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

Avril 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 : _____

13 avril 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 ...

13 avril 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°22208752
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(Lydia) :***

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

2.2 Descriptif des tables

2.2.1 Table 1 : APE_Fusion.csv

Nom colonne	Type	Signification	Caractéristique
Unname d : 0	int	Index ou identifiant de ligne (peut être ignoré dans l'analyse)	
ape	object	Code APE complet de l'activité principale de l'entreprise	Clé primaire
ape_name	object	Nom ou description de l'activité correspondant au code APE	
ape_len	int	Longueur du code APE, indiquant le nombre de caractères qu'il contient	
ape_cat0	int	Premier niveau du code APE (division), composé des 2 premiers chiffres	
ape_cat1	float	Deuxième niveau du code APE (groupe), composé des 3 premiers chiffres	
ape_cat2	float	Troisième niveau du code APE (classe), composé des 4 premiers chiffres	
ape_cat3	object	Dernier niveau du code APE (sous-classe)	
Libellé	object	Description du secteur d'activité auquel appartient le code APE	
Code	object	Code alphabétique supplémentaire associé au secteur d'activité	

2.2.2 Table 2 : Profit and loss - Ontology.csv

Nom colonne	Type	Signification
Columns_(FR/EN)	varchar	Colonnes des états financiers en français et en anglais
Description (FR)	varchar	Explication de ce que chaque colonne représente
Liasse (Id)	int	Identifiant unique des colonnes dans la base INPI
Calcul	varchar	Méthode de calcul pour certaines valeurs dans les colonnes

2.2.3 Table 3 : Data_ Kaggle.csv

N°	Variable	Code
1	year	B
2	Autres impôts, taxes et versements assimilés	C
3	Ventes de marchandises	D
4	Production vendue biens	E
5	Production vendue services	F
6	Chiffres dáffaires nets	G
7	Production stockée	H
8	Production immobilisée	I
9	Subventions déxploitation	J
10	Reprises sur amortissements et provisions, transfert de charges	K
11	Autres produits	L
12	Total des produits déxploitation	M
13	Achats de marchandises (y compris droits de douane)	N
14	Variation de stock (marchandises)	O
15	Achats de matières premières et autres approvisionnements	P
16	Variation de stock (matières premières et approvisionnements)	Q
17	Autres achats et charges externes	R
18	Impôts, taxes et versements assimilés	S
19	Salaires et traitements	T
20	Charges sociales	U
21	Autres charges	V
22	Total des charges déxploitation	W
23	Résultat d'exploitation	X
24	Bénéfice attribué ou perte transférée	Y
25	Perte supportée ou bénéfice transféré	Z
26	Produits financiers de participations	AA
27	Produits des autres valeurs mobilières et créances de láctif immobilisé	AB
28	Autres intérêts et produits assimilés	AC
29	Reprises sur provisions et transferts de charges financier	AD
30	Différences positives de change	AE
31	Produits nets sur cessions de valeurs mobilières de placement	AF
32	Total des produits financiers	AG
33	Dotations financières sur amortissements et provisions	AH
34	Intérêts et charges assimilées	AI
35	Différences négatives de change	AJ
36	Charges nettes sur cessions de valeurs mobilières de placement	AK
37	Total des charges financières	AL
38	Résultat financier	AM
39	Résultat en cours avant impôts	AN
40	Produits exceptionnels sur opérations de gestion	AO
41	Produits exceptionnels sur opérations en capital	AP
42	Reprises sur provisions et transferts de charges exceptionnel	AQ
43	Total des produits exceptionnels	AR
44	Charges exceptionnelles sur opérations de gestion	AS
45	Charges exceptionnelles sur opérations en capital	AT
46	Dotations exceptionnelles aux amortissements et provisions	AU
47	Total des charges exceptionnelles	AV
48	Résultat exceptionnel	AW
49	Participation des salariés aux résultats de léntreprise	AX
50	Impôts sur les bénéfices	AY
51	Total des produits	AZ
52	Total des charges	BA
53	Bénéfices ou perte (Total des produits - Total des charges)	BB
54	Impôts différés (compte de résultat)	BC
55	Résultat net des sociétés mises en équivalence	BD
56	Résultat net des entreprises intégrées	BE
57	Résultat Groupe (Résultat net consolidé)	BF
58	Part des intérêts minoritaires (Résultat hors groupe)	BG
59	Résultat net part du groupe (part de la société mère)	BH
60	Rémunération díntermédiaires et honoraires (hors rétrocessions)	BI
61	Location, charges locatives et de copropriété	BJ
62	Effectif moyen du personnel	BK
63	Sous-traitance	BL
64	Personnel extérieur à léntreprise	BM
65	Rétrocessions dhonoraires, commissions et courtages	BN
66	Taxe professionnelle	BO
67	Montant de la TVA. collectée	BP

N°	Variable	Code
68	Total TVA. déductible sur biens et services	BQ
69	Dividendes	BR
70	siren	BS

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 :

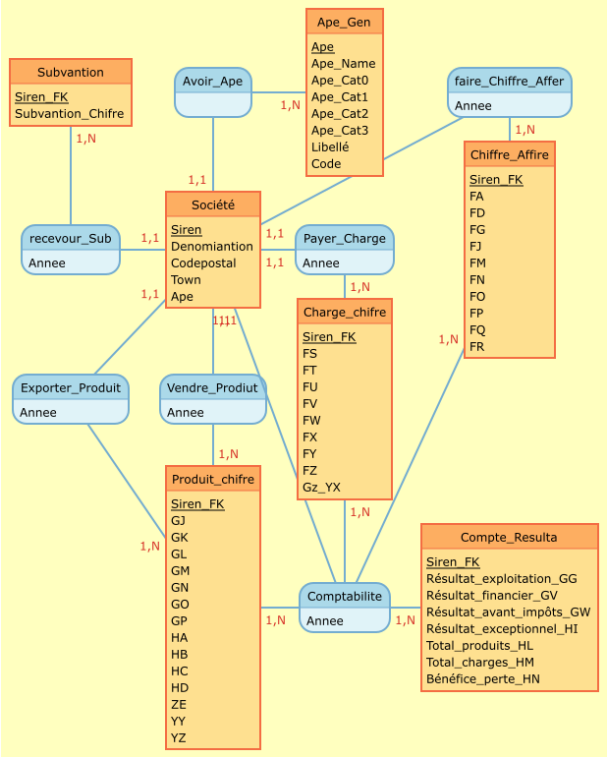

FIGURE 2.1: MCD

— Pour le MOD, inclure une image réalisée avec le logiciel MySQl telle que celle visible sur la Figure 2.2 ci-dessous :

Table	Action	Lignes	Type	Interclassement	Taille	Perte
apegen	Parcourir Structure Rechercher Insérer Vider Supprimer	500	InnoDB	utf8mb3_general_ci	80,0 kio	-
charge_chiffre	Parcourir Structure Rechercher Insérer Vider Supprimer	500	InnoDB	utf8mb3_general_ci	64,0 kio	-
chiffre_affaire	Parcourir Structure Rechercher Insérer Vider Supprimer	500	InnoDB	utf8mb3_general_ci	80,0 kio	-
comptabilit__	Parcourir Structure Rechercher Insérer Vider Supprimer	500	InnoDB	utf8mb3_general_ci	16,0 kio	-
compte_resultat	Parcourir Structure Rechercher Insérer Vider Supprimer	500	InnoDB	utf8mb3_general_ci	80,0 kio	-
produit_chiffre	Parcourir Structure Rechercher Insérer Vider Supprimer	500	InnoDB	utf8mb3_general_ci	16,0 kio	-
societe	Parcourir Structure Rechercher Insérer Vider Supprimer	500	InnoDB	utf8mb3_general_ci	80,0 kio	-
subvention	Parcourir Structure Rechercher Insérer Vider Supprimer	500	InnoDB	utf8mb3_general_ci	16,0 kio	-
8 tables	Somme	4 000	InnoDB	utf8mb4_0900_ai_ci	432,0 kio	0 o

FIGURE 2.2: MOD_2

2.4 Import des données

Dans un premier temps, les données ont été chargées dans R à l'aide de la fonction *read_csv*(). Ensuite, afin de faciliter l'importation de ces données dans phpMyAdmin, nous avons procédé à un **découpage (slicing)** de la base initiale en **sept tables distinctes**, chacune regroupant les variables pertinentes pour son thème spécifique.

Par exemple, pour la table **"société"**, nous avons conservé uniquement les colonnes suivantes : *"siren", "denomination", "postal_ code", "town" et "ape"*. Ce processus a été appliqué à chaque table en sélectionnant les attributs nécessaires, puis nous avons limité chaque sous-table à ses **300 premières lignes** afin de simplifier les tests d'importation.

Enfin, les tables ont été exportées au format CSV à l'aide de la fonction write_csv.

Dans un premier temps, nous avions extrait les **300 premières lignes** de la base de données globale. Toutefois, afin d'obtenir un **échantillon plus équilibré et représentatif dans le temps**, nous avons modifié notre approche : nous avons sélectionné **les 100 premières lignes pour chaque année,** de 2012 à 2016.

Pour cela, nous avons utilisé la bibliothèque **sqldf** dans R, qui permet d'exécuter des requêtes SQL directement sur des data frames. Par exemple, pour extraire les données de l'année 2012, nous avons utilisé la requête suivante :

```
install.packages("sqldf")
library(sqldf)

result <- sqldf("SELECT * FROM data_kaggle WHERE year = 2012")
result
head(result)
annee_2012<-sqldf("SELECT* from data_kaggle WHERE year=2012")
annee_2013<-sqldf("SELECT* from data_kaggle WHERE year=2013")
annee_2014<-sqldf("SELECT* from data_kaggle WHERE year=2014")
annee_2015<-sqldf("SELECT* from data_kaggle WHERE year=2015")
annee_2016<-sqldf("SELECT* from data_kaggle WHERE year=2016")
cor(annee_2012$rentabilite,annee_2012$`Impôts, taxes et versements assimilés`,use = "
plot(annee_2012$rentabilite,annee_2012$`Impôts, taxes et versements assimilés`)
# Table 1 : Société
# Combinaison des 100 premières lignes de chaque sous-ensemble
sous_ensemble <- rbind(
  annee_2012[1:100, ],
  annee_2013[1:100, ],
  annee_2014[1:100, ],
  annee_2015[1:100, ],
  annee_2016[1:100, ]
)
View(sous_ensemble)
```

Cette commande retourne toutes les colonnes de la base data_kaggle pour les lignes dont l'année est **égale à 2012**. Nous avons ensuite appliqué la même méthode

pour chaque année (2013 à 2016), puis extrait **les 100 premières lignes** de chaque sous-ensemble. Enfin, nous avons combiné ces sous-ensembles afin d'obtenir une **table finale regroupant 500 lignes (100 par année)**. Cette nouvelle table est ainsi mieux structurée pour les futures analyses et pourra être **importée dans phpMyAdmin** pour les étapes suivantes de notre projet.

```r
library(readr)
library(dplyr)

# Définir le chemin du fichier source
file_path <- ".....path"

# Charger le fichier CSV
full_data <- read_csv(file_path)

# Vérification des noms de colonnes disponibles
print(colnames(full_data))

# Table 1 : Société
societe <- full_data[1:300, c("siren", "denomination", "postal_code", "town", "ape")]
write_csv(societe, ".....path")

# Table 2 : Subventions
subvention <- full_data[1:300, c("siren", "Subventions d'exploitation")]
write_csv(subvention, ".....path")

# Table 3 : ApeGen
apegen <- full_data[1:300, c("ape", "ape_name", "ape_len", "ape_division", "ape_groupe", "ape_classe", "ape_sous_classe")]
write_csv(apegen, ".....path")

# Table 4 : Chiffre d'Affaires
chiffre_affaire <- full_data[1:300, c("siren", "Chiffres d'affaires nets", "Impôts, taxes et versements assimilés")]
write_csv(chiffre_affaire, ".....path")

# Table 5 : Charges Chiffre
charge_chiffre <- full_data[1:300, c(
  "siren",
  "Reprises sur amortissements et provisions, transfert de charges",
  "Reprises sur provisions et transferts de charges financier",
  "Reprises sur provisions et transferts de charges exceptionnel",
  "Achats de marchandises (y compris droits de douane)",
  "Achats de matières premières et autres approvisionnements",
  "Autres achats et charges externes",
  "Salaires et traitements",
  "Charges sociales"
)]
write_csv(charge_chiffre, ".....path")

# Table 6 : Produits Chiffre
produit_chiffre <- full_data[1:300, c(
  "siren",
  "Total des produits d'exploitation",
  "Total des produits financiers",
  "Total des produits exceptionnels",
  "Autres produits"
)]
write_csv(produit_chiffre, ".....path")

# Table 7 : Compte de Résultat
compte_resulta <- full_data[1:300, c(
  "siren",
  "Chiffres d'affaires nets",
  "Impôts, taxes et versements assimilés",
  "Résultat d'exploitation",
  "Résultat financier",
  "Résultat en cours avant impôts",
  "Résultat exceptionnel",
  "Bénéfices ou perte (Total des produits - Total des charges)"
)]
write_csv(compte_resulta, ".....path")

# Vérifier le répertoire de travail actuel
getwd()
```

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

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

Chapitre 5

Analyse et Résultats

5.1 Comparer les catégories d'entreprises en fonction des Chiffres d'affaires nets

Variables : Chiffres d'affaires nets + catégories (Effectif moyen du personnel)

5.1.1 Étapes pour créer une visualisation :

contuinee ; ; ; ; ; ; ;

Les catégories d'entreprises :

> *L'article 51 de la loi n°2008-776 du 4 août 2008 de modernisation de l'économie (**LME**) détermine, pour les besoins de l'analyse statistique, un classement des entreprises en quatre catégories : les microentreprises, les petites et moyennes entreprises (**PME**), les entreprises de taille intermédiaire (**ETI**) et les grandes entreprises.*
>
> *Le décret n° 2008-1354 du 18 décembre 2008 précise les critères permettant de déterminer l'appartenance à une catégorie d'entreprises.*

- **La microentreprise** est une entreprise dont l'effectif est inférieur à 10 personnes et dont le chiffre d'affaires ou le total du bilan annuel n'excède pas 2 millions d'euros
- **la PME** est une entreprise dont l'effectif est inférieur à 250 personnes et dont le chiffre d'affaires annuel n'excède pas 50 millions d'euros ou dont le total de bilan n'excède pas 43 millions d'euros
- **L'ETI**, entreprise de taille intermédiaire, est une entreprise qui n'appartient pas à la catégorie des PME, dont l'effectif est inférieur à 5000 personnes et dont le chiffre d'affaires annuel n'excède pas 1 500 millions d'euros ou dont le total de bilan n'excède pas 2 000 millions d'euros
- **La grande entreprise** est une entreprise qui ne peut pas être classée dans les catégories précédentes

Visualisation : Boxplot ou Diagramme en barres

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

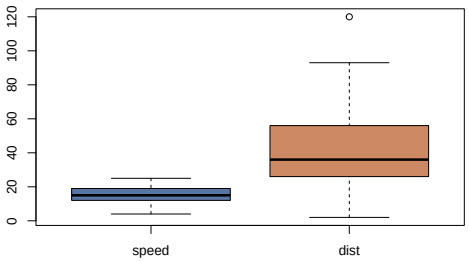

FIGURE 5.1: Deux boxplots.

```r
colMeans(cars)
```

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

```r
#ecrire code R // notre code pour tous calcul or graph etc etc.
```

Test Statistique : ANOVA (Analyse de la Variance) ou les autre

Chauque un/e doit proposer son text :)

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.

5.2 Analyser s'il y a une différence entre les entreprises qui ont recours au refinancement et celles qui n'en ont pas besoin

Variables : indicateur de refinancement + chiffre d'affaires nets

5.2.1 Étapes pour créer une visualisation :

contuinee ; ; ; ; ; ; ;

Les catégories d'entreprises :

Créez deux groupes : Entreprises avec refinancement : Total des charges financières = 0 Entreprises sans refinancement : Total des charges financières = 0

Visualisation : Boxplot ou Diagramme en barres

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

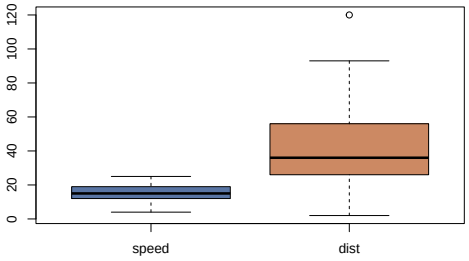

FIGURE 5.2: Deux boxplots.

```
colMeans(cars)
```

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

```
 #ecrire code R // notre code pour tous calcul or graph etc etc.
```

Test Statistique : ANOVA (Analyse de la Variance) ou les autre

Chauque un/e doit proposer son text :) s S 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.

s

5.3 **Analyser la rentabilité des entreprises en fonction de leur localisation géographique** Serdar VAROL

L'objectif de cette étude est d'analyser la rentabilité des entreprises en France à partir de données financières entre 2012 et 2016. La rentabilité a été calculée à partir du *résultat d'exploitation*, du *résultat financier* et du *chiffre d'affaires net* des entreprises et a été examinée par **année** et par **département**.

5.3.1 Préparation des données

1) Un fichier CSV contenant les données financières de l'entreprise de 2012 à 2016 a été utilisé
2) **Variables :**

- Year
- Siren
- résultat d'exploitation
- chiffres d'affaires nets
- résultat financier
- code_postal

$$\text{rentabilité} = \frac{\text{résultat d'exploitation} + \text{résultat financier}}{\text{chiffres d'affaires nets}}$$

3) Création d'une variable catégorielle : La valeur de rentabilité est divisée dans les classes suivantes :

- Rentabilité < 0 : « En perte »
- $0 \leq$ Rentabilité $< 0{,}1$: « Faible rentabilité »
- $0{,}1 \leq$ Rentabilité $< 0{,}3$: « Rentabilité moyenne »
- Rentabilité $\geq 0{,}3$: « Haute rentabilité »

Formuler les hypothèses :

— **Hypothèse nulle (H_0)** : La répartition des catégories de rentabilité (rentabilité) par département ne diffère pas.

— **Hypothèse alternative (H_1)** : La distribution des catégories de rentabilité varie selon le département.

Année	Type de test	Valeur p	Décision ($\alpha = 0{,}05$)	Remarque
2012	Chi² (simulé)	0.2474	H_0 non rejetée	Échantillon insuffisant ou déséquilibré
2013	Chi² (simulé)	0.0019	H_0 rejetée	Différence significative entre départements
2014	Chi² (simulé)	9.999e-05	H_0 rejetée	Différence significative entre départements
2015	Chi² (simulé)	9.999e-05	H_0 rejetée	Forte différence observée
2016	Chi² (simulé)	9.999e-05	H_0 rejetée	Différence significative entre départements

Note : L'année 2012 présente un faible nombre d'observations et des distributions déséquilibrées.

Année	Valeur p	Décision	Interprétation
2012	1	Pas de différence significative	Faible volume de données
2013	< 0.001	Différence significative	Rentabilité varie selon les départements
2014	1	Pas de différence significative	Rentabilité homogène
2015	1	Pas de différence significative	Rentabilité homogène
2016	1	Pas de différence significative	Rentabilité homogène

Conclusion : Seule l'année 2013 présente une variation significative de la rentabilité selon les départements.

Représentation cartographique

— Pour le MOD, inclure une image réalisée avec le logiciel MySQl telle que celle visible sur la Figure 5.3 ci-dessous :

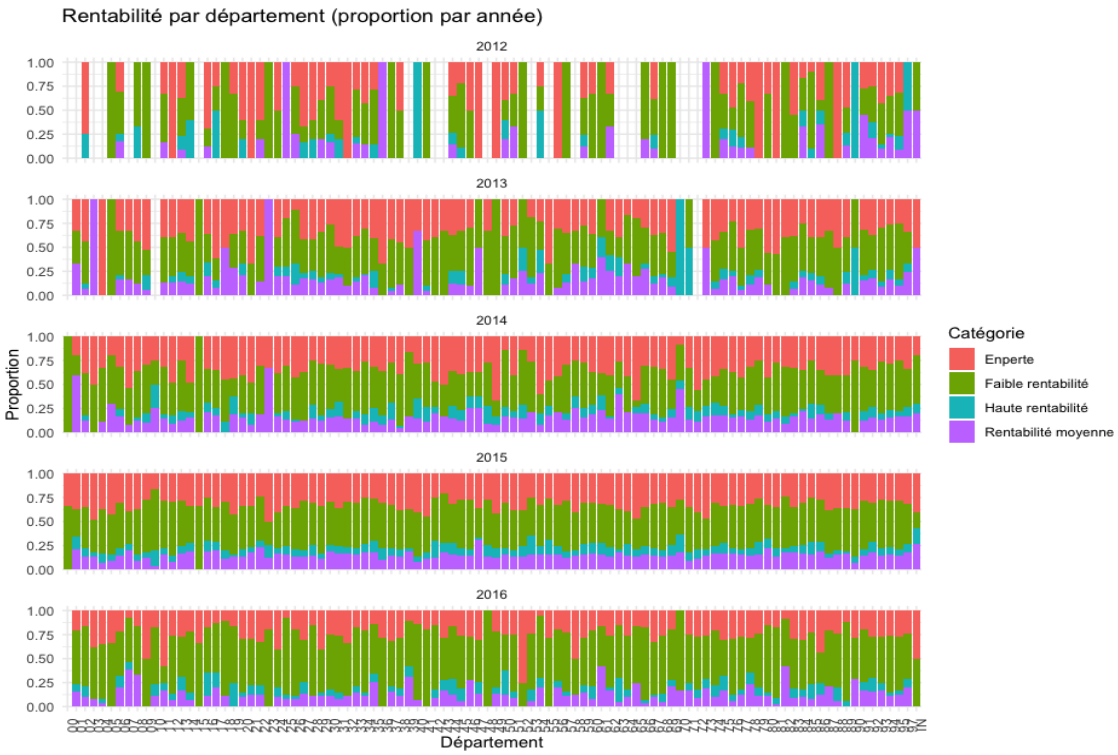

FIGURE 5.3: MOD_2

5.4 La variation de la rentabilité selon le secteur d'activité des entreprises

Variables : le code APE (le secteur d'activité des entreprises) + Chiffres d'affaires nets

5.4.1 Étapes pour créer une visualisation :

Catégorisation des entreprises par secteur d'activité :

Visualisation :

Boxplot : Visualiser la distribution des chiffres d'affaires nets pour chaque secteur d'activité.\ **Diagramme en barres** : Montrer la moyenne des chiffres d'affaires nets par secteur d'activité.

Test Statistique (ANOVA) :

Chauque un/e doit proposer son text :)

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.

5.5 Analyser l'évolution de la rentabilité des entreprises entre 2012 et 2016

Variables : Rentabilité + Année

5.5.1 Étapes pour créer une visualisation :

Préparation des données :

Calcul de la rentabilité : Calculer le ratio de rentabilité pour chaque entreprise et chaque année :

$$\text{Rentabilité} = \frac{\text{Résultat net}}{\text{Chiffre d'affaires net}}$$

Structure des données : Organiser les données dans un tableau avec les colonnes suivantes : Année, Rentabilité.

Visualisation :

Boxplot : Visualiser la distribution des chiffres d'affaires nets pour chaque secteur d'activité.\ **Diagramme en barres** : Montrer la moyenne des chiffres d'affaires nets par secteur d'activité.

Test Statistique (ANOVA) :

Chauque un/e doit proposer son text :)

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.

5.6 Analyser l'impact des taxes sur la rentabilité des entreprises

Variables :Rentabilité Ratio du résultat net sur le chiffre d'affaires net +Taxes (Montant total des taxes payées parles entreprises \Axes : Taxes (X) et Rentabilité (Y).

5.6.1 Étapes pour créer une visualisation :

visualistion

Graphique de dispersion : tracer un graphique de dispersion pour visualiser la relation entre la rentabilité et les taxes.

Test de corrélation :

Chauque un/e doit proposer son text :)

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.

5.7 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.