Projet\_R\_VF

Nasser Kassioui

2024-12-11

## Lecture Base

#Charger les librairies nécessaires  
  
library(data.table)

## Warning: le package 'data.table' a été compilé avec la version R 4.2.3

library(dplyr)

## Warning: le package 'dplyr' a été compilé avec la version R 4.2.3

##   
## Attachement du package : 'dplyr'

## Les objets suivants sont masqués depuis 'package:data.table':  
##   
## between, first, last

## Les objets suivants sont masqués depuis 'package:stats':  
##   
## filter, lag

## Les objets suivants sont masqués depuis 'package:base':  
##   
## intersect, setdiff, setequal, union

file\_path <- "C:\\Users\\kassi\\OneDrive\\Bureau\\M1 IREF\\Projet R\\emp\_offers\_fmtNV.tsv"  
  
emp\_offers\_fmtNV <- fread(file\_path, sep = "\t")

## Warning in fread(file\_path, sep = "\t"): Found and resolved improper quoting  
## out-of-sample. First healed line 4468: <<"CDD - DATA ANALYST Association,  
## Collectivité, Entreprise","",CDD,"",,"SQL, Excel, Statistique, équipe,  
## Proactif","Descriptif du posteL'équipe Données et Pilotage Economique de la  
## Direction Association Collectivités Entreprises de la MAIF recherche un.e Data  
## Analyst. Ce poste est à pourvoir sur Niort dès que possible pour une durée de  
## 12 Mois. Vous êtes passionné(e) par le monde de la Data et de l’Analyse ? Envie  
## de relever des défis de développement et de sécurisation de la tr>>. If the  
## fields are not quoted (e.g. field separator does not appear within any field),  
## try quote="" to avoid this warning.

# Afficher la structure et un aperçu des premières lignes  
unique(emp\_offers\_fmtNV)

## intitule\_poste,entreprise,type\_emploi,secteur,experience\_requise,competences\_requises,poste\_desc,salaire,departement  
## 1: Data Scientist H/F,Safran,CDI,"Industrie Aéronautique, Aérospatial",7,"Excel, modélisation, Deep Learning, Deep Learning, Git, Statistique, collecte de données, Prévision, Optimisation, Bases de donnée, équipe","Les missions du poste  
## 2:   
## 3: Dans le cadre de sa transformation digitale, Safran Nacelles a engagé une première phase de montée en compétence sur le data management, et la mise en place d'outils de collecte de données applicatives et opérationnelles. Safran Nacelles engage maintenant la phase d'analyse et d'exploitation de ces données, afin d'adresser des enjeux importants et variés tels que la maintenance prédictive, l'empreinte carbone, l'excellence opérationnelle et logistique, l'analyse de marché, et la prévision de la demande, entre autres.En tant que Data Scientist, vous aurez comme mission principale de contribuer au succès des projets Data Science sur ces périmètres, en participant activement aux phases d'analyse et de modélisation statistique des données.Vous serez également le référent en terme d'état de l'art des outils utilisés au sein de Safran Nacelles, en terme de data sciences et/ou d'IA. A ce titre, vous serez en charge de proposer et de contribuer à la réalisation de POC sur des sujets pertinents pour le business, afin de valider les perspectives de passage à l'échelle pour la technologie et le contexte considérés.Plus concrètement, votre mission vous mènera à :- Analyser le besoin business et le traduire en modèles mathématique/statistique, en utilisant, ou en concevant, et déployant des modèles d'analyse et de prédiction basés sur des méthodes statistiques (i.e. machine learning / deep learning), d'analyse de série temporelle,- Mettre à disposition du métier les analyses de données répondant à leur problématique, partager les résultats, capitaliser et favoriser l'adoption de la solution développée,- Identifier et mettre en place les optimisations sur les outils et les méthodes existants, en relation étroite avec les autres sociétés du Groupe,- Contribuer à la construction d'un socle technique et applicatif de data science, et à la diffusion de cette connaissance auprès des équipes business,- Cartographier et identifier les données pertinentes pour répondre aux besoins métiers et émettre des recommandations sur les bases de données à consolider ou à modifier,- Proposer les évolutions nécessaires dans la maintenance de la donnée afin d'en améliorer la qualité et la pertinence- Animer des formations internes sur la data science","",76  
## 4: Data Scientist H/F,Orano,CDI,"Secteur Energie, Environnement",1,"Git, Statistique, Optimisation, Rapport, Intelligence artificielle, équipe","Les missions du poste  
## 5: Au sein de la Direction de la Performance Orano, la Direction des Systèmes d'Information et du Digital (SI&D) a la mission de favoriser la transformation numérique pour l'ensemble des entités Orano.Cette transformation de l'ensemble des activités du groupe s'appuie, entre autres, sur un programme de gouvernance des données. L'équipe Data Office opère dans ce cadre des actions permettant d'accompagner les initiatives de valorisation du capital data. Elle s'appuie sur un ensemble d'outils et de technologies permettant d'acquérir, de connecter, de stocker de visualiser des données, et plus généralement de traiter ces données en y appliquant une gouvernance efficiente.Une feuille de route a été définie pour structurer les services fournis par le Data Office Corporate auprès des autres entités du groupe. Parmi les missions définies pour accompagner la mise en oeuvre des cas d'usage data/IA, un rôle de Data Scientist est positionné. Le périmètre à couvrir concerne l'ensemble des missions confiées au Data Office Corporate, auprès des Business Units / Directions centrales, et autres fonctions supports au travers des projets mis en oeuvre. Dans cette perspective vous serez partie prenante pour « faire parler les données » et de mettre à disposition des métiers, des informations élaborées sur la base des données disponibles.En tant que Data Scientist, vos principales missions seront les suivantes :· Exploiter, analyser et interpréter les données à l'aide de techniques statistiques· Fournir des rapports basés sur l'analyse quantitative et le raisonnement scientifique· Identifier, analyser et interpréter des tendances, proposer et exploiter des modèles mathématiques ou statistiques aidant à la prédiction de phénomènes ou comportements techniques· Elaborer des solutions d'aide à la décision ou d'optimisation des processus basés sur les technologies de l'Intelligence Artificielle, ou en accompagnant les métiers dans la mise en oeuvre de ces technologies d'un point de vue opérationnel· Conseiller et accompagner l'identification, la qualification et l'industrialisation des cas d'usage IA","",92  
## ---   
## 25181: 422 000 collaborateurs  
## 25182: er  
## 25183: employeur privé français  
## 25184: dans le monde  
## 25185: Basé·e à Paris, tu travailleras activement sur des problématiques data diverses afin d��aider l’équipe à résoudre les problèmes rencontrés, être plus efficace et itérer plus rapidement grâce à une donnée fiable.

# Convertir en data.table  
setDT(emp\_offers\_fmtNV)  
base\_emp\_new <- emp\_offers\_fmtNV #la nouvelle base -> "base\_emp"  
  
  
#taille  
  
length(base\_emp\_new$entreprise)

## [1] 0

## 2.4 Offre d’emploi

3720 entreprises => lenght

La nouvelle base => ‘base\_emp’ Je commence par traiter la partie “firm\_name”:

1. Harmonisation des noms d’entreprises:
   1. Mettre nom en Majuscule
   2. supprimer les suffixes superflus + espace superflus
   3. supprimer les NA collonnes ‘firme name’

## Code harmonisation:

# a) Mettre les noms en majuscules et supprimer les espaces superflus  
base\_emp\_new[, firm\_name := entreprise] #creation firm name

## Error in eval(jsub, SDenv, parent.frame()): objet 'entreprise' introuvable

base\_emp\_new[, firm\_name := toupper(trimws(firm\_name))]

## Error in is.factor(x): objet 'firm\_name' introuvable

unique(base\_emp\_new)

## intitule\_poste,entreprise,type\_emploi,secteur,experience\_requise,competences\_requises,poste\_desc,salaire,departement  
## 1: Data Scientist H/F,Safran,CDI,"Industrie Aéronautique, Aérospatial",7,"Excel, modélisation, Deep Learning, Deep Learning, Git, Statistique, collecte de données, Prévision, Optimisation, Bases de donnée, équipe","Les missions du poste  
## 2:   
## 3: Dans le cadre de sa transformation digitale, Safran Nacelles a engagé une première phase de montée en compétence sur le data management, et la mise en place d'outils de collecte de données applicatives et opérationnelles. Safran Nacelles engage maintenant la phase d'analyse et d'exploitation de ces données, afin d'adresser des enjeux importants et variés tels que la maintenance prédictive, l'empreinte carbone, l'excellence opérationnelle et logistique, l'analyse de marché, et la prévision de la demande, entre autres.En tant que Data Scientist, vous aurez comme mission principale de contribuer au succès des projets Data Science sur ces périmètres, en participant activement aux phases d'analyse et de modélisation statistique des données.Vous serez également le référent en terme d'état de l'art des outils utilisés au sein de Safran Nacelles, en terme de data sciences et/ou d'IA. A ce titre, vous serez en charge de proposer et de contribuer à la réalisation de POC sur des sujets pertinents pour le business, afin de valider les perspectives de passage à l'échelle pour la technologie et le contexte considérés.Plus concrètement, votre mission vous mènera à :- Analyser le besoin business et le traduire en modèles mathématique/statistique, en utilisant, ou en concevant, et déployant des modèles d'analyse et de prédiction basés sur des méthodes statistiques (i.e. machine learning / deep learning), d'analyse de série temporelle,- Mettre à disposition du métier les analyses de données répondant à leur problématique, partager les résultats, capitaliser et favoriser l'adoption de la solution développée,- Identifier et mettre en place les optimisations sur les outils et les méthodes existants, en relation étroite avec les autres sociétés du Groupe,- Contribuer à la construction d'un socle technique et applicatif de data science, et à la diffusion de cette connaissance auprès des équipes business,- Cartographier et identifier les données pertinentes pour répondre aux besoins métiers et émettre des recommandations sur les bases de données à consolider ou à modifier,- Proposer les évolutions nécessaires dans la maintenance de la donnée afin d'en améliorer la qualité et la pertinence- Animer des formations internes sur la data science","",76  
## 4: Data Scientist H/F,Orano,CDI,"Secteur Energie, Environnement",1,"Git, Statistique, Optimisation, Rapport, Intelligence artificielle, équipe","Les missions du poste  
## 5: Au sein de la Direction de la Performance Orano, la Direction des Systèmes d'Information et du Digital (SI&D) a la mission de favoriser la transformation numérique pour l'ensemble des entités Orano.Cette transformation de l'ensemble des activités du groupe s'appuie, entre autres, sur un programme de gouvernance des données. L'équipe Data Office opère dans ce cadre des actions permettant d'accompagner les initiatives de valorisation du capital data. Elle s'appuie sur un ensemble d'outils et de technologies permettant d'acquérir, de connecter, de stocker de visualiser des données, et plus généralement de traiter ces données en y appliquant une gouvernance efficiente.Une feuille de route a été définie pour structurer les services fournis par le Data Office Corporate auprès des autres entités du groupe. Parmi les missions définies pour accompagner la mise en oeuvre des cas d'usage data/IA, un rôle de Data Scientist est positionné. Le périmètre à couvrir concerne l'ensemble des missions confiées au Data Office Corporate, auprès des Business Units / Directions centrales, et autres fonctions supports au travers des projets mis en oeuvre. Dans cette perspective vous serez partie prenante pour « faire parler les données » et de mettre à disposition des métiers, des informations élaborées sur la base des données disponibles.En tant que Data Scientist, vos principales missions seront les suivantes :· Exploiter, analyser et interpréter les données à l'aide de techniques statistiques· Fournir des rapports basés sur l'analyse quantitative et le raisonnement scientifique· Identifier, analyser et interpréter des tendances, proposer et exploiter des modèles mathématiques ou statistiques aidant à la prédiction de phénomènes ou comportements techniques· Elaborer des solutions d'aide à la décision ou d'optimisation des processus basés sur les technologies de l'Intelligence Artificielle, ou en accompagnant les métiers dans la mise en oeuvre de ces technologies d'un point de vue opérationnel· Conseiller et accompagner l'identification, la qualification et l'industrialisation des cas d'usage IA","",92  
## ---   
## 25181: 422 000 collaborateurs  
## 25182: er  
## 25183: employeur privé français  
## 25184: dans le monde  
## 25185: Basé·e à Paris, tu travailleras activement sur des problématiques data diverses afin d��aider l’équipe à résoudre les problèmes rencontrés, être plus efficace et itérer plus rapidement grâce à une donnée fiable.

length(base\_emp\_new$firm\_name)

## [1] 0

# Supprimer les lignes où `firm\_name` est NA ou vide  
base\_emp\_new <- base\_emp\_new[!is.na(firm\_name) & firm\_name != ""] # on passe à 3032 entreprise

## Error: objet 'firm\_name' introuvable

get\_first\_word <- function(text) {  
 text <- trimws(text) # Supprimer les espaces superflus  
 words <- unlist(strsplit(text, "\\s+")) # Diviser en utilisant un ou plusieurs espaces  
 if (length(words) == 0) return("") # Retourner une chaîne vide si aucun mot n'est trouvé  
 return(words[1]) # Retourner le premier mot  
}  
base\_emp\_new[, firm\_name := sapply(firm\_name, get\_first\_word)]

## Error in lapply(X = X, FUN = FUN, ...): objet 'firm\_name' introuvable

#le nombre d'ent reste pareil=> pb nom composer   
  
#Apres la fonction identiefier des noms supprimer important pour réintegrer  
  
# Liste des noms importants à identifier  
noms\_importants <- c(  
 "CRÉDIT MUTUEL", "AGRICOLE", "RICARD", "DIOR", "THORNTON", "VERITAS",   
 "SEB", "MACLOU", "PSA", "ROULLIER", "ORTEC","DARTY", "DUBREUIL", "LAFAYETTE"  
)  
  
# Réintégrer les mots importants  
# Supprimer le premier mot pour créer `words\_removed`  
base\_emp\_new[, words\_removed := trimws(sub(paste0("^", firm\_name, "\\s\*"), "", firm\_name))]

## Error in is.factor(x): objet 'firm\_name' introuvable

# Réintégrer les mots supprimés dans 'first\_word' si importants  
base\_emp\_new[, firm\_name := ifelse(  
 words\_removed %in% noms\_importants,  
 paste(firm\_name, words\_removed),  
 firm\_name  
)]

## Error in words\_removed %in% noms\_importants: objet 'words\_removed' introuvable

unique(base\_emp\_new)

## intitule\_poste,entreprise,type\_emploi,secteur,experience\_requise,competences\_requises,poste\_desc,salaire,departement  
## 1: Data Scientist H/F,Safran,CDI,"Industrie Aéronautique, Aérospatial",7,"Excel, modélisation, Deep Learning, Deep Learning, Git, Statistique, collecte de données, Prévision, Optimisation, Bases de donnée, équipe","Les missions du poste  
## 2:   
## 3: Dans le cadre de sa transformation digitale, Safran Nacelles a engagé une première phase de montée en compétence sur le data management, et la mise en place d'outils de collecte de données applicatives et opérationnelles. Safran Nacelles engage maintenant la phase d'analyse et d'exploitation de ces données, afin d'adresser des enjeux importants et variés tels que la maintenance prédictive, l'empreinte carbone, l'excellence opérationnelle et logistique, l'analyse de marché, et la prévision de la demande, entre autres.En tant que Data Scientist, vous aurez comme mission principale de contribuer au succès des projets Data Science sur ces périmètres, en participant activement aux phases d'analyse et de modélisation statistique des données.Vous serez également le référent en terme d'état de l'art des outils utilisés au sein de Safran Nacelles, en terme de data sciences et/ou d'IA. A ce titre, vous serez en charge de proposer et de contribuer à la réalisation de POC sur des sujets pertinents pour le business, afin de valider les perspectives de passage à l'échelle pour la technologie et le contexte considérés.Plus concrètement, votre mission vous mènera à :- Analyser le besoin business et le traduire en modèles mathématique/statistique, en utilisant, ou en concevant, et déployant des modèles d'analyse et de prédiction basés sur des méthodes statistiques (i.e. machine learning / deep learning), d'analyse de série temporelle,- Mettre à disposition du métier les analyses de données répondant à leur problématique, partager les résultats, capitaliser et favoriser l'adoption de la solution développée,- Identifier et mettre en place les optimisations sur les outils et les méthodes existants, en relation étroite avec les autres sociétés du Groupe,- Contribuer à la construction d'un socle technique et applicatif de data science, et à la diffusion de cette connaissance auprès des équipes business,- Cartographier et identifier les données pertinentes pour répondre aux besoins métiers et émettre des recommandations sur les bases de données à consolider ou à modifier,- Proposer les évolutions nécessaires dans la maintenance de la donnée afin d'en améliorer la qualité et la pertinence- Animer des formations internes sur la data science","",76  
## 4: Data Scientist H/F,Orano,CDI,"Secteur Energie, Environnement",1,"Git, Statistique, Optimisation, Rapport, Intelligence artificielle, équipe","Les missions du poste  
## 5: Au sein de la Direction de la Performance Orano, la Direction des Systèmes d'Information et du Digital (SI&D) a la mission de favoriser la transformation numérique pour l'ensemble des entités Orano.Cette transformation de l'ensemble des activités du groupe s'appuie, entre autres, sur un programme de gouvernance des données. L'équipe Data Office opère dans ce cadre des actions permettant d'accompagner les initiatives de valorisation du capital data. Elle s'appuie sur un ensemble d'outils et de technologies permettant d'acquérir, de connecter, de stocker de visualiser des données, et plus généralement de traiter ces données en y appliquant une gouvernance efficiente.Une feuille de route a été définie pour structurer les services fournis par le Data Office Corporate auprès des autres entités du groupe. Parmi les missions définies pour accompagner la mise en oeuvre des cas d'usage data/IA, un rôle de Data Scientist est positionné. Le périmètre à couvrir concerne l'ensemble des missions confiées au Data Office Corporate, auprès des Business Units / Directions centrales, et autres fonctions supports au travers des projets mis en oeuvre. Dans cette perspective vous serez partie prenante pour « faire parler les données » et de mettre à disposition des métiers, des informations élaborées sur la base des données disponibles.En tant que Data Scientist, vos principales missions seront les suivantes :· Exploiter, analyser et interpréter les données à l'aide de techniques statistiques· Fournir des rapports basés sur l'analyse quantitative et le raisonnement scientifique· Identifier, analyser et interpréter des tendances, proposer et exploiter des modèles mathématiques ou statistiques aidant à la prédiction de phénomènes ou comportements techniques· Elaborer des solutions d'aide à la décision ou d'optimisation des processus basés sur les technologies de l'Intelligence Artificielle, ou en accompagnant les métiers dans la mise en oeuvre de ces technologies d'un point de vue opérationnel· Conseiller et accompagner l'identification, la qualification et l'industrialisation des cas d'usage IA","",92  
## ---   
## 25181: 422 000 collaborateurs  
## 25182: er  
## 25183: employeur privé français  
## 25184: dans le monde  
## 25185: Basé·e à Paris, tu travailleras activement sur des problématiques data diverses afin d��aider l’équipe à résoudre les problèmes rencontrés, être plus efficace et itérer plus rapidement grâce à une donnée fiable.

# Afficher les valeurs uniques des colonnes firm\_name et first\_word

’Firm\_name a bien été crée avec peut etre l’algo de suppresion a revoir

On passe maintenat sur ‘n\_offres’

On calcule le nombre de fois que chaque entreprise à publier utilisation de la fonction .N puis création d’une nouvelle colonne associé car à la fin on veut une entreprise = une ligne

# Calculer le nombre d'apparitions de chaque entreprise  
base\_emp\_new[, n\_offres := .N, by = firm\_name]

## Error in eval(bysub, parent.frame(), parent.frame()): objet 'firm\_name' introuvable

# Supprimer les doublons pour conserver une seule ligne par entreprise avec la   
  
# Trier les entreprises par nombre d'offres décroissant (facultatif)  
df\_summary <- df\_summary[order(-n\_offres)]

## Error in eval(expr, envir, enclos): objet 'df\_summary' introuvable

# Vérification des résultats  
sum(df\_summary$n\_offres)

## Error in eval(expr, envir, enclos): objet 'df\_summary' introuvable

unique(df\_summary)

## Error in unique(df\_summary): objet 'df\_summary' introuvable

La nouvelle colonone a bien été pris en compte on retoruve bine 3032 en faisant la somme des n\_offres ce qui montre que tous se passe bien

On passe maintentant sur la varibale experience requise avg\_req\_exp:

*Consigne*: Expérience. Au sein de toutes les offres d’une même entreprise, quelle est l’expérience moyenne demandée. Ignorez les valeurs manquantes. Si toutes les valeurs sont manquantes, alors cettevariable est manquante.

# Calculer l'expérience moyenne demandée (avg\_req\_exp) par entreprise  
base\_emp\_new[, avg\_req\_exp := ifelse(  
 all(is.na(experience\_requise)), # Si toutes les valeurs sont NA pour une entreprise  
 NA, # Attribuer NA  
 mean(experience\_requise, na.rm = TRUE) # Sinon, calculer la moyenne en ignorant les NA  
), by = firm\_name]

## Error in eval(bysub, parent.frame(), parent.frame()): objet 'firm\_name' introuvable

# Créer un dataframe avec une seule ligne par entreprise  
df\_summary <- unique(base\_emp\_new[, .(firm\_name, n\_offres, avg\_req\_exp)])

## Error in eval(jsub, SDenv, parent.frame()): objet 'firm\_name' introuvable

# Trier les résultats par nombre d'offres décroissant (facultatif)  
df\_summary <- df\_summary[order(-n\_offres)]

## Error in eval(expr, envir, enclos): objet 'df\_summary' introuvable

# Vérifier les résultats  
head(df\_summary)

## Error in head(df\_summary): objet 'df\_summary' introuvable

unique(df\_summary)

## Error in unique(df\_summary): objet 'df\_summary' introuvable

La variable ’avg\_req\_exp’a bien été crée les résultats sont concluents

Nous passons sur la variable #top\_skill\_req compétence

*consigne*: Les compétences demandées sont au format suivant, ex: “SQL, Spark, Git, database, équipe, Esprit Critique, Collaboration”. C’est à dire que chaque compétence est séparée par une virgule. Les entreprises proposent plusieurs offre d’emploi. La variable top\_skill\_req rapporte les compétences qui apparaissent le plus au sein de toutes les offres. Exemple Si une entreprise a deux offres d’emploi qui listent les compétences suivantes: • “SQL, Spark, Git, Database, équipe, Esprit Critique, Collaboration” • “SQL, Statistique, Power BI, Collaboration” Alors la variable top\_skill\_req = “SQL, Collaboration”. L’ordre des compétences au sein de la chaine de caractères n’a pas d’importance

Étapes pour calculer top\_skill\_req Séparer les compétences :

Transformez la chaîne de caractères en une liste de compétences. Utilisez la virgule comme séparateur. Compter les occurrences des compétences :

Regroupez par entreprise. Comptez la fréquence de chaque compétence au sein de toutes les offres pour cette entreprise. Identifier les compétences les plus fréquentes :

Sélectionnez les compétences ayant le maximum d’occurrences. Créer la colonne top\_skill\_req :

Combinez les compétences les plus fréquentes en une chaîne de caractères séparée par des virgules.

# Étape 1 : Séparer les compétences pour chaque entreprise  
# Convertir les compétences en liste longue (chaque ligne = une compétence unique)  
# Étape 1 : Séparer les compétences pour chaque entreprise  
base\_emp\_long <- base\_emp\_new[, .(  
 skill = unlist(strsplit(competences\_requises, ",\\s\*")) # Séparer les compétences par virgule et espace  
), by = .(firm\_name)]

## Error in eval(bysub, x, parent.frame()): objet 'firm\_name' introuvable

# Étape 2 : Compter les occurrences des compétences pour chaque entreprise  
skill\_counts <- base\_emp\_long[, .N, by = .(firm\_name, skill)] # Compter les compétences par entreprise

## Error in eval(expr, envir, enclos): objet 'base\_emp\_long' introuvable

# Étape 3 : Identifier les compétences les plus fréquentes par entreprise  
top\_skills <- skill\_counts[, .SD[N == max(N)], by = firm\_name] # Compétences avec la fréquence maximale

## Error in eval(expr, envir, enclos): objet 'skill\_counts' introuvable

# Étape 4 : Combiner les compétences les plus fréquentes pour chaque entreprise  
top\_skills\_combined <- top\_skills[, .(  
 top\_skill\_req = paste(skill, collapse = ", ") # Combiner les compétences  
), by = firm\_name]

## Error in eval(expr, envir, enclos): objet 'top\_skills' introuvable

# Étape 5 : Ajouter la colonne `top\_skill\_req` à `base\_emp`  
base\_emp\_new <- merge(base\_emp\_new, top\_skills\_combined, by = "firm\_name", all.x = TRUE)

## Error in is.data.table(y): objet 'top\_skills\_combined' introuvable

# Vérifier les résultats  
head(base\_emp\_new[, .(firm\_name, competences\_requises, top\_skill\_req)]) # Aperçu des colonnes importantes

## Error in eval(jsub, SDenv, parent.frame()): objet 'firm\_name' introuvable

head(base\_emp\_new)

## intitule\_poste,entreprise,type\_emploi,secteur,experience\_requise,competences\_requises,poste\_desc,salaire,departement  
## 1: Data Scientist H/F,Safran,CDI,"Industrie Aéronautique, Aérospatial",7,"Excel, modélisation, Deep Learning, Deep Learning, Git, Statistique, collecte de données, Prévision, Optimisation, Bases de donnée, équipe","Les missions du poste  
## 2:   
## 3:   
## 4: Dans le cadre de sa transformation digitale, Safran Nacelles a engagé une première phase de montée en compétence sur le data management, et la mise en place d'outils de collecte de données applicatives et opérationnelles. Safran Nacelles engage maintenant la phase d'analyse et d'exploitation de ces données, afin d'adresser des enjeux importants et variés tels que la maintenance prédictive, l'empreinte carbone, l'excellence opérationnelle et logistique, l'analyse de marché, et la prévision de la demande, entre autres.En tant que Data Scientist, vous aurez comme mission principale de contribuer au succès des projets Data Science sur ces périmètres, en participant activement aux phases d'analyse et de modélisation statistique des données.Vous serez également le référent en terme d'état de l'art des outils utilisés au sein de Safran Nacelles, en terme de data sciences et/ou d'IA. A ce titre, vous serez en charge de proposer et de contribuer à la réalisation de POC sur des sujets pertinents pour le business, afin de valider les perspectives de passage à l'échelle pour la technologie et le contexte considérés.Plus concrètement, votre mission vous mènera à :- Analyser le besoin business et le traduire en modèles mathématique/statistique, en utilisant, ou en concevant, et déployant des modèles d'analyse et de prédiction basés sur des méthodes statistiques (i.e. machine learning / deep learning), d'analyse de série temporelle,- Mettre à disposition du métier les analyses de données répondant à leur problématique, partager les résultats, capitaliser et favoriser l'adoption de la solution développée,- Identifier et mettre en place les optimisations sur les outils et les méthodes existants, en relation étroite avec les autres sociétés du Groupe,- Contribuer à la construction d'un socle technique et applicatif de data science, et à la diffusion de cette connaissance auprès des équipes business,- Cartographier et identifier les données pertinentes pour répondre aux besoins métiers et émettre des recommandations sur les bases de données à consolider ou à modifier,- Proposer les évolutions nécessaires dans la maintenance de la donnée afin d'en améliorer la qualité et la pertinence- Animer des formations internes sur la data science","",76  
## 5: Data Scientist H/F,Orano,CDI,"Secteur Energie, Environnement",1,"Git, Statistique, Optimisation, Rapport, Intelligence artificielle, équipe","Les missions du poste  
## 6:

unique(base\_emp\_new)

## intitule\_poste,entreprise,type\_emploi,secteur,experience\_requise,competences\_requises,poste\_desc,salaire,departement  
## 1: Data Scientist H/F,Safran,CDI,"Industrie Aéronautique, Aérospatial",7,"Excel, modélisation, Deep Learning, Deep Learning, Git, Statistique, collecte de données, Prévision, Optimisation, Bases de donnée, équipe","Les missions du poste  
## 2:   
## 3: Dans le cadre de sa transformation digitale, Safran Nacelles a engagé une première phase de montée en compétence sur le data management, et la mise en place d'outils de collecte de données applicatives et opérationnelles. Safran Nacelles engage maintenant la phase d'analyse et d'exploitation de ces données, afin d'adresser des enjeux importants et variés tels que la maintenance prédictive, l'empreinte carbone, l'excellence opérationnelle et logistique, l'analyse de marché, et la prévision de la demande, entre autres.En tant que Data Scientist, vous aurez comme mission principale de contribuer au succès des projets Data Science sur ces périmètres, en participant activement aux phases d'analyse et de modélisation statistique des données.Vous serez également le référent en terme d'état de l'art des outils utilisés au sein de Safran Nacelles, en terme de data sciences et/ou d'IA. A ce titre, vous serez en charge de proposer et de contribuer à la réalisation de POC sur des sujets pertinents pour le business, afin de valider les perspectives de passage à l'échelle pour la technologie et le contexte considérés.Plus concrètement, votre mission vous mènera à :- Analyser le besoin business et le traduire en modèles mathématique/statistique, en utilisant, ou en concevant, et déployant des modèles d'analyse et de prédiction basés sur des méthodes statistiques (i.e. machine learning / deep learning), d'analyse de série temporelle,- Mettre à disposition du métier les analyses de données répondant à leur problématique, partager les résultats, capitaliser et favoriser l'adoption de la solution développée,- Identifier et mettre en place les optimisations sur les outils et les méthodes existants, en relation étroite avec les autres sociétés du Groupe,- Contribuer à la construction d'un socle technique et applicatif de data science, et à la diffusion de cette connaissance auprès des équipes business,- Cartographier et identifier les données pertinentes pour répondre aux besoins métiers et émettre des recommandations sur les bases de données à consolider ou à modifier,- Proposer les évolutions nécessaires dans la maintenance de la donnée afin d'en améliorer la qualité et la pertinence- Animer des formations internes sur la data science","",76  
## 4: Data Scientist H/F,Orano,CDI,"Secteur Energie, Environnement",1,"Git, Statistique, Optimisation, Rapport, Intelligence artificielle, équipe","Les missions du poste  
## 5: Au sein de la Direction de la Performance Orano, la Direction des Systèmes d'Information et du Digital (SI&D) a la mission de favoriser la transformation numérique pour l'ensemble des entités Orano.Cette transformation de l'ensemble des activités du groupe s'appuie, entre autres, sur un programme de gouvernance des données. L'équipe Data Office opère dans ce cadre des actions permettant d'accompagner les initiatives de valorisation du capital data. Elle s'appuie sur un ensemble d'outils et de technologies permettant d'acquérir, de connecter, de stocker de visualiser des données, et plus généralement de traiter ces données en y appliquant une gouvernance efficiente.Une feuille de route a été définie pour structurer les services fournis par le Data Office Corporate auprès des autres entités du groupe. Parmi les missions définies pour accompagner la mise en oeuvre des cas d'usage data/IA, un rôle de Data Scientist est positionné. Le périmètre à couvrir concerne l'ensemble des missions confiées au Data Office Corporate, auprès des Business Units / Directions centrales, et autres fonctions supports au travers des projets mis en oeuvre. Dans cette perspective vous serez partie prenante pour « faire parler les données » et de mettre à disposition des métiers, des informations élaborées sur la base des données disponibles.En tant que Data Scientist, vos principales missions seront les suivantes :· Exploiter, analyser et interpréter les données à l'aide de techniques statistiques· Fournir des rapports basés sur l'analyse quantitative et le raisonnement scientifique· Identifier, analyser et interpréter des tendances, proposer et exploiter des modèles mathématiques ou statistiques aidant à la prédiction de phénomènes ou comportements techniques· Elaborer des solutions d'aide à la décision ou d'optimisation des processus basés sur les technologies de l'Intelligence Artificielle, ou en accompagnant les métiers dans la mise en oeuvre de ces technologies d'un point de vue opérationnel· Conseiller et accompagner l'identification, la qualification et l'industrialisation des cas d'usage IA","",92  
## ---   
## 25181: 422 000 collaborateurs  
## 25182: er  
## 25183: employeur privé français  
## 25184: dans le monde  
## 25185: Basé·e à Paris, tu travailleras activement sur des problématiques data diverses afin d��aider l’équipe à résoudre les problèmes rencontrés, être plus efficace et itérer plus rapidement grâce à une donnée fiable.

length(base\_emp\_new$top\_skills\_req)

## [1] 0

#### Étape 6 : Regrouper toutes les nouvelles variables dans un dataframe par entreprise ####  
df\_summary <- base\_emp\_new[, .(  
 n\_offres = .N, # Nombre total d'offres par entreprise  
 avg\_req\_exp = mean(experience\_requise, na.rm = TRUE), # Moyenne de l'expérience requise  
 top\_skill\_req = unique(top\_skill\_req) # Compétences les plus fréquentes  
), by = firm\_name]

## Error in eval(bysub, parent.frame(), parent.frame()): objet 'firm\_name' introuvable

# Trier les résultats par nombre d'offres décroissant (facultatif)  
df\_summary <- df\_summary[order(-n\_offres)]

## Error in eval(expr, envir, enclos): objet 'df\_summary' introuvable

# Vérifier les résultats  
head(df\_summary)

## Error in head(df\_summary): objet 'df\_summary' introuvable

La variable top\_skill\_req a bien été crée et tous semble fonctionner

On passe sur la variable *addr\_dept\_main* département de principale de l’entreprise pour cela on va chercher le nombre d’occurence de chaque département pusi prendre celui avec le plus de’occurence

#### Étape 1 : Compter les occurrences des départements pour chaque entreprise ####  
dept\_counts <- base\_emp\_new[, .N, by = .(firm\_name, departement)] # Compter les occurrences des départements

## Error in eval(bysub, x, parent.frame()): objet 'firm\_name' introuvable

#### Étape 2 : Identifier le département avec le plus d'occurrences pour chaque entreprise ####  
main\_dept <- dept\_counts[, .SD[N == max(N)], by = firm\_name] # Identifier le ou les départements principaux

## Error in eval(expr, envir, enclos): objet 'dept\_counts' introuvable

#### Étape 3 : Combiner les départements principaux (en cas d'égalité) ####  
main\_dept\_combined <- main\_dept[, .(  
 addr\_dept\_main = paste(departement, collapse = ", ") # Combiner les départements en cas d'égalité  
), by = firm\_name]

## Error in eval(expr, envir, enclos): objet 'main\_dept' introuvable

#### Étape 4 : Ajouter la colonne addr\_dept\_main à la base principale ####  
df\_summary <- base\_emp\_new[, .(  
 n\_offres = .N, # Nombre total d'offres  
 avg\_req\_exp = mean(experience\_requise, na.rm = TRUE) # Moyenne de l'expérience requise  
), by = firm\_name]

## Error in eval(bysub, parent.frame(), parent.frame()): objet 'firm\_name' introuvable

# Ajouter la variable addr\_dept\_main  
df\_summary <- merge(df\_summary, main\_dept\_combined, by = "firm\_name", all.x = TRUE)

## Error in merge(df\_summary, main\_dept\_combined, by = "firm\_name", all.x = TRUE): objet 'df\_summary' introuvable

#### Étape 5 : Trier les résultats (facultatif) ####  
df\_summary <- df\_summary[order(-n\_offres)] # Trier par nombre d'offres décroissant

## Error in eval(expr, envir, enclos): objet 'df\_summary' introuvable

#### Étape 6 : Vérifier les résultats ####  
# Aperçu du dataframe avec une ligne par entreprise  
head(df\_summary)

## Error in head(df\_summary): objet 'df\_summary' introuvable

Départements à bien été définis aucune valeur n’est perdu

On passe maintenant par la variable *avg\_wage* qui est le salaire annuel moyen des offres proposées, doit être une variable numérique sachant que la varibale s’appellle ‘salaire’ dans la base.

*consigne du prof*: Salaire. Le salaire est rempli par les entreprises dans un format libre. Ex: “Salaire : 55K à 60K€” ou “50 000 - 63 000 EUR par an”. Il faudra convertir au format numérique. Les étapes sont les suivantes: 1. transformer la chaine de caractère pour qu’elle affiche uniquement un nombre 2. convertir la chaine de caractère en numérique A noter: dans les exemple du dessus il y a une fourchette de salaire. Vous pouvez vous contenter de ne prendre qu’un des deux nombres. Calculer la moyenne est mieux mais vous n’êtes pas obligé de le faire pour avoir le maximum de points.

#### Étape 1 : Extraire et nettoyer les valeurs de salaire ####  
  
# Fonction pour extraire la moyenne de la fourchette de salaire  
extract\_salary <- function(salaire) {  
   
   
 # Remplacer les séparateurs de milliers (espaces, points, etc.) par rien  
 salaire\_clean <- gsub("[^0-9\\-]", "", salaire) # Garder uniquement les chiffres et les tirets  
 salaire\_split <- unlist(strsplit(salaire\_clean, "-")) # Séparer la fourchette (ex: "50000-60000")  
 salaire\_numeric <- as.numeric(salaire\_split) # Convertir en numérique  
   
 # Si une fourchette est présente, calculer la moyenne  
 if (length(salaire\_numeric) == 2) {  
 return(mean(salaire\_numeric, na.rm = TRUE))  
 } else if (length(salaire\_numeric) == 1) {  
 return(salaire\_numeric) # Retourner le salaire unique  
 } else {  
 return(NA) # Si aucune valeur valide, retourner NA  
 }  
}  
  
# Appliquer la fonction à la colonne `salaire`  
base\_emp\_new[, salaire\_clean := sapply(salaire, extract\_salary)]

## Error in lapply(X = X, FUN = FUN, ...): objet 'salaire' introuvable

#### Étape 2 : Calculer le salaire annuel moyen par entreprise ####  
avg\_wage\_by\_firm <- base\_emp\_new[, .(  
 avg\_wage = mean(salaire\_clean, na.rm = TRUE) # Calculer la moyenne des salaires par entreprise  
), by = firm\_name]

## Error in eval(bysub, parent.frame(), parent.frame()): objet 'firm\_name' introuvable

#### Étape 3 : Créer un dataframe avec une ligne par entreprise ####  
  
  
# Ajouter la colonne `avg\_wage`  
df\_summary <- merge(df\_summary, avg\_wage\_by\_firm, by = "firm\_name", all.x = TRUE)

## Error in merge(df\_summary, avg\_wage\_by\_firm, by = "firm\_name", all.x = TRUE): objet 'df\_summary' introuvable

#### Étape 4 : Vérifier les résultats ####  
# Exemple de visualisation des colonnes importantes  
head(df\_summary)

## Error in head(df\_summary): objet 'df\_summary' introuvable

#### Étape 5 : Trier par nombre d'offres (facultatif) ####  
df\_summary <- df\_summary[order(-n\_offres)]

## Error in eval(expr, envir, enclos): objet 'df\_summary' introuvable

#### Étape 6 : Vérification finale ####  
# Résultat attendu avec une ligne par entreprise  
head(df\_summary)

## Error in head(df\_summary): objet 'df\_summary' introuvable

Cela ne ressort pas la bonne sortie valeur bizzare à revoir

dernière varibale *sector\_main* qui est le secteur principal d’activité le nombre d’occurence le plus haut la varibale s’appelle *secteur* dans la base si toutes les offre sont NA on laisse NA

#### Étape 1 : Compter les occurrences des secteurs pour chaque entreprise ####  
sector\_counts <- base\_emp\_new[, .N, by = .(firm\_name, secteur)] # Compter les occurrences des secteurs

## Error in eval(bysub, x, parent.frame()): objet 'firm\_name' introuvable

#### Étape 2 : Identifier le secteur principal avec le plus d'occurrences ####  
sector\_main <- base\_emp\_new[, .(  
 sector\_main = if (all(is.na(secteur))) {  
 as.character(NA) # Retourner NA en tant que caractère  
 } else {  
 secteur[which.max(tabulate(match(secteur, unique(secteur))))]  
 }  
), by = firm\_name]

## Error in eval(bysub, parent.frame(), parent.frame()): objet 'firm\_name' introuvable

#### Étape 3 : Ajouter la colonne `sector\_main` à `base\_emp\_new` ####  
base\_emp\_new <- merge(base\_emp\_new, sector\_main, by = "firm\_name", all.x = TRUE)

## Error in is.data.table(y): objet 'sector\_main' introuvable

unique(base\_emp\_new)

## intitule\_poste,entreprise,type\_emploi,secteur,experience\_requise,competences\_requises,poste\_desc,salaire,departement  
## 1: Data Scientist H/F,Safran,CDI,"Industrie Aéronautique, Aérospatial",7,"Excel, modélisation, Deep Learning, Deep Learning, Git, Statistique, collecte de données, Prévision, Optimisation, Bases de donnée, équipe","Les missions du poste  
## 2:   
## 3: Dans le cadre de sa transformation digitale, Safran Nacelles a engagé une première phase de montée en compétence sur le data management, et la mise en place d'outils de collecte de données applicatives et opérationnelles. Safran Nacelles engage maintenant la phase d'analyse et d'exploitation de ces données, afin d'adresser des enjeux importants et variés tels que la maintenance prédictive, l'empreinte carbone, l'excellence opérationnelle et logistique, l'analyse de marché, et la prévision de la demande, entre autres.En tant que Data Scientist, vous aurez comme mission principale de contribuer au succès des projets Data Science sur ces périmètres, en participant activement aux phases d'analyse et de modélisation statistique des données.Vous serez également le référent en terme d'état de l'art des outils utilisés au sein de Safran Nacelles, en terme de data sciences et/ou d'IA. A ce titre, vous serez en charge de proposer et de contribuer à la réalisation de POC sur des sujets pertinents pour le business, afin de valider les perspectives de passage à l'échelle pour la technologie et le contexte considérés.Plus concrètement, votre mission vous mènera à :- Analyser le besoin business et le traduire en modèles mathématique/statistique, en utilisant, ou en concevant, et déployant des modèles d'analyse et de prédiction basés sur des méthodes statistiques (i.e. machine learning / deep learning), d'analyse de série temporelle,- Mettre à disposition du métier les analyses de données répondant à leur problématique, partager les résultats, capitaliser et favoriser l'adoption de la solution développée,- Identifier et mettre en place les optimisations sur les outils et les méthodes existants, en relation étroite avec les autres sociétés du Groupe,- Contribuer à la construction d'un socle technique et applicatif de data science, et à la diffusion de cette connaissance auprès des équipes business,- Cartographier et identifier les données pertinentes pour répondre aux besoins métiers et émettre des recommandations sur les bases de données à consolider ou à modifier,- Proposer les évolutions nécessaires dans la maintenance de la donnée afin d'en améliorer la qualité et la pertinence- Animer des formations internes sur la data science","",76  
## 4: Data Scientist H/F,Orano,CDI,"Secteur Energie, Environnement",1,"Git, Statistique, Optimisation, Rapport, Intelligence artificielle, équipe","Les missions du poste  
## 5: Au sein de la Direction de la Performance Orano, la Direction des Systèmes d'Information et du Digital (SI&D) a la mission de favoriser la transformation numérique pour l'ensemble des entités Orano.Cette transformation de l'ensemble des activités du groupe s'appuie, entre autres, sur un programme de gouvernance des données. L'équipe Data Office opère dans ce cadre des actions permettant d'accompagner les initiatives de valorisation du capital data. Elle s'appuie sur un ensemble d'outils et de technologies permettant d'acquérir, de connecter, de stocker de visualiser des données, et plus généralement de traiter ces données en y appliquant une gouvernance efficiente.Une feuille de route a été définie pour structurer les services fournis par le Data Office Corporate auprès des autres entités du groupe. Parmi les missions définies pour accompagner la mise en oeuvre des cas d'usage data/IA, un rôle de Data Scientist est positionné. Le périmètre à couvrir concerne l'ensemble des missions confiées au Data Office Corporate, auprès des Business Units / Directions centrales, et autres fonctions supports au travers des projets mis en oeuvre. Dans cette perspective vous serez partie prenante pour « faire parler les données » et de mettre à disposition des métiers, des informations élaborées sur la base des données disponibles.En tant que Data Scientist, vos principales missions seront les suivantes :· Exploiter, analyser et interpréter les données à l'aide de techniques statistiques· Fournir des rapports basés sur l'analyse quantitative et le raisonnement scientifique· Identifier, analyser et interpréter des tendances, proposer et exploiter des modèles mathématiques ou statistiques aidant à la prédiction de phénomènes ou comportements techniques· Elaborer des solutions d'aide à la décision ou d'optimisation des processus basés sur les technologies de l'Intelligence Artificielle, ou en accompagnant les métiers dans la mise en oeuvre de ces technologies d'un point de vue opérationnel· Conseiller et accompagner l'identification, la qualification et l'industrialisation des cas d'usage IA","",92  
## ---   
## 25181: 422 000 collaborateurs  
## 25182: er  
## 25183: employeur privé français  
## 25184: dans le monde  
## 25185: Basé·e à Paris, tu travailleras activement sur des problématiques data diverses afin d��aider l’équipe à résoudre les problèmes rencontrés, être plus efficace et itérer plus rapidement grâce à une donnée fiable.

#### Étape 4 : Créer un dataframe final avec une ligne par entreprise ####  
df\_summary <- base\_emp\_new[, .(  
 n\_offres = .N, # Nombre total d'offres  
 avg\_req\_exp = mean(experience\_requise, na.rm = TRUE), # Moyenne de l'expérience requise  
 addr\_dept\_main = unique(addr\_dept\_main), # Département principal  
 avg\_wage = mean(salaire\_clean, na.rm = TRUE), # Salaire moyen  
 sector\_main = unique(sector\_main) # Secteur principal  
), by = firm\_name]

## Error in eval(bysub, parent.frame(), parent.frame()): objet 'firm\_name' introuvable

#### Étape 5 : Trier par nombre d'offres (facultatif) ####  
df\_summary <- df\_summary[order(-n\_offres)]

## Error in eval(expr, envir, enclos): objet 'df\_summary' introuvable

#### Étape 6 : Vérifier les résultats ####  
# Visualisation des colonnes importantes  
head(df\_summary)

## Error in head(df\_summary): objet 'df\_summary' introuvable