Project Proposal

Due November 17 at 11:59pm

Alex Ackerman, Cindy Gao, Kayla Haeussler, Javidan Karimli

Load Packages

```
library(tidyverse)
```

Dataset 1 (top choice)

Data source: https://raw.githubusercontent.com/fivethirtyeight/data/refs/heads/master/bechdel/movies.csv

Brief description: The Bechdel Test data for movies originates from BechdelTest.com, a website that tracks movies' compliance with the Bechdel Test—a test that evaluates if a movie features at least two women who talk to each other about something other than a man. We captured the csv for this dataset from the 'fivethirtyeight' repository on GitHub. The data is collected and maintained by community contributions and includes information on movies' release years, test results, financial information, and other relevant attributes. This dataset has been periodically updated as users contribute evaluations of new movies. Each row represents a unique movie and includes the movie's title, IMDb identifier, year of release, Bechdel Test results (pass or fail), as well as information on budget, domestic gross, international gross, and other movie attributes.

Research question 1: Is there a significant relationship between a movie's budget and its likelihood of passing the Bechdel Test?

- Outcome variable (include the name/description and type of variable): binary – Bechdel Test pass/fail (Binary)
- Interaction Term**:** budget_2013$ * decade code – to examine if the relationship between budget and Bechdel Test outcome varies by decade (Ordinal). budget_2013$ represents the budget adjusted to 2013 dollar value, adjusting for inflation.

Research question 2: Is there a significant difference in the international box office revenue of movies that pass the Bechdel Test compared to those that fail, after adjusting for genre?

- Outcome variable (include the name/description and type of variable): intgross_2013$ – International Gross Revenue in 2013 dollars (Continuous)
- Interaction Term (optional)**:** For this model, no interaction term is initially included, but an extension could explore binary * genre to see if the effect of passing the Bechdel Test on revenue depends on the genre.

Load the data and provide a `glimpse()`:

```
movies<- read_csv("https://raw.githubusercontent.com/ackerman-alex/IDS_702_Final_Project/refs
```

```
Rows: 1794 Columns: 16
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (7): imdb, title, test, clean_test, binary, code, genre
dbl (9): year, budget, domgross, intgross, budget_2013$, domgross_2013$, int...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(movies)
```

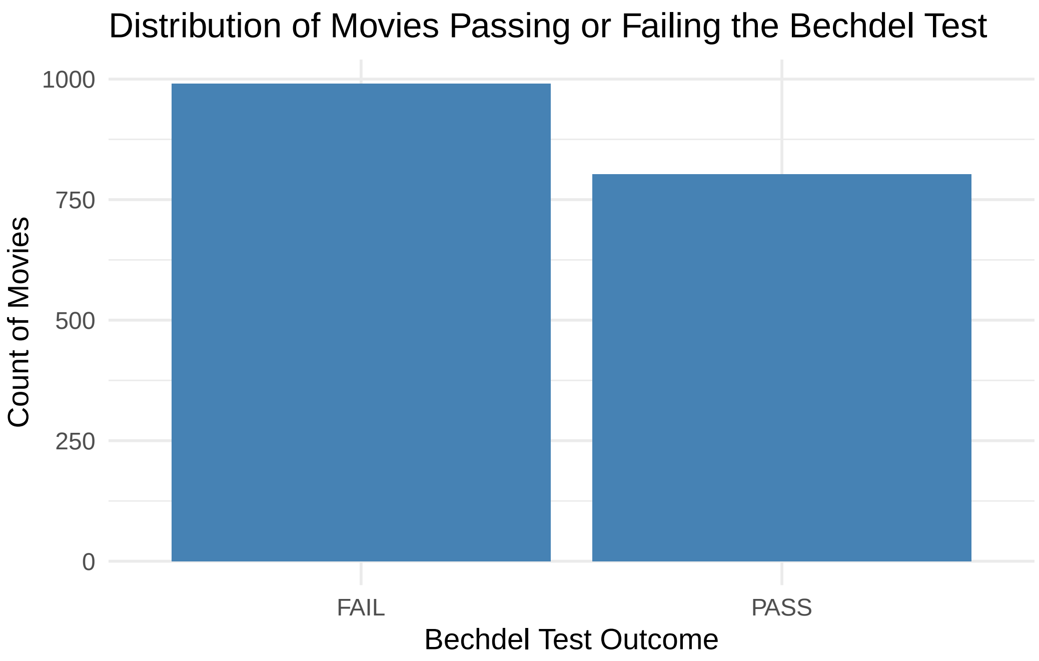
```
Rows: 1,794
Columns: 16
$ year              <dbl> 2013, 2012, 2013, 2013, 2013, 2013, 2013, 2013, 2013,~
$ imdb              <chr> "tt1711425", "tt1343727", "tt2024544", "tt1272878", "~
$ title             <chr> "21 &amp; Over", "Dredd 3D", "12 Years a Slave", "2 G~
$ test              <chr> "notalk", "ok-disagree", "notalk-disagree", "notalk",~
$ clean_test        <chr> "notalk", "ok", "notalk", "notalk", "men", "men", "no~
$ binary            <chr> "FAIL", "PASS", "FAIL", "FAIL", "FAIL", "FAIL", "FAIL~
$ budget            <dbl> 1.30e+07, 4.50e+07, 2.00e+07, 6.10e+07, 4.00e+07, 2.2~
$ domgross          <dbl> 25682380, 13414714, 53107035, 75612460, 95020213, 383~
$ intgross          <dbl> 42195766, 40868994, 158607035, 132493015, 95020213, 1~
$ code              <chr> "2013FAIL", "2012PASS", "2013FAIL", "2013FAIL", "2013~
$ `budget_2013$`    <dbl> 13000000, 45658735, 20000000, 61000000, 40000000, 225~
$ `domgross_2013$`  <dbl> 25682380, 13611086, 53107035, 75612460, 95020213, 383~
$ `intgross_2013$`  <dbl> 42195766, 41467257, 158607035, 132493015, 95020213, 1~
$ `period code`     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
$ `decade code`     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
$ genre             <chr> "Comedy", "Action", "Drama", "Action", "Drama", "Acti~
```

Exploratory Plots:

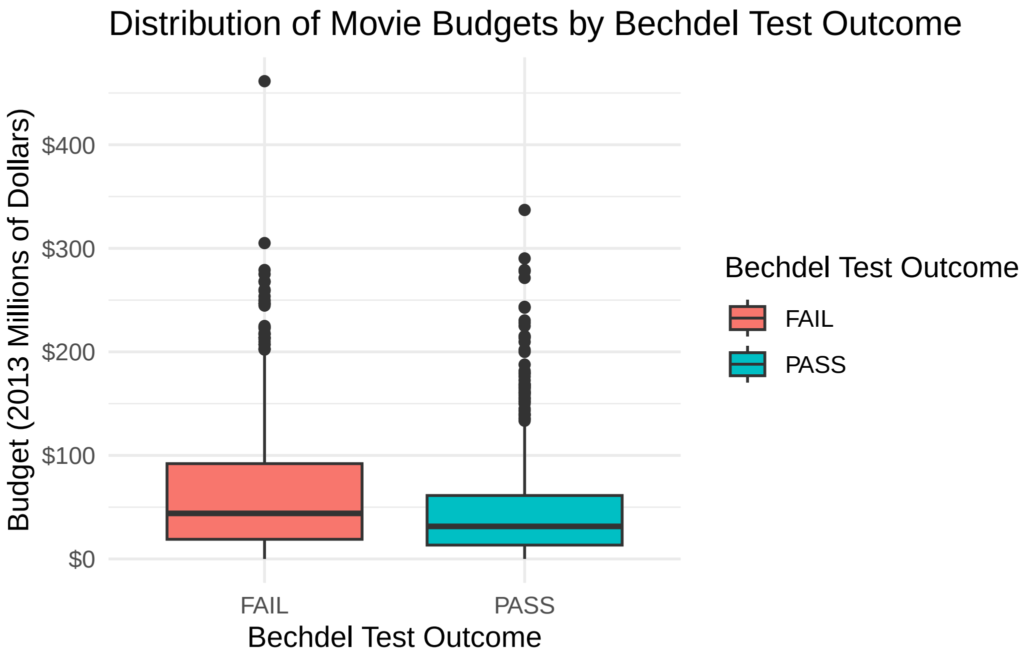
```r
# QUESTION 1

movies$budget_2013_numeric <- as.numeric(movies$`budget_2013$`)
movies$budget_2013_millions <- movies$budget_2013_numeric / 1e6


# OUTCOME VAR: Bar plot for Bechdel Test
ggplot(movies, aes(x = binary)) +
  geom_bar(fill = "steelblue") +
  labs(x = "Bechdel Test Outcome", y = "Count of Movies") +
  ggtitle("Distribution of Movies Passing or Failing the Bechdel Test") +
  theme_minimal()
```

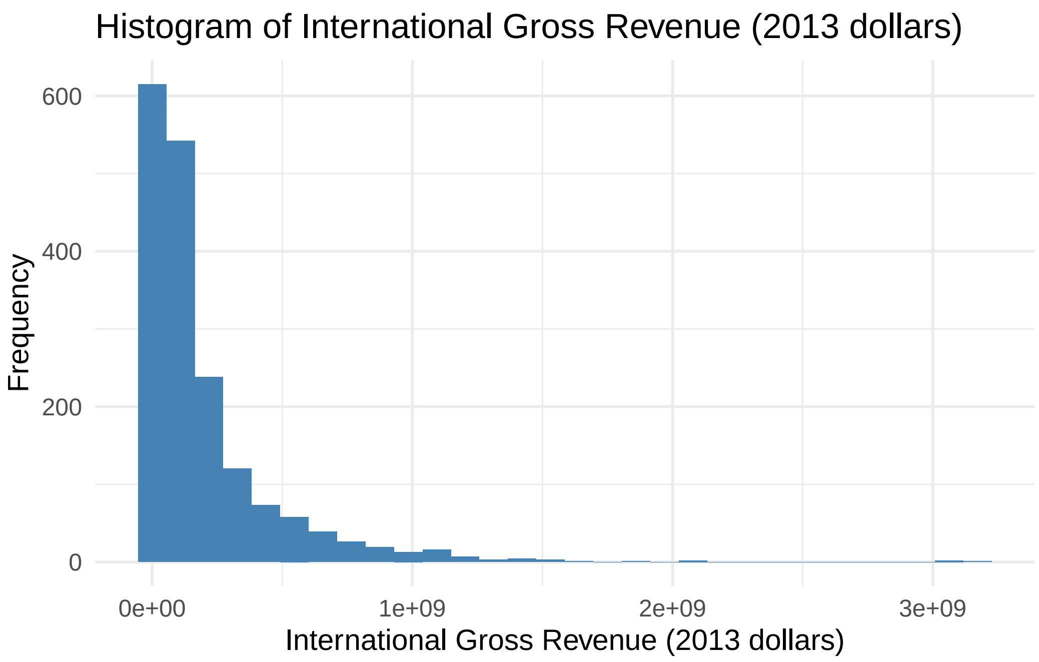


```r
# Box plot: Budget by Bechdel Test outcome
ggplot(movies, aes(x = binary, y = `budget_2013_millions`, fill = binary)) +
  geom_boxplot() +
  scale_y_continuous(labels = scales::dollar) +
  labs(x = "Bechdel Test Outcome", y = "Budget (2013 Millions of Dollars)",
       fill = "Bechdel Test Outcome") +
  ggtitle("Distribution of Movie Budgets by Bechdel Test Outcome") +
  theme_minimal()
```



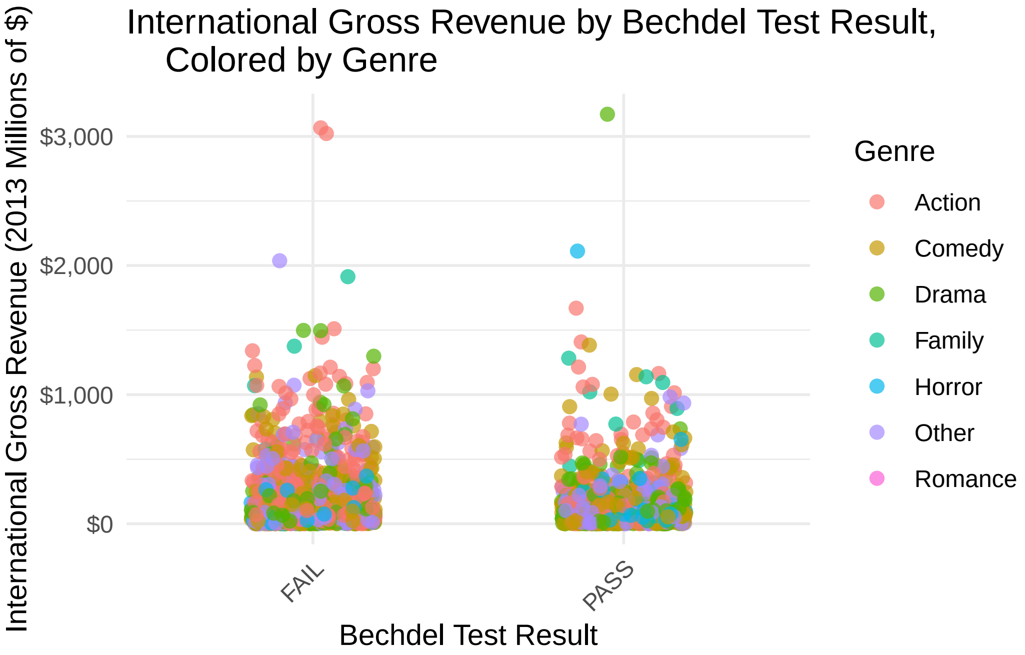
```
# QUESTION 2

# OUTCOME VAR: Histogram for international gross revenue
ggplot(movies, aes(x = `intgross_2013$`)) +
  geom_histogram(fill = "steelblue", bins = 30) +
  labs(x = "International Gross Revenue (2013 dollars)", y = "Frequency") +
  ggtitle("Histogram of International Gross Revenue (2013 dollars)") +
  theme_minimal()
```



```r
# Convert the intgross_2013$ column to numeric after removing any commas
movies$intgross_2013_numeric <- as.numeric(movies$`intgross_2013$`)
movies$intgross_2013_millions <- movies$intgross_2013_numeric / 1e6

# Scatter plot with genre coloring and Bechdel Test result on x-axis
ggplot(movies, aes(x = binary, y = intgross_2013_millions, color = genre)) +
  geom_jitter(position = position_jitter(width = 0.2, height = 0),
              alpha = 0.7, size = 2) +
  scale_y_continuous(labels = scales::dollar) +
  labs(
    title = "International Gross Revenue by Bechdel Test Result,
    Colored by Genre",
    x = "Bechdel Test Result",
    y = "International Gross Revenue (2013 Millions of $)",
    color = "Genre"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Dataset 2

Data source: https://archive.ics.uci.edu/dataset/186/wine+quality

Brief description: The Wine Quality dataset, available from the UCI Machine Learning Repository, was created by researchers at the University of Minho in Portugal. The dataset includes two separate datasets for red and white wine, derived from the Vinho Verde region. It was initially collected in 2009 to explore the relationship between different physicochemical properties of wine and its quality rating.

Each row in the dataset represents a single wine sample, with columns detailing various characteristics of the wine, such as acidity, sugar content, pH, and alcohol content. There is also a quality rating for each sample, given as a score between 0 and 10, which was provided by sensory assessors. The dataset is commonly used to model and predict wine quality based on these physicochemical properties.

Research question 1: Are there differences in the impact of chemical properties on quality ratings between red and white wines?

- Outcome variable (include the name/description and type of variable): quality (Continuous)
- Interaction Terms: wine_type * chemical_property for each chemical property

Research question 2: What chemical properties most significantly distinguish red wines from white wines, and do combinations of properties enhance this distinction?

- Outcome variable (include the name/description and type of variable): wine_type (Categorical: Red, White)
- Interaction Terms: Selected pairs of predictors, such as alcohol * sulphates, density * residual.sugar, pH * fixed.acidity, etc., to capture how combinations of chemical properties might uniquely differentiate red from white wines.

Load the data and provide a `glimpse()`:

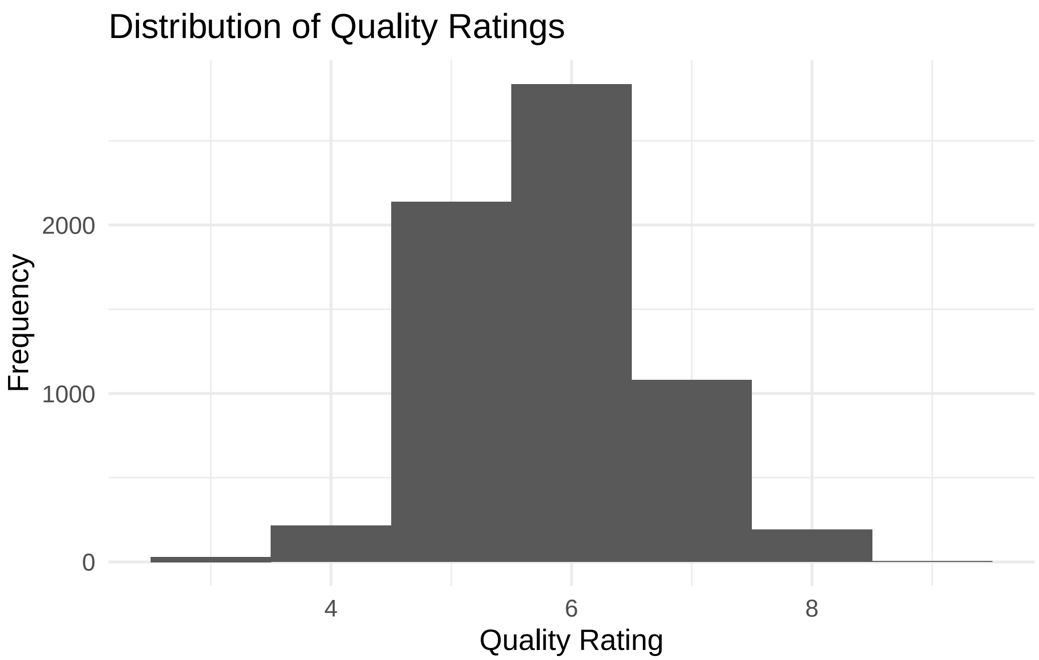
```r
wine <- read.csv("https://raw.githubusercontent.com/ackerman-alex/IDS_702_Final_Project/refs/
glimpse(wine)
```

```
Rows: 6,497
Columns: 13
$ fixed.acidity        <dbl> 7.4, 7.8, 7.8, 11.2, 7.4, 7.4, 7.9, 7.3, 7.8, 7.5~
$ volatile.acidity     <dbl> 0.700, 0.880, 0.760, 0.280, 0.700, 0.660, 0.600, ~
$ citric.acid          <dbl> 0.00, 0.00, 0.04, 0.56, 0.00, 0.00, 0.06, 0.00, 0~
$ residual.sugar       <dbl> 1.9, 2.6, 2.3, 1.9, 1.9, 1.8, 1.6, 1.2, 2.0, 6.1,~
$ chlorides            <dbl> 0.076, 0.098, 0.092, 0.075, 0.076, 0.075, 0.069, ~
$ free.sulfur.dioxide  <dbl> 11, 25, 15, 17, 11, 13, 15, 15, 9, 17, 15, 17, 16~
$ total.sulfur.dioxide <dbl> 34, 67, 54, 60, 34, 40, 59, 21, 18, 102, 65, 102,~
$ density              <dbl> 0.9978, 0.9968, 0.9970, 0.9980, 0.9978, 0.9978, 0~
$ pH                   <dbl> 3.51, 3.20, 3.26, 3.16, 3.51, 3.51, 3.30, 3.39, 3~
$ sulphates            <dbl> 0.56, 0.68, 0.65, 0.58, 0.56, 0.56, 0.46, 0.47, 0~
$ alcohol              <dbl> 9.4, 9.8, 9.8, 9.8, 9.4, 9.4, 9.4, 10.0, 9.5, 10.~
$ quality              <int> 5, 5, 5, 6, 5, 5, 5, 7, 7, 5, 5, 5, 5, 5, 5, 5, 7~
$ wine_type            <chr> "Red", "Red", "Red", "Red", "Red", "Red", "Red", ~
```

Exploratory Plots:

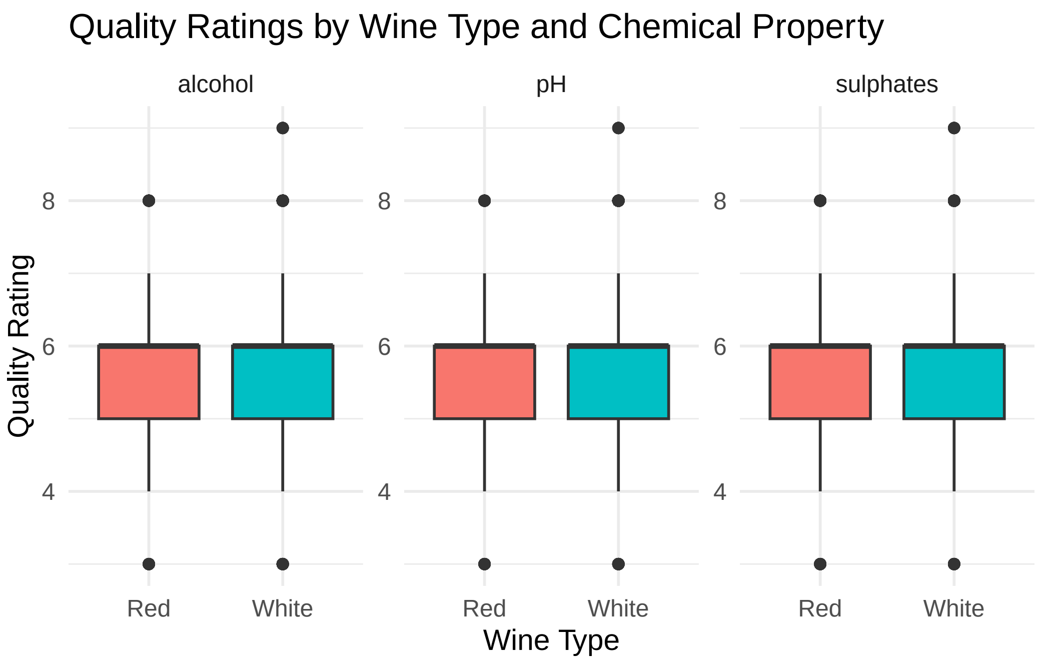
```r
# QUESTION 1

# Histogram of Quality Ratings
ggplot(wine, aes(x = quality)) +
  geom_histogram(binwidth = 1) +
  labs(title = "Distribution of Quality Ratings",
       x = "Quality Rating",
       y = "Frequency") +
  theme_minimal()
```



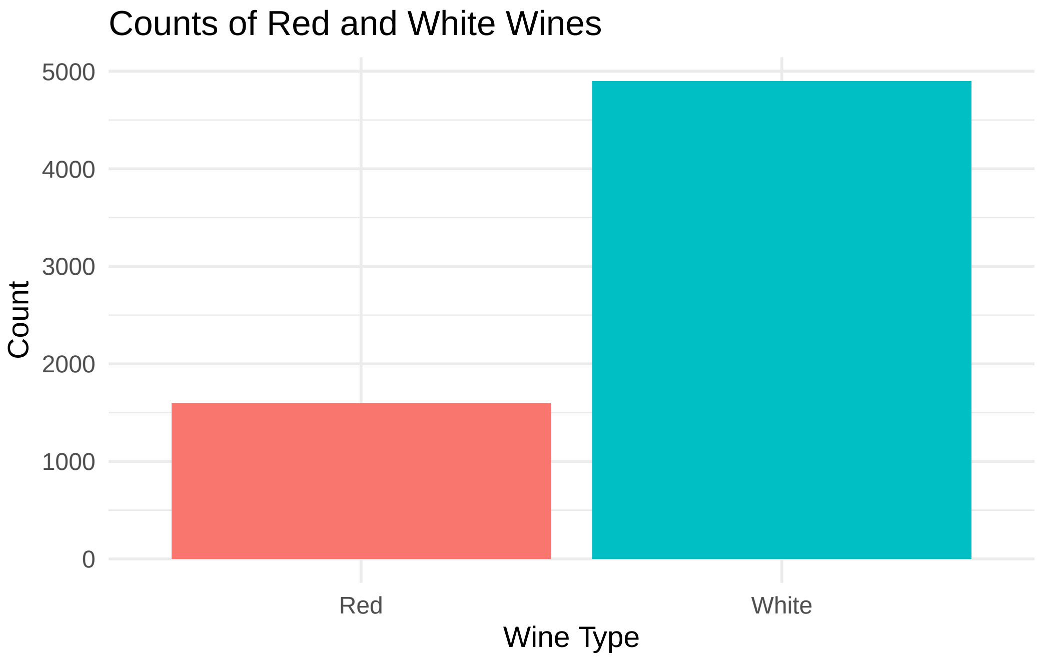
```r
# Convert data to long format for faceting by chemical property
wine_long <- wine %>%
  pivot_longer(cols = pH:alcohol, names_to = "chemical_property",
               values_to = "value")

# Boxplot of Quality Ratings by Wine Type and Chemical Property
ggplot(wine_long, aes(x = wine_type, y = quality, fill = wine_type)) +
  geom_boxplot() +
  facet_wrap(~ chemical_property, scales = "free") +
  labs(title = "Quality Ratings by Wine Type and Chemical Property",
       x = "Wine Type",
       y = "Quality Rating") +
  theme_minimal() +
  theme(legend.position = "none")
```

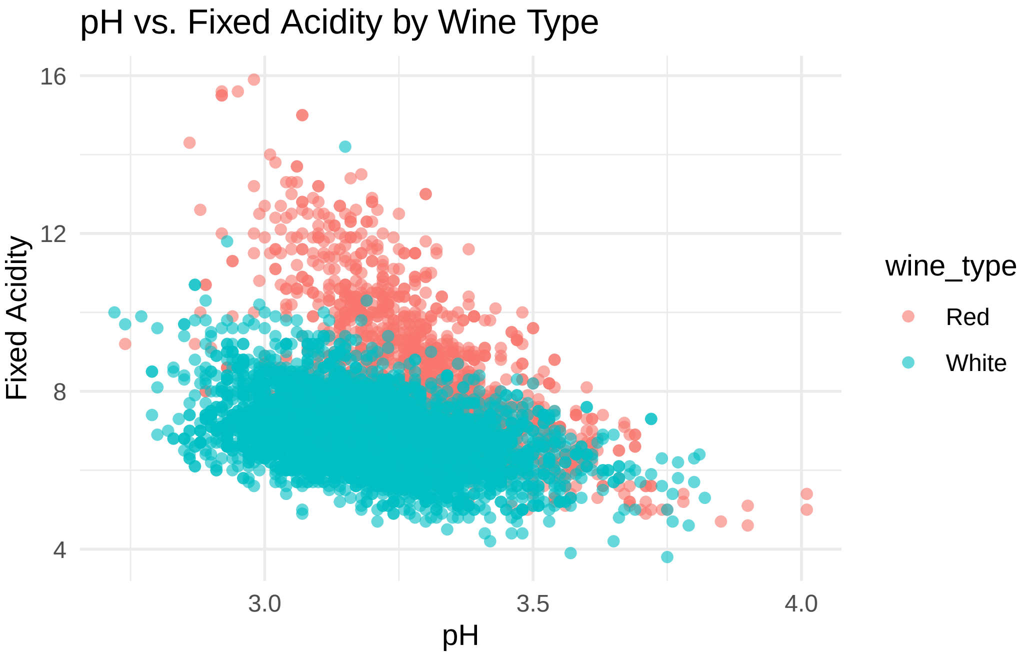


```
# QUESTION 2

# OUTCOME VAR: Bar Plot of Wine Type
ggplot(wine, aes(x = wine_type, fill = wine_type)) +
  geom_bar() +
  labs(title = "Counts of Red and White Wines",
       x = "Wine Type",
       y = "Count") +
  theme_minimal() +
  theme(legend.position = "none")
```



```
# Scatter Plots for Selected Chemical Property Pairs
ggplot(wine, aes(x = pH, y = fixed.acidity, color = wine_type)) +
  geom_point(alpha = 0.6) +
  labs(title = "pH vs. Fixed Acidity by Wine Type", x = "pH",
       y = "Fixed Acidity") +
  theme_minimal()
```



Team Charter

When will you meet as a team to work on the project components? Will these meetings be held in person or virtually?

Our team plans to meet weekly on Wednesday's to check in and allocate work. These meetings will be in-person and will allow us to assign work to members of our group for them to work on individually .

What is your group policy on missing team meetings (e.g., how much advance notice should be provided)?

We understand that now is a very busy time of the year for all of our teammates. However, it is important that we respect each others time. We expect all members to attend meetings as much as they are able. If they are not able to attend, we expect members to let the group know their reason for missing the meeting as well as suggest how they can contribute at a different time to make up for the lost time.

How will your team communicate (email, Slack, text messages)? What is your policy on appropriate response time (within a certain number of hours? Nights/weekends?)?

We have been communicating exclusively over Slack messages, and this has worked well for us thus far. We expect all teammates to respond in a prompt manner, but again, we are all very busy, so we will all be flexible. We are comfortable communicating at night and on the weekends.