MOHAMMED VI POLYTECHNIC UNIVERSITY

EMINES School of Industrial Management

UNIVERSITÉ MOHAMMED VI POLYTECHNIQUE

Rapport final :

Education : Evaluation statistique des systèmes de classement international des universités

Elaboré Par :

ELJABBAR Ayya

ELJABBAR Nassima

MAJDOUBI Nouhayla

MARBOUH Oumaima

Encadré Par :

Khalil SAID

17/05/2020

Plan

I. Introduction et problématique :

1. Introduction

Dans le cadre de notre formation à l'EMINES, particulièrement pendant la deuxième année du cycle ingénieur, nous sommes amenés à effectuer une étude statistique sous forme de projet en groupe. Ceci ne peut être qu'un avantage pour nous dans notre vie professionnelle par la suite. Ce cours vise à nous aider à mettre au point des outils théoriques, comme les méthodes de modélisation mathématique et les différents tests, ainsi que des outils pratiques, à savoir les logiciels R et SAS, pour mieux comprendre des phénomènes réels.

Le but est d'acquérir les connaissances nécessaires pour réussir à traiter, analyser et modéliser correctement les données, de développer une sorte d'intuition pour l'interprétation convenable des résultats, et finalement de les présenter de manière compréhensible et simple.

2. Problématique :

Chaque année, les classements des universités créent un véritable « buzz », plus le rang est élevé, meilleure est la faculté ; Qui sera dans le Top 10 ? Sommes-nous améliorés ? Où en sont les autres universités ? Ces interrogations agitent les comités de direction des établissements d'enseignement supérieur.

Dans ce sens, quelques questions s'imposent :

- Comment parvient-on à un tel classement ?
- Quelles sont les critères qui influencent le score ?
- Pourquoi les résultats divergent-ils autant d'un classement à l'autre ?

C'est à ces questions qu'on essaye de répondre, et c'est dans ce cadre que s'inscrit notre projet statistique. Nous disposons d'une banque de données sur 3 systèmes de classements nettement différents : Times Higher Education World University Ranking, le classement de Shangaï, et le Center for World University Rankings. Notre but est de mener une étude complète sur notre base de données, à savoir nettoyage et description statistique des variables avant de modéliser les différents systèmes de classement. Nous allons ensuite chercher à valider que les classements sont biaisés et nous allons conclure en récapitulant les points essentiels et les conclusions principales de notre étude.

II. Description des variables :

La première des choses, on a importé la data dans R. On dispose de 5 bases de données

```
cwur = read.csv("cwurData.csv")
shanghai = read.csv("shanghaiData.csv")
times = read.csv("timesData.csv")
expenditure = read.csv("education_expenditure_supplementary_data.csv")
attainment = read.csv("educational_attainment_supplementary_data.csv")
```

- **CWUR :** regroupant les classements et les scores de plusieurs universités sur quelques années selon le système cwur.
- **Shanghai :** regroupant les classements et les scores de plusieurs universités sur quelques années selon le système cwur.
- **Times :** regroupant les classements et les scores de plusieurs universités sur quelques années selon le système cwur.
- **Expenditure :** détaillant le pourcentage des dépenses dédié à l'éducation pour plusieurs pays.
- **Attainment :** détaillant le niveau d'accomplissement éducationnel (%) pour plusieurs pays.

Découvrant la structure de chacune des bases de données en plus de détails, et ce en utilisant la fonction str() .

1. Le classement CWUR :

On a 2200 observations de 14 variables :

```
> str(cwur)
'data.frame':   2200 obs. of  14 variables:
 $ world_rank          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ institution         : Factor w/ 1024 levels "Ã–rebro University",..: 194 322 520 653 63 442 83
3 1009 106 643 ...
 $ country             : Factor w/ 59 levels "Argentina","Australia",..: 59 59 59 57 59 59 57 59
 59 59 ...
 $ national_rank       : int  1 2 3 1 4 5 2 6 7 8 ...
 $ quality_of_education: int  7 9 17 10 2 8 13 14 23 16 ...
 $ alumni_employment   : int  9 17 11 24 29 14 28 31 21 52 ...
 $ quality_of_faculty  : int  1 3 5 4 7 2 9 12 10 6 ...
 $ publications        : int  1 12 4 16 37 53 15 14 13 6 ...
 $ influence           : int  1 4 2 16 22 33 13 6 12 5 ...
 $ citations           : int  1 4 2 11 22 26 19 15 14 3 ...
 $ broad_impact        : int  NA NA NA NA NA NA NA NA NA NA ...
 $ patents             : int  5 1 15 50 18 101 26 66 5 16 ...
 $ score               : num  100 91.7 89.5 86.2 85.2 ...
 $ year                : int  2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 ...
```

- World_rank : le classement de l'université à l'échelle internationale selon le système cwur
- Institution : nom de l'université
- Country : pays de l'université
- National_rank : le classement de l'université à l'échelle nationale selon le système cwur
- Quality_of_education : classement de la qualité de l'éducation de l'université
- Alumni_employment : classement de l'emploi des anciens de l'université
- Quality_of_faculty : classement du corps professoral de l'université
- Publications : classement selon les publications de l'université
- Influence : classement selon l'impact de l'université
- Citations : classement selon le nombre des citations de l'université
- Broad_impact : classement selon l'impact global de l'université
- Patents : classement selon les patents de l'université
- Score : Score de l'université selon le système cwur
- Year : L'année dans laquelle les données sont déclarées

2. Le classement Shanghai :

On a 4897 observations de 11 variables :

```
> str(shanghai)
'data.frame':   4897 obs. of  11 variables:
 $ world_rank     : Factor w/ 119 levels "1","10","100",..: 1 21 36 50 65 76 87 98 109 2 ...
 $ university_name: Factor w/ 659 levels "","Aalborg University",..: 102 416 271 407 163 29 55
 235 421 538 ...
 $ national_rank  : Factor w/ 293 levels "","1","1-2","1-3",..: 2 2 103 151 193 216 237 248 264
 103 ...
 $ total_score    : num  100 73.6 73.4 72.8 70.1 67.1 62.3 60.9 60.1 59.7 ...
 $ alumni         : num  100 99.8 41.1 71.8 74 59.2 79.4 63.4 75.6 64.3 ...
 $ award          : num  100 93.4 72.2 76 80.6 68.6 60.6 76.8 81.9 59.1 ...
 $ hici           : num  100 53.3 88.5 69.4 66.7 59.8 56.1 60.9 50.3 48.4 ...
 $ ns             : num  100 56.6 70.9 73.9 65.8 65.8 54.2 48.7 44.7 55.6 ...
 $ pub            : num  100 70.9 72.3 72.2 64.3 52.5 69.5 48.5 56.4 68.4 ...
 $ pcp            : num  72.4 66.9 65 52.7 53 100 45.4 59.1 42.2 53.2 ...
 $ year           : int  2005 2005 2005 2005 2005 2005 2005 2005 2005 2005 ...
```

- world_rank: le classement de l'université à l'échelle internationale selon le système Times
- national_rank : le classement de l'université à l'échelle nationale
- university_name: nom de l'université
- total_score: Score total de l'université utilisé pour déterminer le classement
- Alumni: Alumni Score, basé sur le nombre des étudiants qui ont gagné un prix Nobel ou bien d'autres médailles.
- Award: Award Score,basé sur le nombre de personnel de l'institution qui ont gagné des prix Nobel on physique, chimie, médecine et l'économie et autres médailles en mathématique.
- Hici : HiCi Score, basé sur le nombre de chercheurs les plus cité et sélectionné par Thomson Reuters
- Ns :N&S Score, basé sur le nombre d'articles publiés et qui ont une relation avec la nature et la science.
- Pub :PUB Score, basé sur le nombre total de papiers indexés en " the Science Citation Index-Expanded and Social Science Citation Index"
- Pcp :PCP Score,le score des cinq indicateurs divisé par le nombre du personnel permanents de l'institution.
- Year: l'année du classement (de 2005 à 2015)

3. Le classement Times :

On a 2603 observations de 14 variables :

```
'data.frame':   2603 obs. of  14 variables:
 $ world_rank           : Factor w/ 250 levels "=101","=104",..: 39 150 168 182 194 206 206 229 240 40 ...
 $ university_name      : Factor w/ 818 levels "Ä-rebro University",..: 162 60 260 419 364 536 671 528 177 810 ...
 $ country              : Factor w/ 72 levels "Argentina","Australia",..: 71 71 71 71 71 70 70 71 70 71 ...
 $ teaching             : num  99.7 97.7 97.8 98.3 90.9 90.5 88.2 84.2 89.2 92.1 ...
 $ international        : Factor w/ 804 levels "-","100.0","12.2",..: 582 409 671 162 564 627 622 263 739 454 ...
 $ research             : num  98.7 98 91.4 98.1 95.4 94.1 93.9 99.3 94.5 89.7 ...
 $ citations            : num  98.8 99.9 99.9 99.2 99.9 94 95.1 97.8 88.3 91.5 ...
 $ income               : Factor w/ 613 levels "-","100.0","24.2",..: 98 515 537 380 1 314 447 1 567 1 ...
 $ total_score          : Factor w/ 415 levels "-","41.4","41.5",..: 415 414 413 408 407 389 389 388 385 379 ...
 $ num_students         : Factor w/ 795 levels "","1,211","1,283",..: 292 281 43 169 739 249 275 577 155 61 ...
 $ student_staff_ratio  : num  8.9 6.9 9 7.8 8.4 11.8 11.6 16.4 11.7 4.4 ...
 $ international_students: Factor w/ 54 levels "","0%","1%","10%",..: 20 22 29 17 22 30 30 9 46 15 ...
 $ female_male_ratio    : Factor w/ 70 levels "","-","1 : 99",..: 1 27 31 36 39 40 40 44 31 44 ...
 $ year                 : int  2011 2011 2011 2011 2011 2011 2011 2011 2011 2011 ...
```

- world_rank : le classement de l'université à l'échelle internationale selon le système Times
- university_name : nom de l'université
- country : pays de l'université
- teaching : Score de l'université pour l'enseignement
- international : Score de l'université selon des perspectives internationales (personnel, étudiants, recherche…)

- research : Score de l'université en termes de recherche (volume, revenu, réputation)
- citations : Score de l'université pour les citations (influence de la recherche)
- income : Score de l'université pour le revenu de l'industrie (transfert des connaissances)
- total_score : Score total de l'université utilisé pour déterminer le classement
- num_students : Nombre des étudiants à l'université
- student_staff_ratio : Nombre d'étudiants divisé par le nombre d'employés
- International_students :Pourcentage des étudiants internationaux
- Female_male_ratio : Nombre des étudiantes divisé par le nombre des étudiants
- Year : L'année du classement

4. Bases de « données Expenditure » et « attainment »

Une inspection plus approfondie révèle que les deux bases ne contiennent pas de données dans les années où le classement est disponible (2012,2013,2014,2015). En effet, la base de données 'expenditure' couvre depuis l'année 1995 jusqu'à 2011 :

country	institute_type	direct_expenditure_type	X1995	X2000	X2005	X2009	X2010	X2011

Tandis que la base de données 'attainment' couvre depuis l'année 1985 jusqu'à 2015, mais ne contient aucune observation pour les années 2011-2015 :

X2004	X2005	X2006	X2007	X2008	X2009	X2010	X2011	X2012	X2013	X2015
NA	0.86	NA	NA	NA	NA	1.27	NA	NA	NA	NA

Nous concluons donc qu'elles ne nous seront pas utiles dans notre étude, et nous les utiliserons plus.

III. Nettoyage des bases de données

Avant l'analyse des bases de données, il est indispensable de les nettoyer ; en effet les données présentes peuvent avoir plusieurs types d'erreurs comme des erreurs de frappe, des informations manquantes ou des imprécisions. La partie impropre de la donnée traitée peut être remplacée, modifiée ou supprimée.

D'abord, pour assurer une meilleure compréhension des données, on utilise la fonction summary() pour observer quelques caractéristiques des différentes variables, à savoir le min, le max, le premier et le troisième quartile pour les variables quantitatives, et pour les variables qualitatives on observe le nombre de chaque modalité.

1. Le classement CWUR :

```
> summary(cwur)
   world_rank                                institution          country     national_rank    quality_of_education
Min.   :   1.0   Ã‰cole normale supÃ©rieure - Paris:   4   USA          :573   Min.   :  1.00   Min.   :  1.0
1st Qu.: 175.8   Ã‰cole Polytechnique                :   4   China        :167   1st Qu.:  6.00   1st Qu.:175.8
Median : 450.5   Arizona State University           :   4   Japan        :159   Median : 21.00   Median :355.0
Mean   : 459.6   Boston University                  :   4   United Kingdom:144   Mean   : 40.28   Mean   :275.1
3rd Qu.: 725.2   Brown University                   :   4   Germany      :115   3rd Qu.: 49.00   3rd Qu.:367.0
Max.   :1000.0   California Institute of Technology:   4   France       :109   Max.   :229.00   Max.   :367.0
                 (Other)                           :2176   (Other)      :933
 alumni_employment quality_of_faculty  publications     influence       citations      broad_impact       patents
Min.   :  1.0     Min.   :  1.0      Min.   :  1.0   Min.   :  1.0   Min.   :  1.0   Min.   :  1.0   Min.   :  1.0
1st Qu.:175.8     1st Qu.:175.8      1st Qu.:175.8   1st Qu.:175.8   1st Qu.:161.0   1st Qu.: 250.5  1st Qu.:170.8
Median :450.5     Median :210.0      Median :450.5   Median :450.5   Median :406.0   Median : 496.0  Median :426.0
Mean   :357.1     Mean   :178.9      Mean   :459.9   Mean   :459.8   Mean   :413.4   Mean   : 496.7  Mean   :433.3
3rd Qu.:478.0     3rd Qu.:218.0      3rd Qu.:725.0   3rd Qu.:725.2   3rd Qu.:645.0   3rd Qu.: 741.0  3rd Qu.:714.2
Max.   :567.0     Max.   :218.0      Max.   :1000.0  Max.   :991.0   Max.   :812.0   Max.   :1000.0  Max.   :871.0
                                                                                      NA's   :200
     score             year
Min.   : 43.36   Min.   :2012
1st Qu.: 44.46   1st Qu.:2014
Median : 45.10   Median :2014
Mean   : 47.80   Mean   :2014
3rd Qu.: 47.55   3rd Qu.:2015
Max.   :100.00   Max.   :2015
```

On remarque que seule la variable broad_impact a des valeurs manquantes. En effet, cette variable n'était pas utilisée pendant les années 2012-2013 et n'est ajouté au classement qu'en 2014. Donc si on travaille sur l'année 2012 ou 2013 on doit l'enlever.

2. Le classement Shanghai :

```
> shanghai=read.csv("shanghaiData.csv")
> summary(shanghai)
   world_rank                          university_name national_rank   total_score
 301-400: 600   Queen's University                :  13   1      : 343   Min.   : 23.50
 401-500: 600   University of Maryland, Baltimore :  12   2      : 206   1st Qu.: 27.40
 201-300: 584   Aarhus University                 :  11   3      : 133   Median : 31.30
 151-200: 300   Boston University                 :  11   4      : 122   Mean   : 36.38
 201-302: 204   Brown University                  :  11   1-2    :  86   3rd Qu.: 41.80
 101-150: 200   California Institute of Technology:  11   2-3    :  84   Max.   :100.00
 (Other):2409   (Other)                           :4828   (Other):3923   NA's   :3796
     alumni           award             hici             ns              pub
Min.   :  0.000   Min.   :  0.000   Min.   :  0.00   Min.   :  0.00   Min.   :  7.30
1st Qu.:  0.000   1st Qu.:  0.000   1st Qu.:  7.30   1st Qu.:  8.00   1st Qu.: 28.90
Median :  0.000   Median :  0.000   Median : 12.60   Median : 12.80   Median : 36.00
Mean   :  9.162   Mean   :  7.692   Mean   : 16.22   Mean   : 16.08   Mean   : 38.25
3rd Qu.: 15.600   3rd Qu.: 13.400   3rd Qu.: 21.70   3rd Qu.: 19.80   3rd Qu.: 45.30
Max.   :100.000   Max.   :100.000   Max.   :100.00   Max.   :100.00   Max.   :100.00
NA's   :1         NA's   :2         NA's   :2        NA's   :22       NA's   :2
      pcp              year
Min.   :  8.30   Min.   :2005
1st Qu.: 15.60   1st Qu.:2007
Median : 19.00   Median :2009
Mean   : 21.24   Mean   :2010
3rd Qu.: 24.50   3rd Qu.:2012
Max.   :100.00   Max.   :2015
NA's   :2
```

En utilisant la fonction *which(is.na(x)),* on constate que les variables dont il y a une ou deux valeurs manquantes appartiennent aux mêmes observations :

```
> which(is.na(Shanghai$pcp))
[1] 3798 3897
> which(is.na(Shanghai$pub))
[1] 3798 3897
> which(is.na(Shanghai$hici))
[1] 3798 3897
> which(is.na(Shanghai$award))
[1] 3798 3897
> which(is.na(Shanghai$alumni))
[1] 3897
```

Il s'agit des observations suivantes :

	world_rank	university_name	national_rank	total_score	alumni	award	hici	ns	pub	pcp	year
3798	201-300	University of Oregon	86-109	NA	9	NA	NA	NA	NA	NA	2012
3897	99			NA	NA	NA	NA	NA	NA	NA	2013

On peut donc éliminer ces deux observations de la base de données puisqu'elles n'apportent pas des données suffisantes.

A propos de la variable ns, les valeurs manquantes concernent les universités “Stockholm School of Economics” et “London School of Economics and Political Science” pour toutes les années et l’université “Tilburg University” pour l’année 2014

	world_rank	university_name	national_rank	total_score	alumni	award	hici	ns	pub	pcp	year
218	203-300	London School of Economics and Political Science	20-30	NA	19.8	0.0	15.7	NA	26.1	16.2	2005
349	301-400	Stockholm School of Economics	10-11	NA	0.0	17.1	0.0	NA	11.0	29.4	2005
719	201-300	London School of Economics and Political Science	23-33	NA	19.1	0.0	15.4	NA	25.8	28.6	2006
848	301-400	Stockholm School of Economics	10-11	NA	0.0	16.7	0.0	NA	10.8	29.4	2006
1159	151-202	London School of Economics and Political Science	16-23	NA	18.6	0.0	16.6	NA	25.3	28.4	2007
1348	305-402	Stockholm School of Economics	10	NA	0.0	16.7	0.0	NA	10.4	28.8	2007
1727	201-302	London School of Economics and Political Science	23-33	NA	17.7	0.0	16.3	NA	26.4	17.2	2008
1964	402-503	Stockholm School of Economics	10-11	NA	0.0	16.7	0.0	NA	10.3	29.2	2008
2227	201-302	London School of Economics and Political Science	24-33	NA	17.3	0.0	14.5	NA	26.1	28.5	2009
2467	402-501	Stockholm School of Economics	10-11	NA	0.0	16.7	0.0	NA	10.7	29.8	2009
2731	201-300	London School of Economics and Political Science	20-30	NA	16.9	0.0	16.1	NA	26.0	24.8	2010
2860	301-400	Stockholm School of Economics	10	NA	0.0	16.6	0.0	NA	10.7	39.0	2010
3123	102-150	London School of Economics and Political Science	11-15	NA	22.8	16.2	16.1	NA	28.0	25.9	2011
3365	301-400	Stockholm School of Economics	9-10	NA	0.0	16.6	0.0	NA	10.4	37.0	2011
3623	101-150	London School of Economics and Political Science	10-14	NA	21.0	16.8	16.2	0	28.5	26.0	2012
4007	101-150	London School of Economics and Political Science	9-17	NA	20.4	16.3	15.2	NA	30.5	28.0	2014
4341	401-500	Stockholm School of Economics	11	NA	0.0	16.3	0.0	NA	9.4	37.9	2014
4348	401-500	Tilburg University	13	NA	0.0	0.0	0.0	NA	28.1	21.2	2014
4506	101-150	London School of Economics and Political Science	10-17	NA	19.9	16.3	15.2	NA	31.2	28.7	2015
4844	401-500	Stockholm School of Economics	11	NA	0.0	16.3	0.0	NA	10.2	39.4	2015

Ces 22 valeurs manquantes ne constituent que 0.45% de la base de données ce qui reste insignifiant, on peut donc les éliminer.

Concernant la variable ***total_score,*** il n’y a que 1101 observations qui ont un total score. Après avoir visualisé la base de données, on déduit que ces observations sont les 100 premiers classés pour les 10 années (2005 à 2015)

```
> tapply(Shanghai$total_score , Shanghai$year, min , na.rm = T)
2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
23.9 23.5 23.8 24.1 23.8 24.0 24.2 24.3 24.7 24.0 23.9
```

```
Shanghai100 = subset(Shanghai, (total_score>=   23.9 & year ==2005) |
                               (total_score>=   23.5 & year ==2006) |
                               (total_score>=   23.8 & year ==2007) |
                               (total_score>=   24.1 & year ==2008) |
                               (total_score>=   23.8 & year ==2009) |
                               (total_score>=   24.0 & year ==2010) |
                               (total_score>=   24.2 & year ==2011) |
                               (total_score>=   24.3 & year ==2012) |
                               (total_score>=   24.7 & year ==2013) |
                               (total_score>=   24.0 & year ==2014) |
                               (total_score>=   23.9 & year ==2015) )
```

Les lignes de code suivantes nous ont permis de garder que les 100 premiers de chaque année.

```
> str(Shanghai100)
'data.frame':   1101 obs. of  11 variables:
```

En utilisant la fonction *str()* , on conclut que ces 1101 observations sont les 100 premiers de chaque année.

3. Le classement Times :

```
  world_rank                                    university_name
124    :  3   Ã‰cole Normale SupÃ©rieure                  :  1
132    :  3   Ã‰cole Normale SupÃ©rieure de Lyon          :  1
152    :  3   Ã‰cole Polytechnique                       :  1
174    :  3   Ã‰cole Polytechnique FÃ©dÃ©rale de Lausanne:  1
178    :  3   Aarhus University                         :  1
190    :  3   Alexandria University                     :  1
(Other):176   (Other)                                  :188
                  country      teaching     international    research
United States of America:69   Min.   :29.50   -      :  9   Min.   :28.00
United Kingdom           :27   1st Qu.:44.70   44.9   :  4   1st Qu.:44.12
Germany                  :14   Median :51.45   60.5   :  3   Median :51.50
Netherlands              :10   Mean   :54.87   62.8   :  3   Mean   :55.67
Canada                   : 9   3rd Qu.:62.23   18.4   :  2   3rd Qu.:62.80
Australia                : 7   Max.   :99.70   20.1   :  2   Max.   :99.30
(Other)                  :58                   (Other):171
  citations         income     total_score   num_students student_staff_ratio
Min.   :29.00   -      : 57   46.9   :  3   10,221 :  1   Min.   : 3.60
1st Qu.:59.12   100.0  :  7   48.5   :  3   10,410 :  1   1st Qu.:10.20
Median :70.90   27.4   :  3   49.0   :  3   10,441 :  1   Median :14.50
Mean   :71.52   40.0   :  3   51.2   :  3   10,901 :  1   Mean   :15.95
3rd Qu.:84.62   26.1   :  2   53.3   :  3   10,930 :  1   3rd Qu.:18.18
Max.   :99.90   30.5   :  2   54.4   :  3   11,074 :  1   Max.   :70.40
                (Other):120   (Other):176   (Other):188
international_students female_male_ratio      year
10%    : 10            54 : 46: 18       Min.   :2011
12%    : 10                   : 16       1st Qu.:2011
16%    :  9            52 : 48: 16       Median :2011
25%    :  9            53 : 47: 13       Mean   :2011
8%     :  9            48 : 52: 11       3rd Qu.:2011
13%    :  8            56 : 44: 10       Max.   :2011
(Other):139            (Other):110
```

Il faut tout d'abord changer les variables qui sont définies comme facteurs mais prennent des valeurs numériques en variables numériques.

On a par exemple la variable "international qui prend des valeurs numériques mais elle est définie comme un facteur. Ainsi que la variable total_score , income, female_male_ratio, num_students et international_students.

Pour transformer ces variables, on a utilisé le code suivant :

```
times11$num_students=as.numeric(paste(times11$num_students))
table(times11$num_students)
times11$international=gsub(",",".",times11$international)
table(times11$international)
times11$international=as.numeric(paste(times11$international))
table(times11$female_male_ratio)
times11$income=as.numeric(paste(times11$income))
times11$total_score=as.numeric(paste(times11$total_score))
times11$female_male_ratio=as.character(paste(times11$female_male_ratio))
times11$female_male_ratio = substr(times11$female_male_ratio,1,nchar(times11$female_male_ratio)-1)
times11$female_male_ratio=as.numeric(paste(times11$female_male_ratio))
times11$international_students=gsub("%","",times11$international_students)
times11$international_students=as.numeric(paste(times11$international_students))
```

Pour la variable num_students , elle contient des virgules donc avant de la transformer en une variable numérique il faut changer la virgule en point.

La variable female_male_ratio s'écrit sous la forme : "23 :77". Le nombre 23 représente le pourcentage des femmes et le nombre 77 représente le pourcentage des hommes. Donc pour pouvoir transformer ce ratio en une valeur numérique, on a supprimé les trois derniers charactères de la variable (:77) afin d'avoir le pourcentage des femmes, ensuite, on a utilisé as.numeric(paste(variable)) pour convertir les facteurs en variables numériques.

```
> summary(times11)
    teaching      international       research       citations         income
 Min.   :29.50   Min.   : 15.90   Min.   :28.00   Min.   :29.00   Min.   : 26.10
 1st Qu.:44.70   1st Qu.: 31.70   1st Qu.:44.12   1st Qu.:59.12   1st Qu.: 34.20
 Median :51.45   Median : 53.30   Median :51.50   Median :70.90   Median : 42.50
 Mean   :54.87   Mean   : 54.33   Mean   :55.67   Mean   :71.52   Mean   : 52.09
 3rd Qu.:62.23   3rd Qu.: 72.80   3rd Qu.:62.80   3rd Qu.:84.62   3rd Qu.: 61.50
 Max.   :99.70   Max.   :100.00   Max.   :99.30   Max.   :99.90   Max.   :100.00
                 NA's   :9                                        NA's   :57
  total_score     num_students     student_staff_ratio international_students
 Min.   :46.20   Min.   :  2.218   Min.   : 3.60       Min.   : 1.00
 1st Qu.:51.33   1st Qu.: 13.008   1st Qu.:10.20       1st Qu.:11.00
 Median :56.95   Median : 22.404   Median :14.50       Median :17.00
 Mean   :60.52   Mean   : 24.168   Mean   :15.95       Mean   :19.06
 3rd Qu.:65.97   3rd Qu.: 30.620   3rd Qu.:18.18       3rd Qu.:25.00
 Max.   :96.10   Max.   :127.431   Max.   :70.40       Max.   :54.00
                                                       NA's   :1

 female_male_ratio
 Min.   :13.00
 1st Qu.:47.00
 Median :52.00
 Mean   :49.43
 3rd Qu.:55.00
 Max.   :70.00
 NA's   :16
```

On constate l'apparition des NAs et cela est dû au fait que quelques variables contenaient "-" ou bien " " comme facteurs. Ces valeurs manquantes ne constituent que 2% de la base de données ce qui reste insignifiant, on peut donc les éliminer.

IV. Statistique descriptive des données

1. Méthodologie de travail :

Afin de mener notre étude le plus correcte possible, on propose la démarche de travail suivante qu'on appliquera à toutes les bases de données :

1. Aperçu global sur les variables :

Les bases de données contiennent deux types de variables : Quantitatives et qualitatives. Pour mieux maitriser la data et bien comprendre sa structure, on se servira de maints outils pour afficher les principaux indicateurs pour les variables quantitatives, et pour les variables qualitatives on affichera le nombre de chaque modalité.

2. Aperçu détaillé de la variable dépendante : On se focalisera sur ses propriétés.
3. La variable dépendante en fonction des autres variables :

Maintenant qu'on a plus ou moins une idée sur toutes les variables, nous allons représenter la variable dépendante en fonction des autres pour pouvoir tirer plus de conclusions.

4. Matrices de corrélation linéaires :

Pour évaluer le niveau de dépendance entre les variables, on calculera le coefficient de corrélation linéaire entre eux (coefficient de Pearson).

5. Loi descriptive de la variable dépendante :

Dans cette section, on cherche à préciser le comportement de la variable score. Pour cela, La technique des tests d'hypothèse en statistique inférentielle donne une réponse adaptée aux situations d'échantillonnage. Elle consiste à tester si une loi théorique de probabilité peut représenter au mieux la distribution des fréquences

des valeurs prises par le score dans un échantillon d'éléments prélevés au hasard. De ce fait, nous allons choisir plusieurs lois en se basant sur :

-Les propriétés graphiques : En utilisant l'appréciation visuelle du graphique de la densité, on peut limiter les choix.

-Les paramètres de formes : Un paramètre de forme est un type de paramètre régissant une famille paramétrique de lois de probabilité. Un tel paramètre régit seulement la forme de la distribution.

-Le cœfficient de dissymétrie : (skewness en Anglais) est un moment standardisé qui mesure l'asymétrie de la densité de probabilité d'une variable aléatoire définie sur les nombres réels. En termes généraux, l'asymétrie d'une distribution est positive si la queue de droite (à valeurs hautes) est plus longue ou grosse, et négative si la queue de gauche (à valeurs basses) est plus longue ou grosse. Dans le cas d'une distribution normale, le coefficient est nul. La distribution est symétrique.

Théoriquement ce coefficient est égal à :

$$\frac{E[(X-m)^3]}{\sigma^3}$$

-Le kurtosis (mot d'origine grec), plus fréquemment traduit par cœfficient d'aplatissement, ou cœfficient d'aplatissement de Pearson, correspond à une mesure de l'aplatissement, ou a contrario de la pointicité, de la distribution d'une variable aléatoire réelle. C'est la seconde des caractéristiques de forme, avec le cœfficient de dissymétrie. Elle mesure, hors effet de dispersion (donnée par l'écart-type), la disposition des masses de probabilité autour de leur centre, tel que donné par l'espérance mathématique, c'est-à-dire d'une certaine façon, leur regroupement proche ou loin du centre de probabilité. Pour une variable aléatoire suivant une loi normale, ce cœfficient d'aplatissement vaut 3.

Théoriquement ce coefficient est égal à :

$$\frac{E[(X-m)^4]}{\sigma^4}$$

Avec m l'espérance de la distribution et sigma sa variance

On peut calculer ces coefficients sur R en utilisant la syntaxe suivante :

```
install.packages("moments")
library(moments)
skewness(Shanghai100_2005$total_score)
kurtosis(Shanghai100_2005$total_score)
```

Une fois la loi choisie, on effectuera les tests d'ajustements en utilisant les fonctions disponibles dans la librairie "goftest".

2. Le classement CWUR :

a. Aperçu global sur les variables :

En ce qui concerne les variables qualitatives on peut représenter le nombre de chaque modalité dans les diagrammes en camembert ci-dessous en utilisant la fonction PIE3D:

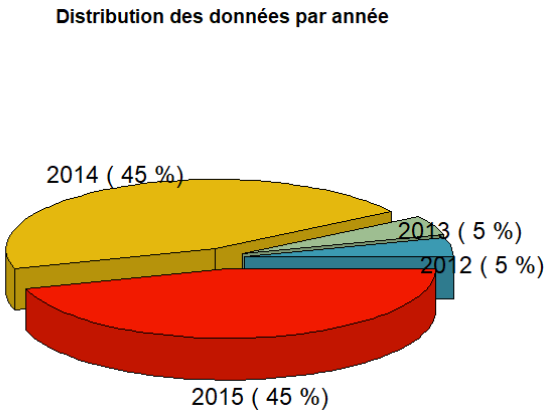


Figure 1: diagrammes en camembert de la distribution des données par année (CWUR)

On remarque que 90% des observations sont dans les années 2014, 2015, et ce car on a ajouté plusieurs universités au classement pendant ces années.

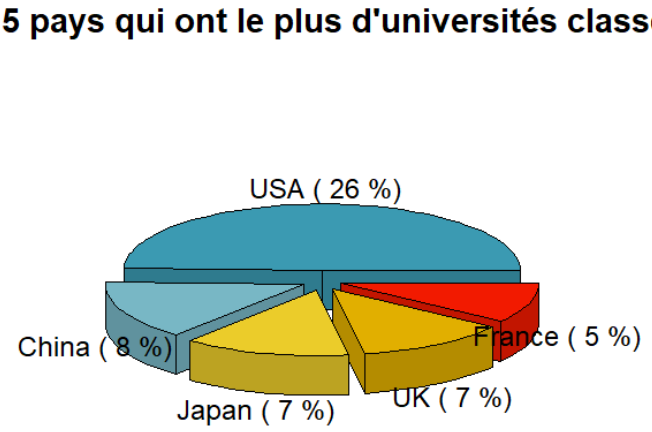


Figure 2:diagrammes en camembert des 5 pays les plus présents (CWUR)

Si on classe les premiers 5 pays par nombre d'observations, on constate que les universités américaines sont les plus présentes dans le classement.

Le graphe ci-dessous représente les cinq premières universités selon le classement CWUR durant les 4 années 2012-2013-2014-2015 :

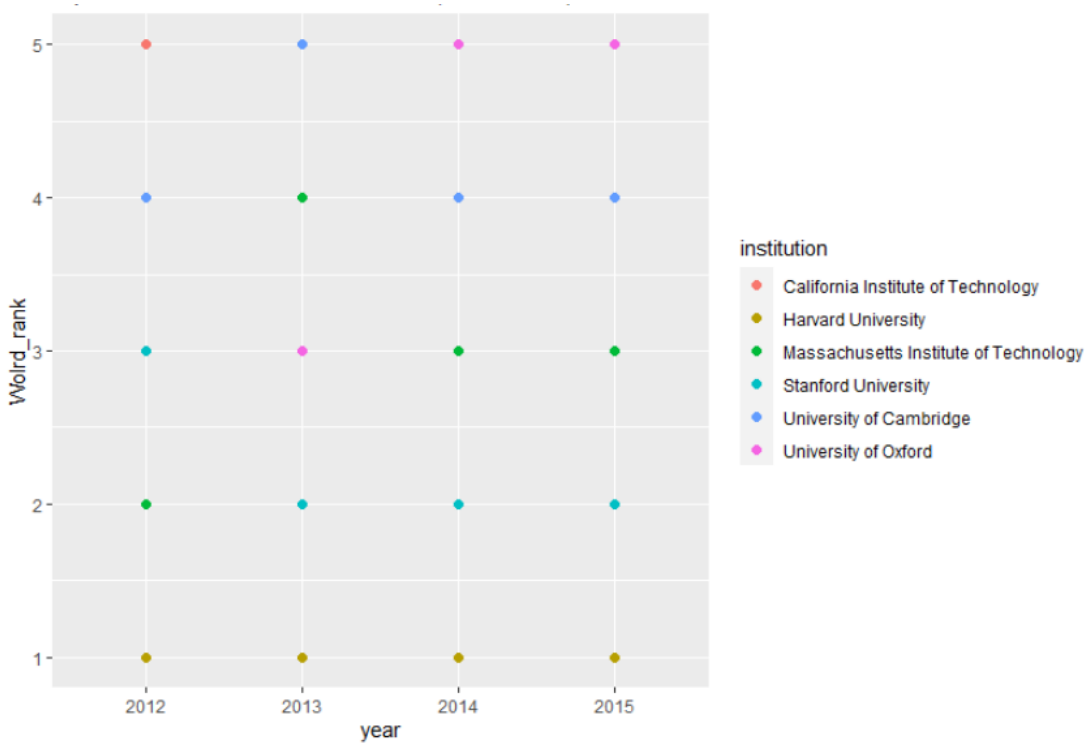


Figure 3: Les 5 premières universités (CWUR 2012-2015)

Les 5 universités les mieux classées selon le classement CWUR sont presque les mêmes pour les 4 années entre 2012-2015. Deux pays sont représentés dans cette liste : Le Royaume-Uni et les Etats-Unis, avec une majorité d'appartenance à ce dernier. La première place est toujours décrochée par l'université de Harvard, tandis que la deuxième et la troisième place sont en gros partagées par l'université de Stanford et MIT.

En utilisant la même fonction qu'avant (summary()), nous pouvons afficher les principaux indicateurs de nos variables quantitatives à savoir, le min, le max, la médiane, la moyenne, le premier et le troisième quartile, par année. Par exemple, pour l'année 2015 :

```
> cwur2015 = subset(cwur, year == 2015)
> summary(cwur2015)
   world_rank                                 institution           country    national_rank
 Min.   :   1.0   Ã-rebro University                :  1   USA           :229   Min.   :  1.00
 1st Qu.: 250.8   Ã…bo Akademi University           :  1   China         : 83   1st Qu.:  7.00
 Median : 500.5   Ã‰cole centrale de Lyon           :  1   Japan         : 74   Median : 22.00
 Mean   : 500.5   Ã‰cole Centrale Paris             :  1   United Kingdom: 65   Mean   : 42.51
 3rd Qu.: 750.2   Ã‰cole normale supÃ©rieure - Paris :  1   Germany       : 55   3rd Qu.: 52.00
 Max.   :1000.0   Ã‰cole normale supÃ©rieure de Cachan:  1   France        : 49   Max.   :229.00
                  (Other)                          :994   (Other)       :445
 quality_of_education alumni_employment quality_of_faculty  publications      influence
 Min.   :  1.0        Min.   :  1.0     Min.   :  1.0      Min.   :   1.0   Min.   :  1.0
 1st Qu.:250.8        1st Qu.:250.8     1st Qu.:218.0      1st Qu.: 250.8   1st Qu.:250.8
 Median :367.0        Median :500.5     Median :218.0      Median : 500.5   Median :500.5
 Mean   :299.8        Mean   :406.5     Mean   :194.3      Mean   : 500.4   Mean   :500.3
 3rd Qu.:367.0        3rd Qu.:567.0     3rd Qu.:218.0      3rd Qu.: 750.0   3rd Qu.:750.2
 Max.   :367.0        Max.   :567.0     Max.   :218.0      Max.   :1000.0   Max.   :991.0

   citations      broad_impact       patents          score            year
 Min.   :  1.0   Min.   :   1.0   Min.   :  1.0   Min.   : 44.02   Min.   :2015
 1st Qu.:234.0   1st Qu.: 250.0   1st Qu.:250.8   1st Qu.: 44.30   1st Qu.:2015
 Median :428.0   Median : 495.0   Median :500.5   Median : 44.78   Median :2015
 Mean   :451.3   Mean   : 496.7   Mean   :491.7   Mean   : 46.86   Mean   :2015
 3rd Qu.:645.0   3rd Qu.: 741.0   3rd Qu.:749.0   3rd Qu.: 46.54   3rd Qu.:2015
 Max.   :812.0   Max.   :1000.0   Max.   :871.0   Max.   :100.00   Max.   :2015
```

Toutes les variables quantitatives (sauf le score, qu'on étudiera en plus de détails par la suite) varient entre un minimum de 1 et un maximum qui ne dépasse pas 1000, et ce parce qu'elles expriment un classement de 1000 observations. Or le maximum peut ne pas atteindre 1000 car on peut trouver plusieurs universités ayants le même classement. On trouve plus ou moins les mêmes indicateurs pour les autres années.

b. Le score :

Pour commencer, on essaie de comprendre la distribution de notre variable dépendante : le score. Pour éviter la redondance, on regroupe d'abord les données par année puis on trace les histogrammes du score.

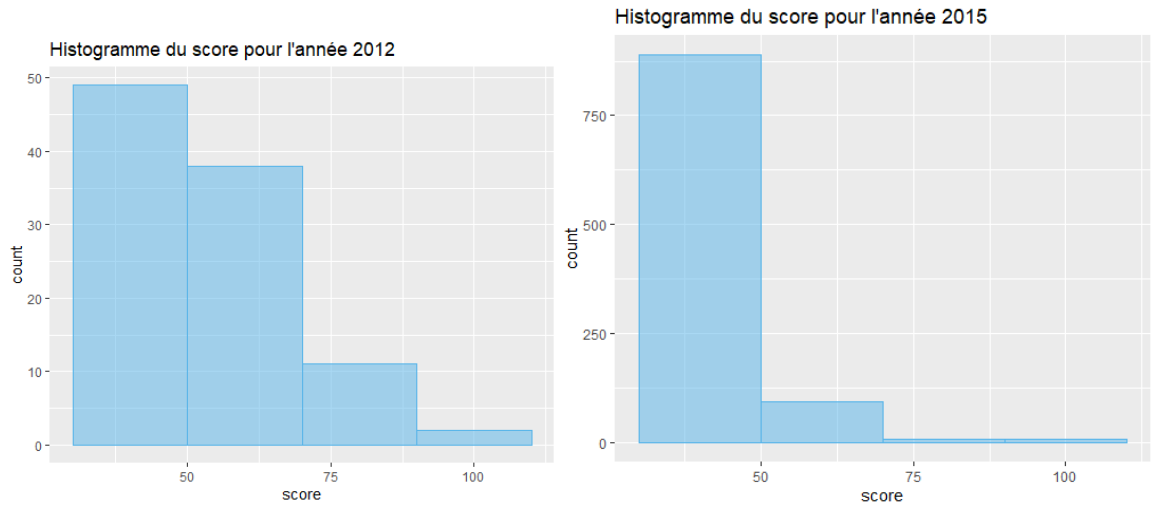


Figure 4: Histogramme du score (CWUR ;2012-2015)

Les histogrammes qu'on retrouve sont biaisés vers le droit, ou positivement désaxés ce qui peut être expliqué par l'existence des valeurs très grandes qui augmentent la valeur de la moyenne, mais n'ont aucun effet sur la médiane. On remarque aussi que la densité des valeurs du score 40-60 est la plus grande, donc la plupart des universités ont des scores dans cette gamme.

Les boîtes à moustaches ci-dessous du score selon l'année confirment nos constatations sur leurs distributions.

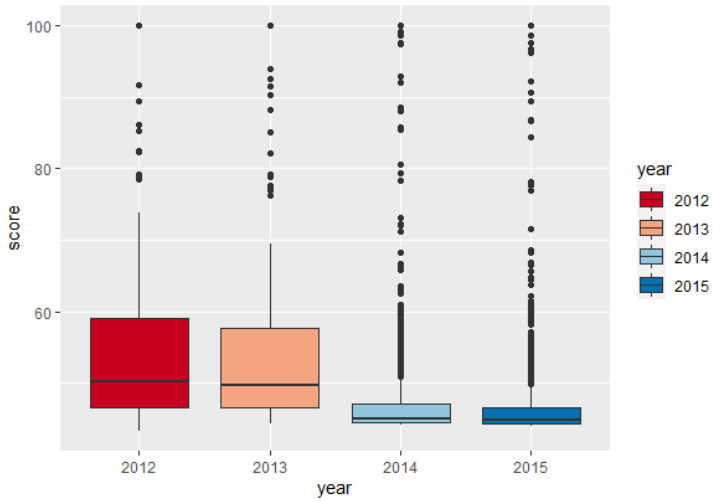


Figure 5: Boîte à moustache du score selon l'année (CWUR)

Une représentation encore plus explicative de cette donnée est une représentation de sa distribution géographique. A l'aide de la librairie 'rworldmap', on peut visualiser cette dernière :

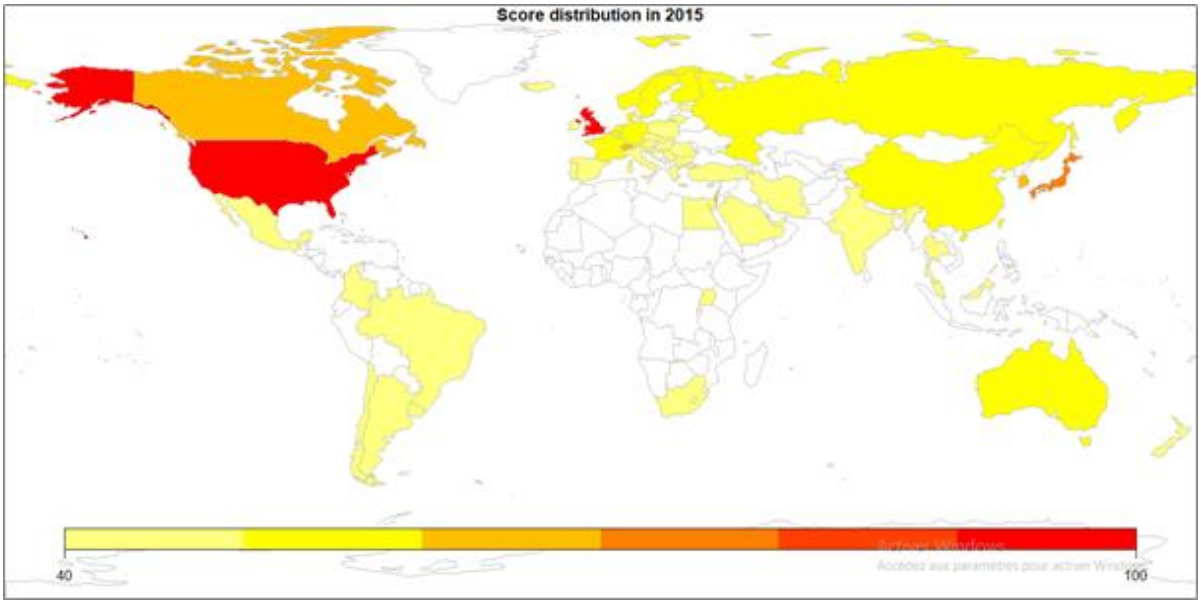


Figure 6: Distribution géographique du score (CWUR;2015)

On remarque que les universités les mieux classées (en rouge) se retrouvent en Amérique du nord et en Angleterre. La majorité des écoles, comme le confirme les histogrammes, ont un score dans la gamme 40-60 et sont répartis partout dans le monde.

c. La variable score en fonction des autres variables

Voici deux exemples de la distribution du score en fonction des variables 'quality_of_education' et 'national_rank' :

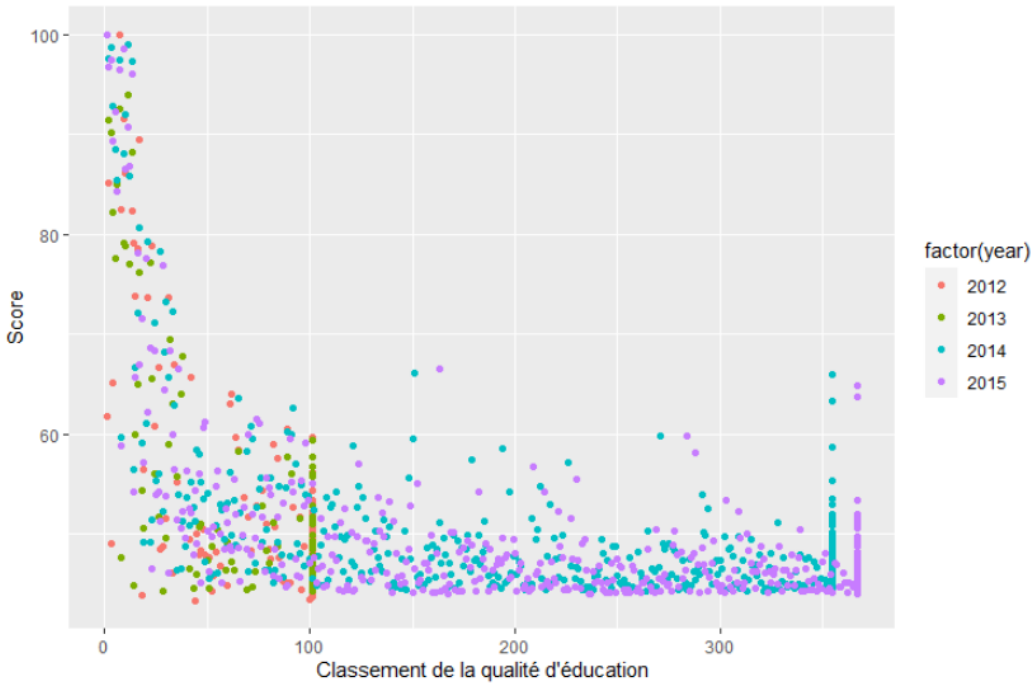


Figure 7: Score en fonction de la qualité d'éduction (CWUR)

On remarque qu'en général, et à travers les années, mieux la qualité d'éducation dans le pays, mieux le score de l'université. Il importe de signaler qu'il existe des exceptions de pays pour lesquels même s'ils sont bien classés pour la qualité d'éducation, leurs universités ne sont pas très bien classées. Si on examine ce sous-groupe on se rend compte très rapidement que même si le rang de la qualité d'éducation peut être élevé, les autres indicateurs sont très mauvais, ce qui baisse le score.

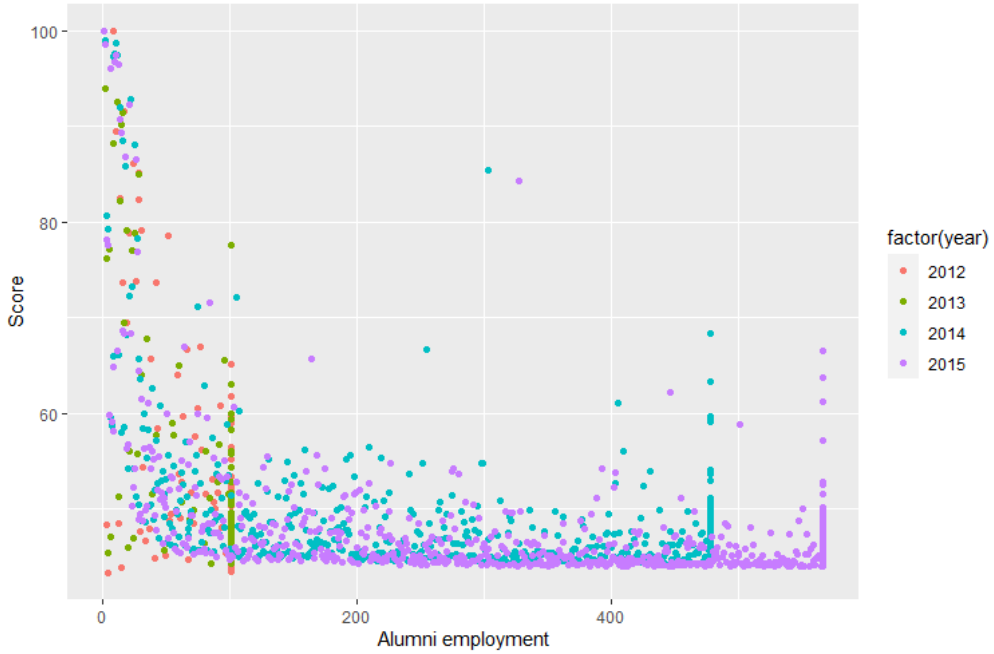
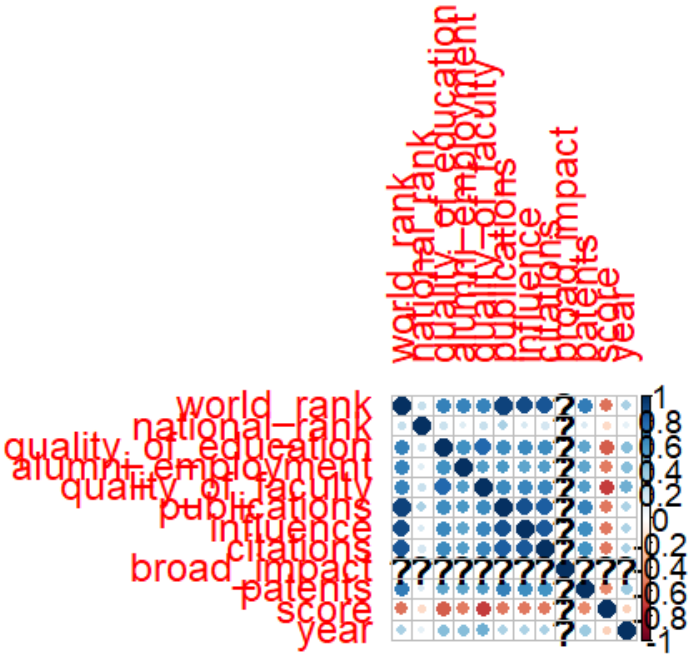


Figure 8:Score en fonction de l'embauche des anciens (CWUR)

On constate la même forme de distribution à peu près, et c'est d'ailleurs le cas avec toutes les autres variables.

d. Matrices de corrélation :

On dispose de 14 variables, donc le résultat obtenu en traçant la matrice de corrélation n'est ni clair ni pratique :



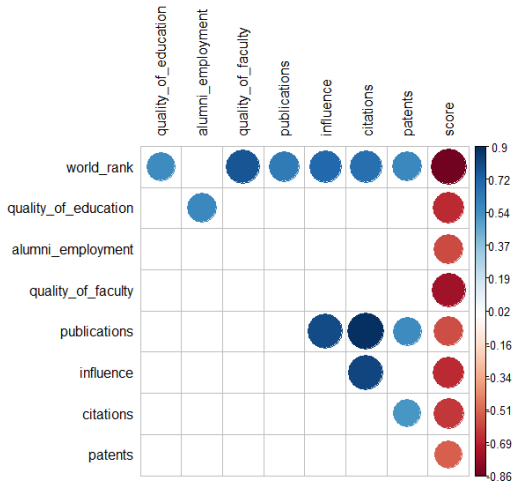
On utilise donc une fonction qui permet d'isoler les variables avec une corrélation supérieure à un seuil (définit par nous à 0.5), en suivant les étapes suivantes :

- Conversion des variables en variables numériques
- Suppression des valeurs doubles
- Suppression des corrélations parfaites (corrélation d'une variable avec elle-même)
- Affichage d'un tableau contenant les corrélations classées en ordre décroissant
- Sélection des variables avec une corrélation supérieure à 0.5
- Affichage des résultats

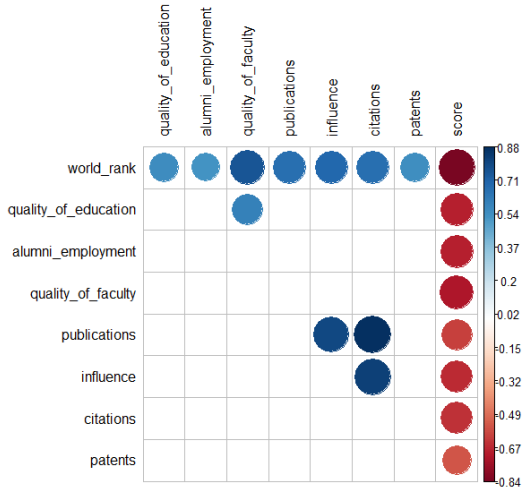
```r
corr_simplifie <- function(data=df,sig=0.5){
  #Conversion des variables en variables numériques
  df_cor <- data %>% mutate_if(is.character, as.factor)
  df_cor <- df_cor %>% mutate_if(is.factor, as.numeric)
  #Calcule de la correlation
  corr <- cor(df_cor)
  #préparation
  corr[lower.tri(corr,diag=TRUE)] <- NA
  #Supression des correlations parfaites
  corr[corr == 1] <- NA
  #Conversion en table
  corr <- as.data.frame(as.table(corr))
  #Suprresion des valeurs NA
  corr <- na.omit(corr)
  #Selection des valeurs significatives
  corr <- subset(corr, abs(Freq) > sig)
  #Classement en ordre décroissant
  corr <- corr[order(-abs(corr$Freq)),]
  #Affiche de la table du classement
  print(corr)
  #Préparation pour le dessin
  mtx_corr <- reshape2::acast(corr, Var1~Var2, value.var="Freq")
  #Dessin de la matrice de correlation
  corrplot(mtx_corr, is.corr=FALSE, tl.col="black", na.label=" ")
}
```

Le résultat, par année est le suivant :

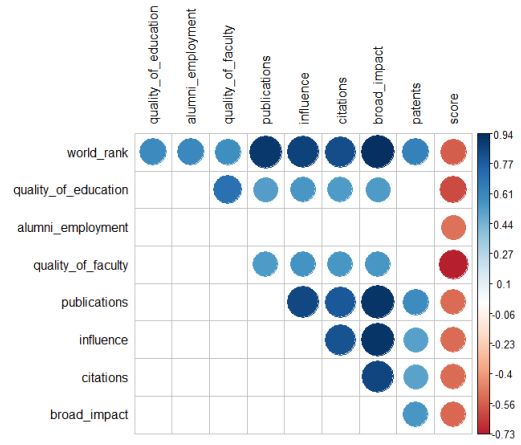
2012



2013



2014



2015

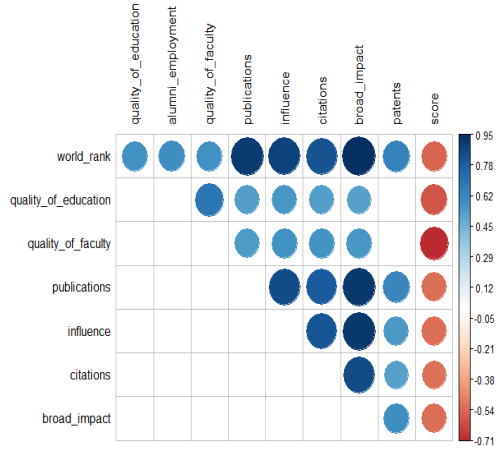


Figure 9: Matrices de corrélation (CWUR)

Le score est significativement corrélé avec 7 variables : 'world_rank' (ce qui n'est pas surprenant vu que ce dernier est calculé à base du score), 'quality_of_education', 'alumni_employment', 'quality_of_faculty', 'publications', 'influence', 'citations' et 'patent' pour les années 2012,2013 et 'broad_impact' pour les années 2014,2015. En effet, la variable 'broad_impact' comme mentionné auparavant n'est introduite qu'en 2014. Cela paraît logique vu que ce sont des facteurs important pour avoir un milieu universitaire favorable. Il y a également une différence légère entre les années ce qui implique que peut être le modèle évolue chaque année.

On remarque aussi une forte corrélation entre quelques variables, ce qui peut poser problème après.

e. Loi descriptive de la variable Score :

Pour avoir une idée sur les modèles à utiliser pour l'ajustement, on trace la densité empirique de la variable score pour toutes les années :

```
> x2 <- cwur2012$score
> x3 <- cwur2013$score
> x4 <- cwur2014$score
> x5 <- cwur2015$score
> plot(density(x5), col="darkred", lwd=2)
> lines(density(x4),col="blue", lwd =1)
> plot(density(x3), col="darkred", lwd=2)
> lines(density(x2),col="blue", lwd =1)
>
```

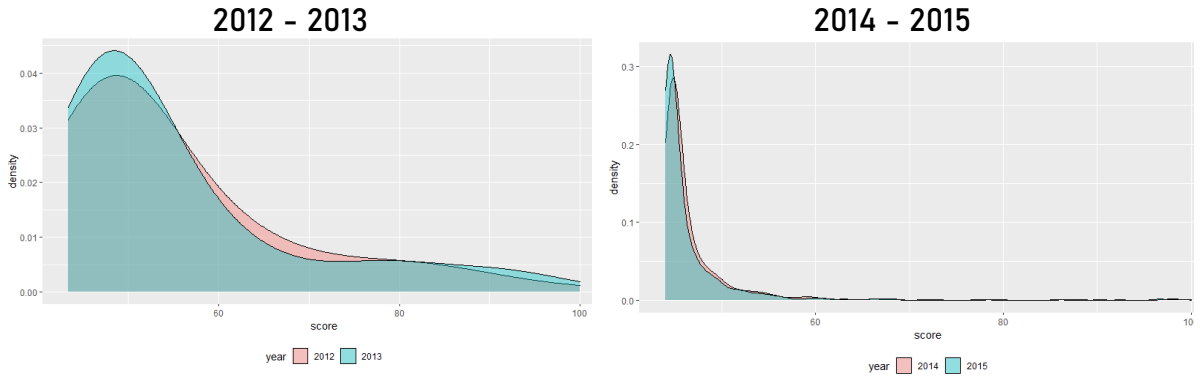


Figure 10:Densité empirique du score (CWUR)

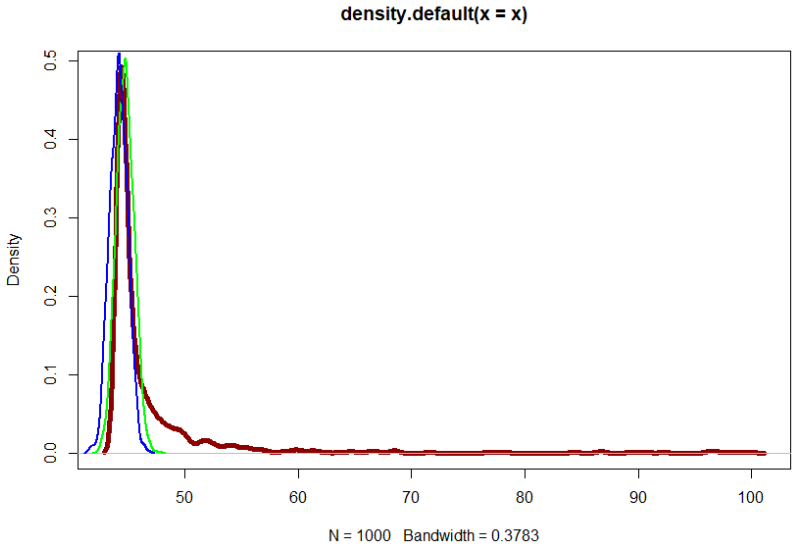
Les années sont regroupées par nombres d'observations (100 obs : 2012 et 2013 / 1000 obs :2014 et 2015). On remarque que la distribution est à peu près la même, donc on se limitera à une seule année, par exemple 2015.

Sur le graphique, on remarque que la fonction densité est désaxée et positive, on exclut par conséquence les lois qui ne répondent pas à ces critères (loi normale par exemple), et on se limite à deux lois : Loi Log-normale et Gamma.

Pour estimer les paramètres de ces lois, on peut se baser sur ceux de la variable dépendante. En utilisant la librairie (moments) susmentionnée, on trouve la valeur de l'asymétrie ainsi que l'aplatissement :

```
> library(moments)
Warning message:
le package 'moments' a été compilé avec la version R 3.5.2
> skewness(x5)
[1] 5.226558
> kurtosis(x5)
[1] 35.33382
>
```

En se basant sur ces valeurs, on estime les paramètres des fonctions Log-normale et Gamma et ci-dessous les meilleures approximations trouvées (Log-normale en vert, Gamma en bleu)



:

Figure 11:Densité empirique et théorique du score (CWUR;2015)

```
#Fonction lognormale
d <- rlnorm(1000, mean=3.800408, sdlog =0.01794214)

#Fonction Gamma
r <- rgamma(1000, shape=3090, rate = 70)
```

Après, on utilise les fonctions disponibles dans la librairie "goftest" pour effectuer les tests d'ajustements et comparer entre les deux lois. On effectuera 3 tests : Kolmogorov-Smirnov, Cramer-Von Mises, et Chi 2.

	Log-normale	Gamma

KS	`> d <- rlnorm(1000, mean=3.800408, sdlog =0.01794214)`
	`> ks.test(x5, d)`
	
	`Two-sample Kolmogorov-Smirnov test`
	
	`data: x5 and d`
	`D = 0.253, p-value < 2.2e-16`
	`alternative hypothesis: two-sided`
	
	`Warning message:`
	`In ks.test(x5, d) :`
	`les valeurs p seront approximées en présence d'ex-aequos`
		`> r <- rgamma(1000, shape=3090, rate = 70)`
		`> ks.test(x5, r)`
		
		`Two-sample Kolmogorov-Smirnov test`
		
		`data: x5 and r`
		`D = 0.444, p-value < 2.2e-16`
		`alternative hypothesis: two-sided`
		
		`Warning message:`
		`In ks.test(x5, r) :`
		`les valeurs p seront approximées en présence d'ex-aequos`

CVM	`> cvm.test(x5,null = "pgamma", shape=3090, rate = 70)`
	
	`Cramer-von Mises test of goodness-of-fit`
	`Null hypothesis: Gamma distribution`
	`with parameters shape = 3090, rate = 70`
	`Parameters assumed to be fixed`
	
	`data: x5`
	`omega2 = 85.685, p-value < 2.2e-16`
		`> cvm.test(x5,null = "plnorm", mean=3.800408, sdlog =0.01794214)`
		
		`Cramer-von Mises test of goodness-of-fit`
		`Null hypothesis: log-normal distribution`
		`with parameter sdlog = 0.01794214`
		`Parameters assumed to be fixed`
		
		`data: x5`
		`omega2 = 15.756, p-value < 2.2e-16`

CHI2	`> d <- rlnorm(1000, mean=3.800408, sdlog =0.01794214)`
	`> tbl2 = table(x5, d)`
	`> chisq.test(tbl2)`
	
	`Pearson's Chi-squared test`
	
	`data: tbl2`
	`X-squared = 415000, df = 414580, p-value = 0.3241`
		`> tbl = table(x5,r)`
		`> chisq.test(tbl)`
		
		`Pearson's Chi-squared test`
		
		`data: tbl`
		`X-squared = 415000, df = 414580, p-value = 0.3241`

Figure 12: Test d'ajustement pour la loi du score (CWUR;2015)

Voici un tableau résumant les comparaisons :

Test	Critère	Loi Log-normale	Loi gamma	Loi à retenir
K-S	D minimal (idéalement inférieur à la valeur critique 0.0136)	D = 0.25175	D = 0.418	Log-normale
CVM	Omega2 minimal	Omega2 = 15.756	Omega2 = 85.685	Log-normale
CHI-2	Statistique Chi2 minimale	Même résultat		Incertain

On peut conclure que la loi la plus adéquate est la loi Log-normale.

3. Le classement Times :

a. Aperçu global sur les variables :

Cette base de données contient aussi deux types de variables : Quantitatives et qualitatives.

En procédant de la même manière, on retrouve les diagrammes suivants :

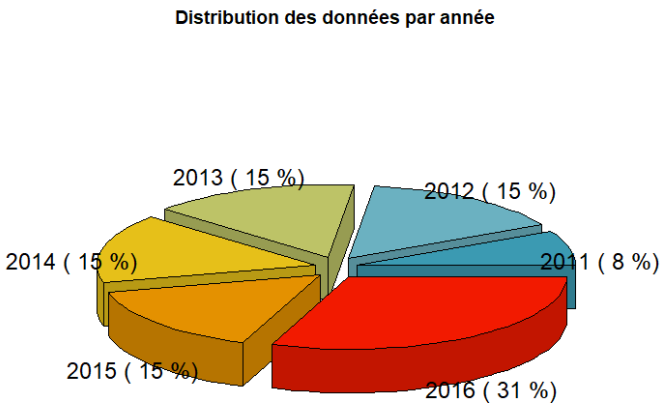


Figure 13:diagrammes en camembert de la distribution des données par année (Times)

On constate que le nombre des observations est le même dans les années 2012, 2013,2014 et 2015, par contre il est le double dans l'année 2016 et la moitié dans l'année 2011.

Si on classe les 10 premiers pays par nombre d'observations, on constate que les universités américaines sont les plus présentes dans le classement

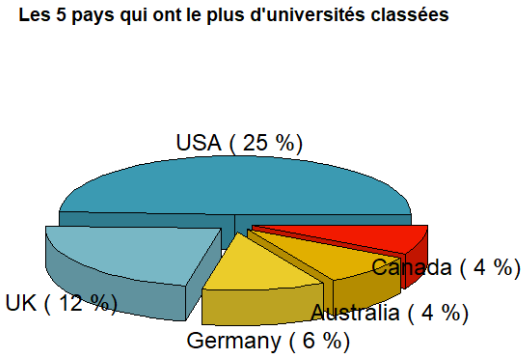


Figure 14:diagrammes en camembert des 5 pays les plus présents (Times)

On représente de même les 5 universités les mieux classés durant les années on trouve :

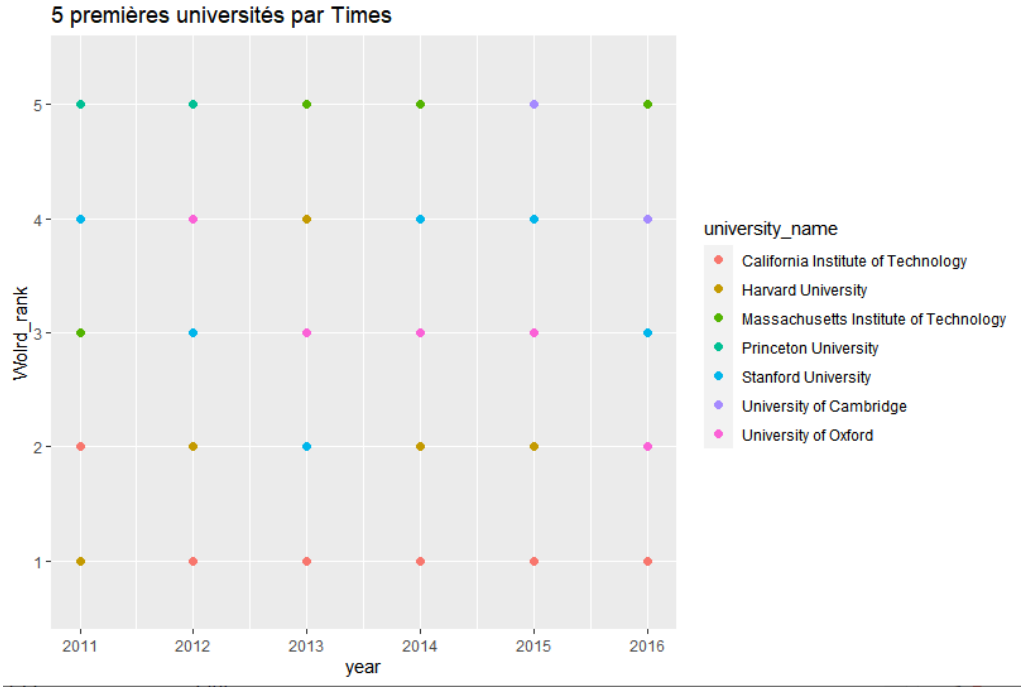


Figure 15: les 5 premières universités (TIMES 2011-2015-2016)

Les 5 universités les mieux classés dans le monde changent légèrement durant les années. Deux pays sont représentés dans cette liste : Etats-Unis et Royaume-Uni.

b. Le score :

On commence par l'histogramme du score :

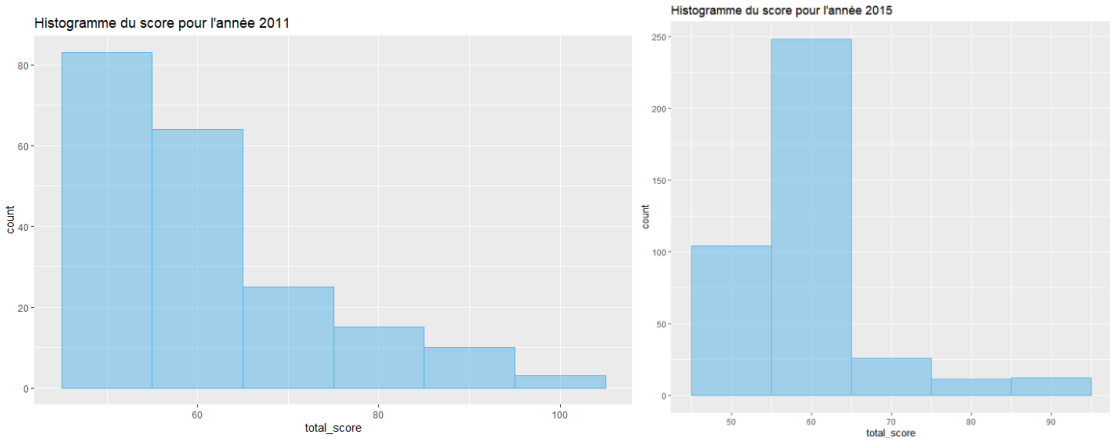


Figure 16:Histogrammes du score (Times;2011-2015)

Conformément à nos attentes, l'histogramme qu'on retrouve ci-dessus est similaire à celui retrouvé auparavant (figure 3-4), c'est-à-dire biaisé vers la droite et donc le score contient des valeurs qui augmentent la valeur de la moyenne, sans avoir un impact sur la médiane. On remarque aussi que la densité des valeurs du score 40-60 est toujours la plus grande, donc la plupart des universités ont des scores dans cette gamme. Les boîtes à moustaches ci-dessous du score selon l'année confirment nos constatations sur leurs distributions :

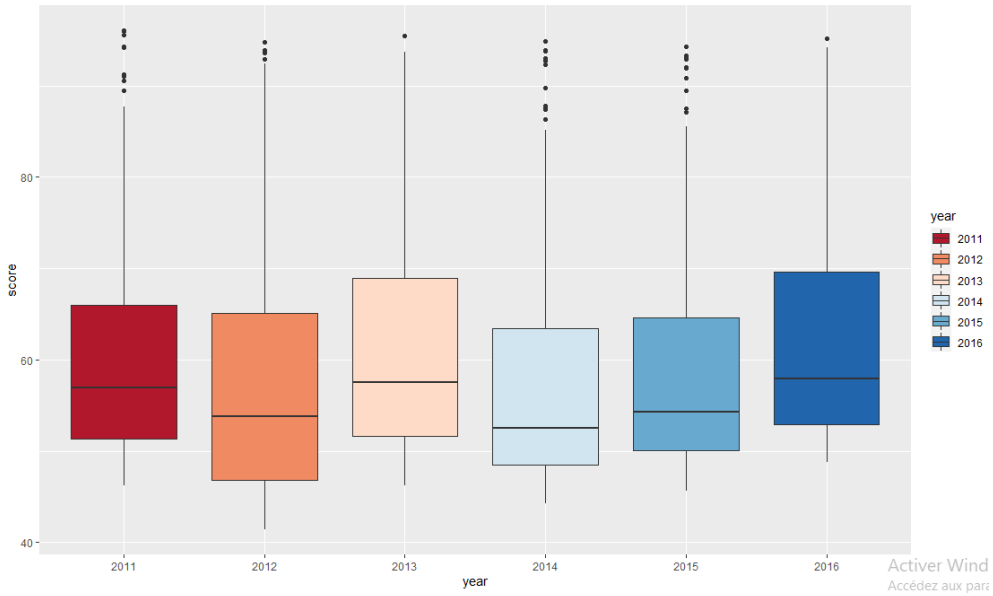


Figure 17: Boîte à moustache du score selon l'année (Times)

Ensuite, on représente la distribution géographique du score pour la même année :

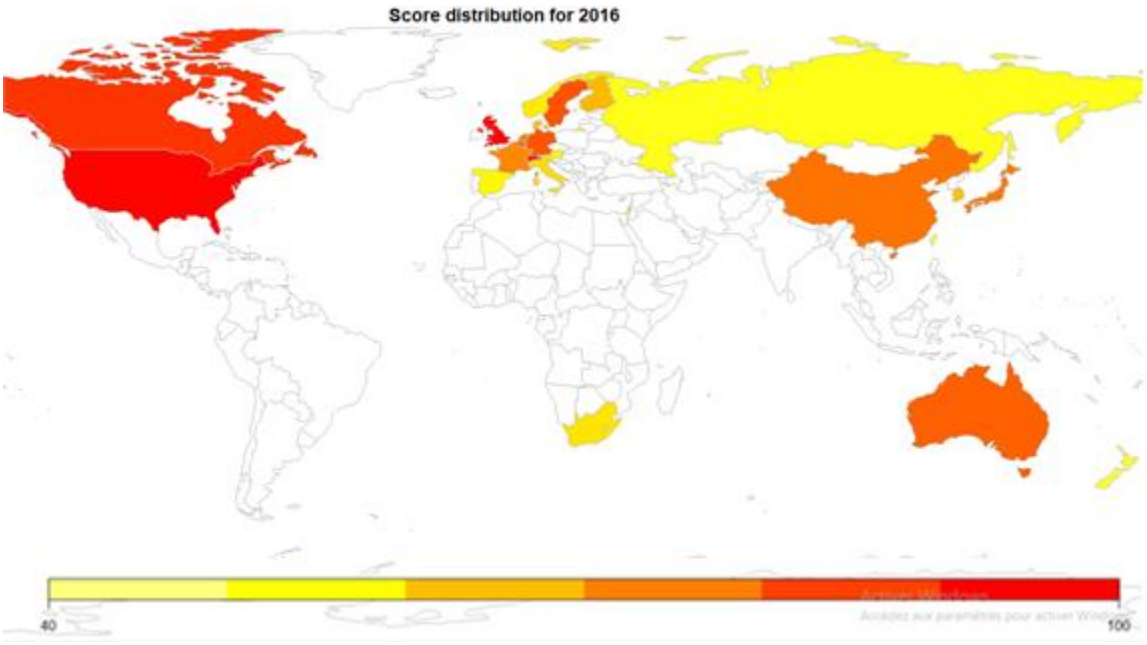


Figure 18:Distribution géographique du score (Times;2016)

On remarque que les universités les mieux classés se trouvent aux États-Unis, suivi par l'Angleterre et le Canada, puis on trouve des pays Européens comme l'Allemagne et le Suède, l'Australie et l'Asie.

c. La variable score en fonction des autres variables

Afin d'analyser la relation entre le score total et les autres scores pour la recherche, la qualité d'éducation et l'influence de la recherche (citations) on réalise les courbes suivants :

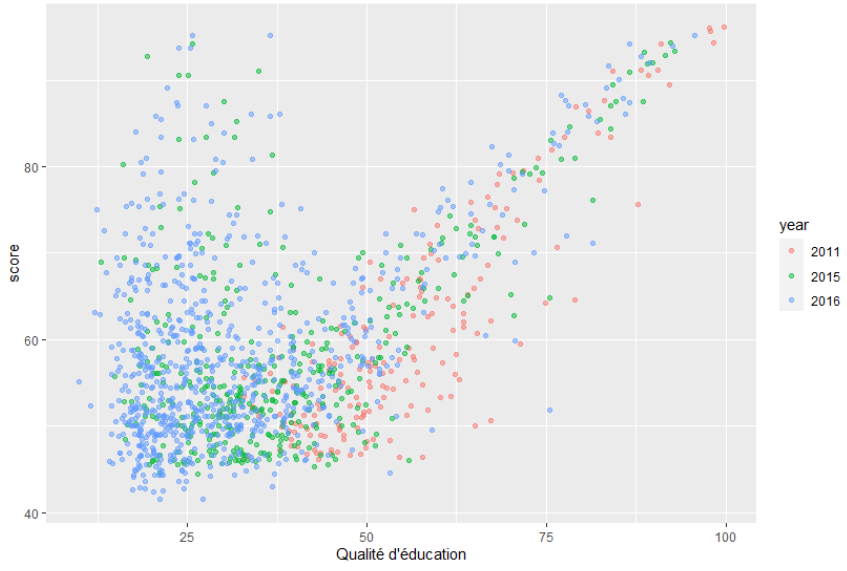


Figure 19: Score en fonction de la qualité d'éduction (Times)

On constate qu'en général, et à travers les années, les universités les moins classés sont ceux qui ont un score pour la qualité d'éducation inférieure par rapport aux autres.

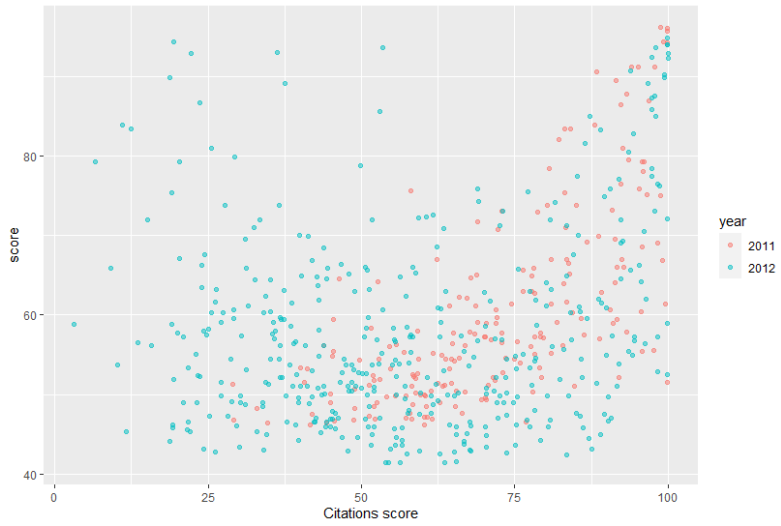


Figure 20: Score en fonction des citations (Times)

Concernant les citations on remarque que mieux la qualité d'éducation dans le pays, mieux le score des citations. Il importe de signaler qu'il existe des exceptions de pays pour lesquels même s'ils sont bien classés pour les citations, leurs universités ne sont pas très bien classées.

Figure 21 : Score en fonction du score de la recherche (Times)

Similairement au premier graphe de la qualité d'éducation on voit que les universités moins classés, on un score de recherche minimale par rapport aux autres.

d. Matrices de corrélation :

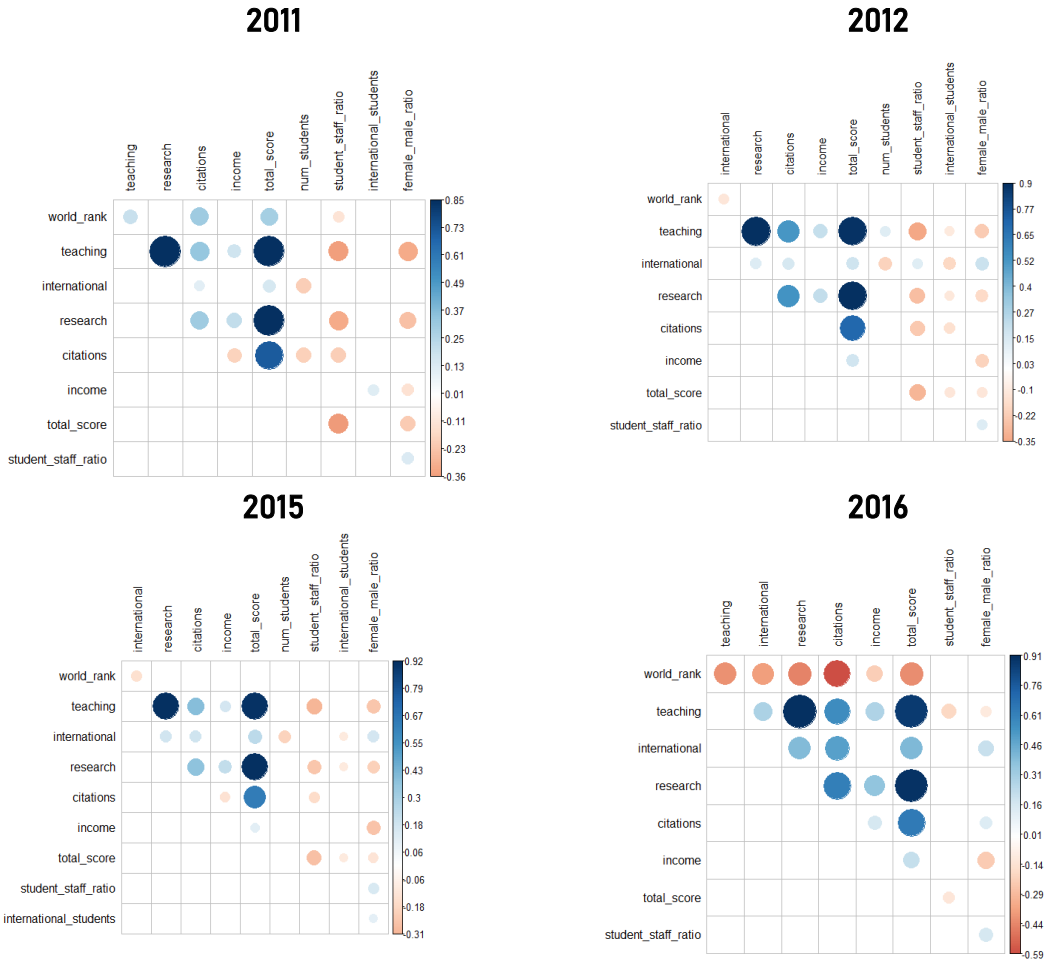


Figure 22:Matrices de corrélation (Times)

En comparant les matrix de corrélation, on constate que celle de l'année 2016 est différente des autres années. Pour cette année, on trouve que total_score est fortement corrélé avec research, teaching, citations :

```
research        total_score  0.8936362
teaching        total_score  0.8676777
citations       total_score  0.6349077
```

De même pour l'année 2011:

```
teaching        total_score  0.8501127
research        total_score  0.8489652
citations       total_score  0.7115317
```

Et l'année 2015:

```
research        total_score  0.9153093
teaching        total_score  0.8992874
citations       total_score  0.6345212
```

e. Loi descriptive de la variable Score :

On commence tout d'abord en traçant la densité empirique de la variable totale_score pour quelques années puisque sa distribution est presque la même. On

peut confirmer cela en regardant les courbes du score pour plusieurs années en utilisant le code suivant :

```
x1=as.numeric(paste(times11$total_score))
x2=na.omit(as.numeric(paste(times13$total_score)))
x3=na.omit(as.numeric(paste(times15$total_score)))
plot(density(x1), col="Red",lwd=2)
lines(density(x2), col="Blue",lwd=2)
lines(density(x3), col="Green",lwd=2)
```

(Courbe rouge : Année 2011 ; Courbe verte : Année 2013 ; Courbe bleue : Année 2015)

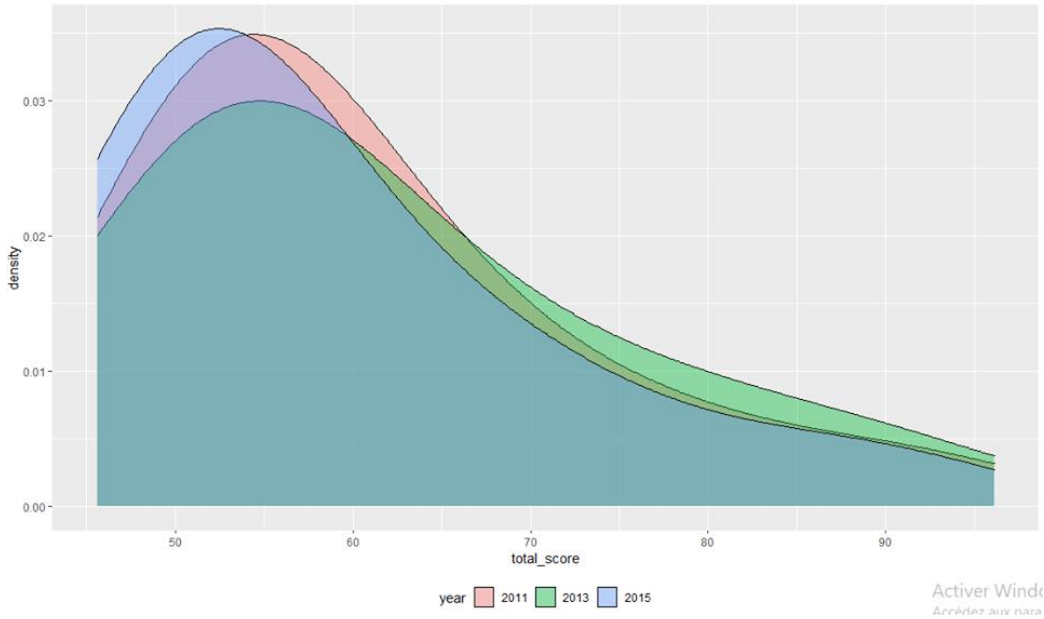


Figure 23:Densité empirique du score (Times;2011-2013-2015)

On détermine les paramètres de formes qui comme mentionné auparavant d'après la forme de la distribution.

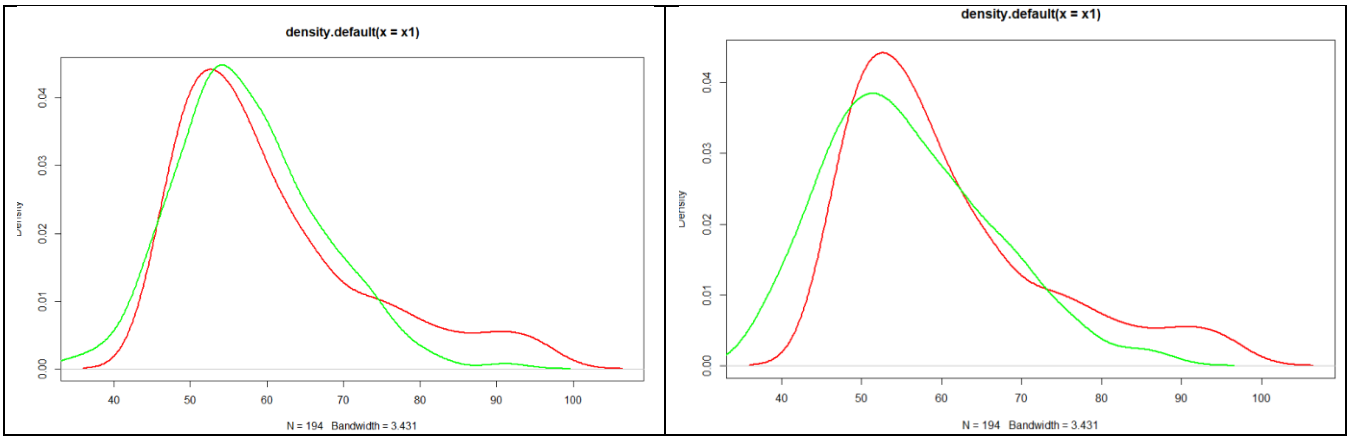
```
> skewness(x1)
[1] 1.184383
> kurtosis(x1)
[1] 3.677347
```

On constate que le coefficient d'asymétrie est égal à 1,184. Il est positif et proche de 0 et le coefficient d'aplatissement est égal à 3,67 qui est très proche de 3. Donc, on peut dire que la loi de distribution de la variable total_score est proche d'une loi normale.

Pour trouver une loi assez proche de celle du total_score, on compare les densités des lois connues à savoir la loi Log-normale et gamma avec la densité de la variable total_score.

Loi Gamma (vert)	Loi Log-normale (vert)

`> l=rgamma(194, shape=37.545, rate=0.63)`
`> plot(density(x1), col="Red",lwd=2)`
`> lines(density(l),col="Green",lwd=2)`
	`l=rlnorm(194, mean=3.9991, sdlog=0.212)`
	`plot(density(x1), col="Red",lwd=2)`
	`lines(density(l),col="Green",lwd=2)`



On effectue de nouveau les 3 tests d'ajustement :

	Log-normale	Gamma
KS	Two-sample Kolmogorov-Smirnov test data: l and x1 D = 0.1701, p-value = 0.007297 alternative hypothesis: two-sided	Two-sample Kolmogorov-Smirnov test data: l and x1 D = 0.082474, p-value = 0.5243 alternative hypothesis: two-sided
CVM	> cvm.test(x1, null="plnorm",mean=3.9991, sdlog=0.212) Cramer-von Mises test of goodness-of-fit Null hypothesis: log-normal distribution with parameter sdlog = 0.212 Parameters assumed to be fixed data: x1 omega2 = 2.4066, p-value = 1.398e-06	> cvm.test(x1, null="pgamma",shape=37.545 ,rate=0.63) Cramer-von Mises test of goodness-of-fit Null hypothesis: Gamma distribution with parameters shape = 37.545, rate = 0.63 Parameters assumed to be fixed data: x1 omega2 = 0.7561, p-value = 0.009245
CHI2	> l=rgamma(194, shape=37.545 ,rate=0.63) > tb1=table(x1,l) > chisq.test(tb1) Pearson's Chi-squared test data: tb1 X-squared = 28518, df = 28371, p-value = 0.268	> l=rlnorm(194, mean=3.9991, sdlog=0.212) > tb1=table(x1,l) > chisq.test(tb1) Pearson's Chi-squared test data: tb1 X-squared = 28518, df = 28371, p-value = 0.268

Le tableau ci-dessous résume les résultats des tests :

Test	Critère	Loi Log-normale	Loi gamma	Loi à retenir
K-S	D minimal	D = 0.1701	D = 0. 0827	Gamma
CVM	Omega2 minimal	Omega2 = 2.4066	Omega2 = 0.7561	Gamma
CHI-2	Statistique Chi2 minimale	Même résultat		incertain

Donc d'après ces tests, on peut dire que la loi gamma, avec les paramètres shape=37.545 et rate=0.63, est la plus proche de la distribution de la variable total_score.

4. Le classement Shanghai :

a. Aperçu global sur les variables :

Cette base de données contient de même des variables quantitatives et qualitatives.

Les variables qualitatives sont: “university_name” et “year”. Mais contrairement aux autres bases de données, celle-ci ne contient pas une colonne “country”.

On commence en visualisant la distribution des observations par année comme d’habitude en traçant le diagramme sectoriel suivant :

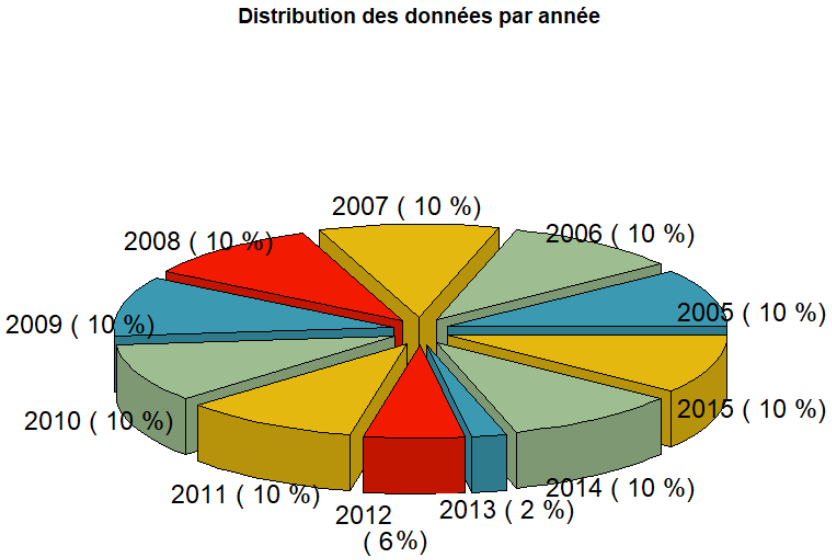


Figure 24:diagrammes en camembert de la distribution des données par année (Shanghai)

Le graphe ci-dessous représente les 5 pays avec le plus grands nombre d’observations. Identiquement aux autres classements, les universités américaines sont les plus présentes, suivi des universités britanniques.

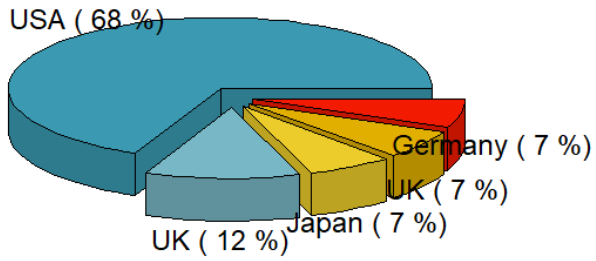


Figure 25:diagrammes en camembert des 5 pays les plus présents (Shanghai)

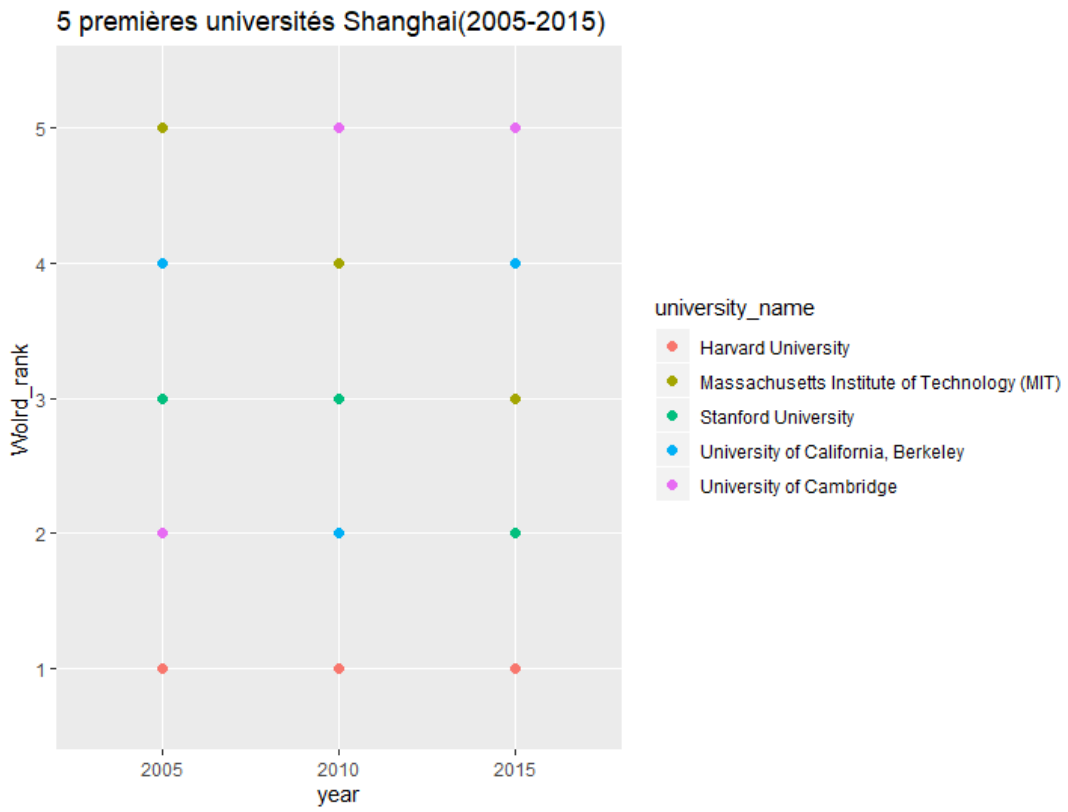


Figure 26:Les 5 premières universités (Shanghai 2005-2015)

Les 5 universités les mieux classées selon le classement Shanghai sont les mêmes pour les 3 années 2005-2010-2015. Comme prévu, les deux pays représentés dans cette liste sont : Le Royaume-Uni (2 universités) et les Etats-Unis (3 universités). La première place est toujours occupée par l'université de Harvard. Les autres quatre changent qu'une année à une autre mais ce sont les mêmes 4 universités qui décrochent ces places.

b. Le score :

La première étape est de tracer l'histogramme du score

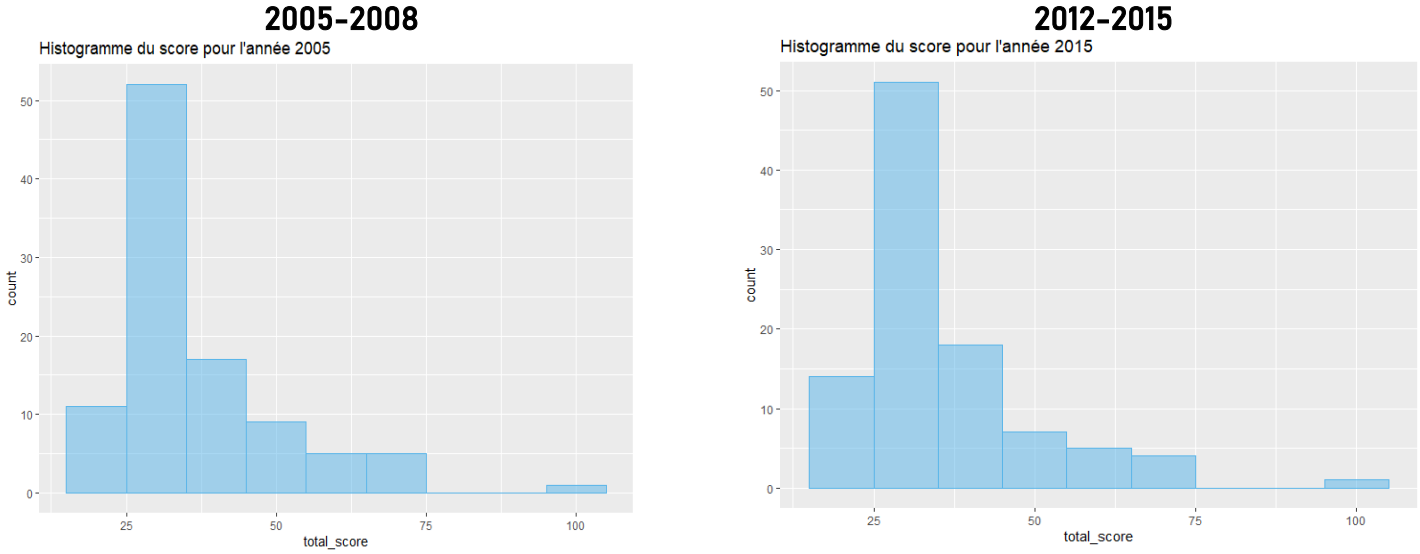


Figure 27: Histogramme du score (Shanghai;2005-2015)

Les histogrammes qu'on retrouve sont biaisés vers la droite, comme ceux retrouvés dans les autres classements. Ceci signifie que les universités les moins classés sont plus fréquents dans la base de données. La boîte à moustache ci-dessous confirme nos observations.

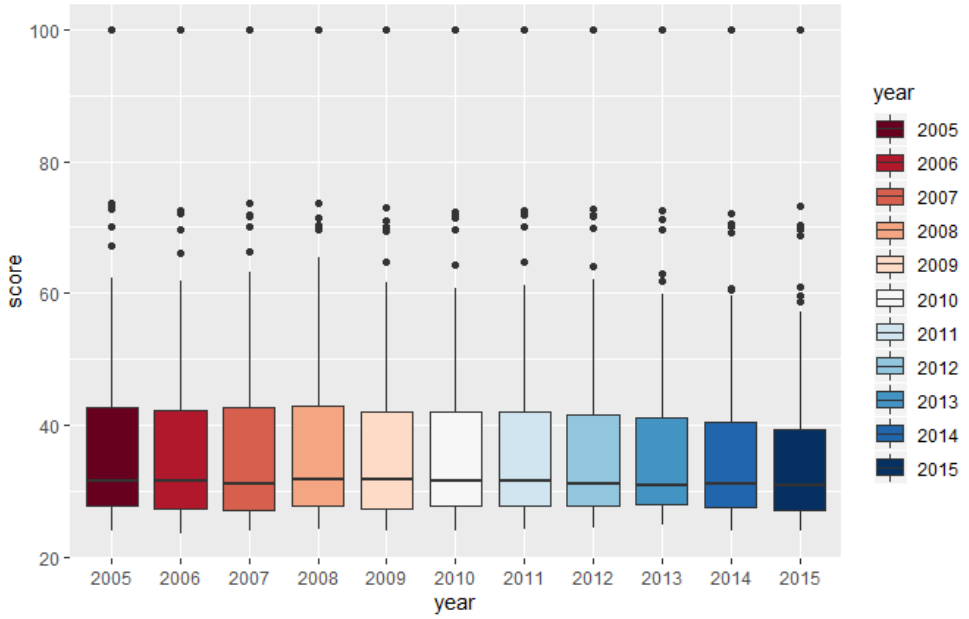


Figure 28 :Boîte à moustache du score selon l'année (Shanghai)

Ci-dessous la distribution géographique du score pour l'année 2005

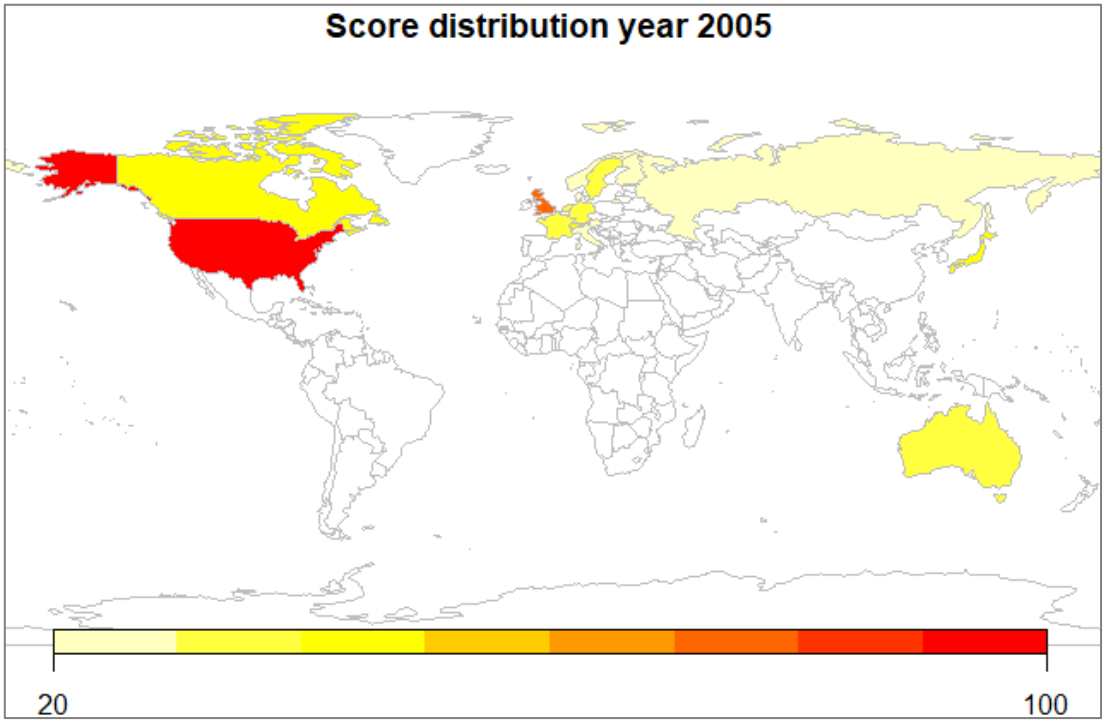


Figure 29: distribution géographique du score (Shanghai 2005)

Conformément aux autres classements, les scores les plus élevés sont localisés dans le continent américain du nord, suivi des pays européens et l'australie.

c. La variable score en fonction des autres variables :

Ci-dessous un exemple de la relation entre le score et les autres variables de la base de données :

- N&S Score, qui est un score basé sur le nombre d'articles publiés et qui ont une relation avec la nature et la science
- Award Score, basé sur le nombre de personnel de l'institution qui ont gagné des prix Nobel on physique, chimie, médecine et l'économie et autres médailles en mathématique.

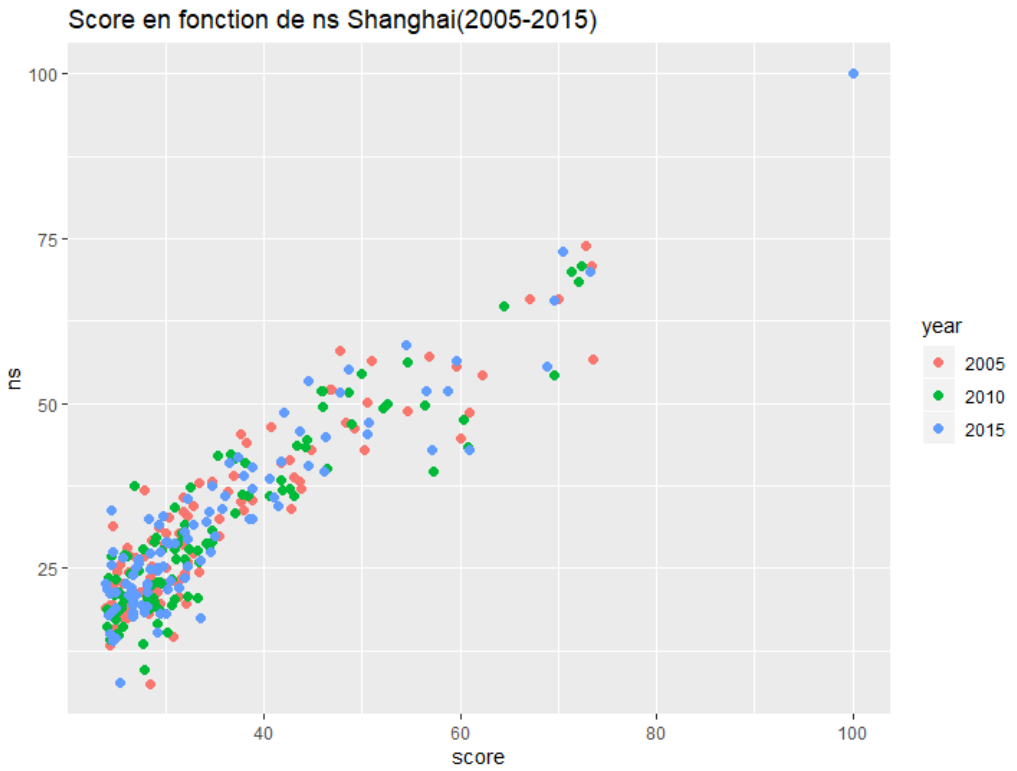


Figure 30; score en fonction de ns (Shanghai 2005-2015)

On constate qu'il y a une forte corrélation entre le score et ns. On remarque également que données reste presque inchangées entre les années. La corrélation est positive, les scores ns les plus élevés correspondent aux scores totaux les plus importants.

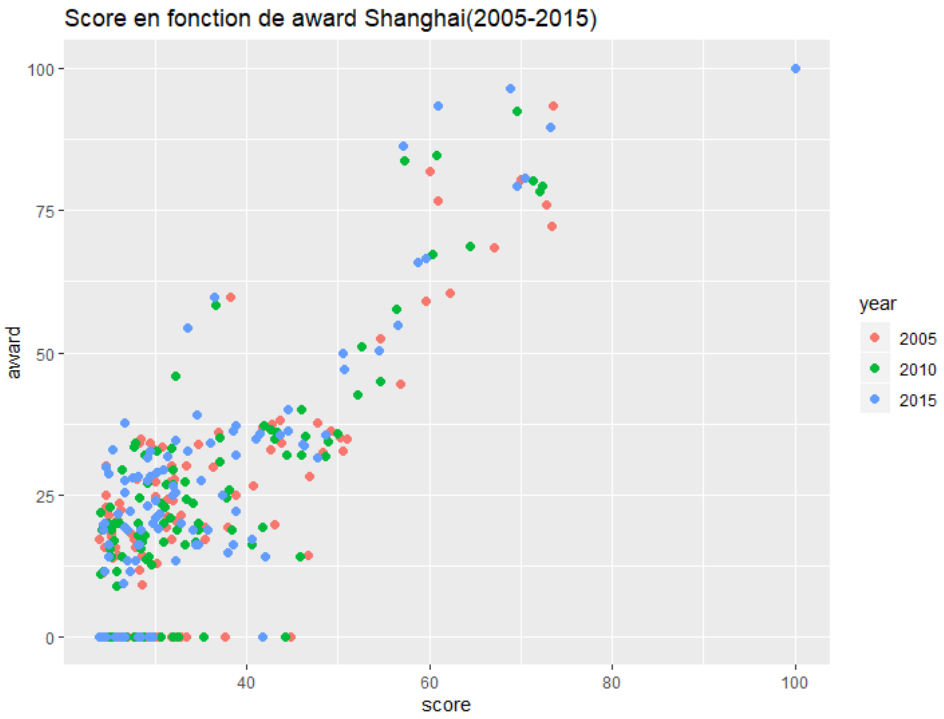


Figure 31: score en fonction de award (Shanghai 2005-2015)

La corrélation entre ces deux variables n'est pas aussi forte que le cas précédent. Néanmoins, on remarque qu'il y a une relation entre le score total et le score award. De même, le résultat est presque le même pour toute les années. On constate également que des observations ont un score award 0 et que ces observations ont un score faible qui ne dépasse pas 45.

En traçant les autres variables en fonction du score, on remarque qu'elles sont toutes corrélées avec le score, unes avec une corrélation plus signifiante que pour d'autres.

d. Matrices de corrélation :

Ces matrices de corrélation indiquent que la variable dépendante “total_score” est très corrélé avec les autres variables numériques.

e. Loi descriptive du score

Dans cette partie, nous avons essayé de comparer la distribution du score avec les lois usuelles en utilisant des tests.

On a commencé tout d'abord par dessiner les graphes de la densité du score pour toutes les années. On a constaté que toutes les densités ont presque la même forme, ci-dessous on a les graphes des années de 2005 à 2015 :

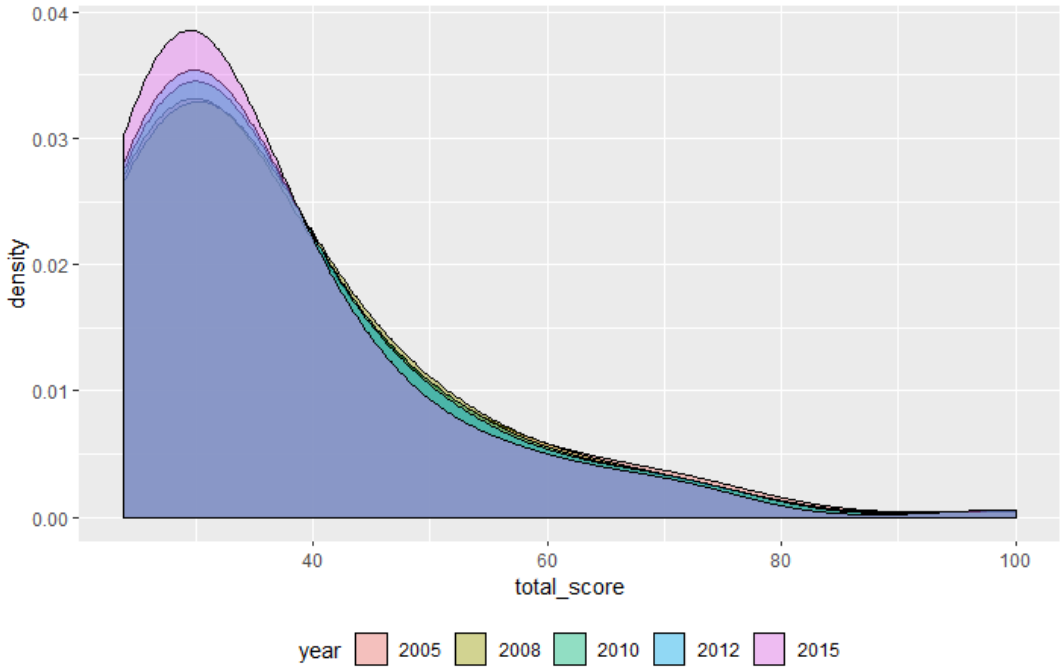


Figure 32:Densité empirique du score (Shanghai; 2005-2008)

Après, on a calculé asymétrie et de l'aplatissement du score de l'année 2005

```
> skewness(Shanghai100_2005$total_score)
[1] 1.847163
> kurtosis(Shanghai100_2005$total_score)
[1] 6.786321
>
```

Ceci nous indique que la densité du score a la forme de celle de la loi du Log-normale ou de gamma.

Après avoir essayé plusieurs valeurs pour les paramètres des deux lois, voici les deux courbes les plus proches de celle du score :

Gamma (Bleu) Log-normale (Rouge)

```
b<-rgamma(100,shape=17.515,rate=0.5842)
plot(density(b))
lines(density(Shanghai100_2005$total_score), col="blue")
      d <- rlnorm(100, mean=3.455618, sdlog = 0.2565072)
      plot(density(x), col="blue")
      lines(density(d), col="red")
      lines(density(b), col="black")
```

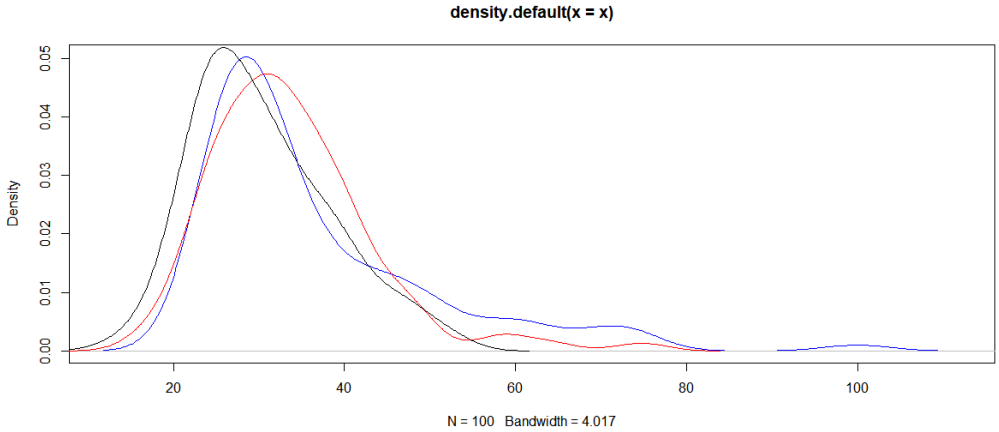


Figure 33:Densité empirique et théorique du score (Shanghai;2005)

Pour valider ceci, on a appliqué les tests pour comparer entre le score et ces deux lois.

	Log-normale	Gamma
KS	Two-sample Kolmogorov-Smirnov test data: x and d D = 0.14, p-value = 0.281 alternative hypothesis: two-sided	Two-sample Kolmogorov-Smirnov test data: x and b D = 0.26, p-value = 0.002318 alternative hypothesis: two-sided
CVM	Cramer-von Mises test of goodness-of-fit Null hypothesis: log-normal distribution with parameter sdlog = 0.2565072 Parameters assumed to be fixed data: x omega2 = 0.4684, p-value = 0.04775	Cramer-von Mises test of goodness-of-fit Null hypothesis: Gamma distribution with parameters shape = 17.515, rate = 0.5842 Parameters assumed to be fixed data: Shanghai100_2005$total_score omega2 = 1.8255, p-value = 2.779e-05
CHI 2	> tb = table(x,b) > chisq.test(tb) Pearson's Chi-squared test data: tb X-squared = 8000, df = 7920, p-value = 0.2615	> tb = table(x,d) > chisq.test(tb) Pearson's Chi-squared test data: tb X-squared = 8000, df = 7920, p-value = 0.2615

Le tableau ci-dessous résume les résultats des tests :

Test	Critère	Loi Log-normale	Loi gamma	Loi à retenir
K-S	D minimal (idéalement inférieur à la valeur critique 0.0136)	D = 0.14	D = 0.26	Log-normale
CVM	Omega2 minimal	Omega2 = 0.4684	Omega2 = 1.8255	Log-normale
CHI-2	Statistique Chi2 minimale	Même résultat	Même résultat	Incertain

Donc d'après ces tests, on peut dire que la loi Log-normale avec les paramètres moyenne =3.45, variance= 0.2, est la plus proche de la distribution de la variable total_score.

5. Synthèse :

	Support	Loi descriptive la plus probable
Score CWUR	$[0;+\infty[$	Log-normale (moyenne =3.8, variance= 0.02)
Score Times	$[0;+\infty[$	Gamma (shape=37.54, rate = 0.63)
Score Shanghai	$[0;+\infty[$	Log-normale (moyenne =3.45, variance= 0.2)

V. Modélisation des bases de données :

1. Méthodologie de travail :

L'objectif de cette partie est de trouver une représentation mathématique de l'association entre les variables exploratoires et la variable dépendante : le score. Notre démarche sera composée de 2 étapes principales : Construction du modèle sur des données d'apprentissage et utilisation des indicateurs statistiques pour choisir entre les modèles retrouvés.

Pour tous les systèmes de classement : On essayera dans un premier temps d'utiliser un modèle glm (generalized linear model) sur R puisqu'il nous permet de modéliser en utilisant les lois trouvées à partir de l'ajustement. Ensuite, nous allons tenter d'utiliser un modèle linéaire pour voir si on aura un meilleur résultat. Et finalement pour avoir encore plus de visibilité et comprendre davantage nos données, nous allons discrétiser la variable score en utilisant des méthodes de classification (clustering), et nous allons également tracer des arbres de décisions permettant de schématiser la relation entre les différentes variables.

2. Étude théorique :

Avant de commencer l'étude, on va donner un aperçu théorique sur les modèles qu'on utilisera.

a. Modèle Linéaire Généralisé :

i. Etude théorique

En ce qui a trait à la question de modélisation de la base de données, la première étape toute naturelle est de définir le meilleur modèle possible en se basant sur les résultats des statistiques descriptives. Dans notre cas, les résultats indiquent que les variables à décrire (le score dans les 3 cas), sont à support positif et les lois les plus proches qu'on a pu déterminer sont la loi gamma et la loi log-normale.

En pratique, ça constitue une généralisation du modèle linéaire (qu'on introduira par la suite), qui permet de le relier à la variable dépendante par une fonction lien. En effet, La moyenne, μ, de la distribution dépend des variables indépendantes, X, à travers la relation suivante :

$$\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta})$$

Avec :

- **E (Y)** est la valeur prévue de Y.
- **Xβ** est le prédicteur linéaire, une combinaison linéaire de paramètres inconnus, β.
- **g** est la fonction de liaison.

Dans ce cadre, la variance est typiquement une fonction, V, de la moyenne :

$$\operatorname{Var}(\mathbf{Y}) = \mathrm{V}(\boldsymbol{\mu}) = \mathrm{V}(g^{-1}(\mathbf{X}\boldsymbol{\beta})).$$

Les β sont généralement estimées en utilisant le maximum de vraisemblance.

ii. Construction du modèle

On utilisera deux modèles, ceux justifiés dans la partie précédente : Lognormale et Gamma. Pour ajuster en utilisant la fonction Gamma, on précise dans la fonction "glm" "family = Gamma", avec la fonction lien "link = inverse". Et pour utiliser la fonction lognormal, en utilisant l'argument "family = Gamma" avec "link = log".

iii. Diagnostic du modèle

La sortie du modèle présente 3 indicateurs importants qui nous donnerons une idée sur la qualité du modèle. D'abord :

Null Deviance et Residual Deviance : La déviance est une mesure de la qualité de l'ajustement du modèle linéaire généralisé, plus les chiffres sont élevés, pire est l'ajustement. La déviance nulle montre à quel point la variable de réponse est prédite par un modèle qui ne comprend que l'ordonnée à l'origine.

AIC(Akaike Information Criterion): Le critère d'information d'Akaike s'écrit comme suit

$$AIC = 2k - 2\ln(L)$$

Où k est le nombre de paramètres à estimer du modèle et L est le maximum de la fonction de vraisemblance du modèle.

Afin de choisir le bon modèle, on calcule les valeurs d'AIC associées. Il y aura toujours une perte d'information, du fait qu'on utilise un modèle pour représenter le processus générant les données réelles, et nous cherchons donc à sélectionner le modèle qui minimise cette perte d'information (ou plus exactement son estimation par l'AIC).

Le AIC propose une estimation de la perte d'information lorsqu'on utilise le modèle considéré pour représenter le processus qui génère les données. Dès lors, plus il est faible, mieux c'est.

Il est calculé sur R en utilisant la fonction :

```
> AIC(model)
```

b. Modèle Linéaire :

i. Etude théorique

On a opté pour la régression dans cette deuxième étape puisqu'elle est une technique très utilisée pour décrire la relation existante entre une variable dépendante et une ou plusieurs variables explicatives, et puisque la variable à expliquer, dans notre cas, est quantitative, on a utilisé la régression linéaire. Cette approche générative permet de fournir une règle de décision compréhensible par un opérateur humain. En effet, la régression linéaire simple est de la forme suivante :

$$y = \beta_0 + \beta_1 x_1 + +\beta_2 x_2 + \cdots .. + + \beta_n x_n + \epsilon$$

Avec :

- **y:** La variable dépendante qui le score total dans notre cas.
- x_i : les variables exploratoires qui doivent être des variables quantitatives.
- β_i: paramètres à estimer
- ε : le terme d'erreur aléatoire du modèle

Il importe de souligner qu'on a jugé admissible d'utiliser ce modèle car les matrices de corrélations renseignent une corrélation positive entre la variable à prédire et les variables indépendantes, donc on s'attend à ce que les coefficients soient positifs, et vu que la valeur des variables est positive, **on garantit des prédictions positives**.

ii. Construction du modèle :

Pour sélectionner le meilleur modèle, on va s'appuyer sur une mesure qui permet de comparer les modèles entre eux (par exemple, le coefficient de détermination ajusté, le critère d'information bayésien, ou encore le critère d'information d'Akaike). La régression pas à pas consiste à ajouter et à supprimer itérativement des prédicteurs, dans le modèle, afin de trouver le sous-ensemble de variables dans l'ensemble de données qui donne comme résultat le modèle le plus performant, c'est-à-dire un modèle qui a le AIC minimal et le R-squared le plus élevé.

Dans la régression pas-à-pas, trois manières de procéder sont les plus employées :

- **Forward :** on part du modèle avec uniquement une constante, et on ajoute les variables unes à unes jusqu'à ce que l'ajout d'une variable supplémentaire se solde par un modèle jugé moins bon en fonction du critère de comparaison sélectionné.
- **Backward :** on part du modèle avec toutes les variables auxquelles on a pensé, en retirant à chaque étape une seule variable jusqu'à ce que la comparaison des modèles indique qu'il est préférable de ne plus retirer de variable.

- **Stepwise :** un mélange des méthodes forward et backward, basée sur le F-statistic. On vérifie que l'ajout d'une variable ne provoque pas la suppression d'une variable déjà introduite.

L'implémentation sur R se fait en utilisant la fonction "train()" de la librairie "caret", elle a une option nommée méthode, qui peut prendre les valeurs suivantes:

- **"leapForward"**, Pour ajuster le modèle en utilisant la méthode Forward
- **"leapBackward",** Pour ajuster le modèle en utilisant la méthode Backward
- **"leapSeq"**, Pour ajuster le modèle en utilisant la méthode Stepwise

On doit également spécifier le paramètre de réglage nvmax, qui correspond au nombre maximal de prédicteurs à intégrer dans le modèle. Par exemple pour nvmax = 9, la fonction commence par chercher différents meilleurs modèles de taille différente, jusqu'au meilleur modèle à 9 variables. Autrement dit, elle cherche le meilleur modèle à 1 variable, le meilleur modèle à 2 variables, …, le meilleur modèle à 9 variables.

Nous utiliserons la méthode K-fold cross validation pour estimer l'erreur de prédiction moyenne (RMSE), et d'autres paramètres de chacun des 9 modèles.

```
library(leaps)
library(MASS)
library(caret)
#Cross validation
train.control <- trainControl(method = "cv", number = 10)
#Training the model
step.model <- train(dependant_variable~., data = data,
                    method = "leapSeq",
                    tuneGrid = data.frame(nvmax = 1:9),
                    trControl = train.control)
```

Il importe de souligner que ces trois méthodes déterminent le meilleur modèle selon le RMSE, R-squard et MAE, mais ne se basent que sur quelques paramètres ce qui n'est pas suffisant pour choisir le meilleur modèle. Il faut prendre en considération les autres critères déjà mentionné (AIC, p-value, F-statistic..)

iii. Diagnostic du modèle :

En regardant plusieurs indicateurs, on peut tirer des conclusions pour confirmer à quel point notre modèle est adéquat ou pas.

T-value : En gros, une T-value plus élevée indique qu'il est moins probable que le coefficient ne soit pas égal à zéro uniquement par hasard. Donc, plus la t-value est élevée, mieux c'est.

P-value : La p-value (probability value en anglais) est la probabilité que le coefficient de la variable associée soit nul, et donc peut être interprété comme la probabilité que le résultat s'est produit en raison d'une variation aléatoire. La fonction summary() nous indique les p-value des variables ainsi que du model, et c'est un facteur très important vu que le model ne peut être considéré statistiquement significatif seulement lorsque les deux sont inférieur à une valeur par défaut (de 0.05).

F-Statistic : La F-statistic est un autre indicateur pour déterminer l'existence d'une relation entre notre variable dépendante et les autres variables. Plus la statistique F est éloignée de 1, mieux c'est.

Residual standard error : Cet indicateur exprime la déviation moyenne de la réponse par rapport à la "vraie ligne de régression". C'est donc une sorte de mesure de la qualité de l'ajustement de notre régression linéaire. Puisque la valeur est très petite, le modèle est bon selon ce critère.

R-squared : On peut l'interpréter de la manière suivante : C'est la quantité de la variance trouvée dans la variable de réponse qui peut être expliquée par la variable prédictive. Donc on cherche à avoir le R-squared le plus élevé, sachant qu'il varie entre 0 et 1.

Le carré moyen des erreurs (MSE pour Mean Square Error ou MCE pour moyenne des carrés des erreurs) : C'est la moyenne arithmétique des carrés des écarts entre prévisions du modèle et observations. C'est la valeur à minimiser dans le cadre d'une régression linéaire. Il est implémenté sur R en utilisant la formule :

```
> mean((model$residuals)^2)
```

AIC(Akaike Information Criterion): (prière de revoir l'explication offerte dans la section du modèle linéaire généralisé.)

BIC(Bayesian Information Criterion): est un autre critère pour la sélection d'un modèle parmi un ensemble fini de modèles. Il est basé, en partie, sur la fonction de vraisemblance, et il est étroitement lié au critère d'information Akaike (AIC). Lors de l'ajustement de modèles, il est possible d'augmenter le R-Squared en ajoutant des paramètres, mais cela peut entraîner le surajustement (overfitting). Le critère BIC résout ce problème en introduisant un terme de pénalité pour le nombre de paramètres dans le modèle. La pénalité est plus élevée en BIC qu'en AIC. Il est calculé sur R en utilisant la fonction :

```
> BIC(model)
```

Cp de Mallow : ce critère aide à choisir entre plusieurs modèles de régression. Il permet de trouver un juste équilibre concernant le nombre de prédicteurs figurant dans le modèle. Le Cp de Mallows compare la précision et le biais du modèle complet à ceux de modèles contenant un sous-ensemble des prédicteurs.

```
> ols_mallows_cp(model, fullmodel)
```

iv. Analyse des résidus :

La différence entre la valeur observée de la variable dépendante et la valeur prédite par le modèle est appelée résidus. L'analyse des résidus joue un rôle important dans la validation du modèle de régression. Étant donné qu'un modèle de régression linéaire n'est pas toujours approprié pour les données, il importe d'évaluer la pertinence du modèle en examinant les graphiques des résidus. Si les termes

d'erreur dans le modèle de régression satisfont les hypothèses suivantes, le modèle sera considéré comme valide :

- Pas de Multi colinéarité : Cela signifie que les variables indépendantes ne doivent pas être trop fortement corrélées les unes avec les autres. On l'a déjà testé avec une matrice de corrélation.
- Pas d'autocorrélation : L'autocorrélation se produit lorsque les résidus ne sont pas indépendants les uns des autres.
- Homoscédasticité : Ce qui signifie que les résidus sont répartis régulièrement sur la ligne de régression, c'est-à-dire au-dessus et en dessous de la ligne de régression et la variance des résidus devrait être la même pour tous les scores prévus le long de la ligne de régression.

Les graphes que nous regarderons sont les suivants :

- **Residuals vs Fitted Plot :** Pour une régression linéaire correcte, les données doivent paraitre à peu près linéaires, ce qui permettra de tester si cette condition est remplie.
- **Normal Q-Q (quantile-quantile) Plot:** Les résidus devraient être normalement distribués et le tracé Q-Q nous permettra de le vérifier. Si les résidus suivent à peu près une ligne droite sur ce graphe, c'est une bonne indication qu'ils sont normalement distribués.
- **Scale-location Plot:** Ce graphique teste l'hypothèse d'homoscédasticité, c'est-à-dire que les résidus ont une variance égale le long de la droite de régression.
- **Residual vs Leverage Plot:** Ce graphique peut être utilisé pour identifier ce qu'on appelle des "points influents" dans l'ensemble de données. Un point influent est une observation qui, si supprimée, changera le modèle, de sorte que son inclusion ou son exclusion doit être prise en considération. Il peut être ou pas une valeur aberrante et le but est d'identifier les points qui ont une forte influence dans le modèle. Les valeurs aberrantes auront tendance à exercer un effet de levier et donc à influencer le modèle. Il apparaîtra en haut à droite ou en bas à gauche du graphique à l'intérieur d'une ligne rouge qui marque la distance du Cook.

c. Classification (Kmeans clustering) :

Pour discrétiser la variable, on opte pour un algorithme de classification, et particulièrement kmeans clustering, et ce pour avoir des classes équilibrées. Le principe est assez simple. L'étape d'initiation consiste à déterminer le nombre de classes de sortie "k", et l'idée est de créer parmi nos observations k nouvelles observations, localisées aléatoirement, appelées «centroïdes». Ensuite, le processus itératif suivant commence :

- Premièrement, pour chaque centroïde, l'algorithme trouve les points les plus proches (en termes de distance qui est généralement une distance euclidienne) de ce centroïde, et les affecte à sa catégorie.

- Deuxièmement, pour chaque catégorie (représentée par un centroïde), l'algorithme calcule la moyenne de tous les points attribués à cette classe. La sortie de ce calcul sera le nouveau centroïde de cette classe.

Le processus s'achève lorsque les centroïdes ne changent plus de position.

Pour déterminer le nombre "k" de classes de sortie, on utilise une méthode très connue et simplificatrice "Elbow Method"; On trace la variation expliquée en fonction des nombres de classes, et le nombre optimal est celui à partir duquel on obtient un rendement décroissant (en terme de variance expliquée) en augmentant k.

d. Arbres de classification :

Les arbres de classification sont des méthodes qui permettent d'obtenir des modèles à la fois explicatifs et prédictifs. Parmi leurs avantages on notera d'une part leur simplicité du fait de la visualisation sous forme d'arbres, d'autre part la possibilité d'obtenir des règles en langage naturel.

Pour implémenter une arbre sur R, il faut tout d'abord charger deux librairies qui nous permettront de créer l'arbre de décision et de le représenter : "rpart" et "rpart.plot".

3. Classement CWUR :

a. Construction des modèles linéaires généralisés (GLM):

D'après les tests d'ajustement, on a trouvé que la loi lognormale est la loi la plus adéquate pour décrire la distribution de la loi de la variable score pour le classement CWUR. Pour reconfirmer ces résultats, nous allons comparer les résultats du modèle glm pour la loi gamma et la loi lognormale et cela en changeant la fonction family par gamma(link="inverse") pour la loi gamma et gamma(link="log") pour la loi lognormale.

Loi lognormale:

```
Call:
glm(formula = score ~ national_rank + quality_of_education +
    alumni_employment + quality_of_faculty + citations + influence +
    patents, family = Gamma(link = log), data = num2012)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.089515  -0.026192  -0.003894   0.027439   0.099995

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.022e-02  1.633e-04  62.599  < 2e-16 ***
national_rank        6.457e-06  5.459e-06   1.183  0.23990
quality_of_education 2.692e-05  3.247e-06   8.289 9.01e-13 ***
alumni_employment    1.900e-05  3.103e-06   6.122 2.24e-08 ***
quality_of_faculty   4.358e-05  3.185e-06  13.683  < 2e-16 ***
citations            2.365e-05  5.101e-06   4.636 1.17e-05 ***
influence            1.463e-05  4.652e-06   3.146  0.00223 **
patents              1.262e-05  2.373e-06   5.315 7.40e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.001612532)

    Null deviance: 4.39611  on 99  degrees of freedom
Residual deviance: 0.14859  on 92  degrees of freedom
AIC: 447.49

Number of Fisher Scoring iterations: 3
```

Loi gamma:

```
Call:
glm(formula = score ~ national_rank + quality_of_education +
    alumni_employment + quality_of_faculty + citations + influence +
    patents, family = Gamma(link = inverse), data = num2012)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.104334  -0.032650  -0.005773   0.037074   0.140638

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.4985098  0.0148435 303.063  < 2e-16 ***
national_rank        -0.0006117  0.0003221  -1.899  0.06064 .
quality_of_education -0.0014970  0.0002087  -7.172 1.83e-10 ***
alumni_employment    -0.0012384  0.0002050  -6.040 3.23e-08 ***
quality_of_faculty   -0.0023721  0.0001977 -11.997  < 2e-16 ***
citations            -0.0014643  0.0003221  -4.546 1.66e-05 ***
influence            -0.0008268  0.0002909  -2.842  0.00552 **
patents              -0.0007716  0.0001628  -4.739 7.81e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.002445842)

    Null deviance: 4.3961  on 99  degrees of freedom
Residual deviance: 0.2240  on 92  degrees of freedom
AIC: 488.55

Number of Fisher Scoring iterations: 4
```

Commentaire :

Pour la loi lognormale, l'ajout des variables indépendantes a diminué la déviance de 4,39611 à 0,14859 avec la perte de 7 degrée de liberté. Alors que pour la loi gamma, la déviance a diminué de 4,3961 à 0,2240. Ainsi que le AIC du premier modèle (loi lognormale) est inférieur à celui du deuxième modèle.

Donc d'après ces résultats, on reconfirme ce qu'on a obtenu en utilisant les tests d'ajustement.

b. Modèles de régression linéaire (LM):

i. Construction des modèles

Nous allons nous focaliser sur l'année 2012.

D'abord, on élimine toutes les variables non-numérique en utilisant la fonction "select_if" de la librairie "dplyr". Ensuite, on se débarrasse de la variable "year" qui n'apporte aucune information supplémentaire vu qu'elle aura une seule valeur (2012), ainsi que la variable "broad_impact" qui n'a pas de valeur pour cette année, et finalement la variable "world_rank" puisqu'elle est calculée à base du score est donc n'est pas une variable indépendante :

```
#Linear_regression
library(dplyr)
#Selection des variables numériques
num2012 = select_if(cwur2012, is.numeric)
num2012$broad_impact = NULL
num2012$year = NULL
num2012$world_rank = NULL
```

Maintenant, puisqu'on dispose de 8 variables, on utilise nvmax = 8 pour les 3 type de la régression pas à pas. Le choix du meilleur modèle est basé sur la minimisation du RMSE.

Méthode "leapForward":

```
> step.model <- train(score~., data = num2012,
+                     method = "leapForward",
+                     tuneGrid = data.frame(nvmax = 1:8),
+                     trControl = train.control)
> step.model$results
  nvmax     RMSE  Rsquared      MAE   RMSESD RsquaredSD     MAESD
1     1 7.714099 0.6858096 6.348805 2.887309 0.14208903 2.3024182
2     2 5.855621 0.8033002 4.865023 1.710353 0.11640205 1.4679710
3     3 4.754570 0.8705536 3.898709 1.632994 0.12003446 1.3876374
4     4 4.772048 0.8816240 4.103391 1.560891 0.12636788 1.3266041
5     5 4.142940 0.9212521 3.490072 1.474729 0.08638350 1.1631842
6     6 3.875909 0.9402426 3.148026 1.178490 0.05053741 0.7444678
7     7 3.790914 0.9429861 3.089039 1.136672 0.05195469 0.7549321
8     8 3.828346 0.9411292 3.131789 1.124649 0.05201936 0.7240217
> step.model$bestTune
  nvmax
7     7
> summary(step.model$finalModel)
Subset selection object
8 Variables  (and intercept)
                    Forced in Forced out
national_rank           FALSE      FALSE
quality_of_education    FALSE      FALSE
alumni_employment       FALSE      FALSE
quality_of_faculty      FALSE      FALSE
publications            FALSE      FALSE
influence               FALSE      FALSE
citations               FALSE      FALSE
patents                 FALSE      FALSE
1 subsets of each size up to 7
Selection Algorithm: forward
         national_rank quality_of_education alumni_employment quality_of_faculty publications influence citations patents
1  ( 1 ) " "           " "                 " "               "*"                " "          " "       " "       " "
2  ( 1 ) " "           " "                 "*"               "*"                " "          " "       " "       " "
3  ( 1 ) " "           " "                 "*"               "*"                " "          "*"       " "       " "
4  ( 1 ) " "           " "                 "*"               "*"                " "          "*"       " "       "*"
5  ( 1 ) " "           "*"                 "*"               "*"                " "          "*"       " "       "*"
6  ( 1 ) " "           "*"                 "*"               "*"                " "          "*"       "*"       "*"
7  ( 1 ) "*"           "*"                 "*"               "*"                " "          "*"       "*"       "*"
```

Méthode "leapBackward" :

```
> step.model <- train(score~., data = num2012,
+                     method = "leapBackward",
+                     tuneGrid = data.frame(nvmax = 1:8),
+                     trControl = train.control)
> step.model$results
  nvmax     RMSE  Rsquared      MAE   RMSESD RsquaredSD     MAESD
1     1 7.941780 0.7039780 6.395731 3.612669 0.26171825 2.6324990
2     2 6.061520 0.8100811 4.964519 2.270377 0.19157224 1.8601680
3     3 5.159294 0.8652703 4.294814 1.766852 0.11947965 1.2752129
4     4 4.166768 0.9145151 3.447676 1.552254 0.07782339 1.0624037
5     5 4.056691 0.9223294 3.350388 1.390199 0.07586656 0.9106548
6     6 3.941434 0.9343854 3.317750 1.475309 0.05585007 1.0532066
7     7 3.692471 0.9468312 3.047403 1.511981 0.04152772 1.0996649
8     8 3.728595 0.9441972 3.093964 1.497998 0.04244823 1.0935637
> step.model$bestTune
  nvmax
7     7
```

```
> summary(step.model$finalModel)
Subset selection object
8 Variables  (and intercept)
                    Forced in Forced out
national_rank           FALSE      FALSE
quality_of_education    FALSE      FALSE
alumni_employment       FALSE      FALSE
quality_of_faculty      FALSE      FALSE
publications            FALSE      FALSE
influence               FALSE      FALSE
citations               FALSE      FALSE
patents                 FALSE      FALSE
1 subsets of each size up to 7
Selection Algorithm: backward
         national_rank quality_of_education alumni_employment quality_of_faculty publications influence citations patents
1  ( 1 ) " "           "*"                 " "                " "                " "          " "       " "       " "
2  ( 1 ) " "           "*"                 " "                " "                " "          " "       "*"       " "
3  ( 1 ) " "           "*"                 " "                "*"                " "          " "       "*"       " "
4  ( 1 ) " "           "*"                 "*"                "*"                " "          " "       "*"       " "
5  ( 1 ) " "           "*"                 "*"                "*"                " "          " "       "*"       "*"
6  ( 1 ) " "           "*"                 "*"                "*"                " "          "*"       "*"       "*"
7  ( 1 ) "*"           "*"                 "*"                "*"                " "          "*"       "*"       "*"
```

Méthode "leapSeq":

```
> step.model <- train(score~., data = num2012,
+                     method = "leapSeq",
+                     tuneGrid = data.frame(nvmax = 1:8),
+                     trControl = train.control)
> step.model$results
  nvmax     RMSE  Rsquared      MAE   RMSESD RsquaredSD     MAESD
1     1 7.956855 0.7082785 6.472004 2.564820 0.15643681 1.9137867
2     2 5.906495 0.8151317 5.023337 2.055126 0.14540033 1.8724625
3     3 5.115190 0.8814279 4.313988 1.954076 0.09667427 1.8780209
4     4 4.317962 0.9092620 3.581490 1.479671 0.08689260 1.0672061
5     5 4.183120 0.9293932 3.576706 1.455707 0.06805774 1.2286290
6     6 3.929799 0.9362005 3.228205 1.485699 0.06633683 1.2411011
7     7 4.052684 0.9285730 3.341841 1.240131 0.06888709 0.9478438
8     8 3.865602 0.9406448 3.206129 1.263981 0.05446358 0.9858173
> step.model$bestTune
  nvmax
8     8
> summary(step.model$finalModel)
Subset selection object
8 Variables  (and intercept)
                    Forced in Forced out
national_rank           FALSE      FALSE
quality_of_education    FALSE      FALSE
alumni_employment       FALSE      FALSE
quality_of_faculty      FALSE      FALSE
publications            FALSE      FALSE
influence               FALSE      FALSE
citations               FALSE      FALSE
patents                 FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: 'sequential replacement'
         national_rank quality_of_education alumni_employment quality_of_faculty publications influence citations patents
1  ( 1 ) " "           " "                 " "                "*"                " "          " "       " "       " "
2  ( 1 ) " "           " "                 "*"                "*"                " "          " "       " "       " "
3  ( 1 ) " "           " "                 "*"                "*"                " "          "*"       " "       " "
4  ( 1 ) " "           "*"                 "*"                "*"                " "          " "       "*"       " "
5  ( 1 ) " "           "*"                 "*"                "*"                " "          " "       "*"       "*"
6  ( 1 ) " "           "*"                 "*"                "*"                " "          "*"       "*"       "*"
7  ( 1 ) "*"           "*"                 "*"                "*"                " "          "*"       "*"       " "
8  ( 1 ) "*"           "*"                 "*"                "*"                "*"          "*"       "*"       "*"
```

Voici les résultats des variables utilisées par le meilleur modèle trouvé dans chaque cas :

	Forward Regression	Backward Regression	Stepwise Regression
national_rank	x	x	x
quality_of_education	x	x	x
alumni_employment	x	x	x
quality_of_faculty	x	x	x
publications			x
influence	x	x	x
citations	x	x	x
patents	x	x	x

ii. <u>Diagnostic t comparaison des modèles :</u>

D'après la régression stepwise, on retient deux modèles :

Modèle 1 (7 variables)

```
> lm1 = lm(score~national_rank + quality_of_education + alumni_employment + quality_of_faculty
+          + citations + influence + patents , data = num2012)
> summary(lm1)

Call:
lm(formula = score ~ national_rank + quality_of_education + alumni_employment +
    quality_of_faculty + citations + influence + patents, data = num2012)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3377 -2.4703 -0.1482  2.3736 15.2195

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         86.76259    1.15787  74.933  < 2e-16 ***
national_rank       -0.05030    0.02512  -2.002 0.048214 *
quality_of_education -0.09159    0.01628  -5.625 1.98e-07 ***
alumni_employment   -0.08978    0.01599  -5.614 2.08e-07 ***
quality_of_faculty  -0.13841    0.01542  -8.973 3.30e-14 ***
citations           -0.09269    0.02513  -3.689 0.000381 ***
influence           -0.04580    0.02269  -2.019 0.046441 *
patents             -0.04115    0.01270  -3.240 0.001666 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.858 on 92 degrees of freedom
Multiple R-squared:  0.9132,    Adjusted R-squared:  0.9065
F-statistic: 138.2 on 7 and 92 DF,  p-value: < 2.2e-16
```

Modèle 2 (8 variables)

```
> lm2 = lm(score~. , data = num2012)
> summary(lm2)

Call:
lm(formula = score ~ ., data = num2012)

Residuals:
    Min      1Q  Median      3Q     Max
-7.4139 -2.4006 -0.3145  2.4214 15.2478

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         86.73160    1.16810  74.250  < 2e-16 ***
national_rank       -0.05530    0.03014  -1.834  0.06985 .
quality_of_education -0.09018    0.01701  -5.300 8.02e-07 ***
alumni_employment   -0.08958    0.01609  -5.568 2.59e-07 ***
quality_of_faculty  -0.13768    0.01569  -8.777 9.22e-14 ***
publications         0.01068    0.03518   0.304  0.76213
influence           -0.04751    0.02349  -2.023  0.04601 *
citations           -0.10106    0.03740  -2.702  0.00822 **
patents             -0.04221    0.01324  -3.189  0.00196 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.877 on 91 degrees of freedom
Multiple R-squared:  0.9132,    Adjusted R-squared:  0.9056
F-statistic: 119.7 on 8 and 91 DF,  p-value: < 2.2e-16
```

Ci-dessous un diagnostic complet des deux modèles qui nous permet de faire une comparaison entre eux :

Indicateur statistique	Critère	Modèle 1	Modèle2	Modèle retenu d'après le critère
T-value	Le plus élevé possible	Entre -8.973 et 74.933	Entre -8.77 et 74.25	Modèle 1
P-value	Le plus faible possible	<2.2e-16	<2.2e-16	---
R-squared	Le plus élevé possible	0.9132	0.9132	----
Residual standard	Le plus faible possible	3.858	3.877	Modèle 1

error				
MSE	Le plus faible possible	13.69201	13.67816	Modèle 2
F-statistic	Le plus élevé possible	138.2	119.7	Modèle 1
AIC	Le plus faible possible	563.469	565.3677	Modèle 1
BIC	Le plus faible possible	586.9155	591.4194	Modèle 1
Cp	Le plus faible possible	7.092167	9	Modèle 1

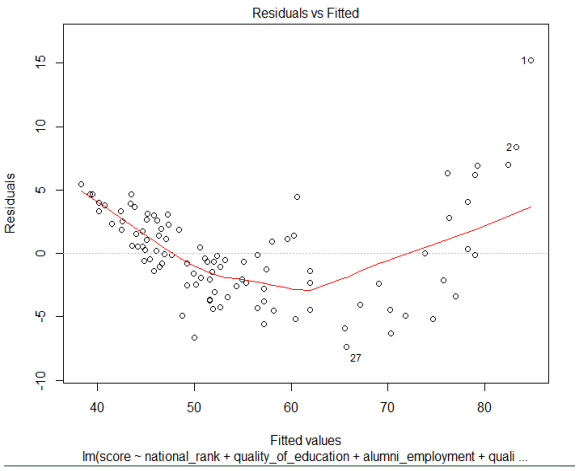
Les résultats sont très proches, ce qui n'est pas choquant vu que les modèles sont très similaires. Mais compte tenu de tous les critères, on opte pour le premier modèle de régression linéaire à 7 variables :

Score ~= -13% x quality_of_faculty - 9% x quality_of_education – 9% x citations - 8% x alumni_employment - 5% x national_rank – 4% x influence – 4% x patents + 87

iii. Residual analysis:

Residuals vs Fitted Plot:

Modèle 1



Modèle 2

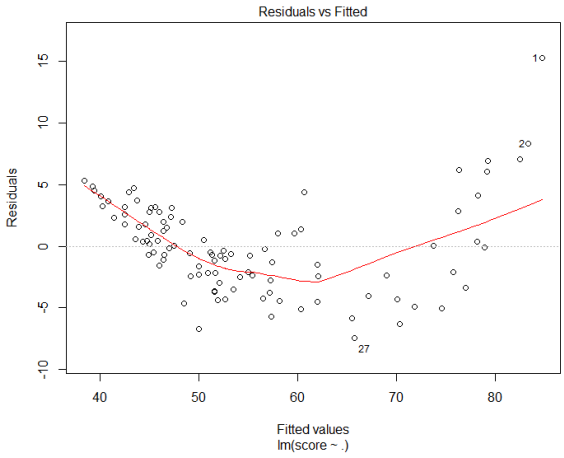


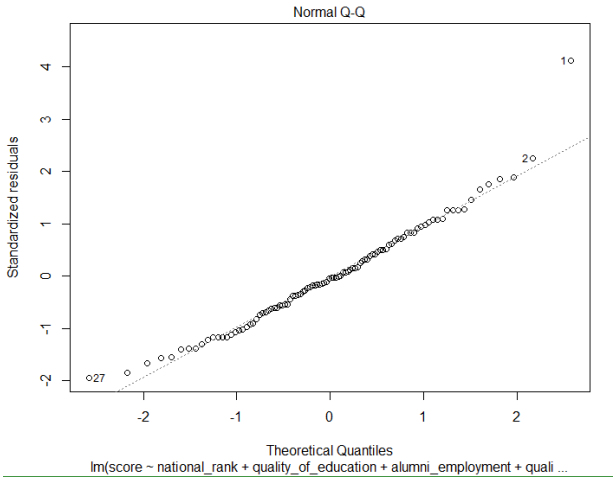
Figure 34:Residuals vs Fitted plot du modèle 1 et 2

Dans notre cas, en regardant le graphique ci-dessus, nous voyons que les données n'ont pas de tendance évidente et claire. Bien qu'il soit légèrement incurvé, les résidus sont répartis autour de la ligne horizontale sans tendance particulière. C'est une bonne indication qu'il ne s'agit pas d'une relation non-linéaire.

Or ces graphes ne permettent pas de conclure lequel des deux modèles est bon, car ils sont trop similaires.

Q-Q plot :

Modèle 1



Modèle 2

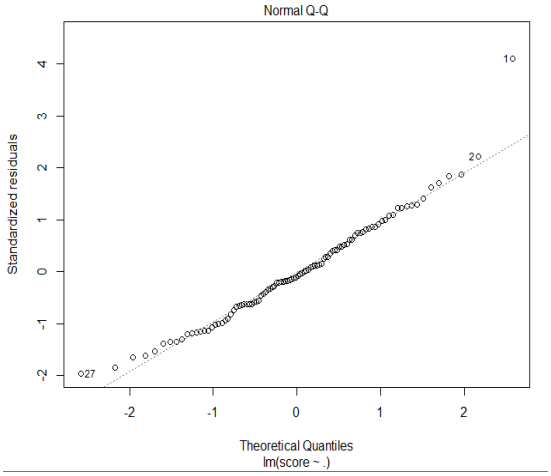
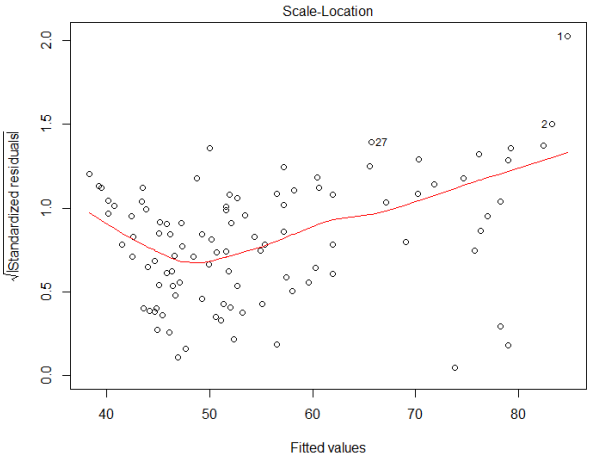


Figure 35: Q-Q plot du modèle 1 et 2

Le graphique montre que les résidus sont pratiquement normalement distribués. Or ces graphes ne permettent pas de conclure lequel des deux modèles est bon, car ils sont trop similaires.

Scale-Location :

Modèle 1



Modèle 2

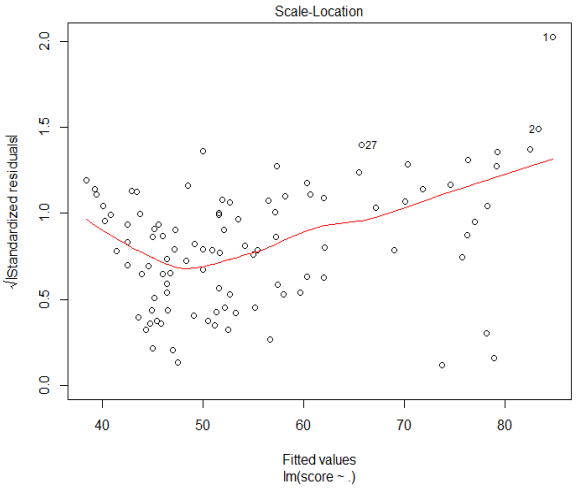


Figure 36: Scale-Location plot du modèle 1 et 2

Dans les deux cas, les résidus sont raisonnablement bien répartis au-dessus et au-dessous d'une ligne assez horizontale, mais la fin de la ligne a moins de points, donc un peu moins de variance.

Residual VS Leverage:

Modèle 1 **Modèle 2**

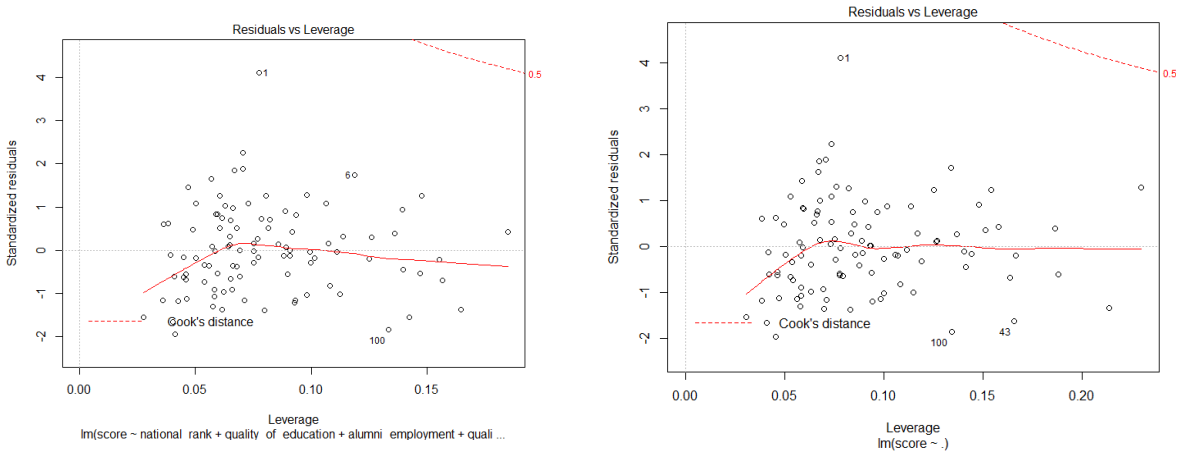


Figure 37: Residual VS Leverage plot du modèle 1 et 2

Dans notre exemple, on n'a aucun "point influent".

c. Création de l'arbre décisionnel :

Avant de créer l'arbre, il faut d'abord transformer la variable dépendante en une variable catégorielle. Pour cela on a créé une nouvelle variable qui permet de mettre les universités dans des catégories selon leur classement, par exemple les 20 premières universités les mieux classées seront placées dans la catégorie (0,20)

```
> num2012$RankCat<-cut(num2012$world_rank, seq(0,100,20))
> table(num2012$RankCat)

 (0,20] (20,40] (40,60] (60,80] (80,100]
     20      20      20      20       20
```

D'après ce qui précède, on constate que les universités sont équitablement réparties selon les différentes catégories. Cela signifie qu'il n'y a pas des différentes universités qui ont le même classement.

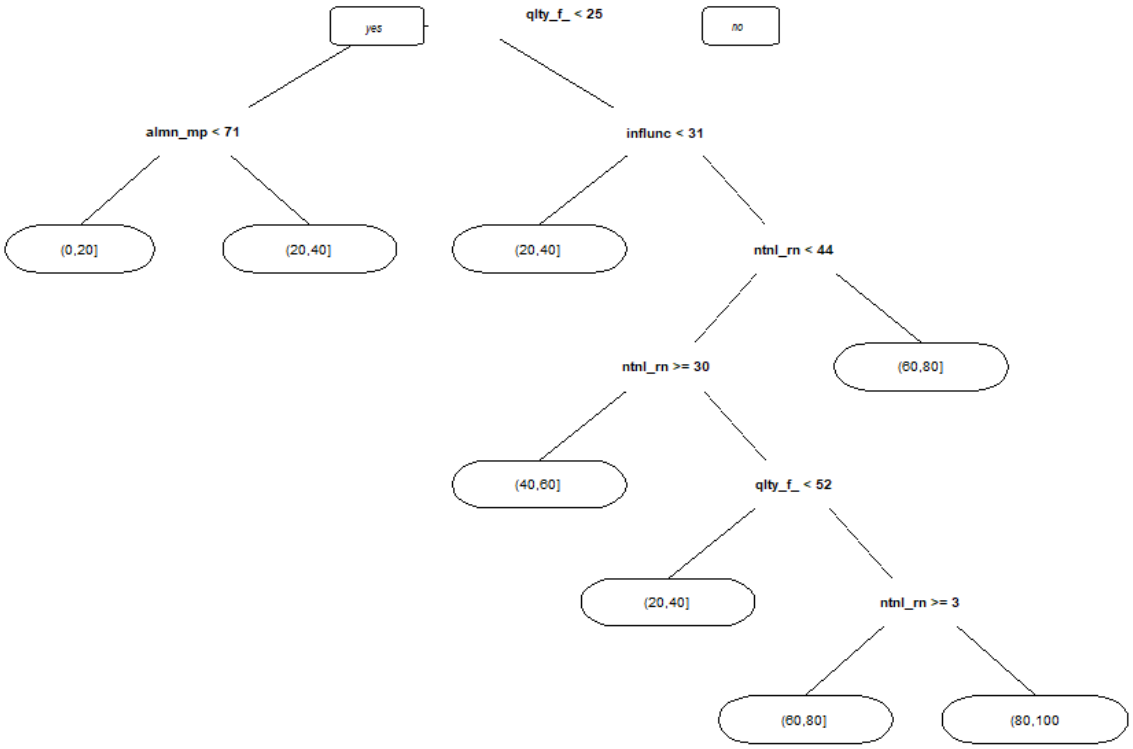


Figure 38: Arbre de classification du classement CWUR

On remarque l'apparition des variables : "quality_of_faculty", "influence","alumni_employment", "patents" et "national_rank" qui sont des variables assez significatives dans les modèles linéaires qu'on a construits.

On constate que cet arbre de classification permet d'avoir une représentation graphique, plus simple à interpréter que des résultats purement numériques.

d. Création des clusters :

Dans ce qui précède, on a créé un arbre décisionnel permettant d'expliquer les réponses de la variable dépendante catégorielle qu'on a créé en transformant la variable "classement des universités" en catégories. Cette fois-ci, on utilisera la méthode de classification kmeans qui permettra de diviser la base de données en classes compacts tout en étant les plus séparées possible.

Avant d'utiliser "elbow method", il est essentiel de normaliser la base de données vu que la distance euclidienne sera très impactée par les grandes différences. Ceci est obtenu en utilisant la fonction "preprocess" :

```
preproc = preProcess(cwur2012)
cwur2012Norm = predict(preproc, cwur2012)
```

Maintenant que nos données sont normalisées, on peut utiliser le code ci-dessous pour tracer la variation expliquée en fonction du nombre des groupes :

```
k.max <- 15
wss <- sapply(1:k.max,
              function(k){kmeans(cwur2012Norm, k, nstart=50,iter.max = 15 )$tot.withinss})
wss
plot(1:k.max, wss,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares")
```

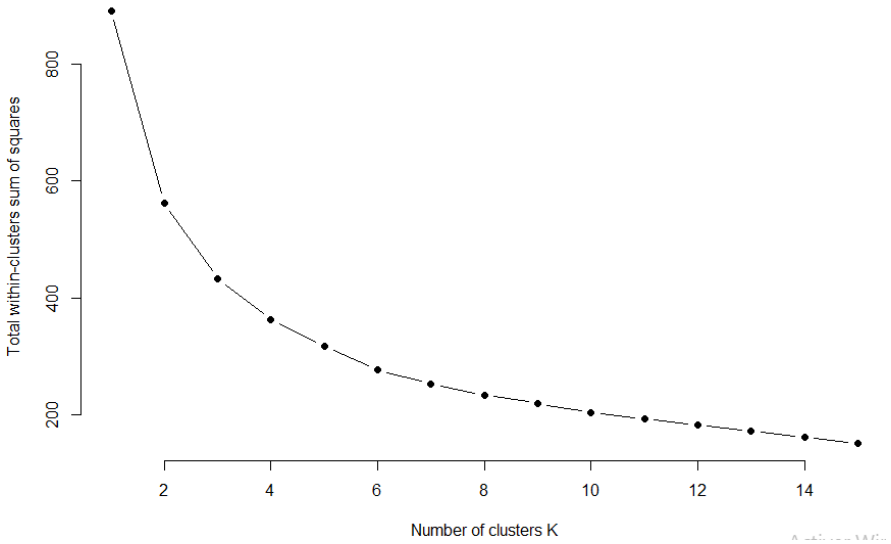


Figure 39: graphe de la Méthode "elbow" (CWUR)

En se basant sur le graphique ci-dessus, on construit 3 clusters :

```
KMC = kmeans(cwur2012Norm, centers = k, iter.max = 1000)
CwurCluster1 = num2012 %>% filter(KMC$cluster == 1)
CwurCluster2 = num2012 %>% filter(KMC$cluster == 2)
CwurCluster3 = num2012 %>% filter(KMC$cluster == 3)
```

Ci-après un tableau résumant les moyennes des variables pour chaque cluster :

	national_rank	quality_of_education	alumni_employment	quality_of_faculty	publications	influence	citations	patents	score	Max_Score	Min_Score	Obs Count
c1	7,14	27,24	37,43	14,86	17,62	16,38	16,05	26,95	75,67	100	58,37	21
c2	29,57	91,26	89,00	66,89	40,66	43,09	37,34	59,40	51,44	60,55	43,88	35
c3	14,30	65,36	82,68	69,09	84,30	82,66	86,32	84,55	47,83	65,09	43,36	44

On remarque que le premier cluster (en vert) est celui regroupant les universités ayant les meilleurs résultats pour toutes les variables. Ensuite, le deuxième cluster contient les universités ayant un bon classement pour les variables “quality of faculty”, “influence”, “citations”, “patents”. Ce qui est surprenant est que les universités appartenant à ce cluster sont en moyenne moins bien classées que les universités du troisième cluster pour les variables “quality of education”, “alumni employment” et “national_rank”. L’interprétation qu’on peut avoir est la suivante :

-Le premier cluster regroupe les universités ayant la meilleure performance globale.

-Le deuxième cluster est celui des universités plus orientées vers la recherche.

-Le troisième cluster est celui des universités ayant une meilleure formation.

4. Classement Shanghai :

a. Construction des modèles linéaires généralisés (GLM) :

D'après les tests d'ajustement, on a trouvé que la loi lognormale est la loi la plus adéquate pour décrire la distribution de la loi de la variable score pour le classement Shanghai. Pour reconfirmer cela, nous allons comparer les résultats du modèle glm pour la loi gamma et la loi lognormale :

Loi lognormale:

```
Call:
glm(formula = total_score ~ alumni + award + hici + ns + pub,
    family = Gamma(link = "log"), data = Shanghai100_2005)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.36302  -0.03675   0.00525   0.04006   0.11780

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.7230911  0.0277088  98.275  < 2e-16 ***
alumni      0.0016025  0.0004665   3.435 0.000883 ***
award       0.0049895  0.0004974  10.031  < 2e-16 ***
hici        0.0050647  0.0007581   6.680 1.66e-09 ***
ns          0.0058833  0.0009583   6.139 1.96e-08 ***
pub         0.0051446  0.0007298   7.050 2.97e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.003432)

    Null deviance: 11.31735  on 99  degrees of freedom
Residual deviance:  0.35224  on 94  degrees of freedom
AIC: 442.11

Number of Fisher Scoring iterations: 5
```

Loi Gamma:

```
Call:
glm(formula = total_score ~ alumni + award + hici + ns + pub,
    family = Gamma(link = "inverse"), data = Shanghai100_2005)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.56391  -0.10618  -0.02454   0.11587   0.25696

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.187e-02  1.615e-03  25.924  < 2e-16 ***
alumni      -4.628e-06  3.053e-05  -0.152  0.87984
award       -1.142e-04  3.450e-05  -3.310  0.00132 **
hici        -1.347e-04  5.049e-05  -2.667  0.00901 **
ns          -5.900e-05  6.194e-05  -0.953  0.34328
pub         -5.252e-05  4.901e-05  -1.072  0.28662
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.02033693)

    Null deviance: 11.3173  on 99  degrees of freedom
Residual deviance:  1.9624  on 94  degrees of freedom
AIC: 614.14

Number of Fisher Scoring iterations: 6
```

Commentaire :

Le AIC = 442.11 du modèle lognormale est inférieur de celui du modèle gamma (AIC: 614.14).

```
> BIC(GammaModel)
[1] 632
> BIC(LogModel)
[1] 460
```

On remarque la même chose en comparant le BIC des deux modèles.

En se basant sur les deux critères AIC et BIC, on peut conclure que la loi log-normale correspond le mieux à la distribution du score du classement Shanghai.

b. Modèles de regression linéaire (LM):

i. Construction des modèles

- **Méthode "leapBackward":**

On utilise maintenant les librairies "MASS" et "leap" pour determiner le meilleur modèle linéaire à partir des modèles linéaires de toutes les combinaisons possibles :

```
> step.model <- train(total_score ~ alumni +
+          award + hici+ ns + pub + pcp , data = Shanghai100_2005,
+                     method = "leapBackward",
+                     tuneGrid = data.frame(nvmax = 1:6),
+                     trControl = train.control)
> step.model$results
  nvmax       RMSE  Rsquared        MAE      RMSESD   RsquaredSD       MAESD
1     1 5.36795500 0.8577558 4.21479572 0.888615528 6.550817e-02 0.754913706
2     2 3.79383913 0.9291093 3.00070999 0.852096371 4.823140e-02 0.562031699
3     3 2.28407979 0.9760364 1.89872679 0.439654327 1.162250e-02 0.359085526
4     4 1.59370277 0.9870971 1.24035276 0.320751626 8.160756e-03 0.271853541
5     5 0.69087657 0.9976405 0.53056711 0.327277870 2.366867e-03 0.179456749
6     6 0.03152927 0.9999946 0.02626627 0.005713982 4.330980e-06 0.005400978
> summary(step.model$finalModel)
Subset selection object
6 Variables  (and intercept)
       Forced in Forced out
alumni     FALSE      FALSE
award      FALSE      FALSE
hici       FALSE      FALSE
ns         FALSE      FALSE
pub        FALSE      FALSE
pcp        FALSE      FALSE
1 subsets of each size up to 6
Selection Algorithm: backward
         alumni award hici ns  pub pcp
1  ( 1 ) " "    " "   " "  "*" " " " "
2  ( 1 ) " "    "*"   " "  "*" " " " "
3  ( 1 ) " "    "*"   " "  "*" "*" " "
4  ( 1 ) " "    "*"   "*"  "*" "*" " "
5  ( 1 ) "*"    "*"   "*"  "*" "*" " "
6  ( 1 ) "*"    "*"   "*"  "*" "*" "*"
> step.model$bestTune
  nvmax
6     6
```

On trouve comme résultat que le meilleur modèle est le modèle qui utilise toutes les variables. Ce choix est basé sur la RMSE la plus faible.

- **Méthode "leapSeq":**

```
> step.model <- train(total_score ~ alumni +
+          award + hici+ ns + pub + pcp , data = Shanghai100_2005,
+                      method = "leapSeq",
+                      tuneGrid = data.frame(nvmax = 1:6),
+                      trControl = train.control)
> step.model$results
  nvmax       RMSE  Rsquared       MAE      RMSESD   RsquaredSD       MAESD
1     1 5.17284012 0.8689989 4.0501670 1.296948692 6.899159e-02 0.808017295
2     2 5.96124192 0.7693338 4.8257166 2.551898675 2.255836e-01 2.226904666
3     3 2.36735992 0.9733496 1.9661705 0.715878592 1.558784e-02 0.598869613
4     4 1.80804591 0.9769512 1.3838383 0.420273926 3.652158e-02 0.299063547
5     5 0.75111169 0.9970946 0.5623696 0.371574811 2.095701e-03 0.177249617
6     6 0.03239834 0.9999937 0.0267861 0.005988113 4.501567e-06 0.005358093
> summary(step.model$finalModel)
Subset selection object
6 Variables  (and intercept)
       Forced in Forced out
alumni     FALSE      FALSE
award      FALSE      FALSE
hici       FALSE      FALSE
ns         FALSE      FALSE
pub        FALSE      FALSE
pcp        FALSE      FALSE
1 subsets of each size up to 6
Selection Algorithm: 'sequential replacement'
         alumni award hici ns  pub pcp
1  ( 1 ) " "    " "   " "  "*" " " " "
2  ( 1 ) " "    "*"   " "  "*" " " " "
3  ( 1 ) " "    "*"   " "  "*" "*" " "
4  ( 1 ) " "    "*"   "*"  "*" "*" " "
5  ( 1 ) "*"    "*"   "*"  "*" "*" " "
6  ( 1 ) "*"    "*"   "*"  "*" "*" "*"
> step.model$bestTune
  nvmax
6     6
```

La méthode Stepwise nous donne le même résultat.

- **Méthode "leapForward":**

```
> step.model <- train(total_score ~ alumni +
+          award + hici+ ns + pub + pcp , data = Shanghai100_2005,
+                      method = "leapForward",
+                      tuneGrid = data.frame(nvmax = 1:6),
+                      trControl = train.control)
> step.model$results
  nvmax       RMSE  Rsquared        MAE      RMSESD   RsquaredSD       MAESD
1     1 5.0105454 0.8756157 4.01531583 1.485502182 5.803378e-02 1.042476245
2     2 3.8401738 0.9088024 3.01056597 0.637813383 6.539104e-02 0.534706041
3     3 2.5437886 0.9573204 2.08101855 0.527996293 3.358168e-02 0.495661050
4     4 1.8259178 0.9803970 1.47735746 0.442090172 9.535192e-03 0.341466212
5     5 0.7047435 0.9970775 0.52726251 0.327887948 2.679442e-03 0.169543766
6     6 0.0322353 0.9999931 0.02691622 0.006192757 4.272070e-06 0.005278948
> summary(step.model$finalModel)
Subset selection object
6 Variables  (and intercept)
       Forced in Forced out
alumni     FALSE      FALSE
award      FALSE      FALSE
hici       FALSE      FALSE
ns         FALSE      FALSE
pub        FALSE      FALSE
pcp        FALSE      FALSE
1 subsets of each size up to 6
Selection Algorithm: forward
         alumni award hici ns  pub pcp
1  ( 1 ) " "    " "   " "  "*" " " " "
2  ( 1 ) " "    "*"   " "  "*" " " " "
3  ( 1 ) " "    "*"   " "  "*" "*" " "
4  ( 1 ) " "    "*"   "*"  "*" "*" " "
5  ( 1 ) "*"    "*"   "*"  "*" "*" " "
6  ( 1 ) "*"    "*"   "*"  "*" "*" "*"
> step.model$bestTune
  nvmax
6     6
```

La méthode Forward donne également le même résultat.

ii. Diagnostic du modèle linéaire :

Dans cette partie, on a réalisé un modèle de régression linéaire pour l'année 2005 en utilisant les variables numériques. Ci-dessous le résultat trouvé :

Modèle1

```
Call:
lm(formula = total_score ~ alumni + award + hici + ns + pub +
    pcp, data = Shanghai100_2005)

Residuals:
      Min        1Q    Median        3Q       Max
-0.062754 -0.022332 -0.002906  0.019353  0.078636

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0059863  0.0164071  -0.365    0.716
alumni       0.1027350  0.0002504 410.326   <2e-16 ***
award        0.2057896  0.0002744 749.895   <2e-16 ***
hici         0.2050699  0.0004029 509.003   <2e-16 ***
ns           0.2062086  0.0005517 373.741   <2e-16 ***
pub          0.2059984  0.0003921 525.328   <2e-16 ***
pcp          0.1027814  0.0004474 229.742   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03112 on 93 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 3.363e+06 on 6 and 93 DF,  p-value: < 2.2e-16
```

On fait la même chose pour le deuxième meilleur modèle qui est le même pour les trois méthodes et qui est le modèle avec 5 variables (toutes les variables sauf "pcp"):

Modèle2

```
Call:
lm(formula = total_score ~ alumni + award + hici + ns + pub,
    data = Shanghai100_2005)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7478 -0.5163 -0.0305  0.3966  4.2849

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.659138   0.349101   4.753 7.21e-06 ***
alumni      0.110948   0.005877  18.878  < 2e-16 ***
award       0.222813   0.006267  35.555  < 2e-16 ***
hici        0.207598   0.009552  21.734  < 2e-16 ***
ns          0.255097   0.012073  21.129  < 2e-16 ***
pub         0.192440   0.009194  20.930  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7381 on 94 degrees of freedom
Multiple R-squared:  0.9974,    Adjusted R-squared:  0.9972
F-statistic:  7155 on 5 and 94 DF,  p-value: < 2.2e-16
```

Le tableau ci-dessous résume les résultats des deux meilleurs modèles selon les trois méthodes :

Indicateur statistique	Critère	Modèle 1	Modèle2	Modèle retenu d'après le critère
T-value	Le plus élevé possible	Entre 229.741 et 749.895	Entre 18.878 et 21.734	Modèle 1
P-value	Le plus faible possible	<2.2e-16	<2.2e-16	--
R-squared	Le plus élevé possible	1	0.9974	Modèle 1
Residual standard error	Le plus faible possible	0.03112	0.7381	Modèle 1
MSE	Le plus faible possible	0.0009006982	0.5120844	Modèle 1
F-statistic	Le plus élevé possible	3.363e+06	7155	Modèle 1
AIC	Le plus faible possible	-401.4463	230.8611	Modèle 1
BIC	Le plus faible possible	-380.605	249.0973	Modèle 1
Cp	Le plus faible possible	7	52786.37	Modèle 1

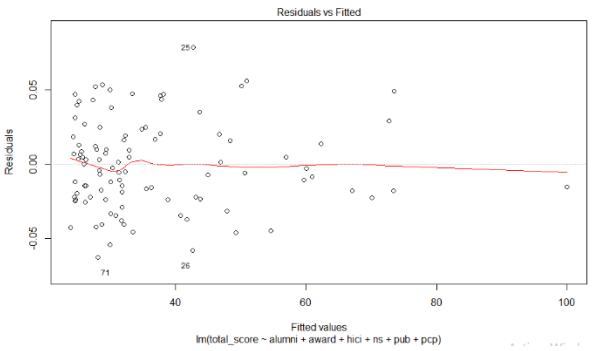
On conclut que le modèle 1 est celui à retenir, et la formule est la suivante :

Score ~= 10% x Alumni +20% x award + 20% x hici + 20% x ns + 20% x pub + 10% x pcp

iii. Residual analysis :

Residuals vs Fitted Plot :

Modèle 1



Modèle 2

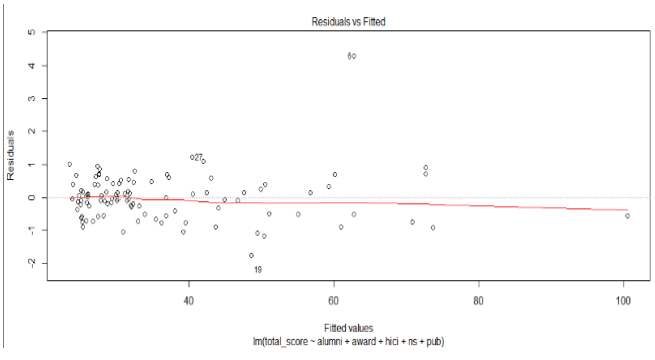
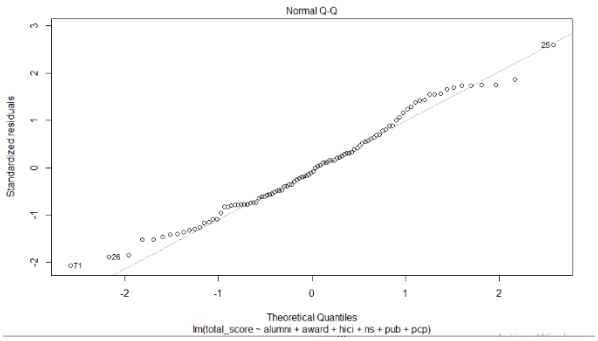


Figure 40: Residual VS Fitted plot du modèle 1 et 2

La ligne rouge des deux modèles sont assez horizontales ce qui signifie que la linéarité est vérifiée pour les deux.

Q-Q plot :

Modèle 1



Modèle 2

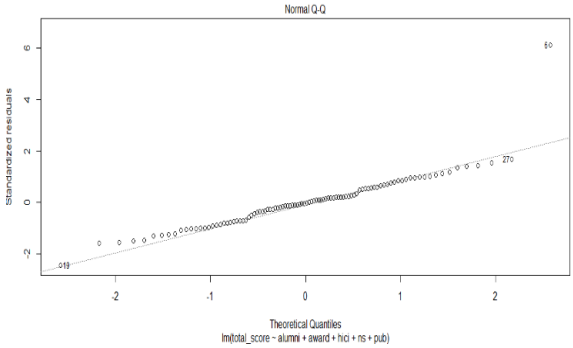
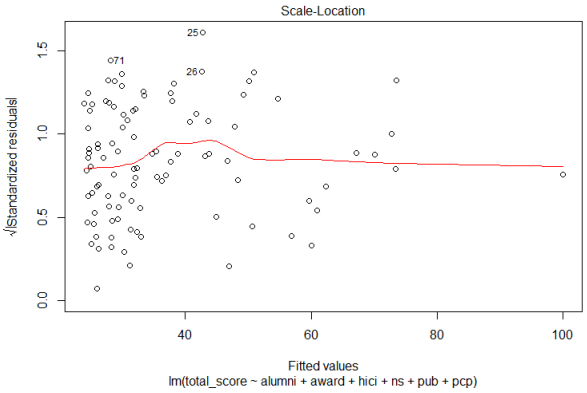


Figure 41: Q-Q plot du modèle 1 et 2

Les résiduels du modèle et les résiduels qui suivent la loi normale tombent approximativement sur la même ligne.

Scale-Location:

Modèle 1



Modèle 2

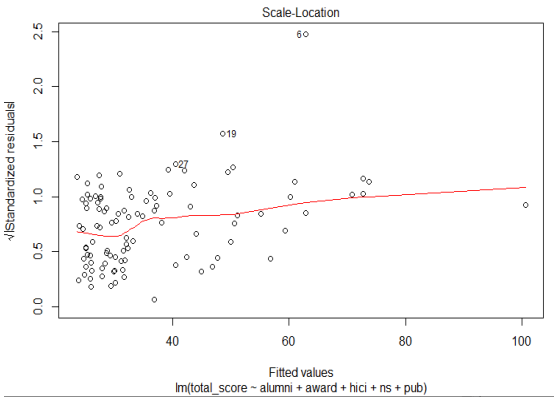
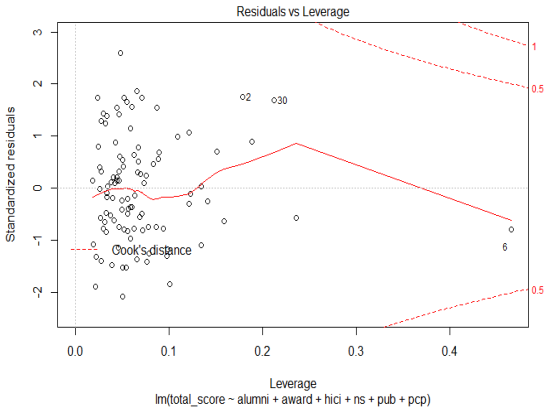


Figure 42: Scale-Location plot du modèle 1 et 2

Les deux courbes montrent que les résidus sont bien répartis au-dessus et au-dessous d'une ligne assez horizontale,

Residual VS Leverage:

Modèle 1



Modèle 2

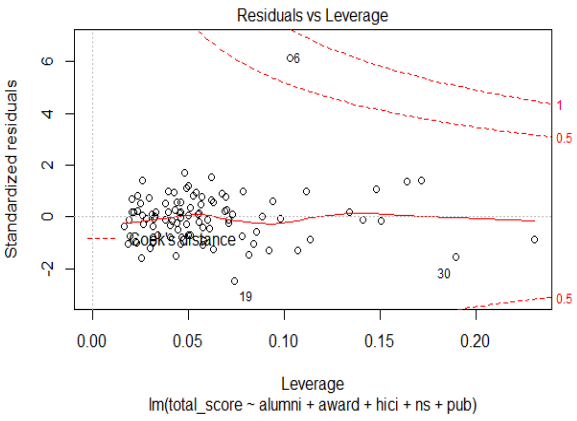


Figure 43: Residual VS Leverage plot du modèle 1 et 2

On n'a aucun "point influent" dans les deux courbes vu qu'on peut à peine voir les lignes "Cook's distance" (les lignes pointillées) et tous les cas sont entre ces lignes pointillées.

c. Création de l'arbre décisionnelle :

On a tout d'abord transformé la variable dépendante en une variable catégorielle en créant une nouvelle variable qui permet de mettre les universités dans des catégories selon leur classement :

On constate que le nombre d'université dans chaque catégorie. Cela signifie qu'il existe des différentes universités qui ont le même classement.

Ce code permet de créer et d'afficher l'arbre de décision suivante :

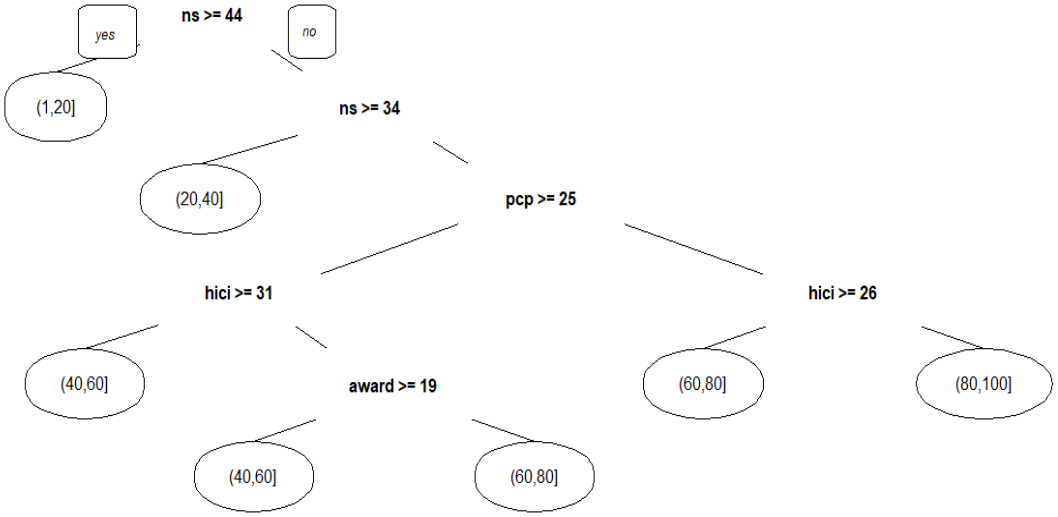


Figure 44: Arbre de classification du classement Shaighai

```
> table(Shanghai100_2005$world_rank2)

   (1,20]   (20,40]   (40,60]   (60,80]  (80,100]
       19        20        20        21        19
```

Le fait de transformer la variable ***world_rank*** en catégorie n'affectent pas les résultats puisque toutes les catégories ont approximativement le même nombre d'observations.

On constate que la variable qui apparait au début de l'arbre ***ns*** est la variable la plus significative dans le modèle linéaire (Coeff = 0.206). La seule variable qui n'est pas présente dans l'arbre est alumni. C'est la variable qui a le plus faible coefficient dans le modèle linéaire.

d. Création des clusters :

Après la sélection des variables numériques, on normalise la base de données, après, on utilise l'***"elbow method"*** pour déterminer le nombre optimal des clusters.

```
library("dplyr")
ShanghaiNumeric2005 = select_if(Shanghai100_2005, is.numeric)
library(caret)
preproc = preProcess(ShanghaiNumeric2005)
Shanghai100_2005Norm = predict(preproc, ShanghaiNumeric2005)
summary(Shanghai100_2005Norm)
set.seed(123)
# Compute and plot wss for k = 2 to k = 15.
k.max <- 15
wss <- sapply(1:k.max,
 function(k){kmeans(Shanghai100_2005Norm, k, nstart=50,iter.max = 15
                    )$tot.withinss})
plot(1:k.max, wss,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares")
```

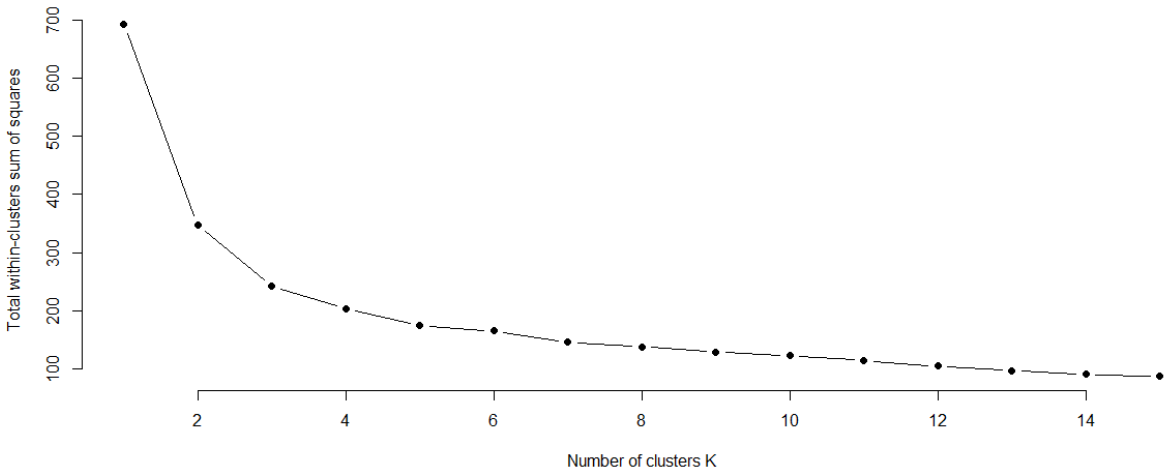


Figure 45: « Elbow method » classement Shanghai

La courbe ci-dessus nous indique que le nombre optimal de clusters est 4.

	Cluster1	Cluster2	Cluster3	Cluster4
total_score	26.96	46.98	70.00	33.93
alumni	21.17	33.75	72.86	21.88
award	14.40	31.91	76.92	20.63
hici	22.95	47.89	65.34	35.75
pub	47.61	66.86	67.50	51.82
pcp	25.48	37.30	60.99	29.58
ns	22.75	46.19	63.62	31.03

Le tableau montre les moyennes des variables de chaque cluster. On constate que les variables de chaque cluster sont plutôt homogènes. Chacun regroupe les universités qui ont les mêmes performances par rapport à tous les critères.

5. Classement times :

a. Construction des modèles linéaires généralisés (GLM):

D'après les tests d'ajustement, on a trouvé que la loi gamma st la loi la plus adéquate pour décrire la distribution de la loi de la variable score pour le classement times. On procède de la même manière pour reconfirmer ces résultats :

Loi gamma:

```
Call:
glm(formula = total_score ~ teaching + international + research +
    citations + income + num_students + international_students,
    family = Gamma(link = "inverse"), data = times11)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.054820  -0.012542   0.003133   0.013121   0.033235

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              3.150e+00  1.069e-02 294.717  < 2e-16 ***
teaching                 4.146e-03  2.669e-04  15.531  < 2e-16 ***
international            8.529e-04  1.172e-04   7.278 4.49e-11 ***
research                 4.698e-03  2.393e-04  19.636  < 2e-16 ***
citations                5.285e-03  1.094e-04  48.303  < 2e-16 ***
income                   4.399e-04  8.812e-05   4.992 2.15e-06 ***
num_students             3.566e-06  1.177e-04   0.030    0.976
international_students -2.536e-04  2.633e-04  -0.963    0.337
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.0003324162)

    Null deviance: 4.624030  on 122  degrees of freedom
Residual deviance: 0.038523  on 115  degrees of freedom
AIC: 374.59

Number of Fisher Scoring iterations: 3
```

Loi lognormale :

```
Call:
glm(formula = total_score ~ teaching + international + research +
    citations + income + num_students + international_students,
    family = Gamma(link = "log"), data = times11)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.097090  -0.027579   0.005547   0.027817   0.076293

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              3.111e-02  3.409e-04  91.237  < 2e-16 ***
teaching                -4.559e-05  8.841e-06  -5.157 1.06e-06 ***
international           -1.456e-05  3.790e-06  -3.840 0.000202 ***
research                -7.713e-05  8.024e-06  -9.613  < 2e-16 ***
citations               -8.608e-05  3.773e-06 -22.815  < 2e-16 ***
income                  -6.321e-06  2.813e-06  -2.247 0.026520 *
num_students            -2.008e-06  4.086e-06  -0.491 0.624170
international_students   1.008e-05  8.393e-06   1.201 0.232076
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.001300477)

    Null deviance: 4.62403  on 122  degrees of freedom
Residual deviance: 0.15062  on 115  degrees of freedom
AIC: 542.32

Number of Fisher Scoring iterations: 3
```

Commentaire :

Pour la loi lognormale, l'ajout des variables indépendantes a diminué la déviance de 4.624030 à 0.0385238 avec la perte de 7 degrée de liberté. Alors que pour la loi gamma, la déviance a diminué de 4.62403 à 0.15062. Ainsi que le AIC du premier modèle (loi gamma) est inférieur à celui du deuxième modèle, ce qui signifie que le modèle adéquat est le premier.

Donc d'après ces résultats, on reconfirme ce qu'on a obtenu en utilisant les tests d'ajustement.

b. Modèles de regression linéaire (LM):

i. Construction des modèles

- **Méthode "leapBackward":**

Afin de choisir les variables à utiliser dans notre modèle de régression linéaire, on utilise les librairies "leaps" et "MASS" qui permettent de calculer le RMSE, R-squard, Adjusted R-squard et MAE (mean absolute error):

```
> library(MASS)
> library(caret)
Le chargement a nécessité le package : lattice
Le chargement a nécessité le package : ggplot2
Warning message:
le package 'ggplot2' a été compilé avec la version R 3.6.3
> train.control <- trainControl(method = "cv", number = 10)
> step.model <- train(total_score~., data = times1,
+                     method = "leapBackward",
+                     tuneGrid = data.frame(nvmax = 1:9),
+                     trControl = train.control)
> step.model$results
  nvmax       RMSE  Rsquared        MAE      RMSESD   RsquaredSD       MAESD
1     1 6.08710880 0.7449095 4.92262054 0.992859956 9.127240e-02 0.934450316
2     2 2.23990883 0.9686829 1.78751578 0.381242150 1.359928e-02 0.348502270
3     3 1.23802033 0.9901434 1.05099264 0.190400461 5.017992e-03 0.207640904
4     4 0.48770585 0.9985523 0.40239522 0.102937432 6.654611e-04 0.081894892
5     5 0.03703176 0.9999917 0.03195337 0.005112345 3.633935e-06 0.005678803
6     6 0.03708783 0.9999917 0.03162222 0.005542176 3.344515e-06 0.006392321
7     7 0.03668270 0.9999919 0.03150616 0.006001543 3.079667e-06 0.006835200
8     8 0.03680029 0.9999918 0.03165945 0.005882690 3.285596e-06 0.006660400
9     9 0.03685600 0.9999918 0.03167233 0.005853333 3.270589e-06 0.006685360
```

D'après les résultats, on constate que le modèle qui utilise 7 variables donne un RMSE minimal et un Rsquared maximal.

Pour savoir ces 7 variables choisies, on utilise la syntaxe suivante :

```
> summary(step.model$finalModel)
Subset selection object
9 Variables  (and intercept)
                   Forced in Forced out
teaching               FALSE      FALSE
international          FALSE      FALSE
research               FALSE      FALSE
citations              FALSE      FALSE
income                 FALSE      FALSE
num_students           FALSE      FALSE
student_staff_ratio    FALSE      FALSE
international_students FALSE      FALSE
female_male_ratio      FALSE      FALSE
1 subsets of each size up to 7
Selection Algorithm: backward
         teaching international research citations income num_students student_staff_ratio international_students female_male_ratio
1  ( 1 ) " "      " "           "*"      " "       " "    " "          " "                " "                    " "
2  ( 1 ) " "      " "           "*"      "*"       " "    " "          " "                " "                    " "
3  ( 1 ) "*"      " "           "*"      "*"       " "    " "          " "                " "                    " "
4  ( 1 ) "*"      "*"           "*"      "*"       " "    " "          " "                " "                    " "
5  ( 1 ) "*"      "*"           "*"      "*"       "*"    " "          " "                " "                    " "
6  ( 1 ) "*"      "*"           "*"      "*"       "*"    "*"          " "                " "                    " "
7  ( 1 ) "*"      "*"           "*"      "*"       "*"    "*"          " "                "*"                    " "
```

- **Méthode "leapSeq":**

```
> step.model <- train(total_score~., data = times1,
+                     method = "leapSeq",
+                     tuneGrid = data.frame(nvmax = 1:9),
+                     trControl = train.control)
> step.model$results
  nvmax        RMSE  Rsquared        MAE      RMSESD   RsquaredSD       MAESD
1     1 5.69730955 0.7543836 4.70423931 0.956793995 1.194214e-01 0.870531958
2     2 2.21651538 0.9687322 1.80167793 0.532100119 1.746498e-02 0.486723842
3     3 5.20709319 0.7852442 4.12482122 0.637408726 1.363312e-01 0.444861287
4     4 0.49009824 0.9982932 0.40537443 0.103397342 9.454748e-04 0.090427544
5     5 0.03695073 0.9999900 0.03174273 0.005722595 7.541412e-06 0.005919727
6     6 0.03785862 0.9999892 0.03237666 0.005798282 8.940267e-06 0.005639128
7     7 0.03677896 0.9999902 0.03119632 0.005238847 7.109421e-06 0.005588911
8     8 0.03726134 0.9999895 0.03168511 0.005296856 8.620006e-06 0.005181673
9     9 0.03735115 0.9999895 0.03176460 0.005317095 8.711687e-06 0.005183155
```

D'après les résultats, on constate que le modèle qui utilise 7 variables donne un RMSE minimal et un Rsquared maximal.

```
> summary(step.model$finalModel)
Subset selection object
9 Variables  (and intercept)
                   Forced in Forced out
teaching               FALSE      FALSE
international          FALSE      FALSE
research               FALSE      FALSE
citations              FALSE      FALSE
income                 FALSE      FALSE
num_students           FALSE      FALSE
student_staff_ratio    FALSE      FALSE
international_students FALSE      FALSE
female_male_ratio      FALSE      FALSE
1 subsets of each size up to 7
Selection Algorithm: 'sequential replacement'
         teaching international research citations income num_students student_staff_ratio international_students female_male_ratio
1  ( 1 ) "*"      " "           " "      " "       " "    " "          " "                " "                    " "
2  ( 1 ) " "      " "           "*"      "*"       " "    " "          " "                " "                    " "
3  ( 1 ) "*"      "*"           "*"      " "       " "    " "          " "                " "                    " "
4  ( 1 ) "*"      "*"           "*"      "*"       " "    " "          " "                " "                    " "
5  ( 1 ) "*"      "*"           "*"      "*"       "*"    " "          " "                " "                    " "
6  ( 1 ) "*"      "*"           "*"      "*"       "*"    "*"          " "                " "                    " "
7  ( 1 ) "*"      "*"           "*"      "*"       "*"    "*"          " "                "*"                    " "
> step.model$bestTune
  nvmax
7     7
```


On constate que le meilleur modèle d'après "leapSeq" est le même modèle trouvé en utilisant "leapBackward".

- **Méthode "leapForward":**

```
> step.model <- train(total_score~., data = times1,
+                     method = "leapForward",
+                     tuneGrid = data.frame(nvmax = 1:9),
+                     trControl = train.control)
> step.model$results
  nvmax        RMSE  Rsquared        MAE      RMSESD   RsquaredSD       MAESD
1     1 5.70541339 0.7483698 4.70869950 1.100887180 1.471111e-01 0.906432760
2     2 2.64126394 0.9476331 2.18313725 0.377767775 3.095666e-02 0.309007534
3     3 1.23917952 0.9895257 1.05285685 0.184886689 6.503331e-03 0.192601356
4     4 0.48346157 0.9982524 0.40388959 0.130181497 1.319309e-03 0.110986887
5     5 0.03678262 0.9999901 0.03168944 0.005728878 6.623772e-06 0.006150626
6     6 0.03722536 0.9999900 0.03180951 0.005182824 6.678311e-06 0.005577063
7     7 0.03654451 0.9999902 0.03124991 0.006186963 7.025720e-06 0.006238363
8     8 0.03718587 0.9999898 0.03194572 0.006379861 7.375962e-06 0.006274352
9     9 0.03722276 0.9999898 0.03194463 0.006430549 7.397182e-06 0.006364378
```

```
> summary(step.model$finalModel)
Subset selection object
9 Variables  (and intercept)
                   Forced in Forced out
teaching               FALSE      FALSE
international          FALSE      FALSE
research               FALSE      FALSE
citations              FALSE      FALSE
income                 FALSE      FALSE
num_students           FALSE      FALSE
student_staff_ratio    FALSE      FALSE
international_students FALSE      FALSE
female_male_ratio      FALSE      FALSE
1 subsets of each size up to 7
Selection Algorithm: forward
         teaching international research citations income num_students student_staff_ratio international_students female_male_ratio
1  ( 1 ) "*"      " "           " "      " "       " "    " "          " "                " "                    " "
2  ( 1 ) "*"      " "           " "      "*"       " "    " "          " "                " "                    " "
3  ( 1 ) "*"      " "           "*"      "*"       " "    " "          " "                " "                    " "
4  ( 1 ) "*"      "*"           "*"      "*"       " "    " "          " "                " "                    " "
5  ( 1 ) "*"      "*"           "*"      "*"       "*"    " "          " "                " "                    " "
6  ( 1 ) "*"      "*"           "*"      "*"       "*"    "*"          " "                " "                    " "
7  ( 1 ) "*"      "*"           "*"      "*"       "*"    "*"          " "                "*"                    " "
> step.model$bestTune
  nvmax
7     7
```

De même, le meilleur modèle selon "leapForward" est celui qui contient 7 variables à savoir : teaching, international, research, citations, income, num_students et international_students.

ii. Diagnostic et comparaison des modèles :

Le modèle linéaire retenu par les méthodes utilisées précédemment ainsi que le modèle qui contient toutes les variables sont respectivement modèle 1 et 2:

Modèle1

```
> summary(model2)

Call:
lm(formula = total_score ~ teaching + international + research +
    citations + income + num_students + international_students,
    data = times1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.078947 -0.030601  0.000389  0.025483  0.082804

Coefficients:
                       Estimate Std. Error  t value Pr(>|t|)
(Intercept)          -0.0174489  0.0209709   -0.832   0.4071
teaching              0.2999453  0.0005238  572.658   <2e-16 ***
international         0.0496945  0.0002299  216.120   <2e-16 ***
research              0.3000319  0.0004695  639.085   <2e-16 ***
citations             0.3251249  0.0002147 1514.462   <2e-16 ***
income                0.0250628  0.0001729  144.943   <2e-16 ***
num_students          0.0003832  0.0002309    1.660   0.0997 .
international_students 0.0006924  0.0005166    1.340   0.1828
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03578 on 115 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 2.067e+06 on 7 and 115 DF,  p-value: < 2.2e-16
```

Modèle2

```
> model=lm(total_score~., data=times1)
> summary(model)

Call:
lm(formula = total_score ~ ., data = times1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.079632 -0.029235  0.000075  0.025836  0.082606

Coefficients:
                       Estimate Std. Error  t value Pr(>|t|)
(Intercept)          -0.0121839  0.0296985   -0.410    0.682
teaching              0.2998717  0.0005656  530.171   <2e-16 ***
international         0.0496958  0.0002357  210.825   <2e-16 ***
research              0.3000941  0.0004953  605.888   <2e-16 ***
citations             0.3251523  0.0002263 1437.021   <2e-16 ***
income                0.0250352  0.0001866  134.152   <2e-16 ***
num_students          0.0003800  0.0002484    1.530    0.129
student_staff_ratio   0.0001102  0.0004244    0.260    0.796
international_students 0.0006882  0.0005211    1.321    0.189
female_male_ratio    -0.0001413  0.0003915   -0.361    0.719
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03606 on 113 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 1.582e+06 on 9 and 113 DF,  p-value: < 2.2e-16
```

Le tableau suivant représente une comparaison entre les différents critères qui permettent de choisir le meilleur modèle :

Indicateur statistique	Critère	Modèle 1	Modèle2	Modèle retenu d'après le critère
T-value	Le plus élevé possible	Entre -0.832 et 1514.462	Entre -0.41 et 1437.021	Modèle1
P-value	Le plus faible possible	<2.2e-16	<2.2e-16	---

R-squared	Le plus élevé possible	1	1	----
Residual standard error	Le plus faible possible	0.03578	0.03606	Modèle1
MSE	Le plus faible possible	0.001194743	0.001196654	Modèle1
F-statistic	Le plus élevé possible	2.067e+06	1.582e+06	Modèle1
AIC	Le plus faible possible	-460.513	-456.7095	Modèle1
BIC	Le plus faible possible	-435.2033	-425.7755	Modèle1
Cp	Le plus faible possible	6.180749	10	Modèle1

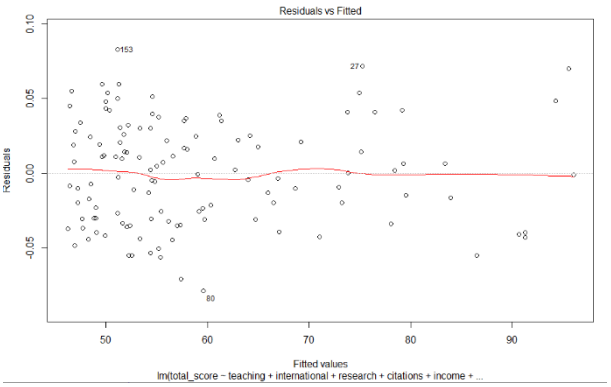
D'après le tableau ci-dessus, on constate que le modèle à retenir est celui qui contient les 7 variables : teaching, international, research, citations, income, num_students et international_students. et la formule est la suivante :

Score ~= 30% x teaching+30% x research + 32% x citation + 2.5% x income+ 5% x international

iii. Residual analysis:

Residuals vs Fitted Plot:

Modèle 1



Modèle 2

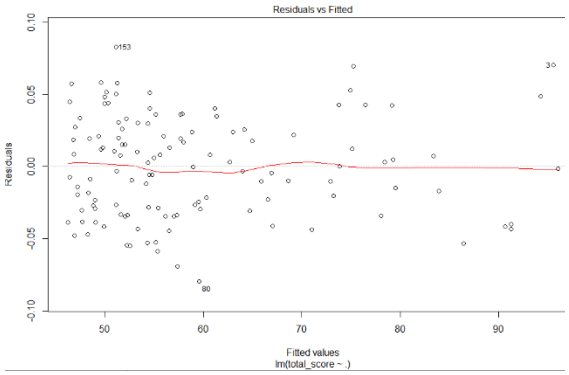


Figure 46: Residuals vs Fitted plot du modèle 1 et 2

La ligne rouge est assez horizontale ce qui signifie que la linéarité est vérifiée, et que les modèles sont bons.

Q-Q plot :

Modèle 1 **Modèle 2**

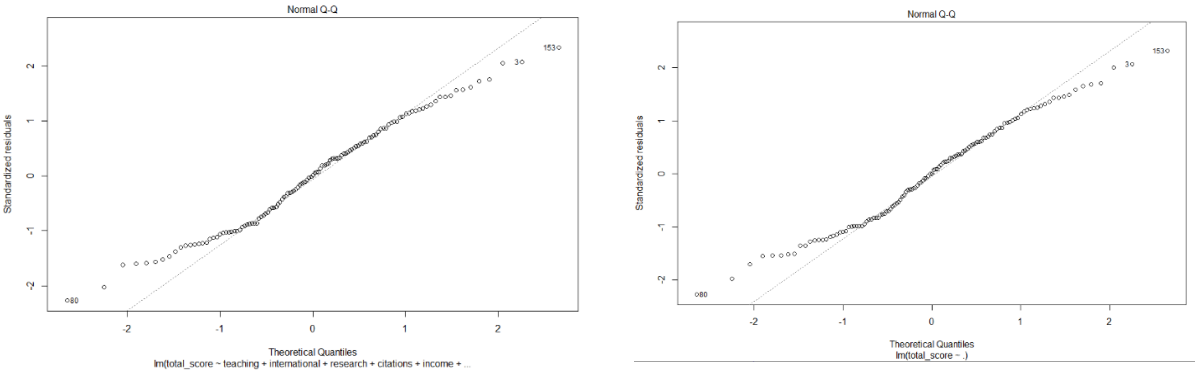


Figure 47: Q-Q plot du modèle 1 et 2

Les graphes ci-dessus montrent que les résidus sont pratiquement normalement distribués.

Scale-Location :

Modèle 1 **Modèle 2**

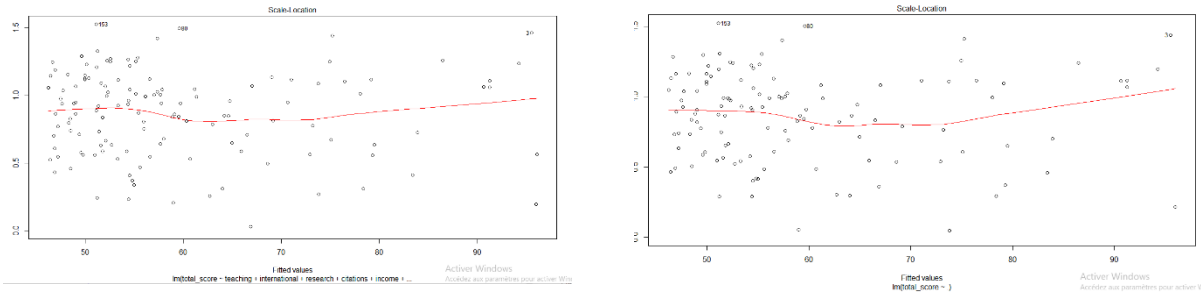


Figure 48: Scale-Location plot du modèle 1 et 2

D'après les deux courbes, on constate que les résidus sont bien répartis au-dessus et au-dessous d'une ligne assez horizontale,

Residual VS Leverage:

Modèle 1 **Modèle 2**

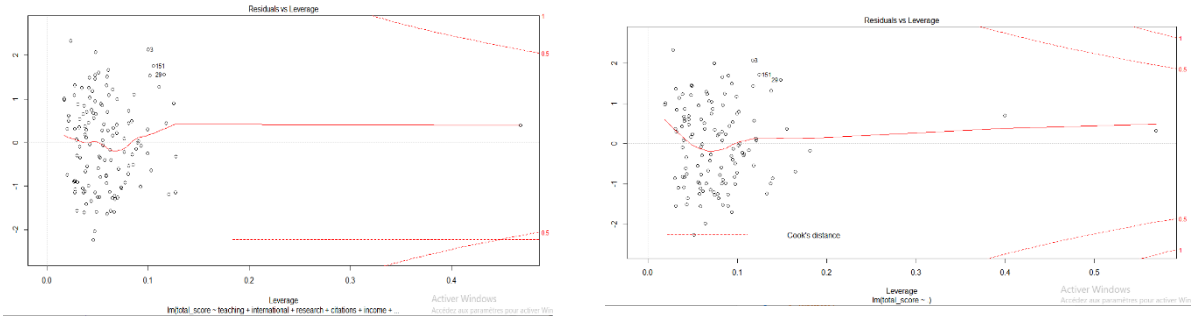


Figure 49: Residual VS leverage plot du modèle 1 et 2

On n'a aucun "point influent" dans les deux courbes vue qu'on peut à peine voir les lignes "Cook's distance" (les ligne pointillées)

c. Création de l'arbre décisionnelle :

Avant de créer l'arbre, on a transformé la variable dépendante en une variable catégorielle en créant une nouvelle variable qui permet de mettre les universités dans des catégories selon leur classement :

```
> times11$world_rank<-cut(times11$world_rank, c(1,20,40,60,80,100,120,140,160,180,200))
> table(times11$world_rank)

   (1,20]   (20,40]   (40,60]   (60,80]  (80,100] (100,120] (120,140] (140,160] (160,180] (180,200]
       11        13         9         9        10        13        14        15        14        15
```

On constate que le nombre d'université dans chaque catégorie varie entre 9 et 15. Cela signifie qu'il existe des différentes universités qui ont le même classement.

```
library(rpart)
library(rpart.plot)
model3=rpart(world_rank~ teaching+international+research+
                    citations+income+num_students+international_students, data = times11)
prp(model3)
```

Ce code permet de créer et d'afficher l'arbre de décision suivante :

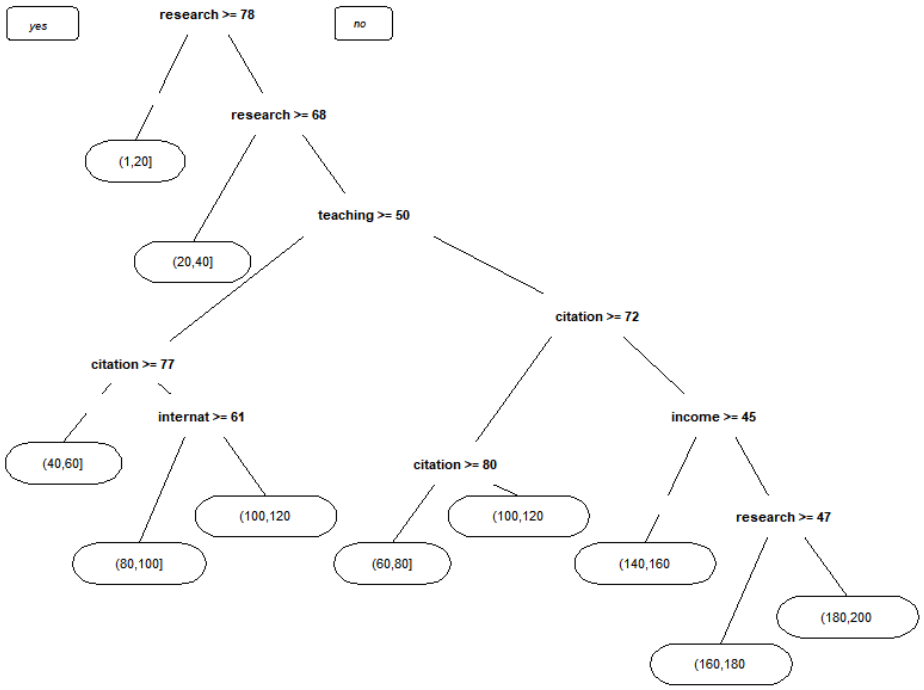


Figure 50: Arbre de classification du classement times

Donc, d'après cet arbre pour qu'une université soit classée parmis les 20 premières, il faut qu'elle ait un pourcentage de "research" supérieur ou égal à 78% sinon, elle peut avoir un classement entre 20 et 40 si le pourcentage de "research" ne diminue pas de 68%. Cela signifie que la variable "research" est assez significative. En effet, cela est compatible avec les résultats obtenus par les modèles de régression linéaire.

d. Création des clusters :

On procède maintenant à la création des classes. On commence tout d'abord par déterminer le nombre de "clusters" optimals et cela en utilisant "elbow method". Mais avant de faire cela, il faut normaliser la base de données :

```
preproc = preProcess(times11)
timesq= predict(preproc,times11)
```

On trace la variation expliquée en fonction des nombres de classes:

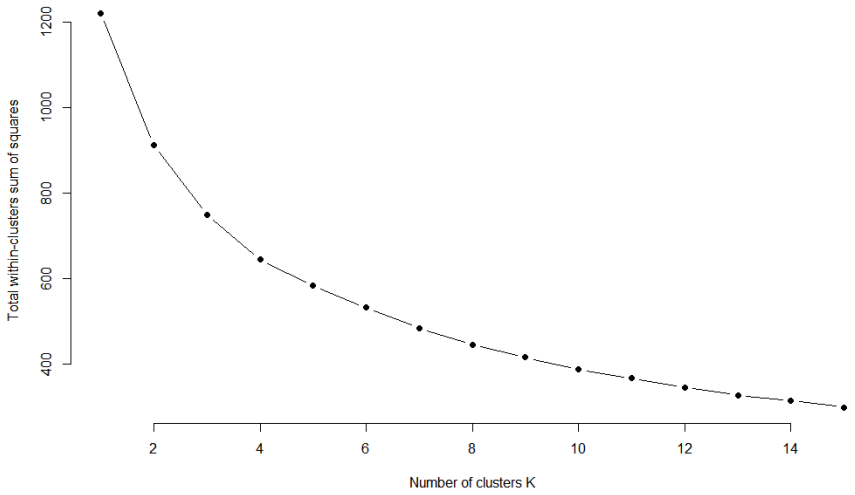


Figure 51: "Elbow method" classement times

En se basant sur le graphe ci-dessus, on choisit k=3 :

```
KmeansCluster = kmeans(timesq, centers=3,iter.max=1000)
library(dplyr)
KmeansCluster1 = times11%>% filter(KmeansCluster$cluster == 1)
KmeansCluster2 = times11 %>% filter(KmeansCluster$cluster == 2)
KmeansCluster3 = times11 %>% filter(KmeansCluster$cluster == 3)
```

Ce programme permet de diviser la base de données en trois catégories. Afin de savoir les caractéristiques de chaque classe, on regarde la répartition des universités selon leur score dans chaque “cluster” :

```
> table(KmeansCluster2$scorecat)

 (40,60]  (60,70]  (70,80] (80,100]
      23        3        1        0
> table(KmeansCluster1$scorecat)

 (40,60]  (60,70]  (70,80] (80,100]
      61       10        0        0
> table(KmeansCluster3$scorecat)

 (40,60]  (60,70]  (70,80] (80,100]
       0        3       13        9
```

On constate que la plupart des universités du 1er cluster ont un score qui varie entre 40% et 60%. Même remarque pour le 2ème cluster. Par contre, le 3ème cluster contient des universités ayant un score supérieur à 60%.

```
> summary(KmeansCluster3$total_score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  66.50   73.80   78.40   80.11   86.40   96.00
> summary(KmeansCluster2$total_score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  46.40   50.30   54.40   54.56   55.75   75.10
> summary(KmeansCluster1$total_score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  46.20   49.75   52.20   53.89   57.70   67.00
```

Afin de savoir la différence entre le cluster 1 et 2, on regarde les caractéristiques des autres variables et on constate que pour le premier cluster la variable "income" a une moyenne de 38.5/100 qui est assez inférieure à celle du deuxième cluster (66.95/100). Par contre, la variable "international" dans le premier cluster a une moyenne de 67.06/100 qui est supérieure à celle du deuxième cluster (35.22/100)

6. Synthèse et validation des modèles :

En entamant cette partie, notre but était de trouver les modèles mathématiques qui représentent au mieux nos données. On a commencé, tout naturellement, en se basant sur les résultats de la partie précédente, par un modèle linéaire généralisé pour ajuster nos données avec une loi Gamma ou Log-normale. Mais vu qu'on n'a pas eu des résultats satisfaisants, on a essayé d'utiliser un modèle linéaire simple, étape qu'on a jugé de valide car les coefficients et les variables sont positifs et donc on s'attend à ce que les prédictions du score le seront aussi. Ci-dessous un récapitulatif des résultats obtenus :

ScoreShanghai ~= 10% x Alumni +20% x award + 20% x hici + 20% x ns + 20% x pub + 10% x pcp,(R^2 =1)

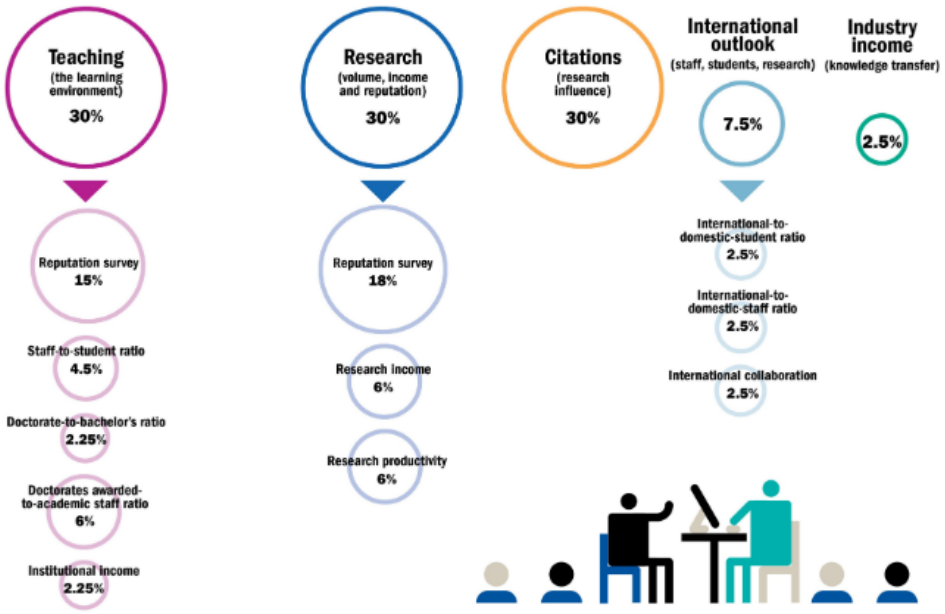
ScoreCWUR ~= -13% x quality_of_faculty - 9% x quality_of_education – 9% x citations -8% x alumni_employment - 5% x national_rank – 4% x influence – 4% x patents + 87; (R^2=0.9)

ScoreTimes ~= 30% x teaching+30% x research + 32% x citation + 2.5% x income+ 5% x international; (R^2=1)

Pour le modèle ayant un R^2=1, on estime que c'est un modèle exact.

Pour valider ces modèles, on a effectué des recherches sur les sites Web officiels des classements :

Classement TIMES :



Source : https://www.timeshighereducation.com/student/advice/world-university-rankings-explained

En comparant les résultats du modèle linéaire et la méthodologie utilisée dans le classement, on constate que les poids obtenus et les pourcentages utilisés en réalité sont les mêmes.

Classement Shanghai :

Criteria	Indicator	Code	Weight
Quality of Education	Alumni of an institution winning Nobel Prizes and Fields Medals	Alumni	10%
Quality of Faculty	Staff of an institution winning Nobel Prizes and Fields Medals	Award	20%
	Highly cited researchers in 21 broad subject categories	HiCi	20%
Research Output	Papers published in Nature and Science*	N&S	20%
	Papers indexed in Science Citation Index-expanded and Social Science Citation Index	PUB	20%
Per Capita Performance	Per capita academic performance of an institution	PCP	10%
Total			100%

Source : http://www.shanghairanking.com/ARWU-Methodology-2017.html

Le modèle linéaire donne les mêmes coefficients que ceux déclarés dans le site web du classement.

Classement CWUR :

1) **Quality of Education**, measured by the number of a university's alumni who have won major international awards, prizes, and medals relative to the university's size (25%)
2) **Alumni Employment**, measured by the number of a university's alumni who have held CEO positions at the world's top companies relative to the university's size (25%)
3) **Quality of Faculty**, measured by the number of academics who have won major international awards, prizes, and medals (10%)
4) **Research Performance:**
i) Research Output, measured by the the total number of research papers (10%)
ii) High-Quality Publications, measured by the number of research papers appearing in top-tier journals (10%)
iii) Influence, measured by the number of research papers appearing in highly-influential journals (10%)
iv) Citations, measured by the number of highly-cited research papers (10%)

Source : https://cwur.org/methodology/world-university-rankings.php

On remarque que dans la base de données, on disposait de plusieurs variables qui ne sont pas utilisées dans le vrai modèle (national_rank et patents) et d'un autre côté, on ne dispose pas de la variable "Research Output". Ceci explique le fait qu'on n'a pas pu trouver le modèle linéaire exact, mais on a tout de même réussi à trouver une bonne approximation.

Conclusion :

En guise de conclusion, bien que les modèles GLM semblent être les plus logiques à utiliser, en réalité les classements utilisent des modèles linéaires, les mêmes que ceux qu'on a trouvé, à part pour le classement CWUR pour lequel on ne disposait pas de toutes les variables utilisées.

VI. Validation des critiques :

1. Classement Shanghai :

a. Hypothèse :

Ce classement se concentre sur la force de recherche en sciences dures, et déprécie les sciences humaines.

b. Analyse :

On a établi dans la partie précédente que le classement Shanghai se base sur un modèle linéaire pour classer les universités. Revisitons les résultats trouvés :

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0059863  0.0164071  -0.365    0.716
alumni       0.1027350  0.0002504 410.326   <2e-16 ***
award        0.2057896  0.0002744 749.895   <2e-16 ***
hici         0.2050699  0.0004029 509.003   <2e-16 ***
ns           0.2062086  0.0005517 373.741   <2e-16 ***
pub          0.2059984  0.0003921 525.328   <2e-16 ***
pcp          0.1027814  0.0004474 229.742   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Toutes les variables sont significatives et ayant toutes les mêmes coefficients sauf pour “alumni” et “pcp”. Voici un rappel des variables :

- Alumni: Alumni Score, basé sur le nombre des étudiants qui ont gagné un prix Nobel ou bien d’autres médailles.
- Award: basé sur le nombre de personnel de l’institution qui ont gagné des prix Nobel en physique, chimie, médecine et l’économie et autres médailles en mathématique.
- Hici : HiCi Score, basé sur le nombre de chercheurs les plus cité et sélectionné par Thomson Reuters, un index scientifique.
- Ns :N&S Score, basé sur le nombre d’articles publiés et qui ont une relation avec la nature et la science.
- Pub :PUB Score, basé sur le nombre total de papiers indexés en “ the Science Citation Index-Expanded and Social Science Citation Index”
- Pcp :PCP Score,le score des cinq indicateurs divisés par le nombre du personnel permanents de l’institution.

En surbrillance les mots-clés en relation avec les sciences dures. On remarque que toutes les variables à coefficients élevés se basent sur des indicateurs scientifiques.

c. Conclusion :

Il n’y a aucun doute que le critère principal pour être bien classé dans le système Shanghai est d’être une université performante en sciences dures. C’est à dire avoir un cadre académique et des étudiants qui ont gagné des prix scientifiques, ainsi que des chercheurs et des publications scientifiques cités dans des index scientifiques prestigieux.

2. Classement Times :

a. Hypothèse :

Ce classement sous-évalue les institutions non-anglo-saxonnes.

b. Analyse :

Test Anova :

L’analyse de la variance (ANOVA) est une méthode qui permet d’étudier la variation de la moyenne du phénomène étudié (variable quantitative qui est dans notre cas le score) selon l’influence d’un ou de plusieurs facteurs d’expérience qualitatifs (Dans notre cas on cherche à voir l’influence du pays sur le score).

```
> summary(anova_one_way)
             Df Sum Sq Mean Sq F value Pr(>F)
country      25   6390   255.6   1.862 0.0114 *
Residuals   168  23066   137.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

P value doit être inférieure à 0.05 (Le seuil de signification) pour dire que le mean du score est différent pour chaque pays, ce qui est vrai dans notre cas. On peut confirmer cela en calculant le mean du score pour chaque pays en 2011:

```
> sort(tapply( times111$total_score,times111$country, mean))
                Austria                  Turkey                 Denmark             Netherlands 
               46.95000                47.70000                51.06667                51.48000 
                  Egypt                   Spain             New Zealand                  Taiwan 
               51.60000                51.60000                51.80000                52.45000 
                Germany                  Norway                  Sweden                 Belgium 
               52.45714                52.70000                53.90000                54.60000 
           South Africa                 Finland                   China Republic of Ireland 
               56.10000                56.60000                58.58333                58.90000 
            South Korea          United Kingdom                  Canada               Australia 
               59.37500                59.59630                59.80000                59.81429 
                  Japan               Singapore                  France             Switzerland 
               60.46000                60.95000                61.82500                62.08333 
              Hong Kong United States of America 
               63.80000                66.80870 
```

Pour l'année 2011, en utilisant une liste des pays qui enseignent en anglais on retrouve

```
> for (i in seq(1,200))
+ { if (times111$country[i] == "United States of America"|
+       times111$country[i] == "Canada"| times111$country[i] == "United Kingdom"|
+       times111$country[i] == "Australia"|times111$country[i] == " Unted Kingdom" |
+       times111$country[i] == "South Africa")
+   times111$english[i] = 1
+ else times111$english[i] = 0
+ }
> table(times111$english)

  0   1 
 82 118 
```

Mais ce n'est pas suffisant car il existe des exceptions même dans les pays qui n'enseignent pas en anglais. Après une étude plus exhaustive, on trouve au total 155 des 200 universités sont anglo-saxonnes. Donc 77,5% des universités incluses dans le classement sont anglo-saxonnes.

Pour mettre cette valeur en perspective, on la compare au pourcentage des universités anglo-saxonnes dans le classement CWUR pour l'année 2015. On trouve que juste 45,8% enseignent en anglais.

c. Conclusion :

En comparant le pourcentage des universités anglo-saxonnes incluses dans ce classement avec celui des autres classements on peut confirmer l'hypothèse que le classement Times privilégie les universités qui enseignent en anglais.

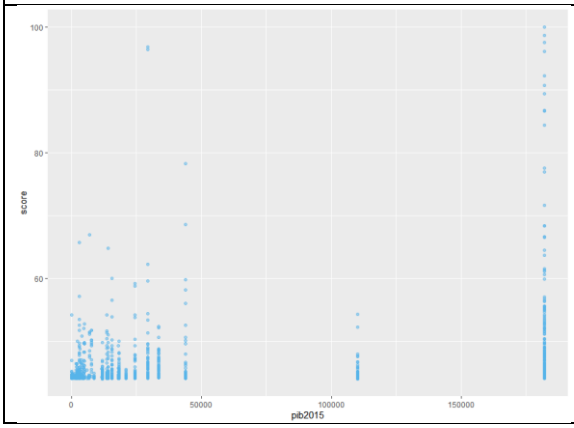
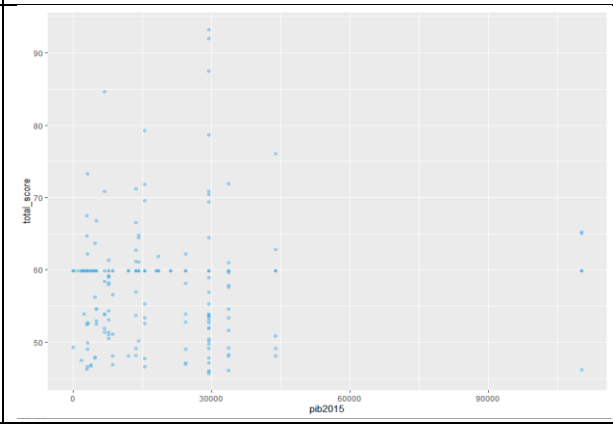
3-Classement cwur :

a. Hypothèse :

Ce classement privilège les pays les plus riches.

b. Analyse :

On remarque que les universités les bien classées dans le monde selon ce classement sont dans 10 pays (United States of America, United kingdom, Japan, China, Germany, Canada). Par suite, on suppose que le PIB représente aussi un critère de classement. Afin de vérifier ce critère On ajoute le PIB de l'année 2015 et on trouve le score de times2015 et de cwur2015 en fonction du PIB par habitant

Cwur	Times
	

Pour le classement cwur on remarque qu'un le plus grand score correspond à des pays où le PIB est élevé, par contre la plupart des score minimales appartiennent à des pays où le PIB est très petit.

Pour le classement times on trouve que le PIB n'influence pas le score.

c. Conclusion :

On peut dire en comparant les deux graphes que le classement cwur privilège les universités qui existent dans des pays avec un PIB très élevé. Par suite, l'hypothèse est vrai pour cwur.

VII. Conclusion :

Pendant presque deux mois, on était amené à réaliser une étude statistique sous la surveillance et le soutien de nos encadrants, portant sur la compréhension, la critique et la modélisation de trois systèmes de classement différents.

Au début, nous avons analysé les bases de données qui contiennent plusieurs variables explicatives selon le classement, et la variable à expliquer est le score qui est continue. Avant d'entamer n'importe quelle étude, il a été indispensable de nettoyer la base de données. Et à travers une étude statistique descriptive on a pu comprendre la répartition des données, leurs distributions géographiques, les différents aspects de la distribution de notre variable dépendante ainsi que sa relation avec les autres variables. Après nous avons analysé les corrélations linéaires entre les variables afin de déceler toute dépendance entre elles. Cette étape était très cruciale et nous a permis une bonne compréhension des données auxquelles nous faisons face.

Ensuite, nous nous sommes lancés dans la modélisation des données à travers plusieurs types de modèles et on a essayé d'améliorer les résultats obtenus à chaque fois. En utilisant plusieurs méthodes (modèles linéaires généralisés,

modèles linéaires, arbres de décisions, classifications), nous avons pu avoir une compréhension plus approfondie des données et nous somme parvenu à trouver les modèles exacts qu'on a validé à partir des sites officiels des classements.

Au début de l'étude on cherchait à comprendre les systèmes de classement et d'analyser la pertinence des divers critiques contre leurs critères. Donc la dernière étape était de valider les critiques adressés à chacun des systèmes de classements maintenant qu'on avait une compréhension approfondie de chacun parmi eux. Et ce qu'il faut retenir d'après cette étude est que ces systèmes présentent souvent des biais de jugements, donc il faut prendre leurs évaluations avec un grain de sel

Ce travail nous a était très bénéfique en termes d'apprentissage et d'élargissement de connaissance sur plusieurs aspects :

- Aspect statistique : Développement d'une bonne intuition nécessaire pour nuancer entre les différentes méthodes statistiques utilisées.
- Aspect analytique : On a été amené au bout de chaque résultat à donner des interprétations propices.
- Aspect collaboratif : la notion du travail collaboratif reste omniprésente.

Liste des figures