

# DataExplorationDraft1

July 14, 2017

## 1 Presentation et Exploitation des Données

Dans ce chapitre nous allons exploiter les données mise en notre disposition par les autorités de l'ULPGL. Celles-ci sont issues du système d'information UAT (University Administrative Tool) et pour des raisons de confidentialité nous n'avons pas eu accès à toute la base des données nous avons juste fait une requête des données dont nous avons besoin pour notre étude et l'administrateur a exécuté une requête vers sa base des données et nous a fourni les données dont nous avons besoin pour l'étude sous forme d'un fichier csv (comma separated values). Comme souligné dans le chapitre premier ce chapitre se basera sur la méthodologie CRISP-DM elle sera subdivisée en différentes sections: - L'exploration et la préparation des données - sélection des algorithmes et leur exécution - l'amélioration et optimisation des algorithmes source : Sklean Handbook Appendix 2

### 1.1 Exploration et la préparation des données

source : <https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/> SUNIL RAY, JANUARY 10, 2016

les spécialistes affirment que 70-80 % du temps consacré à un projet dataMining est alloué à la phase de l'exploration et la préparation des données, il n'y a pas de raccourcis pour cette phase et si on l'a pas bien effectué nous risquons de nous retrouver entraînés à améliorer l'exactitude de notre algorithme mais en vain nous serons toujours obligés de retourner à cette phase et toutes ces techniques de l'exploration des données pourront nous venir en aide.

1. Les étapes de la phase d'exploration et la préparation des données Certaines de ces étapes sont mentionnées sur la figure suivante ; source: [http://www.saedsayad.com/data\\_mining\\_map.htm](http://www.saedsayad.com/data_mining_map.htm)

En bref l'exploration des données consiste à se plonger dans le passé pour prédire l'avenir. Souvenons-nous que la qualité de notre entrée détermine la qualité de notre sortie, ces phases nous permettent d'améliorer la qualité de notre entrée en vue d'avoir une bonne sortie. Voici les étapes de cette phase:

- Identification des variables
- Statistique Descriptive
- Analyse Bi-variée
- Traitement des valeurs manquantes
- Traitement des déviations ou outliers

- Transformation des variables
- creation des nouvelles variables

Comme nous l'avons souligné dans le chapitre 1 ce processus est un processus iteratif et incrementale nous executerons cette phase 2 a 5 fois ou plus en vue d'avoir un bon modele

## 1.2 Identification des Données et des Variables

Comme souligné dans la phase d'introduction les données mise à notre disposition sont sous format csv et nous avons utilisé la librairie pandas de python pour faire l'analyse , nous utiliserons aussi d'autres librairies qui nous permettrons de faire les statistiques ainsi que les visualisations :

voici le code pour charger les librairies.

Nous venons de charger les données et nous pouvons remarquer à quoi ils ressemblent

Out [52]: (9606, 18)

Nous remarquons que les données sont stockées dans une structure de type matricielle appelé dataframe. source : <https://pandas.pydata.org/pandas-docs/stable/dsintro.html#dataframe>

un DataFrame selon la documentation officielle de pandas est une structure des données bidimensionnelle avec des colonnes des données des différents types . Il peut être comparé à une feuille de calcul excel ou une table dans SQL

la commande dataset.shape nous a permis de dire que notre ensemble d'apprentissage de départ comprend 9606 lignes et 22 colonnes ! Analysons de plus près les colonnes

```
Out [6]: Index([u'IDENTIFICATION', u'BIRTHDAY', u'NAME', u'DIPLOMDATE', u'DIPLOMTYPE',
               u'DIPLOMMENTION', u'DIPLOMPERCENTAGE', u'DIPLOMSECTION',
               u'DIPLOMOPTION', u'DIPLOMPLACE', u'SCHOOL', u'SCHOOLPROVINCE',
               u'SCHOOLCODE', u'SCHOOLSTATUS', u'ACADYEAR', u'PERC1', u'MENT1',
               u'PERC2', u'MENT2', u'FAC', u'OPT', u'PROM'],
              dtype='object')
```

chaque ligne comprend les informations d'un étudiant pour une année Académique

1 IDENTIFICATION : contient une identification unique et anonyme d'un étudiant les noms et les matricules réels des étudiants ont été cachés pour des raisons de confidentialité

2 BIRTHDAY : contient la date de naissance de chaque étudiant

3 NAME : contient le sexe de chaque étudiant

4 DIPLOMDATE : L'année d'obtention du diplôme

5 DIPLOMTYPE : le type de diplôme

6 DIPLOMMENTION : mention de diplôme

7 DIPLOMPERCENTAGE: le pourcentage du diplôme

8 DIPLOMSECTION: la section du diplôme

9 DIPLOMOPTION : l'option

10 DIPLOMPLACE : l'endroit d'obtention du diplôme

11 SCHOOL : l'école de provenance

12 SCHOOLPROVINCE : la province de provenance

13 SCHOOLCODE : code de l'école

14 SCHOOLSTATUS : le statut de l'école (privé , public , conventionné ,...)

15 ACADYEAR : l'année académique

16 PERC1 : pourcentage en première session

- 17 MENT1 : mention en premier session
- 18 PERC2 : pourcentage en seconde session
- 19 MENT2 : mention en seconde session
- 20 FAC : la faculté
- 21 OPT : l'option
- 22 PROM : la promotion

Comme nous pouvons le constater les colonnes 1-14 regroupent les informations que chaque étudiant donné à son inscription , ils constitueront nos variables d'entrees les restes seront utilisées pour constituer notre variable de sortie

Nous l'avons aussi signalé que chaque ligne comprend les informations d'un étudiant pour une année académique . Pour mener bien notre analyse nous allons grouper les informations de chaque étudiant en une ligne

nos données seront groupées selon les variables d'entrees ensuite les données de sorties seront groupées selon une fonction d'aggrégation prédéfinie

Nous allons d'abord faire une analyse univariée sur les données en entrees !

Avant la phase de traitement des données regardons notre ensemble pour enlever les colonnes sans informations considérables.

Nous avons écrit une fonction qui nous donne un pourcentage des valeurs manquantes pour chaque colonne et de prime à bord nous allons supprimer certaines colonnes qui ne comportent pas des informations.

```
Out [6] : BIRTHDAY          0
          DIPLOMDATE       98
          DIPLOMPLACE      98
          SCHOOLSTATUS     0
          DIPLOMPERCENTAGE  0
          SCHOOLPROVINCE   0
          FAC              0
          ACADYEAR         0
          DIPLOMMENTION    99
          SCHOOLCODE       70
          PERC1            55
          PERC2            36
          DIPLOMOPTION     0
          OPT              0
          SCHOOL           5
          NAME             0
          DIPLOMSECTION    0
          MENT1            0
          PROM             0
          MENT2            24
          IDENTIFICATION    0
          DIPLOMTYPE       0
          Name: % of missing, dtype: int64
```

Avec cette table nous remarquons que 4 colonnes ne nous serviront à rien dans la suite car elles disposent de plus de 70 % des valeurs manquantes et nous devons les supprimer avant de continuer notre analyse

Nous pouvons remarquer que notre ensemble d'apprentissage change de dimension et (7216,22) à (7216,18)

Out[16]: (9606, 18)

```
Out[13]: Index([u'IDENTIFICATION', u'BIRTHDAY', u'NAME', u'DIPLOMTYPE',  
               u'DIPLOMPERCENTAGE', u'DIPLOMSECTION', u'DIPLOMOPTION', u'SCHOOL',  
               u'SCHOOLPROVINCE', u'SCHOOLSTATUS', u'ACADYEAR', u'PERC1', u'MENT1',  
               u'PERC2', u'MENT2', u'FAC', u'OPT', u'PROM'],  
              dtype='object')
```

les colonnes en entrée deviennent les 10 premières colonnes

Pour une première approche nous allons grouper notre ensemble en fonction des données en entrée et ensuite écrire une fonction qui va grouper les données de sortie

Out[35]: (9606, 18)

voici la fonction qui va grouper les données de sortie

Out[37]: (4715, 8)

après groupement en fonction des matricules nous venons de remarquer que notre ensemble comprend 4715 rows et 18 columns et c'est sera notre ensemble pour notre étude  
cette ensemble est subdivisé en variables d'entrée et variables de sortie!

Out[24]: (4715, 10)

Nous allons afficher notre ensemble d'entrée et de sortie ici

Nous venons de finir avec la présentation de nos données nous allons maintenant débuter avec la phase d'analyse préliminaire des données que nous avons en entrée et ensuite nous faisons une analyse des données en sortie en enfin analyser les données de sortie combinées à celles des données ?

### 1.2.1 Analyse des données

Cette phase comprend une analyse statistique bivariée et univariée nous visualiserons les résultats à l'aide des graphiques. Dans cette partie nous utiliserons beaucoup plus la statistique descriptive et inférentielle.

Comme nous pouvons le remarquer notre ensemble d'apprentissage comprend à la fois des données numériques (continues) ainsi que des données discrètes catégorielles. voici comment nous allons procéder

#### Statistique Descriptive

1. variable Numériques ou continues : Pour les données continues nous allons essayer de comprendre la tendance et la dispersion de nos variables. Les métriques utilisées sont sur la figure suivante: En bref nous allons examiner la moyenne, le mode, l'écart-type et la variance, nous compterons aussi les variables nous faisons les visualisations avec des boxplot! cette étape nous sera aussi utile dans le traitement des valeurs manquantes et des outliers!
2. variable catégorielle ou quantitative Pour les données discrètes nous allons les tables des fréquences pour comprendre la distribution de chaque catégorie nous pourrions aussi voir le pourcentage de chaque catégorie, les histogrammes et bar chart seront utilisés.

**Analyse Bivariée** c'est une technique d'analyse statistique des données, consistant à découvrir la relation pouvant exister entre 2 variables dans le but de tester l'hypothèse d'association et de causalité entre 2 variables ! Par exemple dans notre analyse nous allons essayer de voir la relation existant entre le choix de la faculté de le pourcentage du diplôme à l'état. Elle se déroule en 4 étapes : - Définition de la nature des relations - Identification et direction des relations - Détermination si la relation est importante du point de vue statistique (Intervalle de confiance) - Détermination de la force de relation

Source: Becker, William. Uncertainty propagation through large nonlinear models. Diss. University of Sheffield, 2011. de Smith, M. J. "STATSREF: Statistical Analysis Handbook-a web-based statistics." (2015).

Nous effectuons cette analyse à 3 niveaux : 1. Variables Continues et catégorielles ou quantitatives Pour effectuer cette analyse nous utiliserons le test ANOVA (Analyse of variance):

[formule et décision] 2. variable Catégorielles et Catégorielles

Pour ces types de données nous allons effectuer le test de chi carré: Le chi carré est un test statistique conçu pour déterminer si la différence entre deux distributions de fréquences est attribuable à l'erreur d'échantillonnage (le hasard) ou est suffisamment grande pour être statistiquement significative.

$H_0$  - est, comme son nom l'indique, une hypothèse qui postule qu'il n'y a pas de différence entre les fréquences ou les proportions des deux groupes elle est considérée comme hypothèse nulle.

Si la différence entre les deux distributions est réduite, l'hypothèse nulle sera acceptée. Si la différence est grande, l'hypothèse nulle sera rejetée. Dans ce dernier cas, on parlera d'une différence statistiquement significative parce que l'écart entre les deux distributions est trop important pour être expliqué par le hasard seulement : une différence réelle existe donc. [Insérer la formule]

### 3. Variables Continues et Continues

Pour les variables continues on utilise comme la corrélation et pour notre travail nous allons utiliser le coefficient de corrélation de Pearson: Les coefficients de corrélation permettent de donner une mesure synthétique de l'intensité de la relation entre deux caractères et de son sens lorsque cette relation est monotone. Le coefficient de corrélation de Pearson permet d'analyser les relations linéaires et le coefficient de corrélation de Spearman les relations non-linéaires monotones. Il existe d'autres coefficients pour les relations non-linéaires et non-monotones.

Signalons que Python dispose de multiples bibliothèques pour effectuer ces genres d'analyse.

Commençons par l'analyse des données univariées sur les variables d'entrée

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4038 entries, 0 to 4037
Data columns (total 10 columns):
IDENTIFICATION      4038 non-null int64
BIRTHDAY             4038 non-null object
NAME                 4038 non-null object
DIPLOMTYPE           4038 non-null object
DIPLOMPERCENTAGE     4038 non-null float64
DIPLOMSECTION        4038 non-null object
DIPLOMOPTION         4038 non-null object
SCHOOL               4038 non-null object
SCHOOLPROVINCE       4038 non-null object
```

```
SCHOOLSTATUS      4038 non-null object
dtypes: float64(1), int64(1), object(8)
memory usage: 315.5+ KB
```

Nous remarquons que nous données en entré dispose des 10 colones avec variables quantitatives et qualitatives,

1. Variables continues

a. Attribue Date

Type: String

Pour nous faciliter la tache nous allons remplacer la date de naissance de chaque individu par son age a ce moment ci et ainsi obtenit un attribue continue de plus.

Valeurs Maquantes : Oui , ils sont causées par des erreurs à l'entré

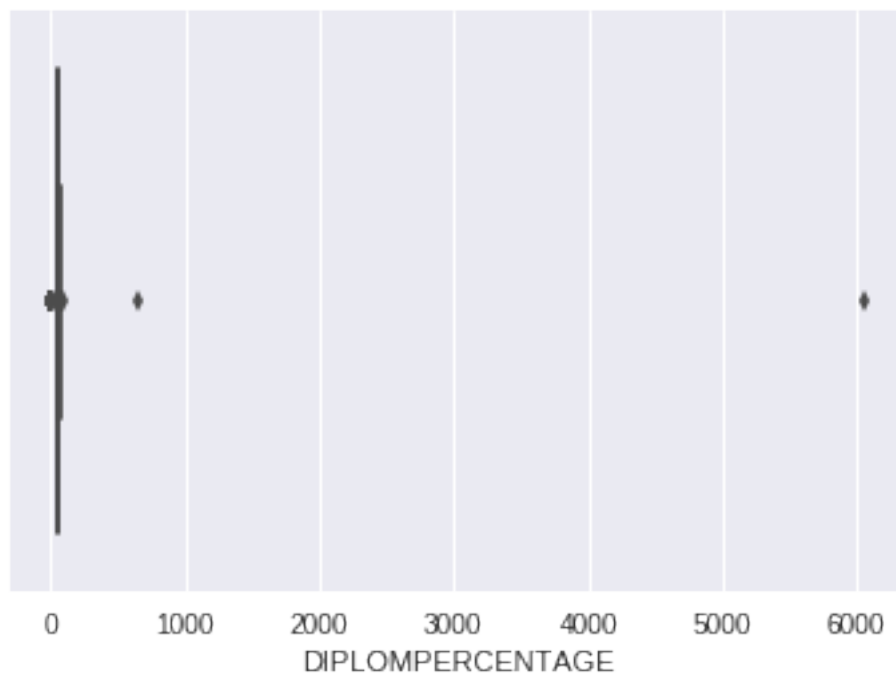
Out [24] : 7

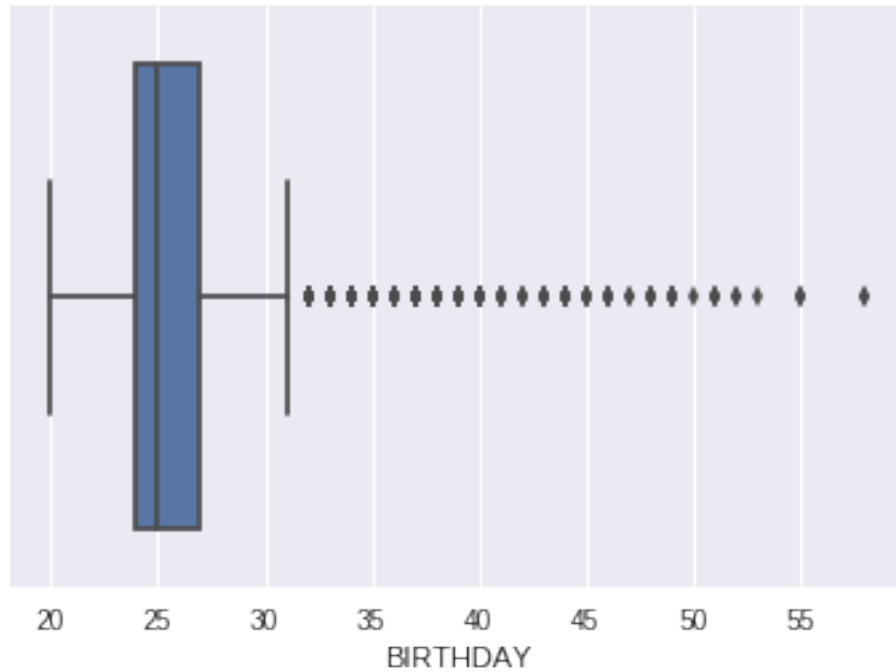
Out [73] : 4038

Out [76] : 0

Solution : Ils sont remplace par la moyenne comme leur proportion est insignifiant

Ce tableau decrit toutes les informations possibles sur les données continues et de prime à bord nous sommes à mesure de constater certaines incoherences sur les diplome percentage qui on un maximum de 6053 et un minimum de 0 qui est vraiment impossible car le diplome en RDC doi etre compris entre 50 et 100 % !Nous allons visualisé ces inchoherence de plus prêt avec des boxplots.

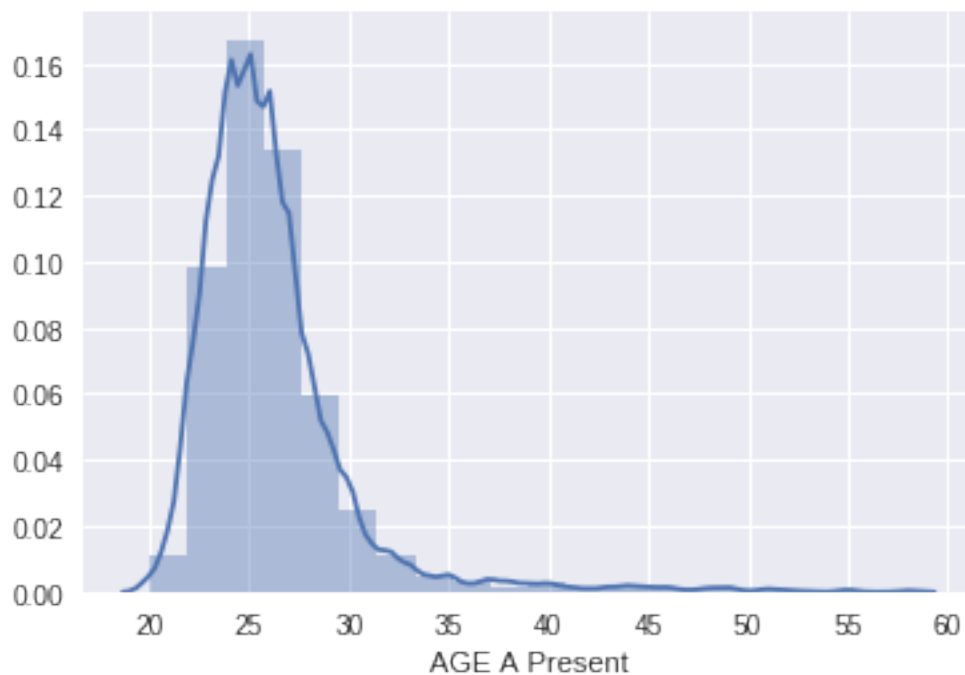




Au vu de ces courbes nous remarquons que l'attribue diplome percentage dispose de beaucoup des deviations.

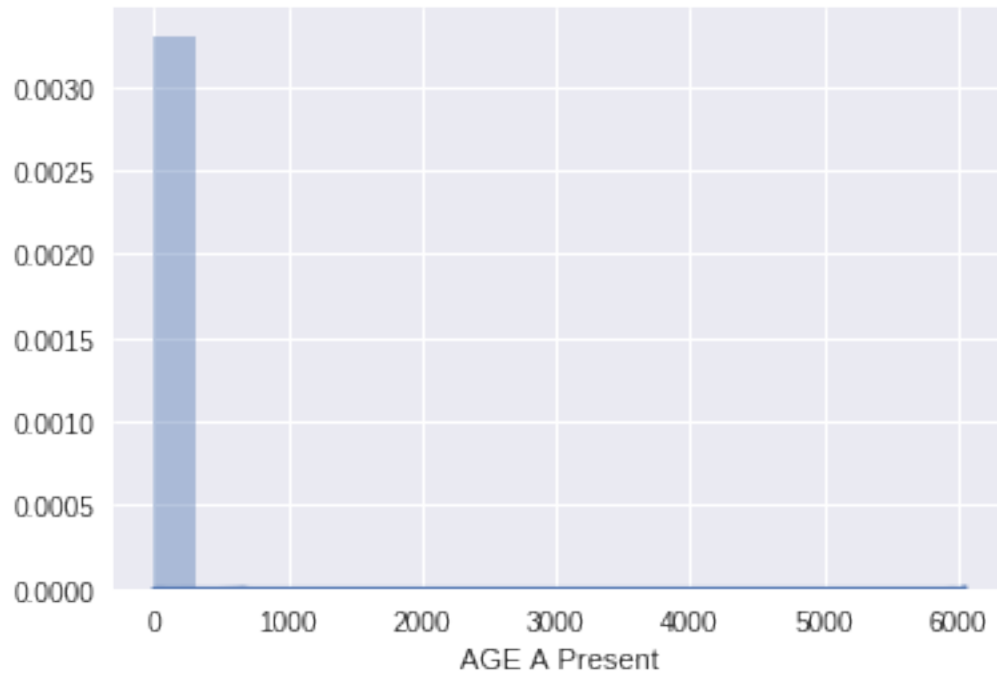
Mais l'attribue Bithday a une distribution presque normale

Out[36]: <matplotlib.axes.\_subplots.AxesSubplot at 0x7fb771ca6410>



Nous pouvons facilement voir que l'age a une distribution presque normale. NB: change norm hist to 1 to see count  
regardons de plus pret celui du diplome percentage

Out [128]: <matplotlib.axes.\_subplots.AxesSubplot at 0x7f1a8e854150>



A cause des dispersions nous ne pouvons pas bien visualis la distribution. essayons d'isoler les distribution pour voir de plus pret les données .

Out [98]: 34

Nous remarquons que nous avons 29 échantillons avec des valeur hors normes nous allons les normaliser dans la phase de préparation

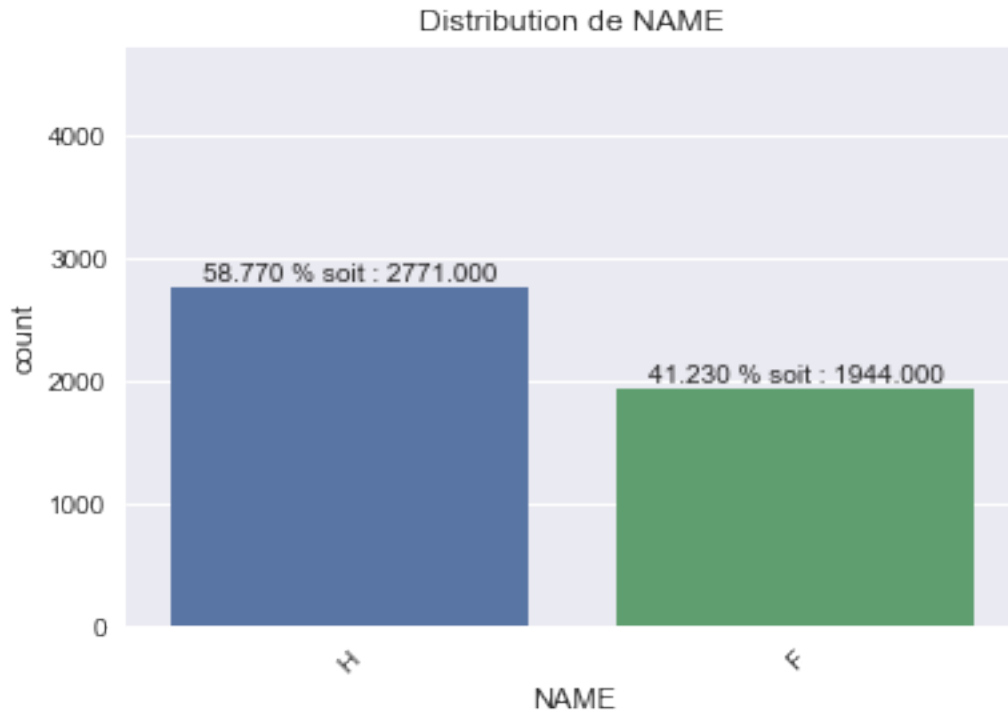
**Variables qualitatives ou categorielles** Pour chaque attribue nous allons faire de count plot voir les differents valeurs

Out [37]: ['BIRTHDAY',  
'NAME',  
'DIPLOMTYPE',  
'DIPLOMSECTION',  
'DIPLOMOPTION',  
'SCHOOL',  
'SCHOOLPROVINCE',  
'SCHOOLSTATUS']



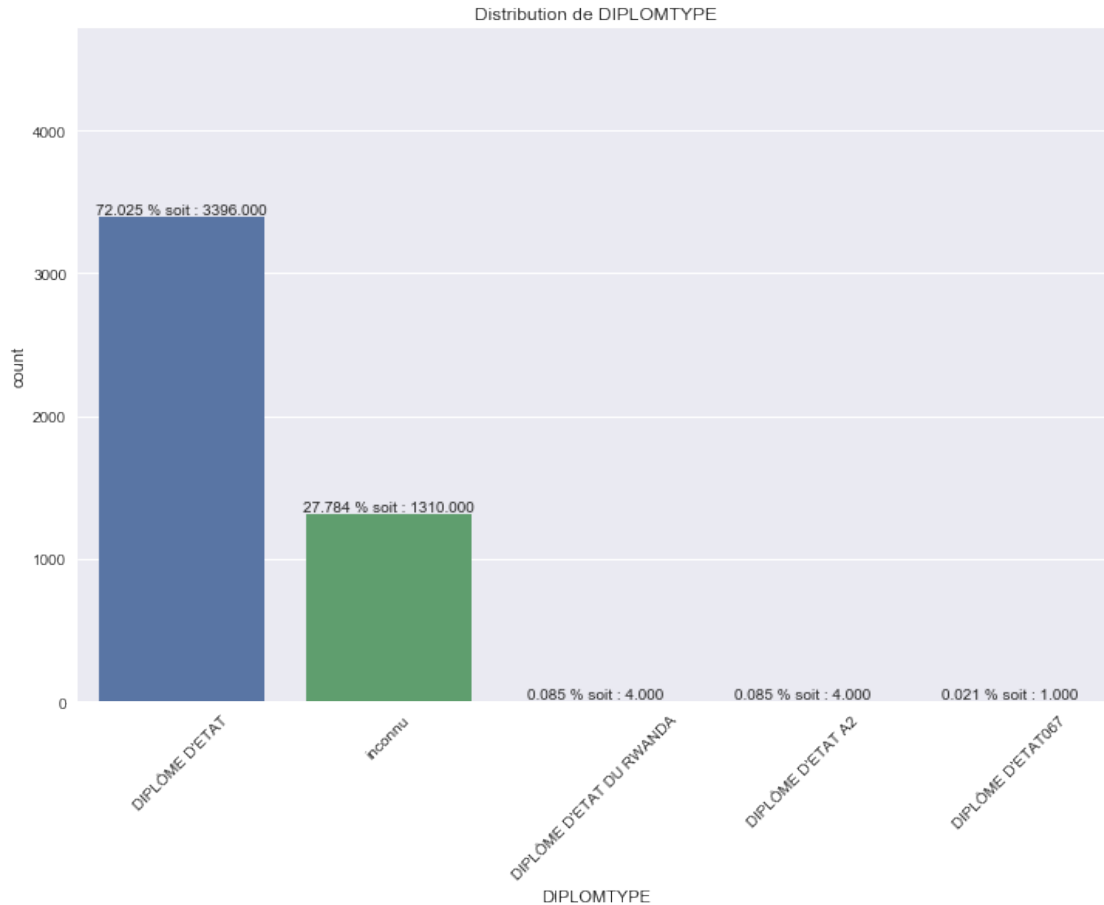
```
Out [39]: H      2771  
         F      1944  
         Name: NAME, dtype: int64
```

Nous pouvons remarqué facilement avec cete comande la repartition des sexes!



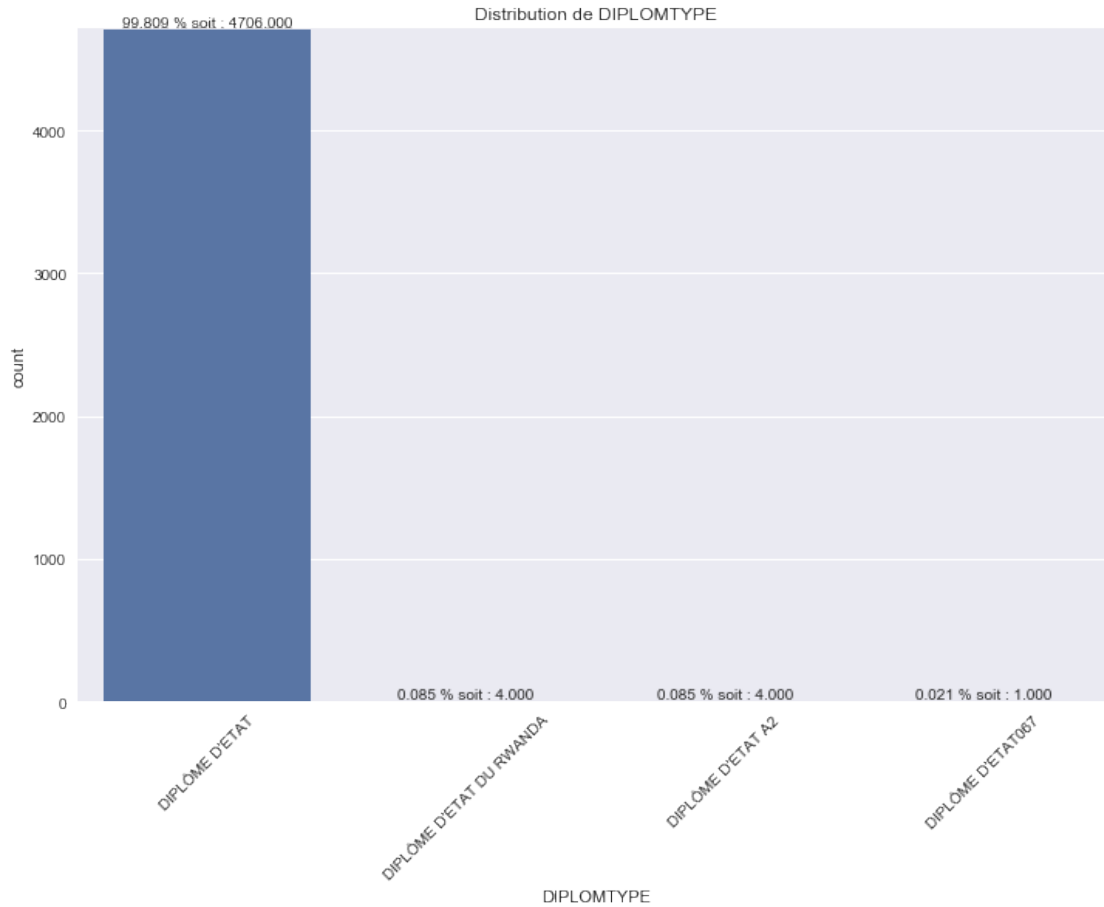
Voici la repartition de sexes dans notre ensemble d'apprentissage on peut aisement constater qu'il n'est pas si desequilibré que ça!, le genre est vraiment respecté avec 41% des nouveaux étudiant etant de sexe feminin.

Attribue : Diplome Type



Nous pouvons remarquer que notre dataset contient plus de 70% d'element avec le diplome d'etat et 26 avec un diplome type inconue nous allons traaiter cela à la suite et quelque echantillon avec des diplomes du Rwanda et d'autre avec des anciens diplomes, les individus dont le diplomé type sont inconu serons considéré comme diplome d'etat, donc nous pouvons le remplacer par le diplome d'etat.

```
Out[50]: DIPLOME D'ETAT          4706
          DIPLOME D'ETAT A2        4
          DIPLOME D'ETAT DU RWANDA  4
          DIPLOME D'ETAT067        1
          Name: DIPLOMTYPE, dtype: int64
```

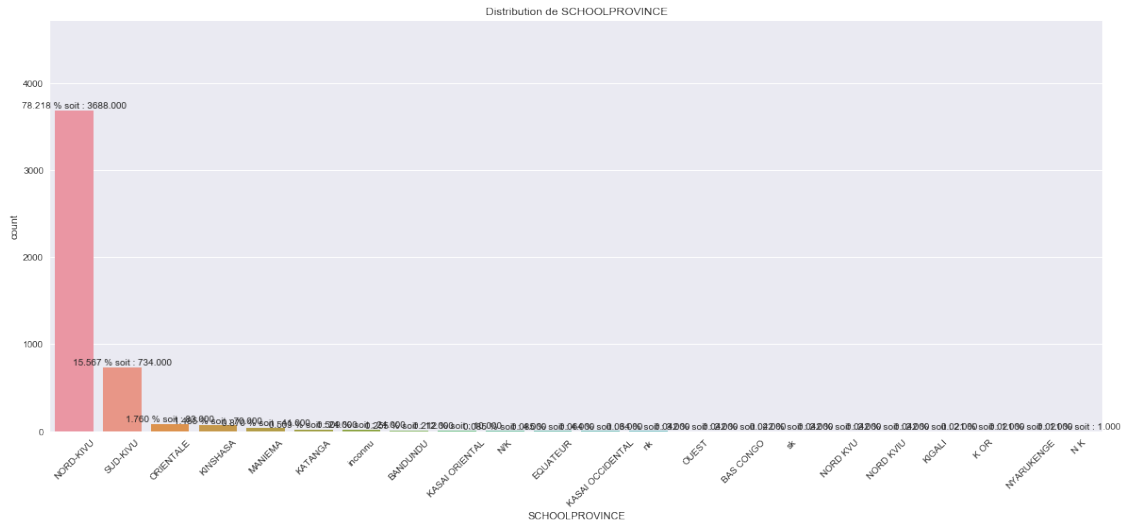


Attribue Diplome option et Diplome Section

Out [104]: 90

Nous avons plus de 90 sections  
une chose importante à remarquer au niveau des attribues 'DIPLOMOPTION' et 'DIPLOM-SECTION' sont tres desorganisées il faut bien les oragnées dans la phase de préparation des données

Attribue School Province



Avec cette image nous pouvons aisement constater que 63 % des étudiants de l'ulpgl proviennent de la province du nord kivu mais il ya une autre categorie qui provient du sud KIVu soit 15%

Attribue SCHOOL ET SCHOOL STATUS

Pour une bonne visualisation on peut prealablement transformer les colones en minuscule

Nous avons un attribue avec 1285 categories differentes ... Nous devons les analyser en details

```
Out[46]: inconnu          3833
         catholique       289
         protestant       260
         privé           170
         publique         152
         musulman         6
         kimbanguiste     3
         autodidacte      2
         Name: SCHOOLSTATUS, dtype: int64
```

ces attribues aussi necessite un nettoyage et une bonne reparation

Nous allons creer une branches à part pour le traitement de ces valeurs pour les diplomes section et diplome option

```
Out[47]: ['IDENTIFICATION',
         'BIRTHDAY',
         'NAME',
         'DIPLOMTYPE',
         'DIPLOMPERCENTAGE',
         'DIPLOMSECTION',
         'DIPLOMOPTION',
         'SCHOOL',
         'SCHOOLPROVINCE',
         'SCHOOLSTATUS']
```

les collones que nous allons traiter sont school ,schoolstatus, dilpomesection,diplomeoption

```
Out [78]:
```

	IDENTIFICATION	DIPLOMSECTION	DIPLOMOPTION \
0	45	PEDAGOGIEQUE	PEDA GENERALE
1	215	SCIENTIFIQUE	MATH-PHYSIQUE
2	343	SCIENTIFIQUE	MATH PHYSIQUE
3	356	ECONOMIE ET COMMERCE	ECONOMIE
4	429	TECHNIQUE	COMMERCIALE ET ADMINISTRATIVE

	SCHOOL	SCHOOLSTATUS
0	INSTITUT MAENDELEO	inconnu
1	INSTITUT VUNGI	inconnu
2	INSTITUT FARAJA	inconnu
3	ESISE/GISENYI	inconnu
4	C,S, UMOJA	inconnu

Nous allons sauvegarder cet nouveau dataframe dans un fichier csv et creer une notebook à part pour le traitement et le raffinement de ces collones

**Nettoyage des attribues bruitées** Nous avons pu remarquer dans la phase de precedante que certaines attribues ont des valeurs très desorganisées et vraiment dispersé et ce qui a une mauvaise influence sur le calcul de l'entropie et ainsi sur les algorithmes du Machine Learning . Nous pouvons aisement constater que ce problèmes est du à des fautes d'orthographes commise lors de la phase de saisie des données et ainsi pour countinues nous devons essayer de corriger ces erreurs et bien organisé les données . Voyons d'abord en chiffre comment cela se presente

Attribue Diplomes sections

```
Out [42]: 101
```

Avec cette commande nous remarquons que cette attribue dispose de 101 valeurs distinctes qui c'es qui es impossible pour la valeur de diplome section

voyons un peu en details ce qui contient cette attribue

Dans cette simple description nous remarquons que les valeurs on été mal saisi comme par exemple les valeurs suivantes : 'TECSC', 'Technique', 'TECHN IQUE', 'technique', 'TCH' qui sont saisie pour la meme et unique section 'techniques' mais avec differentes erreurs d'orthographe

cela n'est qu'un exemple des differentes valeurs mal orthographiées presente dans notre ensemble d'etude

ce genre d'erreur de notation à pour consequence le fait qu'il font augmenter l'entropie de nos colonnes et ainsi penalisent nos algorithmes surtout lorsqu'on travaille avec les arbres de décisions nous avons procédé à un nettoyage automatique qui a consisté en un groupement des vailleurs proches en utilisant la distance de leveinstein :(source : leveinstein) et le clustering par l'agorithme d'affinity propagation,et ainsi qu'un nettoyage manuelle pour arranger les données à la fin de cette phase nous avons obtenus des données moyennement propres et bien netoyer avec un entropie faibe. Nous pouvons le remarquer dans l'ensemble d'apprentissage suivant que ces données sont bien grouper.

```
Out [36]:
```

protestant	1370
catholique	1305

```

publique          726
inconnu           691
privé             536
autodidacte       44
musulman          38
kimbanguiste      5
Name: SCHOOLSTATUS, dtype: int64

```

```

Out[7]: protestant      1370
catholique      1305
publique        726
inconnu         691
privé           536
autodidacte     44
musulman        38
kimbanguiste    5
Name: SCHOOLSTATUS, dtype: int64

```

Nous pouvons maintenant combinner cette ensemble avec notre ensemble d'apprentissage de départ et ainsi continuer nore anlyse univarié poour les collones avce des varaibles qualitatives

```

Out[162]: protestant      1370
catholique      1305
publique        726
inconnu         691
privé           536
autodidacte     44
musulman        38
kimbanguiste    5
Name: SCHOOLSTATUS, dtype: int64

```

Nous allons enfin continuer avec notre anlyser univarié pour les attribues nouvellement net-toyer

```

Out[12]:  IDENTIFICATION  DIPLOMSECTION  DIPLOMOPTION  SCHOOL  SCHOOLSTATUS  \
0          3895.0          TECHN          ca i zanner  protestant

          SCHOOL_CORRECT  SCHOOL_RIGHT          OPTION_RIGHT
0          zanner          zanner  commerciale et adm

```

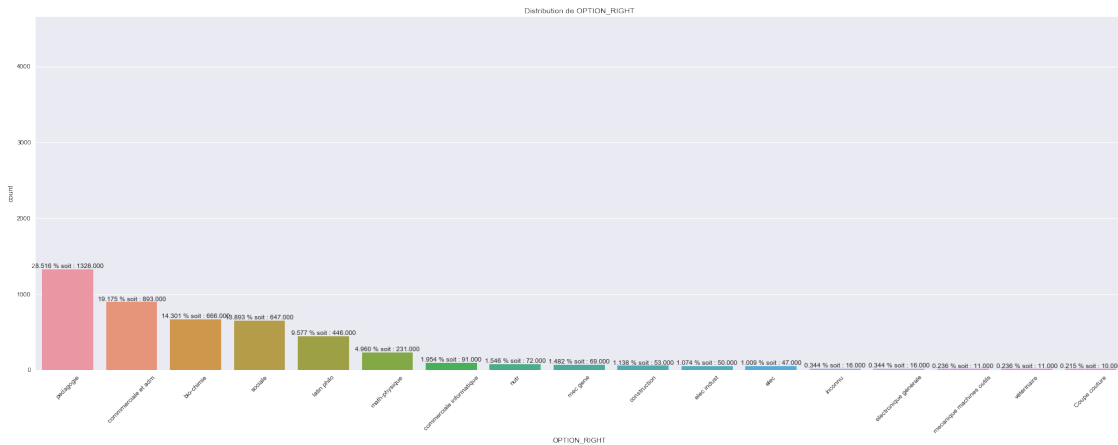
**Attribue SCHOOLSTATUS** Dans cette figure nous pouvons remarquer aisement que 29% des étudiants proviennent des écoles dites protestantes , 27% viennent des écoles catholiques , 11% des écoles privé , 15 des écoles publiques mais aussi il ya des étudiants venant des autodidactes , ceux provenant des écoles musulmanes et kibanguistes mais en proportion vraiment négligeable

**Attribue OPTIONRIGHT** Nous allons maintenant nous attaquer à l'attribue option du diplome qui contient les valeurs de l'option du diplome de 'étudaint'

Essayons de voir de plus pret combien des valeur distinctes il dispose

Out[32]: 33

Nous pouvons aisement remarquer que les étudiants de notre étude proviennent des 33 écoles différentes



Dans cette figure nous pouvons remarquer que la majeure partie des étudiants de notre études proviennent de la section pedagogique avec environ 28% ensuite vient la section commerciale et administrative avec 19% , suivent sociale avec 13%, scientifique bio-chimie avec 14 % ensuite viennent autres differentes options avec des valeurs inferieures à 5%

**Attribut School** cette attribue comprend les valeurs de l'école de provenance des nos finaliste combiné avec l'attribue school status il joue un role important dans notre étude.  
voyons d'abord combien des valeurs differentes il comprend:

Out [13]: 594

nous pouvons aisement remarquer que les eleves proviennent de 594 écoles differentes dans l'image qui va suivre nous allons visulaise les écoles les plus représentées

Nous pouvons remarquer aisement que le top 10 des école de provenance est constituer de grandes écoles de la ville de Goma avec l'insititut metanoia et le college mwanga en tete de liste avec l'institut mwanga et metanoia en tete de liste avec 15% chacun ensuite vienne l'institut bakanja avec 6% ensuite vienne maendelo, le lycée sainte ursule et l'institut de Goma avec 6%, 5% et 5 % respectivement et d'autres éecole se partagent le reste de 50%.

Pour finaliser l'analyse univariée de nos données en entrée nous allons jeter un coup d'oeil sur la colonne Fac qui contient la faculté choisie par l'étudiant.

```
Out[8]:
```

Faculté des Sciences Économiques et de Gestion	1549
Faculté des Sciences et Technologies Appliquées	903
Faculté de Droit	896
Faculté de Santé et Développement Communautaires	758
Faculté de Médecine	242
Faculté de Psychologie et des Sciences de l'Éducation	227
Faculté de Théologie	140

```
Name: FAC, dtype: int64
```

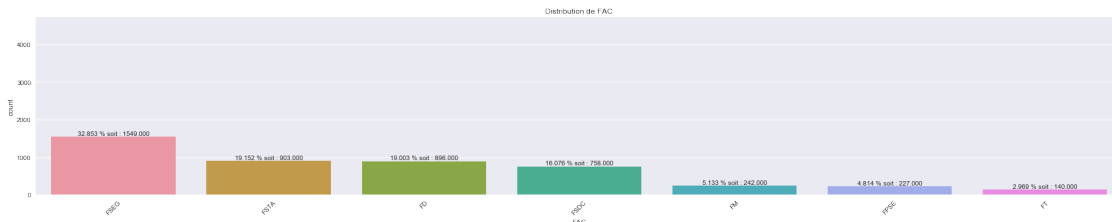
Nous allons remplacer les non des facultés par leurs abreviations respectifs

```
/Users/espyMur/Desktop/Memory-WorkingDir/memoryVenv/lib/python2.7/site-packages/pandas/core/indexing.py:111:
A value is trying to be set on a copy of a slice from a DataFrame
```

See the caveats in the documentation: <http://pandas.pydata.org/pandas-docs/stable/indexing.html#>

```
self._setitem_with_indexer(indexer, value)
```

```
Out[10]: FSEG      1549
         FSTA      903
         FD        896
         FSDC      758
         FM        242
         FPSE      227
         FT        140
         Name: FAC, dtype: int64
```



Dans ce tableau nous remarquons la distribution des valeurs pour l'attribue faculté des étudiants: - FSEG: 32,8% - FSTA : 19,15% - FD :19% - FSDC :16% - FM : 5% - FPSE : 4% - FT :3%

Nous pouvons maintenant passer à l'analyse bi-varié

**Statistique Bivariée** Dans cette partie nous allons nous allons effectué une analyse bivariée entre les attribues en entrée et les attribues en sortie , pour une premiere approche nous allons faire une analyse entre la faculté choisie et les differentes variables d'entres de notre ensemble d'apprentissage dans la seconde approche nous essayerons de le faire la meme chose pour le autres variables de sortie. voici le differentes combinaisons que nous allons effectuer: 1. FAC-Diplome Province : Pour voir la relation entre la faculté et la province d'origine de l'etudiant 2. FAC-DIPPOURCENTAGE: Pour voir la relation entre la faculté et la pourcentage obtenu à l'exetat 3. FAC-AGE: Pour voir la relation entre l'age de l'etudiant et la FAC 4. FAC-DIPLOMEOPTION: Pour voir la relation entre la faculté et l'option du diplome 5. FAC-SEXE: Pour la relation avec le sexe des étudiants 6. FAC-SCHOOL: pour la relation entre l'ecole de provenance la fac 7. FAC-SCHOOLSTATUS : pour la relation entre le status de l'ecole de la FAC

FACULTE DIPLOME PROVINCE

1. Definition et Nature de la relation: Nous allons essayer de decouvrir la relation existante entre la province de l'ecole et le choix de ma section
2. Pour determiner le type nous allons utiliser le test de chi-carré qui nous permettra de trouver la force de la cette relation



Notre hypothèse est la suivante est la suivante : il n'y a pas de relation entre la province du diplôme et la section choisie

Nous nous sommes servis d'une classe spécialisée qui contient les méthodes qui nous permettent de conduire notre test statistique

Prémièrement nous allons chercher la table de contingence qui nous montre comment nos valeurs se répartissent en fonction des sections

Out [128]: SCHOOLPROVINCE

BANDUNDU	2.298197
BAS CONGO	0.383033
EQUATEUR	0.766066
K OR	0.191516
KASAI OCCIDENTAL	0.574549
KASAI ORIENTAL	1.915164
KATANGA	4.596394
KIGALI	0.383033
KINSHASA	13.406151
MANIEMA	7.852174
NORD-KIVU	708.610817
NYARUKENGE	0.191516
ORIENTALE	15.895864
OUEST	0.383033
SUD-KIVU	140.956098
inconnu	4.596394

Name: FSTA, dtype: float64

Dans ces 2 tableaux nous avons les différentes tables l'une c'est pour les valeurs observées et l'autre c'est pour les valeurs attendues ainsi avec ces valeurs nous pouvons trouver la valeur du test statistique chi-carrée ensuite la comparer avec notre valeur critique et ensuite vérifier si nous pouvons confirmer ou infirmer notre hypothèse nulle

Out [124]: ('It indicates that the relationship between the variables is significant at confidence level 0.95')

Out [125]: 113.1452701425554

Notre valeur critique est de 113,14

et notre valeur du test statistique est donnée par :

lorsque nous comparons ces 2 valeurs nous remarquons que notre valeur de  $\chi^2$  est supérieure à notre valeur critique et tombe dans la région de rejet de notre hypothèse nulle et ainsi celle-ci est rejetée et ainsi nous concluons avec 95% de certitude que notre hypothèse alternative est vraie: il y a une relation entre la province et la faculté en d'autres termes ayant un nombre d'étudiants venant d'une province nous pouvons prédire la faculté choisie.

Mais cette relation est à prendre avec réserve car les données sont inégalement réparties entre les provinces!

Nous pouvons remarquer les résultats de nos analyses dans cette figure.

FACULTE DIPLOMEOPTION

nous allons refaire la même analyse que pour le point précédent!!

```

Out[147]: OPTION_RIGHT  agrecole  agronomie  batiment  bio-chimie  \
FAC
FD                1          0          0          28
FM                1          1          0         111
FPSE              0          0          0          4
FSDC              1          0          0         117
FSEG              1          0          1         156
FSTA              5          2          2         244
FT                0          1          0          6

```

```

OPTION_RIGHT  commerciale informatique  commerciale et adm  construction  \
FAC
FD                12                62          0
FM                4                 5          0
FPSE              0                 2          0
FSDC              1                31          0
FSEG              63               724          3
FSTA              11               59          48
FT                0                10          2

```

```

OPTION_RIGHT  coupe couture  diet  economie  ...  machine outil  \
FAC
FD                3          0          1  ...          0
FM                0          0          0  ...          0
FPSE              2          0          0  ...          0
FSDC              10         2          0  ...          0
FSEG              4          0          1  ...          0
FSTA              0          0          0  ...          7
FT                0          0          0  ...          0

```

```

OPTION_RIGHT  math-physique  mec gene  mecanique machines outils  nutr  \
FAC
FD                10          3                0          1
FM                13          0                0          14
FPSE              2          0                0          0
FSDC              11          0                0          47
FSEG              42          2                0          6
FSTA              149         64               11          1
FT                4          0                0          3

```

```

OPTION_RIGHT  pedagogie  relations publiques  secretariat  sociale  \
FAC
FD                327                0          0         169
FM                48                0          1          27
FPSE              189                0          0          16
FSDC              290                0          0         187
FSEG              270                1          5         184
FSTA              114                0          2          51

```

FT	90	0	0	13
----	----	---	---	----

OPTION_RIGHT	vétérinaire
FAC	
FD	0
FM	4
FPSE	0
FSDC	5
FSEG	1
FSTA	0
FT	1

[7 rows x 30 columns]

Out[148]:

OPTION_RIGHT	agrecole	agronomie	batiment	bio-chimie	\
FAC					
FD	1.710286	0.760127	0.570095	126.561188	
FM	0.461930	0.205302	0.153977	34.182821	
FPSE	0.433298	0.192577	0.144433	32.064051	
FSDC	1.446872	0.643054	0.482291	107.068505	
FSEG	2.956734	1.314104	0.985578	218.798303	
FSTA	1.723648	0.766066	0.574549	127.549947	
FT	0.267232	0.118770	0.089077	19.775186	

OPTION_RIGHT	commerciale	informatique	commerciale et adm	construction	\
FAC					
FD		17.292895		169.698409	10.071686
FM		4.670626		45.833722	2.720255
FPSE		4.381124		42.992789	2.551644
FSDC		14.629480		143.561824	8.520467
FSEG		29.895864		293.373701	17.411877
FSTA		17.427996		171.024178	10.150371
FT		2.702015		26.515376	1.573701

OPTION_RIGHT	coupe couture	diet	economie	...	machine outill	\
FAC				...		
FD	3.610604	0.380064	0.380064	...	1.330223	
FM	0.975186	0.102651	0.102651	...	0.359279	
FPSE	0.914740	0.096288	0.096288	...	0.337010	
FSDC	3.054507	0.321527	0.321527	...	1.125345	
FSEG	6.241994	0.657052	0.657052	...	2.299682	
FSTA	3.638812	0.383033	0.383033	...	1.340615	
FT	0.564157	0.059385	0.059385	...	0.207847	

OPTION_RIGHT	math-physique	mec gene	mecanique	machines outills	nutr	\
FAC						
FD	43.897349	13.112195		2.090350	13.682291	
FM	11.856204	3.541463		0.564581	3.695440	

FPSE	11.121315	3.321951	0.529586	3.466384
FSDC	37.136373	11.092683	1.768399	11.574973
FSEG	75.889502	22.668293	3.613786	23.653871
FSTA	44.240297	13.214634	2.106681	13.789183
FT	6.858961	2.048780	0.326617	2.137858

OPTION_RIGHT	pedagogie	relations publiques	secretariat	sociale \
FAC				
FD	252.362248	0.190032	1.520255	122.950583
FM	68.160339	0.051326	0.410604	33.207635
FPSE	63.935525	0.048144	0.385154	31.149311
FSDC	213.493955	0.160764	1.286108	104.013998
FSEG	436.282503	0.328526	2.628208	212.556310
FSTA	254.333828	0.191516	1.532131	123.911135
FT	39.431601	0.029692	0.237540	19.211029

OPTION_RIGHT	vétérinaire
FAC	
FD	2.090350
FM	0.564581
FPSE	0.529586
FSDC	1.768399
FSEG	3.613786
FSTA	2.106681
FT	0.326617

[7 rows x 30 columns]

Out[151]: 4155.0246292496686

Out[152]: ' It indicates that both categorical variable are dependent '

Out[153]: 205.77862677980613

sur base de ce fait on fait la meme conclusion que pour la province : il ya une relation de dependance entre la section du diplome et la faculté ce qui est loique car les étudiant en genral se base sur leur option pour chosir leur option

##### FACULTE SCHOOLSTATUS

nous allons refaire la meme analyse que pour le point precedant!!

Out[223]: 304.71578214658456

Out[224]: ('It indicates that the relationship between the variables is significant at confidence 0.99)

Out[225]: 66.206236283993221

sur base de ce fait on fait la meme conclusion que pour la province : il ya une relation de dependance entre la section du diplome et la faculté ce qui est loique car les étudiant en genral se base sur leur option pour chosir leur option

FACULTE SCHOOL

```
Out[202]: FAC
          FD      29
          FM       2
          FPSE     0
          FSDC     2
          FSEG    42
          FSTA     3
          FT       0
          Name: zanner, dtype: int64
```

```
Out[203]: FAC
          FD    14.822481
          FM     4.003393
          FPSE   3.755249
          FSDC  12.539555
          FSEG  25.625027
          FSTA  14.938282
          FT     2.316013
          Name: zanner, dtype: float64
```

```
Out[197]: 6389.4044131161027
```

```
Out[198]: ('It indicates that the relationship between the variables is significant at confidence
0.95)
```

```
Out[199]: 3697.8815995010605
```

reste à verifier ....  
FACULTE SEXE

```
Out[327]:
          M      F
GENDER   F      H  FAC
FD       356  540  896
FM       109  133  242
FPSE     119  108  227
FSDC     469  289  758
FSEG     726  823 1549
FSTA     140  763  903
FT        25  115  140
```

```
Out[264]: 896
```

```
Out[257]: 4715.0
```

```
Out[171]: 445.85875262024177
```

```
Out[175]: ('It indicates that the relationship between the variables is significant at confidence
0.95)
```

```
Out[173]: 12.591587243743977
```

sur base de ce fait on fait la meme conclusion que pour la province : il ya une relation de dependance entre le sexe et la faculté ce qui est logique car par exemple en faculté de technologie et teologie on trouve moins des hommes que des femmes

FACULTE - DIPLOME AGE et FACULTE DIPLOME POURCENTAGE

comme ces une des colonnes dispose des variables continues nous allons utiliser le test ANOVA

ce test nous permettra de savoir si la moyenne de l'age des étudiants est la meme pour chaque faculté: cela constituera notre hypothèse nulle ,on va chercher la probabilité p est on décidera sur base de cette valeur ! si elle est inférieure à 0.05 on rejettera l'hypothèse nulle

```
Out [314]:
```

	df	sum_sq	mean_sq	F	PR(>F)
C(FAC)	6.0	14857.002658	2476.167110	135.823792	3.721673e-159
Residual	4708.0	85830.284723	18.230732	NaN	NaN

comme la valeur de PR est inférieure à 0.05 on peut conclure que la moyenne de l'age n'est pas la meme au sein de chaque faculté

```
Out [320]:
```

	IDENTIFICATION	DIPLOMPERCENTAGE	AGE
FAC			
FD	8837.433036	56.151372	24.771205
FM	10887.719008	59.434420	21.487603
FPSE	8417.453744	56.202099	28.224670
FSDC	8195.007916	55.329211	25.974934
FSEG	8410.416398	56.832564	24.323434
FSTA	9166.437431	58.941594	23.307863
FT	8125.250000	53.814286	31.428571

Pour prouver le rejet de notre hypothèse nulle on peut remarquer que les facultés de Medecine et celui de technologie ont une moyenne d'age de 21 et 23 respectivement et les facultés de psychologie et celui de theologie ont une moyenne d'age respective de 28 et 31 ans

**Diplome pourcentage** voici comment se presente le test:

```
Out [325]:
```

	df	sum_sq	mean_sq	F	PR(>F)
C(FAC)	6.0	9139.202728	1523.200455	48.757704	2.256165e-58
Residual	4708.0	147078.865397	31.240201	NaN	NaN

Egalement ici on rejette aussi notre hypothèse nulle qui stipulait que la moyenne du pourcentage du diplome est la meme au sein de chaque faculté . Pour prouver le rejet de notre hypothèse nulle on peut remarquer que les facultés de Medecine et celui de technologie ont une moyenne de pourcentage de 59% chacun et celui de theologie à une moyenne de 53%

```
Out [324]: 3.3032675567110044
```

```
File "<ipython-input-1-b7f99248bc48>", line 1
ipython nbconvert --to pdf --template hidecode DataExplorationDraft1.ipynb
^
```

```
SyntaxError: invalid syntax
```