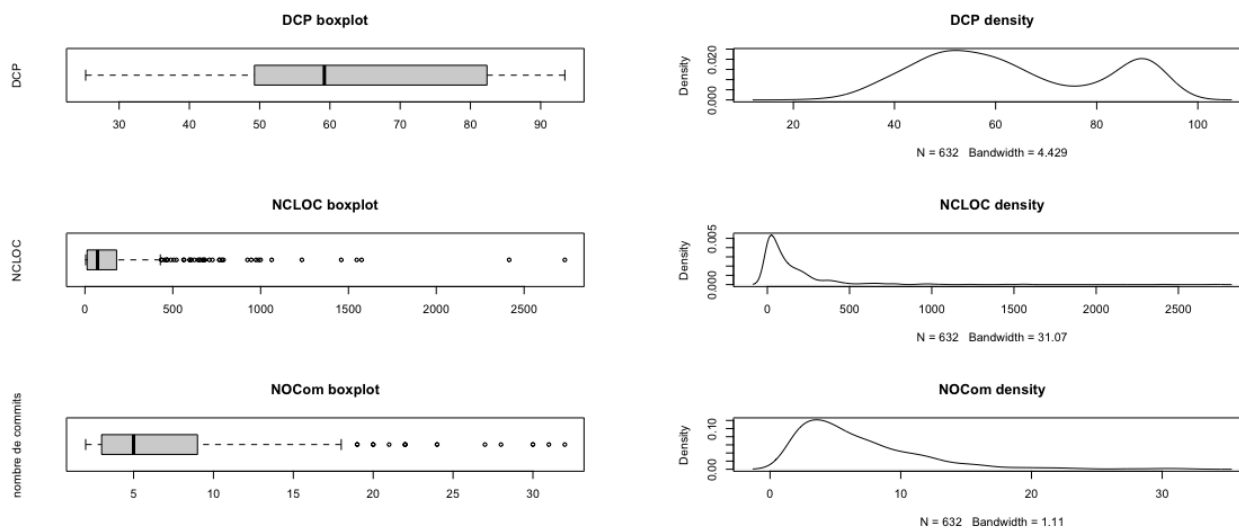


TP3-IFT3913

Wenhao Xu, 20150702

Manping Li, 968527

Tâche 1



Métrique analysis:

	NOCom	NCLOC	DCP
$l=$	3.00	12.00	49.27
$m=$	5.00	71.50	59.21
$u=$	9.00	180.00	82.31
$d=u-l=$	6.00	168.00	33.04
$s=u+1.5d=$	18.00	432.00	132.87
$i=l-1.5d=$	0.00	0.00	0.00

NOCom:

On peut voir qu'il y a beaucoup de points en-dehors du limite supérieure dans le diagramme de boîte (Point extrême). La "queue" est sur le côté droit de la distribution. Donc la distribution est asymétrique à droite.

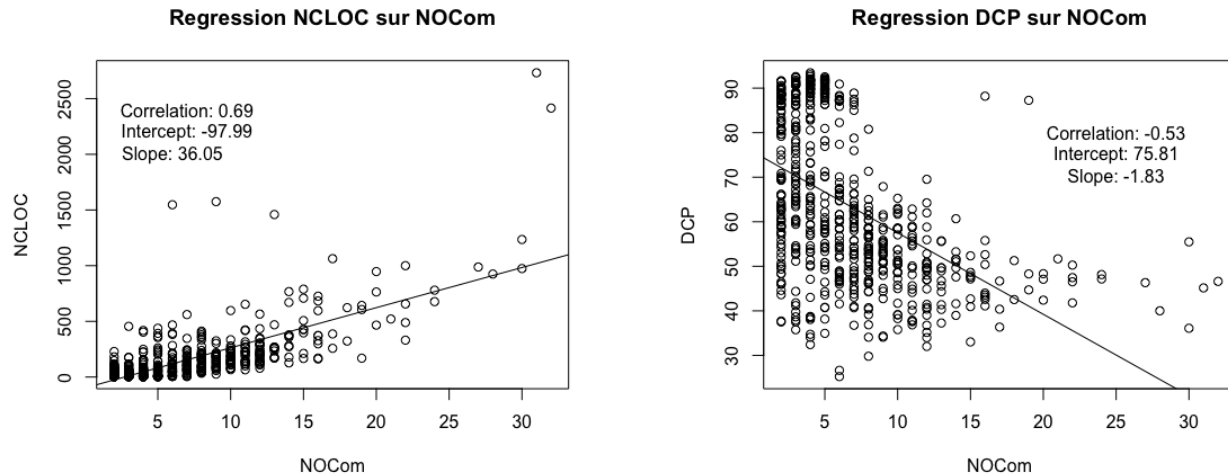
DCP:

Il n'y a pas de point extrême dans la distribution de DCP. Mais la distribution de DCP n'est pas normalement distribué

NCLOC:

On peut aussi voir qu'il y a beaucoup de points en-dehors du limite supérieure dans le diagramme de boîte (Point extrême) pour NCLOC. La "queue" est sur le côté droit de la distribution. Donc la distribution est asymétrique à droite.

Tâche 2



(1) Corrélation:

Comme les variables NCLOC, DCP et NOCom ne sont pas normalement distribués, il faut utiliser "spearman" pour calculer les corrélations entre eux. nous avons:

$\text{Corr}(\text{NOCom}, \text{NCLOC}) = 0.69$;

La corrélation entre NOCom et NCLOC est positive et plus que 0.6, On peut dire que NOCom et NCLOC sont positivement corrélés.

$\text{Corr}(\text{NOCom}, \text{DCP}) = -0.53$;

La corrélation entre NOCom et DCP est négative, c'est pas loin de -0.6, on peut dire que NOCom et DCP sont négativement corrélés.

(2) Regression:

Modèle de regression: $y = a + bx$

On prend variable NOCom comme variable indépendante(x) et NCLOC comme variable dépendante(y), on fait la regression, on a $a = -97.99$, $b = 36.05$. b est la pente de la regression qui est positive. c'est à dire que les classes qui ont été modifiées plus ont NCLOC plus élevé. Ceci correspond aux corrélation entre NCLOC et NOCom.

On prend variable NOCom comme variable indépendante(x) et DCP comme variable dépendante(y), on fait la regression, on a $a = 75.81$, $b = -1.83$. b est la pente de la regression qui est négative. c'est à dire que les classes qui ont été modifiées plus ont DCP moins élevé. ceci correspond aux corrélation entre DCP et NOCom.

Tâche 3

1. Choix d'étude:

La concept de Quasi-expérience

2. Enoncé des hypothèses: les classes qui ont été modifiées plus de 10 fois sont mieux commentées que celles qui ont été modifiées moins de 10 fois

Évaluez l'hypothèse: H_0 : il n'y a pas de différence entre les classes qui ont été modifiées plus de 10 fois et qui ont été modifiées moins de 10 fois

3. Définition des variables:

On utilise le DCP pour mesurer le niveau de commentaire entre des classes. On aussi utilise le NoCom pour diviser des données en deux groupes. Un groupe dont les classes ont été modifiées plus de 10 fois. Un groupe dont les classes ont été modifiées égal ou moins de 10 fois. Le raison qu'on n'a pas choisi NCLOC comme variable c'est que NCLOC mesure pas le niveau de commentaires.

4. Interprétation et généralisation des résultats:

```
> t.test(subdata.plus$DCP,subdata.moin$DCP,alt="less",var.equal = TRUE)
```

Two Sample t-test

```
data: subdata.plus$DCP and subdata.moin$DCP
t = -11.912, df = 630, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
 -Inf -15.96974
sample estimates:
mean of x mean of y
 48.98543  67.51804
```

5. Menaces à la validité: