Lecture 6

Data Exploration

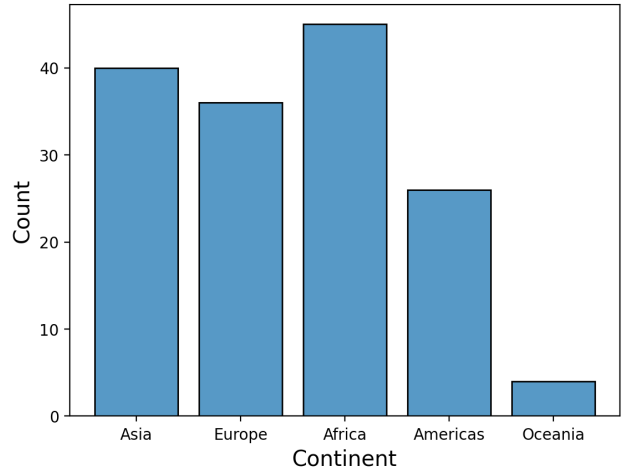
Visualizing a Categorical Feature

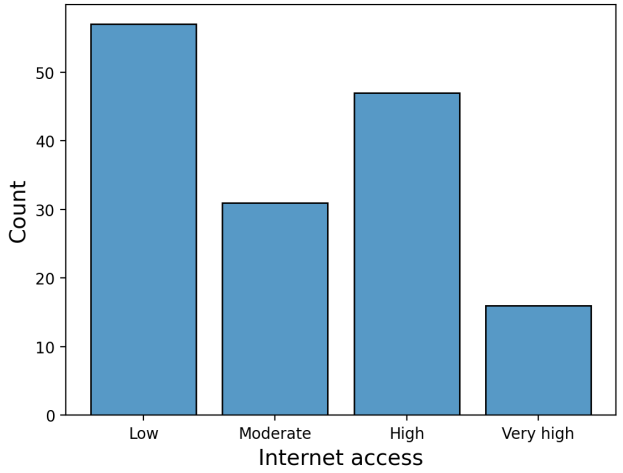
- This section discusses basic data visualization using categorical features.
- It explains how to represent item counts in different groups using bar charts.
- The example dataset used is from Gapminder 2017, including 151 countries and 7 features.
- The features are listed in a table.

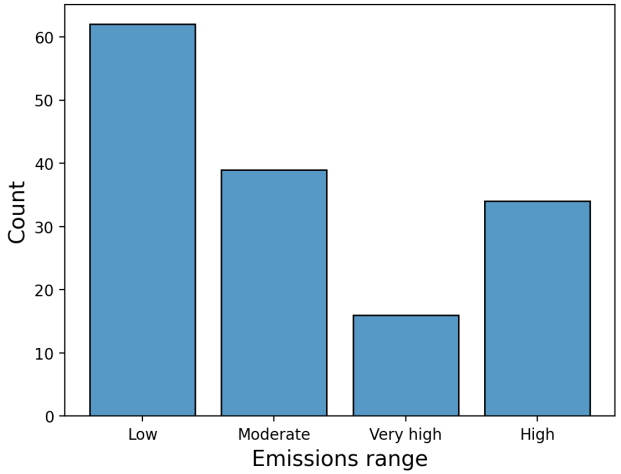
Feature	Type	Description
Years	Numerical	Years of schooling completed
Emissions	Numerical	CO2 emissions per person
Fertility	Numerical	Births per woman
Internet	Numerical	Percent of population with internet access
Continent	Categorical	Continent where the country is located
Internet access	Categorical	Low, Moderate, High, Very high
Emissions range	Categorical	Low, Moderate, High, Very high

Visualizing a Categorical Feature

Feature	Type	Description
Years	Numerical	Years of schooling completed
Emissions	Numerical	CO2 emissions per person
Fertility	Numerical	Births per woman
Internet	Numerical	Percent of population with internet access
Continent	Categorical	Continent where the country is located
Internet access	Categorical	Low, Moderate, High, Very high
Emissions range	Categorical	Low, Moderate, High, Very high







Visualizing a Numerical Feature

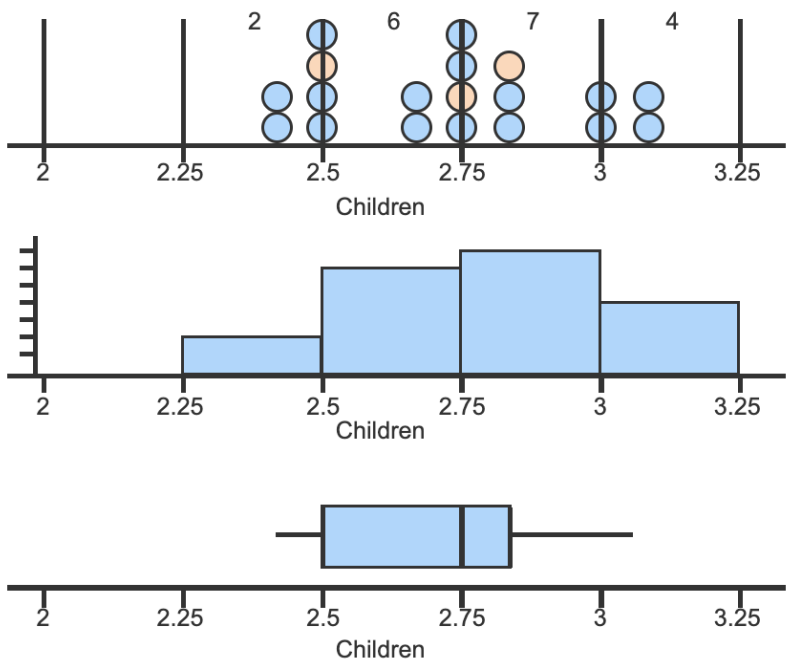
- A numerical feature consists of varying numbers spread across a wide range of values.
- When visualizing such a feature, the goal is to effectively convey this variability.
- Below are outlined several methods for visualizing numerical features.

Type	Description
Histogram	A bar chart that is created by dividing the numerical feature into small regions, or bins, and then counting the number of values in each region.
Density plot	A plot that approximates the density function of the distribution for the feature. Density plots can be thought of as a smoothed histogram. Usually this approximation is done by centering a small normal distribution over each data point and summing all these normal distributions to form the final density curve.
Box plot	A visual representation of the five-number summary; minimum, first quartile, median, third quartile, and maximum of a feature.

Visualizing a Numerical Feature

1. The average number of children born to each woman is shown for a sample of countries.
2. A dot plot is created by representing each country's average as a dot. Dots for countries with the same average are stacked.
3. To make a histogram, averages are placed into bins and counted. In this case, values on the boundary are rounded up.
4. Each bin is transformed into a bar that measures the number of countries in that bin.
5. The three quartiles, which cut the data into quarters, are needed to create a box plot.
6. The first quartile is the left edge of the box, the third quartile is the right edge. The median is the line in the middle. Whiskers extend to the minimum and maximum.

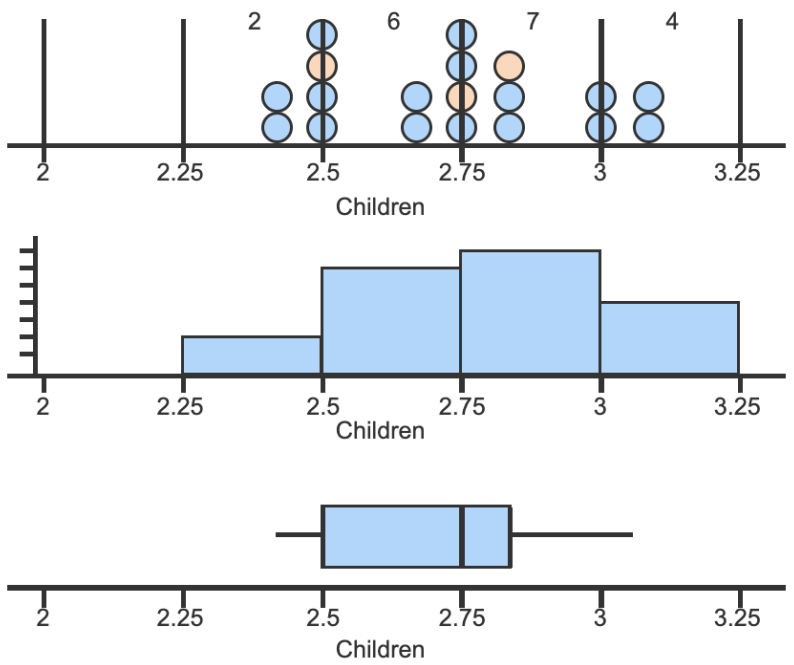
Children
2.5
2.67
2.42
2.5
2.5
2.42
2.08
...
2.75
2.08
2.83
2.67



Visualizing a Numerical Feature

1. The average number of children born to each woman is shown for a sample of countries.
2. A dot plot is created by representing each country's average as a dot. Dots for countries with the same average are stacked.
3. To make a histogram, averages are placed into bins and counted. In this case, values on the boundary are rounded up.
4. Each bin is transformed into a bar that measures the number of countries in that bin.
5. The three quartiles, which cut the data into quarters, are needed to create a box plot.
6. The first quartile is the left edge of the box, the third quartile is the right edge. The median is the line in the middle. Whiskers extend to the minimum and maximum.

Children
2.5
2.67
2.42
2.5
2.5
2.42
2.08
...
2.75
2.08
2.83
2.67



Single Feature Plots in R with `ggplot2`

- The base R's `plot()` function generates plots for numerical and categorical features but lacks easy customization.
- `ggplot2` from `tidyverse` is a favored alternative.
- It follows The Grammar of Graphics, building layered plots.
- Starting with x-axis setup (`ggplot(data, aes(x=x))`), subsequent layers are added with +.
- Plot types are defined by geometries, e.g., `geom_histogram()` for histograms.
- Features changing based on data are aesthetics, like `color`, `fill`, and `pch`.
- Further details are in `ggplot2 docs`.

Visualizing Two Numerical Features

- Single feature visualization uses axes for value and frequency.
- For multiple features, axes are dropped to highlight their relationship. This aids modeling and communication.
- For two numerical features, the common choice is a scatter plot. It places dataset instances as points in a 2D plane, using feature values as coordinates.

Visualizing Two Categorical Features

- A single categorical feature's plot employs axes to segregate categories and show relative frequency.
- Plotting two categorical features enhances information by grouping one based on the other.
- For instance, categorizing male and female customers by their visiting day.

Visualizing Two Categorical Features

1. The cross tabulation relates the categorical features day and smoker. Bar charts visualize the relationship between categorical features.

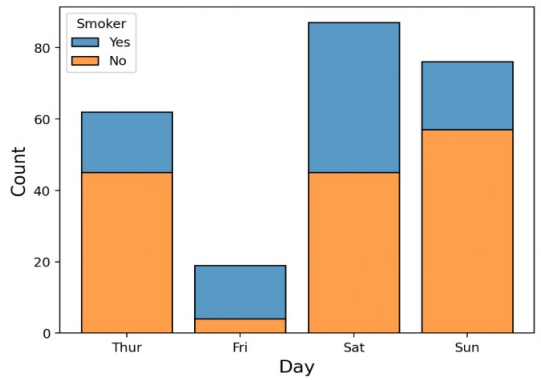
2. A stacked bar chart highlights the total for a category.

3. Segments within a stacked bar chart can be difficult to compare, because the groups often do not start at the same position.

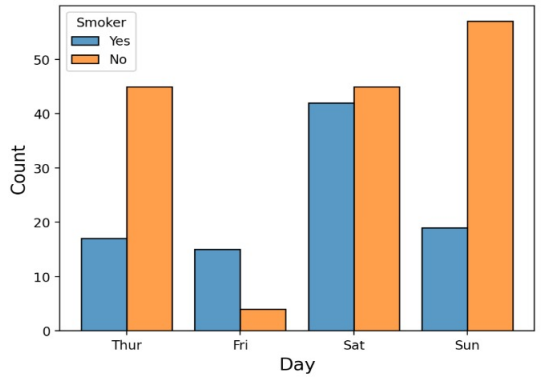
4. A grouped bar chart is helpful for comparisons within each day. A grouped bar chart easily shows that more smokers visited on Saturday.

Cross tabulation

Smoker	Yes	No
Day		
Thur	17	45
Fri	15	4
Sat	42	45
Sun	19	57



Did the restaurant have the most smokers on Thursday, Friday, or Sunday?



The restaurant had the most smokers on Sunday.

Visualizing More than Two Features

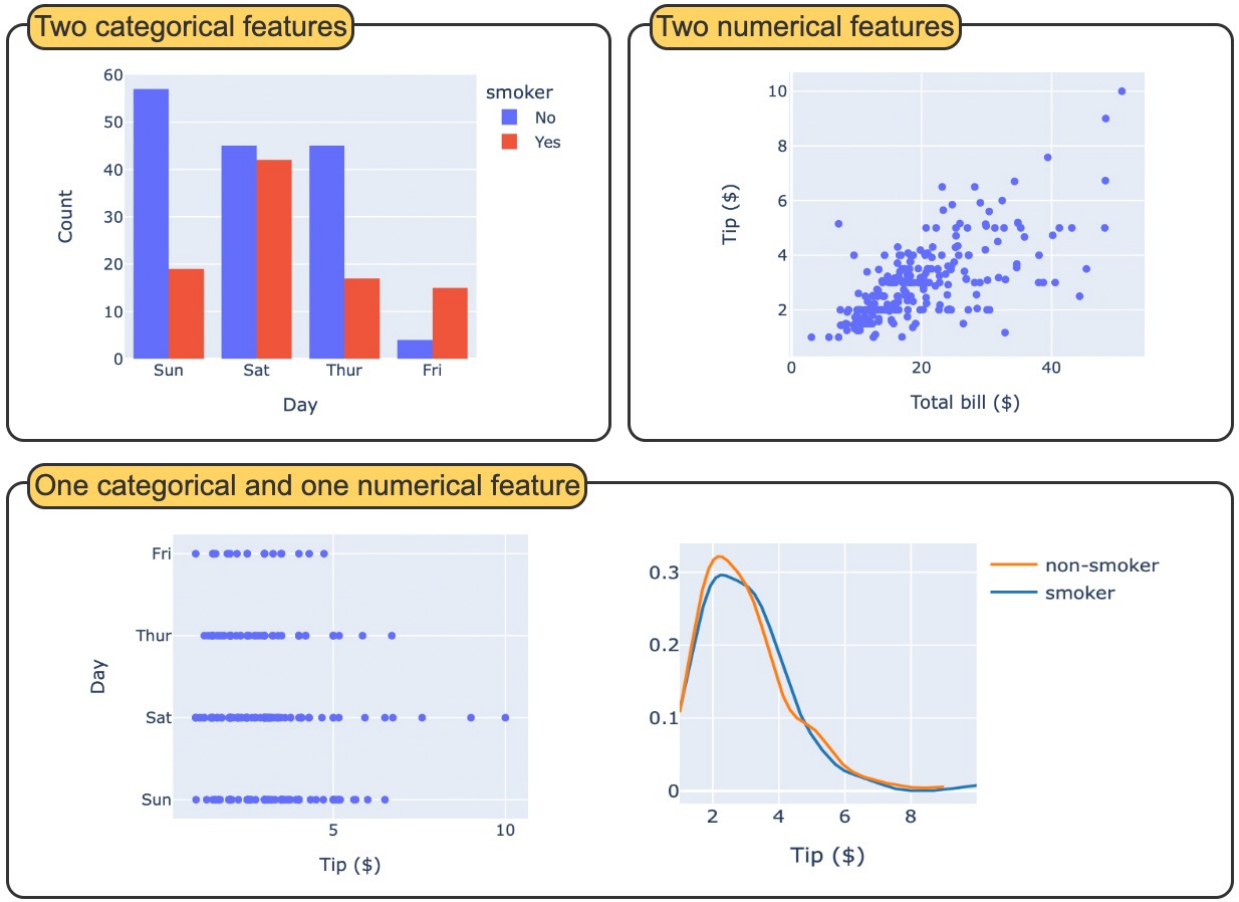
- When dealing with over two features, screens lack ample space for information display.
- Numerical attributes can leverage color, size, transparency, and animation.
- For categorical features, options include color, shading, size, transparency, shape, and faceting.
- Faceting involves arranging multiple plots in an array, with one changing feature across them.

Choosing Base Visualizations

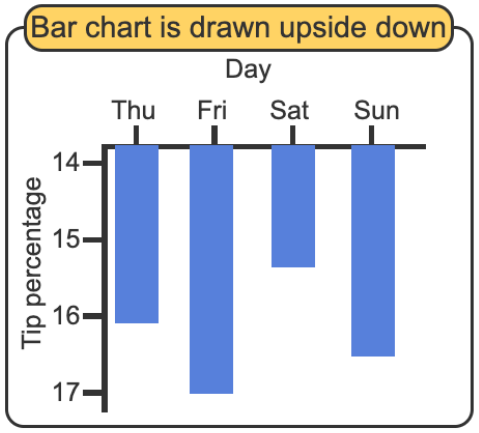
- The features' measurement levels largely dictate the appropriate choice.
- Data scientists must consider the audience and plot narrative.
- Simplicity can outweigh complex visuals.
- The chosen visualization norms depend on the viewer's characteristics.

Tips for Selecting Plots

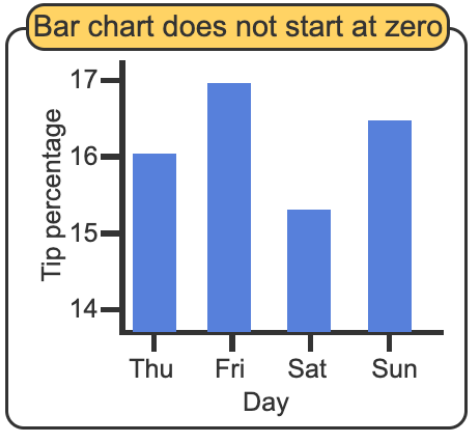
1. If both features are categorical, a bar chart is appropriate.
2. If both features are numerical, a scatter plot is appropriate.
3. For mix of categorical and numerical, plot choice depends on the effect.



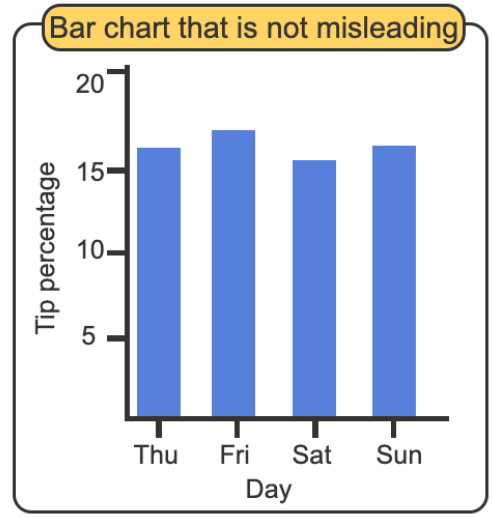
Important Caution: Scales



Vertical scale flipped in chart, distorting Saturday's perception. Flipping rectifies this.



Vertical scale lacks zero start; Friday appears twice better than Saturday.



Bar chart should be upright, avoid deceptive scales.

What Color is this Dress?

The dress was a 2015 online viral phenomenon.

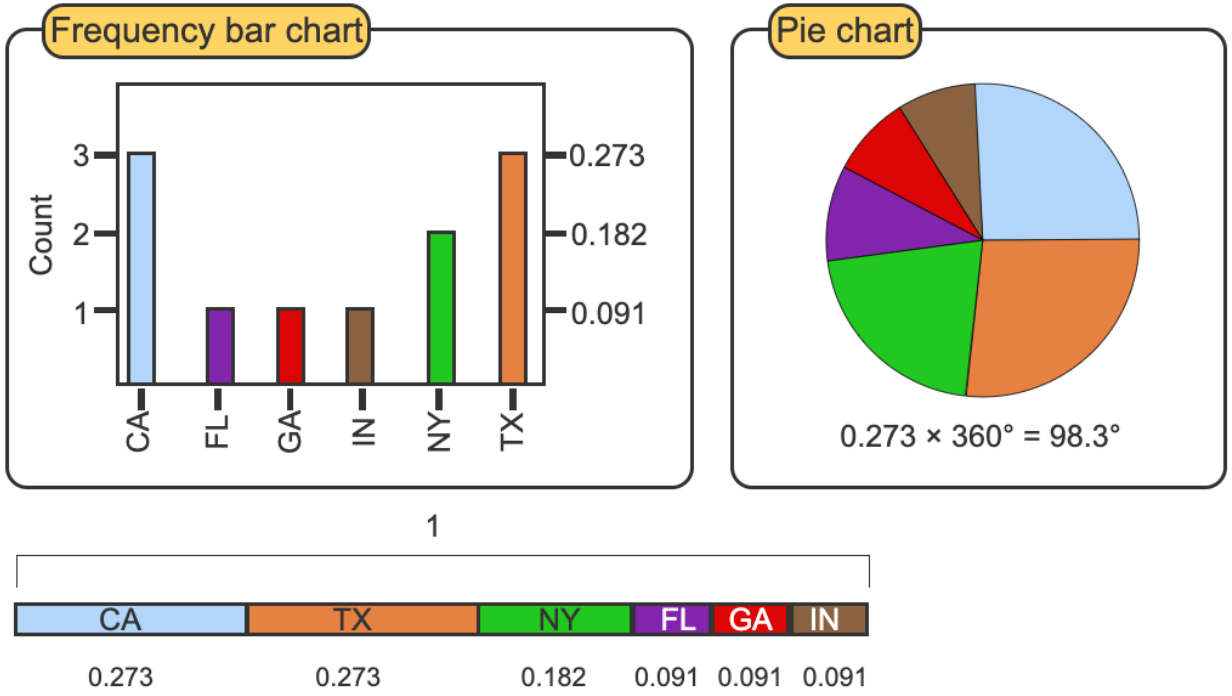
- Viral "The Dress" image debated color perception.
- Some saw white and gold, others blue and black.
- Illumination impact on visual interpretation explored.
- Highlights individual variation in color perception.

Important Caution: Color

- 8% males have color-vision deficiency (CVD), struggle with red-green distinction.
- Visualizations should avoid red-green-only distinctions due to common CVD.
- Contrast, varying brightness, crucial for enhancing color perception.
- Higher contrast aids differentiation in suboptimal conditions like bright sunlight or visual impairment.

Important Caution: Pie Charts

- Pie Chart: Circle divided into wedges for groups.
- Wedge sizes show group proportions.
- Despite aesthetics, pie charts poorly communicate information effectively.



Data Visualization Tools

- Various roles employ data visualization, including managers, analysts, and more.
- Diverse technologies support specific user needs and capabilities.
- Spreadsheets (Excel, Google Sheets, etc.) show data in grids, useful for quick visuals.
- Spreadsheets can be challenging to automate and prone to errors.
- Business intelligence apps (PowerBI, Tableau, etc.) offer reproducible visualizations and data pipelines, suited for dashboarding.
- Python: `Matplotlib`, `Seaborn`, `Plotly`
- R: `ggplot2`, `lattice`, `Shiny`

Python: matplotlib and seaborn

- matplotlib: Python library with MATLAB-like plotting.
- seaborn: Built on matplotlib, offers user-friendly plots.
- Supports pandas data structures, creates clean visuals.
- Various libraries based on matplotlib support maps and interfaces with toolkits and Excel.

R and Python: `plotly`

- `plotly`: Offers visualization services, open source libraries for Python, R, Julia.
- Uses `plotly.js` for interactive visualizations.
- Interactivity helps quickly find data points.
- Provides quicker insights than `ggplot2` or `matplotlib`.

Exploratory Data Analysis (EDA)

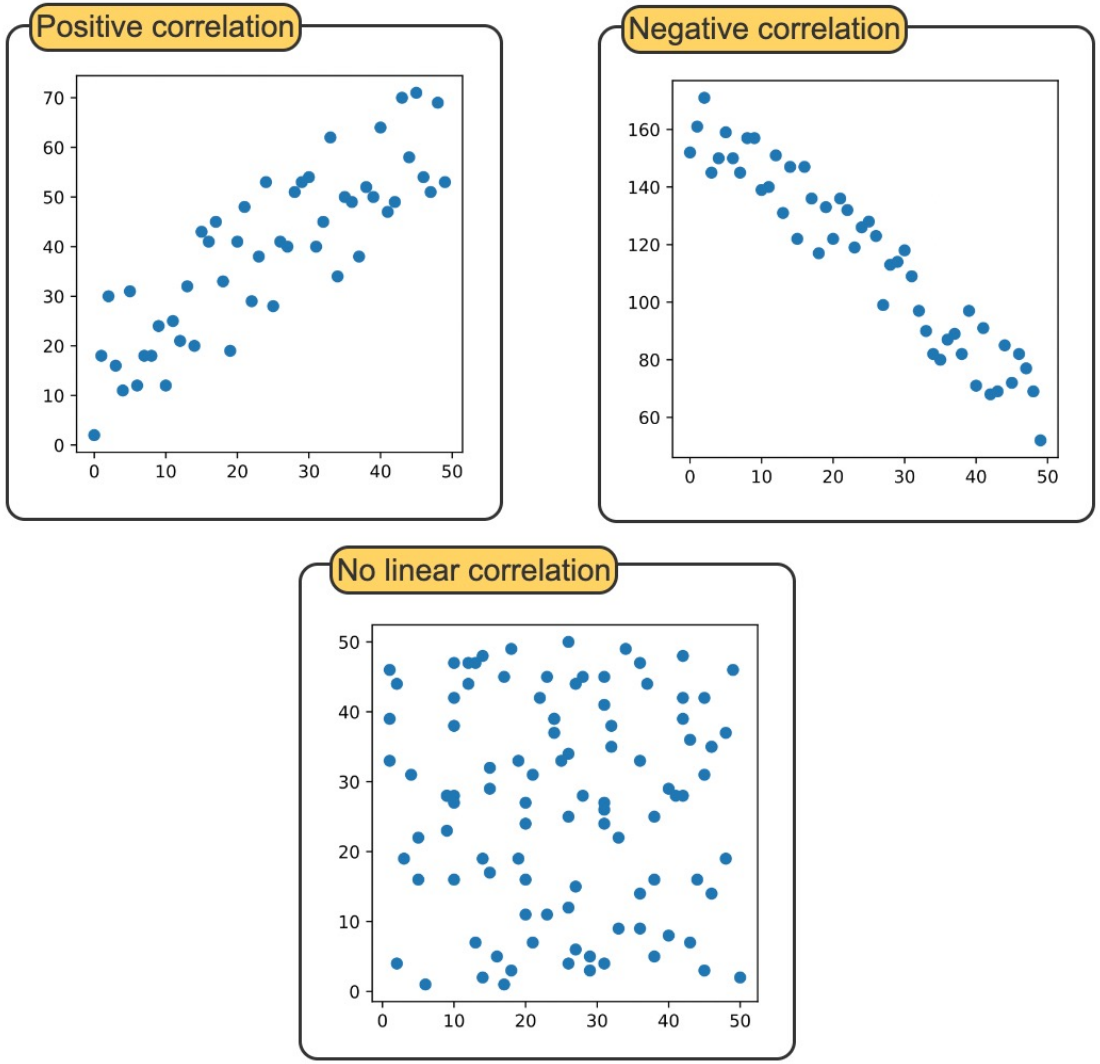
- Exploratory data analysis or EDA is the process of investigating a dataset to understand what is in the dataset.

Step	Description
Step 1: Understand the data	Find the size of the dataset (number of rows and columns), identify and categorize the features (categorical, numerical).
Step 2: Identify relationships between features	Find the direction (positive, negative) and strength (strong, moderate, weak) of correlation between the features.
Step 3: Describe the shape of data	Determine the shape of the distribution (symmetric, skewed).
Step 4: Detect outliers and missing data	Find values that are much higher or lower than the rest of the data or values that strongly affect the shape of the data.

Identifying Relationships Between Features

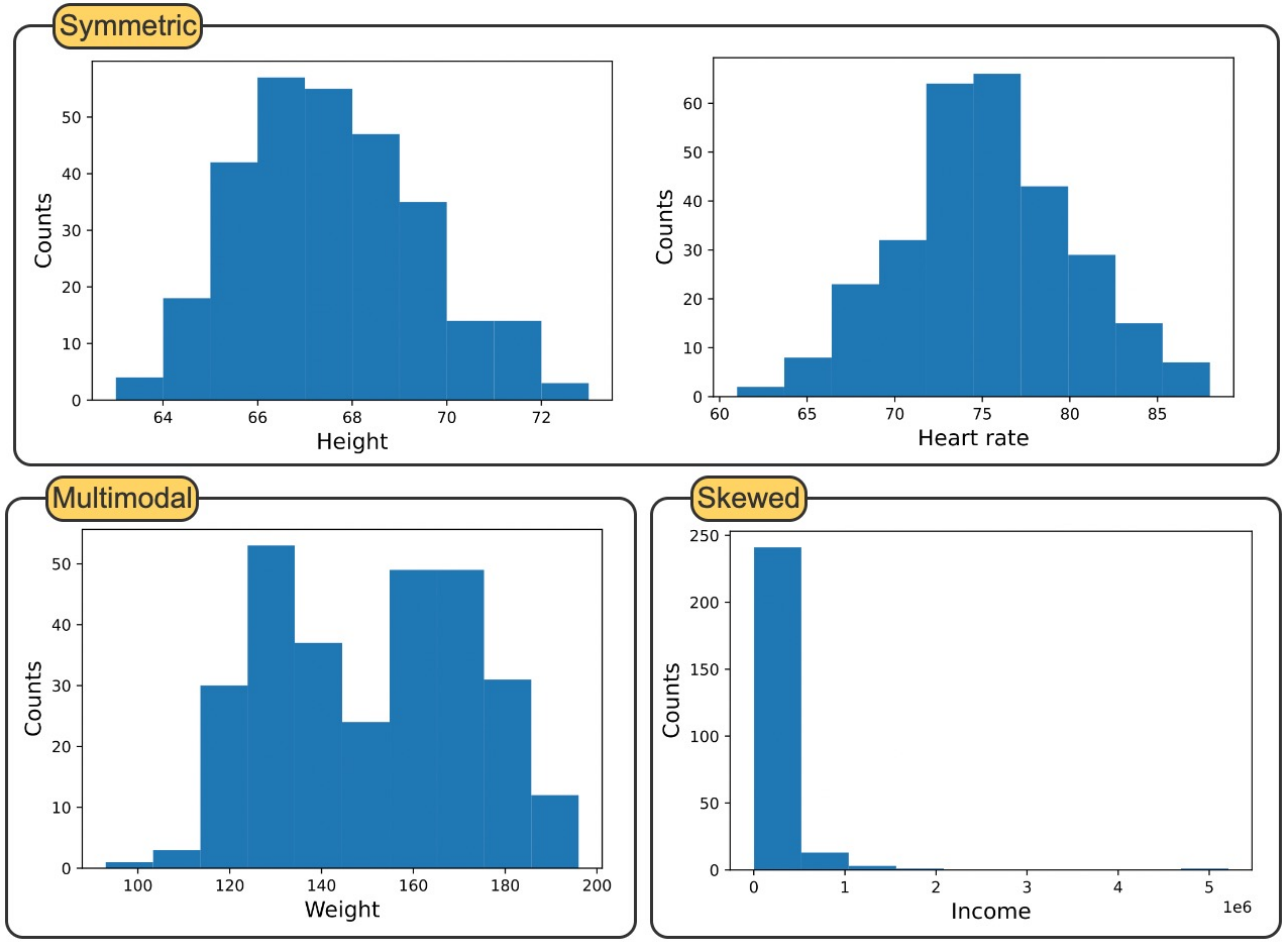
- Data scientists need to grasp feature interactions.
- Linear relationships are simplest; scatter plots help.
- Complex relationships arise, especially for categorical features.
- Linearly related features enable efficient predictions.

Identifying Relationships Between Features



Describing the Shape of Data

- Models often assume symmetric, unimodal feature distributions.
- Deviation impacts model validity.
- Identifying distribution shape validates assumptions.



Missing Data

- At times, datasets might contain missing values stemming from data entry problems, measurement errors, or processing glitches.
- These missing values can be categorized as:
    - Missing completely at random or **MCAR** have the same probability of being missing for all cases.
    - Missing at random or **MAR** have the same probability of being missing for specific observable cases.
    - Missing not at random or **MNAR** have different probabilities of being missing due to unknown reasons.

Exploratory Data Analysis in R

- Base R and `ggplot2` aid feature exploration.
- `ncol()`, `nrow()`, `type()`, and `class()` for `dataframe` dimensions and features.
- base R's `summary()` shows NA count for missing values.
- `ggplot2`'s `geom_boxplot()`, `geom_histogram()` analyze features' shape and outliers.

Outliers

- Outliers are isolated instances in a dataset.
- Isolation can occur in one or multiple dimensions.
- Outliers can disrupt models using averages.
- Identifying outliers is vital to gauge model impact.

Impact of Outliers on Models

- Outliers can heavily impact stats and models.
- Mean, std. dev. can be skewed by outliers.
- Linear models are influenced by outlier values.
- High leverage points can alter model parameters significantly.

Parametric Detection of Outliers

- Tukey's Fences, z-scores detect outliers effectively.
- Both methods rely on far distribution edges.
- Investigate outliers to understand causes, implications.
- Similar to missing data, outlier impact should be explored.

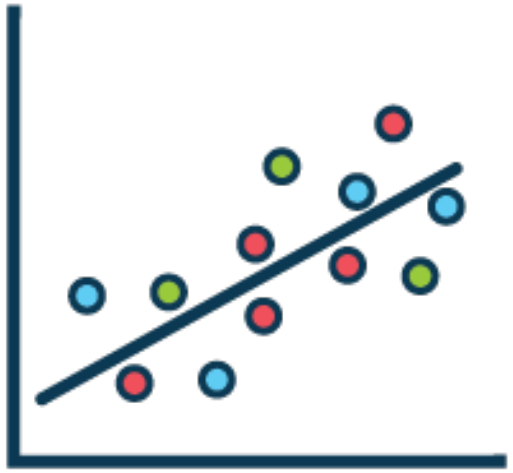
Method	Description
Tukey's Fences	Often used to determine outliers in box plots. 1. Calculate the interquartile range, $\text{IQR} = Q_3 - Q_1$ for a feature. 2. Classify all points that fall 1.5IQR above Q_3 or 1.5IQR below Q_1 as outliers.
z-scores	1. Calculate the z-score $z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$ for each value. 2. Classify all points that have a z-score of $\lvert z \rvert > 3$ as outliers.

Dealing with Outliers

- Pinpointing outlier cause guides data handling.
- Handle outliers carefully; not all need exclusion.
- Natural outliers exist, like solar energy patterns.
- Investigate before removing unexplained outliers.

Case Study

Palmer Penguins



Next Lecture

Regression