SAÉ 1.01 - Reporting à partir de données stockées dans un SGBD relationnel

Fascicule de travail : naissance de reporting

Compétence ciblée	Traiter les données à des fins décisionnelles Niveau 1 : Traiter des données structurées
AC couverts	Correctement interpréter et prendre en compte le besoin du commanditaire ou du client Respecter les formalismes de notation Connaître la syntaxe des langages et savoir l'utiliser Mesurer l'importance de maîtriser la structure des données à exploiter
Objectifs de la SAÉ et problématique professionnelle	La mise à jour et la présentation des tableaux de bord sont essentielles au suivi de l'activité d'une entreprise. En tant que chargé d'analyse et de reporting, l'étudiant pourra être amené à produire de tels tableaux de bord en support aux services de pilotage de l'activité. Il devra pour cela assurer la sélection et l'export des données utiles, notamment celles stockées dans des bases de données, les analyser et les restituer avec les outils adaptés. Les objectifs de cette SAÉ sont les suivants : – Amener l'étudiant à construire des indicateurs de performance ainsi que les restituer sous forme de tableau de bord – Identifier les besoins clients et être force de proposition pour s'adapter à ces besoins. – Se confronter à des difficultés dans les bases de données rencontrées
Description	L'étudiant est mis en situation de production de tableaux de bord à partir de données stockées dans un SGBD relationnel, en respectant les termes d'un cahier des charges fourni (spécification, livrables, délai...). La base de données fournie présente un certain nombre de difficultés que l'on peut rencontrer dans une situation professionnelle réelle (BD plus grande, jointures complexes, ...). Le cahier des charges présente le schéma relationnel de la BD à utiliser, les demandes de tableaux de bords et reporting. L'étudiant doit produire l'ensemble des scripts permettant d'extraire les données nécessaires et réaliser les livrables demandés. Il doit en outre documenter le code et le résultat obtenu
Heures formation	5hTD
Heures de projet tutoré	12h projet
Ressources mobilisées	– R1.01 | Tableur et reporting – R1.02 | Bases de données relationnelles 1 – R1.10 | Projet Personnel et Professionnel 1
Type de rendu : Livrable	Rapport de réalisation + codes commentés + reporting produit SGBD, ACCESS
Semestre	Semestre 1

UNIVERSITÉ DE FRANCHE-COMTÉ
Membre de
UBFC UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

1 Cours

Introdution Il semble aisé de collecter des données et de calculer quelques valeurs, comme des moyennes, puis de les interpréter abusivement. Il est en revanche beaucoup plus complexe de le faire dans un cadre scientifique rigoureux.

Tout citoyen, qu'il soit informaticien, biologiste, chimiste, économiste, etc., est confronté à des données et statistiques qu'il doit savoir analyser avec recul. L'analyste des données doit en plus comprendre les mécanismes sous-jacents qui apparaissent lors des rendus visuels.

Les professionnels des données s'intéressent aux trois questions : *comment collecter les données (quelles dimensions) ? comment les analyser (que permettent les données de conclure) ? et comment les présenter ?*

Les probabilités sont un domaine des mathématiques ayant pour objet l'étude de l'incertitude et du hasard. Il ne faut pas confondre probabilités et statistique. Un processus stochastique (probabiliste) peut servir à produire des données que l'on peut analyser de façon statistique. De même, on cherche souvent, à partir d'un jeu de données, à modéliser leur apparition comme si elle venait d'un processus probabiliste, afin d'en déduire des modèles théoriques et prédicatifs. [1]

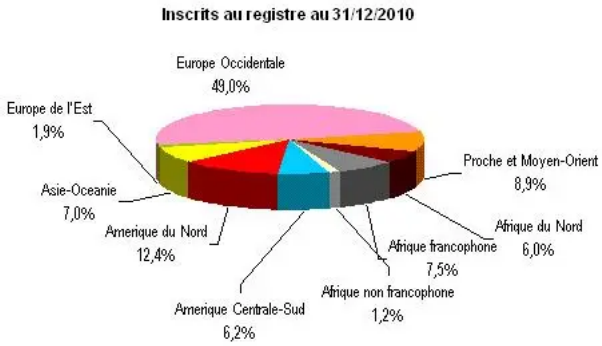
Analyse d'une datavisualisation issue d'un reportage de France3

a) Quelles sont les données ?

b) Quelle visualisation choisie (et pourquoi) ?

c) Quelle(s) erreur(s) ?

d) Conclusion ?



[1] Inspiré du cours de M. Héam

Lorsque l'on présente des résultats statistiques pour comparer des proportions, il est fréquent d'utiliser des camemberts ou des histogrammes. Un exemple est donné dans la Figure 1 [2]. Cette présentation, en relief, est fréquemment utilisée car plus jolie. Cependant, elle est tendancieuse, car ce que l'oeil compare les aires des différents secteurs, et l'utilisation de tranches (pour le relief) donne visuellement un poids plus fort aux données placées devant. Par exemple, sur la Figure 1, l'Europe Occidentale occupe 50% de la surface de l'ellipse marquant le camenbert, mais bien moins de 50% de la surface totale du dessin, alors que cela devrait être le cas. De même, on peut remarque que l'Europe de l'Est paraît occuper une part moins importante que l'Afrique non francophone, ce qui n'est pas le cas. De même l'Amérique Centrale-Sud occupe plus place que le Proche et Moyen-Orient sur le dessin, alors que cela ne devrait pas être le cas. Dans un camembert en relief, les données placées sur l'avant ont tendance à être surestimées.

En pratique il convient donc de ne pas utiliser les camembert en relief, ou, lorsque cela est fait, il faut choisir une très faible épaisseur relativement au rayon du disque et une inclinaison par trop importante, afin de minimiser le biais. De même tout découpage en tranche (en faisant ressortir une tranche) ajoute de l'épaisseur visuelle et accentue l'effet visuel. Il est aussi important de faire attention aux couleurs : des couleurs vives, comme le rouges, attirent plus l'oeil.

[2] http://france3-regions.blog.francetvinfo.fr/ftv-expats/2011/12/04/qui-sont-les-nouveaux-expatries-francais.html

EXERCICE 1 : Analyse de trois datavisualisations de données.

Pour chacune de ces datavisualisation, donnez

a) les éléments mis en avant, à travers les biais

b) ce que la représentation veut montrer

c) un titre

Professions intermédiaires	21,8390804597701
Agriculteurs exploitants	8,76436781609196
Cadres et professions intellectuelles supérieures	36,2068965517241
Artisans, commerçants et chefs d'entreprise	16,5948275862069
Employés et ouvriers qualifiés	9,98563218390805
Employés et ouvriers non qualifiés	6,60919540229885

Figure 1:

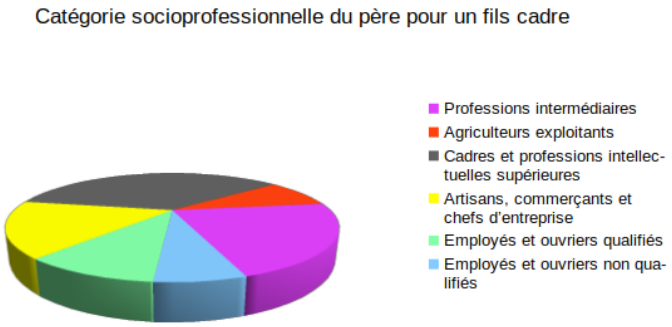


Figure 2:

Figure 3:

EXERCICE 2 : Analyse de trois datavisualisations de données.

Le problème de lien entre une valeur et l'aire représentée est aussi délicat à gérer sur les cartes. On considère par exemple trois communes disposées comme sur la carte ci-dessous et gérée par le même commissariat. Plus la couleur est foncée, plus le nombre de cambriolages est important. Le dessin de gauche représente la situation en 2000 et celle de droite en 2010.

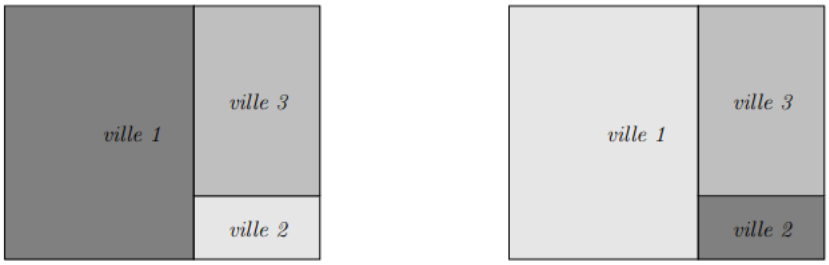


Figure 4: Criminalité

EXERCICE 2.1 : Analyse de la Figure 4

a) Peut-on dire que la situation s'est améliorée ?

b) Comment faudrait il griser la carte pour que cela soit visuellement pertinent ?

EXERCICE 2.2 : Nouvelle analyse

On considère maintenant la figure ci-dessous où plus une ville est foncée, plus il y a de cambriolage par hectare de la ville, encore une fois en 2000 et 2010.

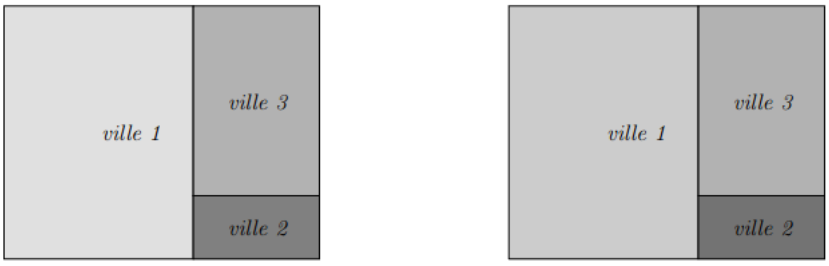


Figure 5: Nouvelle représentation de la criminalité

Cela permet il de savoir où il est préférable d'habiter si l'on craint les cambriolages ?

EXERCICE 3 : Toujours de l'analyse de représentation

Pour cette représentation graphique issue de l'INSEE [a], donnez

a) le(s) biais

b) les éléments mis en avant, à travers le(s) biais

c) ce que la représentation veut montrer

d) un titre

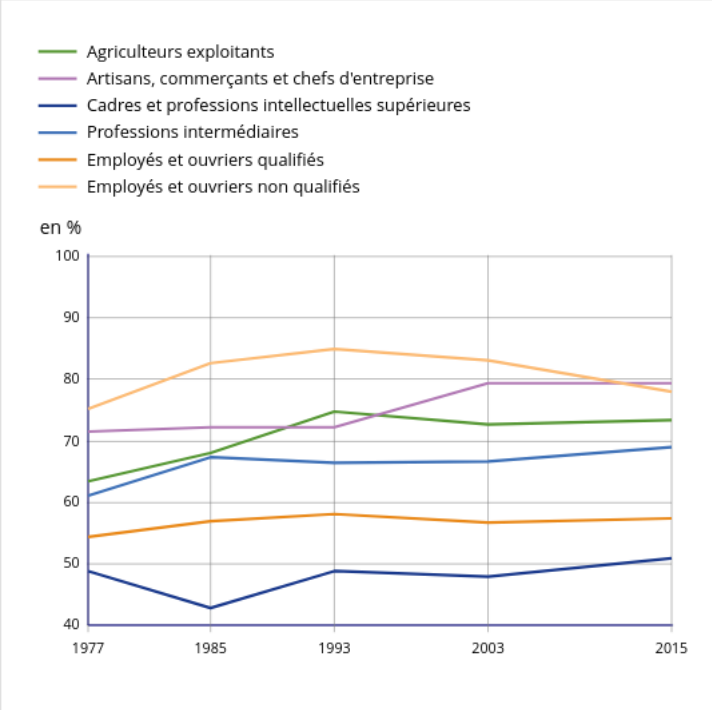


Figure 6:

[a] Source : Insee, enquêtes Formation et qualification professionnelle (FQP) 1977, 1985, 1993, 2003 et 2014-2015.

2 Paradoxe de Simpson

EXERCICE 4 : Paradoxe de Simpson

On considère deux lycées, notés A et B. On observe les taux de réussites suivants dans les deux lycées

Taux de réussite représentation 1

	Lycée A	Lycée B
garçons	70%	71.05%
filles	75%	80%

EXERCICE 4.1 : représentation 1

a) Quel lycée a le meilleur taux de réussite ?

EXERCICE 4.2 : représentation 2

On considère maintenant les effectifs des deux lycées donnés ci-dessous.

Effectif des lycées

	Lycée A	Lycée B
garçons	100	100
filles	190	10

a) Calculer le taux de réussite de chaque lycée.

b) Qu'en déduire ?

Ce qu'il est important de retenir, c'est qu'il est facile (et malheureusement usuel) d'utiliser des représentations graphiques trompeuses. Par ailleurs, il est parfois difficile d'obtenir les valeurs pertinentes permettant d'en tirer des conclusions utiles. Deux exemples :

- il y a en France plus de commotion cérébrale dû à des accidents de la route sur des piétons que sur des vélos. Certaines associations d'utilisateur de vélos l'utilise pour que le port du casque ne soit pas rendu obligatoire. Mais ces chiffres doivent être rapportés aux volume du trafic vélo et du trafic piétons. Se pose alors la question de choisir la mesure du trafic (en temps passé ou en kilomètres parcourus ?) et de savoir le mesurer (actuellement on ne sait le faire qu'avec un facteur 100 près).
- Toujours sur la sécurité routière, il est par exemple aussi difficile de comparer la dangerosité des routes entre deux pays comme la France et le Royaume-Uni. On peut bien entendu compter le nombre d'accidents mortels selon des critères identiques, mais doit on rapporter ce total au nombre de véhicule ? au nombre d'habitants ? au nombre de kilomètres parcourus ? au nombre de kilomètres de route ? Par ailleurs, le trafic n'est pas du tout le même en France (pays de transits entre l'Europe du Nord et du Sud) et le Royaume-Uni qui est une île. De même, on sait que les conditions météorologiques influencent sensiblement le nombre d'accidents. Comment le prendre en compte ?

En conclusion, il faut en statistique avoir les bonnes données (cela demande une expertise métier en général), en nombre suffisant, et savoir restituer ces résultats convenablement.

3 Phénomène de Rogers

Le phénomène de Rogers peut traduire qu'il est possible, avec de même données, de donner des résultats aux conclusion différentes. Cela montre qu'il est possible de truquer des résultat sans pour autant truquer des données, par un moyen purement mathématique.

EXERCICE 5

Imaginons une IUT dans lequel il y a deux groupes de niveaux A et B. Les étudiants du groupe A sont, en général, meilleur que ceux du groupe B. Une année, les groupes A a eu 14.2 de moyenne et le groupe B a eu 8 de moyenne.
La seconde année, les notes du groupe A sont : 16, 15, 15, 13, 11.
Les notes du groupe B sont 13, 8, 6 et 3.

EXERCICE 5.1

- Calculer la moyenne des groupes A et B
- Quelle est l'évolution des moyennes sur les deux années ?
- et quelle est la moyenne des deux moyennes ?

L'enseignant décide alors de faire passer le plus mauvais étudiant du groupe A dans le groupe B.

EXERCICE 5.2

Les notes du groupe A deviennent alors 16, 15, 15 et 13. Les notes du groupe B sont 13, 11, 8, 6 et 3.

- Calculer la moyenne des groupes A et B
- Quelle est l'évolution des moyennes sur les deux années ?
- Et quelle est la moyenne des deux moyennes ?
- Que déduisez-vous ?

Exercice 6 : Les dés non transitifs

On suppose que l'on a trois dés à 6 faces, appelés A, B et C. Lors d'un lancer, pour chaque dé, chaque face sort avec une probabilité 1/6 : les dés ne sont pas pipés.

- le dé A a sur ses faces 2,2,4,4,9,9,
- le dé B a pour sa part 1,1,6,6,8,8,
- le dé C a 3,3,5,5,7,7.

EXERCICE

On s'intéresse à un jeu à deux joueurs où chacun prend un dé, le lance et celui qui a le plus grand résultat gagne.

a) Si j'ai le dé A et mon adversaire le B, quelle est la probabilité que je gagne ?

b) Si j'ai le dé B et mon adversaire le C, quelle est la probabilité que je gagne ?

c) Si j'ai le dé C et mon adversaire le A, quelle est la probabilité que je gagne ?

d) Quel est le dé le plus fort ?

4 Mise en pratique

ENTREE : un SGBD relationnel

Un SGBD relationnel.

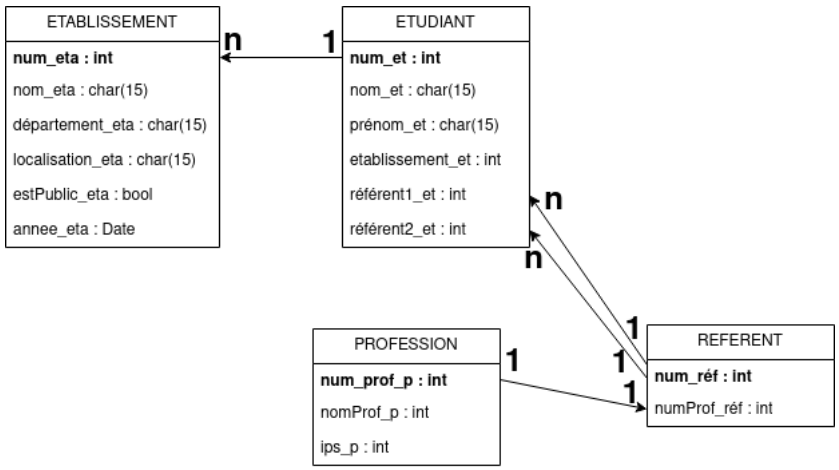


Figure 7: Diagramme UML étudié

Un csv est disponible sous le lien suivant : lien

En sortie : A rendre. Quel modèle de gestion des professeurs pour l'école de demain ?

A rendre : Un classeur Excel qui obtiendra les données nécessaires pour produire un certain nombre de graphiques et tableaux. Vous pouvez d'ors et déjà réfléchir aux données nécessaires et à la façon de concevoir les tableaux et graphiques qui pourraient être demandés.

Traitement des données

Ecrire les requêtes SQL pour :

a) récupérer l'IPS moyen par établissement

b) récupérer l'IPS moyen selon le département de l'établissement

c) récupérer l'IPS moyen selon l'année

d) récupérer l'IPS moyen selon le type d'établissement (public ou privé)

e) récupérer l'IPS moyen selon la localité de l'établissement par année

Réalisation d'un rapport

Réaliser la visualisation des données en respectant 3 cachiers des charges :

a) Un cabinet de conseil peu scrupuleux qui veut une représentation montrant l'égalité des répartitions

b) Un particulier peu scrupuleux qui veut une représentation montrant l'inégalité des répartitions

c) Un comité scientifique qui souhaite une représentation réaliste des répartitions.