PROJET RÉALISÉ PAR L'ÉQUIPE GROUPE 1 - TD 1

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

Malcom CARLET (22106930), Aya MOHAMEDATNI (22106289), Hugo PEREIRA GONÇALVES (22109797), Lucas TRIOZON (22101060)

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.

Lucas Triozon remercie aussi Massimo, Lilian, Noah et Adrien pour leur aide mentale dans cette période délicate de blocage.

27/04/2023.

Résumé

La France possède 13 régions administratives, avec chacune ses spécificités tant culturelles qu'économiques. Elles hébergent toutes des étudiants venant de tout le pays, dans les universités et les écoles en vue d'obtenir un diplôme et entrer sur le marché du travail. Nous voulons déterminer dans quelle région est-on le mieux formé en France. Notre étude concerne les diplômes de Master et de Licence Pro. L'objectif a été de regarder des indicateurs statistiques sur l'emploi et l'insertion des formations et leur localisation, pour conclure quelle région apporte plus de réussite. On a fini par montrer que chacun de nos indicateurs étaient indépendants de la région concernée par des tests de l'ANOVA, et qu'il était donc impossible de déterminer un quelconque classement des régions françaises en termes de formation. Le facteur géographique n'est donc pas déterminant, et le regard sur les disciplines ou les types de diplôme seraient plus appropriés.

Table des matières

Chapitre 1

Introduction

Dans un monde du travail devenant de plus en plus complexe, avec une multitude de formations proposant divers tremplins, il est logique de se demander lesquelles offrent le plus de débouchés à leurs diplômés. Au-delà de l'analyse pour chaque intitulé de formation ou de discipline, on peut s'intéresser à un aspect géographique de l'enseignement supérieur en France.

Dans quelles régions les étudiants sont les mieux formés, entre 2012 et 2019 ?

Nous allons baser notre recherche sur deux jeux de données principaux, proposés par le ministère de l'enseignement supérieur et de la recherche, sur leur plateforme opendata :
https://data.enseignementsup-recherche.gouv.fr/explore/dataset/fr-esr-insertion_professionnelle-lp/information/
https://data.enseignementsup-recherche.gouv.fr/explore/dataset/fr-esr-insertion_professionnelle-master/information/

Ce sont les résultats d'une enquête sur l'insertion professionnelle des étudiants après le master et la licence pro, pour chaque formation de chaque établissement, entre 2012 et 2019. Les données récoltées comprennent le type d'emploi obtenu, les salaires et le chômage.

En complément, nous allons utiliser un autre jeu de données disponible sur la même plateforme, qui contient tous les principaux diplômes et formations préparés dans les établissements publics sous tutelle du ministère en charge de l'Enseignement supérieur :
https://data.enseignementsup-recherche.gouv.fr/explore/dataset/fr-esr-principaux-dipl

Nous sommes aujourd'hui en deuxième année de licence MIASHS, et nous commençons à nous poser des questions sur la poursuite de nos études. Savoir quelle région de France offre le plus de réussite après le diplôme est une donnée importante, qui peut orienter les choix de master de n'importe qui, et c'est pour cette raison que cette question nous intéresse.

Chapitre 2

Base de données

2.1 Descriptif des tables

Parmi les jeux de données proposés initialement figurait une table par type de diplôme : Master, Licence Pro, DUT et Doctorat. Pour aller dans le sens de notre question, nous avons dû écarter les DUT car les données étaient nationales, c'est à dire non détaillées par établissement, et les Doctorats car l'enquête ne concernait que les diplômés de l'année 2014.

Il nous reste donc deux tables avec le bilan de l'insertion professionnelle pour les formations de type Master et Licence Pro en France. La similarité des tables nous permet de les fusionner en une seule table, puisqu'elles possèdent les mêmes colonnes. Cela nous fait un CSV de 28035 lignes à traiter, que nous allons filtrer et basculer dans une base de données SQL.

Afin de pouvoir construire une analyse non perturbée par des valeurs manquantes, nous avons décidé de filtrer au préalable sur Excel puis ensuite sur MySQL en supprimant les lignes contenant des cases vides ou assimilés à la valeur « non-définie ». Après ce prétraitement, sur les 28035 lignes à traiter, il nous en reste 9665.

Nous avons aussi filtré nos colonnes de cette façon :

- Pour garder une certaine cohérence vis-à-vis de notre sujet, nous avons décidé de garder celles décrivant l'année, l'académie ainsi que l'établissement pour obtenir les informations spatio-temporelles nécessaires à notre problématique.
- Nous avons ensuite le type de diplôme : Master ou Licence Pro, le domaine et la discipline pour conserver les informations concernant la formation.
- Pour finir, nous avons préservé les colonnes concernant les statistiques économiques de l'emploi à la sortie des formations : le pourcentage d'emplois cadres, de cadres ou professions intermédiaires, d'emplois stables, et d'emplois à temps plein. Afin de pouvoir davantage analyser la qualité de l'emploi obtenu, nous avons jugé important de conserver deux colonnes concernant la rémunération de ces emplois : salaire brut annuel estimé et salaire net médian des emplois à temps plein. Pour finir, nous avons gardé la variable du taux d'insertion, statistique primordiale pour savoir si une formation offre une perspective d'emploi une fois le diplome obtenu.

Nom colonne	Type	Description
Année	Entier	Année de la fin de formation
Académie	Varchar(16)	Nom de l'Académie de la formation
Etablissement	Varchar(46)	Nom de l'établissement
Diplôme	Varchar(11)	Nom du diplôme obtenu à la fin de la formation
Discipline	Varchar(57)	Nom de la discipline du diplôme
Domaine	Varchar(31)	Nom du domaine du diplôme
% emplois cadre	Entier	Part des emplois cadre parmi tous les emplois obtenus
% emplois cadre ou professions intermédiaires	Entier	Part des emplois ou professions intermédiaires cadre parmi tous les emplois obtenus
% emplois stable	Entier	Part des emplois stables parmi tous les emplois obtenus
% emplois à temps plein	Entier	Part des emplois à temps plein parmi tous les emplois obtenus
Salaire brut annuel estimé	Entier	Valeur du Salaire brut annuel estimé
Salaire Net médian des emplois à temps plein	Entier	Valeur du Salaire net médian des emplois à temps plein
Taux d'insertion	Entier	Taux d'insertion

Table : Insertion Professionnelle (9665×12)

Nos colonnes ne contiennent pas de valeurs manquantes et chacune des valeurs est considérée comme non-uniques.

Avec nos données restantes, nous allons fragmenter notre table en plus petites parties, comme détaillé dans notre MCD qui suit.

2.2 Modèles MCD et MOD

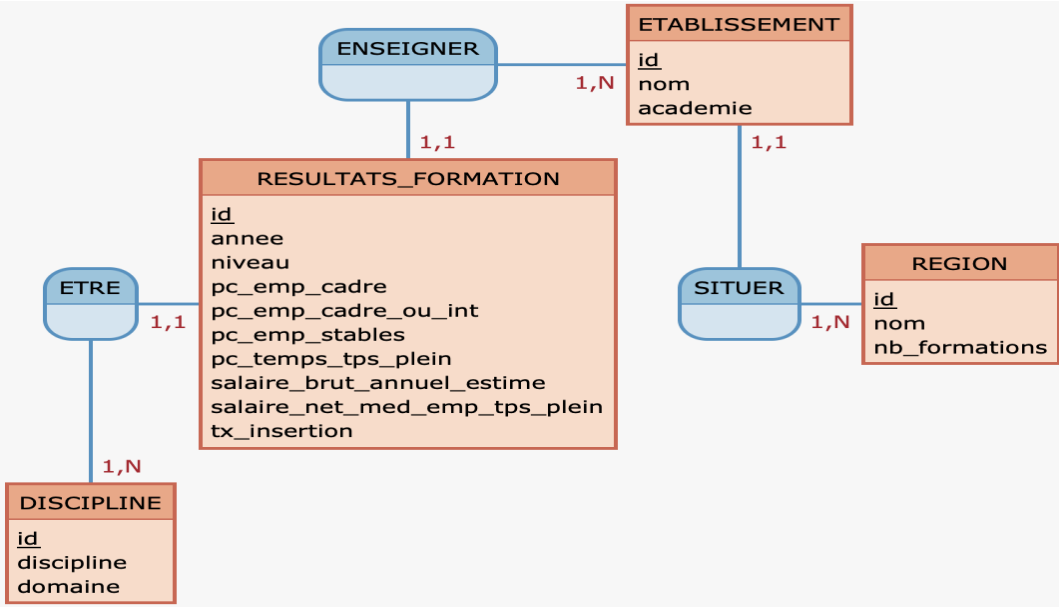


Figure 2.1: Modèle MCD

2.3 Requêtes réalisées

- Nombre d'établissements par region:

```
SELECT region.nom AS Nom, COUNT(*) AS "Nombre"
FROM etablissement, region
WHERE region.id = etablissement.id_region AND region.id != 1
GROUP BY region.id;
```

Table 2.1: Nombre d'établissements par région

Nom	Nombre
Ile-de-France	21
Hauts-de-France	10
Normandie	3
Bretagne	4
Grand-Est	4
Pays de la Loire	3
Centre-Val de Loire	2
Bourgogne-Franche-Comté	2
Nouvelle-Aquitaine	6
Auvergne-Rhône-Alpes	12
Occitanie	8
Provence-Alpes-Côte d'Azur	5
Corse	1

- Nombre d'enregistrements dans la table `resultats_formation` par année :

```
SELECT Annee, COUNT(*) AS 'Nombre'
FROM resultats_formation
GROUP BY Annee
ORDER BY Annee ASC
```

Table 2.2: Nombre d'enregistrements par année

Annee	Nombre
2012	412
2013	586
2014	610
2015	679
2016	679
2017	680
2018	640
2019	668

- Nombre d'enregistrements par année, par région

```
SELECT Annee, id_region, region.nom, COUNT(resultats_formation.id)
AS 'NbEntre'
FROM resultats_formation, etablissement, region
WHERE resultats_formation.id_etablissement = etablissement.id
AND etablissement.id_region = region.id
GROUP BY id_region, Annee
ORDER BY Annee, id_region;
```

Table 2.3: Nombre d'enregistrements par année et par région (10 premiers résultats)

Annee	id_region	nom	NbEntre
2012	1	National	17
2012	2	Ile-de-France	85
2012	3	Hauts-de-France	39
2012	4	Normandie	15
2012	5	Bretagne	25
2012	6	Grand-Est	35
2012	7	Pays de la Loire	18
2012	8	Centre-Val de Loire	12
2012	9	Bourgogne-Franche-Comté	14
2012	10	Nouvelle-Aquitaine	36

- Moyenne du taux d'insertion dans les formations par région et par année

```
SELECT Annee, id_region, region.nom,
AVG(resultats_formation.tx_insertion)AS 'Moyenne'
FROM resultats_formation, etablissement, region
WHERE resultats_formation.id_etablissement = etablissement.id
AND etablissement.id_region = region.id
GROUP BY id_region, Annee
ORDER BY Annee, AVG(resultats_formation.tx_insertion) DESC
```

Table 2.4: Moyenne du taux d'insertion dans les formations par région et par année (10 premiers résultats)

Annee	id_region	nom	Moyenne
2012	3	Hauts-de-France	91.7949
2012	2	Ile-de-France	91.2235
2012	6	Grand-Est	91.0857
2012	8	Centre-Val de Loire	90.9167
2012	7	Pays de la Loire	90.4444
2012	11	Auvergne-Rhône-Alpes	90.3333
2012	5	Bretagne	90.2000
2012	10	Nouvelle-Aquitaine	90.1111
2012	1	National	89.4118
2012	12	Occitanie	89.3939

- Moyenne emplois stables à la sortie des formations par région et par année

```
SELECT Annee, id_region, region.nom,
AVG(pc_emp_stables) AS 'Moyenne'
FROM resultats_formation, etablissement, region
WHERE resultats_formation.id_etablissement = etablissement.id
AND etablissement.id_region = region.id
```

```
GROUP BY id_region, Annee
ORDER BY Annee, AVG(resultats_formation.pc_emp_stables) DESC
```

Table 2.5: Moyenne emplois stables à la sortie des formations par région et par année (10 premiers résultats)

Annee	id_region	nom	Moyenne
2012	3	Hauts-de-France	78.8974
2012	7	Pays de la Loire	75.2778
2012	9	Bourgogne-Franche-Comté	75.1429
2012	12	Occitanie	74.7273
2012	2	Ile-de-France	74.5647
2012	13	Provence-Alpes-Côte d'Azur	74.1724
2012	11	Auvergne-Rhône-Alpes	74.1667
2012	10	Nouvelle-Aquitaine	73.9444
2012	4	Normandie	72.9333
2012	6	Grand-Est	72.4286

- Moyenne salaire net médian pour les emplois à temps plein par région et par année

```
SELECT Annee, id_region, region.nom,
AVG(salaire_net_med_emp_tps_plein) AS Moyenne
FROM resultats_formation, etablissement, region
WHERE resultats_formation.id_etablissement = etablissement.id
AND etablissement.id_region = region.id
GROUP BY id_region, Annee
ORDER BY Annee, Moyenne DESC
```

Table 2.6: Moyenne salaire net médian pour les emplois à temps plein par région et par année (10 premiers résultats)

Annee	id_region	nom	Moyenne
2012	2	Ile-de-France	2034.235
2012	3	Hauts-de-France	1871.282
2012	6	Grand-Est	1854.571
2012	1	National	1854.118
2012	13	Provence-Alpes-Côte d'Azur	1853.448
2012	8	Centre-Val de Loire	1842.500
2012	4	Normandie	1830.667
2012	12	Occitanie	1822.121
2012	11	Auvergne-Rhône-Alpes	1821.111
2012	10	Nouvelle-Aquitaine	1801.667

Chapitre 3

Matériel et Méthodes

3.1 Logiciels

Pour traiter les données, les lire et les manipuler :

- MySQL5.7.40 | 8.0.31(Wickham et al. 2019)[1]
- MySQL Workbench 8.0.31 (Wickham et al. 2019)[2]
- RStudio 4.2.2(Wickham et al. 2019)[3]
- R 4.3(Wickham et al. 2019)[4]
- Microsoft Excel (Wickham et al. 2019)[5]

Pour collaborer:

- filess.io (Wickham et al. 2019)[6]
- Discord (Wickham et al. 2019)[7]
- Google Docs (Wickham et al. 2019)[8]
- Google Sheets (Wickham et al. 2019)[9]

Ordinateurs utilisées :

- Nom : LAPTOP-KUU4P8BA, Processeur : AMD Ryzen 3 3200U with Radeon, Vitesse du Processeur : 2.60 GHz, Mémoire RAM : 8Go, Système D'exploitation : Windows 10 64bits

[1] MySQL est un système de gestion de bases de données relationnelles. [En savoir plus](https://dev.mysql.com/doc/)

[2] MySQL Workbench est un outil visuel unifié pour les architectes de bases de données, les développeurs et les DBA. [En savoir plus](https://www.mysql.com/products/workbench/)

[3] RStudio est un environnement de développement gratuit, libre et multiplateforme pour R, un langage de programmation utilisé pour le traitement de données et l'analyse statistique. [En savoir plus](https://www.r-studio.com/fr/)

[4] R is 'GNU S', a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and non-linear modelling, statistical tests, time series analysis, classification, clustering, etc. [En savoir plus](https://www.r-project.org/)

[5] Microsoft Excel est un logiciel tableur de la suite bureautique Microsoft Office développé et distribué par l'éditeur Microsoft [En savoir plus](https://www.microsoft.com/fr-fr/microsoft-365/excel)

[6] Services de base de données basés sur des plans pour les développeurs. [En savoir plus](https://filess.io/)

[7] Discord est un logiciel propriétaire gratuit de VoIP et de messagerie instantanée. [En savoir plus](https://discord.com/)

[8] Google Docs est un traitement de texte en ligne inclus dans la suite Web gratuite Google Docs Editors proposée par Google. [En savoir plus](https://www.google.fr/intl/fr/docs/about/)

[9] Google Sheets est une application de feuille de calcul incluse dans la suite Web gratuite Google Docs Editors proposée par Google; [En savoir plus](https://www.google.fr/intl/fr/sheets/about/)

3.2 Description des Données

Les données sélectionnées ont été stockées dans une base de données MySQL, qui est hébergée en ligne à l'aide de filess.io, ainsi que localement sur les ordinateurs individuels. L'export de cette base de données pèse 343 Ko. Elle est composée de cinq tables différentes, résultant des jeux de données importés (cf. chapitre 2), et contient dix variables.

3.3 Nettoyage des données

Pour la Table "Insertion Profesionnelle" :

Afin de pouvoir construire une analyse non perturbée par des valeurs manquantes, nous avons décidé de filtrer au préalable sur Excel en supprimant les lignes contenant des cases vides ou assimilés à la valeur « non-définie ». Après ce prétraitement, sur les 28035 lignes à traiter, il nous en restait 9665.

3.4 Étapes de Pré-traitements

Une fois nos fichiers CSV comportant nos données nettoyés, nous les avons importé dans notre base de données SQL. A partir de requêtes SQL nous avons pu extraire les données nécessaire à la création et au remplissage de nos tables : récupération de toutes les régions, création des clés étrangères, sélection et séparation des colonnes.

3.5 Modélisation statistique

Notre modélisation statistique repose sur des tests de l'ANOVA : test d'égalité des moyenne afin d'écarter ou non des hypothèses d'indépendance entre deux variables. Nous avons choisi cette approche après observation de nos données, pour déterminer si la région est déterminante ou non dans la réussite.

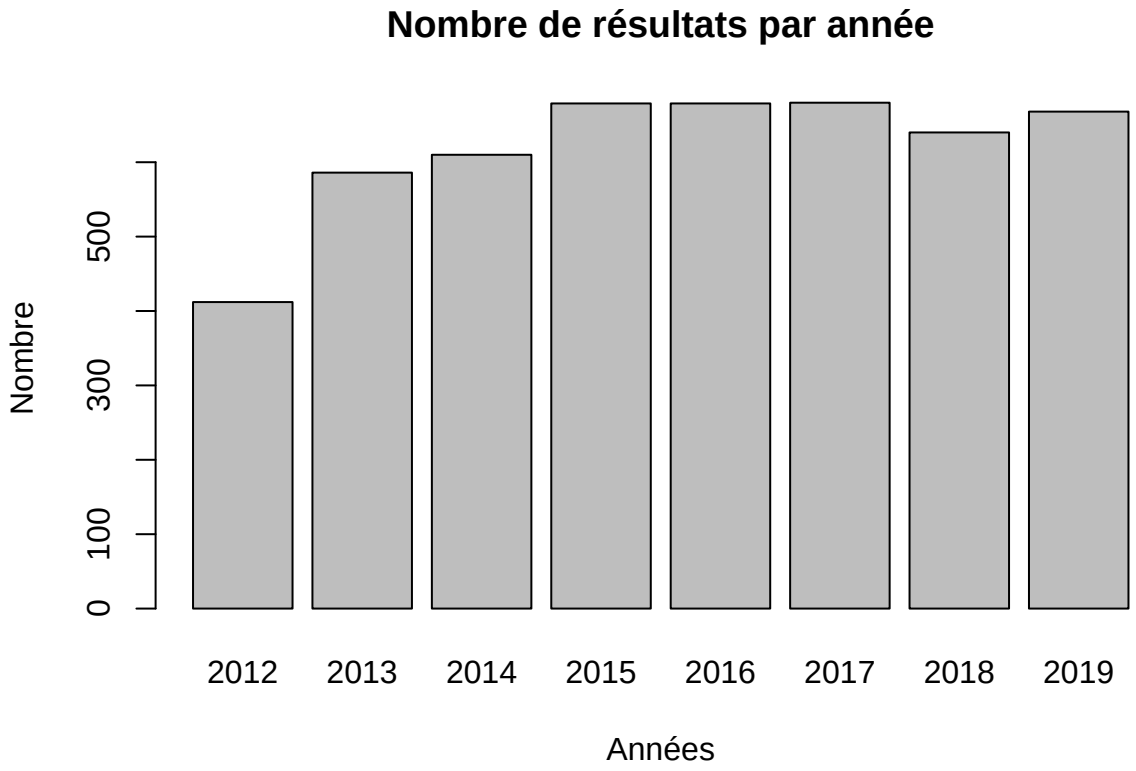
Chapitre 4

Analyse Exploratoire des Données

Pour commencer notre analyse statistique, nous allons regarder comment se comportent nos variables, à l'aide de requêtes SQL et de script R.

4.1 Années

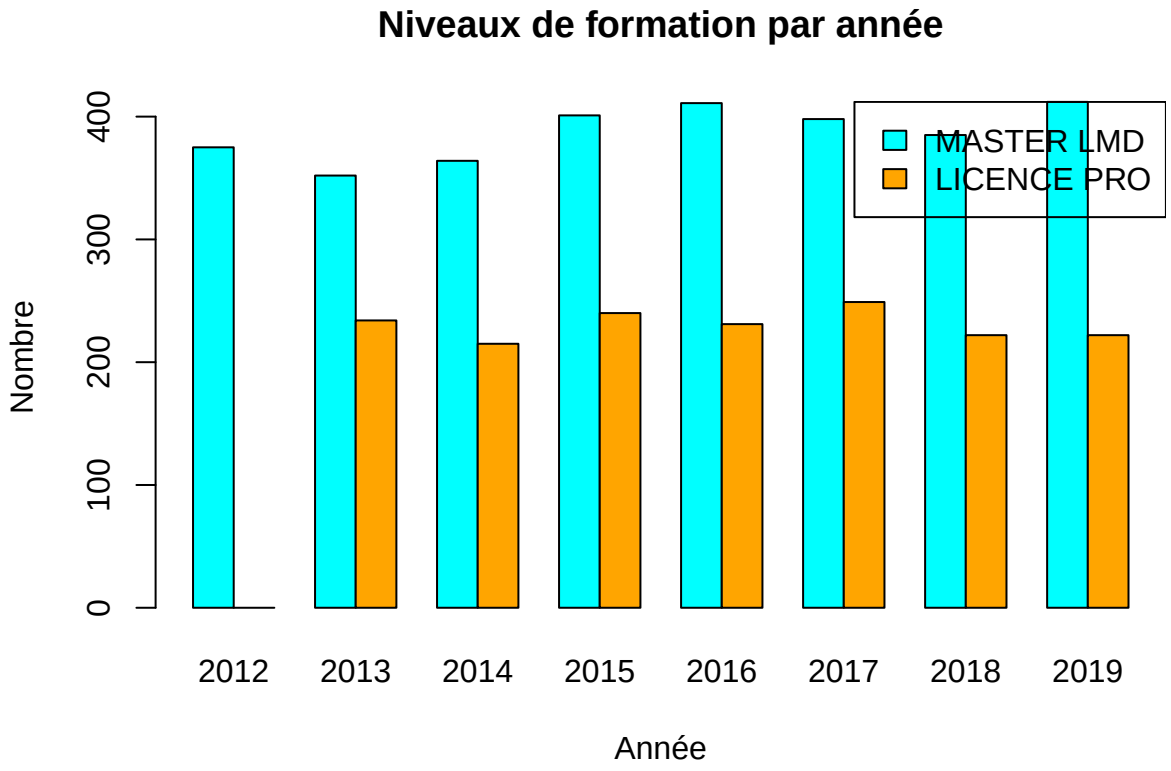
On regarde le nombre d'enregistrements que l'on a par année dans notre table resultats_formation.



On remarque que le nombre d'enregistrements est stable entre 2013 et 2019, mais inférieur de moitié en 2012.

4.2 Niveaux

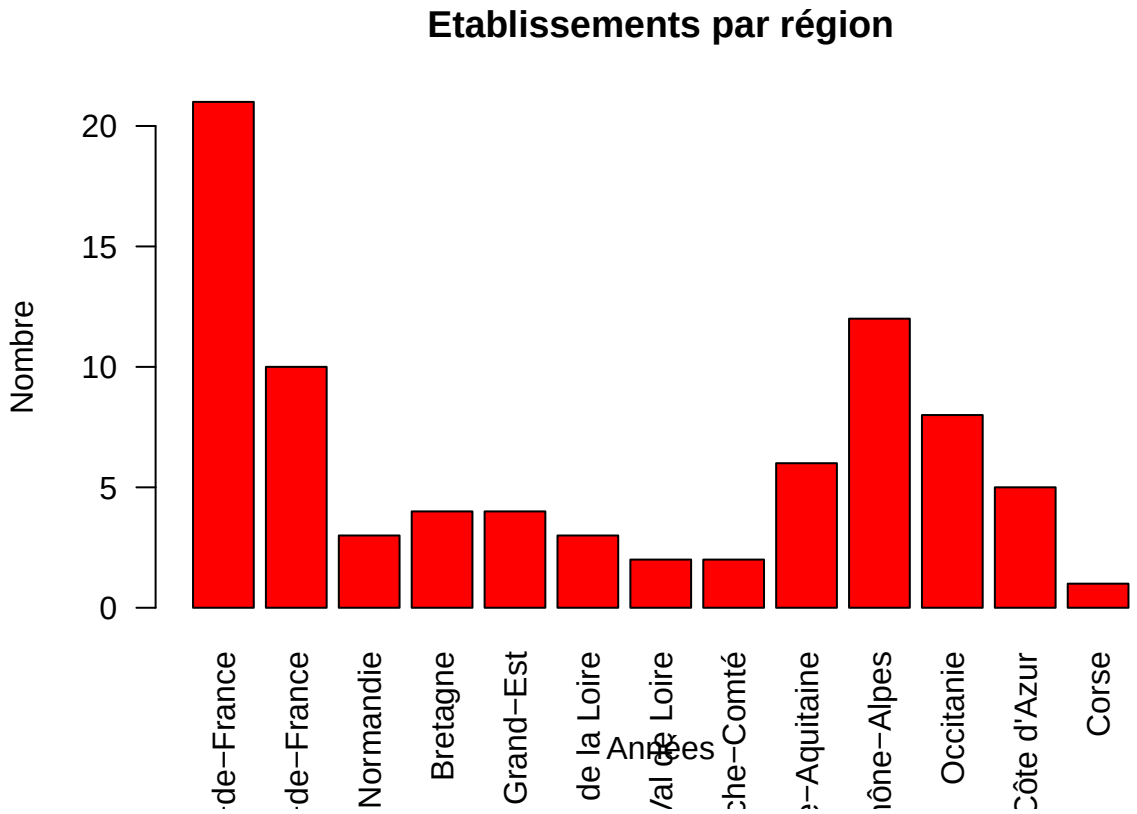
On cherche maintenant à distinguer dans ces enregistrements les niveaux, master et licence pro.



La différence remarquée en 2012 est due à l'absence de réponses pour les Licences Professionnelles au cours de cette période. Nous constatons qu'il y a globalement moins de réponses pour les Licences Professionnelles que pour les Masters.

4.3 Nombre d'établissements

Nous avons 13 régions et 81 établissements, mais on voudrait savoir comment ces derniers sont répartis dans la France.



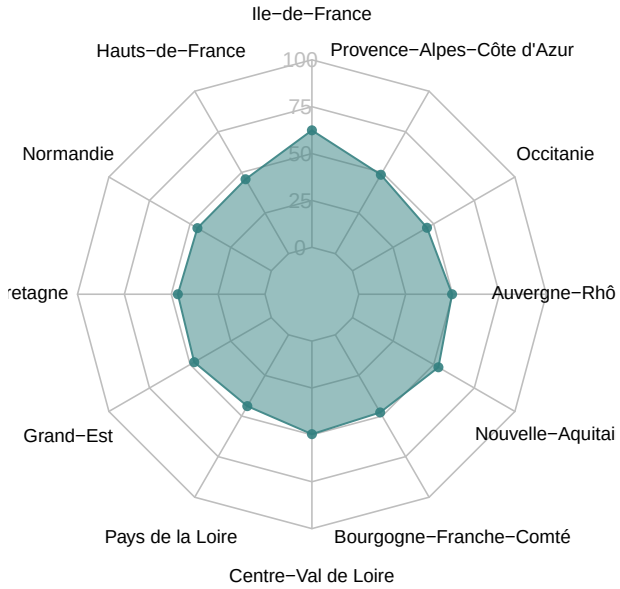
Ce qui est marquant est la supériorité de l'Ile-de-France, qui s'explique par le fait qu'elle est la région la plus riche de France, avec presque le double d'établissements par rapport à l'Auvergne-Rhone-Alpes, deuxième région la plus importante de notre pays.

Après avoir regardé les variables qui subdivisent notre jeu de données, on va analyser les variables qui vont apporter des réponses à notre question. Pour chaque variable, nous faisons la moyenne de chaque enregistrements par région, puis on analyse les disparités entre chaque.

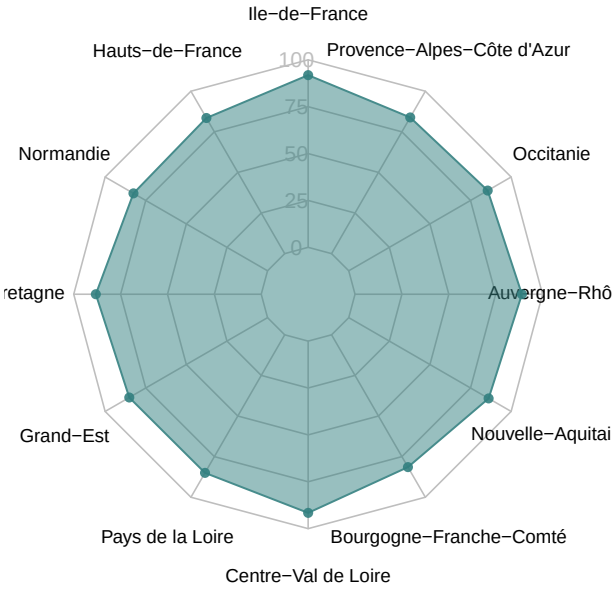
4.4 Taux d'emploi

Tout d'abord, on observe les pourcentages d'emploi : % emploi cadre, % emploi cadre ou intermédiaire, % emploi stable, % emploi à temps plein

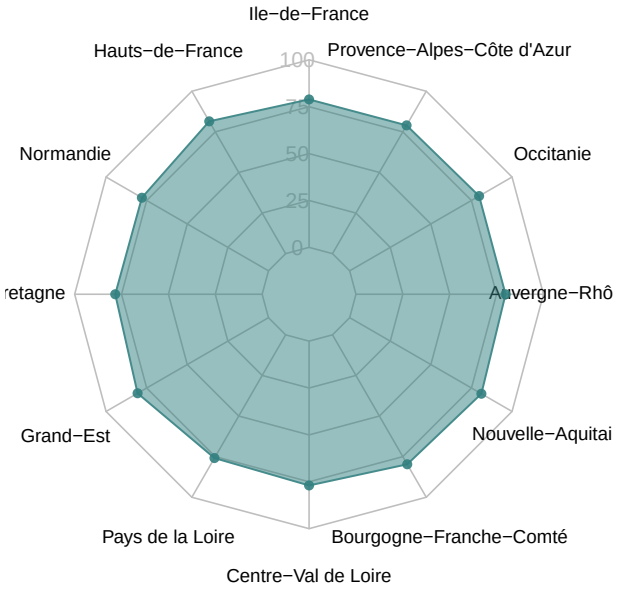
% emploi cadre



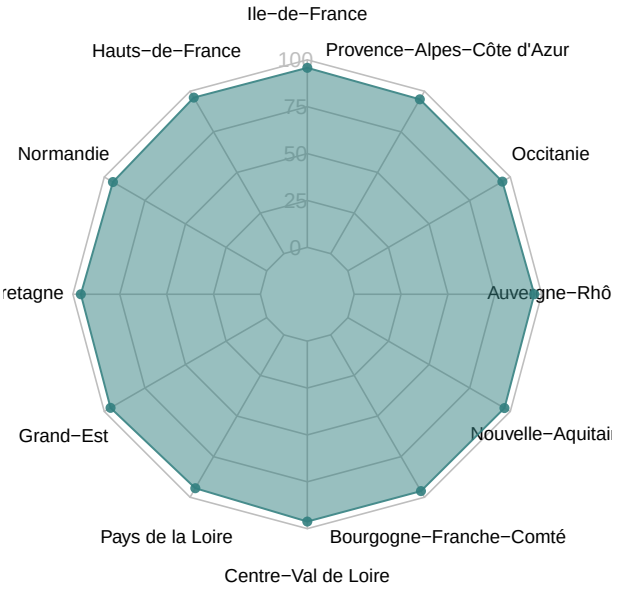
% emploi cadre ou intermediaire



% emploi stable



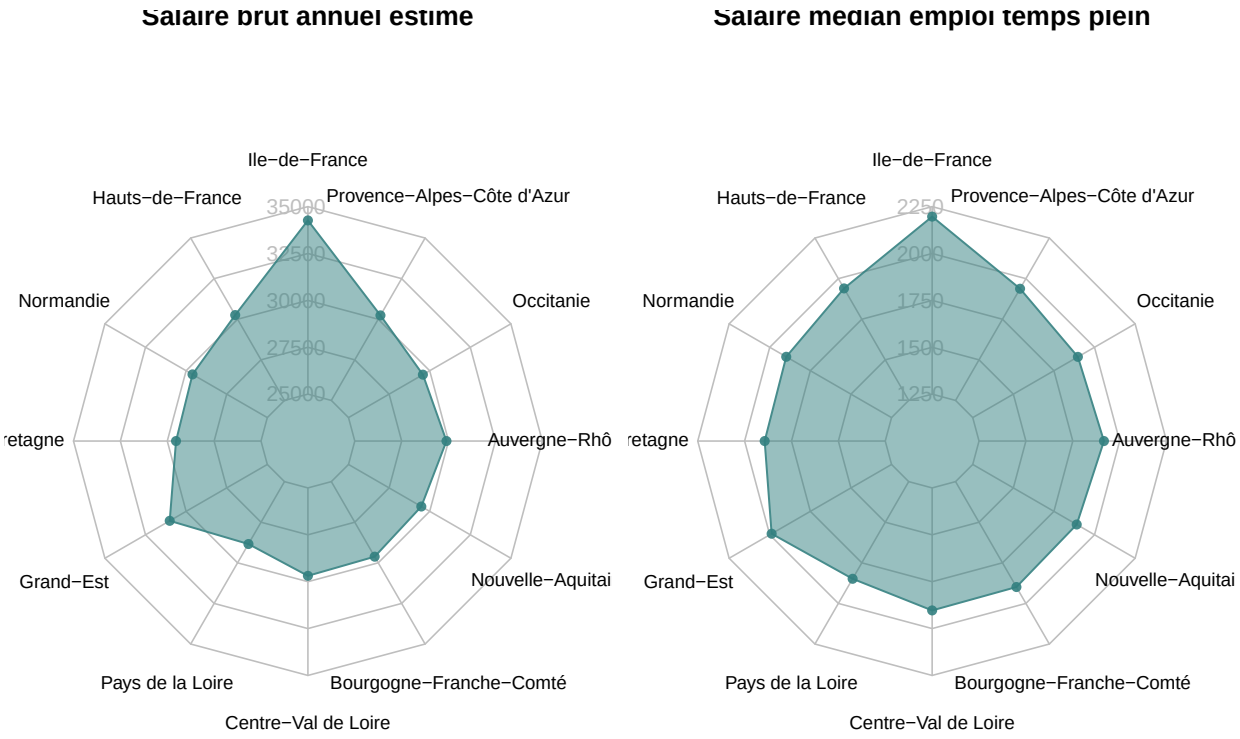
% emploi a temps plein



Nous obtenons quatre 'spider plots'. Nous constatons que les formations en Île-de-France conduisent à davantage d'emplois cadres, avec une différence de 10% par rapport à la Nouvelle-Aquitaine. Pour les autres variables, chaque région présente des résultats assez similaires : entre 94,4% et 96,2% pour les emplois à temps plein et entre 75,87% et 81,47% pour les emplois stables.

4.5 Salaires

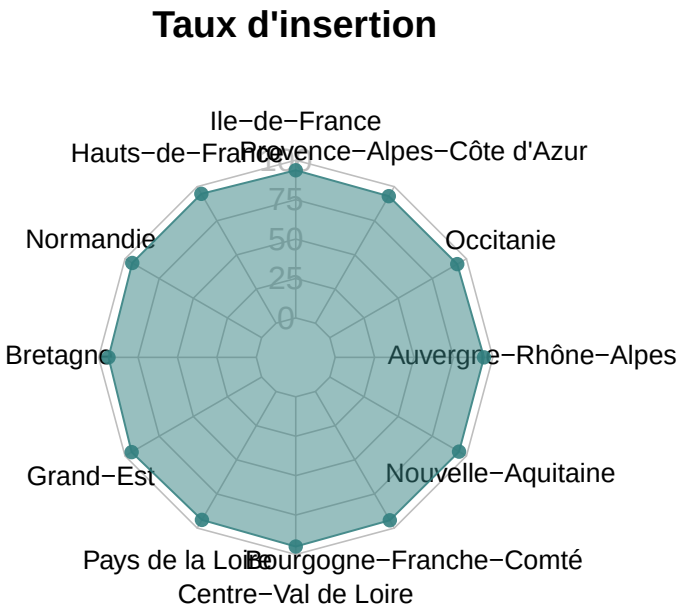
Ensuite, nous allons comparer les salaires entre régions.



En ce qui concerne les salaires, l'Ile-de-France est la région où les diplômés sont mieux rémunérés. Cette situation s'explique par le coût de la vie plus élevé dans cette région, mais indique également une tendance chez les diplômés à rester dans la région où ils ont été formés, quelle que soit cette région. Les autres régions présentent des niveaux de salaires équivalents, sans valeur aberrante comme celle de l'Ile-de-France, et qui suivent graduellement et de manière logique le niveau de richesse de chaque.

4.6 Taux d'insertion

Au final, nous observons les données sur les taux d'insertion.



Toutes les régions ont un taux d'insertion compris entre 90 et 93%, ce qui est une différence non significative.

Au vu de la similarité des résultats de chaque variable sur nos 13 régions, nous sommes en mesure de nous poser la question de l'indépendance entre nos résultats et les régions françaises. C'est ce que nous allons désormais essayer de chercher.

Chapitre 5

Analyse et Résultats

Dans le but de répondre à notre problématique, il est nécessaire d'établir une corrélation entre les différentes régions et le succès d'une formation. Ainsi, pour mener à bien cette étude, nous avons effectué des tests d'ANOVA (Analyse de Variance) sur les données de nos établissements en les regroupant selon leur région (à l'aide de la colonne "id_region"), en vue d'identifier une éventuelle dépendance. Pour réaliser ces tests d'ANOVA sur R, il convient d'utiliser la fonction "aov" (voir Annexe 1).

Indicateur	**Explication**
df	les degrés de liberté associés à chaque source de variation
Sum Sq	la somme des carrés associée à chaque source de variation
Mean Sq	la moyenne des carrés associée à chaque source de variation
F-value	la statistique F associée à chaque source de variation, qui mesure le rapport de la variance expliquée par cette source à la variance résiduelle
Pr(>F)	la valeur p associée à la statistique F, qui indique la significativité des différences entre les groupes ou l'effet d'un facteur

Résultats :

```
anova(data.2019)
```

```
## [1] "% emploi cadre"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1  12.02   12.02   0.477  0.505
## Residuals   10 251.80   25.18
## [1] "% emploi cadre ou intermédiaire"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1   2.67   2.669    0.22  0.649
## Residuals   10 121.28  12.128
## [1] "% emploi stable"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1  0.189  0.1894   0.064  0.806
## Residuals   10 29.819  2.9819
## [1] "% emploi à temps plein"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1  0.007  0.0070   0.014  0.907
## Residuals   10  4.847  0.4847
## [1] "Salaire médian emploi temps plein"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1  19408   19408   2.824  0.124
```

```
## Residuals   10  68734    6873
## [1] "Salaire brut annuel"
##             Df   Sum Sq Mean Sq F value Pr(>F)
## id_region    1  4721418 4721418    2.83  0.123
## Residuals   10 16682898 1668290
## [1] "Taux dinsertion"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1  0.685  0.6853    1.56   0.24
## Residuals   10  4.392  0.4392
```

Si la valeur de Pr(>F) est inférieure à $\alpha = 0{,}05$, nous pouvons rejeter l'hypothèse nulle de l'ANOVA selon laquelle toutes les moyennes conditionnelles sont égales et conclure qu'il existe une différence statistiquement significative entre les moyennes des trois groupes. Toutefois, nous remarquons que la valeur de Pr(>F) est largement supérieure à α pour toutes les variables. Par conséquent, nous ne pouvons pas rejeter l'hypothèse nulle de l'ANOVA, ce qui suggère que la variable Région n'affecte peut-être pas ces variables nécéssaires pour l'évaluation de la réussite d'une formation.

Chapitre 6

Discussion

Étant donné que les variables que nous avons choisies pour évaluer la qualité des formations sont toutes indépendantes de la région, nous pouvons conclure que les résultats d'une formation et donc sa qualité à former les étudiants ne dépendent pas de la région. Par conséquent, la problématique ne pourra être résolue car il ne serait pas pertinent de déterminer qu'une région forme mieux les étudiants qu'une autre, étant donné que cette relation n'est pas avérée. Toutefois, il est important de noter que seuls les diplômes universitaires tels que les Masters LMD et les Licences Professionnelles ont été étudiés. Les résultats auraient pu être différents si d'autres types de diplômes, tels que les grandes écoles (ingénieurs, commerce...) ou les doctorats, avaient été pris en compte. L'étude a également porté sur les résultats généraux des établissements et non pas par domaine, il se peut donc que certaines régions soient spécialisées dans certains domaines plus que d'autres.

Chapitre 7

Conclusion et perspectives

En résumé, l'analyse des diplômes de Master et de Licence Pro montre que la réussite des étudiants ne dépend pas de la région où ils étudient. Cependant, pour une analyse plus complète, il serait pertinent d'inclure d'autres types de diplômes certifiés par l'État, tels que les écoles privées et les écoles d'ingénieurs, et de les comparer entre eux. Une analyse plus fine de la réussite par discipline ou domaine pourrait également conduire à des résultats plus concluants. En outre, d'autres facteurs tels que le niveau socio-économique des étudiants, leur âge, leur sexe, leur expérience professionnelle et leur lieu de résidence pourraient être pris en compte pour comprendre les différences de réussite. Enfin, une comparaison des résultats avec d'autres pays pourrait fournir des informations utiles sur les performances des étudiants en France par rapport à celles des autres pays.

Il est important de noter que l'étude sur le long terme a été incluse en annexe de ce rapport car la conclusion est similaire à celle de l'étude sur l'année 2019. Cette étude longitudinale a permis de suivre la réussite des diplômés sur plusieurs années, et bien qu'elle ne soit pas incluse dans l'analyse principale, elle fournit des informations supplémentaires sur les performances des étudiants à long terme. Les résultats de l'analyse longitudinale suggèrent que la réussite des étudiants est relativement stable sur plusieurs années, ce qui renforce les conclusions de l'étude principale sur la non-dépendance de la réussite des étudiants à la région où ils étudient.

Tout au long de notre travail, différentes difficultés se sont présentées à nous:

- Grand nombre de valeurs manquantes
- Problème d'importation des données car fichier trop volumineux
- Problème de compatibilité des bases de données entre Windows / MacOS
- Tentative de récupération de données supplémentaires finalement avérée inutile (voir Annexe 2)

Bibliographie

WICKHAM, Hadley, AVERICK, Mara, BRYAN, Jennifer, CHANG, Winston, MCGOWAN, Lucy D'Agostino, FRANÇOIS, Romain, GROLEMUND, Garrett, HAYES, Alex, HENRY, Lionel, HESTER, Jim, KUHN, Max, PEDERSEN, Thomas Lin, MILLER, Evan, BACHE, Stephan Milton, MÜLLER, Kirill, OOMS, Jeroen, ROBINSON, David, SEIDEL, Dana Paige, SPINU, Vitalie, TAKAHASHI, Kohske, VAUGHAN, Davis, WILKE, Claus, WOO, Kara et YUTANI, Hiroaki, 2019. Welcome to the tidyverse. *Journal of Open Source Software*. 2019. Vol. 4, n° 43, pp. 1686. DOI 10.21105/joss.01686.

Annexes

Programme R de l'ANOVA

Exemple de l'utilisation de la fonction aov avec le jeu de données "iris":

```r
# Charge le jeu de données iris pour l'exemple
data(iris)

# Effectue le test de l'ANOVA sur la taille des sepal pour chaque espèce
# Sepal.Length est la variable numérique et Species indique les groupes
fit <- aov(Sepal.Length ~ Species, data = iris)

# Montre un résumé des résultats
summary(fit)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  63.21  31.606   119.3 <2e-16 ***
## Residuals   147  38.96   0.265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary(fit) retourne ainsi des dataframes avec 2 lignes: Species (correspond à la variable numérique y): les valeurs prédites pour chaque observation residuals: la différence entre les valeurs observées et les valeurs prédites

Mobilité des diplômés

Structure d'une feuille xlsx contenant les données sur les mobilités des diplômés de Master (semblable à la feuille correspondant aux Licences Pro):

```
library(openxlsx)
data.master <- openxlsx::read.xlsx('data-mobilites.xlsx', sheet = 1)
str(data.master)
```

```
## 'data.frame':    82 obs. of  20 variables:
##  $ membre                 : chr  "malcom" NA NA NA ...
##  $ id.etab                : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ année                  : chr  "2016-2017" "2018" "2019" NA ...
##  $ salaire                : chr  "1700" "1900" "1850" NA ...
##  $ %.emploi.IDF           : chr  NA "20.2" "19" NA ...
##  $ %.autre                : num  44.2 24.9 12 NA 1 0 NA NA 36.2 17.1 ...
##  $ %.hauts.de.france      : chr  NA NA "1" NA ...
##  $ %.normandie            : chr  NA NA "1" NA ...
##  $ %.bretagne             : chr  NA NA "3" NA ...
##  $ %.grand-est            : chr  NA NA "1" NA ...
##  $ %.pays.loire           : chr  NA NA "5" NA ...
##  $ %.centre               : chr  NA NA "5" NA ...
##  $ %.bourgogne.FC         : chr  NA NA "1" NA ...
##  $ %.nv-aquitaine         : num  NA NA 44 NA 2 2 NA NA NA NA ...
##  $ %.auvergne-rhone.alpes: chr  NA NA "4" NA ...
##  $ %.Occitanie            : chr  "52" NA "2" NA ...
##  $ %.PACA                 : chr  NA "54.9" "2" NA ...
##  $ %.Corse                : num  NA NA NA NA 0 0 NA NA NA NA ...
##  $ %.étranger             : chr  "3.8" NA NA NA ...
##  $ sources                : chr  "https://www.univ-jfc.fr/actu/linsertion-professionne
```

Ces données sont celles que nous avions manuellement entrées dans un fichier xls. Nous les avions récupérées dans l'objectif de calculer une espérance de salaire en sortant d'un établissement mais nous nous sommes rendus compte que notre calcul allait finalement nous ramener aux valeurs des salaires déjà présents dans notre base de données et nous n'aurions peut-être fait qu'erroner les valeurs.

Etude longitudinale

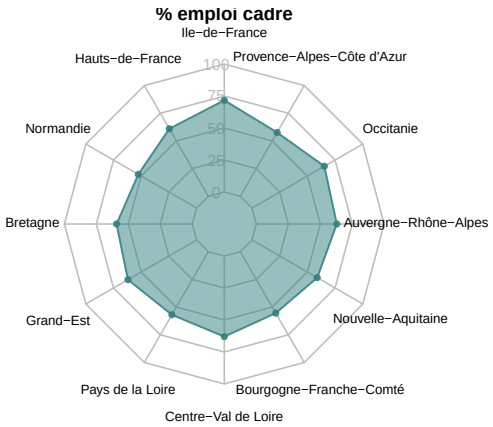
Nous sommes arrivée à la conclusion qu'il n'y avait pas de grandes différences entre l'année 2012 et 2019 au niveau des variables statistiques tel que le pourcentage d'employé cadre, ou encore le taux d'insertion professionnel. Pour visualiser cela, nous vous présentons en annexes les 'spiderplot' des années 2012, 2015 et 2019 qui comparent visuellement les performances des différentes régions de France pour chaque variable statistique.

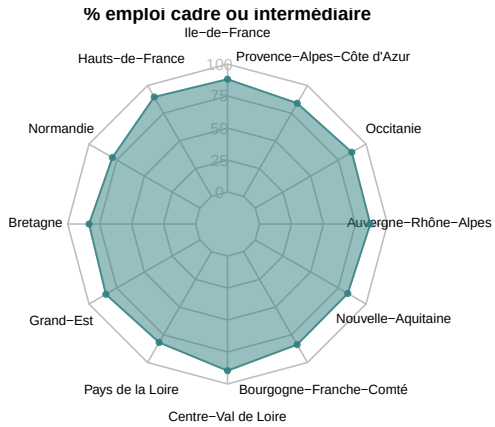
Pour notre étude nous avons établi un test de l'anova sur les donnés de nos établissement en les regroupant selon leur région, sans différenciation entre les années, pour observer une éventuelle dépendance. Nous sommes arrivées à la conclusion que la variable régions n'impact pas l'évaluation de la réussite d'une formation.

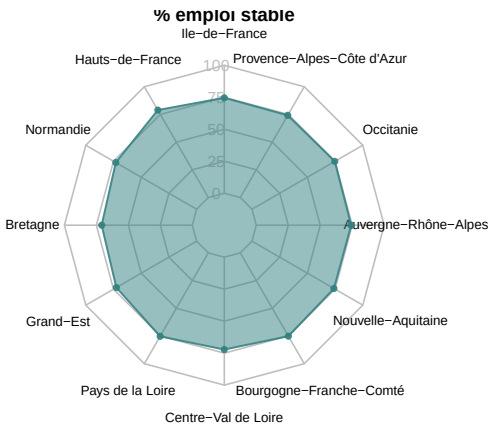
Nous avons tout de même effectuer le test de l'Anova sur les donnés de nos établissement en les regroupant selon leur région et selon les années pour observer une éventuelle dépendance. Nous sommes arrivée à la même conclusion, la variable région n'impact pas l'évaluation de la réussite d'une formation.

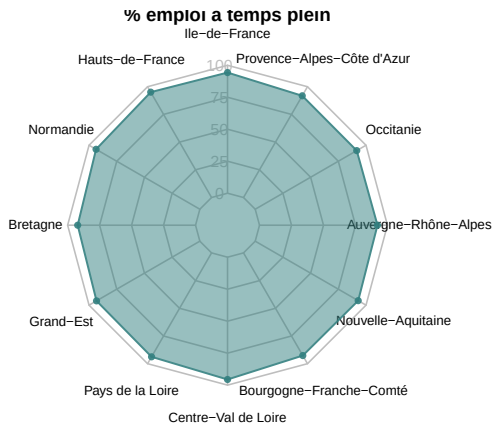
Analyse des variables sur l'année 2012

Emplois

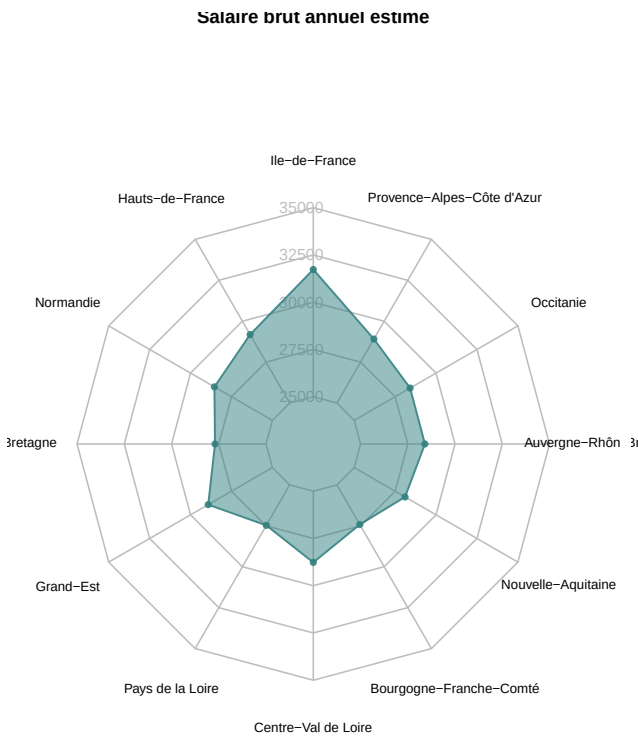


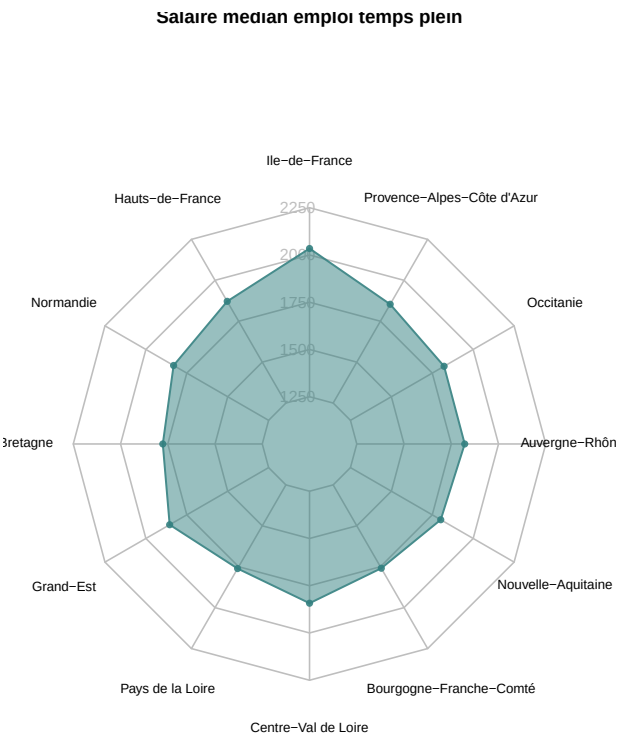




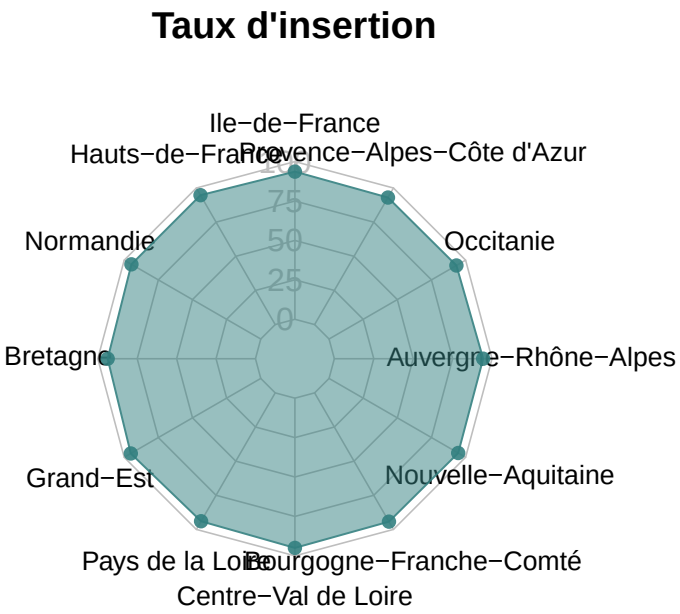


Salaires


Salaire brut annuel estime
Ile−de−France
Provence−Alpes−Côte d'Azur
Occitanie
Auvergne−Rhôn
Nouvelle−Aquitaine
Bourgogne−Franche−Comté
Centre−Val de Loire
Pays de la Loire
Grand−Est
Normandie
Hauts−de−France
35000
32500
30000
27500
25000

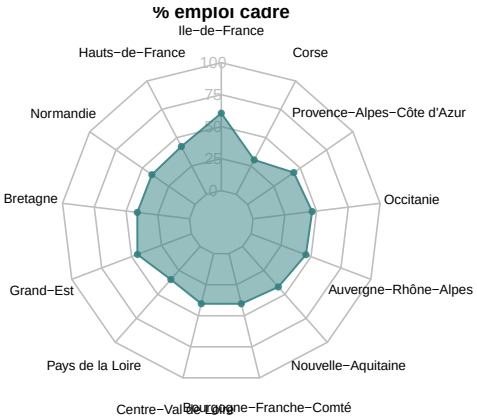

Salaire median emploi temps plein
Ile−de−France
Provence−Alpes−Côte d'Azur
Occitanie
Auvergne−Rhôn
Nouvelle−Aquitaine
Bourgogne−Franche−Comté
Centre−Val de Loire
Pays de la Loire
Grand−Est
Normandie
Hauts−de−France
2250
1750
1500
1250

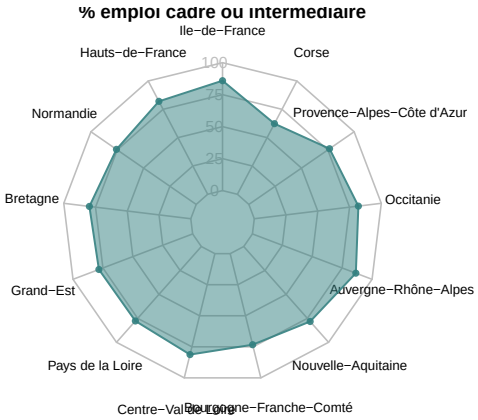
Taux d'insertion

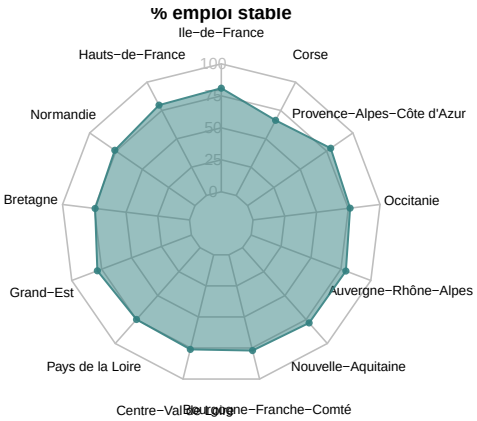


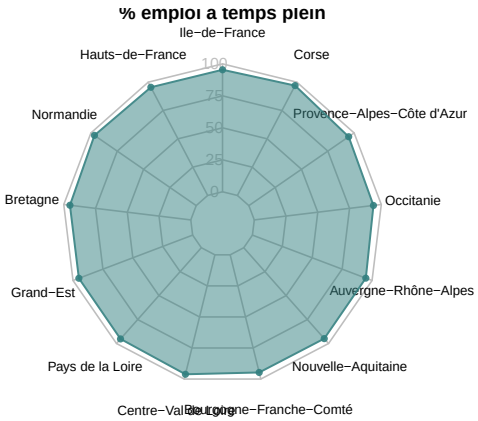
Analyse des variables sur l'année 2015

Emplois

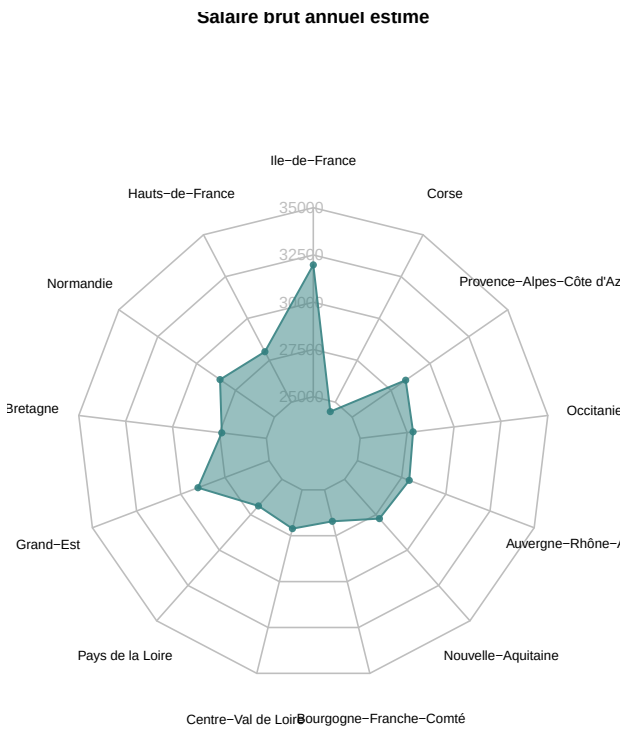


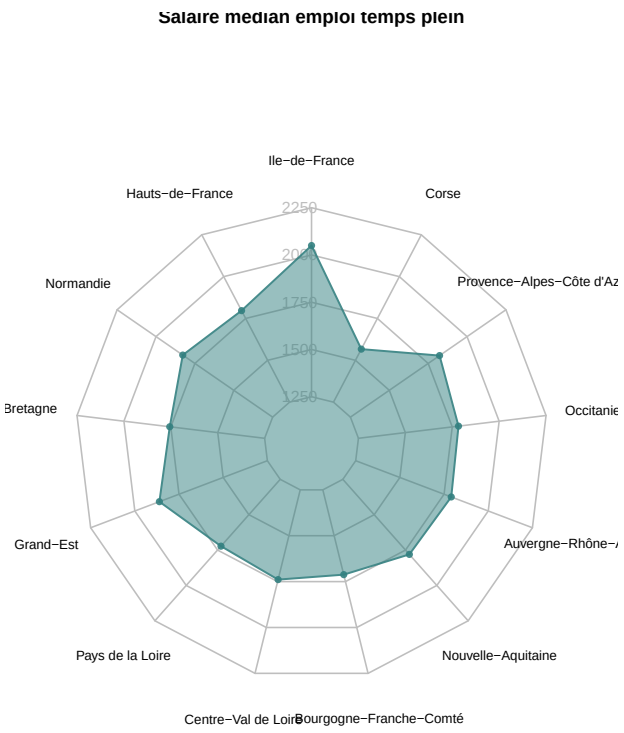




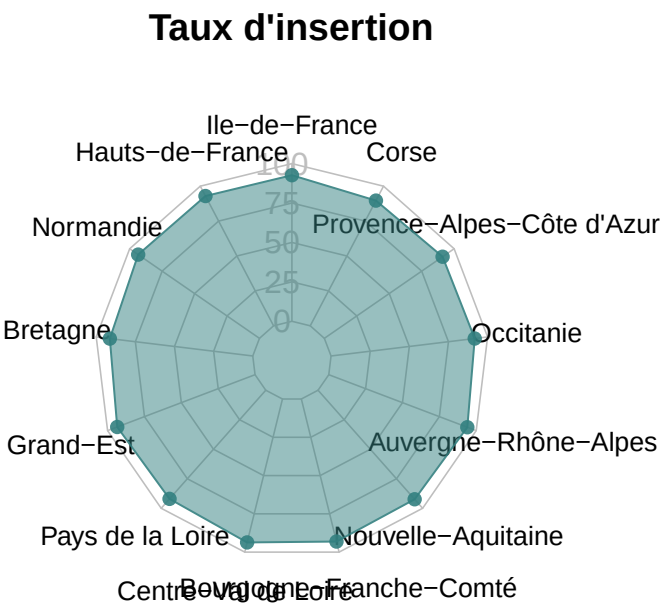


Salaires


Salaire brut annuel estime
Ile−de−France
Corse
Provence−Alpes−Côte d'Az
Occitanie
Auvergne−Rhône−A
Nouvelle−Aquitaine
Bourgogne−Franche−Comté
Centre−Val de Loire
Pays de la Loire
Grand−Est
Bretagne
Normandie
Hauts−de−France
35000
32500
30000
27500
25000

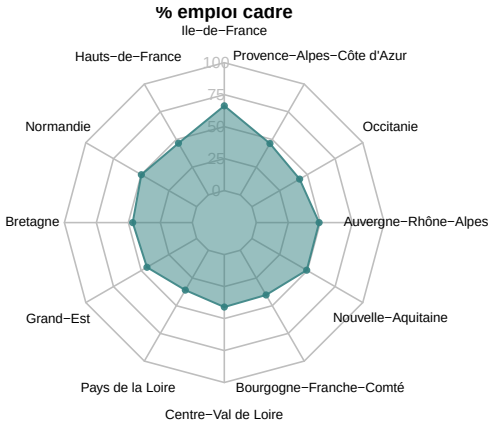

Salaire median emploi temps plein
Ile−de−France
Corse
Provence−Alpes−Côte d'Az
Occitanie
Auvergne−Rhône−A
Nouvelle−Aquitaine
Bourgogne−Franche−Comté
Centre−Val de Loire
Pays de la Loire
Grand−Est
Bretagne
Normandie
Hauts−de−France
2250
2000
1750
1500
1250

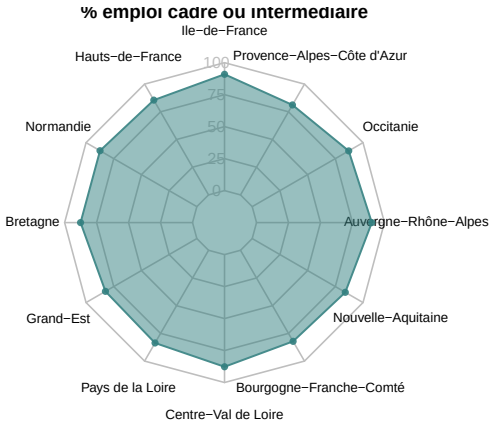
Taux d'insertion

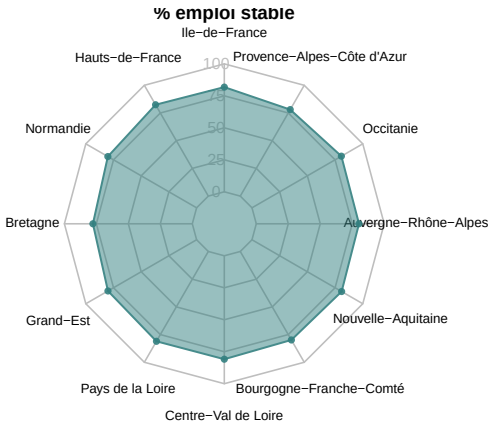


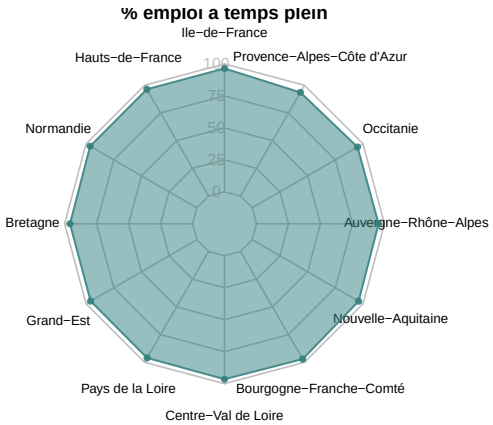
Analyse des variables sur l'année 2017

Emplois:

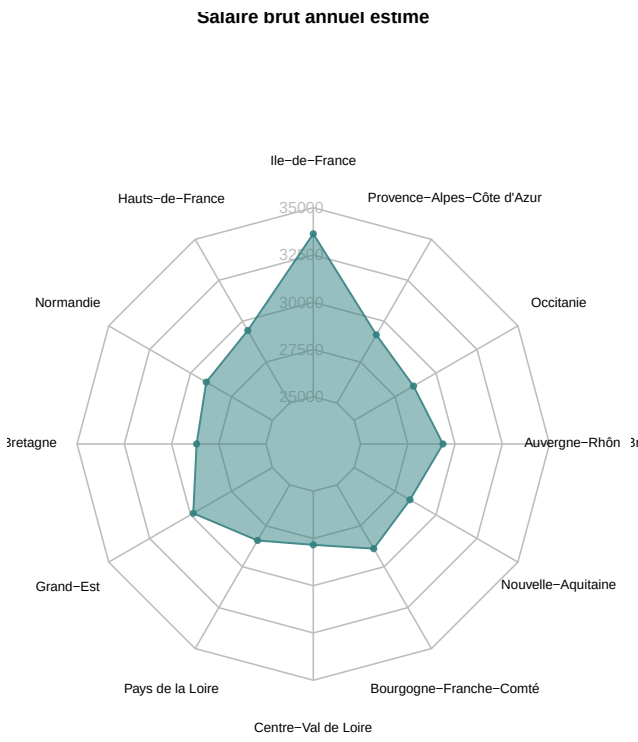


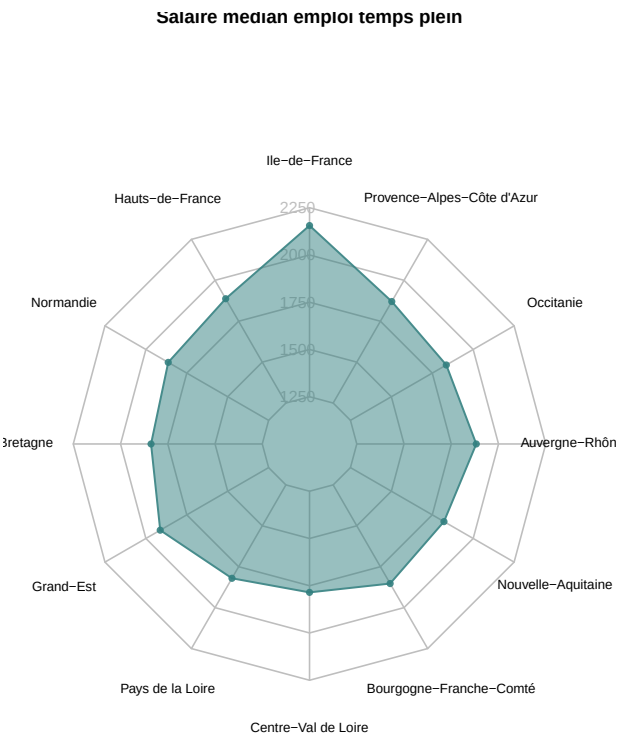




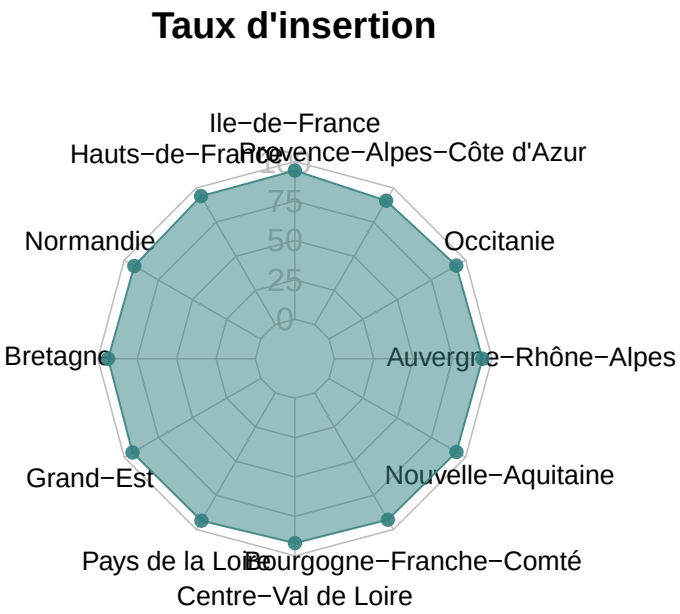


Salaires


Salaire brut annuel estime
Ile−de−France
Provence−Alpes−Côte d'Azur
Occitanie
Auvergne−Rhôn
Nouvelle−Aquitaine
Bourgogne−Franche−Comté
Centre−Val de Loire
Pays de la Loire
Grand−Est
Normandie
Hauts−de−France
35000
32500
30000
27500
25000


Salaire median emploi temps plein
Ile−de−France
Provence−Alpes−Côte d'Azur
Occitanie
Auvergne−Rhôn
Nouvelle−Aquitaine
Bourgogne−Franche−Comté
Centre−Val de Loire
Pays de la Loire
Grand−Est
Normandie
Hauts−de−France
2250
2000
1750
1500
1250

Taux d'insertion



Test de l'ANOVA sur les autres années

Année 2012

```
anova(data.2012)
```

```
## [1] "% emploi cadre"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1   6.03   6.031   0.223  0.647
## Residuals   10 270.86  27.086
## [1] "% emploi cadre ou intermédiaire"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1   0.13   0.133   0.012  0.916
## Residuals   10 114.34  11.434
## [1] "% emploi stable"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1   0.28   0.280   0.063  0.807
## Residuals   10  44.37   4.437
## [1] "% emploi à temps plein"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1  7.335   7.335   6.914 0.0252 *
## Residuals   10 10.609   1.061
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "Salaire médian emploi temps plein"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1  11602   11602   2.506  0.144
## Residuals   10  46298    4630
## [1] "Salaire brut annuel"
##             Df   Sum Sq Mean Sq F value Pr(>F)
## id_region    1  2805945 2805945   2.474  0.147
## Residuals   10 11342331 1134233
## [1] "Taux dinsertion"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1  6.682   6.682   6.927 0.0251 *
## Residuals   10  9.646   0.965
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Année 2015

```
anova(data.2015)
```

```
## [1] "% emploi cadre"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1  102.3  102.27   2.318  0.156
## Residuals   11  485.4   44.12
## [1] "% emploi cadre ou intermédiaire"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1   90.7   90.73   2.764  0.125
## Residuals   11  361.1   32.82
## [1] "% emploi stable"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1  30.73   30.73   2.564  0.138
## Residuals   11 131.84   11.99
## [1] "% emploi à temps plein"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1  0.280  0.2802   0.365  0.558
## Residuals   11  8.439  0.7672
## [1] "Salaire médian emploi temps plein"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1  49839   49839   6.128 0.0308 *
## Residuals   11  89467    8133
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "Salaire brut annuel"
##             Df   Sum Sq  Mean Sq F value Pr(>F)
## id_region    1 12057080 12057080   6.079 0.0314 *
## Residuals   11 21816974  1983361
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] "Taux dinsertion"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1  3.976   3.976   3.101  0.106
## Residuals   11 14.103   1.282
```

Année 2017

```
anova(data.2017)
```

```
## [1] "% emploi cadre"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1  110.6  110.56   2.205  0.168
## Residuals   10  501.5   50.15
## [1] "% emploi cadre ou intermédiaire"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1  13.74   13.73   1.443  0.257
## Residuals   10  95.20    9.52
## [1] "% emploi stable"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1  2.936   2.936   1.681  0.224
## Residuals   10 17.467   1.747
## [1] "% emploi à temps plein"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1  3.463   3.463   6.219 0.0318 *
## Residuals   10  5.569   0.557
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "Salaire médian emploi temps plein"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1  24894   24894   3.418 0.0942 .
## Residuals   10  72841    7284
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "Salaire brut annuel"
##             Df   Sum Sq Mean Sq F value Pr(>F)
## id_region    1  6037184 6037184   3.393 0.0953 .
## Residuals   10 17793064 1779306
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "Taux dinsertion"
##             Df Sum Sq Mean Sq F value Pr(>F)
## id_region    1  3.358   3.358    4.01 0.0731 .
## Residuals   10  8.375   0.837
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] TRUE
```