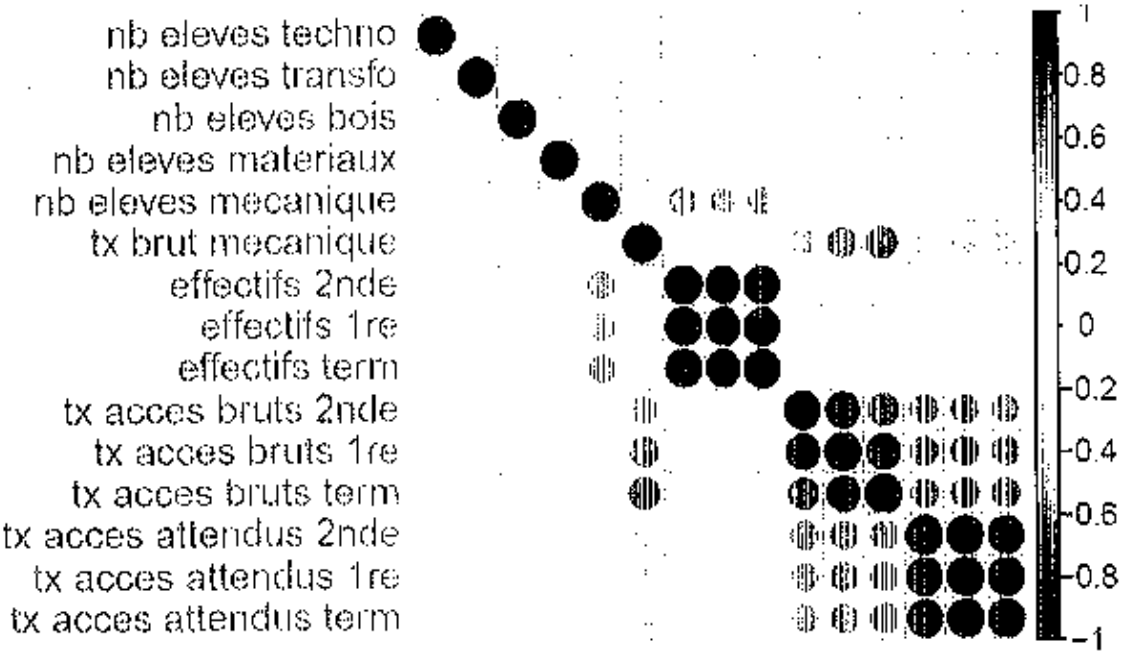
Résultats des lycées professionnels

[redacted]

Le but de ce projet est d'étudier les données issues du ministère de l'éducation nationale concernant les résultats au BAC des filières professionnelles. Nous nous intéressons plus précisément au taux de réussite brut du secteur "Mécanique, Électricité et Électronique" (MEE), puis du secteur "Production". Pour comprendre et prédire ces taux de réussite, nous n'utilisons pas toutes les données du tableau mais seulement celles concernant les effectifs à la rentrée des classes, les taux d'accès bruts à la filière et les taux d'accès attendus, et ceci pour les trois classes (2nde, 1ère, Terminale) pour l'année 2015. De plus certaines données sont propres au secteur d'étude. Ici, nous ne parlerons que du secteur MME. On ajoute alors comme autres données le nombre d'élèves présents au BAC dans les différentes séries du secteur "Production".

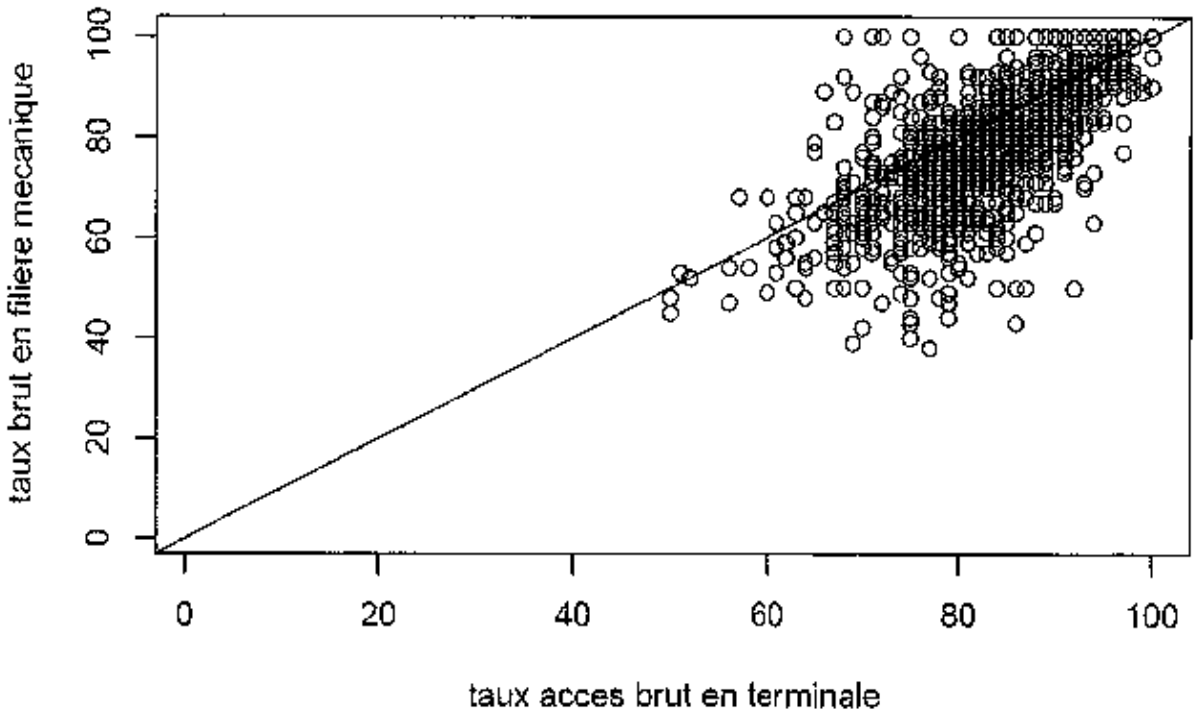
On peut rapidement remarquer que le tableau de données comporte un certain nombre d'erreurs nous empêchant d'appliquer simplement les fonctions de R. Nous avons donc supprimé les lignes mal écrites dans le tableau, qui proviennent principalement de sauts prématurés à la ligne. Le fait que de nombreuses cases du tableau soient vides n'est pas un problème, il n'est pas nécessaire, du moins pas à notre état d'avancement, de nous préoccuper de combler d'une manière ou d'une autre les trous.

Afin de comprendre le lien entre notre variable cible et les autres variables décrites précédemment, une bonne manière de débuter est d'afficher la matrice de corrélation de ces variables. Elle permet de déterminer quelles sont les variables qui ont l'influence la plus grande sur notre variable cible.



Cette matrice de corrélation nous permet d'observer que les variables corrélées avec notre variable cible, à savoir le taux de réussite brut en filière mécanique, électricité et électronique, est corrélé avec les taux d'accès bruts et les taux d'accès attendus. En revanche, elle est très peu corrélée avec le nombre d'élèves dans les différentes filières, ce qui est logique étant donné que nous nous intéressons à un taux et que les filières étudiées ne sont pas forcément la filière Mécanique, Électricité et Électronique.

Nous avons ensuite tracé la répartition des taux de réussite bruts en filière mécanique par rapport aux taux d'accès bruts en terminale afin de visualiser la corrélation de ces deux variables. Si les deux variables étaient parfaitement corrélées, les points seraient alignés sur la droite identité. Dans notre cas, les points sont éparpillés, mais on remarque tout de même une tendance générale à se rapprocher de la droite, ce qui confirme les résultats de la matrice de corrélation. Le fait que les points soient concentrés dans le quart en haut à droite du graphique est normal, c'est en effet dû au fait que les taux de réussite sont toujours supérieurs à 50%.



M.R.R Project : Statistical Analysis and Description

1 Introduction

Every year, the Ministry of National Education publishes information about the different proffessional high schools. They are certainly used by the press to establish high school's prize list. But it would be unfortunate to limit their use only to college students and families able to choose their high school for the next school year. They are useful tools to evaluate the high school that one frequents, that one is about to integrate for lack of real alternative, or that one attended.

To extend the use of this data we're going to study and analyse it in order to make some predections.

2 problematic

This study will have for object :

- Explain and predict the gross success rate for the Mechanical, Electrical and Electronics series.
- Explain and predict the gross success rate of the Production sector.

3 Study of the data set

3.1 Review the organisation and the names of the columns

Our dataset has two lines of headers.The use of this data in this structure may be very difficult and could create a lot of problems in our dataset study. That's why, we decided to clean and reorganize our data before starting the analysis. We first reorganized the header in a more simple way.explain

3.2 Structure of the data

The data is representing the statistical review made on professional high schools in France in the year of 2016, by the french government.It's based on the results of the baccalaureate students and their educational background in the institution. Professional high schools, public and private under contract, are concerned. This data conteins 2027 observations of 56 variables, grouped according to informations concerning High school localisation, number of registers, etc,... The data also gives us significant amount of informations about the registered students as well as the success rate of the students,in different classes.

3.3 Checking NULL/Missing values

Viewing the different rows of the data indicates that there are too many columns which have missing values, others have a ND values.

FIGURE 1 – Number of the missing value for the variables.

FIGURE 2 – Number of the missing value for the variables.

The variables with the largest number of missing values are :

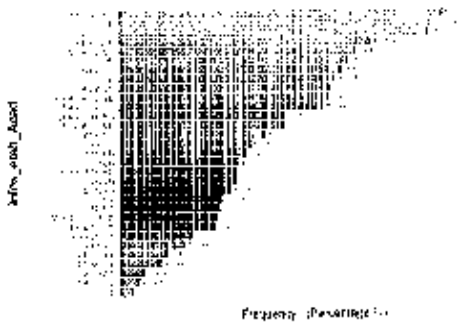
- Nb_elvs_bac_SPT, Nb_elvs_bac_TT, Nb_elvs_bac_CI, Nb_elvs_bac_SC, TRB_SPT, TRB_MS, TRB_CI.

The largest missing values are observed in the part concerning the number of students in the final year of different fields and the success rate. This lack of information will affect the rest of our study. For that, we have several solutions such as, droping the column with highest missing values, replacing the missing values by 0 or, replacing the missing values with the mean value of the column. The fisrt and the second solutions affect substantially the results of our study, because we can not just drop the variable used to describe and predict our target variable, and putting the 0 will affect the success rate. Finally, we chose the third solution.

3.4 Creating corresponding data for each study

According to the help file given by tutorial master, the two studies will have a common basis discriptors, we're going to add specifics discriptors depending on the target variable.

3.5 Analyzing categorical variables



According to our first approch of categorical variable, we notice that some region have more professional high schools than the others, for example, "VERSAILLES",with 142 professionnal high schools. We also notice that the number of PUBLIC professionnal high schools(74.2 percent) is greater than PRIVATE ones(25.8percent).

3.6 Analyzing numeric variables

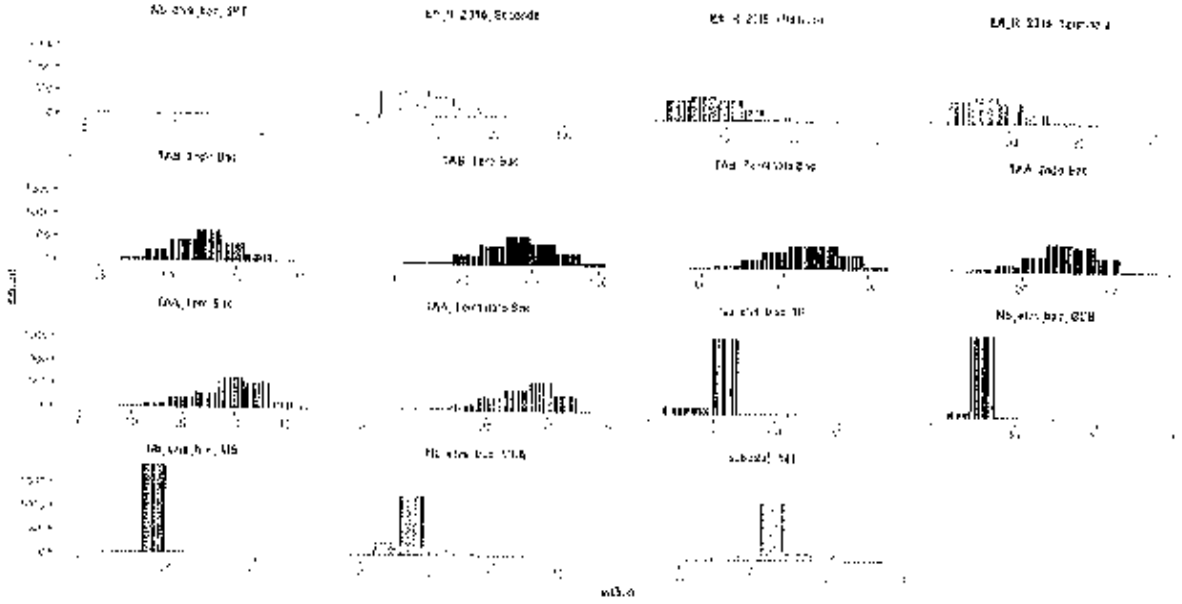


FIGURE 5 – The number of the explanatory variables

We can have an approximation of the mean value of these variables just according to the graphs above.

3.7 Correlation

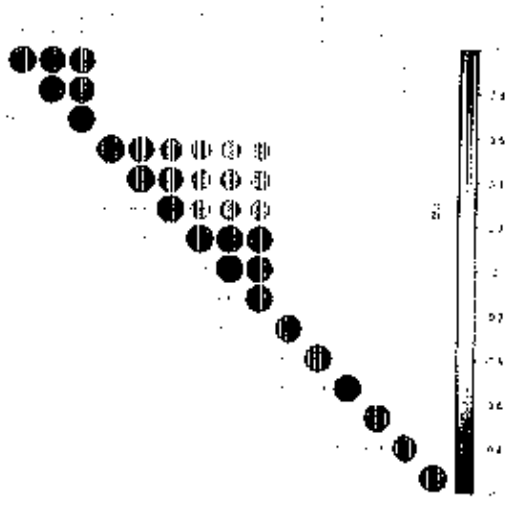


FIGURE 6 – The correlation graph of our target variable and some of the chosen co-variables.

The correlation graph shows that some of the explanatory variables are highly correlated, we should take the strong correlation impact in consideration in the regression analysis.

Projet MRR - Présentation des données

Description des objectifs

L'objectif est de prédire la réussite au bac par lycée pour la série "mécanique, électricité électronique" et pour le secteur "production". Pour cela, on utilise une base de données générée par le gouvernement. Il contient pour chaque lycée professionnel les informations de l'établissement, les effectifs à la rentrée, Les taux d'accès bruts, c'est à dire les taux d'accès calculés en fonction de la réussite des élèves au baccalauréat et les taux d'accès attendus. Les taux d'accès attendus sont des estimations du taux d'accès calculés sans prendre en compte la réussite réelle. Ce sont des prévisions établies à partir de facteurs extérieurs au lycée qui expliqueraient une réussite au baccalauréat: note au DNB, catégorie socioprofessionelle des parents etc ...

Intéret du taux d'accès

Le taux d'accès est une variable plus intéressante que le taux de réussite, il donne plus d'informations : le taux de réussite donne uniquement le pourcentage de personnes ayant obtenu le baccalauréat au cours de l'année, alors que le taux d'accès donne la proportion d'élèves qui aura son bac si il poursuit sa scolarité dans l'établissement.

Intéret des taux attendus

Les taux attendus permettent d'évaluer efficacement la performance d'un lycée: en effet, comparer simplement les taux de réussite de 2 lycées différents n'a pas forcément de sens, il est possible qu'un des deux lycées accueille tout simplement de meilleurs élèves. Pour pallier ce problème, les taux attendus sont introduits. Un lycée est considéré performant si il parvient a avoir de meilleurs résultats que ce qui était attendu avec la qualité de ses élèves.

Matrices de similarité

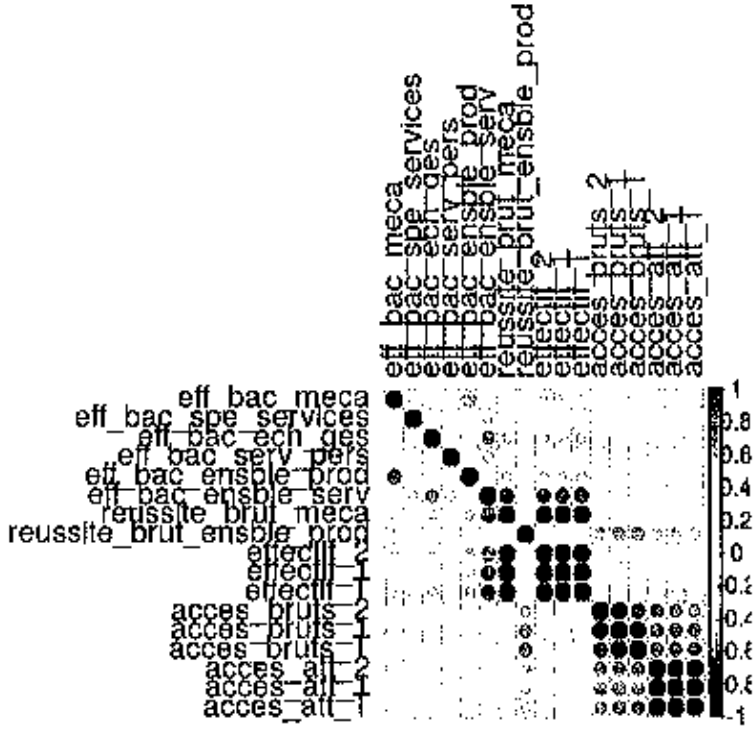
On a décidé de remplacer les valeurs manquantes de chaque colonne du tableau par la moyenne des valeurs présentes dans les colonnes car cela est plus précis que de remplacer tous les NA par la même constante.

Matrice 1 : Modèle ayant pour variable cible la filière 'Mécanique, Électricité et Électronique'

Comme on pouvait s'y attendre, les effectifs de chaque année sont fortement corrélés entre eux puisque généralement les élèves effectuent toute leur scolarité dans le même établissement, de même pour les taux d'accés bruts et les taux d'accés attendus. Enfin, les taux d'accés bruts et attendus sont signifactivement corrélés, ce qui montre que les taux d'accés attendus sont plutôt réalistes.

Matrice 2 : Modèle ayant pour variable cible le secteur Production

On retouve les mêmes corrélations que sur le premier graphique. De plus, la variable cible (taux_reussite_brut_ensble_production) est corrélé avec les taux d'accés. On remarque aussi que le nombre d'éléves dans chaque filière n'est pas autant corélé avec les effectifs des etablissements, ce qu'on peut expliquer par l'absence de certaines filières dans certains établissements.



Résultats des lycées d'enseignement professionnel

Introduction et position du problème

Comment mesurer, pour une filière professionnelle spécifique, le taux de réussite brut au Baccalauréat en fonction de l'établissement et des conditions de passage du Baccalauréat ?

Description de la base de données

La base de données brute est composée de 56 colonnes, regroupées en 7 grands groupes de variables. Pour chaque établissement, nous retrouvons :

1. Informations sur l'établissement ;
2. Nombre d'élèves présents au Bac (par filière, dont **Mécanique-électricité-électronique** et **Production)** ;
3. Taux de réussite bruts (mêmes filières, dont **Mécanique-électricité-électronique** et **Production)** ;
4. Taux de réussite attendus (mêmes filières, dont **Mécanique-électricité-électronique** et **Production)** ;
5. Effectifs à la rentrée 2016 (par classe de seconde, première ou terminale) ;
6. Taux d'accès bruts (accès au Baccalauréat depuis la seconde, la première ou la terminale) ;
7. Taux d'accès attendus (accès au Baccalauréat depuis la seconde, la première ou la terminale) ;

Nettoyage de la Base de Données

1. Nous avons mis en place l'encodage correct (UTF-8) ;
2. Nous avons opéré des modifications sur certaines données qui étaient suivis de « (1) » ;
3. Notre base de données comporte des « ND ». Etant donnée le contexte, nous pouvons considérer ces valeurs comme des NA. De nombreuses valeurs sont manquantes. Nous avons décidé de remplacer celles-ci par les médianes. Cette technique permet d'avoir des valeurs sans biaiser fortement les résultats.
4. La Base de Données contient les variables « Etablissement » et « Code Etablissement ». Ces deux variables font référence à la même chose : l'identité d'un établissement (qui est unique). De ce fait, nous n'avons pas besoin d'utiliser ces deux variables pour l'étude. Il en est de même pour les variables "Villes" et "Code Commune". Parmi tous ces exemples, nous ne garderons donc que les codes.
5. Une des variables (la structure pédagogique en 7 groupes) est constante. De ce fait, nous ne la prendrons pas en compte.

Variables-cibles

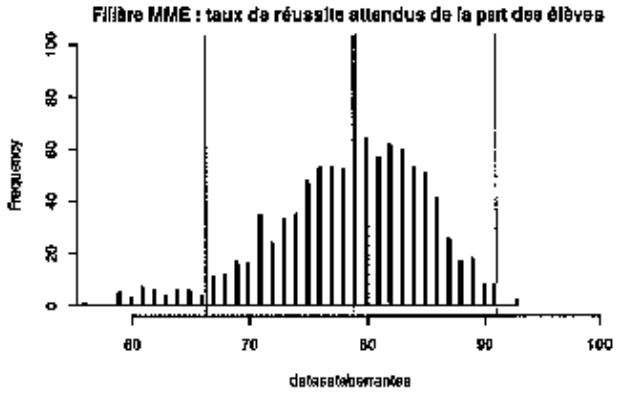
Nous allons mener deux études séparées afin de décrire au mieux deux résultats : le taux de réussite brut pour la filière **Mécanique, électricité, électronique (MEE, colonne Z)** et la filière **Production (colonne AF).**

La nature de ces variables reste la même : il s'agit de taux de réussite bruts. Il s'agit, pour l'année 2016, des résultats effectifs, réels, obtenus par un lycée.

Variables explicatives

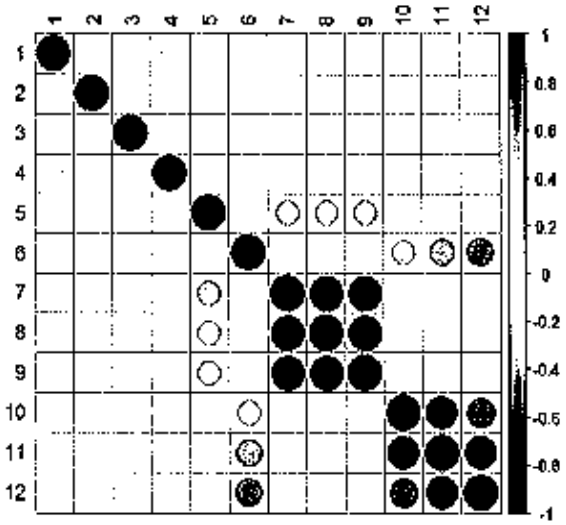
Les taux de réussite « attendus » tiennent compte des caractéristiques sociodémographiques et scolaires des élèves qui le fréquentent ». Les **taux d'accès attendus** (colonnes BB à BD) et les **taux d'accès bruts** (colonnes AY à BA) calculent, pour les classes de Seconde, Première et Terminale, la part d'élèves accédant au Bac (sans prendre en compte la durée de scolarité).

Gestion des valeurs aberrantes des variables explicatives



En bleu, nous avons tracé les « valeurs pivots » (calculées à partir de l'écart interquartile) pour la variable explicative **taux de réussite attendus de la part des élèves**. Les taux attendus en-deçà ou au-delà de ces valeurs sont aberrants : ce sont des outliers et nous les avons supprimés.

Matrice de Corrélation



1. nb_eleve_bac_spé_plurl
2. nb_eleve_bac_transformations
3. nb_eleve_bac_genie_civil
4. nb_eleve_bac_matériaux_souples
5. nb_eleve_Méca/Electricité/Electronique
6. tx_reu_Méca_Electricité_Electronique
7. effectif_2016_seconde
8. effectif_2016_premiere
9. effectif_2016_terminale
10. tx_acces_brutes_2016_seconde
11. tx_acces_brutes_2016_premiere
12. tx_acces_brutes_2016_terminale

Nous voyons que le taux de réussite au baccalauréat « Mécanique, Electrique et Electronique » (6) n'est pas aussi corrélé à l'effectif des classes aux lycées qu'il est corrélé aux taux d'accès bruts dans chacune des classes. Nous verrons que les taux d'accès correspondent effectivement à la valeur ajoutée des lycées, et que la valeur ajoutée est plus forte en première ou terminale qu'en seconde (du fait des changements d'orientation).

Contexte :

Dans le cadre de notre projet de MRR, nous sommes amenés à effectuer une étude complète d'une base de données. Nous sommes tombés sur une base de données regroupant plusieurs informations concernant les résultats au BAC des lycées professionnels de France.

Description de la base de données :

La base de données est décrite en 56 colonnes représentant chacune une variable et en 2027 lignes représentant chaque lycées professionnels.

Les variables sont regroupées selon 7 thèmes :

- Informations établissements : Académie, Département, Etablissement, Ville, Code établissement, Code commune, secteur public ou privé, Structure pédagogique.

- Nombre d'élèves au Bac : Spécialité pluri-technologiques ; transformations ; Genie civil, construction et bois ; Matériaux souples ; Mécanique, électricité, electronique ; spécialitésplurivalentes des services ; échanges et gesstion ; communication et information ; services aux personnes ; services à la collectivité ; ensemble production ; ensemble services ; ensemble tous secteurs ;

- Taux de réussite bruts : Spécialité pluri-technologiques ; transformations ; Genie civil, construction et bois ; Matériaux souples ; Mécanique, électricité, electronique ; spécialitésplurivalentes des services ; échanges et gesstion ; communication et information ; services aux personnes ; services à la collectivité ; ensemble production ; ensemble services ; ensemble tous secteurs ;

- Taux de réussite attendus : Spécialité pluri-technologiques ; transformations ; Genie civil, construction et bois ; Matériaux souples ; Mécanique, électricité, electronique ; spécialitésplurivalentes des services ; échanges et gesstion ; communication et information ; services aux personnes ; services à la collectivité ; ensemble production ; ensemble services ; ensemble tous secteurs ;

- Effectifs à la rentrée : Seconde ; Première ; Terminale-Bac ;

- Taux d'accès brut : Seconde ; Première ; Terminale-Bac ;

- aux d'accès attendus : Seconde ; Première ; Terminale-Bac ;

Hypothèses d'analyse :

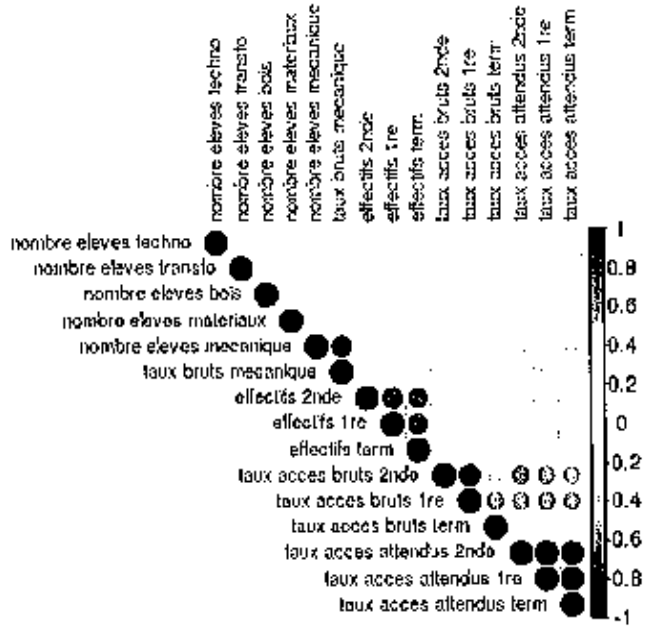
Dans ce paragraphe, nous allons expliciter nos idées d'études au vu des variables présentes. Premièrement, il nous semble intéressant d'expliquer et prédire le taux de réussite brut pour la série Mécanique, Électricité et Électronique en ajoutant comme descripteurs le nombre d'élèves présents au bac dans les séries du secteur Production .

Ensuite, on souhaiterait expliquer et prédire le taux de réussite brut du secteur Production en ajoutant comme descripteurs le nombre d'élèves présents au bac tous secteurs et toutes séries confondus.

Justification :

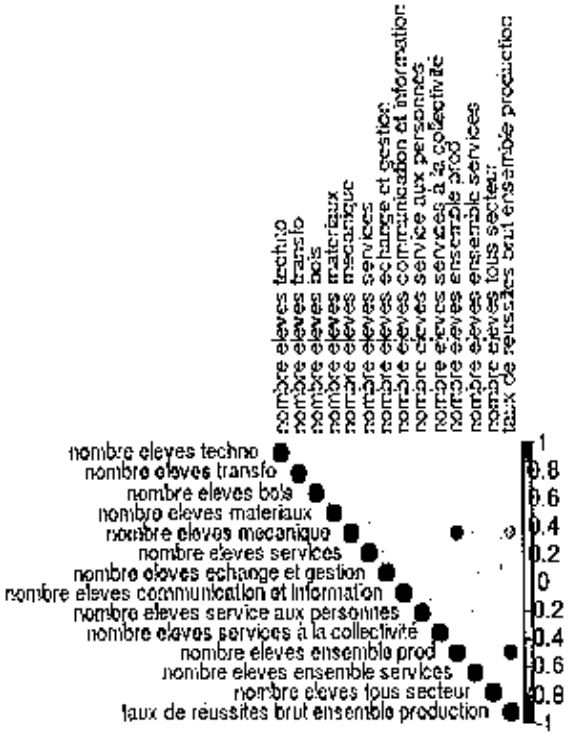
En effet, les hypothèses formulées précédemment semblent logique. Il s'agit maintenant de prouver que c'est bien le cas.

Première Hypothèse :



On constate bien une corrélation entre le taux brut et les nombres d'élèves en secteur production, d'où l'intérêt d'étudier le taux de réussite brut en Mécanique, électricité et électronique.

Deuxième Hypothèse :



On remarque dans la colonne taux de réussites brut une bonne corrélation avec l'ensemble des variables descriptives et particulièrement avec le nombre d'élèves dans l'ensemble de production.

Ainsi, nos hypothèses formulées précédemment sont bien justifiées. Il faut maintenant commencer l'étude des deux variables cibles de manière explicative et essayer de bien interpréter pour fournir un modèle qui sera valable dans le futur.

Projet MRR: évaluation des lycées

Objectifs

Nous allons donc essayer de noter les lycées professionnels. Pour cela, une première manière est de regarder le pourcentage brut de réussite au bac, dans un premier temps nous tenterons donc de modéliser cette variable. Ensuite nous pourrons éventuellement modéliser cette variable par filière pour observer quel est le meilleur lycée par filière. Nous commencerons donc par créer des modèles simple avec lm pour ces variables cible et utiliser la p-value pour avoir une idée de quelles sont les variables importantes. Ensuite nous regarderons la corrélation entre les variables notemment avec la variable cible afin d'avoir une première idée de quelles variables explicatives sont significatives. Cependant comme on a pu le voir dans le sujet, le résultat brut au bac n'est pas suffisant pour déterminer la qualité d'un lycée: si celui-ci renvoie ses plus mauvais éléments au sens scolaire du terme avant la terminal, alors il lui sera aisé d'obtenir les résultats: il faut donc prendre.Il faut donc considérer l'IVAL. Tout d'abord il faut rajouter le taux d'accès au bac, parmis tous les élèves arrivant en secondes, combien passeront leur bac et finiront par l'avoir.

Nous étudierons plus particulièrement les résultats de la filière Mécanique, Electronique et Electricite ainsi que ceux du secteur production.

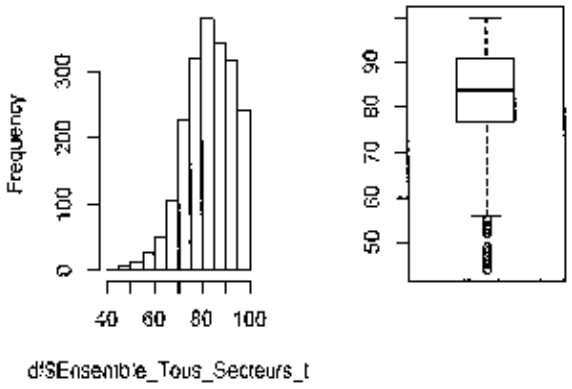
Dataset

Le dataset comporte six types de variables: les informations sur l'établissement Les effectifs à la rentrée,e taux de réussite brut au bac, le taux de réussite attendu, les taux d'accès bruts, les taux d'accès attendus.

La table ne peut être lue par Rstudio sans transformations préalables car beaucoup de lycées ne présentent pas toutes les filières technologiques/professionnelles qui sont proposées au niveau national. Il faut donc que nous trouvions un moyen de pallier ce problème. Nous pensons mettre des zéros à la place des cases vides. Cette technique n'est que partiellement satisfaisante ; autant elle modélise très bien la réalité quant au nombre d'élève présentés par filière mais elle présente une réalité biaisée pour le taux de réussite pour ne citer que cet exemple et cela risquerait de créer des corrélations inexistantes. Cette idée fût donc vite abandonnée. De plus, il y a aussi des des données non déterminées (ND), nous pouvons donc choisir de les retirer des données le temps de la création des modèles (peu d'influence vu que le nombre de données que nous possédons) ou bien de remplacer ces ND par la médiance de la colonne à laquelle ils appartiennent.

Le dataset contient 57 variable avec 2027 données par variables.On constate qu'il ya plus de données que de variables, on pourra donc éventuellement utiliser les

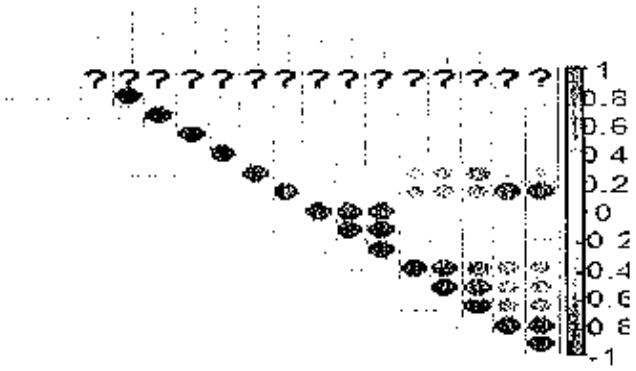
méthodes pénalisation du type forward pour retirer des variables des modèles de régression (les résultats concernant la filière matériaux souple ont peu de chance d'être impactant sur ceux de la série Mécanique) et obtenir des modèles plus simples et pertinents. De plus nous n'avons pas constaté de dépendances linéaires entre les variables, ce qui créerait une redondance, donc nous n'en avons pas



retirée.

Etude des corrélations

Voici une visualisation rapide des variables cibles (on remarque une certaine ressemblance à des lois gaussiennes) avec le résultat général pour comparé. Les résultats de la série Mécanique semblent inférieurs aux résultats moyens d'un lycée au contraire de ceux du secteur production.



Graphiquement, nous pouvons observer que la variable cible Taux de réussite au bac brut de la filière Mécanique, Electronique et Electrique n'est pas corrélée ou peu avec les données des autre filières ainsi que les effectifs de l'école. On pourra éventuellement retirer ces variables à l'avenir, cependant cela n'est pas suffisant pour les retirer: il faudra effectuer une sélection de variables à l'avenir afin de conserver uniquement les variables les plus significatives.