Exercice de Statistiques Descriptives

Martin Duguey[1]

30 novembre 2021

Résumé

L'objet de ce projet est, à partir de données répertoriées, de permettre à un fabriquant de pièces d'identifier les composants sur lesquels il devrait s'interroger. Les données sont présentées en deux tableaux, l'un listant les retours en atelier et l'autre traitant de l'impact financier. L'enjeu de ce problème est d'utiliser à bon escient les notions vues en cours de Statistiques Descriptives avec M. LACAILLE tout en rendant compte et en mettant en lumière les informations du jeu de données permettant de répondre à la question que se pose le fabriquant : Quels composants peuvent-être remis en cause pour réduire les dépenses liées aux réparations ?

Introduction

Dans un contexte industriel, un fabriquant de pièces souhaitant minimiser ses frais liés aux réparations, se pose la question de savoir s'il y a des composants qui tendent à accentuer ces dépenses. L'idée est donc de mener l'étude statistique la plus complète possible pour tenter d'apporter une réponse à cette question. Aussi, pour ce faire, nous nous appuierons, sur les notes du cours de Statistiques Descriptives (voir [1]) et sur les ressources fournis entre autre par M. LACAILLE. Dans un premier temps, on s'intéressera aux variables fournies par le jeu de données afin d'en tirer les premières tendances pour, dans un second temps, cibler les variables explicatives et modéliser leurs dépendances. Enfin dans un dernière partie, nous utiliserons des outils d'analyse factorielle pour identifier les composants et fournir des éléments de réponse au fabriquant.

1 Jeu de données et variables

Le jeu de données est composé de deux tableaux : un tableau Facteurs et un tableau Cout. Dans cette partie, on s'intéresse en détails aux données qu'ils contiennent et aux premières tendances que l'on peut exploiter. Le but est ici de mettre en lumière un certain nombre d'informations pour élaborer le développement de l'étude.

1.1 Présentation du jeu de données

1.1.1 Le tableau Facteurs

Ce tableau contient les informations concernant les pièces rapportées à l'atelier. Pour chaque pièce, on retrouve donc son identifiant noté **Id**, stocké sous la forme d'un entier, la version de la pièce notée **Version**, stockée sous la forme d'une chaîne de caractère. Il y a, pour chaque pièce, trois versions possibles : $v1$, $v2$ ou $v3$. On retrouve aussi dans cette table les durées de fonctionnement des pièces pour chacun des deux modes, notées respectivement **D1** et **D2**. Aussi pour chaque pièce retournée à l'atelier, une **Panne** sour forme de caractère lui est associée. On compte neuf pannes répertoriées dans ce tableau : A, B, C, D, E, F, G, H, I. Enfin, les dernières colonnes correspondent au diagnostic des composants. C'est à dire que, pour chaque pièce retournée à l'atelier, on associe l'entier 1 au diagnostic Xi, $\forall i \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15\}$, si on repère un endommagement du composant i et sinon on lui associe l'entier 0. Il y a donc autant de diagnostics qu'il y a de composants, donc dans la suite, le terme Xi pourra être utilisé, par abus de langage, pour désigné le composant i.

On constate que le tableau contient les informations de 451 pièces différentes. Et une pièce est identifiée par un **Id** et une **Version**.

1. Sup Galilée, Ingénieurs M.A.C.S, Institut Galilée, Université Sorbonne Paris Nord, F-93430, Villetaneuse, France

1.1.2 Le tableau Cout

Ce tableau contient les frais de réparation, noté **Cout**, pour un identifiant, noté **Id**, et une **Panne** donnés. On remarque cependant qu'il n'y a que 75 couples (**Id**, **Panne**) répertoriés dans ce tableau, et plusieurs frais différents peuvent être associés à un même couple. Par ailleurs, les couts sont, pour un **Id**, et une **Panne** donnés, indépendants de la **Version** du tableau Facteurs. En plus de cela, toutes les pannes ne sont pas répertoriées dans ce tableau Cout. En l'occurence, aucune information n'est donnée concernant les frais de réparation, quelque soit la pièce, de la panne H.

Dans la suite, le choix a été fait de travailler avec un tableau Cout un peu différent, on a choisi, lorsque le couple (**Id**, **Panne**) était identique d'une ligne à l'autre, de sommer les frais de réparations. On obtient alors un tableau Cout contenant les dépenses associées à 66 couples (**Id**, **Panne**), tous différents.

Remarque 1.1. *Les deux tableaux étudiés ne présentent aucune valeur manquante, même si d'après leur taille, il est possible qu'un couple* (***Id***, ***Panne***) *de le tableau Facteurs ne soit pas répertorié dans le tableau Cout.*

1.2 Tendance des données

1.2.1 Une première tendance

Dans un premier temps, on peut essayer d'observer pour chaque composant, le nombre de pannes pour lesquelles on a constaté un endommagement, le tout à partir du tableau Facteurs. On a alors tracé la figure 1.

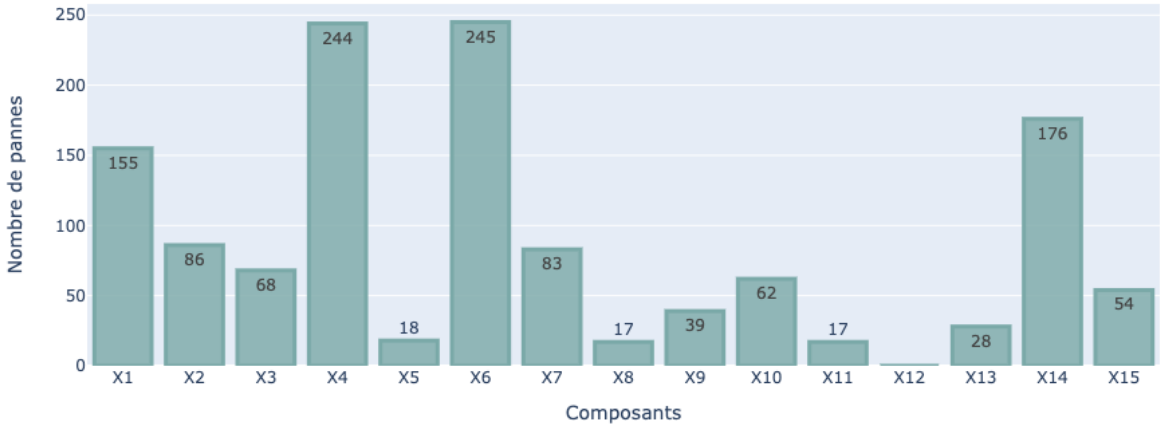

FIGURE 1 – Nombre de pannes par composant.

Observations et interprétation de la figure 1

On remarque qu'il y a quatre composants qui semblent être endommagés fréquemment : X1, X4, X6 et X14. Cependant, on ne peut rien en dire puisqu'aucune variable n'a été prise en compte. Ce graphe sert essentiellement à cibler les composants qui vont éventuellement jouer un rôle dans notre étude. L'idée est maintenant de trouver des variables qui peuvent expliquer ces occurences.

1.2.2 Variables Cout et Durée

À partir des données dont on dispose, on constate qu'à priori les variables **D1**, **D2** du tableau Facteurs et la variable **Cout** du tableau Cout, vont jouer un rôle dans notre étude. On définit la variable **Durée** = **D1** + **D2**, et on choisit d'étudier à la fois la variable **Cout** et la variable **Durée** et d'estimer leur loi.

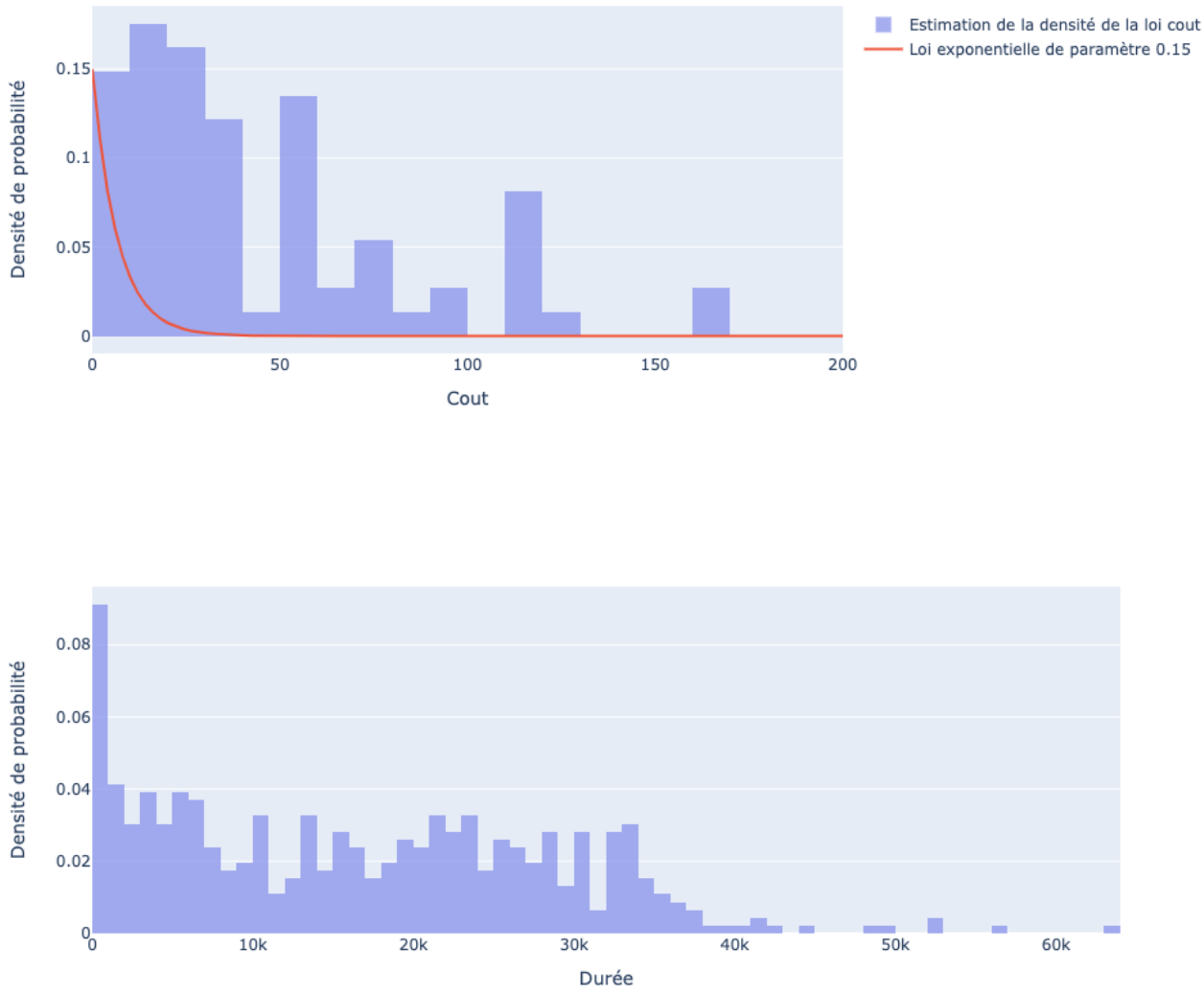

FIGURE 2 – Estimation par histogramme des densités des variables **Cout** et **Durée**.

Observations et interprétation de la figure 2

Les histogrammes nous permettent de voir que les lois des variables **Cout** et **Durée** ne sont à priori pas uni-modale. Une comparaison a été faite pour la variable **Cout** avec la densité de la loi exponentielle ε de paramètre $\lambda = 0.15$. On remarque aussi un nombre important de **Durée** très courte dans la distribution et on peut supposer que le support de la distribution se situe dans l'intervalle $[0; 40k]$. Dans la suite, on ne va pas chercher à plus préciser ces distributions. Néanmoins, on peut se poser les questions suivantes : existe-t-il un lien entre le nombre de pannes dans lequel un composant est impliqué et les coûts de réparation associés à ses pannes ? Existe-t-il un lien entre le nombre de pannes dans lequel un composant est impliqué et la durée d'utilisation des pièces auxquelles il est associé ?

2 Modélisation autour des variables explicatives Cout et Durée

Pour pouvoir étudier ces variables, on va utiliser la variable **Occurences** définie dans la partie 1.2.1. C'est-à-dire qu'on associe à chaque composant le nombre de fois où on a constaté un endommagement dans le tableau Facteurs.

2.1 Tendance et modélisation du lien entre les variables Cout et Occurences

2.1.1 Lien entre Cout et Occurences

On rappelle que l'on souhaite, dans l'idéal, identifier des composants qui potentiellement augmentent les dépenses liées au réparations du fabriquant. Pour ce faire, on associe à chaque composant la somme totale des réparations des pièces pour lesquelles le composant a été diagnostiqué comme endommagé. Pour ce faire, on va étudier uniquement les couples (**Id**, **Panne**) répertoriés dans le tableau Cout. Il faut donc travailler sur un tableau Facteurs réduit aux couples (**Id**, **Panne**) présents sur les deux tableaux. Les résultats obtenus sont tracés sur la figure 3 et sont ordonnés par la valeur de la variable **Occurence**.

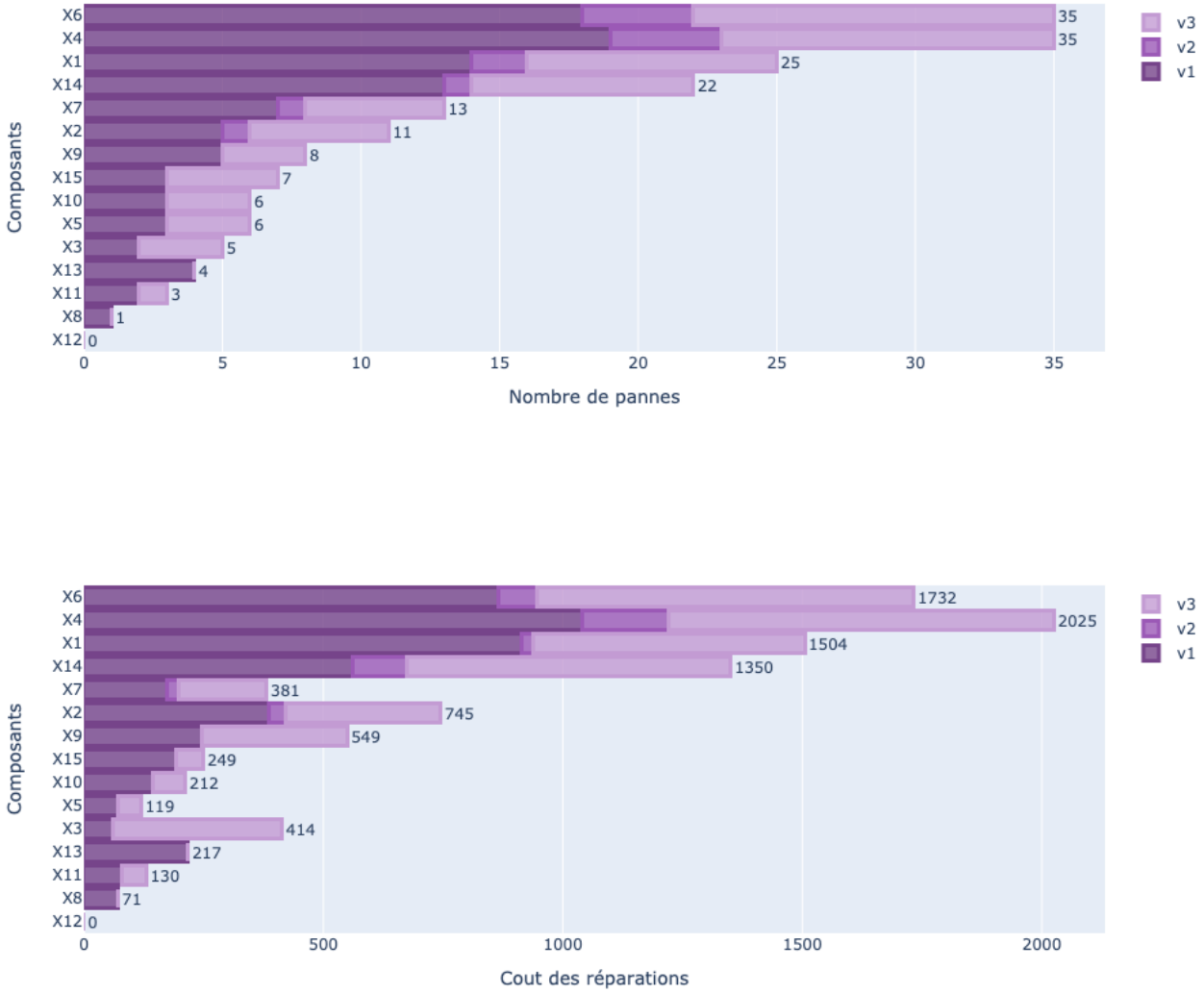

FIGURE 3 – Diagramme représentant le nombre de pannes et le cout des réparations par composant.

Observations et interprétation de la figure 3

Si on regarde le diagramme des coûts, la tendance annoncée par la figure 1, qui est reprise ici sous une autre forme et à partir d'un jeu de données réduit par le diagramme du nombre de pannes, semble se confirmer. En effet, comme on a ordonné les résultats par rapport à la variable **Occurence**, on constate que les coûts suivent la même décroissance, à quelques exceptions près comme les composants X6, X7, X5 et X3. Mais d'une façon générale, les composants répertoriés comme endommagés dans le plus de pannes sont aussi liés aux coûts de réparation les plus élevés.

De la même façon, les coûts liés aux composants X1, X4, X6 et X14 semblent se détacher très nettement, en terme de valeurs, aux coûts liés aux autres composants. Mais puisque cette tendance au niveau des coûts respecte, à quelques exceptions près, la tendance observée sur les occurences, on peut se demander si la dépendance entre la variable **Cout** et la variable **Occurence** peut être modélisé par une relation linéaire ?

2.1.2 Modèle Cout - Occurences

Pour chaque composant, on trace son point correspondant dans le repère **Occurence** - **Cout**, et on obtient notre droite modèle par régression linéaire (voir la figure 4).

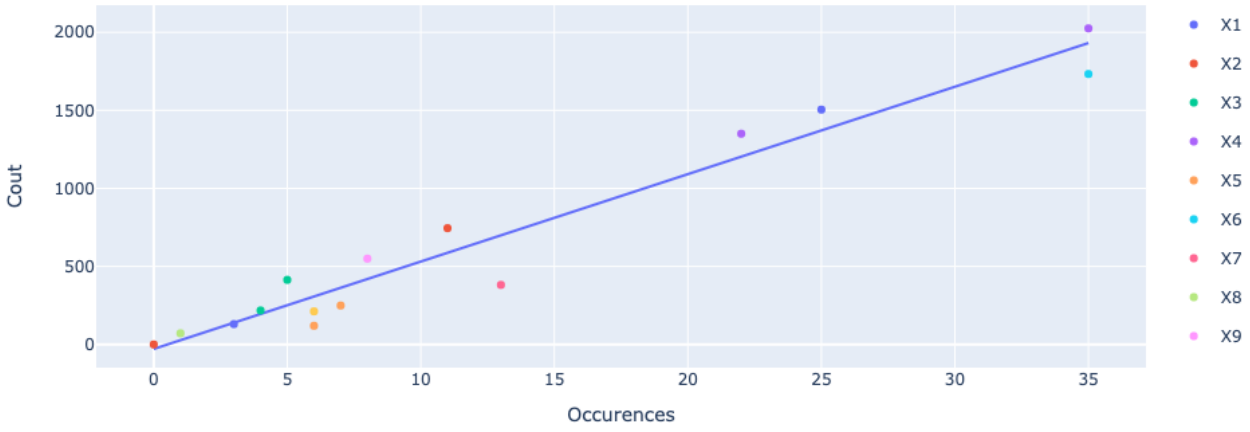

FIGURE 4 – Modèle linéaire pour représenter la dépendance entre les variables **Occurence** et **Cout**.

$$\text{Modèle obtenu :} \quad y(x) = ax + b \quad \text{avec } a = 55.9946 \text{ et } b = -28.7216$$

Observations et interprétation de la figure 4

D'un point de vue graphique, on observe que le modèle linéaire paraît approprié par rapport à la répartition des points dans le plan **Cout** - **Occurences**. On observe aussi qu'il semble y avoir trois groupes parmi les composants, en l'occurence un groupe composé de X4 et X6, plutôt dans les valeurs hautes du plan, un second groupe composé de X1 et X14 autour des valeurs centrales et d'un dernier groupe composé du reste des composants, situé dans des valeurs basses. Par ailleurs, la régression linéaire nous donne la valeur p[1] du modèle, obtenue, d'après la documentation **scipy** par test de Wald. Cette valeur p est exactement la probabilité de rejeter à tort le modèle constant (avec a = 0). Nous obtenons pour ce cas, une valeur p de l'ordre de $8.35e^{-10}$. En plus de cela, comme nous sommes dans le cas univarié (on tente d'expliquer la variable **Cout** par la variable **Occurence**), le coefficient de détermination R^2 est exactement égale au carré de la corrélation des deux variables. On obtient dans notre cas, un R^2 de l'ordre de 0.94. Ainsi, notre modèle explique 94% de la variance de nos données. Maintenant avec ce modèle, on justifie simplement le fait que les composants qui sont associés à des frais de réparation importants sont ceux qui sont les plus souvent endommagés, ce qui est logique. Mais est-ce que les composants qui sont les plus souvent endommagés sont aussi ceux qui sont le plus utilisés ?

2.2 Tendance et modélisation du lien entre les variables Durée et Occurences

2.2.1 Lien entre Durée et Occurences

On réitère l'opération en étudiant cette fois les variables **Occurences** et **Durée**. On fait cette fois-ci l'étude sur le tableau Facteurs en entier, puisque la notion de coût ne rentre pas en compte ici. Les résultats obtenus sont tracés sur la figure 5 et sont ordonnés par la valeur de la variable **Occurence**.

1. En anglais, p-value.

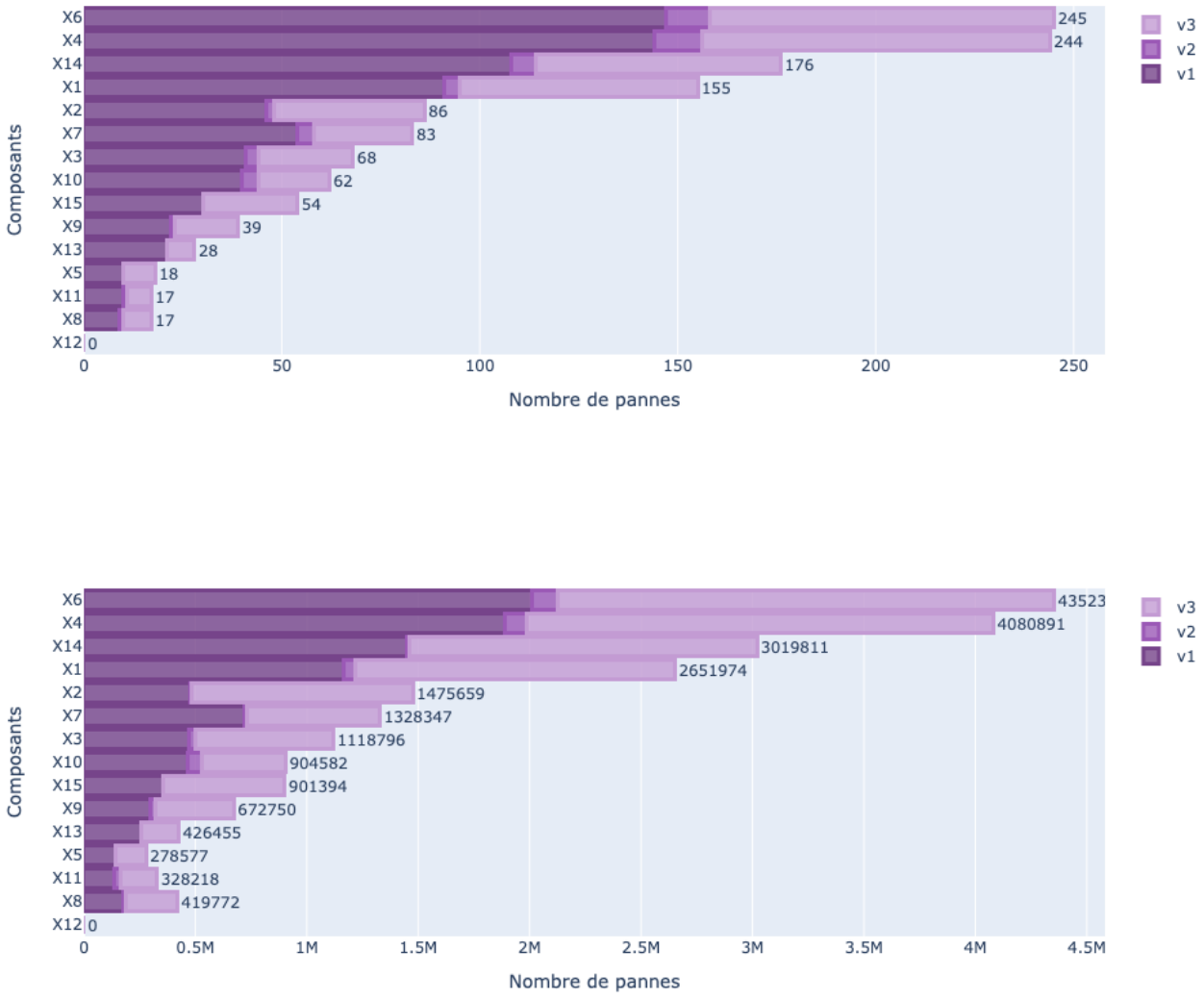

FIGURE 5 – Diagramme représentant le nombre de pannes et la durée d'utilisation par composant.

Observations et interprétation de la figure 5

Sur la figure 5, on voit bien que sans exception, et puisque les deux diagrammes sont ordonnés de la même façon, plus un composant est utilisé plus il est associé à un nombre important de pannes. Là encore, on peut se demander si la dépendance entre la variable **Durée** et la variable **Occurence** peut être modélisée par une relation linéaire ?

2.2.2 Modèle Durée - Occurences

Pour chaque composant, on trace son point correspondant dans le repère **Occurence** - **Durée**, et on obtient notre droite modèle par régression linéaire (voir la figure 6).

$$\text{Modèle obtenu :} \quad y(x) = \alpha x + \beta \quad \text{avec } \alpha = 17221.93 \text{ et } \beta = -19407.78$$

Observations et interprétation de la figure 6

Visuellement, là encore on constate une répartition plutôt linéaire des points, ce qui laisse penser que le choix d'un modèle linéaire n'est pas si mauvais. On observe toujours les trois mêmes groupes que sur la figure 4.

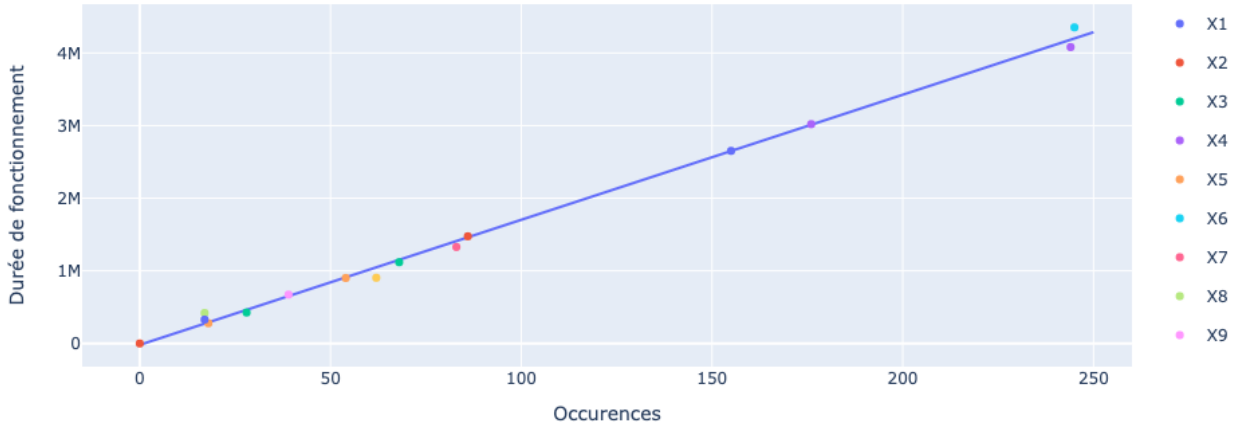

FIGURE 6 – Modèle linéaire pour représenter la dépendance entre les variables **Occurence** et **Durée**.

De la même façon que pour le modèle **Cout** - **Occurence**, la valeur p est très faible, de l'ordre de $1.45e^{-17}$, et le coefficient de détermination R^2, qui n'est autre que le carré du coefficient de corrélation (cas univarié), est de l'ordre de 0.997. On peut en conclure que le modèle semble approprié et qu'il explique 99% de la variance de nos données. On vient donc confirmer le fait que plus un composant est utilisé, plus il est susceptible d'être endommagé, ce qui est là encore logique. Alors comment identifier le(s) composant(s) qui pose(nt) problème(s) ?

3 Identification des composants et proposition de réponse

3.1 Identification et classification des composants

On a montré jusqu'à présent que les variables **Occurence**, **Cout** et **Durée** expliquent bien nos données. L'idée est maintenant, en utilisant une méthode d'analyse factorielle, de proposer une classification des composants. Pour ce faire, on utilisera d'abord une méthode d'analyse en composante principale (ACP) puis une méthode de valeurs moyennes[2] pour classer les composants.

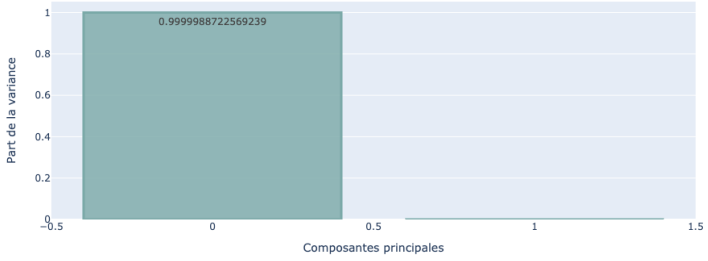

FIGURE 7 – Part de la variance expliquée par composante.

Grâce à l'ACP, on explique, par le biais de deux composantes principales, 99.99% de la variance de nos données regroupant les variables **Occurence**, **Cout** et **Durée**.

En projetant les données de l'ACP dans le plan CP1 - CP2, et en comparant avec celle obtenue par méthode des valeurs moyennes, on peut identifier des groupes de composants. En réalité, on a réalisé deux projections

2. Méthode k-means.

pour la méthode des valeurs moyennes : une avec deux classes, et une avec trois classes. Les résultats obtenus sont visibles sur la figure 8.

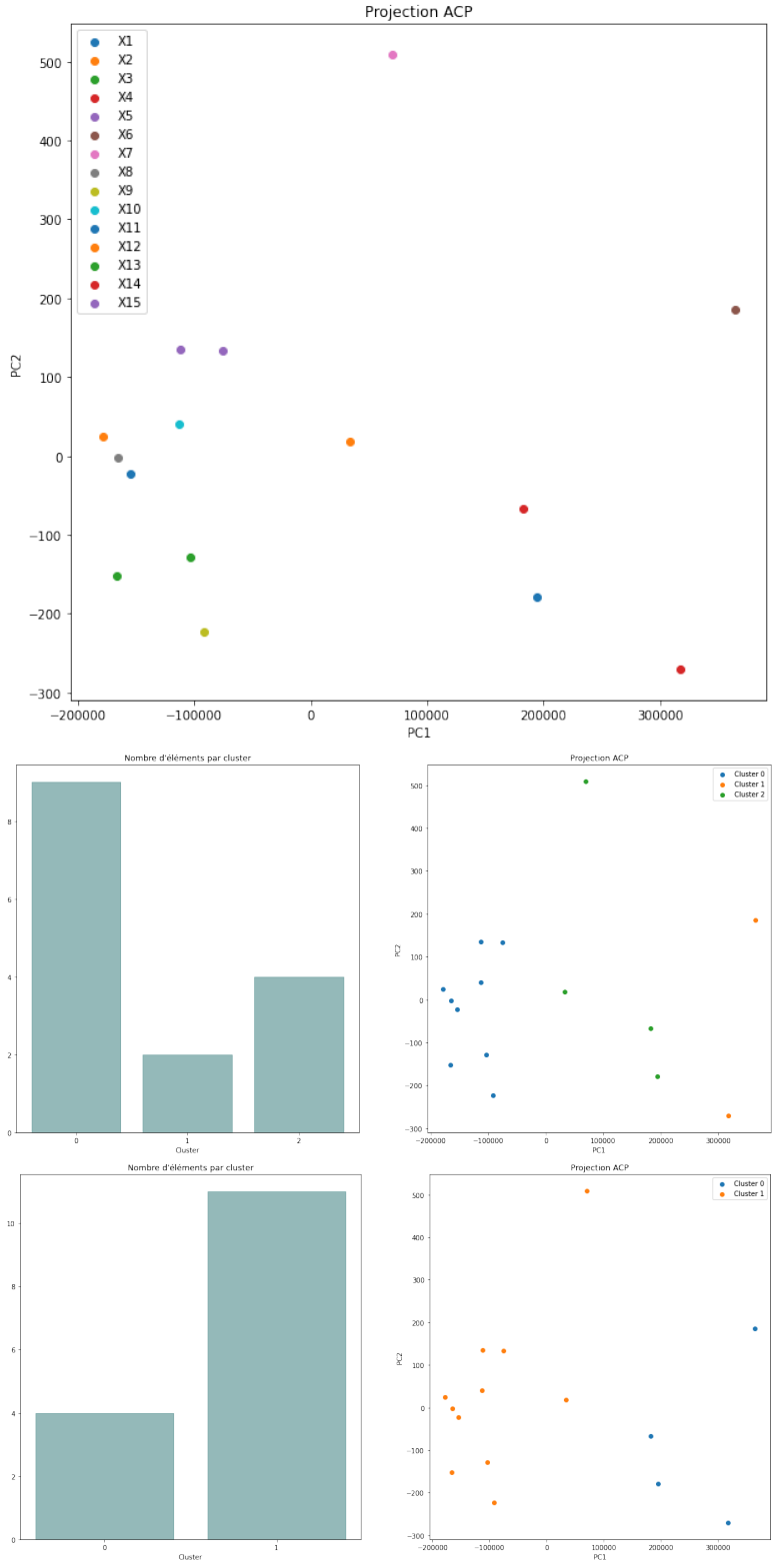

FIGURE 8 – Projection et classification

Observations et interprétation de la figure 8

Avec trois classes, on a dans une classe les composants X4 et X6, dans une autre X7, X2, X14 et X1 et enfin le reste dans une dernière classe. C'est exactement pour les composants cités, ceux qui sont associés au plus

grand nombre de pannes d'après la figure 3. Mais on ne retrouve pas les trois groupes que l'on avait identifié sur les figures 4 et 6. En fait, on peut dire la même chose lorsque l'on classe les composants selon deux entités. On peut donc dire que ces projections, à elles seules, ne permettent pas de proposer une réponse au fabriquant. Donc comment peut-on répondre au besoin du fabriquant ?

3.2 Proposition de réponse

Après avoir observé que les variables **Occurence** et **Cout** et **Occurence** et **Durée** sont liées par une relation linéaire, on peut très bien supposer que les variables **Cout** et **Durée** sont aussi liées par une relation linéaire. À partir des données, on obtient les résultats tracés sur la figure 9.

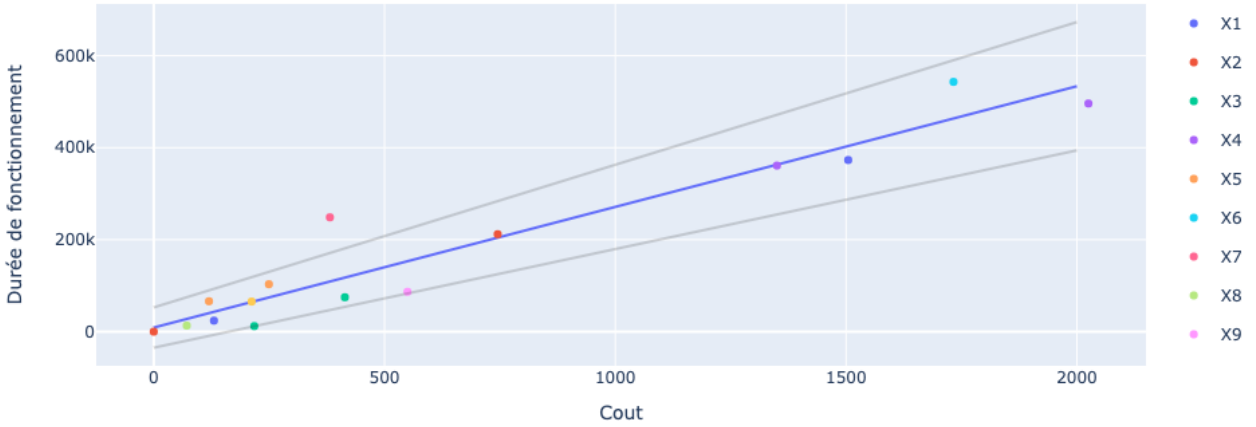

FIGURE 9 – Modèle linéaire pour représenter la dépendance entre les variables **Cout** et **Durée**.

$$\text{Modèle obtenu :} \quad y(x) = Ax + B \quad \text{avec } A = 262.4 \text{ et } B = 8902.2$$

Observations et interprétation de la figure 9

Là encore, la répartition des points semblent se prêter à une modélisation linéaire. Pour l'attester, la valeur p est de l'ordre de $2.44e^{-8}$ et le coefficient de détermination R^2 est de l'ordre de 0.92, donc une large partie des données est expliquée par le modèle. À partir de ce dernier et grâce aux bornes de confiance à 95%, on peut visuellement observer les composants qui ont un rendement peu acceptable. En particulier, on pourraît cibler les composants X9 et X13 qui sont sur la frontière basse de l'intervalle de confiance à 95% du modèle linéaire. En effet, les coûts de réparation associés à ses composants sont très élevés par rapport aux composants X10, X5, X15 pour une durée d'utilisation similaire.

Conclusion

Après avoir étudier le jeu de données, on pourrait fournir au fabriquant la figure 9, et cibler les composants X9 et X13 comme des composants sur lesquels il faudrait s'interroger pour diminuer les dépenses liées aux réparations. Par ailleurs, on pourrait nuancer nos résultats en menant une nouvelle étude qui dissocierait cette fois la durée **D1** de fonctionnement du premier mode et la durée **D2** de fonctionnement du second mode. En effet, on peut très bien imaginer qu'il y ait un ou plusieurs composants qui soient endommagés plus facilement avec le premier mode de fonctionnement qu'avec le second et vice versa.

Références

[1] L. J., *Statistiques descriptives*. Notes de cours, 2021.