PROJET RÉALISÉ PAR L'ÉQUIPE GROUPE 6

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

LOUATI Chamss-Eddine, NDIAYE Ibrahima , EL OUALYDY Mohamed Amine, SARTORI Adrien

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

Mai 2023

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

Déclaration de non plagiat

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: _____

Signature: _____ Date: _____

Signature: _____ Date: _____

Signature: _____ Date: _____

Remerciements

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

03/04/2023.

Résumé

Ce projet consiste à étudier des bases de données pour déterminer les secteurs qui enregistrent le plus gros chiffre d'affaire au cours des trois dernières années (2020, 2021, 2022), en particulier dans quelle région. Nous avons importé nos données dans une base de données SQL, puis effectué des analyses statistiques à l'aide de requêtes SQL et de R. Nous avons utilisé R pour effectuer une analyse exploratoire de nos données et identifier les secteurs d'activité les plus performants. Nous avons ensuite créé des visualisations telles que des diagrammes en barre, des nuages de points et des cartes pour mieux illustrer les tendances économiques en France. Nos résultats peuvent aider les investisseurs à mieux comprendre les tendances économiques en France et à prendre des décisions d'investissement plus éclairées pour maximiser leur rentabilité. .

Table des matières

Chapitre 1

Introduction

1.1 Contexualisation

En France, les secteurs qui enregistrent les plus gros chiffres d'affaires ont une importance particulière en raison de leur contribution significative à l'économie du pays. Connaître les secteurs les plus performants peut aider les entreprises à mieux cibler leurs investissements et à maximiser leur rentabilité. Les gouvernements régionaux peuvent également utiliser ces informations pour développer des politiques économiques qui favorisent les secteurs les plus performants et stimulent la croissance économique régionale. Les trois secteurs économiques principaux sont:

- le secteur primaire : collecte et l'exploitation des ressources naturelles (matériaux, énergie, et certains aliments);
- le secteur secondaire : industries de transformation des matières premières;
- le secteur tertiaire : les industries du service;

Le but de ce projet sera de voir:

Quels sont les secteurs d'activités qui enregistrent les plus gros chiffres d'affaires et dans quelles régions au cours de ces trois dernières années(2020, 2021, 2022) ?

Pour ce faire, nous étudierons les chiffres d'affaires des entreprises aisni que leurs localisations.

Les données utilisées seront celles trouvées sur le site :
https://www.data.gouv.fr/fr/datasets/chiffres-cles-2022/

La question des secteurs d' activities qui enregistrent les plus gros chiffres d'affaires revêt d'une grande importance pour un large éventail d'acteurs économiques, notamment les investisseurs, les entreprises, les gouvernements, les régulateurs et les consommateurs. Ainsi, l'actualité de cette question est constante car les secteurs qui enregistrent les plus gros chiffres d'affaires évoluent régulièrement en fonction des changements économiques, technologiques et sociaux. Par conséquent,les éléments de réponse à cette question pourront être utilisés pour des actions variées.

Chapitre 2

Base de données

2.1 Descriptif des tables

Notre base de données a été créée à partir d'un fichier appelé "Chiffres clés 2022" disponible sur le site web gouvernemental français "Data.gouv.fr"(https://www.data.gouv.fr/fr/datasets/chiffres-cles-2022/). Le fichier initial contenait 167 885 lignes et 42 colonnes. Après analyse et traitement des données, on a choisi de retenir trois tables principales pour notre étude.

- **Table_Activite**(5 colonnes, 443 lignes): Cette table a été créée pour mettre en évidence chaque secteur d'activité dans lequel l'entreprise appartient. On l'a réalisé en regroupant tous les secteurs d'activité pour faire la somme de leur chiffre d'affaire par année. Ce qui fait que dans cette table il n'y a pas de redondances au niveau de nos secteurs d'activité et que et pour chaque secteur on a le total de son chiffre d'affaire par années. Ainsi, elle est composée:
  - code d'activité(qui est de type **varchar** qui signifie caractère variable en francais),
  - secteur d'activite(**varchar**),
  - total chiffre d'affaire 2022(**bigint**, qui signifie gros entier en francais),
  - total chiffre d'affaire 2021(**bigint**),
  - total chiffre d'affaire 2020(**bigint**)
- **Table_Entreprise**(3.264 lignes,11 colonnes) : C'est notre table principale, où il y a le plus de colonnes et c'est la table où l'on retrouve toutes les caractéristiques de chaque entreprise:
  - identifiant de l'entreprise(**int**, qui signifie entier en francais),
  - nom de l'entreprise(**varchar**),
  - SIREN(**int**),
  - code_activite(**int**):clé étrangére référençant table_activite,
  - région (**varchar**), clé étrangére référençant table_region,
  - chiffre d'affaire 2022(**bigint**),
  - chiffre d'affaire 2021(**bigint**),
  - chiffre d'affaire 2020(**bigint**),
  - effectif 2022(**int**),
  - effectif 2021(**int**),
  - effectif 2022(**int**)
- **Table_Region**(17 lignes, 4 colonnes): Ici aussi, on a realisé presque le même calcul pour les secteurs d'activités, on a regroupé toutes les régions de toutes les entreprises, et on a effectué la somme de leur chiffre d'affaire pour chaque année. Voici nos colonnes:

- nom de la region(**varchar**)
- total chiffre d'affaire 2022(**bigint**)
- total chiffre d'affaire 2021(**bigint**)
- total chiffre d'affaire 2022(**bigint**)

En résumé, notre base de données a été construite à partir d'un fichier de données public, elle contient trois tables principales regroupant les chiffres d'affaires et les caractéristiques des entreprises étudiées, et elle est destinée à répondre à notre problématique de recherche des secteurs d'activités les plus performants en termes de chiffres d'affaires et de localisation.

2.2 Modèles MCD et MOD

- MCD :

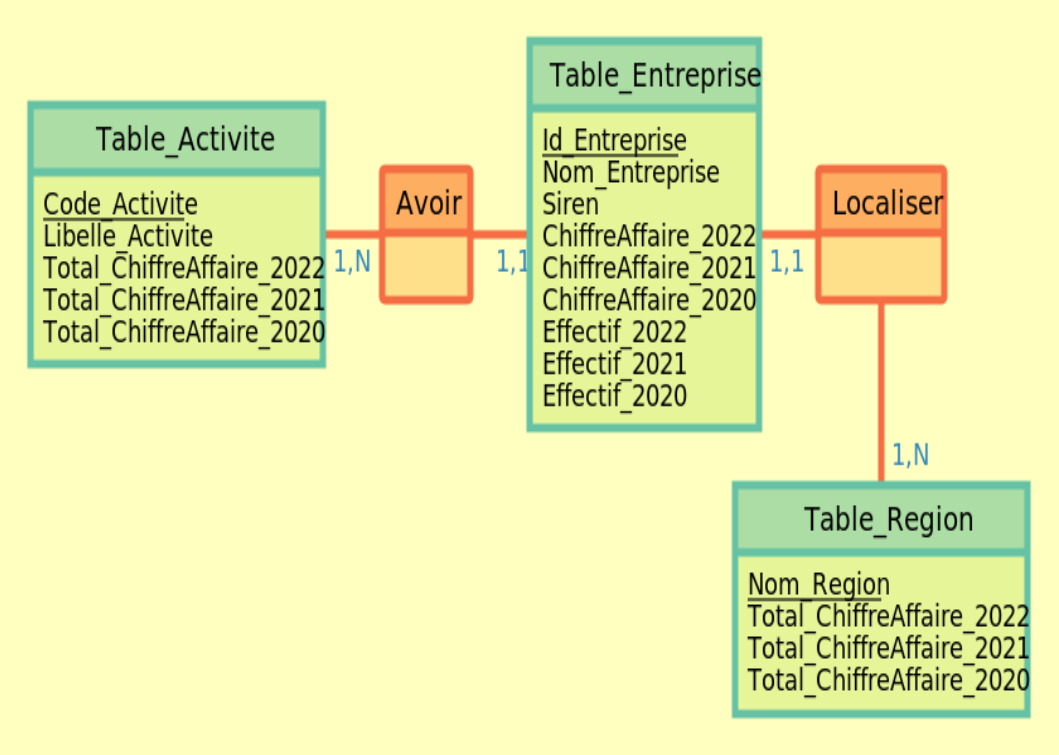
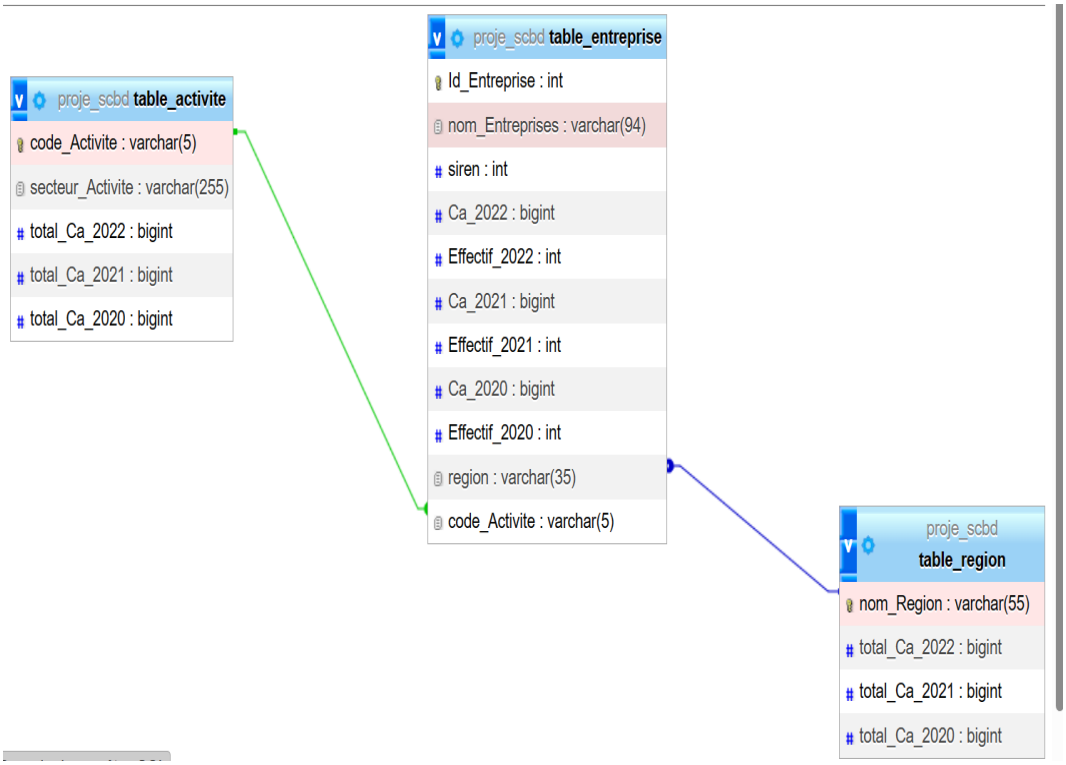


Figure 2.1: MCD.

- MOD:



Legende: Dans l'explication de notre MOD le signe # représente les **clés étrangéres** et les mots soulignés sont les **clés primaires**.

▼ **TABLE_ACTIVITE** (Code_Activite, Libelle_Activite, Total_ChiffreAffaire_2022, Total_ChiffreAffaire_2021, Total_ChiffreAffaire_2020)

- Le champ *Code_Activite* constitue la clé primaire de la table. C'était déjà un identifiant de l'entité *Table_Activite*.
- Les champs *Libelle_Activite*, *Total_ChiffreAffaire_2022*, *Total_ChiffreAffaire_2021* et *Total_ChiffreAffaire_2020* étaient déjà de simples attributs de l'entité *Table_Activite*.

▼ **TABLE_ENTREPRISE** (Id_Entreprise, Nom_Entreprise, Siren, ChiffreAffaire_2022, ChiffreAffaire_2021, ChiffreAffaire_2020, Effectif_2022, Effectif_2021, Effectif_2020, *#Code_Activite*, *#Nom_Region*)

- Le champ *Id_Entreprise* constitue la clé primaire de la table. C'était déjà un identifiant de l'entité *Table_Entreprise*.
- Les champs *Nom_Entreprise*, *Siren*, *ChiffreAffaire_2022*, *ChiffreAffaire_2021*, *ChiffreAffaire_2020*, *Effectif_2022*, *Effectif_2021* et *Effectif_2020* étaient déjà de simples attributs de l'entité *Table_Entreprise*.
- Le champ *Code_Activite* est une clé étrangère. Il a migré par l'association de dépendance fonctionnelle *Avoir* à partir de l'entité *Table_Activite* en perdant son caractère identifiant.
- Le champ *Nom_Region* est une clé étrangère. Il a migré par l'association de dépendance fonctionnelle *Localiser* à partir de l'entité *Table_Region* en perdant son caractère identifiant.

▼ **TABLE_REGION** (Nom_Region, Total_ChiffreAffaire_2022, Total_ChiffreAffaire_2021, Total_ChiffreAffaire_2020)

- Le champ *Nom_Region* constitue la clé primaire de la table. C'était déjà un identifiant de l'entité *Table_Region*.
- Les champs *Total_ChiffreAffaire_2022*, *Total_ChiffreAffaire_2021* et *Total_ChiffreAffaire_2020* étaient déjà de simples attributs de l'entité *Table_Region*.

Figure 2.2: Explication du MOD.

2.3 Import des données

- Le nettoyage des données a été réalisé sur Excel et Python(avec Pandas). Dans le fichier Chiffre clé 2022. Nous avons supprimé toutes les lignes comportants des valeurs manquantes. Pareillement, on a rectifié la forme d'écriture en UTF-8 pour rendre lisible nos données textuelles.
- Nous avons aussi filtré le fichier chiffre clé afin d'avoir nos trois tables. Ainsi, on a pu tirer la table Region en regroupant les regions de toutes les entreprises en faisant la somme de leur chiffre d'affaires pour chaque année avec l'aide de la bibliothéque pandas, puis on a effectué la même tâche pour avoir la table Activité.

2.4 Quelques détails techniques

Nous utilisons ce script pour nous connecter à notre base de données :

```
library(DBI)
library(RMySQL)
conn <- DBI::dbConnect(RMySQL::MySQL(),
                  user = "BDD_arrivedirt",
                  password = "df50a60dc58cfb65eefe321dcd88a5168dfd5083",
                  dbname = "BDD_arrivedirt",
                  host = "60d.h.filess.io",
                 port=3307)
dbListTables(conn)
```

```
## [1] "table_activite"   "table_entreprise" "table_region"
```

Là, on fait un test SQL pour voir, si la connexion avec notre base données est bien établie:

```
show tables;
```

Table 2.1: 3 records

Tables_in_BDD_arrivedirt
table_activite
table_entreprise
table_region

2.5 Requêtes réalisées

1)Quels sont les secteurs d'activités ayant enregistré la plus forte croissance de chiffre d'affaires au cours des trois dernières années(2020,2021,2022)?

```
SELECT secteur_activite,
((total_ca_2022 - total_ca_2020)/total_ca_2020)*100 AS croissance
FROM table_activite
ORDER BY croissance DESC
LIMIT 10;"
```

secteur_activite	croissance 1
Agencement de lieux de vente	6704.5386
Fabrication de matelas	1248.9513
Location et location-bail d'autres biens personnel...	231.5481
Gestion de fonds	200.1801
Promotion immobilière de logements	190.6370
Fabrication d'autres meubles et industries connexe...	188.9064
Commerce de gros (commerce interentreprises) de te...	186.6665
Comm. de gros (comm. interent.) de meubles, de tap...	160.7028
Supports juridiques de programmes	148.8556
Courtage de valeurs mobilières et de marchandises	134.4692

Figure 2.3: Question 1.

- Description: Dans cette requête, on sélectionne la colonne secteur d'activite et on calcule la croissance du chiffre d'affaires entre 2020 et 2022 en pourcentage en utilisant la formule suivante : ((total_ca_2022 - total_ca_2020) / total_ca_2020) * 100. Cela nous a permis de calculer le taux de croissance sur la période de 2020 à 2022. Les résultats sont triés par ordre décroissant de croissance et sont limités aux 10 premiers résultats pour illustrer secteurs d'activité ayant connu la croissance la plus importante.

2)Dans quelles régions se trouvent les secteurs d'activités ayant enregistré la plus forte croissance de chiffre d'affaires entre 2020 et 2022?

```sql
SELECT a.secteur_Activite, r.nom_region,
       (e.ca_2022 - e.ca_2020) AS croissance
FROM Table_Entreprise e
JOIN Table_Region r ON e.region = r.nom_region
JOIN table_activite a on e.code_Activite=a.code_Activite
ORDER BY croissance DESC
LIMIT 10;
```

secteur_Activite	nom_region	croissance_absolu 1
Commerce de voitures et de véhicules automobiles l...	Ile-de-France	2797434094
Commerce de gros d'équipements automobiles	Auvergne-Rhône-Alpes	785127506
Comm. de gros (comm. interent.) d'ordi., d'équi. i...	Languedoc-Roussillon-Midi-Pyrénées	733456394
Centrales d'achat alimentaires	Bretagne	439842175
Fabrication d'articles de joaillerie et bijouterie	Ile-de-France	271997592
Traitement de données, hébergement et activités co...	Ile-de-France	257601806
Production de boissons alcooliques distillées	Aquitaine-Limousin-Poitou-Charentes	242490547
Commerce de combustibles gazeux par conduites	Ile-de-France	231895611
Commerce de gros (commerce interentreprises) de bo...	Ile-de-France	183402449
Commerce de détail de produits surgelés	Ile-de-France	183041393

Figure 2.4: Question 2.

- Description: Cette requête sélectionne les 10 entreprises dont la croissance du chiffre d'affaires entre 2020 et 2022 a été la plus importante. Elle relie les tables **Table_Entreprise**, **Table_Region** et **Table_Activite** pour obtenir le secteur d'activité de chaque entreprise ainsi que le nom de sa région. Le résultat de la requête affiche trois colonnes :

-"secteur_Activite" pour le secteur d'activité de l'entreprise -"nom_region" pour le nom de la région où se situe l'entreprise -"croissance" pour la différence entre le chiffre d'affaires de 2022 et celui de 2020 (positif si la croissance est positive, négatif sinon)

3)Quels sont les secteurs d'activités qui ont enregistré les plus gros chiffres d'affaires entre 2020 et 2022?

```
SELECT secteur_activite, SUM(total_ca_2020) + SUM(total_ca_2021) +
FROM table_activite
GROUP BY secteur_activite
ORDER BY chiffre_affaires_total DESC
LIMIT 10;
```

secteur_activite	chiffre_affaires_total ▼ 1
Hypermarchés	23094581075
Comm. de gros (comm. interent.) d'ordi., d'équi. i...	16112007262
Commerce de voitures et de véhicules automobiles l...	13498677744
Transformation et conservation de la viande de vol...	7029289253
Supermarchés	6537233019
Commerce de détail de carburants en magasin spécia...	5720266855
Commerce de gros (commerce interentreprises) de pr...	5525270607
Commerce de gros d'équipements automobiles	5115696924
Commerce de détail de produits surgelés	4967466344
Production de boissons alcooliques distillées	4695927275

- Description: Cette requête SQL permet d'obtenir le chiffre d'affaires total pour les 10 secteurs d'activité les plus rentables en additionnant les chiffres d'affaires des années 2020, 2021 et 2022 pour chaque secteur d'activité.

4)Dans quelles régions se trouvent les secteurs d'activités ayant enregistré les plus gros chiffres d'affaires au cours des trois dernières années(2020, 2021 et 2022)?

```
SELECT nom_region, SUM(total_ca_2020 + total_ca_2021
+ total_ca_2022) AS chiffre_affaires_total
FROM table_region
```

```
GROUP BY nom_region
ORDER BY chiffre_affaires_total DESC limit 10
```

nom_region	chiffre_affaires_total ▾ 1
Ile-de-France	83839137861
Auvergne-Rhône-Alpes	29637867111
Languedoc-Roussillon-Midi-Pyrénées	17949033635
Pays-de-la-Loire	15616181147
Aquitaine-Limousin-Poitou-Charentes	14169070334
Nord-Pas-de-Calais-Picardie	13937525165
Bourgogne-Franche-Comté	8478873979
Provence-Alpes-Côte d'Azur	8308523140
Alsace-Champagne-Ardenne-Lorraine	7571729875
Bretagne	7521388532
Normandie	5232736055
Centre-Val de Loire	5209336028

Description: Cette requête SQL permet de calculer le chiffre d'affaires total par région en additionnant les chiffres d'affaires des années 2020, 2021 et 2022 de toutes les entreprises de chaque région, puis en triant les résultats par ordre décroissant de chiffre d'affaires total.

5)Quelles sont les régions où se trouvent les entreprises ayant le plus gros chiffre d'affaires sur les trois dernières années(2020, 2021, 2022)?

```
SELECT region,
       MAX(total_ca_2022) AS ca_2022,
       MAX(total_ca_2021) AS ca_2021,
       MAX(total_ca_2020) AS ca_2020
FROM Table_Region,table_entreprise
where Table_Region.nom_region = Table_Entreprise.region
GROUP BY region
ORDER BY ca_2022 DESC, ca_2021 DESC, ca_2020 DESC
LIMIT 10;
```

region	ca_2022 1	ca_2021 2	ca_2020 3
Ile-de-France	32030960828	24888726192	26919450841
Auvergne-Rhône-Alpes	10949204267	9709795617	8978867227
Languedoc-Roussillon-Midi-Pyrénées	6635244454	5938282745	5375506436
Pays-de-la-Loire	5611136015	5046699992	4958345140
Aquitaine-Limousin-Poitou-Charentes	5250251762	4493803368	4425015204
Nord-Pas-de-Calais-Picardie	5086752271	4477083866	4373689028
Bourgogne-Franche-Comté	3024731487	2735781954	2718360538
Bretagne	2994007971	2194004587	2333375974
Provence-Alpes-Côte d'Azur	2915407551	2635227711	2757887878
Alsace-Champagne-Ardenne-Lorraine	2761721799	2373682523	2436325553

Figure 2.5: Question 5.

Cette requête récupère les 10 régions ayant le plus grand chiffre d'affaires en 2022, 2021 et 2020, en prenant les données des tables Table_Region et Table_Entreprise. La clause **where** permet de faire la jointure entre les deux tables sur le champ nom_region de Table_Region et region de Table_Entreprise. Les colonnes renvoyées sont region pour le nom de la région, ca_2022 pour le chiffre d'affaires total de l'année 2022, ca_2021 pour le chiffre d'affaires total de l'année 2021, et ca_2020 pour le chiffre d'affaires total de l'année 2020. Le chiffre d'affaires total est obtenu en calculant la somme des chiffres d'affaires des entreprises de chaque région.

Dans cette requête, on utilise une jointure entre les tables Table_Region et Table_Entreprise pour regrouper les entreprises par région,puis on a calculé le chiffre d'affaire total pour chaque année. Ensuite, on a effectué la sélection des régions avec le chiffre d'affaires le plus élevé pour chaque année en utilisant la fonction d'agrégation **MAX**.Enfin,on ordonne les résultats par chiffre d'affaires décroissant en affichant que les résultats des 10 premières régions.

Chapitre 3

Matériel et Méthodes

3.1 Logiciels

Notre outil principal pour se communiquer entre les membres du groupe est le discord(utilisation de snapchat au debut) ou BBB sur mooodle. On a partagé toutes nos données et toutes les informations importante concernant le projet. Ensuite, On a utilisé Excel pour l'uniformisation des données et le filtrage des colonnes. La bibliothèque Pandas de python a été d'une grande aide pour le néttoyage complet des données bien vraie que le logiciel nous a aussi bien aidé au debut sur le nettoyage de données. En plus R,a été utilisé pour mettre en relation tous les logiciels notamment notre base de données, puis produire ce rapport final(via RMarkdown). Enfin, on a utilisé phpMyAdmin pour réalisé des requêtes sur notre base de données. Voici les informations sur les versions des logiciels et sur l'ordinateur qui a servi pour les analyses.

- Ordinateur:
  - Nom de l'appareil: **LAPTOP-KTKJ2BP4**
  - Processeur: **Intel(R) Core(TM) i7-1065G7 CPU @ 1.30GHz 1.50 GHz**
  - Memoire Ram installé: **8,00Go (7,74Go utilisable)**
  - Type du sytéme: **Système d'exploitation 64bits, processeur x64**
  - Marque: ASUS
- Python:
  - version: 3.9.12
  - google Collab -jupyter(avec anaconda)
- R:
  - language: R

– version.string R version: 4.3.0 (2023-04-21 ucrt) Listons tous les logiciels utiliser pour la partie Base de Données, statistique mais également pour gérer et communiquer entre les membres du projet. On a privilégié les logiciels R (ou Python) pour la Science des Données. Pour assurer une reproductibilité maximale, on a utiliser R Markdown,via RStudio.

Chapitre 4

Lister toutes les connexions actives

4.1 Description des Données

- Comment les données sont-elles stockées? Les données ont été tiré dans un fichier CSV(nommé chiffre clé),puis on les a néttoyé et stocké dans une base de données sql dans phpMyAdmin. Ce qui nous a permis d'effectuer des requêtes SQL et par la suite recupérer les données pour pouvoir faire des analyses statistiques.

- Quelles sont les tailles des fichiers en jeu? Combien y a t-il de fichiers?

Notre fichier avant l'utilisation etait de 67 643 ko contenant 167 885 lignes et 42 colonnes.On a utiliser un seul fichier(chiffre clé 2022) pour l'extraction des données car on a juger que les autres n'étaient pas necessaire, vu notre problématqiue posées.

- Combien d'unités statistiques? Combien de variables? etc. Notre base de données contient trois tables.Voici les informations pour chacune des tables : Table_Activite: 443 unités statistiques et 5 variables, Table_Entreprise: 3,264 unités statistiques et 11 variables, Table_Region: 17 unités statistiques et 4 variables En total, notre base de données contient **3 724 unités statistiques** et **20 variables**.

4.2 Nettoyage des données

Comment gérez-vous les données manquantes, etc. ?

Pour gérer les données manquantes, on a utilisé des fonctions SQL pour filtrer les enregistrements qui contiennent des valeurs manquantes ou pour remplacer les valeurs manquantes par des valeurs par défaut. On a également utilisé des fonctions de traitement de chaînes de caractères avec le logiciel R(données paramétrées avec l'encodage utf-8 pour les rendre plus visible) ou python(Pandas, Numpy,..) pour nettoyer les données en supprimant les espaces inutiles, les caractères spéciaux, etc.

4.3 Étapes de Pré-traitements

-Quelles transformations avez-vous effectuées sur vos données pour les rendre utilisables?

- Pour rendre les données utilisables, on a effectué des transformations telles que la normalisation des données, la discrétisation des variables continues, la transformation des variables catégorielles en variables binaires, etc. Ces transformations ont été effectué à l'aide de fonctions SQL ou de bibliothèques de traitement de données en R ou Python.
- On a également explorer les données à l'aide de graphiques et de résumés statistiques pour identifier des schémas et des tendances, et pour valider la qualité des données avant d'effectuer des analyses

4.4 Modélisation statistique

Quels outils ou méthodes de statistiques allez-vous utiliser? Donner des équations mathématiques sMil 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

Chapitre 5

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. De ce fait,cette partie a pour but de mieux comprendre l'information contenue dans nos données, pour cela nous avons généré différents graphiques et plusieurs valeurs numériques. Ainsi, pour faire des analyses pertinentes nous avons décider de se poser certains bon nombre de questions, afin de répondre à notre problématique. Ces questions vont répondre sur l'analyse de la performance économique des différentes régions et secteurs d'activités au cours des trois dernières années(2020, 2021, 2022). Elles permettront d'identifier les tendances et les modèles dans les données, de comparer les performances des différentes régions et secteurs, et de prendre des décisions éclairées en matière d'investissement et de gestion financière. Voici nos questions:

-Quels sont les secteurs d'activités ayant enregistré la plus forte croissance de chiffre d'affaires au cours des trois dernières années(2020, 2021, 2022)?

-Dans quelles région se trouvent les secteurs d'activités ayant enregistré la plus forte croissance de chiffre d'affaires au cours des trois dernières années(2020, 2021, 2022)?

-Quels sont les secteurs d'activités qui ont enregistré les plus gros chiffres d'affaires au cours des trois dernières années(2020, 2021, 2022)?

-Dans quelles régions se trouvent les secteurs d'activités ayant enregistré les plus gros chiffres d'affaires au cours des trois dernières années(2020, 2021, 2022)?

5.1 **Resumé des données**

On utilise la fonction summary pour obtenir un résumé statistique des données sur nos entreprises. Cela nous permettra d'avoir rapidement un aperçu sur les données tel que les chiffres d'affaires des entreprises

en 2022 qui nous intéresse ici. **Résumé des chiffres d'affaires des entreprises en 2022:**

```
summary(var2$Ca_2022)
```

Voici un résumé des informations obtenues à partir de l'analyse des chiffres d'affaires des entreprises en 2022 :

- La **moyenne** de chiffre d'affaires en 2022 est de **24 971 661 euros**. Cela signifie que si l'on additionne tous les chiffres d'affaires de toutes les entreprises et que l'on divise cette somme par le nombre total d'entreprises, on obtient une moyenne de 24 971 661. Cependant, il est important de noter que la moyenne peut être biaisée par des valeurs extrêmes (appelées "outliers"), ce qui peut fausser la représentativité de cette mesure. C'est pourquoi il ce serait mieux d'utiliser d'autres mesures de tendance centrale en combinaison avec la moyenne, comme la médiane ou le mode.
- Le chiffre d'affaires **minimum** est **de 0 euro**, ce qui pourrait être dû à des entreprises qui n'ont pas généré de revenus au cours de cette période.
- Le **premier quartile** est de **691 914 euros**, ce qui signifie que **25%** des entreprises ont un chiffre d'affaires inférieur à ce montant.
- La **médiane**, qui est la valeur au milieu de la distribution, est de **3 574 482 euros**. Cela signifie que 50% des entreprises ont un chiffre d'affaires inférieur à ce montant et 50% ont un chiffre d'affaires supérieur.
- Le **troisième quartile** est de **15 836 456 euros**, ce qui signifie que 75% des entreprises ont un chiffre d'affaires inférieur à ce montant.
- La valeur **maximale** de chiffre d'affaires est de **2 909 072 591 euros**, ce qui représente l'entreprise ayant généré le plus de revenus en **2022**. **Ecart-type(sd) et Variance(var)**

```
#Variance
var(var2$Ca_2022)
#ecartype
sd(var2$Ca_2022)
```

La variance (var) est une mesure de dispersion qui calcule la moyenne des carrés des écarts à la moyenne d'un échantillon de données. Elle est définie comme la différence entre la moyenne des carrés et le carré de la moyenne de l'échantillon. Plus la variance est grande, plus les données sont dispersées autour de la moyenne.Dans nos données la valeur du chiffre d'affaire 2022 est très élevée (environ 1.4e+16), ce qui indique que les données ont une grande dispersion autour de la moyenne. La déviation standard (sd) ou ecart-type, quant à elle, est une mesure de la dispersion qui calcule la racine carrée de la variance. Elle est utilisée pour mesurer la variation des données par rapport à leur moyenne. La valeur du chiffre d'affaire en 2022 est d'environ 118 549 414, ce qui indique que les données ont une variation relativement élevée par rapport à leur moyenne.

5.2 Réponses aux Questions à l'aide de R

1)Quels sont les secteurs d'activités qui ont enregistré la plus forte croissance entre 2020 et 2022 ?

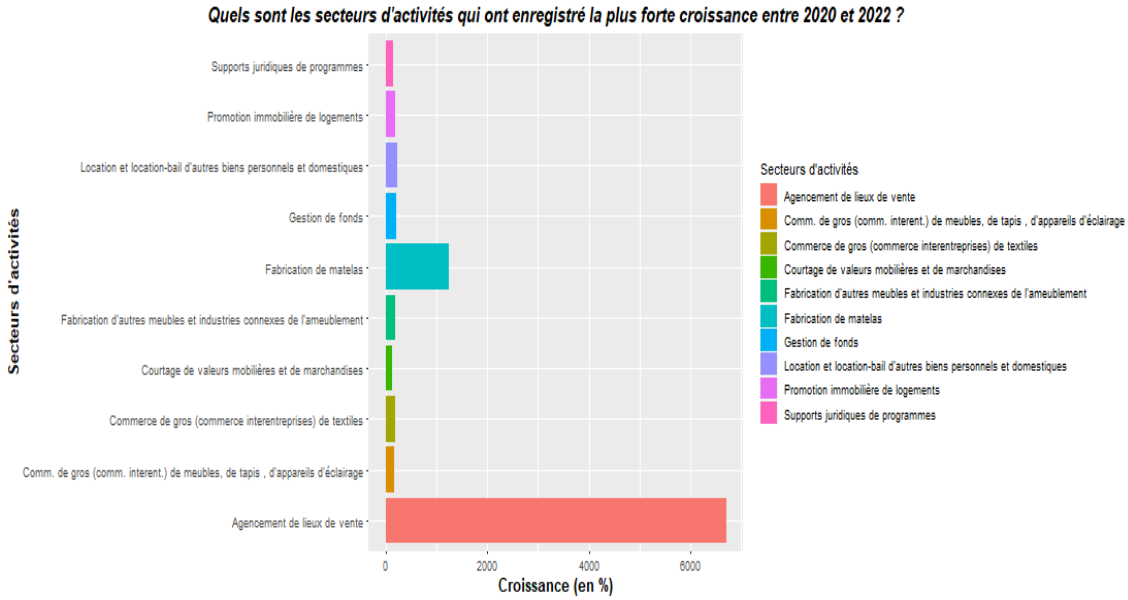


Figure 5.1: Diagramme en barre.

Le graphique affiche les 10 secteurs d'activités avec la plus forte croissance de revenus entre 2020 et 2022 en France. Comme nous pouvons le voir sur le graphique, parmi les 10 secteurs d'activités avec le taux de

croissance le plus élevé, 2 secteurs prédominent. Le secteur de **fabrication de matelas** a enregistré une croissance d'environ **1249%** entre 2020 et 2022 mais c'est surtout le secteur **d'agencement de lieux de vente** qui a connu une forte augmentation du taux de croissance d'environ **6705%** sur la même période. Cela est probablement du à la crise du Covid-19 qui a eu lieu en 2020.

2)Dans quelles régions se trouvent les secteurs d'activités ayant enregistré la plus forte croissance de chiffre d'affaires entre 2020 et 2022?

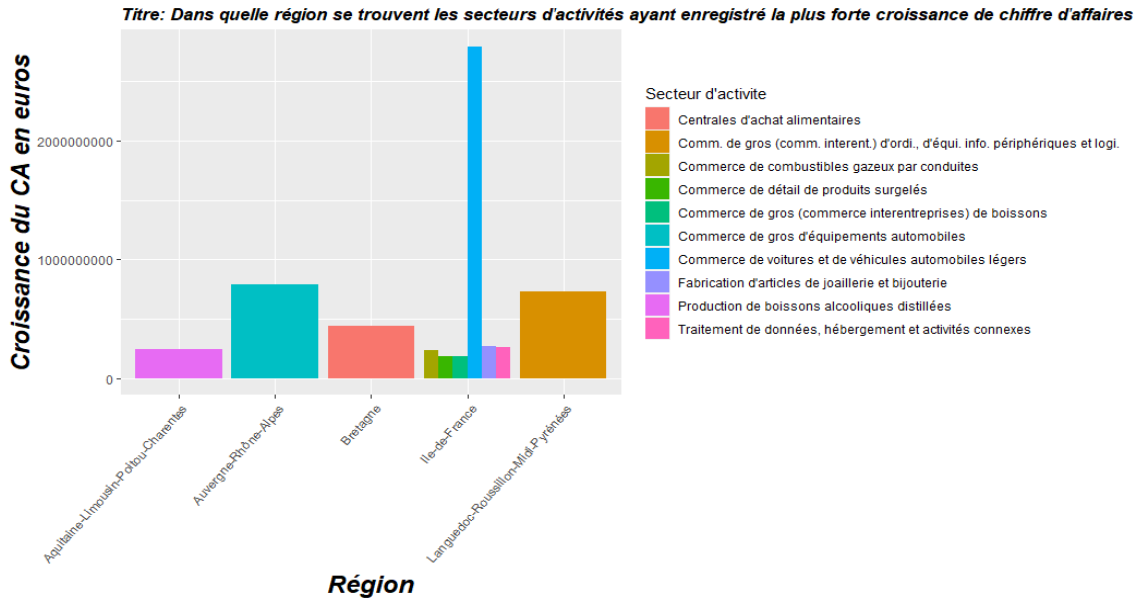
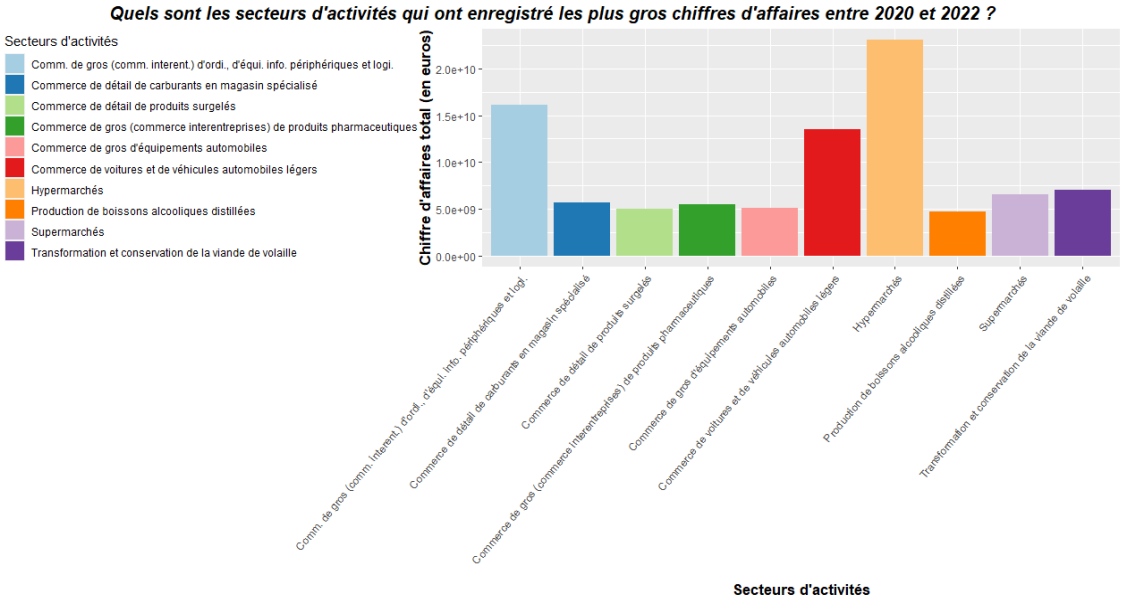


Figure 5.2: Diagramme en barre des régions où on trouvent les secteurs d'activités avec la plus forte croissance entre 2020 et 2022.

Ce graphe permet de répondre à la question "Dans quelle région se trouvent les secteurs d'activités ayant enregistré la plus forte croissance de chiffre d'affaires au cours des trois dernières années(2020 à 2022)?". Il affiche la répartition par région sous forme de diagramme en barre avec la croissance du CA en tant que poids. Cela permet de visualiser les régions où les secteurs d'activités connaissant la croissance la plus rapide en termes de CA et de comprendre les tendances et les opportunités de croissance potentielles pour les entreprises dans ces régions et secteurs.

3)Quels sont les secteurs d'activités qui ont enregistré les plus gros chiffres d'affaires au cours des trois dernières années?



Le graphique affiche les 10 secteurs d'activités avec les plus gros chiffres d'affaires entre 2020 et 2022 en France. Nous remarquons sur le graphique, 3 secteurs d'activités qui ont accumulé plus de 10 milliards d'euros de chiffres d'affaires entre 2020 et 2022. Le **commerce de voitures et de véhicules automobile** a engendré plus de **13 milliards d'euros** durant cette période. Quant au commerce de gros d'ordinateurs, d'équipements informatiques, il a amené plus de **16 milliards d'euros** de chiffre d'affaires. Les **hypermarchés** sont les premiers secteurs d'activités, ils ont rapporté plus de **23 milliards d'euros** de chiffre d'affaires en France ces 3 dernières années.

4)Dans quelles régions se trouvent les secteurs d'activités ayant enregistré les plus gros chiffres d'affaires au cours des trois dernières années?

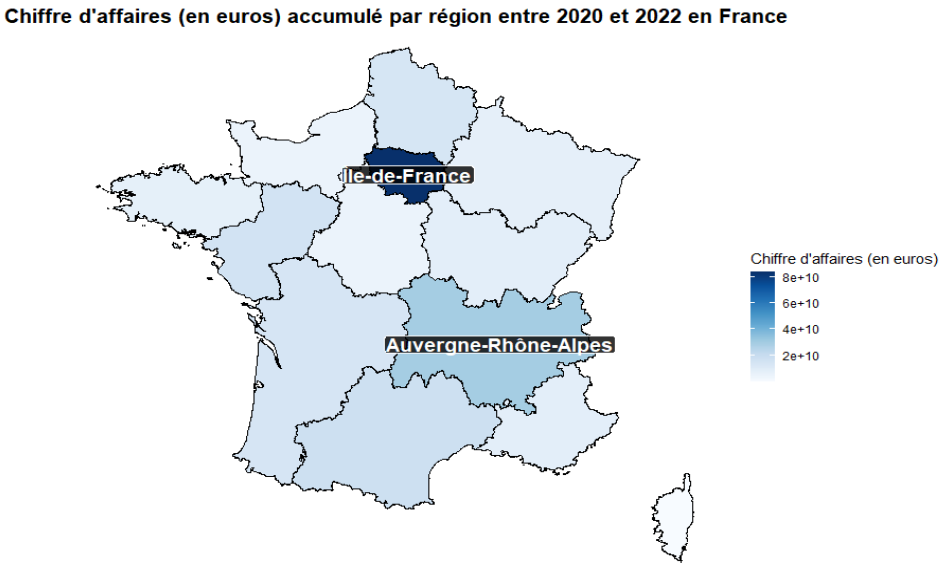


Figure 5.3: Carte.

Le graphique montre les différentes régions de France selon le chiffre d'affaires qu'elles ont obtenu. La deuxième région qui a accumulé le plus de chiffres d'affaires est la région **Auvergne-Rhône-Alpes** qui a généré plus de **29 milliards d'euros** entre 2020 et 2022. C'est beaucoup moins que la région **Île-de-France (bleu foncé sur le graphique)** qui a engendré environ **84 milliards d'euros** de chiffres d'affaires soit presque 3 fois plus que la région **Auvergne-Rhône-Alpes**. **NB**: Nous avons laissé seulement la France métropolitaine pour avoir un graphique plus clair.

Chapitre 6

Conclusion et perspectives

Dans l'ensemble, nous pensons pouvoir proposer deux types d'investissements dans les secteurs d'activité. La première consiste à investir dans des industries à forte influence en fonction de l'évolution de leur chiffre d'affaires sur les trois dernières années. Notre analyse nous amène à deux secteurs qui répondent le mieux à ce critère : l'aménagement de points de vente et la fabrication de matelas. La deuxième stratégie d'investissement consiste à investir dans les régions qui génèrent le plus de chiffre d'affaires. ces zones sont les plus stables sur le long terme car même s'ils n'ont pas connu d'évolutions majeures au cours des 3 années étudiées, leurs chiffre d'affaires reste élevé. Les deux régions les plus importantes sont l'Ile-de-France et l'Auvergne-Rhône-Alpes. Par perspective, en investissant dans une variété d'industries et de secteurs, ainsi que dans différentes régions,l'exposition aux risques spécifiques à une seule entreprise ou à un seul secteur, tout en profitant des opportunités de croissance et de rentabilité dans d'autres domaines. Cette stratégie peut aider aux entreprise à minimiser les risques tout en maximisant les rendements potentiels.

Annexes

Codes

connexion à notre base de données

```
library(DBI)
library(RMySQL)
conn <- DBI::dbConnect(RMySQL::MySQL(),
                       user = "BDD_arrivedirt",
                       password = "df50a60dc58cfb65eefe321dcd88a5168dfd
                       dbname = "BDD_arrivedirt",
                       host = "60d.h.filess.io",
                      port=3307)
dbListTables(conn)
```

- **Partie base de données**

1)Quels sont les secteurs d'activités ayant enregistré la plus forte croissance de chiffre d'affaires au cours des trois dernières années(2020,2021,2022)?

```
SELECT secteur_activite,
((total_ca_2022 - total_ca_2020)/total_ca_2020)*100 AS croissance
FROM table_activite
ORDER BY croissance DESC
LIMIT 10;"
```

2)Dans quelles régions se trouvent les secteurs d'activités ayant enregistré la plus forte croissance de chiffre d'affaires entre 2020 et 2022?

```
SELECT a.secteur_Activite, r.nom_region,
       (e.ca_2022 - e.ca_2020) AS croissance
FROM Table_Entreprise e
JOIN Table_Region r ON e.region = r.nom_region
```

```
JOIN table_activite a on e.code_Activite=a.code_Activite
ORDER BY croissance DESC
LIMIT 10;
```

3)Quels sont les secteurs d'activités qui ont enregistré les plus gros chiffres d'affaires entre 2020 et 2022?

```
SELECT secteur_activite, SUM(total_ca_2020) + SUM(total_ca_2021) +
FROM table_activite
GROUP BY secteur_activite
ORDER BY chiffre_affaires_total DESC
LIMIT 10;
```

4)Dans quelles régions se trouvent les secteurs d'activités ayant enregistré les plus gros chiffres d'affaires au cours des trois dernières années(2020, 2021 et 2022)?

```
SELECT nom_region, SUM(total_ca_2020 + total_ca_2021
+ total_ca_2022) AS chiffre_affaires_total
FROM table_region
GROUP BY nom_region
ORDER BY chiffre_affaires_total DESC limit 10
```

- **Partie Science de données**

1)Quels sont les secteurs d'activités qui ont enregistré la plus forte croissance entre 2020 et 2022 ?

```
requete1 <- "SELECT secteur_activite,
((total_ca_2022 - total_ca_2020)/total_ca_2020)*100 AS croissance
FROM table_activite
ORDER BY croissance DESC
LIMIT 10;"
requete1
View(requete1)
data1 <- dbGetQuery(bd, requete1)

ggplot(data1) +
  aes(
```

```
  x = secteur_activite,
  fill = secteur_activite,
  weight = croissance
) +
geom_bar(position = "dodge") +
scale_fill_hue(direction = 1) +
labs(
  x = "Secteurs d'activités",
  y = "Croissance (en %)",
  title = "Quels sont les secteurs d'activités qui ont enregistr
  fill = "Secteurs d'activités"
) +
coord_flip() +
theme_gray() +
theme(
  plot.title = element_text(size = 15L,
                            face = "bold.italic",
                            hjust = 0.5),
  axis.title.y = element_text(size = 13L,
                              face = "bold"),
  axis.title.x = element_text(size = 13L,
                              face = "bold")
)
```

2)Dans quelles régions se trouvent les secteurs d'activités ayant enregistré la plus forte croissance de chiffre d'affaires entre 2020 et 2022?

```
 diagramme = ggplot(data3) +
aes(
  x = nom_region,
  fill = secteur_Activite,
  weight = croissance
) +
geom_bar(position = "dodge") +
scale_fill_hue(direction = 1) +
labs(
  x = "Région",
```

```
    y = "Croissance du CA en millons d'euros",
    title = "Titre: Dans quelle région se trouvent les secteurs d'
    ayant enregistré la plus forte croissance de chiffre d'affaire
    2020 et 2022?",
    fill = "Secteur d'activite"
  ) +
  theme_gray() +
  theme(
    plot.title = element_text(face = "bold.italic"),
    plot.subtitle = element_text(size = 14L,
    face = "italic",
    hjust = 0.5),
    axis.title.y = element_text(size = 18L,
    face = "bold.italic"),
    axis.title.x = element_text(size = 18L,
    face = "bold.italic")
  )
```

3)Quels sont les secteurs d'activités qui ont enregistré les plus gros chiffres d'affaires au cours des trois dernières années?

```
requete2 <- SELECT secteur_activite, SUM(total_ca_2020) + SUM(tota
FROM table_activite
GROUP BY secteur_activite
ORDER BY chiffre_affaires_total DESC
LIMIT 10;
requete2
View(requete2)
data2 <- dbGetQuery(bd, requete2)
colors <- brewer.pal(10, "Paired")
ggplot(data2) +
  aes(
    x = secteur_activite,
    fill = secteur_activite,
    weight = chiffre_affaires_total
  ) +
  geom_bar() +
  scale_fill_manual(
```

```
    values = colors
  ) +
  labs(
    x = "Secteurs d'activités",
    y = "Chiffre d'affaires total (en euros)",
    title = "Quels sont les secteurs d'activités qui ont enregistr
    les plus gros chiffres d'affaires entre 2020 et 2022 ?",
    fill = "Secteurs d'activités"
  ) +
  theme_gray() +
  theme(
    plot.title = element_text(size = 15L,
                              face = "bold.italic",
                              hjust = 1.5),
    axis.title.y = element_text(size = 13L,
                                face = "bold"),
    axis.title.x = element_text(size = 13L,
                                face = "bold"),
    axis.text.x = element_text(angle = 50, hjust = 1),
    legend.position = "left", legend.justification = "center"
  )
```

4)Dans quelles régions se trouvent les secteurs d'activités ayant enregistré les plus gros chiffres d'affaires au cours des trois dernières années?

```
#install.packages("esquisse")
#install.packages("RColorBrewer")
#install.packages("maps")
#install.packages("sf")
#install.packages("dplyr")
#install.packages("rmapshaper")
library(esquisse)
library(RColorBrewer)
library(RMySQL)
library(DBI)
library(ggplot2)
library(maps)
```

```r
library(sf)
library(dplyr)
library(rmapshaper)
regions <- read_sf("C:/Users/sarto/Desktop/L2 MIASHS 2022-2023/S4/

requete4 <- "SELECT nom_region, SUM(total_ca_2020 + total_ca_2021
             FROM table_region
             GROUP BY nom_region
             ORDER BY chiffre_affaires_total DESC"
data4 <- dbGetQuery(bd, requete4)

regions <- rename(regions, nom_region = nom)

names(regions)

regions1 <- ms_simplify(regions)
format(object.size(regions1),units="Mb")

unique(data4$nom_region)
unique(map_data$nom_region)

dist_matrix <- stringdist::stringdistmatrix(data3$nom_region, map_

for (i in seq_along(data4$nom_region)) {
  closest_match <- which.min(dist_matrix[i,])
  if (dist_matrix[i, closest_match] < 0.2) {
    data3$nom_region[i] <- map_data$nom_region[closest_match]
  }
}

map_data <- dplyr::left_join(regions1, data4, by = c("nom_region"

col <- brewer.pal(6, "Blues")

ggplot() +
  geom_sf(data = map_data, aes(fill = chiffre_affaires_total), col
  scale_fill_gradientn(name = "Chiffre d'affaires", colors = col,
```

```
  theme_void() +
  labs(title = "Chiffre d'affaires (en euros) accumulé par région
  coord_sf(xlim = c(-5.5,10),ylim=c(41,51))+
  theme(plot.title = element_text(hjust = 0.5))

decon <- dbDisconnect(bd)
decon
'''
```