INTRODUCTION

Ce projet est dans la continuité des autres, selon cette vidéo, qui nous encourage développer de nouveaux savoir et base sur différentes technologies pour le métier de Data Analyst.

J'aime apprendre et me perfectionner, et mon envie d'entrer dans le secteur de la data me motive à poursuivre cet objectif. Cette première expérience représente un premier pied dans ce domaine, et je suis convaincu que chaque projet est une nouvelle occasion de progresser. Avec des outils comme l'IA à notre disposition, se former devient plus accessible et puissant, à condition de savoir les utiliser à bon escient.

OBJECTIF

Une entreprise souhaite analyser la répartition des performances des employés pour comprendre les écarts et identifier les outliers.

Nous souhaitons étudier les distributions des scores de performance et des heures travaillées pour détecter les facteurs d'amélioration.

PRESENTATION DU DATASET

Pour se faire, voici le détail de la datasheet :

Colonnes d'identification:

- Employee_Name (Texte) - Nom de l'employé
- EmpID (Numérique/Texte) - Identifiant unique de l'employé

Informations démographiques:

- DOB (Date) - Date de naissance
- Sex (Texte) - Sexe biologique
- GenderID (Numérique) - Identifiant du genre
- MarriedID (Numérique) - Identifiant de statut marital
- MaritalStatusID (Numérique) - Identifiant du statut marital
- MaritalDesc (Texte) - Description du statut marital
- CitizenDesc (Texte) - Statut de citoyenneté
- HispanicLatino (Booléen) - Indicateur si l'employé est hispanique/latino
- RaceDesc (Texte) - Description de l'origine ethnique

Statut professionnel:

- Position (Texte) - Intitulé du poste
- PositionID (Numérique) - Identifiant du poste

- Department (Texte) - Département de travail
- DeptID (Numérique) - Identifiant du département
- EmpStatusID (Numérique) - Identifiant du statut d'emploi
- EmploymentStatus (Texte) - Description du statut d'emploi
- DateofHire (Date) - Date d'embauche
- Salary (Numérique) - Salaire de l'employé

Performance et engagement:

- PerfScoreID (Numérique) - Identifiant du niveau de performance
- PerformanceScore (Texte) - Description du niveau de performance
- LastPerformanceReview_Date (Date) - Date de la dernière évaluation
- EngagementSurvey (Numérique) - Résultats de l'enquête d'engagement
- EmpSatisfaction (Numérique) - Niveau de satisfaction de l'employé
- SpecialProjectsCount (Numérique) - Nombre de projets spéciaux
- DaysLateLast30 (Numérique) - Nombre de jours de retard sur les 30 derniers jours
- Absences (Numérique) - Nombre d'absences

Gestion:

- ManagerName (Texte) - Nom du manager
- ManagerID (Numérique/Texte) - Identifiant du manager

Recrutement et départ:

- RecruitmentSource (Texte) - Source de recrutement
- FromDiversityJobFairID (Booléen/Numérique) - Indicateur si recruté via salon de la diversité
- Termd (Booléen) - Indicateur si l'employé a quitté l'entreprise
- DateofTermination (Date) - Date de départ
- TermReason (Texte) - Raison du départ

Localisation:

- State (Texte) - État de résidence
- Zip (Texte/Numérique) - Code postal

Les données ont été réagencées selon les groupes définis précédemment, améliorant ainsi leur lisibilité et la cohérence de l'analyse.

PREPARER ET NETTOYER LES DONNEES

- Vérifier les valeurs manquantes

Lors de l'analyse des données, une première étape cruciale consiste à identifier et traiter les valeurs manquantes. Pour ce faire, nous avons utilisé des outils de détection des valeurs nulles ou

manquantes dans chaque colonne du dataset. Cela permet non seulement de comprendre l'étendue du problème, mais aussi d'appliquer des solutions appropriées, telles que la suppression des lignes concernées, l'imputation des valeurs manquantes par la moyenne ou la médiane, ou encore le remplacement par des valeurs par défaut selon le contexte.

- Standardiser les formats

La standardisation des formats est essentielle pour garantir la cohérence des données tout au long du processus d'analyse. Nous avons vérifié que toutes les colonnes respectent un format uniforme, en particulier pour les dates, les valeurs numériques, et les catégories textuelles. Cela inclut la conversion des chaînes de caractères en minuscules ou majuscules, l'unification des formats de date (par exemple, passer de "DD/MM/YYYY" à "YYYY-MM-DD") et la normalisation des valeurs numériques (par exemple, en supprimant les espaces et en ajustant les décimales).

- Créer des variables dérivées si nécessaire

Avec les dates à notre disposition, nous pouvons ajouter trois colonnes :

- AgeHired (Numérique) – Age au debut du poste
- CurrentAge (Numérique) – Age acutelle du salarié
- TenureYears (Numérique) – Ancienneté en année dans l'entreprise

Il nous était demandé d'étudier la répartition des scores selon les heures travaillées. Cependant, nous n'avons pas les données suffisantes pour déterminer des valeurs cohérentes et juste. Nous préférons ne pas ajouter cette variable car toute estimation risque d'être approximative.
Afin d'être capable d'avoir des données satisfaisantes, il nous aurait fallu des informations sur le contrat des employés, les congés, arrêtes maladies, etc…

REALISER L'ANALYSE

	CurrentAge	AgeHired	TenureYears	Salary	PerfScoreID	EngagementSurvey	EmpSatisfaction	SpecialProjectsCount	DaysLateLast30	Absences
Min	33.00	19.00	0.00	45046.00	1.00	1.12	1.00	0.00	0.00	1.00
Max	74.00	63.00	19.00	250000.00	4.00	5.00	5.00	8.00	6.00	20.00
Moyenne	46.41	34.10	8.96	69020.68	2.98	4.11	3.89	1.22	0.41	10.24
Écart-type	8.87	8.91	4.43	25156.64	0.59	0.79	0.91	2.35	1.29	5.85

Tableau 1: Distribution des données répartie selon les quartiles : Q1 (25%), Médiane (50%) et Q3 (75%) illustrant les variations et la dispersion des valeurs.

- Analyses descriptives (moyennes, médianes, distributions)

Les valeurs semblent relativement homogènes, avec une répartition équilibrée entre le mix et le max. Seul le salaire se distingue par une ou plusieurs valeurs nettement plus élevées.

On note également que l'ancienneté (*TenureYears*) minimale est à 0, ce qui peut être une donnée intéressante à analyser.

	CurrentAge	AgeHired	TenureYears	Salary	PerfScoreID	EngagementSurvey	EmpSatisfaction	SpecialProjectsCount	DaysLateLast30	Absences
Q1	39.00	28.00	5.00	55501.50	3.00	3.69	3.00	0.00	0.00	5.00
Médiane	45.00	32.00	11.00	62810.00	3.00	4.28	4.00	0.00	0.00	10.00
Q3	52.00	39.00	12.00	72036.00	3.00	4.70	5.00	0.00	0.00	15.00

Tableau 2 : Résumé des données : Moyenne, Écart-Type, Min et Max pour donner un aperçu des valeurs centrales et de la dispersion.

Ce tableau confirme noter précédente observation concernant la variabilité des salaires, avec une différence notable entre Q1 et Q3. Pour le reste, la répartition des valeurs semble relativement équilibrée.

- Analyses de corrélation (salaire vs performance, satisfaction vs absentéisme

EngagementSurvey	0.544927
EmpSatisfaction	0.303579
Salary	0.130903
TenureGroup	0.126562
TenureYears	0.116280
CurrentAge	0.079203
AgeHired	0.070090
Absences	0.046629
SpecialProjectsCount	0.045677
DaysLateLast30	-0.734728

Tableau 3 : Corrélation de « PerfScoreID »

Afin d'identifier les facteurs influençant le score de performance, nous avons effectué une analyse de corrélation entre le **perfscoreID** et différentes variables présentes dans l'ensemble des données. Cette approche nous permet de mieux comprendre quelles variables sont étroitement liées à la performance des employés et d'orienter nos recommandations pour optimiser les résultats.

Dans le tableau ci-contre, on observe une corrélation positive de 0.54 entre l'indice de performance et l'engagement, ce qui suggère une relation entre ces deux variables.

A l'inverse, la variable **'DaysLateLast30'** faisant référence au retard sur les trente derniers jours, présente une forte corrélation négative avec l'indice de performance (-0.73).

Autrement dit, si l'engagement semble avoir un lien avec la performance, la ponctualité est un facteur encore plus fortement associé.

La corrélation n'étant pas transitive, il est pertinent d'explorer plus en détail les deux autres variables figurant dans le top 3, même si leurs corrélations avec la performance sont modestes.

PerfScoreID	0.544927
EmpSatisfaction	0.187105
Salary	0.064966
CurrentAge	0.063384
AgeHired	0.057573
TenureGroup	0.032010
TenureYears	0.031125
SpecialProjectsCount	0.013227
Absences	-0.008771
DaysLateLast30	-0.585232

Tableau 4: Corrélation de « EngagementSurvey

SpecialProjectsCount	0.508333
PerfScoreID	0.130903
AgeHired	0.124498
CurrentAge	0.094360
Absences	0.082382
EngagementSurvey	0.064966
EmpSatisfaction	0.062718
TenureYears	0.036378
TenureGroup	0.034850
DaysLateLast30	-0.069443

Tableau 5: Corrélation du « Salary »

Hormis la corrélation avec le score de performance qui nous était connu, on ne constate rien qui puisse nous aider. Aucune corrélation marquante ne se dégage, ce qui signifie que son influence est soit indépendante des autres variables, soit liée à des facteurs non pris en compte dans cette analyse.

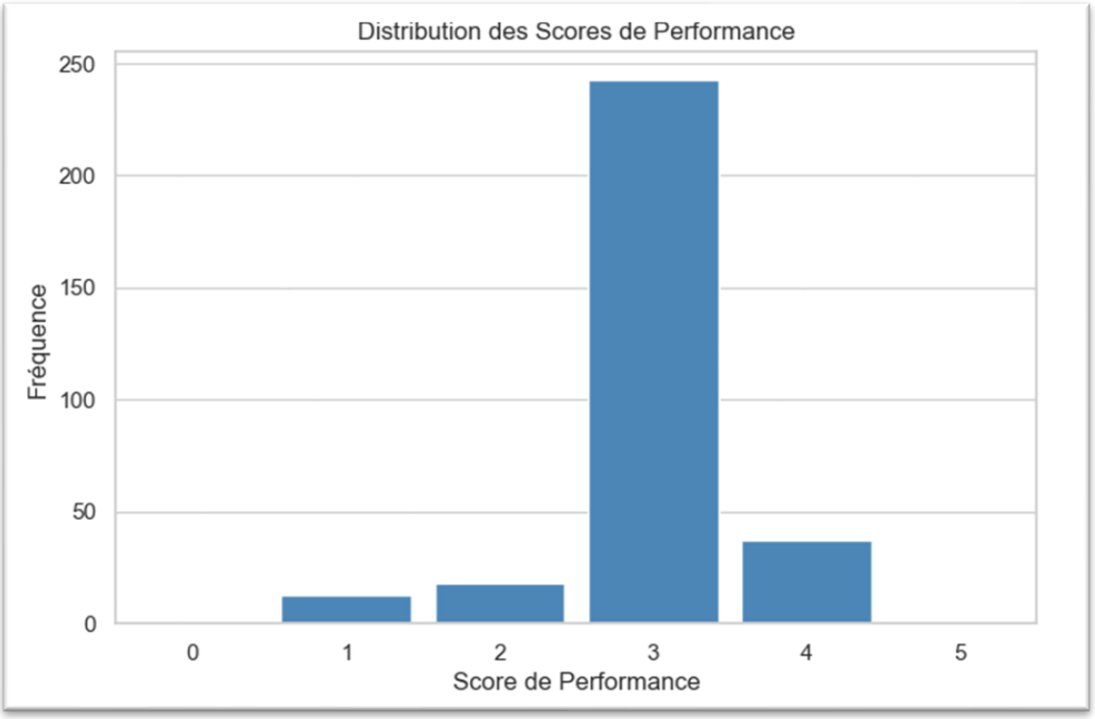
Le salaire présente une corrélation notable avec le nombre de projets spéciaux réalises (0.50), ce qui suggère un lien entre la rémunération et l'implication dans ces projets. En revanche, il n'existe aucune corrélation significative entre le salaire et l'ancienneté (**'TenureYears'**). Cela indique que la rémunération ne semble pas directement être influencer par le temps passer dans l'entreprise.

Absences	0.001833
CurrentAge	-0.053286
AgeHired	-0.058006
Salary	-0.069443
SpecialProjectsCount	-0.092494
TenureYears	-0.111766
TenureGroup	-0.124933
EmpSatisfaction	-0.235412
EngagementSurvey	-0.585232
PerfScoreID	-0.734728

Tableau 6: Corrélation de « DaysLateLast30 »

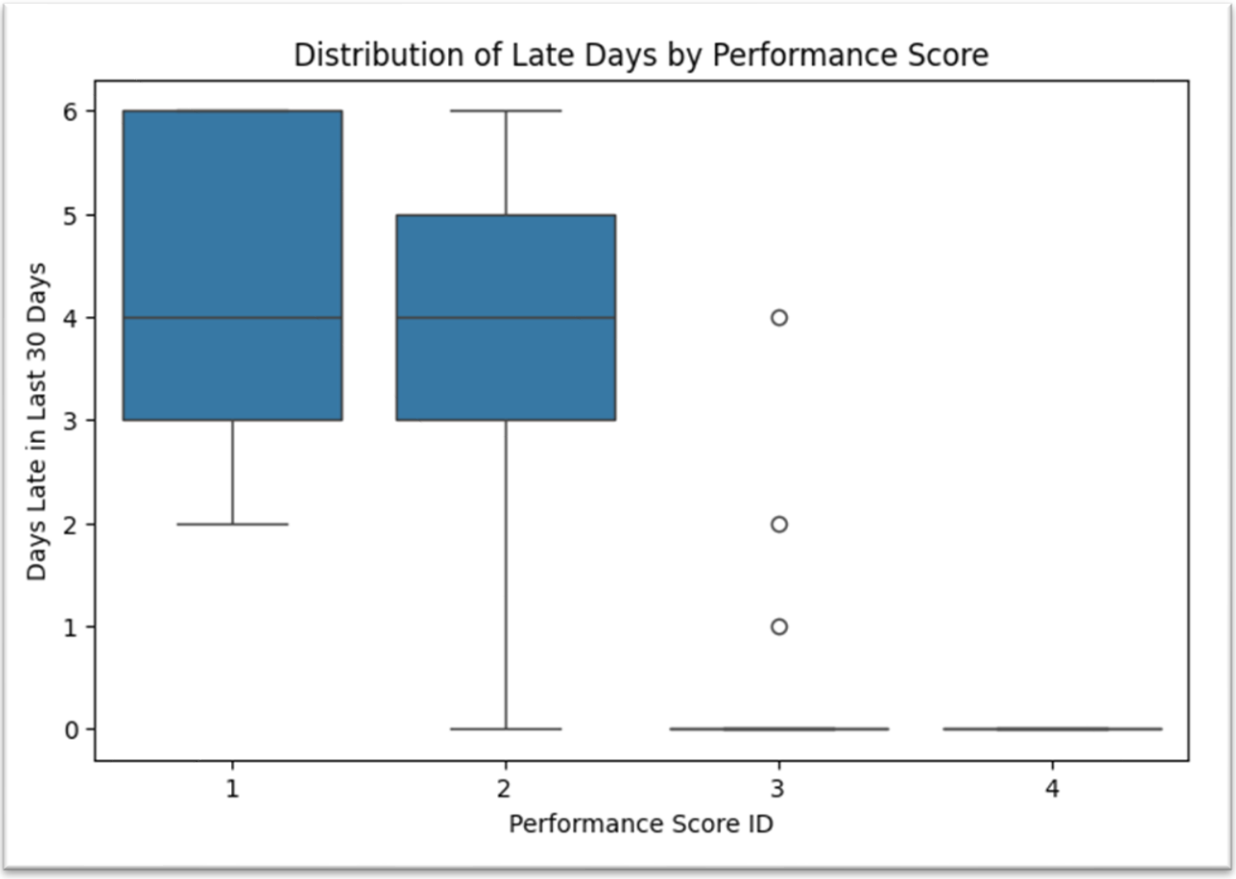
Enfin il parait pertinent d'analyser cette variable, car bien qu'elle soit la moins corrélée avec la performance, elle met en évidence un point intéressante : les personnes ayant été en retard ces 30 derniers jours semblent aussi être celles qui montrent le moins d'engagement et de satisfaction.

- Segmentations (par département, ancienneté, etc.)



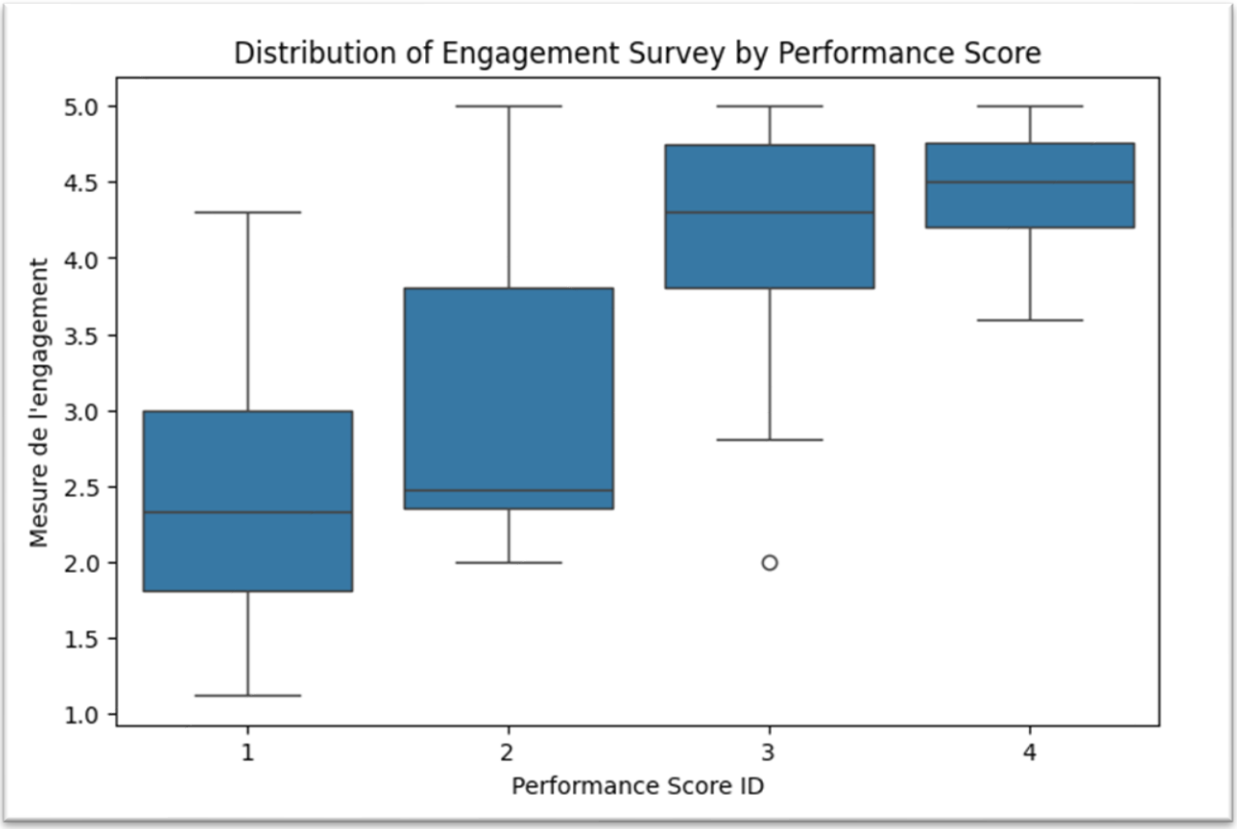
Voici un graphique illustrant la répartition du Score de Performance (PerfScoreID) au sein de l'entreprise.
On observe une concentration des valeurs autour de 3, avec peu ou aucune présence aux extrêmes.

Notre indicateur de performance étant un indicateur discret, il ne nécessite pas de segmentation supplémentaire, nous offrant une segmentation naturelle.



L’analyse montre qu’il existe une tendance claire entre le score de performance (PerfScoreID) et les retards. Plus le score est bas, plus les retards au cours des 30 derniers jours sont fréquents. A partir d’un score de 3, les retards deviennent rares et sont considérés comme des **outilers**.

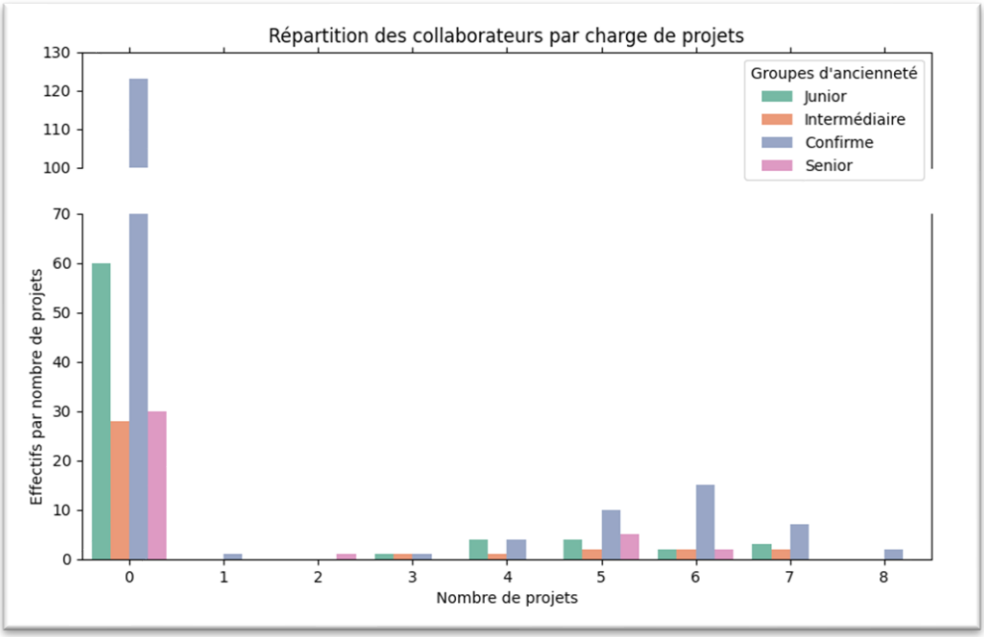
Une relation inverse et observée entre le **PerfScoreID** et **DaysLateLast30**, ce qui laisse présager d’une corrélation négative entre ces deux variables.



Ce diagramme en boite nous montre la relation entre le score de performance et l'engagement. Pour les scores de performance de **1** et **2** on constate une large dispersion des données avec des moustaches relativement longues.

Pour des scores de **3** et **4** ce qui correspond à l'écrasante majorité des employés, nous remarquons des moustaches plus petites avec des médianes situés à l'alentour de 4.5. Cette tendance montre qu'un score de performance plus élevé est généralement associé à un meilleur niveau d'engagement.

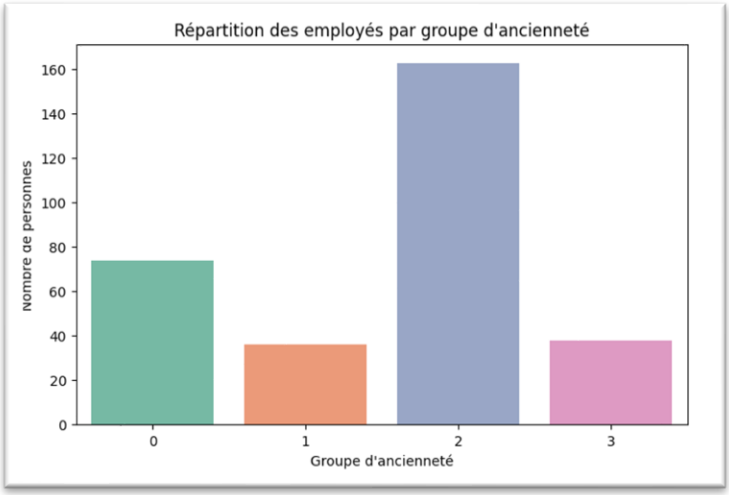
On constate également un faible nombre de valeurs abérrantes (**outlier**), ce qui suggère que les données sont cohérentes et représentatives.



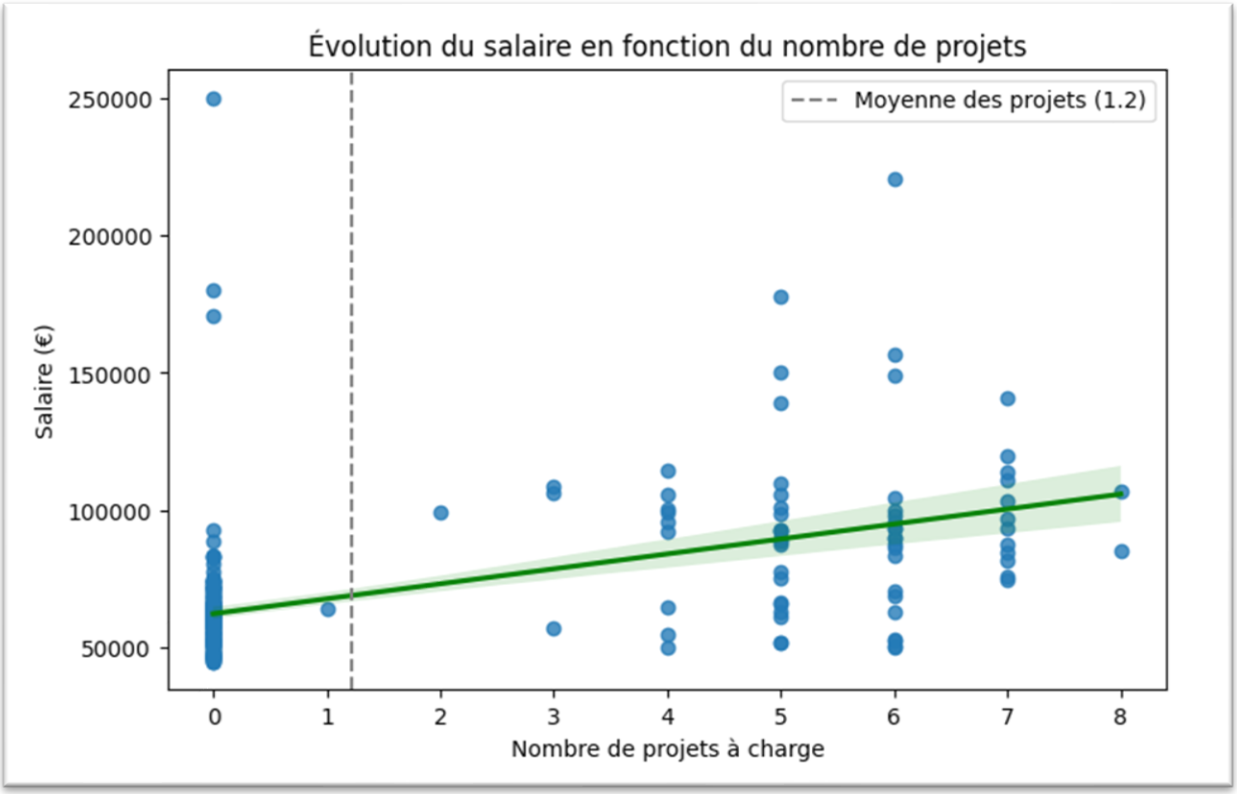
Ce graphique représente la répartition du nombre de projets à charge selon les groupes d'ancienneté au sein de l'entreprise. La segmentation a été effectuée sur la variable **TenureYears**, en utilisant la moyenne et l'écart-type afin de classer les employés en différentes catégories.

Nous observons que la distribution des projets n'est pas optimale, d'autant plus que la corrélation entre l'ancienneté et le nombre de projets est nulle (0.00), ce qui indique une absence de relation entre ces deux variables. Dans cette optique, il serait pertinent de revoir les critères d'attribution des projets afin d'assurer une répartition plus équitable, mieux alignée avec les compétences et l'expérience des employés.

Par ailleurs, le graphique ci-dessous illustre également la distribution des employés dans les différents groupes d'ancienneté, permettant ainsi d'avoir une vision globale de leur répartition au sein de l'entreprise



- Exploration graphique



Le graphique ci-dessous illustre l'évolution des salaires en fonction du nombre de projets à charge. L'objectif de cette visualisation est d'analyser si une relation marquée existe entre ces deux variables. On constate que la progression salariale en fonction du nombre de projets n'est pas significative, suggérant que la charge de travail supplémentaire n'entraîne pas nécessairement une augmentation proportionnelle de la rémunération.

Cette observation peut soulever des questions sur les critères d'évolution salariale au sein de l'entreprise et inciter à une analyse plus approfondie des facteurs influençant la rémunération.

CONCLUSION

Notre analyse met en évidence un relâchement notable au cours des 30 derniers jours chez les employés présentant les niveaux d'investissement les plus faibles de l'entreprise.

L'exploration de l'indice de performance n'a pas révélé de liens forts ou de tendances marquées avec les autres indicateurs à notre disposition. Cette absence de relations significatives suggère soit que les facteurs explicatifs ne figurent pas dans le jeu de données, soit que la variable étudiée dépend de dimensions plus qualitatives ou contextuelles, difficile à capter par des données chiffrées seules

Une première piste d'amélioration consisterait à revoir la distribution des projets entre les employés, en veillant à une répartition plus équilibrée selon l'ancienneté, les compétences ou le niveau d'engagement. Cela pourrait favoriser une meilleure implication individuelle et collective. Par ailleurs, une réévaluation de la grille salariale pourrait être envisagée. En effet les données ne montrent pas d'évolution nette en fonction de l'ancienneté, et certains profils très expérimentés perçoivent des salaires inférieurs à des profils Junior.

Cet écart soulève la question de la valorisation de l'expérience dans l'entreprise et pourrait impacter le sentiment de reconnaissance et d'engagement. Comme nous l'avons constaté, les données disponibles ne permettaient pas de calculer de manière fiable le nombre d'heures travaillées, pourtant au cœur de la problématique initiale. Le manque de précision à ce sujet constitue une limite important e de cette analyse. Pour affiner les résultats, et formuler des conclusions plus solides, il serait pertinent de disposer de données directement liées au volume horaire.