PROJET RÉALISÉ PAR L'ÉQUIPE 2

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

Thomas GOUTIERES, Ivan ARISOY, Axel CAROT,

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

Avril 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: _____

Remerciements

Nos plus sincères remerciements vont à nos encadrants pédagogiques, Sandra Bringey et Pierre Lafaye De Micheaux pour les conseils avisés sur notre travail.

23/11/2022.

Résumé

Notre projet consiste en l'étude de bases de données ayant comme objectif de déterminer le meilleur secteur d'investissement. Nous avons tout d'abord importé nos données dans une base de données SQL. A l'aide de nos requêtes SQL nous avons pu faire nos analyses statistiques sur R. Nous analyserons ensuite les visualisations que nous avons créées pour essayer de répondre à notre problématique. Les visualisations mettront en image les différents paramètres que nous pensons être impactant pour choisir dans quel milieu investir. ¶

¶

Table des matières

Chapitre 1

Introduction

1.1 Contexte

L'investissement désigne une dépense immédiate ayant pour but d'obtenir un effet positif quantifiable à long terme.

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:

Quelles sont les secteurs les plus attractifs au cours des 3 dernières années ?

Pour cela, nous étudierons les chiffres d'affaire des entreprises, leur effectifs ou encore leur localisations.

Les données utilisées seront celles trouvées sur le site :
https://www.data.gouv.fr/fr/datasets/base-sirene-des-entreprises-et-de-leu
https://www.data.gouv.fr/fr/datasets/chiffres-cles-2022/

Chapitre 2

Base de données

2.1 Descriptif des tables

Nous utiliserons trois tables :

- Stock unité légale ;
- Chiffre clé ;
- Code Activité (https://insee.fr/fr/information/2406147)

Tables

Chiffre_cle (8767 lignes)

Colonnes	Types	Significations
Siren	Varchar	**Clé primaire**
num_dept	Varchar	Département de l'entreprise
Année	Varchar	**Clé primaire**
Effectif	Int	Effectif de l'entreprise
CA	Int	Chiffre d'affaire de l'entreprise

Stock_UL (4017 lignes)

Colonnes	Types	Significations
siren	Varchar	**Clé primaire**
categorieEntreprise	Varchar	PME, ETI, GE
etatAdministratifUniteLegale	Varchar	A, C
Activite_PrincipaleUniteLegale	Varchar	Code de l'activité principale

Code_activite_UL (1728)

Colonnes	Types	Significations
code	Varchar	**Clé primaire**
Intitule	Varchar	Nom de la catégorie d'activité principale

2.2 Modèles MCD et MOD

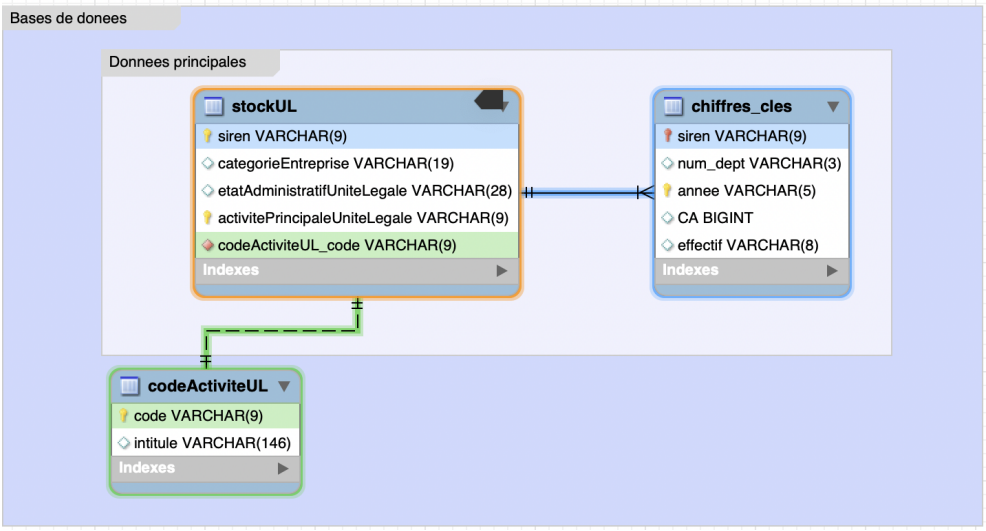


Figure 2.1: MOD.

stockUL(siren, categorieEntreprise, etatAdministratifUniteLegale, activitePrincipale UniteLegale) codeActiviteUL(code, intitule) chiffres_cles(siren, num_dept, annee, CA, effectif)

2.3 Import des données

- Le nettoyage des données a été réalisé sur Excel et Python. Dans la table Chiffre_cles : Nous avons supprimé toutes les lignes avec un CA nul. Nous avons supprimé toutes les lignes ou le numéro de département n'étant pas renseigné en utilisant Pandas. Nous avons avons fait des modifications pour passer de : SIREN-CA 2020-CA 2021-CA 2022 à SIREN-Annee-CA.
- Nous avons ensuite filtré nos deux tables principales : Chiffres_cles et StockUL de sorte à avoir exactement les mêmes SIREN dans les deux tables. Pour cela nous avons effectué deux fonctions RECHERCHEV() pour retenir uniquement les lignes avec des SIREN identiques dans les deux tables.

2.4 Requêtes réalisées

- Les requêtes utilisées seront présentes en annexe et expliquées lors de leurs utilisations.

2.5 Configuration de la base de données dans le Cloud.

- Pour cela on a utilisé le service d'exploitation Amazon relational database Service.
- Puis on a utilisé MySQL Workbench pour se connecter au serveur MySQL sur RDS dans le cloud et ensuite pour créer et gérer notre base de donees.
- Pour les règles de connéctivites du trafic entrant on a tout simplement choisi tout le trafic IPv4 (0.0.0.0/0)

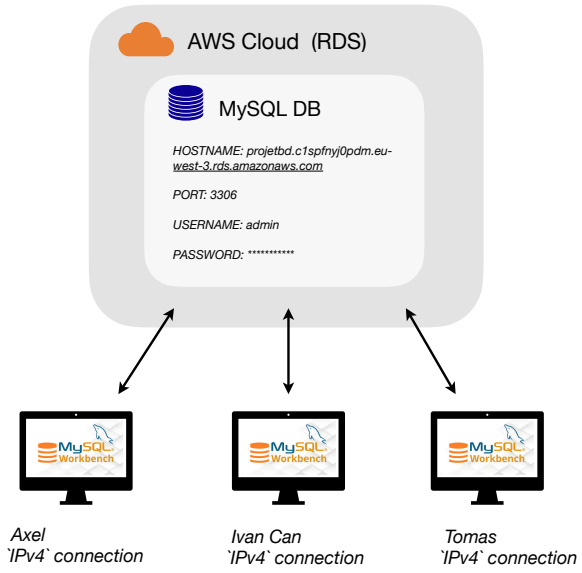


Figure 2.2: Connection.

2.6 Quelques détails techniques

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

```
library(DBI)
library('odbc')
con <- dbConnect(RMySQL::MySQL(),
                 dbname = "projetBD",
                 host = "projetbd.c1spfnyj0pdm.eu-west-3.rds.amazonaws.com",
                 port = 3306,
                 user = "admin",
                 password = "projet2022")


dbListTables(con)
```

Chapitre 3

Matériel et Méthodes

3.1 Logiciels

- Nous avons rédigé les fichiers R et RMarkdown(RMD) dans rStudio
- Nous avons utilisé Python et la library pandas pour nettoyer les données brutes en format csv dans Google collab
- Nous avons utilisé phpmyadmin pour créer la base de données. Puis on a exporté la BD en script SQL dans MySQL Workbench pour avoir notre BD dans le cloud.
- Amazon Relational Database Service (ou Amazon RDS) est un service de base de données relationnelle distribué par Amazon Web que nous avons utilisé.

3.2 Description des Données

Les données sont stockées sous forme de 3 documents Excel distincts sur nos machines personelles localement : Chiffres_cles.xls 238ko, StockUL.xls 94ko et Intitulé.xls 61ko Nous pouvons aussi accèder à ces données sur le Cloud AWS dans la base de données SQL.

3.3 Nettoyage des données

Nous avons fait en sorte de ne pas avoir de données manquantes pour les SIREN, les CA et les num_depts. Pour les autres variables si nous avons des données manquantes cela n'a pas d'impact sur nos analyses.

3.4 Modélisation de la base de données

Modèle conceptuel des données sous forme d'un schéma :

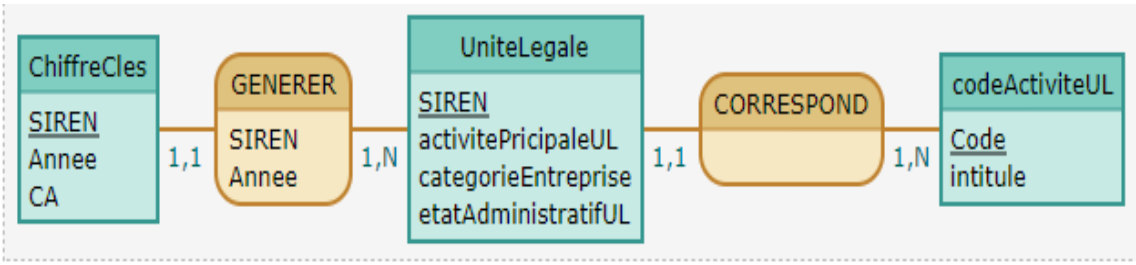


Figure 3.1: Relations.

Chapitre 4

Analyse Exploratoire des Données et Résultats

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.

4.1 Utiliser R

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-1142048    112005    460530  13092482   2412953 3537000000
```

Figure 4.1: Quartiles, écart-type variance.

La valeur minimale est de -1142048. Cela pourrait indiquer que ces entreprises ont déclaré des pertes ou des changements négatifs dans les ventes.
Le premier quartile est 112005, ce qui signifie que 25 % des entreprises ont des chiffres de vente inférieurs ou égaux à 112005. Cela pourrait indiquer que les chiffres de vente sont généralement distribués de manière biaisée ou non uniforme.
La médiane est 460530, ce qui signifie que 50% des entreprises ont des chiffres de ventes inférieurs ou égaux à 460530. Cela pourrait indiquer que les chiffres de ventes sont généralement distribués autour de cette valeur.
La moyenne est de 13092482, ce qui est beaucoup plus grand que la médiane et les premier et troisième quartiles. Cela pourrait indiquer que les chiffres des ventes ont des valeurs très élevées qui tirent la moyenne vers le haut.
Le troisième quartile est 2412953, ce qui signifie que 75 % des entreprises ont des chiffres de ventes inférieurs ou égaux à 2412953.
La valeur maximale est 3537000000, ce qui est beaucoup plus grand que les autres statistiques. Cela pourrait indiquer que l'ensemble de données contient quelques entreprises avec des chiffres de vente très importants qui sont très différents des autres entreprises.
Dans l'ensemble, ces statistiques suggèrent que les chiffres d'affaires des entreprises en France déclarant des chiffres d'affaires négatifs et quelques entreprises affichant des chiffres d'affaires très importants qui sont significativement différents du reste de l'ensemble de données.
Les codes R utilisés pour tracer ces graphiques seront disponible en annexe.

```
> sd(data$CA)
[1] 110314869
>
> var(data$CA)
[1] 1.216937e+16
```

Figure 4.2: Quartiles, écart-type variance.

Ces informations peuvent être utilisées pour identifier les tendances et les modèles dans les données, et pour comparer les chiffres de vente de différentes entreprises. Par exemple, vous pouvez utiliser l'écart type et la variance pour identifier les entreprises dont les ventes sont anormalement élevées ou faibles et pour évaluer les performances globales des entreprises dans l'ensemble de données. Cependant, il est important de noter que ces statistiques ne fournissent pas à elles seules une image complète des chiffres de vente et que d'autres facteurs doivent également être pris en compte lors de la formulation de conclusions sur les données.

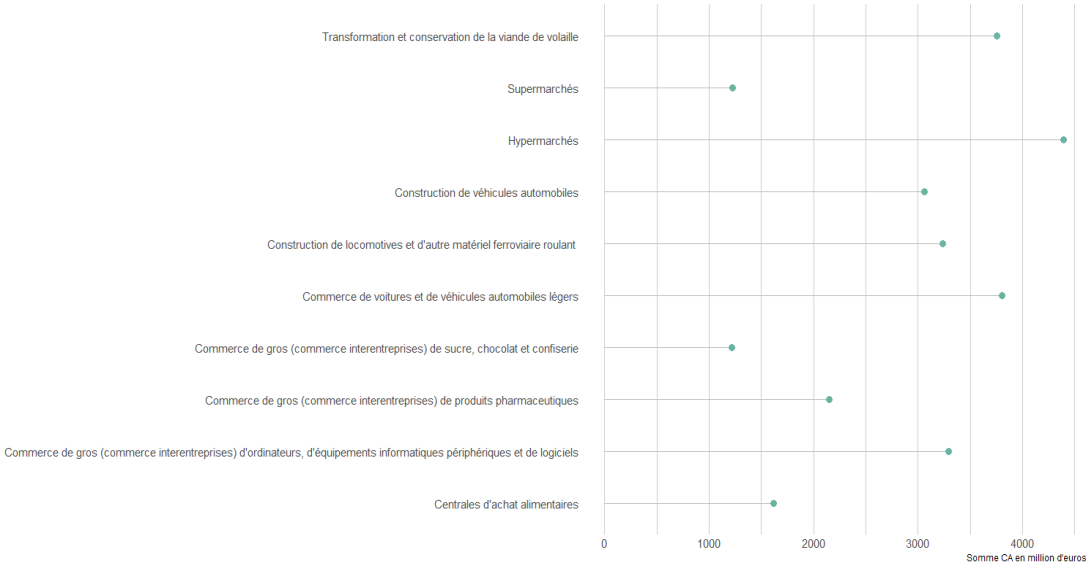


Figure 4.3: Les 10 activités avec le CA le plus élevé en 2022.

Dans ce premier graphique nous avons représenté les 10 activités principales qui ont généré le plus de CA cumulé en 2022. La somme des CA est représentée en million d'euros. Ce graphique nous permet d'avoir une idée sur les activités qui ont dominé le marché cette année. On y voit que les trois principales activités sont les hypermarchés, la transformation et conservation de la viande de volaile et le commerce automobile légers. On constate aussi que même parmis les 10 activités les plus importantes il y a des écarts conséquent : **plus de 2 000 millions d'euros entre les premiers et les derniers**.

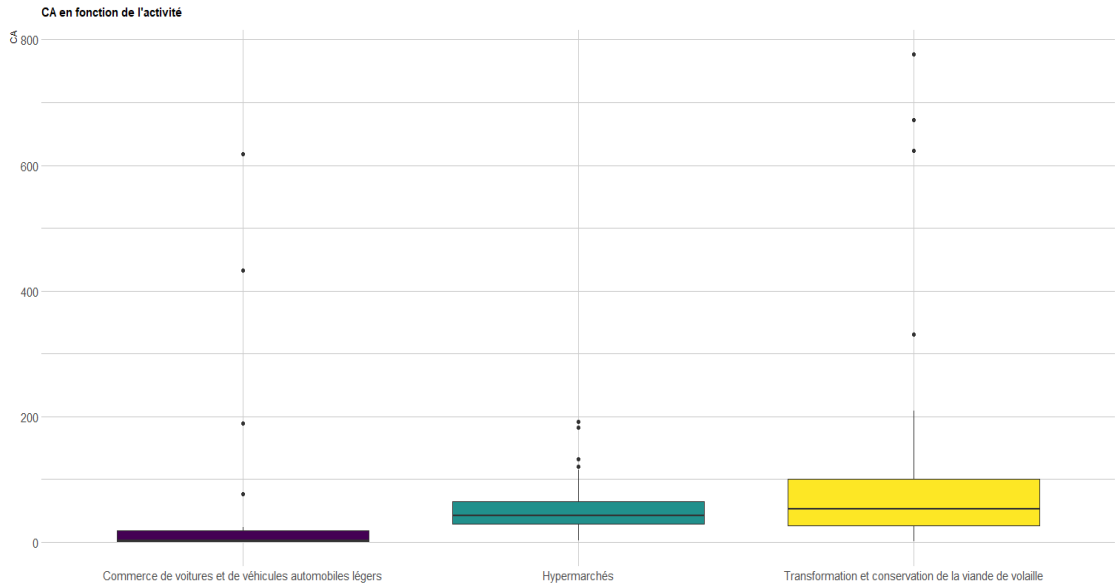


Figure 4.4: CA pour les 3 activités qui génèrent le plus en 2022

Dans ce graphique, nous avons représenté toutes les entreprises qui ont pour activité principale l'une des trois premières (obtenues grâce au graphique précédent). Nous avons donc 3 boxs pour les 3 différentes activités. Nous avons supprimés une valeur qui était extrêmement élevé et qui ruinait le visuel du graphique. Nous voyons ici que l'activité autour des volailes a la moyenne la plus haute, suivi par les hypermarchés puis le domaine automobile.

Il y a peu de valeurs extrêmes (les points hors des boîtes) ce qui montre que pour une activité, les CA générés sont relativement identiques. On peut aussi se dire que l'activité autour du domaine automobile est dans le top 3 des CA les plus élevés grâce au nombre d'entreprise et non à ses valeurs car la boite représentative de cette activité est comprise entre 0 et 20M d'euros.

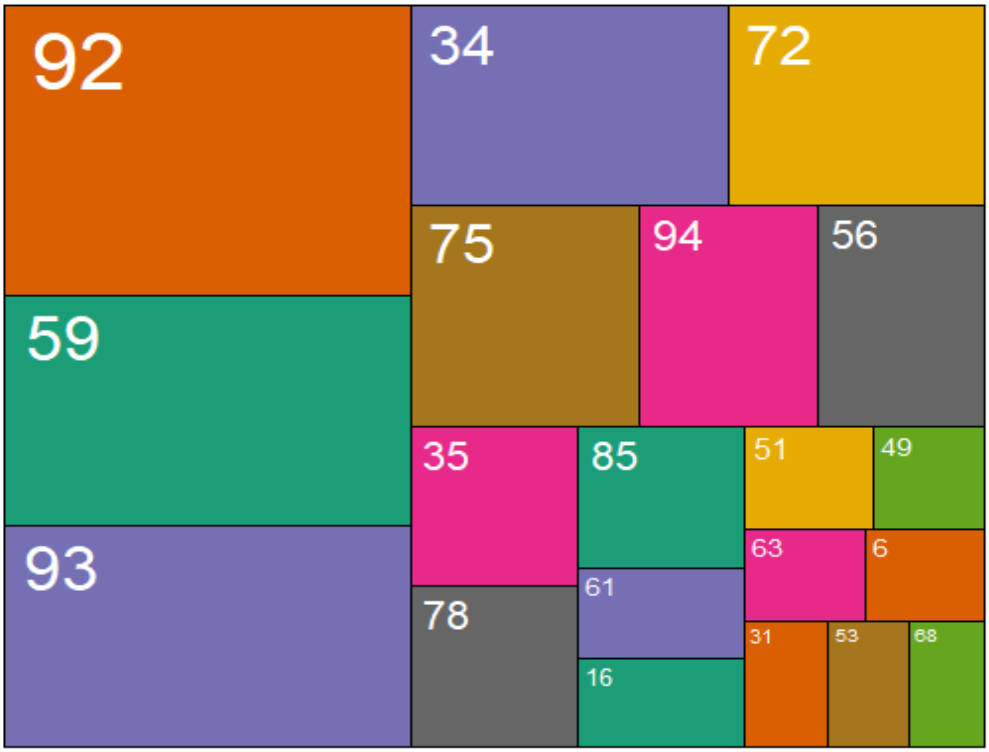


Figure 4.5: Les départements qui génèrent le plus en 2022

Dans ce graphique nous avons voulu représenté la repartition des CA en fonction des départements. Nous avons donc représentés les 20 départements qui ont généré le plus de CA pour l'année 2022, la taille des carrés est proportionnelle avec la somme du CA qu'ils ont généré cette année. Ce graphique nous permet de voir que 8 départements sortent du lot grace a leur CA important comparé aux autres.

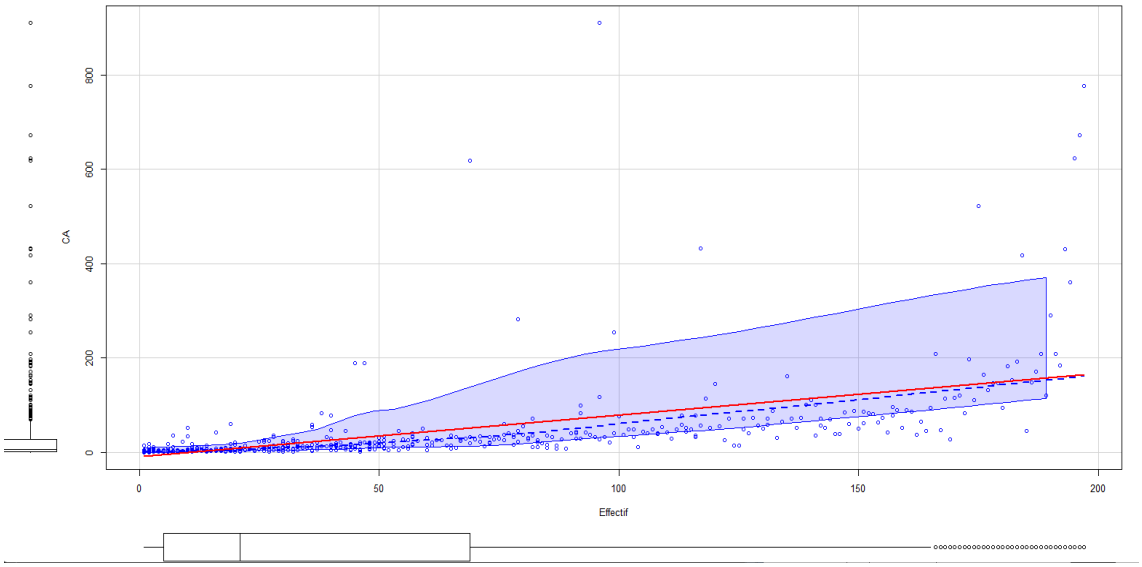
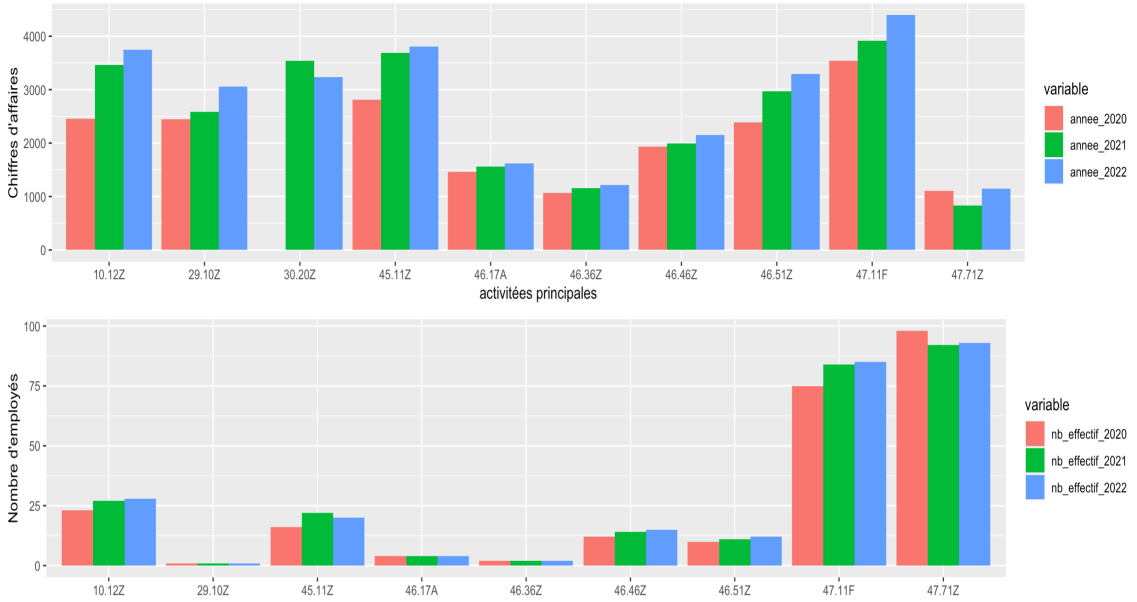


Figure 4.6: CA par rapport à l'effectif pour toutes les entreprises en 2022

Ce graphique représente le CA en million d'euros par rapport au effectif pour toutes les entreprises ayant eu une activité en 2022. Nous avons filtré de sorte à ne pas avoir d'effectif et de CA nul puis nous avons supprimés les valeurs extrêmes situé au dessus de 1 000M d'euros. Les boxplots situés sur le côté et en bas de ce graphique permettent de voir que la plupart des CA se situent entre 0 et approximativement 25M d'euros et que les effectifs sont principalement situés entre 2 et 75 personnes. La droite de regréssion montre ici que plus une entreprise a d'effectif, plus son CA est élevé car la droite de régréssion monte.

Figure 4.7: Les 10 activités ayant la plus grosse évolution de CA de 2020 à 2022

Ce graphique représente les activités avec la meilleure évolution du chiffre d'affaire pour la période de 2020 à 2022 en %. Deux activités sortent particulièrement du lot : la transformation et conservation du poisson et le courtage de valeurs mobilières et de marchandises qui ont une évolution du CA de plus de 20 000%. Tous ces domaines d'activités sont donc en forte croissance depuis les deux dernières années.



Le premier histogramme est un top 10 des secteurs avec les meilleurs chiffres d'affaires. Le deuxième représente le nombre d'employés parmi ce top 10 des meilleurs secteurs. On constate que seulement deux d'entre eux emploient un grand nombre d'employé. On peut supposer que ces secteurs sont plus stables grâce à leur qualité conséquente d'emploi.

Chapitre 5

Conclusion et perspectives

Pour conclure, nous pensons pouvoir proposer deux types d'investissement sur les secteurs d'activités.
La première, serait d'investir dans les secteurs ayant eu une forte évolution de leurs chiffres d'affaires sur les trois dernières années. Nos analyses nous permettent d'avoir les deux secteurs remplissant le plus cette condition: « Courtage de valeurs mobilières et de marchandises" et "Transformation et conservation de poisson ».
La deuxième stratégie d'investissement serait d'investir dans les domaines d'activités qui génèrent le plus de chiffre d'affaires. Ces secteurs sont les plus stables sur le long terme car leurs chiffres d'affaires restent élevés même s'ils ne présentent pas de grosses évolutions sur les 3 années étudiées. Les trois secteurs qui ressortent le plus sont « hypermarchés », « transformation et conservation de volaille » et « commerce de voitures et de véhicules automobiles légers ».
Dans une prochaine étude, il serait intéressant d'analyser les secteurs avec la plus grosse croissance sur ces trois dernières années pour voir si les résultats restent cohérents avec ceux obtenus cette année pour en conclure le retour positif de notre investissement

Bibliographie

https://www.data-to-viz.com/ https://www.data.gouv.fr/fr/datasets/base-sirene-des-entreprises-et-de-leurs-etablissements-siren-siret/ https://www.data.gouv.fr/fr/datasets/chiffres-cles-2022/

Annexes

Codes

- Figure 4.3

```
df1 <- dbGetQuery(con,"SELECT U.intitule AS activite, COUNT(C.siren) AS nb_siren,
(SUM(C.CA)/1000000) AS sum_CA
FROM chiffres_cles C, stockUL S, codeActiviteUL U
WHERE C.siren = S.siren
AND S.activitePrincipaleUniteLegale = U.code
AND C.annee = '2022'
AND C.CA > 1
GROUP BY S.activitePrincipaleUniteLegale
ORDER BY SUM(C.CA) DESC
LIMIT 0, 10")
attach(df1)
df1 %>%
  filter(!is.na(sum_CA)) %>%
  arrange(sum_CA) %>%
  tail(20) %>%
  mutate(Country=factor(activite, activite)) %>%
  ggplot( aes(x=activite, y=sum_CA) ) +
  geom_segment( aes(x=activite ,xend=activite, y=0, yend=sum_CA), color="grey") +
  geom_point(size=3, color="#69b3a2") +
  coord_flip() +
  theme_ipsum() +
  theme(
    panel.grid.minor.y = element_blank(),
    panel.grid.major.y = element_blank(),
    legend.position="none"
  ) +
  xlab("") +
  ylab("Somme CA en million d'euros")
```

- Figure 4.4

```
df3 <- dbGetQuery(con,"SELECT U.intitule AS activite, C.siren AS siren, (C.CA)/1000000 AS CA
FROM chiffres_cles C, stockUL S, codeActiviteUL U
WHERE C.siren = S.siren
AND S.activitePrincipaleUniteLegale = U.code
AND C.annee = '2022'
AND C.CA > 1
AND C.siren != '712034040'
AND (U.code = '47.11F' OR U.code = '45.11Z' OR U.code ='10.12Z')")
attach(df3)
df3 %>%
  ggplot( aes(x=activite, y=CA, fill=activite)) +
  geom_boxplot() +
  scale_fill_viridis(discrete = TRUE) +
  theme_ipsum() +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("CA en fonction de l'activité") +
  xlab("")
```

- Figure 4.5

```
df4 <- dbGetQuery(con,"SELECT S.activitePrincipaleUniteLegale AS Activite, C.num_dept AS Dept, SUM(C.CA) AS Somme_
FROM chiffres_cles C, stockUL S
WHERE C.siren = S.siren
```

```
AND C.annee = '2022'
GROUP BY C.num_dept
ORDER BY SUM(C.CA) DESC
LIMIT 0, 20")
attach(df4)

# Plot
treemap(df4,

        # data
        index="Dept",
        vSize="Somme_CA",
        type="index",

        # Main
        title="",
        palette="Dark2",

        # Borders:
        border.col=c("black"),
        border.lwds=1,

        # Labels
        fontsize.labels=3.5,
        fontcolor.labels="white",
        fontface.labels=1,
        align.labels=c("left", "top"),
        overlap.labels=0.5,
        inflate.labels=T

)
```

- Figure 4.6

```
df5 <- dbGetQuery(con,"SELECT C.effectif as Effectif, (C.CA)/1000000 as CA
FROM chiffres_cles C
WHERE C.annee = '2022'
AND C.effectif > 1
AND C.CA > 1
AND C.CA < 1000000000")
attach(df5)


df5<- transform(df5,Effectif = as.numeric(as.factor(Effectif)))

scatterplot(CA~Effectif, data=df5, regLine=list(method=lm, lty=1, lwd=2, col="red"))
```

- Figure 4.7

```
df6 <- dbGetQuery(con,"SELECT
  codeActiviteUL.intitule,
  SUM(CASE WHEN chiffres_cles.annee = 2020 THEN chiffres_cles.CA ELSE 0 END) AS CA_2020,
  SUM(CASE WHEN chiffres_cles.annee = 2022 THEN chiffres_cles.CA ELSE 0 END) AS CA_2022,
  (SUM(CASE WHEN chiffres_cles.annee = 2022 THEN chiffres_cles.CA ELSE 0 END) - SUM(CASE WHEN chiffres_cles.annee
FROM
  stockUL
INNER JOIN
  codeActiviteUL
    ON stockUL.activitePrincipaleUniteLegale = codeActiviteUL.code
INNER JOIN
  chiffres_cles
    ON stockUL.siren = chiffres_cles.siren
GROUP BY
  codeActiviteUL.intitule
ORDER BY change_in_percentage DESC
LIMIT 10;")
attach(df6)

# Plot
df6 %>%
  filter(!is.na(change_in_percentage)) %>%
```

```
  arrange(change_in_percentage) %>%
  tail(20) %>%
  mutate(Country=factor(intitule, intitule)) %>%
  ggplot( aes(x=intitule, y=change_in_percentage) ) +
  geom_segment( aes(x=intitule ,xend=intitule, y=0, yend=change_in_percentage), color="grey") +
  geom_point(size=3, color="#69b3a2") +
  coord_flip() +
  theme_ipsum() +
  theme(
    panel.grid.minor.y = element_blank(),
    panel.grid.major.y = element_blank(),
    legend.position="none"
  ) +
  xlab("") +
  ylab("Evolution du CA en %")
```

- Figure 4.8

```
#import du dataframe a partir de la requete SQL
df2020 <- dbGetQuery(con,"SELECT stockUL.activitePrincipaleUniteLegale AS activite, (SUM(chiffres_cles.CA)/1000000
FROM chiffres_cles, stockUL
WHERE chiffres_cles.siren = stockUL.siren
AND chiffres_cles.annee = '2020'
AND chiffres_cles.CA > 1
AND (stockUL.activitePrincipaleUniteLegale = '47.11F' OR stockUL.activitePrincipaleUniteLegale = '45.11Z' OR stock
GROUP BY stockUL.activitePrincipaleUniteLegale
ORDER BY SUM(chiffres_cles.CA) DESC
")
attach(df2020)
df2020

df2021 <- dbGetQuery(con,"SELECT stockUL.activitePrincipaleUniteLegale AS activite, (SUM(chiffres_cles.CA)/1000000
FROM chiffres_cles, stockUL
WHERE chiffres_cles.siren = stockUL.siren
AND chiffres_cles.annee = '2021'
AND chiffres_cles.CA > 1
AND (stockUL.activitePrincipaleUniteLegale = '47.11F' OR stockUL.activitePrincipaleUniteLegale = '45.11Z' OR stock
GROUP BY stockUL.activitePrincipaleUniteLegale
ORDER BY SUM(chiffres_cles.CA) DESC
")
attach(df2021)
df2021

df2022 <- dbGetQuery(con,"SELECT stockUL.activitePrincipaleUniteLegale AS activite, (SUM(chiffres_cles.CA)/1000000
FROM chiffres_cles, stockUL
WHERE chiffres_cles.siren = stockUL.siren
AND chiffres_cles.annee = '2022'
AND chiffres_cles.CA > 1
AND (stockUL.activitePrincipaleUniteLegale = '47.11F' OR stockUL.activitePrincipaleUniteLegale = '45.11Z' OR stock
GROUP BY stockUL.activitePrincipaleUniteLegale
ORDER BY SUM(chiffres_cles.CA) DESC
")
attach(df2022)
df2022

# install.packages("tidyverse")
library(tidyverse)
df <- left_join(df2021, df2022, by="activite")
df <- left_join(df, df2020, by="activite")
df

col_order <- c("activite", "annee_2020", "annee_2021",
               "annee_2022")
df <- df[, col_order]
df

# install.packages("reshape2")
library(reshape2)
df <- melt(df, id.vars="activite")
df
```

```r
#plot
p3 <- ggplot(df, aes(activite, value, fill = variable)) + geom_col(position = "dodge") +
  xlab("activitées principales") +
  ylab("Chiffres d'affaires")
p3

#deuxième plot

#import du dataframe a partir de la requete SQL
df2020_e <- dbGetQuery(con,"SELECT stockUL.activitePrincipaleUniteLegale AS activite, COUNT(chiffres_cles.siren) A
FROM chiffres_cles, stockUL
WHERE chiffres_cles.siren = stockUL.siren
AND chiffres_cles.annee = '2020'
AND (stockUL.activitePrincipaleUniteLegale = '47.11F' OR stockUL.activitePrincipaleUniteLegale = '45.11Z' OR stock
GROUP BY stockUL.activitePrincipaleUniteLegale
ORDER BY SUM(chiffres_cles.siren) DESC
")
attach(df2020_e)
df2020_e

#import du dataframe a partir de la requete SQL
df2021_e <- dbGetQuery(con,"SELECT stockUL.activitePrincipaleUniteLegale AS activite, COUNT(chiffres_cles.siren) A
FROM chiffres_cles, stockUL
WHERE chiffres_cles.siren = stockUL.siren
AND chiffres_cles.annee = '2021'
AND (stockUL.activitePrincipaleUniteLegale = '47.11F' OR stockUL.activitePrincipaleUniteLegale = '45.11Z' OR stock
GROUP BY stockUL.activitePrincipaleUniteLegale
ORDER BY SUM(chiffres_cles.siren) DESC
")
attach(df2021_e)
df2021_e

#import du dataframe a partir de la requete SQL
df2022_e <- dbGetQuery(con,"SELECT stockUL.activitePrincipaleUniteLegale AS activite, COUNT(chiffres_cles.siren) A
FROM chiffres_cles, stockUL
WHERE chiffres_cles.siren = stockUL.siren
AND chiffres_cles.annee = '2022'
AND (stockUL.activitePrincipaleUniteLegale = '47.11F' OR stockUL.activitePrincipaleUniteLegale = '45.11Z' OR stock
GROUP BY stockUL.activitePrincipaleUniteLegale
ORDER BY SUM(chiffres_cles.siren) DESC
")
attach(df2022_e)
df2022_e

# install.packages("tidyverse")
library(tidyverse)
df_entreprise <- left_join(df2020_e, df2021_e, by="activite")
df_entreprise <- left_join(df_entreprise, df2022_e, by="activite")
df_entreprise

# install.packages("reshape2")
library(reshape2)
df_entreprise <- melt(df_entreprise, id.vars="activite")
df_entreprise

p2 <- ggplot(df_entreprise, aes(activite, value, fill = variable)) + geom_col(position = "dodge") +
  xlab("activitées principales") +
  ylab("Nombre d'entreprises")
p2

#3eme plot

#import du dataframe a partir de la requete SQL
df2020_effectif <- dbGetQuery(con,"SELECT stockUL.activitePrincipaleUniteLegale AS activite, COUNT(chiffres_cles.e
FROM chiffres_cles, stockUL
WHERE chiffres_cles.siren = stockUL.siren
AND chiffres_cles.annee = '2020'
AND (stockUL.activitePrincipaleUniteLegale = '47.11F' OR stockUL.activitePrincipaleUniteLegale = '45.11Z' OR stock
GROUP BY stockUL.activitePrincipaleUniteLegale
ORDER BY nb_effectif_2020 DESC
```

```
")
attach(df2020_effectif)
df2020_effectif

#import du dataframe a partir de la requete SQL
df2021_effectif <- dbGetQuery(con,"SELECT stockUL.activitePrincipaleUniteLegale AS activite, COUNT(chiffres_cles.e
FROM chiffres_cles, stockUL
WHERE chiffres_cles.siren = stockUL.siren
AND chiffres_cles.annee = '2021'
AND (stockUL.activitePrincipaleUniteLegale = '47.11F' OR stockUL.activitePrincipaleUniteLegale = '45.11Z' OR stock
GROUP BY stockUL.activitePrincipaleUniteLegale
ORDER BY nb_effectif_2021 DESC
")
attach(df2021_effectif)
df2021_effectif

#import du dataframe a partir de la requete SQL
df2022_effectif <- dbGetQuery(con,"SELECT stockUL.activitePrincipaleUniteLegale AS activite, COUNT(chiffres_cles.e
FROM chiffres_cles, stockUL
WHERE chiffres_cles.siren = stockUL.siren
AND chiffres_cles.annee = '2022'
AND (stockUL.activitePrincipaleUniteLegale = '47.11F' OR stockUL.activitePrincipaleUniteLegale = '45.11Z' OR stock
GROUP BY stockUL.activitePrincipaleUniteLegale
ORDER BY nb_effectif_2022 DESC
")
attach(df2022_effectif)
df2022_effectif

# install.packages("tidyverse")
library(tidyverse)
df_effectif <- left_join(df2020_effectif, df2021_effectif, by="activite")
df_effectif <- left_join(df_effectif, df2022_effectif, by="activite")
df_effectif

# install.packages("reshape2")
library(reshape2)
df_effectif <- melt(df_effectif, id.vars="activite")
df_effectif

p1 <- ggplot(df_effectif, aes(activite, value, fill = variable)) + geom_col(position = "dodge") +
  xlab("activitées principales") +
  ylab("Nombre d'employés")
p1

#install.packages("gridExtra")
library("gridExtra")
#install.packages("cowplot")
library("cowplot")

plot_grid(p3, p1, p2, labels=c(" ", " ", " "), ncol = 1, nrow = 3)
```