Analyser et traiter les données d'un fichier CSV

Le protocole suivant fonctionne avec la dernière version d'Excel (Office 365).

Quelques notions de cours pour débuter.

Des données peuvent être **structurées** sous forme de tableaux (ou de listes) qui sont en fait des **tables de données**, ce qui permet de les **manipuler**. Ensuite, pour faciliter le traitement et l'analyse, elles peuvent être présentées sous différentes formes : courbes, graphiques, diagrammes circulaires… Les logiciels appelés **tableurs** (exemple : Excel…) permettant à la fois de **structurer** des données dans des tableaux, de les **traiter** et de les **analyser**, en utilisant leurs représentations graphiques.

Une **base de données** est un moyen de recueillir et présenter des **informations de façon structurée**, généralement sous forme d'un tableau à plusieurs colonnes (voir **tables de données**). Le nom des colonnes correspond aux **descripteurs** et les **valeurs** se trouvent dans le tableau. Un **objet** est un élément d'une table de données. Une **collection** est un regroupement d'objets partageant les mêmes descripteurs.

On peut y **effectuer des recherches** en sélectionnant un ou plusieurs descripteurs, on peut aussi faire des **tris**, des **filtres**, des **calculs** etc.

Une table de données.

Collection

Descripteurs →	Prénom	Nom	Âge
Une valeur du descripteur « Prénom » →	Antoine	Ledoux	15
	Pauline	Darcis	16
	Karim	Chanhoun	15
Un objet →	Sarah	Goldberg	14

D'après SNT 2nde Delagrave 2019

Passons maintenant à la partie pratique.

Les questions nécessitant une réponse écrite ou une copie d'écran sont précédées d'une →

Etape 1. Charger un fichier CSV.

1. **Chercher** le fichier « fr-en-indicateurs-de-resultat-des-lycees-denseignement-general-et-technologique » dans votre casier numérique sur l'ENT. Ce fichier provient du site https://data.education.gouv.fr utilisé lors de la séance précédente. **Télécharger** le fichier.

→ 2. **Faire** un clic droit sur le fichier afin d'en **déterminer** les propriétés : en **donner** les principales métadonnées (= une donnée servant à définir ou décrire une autre donnée).

→ 3. **Ouvrir** le fichier avec le logiciel Bloc-Notes. En **réaliser** une copie d'écran justificative. **Préciser** comment sont représentées les données.

→ 4. Il existe des séparateurs (= signes de ponctuation) entre les différents champs. **Indiquer** quelle est la nature de ces séparateurs.

5. **Fermer** le fichier sans le modifier.

Etape 2. Ouvrir un fichier CSV dans un tableur.

6. Pour **ouvrir** proprement le format csv dans le tableur, **suivre** la procédure suivante :

6a. **Ouvrir** Excel 365 (= ouvrir un nouveau classeur).

6b. Dans le menu « Données », **cliquer** sur « Obtenir des données » puis « A partir d'un fichier » puis « à partir d'un fichier texte/csv ».

Une fenêtre d'exploration s'ouvre pour permettre de sélectionner le fichier « fr-en-indicateurs-de-resultat-des-lycees-denseignement-general-et-technologique » dans Téléchargements. L'**importer**.

6c. Dans la nouvelle fenêtre **aller dans** « origine du fichier », puis **sélectionner** « unicodeUTF_8 » si ce n'est pas le cas.

6d. **Laisser** « Point-virgule » dans le champ « Délimiteur ».

6e. **Cliquer** sur « Charger » pour **valider**. Le tableau est alors mis en forme automatiquement et les outils de filtre et de tri sont créés (ce qui n'était pas le cas dans les précédentes versions d'Excel).

→ 7. **Compter** (et **indiquer**) le nombre de descripteurs du tableau. **Compter** (et **indiquer**) ensuite le nombre de lignes. *Pour gagner du temps, la commande « Ctrl + Maj + Fin » vous permet d'aller directement à la dernière cellule utilisée dans la feuille de calcul (coin inférieur droit). D'une manière générale, la commande « Ctrl + flèche » vous permet d'aller dans chaque coin de la table de données.* **Calculer** alors le nombre de valeurs de la table (attention à ne pas compter les descripteurs). Le calcul doit être détaillé sur la feuille réponse.

Etape 3. Filtrer, trier et masquer les données. Réaliser des calculs.

Vu la taille du tableau, il faut nécessairement filtrer les données pour s'y retrouver. Normalement les filtres se font automatiquement dans la dernière version d'Excel, à condition d'avoir correctement importé les données CSV.

- Sinon, l'option « filtrer » (et l'option « trier ») se trouve dans l'onglet « données ».

8. Pour s'y retrouver par la suite, et garder la ligne de descripteurs toujours présente, je vous conseille de « figer les volets ». Pour cela, aller dans l'onglet « Affichage », puis « Figer les volets » et « Figer la ligne supérieure » qui restera alors toujours à l'écran quel que soit l'endroit où vous vous trouvez dans le tableau.

→ 9. En dessous du tableau, on cherche à connaître le nombre d'objets « HAUTS-DE-SEINE » (c'est-à-dire le nombre de HAUTS-DE-SEINE du tableau). Pour cela, il faut taper la fonction conditionnelle NB.SI sous le tableau. Cette fonction compte le nombre de cellules qui répondent à un critère.

Taper en cellule G27811 : =NB.SI(sélectionner la plage de départements en entier soit par un glisser-déposer, soit avec la cellule de début et la cellule de fin séparées par « : » ; "HAUTS-DE-SEINE"). Les guillemets sont indispensables pour qu'Excel comprenne qu'il s'agit d'une chaîne de caractères. Par ailleurs, il ne faut oublier aucun caractère, sinon le tableur ne comprend pas la sélection (par contre, pour le logiciel, que ce soit écrit en majuscule ou minuscule n'a aucune importance).
Indiquer le nombre d'objets « HAUTS DE SEINE ». **Accompagner** la réponse par la copie d'écran justificative (avec obligatoirement la formule indiquée dans la barre de formule : pour cela, il faut reste sur la cellule G27811 sélectionnée).

10. **Filtrer** par département, et **choisir** « HAUTS DE SEINE ». Seuls les objets correspondant aux Hauts de Seine s'affichent alors.

→ 11. **Ajouter** un deuxième filtre « ANTONY » pour le descripteur « Ville ». **Indiquer** combien de lycées différents apparaissent alors et quels sont-ils.

12. **Filtrer** ensuite par établissement pour ne garder que le lycée Descartes (vous avez donc effectué un triple filtre). Les données apparaissent alors chronologiquement dans le désordre : il faut donc les trier. **Trier** par année croissante (= par ordre croissant de dates).

→ 13. On veut maintenant **masquer** tous les descripteurs inutiles pour ne garder que les descripteurs suivants : établissement, année, ville, ; taux de réussite (série L, série ES, série S, série STMG). Pour **masquer** les descripteurs inutiles, **sélectionner** vos colonnes (vous pouvez faire une sélection groupée) puis **effectuer** un clic droit sur le(s) numéro(s) de colonne puis **masquer** la/ les colonnes inutiles : les colonnes masquées existent toujours, mais ne sont plus visibles (sauf par un double trait). Il est d'ailleurs possible de les démasquer. **Accompagner** la réponse par la copie d'écran justificative des colonnes restantes (normalement : A B C V W X AB).

→ 14. On vous propose maintenant de faire les moyennes des taux de réussite par série. Pour cela, il faut taper en bas de la colonne « taux de réussite au bac L » les trois fonctions suivantes (soit respectivement en V27811, V27812 et V27813) :

- Méthode 1 : = MOYENNE(sélectionner la plage de valeurs V en entier, soit de V2 à V27809)

- Méthode 2 : = MOYENNE(sélectionner les valeurs de V les unes après les autres séparées par des ;)
- Méthode 3 : =SOUS.TOTAL(sélectionner le n° de fonction approprié ; sélectionner la colonne de valeurs V en entier comme pour la première fonction).

Pour chaque méthode, ne **taper** la formule qu'une fois puis tirer le cadre par le petit carré vers la droite pour que les calculs soient faits automatiquement pour les autres descripteurs (c'est ce qui s'appelle la « poignée de recopie » : une grosse croix noire apparaît alors).
Pour réduire le nombre de décimales à 2, aller dans l'onglet « Accueil », puis dans le menu « nombre », réduire le nombre de décimales à deux (vous pouvez faire une sélection groupée).
Les trois méthodes donnent des résultats différents, mais seules deux sont justes. **Trouver** lesquelles et **expliquer** pourquoi l'une est inexacte dans ce contexte.
Accompagner la réponse par la copie d'écran justificative.

Etape 4. Représenter graphiquement et modifier le graphique.

→ 15. On veut représenter graphiquement les taux bruts des quatre séries de baccalauréat concernées en fonction du temps. Pour que le graphique apparaisse correctement, il faut :

- **Sélectionner** l'intégralité du tableau puis le coller dans une nouvelle feuille de calcul (pour se débarrasser des valeurs filtrées) ;
- **Sélectionner** uniquement les colonnes utiles (année et résultats bruts pour les quatre séries). Pour cela **sélectionner** la première colonne puis **ajouter** les autres avec la touche Ctrl maintenue.
- Dans l'onglet insertion **choisir** « nuages de points avec lignes droites et marqueurs ». Cela permet d'avoir les années en abscisse.

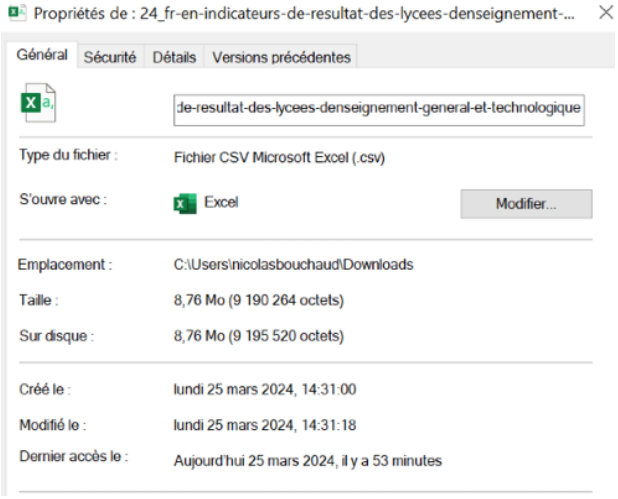
Une fois le graphique effectué :

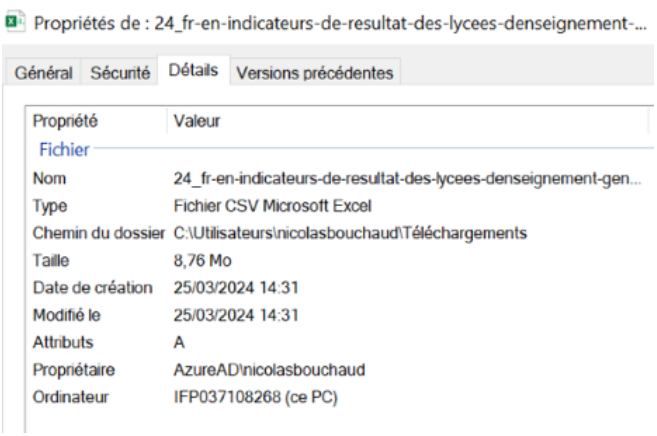
- vous pouvez **adapter** les échelles de l'axe par un clic droit sur les axes (le mieux est de représenter les valeurs en 50 et 100 %) ;
- **légender** les axes (onglet « disposition » accessible lorsque le graphique est sélectionné).
- le **titrer.**

Justifier le travail fait par une copie d'écran.

Correction.

→ 2. **Faire** un clic droit sur le fichier afin d'en déterminer les propriétés : en **déterminer** les métadonnées (= une donnée servant à définir ou décrire une autre donnée) que l'on peut obtenir.





On peut naviguer dans les différents onglets (général, détails) pour trouver différentes métadonnées (= des données en rapport avec le fichier, mais qui n'apparaissent pas dans son contenu) : l'emplacement du fichier, sa taille (y compris sur le disque), la date de création, de modification, le dernier accès, l'auteur…

→ 3. **Ouvrir** le fichier avec le logiciel Bloc-Notes. En **réaliser** une copie d'écran justificative. **Préciser** comment sont représentées les données.

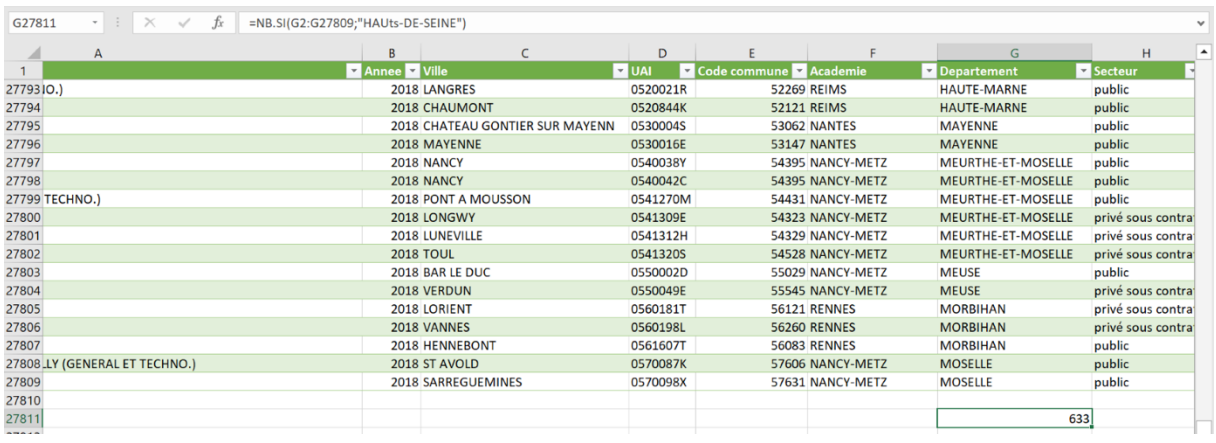
```
eussite attendu acad - Toutes series;Taux de reussite attendu france - L;Taux de reussit
;Taux de mentions attendu - ST2S;Taux de mentions attendu - TMD;Taux de mentions attendu
entions - Toutes series;Valeur ajoutee du taux de reussite - L;Valeur ajoutee du taux de
- Gnle;Taux de reussite - Gnle;Valeur ajoutee du taux de réussite - Gnle;Taux de mentior
ERT DOISNEAU;2013;VAULX EN VELIN;0693619J;69256;LYON;RHONE;PU;22;29;47;;21;;;;;;;;119;82
E DESCARTES;2013;ST GENIS LAVAL;0693654X;69204;LYON;RHONE;PU;31;98;164;88;;;;;59;;;;440;
MAINE TILLION;2013;SAIN BEL;0694069Y;69171;LYON;RHONE;PU;25;66;85;;;;;;;;;;176;88;94;99;
```

Les données sont représentées en ligne, à la suite, sans structuration (fichier texte).

→ 4. Il existe des séparateurs entre les différents champs. **Indiquer** quels sont ces séparateurs.
Ce sont des ; dans le cas présent.

→ 13. **Compter** le nombre de descripteurs du tableau. **Compter** ensuite le nombre d'objets. **Calculer** alors le nombre de valeurs de la table (attention à ne pas compter les descripteurs). Le calcul doit être détaillé sur la feuille réponse.
Descripteurs : jusqu'à EO soit 145 descripteurs et 25 808 lignes si l'on ne compte pas les descripteurs.
Cela représente 145 x 25 808 = 3 742 160 valeurs potentielles.

→ 14. **Indiquer** le nombre d'objets « HAUTS DE SEINE ». **Accompagner** la réponse par la copie d'écran justificative.



Formule indiquée dans la barre de formule. 633 valeurs « HAUTS DE SEINE » apparaissent.

→ 16. **Ajouter** un deuxième filtre « ANTONY ». Indiquer combien de lycées apparaissent alors et quels sont-ils.
Deux lycées différents apparaissent : Sainte Marie Lacroix et Descartes.

→ 18. On veut maintenant **masquer** tous les descripteurs inutiles pour ne garder que les descripteurs suivants : établissement, année, ville, ; taux brut de réussite (série L, série ES, série S, série STMG). **Accompagner** la réponse par la copie d'écran justificative.

	A	B	C	V	W	X	AB
1	Etablissement	Annee	Ville	Taux de reussite - L	Taux de reussite - ES	Taux de reussite - S	Taux de reussite - STMG
2056	LYCEE DESCARTES	2012	ANTONY	95	89	95	
2490	LYCEE DESCARTES	2013	ANTONY	96	92	95	
2976	LYCEE DESCARTES	2014	ANTONY	86	96	96	90
3248	LYCEE DESCARTES	2015	ANTONY	100	95	91	92
3707	LYCEE DESCARTES	2016	ANTONY	95	91	96	82
7038	LYCEE DESCARTES	2017	ANTONY	91	89	94	86
7472	LYCEE DESCARTES	2018	ANTONY	88	93	93	70
8579	LYCEE DESCARTES	2019	ANTONY	96	91	95	71
8993	LYCEE DESCARTES	2020	ANTONY	100	98	99	79
13033	LYCEE DESCARTES	2021	ANTONY				87
16324	LYCEE DESCARTES	2022	ANTONY				85
22200	LYCEE DESCARTES	2023	ANTONY				89

→ 19. On vous propose maintenant de faire les moyennes des taux de réussite par série. Pour cela , il faut taper en bas de la colonne de chaque descripteur la formule moyenne=(sélectionner les valeurs). On vous propose deux méthodes (donc prévoir deux cellules sous chaque colonne) :

- Méthode 1 : moyenne=(sélectionner la colonne de valeurs en entier)
- Méthode 2 : moyenne=(sélectionner les valeurs les unes après les autres séparées par des ;)

Pour chaque méthode, ne taper la formule qu'une fois puis tirer le cadre vers la droite pour que les calculs soient faits automatiquement.
Les deux méthodes donnent des résultats différents, mais une seule est juste. Trouvez laquelle et expliquez pourquoi l'autre est inexacte dans ce contexte.
Accompagner la réponse par la copie d'écran justificative.

	A	B	C	V	W	X	AB
1	ment	Annee	Ville	Taux de reussite - L	Taux de reussite - ES	Taux de reussite - S	Taux de reussite - STMG
2056	SCARTES	2012	ANTONY	95	89	95	
2490	SCARTES	2013	ANTONY	96	92	95	
2976	SCARTES	2014	ANTONY	86	96	96	90
3248	SCARTES	2015	ANTONY	100	95	91	92
3707	SCARTES	2016	ANTONY	95	91	96	82
7038	SCARTES	2017	ANTONY	91	89	94	86
7472	SCARTES	2018	ANTONY	88	93	93	70
8579	SCARTES	2019	ANTONY	96	91	95	71
8993	SCARTES	2020	ANTONY	100	98	99	79
13033	SCARTES	2021	ANTONY				87
16324	SCARTES	2022	ANTONY				85
22200	SCARTES	2023	ANTONY				89
27810							
27811							
27812				92,90	92,02	92,57	91,38
27813				94,11	92,67	94,89	81,43
27814				94,11	92,67	94,89	83,10

La première formule tient compte des valeurs masquées, ce qui n'est pas le cas des deux autres. La meilleure est la dernière (plus rapide et moins de risque d'erreur).

→ 20. On veut représenter graphiquement les taux bruts des quatre séries de baccalauréat concernées en fonction du temps. **Justifier** le travail fait par une copie d'écran.