Rapport d'analyse sur les datasets des vins rouges et blancs

Auteurs : Morpheutt

Introduction

Dans le cadre de ce projet, nous avons choisi d'utiliser deux jeux de données portant sur la qualité des vins : l'un sur le vin rouge (red_wine.csv), l'autre sur le vin blanc (white_wine.csv). Ces données sont issues d'une étude menée par l'Université de Minho (Portugal) et sont accessibles publiquement sur plusieurs plateformes de data science. Elles ont été collectées afin de prédire la qualité des vins à partir de leurs propriétés physico-chimiques.

Rappel du plan d'analyse

L'objectif de notre analyse est d'explorer les relations entre les propriétés physico-chimiques des vins et leur qualité. Nous souhaitons répondre aux questions suivantes :

Questions principales :

1) Existe t'il une corrélation forte entre certaines propriétés chimiques et la qualité du vin ?
2) Y a t'il une différence significative entre les vins rouges et les vins blancs en termes de caractéristiques et de qualité ?
3) Peut-on identifer les variables qui influencent le plus la qualité du vin ?
4) Les vins ayant une teneur élevée en alcool obtiennent-ils une meilleure note ?

1) Existe t'il une corrélation forte entre certaines propriétés chimiques et la qualité du vin ?

==========

Generation de graphique pour répondre à la question (vin rouge)

==========

> Choix d'un scatter plot pour montrer une corrélation entre un élément chimique et la note d'un vin

Chargement des données

```
redwine <- read.csv("~/projet-if36-p25-morphe-utt/data/red_wine.csv", header = TRUE, sep = ';')
whitewine <- read.csv("~/projet-if36-p25-morphe-utt/data/white_wine.csv", header = TRUE, sep = ';')
```

Etape 1 : On ne garde que les variables qui sont reliées à de la chimie

```
variables_chimiques <- c("fixed_acidity", "volatile_acidity", "citric_acid",
              "residual_sugar", "chlorides", "free_sulfur_dioxide",
              "total_sulfur_dioxide", "density", "pH", "sulphates", "alcohol")
```

Etape 1.5 : Ajout des unités aux catégories gardées

```
unit_labels <- c(
  "fixed_acidity"         = "Acidité fixe (g/dm³)",
  "volatile_acidity"      = "Acidité volatile (g/dm³)",
  "citric_acid"           = "Acide citrique (g/dm³)",
  "residual_sugar"        = "Sucre résiduel (g/dm³)",
  "chlorides"             = "Chlorures (g/dm³)",
  "free_sulfur_dioxide"   = "SO2 libre (mg/dm³)",
  "total_sulfur_dioxide"  = "SO2 total (mg/dm³)",
  "density"               = "Densité (g/cm³)",
  "pH"                    = "pH",
  "sulphates"             = "Sulfates (g/dm³)",
  "alcohol"               = "Alcool (% vol.)"
)
```

Etape 2 : On factorise nos différentes variables et leurs valeurs dans un tableau à 2 dimensions

ça va permettre de générer pleins de graphiques en une seule fois plutot que d'en générer plusieurs à la main

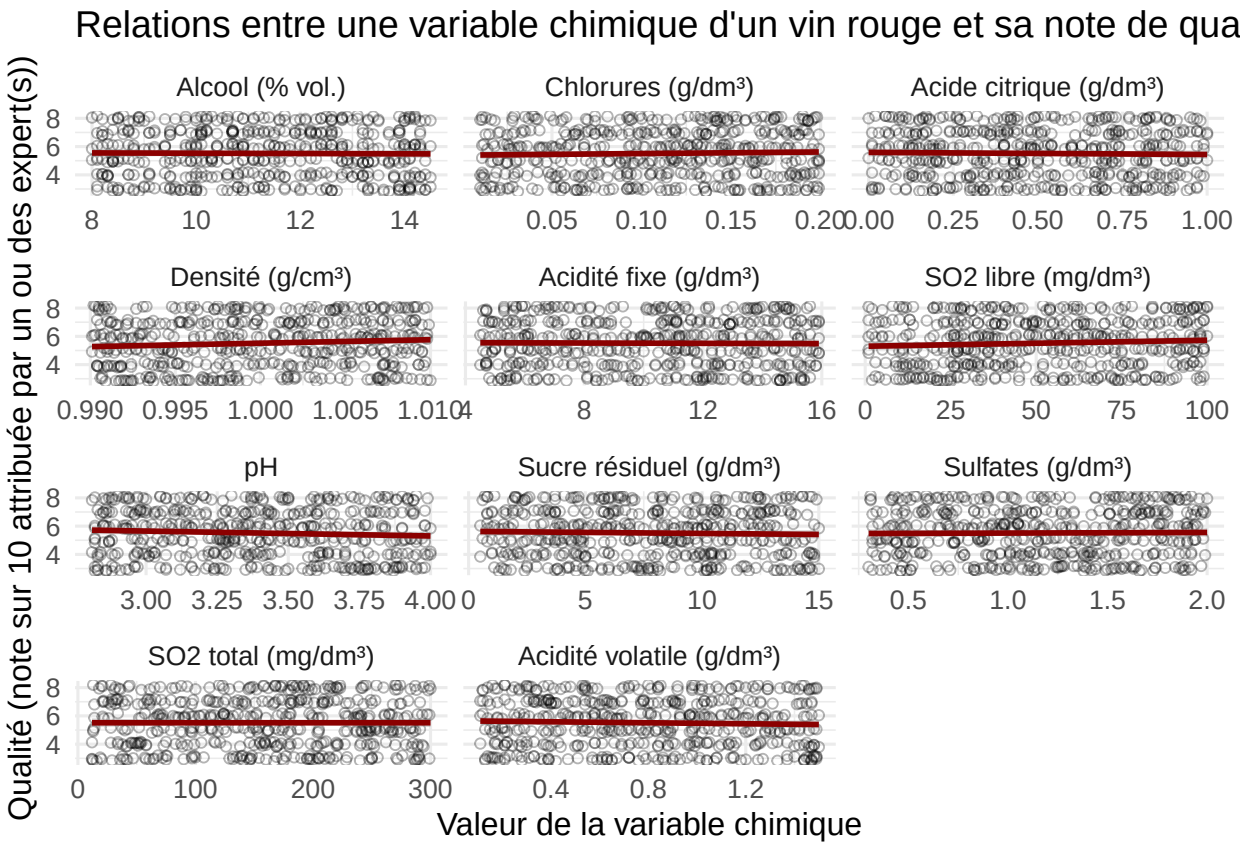
```
redwine2 <- redwine %>%
  select(quality, all_of(variables_chimiques)) %>%
  pivot_longer(
    cols = all_of(variables_chimiques),
    names_to  = "variable",
    values_to = "valeur"
  )
```

Etape 3 : Génération des scatters plots avec le tableau factorisé

```
ggplot(redwine2, aes(x = valeur, y = quality)) +
  # Génération des points
  geom_jitter(alpha = 0.3, shape = 1, width = 0, height = 0.2) +
  # Génération de la droite de regression avec son intervalle de confiance
  geom_smooth(method = "lm", se = TRUE, color = "darkred") +
  # Sert pour la création de nos graphiques, c'est ce qui permet la répartion sur 3 colonnes
  # mais aussi le fait que les abcisses soient libre
```

```
  facet_wrap(~ variable,
             scales = "free_x",
             ncol = 3,
             labeller = labeller(variable = unit_labels)) +
  # Utilise un thème épuré, pour faire ressortir la droite de regression
  #je trouve ça cool
  theme_minimal(base_size = 12) +
  # Gestion des labels
  labs(
    title = "Relations entre une variable chimique d'un vin rouge et sa note de qualité",
    x     = "Valeur de la variable chimique",
    y     = "Qualité (note sur 10 attribuée par un ou des expert(s))"
  )
```

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



==========

Generation de graphique pour répondre à la question (vin blanc)

==========

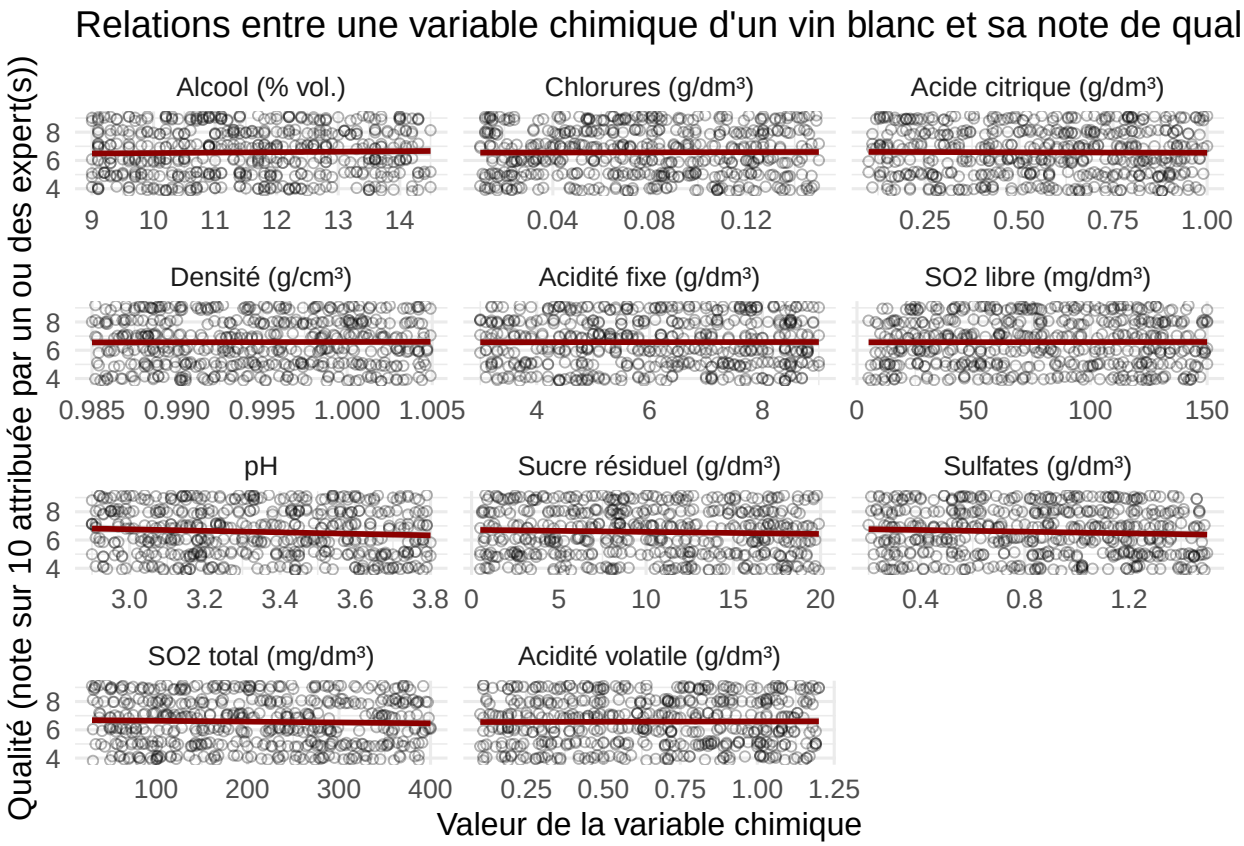
Etape 2 : On factorise nos différentes variables et leurs valeurs dans un tableau à 2 dimensions

```
whitewine2 <- whitewine %>%
  select(quality, all_of(variables_chimiques)) %>%
  pivot_longer(
    cols = all_of(variables_chimiques),
    names_to  = "variable",
    values_to = "valeur"
  )
```

Etape 3 : Génération des scatters plots avec le tableau factorisé

```
ggplot(whitewine2, aes(x = valeur, y = quality)) +
  # Génération des points
  geom_jitter(alpha = 0.3, shape = 1, width = 0, height = 0.2) +
  # Génération de la droite de regression avec son intervalle de confiance
  geom_smooth(method = "lm", se = TRUE, color = "darkred") +
  # Sert pour la création de nos graphiques, c'est ce qui permet la répartion sur 3 colonnes
  # mais aussi le fait que les abcisses soient libre
  facet_wrap(~ variable, scales = "free_x", ncol = 3, labeller = labeller(variable = unit_labels)) +
  # Utilise un thème épuré, pour faire ressortir la droite de regression je trouve ça cool
  theme_minimal(base_size = 12) +
  # Gestion des labels
  labs(
    title = "Relations entre une variable chimique d'un vin blanc et sa note de qualité",
    x     = "Valeur de la variable chimique",
    y     = "Qualité (note sur 10 attribuée par un ou des expert(s))"
  )
```

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



==========

Analyse et réponse à la question donnée

==========

VIN ROUGE : La droite de regression est stable quasiment de partout, on peut donc en conclure qu'il n'y a pas de lien fort entre les éléments chimiques et la note du vin. Une des hypothèses qu'on pourrait tirer serait que la note pourrait être attribuée en fonction du ressentit des experts. Par exemple lors d'une dégustation, une note peut-être attribuée en fonction du gout, de l'odorat, voire aussi du visuel si il y a des dépots en surface etc... Certainement que la provenance du vin joue aussi.

VIN BLANC : Concernant le vin blanc la même analyse peut-être appliquée, la droite de regression est stable quasiment partout. Cependant un lien plus fort peut-être observé entre le pH et la note, ainsi que le sucre residuel et les sulfates. En effet la droite de regression tend de manière négative, donc plus elle descend dans les notes et plus ces valeurs chimiques sont élevées dans le vin. Donc si on reprend notre analyse concernant le vin rouge, on pourrait en déduire que ces variables jouent plus sur les variables de ressentit des experts (gout, visuel etc...) que le reste. Plus elles sont élevées dans un vin, plus elles impactes sur la qualité du vin, et plus ce dernier a une note médiocre

EN CONCLUSION : Il existe bel est bien une corrélation entre certains éléments chimiques et la note du vin, puisque l'évolution des éléments chimiques impactent la note et inversement. En revanche il ne peut pas être catégorisé comme lien fort, puisque aucune variable chimique à elle seule impact de manière majoritaire la note d'un vin en particulier. AVec cette analyse, on comprend surtout que les variables chimiques d'un vin forment un tout, et que c'est leurs associations qui font varier un vin et sa qualité !

2) Y a t'il une différence significative entre les vins rouges et les vins blancs en termes de caractéristiques et de qualité ?

==========

Travail réalisé sur R

==========

Harmoniser les noms de colonnes

```
names(red_wine) <- tolower(names(red_wine))
names(white_wine) <- tolower(names(white_wine))
```

Vérification

```
setdiff(names(red_wine), names(white_wine))
```

```
## [1] "tannins"
```

```
setdiff(names(white_wine), names(red_wine))
```

```
## [1] "mineralite"
```

Ajouter une colonne type_vin pour identifier

```
red_wine$type_vin <- "rouge"
white_wine$type_vin <- "blanc"
```

Ajouter les colonnes manquantes

```
red_wine$mineralite <- NA
white_wine$tannins <- NA
```

Réordonner

```
red_wine <- red_wine[, names(white_wine)]
```

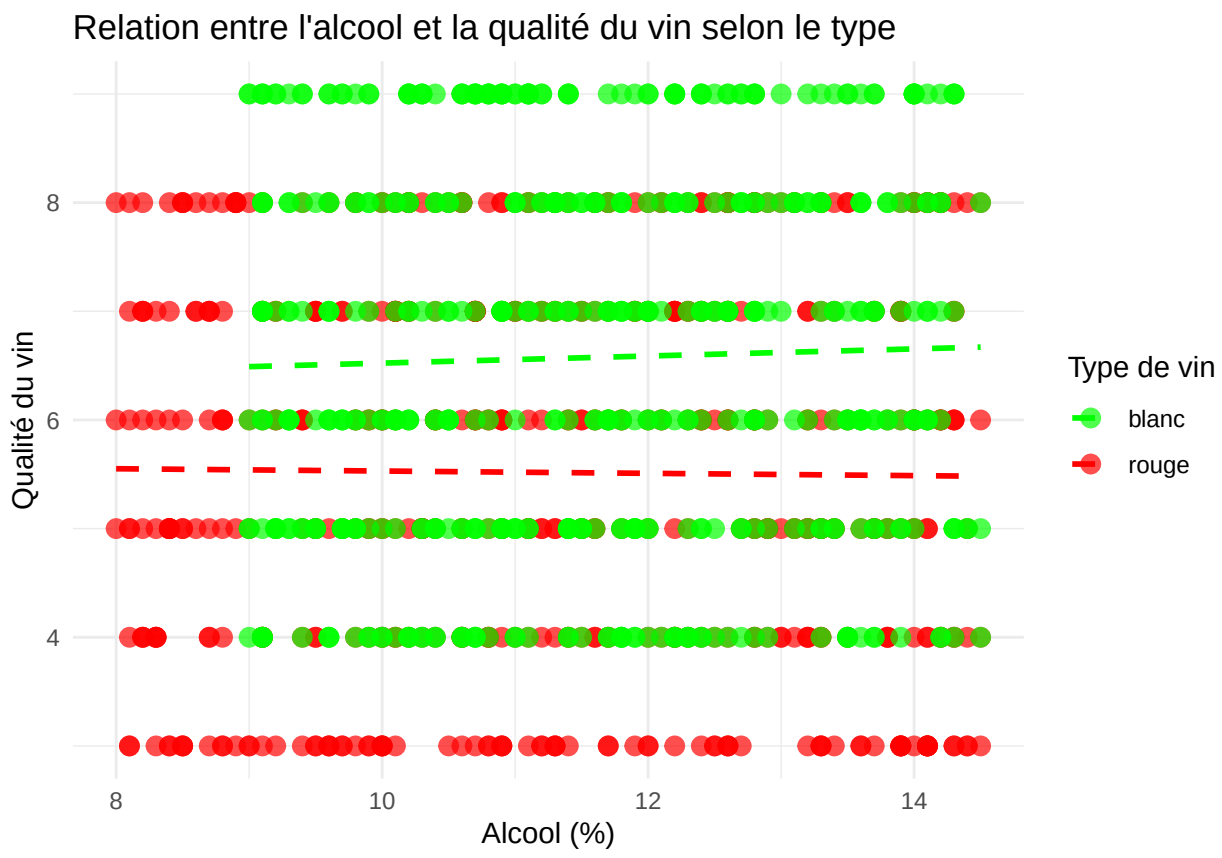
Fusionner

```
vin_total <- rbind(red_wine, white_wine)
```

——— Plot initial ———

```
library(ggplot2)

ggplot(vin_total, aes(x = alcohol, y = quality, color = type_vin)) +
  geom_point(size = 3, alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE, linetype = "dashed") +
  labs(title = "Relation entre l'alcool et la qualité du vin selon le type",
       x = "Alcool (%)",
       y = "Qualité du vin",
       color = "Type de vin") +
  theme_minimal() +
  scale_color_manual(values = c("rouge" = "red", "blanc" = "green"))
```

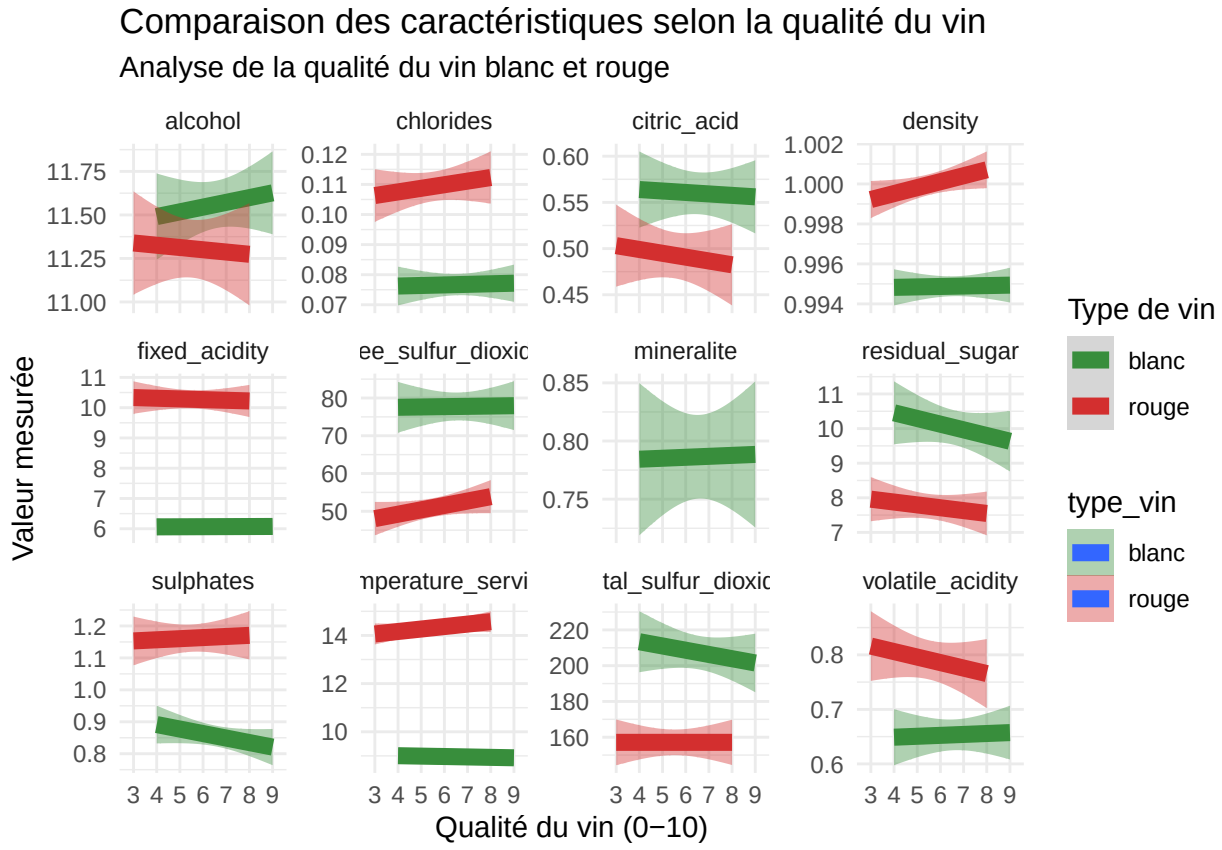
```
## `geom_smooth()` using formula = 'y ~ x'
```



——— Facet Wrap ———

```
library(tidyr)

vin_long <- vin_total %>%
  pivot_longer(cols = c(fixed_acidity, volatile_acidity, citric_acid,
                        residual_sugar, chlorides, free_sulfur_dioxide,
                        total_sulfur_dioxide, density, sulphates,
                        alcohol, temperature_service, mineralite),
               names_to = "variable",
               values_to = "valeur")

ggplot(vin_long, aes(x = as.factor(quality), y = valeur, color = type_vin)) +
  geom_smooth(method = "lm", se = TRUE, linetype = "solid", size = 3,
              aes(group = type_vin, color = type_vin, fill = type_vin)) +
  facet_wrap(~ variable, scales = "free_y") +
  labs(
    title = "Comparaison des caractéristiques selon la qualité du vin",
    subtitle = "Analyse de la qualité du vin blanc et rouge",
    x = "Qualité du vin (0-10)",
    y = "Valeur mesurée",
    color = "Type de vin"
  ) +
  scale_color_manual(values = c("rouge" = "#D32F2F", "blanc" = "#388E3C")) +
  scale_fill_manual(values = c("rouge" = "#D32F2F", "blanc" = "#388E3C")) +
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.

## ‘geom_smooth()‘ using formula = ’y ~ x’

## Warning: Removed 500 rows containing non-finite outside the scale range
## (‘stat_smooth()‘).
```



==========

Analyse et réponse à la question donnée

==========

Oui, il existe des différences significatives entre les vins rouges et les vins blancs en termes de caractéristiques et de qualité.

Caractéristiques chimiques :

Alcool : Les vins blancs ont en moyenne un taux d'alcool plus élevé que les rouges. Pour les blancs, une augmentation du taux d'alcool est associée à une meilleure qualité, alors que c'est l'inverse pour les rouges.

Chlorures : Les vins rouges contiennent davantage de chlorures que les blancs. Pour les rouges, une teneur plus élevée en chlorures est liée à une meilleure qualité, ce qui n'est pas le cas pour les blancs.

Acide citrique : Plus présent dans les blancs que dans les rouges. Dans les deux types de vins, une qualité supérieure est associée à une baisse de la teneur en acide citrique.

Densité : Les rouges sont plus denses que les blancs. La densité influence la qualité des rouges, mais pas celle des blancs.

Acidité fixe : Environ deux fois plus élevée dans les rouges que dans les blancs.

Dioxyde de soufre total : Plus présent dans les blancs. Dans les rouges, une augmentation est corrélée à une meilleure qualité.

Minéralité : Présente uniquement dans les blancs, sans lien clair avec la qualité.

Sucres résiduels : Plus élevés de 25 % dans les blancs que dans les rouges. Dans les deux cas, une baisse des sucres est associée à une meilleure qualité.

Sulphates : Plus présents dans les rouges. Leur légère augmentation améliore la qualité des rouges, tandis que dans les blancs, leur diminution est corrélée à une meilleure qualité.

Dioxyde de soufre libre : Plus abondant dans les blancs. Sa diminution est associée à une amélioration de la qualité dans les blancs, mais elle n'influence pas la qualité des rouges.

Acidité volatile : Plus présente dans les rouges. Sa baisse est liée à une amélioration de la qualité pour les rouges, alors que pour les blancs, c'est sa hausse qui semble bénéfique.

Température de service : Les vins rouges sont servis en moyenne 4°C plus chauds que les blancs.

CONCLUSION

Les vins rouges et blancs présentent des différences marquées à la fois dans leur composition chimique et dans l'impact de ces éléments sur leur qualité perçue. Ces distinctions sont importantes pour leur production, leur évaluation et leur dégustation.

3) Peut-on identifer les variables qui influencent le plus la qualité du vin ?

==========

COMMENTAIRE D'INTERPRÉTATION DU TRAVAIL RÉALISÉ

==========

1. Chargement des données

Nous avons chargé deux jeux de données distincts contenant les caractéristiques physico-chimiques du vin rouge et du vin blanc. Chaque échantillon est accompagné d'un score de qualité.

2. Identification du type de vin

Afin de différencier les échantillons, une nouvelle colonne ('type_vin') a été ajoutée pour indiquer s'il s'agit de vin rouge ou blanc.

3. Fusion des données

Les deux datasets ont été combinés en un seul ensemble de données global pour faciliter l'analyse.

4. Préparation des données

Nous avons remplacé toutes les valeurs manquantes (NA) par 0 pour éviter toute perturbation lors du calcul des corrélations.

5. Sélection des variables pertinentes

Seules les variables numériques (caractéristiques physico-chimiques mesurables) ont été conservées pour le calcul de corrélation avec la qualité.

6. Calcul des corrélations

Pour chaque type de vin (rouge et blanc), nous avons calculé la corrélation de chaque variable physico-chimique avec la qualité. La corrélation mesure ici la force et le sens de la relation entre chaque caractéristique et la qualité :

- Une corrélation positive indique qu'une augmentation de la variable est associée à une augmentation de la qualité.
- Une corrélation négative indique qu'une augmentation de la variable est associée à une diminution de la qualité.

7. Visualisation

Trois types de graphiques ont été générés : - Un graphique pour le vin blanc seul - Un graphique pour le vin rouge seul - Un graphique comparatif entre vin blanc et vin rouge

Chaque graphique présente : - Les variables physico-chimiques en abscisse, - La corrélation avec la qualité en ordonnée, limitée entre -0.10 et 0.10, - Les couleurs bleu foncé pour le vin blanc et vert pour le vin rouge.

8. Réglages graphiques spécifiques

Nous avons choisi une échelle restreinte (-0.10 à +0.10) pour mieux visualiser les faibles corrélations, caractéristiques typiques des études sur le vin. Les intervalles de 0.05 permettent une lecture précise sans surcharger l'axe.

Conclusion

Cette méthodologie nous permet d'identifier rapidement quelles variables influencent le plus (positivement ou négativement) la qualité du vin, de façon distincte entre le vin rouge et le vin blanc.

==========

Ajouter une colonne pour indiquer le type de vin

```
red_wine <- red_wine %>% mutate(type_vin = "rouge")
white_wine <- white_wine %>% mutate(type_vin = "blanc")
```

Fusionner les deux datasets

```
wine_data <- bind_rows(red_wine, white_wine)
```

Remplacer les NA par 0

```
wine_data <- wine_data %>% mutate(across(everything(), ~replace_na(., 0)))
```

Sélection des variables numériques

```
num_vars <- wine_data %>%
  select(-type_vin) %>%
  select_if(is.numeric) %>%
  names()
```

Calcul de la corrélation entre chaque variable et la qualité, par type de vin

```
correlations <- wine_data %>%
  select(all_of(num_vars), type_vin) %>%
  pivot_longer(cols = -c(quality, type_vin), names_to = "variable", values_to = "valeur") %>%
  group_by(type_vin, variable) %>%
  summarize(correlation = cor(valeur, quality), .groups = "drop") %>%
  filter(correlation != 0)
```

```
## Warning: There were 2 warnings in ‘summarize()‘.
## The first warning was:
## i In argument: ‘correlation = cor(valeur, quality)‘.
## i In group 13: ‘type_vin = "blanc"‘ ‘variable = "tannins"‘.
## Caused by warning in ‘cor()‘:
## ! l’écart type est nul
## i Run ‘dplyr::last_dplyr_warnings()‘ to see the 1 remaining warning.
```

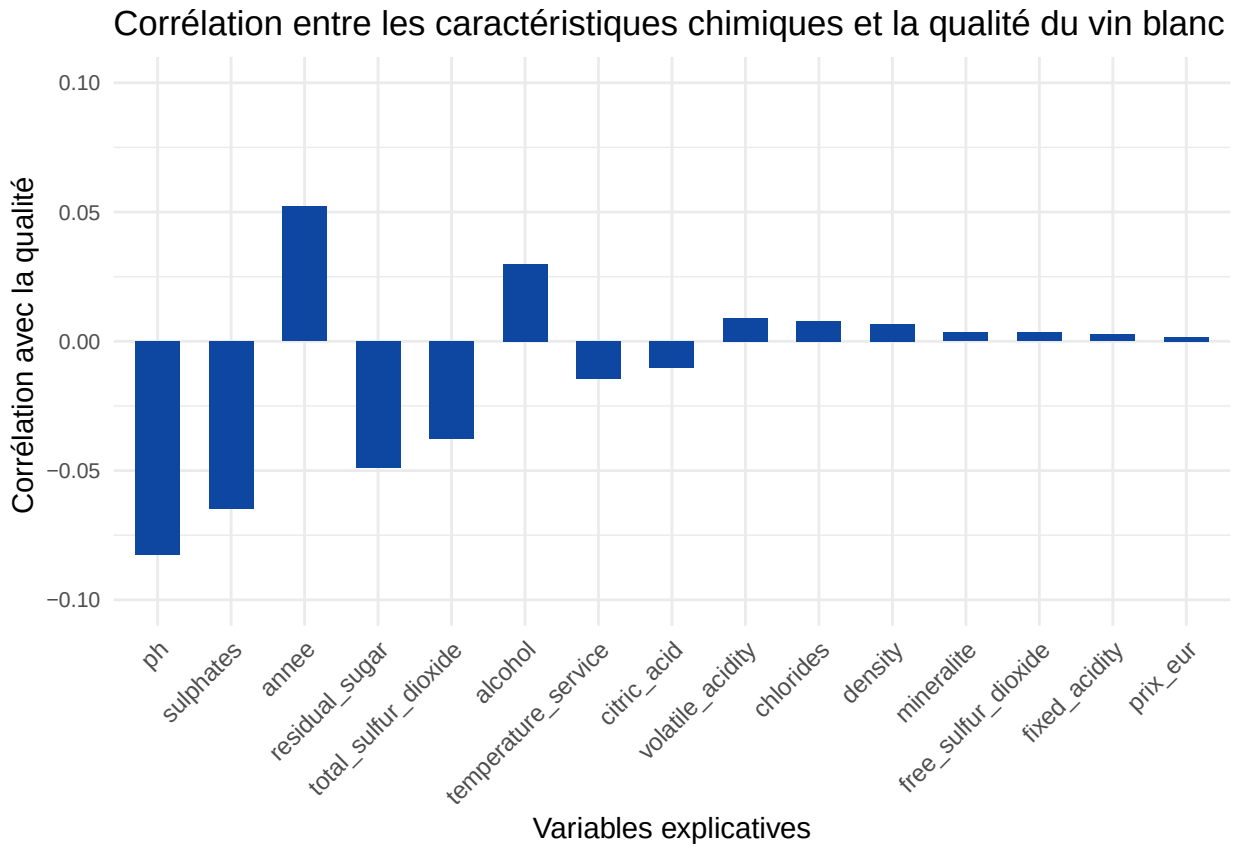
==========

1. Graphe pour le vin blanc

==========

```
cor_blanc <- correlations %>%
  filter(type_vin == "blanc") %>%
  mutate(variable = forcats::fct_reorder(variable, abs(correlation), .desc = TRUE))

ggplot(cor_blanc, aes(x = variable, y = correlation, fill = type_vin)) +
  geom_bar(stat = "identity", width = 0.6) +
  scale_fill_manual(values = c("blanc" = "#0d47a1")) +  # Bleu foncé
  scale_y_continuous(
    breaks = seq(-0.10, 0.10, by = 0.05),
    limits = c(-0.10, 0.10)
  ) +
  labs(
```

```
    title = "Corrélation entre les caractéristiques chimiques et la qualité du vin blanc",
    x = "Variables explicatives",
    y = "Corrélation avec la qualité ",
    fill = "Type de vin"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    axis.text.y = element_text(size = 8),
    legend.position = "none"
  )
```



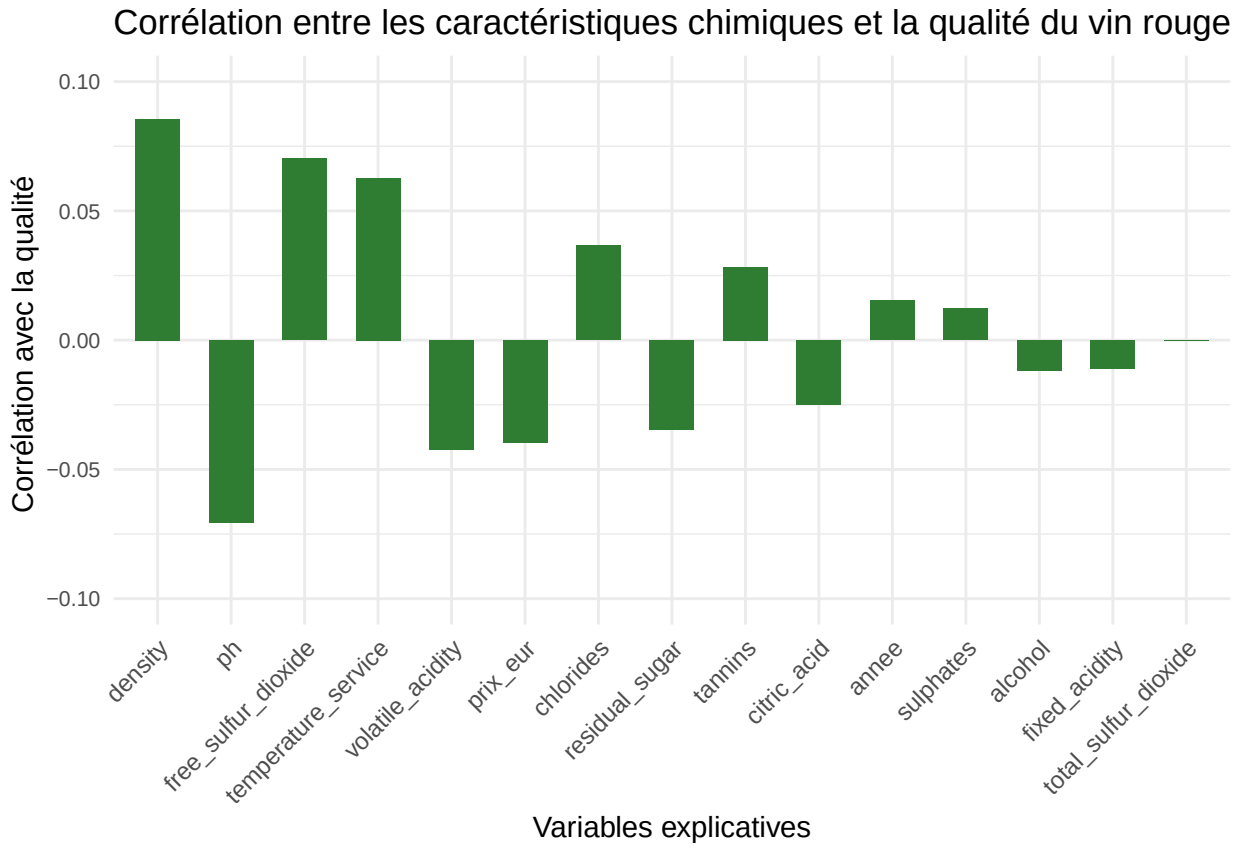
==========

2. Graphe pour le vin rouge

==========

```
cor_rouge <- correlations %>%
  filter(type_vin == "rouge") %>%
  mutate(variable = forcats::fct_reorder(variable, abs(correlation), .desc = TRUE))

ggplot(cor_rouge, aes(x = variable, y = correlation, fill = type_vin)) +
```

```r
  geom_bar(stat = "identity", width = 0.6) +
  scale_fill_manual(values = c("rouge" = "#2e7d32")) +  # Vert foncé
  scale_y_continuous(
    breaks = seq(-0.10, 0.10, by = 0.05),
    limits = c(-0.10, 0.10)
  ) +
  labs(
    title = "Corrélation entre les caractéristiques chimiques et la qualité du vin rouge",
    x = "Variables explicatives",
    y = "Corrélation avec la qualité ",
    fill = "Type de vin"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    axis.text.y = element_text(size = 8),
    legend.position = "none"
  )
```



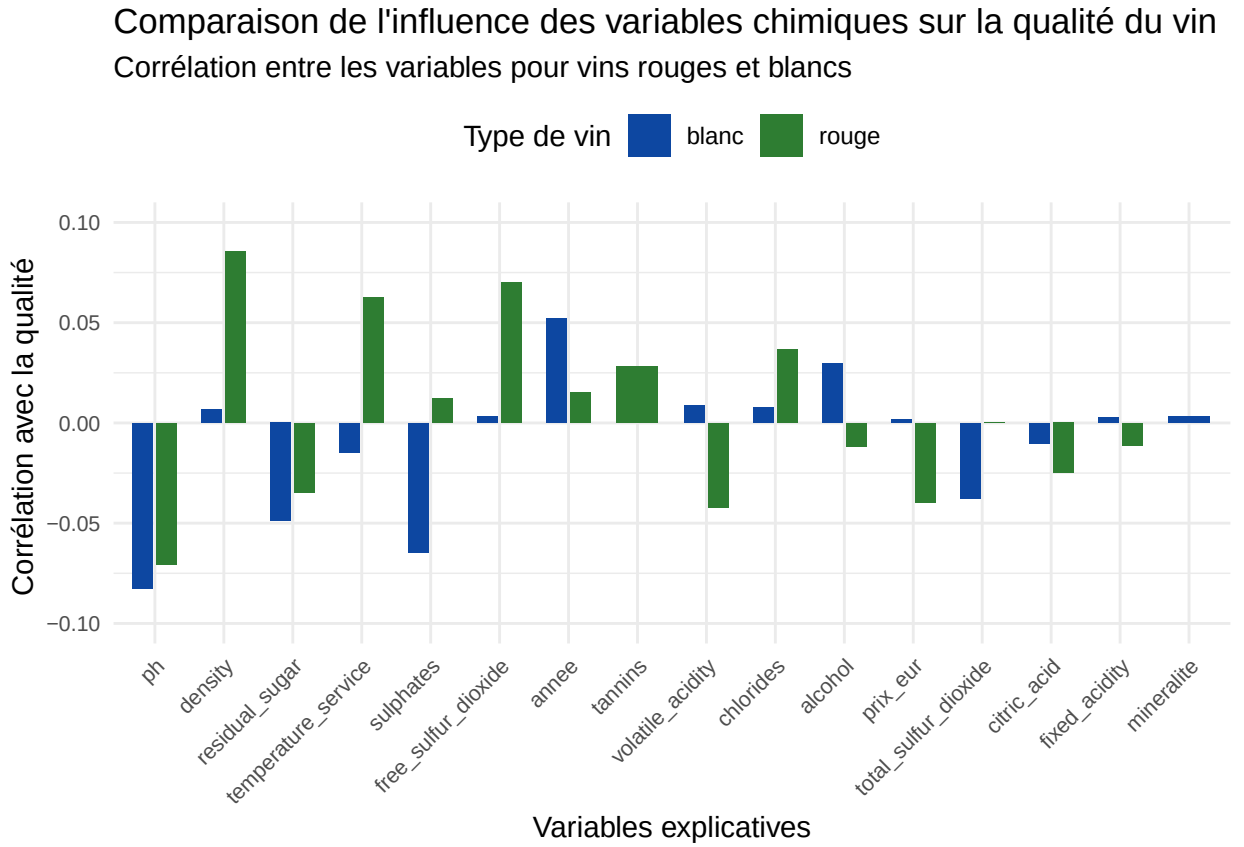
==========

3. Graphe comparatif vin rouge vs vin blanc

==========

Re-trier les variables pour la comparaison

```r
correlations <- correlations %>%
  mutate(variable = forcats::fct_reorder(variable, abs(correlation), .desc = TRUE))

ggplot(correlations, aes(x = variable, y = correlation, fill = type_vin)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.7), width = 0.6) +
  scale_fill_manual(values = c("blanc" = "#0d47a1", "rouge" = "#2e7d32")) +
  scale_y_continuous(
    breaks = seq(-0.10, 0.10, by = 0.05),
    limits = c(-0.10, 0.10)
  ) +
  labs(
    title = "Comparaison de l'influence des variables chimiques sur la qualité du vin",
    subtitle = "Corrélation entre les variables pour vins rouges et blancs",
    x = "Variables explicatives",
    y = "Corrélation avec la qualité ",
    fill = "Type de vin"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 8),
    axis.text.y = element_text(size = 8),
    legend.position = "top"
  )
```



Le pH et le sucre résiduel affectent négativement la qualité des vins rouges comme des vins blancs. À l'inverse, l'acide amree contribue à améliorer la qualité pour les deux types de vins. La volatilité nuit particulièrement à la qualité du vin rouge, tandis que la densité et le dioxyde de soufre libre exercent un effet bénéfique, surtout sur les vins rouges. Enfin, les sulfates présentent un comportement opposé : ils impactent positivement la qualité du vin rouge, mais négativement celle du vin blanc.

4) Les vins ayant une teneur élevée en alcool obtiennent-ils une meilleure note ?

Hypothèses de départ Nous formulons l'hypothèse suivante : > Les vins ayant une teneur en alcool plus élevée pourraient obtenir de meilleures notes de qualité, car l'alcool peut influencer positivement la perception sensorielle du vin (structure, chaleur en bouche, équilibre).

Cependant, un excès d'alcool pourrait également déséquilibrer le vin et nuire à son évaluation.

1. **Visualisation de la relation entre `alcohol` et `quality`** :
   - Création d'un nuage de points (`scatterplot`) avec `alcohol` en abscisse et `quality` en ordonnée.
   - Ajout d'une droite de tendance pour observer une éventuelle corrélation visuelle.
2. **Comparaison entre types de vins** :

- Réaliser les analyses séparément pour les vins rouges et les vins blancs.
- Vérifier si la relation entre alcool et qualité diffère selon le type de vin.

Résultats attendus

- Nous nous attendons à observer une **corrélation positive modérée** entre la teneur en alcool et la qualité perçue.
- Il est probable que les vins ayant un **taux d'alcool plus élevé** aient, en moyenne, une **meilleure note** que les vins faiblement alcoolisés.
- Toutefois, la relation pourrait ne pas être linéaire parfaite : un vin excessivement alcoolisé pourrait être moins apprécié.