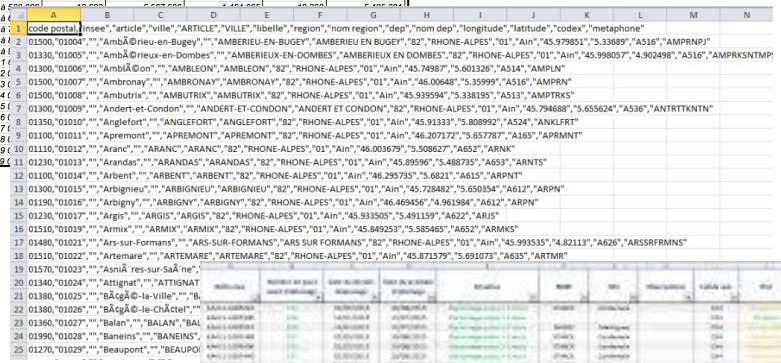
- Contexte & Définitions
- Objectifs
- Démarche & livrables
- Planning
- Organisation

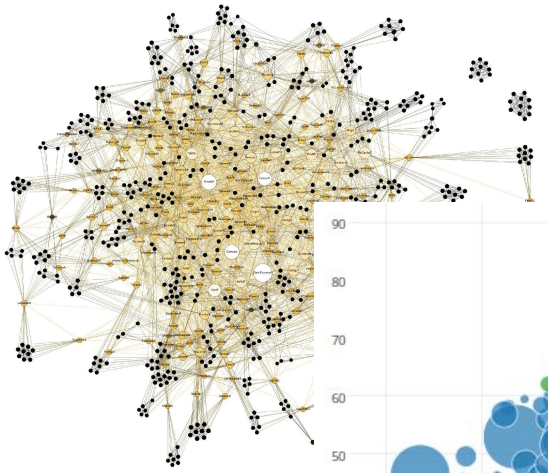
Contexte & Définitions

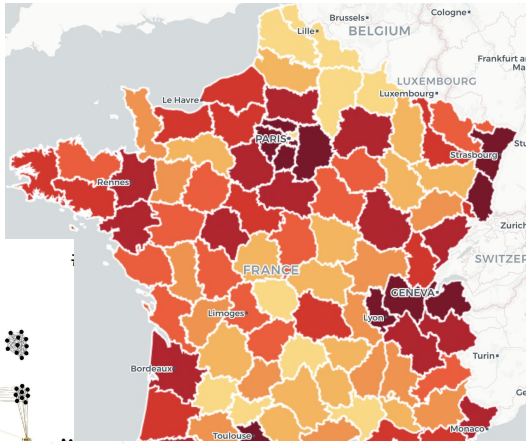
Data Visualization : représentation visuelle des données pour une **meilleure compréhension**

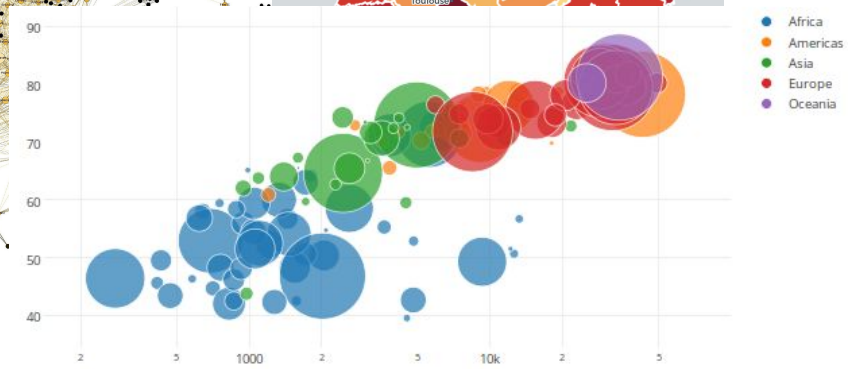
- Graphiques
- Cartographies
- Listes / Chronologie
- Organigramme

Revenu fiscal de référence par tranche (en euros)	Nombre de foyers fiscaux	Revenu fiscal de référence des foyers fiscaux	Impôt net (total)*	Nombre de foyers fiscaux imposés	Revenu fiscal de référence des foyers fiscaux imposés
0 à 10 000	8 779 578	37 017 353	-120 471	64 840	286 206
10 001 à 12 000	2 141 456	23 577 329	-52 668	5 831	63 719
12 001 à 15 000	3 415 487	46 459 161	-97 920	236 299	3 491 606
15 001 à 20 000	5 907 523	102 784 312	1 757 683	3 105 770	53 996 313
20 001 à 30 000	6 830 792	167 947 232	5 647 709	3 927 006	96 705 054
30 001 à 50 000	6 553 656	250 560 654	13 309 362	5 134 397	199 437 842
50 001 à 100 000	3 305 940	217 391 802	20 630 441	3 121 015	205 724 469
Plus de 100 000 dont:	749 163	140 216 578	28 027 792	727 817	136 965 445
100 001 à 200 000	*597 946*	*78 515 622*	*12 812 674*	*579 100*	*76 088 011*
200 001 à 300 000	*88 183*	*21 094 944*	*4 697 851*	*86 452*	*20 685 165*
300 001 à 400 000	*28 243*	*9 670 624*	*2 422 619*	*27 854*	*9 537 660*
400 001 à 500 000					
500 001 à 6					
600 001 à 7					
700 001 à 8					
800 001 à 9					
900 001 à 1					
1 000 001 à 2					
2 000 001 à 3					
3 000 001 à 4					
4 000 001 à 5					
5 000 001 à 6					
6 000 001 à 7					
7 000 001 à 8					
8 000 001 à 9					
Plus de 9					







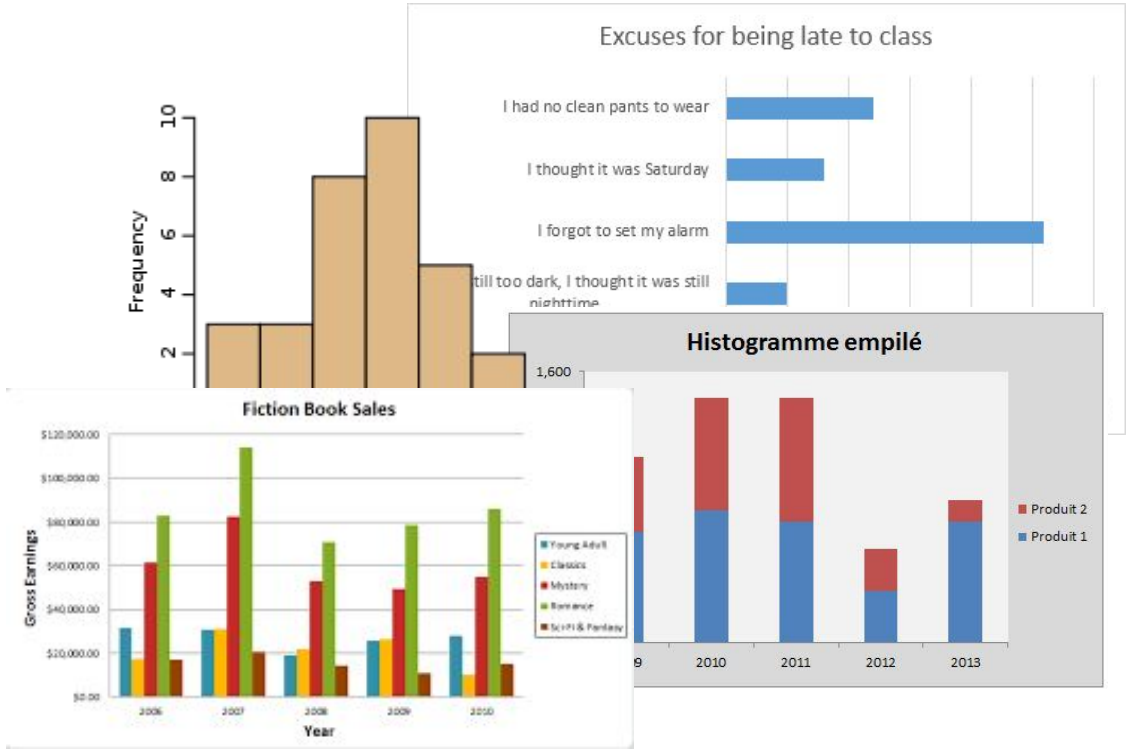


Contexte & Définitions

Limites :

➢ Multitudes de graphiques disponibles et imaginables en fonction du jeu de données

➢ Nommage de ces graphiques non arrêté :
- un graphe renvoie à plusieurs mot-clés (histogram / bar chart)
- un mot-clé renvoie à plusieurs graphiques (bar chart)



Comment peut-on labéliser / normer les visuels de data-visualization ?

Objectifs

Collecter un jeu de données de data vizualisation et proposer une **méthode de classification** automatique

- Créer un jeu de données
- Classifier les visuels
- Automatiser la classification (ML)
- Etendre l'étude à de plus grands jeux de données

Démarche & Livrable

1

Créer un jeu de données

Collecter un jeu de données

Script de scraping Google Image

Connection API Bing

Obtenir un jeu de données propre

Sélection d'image

Redimensionnement, cropper

2

Classer les visuels

Sélectionner des visuels

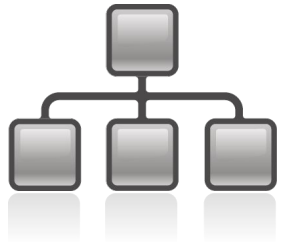
Recherche des visuels les plus récurrents

Sélection de visuels analysables (pour ML)

Annoter les visuels

Définition de labels précis

Nommer les classes



3

Méthode de classification automatique (Machine Learning)

Rechercher les algorithmes existants (Google, etc.)

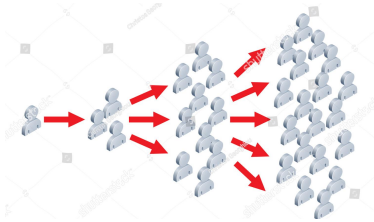
Choix de l'algorithme et des méthodes les plus pertinents (KMeans,..)
Définition des seuils de validation

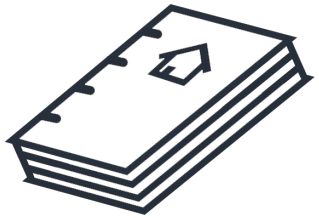
Classification binaire & classification multi-catégories

Démarche & Livrable

4

Étendre à de plus grands jeux de données, rendre public les résultats





Livrables :

- Une base de données da data visualizations
- Un programme de classification des data visualizations
- Une documentation explicative du projet détaillant:
  - La démarche du projet
  - Le fonctionnement du programme

Important et limites possibles:

identifier l'aspect légal du traitement des images, notamment par rapport aux droits des images

Planning

Décembre 2017	Jusqu'au 18 janvier	Jusqu'au 30 janvier	Jusqu'au 15 février	Jusqu'au 15 mars	Jusqu'au 29 mars
Constitution de l'équipe et prise de contact avec le tuteur	Phase de cadrage : définition des rôles, des livrables, des méthodes de travail Identification des problématiques non-techniques pouvant impacter le projet(légales par exemple) ***Premier reporting le 18 janvier***	Constitution de la base de données de data vizualisation et premiers traitements en vue de l'obtention du jeu de données propres ***Objectif : premier livrable pour le 30 janvier***	Classification des visuels sur la base d'une méthodologie de sélection et de normage ***Second reporting le 15 février***	Application des méthodes de machine learning afin d'obtenir notre programme de classification de data vizualisation adapté au grand jeu de données ***Objectif : second livrable pour le 15 mars***	Préparation de la documentation, préparation de la présentation finale, rattrapage d'un éventuel retard et correction des possibles défauts ***Troisième reporting le 29 mars***

Organisation

Chef de projet

Alexis MARTIN - DELAHAYE

Fonctionnel

Marc ARNAL

Dev Team

Luca GUÉRY

Louis KRAEMER

Arnaud BRUGIERE

Des questions ?