Visualizing Data Distributions: Takeaways

Syntax

Creating a bar chart:

```
data_frame %>%
      ggplot(aes(x = variable_1, y = variable_2)) +
      geom_col()
```

Creating a histogram:

```
reviews %>%
      ggplot(aes(x = Rating)) +
      geom_histogram(bins = 30)
```

Creating a density plot:

```
reviews %>%
      ggplot(aes(x = Rating)) +
      geom_density()
```

Adding color to distinguish between groups:

```
reviews %>%
      ggplot(aes(x = Rating, fill= Rating_Site)) +
      geom_histogram(bins = 30)
```

Creating a boxplot:

```
reviews %>%
      ggplot(aes(x = Rating_Site, y = Rating)) +
      geom_boxplot()
```

Using factor() to create a factor variable:

```
reviews <- reviews %>%
      mutate(
        Rating_Site_cat = factor(Rating_Site, levels =
```

```
c("Rotten_Tomatoes", "Metacritic", "IMDB", "Fandango"))
      )
```

Concepts

- Bar charts:
    - Represent grouped data summaries using bars with heights proportional to values of a summary variable such as the average.
    - Do not provide information about the distribution of variables.
- Using `stat = "identity"` overrides the default behavior of the height of the bars corresponding to the number of values, and instead creates bars equal to the value of the y-variable.
- Histograms depict the frequency with which values of a variable occur. Unlike bar charts and line graphs, histograms are used to understand characteristics of one variable rather than the relationship between two variables.
- You can specify two different arguments in the `geom_histogram()` layer to specify the number of categories for binning the independent variable:
    - `binwidth =` allows you to specify the *size* of the bins, and is useful for instances, such as this example, where you want categories to span specific intervals.
    - `bins =` allows you to specify the *number* of bins, which can be useful to experiment with when deciding how much detail you want to use to display your data.
- Box plots provide a summary of data for each group, as well as provide information about how data is spread.
- Box plots present the following data:
    - The largest value: Represented by the top of the black line extending from the top of the box. These lines are also known as "whiskers".
    - The third quartile (Q3): Represented by the top of the box. 75% of the values are smaller than the third quartile.
    - The median: Represented by the thick black line. The median is the value that falls in the middle of the data.
    - The first quartile (Q1): Represented by the bottom of the box. 25% of the values are smaller than the third quartile.
    - The smallest value: Represented by the bottom of the black line extending from the bottom of the box.
- General guidelines for picking a visualization:
    - Bar charts may be used for showing a quick summary of your data, such as averages or counts of the number of instances of a value that occur for a given variable.
    - Histograms are useful for visualizing distributions of data when you want to know the *shape* of a distribution (in other words, where most values are clustered).
    - Box plots provide an informative summary of the shape, spread, and center of your data.

Resources

- Five Number Summary
- Data Visualization 101
- Design Tips for Data Visualization