1 Introduction

Ce rapport présente les résultats d'une analyse de données exploratoire et prédictive menée sur un jeu de données d'entreprise, le fichier `Salaries.csv`. L'objectif principal de cette étude est de comprendre les facteurs qui influencent l'accès au statut de cadre au sein de cette organisation.

L'analyse est structurée en quatre étapes principales : une étude descriptive du jeu de données, l'évaluation des dépendances linéaires entre les variables, la modélisation du statut de cadre par régression logistique, et enfin, une application de prédiction sur un profil d'employé donné. Ce rapport est rédigé dans un style pédagogique, visant à rendre les concepts de l'analyse de données accessibles à un public non expert.

2 Description du Jeu de Données (Question 1 – Étude Descriptive)

2.1 Présentation des Variables

Le jeu de données contient des informations sur **474 employés** et est caractérisé par **six variables** :

Table 1 – Description des variables du jeu de données

Variable	Description	Type	Codage
Salaire	Salaire annuel en dollars	Quantitative continue	-
Education	Nombre d'années d'étude	Quantitative	-
Anciennete	Ancienneté dans l'entreprise (en mois)	Quantitative	-
Minorite	Appartenance à une minorité	Qualitative binaire	0 = non, 1 = oui
Genre	Sexe de l'employé	Qualitative binaire	0 = homme, 1 = femme
Cadre	Statut professionnel	Qualitative binaire	0 = non-cadre, 1 = cadre

Il est important de noter que le jeu de données est de bonne qualité : il ne présente **aucune valeur manquante** et **aucune ligne dupliquée**.

2.2 Statistiques Descriptives Globales

Les statistiques descriptives fournissent un premier aperçu des distributions des variables quantitatives :

Table 2 – Statistiques descriptives des variables quantitatives

Variable	Unité	Moyenne	Médiane	Écart-type	Min	Max
Salaire	$/an	34 420	28 875	17 076	15 750	135 000
Education	Années	13,49	12	2,88	8	21
Anciennete	Mois	176,97	140	105,10	65	554

Le **Salaire** présente une forte dispersion (écart-type élevé) et une moyenne supérieure à la médiane, ce qui suggère une distribution asymétrique avec quelques salaires très élevés. L'Éducation moyenne est d'environ 13,5 ans, soit un niveau post-secondaire. L'Ancienneté moyenne est d'environ 177 mois (près de 15 ans), avec une grande variabilité.

2.3 Répartition des Variables Binaires

La répartition des variables qualitatives binaires est la suivante :

TABLE 3 – Répartition des variables binaires

Variable	Catégorie	Effectif	Pourcentage
Cadre	Cadre (1)	84	17,7 %
	Non-cadre (0)	390	82,3 %
Genre	Femme (1)	216	45,6 %
	Homme (0)	258	54,4 %
Minorite	Oui (1)	104	21,9 %
	Non (0)	370	78,1 %

Seulement **17,7 %** des employés ont le statut de cadre. La population est légèrement majoritairement masculine (54,4 % d'hommes).

2.4 Comparaison des Profils "Cadre" vs "Non-Cadre"

La comparaison des profils moyens entre les deux groupes révèle des différences majeures :

TABLE 4 – Comparaison des profils Cadres et Non-Cadres

Variable	Non-Cadres (n=390)	Cadres (n=84)
Salaire moyen	28 053 $	63 978 $
Éducation moyenne	12,68 ans	17,25 ans
Ancienneté moyenne	180,9 mois	158,8 mois
Proportion de femmes	52,8 %	11,9 %
Proportion de minorités	25,6 %	4,8 %

Interprétation des résultats :

— Les cadres ont un salaire moyen plus du double de celui des non-cadres (63 978 $ contre 28 053 $) et un niveau d'études nettement supérieur (17,25 ans contre 12,68 ans).

— L'ancienneté moyenne est légèrement plus faible chez les cadres (158,8 mois) que chez les non-cadres (180,9 mois). Le statut de cadre est donc davantage lié au niveau de diplôme et au salaire qu'à la seule ancienneté.

— On observe une très forte **sous-représentation** des femmes (11,9 % des cadres) et des minorités (4,8 % des cadres). Ces chiffres soulignent une possible inégalité d'accès aux postes de responsabilité, souvent appelée **"plafond de verre"**.

2.5 Explication du Code (Q1)

Pour obtenir ces résultats, le *data scientist* utilise la librairie `pandas` de Python.

```python
df = pd.read_csv("Salaries.csv")

# Taille du jeu de données
print(df.shape)
# Types des variables et valeurs manquantes
print(df.info())
# valeurs manquantes par colonne
print(df.isna().sum())
# Verifier les doublons
df.duplicated().sum()
```

✓ 0.0s

FIGURE 1 – Code Python pour l'étude descriptive

```python
df.groupby('Cadre').agg({
    'Salaire': ['mean','median'],
    'Education': ['mean','median'],
    'Anciennete': ['mean','median'],
    'Genre': 'mean',
    'Minorite': 'mean'
})
```

✓ 0.0s

	Salaire		Education		Anciennete		Genre	Minorite
	mean	median	mean	median	mean	median	mean	mean
Cadre								
0	28053.179487	27000.0	12.682051	12.0	180.889744	140.0	0.528205	0.256410
1	63977.797619	60500.0	17.250000	17.0	158.773810	137.0	0.119048	0.047619

FIGURE 2 – Suite - Code Python pour l'étude descriptive

Ce bloc de code permet de lire le fichier, de vérifier sa qualité et de calculer les statistiques de base, notamment en utilisant la fonction `groupby("Cadre").agg(...)` pour comparer les moyennes des variables entre les deux groupes.

3 Étude de la Dépendance Linéaire (Question 2 – Corrélations)

3.1 Le Concept de Corrélation

La **corrélation de Pearson** mesure la force et la direction de la relation linéaire entre deux variables quantitatives. Le coefficient de corrélation (r) est compris entre -1 et +1. Une valeur proche de 1 ou -1 indique une relation forte, tandis qu'une valeur proche de 0 indique une absence de relation linéaire.

3.2 Résultats de la Matrice de Corrélation

Les corrélations clés sont les suivantes :

TABLE 5 – Corrélations clés avec le statut de Cadre

Relation	Coefficient de Corrélation (r)	Force et Direction
Cadre - Salaire	0,804	Très forte et positive
Cadre - Éducation	0,605	Forte et positive
Cadre - Genre	-0,314	Modérée et négative
Cadre - Minorite	-0,193	Faible et négative
Cadre - Anciennete	-0,080	Quasi nulle
Salaire - Éducation	0,661	Forte et positive

3.3 Interprétation

— **Facteurs Clés du Statut de Cadre :** Le **Salaire** et l'**Éducation** sont très fortement liés au statut de cadre, confirmant les observations de l'étude descriptive.
— **Inégalités :** Les corrélations négatives avec le **Genre** (rappel : 1 = femme) et la **Minorité** indiquent que les femmes et les minorités sont moins susceptibles d'être cadres.
— **Ancienneté :** La corrélation entre **Cadre** et **Ancienneté** est quasi nulle, ce qui signifie que l'ancienneté n'est pas un facteur déterminant pour l'accès au statut de cadre.
— **Salaire et Éducation :** La forte corrélation entre Salaire et Éducation ($r \approx 0,661$) est classique : la progression salariale est fortement associée au niveau de diplôme.

3.4 Explication du Code (Q2)

La matrice de corrélation est calculée simplement avec la méthode `.corr()` de `pandas`.

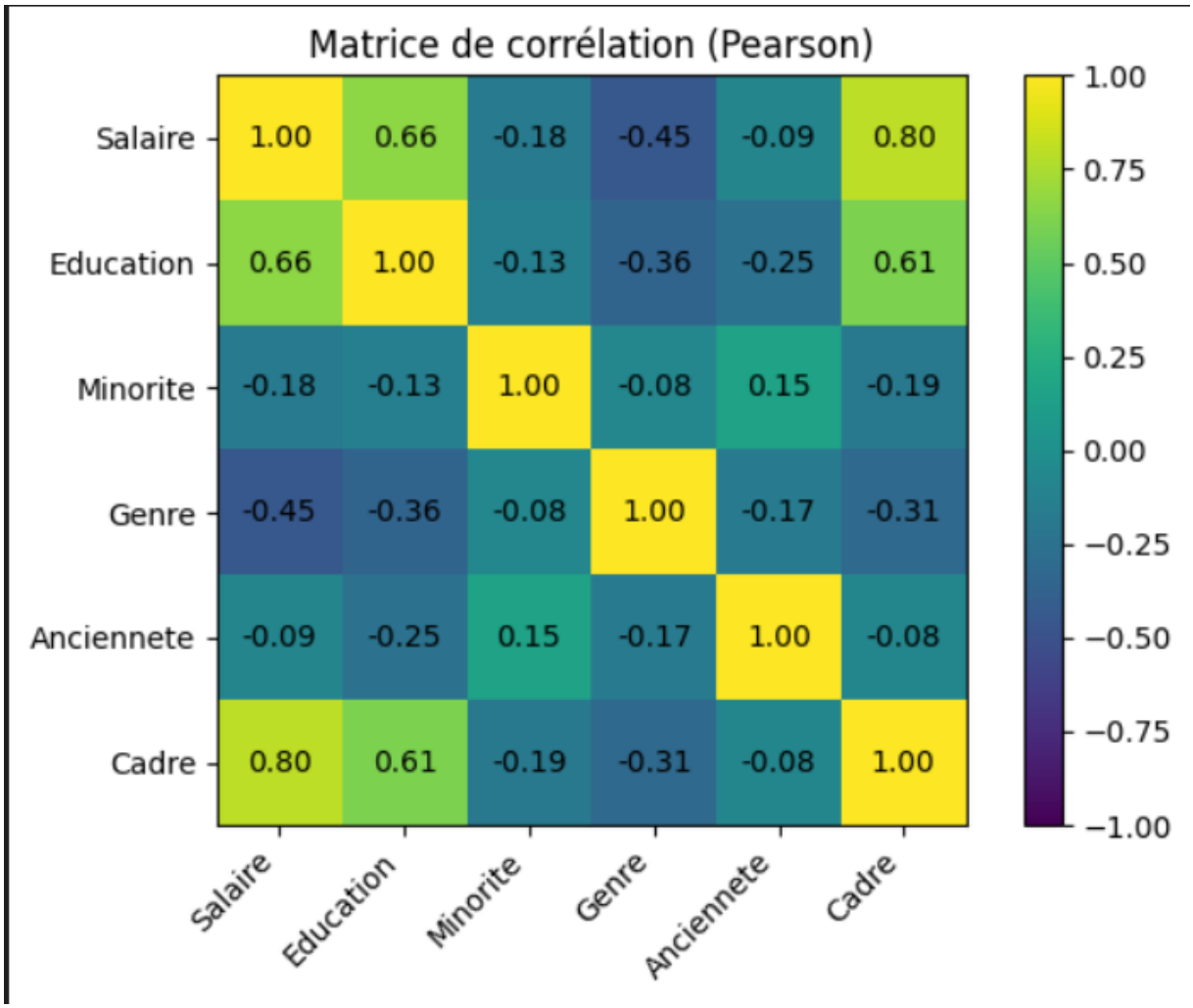

FIGURE 3 – le calcul de la matrice de corrélation

4 Modélisation de la Variable Cadre par Régression (Question 3 – Régression Logistique)

4.1 Choix du Modèle

La variable à expliquer, **Cadre**, étant binaire (0 ou 1), nous utilisons une **Régression Logistique** pour modéliser la probabilité qu'un employé soit cadre. Ce modèle garantit que la probabilité prédite se situe toujours entre 0 et 1.

4.2 Comparaison des Modèles

Deux modèles ont été testés. Le **Modèle Réduit** a été sélectionné car il est plus **parcimonieux** (plus simple) tout en conservant une performance très proche du modèle complet, comme l'indique un meilleur **BIC** (Critère d'Information Bayésien).

```
------------------------------------------------------------------
            Coef.    Std.Err.      z      P>|z|    [0.025    0.975]
------------------------------------------------------------------
const      -19.5239    3.4480  -5.6624  0.0000  -26.2818  -12.7660
SalaireK     0.1891    0.0298   6.3526  0.0000    0.1308    0.2475
Education    0.6718    0.2213   3.0356  0.0024    0.2380    1.1055
==================================================================

AIC modele complet : 108.63173294969616
AIC modele reduit  : 109.45373361480418
BIC modele complet : 133.59897687986663
BIC modele reduit  : 121.93735557988941
Accuracy modele reduit : 0.9535864978902954
Matrice de confusion (modele reduit) :
[[382   8]
 [ 14  70]]
```

FIGURE 4 – Les indicateurs de performance des modèles

Le **Modèle Logistique Réduit** est retenu : Cadre $\sim$ SalaireK + Éducation.

4.3 Interprétation du Modèle Réduit

L'équation du modèle, qui prédit le *logit* (le logarithme de l'odds de devenir cadre), est :

$$\text{logit}(P(\text{Cadre} = 1)) = -19,5239 + 0,1891 \times \text{SalaireK} + 0,6718 \times \text{Éducation} \quad (1)$$

Interprétation des Odds Ratios (OR) :

```
import numpy as np
np.exp(res_red.params)
```
✓ 0.0s

```
const         3.317978e-09
SalaireK      1.208178e+00
Education     1.957676e+00
dtype: float64
```

FIGURE 5 – Odds Ratios

— **OR(SalaireK)** $\approx 1,20$ **:** Pour une augmentation de 1 000 $ de salaire (à niveau d'éducation fixé), les chances d'être cadre sont multipliées par environ **1,20**.
— **OR(Éducation)** $\approx 1,96$ **:** Pour une année d'étude supplémentaire (à salaire fixé), les chances d'être cadre sont multipliées par environ **1,96** (soit presque le double).

Le modèle confirme le rôle prépondérant du salaire et du niveau d'étude dans l'accès au statut de cadre.

4.4 Explication du Code (Q3)

La régression logistique est réalisée avec la librairie `statsmodels`.

```
# Modèle reduit : Salaire + Education
X_red = df[["SalaireK", "Education"]]
X_red = sm.add_constant(X_red)

model_red = sm.Logit(y, X_red)
res_red = model_red.fit()
print(res_red.summary2())

# Qualité de prédiction
p_red = res_red.predict(X_red)
y_pred_red = (p_red >= 0.5).astype(int)

print("AIC modèle complet :", res_full.aic)
print("AIC modèle reduit  :", res_red.aic)
print("BIC modèle complet :", res_full.bic)
print("BIC modèle reduit  :", res_red.bic)

print("Accuracy modèle reduit :", accuracy_score(y, y_pred_red))
print("Matrice de confusion (modele reduit) :")
print(confusion_matrix(y, y_pred_red))
```

FIGURE 6 – Code Python pour la régression logistique modele reduite

5 Prédiction pour un Salarié Donné (Question 4)

Nous appliquons le Modèle Réduit au profil suivant : **Salaire** = 40 000 $/an (SalaireK = 40) et **Éducation** = 15 ans d'études.

5.1 Calcul de la probabilité

En utilisant l'équation 1 :

$$z = -19,5239 + 0,1891 \times 40 + 0,6718 \times 15 \approx -1,8829$$

La probabilité $P(\text{Cadre} = 1)$ est obtenue en appliquant la **fonction logistique** (ou sigmoïde) :

$$P(\text{Cadre} = 1) = \frac{1}{1 + e^{-z}} \approx 0,132$$

```python
new_emp = pd.DataFrame({
    "const": [1],
    "SalaireK": [40000 / 1000],
    "Education": [15]
})

# On ajoute l'intercept
new_emp = sm.add_constant(new_emp)

# Selon le modèle réduit
p_new = res_red.predict(new_emp)
print("Probabilite d'etre cadre :", float(p_new))
print("Prediction (0/1) :", int(p_new >= 0.5))
```

✓ 0.1s

```
Probabilite d'etre cadre : 0.13204648449078768
Prediction (0/1) : 0
```

FIGURE 7 – Prédiction pour 40 000 /*an*

5.2 Résultat de la prédiction

La probabilité que cet employé soit cadre est d'environ **13,2 %**. Avec un seuil de classification standard de 0,5, le modèle prédit que cet employé **n'est pas cadre**.

Ce résultat montre que, malgré un niveau d'études et un salaire corrects, la probabilité d'accéder au statut de cadre reste modérée, soulignant que le statut est fortement concentré chez les employés ayant des niveaux très élevés de ces deux variables.

6 Discussion Générale : Interprétation, Causes Possibles, Liens avec le Monde Réel

6.1 Liens avec le Monde du Travail

Les résultats confirment que les cadres sont en moyenne beaucoup plus diplômés et mieux payés. Cependant, l'étude a révélé une **sous-représentation alarmante des femmes et des minorités** parmi les cadres, faisant écho au phénomène du **"plafond de verre"**.

6.2 Causes Possibles

Ces inégalités peuvent être dues à :
— La **discrimination** (directe ou indirecte).
— Des différences d'**accès aux opportunités** (formation, réseaux professionnels).

— Le fait que les variables **Salaire** et **Éducation** elles-mêmes peuvent être le **résultat** d'inégalités en amont (politique salariale, accès aux études supérieures).

7 Limites de l'Étude et Pistes d'Amélioration

7.1 Limites

— Les données proviennent d'une seule entreprise, limitant la généralisation des conclusions.
— Le **Genre** est codé de manière binaire, ce qui est une simplification.
— L'absence de variables cruciales comme l'âge, le département ou les évaluations de performance.
— La **corrélation ne prouve pas la causalité**.

7.2 Pistes d'Amélioration

Il serait pertinent de :
— **Collecter plus de variables** (poste, performance).
— **Tester d'autres modèles** de classification (arbres de décision, *Random Forest*).
— **Ajouter une validation croisée** pour évaluer la robustesse du modèle.

8 Conclusion

Cette analyse a permis de confirmer le rôle prépondérant du **Salaire** et du **Niveau d'Éducation** dans l'accès au statut de cadre. Le modèle logistique réduit, simple et performant, a confirmé cette relation.

Néanmoins, l'étude a mis en lumière des **inégalités de représentation** importantes pour les femmes et les minorités. Le modèle statistique fournit une vision agrégée qui doit servir de base à une réflexion sur les politiques de Ressources Humaines, l'égalité des chances et la lutte contre les biais au sein de l'entreprise.