Comprendre les déterminants du turnover des salariés

Analyse exploratoire basée sur des données RH – Projet People Analytics

1. Objectif du projet

Ce projet a pour but d'explorer les facteurs associés au départ volontaire des salariés (turnover) à partir d'un jeu de données RH issu d'un contexte fictif fourni par IBM.
Notre analyse vise à mettre en lumière les variables clés influençant ce phénomène, dans une optique de prévention et d'aide à la décision pour les directions RH.

2. Données utilisées

Le jeu de données comporte 1 470 observations et plus de 30 variables couvrant plusieurs dimensions:

- **Informations démographiques :** âge, sexe, état civil
- **Parcours professionnel :** poste, ancienneté, département, niveau d'études
- **Facteurs financiers :** salaire mensuel, bonus, stock options
- **Engagement et satisfaction :** satisfaction au travail, équilibre vie pro/perso
- **Comportements:** heures supplémentaires, formations suivies, promotions

La variable cible est **Départ**, indiquant si l'employé a quitté l'entreprise (Yes / No).

Remarque importante: les noms de variables ont été traduits en français pour renforcer la lisibilité et faciliter la compréhension pour un public non anglophone. Cela permet également de

présenter les résultats dans un format plus professionnel adapté aux recruteurs francophones.

```python

## IMPORTATION ET APERCU
```

```python

### Importation des Bibliothèques utiles
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

### Pour affichage clair
sns.set(style="whitegrid")
plt.rcParams['figure.figsize'] = (6, 4)

```

```python

### Chargement du jeu de données
df = pd.readcsv("WAFn-UseC_-HR-Employee-Attrition.csv")

### Afficher les 5 premières lignes
df.head()

```

```python
```

Aperçu des dimensions & types

```

```python

## Nombre de lignes et colonnes

df.shape

```

(1470, 35)

```python

## Types de données

df.dtypes

```

Age int64
Attrition object
BusinessTravel object
DailyRate int64
Department object
DistanceFromHome int64
Education int64
EducationField object
EmployeeCount int64
EmployeeNumber int64
EnvironmentSatisfaction int64
Gender object
HourlyRate int64
JobInvolvement int64
```

JobLevel int64
JobRole object
JobSatisfaction int64
MaritalStatus object
MonthlyIncome int64
MonthlyRate int64
NumCompaniesWorked int64
Over18 object
OverTime object
PercentSalaryHike int64
PerformanceRating int64
RelationshipSatisfaction int64
StandardHours int64
StockOptionLevel int64
TotalWorkingYears int64
TrainingTimesLastYear int64
WorkLifeBalance int64
YearsAtCompany int64
YearsInCurrentRole int64
YearsSinceLastPromotion int64
YearsWithCurrManager int64
dtype: object

```python
```

Quelques statistiques descriptives générales

```

```python

## Statistiques descriptives

df.describe()

```

```python

## Statistiques pour les colonnes catégorielles

df.describe(include='object')

```

```python

## Analyse de la répartition de la variable cible 'Départ'

```

```python

## Distribution de la variable cible 'Départ'

**df['Attrition'].value_counts()**

```

Attrition
No 1233
Yes 237
Name: count, dtype: int64

```python

## En pourcentage

df['Attrition'].value_counts(normalize=True) * 100
```

Attrition
No 83.877551
Yes 16.122449
Name: proportion, dtype: float64

```python

## Renommage en français de la variable cible et de 14 variables explicatives pertinentes

```

```python
df = df.rename(columns={
    'Attrition': 'Départ', # Variable cible
    'BusinessTravel': 'Déplacements professionnels',
    'Age': 'Âge',
    'Department': 'Département',
    'MonthlyIncome': 'Salaire mensuel',
    'EducationField': 'Domaine d’études',
    'TotalWorkingYears': 'Années de carrière',
    'Gender': 'Sexe',
    'YearsAtCompany': 'Ancienneté dans l’entreprise',
    'JobRole': 'Poste',
    'JobSatisfaction': 'Satisfaction au travail',
    'MaritalStatus': 'État civil',
    'EnvironmentSatisfaction': "Satisfaction environnement",
    'OverTime': 'Heures supplémentaires',
    'WorkLifeBalance': 'Équilibre vie pro/perso'
})

```

```python

## Affichage des variables mises à jour

df.columns

```

```
Index(['Âge', 'Départ', 'Déplacements professionnels', 'DailyRate',
       'Département', 'Distance domicile-travail', 'Niveau d’études',
       'Domaine d’études', 'EmployeeCount', 'EmployeeNumber',
       'Satisfaction environnement', 'Sexe', 'HourlyRate', 'JobInvolvement',
       'JobLevel', 'Poste', 'Satisfaction au travail', 'État civil',
       'Salaire mensuel', 'MonthlyRate', 'NumCompaniesWorked', 'Over18',
       'Heures supplémentaires', 'PercentSalaryHike', 'PerformanceRating',
       'RelationshipSatisfaction', 'StandardHours', 'StockOptionLevel',
       'Années de carrière', 'TrainingTimesLastYear',
       'Équilibre vie pro/perso', 'Ancienneté dans l’entreprise',
       'YearsInCurrentRole', 'YearsSinceLastPromotion',
       'YearsWithCurrManager'],
```
```

dtype='object')

```python
## Affichage de quelques variables utiles
colonnesutiles = ['Départ', 'Âge', 'Sexe', 'Département', 'Salaire mensuel', 'Années de carrière']
df[colonnesutiles].head()
```

```python
## Exploration de la variable cible 'Départ'
```

```python
## Répartition en nombres (valeurs entières)
df['Départ'].value_counts()
```

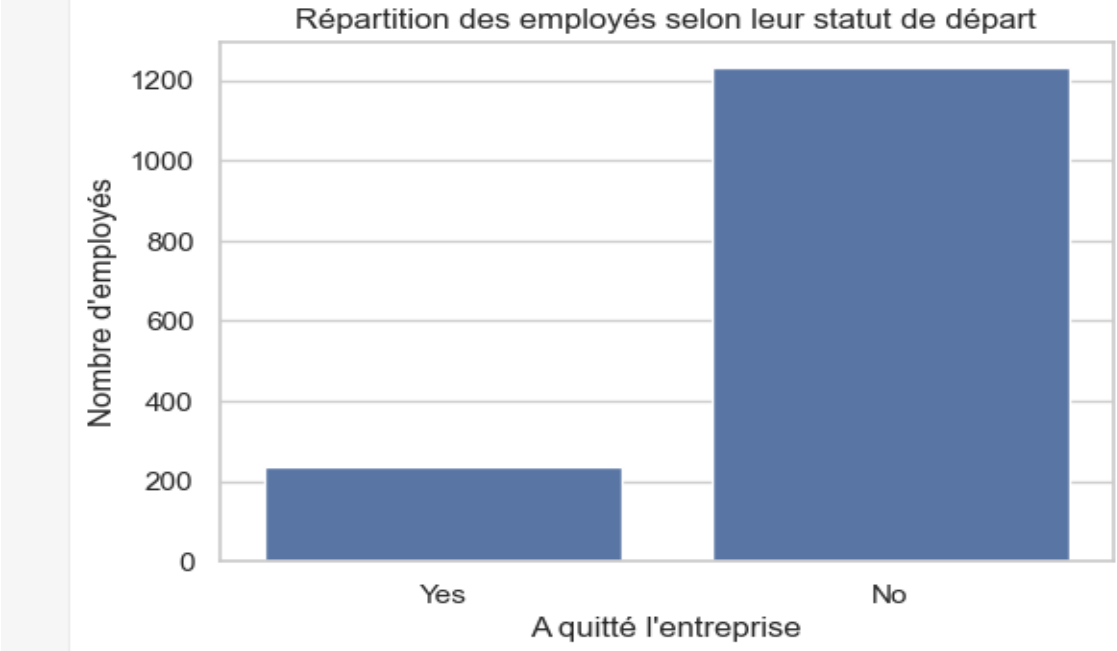
Départ
No 1233
Yes 237
Name: count, dtype: int64

```python
## Répartition en pourcentage
df['Départ'].value_counts(normalize=True) * 100
```

Départ
No 83.877551
Yes 16.122449
Name: proportion, dtype: float64

```python

## Visualisation de la répartition
sns.countplot(x='Départ', data=df)
plt.title("Répartition des employés selon leur statut de départ")
plt.ylabel("Nombre d'employés")
plt.xlabel("A quitté l'entreprise")
plt.show()
```

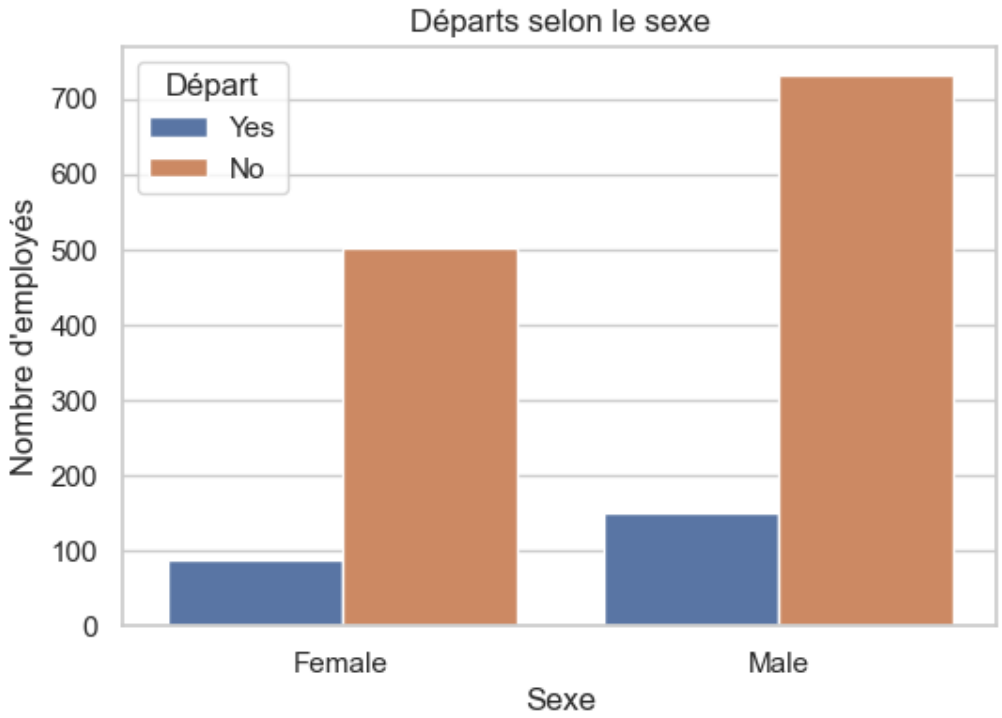


La majorité des employés (environ 84 %) sont restés dans l'entreprise, tandis que 16 % ont quitté leur poste, indiquant un taux d'attrition modéré nécessitant une attention RH ciblée.
```

```python

## Découverte du profil des employés qui quittent l'entreprise
```

```python

## Départ selon le sexe
```

```python

## Tableau croisé
pd.crosstab(df['Sexe'], df['Départ'], normalize='index') * 100
```

```python

## Visualisation
sns.countplot(x='Sexe', hue='Départ', data=df)
plt.title("Départs selon le sexe")
plt.ylabel("Nombre d'employés")
plt.xlabel("Sexe")
plt.show()
```
```



Les hommes présentent un nombre de départs légèrement plus élevé que les femmes,ce qui peut refléter des différences d'attentes, de conditions ou de perception du travail selon le genre.

```

```python

## Départ selon les tranches d'âge

```

```python

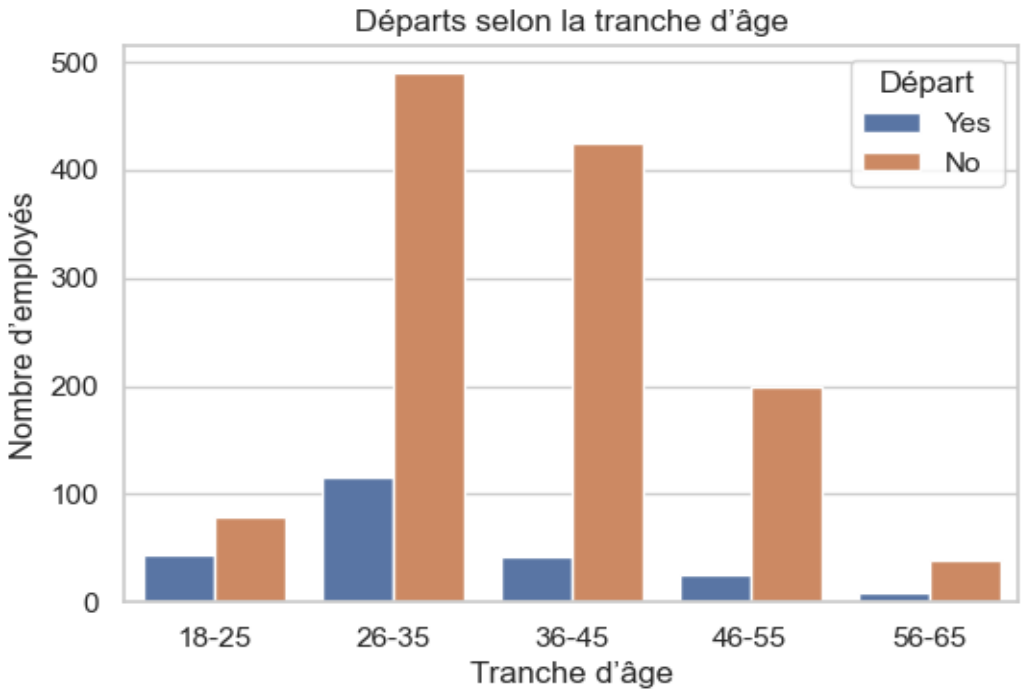
## Créer des tranches d'âge

df['Tranche d'âge'] = pd.cut(df['Âge'], bins=[17, 25, 35, 45, 55, 65],
                             labels=['18-25', '26-35', '36-45', '46-55', '56-65'])

## Taux de départ par tranche d'âge

pd.crosstab(df['Tranche d'âge'], df['Départ'], normalize='index') * 100
```
```

```python

## Visualisation

sns.countplot(x='Tranche d’âge', hue='Départ', data=df)
plt.title("Départs selon la tranche d’âge")
plt.xlabel("Tranche d’âge")
plt.ylabel("Nombre d’employés")
plt.show()
```



Les départs sont particulièrement élevés dans les tranches d’âge 26–35 ans et 18–25 ans, ce qui suggère une volatilité accrue en début de carrière et un enjeu de rétention des jeunes talents.

```python
```

Départ selon le salaire mensuel

```
```

```python
```

Boxplot pour comparer les salaires

```
sns.boxplot(x='Départ', y='Salaire mensuel', data=df)
plt.title("Distribution des salaires selon le statut de départ")
plt.show()
```

Les employés ayant quitté l'entreprise affichent une rémunération médiane plus faible, suggérant une possible influence du niveau de salaire sur la décision de départ.

```
```

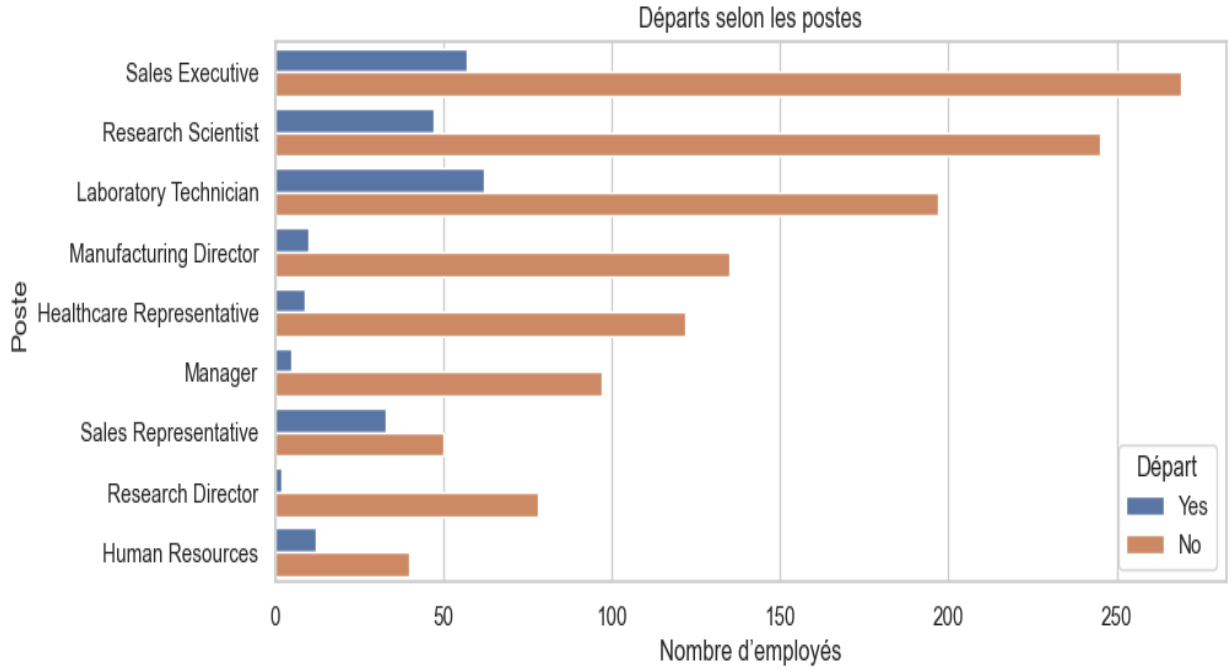
```python
```

Départ selon le poste occupé

```

```python

## Taux de départ par poste

pd.crosstab(df['Poste'], df['Départ'], normalize='index') * 100
```

```python

## Graphique

plt.figure(figsize=(12,6))
sns.countplot(y='Poste', hue='Départ', data=df)
plt.title("Départs selon les postes")
plt.xlabel("Nombre d'employés")
plt.ylabel("Poste")
plt.show()
```


```

Les départs sont plus fréquents dans les postes opérationnels comme Laboratory Technician et Sales Representative, suggérant des zones de tension potentielles en lien avec les conditions ou perspectives d'évolution.
```

```python

## NETTOYAGE DES DONNEES
```

```python

### Détection des valeurs manquantes
```

```python

### Nombre total de valeurs manquantes par colonne
df.isnull().sum()
```

Âge 0
Départ 0
Déplacements professionnels 0
DailyRate 0
Département 0
Distance domicile-travail 0
Niveau d'études 0
Domaine d'études 0
EmployeeCount 0
EmployeeNumber 0
Satisfaction environnement 0
```

```
Sexe                        0
HourlyRate                  0
JobInvolvement              0
JobLevel                    0
Poste                       0
Satisfaction au travail     0
État civil                  0
Salaire mensuel             0
MonthlyRate                 0
NumCompaniesWorked          0
Over18                      0
Heures supplémentaires      0
PercentSalaryHike           0
PerformanceRating           0
RelationshipSatisfaction    0
StandardHours               0
StockOptionLevel            0
Années de carrière          0
TrainingTimesLastYear       0
Équilibre vie pro/perso     0
Ancienneté dans l'entreprise  0
YearsInCurrentRole          0
YearsSinceLastPromotion     0
YearsWithCurrManager        0
Tranche d'âge               0
dtype: int64
```

```python
```

Recherche des doublons

```

```python

## Vérifier les doublons

df.duplicated().sum()
```

np.int64(0)

```python

## Vérification de la cohérence des variables constantes

```

python
df.nunique()

Âge 43
Départ 2
Déplacements professionnels 3
DailyRate 886
Département 3
Distance domicile-travail 29
Niveau d'études 5
Domaine d'études 6
EmployeeNumber 1470
Satisfaction environnement 4
Sexe 2
HourlyRate 71
JobInvolvement 4
```

JobLevel 5
Poste 9
Satisfaction au travail 4
État civil 3
Salaire mensuel 1349
MonthlyRate 1427
NumCompaniesWorked 10
Heures supplémentaires 2
PercentSalaryHike 15
PerformanceRating 2
RelationshipSatisfaction 4
StockOptionLevel 4
Années de carrière 40
TrainingTimesLastYear 7
Équilibre vie pro/perso 4
Ancienneté dans l'entreprise 37
YearsInCurrentRole 19
YearsSinceLastPromotion 16
YearsWithCurrManager 18
Tranche d'âge 5
dtype: int64

```python
```

Suppression des variables constantes

```
df = df.drop(['EmployeeCount', 'Over18', 'StandardHours'], axis=1)
```

Ces colonnes ont été supprimées car elles ne fournissent aucune information pour l'analyse.

EmployeeCount 1 Constante inutile
Over18 1 Tous ont plus de 18 ans
StandardHours 1 Tous travaillent 80h (?)
```

```python

## ANALYSE EXPLORATOIRE DES DONNEES(EDA)
```

```python

## Hypothèse 1: Les employés les moins satisfaits quittent davantage l'entreprise
```

```python

## Barres empilées : pourcentage de départs selon le niveau de satisfaction

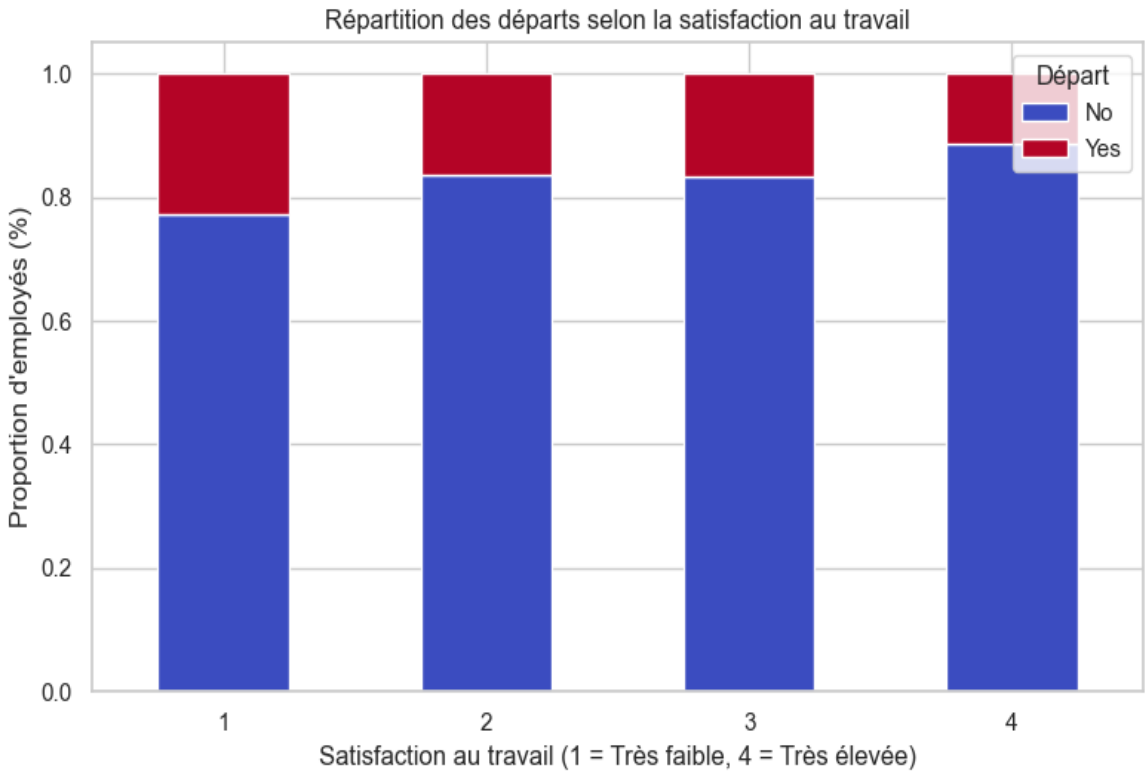
```
table_croisee = pd.crosstab(df['Satisfaction au travail'], df['Départ'], normalize='index')
```

```python
```

Affichage du graphique

```
table_croisee.plot(kind='bar', stacked=True, colormap='coolwarm',
figsize=(8, 5))

plt.title("Répartition des départs selon la satisfaction au travail")
plt.xlabel("Satisfaction au travail (1 = Très faible, 4 = Très élevée)")
plt.ylabel("Proportion d'employés (%)")
plt.legend(title="Départ", loc="upper right")
plt.xticks(rotation=0)
plt.tight_layout()
plt.show()
```



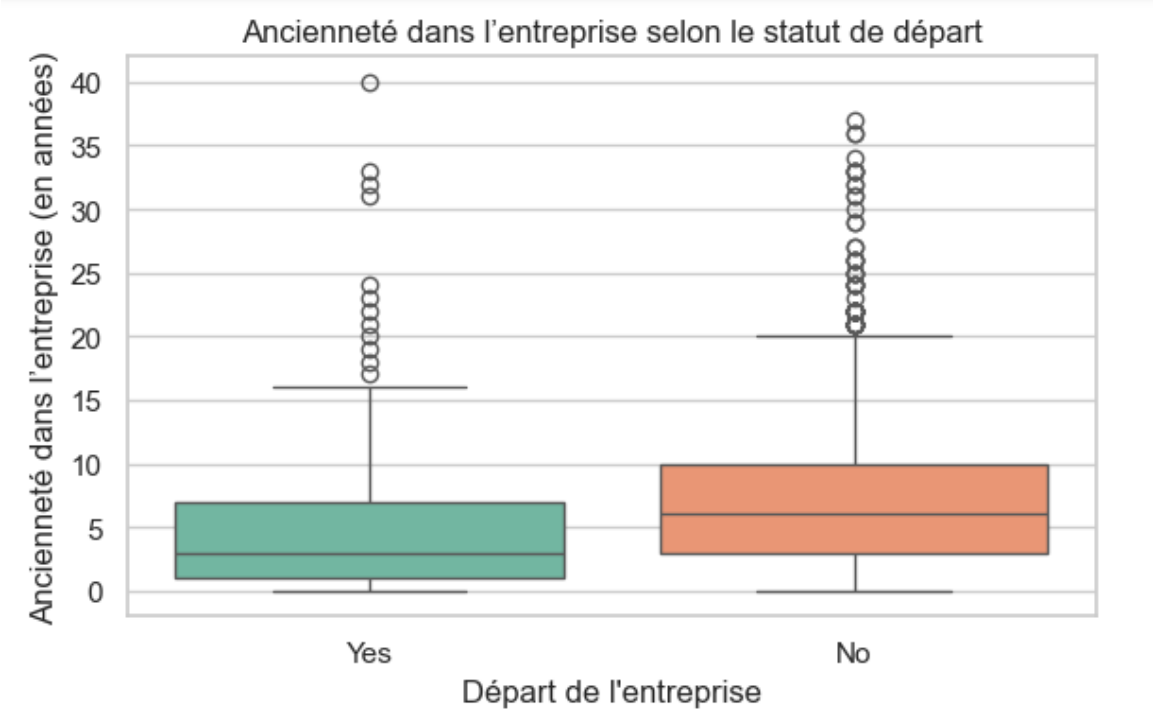
On observe que la proportion de départs diminue à mesure que la satisfaction au travail augmente. Le taux de départ est nettement plus élevé parmi les employés ayant une satisfaction très faible (niveau 1), suggérant donc une corrélation négative entre satisfaction et attrition.

Cette tendance valide l'hypothèse selon laquelle l'insatisfaction constitue un facteur de risque important de départ.

```

```python

## Hypothèse 2 : Les employés les plus récents partent davantage

```

```python

### Graphique : distribution de l'ancienneté selon le statut de départ

```
sns.boxplot(x='Départ', y="Ancienneté dans l'entreprise", hue='Départ', data=df, palette='Set2', legend=False)

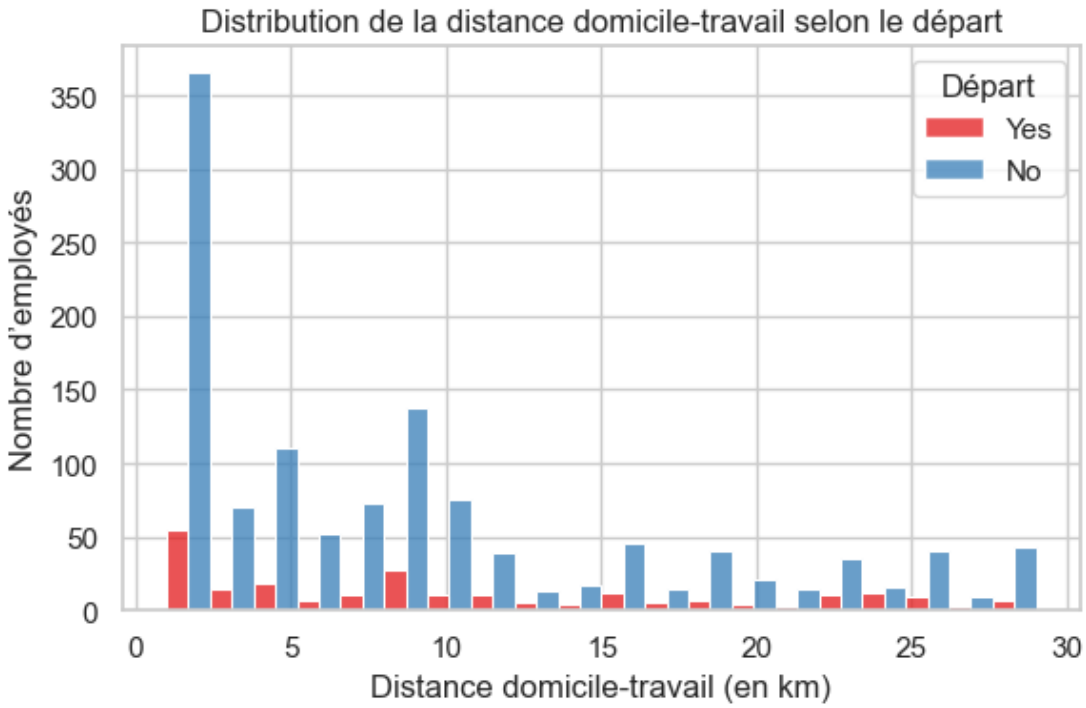
plt.title("Ancienneté dans l'entreprise selon le statut de départ")
plt.xlabel("Départ de l'entreprise")
plt.ylabel("Ancienneté dans l'entreprise (en années)")
plt.tight_layout()
plt.show()
```



Le graphique montre que les employés ayant quitté l'entreprise ("Yes") ont une ancienneté médiane significativement plus faible que ceux restés ("No"). La majorité des départs concerne des collaborateurs ayant moins de 5 ans d'ancienneté.

Cette tendance confirme notre hypothèse: les salariés les plus récents sont plus susceptibles de quitter l'organisation.

```

```python

## Hypothèse 3: Plus la distance est grande, plus les salariés quittent l'entreprise

```

```python

## Histogramme groupé : distribution de la distance selon le statut de départ

```
sns.histplot(data=df, x='Distance domicile-travail', hue='Départ',
multiple='dodge', kde=False, palette='Set1', bins=20)

plt.title("Distribution de la distance domicile-travail selon le départ")
plt.xlabel("Distance domicile-travail (en km)")
plt.ylabel("Nombre d'employés")

plt.legend(title="Départ") ← supprime cette ligne

plt.tight_layout()
plt.show()
```



Le graphique révèle une concentration importante de départs parmi les salariés habitant à proximité immédiate (0–5 km). Toutefois, on observe également une proportion non négligeable de départs jusqu'à environ 25 km.

Contrairement à l'hypothèse initiale, l'attrition ne semble pas croître de manière linéaire avec la distance.

```

```python

## Hypothèse 4: Les salariés faisant des heures supplémentaires quittent davantage l'entreprise.

```

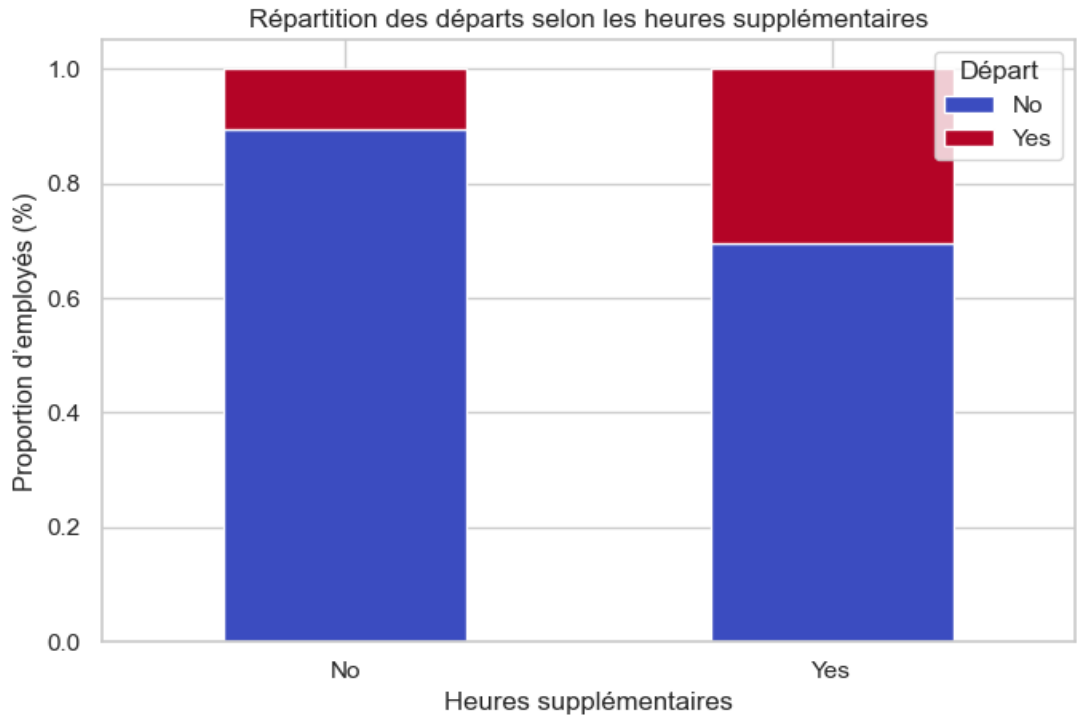
Tableau croisé : proportion de départs selon les heures supplémentaires

```
tableheuresupp = pd.crosstab(df['Heures supplémentaires'], df['Départ'], normalize='index')
```
```

Graph ique à barres empilées

```
tableheuresupp.plot(kind='bar', stacked=True, colormap='coolwarm',
figsize=(7, 5))

plt.title("Répartition des départs selon les heures supplémentaires")
plt.xlabel("Heures supplémentaires")
plt.ylabel("Proportion d'employés (%)")
plt.legend(title="Départ", loc="upper right")
plt.xticks(rotation=0)
plt.tight_layout()
plt.show()
```



Le graphique met en évidence une proportion nettement plus élevée de départs chez les employés effectuant des heures supplémentaires. Alors que les salariés ne faisant pas d'heures sup présentent un faible taux de départ, ceux soumis à une surcharge horaire semblent significativement plus enclins à quitter l'entreprise.

```
```

```python

## Hypothèse 5: L'entreprise affecte davantage les heures supplémentaires aux salariés récents, ce qui pourrait contribuer à leur départ.

```

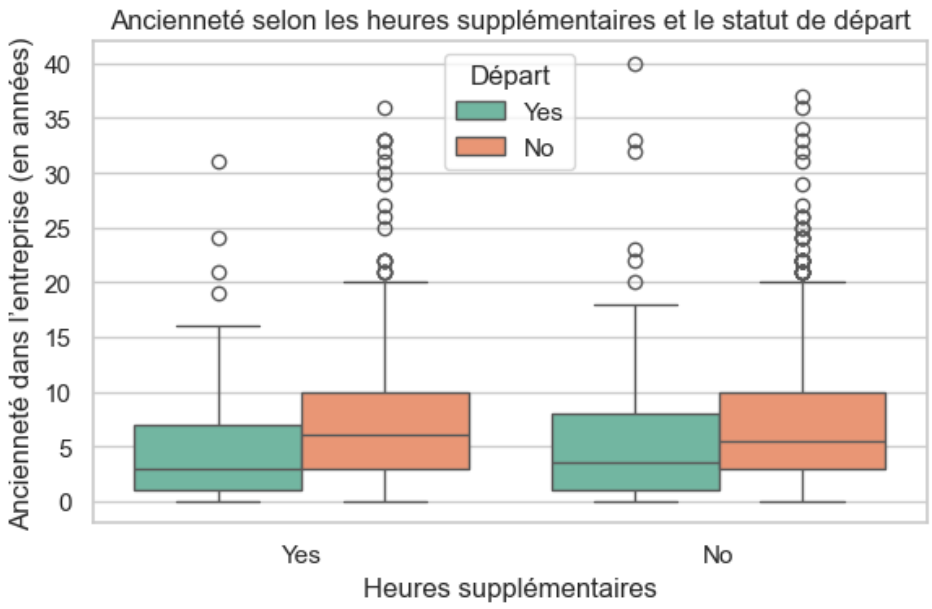
```python

### Comparer l'ancienneté en fonction des heures sup, et en séparant selon le départ

sns.boxplot(x='Heures supplémentaires', y="Ancienneté dans l'entreprise", hue='Départ', data=df, palette='Set2')

plt.title("Ancienneté selon les heures supplémentaires et le statut de départ")
plt.xlabel("Heures supplémentaires")
plt.ylabel("Ancienneté dans l'entreprise (en années)")
plt.legend(title="Départ")
plt.tight_layout()
plt.show()

```



les employés ayant quitté l'entreprise et déclarant effectuer des heures supplémentaires ont en moyenne une ancienneté plus faible que les autres groupes. Les médians d'ancienneté les plus bas concernent justement les salariés partants soumis aux heures supplémentaires,ce qui suggère un effet cumulatif : surcharge + faible expérience = départ accéléré.
```

```python

## Proportion d'heures sup par tranche d'ancienneté

### Création de tranches d'ancienneté

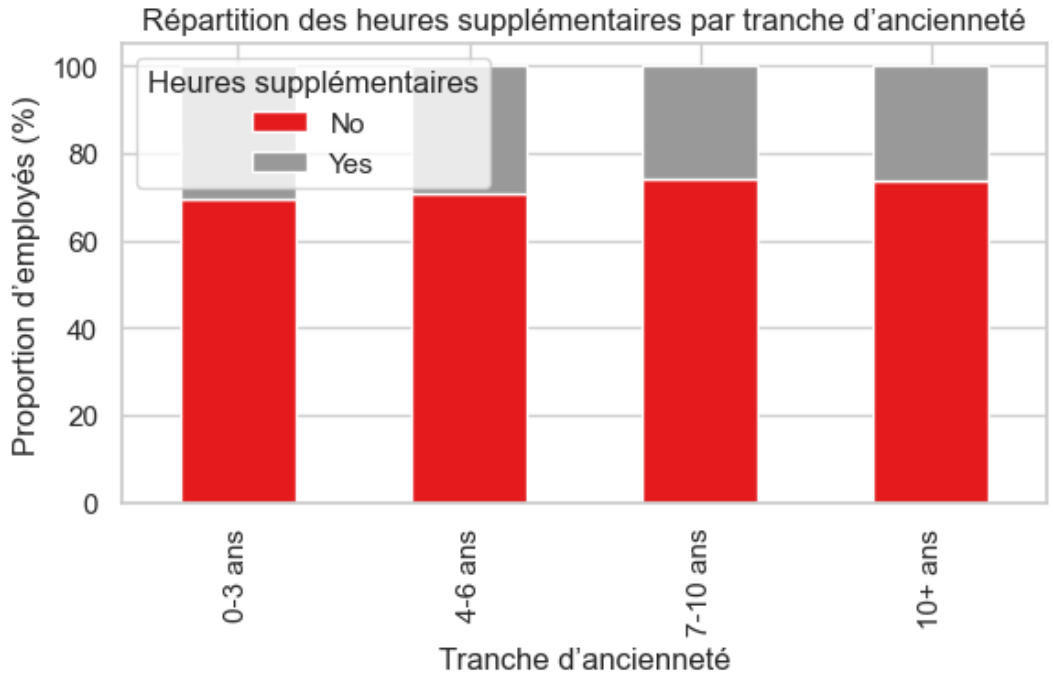
df['Tranche ancienneté'] = pd.cut(df["Ancienneté dans l'entreprise"],
bins=[0, 3, 6, 10, df["Ancienneté dans l'entreprise"].max()],
labels=['0-3 ans', '4-6 ans', '7-10 ans', '10+ ans'])

### Tableau croisé Heures sup / Tranches

table = pd.crosstab(df['Tranche ancienneté'], df['Heures supplémentaires'], normalize='index') * 100

### Affichage

table.plot(kind='bar', stacked=True, colormap='Set1')
plt.title("Répartition des heures supplémentaires par tranche d'ancienneté")
plt.xlabel("Tranche d'ancienneté")
plt.ylabel("Proportion d'employés (%)")
plt.legend(title="Heures supplémentaires")
plt.tight_layout()
plt.show()
```
```



Ce graphique révèle que la répartition des heures supplémentaires est relativement stable selon l'ancienneté, avec environ 25 à 30 % d'employés concernés dans chaque tranche.
Il n'y a pas de concentration exclusive sur les plus récents, mais la combinaison "récents + exposés aux heures sup" reste à surveiller, car ils sont plus vulnérables au départ.
```

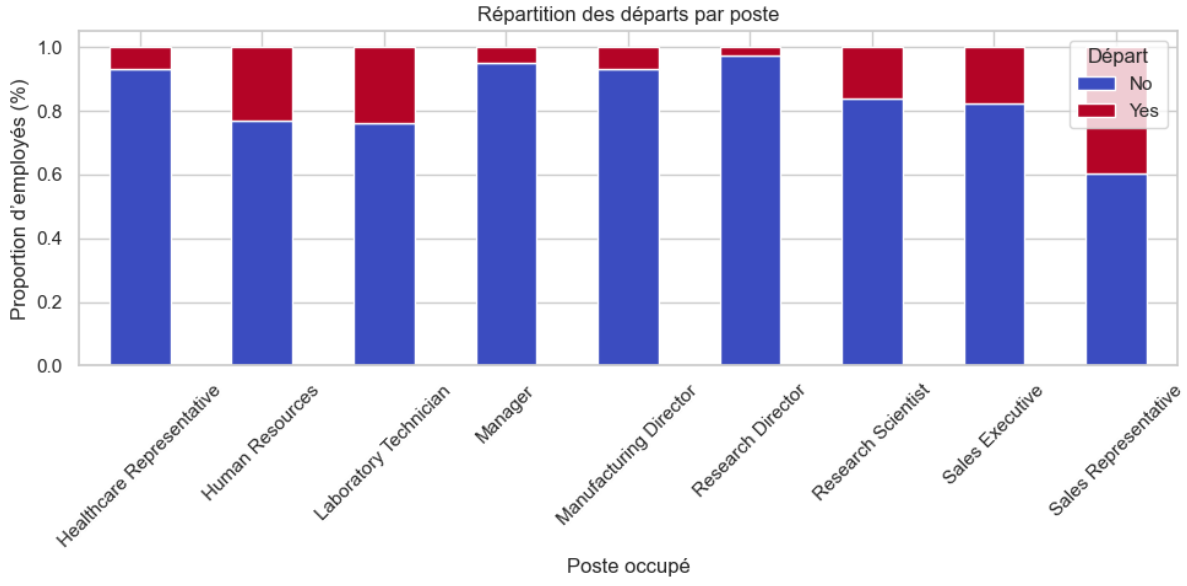
**L'hypothèse 5 n'est pas vérifiée dans sa forme littérale, cependant elle soulève un vrai enjeu RH : les salariés récents sont plus sensibles à la charge horaire, et plus exposés au départ quand elle est présente.**
```

```python
```

Hypothèse 6: Certains postes présentent des taux de départ plus élevés que d'autres, ce qui suggère des zones à risque métier au sein de l'entreprise.

```

```python

### Tableau croisé des départs par poste

table_poste = pd.crosstab(df['Poste'], df['Départ'], normalize='index')

### Graphique à barres empilées

tableposte.plot(kind='bar', stacked=True, colormap='coolwarm',
figsize=(10,5))
plt.title("Répartition des départs par poste")
plt.xlabel("Poste occupé")
plt.ylabel("Proportion d'employés (%)")
plt.legend(title="Départ")
plt.xticks(rotation=45)
plt.tightlayout()
plt.show()

```

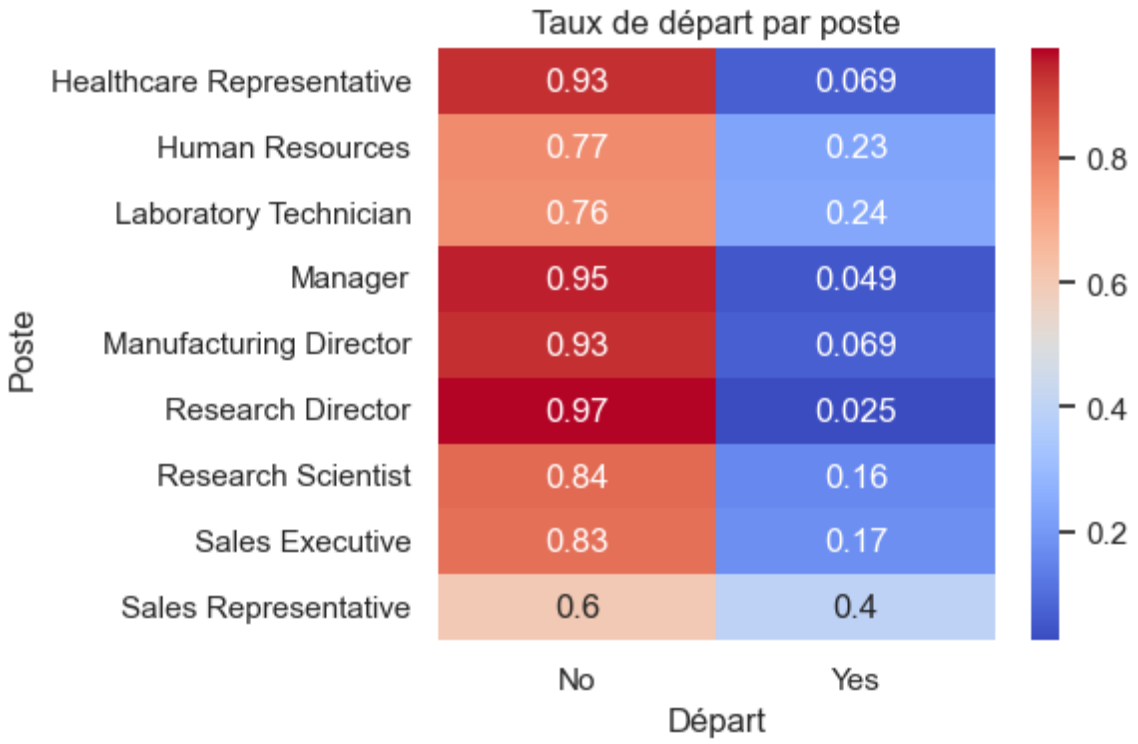

```

Le graphique montre une répartition hétérogène du taux de départ selon les postes. Les postes de Sales Representative, Human Resources et Laboratory Technician présentent les proportions les plus élevées de départs. À l'inverse, les fonctions de Manager, Research Director et Manufacturing Director affichent un taux de fidélisation plus élevé.
```

```python

## Découvrir dans quels postes les heures sup. coïncident avec un haut taux de départ.

### Tableau croisé à 2 dimensions
heatmap_data = pd.crosstab(df['Poste'], df['Départ'], normalize='index')

### Affichage de la heatmap
sns.heatmap(heatmapdata, annot=True, cmap='coolwarm')
plt.title("Taux de départ par poste")
plt.ylabel("Poste")
plt.xlabel("Départ")
plt.tightlayout()
plt.show()
```
```



La heatmap confirme visuellement les écarts de taux de départ entre les fonctions. Le poste de Sales Representative affiche le taux de départ le plus élevé : 40 % des salariés quittent l'entreprise. Les postes de Laboratory Technician (24 %) et Human Resources (23 %) sont également au-dessus de la moyenne. En revanche, les postes de Manager (4,9 %) et Research Director (2,5 %) sont très stables.

```

## SYNTHESE GENERALE

Le départ n'est pas uniforme: il touche prioritairement les salariés récents, exposés à la surcharge, et occupant certaines fonctions sensibles. Une politique RH efficace doit combiner détection précoce, équité dans les conditions de travail et accompagnement différencié selon le poste et l'ancienneté.
```