EDA

Goal

- Examine distributions of variables, identify skewness, and determine whether transformation is needed.
- Examine correlation structure to identify potential multicollinearity.
- Compare distributions of chemical attributes between red and white wines to identify potential interactions.
- Use scatterplots to assess the direction and strength of linear relationships.

```
# prep
library(tidyverse)
library(ggplot2)
library(tidyr)
library(corrplot)

wine <- read_csv("data/wine.csv")
wine <- wine |>
  mutate(
    type = factor(
      type,
      levels = c(0, 1),
      labels = c("red", "white")
    )
  )
```

Histogram

For each chemical factors:

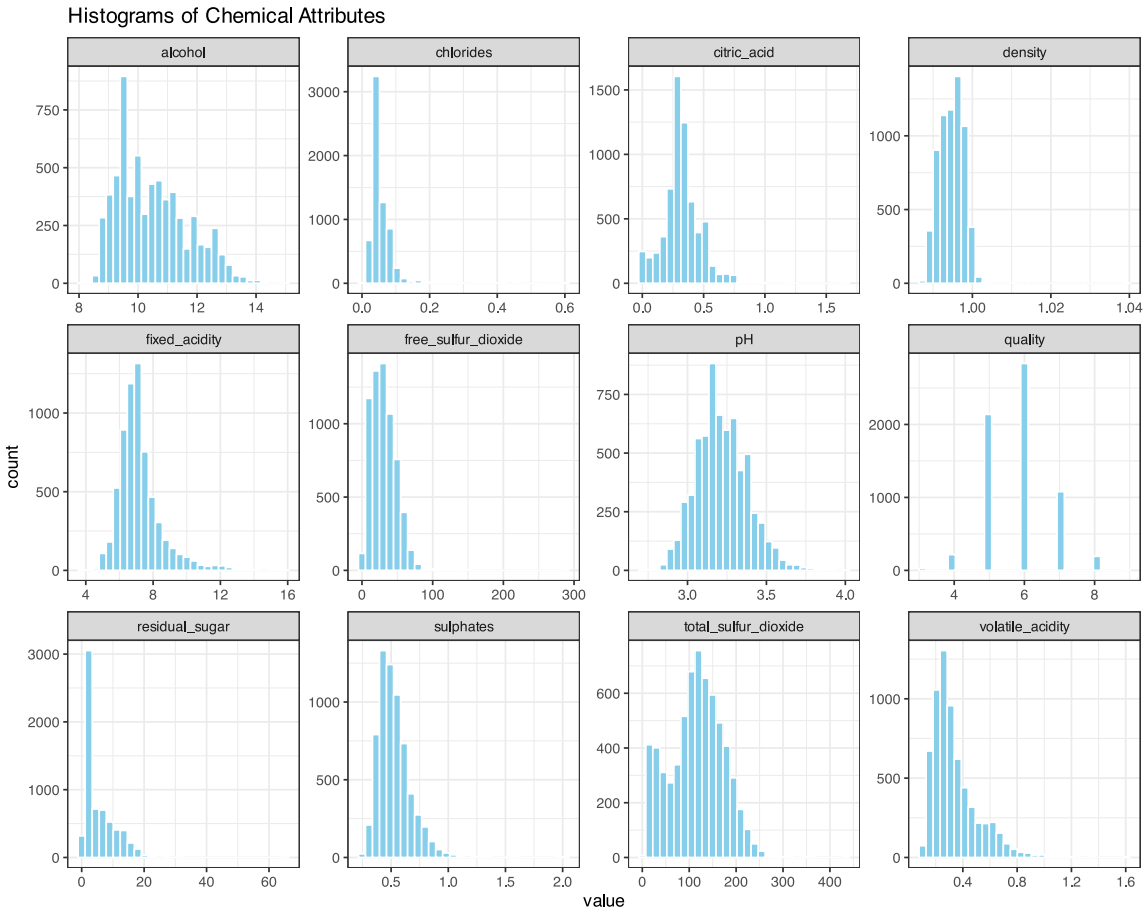
skewness, outliers, long tail, multimodality -> transformations (log, sqrt)

对 predictor 做 transformation 的目的只有三类：

- 改善 predictor–response（quality）的线性关系
- 减少 right-skew ⟶ 减轻 heteroscedasticity
- 改善 model interpretability（不影响科学解释）

```
# only numerical variables, excluding types
wine_numeric <- wine[, sapply(wine, is.numeric)]

wine_long <- pivot_longer(wine_numeric,
                          cols = everything(),
```

```
                        names_to = "variable",
                        values_to = "value")

ggplot(wine_long, aes(x = value)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "white") +
  facet_wrap(~ variable, scales = "free") +
  theme_bw() +
  labs(title = "Histograms of Chemical Attributes")
```



Conclusion

Variable	Transformation Recommendation
residual_sugar	log(x + 0.1)
total_sulfur_dioxide	log(x + 1)
free_sulfur_dioxide	log(x + 1)
chlorides	log(x)

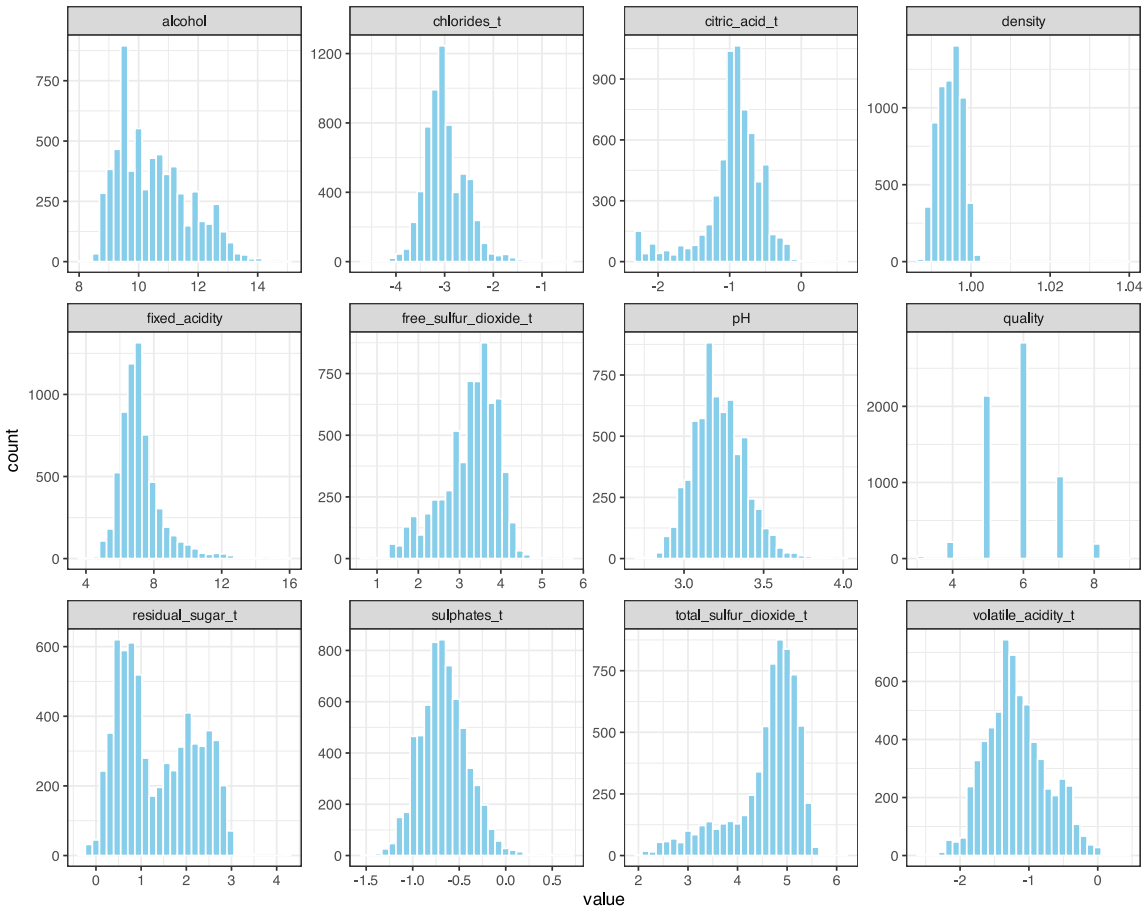
Variable	Transformation Recommendation
volatile_acidity	log(x)
sulphates	log(x)
citric_acid	log(x + 0.01)
alcohol	no transform
fixed_acidity	no transform
density	no transform (consider remove later)
pH	no transform
quality	response, no transform

Transformation

```
wine_tf <- data.frame(
  # response
  quality = wine$quality,
  wine_type = wine$type,

  # transformed predictors
  residual_sugar_t        = log(wine$residual_sugar + 0.1),
  total_sulfur_dioxide_t  = log(wine$total_sulfur_dioxide + 1),
  free_sulfur_dioxide_t   = log(wine$free_sulfur_dioxide + 1),
  chlorides_t             = log(wine$chlorides),
  volatile_acidity_t      = log(wine$volatile_acidity),
  sulphates_t             = log(wine$sulphates),
  citric_acid_t           = log(wine$citric_acid + 0.1),

  # raw predictors kept as-is
  alcohol          = wine$alcohol,
  fixed_acidity    = wine$fixed_acidity,
  pH               = wine$pH,
  density          = wine$density
)
```

```
wine_tf_numeric <- wine_tf[, sapply(wine_tf, is.numeric)]

wine_tf_long <- pivot_longer(wine_tf_numeric,
                             cols = everything(),
                             names_to = "variable",
                             values_to = "value")

ggplot(wine_tf_long, aes(x = value)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "white") +
```

```
  facet_wrap(~ variable, scales = "free") +
  theme_bw()
```

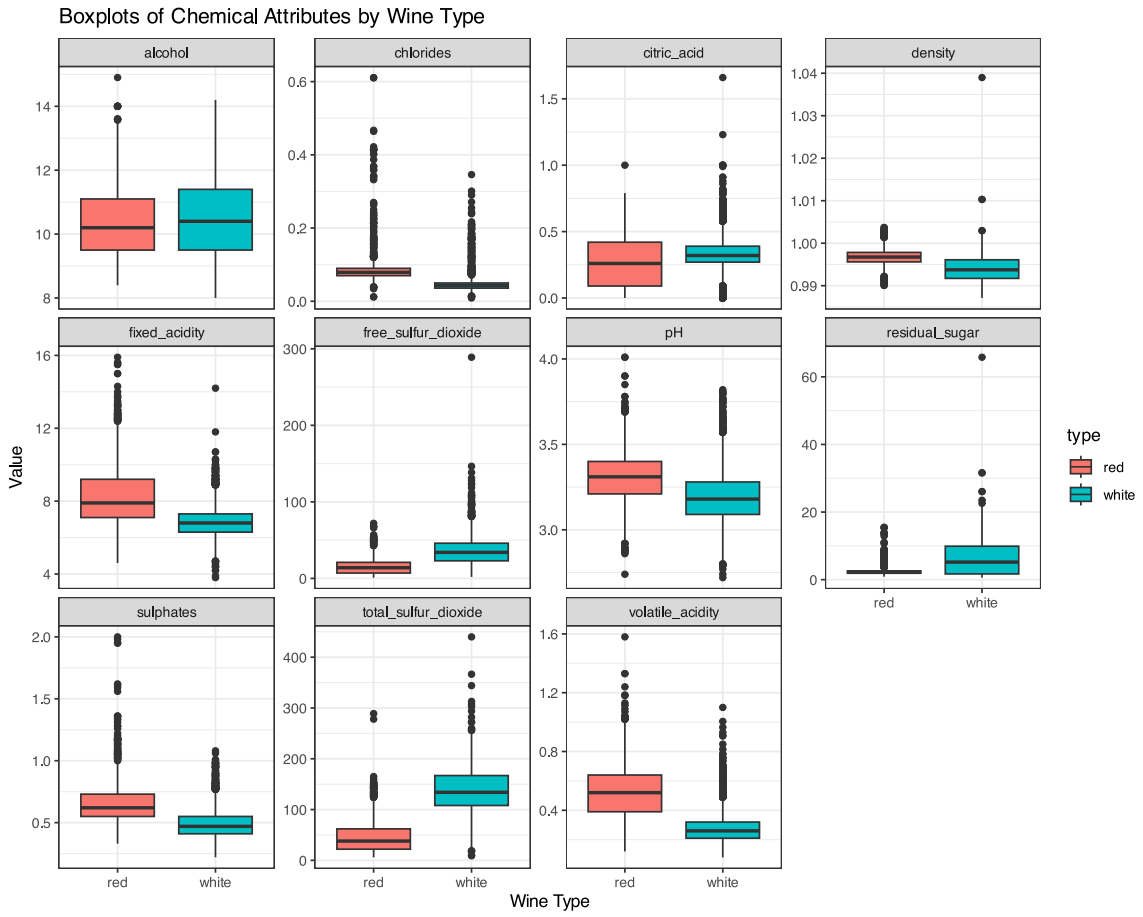


Boxplot by Type

Diffrence in red and white wine according to each predictor -> look for potential interaction with type

```
# Boxplot: chemical variable by type
wine_long2 <- pivot_longer(wine,
                           cols = -c(type, quality),
                           names_to = "variable",
                           values_to = "value")

ggplot(wine_long2, aes(x = type, y = value, fill = type)) +
  geom_boxplot() +
  facet_wrap(~ variable, scales = "free_y") +
  theme_bw() +
```

```
  labs(title = "Boxplots of Chemical Attributes by Wine Type",
       x = "Wine Type", y = "Value")
```



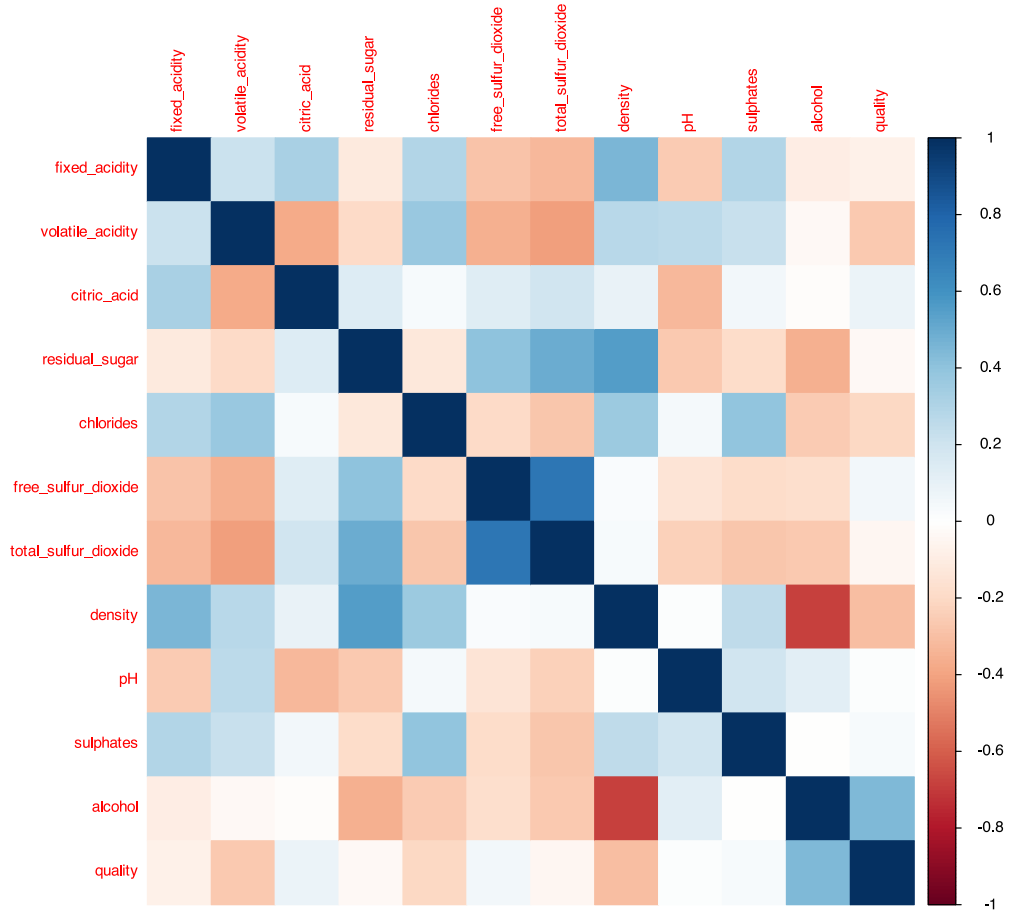
Conclusion

- residual_sugar * type
- total_sulfur_dioxide * type
- free_sulfur_dioxide * type
- volatile_acidity * type
- fixed_acidity * type
- chlorides * type

Correlation Heatmap

linear collinearity in between variables -> potential multicollinearity; strongest marginal relationships to quality

```
corr_matrix <- cor(wine_numeric)   # numeric variables only
corrplot(corr_matrix, method = "color", tl.cex = 0.8)
```

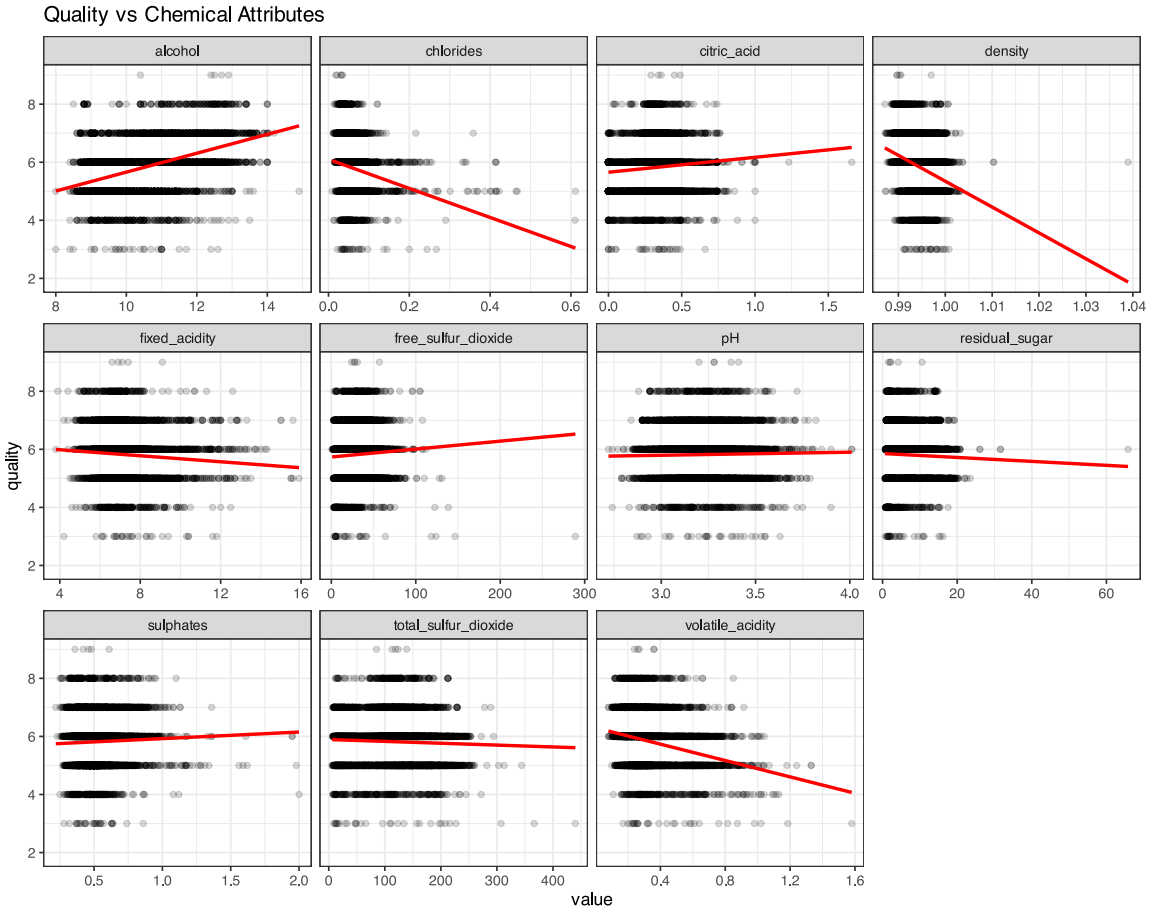


Conclusion

保留两者但不要在同一模型中出现 strong: - residual_sugar vs density - free_sulfur_dioxide vs total_sulfur_dioxide moderate: - fixed_acidity vs citric_acid - volatile_acidity vs citric_acid - sulphates vs total_sulfur_dioxide weak: - alcohol vs anything - pH vs anything

Scatterplots: Quality vs Each Predictor

```
ggplot(wine_long2, aes(x = value, y = quality)) +
  geom_point(alpha = 0.15) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  facet_wrap(~ variable, scales = "free_x") +
  theme_bw() +
  labs(title = "Quality vs Chemical Attributes")
```

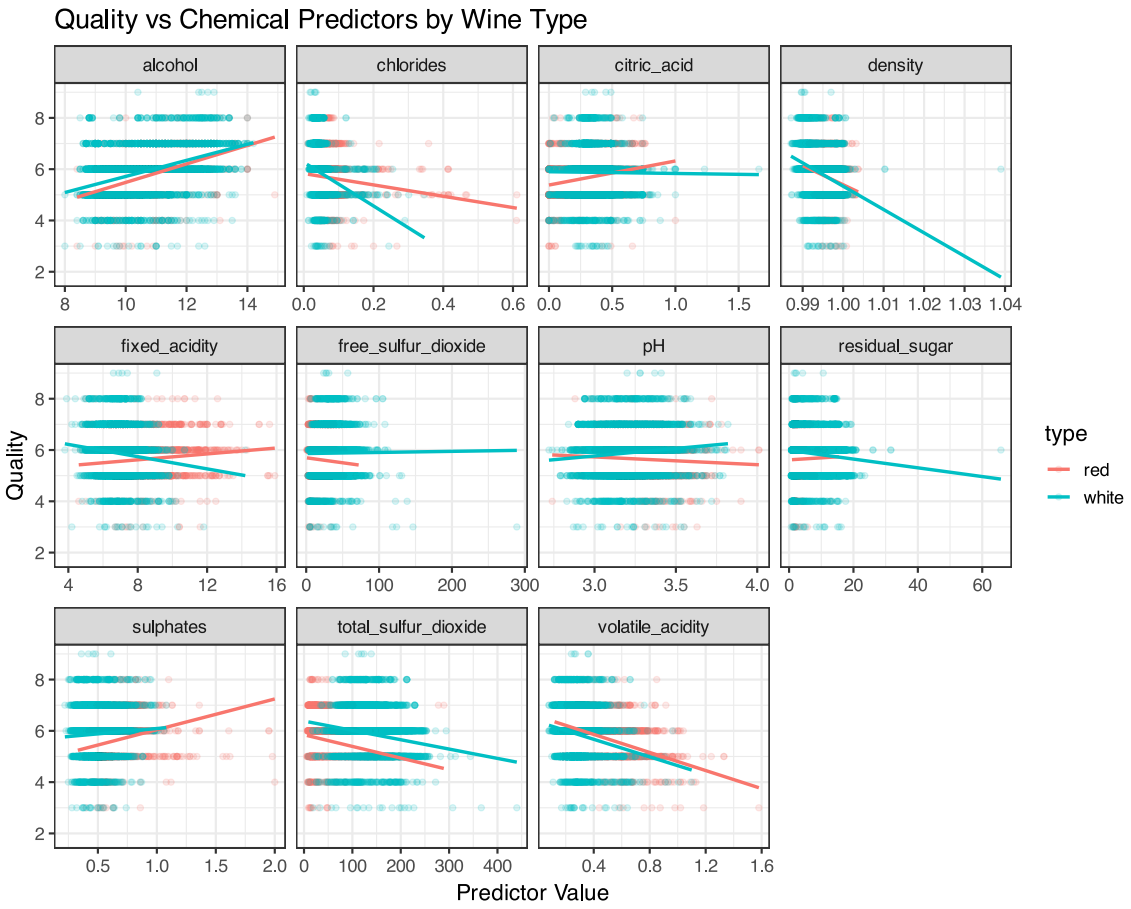


Quality vs Predictor by Wine Type

investigate on slope difference of red and white wine -> as potential proof for type × variable interaction

```
# modified: 只选择原始数据中的数值型化学变量，排除 quality 和 type
predictors <- wine %>%
  select(-quality, -type) %>%
  select(where(is.numeric)) %>%
  colnames()

# modified: 仅对上述 predictors 做 pivot_longer，不使用 transformed data
wine_long <- wine %>%
  pivot_longer(
    cols = all_of(predictors),
    names_to = "variable",
    values_to = "value"
  )
```

```r
ggplot(wine_long, aes(x = value, y = quality, color = type)) +
  geom_point(alpha = 0.15) +
  geom_smooth(method = "lm", se = FALSE) +
  facet_wrap(~ variable, scales = "free_x") +
  theme_bw(base_size = 14) +
  labs(
    title = "Quality vs Chemical Predictors by Wine Type",
    x = "Predictor Value",
    y = "Quality"
  )
```



Conclusion

Strong: - volatile_acidity * type - chlorides * type Moderate: - total_sulfur_dioxide * type - residual_sugar * type Weak: - citric_acid * type - sulphates * type

Quality Distribution by Type

```
ggplot(wine, aes(x = quality, fill = type)) +
  geom_histogram(position = "identity", alpha = 0.5, bins = 20) +
  theme_bw() +
  labs(title = "Distribution of Wine Quality by Type")
```