Projet Statistiques

Rapport

F3B101E

S5

Année scolaire 2016-2017

Anas IRHBOULA

Ilham LAASSAIRI

Jose Alcides OTINIANO REGALADO

Khalil IBRAHIMI

Nabil RATBI

TELECOM Bretagne

Sommaire

1. INTRODUCTION

1.1 CONTEXTE

Dans un domaine où le travail en groupe est indispensable et où les résultats dépendent essentiellement de la synchronisation des idées et des tâches au sein de l'équipe, il est pertinent de chercher à savoir quels sont, au sein d'une population, les critères qui permettent un regroupement optimal des individus pour atteindre les meilleurs résultats possibles. En outre, le but du projet est de déterminer pour une combinaison donnée d'individus, si le groupe est susceptible d'apporter des résultats satisfaisants ou non.

1.2 BUT DU PROJET

L'analyse des résultats de l'enquête a pour but de définir des critères de performances permettant de prédire le succès ou l'échec d'un groupe d'étudiants sur un projet lambda.

2. MÉTHODE D'ÉCHANTILLONNAGE :

Le traitement du sujet a commencé sans aucune connaissance au préalable de la population. Nous avons opté pour un ciblage des étudiant de Télécom bretagne car étant les plus facile d'accès pour nous et dont nous pouvions avoir le plus de réponses.

Le recueil de données a donc été fait grâce à un questionnaire pour multiples raisons :

- Nous avions une bonne connaissance du sujet de l'étude.

-Comme cité avant nous cherchions à valider ou à généraliser des résultats.

2.1 QUESTIONS :

Pour construire tout questionnaire, nous avons commencé par identifier des dimensions pour étendre le concept :

- Informations générales sur le répondeur.
- Informations sur le cursus.
- Informations sur le projet choisi.
- Informations sur le groupe de projet.
- Informations sur le succès du projet.

Nous avons opté pour des questions fermées pour la plupart du questionnaire et quelques questions ouvertes là où l'implication du répondeur nous paraissait nécessaire.

Finalement, après consultation des tuteurs et experts en statistiques, nos questions ne posaien't aucun biais ou mal formulation poussant à une mal compréhension.

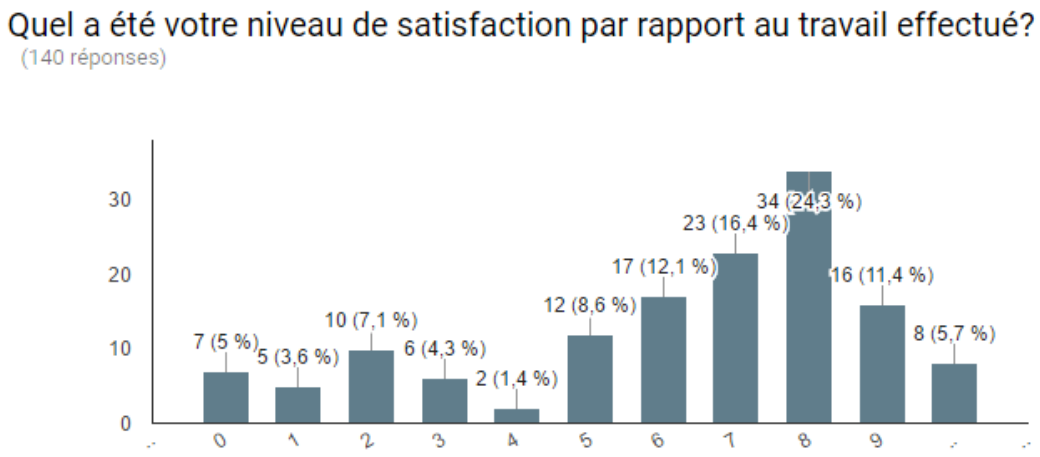
2.2 ÉCHANTILLONNAGE :

A l'inverse d'un recensement, le sondage auquel on a procédé ne vise qu'une petite partie qu'on pense représentative de la population cible. Ceci a été fait par volontariat mais nous pensons, basés sur une comparaison de la proportion des filles et nationalités obtenues avec les proportions réelles, que notre échantillon est plus ou moins effectivement représentatif.

Le problème de saturation est souvent posé dans ce cas de sondage, mais nous en étions bien loin. Nous avions un problème bien plus grave, celui du manque d'information.

Effectivement, avec 140 réponses au questionnaire dont seulement 9 (7,7%) en dessous de la note nécessaire pour la validation (notre critère initial de succès). Ce nombre n'était pas très encourageant étant donné que notre but était de dégager les critères d'échec/succès dans un groupe.

Nous avons donc pensé à utiliser le niveau de satisfaction à la place. Ce dernier était beaucoup plus équilibré :

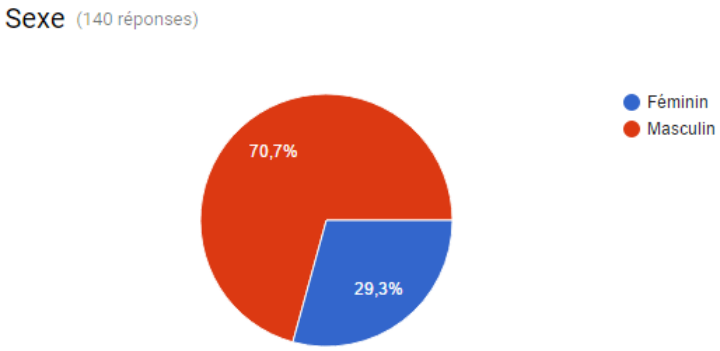


2.3 MÉTHODE DES QUOTAS :

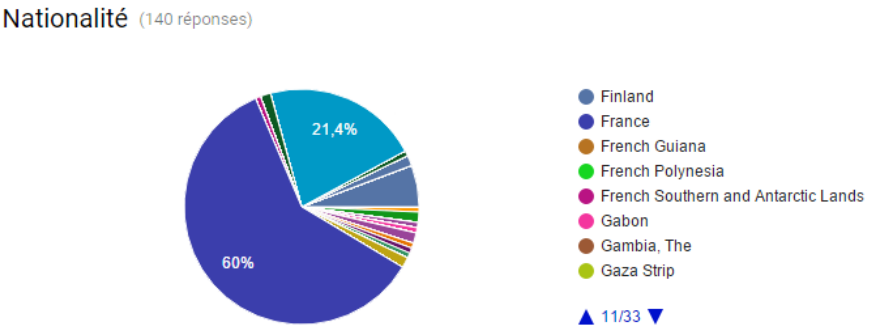
Pour mieux préciser notre méthode d'échantillonnage et s'assurer de sa représentativité :

Comme mentionné précédemment, la répartition des filles dans l'échantillon ainsi que la répartition des nationalités était similaire à la proportion réelle :

Représentation réelle des filles : 25 % (http://www.letudiant.fr/palmares/palmares-des-ecoles-d-ingenieurs/telecom-bretagne-brest.html)



Pourcentage des étudiants étrangers : 39.1 % (http://www.letudiant.fr/palmares/palmares-des-ecoles-d-ingenieurs/telecom-bretagne-brest.html)



3. TEST STATISTIQUES

3.1 NORMALITÉ DES DONNÉES

Dans un premier temps nous commençons par tester la normalité des données. La contrainte de la loi normale est forte car elle nécessite de vérifier les conditions d'application du test.

Les tests paramétriques sont globalement robustes, c'est-à-dire que leurs conclusions restent valables même en cas de faible écart aux conditions d'application.

Notre attention étant rivée sur la variable Satisfaction, c'est sur cette variable que s'effectuera le test de la condition de normalité :

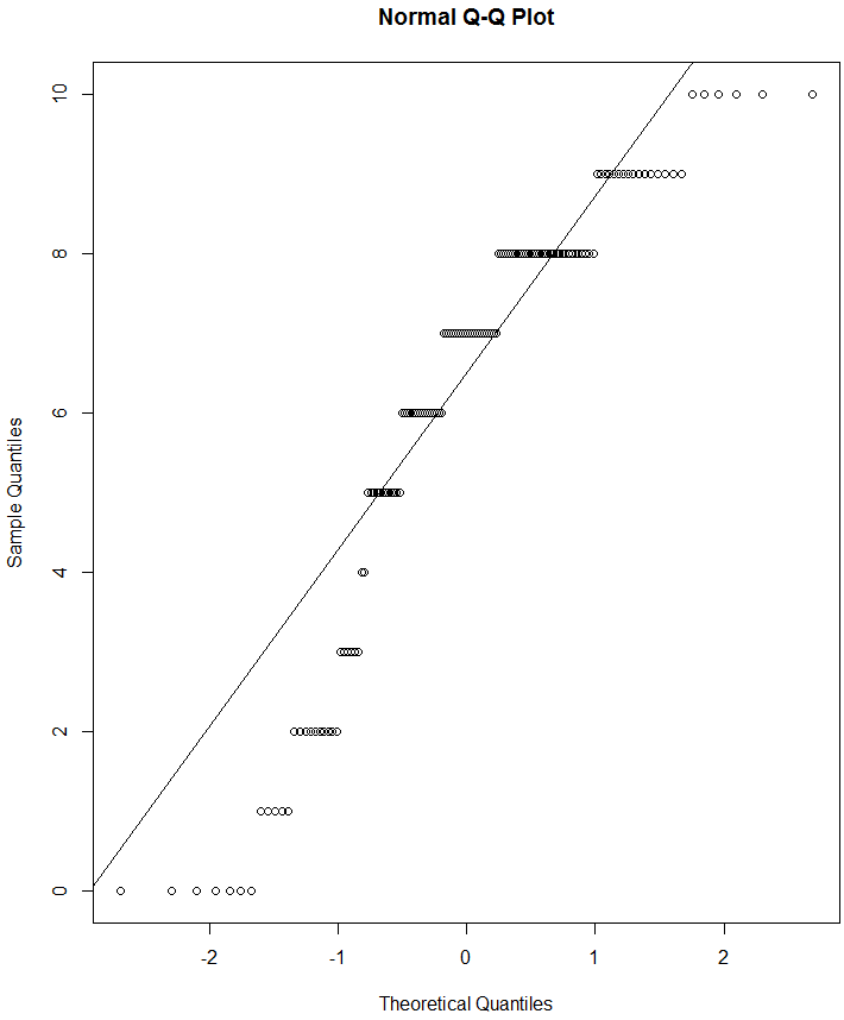
```
> shapiro.test(Stats_Renamed$Satisfaction)

        Shapiro-Wilk normality test

data:  Stats_Renamed$Satisfaction
W = 0.88916, p-value = 1.008e-08
```

Le test shapiro permet de rejeter l'hypothèse selon laquelle l'échantillon suit la loi normal.

Pour plus de certitude, le Q-Q plot suivant confirme la non-normalité :



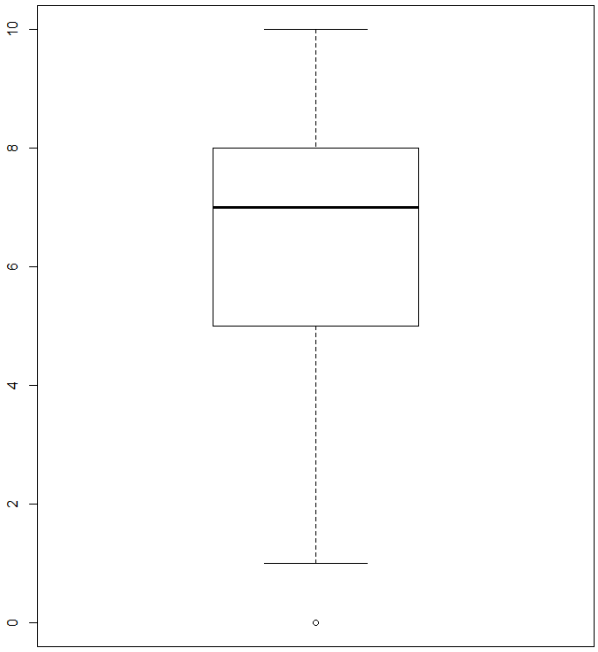
Q-Q plot de la variable Satisfaction

Identification des Outliers

*Faute d'utiliser un test outliers de Walsh on utilise une simple fonction se basant sur le methode de Hampel : (+/- 5.2 * Median Absolute Deviation)*

```
hampel.outlier = function (x) {
    x <- na.omit(x)
    lims <- median(x) + c(-1, 1) * 5.2 * mad(x, constant = 1)
    x < lims[1] | x > lims[2]
}
```

Nombre des outliers : 12



Boîte à moustaches de la Satisfaction

On se dirige donc vers des tests non paramétriques.

3.2 TESTS KHI2

Nous allons étudier ci dessous l'influence de plusieurs éléments de notre questionnaire sur de la note de projet d'un groupe quelconque .
Question 20 : Dans quelle tranche se trouve la note que vous avez obtenu dans votre projet ?
Nous avons réparti les notes en 5 tranches :

- 0 à 4,99
- 5 à 8,99
- 9 à 9,99
- 10 à 14,99
- 15 à 20

Pour la suite de notre étude, nous avons décidé de regrouper les notes en deux catégories comme indiqué sur le tableau ci dessous.

Hypothèse1 : La note d'un projet ne dépend pas du nombre d'année passé au sein de Télécom Bretagne.

Note du projet	Une année	Deux années	Trois années	Total
0 à 9,99	2	2	5	9
10 à 20	18	40	48	106
Total	20	42	53	115

Sur R, le test de Khi deux donne une valeur de p-value= 0.6477 et avec un risque de 5% , on accepte l'hypothèse pour le nombre d'années à Télécom Bretagne.

```
chisq.test(Note,AnneesTB, simulate.p.value = TRUE)

        Pearson's Chi-squared test with simulated p-value (based on 2000
        replicates)

data:  Note and AnneesTB
X-squared = 0.86764, df = NA, p-value = 0.6477
```

Hypothèse 2 : La note d'un projet dépend de l'année de césure.

Note du projet	Sans césure	Avec césure	Total
0 à 9,99	7	2	9
10 à 20	78	28	106
Total	85	30	115

Sur R, le test de Khi deux donne une valeur de p-value= 1 et avec un risque de 5% , on accepte l'hypothèse pour l'année de césure.

```
chisq.test(Note,Cesure, simulate.p.value = TRUE)

        Pearson's Chi-squared test with simulated p-value (based on 2000
        replicates)

data:  Note and Cesure
X-squared = 0.075636, df = NA, p-value = 1
```

Hypothèse 3 : La note d'un projet dépend du nombre de projets déjà effectués.

Note du projet	1 projet	2 projets	3 projets	4 projets ou plus	Total
0 à 9,99	2	1	0	6	9
10 à 20	13	4	18	71	106
Total	15	5	18	77	115

Sur R, le test de Khi deux donne une valeur de p-value= 0.3018 et avec un risque de 5% , on accepte l'hypothèse pour le nombre de projets effectués.

```
chisq.test(Note,NombreProjets, simulate.p.value = TRUE)

        Pearson's Chi-squared test with simulated p-value (based on 2000
        replicates)

data:  Note and NombreProjets
X-squared = 3.1864, df = NA, p-value = 0.3018
```

Hypothèse 4 : La note d'un projet dépend du projet choisi dans le questionnaire.

On peut numéroter les noms des projets figurés dans notre questionnaire de la manière suivante :

1. Projet Découverte
2. Projet Développement
3. Projet Innovation
4. Projet Ingénieur
5. Projet S5
6. Projet Développement Durable
7. Autres

Note du projet	1	2	3	7	5	6	7	Total
0 à 9,99	3	1	1	3	1	0	0	9
10 à 20	13	40	16	22	5	5	5	106
Total	16	41	17	25	6	5	5	115

Sur R, le test de Khi deux donne une valeur de p-value= 0.3278 et avec un risque de 5% , on accepte l'hypothèse pour le projet choisi dans le questionnaire.

```
chisq.test(Note, Projet, simulate.p.value = TRUE)

        Pearson's Chi-squared test with simulated p-value (based on 2000
        replicates)

data:  Note and Projet
X-squared = 6.4882, df = NA, p-value = 0.3278
```

Hypothèse 5 : La note d'un projet dépend de l'effectif des membres du groupe.

Note du projet	2	3	4	5	6	7	8	9	10 ou plus	Total
0 à 9,99	0	1	1	3	0	1	2	1	0	9
10 à 20	11	22	21	17	4	2	23	6	0	106
Total	11	23	22	20	4	3	25	7	0	115

Sur R, le test de Khi deux donne une valeur de p-value= 0.4328 et avec un risque de 5% , on accepte l'hypothèse pour l'effectif des membres du groupe.

```
chisq.test(Note, Effectif, simulate.p.value = TRUE)

    Pearson's Chi-squared test with simulated p-value (based on 2000
        replicates)

data:  Note and Effectif
X-squared = 6.5262, df = NA, p-value = 0.4328
```

Hypothèse 6 : La note d'un projet dépend de l'effet de la barrière de la langue dans le groupe.

Note du projet	0	1	2	3	4	5	Total
0 à 9,99	2	2	0	3	0	2	9
10 à 20	55	26	12	7	5	1	106
Total	57	28	12	10	5	3	115

Sur R, le test de Khi deux donne une valeur de p-value= 0.002499 et avec un risque de 5% , on rejette l'hypothèse pour la barrière de la langue dans le groupe.

```
chisq.test(Note, LangageBarriere, simulate.p.value = TRUE)
```

```
        Pearson's Chi-squared test with simulated p-value (based on 2000
        replicates)

data:  Note and LangageBarriere
X-squared = 24.149, df = NA, p-value = 0.002499
```

Hypothèse 7 : La note d'un projet dépend du nombre de nationalité dans le groupe.

Note du projet	1	2	3	4	5	6	Total
0 à 9,99	0	2	2	3	2	0	9
10 à 20	14	38	28	17	8	1	106
Total	14	40	30	20	10	1	115

Sur R, le test de Khi deux donne une valeur de p-value= 0.3523 et avec un risque de 5% , on accepte l'hypothèse pour le nombre de nationalité dans le groupe de projet

```
chisq.test(Note,Diversite, simulate.p.value = TRUE)
```

```
        Pearson's Chi-squared test with simulated p-value (based on 2000
        replicates)

data:  Note and Diversite
X-squared = 5.2538, df = NA, p-value = 0.3523
```

Hypothèse 8 : La note d'un projet dépend de l'implication des membres du groupe.

Note du projet	1	2	3	4	5	Total
0 à 9,99	1	2	5	1	0	9
10 à 20	7	14	26	37	22	106
Total	8	16	31	38	22	115

Sur R, le test de Khi deux donne une valeur de p-value= 0.1199 et avec un risque de 5% , on accepte l'hypothèse pour l'implication du groupe dans le projet.

```
chisq.test(Note,Implication, simulate.p.value = TRUE)
```

```
        Pearson's Chi-squared test with simulated p-value (based on 2000
        replicates)

data:  Note and Implication
```

```
X-squared = 6.9787, df = NA, p-value = 0.1199
```

Hypothèse 9 : La note d'un projet dépend du nombre de filles dans le groupe.

A partir des résultats du questionnaire pour le nombre de filles dans le groupe de projet choisi, nous avons calculé le pourcentage de filles dans chaque groupe. Nous avons classer ensuite les données en pourcentage selon quatre catégories.

Note du projet	0 à 15%	15% à 30%	30% à 45%	45% à 100 %	Total
0 à 9,99	2	4	2	1	9
10 à 20	35	40	20	10	105
Total	37	44	22	11	114

Sur R, le test de Khi deux donne une valeur de p-value= 0.9115 et avec un risque de 5% , on accepte l'hypothèse H0.

```
chisq.test(Note,Filles, simulate.p.value = TRUE)

Pearson's Chi-squared test with simulated p-value (based on 2000
replicates)
data: D
X-squared = 0.46683, df = NA, p-value = 0.9115
```

Hypothèse 10 : La note d'un projet dépend du choix du groupe.

Note du projet	Non choix du groupe	Choix du groupe	Total
0 à 9,99	9	0	9
10 à 20	72	34	106
Total	81	34	115

Sur R, le test de Khi deux donne une valeur de p-value= 0.06497 et avec un risque de 5% , on accepte l'hypothèse pour le choix du groupe.

```
chisq.test(Note, ChoisiGroupe, simulate.p.value = TRUE)

        Pearson's Chi-squared test with simulated p-value (based on 2000
        replicates)

data:  Note and ChoisiGroupe
X-squared = 4.0985, df = NA, p-value = 0.06497
```

Hypothèse 11 : La note d'un projet dépend du choix du sujet.

Note du projet	Non choix du sujet	Choix du sujet	Je ne me rappelle pas	Total
0 à 9,99	7	2	0	9
10 à 20	35	68	3	106
Total	42	70	3	115

Sur R, le test de Khi deux donne une valeur de p-value= 0.03298 et avec un risque de 5% , on rejette l'hypothèse pour le choix du sujet.

```
chisq.test(Note, ChoisiSujet, simulate.p.value = TRUE)
```

```
        Pearson's Chi-squared test with simulated p-value (based on 2000
        replicates)

data:  Note and ChoisiSujet
X-squared = 7.2011, df = NA, p-value = 0.03298
```

Hypothèse 12 : La note d'un projet dépend de certains critères de choix précis .

On peut numéroter les résultats des choix obtenus de la manière suivante :

1. Affinités personnelles
2. Proposition de la part d'un groupe préconçu
3. Compétences des membres du groupe
4. Autres
5. Affinités personnelles et Proposition de la part d'un groupe préconçu
6. Affinités personnelles et Compétences des membres du groupe
7. Proposition de la part d'un groupe préconçu et Compétences des membres du groupe

Note du projet	1	2	3	7	5	6	7	Total
0 à 9,99	0	0	0	3	0	0	0	3
10 à 20	15	7	2	17	5	5	1	52
Total	15	7	2	20	5	5	1	55

Sur R, le test de Khi deux donne une valeur de p-value= 0.3793 et avec un risque de 5% , on accepte l'hypothèse pour les critères du choix des groupes.

```
chisq.test(Note, CriteresChoix, simulate.p.value = TRUE)

        Pearson's Chi-squared test with simulated p-value (based on 2000
        replicates)

data:  Note and CriteresChoix
X-squared = 5.5529, df = NA, p-value = 0.3793
```

Hypothèse 13 : La note d'un projet dépend de la connaissance antérieure d'un membre ou plusieurs du groupe.

Note du projet	Pas connaissance	Avec connaissance	Je ne me rappelle pas	Total
0 à 9,99	5	4	0	9
10 à 20	24	81	1	106
Total	29	85	1	115

Sur R, le test de Khi deux donne une valeur de p-value= 0.1174 et avec un risque de 5% , on accepte l'hypothèse pour la connaissance antérieure avec un ou plusieurs membres du groupe.

```
chisq.test(Note,Affinite, simulate.p.value = TRUE)

        Pearson's Chi-squared test with simulated p-value (based on 2000
        replicates)

data:  Note and Affinite
X-squared = 4.7959, df = NA, p-value = 0.1174
```

Hypothèse 14 : La note d'un projet dépend du travail antérieur avec un ou plusieurs membres du groupe.

Note du projet	Pas de travail antérieur	Avec un travail antérieur	Je ne me rappelle pas	Total
0 à 9,99	5	4	0	9
10 à 20	56	45	5	106
Total	61	49	5	115

Sur R, le test de Khi deux donne une valeur de p-value= 1 et avec un risque de 5% , on accepte l'hypothèse pour le travail antérieur avec un ou plusieurs membres du groupe.

```
chisq.test(Note,DejaBosseAvec, simulate.p.value = TRUE)
```

```
        Pearson's Chi-squared test with simulated p-value (based on 2000
        replicates)

data:  Note and DejaBosseAvec
X-squared = 0.44387, df = NA, p-value = 1
```

Hypothèse 15 : La note d'un projet dépend de la première impression lors du premier contact avec le groupe.

Note du projet	Négative	Neutre	Positive	Total
0 à 9,99	5	3	1	9
10 à 20	2	33	43	78
Total	7	36	44	87

Sur R, le test de Khi deux donne une valeur de p-value= 0.0004998 et avec un risque de 5% , on rejette l'hypothèse pour la première impression lors du premier contact avec le groupe.

```
chisq.test(Note,Impression, simulate.p.value = TRUE)
```

```
        Pearson's Chi-squared test with simulated p-value (based on 2000
        replicates)

data:  Note and Impression
X-squared = 31.409, df = NA, p-value = 0.0004998
```

La réussite d'un projet peut être jugé par l'obtention d'une bonne note mais peut être aussi jugé par la satisfaction des membres du groupe par rapport au travail effectué. Pour vérifier cette hypothèse nous allons donc étudier s'il y a une dépendance entre la note du groupe et la satisfaction des membres du groupe.

Hypothèse 16 : La satisfaction d'un membre du groupe par rapport au travail effectué dépend de la note finale du projet.

Note du projet	0	1	2	3	4	5	6	7	8	9	10	Total
0 à 9,99	3	1	2	1	1	0	1	0	0	0	0	9
10 à 20	3	3	7	5	1	10	13	18	28	12	6	106

Total	6	4	9	6	2	10	14	18	28	12	6	115

Sur R, le test de Khi deux donne une valeur de p-value= 0.007496 et avec un risque de 5% , on rejette l'hypothèse pour la satisfaction par rapport au travail effectué.

```
chisq.test(Note,Satisfaction, simulate.p.value = TRUE)
```

```
        Pearson's Chi-squared test with simulated p-value (based on 2000
        replicates)

data:  Note and Satisfaction
X-squared = 30.889, df = NA, p-value = 0.007496
```

Tableau récapitulatif :

Élément traité dans le questionnaire :	Note du Projet :
Nombre d'année passé au sein de Télécom Bretagne	Dépendance
Année de césure	Dépendance
Nombre de projets déjà effectués	Dépendance
Projet choisi dans le questionnaire	Dépendance
L'effectif des membres du groupe	Dépendance
Barrière de la langue dans le groupe	Indépendance
Nombre de nationalité dans le groupe	Dépendance
Implication des membres du groupe	Dépendance
Nombre de filles dans le groupe	Dépendance
Choix du groupe	Dépendance
Choix du sujet	Indépendance
Critères de choix précis	Dépendance
La connaissance antérieure d'un membre ou plusieurs du groupe	Dépendance
Travail antérieur avec un ou plusieurs membres du groupe.	Dépendance
Première impression lors du premier contact avec le groupe	Indépendance
Satisfaction d'un membre du groupe par rapport au travail effectué	Indépendance

Néanmoins la quantité de valeurs pour les notes inférieures à 10 étant extrêmement faible par rapport aux valeurs de succès, la significativité des tests est ici très faible.

4. ACP

4.1 CLASSIFICATION DES VARIABLES

Pour réaliser une analyse en composantes principales nous avons décidé la classification suivante pour les variables:

Variables Quantitatives (avec leur ordre dans les réponses aux questions)

- 3: Age
- 4: AnneesTB
- 7: NombreProjets
- 9: Effectif
- 11: Diversite

- 12: LangageBarrier
- 13: Filles
- 14: Implication
- 19: Note
- 20: Satisfaction

Variables Qualitatives Illustratives (avec leur ordre dans les réponses aux questions)

- 1: Cursus
- 2: Sexe
- 5: Cesure
- 6: Pays
- 8: Projet
- 10: ChoisiProjet
- 15: ChoisiGroupe
- 16: Affinite
- 17: Impression
- 18: DejaBosseAvec

Le code utilisé a été le suivant:

```
resPCA=PCA(dataACP,
quali.sup = c(1,2,5,6,8,10,15,16,17,18,19),
scale.unit=TRUE, graph=T)
```

L'ACP nous permet de faire 3 études: analyse des individus, des variables et le lien entre les deux.

4.2 ANALYSE DES VARIABLES

Initialement, on montre le graphique des variables qui affichera les variables les plus corrélées entre elles comme les plus proches. Ici nous pouvons voir que les variables les plus corrélées sont Diversité et Langage Barrière. C'est évident donc que plus de diversité au sein d'un groupe conduit à une barrière de langage plus significatif.

Corrélations

Dans le résultat suivant, nous affichons les corrélations entre les variables et les dimensions. Nous pouvons voir le degré de proximité très grand entre les variables Diversité et Langage Barrière, ainsi comme Années TB et Nombre Projets.

```
> resPCA$var$cor
                    Dim.1    Dim.2    Dim.3    Dim.4    Dim.5
Age            0.684785724  0.17379445  0.17247544 -0.21167508  0.10191923
AnneesTB       0.879702491 -0.03424609  0.06628814  0.05139677  0.21273209
NombreProjets 0.859972365 -0.07226647 -0.08597207  0.04781028  0.07883872
Effectif      -0.555059293  0.15962960  0.55653439  0.26376365  0.31281880
Diversite      0.121179297  0.20556830  0.82523673  0.01679068  0.25740917
LangageBarriere  0.054568044  0.01388251  0.60790159 -0.53679964 -0.52840398
Filles         0.297609697  0.33295881  0.20472219  0.72482214 -0.45379863
Implication    0.006172471  0.85787564 -0.26079299 -0.07006469 -0.15298246
Satisfaction -0.080859382  0.84331179 -0.17226753 -0.21027804  0.21592380
```

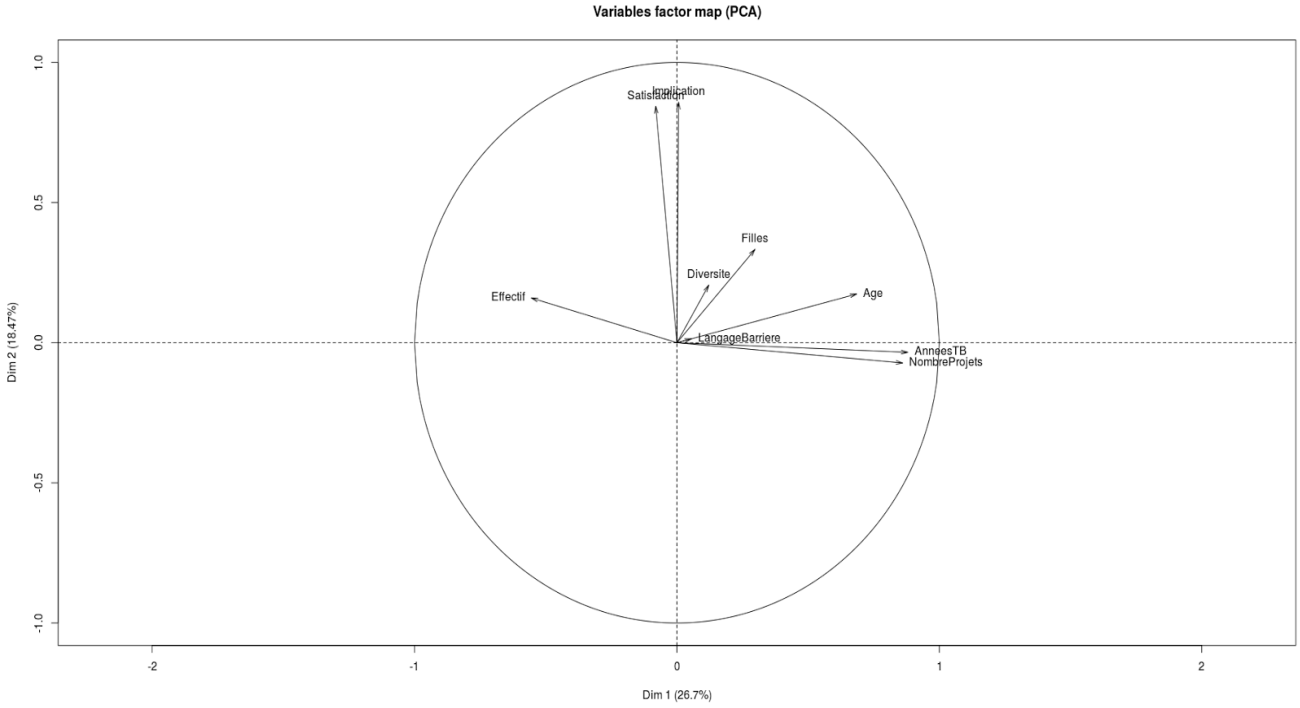
Inertie

Ensuite, dans le code suivant nous allons afficher les degrés d'inertie de chaque composant.

```
> resPCA$eig
   eigenvalue percentage of variance cumulative percentage of variance
comp 1  2.4032612          26.702902                    26.70290
comp 2  1.6625194          18.472437                    45.17534
comp 3  1.5414239          17.126932                    62.30227
comp 4  0.9822339          10.913710                    73.21598
comp 5  0.7811438           8.679375                    81.89536
comp 6  0.6403384           7.114871                    89.01023
comp 7  0.4463844           4.959826                    93.97005
comp 8  0.3335175           3.705750                    97.67580
comp 9  0.2091776           2.324196                   100.00000
```

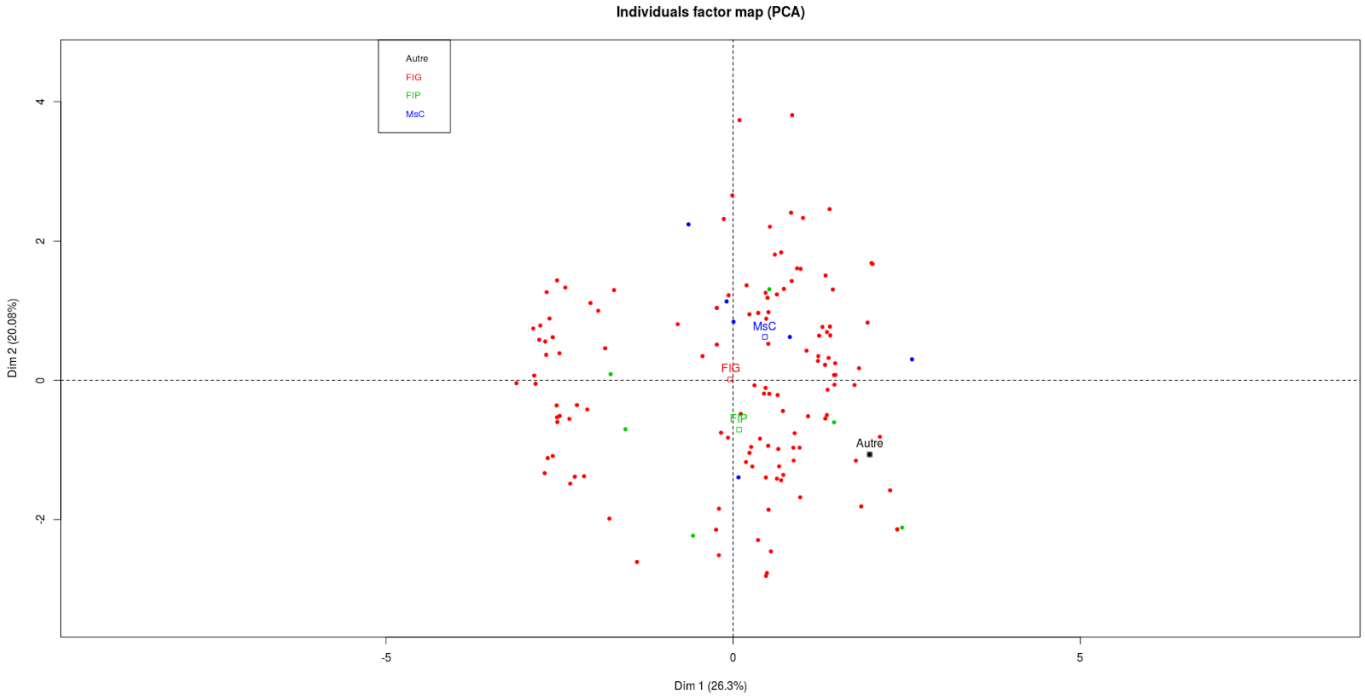
Selon la règle de Kaiser, on garde les 3 premiers composants car elles ont des valeurs propres supérieurs à 1.

Finalement, nous avons le graphique des variables avec le cercle de corrélations.

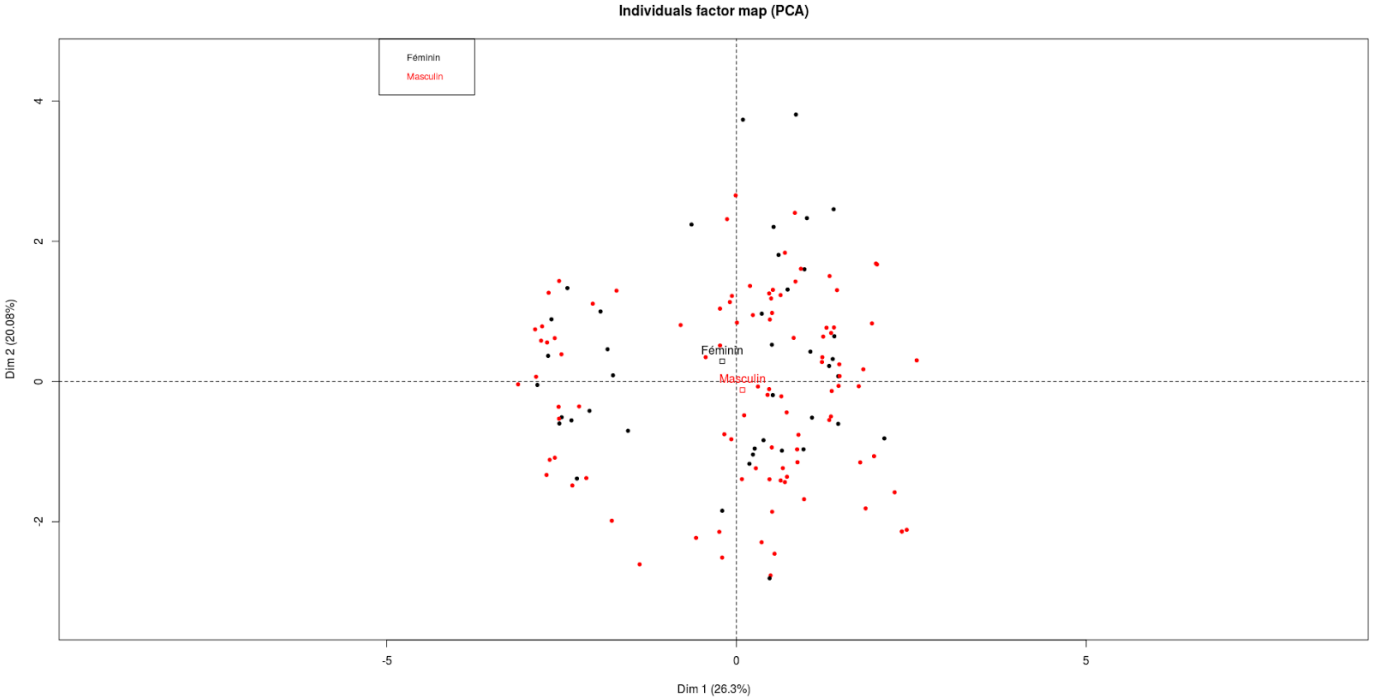


4.3 Analyse des Individus et lien avec variables

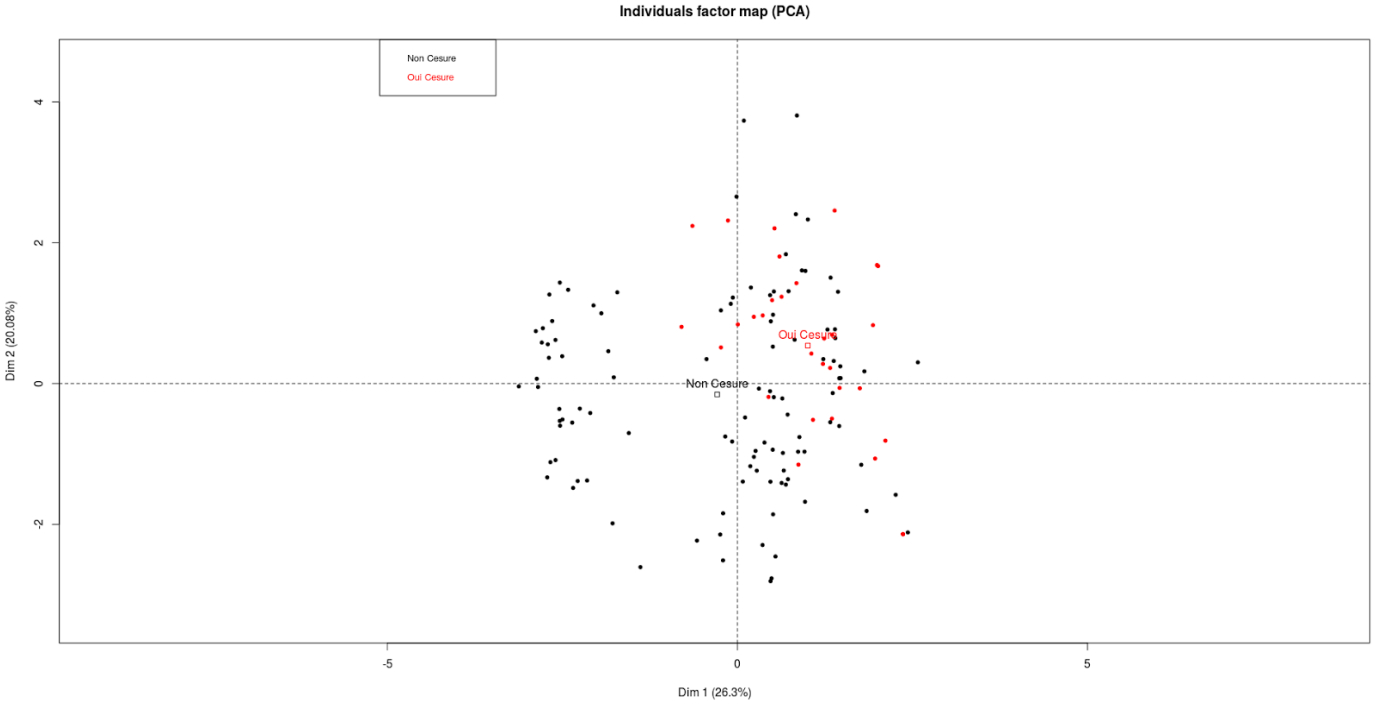
Nous allons analyser les individus selon leurs coordonnées et avec l'aide d'une variable qualitative illustrative pour chaque graphique:



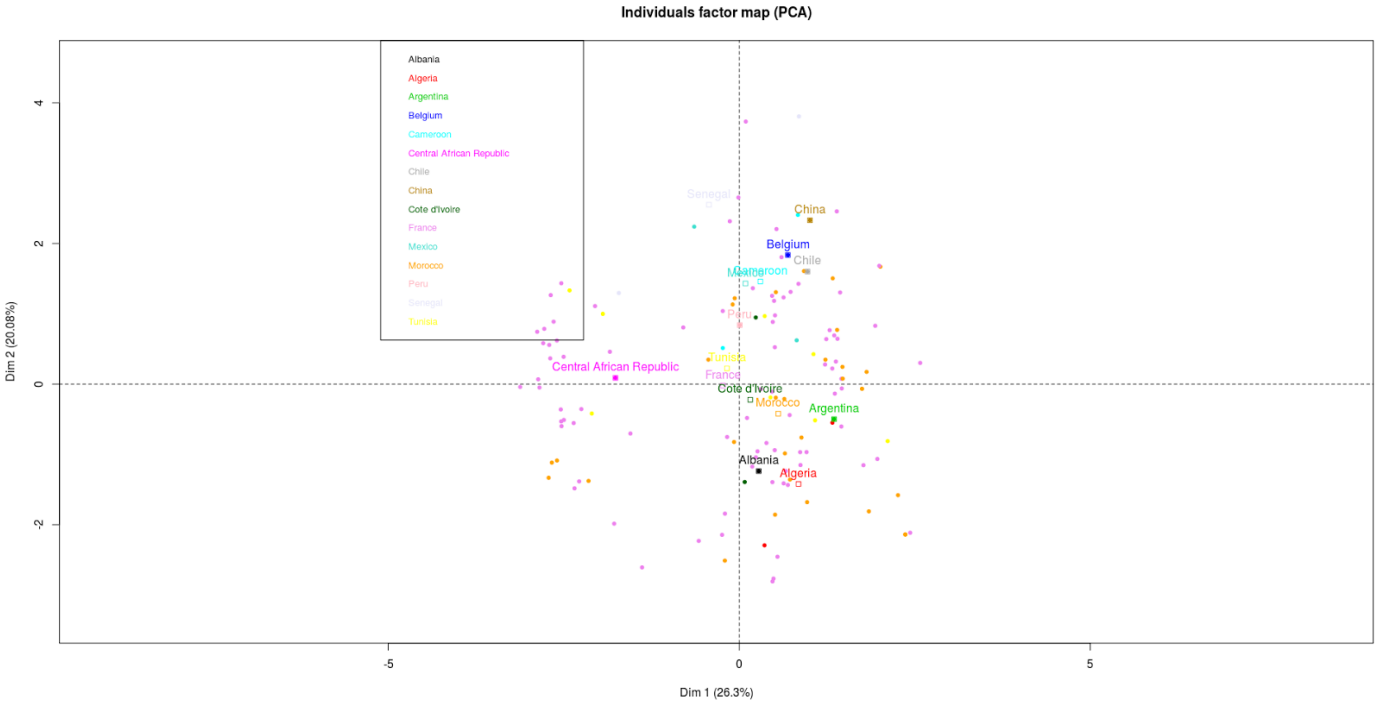
Selon le type de cursus



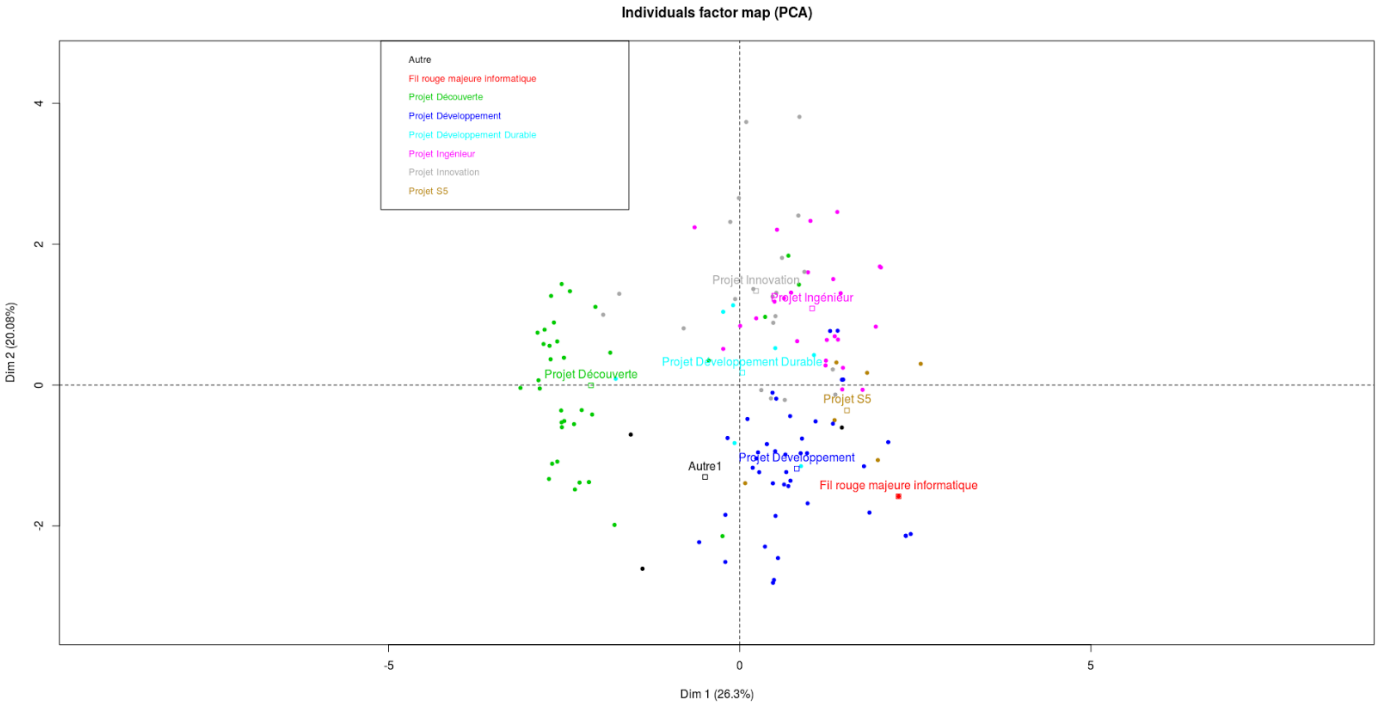
Selon le sexe



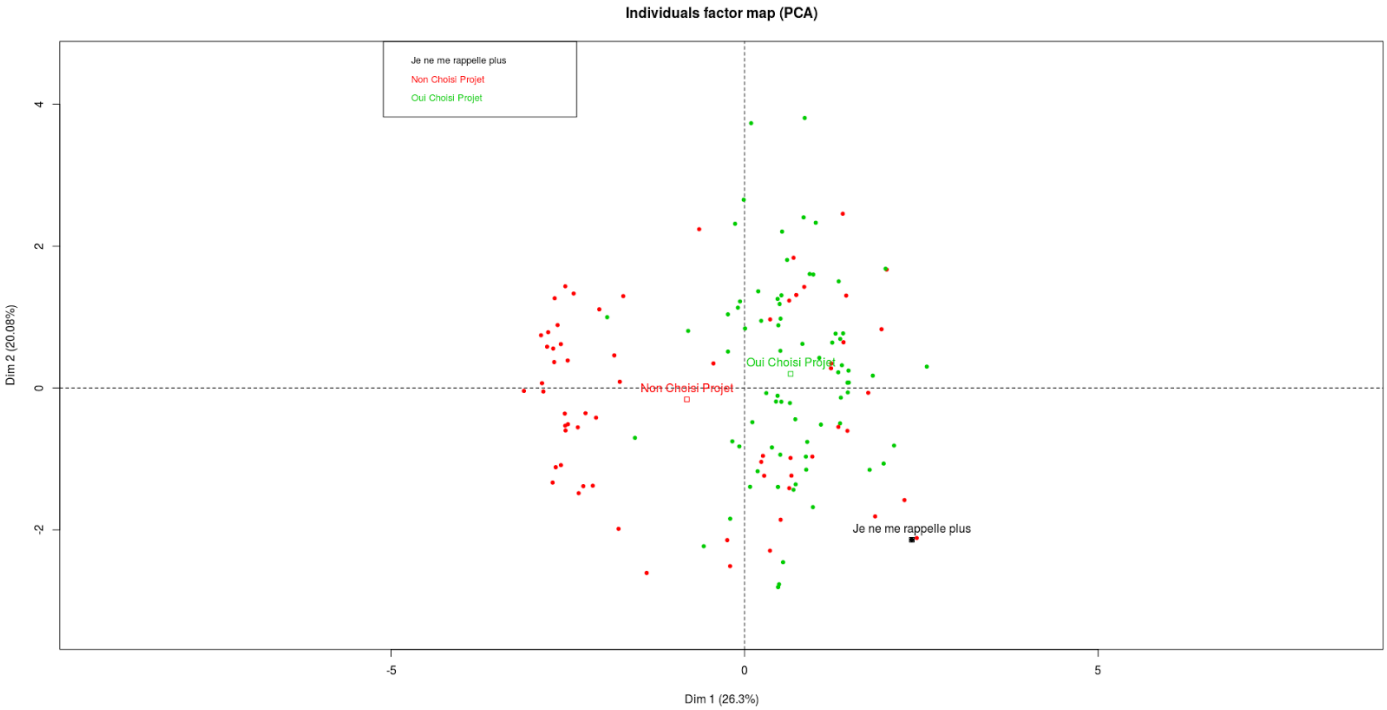
Selon la prise de césure



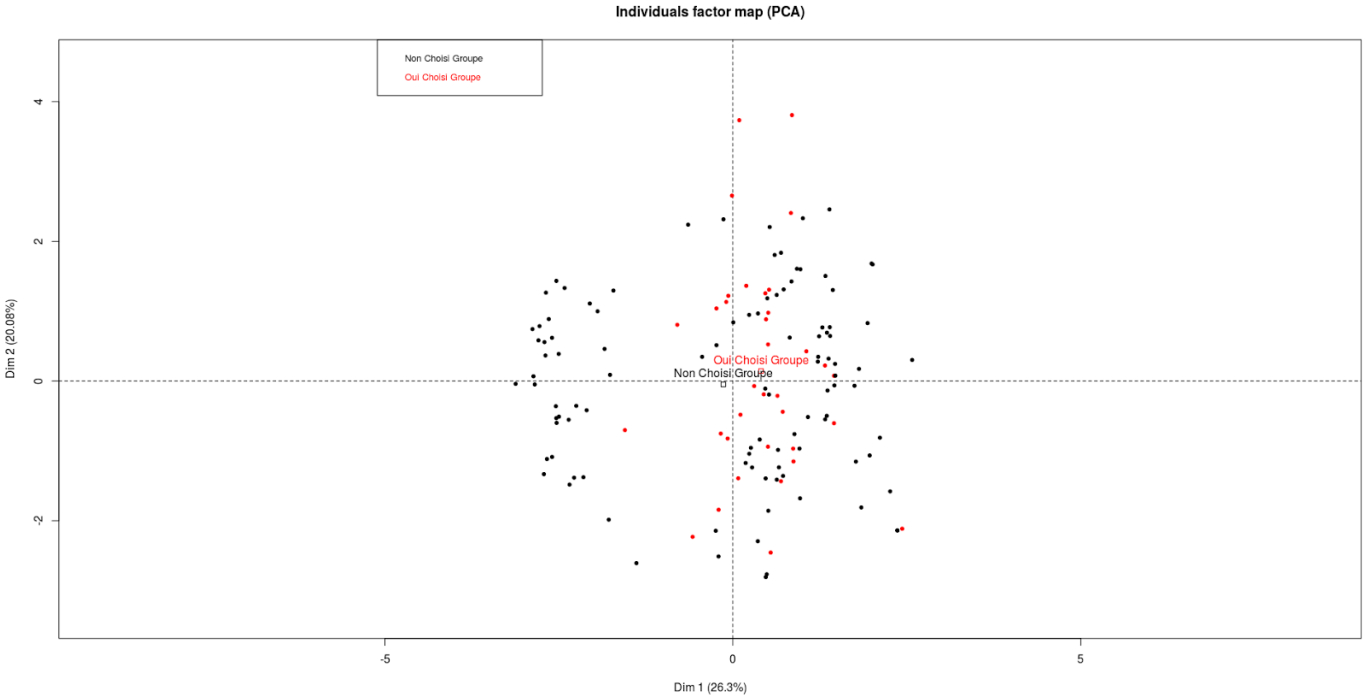
Selon leur nationalité



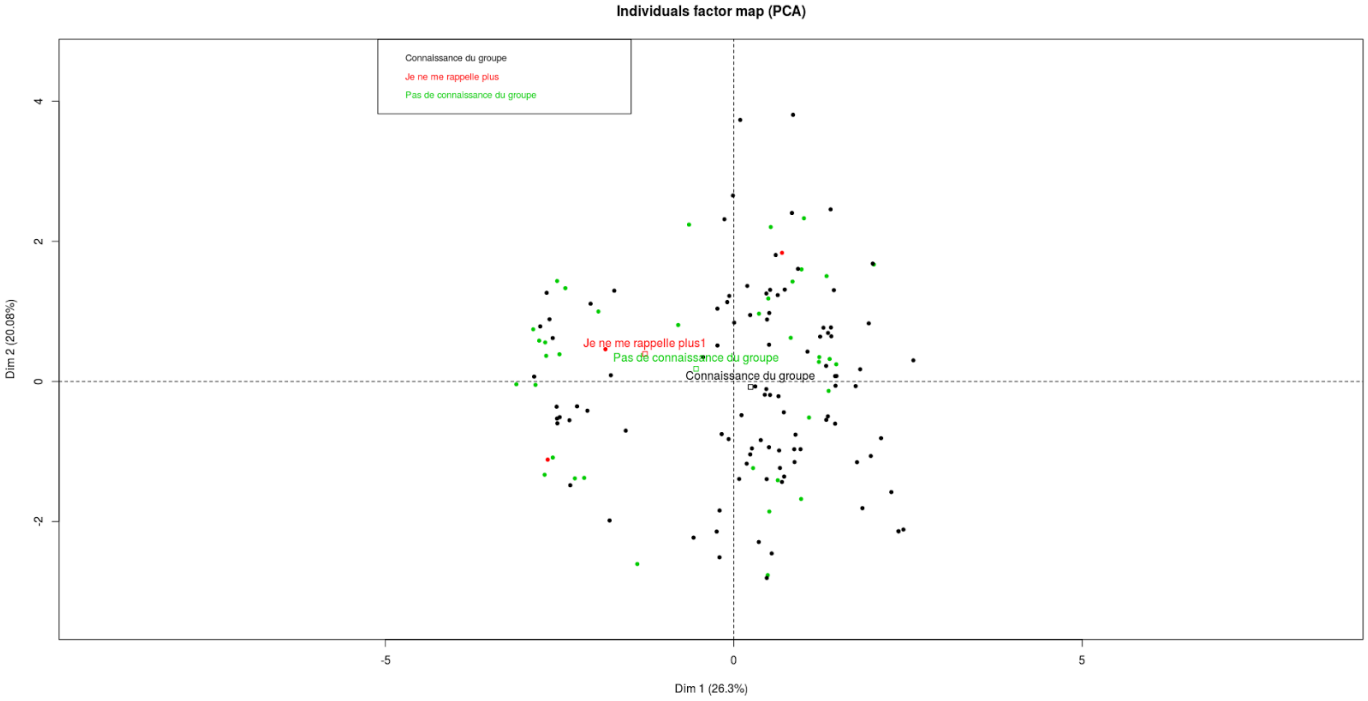
Selon le type de projet



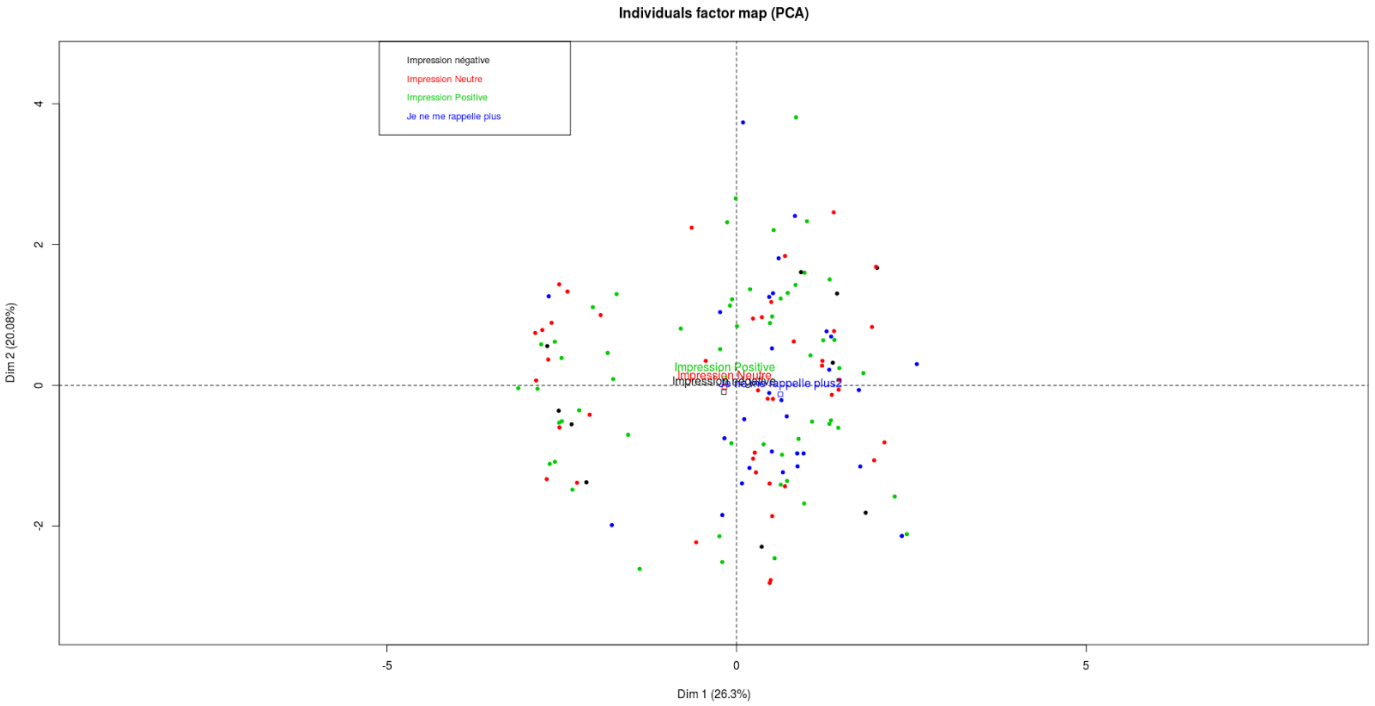
Selon leur choix du sujet



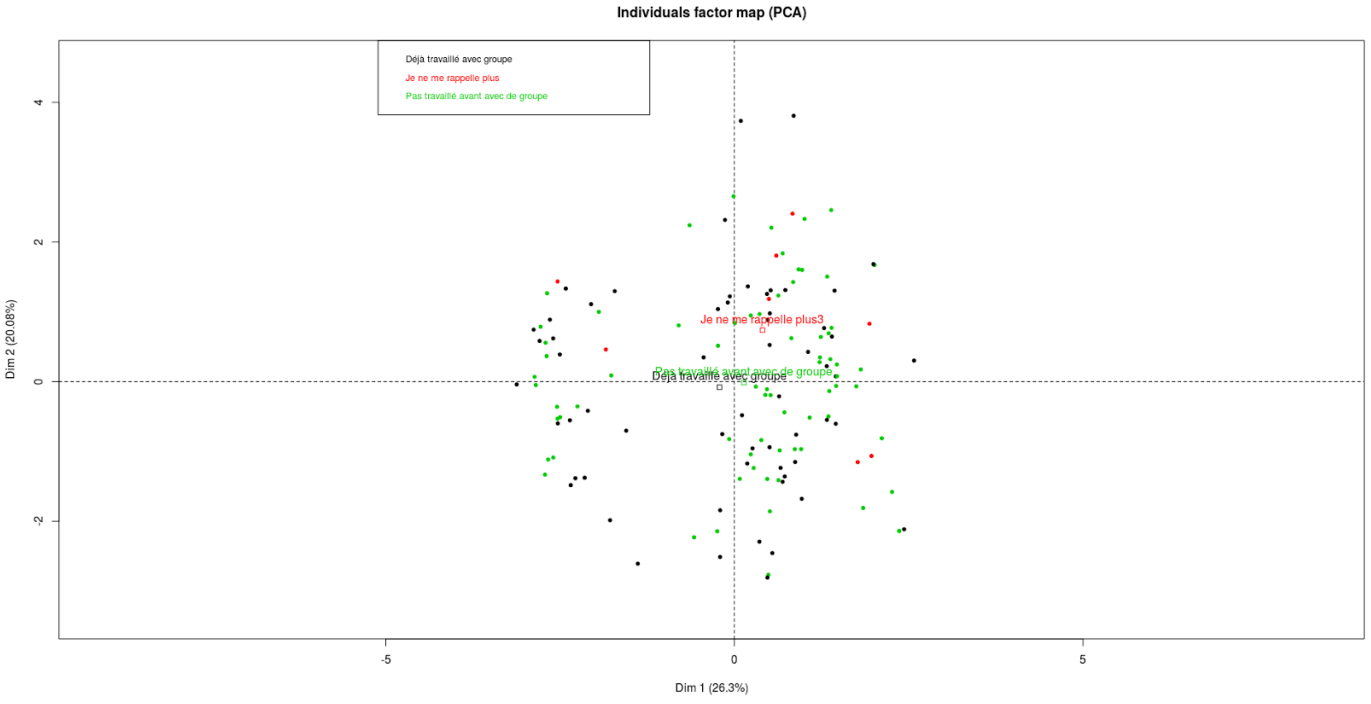
Selon leur choix du groupe



Selon leur connaissance du groupe

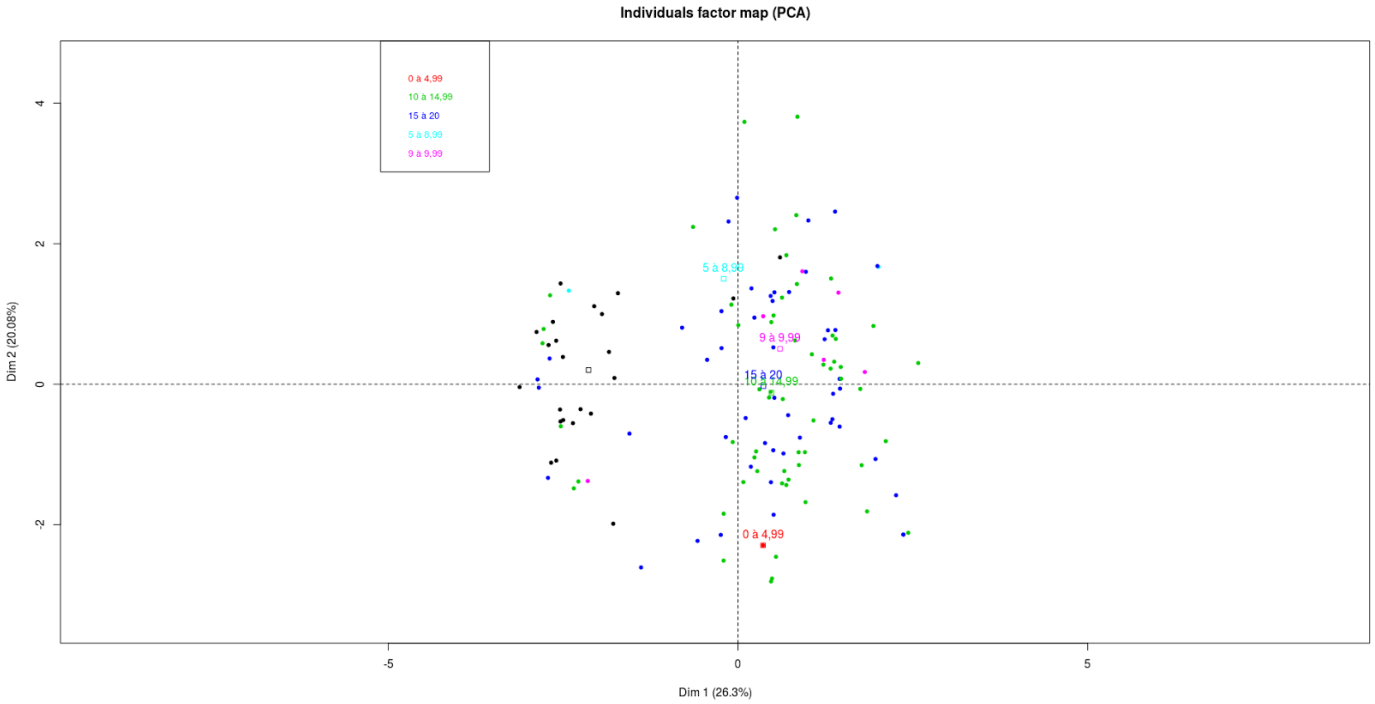


Selon leur impression du groupe



Selon leur niveau d'expérience

(Indique si les individus ont déjà travaillé avec les membres de leur groupes)



Selon leur tranche de notes

5. RÉGRESSION LOGISTIQUE BINAIRE (VARIABLE D'INTÉRÊT = SATISFACTION) :

Comme nous l'avions indiqué lors de la présentation du sujet, le but du projet est de définir des critères de performances permettant de prédire le succès ou l'échec d'un groupe d'individus sur un projet lambda. Or nous avons élaboré notre questionnaire de telle sorte qu'il y ait deux variables qui caractérisent le succès ou l'échec sur un projet, à savoir la note obtenue et la satisfaction des étudiants à l'issu du projet.

Compte-tenu de la très forte disparité des données relatives aux notes, nous avons choisi d'appliquer le modèle de régression logistique à la variable satisfaction dans la mesure où le nombre de notes inférieures à 10 est très faible (environ une quinzaine = non significatif).
Nous avons opté pour une régression logistique binaire par soucis de simplification ce qui nous a amené à transformer notre échelle de satisfaction en une variable d'intérêt binaire : une satisfaction de 0 à 5 vaut 0, et celle-ci vaut 1 pour les valeurs allant de 6 à 10.
Les variables explicatives susceptibles d'expliquer la satisfaction sont nombreuses, car nous pouvons expliquer la variable par toutes les combinaisons de colonnes possibles (en retirant les colonnes qu'on juge non pertinentes) et la qualité de la régression qui en découle varie d'un modèle à l'autre.

Nous avons donc implémenté un algorithme sous R permettant de générer toutes les régressions possibles et de renvoyer le modèle avec la meilleure qualité, cette-dernière étant évaluée grâce au critère d'information d'Akaike (AIC) : en effet, plus l'AIC est faible, meilleure est la régression.

```
setwd("/Users/nabil/Desktop/stats")
stats = read.csv("data_note.csv", sep = "", stringsAsFactors =  FALSE)
attach(stats)
library(MASS)
library(effects)
DATA = na.omit(stats)
str_constant = "~ 1"
s1 = "~ Age + AnneesTB + NombreProjets + Projet + Effectif + Diversite + "
s2 = "LangageBarriere + Filles + Implication + Affinite + Impression + DejaBosseAvec"
str_full = paste(s1,s2, sep = "")
regLog = glm(Satisfaction ~ 1, data = DATA, family = binomial(logit))
regLog.forward = stepAIC(regLog, scope = list(lower = str_constant, upper = str_full),
                                 trace = TRUE, data = DATA, direction = "forward")
summary(regLog.forward)
plot(allEffects(regLog.forward))
```

Script de génération du modèle de régression logistique adopté pour la variable d'intérêt Satisfaction

Grâce à la commande summary nous pouvons récupérer l'ensemble des informations relatifs à notre régression :

```
> summary(regLog.forward)

Call:
glm(formula = Satisfaction ~ Implication + Impression + Filles +
    Effectif, family = binomial(logit), data = DATA)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5006  -0.5786   0.2579   0.6226   1.6055

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.4230     1.3873  -3.188 0.001432 **
Implication   1.3208     0.3620   3.649 0.000264 ***
Impression    1.5513     0.5229   2.967 0.003008 **
Filles       -3.7601     1.6719  -2.249 0.024514 *
Effectif      0.2492     0.1454   1.714 0.086497 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 108.533  on 85  degrees of freedom
Residual deviance:  68.941  on 81  degrees of freedom
AIC: 78.941

Number of Fisher Scoring iterations: 5
```

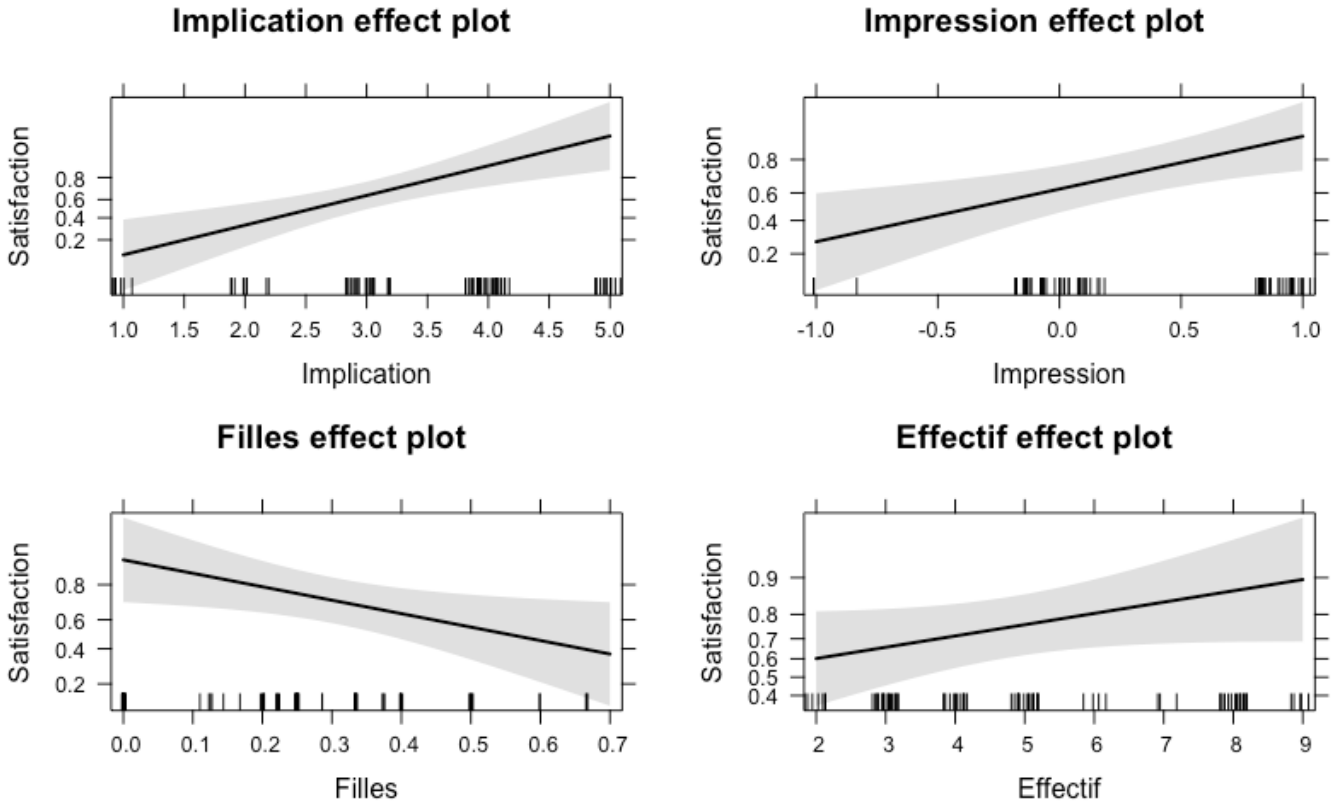
Summary du modèle retenu

Nous pouvons déjà en déduire la significativité de chacune des variables explicatives :

Variable explicative	Significativité
Implication	Très forte
Impression	Forte
Filles	Moyenne
Effectif	Faible

Significativité des variables explicatives

La commande plot(allEffects(regLog.forward)) permet de visualiser l'effet de chacune de ces variables sur la satisfaction :



Graphes des effets des variables explicatives sur la satisfaction

L'analyse des graphes permet d'affirmer qu'il y a une très forte influence de l'implication des membres du groupes et de la première impression sur le niveau de satisfaction (forte pente) : on peut en déduire qu'être déterminé à être assidu et impliqué dans un projet et avoir un premier bon contact avec l'équipe sont deux critères primordiaux pour le succès d'un projet. En ce qui concerne la variable Filles, i.e le pourcentage de filles dans un groupe, nous constatons que plus ile pourcentage de filles est élevé, moins la satisfaction est importante. Ceci n'est sans doute pas du sexisme de la part du modèle, mais on pourrait plutôt imaginer des contraintes personnelles que les filles ont et que les garçons n'ont pas et qui ferait qu'elles seraient par exemple moins impliquées dans un projet. De plus la significativité de la variable Filles est moyenne donc nous pouvons bien évidemment remettre en cause ce résultat compte-tenu de la faible taille de l'échantillon (N = 138).

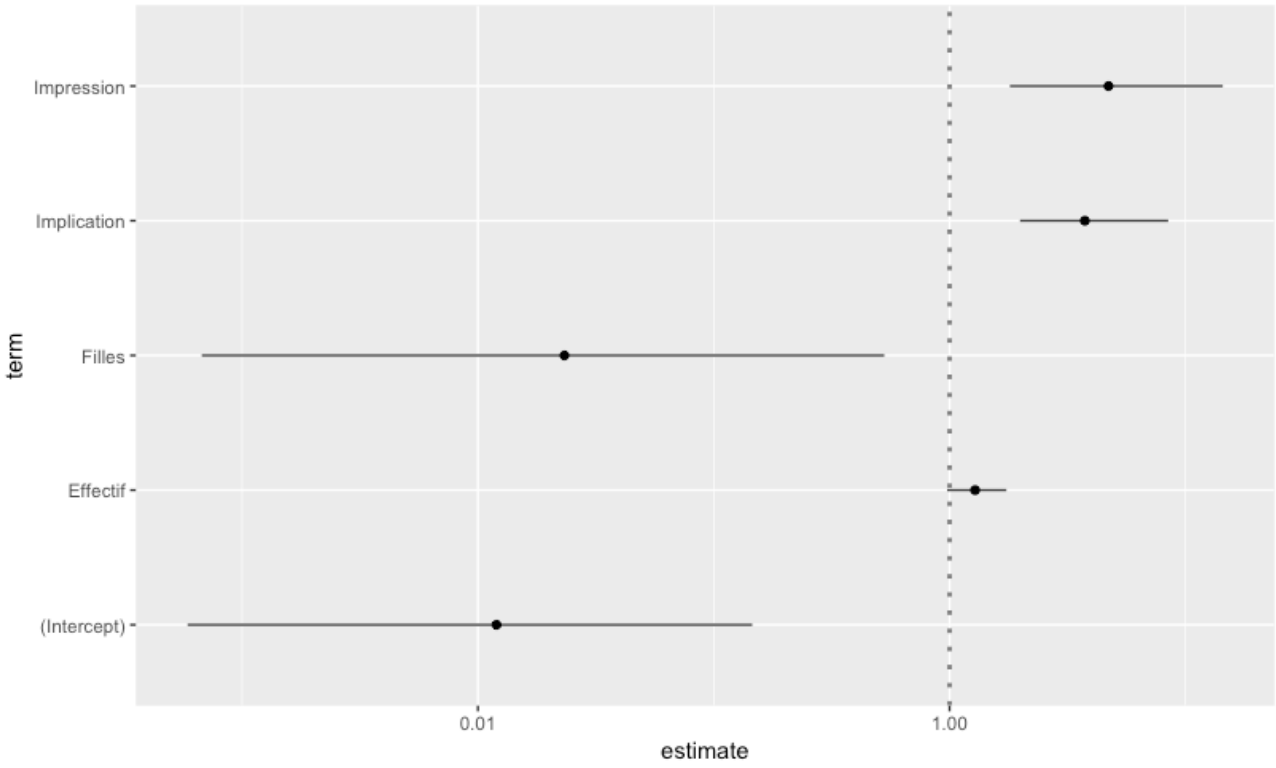
Enfin la variable Effectif, qui est le nombre de personnes dans un groupe, a une influence positive sur la satisfaction dans la mesure où plus le groupe est grand, et plus la satisfaction est importante. Néanmoins la faible significativité de la variable nous amène à une certaine prudence quant aux informations qu'on en tire.

Un autre moyen de connaître l'influence des variables sur la satisfaction est d'analyse les coefficients du modèle :

```
> exp(coef(regLog.forward))
(Intercept) Implication  Impression      Filles    Effectif
 0.01199781  3.74648142  4.71762681  0.02328058  1.28299160
```

Exponentiels des coefficients du modèle

Ces valeurs doivent être comparées à la valeur 1 qui est à la valeur de référence et qui indique l'absence d'effet. Pour une valeur supérieur à 1, on a un effet d'augmentation de la variable vis-à-vis de la satisfaction, et pour une valeur inférieur à 1, on a un effet de diminution. Le graphe suivant représente l'intervalle de confiance à 5% de risque de chacun des coefficients (exponentiels) évoqués ci-dessus :



Intervalles de confiance des coefficients (exponentiels)

On voit que seul la borne inférieure de l'intervalle de confiance du coefficient relatif à l'Effectif est d'environ 1 ce qui nous amène à une faible significativité de cette variable. Pour les autres, en revanche, on obtient des résultats convenables.

6. CONCLUSION

Le projet a suivi des étapes de recueil et d'analyse de données. Nous pouvons finir par citer des remarques ou des pistes d'amélioration sur ces deux aspects.

Au niveau de la collecte de données, dans un premier temps, on aurait pu réduire la taille du questionnaire ; enlever les questions non utilisées, trop vague ou trop ouvertes. Ceci aurait éventuellement résulté en une base de données plus riche ou plus précise. Par exemple, pour ce qui concerne la question de la note du projet, on aurait pu demander aux personnes de fournir une note plus précise au lieu de proposer des tranches. Cela a notamment contribué aux difficultés rencontrées lors du traitement de cette dernière.

Dans le même contexte, nous aurions pu être en mesure de demander à la direction de la formation de nous fournir des informations sur les notes réellement obtenues lors des différents projets à Télécom Bretagne. Ce qui, conjugué à des notes précises, aurait été pertinent pour confirmer la représentativité de notre échantillon.

Pour continuer dans la direction de la précision des données collectées, les questions ouvertes incluses dans le questionnaire n'ont presque été d'aucune utilité lors de l'analyse. On aurait pu les supprimer visant un questionnaire plus léger, ou les remplacer par une échelle (sur l'affinité personnelle entre les membre du groupe).

Quant au niveau de la partie analyse, les outliers détectés auparavant auraient pu être retirés pour permettre l'utilisation de tests paramétriques.

7. BIBLIOGRAPHIE

Recueil, analyse & traitement de données : Le questionnaire (05/2014)
http://rb.ec-lille.fr/l/Analyse_de_donnees/Methodologie_Conception_et_administration_de_questionnaires.pdf

Checking data for outliers (Consulté le 15/11/2016)
https://www.seedtest.org/upload/cms/user/presentation2Remund2.pdf

Régression logistique (Consulté le 15/11/2016)
http://larmarange.github.io/analyse-R/regression-logistique.html

Initiation à la statistique avec R Cours, exemples, exercices et problèmes corrigés - Licence 3, Master 1, écoles d'ingénieurs de Frédéric Bertrand, Myriam Maumy-Bertrand
http://sibib-brest.it-sudparis.eu/cgi-bin/koha/opac-detail.pl?biblionumber=29820

w w w . t e l e c o m - b r e t a g n e . e u

Campus de Brest
Technopôle Brest-Iroise
CS 83818
29238 Brest Cedex 3
France
Tél. : + 33 (0)2 29 00 11 11
Fax : + 33 (0)2 29 00 10 00

Campus de Rennes
2, rue de la Châtaigneraie
CS 17607
35576 Cesson Sévigné Cedex
France
Tél. : + 33 (0)2 99 12 70 00
Fax : + 33 (0)2 99 12 70 19

Campus de Toulouse
10, avenue Edouard Belin
BP 44004
31028 Toulouse Cedex 04
France
Tél. : +33 (0)5 61 33 83 65
Fax : +33 (0)5 61 33 83 75

TELECOM Bretagne