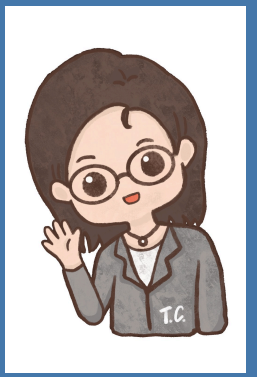
MTH 208 Exploratory Data Analysis

Lesson 04: Data Visualization Fundamentals

Ying-Ju Tessa Chen, PhD

Associate Professor
Department of Mathematics
University of Dayton

@ying-ju
ying-ju
ychen4@udayton.edu



Learning Objectives

- The importance of data visualization in EDA
- Basic chart types: histograms, bar charts, pie charts
- Principles of effective data visualization
- Advanced chart types
- Multi-dimensional data visualization
- Interpreting visualizations

The importance of data visualization in EDA

Data Visualization in EDA

- `Definition:` Data visualization in EDA involves using visual tools to explore and analyze datasets to uncover underlying patterns, correlations, outliers, and trends that might not be apparent from raw data alone.
- `Importance:` Visual methods enable quick, intuitive understanding and analysis of complex data, making them indispensable in EDA.

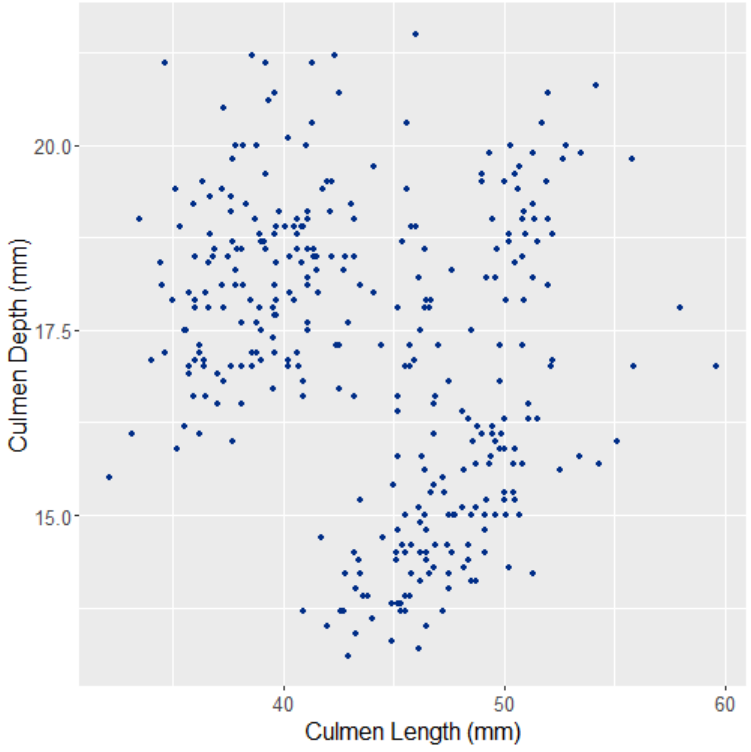
Roles of Visualization

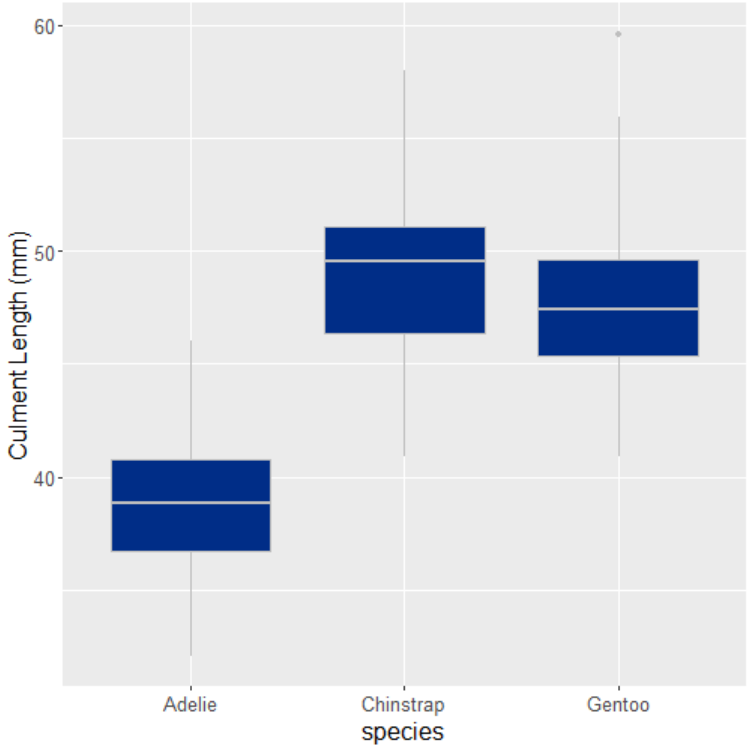
- `Insight Generation:` Visualization helps identify areas that require further analysis, guiding the EDA process.
- `Communication:` Effective visuals communicate complex data stories in an understandable manner to stakeholders with varied expertise.

The importance of data visualization in EDA (Continued)

Visualization vs. Traditional Statistical Methods

While traditional statistical methods quantify relationships and trends within data, visualization complements these by providing a qualitative perspective that can reveal unexpected insights.





Basic chart types: histograms, bar charts, pie charts

Histograms

- `Focus:` Distribution of a single continuous variable. Look for the shape of the distribution (normal, skewed, bimodal, multimodeal, uniform), identify any outliers, and assess the central and spread tendency.

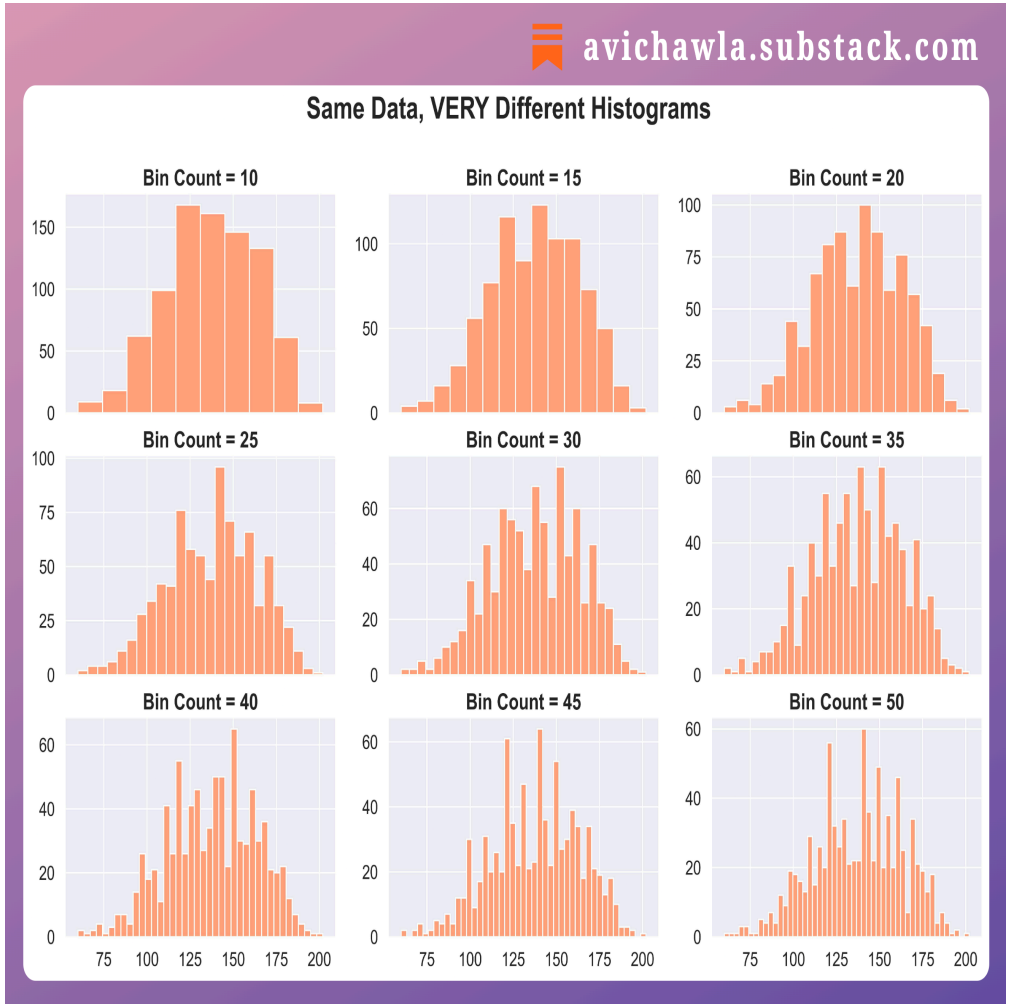
Bar Charts

- `Focus:` Comparison of discrete categories. Focus on the relative sizes of categories, trends across categories, and any anomalies.

Pie Charts

- `Focus:` Proportions of a whole. Ensure that the categories are mutually exclusive and collectively exhaustive. Pie charts are best for a small number of categories.

Basic chart types: histograms, bar charts, pie charts (Continued)

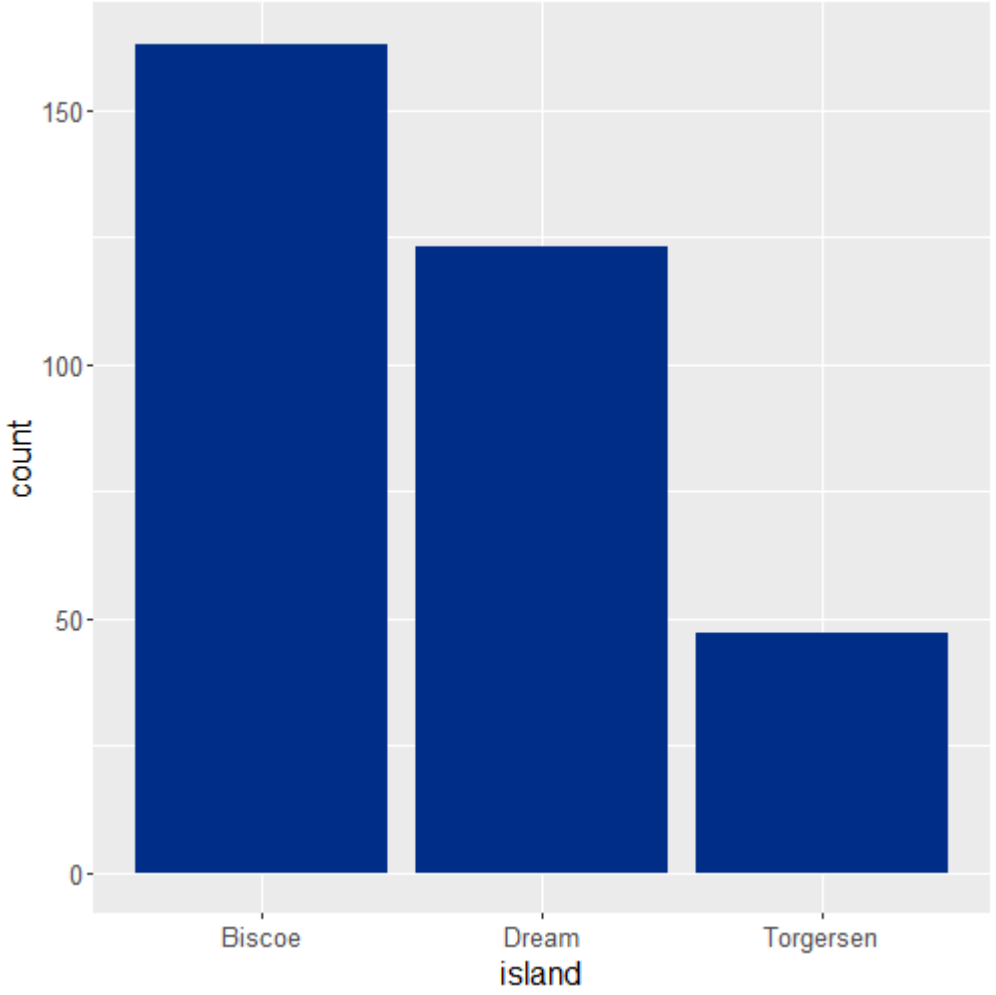


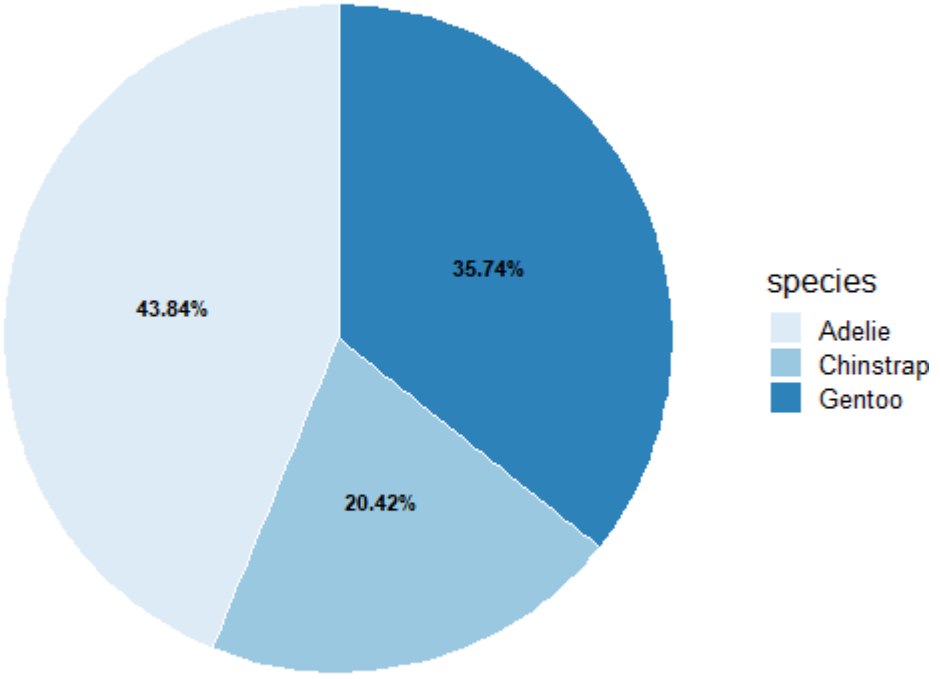
- Be aware that making the bins wider hides detail and making the bins smaller can show too much detail.
- Characterizing the shape of a distribution is often a personal judgment call. Statistical displays (for example, a histogram) only help us see what a variable's distribution "may" look like.
- A violin plot and a KDE plot can help. (Will be introduced later.)

This figure is from Daily Dose of Data Science

Basic chart types: histograms, bar charts, pie charts(Continued)

Examples of Bar charts and Pie charts





Principles of effective data visualization

Key Principles

- `Simplicity:` Avoid clutter to make the message clear.
- `Clarity:` Ensure the visualization communicates the data accurately.
- `Consistency:` Use consistent design elements for coherence.
- `Accessibility:` Design for all audiences, including those with color vision deficiencies.
- `Focus:` Guide the viewer's attention to the most important parts of the visualization.

Principles of effective data visualization (Continued)

Good Data Visualization Practices

- `Use Appropriate Chart Types:` Matching the chart type to the data's nature and the story we want to tell ensures clarity and effectiveness.
- `Keep It Simple:` Simplify the design to focus on the key message. Avoid unnecessary decorations or elements (chart junk) that do not add value to the understanding of the data.
- `Consistent and Clear Labeling:` Use clear, concise labels for axes, legends, and data points. Consistent use of terminology and units makes charts easier to understand.
- `Use Color Effectively:` Utilize color to highlight important data points or categories, not to overwhelm. Consider colorblind-friendly palettes to ensure accessibility for all audiences.
- `Provide Context:`Always provide context for our data. This could include a brief description of the data source, how the data was collected, and any assumptions made during analysis.

Principles of effective data visualization (Continued)

Good Data Visualization Practices

- `Start Axes at Zero (when applicable):` Starting y-axes at zero for bar charts ensures that the visual representation of data isn't misleading, providing a true sense of scale and difference.
- `Highlight Key Insights:` Use annotations, contrasting colors, or focus techniques to draw attention to key insights or important data points within the visualization.
- `Ensure Accuracy:` Verify that visualizations accurately represent the data without distortion. This includes checking for correct scales on axes and ensuring that data transformations are appropriate and transparent.
- `Responsive and Interactive Elements (when possible):` In digital contexts, add interactive elements like tooltips, filters, or drill-down capabilities to allow users to explore the data in more depth.
- `Adopt a Grid Layout:` Organize visual elements in a grid layout to create a clean, aligned structure that's easy to navigate and understand.

Principles of effective data visualization (Continued)

Bad Data Visualization Practices

- `Presenting Qualitative Data by Showing the DataTable`: Directly displaying raw data tables for qualitative information can overwhelm the audience, making it hard to discern patterns or key insights without visual aids.
- `Pie Chart with Too Many Categories`: Pie charts are most effective with a small number of categories.
- `Multi-colored Bar Charts:` Using more colors can make a visualization harder to understand because it increases the number of categories the brain needs to process.
- `Too Much Information:` Bombarding a visualization with excessive detail, data points, or variables can overwhelm the viewer, leading to confusion rather than clarity.
- `3D Graphs:` While aesthetically appealing, 3D graphs can distort perception, making it difficult to accurately compare values and identify trends. They often introduce unnecessary complexity in interpreting data.
- `Charts That Don’t Start at Zero (Misleading)`: This practice can make small variations appear more significant than they are, leading to misinterpretation.
- `Tables with No Context:` Presenting data tables without explanatory context, annotations, or highlights can leave the audience puzzled about what's important. Tables should be accompanied by clear narratives or visual cues that guide the viewer's understanding.

Advanced chart type

In this section, we introduce some advanced chart types listed below.

- **Scatter Plots:** While basic in concept, when used with large datasets or enhanced with interactive features, scatter plots can reveal complex relationships and trends.
- **Box Plots:** For detailed distribution analysis, highlighting median, quartiles, and distribution density.
- **Kernel Density Estimation (KDE) Plots:** For visualizing the probability density function of a continuous variable, offering a smooth representation of data distribution to identify patterns and outliers.
- **Violin Plots:** For visualizing the distribution of data across different categories, combining the features of a box plot with a kernel density plot to show the distribution's shape, central tendency, and variability.
- **Heatmaps:** Effective for visualizing complex matrices of data, such as correlation matrices or geographical data density.
- **Tree Maps and Sunburst Charts:** Useful for visualizing hierarchical data or parts of a whole in nested structures.
- **Sankey Diagrams:** To visualize flow and transfer across systems or processes.

Advanced chart type: Scatter Plots

Scatter Plots: Scatter plots are fundamental for exploring relationships between two continuous variables, offering insights into correlation, trends, and outlier presence.

Key Features to Focus On

- `Correlation Strength and Direction:` Assess whether the relationship between variables is positive, negative, or nonexistent based on the data point arrangement.
- `Trend Patterns:` Look for patterns that may indicate a linear, exponential, or other type of relationship. Utilize trend lines when available to aid in this assessment.
- `Outliers and Anomalies:` Identify data points that fall far from the general pattern, which could indicate anomalies or outliers requiring further investigation.
- `Data Density:` Note areas of high data point concentration, which could indicate common value combinations or preferred states.

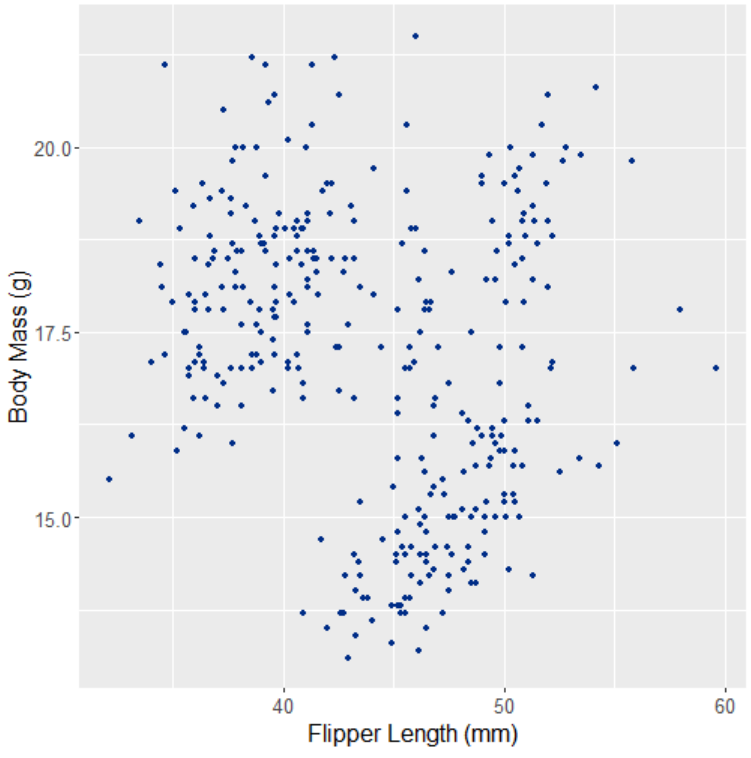
Advanced chart type: Box Plots

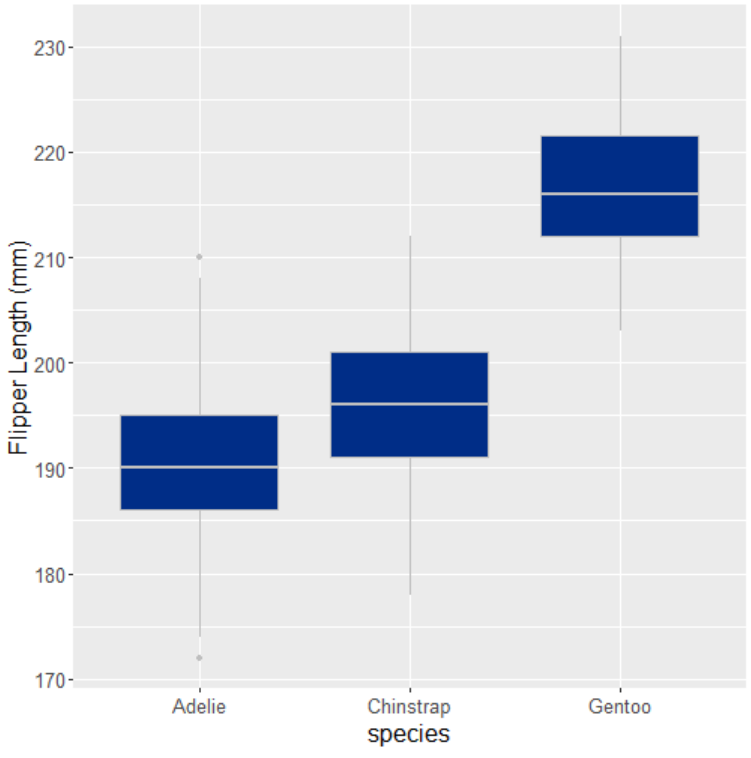
Box Plots: Box plots (or box-and-whisker plots) succinctly summarize the distribution of a continuous variable, highlighting central tendencies, dispersion, and potential outliers.

Key Features to Focus On

- `Central Tendency and Spread:` The box represents the interquartile range (IQR), with the line inside the box denoting the median. Assess the data's center and variability at a glance.
- `Outliers:` Points outside the "whiskers" (lines extending from the box) are considered outliers. Evaluate their potential impact on the analysis.
- `Symmetry and Skewness:` The box's symmetry around the median line indicates skewness. A box skewed toward the whiskers suggests a skewed distribution.
- `Comparison Across Groups:` When multiple box plots are aligned side by side, compare their medians, IQRs, and overall ranges to understand distribution differences across categories.

Examples of Scatter and Box Plots





Advanced chart types: KDE Plots

Kernel Density Estimation (KDE) Plots: KDE plots are a way to estimate the probability density function of a continuous variable. KDE plots are useful for visualizing the underlying distribution of data points, highlighting the density where data points occur more frequently.

Key Features to Focus On

- `Density Peaks:` Peaks in a KDE plot represent areas where data points are concentrated, indicating the mode(s) of the distribution.
- `Spread:` The width of the KDE plot gives an idea of the spread of the distribution. A wider plot suggests a larger variance in the data.
- `Tails:` Pay attention to the tails of the KDE plot, as they can indicate the presence of outliers or long-tailed distributions.
- `Comparison of Distributions:` When overlaying KDE plots for different groups, focus on differences in peaks and spreads to compare distributions directly.
- `Example Use Case:` Analyzing the distribution of product prices on an e-commerce platform or comparing the age distribution of different population groups.

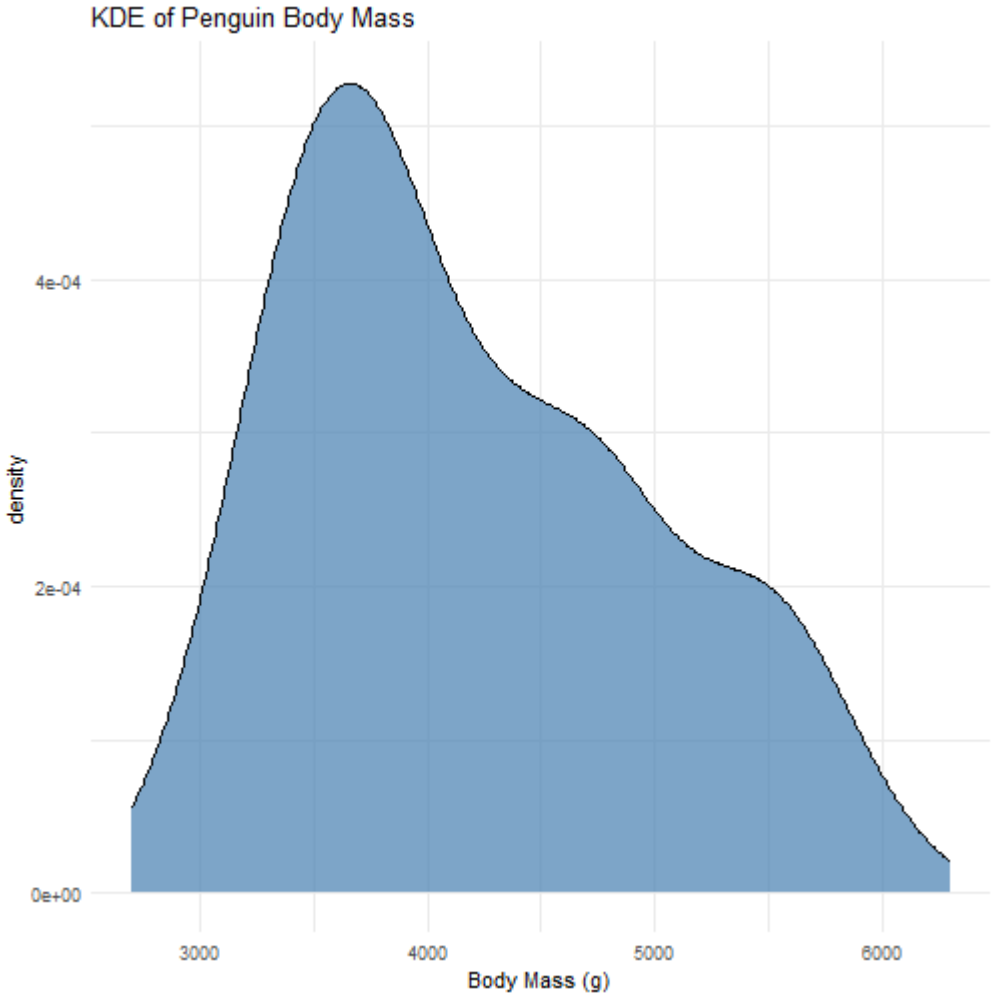
Advanced chart types: Violin Plots

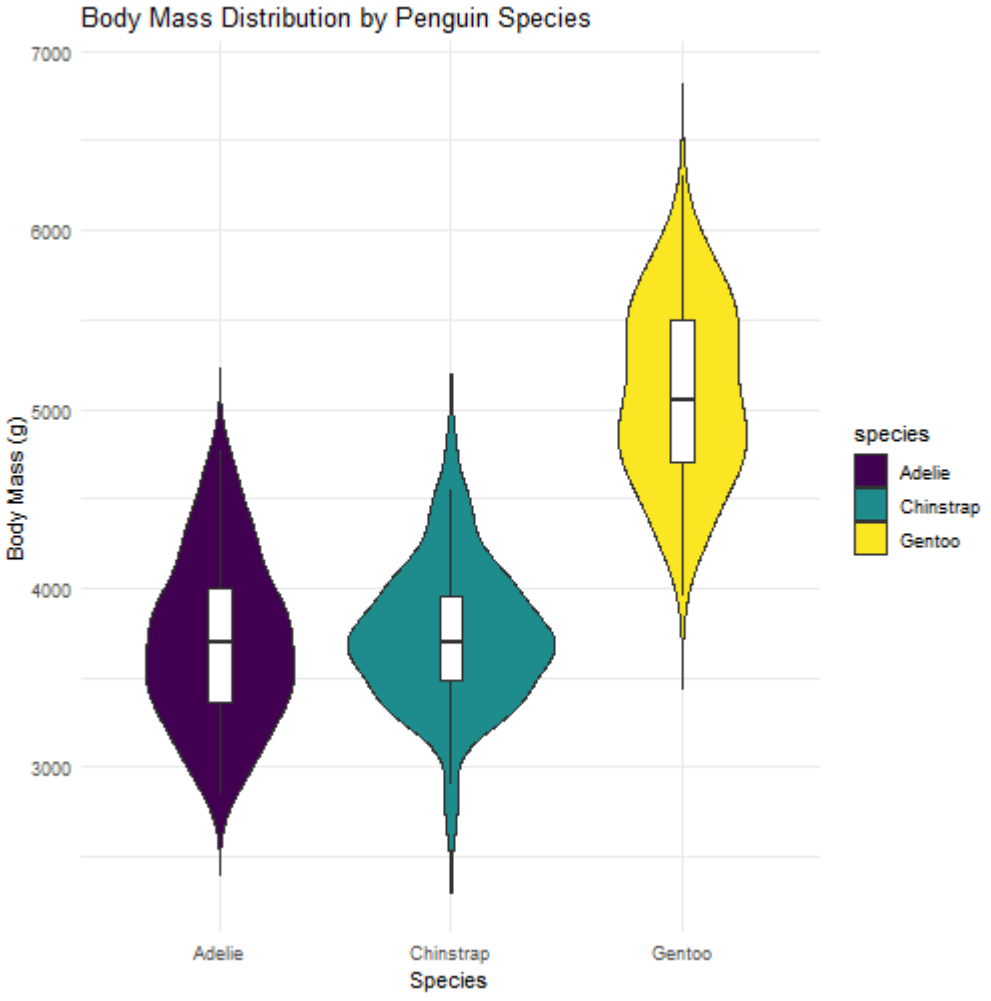
Violin Plots: A violin plot is a method of plotting numeric data and can be seen as a combination of a box plot and a kernel density plot (KDE). It provides a detailed representation of the data distribution, showing peaks, valleys, and tails within the data.

Key Features to Focus On

- `Overall Shape:` The wider sections of the violin plot represent a higher probability that members of the population will take on the given value; the skinnier sections represent a lower probability.
- `Median and Quartiles:` Inside the violin, look for markers or lines that indicate the median and interquartile range, similar to a box plot. These provide a summary of the central tendency and spread.
- `Symmetry:` The symmetry of the plot can indicate skewness in the data. Asymmetrical violins suggest a skewed distribution.
- `Multiple Peaks:` Peaks and valleys within the violin shape can indicate multimodality in the data, suggesting the presence of multiple groups or behaviors within the dataset.
- `Example Use Case:` Comparing the distribution of customer satisfaction scores across multiple stores or comparing the distribution of exam scores across different subjects.

Examples of KDE and Violin Plots





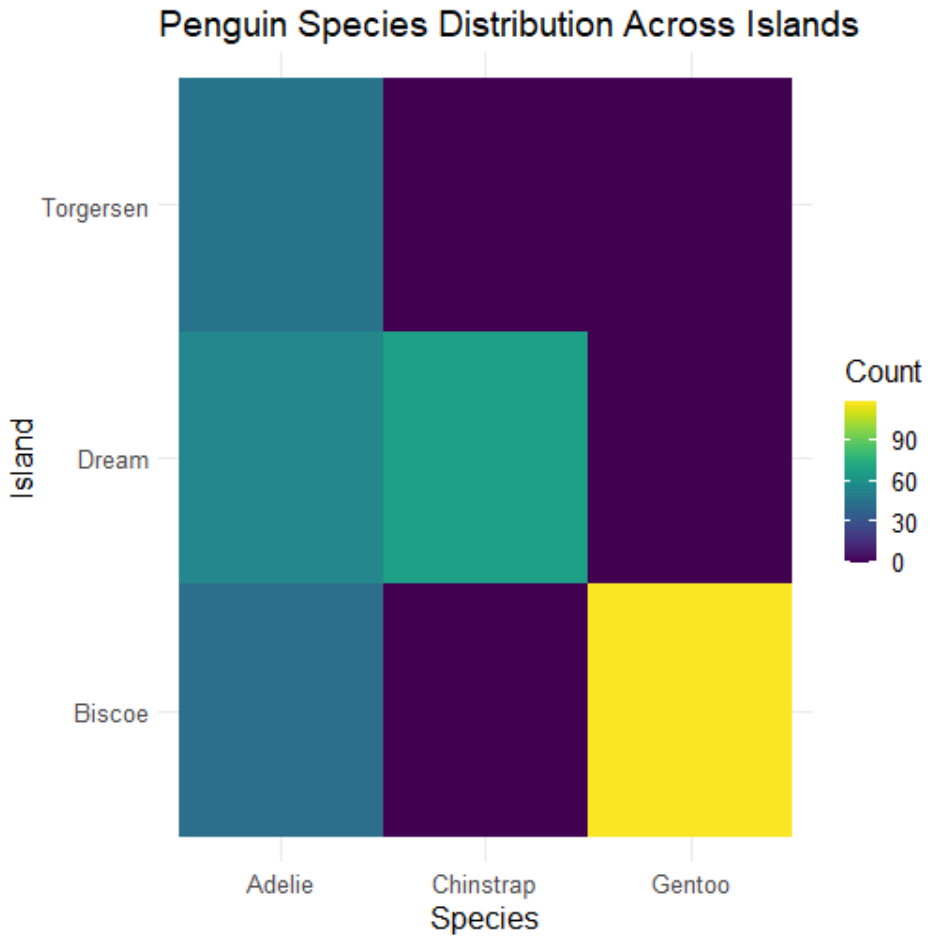
Advanced chart type: Heatmaps

Heatmaps: Heatmaps effectively represent complex data matrices, making them ideal for spotting patterns, correlations, and trends across two dimensions.

Key Features to Focus On

- `Color Intensity:` The color scale in a heatmap indicates magnitude, with darker or more intense colors typically representing higher values. This visual cue helps identify areas of concentration or significance within the data matrix.
- `Patterns and Clusters:` Look for patterns or clusters in the heatmap, which can reveal correlations, trends, or groupings in the data. These patterns can guide further analysis or hypothesis generation.
- `Anomalies:` Deviations from the general color pattern might indicate anomalies or outliers. These areas warrant closer inspection for insights or data issues.
- `Comparisons:` In heatmaps comparing different categories or time periods, focus on shifts in color patterns to identify changes over time or differences between groups.

An Example of Heatmaps



Advanced chart types: Tree Maps and Sunburst Charts

Tree Maps and Sunburst Charts: Tree Maps and Sunburst Charts are excellent for displaying hierarchical data and proportions in a nested format, where each level of the hierarchy is represented as a rectangle (Tree Map) or ring segment (Sunburst Chart).

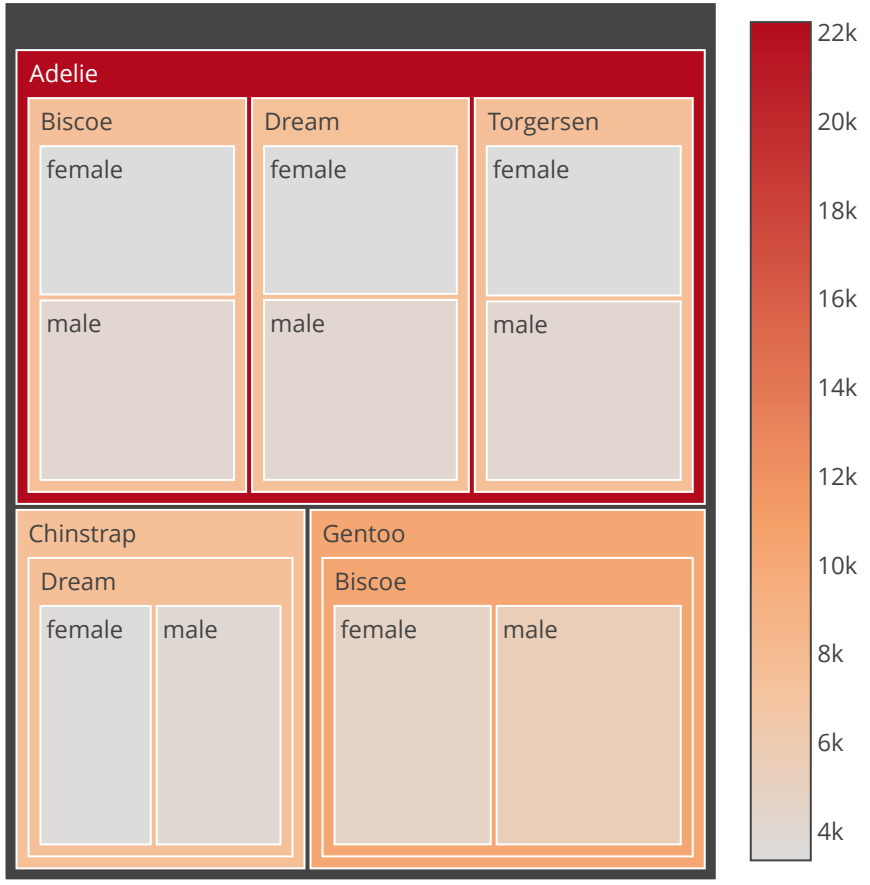
Key Features to Focus On

- `Hierarchy Structure:` Observe how the data is organized hierarchically, with larger blocks or segments representing higher values or larger proportions.
- `Proportions:` Focus on the size of each block or segment to understand the proportion of each category relative to the whole.
- `Comparison Within Hierarchy:` Compare sizes of blocks or segments within the same level to assess relative importance or value.
- `Color Coding:` Utilize color to distinguish between different levels of the hierarchy or different categories within the same level, enhancing the readability and interpretability of the data.

An Example of Tree Maps

A tree map could be used to visualize the distribution of penguins across different species and islands, with the size of each section representing a numerical value such as the average body mass, count of penguins, or another aggregate metric. For example:

- The top level of the tree map could divide penguins by species.
- The next level could further divide each species by island.
- The size of each section could represent the number of penguins in each category or their average body mass. In our example, we show the average body mass.

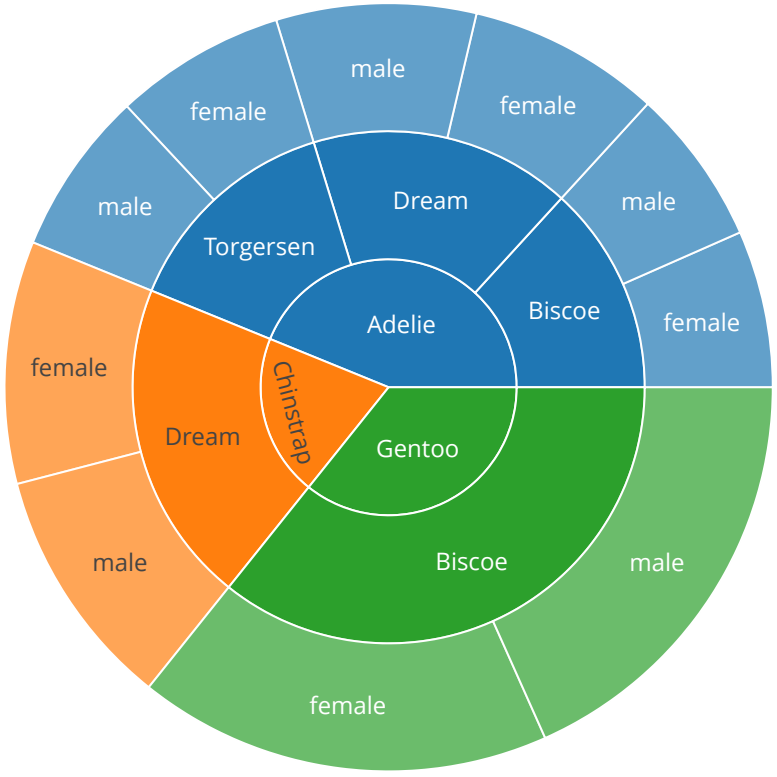


An Example of Sunburst Charts

A sunburst chart is similar to a tree map but is radial and can effectively show hierarchical data and proportions.

For the penguins dataset:

- The innermost circle could represent the different penguin species.
- The next layer could represent the islands for each species.
- Further layers could represent additional categorizations, such as the sex of the penguins, if desired. The size or angle of each segment could represent the count of penguins, average flipper length, or body mass.



Advanced chart types: Sankey Diagrams

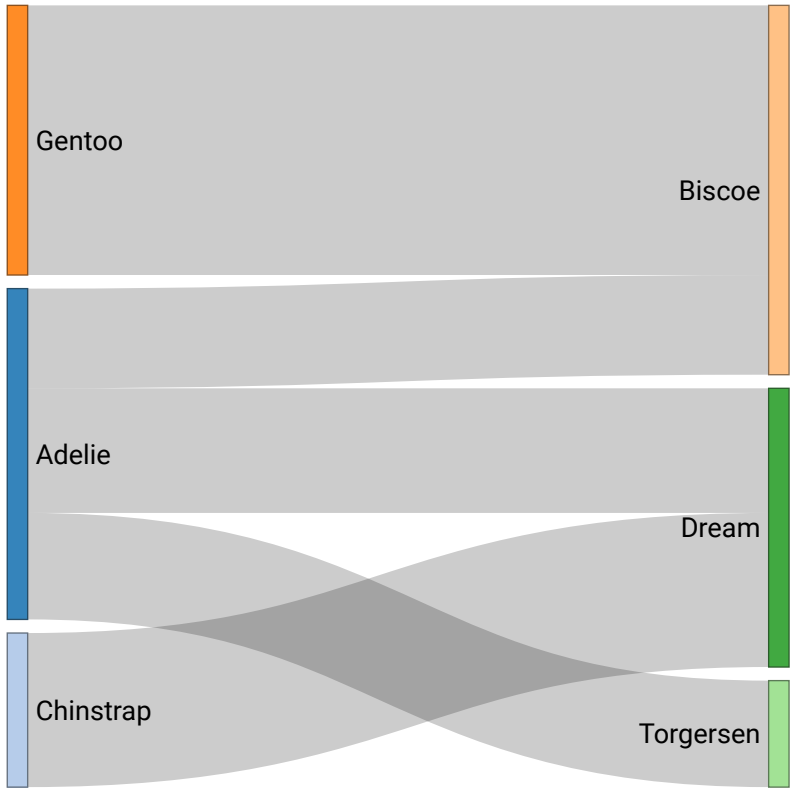
Sankey Diagrams: Sankey Diagrams are specialized charts used to visualize the flow of resources, energy, or information between different stages in a system or process. They highlight the volume of flow and how it splits and converges through different paths.

Key Features to Focus On

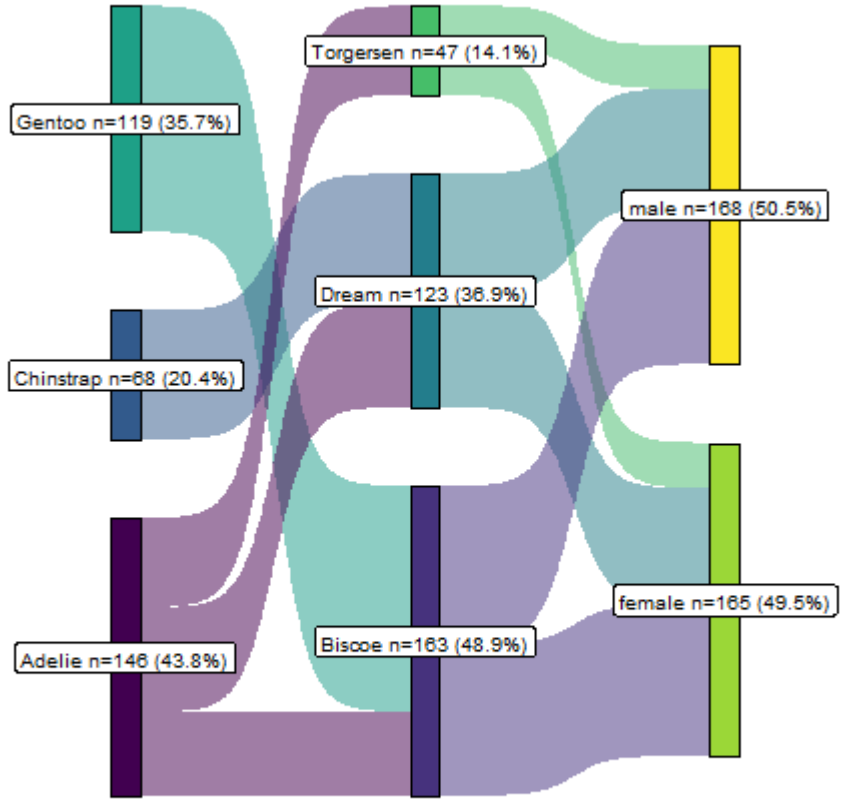
- `Flow Volume:` The thickness of the lines or paths in a Sankey Diagram represents the volume of flow, allowing you to assess the magnitude of flow between different nodes.
- `Flow Direction:` Pay attention to the direction of the flow, which shows how resources, energy, or information moves from one stage or category to another.
- `Bottlenecks and Major Paths:` Identify areas where the flow converges or diverges significantly, indicating bottlenecks or major pathways within the system.
- `Balance of Input and Output:` In systems where balance is expected (e.g., energy systems), compare the total input and output to identify potential losses or inefficiencies.

An Example of Sankey Diagrams

- The Distribution of Penguins by Species to Islands: Showing how different species of penguins are distributed across the islands.



- Sex Distribution within Species and Islands: Demonstrating the flow from species to sex, to show the gender distribution within each species and location.



Multi-dimensional data visualization

Visualizing multi-dimensional data—that is, data with more than three variables—presents unique challenges due to the limitations of two-dimensional screens and the human ability to process complex information. Here's an overview of these challenges and strategies to address them, including the use of dimensionality reduction techniques.

Challenges

- `Complexity:` As the number of dimensions increases, so does the complexity of the visualization. This can make it difficult for the audience to understand and extract useful insights.
- `Overplotting:` Displaying a large number of variables simultaneously can lead to overplotting, where too many data points overlap, making it hard to distinguish between them.
- `Loss of Information:` Trying to represent multi-dimensional data in two dimensions can result in loss of information or misleading representations.
- `Cognitive Load:` Multi-dimensional visualizations can be overwhelming and increase the cognitive load on the viewer, hindering the ability to make quick and accurate interpretations.

Multi-dimensional data visualization (Continued)

Strategies

- `Faceting:` One approach to simplifying multi-dimensional data visualization is to use faceting, where separate plots are created for subsets of the data, allowing for comparison across different dimensions without overloading a single chart.
- `Parallel Coordinates:` This technique involves plotting each dimension on its own vertical axis, all parallel to each other. Lines connect each data point across the axes, making it possible to observe patterns and relationships across multiple dimensions.
- `Radar Charts:` Useful for comparing multiple variables of a dataset to reveal strengths and weaknesses across different categories. Each axis represents one variable, and the data points are plotted as points along each axis, connected to form a polygon.
- `Interactive Visualizations:` Leveraging interactivity to explore multi-dimensional data, using tools like Tableau, Plotly, or D3.js, where users can filter, zoom, or adjust what data is being displayed to uncover deeper insights.

Parallel Coordinates Plots

Parallel Coordinates Plots: Parallel Coordinates Plots are used for visualizing and analyzing multivariate data. Each variable is represented by a vertical line (or axis), and data points are plotted as lines crossing each axis at the point corresponding to the data value.

Key Features to Focus On:

- `Variable Relationships:` Observe how lines intersect across different axes. Parallel lines indicate a positive correlation between variables, while intersecting lines suggest a negative correlation.
- `Cluster Identification:` Look for groups of lines that follow a similar path across axes, indicating clusters within the data. These clusters can reveal patterns or relationships across multiple variables.
- `Outliers and Anomalies:` Identify lines that diverge significantly from the rest, as these may represent outliers with unique characteristics compared to the bulk of the data.
- `Data Density and Overlap:` Areas where lines densely overlap can indicate common value combinations across variables, while sparse areas suggest less common combinations.

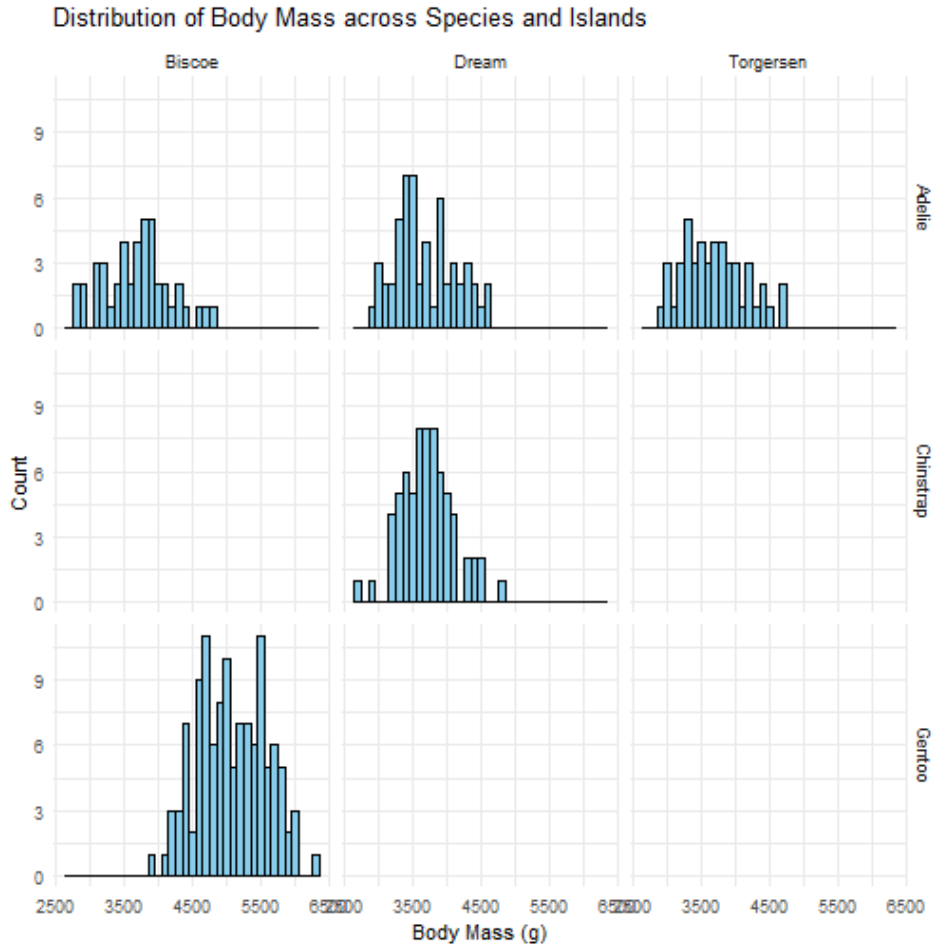
Radar Charts

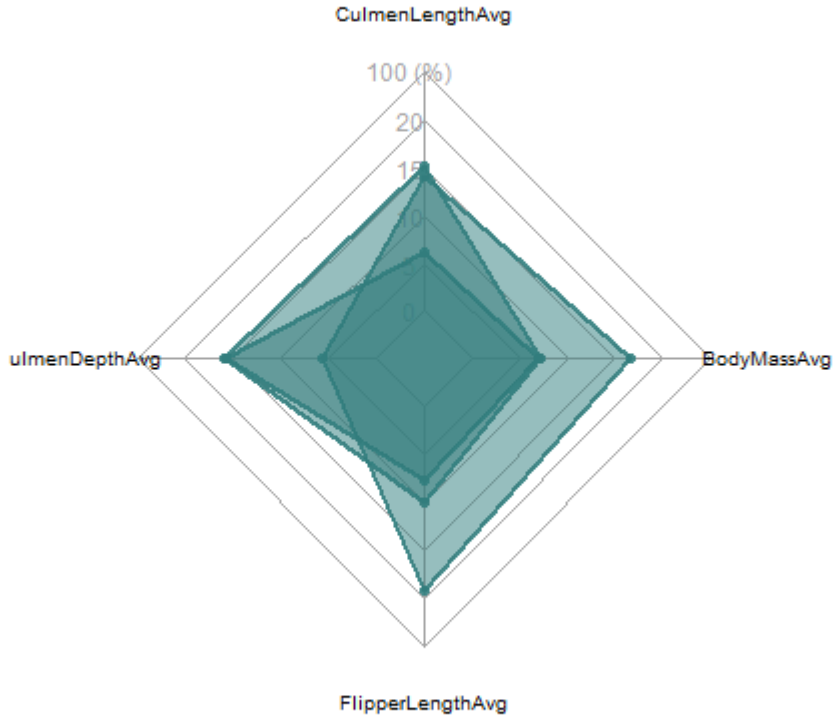
Radar Charts: Radar Charts, also known as Spider or Web charts, are two-dimensional charts used to plot one or more groups of values over multiple variables. They are particularly useful for displaying multivariate observations with an equal number of variables.

Key Features to Focus On:

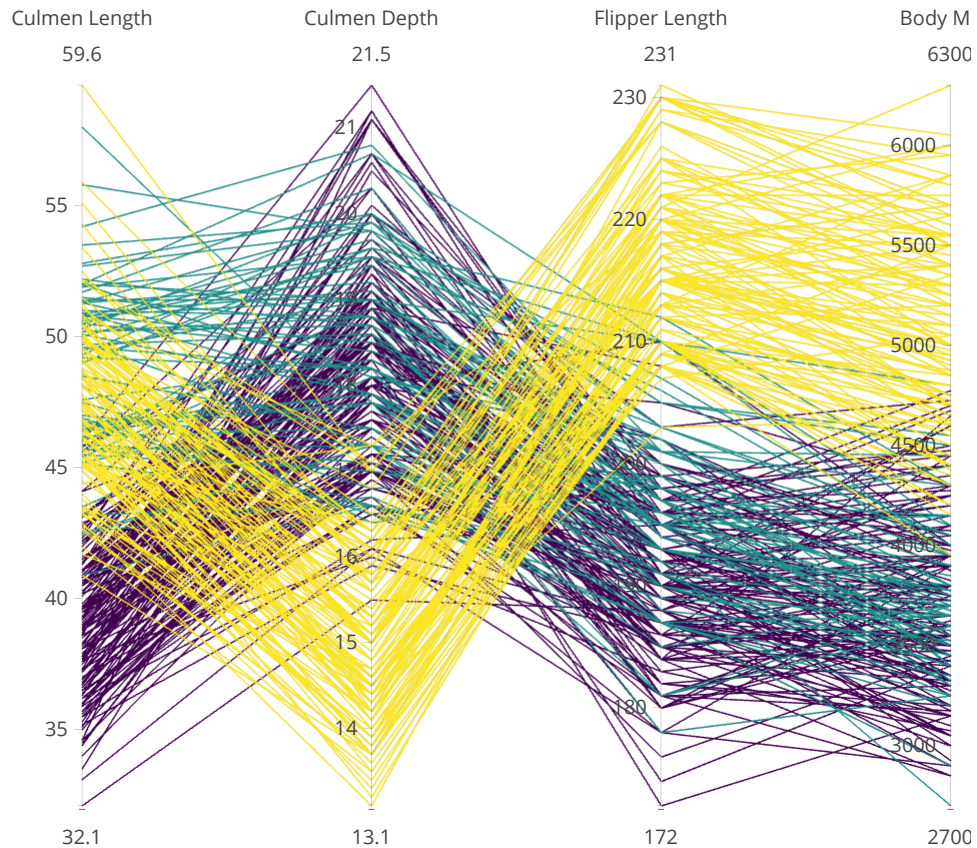
- `Relative Performance:` Assess how each group or category performs across different variables. This is useful for comparing the strengths and weaknesses of each group.
- `Variable Contribution:` Notice the contribution of each variable to the overall profile. Variables where the data points extend further from the center indicate higher values or performance in that area.
- `Shape Patterns:` The shape formed by connecting data points for each group can indicate similarities or differences between groups. Similar shapes suggest similar performance across variables, while distinct shapes highlight differences.
- `Balance and Symmetry:` In assessments where balance across variables is desired (e.g., skill assessments), the symmetry of the radar chart can indicate well-roundedness or a balanced profile.

Examples of Faceting & Radar Charts





An Example of Parallel Coordinates Plots



Multi-dimensional data visualization (Continued)

Dimensionality Reduction Techniques

When the visualization becomes too complex or the data too high-dimensional, dimensionality reduction techniques can be employed to simplify the data without losing its essential features. These techniques are particularly useful in exploratory data analysis to uncover patterns or in preparation for machine learning modeling.

- `Principal Component Analysis (PCA):` PCA reduces the dimensionality of the data by transforming it into a new set of variables, the principal components, which are uncorrelated and which capture the most variance in the data.
- `t-Distributed Stochastic Neighbor Embedding (t-SNE):` t-SNE is a non-linear technique particularly well-suited for the visualization of high-dimensional datasets. It reduces dimensions while maintaining the local structure of the data, making it useful for identifying clusters.
- `Uniform Manifold Approximation and Projection (UMAP):` Similar to t-SNE, UMAP is a dimension reduction technique that is particularly effective at preserving both the local and global structure of the data, making it useful for a wide range of visualization and machine learning tasks.

Addressing the challenges of multi-dimensional visualization requires a combination of thoughtful data preparation, choosing the right visualization techniques, and potentially applying dimensionality reduction methods. By carefully considering the goal of the visualization and the audience's needs, it's possible to convey complex, multi-dimensional insights effectively.

Interpreting visualizations

Insight Extraction: Asking the Right Questions

- Start with broad questions about trends, patterns, and anomalies observed in the visualization. For example, "What trend does this line chart show over time?"
- Move to specific inquiries related to the data's implications, such as, "Why is there a spike in user engagement during this period?"
- Consider the context and external factors that could influence the data, asking, "Could external events have influenced these results?"

Interpreting visualizations (Continued)

Narrative Building: Crafting a Story from Data

- `Identify the Core Message:` Every data visualization should convey a clear, central message or insight.
- `Connect the Dots:` Show how individual data points relate to each other and contribute to the overall story. This could involve highlighting cause-and-effect relationships or trends. However, the cause-and-effect relationships must be justified. Correlation is NOT causation.
- `Engage the Audience:` Tailor the narrative to their audience's interests and level of understanding, making the data relatable and impactful.

Interpreting visualizations (Continued)

Critical Evaluation: Fostering Skepticism and Awareness

- `Question Data Sources:` It is important to consider where and how data was collected, and potential biases in data sources.
- `Identify Misleading Visuals:` Make sure to spot visualizations that may distort the data, such as improper scaling, cherry-picked time frames, or inappropriate chart types.
- `Consider the Creator's Intent:` The creator's perspective and goals might influence a visualization's design and the data presented.

Useful Resources

- The R Graph Gallery
- The Python Graph Gallery

In the realm of data visualization, the core value resides not in the sophistication or brand of the tools we employ but in the underlying ideas and insights we seek to communicate. Whether we choose to harness the capabilities of Tableau, Excel, JMP, or delve into programming languages like R or Python, these are merely instruments at our disposal. What truly matters is the conceptual vision behind our visualizations—the story we aim to tell with our data. Each tool, with its unique features and capabilities, serves as a conduit for translating complex data into comprehensible and impactful visual narratives. Thus, the art of data visualization is fundamentally about leveraging the tool that best aligns with our comfort level and the specific demands of our data story, ensuring that the essence of our insights is conveyed with clarity and precision

References

The lectures of this course are based on the ideas from the following references.

- Exploratory Data Analysis by John W. Tukey
- A Course in Exploratory Data Analysis by Jim Albert
- The Visual Display of Quantitative Information by Edward R. Tufte
- Data Science for Business: what you need to know about data mining and data-analytic thinking by Foster Provost and Tom Fawcett
- Storytelling with Data: A Data Visualization Guide for Business Professionals by Cole Nussbaumer Knaflic