



Analyses des Ressources Humaines

R language

- 1 Introduction
- 2 Description du Problème
- 3 Objectifs
- 4 Dataset
- 5 Méthodologie
- 6 Conclusion



INTRODUCTION

Le Data Mining est l'acte d'analyser d'énormes index d'informations afin de créer de nouvelles données.

En ressources humaines (RH), le Data Mining est un dispositif fondamental pour faire face au défi qui se développe rapidement. Cela se fait par l'enthousiasme croissant pour l'utilisation des informations sociales basées sur les travailleurs pour faire des prévisions et obtenir des données à un rythme rapide, ce qui facilite le processus de leadership de base dans la gestion des ressources humaines.

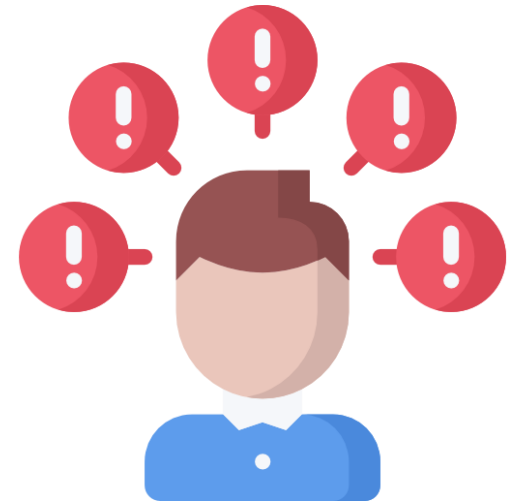
Description du Problème

L'objectif des services RH et managers associés au processus de recrutement consiste effectivement à bien sélectionner les candidats les plus adaptés aux postes afin de maximiser leurs chances de réussite dans l'entreprise.

L'entreprise a investi du temps et de l'argent pour les former, Pourtant quel quittent l'entreprise volontairement avant leur premier anniversaire

Il est important pour la direction de connaître les variables responsables des départs des employés et d'avoir également une prédiction sur les employés qui quitteront leur emploi à l'avenir.

Dès lors, comment remédier à ces départs prématurés ?



Objectifs



L'objectif de ce projet est de concevoir différents modèles pour prédire si un employé restera ou quittera l'entreprise au cours de la prochaine année et d'analyser l'exactitude des modèles.

Dataset



Dataset (humanresources.csv) sur les ressources humaines de Kaggle.com avec 11111 observations et 8 variables.

C'est un des données historiques nous donnant les informations qui ont quitté l'entreprise et qui n'ont pas quitté l'entreprise au cours de la dernière année.

Dans cet dataset, nous allons prédire la variable "vol_leave" (0 = rester, 1 = partir) en utilisant les autres variables.

Méthodologie

Prétraitement de données

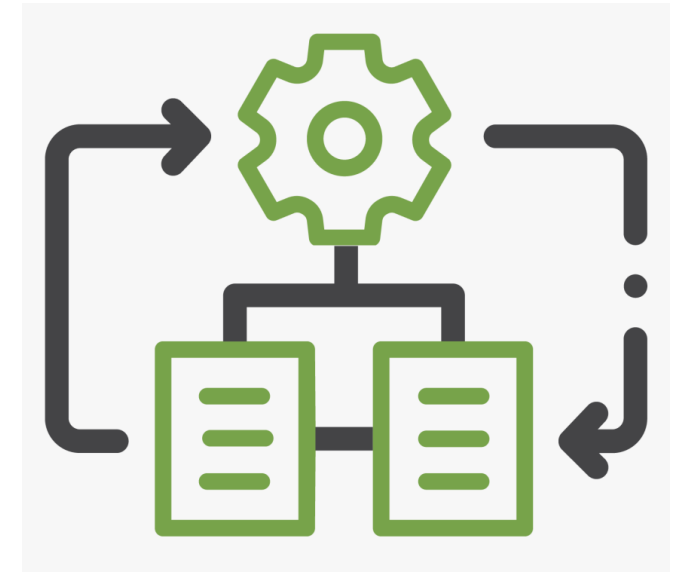
1. Exploration initiale des données
2. Préparation des données

Analyse de données

1. Performance vs Départ volontaire
2. Genre avec vs Départ volontaire
3. Département vs Départ volontaire
4. Role vs Départ volontaire
5. Age vs Départ volontaire
6. Salaire vs Départ volontaire

Construction des modèles prédictives

1. Modèle de Regression Logistique
 - Evaluation du pattern
2. Arbres de décisions
 - Evaluation du pattern



Méthodologie

Prétraitement de données : Exploration initiale des données

1. Description des données

11111 observations, et chaque observation contient des informations sur 8 variables :

- **Id** : identifiant de chaque employé dans l'entreprise.
- **Role** : le rôle de chaque employé dans l'entreprise.
- **Perf** : la performance de chaque employé dans l'entreprise.
- **Vol_leave** : boolean(1/0) si l'employé quitte l'entreprise volontairement cela vaut 1, sinon 0.
- **Sex** : le sexe de l'employé (male/female)
- **Area** : le département du travail.
- **Age** : l'âge de l'employé.
- **Salary**: Salaire de l'employé.

Méthodologie

Prétraitement de données : Exploration initiale des données

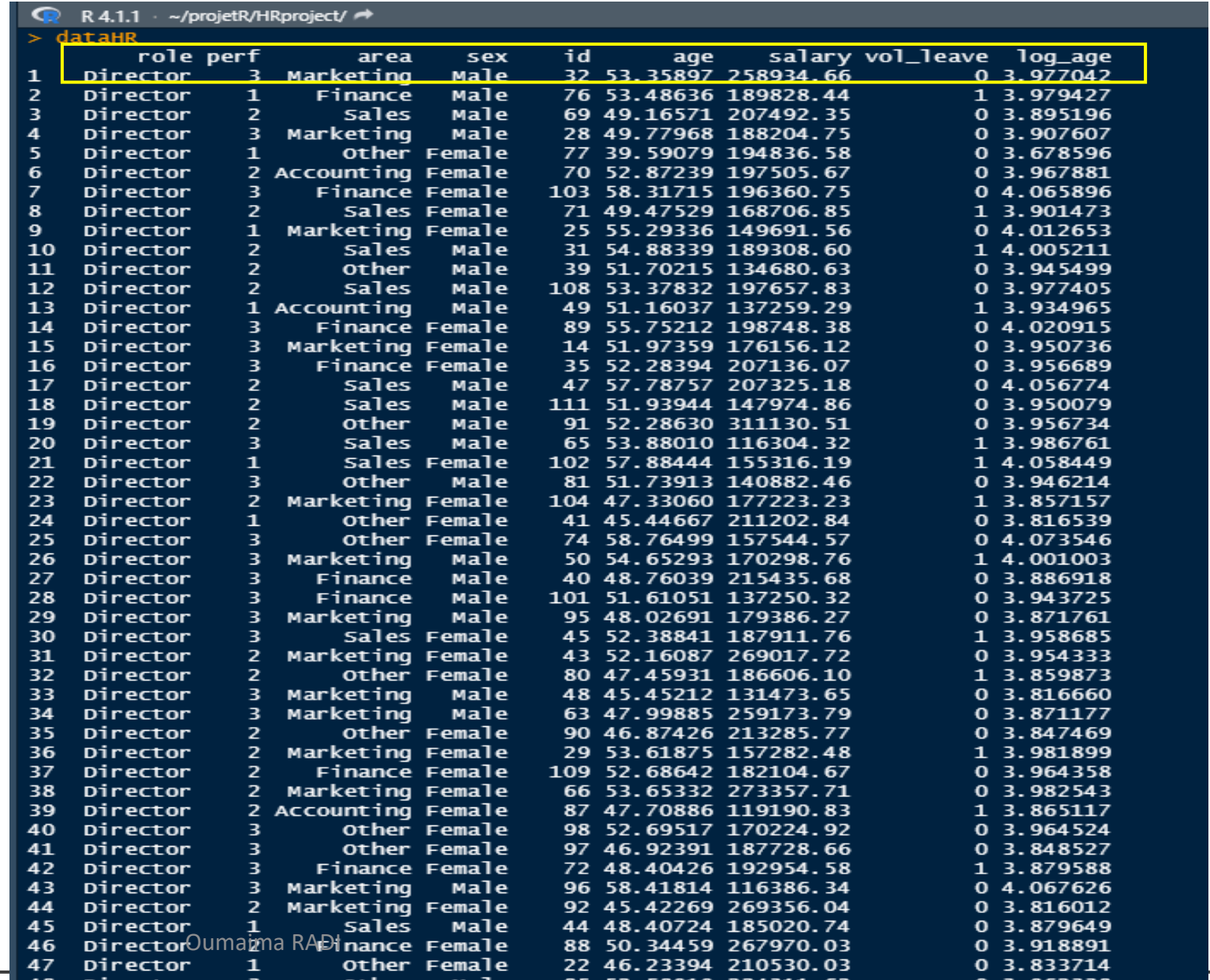
2. Chargement des données dans R

```
R 4.1.1 · ~/projetR/HRproject/ ↵  
> dataHR <- read.csv("C:/dataHR.csv")  
>  
> str(dataHR)  
'data.frame':  11111 obs. of  8 variables:  
 $ role      : chr  "CEO" "Director" "Director" "Director" ...  
 $ perf      : int   3 3 1 2 3 1 2 3 2 1 ...  
 $ area      : chr  "Sales" "Marketing" "Finance" "Sales" ...  
 $ sex       : chr  "Male" "Male" "Male" "Male" ...  
 $ id        : int   1 32 76 69 28 77 70 103 71 25 ...  
 $ age       : num   62 53.4 53.5 49.2 49.8 ...  
 $ salary    : num  1000000 258935 189828 207492 188205 ...  
 $ vol_leave: int    0 0 1 0 0 0 0 0 1 0 ...  
> |
```

Méthodologie

Prétraitement de données :
Exploration initiale des données

* Visualiser le dataset :



R 4.1.1 · ~/projetR/HRproject/ ↗

```
> dataHR
```

	role	perf	area	sex	id	age	salary	vol_leave	log_age
1	Director	3	Marketing	Male	32	53.35897	258934.66	0	3.977042
2	Director	1	Finance	Male	76	53.48636	189828.44	1	3.979427
3	Director	2	Sales	Male	69	49.16571	207492.35	0	3.895196
4	Director	3	Marketing	Male	28	49.77968	188204.75	0	3.907607
5	Director	1	Other	Female	77	39.59079	194836.58	0	3.678596
6	Director	2	Accounting	Female	70	52.87239	197505.67	0	3.967881
7	Director	3	Finance	Female	103	58.31715	196360.75	0	4.065896
8	Director	2	Sales	Female	71	49.47529	168706.85	1	3.901473
9	Director	1	Marketing	Female	25	55.29336	149691.56	0	4.012653
10	Director	2	Sales	Male	31	54.88339	189308.60	1	4.005211
11	Director	2	Other	Male	39	51.70215	134680.63	0	3.945499
12	Director	2	Sales	Male	108	53.37832	197657.83	0	3.977405
13	Director	1	Accounting	Male	49	51.16037	137259.29	1	3.934965
14	Director	3	Finance	Female	89	55.75212	198748.38	0	4.020915
15	Director	3	Marketing	Female	14	51.97359	176156.12	0	3.950736
16	Director	3	Finance	Female	35	52.28394	207136.07	0	3.956689
17	Director	2	Sales	Male	47	57.78757	207325.18	0	4.056774
18	Director	2	Sales	Male	111	51.93944	147974.86	0	3.950079
19	Director	2	Other	Male	91	52.28630	311130.51	0	3.956734
20	Director	3	Sales	Male	65	53.88010	116304.32	1	3.986761
21	Director	1	Sales	Female	102	57.88444	155316.19	1	4.058449
22	Director	3	Other	Male	81	51.73913	140882.46	0	3.946214
23	Director	2	Marketing	Female	104	47.33060	177223.23	1	3.857157
24	Director	1	Other	Female	41	45.44667	211202.84	0	3.816539
25	Director	3	Other	Female	74	58.76499	157544.57	0	4.073546
26	Director	3	Marketing	Male	50	54.65293	170298.76	1	4.001003
27	Director	3	Finance	Male	40	48.76039	215435.68	0	3.886918
28	Director	3	Finance	Male	101	51.61051	137250.32	0	3.943725
29	Director	3	Marketing	Male	95	48.02691	179386.27	0	3.871761
30	Director	3	Sales	Female	45	52.38841	187911.76	1	3.958685
31	Director	2	Marketing	Female	43	52.16087	269017.72	0	3.954333
32	Director	2	Other	Female	80	47.45931	186606.10	1	3.859873
33	Director	3	Marketing	Male	48	45.45212	131473.65	0	3.816660
34	Director	3	Marketing	Male	63	47.99885	259173.79	0	3.871177
35	Director	2	Other	Female	90	46.87426	213285.77	0	3.847469
36	Director	2	Marketing	Female	29	53.61875	157282.48	1	3.981899
37	Director	2	Finance	Female	109	52.68642	182104.67	0	3.964358
38	Director	2	Marketing	Female	66	53.65332	273357.71	0	3.982543
39	Director	2	Accounting	Female	87	47.70886	119190.83	1	3.865117
40	Director	3	Other	Female	98	52.69517	170224.92	0	3.964524
41	Director	3	Other	Female	97	46.92391	187728.66	0	3.848527
42	Director	3	Finance	Female	72	48.40426	192954.58	1	3.879588
43	Director	3	Marketing	Male	96	58.41814	116386.34	0	4.067626
44	Director	2	Marketing	Female	92	45.42269	269356.04	0	3.816012
45	Director	1	Sales	Male	44	48.40724	185020.74	0	3.879649
46	Director	2	Finance	Female	88	50.34459	267970.03	0	3.918891
47	Director	1	Other	Female	22	46.23394	210530.03	0	3.833714
48	Director	2	Other	Male	85	52.58810	224211.62	0	3.862228

Oumama RADI

Méthodologie

Prétraitement de données :Exploration initiale des données :

3. Summarisation

```
R 4.1.1 · ~/projetR/HRproject/ ↵
> #=====Summarisation=====
> summary(dataHR)
  role          perf          area          sex          id          age
Length:11111   Min.   :1.000   Length:11111   Length:11111   Min.    :    1   Min.   :22.02
Class :character 1st Qu.:2.000   Class :character 1st Qu.:2778   1st Qu.: 2778   1st Qu.:24.07
Mode  :character Median :2.000   Mode  :character Median :5556   Median : 5556   Median :25.70
              Mean  :2.198              Mean  : 5556   Mean  : 5556   Mean  :27.79
              3rd Qu.:3.000              3rd Qu.: 8334   3rd Qu.: 8334   3rd Qu.:28.49
              Max.   :3.000              Max.   :11111   Max.   :11111   Max.   :62.00

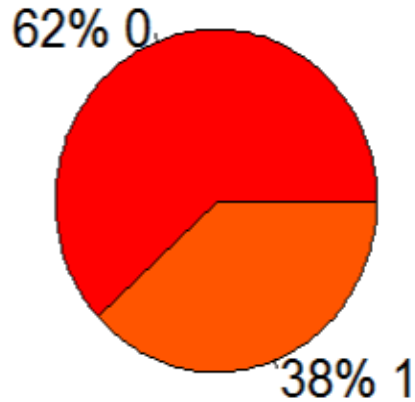
 salary          vol_leave
Min.   : 42168   Min.   :0.0000
1st Qu.: 57081   1st Qu.:0.0000
Median : 60798   Median :0.0000
Mean   : 65358   Mean   :0.3812
3rd Qu.: 64945   3rd Qu.:1.0000
Max.   :1000000   Max.   :1.0000
> |
```

Méthodologie

Prétraitement de données : Exploration initiale des données :

4. Visualiser les employés qui quittent volontairement

vol_leave proportions - n: 400



L'entreprise perd 38% de son temps et argent pour les former, et pourtant ils quittent l'entreprise volontairement avant leur premier anniversaire.

Méthodologie

Prétraitement de données : Préparation des données

1. Nettoyage de données : Valeurs manquantes , inconsistante, noise

```
R 4.1.1 · ~/projetR/HRproject/ ↵  
> #=====Missing values=====  
>  
> which(is.na(dataHR)) #--> pas de valeurs manquantes  
integer(0)  
> |
```

Méthodologie

Prétraitement de données : Préparation des données

2. Selection de données

```
> data[data$role=="CEO",]
  role perf area sex id age salary vol_leave
1  CEO    3 sales Male 1 62 1e+06          0
```

⇒ Le CEO n'a pas quitté son poste donc ce rôle n'a pas d'influence sur le résultat donc on va éliminer le CEO.

```
> data[data$role=="VP",]
  role perf area sex id age salary vol_leave
11102 VP    2 Finance Male 7 51.70082 546598.8          0
11103 VP    2 Sales Female 3 54.55917 549328.9          0
11104 VP    2 Accounting Female 6 56.57139 503215.2          0
11105 VP    3 Marketing Female 8 56.63474 513601.1          0
11106 VP    2 Marketing Male 4 50.91491 517050.3          0
11107 VP    2 Sales Female 11 50.91673 494109.2          1
11108 VP    2 Other Male 10 55.04723 555182.7          0
11109 VP    3 Sales Male 5 57.86106 526169.9          0
11110 VP    2 Marketing Male 9 60.40083 534392.7          0
11111 VP    2 Accounting Male 2 55.37020 508399.1          0
```

⇒ 90% des VP n'ont pas quitté leur poste et pourcentage des VP est $(10/11111) \times 100 = 0,09\%$ donc le rôle de VP a une influence négligée donc on va éliminer le VP

Méthodologie

Prétraitement de données : Préparation des données

2. Selection de données

Comme nous le voyons, il y a 5 types de rôles dans l'ensemble de données, à savoir, PDG, Directeur, Ind, Manager et VP. Mais depuis, PDG et VP tombent dans un segment distinct des autres postes, nous ne les incluons pas dans notre modèle. Par conséquent, maintenant appeler à nouveau les données et les résumer.

```
R 4.1.1 · ~/projetR/HRproject/ ↗
> dataHR = filter(dataHR, dataHR$role == "Ind" | dataHR$role == "Manager" | dataHR$role == "Director")
> dataHR$role <- factor(dataHR$role)
>
> summary(dataHR)
```

role	perf	area	sex	id	age	salary
Director: 100	Min. :1.000	Length:11100	Length:11100	Min. : 12	Min. :22.02	Min. : 42168
Ind :10000	1st Qu.:2.000	Class :character	Class :character	1st Qu.: 2787	1st Qu.:24.07	1st Qu.: 57080
Manager : 1000	Median :2.000	Mode :character	Mode :character	Median : 5562	Median :25.70	Median : 60788
	Mean :2.198			Mean : 5562	Mean :27.77	Mean : 64860
	3rd Qu.:3.000			3rd Qu.: 8336	3rd Qu.:28.48	3rd Qu.: 64928
	Max. :3.000			Max. :11111	Max. :61.67	Max. :311131

vol_leave
Min. :0.0000
1st Qu.:0.0000
Median :0.0000
Mean :0.3815
3rd Qu.:1.0000
Max. :1.0000

```
> |
```

Méthodologie

Analyse de données :

1. Performance vs départ volontaire

Étant donné que la variable de sortie de réponse se compose de deux groupes, c'est-à-dire (0, 1), la comparer avec d'autres colonnes serait beaucoup plus facile si nous utilisons une fonction d'agrégat.

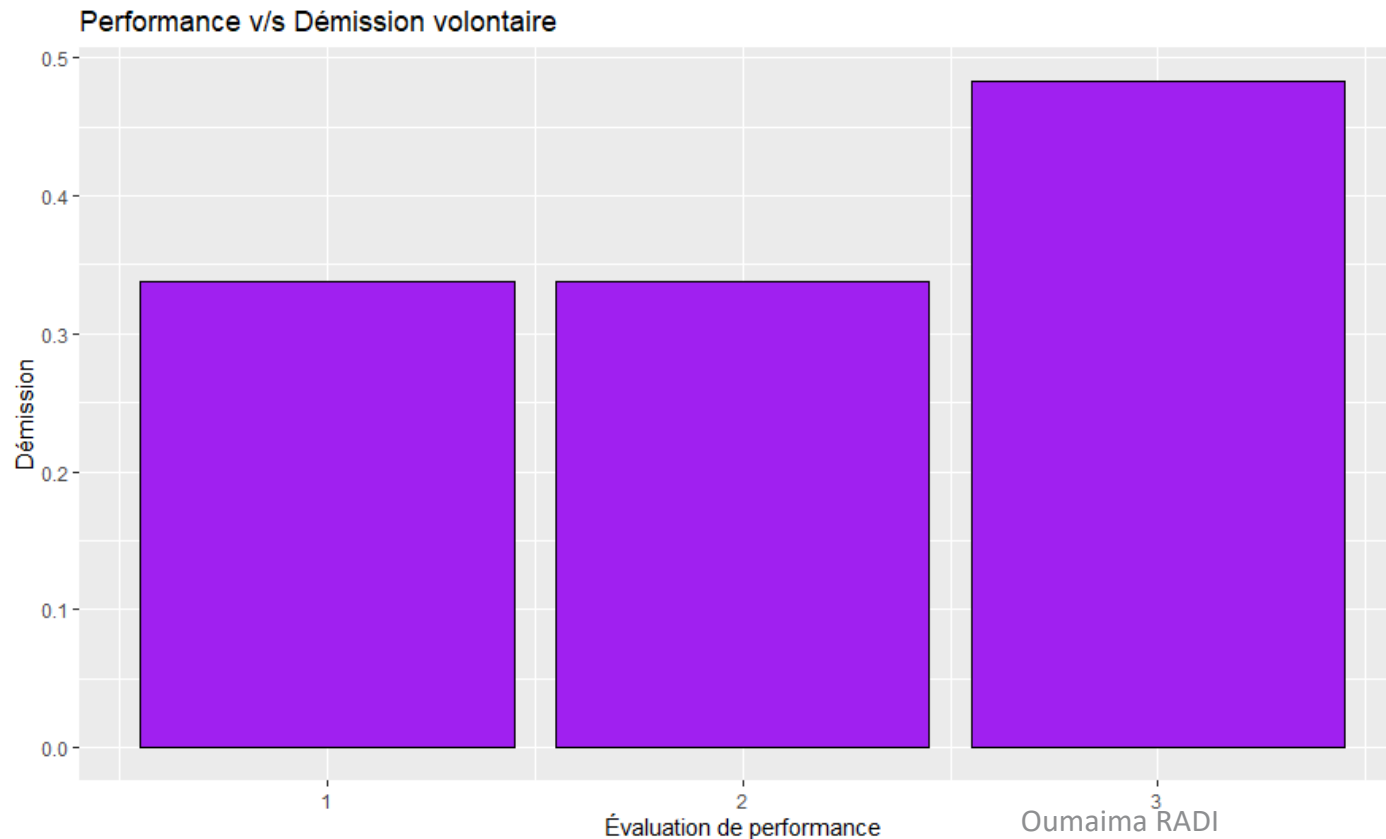
```
R 4.1.1 · ~/projetR/HRproject/ ↵  
> #++++++Performance vs quitter volontairement++++++  
> agg_perf = aggregate(vol_leave ~ perf, data = dataHR, mean)  
> agg_perf  
  perf vol_leave  
1     1 0.3375112  
2     2 0.3383831  
3     3 0.4831122  
> |
```


Méthodologie

Analyse de données :

1. Performance vs Départ volontaire

```
R 4.1.1 · ~/projetR/HRproject/ ↗  
> ggplot(agg_perf, aes(x = perf, y = vol_leave)) + geom_bar(stat = "identity", fill = 'purple', colour = 'black') + ggtitle  
("Performance v/s Quitter volontairement") + labs(y = "Quitter volontairement", x = "Évaluation de performance")  
> |
```



L'histogramme montre que les employés ayant une note de **performance plus élevée** sont plus susceptibles de quitter l'entreprise.

Méthodologie

Analyse de données :

2. Genre vs Départ volontaire

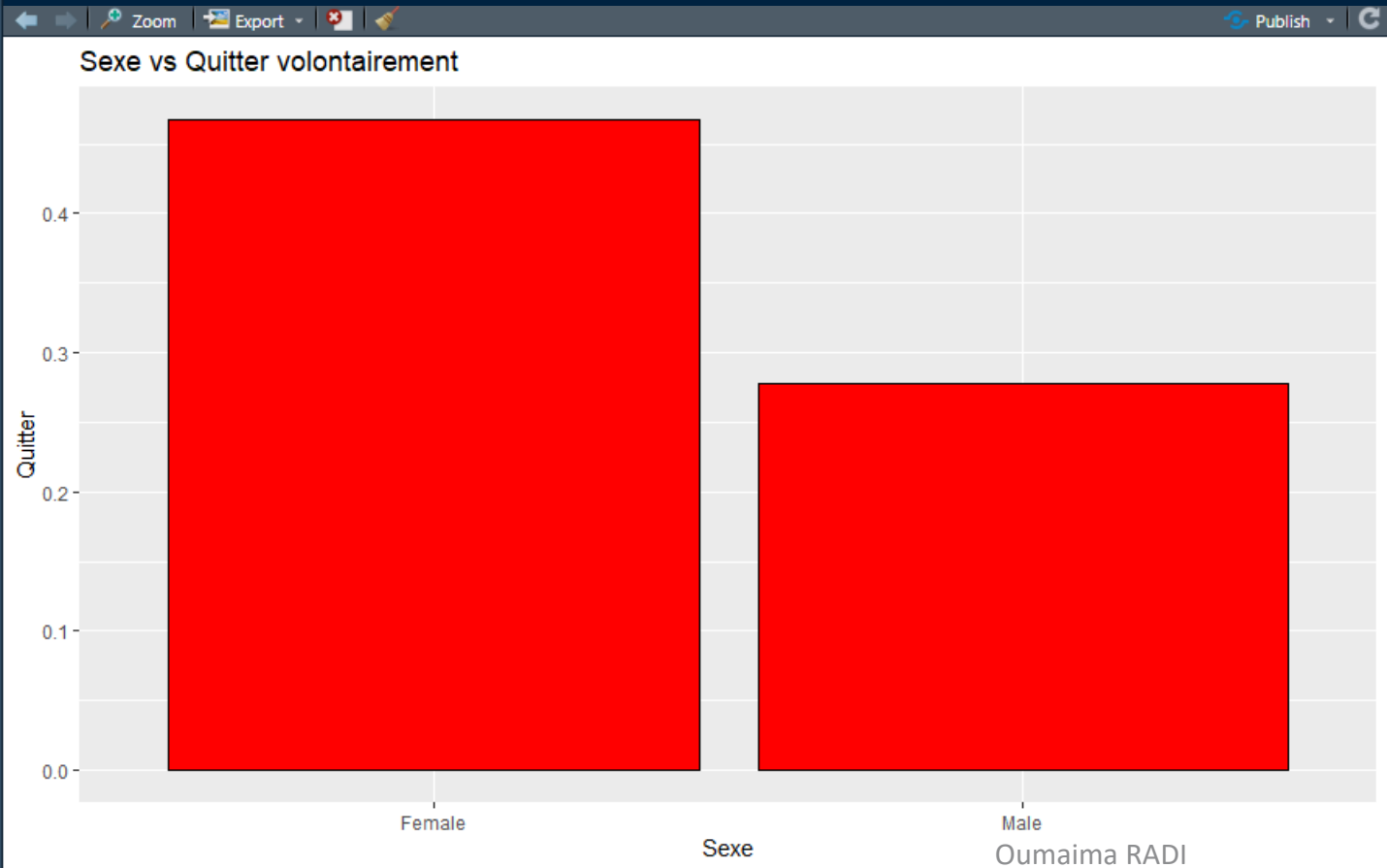
```
> #+++++++quitter volontairement vs sexe+++++++  
> agg_sex = aggregate(vol_leave ~ sex, data = dataHR, mean)  
> agg_sex  
      sex vol_leave  
1 Female 0.4673483  
2   Male 0.2781970  
> |
```

Méthodologie

Analyse de données :

2. Genre vs Départ volontaire

```
R 4.1.1 · ~/projetR/HRproject/ ↗  
> ggplot(agg_sex, aes(x = sex, y = vol_leave)) + geom_bar(stat = "identity", fill = 'red', colour = 'black') + ggtitle("Sexe  
vs Quitter volontairement") + labs(y = "Quitter", x = "Sexe")  
> |
```



Le plot montre que les employées femmes sont plus susceptibles de quitter leur emploi que les hommes.

Méthodologie

Analyse de donnees :

3. Département vs départ volontaire

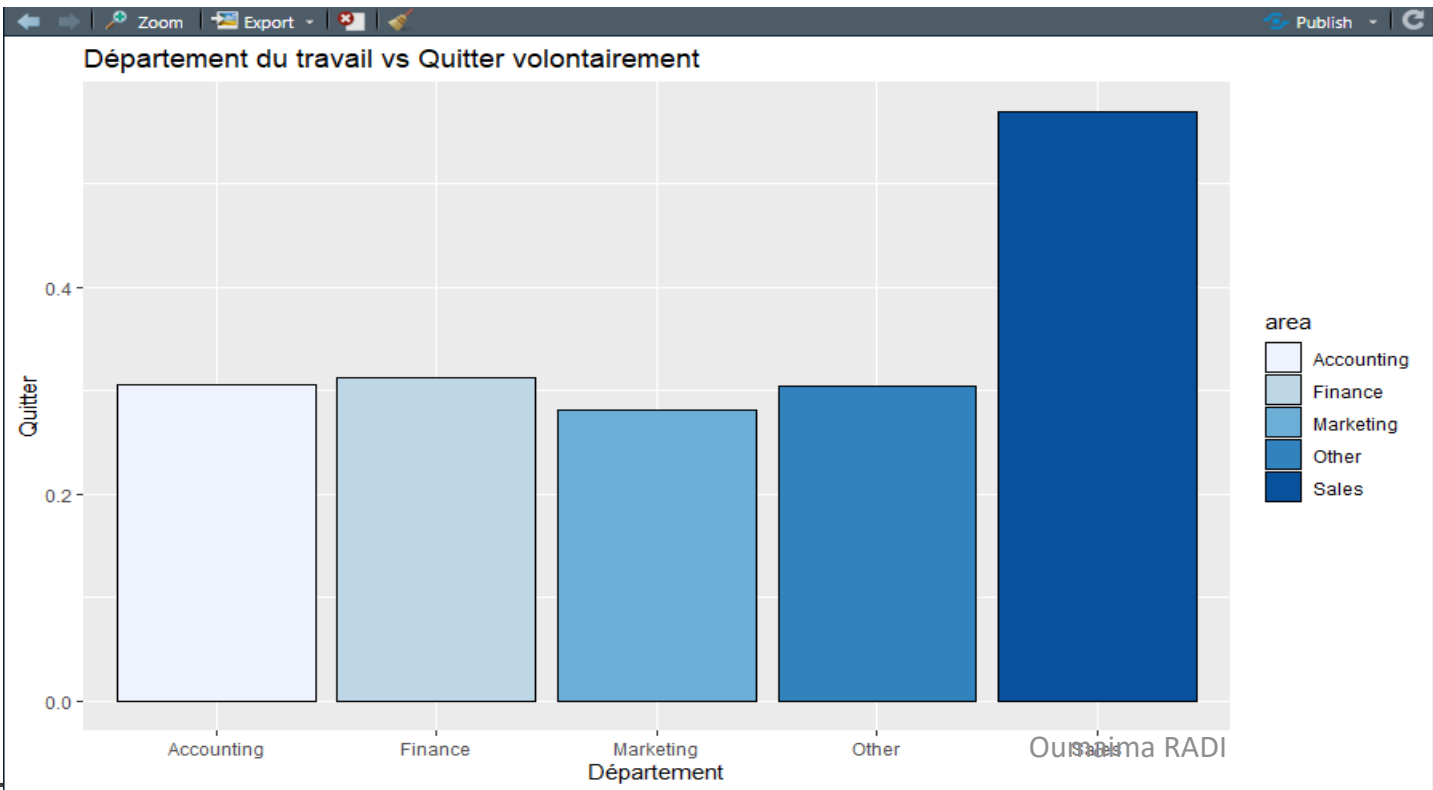
```
R 4.1.1 · ~/projetR/HRproject/ ↵  
> #+++++++quitter volontairement vs département du travail+++++++  
>  
> agg_area = aggregate(vol_leave ~ area, data = dataHR, mean)  
> agg_area  
      area vol_leave  
1 Accounting 0.3055383  
2   Finance 0.3126492  
3 Marketing 0.2815965  
4    other 0.3040510  
5    Sales 0.5696880  
> |
```

Méthodologie

Analyse de données :

3. Département vs départ volontaire

```
R 4.1.1 · ~/projetR/HRproject/ ↗  
> ggplot(agg_area, aes(x = area, y = vol_leave, fill = area)) + geom_bar(stat =  
+ "identity", colour = "black") + scale_fill_brewer() + ggtitle("Département du travail vs Quitter volontairement") + labs  
(y = "Quitter", x = "Département")  
*** recursive gc invocation  
Warning: stack imbalance in 'lapply', 73 then 71  
Warning: stack imbalance in 'lapply', 62 then 60  
> |
```



Le plot montre que les employés du service des ventes sont les plus susceptibles de quitter leur emploi par rapport aux autres domaines d'activité dans la société.

Méthodologie

Analyse de données :

4. Role vs Départ volontaire

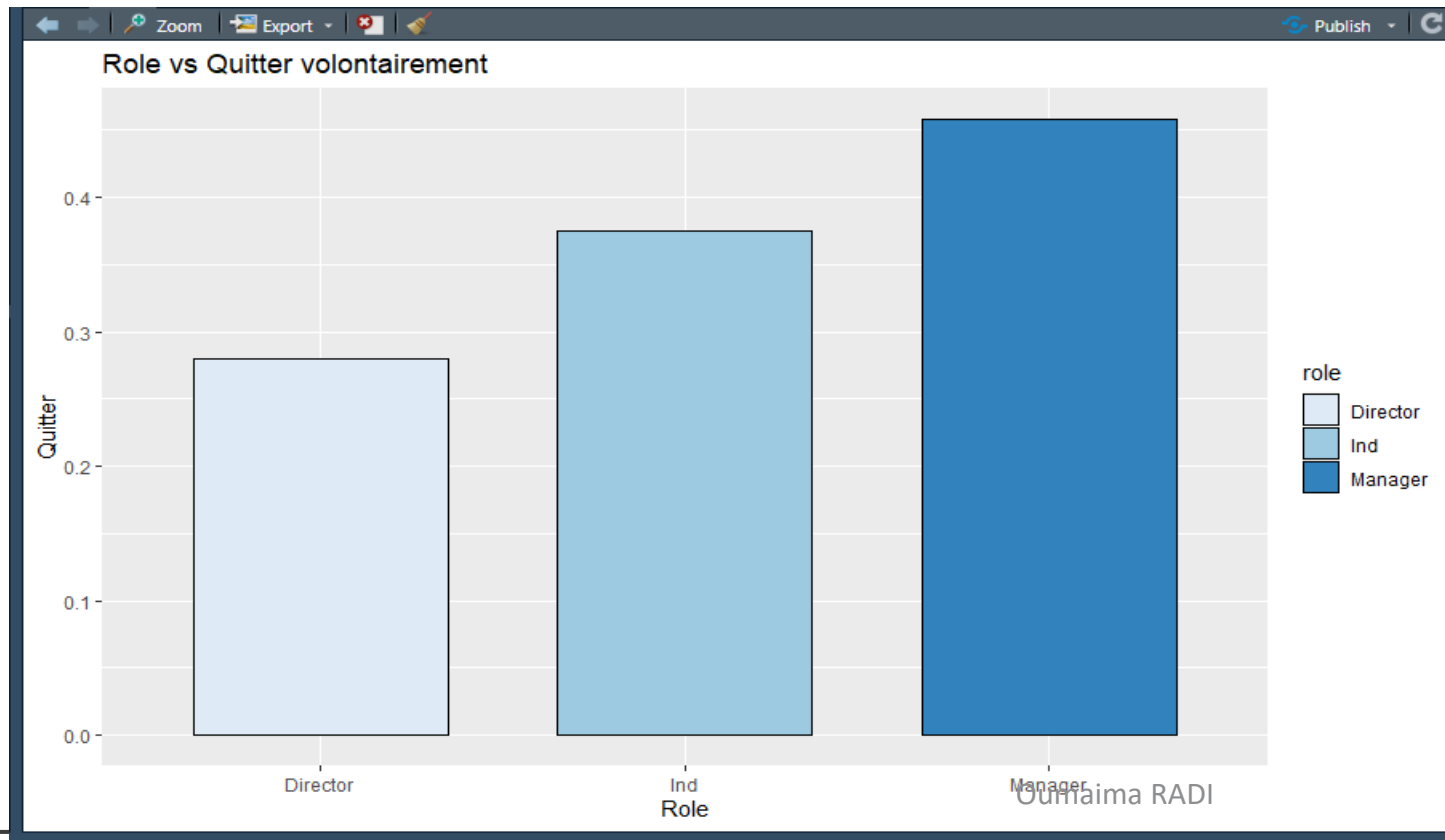
```
R 4.1.1 · ~/projetR/HRproject/ ↵  
> #+++++++quitter volontairement vs Role+++++++  
> agg_role = aggregate(vol_leave ~ role, data = dataHR, mean)  
> agg_role  
      role vol_leave  
1 Director    0.2800  
2      Ind    0.3749  
3  Manager    0.4580  
> |
```

Méthodologie

Analyse de données :

4. Role vs Départ volontaire

```
R 4.1.1 · ~/projetR/HRproject/ ↗  
> ggplot(agg_role, aes(x = role, y = vol_leave, fill = role)) + geom_bar(stat = "identity", width = .7,  
+ colour = 'black') + scale_fill_brewer() + ggtitle("Role vs Quitter volontairement") + labs (y = "Quitter", x= "Role")  
> |
```



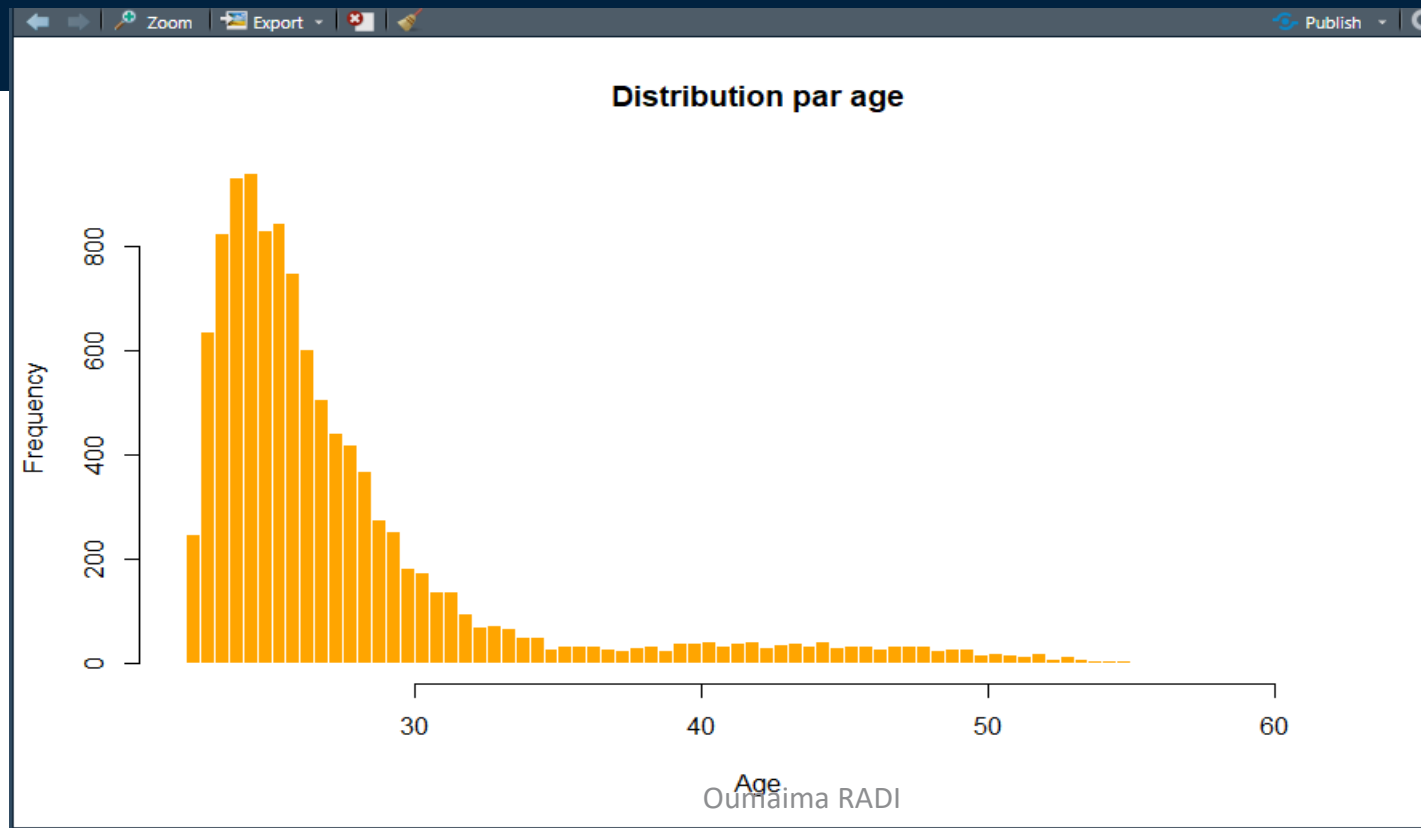
Le plot montre que **les managers sont les plus susceptibles de quitter leurs emplois** tandis que **les administrateurs sont les moins susceptibles de quitter leurs emplois.**

Méthodologie

Analyse de données :

5. Age avec Départ volontaire

```
R 4.1.1 · ~/projetR/HRproject/ ↗  
> #+++++Quitter volontairement vs Age+++++  
> hist(dataHR$age, breaks = 100, main = "Distribution par age", border = F,  
+       xlab = "Age", col = 'orange')  
> |
```



Age
Ourmaïma RADI

Méthodologie

Analyse de données :

5. Age vs Départ volontaire

```
R 4.1.1 · ~/projetR/HRproject/ ↵  
> quantile(dataHR$age, probs = seq(0,1,.1))  
 0%    10%   20%   30%   40%   50%   60%   70%   80%   90%  100%  
22.02289 23.14094 23.76757 24.36880 25.01564 25.69533 26.55048 27.73737 29.51513 35.70077 61.67132  
> |
```



90 % des salariés se situent dans la tranche d'âge de 22 à 36 ans.
Cette catégorisation semble faussée.

```
R 4.1.1 · ~/projetR/HRproject/ ↵  
> library(e1071)  
> skewness(dataHR$age)  
[1] 2.2669  
> |
```



Nous voyons que la distribution par âge est positive/versée à droite, ce qui implique que la moyenne est inférieure à la médiane.

Par conséquent, on prend le **log** de la variable âge.

Méthodologie

Analyse de données :

5. Age vs Départ volontaire

```
R 4.1.1 · ~/projetR/HRproject/ ↵  
> dataHR$log_age = log(dataHR$age)  
> summary(dataHR$log_age)  
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   
 3.092  3.181  3.246  3.304  3.349  4.122   
> |
```

Catégorisons davantage la répartition par âge en termes de rôles des employés dans l'entreprise.

```
R 4.1.1 · ~/projetR/HRproject/ ↵  
> boxplot(age ~ role, data = dataHR, col = 'pink', xlab = "Role dans la société",  
+         ylab = "Age du salarié", main = 'Distribution par Age en termes de Role')  
> |
```

Méthodologie

Analyse de données :

5. l'Age vs Départ volontaire



Le box plot ci-joint montre qu'il existe une **relation** entre le rôle du salarié dans l'entreprise et son âge. Les administrateurs se situent dans la tranche d'âge la plus élevée tandis que les salariés **Ind** se situent dans la tranche d'âge inférieure à moyenne.

Méthodologie

Analyse de données :

5. l'Age vs Départ volontaire

Agrégeons maintenant la variable âge pour voir la relation avec le départ des salariés.

```
Console Terminal x Jobs x
R 4.1.1 · ~/projetR/HRproject/ ➔
> #+++++agrèger la variable d'âge pour voir la relation avec le départ des employés+++++
> agg_age = aggregate(x = dataHR$vol_leave, by = list(cut(dataHR$age, 10)), mean)
> agg_age
```

	Group.1	x
1	(22,26]	0.3866177
2	(26,30]	0.3645902
3	(30,33.9]	0.3374536
4	(33.9,37.9]	0.3992806
5	(37.9,41.8]	0.4155405
6	(41.8,45.8]	0.4640288
7	(45.8,49.8]	0.4333333
8	(49.8,53.7]	0.4260870
9	(53.7,57.7]	0.4666667
10	(57.7,61.7]	0.2727273

```
> |
```

Méthodologie

Analyse de données :

5. l'Age vs Départ volontaire

R 4.1.1 · ~/projetR/HRproject/ →

```
> names(agg_age) = c("Age", "Probabilité")  
> ggplot(agg_age, aes(x = Age, y = Probabilité, fill = Age)) + geom_bar(stat =  
+ "identity", width = .1, colour = 'black') + scale_fill_brewer() +  
+ ggtitle("Age vs Départ volontaire") + labs(y = "Quitter", x = "Age")
```

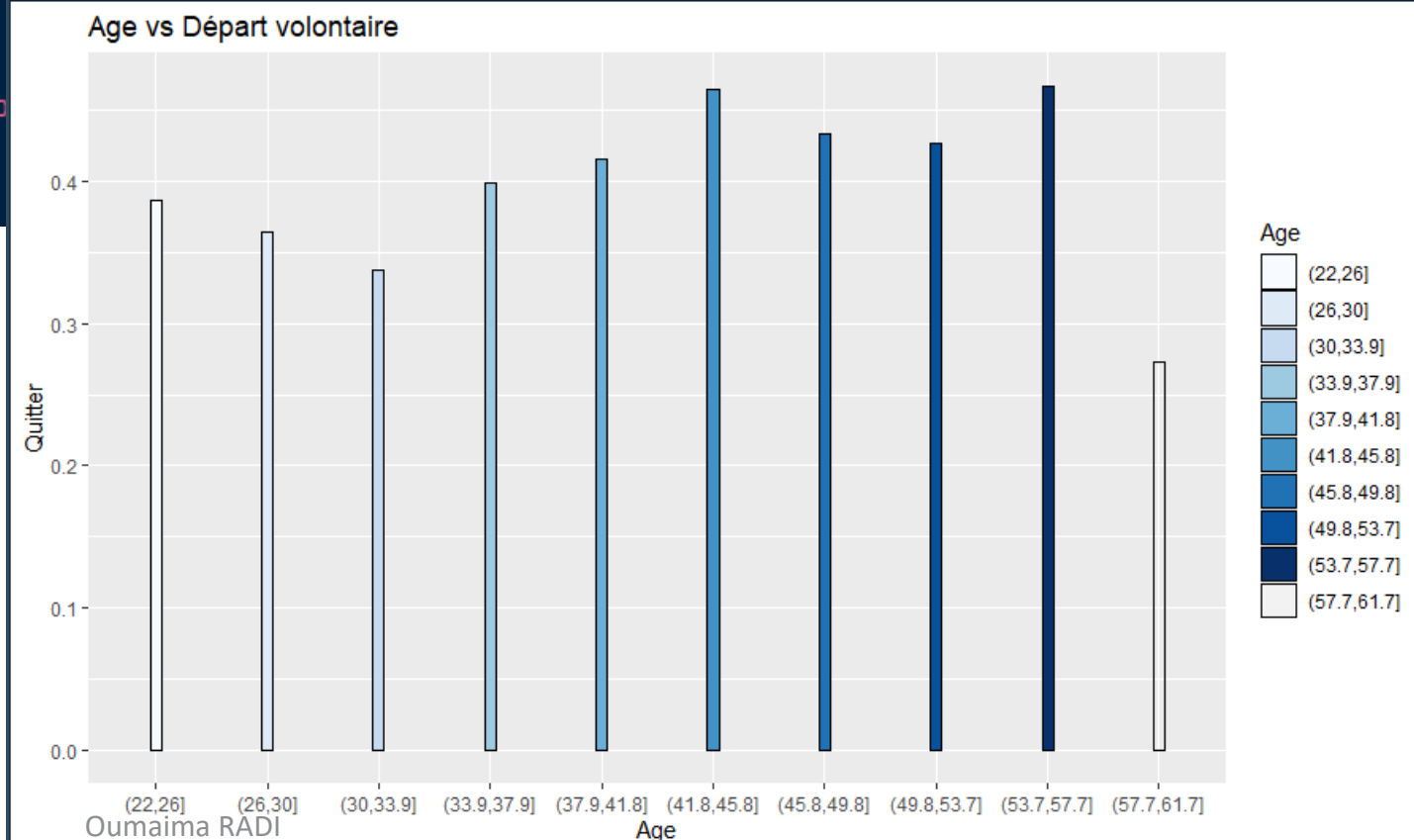
warning message:

```
In RColorBrewer::brewer.pal(n, pal) :  
n too large, allowed maximum for palette Blues is 9  
Returning the palette you asked for with that many colors
```

```
> |
```

Le graphique ci-joint montre que les employés âgés de **42 à 57 ans sont les plus susceptibles de quitter leur emploi** par rapport aux employés de 22 à 41 ans.

Et les employés de **plus de 57 ans sont les moins susceptibles de quitter leur emploi**, puisque c'est généralement le rôle du PDG et du directeur.



Méthodologie

Analyse de données :

6. Salaire vs Départ volontaire

```
R 4.1.1 · ~/projetR/HRproject/ ↗  
> #+++++Départ volontaire vs Salaire+++++  
> summary(dataHR$salary)  
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   
 42168   57080   60788   64860   64928  311131   
> quantile(dataHR$salary, probs = seq(0,1,.2))  
      0%      20%      40%      60%      80%     100%   
42168.22  56189.17  59385.03  62307.14  66151.43 311130.51   
> |
```

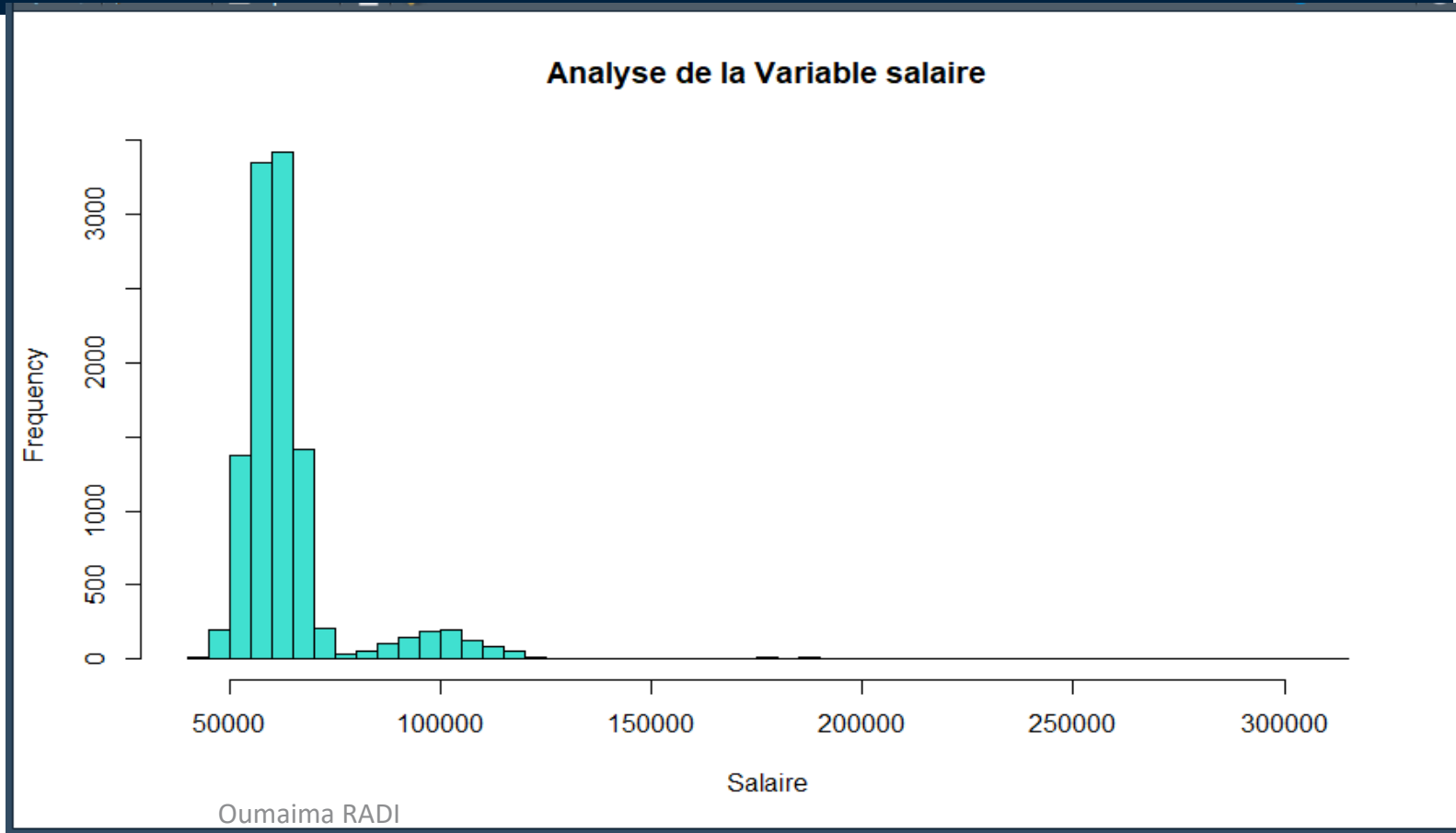
Méthodologie

Analyse de données :

6. Salaire vs Départ volontaire

```
R 4.1.1 · ~/projetR/HRproject/ ➔  
> hist(dataHR$salary, breaks = 50, col = 'turquoise', main = "Analyse de la variable salaire", xlab = "Salaire")  
> |
```

La médiane du salaire des employés est 60788 \$. Mais le maximum des employés (80%) gagnent jusqu'à 66173,65 \$.



Méthodologie

Construction du modèle prédictive

1. Modèle de regression logistique

Tout d'abord, nous devons diviser nos données en un ensemble d'apprentissage et un ensemble de test. Les deux tiers des données sont dédiés à l'ensemble de données d'entraînement et un tiers est dédié à l'ensemble de données de test.

```
R 4.1.1 · ~/projetR/HRproject/ ↵  
> #-----Diviser le dataset en 2 (test & apprentissage)-----  
> set.seed(42)  
> div_dataHR = sample.split(dataHR$vol_leave, 2/3)  
> train = dataHR[div_dataHR,]  
> test = dataHR[!div_dataHR,]  
> |
```


Méthodologie

Construction du modèle prédictive

1. Modèle de regression logistique

Nous avons un problème de classification, les résultats étant « Rester » ou « Partir » prédit par les variables significatives. Nous utilisons donc la régression logistique pour ajuster le modèle.

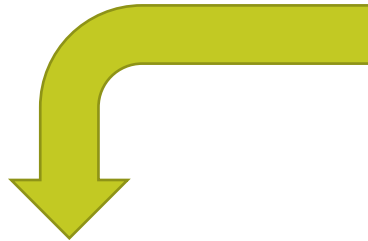
```
>  
>  
> test_mean = mean(test$vol_leave)  
> train_mean = mean(train$vol_leave)  
> print(c(test_mean, train_mean))  
[1] 0.3816216 0.3814865  
> |
```

Méthodologie

Construction du modèle prédictive

1. Modèle de regression logistique

Ajustement du modèle en utilisant
generalized linear model (GLM)



Maintenant, en vérifiant la valeur p pour toutes les variables indépendantes, nous voyons que **areaFinance**, **areaMarketing**, **areaOther** sont des facteurs non significatifs car la valeur p est supérieure à 0,05.

```
R 4.1.1 · ~/projetR/HRproject/ ↗
> #Ajustement du modèle en utilisant GLM
> ajst = glm(vol_leave ~ role + perf + area + sex + log_age + salary, data= dataHR, family = 'binomial')
> summary(ajst)

Call:
glm(formula = vol_leave ~ role + perf + area + sex + log_age +
    salary, family = "binomial", data = dataHR)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4737  -0.9123  -0.6068   1.0906   3.2238

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.290e+01  1.100e+00  11.725 < 2e-16 ***
roleInd      -8.146e+00  5.573e-01 -14.617 < 2e-16 ***
roleManager  -4.865e+00  4.327e-01 -11.242 < 2e-16 ***
perf         4.931e-01  3.598e-02  13.703 < 2e-16 ***
areaFinance   3.517e-02  7.920e-02   0.444 0.657003
areaMarketing -9.517e-02  7.490e-02  -1.271 0.203862
areaOther     -9.540e-05  7.471e-02  -0.001 0.998981
areaSales     1.239e+00  6.799e-02  18.230 < 2e-16 ***
sexMale      -9.435e-01  4.374e-02 -21.571 < 2e-16 ***
log_age      -7.516e-01  2.037e-01  -3.689 0.000225 ***
salary       -6.515e-05  3.723e-06 -17.501 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 14759  on 11099  degrees of freedom
Residual deviance: 13004  on 11089  degrees of freedom
AIC: 13026

Number of Fisher Scoring iterations: 4

> | Oumaima RADI
```

Méthodologie

Construction du modèle prédictive

1. Modèle de regression logistique

Maintenant, nous allons analyser la déviance pour tester les différences entre deux ou plusieurs moyennes par ANOVA (Analyse de la variance) en utilisant la méthode du Chi-Square.


```
R 4.1.1 ~./projetR/HRproject/ ↗
> anova(ajst, test = "chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: vol_leave

Terms added sequentially (first to last)

      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                11099      14759
role      2      30.69     11097      14728 2.162e-07 ***
perf      1     161.14     11096      14567 < 2.2e-16 ***
area      4     735.02     11092      13832 < 2.2e-16 ***
sex       1     466.69     11091      13365 < 2.2e-16 ***
log_age   1      11.21     11090      13354 0.0008158 ***
salary    1     350.08     11089      13004 < 2.2e-16 ***
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```



La déviance est une mesure de la qualité de l'ajustement pour un modèle.

La différence entre la déviance nulle et la déviance résiduelle ainsi que les faibles valeurs de p montre toutes les variables significatives

Méthodologie

Construction du modèle prédictive

1. Modèle de regression logistique

Maintenant, analyser la capacité prédictive de notre modèle via la matrice de confusion.

```
R 4.1.1 · ~/projetR/HRproject/ ↵  
> pred_model = predict(fit, test, type = 'response')  
> pred_model = ifelse(pred_model > 0.5,1,0)  
> MCE = mean(pred_model != test$vol_leave)  
> table(actual = test$vol_leave, prediction = pred_model)  
      prediction  
actual    0     1  
    0 1919  369  
    1  780  632  
> |
```

Méthodologie

Construction du modèle prédictive

1. Modèle de regression logistique

* Evaluation du pattern :

```
> #Calcul de la précision du modèle  
> print(paste('Accuracy', 1 - MCE))  
[1] "Accuracy 0.689459459459459"  
> |
```



L'ACCURACY de ce model = 68.94%

Méthodologie

Construction du modèle prédictive

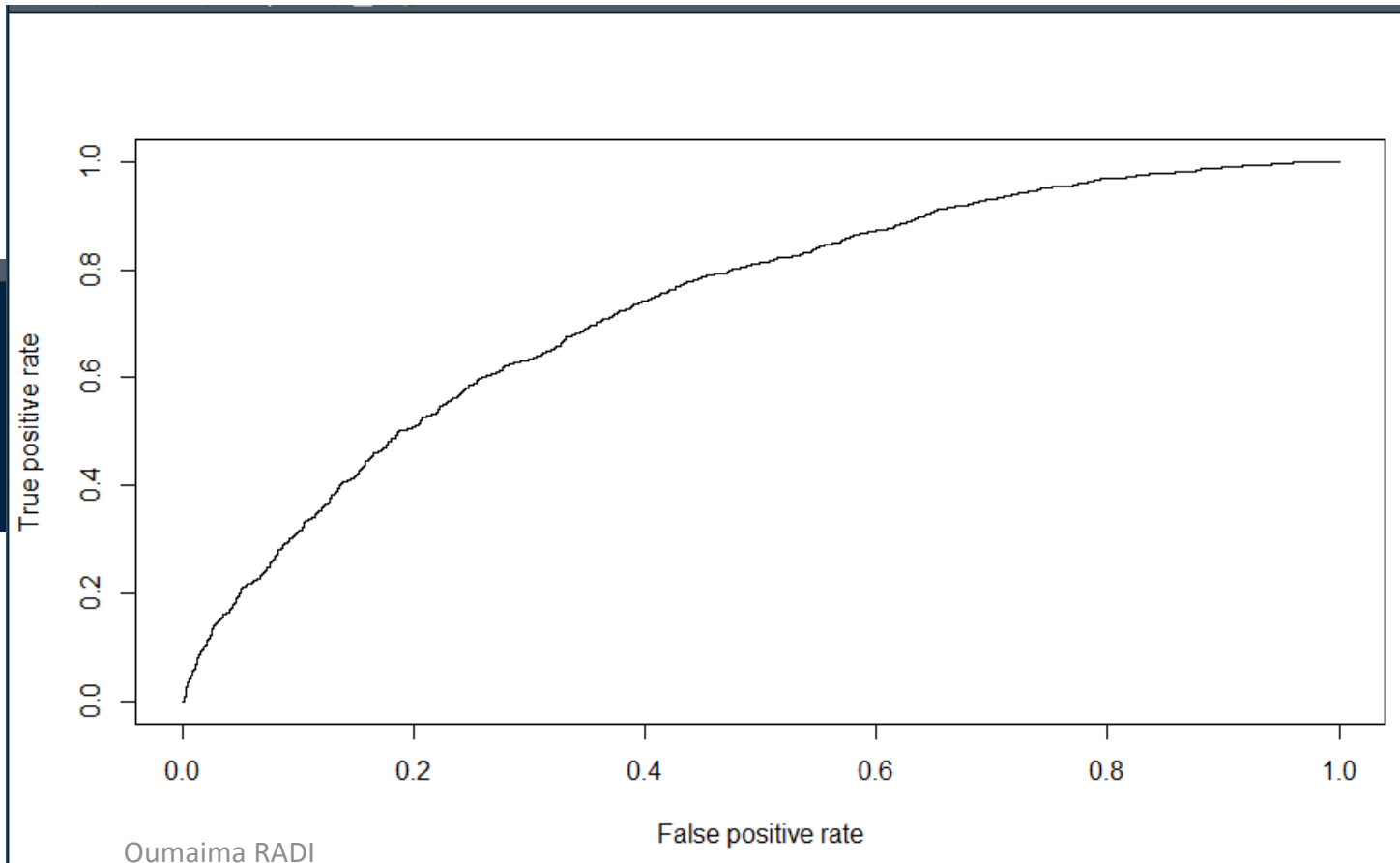
1. Modèle de regression logistique

* Evaluation du pattern :

Nous allons tracer la courbe ROC et calculer l'AUC qui sont des mesures de performances typiques pour un classificateur binaire.

1- Tracer la courbe ROC :

```
R 4.1.1 ~ /projetR/HKproject/
> #Courbe ROC
warning message:
In diff.default(xscale) : reached elapsed time limit
> plot1 = predict(ajst, test, type = "response")
> plot2 = prediction(plot1, test$vol_leave)
> plot3 = performance(plot2, measure = "tpr", x.measure = "fpr")
> plot(plot3)
>
```



Méthodologie

1. Modèle de regression logistique

* Evaluation du pattern :

2- Calculer AUC :

```
R 4.1.1 ~/projetR/HRproject/ ↗  
> #Calculer AUC (air sous la courbe)  
> AUC = performance(plot2, measure = "auc")  
> AUC = AUC@y.values[[1]]  
> AUC  
[1] 0.7326298  
> |
```



Sur la base de la règle empirique, un modèle a une bonne capacité prédictive si l'AUC est plus proche de 1. Selon notre analyse, l'AUC = 0,73 est plus proche de 1, par conséquent, le modèle de régression logistique a une bonne capacité prédictive.

Méthodologie

Construction du modèle prédictive

2. Arbre de décisions

Commençons par ajuster le modèle

```
R 4.1.1 ~ /projets/HRproject > #Ajuster le modèle
> set.seed(42)
> dt = rpart(vol_leave ~ role + perf + age + sex + area + salary, data = train, method = "class")
> dt
n= 7400

node), split, n, loss, yval, (yprob)
  * denotes terminal node

1) root 7400 2823 0 (0.6185135 0.3814865)
 2) area=Accounting,Finance,Marketing,other 5188 1544 0 (0.7023901 0.2976099) *
 3) area=Sales 2212 933 1 (0.4217902 0.5782098)
   6) sex=Male 1015 479 0 (0.5280788 0.4719212)
     12) perf< 2.5 682 281 0 (0.5879765 0.4120235) *
     13) perf>=2.5 333 135 1 (0.4054054 0.5945946) *
     7) sex=Female 1197 397 1 (0.3316625 0.6683375) *
> |
```


Méthodologie

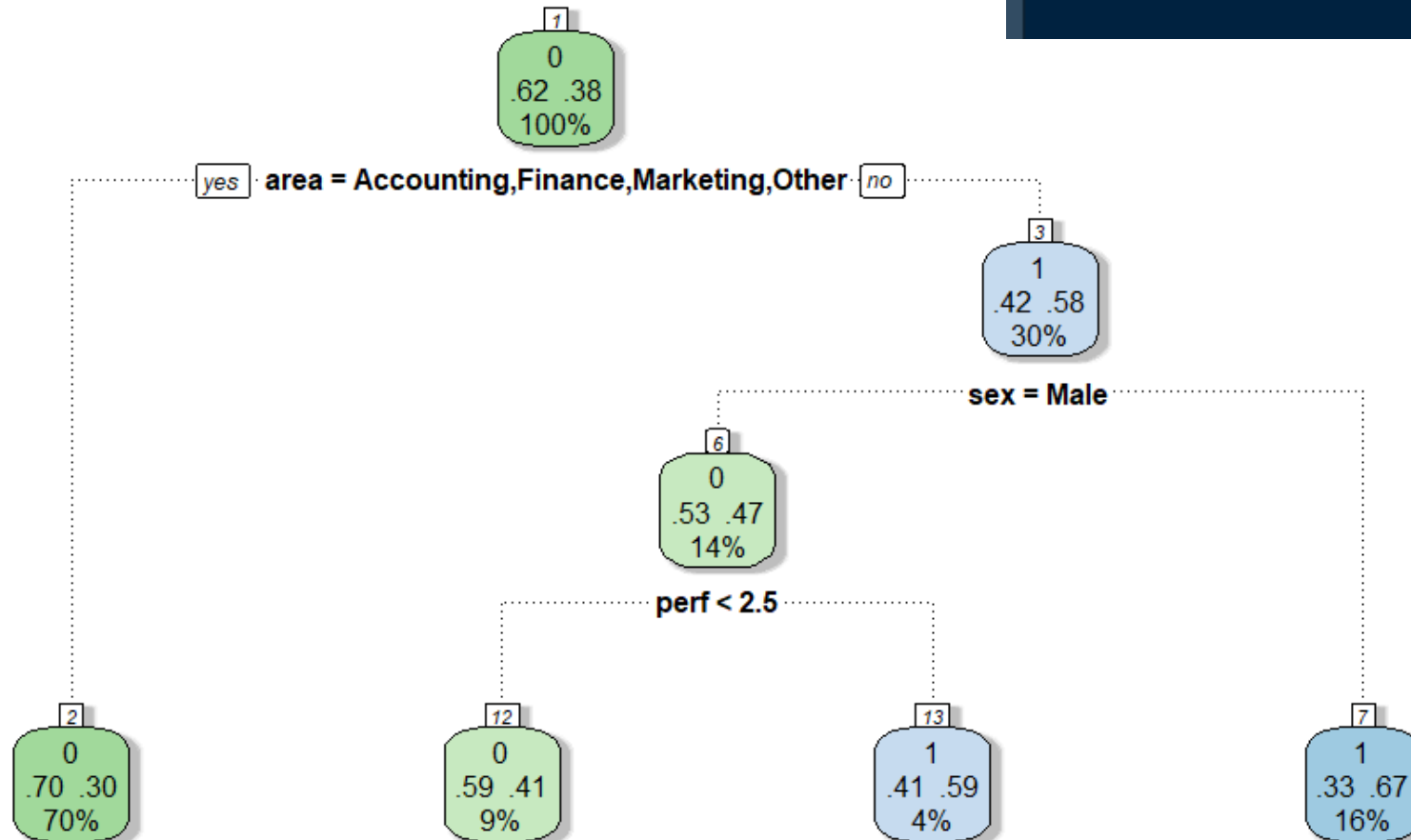
Construction du modèle prédictive

2. Arbre de décisions

Traçage de l'arbre de décision :

```
R 4.1.1 ~./projetR/HRproject/
> #Tracer l'arbre de décision
> library(rattle)
> par(mar = c(5,4,1,2))
> fancyRpartPlot(dt, sub = NULL, main = "Arbre de déciion de base")
> |
```

Arbre de déciion de base



Méthodologie

Construction du modèle prédictive

2. Arbre de décisions

Analyse

- Le premier nœud signifie la racine. Ici, **62 %** des personnes dans nos données d'entraînement ont 0 (Rester) pour la variable de réponse et **38 %** ont un 1 (Partir).
- En dessous, nous voyons notre premier nœud de décision. Dans le cas où nos employés se trouvent dans les régions Comptabilité, Finance, Marketing ou Autres, nous disons « oui » et prenons la branche de gauche.
- Nous avons une chance que la réponse soit « non » (c'est-à-dire qu'ils sont dans les ventes), alors nous prenons la bonne branche.
- Après la branche de gauche, nous voyons qu'elle se termine par un nœud solitaire pour tous ceux qui ne sont pas dans les ventes. Pour toutes ces personnes, la réponse la plus courante est « 0 » (Rester), avec **70 %** des employés qui resteront dans l'entreprise et seulement **30 %** dans ce seau quitteront l'entreprise.
- Les « **70 %** » signalés au bas du nœud nous indiquent que ce seul seau représente **70 %** de l'échantillon total que nous modélisons.

- En suivant la branche de droite, on constate que la réaction la plus connue est le « 1 » pour le salarié qui va quitter l'entreprise. De plus, le nœud nous fait également savoir que **42%** des employés de ce compartiment resteront tandis que **58%** partiront.
- En procédant avec la branche droite, si l'employé est un homme, nous disons « oui » et allons vers le côté gauche. Au cas où l'employé soit une femme, on va à droite. Pour les femmes, nous nous retrouvons dans un nœud de terminaison qui a une réponse dominante de 1 (**33% - Stay et 67% - Leave**).
- Ce nœud final représente **16%** de la population globale. Pour les hommes, nous descendons encore à la variable de performance. Si la performance est inférieure à **2,5**, nous allons à gauche, sinon nous allons à droite.
- Pour des performances inférieures à **2,5**, nous nous retrouvons dans un nœud de terminaison qui a une réponse dominante de 0 (**59% - Stay et 41% - Leave**). Ce nœud de terminaison représente **16%**.
- Pour des performances supérieures à **2,5**, nous nous retrouvons dans un nœud de terminaison qui a une réponse dominante de 1 (**33% - Stay et 67% - Leave**). Ce nœud de terminaison représente **4%**.

Méthodologie

Construction du modèle prédictive

2. Arbre de décisions

Maintenant, analyser la capacité prédictive du modèle à l'aide de la matrice de confusion :

```
> #Analyser la capacité prédictive du modèle en utilisant la matrice de confusion
> pred_dt = predict(dt, test, type = 'class')
> cm_dt = table(actual = test$vol_leave, prediction = pred_dt)
> cm_dt
      prediction
actual      0      1
0 2006    282
1  930    482
> |
```

Méthodologie

Construction du modèle prédictive

2. Arbre de décisions

Calculer la précision (Accuracy) de cet arbre de décision :

```
> #Calculer l'accuracy du modèle  
> accuracy = sum(diag(cm_dt))/sum(cm_dt)  
> accuracy  
[1] 0.6724324  
> |
```



La précision de ce model = 67.24%

CONCLUSION



La régression logistique est meilleure que l'arbre de décision pour prédire la variable de réponse de sortie pour prédire si l'employé restera dans l'entreprise ou quittera l'entreprise à l'avenir.