

Analyse des Combats de MMA dans l'UFC

IF36 - Les Yonkos

UTT - P25

Table des matières

Introduction.....	2
I. Données	3
1. Complete UFC Data	3
2. Pro MMA Fighters	3
II. Exploration	4
Question 1 : Durant quel round la majorité des combats se terminent ?.....	4
Question 2 : L'allonge est-elle un avantage significatif pour gagner un combat ?.....	5
Question 3 : Quel type de garde est la plus dominante (southpaw, orthodox...) ?	11
Question 4 : Les notions de favori et d'underdog sont-elles respectées réellement ? Les bookmakers ont-ils raison ?.....	15

Introduction

Pour notre projet, on a travaillé avec deux datasets sur les combats de MMA dans l'UFC. Le premier, "**Complete UFC Data**", contient des informations sur les événements, les combattants et leurs statistiques de combat, ainsi que les cotes des paris. Ce dataset couvre une période de **1994 à 2023**, ce qui nous donne beaucoup de matière pour analyser les performances et les tendances dans le MMA.

Le second dataset, "**Pro MMA Fighters**", se focalise plutôt sur les combattants eux-mêmes, avec leurs caractéristiques personnelles, physiques et leur bilan de carrière. Par contre, les données ne sont à jour que jusqu'en **2022**, donc un peu limité pour les analyses récentes.

On a choisi ces deux jeux de données pour explorer ce qui influence les résultats des combats et pour comprendre l'impact de différents facteurs comme les attributs physiques, les styles de combat et les cotes des paris.

Pour les premières questions, on a principalement utilisé le premier dataset, tandis que le second nous servira pour des problématiques plus complexes.

I. Données

1. Complete UFC Data

Ce dataset contient des infos détaillées sur les combats de l'UFC : - Les événements (date, lieu, etc.) - Les combattants (nom, âge, taille, poids, etc.) - Les statistiques de combat (coups portés, takedowns, etc.) - Les cotes des paris

Les données couvrent **1994 à 2023**, ce qui nous permet d'analyser les tendances sur une bonne période de temps.

2. Pro MMA Fighters

Ce dataset se concentre sur les combattants : - Infos personnelles (âge, nationalité, etc.) - Attributs physiques (taille, poids, allonge, etc.) - Bilan de carrière (victoires, défaites, etc.)

Les données sont disponibles jusqu'en **2022**, donc pas très à jour pour les analyses récentes.

II. Exploration

Voici les questions qu'on a essayé de résoudre avec nos visualisations et nos interprétations.

D'abord, on charge les bibliothèques et les données dont on a besoin :

```
library(ggplot2)
library(dplyr)
library(tidyr)
library(scales)
library(readr)
library(stringr)

data <- read_csv("data/complete_ufc_data.csv", show_col_types = FALSE)
```

Question 1 : Durant quel round la majorité des combats se terminent ?

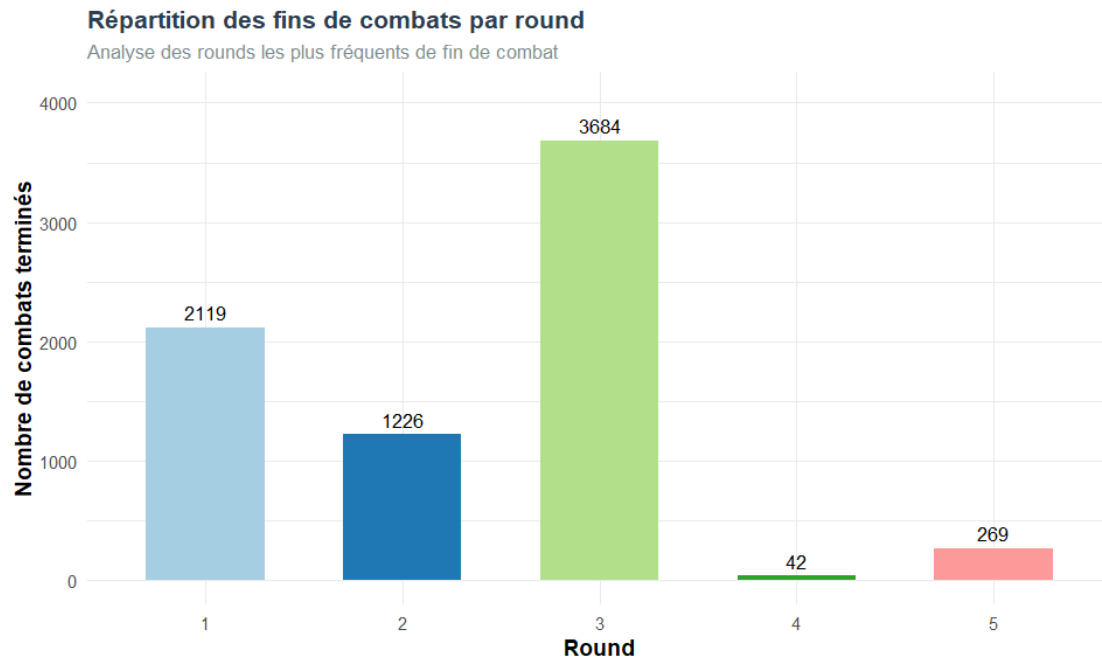
Nous supposons que la majorité des rounds se terminent vers le milieu du combat : au 2ème ou 3ème round. Il y a aussi sûrement beaucoup de KO où le combat se finit vite.

Commençons par nettoyer les données pour la question : enlever les NA et convertir la colonne 'round' en facteur.

```
df_clean <- data %>%
  filter(!is.na(round)) %>%
  mutate(round = as.factor(round))
```

Visualisation : Histogramme de répartition des fins de combats par round

```
ggplot(df_clean, aes(x = round)) +
  geom_bar(aes(fill = round), width = 0.6, show.legend = FALSE) +
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5, size = 4.5)
+
  scale_fill_brewer(palette = "Paired") +
  labs(
    title = "Répartition des fins de combats par round",
    subtitle = "Analyse des rounds les plus fréquents de fin de combat",
    x = "Round",
    y = "Nombre de combats terminés",
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", size = 16, color = "#2c3e50"),
    plot.subtitle = element_text(size = 13, color = "#7f8c8d"),
    axis.title.x = element_text(face = "bold"),
    axis.title.y = element_text(face = "bold")
  ) +
  expand_limits(y = max(table(df_clean$round)) * 1.1) # ajoute de l'espace
  au-dessus des barres
```



Interprétation : Le graphique permet de répondre clairement à la question posée, le round 3 est celui où la majorité des combats se terminent. Cela montre que les combats ne s'arrêtent pas forcément sur un KO ou une soumission précoce comme certains qui pourraient le penser.

Conclusion question 1

D'après le graphique, **le round 3 est le round de fin le plus fréquent.**

On observe également que plusieurs combats prennent fin lors des deux premiers rounds.

Néanmoins, les rounds 4 et 5 s'avèrent être beaucoup plus rares, ce qui est logique puisqu'ils ne concernent généralement que les combats principaux (*main events* ou *title fights*). De même, les combats semblent durer au moins un certain temps avant de se conclure, ce qui peut représenter des affrontements équilibrés, des stratégies de fatigue progressive ou des combats à l'usure.

Question 2 : L'allonge est-elle un avantage significatif pour gagner un combat ?

Nous cherchons à déterminer si l'allonge représente un avantage significatif pour remporter un combat et comment cet avantage potentiel peut se manifester dans les stratégies employées par les combattants. Nous tenons l'hypothèse que l'allonge est un facteur très important et qu'elle impacte la stratégie des combattants.

Visualisation 1: Proportion de victoires par catégorie de différence d'allonge

Calcul de la différence d'allonge

```
data$reach_diff <- data$fighter1_reach - data$fighter2_reach
```

Création de catégories de différence d'allonge (valeur absolue)

```
data$reach_category <- cut(abs(data$reach_diff),
```

```

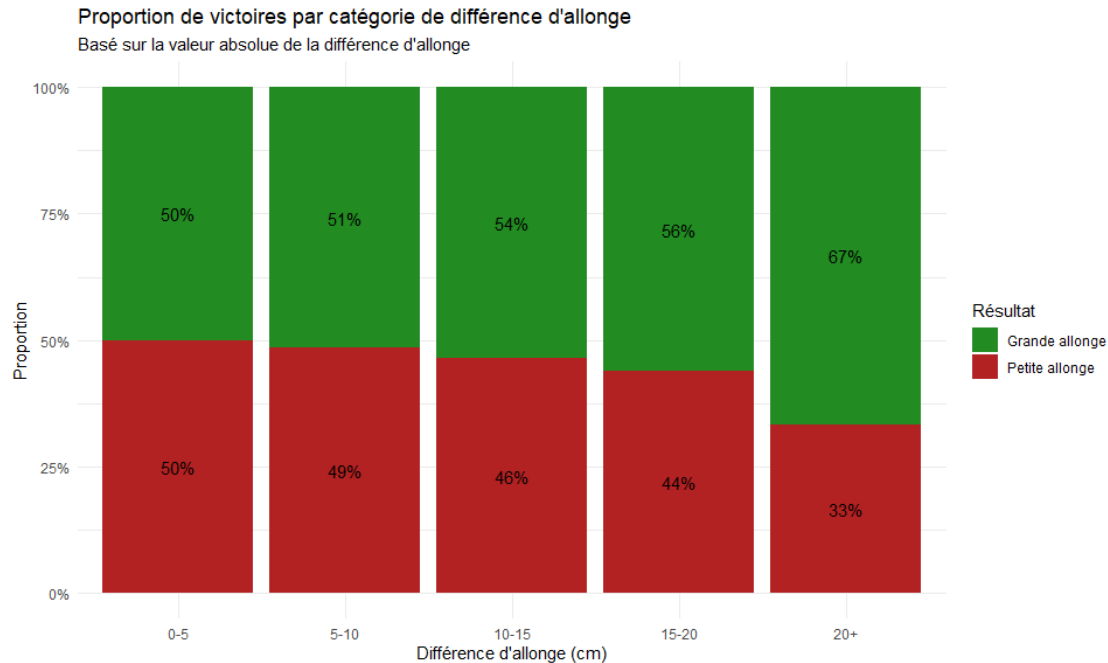
breaks = c(0, 5, 10, 15, 20, 25),
labels = c("0-5", "5-10", "10-15", "15-20",
"20+"))

# Calcul des proportions de victoires par catégorie
victory_proportion <- data %>%
  filter(!is.na(reach_category)) %>%
  group_by(reach_category) %>%
  summarize(
    total_combats = n(),
    victoires_plus_grande_allonge = sum(reach_diff > 0),
    victoires_plus_petite_allonge = sum(reach_diff < 0),
    proportion_victoire_grande_allonge = victoires_plus_grande_allonge /
total_combats,
    proportion_victoire_petite_allonge = victoires_plus_petite_allonge /
total_combats
  )

# Restructuration pour la visualisation
victory_data <- victory_proportion %>%
  select(reach_category,
    `Grande allonge` = proportion_victoire_grande_allonge,
    `Petite allonge` = proportion_victoire_petite_allonge) %>%
  pivot_longer(cols = c(`Grande allonge`, `Petite allonge`),
    names_to = "type_victoire",
    values_to = "proportion")

# Création du barplot empilé
ggplot(victory_data, aes(x = reach_category, y = proportion, fill =
type_victoire)) +
  geom_bar(stat = "identity", position = "stack") +
  geom_text(aes(label = scales::percent(proportion, accuracy = 1)),
    position = position_stack(vjust = 0.5)) +
  labs(title = "Proportion de victoires par catégorie de différence
d'allonge",
    subtitle = "Basé sur la valeur absolue de la différence d'allonge",
    x = "Différence d'allonge (cm)",
    y = "Proportion",
    fill = "Résultat") +
  scale_fill_manual(values = c("Grande allonge" = "forestgreen",
    "Petite allonge" = "firebrick")) +
  scale_y_continuous(labels = scales::percent) +
  theme_minimal()

```



Interprétation : Ce graphique montre, pour chaque catégorie de différence d'allonge, quelle proportion de combats est remportée par des combattants avec une allonge supérieure ou inférieure. Il permet de se rendre compte que l'avantage d'allonge devient de plus en plus significatif sur la victoire dans un combat à mesure qu'elle augmente.

Visualisation 2 : Nombre brut de victoires selon l'avantage d'allonge

Pour compléter notre analyse, examinons également le nombre brut de victoires selon l'avantage d'allonge.

```
# Créer une variable catégorielle pour l'avantage d'allonge
data$reach_advantage <- ifelse(data$reach_diff > 0,
                                "Gagnant avec allonge supérieure",
                                ifelse(data$reach_diff < 0,
                                        "Gagnant avec allonge inférieure",
                                        "Allonge égale"))

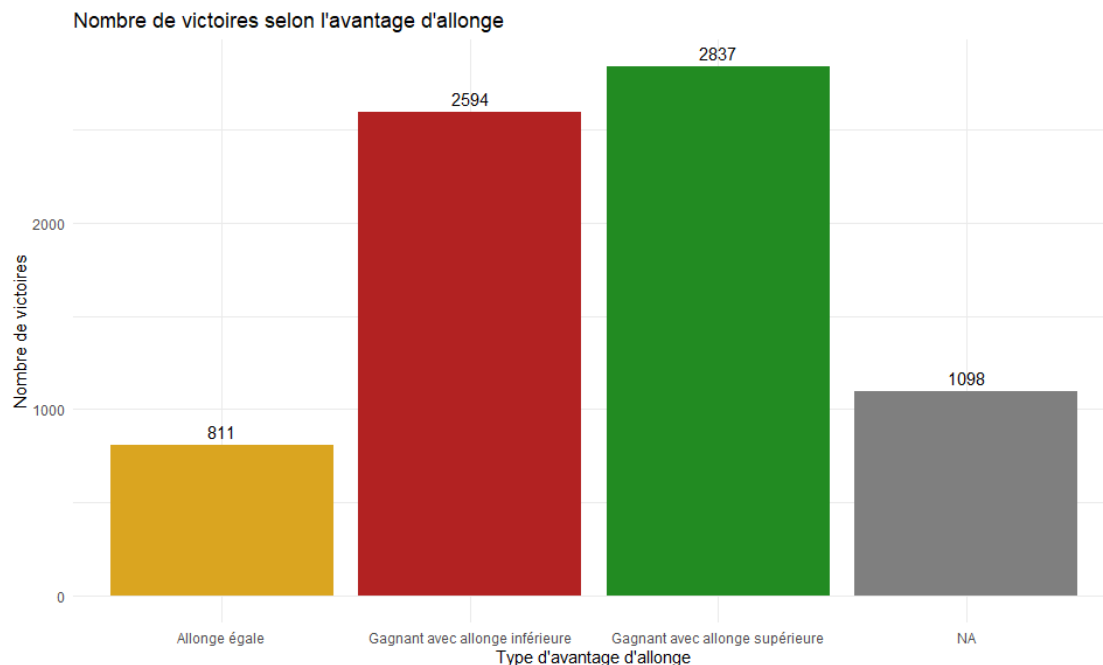
# Compter le nombre de combats dans chaque catégorie
reach_victory_counts <- data %>%
  group_by(reach_advantage) %>%
  summarize(
    nombre_victoires = n()
  )

# Créer le barplot
ggplot(reach_victory_counts, aes(x = reach_advantage, y = nombre_victoires,
fill = reach_advantage)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = nombre_victoires), vjust = -0.5) +
  labs(title = "Nombre de victoires selon l'avantage d'allonge",
```

```

x = "Type d'avantage d'allonge",
y = "Nombre de victoires") +
scale_fill_manual(values = c("Gagnant avec allonge supérieure" =
"forestgreen",
"Gagnant avec allonge inférieure" =
"firebrick",
"Allonge égale" = "goldenrod")) +
theme_minimal() +
theme(legend.position = "none")

```



Interprétation : On voit ici que ceux ayant une allonge supérieure gagnent plus souvent que leurs adversaires, avec une allonge plus faible, même si le résultat n'est pas unanime.

Visualisation 3: Impact de l'allonge sur les coups significatifs portés

Ce graphique analyse la relation entre la différence d'allonge du fighter 1 face au fighter 2 (en pourcentage) et la différence de coups significatifs portés par minute entre les combattants. On pourrait penser que plus un combattant possède une grande allonge, plus il l'utiliserait pour porter des coups à son adversaire.

```

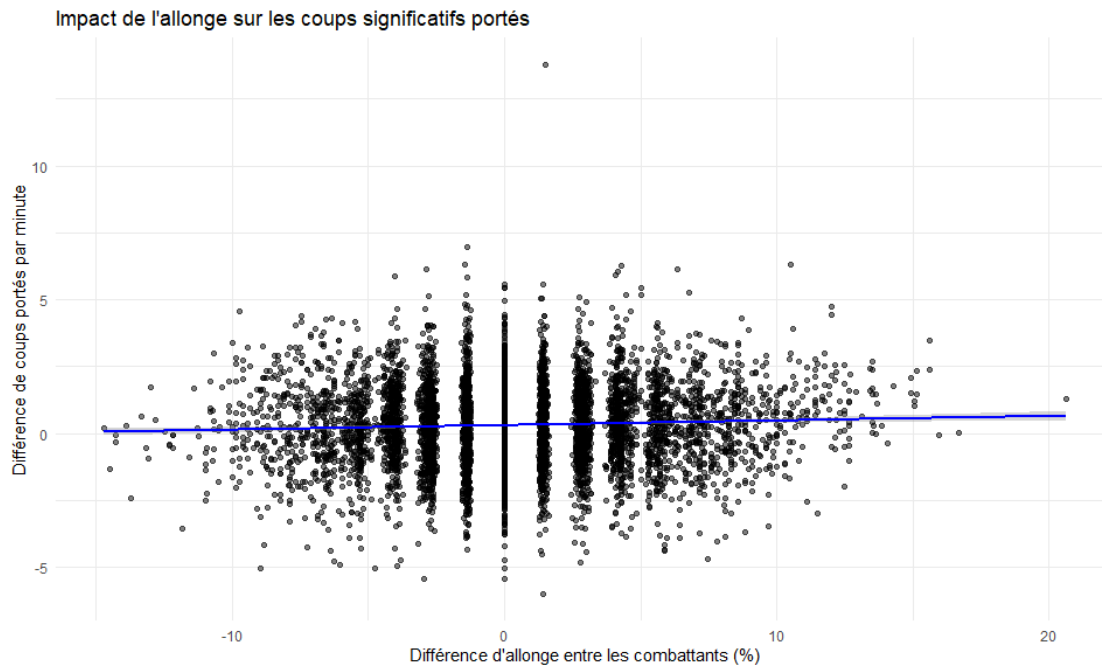
# Calcul des différences relatives d'allonge et de striking
data$reach_diff_pct <- (data$fighter1_reach - data$fighter2_reach) /
data$fighter2_reach * 100
data$strike_diff <- data$fighter1_sig_strikes_landed_pm -
data$fighter2_sig_strikes_landed_pm

# Graphique de dispersion avec ligne de régression
ggplot(data, aes(x = reach_diff_pct, y = strike_diff)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  labs(title = "Impact de l'allonge sur les coups significatifs portés",

```



```
x = "Différence d'allonge entre les combattants (%)",
y = "Différence de coups portés par minute" +
theme_minimal()
```



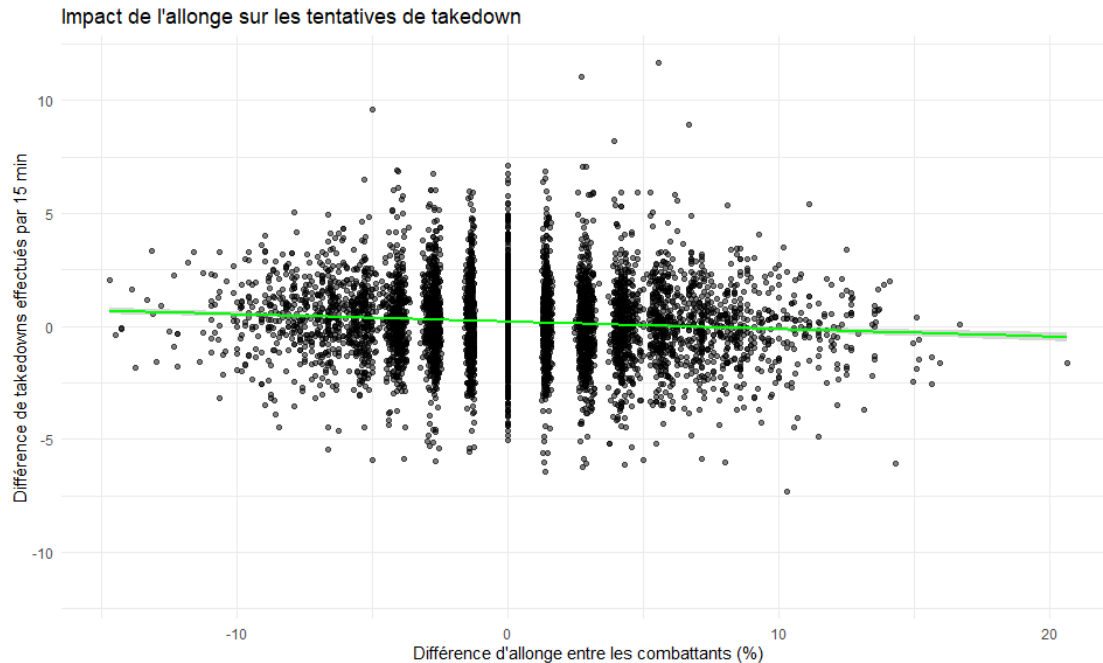
Interprétation : Cette visualisation nous permet de nous rendre compte qu'une plus grande allonge ne se traduit pas forcément par un avantage en termes de coups portés. La ligne de régression indique la tendance générale (la pente) est très légèrement positive. L'allonge a donc un impact, mais il est plus léger que ce qu'on pourrait penser.

Visualisation 4: Impact de l'allonge sur les tentatives de takedown

Cette visualisation examine, de la même manière, la relation entre la différence d'allonge du fighter 1 face au fighter 2 (en pourcentage) et la différence de takedowns tentés par 15 minutes entre les combattants. On pourrait penser qu'un combattant ayant une plus petite allonge que son adversaire va tenter de l'emmener au sol, afin de compenser le déséquilibre debout.

```
# Calcul de la différence de takedowns
data$takedown_diff <- data$fighter1_takedown_avg_per15m -
data$fighter2_takedown_avg_per15m

# Graphique de dispersion avec ligne de régression
ggplot(data, aes(x = reach_diff_pct, y = takedown_diff)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = TRUE, color = "green") +
  labs(title = "Impact de l'allonge sur les tentatives de takedown",
       x = "Différence d'allonge entre les combattants (%)",
       y = "Différence de takedowns effectués par 15 min") +
  theme_minimal()
```



Interprétation : Ce graphique nous permet de voir déjà plus clairement la stratégie utilisée par les combattants concernant les takedowns. La pente est négative, elle indique donc que les combattants avec une allonge inférieure compensent en tentant plus de projections au sol. Notre hypothèse est donc vérifiée.

Conclusion question 2

Ces visualisations nous permettent d'analyser sous différents angles l'impact de l'allonge sur les performances et les résultats des combats UFC. Nous pouvons désormais répondre à nos questions initiales:

1. **À quel point l'allonge peut être un avantage significatif pour gagner un combat?**
 - Les visualisations 1 et 2 offrent des perspectives claires sur cette question, en montrant la différence entre gagnants et perdants ainsi que les proportions de victoires par catégorie d'allonge. On se rend compte que plus la différence d'allonge est grande, plus elle a un impact sur la finalité du combat
2. **Un combattant devrait-il accepter un combat face à un adversaire disposant d'une plus grande allonge?**
 - La visualisation 1 est particulièrement utile pour répondre à cette question, car elle montre les proportions de victoires en fonction de l'écart d'allonge. Un combattant ne devrait ainsi pas accepter un combat lorsqu'il a une allonge plus de 20 centimètres plus petite que son adversaire, les chances de victoire étant grandement réduites.
3. **Avoir une grande allonge pousse-t-elle le combattant à utiliser son striking ou à tenter des takedowns?**
 - Les visualisations 3 et 4 explorent cette question en montrant comment l'allonge influence les stratégies de frappe et de projection. Elle n'est pas si

significative pour les coups, mais on se rend compte qu'une plus petite allonge engendre un nombre de takedowns plus élevé.

Question 3 : Quel type de garde est la plus dominante (southpaw, orthodox...) ?

Cette question vise à découvrir si l'utilisation d'une stance (Garde de combat) peut être un facteur de réussite lors des combats.

Dans le dataset, les stances sont réparties en deux colonnes : fighter1_stance et fighter2_stance. On va donc commencer par regrouper ces deux colonnes en une seule et pour chacune des lignes on indiquera si l'utilisation de la stance a mené à une victoire ou à une défaite.

```
winners <- data %>%
  select(stance = fighter1_stance) %>%
  mutate(result = "Victoire")

losers <- data %>%
  select(stance = fighter2_stance) %>%
  mutate(result = "Défaite")

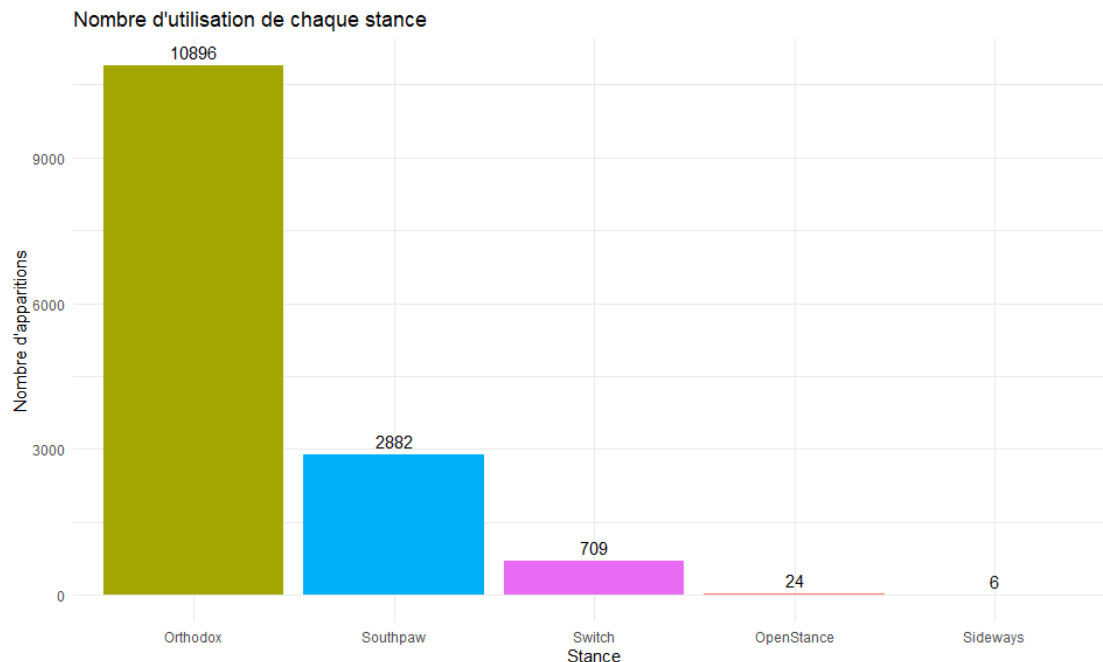
stances <- bind_rows(winners, losers)
stances <- stances %>%
  filter(!is.na(stance) & stance != "")
```

Visualisation 1 : Nombre d'utilisation de chaque stance

Il est difficile d'établir un véritable classement ici, mais la garde orthodoxe devrait logiquement être la plus utilisée. C'est la plus répandue, notamment en boxe où elle est très courante.

```
stance_usage <- stances %>%
  group_by(stance) %>%
  summarise(count = n()) %>%
  arrange(desc(count))

ggplot(stance_usage, aes(x = reorder(stance, -count), y = count, fill =
stance)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = count), vjust = -0.5, size = 4) +
  labs(
    title = "Nombre d'utilisation de chaque stance",
    x = "Stance",
    y = "Nombre d'apparitions"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```



Interprétation : Ce graphique démontre que l'écrasante majorité des combats utilise une garde orthodoxe. Le seul argument qui pourrait nous empêcher de considérer la garde orthodoxe comme étant la meilleure garde à utiliser serait que la southpaw ou la switch possède un taux de victoires très élevé. En revanche, on peut d'ores et déjà éliminer les autres gardes de notre analyse car leur nombre d'utilisation est négligeable.

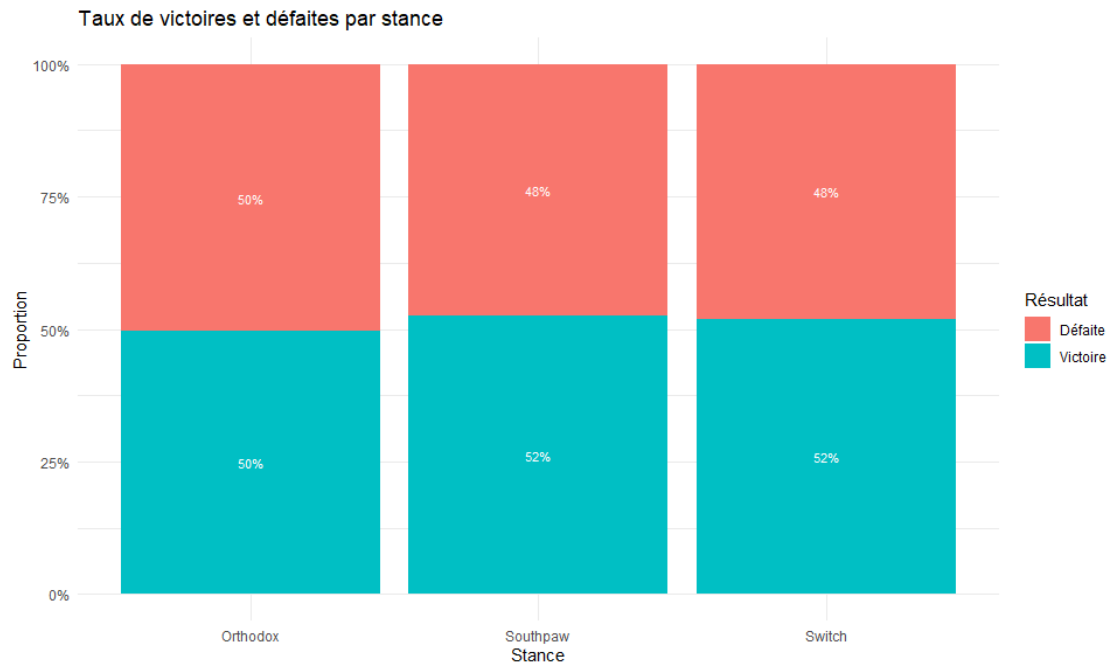
Visualisation 2 : Taux de victoire

On cherche donc le taux de victoire des trois gardes les plus utilisées (Orthodox, Southpaw et Switch). On pourrait supposer, par sa grande domination, que l'orthodox a un taux de victoire plus élevé.

```
stance_summary <- stances %>%
  filter(stance %in% c("Orthodox", "Switch", "Southpaw")) %>%
  group_by(stance, result) %>%
  summarise(count = n()) %>%
  group_by(stance) %>%
  mutate(percent = count / sum(count)) %>%
  ungroup()

ggplot(stance_summary, aes(x = stance, y = percent, fill = result)) +
  geom_bar(stat = "identity", position = "fill") +
  geom_text(aes(label = scales::percent(percent, accuracy = 1)),
            position = position_fill(vjust = 0.5),
            size = 3, color = "white") +
  scale_y_continuous(labels = scales::percent) +
  labs(
    title = "Taux de victoires et défaites par stance",
    x = "Stance",
    y = "Proportion",
```

```
fill = "Résultat"
) +
theme_minimal()
```



Interprétation : Ici, on peut voir qu'aucune des gardes ne garantit un taux de victoire significativement supérieur à 50%.

Visualisation 3 : Taux de victoire entre les gardes

Essayons donc de voir si une de ces gardes a un avantage significatif sur une autre garde.

```
df_filtered <- data %>%
  filter(fighter1_stance %in% c("Orthodox", "Southpaw", "Switch"),
         fighter2_stance %in% c("Orthodox", "Southpaw", "Switch"))

# Compter Les matchs gagnés par Fighter1
stance_matchups <- df_filtered %>%
  group_by(fighter1_stance, fighter2_stance) %>%
  summarise(victory_count = n()) %>%
  ungroup()

# Créer toutes les confrontations
df_duels <- df_filtered %>%
  mutate(matchup = paste(pmin(fighter1_stance, fighter2_stance),
                          pmax(fighter1_stance, fighter2_stance),
                          sep = " vs "))

# Total de combats par matchup
```

```

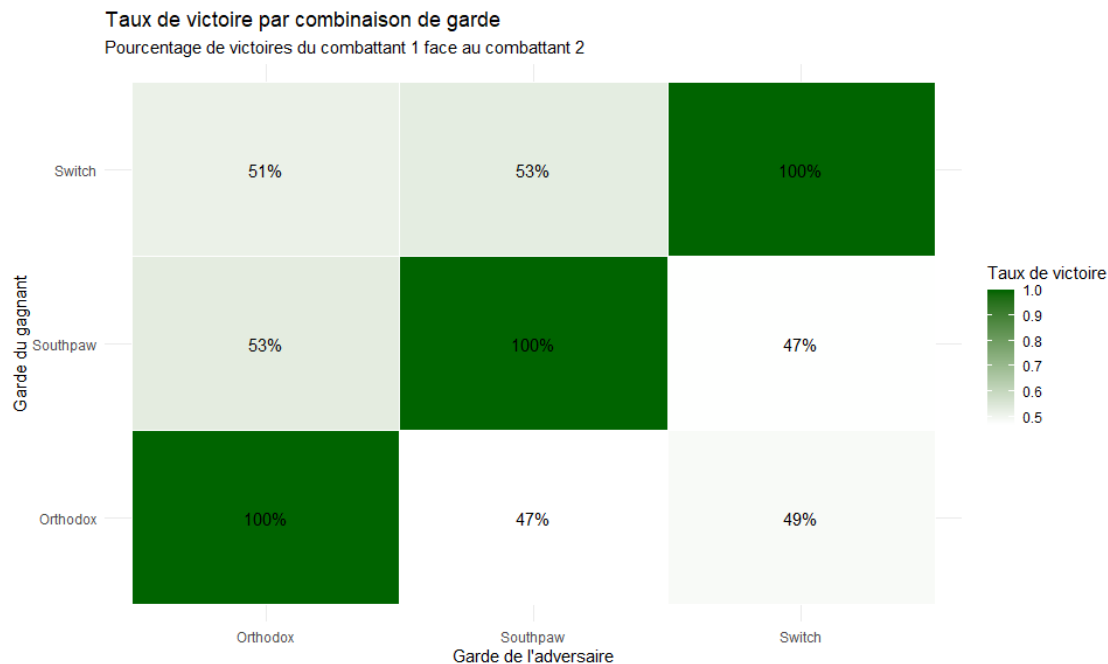
total_fights <- df_duels %>%
  group_by(matchup) %>%
  summarise(total = n())

# Victoires par combinaison exacte
wins <- df_duels %>%
  group_by(fighter1_stance, fighter2_stance) %>%
  summarise(victory = n()) %>%
  mutate(matchup = paste(pmin(fighter1_stance, fighter2_stance),
                          pmax(fighter1_stance, fighter2_stance),
                          sep = " vs ")) %>%
  left_join(total_fights, by = "matchup") %>%
  mutate(victory_rate = victory / total)

# Pour affichage correct dans la heatmap
wins_plot <- wins %>%
  select(fighter1_stance, fighter2_stance, victory_rate)

ggplot(wins_plot, aes(x = fighter2_stance, y = fighter1_stance, fill =
victory_rate)) +
  geom_tile(color = "white") +
  geom_text(aes(label = scales::percent(victory_rate, accuracy = 1)), color =
"black", size = 4) +
  scale_fill_gradient(low = "white", high = "darkgreen") +
  labs(
    title = "Taux de victoire par combinaison de garde",
    subtitle = "Pourcentage de victoires du combattant 1 face au combattant
2",
    x = "Garde de l'adversaire",
    y = "Garde du gagnant",
    fill = "Taux de victoire"
  ) +
  theme_minimal()

```



Interpretation : Aucune domination significative ne peut être notée. Il semble que la garde n'est pas l'élément qui influence le plus l'issue d'un combat.

Conclusion question 3

Finalement, on pourrait considérer la garde orthodoxe comme étant la meilleure garde à utiliser. Bien qu'elle possède le même taux de réussite que les autres gardes, elle est tout de même beaucoup plus utilisée. Cela veut donc dire que numériquement elle permet à un nombre important de combattants de remporter leur combat. Une analyse plus profonde serait d'étudier les gardes utilisées selon le niveau des combattants.

Question 4 : Les notions de favori et d'underdog sont-elles respectées réellement ? Les bookmakers ont-ils raison ?

Commençons par nettoyer le dataset pour cette question. Nous avons besoin des combats où les cotes sont existantes et le résultat l'est aussi.

```
odds_df <- data %>%
  filter(
    !is.na(favourite_odds),
    !is.na(underdog_odds),
    betting_outcome %in% c("favourite", "underdog")
  )
```

Visualisation 1 :

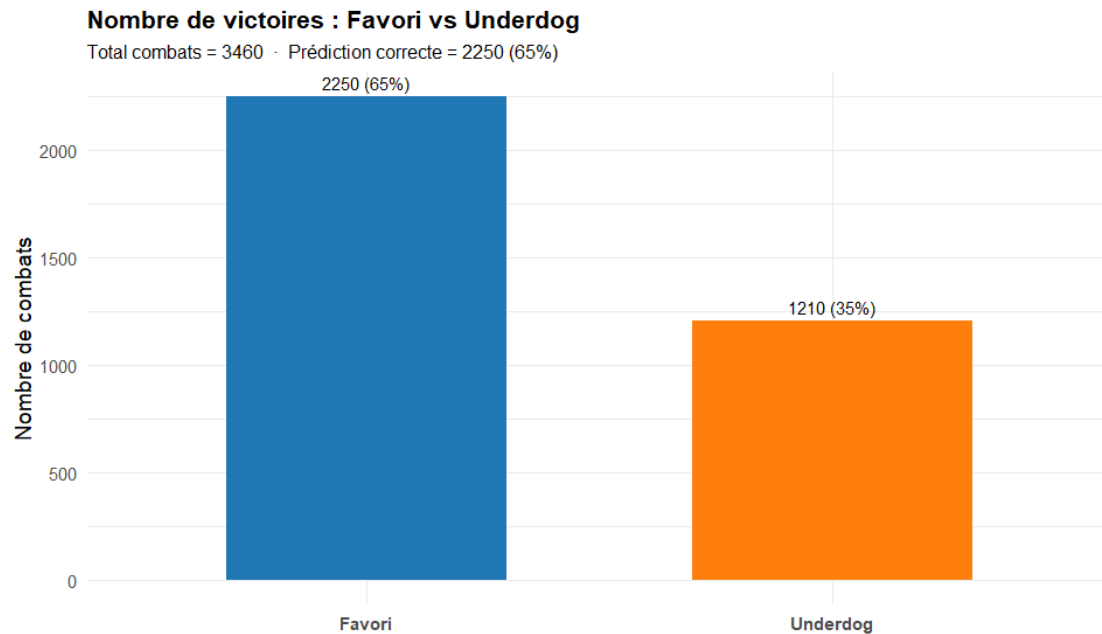
Intuitivement, nous nous attendons à ce que les favoris remportent la majorité des combats, sinon les cotes n'auraient pas de sens.

```

win_summary <- odds_df %>%
  count(betting_outcome) %>%
  mutate(
    percent = n / sum(n) * 100,
    label = str_c(n, " (", round(percent, 1), "%)")
  )

ggplot(win_summary, aes(x = betting_outcome, y = n, fill = betting_outcome))
+
  geom_col(width = 0.6, show.legend = FALSE) +
  scale_x_discrete(labels = c("favourite" = "Favori", "underdog" =
"Underdog")) +
  scale_fill_manual(
    values = c("favourite" = "#1f77b4", "underdog" = "#ff7f0e")
  ) +
  labs(
    title = "Nombre de victoires : Favori vs Underdog",
    subtitle = str_c(
      "Total combats = ", sum(win_summary$n),
      ". ",
      "Prédiction correcte = ", win_summary$label[win_summary$betting_outcome
== "favourite"]
    ),
    x = "",
    y = "Nombre de combats"
  ) +
  geom_text(aes(label = label), vjust = -0.5) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(size = 12),
    axis.text.x = element_text(face = "bold")
  )

```

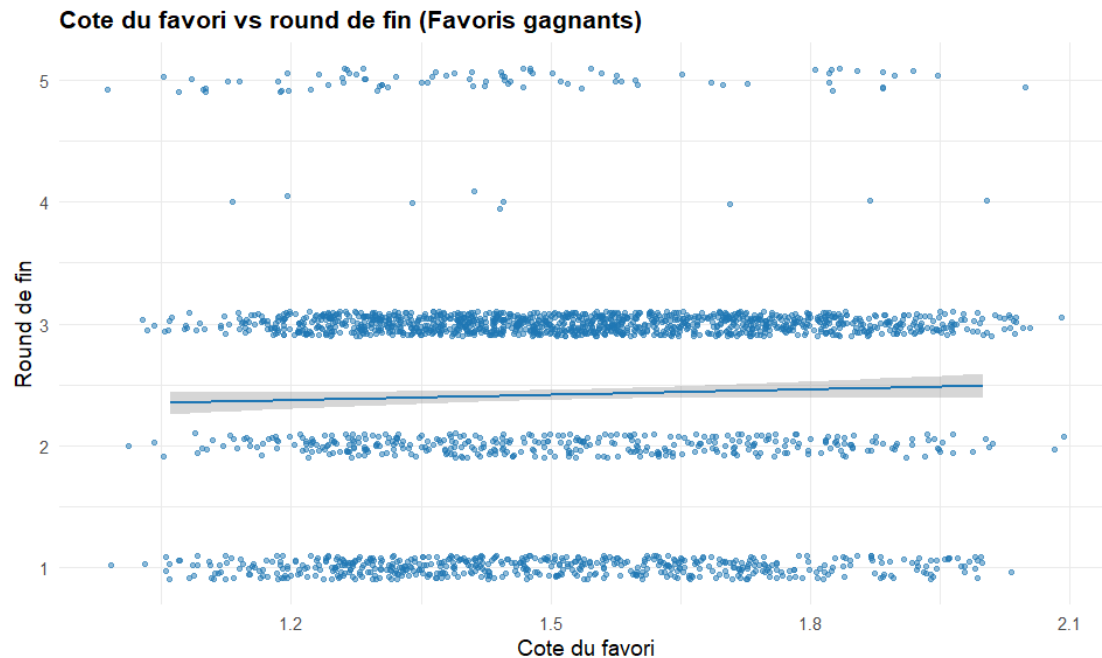
Interprétation : Le graphique montre clairement que les favoris remportent environ 65% des combats, contre 35% pour les underdogs. Cette différence significative confirme que les bookmakers ont globalement raison dans leurs prédictions. Un taux de précision de 65% est considérable dans un sport aussi imprévisible que le MMA, où un seul coup peut changer l'issue d'un combat, néanmoins il aurait été facile d'imaginer que ce nombre soit plus grand. En effet, le fait que les underdogs gagnent dans plus d'un tiers des cas soulève une question intéressante: existe-t-il des situations particulières où les underdogs ont plus de chances de l'emporter ?

Visualisation 2 : Relation entre côtes et durée du combat

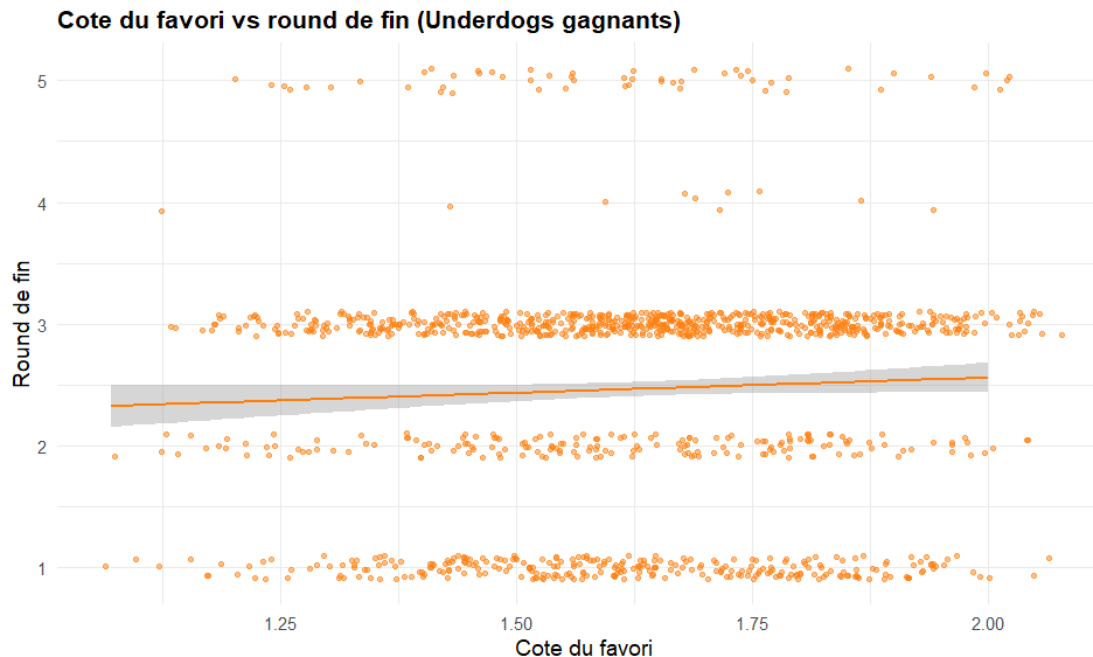
Nous cherchons à quantifier cette tendance. Nous avons déjà tracé précédemment la durée moyenne des combats (qui était de 3 rounds), y a t-il une différence majeure si l'on prend en considération les côtes ? Ici, on s'attendrait à ce que les combats soient plus longs si les côtes étaient hautes (adversaires égaux).

```
# Scatterplot cotes vs round pour Les favoris gagnants
ggplot(odds_df %>% filter(betting_outcome == "favourite", favourite_odds <=
2),
  aes(x = favourite_odds, y = round)) +
  geom_jitter(width = 0.1, height = 0.1, alpha = 0.5, color = "#1f77b4") +
  geom_smooth(method = "lm", se = TRUE, color = "#1f77b4") +
  labs(
    title = "Cote du favori vs round de fin (Favoris gagnants)",
    x = "Cote du favori",
    y = "Round de fin"
  ) +
  theme_minimal(base_size = 14) +
  theme(
```

```
plot.title = element_text(face = "bold", size = 16)
)
```



```
# Scatterplot cotes vs round pour Les underdogs gagnants
ggplot(odds_df %>% filter(betting_outcome == "underdog", favourite_odds <=
2),
  aes(x = favourite_odds, y = round)) +
  geom_jitter(width = 0.1, height = 0.1, alpha = 0.5, color = "#ff7f0e") +
  geom_smooth(method = "lm", se = TRUE, color = "#ff7f0e") +
  labs(
    title = "Cote du favori vs round de fin (Underdogs gagnants)",
    x = "Cote du favori",
    y = "Round de fin"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", size = 16)
  )
)
```



Interprétation : On a ici tracé la durée des combats en fonction des cotes du favori lorsque le favori gagne (graphique 1) ou lorsque l'underdog gagne (graphique 2). Les lignes de tendance ne montre pas une grande évolution (surtout en considérant l'intervalle de confiance), mais il est tout de même possible d'en tirer des potentielles conclusions intéressantes :

- Sur le premier graphique, on voit que plus les cotes augmentent, plus la durée du combat semble augmenter. Cela confirme notre hypothèse de départ : si la différence de niveau est moins grande, les combats durent plus longtemps (de peu).
- Sur le deuxième graphique, on observe les rounds où l'underdog a gagné selon les cotes du favori. Ici, la tendance est la même que sur le premier graphique (voir plus marqué vers la croissance), ce qui est assez surprenant. Ainsi, si la différence de niveau est grande, l'underdog a plus de chances de gagner avec un combat rapide. Cela est sûrement dû à l'effet de surprise et au fait de ne pas laisser le favori imposer son rythme.

Visualisation 3 : Précision des cotes

Nous cherchons maintenant à voir la précision des cotes. En moyenne, avoir une cote plus basse signifie-t-il effectivement avoir plus de chances de gagner ? Bien évidemment, c'est ce qui est attendu.

```
# Taux de réussite par cotes
odds_df <- odds_df %>%
  mutate(cote_bin = case_when(
    favourite_odds > 2 ~ "Plus de 2.0",
    favourite_odds >= 1.8 ~ "1.8 à 2.0",
    favourite_odds >= 1.6 ~ "1.6 à 1.8",
    favourite_odds >= 1.4 ~ "1.4 à 1.6",
    favourite_odds >= 1.2 ~ "1.2 à 1.4",
    TRUE ~ "1.0 à 1.2")
```

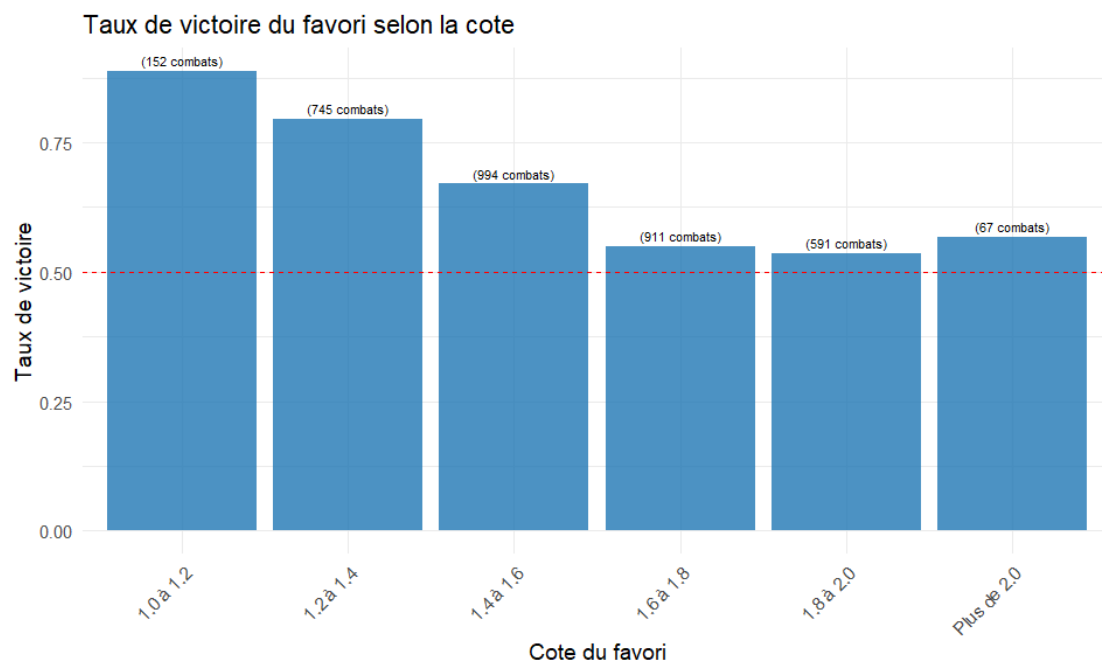
```

))

winrate_df <- odds_df %>%
  group_by(cote_bin) %>%
  summarise(
    total = n(),
    wins = sum(betting_outcome == "favourite"),
    winrate = wins / total
  ) %>%
  filter(!is.na(cote_bin)) %>%
  mutate(
    label = paste0("(", total, " combats)")
  )

ggplot(winrate_df, aes(x = factor(cote_bin, levels = c("1.0 à 1.2", "1.2 à 1.4", "1.4 à 1.6", "1.6 à 1.8", "1.8 à 2.0", "Plus de 2.0")), y = winrate)) +
  geom_col(fill = "#1f77b4", alpha = 0.8) +
  geom_hline(yintercept = 0.5, linetype = "dashed", color = "red") +
  geom_text(aes(label = label), vjust = -0.5, size = 3) +
  labs(
    title = "Taux de victoire du favori selon la cote",
    x = "Cote du favori",
    y = "Taux de victoire",
  ) +
  theme_minimal(base_size = 14) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```



Interprétation : On observe bien une pente plus ou moins linéaire. Plus la côte augmente et moins il y a de chances de gagner, se rapprochant même à 50% sur les cotes qui tendent

vers 2. Les combats où le favori a une cote de plus de 2 ont un taux de victoire supérieur aux combats entre 1.8 et 2. Cette différence vient sûrement du peu de combats où le favori a une cote aussi élevée. Le résultat est alors légèrement biaisé, mais la visualisation confirme l'hypothèse.

Conclusion question 4

Les favoris remportent environ **65% des combats**, ce qui confirme que les bookmakers ont globalement raison dans leurs prédictions. Cependant, les underdogs gagnent dans plus d'un tiers des cas, ce qui montre que le MMA reste un sport imprévisible. De plus, les cotes reflètent bien les probabilités de victoire, avec une corrélation claire entre des cotes plus basses et un taux de victoire plus élevé.