Laurent Jégou
Dépt. de Géographie, UT2J
jegou@univ-tlse2.fr

5 oct 2020

L3 APTER

UE 506 -Informatique

2.1.2 Acquisition par importation de fichier texte

Lorsque les données ne sont pas fournies dans un format de tableur, il faut les importer pour les convertir. Très souvent le format d'importation est un format texte, dans lequel les lignes et les colonnes du tableau sont identifiées par certains caractères.

Les formats texte de tableau de données présentent les paramètres de lecture suivants :

- largeur fixe (rare et ancien) ou délimité
- si délimité, type de séparateur (délimitation) :
  - de champ (colonne)
  - de contenu texte, généralement des guillemets (on trouve aussi des apostrophes)

Le fait de préciser si les champs détectés dans le fichier d'importation sont de type texte ou standard (numérique) permet d'éviter d'avoir à le faire ensuite.

- Télécharger le fichier du nombre de logements commencés entre 2009 et 2013 en Haute-Garonne, par canton, sur le site de l'Open-Data régional, au format CSV :

https://www.data.gouv.fr/fr/datasets/logements-commences-par-canton-de-2009-a-2013/

Aussi disponible sur GéoTests : http://www.geotests.net/cours/apter/l3/donnees/logements-commences-par-canton-de-2009-a-2013.csv

Extrait du fichier csv :

```
Code canton;Canton;Année;Nombre;Surface (m²);Taille Moyenne (m²)
3101;AUTERIVE;2009;361;42 562;118
3104;CASTANET-TOLOSAN;2009;192;22 133;115
3105;CASTELGINEST;2009;193;19 163;99
3114;SAINT-GAUDENS;2009;113;14 222;126
3123;TOULOUSE-9 (canton partiel);2010;87;6 118;70
3126;TOURNEFEUILLE;2010;321;27 266;85
```

Le format CSV est à l'origine un format dans lequel le séparateur est une virgule ("*comma separated values*", en anglais). Le souci c'est qu'en France, la virgule sert de séparateur décimal. Ainsi, dans la version française d'Excel, le format CSV est censé

comporter des points-virgules comme séparateurs. En cas de doute, choisir le format « texte simple » qui offre le choix du séparateur dans l'étape suivante de l'assistant.

Dans Excel, il faut choisir Fichier / Nouveau, puis Fichier / Importer, ensuite, dans l'assistant, choisir les options suivantes :

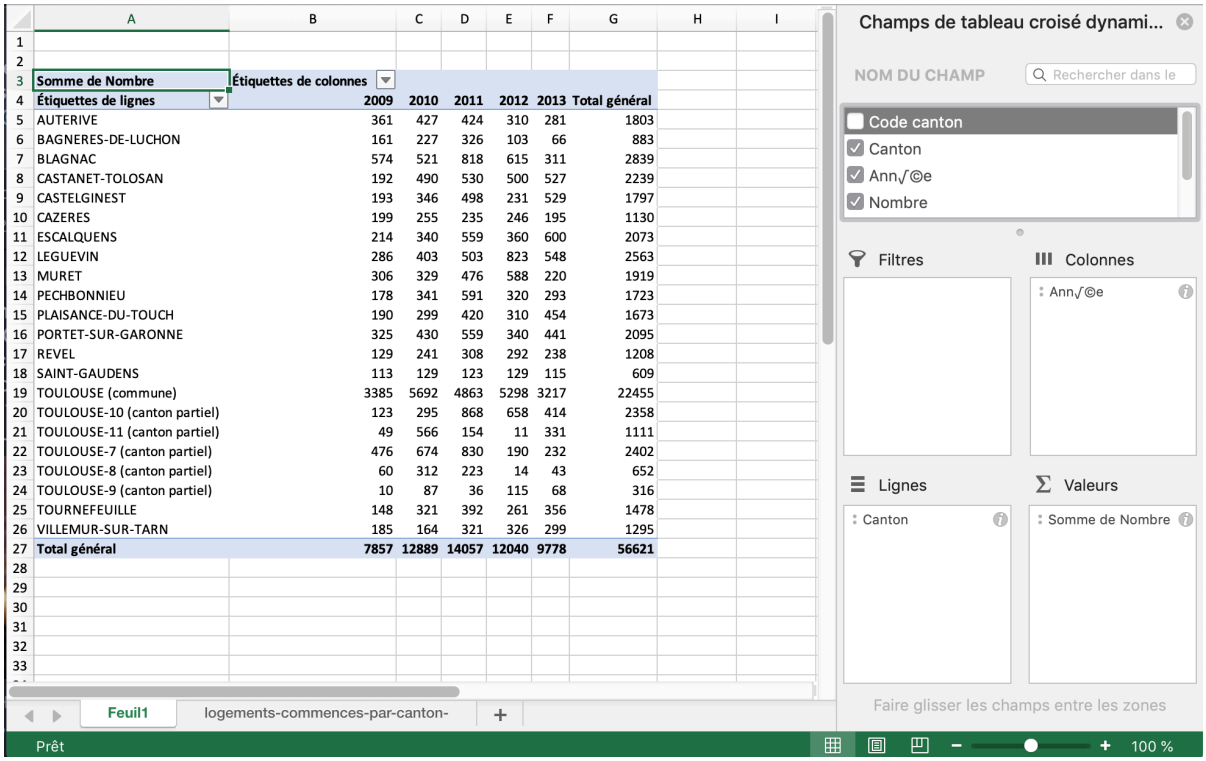
- format délimité
- origine du fichier (codage des caractères) : UTF-8
- séparateur point-virgule
- les champs Code_Canton et Canton doivent être définis au **format « texte »** (et non « standard », qui correspond à une valeur numérique).

3. Traitement et analyses de données simples

En partant des données que nous venons d'importer, nous allons réaliser quelques traitements et analyses simples, pour y voir plus clair concernant cette thématique dans ce département, durant cette période.

Tout d'abord, comment présenter ces données sous la forme d'un tableau des nombres de logements avec les cantons en lignes, les années en colonnes ? Comment visualiser les valeurs des cantons et leurs évolutions ?

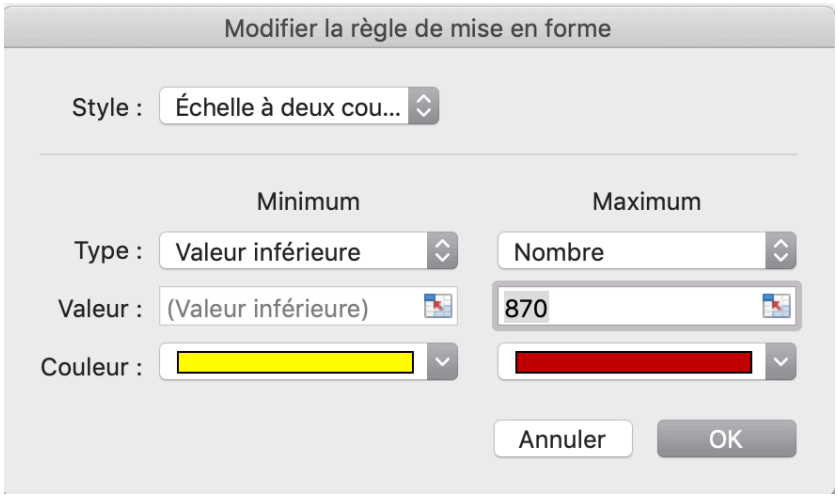
→ Un tableau croisé dynamique (Menu "Insertion" puis "Tableau croisé dynamique")



Somme de Nombre	Étiquettes de colonnes					
Étiquettes de lignes	2009	2010	2011	2012	2013	Total général
AUTERIVE	361	427	424	310	281	1803
BAGNERES-DE-LUCHON	161	227	326	103	66	883
BLAGNAC	574	521	818	615	311	2839
CASTANET-TOLOSAN	192	490	530	500	527	2239
CASTELGINEST	193	346	498	231	529	1797
CAZERES	199	255	235	246	195	1130
ESCALQUENS	214	340	559	360	600	2073
LEGUEVIN	286	403	503	823	548	2563
MURET	306	329	476	588	220	1919
PECHBONNIEU	178	341	591	320	293	1723
PLAISANCE-DU-TOUCH	190	299	420	310	454	1673
PORTET-SUR-GARONNE	325	430	559	340	441	2095
REVEL	129	241	308	292	238	1208
SAINT-GAUDENS	113	129	123	129	115	609
TOULOUSE (commune)	3385	5692	4863	5298	3217	22455
TOULOUSE-10 (canton partiel)	123	295	868	658	414	2358
TOULOUSE-11 (canton partiel)	49	566	154	11	331	1111
TOULOUSE-7 (canton partiel)	476	674	830	190	232	2402
TOULOUSE-8 (canton partiel)	60	312	223	14	43	652
TOULOUSE-9 (canton partiel)	10	87	36	115	68	316
TOURNEFEUILLE	148	321	392	261	356	1478
VILLEMUR-SUR-TARN	185	164	321	326	299	1295
Total général	7857	12889	14057	12040	9778	56621

→ Une mise en forme conditionnelle pour colorer les cellules. (Menu « Accueil », puis « Mise en forme conditionnelle »).

On choisit une "échelle à deux couleurs" , c'est-à-dire un dégradé de couleurs entre deux valeurs, avec la valeur minimale en jaune clair et 870 pour la valeur maximale (ce qui permet d'éviter que les chiffres de Toulouse commune ne distordent le dégradé) en rouge foncé.



En modifiant la couleur du texte des cellules foncées, la lisibilité est améliorée :

Somme de Nombre	Étiquettes de colonnes					
Étiquettes de lignes	2009	2010	2011	2012	2013	Total général
AUTERIVE	361	427	424	310	281	1803
BAGNERES-DE-LUCHON	161	227	326	103	66	883
BLAGNAC	574	521	818	615	311	2839
CASTANET-TOLOSAN	192	490	530	500	527	2239
CASTELGINEST	193	346	498	231	529	1797
CAZERES	199	255	235	246	195	1130
ESCALQUENS	214	340	559	360	600	2073
LEGUEVIN	286	403	503	823	548	2563
MURET	306	329	476	588	220	1919
PECHBONNIEU	178	341	591	320	293	1723
PLAISANCE-DU-TOUCH	190	299	420	310	454	1673
PORTET-SUR-GARONNE	325	430	559	340	441	2095
REVEL	129	241	308	292	238	1208
SAINT-GAUDENS	113	129	123	129	115	609
TOULOUSE (commune)	3385	5692	4863	5298	3217	22455
TOULOUSE-10 (canton partiel)	123	295	868	658	414	2358
TOULOUSE-11 (canton partiel)	49	566	154	11	331	1111
TOULOUSE-7 (canton partiel)	476	674	830	190	232	2402
TOULOUSE-8 (canton partiel)	60	312	223	14	43	652
TOULOUSE-9 (canton partiel)	10	87	36	115	68	316
TOURNEFEUILLE	148	321	392	261	356	1478
VILLEMUR-SUR-TARN	185	164	321	326	299	1295
Total général	**7857**	**12889**	**14057**	**12040**	**9778**	**56621**

Ici, nous avons des chiffres bruts de nombre de logements : comment comparer les cantons d'une année sur l'autre ? Comment comparer les cantons entre eux ?

Ensuite, comment se faire une idée rapidement, obtenir une synthèse, résumant la situation ? Quels indicateurs utiliser ?

3.1 Indicateurs synthétiques

En statistiques descriptives, les indicateurs peuvent se regrouper en deux types : de centralité et de dispersion. Pour analyser plus finement la situation récente de la construction de logements en Haute-Garonne, nous allons travailler sur un fichier par communes, entre 2010 et 2018, issu de cette source : https://www.data.gouv.fr/fr/datasets/statistiques-sur-les-permis-de-construire-pc-permis-damenager-pa-et-declaration-prealable-dp-base-sitadel/#_

Ce site permet de télécharger les données année par année. Pour simplifier le travail, a été préparé un fichier synthétique regroupant plusieurs variables : le nombre total de logements *commencés*, leur superficie totale, ces mêmes valeurs pour les types de

logements : individuels purs, individuels regroupés (lotissements), collectifs, en résidences (cités U, résidences médicales, EHPADs…), pour les années 2009-2010 et 2017-2018. Comme les mises en construction de logement peuvent varier fortement d'une année à l'autre en fonction des chantiers, on a ici regroupé les valeurs de deux années par des moyennes mobiles.
http://www.geotests.net/cours/apter/l3/donnees/
sitadel_31coms_commences_2009-2018.xlsx

Le surfaces sont en mètres carrés, vous pourrez ainsi calculer la surface moyenne des logements commencés par commune et par année. Les communes sont celles de 2010, certaines ont depuis été fusionnées (comme Pradère-les-Bourguets et Lasserre, par exemple).

→ Dans une nouvelle feuille du classeur, réalisez un tableau croisé dynamique pour isoler les valeurs de Tournefeuille et des communes comparables de l'agglomération (ou des communes proches comme Cugnaux), qui permette d'analyser l'évolution du nombre et des surfaces des logements individuels purs par rapport aux logements collectifs.

3.1.1 Indicateurs de centralité

Pour une lecture plus intéressante du tableau, nous allons trier les lignes en fonction des valeurs décroissantes du nombre total de logements commencés (menu "Données", puis "Trier").

Il existe deux indicateurs de centralité : la *moyenne* et la *médiane*. La moyenne est la valeur moyenne de l'échantillon de données, la médiane la valeur qui se situe à la moitié du nombre total de valeurs lorsqu'on la trie dans l'ordre croissant. Ainsi, la médiane « coupe » la variable en deux moitiés de nombre égal de valeurs (en cas de nombre pair de valeurs, on prend la moyenne des deux valeurs qui encadrent la médiane).

Dans notre exemple, pour calculer la moyenne d'un groupe de cellules Excel, il suffit d'entrer la formule dans une cellule vierge, sous la colonne par exemple, pour obtenir la moyenne d'une année pour le département. En bout de ligne, on obtient la moyenne pour une commune sur toute la période.

Pour la première colonne, correspondant aux logements individuels purs de l'année 2009-2010 (colonne « B » de la feuille), saisir dans les cellules B595 et B596 :

=MOYENNE(B5:B594)
=MEDIANE(B5:B594)

(On commence à la ligne 5 pour éviter de prendre en charge les valeurs de l'Occitanie et de l'ancienne région Midi-Pyrénées. Une ou deux décimales suffisent pour la lecture simple des résultats).

Pour la première ligne de données du département, correspondant à la commune d'Agassac (ligne 5 de la feuille), pour calculer la moyenne sur les logements individuels sur la période, saisir dans la cellules V5 :

=MOYENNE(B5 ;L5)

Pour les petites communes, avec des années sans constructions, ces formules renvoient des erreurs comme des divisions par zéro. Il est possible de les éviter en utilisant la fonction « SI() » pour ne calculer la moyenne ou la médiane que si les cases ne sont pas vides.

La différence entre la médiane et la moyenne (notamment en colonne, entre les différentes communes de Haute-Garonne) indique que cette dernière est « tirée » par des valeurs élevées isolées (notamment celle de Toulouse). Comme les logements commencés peuvent fluctuer d'une année sur l'autre, le calcul de moyenne peut être intéressant, éventuellement sur deux années en début et en fin de période.

Pour caractériser ces écarts entre valeurs, avoir une vision synthétique de l'hétérogénéité de la variable, il faut utiliser les indicateurs de *dispersion*.

3.1.2. Indicateurs de dispersion : variance et écart-type

La calcul de la variance entre cantons du département permet de confirmer la grande dispersion des valeurs.

L'écart-type (la racine carrée de la variance) est un indicateur plus souvent utilisé pour estimer la dispersion d'une variable. Il correspond à l'écart moyen à la moyenne. Il présente l'avantage d'être exprimé dans la même unité que les valeurs (ici des logements), contrairement à la variance (dont la valeur est abstraite).

Sous Excel, la variance se calcul en utilisant la fonction VARPA(), car la fonctionVAR() tout court estime la variance d'une population dont on n'aurait les valeurs que d'un échantillon. De même, l'écart-type se calcule en utilisant la fonction ECARTYPEP() ou ECARTYPE.PEARSON(), car la fonction ECARTYPE() tout court ou ECARTYPE.STANDARD() est réservée à l'estimation d'un échantillon.

Le calcul de l'écart-type permet de repérer les communes dont le nombre de logements construits est très différent de la moyenne départementale (en plus ou en moins).

Dans la cellule B597, on calcule la variance des communes pour les logements individuels purs en 2009-2010 :
=VAR.P(B5:B594)
Puis l'écart-type en cellule B598 :
=ECARTYPEP(B5:B594)

Une fois les variables mieux connues par des indicateurs de synthèse, on peut tenter de les croiser avec d'autres pour produire des indicateurs analytiques qui les relativisent et permettent de les comparer plus directement.

3.2 Indicateurs analytiques

Pour étudier plus dynamiquement les valeurs du nombre de logements commencés, nous allons les observer dans leur évolution sur la période. Pour éviter de calculer des indicateurs trop sensibles à des variations inter-annuelles, nous allons "lisser" l'analyse en tr vaillant sur des moyennes mobiles sur deux ans.

3.2.1. Taux d'évolution

Dans notre fichier, nous pouvons calculer, pour la première commune :

- l'évolution brute entre 2009-2010 et 2017-2018, dans la case W2 :
  =L2-B2
- l'évolution relative (%) entre 2009-2010 et 2017-2018, dans la case X2 :
  `=V2*100/B2` (attention aux valeurs zéro).
- l'évolution annuelle moyenne (EAM) en % entre 2010 et 2018, dans la case Y2 :
  `=(((L2/B2)^(1/8))-1)*100`

$$\text{EAM} = \left(\sqrt[n]{\frac{\text{valeur finale}}{\text{valeur initiale}}} - 1 \right) \times 100$$

Avec *n* le nombre d'années de la période, ici 8 ans.

L'évolution annuelle moyenne permet de comparer des évolutions ne portant pas sur les mêmes périodes.

3.2.2. Projections

Le taux d'évolution annuelle moyen (EAM) permet de projeter des valeurs dans le futur.

Projection = Valeur_connue x $(1+EAM/100)^n$
(avec n = nombre d'années dans le futur)

→ Projetez l'évolution du nombre de logements collectifs en 2050.

3.3 Croisements et comparaisons

Pour comparer la construction de logements et l'évolution de la population, nous avons rassemblé les données dans le même fichier, par onglets.

Attention, comme les communes évoluent d'année en année, les données peuvent ne plus être compatibles les unes avec les autres ! Vérifiez bien, en triant selon l'ordre alphabétique des codes INSEE, que les communes d'un onglet correspondent bien à celles du second onglet (on peut utiliser la fonction Excel « SI() » pour vérifier, ou recherchev() pour retrouver les bonnes correspondances).

Comment comparer l'évolution de la population et celle des logements mis en construction ?

- Par le tri des données et la comparaison visuelle.
- Par le calcul des rangs (classements) ou d'un indicateur spécifique.
- Par le regroupement en classes de variation aux évolutions (forte / faible / stable).
- Par le calcul d'un indicateur de comparaison directe (nb. d'habitants par logement construit sur la période 2009-2018). *Attention, vérifiez que les communes sont les mêmes dans les deux tableaux de données ! (ici, oui).*

Ainsi, en comparant le nombre de nouveaux logements et celui de nouveaux habitants, on peut se faire une idée du type de croissance urbaine dans les communes (notamment de sa densité, donc se sa forme concrète : logement individuel ou collectif, taille des logements).

La comparaison visuelle des valeurs peut être grandement améliorée en utilisant une coloration graduée dans les cellules, avec la mise en forme conditionnelle.

- Sélectionner la colonne à colorer, par exemple celle de l'évolution relative du nombre de logements commencés.
- Aller dans le menu Format, choisir Mise en forme conditionnelle.
- Cliquer sur le bouton + en bas à gauche de la fenêtre des règles de coloration.
- Choisir « échelle à deux couleurs » dans la liste des styles.
- Pour le type de minimum, choisir valeur puis 0, puis une couleur jaune très clair.
- Pour le type de maximum, choisir valeur puis 100, puis un rouge sombre.

Enfin, il existe des outils statistiques pour mesurer quantitativement la corrélation entre deux variables, le plus simple d'entre eux étant le coefficient de corrélation de Bravais-Pearson (régression selon un modèle linéaire).

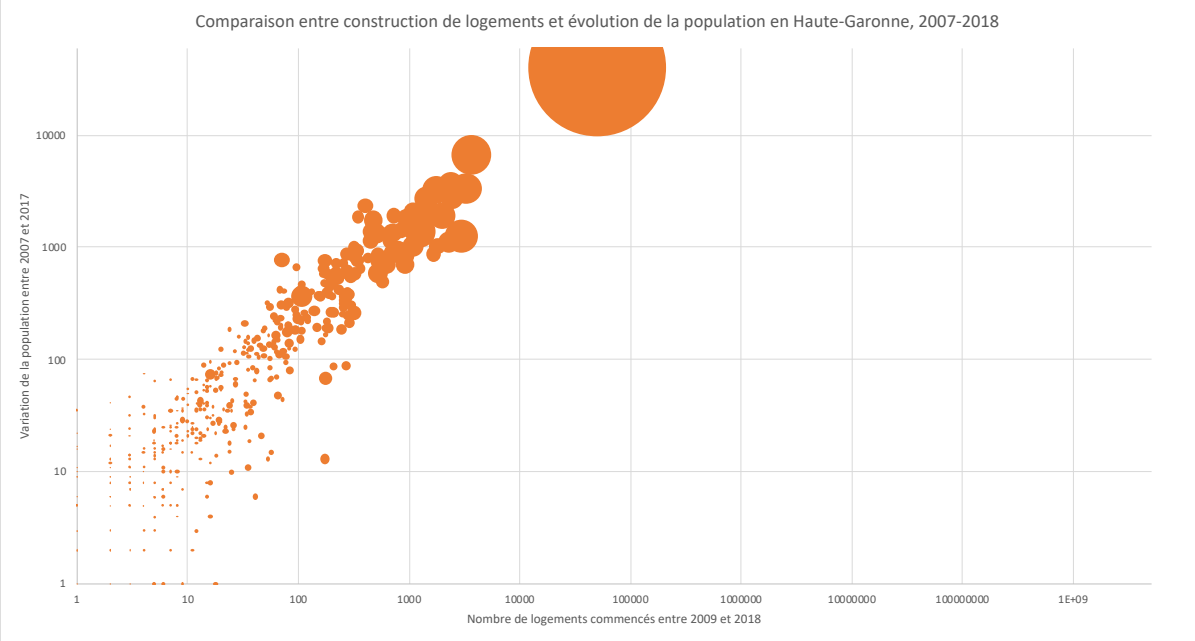
$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

La formule ci-dessus se lit comme la somme des produits des écartes à la moyenne ($\overline{x}$ et $\overline{y}$) divisée par la racine carrée du produit des sommes des écarts à la moyenne au carré.

Excel propose une fonction dédiée : COEFFICIENT.CORRELATION

Ici la corrélation est très élevée (0,995), ce qui est normal, plus intéressant serait d'observer les écarts à ce modèle (ie. les communes dont les gains en habitants sont assez différents de ceux en logements, dans les deux sens), ce qu'on appelle les "résidus" en statistiques.

Pour aller plus loin, on peut chercher à modéliser la relation entre nos deux valeurs, comme la corrélation est très forte on peut rester au modèle le plus simple, la régression linéaire : une droite représente le lien entre les valeurs de population et de logements.



Représentation graphique de la relation entre les évolutions de la population et la construction de logements entre 2007 et 2018, en échelle logarithmique

Excel permet le calcul automatique de la droite de régression linéaire, en utilisant la fonction DROITEREG, « enrobée » dans la fonction INDEX pour récupérer les deux paramètres de pente (paramètre a, d'index 1) et de constante de décalage en Y (paramètre b, d'index 2), pour obtenir l'équation de la droite Y = ax+b.

Avec ces paramètres, on peut calculer la valeur modélisée des logements par rapport à la population, puis l'écart à cette valeur, le résidu. L'analyse des résidus, des écarts à ce qui est attendu en général dans le département, peut aider à repérer les communes qui sont loin de la situation moyenne, dans les deux sens (manque ou surplus de logements).

4. Représentation graphique des données

Les graphiques statistiques peuvent servir à :

- Analyser visuellement l'information, pour se rendre compte plus rapidement de son contenu ;
- Communiquer les résultats de cette analyse de manière rapidement lisible.

Comme les logiciels permettent de dessiner très rapidement des graphiques variés, on peut d'abord les utiliser comme outils d'exploration de l'information, complémentaires au calcul d'indicateurs, pour une meilleure lecture des répartitions des valeurs.

Cependant, il existe des règles à respecter pour choisir le bon type de graphique et le paramétrer pour obtenir une représentation objective et lisible.

Par exemple, représentons graphiquement :

- Le nombre de logements commencés en 2009-2010 et en 2017-2018.
- Le ratio du nombre d'habitants nouveaux par logement commencés entre 2007 et 2018.

Quel type de représentation effectuer, parmi tous les types de graphiques disponibles dans Excel ? Il faut se demander ce qu'il est intéressant de montrer avec ces données.

- L'évolution dans le temps ? (Courbes)
- La répartition très inégale du ratio hbts/logt (bâtons / histogrammes)

Les graphiques en courbes ne se justifient que si les valeurs entre les points de la courbe ont un sens (le plus souvent : une évolution dans le temps).

L'intérêt d'une carte ne se justifie que si la lecture géographique de l'information est utile, c'est-à-dire si la position des valeurs dans l'espace est une variable explicative à part entière : distance au centre de l'agglomération, proximité d'infrastructures, lien avec le contexte géopolitique ou d'autres facteurs spatialisés.

Le tri d'un histogramme permet de classer l'information et de la rendre plus lisible. Cela évite d'avoir à faire de nombreux allers-retours visuels entre les catégories.

Un diagramme en secteurs serait-il utile ici ? Pourquoi ?

Exercices :

- Calculer la **part** de logements individuels dans les logements commencés.
- Calculer la **surface moyenne** des logements totaux et son évolution.
- Calculer la **surface moyenne** des logements individuels et son évolution.
- Comparer les **surfaces** construites entre 2009-2010 et 2017-2018 avec les variations de population sur cette période (ie. les m² par habitant nouveau), pour les logements totaux et les logements individuels purs.