IFT3913 - TP3

Abderrahmane MANSEUR 20148685

Tobias LEPOUTRE 20177637

18 novembre 2023

1 Première tâche : visualisation et analyse de la distribution des variables

Figure 1 – Boîte à moustache pour la variable TLOC

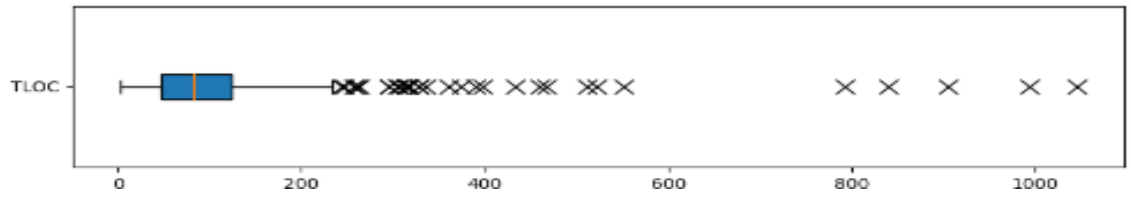


Figure 2 – Boîte à moustache pour la variable WMC

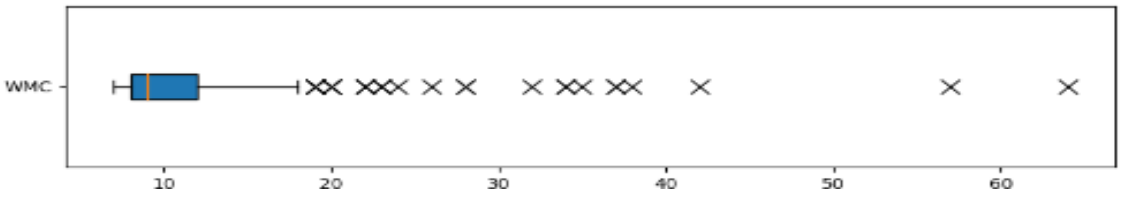


Figure 3 – Boîte à moustache pour la variable TASSERT

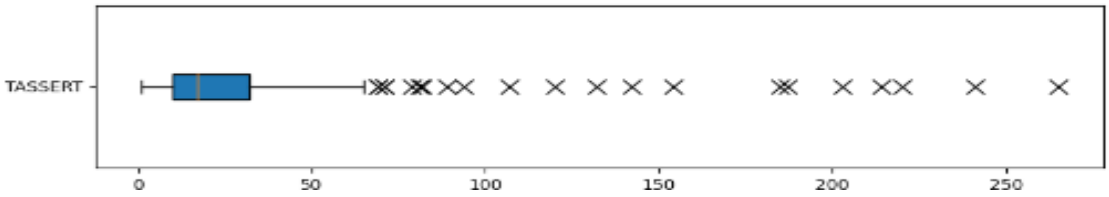


Table 1 – Données statistiques pour TLOC

Données	Valeurs
Moyenne	115.13
Mediane	83.00
Premier quartile	47.50
Troisième quartile	124.50
Longueur boîte	77.00
Limite supérieur	240.00

Table 2 – Données statistiques pour WMC

Données	Valeurs
Moyenne	11.58
Mediane	9.00
Premier quartile	8.00
Troisième quartile	12.00
Longueur boîte	4.00
Limite supérieur	18.00

Table 3 – Données statistiques pour TASSERT

Données	Valeurs
Moyenne	27.19
Mediane	17.00
Premier quartile	10.00
Troisième quartile	32.00
Longueur boîte	22.00
Limite supérieur	65.00

D'après les boites à moustaches et les données statistiques, nous pouvons fournir une petite descirption sur la distributions des données pour chaque variable.

TLOC

- Distribution très large, avec une médiane qui est relativement plus proche du premier quartile que du troisième quartile.
- Il y a énormément de points aberrants qui sont plus élevée que la majorité des données, certains sont même extrêmement élevées.
- L'écart entre le premier quartile et la médiane est plus petit que celui entre la médiane et le troisième quartile, ce qui suggère une distribution asymétrique vers les valeurs supérieures.

WMC

- Distribution étroite, avec une médiane qui est relativement plus proche du premier quartile que du troisième quartile.
- Il y a quelques points aberrants, et très peu de valeur très élevée par rapport à la majorité.
- L'écart entre le premier quartile et la médiane est plus petit que celui entre la médiane et le troisième quartile, ce qui suggère une distribution asymétrique vers les valeurs supérieures.

TASSERT

- Distribution large, avec une médiane qui semble être plus équilibrée entre le premier et le troisième quartile.
- Il y a également une présence notable de points aberrants, indiquant des valeurs significativement plus élevées par rapport à la distribution générale.
- L'écart entre le premier quartile et la médiane est presque égal à celui entre la médiane et le troisième quartile, ce qui suggère une distribution symétrique entre les valeurs inférieurs et supérieures.

2 Deuxième tâche : étude des corrélations entre les variables

Pour cette tache, nous allons commencer par expliquer les étapes de l'étude de la corrélation entre deux variables, ensuite nous allons les appliquer pour notre étude.

2.1 Analyse du diagramme de nuage de points

Ce diagramme nous permet d'avoir une idée visuel à propos de l'existence d'une relation entre deux variable, cependant, ce n'est pas un moyen sur de le confirmer. C'est pour cela que l'étude de certains coefficients de corrélation est important, mais avant cela une étape intermédiaire est nécessaire.

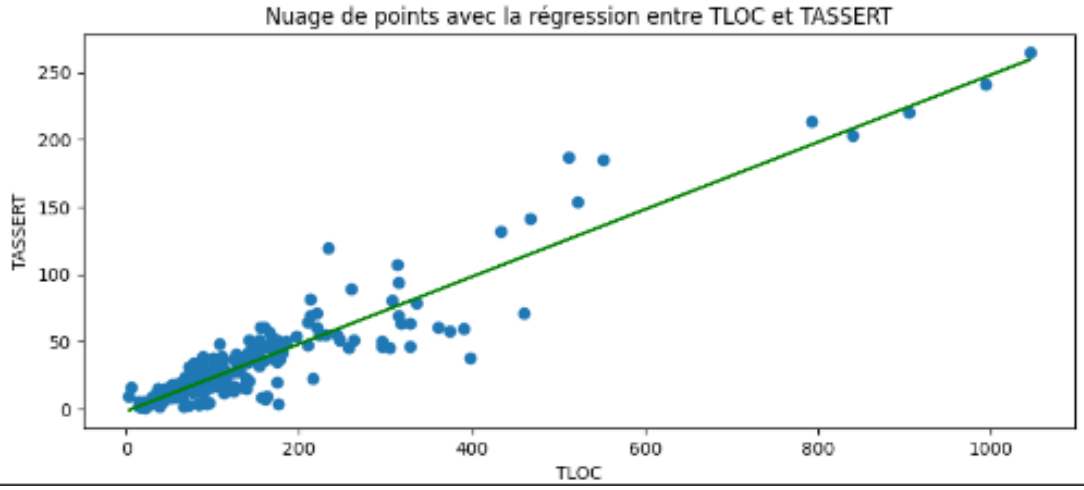
2.2 Analyse des coefficients de corrélation

Avant de procéder au calcul d'un coefficient, il est judicieux d'analyser la distribution des données de ces variables, afin de savoir si les données sont normalement distribués. Car certain coefficient comme celui de Pearson qui suggère une relation linéaire, sont sensibles face à des valeurs aberrantes. Alors, une fois le test effectuer, à partir de là nous avons deux cas :

- **Les donnés sont normalement distribués :** Dans ce cas, nous pouvons directement calculer le coefficient de Pearson et étudier sa valeur, s'il est proche de 1 ou -1, cela suggère la présence d'une relation linaire entre les deux variables. S'il est proche de 0, dans ce cas nous procédons à l'analyse du coefficient de Spearman, pour vérifier l'existence d'une possible relation non linéaire entre les deux variables. Si sa valeur est proche de 1 ou -1, cela suggère l'existence d'une relation non linéaire.
- **Les donnés ne sont pas normalement distribués :** Dans ce cas, nous ne pouvons pas nous fier au coefficient de Person afin de vérifier l'existence d'une relation linéaire, nous somme obliger de calculer le coefficient de Spearman. Ainsi, si la valeur de ce dernier est proche de 1 ou -1, cela suggere l'existence d'une relation entre les deux variables, or pour connaître la nature de cette relation, nous pouvons sois simplement observer la fonction de régression calculer à partir du nuage de points, sois une méthode serais de procéder au calcul du coefficient de Pearson, malgré que ce coefficient n'est pas fiable quand les données ne sont pas normalement répartie, si sa valeur absolue est plus élevée que celle de Spearman alors ceci suggère que la relation entre les deux variables est linéaire, sinon la relation est de nature non linéaire.

2.3 Étude de la corrélation entre nos variables

2.3.1 TLOC et TASSERT

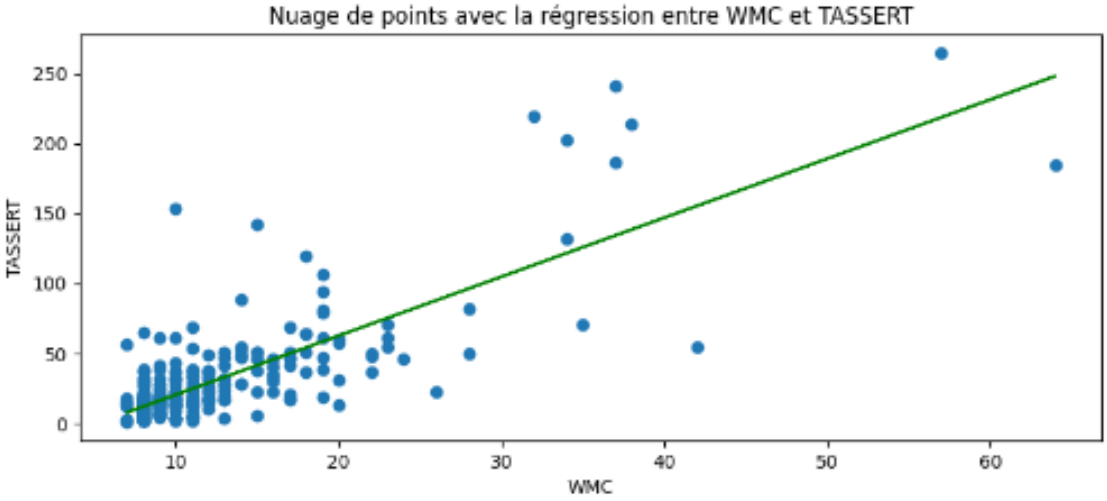


D'aprés le graphe on remarque que sur le diagramme de nuage de point nous obtenons une droite de régression qui semble linéaire, ceci suggère que TLOC et TASSERT ont une relation linéaire.

Cependant afin de confirmer ceci, nous avons fais un test de normalité grace au test de Anderson-Darling qui est un test efficace sur des échantillon de grandes tailles, qui a conclue que les données ne sont pas normalement distribué, nous avons alors calculer le coefficient de corrélation de Spearman qui est de 0.83, sa valeur semble proche de 1, ce qui signifie qu'il existe une relation positive entre ces deux variables.

De plus, afin de vérifier également que la nature de cette relation est bien linéaire, nous avons calculer le coefficient de Pearson qui est de 0.93. Sa valeur est supérieur à celle de Spearman, nous avons donc conclue que ces deux variables sont corréler par une relation linéaire positive.

2.3.2 WMC et TASSERT



D'après le graphe on remarque également pour ces deux variables que sur le diagramme de nuage de point nous obtenons une droite de régression qui semble linéaire, ceci suggère que WMC et TASSERT ont également une relation linéaire. Confirmons cela.

Ces deux variables ont également des données qui ne sont pas normalement distribué, nous avons alors calculer le coefficient de Spearman, qui est de 0.61, sa valeur est relativement proche de 1, ce qui

signifie que ces deux variables sont corrélé par une relation positive modérée.

Quand à la nature de cette relation, le coefficient de Pearson a donner une valeur de 0.79 qui est supérieur à la valeur de coefficient de Spearman. Ceci nous a donc mener à conclure que ces deux variables sont également corrélés par une relation linéaire positive.

3 Troisième tâche : quasi-expérience

Choix d'étude : Quasi-expérience comparant la complexité de classes de code en par les métriques TLOC et WMC pour des classes avec plus de 20 assertions contre celles avec 20 assertions ou moins.

Énoncé des hypothèses :
Les classes qui contiennent plus de 20 assertions sont plus complexes que celles qui contiennent moins de 20 assertions

Définition des variables :
– Variable indépendantes : TASSERT (classes avec + de 20 assertions, et celles avec - de 20 assertions)
– Variables dépendantes : TLOC et WMC (deux mesures de complexité des classes)

Interprétation et généralisation des résultats :

TABLE 4 – Stats + de 20 assertions (TLOC)

Données	Valeurs
Moyenne	202,52
Mediane	141,00
Ecart-type	174.95

TABLE 5 – Stats + de 20 assertions (WMC)

Données	Valeurs
Moyenne	15,50
Mediane	13,00
Ecart-type	8,99

TABLE 6 – Stats de 20 assertions et - (TLOC)

Données	Valeurs
Moyenne	60,52
Mediane	55,50
Ecart-type	31,87

TABLE 7 – Stats de 20 assertions et - (WMC)

Données	Valeurs
Moyenne	9,14
Mediane	8,00
Ecart-type	1,87

On remarque que en moyenne la complexité mesuré par TLOC est 335% plus grande pour les classes de plus de 20 assertions et 170% plus grande pour les la complexité mesurée par WMC.
De plus, les grand écart-types de classes de + de 20 assertions interrogent sur la pertinence d'avoir choisie le nombre de 20 assertions comme référence de variable indépendante. Surtout que ce nombre ne correspond ni à la médianne(17) des TASSERT ni à sa moyenne(27,19). Néanmois, ce nombre est relativement proche des ces deux valeurs et prouve bien que même si la majorité des classes ont moins de 20 assertions, celles qui ont plus de 20 assertion sont généralement bien plus complexes.

Test de T (Voir code sur Github) :
– Pour TLOC : Statistique T = 9,334 (trés élevé) et la Valeur P = $2,14 * 10^{-16}$ (très faible)
– Pour WMC : Statistique T = 8,108 (trés élevé) et la Valeur P = $2,22 * 10^{-13}$ (très faible)

Les résultats du test de T renforce la validité de l'hypothèse que les classes qui contiennent plus de 20 assertions sont plus complexes que celles qui contiennent moins de 20 assertions. Les valeurs trés élevées (>2) des statistiques T et les valeurs très faibles (0) des valeurs P indiquent qu'il est peu probable que la différence observée entre les classe de + de 20, et moins de 20 assertions, soit du au hasard.

Nous pouvons généraliser ces résultats à toutes les classes test Java de jfreechart puisque la taille de l'échantillon analysé est significatif (351 classes). Néanmoins, il faut remarqué que les classes Java analysées ne sont propres qu'à un seul logiciel (jfreechart), et non pas à une multitude de logiciels utilisant des classes Java. De plus, on ne s'intéresse qu'à des classes test et cette étude ne pourrait donc pas utiliser TASSERT comme variable indépendantes dans le cas de classes Java sans test. Notre étude est donc trop limité dans la variétés des logiciels et des types de classes pour pouvoir en faire une généralisation au delà de jfreechart-test.

Discussion des menaces à la validité :
Validité de construction :
- Variables confondantes à prendre en compte : la qualité des fonction TLOC, WMC et TASSERT, ou encore la complexité intrinsèque d'une classe au delà de ses assertions et lignes de code
- Biais dans la conception expérimentale : Il n'est pas pertinent de diviser les classes par nombre d'assertions pour des classes non-test. L'assertion n'est donc pas toujours représentative de la complexité d'une classe.
Validité interne :
- Régression vers la moyenne : il se peut que 20 assertions par classes n'est pas un nombre représentatif de l'ensemble de données habituel.