UV101E : Projet Statistiques 1

Lisez-vous vos mails ?

Rapport

Version 1

F3B - Année scolaire 2018-2019

Groupe 1 :

El Mahdi AIT SIDI ALI

Hiba KSOURI

Ahmed TRIFA

Etienne JENNER

Afef Ben KHALIFA

Amal SOUFIANI

Encadrants techniques :

BILLOT Romain

COPPIN Gilles

GOUVERNNEC Bernard

IMT Atlantique

Bretagne-Pays de la Loire

École Mines-Télécom

Sommaire

1. INTRODUCTION

1.1 CONTEXTE

Avec l'ère digitale, qui parmi nous ne possède pas d'adresse mail de nos jours ? Que ce soit pour recevoir toutes les dernières nouveautés de votre produit préféré ou pour communiquer avec votre entourage professionnel, des mails, nous en recevons presque tous quotidiennement ! Ils sont devenus le principal moyen de communication dans le monde professionnel mais aussi le plus formel des outils. Ceci étant, le droit à la déconnexion est désormais reconnu par la loi afin d'assurer le respect des temps de repos et de congé ainsi que de la vie personnelle et familiale de tout un chacun. Malgré cela, certaines personnes demeurent obsédées et ne peuvent s'empêcher d'actualiser leur boîte mail toutes les 10 secondes pendant que d'autres peuvent accumuler des milliers de mails non lus.

Dans ce contexte, notre groupe s'est constitué dans le cadre du projet statistiques afin de mener une enquête dont le but est de dégager les principales tendances du personnel et des étudiants de IMT Atlantique dans le traitement de leurs mails. Dès le choix de notre sujet, nous avons déjà pu remarquer qu'au sein de notre groupe même nous n'interagissons pas de la même manière avec nos mails, nous étions donc curieux de voir les différences de comportement à IMT Atlantique.

1.2 METHODOLOGIE POUR LE QCM ET L'ECHANTILLONNAGE

Notre enquête cible toutes les personnes utilisant régulièrement leur boîtes mail. Et pour réduire notre périmètre, nous avons supposé que les étudiants et le personnel de l'IMT Atlantique représentent ensemble un bon sous échantillon de la population mère ciblée par notre enquête. Cet échantillon comprend à la fois des étudiants, des enseignants mais aussi le personnel de la cantine et autres, il est donc bien varié et correspond à notre enquête. Nous avons ensuite construit un sondage qui s'adresse à toute personne possédant une adresse mail IMT, c'est-à-dire à toute personne étudiant ou travaillant à l'école. Ainsi, nous avons fait le choix de procéder par un échantillonnage par grappes, c'est à dire que l'on considère uniquement IMT Atlantique pour représenter la population mère et nous envoyons le sondage à tous les individus de l'échantillon.

La réponse au sondage était basée sur le volontariat.

Pour cette enquête, nous avons alors choisi de procéder par un échantillonnage par grappes basé sur le volontariat.

2. NETTOYAGE ET DESCRIPTION DES DONNEES

2.1 NOMENCLATURE DES COLONNE ET SUPPRESSION DES OUTLIERS

Dans un premier temps nous avons renommé les noms des colonnes pour faciliter nos analyses. Puis, nous avons cherché à supprimer les valeurs aberrantes parmi nos réponses.

En regardant les données (min, max, moyenne, summary), on s'aperçoit qu'il y a peu de valeurs qui semblent aberrantes par rapport aux questions posées. A première vue, seules 2 colonnes sont concernées (le temps.passé.moyen.sur.internet avec des valeurs supérieurs à 24h et le nombre.de.boîtes.mail avec un maximum de 2000).

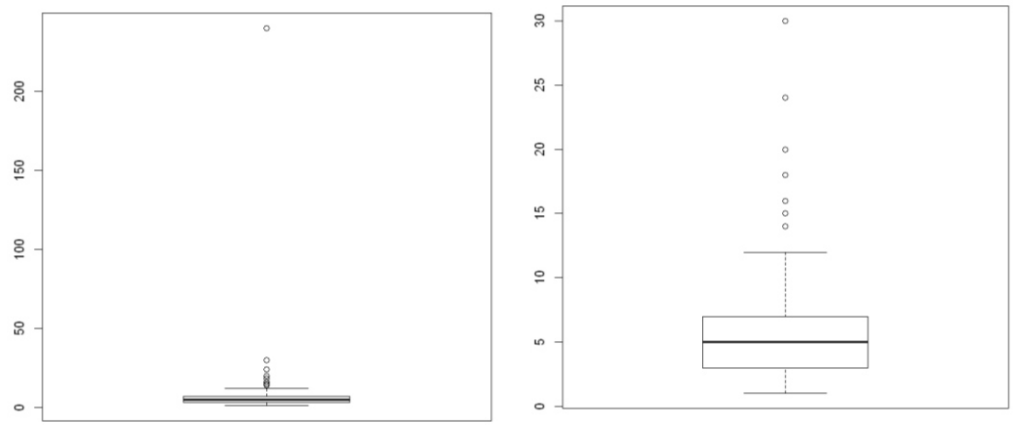
>> summary(mail)

```
                statut            age              expérience.professionnelle
1A                    :96   Min.   :18.00   Ça se compte en années:123
3A                    :96   1st Qu.:21.00   Entre 1 et 6 mois     :131
2A                    :77   Median :22.00   Expérience > 6 mois : 79
Autre                 :42   Mean   :26.06   Sans Expérience       : 89
Personnel Administratif :36   3rd Qu.:25.00
Enseignant            :34   Max.   :65.00
(Other)                     :41

temps.passé.moyen.sur.internet appareil.préféré.de.consultation  nombre.de.boites.mail
Min.   :  1.000                Un ordinateur      :209          Min.   :   1.000
1st Qu.:  3.000                Un smartphone      :206          1st Qu.:   3.000
Median :  5.000                Une montre connectée:  5         Median :   3.000
Mean   :  6.204                Une tablette       :  2          Mean   :   8.815
3rd Qu.:  7.000                                                 3rd Qu.:   4.000
Max.   :240.000                                                 Max.   :2000.000

nombre.de.messagerie.  fréquence.de.consultation.  nombre.de.mails.   concerné consultées.régulièrement
de.la.messagerie       reçus.par.jour
Min.   : 0.000          < 10  :168                 Min.   :  1.00     Jamais  :  2
1st Qu.: 2.000          < 3   : 79                 1st Qu.:  5.00     Rarement:138
Median : 2.000          > 10  :118                 Median : 10.00     Souvent :272
Mean   : 2.524          > 30  : 57                 Mean   : 16.76     Toujours: 10
3rd Qu.: 3.000                                     3rd Qu.: 18.75
Max.   :12.000                                     Max.   :300.00

notification.  lecture.de.tous.les.mails  lecture.de.l.objet.expéditeur  nombre.de.mails.non.lus
de.mails       En grande partie:210       Non: 26                       Min.   :    0.0
Non:158        Quasiment tous  : 67       Oui:396                       1st Qu.:    0.0
Oui:264        Rarement        :145                                     Median :    2.0
                                                                        Mean   :  416.3
                                                                        3rd Qu.:   57.5
                                                                        Max.   :14300.0

nombre.de.mails.  réponse.à.des.mails.  consultation.de.mails.  personnalisation.
envoyés.par.semaine  nécessitant.une.réponse  en.dehors.des.heures.  de.la.messagerie
Min.   :  0.00        Non    : 37          de.travail.études       Non:186
1st Qu.:  2.00        Oui    :299          Non    : 22             Oui:236
Median :  5.00        Parfois: 86          Oui    :354
Mean   : 18.98                             Parfois: 46
3rd Qu.: 15.00
Max.   :350.00

importance.de.l.ergonomie  sentiment.envers.de.zéro.inbox  autre.fonctionnalité.de.la.boite.mail
Pas tellement importante: 26  Avoir 10.000 mails non lus: 41   Ah bonne idée                : 14
Relativement importante :184  Lire tous les mails      :193   Non, j'ai d'autres outils pour ça:150
Très importante         :212  Marquer tout en lu       :188   Oui, parfois                 :165
                                                             Oui, souvent                 : 93
```

On décide alors de supprimer la réponse entière, lorsqu'il y a une valeur aberrante, puisque cela ne concerne que peu de cas.

Pour déterminer les valeurs que l'on juge aberrantes, on trace des boxplots sur les 2 colonnes concernées et on regarde les valeurs supérieures à Q3 + 1,5*(Q3-Q1).

>>boxplot((mail$temps.passé.moyen.sur.internet))
(avec ou sans la valeur 240)



>> outliers_val=boxplot.stats(mail$temps.passé.moyen.sur.internet)$out

15 30 15 16 16 15 24 16 18 24 20 30 15 16 16 16 18 14 240 14

Pour le temps.passé.moyen.sur.internet, on décide de supprimer toutes les valeurs supérieures à Q3+1,5*(Q3-Q1), c'est à dire à partir de 14h.

>> outliers_val=boxplot.stats(mail$nombre.de.boites.mail)$out
1 8 1 6 6 6 10 6 6 6 1 7 1 7 10 1 15 7 6 8 7 6 7 12 6 6 6 7 2000 1 230 6 9 7 7 8

Pour le nombre.de.boites.mails, on estime qu'il est possible d'avoir jusqu'à 15 boites mails et on ne supprime que les réponses 2000 et 230 boites mails.

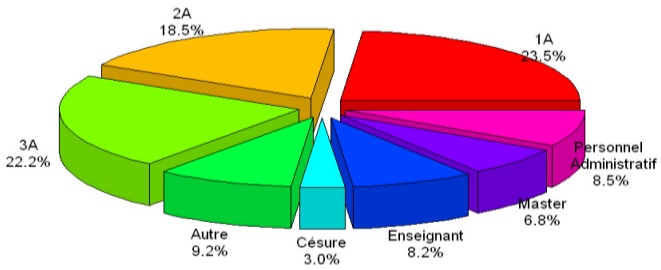
Après ce nettoyage il reste 400 réponses sur les 422 présentes initialement. Nous allons donc baser notre étude sur ces données nettoyées.

2.2 DESCRIPTION DES DONNEES

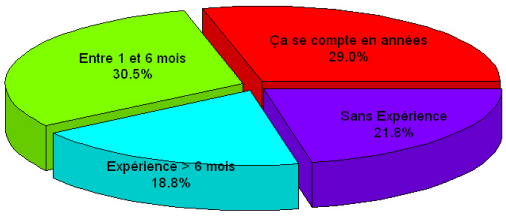
2.2.1 Le profil des sondés

Le sondage a été diffusé à l'ensemble des membres de l'IMT Atlantique, sans distinction de campus ou de statut au sein de l'école. Sans surprise donc on retrouve une grande majorité d'étudiants (notamment 1A, 2A, 3A) et un âge médian de 22 ans. La représentativité de l'échantillon au niveau du statut et de l'âge est bonne par rapport à la répartition réelle au sein de l'IMT Atlantique.

Répartition par profession et statut au sein de l'école



Répartition par expérience professionnelle



>> summary(age)

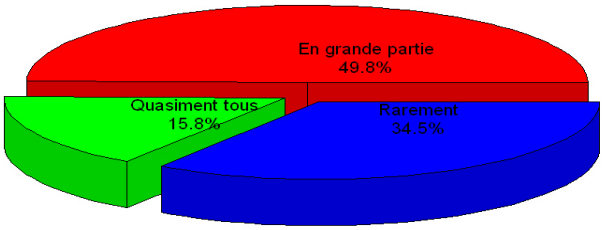
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
18.00	21.00	22.00	25.98	25.00	65.00

2.2.2 Quelques analyses directes des réponses

a- Les sondés lisent-ils leurs mails ?

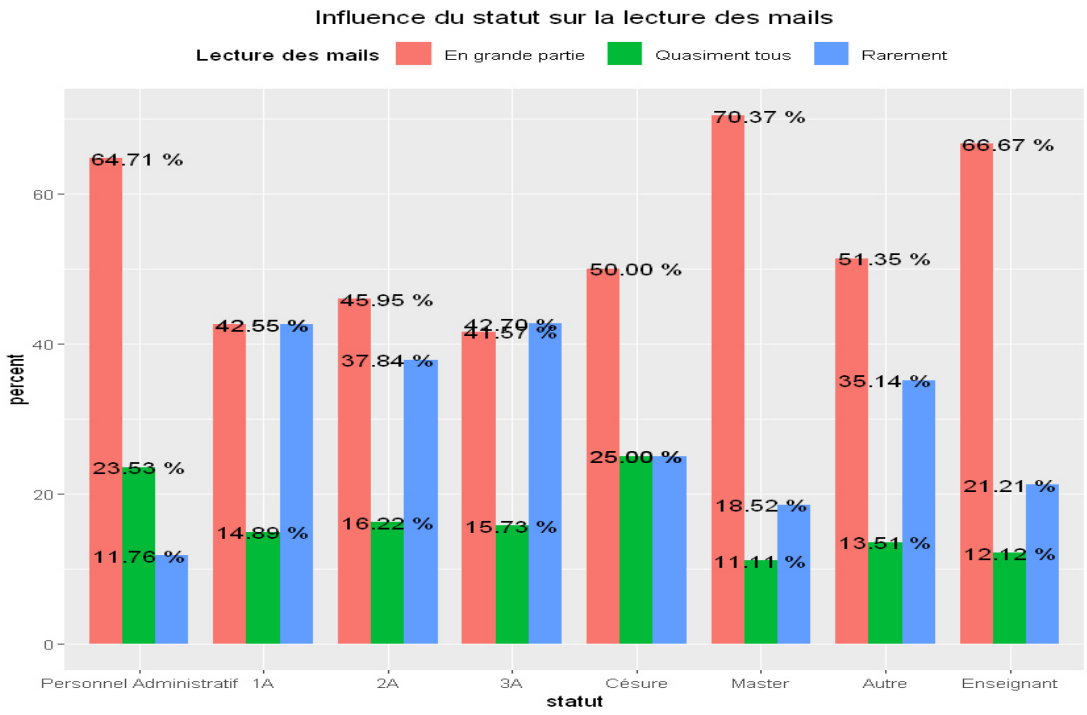
On observe sur ce premier graphique qu'une part importante des sondés (34,5%) lisent rarement tous leurs mails, et la part qui les lisent « quasiment tous » est nettement plus faible (15,8%). L'objet de l'étude va donc être de savoir si ces 2 populations se différencient du reste des sondés et par quelles caractéristiques.

Les sondés lisent-ils leurs mails ?



Un premier graphique montrant l'influence du statut (1A, 2A, 3A, Césure, Master, Enseignant, Personnel) sur le niveau de lecture des mails, indique que le pourcentage de personne lisant rarement tous leurs mails est nettement plus élevé chez les étudiants en 1A, 2A, 3A.

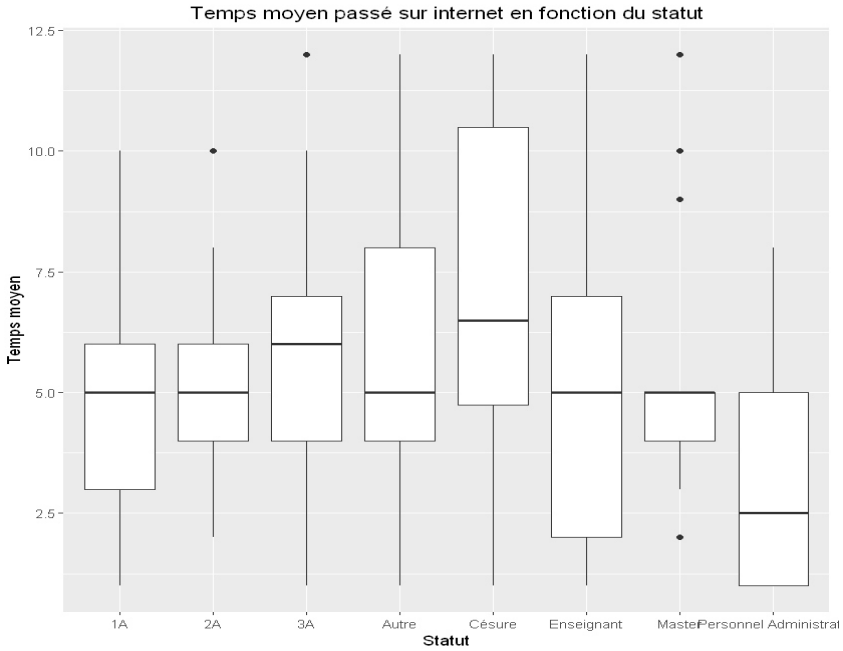
Il nous faut cependant faire attention à bien réaliser que ces valeurs ont probablement un biais car les personnes qui lisent moins leurs mails auront probablement eu moins tendance à répondre à notre questionnaire.



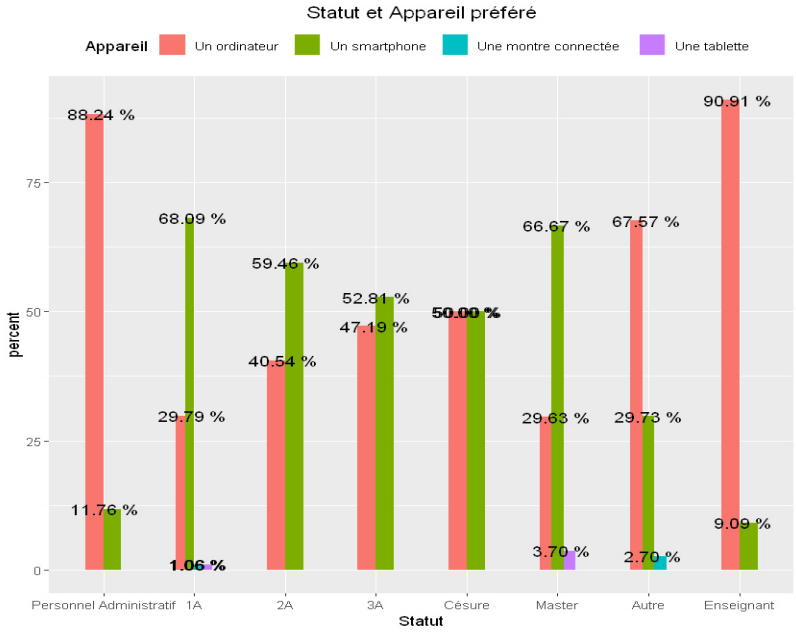
b- Quel comportement pour quels statuts ?

Nous nous sommes intéressés au profil digital des différentes catégories, et ici notamment le temps moyen passé sur internet par jour et l'appareil préféré pour la lecture des mails.

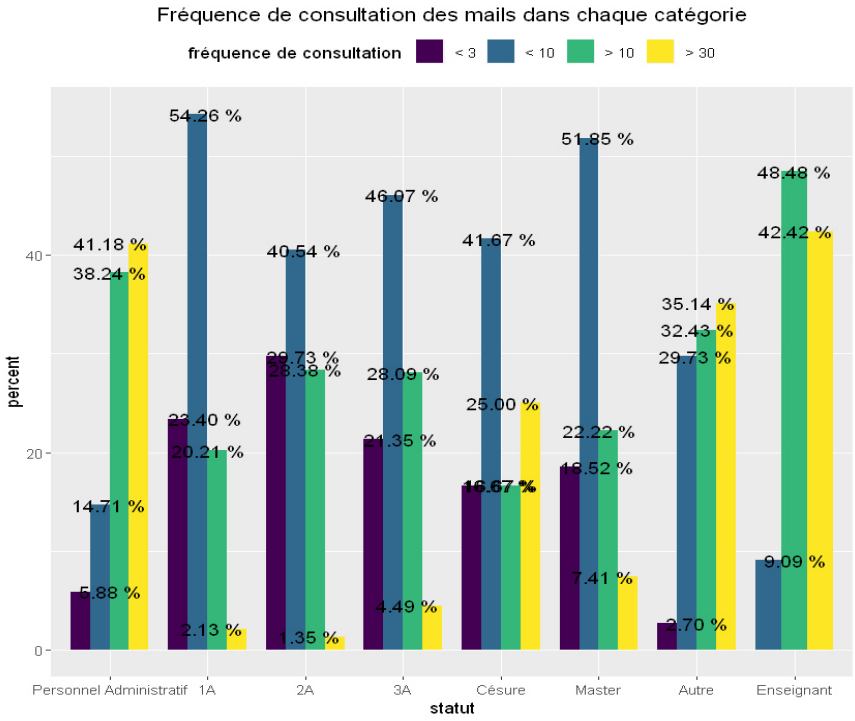
On s'aperçoit pour le temps moyen que 2 sous-populations se différencient. Les étudiants en césure passent nettement plus de temps connecté, probablement parce que dans le cadre de leur stage, la majorité de leur temps de travail est connecté. En revanche le personnel de l'école est moins connecté que la moyenne.



Concernant l'appareil utilisé, la séparation est très nette entre les étudiants qui ont tendance à privilégier le smartphone et les non-étudiants qui privilégient l'ordinateur.

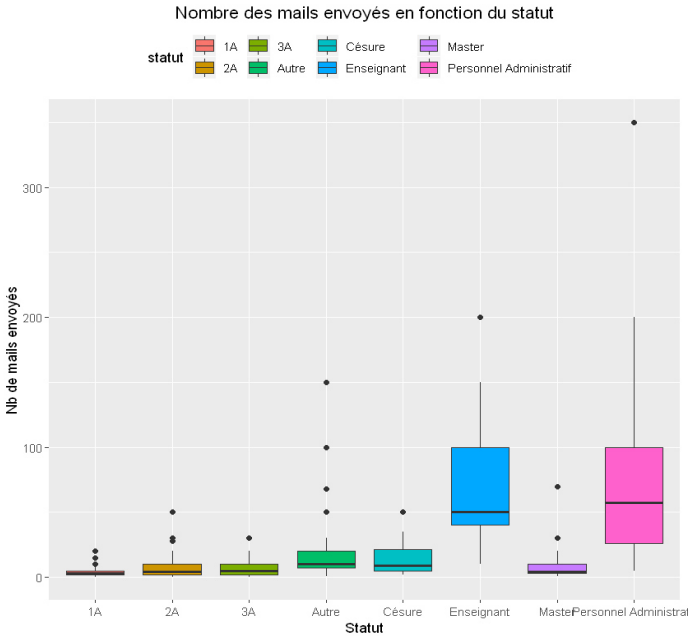


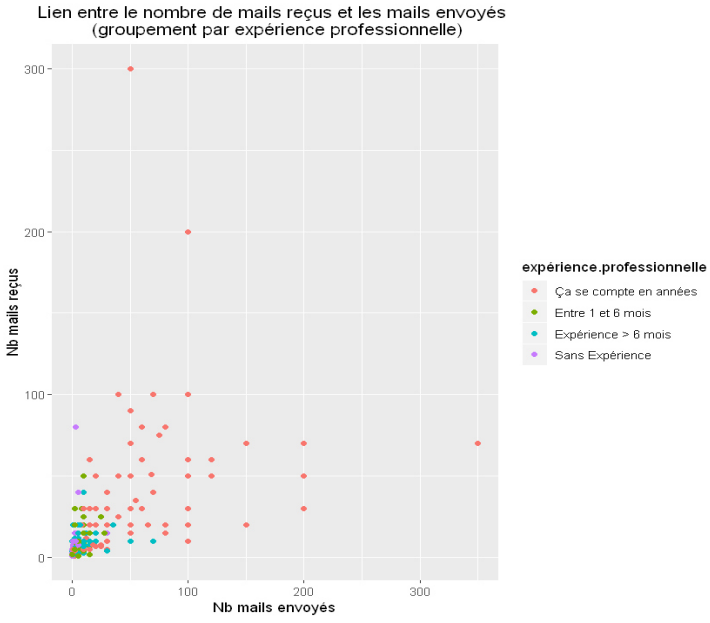
En ce qui concerne le nombre de consultation des mails par jour, les étudiants ont tendance à consulter leurs mails moins souvent que les non-étudiants de manière générale.



c- Impact de l'expérience professionnelle sur le nombre de mails reçus et envoyés

Ce graphique nous permet de voir que : la catégorie de personnes ayant une expérience professionnelle qui « se compte en années » se détache du reste des sondés avec des nombres de mails reçus (par jour) et envoyés (par semaine) plus importants. Concernant, le nombre de mails envoyés, 2 catégories se détachent : les enseignants et le personnel administratif. Les étudiants envoient peu de mails.





3. TESTS STATISTIQUES

3.1 GESTION DE LA MESSAGERIE PAR RAPPORT AU PROFIL

3.1.1 L'envoi des mails augmente avec le niveau d'expérience professionnelle

Hypothèse nulle : *Il n'y a pas de différence significative entre les mails envoyés des groupes ayant différents niveaux d'expérience pro.*
Hypothèse alternative : *Il existe au moins une moyenne de mails envoyés pour un échantillon (un niveau d'exp pro) qui n'est pas égale aux autres.*

a- Test ANOVA unidirectionnel

La fonction summary.aov() donne un résumé du modèle d'analyse de variance ANOVA

```
# Compute the analysis of variance
res.aov <- aov(nombre.de.mails.envoyés.par.semaine ~ expérience.professionnelle, data = df)
# Summary of the analysis
summary(res.aov)
```

	Df	*Sum Sq*	*Mean Sq*	*F value*	*Pr(>F)*
Expérience.professionnelle	***3***	***136929***	***45643***	***49.27***	***<2e-16 ******
Residuals	***396***	***366862***	***926***		

Comme la valeur P-value est inférieure au seuil de signification de 0.05, nous pouvons conclure qu'il existe des différences significatives entre les groupes de niveaux d'expérience pro.

b- Test Turkey de comparaison des moyennes multiples

Dans le test ANOVA unidirectionnel, une valeur P-value significative indique que certaines des moyennes de groupe 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.

Comme le test ANOVA est significatif, nous pouvons calculer Tukey HSD (Tukey Honest Significant Differences) pour effectuer plusieurs comparaisons par paires entre les moyennes de groupes. La fonction TukeyHSD() prend l'ANOVA adaptée en argument.

TukeyHSD(res.aov)

	diff	lwr	upr	p adj
Entre 1 et 6 mois-Sans Expérience	**2.447051**	**-8.572216**	**13.46632**	**0.9401154**
Expérience > 6 mois-Sans Expérience	**4.661149**	**-7.712174**	**17.03447**	**0.7655355**
Ça se compte en années-Sans Expérience	**42.925287**	**31.788023**	**54.06255**	**0.0000000**
Expérience > 6 mois-Entre 1 et 6 mois	**2.214098**	**-9.308268**	**13.73646**	**0.9600037**
Ça se compte en années-Entre 1 et 6 mois	**40.478236**	**30.294698**	**50.66177**	**0.0000000**
Ça se compte en années-Expérience > 6 mois	**38.264138**	**26.628876**	**49.89940**	**0.0000000**

Quand l'expérience est assez importante, qu'elle se compte en nombre d'année, le nombre de mails envoyés devient significativement élevé par rapport à des niveaux d'expérience plus faibles.

3.1.2 Les utilisateurs de PC pour la consultation de mails lisent la totalité des messages

Hypothèse nulle : *La tendance à lire la totalité des messages est indépendante de l'appareil de consultation préféré.*
Hypothèse alternative : *La tendance à lire la totalité des messages dépend de l'appareil de consultation préféré.*

a- Test khi2 de dépendance **X-squared = 6.4431, df = 6, p-value = 0.3754**

P-value > 0,05, on accepte H0 et donc les deux variables catégorielles sont indépendantes.

3.1.3 Les utilisateurs de Smartphone pour la consultation de mails lisent souvent que l'objet et l'expéditeur

Hypothèse nulle : *La tendance à lire que l'objet et l'expéditeur des mails est indépendante de l'appareil de consultation préféré.*
Hypothèse alternative : *La tendance à lire que l'objet et l'expéditeur des mails dépend de l'appareil de consultation préféré.*

a- Test khi2 de dépendance **X-squared = 2.4581, df = 3, p-value = 0.4829**

P-value > 0,05, on accepte H0 et donc les deux variables catégorielles sont indépendantes.

3.1.4 La lecture des mails dépend du statut de l'individu

> ***Hypothèse nulle :*** *La tendance de lecture de mails est indépendante du statut.*
> ***Hypothèse alternative :*** *La tendance de lecture de mails dépend du statut.*

a- Test khi2 de dépendance

Warning message in chisq.test(lecture.de.tous.les.mails, statut):
"Chi-squared approximation may be incorrect"

Pearson's Chi-squared test
X-squared = 23.758, df = 14, p-value = 0.04901

Le warning ci-dessus nous indique que le test du khi2 n'est pas valable dans le cas présent. Nous devons utiliser le test exact de Fisher:

a- Test de fisher

Fisher's Exact Test for Count Data with simulated p-value
p-value = 0.03998
alternative hypothesis: two.sided

Avec une P-value petite, nous pouvons conclure qu'il y a un lien entre les variables. Nous pouvons également valider cette relation à l'aide d'un tableau des effectifs :

Contingency table (statut & lecture des mails)

	En grande partie	Quasiment tous	Rarement
1A	40	14	40
2A	34	12	28
3A	37	14	38
Autre	19	5	13
Césure	6	3	3
Enseignant	22	4	7
Master	19	3	5
Personnel Administratif	22	8	4

Pour l'échantillon des étudiants en 1A, 2A et 3A, il y pratiquement autant de personnes qui lisent leurs mails rarement que des personnes qui le font en grande partie. Par contre pour les enseignants, personnel et les étudiants en Master, il y en a peu qui lisent rarement leurs mails. Ainsi, nous pouvons conclure qu'il y a une dépendance entre la lecture de mail et le statut : étudiant / personnel.

3.1.5 La fréquence de consultation de la messagerie change avec le niveau d'expérience

Hypothèse nulle : *La fréquence de consultation et le niveau d'expérience pro sont indépendants.*
Hypothèse alternative : *La fréquence de consultation dépend du niveau de l'expérience pro.*

a- Test khi2 de dépendance

X-squared = 72.302, df = 9, p-value = 5.375e-12

La P-value étant très petite, nous pouvons confirmer notre hypothèse alternative. Ainsi, pour comprendre encore plus l'effet de l'expérience sur la consultation nous avons tracé la table de contingence de notre échantillon.

Contingency table (fréquence de consultation & expérience pro)

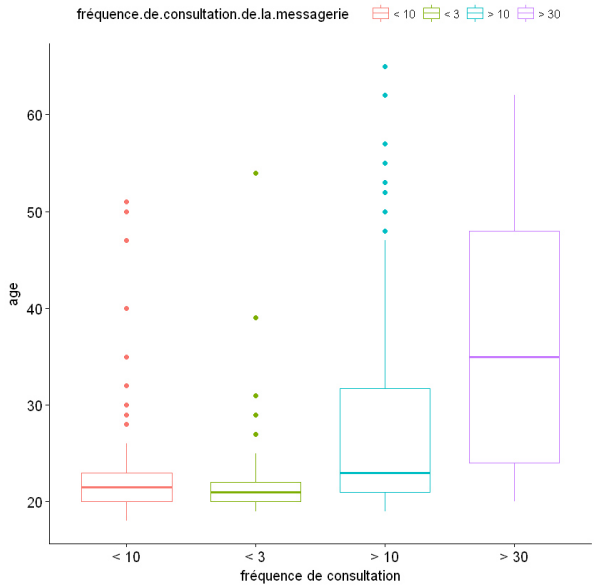
	< 10	< 3	> 10	> 30
Ça se compte en années	25	13	42	36
Entre 1 et 6 mois	65	27	23	7
Expérience > 6 mois	28	12	29	6
Sans Expérience	42	21	20	4

La fréquence de consultation devient plus importante avec l'expérience pro. En effet, quand cette dernière se compte en années, la fréquence est significativement plus élevée.

3.1.6 La fréquence de consultation de la messagerie dépend de l'âge de l'individu

Hypothèse nulle : *Il n'y a pas de différence significative d'âge entre les catégories de fréquences de consultation.*
Hypothèse alternative : *Il existe au moins une moyenne d'âge pour un échantillon (une fréq de consultation) qui n'est pas égale aux autres.*

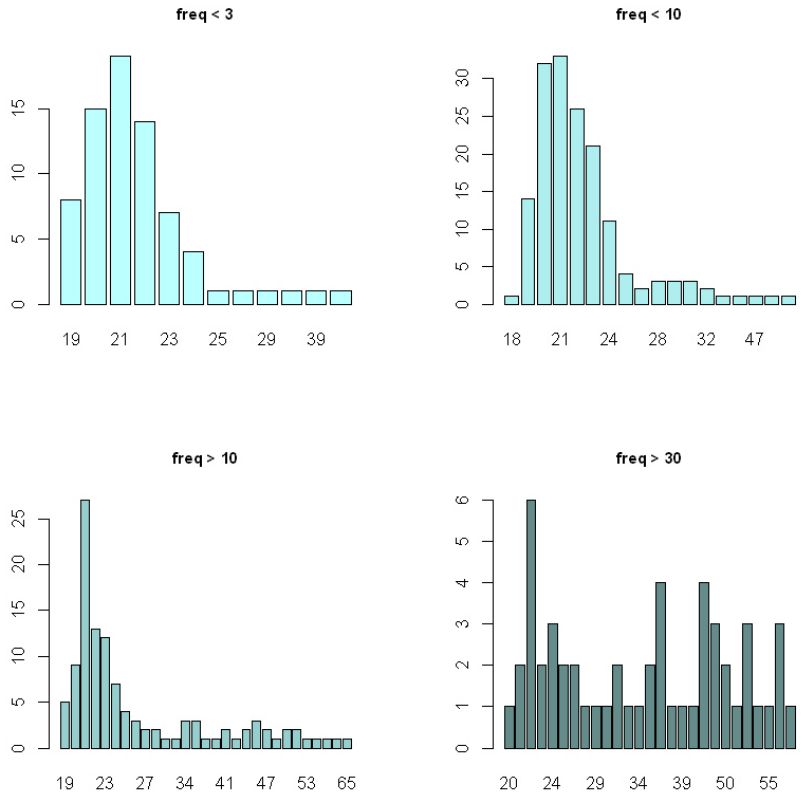
Avant de procéder au test, nous avons tracé un diagramme de boîte à moustaches afin de visualiser la répartition d'âge pour chaque catégorie de fréquence de consultation par jour :



a- Test Anova

	Df	*Sum Sq*	*Mean Sq*	*F value*	*Pr(>F)*
fréquence.de.consultation.de.la.messagerie	***3***	***9156***	***3052.0***	***44.54***	***<2e-16***
Residuals	***396***	***27135***	***68.5***		

La P-value étant faible, les variances ne peuvent pas être considérés égales. Vérifions la distribution :



La normalité n'est vraiment pas respectée. Il est donc nécessaire d'utiliser un test non paramétrique.

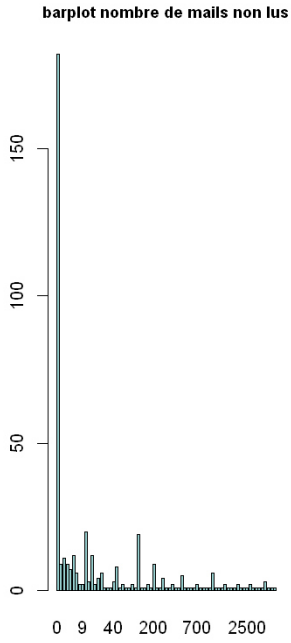
b- Test de Wallis:

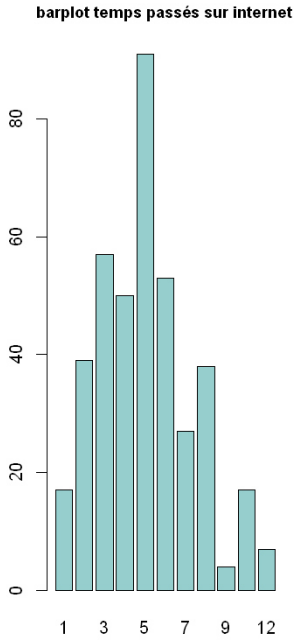
```
Kruskal-Wallis rank sum test
data:  age by fréquence.de.consultation.de.la.messagerie
Kruskal-Wallis chi-squared = 86.824, df = 3, p-value < 2.2e-16
```

La P-value étant significative, nous pouvons confirmer le lien entre l'âge de l'individu et sa fréquence de consultation de sa boîte de mails.

3.1.7 Plus on est connecté à internet moins on a de mails non lus

Hypothèse nulle : *Le nombre de mails non lus et le temps de connexion sur internet par jour ne sont pas corrélées.*
Hypothèse alternative : : *Le nombre de mails non lus est corrélé avec le temps de connexion sur internet.*





La normalité n'étant pas respectée :

a- Test Spearman de corrélation:

```
Spearman's rank correlation rho
S = 10715000, p-value = 0.9282
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.004521631
```

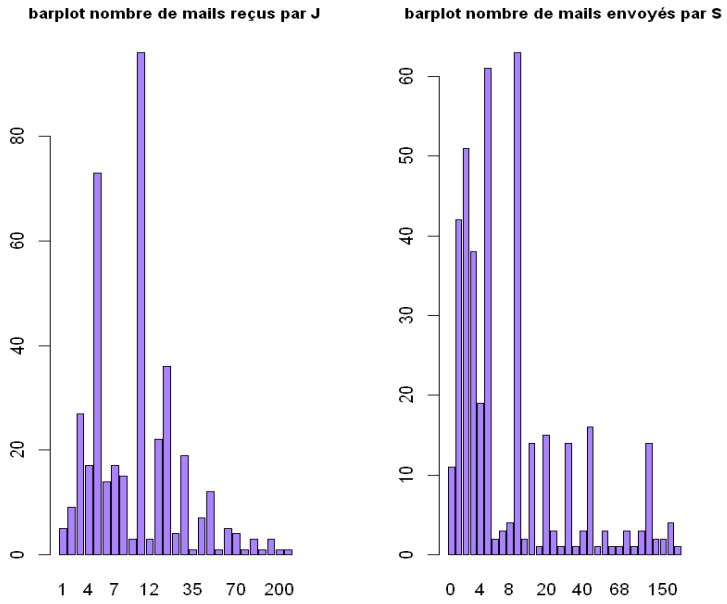
On remarque la P value n'est pas significative. On peut alors déduire qu'il n'existe pas de lien entre le temps passé sur internet et le nombre de mails non lus.

3.2 FLUX DE MAILS : CORRELATION ENTRE LES MAILS ENTRANTS ET SORTANTS

3.2.1 Plus on envoie de mails plus on en reçoit

> ***Hypothèse nulle :*** *Le nombre de mails envoyés par semaine et le nombre de mails reçus par jour ne sont pas corrélées.*
> ***Hypothèse alternative :*** *Le nombre de mails envoyés par semaine est corrélé avec le nombre de mails reçu par jour.*

Nous avons commencé par tracer le barplot de nos deux variables :



Aucune des deux variables ne suit une loi normale. Donc, pour vérifier la corrélation entre les deux, on doit donc utiliser un test de corrélation de Spearman :

a- Test Spearman de corrélation

```
Spearman's rank correlation rho
        S = 4252600, p-value < 2.2e-16
        alternative hypothesis: true rho is not equal to 0
        sample estimates:  rho
                          0.601317
```

La P-value étant significative (2.2e-16), on peut conclure l'existence d'un lien entre le nombre de mails reçus par jour et le nombre de mails envoyés par semaine.

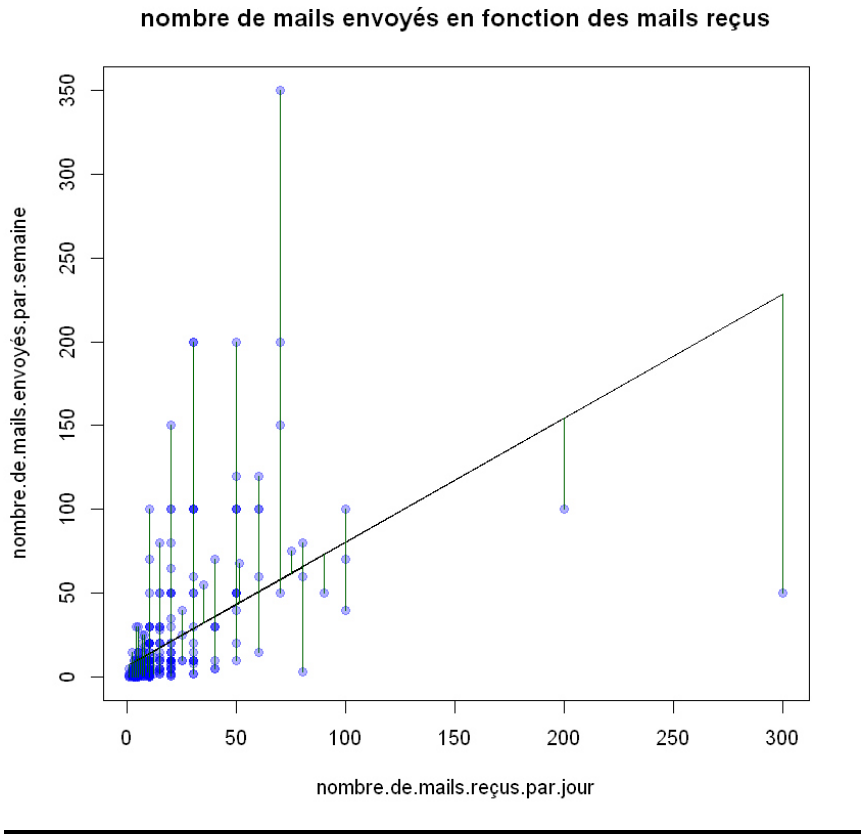
b- Test de Régression linéaire

```
Call:
lm(formula = nombre.de.mails.envoyés.par.semaine ~ nombre.de.mails.reçus.par.jour)

Residuals:
    Min      1Q  Median      3Q     Max
-178.209  -10.021   -7.328   -1.067  291.664

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               6.63515    1.85293   3.581 0.000385 ***
nombre.de.mails.reçus.par.jour  0.73858    0.06456  11.439  < 2e-16 ***
```

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

Residual standard error: 30.86 on 398 degrees of freedom
Multiple R-squared: 0.2474, Adjusted R-squared: 0.2455
F-statistic: 130.9 on 1 and 398 DF, p-value: < 2.2e-16

Comme on a pu constater dans le test de spearman, la P value et la F-statistic de la régression nous confirment que nos variables d'étude sont bien corrélées. Observons graphiquement cette régression linéaire:



Selon cette régression on peut confirmer notre hypothèse de départ : plus on envoie des mails plus on en reçoit.

3.3 IMPORTANCE DE CERTAINES FONCTIONNALITES DANS LA GESTION SON MAIL

3.3.1 Recevoir des notifications pour des nouveaux mails reçus réduit le nombre de mails non lus

Hypothèse nulle : *Le nombre moyen de mails non lus est le même pour ceux qui utilisent l'option notification et d'autres qui sont indifférents.*
Hypothèse alternative : *Le nombre moyen de mails non lus est différent entre les deux échantillons : notifiés de mails reçus / non notifiés.*

a- Description des deux échantillons avec la moyenne et écart type de la variable nombre de mails non lus

	Moyenne	**Ecart type**
Notifiés de mails reçus	**311.8**	**1314**
Non notifiés de mails reçus	**480.5**	**2038.7**

b- Test Shapiro de normalité

- *Notifiés de mails reçus*
  Shapiro-Wilk normality test
  W = 0.23947, p-value < 2.2e-16

Avec un risque de 10%, on rejette l'hypothèse nulle de normalité pour l'échantillon des utilisateurs de la fonctionnalité notification sur les mails reçus.

- *Non notifiés de mails reçus*
  Shapiro-Wilk normality test
  W = 0.24557, p-value < 2.2e-16

Avec le même risque de 10%, on rejette l'hypothèse nulle de normalité pour l'échantillon n'utilisant pas cette fonctionnalité.

c- Test Wilcoxon de moyenne

Wilcoxon rank sum test with continuity correction
W = 19010, p-value = 0.9518

Avec un risque de 10%, on accepte H0 et donc il n'y a pas de différence de mails non lus entre les deux échantillons. Sous réserve de conformité de la représentativité de notre échantillon, on peut conclure que la bonne gestion de la boite mail, reflété par le nombre de mails non lus, est inchangée par rapport à la fonctionnalité de notification. De plus, une autre réserve sur ce résultat est dû à l'option de marquage comme lu pour tous les mails : la variable de nombre de mails non lu est biaisée.

3.3.2 Recevoir des notifications pour des nouveaux mails reçus influe la fréquence de consultation de la boite mails

Hypothèse nulle : *La fréquence de consultation et l'utilisation des notifications sont indépendantes.*
Hypothèse alternative : *La fréquence de consultation dépend de l'utilisation de la fonctionnalité de notification.*

a- Test khi2 de dépendance

Pearson's Chi-squared test
X-squared = 1.3525, df = 3, p-value = 0.7167

Le test de $\chi 2$ nous renvoie une P value non significative. On conclut alors que la fréquence de consultation de la boite mail reste inchangée si on reçoit des notifications pour les nouveaux mails ou non.

H4 Ceux qui personnalisent leurs boites mails envoient plus de mails

Hypothèse nulle : *Le nombre de mails envoyés par semaine est le même entre ceux qui personnalisent leurs boites mails et ceux qui ne le font pas.*
Hypothèse alternative : *Le nombre de mails envoyés par semaine est plus élevé pour ceux qui personnalisent leurs boîtes mails.*

a- Description des deux échantillons avec la moyenne et écart type de la variable nombre de mails envoyés par semaine

	Mean	**Sd**
Personnalisation	*24.5*	*41.7*
Pas de personnalisation	*10.5*	*23.5*

b- Test de normalité sur chaque échantillon

- Personnalisation

Shapiro-Wilk normality test
W = 0.57478, p-value < 2.2e-16

Avec un risque de 10%, on rejette l'hypothèse nulle de normalité pour l'échantillon contenant ceux qui personnalisent leurs boîtes mails.

- Pas de personnalisation

Shapiro-Wilk normality test
W = 0.40099, p-value < 2.2e-16

Avec un risque de 10%, on rejette l'hypothèse nulle de normalité pour l'échantillon contenant ceux qui ne personnalisent pas leurs boîtes mails.

c- Test de moyenne entre les deux échantillons

Wilcoxon rank sum test with continuity correction
W = 13286, p-value = 8.088e-09
alternative hypothesis: true location shift is less than 0

Pour ce test, nous avons choisi l'hypothèse nulle : La moyenne de mails envoyés pour l'échantillon de la non personnalisation est supérieure à celui de personnalisation. P-value est très faible par rapport à 0,05, on rejette l'hypothèse 0 et donc les gens qui personnalisent leurs messageries envoient plus de mails.

4. ACM

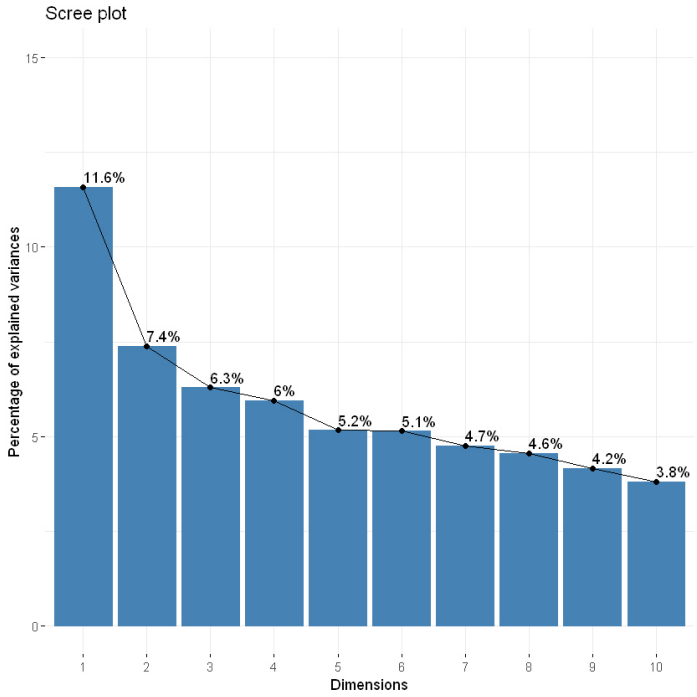
Le jeu de données issu du sondage comporte 14 variables catégorielles et 8 variables quantitatives. De ce fait, une Analyse des Correspondances Multiples serait donc plus adaptée.
Pour que notre analyse soit pertinente, on a choisi les variables et les individus actifs qui vont nous permettre de bien distinguer les profils similaires d'utilisation des mails :

- Variables actives : toutes les variables catégorielles sauf "autres fonctionnalités de la boîte mail" et "lecture l'objet / expéditeur" (seulement 6% ont répondu non donc cette variable n'apporte pas une information supplémentaire)
- Individu actifs : les lignes 1:150
- Variables supplémentaires : "nombre.de.mails.non.lus", "age", "'nombre.de.mails.envoyés.par.semaine'". Les coordonnées de celle-ci seront prédites.

4.1 LES VALEURS PROPRES ET LEURS PROPORTIONS DES VARIANCES

Comme on a plusieurs modalités pour chaque variable, le nombre de dimensions obtenu, à savoir 29, est bien élevé. Ainsi les parts de variances sont faibles et se répartissent sur les modalités des différentes variables. On peut donc se contenter des 10 premières regroupent à peu près 59% de l'inertie totale.

	eigenvalue	variance percent	cumulative variance percent
Dim.1	0.27983411	11.5793427	11.57934
Dim.2	0.17827833	7.3770345	18.95638
Dim.3	0.15209555	6.2936090	25.24999
Dim.4	0.14386352	5.9529732	31.20296
Dim.5	0.12466615	5.1585995	36.36156
Dim.6	0.12433207	5.1447753	41.50633
Dim.7	0.11474386	4.7480217	46.25436
Dim.8	0.10997153	4.5505460	50.80490
Dim.9	0.10056007	4.1611062	54.96601
Dim.10	0.09178247	3.7978955	58.76390

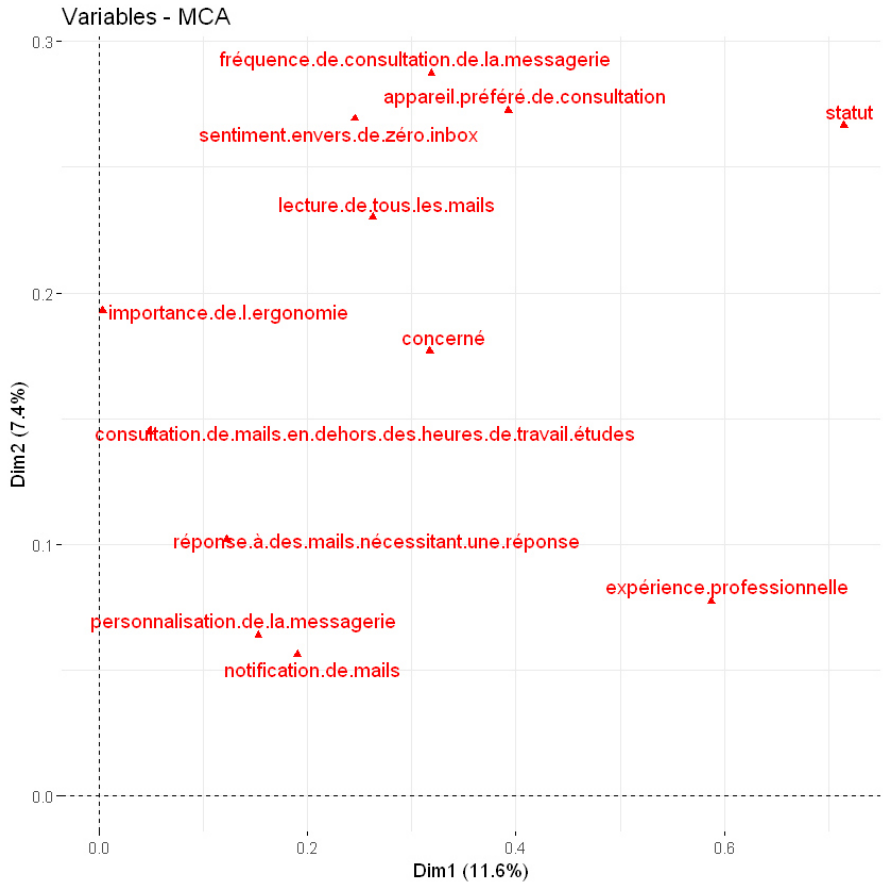


Pourcentage de variances expliquées

4.2 CORRELATION ENTRE LES VARIABLES ET LES AXES PRINCIPAUX

Le graphique ci-dessus permet d'identifier les variables les plus corrélées avec chaque axe. Les corrélations au carré entre les variables et les axes sont utilisés comme coordonnées.

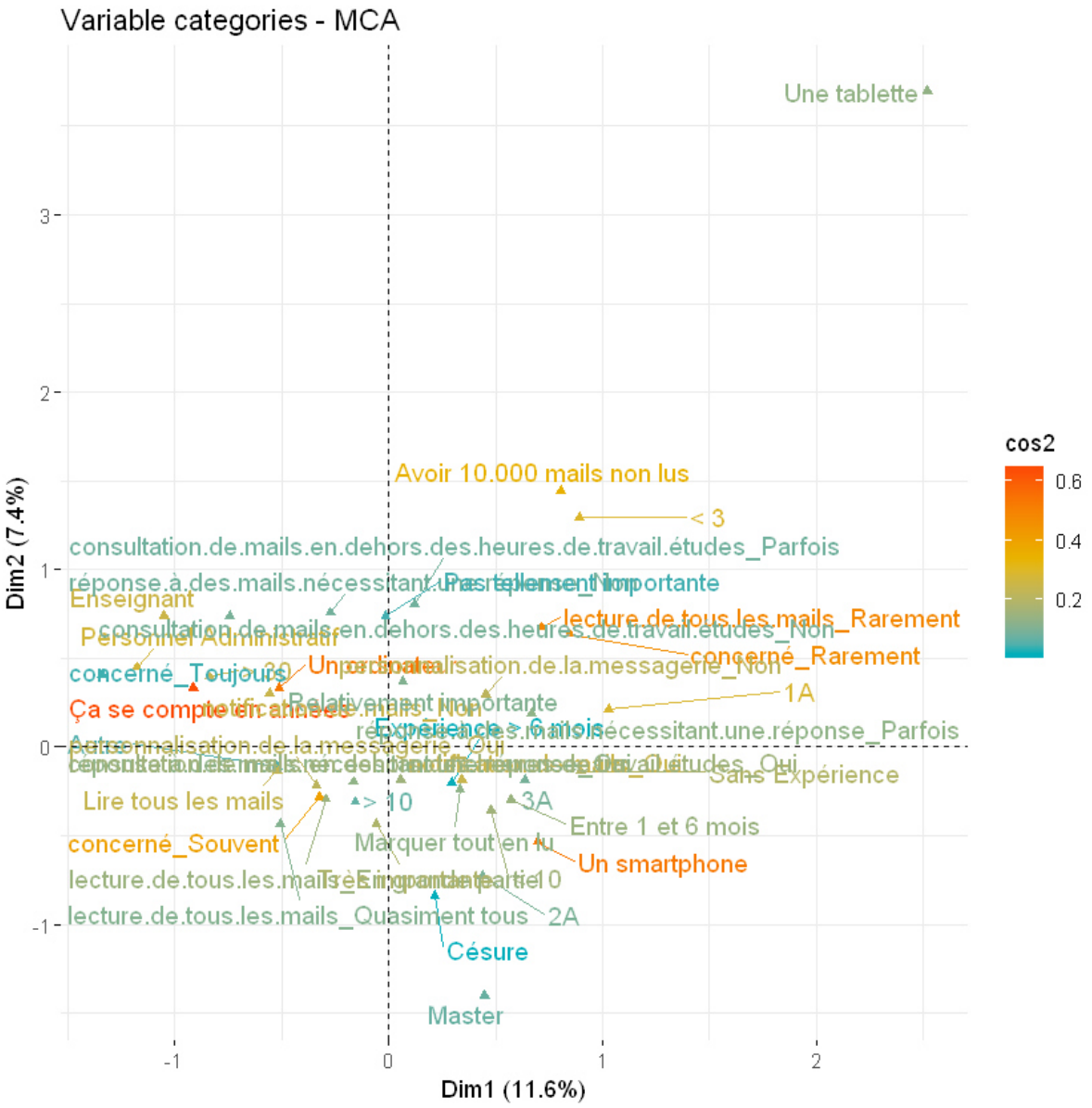
On constate que les variables "expérience professionnelle" et "statut" sont les plus corrélées avec la dimension 1. De même, les variables "importance de l'ergonomie", "fréquence de consultation de la messagerie" et "consultation en dehors des heures de travails/études" sont les plus corrélées avec la dimension 2.



4.3 CATEGORIES DES VARIABLES

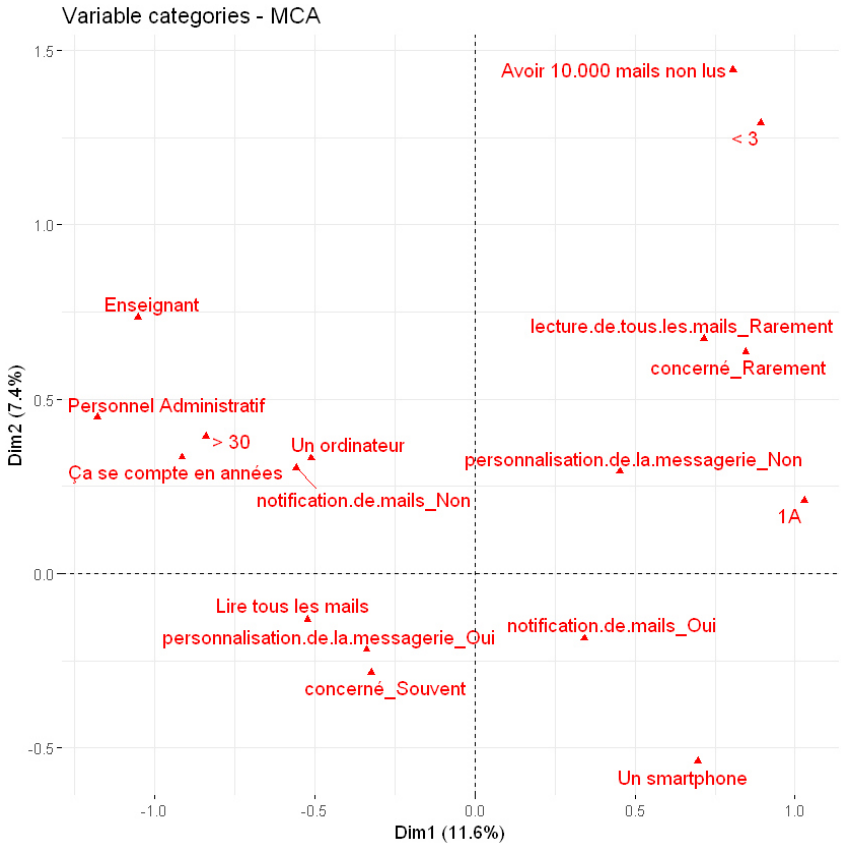
Le graphique ci-dessus montre les relations entre les catégories des variables. Il peut être interprété comme suit :

- Les catégories avec un profil similaire sont regroupées.
- Les catégories corrélées négativement sont positionnées sur les côtés opposés de l'origine du graphique.
- Tous les points ne sont pas aussi bien représentés par les deux dimensions.
- La qualité de représentation, appelée cosinus carré (cos2), mesure le degré d'association/contribution entre les catégories des variables et les dimensions.

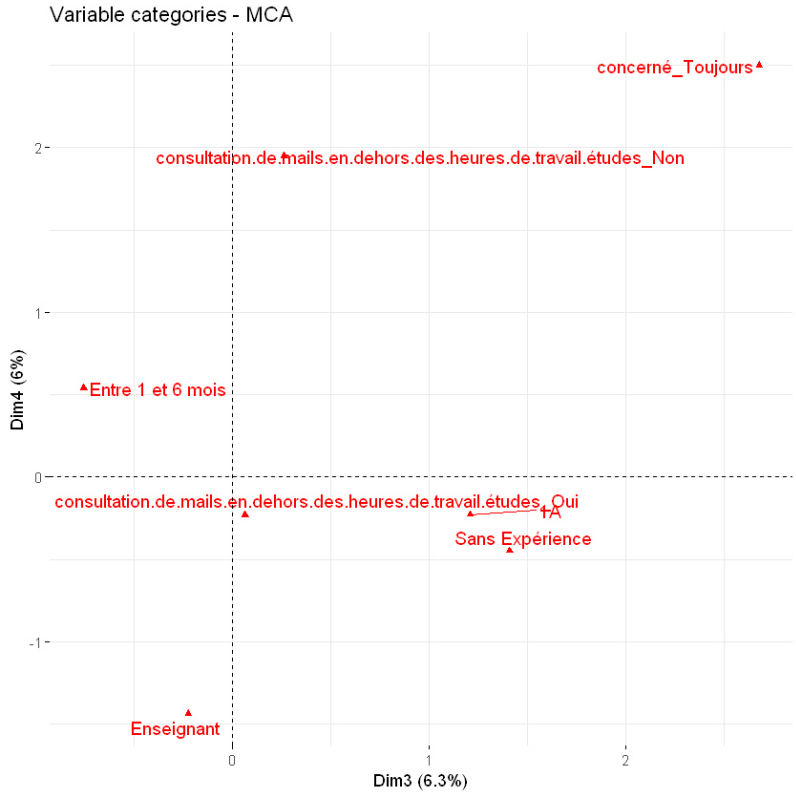


→ Dim 1 : On voit bien sur ce graphique que les réalisations négatives sur la dimension correspondent à des profils de personnes qui sont plus responsables que le reste de la population : lecture de tous les mails majoritairement sur un ordinateur, souvent concerné par les mails qu'ils reçoivent, une bonne expérience professionnelle. Ceux sont généralement des enseignants ou des personnels administratifs. Cela se confirment avec les réalisations positivent qui décrivent un profil jeune et n'aime pas lire tous ses mails.

On pourra bien analyser et caractériser la dimension 2 en ne gardant que les catégories bien représenté (cos2 > 0.2) :



→ Dim 2 : La dimension 2 décrit l'expérience utilisateur et sa relation avec sa messagerie : personnalisation de la messagerie, activation des notifications, tendance à lire tous les mails et se sentir concerné par les mails reçus.



En utilisant la dimension dimdesc(), on peut identifier les variables et les catégories les plus corrélés avec les dimensions 3 et 4. Le résultat obtenu correspond aux résultats d'un test Fisher pour voir si la variable a un effet significatif sur la dimension et aux tests de Student réalisés modalité par modalité.
Ceci nous permet de confirmer que :

- La dimension 3 est décrite par un profil débutant dans l'utilisation des mails au quotidien (étudiant avec une expérience professionnelle de quelques mois ou absente)
- La dimension 4 est constitués de deux pôles d'individus qui travaillent chez l'IMT Atlantique : l'enseignant et le personnel administratif qui se comportent d'une manière différente vis-à-vis de la lecture des mails en dehors des heures de travail.

$`Dim 3`
$quali

	R2	p.value
expérience.professionnelle	**0.58833090**	**5.435142e-28**
statut	**0.58927083**	**1.354333e-24**
importance.de.l.ergonomie	**0.18305418**	**3.517257e-07**
concerné	**0.14677237**	**8.575724e-06**
lecture.de.tous.les.mails	**0.11250279**	**1.549795e-04**
appareil.préféré.de.consultation	**0.07955669**	**2.258306e-03**
notification.de.mails	**0.04811808**	**6.995980e-03**
consultation.de.mails.en.dehors.des.heures.de.travail.études	**0.05699396**	**1.339102e-02**

$category

	Estimate	p.value
Sans Expérience	**0.54239530**	**1.458397e-22**
1A	**0.49762794**	**4.079190e-17**
concerné_Toujours	**0.71230005**	**1.315503e-06**
Relativement importante	**0.27206715**	**3.226418e-04**
lecture.de.tous.les.mails_Quasiment tous	**0.17978845**	**1.275762e-03**
notification.de.mails_Non	**0.08812412**	**6.995980e-03**
Une tablette	**0.69218838**	**7.006527e-03**
Personnel Administratif	**0.23169277**	**7.194144e-03**
Un smartphone	**-0.27552535**	**4.753236e-02**
Un ordinateur	**-0.41666303**	**1.519191e-02**
notification.de.mails_Oui	**-0.08812412**	**6.995980e-03**
consultation.de.mails.en.dehors.des.heures.de.travail.études_Parfois	**-0.21194668**	**4.167666e-03**
Expérience > 6 mois	**-0.24517256**	**1.821156e-03**

Autre	-0.23106344	3.568248e-04
lecture.de.tous.les.mails_Rarement	-0.18325012	3.245278e-04
Pas tellement importante	-0.37898130	2.061930e-06
3A	-0.31813124	2.924961e-07
Entre 1 et 6 mois	-0.30277292	8.526682e-09

$`Dim 4`

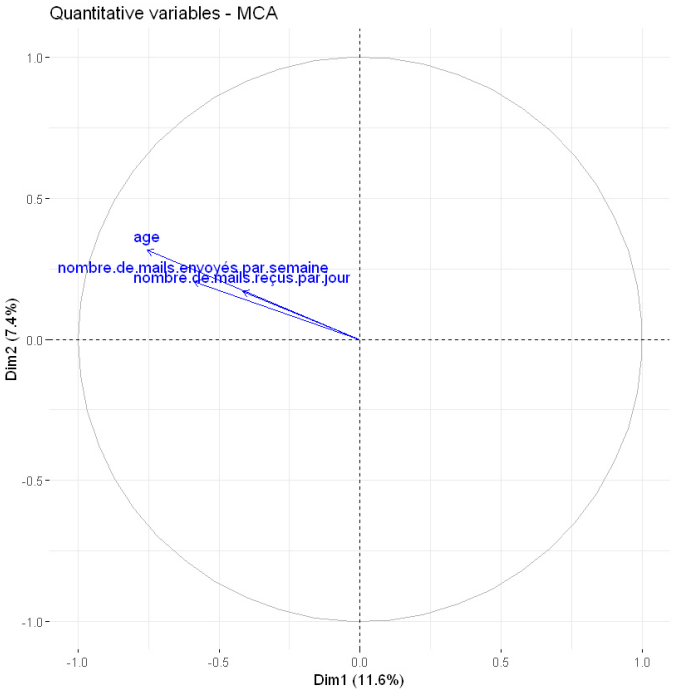
$quali

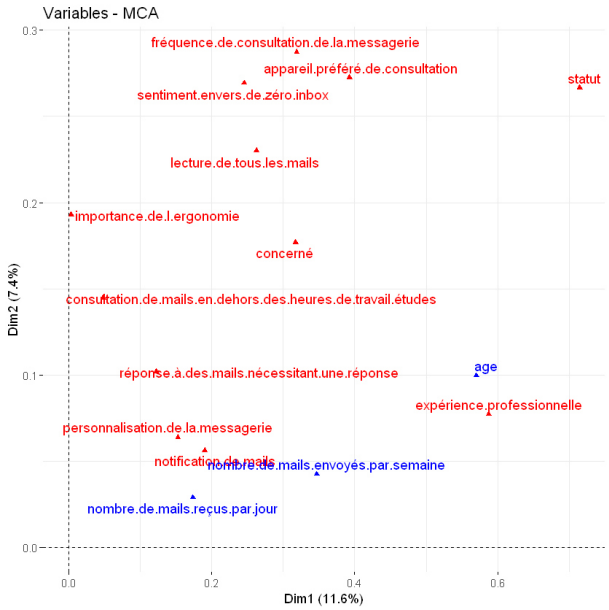
	R2	p.value
statut	0.53607034	6.126357e-21
consultation.de.mails.en.dehors.des.heures.de.travail.études	0.35580753	9.175926e-15
concerné	0.16481780	1.782051e-06
lecture.de.tous.les.mails	0.14920272	6.953728e-06
expérience.professionnelle	0.14441689	4.342478e-05
réponse.à.des.mails.nécessitant.une.réponse	0.12741416	4.460667e-05
fréquence.de.consultation.de.la.messagerie	0.08800764	3.678637e-03
sentiment.envers.de.zéro.inbox	0.05422214	1.661506e-02
importance.de.l.ergonomie	0.04991207	2.320808e-02

$category

	Estimate	p.value
consultation.de.mails.en.dehors.des.heures.de.travail.études_Non	0.48709631	1.494365e-14
Personnel Administratif	0.31616988	1.962065e-06
concerné_Toujours	0.62324869	7.290709e-06
lecture.de.tous.les.mails_Quasiment tous	0.20389882	2.500930e-05
Entre 1 et 6 mois	0.19439091	6.240560e-05
Césure	0.61445483	5.015894e-04
réponse.à.des.mails.nécessitant.une.réponse_Parfois	0.22412345	2.130897e-03
< 3	0.18564984	1.439085e-02
< 10	0.06235985	4.017363e-02
> 10	-0.12052964	3.223809e-02
Master	-0.39039967	3.025415e-02
Ça se compte en années	-0.09687884	2.342478e-02
Pas tellement importante	-0.20732528	9.026728e-03
Sans Expérience	-0.18058499	7.100833e-03
Avoir 10.000 mails non lus	-0.17832190	5.850889e-03
concerné_Souvent	-0.39245566	1.078880e-03
réponse.à.des.mails.nécessitant.une.réponse_Non	-0.26422015	2.023564e-04
lecture.de.tous.les.mails_En grande partie	-0.18326002	5.052399e-05
consultation.de.mails.en.dehors.des.heures.de.travail.études_Oui	-0.33995548	8.961956e-10
Enseignant	-0.57272977	6.918176e-14

4.4 VARIABLES SUPPLEMENTAIRES





On voit bien sur ce graphique que l'âge est négativement corrélé avec la dimension 1. Ce qui est bien logique étant donné l'âge est un facteur participant au sens de responsabilité et maturité de la personne. Le nombre de mails reçus et envoyés par semaine contribue également à cette dimension mais d'une façon moins importante que l'âge.

5. DISCUSSION CRITIQUE (CE QUI AURAIT PU ETRE AMELIORE)

Avec un peu de recul sur le projet, on réalise que certaines choses auraient pu être améliorées dans notre manière de procéder, notamment en ce qui concerne le questionnaire en lui-même.
Certaines formulations des questions ou des réponses n'étaient pas suffisamment claires ce qui a diminué la force des interprétations. Par exemple dans nos catégories de statut dans l'école des individus sondés, il était difficile de savoir à quoi correspondait la catégorie "autre" (des doctorants peut être) et d'en tirer des conclusions.

Il faut également rester critique sur la représentativité de notre échantillon. En effet, un biais a pu être introduit par le fait que nous avons sondés via un mail alors que nous nous intéressons notamment aux gens qui ne lisent pas leurs mails (et qui ont donc moins tendance à répondre).

6. CONCLUSION

Le but de cette étude était d'investiguer plusieurs aspects de la gestion d'une boite mail et du traitement des messages. A travers un questionnaire conçu autour des principaux aspects de la problématique, nous avons pu collecter des informations variées sur notre échantillon. Ensuite, grâce à des tests statistiques, nous avons fait des conclusions plus générales avec un risque de + -10%.

Les données collectées sur un individu permettaient de décrire son profil, son comportement digital, le flux de mails dans sa boite et sa manière de gérer sa messagerie. Ces données, qualitatives et quantitatives, nous ont amené à répondre aux hypothèses que nous avons énoncées au début et en cours de notre étude.

Dans ce document, nous avons détaillé les méthodes que nous avons utilisées afin de tester nos hypothèses aussi bien que les résultats obtenus. Finalement, voici un récapitulatif de certains clichés que nous avons réussi confirmer ou infirmer :

L'e-mailing et le profil :

- Plus on a d'expérience professionnelle, plus on envoie des mails. Hypothèse 3.1.1
- La tendance à lire seulement l'objet et l'expéditeur des mails ou à lire la totalité de ses messages est indépendante de l'appareil de préférence (PC ou Smartphone). Hypothèse 3.1.2 et Hypothèse 3.1.3
- Le personnel et les enseignants portent plus attention à la lecture de leurs mails que les jeunes étudiants. Hypothèse 3.1.4
- La fréquence de consultation devient plus importante avec l'expérience pro. Hypothèse 3.1.5
- La fréquence de consultation varie avec l'âge. Hypothèse 3.1.6

Le flux de mails :

- Plus on envoie des mails plus on en reçoit. Hypothèse 3.2

L'importance de certaines fonctionnalités de la boite mail :

- L'option de notification n'affecte pas la réactivité de l'individu : les mails non lus et la fréquence de consultation : Hypothèse 3.3.1 et Hypothèse 3.3.2

La personnalisation de la boite mail rend plus réactif avec une fréquence de consultation plus élevé et plus de mails envoyés. Hypothèse 3.3.3 et Hypothèse 3.3.4.

L'ACM nous a permis de conclure que la lecture de tous les mails dépend principalement de son sens de responsabilité et de la maturité de son expérience d'utilisation de sa boite mail.

7. BIBLIOGRAPHIE

Tests statistiques : http://www.sthda.com/french/wiki/tests-statistiques-avec-r , consulté le 15-16-17 novembre 18. [1]

R Basic Statistics - Easy Guides - Wiki - STHDA. (s.d.). Récupéré 17 novembre, 2018, de : http://www.sthda.com/english/wiki/r-basic-statistics . [2]

[2] MCA - Multiple Correspondence Analysis in R : Essentials - Articles - STHDA. (2017, 24 septembre). Récupéré 17 novembre, 2018, de : http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/114-mca-multiple-correspondence-analysis-in-r-essentials/ [3]

8. ANNEXES

Lisez-vous vos mails ?

Le but du projet est d'étudier, d'un point de vue statistique, le comportement vis-à-vis des mails à l'IMT Atlantique.
Merci d'avance pour vos réponses.

*Obligatoire

Informations Générales

1. **Quel est votre statut au sein de l'école ?** *
   Une seule réponse possible.

   - ◯ Personnel Administratif
   - ◯ Enseignant
   - ◯ 1A
   - ◯ 2A
   - ◯ 3A
   - ◯ Césure
   - ◯ Master
   - ◯ Autre

2. **Quel âge avez-vous ?** *
   Veuillez saisir des chiffres.

Expérience professionnelle

3. **Quel est votre niveau d'expérience professionnelle ?** *
Une seule réponse possible.

- ◯ Sans Expérience
- ◯ Entre 1 et 6 mois
- ◯ Expérience > 6 mois
- ◯ Ça se compte en années

Addiction à Internet

4. **Combien d'heures par jour êtes-vous actifs sur internet ?** *
En moyenne, un français passe 4h48min à surfer sur le net (Janvier 2018). Saisissez votre estimation en chiffres.

_____

5. **Pour consulter vos mail, vous utilisez le plus souvent :** *
Une seule réponse possible.

- ◯ Un ordinateur
- ◯ Un smartphone
- ◯ Une tablette
- ◯ Une montre connectée

Traitement des mails

6. **Combien avez-vous de boites mails ?** *
Veuillez saisir un nombre.

_____

7. **Combien en consultez-vous régulièrement ?** *
   Veuillez saisir un nombre.

8. **Combien de fois par jour consultez-vous vos mails ?** *
   Une seule réponse possible.

   - [ ] < 3
   - [ ] < 10
   - [ ] > 10
   - [ ] > 30

9. **En moyenne, combien de mails recevez-vous par jour sur votre adresse mail principale ?** *
   Une estimation numérique.

10. **Vous sentez-vous concerné par les mails que vous recevez ?** *
    Une seule réponse possible.

    - [ ] Toujours
    - [ ] Souvent
    - [ ] Rarement
    - [ ] Jamais

11. **Recevez-vous des notifications pour vos mails ?** *
    Une seule réponse possible.

    - [ ] Oui
    - [ ] Non

12. **Lisez-vous tous vos mails intégralement ?** *
Une seule réponse possible.

- ◯ Quasiment tous
- ◯ En grande partie
- ◯ Rarement

13. **Lisez-vous au moins l'objet et l'expéditeur pour chaque mail ?** *
Une seule réponse possible.

- ◯ Oui
- ◯ Non

14. **Combien estimez-vous avoir de mails non lus dans votre boîte de réception ? (si vous êtes courageux, allez vérifier sur Zimbra)** *
Veuillez saisir une estimation.

_____

15. **Combien de mails envoyez-vous par semaine ?** *
Veuillez saisir une estimation numérique.

_____

16. **Répondez-vous toujours aux mails qui nécessitent une réponse ?** *
Une seule réponse possible.

- ◯ Oui
- ◯ Non
- ◯ Parfois

OUR WORLDWIDE PARTNERS UNIVERSITIES - DOUBLE DEGREE AGREEMENTS

3 CAMPUS, 1 SITE

IMT Atlantique Bretagne–Pays de la Loire – **http://www.imt-atlantique.fr/**

Campus de Brest
Technopôle Brest-Iroise
CS 83818
29238 Brest Cedex 3
France
T +33 (0)2 29 00 11 11
F +33 (0)2 29 00 10 00

Campus de Nantes
4, rue Alfred Kastler
CS 20722
44307 Nantes Cedex 3
France
T +33 (0)2 51 85 81 00
F +33 (0)2 99 12 70 08

Campus de Rennes
2, rue de la Châtaigneraie
CS 17607
35576 Cesson Sévigné Cedex
France
T +33 (0)2 99 12 70 00
F +33 (0)2 51 85 81 99

Site de Toulouse
10, avenue Édouard Belin
BP 44004
31028 Toulouse Cedex 04
France
T +33 (0)5 61 33 83 65

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Imprimé à IMT Atlantique
Dépôt légal : Mois Année
ISSN 2556-5060