Rapport du projet R

Ce travail est réalisé par : Khlifi Amen

Explication de notre dataset :

Cet ensemble de données contient les réponses à des sondages réalisés auprès de personnes du secteur technologique sur leur santé mentale. Les questions portent notamment sur le traitement, les ressources offertes en milieu de travail et l'attitude face à la discussion de la santé mentale au travail.

En analysant cet ensemble de données, nous pouvons mieux comprendre la fréquence des problèmes de santé mentale chez les personnes qui travaillent dans le secteur technologique - et quels types de ressources elles utilisent pour trouver de l'aide - afin de créer un environnement de travail plus sain pour tous.

Cet ensemble de données suit des mesures clés telles que l'âge, le sexe et le pays pour déterminer la prévalence globale, ainsi que les réponses concernant l'accès des employés aux options de soins; la question de savoir si la

santé mentale ou les maladies physiques sont prises aussi au sérieux par les employeurs; la protection de l'anonymat pour la recherche d'aide; et la façon dont les collègues peuvent percevoir les personnes aux prises avec des problèmes de santé mentale tels que la dépression ou l'anxiété.

Objectif de l’analyse de cette dataset :

- Comprendre la prévalence des problèmes de santé mentale dans le secteur technologique.
- Identifier les ressources utilisées par les employés pour obtenir de l'aide.
- Améliorer l'environnement de travail pour la santé mentale.

-->L'objectif est de créer un environnement de travail plus sain en promouvant la santé mentale et en luttant contre la stigmatisation.

variables de mon ensemble de données :

1. **Timestamp**: Horodatage de la soumission de l'enquête. (DateTime)

2. **Age**: Âge du répondant. (Numérique)

3. **Gender**: Genre du répondant. (Catégorique)

4. **Country**: Pays de résidence du répondant. (Catégorique)

5. **state**: État américain de résidence du répondant. (Catégorique)

6. **self_employed**: Indique si le répondant est travailleur indépendant. (Catégorique)

7. **family_history**: Indique si le répondant a des antécédents familiaux de problèmes de santé mentale. (Catégorique)

8. **treatment**: Indique si le répondant a recherché un traitement pour un problème de santé mentale. (Catégorique)

9. **work_interfere**: Indique dans quelle mesure le travail est perturbé en raison de problèmes de santé mentale. (Catégorique)

10. **no_employees**: Nombre d'employés dans l'entreprise du répondant. (Numérique)

11. **remote_work**: Indique si le répondant travaille à distance. (Catégorique)

12. **tech_company**: Indique si le répondant travaille dans une entreprise technologique. (Catégorique)

13. **benefits**: Indique si le lieu de travail du répondant propose des avantages en matière de santé mentale. (Catégorique)

14. **care_options**: Indique si le lieu de travail du répondant propose des options de soins de santé mentale. (Catégorique)

15. **wellness_program**: Indique si le lieu de travail du répondant propose un programme de bien-être. (Catégorique)

16. **seek_help**: Indique si le répondant a cherché de l'aide pour un problème de santé mentale. (Catégorique)

17. **anonymity**: Indique si le lieu de travail du répondant permet l'anonymat lors de la recherche d'aide pour un problème de santé mentale. (Catégorique)

18. **leave**: Indique si le lieu de travail du répondant propose un congé pour les problèmes de santé mentale. (Catégorique)

19. **mental_health_consequence**: Indique si le répondant a subi des conséquences négatives en raison de la discussion de problèmes de santé mentale au travail. (Catégorique)

20. **phys_health_consequence**: Indique si le répondant a subi des conséquences négatives en raison de la discussion de problèmes de santé physique au travail. (Catégorique)

21. **coworkers**: Indique si le répondant a discuté de problèmes de santé mentale avec des collègues. (Catégorique)

22. **supervisor**: Indique si le répondant a discuté de problèmes de santé mentale avec son superviseur. (Catégorique)

Commençons l’analyse :

-data = read.csv ("survey.csv",sep=";",dec=".",header=TRUE,stringsAsFactors=TRUE)

Cette commande importe les données à partir d'un fichier CSV dans RStudio et les stocke dans un objet appelé **data**.

-Dim(data)

```
> dim(data)
[1] 1250   22
>
```

1250 lignes et 22 colonnes (variables d’observations)

-Names(data)

```
> names(data)
 [1] "index"                  "Age"                    "Gender"                 "Country"
 [5] "state"                  "self_employed"          "family_history"         "treatment"
 [9] "work_interfere"         "remote_work"            "tech_company"           "benefits"
[13] "care_options"           "leave"                  "mental_health_consequence" "phys_health_consequence"
[17] "coworkers"              "supervisor"             "mental_health_interview"   "phys_health_interview"
[21] "mental_vs_physical"     "obs_consequence"
>
```

Affiche les noms des colonnes de ma dataset (les variables d’observations)

-Str(data)

-View(data)

idex	Age	Gender	Country	state	self_employed	family_history	treatment	work_interfere	remote_work	tech_company	benefits	ca
0	37	Female	United States	IL	NA	No	Yes	Often	No	Yes	Yes	N
1	44	M	United States	IN	NA	No	No	Rarely	No	No	Don't know	N
2	32	Male	Canada	NA	NA	No	No	Rarely	No	Yes	No	N
3	31	Male	United Kingdom	NA	NA	Yes	Yes	Often	No	Yes	No	Y
4	31	Male	United States	TX	NA	No	No	Never	Yes	Yes	Yes	N
5	33	Male	United States	TN	NA	Yes	No	Sometimes	No	Yes	Yes	N
6	35	Female	United States	MI	NA	Yes	Yes	Sometimes	Yes	Yes	No	N
7	39	M	Canada	NA	NA	No	No	Never	Yes	Yes	No	Y

Data[1,3] : afficher la 1ere ligne du 3eme colonne

-Head(data)

Cela affiche les premières lignes de l'ensemble de données pour avoir un aperçu des variables et de leur format.

```
> head(data)
  index Age Gender        Country state self_employed family_history treatment work_interfere remote_work tech_company
1     0  37 Female  United States    IL          <NA>             No       Yes          Often          No          Yes
2     1  44      M  United States    IN          <NA>             No        No         Rarely          No           No
3     2  32   Male         Canada  <NA>          <NA>             No        No         Rarely          No          Yes
4     3  31   Male United Kingdom  <NA>          <NA>            Yes       Yes          Often          No          Yes
5     4  31   Male  United States    TX          <NA>             No        No          Never         Yes          Yes
6     5  33   Male  United States    TN          <NA>            Yes        No      Sometimes          No          Yes
    benefits care_options              leave mental_health_consequence phys_health_consequence    coworkers supervisor
1        Yes     Not sure      Somewhat easy                         No                       No Some of them        Yes
2 Don't know           No         Don't know                      Maybe                       No           No         No
3         No           No Somewhat difficult                         No                       No          Yes        Yes
4         No          Yes Somewhat difficult                        Yes                      Yes Some of them         No
5        Yes           No         Don't know                         No                       No Some of them        Yes
6        Yes     Not sure         Don't know                         No                       No          Yes        Yes
  mental_health_interview phys_health_interview mental_vs_physical obs_consequence
1                      No                  Maybe                Yes              No
2                      No                     No         Don't know              No
3                     Yes                    Yes                 No              No
4                   Maybe                  Maybe                 No             Yes
5                     Yes                    Yes         Don't know              No
6                      No                  Maybe         Don't know              No
> |
```

-tail(data)

Cela affiche les dernières lignes de l'ensemble de données pour avoir un aperçu des variables et de leur format.

-Data$Age:

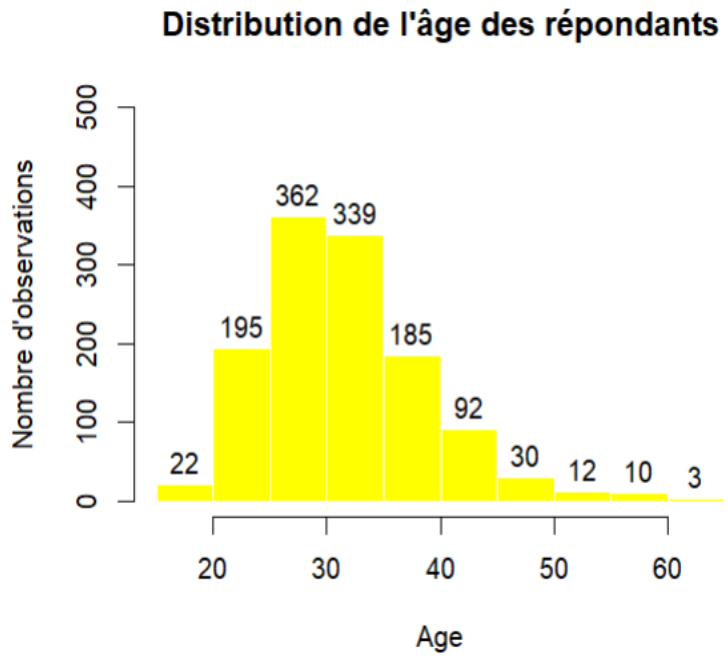
La commande **Data$Age** permet d'accéder à la colonne "Age" du dataframe "Data", fournissant ainsi un aperçu des âges des individus dans l'ensemble de données.

-Summary(data):

La commande **summary** fournit un résumé statistique de la variable d'âge, tandis que la commande **hist** crée un histogramme pour visualiser la distribution des âges des répondants.

```
> summary(data$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   27.00   31.00   32.04   36.00   65.00
> hist (data$Age, col= "yellow", border= "white",ylim = c(0,500), labels=TRUE, xlab="Age", ylab= "Nombre d'observations",mai
n= "Distribution de l'âge des répondants")
```

-hist (data$Age, col= "yellow", border= "white",ylim = c(0,500), labels=TRUE, xlab="Age", ylab= "Nombre d'observations",main= "Distribution de l'âge des répondants")



-table (data$work_interfere): génère un tableau de fréquences des différentes modalités de la variable "work_interfere" dans le dataframe "dfc", fournissant ainsi une vue détaillée de la répartition des réponses concernant l'impact du travail sur les interférences liées à la santé mentale.

-glimpse(data):

La fonction **glimpse(data)** fournit un aperçu concis des premières observations du dataframe "data", montrant le type de données et les premières valeurs de chaque colonne, offrant ainsi un résumé rapide de la structure et du contenu du dataframe.

-class(data$age): renvoie la classe de la variable "age" dans le dataframe "data", indiquant ainsi le type de données de cette variable.

-unique(data$age):

renvoie les valeurs uniques de la variable "age" dans le dataframe "data", fournissant ainsi une liste des différentes valeurs d'âge présentes dans les données, sans répétition.

```
> unique(data$Age)
 [1]          37          44          32          31          33          35          39          42          23
[10]          29          36          27          46          41          34          30          40          38
[19]          50          24          18          28          26          22          19          25          45
[28]          21         -29          43          56          60          54         329          55 99999999999
[37]          48          20          57          58          47          62          51          65          49
[46]       -1726           5          53          61           8          11          -1          72
> |
```

- data <- data[data$Age <= 70 & data$Age >= 15,]: filtre les données dans le dataframe "data" pour inclure uniquement les observations où l'âge est compris entre 15 et 70 ans

```
> unique(data$Age)
 [1] 37 44 32 31 33 35 39 42 23 29 36 27 46 41 34 30 40 38 50 24 18 28 26 22 19 25 45 21 43 56 60 54 55 48 20 57 58 47
[39] 62 51 65 49 53 61
>
```

Cleaning :

unique_values <- lapply(data, unique)
> unique_values :

crée une liste nommée "unique_values" où chaque élément correspond aux valeurs uniques de chaque variable dans le dataframe "data". Cela permet d'explorer les valeurs uniques de chaque variable dans le dataframe.

Changer les espaces vides par NA :

data[data == ""] <- NA

Verifier l'existence des valeurs NA :

complete.cases(data)

```
> complete.cases(data)
  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
 [20] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE
 [39] FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE  TRUE FALSE
 [58]  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE FALSE
 [77] FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE
 [96]  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[115] FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
[134] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE
[153]  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
[172] FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE
[191]  TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE
[210] FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
[229] FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE
[248]  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE
[267] FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE
[286]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[305] FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
[324]  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE
```

Les noms des colonnes contenant NA comme valeurs :

names_na <- colnames(data)[apply(data, 2, anyNA)]

```
> names_na <- colnames(data)[apply(data, 2, anyNA)]
> names_na
[1] "state"          "self_employed"  "work_interfere"
```

Supprimer les colonnes qui ne sont pas nécessaires dans nos analyses :

data <- subset(data, select = -comments)

data <- subset(data, select = -no_employees)

data <- subset(data, select = -seek_help)

data <- subset(data, select = -wellness_program)

data <- subset(data, select = -anonymity)

data <- subset(data, select = -Timestamp)

dfc<- na.omit(data):

crée un nouveau dataframe appelé "dfc" en supprimant toutes les lignes contenant des valeurs manquantes du dataframe "data". Cela permet de nettoyer les données en éliminant les observations avec des valeurs manquantes.

na <- any(is.na(dfc$state)):

Verifier ci la colonne state à pour valeur na

```
> na <- any(is.na(dfc$state))
> na
[1] FALSE
>
```

Rearanger les colonnes :

1) Gender

Avant le cleaninng :

```
> unique(data$Gender)
 [1] Female                          M
 [3] Male                            male
 [5] female                          m
 [7] Male-ish                        maile
 [9] Trans-female                    Cis Female
[11] F                               something kinda male?
[13] Cis Male                        Woman
[15] f                               Mal
[17] Male (CIS)                      queer/she/they
[19] non-binary                      Femake
[21] woman                           Make
[23] Nah                             Enby
[25] fluid                           Genderqueer
[27] Female                          Androgyne
[29] Agender                         cis-female/femme
[31] Guy (-ish) ^_^                  male leaning androgynous
[33] Male                            Man
[35] Trans woman                     msle
[37] Neuter                          Female (trans)
```

Remplacer avec female and male :

```
dfc$Gender<-replace(dfc$Gender,dfc$Gender=="Female","female")
```

Supprimer les autres lignes ayant des valeurs or (female , male , non_binary):

```
dfc <- dfc[dfc$Gender %in% c("female", "male", "non-binary"), ]
```

dfc$Gender <- tolower(dfc$Gender)

Supprimer un dataframe que je n'utilise plus :

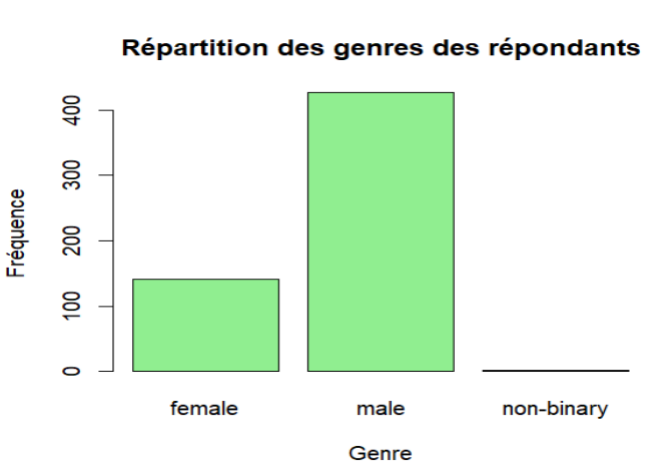
rm(dc)

After cleaning dfc$gender:

```
> unique(dfc$Gender)
[1] "male"       "female"     "non-binary"
>
```

Analyses apres le cleaning :

```
> table(dfc$Gender)

    female       male non-binary
       141        427          1
> barplot(table(dfc$Gender), main = "Répartition des genres des répondants", xlab = "Genre", ylab = "Fréquence", col = "ligh
tgreen")
>
```



La commande **`table`** compte le nombre de répondants par genre, tandis que **`barplot`** crée un graphique à barres pour visualiser la répartition des genres.

Avant le data cleaning des lignes qui contiennet NA comme valeurs dans quelques colonnes :

```
table(data$Country)
```

```
        Australia           Austria      Bahamas, The           Belgium Bosnia and Herzegovina 
               21                 3                 0                 6                 1 
           Brazil          Bulgaria            Canada             China          Colombia 
                6                 4                72                 1                 2 
       Costa Rica           Croatia    Czech Republic           Denmark           Finland 
                1                 2                 1                 2                 3 
           France           Georgia           Germany            Greece           Hungary 
               13                 1                45                 2                 1 
            India           Ireland            Israel             Italy             Japan 
               10                27                 5                 7                 1 
           Latvia            Mexico           Moldova       Netherlands       New Zealand 
                1                 3                 1                27                 8 
          Nigeria            Norway       Philippines            Poland          Portugal 
                1                 1                 1                 7                 2 
          Romania            Russia         Singapore          Slovenia      South Africa 
                1                 3                 4                 1                 6 
            Spain            Sweden       Switzerland          Thailand    United Kingdom 
                1                 7                 7                 1               184 
    United States           Uruguay          Zimbabwe 
              745                 1                 0 
```

Aprés :

```
> table(dfc$Country)

          Australia             Austria       Bahamas, The             Belgium Bosnia and Herzegovina
                  0                   0                   0                   0                   0
             Brazil            Bulgaria              Canada               China            Colombia
                  0                   1                   0                   0                   0
         Costa Rica             Croatia      Czech Republic             Denmark             Finland
                  0                   0                   0                   0                   0
             France             Georgia             Germany              Greece             Hungary
                  0                   0                   0                   0                   0
              India             Ireland              Israel               Italy               Japan
                  0                   0                   1                   0                   0
             Latvia              Mexico             Moldova         Netherlands         New Zealand
                  0                   0                   0                   0                   0
            Nigeria              Norway         Philippines              Poland            Portugal
                  0                   0                   0                   0                   0
            Romania              Russia           Singapore            Slovenia        South Africa
                  0                   0                   0                   0                   0
              Spain              Sweden         Switzerland            Thailand      United Kingdom
                  0                   0                   0                   0                   0
      United States             Uruguay            Zimbabwe
                567                   0                   0

> table(dfc$self_employed)

    No Yes
  0 525  44
>
```

--> 44 répondants qui sont travailleurs indépendants, 525 non independants

```
> table(dfc$tech_company)

 No Yes
101 468
```

468 répondants travaillant dans une entreprise technologique tandis que 101 n’y travaillent pas .

Test de corrélation de Pearson:

-cor.test(dfc$Age, as.numeric(dfc$treatment))

```
> cor.test(dfc$Age, as.numeric(dfc$treatment))

        Pearson's product-moment correlation

data:  dfc$Age and as.numeric(dfc$treatment)
t = -0.18103, df = 567, p-value = 0.8564
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.08974364  0.07464203
sample estimates:
         cor
-0.007602167
```

Cette analyse de corrélation utilise le test de corrélation de Pearson pour évaluer la relation entre l'âge des répondants et leur décision de chercher un traitement pour un problème de santé mentale.

Étant donné que la valeur p (p-value) est significativement supérieure à 0.05 (niveau de signification courant),cela suggère qu'il n'y a pas de corrélation significative entre l'âge des répondants et leur décision de chercher un traitement pour un problème de santé mentale. La corrélation estimée est très proche de zéro (cor = -0.0076), ce qui confirme cette conclusion.

```
> cor.test(dfc$Age, as.numeric(dfc$remote_work))

        Pearson's product-moment correlation

data:  dfc$Age and as.numeric(dfc$remote_work)
t = 4.3346, df = 567, p-value = 1.728e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.09834248 0.25749926
sample estimates:
      cor
0.1790923
```

Cette analyse de corrélation utilise le test de corrélation de Pearson pour évaluer la relation entre l'âge des répondants et le fait de travailler à distance.

Étant donné que la valeur p est très faible (inférieure à 0.05), indique qu'il existe une corrélation significative entre l'âge des répondants et le fait de travailler à distance. La corrélation estimée est de 0.1791, ce qui suggère une corrélation positive modérée entre l'âge et le travail à distance.

```
> cor.test(dfc$Age, as.numeric(dfc$leave))

        Pearson's product-moment correlation

data:  dfc$Age and as.numeric(dfc$leave)
t = 0.15819, df = 567, p-value = 0.8744
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.07559551  0.08879240
sample estimates:
        cor
0.006643329
```

Étant donné que la valeur p est supérieure à 0.05, Cela suggère qu'il n'y a pas de corrélation significative entre l'âge et la politique de congé pour les problèmes de santé mentale dans le lieu de travail. La corrélation estimée est de 0.0066, ce qui indique une corrélation très faible entre l'âge et la politique de congé pour les problèmes de santé mentale.

```
> cor.test(dfc$Age, as.numeric(dfc$benefits))

        Pearson's product-moment correlation

data:  dfc$Age and as.numeric(dfc$benefits)
t = 3.996, df = 567, p-value = 7.292e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.0844541 0.2443758
sample estimates:
      cor
0.1655028
```

Étant donné que la valeur p est inférieure à 0.05, ,Cela suggère qu'il existe une corrélation significative entre l'âge et la disponibilité d'avantages en matière de santé mentale dans le lieu de travail. La corrélation estimée est de 0.1655, ce qui indique une corrélation positive modérée entre l'âge et la disponibilité d'avantages en matière de santé mentale.

Test du khi-deux entre 2 variables :

cette commande effectue un test du khi-deux pour évaluer s'il y a une association significative entre 2 variables .

```
> chisq.test(table(dfc$Gender, dfc$treatment))

        Pearson's Chi-squared test

data:  table(dfc$Gender, dfc$treatment)
X-squared = 21.565, df = 2, p-value = 2.076e-05
```

Étant donné que la valeur p est inférieure à 0.05, Cela suggère qu'il existe une association significative entre le genre des répondants et le fait d'avoir recherché un traitement pour un problème de santé mentale.

```
> chisq.test(table(dfc$family_history, dfc$treatment))

      Pearson's Chi-squared test with Yates' continuity correction

data:  table(dfc$family_history, dfc$treatment)
X-squared = 54.592, df = 1, p-value = 1.483e-13
```

Étant donné que la valeur p est extrêmement faible, bien inférieure à 0.05, Cela suggère qu'il existe une association significative entre les antécédents familiaux de problèmes de santé mentale et le fait d'avoir recherché un traitement pour un problème de santé mentale.

```
> chisq.test(table(dfc$family_history, dfc$mental_health_consequence))

      Pearson's Chi-squared test
```

```
data:  table(dfc$family_history, dfc$mental_health_consequence)
X-squared = 8.9499, df = 2, p-value = 0.01139
```

Étant donné que la valeur p est inférieure à 0.05,Cela suggère qu'il existe des preuves statistiquement significatives d'une association entre les antécédents familiaux de problèmes de santé mentale et les conséquences négatives en matière de santé mentale au travail.

Vérification de la distribution normale d'une variable:

On effectue un test de normalité de Shapiro-Wilk pour vérifier si la variable suit une distribution normale.

```
> shapiro.test(dfc$Age)

        Shapiro-Wilk normality test

data:  dfc$Age
W = 0.9616, p-value = 4.961e-11
```

Étant donné que la valeur p est significativement inférieure à 0.05, Cela suggère que l'âge des répondants ne suit pas une distribution normale.

Pour mieux structurer notre analyse on vous présente ci dessous nos questions de recherches précises :

1. Effets des avantages en santé mentale sur la recherche de traitement :

Les employés bénéficiant d'avantages en santé mentale sont-ils plus susceptibles de rechercher un traitement pour leurs problèmes de santé mentale que ceux n'en bénéficiant pas ?

- on utilisera le test de chi carré pour comparer la proportion de personnes ayant recherché un traitement entre les groupes bénéficiant et non bénéficiant d'avantages en santé mentale.

```
> chisq.test(table(dfc$benefits, dfc$treatment))

        Pearson's Chi-squared test

data:  table(dfc$benefits, dfc$treatment)
X-squared = 27.344, df = 2, p-value = 1.154e-06
```

Étant donné que la valeur de statistique de test (X-squared) est élevée, cela indique

La valeur p (p-value) de 1.154e-06 est extrêmement faible, ce qui suggère que la relation observée entre ces deux variables n'est pas due au hasard. En d'autres termes, il y a une association significative entre le fait de bénéficier d'avantages en santé mentale au travail et la probabilité de rechercher un traitement pour des problèmes de santé mentale.

Étant donné que la valeur de statistique de test (X-squared) est élevée, cela indique qu'il existe des différences significatives entre les groupes de traitement et de non-traitement parmi ceux qui bénéficient d'avantages en santé mentale. En conséquence, on peut supposer que les employés qui bénéficient d'avantages en santé mentale sont plus enclins à rechercher un traitement pour leurs problèmes de santé mentale par rapport à ceux qui n'en bénéficient pas.

2. Impact du télétravail sur les interférences liées à la santé mentale :

Les employés travaillant à distance sont-ils moins susceptibles de subir des interférences dans leur travail dues à des problèmes de santé mentale par rapport à ceux travaillant sur site ?

Pour évaluer l'impact du télétravail sur les interférences liées à la santé mentale, nous avons utilisé le test de Kruskal-Wallis pour comparer les niveaux de perturbation du travail (**work_interfere**) entre les groupes de travail à distance et sur site. Voici la commande utilisée :

```
> kruskal.test(dfc$work_interfere ~ dfc$remote_work)

        Kruskal-Wallis rank sum test

data:  dfc$work_interfere by dfc$remote_work
Kruskal-Wallis chi-squared = 0.43164, df = 1, p-value = 0.5112

>
>
```

Le résultat du test indique qu'il n'y a pas de différence significative dans les niveaux de perturbation du travail entre les groupes de travail à distance et sur site (p-valeur = 0.5112). Cela suggère que le télétravail n'a pas un impact significatif sur les interférences liées à la santé mentale telles que perçues par les répondants de l'enquête.

3. Corrélation entre la discussion de problèmes de santé mentale et la perception des conséquences :

Existe-t-il une corrélation entre le fait de discuter de problèmes de santé mentale avec des collègues ou des superviseurs et la perception des conséquences négatives pour la santé mentale ou physique au travail ?

- On utilisera le test de chi carré pour examiner la relation entre la discussion de problèmes de santé mentale avec des collègues ou des superviseurs et la perception des conséquences négatives pour la santé mentale ou physique au travail.

```
> chisq.test(table(dfc$coworkers, dfc$mental_health_consequence))

        Pearson's Chi-squared test

data:  table(dfc$coworkers, dfc$mental_health_consequence)
X-squared = 132.28, df = 4, p-value < 2.2e-16
```

Le test du Chi-deux suggère qu'il existe une association significative entre la discussion de problèmes de santé mentale avec des collègues et la perception des conséquences mentales au travail. La valeur de p très faible (< 2.2e-16) indique que la relation observée est très improbable sous l'hypothèse nulle d'indépendance entre ces deux variables. Cela signifie qu'il est peu probable que la relation observée soit due au hasard, suggérant qu'il existe une association entre ces deux variables dans la population sous-jacente.

4. Comparaison des politiques de santé mentale entre les entreprises technologiques et non-technologiques :

Les entreprises technologiques offrent-elles davantage d'avantages en santé mentale ou de programmes de bien-être par rapport aux entreprises non-technologiques ?

- On utilise le test de chi carré pour comparer la proportion d'entreprises offrant des avantages en santé mentale ou des programmes de bien-être entre les entreprises technologiques et non-technologiques.

```
> chisq.test(table(dfc$tech_company, dfc$benefits))

        Pearson's Chi-squared test

data:  table(dfc$tech_company, dfc$benefits)
X-squared = 1.7895, df = 2, p-value = 0.4087
```

Le test du Chi-deux indique qu'il n'y a pas de différence significative dans les politiques de santé mentale entre les entreprises technologiques et non-technologiques. La valeur de p élevée (0.4087) suggère qu'il n'y a pas suffisamment de preuves pour rejeter

l'hypothèse nulle d'indépendance entre ces deux variables. Cela signifie que la présence ou l'absence de politiques de santé mentale ne semble pas dépendre du fait que l'entreprise soit technologique ou non.

5. Analyse des variations régionales dans la perception de la santé mentale en milieu de travail:

Existe-t-il des différences significatives dans la manière dont les problèmes de santé mentale sont perçus et traités entre les différents États américains ou entre les différents pays ?

- On utilise test de Kruskal-Wallis pour comparer les réponses sur la perception de la santé mentale en milieu de travail entre les différents États américains.

```
> kruskal.test(dfc$mental_health_consequence ~ dfc$state)

        Kruskal-Wallis rank sum test

data:  dfc$mental_health_consequence by dfc$state
Kruskal-Wallis chi-squared = 38.632, df = 42, p-value = 0.6196
```

Le test Kruskal-Wallis ne montre pas de différences significatives dans la perception des conséquences liées à la santé mentale en milieu de travail entre les différents États américains. Avec une valeur de p élevée (0.6196), nous ne pouvons pas rejeter l'hypothèse nulle selon laquelle les médianes des distributions des perceptions des conséquences sont égales dans tous les États étudiés. Cela suggère que la perception des conséquences liées à la santé mentale au travail semble être similaire quel que soit l'État de résidence des répondants.

6. Impact des congés pour problèmes de santé mentale sur la recherche du traitement :

Existe-t-il une association entre le fait d'avoir recherché un traitement pour des problèmes de santé mentale et la disponibilité de congés spécifiques pour la santé mentale dans le lieu de travail ?

- on utilise la régression robuste pour examiner la relation entre la disponibilité des congés pour les problèmes de santé mentale et la recherche du traitement , en contrôlant d'autres variables pertinentes.

1ere etape on change les valeurs de la colonne treatement pour pouvoir faire le test.(dc une nouvelle dataframe)

```
dc$treatment <- as.numeric(ifelse(dc$treatment == "Yes", 1, ifelse(dc$treatment == "No", 0, dc$treatment)))
```

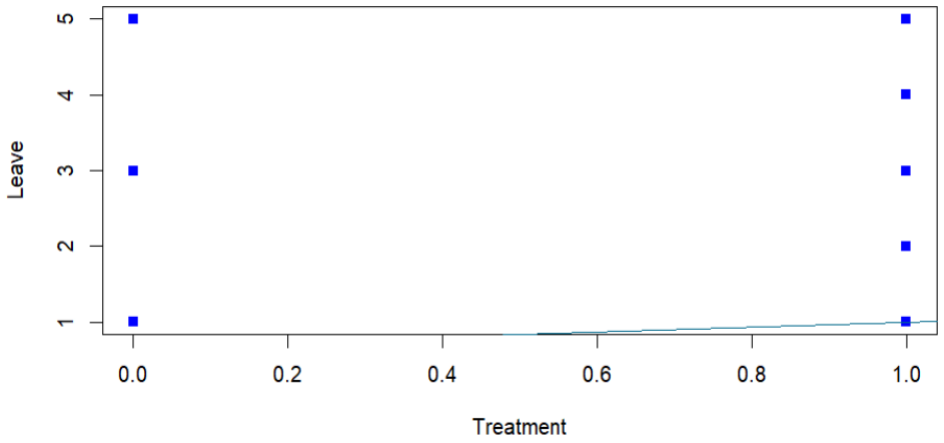
On a changé les valeurs contenant yes par 1 et les valeurs contenant No par 0.

```
summary(lm(treatment ~ leave, data = dc))
```

```
Call:
lm(formula = treatment ~ leave, data = dc)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7895  0.0000  0.2105  0.3226  0.4000

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.67742    0.07283   9.301 3.98e-14 ***
leaveSomewhat difficult  0.32258    0.15354   2.101   0.0390 *
leaveSomewhat easy      -0.07742    0.14747  -0.525   0.6011
leaveVery difficult      0.32258    0.14231   2.267   0.0263 *
leaveVery easy           0.11205    0.11814   0.948   0.3459
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4055 on 75 degrees of freedom
Multiple R-squared:  0.116,     Adjusted R-squared:  0.06883
F-statistic:  2.46 on 4 and 75 DF,  p-value: 0.05256
```

Dans ce modèle linéaire, nous examinons la relation entre le traitement (treatment) pour les problèmes de santé mentale et la politique de congé (leave). Les résultats montrent que le modèle est significatif (F = 2.46, p = 0.05256), ce qui indique une relation globale entre le traitement et le type de congé. Cependant, les coefficients pour les différents niveaux de congé ne sont pas tous significatifs. Les employés qui ont trouvé leur congé "Somewhat difficult" ou "Very difficult" sont plus susceptibles de rechercher un traitement que ceux qui ont trouvé leur congé "Somewhat easy" ou "Very easy". Cela suggère que la perception de la difficulté du congé peut influencer la décision de rechercher un traitement pour les problèmes de santé mentale.

Représentation graphique :



7. Analyse des interactions entre les variables démographiques et les politiques de l'entreprise sur la recherche de traitement pour les problèmes de santé mentale :

Quel est l'impact du genre sur la propension des individus à rechercher un traitement pour les problèmes de santé mentale, en tenant compte d'autres facteurs tels que l'âge et les avantages en santé mentale dans le lieu de travail ?

-on utilise l'analyse de variance (ANOVA) ou la régression linéaire multiple pour examiner les interactions entre les politiques de l'entreprise sur la santé mentale et les variables démographiques des employés, telles que l'âge, le genre, etc.

```
> summary(aov(treatment ~ Age * Gender * benefits, data = dc))
                    Df Sum Sq Mean Sq F value Pr(>F)
Age                  1  0.503  0.5029   3.100 0.0828 .
Gender               1  0.942  0.9420   5.807 0.0187 *
benefits             2  0.494  0.2472   1.524 0.2252
Age:Gender           1  0.044  0.0442   0.273 0.6032
Age:benefits         2  0.696  0.3482   2.146 0.1248
Gender:benefits      2  0.054  0.0269   0.166 0.8474
Age:Gender:benefits  2  0.186  0.0930   0.573 0.5663
Residuals           68 11.030  0.1622
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Ce résumé ANOVA présente les résultats de l'ajustement d'un modèle où le traitement est régressé sur l'âge, le genre et les avantages, ainsi que leurs interactions. Voici comment interpréter les résultats :

- Âge : La valeur p (0,0828) suggère que l'âge pourrait ne pas être significativement associé au traitement.
- Genre : La valeur p (0,0187) indique que le genre est significativement associé au traitement, car elle est inférieure au seuil de signification typique de 0,05.
- benefit : La valeur p (0,2252) suggère que la variable des avantages pourrait ne pas être significativement associée au traitement.
- Interactions : Les valeurs p pour les termes d'interaction (Âge:Genre, Âge:benefit, Genre:benefit et Âge:Genre:benefit) sont toutes supérieures à 0,05, ce qui indique que ces interactions pourraient ne pas être statistiquement significatives.

Dans l'ensemble, le genre semble avoir un effet significatif sur le traitement, tandis que l'âge et les avantages peuvent ne pas avoir d'effets significatifs. Cependant, il est essentiel de considérer le contexte et les limitations potentielles de l'analyse.

Intervalle de confiance d'une moyenne:

```
> t.test(dv$Age)

        One Sample t-test

data:  dv$Age
t = 44.719, df = 83, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 33.83005 36.97947
sample estimates:
mean of x
 35.40476
```

Le test t à un échantillon a été utilisé pour évaluer si la moyenne de l'âge dans votre ensemble de données diffère significativement de zéro. La sortie indique un t de 44.719 avec un degré de liberté de 83. Le p-value est très proche de zéro (p-value < 2.2e-16), ce qui indique une forte significativité statistique. Par conséquent, nous rejetons l'hypothèse nulle selon laquelle la moyenne de l'âge est égale à zéro. La moyenne de l'âge dans l'échantillon est estimée à environ 35ans, avec un intervalle de confiance à 95 % entre environ 33 ans et 36 ans. En bref, cela suggère qu'il existe une différence significative dans l'âge par rapport à zéro dans l'échantillon analysé, avec une moyenne d'environ 35 ans.

Data visualisation :

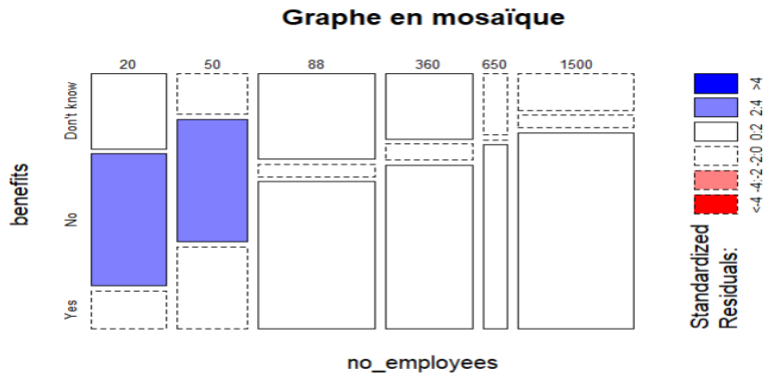
On reprendra le dataframe d'origine aprés le cleaning des valeurs NA :

1ere etape on va se focaliser sur la colonne no_employees dans le cleaning:resultat aprés le cleaning.

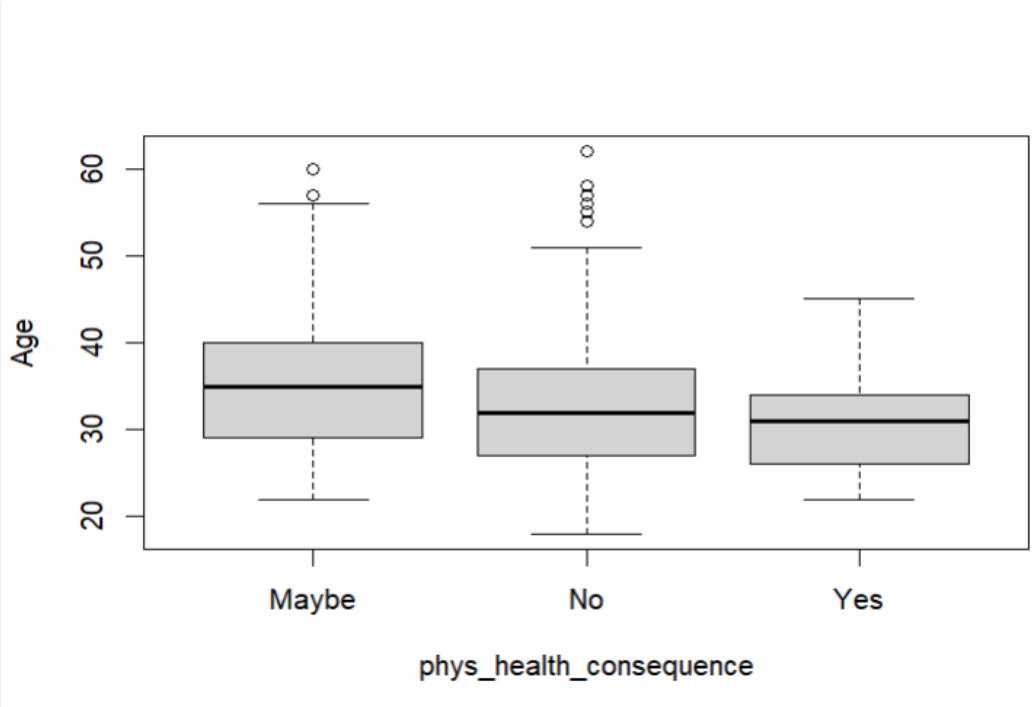
```
> unique(dv$no_employees)
[1]   88 1500   20   50  650  360
> 
```

```
> mosaicplot(no_employees~benefits, data=dv, shade = TRUE, main
+           ="Graphe en mosaïque")
> fisher.test (dv$no_employees, dv$benefits,simulate.p.value = TRUE)

        Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

data:  dv$no_employees and dv$benefits
p-value = 0.001999
alternative hypothesis: two.sided
```

Le test exact de Fisher pour les données de comptage avec une valeur p simulée (basée sur 2000 réplicats) indique une valeur p de 0.001999, suggérant une association significative entre le nombre d'employés et les avantages en matière de santé mentale dans votre ensemble de données.




```
boxplot (Age ~ phys_health_consequence, data = dfc)
```



Le diagramme en boîte représente la distribution des âges en fonction des conséquences de santé physique perçues au travail.

La population qui a indiqué “maybe” est plus agée que les répondants par “NO” qui sont de leurs tour plus agée que les répondants par “yes “.

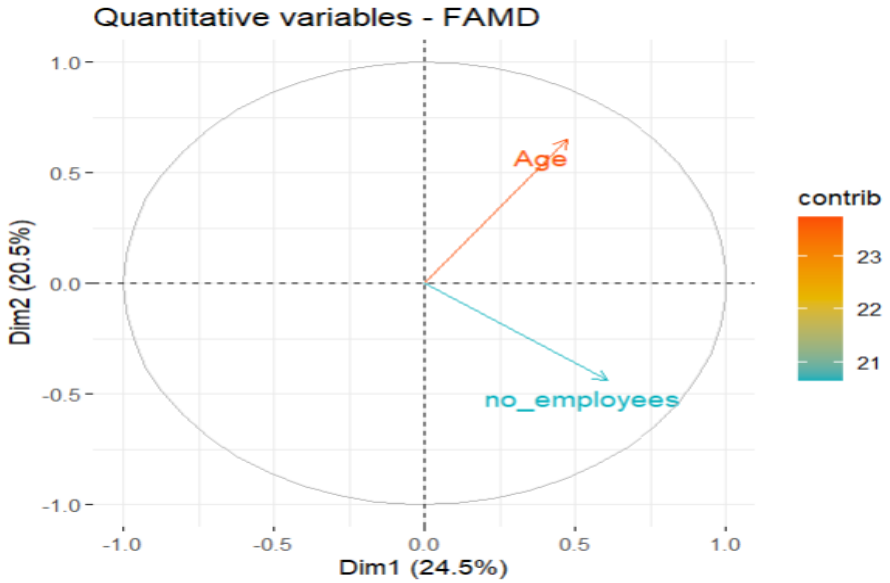
Fonction FAMD:

Avant de debuter il faut installer les packages :

install.packages("FactoMineR") et install.packages("factoextra")

```
> variables <- dv[, c("Age", "no_employees", "benefits", "phys_health_consequence")]
> famd_result <- FAMD(variables)
```

```
fviz_famd_var (famd_result, "quanti.var", col.var = "contrib",
+                gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel =
+                    TRUE)
```



```
> summary(famd_result)

Call:
FAMD(base = variables)


Eigenvalues
                       Dim.1  Dim.2  Dim.3  Dim.4  Dim.5
Variance               1.470  1.231  1.008  0.963  0.765
% of var.             24.506 20.511 16.807 16.044 12.758
Cumulative % of var.  24.506 45.016 61.823 77.867 90.625
```

Les résultats de l'analyse FAMD montrent les valeurs propres et l'inertie associée à chaque dimension :

Dim.1: Variance = 1.470, % de variance = 24.506%

Dim.2: Variance = 1.231, % de variance = 20.511%

Dim.3: Variance = 1.008, % de variance = 16.807%

Dim.4: Variance = 0.963, % de variance = 16.044%

Dim.5: Variance = 0.765, % de variance = 12.758%

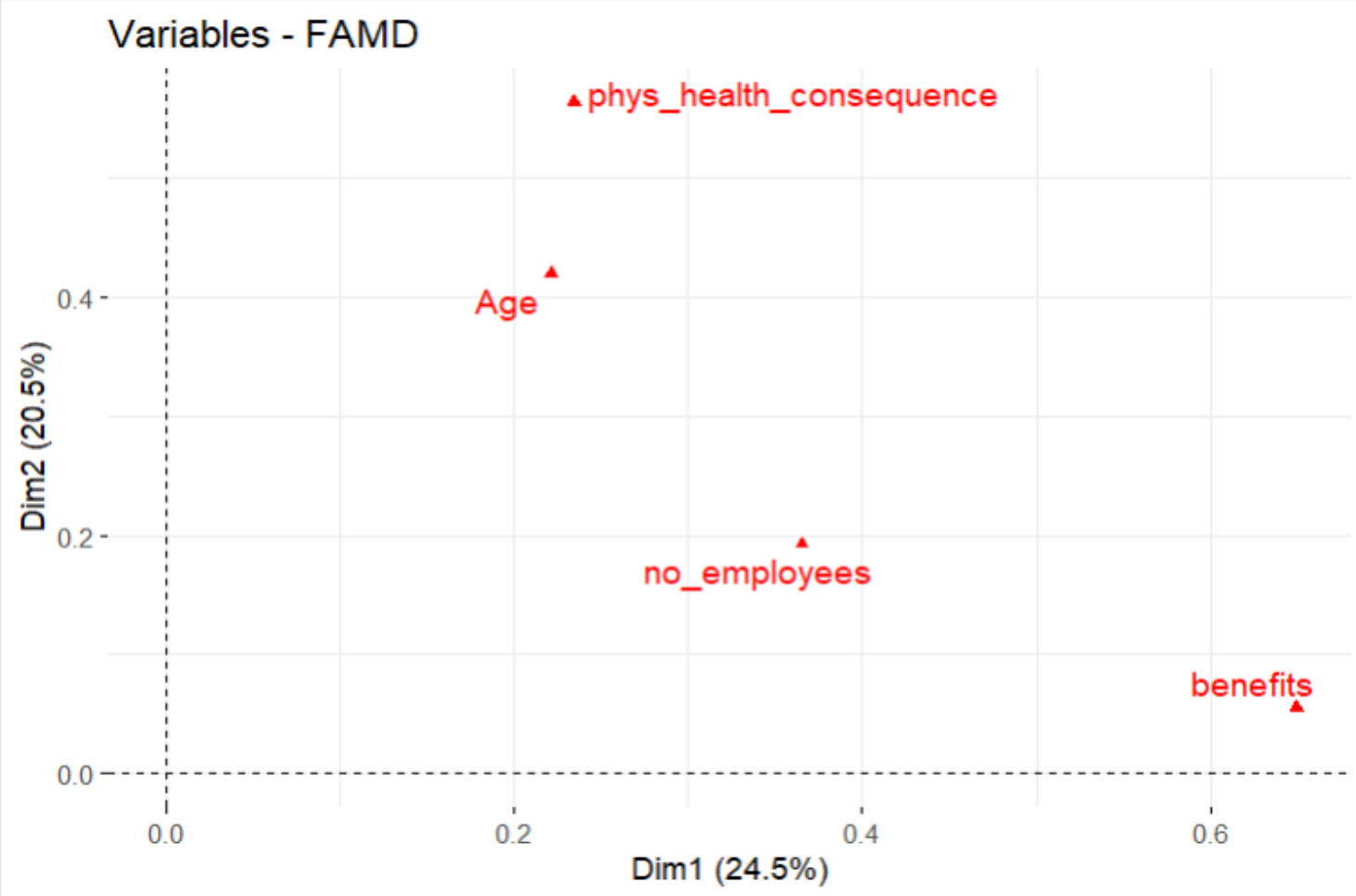
Ces valeurs indiquent la proportion de variance expliquée par chaque dimension et fournissent des informations sur l'importance relative des dimensions dans la représentation des données.

```
Continuous variables
                      Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
Age                 |  0.470 15.054  0.221 |  0.648 34.094  0.420 |  0.068  0.459  0.005 |
no_employees        |  0.604 24.843  0.365 | -0.438 15.616  0.192 |  0.186  3.445  0.035 |

Categories
                      Dim.1    ctr   cos2 v.test    Dim.2    ctr   cos2 v.test    Dim.3    ctr
Don't know          | -0.705  5.753  0.161 -3.060 |  0.032  0.017  0.000  0.151 |  1.411 48.947
No                  | -1.522 20.419  0.490 -5.548 |  0.503  3.178  0.053  2.003 | -1.214 27.583
Yes                 |  0.833 17.976  0.750  7.057 | -0.185  1.269  0.037 -1.715 | -0.217  2.599
Maybe               |  0.851  7.586  0.198  3.459 |  1.341 26.866  0.491  5.955 | -0.113  0.282
No                  | -0.129  0.552  0.039 -1.535 | -0.527 13.081  0.654 -6.838 | -0.102  0.730
Yes                 | -1.685  7.818  0.173 -3.185 |  1.223  5.879  0.091  2.527 |  1.651 15.955
                     cos2 v.test
Don't know          0.643  7.391 |
No                  0.311 -5.340 |
Yes                 0.051 -2.222 |
Maybe               0.003 -0.553 |
No                  0.024 -1.462 |
Yes                 0.166  3.768 |
> |
```

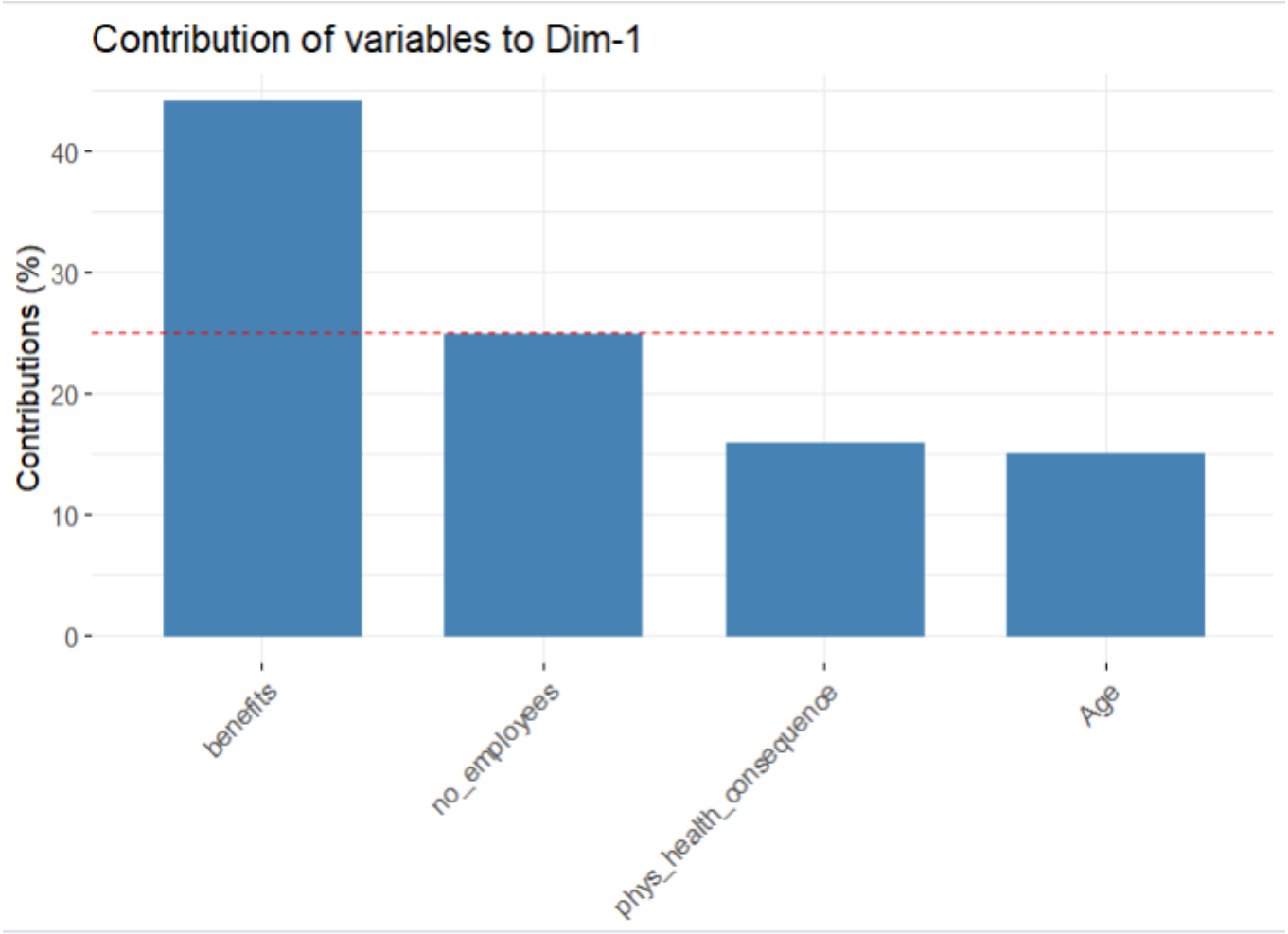
Les variables continues telles que l'âge et le nombre d'employés ont une contribution significative aux deux premières dimensions, avec des contributions relatives (ctr) de 15.054% à 24.843%. Les catégories telles que "Don't know", "No" et "Yes" dans différentes variables catégorielles contribuent également de manière significative à l'explication de la variance dans les dimensions correspondantes, avec des valeurs de cos2 allant jusqu'à 0.750, indiquant leur poids dans la représentation des données.

Representation graphique :

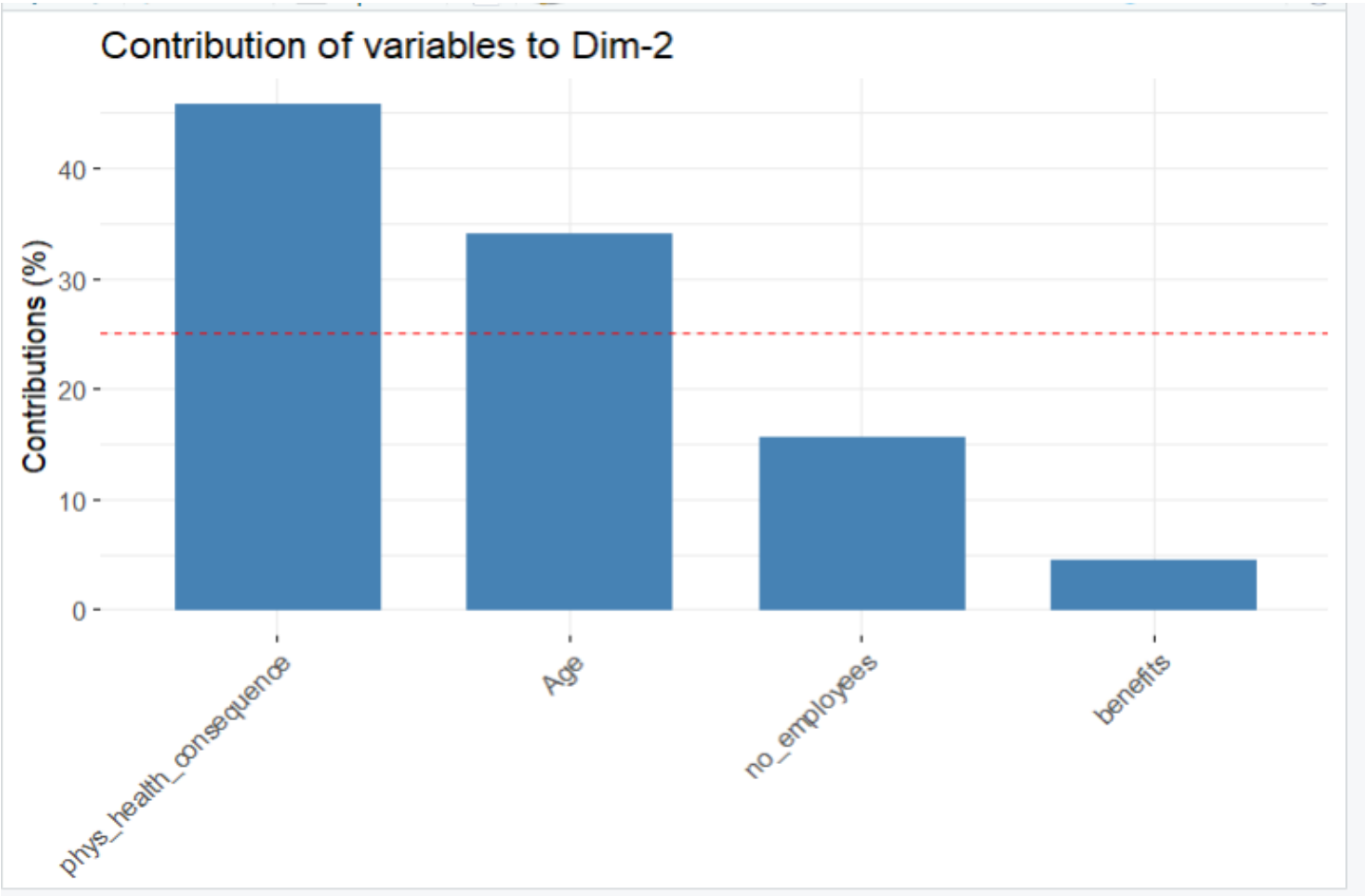
```
fviz_famd_var(famd_result, repel = TRUE)
```



Contribution a la 1ere dimension :



Contribution a la 2eme dimension :



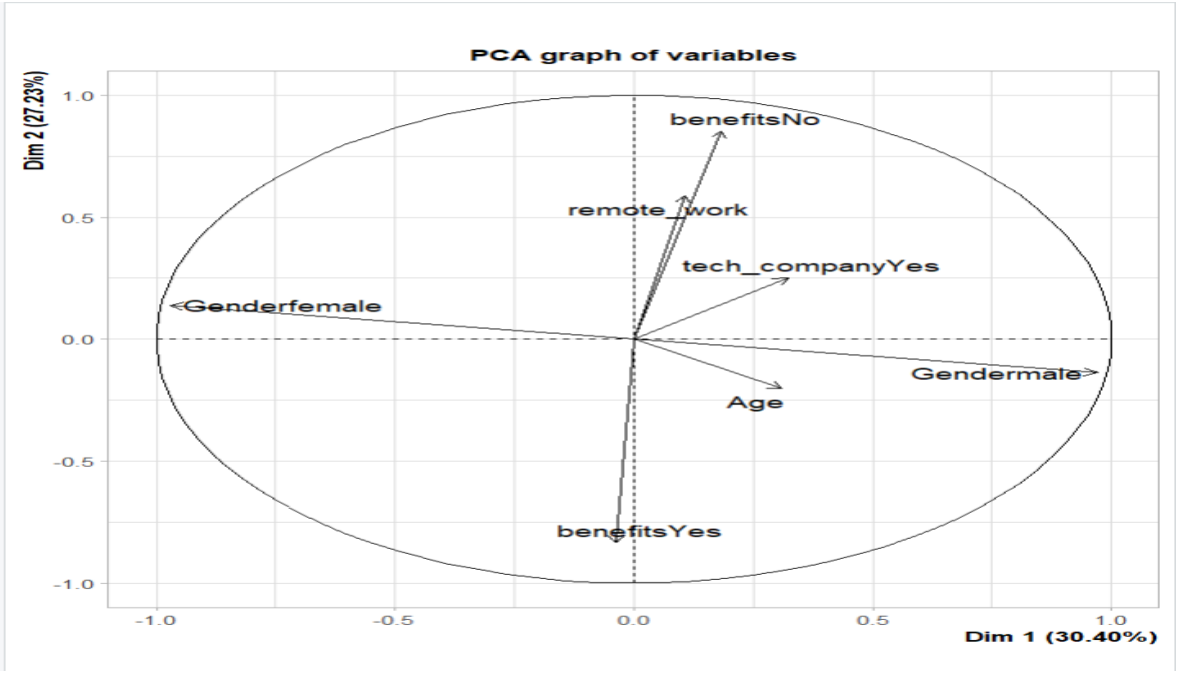
La ligne en pointillé rouge sur le graphique ci-dessus indique la valeur moyenne attendue, si les contributions étaient uniformes.

Analyse en Composantes Principales

Pour effectuer une analyse en composantes principales (PCA), il est nécessaire que toutes les variables soient quantitatives. Vous devrez convertir ces variables en format numérique ou en variables indicatrices (dummy variables) si elles sont catégoriques.

1ere etape :**Conversion des variables catégoriques en variables indicatrices** :

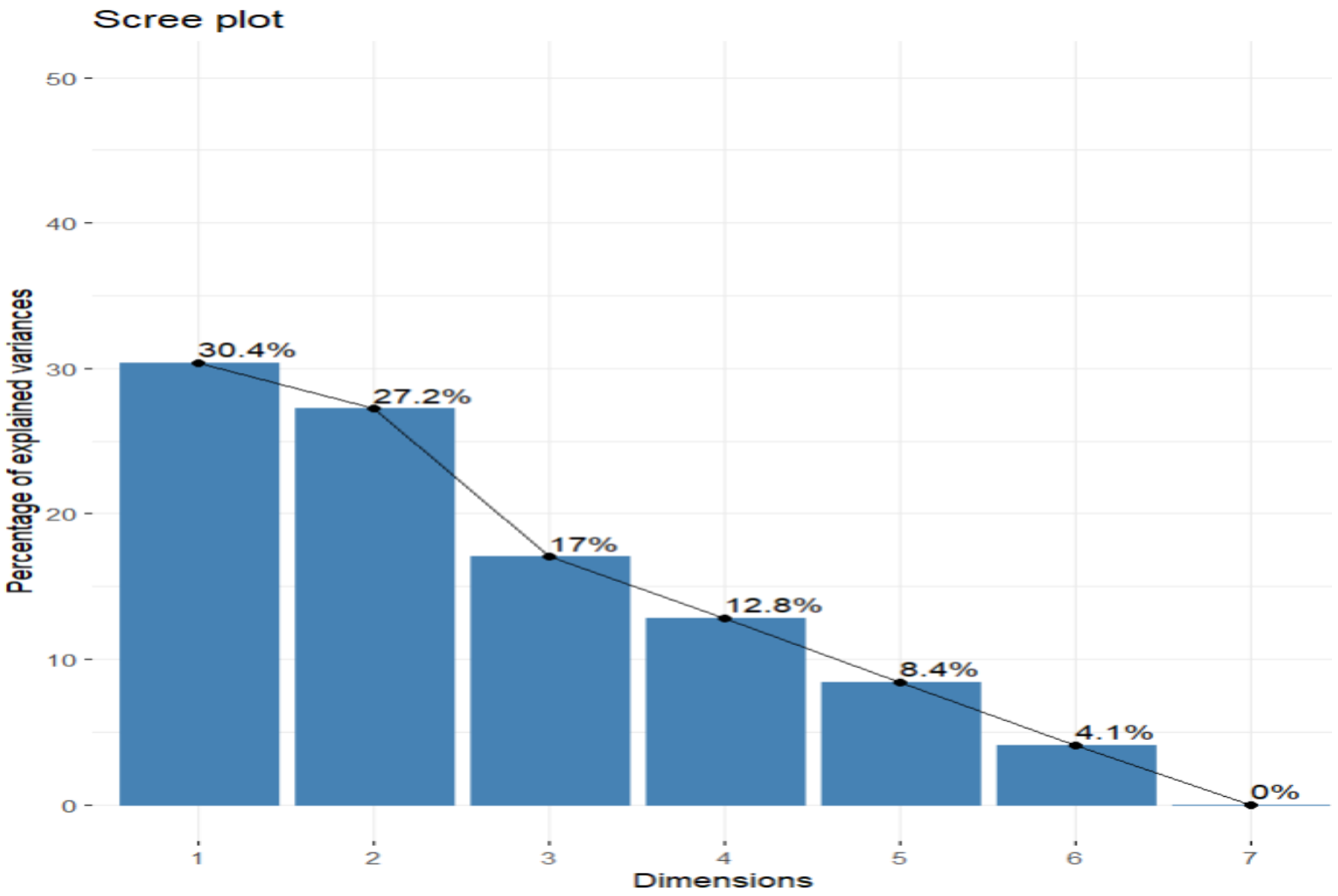
```
>  variables <- dfc[dfc$state == "WA", c("Age", "Gender", "remote_work", "tech_company", "benefits")]
>  variables_indicatrices <- model.matrix(~ Gender + remote_work + tech_company + benefits - 1, data = variables)
> colonnes_quantitatives <- setdiff(names(variables), c("Gender", "remote_work", "tech_company", "benefits"))
> variables_quantitatives <- cbind(variables[colonnes_quantitatives], variables_indicatrices)
> resultat_pca <- PCA(variables_quantitatives, scale.unit = TRUE, graph = TRUE)
```




```
> eigval = get_eigenvalue(resultat_pca)
> eigval
       eigenvalue variance.percent cumulative.variance.percent
Dim.1 2.128160e+00     3.040228e+01                    30.40228
Dim.2 1.905863e+00     2.722661e+01                    57.62890
Dim.3 1.193168e+00     1.704526e+01                    74.67415
Dim.4 8.977103e-01     1.282443e+01                    87.49859
Dim.5 5.898546e-01     8.426494e+00                    95.92508
Dim.6 2.852443e-01     4.074919e+00                   100.00000
Dim.7 2.697017e-31     3.852882e-30                   100.00000
```

Les résultats de l'analyse PCA montrent que les sept premières dimensions expliquent ensemble environ 100 % de la variance des données. La première dimension (Dim.1) a la plus grande valeur propre et explique environ 30,40 % de la variance totale. Les dimensions suivantes contribuent également de manière significative à la variance, avec la deuxième dimension (Dim.2) expliquant environ 27,23 % de la variance totale. Les dimensions restantes expliquent chacune une part décroissante de la variance, avec la septième dimension (Dim.7) ayant une valeur propre très proche de zéro, ce qui indique une faible contribution à l'explication de la variance des données.

Representation graphique avec plot:



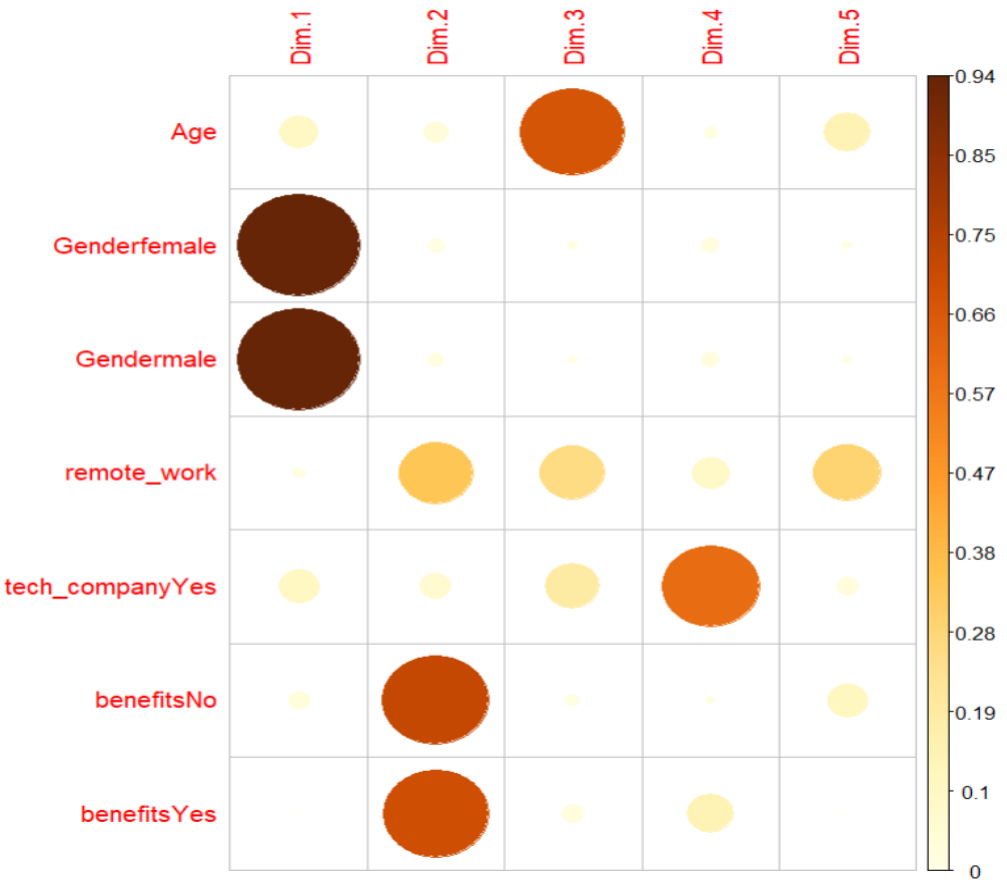
Le graphe des variables:

install.packages("corrplot")

library(corrplot)

```
> var = get_pca_var(resultat_pca)
> var
Principal Component Analysis Results for variables
 ===================================================
  Name       Description
1 "$coord"   "Coordinates for the variables"
2 "$cor"     "Correlations between variables and dimensions"
3 "$cos2"    "Cos2 for the variables"
4 "$contrib" "contributions of the variables"
```

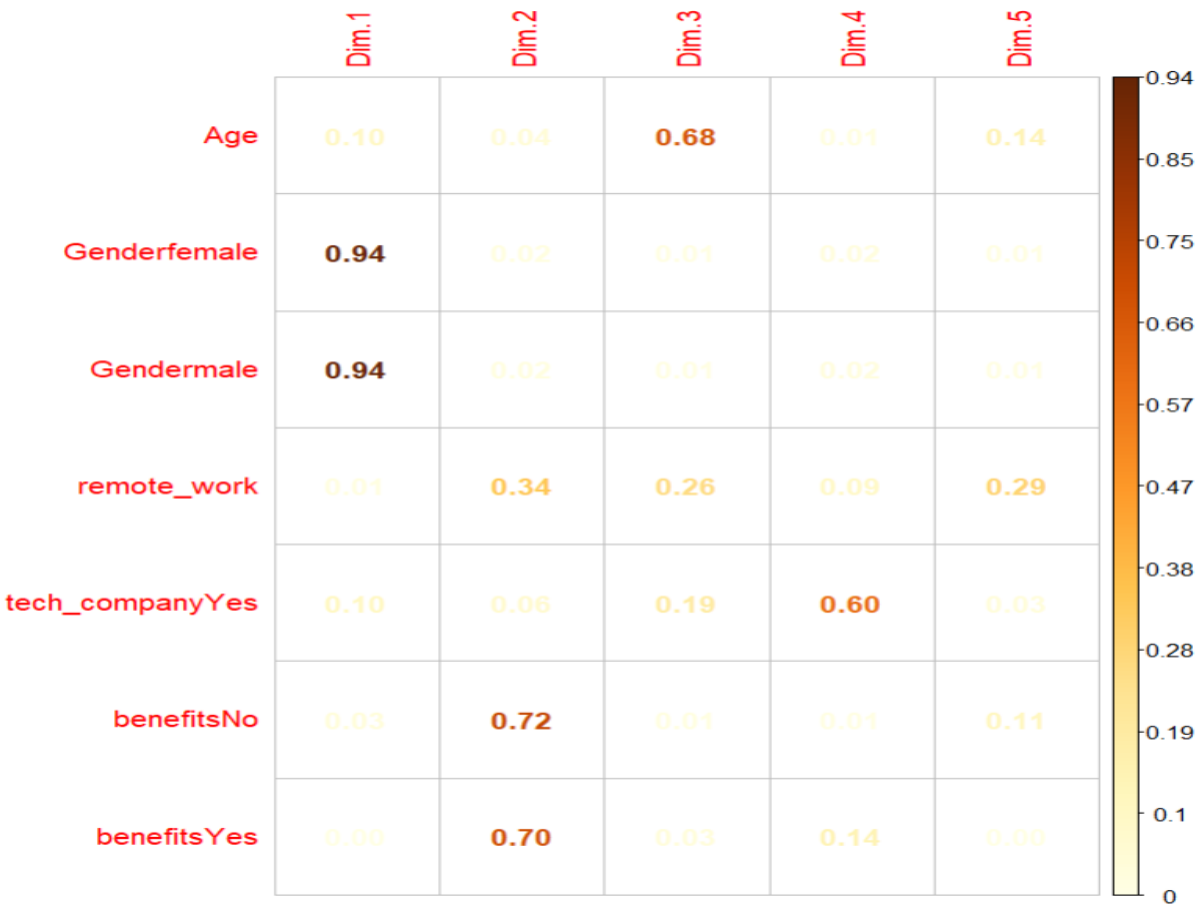
```
(var$cos2, is.corr=FALSE)
```



Lorsque les carrés des cosinus sont représentés graphiquement, chaque cercle correspond à une variable, et sa couleur indique la qualité de sa représentation sur les composantes principales. Les couleurs plus foncées indiquent une meilleure représentation, c'est-à-dire que la variable est bien alignée avec la composante principale correspondante.

```
corrplot(var$cos2, is.corr=FALSE, method="number")
```

L'ajout de valeurs numériques dans la seconde commande permet de voir précisément les valeurs des carrés des cosinus pour chaque variable, fournissant ainsi une évaluation quantitative de la qualité de représentation.



Conclusion :

Après avoir analyser les données et interpréter les résultats, voici nos conclusions tirées du projet :

1. Impact des facteurs démographiques sur la santé mentale:

L'âge semble jouer un rôle dans la perception et la gestion des problèmes de santé mentale au travail. Les travailleurs plus âgés peuvent être plus enclins à rechercher de l'aide ou à discuter de leurs problèmes de santé mentale avec leurs collègues et superviseurs.

2. Influence de l'environnement de travail:

Des facteurs tels que la taille de l'entreprise, le secteur d'activité et la disponibilité des avantages en santé mentale peuvent avoir un impact significatif sur la façon dont les problèmes de santé mentale sont perçus et traités en milieu professionnel.

3. Nécessité de politiques de santé mentale adaptées :

Les résultats mettent en évidence l'importance pour les entreprises d'adopter des politiques de santé mentale qui répondent aux besoins de leurs employés, notamment en offrant des congés spécifiques pour les problèmes de santé mentale et en fournissant des ressources pour le bien-être mental.

4. Importance de la sensibilisation et du soutien:

Encourager la discussion ouverte sur les problèmes de santé mentale au travail et fournir un soutien adéquat, y compris des programmes de bien-être et des options de traitement, peut contribuer à réduire la stigmatisation et à promouvoir un environnement de travail sain.

En résumé, ce projet met en lumière l'importance croissante de la santé mentale en milieu de travail et souligne la nécessité d'adopter des politiques et des pratiques qui favorisent le bien-être mental des employés.