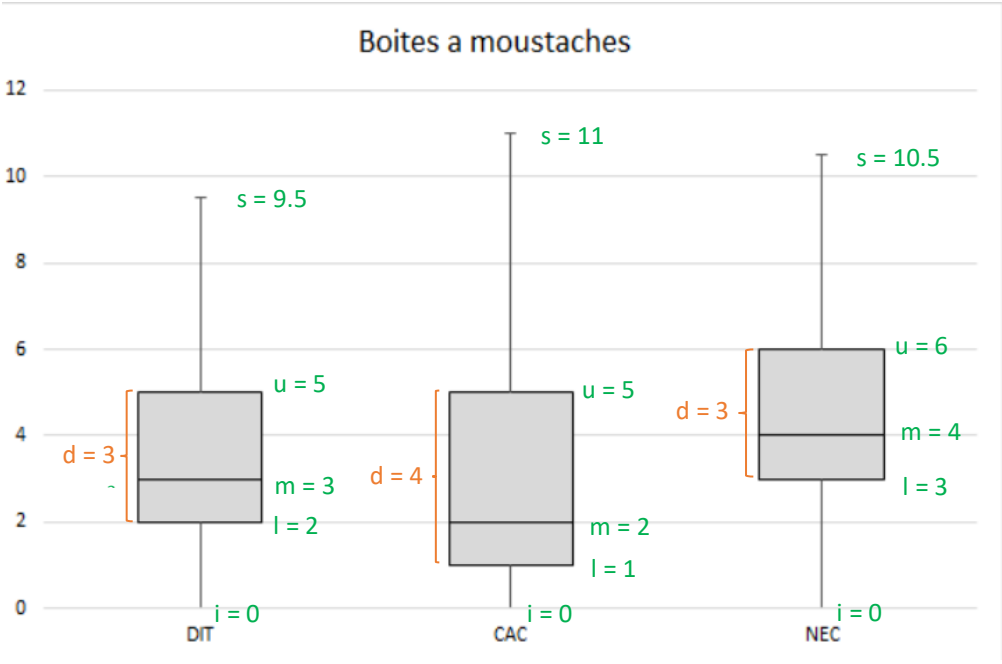
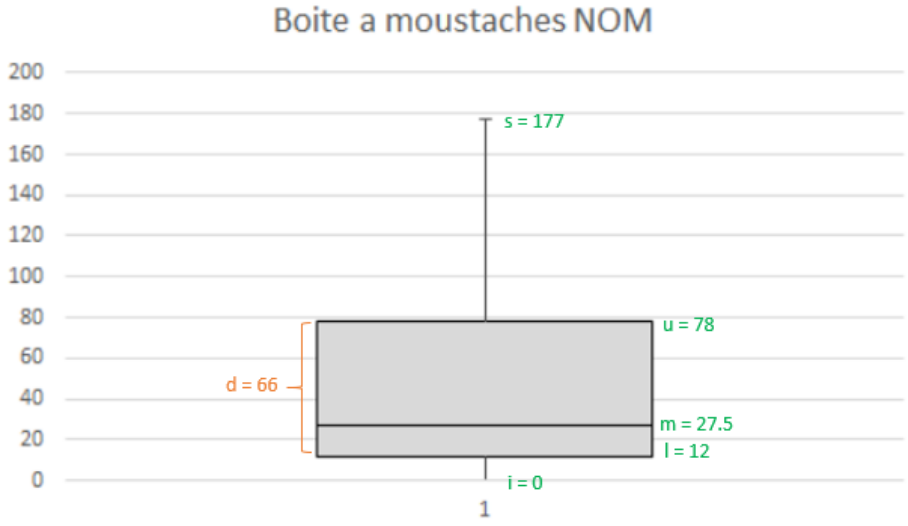
Julien Lanctôt (20140970)
Laurence Fortin (20088891)

T1.
Ici, nous pouvons voir la représentation des boites à moustaches pour DIT, CAC, NEC et NOM avec leurs informations pertinentes.





Pour ce qui concerne la distribution, on peut remarquer que NOM et CAC contiennent des données aberrantes respectivement C30 et C21Selon l'énoncé, les données suivent une distribution normale, cependant lorsque nous regardons les graphiques fréquences (voir Annexe 1), nous pouvons voir qu'aucune des métriques n'est complètement normale. Malgré cela, nous allons assumer qu'elles sont normales pour la suite du travail.

T2.

Le choix d'étude pour ce cas précis est une étude de cas. Cela s'explique en regardant les données qui nous sont présentées. Nous savons qu'il y n'a qu'une seule occurrence d'un phénomène, ce dernier est la collecte unique 30 classes, en plus d'avoir été obtenus lors de la mesure d'un logiciel hypothétique. Comme nous devons confirmer ou infirmer une théorie, cela nous permet de vérifier que c'est bel et bien une étude de cas que l'on examine.
En analysant l'hypothèse donnée dans l'énoncé, nous arrivons à l'hypothèse nulle qui est que les classes ayant un nombre de méthodes inférieur ou égal à 30 n'ont pas moins d'erreurs que celles ayant plus de 30 méthodes

Nous pouvons observer que la variable d'état dans notre cas est le nombre de méthodes (NOM) puisque c'est la variable qui peut être manipulée (en augmentant ou diminuant le nombre de méthodes), c'est également la variable qui influence les résultats de l'étude. Puisque des valeurs du nombre d'erreurs (NEC) sont le résultat de notre étude de cas, nous sommes en mesure de conclure que ce dernier est la variable dépendante de notre étude.

Pour évaluer les 2 métriques (NOM & NEC), nous avons choisi la moyenne et la médiane comme outil de comparaison. Lorsque l'on regarde les classes qui ont moins de 30 méthodes (C1 à C16) et qu'on les compare avec les classes de plus de 30 méthodes (C17 à C30) la moyenne et la médiane du nombre de méthodes (NOM) passent respectivement de 15.25 à 90.79 et de 13.31 à 7.79. Pour la moyenne et la médiane du nombre d'erreurs (NEC) qu'en a eu de 13.5 à 83 et de 3.5 à 5.5 (voir Annexe 2). Nous pouvons apercevoir que le nombre d'erreurs augmente bel et bien lorsque le nombre de méthodes augmente aussi. Cela étant dit, cette mesure n'est pas précise, nous ne pouvons pas dire qu'il y a une corrélation directe entre ces 2 métriques seulement avec cette mesure.

Pour les menaces à la validité, on peut voir que la validité interne est menacée, car les changements dans la variable dépendante ne peuvent pas forcément être raisonnablement attribués aux changements dans les variables indépendantes : l'augmentation du nombre d'erreurs n'est pas forcément liée aux nombres de méthodes, même si on peut voir une légère augmentation dans ce cas. On peut également voir des menaces à la validité de construction, car il se peut que d'autres variables aient influencé le nombre d'erreurs dans les méthodes, par exemple l'expérience et l'âge de la personne qui les a écrites.

T3.

Tout d'abord, en regardant les valeurs des boites à moustaches, on peut voir que 2 classes possèdent des valeurs aberrantes [les valeurs qui dépassent la valeur de la limite supérieur(s)] dans 2 attributs : C21, qui possède un attribut CAC de 17 (plus grand que 11) et C30, qui possède un NOM de 184 (plus grand que 177). Il sera intéressant d'étudier nos corrélations en tenant compte de ses valeurs et en ne les prenant pas en compte.

Comme indiqué dans l'énoncé, il faut utiliser le coefficient de Pearson (r) pour effectuer nos calculs. Le coefficient ce calcul avec les moyennes respectives de $x_i s$ et $y_i s$, (voir Annexe 3). Ensuite, nous avons été en mesure de calculer de coefficient de corrélation de rang de Pearson comme suit :

$$r = \frac{\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y)}{\sqrt{\sum_{i=1}^{n}(x_i - m_x)^2}\sqrt{\sum_{i=1}^{n}(y_i - m_y)^2}} = \frac{(estimation\ des\ covariances)}{(produit\ des\ écarts - type}$$

On arrive ensuite à ces résultats :

	NEC et NOM	NEC et CAC	NEC et DIT
Pearson Coefficient	0.336179047	0.156165298	0.225347199

Avec données aberrantes

	NEC et NOM	NEC et CAC	NEC et DIT
Pearson Coefficient	0.396937052	0.242472242	0.282848599

Sans données aberrantes

On peut voir ci-dessus que les corrélations entre les métriques de structure et NEC, nous apercevons ces coefficients sont minimes. Que ce soit en considérant ou pas les valeurs aberrantes, les coefficients changent de façon négligeable. Il nous est donc possible de conclure qu'il n'y a pratiquement aucune corrélation entre les variables dans notre cas.

En regardant les droites de régressions linéaires des métriques (voir Annexe 4-5), nous sommes en mesure de voir qu'aucune de ces dernières ne semble linéaire. En plus, lorsque l'on regarde leur R^2, qui sont tous loin de 1, il est facile de conclure qu'aucune des métriques n'est linéaire.

T4.

a. Comme nous avons vu dans le T2, on traite d'une seule occurrence, donc nous avons bel et bien une étude de cas. La variable dépendante sera NEC et la variable d'état sera NOM pour les mêmes raisons qu'énoncées en T2. Pour évaluer l'hypothèse [le nombre d'erreurs est une fonction linéaire du NOM], on choisit de travailler avec la visualisation graphique de cette fonction (Annexe 4 et 5). On voit tout de suite que les données sont beaucoup trop dispersées pour être des fonctions linéaires, donc on rejette l'hypothèse.

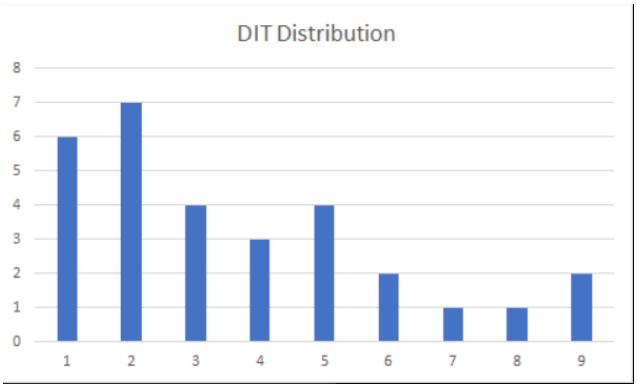
b. Comme nous avons les mêmes données à évaluer ainsi que les mêmes variables, nous avons ici le même type d'étude et les mêmes types de variables que trouvé en a : étude de cas, NEC comme variable dépendante et NOM comme variable d'état. Pour évaluer l'hypothèse [le nombre d'erreurs est une fonction linéaire du DIT nous choisissons encore de travailler avec la visualisation graphique de cette fonction (Annexe 4 et 5). On voit tout de suite que les données sont beaucoup trop dispersées pour être des fonctions linéaires, donc on rejette ici aussi l'hypothèse.

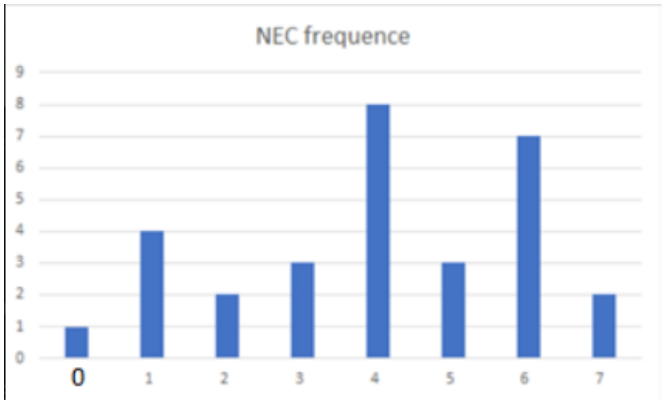
c. Comme en b, nous pouvons conclure que le type d'étude est une étude de cas et NEC est la variable dépendante et NOM la variable d'état. Pour évaluer l'hypothèse [le nombre d'erreurs est une fonction linéaire du CAC], nous restons avec la méthode de visualisation graphique pour être constants (Annexe 4 et 5). On peut remarque qu'ici aussi, les données sont beaucoup trop dispersées de façon aléatoire pour que cette fonction puisse être linéaire. Comme en a-b, ici nous rejetons aussi l'hypothèse.

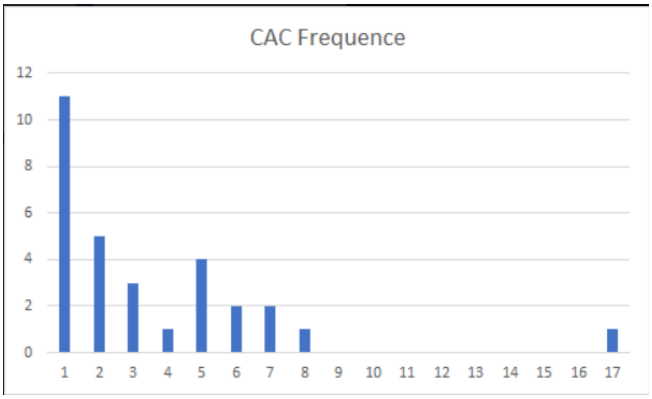
T5.

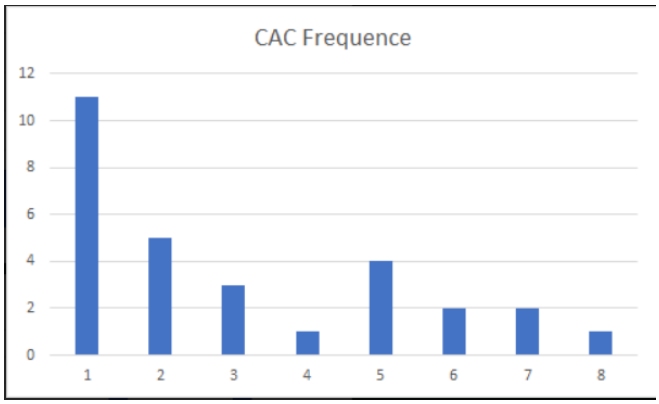
En conclusion, on a démontré de manière assez évidente qu'un nombre limité de variables peut forcément nuire à une étude empirique : dans notre cas, on a prouvé que le nombre d'erreurs avait peu ou pas de lien avec nos 3 autres variables. On peut alors se demander ce qui manquait à cette étude pour obtenir de plus fortes corrélations : est-ce que le stress des personnes qui ont écrit les classes pourrait avoir un impact sur leur qualité? Qu'en est-il du genre ou du milieu social de la personne? Il pourrait y avoir plusieurs autres facteurs qui auraient une beaucoup plus forte corrélation avec le nombre d'erreurs, mais malheureusement, nous n'avions accès qu'à un nombre limité de variables. Cette évidence s'applique fort probablement à de nombreuses études similaires et forcera bien des chercheurs à approfondir leur champ d'études pour obtenir des conclusions robustes et élégantes

ANNEXE 1 : Fréquence des métriques pour évaluer la distribution

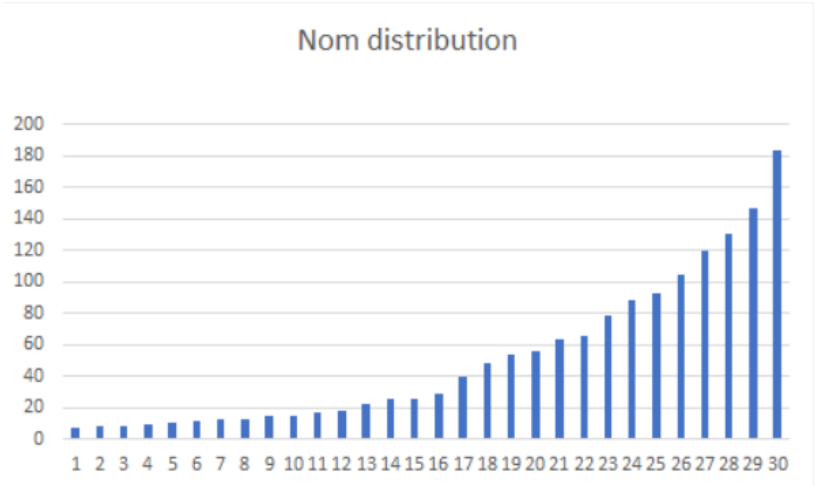








CAC avec données aberrantes et sans données aberrantes



ANNEXE 2 : Calcul de la moyenne et médiane

Comparaison des métriques +/- 30 méthodes

Classes	NOM	NEC
C1	7	4
C3	8	3
C2	8	4
C4	9	2
C5	10	7
C6	11	3
C8	12	3
C7	12	5
C9	15	4
C10	15	5
C11	17	4
C12	18	1
C13	22	1
C14	25	0
C15	26	1
C16	29	6
Moyenne	15.25	3.3125
Médiane	13.5	3.5

Classes	NOM	NEC
C17	40	6
C18	48	4
C19	54	6
C20	56	6
C21	63	4
C22	66	2
C23	78	1
C24	88	4
C25	93	6
C26	104	5
C27	120	6
C28	130	7
C29	147	6
C30	184	4
Moyenne	90.78571429	4.785714286
Médiane	83	5.5

ANNEXE 3 : Coefficient de Pearson

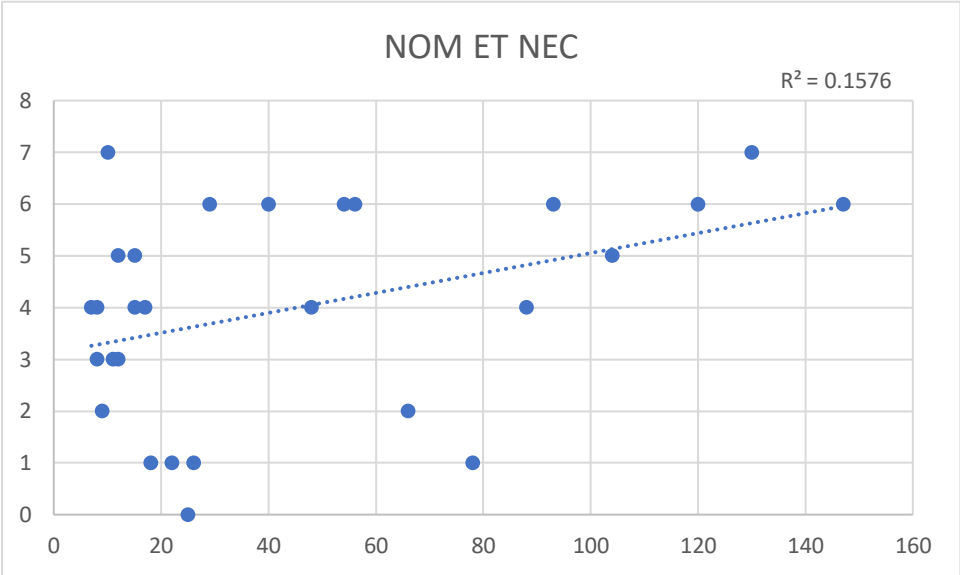
Avec valeurs aberrantes

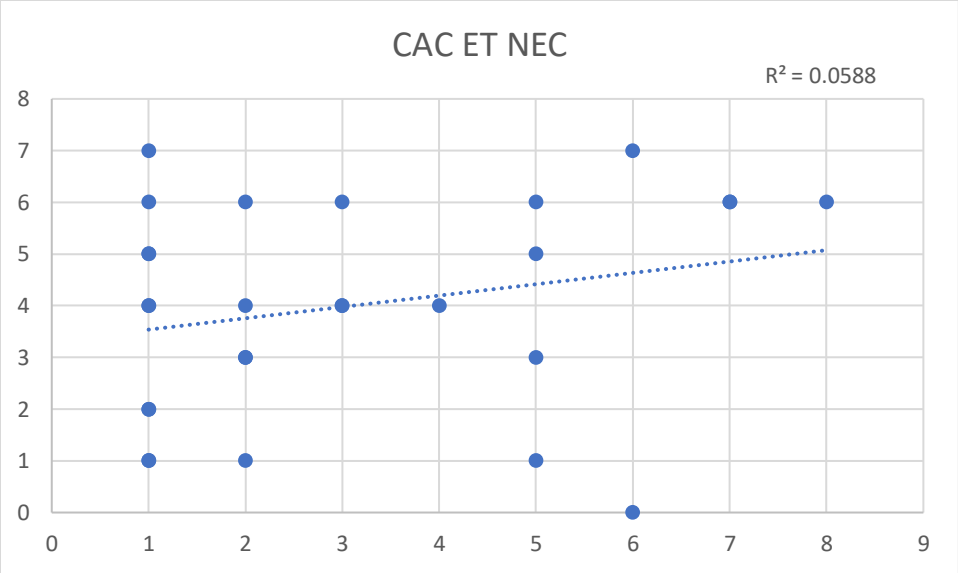
Classes	NOM	DIT	CAC	NEC	Rang NOM	Rang DIT	Rang CAC	Rang NEC
C1	7	2	2	4	30	21	17	16.5
C3	8	2	2	3	28.5	21	17	22
C2	8	1	1	4	28.5	27.5	25	16.5
C4	9	1	1	2	27	27.5	25	24.5
C5	10	1	1	7	26	27.5	25	1.5
C6	11	2	2	3	25	21	17	22
C8	12	5	5	3	23.5	8.5	8.5	22
C7	12	1	1	5	23.5	27.5	25	11
C9	15	2	3	4	21.5	21	13	16.5
C10	15	1	1	5	21.5	27.5	25	11
C11	17	4	4	4	20	12	11	16.5
C12	18	2	2	1	19	21	17	27.5
C13	22	1	1	1	18	27.5	25	27.5
C14	25	6	6	0	17	5.5	5.5	30
C15	26	2	1	1	16	21	25	27.5
C16	29	2	2	6	15	21	17	6
C17	40	3	3	6	14	15.5	13	6
C18	48	4	1	4	13	12	25	16.5
C19	54	7	7	6	12	4	3.5	6
C20	56	5	5	6	11	8.5	8.5	6
C21	63	9	17	4	10	1.5	1	16.5
C22	66	3	1	2	9	15.5	25	24.5
C23	78	3	5	1	8	15.5	8.5	27.5
C24	88	3	3	4	7	15.5	13	16.5
C25	93	4	7	6	6	12	3.5	6
C26	104	5	5	5	5	8.5	8.5	11
C27	120	5	1	6	4	8.5	25	6
C28	130	6	6	7	3	5.5	5.5	1.5
C29	147	8	8	6	2	3	2	6
C30	184	9	1	4	1	1.5	25	16.5

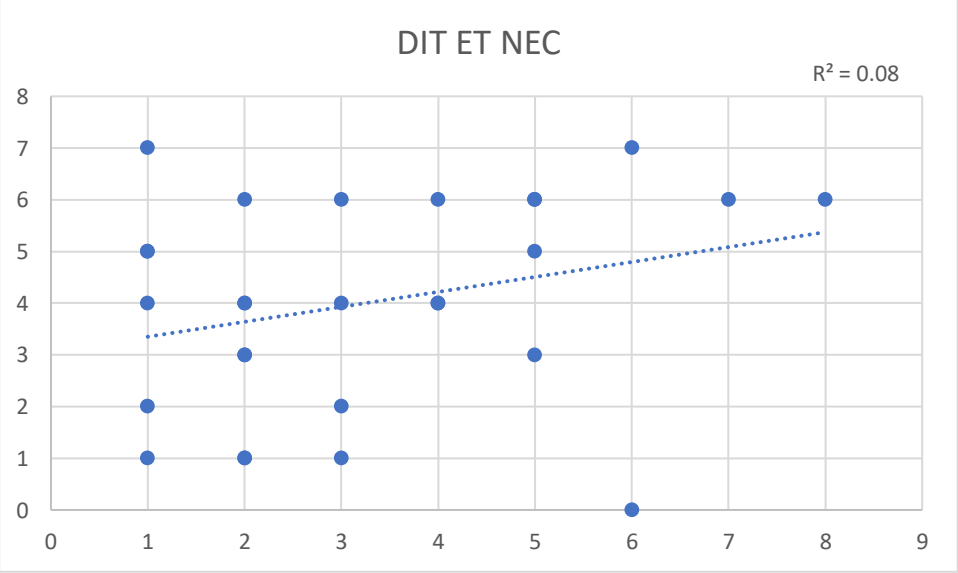
Sans valeurs aberrantes

Classes	NOM	DIT	CAC	NEC	Rang NOM	Rang DIT	Rang CAC	Rang NEC
C1	7	2	2	4	28	19	16	15.5
C3	8	2	2	3	26.5	19	16	20
C2	8	1	1	4	26.5	25.5	23.5	15.5
C4	9	1	1	2	25	25.5	23.5	22.5
C5	10	1	1	7	24	25.5	23.5	1.5
C6	11	2	2	3	23	19	16	20
C8	12	5	5	3	21.5	6.5	7.5	20
C7	12	1	1	5	21.5	25.5	23.5	11
C9	15	2	3	4	19.5	19	12	15.5
C10	15	1	1	5	19.5	25.5	23.5	11
C11	17	4	4	4	18	10	10	15.5
C12	18	2	2	1	17	19	16	25.5
C13	22	1	1	1	16	25.5	23.5	25.5
C14	25	6	6	0	15	3.5	4.5	28
C15	26	2	1	1	14	19	23.5	25.5
C16	29	2	2	6	13	19	16	6
C17	40	3	3	6	12	13.5	12	6
C18	48	4	1	4	11	10	23.5	15.5
C19	54	7	7	6	10	2	2.5	6
C20	56	5	5	6	9	6.5	7.5	6
C22	66	3	1	2	8	13.5	23.5	22.5
C23	78	3	5	1	7	13.5	7.5	25.5
C24	88	3	3	4	6	13.5	12	15.5
C25	93	4	7	6	5	10	2.5	6
C26	104	5	5	5	4	6.5	7.5	11
C27	120	5	1	6	3	6.5	23.5	6
C28	130	6	6	7	2	3.5	4.5	1.5
C29	147	8	8	6	1	1	1	6

ANNEXE 4 : Droite de régression linéaire avec valeur aberrante







ANNEXE 5 Droite de régression linéaire sans valeur aberrante