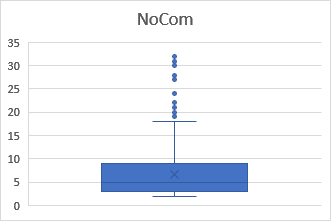
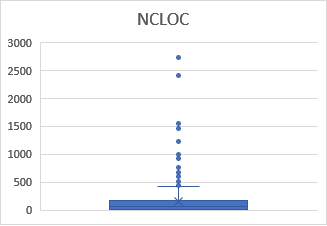
Par : Hugo Carrier 20197563 et Maggie Robert: 20182443

# Visualiser les données

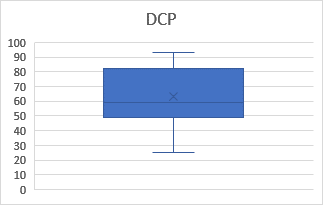
|  |  |  |
| --- | --- | --- |
| m | Médian | 5 |
| u | Quartil1 | 3 |
| l | Quartil3 | 9 |
|  | Min | 2 |
|  | Max | 32 |
| d | Longueur | 6 |
| s | Limite supérieure | 12 |
| i | Limite inférieure | 2 |



|  |  |  |
| --- | --- | --- |
| m | Médian | 71.5 |
| u | Quartil | 12 |
| l | Quartil | 180 |
|  | Min | 4 |
|  | Max | 2732 |
| d | Longueur | 168 |
| s | Limite supérieure | 264 |
| i | Limite inférieure | 4 |



|  |  |  |
| --- | --- | --- |
| m | Médian | 59.21 |
| u | Quartil | 49.27 |
| l | Quartil | 82.36 |
|  | Min | 25.24 |
|  | Max | 93.44 |
| d | Longueur | 33.09 |
| s | Limite supérieure | 98.905 |
| i | Limite inférieure | 32.725 |



Nous avons utilisé le test de Jarque–Bera afin de déterminer rapidement si les données sont distribuées de façon normale. Dans les 3 cas, elles ne le sont pas. Nous avons un résultat de 491 pour NoCom, 37861 pour NCLOC et 471 pour dcp. Visuellement, nous pouvoir voir que les variables sont soit très biaisées (NoCom et NCLOC) ou aplaties (DCP).

# Étudier les corrélations

|  |  |
| --- | --- |
| Spearman | 0.688037 |

|  |  |
| --- | --- |
| Spearman | -0.53352 |

Étant donné que nos distributions ne semblent pas être normales, nous ne pouvons pas utiliser la corrélation de Pearson. Nous avons utilisé celle de Spearman. Elle est aussi appelée la corrélation de rang. Elle compare les positions relatives. Selon la corrélation, NCLOC semble varié dans le même sens que NoCom et DPC semble varié dans le sens contraire de DCP. On peut faire cette même remarque en observant les pentes des régressions. Nous avons aussi sortie le coefficient de détermination . On remarque que les deux régressions semblent être de mauvaise qualité. Logiquement, les données n’étaient pas normales et lorsqu’observée, elles ne semblaient pas avoir la même distribution. Ainsi, on s’attendait à de mauvaise régression.

# Est-ce que les classes qui ont été modifiées plus de 10 fois sont mieux commentées que celles qui ont été modifiées moins de 10 fois ?

Afin de répondre à cette question, nous allons prendre nos données et les séparer en deux groupes. Un groupe qui contient les classes qui ont été modifiées plus de 10 fois et un groupe qui ont été 10 fois et moins. Une fois ces groupes créer, nous allons faire un test statique (un t-test) à savoir si en moyenne des groupes ont été commenté de façon significativement différente. Notre hypothèse null est que les classes qui ont été modifiées plus de 10 fois sont mieux commentées que celles qui ont été modifiées moins de 10 fois. Notre hypothèse alternative est qu’elles ne sont pas mieux. Afin de vérifier cette hypothèse nous allons regarder la densité des commentaires selon le groupe qu’ils ont été affectés en regardant leur nombre de commit. Nous allons prendre un seuil de 5 % comme valeur significative. Nous voulons regarder un côté de la courbe car l’hypothèse indique que les classes modifiées plus de 10 fois sont mieux commentées. Si on avait dit commenté de façon différente, on aurait eu un « two-tail test ».

Le test qui suit nous montre un p-values de . Ainsi, au risque de se tromper une fois sur 20 (seuil de 5% mentionné précédemment) on peut conclure que les classes sont mieux commenter lorsqu’ils ont été modifiées plus de 10 fois.

Les menaces à la qualité de nos résultats sont l’utilisation de proxy comme variable. La densité de commentaires nous donne une bonne image pour décrire la qualité des commentaires, mais ne reflète pas entièrement la vérité. On suppose que mieux commenté égal plus commenté, hors, pour chacune des classes, il y a probablement un niveau de commentaire parfait qui différente de l’une à l’autre. Aussi, il pourrait y avoir trop de commentaires. Ceci est quelque que nous ne regardons pas avec le test.

|  |  |  |
| --- | --- | --- |
|  | *<10* | *>10* |
| Mean | 48.18527778 | 66.62192748 |
| Variance | 81.21956534 | 310.5420584 |
| Observations | 108 | 524 |
| Hypothesized Mean Difference | 0 |  |
| df | 304 |  |
| t Stat | -15.89915565 |  |
| P(T<=t) one-tail | 3.81412E-42 |  |
| t Critical one-tail | 1.649881428 |  |
| P(T<=t) two-tail | 7.62824E-42 |  |
| t Critical two-tail | 1.967798141 |  |