

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

SOU MIS 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 : _____

23 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 ...

23 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
1.1 Présentation du projet	1
1.2 Responsabilités et composition de l'équipe	1
1.3 Objectifs et questions de recherche	1
Chapitre 2 Base de données	3
2.1 Provenance des données	3
2.2 Descriptif des tables	3
2.2.1 Table 1 : APE_Fusion.csv	3
2.2.2 Table 2 : Profit and loss - Ontology.csv	3
2.2.3 Table 3 : Data_Kaggle.csv	4
2.3 Modèles MCD et MOD	6
2.4 Import des données	10
2.5 Requêtes réalisées	11
Chapitre 3 Matériel et Méthodes	13
3.1 Logiciels	13
3.2 Modélisation statistique	13
Chapitre 4 Analyse et Résultats	14
4.1 Comparer les catégories d'entreprises en fonction des Chiffres d'affaires nets	14
4.1.1 Test Statistique : ANOVA	15
4.1.2 Conclusion	22
4.2 Analyse des subventions et du chiffre d'affaires	23
4.3 Analyser des entreprises en fonction de leur localisation géographique	28
4.3.1 Préparation des données	28
4.3.2 Analyse Univariate	30
4.3.3 Analyse Bivariate	32
4.4 La variation de la rentabilité selon le secteur d'activité des entreprises	34
4.4.1 Étapes pour créer une visualisation :	34
Chapitre 5 Discussion	35
Chapitre 6 Conclusion et perspectives	36
Bibliographie	37

Annexes	38
Codes	38
Tables	40

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 :

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 ?

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 ?

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 ?

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
Unnamed : 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 d'affaires nets	G
7	Production stockée	H
8	Production immobilisée	I
9	Subventions d'exploitation	J
10	Reprises sur amortissements et provisions, transfert de charges	K
11	Autres produits	L
12	Total des produits d'exploitation	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 d'exploitation	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 l'actif 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 l'entreprise	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 d'intermédiaires 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 l'entreprise	BM
65	Retrocessions d'honoraires, 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 :

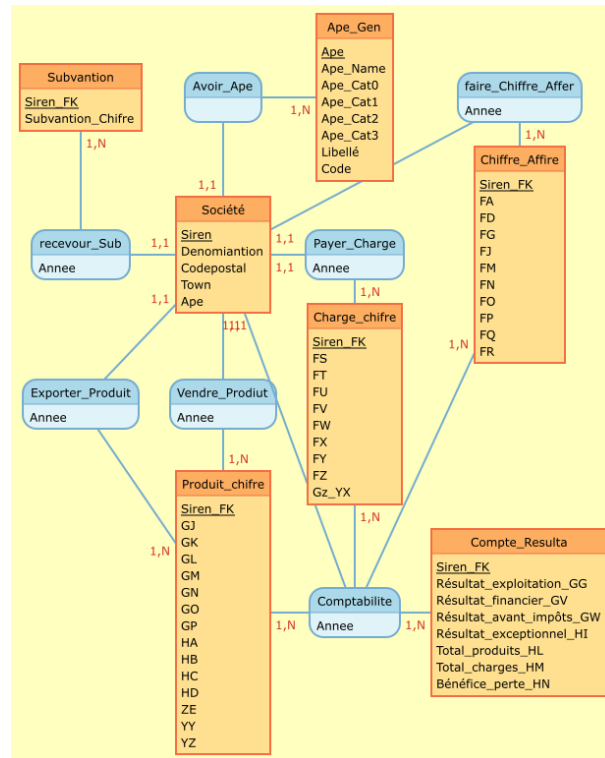


FIGURE 2.1: MCD

- Pour le MOD, inclure les images réalisées avec le logiciel MySQL, telle que celles visible sur la Figure ci-dessous :

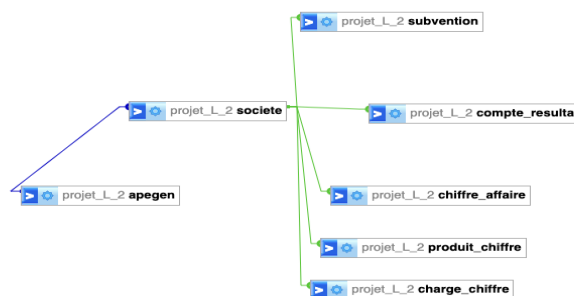


FIGURE 2.2: MOD

Table	Action	Lignes	Type	Interclassement	Taille	Perte
<input type="checkbox"/> apegen	Parcourir Structure Rechercher Insérer Vider Supprimer	188	InnoDB	utf8mb3_general_ci	64,0 kio	-
<input type="checkbox"/> charge_chiffre	Parcourir Structure Rechercher Insérer Vider Supprimer	500	InnoDB	utf8mb3_general_ci	64,0 kio	-
<input type="checkbox"/> chiffre_affaire	Parcourir Structure Rechercher Insérer Vider Supprimer	412	InnoDB	utf8mb3_general_ci	80,0 kio	-
<input type="checkbox"/> compte_resultat	Parcourir Structure Rechercher Insérer Vider Supprimer	500	InnoDB	utf8mb3_general_ci	80,0 kio	-
<input type="checkbox"/> produit_chiffre	Parcourir Structure Rechercher Insérer Vider Supprimer	500	InnoDB	utf8mb3_general_ci	16,0 kio	-
<input type="checkbox"/> societe	Parcourir Structure Rechercher Insérer Vider Supprimer	412	InnoDB	utf8mb3_general_ci	80,0 kio	-
<input type="checkbox"/> subvention	Parcourir Structure Rechercher Insérer Vider Supprimer	500	InnoDB	utf8mb3_general_ci	16,0 kio	-
7 tables	Somme	3 012	InnoDB	utf8mb4_0900_ai_ci	400,0 kio	0 o

FIGURE 2.3: MOD Table Total

#	Nom	Type	Interclassement	Attributs	Null	Valeur par défaut	Commentaires	Extra	Action
<input type="checkbox"/> 1	siren	int			Non	Aucun(e)			Modifier Supprimer Plus
<input type="checkbox"/> 2	denomination	varchar(60)	utf8mb3_general_ci		Oui	NULL			Modifier Supprimer Plus
<input type="checkbox"/> 3	postal_code	varchar(5)	utf8mb3_general_ci		Oui	NULL			Modifier Supprimer Plus
<input type="checkbox"/> 4	town	varchar(31)	utf8mb3_general_ci		Oui	NULL			Modifier Supprimer Plus
<input type="checkbox"/> 5	ape	varchar(5)	utf8mb3_general_ci		Oui	NULL			Modifier Supprimer Plus

FIGURE 2.4: Societe Table

#	Nom	Type	Interclassement	Attributs	Null	Valeur par défaut	Commentaires	Extra	Action
<input type="checkbox"/> 1	year	int			Non	Aucun(e)			Modifier Supprimer Plus
<input type="checkbox"/> 2	siren	int			Non	Aucun(e)			Modifier Supprimer Plus
<input type="checkbox"/> 3	Reprises sur amortissements et provisions, transfert de charges	varchar(8)	utf8mb3_general_ci		Oui	NULL			Modifier Supprimer Plus
<input type="checkbox"/> 4	Reprises sur provisions et transferts de charges financier	varchar(7)	utf8mb3_general_ci		Oui	NULL			Modifier Supprimer Plus
<input type="checkbox"/> 5	Reprises sur provisions et transferts de charges exceptionnel	varchar(7)	utf8mb3_general_ci		Oui	NULL			Modifier Supprimer Plus
<input type="checkbox"/> 6	Achats de marchandises (y compris droits de douane)	varchar(10)	utf8mb3_general_ci		Oui	NULL			Modifier Supprimer Plus
<input type="checkbox"/> 7	Achats de matières premières et autres approvisionnements	varchar(9)	utf8mb3_general_ci		Oui	NULL			Modifier Supprimer Plus
<input type="checkbox"/> 8	Autres achats et charges externes	varchar(9)	utf8mb3_general_ci		Oui	NULL			Modifier Supprimer Plus
<input type="checkbox"/> 9	Salaires et traitements	varchar(9)	utf8mb3_general_ci		Oui	NULL			Modifier Supprimer Plus
<input type="checkbox"/> 10	Charges sociales	varchar(8)	utf8mb3_general_ci		Oui	NULL			Modifier Supprimer Plus

FIGURE 2.5: Charge Chiffre

#	Nom	Type	Interclassement	Attributs	Null	Valeur par défaut	Commentaires	Extra	Action
<input type="checkbox"/>	1 siren	int			Non	Aucun(e)			Plus
<input type="checkbox"/>	2 Chiffres d'affaires nets	varchar(10)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	3 Impôts, taxes et versements assimilés	varchar(8)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	4 Ventes de marchandises	varchar(9)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	5 Production vendue biens	varchar(10)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	6 Production vendue services	varchar(8)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	7 Chiffres d'affaires nets.1	varchar(10)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	8 Production stockée	varchar(9)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	9 Production immobilisée	varchar(7)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	10 Subventions d'exploitation	varchar(7)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	11 Reprises sur amortissements et provisions, transfert de charges	varchar(8)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	12 Autres produits	varchar(7)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	13 Total des produits d'exploitation	varchar(10)	utf8mb3_general_ci		Oui	NULL			Plus

FIGURE 2.6: Chiffre affaire

#	Nom	Type	Interclassement	Attributs	Null	Valeur par défaut	Commentaires	Extra	Action
<input type="checkbox"/>	1 year	int			Non	Aucun(e)			Plus
<input type="checkbox"/>	2 siren	int			Non	Aucun(e)			Plus
<input type="checkbox"/>	3 Chiffres d'affaires nets	varchar(10)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	4 Impôts, taxes et versements assimilés	varchar(8)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	5 Résultat d'exploitation	varchar(8)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	6 Résultat financier	varchar(8)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	7 Résultat en cours avant impôts	varchar(8)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	8 Résultat exceptionnel	varchar(8)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	9 Bénéfices ou perte (Total des produits - Total des charges)	varchar(8)	utf8mb3_general_ci		Oui	NULL			Plus

FIGURE 2.7: Compte de Résultat

#	Nom	Type	Interclassement	Attributs	Null	Valeur par défaut	Commentaires	Extra	Action
<input type="checkbox"/>	1 year	int			Non	Aucun(e)			Plus
<input type="checkbox"/>	2 siren	int			Non	Aucun(e)			Plus
<input type="checkbox"/>	3 Total des produits d'exploitation	varchar(10)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	4 Total des produits financiers	varchar(7)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	5 Total des produits exceptionnels	varchar(8)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	6 Autres produits	varchar(7)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	7 Produits financiers de participations	varchar(7)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	8 Produits des autres valeurs mobilières	varchar(6)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	9 Autres intérêts et produits assimilés	varchar(7)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	10 Reprises sur provisions et transferts de charges financier	varchar(7)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	11 Différences positives de change	varchar(7)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	12 Dotations financières sur amortissements et provisions	varchar(7)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	13 Total des produits financiers.1	varchar(7)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	14 Produits exceptionnels sur opérations de gestion	varchar(7)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	15 Produits exceptionnels sur opérations en capital	varchar(7)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	16 Reprises sur provisions et transferts de charges exceptionnel	varchar(7)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	17 Total des produits exceptionnels.1	varchar(8)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	18 Dividendes	varchar(7)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	19 Montant de la TVA. collectée	varchar(7)	utf8mb3_general_ci		Oui	NULL			Plus
<input type="checkbox"/>	20 Total TVA. déductible sur biens et services	varchar(7)	utf8mb3_general_ci		Oui	NULL			Plus

FIGURE 2.8: Produit Chiffre

#	Nom	Type	Interclassement	Attributs	Null	Valeur par défaut	Commentaires	Extra	Action
<input type="checkbox"/>	1 year 🔑	int			Non	Aucun(e)			Modifier Supprimer Plus
<input type="checkbox"/>	2 siren 🔑	int			Non	Aucun(e)			Modifier Supprimer Plus
<input type="checkbox"/>	3 Subventions d'exploitation	varchar(7)	utf8mb3_general_ci		Oui	NULL			Modifier Supprimer Plus

FIGURE 2.9: Subvention

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 avons 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 :

```
annee_2012 <- sqldf("SELECT * FROM data_kaggle WHERE year = 2012")
```

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.

Lors de l'importation de données sur phpMyAdmin, on a rencontré un petit problème : la colonne était trop longue pour être insérée dans la table. Du coup, on a modifié simplement le nom de la colonne. Elle était : « Produits des autres valeurs mobilières et créances de l'actif immobilisé » et on l'a changée pour : « Produits des autres valeurs mobilières ».

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.

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.

- 1) Pour comparer et trouver selon leur code postal

```
SELECT
  LEFT(postal_code, 2) AS departement,
  COUNT(*) AS nombre_entreprises
FROM
  societe
WHERE
  postal_code IS NOT NULL
GROUP BY
  departement
ORDER BY
  nombre_entreprises DESC;
```

departement	nombre_entreprises
13	64
44	45
75	33
06	33
80	25
04	22
94	14
59	14
21	11
97	11
NA	11
30	9
92	8
20	7
95	6
38	6
85	5
83	5
12	4
67	4
87	4
28	4
16	4
45	3
69	3

FIGURE 2.10: Code postal

```

SELECT
    s.siren,
    s.denomination,
    s.postal_code,
    cr.`Chiffres d'affaires nets` AS chiffre_affaires_net
FROM
    societe s
JOIN
    compte_resultat cr ON s.siren = cr.siren
WHERE
    LEFT(s.postal_code, 2) = '44';

```

siren	denomination	postal_code	chiffre_affaires_net
5580113	L'ABRI FAMILIAL	44603	- Cliquer pour trier les résultats
5680541	IMMOBILIERE ATELIERS DAVID	44350	colonne.
5780044	PEINTURES SYNTHETIQUES MODERNES	44600	- Maj+Clc pour ajouter cette co
5780283	STE HALLEREAU & CIE	44350	ORDER BY ou pour basculer AS
5780390	TRANSPORTS DE LA BRIERE	44600	- Ctrl+Clc ou Alt+Clc (sur Mac
5780960	SOCIETE IMMOBILIERE TOURISTIQUE ET HOTELIERE DE LA...	44500	Maj+Option+Clc) pour enlever
5781133	IDEA GROUPE	44550	clause ORDER BY
5880596	GEDIMO HOLDING	44460	
5980016	SOCIETE NAZAIRIENNE DE MECANIQUE	44550	
5980016	SOCIETE NAZAIRIENNE DE MECANIQUE	44550	
5980016	SOCIETE NAZAIRIENNE DE MECANIQUE	44550	
6076434	MENUISERIE CHARPENTE PIED PERRAUD	44380	
6180301	SOCIETE FINANCIERE ATLANTIC SOFIA	44600	
6280234	LE GAL	44550	
6280309	COMPTOIR ATLANTIQUE DONGEOIS DE DISTRIBUTION ET D ...	44480	
6280309	COMPTOIR ATLANTIQUE DONGEOIS DE DISTRIBUTION ET D ...	44480	
6280309	COMPTOIR ATLANTIQUE DONGEOIS DE DISTRIBUTION ET D ...	44480	
6380117	IMPRIMMO	44350	
6380158	ESPACE DOMICILE SOCIÉTÉ ANONYME D'HABITATIONS À LO...	44570	
6380315	HORIZON AUTOMOBILES	44570	
6480065	TRANSPORTS MORAND FRIGORIFIQUES	44550	
6580161	SOCIETE DES PRODUITS EN BETON ET MATERIAUX DE CONS...	44160	
6580195	SOCIETE INDUSTRIELLE POUR LE DEVELOPPEMENT DE LA S...	44600	
6780514	SOCIETE FRANCAISE D'OUTILS PROFESSIONNELS SOFOP	44550	
6880090	PONTCHATELAIN D'EQUIPEMENTS CIDRICOLLES	44160	

FIGURE 2.11: Code postal

CHAPITRE 3

Matériel et Méthodes

3.1 Logiciels

Nous avons utilisé ces logiciels lors de la réalisation de notre projet :

- Pretraitement : Libre Office
- Stocker Base Donnée : Mysql /MAMP
- Traitement statistique et écrit : R
- Correction/ remplacement : Regex
- Version control : GitHub
- L'IA : OpenAI et MISTRAL

Voici lien de github notre projet : [TDDT github](#)

3.2 Modélisation statistique

<http://biostatisticien.eu/springer/livreR.pdf>

CHAPITRE 4

Analyse et Résultats

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

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

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

4.1.1 Test Statistique : ANOVA

2012

```
##              Df      Sum Sq   Mean Sq F value Pr(>F)
## categorie      2 3.406e+16 1.703e+16   817.9 <2e-16 ***
## Residuals    204 4.248e+15 2.082e+13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

— Détail du test :

— **Df** : Cela indique les degrés de liberté pour chaque facteur (ici *categorie*) et les résidus.

Nous avons 2 degrés de liberté pour les catégories et 204 pour les résidus.

— **Sum Sq** : La somme des carrés, qui mesure la variation expliquée par chaque facteur (ici, la variation expliquée par la catégorie d'entreprise) et par les résidus (erreur).

— **Mean Sq** : La moyenne des carrés, obtenue en divisant la somme des carrés par les degrés de liberté. Cela permet d'avoir une idée de la variance.

— **F value** : Le rapport entre la variance expliquée par les catégories et la variance résiduelle (erreur).

Un F élevé (ici 817.9) indique qu'il existe une grande différence entre les groupes.

— **Pr(>F)** : La p-value associée au test F. Elle est ici inférieure à 2e-16, ce qui est très significatif.

— **Test de Signification**

La p-value obtenue est extrêmement faible ($< 2e-16$), ce qui signifie que nous rejetons l'hypothèse nulle (H_0) au seuil de signification de 5 % (0.05).

En d'autres termes, il y a une différence statistiquement significative entre les moyennes des Chiffres d'affaires nets des différentes catégories d'entreprises.

— **Que signifie cette différence ?**

Cela signifie que les entreprises de catégories différentes (par exemple, Microentreprise, PME, ETI, Grande entreprise) ont des Chiffres d'affaires nets très différents.

Il est donc évident qu'une entreprise de type *Microentreprise* n'a pas la même performance en termes de chiffre d'affaires net qu'une PME ou une Grande entreprise.

— **Résumé**

En 2012, les Chiffres d'affaires nets varient de manière significative en fonction de la catégorie d'entreprise.

Cela signifie que les Microentreprises ne génèrent pas le même chiffre d'affaires net que les PME ou les Grandes entreprises.

2013

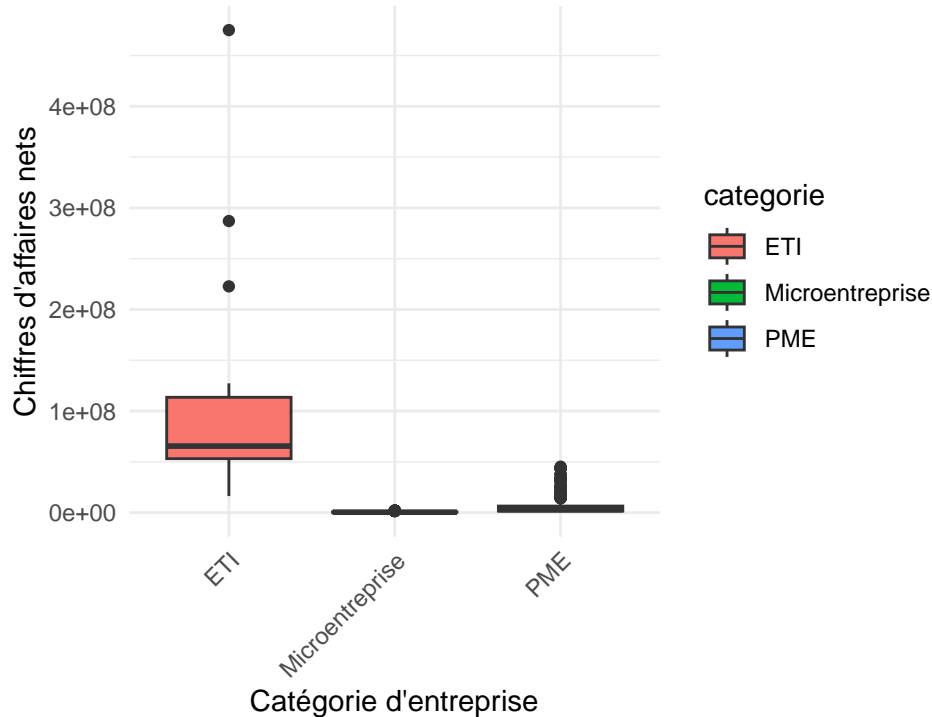
```
##              Df      Sum Sq   Mean Sq F value Pr(>F)
## categorie      2 2.160e+17 1.080e+17   491.2 <2e-16 ***
## Residuals    1102 2.423e+17 2.198e+14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

— Détails du test :

- **Df** (degrés de liberté) : categorie : 2 (ce qui correspond à 3 catégories de classification : Microentreprise, PME, Grande entreprise)
- **Residuals** : 1102 (le nombre de données restantes après l'application des catégories)
- **Sum Sq** (somme des carrés) : Pour la variable categorie, la somme des carrés est de 2.160e+17. Cela représente la variation expliquée par les différences entre les catégories. Pour les résidus, la somme des carrés est de 2.423e+17. Cela représente la variation non expliquée, c'est-à-dire celle qui est attribuée aux erreurs ou à des facteurs non pris en compte dans le modèle.
- **Mean Sq** (moyenne des carrés) : Pour categorie, la moyenne des carrés est de 1.080e+17. Pour les résidus, la moyenne des carrés est de 2.198e+14.
- **F value** : La valeur de F est de 491.2, ce qui est très élevé. Cela indique une forte variation entre les catégories par rapport à la variation résiduelle.
- **Pr(>F)** : La valeur p associée à ce test est $< 2e-16$, ce qui est très inférieur à 0.05. Cela signifie que les différences entre les catégories sont statistiquement significatives.
- **Interprétation des résultats** : Le test ANOVA montre que la variable "Chiffres d'affaires nets" varie de manière significative en fonction de la catégorie d'entreprise (Microentreprise, PME, Grande entreprise). La valeur p très faible ($< 2e-16$) nous permet de conclure que les moyennes des chiffres d'affaires nets sont significativement différentes selon les catégories d'entreprises. En d'autres termes, il y a des différences importantes dans les chiffres d'affaires nets entre les différentes catégories d'entreprises (Microentreprises, PME et Grandes entreprises) pour l'année 2013.
- **Conclusion** : Le test ANOVA révèle que la taille de l'entreprise (catégorie) est un facteur qui a un impact important sur le chiffre d'affaires net des entreprises pour l'année 2013.

Représentation graphique boîte à moustaches :

Boîte à moustaches des Chiffres d'affaires nets selon la catégorie (2013)



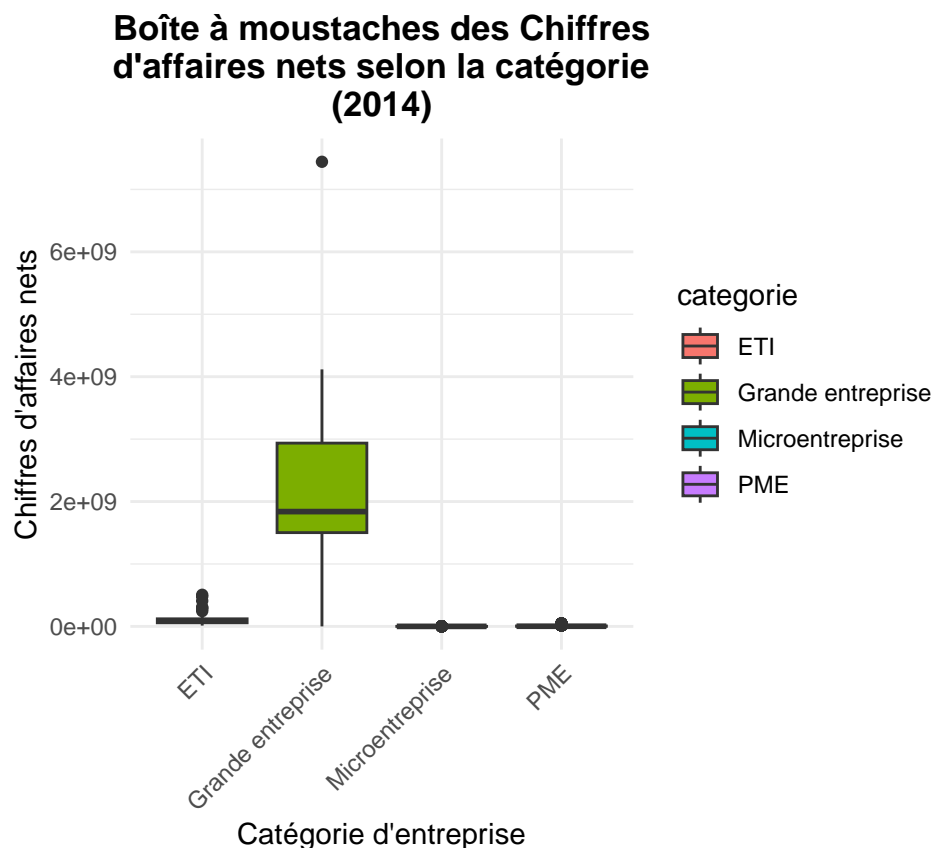
2014

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## categorie      3 5.252e+19  1.751e+19   1118 <2e-16 ***
## Residuals    2716 4.253e+19  1.566e+16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

— Détails du test :

- **Df (degrés de liberté)** : La variable 'categorie' a 3 degrés de liberté, ce qui correspond à 4 catégories : Microentreprise, PME, ETI, et Grande entreprise. Les résidus ont 2716 degrés de liberté, correspondant au nombre de données restantes après l'application des catégories.
- **Sum Sq (somme des carrés)** : Pour la variable 'categorie', la somme des carrés est de 5.252e+19, ce qui représente la variation expliquée par les différences entre les catégories. Pour les résidus, la somme des carrés est de 4.253e+19, représentant la variation non expliquée par le modèle.
- **Mean Sq (moyenne des carrés)** : La moyenne des carrés pour 'categorie' est de 1.751e+19, La moyenne des carrés pour les résidus est de 1.566e+16.
- **F value** : La valeur de F est très élevée, à 1118, ce qui montre qu'il existe une grande différence entre les groupes par rapport à la variation résiduelle. Une valeur de F élevée signifie que les différences entre les groupes sont beaucoup plus grandes que la variation interne à chaque groupe.

- **Pr(>F)** : La valeur p associée à ce test est inférieure à $2e-16$, ce qui est bien en dessous du seuil de 0.05. Cela montre que les différences entre les catégories sont très significatives, ce qui nous permet de rejeter l'hypothèse nulle selon laquelle il n'y a pas de différence entre les catégories.
- **Interprétation des résultats** : Le test ANOVA montre que la variable 'Chiffres_d'affaires_nets' varie de manière significative en fonction de la catégorie d'entreprise (Microentreprise, PME, ETI, Grande entreprise). La valeur p très faible ($< 2e-16$) indique que les moyennes des chiffres d'affaires nets sont statistiquement différentes selon les catégories d'entreprises. Cela signifie que l'appartenance à une catégorie d'entreprise (Microentreprise, PME, ETI, Grande entreprise) a un impact significatif sur les chiffres d'affaires nets pour l'année 2014.
- **Conclusion** : Le test ANOVA confirme que la taille de l'entreprise (catégorie) influence le chiffre d'affaires net des entreprises en 2014. En comparant les moyennes des différentes catégories (Microentreprises, PME, ETI, et Grandes entreprises), on peut dire qu'il y a des différences marquées dans les chiffres d'affaires entre ces groupes.
- Représentation graphique boîte à moustaches :



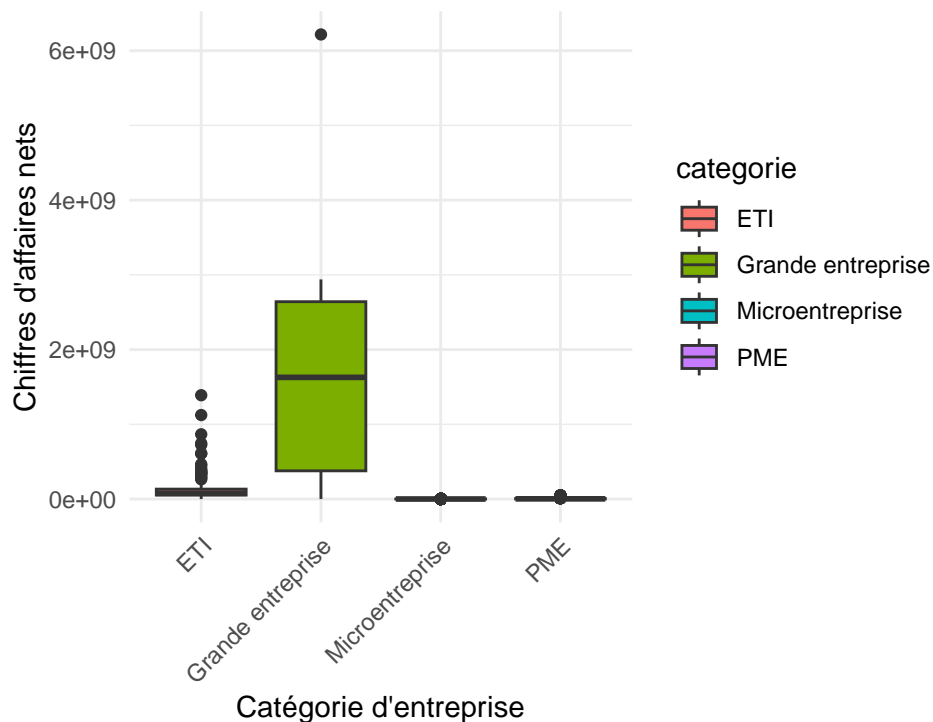
2015

```
##              Df      Sum Sq   Mean Sq F value Pr(>F)
## categorie      3 2.864e+19 9.546e+18   2780 <2e-16 ***
## Residuals    9588 3.293e+19 3.434e+15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

— Détails du test :

- **Df (Degrés de liberté)** : Le facteur “categorie” a 3 degrés de liberté, et les résidus (erreurs) ont 9588 degrés de liberté. Cela reflète le nombre de groupes comparés (les différentes catégories d’entreprises) et la quantité de données disponibles pour les erreurs.
- **Sum Sq (Somme des carrés)** : La somme des carrés représente la variation dans les données. La variation expliquée par la catégorie d’entreprise est de 2.864e+19, tandis que la variation résiduelle (l’erreur) est de 3.293e+19.
- **Mean Sq (Moyenne des carrés)** : La moyenne des carrés est calculée en divisant la somme des carrés par les degrés de liberté. Pour la catégorie, la moyenne des carrés est de 9.546e+18, et pour les résidus, elle est de 3.434e+15.
- **F value** : Le F-value est le rapport entre la variation expliquée par les catégories et celle des résidus. Ici, il est de 2780, ce qui est très élevé, indiquant une différence marquée entre les groupes.
- **Pr(>F)** : La p-value associée au test F. Elle est inférieure à 2e-16, ce qui est très significatif. Cela montre que la probabilité que les différences observées soient dues au hasard est extrêmement faible.
- **Test de Signification** : La p-value obtenue est extrêmement faible (< 2e-16), ce qui signifie que l’hypothèse nulle (H0), qui stipule qu’il n’y a pas de différence entre les groupes, est rejetée. Ainsi, au seuil de 5 % de signification, nous concluons qu’il existe une différence statistiquement significative entre les moyennes des chiffres d’affaires nets des différentes catégories d’entreprises.
- **Que signifie cette différence ?** : Les entreprises de catégories différentes (comme les Microentreprises, PME, ETI, et Grandes entreprises) ont des chiffres d’affaires nets très différents. Par exemple, une Microentreprise n’a clairement pas les mêmes performances financières qu’une PME ou une Grande entreprise.
- **Résumé** : En 2015, les chiffres d’affaires nets varient de manière significative selon la catégorie d’entreprise. Cela suggère que les petites entreprises (comme les Microentreprises) génèrent beaucoup moins de chiffre d’affaires net comparées aux entreprises plus grandes.

Boîte à moustaches des Chiffres d'affaires nets selon la catégorie (2015)



2016

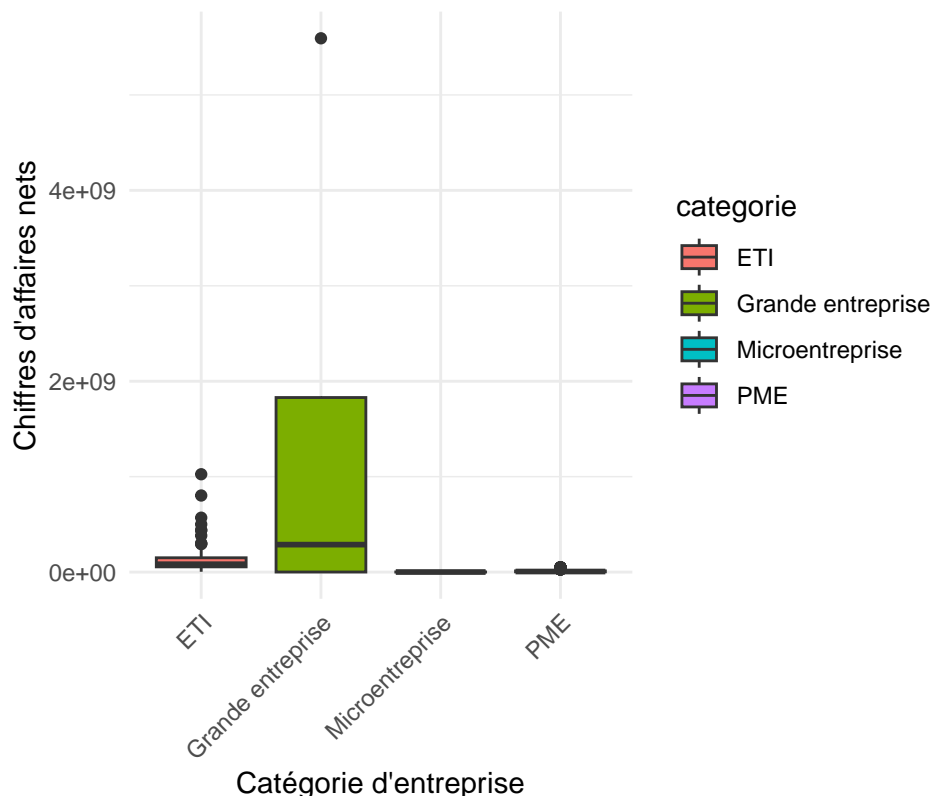
```
##               Df      Sum Sq   Mean Sq F value Pr(>F)
## categorie      3  1.092e+19  3.639e+18   200.9 <2e-16 ***
## Residuals    1358  2.460e+19  1.812e+16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

— Détails du test :

- **Df (Degrés de liberté)** : Le facteur “categorie” a 3 degrés de liberté, et les résidus (erreurs) ont 1358 degrés de liberté. Cela représente le nombre de groupes comparés (les différentes catégories d’entreprises) et la quantité de données restantes pour l’erreur.
- **Sum Sq (Somme des carrés)** : La somme des carrés représente la variation dans les données. La variation expliquée par la catégorie d’entreprise est de 1.092e+19, tandis que la variation résiduelle (l’erreur) est de 2.460e+19.
- **Mean Sq (Moyenne des carrés)** : La moyenne des carrés est calculée en divisant la somme des carrés par les degrés de liberté. Pour la catégorie, la moyenne des carrés est de 3.639e+18, et pour les résidus, elle est de 1.812e+16.
- **F value** : Le F-value est le rapport entre la variation expliquée par les catégories et celle des résidus. Ici, il est de 200.9, ce qui est relativement élevé, indiquant une différence entre les groupes.

- **Pr(>F)** : La p-value associée au test F. Elle est inférieure à $2e-16$, ce qui est extrêmement significatif. Cela montre que la probabilité que les différences observées soient dues au hasard est très faible.
- **Test de Signification** : La p-value obtenue est extrêmement faible ($< 2e-16$), ce qui signifie que l'hypothèse nulle (H_0), qui stipule qu'il n'y a pas de différence entre les groupes, est rejetée. Ainsi, au seuil de 5 % de signification, nous concluons qu'il existe une différence statistiquement significative entre les moyennes des chiffres d'affaires nets des différentes catégories d'entreprises.
- **Que signifie cette différence ?** : Les entreprises de catégories différentes (comme les Microentreprises, PME, ETI, et Grandes entreprises) ont des chiffres d'affaires nets très différents. Une Microentreprise n'a pas les mêmes performances financières qu'une PME ou une Grande entreprise.
- **Résumé** : En 2016, les chiffres d'affaires nets varient de manière significative en fonction de la catégorie d'entreprise. Cela suggère que les petites entreprises (comme les Microentreprises) génèrent beaucoup moins de chiffre d'affaires net comparées aux entreprises plus grandes.
- Représentation graphique (boîte à moustaches) :

Chiffres des Chiffres d'affaires nets selon la catégorie (2016)



4.1.2 Conclusion

Dans cette analyse, nous avons examiné les différences de chiffre d'affaires net entre les différentes catégories d'entreprises (**Microentreprise, PME, ETI, Grande entreprise**) sur la période 2012-2016. Les tests ANOVA ont permis de confirmer que les différences observées sont statistiquement significatives, ce qui montre que la taille de l'entreprise a un impact notable sur ses performances économiques. Les résultats ont montré que les Microentreprises génèrent des chiffres d'affaires nettement inférieurs à ceux des PME et des Grandes entreprises. Les boîtes à moustaches ont illustré visuellement ces différences, renforçant ainsi les conclusions des tests statistiques. En conclusion, cette analyse confirme que la taille de l'entreprise est un facteur déterminant dans ses résultats économiques. Ces résultats peuvent être utilisés pour guider des décisions politiques et stratégiques, notamment pour le soutien aux entreprises en fonction de leur taille.

4.2 Analyse des subventions et du chiffre d'affaires

L'objectif de cette section est de comprendre si les subventions reçues par les entreprises françaises ont un impact sur leur chiffre d'affaires.

Les montants des subventions et du chiffre d'affaires varient énormément entre les entreprises : certaines ne reçoivent presque rien, d'autres ont des montants très élevés. Pour rendre les données plus lisibles et éviter que quelques cas extrêmes ne faussent les résultats, on utilise une transformation logarithmique. Cela permet d'observer les tendances de manière plus équilibrée et de voir si, proportionnellement, plus de subventions entraîne plus de chiffre d'affaires.

1. Nettoyage et préparation des données

Nous avons utilisé connexion MySQL pour examiner cette party, voici le code R :

```
con <- dbConnect(
  MySQL(),
  user = "root",
  password = "root", # même mot de passe créé plus haut
  dbname = "projet_L_2",
  host = "127.0.0.1",
  port = 8889
)

# Vérifie immédiatement la connexion :
#dbListTables(con)

df <- dbGetQuery(con, "
SELECT
  s.siren,
  s.denomination,
  s.town,
  s.ape,
  a.ape_name,
  cr.year,
  cr.`Chiffres d'affaires nets` AS chiffre_affaires,
  sb.`Subventions d'exploitation` AS subvention
FROM projet_L_2.societe s
JOIN projet_L_2.apegen a
  ON s.ape = a.ape
JOIN projet_L_2.compte_resultat cr
  ON s.siren = cr.siren
JOIN projet_L_2.subvention sb
  ON s.siren = sb.siren AND cr.year = sb.year
WHERE cr.`Chiffres d'affaires nets` IS NOT NULL
  AND sb.`Subventions d'exploitation` IS NOT NULL
")
```

```

library(dplyr)
library(ggplot2)

# Nettoyage
df_clean <- df %>%
  mutate(
    chiffre_affaires = as.numeric(gsub("[^0-9]", "", chiffre_affaires)),
    subvention = as.numeric(gsub("[^0-9]", "", subvention))
  ) %>%
  filter(!is.na(chiffre_affaires) & !is.na(subvention))

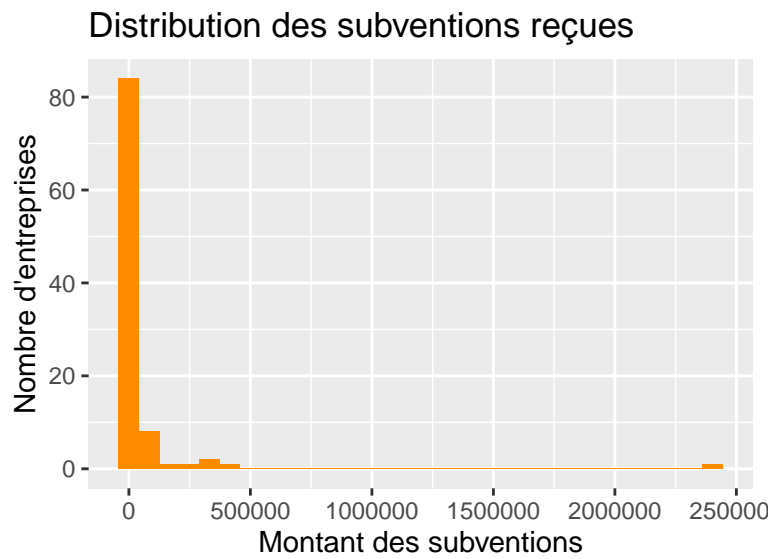
# Création des colonnes logarithmiques
df_log <- df_clean %>%
  filter(chiffre_affaires > 0, subvention > 0) %>%
  mutate(
    log_CA = log10(chiffre_affaires),
    log_subvention = log10(subvention)
  )

# Régression log-log
modele_log <- lm(log_CA ~ log_subvention, data = df_log)
summary(modele_log)

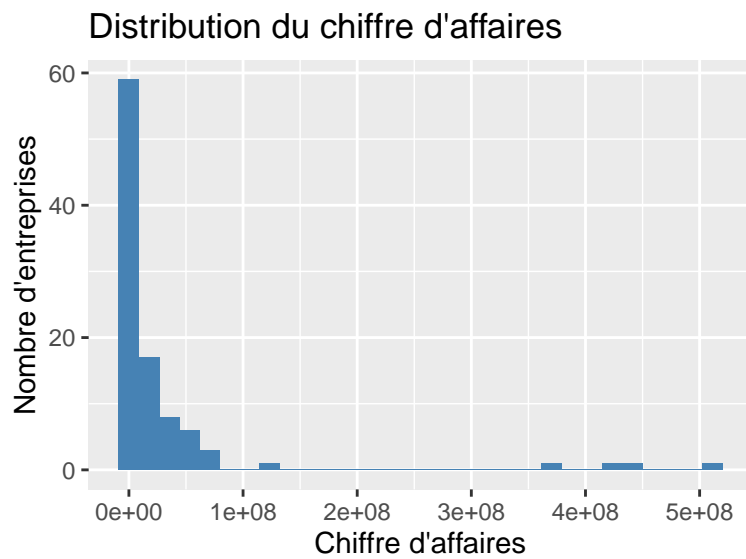
##
## Call:
## lm(formula = log_CA ~ log_subvention, data = df_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.35059 -0.46289  0.02854  0.46093  1.64308
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.11391    0.31714   16.125 < 2e-16 ***
## log_subvention  0.44712    0.08262    5.412 4.59e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6789 on 96 degrees of freedom
## Multiple R-squared:  0.2337, Adjusted R-squared:  0.2258
## F-statistic: 29.28 on 1 and 96 DF,  p-value: 4.589e-07

```

2. Analyse univariée

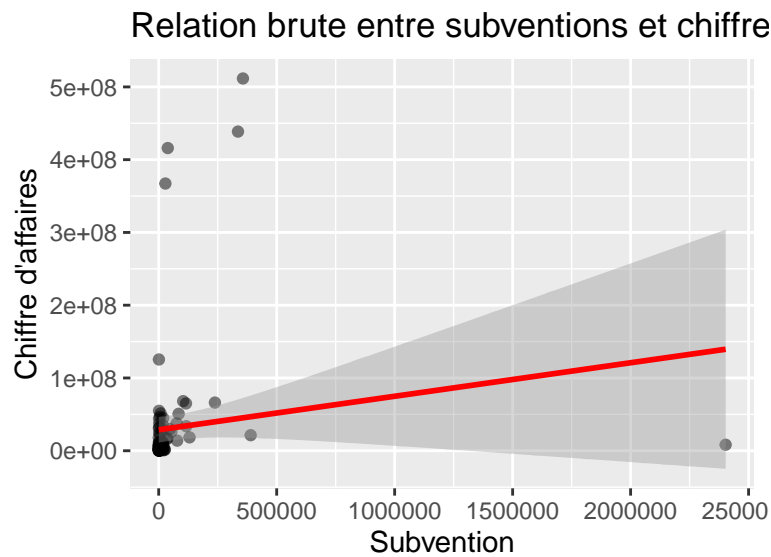


La majorité des entreprises reçoivent de petites subventions. Quelques-unes, plus rares, touchent des montants bien plus élevés.



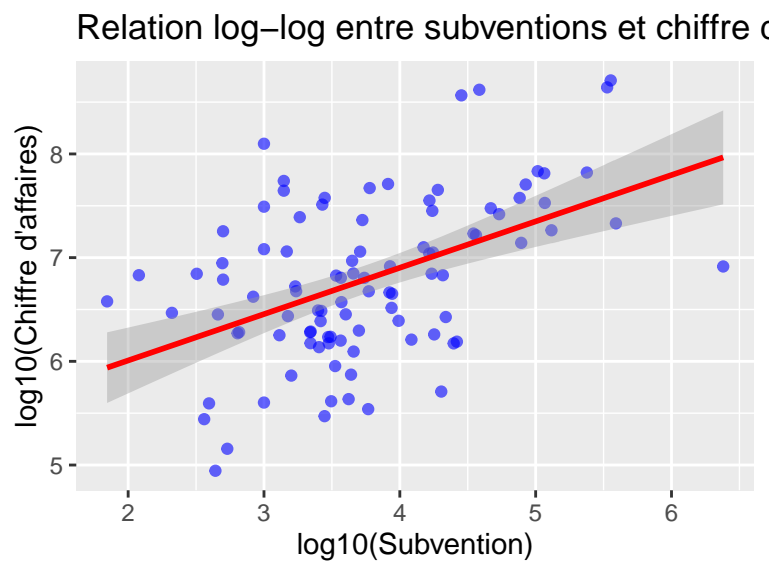
Le chiffre d'affaires est aussi très variable, certaines entreprises réalisant plusieurs dizaines de millions d'euros.

3. Analyse bivariée brute



À première vue, il n'y a pas de lien clair entre les subventions et le chiffre d'affaires. Le graphique montre beaucoup de dispersion, ce qui empêche une vraie lecture de tendance.

4. Régression log-log : Visualisation



Une fois les données transformées en logarithme, une tendance apparaît : les entreprises qui reçoivent plus de subventions ont, en moyenne, un chiffre d'affaires plus élevé. La relation est significative et le modèle indique qu'une augmentation de 10% des subventions correspond à une hausse d'environ 4,5% du chiffre d'affaires.

Conclusion

En conclusion, même si à l'échelle brute le lien entre subventions et chiffre d'affaires n'est pas visible, la transformation log-log montre qu'il existe une **relation proportionnelle significative**. Cela signifie que les subventions peuvent avoir un impact positif mesurable sur le développement économique des entreprises, en particulier si l'on considère les effets relatifs plutôt que les montants absolus. sans refinancement : Total des charges financières = 0

4.3 Analyser des entreprises en fonction de leur localisation géographique

L'objectif de cette étude est d'analyser des entreprises en France à partir de données financières entre 2012 et 2016. Nous avons examiné **le chiffre d'affaires net**, **le résultat d'exploitation**. Des entreprises et a été examinée par **année** et par **region**. La base de données ne contenait que des codes postaux. Nous avons d'abord regroupé les départements puis les régions en fonction des deux chiffres du code postal. Nous avons examiné chaque année indépendamment.

Dans ce chapitre pour analyse univariée, nous allons présenter l'année 2015 car l'année avec le plus de données est 2015. Nous avons constaté que les différences et les incohérences dans la taille des données selon les années et les régions constituaient un obstacle.

Pivot Table_resulta_data_kaggle_2012_2016_order_1

Count - ca	year				
region	2012	2013	2014	2015	2016
Auvergne-Rhône-Alpes	22	242	901	4767	1065
Bourgogne-Franche-Comté	29	89	215	1288	313
Bretagne	26	188	422	1644	160
Centre-Val de Loire	29	218	553	1656	219
Grand Est	59	332	775	3756	511
Hauts-de-France	187	409	947	4082	445
Île-de-France	757	3510	7527	26978	1709
Normandie	54	180	485	1641	189
Nouvelle-Aquitaine	152	461	1039	3934	553
Occitanie	73	633	1507	5544	465
Pays de la Loire	69	225	930	1919	393
Provence-Alpes-Côte d'Azur	182	608	1492	5784	1026
Total Result	1639	7095	16793	62993	7048

FIGURE 4.1: Pivot Table

4.3.1 Préparation des données

- 1) Un fichier CSV (créé par code R) contenant les données financières de l'entreprise de 2012 à 2016 a été utilisé
- 2) Les Variables sont utilise :

Variable	source
Year	Datakagle.csv
Siren	Datakagle.csv
résultat d'exploitation	Datakagle.csv
chiffres d'affaires nets	Datakagle.csv
code_postal	Datakagle.csv
rentabilite	créé
categorie_rentabilite	créé
Department	créé
region	créé

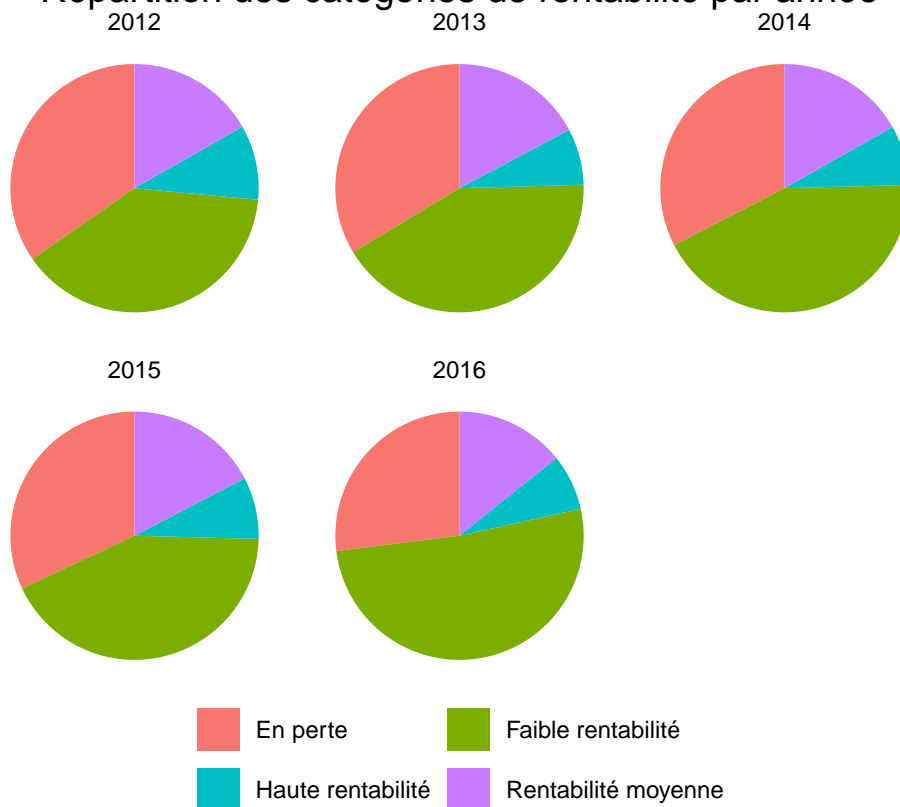
- 3) Création d'une variable catégorielle : La valeur de rentabilité est divisée dans les classes suivantes :

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

TABLE 4.2: Classification des entreprises selon leur rentabilité

Intervalle de rentabilité	Catégorie
Rentabilité < 0	En perte
$0 \leq \text{Rentabilité} < 0,1$	Faible rentabilité
$0,1 \leq \text{Rentabilité} < 0,3$	Rentabilité moyenne
$\text{Rentabilité} \geq 0,3$	Haute rentabilité

Répartition des catégories de rentabilité par année

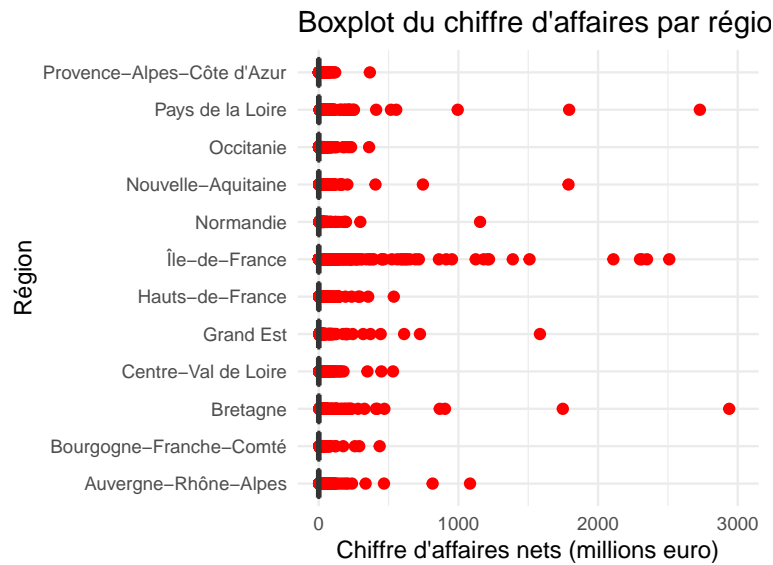


4.3.2 Analyse Univariate

Chiffre d'affaire

TABLE 4.3: Résumé des statistiques du chiffre d'affaires nets (en millions d'euros) – 2015

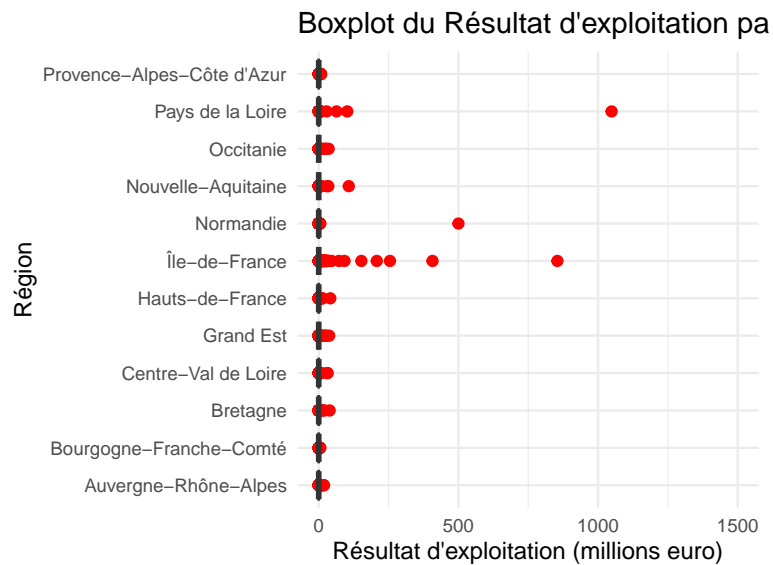
region	avarage	median	min	max	ecart_type
Auvergne-Rhône-Alpes	3.18	0.33	-0.06	1083.39	25.26
Bourgogne-Franche-Comté	4.21	0.56	0.00	436.49	20.71
Bretagne	11.57	0.74	0.00	2938.86	101.94
Centre-Val de Loire	4.06	0.44	-0.12	531.87	23.78
Grand Est	4.36	0.53	-0.04	1584.02	36.31
Hauts-de-France	3.26	0.49	-0.11	537.71	17.26
Normandie	4.52	0.59	-0.01	1155.42	34.62
Nouvelle-Aquitaine	3.51	0.44	-0.05	1788.57	35.47
Occitanie	2.17	0.36	-0.14	360.82	10.56
Pays de la Loire	10.93	0.78	-0.01	2729.00	90.24
Provence-Alpes-Côte d'Azur	1.46	0.35	-6.59	366.13	7.17
Île-de-France	3.91	0.35	-0.31	6217.25	64.37



Resultat d'Explation

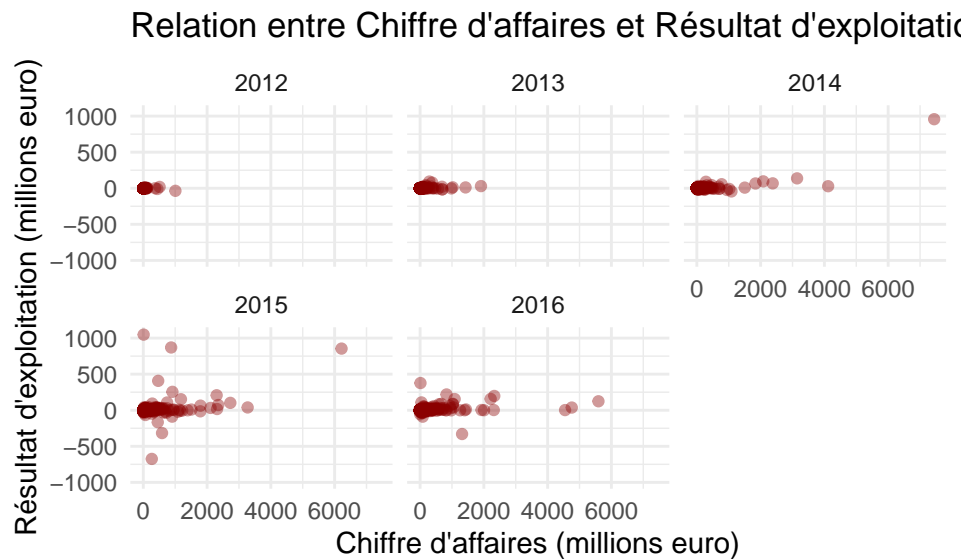
TABLE 4.4: Résumé des statistiques du Résultat d'exploitation (en millions d'euros) – 2015

region	avarage	median	min	max	ecart_type
Auvergne-Rhône-Alpes	3.18	0.33	-0.06	1083.39	25.26
Bourgogne-Franche-Comté	4.21	0.56	0.00	436.49	20.71
Bretagne	11.57	0.74	0.00	2938.86	101.94
Centre-Val de Loire	4.06	0.44	-0.12	531.87	23.78
Grand Est	4.36	0.53	-0.04	1584.02	36.31
Hauts-de-France	3.26	0.49	-0.11	537.71	17.26
Normandie	4.52	0.59	-0.01	1155.42	34.62
Nouvelle-Aquitaine	3.51	0.44	-0.05	1788.57	35.47
Occitanie	2.17	0.36	-0.14	360.82	10.56
Pays de la Loire	10.93	0.78	-0.01	2729.00	90.24
Provence-Alpes-Côte d'Azur	1.46	0.35	-6.59	366.13	7.17
Île-de-France	3.91	0.35	-0.31	6217.25	64.37



4.3.3 Analyse Bivarie

Chiffre d'affaires nets et Résultat d'exploitation

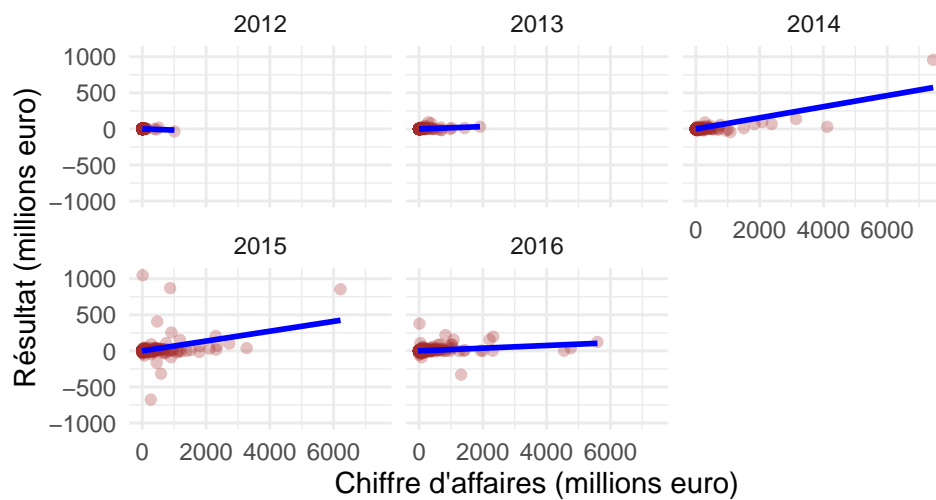


Coefficient de corrélation de Pearson

Année	Corrélation (Pearson)	Interprétation
2012	-0.45	Corrélation négative modérée
2013	0.35	Relation positive modérée
2014	0.80	Forte corrélation positive
2015	0.38	Tendance positive similaire à 2013
2016	0.29	Corrélation positive mais faible

La corrélation doit être comprise entre **-1 et 1**, donc les valeurs que nous avons trouvées sont correctes. On peut dire qu'en 2012, il existe une relation **négative** entre les deux variables. Pour les années 2013, 2015 et 2016, la corrélation est **faible mais positive** : cela signifie que lorsque le chiffre d'affaires augmente, le résultat d'exploitation a tendance à augmenter également. En 2014, on observe une **forte** corrélation positive, ce qui indique une relation claire entre les deux variables : elles ont tendance à évoluer dans le même sens, c'est-à-dire à augmenter en même temps.

Régression linéaire entre CA et Résultat d'exploitation



Formuler les hypothèses :

- Hypothèse nulle (H_0) : La répartition du chiffre d'affaires net et du résultat d'exploitation ne varie pas selon les régions.
- Hypothèse alternative (H_1) : Il existe une différence significative du chiffre d'affaires net et du résultat d'exploitation entre les régions.

##	Annee	F_value	p_value	Decision
## 2012	2012	11.49	0.000000000	H0 rejetée
## 2013	2013	6.25	0.000000000	H0 rejetée
## 2014	2014	3.77	0.000020167	H0 rejetée
## 2015	2015	7.46	0.000000000	H0 rejetée
## 2016	2016	2.76	0.001434238	H0 rejetée

##	Annee	F_value	p_value	Decision
## 2012	2012	0.72	0.7195	H0 non rejetée
## 2013	2013	0.61	0.8213	H0 non rejetée
## 2014	2014	0.18	0.9987	H0 non rejetée
## 2015	2015	1.95	0.0289	H0 rejetée
## 2016	2016	0.44	0.9379	H0 non rejetée

Conclusion

4.4 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

4.4.1 *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) :

Chaque 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.

CHAPITRE 5

Discussion

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

CHAPITRE 6

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

```
#Charger la bibliothèque
library(readr)

#Lire le fichier CSV
data_kaggle <- read_csv("csv/data_kaggle.csv")

#Fonction pour calculer la rentabilité (PAS de référence directe à data_kaggle dans les paramètres)
fonction_rentabilite <- function(resultat_financier, ca_net) {
  if (!is.na(resultat_financier) && !is.na(ca_net) && ca_net != 0) {
    return(resultat_financier / ca_net)
  } else {
    return(NA)
  }
}

#Appliquer la fonction ligne par ligne
data_kaggle$rentabilite <- apply(data_kaggle, 1, function(row) {
  resultat <- as.numeric(row["Résultat financier"])
  ca <- as.numeric(row["Chiffres d'affaires nets"])
  return(fonction_rentabilite(resultat, ca))
})

#Sauvegarder le résultat dans un nouveau fichier CSV
write_csv(data_kaggle,
  "csv/data_kaggle_new.csv",
  row.names = FALSE)

#Affichage pour vérification
View(data_kaggle)
data_kaggle$rentabilite()
annee_2012 <- data_kaggle$year[]
annee_2012

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")
View(annee_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 = "complete.obs")
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)

# Sélection des colonnes souhaitées
societe <- sous_ensemble[, c("siren", "denomination", "postal_code", "town", "ape")]
View(societe)
write_csv(societe, "csv/societe.csv")

# Table 2 : Subventions
```

```

sous_ensemble_2 <- rbind(
  annee_2012[1:100, ],
  annee_2013[1:100, ],
  annee_2014[1:100, ],
  annee_2015[1:100, ],
  annee_2016[1:100, ]
)
subventions<-sous_ensemble_2[, c("year","siren", "Subventions d'exploitation")]

write_csv(subventions, "csv/subvention.csv")

# Table 3 : ApeGen
sous_ensemble_3 <- rbind(
  annee_2012[1:100, ],
  annee_2013[1:100, ],
  annee_2014[1:100, ],
  annee_2015[1:100, ],
  annee_2016[1:100, ]
)

apegen<-sous_ensemble_3[, c("ape", "ape_name", "ape_len", "ape_division", "ape_groupe", "ape_classe", "ape_sous_classe")]
write_csv(apegen, "csv/apegen.csv")

# Table 4 : Chiffre d'Affaires
sous_ensemble_4<- rbind(
  annee_2012[1:100, ],
  annee_2013[1:100, ],
  annee_2014[1:100, ],
  annee_2015[1:100, ],
  annee_2016[1:100, ]
)
chiffre_affaire<-sous_ensemble_4[,c("siren", "Chiffres d'affaires nets", "Impôts, taxes et versements assimilés")]

write_csv(chiffre_affaire, "csv/chiffre_affaire.csv")

# Table 5 : Charges Chiffre
sous_ensemble_5<- rbind(
  annee_2012[1:100, ],
  annee_2013[1:100, ],
  annee_2014[1:100, ],
  annee_2015[1:100, ],
  annee_2016[1:100, ]
)
charge_chiffre <- sous_ensemble_5[, c("year",
"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, "csv/charge_chiffre.csv")

# Table 6 : Produits Chiffre
sous_ensemble_6<- rbind(
  annee_2012[1:100, ],
  annee_2013[1:100, ],
  annee_2014[1:100, ],
  annee_2015[1:100, ],
  annee_2016[1:100, ]
)

produit_chiffre <- sous_ensemble_6[, c("year",
"siren",
"Total des produits d'exploitation",
"Total des produits financiers",
"Total des produits exceptionnels",
"Autres produits"
)]
write_csv(produit_chiffre, "csv/produit_chiffre.csv")

# Table 7 : Compte de Résultat
sous_ensemble_7<- rbind(
  annee_2012[1:100, ],
  annee_2013[1:100, ],
  annee_2014[1:100, ],
  annee_2015[1:100, ],
  annee_2016[1:100, ]
)

compte_resultat <- sous_ensemble_7[, c("year",
"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_resultat, "csv/compte_resultat.csv")
stock <- sous_ensemble_7[ c("year")]
write_csv(stock, "csv/stock.csv")
```

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 L^AT_EX.

TABLE 6.1: 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.