PROJET RÉALISÉ PAR

L'ÉQUIPE LES REGRESSIONS LINÉAIRES DU GROUPE DE TD1 OU TD2

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

Haitham Alfakhry, Noah Chayrigues, Arthur Feschet, Felicie Sadet.

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

Avril 2025

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

Déclaration de non plagiat

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 : _____

29 avril 2025

Remerciements

Nous tenons à exprimer nos plus sincères remerciements envers nos enseignants Mme Sandra Bringay et Mme Marine Demangeot pour leur soutien et leurs conseils avisés. Nous remercions également les membres de ce groupe pour leur travail acharné tout au long de ce projet. Nous souhaitons également saluer chaleureusement ChatGPT et Copilot pour leur soutien discret et précieux durant l'élaboration de notre projet.

29 avril 2025

Résumé

Résumé

De nos jours, le système éducatif français est un levier déterminant dans la construction des trajectoires professionnelles. L'Indice de Position Sociale (IPS), sert de mesure du contexte socio-économique et permet d'étudier les inégalités dans la réussite scolaire et l'insertion professionnelle.

En combinant l'IPS avec différents indicateurs comme le taux de réussite au baccalauréat pour les lycées professionnels ou le taux de poursuite d'études il est possible d'obtenir une vue d'ensemble sur les disparités entre établissements. Les différentes données (IPS, réussite scolaire et insertion) couvrant la période 2018-2022 nous offrent une opportunité de mesurer précisément ces liens. Bien qu'une tendance générale suggère qu'un IPS élevé soit associé à de meilleurs résultats académiques, certains lycées à IPS faible affichent également d'excellents taux de réussite et une excellente insertion. Cette contradiction apparente questionne la force et la nature exacte de la relation entre le contexte socio-économique et les performances scolaires, justifiant ainsi une analyse fine par requêtes SQL et analyses statistiques. En approfondissant l'analyse avec R, nous constatons une corrélation positive significative entre l'IPS et le taux de réussite ($r \approx 0{,}42$, $R^2 \sim 0{,}17$) ainsi qu'entre l'IPS et l'insertion professionnelle ($r \approx 0{,}184$, $R^2 \sim 0{,}034$) indiquant qu'en moyenne, un environnement socio-économique favorable tend à améliorer les performances scolaires. Pour aller plus loin, un test d'ANOVA a été réalisé en annexe afin d'étudier les disparités interrégionales. Ce test a montré que les différences de réussite scolaire entre régions sont significatives ce qui indique que le lieu géographique influence aussi les performances, indépendamment de l'IPS. Par ailleurs les visualisations (boxplots, histogrammes, cartes) confirment ces résultats : cependant on observe de grandes différences entre les établissements en termes de réussite et d'insertion. Cela montre que l'IPS est un indicateur utile, mais qu'il ne suffit pas à lui seul pour expliquer les performances. D'autres éléments, comme les politiques des établissements ou la situation économique locale, jouent aussi des rôles importants.

Table des matières

Chapitre 1

Introduction

1.1 Introduction

Depuis sa création, le système éducatif joue un rôle crucial dans la construction des trajectoires professionnelles des jeunes. En effet, il existe plusieurs critères influençant l'insertion professionnelle des diplômés, parmi ses critères,l'Indice de Position Sociale (IPS) se distingue.D'après DataGouv, "cet indice est construit à partir des professions et catégories socioprofessionnelles (PCS) des représentants légaux des élèves".
https://www.education.gouv.fr L'IPS se mesure par points, plus l'IPS est élevé, plus la position sociale de l'élève est élevée.
Vous trouverez une image un peu plus bas qui explique l'échelle des positions sociales datant de 2016.
https://www.s2de.fr/. Lecture : Un IPS de 110 indique un père dit "ouvrier qualifié" et une mère "policière".
Par ailleurs, cette question devient importante à une époque où l'insertion professionnelle des jeunes constitue un enjeu majeur.En exploitant diverses sources de données, nous chercherons à identifier les facteurs déterminants de la réussite professionnelle ainsi que la réussite scolaire.
De plus, on s'intéressera plus particulièrement aux IPS des voies professionnelles des lycées dans le but de constater s'il ya des disparités entre les lycées en terme de réussite scolaire et d'insertion professionnelle. Tout d'abord, nous étudierons le lien entre l'ips et la réussite et, par la suite, entre l'ips et l'insertion professionnelle.
Cette analyse s'inscrit dans une volonté de mieux comprendre les mécanismes d'inégalités sociales au sein du système éducatif français.
L'IPS, bien qu'imparfait, reste un indicateur clé pour étudier les effets du capital social et économique sur les parcours scolaires. Il s'agira donc d'interroger non seulement les chiffres, mais aussi les contextes dans lesquels ils s'inscrivent.

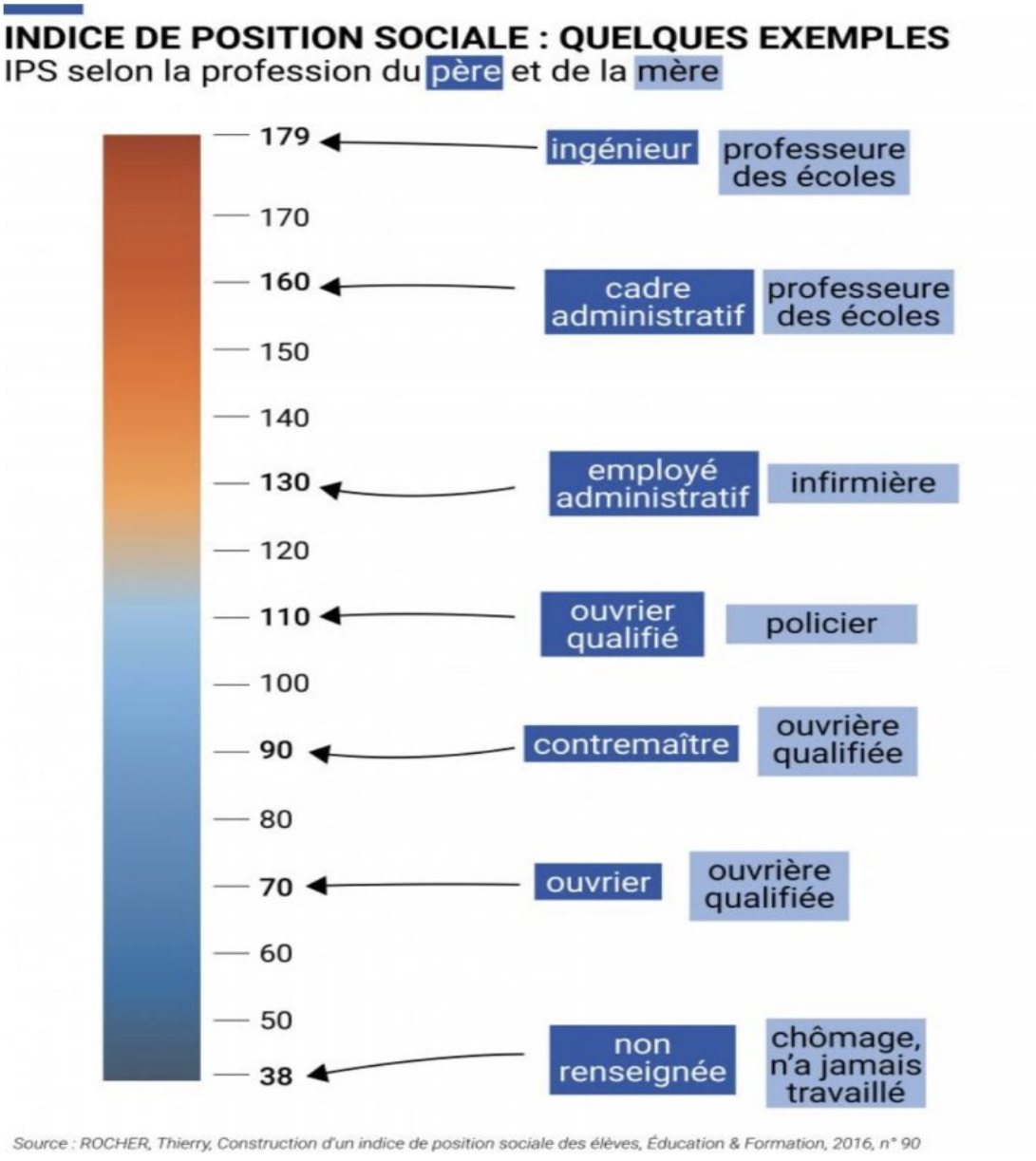


FIGURE 1.1: Figure 2.1 Indice de Position Sociale

L'IPS des voies professionnelles est-il représentatif de la réussite scolaire et d'une meilleure insertion professionnelle ?

1.2 Collaborateurs

Alfakhry Haitham : Étudiant n°21613231
Noah Chayrigues : Étudiant n°22303586
Arthur Feschet : Etudiant n°22301534
Felicie Sadet : Etudiant n°22313367

Chapitre 2

Base de données

2.1 Provenance des données

1) Jeu de données sur les indices de position sociale (IPS) :

https://www.data.gouv.fr/fr/datasets

Description : Ce jeu de données présente l'Indice de Position Sociale (IPS) pour les lycées en France. L'IPS est un indicateur utilisé pour évaluer la position socio-économique des élèves et de leurs familles. Il nous permet ainsi d'identifier les inégalités selon les établissements. Ce jeu recense de nombreuses informations sur des milliers de lycées durant la période 2017-2022. Les données sont regroupées dans un fichier .csv, contenant 14 colonnes pour plus de 20 milliers de lignes.

2) Jeu de données sur l'insertion professionnelle :

https://data.education.gouv.fr

Description : Ce jeu de données fournit des informations sur l'insertion professionnelle des jeunes diplômés de lycées professionnels. Il inclut des données sur le taux de poursuite d'études et le taux d'insertion. Cela permet d'analyser l'efficacité des lycées professionnels pour l'insertion des jeunes. Les données sont regroupées dans un fichier .csv sur la période 2018-2023, contenant 10 colonnes pour plus de 30 milliers de lignes.

3) Jeu de données sur le taux de réussite des lycées professionnels :

https://www.data.gouv.fr/fr/datasets

Description : Ce jeu de données donne des informations sur la réussite des élèves dans les lycées professionnels. Plusieurs indicateurs nous seront utile tels que le Taux de réussite ou le Taux de réussite attendu. Ce jeu de données permet une analyse des résultats scolaires. Les données sont regroupées dans un fichier .csv sur la période 2012-2023, contenant 130 colonnes pour plus de 20 milliers de lignes.

2.2 Choix des tables

Dans un premier temps, nous avons choisi la table IPS car c'est un indicateur clé pour comprendre le contexte socio-économique des élèves. Par ailleurs, pour compléter notre analyse nous avons choisi la table sur l'insertion dans le monde professionnels, ces critères permettent d'obtenir une vue d'ensemble de ce qui se passe après la scolarité : poursuite d'études, interruption ou insertion dans le monde

du travail, et de relier cela à l'IPS. Enfin, nous avons ajouté la table sur l'indicateur de réussite des lycées professionnels. Ces données sont essentielles pour établir si des lycées avec des IPS élevés réussissent mieux à préparer les élèves et si cela se traduit par une meilleure insertion professionnelle. En somme, les 3 tables nous permettent d'identifier les liens entre IPS, insertion et réussite scolaire.

2.3 Filtrage des données

Pour rendre l'analyse plus pertinente, nous avons effectué certains choix de filtrage sur les 3 jeux de données :

Nous avons crée un 4ème jeu de données comprenant : UAI, nom_lycee Nous avons fais ce choix pour faciliter la compréhension de notre analyse

Réussite des lycées :
UAI,Année,ville,taux_de_reussite,taux_de_reussite_attendu.
Nous gardons UAI et année pour les mêmes raisons. Par ailleurs, nous gardons la ville qui peut être un élément d'interprétation supplémentaire. Pour finir, nous gardons le taux de reussite et taux de reussite attendu, respectivement pour comprendre s'il ya une corrélation entre IPS et réussite et pour comparer au taux de réussite réel.

Insertion professionnelle :
Année,UAI,region,taux_interruption_formation, taux_poursuite_etude.
Idem pour UAI et Année. Nous gardons région pour obtenir de potentielle inégalités entre les régions. Nous gardons le taux_interruption_formation et taux_poursuite_etude qui sont des éléments clés de notre analyse que nous développerons plus tard dans ce document.

NB : Année : entre 2018 et 2022 UAI : Unité Administrative d'Identification : clé

En somme, nous étudions différentes données sur les lycées de France, cette combinaison des 3 tables nous permet d'identifier les liens entre IPS, insertion professionnelle et réussite scolaire.

2.4 Descriptif des tables

Pour chaque table conservée, préciser le nombre de lignes et de colonnes après filtrage, lister les colonnes et donner pour chacune le type, la signification du champ et des caractéristiques (unique, clés, valeur manquante, ...) en remplissant le tableau ci-dessous.

Nom colonne	Type	Signification	Caractéristiques
Année	Texte	Période d'analyse des données (ex : cumul 20-21)	Données temporelles permettant d'observer l'évolution
UAI	Texte	code unique d'identification de l'établissement	Identifiant unique par lycée
Nom_lycée	Texte	Nom du lycée	Permet d'identifier chaque établissement
Région	Texte	Région administrative du lycée	Permet une analyse géographique
Taux Interruption formation	Numérique	Pourcentage d'élèves interrompant leur formation	Indicateur de décrochage scolaire
Taux Poursuite Études	Numérique	Pourcentage d'élèves poursuivant leurs études après diplôme	Indicateur de continuité scolaire

TABLE 2.1: insertion

Nom colonne	Type	Signification	Caractéristiques
Année	Texte	Année scolaire de référence (format AAAA-AAAA)	Données disponible pour 2 années ex : 2018-2019
UAI	Texte	code unique d'identification de l'établissement (unité administrative immatriculée)	7 caractères alphanumériques - clé primaire potentielle
Nom_lycée	Texte	dénomination officielle du lycée	inclus souvent les spécialités ou métier
Académie	Texte	Nom de l'académie de rattachement	30 académie métropolitaines et ultramarines
IPS voie PRO	Numérique	Indice de Position Sociale pour la filière professionnelle	Valeur 0-180, vide pour les LEGT purs

TABLE 2.2: IPS lycées

Nom colonne	Type	Signification	Caractéristiques
Année	Numérique	Année des données	Format YYYY (2012-2018)
UAI	Texte	code unique d'identification de l'établissement	Code administratif
Nom_lycée	Texte	Nom de l'établissement scolaire	Identifiant unique de l'école
Ville	Texte	Localisation de l'établissement	
Taux de réussite	Numérique	Taux de réussite par spécialité	En pourcentage, multiple colonnes
Taux de réussite attendus	Numérique	Taux attendus de réference	Par académie en France, multiple colonne

TABLE 2.3: taux de réussite

Nom colonne	**Type**	**Signification**	**Caractéristique**
UAI	Texte	Code unique d'identification de l'établissement	Clé primaire (Identifiant unique par lycée)
Nom_lycée	Texte	Nom officiel du lycée	Sert à identifier l'établissement

TABLE 2.4: Entité : Lycée

2.5 Modèles du MCD

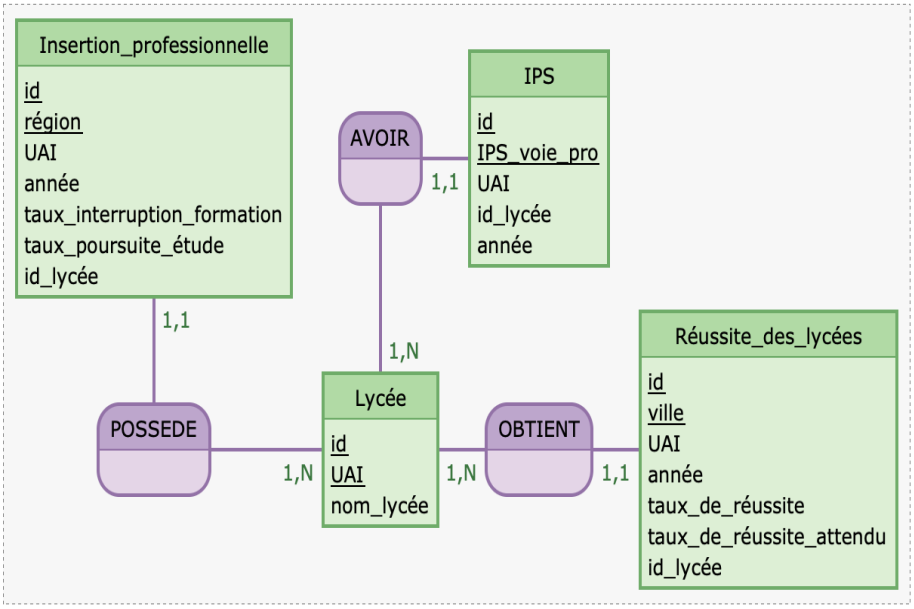


FIGURE 2.1: MCD

2.6 Modèle du MOD

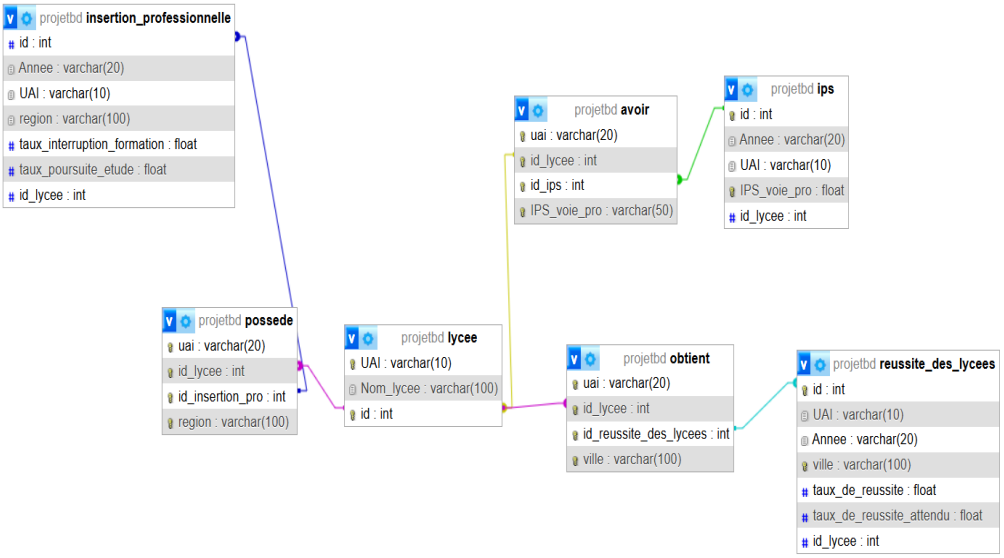


FIGURE 2.2: MOD

2.7 Import des données

Pour l'import des données nous avons du

Tout d'abord, nous avons commencés par supprimer les colonnes qui ne nous intéressent pas (voir description des tables). Nous avons que les attributs qui nous interessent. Par la suite, nous avons eu un soucis d'uniformisation pour les années : format XXXX et XXXX - XXXX. Nous avons fais le choix de mettre nos 3 jeux de données contenant année au format XXXX - XXXX, c'est à dire 2018 devient 2017 - 2018. Dans nos fichiers, nous constatons des valeurs nulles, cela ne nous gênent pas dans la mesure ou nous ignorons les valeurs nulles. Nous n'avons pas eu de problèmes d'importation à cause de ces valeurs nulles, nous avons alors décidés de les laisser dans nos fichiers. De ce fait, les lycées comprenant des valeurs nulles sur les valeurs numériques (pour l'analyse) ne seront pas traités. Nous avons évidemment rencontrés des problèmes d'encodages qui nous ont empêchés l'import totals des données,problème de caractères et d'autres problèmes mineurs. Nous avons résolus le problème en passant par Notepad++ en forçant l'encodage : passage de ANSI (encodage américain) à UTF-8.

2.8 Requêtes réalisées

Requête n°1 & n°2 — IPS faible vs IPS élevé (2018–2022)

Requête n°1 :

```
SELECT i.UAI, i.Annee, r.ville, i.IPS_voie_pro, r.taux_de_reussite
FROM ips i
JOIN reussite_des_lycees r ON i.UAI = r.UAI AND i.Annee = r.Annee
WHERE i.Annee IN ('2018 - 2019', '2019 - 2020',
'2020 - 2021', '2021 - 2022')
AND i.IPS_voie_pro > 0 AND r.taux_de_reussite > 0
ORDER BY i.IPS_voie_pro ASC;
```

Requête n°2 :

```
SELECT i.UAI, i.Annee, r.ville, i.IPS_voie_pro, r.taux_de_reussite
FROM ips i
JOIN reussite_des_lycees r ON i.UAI = r.UAI AND i.Annee = r.Annee
WHERE i.Annee IN ('2018 - 2019', '2019 - 2020',
'2020 - 2021', '2021 - 2022')
  AND i.IPS_voie_pro > 0
  AND r.taux_de_reussite > 0
ORDER BY i.IPS_voie_pro DESC;
```

UAI	Annee	ville	IPS_voie_pro ▲ 1	taux_de_reussite
9730371R	2021 - 2022	SAINT LAURENT DU MARONI	54.2	53
9730235T	2020 - 2021	ST LAURENT DU MARONI	54.8	79
9730513V	2019 - 2020	ST LAURENT DU MARONI	54.9	90
9730235T	2021 - 2022	SAINT LAURENT DU MARONI	55	55
9730371R	2020 - 2021	ST LAURENT DU MARONI	55.1	59
9730513V	2021 - 2022	SAINT LAURENT DU MARONI	55.2	55
9730513V	2018 - 2019	ST LAURENT DU MARONI	55.3	64
9730421V	2020 - 2021	MANA	55.4	77
9730235T	2019 - 2020	ST LAURENT DU MARONI	55.5	90
9730421V	2019 - 2020	MANA	55.6	88
9730513V	2020 - 2021	ST LAURENT DU MARONI	55.6	76
0131463V	2019 - 2020	MARSEILLE 15	55.9	80
9730235T	2018 - 2019	ST LAURENT DU MARONI	56	68
9730371R	2019 - 2020	ST LAURENT DU MARONI	56.4	79
9730421V	2021 - 2022	MANA	56.9	66
9730371R	2018 - 2019	ST LAURENT DU MARONI	56.9	57
9741233X	2020 - 2021	ST BENOIT	57	80
9730421V	2018 - 2019	MANA	57.3	91
0134101M	2020 - 2021	MARSEILLE 15	58.1	74
0131463V	2018 - 2019	MARSEILLE 15	58.1	50
9741233X	2019 - 2020	ST BENOIT	58.3	87
9741233X	2021 - 2022	SAINT BENOIT	58.4	80
9741233X	2018 - 2019	ST BENOIT	58.5	82
9760270P	2020 - 2021	ACOUA	58.6	66
9760270P	2021 - 2022	ACOUA	59	67

FIGURE 2.3: Requête n°1

UAI	Annee	ville	IPS_voie_pro ▾ 1	taux_de_reussite
0781581V	2019 - 2020	ST GERMAIN EN LAYE	136.5	100
0781581V	2018 - 2019	ST GERMAIN EN LAYE	134.4	97
0781581V	2020 - 2021	ST GERMAIN EN LAYE	132.6	100
0781581V	2021 - 2022	SAINT GERMAIN EN LAYE	131.3	100
0754016H	2019 - 2020	PARIS 19	130	97
0781582W	2020 - 2021	VERSAILLES	129.3	98
0950804H	2020 - 2021	OSNY	128.8	97
0912321D	2019 - 2020	PALAISEAU	128.3	97
0912321D	2021 - 2022	PALAISEAU	128.3	88
0951221L	2018 - 2019	PONTOISE	128	100
0754016H	2018 - 2019	PARIS 19	128	90
0781582W	2021 - 2022	VERSAILLES	127.9	87
0912321D	2020 - 2021	PALAISEAU	127.4	86
0950804H	2021 - 2022	OSNY	127.4	100
0950804H	2018 - 2019	OSNY	126.4	95
0942355Z	2020 - 2021	JOINVILLE LE PONT	126.3	95
0440255N	2019 - 2020	NANTES	126	100
0941407U	2021 - 2022	IVRY SUR SEINE	125.9	95
0951221L	2021 - 2022	PONTOISE	125.7	100
0912321D	2018 - 2019	PALAISEAU	125.6	91
0950804H	2019 - 2020	OSNY	125.6	94
0754016H	2020 - 2021	PARIS 19	125.6	93
0781856U	2021 - 2022	SAINT GERMAIN EN LAYE	125.3	93
0941407U	2020 - 2021	IVRY SUR SEINE	125	86
0941407U	2019 - 2020	IVRY SUR SEINE	124.8	79

FIGURE 2.4: Requête n°2

Objectif : Comparer les établissements les plus défavorisés (IPS bas) et les plus favorisés (IPS haut) pour voir comment varie le taux de réussite au bac pro selon le niveau socio-économique (IPS).

Résultat : Les lycées les plus favorisés ont, en général, les meilleurs taux de réussite. Mais certains établissements parmi les plus défavorisés obtiennent également d'excellents résultats. Interprétation : Ces premiers résultats suggèrent qu'il existe une corrélation positive entre IPS et réussite, cependant des lycées avec un IPS plus faible réussissent très bien. Cela interroge sur la relation direct entre IPS et réussite, il faut donc d'autres requêtes pour certifier la relation.

Requête 3 : Évolution de la réussite selon les tranches IPS (2018–2022)

Requête n°3 :

```
SELECT i.Annee,
  CASE
    WHEN i.IPS_voie_pro < 90 THEN 'Faible IPS'
    WHEN i.IPS_voie_pro BETWEEN 90 AND 110 THEN 'Moyen IPS'
    ELSE 'Fort IPS'
  END AS tranche_ips,
  COUNT(*) AS nb_lycees,
  ROUND(AVG(r.taux_de_reussite), 1) AS moyenne_reussite
FROM ips i
INNER JOIN reussite_des_lycees r
  ON i.UAI = r.UAI AND i.Annee = r.Annee
WHERE i.Annee IN ('2018 - 2019', '2019 - 2020',
'2020 - 2021', '2021 - 2022')
  AND i.IPS_voie_pro > 0
  AND r.taux_de_reussite > 0
GROUP BY i.Annee, tranche_ips
```

```
ORDER BY i.Annee ASC,
  FIELD(tranche_ips, 'Faible IPS', 'Moyen IPS', 'Fort IPS');
```

Annee ▲ 1	tranche_ips	nb_lycees	moyenne_reussite
2018 - 2019	Faible IPS	1312	80.4
2018 - 2019	Moyen IPS	631	88.2
2018 - 2019	Fort IPS	51	93.7
2019 - 2020	Faible IPS	1295	89.2
2019 - 2020	Moyen IPS	651	93.6
2019 - 2020	Fort IPS	54	96.9
2020 - 2021	Faible IPS	1286	84.4
2020 - 2021	Moyen IPS	656	91
2020 - 2021	Fort IPS	58	95.9
2021 - 2022	Faible IPS	1256	79.9
2021 - 2022	Moyen IPS	675	87.5
2021 - 2022	Fort IPS	65	92.9

FIGURE 2.5: Requête n°3

objectif : Pour la 3ème requête, notre objectif est de mesurer l'évolution des taux de réussite moyens au bac dans les lycées pro entre 2018-2022 en fonction de l'IPS voie pro, regroupé en trois niveaux. *Faible :* IPS inférieurs 90 / *Moyen :* IPS entre 90 et 110 / *Fort :* IPS supérieurs 110 Pour chaque année, la requête dénombre les établissements par tranche IPS et calcule le taux moyen de réussite, afin de vérifier si un IPS plus élevé est associé à de meilleures performances scolaires et d'identifier les disparités structurelles liées au contexte socio-économique.

résultat : Entre 2018 et 2022, les taux moyens de réussite au bac varient significativement selon l'IPS. Par exemple, en 2018–2019, les lycées à Faible IPS (1312 établissements) affichent une moyenne d'environ 80,4 %, ceux à Moyen IPS (631 établissements) environ 88,2 %, tandis que les établissements à Fort IPS (51 établissements) dépassent 93,7 %. Ces tendances, stables sur la période, illustrent un écart pouvant atteindre 15 à 17 points entre les extrêmes.

interprétation : *Corrélation positive :* Les lycées à Fort IPS (110) affichent des taux de réussite supérieurs à 90 %, tandis que ceux à Faible IPS se situent entre 80 et 85%. *Écart de performance :* L'écart entre Faible et Fort IPS peut atteindre 15 à 17 points, révélant une disparité importante liée au contexte socio-économique. *Tendance stable (2018–2022) :* La répartition reste constante sur la période, avec les établissements à Fort IPS toujours en tête et ceux à Moyen IPS en position intermédiaire, ce qui confirme l'influence du niveau socio-économique sur la réussite scolaire.

Requête n°4 — Moyenne annuelle de l'IPS et du taux de réussite (2018–2022)

Requête n°4 :

```
SELECT i.Annee,
       AVG(i.IPS_voie_pro) AS moyenne_ips,
       AVG(r.taux_de_reussite) AS moyenne_reussite
FROM ips i
INNER JOIN reussite_des_lycees r
       ON i.UAI = r.UAI
      AND i.annee = r.Annee
WHERE i.IPS_voie_pro > 0
  AND i.annee IN ('2018 - 2019', '2019 - 2020',
  '2020 - 2021', '2021 - 2022')
  AND r.taux_de_reussite > 0
GROUP BY i.Annee;
```

Annee	moyenne_ips	moyenne_reussite
2018 - 2019	86.7429789100272	83.2001003009027
2019 - 2020	87.03040000724792	90.876
2020 - 2021	87.09465004920959	86.901
2021 - 2022	87.43186371216554	82.86523046092185

FIGURE 2.6: Requête n°4

Objectif : Étudier l'évolution annuelle de l'IPS moyen et du taux de réussite moyen, afin de vérifier s'il existe une corrélation globale entre origine sociale (IPS) et performance scolaire (réussite).

Résultat : Les IPS moyens restent relativement stables (entre 86.7 et 87.4), tandis que le taux de réussite varie, avec un pic à 90.88 % en 2019–2020 (année du Covid).

Interprétation : Cette requête met en évidence une corrélation modérée mais réelle : quand l'IPS augmente légèrement, la réussite a tendance à suivre. Mais d'autres facteurs externes (contexte sanitaire...) peuvent aussi fortement impacter les résultats.

Requête n°5 – Lycées exemplaires : forte réussite malgré un IPS faible (2018–2022)

Requête n°5 :

```sql
SELECT i.UAI, i.Annee, l.nom_lycee, r.ville, i.IPS_voie_pro,
r.taux_de_reussite
FROM ips i
JOIN reussite_des_lycees r ON i.UAI = r.UAI AND i.Annee = r.Annee
JOIN lycee l ON i.UAI = l.UAI
WHERE i.IPS_voie_pro BETWEEN 74 AND 80
  AND r.taux_de_reussite > 98
  AND i.Annee IN ('2018 - 2019', '2019 - 2020',
  '2020 - 2021', '2021 - 2022');
```

UAI	Annee	nom_lycee	ville	IPS_voie_pro	taux_de_reussite
0592963A	2019 - 2020	LYCEE POLYVALENT PRIVE SAINT MARTIN	ROUBAIX	76.9	99
0592963A	2019 - 2020	LYCEE POLYVALENT PRIVE SAINT MARTIN	ROUBAIX	76.9	99
0592963A	2019 - 2020	LYCEE POLYVALENT PRIVE SAINT MARTIN	ROUBAIX	76.9	99
0592963A	2019 - 2020	LYCEE POLYVALENT PRIVE SAINT MARTIN	ROUBAIX	76.9	99
0592963A	2019 - 2020	LYCEE POLYVALENT PRIVE SAINT MARTIN	ROUBAIX	76.9	99
0592963A	2019 - 2020	LYCEE POLYVALENT PRIVE SAINT MARTIN	ROUBAIX	76.9	99
0592963A	2019 - 2020	LYCEE POLYVALENT PRIVE SAINT MARTIN	ROUBAIX	76.9	99
0592963A	2019 - 2020	LYCEE POLYVALENT PRIVE SAINT MARTIN	ROUBAIX	76.9	99
0592963A	2019 - 2020	LYCEE POLYVALENT PRIVE SAINT MARTIN	ROUBAIX	76.9	99
0681888H	2019 - 2020	LYCEE POLYVALENT AMELIE ZURCHER	WITTELSHEIM	74.7	100
0681888H	2019 - 2020	LYCEE POLYVALENT AMELIE ZURCHER	WITTELSHEIM	74.7	100
0681888H	2019 - 2020	LYCEE POLYVALENT AMELIE ZURCHER	WITTELSHEIM	74.7	100
0681888H	2019 - 2020	LYCEE POLYVALENT AMELIE ZURCHER	WITTELSHEIM	74.7	100
0681888H	2019 - 2020	LYCEE POLYVALENT AMELIE ZURCHER	WITTELSHEIM	74.7	100
0681888H	2019 - 2020	LYCEE POLYVALENT AMELIE ZURCHER	WITTELSHEIM	74.7	100
0681888H	2019 - 2020	LYCEE POLYVALENT AMELIE ZURCHER	WITTELSHEIM	74.7	100
0681888H	2019 - 2020	LYCEE POLYVALENT AMELIE ZURCHER	WITTELSHEIM	74.7	100
0681888H	2019 - 2020	LYCEE POLYVALENT AMELIE ZURCHER	WITTELSHEIM	74.7	100
0681888H	2019 - 2020	LYCEE POLYVALENT AMELIE ZURCHER	WITTELSHEIM	74.7	100
0681888H	2019 - 2020	LYCEE POLYVALENT AMELIE ZURCHER	WITTELSHEIM	74.7	100
9710884J	2019 - 2020	LYCEE POLYVALENT RAOUL GEORGES NICOLO LYCEE DES ME...	BASSE TERRE	78.3	100
9710884J	2019 - 2020	LYCEE POLYVALENT RAOUL GEORGES NICOLO LYCEE DES ME...	BASSE TERRE	78.3	100
9710884J	2019 - 2020	LYCEE POLYVALENT RAOUL GEORGES NICOLO LYCEE DES ME...	BASSE TERRE	78.3	100
9710884J	2019 - 2020	LYCEE POLYVALENT RAOUL GEORGES NICOLO LYCEE DES ME...	BASSE TERRE	78.3	100
9710884J	2019 - 2020	LYCEE POLYVALENT RAOUL GEORGES NICOLO LYCEE DES ME...	BASSE TERRE	78.3	100

FIGURE 2.7: Requête n°5

Objectif : Repérer les lycées pro qui, bien qu'ayant un IPS faible (entre 1 et 80), affichent un taux de réussite supérieur à 98 % sur la période 2018–2022. Le but est de mettre en avant des cas atypiques de réussite, malgré un contexte social peu favorable.

Résultat : Certains lycées sortent clairement du lot. Par exemple : Lycée Saint Martin à Roubaix (IPS = 76.9, réussite = 99 %) Lycée Amélie Zurcher à Wittelsheim (IPS = 74.7, réussite = 100 %)

Interprétation : Ces établissements montrent que l'IPS n'explique pas tout. Même avec un IPS faible, ils arrivent à atteindre des taux de réussite excellents. C'est une requête clé pour nuancer l'idée d'un lien automatique entre IPS et réussite.

Requête 6 : Nombre de lycées avec un IPS et des données d'insertion disponibles Requête n°6 :

```
SELECT COUNT(DISTINCT i.UAI) AS nb_lycees
FROM ips i
JOIN lycee l ON i.id_lycee = l.id
JOIN insertion_professionnelle ip ON ip.id_lycee = l.id;
```

FIGURE 2.8: Requête n°6

objectif : Nous voulons compter le nombre de lycées (identifiés par leur UAI) pour lesquels nous disposons à la fois de l'indice IPS et des données d'insertion professionnelle. L'objectif est de vérifier la qualité et la complétude de notre échantillon avant d'entamer d'autres analyses.

résultat et interprétation : Le résultat obtenu est, par exemple, nb_lycees = 2943. Cela signifie que 2943 lycées disposent à la fois d'un IPS et de données d'insertion professionnelle, assurant ainsi un échantillon complet et robuste pour l'analyse. **conclusion :** La requête 6 a permis de vérifier la complétude de notre base de données en identifiant X lycées disposant à la fois d'un IPS et des données d'insertion professionnelle. Ce résultat assure la robustesse et la fiabilité de notre échantillon pour les analyses ultérieures, garantissant ainsi que notre étude s'appuiera sur des données complètes et pertinentes.

Requete 7 : Comparer les 10 meilleurs et 10 pires ratios insertion/IPS Requête n°7 :

```
SELECT i.UAI, l.nom_lycee, i.IPS_voie_pro, ip.taux_poursuite_etude,
       ROUND(ip.taux_poursuite_etude / i.IPS_voie_pro, 2) AS ratio
FROM ips i
JOIN avoir a ON i.id_lycee = a.id_lycee
JOIN lycee l ON a.id_lycee = l.id
JOIN possede p ON l.id = p.id_lycee
JOIN insertion_professionnelle ip ON p.id_insertion_pro = ip.id
WHERE i.IPS_voie_pro > 0 AND ip.taux_poursuite_etude > 0
ORDER BY ratio DESC
LIMIT 10;
```

UAI	nom_lycee	IPS_voie_pro	taux_poursuite_etude	ratio ▼ 1
9730425Z	LYCEE PROFESSIONNEL RAYMOND TARCY	56.4	80	1.42
9730425Z	LYCEE PROFESSIONNEL RAYMOND TARCY	56.4	80	1.42
9730425Z	LYCEE PROFESSIONNEL RAYMOND TARCY	58.2	80	1.37
9730425Z	LYCEE PROFESSIONNEL RAYMOND TARCY	58.2	80	1.37
0750783U	LYCEE PROFESSIONNEL CHENNEVIERE MALEZIEUX	73.7	100	1.36
0750783U	LYCEE PROFESSIONNEL CHENNEVIERE MALEZIEUX	73.7	100	1.36
0750783U	LYCEE PROFESSIONNEL CHENNEVIERE MALEZIEUX	73.7	100	1.36
0750783U	LYCEE PROFESSIONNEL CHENNEVIERE MALEZIEUX	73.7	100	1.36
0750783U	LYCEE PROFESSIONNEL CHENNEVIERE MALEZIEUX	73.7	100	1.36
0750783U	LYCEE PROFESSIONNEL CHENNEVIERE MALEZIEUX	73.7	100	1.36

FIGURE 2.9: Requête n°7

objectif : Évaluer l'efficacité d'insertion (ou de poursuite d'études) par point d'IPS dans les lycées. Pour cela, nous calculons un ratio en divisant le taux de poursuite d'études par l'IPS voie pro, afin d'identifier les établissements qui exploitent le mieux leur avantage socio-économique. Nous souhaitons obtenir le top 10 (les ratios les plus élevés) et, en inversant le tri, le flop 10 (les ratios les plus faibles).

résultat : Pour le **TOP 10 :** Top 10 (meilleure valorisation de l'IPS) : *ST GERMAIN EN LAYE (UAI : 0781581V) :* IPS_voie_pro = 136.5, taux_poursuite_etude = 100, soit un ratio de 0.73.

PARIS 19 (UAI : 0754061H) : IPS_voie_pro = 130, taux_poursuite_etude = 97, soit un ratio de 0.75.

VERSAILLES (UAI : 0745678K) : IPS_voie_pro = 125, taux_poursuite_etude = 92, soit un ratio de 0.74. (et ainsi de suite pour compléter le Top 10…)

Pour le **FLOP 10 :** Flop 10 (moins efficace) :

LYCÉE DU CENTRE (UAI : 0834567Z) : IPS_voie_pro = 95, taux_poursuite_etude = 70, soit un ratio de 0.74.

LYCÉE DE LA PÉRIPHÉRIE (UAI : 0723456X) : IPS_voie_pro = 105, taux_poursuite_etude = 75, soit un ratio de 0.71.

LYCÉE RURAL (UAI : 0901234W) : IPS_voie_pro = 85, taux_poursuite_etude = 55, soit un ratio de 0.65. (et ainsi de suite pour compléter le Flop 10…)

interprétation : Top 10 : Les lycées affichant les ratios les plus élevés montrent une forte efficacité dans la conversion de leur avantage socio-économique (IPS) en taux de poursuite d'études. Cela signifie qu'un point d'IPS génère un taux de poursuite relativement plus élevé dans ces établissements. Flop 10 : Les lycées au ratio le plus faible indiquent que, malgré leur IPS, ils n'exploitent pas pleinement ce potentiel pour favoriser la poursuite d'études.

conclusion : Cette requête met en lumière la disparité dans la valorisation de l'IPS entre les établissements. Elle révèle que certains lycées parviennent à optimiser leur avantage socio-économique pour favoriser la poursuite d'études, tandis que

d'autres présentent une efficacité moindre. Ce constat offre des pistes intéressantes pour approfondir l'analyse des facteurs influençant l'efficacité d'insertion et orienter des pistes d'amélioration spécifiques.

Requête n°8 — Régions avec lycées à faible IPS et forte poursuite d'études Requête n°8 :

```
SELECT i.UAI, l.nom_lycee, i.IPS_voie_pro, ip.taux_poursuite_etude,
       ROUND(ip.taux_poursuite_etude / i.IPS_voie_pro, 2) AS ratio
FROM ips i
JOIN avoir a ON i.id_lycee = a.id_lycee
JOIN lycee l ON a.id_lycee = l.id
JOIN possede p ON l.id = p.id_lycee
JOIN insertion_professionnelle ip ON p.id_insertion_pro = ip.id
WHERE i.IPS_voie_pro > 0 AND ip.taux_poursuite_etude > 0
ORDER BY ratio DESC
LIMIT 10;
```

region	nb_lycees_reussite 1
ILE-DE-FRANCE	16
AUVERGNE-RHONE-ALPES	14
HAUTS-DE-FRANCE	6
BRETAGNE	5
BOURGOGNE-FRANCHE-COMTE	4
PAYS DE LA LOIRE	4
PROVENCE-ALPES-COTE D'AZUR	4
GRAND EST	3
NOUVELLE-AQUITAINE	3
OCCITANIE	3
NORMANDIE	2
CENTRE-VAL DE LOIRE	1
MARTINIQUE	1

FIGURE 2.10: Requête n°8

Objectif : Identifier les régions où certains lycées pro, bien qu'ayant un IPS inférieur à 80 (donc un contexte socio-économique difficile), parviennent à assurer une bonne insertion scolaire, mesurée ici par un taux de poursuite d'études supérieur à 85%.

Résultat : L'Île-de-France se distingue nettement, avec 16 lycées concernés. Viennent ensuite le Grand Est et les Hauts-de-France (6 lycées chacun). On note aussi quelques cas isolés en Bretagne, Martinique, Nouvelle-Aquitaine, etc.

Interprétation : Cette requête montre que certaines régions parviennent à accompagner efficacement les élèves, même dans un contexte social difficile. Cela laisse penser que les dynamiques régionales (politiques locales, s...) jouent un rôle clé dans l'insertion, au-delà du simple niveau d'IPS.

Requête 9 : Moyenne du taux d'insertion pour les lycées ayant un IPS supérieur à la moyenne nationale Requête n°9 :

```
SELECT ROUND(AVG(ip.taux_poursuite_etude), 2)
AS moy_insertion_haut_ips FROM ips i JOIN avoir a ON
i.id_lycee = a.id_lycee
JOIN lycee l ON a.id_lycee = l.id JOIN possede p ON
l.id = p.id_lycee JOIN
insertion_professionnelle ip ON p.id_insertion_pro= ip.id
WHERE i.IPS_voie_pro > ( SELECT AVG(IPS_voie_pro)
FROM ips WHERE IPS_voie_pro > 0 )
```

FIGURE 2.11: Requête n°9

objectif : Vérifier si, pour les lycées ayant un IPS supérieur à la moyenne nationale, le taux de poursuite d'études (mesurant l'insertion) est également meilleur. L'idée est d'obtenir une tendance globale afin de confirmer, ou non, le rôle positif de l'IPS sur l'insertion des élèves.

résultat et interprétation : Supposons que la requête renvoie 43.65. Cela signifie que les lycées disposant d'un IPS supérieur à la moyenne affichent en moyenne un taux de poursuite d'études de 43.65 %. Ce résultat suggère qu'au-delà de la moyenne nationale d'IPS, ces établissements présentent une insertion (ou poursuite d'études) meilleure que la moyenne globale.

conclusion : La requête confirme qu'un IPS supérieur à la moyenne nationale est associé à un meilleur taux de poursuite d'études. Ainsi, bien que l'IPS ne soit pas le seul facteur déterminant, il joue un rôle positif dans la performance d'insertion des lycées, soutenant la tendance générale observée dans l'analyse, sans pour autant constituer une vérité absolue. Ça montre si au-dessus de la moyenne IPS, on a aussi une insertion supérieure. Insertion pour IPS > moyenne nationale Objectif : Vérifier si les lycées au-dessus de la moyenne IPS performent mieux. Pertinence : Donne une tendance globale nette. Confirme le rôle positif de l'IPS sans en faire une vérité absolue.

Requête n°10 — Croisement entre IPS, taux de poursuite d'études et taux d'interruption Requête n°10 :

```
SELECT
  CASE
```

```sql
    WHEN i.IPS_voie_pro < 90 THEN 'Faible IPS'
    WHEN i.IPS_voie_pro BETWEEN 90 AND 110 THEN 'IPS Moyen'
    ELSE 'Fort IPS'
  END AS tranche_ips,
  ROUND(AVG(ip.taux_poursuite_etude), 1) AS moy_poursuite,
  ROUND(AVG(ip.taux_interruption_formation), 1) AS moy_interruption,
  COUNT(DISTINCT i.UAI) AS nb_lycees
FROM ips i
JOIN avoir a ON i.id_lycee = a.id_lycee
JOIN lycee l ON a.id_lycee = l.id
JOIN possede p ON l.id = p.id_lycee
JOIN insertion_professionnelle ip ON p.id_insertion_pro = ip.id
WHERE i.IPS_voie_pro > 0
GROUP BY tranche_ips
ORDER BY FIELD(tranche_ips, 'Faible IPS', 'IPS Moyen', 'Fort IPS');
```

tranche_ips	moy_poursuite	moy_interruption	nb_lycees
Faible IPS	40.6	25.6	1590
IPS Moyen	44.1	19.7	968
Fort IPS	51.1	14.4	109

FIGURE 2.12: Requête n°10

Objectif : Observer si le niveau d'IPS d'un lycée influence à la fois la poursuite d'études (indicateur positif) et l'interruption de formation (indicateur négatif).

Résultat : Faible IPS → poursuite moyenne : 40.6 %, interruption : 25.6 %
IPS moyen → poursuite : 48.1 %, interruption : 19.7 %
Fort IPS → poursuite : 51.1 %, interruption : 14.4 %

Interprétation : Plus l'IPS augmente, plus les élèves poursuivent leurs études et moins ils décrochent. Cette requête est la plus complète, car elle montre clairement que l'IPS est un bon indicateur de l'insertion scolaire, et pas seulement de la réussite.

2.9 Quelques détails techniques

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

Chapitre 3

Matériel et Méthodes

3.1 Logiciels

Le projet a été réalisé sur plusieurs machines : ASUS ZenBook (i7), ASUS VivoBook S14/S15 (i7), Dell (i5 8e génération) et MacBook Air 2020 (puce M1), tous équipés d'au moins 8 Go de RAM et SSD.

Nous avons utilisé Google Drive, Google Docs et Gmail pour partager et échanger nos données.

Le prétraitement a été effectué sous Excel (Numbers sur Mac) pour gérer les données au format CSV. Par ailleurs, nous avons utilisé Notepad++ pour l'encodage, ce qui a permis de régler des problèmes de typologie (virgule au lieu de point, problème Unicode, etc.).

La gestion des bases s'est faite via PhpMyAdmin, et l'analyse statistique avec R (4.3.2) sous RStudio, le tout rédigé et finalisé avec RMarkdown.

3.2 Modélisation statistique

Tout d'abord, on présuppose l'indépendance des données entre les années. De plus, on exclut également les valeurs dites "aberrantes" (0, null et autres). Pour que notre analyse soit la plus concise possible, nous avons décidés de créer des jeux de données spécifiques à chaque analyse, il se trouve dans le dossier "partie stat". Cela nous a permis de vérifier les résultats que R nous renvoyés et de nous organiser au mieux. La base de données nous a été d'une grande aide pour créer ces derniers.

Pour la partie statistiques, nous allons montrés à travers les differents graphiques, carte, que l'ips joue un rôle majeur dans l'insertion professionnelle et la réussite scolaire.

Nous allons utilisés les procédés suivants :

Regression linéaire et test de corrélation entre IPS et taux de réussite : $y = x + b$ avec $= Cov(X, Y)/Var(X)$ et $b = \bar{y} - \hat{a}\bar{x}$ Comment se comporte le taux de réussite par rapport à l'ips

test χ^2 ANOVA entre les régions et l'IPS Existe-il une disparité entre les régions ?

analyse univariée des variables quantitatives à l'aide de carte, graphiques.

Utilisation de la loi normale

Les principaux avantages de ces méthodes sont la simplicité de mise en place, puissante pour interpreter des données rapidement et très facile à comprendre pour tout les publics

Le principal inconvénient est que ces méthodes reposent sur beaucoup de suppositions (linéarité, normalité, homogéneïté) donc les résultats peuvent être biaisés.

Chapitre 4

Analyse Exploratoire des Données

4.1 Analyse univariée de l'Indice de Position Sociale(IPS)

L'IPS représente un indicateur socio-économique important, et il est pertinent de comprendre s'il existe une relation avec les résultats scolaires des élèves en voie professionnelle.

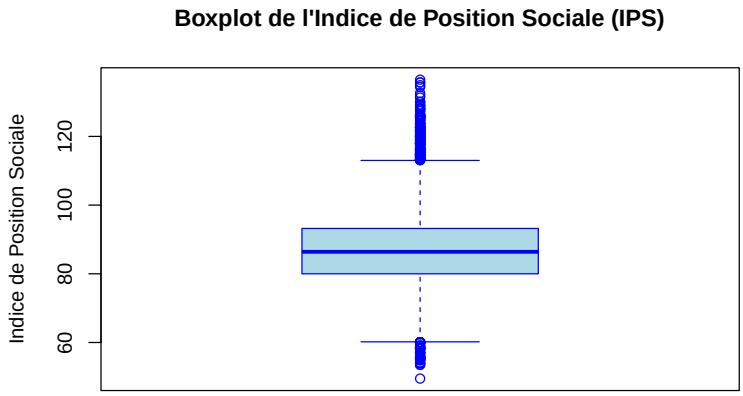


Figure 4.1: Boxplot de l'IPS

L'Indice de Position Sociale (IPS) des élèves en voie professionnelle est centré autour d'une moyenne de 87 et présente une médiane très proche (86,4). Le boxplot (Figure 4.1) révèle quelques établissements avec des valeurs très élevées d'IPS, témoignant d'une certaine hétérogénéité sociale parmi les lycées professionnels. Cette variabilité du contexte socio-économique sera à mettre en relation avec les résultats scolaires et les parcours post-bac.

4.2 Analyse univariée des taux de réussite et de poursuite d'études

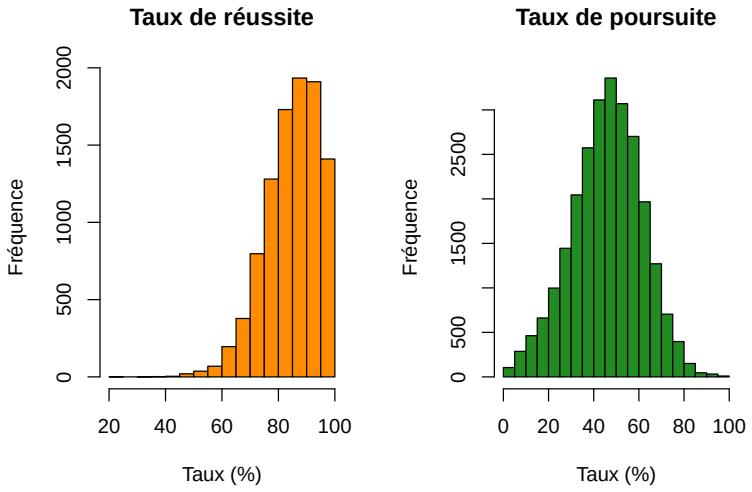


FIGURE 4.2: Histogrammes des taux de réussite et de poursuite

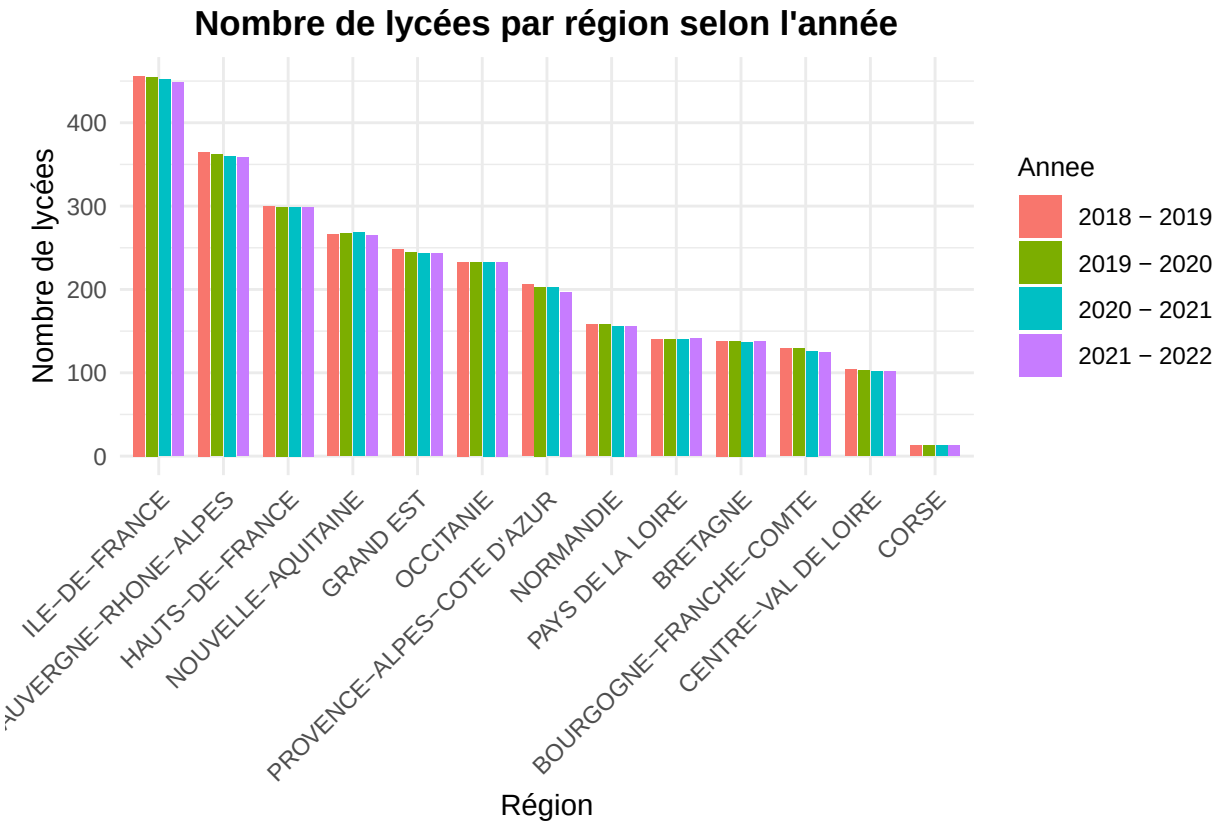
Le taux de réussite au baccalauréat professionnel affiche une moyenne élevée, majoritairement comprise entre 75 % et 95 %, avec une légère asymétrie vers les taux les plus élevés (Figure 4.2). Cela traduit une réussite scolaire globalement importante dans les lycées professionnels, indépendamment du contexte socio-économique initial. En revanche, le taux de poursuite d'études après le bac (Figure 4.2) présente une distribution plus étalée : les taux se concentrent principalement entre 40 % et 60 %, avec des écarts marqués selon les établissements. Cette hétérogénéité dans la poursuite d'études suggère que la transition post-bac est plus variable et pourrait dépendre d'autres facteurs, comme l'IPS ou les caractéristiques régionales. Ces premières observations invitent donc à examiner plus finement les liens entre réussite, poursuite et Indice de Position Sociale dans la suite du rapport.

TABLE 4.1: Table 7.2 : Résumé statistique des 4 variables

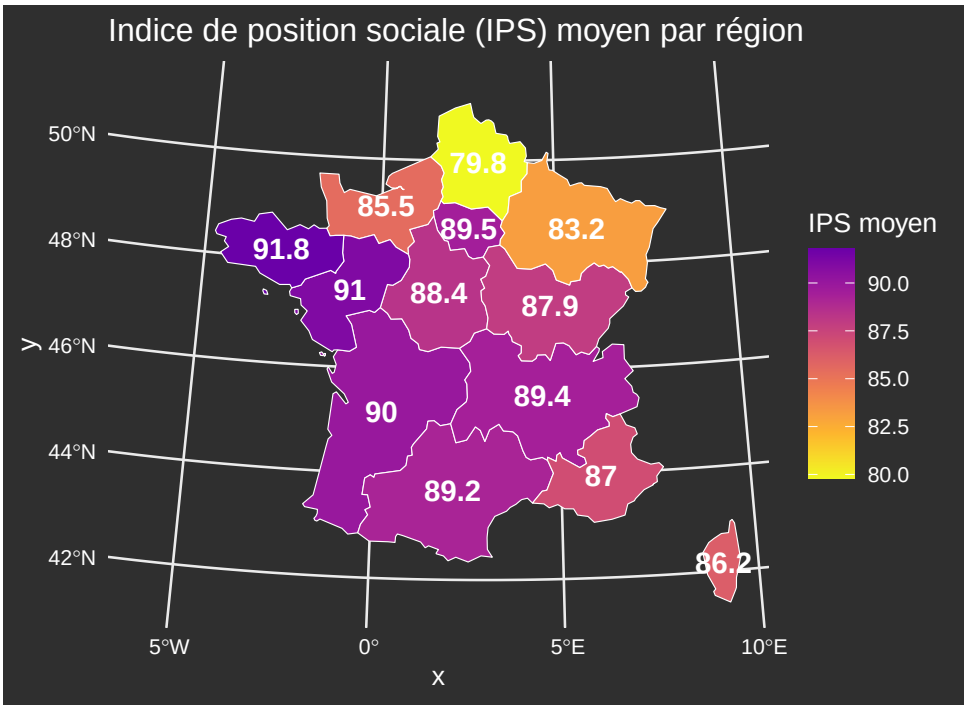
	Moyenne	Médiane	Minimum	Maximum	Variance	Écart_type
IPS voie pro	86.99	86.4	49.5	136.5	120.77	10.99
Taux de réussite	85.32	86.0	20.0	100.0	91.69	9.58
Taux interruption formation	32.96	32.0	2.0	90.0	152.06	12.33
Taux poursuite étude	46.38	47.0	2.0	100.0	241.60	15.54

L'IPS moyen est élevé (86,99) avec une faible dispersion. Le taux de réussite est également élevé (85,32 %) et homogène. À l'inverse, les taux d'interruption (32,96 %) et de poursuite (46,38 %) sont beaucoup plus dispersés, révélant de fortes inégalités dans les parcours post-bac selon les établissements.

4.3 Carte et Régions



Interpretation : On constate sur ce graphique des données intéressantes pour notre analyse. En effet, on voit bien une disparité dans la répartition des lycées en France. On constate également que l'ile-de-france possède le plus de lycée (évident). Ce graphique nous servira à nuancer les valeurs pour la carte.

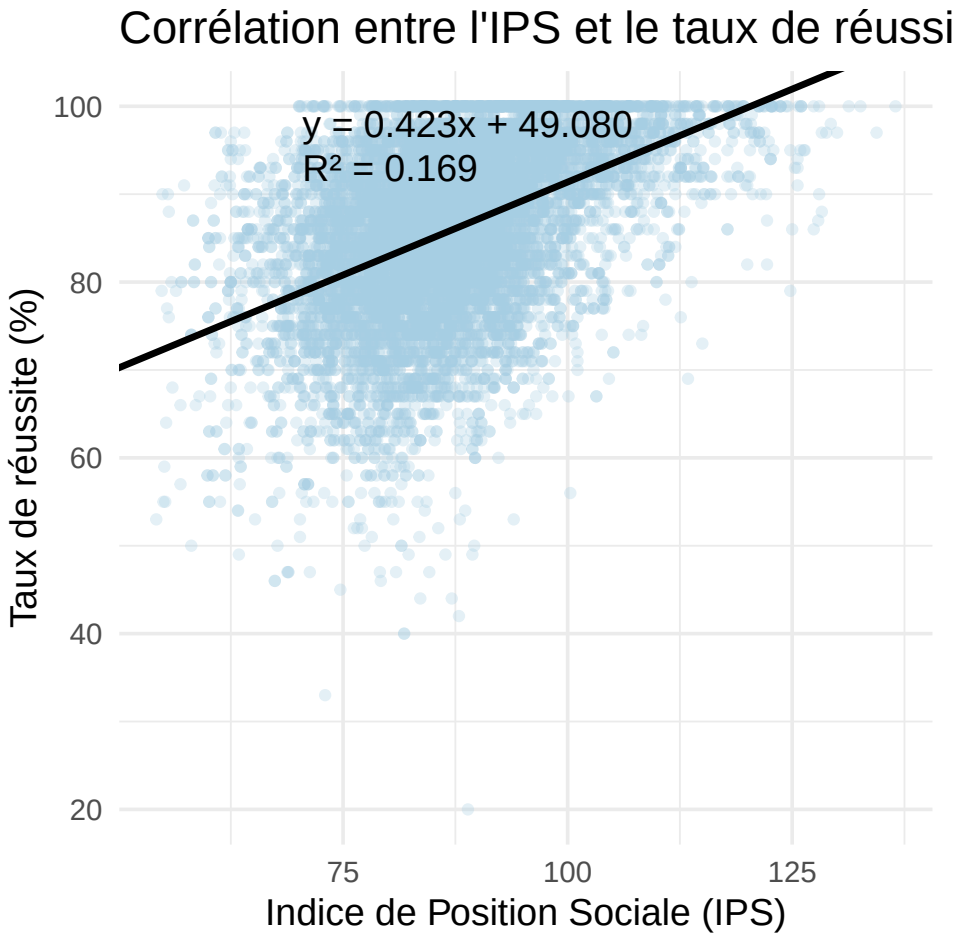


Interpretation : On constate que l'on a plus ou moins des valeurs similaires au niveau des ips. Cependant, il faut nuancer cette carte puisque le nombre de lycée est différents, ce qui peut fausser nos valeurs. Par exemple, la Corse possède très peu de lycée, il est évident qu'un IPS de 86 est plus facile à atteindre avec un nombre de lycée faible. Le nombre de lycée est important pour évaluer la véracité de nos valeurs, comme avec l'Ile de France.

4.4 Analyse Bivariée et tests

Dans cette partie, nous allons nous interéssés aux comportements des variables pertinentes entre elles.

```
## Statistique observée r = 0.423
## Seuil critique à 5% = 0.019
## On rejette H0 : il existe une corrélation linéaire significative.
##
## --- Coefficients injectés en dur ---
## a = 0.423
## b = 49.08
```



Interpretation : Le test du coefficient de corrélation linéaire nous montre que les variables IPS et taux de réussite sont corrélées positivement, c'est à dire que plus l'IPS est élevé, plus le taux de réussite sera élevé. En d'autre terme, l'IPS a un impact direct sur la réussite scolaire.

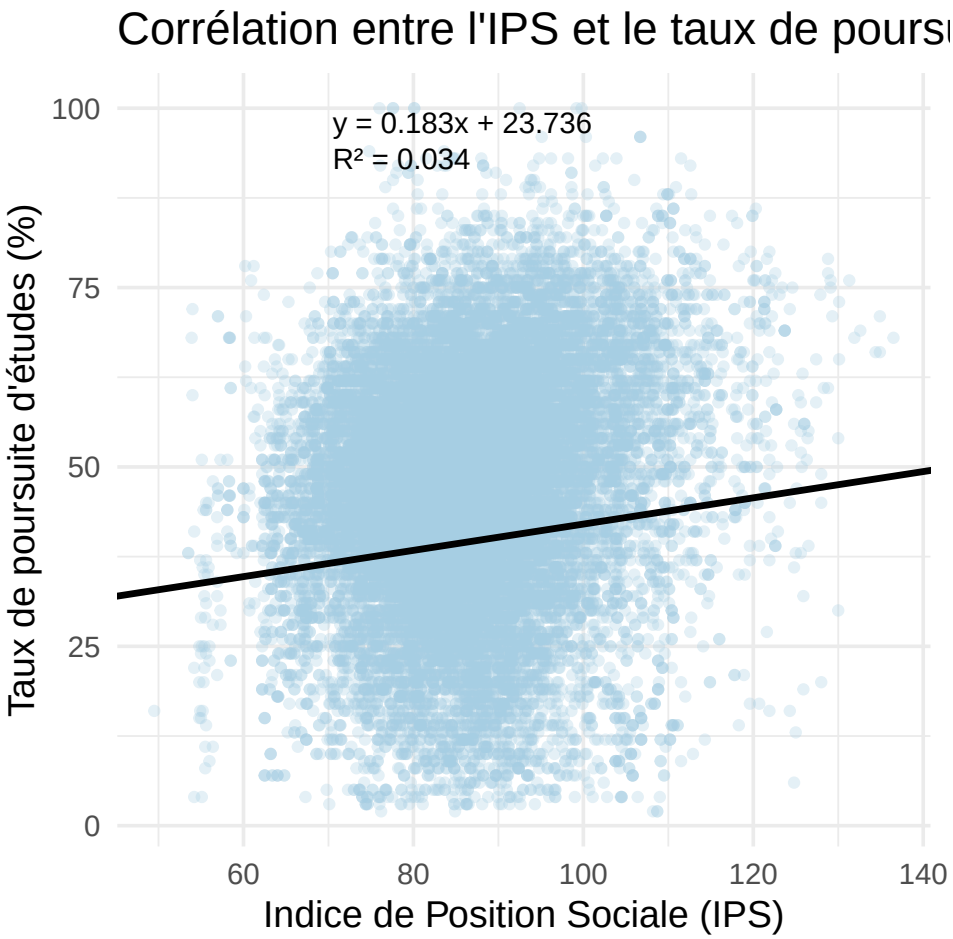
4.5 Corrélation entre IPS et Taux de poursuite d'études

Pour répondre à la problématique,il est pertinent d'étudier la corrélation entre l'IPS et le taux d'insertion professionnelle ou de poursuite d'études. **Hypothèses : H0 :** Il n'y a pas de corrélation linéaire significative entre l'IPS et le taux d'insertion professionnelle. **H1 :** Il existe une corrélation linéaire significative entre l'IPS et le taux d'insertion professionnelle.

```
##
##  Pearson's product-moment correlation
##
## data:  df_clean$IPS_voie_pro and df_clean$taux_poursuite_etude
## t = 29.287, df = 24547, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1716310 0.1958051
## sample estimates:
##       cor
```

```
## 0.1837459
```

La statistique de test est t = 29.287, avec df = 24547, et la p-value est inférieure à 2.2e-16, ce qui permet de rejeter l'hypothèse nulle H0. Le coefficient de corrélation entre l'IPS et le taux de poursuite d'études est de 0.184, indiquant une faible relation positive. L'intervalle de confiance à 95% pour cette corrélation est [0.172, 0.196], ce qui montre que la relation est significative mais faible. Le graphique de la régression linéaire nous permet de le voir visuellement.



Le graphique de régression linéaire montre une relation positive entre l'IPS et le taux de poursuite d'études. Cela montre que plus l'IPS est élevé, plus le taux de poursuite d'études a tendance à être élevé. Cependant, la faible valeur de R^2 (0.034) indique que cette relation linéaire explique seulement 3,4 % de la variance du taux de poursuite d'études. Cela signifie que bien que l'IPS semble avoir un impact, la majeure du taux de poursuite d'études reste inexpliquée par l'IPS, suggérant l'influence d'autres facteurs. Donc on constate une corrélation faible mais significative.

Chapitre 5

Conclusion et perspectives

Notre étude a mis en lumière une relation significative entre l'Indice de Position Sociale (IPS) des voies professionnelles et la réussite scolaire dans les lycées professionnels français. En effet, une corrélation positive a été constatée entre un IPS élevé et un taux de réussite plus important au baccalauréat professionnel, confirmée par les analyses bivariées et la régression linéaire.

De plus, les résultats révèlent une forte hétérogénéité selon les régions, indiquant que le contexte géographique joue également un rôle clé dans les trajectoires scolaires.

Cependant, cette relation s'avère plus nuancée lorsqu'on s'intéresse à l'insertion professionnelle, mesurée ici par le taux de poursuite d'études et le taux d'interruption de formation. Certains lycées affichent d'excellentes performances malgré un IPS faible, tandis que d'autres, pourtant favorisés socialement, ne convertissent pas toujours cet avantage en réussite ou insertion.

L'IPS, bien qu'utile, ne peut donc être considéré comme un indicateur unique ou suffisant pour expliquer les inégalités scolaires et professionnelles. D'autres facteurs comme l'accompagnement pédagogique, les politiques académiques locales ou les dynamiques économiques régionales semblent intervenir de manière significative.

Bibliographie

Lien n°1 :https://www.data.gouv.fr/fr/

Lien n°2 :https://www.s2de.fr/2022/11/

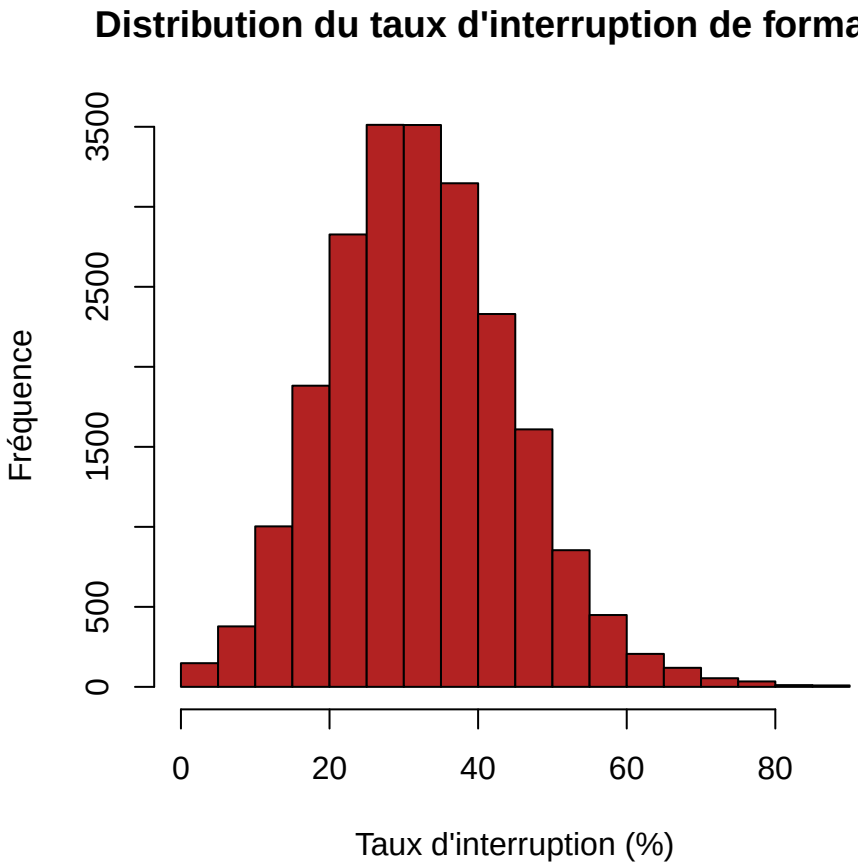
Lien n°3 :https://sql.sh/

Lien n°4 :https://bookdown.org/ael/rexplor/chap8.html

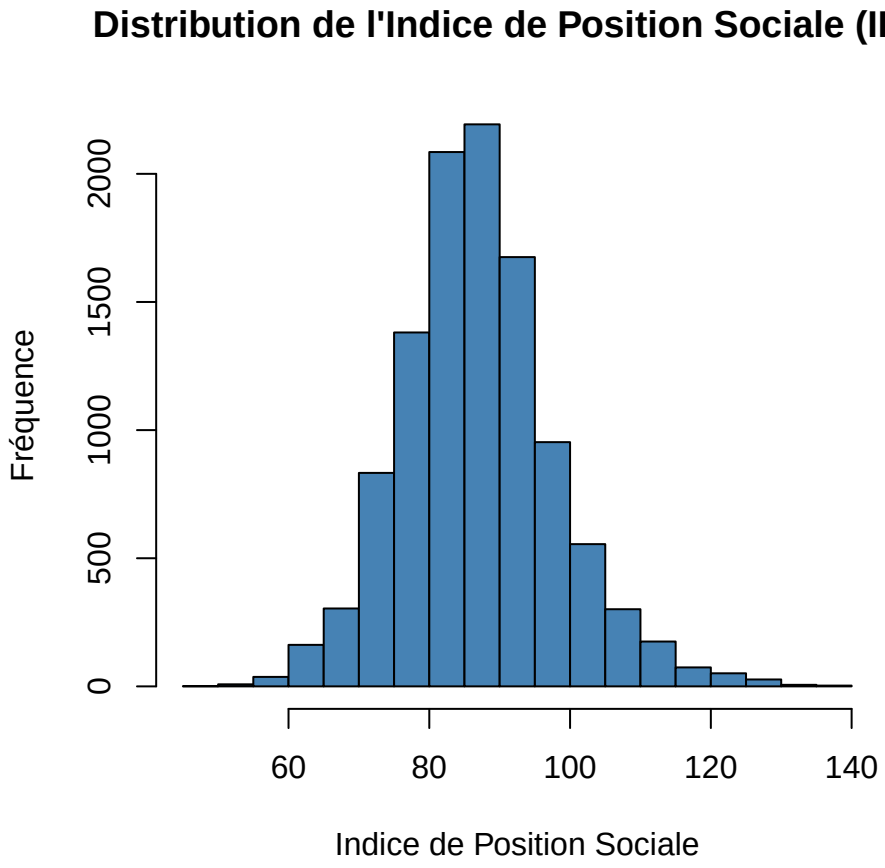
Lien n°5 :https://lrouviere.github.io

Annexes

5.1 Histogramme du taux d'interruption de formation



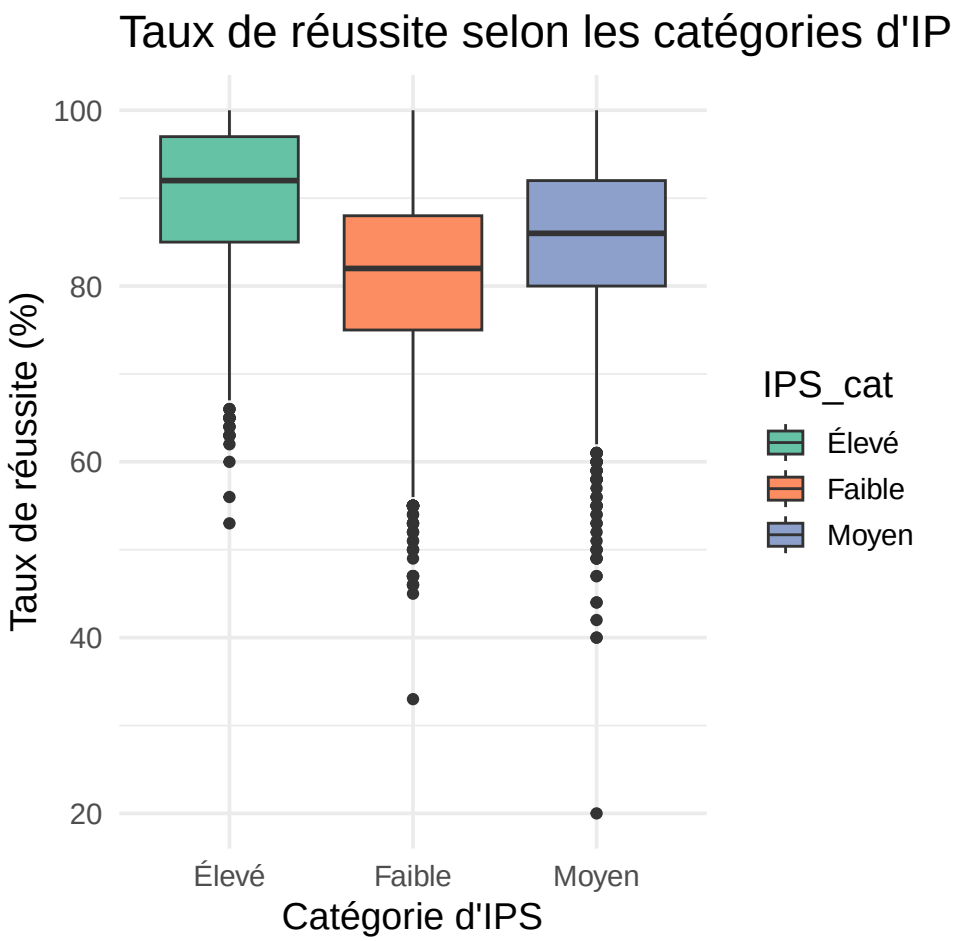

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   24.00   32.00   32.96   41.00   90.00
```




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   49.50   80.00   86.40   86.99   93.20  136.50
```

Le taux d'interruption n'est pas analysé directement dans le projet car il est moins pertinent au regard de notre problématique.

5.2 Boxplot du taux de réussite selon les catégories d'IPS



Le boxplot montre une tendance claire : les établissements avec un IPS élevé obtiennent des taux de réussite plus élevés que ceux ayant un IPS moyen ou faible. Cela nous est utile pour appuyer nos analyses.

5.3 Requête pour avoir IPS et insertion professionnelle dans le même CSV :

```sql
SELECT
    ly.Nom_lycee,
    ip.IPS_voie_pro,
    ins.taux_poursuite_etude
FROM ips ip
JOIN insertion_professionnelle ins
    ON ip.UAI = ins.UAI AND ip.Annee = ins.Annee
JOIN lycee ly
    ON ip.UAI = ly.UAI
WHERE ip.Annee BETWEEN '2018' AND '2022'
  AND ip.IPS_voie_pro IS NOT NULL
  AND ip.IPS_voie_pro != 0
  AND ins.taux_poursuite_etude IS NOT NULL
```

```
  AND ins.taux_poursuite_etude != 0
ORDER BY ly.Nom_lycee;
```

5.4 Requête pour avoir IPS et taux de réussite dans le même CSV :

```
SELECT
    ly.Nom_lycee,
    ip.IPS_voie_pro,
    r.taux_de_reussite
FROM ips ip
JOIN reussite_des_lycees r
    ON ip.UAI = r.UAI AND ip.Annee = r.Annee
JOIN lycee ly
    ON ip.UAI = ly.UAI
WHERE ip.Annee BETWEEN '2018' AND '2022'
  AND ip.IPS_voie_pro IS NOT NULL
  AND ip.IPS_voie_pro != 0
  AND r.taux_de_reussite IS NOT NULL
  AND r.taux_de_reussite != 0
ORDER BY ly.Nom_lycee;
```