Handling Overlapping

Data Visualization Techniques for Overplotting

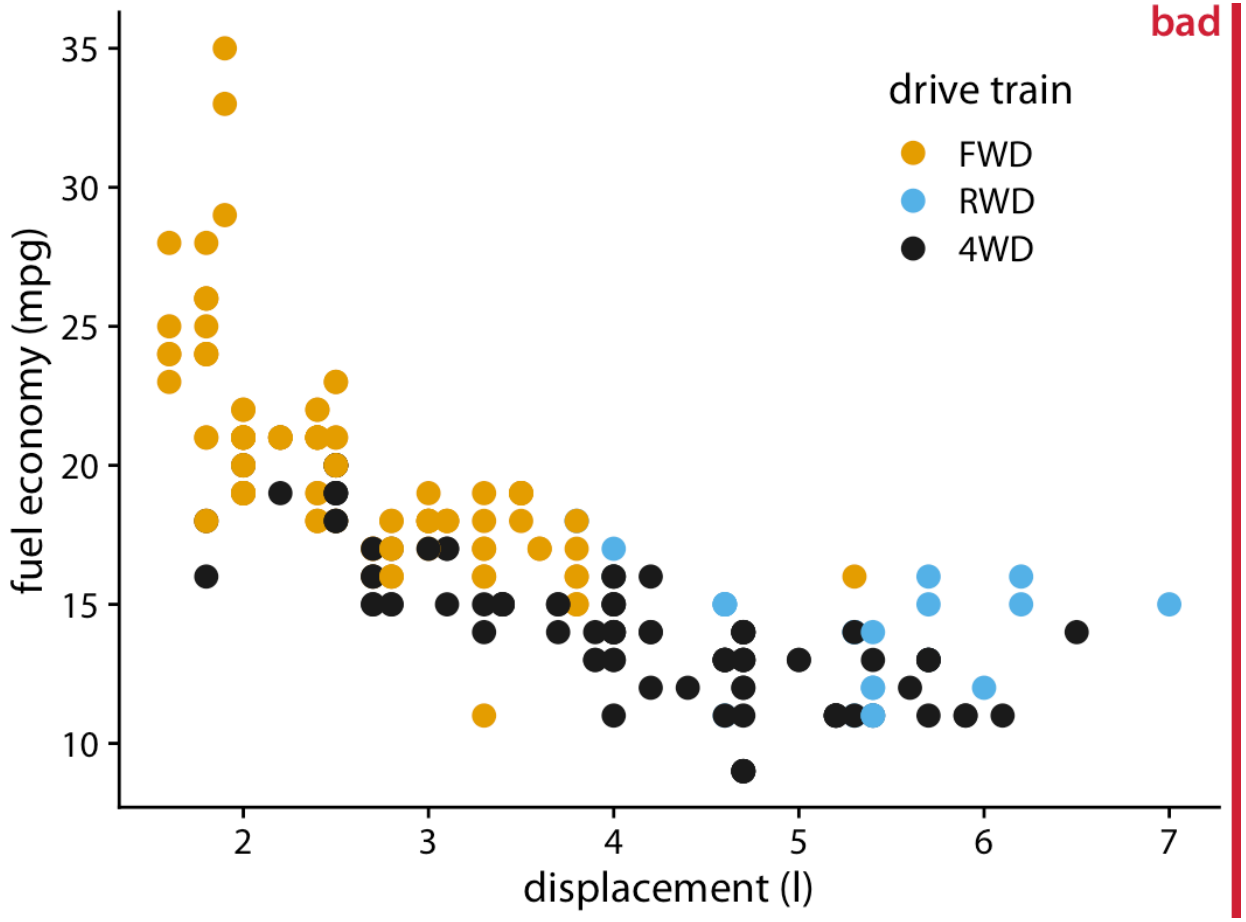
Introduction to Overplotting

- **Definition:** Overplotting occurs when many data points overlap in scatter plots, making it hard to discern patterns.
- **Problem Context:**
  - Large datasets with similar values cause points to overlap.
  - Loss of information when multiple data points share identical x–y coordinates.

Example of Overplotting

- **Visualization:**
- Display: "Bad" scatter plot showing overlapping points.
- **Explanation:**
  - Many cars with the same fuel economy and engine displacement values overlap.
  - Overlap obscures meaningful data, especially for less frequent categories like four-wheel drive cars.

City Fuel Economy vs. Engine Displacement

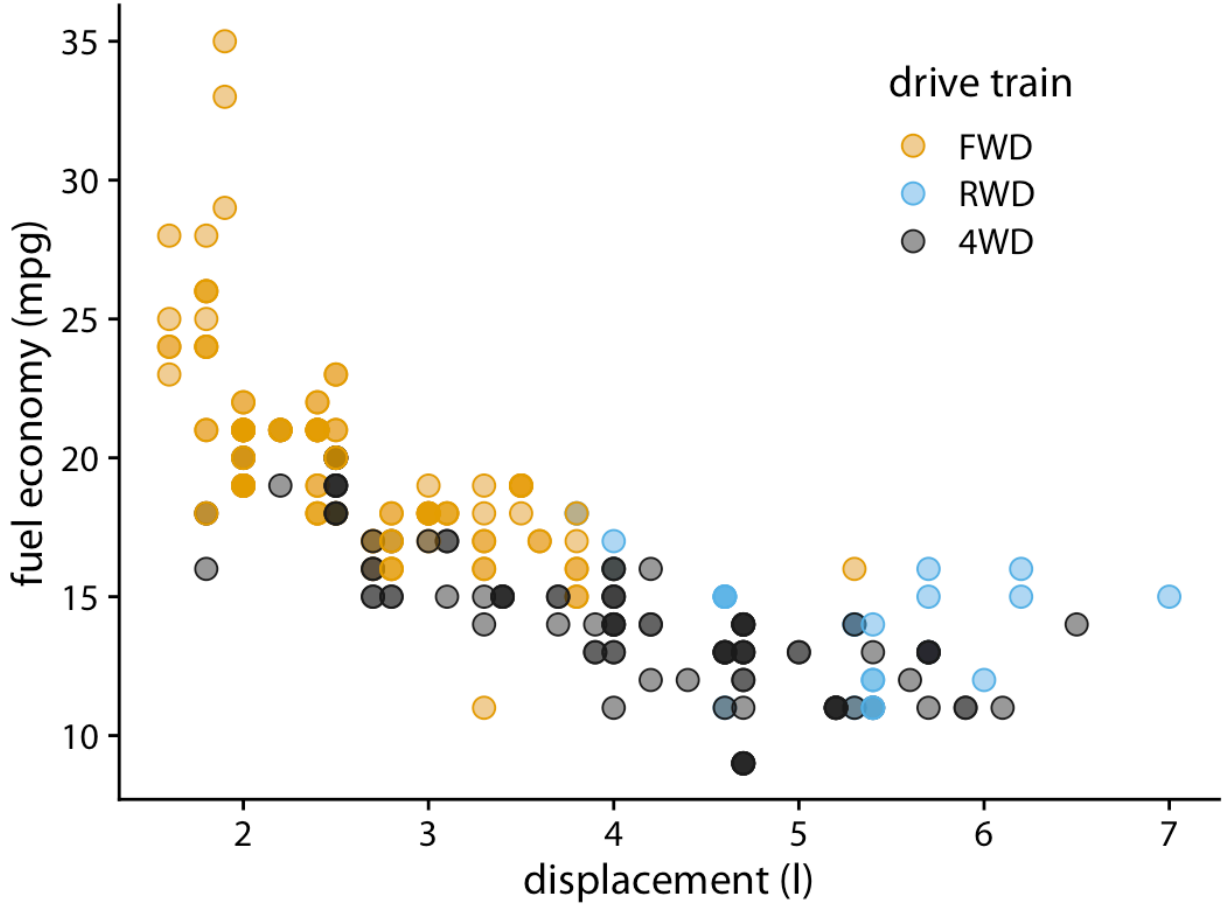


City fuel economy versus engine displacement for popular cars (1999-2008). Each point represents one car, with color indicating the drive train: FWD, RWD, or 4WD. The figure is labeled 'bad' due to overplotted points obscuring each other.

Strategy 1: Partial Transparency

- **Approach:** Make points semi-transparent to allow for visibility of overlapping points.
- **Visual Example:**
  - Display: Improved plot with partial transparency.
- **Benefit:**
  - Darker points indicate higher density.

Fuel Economy vs. Engine Displacement (Transparency)

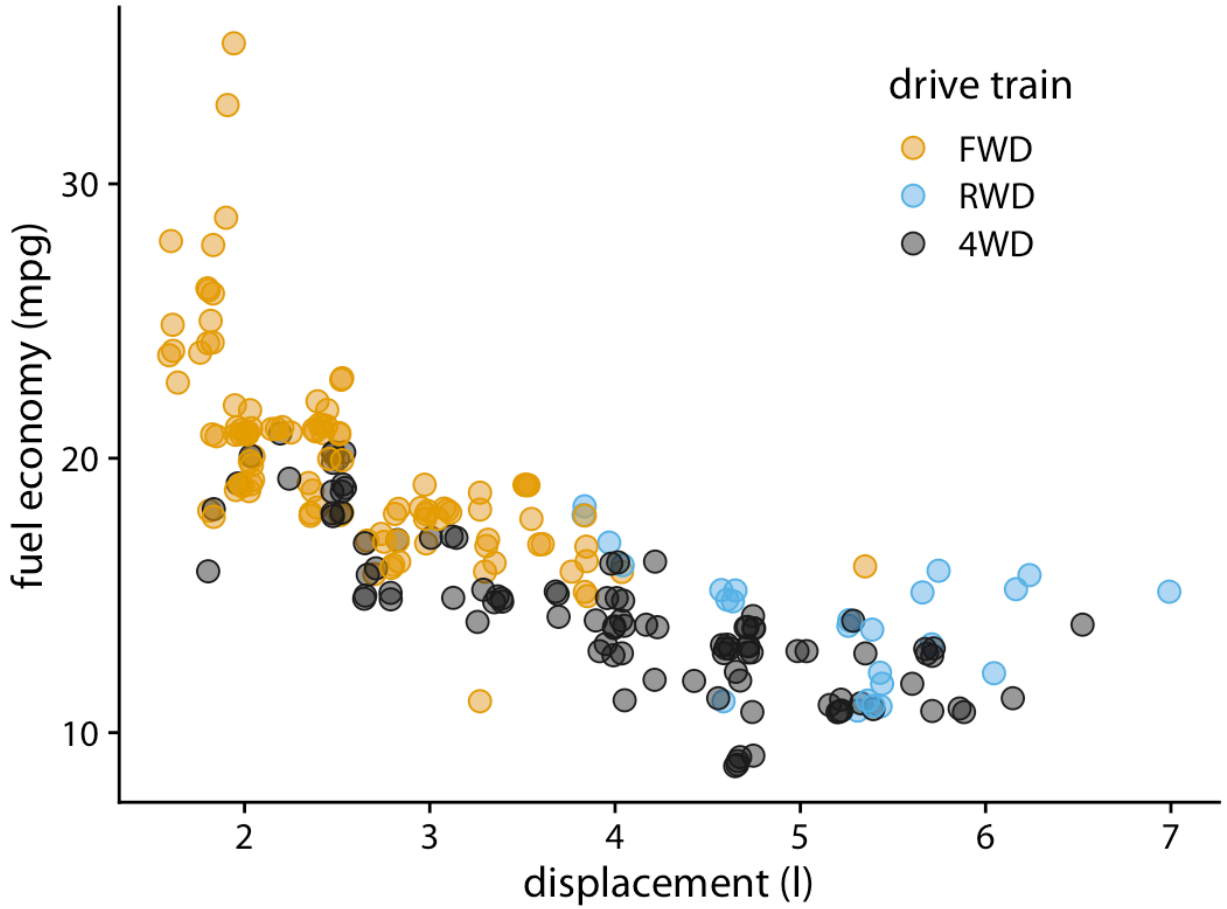


City fuel economy vs. engine displacement with partially transparent points, showing overlapping points as darker shades.

Limitations of Transparency

- **Challenge:** Even with transparency, point density can be hard to quantify.
- **Solution:** Combine transparency with jittering for better visibility.
- **Visual Example:**
  - Display: Plot with jittered points.

Fuel Economy vs. Engine Displacement (With Jitter)

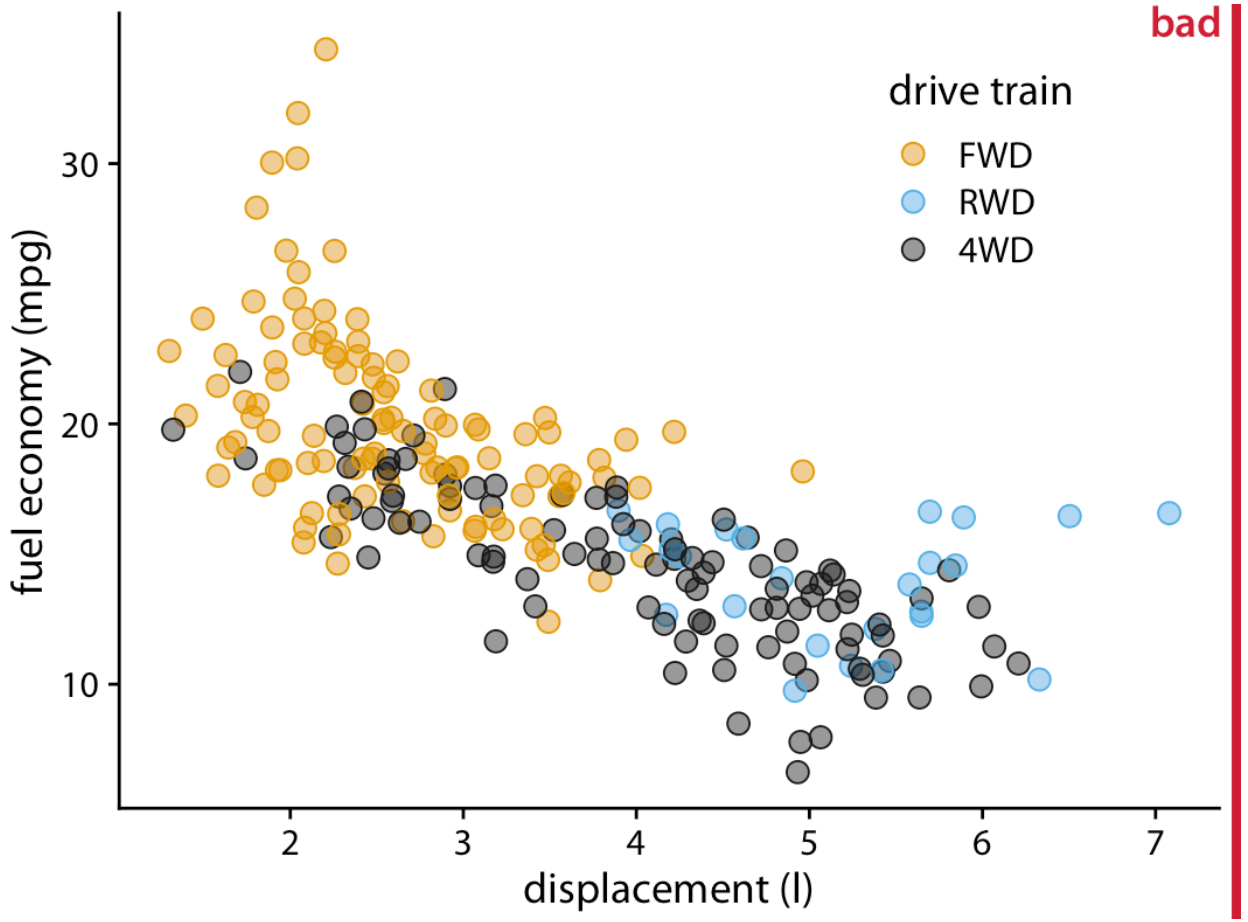


City fuel economy versus engine displacement. Adding jitter to each point makes overplotted points more visible without distorting the plot's message.

Strategy 2: Jittering

- **Approach:** Add random noise (jitter) to points to avoid exact overlap.
- **Visual Example:**
  - Display: Jittering showing better distribution of points.
- **Note:** Excessive jittering can distort the data and lead to misleading representations.
- **Visual Example:**
  - Display: Over-jittered plot that misrepresents data.

Excessive Jitter Distortion

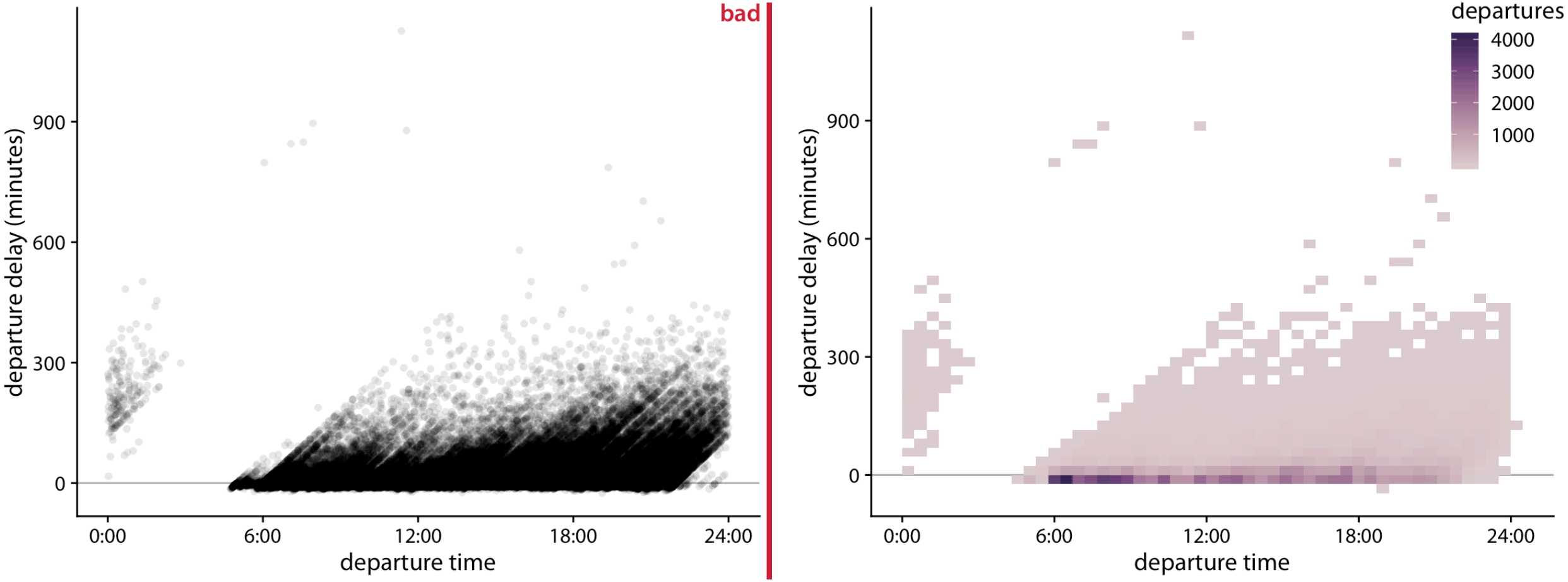


City fuel economy versus engine displacement. Excessive jitter creates a visualization that misrepresents the dataset.

Strategy 3: 2D Histograms

- **Approach:** For large datasets, use 2D histograms to bin data into grid cells.
- **Visual Example:**
  - Display: Flight departure delays visualized using a 2D histogram.
- **Benefit:**
  - Allows clearer visualization of data density.
  - Better for large datasets where individual points are difficult to see.

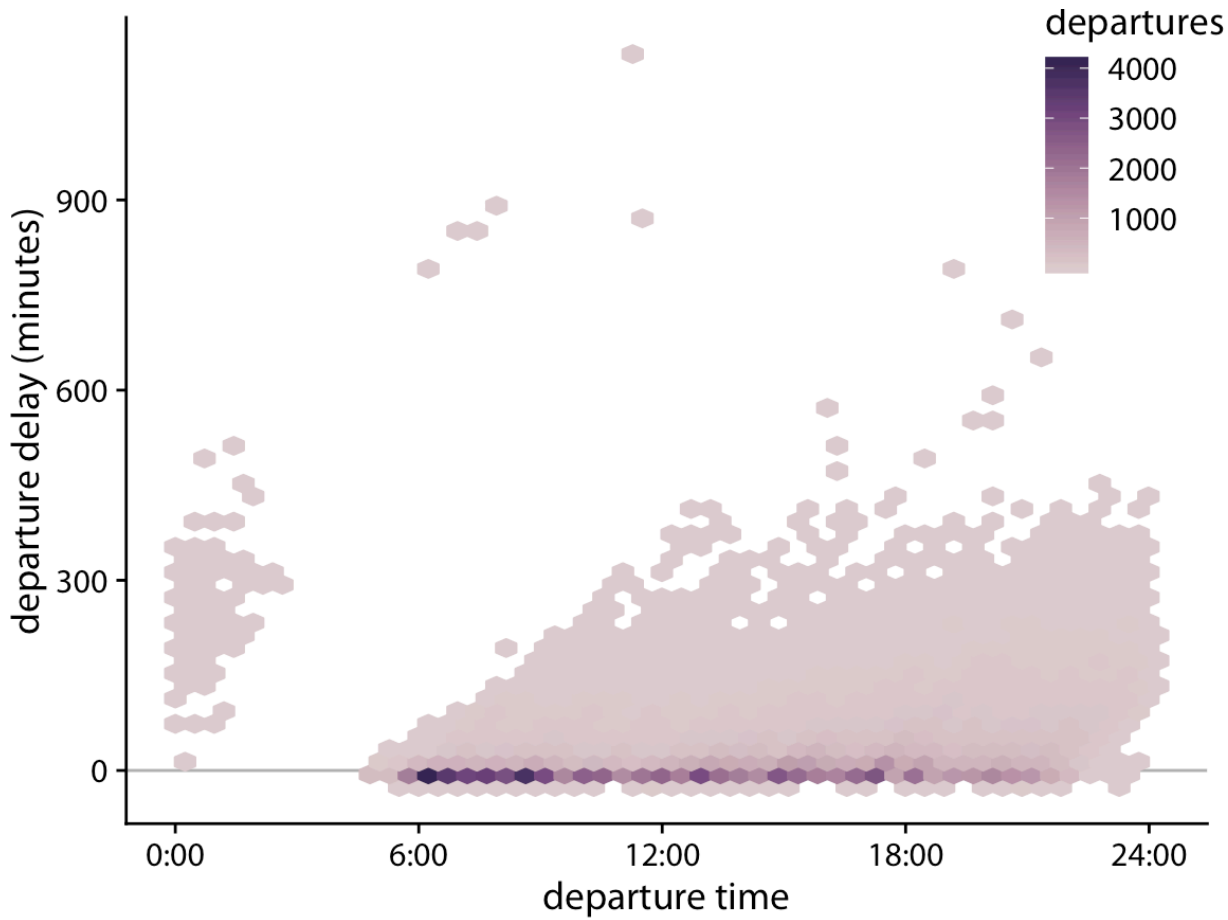
Flight Departure Delays



Departure delay versus flight departure time for all Newark airport flights (2013). Dots represent individual departures, while colored rectangles show flight counts by time and delay.

Strategy 4: Hexagonal Binning

- **Approach:** Instead of rectangular bins, use hexagons for more accurate data representation.
- **Visual Example:**
  - Display: Hexagonal binning for flight departure delays.
- **Benefit:**
  - Provides a more uniform and compact representation of point density.

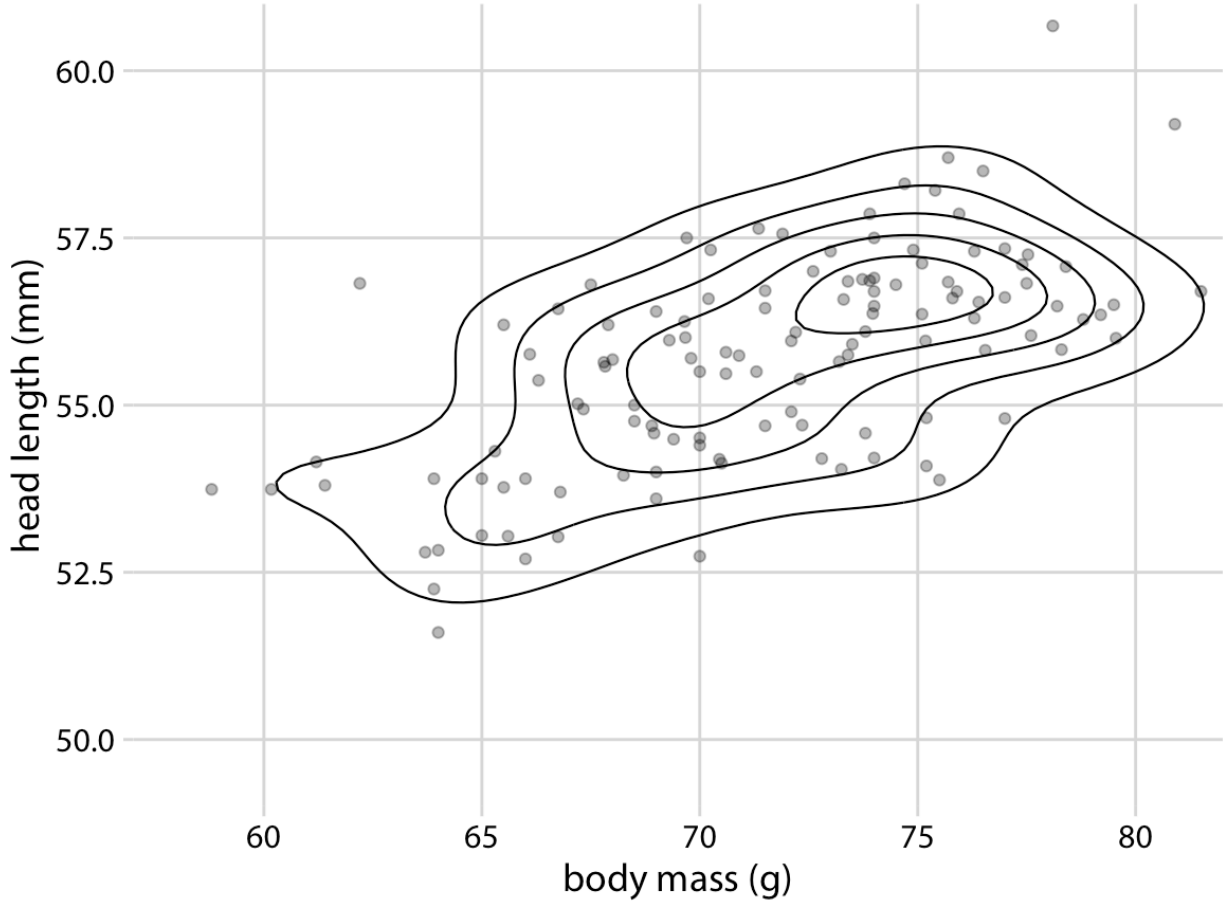


Departure delay versus flight departure time. Each colored hexagon represents flights by time and delay, with color indicating flight count.

Strategy 5: Contour Lines

- **Approach:** Estimate point density and use contour lines to represent regions of similar density.
- **Visual Example:**
  - Display: Plot with contour lines showing point density.
- **Benefit:**
  - Visualizes areas of high and low point density without cluttering the plot.

Head Length vs Body Mass Density

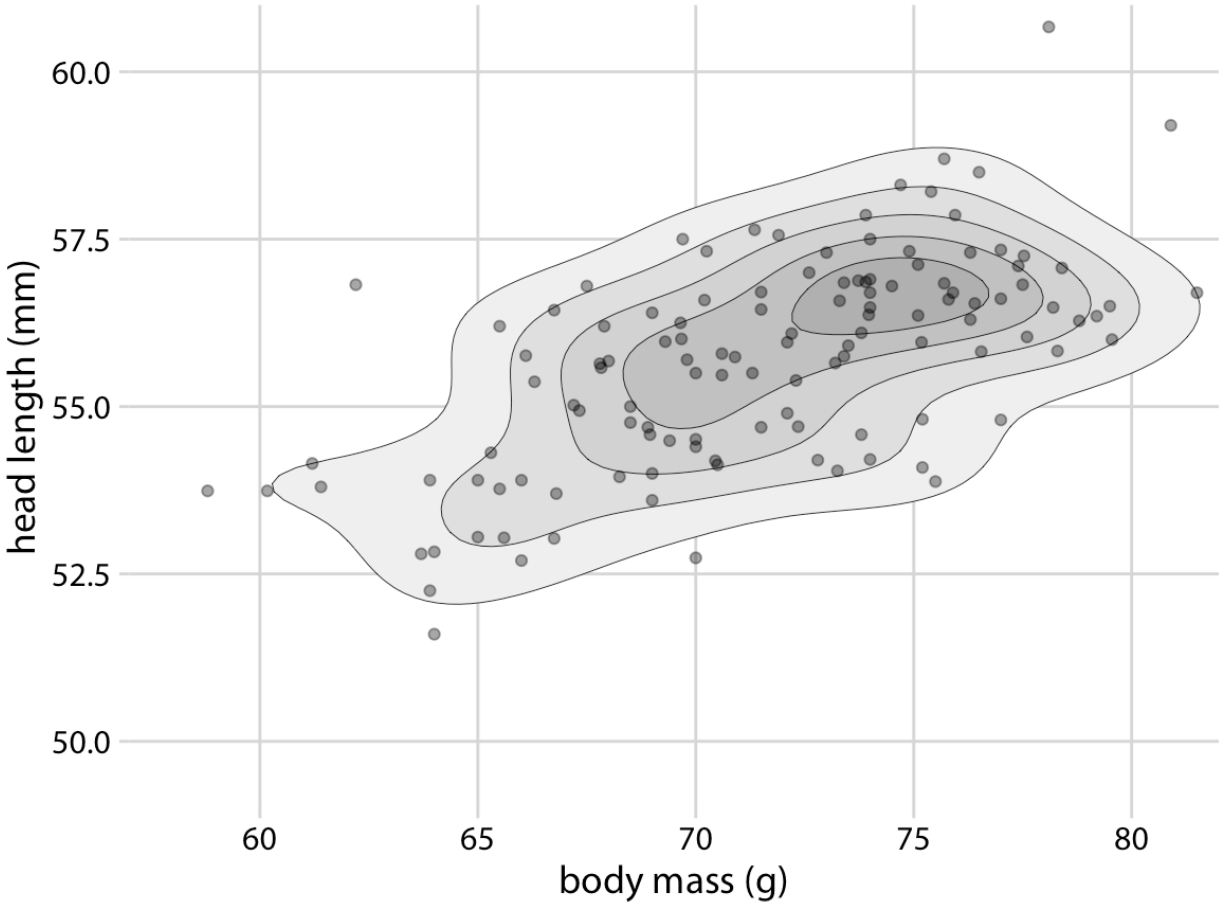


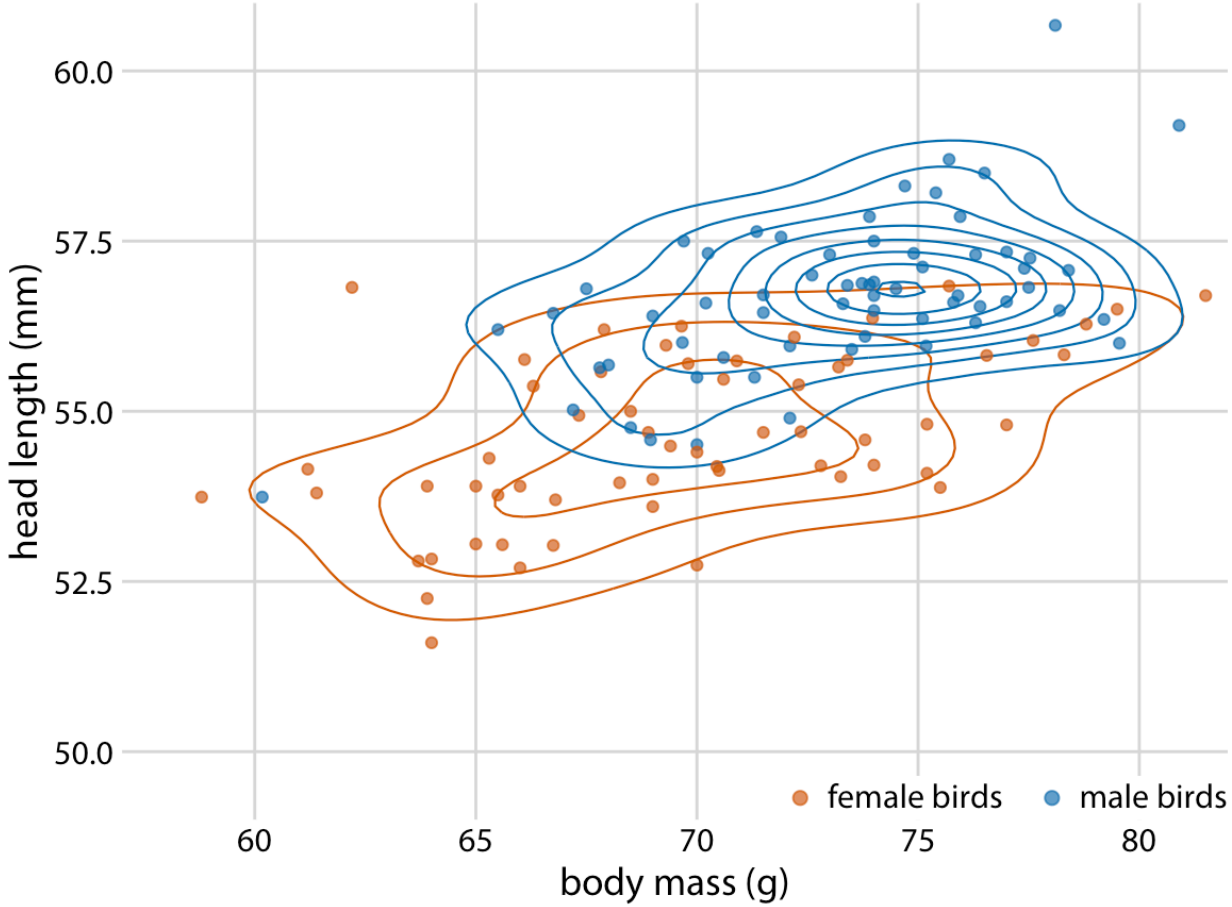
Head length versus body mass for 123 blue jays. Lines indicate regions of similar point density, highest near 75g mass and 55-57.5mm head length.

Case Study: Multiple Contour Lines

- **Example:**
  - Display: Head length versus body mass for blue jays with separate contour lines for males and females.
- **Challenge:** Contour lines work best when there are few groups and they are clearly separated.

Blue jay head length vs. body mass with sex differentiation



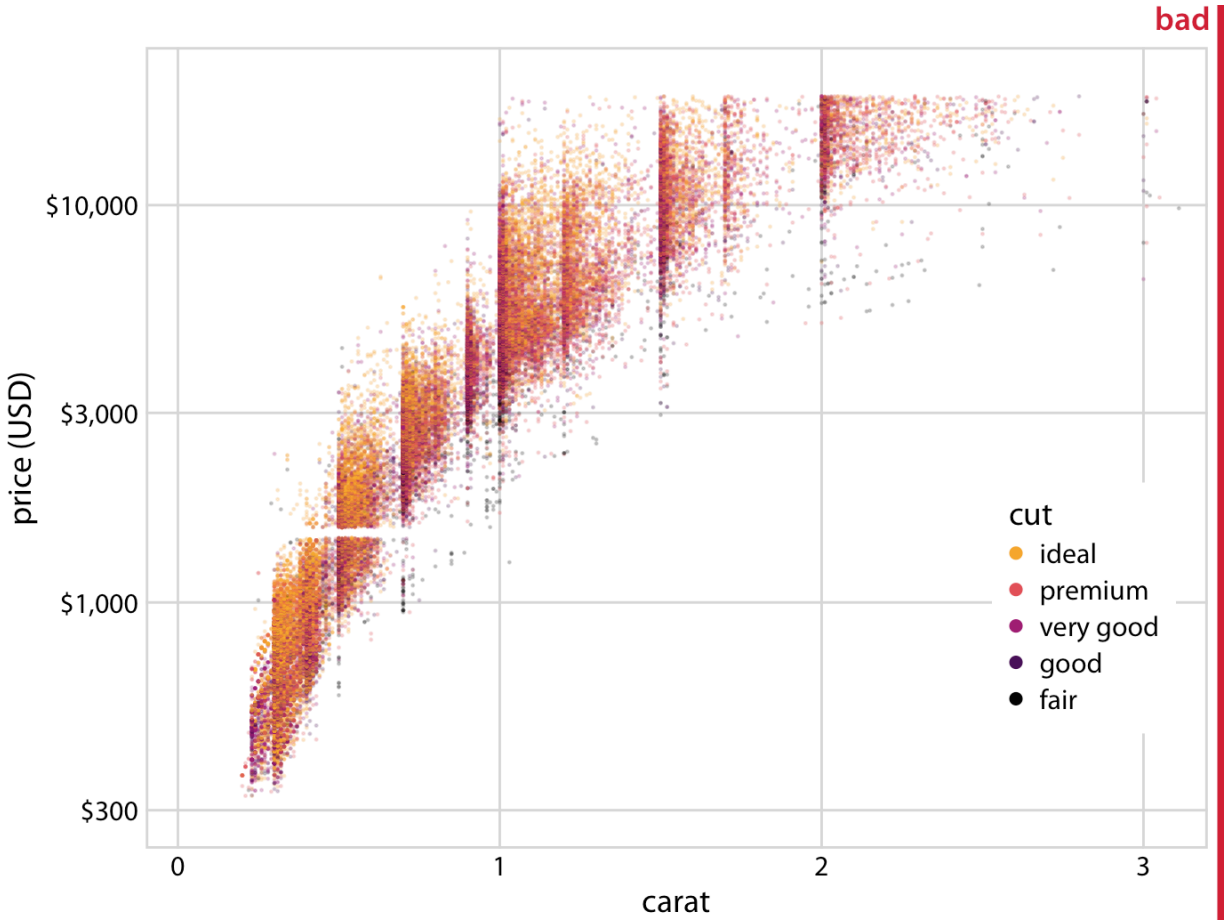


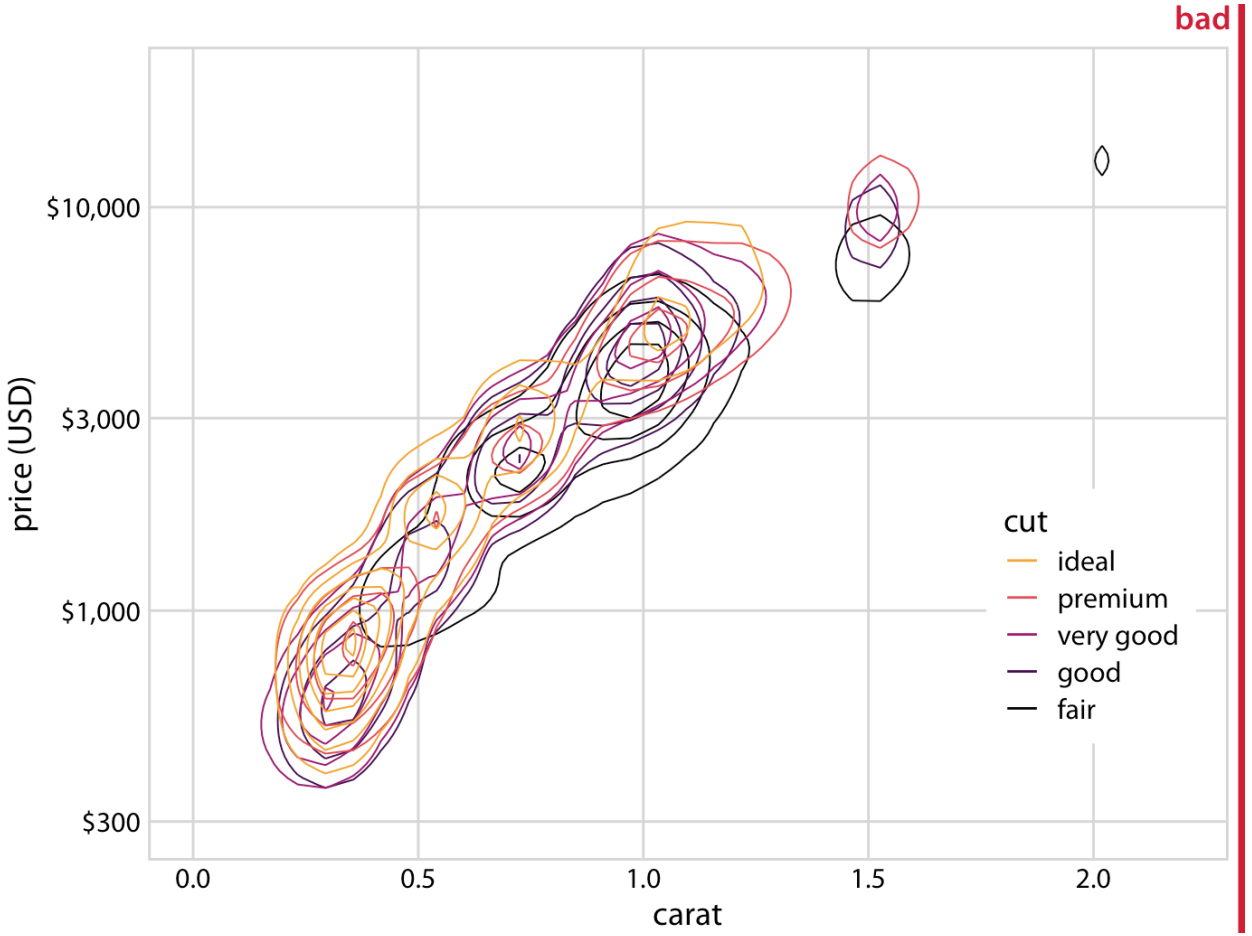
Head length vs. body mass for 123 blue jays, with shaded contours and color-coded by sex, showing clustering of males and spread of females.

Handling Multiple Groups

- **Example:**
  - Display: Overlapping scatter plot for diamonds dataset.
- **Challenge:** Multiple categories overlap, making it difficult to discern patterns.
- **Solution:** Use separate contour plots for each group.
- **Visual Example:**
  - Display: Contour lines shown separately for different diamond cuts.

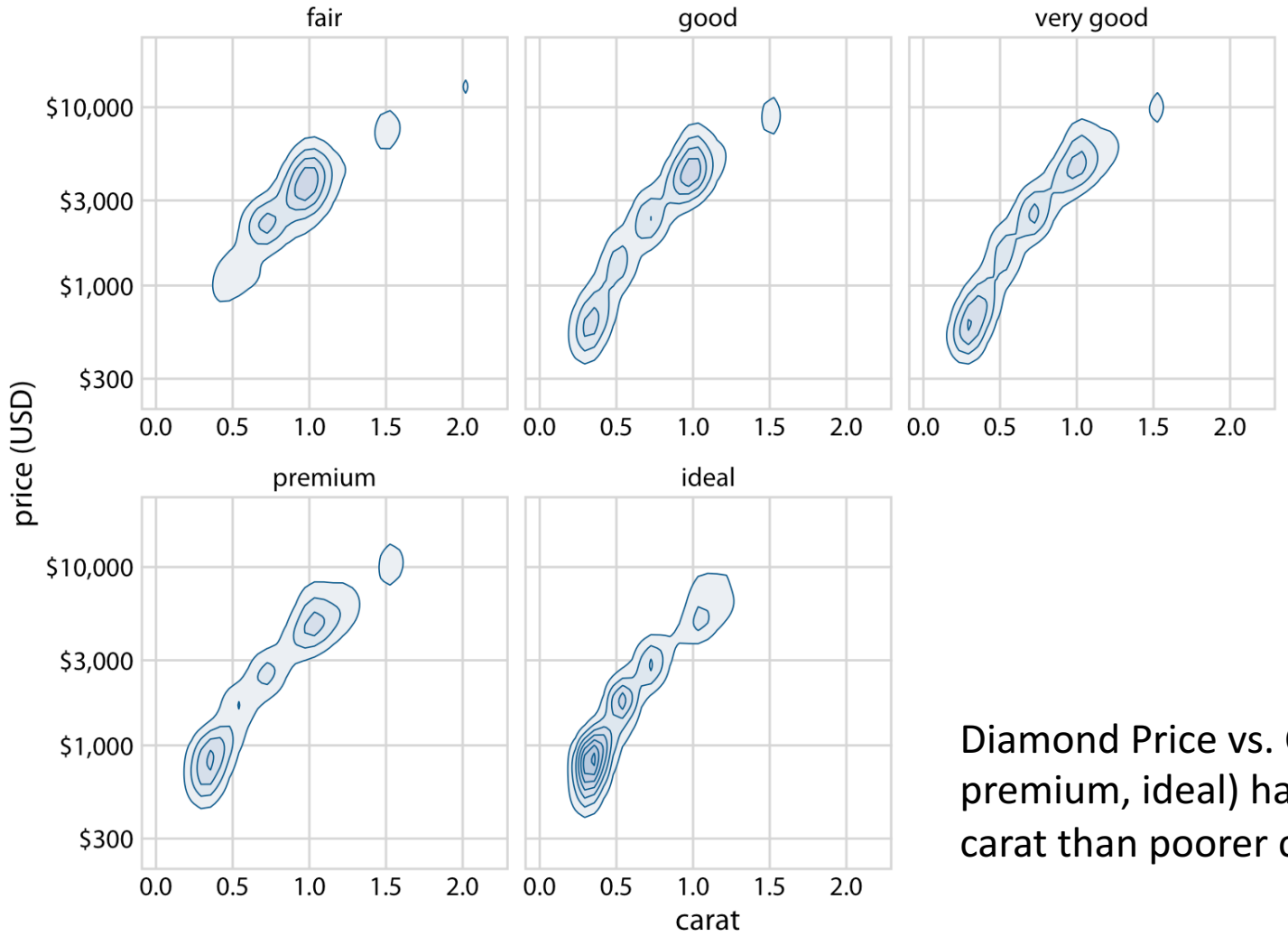
Diamond Price vs. Carat Value with Overplotting





Price of diamonds versus carat value for 53,940 diamonds, with cut indicated by color. Plot labeled “bad” due to overplotting, making it difficult to discern patterns, even after replacing individual points with contour lines, as they overlap completely.

Diamond Price vs. Carat by Cut



Diamond Price vs. Carat Value by Cut: Better cuts (very good, premium, ideal) have lower carat values but higher prices per carat than poorer cuts (fair, good).

Conclusion

- **Summary of Techniques:**
  - Partial transparency.
  - Jittering.
  - 2D histograms and hexagonal binning.
  - Contour lines.
- **Final Thoughts:**
  - Choose the strategy based on the dataset size and clarity of patterns.
  - Carefully consider how each technique might affect the accuracy of the data representation.