

Jedha

CERTIFICATION BLOC 6 - CDSD

PRÉDIRE LE DÉPART DES SALARIÉS PAR MACHINE LEARNING

PRISE DE NOTES

Non chronologique

HEINRY ELODIE

Table des matières

INTRODUCTION.....	4
I. Importation des Librairies	5
II. Importation du jeu de données.....	5
III. Observations générales	5
IV. Choix du sujet.....	6
V. Elimination des doublons	6
VI. Traitement des valeurs manquantes	7
VII. Elimination de variables	7
VIII. Repérage des valeurs aberrantes ou atypiques.....	7
IX. Changement du type de certaines variables.....	8
X. Création d'un nouveau fichier .csv.....	8
XI. EDA avec le logiciel TABLEAU	8
1. <i>Analyses des différentes variables en fonction de l'attrition.....</i>	8
a) L'attrition dans l'entreprise.....	9
b) Où retrouve-t- on le plus d'attrition dans l'entreprise ?.....	10
c) Quels talents sont perdus ?.....	12
d) Rémunération	16
e) Ancienneté	24
f) Years With Current Manager.....	30
g) Job Involvement.....	31
h) Business Travel.....	32
i) Marital Status	34
j) Questionnaire de satisfaction	34
k) Training Times Last Year.....	36
l) Distance from Home.....	37
2. <i>Conclusions sur la visualisation des données.....</i>	38
XII. Analyses des corrélations entre les variables	39
1. Corrélogramme	40
2. Variance Inflation Factor	41
3. Corrélation du point bisérial.....	43
4. Test de Chi-2.....	44
XIII. Choix des variables pour le Machine Learning	46
XIV. Machine Learning.....	48

1. Préparation des données	48
a) Nettoyage.....	48
b) Modifications	49
c) Features - Target SPLIT	49
d) Train - Test SPLIT	49
e) Transformation des colonnes.....	50
2. Modèles testés	51
3. Amélioration des modèles	53
4. Résultats du Machine Learning	56
5. Importance des variables prédictives	58
a) Les coefficients de feature de la régression logistique	58
b) Les odds-ratio des features et p-value.....	59
c) Résultats.....	59
d) Interprétation	60
6. Conclusion sur le Machine Learning	62
7. Perspectives d'amélioration du modèle de prédiction	62
XV. Conclusion de l'étude	63
XVI. Recommandations	66
1. Accès aux formations.....	66
2. Politique de rémunération.....	66
3. Contraintes professionnelles	67
4. Embauche.....	67
5. Feedback	67
6. Autres idées.....	68

INTRODUCTION

La fidélisation des talents et la gestion des compétences représentent des enjeux stratégiques majeurs pour les entreprises. Un taux d'attrition élevé nuit non seulement à la productivité et à la stabilité des équipes, mais génère également des coûts significatifs liés au recrutement et à la formation de nouveaux collaborateurs.

L'attrition désigne l'ensemble des départs d'une entreprise, qu'ils soient volontaires (démissions, retraites) ou non (licenciements, mutations, décès). En général, un taux d'attrition autour de 10 % est considéré comme « normal ».

Nous avons choisi d'analyser ce phénomène au sein d'IBM, une multinationale américaine qui est spécialisée dans la vente de matériel, de logiciels et de solutions informatiques.

L'objectif de cette étude est d'identifier les principaux facteurs responsables des départs chez IBM à travers des analyses statistiques sur les données des salariés. Comprendre l'impact de chaque variable sur l'attrition nous permettra de proposer des solutions et des recommandations pour diminuer leur impact sur les départs et ainsi réduire le phénomène d'attrition.

Enfin, la construction d'un modèle de Machine Learning fournira un outil performant pour prédire les départs potentiels des salariés, facilitant ainsi la gestion anticipative de l'attrition au sein de l'entreprise.

I. Importation des Librairies

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from statsmodels.stats.outliers_influence import variance_inflation_factor
from scipy.stats import pointbiserialr, f_oneway, chi2_contingency
from sklearn.model_selection import train_test_split
from sklearn.impute import SimpleImputer          # utilisé ni sur le training ni sur le test
from sklearn.preprocessing import StandardScaler, OneHotEncoder, LabelEncoder
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
import time
from sklearn.metrics import accuracy_score, roc_auc_score, f1_score, precision_score, recall_score
```

II. Importation du jeu de données

Fichier "WA_Fn-UseC_-HR-Employee-Attrition.csv"

III. Observations générales

- C'est une description annuelle de l'ensemble des employés travaillant chez IBM.
- Il y a 1470 employés (=lignes).
- La description pour chaque employé est réalisée grâce à 35 variables (=colonnes) :
 - 26 variables numériques
 - 9 catégorielles

Variables numériques :

Age : Âge de l'employé.

DailyRate : Taux journalier de l'employé.

DistanceFromHome : Distance entre le domicile de l'employé et le lieu de travail.

Education : Niveau d'éducation de l'employé (échelle de 1 à 5).

EmployeeCount : Champ constant avec une valeur de "1" pour tous les employés.

EmployeeNumber : Identifiant unique attribué à chaque employé.

EnvironmentSatisfaction : Satisfaction de l'employé vis-à-vis de son environnement de travail (note de 1 à 4).

HourlyRate : Taux horaire de l'employé.

JobInvolvement : Degré d'implication de l'employé dans son poste (note de 1 à 4).

JobLevel : Niveau de poste de l'employé, de débutant (1) à senior (5).

JobSatisfaction : Satisfaction de l'employé vis-à-vis de son travail (note de 1 à 4).

MonthlyIncome : Salaire mensuel de l'employé.

MonthlyRate : Autre variable financière indiquant le taux mensuel de l'employé.

NumCompaniesWorked : Nombre d'entreprises pour lesquelles l'employé a travaillé.

PercentSalaryHike : Pourcentage d'augmentation du salaire de l'employé.
PerformanceRating : Évaluation de la performance de l'employé (note de 1 à 4).
RelationshipSatisfaction : Satisfaction vis-à-vis des relations professionnelles (note de 1 à 4).
StandardHours : Nombre standard d'heures de travail (toujours 40).
StockOptionLevel : Niveau d'options sur actions accordées à l'employé (de 0 à 3).
TotalWorkingYears : Nombre total d'années d'expérience professionnelle.
TrainingTimesLastYear : Nombre de sessions de formation suivies l'année précédente.
WorkLifeBalance : Évaluation de l'équilibre travail-vie personnelle (note de 1 à 4).
YearsAtCompany : Nombre d'années passées dans l'entreprise actuelle.
YearsInCurrentRole : Nombre d'années passées dans le rôle actuel.
YearsSinceLastPromotion : Nombre d'années depuis la dernière promotion.
YearsWithCurrManager : Nombre d'années passées avec le manager actuel.

Variables catégorielles :

Attrition : Indique si l'employé a quitté l'entreprise ("Yes") ou est resté en poste ("No").
BusinessTravel : Fréquence des déplacements professionnels.
Department : Département dans lequel l'employé travaille.
EducationField : Domaine d'études de l'employé.
Gender : Genre de l'employé.
JobRole : Métier de l'employé au sein de l'organisation.
MaritalStatus : Statut marital de l'employé.
OverTime : Indique si l'employé effectue des heures supplémentaires.
Over18 : Indique si l'employé a plus de 18 ans.

IV. Choix du sujet

L'étude portera sur l'attrition des salariés : l'objectif sera de déterminer les facteurs favorisant ou limitant l'attrition, autrement dit, les variables ayant une influence sur l'attrition d'un salarié. Ainsi, nous pourrons cibler les actions à mener et proposer des recommandations à l'entreprise.

Nous proposons de mettre au point un modèle de Machine Learning dont l'objectif sera de prédire le départ d'un salarié.

Pour l'étude, la variable cible est "Attrition" qui est divisée en deux classes :

- 0 = No = salarié resté
- 1 = Yes = salarié parti

V. Elimination des doublons

Le jeu de données ne présentait pas de doublons.

VI. Traitement des valeurs manquantes

Le jeu de données n'avait pas de valeurs manquantes.

VII. Elimination de variables

- La variable "Over18" et la variable "StandardHours" montrent une seule valeur pour l'ensemble des salariés, elles sont donc éliminées de notre jeu de données.
- La variable "EmployeeCount" montre également une valeur unique, mais elle va nous servir pour calculer la somme des employés avec le logiciel 'Tableau'.
- La variable "EmployeeNumber" donne une valeur différente à chacun des salariés, cette colonne aurait pu nous être utile pour des jointures de tables, mais dans notre étude elle ne sera pas nécessaire, elle est donc éliminée.
- La variable numérique 'PerformanceRating' n'est dispersée que sur deux valeurs (3 et 4), il s'agit d'une variable catégorielle codée (=variable catégorielle discrète et ordinale). Nous l'éliminons car elle n'est pas assez dispersée.

Il nous reste donc :

- la variable cible qui est catégorielle : "Attrition"
- 7 variables catégorielles potentiellement prédictives
- 23 variables numériques potentiellement prédictives

On peut déjà remarquer à ce stade que nous avons un déséquilibre de classe sur deux variables : "Attrition" et "OverTime" qui montrent une fréquence de 'No' très élevée dans les deux cas.

VIII. Repérage des valeurs aberrantes ou atypiques

Grâce au module Seaborn, nous faisons des boxplot pour les colonnes ayant des valeurs numériques continues non ordonnées :

- "MonthlyIncome" : On peut voir que les valeurs supérieures à 16 581 réagissent comme des valeurs atypiques, cependant, bien qu'extrêmes, ces valeurs ne sont pas aberrantes. Elles sont le résultat de la politique RH de l'entreprise.
- "PercentSalaryHike" : pas de valeurs aberrantes ou atypiques.
- "MonthlyRate" : pas de valeurs aberrantes ou atypiques.
- "DailyRate" : pas de valeurs aberrantes ou atypiques.
- "HourlyRate" : pas de valeurs aberrantes ou atypiques.

IX. Changement du type de certaines variables

Pour une facilité de lecture des graphiques construits avec le logiciel ‘Tableau’, nous avons changé les variables numériques discrètes et ordinaires en variables catégorielles.

```
df["Education"] = df["Education"].replace({1:"1_Below  
College",2:"2_College",3:"3_Bachelor",4:"4_Master",5:"5_Doctor"})  
df["EnvironmentSatisfaction"] =  
df["EnvironmentSatisfaction"].replace({1:"1_Faible",2:"2_Moyen",3:"3_Elevée",4:"4_Très  
élevée"})  
df["JobInvolvement"] =  
df["JobInvolvement"].replace({1:"1_Faible",2:"2_Moyen",3:"3_Elevée",4:"4_Très élevée"})  
df["JobLevel"] =  
df["JobLevel"].replace({1:"1_Débutant",2:"2_Junior",3:"3_Confirmé",4:"4_Senior",  
5:"5_Expert"})  
df["JobSatisfaction"] =  
df["JobSatisfaction"].replace({1:"1_Faible",2:"2_Moyen",3:"3_Elevée",4:"4_Très elevée"})  
df["RelationshipSatisfaction"] =  
df["RelationshipSatisfaction"].replace({1:"1_Faible",2:"2_Moyen",3:"3_Elevée",4:"4_Très  
élevée"})  
df["WorkLifeBalance"] = df["WorkLifeBalance"].replace({1:"1_Mauvais",2:"2_Bon",3:"3_Très  
bon",4:"5_Parfait"})
```

À ce stade, nous avons donc 16 variables numériques et 15 variables catégorielles dont notre variable cible.

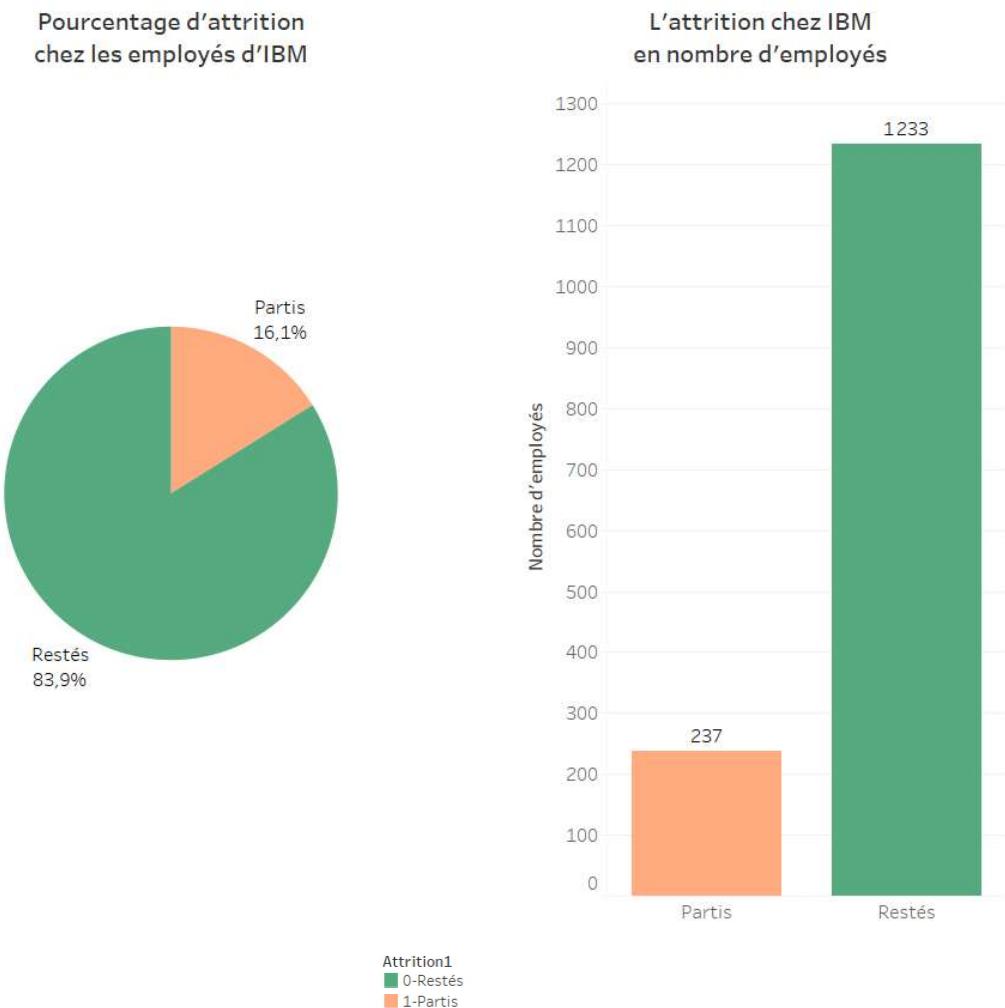
X. Création d'un nouveau fichier .csv

Pour poursuivre l'analyse du jeu de données avec le logiciel TABLEAU, nous avons créé un nouveau fichier : “ IBM_EDA-fichier_pour_TABLEAU.csv”.

XI. EDA avec le logiciel TABLEAU

1. Analyses des différentes variables en fonction de l'attrition

a) L'attrition dans l'entreprise



Nous constatons que l'attrition chez IBM est plus élevée qu'une attrition dite 'normale'. En effet, d'après les informations trouvées, le taux d'attrition moyen d'une entreprise s'élève aux alentours de 10%, tandis que chez IBM, il s'élève à 16,1 %.

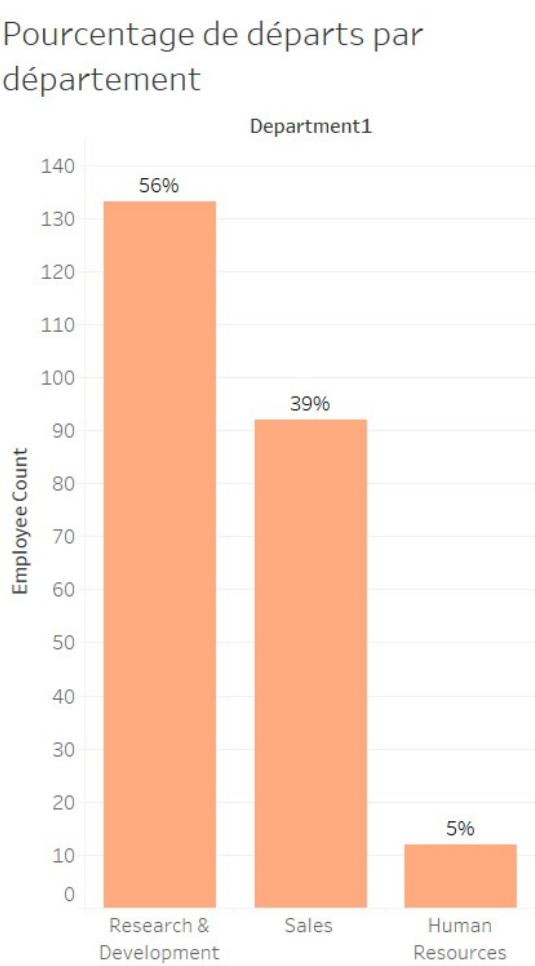
Nous pouvons donc dire que le taux d'attrition est alarmant et qu'il serait judicieux de mettre en place des mesures afin de le diminuer.

Nous remarquons qu'il y a 237 salariés partis pour 1233 salariés restés en poste. Cela posera un problème de déséquilibre des classes sur la variable 'Attrition', il faudra donc pallier à ce problème lors de la construction de notre modèle de Machine Learning.

b) Où retrouve-t-on le plus d'attrition dans l'entreprise ?

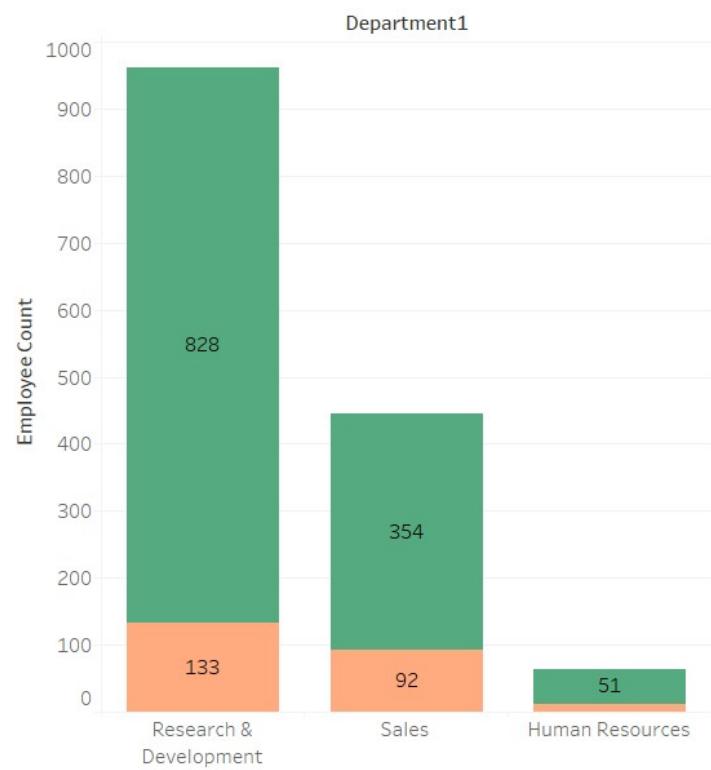
i. Department

Pourcentage de départs par département

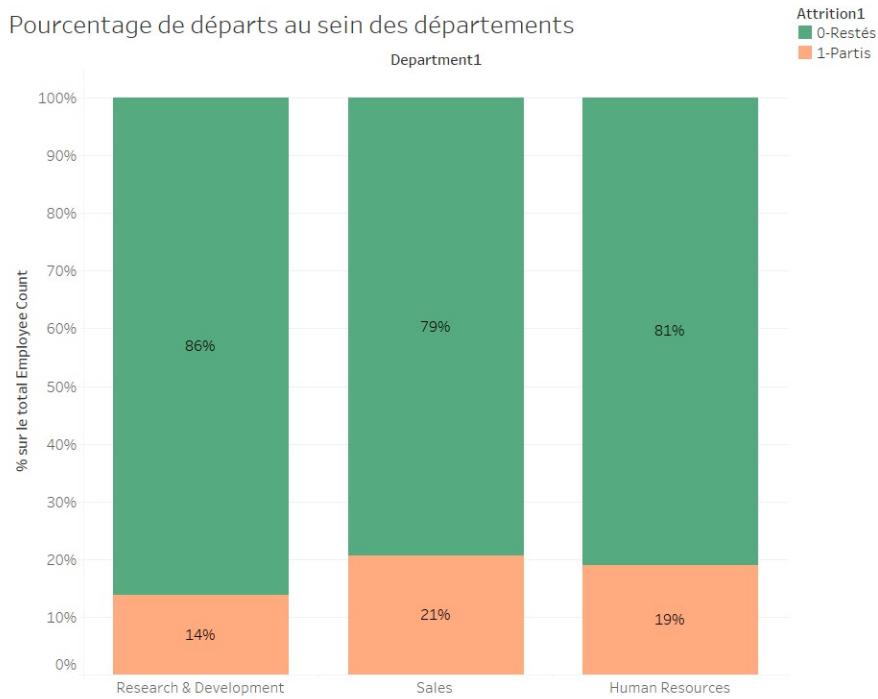


Attrition1
0-Restés
1-Partis

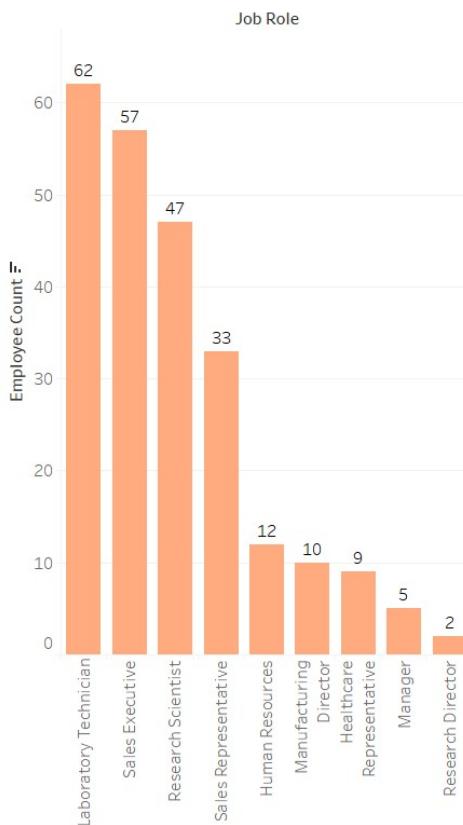
Répartition des salariés dans les différents départements



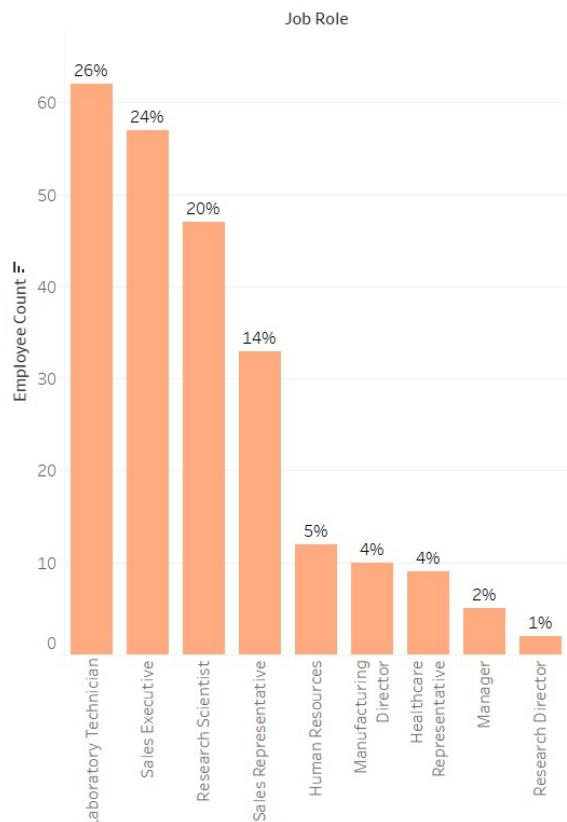
L'attrition est principalement retrouvée au sein du département Recherche & Développement. Mais c'est également le département où on retrouve la majorité des salariés. En proportion du nombre d'employés, c'est au département Ventes que l'attrition est la plus élevée et atteint 21 %, suivi par le département des Ressources Humaines qui présente un taux d'attrition de 19 %.



ii. Job Role



Nombre de salariés partis par métier

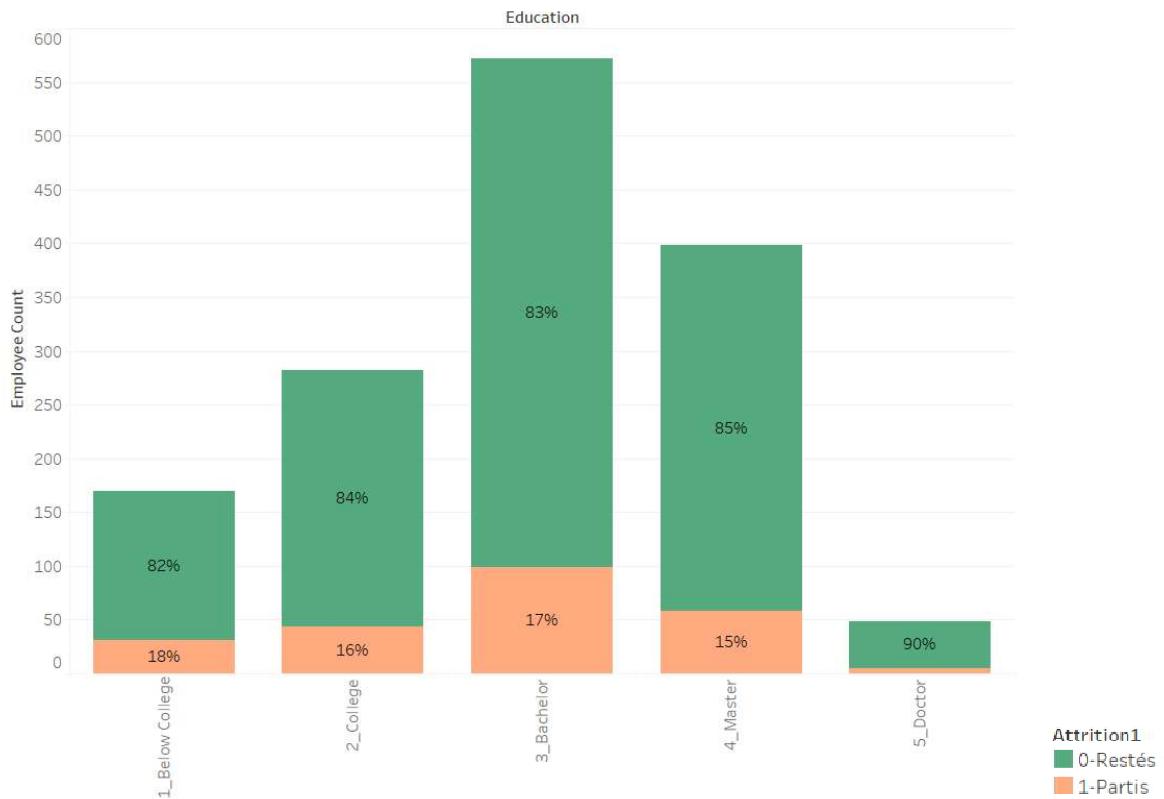


Pourcentage des départs de chez IBM par métier

Les taux d'attrition les plus élevés sont pour les métiers : Technicien de laboratoire, Directeur des ventes et Chercheur scientifique. Les taux dépassent largement le taux d'attrition de l'entreprise (16 %) puisqu'ils s'élèvent respectivement à 26 %, 24 %, 20 %.

c) Quels talents sont perdus ?

i. Education



Répartition des salariés en fonction de leur diplôme avec visualisation de l'attrition

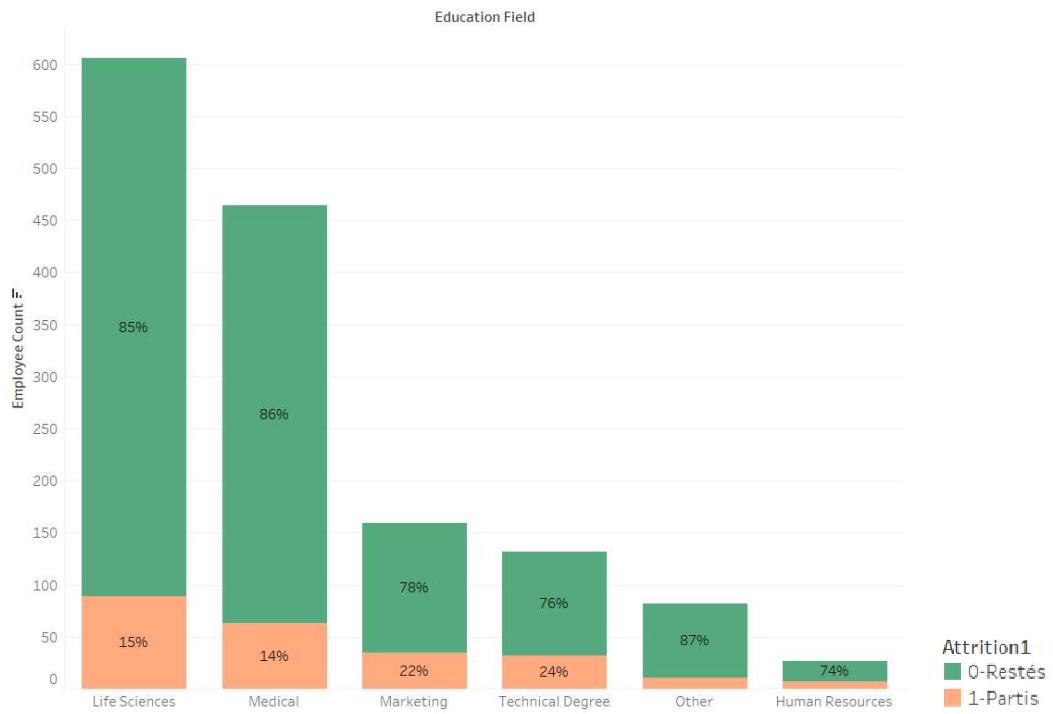
La plupart des employés de l'organisation a obtenu un baccalauréat ou une maîtrise comme diplôme.

Nous pouvons observer une tendance à la diminution du taux d'attrition à mesure que le diplôme est élevé.

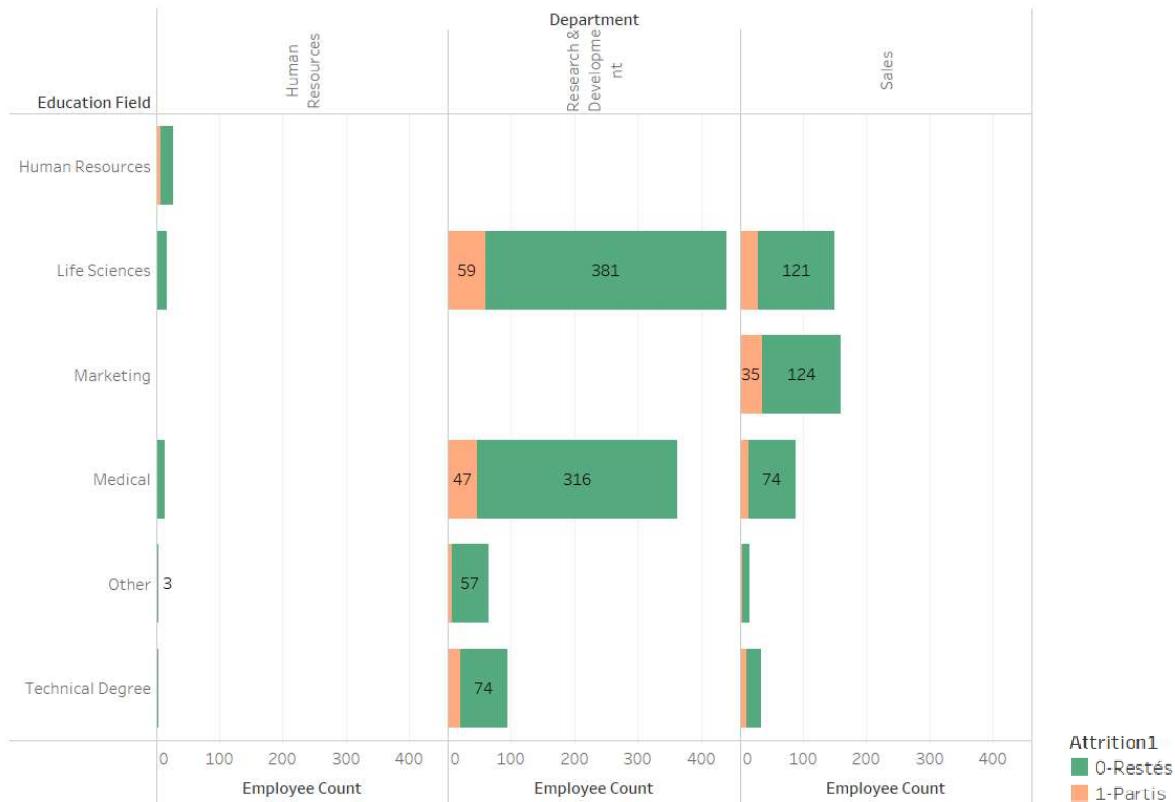
ii. Education Field

Les employés ayant étudié le marketing ou ayant un diplôme de technicien ont une forte tendance à quitter l'entreprise (respectivement 22 et 24%). Il serait intéressant de pousser un peu plus les analyses sur ces diplômés.

La majorité des employés sont formés dans les domaines 'life sciences' et 'médical' et elle a un taux d'attrition légèrement inférieur à celui de l'entreprise qui est à 16%.

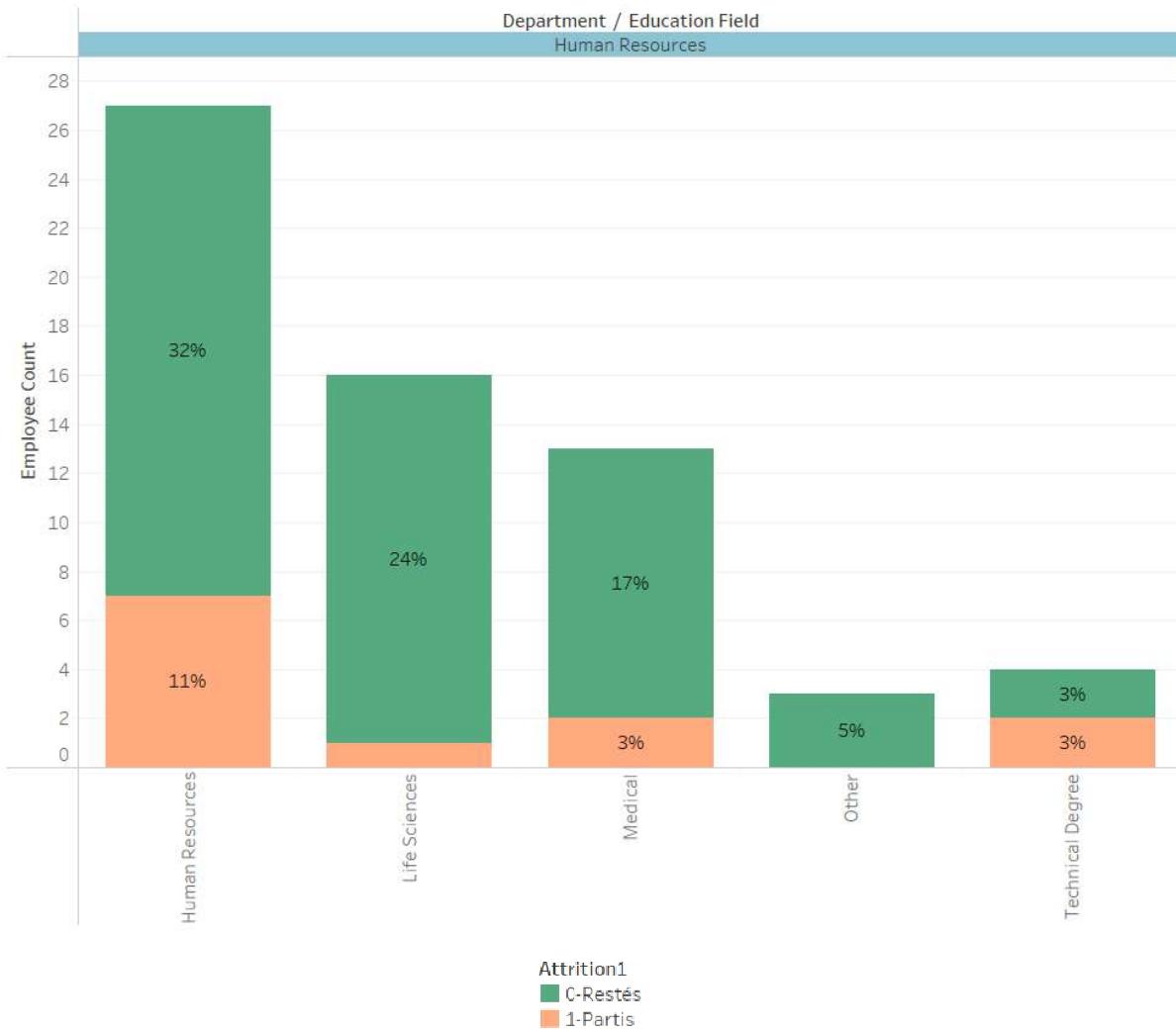


Répartition des salariés en fonction du domaine d'étude avec visualisation de l'attrition



Répartition des salariés en fonction du domaine d'étude par département avec visualisation de l'attrition

Les personnes recrutées viennent de secteurs cohérents par rapport au département auquel ils appartiennent. On remarque tout de même que, dans le département Ressources Humaines, très peu d'employés proviennent du domaine d'éducation 'Human Resources' et qu'il s'agit d'un département avec peu d'individus. L'attrition y est très élevée (26%).



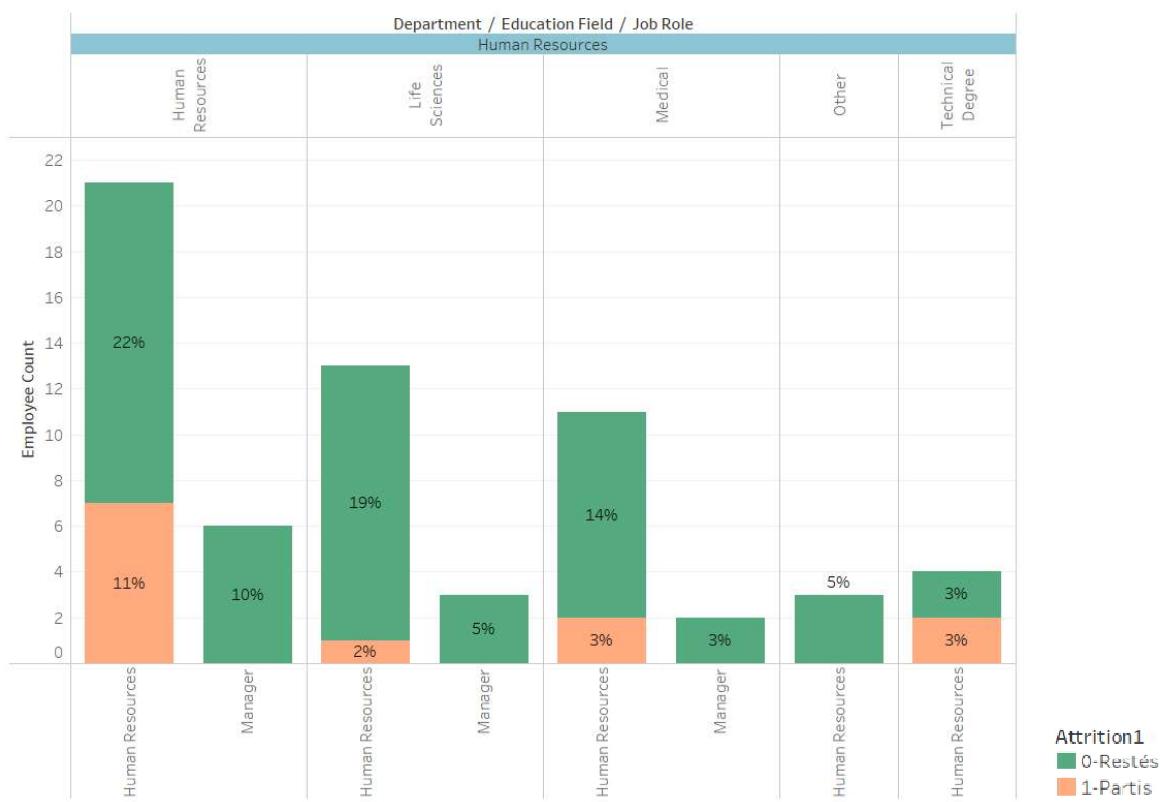
Répartition des salariés en fonction du domaine d'étude au sein du département RH avec visualisation de l'attrition

On a une forte proportion de personnels formés en sciences ou au médical mais travaillant pourtant au département RH.

Ceci est assez étonnant, d'autant plus qu'ils ont une attrition plus faible que les salariés formés aux RH. Ce sont donc les spécialistes des Ressources Humaines au sein du département RH qui ont tendance à quitter l'entreprise.

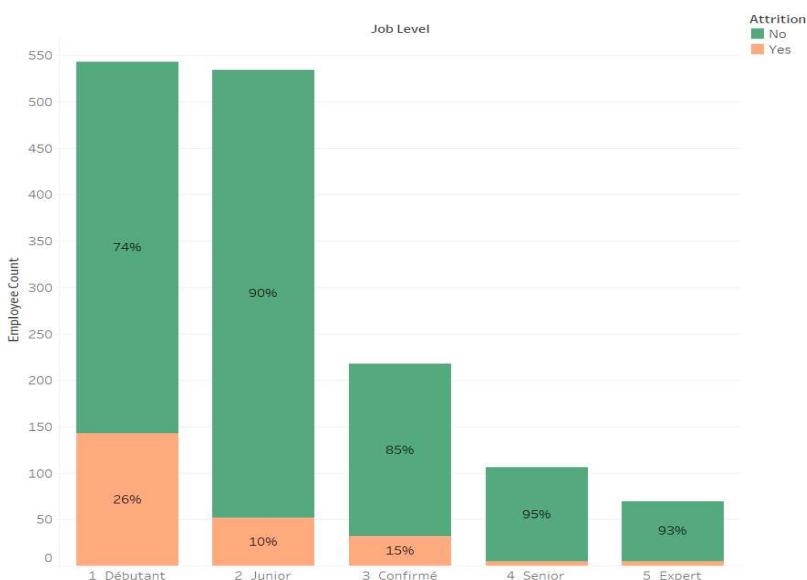
D'autre part, les managers des RH ont une attrition de 0. La moitié des managers sont issus d'un parcours de formation aux RH tandis que l'autre moitié a été formée à d'autres domaines d'étude. Il serait intéressant de mener une enquête de satisfaction ciblée auprès du personnel des RH au métier appelé 'Human Resources'. En faisant la distinction entre les

personnes qui ont étudié dans le domaine RH et ceux qui ont étudié dans d'autres domaines, il serait possible de comprendre le taux d'attrition élevé concernant les spécialistes RH (1/4 de partis).



Répartition des salariés par métier et domaine d'étude au sein du département RH avec visualisation de l'attrition

iii. Job Level



Répartition des salariés en fonction de leur niveau dans l'entreprise avec visualisation de l'attrition

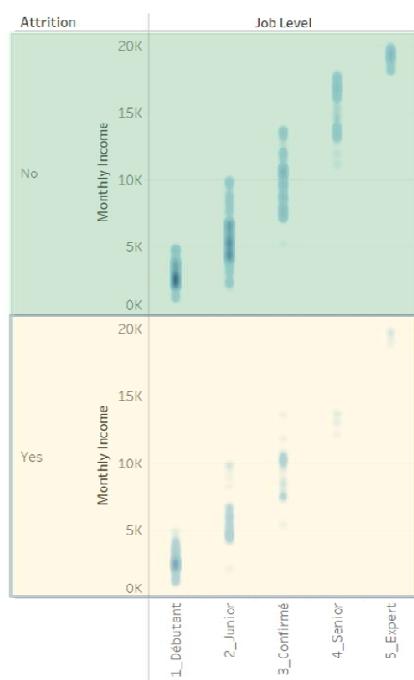
La plupart des employés de l'organisation sont au niveau débutant ou junior.

L'attrition la plus élevée se situe chez les débutants et est bien supérieure au taux d'attrition de l'entreprise.

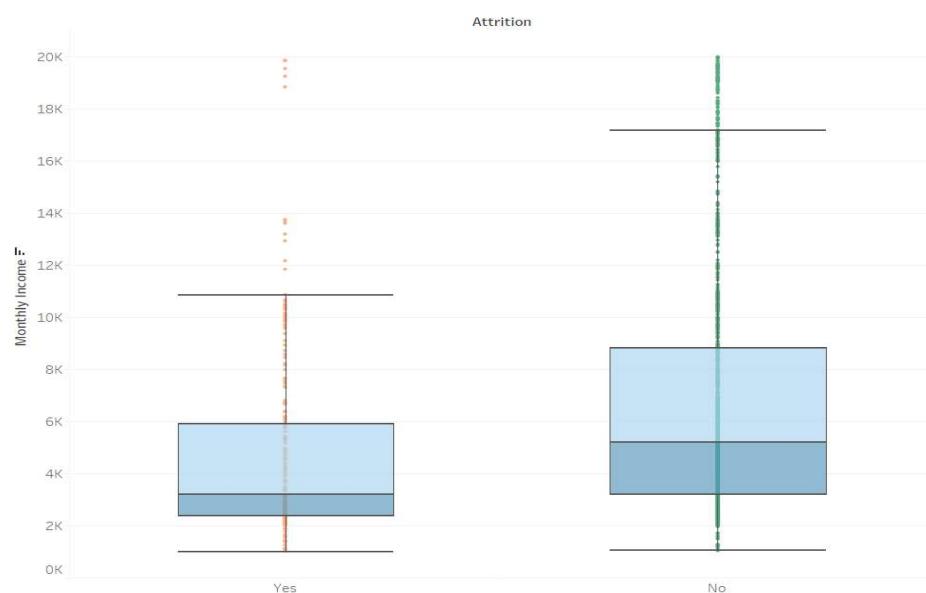
À mesure que le niveau augmente, le taux d'attrition diminue.

d) Rémunération

i. Monthly Income



Salaires mensuels des salariés attritionnistes et non-attritionnistes en fonction de leur métier



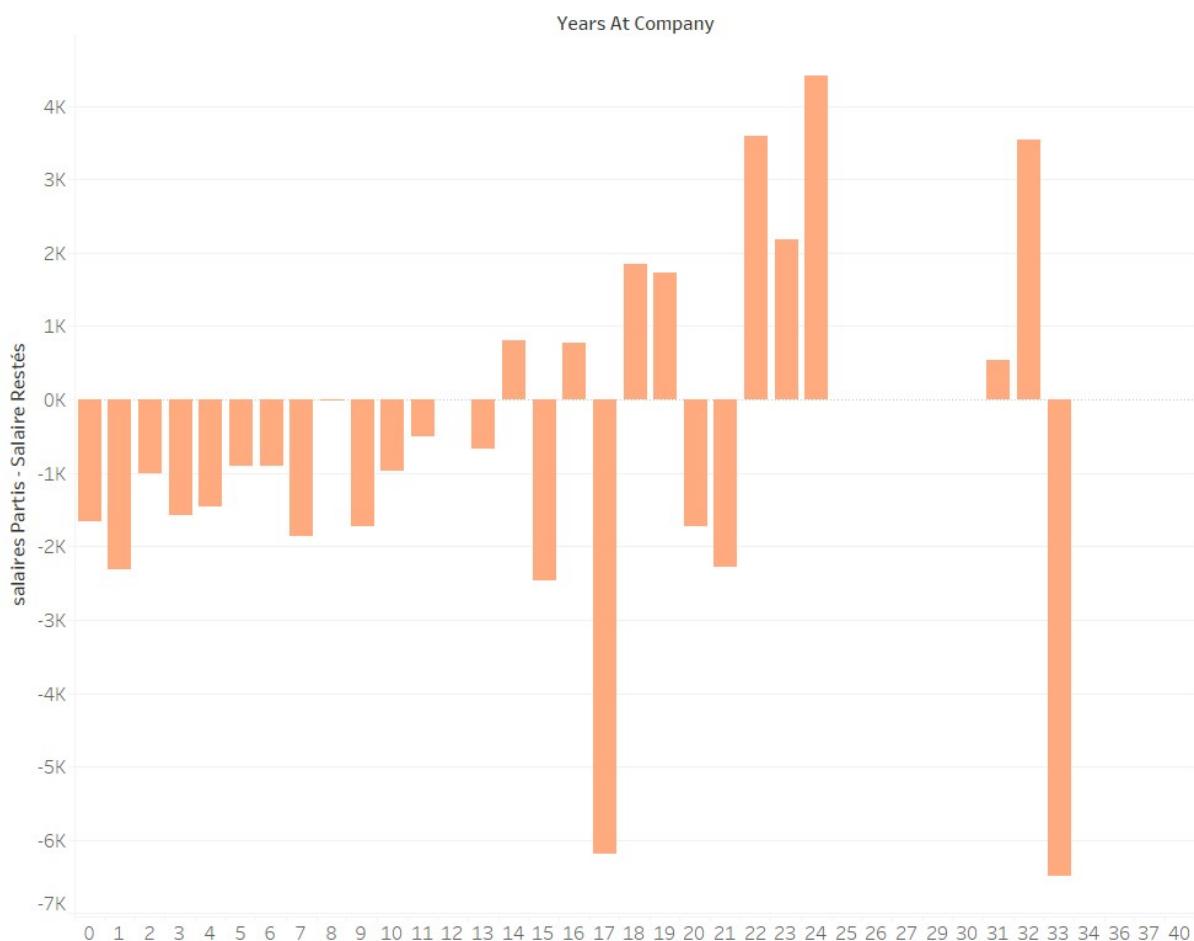
Distribution des salaires mensuels chez les attritionnistes et les non-attritionnistes.

La majorité des salariés perçoit un salaire mensuel inférieur à 10 K et l'attrition y est élevée, il s'agit des débutants et junior de l'entreprise.

On remarque que les salariés qui partent ont tendance à faire partie des moins bien payés dans leur niveau (débutant, junior, confirmé et senior).

À mesure que le revenu mensuel augmente, l'attrition diminue.

Le salaire médian mensuel des attritionnistes est très inférieur au salaire médian des non-attritionnistes. D'autre part, on remarque qu'il y a très peu d'attrition pour le personnel touchant des hauts salaires.



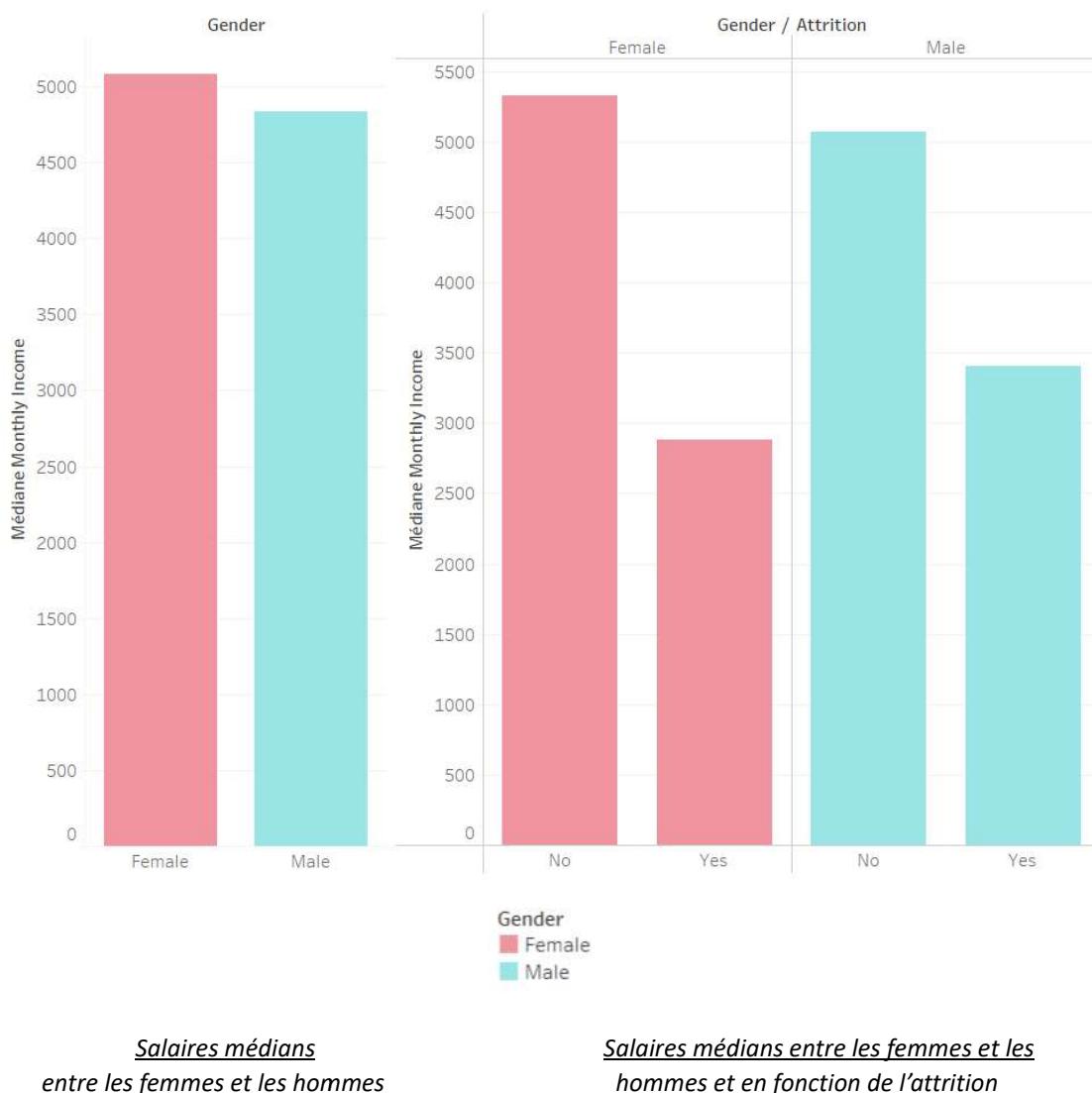
Ecart de salaires par année d'ancienneté chez IBM pour les salariés partis

On voit que la différence de salaire entre les individus partis et ceux restés est en défaveur des individus partis, ce qui peut expliquer leur départ. Pour les 13 premières années dans l'entreprise, les employés partants sont ceux dont le salaire est moindre en comparaison des employés restés avec la même ancienneté.

ii. Gender / Monthly Income

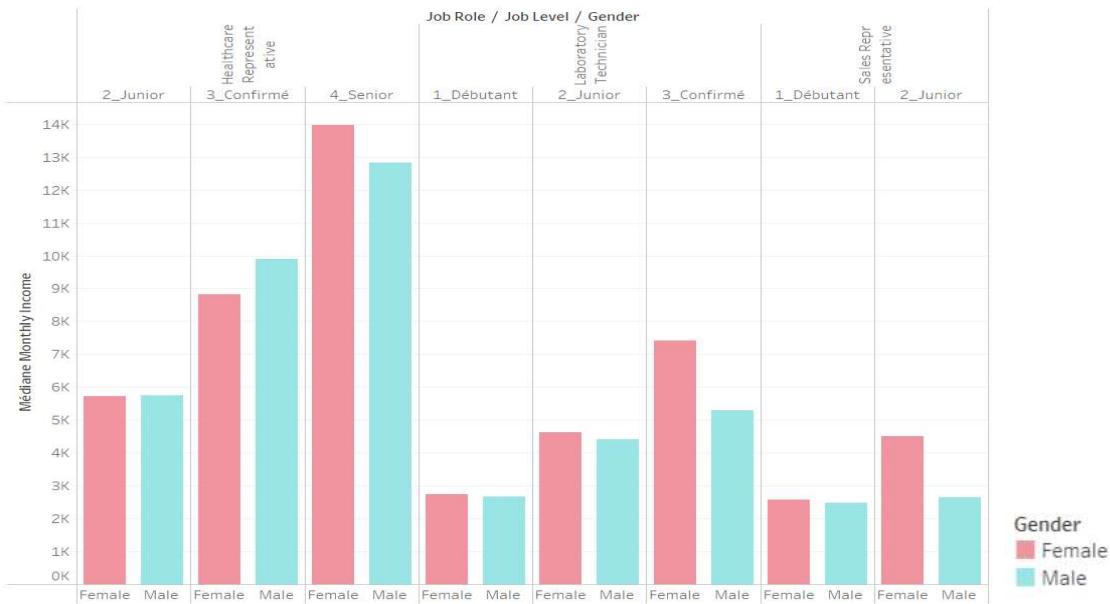
On remarque que les femmes ont un salaire médian légèrement supérieur à celui des hommes.

Pour chaque genre, le salaire médian des employés partis était plus bas que celui des restants.



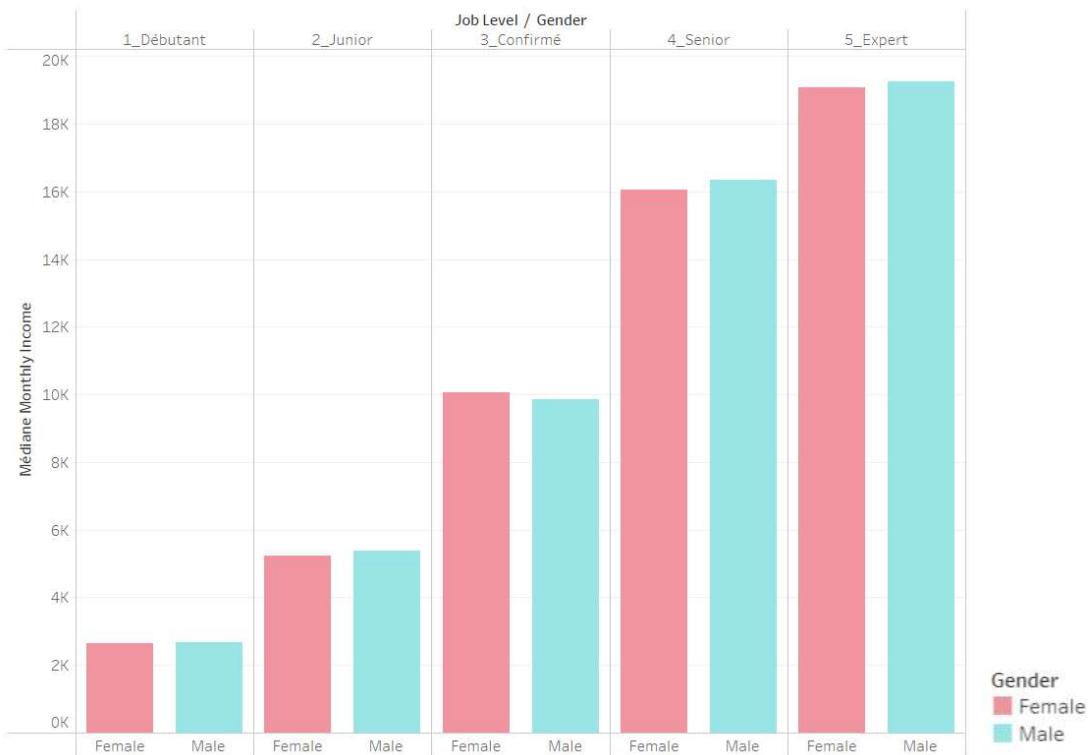
Salaires médians entre les femmes et les hommes

Salaires médians entre les femmes et les hommes et en fonction de l'attrition



Salaires médians en fonction du genre, du métier et du niveau des employés

On voit qu'il y a 3 situations professionnelles où les femmes sont mieux rémunérées (Représentantes des soins de santé en senior / techniciennes de laboratoire junior et confirmées / Représentantes des ventes junior), c'est la seule configuration graphique qui explique le salaire médian plus élevé chez les femmes, car dans le cas général, il est toujours inférieur :



Salaires médians pour les femmes et les hommes en fonction de leur niveau

Dans le cas général, seul le niveau ‘Confirmé’ montre un salaire médian supérieur pour les femmes en comparaison à celui des hommes. Pour l’ensemble des autres niveaux, le salaire médian des hommes est supérieur au salaire médian des femmes.

iii. Comparaison avec Daily Rate, Hourly Rate et Monthly Rate

Lorsque nous comparons les données des variables ‘Monthly Income’, ‘Daily Rate’, ‘Hourly Rate’ et ‘Monthly Rate’ en fonction de l’attrition et du niveau des employés on remarque que la variable ‘DailyRate’ est très dispersée dans chaque classe de Job Level (niveau). Nous constatons la même chose pour la variable ‘Hourly Rate’.

On remarque encore que ce sont les débutants et les juniors qui montrent le plus d’attrition.

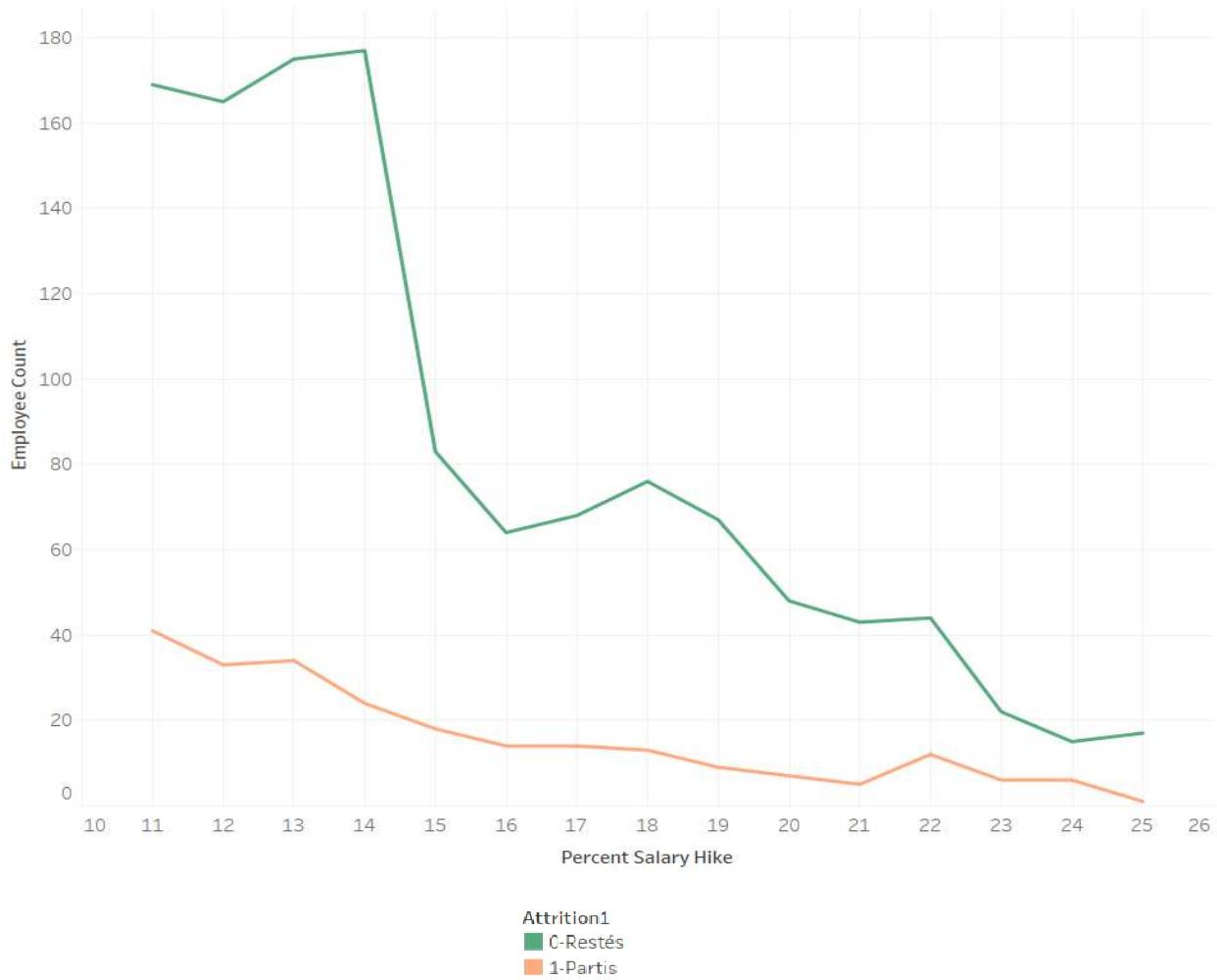
Concernant la variable ‘Monthly Rate’, le taux mensuel ne semble pas logique, ou je ne comprends pas à quoi correspond la variable.



Comparaison des variables ‘Monthly Income’, ‘Daily Rate’, ‘Hourly Rate’ et ‘Monthly Rate’ en fonction de l’attrition et du niveau des employés

On ne gardera que la variable ‘Monthly Income’ comme indication de salaire en gardant à l'esprit que les disparités peuvent être dues à des emplois à 80%, ou autres particularités de contrat de travail.

iv. Percentage Salary Hike

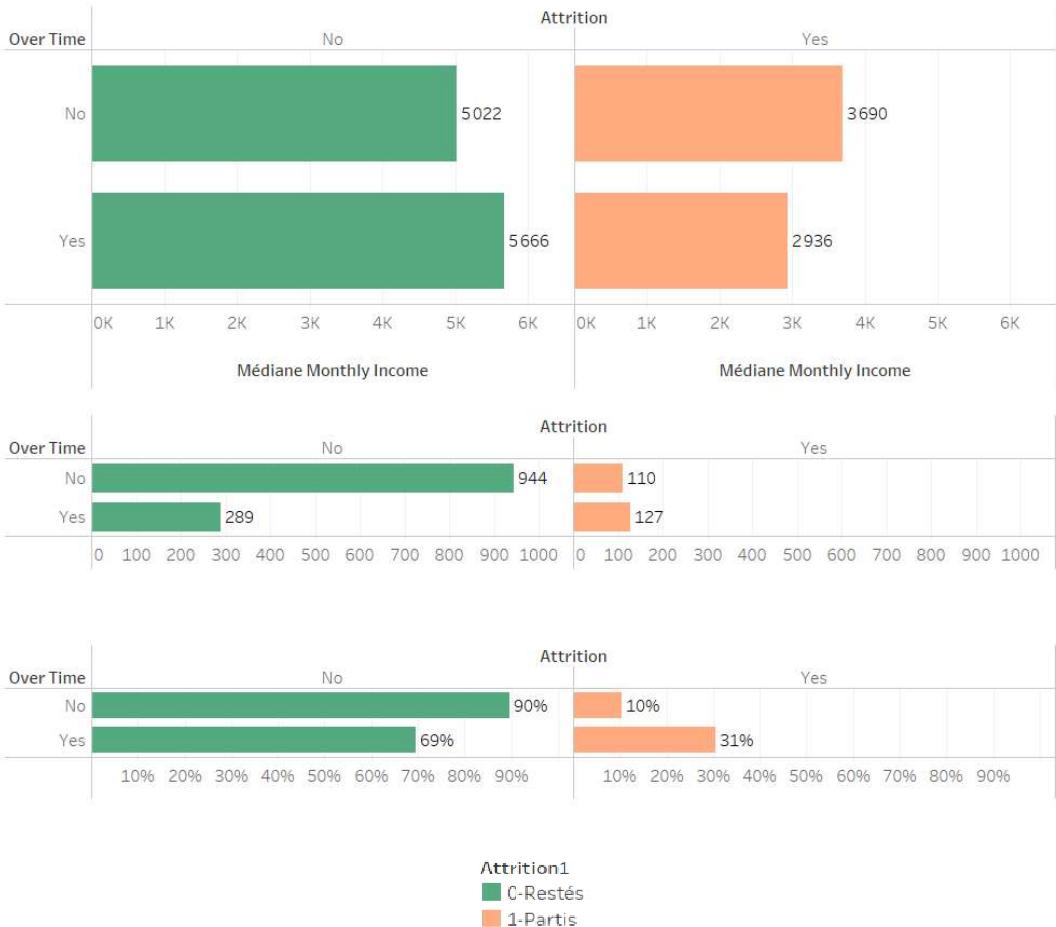


Répartition des augmentations de salaire exprimées en pourcentage

Les augmentations de salaire débutent à 11% avec une majorité entre 11 et 14% inclus.

À mesure que le pourcentage d'augmentation du salaire augmente, le taux d'attrition diminue.

v. Over Time



Comparaison des salaires médians mensuels et Répartition des employés en fonction de l'exécution d'heures supplémentaires et de l'attrition

Les salariés partis sont ceux qui faisaient des heures supplémentaires et dont le salaire médian était inférieur au salaire médian des employés ne faisant pas d'heures supplémentaires.

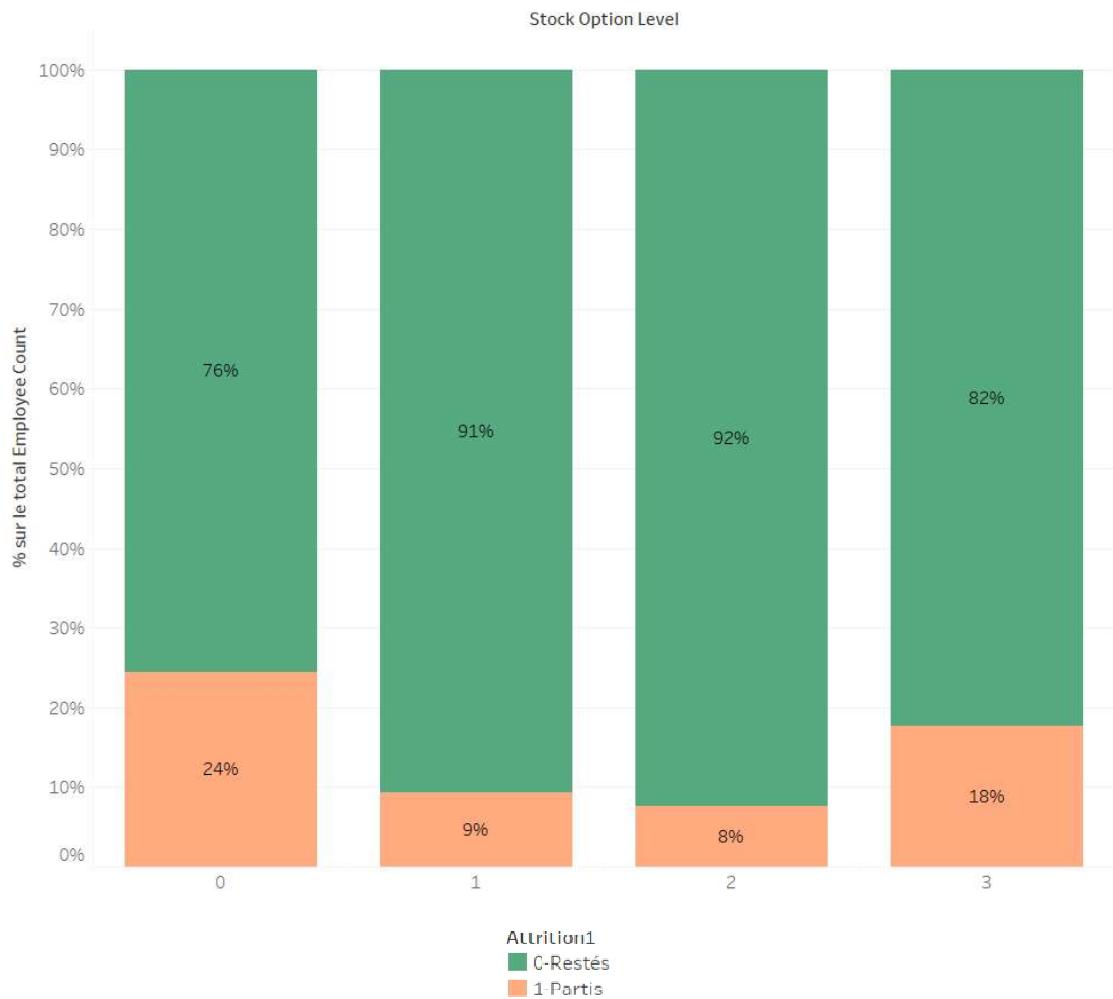
Les hauts salaires semblent enclins à faire des heures supplémentaires.

La variable 'Over Time' présente un déséquilibre de classe important. En effet, nous n'avons que 416 individus faisant des heures supplémentaires sur 1470 employés, 127 d'entre eux ont quitté l'entreprise tandis que les autres (289) sont restés.

Ceci dit, en proportion, c'est 31% des employés faisant des heures supplémentaires qui quittent l'entreprise

Si on regarde la répartition des heures supplémentaires en fonction des métiers (Job Role) et du niveau (Job Level), on voit qu'il y a des heures supplémentaires à faire dans chaque secteur d'activité et aussi bien pour les débutants que pour les séniors.

vi. Stock Option Level



Répartition de l'attrition en fonction du niveau de stock option

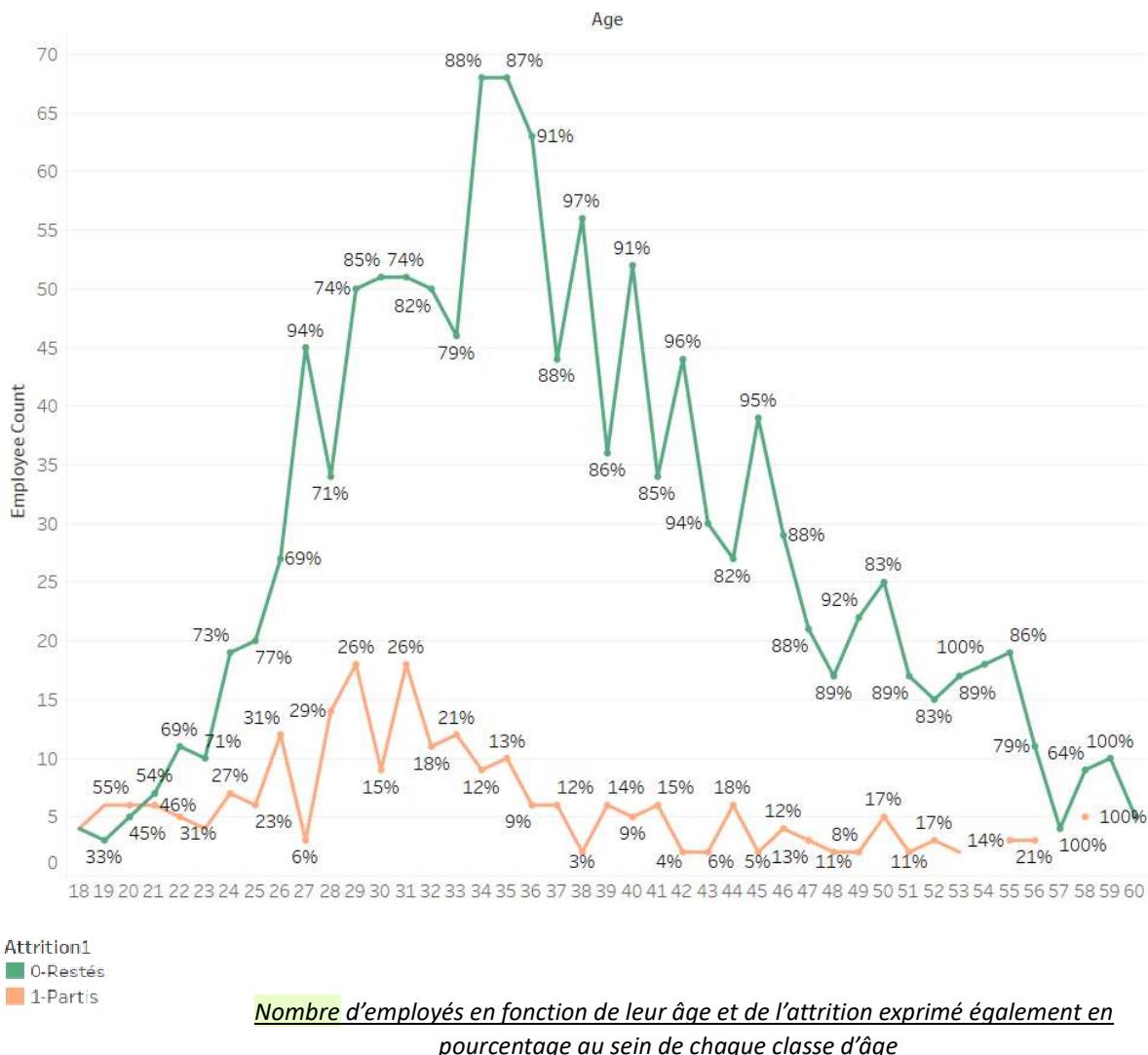
La majorité des employés (83,5 %) a moins de 2 niveaux de stock options.

L'obtention de stock options n'est pas dépendante du nombre d'années d'ancienneté dans l'entreprise et peuvent être négociées au recrutement. On le remarque chez les nouveaux arrivants (= 0 année d'ancienneté).

On observe 24% d'attrition chez les personnes n'ayant pas de stock options, mais nous avons également 18% d'attrition chez les personnes disposant du maximum possible. Dans les deux cas, on remarque une attrition supérieure à l'attrition moyenne dans l'entreprise.

e) Ancienneté

i. Age



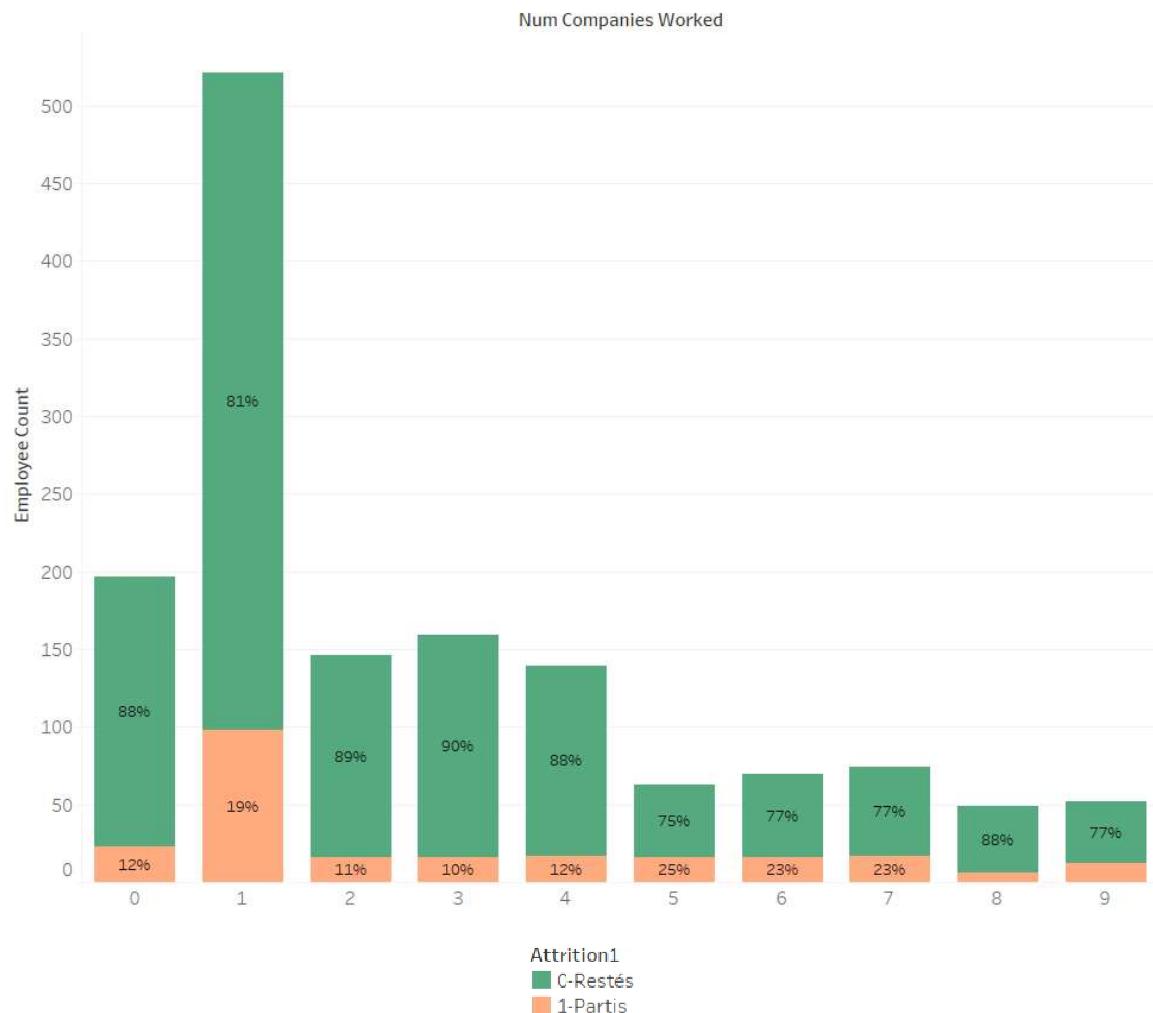
À mesure que l'âge des salariés augmente, l'attrition diminue.

Nous pouvons également observer que l'âge moyen des employés partis de l'organisation est inférieur à l'âge moyen des employés restés.

Les jeunes employés (âge inférieur ou égal à 34 ans) quittent davantage l'entreprise que les employés plus âgés.

C'est une dynamique habituelle dans une entreprise.

ii. Number of Companies Worked



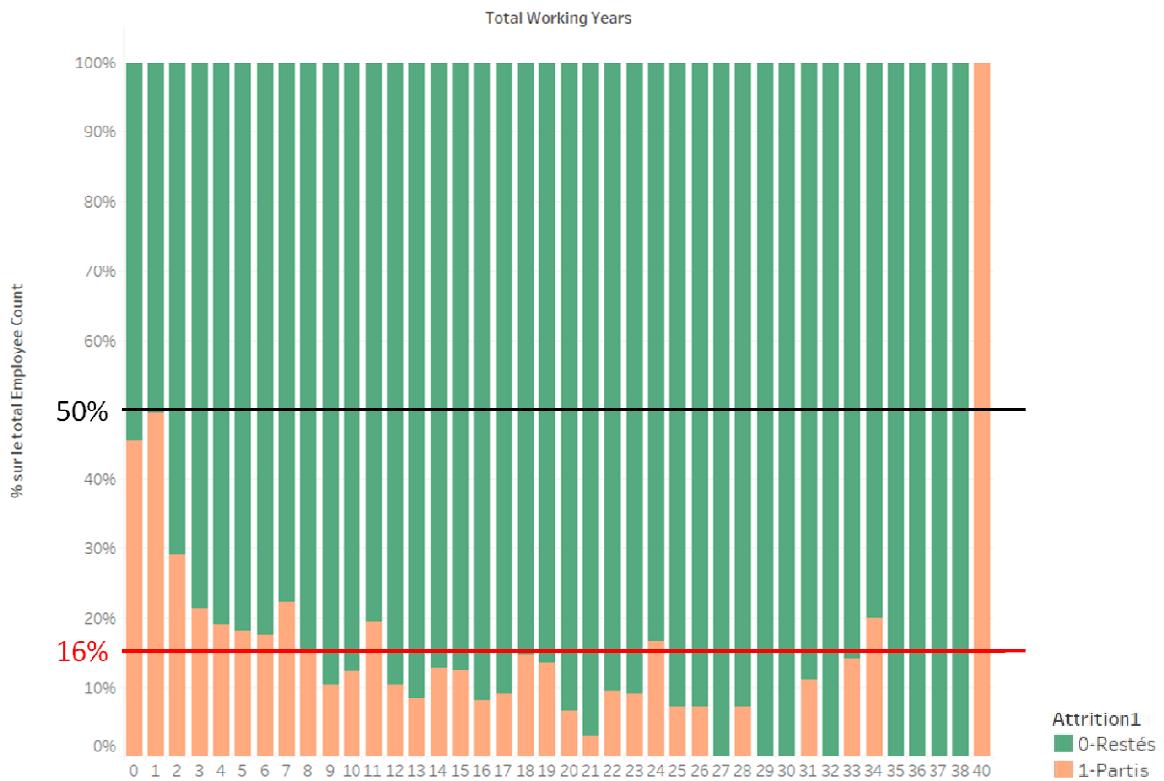
Répartition des employés par nombre d'entreprises dans lesquelles ils ont travaillés

La plupart des salariés ont travaillé pour moins de 2 entreprises.

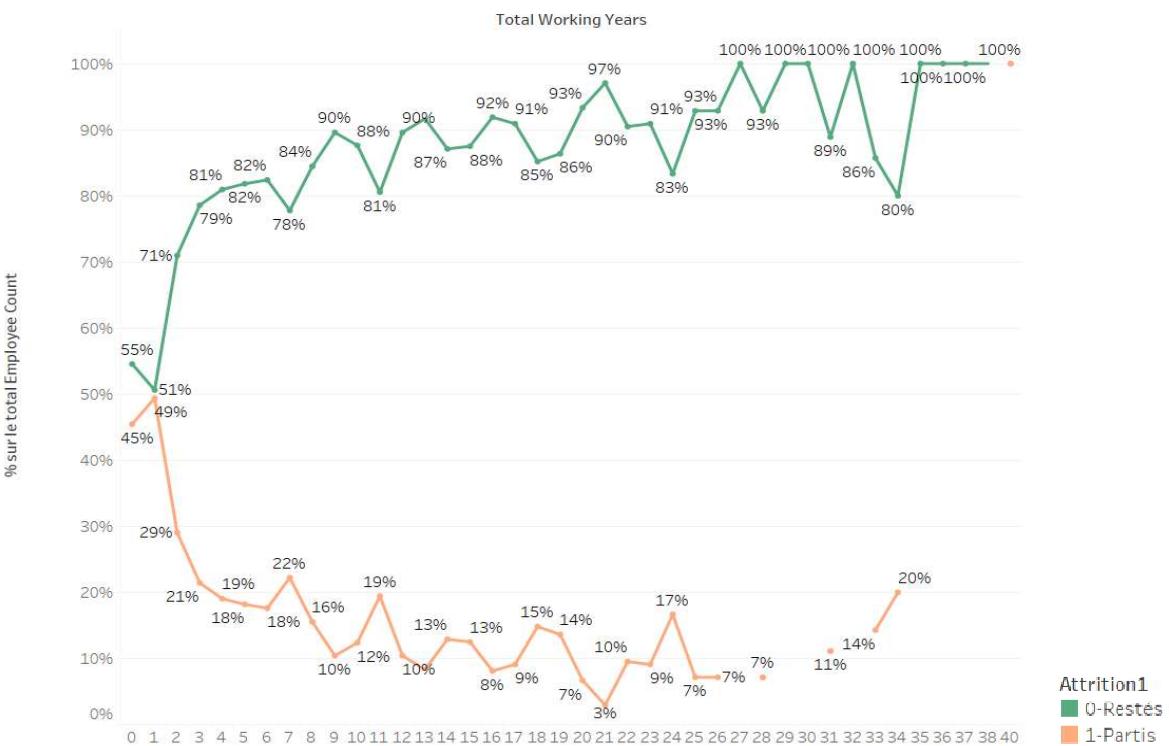
La majorité des employés ont travaillé pour une autre entreprise avant d'intégrer IBM.

On peut voir que l'attrition augmente lorsque le nombre d'entreprises fréquentées par l'employé augmente. Il faut tout de même rester prudent sur cette interprétation, en effet, nous avons peu de valeurs dans ces classes.

iii. Total Working Years



Répartition des salariés exprimée en pourcentage par nombre total d'années travaillées par le salarié.



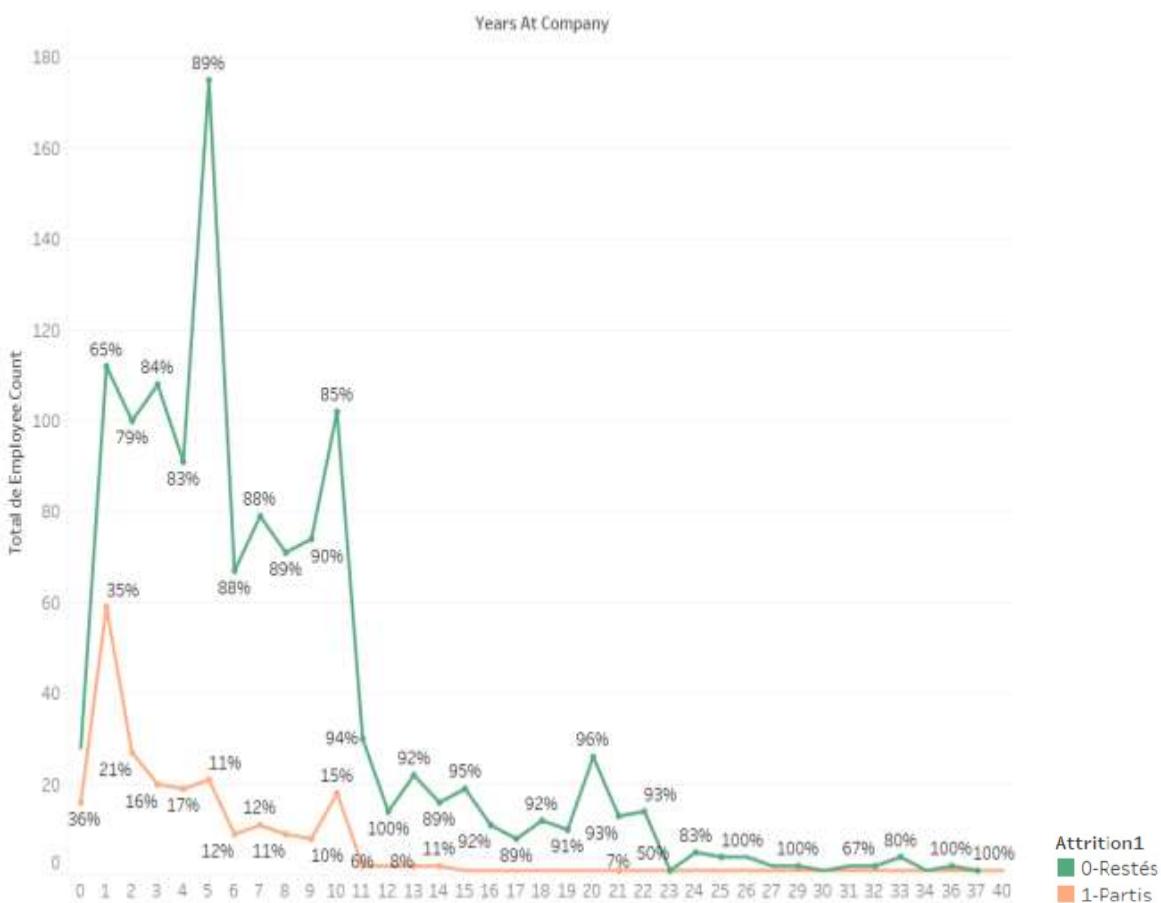
Pourcentage d'attrition par nombre d'années d'expérience professionnelle

Indépendamment du nombre d'employés dans chaque classe (nombre d'années de travail), on remarque que les employés ayant le moins d'expérience professionnelle ont une attrition plus élevée.

Pour l'expérience professionnelle comprise entre 0 et 7 ans, l'attrition est plus élevée que la moyenne de l'entreprise qui est à 16%.

À 40 ans d'expérience professionnelle, on observe 100% d'attrition, il s'agit des départs en retraite.

iv. Years at Company



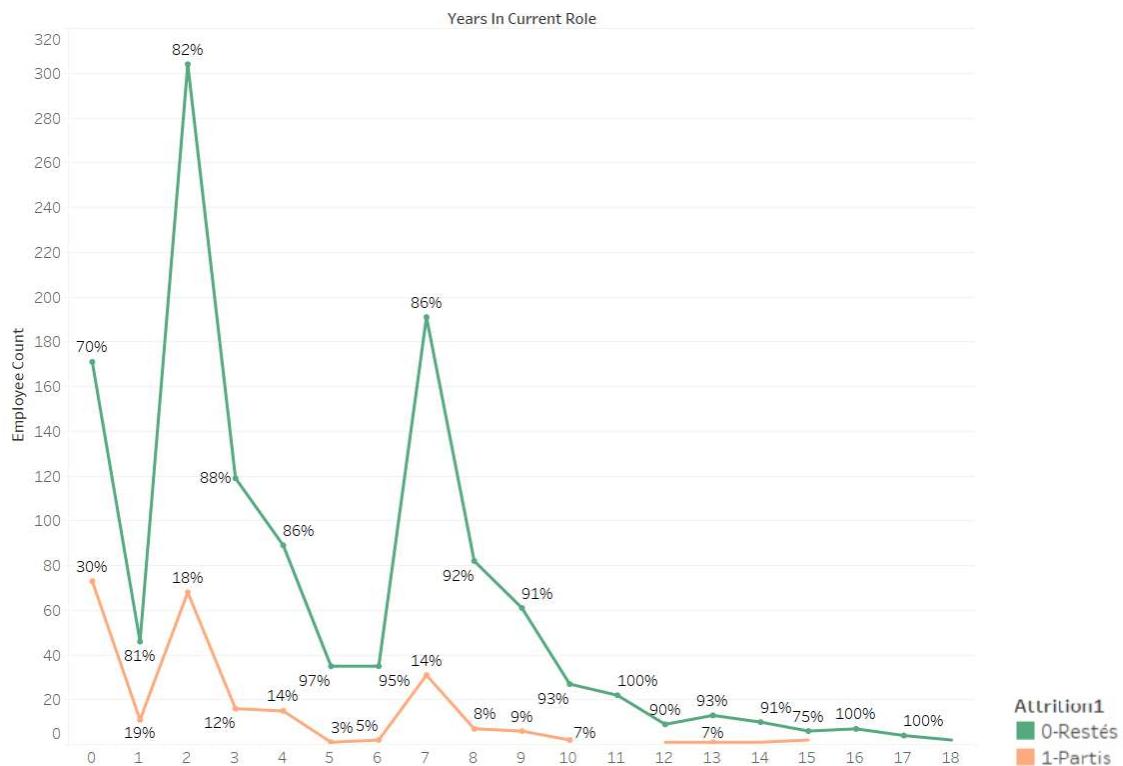
Répartition des employés en fonction de l'attrition et du nombre d'années d'ancienneté dans l'entreprise avec indication de la proportion d'attrition exprimée en pourcentage pour chaque classe

La majorité des employés travaillent depuis moins de 12 ans chez IBM.

Durant les 12 premières années de travail chez IBM, l'attrition diminue lorsque le nombre d'années d'ancienneté augmente.

L'attrition est très élevée les deux premières années d'intégration des salariés (1^{ère} année : 16 salariés partis sur 44 arrivés, 2^{ème} année : 59 départs sur 171 arrivées).

v. Years In Current Role



Répartition des employés en fonction de l'attrition et du nombre d'années sur leur poste actuel avec indication de la proportion d'attrition exprimée en pourcentage pour chaque classe

Nous avons constaté que les salariés sont plus enclins à quitter l'organisation au cours des premières années dans l'entreprise et donc, nécessairement, sur le poste actuel. Il faut donc retirer les nouveaux arrivants sur les valeurs obtenues sur la première année, ceci afin de mieux comprendre l'attrition au moment d'un changement de poste.

Pour rappel, les nouveaux arrivants étaient 44, 28 sont restés et 16 sont partis.

Ici, pour la première année sur un nouveau poste, on observe 171 salariés restés, mais seulement 143 salariés (171-28) sont des collaborateurs ayant changé de poste.

De même, 57 salariés (73-16) sont partis lorsqu'ils étaient sur un nouveau poste de travail.

En proportion, sur 200 salariés (143+57) concernés par un changement de poste, 57 sont attritionnistes, cela représente un taux d'attrition de 28,5 %.

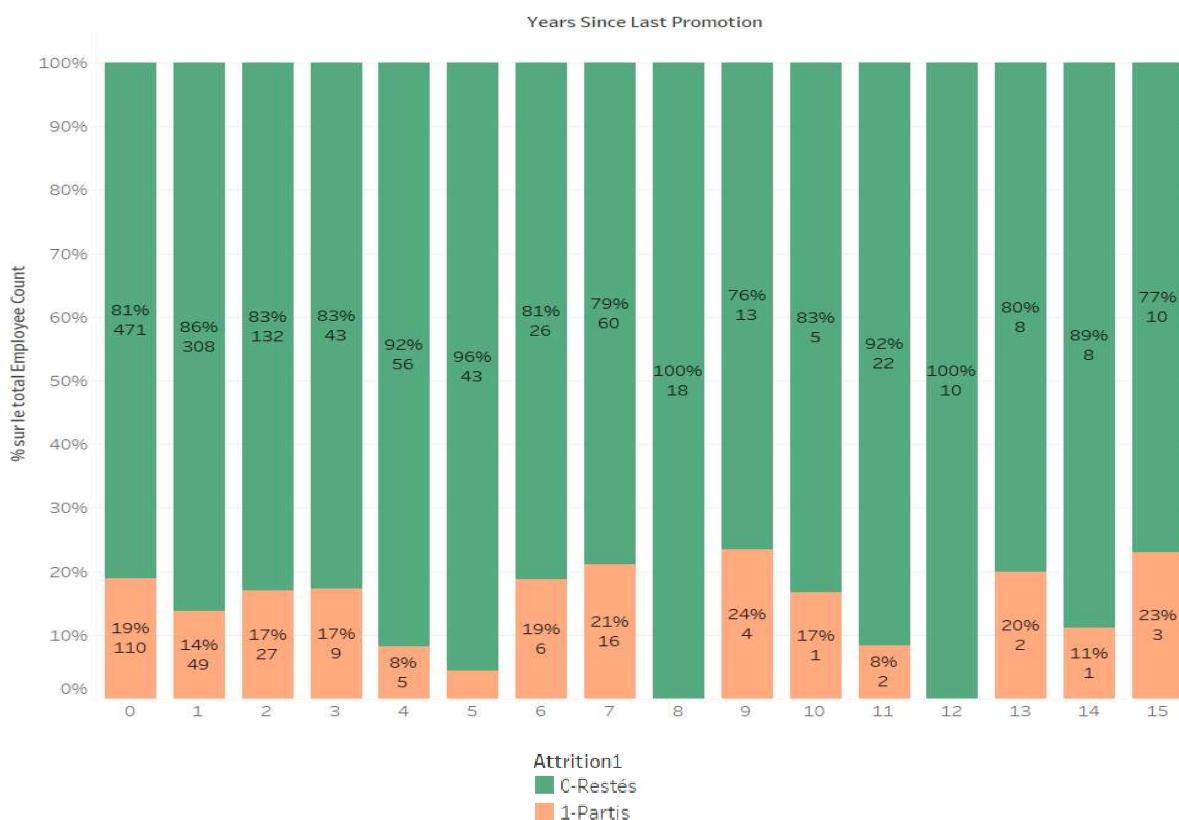
vi. Years Since Last Promotion

Les nouveaux arrivants devraient avoir une classe définie pour signaler qu'ils n'ont pas la possibilité d'avoir eu d'augmentation, en effet, ici, la case remplie est nécessairement '0

année' puisque toutes les valeurs de cette variable étaient renseignées. Or ça suggère une augmentation récente pour les nouveaux arrivants alors qu'il n'y en a pas eu, cela fausse l'interprétation du graphique.

La valeur 0 comprend donc les salariés qui ont été promus cette année et les nouveaux arrivants. Les vraies valeurs à obtenir sont : (Pour rappel, les nouveaux arrivants étaient 44, 28 sont restés et 16 sont partis)

- 443 salariés augmentés cette année sont restés (471-28)
- 94 salariés augmentés cette année sont partis (110-16)
- Pour un total de salariés augmentés cette année de 537 individus
- Attrition de 17,5 % chez les personnes venant d'être augmentées (au lieu de 19%)



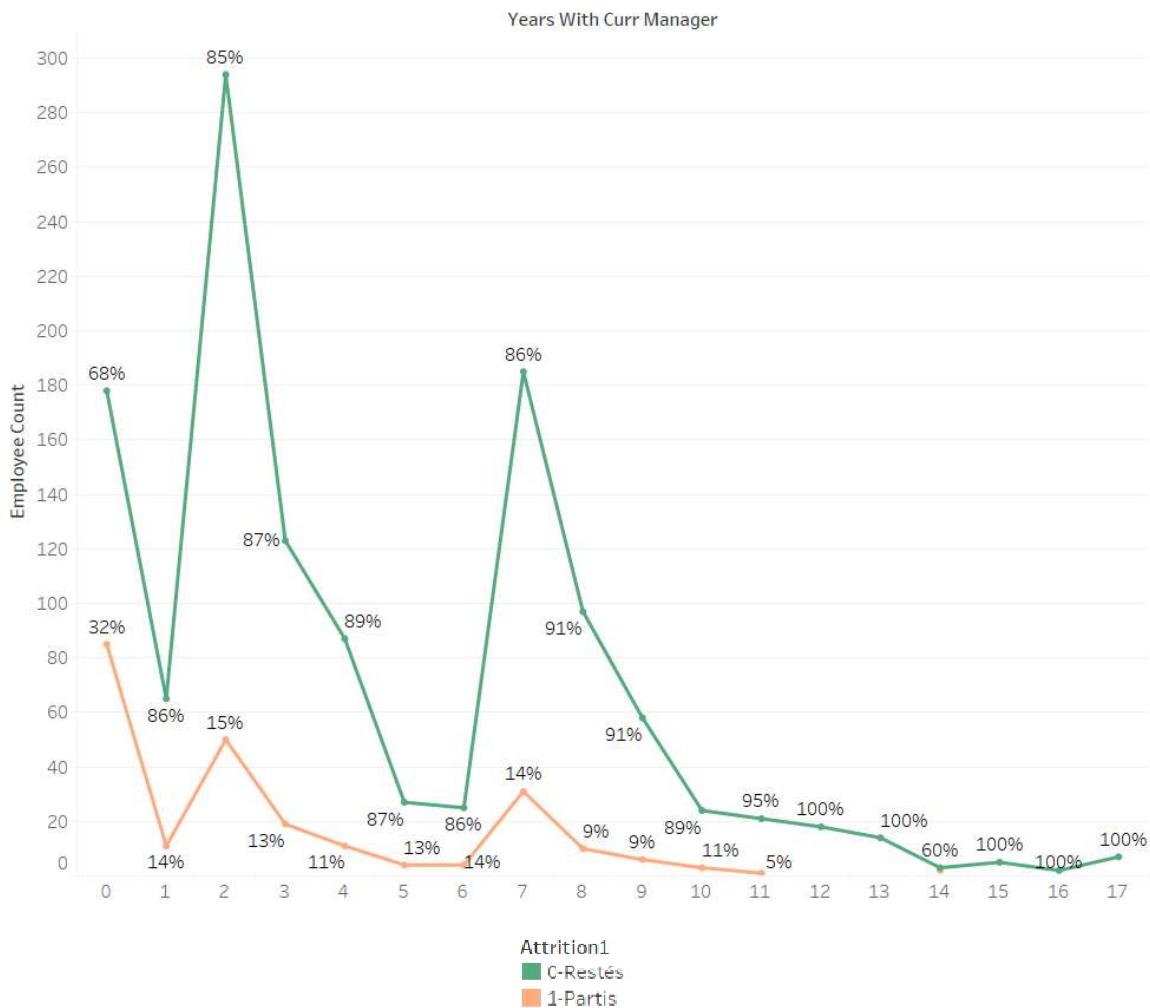
Répartition des employés en fonction du nombre d'années depuis la dernière promotion avec indication de la proportion d'attrition exprimée en pourcentage et en nombre d'individus pour chaque classe

Il est compliqué d'interpréter les résultats par classe 'années depuis dernière promotion' car les taux d'attrition les plus élevés observés sont dans les classes représentées par peu d'individus, donc chaque individu prend une part importante sur le pourcentage.

On observe une attrition assez proche de l'attrition moyenne de l'entreprise pour l'ensemble des classes.

f) Years With Current Manager

Pour la première année passée avec un nouveau manager (mais aussi la première année dans l'entreprise), on voit un pic d'attrition à 32%. Ici 85 partants et 178 restants soit 263 personnes.



Répartition des salariés en fonction de l'attrition et du nombre d'années passées avec le même manager avec indication de la proportion d'attrition exprimée en pourcentage pour chaque classe

Si on retire les nouveaux arrivants (28 restés et 16 partis), afin d'obtenir uniquement les cas où le salarié a eu un changement de manager, on obtient :

- 219 salariés ayant au moins une année d'ancienneté ont changés de manager (263-44).
- 69 d'entre eux sont partis (85-16)

Soit 31,5 % de départs liés à un changement de manager.

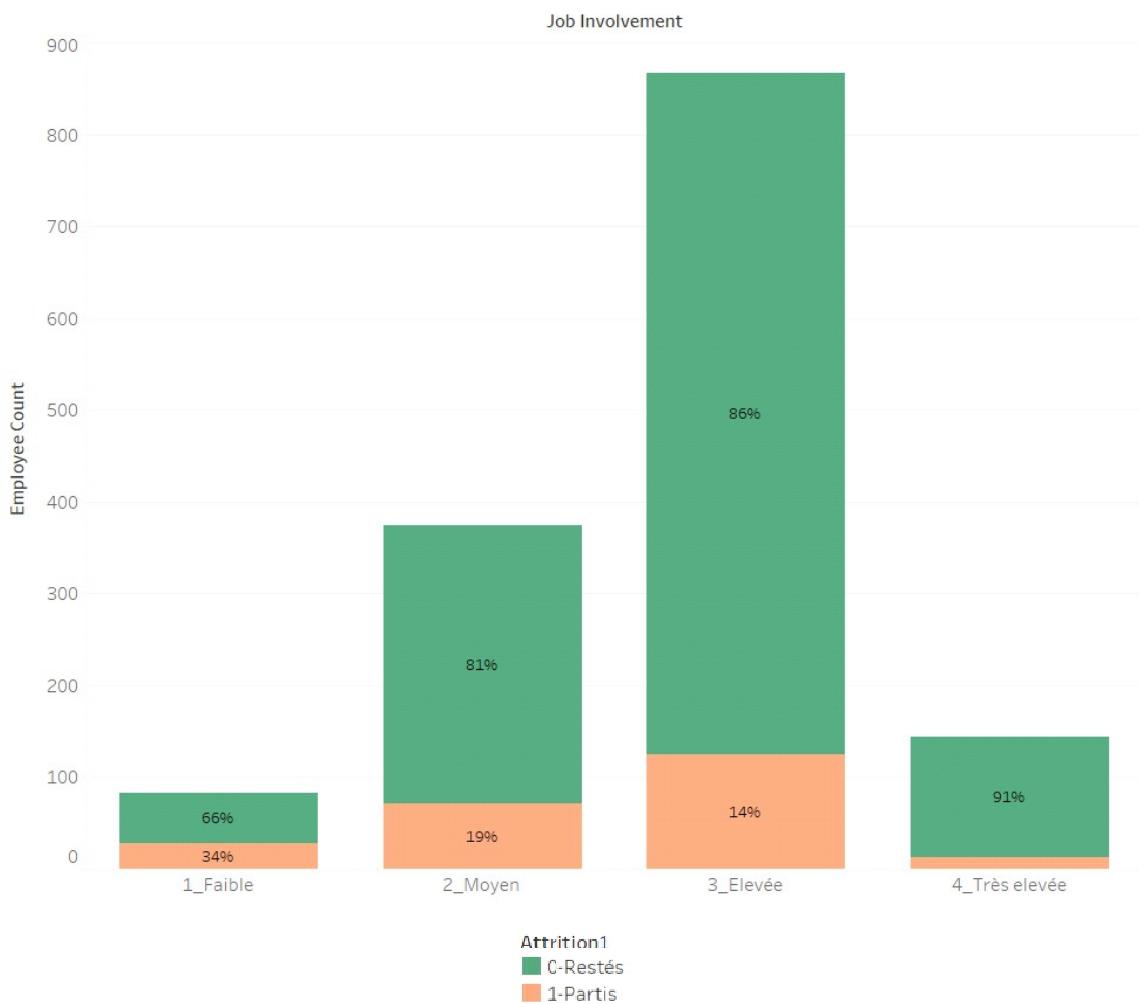
Après un an avec le même manager, l'attrition reste approximativement la même.

g) Job Involvement

Peu d'employés (83) ont été qualifiés de faiblement impliqués dans leur travail, 34% d'entre eux sont partis.

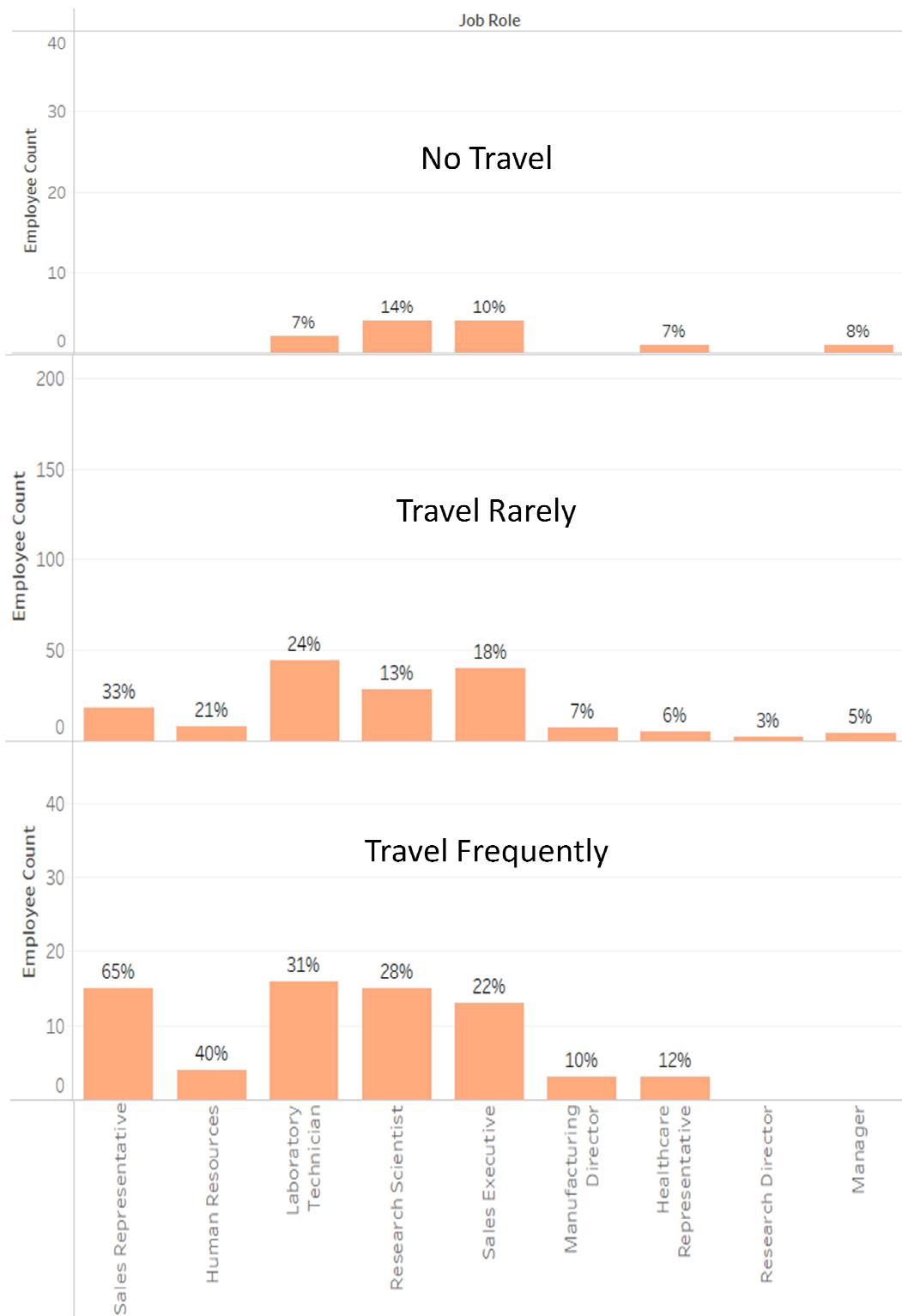
Les employés dont l'implication professionnelle a été jugée 'moyenne' ont un taux d'attrition supérieur au taux de l'entreprise (19% pour 16% sur l'entreprise entière).

La majorité des employés sont considérés comme ayant une implication professionnelle élevée.

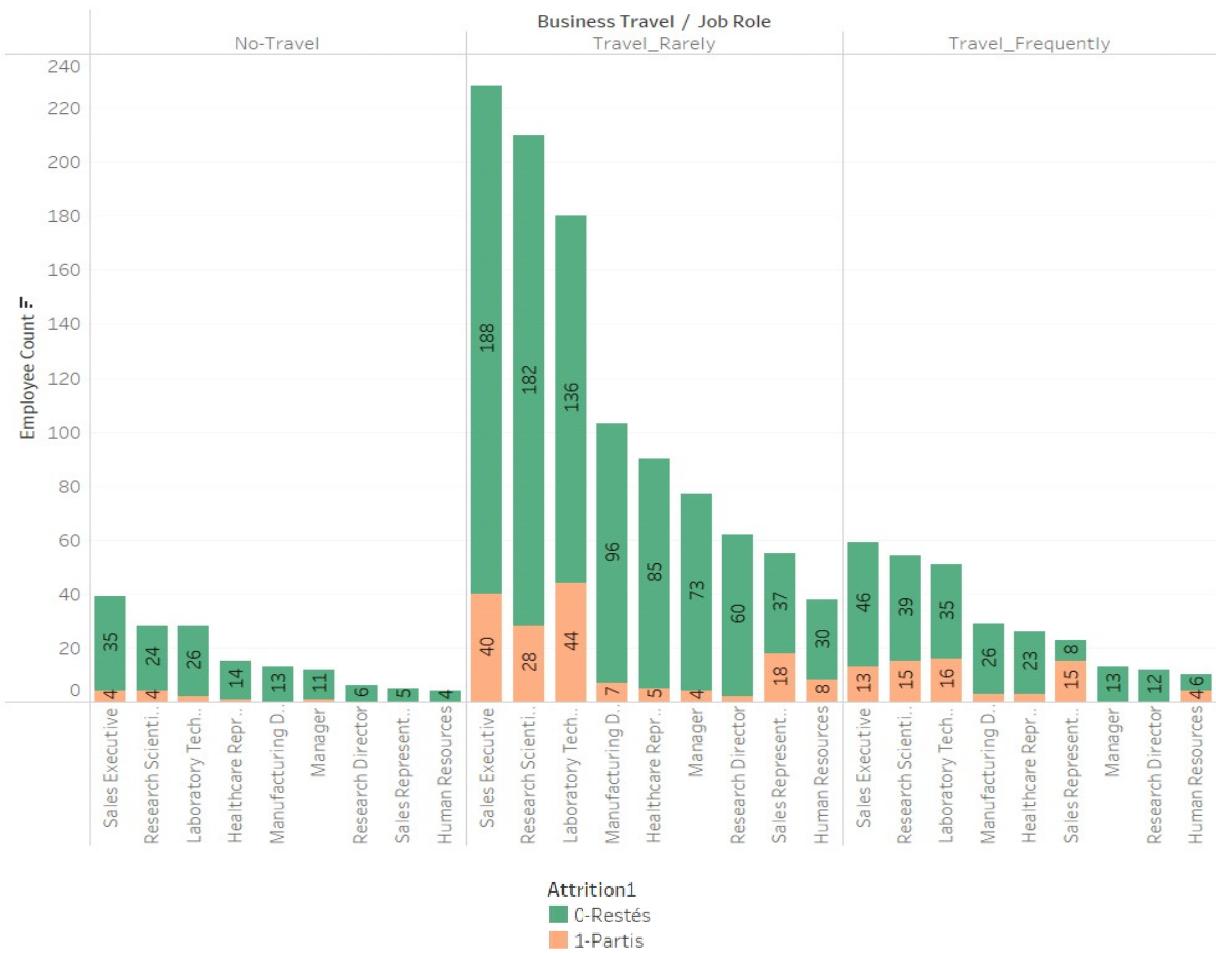


Répartition des salariés en fonction de l'implication professionnelle estimée avec indication de la proportion d'attrition exprimée en pourcentage pour chaque classe

h) Business Travel



Taux d'attrition par métier en fonction de la fréquence des déplacements professionnels



Répartition des salariés et attrition en fonction du poste occupé et de la fréquence des déplacements professionnels

Des déplacements professionnels sont effectués dans chaque département et pour chaque poste (JobRole). La majorité des employés voyagent rarement pour le travail.

Plus les voyages augmentent en fréquence plus l'attrition devient importante pour les métiers :

- Sale representative
- Human Resources
- Laboratory Technician
- Research Scientist
- Sales executive
- Manufactory Director
- Healthcare Representative

Pour un même métier, certains employés font plus de déplacements professionnels que d'autres. Il serait intéressant de comprendre pour quelles raisons certains employés ont beaucoup de déplacements professionnels tandis que d'autres n'en n'ont pas ou peu.

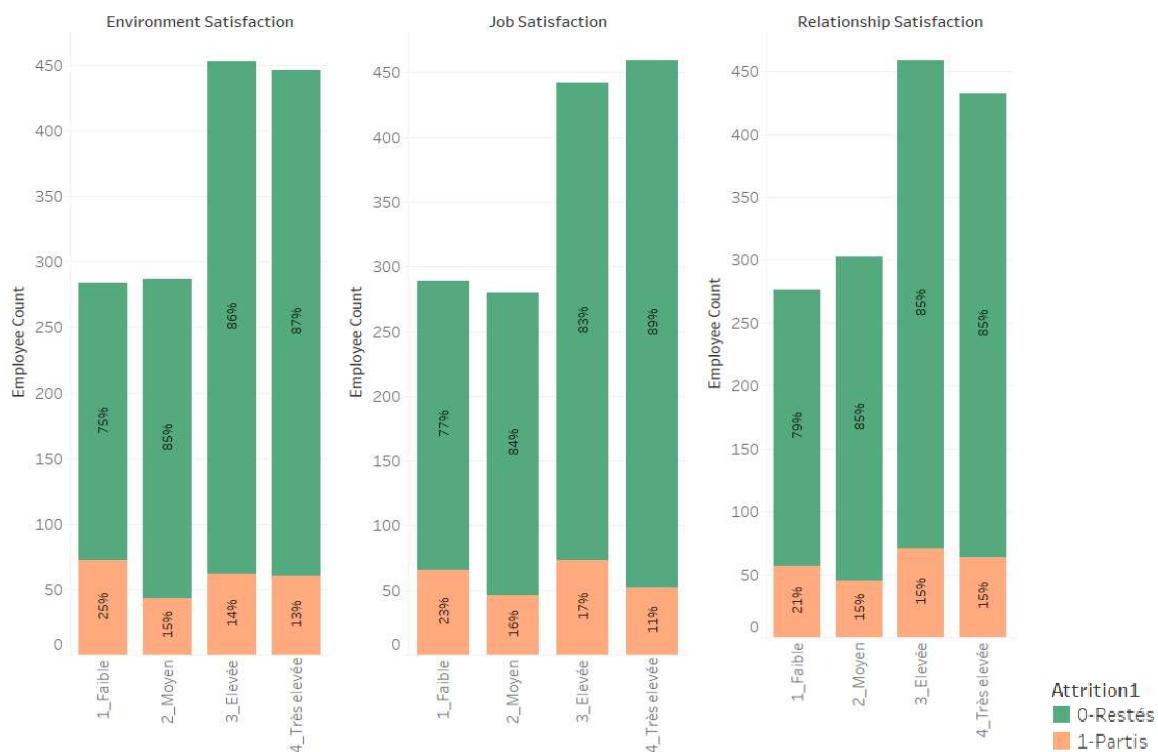
Le taux d'attrition augmente lorsque la fréquence de déplacement augmente à l'exception des métiers 'Research Director' et 'Manager'.

i) Marital Status

L'entreprise n'a pas d'impact sur le statut marital des salariés et il est illégal d'en tenir compte au recrutement, donc nous ne tiendrons pas compte de cette variable.

j) Questionnaire de satisfaction

i. Environment / Job / Relationship Satisfaction



Evaluation de la satisfaction par les employés sur leur environnement, leur métier et leurs relations professionnels avec les proportions d'attrition exprimée en pourcentage

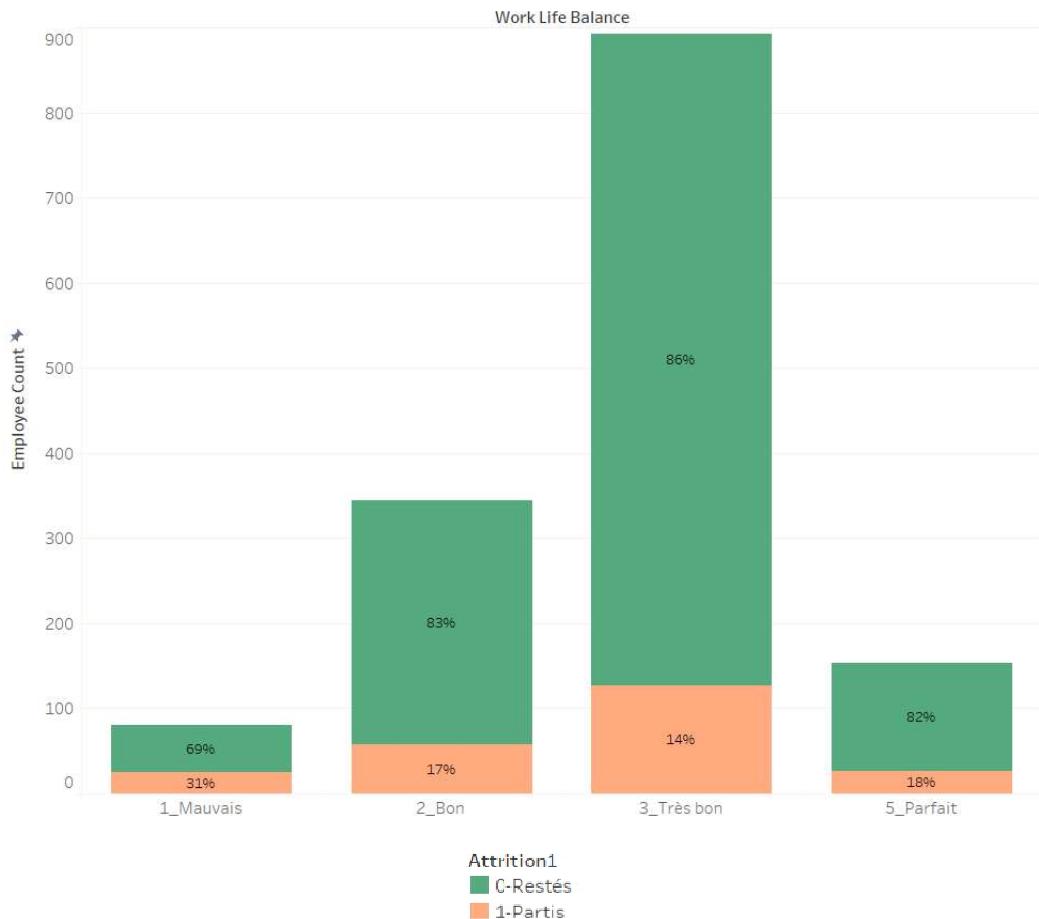
Les employés dont la satisfaction, quelle soit de l'environnement, du métier ou des relations professionnelles, a été évaluée comme 'faible' ont un taux d'attrition supérieur au taux de l'entreprise (respectivement 25%, 23 %, 21% pour 16% sur l'entreprise entière).

A partir d'une note de 'moyen' pour chacune des évaluations de la satisfaction, l'attrition revient à la normale. On remarque tout de même que la majorité des salariés sont satisfaits

dans l'entreprise puisqu'ils notent "élevée" et « très élevée » les différentes satisfactions mesurées.

Cette notation effectuée par les salariés semble être liée à l'attrition lorsque la case « faible » est cochée.

ii. Work Life Balance

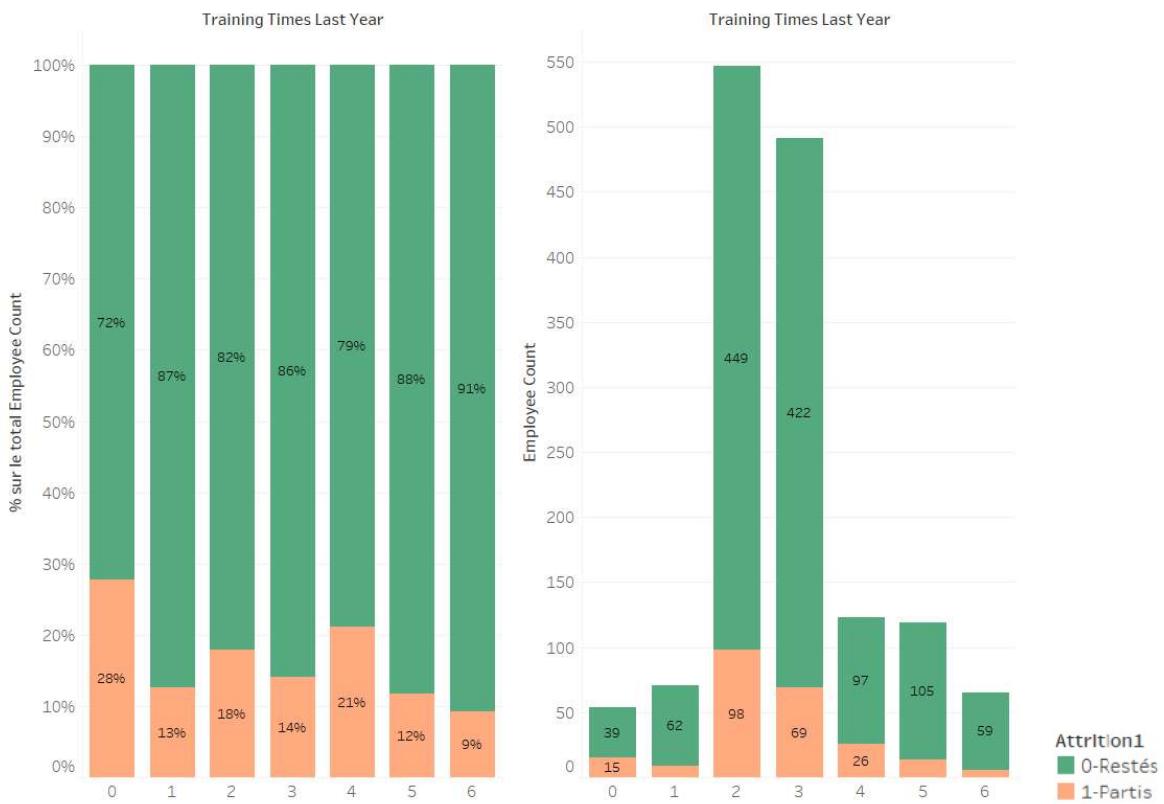


Evaluation de l'équilibre entre vie professionnelle et vie privée par les employés avec les proportions d'attrition exprimée en pourcentage

La grande majorité des employés ont un très bon équilibre entre leur vie professionnelle et leur vie personnelle.

Les employés dont l'équilibre est jugé mauvais ont un taux d'attrition très élevé (31%). Cependant il est faut remarquer que malgré une estimation « parfaite » de leur équilibre, certains salariés quittent l'entreprise (18% contre 16% en moyenne chez IBM).

k) Training Times Last Year



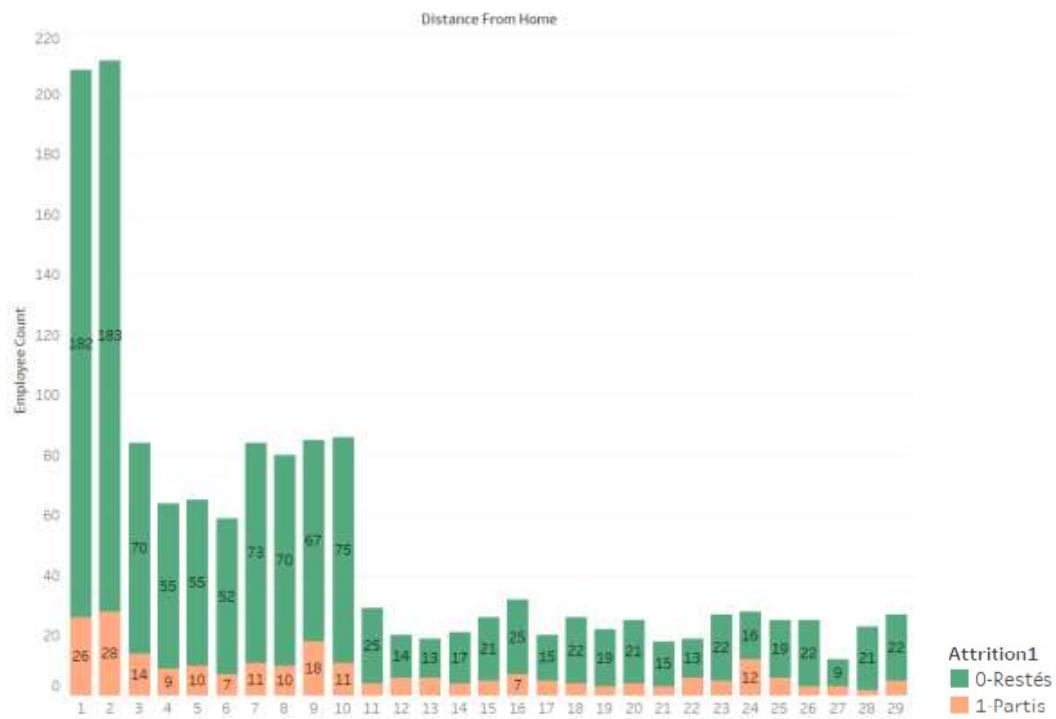
Répartition des salariés en fonction du nombre de formations suivies l'année précédente avec indication de la proportion d'attrition exprimée en pourcentage pour chaque classe

Nous constatons que la valeur attribuée au nombre de formations suivies l'année précédente ne correspond pas à des formations reçues par le biais de l'entreprise puisque les salariés arrivants cette année ne sont pas inclus dans la classe '0 formation'. En effet, il y avait 16 attritionnistes chez les nouveaux arrivants tandis qu'on ne comptabilise que 15 attritionnistes pour la classe '0 formation'.

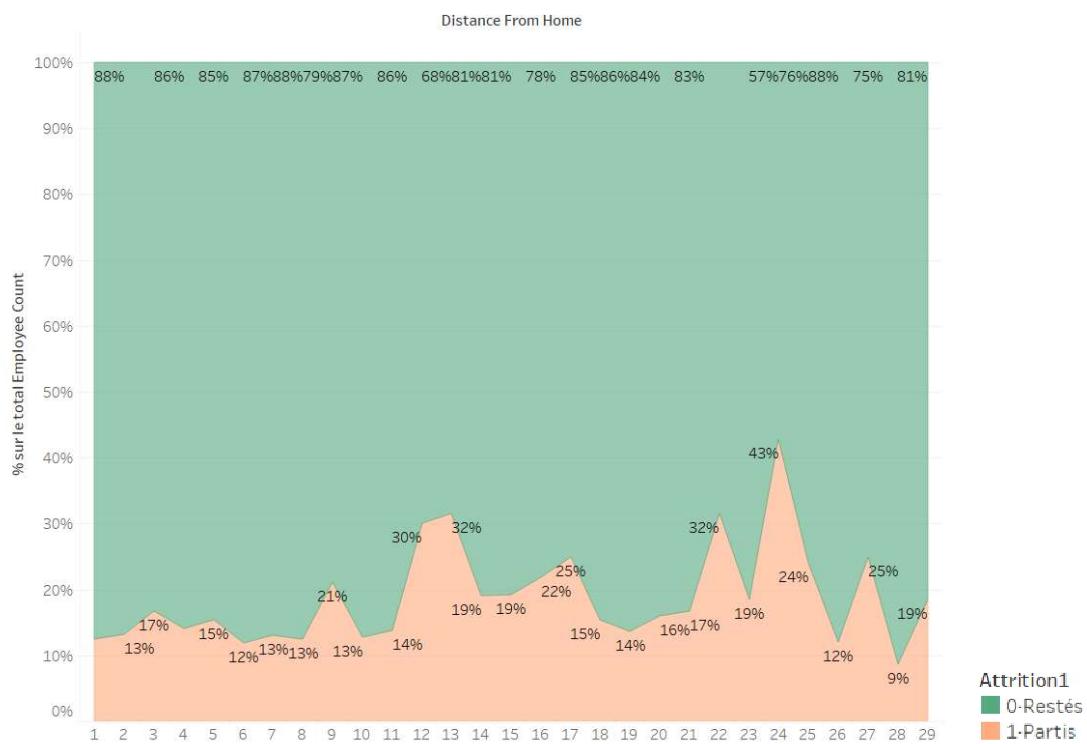
Il s'agit bien de tout type de formation puisqu'en regardant le jeu de données, on remarque des employés avec 0 année d'ancienneté dans l'entreprise ayant plusieurs formations l'année précédente.

Cela implique que les résultats pour cette variable sont moins intéressants à étudier. En effet, l'intérêt de l'accès à la formation (pour réduire l'attrition) n'est appréciable qu'à condition que ce soit l'entreprise qui propose la formation continue.

I) Distance from Home



Répartition des salariés en fonction de la distance parcourue pour se rendre au travail avec indication de la proportion d'attrition exprimée en nombre d'individus pour chaque classe



Proportion d'attrition en fonction de la distance travail-maison à parcourir

La majorité des employés ont moins de 10 km à parcourir pour se rendre sur leur lieu de travail.

Les pics d'attrition observés sur le second graphique ne sont pas significatifs, en effet, peu d'individus sont dans les classes correspondantes et donc chaque attritionniste à un poids trop élevé sur la proportion d'attrition.

2. Conclusions sur la visualisation des données

Nous observons la fluctuation du personnel sur une seule année, et nous n'avons que 1470 données par variable. Il faudra rester vigilant sur l'interprétation des résultats.

Il est important de rappeler que l'attrition représente tous les départs de l'entreprise, qu'ils soient volontaires ou involontaires et donc qu'il ne s'agit pas toujours d'une décision prise par le salarié en lien avec les facteurs mis en exergue dans le jeu de données.

Le taux d'attrition moyen d'une entreprise s'élève aux alentours de 10%, tandis que chez IBM, il s'élève à 16,1 %.

L'étude des variables en fonction de l'attrition grâce au logiciel TABLEAU nous a permis de révéler quelques facteurs pouvant être à l'origine des départs des salariés :

- Le département dont dépend un salarié pourrait être lié à son départ.
- La mission de l'employé (JobRole) semble très impactante sur l'attrition.
- Le domaine d'étude semble déterminant.
- Le niveau d'emploi serait inversement proportionnel au taux d'attrition.
- Le salaire médian mensuel des attritionnistes, très inférieur au salaire médian des non-attritionnistes, traduit l'importance du salaire.
- Les employés qui font beaucoup d'heures supplémentaires ont plus tendance à quitter l'entreprise.
- À mesure que l'âge des salariés augmente, l'attrition diminue.
- Les employés ayant plus d'années d'expérience sont moins susceptibles de quitter l'entreprise.
- L'attrition diminue lorsque le nombre d'années d'ancienneté augmente.
- Plus un employé reste longtemps sur la même mission (YearsInCurrentRole), plus il a tendance à rester dans l'entreprise.
- Le changement de manager semble faire augmenter l'attrition.
- Les employés plus impliqués dans leur travail sont moins enclins à quitter l'entreprise.
- Le taux d'attrition augmente lorsque la fréquence des déplacements professionnels augmente.
- Les employés satisfaits de leur travail sont moins susceptibles de quitter l'entreprise.

- Les employés ayant un faible niveau de satisfaction de leur environnement semblent plus enclins à partir.
- Un bon équilibre entre la vie privée et la vie professionnelle réduit l'attrition.

À ce stade, et avant les analyses statistiques, nous supprimons des colonnes du jeu de données :

- **MaritalStatut** n'est pas une variable sur laquelle il est possible d'apporter un changement, cette variable est donc éliminée.
- **EmployeeCount** n'était utile que pour comptabiliser les salariés dans les catégories des différentes variables, nous pouvons donc la supprimer.
- Nous envisageons d'éliminer trois des variables liées au salaire des employés (DailyRate, HourlyRate, MonthlyRate et MonthlyIncome) car elles semblent redondantes. Cependant, nous n'avons pas encore déterminé laquelle conserver. En effet, les différences salariales observées avec MonthlyIncome pourraient résulter de clauses contractuelles différentes en termes d'heures de travail hebdomadaire. D'autre part, les valeurs des trois autres variables salariales suscitent des interrogations quant à la répartition financière des employés. Nous comptons sur les résultats des tests statistiques pour guider notre choix de manière éclairée.

XII. Analyses des corrélations entre les variables

Les variables ayant une faible relation, ou pas de relation, avec l'attrition peuvent être éliminées de notre modèle de Machine Learning, en effet, le modèle sera plus robuste et plus performant.

Avec l'analyse faite par le logiciel TABLEAU, nous avons pu remarquer que certaines variables ne semblaient pas avoir d'impact sur l'attrition.

Afin de vérifier ces relations, nous a fait plusieurs tests statistiques :

* Vérification de la multi-colinéarité :

Entre les variables numériques et à l'aide d'un **corrélogramme et du calcul du Variance Inflation Factor**.

=> Si des variables sont trop corrélées entre elles, nous garderons celle qui semble la plus pertinente ou la plus interprétable.

* Analyser la relation avec la cible :

- Tests de corrélation du Point-Bisérial pour les variables numériques.

- Tests statistiques Chi² pour les variables catégorielles.

=> Si des variables sont statistiquement non significatives, et que nous jugeons qu'elles ne peuvent pas servir à un calcul d'une nouvelle variable, nous ne les garderons pas.

* Importance des variables :

Nous avons utilisé un modèle de Random Forest (fait rapidement) pour déterminer l'importance de chaque variable qu'elle soit catégorielles ou numériques.

1. Corrélogramme

Le corrélogramme ne permet pas de voir les corrélations avec la cible (Attrition) puisqu'il s'agit d'une variable catégorielle. Cependant, il nous montre les relations entre les variables numériques et nous permet ainsi de vérifier qu'il n'y a pas de trop fortes relations entre elles traduisant des informations redondantes (multi-colinéarité).

i. Résultats du corrélogramme

Le corrélogramme (cf fichier ipynb du projet) montre une relation importante entre :

- Age et Ancienneté chez IBM (0.31) :
- Age et Nombre d'années travaillées (0.68)
- Age et Salaire mensuel (0.5)
- Ancienneté chez IBM et Nombre d'années travaillées (0.63)
- Ancienneté chez IBM et Salaire mensuel (0.51)
- Nombre d'années travaillées et Salaire mensuel (0.77)
- Ancienneté chez IBM et Nombre d'années avec l'actuel manager (0.77)
- Ancienneté chez IBM et Nombre d'années sur le même poste (0.76)

ii. Interprétations

Une règle courante est de considérer que des variables avec une corrélation absolue supérieure à 0.8 ou 0.9 sont fortement corrélées et traduisent une multi-colinéarité.

Dans notre cas, nous n'observons aucune corrélation supérieure à 0.77.

iii. Conclusion

À ce stade, nous ne pouvons pas envisager d'éliminer des variables pour notre modèle de Machine Learning.

2. Variance Inflation Factor

Le Facteur d'Inflation de la Variance (VIF) est une mesure utile pour détecter la multicolinéarité dans les modèles de régression, autrement dit il permet la détection de variables redondantes.

La multicolinéarité se produit lorsque deux ou plusieurs variables indépendantes sont fortement corrélées entre elles et montrent donc un même phénomène.

En Machine Learning, cela peut réduire la fiabilité des coefficients estimés du modèle et nuire à l'interprétation des résultats.

i. Résultats du VIF

Variables	VIF
Age	1.993
DailyRate	1.012
DistanceFromHome	1.009
HourlyRate	1.007
MonthlyIncome	2.529
MonthlyRate	1.009
NumCompaniesWorked	1.244
PercentSalaryHike	1.006
StockOptionLevel	1.014

Variables	VIF
TotalWorkingYears	4.650
TrainingTimesLastYear	1.008
YearsAtCompany	4.565
YearsInCurrentRole	2.693
YearsSinceLastPromotion	1.669
YearsWithCurrManager	2.748

ii. Interprétations

Certaines de nos variables prédictives obtiennent un VIF proche de 1. Cela signifie que la variance de la variable étudiée n'est pas augmentée par les autres variables. Ainsi, nous pouvons dire qu'elles ne présentent pas de colinéarité avec les autres variables du modèle. C'est le cas pour les variables suivantes :

- DailyRate
- DistanceFromHome
- HourlyRate
- MonthlyRate
- PercentSalaryHike
- StockOptionLevel
- TrainingTimesLastYear

Les autres variables étudiées obtiennent un VIF compris entre 1 et 5. Il y a donc une colinéarité modérée avec les autres variables pour chacune d'entre elles.

Deux variables prédictives ont un VIF très élevé, proche de 5, TotalWorkingYears et YearsAtCompany. Bien que des valeurs de VIF inférieures à 5 ne permettent pas de présager d'un problème de multicolinéarité dans le modèle de prédiction, il convient de surveiller ces prédicteurs afin de voir si la multicolinéarité devient un problème.

iii. Conclusion

À ce stade, nous ne pouvons pas envisager d'éliminer des variables pour notre modèle de Machine Learning mais nous surveillerons les variables TotalWorkingYears et YearsAtCompany.

3. Corrélation du point bisérial

Afin d'évaluer la relation entre chaque variable numérique et la cible, il est recommandé de calculer la corrélation du point bisérial plutôt que le coefficient de corrélation de Pearson car notre variable cible est catégorielle et binaire (dichotomique).

i. Résultats de la corrélation du point bisérial

Variables	point bisérial	p_value
Entre Age et Attrition :	-0.159	p-value: 0.0
Entre DailyRate et Attrition :	-0.057	p-value: 0.03
Entre DistanceFromHome et Attrition :	+0.078	p-value: 0.003
Entre HourlyRate et Attrition :	-0.007	p-value: 0.793
Entre MonthlyIncome et Attrition :	-0.16	p-value: 0.0
Entre MonthlyRate et Attrition :	+0.015	p-value: 0.561
Entre NumCompaniesWorked et Attrition :	+0.043	p-value: 0.096
Entre PercentSalaryHike et Attrition :	-0.013	p-value: 0.606
Entre StockOptionLevel et Attrition :	-0.137	p-value: 0.0
Entre TotalWorkingYears et Attrition :	-0.171	p-value: 0.0
Entre TrainingTimesLastYear et Attrition :	-0.059	p-value: 0.023
Entre YearsAtCompany et Attrition :	-0.134	p-value: 0.0
Entre YearsInCurrentRole et Attrition :	-0.161	p-value: 0.0
Entre YearsSinceLastPromotion et Attrition :	-0.033	p-value: 0.206
Entre YearsWithCurrManager et Attrition :	-0.156	p-value: 0.0

ii. Interprétations

Les variables dont la p-value est supérieure à 0,05 ne montrent pas de corrélation significative avec la variable cible Attrition. C'est le cas des variables HourlyRate, MonthlyRate, NumCompaniesWorked, PercentSalaryHike, et YearsSinceLastPromotion.

Les variables dont la valeur de p-value est inférieure à 0,05 ont une corrélation statistiquement significative avec la cible. Elles peuvent être divisées en 3 catégories :

- Une corrélation comprise entre -0,1 et +0,1 signifie que la relation linéaire entre la variable étudiée et la cible est faible ou négligeable. Cela signifie qu'il n'y a pratiquement pas de relation linéaire entre les deux variables mais puisque la p-value

était inférieure à 0,05, l'hypothèse d'une corrélation non significative a été écartée. C'est le cas pour les variables DailyRate, DistanceFromHome, TrainingTimesLastYear.

- Une corrélation comprise entre +0,1 et + 1 signifie une relation positive. Ce n'est le cas d'aucune variable.

- Une corrélation comprise entre -0,1 et -1 signifie une relation négative, c'est le cas pour les variables Age, MonthlyIncome, StockOptionLevel, TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsWithCurrManager.

ii. Conclusion

Les variables HourlyRate, MonthlyRate, NumCompaniesWorked, PercentSalaryHike, et YearsSinceLastPromotion ne montrent pas d'impact significatif sur l'attrition, ce qui signifie qu'elles pourraient ne pas être utiles pour prédire si un employé quittera l'entreprise.

Les variables Age, MonthlyIncome, StockOptionLevel, TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsWithCurrManager ont un impact négatif sur l'attrition. Autrement dit, les employés :

- plus âgés,
- ayant un salaire plus élevé
- et une quantité de stock options plus importante
- plus d'années d'expérience professionnelle à leur cursus
- ou travaillant depuis plus longtemps dans l'entreprise
- ou plus longtemps sur le même emploi
- et ayant le même manager depuis plus longtemps

... ont tendance à rester plus longtemps que les autres salariés (l'attrition est diminuée dans chacun de ces cas).

Ces variables peuvent être utilisées dans les modèles prédictifs de Machine Learning.

4. Test de Chi-2

Le test du Chi² permet de tester s'il existe une relation significative entre deux variables catégorielles. Dans notre cas, nous allons tester les variables catégorielles avec notre cible.

Les résultats peuvent aider à déterminer quelles variables sont les plus importantes pour prédire l'attrition des employés dans un modèle de Machine Learning. En effet, des variables

avec une p-value significative peuvent être de bons candidats pour la prédiction des départs chez IBM.

i. Résultats du Chi²

Variables	Chi ²	p_value
Entre OverTime et Attrition :	87.564	p-value: 0.0
Entre JobRole et Attrition :	86.19	p-value: 0.0
Entre JobLevel et Attrition :	72.529	p-value: 0.0
Entre JobInvolvement et Attrition :	28.492	p-value: 0.0
Entre BusinessTravel et Attrition :	24.182	p-value: 0.0
Entre EnvironmentSatisfaction et Attrition :	22.504	p-value: 0.0
Entre JobSatisfaction et Attrition :	17.505	p-value: 0.001
Entre WorkLifeBalance et Attrition :	16.325	p-value: 0.001
Entre EducationField et Attrition :	16.025	p-value: 0.007
Entre Department et Attrition :	10.796	p-value: 0.005
Entre RelationshipSatisfaction et Attrition :	5.241	p-value: 0.155
Entre Education et Attrition :	3.074	p-value: 0.546
Entre Gender et Attrition :	1.117	p-value: 0.291

ii. Interprétations

Les variables Gender, Education et RelationshipSatisfaction ont une p-value supérieure à 0,05, elles ne sont donc pas significativement associées à l'attrition.

Les variables significativement associées à l'attrition (p-value < 0.05) sont :

- BusinessTravel
- Department
- EducationField
- EnvironmentSatisfaction
- JobInvolvement
- JobLevel
- JobRole
- JobSatisfaction
- OverTime
- WorkLifeBalance

ii. Conclusion

Le genre, le niveau d'éducation et la satisfaction dans les relations de travail ne semblent pas être en lien avec la décision de départ d'un salarié. Il est donc envisageable d'écartier ces variables du modèle de Machine Learning.

La réalisation d'heures supplémentaires, le rôle occupé ainsi que le niveau d'emploi d'un salarié ont un impact considérable sur son attrition.

Les employés les plus impliqués dans leur travail sont probablement moins enclins à quitter leur emploi tandis que les salariés ayant un faible niveau de satisfaction concernant leur environnement de travail semblent plus enclins à partir de l'entreprise.

La fréquence des déplacements professionnels pourrait influencer l'attrition d'un individu. En effet, une fréquence élevée de déplacements est souvent associée à un taux d'attrition plus important que des déplacements moins fréquents.

Les employés satisfaits de leur métier sont moins susceptibles de quitter l'entreprise. Un bon équilibre entre vie professionnelle et vie privée réduit l'attrition. Le domaine d'étude d'un employé peut avoir un impact sur son taux d'attrition.

Les résultats ont montré que le département auquel l'employé appartient peut avoir un impact sur l'attrition.

XIII. Choix des variables pour le Machine Learning

Afin de construire un modèle de Machine Learning robuste et performant, il est intéressant de réduire les variables à intégrer. En effet, seules les variables pertinentes et ayant une forte relation avec la cible permettront une prédiction juste et une réduction du risque de surapprentissage (overfitting).

L'analyse statistique descriptive avec python, la visualisation des données avec le logiciel TABLEAU, l'analyse de colinéarité et l'analyse des corrélations entre les variables et la cible nous ont permis de faire une sélection éclairée des variables choisies pour notre modèle de Machine Learning.

Concernant le choix à faire entre les variables financières MonthlyIncome et la DailyRate, nous avons choisi la variable MonthlyIncome qui a présenté une plus grande importance lors

d'un premier essai de modèle de Machine Learning. Autrement dit, elle contribue mieux à la prédiction globale du modèle.

Listes des variables éliminées pour le Machine Learning :

DailyRate	NumCompaniesWorked
Education	Over18
EmployeeCount	PercentSalaryHike
EmployeeNumber	PerformanceRating
Gender	RelationshipSatisfaction
HourlyRate	StandardHours
MaritalStatus	YearsSinceLastPromotion
MonthlyRate	

Listes des variables utilisées comme 'prédicteurs' pour le Machine Learning :

Variables numériques (9)

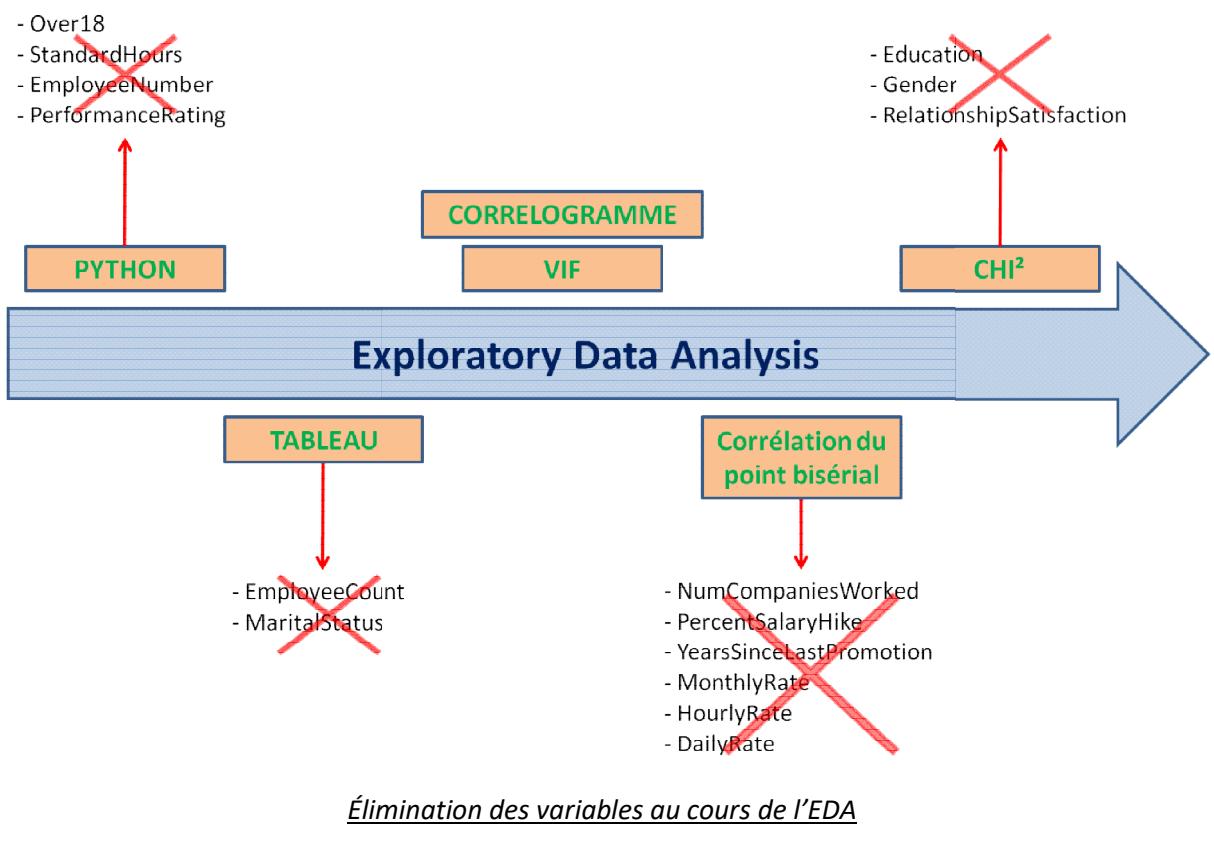
Age
DistanceFromHome
MonthlyIncome
StockOptionLevel
TotalWorkingYears
TrainingTimeLastYear
YearsAtCompany
YearsInCurrentRole
YearsWithCurrManager

Variables catégorielles (10)

BusinessTravel
Department
EducationField
EnvironmentSatisfaction
JobInvolvement
JobLevel
JobRole
JobSatisfaction
OverTime
WorkLifeBalance

Nous décidons de supprimer les personnes de plus de 55 ans puisqu'ils ne prennent plus le risque de partir si proche de la retraite.

À l'issue de la suppression d'individus et de variables, il reste 1401 individus et 19 variables prédictives pour le Machine Learning.



XIV. Machine Learning

1. Préparation des données

a) Nettoyage

Le nettoyage du jeu de données (valeurs manquantes, aberrantes ou atypiques, doublons) avait été fait avant l'EDA.

L'EDA nous permet d'éliminer les variables non pertinentes et les variables sans relation avec la cible, celles-ci ont été listées au chapitre « XIII. Choix des variables pour le Machine Learning » ainsi qu'une partie de la population (≥ 55 ans).

b) Modifications

Nous avons repéré que la valeur '-HumanResources-' correspond à une catégorie de la colonne 'JobRole', à une catégorie de la colonne 'Department' et à une catégorie de la colonne 'EducationField'. Nous avons donc apporté les modifications suivantes :

- HumanResources_Job pour 'JobRole'
- HumanResources_Dpt pour 'Department'
- HumanResources_Studies pour 'EducationField'

D'autre part, nous avons une variable dont les valeurs sont 'Yes' ou 'No' , donc nous avons changé :

- OverTime pour le Yes de la colonne 'OverTime'
- No OverTime pour le No de la colonne 'OverTime'

Et enfin, nous avons mis les valeurs Attrition-Yes pour le '1' et Attrition-No pour le '0' pour la variable cible 'Attrition'.

Ces modifications permettront de faciliter la lecture de l'importance des variables.

c) Features - Target SPLIT

Nous avons attribué les variables prédictives à X (Features) et avons désigné la cible 'Attrition' comme étant y (Target).

d) Train - Test SPLIT

Le jeu de données est divisé en deux ensembles : un **Training set** et un **Testing set**. Le Training set est utilisé pour entraîner le modèle sur une partie des données. Le Testing set est utilisé pour évaluer les performances de ce modèle, sur l'autre partie des données.

Test-size : détermine un nombre d'exemples du jeu de données, ici 30%.

Le paramètre **Stratify** choisit si les données sont séparées de façon à garder les mêmes proportions d'observations dans chaque classe dans les ensembles Train et Test que dans le dataset initial. Ce paramètre est particulièrement utile face à des données « unbalanced »

avec des proportions très déséquilibrées entre les différentes classes comme nous avons pu l'observer avec 'Attrition' et avec la variable 'OverTime'. Stratifier selon la **variable y** (=cible) signifie qu'il essaiera de répartir uniformément les valeurs de y dans chaque division.

Le **random state** est un nombre qui contrôle la façon dont le générateur pseudo-aléatoire divise les données. Choisir un nombre entier comme random state permet de séparer les données de la même manière à chaque appel de la fonction. Cela rend donc le code reproductible.

e) Transformation des colonnes

À l'aide des outils de préprocessing de scikit-learn, nous avons préparé les colonnes pour le Machine Learning.

i. Standard scaling

L'outil **StandardScaler** recalibre les données des variables numériques pour obtenir des répartitions normalisées.

ii. Encoding

Le **OneHotEncoding** est une technique utilisée pour convertir les données catégorielles en un format binaire.

`drop="first"` : oublie la première catégorie puisqu'elle peut être prédite par l'absence des autres ainsi, nous évitons la colinéarité des features.

`handle_unknown="ignore"` indique à l'encodeur d'ignorer les catégories inconnues (catégories qui n'ont pas été vues pendant le processus d'adaptation) pendant la transformation.

iii. Fit et Transform

Nous utilisons la fonction **fit()** sur le Training set pour que lors de l'apprentissage du modèle, l'ajustement se fasse en estimant les paramètres (μ et σ) à partir du Training set.

La fonction `fit()` n'est donc pas appliquée au Testing set.

La fonction `transform()` permet de transformer les données en une forme plus adaptée au modèle. Cette fonction est appliquée au Training set et au Testing set.

iv. Label Encoding

Sur la cible `y`, nous appliquons la fonction `LabelEncoding`. Elle convertit les données de notre variable catégorielle cible en données numériques. Nous l'appliquons sur le Training set et le Testing set.

2. Modèles testés

Nous avons testé 3 modèles de Machine Learning, la Logistic Regression (LR), Le Decision Tree (DT) et le Random Forest (RF).

Matrice de confusion

		PRÉDICTION	
		Restés - 0	Partis - 1
R É A L I T É	Restés 0	TN <i>Restés mais vus comme Partis</i>	FP
	Partis 1	FN <i>Partis mais vus comme Restés</i>	TP

Nous cherchons à repérer les individus partis dans notre population. De ce fait, nous nous intéressons aux TP (True Positives) qui sont les individus partis et reconnus 'partis' par nos modèles. Mais nous nous intéressons également aux FN (False Negatives) qui sont les individus partis mais n'ont pas été reconnus comme tels par les modèles.

Nous allons donc essayer d'optimiser l'accuracy de nos modèles (pourcentage de prédictions correctes) mais également le recall (pourcentage des individus réellement partis repérés par le modèle).

Le calcul de l'**ROC-AUC** va également avoir un intérêt. En effet, il nous aidera à juger de la performance de notre modèle puisqu'il mesure la capacité d'un modèle à discriminer les classes dans le cas d'une classification binaire. Une AUC de 1 indique une séparation parfaite entre les classes, tandis qu'une AUC de 0.5 signifie que le modèle ne fait pas mieux qu'un tirage aléatoire.

Les variables prédictives du modèle seront (43):

```
'cat__BusinessTravel_Travel_Frequently',
'cat__BusinessTravel_Travel_Rarely',
'cat__Department_Research &
Development',
'cat__Department_Sales',
'cat__EducationField_Life Sciences',
'cat__EducationField_Marketing',
'cat__EducationField_Medical',
'cat__EducationField_Other',
'cat__EducationField_Technical Degree',
'cat__EnvironmentSatisfaction_2_Moyen',
'cat__EnvironmentSatisfaction_3_Elevée',
'cat__EnvironmentSatisfaction_4_Très
élevée',
'cat__JobInvolvement_2_Moyen',
'cat__JobInvolvement_3_Elevée',
'cat__JobInvolvement_4_Très élevée',
'cat__JobLevel_2_Junior',
'cat__JobLevel_3_Confirmé',
'cat__JobLevel_4_Senior',
'cat__JobLevel_5_Expert',
'cat__JobSatisfaction_2_Moyen',
'cat__JobSatisfaction_3_Elevée',
'cat__JobSatisfaction_4_Très élevée',
'cat__JobRole_HumanResources_Job',
'cat__JobRole_Laboratory Technician',
'cat__JobRole_Manager',
'cat__JobRole_Manufacturing Director',
'cat__JobRole_Research Director',
'cat__JobRole_Research Scientist',
'cat__JobRole_Sales Executive',
'cat__JobRole_Sales Representative',
'cat__OverTime_OverTime-Yes',
'cat__WorkLifeBalance_2_Bon',
'cat__WorkLifeBalance_3_Très bon',
'cat__WorkLifeBalance_5_Parfait',
'num__Age',
'num__DistanceFromHome',
'num__MonthlyIncome',
'num__StockOptionLevel',
'num__TotalWorkingYears',
'num__TrainingTimesLastYear',
'num__YearsAtCompany',
'num__YearsInCurrentRole',
'num__YearsWithCurrManager'
```

Il est à noter que le OneHotEncoding a éliminé une catégorie par variable encodée pour éviter la multicolinéarité. Il sera important de rester vigilant lors de l'interprétation de l'importance des features.

3. Amélioration des modèles

- * Puisque notre jeu de données est déséquilibré, nous nous sommes concentrés sur l'amélioration des modèles LR et RF, qui sont des algorithmes plus robustes pour la gestion des déséquilibres de classes.
- * Le déséquilibre de notre jeu de données ne permettait pas de baser nos résultats sur l'accuracy, donc nous avons regardé le recall et l'AUC. Maximiser le recall des modèles permet d'améliorer leur capacité à détecter la classe positive de notre cible.
- * Le jeu de données étant déséquilibré, nous avons essayé un SMOTE, mais d'autres méthodes peuvent être plus efficaces. Le SMOTE est une technique de suréchantillonnage des minorités synthétiques, autrement dit, il s'agit d'une technique visant à équilibrer la répartition des classes en augmentant de manière aléatoire les exemples de classes prioritaires présents dans le jeu de données en les reproduisant. Le SMOTE n'est appliqué que sur le Train-set et donc ne biaise pas les résultats du Test.
- * Afin d'ajuster le poids des classes sur les modèles testés, une fois de plus dans le but de traiter les déséquilibres de classes, nous associons le paramètre `class_weight='balanced'` aux algorithmes des modèles.
- * Nous avons éliminé des variables du jeu de données pour améliorer nos modèles en les rendant plus robustes et plus performant.
- * Nous avons retiré une partie de la population d'origine, en effet, les individus de plus de 54 ans pouvaient avoir trop de poids sur l'entraînement de nos modèles.
- * Nous avons utilisé l'outil GridSearchCV pour rechercher les meilleurs hyperparamètres des modèles évalués. Nous souhaitions diminuer les Faux Négatifs, puisque ce sont ceux qui partent (même s'ils sont donnés comme restants). Pour cela, nous avons fait une recherche qui optimisait le recall.
- * Chaque changement en vue d'amélioration a été effectué sur les 3 modèles, ici, ne sont présentés que les résultats sur le modèle de Régression Logistique.

		Logistic Regression	Decision Tree	Random Forest
TRAIN	Matrice de confusion Train set	[[678 144] [117 705]]	[[823 40] [216 647]]	[[863 0] [7 856]]
	ACCURACY_Train-set	0.841	0.852	0.996
TEST	Matrice de confusion Test set	[[289 64] [20 48]]	[[349 21] [57 14]]	[[355 15] [50 21]]
	ACCURACY_Test set	0.800	0.823	0.85
		RECALL	71%	20%
		AUC	81%	70%
		F1_score	53%	26%
		Precision	43%	40%
		Train-time	0,04	0,01
		Test-time	0,00	0,01

Tableau de visualisation des 3 modèles testés après amélioration

$$\begin{bmatrix} \text{TN} & \text{FP} \\ \text{FN} & \text{TP} \end{bmatrix}$$

		Logistic Regression AVANT SMOTE	Logistic Regression AVANT HYPERPARAMETRAGES
T R A I N	Matrice de confusion Train set	[[849 14] [91 75]]	[[681 182] [133 730]]
	ACCURACY_Train-set	0,778	0,817
T E S T	Matrice de confusion Test set	[[357 13] [47 24]]	[[307 63] [29 42]]
	ACCURACY_Test set	0,794	0,791
RECALL		34%	59%
AUC		80%	80%
F1_score		44%	48%
Precision		65%	40%
Train-time		0,06	0,02
Test-time		0,00	0,00

		Logistic Regression AVANT SELECTION INDIVIDUS	Logistic Regression AVANT SELECTION VARIABLES
T R A I N	Matrice de confusion Train set	[[849 14] [91 75]]	[[686 133] [109 710]]
	ACCURACY_Train-set	0,824	0.844
T E S T	Matrice de confusion Test set	[[357 13] [47 24]]	[[296 56] [23 45]]
	ACCURACY_Test set	0,790	0.811
RECALL		68%	66%
AUC		82%	81%
F1_score		51%	53%
Precision		41%	45%
Train-time		0,06	0,03
Test-time		0,00	0,00

Résultats de quelques tests fait sur les 3 modèles mais présentés

pour le modèle de Régression Logistique

4. Résultats du Machine Learning

		Logistic Regression
TRAIN	Matrice de confusion Train set	[[678 144] [117 705]]
	ACCURACY_Train-set	0.841
TEST	Matrice de confusion Test set	[[289 64] [20 48]]
	ACCURACY_Test set	0.800
		RECALL
		81%
		AUC
		53%
		Precision
		43%
		Train-time
		0,04
		Test-time
		0,00

Résultats retenus pour notre étude

DECISION TREE : Nous avons eu de l'overfitting avec ce modèle. En effet, nous avions un accuracy-score en training à 1.0 lors de nombreux essais. L'amélioration du modèle avec des accuracy-score acceptables ne nous permet pas de nous intéresser à ce modèle puisque le F1-score n'est qu'à 26%, nous indiquant que le modèle est peu performant. Ce résultat était prévisible, il reflète que ce modèle n'est pas adapté au problème de déséquilibre de classes. Augmenter le F1-score impliquait du surapprentissage.

RANDOM FOREST : Nous avons du mal à augmenter le F1-score sans atteindre l'overfitting. Malgré l'optimisation du recall avec l'hyperparamétrage de GridSearchCV, nous n'arrivons pas à des résultats satisfaisants.

LOGISTIC REGRESSION : La régression logistique est le modèle le plus performant malgré qu'il ne soit pas parfait. Nous ne frôlons pas l'overfitting et les résultats sont intéressants. Voici une interprétation détaillée de chaque métrique :

- * Le matrice de confusion du Training montre un bon équilibre de classes mais montre également que le modèle fait beaucoup d'erreurs.
- * La matrice de confusion du Testing montre qu'il y a encore beaucoup d'individus restés étant prédis comme partis (FP).
- * Lors du Training, le modèle a correctement classé 84,1% des individus et lors du Testing, le modèle a classé correctement 80,0% des cas.
- * Recall à 71% : le modèle parvient à repérer 71% des employés ayant réellement quitté l'entreprise. Le modèle a une bonne capacité à détecter les vrais positifs (TP) même si il peut encore être amélioré puisque 29% des départs ne sont pas prédis.
- * L'AUC est relativement bonne (81%) , cela signifie que le modèle réussi à bien différencier les classes de la cible (Partis et Restés).
- * Le F1-score (53%) est la moyenne harmonique du recall et de la précision de notre modèle. Nous constatons que la précision n'est pas bonne et fait chuter le F1-score, ceci est dû au fait que nous avons dirigé la recherche des meilleurs hyperparamètres du modèle vers un meilleur recall. Ce résultat montre que notre modèle, bien qu'il détecte correctement les vrais positifs (TP), il fait beaucoup d'erreurs sur les faux positifs (FP).
- * La précision du modèle est faible (43%). Seuls 43% des départs prédis sont correctement classés (salariés vraiment partis). Il y a donc beaucoup de faux positifs (FP) générés.
- * Le modèle prend très peu de temps à s'entraîner et à prédire.

Pour conclure, avec un AUC à 81%, le modèle montre une bonne capacité de discrimination entre les employés qui partent et ceux qui ne partent pas, mais il y a de la place pour l'amélioration. Le modèle est sensible à la détection des départs réels (recall de 71%) mais a du mal à limiter les faux positifs puisqu'il a une précision relativement basse (43%) ce qui est une faiblesse importante. D'autre part, il obtient un F1-score à 53%, or il s'agit d'une métrique cruciale qui doit être améliorée.

5. Importance des variables prédictives

a) Les coefficients de feature de la régression logistique

Les coefficients d'un modèle de régression logistique indiquent l'effet de chaque variable sur la probabilité que la cible prenne une certaine valeur, dans notre cas, qu'un employé quitte l'entreprise. Il est possible de comparer l'importance des différentes variables sur la prédiction de la variable cible grâce à la valeur absolue des coefficients de régression logistique (Magnitude). Les features avec des coefficients de plus grande magnitude ont généralement un impact plus important sur la prédiction de la cible. L'interprétation des résultats obtenus permet de formuler des recommandations pratiques sur les politiques de rétention des employés.

Pour ce faire, il faut veiller à éliminer la multicolinéarité entre les variables et à faire la standardisation des données :

** Les résultats de l'importance des features peuvent être biaisés par la multicolinéarité entre certaines variables. Si plusieurs features sont fortement corrélées entre elles, leurs coefficients peuvent être instables et difficiles à interpréter. C'est pourquoi nous avons analysé les corrélations entre les variables avant la mise au point du modèle de Machine Learning et que nous avons utilisé la fonction drop : « first » lors de l'encodage des variables catégorielles avec l'outil « OneHotEncoder ».

** Les features ayant des échelles différentes peuvent avoir des coefficients difficilement comparables. La standardisation des données était donc essentielle. Pour palier à cet obstacle, nous avons fait une standardisation des variables numériques avec l'outil « StandardScaler ».

Après ces étapes, il est possible de comparer les variables numériques entre elles et les variables catégorielles entre elles. Cependant, il n'est pas possible de comparer les variables numériques avec les variables catégorielles.

En effet, les coefficients obtenus pour les variables numériques représentent l'effet d'une variation d'une unité dans la variable sur la probabilité de la classe cible.

Les coefficients des variables catégorielles, quant à eux, n'ont ni la même échelle ni la même unité. Ils permettent d'interpréter l'impact de chaque catégorie (au sein d'une même variable) par rapport à la catégorie de référence, celle qui a été supprimée lors de l'OneHotEncoding. L'importance d'une des catégories montre l'impact de la variable codée de manière générale.

Autrement dit, si une catégorie a une importance très élevée, cela signifiera que la catégorie a un fort impact sur la cible en comparaison de l'impact qu'à la catégorie de référence

(catégorie supprimée lors de l'OneHotEncoding). Sans cette comparaison, l'interprétation perd son sens.

Le plus simple est donc de comparer les variables numériques entre elles et de comparer les catégories de chaque variable catégorielle entre elles.

Toutefois, il est possible d'envisager une comparaison plus générale en calculant l'odds-ratio pour chaque feature du modèle de Machine Learning. Les odds-ratios sont le résultat de l'exponentiation des coefficients de la régression logistique. Ces odds-ratios sont plus faciles à interpréter que les coefficients bruts.

b) Les odds-ratio des features et p-value

En examinant les odds-ratios, nous pouvons comprendre l'importance des features puisqu'ils permettent de quantifier l'impact d'une variable, qu'elle soit numérique ou catégorielle, sur la probabilité de l'événement cible.

Puisque la standardisation des variables numériques a été effectuée, nous pouvons comparer l'influence des variables catégorielles et numériques pour prédire l'attrition chez IBM. Nous n'interpréterons que les odds-ratios dont les variables ont une p-value inférieure à 0,05, c'est-à-dire qui sont statistiquement significatives dans notre modèle de Machine Learning.

c) Résultats

feature_names	Odds_Ratio	p_values
cat__BusinessTravel_Travel_Frequently	22,047405	0,000
cat__BusinessTravel_Travel_Rarely	7,034252	0,000
cat__Department_Research & Development	2,328993	0,773
cat__Department_Sales	2,204404	0,796
cat__EducationField_Life Sciences	0,16976	0,018
cat__EducationField_Marketing	0,320699	0,118
cat__EducationField_Medical	0,231288	0,046
cat__EducationField_Other	0,199451	0,040
cat__EducationField_Technical Degree	0,492606	0,278
cat__EnvironmentSatisfaction_2_Moyen	0,234346	0,000
cat__EnvironmentSatisfaction_3_Elevée	0,192807	0,000
cat__EnvironmentSatisfaction_4_Très élevée	0,092927	0,000
cat__JobInvolvement_2_Moyen	0,232393	0,000
cat__JobInvolvement_3_Elevée	0,184345	0,000

feature_names	Odds_Ratio	p_values
cat__JobInvolvement_4_Très élevée	0,045145	0,000
cat__JobRole_HumanResources_Job	1,654023	0,832
cat__JobRole_Laboratory Technician	3,850782	0,008
cat__JobRole_Manager	0,304373	0,195
cat__JobRole_Manufacturing Director	0,915035	0,848
cat__JobRole_Research Director	0,064088	0,008
cat__JobRole_Research Scientist	1,557245	0,361
cat__JobRole_Sales Executive	7,811752	0,087
cat__JobRole_Sales Representative	6,103887	0,142
cat__JobSatisfaction_2_Moyen	0,308261	0,000
cat__JobSatisfaction_3_Elevée	0,390061	0,000
cat__JobSatisfaction_4_Très élevée	0,179257	0,000
cat__OverTime_OverTime-Yes	13,114347	0,000
cat__WorkLifeBalance_2_Bon	0,258477	0,000
cat__WorkLifeBalance_3_Très bon	0,16617	0,000
cat__WorkLifeBalance_5_Parfait	0,583458	0,170
num__Age	1,023764	0,808
num__DistanceFromHome	1,54325	0,000
num__MonthlyIncome	0,479845	0,039
num__StockOptionLevel	0,466779	0,000
num__TotalWorkingYears	0,446459	0,000
num__TrainingTimesLastYear	0,714013	0,000
num__YearsAtCompany	1,161962	0,505
num__YearsInCurrentRole	0,731635	0,066
num__YearsWithCurrManager	0,898903	0,513

d) Interprétation

* Certaines variables ne sont pas statistiquement significatives (p-value > 0,05), cela suggère qu'il n'y a pas suffisamment de preuves pour affirmer qu'elles sont liées à l'attrition. C'est le cas pour les variables numériques 'Department', 'Age', 'YearsAtCompany', 'YearsInCurrentRole' et 'YearsWithCurrManager'.

* Les variables 'TrainingTimesLastYear', 'MonthlyIncome', 'StockOptionLevel' et 'TotalWorkingYears' diminuent l'attrition à mesure que leur valeur augmente. En effet, elles ont des odds-ratios inférieurs à 1 signifiant que lorsque la valeur de la variable augmente, la probabilité que l'employé quitte l'entreprise diminue. Elles sont, ici, classées dans l'ordre inverse de l'importance de leur impact sur l'attrition, autrement dit, la variable 'TrainingTimesLastYear' a plus d'impact sur l'attrition que 'TotalWorkingYears'.

* Les odds-ratios des variables :

cat_BusinessTravel_Travel_Frequently
cat_OverTime_OverTime-Yes
cat_BusinessTravel_Travel_Rarely
cat_JobRole_Laboratory Technician

... sont très élevés. Ces catégories ont donc une très forte influence sur le départ des salariés. Ici, elles sont classées par ordre décroissant de leur impact relatif sur l'attrition.

* La distance entre le domicile et l'entreprise entraîne le départ des salariés dès lors qu'elle augmente.

* Voici l'ordre décroissant de l'importance des catégories qui diminuent l'attrition :

cat_JobSatisfaction_3_Elevée
cat_JobSatisfaction_2_Moyen
cat_WorkLifeBalance_2_Bon
cat_EnvironmentSatisfaction_2_Moyen
cat_JobInvolvement_2_Moyen
cat_EducationField_Medical
cat_EducationField_Other
cat_EnvironmentSatisfaction_3_Elevée
cat_JobInvolvement_3_Elevée
cat_JobSatisfaction_4_Très élevée
cat_EducationField_Life Sciences
cat_WorkLifeBalance_3_Très bon
cat_EnvironmentSatisfaction_4_Très élevée
cat_JobRole_Research Director
cat_JobInvolvement_4_Très élevée

Pour exemple, un salarié qui a coché 'élevée' pour évaluer la satisfaction qu'il éprouve dans son travail, présentera moins de risque de quitter l'entreprise que celui qui aura coché une autre catégorie.

* Un employé qui fait des déplacements professionnels fréquents a 22 fois plus de risques de quitter l'entreprise qu'un employé qui n'a pas de déplacements professionnels à faire. Pareillement, un employé faisant rarement des déplacements, a 7 fois plus de chance de quitter l'entreprise.

* Concernant le domaine d'étude, les employés ayant un diplôme dans Life Sciences, Other et Medical (dans cet ordre) ont moins de chances de quitter l'entreprise par rapport à ceux ayant un diplôme en Human Resources. Cependant, pour Marketing et Technical Degree, il n'y a pas suffisamment de preuve statistique pour établir un lien avec l'attrition.

* Plus les employés sont satisfaits de leur environnement et de leur travail moins ils risquent de quitter l'entreprise.

- * Plus les salariés sont impliqués moins ils risquent de quitter l'entreprise.
- * Un bon équilibre entre la vie privée et la vie professionnelle implique que les salariés ont moins tendance à partir.
- * Un odds ratio de 13,1 pour OverTime-Yes indique que les employés qui font des heures supplémentaires ont 13,1 fois plus de risques de quitter l'entreprise que ceux qui ne font pas d'heures supplémentaires et donc que cette variable est très influente sur l'attrition.
- * On peut également remarquer que le métier a peu ou pas d'influence sur le départ des salariés à l'exception des techniciens de laboratoire qui ont une attrition plus probable que les autres métiers et des directeurs de recherches qui semblent être plus à même de rester salariés.

6. Conclusion sur le Machine Learning

Nous avons eu du mal à obtenir un modèle aux résultats satisfaisants et il a été difficile d'interpréter l'impact des variables sur l'attrition chez IBM. Cependant, nous obtenons des résultats cohérents avec l'analyse exploratoire des données faite à l'aide du logiciel TABLEAU, les résultats du Machine Learning ont permis d'affiner les conclusions et vont nous permettre de faire des recommandations plus ciblées et plus pertinentes.

7. Perspectives d'amélioration du modèle de prédiction

- * Il serait intéressant d'essayer d'autres techniques de suréchantillonnage pour la gestion du déséquilibre de classes.
- * Essayer d'autres modèles de Machine Learning serait également à envisager.
- * Tenter d'autres approches pour l'amélioration et l'ajustement des hyperparamètres des modèles. Un Random Forest sur ce type de sujet est probablement plus adapté.
- * Améliorer la database semble essentiel, l'enrichir avec les données collectées les années précédentes et ajouter de nouvelles variables pourrait grandement augmenter les probabilités d'obtenir de bonnes prédictions.

* Mieux catégoriser la variable Attrition semble également nécessaire. En effet, l'attrition pouvant être volontaire, involontaire, démographique, (...), il est difficile de conclure sur les facteurs ayant induit la décision d'un salarié alors que parfois elle est imposée au salarié.

* Un travail commun avec le département des Ressources Humaines permettrait certainement d'éliminer d'autres variables ou de mieux en comprendre certaines, l'intérêt étant d'alléger le modèle pour éviter l'overfitting.

XV. Conclusion de l'étude



L'individu le plus représenté chez IBM est un homme marié qui a étudié dans le domaine de 'Life-Science' et travaille au département Recherche & Développement ou au service commercial. Il fait de rares déplacements professionnels et ne fait pas d'heures supplémentaires.

Nous avons cherché à expliquer l'attrition dans l'entreprise afin de construire un modèle de Machine Learning capable de prédire le départ d'un salarié.

Dans notre cas, les types d'attrition d'intérêt sont l'attrition volontaire et l'attrition démographique. En effet, ce sont les types d'attrition où il s'agit d'une décision prise par le salarié, or l'objectif de notre modèle est d'anticiper le départ des talents de l'entreprise.

Cependant, nos données sur la cible regroupaient l'attrition volontaire, l'attrition démographique, l'attrition involontaire, les retraites, l'attrition interne, du moins c'est l'interprétation que nous en avons fait, par manque d'informations sur l'origine des données. Cette méconnaissance induit de potentielles erreurs d'interprétation et démontre qu'il est essentiel que l'analyste de données échange avec le personnel de l'entreprise afin de bien comprendre le contexte de l'étude.



L'étude a montré que les déplacements professionnels étaient le facteur prépondérant de l'augmentation de l'attrition.

Les heures supplémentaires ont également un rôle très impactant sur le départ des salariés.

Nous avons également pu mettre en évidence qu'être technicien de laboratoire chez IBM augmente la probabilité de partir de l'entreprise.

Enfin, plus la distance entre le lieu de vie et le lieu de travail est grande plus l'attrition est importante.

A contrario, certains facteurs semblent influer sur le fait qu'un salarié reste en poste dans l'entreprise.

En effet, le nombre de formations suivies l'année précédente semble avoir un impact positif sur la rétention du personnel puisque l'attrition diminue lorsqu'il augmente.

L'augmentation du salaire médian mensuel et du niveau de stock options entraîne également une diminution de l'attrition.

Nous avons également constaté que les salariés ayant le plus d'années d'expérience professionnelle sont moins sujets à quitter leur travail chez IBM.

D'autres facteurs, avec une moins grande influence, favorise la rétention du personnel : la satisfaction de l'environnement de travail et celle du métier, un bon équilibre entre la vie privée et la vie professionnelle et l'implication du salarié dans son travail.

Les salariés dont les domaines d'étude sont 'Medical' et 'LifeScience' semblent moins enclins à quitter l'entreprise tout comme les directeurs de recherche qui ont une attrition moins élevée que les autres métiers présents chez IBM.



Grâce à cette étude, il est possible de dégager des insights utiles pour la gestion des ressources humaines dans l'entreprise.

XVI. Recommandations

Les recommandations sont listées par ordre d'importance décroissante, dans l'ordre où il est préférable de mener les actions.

1. Accès aux formations

Nous avons constaté que la majorité des employés attritionnistes sont des débutants ou des juniors et que l'augmentation du nombre de formations suivies l'année précédente avait une forte influence sur la rétention du personnel.

Il peut être impactant de proposer des formations aux salariés, en effet, l'opportunité d'apprentissage permet d'envisager un avancement professionnel dans l'entreprise et d'améliorer le parcours de carrière des employés.

2. Politique de rémunération

Il est essentiel de surveiller le système de rémunération pour garantir l'équité et la cohérence.

Même si l'augmentation des salaires ne réduit pas toujours l'attrition, il faut être certain que la grille de rémunération soit juste. Nous avons perçu quelques incohérences. Un ajustement des salaires pourrait diminuer le sentiment d'injustice, de manque de reconnaissance et de manque d'appréciation du travail effectué.

Le fait que les attritionnistes perçoivent un salaire médian très inférieur à celui des salariés restés nous indique tout de même que des propositions d'avantages financiers pourraient être bénéfiques sur la rétention du personnel dans le cas d'IBM. Cela peut se présenter sous la forme d'avantages sociaux, de jours de congés payés, d'un bon régime de retraite, d'une assurance santé de qualité, de bonus, ou même de prime sur la productivité (puissant facteur de motivation).

Le niveau de stock-options a également un effet néfaste sur la rétention du personnel lorsqu'il est bas ou inexistant. Certains contrats sont proposés avec de hauts niveaux de stock-options dès la signature tandis que d'autres salariés n'en ont pas. Une répartition plus juste ou un accès mieux mesuré à cet avantage pourraient impliquer davantage les salariés et diminuer l'attrition.

3. Contraintes professionnelles

Nous avons mis en évidence que les déplacements professionnels et les heures supplémentaires avaient un impact fort sur l'attrition.

Envisager un avantage financier plus conséquent pour les employés contraints par ces deux facteurs peut apporter une solution mais n'empêchera pas l'épuisement professionnel. En effet, ces deux contraintes professionnelles sont des causes de déséquilibre de la vie personnelle et vie professionnelle et peut diminuer la satisfaction des salariés. Les limiter ou mieux les répartir sur l'ensemble des salariés pourraient diminuer leur impact.

La distance pour se rendre au travail semble favoriser l'attrition lorsqu'elle augmente. Pour réduire son effet, il est possible d'envisager une prime au kilométrage. De plus en plus d'entreprises proposent désormais des possibilités de télétravail. Cela permet de réduire les temps et les coûts de trajet pour le salarié.

4. Embauche

Le domaine d'étude est apparu comme un facteur diminuant l'attrition dans le cas des salariés ayant étudié dans le médical et les sciences de la vie. Il est possible qu'il y ait une inadéquation entre le métier et les compétences des salariés ayant étudié dans d'autres domaines, il peut être important de mieux cibler les candidats lors du recrutement et de bien présenter le travail proposé.

Cette recommandation peut également être appliquée pour le facteur « années d'expérience professionnelle ». En effet, plus les salariés ont d'expérience professionnelle moins ils sont susceptibles de quitter l'entreprise. Cibler à l'embauche des candidats avec une expérience professionnelle plus grande permettrait peut-être de stabiliser ou même réduire l'attrition. Il serait tout de même intéressant de comprendre pourquoi ce facteur est important.

5. Feedback

S'il est anonyme, le feedback peut largement contribuer à identifier les raisons de départ volontaire du personnel, et permet ainsi de prendre des mesures pour y remédier.

L'anonymat est difficile à garantir pour les métiers dont l'effectif est faible, nous imposant de prendre le risque que certaines données remontées par les salariés soient mensongères (les salariés souhaitant préserver leur anonymat ou ayant peur de perdre leur travail).

D'autre part, certaines informations démographiques, telles que la situation familiale, le nombre d'enfants, ...), peuvent être cachées, prétextant que l'information peut être utilisée à des fins discriminatoires.

Cependant, mettre en place un questionnaire de satisfaction en lien avec les informations des salariés permettrait de déceler les signes de mécontentement et de mieux comprendre les raisons qui poussent certains employés à quitter l'entreprise.

Créer un nouveau questionnaire serait l'occasion de catégoriser l'attrition en plusieurs classes :

- Volontaire : démission ou rupture conventionnelle,
- Involontaire : les différents types de licenciements, inaptitude,
- Retraite.

Cela permettrait également d'augmenter les informations personnelles sur les salariés et ainsi s'assurer qu'aucune raison démographique n'entraîne la hausse de l'attrition.

Connaitre l'avis sur le management semble pouvoir nous éclairer sur les raisons expliquant pourquoi la première année d'un changement de manager l'attrition est forte.

Être en mesure de préciser la raison d'un départ volontaire en posant la question directement aux salariés attritionnistes avant leur départ définitif serait un atout important pour l'étude des données.

Demander aux salariés quelles sont leurs attentes en terme de plan de carrière afin de nous assurer qu'ils sont satisfaits sur ce point crucial et poser des questions plus précises sur la satisfaction au travail enrichiraient le jeu de données et aboutiraient à des recommandations plus éclairées.

Le feedback doit être un outil de communication entraînant le dialogue ouvert et à double sens pour devenir un outil d'amélioration continue.

6. Autres idées

Il serait intéressant de demander les raisons qui poussent les techniciens de laboratoire à quitter l'entreprise afin de palier aux problèmes qui pourraient être remontés.

Assurer une intégration efficace pourrait limiter le départ des nouveaux arrivants.

Enfin, nous fournir les données des années précédentes permettrait d'améliorer notre modèle et de mieux cibler les recommandations pour limiter l'attrition.

