Module 4- Data Visualization and Data Exploration

Module 4 Syllabus	**Introduction:** Data Visualization, Importance of Data Visualization, Data Wrangling, Tools and Libraries for Visualization **Comparison Plots:** Line Chart, Bar Chart and Radar Chart**; Relation Plots:** Scatter Plot, Bubble Plot, Correlogram and Heatmap; **Composition Plots:** Pie Chart, Stacked Bar Chart, Stacked Area Chart, Venn Diagram; **Distribution Plots:** Histogram, Density Plot, Box Plot, Violin Plot**; Geo Plots:** Dot Map, Choropleth Map, Connection Map**; What Makes a Good Visualization?** **Textbook 2: Chapter 1, Chapter 2**

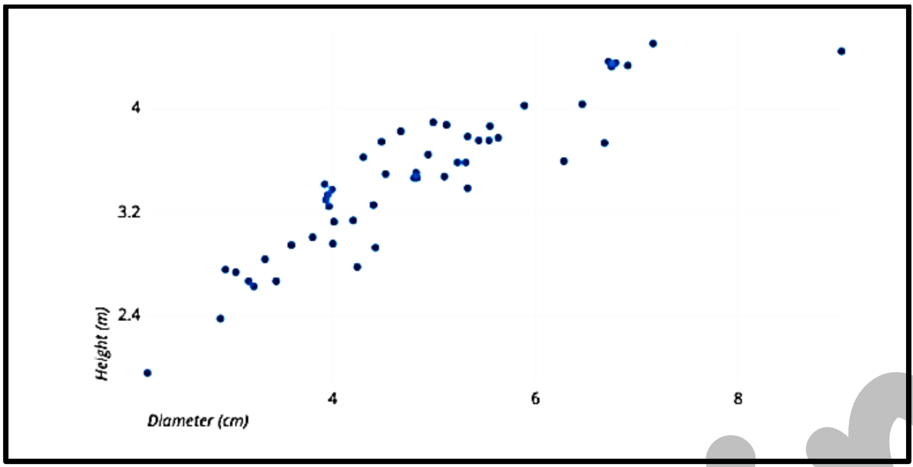
Handouts for Session 1: Introduction: Data Visualization, Importance of Data Visualization

4.1 Introduction to Data Visualization

- Computers and smartphones store names and numbers digitally.
- Data representation involves the forms used to store, process, and transmit data.
- Effective representations convey stories and fundamental discoveries, enhancing the data's value.
- By modeling information properly, we gain clearer, more concise, and understandable insights.
- Representations convert data into useful information, helping to derive meaningful insights.

4.2 The Importance of Data Visualization

- Instead of merely viewing data in Excel columns, visualization helps us better understand our data.
- For instance, visualizing data through charts and graphs can reveal patterns, trends, and insights that are not immediately obvious in raw data form.
- It's easy to see a pattern emerge from the numerical data that's given in the following scatter plot.
- It shows the correlation between diameter and the height of various trees. There is a positive correlation between diameter and height.



Questions

1. Briefly explain Data Visualization.

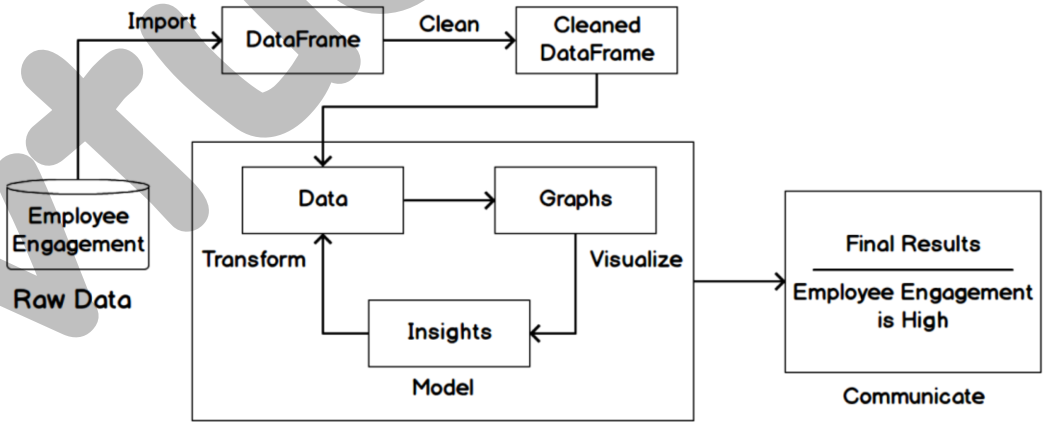
2. Why Data Visualization is Important/Significant?

Handouts for Session 2: Data Wrangling, Tools and Libraries for Visualization

4.3 Data Wrangling

Data wrangling is the process of transforming raw data into a suitable representation for various tasks. It is the discipline of augmenting, cleaning, filtering, standardizing, and enriching data in a way that allows it to be used in a downstream task, which in our case is data visualization.

Examine the following flow diagram of the data wrangling process to understand how precise and actionable data is prepared for business analysts to utilize.



Data wrangling process to measure employee engagement

The following steps explain the flow of the data wrangling process:

1. First, the Employee Engagement data is in its raw form.

2. Then, the data gets imported as a DataFrame and is later cleaned.
3. The cleaned data is then transformed into graphs, from which findings can be derived.
4. Finally, we analyze this data to communicate the final results.

For example, employee engagement can be measured based on raw data gathered from feedback surveys, employee tenure, exit interviews, one-on-one meetings, and so on. This data is cleaned and made into graphs based on parameters such as referrals, faith in leadership, and scope of promotions. The percentages, that is, information derived from the graphs, help us reach our result, which is to determine the measure of employee engagement.

4.4 Tools and Libraries for Visualization

- Several tools are available for creating data visualizations to suit different needs.
- Non-coding tools like Tableau provide an intuitive interface for exploring and understanding data.
- Alongside Python, MATLAB and R are also commonly used in data analytics.
- Python stands out as the industry's preferred language due to its user-friendly nature and efficiency in data manipulation and visualization.
- Its extensive library ecosystem further enhances Python's appeal, making it the optimal choice for robust data visualization tasks.

Questions:

1. What is Data Wrangling?

2. Explain the data wrangling process with an example of employee engagement.

3. With a neat diagram explain the steps involved in the Data Wrangling process.

Handouts for Session 3: Comparison Plots: Line Chart, Bar Chart and Radar Chart

4.5 Comparison Plots

- Comparison plots include charts that are ideal for comparing multiple variables or variables over time.
- **Line charts** are great for visualizing variables over time.

- For comparison among items, **bar charts** (also called column charts) are the best way to go. For a certain time period (say, fewer than 10-time points), vertical bar charts can be used as well.
- **Radar charts** or spider plots are great for visualizing multiple variables for multiple groups.

1. Line Chart

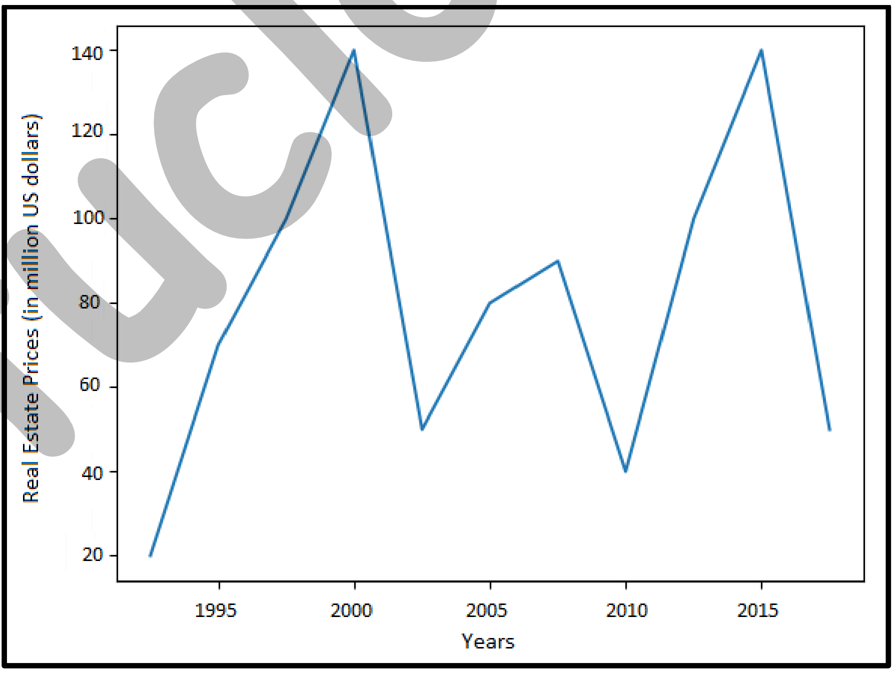
- Line charts are used to display quantitative values over a continuous time period and show information as a series.
- A line chart is ideal for a time series that is connected by straight-line segments.
- The value being measured is placed on the y-axis, while the x-axis is the timescale.

Uses

✓ Line charts are great for comparing multiple variables and visualizing trends for both single as well as multiple variables, especially if your dataset has many time periods (more than 10).

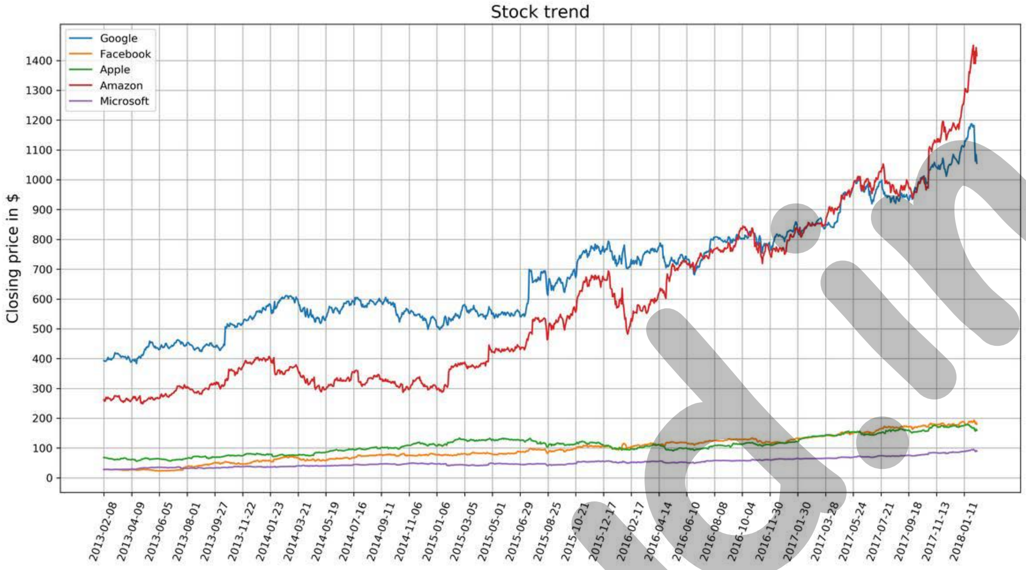
✓ For smaller time periods, vertical bar charts might be the better choice.

Example 1: The following diagram shows a trend of real estate prices (per million US dollars) across two decades. Line charts are ideal for showing data trends:



Line chart for a single variable

Example 2: The following figure is a multiple-variable line chart that compares the stock-closing prices for Google, Facebook, Apple, Amazon, and Microsoft. A line chart is great for comparing values and visualizing the trend of the stock. As we can see, Amazon shows the highest growth:



Line chart showing stock trends for five companies

Design Practices

- ✓ Avoid too many lines per chart.
- ✓ Adjust your scale so that the trend is clearly visible.

2. Bar Charts

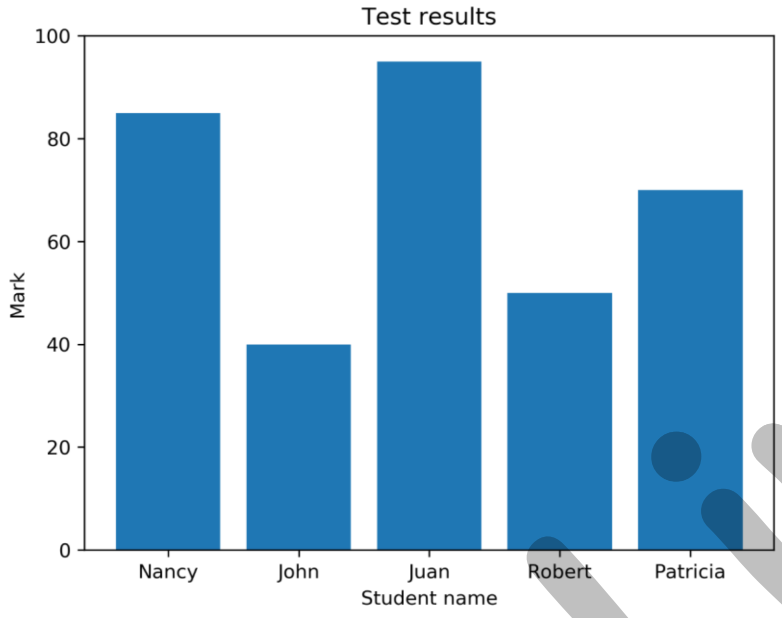
- In a bar chart, the bar length encodes the value. There are two variants of bar charts: **vertical bar charts** and **horizontal bar charts.**

Uses

- While they are both used to compare numerical values across categories, vertical bar charts are sometimes used to show a single variable over time.

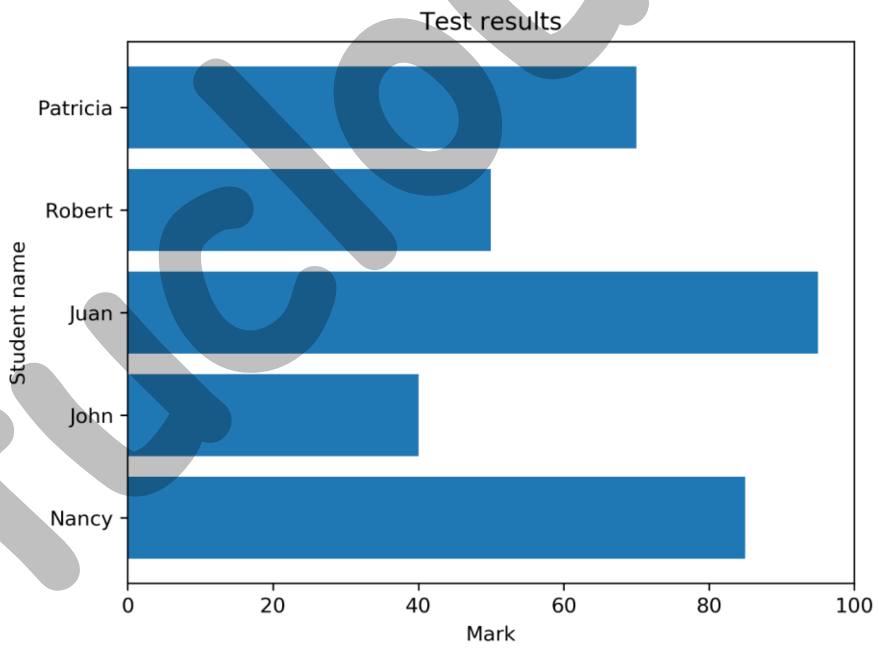
Example 1:

The following diagram shows a vertical bar chart. Each bar shows the marks out of 100 that 5 students obtained in a test:



Vertical bar chart using student test data

The following diagram shows a horizontal bar chart. Each bar shows the marks out of 100 that 5 students obtained in a test:

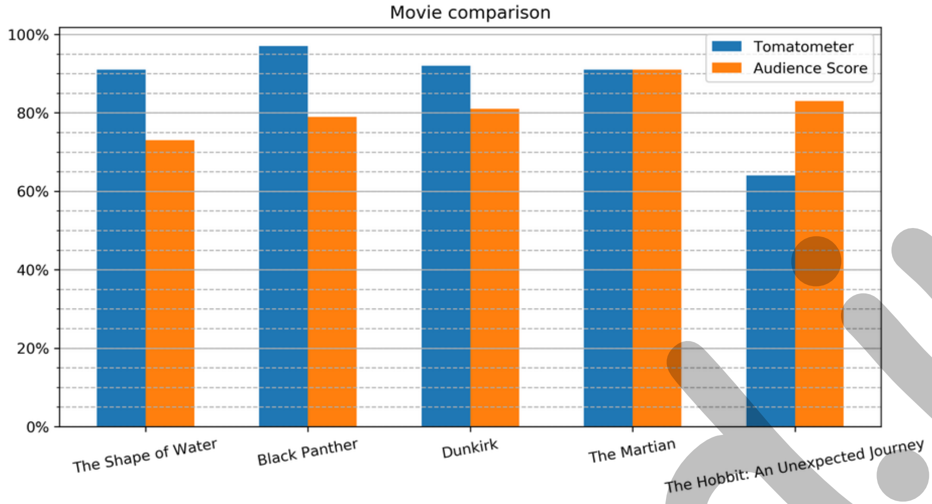


Horizontal bar chart using student test data

Example 2:

The below graph compares movie ratings with two scores: the Tomatometer, representing the percentage of approved critic reviews, and the Audience Score, representing the percentage of users rating 3.5 or higher out of 5. Notably, The

Martian has high scores on both metrics. The Hobbit: An Unexpected Journey has a high Audience Score despite a lower Tomatometer score, likely due to its large fan base.



Comparative bar chart

Design Practices

1. When creating bar charts, ensure the numerical axis starts at zero to avoid misleading representations.
2. Use horizontal labels if the chart isn't too cluttered.
3. If space is limited, rotate the labels at different angles, as seen on the x-axis of the preceding diagram.

3. Radar Charts

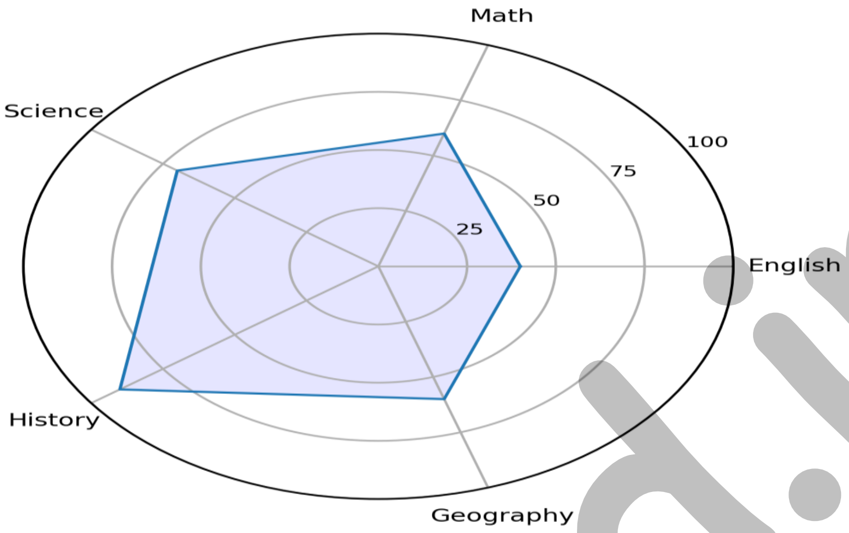
- Radar charts (also known as spider or web charts) visualize multiple variables with each variable plotted on its own axis, resulting in a polygon.
- All axes are arranged radially, starting at the center with equal distances between one another, and have the same scale.

Uses

- ✓ Radar charts are great for comparing multiple quantitative variables for a single group or multiple groups.
- ✓ They are also useful for showing which variables score high or low within a dataset, making them ideal for visualizing performance.

Example 1:

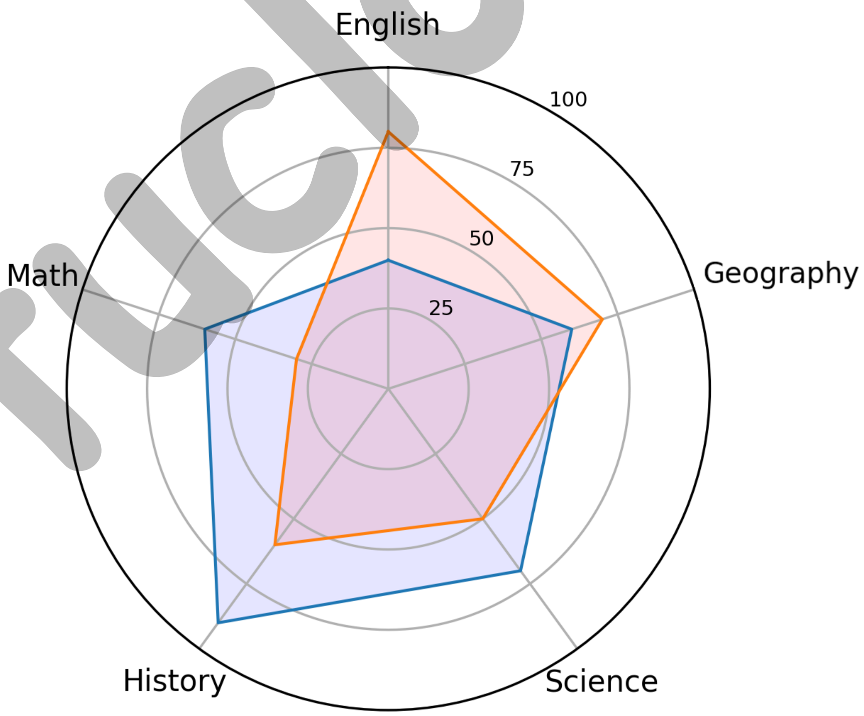
The following diagram shows a radar chart for a single variable. This chart displays data about a student scoring marks in different subjects:



Radar chart for one variable (student)

Example 2:

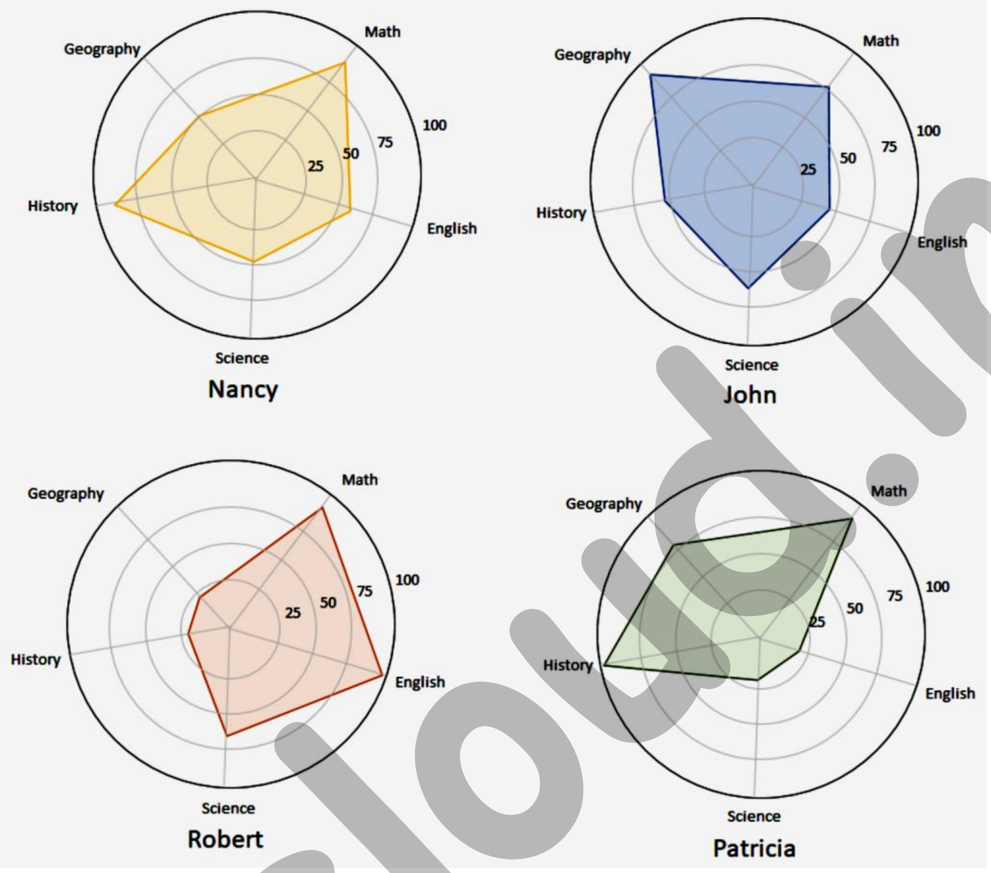
The following diagram shows a radar chart for two variables/groups. Here, the chart explains the marks that were scored by two students in different subjects:



Radar chart for two variables (two students)

Example 3:

The following diagram shows a radar chart for multiple variables/groups. Each chart displays data about a student's performance in different subjects:



Radar chart with faceting for multiple variables (multiple students)

Design Practices

1. Try to display 10 factors or fewer on a single radar chart to make it easier to read.
2. Use faceting (displaying each variable in a separate plot) for multiple variables/ groups, as shown in the preceding diagram, in order to maintain clarity.

Questions:

1. Discuss various comparison plots.

2. Explain what Line, Bar and Radar charts are. Also explain their uses and design practices with examples.

3. Explain the variants of Bar charts with examples.

Handouts for Session 4: Relation Plots: Scatter Plot, Bubble Plot , Correlogram and Heatmap

4.6 Relation Plots

- **Relation plots** are used to show **relationships among variables**.
- A **scatter plot** visualizes the correlation between two variables for one or multiple groups.
- **Bubble plots** can be used to show relationships between three variables. The additional third variable is represented by the dot size.
- **Heatmaps** are great for revealing patterns or correlations between two qualitative variables.
- A **correlogram** is a perfect visualization for showing the correlation among multiple variables.

1. Scatter Plot

- **Scatter plots** show data points for two numerical variables, displaying a variable on both axes.

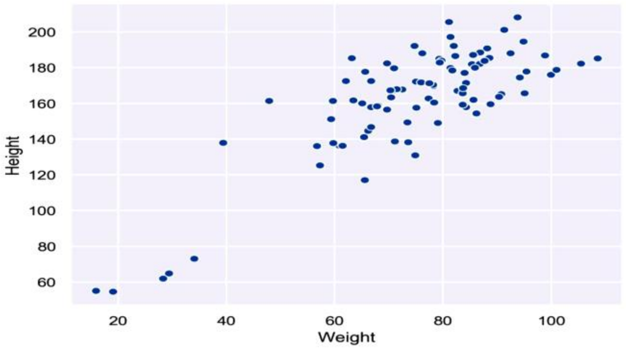
Uses

- ✓ Used to determine if a correlation (relationship) exists between two variables.
- ✓ Used to plot the relationship between multiple groups or categories using different colors.

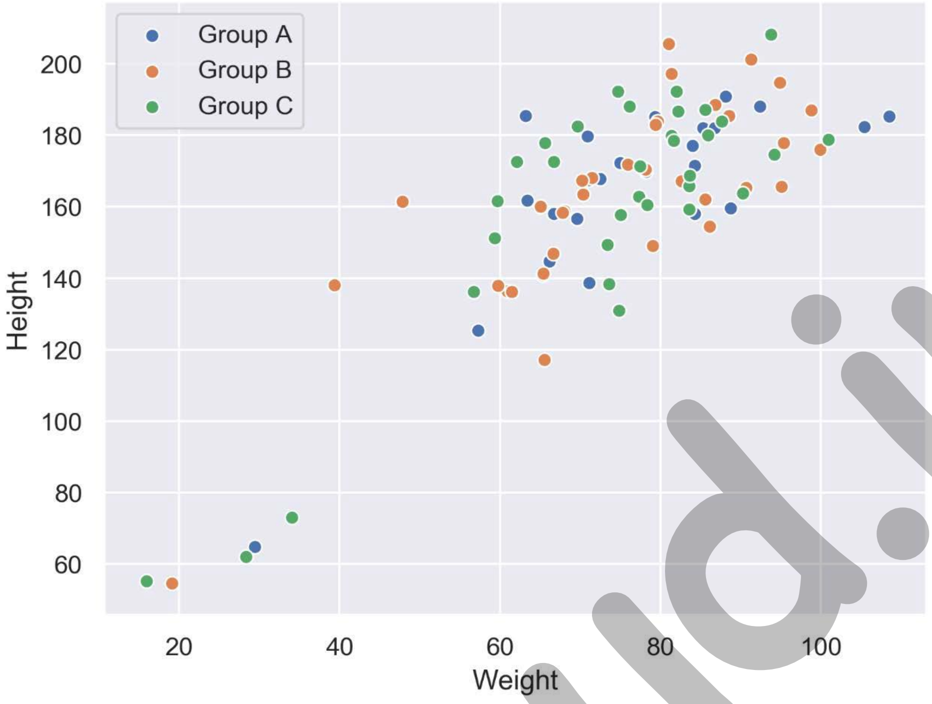
A bubble plot, which is a variation of the scatter plot, is an excellent tool for visualizing the correlation of a third variable.

Examples

The following diagram shows a scatter plot of height and weight of persons belonging to a single group: **Scatter plot with a single group**



The following diagram shows the same data as in the previous plot but differentiates between groups. In this case, we have different groups: A, B, and C:



Scatter plot with multiple groups

Design Practices

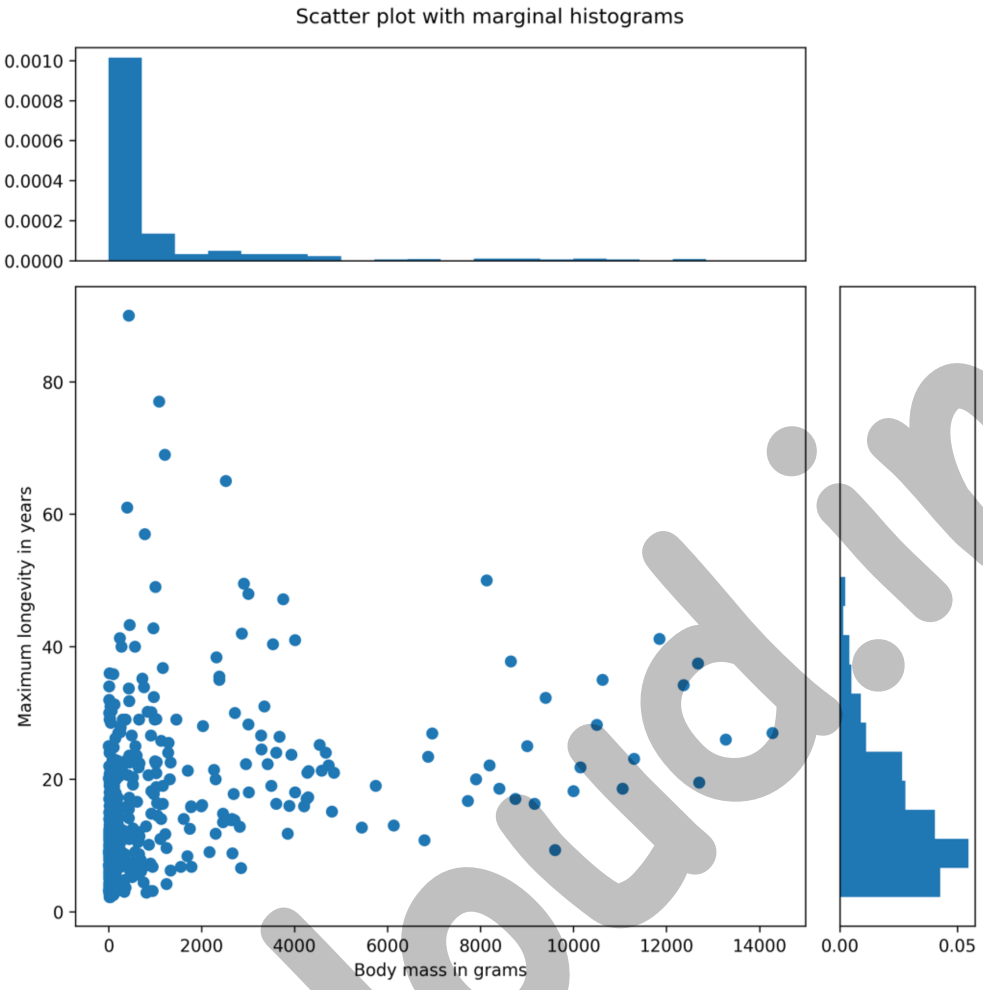
1. Start both axes at zero to represent data accurately.
2. Use contrasting colors for data points and avoid using symbols for scatter plots with multiple groups or categories.

Variants: Scatter Plots with Marginal Histograms

In addition to the scatter plot, which visualizes the correlation between two numerical variables, you can plot the marginal distribution for each variable in the form of histograms to give better insight into how each variable is distributed.

Example

The following diagram shows the correlation between body mass and the maximum longevity for animals in the **Aves** class. The marginal histograms are also shown, which helps to get a better insight into both variables:



Correlation between body mass and maximum longevity of the Aves class with marginal histograms

2. Bubble Plot

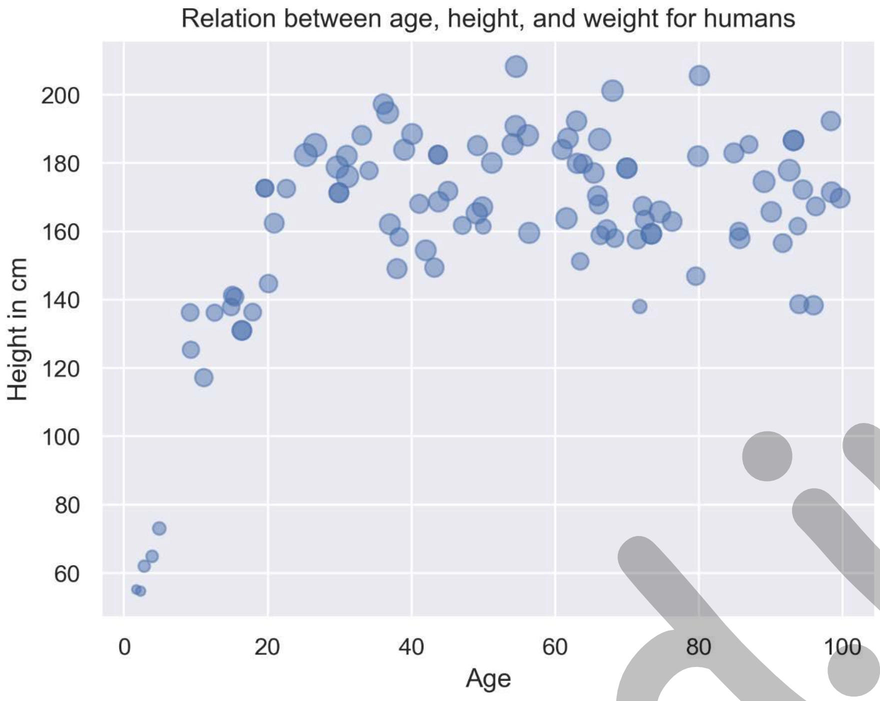
- A bubble plot **extends a scatter plot by introducing a third numerical variable**.
- The **value of the variable** is represented by **the size of the dots**.
- The area of the dots is proportional to the value.
- **A legend is used to link the size of the dot to an actual numerical value**.

Uses

- Bubble plots help to show a correlation between three variables.

Example

The following diagram shows a bubble plot that highlights the relationship between heights and age of humans to get the weight of each person, which is represented by the size of the bubble:



Bubble plot showing the relation between height and age of humans

Design Practices

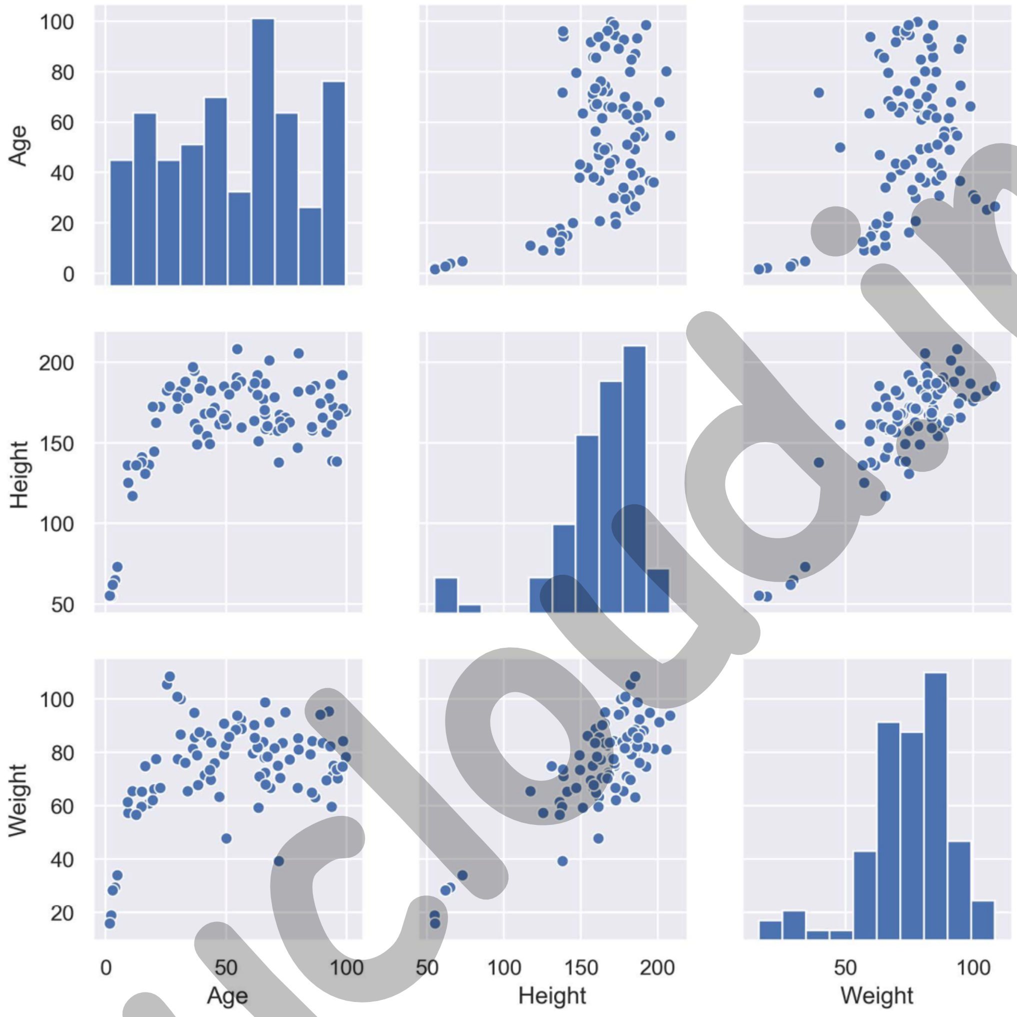
1. The design practices for the scatter plot are also applicable to the bubble plot.
2. Don't use bubble plots for very large amounts of data, since too many bubbles make the chart difficult to read.

3. Correlogram

- A **correlogram** is a combination of scatter plots and histograms.
- A **correlogram** or **correlation matrix** visualizes the **relationship between each pair of numerical variables using a scatter plot.**
- The **diagonals** of the correlation matrix represent the **distribution of each variable in the form of a histogram.**
- Different colors can also be used to plot the relationship between multiple groups or categories.
- A correlogram is **a great chart for exploratory data analysis** to get a feel for your data, especially the correlation between variable pairs.

Examples

The following diagram shows a correlogram for the height, weight, and age of humans. The diagonal plots show a histogram for each variable. The off-diagonal elements show scatter plots between variable pairs:

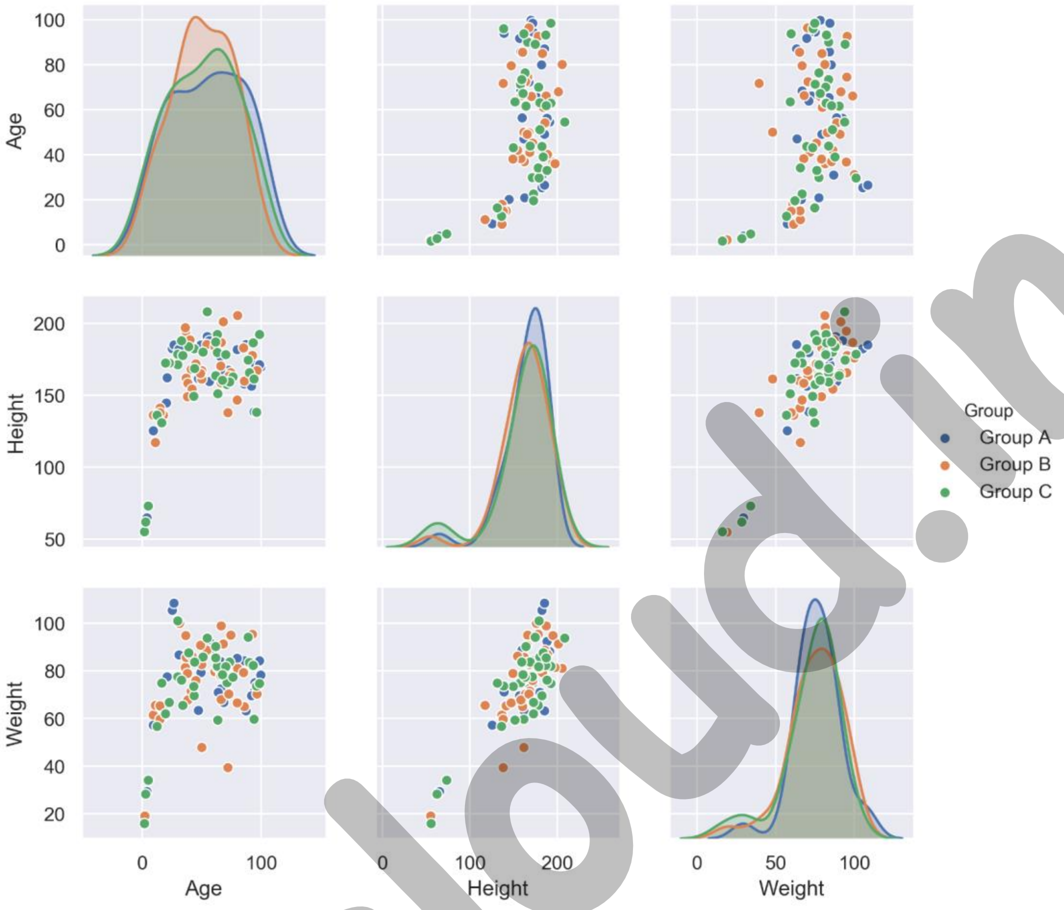


Correlogram with a single category

Design Practices

1. Start both axes at zero to represent data accurately.
2. Use contrasting colors for data points and avoid using symbols for scatter plots with multiple groups or categories.

The following diagram shows the correlogram with data samples separated by color into different groups:



Correlogram with multiple categories

4. Heatmap

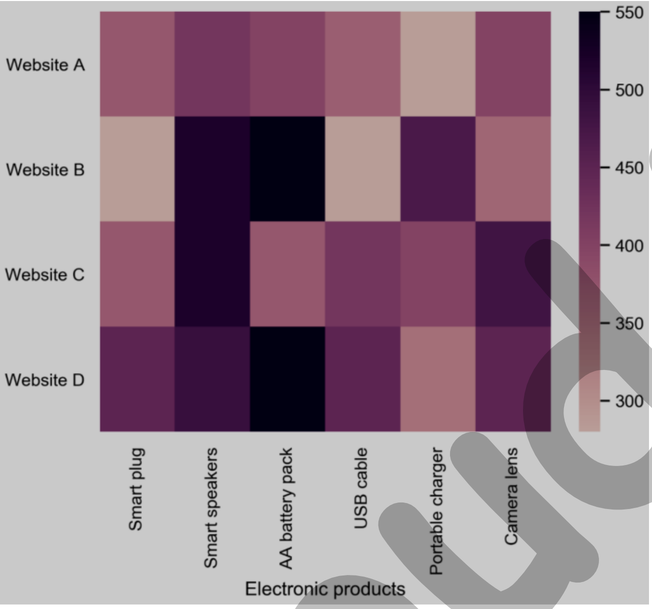
- A **heatmap** is a visualization where **values** contained in a matrix **are represented as colors** or **color saturation**.
- Heatmaps are great for **visualizing multivariate data** (data in which analysis is based on more than two variables per observation).
- In heatmaps **categorical variables are placed in the rows and columns** and **a numerical or categorical variable is represented as colors** or **color saturation**.

Use

- The visualization of multivariate data can be done using heatmaps as they are great for finding patterns in your data.

Example:

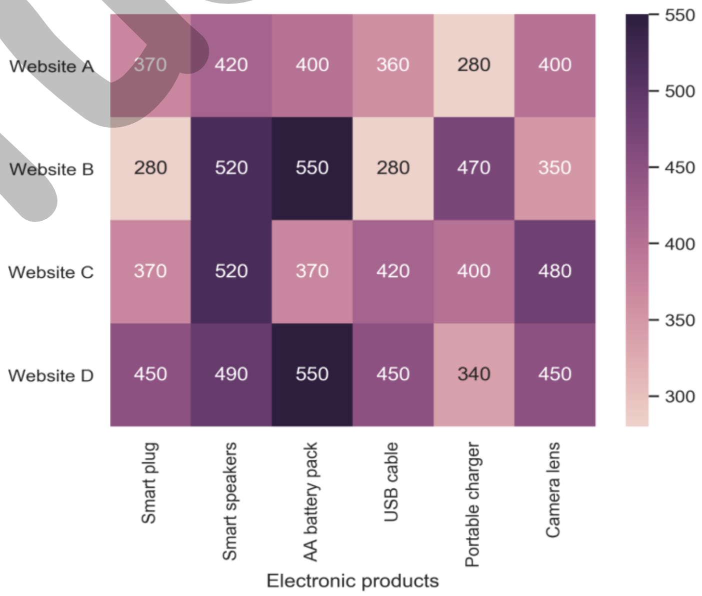
The following diagram shows a heatmap for the most popular products on the electronics category page across various e-commerce websites, where the color shows the number of units sold. In the following diagram, we can analyze that the darker colors represent more units sold, as shown in the key:



Heatmap for popular products in the electronics category

Variants: Annotated Heatmaps

Let's see the same example we saw previously in an annotated heatmap, where the color shows the number of units sold:



Questions:

1. Discuss various Relation plots.

2. Explain what Scatter Plot, Bubble Plot, Correlogram and Heatmap are. Also explain their uses and design practices with examples.

3. Explain the Scatter Plots with Marginal Histograms with an example.

4. Explain Heatmaps and its variant in detail.

Handouts for Session 5: Composition Plots: Pie Chart, Stacked Bar Chart, Stacked Area Chart, Venn Diagram

4.7 Composition Plots

- **Composition plots** are ideal **to represent some data as a part of a whole.**
- For **static data**, you can use **pie charts**, **stacked bar charts**, or **Venn diagrams**.
- **Pie charts** or **donut charts** help show **proportions and percentages for groups**.
- If you need an additional dimension, **stacked bar charts** are great.
- **Venn diagrams** are the best way **to visualize overlapping groups**, where each group is represented by a circle.
- For **data that changes over time**, you can use either **stacked bar charts** or **stacked area charts.**

1. Pie Chart

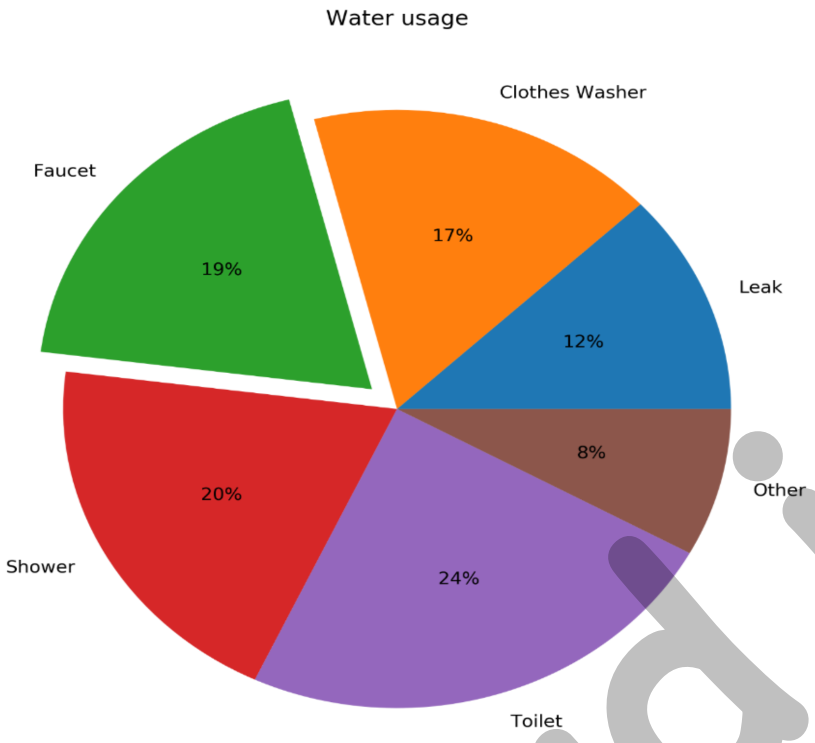
- Pie charts illustrate numerical proportions by dividing a circle into slices.
- Each arc length represents a proportion of a category.
- The full circle equates to 100%.
- For humans, it is easier to compare bars than arc lengths; therefore, it is recommended to use bar charts or stacked bar charts the majority of the time.

Use

- To compare items that are part of a whole.

Examples

The following diagram shows household water usage around the world:



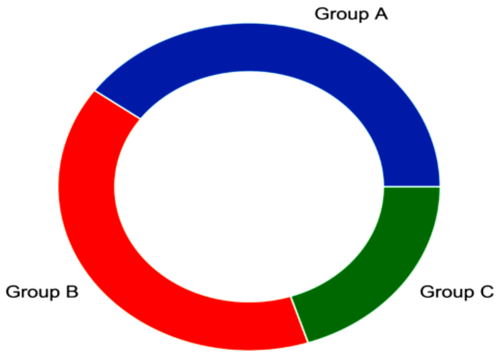
Pie chart for global household water usage

Design Practices

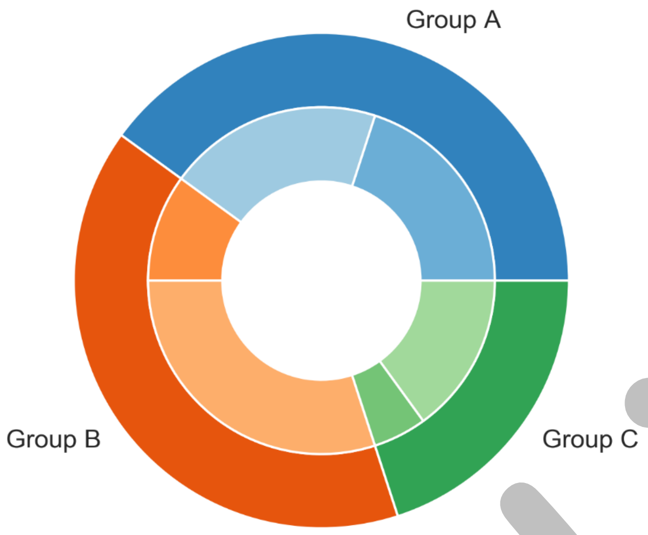
1. Arrange the slices according to their size in increasing/decreasing order, either in a clockwise or counter clockwise manner.
2. Make sure that every slice has a different color.

Variants: Donut Chart

- An alternative to a pie chart is a **donut chart.**
- In **contrast to pie charts**, it is **easier to compare the size of slices**, since the **reader focuses more on reading the length of the arcs** instead of the area.
- Donut charts are also more space-efficient because the center is cut out, so it can be used to display information or further divide groups into subgroups.
- The following diagram shows a basic donut chart:



The following diagram shows a donut chart with subgroups:



Design Practice

1. Use the same color that's used for the category for the subcategories.
2. Use varying brightness levels for the different subcategories.

2. Stacked Bar Chart

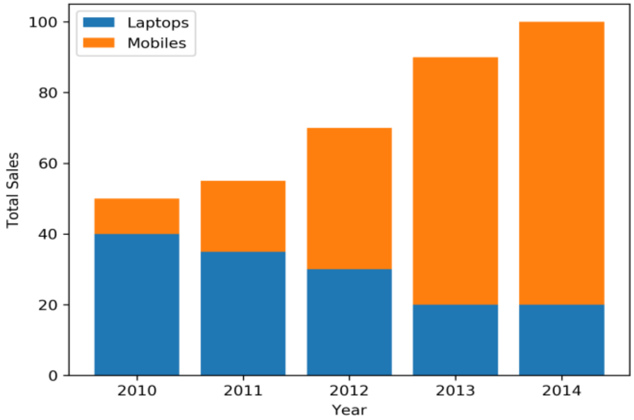
- **Stacked bar charts** are used to show how a category is divided into subcategories and the proportion of the subcategory in comparison to the overall category.
- Total amounts can be compared across each bar, or the percentage of each group can be displayed.
- The latter is also referred to as a **100% stacked bar chart** and makes it easier to see relative differences between quantities in each group.

Use

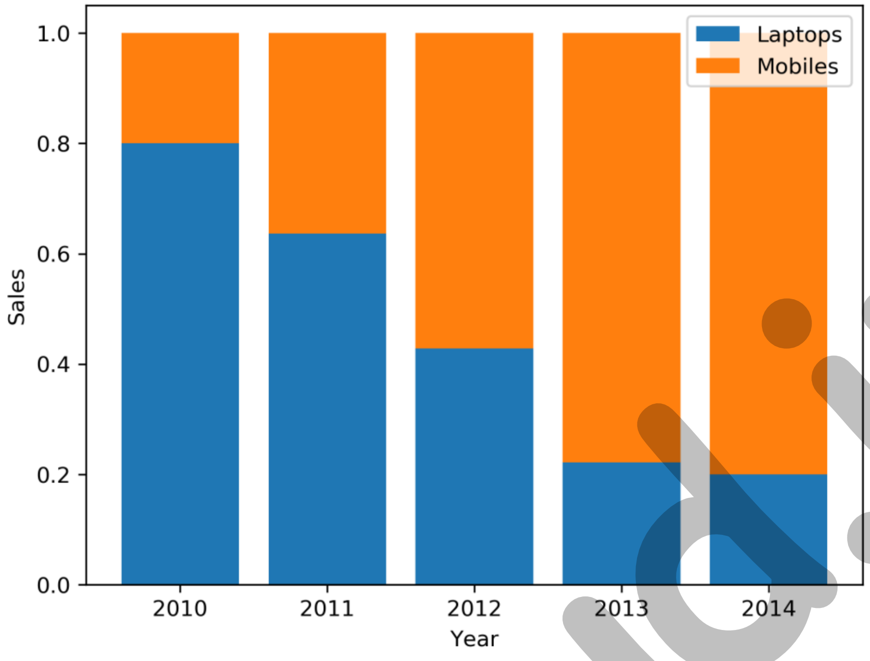
- To compare variables that can be divided into sub-variables.

Example 1:

The following diagram shows a generic stacked bar chart with five groups: Stacked bar chart to show sales of laptops and mobiles



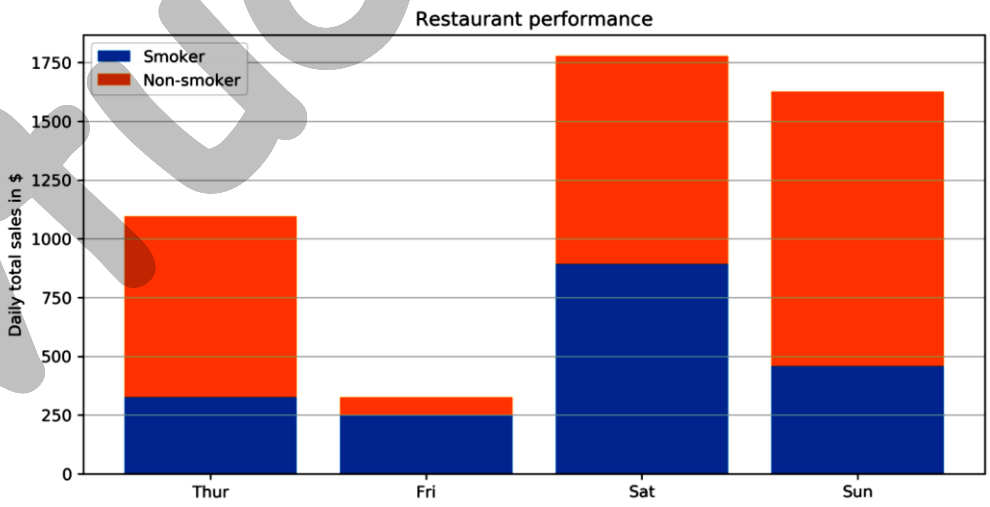
The following diagram shows a 100% stacked bar chart with the same data that was used in the preceding diagram:



100% stacked bar chart to show sales of laptops, PCs, and mobiles

Example 2:

The following diagram illustrates the daily total sales of a restaurant over several days. The daily total sales of non-smokers are stacked on top of the daily total sales of smokers:



Daily total restaurant sales categorized by smokers and non-smokers

Design Practices

- Use contrasting colors for stacked bars.
- Ensure that the bars are adequately spaced to eliminate visual clutter.
- The ideal space guideline between each bar is half the width of a bar.
- Categorize data alphabetically, sequentially, or by value, to uniformly order it and make things easier for your audience.

3. Stacked Area Chart

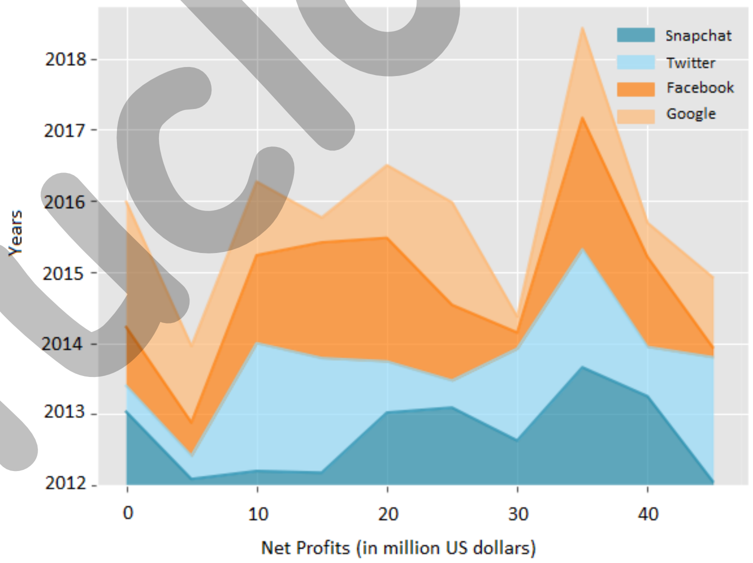
- Stacked area charts show trends for part-of-a-whole relations.
- The values of several groups are illustrated by stacking individual area charts on top of one another.
- It helps to analyze both individual and overall trend information.

Use

- To show trends for time series that are part of a whole.

Examples

The following diagram shows a stacked area chart with the net profits of Google, Facebook, Twitter, and Snapchat over a decade:



Stacked area chart to show net profits of four companies

Design Practice

1. Use transparent colors to improve information visibility.
2. This helps in analyzing overlapping data and makes the grid lines visible.

4. Venn Diagram

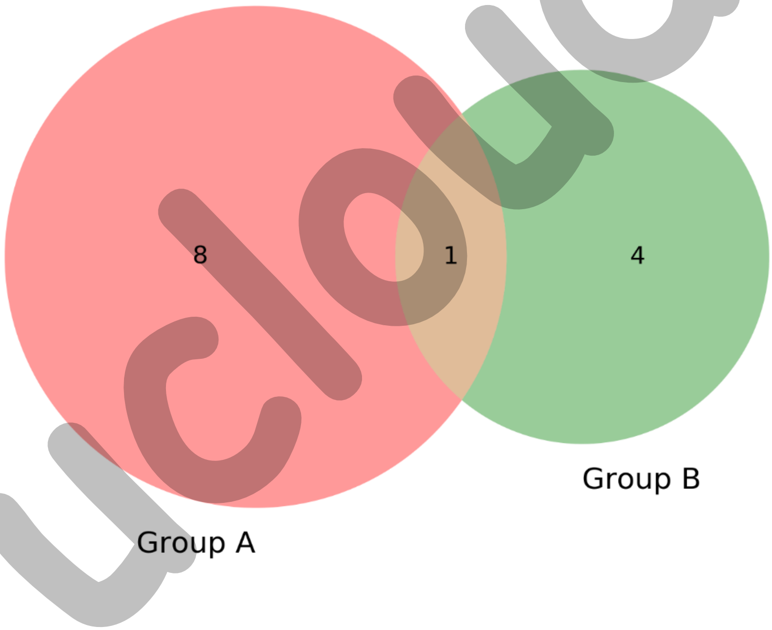
- **Venn diagrams**, also known as **set diagrams**, show all possible logical relations between a finite collection of different sets.
- Each set is represented by a circle.
- The **circle size illustrates the importance of a group**.
- The size of overlap represents the intersection between multiple groups.

Use

- ✓ To show overlaps for different sets.

Example

Visualizing the intersection of the following diagram shows a Venn diagram for students in two groups taking the same class in a semester:



Venn diagram showing students taking the same class

From the preceding diagram, we can note that there are eight students in just group A, four students in just group B, and one student in both groups.

Design Practice

1. It is not recommended to use Venn diagrams if you have more than three groups. It would become difficult to understand.
2. Moving on from composition plots, we will cover distribution plots in the following section.

Questions

1. Discuss various Composition plots.

2. Explain in detail Pie Charts. How Donut Chart can be more convenient than Pie Chart?

3. Explain the Stacked Bar Charts with an example. Also explain the uses and the design practices to ne followed.

4. Compare Stacked Bar Charts and Stacked Area Charts.

5. Discuss Venn Diagram with an example.

Handouts for Session 6: Distribution Plots: Histogram, Density Plot, Box Plot, Violin Plot

4.8 Distribution Plots

- Distribution plots give a deep insight into how your data is distributed.
- For a single variable, a histogram is effective.
- For multiple variables, you can either use a box plot or a violin plot.
- The violin plot visualizes the densities of your variables, whereas the box plot just visualizes the median, the interquartile range, and the range for each variable.

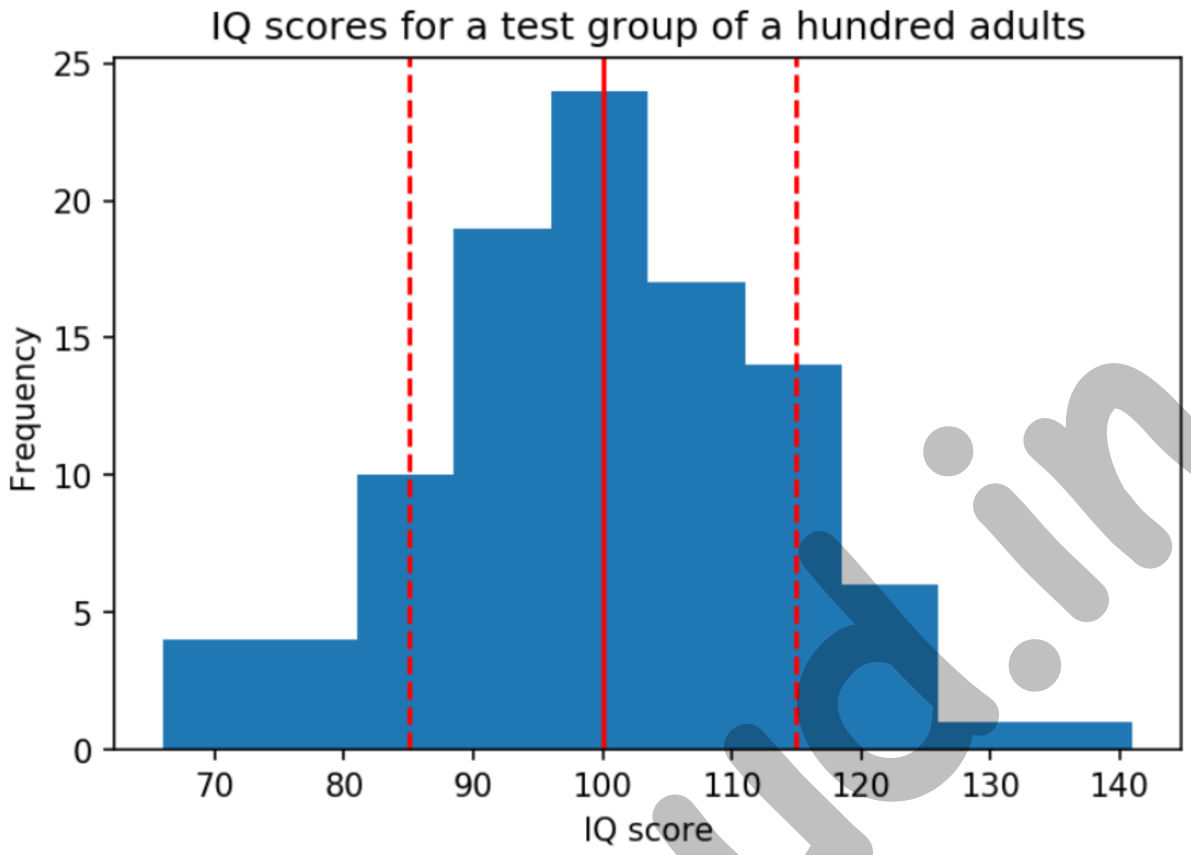
1. **Histogram**
   - A histogram visualizes the distribution of a single numerical variable.
   - Each bar represents the frequency for a certain interval.
   - Histograms provide an estimate of statistical measures, revealing where values are concentrated and making it easy to detect outliers.
   - A histogram can be plotted using absolute frequency values or, alternatively, by normalizing the values.
   - Different colors for the bars can be used to compare distributions of multiple variables.

Use

- Get insights into the underlying distribution for a dataset.

Example

The following diagram shows the distribution of the **Intelligence Quotient** (**IQ**) for a test group. The dashed lines represent the standard deviation each side of the mean (the solid line):



Distribution of IQ for a test group of a hundred adults

Design Practices

1. Try different numbers of bins (data intervals), since the shape of the histogram can vary significantly.

2. Density Plot

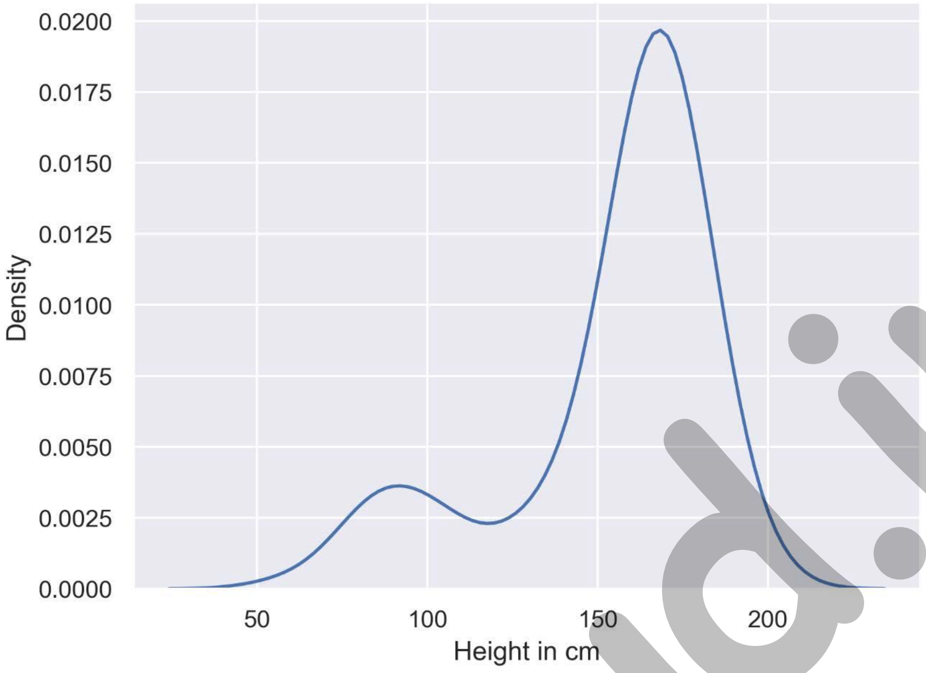
- A density plot shows the distribution of a numerical variable.
- It is a variation of a histogram that uses kernel smoothing, allowing for smoother distributions.
- One advantage these have over histograms is that density plots are better at determining the distribution shape since the distribution shape for histograms heavily depends on the number of bins (data intervals).

Use

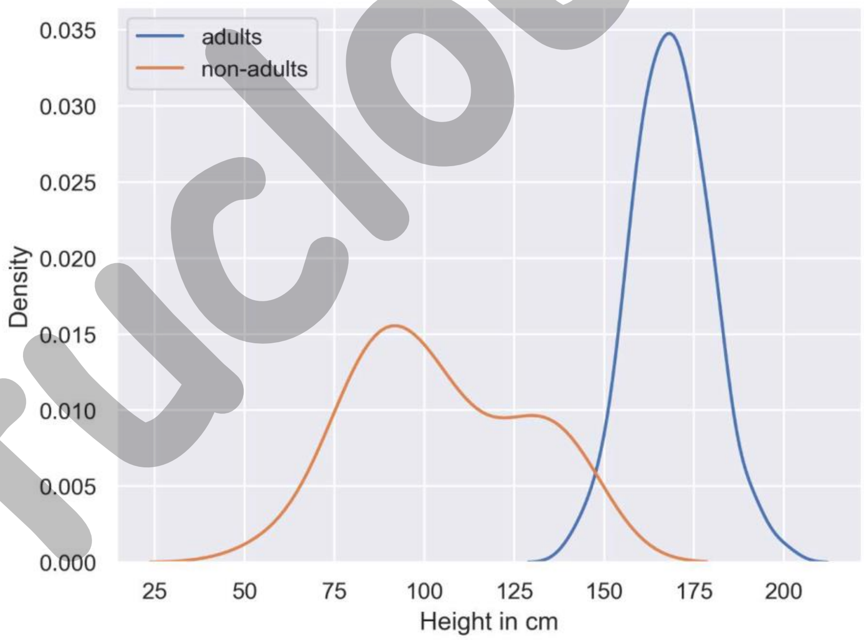
- To compare the distribution of several variables by plotting the density on the same axis and using different colors.

Example

The following diagram shows a basic density plot:



The following diagram shows a basic multi-density plot:



Design Practices

1. Use contrasting colors to plot the density of multiple variables.

3. Box Plot

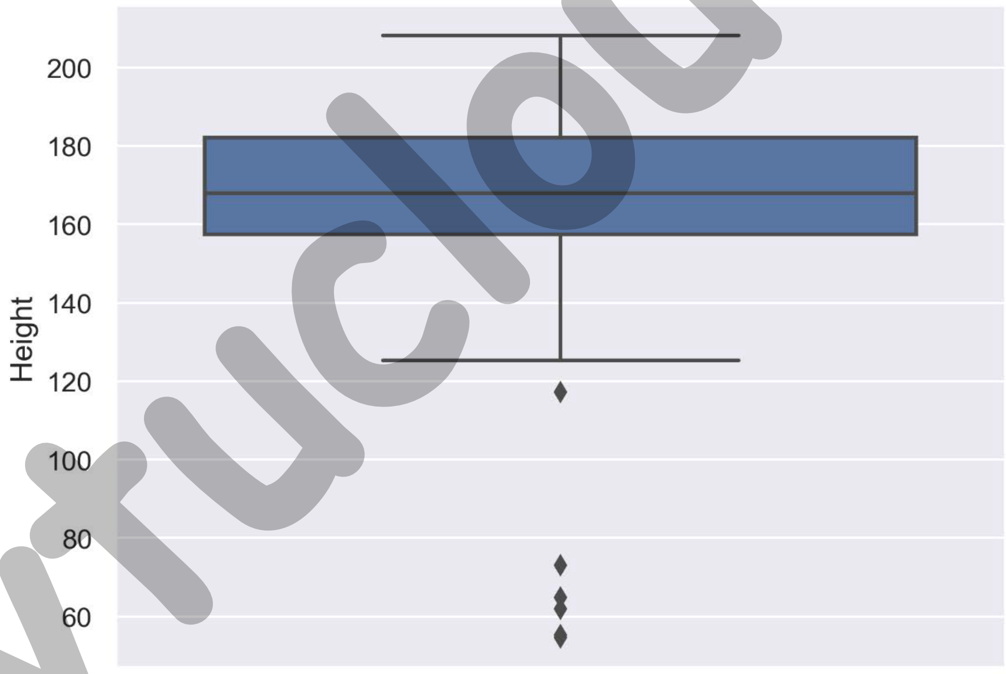
- The box plot shows multiple statistical measurements.
- The box extends from the lower to the upper quartile values of the data, thus allowing us to visualize the interquartile range (IQR).
- The horizontal line within the box denotes the median.
- The parallel extending lines from the boxes are called whiskers; they indicate the variability outside the lower and upper quartiles.
- There is also an option to show data outliers, usually as circles or diamonds, past the end of the whiskers.

Use

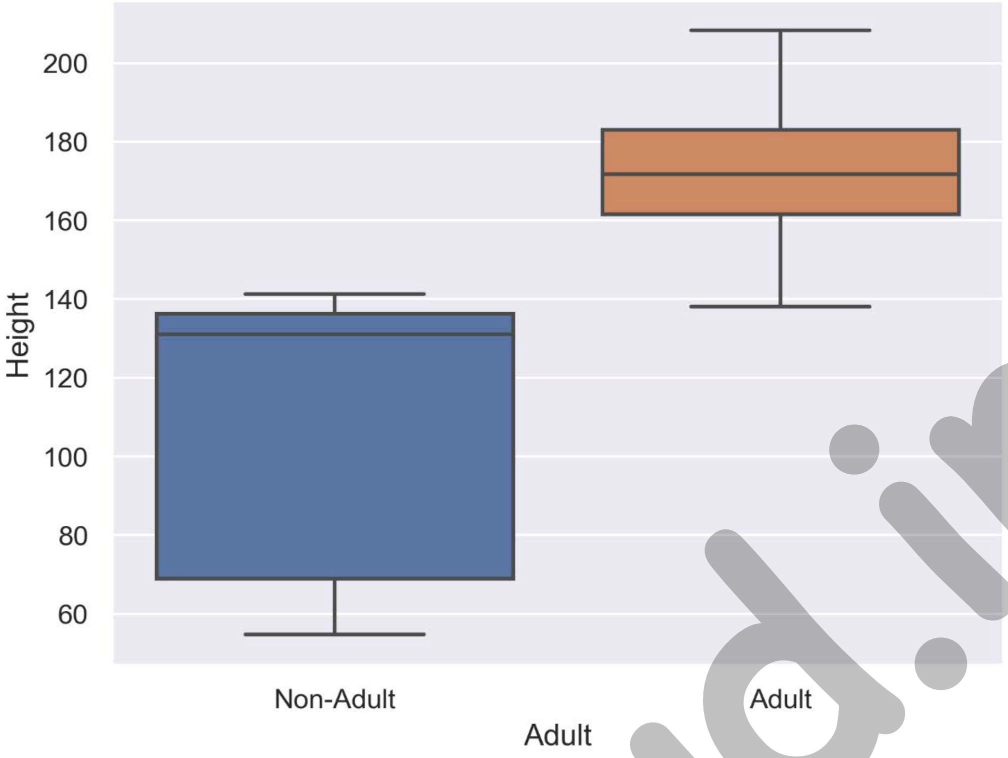
✓ Compare statistical measures for multiple variables or groups.

Example

The following diagram shows a basic box plot that shows the height of a group of people:



The following diagram shows a basic box plot for multiple variables. In this case, it shows heights for two different groups – adults and non-adults:



4. Violin Plot

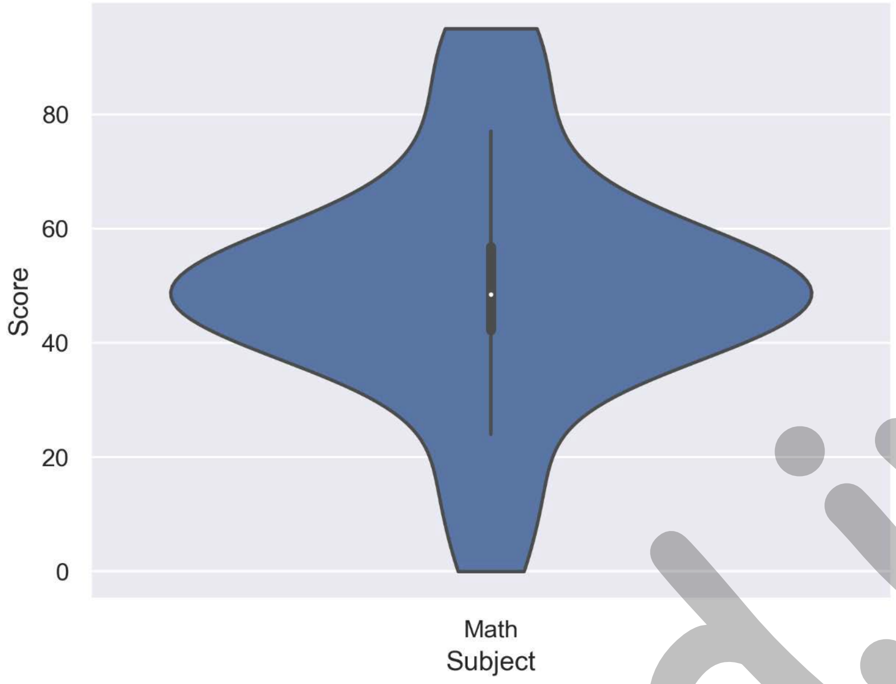
- Violin plots are a combination of box plots and density plots.
- Both the statistical measures and the distribution are visualized.
- The thick black bar in the center represents the interquartile range, while the thin black line corresponds to the whiskers in a box plot.
- The white dot indicates the median.
- On both sides of the centerline, the density is visualized.

Use

✓ Compare statistical measures and density for multiple variables or groups.

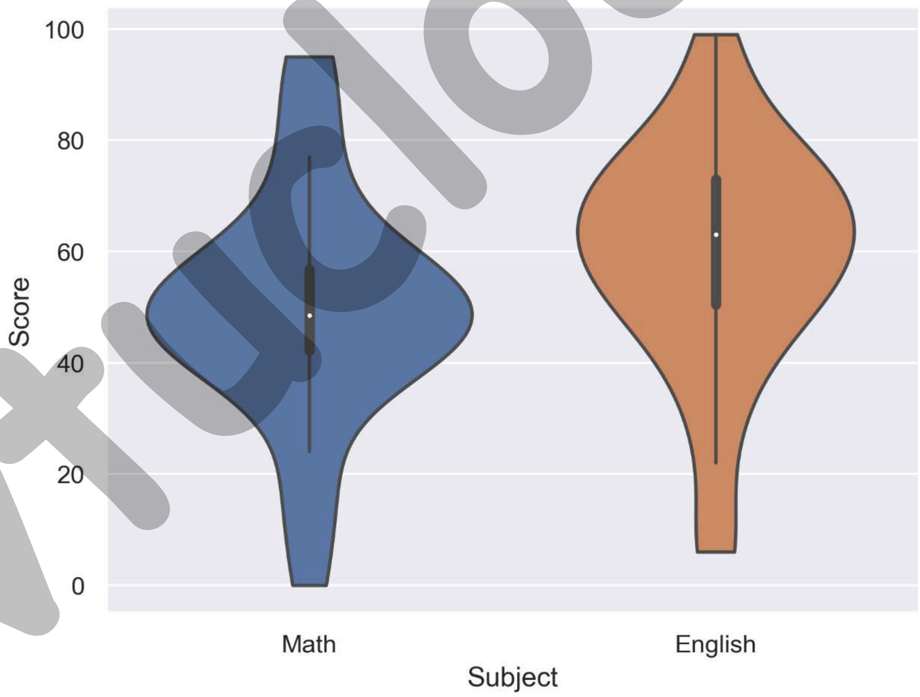
Example

The following diagram shows a violin plot for a single variable and shows how students have performed in Math: From the diagram, we can analyze that most of the students have scored around 40-60 in the Math test.



Violin plot for a single variable (Math)

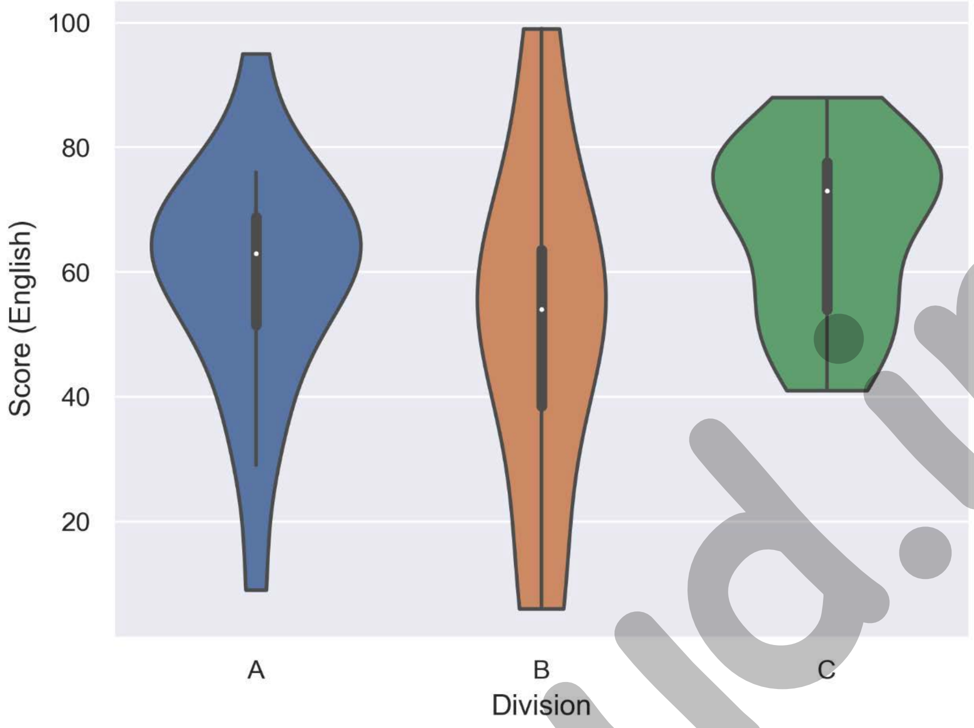
The following diagram shows a violin plot for two variables and shows the performance of students in English and Math:



Violin plot for multiple variables (English and Math)

From the preceding diagram, we can say that on average, the students have scored more in English than in Math.

The following diagram shows a violin plot for a single variable divided into three groups, and shows the performance of three divisions of students in English based on their score:



Violin plot with multiple categories (three groups of students)

From the preceding diagram, we can note that on average, division C has scored the highest, division B has scored the lowest, and division A is, on average, in between divisions B and C.

Design Practices

1. Scale the axes accordingly so that the distribution is clearly visible and not flat.

Questions

1. Discuss various Distribution plots.

2. Explain in detail Histogram Plots in detail with example.

3. Explain the Density Plots how is it different from Histogram Plots.

4. Compare Box Plots and Violin Plots.

Handouts for Session 7: Geo Plots: Dot Map, Choropleth Map, Connection Map

4.9 Geo Plots

- ✓ Geological plots are a great way to visualize geospatial data.
- ✓ Choropleth maps can be used to compare quantitative values for different countries, states, and so on.
- ✓ Connections between different locations can be represented using connection maps.

1. Dot Map

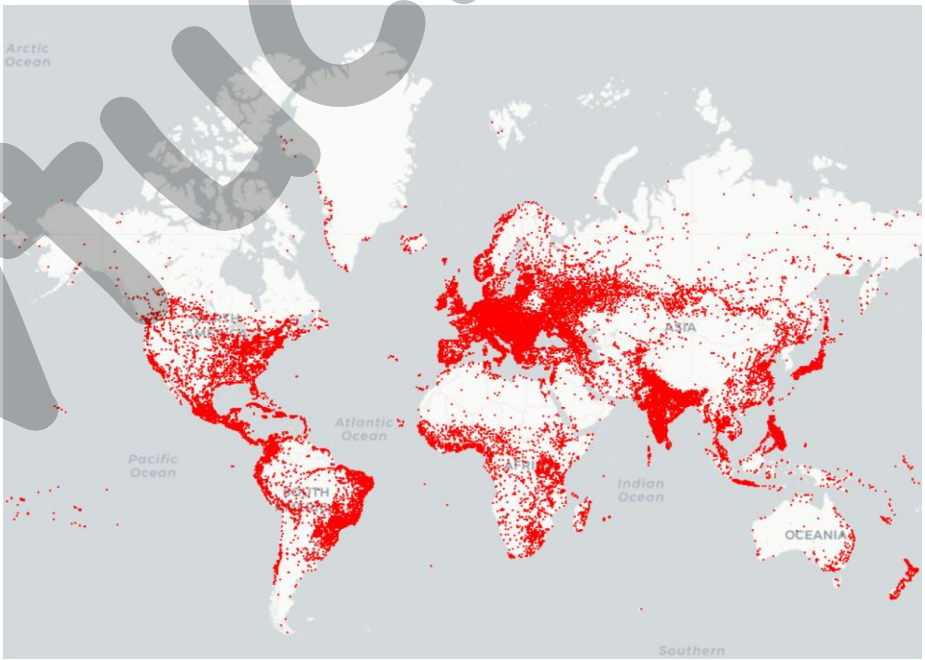
- In a dot map, each dot represents a certain number of observations.
- Each dot has the same size and value (the number of observations each dot represents).
- The dots are not meant to be counted; they are only intended to give an impression of magnitude.
- The size and value are important factors for the effectiveness and impression of the visualization.
- Different colors or symbols can be used for the dots to show multiple categories or groups.

Use

- ✓ To visualize geospatial data.

Example

The following diagram shows a dot map where each dot represents a certain amount of bus stops throughout the world:



Dot map showing bus stops worldwide

Design Practices

1. Avoid displaying too many locations to ensure the map remains clear and the actual locations are discernible.
2. Select an appropriate dot size and value to ensure that in dense areas, the dots blend together, providing a clear impression of the underlying spatial distribution.

2. Choropleth Map

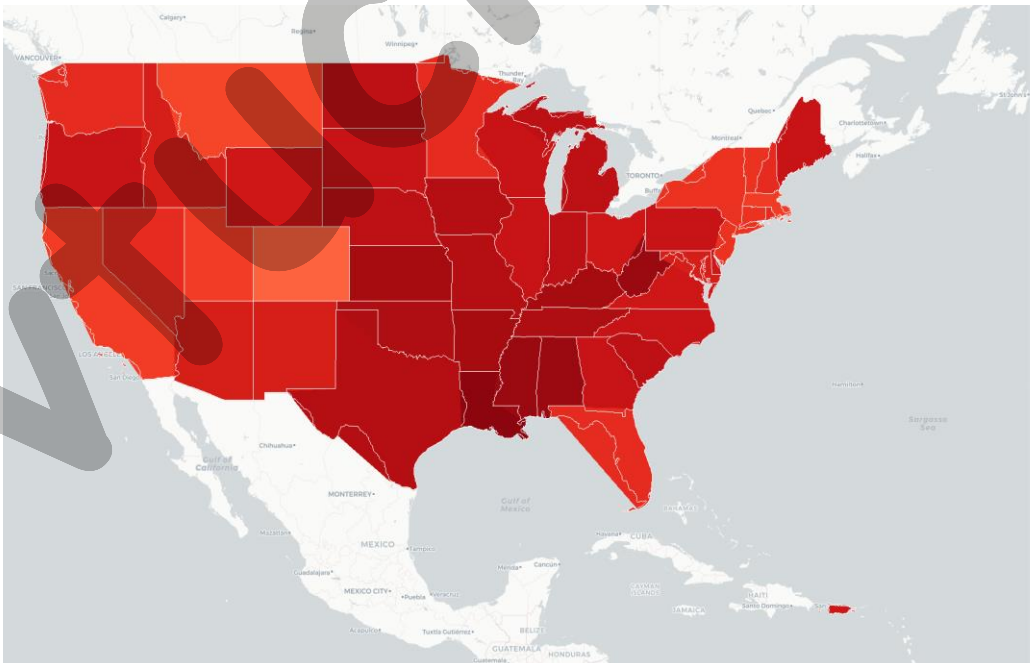
- In a choropleth map, each tile is colored to encode a variable.
- For example, a tile represents a geographic region for counties and countries.
- Choropleth maps provide a good way to show how a variable varies across a geographic area.
- One thing to keep in mind for choropleth maps is that the human eye naturally gives more attention to larger areas, so you might want to normalize your data by dividing the map area-wise.

Use

✓ To visualize geospatial data grouped into geological regions—for example, states or countries.

Example

The following diagram shows a choropleth map of a weather forecast in the USA:



Choropleth map showing a weather forecast for the USA

Design Practices

1. Use darker colors for higher values, as they are perceived as being higher in magnitude.
2. Limit the color gradation, since the human eye is limited in how many colors it can easily distinguish between. Seven color gradations should be enough.

3. Connection Map

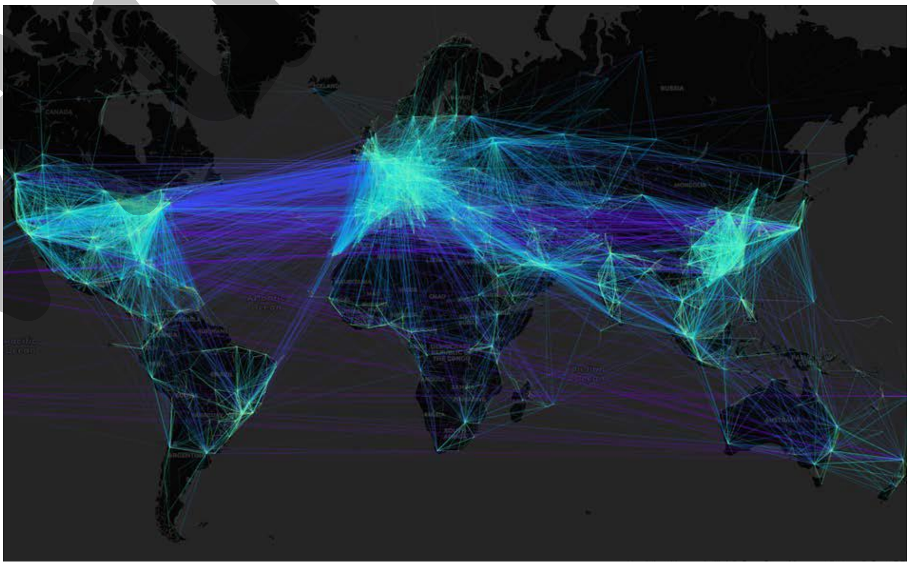
- In a **connection map**, each line represents a certain number of connections between two locations.
- The link between the locations can be drawn with a straight or rounded line, representing the shortest distance between them.
- Each line has the same thickness and value (the number of connections each line represents).
- The lines are not meant to be counted; they are only intended to give an impression of magnitude.
- The size and value of a connection line are important factors for the effectiveness and impression of the visualization.

Use

✓ To visualize connections.

Example

The following diagram shows a connection map of flight connections around the world:



Design Practices

1. Avoid displaying too many connections, as it can make data analysis challenging. Ensure the map remains clear enough to identify the actual locations of the start and end points.
2. Choose a line thickness and value so that the lines start to blend in dense areas. The connection map should give a good impression of the underlying spatial distribution.

Questions:

1. Discuss various Geo plots.
2. Explain in detail Dot Plots in detail with example.
3. Explain the uses of Choropleth Map and what are the design practices to be followed.

Handouts for Session 8: What Makes a Good Visualization?

4.10 What Makes a Good Visualization?

There are multiple aspects to what makes a good visualization:

1. Most importantly, the visualization should be self-explanatory and visually appealing. To make it self-explanatory, use a legend, descriptive labels for your x-axis and y-axis, and titles.
2. A visualization should tell a story and be designed for your audience. Before creating your visualization, think about your target audience; create simple visualizations for a non-specialist audience and more technical detailed visualizations for a specialist audience. Think about a story to tell with your visualization so that your visualization leaves an impression on the audience.

Common Design Practices

- ✓ Use colors to differentiate variables/subjects rather than symbols, as colors are more perceptible.
- ✓ To show additional variables on a 2D plot, use color, shape, and size.
- ✓ Keep it simple and don't overload the visualization with too much information.