PROJET DATA SCIENCE

Nayel KETFI - Said SEFIANE - Jackson MBELLA

Analyse préliminaire

	date	quarter	department	day	team	targeted_productivity	smv	wip	over_time
1	1/1/2015	Quarter1	sweing	Thursday	8	0.80	26.16	1108	7080
2	1/1/2015	Quarter1	finishing	Thursday	1	0.75	3.94	NA	960
3	1/1/2015	Quarter1	sweing	Thursday	11	0.80	11.41	968	3660
4	1/1/2015	Quarter1	sweing	Thursday	6	0.80	25.90	1170	1920
5	1/1/2015	Quarter1	sweing	Thursday	7	0.80	25.90	984	6720
6	1/1/2015	Quarter1	finishing	Thursday	2	0.75	3.94	NA	960
7	1/1/2015	Quarter1	sweing	Thursday	10	0.75	19.31	578	6480
8	1/1/2015	Quarter1	finishing	Thursday	8	0.75	2.90	NA	960
9	1/1/2015	Quarter1	finishing	Thursday	4	0.75	3.94	NA	2160
10	1/1/2015	Quarter1	finishing	Thursday	7	0.80	2.90	NA	960
11	1/3/2015	Quarter1	finishing	Saturday	4	0.80	4.15	NA	6600
12	1/3/2015	Quarter1	finishing	Saturday	11	0.75	2.90	NA	5640
13	1/3/2015	Quarter1	sweing	Saturday	1	0.80	28.08	772	6300
14	1/3/2015	Quarter1	sweing	Saturday	3	0.80	28.08	913	6540
15	1/3/2015	Quarter1	sweing	Saturday	11	0.80	11.61	1005	7080
16	1/3/2015	Quarter1	finishing	Saturday	2	0.80	4.15	NA	960
17	1/3/2015	Quarter1	sweing	Saturday	2	0.75	19.87	944	6600

QUESTION N°1

Combien y a-t-il d'observations et de variables disponibles ?

600 observations et 15 variables.

QUESTION N°2

Le jeu de données contient-il des valeurs manquantes ? Si oui, combien d'observations et de variables sont concernées ? Donnez le nom de ces variables.

Oui, 254 observations contiennent des valeurs manquantes. 1 variable (wip) est concernée par ces données manquantes.

QUESTION N°3

Question théorique : suggérez au moins deux méthodes pour traiter ces données manquantes.

Nous pourrions supprimer les observations qui contiennent des valeurs manquantes. Nous pourrions également remplacer les données manquantes par des données artificielles (par exemple, une moyenne des observations).

Analyse préliminaire

QUESTION N°4

Calculez les statistiques descriptives pour la variable cible *actual_productivity*. Interprétez les résultats.

Les statistiques descriptives calculées semblent cohérentes. Pas de valeurs aberrantes. La moyenne et la médiane sont assez proches, avec une valeur acceptable. 75 % des productivités sont au-dessus du premier quartile et 25 % au-dessus du troisième quartile.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.2494	0.6648	0.7506	0.7262	0.8004	1.1005

QUESTION N°5

Calculez le coefficient de corrélation entre la variable *actual_productivity* et les autres variables numériques disponibles. Quelles sont les variables les plus corrélées à la productivité actuelle ?

Après une étude sur les coefficients de correlation, les variables les plus corrélées avec la «actual_productivity» sont «incentive» et la «targeted-productivity» (0.81 et 0.7077)

Analyse en Composantes Principales

QUESTION N°6

Si deux variables sont parfaitement corrélées dans le jeu de données, serait-il judicieux d'inclure ces deux variables dans l'analyse en faisant une ACP ? Justifiez.

Non car le but de la réalisation de l'ACP est de réduire le nombre de variables car si deux variables sont parfaitement corrélées, elles deviennent redondantes et la deuxième devient inutile.

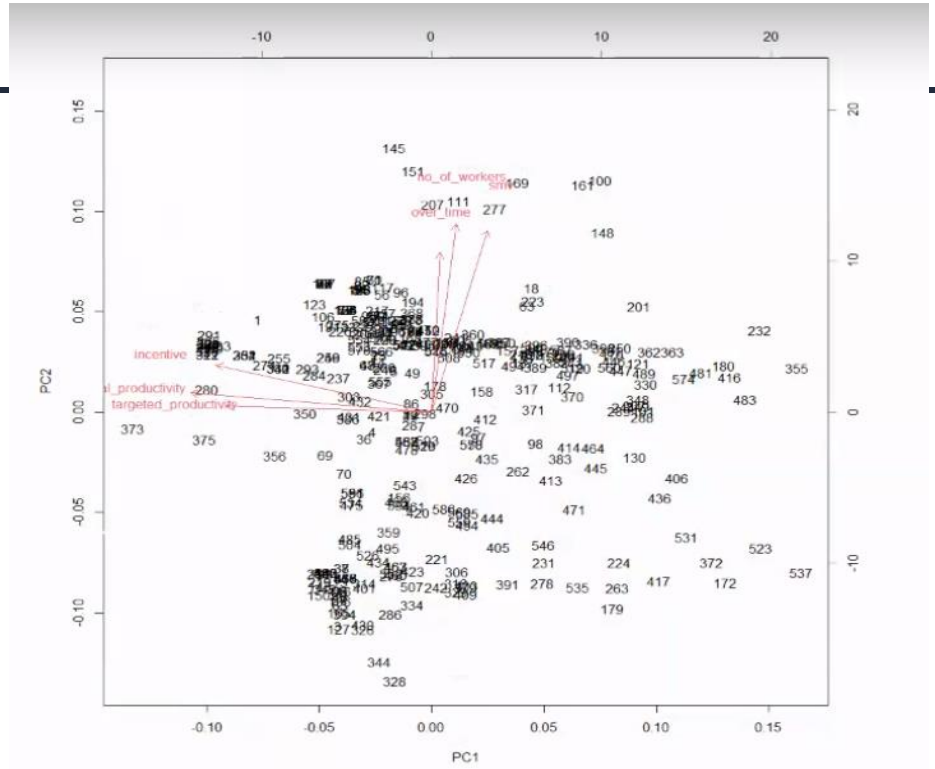
Analyse en Composantes Principales

QUESTION N°7

Calculez la variance de chaque variable and interprétez les résultats. Vous semble-t-il nécessaire de standardiser les variables avant de réaliser une ACP sur ce jeu de données ? Pourquoi ?

Nous pensons qu'il est nécessaire de standardiser les variables avant d'effectuer l'ACP pour ce jeu de données car les variables n'ont pas les mêmes dimensions.

Variable	Valeur	Description
Variance team	12.59	Normal +
Variance t_p	0.0096	Très petit
Variance smv	52.4024	Normal +
Variance wip	2451472	Très grand
Variance over_time	8398011	Très grand
Variance incentive	787.6156	Grand
Variance idle_time	283.40	Normal ++
Variance idle_men	17.99	Normal +
Variance no_of_style	0.25	Petit
Variance ino_of_workers	91.87	Normal +



QUESTION N°8

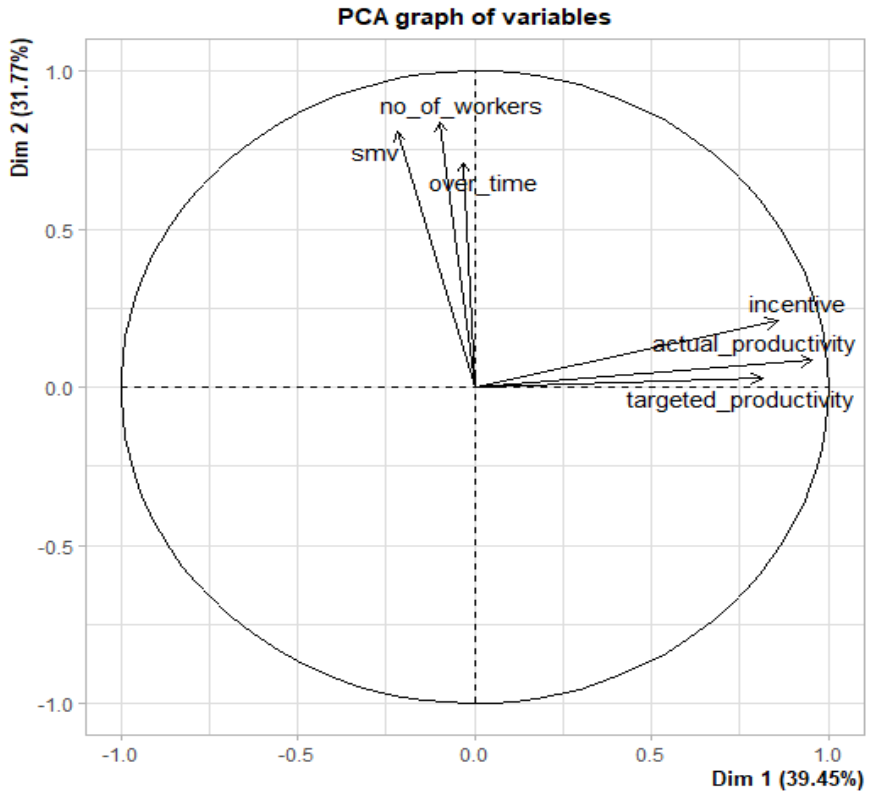
Appliquez une ACP en considérant uniquement les variables suivantes : *targeted_productivity, smv, over_time, incentive, no_of_workers* et *actual_productivity*. Normalisez préalablement les variables sélectionnées ci besoin (cf question 7). Affichez les deux premières composantes principales sous forme d'un nuage de points pour visualiser les résultats. Commentez.

La deuxième composante principale montre une forte association avec « no_of_workers », « smv » et « over_time ». Il mesure principalement le temps passé par équipe pour la réalisation d'une tâche. La première composante principale montre une forte association avec « incitation », « productivité_réelle » et « productivité_ciblée ». Il mesure la capacité et la motivation à atteindre des objectifs.

Analyse en Composantes Principales

QUESTION N°9

Affichez et commentez le cercle de corrélation.

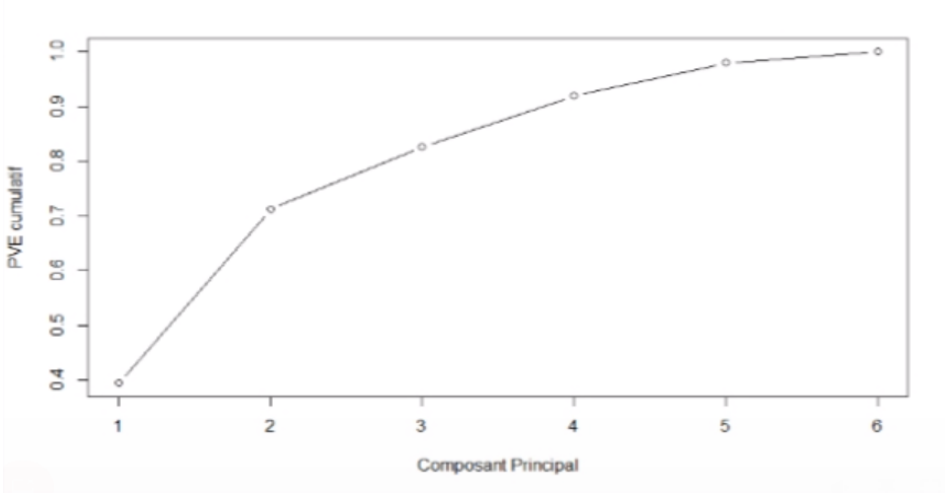


QUESTION N°10

Que représente le Pourcentage de Variance Expliquée (PVE) par une composante principale ? Calculez le PVE par chaque composante, ainsi que le PVE cumulatif. En déduire les composantes principales retenues.

$$PVE_m = \frac{\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}$$

```
> PVE
[1] 0.39445913 0.31770579 0.11335943 0.09369037 0.06056467 0.02022061
> cumsum(PVE)
[1] 0.3944591 0.7121649 0.8255244 0.9192147 0.9797794 1.0000000
```



Regression Linéaire

QUESTION N°11

Question théorique : nous souhaitons appliquer une régression linéaire pour prédire Y connaissant X. Quelle est le lien entre r(X, Y), le coefficient de corrélation entre X et Y, et le coefficient de détermination R2 obtenu en ajustant le modèle ? Quel est l'intervalle de valeurs admissibles pour R2 et pour r ?

Le lien reliant r et R2 est que $R2 = r^2$ et prenne comme intervalle [0;1]

Regression Linéaire

QUESTION N°12

Quelles valeurs de β0 et β1 ont été prédites par le modèle ? Donnez une interprétation visuelle. Quelle est la valeur moyenne attendue de l'estimation de l'erreur ε ? Vérifiez expérimentalement.

```
> #12)
> prod.regsimple$coefficients #val b0 et b1
(Intercept)           x
-0.06468112  1.08886089
```

```
> e=prod.regsimple$residuals #ttes les val de e
> mean(e) #-2.285572*10^-18
[1] -2.285572e-18
```

Regression Linéaire

QUESTION N°13

Calculez l'intervalle de confiance à 90% pour β0 et β1. Interprétez les résultats

```
> #13)int conf 90%
> confint(prod.regsimple,1,0.90) #int de b0
                   5 %         95 %
(Intercept) -0.13553 0.006167787
> confint(prod.regsimple,2,0.90) #int de b1
      5 %     95 %
x 0.99219 1.185532
```

Regression Linéaire

QUESTION N°14

Evaluez l'hypothèse de pente nulle pour le coefficient β1 et concluez sur l'existence d'une relation entre la productivité actuelle et ciblée. Le coefficient β1 est-il significativement non nul ?

Les hypothèses que l'on teste sont : $H_0 : b = 0$ contre l'hypothèse alternative $H_1 : b \neq 0$.

La statistique de ce test est donnée par $t = \frac{\widehat{b}-0}{\sqrt{var(\widehat{b})}}$, où $var(\widehat{b}) = \frac{\widehat{\sigma}_\varepsilon^2}{\sum_{i=1}^{n}(x_i-\bar{\mathbf{x}})^2}$ avec $\widehat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^{n}\widehat{e}_i^2}{n-2}$.

On peut montrer [2], que sous H_0, cette statistique suit une loi de Student à $n-2$ degrés de liberté. Ainsi, si $|t|$ est plus grand que le quantile d'ordre $1-\alpha/2$ (α étant le risque de première espèce que l'on s'est fixé, en général égal à 0.05) d'une loi de Student à $n-2$ degrés de liberté, alors on rejette H_0 : on dit alors que la pente est significativement non nulle (au risque α) et il existe un lien entre les deux variables **y** et **x**. Cela ne signifie pas pour autant que le modèle linéaire estimé soit le bon modèle à utiliser ni le seul.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.06468    0.04296  -1.506    0.133
x            1.08886    0.05861  18.577   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1069 on 344 degrees of freedom
Multiple R-squared:  0.5008,    Adjusted R-squared:  0.4993
F-statistic: 345.1 on 1 and 344 DF,  p-value: < 2.2e-16
```

Regression Linéaire

QUESTION N°15

Quelle est la valeur du coefficient de détermination R2 ? Interprétez ce résultat. Ce modèle est-il approprié pour prédire la productivité actuelle ?

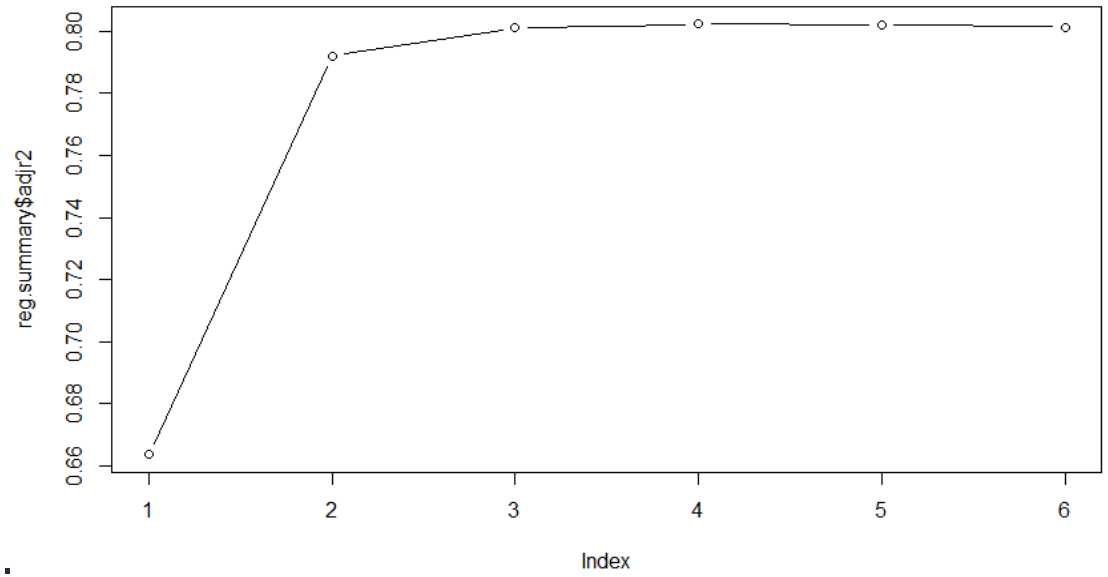
```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.49891 -0.00616 -0.00120  0.04373  0.29408

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.06468    0.04296  -1.506    0.133
x            1.08886    0.05861  18.577   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1069 on 344 degrees of freedom
Multiple R-squared:  0.5008,    Adjusted R-squared:  0.4993
F-statistic: 345.1 on 1 and 344 DF,  p-value: < 2.2e-16
```

R2 = 0.5008

Regression Linéaire

QUESTION N°16

Appliquez une régression linéaire entre la variable cible actual_productivity et chaque combinaison possible de variables parmi les 6 proposées. Affichez le coefficient de détermination ajusté R̄2 en fonction des variables sélectionnées. Sélectionnez alors le modèle pour lequel le coefficient de détermination ajusté R̄2 est maximal.




```
targeted_productivity
smv
wip
over_time
incentive
no_of_workers
```

Regression Linéaire

QUESTION N°17

Quelles sont les variables sélectionnées ?

```
coef(select_var,4)
       (Intercept) targeted_productivity                 smv           incentive         no_of_workers
      0.1371719498          0.6345878080       -0.0026810188       0.0032581455          0.0008124622
```

Regression Linéaire

QUESTION N°18

Pourquoi est-il préférable d'utiliser le coefficient de détermination ajusté plutôt que le coefficient de détermination en comparant deux modèles avec un nombre différents de variables ?

Le coefficient de détermination ajusté est plus fidèle que le coefficient de détermination de plus, le coefficient de détermination simple ne peut pas montrer qu'elle variable de régression est plus importante qu'une autre ettest les differentes variables independante ce que ne fait pas le coefficient simple.

Regression Linéaire

QUESTION N°19

Quelle sont les valeurs des coefficients estimés pour le modèle sélectionné ? Commentez. Que vaut la somme des carrés des résidus (RSS en anglais) ?

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.1371719  0.0367476   3.733 0.000222 ***
smv                     -0.0026810  0.0006214  -4.314  2.1e-05 ***
targeted_productivity    0.6345878  0.0426190  14.890  < 2e-16 ***
incentive                0.0032581  0.0001504  21.664  < 2e-16 ***
no_of_workers            0.0008125  0.0004759   1.707 0.088677 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06717 on 341 degrees of freedom
Multiple R-squared:  0.8046,    Adjusted R-squared:  0.8023
F-statistic:   351 on 4 and 341 DF,  p-value: < 2.2e-16
```

Regression Linéaire

QUESTION N°20

Pour le modèle sélectionné, testez l'hypothèse de pente nulle en excluant β0 et conclure.

Comme pour la regression simple notre hypothese est rejete mais maintenant on a un R2 equivalent a 80% ce qui est tres fiable

```
p-value: < 2.2e-16
```

Conclusion