PROJET RÉALISÉ PAR

L'ÉQUIPE TDDT DU GROUPE DE TD1

RAPPORT DE GROUPE DES UE BASES DE DONNÉES + SCIENCES DES DONNÉES 2

MOUTCHACHOU Lydia
IBNMTAR Hazem
BERETTI–PRENANT Esteban
VAROL Serdar

Département MIASHS, UFR 6 Informatique, Mathématique et Statistique
Université Paul Valéry, Montpellier 3

Mars 2025

Soumis comme contribution partielle
pour le cours Science des données 2 et Bases de données

Déclaration de non plagiat

À compléter avant la remise du rapport.

Nous déclarons que ce rapport est le fruit de notre seul travail, à part lorsque cela est indiqué explicitement.

Nous acceptons que la personne évaluant ce rapport puisse, pour les besoins de cette évaluation :

- la reproduire et en fournir une copie à un autre membre de l'université ; et/ou,
- en communiquer une copie à un service en ligne de détection de plagiat (qui pourra en retenir une copie pour les besoins d'évaluation future).

Nous certifions que nous avons lu et compris les règles ci-dessus.

En signant cette déclaration, nous acceptons ce qui précède.

Signature : _____ Date : _____

25 mars 2025

Remerciements

À compléter avant la remise du rapport.

Nos plus sincères remerciements vont à notre encadrant pédagogique pour les conseils avisés sur notre travail.

Nous remercions aussi ...

25 mars 2025

Résumé

Notre projet vise à analyser les performances financières des entreprises françaises entre 2018 et 2022 à partir des données du Registre National du Commerce et des Sociétés (RNCS). Nous cherchons à comprendre quels sont les facteurs qui influencent la rentabilité des entreprises et comment ces dernières évoluent en fonction de leur secteur d'activité. Plus précisément, nous allons : - Comparer les performances des entreprises selon leur chiffre d'affaires et leur rentabilité. - Étudier l'impact de la fiscalité sur la profitabilité des entreprises. - Analyser l'évolution des ventes, des stocks et des taxes pour identifier des tendances économiques.

Table des matières

Chapitre 1

Introduction

1.1 Présentation du projet

Les données financières des entreprises jouent un rôle crucial dans la compréhension de leur santé économique. Ce projet se concentre sur l'analyse des performances financières des entreprises françaises entre 2018 et 2022, en utilisant les données fournies par le Registre National du Commerce et des Sociétés (RNCS).

Comparer les performances des entreprises selon leur chiffre d'affaires et leur rentabilité.

Étudier l'impact de la fiscalité sur la profitabilité des entreprises.

Analyser l'évolution des ventes, des stocks et des taxes pour identifier des tendances économiques.

1.2 Responsabilités et composition de l'équipe

MOUTCHACHOU Lydia : Étudiant n°22212656
IBNMTAR Hazem : Étudiant n°22309227
BERETTI–PRENANT Esteban : Étudiant n°XXXX
VAROL Serdar : Étudiant n°22009668

1.3 Objectifs et questions de recherche

Notre projet vise à analyser les performances financières des entreprises françaises entre 2018 et 2022. Pour ce faire, nous allons examiner plusieurs facteurs qui pourraient influencer la rentabilité des entreprises. Les questions spécifiques que nous allons aborder sont les suivantes :

1.3.1 ***Comparaison de la rentabilité par rapport au chiffre d'affaires :***

a. Comment la rentabilité varie-t-elle en fonction de la taille de l'entreprise ?

b. Y a-t-il une différence notable entre les entreprises qui ont recours au refinancement et celles qui n'en ont pas besoin ?

1.3.2 ***Comparaison de la rentabilité par rapport au chiffre***

a. La rentabilité des entreprises diffère-t-elle selon la ville où elles sont implantées ?

b. Les entreprises qui exportent leurs produits ou services sont-elles plus rentables que celles qui opèrent uniquement sur le marché national ?

1.3.3 ***Impact fiscal et sectoriel :***

a. Quel est l'impact des taxes sur la rentabilité des entreprises ?

b. Comment la rentabilité varie-t-elle selon le secteur d'activité des entreprises ?

1.3.4 ***Évolution temporelle :***

a. Comment la rentabilité des entreprises a-t-elle évolué entre 2012 et 2016 ?

b. Peut-on identifier des tendances spécifiques ou des périodes de croissance/déclin dans les performances financières des entreprises ?

En répondant à ces questions, nous espérons identifier les principaux facteurs influençant la rentabilité des entreprises françaises et fournir des insights précieux pour les décideurs économiques et les gestionnaires d'entreprises.

Chapitre 2

Base de données

2.1 Provenance des données

Les données utilisées dans ce projet proviennent du jeu de données Kaggle :

- **Profit and loss - Ontology.csv :** Contient les comptes de résultat de 100 000 entreprises françaises, avec des informations détaillées sur les revenus, les dépenses et les bénéfices.
- **APE_Fusion.csv :** Utilise le code APE pour classer les entreprises selon leur secteur d'activité, permettant des comparaisons sectorielles précises.
- **Data_Kaggle.csv :** Fournit des données globales sur les entreprises, incluant les ventes, les stocks et les taxes, permettant d'analyser l'évolution des performances financières sur plusieurs années.

Lien vers les données : Kaggle Dataset[https://www.kaggle.com/datasets/briaclg/financial-data-of-french-compagnies/data?select=Profit+and+loss+-+Onthology.csv]

2.2 Descriptif des tables

Table 2.1: Profit and loss - Ontology.csv (nombre de lignes × nombre de colonnes)

Nom colonne	Type	Signification	Caractéristique
Columns (FR/EN)	varchar	Colonnes des états\financiers en français et en anglais	

2.3 Modèles MCD et MOD

- Pour le MCD, inclure une image réalisée avec le logiciel Mocodo [https://www.mocodo.net/] telle que celle visible sur la Figure 2.1 ci-dessous :

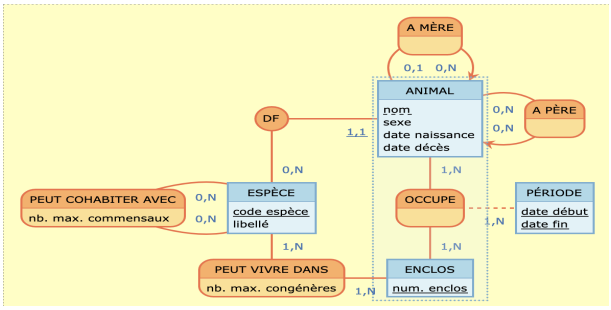

FIGURE 2.1: Relations.

— Pour le MOD, inclure une image réalisée avec le designer de phpmyadmin

Noter en passant qu'il est possible de créer des diagrammes en R Markdown au moyen du package `DiagrammeR` [https://rich-iannone.github.io/DiagrammeR/graphviz_and_mermaid.html] comme on peut le voir ci-dessous.

```
# install.packages("webshot",dependencies = TRUE)
# library(webshot)
# webshot::install_phantomjs()
Sys.setenv(OPENSSL_CONF="/dev/null")
DiagrammeR::grViz("
digraph boxes_and_circles {

  # a 'graph' statement
  graph [overlap = true, fontsize = 10]

  # several 'node' statements
  node [shape = box,
        fontname = Helvetica]
  A; B; C; D; E; F

  node [shape = circle,
        fixedsize = true,
        width = 0.9] // sets as circles
  1; 2; 3; 4; 5; 6; 7; 8

  # several 'edge' statements
  A->1 B->2 B->3 B->4 C->A
  1->D E->A 2->4 1->5 1->F
  E->6 4->6 5->7 6->7 3->8
}
")
```

2.4 Import des données

— Préciser les nettoyages réalisés avant l'import comme l'uniformisation des valeurs des champs (*e.g.*, Mr, M., Monsieur, ...) ou le remplissage des valeurs manquantes par une valeur moyenne ...

— Source de données 1 :
  — Suppression des colonnes XXX, car XXX
  — Suppression des doublons dans les colonnes XXX
  — Filtrage en fonction de la colonne XXx, nous n'avons conservé que....

2.5 Requêtes réalisées

Pour chaque requête, l'exprimer en langage naturel puis en SQL. Puis donner le résultat obtenu (ou un extrait) et expliquer ce résultat.

L'objectif est de varier le type de requêtes et de répondre à votre problématique initiale.

2.6 Quelques détails techniques

On peut interagir avec une base de données directement depuis RMarkdown. Un fichier .Rmd sera fourni pour donner des exemples.

Chapitre 3

Matériel et Méthodes

3.1 Logiciels

Lister tous les logiciels utilisé pour la partie statistique du rapport (et également ceux pour gérer et communiquer entre les membres du projet s'il y en a en particulier)

R (ou Python) est le logiciel à privilégier pour la Science des Données. Pour assurer une reproductibilité maximale, vous devriez utiliser R Markdown (ou un Notebook Jupyter, et éventuellement un outil de gestion des versions tel que `Git`), par exemple via Google Colab ou RStudio dans les nuages. Évitez d'utiliser Word !

Il est de votre responsabilité de donner les versions des logiciels que vous utilisez, ainsi que de donner des informations techniques sur l'ordinateur qui vous a servi pour les analyses (système d'exploitation, vitesse du processeur, etc.). Penser à fournir des citations pour les logiciels utilisés, par exemple [1].

3.2 Modélisation statistique

Quels outils ou méthodes de statistiques allez-vous utiliser ? Donner des équations mathématiques s'il y a lieu et lister les éventuels présupposés («assumptions» en anglais) que vous devez faire sur les données afin d'utiliser ces outils ou méthodes (*e.g.*, normalité, absence de valeurs aberrantes, etc.).

Il est également bon d'indiquer quelles sont les avantages et les limites de ces méthodes.

Vous pourrez consulter avec profit les Chapitre 11–13 du livre sur R utilisé pendant le cours :

http://biostatisticien.eu/springeR/livreR.pdf

1. L'entrée BibTeX ajoutée dans le fichier `references.bib` a été obtenue grâce à la commande `citation(package = "tidyverse")` tapée dans la console de R.

Chapitre 4

Analyse Exploratoire des Données

Toute étude impliquant des données doit **obligatoirement** inclure une analyse exploratoire préalable. Celle-ci permet de mieux comprendre l'information contenue dans les données.

Il faut produire de nombreux résumés graphiques (*e.g.*, histogrammes, nuages de points, boxplots, etc.) et numériques (*e.g.*, médiane, moyenne, variance, etc.). Ainsi, il faut faire une analyse descriptive uni- et bivariée systématique de toutes les variables du jeu de données. Puis, il faut uniquement conserver les plus pertinents (les autres pouvant être gardés en Annexe), c'est-à-dire ceux qui permettront de dégager des éléments de réponse pour la question de recherche envisagée. Chaque figure et tableau doit être commenté. Mais il ne faut pas extrapoler et dire des choses qui ne sont pas visibles dans ces graphiques ou tableaux. Pour chaque analyse, vous pourrez préciser le nombre d'individus/ d'unités statistiques concernés au total.

Vous pourrez consulter avec profit le Chapitre 9 du livre sur R utilisé pendant le cours :

http://biostatisticien.eu/springeR/livreR.pdf

4.1 Utiliser R

Il est facile d'inclure des codes R dans votre rapport, qui seront exécutés à la volée (*i.e.*, lors de la traduction de votre fichier `Rmd` en fichier `PDF` ou `DOC`). Par exemple :

```
boxplot(cars, col = c("#5975a4", "#cc8963"))
```

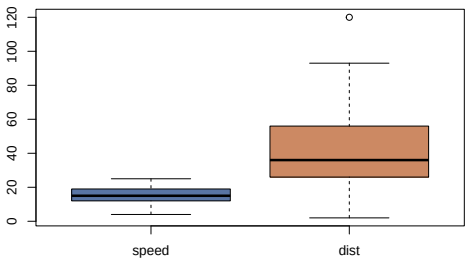

Figure 4.1: Deux boxplots.

```
colMeans(cars)
```

```
## speed  dist
## 15.40 42.98
```

Les lignes de code ne doivent pas dépasser dans la marge de droite. Ainsi on pourrait remplacer le chunk ci-dessous :

```
boxplot(cars, main = "Un titre qui est vraiment beaucoup trop long et qui dépas
```

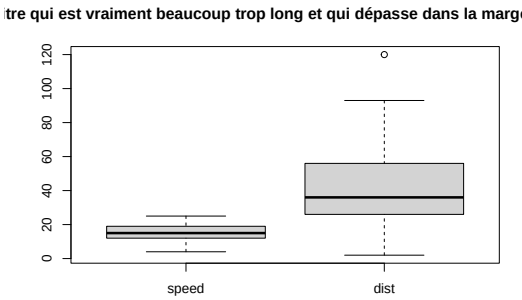

FIGURE 4.2: Pas super.

par celui-ci :

```
boxplot(cars,
        main = "Un titre qui est vraiment beaucoup trop long\n mais qui ne dépasse plus dans la marge de droite")
```

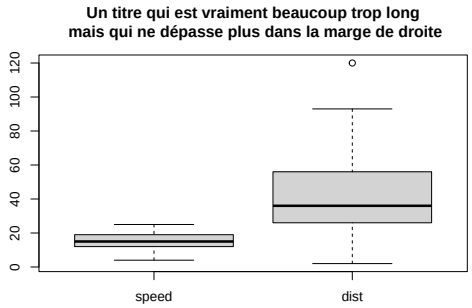

FIGURE 4.3: Déjà mieux.

où l'on a :

— utilisé la commande LaTeX `\tiny` pour changer la taille de la police (suivi de de `\normalsize` pour revenir à la taille normale),
— mis l'instruction `main = ...` sur la deuxième ligne,
— utilisé `\n` pour afficher le titre sur deux lignes.

Chapitre 5

Analyse et Résultats

Dans cette partie, vous pourrez utiliser les outils et méthodes vus au semestre précédent pour analyser les liens entre les variables.

Pour cela, vous pourrez utiliser les tests du χ^2, test du coefficient de corrélation linéaire, test d'Anova, la droite de régression linéaire.

Vous pourrez également proposer des modèles pour faire du clustering (k-means, CAH), de la classification (K plus proches voisins par exemple) comme vu en Science des données 1.

5.1 La droite de régression linéaire : un premier exemple

Si on souhaite expliquer les variations d'une variables réponse Y en fonction d'un certain nombre de prédicteurs $x_1, \ldots, x_p$, on peut utiliser un modèle de régression linéaire simple ($p = 1$) ou multiple ($p > 1$)

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i, \qquad i = 1, \ldots, n.$$

où l'on présuppose que les ϵ_i sont i.i.d. $N(0, 1)$ pour tout $i = 1, \ldots, n$ (n étant la taille de l'échantillon).

Vous pourrez toujours consulter avec profit les Chapitre 11–13 du livre sur R utilisé pendant le cours :

http://biostatisticien.eu/springeR/livreR.pdf

Ces chapitres détaillent l'utilisation de certains tests et modèles sous `R`.

Chapitre 6

Discussion

Placer les résultats que vous avez obtenus dans le chapitre précédent en perspective par rapport au problème étudié.

Chapitre 7

Conclusion et perspectives

Quelles sont les conclusions principales ? Quelles sont vos recommandations pour le commanditaire ? Quelles analyses subséquentes pourraient être faites dans le futur ?

On attend de vous deux types de perspectives : des perspectives à court terme pour améliorer rapidement votre approche et des perspectives à plus long terme qu'elles soient liées à la science des données ou au domaine métier pour lequel vous avez travaillé.

Lister également les difficultés rencontrées dans la partie BD (e.g., taille de la base, manque de données, ...) et dans la partie statistique.

Bibliographie

Annexes

Il faut utiliser les annexes de façon judicieuse. C'est ici que l'on place des résultats trop volumineux pour apparaître dans le corps du rapport. Ou bien des résultats (e.g., graphiques) moins intéressants que les autres. Cela permet de limiter le nombre de pages du coeur du rapport, et d'ajouter des détails dans cette partie pour le lecteur désireux d'en savoir plus.

Codes

Ajouter vos codes informatique ici. Les codes doivent être correctement indentés et commentés.

Tables

Si vous avez des tableaux supplémentaires, vous pouvez les ajouter ici.

Utiliser https://www.tablesgenerator.com/markdown_tables pour créer des tables Markdown simples, ou bien utiliser LaTeX.

TABLE 7.1: une légende au-dessus du tableau.

Les tables	sont	cool
col 1 est	alignée à gauche	$1600
col 2 est	centrée	$12
col 3 est	alignée à droite	$1

Aligner les nombres de la troisième colonne sur la droite permet d'afficher les unités au-dessus des unités, les dizaines au-dessus des dizaines, etc. Il faut toujours privilégier cette présentation.