**Christian Lungescu (20079725)**

**Malik Abada (20173066)**

**TP3 du cours IFT-3913**

**Tache 1 : Visualisation des métriques de l’échantillon en créant des boites à moustaches et déscription des distributions.**

**Notes pour la tache 1 :** Nos valeurs de « boite à moustache » ont été calculé à l’aide d’un programme qu’on a écrit pour ce but (disponible dans le projet GitHub et la remise). Tandis que le diagramme a été généré par un le programme en ligne «datatab» en fonction du data set.

|  |
| --- |
| A graph with a bar and a line  Description automatically generated with medium confidence**Métrique “TLOC”**  **Description de base :** En analysant les données de l’image de gauche, on remarque ici que 50% des classes “test” ont un nombre de ligne(TLOC) entre 47 et 125 (“range” entre le quartile inferieur et supérieur ). On remarque aussi que ∼ 50% des classes “test” on un nombre de ligne inférieur à 83 et que l’autre ∼50% a un nombre de ligne supérieur à 83. |
| A graph with a line and dots  Description automatically generated with medium confidence**Métrique “WMC”**  **Description de base :** En analysant les données de l’image de gauche, on remarque ici que 50% des classes “test” ont un WMC entre 8 et 12 (“range” entre le quartile inferieur et supérieur ). On remarque aussi que ∼ 50% des classes “test” ont un WMC inférieur à 9 et que l’autre ∼50% ont un WMC supérieur à 9.  **Remarques :** On peut aussi voir que qu’il y au moins 25% des classes « test » qui ont un WMC de 8 ou 9 , vu que le « quartile inferieur » est egal a 8 est la mediane est de 9. Ce qui est interessant vu que ce que ces 2 valeurs (8 et 9) representent le quart des informations dans notre ensemble de données WMC tandis qu’il ya plus 60 valeurs possibles dans l’ensemble d’après le schema de gauche(16 valeurs possibles si l’on rejette les données à l’exterieur des limites inferieur et supérieur ). |
| A graph with a line and a line  Description automatically generated with medium confidence**Métrique “TASSERT”**  **Description de base :** En analysant les données de l’image de gauche ,on remarque ici que 50% des classes “test” ont un nombre de assertions JUnit entre 10 et 32 (“range” entre le quartile inferieur et supérieur ). On remarque aussi que ∼ 50% des classes “test” ont un nombre de assertions JUnit inférieur à 17 et que l’autre ∼50% a un nombre d’assertions supérieur à 17. |

**Tache 2 : Étude des corrélations entre TLOC et TASSERT, WMC et TASSERT ; calcul des droits de régression et des coefficients de corrélation qui ont du sens.**

Tout d’abord, il faut déterminer si les ensembles de chaque métrique est normalement distribué. Pour cela on va se servir du “P value” obtenu par différent test de normalité. Comme on peut le voir dans le tableau ci-dessous, tous les **“P values”** trouvées avec nos tests sont < 0.05 donc on peut rejeter l’hypothèse nulle et assumer que les «sets » ne sont pas distribués normalement.Pour calculer ces “P values”, nous avons juste copier chacun des sets dans le calculateur en ligne « <https://www.statskingdom.com/>» (section fit test).

|  |  |  |
| --- | --- | --- |
| **Nom du test**  **Métrique** | **“P value” avec Shapiro-Wilk** | **“P value” avec Kolmogorov** |
| **TLOC** | **∼0** | **2.539e-7** |
| **WMC** | **∼0** | **∼0** |
| **TASSERT** | **∼0** | **∼0** |

Maintenant qu’on a déterminé que les « sets » ne sont probablement pas distribués normalement, on sait que qu’il faut calculer le coefficient de corrélation (r) avec la « méthode de Spearman » entre les métriques TLOC et TASSERT, WMC et TASSERT. Voici un tableau contenant les coefficients de corrélation avec le filtrage de points extrêmes qui se trouvent à l’extérieur des limites inférieures et supérieures calculées dans les boites à moustache de la tache 1, et sans filtrage de points extrêmes.

|  |  |  |
| --- | --- | --- |
| **Données avec filtrage ?**  **Pair de métriques étudiées** | **Données avec filtrage** | **Données sans filtrage** |
| **TLOC ET TASSERT** | r = **0.7761**  Fichier de donnes utilisées = « tlocTassertFilteredData.csv » | r = **0.8346**  Fichier de donnes utilisées = « tlocTassertUnfilteredData.csv » |
| **WMC ET TASSERT** | r = **0.4930**  Fichier de donnes utilisées = « wmcTassertFilteredData.csv » | r = **0.61**  Fichier de donnes utilisées = « wmcTassertUnfilteredData.csv » |

D’après les coefficients de corrélations(r) du tableau ci-dessus, on constate que tous pairs de métriques étudiées (filtrées ou pas) sont corrélés positivement. On voit aussi qu’il y a une plus forte corrélation lorsqu’on ne filtre pas les donnes ce qui peut être expliqué par les données supplémentaire qui raffine la droite de régression ??(**big sauce jai aucune idée pk c plus fortement corrélé xd**).

**Ci-dessous on retrouve les équations des droites de régressions avec leur représentation graphique :**

|  |  |  |
| --- | --- | --- |
| **Informations**  **Pair de métriques étudiées** | **Représentation graphique** | **Équation** |
| **Tloc et tassert avec filtrage de données** |  | y = 0.2294x + 0.06986 |
| **Tloc et tassert sans filtrage de données** |  | y = 0.2498x - 1.568 |
| **Wmc et tassert avec filtrage de données** |  | y = 3.116x - 11.94 |
| **Wmc et tassert sans filtrage de données** |  | y = 4.221x - 21.70 |

**T3. Nous voulons évaluer l’hypothèse : « les classes qui contiennent plus de 20 assertions sont plus complexes que celles qui contiennent moins de 20 assertions ». Décrivez la conception d’un quasi-expérience qui vous permettra de le faire. Ensuite, évaluez l’hypothèse, discutez les résultats et décrivez vos conclusions.**