TP 3 - Variables quantitatives

Lucas Leroy ET Vincent Dubuc

2023-2024

Nombre d'enfants

Question

On cherche à répondre à la question : "Les couples pacsés ont-ils plus d'enfants de moins de 25 ans que les couples mariés en France en 2017 ?"

Définissez les individus, la population, et la variable mesurée, en précisant son type et ses modalités.

Définition des termes :

- Individus : Couples en France
- Population : Tous les couples en France en 2017
- Variable mesurée : Nombre d'enfants de moins de 25 ans
- Type : Quantitative discrète
- Modalités : Nombre entier d'enfants (0, 1, 2, ...)

Données

```
# Chargement des données
donnees <- read.csv("rp2017_td_fam2.csv")

# Affichage des premières lignes pour comprendre la structure
head(donnees)
```

```
##                                                                                FAM2...Couples.selon.le.statut.
## 1
## 2
## 3                                                                                        FAM2 - Couples selon le
## 4 \tAucun enfant de moins de 25 ans\t1 enfant de moins de 25 ans\t2 enfants de moins de 25 ans\t3 en
## 5                                                                                      Couple de deu
## 6                                                                                            Couple
```

```
# Résumé statistique des données
summary(donnees)
```

```
##  FAM2...Couples.selon.le.statut.conjugal.des.conjoints.et.le.nombre.d.enfants.de.moins.de.25.ans.en.
##  Length:12
##  Class :character
##  Mode  :character
```

Formatage

```
# Chemin du fichier
file_path <- "rp2017_td_fam2.csv"

# Lecture du fichier en utilisant le séparateur de tabulation et en spécifiant l'encodage
data_raw <- read.csv(file_path, skip = 7, header = FALSE, sep = "\t", fileEncoding = "ISO-8859-1")

# Suppression de la dernière ligne (pied de page)
data_clean <- data_raw[-nrow(data_raw), ]

head(data_clean)
```

```
##                                                        V1      V2      V3
## 1                        Couple de deux personnes mari?es 6448133 1644613
## 2                        Couple de deux personnes pacs?es  407144  337083
## 3  Couple de deux personnes en concubinage ou union libre 1304386  673141
## 4 Couple de deux personnes ayant un autre statut conjugal  177322   53585
## 5                                                Ensemble 8336985 2708422
##        V4      V5     V6       V7
## 1 1975639  798166 263408 11129960
## 2  335833   62577  11225  1153862
## 3  564489  167676  61358  2771049
## 4   41589   17237   8778   298511
## 5 2917549 1045657 344769 15353382
```

```
# Attribution de noms de colonnes
col_names <- c("Type de Couple", "Aucun enfant", "1 enfant", "2 enfants", "3 enfants", "4 enfants ou plu
colnames(data_clean) <- col_names

# Renommer la colonne pour la situation maritale en 'situation'
data_long <- data_clean %>%
  pivot_longer(cols = -c(`Total`, `Type de Couple`),
               names_to = "enfants",
               values_to = "compte") %>%
  rename(situation = `Type de Couple`)

# Vérification des noms de colonnes
print(colnames(data_long))
```

```
## [1] "situation" "Total"     "enfants"   "compte"
```

Polygone des fréquences

```
# Chargement de la bibliothèque ggplot2 pour la visualisation
library(ggplot2)

# Convertissons la variable 'enfants' en un facteur si ce n'est pas déjà fait
data_long$enfants <- factor(data_long$enfants, levels = c("Aucun enfant", "1 enfant", "2 enfants", "3 en
```

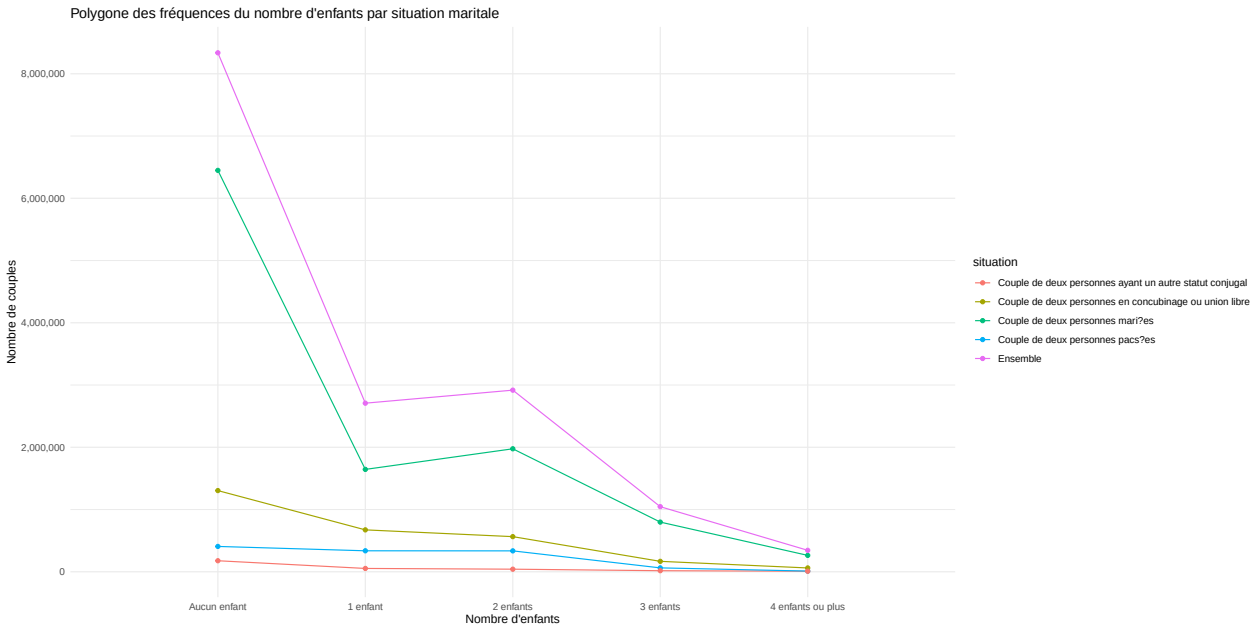
```r
head(data_long)
```

```
## # A tibble: 6 x 4
##   situation                          Total enfants            compte
##   <chr>                              <int> <fct>               <int>
## 1 Couple de deux personnes mari?es 11129960 Aucun enfant      6448133
## 2 Couple de deux personnes mari?es 11129960 1 enfant          1644613
## 3 Couple de deux personnes mari?es 11129960 2 enfants         1975639
## 4 Couple de deux personnes mari?es 11129960 3 enfants          798166
## 5 Couple de deux personnes mari?es 11129960 4 enfants ou plus  263408
## 6 Couple de deux personnes pacs?es  1153862 Aucun enfant       407144
```

```r
head(data_clean)
```

```
##                                         Type de Couple Aucun enfant 1 enfant
## 1                       Couple de deux personnes mari?es      6448133  1644613
## 2                       Couple de deux personnes pacs?es       407144   337083
## 3  Couple de deux personnes en concubinage ou union libre      1304386   673141
## 4 Couple de deux personnes ayant un autre statut conjugal       177322    53585
## 5                                                Ensemble      8336985  2708422
##   2 enfants 3 enfants 4 enfants ou plus    Total
## 1   1975639    798166            263408 11129960
## 2    335833     62577             11225  1153862
## 3    564489    167676             61358  2771049
## 4     41589     17237              8778   298511
## 5   2917549   1045657            344769 15353382
```

```r
# Nous utilisons 'geom_line' pour tracer un polygone des fréquences
# Utilisation de la fonction scale_x_discrete pour ajuster les limites de l'axe des x
# Utilisation de la fonction expand_limits pour ajouter de l'espace à droite
ggplot(data_long, aes(x = enfants, y = compte, group = situation, color = situation)) +
  geom_line() +
  geom_point() +
  scale_x_discrete(expand = expansion(add = 1)) + # Ajoute de l'espace sur les côtés de l'axe des x
  scale_y_continuous(labels = scales::comma) +
  labs(title = "Polygone des fréquences du nombre d'enfants par situation maritale",
       x = "Nombre d'enfants",
       y = "Nombre de couples") +
  theme_minimal()
```




```
# Justification du choix de la représentation :
# Le polygone des fréquences permet de comparer visuellement les distributions du nombre d'enfants
# entre différentes situations maritales. Cela aide à répondre à la question de savoir
# si les couples pacsés ont plus d'enfants de moins de 25 ans que les couples mariés.
```

#Question: Quelle est la répartition du nombre total de couples par nombre d'enfants pour chaque situation maritale?

Pour répondre à cette question, nous pouvons utiliser un graphique en barres empilées qui montre le nombre total de couples pour chaque nombre d'enfants, avec des couleurs différentes représentant chaque situation maritale. Cela permettrait de voir non seulement le nombre total de couples mais aussi de comparer la répartition entre les différentes situations maritales.

Autre Analyse

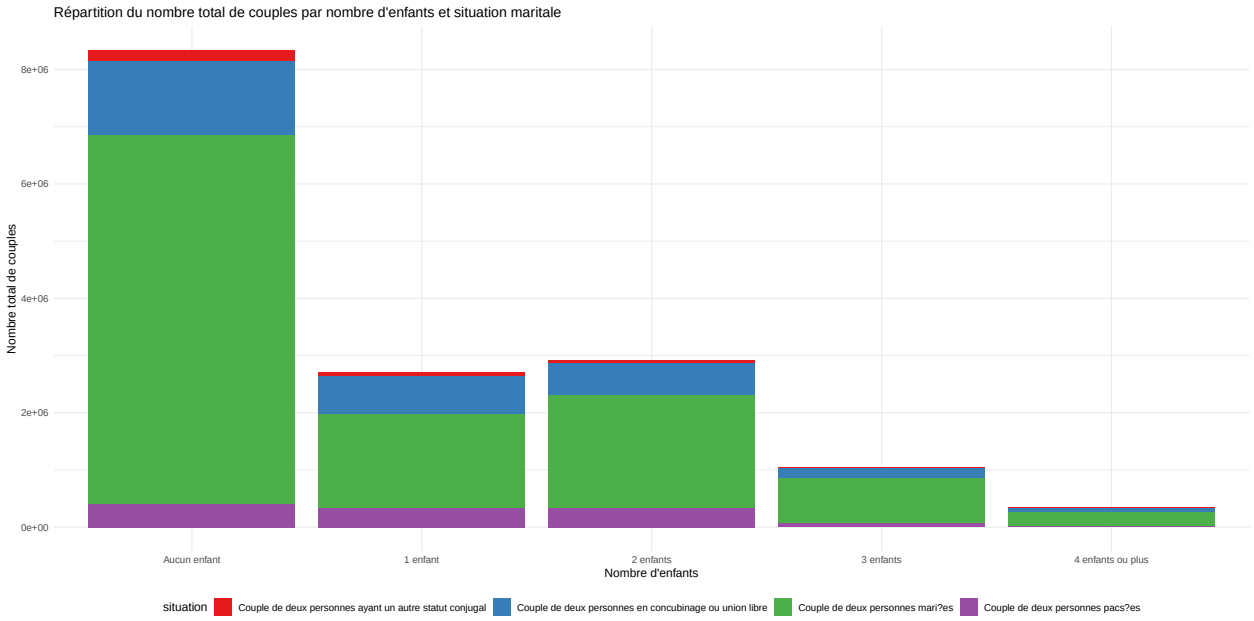
Une autre question intéressante à explorer est la répartition du nombre total de couples par nombre d'enfants pour chaque situation maritale.

```
# Chargement de la bibliothèque ggplot2 pour la visualisation
library(ggplot2)

# Filtrer pour exclure la catégorie 'Ensemble'
data_long_filtered <- data_long %>%
  filter(situation != "Ensemble")

# Recréer le graphique en barres empilées sans la catégorie 'Ensemble'
ggplot(data_long_filtered, aes(x = enfants, y = compte, fill = situation)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(title = "Répartition du nombre total de couples par nombre d'enfants et situation maritale",
       x = "Nombre d'enfants",
       y = "Nombre total de couples") +
  scale_fill_brewer(palette = "Set1") +
```

```
  theme_minimal() +
  theme(legend.position = "bottom")
```



Expéditions sur l'Everest

On s'intéresse à un jeu de données sur les expéditions dans la chaîne himalayenne, fourni par TidyTuesday.

Données

Lisez et chargez le jeu de données `members`. Décrivez-le en deux mots.

```
# Chargement des données
donnees <- read.csv("2020-09-22/members.csv")

# Affichage des premières lignes pour comprendre la structure
head(donnees)
```

```
##   expedition_id    member_id peak_id  peak_name year season sex age citizenship
## 1     AMAD78301 AMAD78301-01    AMAD Ama Dablam 1978 Autumn   M  40      France
## 2     AMAD78301 AMAD78301-02    AMAD Ama Dablam 1978 Autumn   M  41      France
## 3     AMAD78301 AMAD78301-03    AMAD Ama Dablam 1978 Autumn   M  27      France
## 4     AMAD78301 AMAD78301-04    AMAD Ama Dablam 1978 Autumn   M  40      France
## 5     AMAD78301 AMAD78301-05    AMAD Ama Dablam 1978 Autumn   M  34      France
## 6     AMAD78301 AMAD78301-06    AMAD Ama Dablam 1978 Autumn   M  25      France
##   expedition_role hired highpoint_metres success  solo oxygen_used  died
## 1          Leader FALSE              NA   FALSE FALSE       FALSE FALSE
## 2   Deputy Leader FALSE            6000   FALSE FALSE       FALSE FALSE
## 3         Climber FALSE              NA   FALSE FALSE       FALSE FALSE
## 4      Exp Doctor FALSE            6000   FALSE FALSE       FALSE FALSE
## 5         Climber FALSE              NA   FALSE FALSE       FALSE FALSE
```

```
## 6         Climber FALSE             6000   FALSE FALSE       FALSE FALSE
##   death_cause death_height_metres injured injury_type injury_height_metres
## 1        <NA>                  NA   FALSE        <NA>                   NA
## 2        <NA>                  NA   FALSE        <NA>                   NA
## 3        <NA>                  NA   FALSE        <NA>                   NA
## 4        <NA>                  NA   FALSE        <NA>                   NA
## 5        <NA>                  NA   FALSE        <NA>                   NA
## 6        <NA>                  NA   FALSE        <NA>                   NA
```

```
# Résumé statistique des données
summary(donnees)
```

```
##  expedition_id       member_id           peak_id           peak_name
##  Length:76519       Length:76519       Length:76519       Length:76519
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##       year         season              sex                 age
##  Min.   :1905   Length:76519       Length:76519       Min.   : 7.00
##  1st Qu.:1991   Class :character   Class :character   1st Qu.:29.00
##  Median :2004   Mode  :character   Mode  :character   Median :36.00
##  Mean   :2000                                         Mean   :37.33
##  3rd Qu.:2012                                         3rd Qu.:44.00
##  Max.   :2019                                         Max.   :85.00
##                                                       NA's   :3497
##  citizenship        expedition_role      hired         highpoint_metres
##  Length:76519       Length:76519       Mode :logical   Min.   :3800
##  Class :character   Class :character   FALSE:60788     1st Qu.:6700
##  Mode  :character   Mode  :character   TRUE :15731     Median :7400
##                                                       Mean   :7471
##                                                       3rd Qu.:8400
##                                                       Max.   :8850
##                                                       NA's   :21833
##   success           solo          oxygen_used        died
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:47320     FALSE:76398     FALSE:58286     FALSE:75413
##  TRUE :29199     TRUE :121       TRUE :18233     TRUE :1106
##
##
##
##
##  death_cause        death_height_metres  injured        injury_type
##  Length:76519       Min.   : 400        Mode :logical   Length:76519
##  Class :character   1st Qu.:5800        FALSE:74806     Class :character
##  Mode  :character   Median :6600        TRUE :1713      Mode  :character
##                     Mean   :6593
##                     3rd Qu.:7550
##                     Max.   :8830
##                     NA's   :75451
##  injury_height_metres
##  Min.   : 400
```

```
##  1st Qu.:6200
##  Median :7100
##  Mean   :7050
##  3rd Qu.:8000
##  Max.   :8880
##  NA's   :75510
```

Les données présentent des informations détaillées sur les membres des expéditions dans l'Himalaya, incluant l'identité des expéditions et des membres, le pic visé, le nom, l'année, la saison, le sexe, l'âge, la nationalité, le rôle dans l'expédition, si le membre était engagé (hired), l'altitude atteinte, le succès de l'ascension, l'utilisation d'oxygène, si le membre est décédé et d'autres détails liés à des incidents. Les données s'étendent de 1905 à 2019 et révèlent des tendances sur des décennies d'expéditions alpines.

Age des membres d'une expédition réussie

On se pose la question suivante : "Comment se répartit l'âge des membres d'une expédition réussie vers le Mont Everest ?"

Décrivez l'expérience statistique (individu, population, échantillon, variable mesurée, ...).

Sélectionnez dans le tableau uniquement les lignes répondant à ces critères, et dont l'âge n'est pas manquant.

Représentez ces données sous la forme d'un histogramme. Justifiez le choix de la largeur des classes.

Représentez ces même données sous la forme d'une boîte à moustache (*boxplot*).

Laquelle de ces représentations est la plus informative ? Justifiez.

Que pouvez-vous dire sur l'âge des membres d'une expédition réussie vers le Mont Everest ?

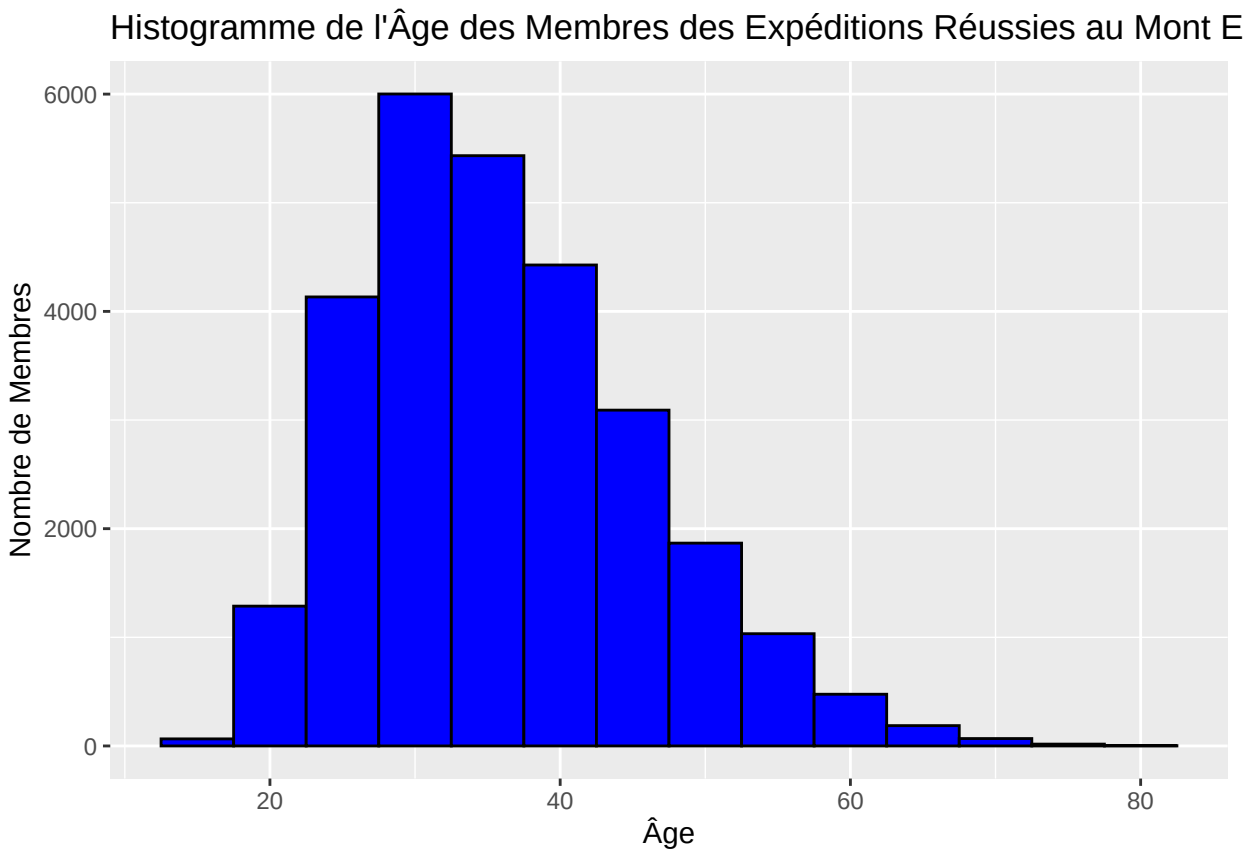
Sélection des Membres d'Expéditions Réussies avec Âge Connu

```r
donnees_reussies <- donnees %>%
  filter(success == TRUE, !is.na(age))
```

Description des données :

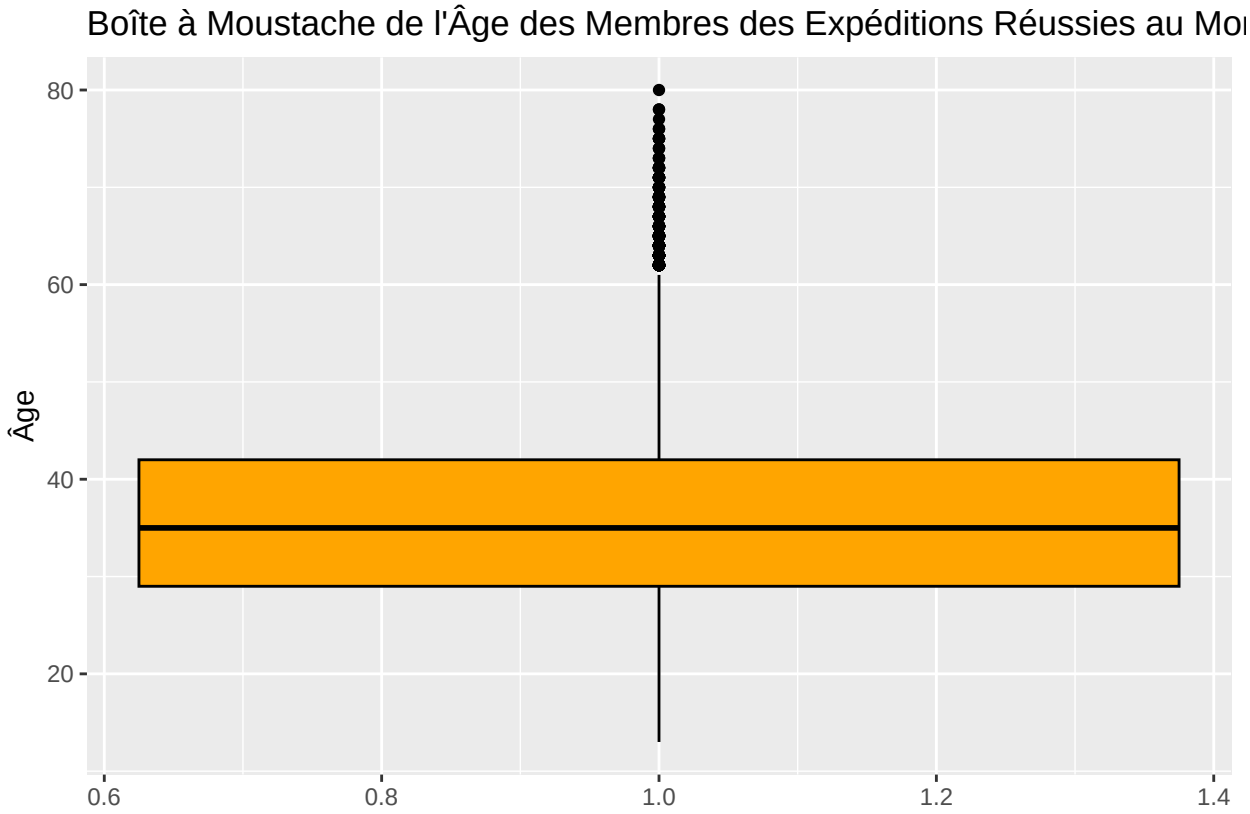
Histogramme de l'Âge

```r
ggplot(donnees_reussies, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  labs(title = "Histogramme de l'Âge des Membres des Expéditions Réussies au Mont Everest",
       x = "Âge",
       y = "Nombre de Membres")
```



Boîte à Moustache de l'Âge

```
ggplot(donnees_reussies, aes(y = age, x = 1)) +
  geom_boxplot(fill = "orange", color = "black") +
  labs(title = "Boîte à Moustache de l'Âge des Membres des Expéditions Réussies au Mont Everest",
       x = "",
       y = "Âge")
```



Conclusion

L'histogramme et la boîte à moustache montrent la distribution de l'âge des membres des expéditions réussies au Mont Everest. L'histogramme révèle une distribution unimodale avec la majorité des grimpeurs dans la tranche d'âge de 30 à 40 ans, tandis que la boîte à moustache met en évidence la médiane, les quartiles et les valeurs aberrantes. Bien que l'histogramme donne un aperçu détaillé de la distribution des fréquences, la boîte à moustache fournit une synthèse concise des tendances centrales et de la dispersion. Les deux graphiques suggèrent que les membres d'expéditions réussies sont généralement dans leur trentaine ou leur quarantaine, avec des exceptions notables aux deux extrémités de l'échelle d'âge.

Age en fonction des années d'ascension

En reprenant le même jeu de données (expéditions réussies vers le Mont Everest), on se pose la question suivante : "L'âge des membres d'une expédition réussie vers le Mont Everest change-t-il au cours du temps ?"

Pour répondre à cette question, reprenez le jeu de données de la section précédente, et représentez-le sous forme de boxplots, un par année.

Interprétez.

Intuition

Pour étudier la variation de l'âge des membres d'expéditions réussies au Mont Everest au fil des ans, il faudrait créer une série de boîtes à moustaches, chacune représentant une année différente. Les boîtes à

moustaches permettraient d'observer les tendances centrales et les dispersions de l'âge pour chaque année et de détecter d'éventuels changements ou tendances au fil du temps.

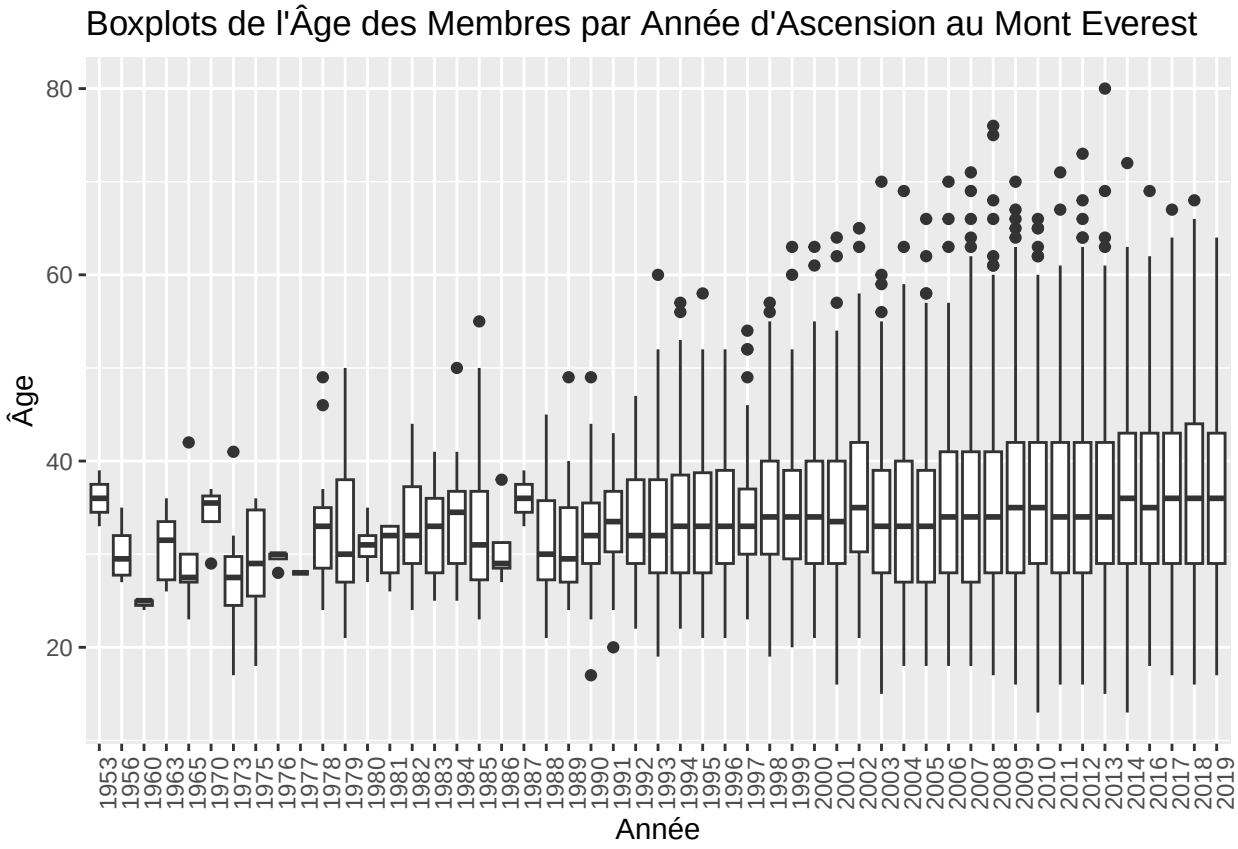
Si on observe que la médiane de l'âge se déplace vers le haut ou vers le bas au cours des années, ou que la dispersion des âges change, cela pourrait indiquer une tendance dans l'âge des alpinistes réussissant l'ascension. Par exemple, si les médianes augmentent avec le temps, on pourrait conclure que les alpinistes réussissant sont de plus en plus âgés. En revanche, si la dispersion augmente, cela pourrait suggérer que l'Everest attire une gamme d'âges plus large au fil des années.

Filtration des Données pour Expéditions Réussies avec Âge Connu

```
donnees_reussies <- donnees %>%
  filter(success == TRUE, !is.na(age), peak_name == "Everest")
```

Boxplots de l'Âge par Année d'Ascension

```
donnees_reussies %>%
  ggplot(aes(x = factor(year), y = age)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Boxplots de l'Âge des Membres par Année d'Ascension au Mont Everest",
       x = "Année",
       y = "Âge")
```



Interpretation

L'analyse de la boîte à moustaches par année indique que l'âge médian des membres d'expéditions réussies au Mont Everest est resté relativement stable au fil des décennies. La plupart des médianes se situent dans la trentaine, avec une légère tendance à la hausse dans les années récentes. Les gammes interquartiles et les valeurs aberrantes montrent que, bien que la majorité des grimpeurs soient de jeunes adultes à adultes d'âge moyen, il y a toujours eu une présence significative de membres plus âgés, et cette diversité d'âge semble s'être maintenue au fil du temps.

Age des membres d'une expédition réussie ou non

On se pose la question suivante : "Y-a-t-il une différence d'âge entre les membres d'une expédition réussie, et ceux d'une expédition qui a échouée, avec ou sans oxygène ?"

Décrivez l'expérience statistique. Sélectionnez dans le tableau uniquement les lignes répondant à ces critères, et dont l'âge n'est pas manquant.

Représentez ces même données sous la forme de boîtes à moustaches (*boxplot*). Il devrait y avoir en tout 4 boîtes, distinguées par des positions (en `x`) et des facettes. Pour créer et renommer les facettes, vous pourrez utiliser la commande suivante:

```
  ... +
  facet_wrap(success ~ .,
             labeller = as_labeller(c(`TRUE` = "Succès", `FALSE` = "Echec"))) +
  ...
```

Vous pourrez renommer les axes avec la fonction `scale_x_discrete`.

Y-a-t-il une différence dans la distribution des âges ?

Renseignez-vous sur les graphes "en violons" (*violin plot*, voir `?geom_violin`). Reproduisez le graphique précédent, mais avec des violons à la place des boîtes.

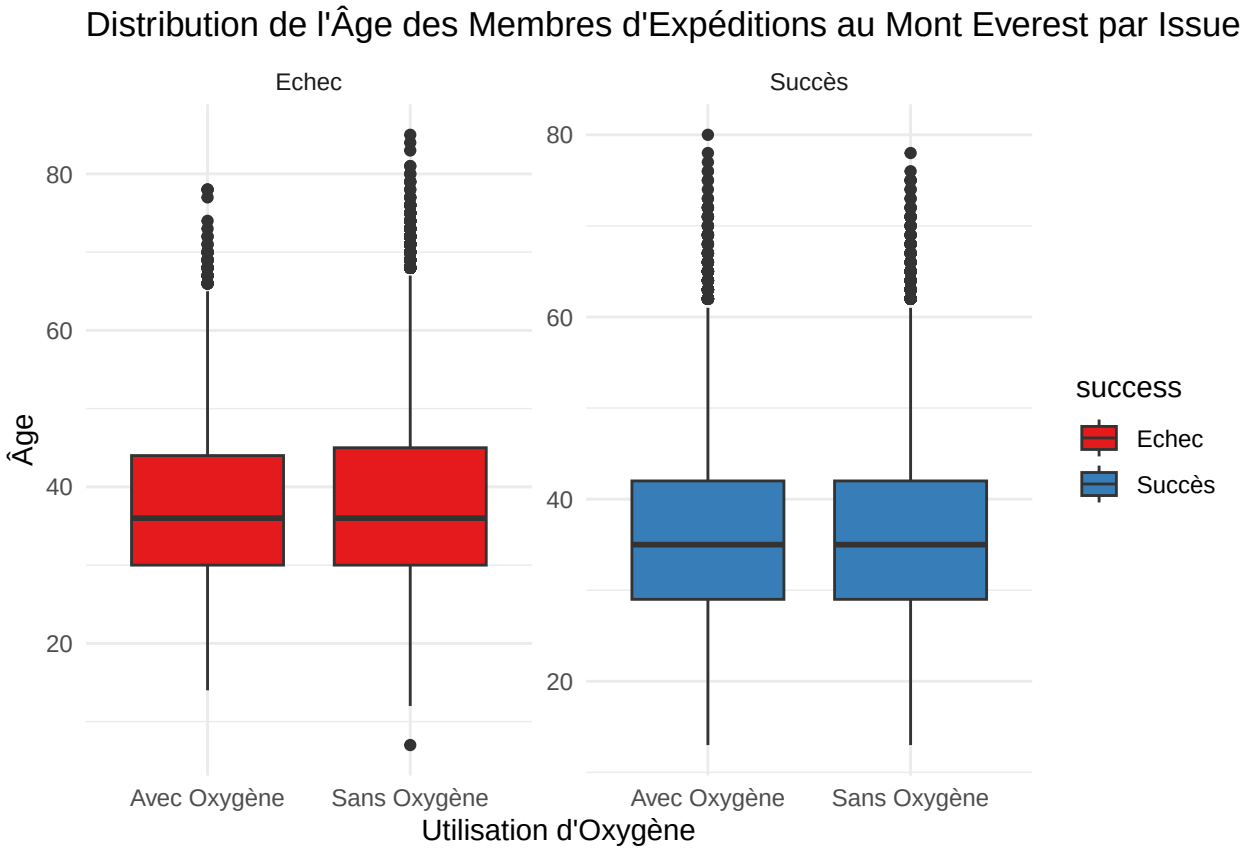
Cela change-t-il votre réponse à la question ? Quel graphique trouvez-vous le plus informatif ?

Sélection des Membres avec Âge Connu par Issue de l'Ascension et Utilisation d'Oxygène

```
donnees_filtrees <- donnees %>%
  filter(!is.na(age)) %>%
  mutate(success = ifelse(success == TRUE, "Succès", "Echec"),
         oxygen = ifelse(oxygen_used == TRUE, "Avec Oxygène", "Sans Oxygène"))
```

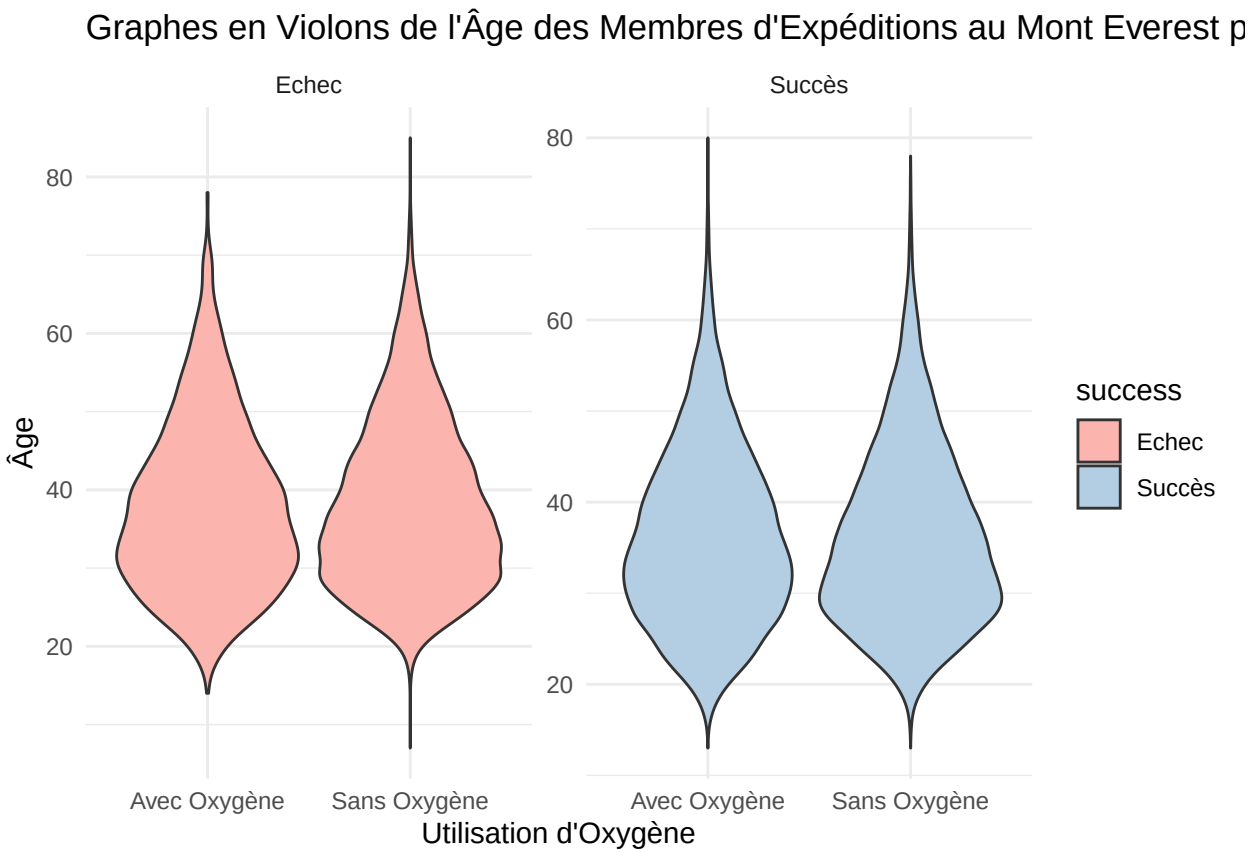
Boîtes à Moustaches de l'Âge par Réussite et Utilisation d'Oxygène

```
ggplot(donnees_filtrees, aes(x = oxygen, y = age, fill = success)) +
  geom_boxplot() +
  facet_wrap(~success, scales = "free", labeller = as_labeller(c("Succès" = "Succès", "Echec" = "Echec"
  labs(title = "Distribution de l'Âge des Membres d'Expéditions au Mont Everest par Issue et Utilisatio
       x = "Utilisation d'Oxygène",
       y = "Âge") +
  scale_fill_brewer(palette = "Set1") +
  theme_minimal()
```



Graphes en Violons de l'Âge par Réussite et Utilisation d'Oxygène

```
ggplot(donnees_filtrees, aes(x = oxygen, y = age, fill = success)) +
  geom_violin() +
  facet_wrap(~success, scales = "free", labeller = as_labeller(c("Succès" = "Succès", "Echec" = "Echec"
  labs(title = "Graphes en Violons de l'Âge des Membres d'Expéditions au Mont Everest par Issue et Utili
       x = "Utilisation d'Oxygène",
       y = "Âge") +
  scale_fill_brewer(palette = "Pastel1") +
  theme_minimal()
```



Conclusion

Dans le premier graphique, il semble y avoir une légère différence dans la distribution de l'âge entre les membres des expéditions réussies et celles qui ont échoué, à la fois pour les groupes utilisant de l'oxygène et ceux qui n'en utilisent pas. Les médianes sont similaires, mais la dispersion (l'étendue interquartile et les valeurs extrêmes) diffère légèrement, avec une tendance vers une dispersion plus grande pour les expéditions réussies.

Graphes en Violons par rapport aux Boxplots :

La représentation des données avec des graphes en violon apporte une dimension supplémentaire à l'analyse en illustrant non seulement les statistiques résumées (comme dans les boxplots) mais aussi la densité de la distribution des âges. Les "violons" montrent où les données sont plus concentrées, ce qui peut indiquer des plages d'âges plus communes. Implication des Violon Plots sur l'Analyse : Oui, cela change l'analyse. Les graphes en violon fournissent une vue plus complète de la distribution des âges, révélant des détails sur la concentration des âges qui ne sont pas évidents dans les boxplots.

Graphique le Plus Informatif :

Les graphes en violon sont plus informatifs pour comprendre la distribution complète des âges. Ils montrent non seulement la médiane et les quartiles mais aussi la densité des âges à différents niveaux, ce qui est particulièrement utile pour identifier les plages d'âges les plus courantes ou les plus rares au sein des groupes ## Autre question

Posez une autre question sur le jeu de données, et répondez-y à l'aide d'un graphique. Décrivez bien votre démarche statistique, le choix de la représentation, et les conclusions que vous en tirez.

Introduction

Dans cette analyse, nous cherchons à comprendre s'il existe une relation entre l'âge des grimpeurs et l'altitude maximale atteinte lors des expéditions au Mont Everest. Nous posons l'hypothèse que l'expérience acquise avec l'âge pourrait influencer la capacité à atteindre des altitudes plus élevées.

Méthodologie

Pour explorer cette question, nous utiliserons un scatter plot pour visualiser la corrélation potentielle entre l'âge et l'altitude maximale atteinte par les grimpeurs. Nous commencerons par nettoyer les données pour exclure les entrées incomplètes.

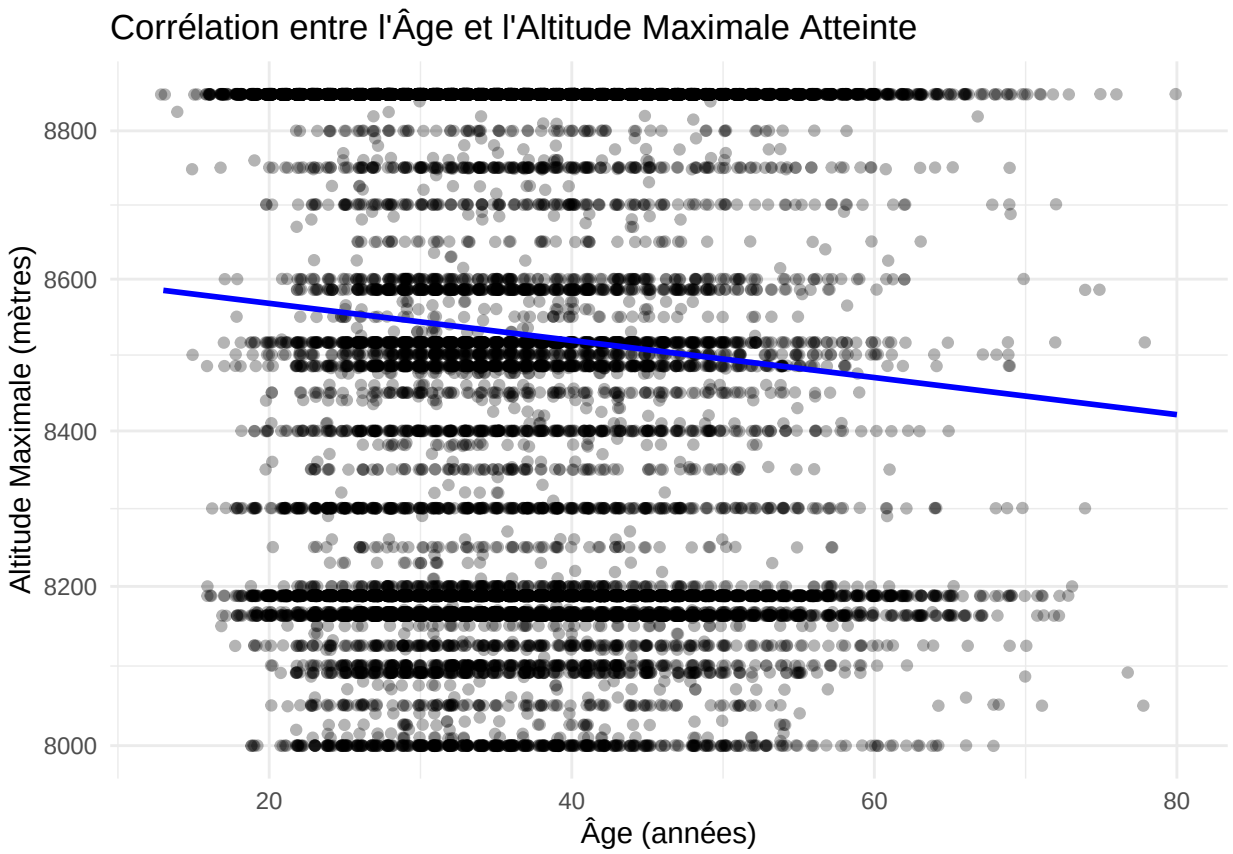
```
donnees_clean <- donnees %>%
  filter(!is.na(age) & !is.na(highpoint_metres) & highpoint_metres >= 8000)
```

Scatter Plot de l'Âge par rapport à l'Altitude Maximale Atteinte

Nous construisons ensuite un scatter plot avec une ligne de tendance pour évaluer visuellement la corrélation.

```
ggplot(donnees_clean, aes(x = age, y = highpoint_metres)) +
  geom_jitter(alpha = 0.3, size = 1.5, width = 0.25) + # Jittering avec transparence
  geom_smooth(method = "lm", se = FALSE, color = "blue") + # Ligne de tendance
  scale_y_continuous(trans = 'log10') +
  labs(title = "Corrélation entre l'Âge et l'Altitude Maximale Atteinte",
       x = "Âge (années)",
       y = "Altitude Maximale (mètres)") +
  theme_minimal() +
  theme(legend.position = "bottom")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Analyse et Conclusions

Après avoir créé le graphique, nous analyserons la disposition des points et la pente de la ligne de tendance. Un coefficient de corrélation sera également calculé pour quantifier la relation.

```
# Calcul du coefficient de corrélation de Pearson
correlation <- cor(donnees_clean$age, donnees_clean$highpoint_metres, method = "pearson")
correlation
```

```
## [1] -0.07344491
```

La valeur du coefficient de corrélation de Pearson est de -0.0734491, ce qui indique une très faible corrélation négative entre ces deux variables. Cela signifie qu'il n'y a pas de relation linéaire forte entre l'âge des grimpeurs et l'altitude atteinte.
