كلية العلوم – القنيطرة
Faculté des Sciences – Kenitra

Master Big Date et Cloud Computing

Data Mining

Rapport de mini-projet

Pr : Idriss Moumen

Réaliser par : Bouroummana Ismail

PLAN

Introduction

Dans le cadre de notre formation en Master Big Data et Cloud j'ai l'occasion de réaliser un mini-projet dans le domaine de Data mining,

Ce domaine est une composante essentielle des technologies Big Data et des techniques d'analyse de données volumineuses.

Il s'intéresse à l'extraction des connaissances à partir des donnés.

Durant la réalisation de ce mini projet j'ai utilisé les différentes notions et techniques du data mining.

Dans ce rapport on va implémenter un modèle de classification linéaire (logistic regression) avec le langage R.

En somme :

- On va apprendre à lire un jeu de données, à le visualiser de différentes manières, à sélectionner et nettoyer nos données.
- On va mettre en place une régression logistique qui va nous permettre de résoudre des problèmes de classification.

Prétraitement de données

1. Description du problème

Depuis sa création, la célèbre école de sorciers, Hogwarts, n'avait jamais connu un tel délit.

Les forces du mal ont ensorcelé le Choixpeau.

Ce Choixpeau ne répond plus, il est incapable de trier les nouveaux élèves vers les 4 maisons de l'école.

La nouvelle année scolaire approche et il est impossible pour **Hogwarts** de ne pas accueillir de nouveaux étudiants.

McGonagall a décidé de faire appel un datascientist pour régler ce probléme, En basant sur les données des étudiants.

2. Description des données

- Dataset contient 1600 observations.
- Chaque observation correspond à un élève de l'école.
- Pour chaque élève on a différentes informations dans notre Dataset.
- Parmi ces informations on distingue :

a. House : À **Hogwarts**, il existe quatre maisons distinctes : « Gryffondor », « Poufsouffle », « Serdaigle » et « Serpentard »

b. Firstname

c. Lastname

d. Birthday

e. Best Hand

f. Courses :

- Arithmancy
- Astronomy
- Herbology
- Transfiguration
- Defence Against the Dark Arts
- Divination
- flying
- Muggle Studies
- Ancient Runes
- History of magic
- Potions
- Care of magical creatures
- charms

3. Chargement des données dans R

```
readData <- function()
{
  tryCatch(
    df <- read.csv("./dataset_train.csv",header=TRUE),
    msgErr <- paste("Error: File","./data.csv ","doesn't exist",sep = " "),
    error = function(e) {print(msgErr)},
    warning = function(w) {print(msgErr)}
  )
  return (df)
}
df <- readData()
```

Index	Hogwarts.House	First.Name	Last.Name	Birthday	Best.Hand	Arithmancy	Astronomy	Herbology	Defense.Against.the.Dark.Arts	Divination	Muggle.Studies	Ancient.Runes	History.of.Magic	Transfiguration	Potions	Care.of.Magical.Creatures	Charms	Flying
1358	Gryffindor	Leif	Cantu	1998-02-08	Left	44888	1016.2119	-2.9887054	-10.162119	4.990	-703.4684734	601.2798	-5.3269468	938.4623	4.15085793	-1.103804875	-256.4497	256.09
308	Hufflepuff	Jovita	Regalado	1999-04-07	Right	13733	970.6177	5.7832078	-9.706177	6.736	-709.6679238	394.6110	7.3072758	1055.0903	10.20582696	-0.417413950	-244.0326	43.22
487	Gryffindor	Addie	Pedroza	1999-04-06	Right	32033	956.4844	-4.2366708	-9.564844	5.863	-336.0640951	604.3111	-3.4174427	937.8780	5.00904932	0.458752073	-253.7333	222.84
820	Gryffindor	Sidney	Alcorn	1997-12-03	Left	-2464	940.3961	2.9243478	-9.403981	6.958	-262.1416646	435.1914	9.9138277	1018.8519	11.82777763	0.371717746	-242.5201	39.62
1488	Hufflepuff	Lola	Clemons	1999-03-17	Right	20132	873.4851	3.5902832	-8.734851	6.501	-381.9040643	406.2925	5.9978366	1038.8797	7.73545118	-0.213175580	-244.9104	29.62
1177	Gryffindor	Lowell	Lerma	1999-06-05	Left	35676	870.0722	-3.1929928	-8.700722	5.747	-688.0766695	584.8399	-6.1901866	932.4009	2.19777937	0.703536089	-254.8599	260.75
1312	Hufflepuff	Curtis	Nunez	1997-09-17	Left	19531	860.8582	5.4817107	-8.608582	5.329	-371.1708514	345.0504	4.7344780	1039.4097	4.36091189	0.436388461	-245.7926	30.00
1067	Gryffindor	Mamie	Hightower	1997-12-28	Left	64197	849.1319	-0.4183538	-8.491319	6.582	-442.0018219	689.6572	-7.4299149	955.8738	3.27606715	-2.387181485	-251.9111	270.27
384	Gryffindor	Rachael	Coyle	2000-10-14	Right	56746	842.5382	-4.0490537	-8.425382	3.920	-772.0903291	626.9812	-4.9549426	940.6669	5.31146380	-1.243991900	-256.7991	240.85
196	Hufflepuff	Julianna	Doran	2000-08-01	Right	28482	825.5019	8.5735614	-8.255019	2.796	-148.6423185	323.6683	1.1606146	1020.9309	-1.09147950	-1.084252025	-247.4226	75.19
315	Gryffindor	Jimmie	Mckeown	1998-04-25	Left	44314	820.7534	-5.6717402	-8.207534	3.785	-643.2946212	572.2815	-5.1191658	912.5344	2.03792590	-1.733968668	-258.0231	234.22
1045	Gryffindor	Rodrick	Lemons	2001-03-15	Right	42075	814.0572	-3.3733995	-8.140572	5.012	-599.1452454	649.4006	-5.1396113	948.4760	6.19655332	0.985468034	-253.6724	257.42
1252	Hufflepuff	Murray	Evans	1996-11-27	Left	27490	813.0633	5.4618712	-8.130633	5.011	-466.1214918	361.4712	1.8117965	1049.9508	3.38166193	-0.426739378	-246.9474	52.84
969	Hufflepuff	Corey	Villarreal	1997-12-26	Left	28654	808.2610	6.9331383	-8.082610	2.348	-143.3897942	317.6980	6.2305630	1024.0582	3.04651265	0.310679296	-246.8798	10.13
1365	Hufflepuff	Fidel	Ginn	2000-12-10	Left	28456	806.6085	5.5836069	-8.066085	3.646	-100.1684434	348.0137	5.9017700	1020.8832	3.45022206	0.156932127	-246.3140	12.97
667	Hufflepuff	Leilani	Crisp	1997-09-21	Left	15925	805.5828	5.3393146	-8.055828	6.095	-350.6845090	360.1667	6.9301883	1035.5423	5.95379429	1.350978190	-243.6573	14.97
760	Hufflepuff	Alaina	Lemay	1999-12-13	Left	28583	802.7269	2.9359231	-8.027269	8.644	-332.4425542	328.7538	8.1218481	1046.2646	4.52389829	-0.221335972	-243.6330	-65.78
876	Hufflepuff	Jody	Obrien	2001-02-09	Right	27663	796.0466	3.3474175	-7.960466	NA	-346.0858850	411.7428	4.6646741	1038.6302	7.03448092	0.486110044	-246.3994	36.09
744	Gryffindor	Rosalba	Timmerman	1998-01-04	Right	49913	794.9030	-2.2149851	-7.949030	6.268	-767.7463351	598.3041	-7.4432778	934.3434	0.50042345	1.337030978	-254.4445	263.53
87	Hufflepuff	Shanna	Reuter	2000-02-06	Left	32491	793.0760	5.6138960	-7.930760	4.977	-533.3751884	428.3497	7.1111352	1041.2456	9.23555349	0.245646194	-244.5067	35.69
1159	Gryffindor	Max	Craft	1996-11-30	Left	35454	791.2177	-2.2772903	-7.912177	6.672	-682.6432681	621.7268	-6.5702578	941.9620	3.06492473	-0.920632405	-252.3502	273.96
152	Gryffindor	Jeff	Ernst	1998-02-23	Left	32262	791.1833	-2.6006438	-7.911833	8.042	-542.2729290	536.7796	-6.2720482	937.8606	-1.40488371	0.451197396	-252.1266	221.15
1235	Hufflepuff	Dwight	Jamieson	1998-10-30	Right	32303	788.1177	6.5251978	NA	4.154	-481.6983604	400.6345	3.2433311	1042.9832	5.44321415	0.344548301	-246.1315	66.70
425	Gryffindor	Delilah	Cade	2001-03-18	Left	21274	786.9514	-4.9320493	-7.869514	6.533	-226.2429992	564.6225	-3.3963163	941.5808	3.21110263	-0.077073380	-251.8529	194.36
313	Hufflepuff	Lynn	Tovar	2000-01-15	Right	22301	785.2455	6.8177400	-7.852455	4.398	-437.5178744	446.8921	2.8603664	1029.5012	6.61601427	0.927903404	-244.7866	113.99
556	Gryffindor	Chester	Fry	2000-01-13	Left	62636	783.2780	-5.4739514	-7.832780	2.576	NA	582.8451	-5.5340453	922.5584	1.77365800	-0.321919648	-259.1816	212.83
377	Hufflepuff	Armand	Bolling	1998-11-13	Left	6407	773.4238	6.0795444	-7.734238	7.233	-393.2799378	391.5263	5.6095468	1031.8274	5.96636468	-1.768941833	-241.7049	58.73
1524	Hufflepuff	Alberto	Radford	2000-09-23	Left	45610	769.8166	3.3208157	-7.698166	4.901	-470.4141771	NA	2.6043265	1052.0808	5.86085630	1.119573810	-247.8753	32.10
748	Hufflepuff	Malcolm	Mcqueen	1997-01-02	Right	24891	767.3346	4.8996946	-7.673346	4.845	-170.4219886	354.5007	8.9487568	1024.1392	5.96255382	0.434290590	-244.2249	-19.90
1360	Hufflepuff	Dianne	Vaught	2000-08-14	Left	27198	764.0716	5.2434345	-7.640716	5.404	-357.4384517	347.3995	6.2149278	1040.7550	4.74299234	-0.746694446	-244.9033	-1.87

Showing 1 to 31 of 1,600 entries, 19 total columns

4. Summarization des données dans R

```
> summary(df)
     Index        Hogwarts.House      First.Name         Last.Name          Birthday          Best.Hand          Arithmancy     
 Min.   :   0.0   Length:1600        Length:1600        Length:1600        Length:1600        Length:1600        Min.   :-24370  
 1st Qu.: 399.8   Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.: 38512  
 Median : 799.5   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median : 49014  
 Mean   : 799.5                                                                                                 Mean   : 49635  
 3rd Qu.:1199.2                                                                                                 3rd Qu.: 60811  
 Max.   :1599.0                                                                                                 Max.   :104956  
                                                                                                                NA's   :34      
   Astronomy        Herbology       Defense.Against.the.Dark.Arts   Divination      Muggle.Studies    Ancient.Runes    History.of.Magic
 Min.   :-966.7   Min.   :-10.296   Min.   :-10.1621              Min.   :-8.727   Min.   :-1086.5   Min.   :283.9   Min.   :-8.859  
 1st Qu.:-489.6   1st Qu.: -4.308   1st Qu.: -5.2591              1st Qu.: 3.099   1st Qu.: -577.6   1st Qu.:397.5   1st Qu.: 2.219  
 Median : 260.3   Median :  3.469   Median : -2.5893              Median : 4.624   Median : -419.2   Median :463.9   Median : 4.378  
 Mean   :  39.8   Mean   :  1.141   Mean   : -0.3879              Mean   : 3.154   Mean   : -224.6   Mean   :495.7   Mean   : 2.963  
 3rd Qu.: 524.8   3rd Qu.:  5.419   3rd Qu.:  4.9047              3rd Qu.: 5.667   3rd Qu.:  255.0   3rd Qu.:597.5   3rd Qu.: 5.825  
 Max.   :1016.2   Max.   : 11.613   Max.   :  9.6674              Max.   :10.032   Max.   : 1092.4   Max.   :745.4   Max.   :11.890  
 NA's   :32       NA's   :33        NA's   :31                    NA's   :39       NA's   :35        NA's   :35      NA's   :43      
 Transfiguration     Potions        Care.of.Magical.Creatures     Charms           Flying        
 Min.   : 906.6   Min.   :-4.697   Min.   :-3.31368          Min.   :-261.0   Min.   :-181.470  
 1st Qu.:1026.2   1st Qu.: 3.647   1st Qu.:-0.67161          1st Qu.:-250.7   1st Qu.: -41.870  
 Median :1045.5   Median : 5.875   Median :-0.04481          Median :-244.9   Median :  -2.515  
 Mean   :1030.1   Mean   : 5.950   Mean   :-0.05343          Mean   :-243.4   Mean   :  21.958  
 3rd Qu.:1058.4   3rd Qu.: 8.248   3rd Qu.: 0.58992          3rd Qu.:-232.6   3rd Qu.:  50.560  
 Max.   :1099.0   Max.   :13.537   Max.   : 3.05655          Max.   :-225.4   Max.   : 279.070  
 NA's   :34       NA's   :30       NA's   :40                                                   
```

5. Gestion des données manquantes dans R

D’abord on va calculer le pourcentage de données manquantes pour chaque attribut.

```
> p <- function(x) {sum(is.na(x))/length(x)*100}
> apply(df, 2, p)
                      Index                Hogwarts.House
                     0.0000                        0.0000
                 First.Name                   Last.Name
                     0.0000                        0.0000
                   Birthday                   Best.Hand
                     0.0000                        0.0000
                 Arithmancy                   Astronomy
                     2.1250                        2.0000
                  Herbology Defense.Against.the.Dark.Arts
                     2.0625                        1.9375
                 Divination                Muggle.Studies
                     2.4375                        2.1875
              Ancient.Runes             History.of.Magic
                     2.1875                        2.6875
             Transfiguration                      Potions
                     2.1250                        1.8750
  Care.of.Magical.Creatures                      Charms
                     2.5000                        0.0000
                    Flying
                     0.0000
```

Par exemple l’attribut **Flying** ne contient aucune valeur manquante, Par contre l’attribut **Divination** contient 2,43 % des valeurs manquantes

Mean imputation

Voici la syntaxe da la fonction **Mice** qui implémente plusieurs méthodes pour gérer les données manquantes.

```
> impute <- mice(df, m=3)

 iter imp variable
  1   1  Arithmancy  Herbology  Defense.Against.the.Dark.Arts  Divination  Muggle.Studies  Ancient.Runes
ry.of.Magic  Transfiguration  Potions  Care.of.Magical.Creatures
  1   2  Arithmancy  Herbology  Defense.Against.the.Dark.Arts  Divination  Muggle.Studies  Ancient.Runes
ry.of.Magic  Transfiguration  Potions  Care.of.Magical.Creatures
  1   3  Arithmancy  Herbology  Defense.Against.the.Dark.Arts  Divination  Muggle.Studies  Ancient.Runes
ry.of.Magic  Transfiguration  Potions  Care.of.Magical.Creatures
  2   1  Arithmancy  Herbology  Defense.Against.the.Dark.Arts  Divination  Muggle.Studies  Ancient.Runes
ry.of.Magic  Transfiguration  Potions  Care.of.Magical.Creatures*
  2   2  Arithmancy  Herbology  Defense.Against.the.Dark.Arts  Divination  Muggle.Studies  Ancient.Runes
ry.of.Magic  Transfiguration  Potions  Care.of.Magical.Creatures
  2   3  Arithmancy  Herbology  Defense.Against.the.Dark.Arts  Divination  Muggle.Studies  Ancient.Runes
ry.of.Magic  Transfiguration  Potions  Care.of.Magical.Creatures*
  3   1  Arithmancy  Herbology  Defense.Against.the.Dark.Arts*  Divination*  Muggle.Studies*  Ancient.Rune
story.of.Magic*  Transfiguration*  Potions  Care.of.Magical.Creatures*
  3   2  Arithmancy  Herbology  Defense.Against.the.Dark.Arts  Divination  Muggle.Studies  Ancient.Runes
```

- En exécutant la fonction **mice** sur notre data on obtient les 3 choix pour chaque attribut .
- Par exemple , si on prend la ligne 21 de l'attribut **Arithmancy** on voie que a 3 valeurs 45029 , 44965 , 45055 .
- Alors il faut choisir l'un des trois valeurs , c'est pour cela on va calculer le mean de chaque choix .
- Comme vous voyez on obtient les 3 moyennes 52713,06 52595,79 51259,79 .
- Alors on va choisir la valeur la plus proche du moyenne de l'attribut .
- On a choisi le 3 eme choix pour obtenir les valeurs manquantes dans notre data

```
> head(impute$imp$Arithmancy)
        1      2      3
21  45029 44965 45055
71  57277 57109 57294
136 77257 76651 77286
160 61581 61754 61931
169 89262 91006 63666
177 35060 33672 33560
> summary(df$Arithmancy)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 -24370   38512   49014   49635   60811  104956      34
> apply(impute$imp$Arithmancy,2,mean)
       1        2        3
52713.06 52595.79 51259.79
> newDATA <- complete(impute, 3)
```

Notre data alors est prêt pour la partie suivante

Analyse de données exploratoire

1. Visualisation des données dans R

La visualisation des données est un outil puissant pour le datascientist.

Cela vous permet d'acquérir une intuition sur comment les données sont connectées les unes aux autres.

Visualiser vos données vous permet aussi de déceler plusieurs défauts.

À partir de cette visualisation, on va choisir les attribues pour entraîner notre prochaine régression logistique

On va implémenter 2 méthodes de visualisation :

a) ***Histogram***
b) ***Scatter plot***

a. Histogram

Notre objectif est de répondre sur la question suivante :

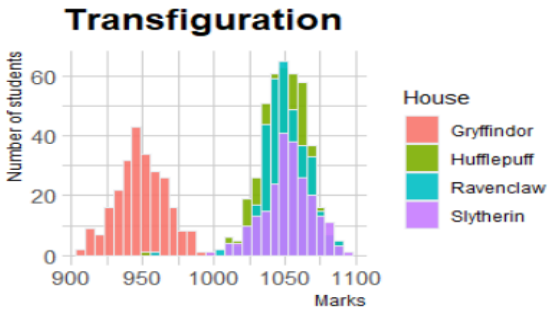
Quel cours de **Hogwarts** a une répartition des notes homogènes entre les quatres maisons ?

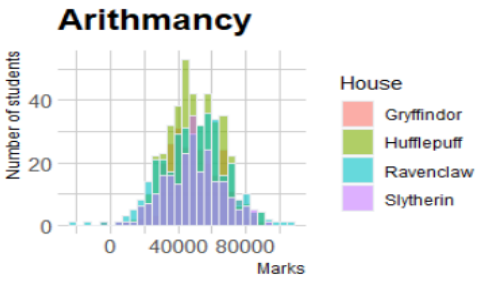
C'est pour cela on va tracer à chaque cour un histogramme, et on va choisir le cours le plus homogènes.

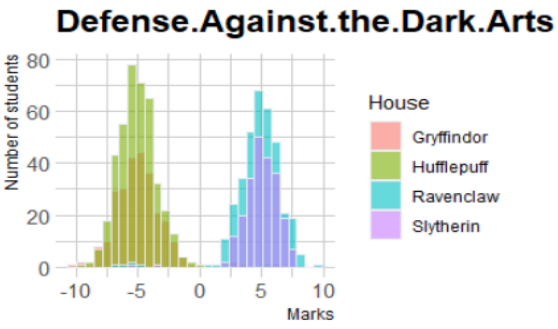
On prend l'exemple des 4 cours suivantes :

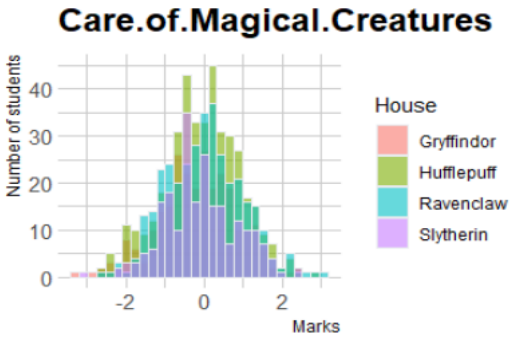
- Transfiguration
- Defense.Against.the.Dark.Arts
- Arithmancy
- Care.of.Magical.Creatures

```
library(ggplot2)
library(hrbrthemes)
 ggplot(obj, aes(x=Care.of.Magical.Creatures, fill=Hogwarts.House)) +
  geom_histogram( color="#e9ecef", alpha=0.6, position = 'identity',bins=30) +
  theme_ipsum() +
  labs(title="Care.of.Magical.Creatures",fill="House",x="Marks",y="Number of students")
```







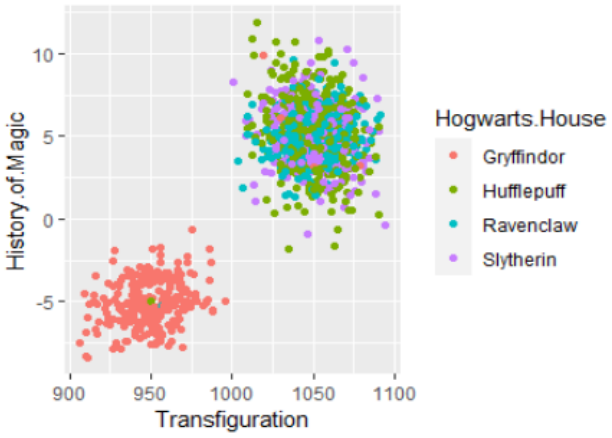


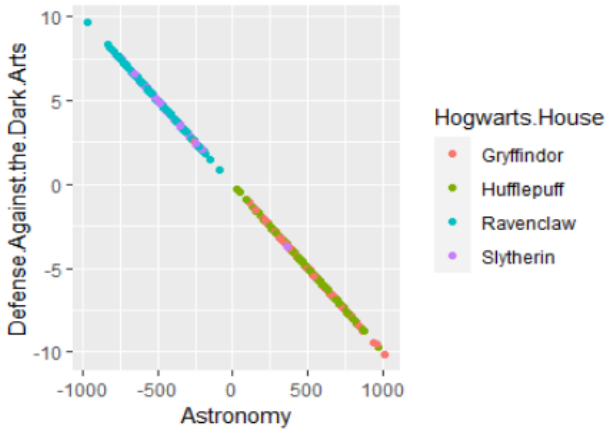
D'après les 4 histogrammes on déduit que **Transfiguration** et **Defense Against the dark arts** ne sont pas homogènes par contre les deux autres cours sont homogènes

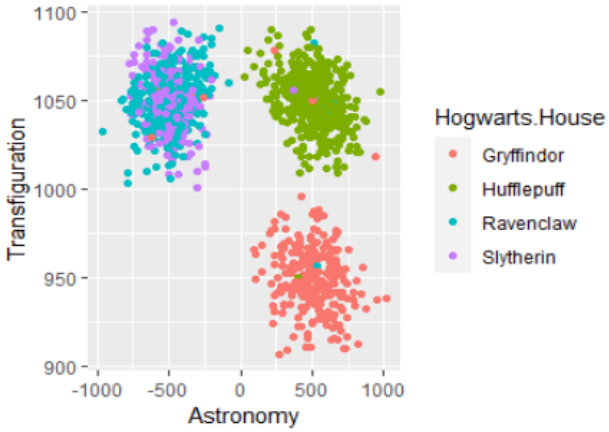
b. Scatter Plot

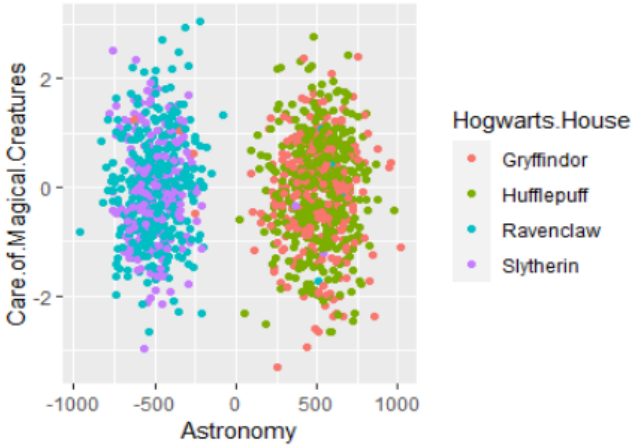
- Un graphique de points tracés qui montre la relation entre deux attributs de la même Dataset.
- Notre objectif est de répondre sur la question suivante :
  - ➢ Quelles sont les deux cours qui sont semblables ?
- Pour répondre sur cette question on va tracer des « Scatter Plot »

```
library(ggplot2)
ggplot(obj, aes(x = Astronomy, y = Care.of.Magical.Creatures, color = Hogwarts.House)) +
 geom_point()
```









On déduit que **Astronomy** et **Defense Against the dark arts** sont semblables par contre les autres cours ne sont pas semblables.

Construction de modèles prédictifs

1. Logistic Regression

- La régression logistique est une technique prédictive.
- On distingue 2 types de régression logistique :
  - La régression logistique ordinaire ou régression logistique binaire vise à expliquer une variable d'intérêt binaire (c'est-à-dire de type « oui / non » ou « vrai / faux »). Les variables explicatives qui seront introduites dans le modèle peuvent être quantitatives ou qualitatives.
  - La régression logistique multinomiale est une extension de la régression logistique aux variables qualitatives à trois modalités ou plus, la régression logistique ordinale aux variables qualitatives à trois modalités ou plus qui sont ordonnées hiérarchiquement
- Dans notre cas on va travailler par la régression logistique multinomial

2. Training model

On va travailler par la Library **nnet** pour implémenter le modèle prédictif

D'abord il faut changer l'attribut **Howarts.House** qui est notre **Target** comme un Factor

Puis il faut définir la référence de notre model, dans notre cas j'ai choisi la maison **Grynffindor**

```
library(nnet)
obj$Hogwarts.House = as.factor(obj$Hogwarts.House)
obj$Hogwarts.House <- relevel(obj$Hogwarts.House, ref = "Gryffindor")
mymodel <- multinom(Hogwarts.House~.-First.Name -Last.Name-Birthday-Best.Hand-Index , data = obj)
mymodel
```

```
> mymodel <- multinom(Hogwarts.House~.-First.Name -Last.Name-Birthday-Best.Hand-Index , data = obj)
# weights:  60 (42 variable)
initial  value 1734.254246
iter  10 value 276.905191
iter  20 value 254.953418
iter  30 value 174.521265
iter  40 value 130.239374
iter  50 value 125.298111
iter  60 value 125.062968
iter  70 value 125.015161
iter  80 value 125.011291
final  value 125.009041
converged
```

3. Summarization du model prédictif

La fonction **summary** donne des informations de notre modèle prédictif comme les attributs utilisés ainsi les erreurs de chaque attribut

```
> summary(mymodel)
Call:
multinom(formula = Hogwarts.House ~ . - First.Name - Last.Name -
    Birthday - Best.Hand - Index, data = obj)

Coefficients:
           (Intercept)   Arithmancy    Astronomy   Herbology Defense.Against.the.Dark.Arts
Hufflepuff -3.879588e-04 5.462484e-05  0.001023770  0.30084366                -1.023772e-05
Ravenclaw  -3.702093e-05 7.605456e-05 -0.002804111 -0.06048854                 2.804119e-05
Slytherin  -4.796734e-03 1.596634e-04 -0.005820304 -0.16818462                 5.820300e-05
           Divination Muggle.Studies Ancient.Runes History.of.Magic Transfiguration   Potions
Hufflepuff -0.2665814    0.0001049543   -0.02946548       0.08185191      0.01318936 0.1897988
Ravenclaw  -0.4713519    0.0045810771   -0.02459571       0.17471560      0.04188871 0.1136025
Slytherin  -0.5608690    0.0035647639   -0.06828988      -0.03556256      0.06145938 1.0444467
           Care.of.Magical.Creatures      Charms      Flying
Hufflepuff                0.01627779 0.005385287 0.001874628
Ravenclaw                -0.13424365 0.131877456 0.024943165
Slytherin                -0.14529815 0.173451169 0.056625595

Std. Errors:
            (Intercept)   Arithmancy    Astronomy   Herbology Defense.Against.the.Dark.Arts
Hufflepuff 3.544714e-05 2.839315e-05 0.001808159 0.005716880                1.808159e-05
Ravenclaw  4.654066e-05 2.840877e-05 0.001693579 0.009802524                1.693579e-05
Slytherin  1.409819e-05 3.568200e-05 0.002102249 0.003476673                2.102249e-05
           Divination Muggle.Studies Ancient.Runes History.of.Magic Transfiguration     Potions
Hufflepuff 0.002510965    0.002025846    0.01146458     0.0008031504     0.005579306 0.001281121
Ravenclaw  0.005024593    0.002186330    0.01200201     0.0013555496     0.006517744 0.002185218
Slytherin  0.002715255    0.002764485    0.01642508     0.0014005642     0.006931703 0.001873471
           Care.of.Magical.Creatures      Charms     Flying
Hufflepuff               1.346319e-04 0.014993447 0.01236457
Ravenclaw                8.172894e-05 0.022585642 0.01354879
Slytherin                9.883684e-05 0.007863043 0.01780954

Residual Deviance: 250.0181
AIC: 328.0181
```

4. Model Testing

On va calculer notre **cost fucntions errors** qui nous aide à savoir les attributs qui n'aide pas notre modèle prédictif.

Tous les valeurs qui sont supérieur à 0.05 sont des attributs non significatifs pour notre modèle prédictif.

Et aussi nous donne les attributs qui sont semblable comme **Astronomy** et **Defense.Against.the.dark.arts.**

Donc il faut répéter l'entrainement du modèle sans les attributs qu'on a obtenu et aussi les attributs qui sont semblable il faut prendre juste un.

```
> z <- summary(mymodel)$coefficients/summary(mymodel)$standard.errors
> p <- (1- pnorm(abs(z),0,1)) * 2
> p
           (Intercept)   Arithmancy   Astronomy    Herbology Defense.Against.the.Dark.Arts Divination
Hufflepuff   0.0000000 5.437038e-02 0.571261382 0.000000e+00                  0.571260608         0
Ravenclaw    0.4263498 7.425114e-03 0.097776281 6.798373e-10                  0.097775397         0
Slytherin    0.0000000 7.654712e-06 0.005629627 0.000000e+00                  0.005629663         0
           Muggle.Studies Ancient.Runes History.of.Magic Transfiguration Potions Care.of.Magical.Creatures
Hufflepuff     0.95868196  1.016600e-02               0    1.807989e-02       0                         0
Ravenclaw      0.03614194  4.043290e-02               0    1.302560e-10       0                         0
Slytherin      0.19722933  3.215269e-05               0    0.000000e+00       0                         0
               Charms      Flying
Hufflepuff 7.194634e-01 0.879492262
Ravenclaw  5.251654e-09 0.065623320
Slytherin  0.000000e+00 0.001475243
```

5. Predicting the values for train dataset

On a utilisé la fonction **predict** pour prédire notre **Target**.

```
> pred <- predict(mymodel,obj)
> head(pred)
[1] Ravenclaw  Slytherin  Ravenclaw  Gryffindor Slytherin  Gryffindor
Levels: Gryffindor Hufflepuff Ravenclaw Slytherin
> head(obj$Hogwarts.House)
[1] Ravenclaw  Slytherin  Ravenclaw  Gryffindor Slytherin  Gryffindor
Levels: Gryffindor Hufflepuff Ravenclaw Slytherin
```

6. Confusion Matrix & Accuracy

La matrix de Confusion nous donne les valeurs correctes de notre modèle prédictif et aussi les échecs qui a commis notre modèle prédictif.

```
> tab <- table(pred,obj$Hogwarts.House)
> tab

pred          Gryffindor Hufflepuff Ravenclaw Slytherin
  Gryffindor         253          1         1         0
  Hufflepuff           4        413         4         2
  Ravenclaw            4          1       344         2
  Slytherin            0          0         2       220
> # Calculating accuracy
> acc <- sum(diag(tab))/sum(tab)
> acc
[1] 0.9832134
```

Notre Modèle prédictif a une précision de 98%, alors il a réussi de trier les étudiants par les 4 maisons de l'école.

Conclusion

Pour conclure durant la réalisation de ce mini projet j'ai réussi d'appliquer tous les étapes du datamining, en passant par l'importation, l'interprétation des données, vers la visualisation, puis l'analyse des données puis la création du modèle prédictif.