



Groupe 8

BENCHAAABANE Mahmoud
BOUAIDA Mouad
EL YOUNSI Abdellah
HARIM Fatim Zahra
HARTMANN Guillaume
MORSLI Omar
TAKHCHI Mehdi

3°A FIG ISA Brest

Etudiants en troisième année à l'IMT Atlantique – campus de Brest

Sous la tutelle de Romain Billot, Gilles Coppin, Bernard Gourvennec, enseignants-chercheurs à l'IMT Atlantique – campus de Brest

RAPPORT UV 101E : PROJET STATISTIQUE

DATA SCIENCE : PERCEPTION ET UTILISATION PAR LES ETUDIANTS DE L'IMT ATLANTIQUE

Rapport

Version 1

Le 17/11/2018



IMT Atlantique

Bretagne-Pays de la Loire
École Mines-Télécom

Table des matières

1) Introduction	1
1.1) Contexte	1
1.2) But du projet	1
2) Echantillonnage et conception du questionnaire.....	1
2.1) Identification des informations nécessaires	1
2.2) Conception du questionnaire.....	2
2.3) Nettoyage des données	3
3) Représentativité	4
3.1) Représentativité par sexes	4
3.2) Représentativité par origines.....	4
3.3) Représentativité des étudiants par filière.....	4
4) Tests statistiques.....	5
4.1) Tests sur l'intérêt qu'ont les étudiants pour les Data Sciences.....	5
4.2) Tests du Khi-2.....	10
4.3) Tests sur les projets/MOOC en Data Science	16
5) Discussion critique de l'étude	18
6) Conclusion	19
7) Bibliographie	20
8) Annexes	21

1) INTRODUCTION

1.1) Contexte

Les data sciences sont un thème que l'on voit de plus en plus apparaître à l'heure d'aujourd'hui. Cependant, elles restent un thème qui est apparu récemment et qui n'est pas forcément bien connu de tous de manière générale. Les data sciences désignent la science des données qui permet à une entreprise d'analyser une certaine quantité de données et d'en extraire des informations utiles permettant de prendre des décisions ou de résoudre des problèmes pour l'entreprise.

Les data sciences étant une discipline qui peut être choisie à l'école, elles sont un thème qui est à priori plus connu par les étudiants de l'IMT Atlantique que par les étudiants de manière générale. Nous nous sommes alors intéressés à la perception mais aussi à l'utilisation que pouvaient en faire les étudiants de l'école.

1.2) But du projet

La finalité de ce sondage est donc d'avoir une analyse globale sur la vision qu'ont les étudiants de l'école sur les data sciences selon qu'ils soient ou souhaitent entrer dans une filière les utilisant ou qu'ils soient dans une autre filière ne les utilisant pas.

2) ECHANTILLONNAGE ET CONCEPTION DU QUESTIONNAIRE

Pour réaliser le questionnaire, nous sommes partis d'idées que se faisaient les membres de notre groupe sur les data sciences et leur perception et utilisation par les étudiants. Nous avons alors choisi de retenir certaines hypothèses qui nous paraissaient les plus pertinentes. Nous avons présenté nos hypothèses à nos encadrants en amont pour qu'ils puissent les valider.

2.1) Identification des informations nécessaires

Dans un premier temps, il a été nécessaire d'identifier les informations dont nous allions avoir besoin pour vérifier les hypothèses que nous avons formulées à partir de notre problématique.

Il était alors nécessaire de connaître la population que nous allions interrogée. Nous avons choisi d'interroger uniquement les étudiants de l'IMT Atlantique (campus de Brest, Rennes et Nantes) car c'était la population étudiante la plus facile à toucher et aussi une population qui devait connaître le thème des data sciences.

Ensuite, il fallait connaître des informations simples sur cette population telle que la proportion de personnes selon le sexe, leur nationalité (française ou étrangère) ainsi que leur niveau actuel de formation à l'école (1^{ère}, 2^{ème}, 3^{ème} année ou master).

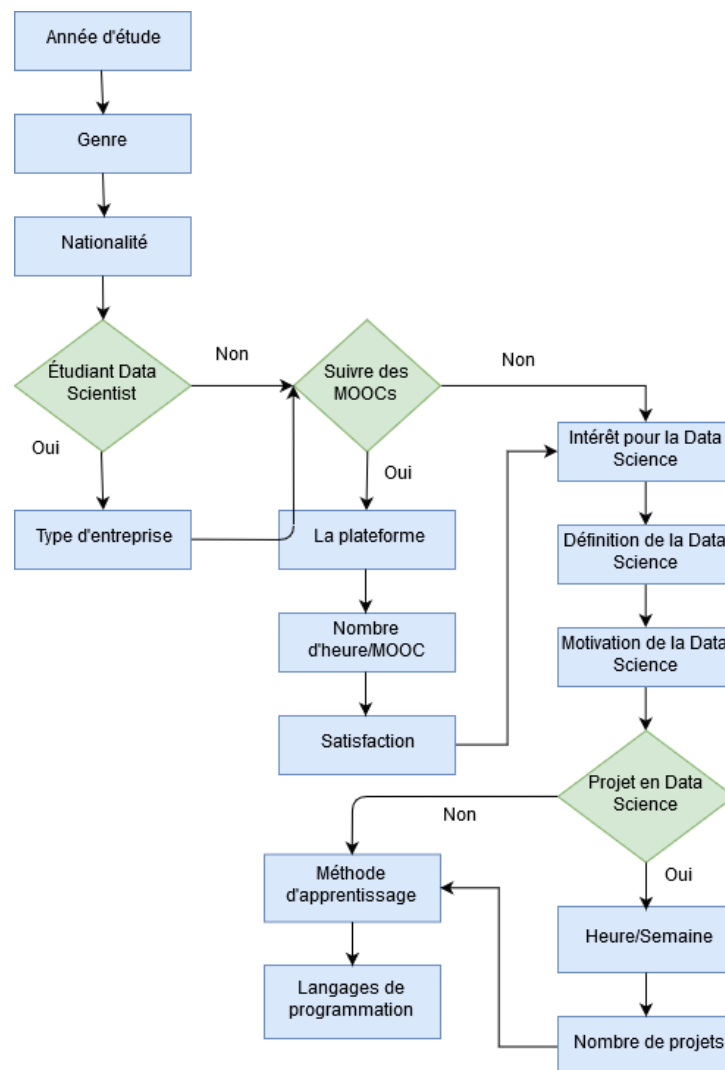
Enfin, nous avons commencé à étudier les informations nécessaires pour répondre aux différentes hypothèses formulées.

2.2) Conception du questionnaire

Nous nous sommes inspirés de la présentation de M. Gourvennec sur les enquêtes ce qui nous a permis de prendre les décisions suivantes :

- ✚ Pour garder l'attention des personnes sondées, nous nous sommes limités à 19 questions.
- ✚ Pour obtenir des réponses précises, nous avons préférés des questions semi-ouvertes.
- ✚ Pour que les questions soient cohérentes nous avons utilisé un arbre de décisions
- ✚ Pour respecter l'aspect multiculturel de l'école, nous avons utilisé des expressions adéquates et universelles.

Ci-dessous, la structure du questionnaire :



Pour parvenir aux sondés, Nous avons opté pour un questionnaire en ligne. Nous avons préféré d'utiliser l'outil de l'école, LimeSurvey, plutôt que Google Form pour des raisons de sécurité de données. Nous étions sûrs de la manière dont allaient être sauvegardées nos données.

Nous avons utilisé une méthode non aléatoire. Plus précisément **le mode d'échantillonnage est « volontaire »**. Théoriquement moins fiable, ce mode nous permet de nous passer de base de sondage et de proposer un questionnaire à moindre coût.

Après l'avoir envoyé par mail aux élèves des campus de Brest, Rennes et Nantes, nous avons obtenu 221 réponses brutes qu'il nous a fallu traiter afin d'en dégager des premiers résultats que nous avons analysés grâce à R.

2.3) Nettoyage des données

a) Mise en forme des données

Les données récupérées sur LimeSurvey ont nécessité un traitement afin de pouvoir être analysées. En effet chaque réponse (ligne) est associée à 58 colonnes différentes. Chaque colonne correspond à un des choix de réponse donc il a fallu les ressembler pour pouvoir les traiter. Pour mieux comprendre, voici un exemple :

Femme	Homme	Autre		Genre
x			→	Femme
	x			Homme

Même si en fonction de l'arbre de décisions, certaines questions n'ont pas été posées, nous avons des colonnes avec des NA qui correspondent à une personne pour qui la question n'a pas été posée du fait de ces réponses précédentes.

Nous avons également enlevé les réponses des personnes qui ne sont pas allées jusqu'au bout du sondage.

b) Données aberrantes

Dans un second temps, nous avons identifié des réponses qui faussaient l'étude telle qu'une personne qui a passé 999999999 heures, ce qui représente 3805 années, sur un MOOC. Il s'agit là d'éliminer les réponses aberrantes en elles-mêmes.

c) Résultats

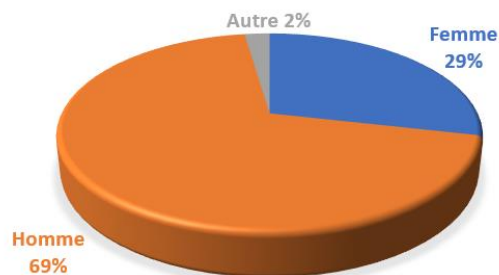
Une fois ce traitement des données brutes effectué nous avons pu dégager les premiers résultats de notre étude. Notre échantillon est composé de 173 personnes qui ont répondu entièrement et correctement au sondage.

3) REPRESENTATIVITE

3.1) Représentativité par sexes

A partir des informations fournies par l'Etudiant¹, le pourcentage de femme parmi les élèves de l'IMT Atlantique est de 23,5% théoriquement.

Parmi les personnes ayant répondu à notre questionnaire, on trouve les chiffres suivants :



Sexe	Pourcentage	Représentativité
Femme	28,57%	121,57%
Homme	69,14%	92,18%
Autre	2,29%	X

A noter que la catégorie « Autre » n'est pas présente dans les informations recueillies sur l'Etudiant.

On peut constater que la population des femmes est surreprésentée par rapport à celle des hommes puisqu'elle a un pourcentage de représentativité de 121,57%.

3.2) Représentativité par origines

De la même manière, à partir des informations fournies par Campus France², le pourcentage de personnes de nationalité non française parmi les élèves de l'IMT Atlantique est de 40% en septembre 2018 théoriquement.

Parmi les personnes ayant répondu à notre questionnaire, on trouve les proportions suivantes :



Nationalité	Pourcentage	Représentativité
Française	71.68%	119,47%
Etrangère	28.32%	70.8%

L'échantillon n'est pas très représentatif du point de vue des origines puisque le pourcentage d'étudiants étrangers ayant répondu à notre questionnaire est bien inférieur au pourcentage réel et a une représentativité de seulement 70,8%.

3.3) Représentativité des étudiants par filière

¹ D'après l'Etudiant, pourcentage d'étudiante à la rentrée 2017, accessible à l'adresse suivante : <https://www.letudiant.fr/palmares/palmares-des-ecoles-d-ingenieurs/imt-atlantique-mines-nantes.html>

² D'après Campus France, pourcentage d'étudiants étrangers en septembre 2018, accessible à l'adresse suivante :

https://ressources.campusfrance.org/guides_etab/etablissements/fr/ing_mines_nantes_fr.pdf

Nous avons aussi posé une question préliminaire pour connaître le pourcentage de personnes ayant répondu dans chaque filière :

Filière	Pourcentage
1 ^{ère} année	24,86%
2 ^{ème} année	23,70%
3 ^{ème} année	40,46%
Master of Science	4,62%
Master spécialisé	5,20%
Master	1,16%

On remarque alors que le pourcentage de 3^{ème} année est bien plus important que celui de 1^{ère} et 2^{ème} année alors qu'il y a à-peu-près le même nombre d'étudiants dans chaque année. Les étudiants de 3^{ème} année sont ceux qui se sont sentis le plus concernés, ayant eux aussi des sondages à réaliser, mais c'est aussi ceux qui sont le plus susceptibles de connaître le sujet étudié.

4) TESTS STATISTIQUES

4.1) Tests sur l'intérêt qu'ont les étudiants pour les Data Sciences

- **Hypothèse 1 :** Le taux de français intéressés par le domaine de la data science est inférieur à celui des étrangers.

	Moyenne	Ecart-type
Français	3.73	0.87
Etrangers	4.0	0.88

Test de normalité sur les deux échantillons :

Français

```
shapiro.test(fr_int)
```

```
# Shapiro-Wilk normality test
# data: fr_int
# W = 0.86054, p-value = 1.928e-09
```

Donc avec un risque de 10%, on rejette l'hypothèse nulle de normalité pour l'échantillon des français.

Etrangers

```
shapiro.test(etr_int)
```

```
# Shapiro-Wilk normality test
# data: etr_int
# W = 0.72681, p-value = 3.247e-08
```

Donc aussi avec un risque de 10%, on rejette l'hypothèse nulle de normalité pour l'échantillon des étrangers.

Les deux échantillons ne suivent pas des lois normales. Ainsi, on va utiliser le test non paramétrique de Wilcoxon.

Test de moyenne

```
wilcox.test(fr_int,etr_int,alternative = 'g')
```

```
# Wilcoxon rank sum test with continuity correction
# data: fr_int and etr_int
# W = 2427, p-value = 0.9878
```

Conclusion : Avec un risque de 10%, on accepte H_0 : Les étrangers sont plus intéressés par la data science que les français.

- **Hypothèse 2 :** Le taux d'étudiants hommes intéressés par la Data Science est égal à celui étudiantes femmes intéressées par la Data Science.

	Moyenne	Ecart-type
Etudiants hommes	3.84	0.85
Etudiants Femmes	3.72	0.96

Test de normalité sur les deux échantillons :

Homme :

```
shapiro.test(homme_int)
```

```
Shapiro-Wilk normality test
data: homme_int
W = 0.80936, p-value = 3.144e-11
```

Conclusion : Avec un risque de 10%, on rejette l'hypothèse nulle de normalité pour l'échantillon des hommes.

Femme :

```
shapiro.test(femme_int)
```

```
Shapiro-Wilk normality test
data: femme_int
W = 0.88387, p-value = 0.0001469
```

Donc avec un risque de 10%, on rejette l'hypothèse nulle de normalité pour l'échantillon des femmes.

Les deux échantillons ne suivent pas des lois normales. Ainsi, on va utiliser le test non paramétrique de Wilcoxon.

```
wilcox.test(femme_int, homme_int)
```

Wilcoxon rank sum test with continuity correction

data: femme_int and homme_int

W = 2789.5, p-value = 0.3846

alternative hypothesis: true location shift is not equal to 0

Avec un risque de 10%, on accepte H_0 . L'intérêt des hommes et des femmes pour la data science est le même.

Analyse de l'intérêt par la méthode ANOVA

```
> summary(AnovaModel.1)
              Df Sum Sq Mean Sq F value Pr(>F)
Year           5    3.69   0.7390   0.885  0.492
Residuals    167  139.43   0.8349

> with(anova, numSummary(Interest, groups=Year, statistics=c("mean", "sd")))
              mean          sd data:n
1ère année    3.046512 0.8716019     43
2ème année    3.341463 0.8546858     41
3ème année    3.400000 0.9231327     70
Master        3.000000 0.0000000      2
Master of Science 3.250000 1.2817399      8
Master spécialisé 3.333333 1.0000000      9
```

Interprétation des résultats du tests ANOVA

La p-value est supérieure au seuil de risque 10%, nous pouvons conclure qu'il n'y a pas de différences entre les groupes (niveau d'études).

Dans ce cas, nous pouvons conclure que les moyennes de ces catégories sont presque identiques.

D'autre part, une p-value significative (inférieure au seuil de risque - ce qui n'est pas notre cas), indique que certaines moyennes sont différentes, mais nous ne savons pas quelles paires de groupes sont différentes.

Il est possible d'effectuer plusieurs comparaisons par paires afin de déterminer si la différence moyenne entre des paires de groupes spécifiques est statistiquement significative.

Même si le test ANOVA n'est pas significatif, nous avons calculé le Tukey HSD (Différence significative honnête de Tukey, fonction R : `TukeyHSD()`) pour effectuer plusieurs comparaisons par paires entre les moyennes de groupes.

La fonction `TukeyHD()` prend l'ANOVA adaptée en argument (`AnovaModel.1`).

Les paires les plus significatives retenues par `TukeyHD` sont :

```
> TukeyHSD(AovaModel.1)
Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = Interest ~ Year, data = anova)

$`Year`
              diff      lwr      upr    p adj
2ème année-1ère année  0.294951787 -0.2800735  0.8699770  0.6782182
3ème année-1ère année  0.353488372 -0.1569336  0.8639104  0.3486344
Master-1ère année     -0.046511628 -1.9521060  1.8590827  0.9999998
Master of Science-1ère année  0.203488372 -0.8108417  1.2178184  0.9923275
Master spécialisé-1ère année  0.286821705 -0.6788280  1.2524714  0.9561341
3ème année-2ème année  0.058536585 -0.4595400  0.5766132  0.9995056
Master-2ème année     -0.341463415 -2.2491223  1.5661955  0.9954929
Master of Science-2ème année -0.091463415 -1.1096668  0.9267400  0.9998394
Master spécialisé-2ème année -0.008130081 -0.9778476  0.9615874  1.0000000
Master-3ème année     -0.400000000 -2.2891901  1.4891901  0.9901633
Master of Science-3ème année -0.150000000 -1.1331657  0.8331657  0.9978876
Master spécialisé-3ème année -0.066666667 -0.9995271  0.8661938  0.9999481
Master of Science-Master  0.250000000 -1.8326363  2.3326363  0.9993360
Master spécialisé-Master  0.333333333 -1.7260325  2.3926992  0.9971994
Master spécialisé-Master of Science  0.083333333 -1.1967305  1.3633972  0.9999673
```

- **Hypothèse 3 :** La data science est un domaine aussi connu que l'informatique, l'électronique ou les mathématiques.

Test de normalité

```
shapiro.test(data$Connaissance_DataScience)
```

Shapiro-Wilk normality test
data: data\$Connaissance_DataScience
W = 0.89998, p-value = 1.981e-09

```
shapiro.test(data$Connaissance_Electronique)
```

Shapiro-Wilk normality test
data: data\$Connaissance_Electronique
W = 0.89949, p-value = 1.852e-09

```
shapiro.test(data$Connaissance_Informatique)
```

Shapiro-Wilk normality test
data: data\$Connaissance_Informatique
W = 0.87251, p-value = 5.982e-11

```
shapiro.test(data$Connaissance_Mathematiques)
```

Shapiro-Wilk normality test
data: data\$Connaissance_Mathematiques
W = 0.88401, p-value = 2.423e-10

Donc, avec un risque de 10% on rejette l'hypothèse nulle de normalité pour chacun des 4 échantillons.

Les 4 échantillons ne suivent pas des lois normales. Par conséquent, on va utiliser le test non paramétrique de Wilcoxon.

Test de moyenne

Data science vs Electronique

```
wilcox.test(data$Connaissance_DataScience,data$Connaissance_Electronique)
```

```
#Wilcoxon rank sum test with continuity correction
# data: data$Connaissance_DataScience,data$Connaissance_Electronique
# W = 15906, p-value = 0.2974
# alternative hypothesis: true location shift is not equal to 0
```

Avec un risque de 10% on accepte H_0 : La Data Science est aussi connue que l'électronique.

Data science vs Informatique

```
wilcox.test(data$Connaissance_DataScience,data$Connaissance_Informatique)
```

```
# W = 7544, p-value < 2.2e-16
# alternative hypothesis: true location shift is not equal to 0
```

Avec un risque de 10% on rejette H_0 : La Data Science n'est pas aussi connue que l'Informatique.

Data science vs Mathématiques

```
wilcox.test(data$Connaissance_DataScience,data$Connaissance_Mathematiques)
```

```
# W = 8160.5, p-value = 5.422e-14
# alternative hypothesis: true location shift is not equal to 0
```

Avec un risque de 10% on rejette H_0 : la Data Science n'est pas aussi connue que les Mathématiques.

Conclusion : La data science est aussi connue que l'électronique, mais moins que l'informatique et les mathématiques

► **Hypothèse 4 :** La Data Science est perçue comme une nouvelle science par les étudiants.

Test de normalité sur les deux échantillons

```
shapiro.test(data$Combinaison_de_disciplines)
```

```
Shapiro-Wilk normality test
data: data$Combinaison_de_disciplines
W = 0.28919, p-value < 2.2e-16
```

```
shapiro.test(data$Nouvelle_sciences)
```

```
Shapiro-Wilk normality test
data: data$Nouvelle_sciences
W = 0.28919, p-value < 2.2e-16
```

Avec un risque de 10% on rejette l'hypothèse nulle de normalité.

Aucun des deux échantillons ne suivent pas des lois normales. Par conséquent, on va utiliser le test, non paramétrique : test de Wilcoxon.

Test de moyenne

```
wilcox.test(data$Nouvelle_science, data$Combinaison_de_disciplines, alternative = "l")
```

Wilcoxon rank sum test with continuity correction

data: data\$Nouvelle_science and data\$Combinaison_de_disciplines

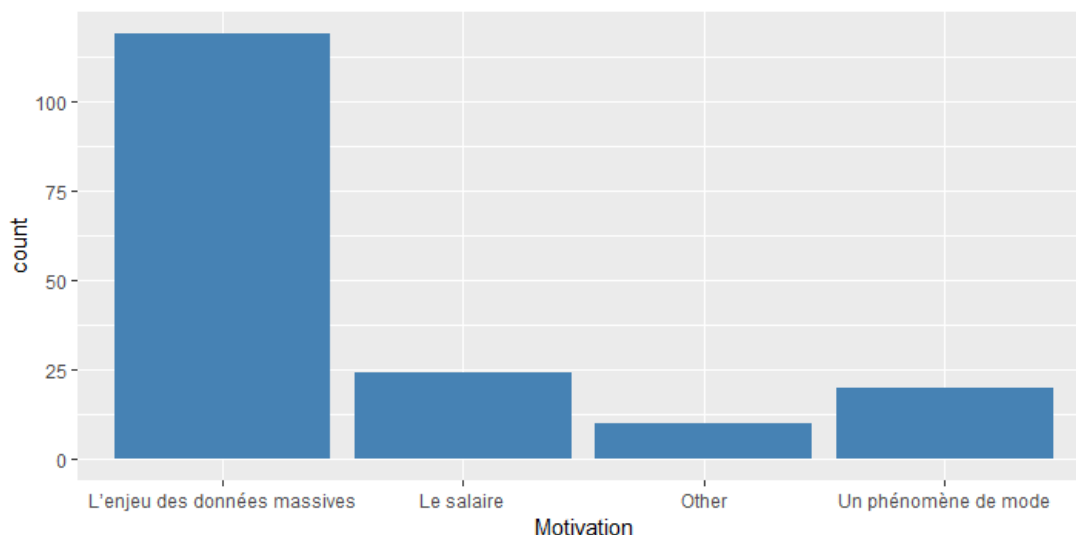
W = 2249, p-value < 2.2e-16

alternative hypothesis: true location shift is less than 0

Avec un risque de 10% on rejette H_0 . La Data Science est donc perçue comme une combinaison de sciences existantes (Mathématiques, Statistiques et Informatique).

4.2) Tests du Khi-2

- **Hypothèse 5 :** Le choix d'être/devenir un Data Scientist est motivé par un phénomène de mode



D'après l'histogramme le choix d'être/devenir un Data Scientist est motivé plutôt par l'enjeu des données massives. On rejette l'hypothèse 5.

- **Hypothèse 5 bis :** La motivation de devenir un Data Scientist ne dépend pas du genre.

```
tableau_motivation<-matrix(c(motivation_homme,motivation_femme),2,4,byrow=T)
```

```
chisq.test(tableau_motivation)
```

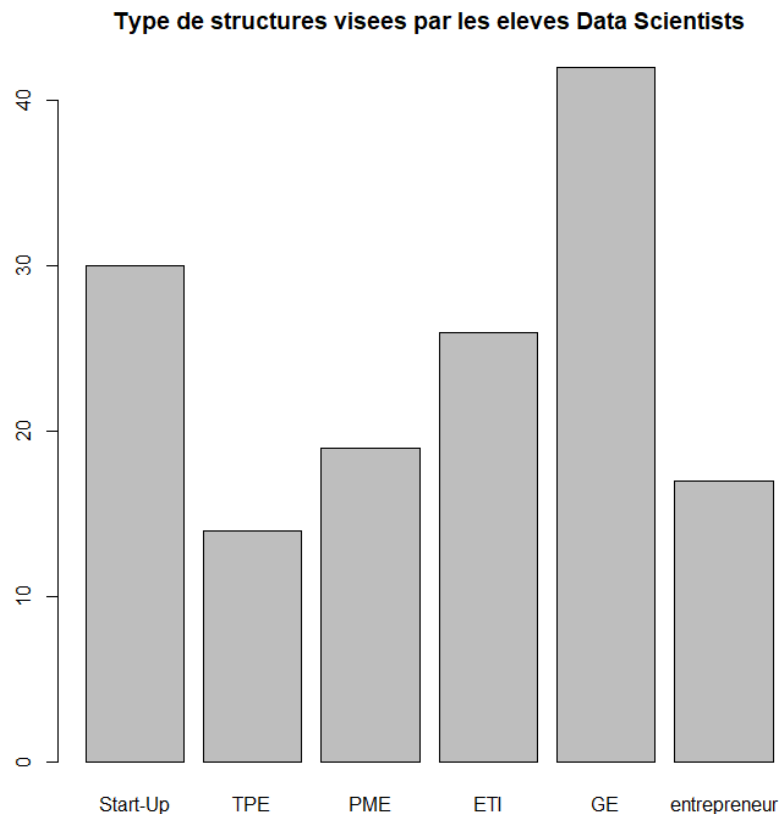
Pearson's Chi-squared test

data: tableau

X-squared = 4.6545, df = 3, p-value = 0.1989"

Avec un risque de 10%, on accepte H_0 : la motivation est indépendante du genre.

- **Hypothèse 6 :** La répartition de la préférence de type de structure pour le travail des étudiants data scientists est uniforme.



```
proba <- c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)
chisq.test(preferences_structure_etudiant_data_obs, p=proba)
```

Chi-squared test for given probabilities
 data: preferences_structure_etudiant_data_obs
 X-squared = 21.703, df = 5, p-value = 0.0005963

Avec un risque de 5%, on rejette H_0 . Donc la répartition de la préférence de type de structure des étudiants data scientist n'est pas uniforme.

- **Hypothèse 7 :** La répartition de la préférence des langages de programmation utilisés en Data Science est uniforme

```
preferences_langage_etudiant_data_obs <- c(63, 34, 15, 12, 21) #Dans cet ordre : Python, R, Java,
SQL, Other (Matlab, C++, Scala...)
proba_uniforme <- c(0.2, 0.2, 0.2, 0.2, 0.2)
chisq.test(preferences_langage_etudiant_data_obs, p=proba) #p-value = 3.428e-12
```

Chi-squared test for given probabilities
 data: preferences_langage_etudiant_data_obs
 X-squared = 59.655, df = 4, p-value = 3.428e-12

Avec un risque 10%, on rejette H_0 : la préférence des langages de programmation pour les étudiants en Data Science n'est pas équirépartie.

- **Hypothèse 7 bis** : La répartition de la préférence des langages de programmation utilisés en Data Science est uniforme (pour les étudiants n'étant pas en filière Data Science : ceci est donc ce qu'ils estiment)

```
preferences_langage_etudiant_non_data_obs <- c(97,38,36,46,61) #Meme ordre que le precedent
chisq.test(preferences_langage_etudiant_non_data_obs,p=proba) #p-value = 3.146e-09
```

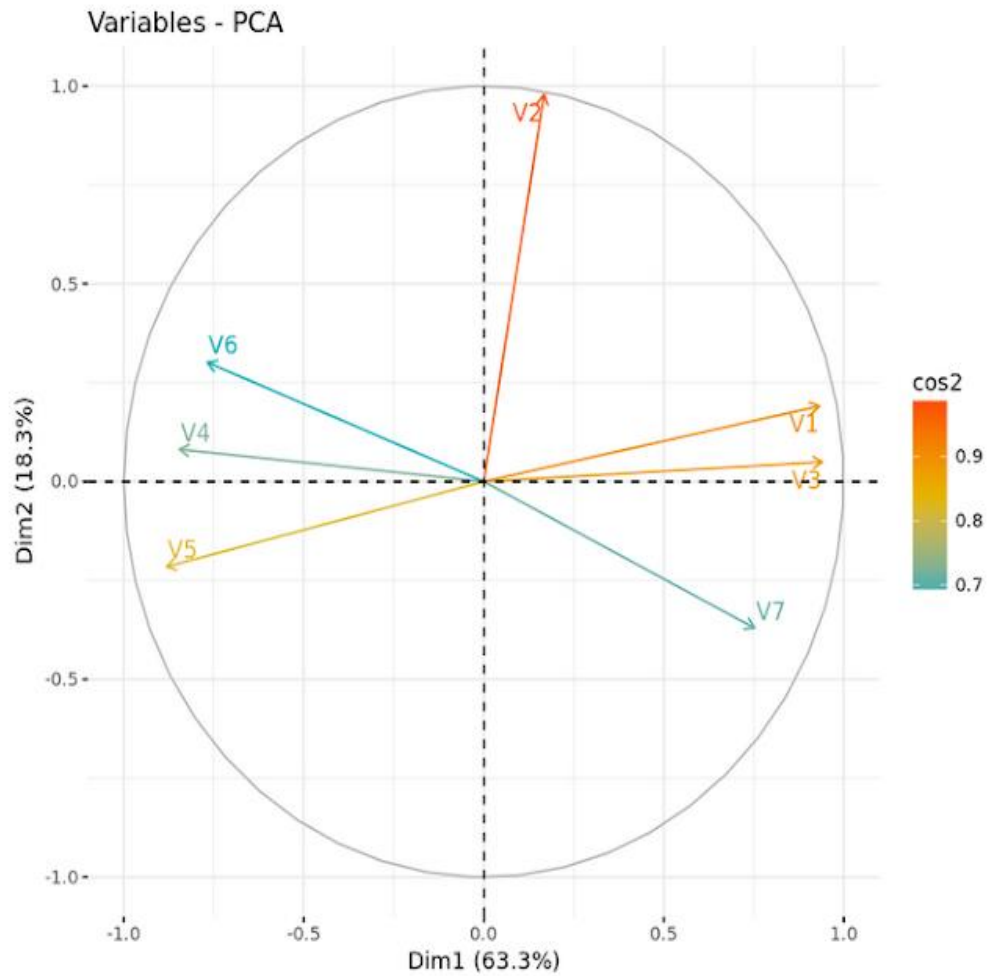
Chi-squared test for given probabilities
 data: preferences_langage_etudiant_non_data_obs
 $X\text{-squared} = 45.489$, $df = 4$, $p\text{-value} = 3.146e-09$

Avec un risque de 10%, on rejette H_0 : La préférence des langages de programmation pour les étudiants qui ne sont pas en Data Science n'est équirépartie.

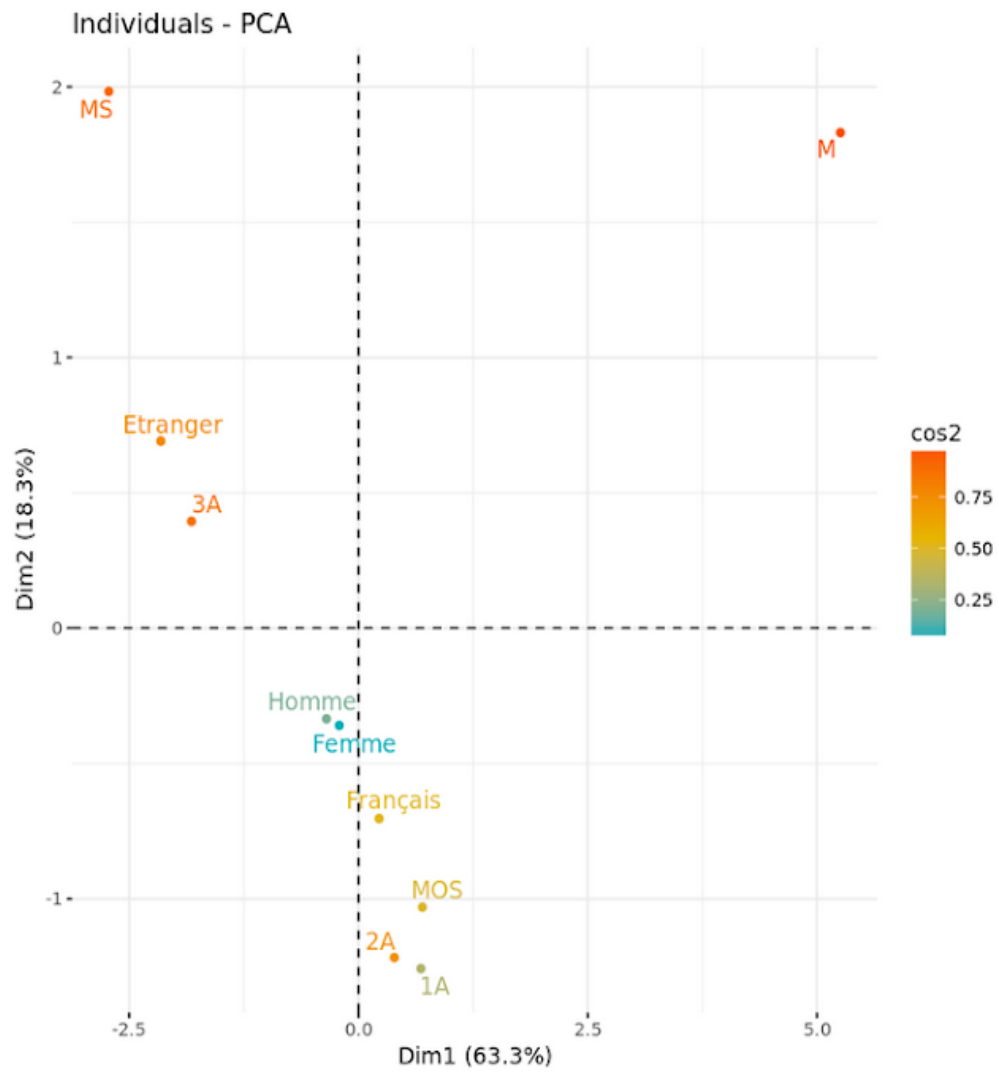
Analyse approfondie : ACP

Afin d'avoir une analyse plus fine, on utilise l'analyse par composantes principales (PCA). On obtient une ACP en deux composantes, avec deux bonnes inerties dim1(63,30%) et dim2(18,25%).

	Français	Etranger	Homme	Femme	1A	2A	3A	MS	MOS	M
Langage										
C	0.258065	0.0816327	0.214876	0.192308	0.470588	0.243902	0.1	0	0.125	1
R	0.33871	0.612245	0.421488	0.384615	0.264706	0.243902	0.571429	0.888889	0.375	1
java	0.306452	0.265306	0.256198	0.384615	0.617647	0.341463	0.128571	0.111111	0.5	1
matlab	0.16129	0.204082	0.157025	0.173077	0.235294	0.121951	0.185714	0.333333	0.125	0
python	0.693548	0.897959	0.752066	0.711538	0.794118	0.731707	0.842857	0.777778	0.75	0.5
scala	0.0483871	0.0816327	0.0330579	0.0961538	0.0588235	0.0243902	0.0857143	0.111111	0	0
sql	0.403226	0.163265	0.280992	0.442308	0.558824	0.365854	0.257143	0.222222	0.375	0.5



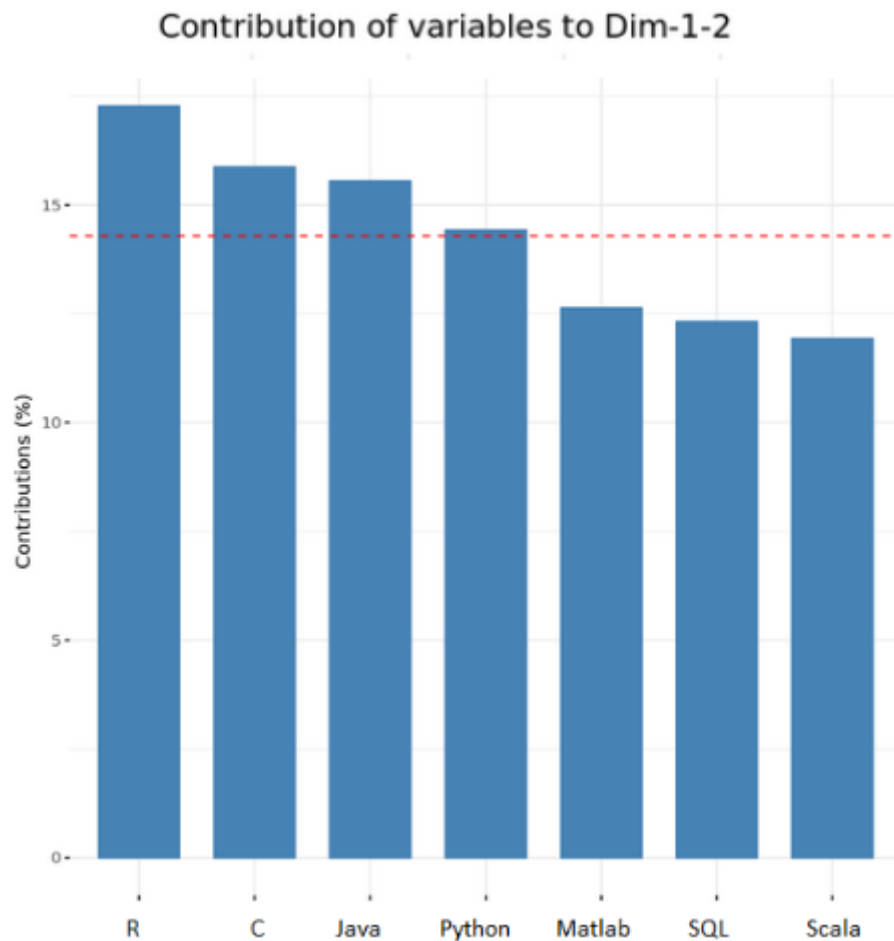
V1 : C
V2 : R
V3 : Java
V4 : Matlab
V5 : Python
V6 : Scala
V7 : SQL



MOS : Master of Science

MS : Mastère spécialisée

M : Master



Master est un point aberrant, seulement deux personnes en Master ont répondu au questionnaire.

C, Java et SQL sont positivement corrélés avec la dim1 ($\cos^2 > 0.6$)

R est positivement corrélé avec la dim2 ($\cos^2 > 0.6$)

Interprétation de l'ACP

- 1- R est très utilisé par les Mastères spécialisés, alors qu'ils utilisent très peu C et Java. On peut raisonnablement supposer que les Mastères spécialisés ont un profil de Statisticiens et utilisent peu Java ou C qui sont plus des langages de développement informatique.
- 2- Les 1^{ères} années et les 2^{èmes} années sont négativement corrélées à la dim2 et Python et SQL sont négativement corrélés à dim2. Donc les 1^{ères} années et les 2^{èmes} années préfèrent Python et SQL. Ce qui peut s'expliquer par le fait que Python et SQL sont les langages qu'ils apprennent en classes préparatoires (vu que la majorité des 1^{ères} années et des 2^{èmes} années est recrutée après les classes préparatoires).
- 3- Le genre n'a pas d'influence sur le choix du langage car Homme et Femme sont proches dans le diagramme d'individual factor map.
- 4- Étranger est négativement corrélé avec dim1. Matlab est positivement corrélé avec dim2 et négativement avec dim1. Et SQL est négativement corrélé avec dim1 et positivement avec dim2.
On peut supposer que les étrangers préfèrent Matlab et n'utilisent pas SQL.

4.3) Tests sur les projets/MOOC en Data Science

- **Hypothèse 8 :** Un élève ingénieur (toutes filières confondues) travaille en moyenne sur trois projets Data durant sa formation.

```
t.test(data_projet_oui$Nbr_projet_data, mu=3)
```

```
One Sample t-test
data: data_projet_oui$Nbr_projet_data
t = 0.56743, df = 52, p-value = 0.5729
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
 2.473578 3.941516
sample estimates:
mean of x
 3.207547
```

Avec un risque de 5%, on accepte l'hypothèse H_0 : les élèves ingénieurs travaillent en moyenne sur trois projets Data Science durant leur formation.

- **Hypothèse 9 :** les étudiants Data Scientists passent en moyenne 4h par semaine à programmer dans le cadre d'un projet Data Science.

```
shapiro.test(tps)
"Shapiro-Wilk normality test
```

```
data: tps
W = 0.75904, p-value = 6e-08 "
```

Avec un risque de 10%, on rejette H_0 . Donc, la distribution du temps moyen consacré à un projet Data science ne suit pas une loi normale

```
wilcox.test(tps, mu=4)
```

```
"Wilcoxon signed rank test with continuity correction
data: tps
V = 507, p-value = 0.06528
alternative hypothesis: true location is not equal to 10.68868"
```

Avec un risque de 5%, on accepte H_0 . La moyenne par semaine du nombre d'heures passé à programmer dans le cadre d'un projet Data Science est 4 heures.

- **Hypothèse 10 :** les étudiants apprennent la Data science à travers des MOOCs.

```
shapiro.test(form_MOOC)
```

```
Shapiro-Wilk normality test
data: form_MOOC
W = 0.49214, p-value < 2.2e-16
```

On rejette H_0 . La distribution n'est pas normale.

```
shapiro.test(form_init)
```

Shapiro-Wilk normality test

data: form_init

$W = 0.6275$, $p\text{-value} < 2.2e-16$

#p-value = $2.2e-16$

On rejette H_0 . La distribution n'est pas normale

```
wilcox.test(form_MOOC,form_init,alternative = 'g')
```

On accepte H_0 avec $p\text{-value} = 1$. Les étudiants data scientists formés en formation initiale sont plus nombreux que les étudiants data scientists formés par les MOOC.

- **Hypothèse 11 :** pour les étudiants ayant suivi un MOOC Data science, les étudiants passent en moyenne 10 heures de formation en ligne sur un MOOC.

```
shapiro.test(data_MOOC_oui$Heures_mooc)
```

Shapiro-Wilk normality test

data: data_MOOC_oui\$Heures_mooc

$W = 0.82289$, $p\text{-value} = 4.465e-06$

- **Hypothèse 12 :** pour les étudiants ayant suivi un MOOC Data Science, le taux de satisfaction est 70% (7/10)

Dans notre questionnaire nous avons choisi de poser deux questions :

Avez-vous déjà suivi un MOOC de Data Science ?

Quel est votre degré de satisfaction vis-à-vis des MOOC ?

On doit tout d'abord sélectionner les étudiants qui ont déjà suivi un MOOC et après voir leur degré de satisfaction.

```
etudiant_mooc <- results[ which(results$Mooc=="Oui"),]
```

```
s?tis <- as.numeric(etudiant_mooc$Satisfaction)
```

En supposant que la variable satisfaction MOOC suit une loi normale pour n grand, nous avons effectué un test de Student d'égalité de moyenne :

```
t.test(satis,mu = 7)
```

One Sample t-test

data: satis

$t = 0.081088$, $df = 48$, $p\text{-value} = 0.9357$

alternative hypothesis: true mean is not equal to 7

95 percent confidence interval:

6.473560 7.485624

sample estimates:

mean of x

6.979592

Avec un risque de 5% on accepte H_0 , le taux de satisfaction moyen vis-à-vis des étudiants qui ont déjà suivi un MOOC est de 70%.

5) DISCUSSION CRITIQUE DE L'ETUDE

Le but de ce projet était de mettre en œuvre les tests statistiques vus en cours.

Nous pouvons raisonnablement penser que les données que nous avons recueillies sont biaisées. En effet, il y a une sur-représentativité de certaines populations alors que d'autres sont sous-représentées. Par exemple, les femmes et les français sont sur-représentées avec respectivement 121,57% et 92,18%, alors que les hommes et les étrangers sont sous-représentés avec des taux de 119.47% et 70,8%.

L'autre biais introduit est celui des questions qui sont dirigées à des populations d'élèves qui n'ont peut-être pas le recul nécessaire pour aborder des thèmes complexes comme la Data Science.

Le troisième biais introduit découle du fait que notre questionnaire était adressé à des élèves de l'IMT Atlantique, qui propose une formation en Data Science, là où d'autres établissements ne proposent pas de formation en Data Science et par conséquent leurs élèves auront (à priori) une connaissance différente de cette science. C'est pour cela que nous parlons tout au long de l'étude de la perception des élèves de l'IMT Atlantique, il ne semble pas judicieux d'appliquer ces résultats pour l'ensemble des étudiants ingénieurs.

Pour pallier à ces biais, nous proposons les démarches suivantes :

- Utiliser un sondage par quota, afin de respecter les représentativités des différentes catégories (Genre, Nationalité, Année d'études...).
- Envoyer les questions à des groupes d'écoles ou à des universités représentatives de l'enseignement supérieur français.
- Améliorer la stratégie de communication en vue d'augmenter le nombre de réponses reçues.

6) CONCLUSION

Cette étude que nous avons menée nous a apporté un certain nombre d'enseignements sur la perception et l'utilisation des Data Sciences par les élèves de l'IMT Atlantique. En effet, cette discipline étant de plus en plus demandée dans le monde professionnel, il est intéressant de comprendre la perception qu'en ont des étudiants qui peuvent suivre des cours sur la Data Science au cours de leur formation. Les conclusions majeures que nous pouvons en tirer sont :

Les femmes et les hommes sont tout autant intéressés par les Data Sciences, en revanche les étrangers semblent être plus attirés par ce domaine.

Les Data Sciences sont une discipline qui reste moins connue que des disciplines phares de l'école que sont les Mathématiques et l'informatique. En revanche, elles sont toutes aussi connues que l'électronique.

Les étudiants ont compris que l'émergence des Data Sciences est due à l'enjeu des données massives à l'heure d'aujourd'hui.

De plus, les élèves de l'école pensent que les Data Scientists utilisent plus des langages de programmation tels que Python ou Matlab. Cela semble logique au vu de l'apprentissage par la majorité des élèves de ces langages en classes préparatoires.

Enfin en ce qui concerne la formation des élèves, la plus grande partie se forme via les filières proposées dans l'école. Cependant, pour ceux qui ont suivi un ou plusieurs MOOC sur les Data Sciences ils sont satisfaits du contenu, leur degré de satisfaction est de 70%.

En somme, les Data Sciences qui sont une discipline de plus en plus recherchée, semblent bien être un sujet d'intérêt pour les élèves de l'école.

7) BIBLIOGRAPHIE

XLStat [en ligne]. 2018 (consulté le 12 Novembre 2018). Disponible sur : https://help.xlstat.com/customer/fr/portal/articles/2062457-guide-de-choix-de-test-statistique?b_id=9283


Statistical Tools For High-Throughput Data Analysis [en ligne] (consulté le 15 Novembre 2018) Disponible sur : <http://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique/73-acp-analyse-en-composantes-principales-avec-r-l-essentiel/>

Google [en ligne] (consulté en ligne le 12 Novembre 2018). Disponible sur : <https://sites.google.com/site/rgraphiques/4--stat/anova-analyse-de-la-variance>

8) ANNEXES

Le questionnaire

a) Questions préliminaires



90% des données aujourd'hui ont été créés il y a moins de deux ans, ce qui justifie en partie l'explosion de la Data Science. Entre 2000 et 3000 profils Data Scientist sont recherchés par an en France, et pour répondre à cette demande, des formations en Data Science ont été conçues et c'est le cas de l'IMT Atlantique. Néanmoins, la perception et les pratiques de la Data Science restent diverses et variées.

Au Coeur de la Data Science

0% 100%

Questions préliminaires

***Vous êtes en quelle année d'études ?**
Choose one of the following answers

☐ 1ère année
☐ 2ème année
☐ 3ème année
☐ Master of Science
☐ Master
☐ Master spécialisé

***Quel est votre genre ?**
Choose one of the following answers

☐ Homme
☐ Femme
☐ Other

***Quelle est votre nationalité ?**
Choose one of the following answers

☐ Française
☐ Etrangère

***Quelle est votre nationalité ?**
Choose one of the following answers

☐ Française
☐ Etrangère

***Êtes-vous étudiant en Data Science ou comptez-vous étudier dans cette filière ?**
Choose one of the following answers

☐ Oui
☐ Non

***Avez-vous déjà suivi un MOOC en lien avec les Data Sciences ?**
Choose one of the following answers

☐ Oui
☐ Non

b) 1^{ère} partie du questionnaire0%  100%

Questions (1ère partie)

* Quel est votre intérêt pour la Data Science ? (Par intérêt on entend la lecture de façon générale, la lecture d'articles scientifiques, l'utilisation fréquente ou pas de la Data Science...)

Choose one of the following answers

- ☐ Pas du tout intéressé
- ☐ Pas Intéressé
- ☐ Indifférent
- ☐ Intéressé
- ☐ Tout à fait intéressé

* Quelle est votre perception de la data science ? C'est une :

Choose one of the following answers

- ☐ Une Nouvelle science qui vient d'être inventée.
- ☐ Une combinaison de disciplines existantes (maths, stats, info...)

* Pour chacun des domaines suivants : Cochez la case de ce que vous sauriez faire :

	1	2	3	4	5
Data Science	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Electronique	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Informatique	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Mathématiques	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

? 1- Vous pouvez définir le sujet d'une façon basique et vague. 2- Vous pouvez donner quelques exemples d'utilisation du sujet. 3- Vous pouvez citer les outils utilisés dans le domaine. 4- Vous pouvez travailler dans le domaine. 5- Vous pouvez avoir une discussion approfondie avec un professionnel du domaine.

* Pour vous, travailler en Data Science est un choix motivé par :

Choose one of the following answers

- ☐ Un phénomène de mode
- ☐ L'enjeu des données massives
- ☐ Le salaire
- ☐ Other

c) 2^{ème} partie du questionnaire

Dans le cas où la personne est en filière Data Science, a déjà fait un MOOC sur ce domaine, et a déjà fait un projet en lien avec les Data Sciences. Sinon certaines questions ne sont pas posées (voir l'arbre de décision).

Questions (2ème partie)

*** Avez-vous déjà travaillé sur un projet Data science**
Choose one of the following answers

Oui

*** Combien de temps consacrez-vous à un projet Data science par semaine ? (en heure)**

Only numbers may be entered in this field

*** Sur combien de projets Data Science avez-vous déjà travaillé ?**

Only numbers may be entered in this field

*** Dans quel cadre avez-vous appris la data science ?**
Check any that apply

☐ MOOC

☐ Ecole d'ingénieur ou formation universitaire

☐ Stages / projets

☐ Je n'ai encore rien appris en lien avec la data-science

*** Quelles plateformes avez-vous utilisées pour vos MOOC en lien avec la data-science ?**
Check any that apply

☐ Coursera

☐ Edx

☐ Fun

☐ Udemy

☐ Udacity

☐ Datacamp

☐ Other:

*** Combien d'heures avez-vous consacré pour un MOOC data-science au total ? (moyenne du nombre d'heure passé par MOOC)**

Only numbers may be entered in this field

*** Quel est votre niveau de satisfaction vis-à-vis des MOOCs suivis en data-science ?**

	1	2	3	4	5	6	7	8	9	10
Satisfaction de 1 (pas satisfait) à 10 (très satisfait)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

***Quels langages de programmation utilisez-vous en priorité en Data Science?**

Check any that apply

- ☐ Python
- ☐ R
- ☐ Matlab
- ☐ SQL
- ☐ Scala
- ☐ Java
- ☐ C++
- ☐ Other:

***En tant que Data Scientist, dans quel type de structure souhaitez-vous travailler ?**

Check any that apply

- ☐ Start-up
- ☐ TPE (très petite entreprise) (effectifs compris entre 0 et 19 salariés)
- ☐ PME (petite et moyenne entreprise) (effectifs compris entre 20 et 249 salariés)
- ☐ ETI (entreprises de taille intermédiaire) (effectifs compris entre 250 et 4999 salariés)
- ☐ Grandes entreprises (plus de 5000 salariés)
- ☐ Chez-soi - entrepreneur
- ☐ Other: