Comparing Frequency Distributions: Takeaways

Syntax

- Generating a grouped bar plot from raw data:

```
library(ggplot2)

ggplot(data = df,

         aes(x = col_1, fill = col_2)) +

  geom_bar(position = "dodge")
```

- Generating a grouped bar plot from a frequency distribution table grouped by two variables:

```
frequency_table <- df %>%

  group_by(col_1, col_2) %>%

  summarize(Freq = n())

ggplot(data = frequency_table,

         aes(x = col_2, y = Freq, fill = col_1)) +

  geom_bar(position = "dodge", stat = "identity")
```

- Generating overlapping histograms for two categories from a single variable:

```
ggplot(data = df,

         aes(x = continuous_variable,

             fill = categorical_variable)) +

  geom_histogram(bins = 10,

                 position = "identity",

                 alpha = 0.5)
```

- Overlapping histograms with mean value displayed as vertical line:

```
ggplot(data = df,

         aes(x = continuous_variable,

             fill = categorical_variable)) +

  geom_histogram(bins = 10,

                 position = "identity",

                 alpha = 0.5) +

  geom_vline(aes(xintercept = mean(df$continuous_variable),

                 linetype = "Descriptive title"),

             color = "black")
```

- Side-by-side histograms with mean value displayed as vertical line:

```
ggplot(data = df,
         aes(x = continuous_variable,
             fill = categorical_variable)) +
  geom_histogram(bins = 10,
                 position = "identity",
                 alpha = 0.5) +
  geom_vline(aes(xintercept = mean(df$continuous_variable),
                 linetype = "Descriptive title"),
             color = "black") +
  facet_wrap(~ categorical_variable)
```

- Generating kernel density plots for two categories from a single variable:

```
ggplot(data = wnba,
         aes(x = continuous_variable,
             color = categorical_variable)) +
  geom_density()
```

- Generating strip-style scatter plots with jitter:

```
ggplot(data = df,
       aes(x = categorical_variable,
           y = continuous_variable,
           color = categorical_variable)) +
  geom_point() +
  geom_jitter()
```

- Generating multiple box plots:

```
ggplot(data = df,
         aes(x = categorical_variable,
             y = continuous_variable,
             color = categorical_variable)) +
  geom_boxplot()
```

Concepts

- To compare visually frequency distributions for nominal and ordinal variables, we can use **grouped bar charts**.
- To compare visually frequency distributions for variables measured on an interval or ratio scale, we can use:
    - **Overlaid histograms**
    - **Kernel density plots**
    - **Scatter plots**
    - **Box plots**
- A value that is much lower or much larger than the rest of the values in a distribution is called an **outlier**. A value is an outlier if:
    - It's larger than the upper quartile by 1.5 times the interquartile range.
    - It's lower than the lower quartile by 1.5 times the interquartile range.

Resources

- tidyverse documentation and examples of bar charts.
- tidyverse documentation and examples of histograms and frequency polygons.
- tidyverse documentation and examples of scatter plots.
- tidyverse documentation and examples of box plots.