projet2_data

January 22, 2025

Projet 2 :Explorer et comprendre les performances des employés (Python ou R)</h1..>

Problématique métier

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

Objectif

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

Dans mon travail je me contenterai de faire les analyses univariées de certaines variables de l'entreprise.

0.1 importation des bibliothèques

[ ]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

0.1.1 Récupération des données depuis le fichiers csv

[12]:
```
data = pd.read_csv(r"C:
↪\Users\mabou\Documents\12_projet_analyse_de_donnees\projet_2\HRDataset_v14.
↪csv")
```

[10]:
```
data.head(10)
```

[10]:

	Employee_Name	EmpID	MarriedID	MaritalStatusID	GenderID	\
0	Adinolfi, Wilson K	10026	0	0	1	
1	Ait Sidi, Karthikeyan	10084	1	1	1	
2	Akinkuolie, Sarah	10196	1	1	0	
3	Alagbe,Trina	10088	1	1	0	
4	Anderson, Carol	10069	0	2	0	
5	Anderson, Linda	10002	0	0	0	
6	Andreola, Colby	10194	0	0	0	
7	Athwal, Sam	10062	0	4	1	
8	Bachiochi, Linda	10114	0	0	0	

```
9        Bacong, Alejandro   10250          0                 2          1

   EmpStatusID  DeptID  PerfScoreID  FromDiversityJobFairID  Salary  …  \
0            1       5            4                       0   62506  …
1            5       3            3                       0  104437  …
2            5       5            3                       0   64955  …
3            1       5            3                       0   64991  …
4            5       5            3                       0   50825  …
5            1       5            4                       0   57568  …
6            1       4            3                       0   95660  …
7            1       5            3                       0   59365  …
8            3       5            3                       1   47837  …
9            1       3            3                       0   50178  …

       ManagerName  ManagerID   RecruitmentSource PerformanceScore  \
0   Michael Albert       22.0           LinkedIn          Exceeds
1       Simon Roup        4.0             Indeed      Fully Meets
2   Kissy Sullivan       20.0           LinkedIn      Fully Meets
3     Elijiah Gray       16.0             Indeed      Fully Meets
4   Webster Butler       39.0      Google Search      Fully Meets
5         Amy Dunn       11.0           LinkedIn          Exceeds
6  Alex Sweetwater       10.0           LinkedIn      Fully Meets
7    Ketsia Liebig       19.0   Employee Referral      Fully Meets
8   Brannon Miller       12.0  Diversity Job Fair      Fully Meets
9     Peter Monroe        7.0             Indeed      Fully Meets

   EngagementSurvey EmpSatisfaction SpecialProjectsCount  \
0              4.60               5                    0
1              4.96               3                    6
2              3.02               3                    0
3              4.84               5                    0
4              5.00               4                    0
5              5.00               5                    0
6              3.04               3                    4
7              5.00               4                    0
8              4.46               3                    0
9              5.00               5                    6

  LastPerformanceReview_Date DaysLateLast30 Absences
0                  1/17/2019              0        1
1                  2/24/2016              0       17
2                  5/15/2012              0        3
3                   1/3/2019              0       15
4                   2/1/2016              0        2
5                   1/7/2019              0       15
6                   1/2/2019              0       19
7                  2/25/2019              0       19
```

```
8                  1/25/2019              0        4
9                  2/18/2019              0       16

[10 rows x 36 columns]
```

Notre analyse sera fait sur les variables suivantes : PerformanceScore, RaceDesc, Sex, MaritalDesc, Salary. Toutes nos analyses seront faites sur les ayant le status active.

[369]:
```
# création du dataframe des employés ayant le 'EmploymentStatus' active
# ie ceux qui travaille encore dans l'entreprise
dfEmployeActif = data[data['EmploymentStatus'] == "Active"]
```

Partie 1 : Statistiques descriptives univariés

Dans cette partie nous analysons les distributions de nos variables dans l'entreprise.

1 : La variable sexe

[373]:
```
dfEmployeActif['Sex'].value_counts(normalize=True) * 100
```

[373]:
```
Sex
F     56.038647
M     43.961353
Name: proportion, dtype: float64
```

les résultats nous montrent que le nombre d'employés de sexe féminin est légèrement supérieur au nombre d'employés de sexe masculin.

[ ]:
```

```

2 : La variable RaceDesc

[378]:
```
dfEmployeActif['RaceDesc'].value_counts(normalize=True) * 100
```

[378]:
```
RaceDesc
White                                59.903382
Black or African American            24.637681
Asian                                 9.661836
Two or more races                     3.864734
American Indian or Alaska Native      1.449275
Hispanic                              0.483092
Name: proportion, dtype: float64
```

Les résultats nous montrent la race blanches est la plus domminant avec près de 60% des employés suivit de la race noire qui représente un peu moins de 25% des employés.

3 : La variable MaritalDesc

[382]:
```
dfEmployeActif['MaritalDesc'].value_counts(normalize=True) * 100
```

[382]:
```
MaritalDesc
Single       48.792271
Married      37.198068
Divorced      6.763285
Separated     5.314010
Widowed       1.932367
Name: proportion, dtype: float64
```

Les célibataires sont majoritaires dans l'entreprise soit plus de 48% suivit des mariés qui repésentent une part non négligeagle des employés soit un peu plus 37%. Les autre status matrimoniale ont une part négligeable dans l'entreprise soit un peut moins de 14%.

4 : La variable PerformanceScore

[386]:
```
dfEmployeActif['PerformanceScore'].value_counts(normalize=True) * 100
```

[386]:
```
PerformanceScore
Fully Meets          78.260870
Exceeds              14.009662
Needs Improvement     3.864734
PIP                   3.864734
Name: proportion, dtype: float64
```

Plus de 92% des employés repondent pleinement au attente de performence avec un peu plus de 14% qui ont des performance plus que satisfaissante. Neanmoins une part presque marginale (plus de 7%) des employés restent en dessous des performances attendus.

5 : La variable Salary

[390]:
```
dfEmployeActif['Salary'].describe()
```

[390]:
```
count       207.000000
mean      70694.033816
std       27739.416425
min       45046.000000
25%       56593.000000
50%       63051.000000
75%       72816.000000
max      250000.000000
Name: Salary, dtype: float64
```

Ces résultats nous montrent que 50% des employés ont un salaire compris entre 56593$ et 72816$ mais l'ecart entre le salaire moyen et le salaire maximal est très signification ce qui peux signifier des grand écart entre les salaires.

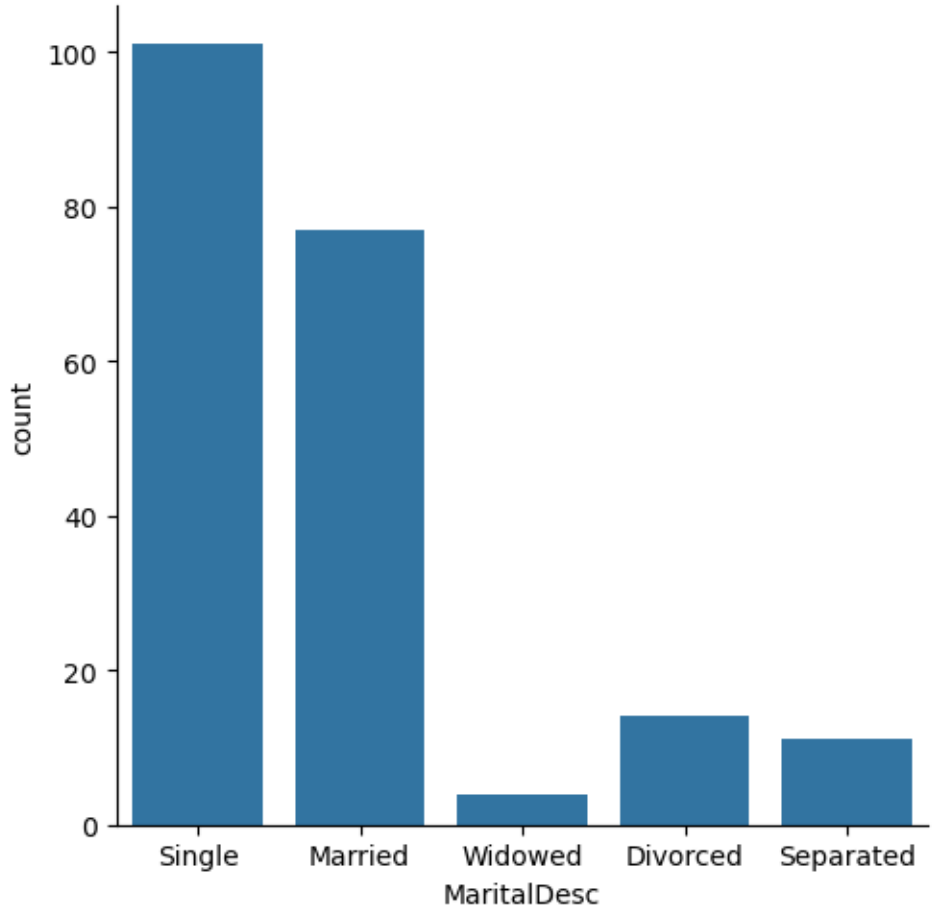
Partie 2 : Visualisations des distributions

Dans cette partie nous visualisons les distributions de nos variables dans l'entreprise à travers les histogrammes, les boxplots et les diagrammes à barre.

1 : La variable MaritalDesc

Diagramme à barre de la variable MaritalDesc qui représente les éffectifs pour chaque sous groupe

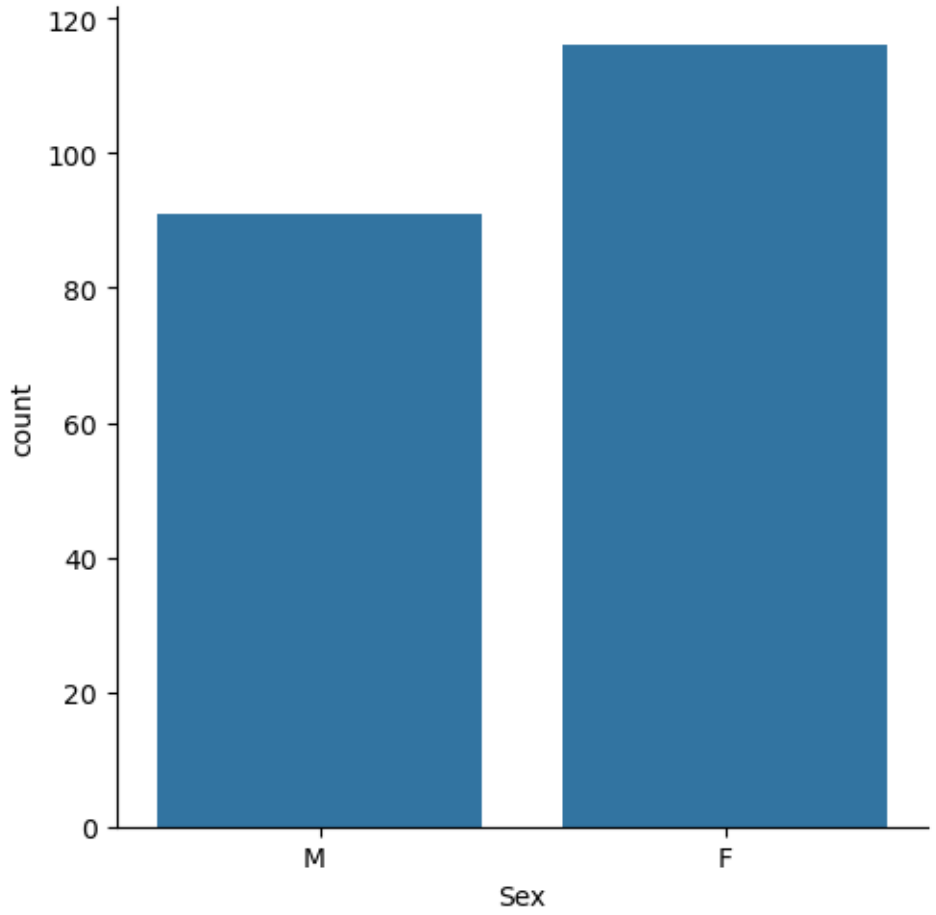
[395]:
```
sns.catplot(data=dfEmployeActif, x='MaritalDesc', kind='count')
plt.show()
```



2 : La variable Sexe

Diagramme à barre de la variable Sex qui représente les éffectifs pour chaque sous groupe

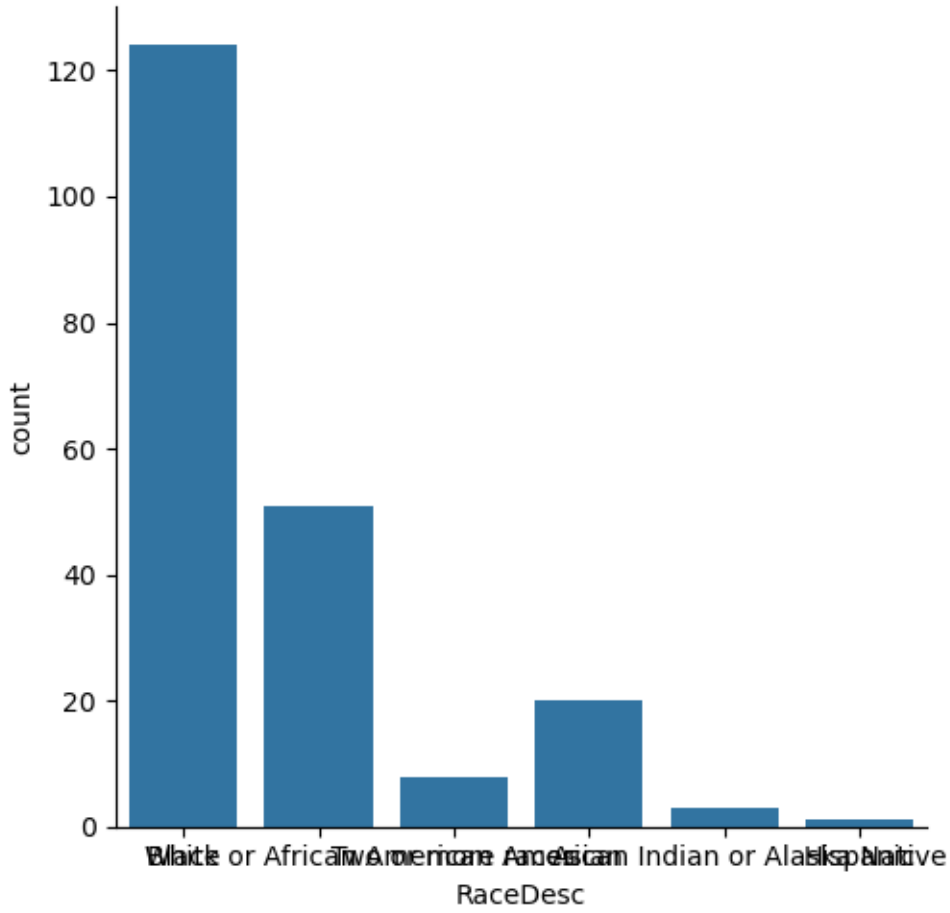
[398]:
```
sns.catplot(data=dfEmployeActif, x='Sex', kind='count')
plt.show()
```



3 : La variable RaceDesc

Diagramme à barre de la variable RaceDesc qui représente les éffectifs pour chaque sous groupe

[401]:
```
sns.catplot(data=dfEmployeActif, x='RaceDesc', kind='count')
plt.figure(figsize=(15, 3))
plt.show()
```



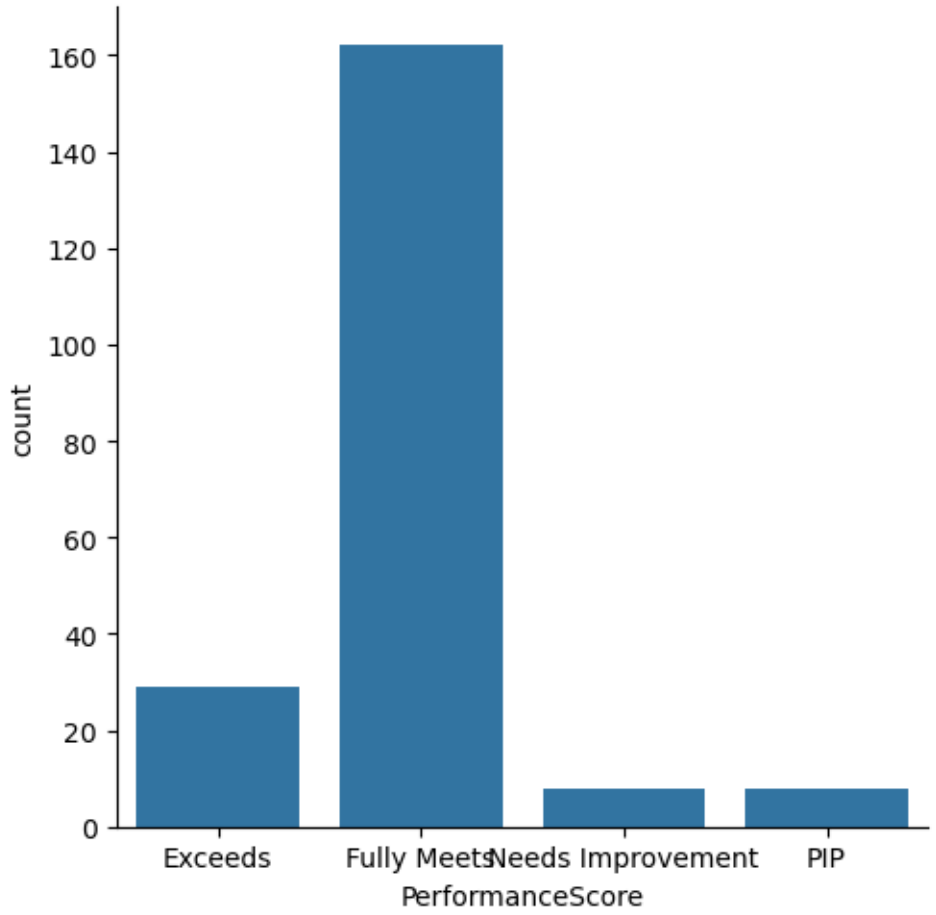

```
<Figure size 1500x300 with 0 Axes>
```

4 : La variable PerformanceScore

Diagramme à barre de la variable PerformanceScore qui représente les éffectifs pour chaque sous groupe

[404]:
```
sns.catplot(data=dfEmployeActif, x='PerformanceScore', kind='count')
```

[404]:
```
<seaborn.axisgrid.FacetGrid at 0x2379ece3530>
```

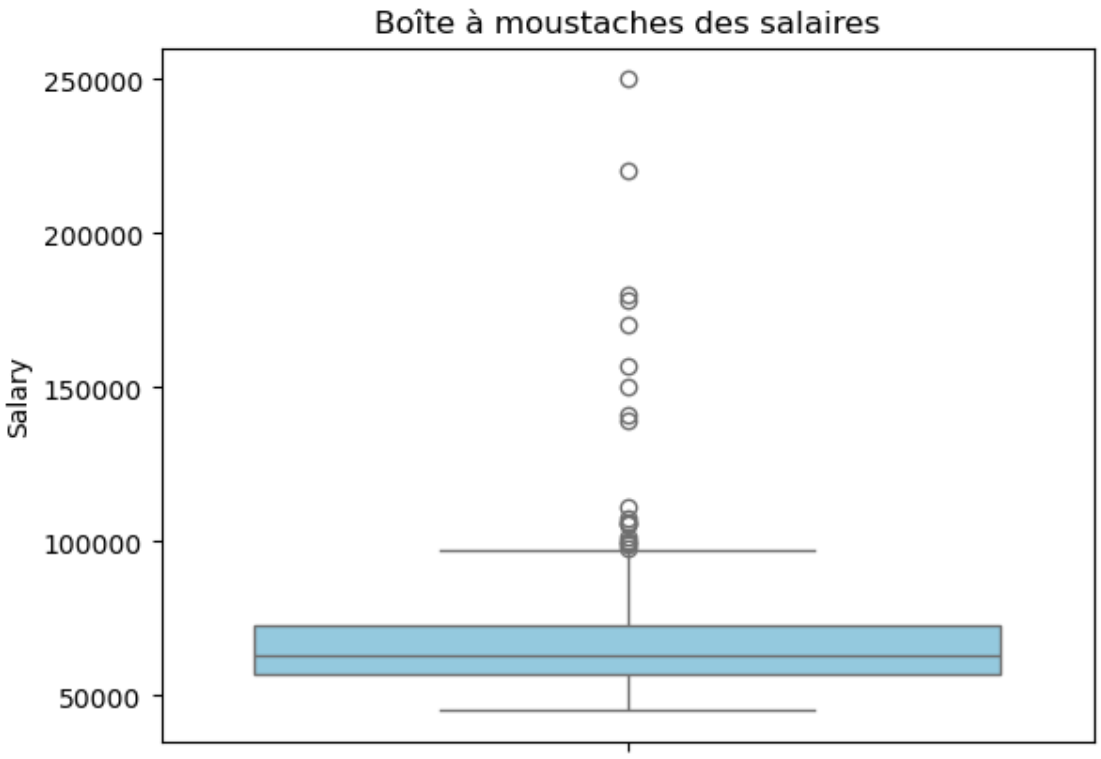


5 : La variable Salaire

Boite à moustache de la variable Salaire pour visualiser les salaires extrèmes

[407]:
```
sns.boxplot(data=dfEmployeActif['Salary'], color="skyblue")

# Ajout des titres
plt.title("Boîte à moustaches des salaires")
plt.show()
```



Ce graphique nous présentre un très grand nombre de salaires abèrant ce qui présente une très grande variabilité dans les salaires de l'entreprise.